Desafios na avaliação conjunta do reconhecimento de entidades mencionadas:

O Segundo HAREM

Cristina Mota e Diana Santos
editoras

Linguateca, 2008

Desafios na avaliação conjunta do reconhecimento de entidades mencionadas:

O Segundo HAREM

Cristina Mota e Diana Santos
editoras

Linguateca, 2008

1ªEdição, Dezembro de 2008.

Publicação Digital.
ISBN: 978-989-20-1656-6

Prefácio

Após o encontro do Segundo HAREM realizado no dia 7 de Setembro de 2008, em Aveiro, demos início ao processo de edição da presente obra, que documenta o Segundo HAREM na perspectiva quer dos organizadores quer dos participantes.

À semelhança do que aconteceu com outros livros editados pela Linguateca, como seja o primeiro livro sobre avaliação conjunta em português, que inclui a descrição das primeiras Morfolimpíadas, e o livro que documenta o Primeiro HAREM, esse processo decorreu em várias fases, visando aumentar a qualidade científica do material que agora apresentamos. Assim, após termos recebido a primeira versão dos capítulos, demos início à sua revisão cruzada pelos outros autores. Convidámos outros investigadores não envolvidos no Segundo HAREM, constituindo uma comissão científica, para colaborar também nesse processo, de forma a que, cada capítulo fosse revisto por dois investigadores pelo menos. Em seguida, todos os capítulos foram revistos pelas editoras e, por fim, efectuámos as tarefas de edição propriamente dito.

No fim deste processo, não poderíamos estar mais gratas a todos os autores e membros da comissão científica, constituída por Alberto Simões, Cláudia Oliveira, Graça Nunes, Luís Costa, Mário J. Silva, Max Silberztein e Violeta Quental, pelas recensões construtivas que elaboraram e por terem cumprido a nossa sugestão de calendário apertado.

Dedicamos um agradecimento especial a Cláudia Freitas, Hugo Gonçalo Oliveira e Paula Carvalho pelo seu empenho entusiasmado e criativo na organização do Segundo HAREM, e a todos os participantes pelos muitos comentários, propostas e sugestões que deram sentido a esta avaliaçõa conjunta.

Estamos também gratas ao David Cruz pela sua colaboração inicial na organização, ao Paulo Rocha por ter dado início ao serviço de catalogação de publicações da Linguateca, e ao Luís Miguel Cabral por ter criado a partir daí o SUPeRB que facilitou muito o processo de elaboração da bibliografia.

Em meu nome individual, agradeço à Diana Santos por me ter dado não só a oportunidade de colaborar na organização do Segundo HAREM mas, sobretudo, por me ter convidado a editar com ela este livro, e por ter tido a paciência e disponibilidade que me permitiram concluir o doutoramento, o que, infelizmente, estendeu o processo de edição além do inicialmente previsto.

Embora de facto a versão final do livro só tenha ficado pronta em 2009, insistimos em publicá-lo com data de 2008 devido a nos termos comprometido a tal através da Linguateca (cujo financiamento terminou em Dezembro) e o termos publicitado junto dos autores com essa data.

Resta-nos agradecer a todos quantos tornam a Linguateca possível, e acusar com gratidão o financiamento recebido do governo português e da União Europeia, e executado

pela FCCN, através dos projectos POSC 339/1.3/C/NAC (2006-2008) e assim como o da UMIC, em 2009.

Esperamos que este livro possa constituir uma prova de que houve avanço substancial na área do reconhecimento de entidades mencionadas em relação ao anterior HAREM, assim como servir demonstração plena de que a Linguateca teve real utilidade e tem capacidade de organização de avaliações conjuntas. Desejamos, assim, que o conhecimento oferecido neste livro venha contribuir para a organização de um Terceiro HAREM, se tal se vier a considerar suficientemente útil.

Porto, Julho de 2009

As editoras,
Cristina Mota e Diana Santos

Lista de autores

Afonso Mendes Priberam Informática, Portugal.

António Teixeira Departamento de Electrónica, Telecomunicações e Informática, Instituto de Engenharia Electrónica e Telecomunicações de Aveiro, Universidade de Aveiro, Portugal.

Bruno Martins Departamento de Engenharia Informática, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal.

Carlos Amaral Priberam Informática, Portugal.

Caroline Hagège Xerox Research Centre Europe, Grenoble, França.

Cláudia Freitas Linguateca, Pólo de Coimbra, CISUC, Departamento de Engenharia Informática, Faculdade de Ciências e Tecnologia, Universidade de Coimbra, Portugal.

Cláudia Pinto Priberam Informática, Portugal.

Cristina Mota Linguateca/FCCN, Portugal.

Diana Santos Linguateca, SINTEF ICT, Noruega.

Helena Figueira Priberam Informática, Portugal.

Henrique Madeira CISUC, Departamento de Engenharia Informática, Universidade de Coimbra, Portugal.

Hugo Gonçalo Oliveira Linguateca, Pólo de Coimbra, CISUC, Departamento de Engenharia Informática, Faculdade de Ciências e Tecnologia, Universidade de Coimbra, Portugal.

João Paulo da Silva Cunha Departamento de Electrónica, Telecomunicações e Informática, Instituto de Engenharia Electrónica e Telecomunicações de Aveiro, Universidade de Aveiro, Portugal.

Joaquim Macedo Departamento de Informática, Universidade do Minho, Portugal.

Jorge Baptista Universidade do Algarve, Faro, Portugal & L2F - Spoken Language Systems Laboratory, INESC-ID Lisboa, Portugal.

José Guilherme Camargo de Souza Godigital Tecnologia e Participações, GODIGITAL, Brasil.

Liliana Ferreira Departamento de Electrónica, Telecomunicações e Informática, Instituto de Engenharia Electrónica e Telecomunicações de Aveiro, Universidade de Aveiro, Portugal.

Marcirio Silveira Chaves Linguateca, Pólo do XLDB, LaSIGE - Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Portugal.

Mírian Bruckschen Pontifícia Universidade Católica do Rio Grande do Sul, Brasil.

Nuno Cardoso LaSIGE - Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Portugal.

Nuno Mamede Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal & L2F - Laboratório de sistemas de Língua Falada, INESC-ID Lisboa, Portugal.

Olga Craveiro Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Leiria, Portugal & CISUC, Departamento de Engenharia Informática, Universidade de Coimbra, Portugal.

Paula Carvalho Linguateca, Pólo do XLDB, LaSIGE - Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Portugal.

Pedro Mendes Priberam Informática, Portugal.

Renata Vieira Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Brasil.

Sandro Rigo UNISINOS - Universidade do Vale do Rio dos Sinos, Brasil.

Tiago Veiga Priberam Informática, Portugal.

Glossário

Paula Carvalho, Cristina Mota, Diana Santos, Cláudia Freitas e Hugo Gonçalo Oliveira

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, p. v–ix.

Abrangência Métrica que afere a qualidade da participação, em termos da proporção de respostas correctas no universo de respostas possíveis.

Análises alternativas Diferentes análises concorrentes, dentro do mesmo ALT, associadas a uma dada sequência.

ALT Etiqueta que identifica um conjunto de análises alternativas associadas a uma mesma sequência, passível de ser segmentada de formas distintas.

Atributos do HAREM clássico Conjunto de atributos previstos nas directivas do HAREM clássico, designadamente, CATEG, TIPO e SUBTIPO, os quais também são adoptados nas directivas do TEMPO para caracterizar as expressões temporais.

Atributos do ReRelEM Conjunto de atributos especificamente utilizados na pista do ReRelEM, designadamente, TIPOREL e COREL.

Atributos do TEMPO estendido Conjunto de atributos específicos da categoria TEMPO, designadamente, TEMPO_REF, SENTIDO, VAL_NORM e VAL_DELTA.

Avaliação conjunta Modelo de avaliação que permite a comparação do progresso dos sistemas participantes numa dada área, usando, para tal, um conjunto de recursos comum e métricas e medidas bem definidas.

Avaliação estrita de ALT Contabilização de todas as análises alternativas possíveis para um dado segmento do texto, recebendo cada elemento do ALT um peso igual ao inverso do número de alternativas associadas a esse segmento.

Avaliação relaxada de ALT Avaliação do HAREM clássico onde, dentro de cada ALT, é seleccionado apenas o elemento que maximiza o valor da medida de classificação do sistema, tal como foi feito no Primeiro HAREM.

Cenário selectivo Subconjunto de categorias, tipos, subtipos e/ou outros atributos do cenário total definido pelo participante ou pela organização.

Cenário de avaliação Subconjunto de categorias, tipos, subtipos e/ou outros atributos do cenário total sobre o qual vai incidir a avaliação. Pode coincidir com o cenário total.

Cenário de participação Subconjunto de categorias, tipos, subtipos e/ou outros atributos do cenário total em que um dado sistema participou e pretende ser avaliado. Pode coincidir com o cenário total.

Cenário total Conjunto total de categorias, tipos e subtipos propostos no Segundo HAREM.

Classificação Classificação das EM de acordo com as categorias, tipos, subtipos e/ou outros atributos previstos nas directivas.

Colecção dourada (CD) do Segundo HAREM Subconjunto de documentos da colecção HAREM anotados de acordo com as directivas do HAREM clássico (e do TEMPO quanto aos atributos CATEG, TIPO e SUBTIPO), e com base nos quais os sistemas participantes são avaliados no HAREM clássico, que inclui a avalição da pista do TEMPO em relação aos atributos mencionados.

Colecção dourada do ReRelEM (ou CD do ReRelEM) Subconjunto de documentos da CD do Segundo HAREM anotados de acordo com as directivas do ReRelEM, e com base nos quais os sistemas participantes são avaliados na tarefa de reconhecimento de relações entre EM.

Colecção dourada do TEMPO (ou CD do TEMPO) Subconjunto de documentos da CD do Segundo HAREM anotados de acordo com as directivas do TEMPO, e com base nos quais os sistemas participantes são avaliados na pista do TEMPO.

Colecção do Segundo HAREM Conjunto de documentos, com géneros e registos variados, que os sistemas participantes tiveram de anotar no âmbito do Segundo HAREM.

Corrida no Segundo HAREM (ou corrida) Colecção do Segundo HAREM anotada por um sistema participante. O mesmo que **participação no Segundo HAREM**.

Directivas do HAREM clássico Conjunto de regras e indicações que foram adoptadas na identificação e classificação das EM, no âmbito da pista do HAREM clássico, e segundo as quais os sistemas foram avaliados.

Directivas do ReRelEM Conjunto de regras e indicações que foram adoptadas na identificação e classificação das relações entre EM, no âmbito da pista do ReRelEM, e segundo as quais os sistemas foram avaliados.

Directivas do TEMPO Conjunto de regras e indicações que foram adoptadas na identificação, classificação e normalização das expressões temporais, no âmbito da pista do TEMPO, e segundo as quais os sistemas foram avaliados.

Elemento do `ALT` Elemento do conjunto de análises alternativas possíveis.

Encaixe (de EM) Existência de EM dentro de uma EM maior.

Entidade mencionada (EM) No HAREM clássico, as EM correspondem, em geral, a uma entidade referida no texto por um nome próprio ou por uma expressão numérica, cuja caracterização mais pormenorizada se encontra nas directivas.

EM espúria Sequência identificada pelos sistemas participantes como EM, que não se encontra anotada na CD.

EM vaga Entidade mencionada que tem mais de uma interpretação (e como tal recebe mais de uma classificação no HAREM). Cada interpretação é designada por faceta.

Expansão de relações Explicitação das consequências das propriedades de simetria, existência de inversa e transitividade, e aplicação das regras de expansão, a um conjunto de relações numa colecção.

Faceta Interpretação parcial do conteúdo de uma EM vaga, seleccionando apenas uma categoria (ou tipo, ou subtipo) do conjunto de categorias (ou tipos, ou subtipos) postulado para essa EM.

HAREM Acrónimo para "HAREM - Avaliação de Reconhecimento de Entidades Mencionadas", uma série de avaliações conjuntas na área do reconhecimento de entidades mencionadas em português.

HAREM clássico Pista de avaliação de REM no âmbito do Segundo HAREM, baseada nas directivas do HAREM clássico, cujo objectivo é anotar as EM de acordo com o leque de categorias, tipos e subtipos especificados nas directivas. A avaliação do HAREM clássico tem também em conta a avaliação da pista do TEMPO para os atributos do HAREM clássico.

HAREM clássico na CD do TEMPO Uso da CD do TEMPO como base de avaliação segundo as directivas do HAREM clássico e do TEMPO quanto aos atributos do HAREM clássico.

HAREM clássico na CD do ReRelEM Uso da CD do ReRelEM como base de avaliação segundo as directivas do HAREM clássico e do TEMPO quanto aos atributos do HAREM clássico.

Identificação Reconhecimento de que uma dada expressão constitui uma EM, independentemente da sua classificação.

Medida Função que permite a transformação quantitativa da pontuação num valor numérico.

Medida-F Métrica que combina a precisão e a abrangência para cada tarefa, de acordo com a fórmula estabelecida.

Métrica Conjugação dos valores produzidos por uma dada medida num único valor acumulado.

Mini-HAREM Segundo evento de avaliação do Primeiro HAREM.

Modo de avaliação Selecção de atributos do `TEMPO` que são alvo de avaliação. Existem quatro modos de avaliação: clássico, estendido completo, estendido sem normalização e estendido só com normalização.

Participação no Segundo HAREM (ou participação) Colecção do Segundo HAREM anotada por um sistema participante. O mesmo que **corrida no Segundo HAREM**.

Peso Valor que é atribuído a de cada um dos atributos de uma EM ou a cada um dos elementos do `ALT`, no cálculo da medida atribuída a uma dada EM.

Pistas do Segundo HAREM Conjunto das diferentes pistas integradas no Segundo HAREM, designadamente, o HAREM clássico, a pista do ReRelEM e a pista do TEMPO.

Pista do ReRelEM Pista específica do Segundo HAREM, que tem por finalidade o reconhecimento de relações entre entidades mencionadas.

Pista do TEMPO Pista específica do Segundo HAREM, baseada num conjunto de directivas independentes das directivas do HAREM clássico, as directivas do TEMPO (Hagège et al., 2008), que tem em vista o reconhecimento e a normalização de expressões temporais.

Pontuação Avaliação qualitativa da relação entre a resposta do sistema e o que está na CD.[1]

Precisão Métrica que afere a qualidade da participação, em termos da proporção de respostas correctas dentro do total de respostas dadas.

Primeiro HAREM Primeira edição do HAREM, que englobou duas avaliações distintas: o primeiro evento do Primeiro HAREM e o Mini-HAREM.

Reconhecimento de entidades mencionadas (REM) Tarefa de identificação e classificação automática de entidades mencionadas.

Reconhecimento de relações entre entidades mencionadas (ReRelEM) Tarefa de identificação e classificação de um conjunto de relações entre EM, previamente estabelecidas nas directivas do ReRelEM.

Relação entre EM Ligação entre duas entidades, de acordo com os critérios sugeridos nas directivas do ReRelEM.

ReRelEM Acrónimo para Reconhecimento de Relações entre Entidades Mencionadas.

Saída Resultado de um sistema. No contexto do Segundo HAREM, uma saída refere-se à anotação da colecção do Segundo HAREM, mas distingue-se de corrida por esta última ser uma saída que foi seleccionada pelo participante para ser avaliada no contexto do HAREM.

Sistema participante Sistema que participou no Segundo HAREM.

Segundo HAREM Segunda edição do HAREM.

TEMPO clássico Modo de avaliação da pista do TEMPO que tem em consideração apenas o preenchimento dos atributos `CATEG`, `TIPO` e `SUBTIPO`.

TEMPO estendido completo Modo de avaliação da pista do TEMPO em que todos os atributos de `TEMPO` são avaliados, incluindo os atributos específicos do `TEMPO`.

TEMPO estendido sem normalização Modo de avaliação da pista do TEMPO em que todos os atributos de `TEMPO` são avaliados, excepto os referentes à normalização.

TEMPO estendido só com normalização Modo de avaliação da pista do TEMPO em que são avaliados os atributos de `TEMPO` previstos no HAREM clássico conjuntamente com os atributos de normalização.

[1] No Segundo HAREM, adoptámos o trio de termos do Primeiro HAREM "pontuação", "medida" e "métrica", embora consideremos agora que os termos "apreciação qualitativa", "pontuação" e "medida", respectivamente, teriam sido uma melhor escolha.

Enquadramento e historial do Segundo HAREM

Diana Santos

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, p. 1–7.

O HAREM é uma avaliação conjunta na área do reconhecimento de entidades mencionadas (REM) em português, organizada pela Linguateca. Uma avaliação conjunta, em poucas palavras, é uma tarefa que vários sistemas concordam em tentar executar, de forma a comparar o desempenho entre eles, com base em medidas consensuais e com recursos criados por uma comissão idónea – veja-se Santos (2007b) para mais informação sobre este paradigma de avaliação.

Neste texto pretendo fazer um historial da actividade da Linguateca nesta área e deixar expresso quais as muitas pessoas que participaram, e em que grau, na já longa história do HAREM.

O interesse pela avaliação na área foi inicialmente levantado no Encontro Preparatório sobre avaliação conjunta em 2002 (EPAV 2002), e uma parte significativa do trabalho da Cristina Mota, contratada na altura pela Linguateca para trabalhar no então pólo da Linguateca no LabEL, foi o de fazer um estudo preliminar das necessidades e interesses da comunidade, cujo relato foi feito no Avalon 2003 (Mota, 2003).

Quando houve finalmente oportunidade para organizar o que veio a ser chamado o Primeiro HAREM, já a Cristina se encontrava a trabalhar na sua tese de doutoramento sem ligação à Linguateca, e coube ao pólo da Linguateca no XLDB, em colaboração com o pólo de Oslo, a organização do Primeiro HAREM (de Setembro de 2004 a Julho de 2006).

A parte de leão coube ao Nuno Cardoso e a mim, embora contássemos com o apoio e colaboração de várias pessoas em fases diferentes dessa avaliação conjunta, que foi descrita em forma de livro em Santos e Cardoso (2007a) e onde essas contribuições estão bem documentadas. O Nuno acabou também por fazer a sua tese de mestrado, orientada pelo Mário J. Silva pelo XLDB e pelo Eugénio Oliveira pela FEUP sobre a organização e validação dos resultados do HAREM (Cardoso, 2006), num exemplo raro de uma tese em avaliação em Portugal, descrevendo, não um protótipo nem uma ideia para confirmação futura, mas uma actividade passada de interesse para toda a comunidade.

Quando a Linguateca recebeu financiamento para a sua terceira fase, de 15 de Dezembro de 2006 até fim de Dezembro de 2008, uma das iniciativas que constava do seu plano de actividades era precisamente a organização de um Segundo HAREM. Infelizmente era agora a vez de o Nuno Cardoso estar em trabalhos de doutoramento – aliás a maior parte das pessoas ligadas ao Primeiro HAREM também tinham deixado de ter ligação contratual à Linguateca – e, portanto, foi preciso criar uma nova equipa para organizar o Segundo HAREM.

Trabalho que me coube mais uma vez, mas com cujo resultado não me podia dar por mais satisfeita: A primeira "aquisição" foi a Cláudia Freitas, recém-doutorada em semântica computacional (Freitas, 2007) e uma fã do HAREM, que tinha vindo trabalhar para a Linguateca desde Julho de 2007 e graças em grande parte ao HAREM (com o qual tomara conhecimento no PROPOR 2006 e seus satélites). O segundo foi o Hugo Oliveira, também a trabalhar em semântica computacional (Gonçalo Oliveira et al., 2008) e chegado à Linguateca, em boa hora, em Setembro de 2007. Ambos no pólo de Coimbra. Quando o Segundo HAREM arrancou, por essa altura, éramos portanto apenas três, mas a vinda da Paula Carvalho, também recém-doutorada (Carvalho, 2007), em Dezembro de 2007, para engrossar as fileiras do HAREM, foi providencial, porque à data já estávamos a compreender que nos tínhamos comprometido a muito (talvez demais). A Paula encontrava-se afecta ao pólo da Linguateca no XLDB, de onde também o David Cruz foi chamado para ajudar à parte informática, tendo contudo sido substituído pela Cristina Mota, que ao acabar a sua bolsa de doutoramento da FCT foi recontratada pela Linguateca para o HAREM.

Note-se aliás que a Cristina apenas começou a trabalhar para a organização do HAREM a partir da data da avaliação conjunta, visto que participou no HAREM com o seu sistema. E ficou com a avaliação do TEMPO porque foi condição para a sua contratação não concorrer ao TEMPO.

Desde essa altura, esta equipa de cinco pessoas esteve a dar o seu melhor (embora em percentagens de afectação diferentes) à organização do HAREM, só se considerando o Segundo HAREM terminado à data da disponibilização do seu pacote de recursos, a LÂMPADA, o que ocorreu a 17 de Novembro de 2008.

Existem algumas diferenças fundamentais entre o Primeiro e o Segundo HAREM que convém serem destacadas aqui, de forma a este livro poder ser comparado de forma justa com o anterior.

1. Em primeiro lugar, o Segundo HAREM estava limitado pelo tempo escasso que restava à Linguateca, e que é mais compreensível na secção que faz o historial detalhado do Segundo HAREM, mas que é sumarizado na figura 1.

2. Em segundo lugar, por se tratar de uma segunda edição, tivemos em muitos casos que efectuar compromissos entre razões históricas e de continuidade e razões científicas e de progresso, compromissos esses que não são fáceis e que o presente livro tenta documentar.

3. Em terceiro lugar, apostámos numa internacionalização do HAREM que não funcionou, tendo mandado todas as mensagens (e principal documentação) sempre em paralelo, em português e em inglês. Contudo, e ao contrário do Primeiro HAREM em que até houve quatro capítulos em inglês no livro, não houve nesta edição qualquer participação não-lusófona.

4. Em quarto lugar, tivemos a oferta de uma nova tarefa para realizar no âmbito do HAREM, a normalização do TEMPO, que foi muito gratificante e enriquecedora mas que deu origem a muitíssimos problemas e dificuldades que não podíamos prever, e de que fazemos aqui o historial mais detalhado, para explicar, entre outras coisas, a existência de dois capítulos do presente volume dedicados a essa pista. Convém contudo indicar que larga troca de correspondência e discussão de diversos pormenores da proposta do TEMPO teve lugar nos bastidores, entre o grupo e a organização do HAREM, e em que não só ficou clara a nossa divergência teórica como a autoria diferenciada da proposta. Quanto à forma de avaliar o TEMPO, devido à necessidade de integração com o resto do HAREM, essa ficou desde sempre atribuída à organização do HAREM, que teria também a obrigação de calcular os resultados e publicá-los. Isso explica, como já mencionado, a existência de dois capítulos sobre a organização desta pista: o capítulo 2 e o 3.

5. Em quinto lugar, e para garantir algum progresso também ao nível das tarefas a que os sistemas eram desafiados, lançámos mais uma pista, o ReRelEM, para identificar relações entre EM (capítulo 4).

6. Em sexto lugar, e talvez esta seja a diferença mais importante, muito do trabalho de estudo e reflexão sobre o Segundo HAREM terá de ficar por fazer – ou terá de ser feito no âmbito de um terceiro se vier a ser organizado, devido ao pouco tempo a que já aludi.

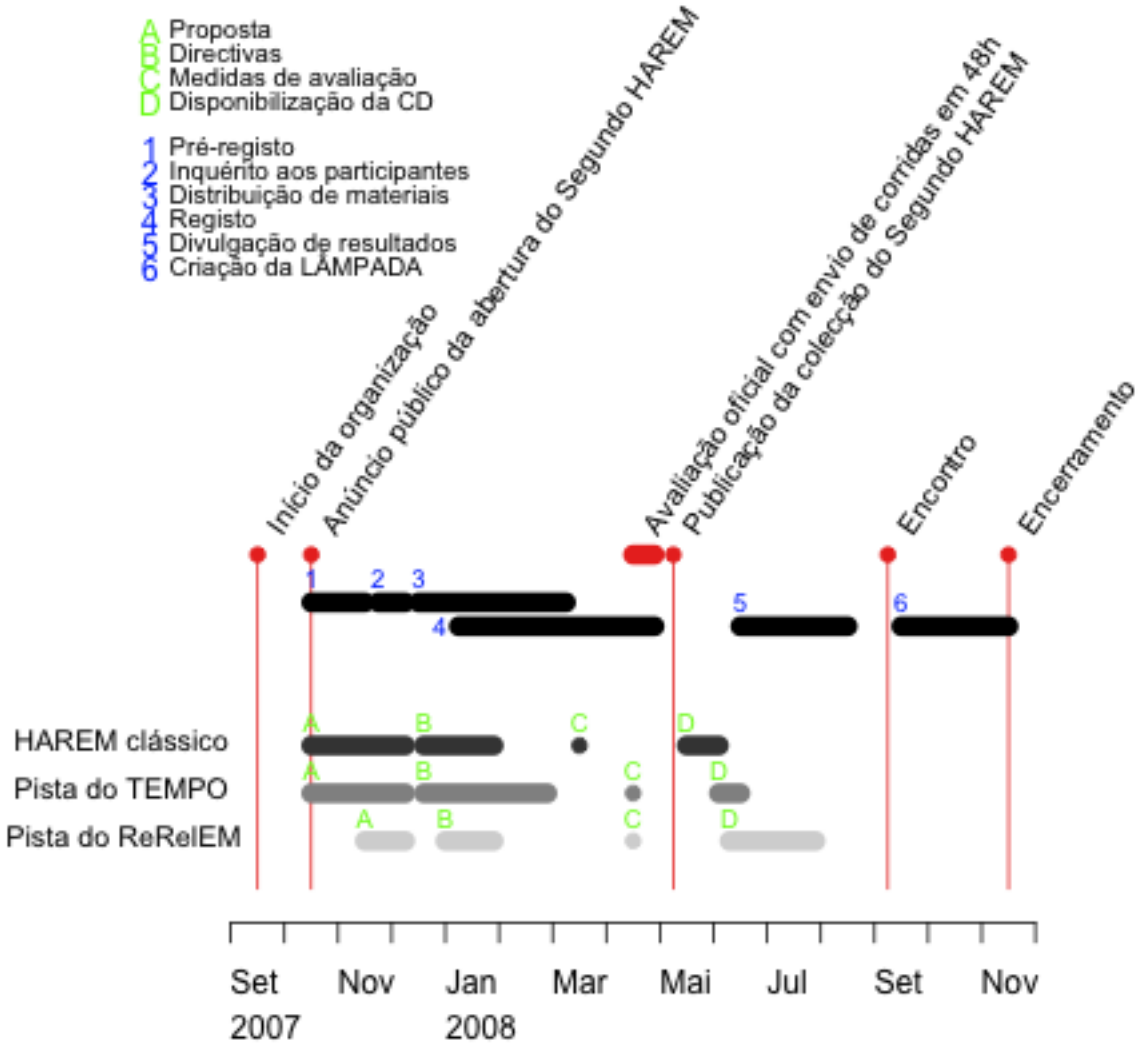


Figura 1: Diagrama temporal do Segundo HAREM

Em relação ao conteúdo do presente livro, a organização do HAREM como um todo considerou que, embora houvesse diferentes datas de início do trabalho e diferentes percentagens de afectação ao mesmo (a única que esteve a 100% no HAREM foi a Cristina), a equipa foi tão coesa e trabalhou tão bem em conjunto que não havia razão para apresentar artigos em separado, e por isso todos os artigos têm os cinco autores como autores. Refira-se, contudo, que a Cristina não participou na elaboração dos programas de avaliação do HAREM clássico visto que foi participante. Isso não nos pareceu razão suficiente para não ser co-autora do capítulo 1, que redigiu em conjunto com os outros autores.

Mais uma vez, quero realçar que o trabalho apresentado aqui é fruto indirecto de toda a equipa da Linguateca, e que em particular toda a infraestrutura, apoio organizativo e informático muito devem ao Luís Costa e ao Luís Miguel Cabral do pólo de Oslo.

Finalmente nunca é demais salientar que uma avaliação conjunta só funciona graças aos seus participantes, e que a comunidade do processamento da língua portuguesa é a principal beneficiada, mas também responsável e autora, do HAREM.

Historial detalhado

Aqui apresento as datas mais importantes do Segundo HAREM, para que fique registado o processo.

15 de Outubro de 2007 Abertura do Segundo HAREM, publicitada em várias listas.

15 de Outubro de 2007 Abordagem, por parte do Nuno Mamede, Caroline Hagège e Jorge Baptista (mais tarde chamados "grupo do TEMPO"), anunciando-nos a sua intenção de propor uma tarefa nova em relação ao tempo.

10 de Novembro de 2007 Prazo limite para registo dos participantes, tivemos manifestações de interesse de 22 grupos.

12 de Novembro de 2007 Mencionada a possibilidade da existência de uma pista de co-referência, em colaboração com a Renata Vieira (que mais tarde preferiu apenas participar).

14 de Novembro de 2007 Envio da primeira versão das directivas relativas ao TEMPO pelo respectivo grupo, por nós reenviada para a lista.

22 de Novembro de 2007 Inquérito aos participantes sobre os moldes de participação no TEMPO e no ReRelEM assim como outras questões.

3 de Dezembro de 2007 Prazo limite para discussão sobre os moldes do Segundo HAREM.

4 de Dezembro de 2007 Conclusões da discussão tornadas públicas, em particular com a proposta do TEMPO cabalmente aprovada.

10 de Dezembro de 2007 Exemplos preliminares do HAREM (primeira colecção de exemplo).

18 de Dezembro de 2007 Directivas preliminares publicadas, do HAREM e do TEMPO.

4 de Janeiro de 2008 Directivas iniciais do ReRelEM publicadas.

9 de Janeiro de 2008 Foi criado um formulário para que todos os participantes se registassem com uma definição precisa das tarefas que iam tentar.

17 de Janeiro de 2008 Como resultado, dois grupos afirmaram querer tentar o TEMPO completo, e cinco apenas o TEMPO clássico (ou seja, categoria, tipo e subtipo de EM tipo `TEMPO`). Em relação ao ReReLEM, houve quatro grupos inicialmente interessados. Finalmente, quanto a SUBTIPOS, apenas os dos `LOCAL` foram aceites, tendo sido rejeitados os de `PESSOA` e `ORGANIZACAO` propostos.

31 de Janeiro de 2008 Directivas finais publicadas, assim como material de teste congelado.

15 de Fevereiro de 2008 Adenda sobre o TEMPO divulgada.

15 de Fevereiro de 2008 Primeira versão do validador do Segundo HAREM disponível.

26 de Fevereiro de 2008 Nova versão das directivas do TEMPO divulgada.

4 de Março de 2008 Disponibilização do exemplário (para o HAREM clássico).

10 de Março de 2008 Primeira versão das medidas de avaliação, para o HAREM clássico.

10 de Abril de 2008 Primeira versão das medidas de avaliação para o TEMPO e para o ReRelEM.

14 de Abril de 2008 Abertura do prazo oficial do Segundo HAREM, tendo os participantes 48 horas desde que fossem buscar a colecção para a devolver anotada.

28 de Abril de 2008 Encerramento do prazo oficial do Segundo HAREM.

8 de Maio de 2008 Publicação da colecção do Segundo HAREM.

16 de Maio de 2008 Disponibilização da primeira versão da colecção dourada do Segundo HAREM para inspecção.

4 de Junho de 2008 Disponibilização da primeira versão da colecção dourada para o TEMPO completo para inspecção.

4 de Junho de 2008 Disponibilização da versão final[2] da colecção dourada do Segundo HAREM.

6 de Junho de 2008 Disponibilização da primeira versão da colecção dourada para o ReRelEM para inspecção.

12 de Junho de 2008 Disponibilização da versão final da colecção dourada para o TEMPO completo.

18 de Junho de 2008 Disponibilização da primeira versão dos programas de avaliação.

[2] Na altura foi divulgada como final, contudo foram sendo descobertos alguns problemas e houve uma CD oficial – para os resultados – e mais tarde ainda foram feitas umas últimas correcções até à versão derradeira disponibilizada na LÂMPADA. Note-se, todavia, que a lista dessas alterações foi enviada a todos os participantes, o memso acontecendo em relação às outras colecções douradas.

19 de Junho de 2008 Divulgação dos resultados preliminares do HAREM clássico.

25 de Junho de 2008 Divulgação dos resultados preliminares da avaliação do TEMPO.

31 de Julho de 2008 Disponibilização da versão final da colecção dourada para o ReRelEM.

6 de Agosto de 2008 Divulgação dos resultados preliminares da pista ReRelEM.

8 de Agosto de 2008 Divulgação dos relatórios finais individuais (exceptuando a pista ReRelEM).

21 de Agosto de 2008 Divulgação de novos resultados da pista ReRelEM.

7 de Setembro de 2008 Encontro do Segundo HAREM, como satélite do PROPOR 2008.

12 de Novembro de 2008 Divulgação das últimas alterações às colecções douradas.

13 de Novembro de 2008 Divulgação das últimas alterações aos programas de avaliação.

17 de Novembro de 2008 Pacote de recursos finais do Segundo HAREM, a LÂMPADA, disponibilizado.

Parte I

O HAREM pela organização

Capítulo 1

Segundo HAREM: Modelo geral, novidades e avaliação

Paula Carvalho, Hugo Gonçalo Oliveira, Diana Santos, Cláudia Freitas e Cristina Mota

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 1, p. 11–31.

No Segundo HAREM, foi mantida a filosofia subjacente ao Primeiro HAREM, nomeadamente o modelo semântico (Santos, 2007d) e o modelo geral de avaliação (Santos et al., 2007). Contudo, e como seria de esperar, procurou-se corrigir e aperfeiçoar algumas arestas em relação à edição anterior, o que se reflectiu numa caracterização mais precisa e linguisticamente motivada de certas entidades mencionadas (EM), bem como numa avaliação mais justa dos sistemas. Esta segunda edição do HAREM passou também a incluir duas novas tarefas/pistas, designadamente a tarefa de reconhecimento e normalização de expressões temporais e a tarefa de reconhecimento de relações semânticas entre EM, o ReRelEM, a que dedicamos os capítulos 2 e 4, respectivamente, deste livro.

Neste capítulo, discutimos especificamente a pista geral de reconhecimento de entidades mencionadas no Segundo HAREM, a que nos referiremos, daqui em diante, como HAREM clássico. Mais especificamente, na secção 1.1, apresentamos, de forma sucinta, o modelo semântico subjacente ao HAREM. Em 1.2, centramo-nos na proposta de classificação das EM tida em consideração no Segundo HAREM, bem como nas alterações que esta sofreu em relação à proposta de classificação utilizada no Primeiro HAREM. Na secção 1.3, discutimos as melhorias introduzidas no Segundo HAREM, face à primeira edição. Em 1.4, descrevemos o processo de constituição das colecções usadas especificamente no âmbito desta avaliação, nomeadamente a colecção do Segundo HAREM e a respectiva colecção dourada (CD). Fazemos, ainda, uma breve caracterização de ambas as colecções, e enumeramos as principais fases inerentes ao processo de anotação e revisão da CD. Por fim, na secção 1.5, discutimos os resultados obtidos pelos sistemas participantes, nos diferentes tipos de avaliação tidos em conta no Segundo HAREM.

1.1 Filosofia do HAREM

O modelo semântico do HAREM assenta em dois aspectos essenciais, que o distinguem de outros modelos vulgarmente utilizados na avaliação de REM[1]. Esses aspectos prendem-se nomeadamente com (i) a ideia de que identificação e classificação de uma dada expressão como entidade mencionada depende exclusivamente do seu uso em contexto, não estando lexicalmente "presa" a nenhum dos atributos a que possa estar associada noutros recursos linguísticos, por exemplo, dicionários, almanaques, ontologias e com (ii) o facto de ser possível atribuir mais do que uma classificação (categoria, tipo e/ou subtipo) a uma mesma EM (considerando-a portanto vaga entre as várias classificações), se o contexto em que a mesma se encontra não permitir escolher apenas uma delas.

Embora, na maioria das avaliações levadas a cabo neste domínio, a classificação das entidades mencionadas esteja intimamente relacionada com a sua caracterização (semântica) nos recursos lexicais, no HAREM considera-se que essa caracterização só pode ser feita numa situação de uso concreto da língua. Não consideramos, portanto, que uma EM possui, intrinsecamente, um dado significado, que pode eventualmente ganhar diferentes nuances conforme o contexto que essa EM integre. Isso implicaria, entre outras coisas, assumir a existência de "um significado de base" e de "um significado derivado do uso". Como referimos antes, a nossa posição é a de que o significado de qualquer EM é, à partida, quase imprevisível, e só pode ser compreendido através da sua função em contexto. De facto, apesar de poder parecer fazer sentido definir lexicalmente algumas categorias

[1] Para uma análise contrastiva entre o HAREM e outras avaliações realizadas neste domínio, em particular o MUC e o CoNLL, veja-se Santos (2007c), Santos e Cardoso (2007b) e Seco (2007).

semânticas, como é o caso paradigmático de *país*[2], não é obrigatório que exista uma relação de univocidade entre esse conceito e uma única categoria ou conjunto de categorias que considerámos pertinentes no HAREM, nomeadamente, LOCAL e/ou ORGANIZACAO. Por exemplo, *Portugal* pode ser usado para fazer referência a um conjunto variado de sentidos (como ilustrado nos exemplos (1.1) a (1.5)[3]), sem que nenhum deles tenha necessariamente primazia sobre os outros.

(1.1) Regressou então a <EM ID="ub-67792-10" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="PAIS">**Portugal**</EM>, onde iniciou meteórica carreira na experimentação de novas formas de expressão

(1.2) O acordo político quanto à revisão foi obtido durante a Presidência Alemã, tendo cabido a <EM ID="a46996-5" CATEG="ORGANIZACAO" TIPO="ADMINISTRACAO">**Portugal**</EM> concluir o processo de revisão.

(1.3) Este debate passou completamente ao lado de <EM ID="2-dftre765-" CATEG="PESSOA" TIPO="POVO">**Portugal**</EM>

(1.4) o problema do PSD é começar a ter só um <EM ID="ub-24360-32" CATEG="ABSTRACCAO" TIPO="IDEIA">**Portugal**</EM> ou dois dentro de si

(1.5) <EM ID="x-1G" CATEG="PESSOA" TIPO="GRUPOMEMBRO">**Portugal**</EM> perdeu com a Suíça por 2-0

Mas, se para o exemplo de *Portugal* não é difícil acordar sobre uma definição, a de "país" (a qual, segundo uma certa visão da língua, estaria, pelo menos, associada às "variações" LOCAL e ORGANIZACAO), o mesmo não acontece para EM mais abstractas. Por exemplo, *Big-Bang* tanto pode ser definida como uma "teoria" sobre a criação do universo (exemplo (1.6)) ou como uma "explosão cósmica" (exemplo (1.7)), sendo, respectivamente, classificada como ABSTRACCAO e ACONTECIMENTO.

(1.6) A radiação de origem cósmica, prevista pelo <EM ID="bb1" CATEG="ABSTRACCAO">**Big Bang**</EM> seria descoberta em 1964, quase acidentalmente, por Arno Penzias e Robert Wilson.[4]

(1.7) Esse ponto deve ter sido o começo dos tempos, pelo qual tem início a expansão das galáxias, que os cosmologistas descrevem como uma explosão, ou seja, o <EM ID="bb2" CATEG="ACONTECIMENTO">**Big Bang**</EM>[5]

Diferentemente de outras avaliações de REM, em que se considera que as entidades devem receber uma única classificação, mesmo que arbitrária em última análise, no HAREM propomos que as entidades poderão (e deverão) estar associadas a mais do que uma etiqueta, sempre que o contexto em que essas EM ocorrem não permita seleccionar uma de

[2] Por exemplo, na Wikipédia, *país* é definido como um "território social, política, cultural e geograficamente delimitado" e na Infopédia como um "espaço demarcado por fronteiras geográficas e dotado de soberania própria; estado; nação".
[3] Para mais pormenores sobre o esquema de anotação, veja-se a próxima secção ou o apêndice A.
[4] http://www.if.ufrj.br/teaching/cosmol/exprim1.html, em 24 de Outubro de 2008
[5] http://www.coladaweb.com/astronomia/bigbang.htm, em 24 de Outubro de 2008

entre as várias análises possíveis (Santos, 2007d). Trata-se, pois, de preservar aquilo que consideramos uma propriedade essencial da linguagem natural, a vagueza, que não pode ser resolvida nem eliminada, de modo a não se perder informação (Santos, 1997, 2006).

Ilustramos, em seguida, alguns exemplos de EM vagas – extraídas da CD do Segundo HAREM – retomando o caso de *Portugal*.

(1.8) Pela mão do ministro Freitas do Amaral, e sem necessidade alguma, <EM ID="a66435-10" CATEG="ORGANIZACAO|PESSOA" TIPO="ADMINISTRACAO|POVO">**Portugal**</EM> foi enxovalhado, coberto de vergonha e de cobardia, por um dos mais tristes textos políticos que já alguém escreveu.

(1.9) Mais de 32 mil pessoas poderiam morrer se uma pandemia de gripe humana de origem aviária atingisse <EM ID="ub-28874-3" CATEG="PESSOA|LOCAL" TIPO="POVO|HUMANO" SUBTIPO="|PAIS">**Portugal**</EM>

(1.10) Os dois reinos católicos, <EM ID="a66435-5" CATEG="PESSOA|ORGANIZACAO" TIPO="GRUPOIND|ADMINISTRACAO">**Portugal**</EM> e Espanha, partiram à conquista do mundo e tornaram-se Impérios marítimos do <EM ID="aa66435-54" CATEG="LOCAL|LOCAL" TIPO="FISICO|HUMANO" SUBTIPO="REGIAO|DIVISAO">**Novo Mundo**</EM>

Em (1.8), *Portugal* tanto pode referir o governo (ORGANIZACAO ADMINISTRACAO) como o povo (PESSOA POVO) português; em (1.9), a vagueza observa-se entre esta última análise (a de PESSOA POVO) e a de LOCAL; por fim, no exemplo (1.10), *Portugal*, tanto pode referir o governo português como um grupo indeterminado de pessoas individuais que não possuem um nome convencional (PESSOA GRUPOIND). A vagueza não se observa simplesmente ao nível da categoria (CATEG) das EM; em muitos casos, esta propriedade estabelece-se a um nível de subcategorização mais fino das entidades, nomeadamente no que respeita aos tipos e subtipos envolvidos. Por exemplo, em (1.10), *Novo Mundo*, que no contexto em questão faz menção a um LOCAL, pode representar tanto um local da geografia física (LOCAL FISICO REGIAO) como da geografia humana (LOCAL HUMANO DIVISAO).

Não queremos dar, contudo, a ideia (completamente errada) de que esta situação se passa sobretudo no caso dos nomes de países ou cidades, embora este seja um exemplo tão discutido na literatura que é incontornável não o referir (veja-se, a propósito, a vasta literatura citada em Santos (2007d)). Apresentamos, em seguida, outros casos, completamente distintos dos anteriormente ilustrados, em que, uma vez mais, conceitos complexos se desdobram em sentidos múltiplos, no texto.

(1.11) O carácter diferente da <EM ID="H2-dftre765-41" CATEG="ABSTRACCAO|ACONTECIMENTO" TIPO="IDEIA|EFEMERIDE">**Reforma Inglesa**</EM> deve-se ao facto de ter sido promovida inicialmente pelas necessidades políticas de Henrique VIII.

(1.12) Assim aceitam os dois sacramentos do <EM ID="H2-dftre765-122" CATEG="ABSTRACCAO|OBRA" TIPO="IDEIA|PLANO">**Evangelho**</EM>: o Santo Batismo, através do qual a pessoa é feita membro da Igreja de Cristo.

No exemplo (1.11), tanto podemos entender *Reforma Inglesa* como um ACONTECIMENTO ou como uma ABSTRACCAO, mais especificamente uma IDEIA, e nenhuma das interpretações

exclui a outra. O mesmo se passa em relação a *Evangelho*, no exemplo (1.12), que pode corresponder quer a uma ABSTRACCAO quer a uma OBRA.

Naturalmente, a existência de vagueza entre várias interpretações depende do número de interpretações que o modelo semântico reputa como relevantes. Quanto mais diferenças finas de sentido quisermos reconhecer e anotar, maior será a possibilidade de não nos virmos obrigados a decidir por uma única interpretação, ou, por outras palavras, maior será a probabilidade de as EM serem consideradas vagas.

Esta questão não é meramente teórica e corresponde a uma fatia significativa dos casos que tivemos de anotar. Para um resumo quantitativo, veja-se a tabela 1.1, mais à frente, em que apresentamos a quantificação dos casos de vagueza presentes na CD, isto é, as EM em que não foi possível atribuir uma única classificação.

1.2 Esquema de anotação no Segundo HAREM

Nesta secção, procuramos, por um lado, fazer uma breve descrição do formato das etiquetas utilizado no Segundo HAREM, e, por outro, apresentar a proposta de classificação adoptada na anotação das EM, apontando as principais diferenças entre esta proposta e a que foi utilizada no Primeiro HAREM. Para informações mais detalhadas, sugerimos a consulta das directivas, no apêndice A.

1.2.1 Sintaxe das anotações

A anotação no Segundo HAREM foi feita de acordo com o formato XML. No que se refere às EM, todas as etiquetas começam com <EM ID="xxx"> e acabam com </EM>. O único atributo obrigatório é o identificador (ID), que, para facilidade de processamento, restringimos a uma combinação de apenas letras não acentuadas (maiúsculas ou minúsculas), algarismos, e os caracteres "-" e "_". Contrariamente ao que acontecia no Primeiro HAREM, cuja sintaxe de anotação das EM obrigava à explicitação da respectiva categoria (a qual incluía a etiqueta de abertura e de fecho da EM, por exemplo <PESSOA> e </PESSOA>), no Segundo HAREM a sintaxe das anotações é mais flexível, combinando numa mesma caracterização de saída (i) apenas a identificação (ii) a identificação e classificação de categorias, (iii) a identificação e classificação de categorias e tipos, (iv) a identificação e classificação de categorias, tipos e subtipos e (v) a identificação e categorias, tipos, subtipos e outros atributos previstos na classificação das EM (em concreto, os atributos previstos na classificação de expressões temporais ou na identificação de relações entre EM), sendo todas estas classificações opcionais.

Nos casos em que existem diferentes possibilidades de segmentação de uma dada sequência no texto, as diferentes análises alternativas associadas a essa sequência encontram-se compreendidas entre as etiquetas <ALT> e </ALT>, estando separadas entre si pelo símbolo "|"[6]; as diferentes EM identificadas no âmbito dessas análises recebem cada uma delas um ID distinto (cf. exemplo (1.13)).[7]

(1.13) aproximava a <ALT> <EM ID="2-dftre765-10" CATEG="ABSTRACCAO" TIPO="DISCIPLINA">**Igreja de Inglaterra**</EM> | <EM ID="2-dftre765-106-a" CATEG="ABSTRACCAO"

[6] Neste caso, o "|" não faz parte da linguagem XML, é uma representação própria do HAREM.

[7] Embora a notação sejam muito parecida com a do MUC-7 (Chinchor, 1998), chamamos a atenção para que nem o sentido de ALT nem o uso do símbolo "|" correspondem ao desta última avaliação conjunta.

TIPO="DISCIPLINA">**Igreja**</EM> de <EM ID="2-dftre765-1" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="PAIS">**Inglaterra**</EM> </ALT> do calvinismo.

O mesmo símbolo é também utilizado para separar as diferentes possibilidades de análise associadas a uma mesma EM, uma EM vaga (como ilustrado nos exemplos (1.8)-(1.10), anteriormente apresentados).

1.2.2 Classificação das EM

O conjunto de etiquetas usado no Segundo HAREM não é significativamente distinto do usado no Primeiro HAREM (cf. figura 1.1). O número de categorias nas duas avaliações é idêntico: dez categorias, as quais permaneceram intactas em relação à sua designação, excepto no que respeita a VARIADO, que foi substituída por OUTRO. Estas categorias pareceram-nos, pois, as mais pertinentes no âmbito de uma avaliação de REM em português, mas não rejeitamos a possibilidade de outras o poderem ser também, nomeadamente tendo em conta os interesses específicos de cada participante. Nesta perspectiva, a categoria, tipo ou subtipo OUTRO serve precisamente para dar conta de outras possibilidades de classificação das EM que não estejam contempladas no elenco de categorias (e/ou respectivos tipos e/ou subtipos) que definimos.

As categorias ACONTECIMENTO, VALOR e COISA não sofreram quaisquer alterações, exceptuando-se a inclusão do tipo OUTRO, que passou a ser um tipo possível de qualquer categoria.

Pelo contrário, as categorias LOCAL e TEMPO foram as que sofreram alterações mais substanciais, tendo sido alterados e/ou rebaptizados a maioria dos tipos anteriormente previstos. Além disso, estas categorias passaram ainda a incluir subtipos.

A categoria TEMPO encontra-se detalhadamente descrita no capítulo 2, pelo que não nos ocuparemos dela aqui.

No que respeita a LOCAL, deixámos de considerar o tipo CORREIO como uma EM, preferindo a marcação separada de ruas, estados e países dentro de moradas. Além disso, a informação abrangida, no Primeiro HAREM, pela etiqueta LOCAL ALARGADO passou a ser considerada como informação adicional em relação aos tipos ADMINISTRATIVO ou GEOGRAFICO (agora rebaptizados de HUMANO ou FISICO).

Deste modo, criou-se uma tripartição da categoria LOCAL em FISICO, HUMANO e VIRTUAL, em que FISICO substitui o anterior termo GEOGRAFICO, e HUMANO o anterior termo ADMINISTRATIVO.

Além da categoria TEMPO, esta foi a única categoria em que os participantes demonstraram interesse numa classificação mais fina em subtipos. A definição destes subtipos resultou de uma discussão entre os participantes especificamente interessados nesta categoria e a organização, reflectindo, assim, a soma das várias sensibilidades, experiências e opiniões das duas partes envolvidas.

A categoria PESSOA passou a incluir um novo tipo, que designámos como POVO, para dar conta de casos em que uma dada entidade, geralmente associada a um determinado local, é usada para referir a população desse local. Este conceito não era integralmente captado por nenhum dos tipos contemplados nas anteriores directivas.

A categoria ORGANIZACAO deixou de incluir o tipo SUB, que, na verdade, correspondia a uma subespecificação (ou se quisermos, subtipo) dos tipos ADMINISTRACAO, INSTITUICAO ou EMPRESA. Estes três tipos, já presentes no Primeiro HAREM, foram mantidos, e usados quer para a instituição (ou empresa, etc.) completa quer para uma subparte dela.

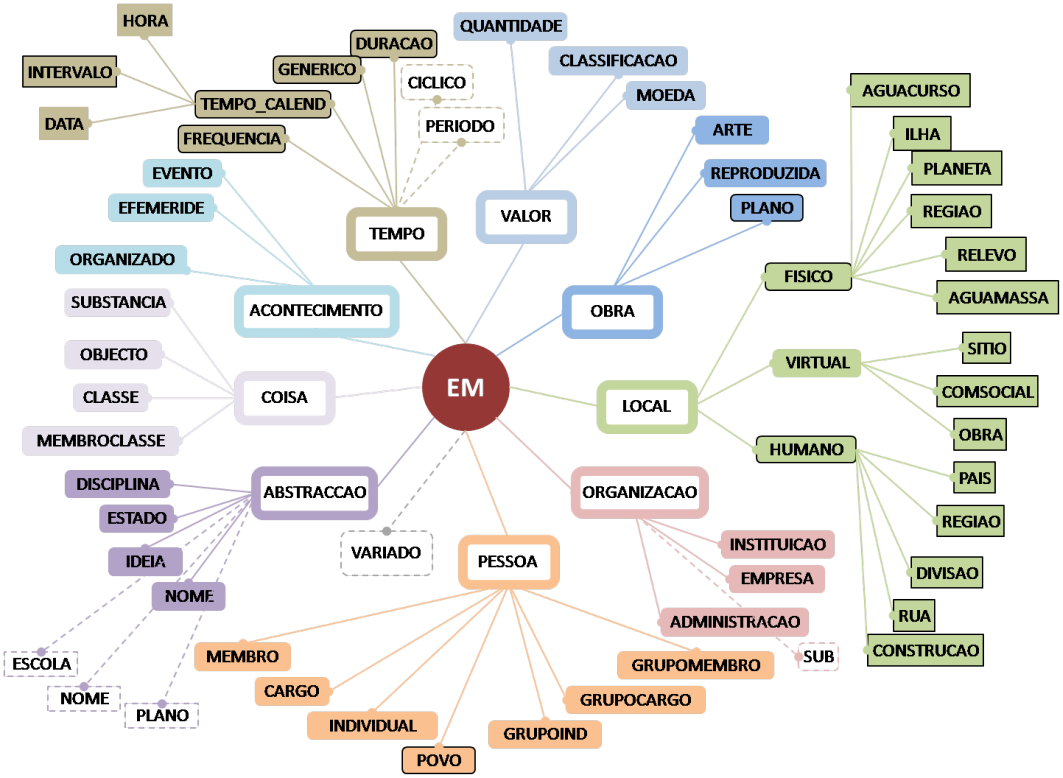


Figura 1.1: Árvore de categorias no Segundo HAREM: as categorias, tipos e subtipos representados nas caixas com contorno sólido preto só existem no Segundo HAREM; as categorias, tipos e subtipos representados nas caixas com contorno pontilhado só existem no Primeiro HAREM

A categoria OBRA passou a incluir o tipo PLANO (que anteriormente correspondia a um tipo da categoria ABSTRACCAO), deixando de parte o tipo PUBLICACAO, que, tal como CORREIO, correspondia a uma estrutura complexa, que preferimos não contemplar como EM.

A categoria ABSTRACCAO foi consideravelmente simplificada, retendo apenas os tipos DISCIPLINA, ESTADO, IDEIA e NOME. Por um lado, foram retirados desta categoria os tipos MARCA (convertido para a categoria COISA de tipo CLASSE ou IDEIA) e PLANO (transferido para categoria OBRA de tipo PLANO). Por outro lado, os tipos DISCIPLINA, ESCOLA e OBRA passaram a ser todos eles representados por um único tipo, DISCIPLINA.

Cada uma das categorias, tipos e subtipos referidos encontram-se ilustrados no apêndice E.

1.3 Melhorias no Segundo HAREM

A repetição de qualquer evento, neste caso, um evento de avaliação, não pode/deve corresponder, na nossa perspectiva, a uma mera cópia do evento anterior, sobretudo se considerarmos que há espaço para introdução de melhorias. É assim que entendemos o Se-

gundo HAREM: uma avaliação que tenta reter os aspectos positivos do Primeiro HAREM, mas que, naturalmente, procura melhorar os aspectos menos positivos, alguns dos quais previamente identificados aquando da realização do balanço do Primeiro HAREM (Santos e Cardoso, 2007b). Nas próximas subsecções, abordaremos as principais melhorias, em nosso entender, introduzidas especificamente no HAREM clássico.

1.3.1 Delimitação e classificação das EM

Ainda que, na maior parte dos casos, os critérios para a identificação e classificação de EM propostos no Primeiro HAREM tenham sido aplicados com sucesso ao reconhecimento de entidades mencionadas em português, considerámos que, em casos pontuais, a definição operacional de EM deveria ser ligeiramente modificada, de modo a ter uma classificação mais coerente e precisa, a qual pudesse, ao mesmo tempo, servir adequadamente os propósitos das aplicações em extracção e/ou recuperação de informação.

Neste sentido, as EM estruturalmente complexas, como moradas (anterior `LOCAL CORREIO`) e referências bibliográficas (anterior `OBRA PUBLICACAO`), embora relevantes num contexto de extracção de informação, deixaram de ser consideradas no Segundo HAREM, dada a dificuldade em motivar a sua identificação como entidades, numa tarefa de REM. De facto, neste contexto, parece-nos mais adequado privilegiar a análise autónoma das EM que constituem estas sequências, do que as sequências em si mesmo.

Numa outra perspectiva, mas tendo igualmente em linha de conta a própria noção de unidade lexical e semântica das EM, deixámos de fragmentar palavras ou expressões (compostas) cujos constituintes não obedeciam ao critério formal (das maiúsculas) previamente definido no HAREM para a identificação das EM. Concluímos que, nuns casos, as palavras ou expressões que anteriormente haviam sido classificadas como EM não o eram de facto (caso de *de Belém* para identificar *pastel de Belém* como EM, que agora não foi assim considerado) e que, noutros casos, toda a expressão deveria ser identificada como EM, desde que os elementos grafados em minúsculas integrassem a lista das minúsculas permitidas (cf. apêndice A, secção A.6), a qual foi criada para o efeito no âmbito desta avaliação (caso de *doença* em *doença de Chagas*).

Um outro caso em que decidimos refinar a identificação das EM está directamente relacionado com a representação de intervalos de valores e/ou especificação mais fina desses valores. Em particular, passámos a considerar intervalos de valores, tais como *entre 3 e 4%* ou *de 5 a 10 kg*, como uma única EM, e não duas como acontecia no Primeiro HAREM. Os quantificadores ou modificadores que permitem precisar o valor da entidade, como acontece em *cerca de 200 gramas*, *menos de 10%* ou *aproximadamente 15 euros*, também passaram a ser incluídos no âmbito da EM.

1.3.2 Representação sistemática das análises alternativas

No Primeiro HAREM, demos conta da possibilidade de uma dada sequência poder ser segmentada de forma distinta, nomeadamente nos casos em que essa sequência corresponde a uma EM estruturalmente ambígua, como ilustrado em (1.13), ou, numa outra perspectiva, quando não há certeza de que a sequência em análise corresponda efectivamente a uma entidade mencionada, explicitando-se, assim, a possibilidade de a mesma ser, ou não, identificada como EM, como ilustrado em (1.14).

(1.14) Portugal e Espanha, partiram à conquista do mundo e tornaram-se <ALT> <EM ID="a66435-5" CATEG="OUTRO">**Impérios**</EM> | Impérios </ALT> marítimos;

No Segundo HAREM, a etiqueta ALT passsou ainda a ser utilizada para representar, de forma sistemática, a estrutura interna das entidades constituídas por outras EM, como é o caso da EM que apresentamos em (1.15).

(1.15) <ALT> <EM ID="a55968-47" CATEG="PESSOA" TIPO="CARGO">**presidente da Câmara de Nova Iorque**</EM> | presidente da <EM ID="a55968-" CATEG="ORGANIZACAO" TIPO="ADMINISTRACAO">**Câmara de Nova Iorque**</EM> | presidente da <EM ID="a55968-475a" CATEG="ORGANIZACAO" TIPO="ADMINISTRACAO">**Câmara**</EM> de <EM ID="a55968-47" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="DIVISAO">**Nova Iorque**</EM> </ALT>

Este procedimento pode ser, de certo modo, encarado como uma forma de representar o encaixe de EM, uma situação não contemplada no Primeiro HAREM e que pode ter interesse a vários níveis. Por exemplo, além de permitir uma análise mais fina sobre o próprio mecanismo de composição de certas EM, possibilita a identificação de EM que, de outro modo, não seriam analisadas. Tendo em conta que uma das indicações fornecidas nas directivas do Primeiro HAREM apontava no sentido de marcar preferencialmente a EM mais longa (Cardoso e Santos, 2007), a identificação de, por exemplo, *Câmara de Nova Iorque* no exemplo acima não seria considerada. Isto poderia trazer inconvenientes, por exemplo, aos participantes que estivessem interessados em reconhecer especificamente organizações.

Apresentamos, no apêndice D, a lista de regras criadas para o efeito. Estas regras gerais foram, em alguns casos, refinadas, em função das propriedades lexicais e/ou semânticas dos constituintes de certas EM. Por exemplo, a regra PESSOA de LOCAL não deve ser empregue nos casos em que o indivíduo, referido pelo seu título nobiliário (que marcámos como CARGO) corresponde a uma das seguintes palavras: *conde*, *duque* e *marquês*. Esta opção deve-se ao facto de termos considerado como demasiado remota, e daí pouco pertinente, a relação que se estabelece entre a menção ao título e ao nome do local (caso de *Conde de Ourém*, *Duque de Bragança* e *Marquês de Pombal*). Não segmentámos também em constituintes menores as expressões classificadas como OBRA, se estas estiverem delimitadas por aspas ou plicas. Além disso, também não considerámos possível a segmentação de locais do tipo *Mosteiro dos Jerónimos*, no sentido em que se considera que é esta EM (CONSTRUCAO) que está na base da denominação de um dos seus constituintes, *Jerónimos* (DIVISAO), e não o contrário.[8]

1.4 Recursos

No Segundo HAREM, foram desenvolvidos e disponibilizados vários recursos, tanto para treino como para a avaliação propriamente dita dos sistemas. Para treino, disponibili-

[8] Para os nossos leitores não familiarizados com a história de Lisboa, convém talvez referir que o Mosteiro dos Jerónimos foi assim baptizado devido ao facto de este ter sido habitado pelos Jerónimos, os frades pertencentes à ordem de São Jerónimo, após ter sido erigido no século XVI. Actualmente, *Jerónimos* é usado (pelo menos, pelos lisboetas) para designar tanto o mosteiro como a zona onde este se encontra. Temos pois um caso em que o LOCAL vago *Jerónimos* provém do local (construção) *Mosteiro dos Jerónimos*, não sendo, por isso, parafraseável por "Mosteiro que se situa nos Jerónimos" (contrariamente, ao caso da *Torre de Pisa*, que é parafraseável por "Torre que se situa em Pisa").

zámos diferentes colecções anotadas de acordo com as directivas do HAREM clássico e da pista do TEMPO, assim como um *Exemplário* (cf. apêndice E), isto é, um conjunto de exemplos com EM ilustrativas de cada uma das categorias, tipos e subtipos previstos nas directivas do HAREM clássico (cf. apêndice A).

Para efectuar a própria avaliação, criámos a colecção do Segundo HAREM – a colecção que todos os sistemas tiveram de anotar – e a colecção dourada, um subconjunto da colecção do Segundo HAREM em que foi feita a anotação humana de tudo o que pretendíamos avaliar. Em seguida, descrevemos estes recursos com mais pormenor.

1.4.1 Constituição das colecções do Segundo HAREM

A colecção do Segundo HAREM é constituída por 1040 documentos (15737 parágrafos, 670610 palavras), entre os quais se encontram, como referimos antes, os documentos seleccionados para a colecção dourada. A colecção dourada é constituída por 129 documentos (correspondendo a 2274 parágrafos perfazendo 147991 palavras), representando cerca de 12% dos documentos que compõem a colecção do Segundo HAREM.

Os documentos da colecção do Segundo HAREM foram seleccionados tendo essencialmente em consideração os seguintes requisitos: (i) o português de Portugal e o do Brasil deveriam estar equitativamente representados na colecção, (ii) os documentos deveriam contemplar diferentes géneros e registos textuais, e (iii) a colecção deveria incluir algum material utilizado no Primeiro HAREM (nomeadamente, de forma a permitir comparar o desempenho dos sistemas nesses documentos) e noutras avaliações, como é o caso da colecção CHAVE (Santos e Rocha, 2005), a qual tem vindo a ser usada na avaliação de sistemas de respostas automáticas a perguntas (QA@CLEF (Giampiccolo et al., 2008)) e de recolha de informação geográfica. Neste último caso, os textos foram escolhidos com base na penúltima edição do GeoCLEF: para cada um dos 25 tópicos do GeoCLEF 2007 (Mandl et al., 2008), foram incluídos todos os documentos classificados como relevantes e dez documentos classificados como irrelevantes. Tal permitirá, no futuro, estudar, por exemplo, a influência e a relevância de REM na recuperação de informação geográfica.

A cada documento da colecção foram associadas diversas informações que caracterizam o documento. Entre outras propriedades, destacamos: variante de português, género e nome da fonte. A distribuição dos valores dessas propriedades na colecção do Segundo HAREM, bem como em cada uma das colecções douradas, encontra-se no apêndice H.

1.4.2 Processo de anotação da CD

A colecção dourada, como referimos anteriormente, constitui um subconjunto da colecção do Segundo HAREM, com base na qual os sistemas são avaliados. Numa primeira fase, o processo de anotação da CD foi cruzado, isto é, duas anotadoras anotaram o mesmo conjunto de textos. Esse processo foi levado a cabo com a ajuda da ferramenta Etiquet(H)AREM (ver apêndice F para informações mais detalhadas sobre esta ferramenta). As anotações foram posteriormente confrontadas/comparadas, recorrendo a um programa que apresentava as diferenças, com base na saída do programa `Alinhador` (capítulo 5). As diferenças encontradas por este programa foram então reanalisadas e discutidas pelas anotadoras (e, em alguns casos, por toda a organização), de forma a chegar a uma anotação consensual. Numa fase posterior, em que as directivas já se encontravam afinadas, os textos da CD passaram a ser alternadamente anotados por cada uma das

anotadoras. Casos problemáticos ou duvidosos eram expostos a (e discutidos por) toda a organização, de modo a tentar encontrar uma solução de anotação em que, pelo menos, a maioria estivesse de acordo.

Depois de anotada toda a CD, procedemos à sua revisão, a qual foi realizada em três fases distintas, mas complementares: numa primeira fase, levámos a cabo uma revisão sequencial dos documentos de toda a CD; seguidamente, efectuámos uma revisão fina e exaustiva das EM por categoria (tendo sempre, naturalmente, em conta o contexto em que estavam integradas), revisão essa levada a cabo por três pessoas[9]; finalmente, revimos especificamente os casos das EM compreendidas entre as etiquetas <ALT> e </ALT>.

Já após a apresentação dos resultados oficiais, mas antes da disponibilização dos recursos finais do Segundo HAREM, fizemos uma última revisão de todas as entidades espúrias nas participações dos sistemas, de modo a garantir, por um lado, que não tínhamos problemas que pudessem prejudicar indevidamente os sistemas, e, por outro, a disponibilizar um recurso final o mais correcto possível. Essa revisão foi feita por quatro pessoas (cada qual revendo um quarto dos quase 10 mil casos espúrios). Os casos problemáticos foram discutidos por toda a equipa, e aqueles que classificámos como erro foram alterados na CD de modo a produzir o recurso que reputamos de final[10].

O processo de anotação e revisão da CD levou à identificação de 7836 entidades mencionadas, distribuídas pelas diversas categorias, de acordo com o gráfico da figura 1.3(b). Observa-se que a categoria mais frequente na CD é a categoria PESSOA, seguida das categorias LOCAL, TEMPO e ORGANIZACAO, com proporções de 27,11%, 18,15%, 15,21% e 14,02%, respectivamente. De referir que, no Primeiro HAREM, a categoria com maior representatividade na CD do Primeiro HAREM é a categoria LOCAL (24,6%), seguida, respectivamente, de PESSOA (21,0%) e ORGANIZACAO (17,8%), como indicado na figura 1.2. Tendo em consideração que a análise do TEMPO mudou radicalmente de uma edição para a outra, a proporção de EM reconhecidas nas duas edições de avaliação (apenas 9,0%, no Primeiro HAREM) não é naturalmente comparável.

No que diz respeito à vagueza, se tivermos apenas em conta a categoria, 535 entidades são vagas (6,38% dos casos). No entanto, observa-se que 633 EM da CD correspondem a EM vagas quanto a pelo menos um dos atributos CATEG, TIPO ou SUBTIPO (cerca de 8% dos casos). Ao nível da categoria, foram identificadas 52 classes de vagueza, encontrando-se na tabela 1.4 todas as classes que ocorrem mais de duas vezes[11] e na figura 1.4 a distribuição das categorias vagas. Na sua grande maioria (91,8% dos casos), a vagueza estabelece-se entre duas categorias. Os três casos mais frequentes foram: LOCAL|ORGANIZACAO (23,18% das entidades vagas), ORGANIZACAO|PESSOA (14,02%) e ABSTRACCAO|PESSOA (10,66%).

Relativamente às análises alternativas de identificação, observa-se que 372 sequências podem ser segmentadas de duas formas distintas, registando-se que apenas 11 sequências se encontram associadas a três possibilidades alternativas de segmentação. Das 7836 entidades existentes na CD, 1022 encontram-se dentro de um ALT (cerca de 13,8%).

Os casos acordados por maioria, e não por unanimidade (122 casos), foram devidamente identificados na CD, através da notação 2/3, que foi guardada no campo COMENT (um atributo opcional previsto na sintaxe de anotação das EM). A tabela 1.1 ilustra os casos de discordância registados. Nos casos em que não foi possível encontrar uma classificação

[9] E que, por essa razão, permitiu a marcação dos casos de decisão por maioria como 2/3.

[10] De referir, no entanto, que os resultados oficiais do Segundo HAREM se baseiam na CD que divulgámos no momento próprio, e, portanto, as mudanças referidas não influenciam a avaliação.

[11] Embora a tabela não mostre, verifica-se também vagueza entre 4 e 5 categorias.

Tabela 1.1: Distribuição de categorias, e discordância na anotação: D2/3 - Número de vezes em que a decisão de anotação não foi unânime; % - percentagem de entidades dessa categoria em que a decisão não foi unânime; DT: Número de vezes em que não houve acordo quanto à categoria

Categoria	Quant.	D2/3	%	DT
PESSOA	2036	13	0,64	2
LOCAL	1311	15	1,14	-
TEMPO	1189	35	2,94	-
ORGANIZACAO	961	16	1,66	2
OBRA	449	5	1,11	5
VALOR	353	-	-	-
COISA	308	5	1,62	1
ACONTECIMENTO	300	-	-	-
ABSTRACCAO	286	2	0,7	-
LOCAL|ORGANIZACAO	124	2	1,61	-
OUTRO	79	4	5,06	-
ORGANIZACAO|PESSOA	75	2	2,67	1
ABSTRACCAO|PESSOA	57	2	3,51	-
LOCAL|OBRA	33	1	3,03	-
ABSTRACCAO|ORGANIZACAO	31	4	12,9	-
EM	29	-	-	-
COISA|OBRA	24	1	4,17	-
LOCAL|PESSOA	14	-	-	-
COISA|LOCAL	14	7	50	-
OBRA|ORGANIZACAO	12	-	-	-
ACONTECIMENTO|LOCAL	11	-	-	1
ABSTRACCAO|LOCAL	11	-	-	-
ACONTECIMENTO|OUTRO	10	-	-	-
LOCAL|ORGANIZACAO|PESSOA	9	-	-	-
ACONTECIMENTO|OBRA	9	1	11,11	-
ABSTRACCAO|ACONTECIMENTO	9	-	-	-
COISA|ORGANIZACAO	8	-	-	-
ABSTRACCAO|COISA	6	-	-	-
ACONTECIMENTO|PESSOA	6	-	-	1
COISA|PESSOA	6	-	-	-
ABSTRACCAO|ACONTECIMENTO|ORGANIZACAO	6	-	-	-
ABSTRACCAO|ORGANIZACAO|PESSOA	4	2	50	-
COISA|OUTRO	4	1	25	-
TEMPO|VALOR	4	-	-	-
ACONTECIMENTO|ORGANIZACAO	3	-	-	-
LOCAL|OUTRO	3	-	-	1
ABSTRACCAO|OBRA	3	-	-	-
OBRA|PESSOA	3	-	-	-
OBRA|OUTRO	2	1	50	-
ABSTRACCAO|OUTRO	2	1	50	-
ABSTRACCAO|LOCAL|PESSOA	2	1	50	-
Outros casos de vagueza que ocorrem 2 vezes	16	-	-	-
Outros casos de vagueza que ocorrem 1 vez	14	-	-	-

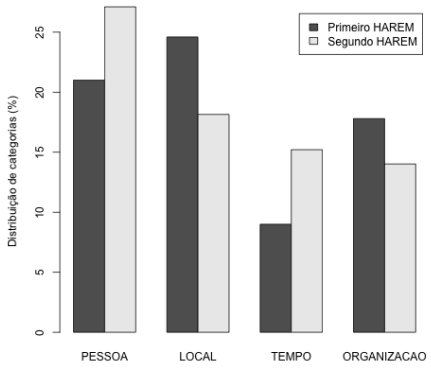
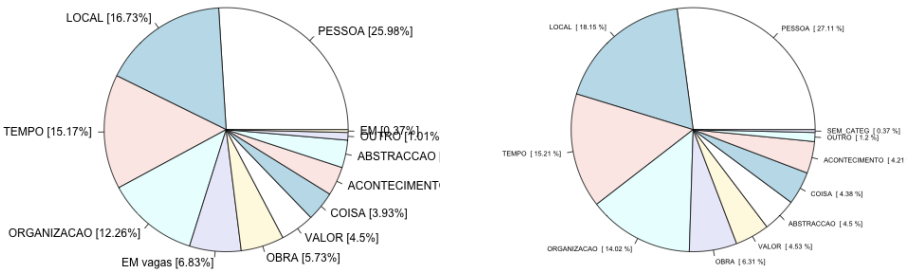


Figura 1.2: Distribuição das categorias mais frequentes na CD do HAREM em comparação com as mesmas categorias na CD do Primeiro HAREM



(a) A combinação de categorias de uma entidade vaga conta com uma única categoria, não contribuindo para cada categoria individualmente

(b) Para esta contabilização, cada categoria de uma entidade vaga contribuiu com $1/n$, sendo n número de categorias vagas dessa entidade

Figura 1.3: Distribuição de categorias na CD

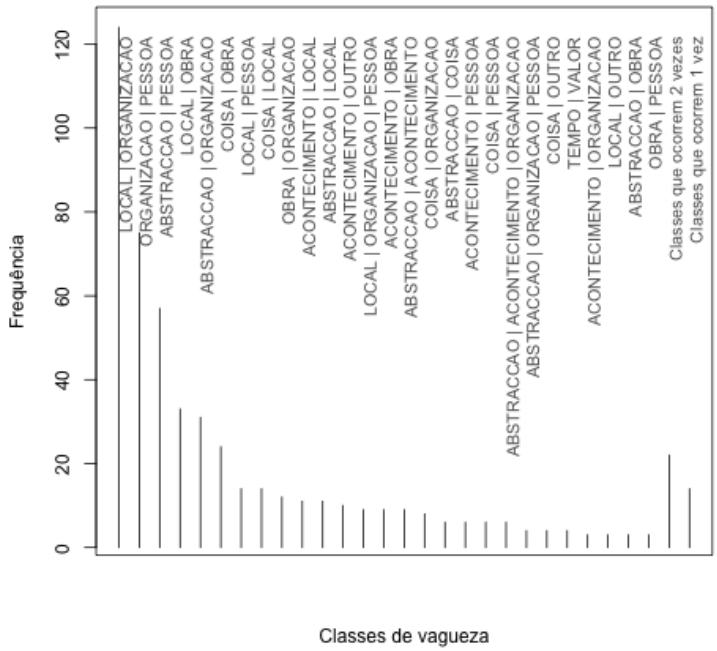


Figura 1.4: Distribuição das categorias vagas

consensual ou maioritária, optámos por omitir as EM em ambas as colecções (fazendo uso das etiquetas <OMITIDO> e </OMITIDO>), de modo a que as mesmas deixassem de ser alvo de avaliação (101 casos). De referir ainda que, em alguns casos, a discussão mostrou que as diferentes análises de interpretação em discordância eram possíveis, o que significa que todas elas passaram a ser representadas na CD, tirando partido dos mecanismos já anteriormente descritos para representação de EM vagas ou de EM que podem fazer parte de análises alternativas (em termos de segmentação).

1.5 Resultados da avaliação

Uma característica que consideramos inovadora e essencial no modelo de avaliação do HAREM diz respeito à flexibilidade oferecida aos sistemas em termos de participação e avaliação. Em concreto, os sistemas têm a possibilidade de escolher as categorias, tipos, subtipos ou outros atributos que pretendem etiquetar e ver avaliados, em função do interesse, pertinência ou adequação que essas anotações possam ter no âmbito de outras aplicações desenvolvidas ou a desenvolver por parte dos participantes, e que dependem directa ou indirectamente dessas informações. A cada conjunto diferente de categorias a que os participantes se propuseram ser avaliados (que aprofundaremos mais adiante), demos o nome de **cenário selectivo de participação**.

Tabela 1.2: Sistemas participantes no HAREM clássico e dados de participação

Sistema	N. corridas	Cenário	ALT
CaGE2	4	Selectivo 2	-
DobrEM	1	PESSOA	-
PorTexTO	4	TEMPO	-
Priberam	1	Total	-
R3M	2	Selectivo 3	-
REMBRANDT	3	Total	Sim
REMMA	3	Selectivo 4	Sim
SEI-Geo	4	Selectivo 5	-
SeRELeP	1	Total só Id	-
XIP-L2F/XEROX	4	Selectivo 6	-

Além disso, no Segundo HAREM implementámos outro tipo de cenários, os **cenários selectivos de avaliação**, que permitem a avaliação num subconjunto de categorias e tipos que não necessariamente o proposto pelo sistema.

A avaliação em cenários selectivos permite, entre outros aspectos, comparar o desempenho dos diferentes sistemas com base em cada uma das categorias que se propuseram reconhecer, assim como noutros conjuntos de categorias que possam fazer sentido.

Dito de outro modo, a avaliação levada a cabo no HAREM não se cinge a avaliar sistemas no âmbito de uma tarefa geral de REM, mas também, e fundamentalmente, a analisar mais detalhadamente o comportamento dos sistemas em tarefas mais específicas, previamente definidas pelos participantes, no âmbito da tarefa geral proposta pela organização. Deste modo, torna-se igualmente possível comparar os sistemas em cenários diferentes do cenário para o qual foram desenvolvidos.

Assim, todos os sistemas foram avaliados no cenário total e em cada um dos cenários selectivos de participação descritos na tabela 1.2. Além disso, todos os sistemas foram avaliados por categoria, o que corresponde a fazer a avaliação utilizando um cenário selectivo constituído apenas por cada uma dessas categorias. Em qualquer dos cenários referidos, os sistemas foram avaliados com avaliação estrita e relaxada de ALT (cf. capítulo 5).

O modelo e programas de avaliação do Segundo HAREM encontram-se descritos em detalhe no capítulo 5. Nesta secção, apenas apresentamos os sistemas participantes no HAREM clássico e os resultados de desempenho das corridas enviadas por esses sistemas.

1.5.1 Sistemas participantes

A tabela 1.2 mostra os dez sistemas participantes (que em conjunto enviaram 27 corridas[12]) e outros dados referentes à forma de participação. Por exemplo, se fez apenas identificação ou também classificação, e quais os cenários em que concorreu[13]. Como ilustra o quadro, os participantes envolveram-se de formas muito distintas na tarefa de reconhecimento de entidades mencionadas, uma situação que pode ter sido motivada pelo facto de o HAREM permitir a avaliação por cenários selectivos.

[12] Cada participante podia enviar no máximo quatro corridas.

[13] Ou seja, na terminologia técnica do HAREM, o cenário selectivo de participação de cada corrida (ver capítulo 5).

Tabela 1.3: Cenários de participação: I - apenas EM; C - classificação usando todos os atributos; CAT - apenas CATEG; CAT/T - sem SUBTIPO; F+H - LOCAL cujo TIPO seja FISICO e HUMANO

Cenário	PES	ORG	LOC	OBR	ACO	ABS	COI	TEM	VAL
PESSOA	I								
TEMPO								C	
Selectivo 2	CAT	CAT	F + H					CAT	
Selectivo 3	I	I	I	I	I	I	I		
Selectivo 4	C	C	CAT/T	C	C	C	C	CAT/T	C
Selectivo 5			F + H						
Selectivo 6	C	C	C	C	C			C	C
Total	C	C	C	C	C	C	C	C	C
Total só Id	I	I	I	I	I	I	I	I	I

1.5.2 Resultados

Apesar da diversidade da participação, a tarefa alvo em avaliação é o reconhecimento de entidades mencionadas. Como tal, começamos por analisar o desempenho dos sistemas no reconhecimento de todas as entidades existentes na CD, em termos de medida F, precisão e abrangência, no cenário total com avaliação estrita de ALT (figura 1.5[14]).

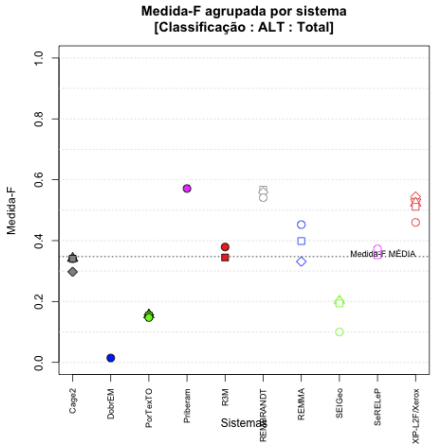
Note-se que não levámos a cabo, por enquanto, nenhum estudo estatístico dos resultados, como será referido no capítulo 6, e por isso a análise apresentada aqui será apenas uma primeira análise, bastante superficial.

O sistema da Priberam (cf. capítulo 9) foi o sistema com melhor medida F (0,5711), tendo ficado, no entanto, muito próximo do segundo melhor sistema, o REMBRANDT (cf. capítulo 11), cuja melhor corrida obteve 0,5674. Estes dois sistemas juntamente com o XIP-L2F/Xerox foram os únicos a obter valores de medida F superiores a 0,5.

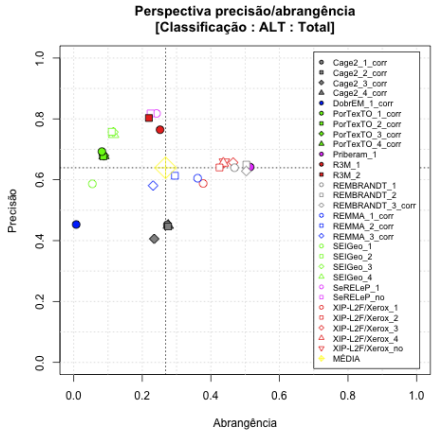
Relativamente às restantes corridas, apenas uma, enviada pelo REMMA (cf. capítulo 12) teve uma medida F superior a 0,4. De referir, no entanto, que isso tem naturalmente a ver com o facto de os cenários de participação dos restantes sistemas incluirem menos categorias (como é o caso do CaGE2 (cf. capítulo 7)) ou menos subtipos (caso do REMMA) e de alguns desses sistemas (caso do R3M (cf. capítulo 10) e do SeRELeP (cf. capítulo 14)) só terem feito identificação de entidades.

Uma explicação que se impõe em relação à interpretação dos resultados prende-se com justificar por que razão, na avaliação da classificação, sistemas que fizeram unicamente identificação têm valores de medida F próximos dos valores de sistemas que fizeram classificação. Compare-se, por exemplo, o desempenho dos sistemas R3M e SeRELeP, que fizeram apenas identificação, com o do sistema REMMA, que também fez classificação. Ao observarmos o gráfico que representa os resultados da avaliação da identificação (figura 1.5(c)), verificamos que os sistemas R3M e SeRELeP se encontram entre os melhores, o que não acontece com o sistema REMMA, que tem claramente um pior desempenho na identificação, o que também se reflecte na avaliação da classificação. Assim, podemos desde já afirmar que ainda estamos insatisfeitos com o peso atribuído à identificação, que acaba por penalizar indevidamente sistemas que fazem classificação – veja-se o capítulo 6 para mais discussão sobre este assunto.

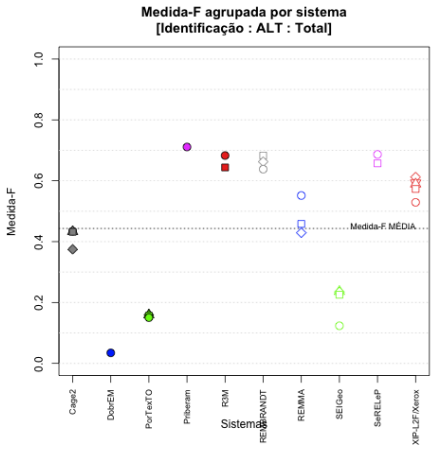
[14] Os valores correspondentes a esta figura e seguintes encontram-se no apêndice I (e no sítio do HAREM).



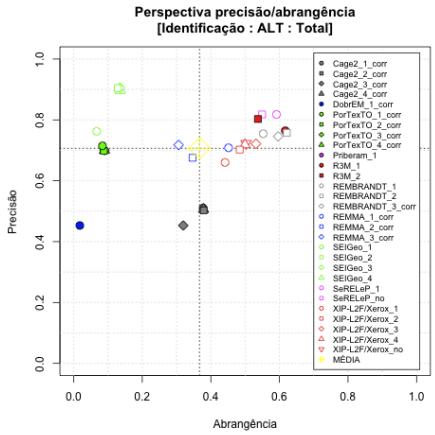
(a) Medida F na classificação



(b) Precisão e abrangência na classificação

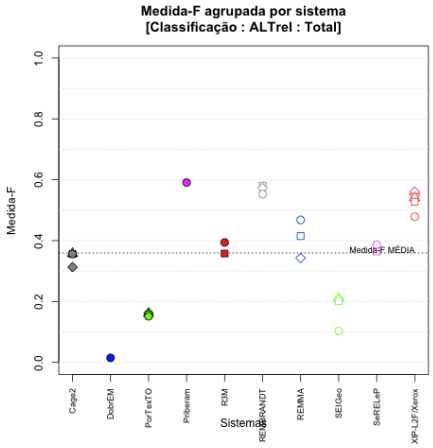


(c) Medida F na identificação

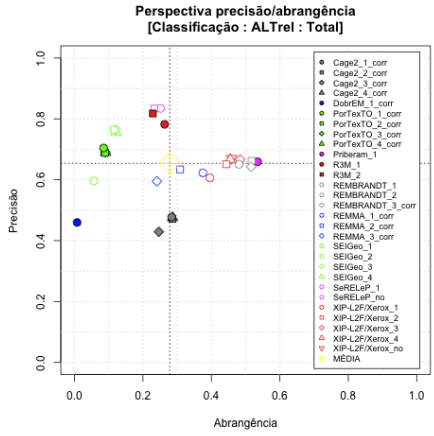


(d) Precisão e abrangência na identificação

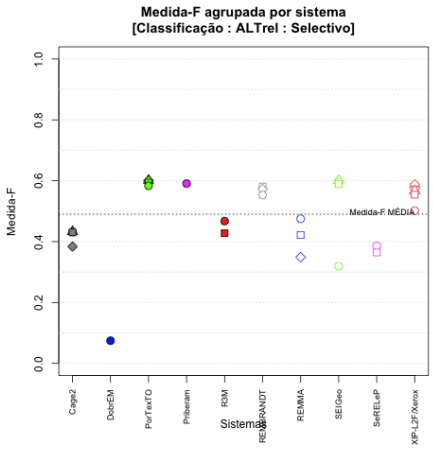
Figura 1.5: Avaliação no cenário total com avaliação estrita de `ALT`



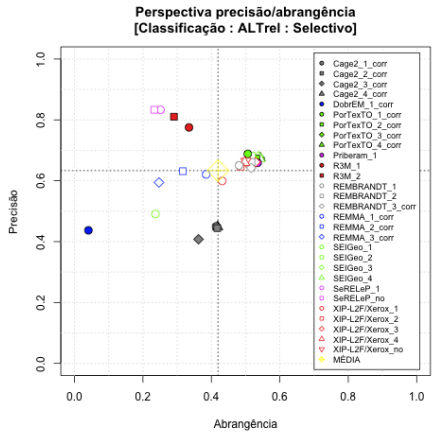
(a) Medida F no cenário total



(b) Precisão e abrangência no cenário total



(c) Medida F no cenário selectivo



(d) Precisão e abrangência no cenário selectivo

Figura 1.6: Classificação com avaliação relaxada de ALT

Relativamente ao desempenho na classificação com avaliação relaxada de ALT, vemos, na figura 1.6(a), que a medida F melhora ligeiramente para todos os sistemas. Em particular, os melhores sistemas, o sistema da Priberam e a melhor corrida do REMBRANDT, obtêm 0,5908 e 0,5808, respectivamente, aumentando um pouco mais a diferença de desempenho entre os dois sistemas. Esse aumento deve-se ao facto de apenas o sistema REMBRANDT ter utilizado ALT nas suas corridas.

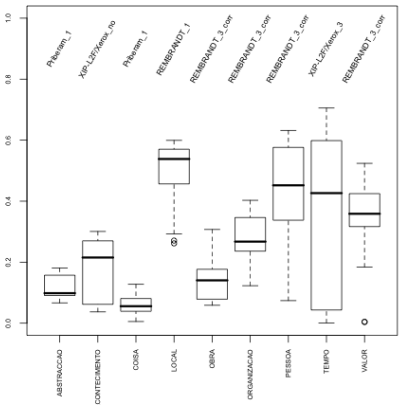
Analisemos agora o desempenho dos sistemas nos seus cenários selectivos, também tendo em conta a avaliação relaxada de ALT (já que apenas o sistema REMBRANDT e REMMA fizeram marcação de análises alternativas). Enquanto os gráficos anteriores ilustravam uma situação em que os sistemas estavam a ser todos avaliados no mesmo cenário, o cenário total, o que naturalmente desfavorece os sistemas que não participaram em todas as categorias, a figura 1.6(c) compara os sistemas tendo em consideração os respectivos cenários selectivos de participação.

Como seria de esperar, os sistemas que têm cenários de participação coincidentes com o cenário total, como seja o REMBRANDT e o da Priberam, não sofreram quaisquer alterações. Quanto aos restantes sistemas, vemos claramente melhores valores de medida F, sobretudo no caso de sistemas como o PorTexTO (cf. capítulo 8) e o SEI-Geo (cf. capítulo 13), que tentaram reconhecer apenas uma categoria, respectivamente TEMPO e LOCAL. Isto significa que, em relação ao objectivo que se propuseram alcançar, obtiveram um desempenho equiparável ao de outros sistemas que tinham objectivos mais ambiciosos. Ou, por outras palavras, estes sistemas podem ter reconhecido apenas uma categoria, mas, em termos relativos, foram tão bons a executar esse reconhecimento como os sistemas que tentaram reconhecer várias categorias.

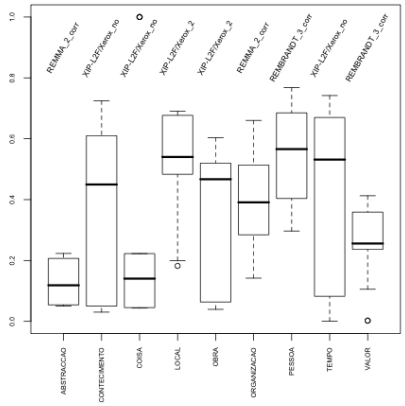
Com isto, não estamos a dizer que, no caso das categorias TEMPO e LOCAL, os sistemas PorTexTO e SEI-Geo, respectivamente, foram os melhores a reconhecer entidades com essa categoria. De facto, não o foram, como se pode ver na figura 1.7, que apresenta os melhores sistemas em cada uma das categorias. No caso da categoria TEMPO, o melhor sistema foi o XIP-L2F/Xerox (corrida 3), com 0,7054, que foi também o melhor sistema a reconhecer entidades ACONTECIMENTO; quanto à categoria LOCAL, o melhor sistema foi o sistema REMBRANDT (corrida 1), com 0,5993, que também foi, aliás, o melhor sistema, embora com uma corrida diferente, a reconhecer as restantes categorias, excepto ABSTRACCAO e COISA. Nestes últimos casos, o melhor sistema foi o da Priberam.

Se pensarmos que o melhor desempenho no reconhecimento de uma categoria traduz a facilidade no reconhecimento dessa categoria, podemos concluir que a entidade mais fácil de identificar é TEMPO, pois foi aquela onde foi obtido o melhor desempenho, imediatamente seguida de PESSOA e LOCAL. Nesta linha de interpretação, entidades como ABSTRACCAO e COISA seriam as mais difíceis de reconhecer, o que de certo modo faz algum sentido, na medida que se tratam de entidades mais abstractas ou, noutra perspectiva, mais abrangentes, e, por isso, mais difíceis de modelar.

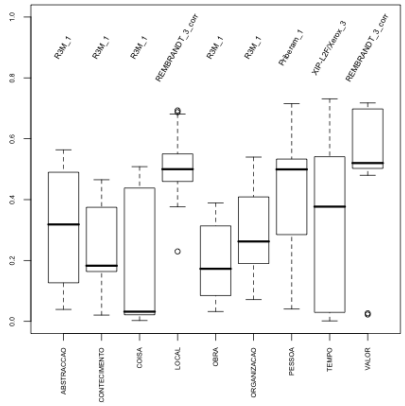
Parece-nos, no entanto, que a categoria onde houve de facto mais sucesso foi LOCAL. Algo que não é completamente surpreendente, uma vez que os autores de três sistemas participantes se dedicam a reconhecimento geográfico. Note-se, por exemplo, que a grande maioria das corridas obteve valores de medida F acima de 0,5, e que o pior sistema tem melhor desempenho na categoria LOCAL do que a maioria dos sistemas noutras categorias, sendo mesmo o melhor desempenho entre os piores das várias categorias. Esta situação contrasta com o desempenho na categoria TEMPO, onde se observa que a maioria dos sistemas está abaixo de 0,5 e onde se verifica uma maior dispersão dos valores, apesar



(a) Medida F



(b) Precisão



(c) Abrangência

Figura 1.7: Resumo de estatísticas da avaliação por categorias com avaliação estrita de `ALT`: máximo, mínimo, mediana, primeiro e terceiro quartis.

do melhor sistema ter obtido acima de 0,7.

Resta referir que estamos conscientes de que esta análise é bastante superficial e que, antes de tecer quaisquer conclusões definitivas sobre o que é fácil ou difícil, é também necessário fazer uma análise sistemática e aprofundada dos textos anotados, que passa, nomeadamente, pelo estudo das discordâncias de interpretação de certas entidades. Remetemos, pois, o leitor para o capítulo 6 para mais considerações sobre estas questões.

Capítulo 2

Identificação, classificação e normalização de expressões temporais do português: A experiência do Segundo HAREM e o futuro

Jorge Baptista, Caroline Hagège e Nuno Mamede

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 2, p. 33–54.

Neste capítulo apresentamos a proposta que elaborámos (Hagège et al., 2008) para a tarefa de reconhecimento, classificação e normalização de expressões temporais (ET) no âmbito da segunda avaliação conjunta de sistemas de reconhecimento entidades mencionadas (EM) do português – o Segundo HAREM. Procurámos, além disso, reflectir sobre a experiência desta avalição conjunta para que, baseados na forma de intervenção dos vários sistemas participantes e nos resultados globais desta pista de avaliação, pudéssemos sugerir futuros desenvolvimentos e novas iniativas de avaliação desta tarefa.

2.1 Introdução

2.1.1 Generalidades

O reconhecimento, classificação e representação das ET não é uma tarefa trivial. Apesar de o conjunto de elementos lexicais (pelo menos em termos de palavras simples) envolvidos ser relativamente extenso, é, ainda assim, suficientemente bem limitado para que se conceba como meta exequível atingir-se uma cobertura lexical próxima da exaustividade. Já o mesmo não se passa com o conjunto de construções em que se podem combinar estes elementos lexicais associados à expressão do tempo, que poderão representar várias centenas de construções diferentes[1], os quais se podem combinar entre si segundo padrões sintácticо-semânticos que, tanto quanto sabemos, ainda não foram sistematicamente recenseados.

Este tipo de expressão apresenta também a dificuldade suplementar que resulta na diversidade de valores semânticos (interpretação) que podem ser associados aos elementos gramaticais ou formais que introduzem a expressão temporal. Assim, por exemplo, nas ET *no próximo ano* e *em duas semanas*, não é possível fazer depender apenas da presença da preposição *em* a interpretação global de cada uma destas expressões. Pelo contrário, só levando em consideração toda a expressão bem como o preenchimento lexical das várias posições estruturais (preposição, determinante, nome de tempo e eventual modificador) é possível classificá-las de forma adequada, nomeadamente, considerando a primeira ET como uma **data** (exemplo (2.1)) e a segunda como uma **duração** (exemplo (2.2)).

(2.1) O João só vai fazer isso *no próximo ano*.

(2.2) O Pedro concluiu a tarefa *em duas semanas*.

Além disso, e como em muitos outros aspectos da linguagem natural, verifica-se um determinado grau de vagueza na interpretação de muitas ET. Assim, por exemplo, uma ET como *há dois anos* deverá ser interpretada como se referindo ao intervalo de tempo entre 1 de Janeiro e 31 de Dezembro de 2006 ou a uma data exacta nesse ano, mas relativamente ao momento da enunciação (*hoje*)? Repare-se que as línguas têm geralmente mecanismos (quantificadores) que tanto permitem controlar (contrariar?) como reforçar esta dimensão (vagueza) intrínseca do discurso:

(2.3) O João fez isso *há precisamente/aproximadamente/mais de dois anos*.

(2.4) O João fez isso *há imensos/vários/alguns/poucos/uns poucos de anos*.

[1] Como exemplo de uma exploração sistemática de famílias de expressões temporais em português, veja-se, entre outros, Móia (2000) e Baptista (2003).

A indefinição, a que acima nos referimos, poderá eventualmente ser esclarecida pelo contexto comunicativo ou discursivo. Contudo, noutros casos, ela é um mecanismo expressivo da língua, dando origem a formas cuja interpretação não é necessariamente literal, como acontece em situações de hipérbole (como em (2.5)) ou de eufemismo (como em (2.6)).

(2.5) O Pedro fez isso *há séculos/mais de três quinze dias*!

(2.6) Espera só *um minuto* que eu já te faço isso.

Finalmente, salientamos que uma adequada interpretação das ET depende muitas vezes da frase em que se insere. Assim, por exemplo, até uma data como *5 de Dezembro* só pode ser localizada relativamente ao momento da enunciação se se levar em conta o tempo-modo do verbo que a ET modifica:

(2.7) O avião aterrou em Lisboa *no dia 5 de Dezembro*.

(2.8) O avião vai aterrar em Lisboa *no dia 5 de Dezembro*.

Por outro lado, esta expressão tem, nas frases acima, um valor aspectual pontual, resultado da combinatória com um predicado como *aterrar* (avião); se se tratar de outro tipo de predicado, com outro valor aspectual, a modificação que o advérbio exerce parece ser aspectualmente diferente:

(2.9) O Pedro esteve em casa doente *no dia 5 de Dezembro*.

Um caso semelhante, ocorre nas construções temporais com *haver*, que podem ter leituras diferentes consoante o tempo-modo do verbo da frase que modificam: **data** em (2.10) e **duração** em (2.11).

(2.10) O João fez isso *há 5 anos*.

(2.11) O João faz isso *há 5 anos*.

A proposta de avaliação da categoria TEMPO apresentada ao Segundo HAREM procurou abordar algumas destas questões, dando particular ênfase ao tratamento da referência e tentando contribuir no sentido da construção de um standard de normalização das ET.

2.1.2 Motivação da proposta

Com a normalização de ET, temos como objectivo final a tarefa, bem mais complexa, de reconhecer as ET presentes no texto para as associar aos eventos e estados de coisas que aquelas modificam, de modo a podermos ordenar parcialmente, segundo uma sequência cronológica, esses mesmos eventos e estados de coisas.

Naturalmente, esta meta constitui um objectivo demasiado ambicioso, em particular no quadro de um evento como o HAREM, cujo foco é o reconhecimento e a classificação de EM. Pretendemos, pois, com a nossa proposta dar um passo naquela direcção, passando pela incontornável tarefa de reconhecimento e classificação de ET, na continuidade do Primeiro HAREM, ao mesmo tempo que fazemos uma primeira abordagem a um dos

grandes problemas levantados por este tipo de expressões, nomeadamente o problema da referência temporal.

A proposta de reconhecimento, classificação e normalização de expressões temporais que fizemos no âmbito do Segundo HAREM (Hagège et al., 2008) encontra a sua principal motivação em trabalhos recentes e num interesse renovado da comunidade do processamento de linguagem natural (PLN) pela problemática do tratamento do tempo, no domínio mais vasto da extracção de informação. Com efeito, é necessário tomar em conta a dimensão temporal veiculada nos textos para levar a cabo de maneira satisfatória diversas tarefas que visam a extracção de informação a partir de textos. Por exemplo, as respostas a perguntas como *Qual é a capital da Alemanha?* *Quem era o vice presidente de Bush?* serão diferentes conforme os momentos da história a que se possam referir e, naturalmente, consoante a data dos textos que estarão acessíveis para poder responder a estas perguntas. Para aplicações de PLN que trabalham com vários documentos como, por exemplo, a sumarização, uma representação adequada da dimensão temporal dos textos deverá permitir relacionar entre si os eventos neles referidos.

Vários indicadores mostram o interesse crescente na área do processamento do tempo: é disto exemplo a primeira avaliação conjunta TempEval[2], em 2007 (Verhagen et al., 2007), que teve lugar no âmbito da conferência Senseval 2007[3]. A Google também oferece na Google Trends[4] a possibilidade de visualizar o resultado de uma pesquisa usando a dimensão temporal. Além do mais, já foram feitas propostas para anotação fina de ET e, para o inglês, existem alguns recursos, tais como os textos anotados com a norma TimeML (Saurí et al., 2006)[5]. Para outras línguas (o francês e o romeno, pelo menos), estão já em desenvolvimento diversos trabalhos nesta área (ver, por exemplo, Battistelli et al. (2008)).

Pareceu-nos importante abordar este problema para o português e a avaliação conjunta do HAREM constituiu uma excelente plataforma para o fazer, embora a nossa proposta ultrapasse o quadro estrito de reconhecimento de entidades mencionadas (REM).

2.1.3 Questões operacionais da proposta

Nesse sentido, na elaboração da proposta, procurámos seguir alguns princípios norteadores que aqui apresentamos sucintamente, embora tenhamos de retomar alguns deles mais adiante:

(i) uma tarefa executável em seis meses de desenvolvimento, a fim de permitir não só a continuidade dos anteriores participantes, dando-lhes tempo de reconverterem os seus sistemas, se necessário, mas também incentivar a participação de novos actores;

(ii) compatibilidade com propostas já existentes, garantido uma continuidade natural com a tarefa da anterior edição do HAREM (Cardoso e Santos, 2007), aproximando-a ou adaptando-a, no entanto, aos standards que se estão a constituir em torno das mais recentes avaliações conjuntas internacionais;

(iii) limitação da dependência entre eventos e ET, procurando minimizar as por vezes complexas interacções entre o tipo de construção e a ET que a modifica;

[2] http://www.timeml.org/tempeval/
[3] http://nlp.cs.swarthmore.edu/semeval/
[4] http://www.google.com/trends
[5] http://www.timeml.org/site/index.html

(iv) independência entre a tarefa de delimitação das ET e o tratamento da subcategorização verbal, o que nos levou a propor a inclusão de certas preposições na EM;

(v) adopção de critérios claros de atomização das ET;

(vi) adesão ao princípio de classificar antes de resolver a referência temporal;

(vii) normalização parcial das ET, isto é, apresentar para um conjunto de situações, suficientemente claras, uma proposta de normalização, deixando para momento posterior o tratamento de outras expressões; do mesmo modo, permitir que uma expressão para a qual está disponível apenas parte da informação necessária à sua adequada normalização seja, ainda assim, normalizada pelo menos parcialmente;

(viii) os agregados temporais[6] não são, por ora, considerados, dada a sua especifidade;

(ix) tentar assegurar o critério de intersubjectividade máxima na anotação, procedendo sempre que possível à listagem e/ou descrição intensional dos elementos lexicais que entram na formação das ET.

2.2 Proposta para o Segundo HAREM

2.2.1 Delimitação das ET

A fim de se poder anotar de maneira unívoca as entidades da categoria TEMPO, convém definir rigorosamente os critérios sintáctica e semanticamente motivados que deverão ser seguidos a fim de se poder delimitar com precisão as fronteiras das entidades a anotar. Neste sentido, a proposta que apresentámos representa uma evolução e modificação importantes relativamente à estratégia adoptada no Primeiro HAREM (Cardoso e Santos, 2007, pp. 223-225).

Assim, nesta proposta, considera-se que deverá ser delimitada entre as balizas <EM ID=... CATEG="TEMPO"> e </EM> a totalidade da expressão temporal, isto é, **incluindo a preposição** que a introduzir, no caso da expressão temporal ser um sintagma preposicional (e.g. *no ano passado*), **ou o determinante** no caso de ser um sintagma nominal (e.g. *todos os dias*).

Por detrás desta opção está a noção de que na maioria das ET, os elementos ditos gramaticais (preposições e determinantes, sobretudo) são não apenas parte integrante destas locuções, apresentando muitas delas um elevado grau de fixidez combinatória interna, como contribuem de modo crucial para a classificação das ET nos diferentes tipos da categoria TEMPO. Naturalmente, este tipo de decisão acarretou, sobretudo por uma questão de coerência mas também de simplicidade, que se incluíssem nas EM certas preposições que não fazem parte da ET propriamente dita mas que são seleccionadas (regidas) por outros elementos lexicais (operadores). Tal sucede, sobretudo, nos casos das ET genéricas, como se pode ver em (2.12).

(2.12) Eu gosto <EM ID="..." CATEG="TEMPO" TIPO="GENERICO">**do Verão**</EM>.

[6] Um agregado temporal é uma expressão complexa que inclui simultaneamente valores de DATA e de FREQUENCIA, como, por exemplo: no primeiro domingo de cada mês.

A preposição *de*, neste caso, enquanto elemento que introduz o complemento de *gostar*, em nada contribui para a interpretação da EM. Considerámos, no entanto, que o tratamento das regências verbais (para usar um termo mais tradicional) deveria constituir um problema distinto, a resolver independemente do reconhecimento das EM.

2.2.2 Delimitação das ET complexas

Decidimos também integrar na EM certos elementos gramaticais, tradicionalmente analisados como advérbios, que entram na formação de ET complexas:

(2.13) O Pedro fez isso <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**alguns dias depois**</EM>.

De facto, este tipo de ET complexa é formado por dois elementos: uma expressão quantificadora do tipo DURACAO (*alguns dias*) e o adverbial *depois*. Esta última forma pode introduzir outros constituintes ligando-se-lhes por meio da preposição *de* e, assim, receber diferentes análises consoante seja seguida de uma oração (conjunção subordinativa temporal), como em (2.14), ou de um grupo nominal (locução prepositiva ou preposição composta), como em (2.15).

(2.14) O Pedro fez isso <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**alguns dias**</EM> depois de ter ido ver o futebol.

(2.15) O Pedro fez isso <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**alguns dias**</EM> depois do jogo.

Na construção com a locução prepositiva, distinguimos ainda duas situações: a primeira, como no exemplo (2.15), em que o núcleo do sintagma nominal é um nome qualquer; e uma segunda situação, ilustrada no exemplo (2.16), em que esse sintagma é preenchido por um nome de tempo (voltaremos a este último caso já adiante).

(2.16) O Pedro fez isso <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**alguns dias depois do domingo**</EM>.

Considerámos, por princípio, que não nos cabia propor qualquer análise unificada (e coerente) deste tipo de fenómeno, mas sim determinar com rigor as regras de delimitação das EM.

Assim, uma vez que, nesta fase de desenvolvimento da tarefa do HAREM dedicada à categoria TEMPO, tomámos a decisão de excluir as orações subordinadas, apenas utilizamos a informação da conjunção para determinar o atributo SENTIDO com que será anotada a EM (ver adiante).

No caso da locução prepositiva, seguimos critério idêntico, excluindo apenas os casos que envolvem um complemento com nomes de tempo, na medida em que estas expressões complexas exigem uma análise mais subtil.

De facto, no caso de expressões complexas como *dois dias depois do Natal*, a questão que se coloca é a de se saber se esta expressão deverá ser considerada como uma só EM ou, então, segmentada em duas subexpressões *dois dias* + *depois do Natal* (obedecendo tanto a

expressão mais longa como ambas as subexpressões aos critérios definitórios mencionados acima). Neste sentido, verifica-se que uma expressão como *dois dias depois do Natal*, ilustrada no exemplo (2.17) é ambígua podendo ter duas leituras distintas, a que correspondem duas análises sintácticas diferentes (e logo diferentes atomizações).

(2.17) Vimo-nos *dois dias depois do Natal*.

(a) *Vimo-nos no dia 27 de Dezembro*
Vimo-nos <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**dois dias depois do Natal**</EM>.

(b) *Vimo-nos durante dois dias, a seguir ao 25 de Dezembro*
Vimo-nos <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">**dois dias**</EM> <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**depois do Natal**</EM>.

Os critérios adoptados para a segmentação são os definidos em Hagège e Tannier (2007) e que aqui foram reproduzidos:

> Uma expressão temporal complexa **deverá ser dividida** em unidades menores se se verificarem **simultaneamente** os critérios seguintes:
>
> 1. cada expressão componente é sintacticamente válida quando combinada independentemente com o evento que modifica.
> 2. cada expressão componente, combinada com o evento que modifica, está logicamente implicada na expressão complexa. Ou seja, cada combinação "evento mais expressão_temporal_mínima" deve ser logicamente implicada pela combinação "evento + expressão_temporal_complexa".

Ora, no caso da frase ambígua (2.17), o primeiro critério pode aplicar-se tanto na leitura (a) como na leitura (b), acima:

> Vimo-nos *dois dias* (DURACAO).
> Vimo-nos *depois do Natal* (DATA).

mas o segundo critério não se observa, já que o valor de duração está ausente da leitura complexa (a), que acima glosamos. Ainda assim, neste caso, parece-nos que, embora a presença do segundo membro tenha tendência em "forçar" a leitura complexa da expressão temporal (DATA), em última análise, a ambiguidade deverá ficar expressa na anotação a adoptar futuramente.

As expressões de tempo foram organizadas em quatro grandes tipos:

- as expressões de **localização temporal**, de tipo TEMPO_CALEND;
- as expressões de **quantificação temporal**, de tipo DURACAO;
- as expressões de **frequência**, de tipo FREQUENCIA;
- as ET **genéricas**, de tipo GENERICO.

De um modo geral, esta organização clássica das ET conserva, no essencial, as definições do Primeiro HAREM (Cardoso e Santos, 2007)[7], conquanto se tenha procurado, nesta proposta, precisar e definir com maior rigor alguns dos seus aspectos. Em seguida, apresentaremos, de forma sucinta, cada um destes tipos, remetendo o leitor para o texto da proposta (Hagège et al., 2008), que também se encontra reproduzido no anexo B.

2.2.3 TEMPO_CALEND

As entidades de tipo TEMPO_CALEND são expressões que permitem inserir ou localizar o predicado que elas modificam numa linha temporal (como um ponto ou um intervalo). Correspondem aos seguintes subtipos:

- **datas**, sejam elas **absolutas** (fórmulas contendo os três campos ANO-MES-DIA, nas quais até dois campos no máximo podem ser omitidos) ou **referenciais** (ET cuja resolução implica conhecer a data do momento da enunciação, ou conhecer a data de um outro evento que funciona então como referência temporal para a expressão a calcular).
- **horas** (ET com valor de DATA mas com granularidade inferior à unidade *dia*).
- **intervalos** (expressões denotando uma duração no tempo e que têm explicitamente dois limites).

2.2.3.1 Data

As expressões deste subtipo podem representar *datas absolutas* ou *datas referenciais*[8].

Datas absolutas

As ET constituem *datas absolutas* quando contêm a informação necessária para localizar essa data num calendário. Assim, por exemplo, na expressão *em 23 de Outubro de 2007*, a informação está totalmente especificada em relação aos três campos <dia>, <mês> e <ano>; pelo contrário, nas expressões *em 23 de Outubro* e *em 2007*, a informação está apenas parcialmente especificada em relação aos três campos. Apresentam-se de seguida alguns exemplos de ET do tipo TEMPO_CALEND e subtipo DATA:

- Data absoluta completa (campos dia, mês e ano preenchidos):
  Vou viajar <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">***no dia 19 de Outubro de 2007***</EM>.
- Data absoluta incompleta (campos dia e mês não preenchidos)[9]:
  Trabalhei em Londres <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">***em 1998***</EM>.

[7] Relativamente ao Primeiro HAREM, são eliminados os tipos PERIODO, CICLICO, que passam, de um modo geral, a estar integrados em TEMPO_CALEND.

[8] A organização da referência das expressões temporais aqui sucintamente apresentada já é, de resto, bem conhecida. Veja-se, entre outros, Gross (1986) com especial referência a advérbios compostos (ou expressões adverbiais multipalavras), sobretudo de natureza idiomática, e Molinier e Levrier (2000), este último a propósito de advérbios de tempo terminados em *–mente* (*futuramente*, *anteriormente*, *posteriormente*, etc.).

[9] As ET com datas em que apenas os campos <dia>, <mês> ou <dia><mês> estão preenchidos (e.g. *no dia 8*, *em Setembro*, *a 8 de Setembro*) são, em rigor, datas referenciais, cujo valor exacto é relativo ao momento da enunciação. Nesse sentido, será necessário modificar o critério que determina se o valor de TEMPO_REF deve ser ABSOLUTO (ver adiante).

Datas referenciais

Também são consideradas como abrangidas pelo subtipo DATA as expressões que exprimem *datas referenciais*, isto é, para as quais é necessário determinar um ponto de referência para poder localizá-las na linha temporal (e.g. *dois dias mais tarde, na quinta-feira passada, ontem, na próxima terça feira*, etc.). Vejamos, agora, os dois tipos de ET referenciais consideradas: as ET que fazem referência ao momento da enunciação e aquelas que se referem ao tempo de um evento presente no discurso. Um exemplo típico desta distinção pode ser dado através dos exemplos (2.18) e (2.19), respectivamente.

(2.18) O Pedro chegou *ontem*.

(2.19) O barco chegou *no dia anterior*.

Nestes dois exemplos, estamos perante ET que permitem localizar no calendário o evento a que estão associadas, respondendo adequadamente à interrogativa *quando?*. Pode-se, pois, associar a estas expressões o valor SUBTIPO="DATA". Contudo, não se trata aqui de datas absolutas mas sim de expressões referenciais cujo valor tem de ser calculado relativamente a outra referência temporal.

No primeiro exemplo, (2.18), esta referência é o momento da enunciação. Com efeito, se a asserção *O Pedro chegou ontem* for produzida no dia 4/12/2007, pode-se inferir que o evento *chegou* ocorreu no dia 3/12/2007. O tempo em que o evento ocorre, neste exemplo, é função do tempo do momento da enunciação (tempo_enunciação – 1 dia). Fala-se, pois, neste caso, de uma *expressão temporal referencial relativa ao momento da enunciação*.

No segundo exemplo, (2.19), embora também se trate de uma data referencial, a sua referência não é o momento da enunciação, já que a localização temporal de *chegou* é independente do momento em que for produzida a asserção. Neste caso, a referência é outra data/evento que aparece no contexto discursivo. A título ilustrativo, considere-se o exemplo (2.20).

(2.20) O barco só devia chegar ao porto *no dia 25 de Novembro*, no entanto chegou *no dia anterior*.

Como se pode ver, a referência da expressão *no dia anterior* é a data do evento da chegada do barco ao porto, que deveria ter ocorrido no dia 25/11. Conhecendo esta referência pode-se então deduzir que o evento *chegou* ocorreu no dia 24/11. Assim, neste caso está-se em presença de uma *expressão temporal com referência textual*, isto é, uma data relativa a uma outra data explícita no texto.

Esta distinção entre data absoluta, data referencial relativa ao momento de enunciação e data referencial relativa a uma referência textual é formalizada através do atributo TEMPO_REF. No caso de datas absolutas, o valor do atributo TEMPO_REF é ABSOLUTO. No caso de datas referenciais, conforme o tipo da referência o valor do atributo TEMPO_REF é, respectivamente, ENUNCIACAO ou TEXTUAL.

Finalmente, no caso de algumas ET referenciais, é ainda possível acrescentar outra informação complementar com vista à normalização das ET. Trata-se dos atributos SENTIDO e VAL_DELTA. O atributo SENTIDO indica se o seu valor temporal se situa cronologicamente *antes*, *em simultâneo* ou *depois* do *tempo de referência*.

O atributo VAL_DELTA tem por valor uma expressão que indica a distância temporal entre o tempo do evento denotado pela expressão temporal e o momento de referência, seja este o tempo da enunciação ou outro, quando esta distância temporal aparece explicitamente no texto (sobre a normalização destas expressões, ver adiante).

Os exemplos (2.21) a (2.24) ilustram o uso dos atributos TEMPO_REF, SENTIDO e VAL_DELTA e alguns dos seus possíveis valores.

(2.21) O Pedro nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO">**a 3 de Janeiro de 1986**</EM>.

(2.22) O Pedro nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="POSTERIOR" VAL_DELTA="A0M0S0D2H0M0S0">**dois dias depois**</EM>.

(2.23) O Pedro nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="ANTERIOR" VAL_DELTA="A0M0S0D2H0M0S0">**dois dias antes do Natal**</EM>.

(2.24) O Pedro nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ENUNCIACAO" SENTIDO="ANTERIOR">**na sexta-feira passada**</EM>.

2.2.3.2 Hora

Trata-se de ET com valor de DATA mas com granularidade inferior à unidade *dia* (ver exemplo (2.25)).

(2.25) O Pedro está disponível <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="HORA" VAL_NORM="+--------T1500E--LMA">**às 15:00**</EM>.

A existência deste subtipo de datas pode justificar-se pelo facto de constituírem uma classe natural de expressões, que seguem um conjunto de convenções gráficas particulares, facilmente modelizáveis por uma gramática própria, distinta da dos outros tipos de datas. Neste sentido, a proposta apresentada ao Segundo HAREM conservou esta distinção entre data e hora.

2.2.3.3 Intervalo

Corresponde a uma expressão complexa, isto é, composta por duas ET elementares/simples mas que, semanticamente, formam um única EM, e que tem explicitamente dois limites temporais (um limite inicial e um limite final), como ilustram os exemplos (2.26) e (2.27).

(2.26) Trabalhei em Londres <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO">**entre 2000 e 2003**</EM>.

(2.27) Trabalhei em Londres <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO">**de Outubro a Dezembro de 2007**</EM>.

Note-se que, nesta avaliação conjunta, não se levou em consideração a granularidade das expressões de tempo que constituem os limites explícitos do intervalo. Assim, por exemplo, integram este tipo de ET formas com granularidade inferior à unidade *dia*, tal como em (2.28).

(2.28) O escritório fecha para almoço <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO">**das 12:00 às 14:00 horas**</EM>.

Por outro lado, incluímos ainda no tipo INTERVALO não só expressões complexas com datas, como as dos exemplos acima, mas combinações que exprimem outros valores temporais como, por exemplo, a duração, em (2.29).

(2.29) Vai demorar <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO">**de 3 a 6 meses**</EM>.

As ET do tipo INTERVALO não foram normalizadas nesta avaliação conjunta, dada a complexidade de que se revestem algumas das suas formas, nomeadamente as que combinam ET dos tipos DATA e HORA. Veja-se o exemplo (2.30).

(2.30) O Pedro esteve a fazer isso isso *desde a meia-noite de 5 de Dezembro de 2007 até ao dia de Natal, ao meio-dia*.

Nesse sentido, será de esperar que algumas destas questões venham a ser resolvidas pelas propostas que apresentamos no fim deste capítulo. Tal permitiria igualmente dar também alguns passos no sentido da normalização das ET do tipo INTERVALO.

Além das expressões TEMPO_CALEND, consideraram-se ainda dentro da categoria TEMPO as expressões de **duração** e de **frequência**, de que trataremos já a seguir.

2.2.3.4 Duração

Corresponde a uma expressão TEMPO que se refere a uma duração de tempo contínuo. Ao contrário das datas, trata-se de expressões que não exprimem propriamente a localização (ou calendarização) de um evento, mas sim uma *quantificação temporal*, sendo constituídas por nomes de unidades de medida de tempo e determinantes com função de quantificadores (numerais, por exemplo). Podem, por vezes, ser introduzidas, facultativamente, pela preposição *durante* (encontrando-se também outras preposições) e respondem adequadamente à interrogativa *(prep) quanto tempo?*. Ver exemplos (2.31) a (2.35).

(2.31) Fiquei <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">**dois meses**</EM> em Lisboa.

(2.32) O urso fica <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">**todo o inverno**</EM> na toca.

(2.33) O Pedro trabalhou <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">**várias semanas**</EM> no restaurante.

(2.34) O Pedro trabalhou <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">**durante três anos**</EM> na tese.

(2.35) A aplicação da lei será suspensa <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">**por dez anos**</EM>.

2.2.3.5 Frequência

O tipo FREQUENCIA corresponde a expressões TEMPO que exprimem uma repetição de um evento no tempo. Estas expressões respondem adequadamente às interrogativas do tipo *com que frequência?*, como ilustram os exemplos (2.36) a (2.40).

(2.36) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">**amiúde** </EM>.

(2.37) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">**diariamente** </EM>.

(2.38) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">**todos os dias**</EM>.

(2.39) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">**duas vezes por semana**</EM>.

(2.40) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">**dia sim dia não**</EM>.

Como se pode ver pelos exemplos acima, as ET deste tipo podem ser advérbios simples, derivados de adjectivos (*diariamente*) ou não (*amiúde*), locuções adverbiais mais ou menos cristalizadas (*dia sim dia não*), certas expressões com forma de sintagma nominal (*todos os dias*) e outras construções em torno de nomes como *vez* (*duas vezes por semana*). Incluem-se ainda neste tipo de ET certos advérbios que têm sobretudo um valor aspectual (*frequentemente*, *pontualmente*, *ocasionalmente*, *raramente*). Contudo, a definição deste tipo de ET é ainda insuficiente para dar conta de expressões cujo significado global parece combinar o valor de frequência com o de localização temporal, como acontece em (2.41).

(2.41) A reunião de pais tem lugar *todas as primeiras segundas-feiras de cada mês*.

2.2.3.6 ET genéricas

Trata-se de expressões TEMPO que não se referem a uma data específica embora a expressão linguística integre elementos lexicais que denotam um valor temporal, como nos exemplos (2.42) e (2.43).

(2.42) Adoro <EM ID="..." CATEG="TEMPO" TIPO="GENERICO">**o Verão**</EM>.

(2.43) <EM ID="..." CATEG="TEMPO" TIPO="GENERICO">**Fevereiro**</EM>é o mês mais curto do ano.

Estas expressões genéricas podem, como se sabe, ter um papel relevante no cálculo de referências temporais, pelo que importa identificá-las adequadamente. Por ora, contudo, elas não são normalizadas.

2.3 Normalização

A normalização das datas absolutas e horas, como ilustrado no exemplo (2.44), obedece ao seguinte formato:

<Era><Ano><Mes><Dia>**T**<Hora><Minuto>**E**<ESTACAO>**LM**<limite_aberto>

Onde:

- <Era> corresponde a um caracter que indica se a data é depois ou antes da nossa era;
- <Ano> corresponde a quatro algarismos que representam o valor do ano;
- <Mes> corresponde a dois algarismos que representam o valor do mês;
- <Dia> corresponde a dois algarismos que representam o valor do dia;
- <Hora> corresponde a dois algarismos que representam o valor da hora;
- <Minuto> corresponde a dois algarismos que representam o valor dos minutos;
- <ESTACAO> corresponde a duas letras maiúsculas referentes às estações do ano;
- <limite_aberto> indica se a expressão normalizada de data absoluta introduz um intervalo de tempo com limite anterior ou limite posterior não determinado (em aberto). Os valores respectivos são "A", no caso de limite *anterior* em aberto; ou "P", no caso de limite *posterior* em aberto.

Exemplo:

(2.44) Nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+19860103T----E--LM-">**a 3 de Janeiro de 1986**</EM>.

2.3.1 Normalização de datas referenciais

Como já se disse atrás, algumas ET referenciais recebem uma outra informação complementar com vista à sua normalização. Trata-se dos atributos SENTIDO e VAL_DELTA. O atributo SENTIDO indica se o seu valor temporal se situa cronologicamente *antes*, *em simultâneo* ou *depois* do *tempo de referência*. Os possíveis valores do atributo SENTIDO são, pois:

ANTERIOR, POSTERIOR, SIMULT, ANTERIOR_OU_SIMULT, POSTERIOR_OU_SIMULT.

O atributo VAL_DELTA corresponde ao valor temporal que se deve incrementar ou subtrair a partir do tempo de referência para obter o valor temporal do evento associado à expressão temporal a anotar, quando esta distância temporal aparece explicitamente no texto. No caso de esta distância temporal não estar explícita, o valor de VAL_DELTA é omitido. Tal como ilustrado em (2.45), os valores possíveis de VAL_DELTA são representados da maneira seguinte:

A<digitos>**M**<digitos>**S**<digitos>**D**<digitos>**H**<digitos>**M**<digitos>**S**<digitos>

Onde:

- as letras **A**, **M**, **S**, **D**, **H**, **M**, **S** são constantes que devem aparecer nesta ordem e marcam, respectivamente, a posição dos valores de anos, meses, semanas, dias, horas, minutos e segundos.
- os <digitos> à direita das letras constantes correspondem ao número de anos, meses, semanas, dias, horas, minutos e segundos que se devem adicionar ou diminuir à data de referência para obter o valor temporal da expressão anotada.

Exemplo:

(2.45) Apareceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="POSTERIOR" VAL_DELTA="A0M0S2D0H0M0S0">**duas semanas**</EM> depois da festa.

2.3.2 Normalização da DURACAO

Para expressões de tipo DURACAO, a normalização exprime uma distância temporal representada com o seguinte formato:

A<digitos>**M**<digitos>**S**<digitos>**D**<digitos>**H**<digitos>**M**<digitos>**S**<digitos>

Onde:

- as letras **A**, **M**, **S**, **D**, **H**, **M**, **S** são constantes que devem aparecer nesta ordem e marcam, respectivamente, a posição dos valores de anos, meses, semanas, dias, horas, minutos e segundos;
- os <digitos> à direita das letras constantes correspondem ao número de anos, meses, semanas, dias, horas, minutos e segundos que se devem adicionar ou diminuir à data de referência para obter o valor temporal da expressão anotada.

Exemplo:

(2.46) Fiquei <EM ID="..." CATEG="TEMPO" TIPO="DURACAO" VAL_NORM="A0M2S0D0H0M0S0">**dois meses**</EM> em Lisboa.

Para terminar esta secção, uma breve nota apenas para indicar que a proposta de normalização das ET ainda *não* contemplou, neste momento, as expressões do tipo FREQUENCIA nem o subtipo INTERVALO do tipo TEMPO_CALEND. Estes dois aspectos deverão ser aprofundados em futuras edições do HAREM. Por um lado, é possível normalizar, pelo menos parcialmente, alguma da informação veiculada pelas expressões de FREQUENCIA, indicando, nomeadamente, entre outros valores, a granularidade do intervalo entre instâncias do evento modificado e o número de repetições desse evento. Por outro lado, no caso dos intervalos, é possível normalizar cada um dos limites temporais.

2.4 A experiência do Segundo HAREM

Com uma primeira versão, nos seus traços gerais, já bastante próxima da versão final, que ficou disponível logo a 18 de Dezembro de 2007, a elaboração, discussão e redacção final da proposta foi um processo longo e complexo que culminou no documento ora disponível no sítio da avaliação conjunta do Segundo HAREM (13 de Abril de 2008). Produziu-se nessa altura (14 de Abril) uma versão dos primeiros 10% da CD do Mini-HAREM anotada segundo as directivas do TEMPO, que foi distribuída aos participantes para treino e discussão[10].

Participaram na pista do TEMPO sete dos dez participantes no HAREM, embora se verifiquem diferenças relativamente à forma como cada um se apresentou[11]:

- seis sistemas com `TIPO`;
- cinco sistemas com `SUBTIPO`;
- dois sistemas com `TEMPO_REF` (tipo de referência para datas referenciais);
- um sistema com a normalização.

Como primeira conclusão a tirar, inevitavelmente, deste perfil de participação, recomenda-se prudência e moderação no desenvolvimento da tarefa para futuras avaliações conjuntas de TEMPO, o que não impede, naturalmente, que se introduzam melhoramentos ou mesmo correcções.

Do ponto de vista dos resultados[12], e de acordo com o modo de avaliação em que todos os sistemas participaram (TEMPO clássico), é possível fazer algumas observações gerais: na tarefa de classificação (cf. figura 2.1), verifica-se que apenas dois sistemas apresentam resultados de precisão consistententemente acima de 0,7 (máximo 0,767); em termos de abrangência, apenas um sistema apresenta valores acima de 0,7 (máx. 0,758), embora duas das respectivas corridas apresentem valores cerca de dez por cento inferiores; já o segundo melhor sistema em abrangência, embora com resultados consistentes, só consegue valores pouco superiores a 0,5 (entre 0,533 e 0,489); finalmente, os mesmos dois sistemas apresentam resultados consistentes em termos de medida F: o primeiro, com valores superiores a 0,7 (máx. 0,748) e o segundo na casa dos 0,6 (máx. 0,618).

Na tarefa de identificação, e como se pode verificar pela figura 2.2, os melhores sistemas apresentam resultados relativos em grande medida semelhantes aos acima relatados (verificam-se os máximos de 0,769 de precisão, 0,758 de abrangência e 0,747 de medida F).

2.5 Próximos passos e perspectivas futuras

Nesta secção, apresentamos os aspectos que, na sequência da experiência do Segundo HAREM, nos parece relevante tratar, em termos de perspectivas de investigação e desen-

[10] Este fragmento anotado faz parte da LÂMPADA - Pacote de Recursos do Segundo HAREM (http://www.linguateca.pt/HAREM/PacoteRecursosSegundoHAREM.zip).

[11] Remetemos o leitor para os capítulos 1 e 3, para uma descrição mais pormenorizada dos cenários de participação dos sistemas e modos de avaliação, bem como para o capítulo 5, que inclui a descrição da avaliação da pista do TEMPO.

[12] Os valores aqui apresentados correspondem aos disponíveis em http://www.linguateca.pt/HAREM, ver Resultados do Segundo HAREM, e são arredondados à terceira casa decimal.

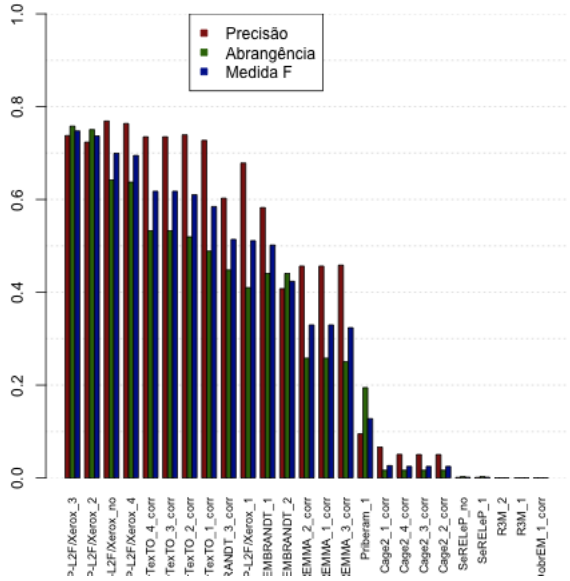


Figura 2.1: Resultados do HAREM clássico no cenário selectivo TEMPO na CD do TEMPO, tarefa de classificação.

volvimento futuros. Duas preocupações norteiam estas sugestões que assim submetemos à apreciação da comunidade de PLN do português:

Em primeiro lugar, corrigir ou melhorar alguns aspectos da proposta actual da avaliação conjunta do Segundo HAREM. Trata-se de observações que fomos recolhendo ao longo do trabalho desenvolvido, bem como várias sugestões recebidas tanto de outros participantes como da parte da organização.

Em segundo lugar, garantir uma continuidade, tanto quanto possível suave, entre as sucessivas edições das avaliações conjuntas de sistemas de REM/TEMPO, por forma a garantir a novos actores uma mais fácil integração neste processo, estabilizando os standards e potenciando os recursos e ferramentas entretanto construídos. Não esquecemos que, nesta edição do Segundo HAREM, parte dos sistemas participantes (ainda?) não integrou todas as dimensões da nossa proposta, nomeadamente aquele que era o seu principal desafio: o de ir além da tarefa de REM e tratar também a normalização das ET. Seria, no mínimo, inadequado fazer evoluir a proposta sem um consenso e participação alargados da comunidade[13]. Neste sentido, as linhas que se seguem podem ser interpretadas como um mapa do caminho para uma futura edição HAREM/TEMPO.

[13] Neste sentido, a equipa L2F/Xerox veria com naturalidade que uma nova avaliação conjunta, Terceiro HAREM, se realizada num prazo relativamente curto, se limitasse para já a repetir a experiência do Segundo HAREM.

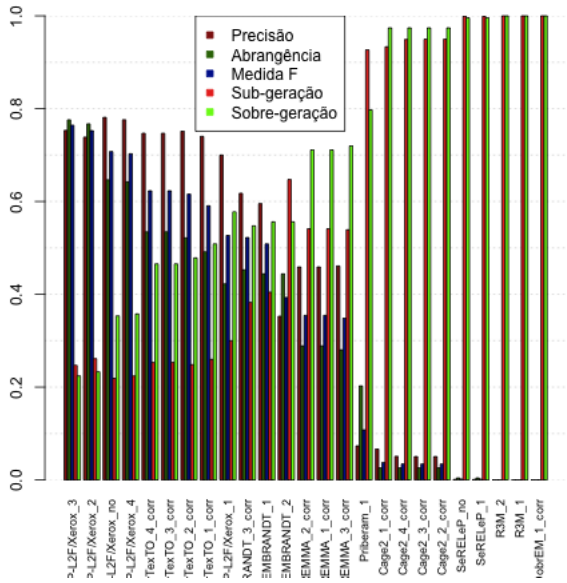


Figura 2.2: Resultados do HAREM clássico no cenário selectivo TEMPO na CD do TEMPO, tarefa de identificação.

2.5.1 TEMPO_CALEND

2.5.2 Novo subtipo=DATA

Propõe-se agregar no tipo `DATA` os actuais subtipos `DATA` e `HORA`. São várias as motivações para esta evolução: em primeiro lugar, semanticamente, ambos os tipos correspondem à localização dos eventos numa linha do tempo, a única diferença entre eles é a granularidade da unidade temporal; a normalização é basicamente a mesma: há campos comuns em cada um dos subtipos e uma representação única irá simplificar a normalização das ET do tipo `INTERVALO` quando os respectivos limites são expressos simultaneamente com datas e horas.

Note-se que uma das motivações principais para conservação da distinção dos subtipos `DATA` e `HORA` prendia-se com as gramáticas (ou regras) usadas para a sua identificação. Na medida em que se pretende orientar a actual proposta no sentido de evoluir para lá da tarefa de REM e passar a incluir também a normalização, não só essa motivação perde alguma da sua razão de ser como se ganha em obter uma normalização uniforme.

Esta alteração implica a (relativamente ligeira) reformulação dos critérios de atomização das ET. Assim, no quadro da actual proposta, considerou-se que nos casos (2.47) e (2.48) se estava em presença de várias ET.

(2.47) Isto aconteceu <EM ... SUBTIPO="DATA">**na sexta-feira**</EM>, <EM ... SUBTIPO="DATA">**23 de Abril de 2008**</EM>, <EM ... SUBTIPO="HORA">**pelas 18:30**</EM>

(2.48) Isto aconteceu <EM ... SUBTIPO="DATA">**na sexta-feira**</EM>, <EM ... SUBTIPO="DATA">**23 de Abril de 2008**</EM>, <EM ... SUBTIPO="DATA">**dia de São Jorge**</EM>, <EM ... SUBTIPO="HORA">**pelas 18:30**</EM>

Nestes exemplos, cada ET é, de acordo com os actuais critérios de atomização, identificada e normalizada separadamente. Contudo, esta forma de representação não é inteiramente adequada, pois trata-se de sequências de ET numa cadeia de aposição, em que cada nova ET precisa ou desenvolve as referências temporais das ET anteriores, pelo que deveriam constituir *uma única referência temporal*. Por outro lado, a estrutura de aposição permite resolver imediatamente alguns dos valores referenciais não absolutos: por exemplo, enquanto a ET *na sexta feira* teria, à partida, um valor refencial relativo ao momento de enunciação, quando integrado nesta sequência apositiva ela é mera informação complementar, dispensando o cálculo da referência temporal, na medida em que se subordina ao valor referencial absoluto da ET de data adjacente, e.g., *23 de Abril de 2008*. Além disso, certas dificuldades de classificação levantadas por ET como *dia de São Jorge*, que poderiam ser incorrectamente classificadas no tipo `GENERICO` podem ser evitadas, já que também esta ET é mera informação adicional à data absoluta adjacente.

A aceitar-se estes argumentos, o critério geral para separar/juntar ET deverá ser alterado de modo a permitir tratar instâncias de `DATA` e `HORA` em aposição como uma única ET, desde que a sua normalização seja complementar:

(2.49) Isto aconteceu <EM ... SUBTIPO="*DATA">**na sexta-feira, 23 de Abril de 2008, pelas 18:30**</EM>

(2.50) Isto aconteceu <EM ... SUBTIPO="*DATA">**na sexta-feira, 23 de Abril de 2008, dia de São Jorge, pelas 18:30**</EM>

em que `*DATA` corresponde ao novo tipo unificado.

2.5.2.1 subtipo=INTERVALO

Propõe-se a normalização das ET do subtipo `INTERVALO`, tais como as apresentadas nos exemplos (2.51) a (2.54), que neste momento não são normalizadas. Para este tipo de situações, como se vê nos exemplos, a normalização poderia ser feita duplicando os pares atributo-valor e dando índices numéricos a cada um dos limites temporais explícitos do `INTERVALO`.

(2.51) O Pedro está de férias <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO" TEMPO_REF1="ABSOLUTO" VAL_NORM1="+----0423T----E--LM" TEMPO_REF2="ABSOLUTO" VAL_NORM2="+----0529T----E--LM">**de 23 de Abril a 29 de Maio**</EM>.

(2.52) O Pedro está de férias <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO" TEMPO_REF1="ABSOLUTO" VAL_NORM1="+ 20090423T - - - - E - - LM" TEMPO_REF2="ABSOLUTO" VAL_NORM2="+ 20090529T - - - - E - - LM">**entre 23 de Abril e 29 de Maio de 2009**</EM>.

(2.53) O Pedro está de férias <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO" TEMPO_REF1="ENUNCIACAO" SENTIDO1="SIMULT" VAL_DELTA1="A0M0S0D0H0M0S0"

TEMPO_REF2="ENUNCIACAO" SENTIDO2="POSTERIOR" VAL_DELTA2="A0M0S1D0H0M0S0">**desde hoje até à próxima semana**</EM>.

(2.54) O Pedro está de baixa <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO" TEMPO_REF1="ENUNCIACAO" SENTIDO1="SIMULT" VAL_DELTA1="A0M0S0D0H0M0S0" TEMPO_REF2="ENUNCIACAO" SENTIDO2="POSTERIOR" VAL_DELTA2="A0M0S0D2H0M0S0">**entre hoje e depois de amanhã**</EM>.

2.5.2.2 Novo subtipo=COMPLEXO

Sugere-se a eventual criação de um novo subtipo `COMPLEXO` dentro do tipo `TEMPO_CALEND`, que deverá capturar ET que incluem os conceitos de `DATA` e `VAL_DELTA`:

(2.55) *Faz (hoje, no dia 21 de Dezembro) quinze dias* que isso aconteceu.

(2.56) Isso acontecerá *de (hoje, ontem) a quinze dias*.

2.5.3 DURACAO

Propõe-se que a normalização das ET do tipo `DURACAO` passe a incluir uma unidade menor que o segundo (milissegundos), a fim de permitir o tratamento adequado de, por exemplo, resultados desportivos.

2.5.3.1 tipo=DURACAO subtipo=INTERVALO

O subtipo `INTERVALO` é, na actual proposta, um tipo híbrido pois não integra apenas ET que exprimem uma localização temporal (`TIPO="TEMPO_CALEND"`), desde que apresentem dois limites temporais explícitos, como também abrange expressões de tempo que denotam outras formas de modicação temporal, nomeadamente expressões de `DURACAO`:

(2.57) Isso durou *entre 2 e 3 horas*.

Tal solução não é, pois, inteiramente adequada. Uma solução possível seria que o tipo `DURACAO` passasse a incluir o subtipo `INTERVALO`, por forma a dar conta de situações como as ilustradas no exemplo acima. A normalização deste tipo de intervalos far-se-ia de modo análogo ao dos intervalos com datas (ver acima), através da duplicação de `VAL_NORM` e atribuição de índices aos pares atributo valor:

(2.58) Isso durou <EM ID="..." CATEG="TEMPO" TIPO="DURACAO" SUBTIPO="INTERVALO" VAL_NORM1="A0M0S0D0H2M0S0" VAL_NORM2="A0M0S0D0H3M0S0">**entre 2 e 3 horas**</EM>.

2.5.4 FREQUENCIA

Propõe-se passar a normalizar de forma explícita um determinado conjunto de ET do tipo `FREQUENCIA`. Tomamos como modelo deste tipo de ET expressões como a do exemplo seguinte:

(2.59) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">**duas vezes por semana**</EM>.

Para normalização da FREQUENCIA propõe-se usar dois atributos suplementares de EM[14]:

- VAL_QUANT, que indica o número de vezes em que o evento/processo se repete; e
- VAL_MODULO, que representa a *granularidade* dessa frequência.

O primeiro atributo seria preenchido por valores numéricos e o segundo por uma notação semelhante à já usada na normalização da DURACAO:

A<digitos>M<digitos>S<digitos>D<digitos>H<digitos>M<digitos>S<digitos>

Deste modo, a expressão acima ilustrada seria normalizada como em (2.60). Da mesma forma, as ET do tipo FREQUENCIA ilustradas nos exemplos (2.61) a (2.65), passariam a ser normalizadas de acordo com este formato.

(2.60) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_QUANT="2" VAL_MODULO="A0M0S1D0H0M0S0">**duas vezes por semana**</EM>.

(2.61) O Pedro faz isso <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_QUANT="1" VAL_MODULO="A0M0S1D0H0M0S0">**semanalmente**</EM>.

(2.62) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_QUANT="1" VAL_MODULO="A0M0S0D1H0M0S0">**diariamente**</EM>.

(2.63) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_NORM="1">**todos os dias**</EM>A0M0S0D1H0M0S0.

(2.64) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_NORM="1">**dia sim dia não**</EM>A0M0S0D2H0M0S0.

(2.65) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_NORM="1">**todas as semanas**</EM>A0M1S0D0H0M0S0.

Para expressões complexas (agregados temporais) como:

na primeira quinta-feira de cada mês, *quatro domingos seguidos*, *dez dias interpolados*

que incluem tanto o conceito de DATA como de FREQUENCIA, ou para expressões em que nomes como *vez(es)* aparecem determinados por um quantificador indefinido:

(*várias*, *muitas*, *algumas*, *umas poucas*, *poucas*, *bastantes*, *imensas*) *vezes por semana*

ou ainda para expressões em que não é possível determinar com rigor esse quantificador, como sucede na ET *todas as semanas*, sugere-se que só o campo MODULO seja normalizado, como se ilustra nos exemplos (2.66) e (2.67).

(2.66) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_QUANT="not_defined" VAL_MODULO="A0M1S0D0H0M0S0">**na primeira quinta-feira de cada mês**</EM>.

[14] Esta proposta é fortemente inspirada na TimeML (Boguraev et al., 2005).

(2.67) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_NORM="not_defined">**algumas vezes por semana** </EM>A0M1S0D0H0M0S0.

Naturalmente, continuariam por normalizar expressões que veiculam valores vagos ou imprecisos, sobretudo os que são expressos por certos adverbiais como *amiúde*, *frequentemente*, *ocasionalmente*, etc.:

(2.68) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">**amiúde** </EM>.

2.5.5 Outras sugestões

Além das sugestões acima apresentadas, julgamos que seria oportuno e não demasiado complexo introduzir alguns pequenos melhoramentos na normalização das ET.

2.5.5.1 Not_Norm

Propõe-se a inclusão de uma propriedade que explicite a distinção entre, por um lado, as expressões que, por qualquer razão, não foram normalizadas pelo sistema, das que se definiu como não devendo ser normalizadas de todo. Assim, por exemplo, apenas o advérbio de `FREQUENCIA` *frequentemente* deveria receber este traço, ao contrário de *semanalmente*, que deveria ser normalizado pelos diferentes sistemas.

2.5.5.2 Indefinição (ou vagueza)

Propõe-se a inclusão de uma propriedade que explicite a existência de vagueza em algumas categorias, como acontece com as ET dos exemplos (2.69) a (2.71).

(2.69) O Pedro fez isso *por volta do dia 23 de Abril de 2008*.

(2.70) O Pedro fez isso *perto das 3 da tarde*.

(2.71) O Pedro fez isso *em pouco tempo*.

Em futuras avaliações conjuntas é necessário estender este conceito para tornar mais clara a granularidade da imprecisão temporal da ET.

2.6 Conclusões

Apresentámos neste capítulo a proposta de tarefa de reconhecimento, classificação e normalização de expressões temporais para a segunda avaliação conjunta de sistemas de reconhecimento de entidades mencionadas – o Segundo HAREM. Trata-se de uma proposta de algum modo conservadora na medida em que preserva, embora procure definir com maior precisão, grande parte da estrutura de classificação de ET do Primeiro HAREM.

Ao mesmo tempo, a proposta introduz diversos aspectos inovadores, sobretudo no que diz respeito à delimitação das ET e a normalização das ET, esta última tendo em vista o cálculo de referências temporais. Tratou-se de dar um primeiro passo no sentido de associar as ET aos eventos e estados de coisas que elas modificam, a fim de os ordenar parcialmente, numa sequência cronológica. Contudo, procurámos intencionalmente garantir que

estes aspectos inovadores mantivessem um certo grau de simplicidade, evitando uma excessiva (porque demasiado súbita) descontinuidade com a tarefa do Primeiro HAREM e permitindo uma participação o mais abrangente possível da comunidade do PLN.

Procurámos, além disso, reflectir, ainda que de forma breve, sobre a experiência deste Segundo HAREM. Com base no perfil de participação dos vários sistemas em jogo, parece-nos necessário adoptar prudência e moderação no desenvolvimento da tarefa para futuras avaliações conjuntas de TEMPO, o que não impede, naturalmente, que se introduzam melhoramentos ou mesmo correcções. Do ponto de vista dos resultados, é possível considerar que, de um modo geral, a fasquia do estado da arte, para a classificação de ET, se situa em valores na ordem dos 0,75 para a precisão, abrangência e medida F. Contudo, o conjunto dos sistemas participantes apresenta ainda grandes disparidades nos resultados obtidos, quer entre si, quer entre as diferentes medidas.

Como resultado da experiência deste Segundo HAREM, apresentámos, finalmente, um conjunto de propostas que procuram corrigir ou melhorar aspectos da classificação e normalização das ET, na perspectiva de uma nova avaliação conjunta de entidades mencionadas. Como nota final, referimos que nestas propostas se deixa para um outro ciclo de avaliação o cálculo da referência temporal: não porque não se julgue esta tarefa importante – lembramos ser este o objectivo que pretendemos alcançar com a proposta de normalização das ET –, mas porque consideramos ser necessário e mais proveitoso adoptar uma estratégia de progressão em pequenos (mas firmes) passos, a fim de que se possa manter um grupo de investigadores interessados e activos nesta linha de avaliação.

Capítulo 3

É tempo de avaliar o TEMPO

Cristina Mota, Paula Carvalho, Cláudia Freitas, Hugo Gonçalo Oliveira e Diana Santos

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 3, p. 55–75.

No capítulo 2 foi apresentada a pista de identificação, classificação e normalização de expressões temporais adoptada no Segundo HAREM, que designaremos daqui em diante como pista do TEMPO, e cujas directivas (Hagège et al., 2008) se encontram republicadas no apêndice B. Uma vez que os autores da proposta referiram desde o início a sua intenção de serem participantes nesta avaliação conjunta, a tarefa de anotação da colecção dourada e de avaliação dos sistemas ficou inteiramente a cargo da organização do Segundo HAREM. O objectivo deste capítulo é então documentar essas duas actividades, dando também conta dos resultados obtidos.

Começamos por discutir as principais questões que tivemos de resolver durante o processo de anotação das expressões temporais na colecção dourada e que se resumem a dois tipos de situações: casos que não nos pareceram suficientementes claros nas directivas e casos que, por verificarem mais do que um critério, tiveram de ser decididos por nós. Embora alguns desses casos tenham sido posteriormente esclarecidos pelos autores da proposta, mantivemos a sua documentação aqui dado que os participantes e outros leitores podem ter as mesmas dúvidas. Além disso, apresentamos critérios adicionais que estabelecemos para situações não previstas nas directivas. Em vários casos, aproveitamos para exprimir o nosso ponto de vista discordante, na perspectiva de enriquecer genuinamente o estudo e processamento temporal do português.

Na próxima secção, fornecemos igualmente dados quantitativos sobre as entidades anotadas com a categoria `TEMPO`. Em seguida, apresentamos sucintamente os modos de avaliação desta pista (o leitor poderá encontrar informação mais detalhada sobre a avaliação no capítulo 5) e os resultados de desempenho obtidos pelos sistemas. Finalmente, tecemos algumas sugestões sobre novas versões de uma nova pista do TEMPO, quer fazendo uma autocrítica sobre a forma de avaliação como propondo alguns trabalhos futuros.

3.1 Anotação da colecção dourada

O processo de anotação das expressões temporais na colecção dourada (CD) do Segundo HAREM decorreu em duas fases. Numa primeira fase, todas as expressões que verificavam os critérios estabelecidos nas directivas do TEMPO foram anotadas com os atributos do HAREM clássico, ou seja, `CATEG`, `TIPO` e `SUBTIPO`. Essa anotação decorreu em simultâneo com a anotação das entidades pertencentes às restantes categorias previstas nas directivas do HAREM clássico, tal como descrito no capítulo 1.[1]

Numa fase posterior, foi seleccionado um subconjunto de documentos, mais precisamente trinta (cf. secção 3.2 para uma descrição detalhada), nos quais se adicionou, às entidades classificadas como `TEMPO`, os atributos `SENTIDO`, `TEMPO_REF`, `VAL_DELTA` e `VAL_NORM`, específicos das directivas do TEMPO, e que designaremos como atributos estendidos. Esta colecção de documentos foi baptizada CD do TEMPO.

Tendo em conta que os autores das directivas foram também participantes, o processo de anotação não pôde usufruir da colaboração directa de pelo menos um dos proponentes da pista do TEMPO. Assim, mesmo tendo como material de apoio seis documentos ano-

[1] O HAREM clássico tem, contudo, uma questão que não é clássica no sentido de que o critério de delimitação de entidades temporais, proposto nas directivas do TEMPO, é muito diferente do das outras categorias – como será mais amplamente discutido no capítulo 6.

tados pelos autores das directivas[2], um dos desafios da anotação feita por nós esteve em esclarecer as diversas dúvidas que foram surgindo, mas que não podiam ser discutidas (explicitadas) de forma directa, visto que não podíamos revelar o conteúdo da CD nem permitir a localização dos documentos correspondentes à colecção dourada na colecção HAREM. Tivemos sempre essa precaução, para não criar uma situação de desigualdade em relação aos restantes participantes.

Na maioria dos casos, as dúvidas com que nos deparámos prenderam-se com o facto de as entidades em análise parecerem não se enquadrar perfeitamente nos critérios previstos nas directivas de anotação do TEMPO, ou então poderem encaixar-se em mais do que um critério quanto à sua classificação. Naturalmente, o facto de ser a primeira vez que as directivas estavam a ser seguidas num processo de anotação contribuiu para que alguns pontos não estivessem ainda suficientemente explicitados.

Contudo, e como já referido, muitas vezes considerámos que as opções tomadas pela proposta do TEMPO não foram as melhores, e vozeamos a nossa crítica no presente capítulo. É muito importante contudo salientar que essa crítica tem como objectivo o futuro, e que durante a anotação tentámos sempre seguir da forma mais próxima possível as directivas acordadas. Ou seja, nunca tentámos anotar as expressões temporais de acordo com a nossa opinião quando esta divergia, mas simplesmente obedecer às directivas.

3.1.1 Opções relativas aos atributos do HAREM clássico de TEMPO

3.1.1.1 Delimitação da entidade quando a expressão temporal verifica os critérios 1 e 2-6

Quando uma expressão temporal verifica o subcritério 2-6, ou seja, quando

> um sintagma preposicional cujo núcleo seja uma das palavras *altura*, *tempo*, *momento*, *período*, *era*, etc., quando
>
> - estas palavras forem determinadas por um demonstrativo (por exemplo: *nesse tempo*),
> - ou especificados por uma relativa (por exemplo: ***na altura*** *em que ela adoeceu*),
> - um possessivo (por exemplo: *durante a nossa era*)
> - ou modificado por outro sintagma preposicional introduzido por *de* (por exemplo: *durante a era dos dinossauros*)
> - ou então por um adjectivo capitalizado (por exemplo: *durante o período Barroco, Cretáceo*, etc.);
>
> (Hagège et al., 2008, reformatação nossa)

estamos perante uma expressão temporal que constitui toda ela (núcleo e respectivos modificadores) uma unidade sintáctica que responde adequadamente às interrogações *<prep> quando?* ou *quando?*.

[2] Estes documentos correspondem a 10% da CD do Mini-HAREM e foram disponibilizados aos participantes como material de treino. A anotação destes documentos foi copiosamente discutida e mutuamente esclarecida entre a organização do HAREM e o grupo do TEMPO em Novembro-Dezembro de 2007, o que permitiu tanto o refinamento das directivas como uma maior clarificação, embora contudo e como descrevemos no presente capítulo, ainda não totalmente suficiente para efectuarmos a anotação sem dúvidas.

Embora da nossa interpretação das directivas, para este caso específico, não fosse claro se se deveria incluir ou não os modificadores, e, de um ponto de vista semântico, a sua não inclusão iria forçar-nos a marcar expressões sem sentido, foi-nos indicado pelos proponentes que os modificadores preposicionais e oracionais não deveriam, de facto, ser tidos em consideração na etiquetagem das expressões temporais, excepto no caso dos modificadores adjectivais em maiúscula, com a seguinte justificação:

> se assim não fosse, isso implicaria um processamento sintáctico complexo (que se prende, nomeadamente, com a identificação dos limites temporais das orações subordinadas), desviando-se da proposta de REM,

tal como discutido no capítulo 2.

Assim, nos exemplos (3.1) e (3.2) apenas anotámos o núcleo da expressão temporal, em vez de alargar a anotação também às expressões que representamos em itálico, o que no seu todo nos parecia ser de facto as expressões temporais; no exemplo (3.3), pelo contrário, a expressão temporal inclui então o adjectivo.

(3.1) Muitos milhões acabam, como <EM ID="cver-8" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**no tempo**</EM> *de Arafat*, em contas secretas

(3.2) acompanhando «o percurso da decadência, da perdição», do autarca alentejano <EM ID="hub-57257-6" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**até ao momento**</EM> *em que aceita dinheiro em troca de um favor*

(3.3) refrões fortes que traduziam o sentimento da juventude <EM ID="hub-77558-120" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**da era Cavaquista**</EM>

3.1.1.2 Delimitação da entidade quando a expressão temporal é constituída por DATA e HORA

Quando uma expressão é internamente composta por dois constituintes, um do tipo `DATA` e outro do tipo `HORA`, identificámos cada um desses constituintes como EM independentes, mesmo que um deles não se combine isoladamente com o predicado que modifica, como ilustram os exemplos (3.4) e (3.5).

(3.4) O provável primeiro bebé português do ano é do sexo masculino e nasceu <EM ID="hub-71248-191" CATEG="TEMPO" TIPO="TEMPO_CALEND">**aos 30 segundos**</EM> <EM ID="hub-71248-192" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**de hoje**</EM> na Maternidade Alfredo da Costa

(3.5) <EM ID="aa58069-369" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="HORA">**Às 17h20**</EM> <EM ID="aa58069-370" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**de ontem**</EM> em Lisboa, a vaga especulativa do mercado do petróleo convergiu com as previsões dos últimos meses

Em casos como estes, embora não se verifique um dos critérios para segmentação de uma expressão complexa em duas entidades[3], pareceu-nos que este procedimento se justi-

[3] O critério que não se verifica é o das expressões componentes serem ambas sintacticamente válidas quando combinadas com o evento que modificam. No primeiro caso não podemos ter "nasceu *de hoje*" (e parece-nos pouco aceitável ter "nasceu *aos 30 segundos*"), e no segundo não podemos ter "*de hoje* em Lisboa, ... convergiu".

ficava porque da nossa interpretação das directivas não era claro se o subtipo DATA também englobava expressões complexas que incluissem hora. Assim, tomando esta opção poderia dar-se mais valorização aos sistemas, nomeadamente se tivessem tido dificuldade em amalgamar estas duas entidades numa.

Naturalmente, poderíamos ter usado a notação dos ALT para produzir as duas segmentações na colecção dourada. Não o fizemos, principalmente, porque isso não estava previsto na proposta do TEMPO. Estamos agora convencidos que essa solução teria sido muito melhor, em vez de uma decisão arbitrária.

Ao contrário do proposto no capítulo 2 como futura melhoria, de existir apenas um subtipo DATA* que permitiria dar conta deste tipo de expressões complexas, defenderíamos uma notação que permitisse o encaixe das entidades HORA nas entidades TEMPO. Embora concordemos que semanticamente uma HORA pode ser uma DATA, estas expressões têm de facto uma sintaxe muito diferente, e julgamos que seria mais adequado manter a distinção entre ambas.

3.1.1.3 Classificação como GENERICO

Um dos critérios que nos levantou mais dúvidas e que, consequentemente, provocou maior discordância entre anotadores foi o da classificação de uma entidade como GENERICO[4]: em 92 entidades deste tipo, 22 têm o atributo COMENT preenchido com os valores 2/3 ou DUVIDA_DIRECTIVASTEMPO[5], num total de 1204 expressões temporais das quais 83 foram marcadas da mesma forma.

De acordo com o critério 3, uma expressão temporal deveria ser deste tipo se verificasse um dos subcritérios do critério 2 e não verificasse o critério 1. Ou seja, se a expressão (lexicalmente) contivesse elementos temporais, mas não fosse uma resposta adequada a uma das interrogativas previstas no critério 1: *(<prep>) quando?*, *(<prep>) quanto tempo?*, *(<haver>) quanto tempo?* ou *com que frequência?*.

Considere-se o exemplo (3.6).

(3.6) Lápis-lazúli, conhecido também como lápis, é uma rocha metamórfica de cor azul utilizada como gema ou como rocha ornamental utilizada desde antes de 7000 a.C. em Mehrgarh na Índia, situado <EM ID="H2-dhy6432-141" CATEG="TEMPO" TIPO="GENERICO">**nos dias de hoje**</EM> no Paquistão

Embora a expressão *nos dias de hoje* corresponda a um locativo temporal, o critério sintáctico supra-mencionado não parece poder aplicar-se:

**quando é que estava [Mehrgarh] situada no Paquistão? / nos dias de hoje*

Por esse motivo, e de acordo com as directivas, acabámos por atribuir a classificação de GENERICO a esta expressão e a outras cujo par pergunta-resposta nos parecia duvidoso ou mal-formado.

[4] Achamos, a este respeito, que a denominação GENERICO não é apropriada e pode induzir em erro, já que genérico tem um sentido concreto bem diferente em linguística e na área do tempo e do aspecto em particular (consulte-se, por exemplo, Krifka et al. (1995), Dahl (1973) e, em português, Lopes e Santos (1993)).

[5] Como referido no capítulo 1, o valor 2/3 é usado para marcar entidades cuja classificação não resultou do total acordo dos anotadores; usámos DUVIDA_DIRECTIVASTEMPO para indicar entidades em que as anotadoras tiveram dúvidas, geralmente associadas a diferentes interpretações possíveis das directivas.

Contudo, houve casos em que apesar de termos achado pouco natural a formulação das interrogações com *<prep> quando?* ou *quando?*, considerámos que esta era mesmo assim mais aceitável do que no caso anterior, como é o caso de expressões temporais que modificam um outro sintagma nominal sem valor temporal.

Veja-se, por exemplo, as expressões *da década de 1920* e *nos anos 1950* na frase (3.7).

(3.7) No Brasil, eles remontam ao século dezenove, com o grupo dos românticos em São Paulo, os grupelhos de poetas simbolistas, os modernistas <EM ID="gtqqq-168" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**da década de 1920**</EM>, o grupo antropofágico, os concretistas <EM ID="gtqqq-169" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**nos anos 1950**</EM>, o coletivo Rex de artistas na década seguinte

Tanto no primeiro caso, como no segundo, pareceu-nos possível, apesar de questionável, formular o par pergunta-resposta:

*?*os modernistas de quando? / da década de 1920*

*?*os concretistas quando? / nos anos 1950*

Assim, optámos por não classificar estas expressões como `GENERICO`.

Como ilustra este exemplo, a questão da aceitabilidade/inaceitabilidade dos pares pergunta-resposta interfere não só com a classificação como `GENERICO`, como também com as restantes classificações, pois se o par pergunta-resposta não for aceitável então a expressão temporal tem o tipo `GENERICO`, noutros casos a expressão receberá outra classificação (`DATA`, `FREQUENCIA`, etc.).[6]

3.1.1.4 Classificação como DURACAO

De acordo com as directivas, as entidades com tipo `DURACAO` referem "uma duração de tempo contínuo" e correspondem a entidades que exprimem

> quantificação temporal, sendo constituídas por nomes de unidades de medida de tempo e determinantes com função de quantificadores (e.g.. numerais). Podem, por vezes, ser introduzidas, facultativamente, pela preposição *durante* e respondem adequadamente à interrogativa *(<prep>) quanto tempo?*".
>
> (Hagège et al., 2008)

Como nos pareceu que a leitura estrita deste critério levaria à exclusão de expressões que não contivessem unidades de medida de tempo, e como são dados dois exemplos de expressões deste tipo que não as incluem (*todo o inverno* e *três manhãs*), de facto, acabámos por classificar como `DURACAO`, expressões que não incluíam unidades de medida temporal.

[6] Refira-se, ainda, que Jorge Baptista, na sua recensão ao presente capítulo, discordou dos nossos juízos de valor quanto à gramaticalidade/aceitabilidade dos pares pergunta-resposta, até no modo como os pares foram formulados. No caso (3.6), sugeriu a formulação da pergunta usando outro verbo copulativo, o que tornaria, no seu entender, o par mais aceitável:

?**quando é que* [*Mehrgarh*] *passou a estar situado no Paquistão? / nos dias de hoje*

Parece-nos pois importante referir que poderá haver mais casos de perguntas que permitam estabelecer que uma dada expressão representa uma data do que aquelas mencionadas nas directivas.

Para tal, teriam de responder adequadamente à interrogação *(<prep>) quanto tempo?* e verificar pelo menos um dos subcritérios 2. Como exemplos, veja-se (3.8) e (3.9).

(3.8) Detroit tem <EM ID="2ght33-10" CATEG="TEMPO" TIPO="DURACAO">**por longo tempo**</EM> sido um lugar de referência na imaginação sônica.

(3.9) Passa a viver com a avó Dionísia e as duas tias na Rua da Bela Vista, 17. A mãe e o padrasto também retornam a Lisboa <EM ID="aa87333-155" CATEG="TEMPO" TIPO="DURACAO">**durante um período**</EM> de férias <EM ID="aa87333-156" CATEG="TEMPO|TEMPO" TIPO="DURACAO|GENERICO">**de um ano**</EM>

3.1.1.5 Classificação de expressões iniciadas por *há*

No caso de expressões temporais iniciadas por *há*, optámos pelo valor durativo sempre que a formulação com *durante* fosse possível, e por tempo de calendário nos casos em que a expressão de tempo respondesse exclusivamente a *(<prep>) quando?*. Vejam-se os exemplos (3.10) a (3.12).

(3.10) o pensamento cartesiano <EM ID="H2-bbb-231" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**de há quatro séculos**</EM>

(3.11) o CCB iniciava, <EM ID="Ntyr-78-400" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**há quinze anos**</EM>, a sua actividade

(3.12) é um projeto que vem sendo realizado <EM ID="bob-14949-607" CATEG="TEMPO" TIPO="DURACAO">**há mais de dois anos**</EM>

Em dois casos (um deles ilustrado no exemplo (3.13)), em que ambas as interpretações nos pareceram possíveis, marcámos ambas. Apesar de as directivas do tempo não preverem a marcação de vagueza, pareceu-nos adequado nesta situação tirar partido dessa característica do esquema de anotação do HAREM, em vez de ter de optar arbitrariamente por uma das análises.

(3.13) nesta comemoração de uma data que deve ser pretexto para uma renovação da nossa ligação com o público que connosco habita, todos os dias, <EM ID="Ntyr-78-100" CATEG="TEMPO|TEMPO" TIPO="TEMPO_CALEND|DURACAO" SUBTIPO="DATA|">**de há 15 anos para cá**</EM>

Refira-se ainda a propósito do exemplo (3.13) que, de certo modo, vemos toda a sequência *todos os dias, de há 15 anos para cá* como uma expressão temporal complexa que denota um valor de frequência que só está completamente definido se tivermos em conta toda a expressão. Ou seja, esta frequência só é válida *de há 15 anos para cá* e como tal os dois valores de frequência e duração são indissociáveis.[7]

[7] De acordo com as directivas, contudo, apenas a expressão *todos os dias* deveria ser anotada com o tipo `FREQUENCIA`.

3.1.1.6 Ausência de anotação relativa a TEMPO

O subcritério 2-2 permite a identificação de expressões temporais no caso de

> uma unidade de medida temporal (*dia*, *mês*, *trimestre*, *ano*, *século*, etc.) ou um advérbio terminado em *mente* derivado destas expressões (*diariamente*, *semanalmente*, *mensalmente*, etc.).
>
> (Hagège et al., 2008)

No entanto, não está previsto que adjectivos derivados dos nomes de unidades de medida o sejam. Por esse motivo, não anotámos expressões temporais que envolvessem esses adjectivos, nem os próprios adjectivos, mesmo que a expressão completa tivesse um valor de frequência, como no exemplo (3.14).

(3.14) A partir de maio de 2000 estará sendo lançada a Revista de Direitos Difusos, com mais de 100 páginas e *periodicidade bimestral*

Pareceu-nos que, nesses casos, também não seria aplicável o critério 2-8:

> expressões de frequência, como as seguintes: de vez em quando, às vezes, de quando em quando, frequentemente, etc.,
>
> (Hagège et al., 2008)

Um outro caso em que não anotámos como TEMPO foi o de nomes de acontecimentos usados com valor temporal. O problema destas expressões foi que correspondiam a casos em que nos parecia que o sentido era temporal, mas não estavam previstos nas directivas do TEMPO. No entanto, só nos apercebemos disso depois de termos anotado consistentemente esses casos com TEMPO na primeira versão da CD.

Uma vez que não nos parecia muito aceitável adicionar tarde demais esta cláusula, e como iria flagrantemente contra a filosofia do HAREM anotá-los como ACONTECIMENTO se estavam a indicar tempo, anotámos então essas entidades como OUTRO, em vez de termos usado a possibilidade oferecida pelo Segundo HAREM de classificar como CATEG="TEMPO" TIPO="OUTRO".

Assim, não anotámos como pertencendo à categoria TEMPO períodos históricos, como sejam os casos de *Idade Média* (veja-se o exemplo (3.15)) ou *Descobrimentos*.

(3.15) No seguimento do colapso de instituições monásticas e do escolasticismo nos finais da <EM ID="H2-dftre765-102" CATEG="OUTRO">**Idade Média**</EM>

No caso de *Idade Média*, em particular, considerámos que não se verificava o critério 2-6, como nos casos de:

> "*altura*, *tempo*, *momento*, *período*, *era*, etc."

pois não achamos que a palavra *idade* na expresssão *Idade Média* tenha as mesmas propriedades das palavras lá referidas: com efeito, ao contrário delas, *idade* não nos parece poder ser determinada por um possessivo (*nessa idade...*), modificada por uma relativa (*na idade em que...*) ou complementada por um sintagma preposicional (*na idade de...*). Ainda se poderia argumentar que está a ser modificada por um adjectivo em maiúscula, mas trata-se de um composto. Como o núcleo nominal (o próprio nome composto) não está a

ser também ele modificado por um adjectivo, não é modificado por uma relativa, nem é complementado por um outro sintagma preposicional, optámos por não anotar *nos finais da Idade Média* como TEMPO. Em vez disso, conforme já dito, anotámos expressões como esta com a categoria OUTRO.

Finalmente, também não anotámos expressões que, embora verificassem pelo menos um dos subcritérios 2, se encontrassem no contexto de uma pergunta. Nestes casos, não nos pareceu fazer sentido formular um par pergunta-resposta para poder verificar se estávamos perante o critério 1 ou 3. Veja-se, a título de exemplo, a frase interrogativa (3.16).

(3.16) Em que *ano* é que Torquemada foi nomeado Inquisidor Geral?

3.1.2 Opções relativas aos atributos do TEMPO estendido

3.1.2.1 Tensão entre dois tipos de DATA

Por vezes, tomámos opções de delimitação com consequências inesperadas, ou limitadoras, na marcação dos atributos estendidos do TEMPO. Por exemplo, considere-se os exemplos (3.17) e (3.18)

(3.17) o que traduz um crescimento de 2,4%, <EM ID="hub-51467-348" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+—01–T—E–LMP">**a partir de Janeiro do próximo ano**</EM>

(3.18) promulgado <EM ID="hub-18050-209" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABOLUTO" VAL_NORM="+——30T—E–LM-">**a 30 desse mês**</EM>

O problema destes casos é que são datas que obedecem a ambos os critérios definidos nas proposta do TEMPO como mutuamente exclusivos, e que repetimos aqui:

> datas, sejam elas absolutas (fórmulas contendo os três campos ANO-MES-DIA, nas quais até dois campos no máximo podem ser omitidos) ou referenciais (ET cuja resolução implica conhecer a data do momento da enunciação, ou conhecer a data de um outro evento que funciona então como referência temporal para a expressão a calcular).
>
> (Hagège et al., 2008)

Ou seja, os dois critérios definidos são, num primeiro caso, puramente sintácticos (ou mesmo lexicais), e no segundo totalmente semânticos, e nada garante a sua mútua exclusão.

Como se pode verificar, o caso da expressão *a partir de Janeiro do próximo ano* tanto inclui uma referência ao nome de um mês como exige saber em que momento foi enunciada para que se possa proceder à sua resolução. O mesmo se passa com *a 30 desse mês*, que inclui a referência a um dia (30) de um mês cuja localização na linha temporal é fornecida pelo resto do discurso.

Dado que, de acordo com as directivas do TEMPO, a divisão de tais expressões em duas EM temporais não era possível, tivemos de decidir se considerávamos as EM em causa como "referenciais" ou como "absolutas". A decisão foi arbitrária e recaiu na segunda escolha.

Não nos parece, contudo, que a correcta análise das expressões em questão passasse pela separação em duas partes independentes. Pelo contrário, o que nós defenderíamos era a interpretação da entidade como um todo (que, nos dois casos, implicava que a data não fosse absoluta).

3.1.2.2 Expressões com valor de data sem nenhum dos campos ANO-MES-DIA especificado

De acordo com as directivas, as datas com valor absoluto devem ter pelo menos explicitado um dos campos `ANO-MES-DIA`. No entanto, pareceu-nos que, tal como no exemplo dado nas directivas, *na era* [*dos dinossauros*], que tem valor absoluto sem ter nenhum desses campos explicitados, expressões temporais que não dependessem de outras expressões temporais no texto seriam igualmente anotadas como datas absolutas.

As expressões *No início do século XVI* e *nos anos 90* nas frases (3.19) e (3.20), respectivamente, são exemplos de casos anotados como datas absolutas.

(3.19) <EM ID="hub-66526-557" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="">**No início do século XVI**</EM> o rei D. Manuel I ordena uma grande reforma

(3.20) Carlos Gerbase faz parte de uma geração de cineastas que apareceu em Porto Alegre <EM ID="ric-54609-190" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="">**nos anos 90**</EM>

O problema com estes casos é que os campos do atributo `VAL_NORM` não permitem normalizar estas datas. No caso de *nos anos 90* (exemplo (3.20)), temos uma referência à década de 90, mas o ano em causa pode não ser 1990. Por esse motivo, não especificámos o campo ano do atributo `VAL_NORM` com esse valor. Note-se, além disso, que apesar de ter valor absoluto, por se tratar de uma referência à decada de 90, cuja resolução não depende de uma outra data, no contexto em questão não se trata de uma referência a um ponto específico na década de 90, mas sim a um intervalo de tempo[8]. Por exemplo, na frase (3.21), a mesma expressão refere-se a um ponto específico no tempo, por se tratar de um estreia que ocorre num dia particular, ou até a mais pontos se pensarmos que o mesmo filme pode estrear em vários sítios em datas diferentes.

(3.21) O primeiro filme de Carlos Gerbase estreou <EM ID="ex-903" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="">**nos anos 90**</EM>

3.1.2.3 Preenchimento de VAL_DELTA e VAL_NORM na ausência total de informação

Quando não existia informação que permitisse preencher pelo menos um dos campos de `VAL_DELTA` ou `VAL_NORM`, optámos por preencher esse atributo na colecção dourada com o valor "", em vez de o omitir[9].

(3.22) <EM ID="aa94781-176" CATEG="TEMPO" TIPO="DURACAO" VAL_NORM="">**Há anos**</EM> que se discute se a fotografia da autoria do famoso fotojornalista

[8] Usamos aqui uma noção mais alargada de intervalo do que a definida nas directivas do tempo, não obrigando a que os limites sejam explícitos.

[9] As directivas sugeriam que se poderia optar pela omissão do parâmetro em vez do preenchimento com "".

Tabela 3.1: Dados quantitativos sobre a CD do TEMPO

Parâmetro	Valor
Documentos	30
Parágrafos	304
Palavras	12992
Entidades	1508
Entidades vagas	118
Entidades TEMPO	232

(3.23) o miliciano fotografado não é o que foi morto <EM ID="aa94781-192" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="SIMULT" VAL_DELTA="">**naquele dia**</EM>

Este é um pormenor meramente técnico, mas que documentamos aqui para uma maior clareza na especificação da anotação efectuada.

3.2 O TEMPO em números no Segundo HAREM

No capítulo 1 caracterizámos a CD do Segundo HAREM, a qual inclui 1195 entidades TEMPO (das quais seis são vagas com outra categoria) que foram anotadas com os atributos do HAREM clássico: CATEG, TIPO e SUBTIPO.

A CD do TEMPO, tal como mencionámos anteriormente, foi constituída com o objectivo de adicionar às expressões temporais os atributos estendidos do TEMPO e corresponde a uma sub-colecção de trinta documentos da colecção dourada. Desses documentos, doze correspondem aos documentos que constituem a CD do ReRelEM, e os restantes dezasseis são os primeiros documentos da colecção dourada que não incluem entidades TEMPO marcadas com 2/3 ou DUVIDA_DIRECTIVASTEMPO.

Tínhamos inicialmente previsto anotar apenas 10% dos documentos da colecção dourada com os atributos estendidos do TEMPO e também do ReRelEM, o que corresponderia a doze documentos. No entanto, esses documentos incluíam apenas 85 entidades TEMPO (o que corresponde a menos de 10% das entidades TEMPO) e sobre algumas delas pesavam ainda dúvidas de anotação ou o não total acordo dos anotadores. Optámos, então, por seleccionar mais uma série de documentos que não tivessem essa última característica e que além disso incluíssem cerca de 10% das entidades TEMPO da colecção dourada. Foram igualmente considerados três documentos que permitissem aumentar um pouco a representividade das entidades do tipo HORA, já que o objectivo desta CD era ser usada na avaliação dos atributos estendidos. Acabámos por juntar todos os documentos na CD do TEMPO, cobrindo assim cerca de 19% das entidades TEMPO.

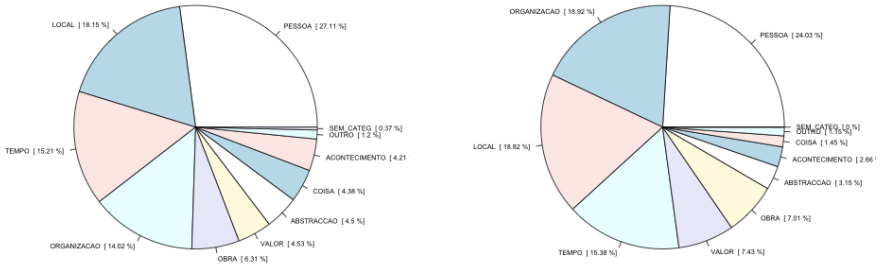
Esta sub-colecção é assim constituída por 304 parágrafos e 12992 palavras; das 1508 entidades anotadas, 232 estão anotadas com a categoria TEMPO (cf. tabela 3.1). Além disso, das 118 entidades vagas, nenhuma envolve a categoria TEMPO e existem 89 sequências delimitadas com ALT, das quais cinco envolvem a categoria TEMPO.

Comparando a distribuição das categorias na CD (já ilustrada no capítulo 1, mas que aqui se reproduz na figura 3.1(a)), e na CD do TEMPO (cf. figura 3.1(b)), podemos ver que as entidades marcadas com a categoria TEMPO correspondem, em ambos os casos, a

cerca de 15% do total de entidades. Notamos, no entanto, que isso não foi um critério que tivéssemos estabelecido à partida.

Pode igualmente ver-se que a distribuição das categorias na CD do TEMPO é diferente da observada na CD do Segundo HAREM. Por exemplo, ORGANIZACAO é a quarta categoria mais frequente na CD do Segundo HAREM, correspondendo a cerca de 14% das entidades, enquanto, na CD do TEMPO, ORGANIZACAO é a segunda mais frequente, compreendendo quase 19% das entidades. No que se refere à categoria TEMPO, verifica-se que esta é a quarta categoria mais frequente na CD do TEMPO, apesar de ser a terceira mais frequente na CD do Segundo HAREM.

Não temos contudo a certeza de tal constatação ser relevante, a não ser pelo facto de poder ser uma explicação para a avaliação na CD do TEMPO - em relação ao HAREM clássico (ou seja, excluindo a avaliação dos atributos estendidos do TEMPO) - ter resultado numa ordenação diferente dos sistemas em termos de desempenho.



(a) CD do Segundo HAREM (b) CD do TEMPO

Figura 3.1: Distribuição de categorias na CD do Segundo HAREM e na CD do TEMPO

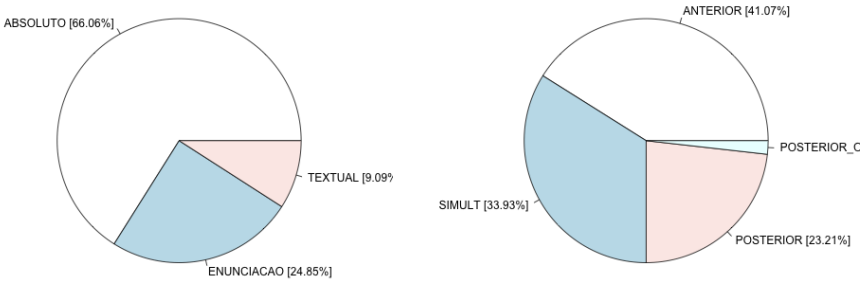
A distribuição de tipos da categoria TEMPO na tabela 3.2 mostra uma prevalência de expressões temporais do tipo TEMPO_CALEND (mais de 80%) tanto na CD como na CD do TEMPO, das quais a grande maioria são datas (cerca de 89% na CD e aproximadamente 84% na CD do TEMPO). Talvez por termos forçado a inclusão de mais entidades com o subtipo HORA na CD do TEMPO, como referido acima, este subtipo é mais frequente do que o subtipo INTERVALO nessa CD do que na CD do Segundo HAREM.

A figura 3.2 ilustra a distribuição dos atributos estendidos na CD do TEMPO. Em particular, a figura 3.2(a) ilustra a distribuição dos valores do atributo TEMPO_REF. Como se pode observar, existe um maior uso de expressões temporais com valor absoluto (cerca de 66% das entidades tiveram o atributo TEMPO_REF preenchido com valor ABSOLUTO) do que com valor referencial: cerca de 25% das vezes o atributo TEMPO_REF tem o valor ENUNCIACAO e cerca de 9% tem o valor TEXTUAL. No que diz respeito ao atributo SENTIDO (ver figura 3.2(b)), temos que, na maioria das entidades com valor referencial, esse atributo foi preenchido com

Tabela 3.2: Distribuição de tipos (T) e subtipos (ST) da categoria TEMPO: dentro de parêntesis encontram-se tipos e subtipos de outras categorias que não TEMPO

		CD		CD do TEMPO	
TIPO	SUBTIPO	T	ST	T	ST
TEMPO_CALEND		973		195	
	DATA		873		164
	INTERVALO		63		12
	HORA		37		19
GENERICO		89		12	
FREQUENCIA		71		15	
DURACAO		56		12	
DURACAO|(QUANTIDADE)		3		–	
TEMPO_CALEND|DURACAO		4		–	
	DATA|		4		–
DURACAO|GENERICO		1		–	
(EFEMERIDE)|GENERICO		1		–	
TEMPO_CALEND|(QUANTIDADE)		1		–	
	INTERVALO|		1		–
TEMPO_CALEND|GENERICO		1		–	
	DATA|		1		–

o valor ANTERIOR (41%) e SIMULT (34%), enquanto apenas em 23% dos casos foi preenchido com o valor POSTERIOR.



(a) TEMPO_REF (b) SENTIDO

Figura 3.2: Distribuição dos atributos TEMPO_REF e SENTIDO na CD do TEMPO

Quanto aos atributos de normalização, a tabela 3.3 mostra que uma parte significativa das expressões temporais referenciais não explicita a distância temporal ao momento da

Tabela 3.3: Preenchimento dos atributos de normalização VAL_DELTA e VAL_NORM na CD do TEMPO

Atributo	Preenchido	Vazio	Total
VAL_DELTA	36	19	55
VAL_NORM de DURACAO	8	4	12
VAL_NORM de HORA	19	0	19
VAL_NORM de ABSOLUTO	101	8	109

referência, já que cerca de um terço dos valores de VAL_DELTA correspondem à sequência vazia. Numa proporção muito menor, vê-se na mesma tabela que para as data absolutas em cerca de 7% dos casos também não foi possível determinar nenhum dos campos do atributo VAL_NORM.

Finalmente, pode-se observar na figura 3.3, que as datas absolutas cujo campo referente ao ano se encontra especificado se distribuem entre 1131 e 2011. As datas concentram-se sobretudo na passagem do século XV para o século XVI, em meados do século XX e no início do século XXI, o que sugere que os documentos da CD do TEMPO descrevem sobretudo eventos decorridos nesses períodos.

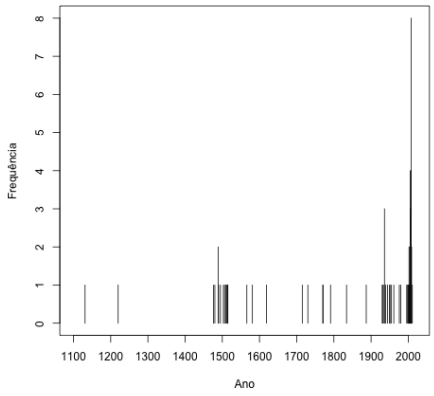


Figura 3.3: Histograma dos anos especificados nas datas absolutas na CD do TEMPO

3.3 Avaliação

Embora a proposta de reconhecimento e normalização de expressões temporais (pista do TEMPO) tenha sido feita de forma independente da proposta de reconhecimento das entidades pertencentes a outras categorias (HAREM clássico), a avaliação dos sistemas no que respeita às entidades TEMPO foi levada a cabo de forma integrada com o HAREM clássico.

Queremos com isto dizer que:

- não desenhámos todo um novo processo de avaliação exclusivo das entidades TEMPO. Pelo contrário, como descrito em pormenor no capítulo 5, integrámos apenas na

sequência de avaliação do HAREM clássico um novo módulo para atribuir uma pontuação adicional às entidades TEMPO, no caso dos atributos estendidos do TEMPO (SENTIDO, TEMPO_REF, VAL_NORM e VAL_DELTA) estarem correctamente preenchidos. Aliás, se não fosse o facto de existirem atributos específicos de TEMPO, as entidades TEMPO teriam sido avaliadas como se se tratassem de entidades de outra categoria qualquer.

- não separámos a avaliação das entidades pertencentes a outras categorias da avaliação das entidades TEMPO. Consequentemente, como vimos no capítulo 1, o HAREM clássico inclui diversos cenários em que as entidades da categoria TEMPO foram avaliadas, e, em particular, um dos cenários inclui apenas entidades TEMPO. Com ou sem separação, a avaliação da pista do TEMPO pode ser vista simplesmente como a avaliação num cenário selectivo constituído apenas por essa categoria.

Esta forma integrada de fazer a avaliação tem a vantagem de permitir ver a tarefa de reconhecimento de entidades mencionadas como um todo, não atribuindo um estatuto especial às entidades TEMPO por terem sido o alvo de uma proposta de anotação independente da das entidades pertencentes às restantes categorias.

Por essa razão, mesmo na CD do TEMPO fizemos a avaliação sem separar a avaliação do HAREM clássico da avaliação da pista do TEMPO.

Uma outra consequência adicional é que usaremos o termo "*avaliação da pista do TEMPO*" para designar apenas a avaliação no cenário TEMPO; em todos os outros casos de cenários que contêm a categoria TEMPO, referir-nos-emos à "*avaliação das entidades TEMPO*".

Contudo, como as expressões temporais têm outros atributos, além dos atributos do HAREM clássico, os sistemas foram avaliados, no que respeita às entidades TEMPO, de quatro modos distintos:

Clássico, tendo em conta apenas os atributos CATEG, TIPO e SUBTIPO do HAREM clássico;

Estendido completo, tendo em conta todos os atributos;

Estendido sem normalização, igorando os atributos VAL_NORM e VAL_DELTA;

Estendido só com normalização, ignorando os atributos SENTIDO e TEMPO_REF.

A avaliação no modo clássico foi feita tendo como referência tanto a CD do Segundo HAREM completa (cujos resultados no cenário total e nos vários cenários selectivos foram apresentados no capítulo 1), como apenas o subconjunto de documentos pertencentes à CD do TEMPO (neste último caso, quer dizer que os atributos estendidos de TEMPO não foram tidos em conta na avaliação)[10]; os restantes modos eram apenas aplicáves no caso de se usar a CD do TEMPO.

Para cada um dos modos de avaliação usando a CD do TEMPO, avaliámos os sistemas no cenário total (ou seja, tendo em conta todas as categorias) e em todos os cenários selectivos de avaliação que incluem a categoria TEMPO (ou seja, nos cenários selectivos 2, 4 e 6, cuja descrição se encontra na tabela 1.3). Além disso, avaliámos ainda os sistemas apenas relativamente à categoria TEMPO.

Salientamos ainda que a avaliação na CD do TEMPO foi feita com base na avaliação estrita de ALT, apesar de também poder ser feita a partir da avalição relaxada de ALT (ver

[10] A avaliação no modo clássico foi feita também na CD do TEMPO para servir de referência à avaliação tendo em conta os atributos estendidos de TEMPO.

Tabela 3.4: Panorâmica da avaliação das entidades TEMPO: CD indica que a avaliação foi feita com a CD do Segundo HAREM e CDT que a avaliação foi feita com a CD do TEMPO

	Modo			
Cenário	Clássico	Estendido Completo	Sem Norm.	Só Norm.
Total	CD e CDT	CDT	CDT	CDT
Selectivos 2, 4 e 6	CD e CDT	CDT	CDT	CDT
TEMPO	CD e CDT	CDT	CDT	CDT

Tabela 3.5: Sistemas participantes na categoria TEMPO

Sistema	TIPO	SUBTIPO	SENTIDO	TEMPO_REF	Normalização
CaGE2					
PorTexTO	x	x			
PRiberam	x	x		x	
REMBRANDT	x	x			
REMMA	x				
SeRELeP	x	x			
XIP-L2F/Xerox	x	x	x	x	x

capítulos 1 e 5 para mais pormenores sobre estas duas formas de avaliação das análise alternativas). Esta escolha foi em parte arbitrária, porque contávamos inicialmente ter feito a avaliação de ambas as formas, o que acabou por não se verificar por falta de tempo. No entanto, como vimos no capítulo 1, não existe uma diferença significativa de desempenho no HAREM clássico entre essas duas formas de avaliação.

A tabela 3.4 sumariza as várias formas de avaliação para a pista do TEMPO.

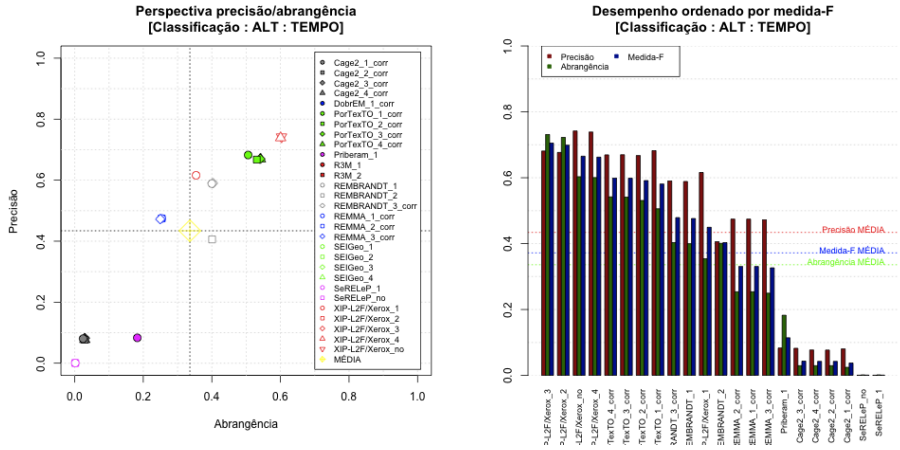
3.3.1 Sistemas participantes

Os sistemas que participaram na pista do TEMPO e o seu nível de envolvimento na tarefa, ou seja os atributos de TEMPO que foram preenchidos pelos sistemas, encontram-se descritos na tabela 3.5.

Dos dez sistemas participantes no Segundo HAREM, somente três não fizeram reconhecimento de expressões temporais, o que demonstra um claro interesse em reconhecer este tipo de entidades. No entanto, apenas dois sistemas foram além do preenchimento dos atributos do HAREM clássico: o sistema da Priberam atribuiu ainda valores ao atributo TEMPO_REF, e o sistema do grupo proponente (XIP-L2F/Xerox) preencheu todos os atributos.

3.3.2 Resultados

Apesar de no capítulo 1 termos apresentado os resultados do HAREM clássico, que incluem as entidades da categoria TEMPO, não demos destaque à avaliação da categoria TEMPO em particular, pois isso corresponde, como já referimos, à avaliação da pista do TEMPO. Começamos pois por apresentar na figura 3.4 os resultados de avaliação no cenário selectivo constituído pela categora TEMPO (e atributos TIPO e SUBTIPO) na CD do Segundo HAREM.



(a) Precisão e abrangência

(b) Precisão, abrangência e medida F

Figura 3.4: Resultados dos sistemas na classificação da categoria TEMPO e respectivos atributos TIPO e SUBTIPO

As quatro melhores corridas pertencem ao sistema XIP-L2F/Xerox, com valores de medida F de 0,7054, 0,6989, 0,6654 e 0,6625, enquanto as quatro corridas seguintes, pertencentes ao sistema PorTexTO, têm valores de medida F claramente mais baixos que variam entre 0,58 e 0,60; as restantes corridas obtiveram valores abaixo de 0,479.

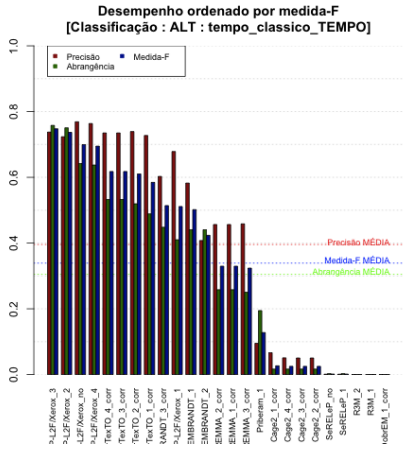
Destaca-se ainda que as duas melhores corridas apresentam valores de precisão e abrangência equilibrados: no melhor caso, cerca de 0,68 e 0,73 respectivamente, o que não acontece com as restantes corridas, que apresentam maiores diferenças entres as duas métricas: por exemplo, a terceira melhor corrida tem uma abrangência 0,14 abaixo da precisão que se situa em 0,74.

Centramo-nos agora na análise dos resultados na CD do TEMPO, também no cenário selectivo composto apenas pela categoria TEMPO.

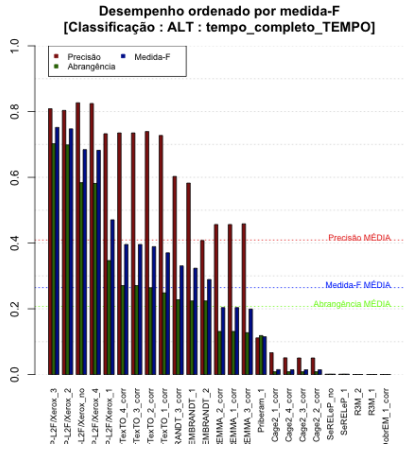
Uma vez que a avaliação das entidades do TEMPO nos modos estendidos tem por base a avaliação no modo clássico, e também para efeitos de comparação com o desempenho na CD do Segundo HAREM, apresentamos em primeiro lugar na figura 3.5(a) o desempenho nesse modo na CD do TEMPO.

Comparando então o desempenho nas duas colecções douradas, podemos concluir que os resultados para a categoria TEMPO no HAREM clássico são melhores para a maioria dos sistemas (cerca de 0,4 mais alto em termos de medida F, no caso da melhor corrida) na CD do TEMPO (figura 3.5(a)) do que na CD do Segundo HAREM (figura 3.4(a)). Em todo o caso, a ordem de desempenho das nove melhores corridas não é alterada e em média estas corridas têm valores de medida F 0,0269 mais altos.

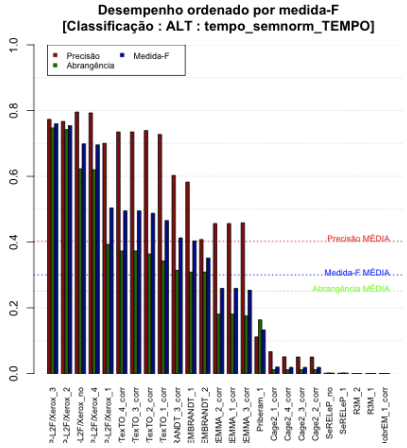
No que diz respeito à avaliação no modo estendido completo (cf. figura 3.5(b)), a comparação com a avaliação clássica na CD do TEMPO (cf. figura 3.5(a)) denota um decréscimo nas classificações de todas as corridas (cerca de 0,09 em média) excepto nas duas melhores corridas do sistema XIP-L2F/Xerox, que subiram ligeiramente, visto que foi o único que tentou atribuir todos os atributos de TEMPO. Isso explica-se porque ao modo estendido



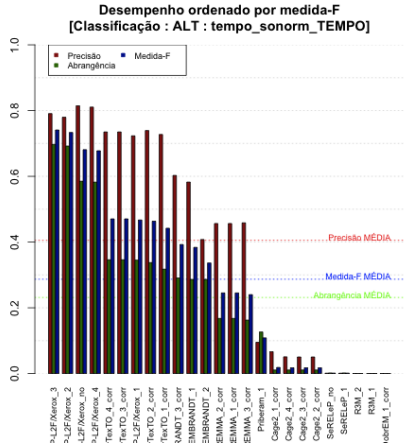
(a) HAREM clássico na CD do TEMPO



(b) Modo estendido completo



(c) Modo estendido sem normalização



(d) Modo estendido só normalização

Figura 3.5: Precisão, abrangência e medida F na avaliação da pista do TEMPO (na CD do TEMPO)

Tabela 3.6: Resumo dos resultados da pista do TEMPO na CD e na CD do TEMPO

Modo	XIP-L2F/Xerox					Priberam
	3	2	no	4	1	1
Clássico na CD	0,7054	0,6989	0,6654	0,6626	0,4496	0,1143
Clássico na CD do TEMPO	0,7477	0,7369	0,6997	0,6949	0,5112	0,1277
Estendido completo	0,7518	0,7475	0,6843	0,6821	0,4707	0,1150
Estendido sem normalização	0,7600	0,7543	0,6987	0,6958	0,5035	0,1326
Estendido só com normalização	0,7400	0,7334	0,6811	0,6778	0,4672	0,1085

está asssociada um maior valor máximo do que no caso de se ter em conta apenas os atributos do HAREM clássico de TEMPO (CATEG, TIPO e SUBTIPO), visto que se atribui uma pontuação adicional às entidades de TEMPO pela existência dos atributos estendidos (TEMPO_REF, SENTIDO, VAL_DELTA, e VAL_NORM), o que faz naturalmente diminuir o valor da abrangência se esses atributos não forem preenchidos.

Mostramos ainda os resultados da avaliação no modo estendido sem normalização (cf. figura 3.5(c)) e só com normalização (cf. figura 3.5(d)). O aspecto que mais se evidencia é que comparando estes resultados com os mostrados para a avaliação no modo estendido completo (cf. figura 3.5(b)), as corridas de sistemas que tentaram resolver a tarefa têm melhores valores de medida F nestes dois modos. Voltamos a frisar que isso se deve, neste caso, a uma diminuição do valor máximo alcançável, pois estão a ser avaliados menos atributos do que no modo estendido completo.

Dado que apenas dois sistemas participaram na atribuição dos atributos estendidos, destacamos na tabela 3.6, o desempenho destes dois sistemas em termos de medida F.

3.4 Sugestões para o futuro da avaliação do TEMPO

Como notas finais abordaremos brevemente algumas questões que esperamos possam contribuir para uma melhor avaliação do TEMPO. Remetemos o leitor para o capítulo 6 para um balanço por pista e também para o capítulo 2 para o balanço dos proponentes desta pista.

3.4.1 Medida de avaliação

Um primeiro aspecto prende-se com a medida de avaliação que propusemos para ter em conta também os atributos estendidos, em torno da qual acabou por não haver qualquer discussão.

Esta medida combina o valor da medida obtido por avaliar os atributos estendidos com o que é obtido por avaliar os atributos do HAREM clássico da entidade. Embora seja possível atribuir pesos diferentes às parcelas dos vários atributos, não chegámos a experimentar variar o valor dos mesmos. No entanto, pensamos que para uma análise mais detalhada dos atributos estendidos poder-se-ia colocar a zero os pesos dos atributos CATEG, TIPO e SUBTIPO. Com esse procedimento, que pode ser facilmente testado no serviço de avaliação SAHARA (consulte-se o apêndice G para mais informações), as expressões temporais passariam então a ser apenas avaliadas em relação aos atributos estendidos.

Além disso, também julgamos que seria útil ter uma forma de avaliação relativa. Nesse caso, seriam avaliados os atributos estendidos apenas de expressões temporais correctamente classificadas quanto aos atributos do HAREM clássico, concentrando assim a análise nas entidades em que existe alguma possibilidade dos atributos estarem correctos (pois com certeza, nas que estão incorrectamente classificadas os atributos estão ausentes ou mal atribuídos). Seria então necessário colocar a zero os pesos dos atributos do HAREM clássico de entidades incorrectamente classificadas (de outro modo, entidades incorrectamente classificadas teriam uma penalização pela classificação espúria).

3.4.2 Estudos empíricos iluminàveis pela LÂMPADA

Tendo em conta que documentámos as várias dúvidas de anotação acrescentando `COMENT` na colecção dourada disponibilizada na LÂMPADA, o pacote de recursos do Segundo HAREM, pensamos que a análise dessas expressões temporais poderia contribuir para uma melhoria na definição da tarefa. Além disso, serviria igualmente para melhorar a própria anotação.

Mais concretamente, seria interessante:

- refazer a avaliação depois de o grupo do TEMPO corrigir esses casos, ou seja, decidir em cada caso como é que as dúvidas relatadas seriam esclarecidas;
- refazer a avaliação removendo (através da marcação de `OMITIDO`) todos os `COMENT` relativos ao `TEMPO`, para ver se em relação a estes casos também houve mais dispersão em relação ao que cada sistema marcou (e se portanto também são os casos mais discriminadores ou difíceis);
- refazer a avaliação depois de escolher e reanotar os casos das preposições, ou seja, apenas incluir nas expressões temporais as preposições que marcam tempo, e excluir das expressões temporais as preposições que vêm, por exemplo, da subcategorização do verbo.

3.4.3 Opiniões diferentes sobre o REM temporal, ou melhor, sobre o RET

Salientamos que, mesmo assim, temos uma opinião diferente em relação a muitas questões linguísticas da proposta do TEMPO. Algumas dessas questões foram sendo mencionadas na secção 3.1, a propósito das dúvidas que tivemos durante o processo de anotação; em Carvalho e Mota (2009), que ainda está em preparação, fazemos uma análise comparativa entre a anotação do `TEMPO` no Primeiro e no Segundo HAREM, em que mostramos os aspectos positivos e negativos da actual proposta de identificação, classificação e normalização de expressões temporais, e focamos em mais detalhe os aspectos em que estamos em desacordo.

Devemos, no entanto, ao concluir o capítulo, referir uma questão que nos parece central. Do nosso ponto de vista, o reconhecimento de entidades temporais (RET) não é, na sua maior parte, redutível ao REM no sentido em que extravasa em muito a identificação e classificação de entidades únicas e bem denominadas, o que aliás é particularmente visível em português na própria grafia: ao invés de marcar em maiúsculas, as expressões temporais são maioritariamente escritas em minúsculas. Por outro lado, tem também afinidades com a outra tarefa que foi proposta no MUC e que englobámos talvez erradamente

no HAREM, a do REN (reconhecimento de entidades numéricas), e que seria também de explorar e melhorar em futuras avaliações conjuntas para o português.

Estamos contudo convencidos, tal como o grupo do TEMPO, que a área do RET é fascinante e relevante para o processamento da nossa língua, e gostaríamos de a continuar a desbravar em português no futuro.

Agradecimentos

Estamos gratos à Caroline Hagège, Graça Volpes Nunes e Olga Craveiro, bem como ao Jorge Baptista, pelos seus comentários e sugestões que ajudaram certamente a criar um texto mais claro.

Capítulo 4

Relações semânticas do ReRelEM: além das entidades no Segundo HAREM

Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho e Cristina Mota

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 4, p. 77–96.

Neste capítulo apresentamos a pista do ReRelEM (Reconhecimento de Relações entre Entidades Mencionadas), integrada no Segundo HAREM. Essa pista tem como objetivo a avaliação de sistemas que identifiquem e classifiquem relações semânticas entre entidades mencionadas (EM) em um conjunto de textos da língua portuguesa, uma tarefa complementar à que é avaliada no HAREM clássico. No ReRelEM são consideradas apenas relações entre EM; ou seja, relações entre EM e pronomes ou outros tipos de sintagmas nominais, por exemplo, não são anotadas. Além disso, apenas consideramos relações entre EM em um mesmo documento.

Como tarefa dependente e integrada no HAREM, compartilha com este os mesmos pressupostos, apresentados e discutidos no capítulo 1[1], em particular a definição de EM e as categorias em que esta se enquadra, o que se reflete (i) na classificação das relações em contexto e (ii) no processo de escolha dos tipos de relação considerados.

Quanto ao primeiro ponto, o ReRelEM depende de uma anotação que considera o valor semântico das relações entre EM apenas quando inseridas em um contexto. Por isso, relações que, embora possam fazer sentido de um ponto de vista "puramente lexical" (ou de conhecimento de dicionário/almanaque), se não aparecerem num contexto apropriado não são consideradas válidas. Considere-se o seguinte exemplo fictício:

(4.1) *Portugal* perdeu para a *Alemanha* nas quartas-de-final da Eurocopa. Eu vi o jogo com os amigos na *Praça da República*. Depois da derrota, os bares de *Coimbra* estavam cheios.

Em (4.1), as entidades mencionadas pelas designações *Portugal* e *Alemanha* não são locais, são equipes (ou seja, as palavras *Portugal* e *Alemanha* constituem no contexto acima uma menção aos jogadores), e portanto as EM devem ser classificadas como pertencendo à categoria PESSOA, conforme as directivas do Segundo HAREM. Deste modo, embora exista uma relação de inclusão entre os **locais** *Praça da República* e *Coimbra*, não existe relação de inclusão entre estas ocorrências de *Coimbra* (ou de *Praça da República*) e a EM *Portugal*, visto que a relação de inclusão no ReRelEM foi apenas definida entre entidades da mesma categoria.[2]

Quanto ao segundo ponto, a escolha das relações que seriam alvo da tarefa, tínhamos duas opções: adotar um conjunto de relações lexicais existentes na literatura (veja-se, por exemplo, as propostas de Cruse (1986) ou Fellbaum (1998), ou as relações comuns da extração de informação (Chu-Carroll e Prager, 2007; Culotta e Sorensen, 2004; Roth e tau Yih, 2004; Zhao e Grishman, 2005), ou, pelo contrário, partir da análise dos textos, sem categorias pré-definidas. Embora mais morosa e ambiciosa, preferimos a segunda opção por dois principais motivos: a) a literatura sobre a análise e processamento de relações linguísticas entre palavras ou expressões não costuma tratar especificamente de relações entre EM (ou, dito de uma forma simplista, de relações entre nomes próprios), e a literatura de extração de informação pareceu-nos demasiado limitada para a escolha das relações; b) acreditamos que a tarefa humana de análise de textos, vistos como uma fonte da representação

[1] Fica, pois, aqui o aviso ao leitor que para compreender totalmente os exemplos do presente texto terá de se familiarizar um pouco com os pressupostos e categorias usados no HAREM.

[2] Note-se que a decisão de não relacionar estas duas entidades é defensável, mesmo que a relação de inclusão se estabelecesse entre entidades com categorias diferentes, pois, parece-nos, nenhum pesquisador defenderia a inclusão de local em equipe. Basta substituir Portugal por Sporting para compreender que "Praça da República incluído em Sporting"não é uma relação que queiramos aceitar como válida.

de conhecimento de uma dada língua, seria capaz de nos oferecer um vasto material, não apenas das relações, mas também das relações entre EM em língua portuguesa – e nisso estamos em sintonia com o que já é feito no HAREM com relação à escolha das categorias para a classificação de EM (Santos, 2007d).

Assim, as relações semânticas consideradas no ReRelEM foram obtidas a partir da leitura de textos da própria coleção do Segundo HAREM, bem como de alguns textos do corpo SUMMIT[3], e de outros que usamos para criar os textos de exemplo, em um processo cuidadoso de seleção e generalização. Um dos maiores desafios na definição da tarefa estava justamente em buscar um equilíbrio entre, por um lado, a especificidade com o conseqüente detalhamento de informação e, por outro, a generalidade, com o conseqüente maior poder descritivo das relações.

Sabemos que a decisão será sempre arbitrária mas, como um fator suavizante, é possível invocar a noção de relevância: uma determinada relação deve ser mantida específica ou, por outro lado, deve ser generalizada, na medida em que for relevante para o domínio a que se aplica. Esse critério, porém, não nos ajuda muito, uma vez que estamos no ambiente artificial de um contexto de avaliação de sistemas, atuando sobre um corpo genérico. Por isso, temos a consciência de que, embora as opções tomadas possam não ser as ideais de acordo com pontos de vista diversos, foram as que nos pareceram, durante o processo de identificação e análise, atender minimamente ao que nos propusemos: serem informativas e, ao mesmo tempo, com potencial de aplicação a diferentes domínios.

4.1 Relações do ReRelEM: o que anotar

Nesta secção, apresentamos as relações semânticas que definimos como o alvo do ReRelEM (e que estão conseqüentemente presentes na coleção dourada (CD) do ReRelEM), e discutimos as opções e dificuldades encontradas no seu estabelecimento.

Após a análise inicial dos textos, e tendo em vista os fatores já mencionados – generalidade e informatividade –, estabelecemos as seguintes relações entre EM: **identidade**, **inclusão** e **localização** (que podemos também chamar de **ocorrência em**). Além disso, englobamos inicialmente todas as restantes relações que consideramos relevantes, mas que não correspondem a nenhum dos tipos anteriormente explicitados, sob a designação de **outra** (relação).

4.1.1 Identidade

A relação de identidade estabelece-se entre EM que tenham o mesmo referente, ou seja, que designem a mesma entidade. Daí decorre que só pode existir entre EM que pertencem à mesma categoria. Isso quer dizer que a relação de identidade se estabelece não apenas entre expressões textuais formalmente idênticas ou que possam ser obtidas por transformações lexicais (como o apagamento (ou redução) lexical de um elemento), mas também entre EM relacionadas por abreviaturas, acrônimos, traduções ou "nomes alternativos", como o ilustram os seguintes exemplos, extraídos da CD do ReRelEM.

(4.2) assinam *Carta dos Direitos Fundamentais* (...). (...) esta Carta vai para além dos cidadãos...

[3] O SUMMit é um corpo marcado com co-referência, descrito em Collovini et al. (2007) e publicamente acessível de http://www.inf.pucrs.br/~linatural/Docs/Summ-it_v3.0.zip.

(4.3) Um simples teste de *ADN* (DNA)

(4.4) O coração da *Terra do Pão de Queijo* (...) trocou Nikiti por BH para suar...[4]

Nas frases (4.2), (4.3) e (4.4), entre as EM de cada um dos pares, *ADN*/*DNA*, *Carta dos Direitos Fundamentais*/*Carta* e *Terra do Pão de Queijo*/*BH*, existe uma relação de identidade.

Por outro lado, a identidade formal de expressões textuais não justifica por si só, naturalmente, a marcação da relação de identidade, que só pode ser aferida através de uma análise semântica dos textos em que essas expressões ocorrem, como é demonstrado no seguinte exemplo fictício:

(4.5) Os adeptos do *Porto* invadiram a cidade do Porto em júbilo.

Com efeito, as duas ocorrências da palavra *Porto* em (4.5) designam entidades distintas: respectivamente, um clube e um local.

4.1.2 Relação de inclusão

A relação de inclusão é bastante genérica e abrangente e, como o nome indica, deve ser estabelecida entre EM quando uma delas faz parte da outra. Esta relação tem como única restrição a exigência de que as EM relacionadas sejam da mesma categoria. Quando a entidade descrita por uma EM inclui a entidade descrita por outra, a relação entre essas duas EM é marcada como `inclui`. Quando a relação é inversa, ou seja, quando a entidade descrita por uma EM está incluída numa entidade descrita por outra, é marcada como `incluido`. (Ambas as formulações são válidas, e totalmente equivalentes no âmbito do ReRelEM, como será explicitado mais tarde.)

(4.6) *Lobos* recebidos em apoteose. (...) o capitão Vasco Uva explicou por que houve uma empatia tão grande entre...

(4.7) No *Terceiro Mundo*, os cientistas se desobrigariam de fornecer aos doentes o melhor tratamento médico conhecido. (...) O debate surgiu após estudos em Ruanda e na Tailândia

(4.8) Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o *Santanaraptor* ocuparia uma posição no grupo Tyrannoraptora, o mesmo do Tyrannossaurus rex

Tomando como exemplo as frases (4.6), (4.7) e (4.8), temos que:

Vasco Uva `incluido` Lobos[5]

Ruanda `incluido` Terceiro Mundo

Tailândia `incluido` Terceiro Mundo

Tyrannoraptora `inclui` Santanaraptor

Tyrannossaurus rex `incluido` Tyrannoraptora

[4] Para leitores que não conheçam suficientemente bem a geografia e cultura brasileiras, convém referir que o estado de Minas Gerais, bem como sua capital Belo Horizonte (BH), são conhecidos no Brasil como a Terra do Pão de Queijo.

[5] Ou Lobos `inclui` Vasco Uva. Por uma questão de economia escolhemos neste capítulo sempre apenas uma das duas possíveis formulações.

A relação de inclusão também vincula EM que, embora expressas pela mesma palavra, não apresentam uma relação de identidade, mas antes uma relação entre EM superficialmente idênticas, representando uma delas um elemento de uma classe e a outra a própria classe. Veja-se por exemplo as EM *Gemini* na frase (4.9).

(4.9) Astrônomos brasileiros esperam fotografar os primeiros planetas fora do Sistema Solar com a ajuda do maior telescópio do mundo, o *Gemini* (...) os telescópios Gemini têm capacidade científica...

Por fim, uma simplificação que propusemos neste primeiro ReRelEM foi a de que o valor dos atributos `TIPO` e `SUBTIPO` da categoria `LOCAL` não fosse levado em consideração na especificação das relações de inclusão. Conseqüentemente, um `LOCAL FISICO` pode, por exemplo, incluir um `LOCAL HUMANO`. No trecho abaixo, *Pampulha*, `LOCAL HUMANO`, inclui *Lago da Pampulha*, um `LOCAL FISICO`:

(4.10) Volta Internacional da *Pampulha* (...) Antonio Ricardo e mais uma turma da Araribia Runners trocou Nikiti por BH para suar ao redor do Lago da Pampulha

Deixamos para reflexão futura se esta decisão, que nos pareceu correta em termos da especificação das relações em português, tem conseqüências (teóricas ou práticas) para a categorização dos locais.

4.1.3 Relação de localização, ou de ocorrência em

A relação de localização (ou de ocorrência em) ocorre entre EM das categorias `ORGANIZACAO` ou `ACONTECIMENTO` e EM da categoria `LOCAL`, indicando a localização espacial de um evento ou de uma organização. É expressa por `ocorre_em`[6], enquanto a sua relação inversa é marcada através do nome `sede_de`.

(4.11) Em 9 de Setembro de 1895, foi organizado em *New York* o Congresso Americano de Bowling.

(4.12) A *IBM Research*, com o seu quartel general em Yorktown Heights, lidera o ranking das publicações americanas na indústria.

A partir das frases (4.11) e (4.12), obtêm-se as seguintes relações:

Congresso Americano de Bowling `ocorre_em` New York

Yorktown Heights `sede_de` IBM Research

[6] Embora a designação `ocorre_em` seja mais apropriada em português para acontecimentos do que organizações, optamos por ter apenas um nome de relação, visto que a diferença é visível por meio da categoria a que pertence a entidade relacionada. Leia-se portanto `localizado_em` quando a relação é entre uma `ORGANIZACAO` e um `LOCAL`.

4.1.4 Relação outra e outras relações

A relação `outra`, assim como a categoria `OUTRO` no Segundo HAREM, permitiu estabelecer relações não contempladas no elenco de relações do ReRelEM (já caracterizadas neste capítulo), mas que nos pareceram relevantes e que, por isso, deveriam ser identificadas. É importante salientar, contudo, que a relação `outra` tem de ser linguisticamente motivada, ficando de fora, por exemplo, uma eventual relação de co-ocorrência de EM no mesmo texto ou no mesmo parágrafo.

Ainda assim, decidir o que deve ser ou não anotado como `outra` é uma tarefa altamente subjetiva, e que esbarra inevitavelmente na discussão sobre os limites entre conhecimento lingüístico, conhecimento enciclopédico e conhecimento de mundo, e mesmo sobre a possibilidade de tais distinções (Peeters, 2000). Esbarra, ainda, na própria noção de relevância, que, como já dissemos, é dependente do contexto.

Atente-se no seguinte excerto:

(4.13) Depois de ser exibida no Rio, chega a São Paulo a mostra Carmen Miranda Para Sempre, que será inaugurada hoje para convidados e amanhã para o público no Memorial da **América Latina**[7]. Fotos, roupas, objetos, são mais de 700 peças reunidas para contar a história da "**Pequena Notável**" ou a **Brazilian Bombshell**- não há no mundo quem não conheça essa genial estrela que conquistou o **Brasil**, a **Broadway** e **Hollywood**.

A mostra tem percurso cronológico e está dividida em núcleos. Inicia com o nascimento em Portugal e inclui imagens de sua família. Depois, vem a fase brasileira (...).Era uma "mulher art déco dos anos 30", que usava calças, ternos e vestidos belos – em particular, há uma sala especial com retratos da artista feitos em 1931, em **Buenos Aires**, pela alemã Annemarie Heinrich.

Neste trecho, por exemplo, seria possível (ou desejável) relacionar os locais *América Latina* e *Buenos Aires*? Seria possível (ou desejável) relacionar *Pequena Notável* ou *Brazilian Bombshell*, por um lado, e *Brasil*, *Broadway* ou *Hollywood*, por outro lado, por meio de alguma relação como `conhecida em`?[8]

Conforme dissemos anteriormente, para que uma dada relação seja considerada, deve ser suficientemente informativa, por um lado, e capaz de permitir generalizações, por outro. Deste modo, a relação entre *Pequena Notável* (ou *Brazilian Bombshell*) e *Brasil* (ou *Broadway* ou *Hollywood*) não foi marcada, por nos parecer uma relação pouco produtiva, pelo menos nos textos analisados.

A relação de inclusão entre *América Latina* e *Buenos Aires*, por sua vez, embora irrelevante – nesse contexto – para a compreensão do texto (não há diferença se os retratos foram feitos em Buenos Aires ou, por exemplo, na Nova Zelândia), ou, dito de outra ma-

[7] A EM *América Latina* é uma análise alternativa à segmentação *Memorial da América Latina*, em que a entidade pode ser segmentada em *Memorial* e *América Latina*, conforme descrito no capítulo 1.

[8] Há, obviamente, outras relações que foram estabelecidas entre as EM desse trecho, e que podem ser consultadas na CD do ReRelEM, mas que omitimos aqui por uma questão de simplicidade na exposição.

neira, ainda que não seja uma relação que esteja no texto[9], deve ser marcada, e esperamos que seja reconhecida pelos sistemas.

Sob um outro ângulo, podem existir necessidades de informação tão excêntricas que a relação entre *Buenos Aires* e *América Latina* pudesse ser útil. Atente-se na frase abaixo:

(4.14) Visitei uma exposição de cavalos, no Peru, e vi raças que só conhecia de fotografia: Falabella, Hunter, Berbere, Andaluz e Paso.

Um leitor especialista em cavalos poderia ver como relevante uma relação `origem_de` entre *Paso* e *Peru*, uma vez que Paso é uma raça de origem peruana. No entanto, não existe forma de inferir essa relação a partir do texto, nem o estabelecimento dessa relação é importante para a compreensão do mesmo. Porém, do ponto de vista de uma aplicação de recolha de informação, é possível imaginar pessoas interessadas em pesquisar textos sobre exposições de cavalos em que alguma das raças fosse característica da região onde a exposição foi realizada.

Assim, a fim de compatibilizar uma anotação linguisticamente (e humanamente) motivada com as possíveis capacidades e interesses dos sistemas, optamos por marcar todas as relações – desde que estivessem contempladas nas directivas – distinguindo com a indicação `INDEP`[10] as que não podem ser inferidas mediante a interpretação do texto (como as relações acima mencionadas entre *Buenos Aires* e *América Latina*, ou entre *Paso* e *Peru*).

Por fim, durante a anotação, distinguimos ainda as relações que apenas acontecerão no futuro (dado que essa relação, de acordo com a informação do texto, ainda não aconteceu) com a indicação `FUTURO`[11].

Uma vez estabelecido de forma genérica o que deveria ser anotado como `outra`, a sua análise posterior permitiu aos anotadores examinar, com maior detalhe e com mais tempo, o tipo de relações abrangidas por essa relação, apontando casos gerais, produtivos, e/ou interessantes.

De fato, essa análise mais fina das relações `outra` levou a um total de 22 sub-categorias, que usamos na anotação das relações na CD do ReRelEM, a saber: natural de, povo de, residente de, vínculo institucional, relação profissional, relação familiar, autor de, produtor de, proprietário de, datado de, causa de, outra edição, representante de, praticado em, participante em, nome de, data de nascimento, data da morte, período de vida, personagem de, localizada em, e outra relação. Embora a especificação de tais categorias não tenha sido alvo de avaliação do ReRelEM (visto que ocorreu posteriormente à definição da tarefa), permitiu criar um recurso semântico mais rico e informativo para servir de base a outros estudos e aplicações futuras (cf. tabela 4.4, na secção 4.3, que lista a distribuição dos 156 casos de relações previamente classificadas como `outra`, indicando também a que categorias se podem aplicar).

Embora algumas relações sejam pouco freqüentes na CD do ReRelEM, nos pareceram potencialmente produtivas, com possibilidades de ocorrência em outros textos. Por isso, decidimos mantê-las na CD do ReRelEM.

---

[9] Essa opção pode, à primeira vista, parecer incoerente com o que já afirmamos sobre a dependência entre o contexto e o estabelecimento de uma relação. Lembramos, mais uma vez, que a informação contextual diz respeito à classificação das EM, tarefa anterior ao estabelecimento das relações semânticas.

[10] `INDEP` corresponde a "conhecimento independente" e é anotado no campo específico da CD do ReRelEM para comentários. Só foram marcados seis casos na CD do ReRelEM.

[11] A marcação é anotada no campo específico da CD do ReRelEM para comentários. Só foram anotados sete casos na CD do ReRelEM. Relações marcadas desta forma não foram contabilizadas como relações diferentes na tabela 4.4.

Cabem assim algumas notas sobre algumas destas relações:

- A relação `autoria` também compreende, por exemplo, um diretor[12] do filme e o filme.
- Embora as relações `autoria` e `produzido_por` sejam próximas (talvez a distinção esteja mais na dimensão intelectual embutida na noção de autoria), preferimos, por ora, manter a separação. E, embora a relação `produtor_de` não tenha aparecido nos documentos da CD do ReRelEM, esteve presente nos documentos analisados anteriormente.
- As relações que envolvem a categoria `ABSTRACCAO NOME` são mais especificadas que as demais relações. Como uma relação do tipo `nome_de`/`nomeado_por` é pouco informativa, pois explicita apenas que uma dada EM é lexicalizada de uma determinada maneira, optamos por refinar ainda mais a informação, especificando a relação existente entre as entidades envolvidas além do nome. Por exemplo, em (4.15), a informação de que a EM *Portugal* (`ABSTRACCAO NOME`) nomeia (e, portanto, é `nome_de`) a *Selecção* (uma EM do tipo `GRUPOMEMBRO`), pode ser enriquecida se indicarmos, neste caso, que há uma relação de identidade subjacente ao uso do nome. Por isso, neste exemplo, a relação é anotada `nome_de_ident`[13].

  (4.15) *SELECÇÃO* DE REGRESSO APÓS BOA PRESTAÇÃO NO MUNDIAL ... a maioria dos adeptos a gritar o nome de *Portugal* de forma entusiasmada.

- Embora não tenhamos encontrado nenhuma ocorrência da relação `data_nascimento`, entre entidades `PESSOA` e `TEMPO`, na CD do ReRelEM, nos parece produtiva, principalmente se considerada em conjunto com a relação `data_morte`.
- O mesmo se passa com a relação `localizado_em/localizacao_de`, para relacionar obras e os locais onde se encontram (p. ex., a *Mona Lisa* está *no Louvre*), e que apenas ocorre uma vez na CD do ReRelEM.

4.2 Relações do ReRelEM: como anotar

Além dos atributos do HAREM clássico (que aliás são todos opcionais, exceto o `ID`), no ReRelEM foram usados mais dois atributos: `COREL` e `TIPOREL`. O valor do primeiro é preenchido com um ou mais identificadores (ID), correspondentes à(s) entidade(s) com que a EM anotada se relaciona; o segundo é preenchido com um ou mais tipos (tantos quanto o número de `ID` usados em `COREL`) que especificam o tipo de relação em questão.

(4.16) Um dos telescópios já está pronto e em funcionamento no <EM ID="a1" CATEG="LOCAL">**Havaí**</EM>, <EM ID="a3" COREL="a1" TIPOREL="inclui">**EUA**</EM>

Na frase (4.16), `COREL="a1"` indica que a EM em causa (*EUA*) se relaciona com a EM cujo ID é `a1` (isto é, *Havaí*), através da relação de `TIPOREL="incluido"`. A informação codificada

[12] realizador, em português de Portugal
[13] Na tabela 4.4 esta relação foi contabilizada como `nome_de`.

pode ser lida da seguinte maneira: *EUA inclui Havaí*, ou, por simetria, *Havaí incluído em EUA* (ver secção 4.2.4).

Note-se que o valor de COREL pode ser preenchido com o ID de uma entidade que ainda não foi mencionada no texto, desde que essa entidade exista. Isso permite que os sistemas possam analisar e anotar os textos da forma que acharem mais conveniente, segundo qualquer tipo de algoritmo.

4.2.1 Relações múltiplas entre EM

É naturalmente possível que uma dada EM possua relações diferentes com mais de uma EM. Nesses casos, anotamos as diferentes relações em uma estrutura de lista, ou seja, tanto o valor de COREL como o de TIPOREL são preenchidos com uma sequência de identificadores e de tipos de relação, respectivamente, separados por espaços. As correspondências entre os atributos de TIPOREL e COREL estabelecem-se em função da ordem em que estão especificadas, sendo esta ordenação uma exigência.

(4.17) depois de partir em vantagem pontual no <EM ID="b13" CATEG="ACONTECIMENTO" TIPO="ORGANIZADO" COREL="b3 b5 b11" TIPOREL="ident ident ocorre_em">**Campeonato do Mundo**</EM>

No exemplo (4.17), a EM cujo ID é b13 (*Campeonato do Mundo*) está relacionada com as entidades:

b3, e a relação é do tipo ident;

b5, e a relação é do tipo ident;

b11, e a relação é do tipo ocorre_em.

4.2.2 ReRelEM e análises alternativas (ALT)

Não anotamos relações entre EM que se encontrem em alternativa dentro do mesmo ALT.

(4.18) <ALT> <EM ID="hub-94570-118" CATEG="LOCAL|ORGANIZACAO" TIPO="HUMANO|INSTITUICAO" SUBTIPO="CONSTRUCAO|">**Universidade de Lisboa**</EM>
|
<EM ID="hub-94570-118-aa" CATEG="LOCAL|ORGANIZACAO" TIPO="HUMANO|INSTITUICAO" SUBTIPO="CONSTRUCAO|">**Universidade**</EM> de <EM ID="hub-94570-131" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="DIVISAO" COREL="hub-94570-118-aa" TIPOREL="outra">**Lisboa**</EM>
<|ALT>

Por exemplo, como se vê pela anotação da sequência *Universidade de Lisboa* (exemplo (4.18)), não existe qualquer relação entre *Universidade de Lisboa* e *Universidade* (ou *Lisboa*), dado que não se trata efectivamente de duas entidades distintas no documento, mas tão só de duas formas diferentes de representar a mesma entidade.

4.2.3 ReRelEM e a vagueza do HAREM

Uma das características mais interessantes do HAREM é o tratamento que se dá à vagueza: o fato de uma mesma EM representar, em um mesmo contexto, mais do que uma das classes semânticas pré-definidas no modelo de classificação (ver capítulo 1). Na frase (4.19), *Portugal* pode ser simultaneamente entendido como uma organização e um local:

(4.19) Expressando ainda a "honra" por *Portugal* ficar associado a "uma importante etapa da cidadania europeia" – foi durante a *Presidência*, em 2000, que se iniciou a...

Nesses casos, que correspondem a cerca de 10% das entidades da CD do ReRelEM, consideramos que a co-relação se pode estabelecer entre as diferentes facetas de uma EM, ou apenas entre algumas delas. Isto é, embora em um dado contexto uma EM possa ser vaga entre duas ou mais leituras, nada impede que, no decorrer do texto, quando referida por outra EM, tenha o seu significado refinado, levando a que apenas uma das suas facetas esteja envolvida na relação.

Por exemplo, em (4.19), embora a EM *Portugal* seja vaga entre as categorias `ORGANIZACAO` e `LOCAL`, a EM *Presidência* (anotada como `ACONTECIMENTO`) estabelece uma relação com *Portugal* relativa apenas à faceta `LOCAL`, e portanto refina na relação o significado de *Portugal* mencionado anteriormente.[14]

Tendo em conta estas considerações, optamos por explicitar as relações não apenas entre EM, mas também entre facetas de EM no caso de EM vagas. Para tal, adoptamos um tipo de anotação ligeiramente diferente do inicialmente proposto, a fim de diferenciar as relações entre EM não vagas das relações que envolvem vagueza. Em particular, essa anotação passa por explicitar no campo `TIPOREL` não apenas o nome da relação, como também as facetas (categorias) das EM participantes. Temos, portanto, a seguinte anotação para o trecho já referido:

(4.20) Expressando ainda a "honra" por <EM ID="a97" CATEG="ORGANIZACAO|LOCAL" TIPO="ADMINISTRACAO|HUMANO" SUBTIPO="|PAIS">**Portugal**</EM> ficar associado a "uma importante etapa da cidadania europeia" – foi durante a <EM ID="a98" CATEG="ACONTECIMENTO" TIPO="ORGANIZADO" COREL="a97" TIPOREL="ACONTECIMENTO**ocorre_em**a97**LOCAL">**Presidência**</EM>

Com a especificação das relações entre categorias vagas, explicitamos também todas as relações que possam existir (na CD do ReRelEM) entre EM expressas por o mesmo item lexical, mas com referentes distintos. Ou seja, nada impede que uma EM *União Europeia* (`LOCAL`) seja sede de *União Europeia* (`ORGANIZACAO`).

4.2.4 Simetria, inversão e transitividade

Algumas das relações que apresentamos possuem determinadas propriedades, em particular, simetria, existência de relação inversa e transitividade, o que leva a que não seja necessário anotar exaustivamente todas as relações que existem no texto.

[14] Como a Renata Vieira referiu, a *Presidência*, fora de contexto, também podia ser considerada como uma organização, entrando pois em relação com a faceta `ORGANIZACAO` de *Portugal*. Contudo, não foi essa a leitura que as anotadoraas da CD do ReRelEM fizeram neste caso, quando concluiram que o contexto de *durante* força a leitura única de `ACONTECIMENTO`.

Tabela 4.1: Regras de expansão

A `ident` B e B `ident` C	⇒ A `ident` C
A `inclui` B e B `inclui` C	⇒ A `inclui` C
A `inclui` B e B `sede_de` C	⇒ A `sede_de` C
A `ident` B e B `qualquer_relação` C	⇒ A `qualquer_relação` C

Tal como referimos anteriormente, a relação de identidade é simétrica, ou seja, se a entidade *A* é a mesma que a entidade *B*, então também existe uma relação de identidade entre *B* e *A*. O que significa que, desde que os nossos programas sejam inteligentes, apenas é necessário anotar uma das entidades com a relação `ident`. Da mesma forma (como apontado por Vilain et al. (1995)), se existirem quatro EM com o mesmo referente, basta especificar três relações, e não doze.

Relativamente aos pares de relações `inclui`/`incluido` e `ocorre_em`/`sede_de`, como também já mencionamos, cada relação do par é a relação inversa da outra relação no mesmo par. Ou seja, se tivermos a relação *A* `inclui` *B*, então também podemos inferir que *B* está `incluido` em A.

Além disso, a relação de identidade e a de inclusão são transitivas. Quer isto dizer que, em uma relação de identidade, por exemplo, se tivermos que uma entidade *A* é idêntica a *B* e que *B* é idêntica a *C*, então também existe uma relação de identidade entre as entidades *A* e *C*.

Temos a conjugação de várias destas regras de forma a podermos concluir mais informação do que a que é necessário explicitar. A tabela 4.1 lista as regras utilizadas.

Isso leva a que possam existir dois textos anotados de maneira diferente, mas que codificam o mesmo conhecimento, ou, dito de outro modo, que são equivalentes depois de inferidas todas as relações por meio da explicitação das relações simétricas e inversas e através da aplicação de regras de expansão a essas relações.

Veja-se um exemplo de duas maneiras equivalentes de anotar a mesma frase:

(4.21) a. Em 9 de Setembro de 1895, foi organizado em <EM ID="15">**New York**</EM> o <EM ID="16" COREL="15" TIPOREL="ocorre_em">**Congresso Americano de Bowling**</EM> (“<EM ID="17" COREL="16 15" TIPOREL="ident ocorre_em">**ABC**</EM> – <EM ID="18" COREL="16 15" TIPOREL="ident ocorre_em">**American Bowling Congress**</EM>”), sediado em <EM ID="19" COREL="15 16 17 18" TIPOREL="incluido sede_de sede_de sede_de">**Milwaukee**</EM>, com o objetivo de aplicar medidas corretivas contra os excessos de jogatina e aperfeiçoar ainda mais as regras.

b. Em 9 de Setembro de 1895, foi organizado em <EM ID="15" COREL="19" TIPOREL="inclui">**New York**</EM> o <EM ID="16" COREL="15" TIPOREL="ocorre_em">**Congresso Americano de Bowling**</EM> (“<EM ID="17" COREL="16">**ABC**</EM> – <EM ID="18" COREL="16">**American Bowling Congress**</EM>”), sediado em <EM ID="19" COREL="16" TIPOREL="sede_de">**Milwaukee**</EM>, com o objetivo de aplicar medidas corretivas contra os excessos de jogatina e aperfeiçoar ainda mais as regras.

Salientamos que a não obrigatoriedade de anotar exaustivamente todas as relações se aplica tanto à anotação humana como à anotação feita pelos sistemas. Como veremos mais

adiante (cf. capítulo 5), durante o processo de avaliação existe um módulo responsável por expandir, ou seja, explicitar, todas as relações de acordo com as propriedades de simetria e transitividade.

4.3 A coleção dourada do ReRelEM

A CD do ReRelEM é um subconjunto da coleção dourada do Segundo HAREM. Por esse motivo, contém, além das informações referentes à classificação das entidades mencionadas[15], informação relativa às relações semânticas entre as EM. Esta informação é usada como termo de comparação para medir o desempenho dos sistemas no ReRelEM.

A anotação humana das relações foi feita com auxílio da ferramenta `Etiquet(H)AREM`, que permite a anotação dos atributos `COREL` e `TIPOREL` (veja-se o apêndice F para uma descrição detalhada da ferramenta).

A anotação dos textos desta CD decorreu em duas etapas principais. Numa primeira etapa, cada uma das anotadoras anotou uma parte dos textos da CD, tendo como base as relações-alvo definidas no ReRelEM (identidade, inclusão, localização e outra). Numa segunda etapa, os textos foram alternadamente anotados por cada uma das anotadoras, visando a especificação das categorias derivadas das relações `outra`. Tanto numa como noutra fase, os textos passaram por uma revisão cruzada, e os casos problemáticos ou duvidosos foram discutidos pela organização, de forma a encontrar uma solução de anotação consensual ou maioritária.

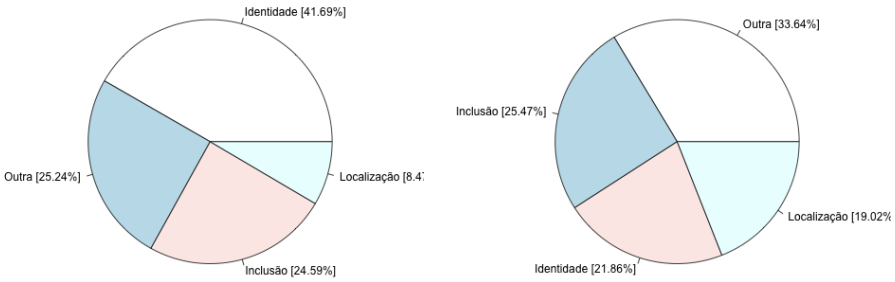
A CD do ReRelEM é composta por doze textos, 4417 palavras, 573 entidades mencionadas e 614 relações manualmente anotadas. Após a expansão das relações, tal como mencionado na secção anterior, a CD do ReRelEM passa a ter 6477 relações. A tabela 4.2 apresenta a distribuição das relações, antes e depois da expansão, e a figura 4.1 apresenta a mesma informação graficamente.

Tabela 4.2: Tipos de relação na coleção dourada do ReRelEM

Relação	Antes da expansão	Depois da expansão
identidade	256	1416
inclusão	151	1650
localização	52	1232
outra	155	2179
Total	614	6477

Como se pode constatar, a distribuição das relações não é idêntica antes e depois da expansão. Em particular, e embora a relação de localização seja a menos freqüente nos dois casos, existem proporcionalmente mais relações deste tipo depois da expansão do que antes. Além disso, na CD com as relações expandidas, a relação `outra` é a mais freqüente, e na CD antes da expansão a relação mais freqüente é a de identidade. Observa-se ainda que a relação de inclusão tem proporcionamente o mesmo número de relações nas duas versões da CD.

[15] De fato, a CD do ReRelEM é um subconjunto da CD do TEMPO, contendo igualmente informações referentes à normalização de expressões temporais



(a) Antes da expansão (b) Depois da expansão

Figura 4.1: Distribuição de relações

A tabela 4.3 apresenta a distribuição, nos doze textos da CD ReRelEM, do número de pares de relações[16] (por tipo de relação) assim como o número de (facetas de) EM envolvidas.

Tabela 4.3: Tipos de relação por documento

Documento	Identidade	Inclusão	Localização	Outra	Total	Facetas	Facetas em relações
aa56088	862	818	756	1378	3814	146	131
bob-14949	92	158	12	116	378	89	56
hub-21881	22	32	4	2	60	36	23
hub-41899	42	26	16	117	201	75	39
hub-49343	60	160	112	100	432	127	62
hub-66526	110	86	158	48	402	84	47
hub-71248	22	16	0	0	38	33	14
hub-78051	18	42	4	34	98	28	19
hub-94570	8	8	2	39	57	39	20
hub-96408	82	132	56	242	512	67	40
ric-54609	14	4	12	74	104	31	19
ric-92221	84	168	100	29	381	64	42
Total	1416	1650	1232	2179	6477	819	512

Podemos assim observar que os textos diferem muito em termos de densidade de EM e de relações entre elas. Quanto ao tipo de relações, na maioria dos textos a identidade é a mais frequente, mas noutros (três) a inclusão é mais comum, sendo que no texto mais relacionado é a outra relação a mais frequente.

[16] Por "*par de relação*" designamos a relação e a sua inversa.

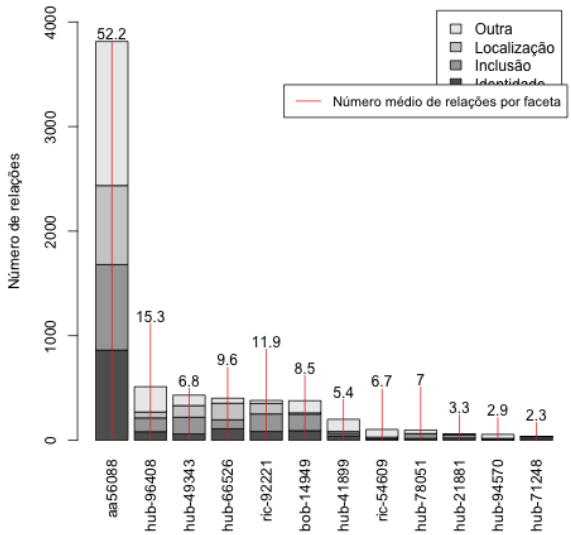


Figura 4.2: Distribuição de relações pelos documentos da CD do ReRelEM

A figura 4.2 mostra o total de relações por documento (distribuídas por tipo de relação) e o número médio de relações por faceta. Os documentos estão ordenados por ordem decrescente do total de relações, sendo possível ver que as entidades de documentos com mais relações não estão necessariamente mais envolvidas em média numa relação do que as entidades de documentos menores. Compare-se, por exemplo, os documentos `hub-49343` e `ric-54609`, em que o primeiro documento tem quatro vezes mais entidades do que o segundo, mas em média cada entidade participa em cerca de sete relações nos dois casos.

Para dar uma visão mais clara do que está envolvido na relação `outra` na CD do ReRelEM, a tabela 4.4 apresenta a distribuição por tipo de relação, antes e depois da expansão. Salientamos que, embora algumas relações sejam apresentadas como um par de relações, esta complementaridade não implica, necessariamente, simetria. Isto é, a relação número 11, `causa_de` / `consequencia_de`, por exemplo, é considerada um mesmo tipo de relação por veicular informação de natureza semelhante, mas esse agrupamento não significa que as relações envolvidas sejam simétricas. No exemplo (4.22), embora seja possível estabelecer uma relação de consequência entre as EM *Carta* e *Convenção*, não nos parece natural, a partir da leitura do texto, uma relação de causa entre *Convenção* e *Carta*.

(4.22) (...) foi durante a Presidência, em 2000, que se iniciou a *Convenção* que deu origem à *Carta*.

Estamos conscientes de que esta é uma questão que merece um tratamento mais aprofundado. Deixamos para discussão futura a validade desta tipologia de relações, assim como a pertinência de definir (ou explicitar) a inversa de uma relação do ReRelEM.

Tabela 4.4: Subdivisão das relações `outra`

	Relação	Categorias a que se aplica	Anotadas	Após expansão
1	natural_de /	PESSOA e LOCAL	5	48
	local_nascimento_de		11	48
2	povo_de /	PESSOA POVO e LOCAL	5	34
	local_de		5	35
3	residente_de /	PESSOA e LOCAL	1	15
	residencia_de		3	15
4	vinculo_inst	PESSOA e ORGANIZACAO	42	783
5	relacao_profissional	PESSOA e PESSOA	7	106
6	relacao_familiar	PESSOA e PESSOA	17	90
7	autor_de /	PESSOA e OBRA	3	300
	obra_de		3	300
8	produtor_de /	PESSOA ou ORGANIZACAO e COISA	0	0
	produzido_por		0	0
9	proprietario_de /	PESSOA ou ORGANIZACAO e COISA ou ORGANIZACAO	1	10
	propriedade_de		2	10
10	datado_de /	OBRA ou ACONTECIMENTO e TEMPO	0	0
	data_de		6	78
11	causa_de /	ACONTECIMENTO e ACONTECIMENTO	0	0
	consequencia_de		1	17
12	outra_edicao	ACONTECIMENTO ORGANIZADOe ACONTECIMENTO ORGANIZADO	1	2
13	representante_de /	PESSOA e DISCIPLINA ou LOCAL ou COISA	6	13
	representado_por		2	7
14	praticado_em /	DISCIPLINA ou COISA e LOCAL ou ACONTECIMENTO	1	3
	pratica_se		2	3
15	participante_em /	PESSOA e OBRA ou EVENTO	12	113
	ter_participacao_de		7	113
16	nome_de /	ABSTRACCAO NOME e qualquer CATEG	1	4
	nomeado_por		0	0
17	data_nascimento	PESSOA e TEMPO	0	0
18	data_morte	PESSOA e TEMPO	1	1
19	periodo_vida	PESSOA e TEMPO	2	11
20	personagem_de	PESSOA e OBRA	4	12
21	localizado_em /	OBRA e LOCAL	1	1
	localização_de		0	0
22	outrarel	Todas	4	7

4.4 Avaliação

Nesta secção, descrevemos brevemente os aspectos gerais do processo de avaliação do ReRelEM, que estão detalhados no capítulo 5. Em seguida, destacamos os sistemas participantes nesta pista e por fim mostramos os resultados obtidos pelos sistemas, ou seja, o seu desempenho no ReRelEM.

Tabela 4.5: Sistemas participantes no ReRelEM e dados de participação

Sistema	Cenários selectivos do HAREM clássico	Cenários do ReRelEM	N. de corridas
REMBRANDT	Total	Total	3
SEI-Geo	Só LOCAL (Sel5)	Inclusão	4
SeRELeP	Total (Identificação)	Todas menos `outra`	2

4.4.1 Processo de avaliação

Na avaliação do ReRelEM, é importante separar a avaliação da identificação e classificação de relações da tarefa de classificação de EM, objecto de avaliação do HAREM clássico. Ou seja, uma das nossas preocupações esteve em não penalizar duplamente uma participação.

Assim, é retirado da avaliação do ReRelEM aquilo que já foi considerado erro no HAREM clássico: são retiradas as EM que não foram identificadas e as que foram mal classificadas, bem como as relações em que estas participam.

Simplificadamente, durante a avaliação do ReRelEM, é preciso que as corridas dos sistemas sejam alinhadas com a CD do ReRelEM, para que sejam comparadas.

O passo seguinte é a explicitação (ou expansão) das relações, nomeadamente das relações de identidade, das relações inversas e das relações decorrentes da aplicação das regras de transitividade.

Visto que a CD, devido à análise em facetas, possui uma anotação mais detalhada (e portanto ligeiramente diferente) que as corridas dos participantes, foi preciso converter esta anotação para um formato pseudo-facetas e adicionar à comparação dos alinhamentos a questão da compatibilidade entre facetas.

Só depois se aplicam os véus para o ReRelEM, para considerar o caso de os participantes estarem apenas a marcar um subconjunto das relações na CD.

Finalmente, as relações da participação são avaliadas, por meio de uma comparação com as relações da CD. O resultado da comparação é um conjunto de relações corretas, espúrias ou em falta.

Embora tenhamos apresentado, por ocasião dos resultados oficiais, os resultados de acordo com três medidas diferentes, consideramos agora que a única medida que faz sentido é aquela em que tanto os argumentos como o tipo de relação estão corretos, chamada **avaliação de relações**. Ou seja, parece-nos que um sistema que marca uma relação de localização entre A e B quando a relação correta entre A e B é a de identidade não merece qualquer valorização adicional e que portanto não faz sentido a anteriormente denominada **avaliação de COREL**[17].

4.4.2 Sistemas participantes

Três sistemas, dos dez participantes no HAREM clássico, participaram na pista do ReRelEM. A tabela 4.5 mostra os participantes no ReRelEM com alguns dados sobre a respectiva participação.

Como se pode ver na tabela, para além de um dos sistemas ter participado no HAREM clássico num cenário seletivo diferente do dos outros dois sistemas, os três sistemas par-

[17] Esta avaliação premiaria sistemas que tivessem marcado uma relação entre A e B, mesmo que o tipo da relação não estivesse correto. Essa relação teria, em todo o caso, uma valorização inferior à atribuída se o tipo de relação estivesse correto.

ticiparam de formas distintas no ReRelEM. Isso levou a que também se criassem cenários seletivos para as relações do ReRelEM, como mencionado na secção anterior.

4.4.3 Resultados

Começamos por apresentar na figura 4.3 os resultados de desempenho dos sistemas no cenário total, tomando em conta todas as relações anotadas na CD do ReRelEM. Em todo o caso, salientamos que os sistemas, mesmo quando avaliados no cenário com todas as relações, acabam por ser classificados em função de sub-conjuntos diferentes de relações. Isto acontece porque apenas são avaliadas as relações cujas entidades participantes estão bem classificadas.

Como se pode observar, os resultados dos sistemas ainda estão muito aquém do que seria desejável: a melhor corrida, a corrida 1 do sistema REMBRANDT, obteve apenas 0,45 de medida F, enquanto a média dos vários sistemas se situou em 0,29. Relembramos, no entanto, quão complexa é a tarefa e o fato de se tratar de uma tarefa piloto.

Vê-se igualmente que o sistema SEI-Geo tem uma precisão muito alta em três das suas quatro corridas (pelo menos 0,91), mas por outro lado teve uma abrangência muito baixa (inferior a 0,16). Os outros dois sistemas mostram um maior equilíbrio entre abrangência e precisão, embora o sistema REMBRANDT (com excepção da sua melhor corrida) tenha mais abrangência (cerca de 0,4) do que precisão (abaixo de 0,27) e o SeRELeP se encontre na situação inversa (abrangência e precisão acima de 0,26 e 0,46, respectivamente).

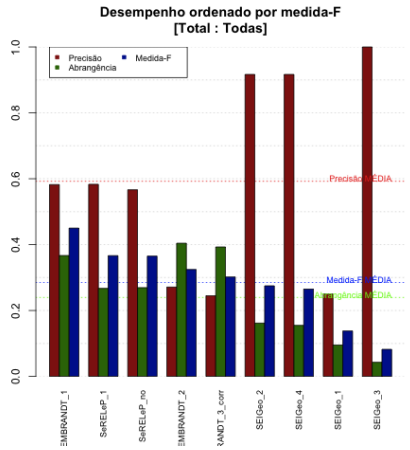
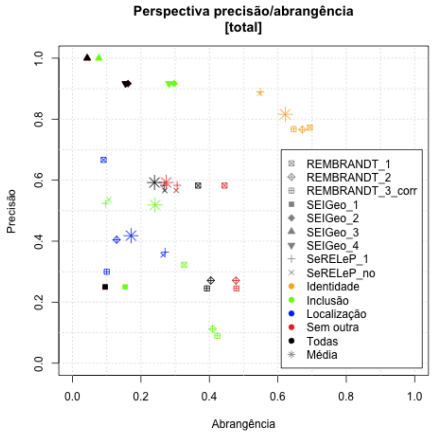


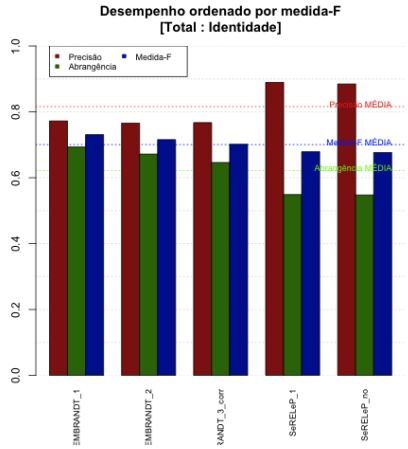
Figura 4.3: Avaliação de todas as relações no cenário total

Na figura 4.4 mostramos os resultados da avaliação nos cenários seletivos do ReRelEM, ou seja, usando um subconjunto das relações anotadas na coleção dourada.

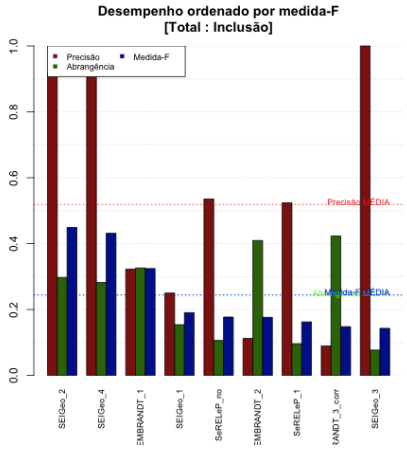
No primeiro gráfico dessa figura (4.4(a)) são comparados os vários cenários do ReRelEM em termos de precisão e abrangência: todas as relações (cenário *todas*), todas as relações menos a relação `outra` (cenário *Sem outra*), só relações de identidade (cenário *Identidade*), só relações de inclusão (cenário *Inclusão*) e só relações de localização (cenário *Localização*).



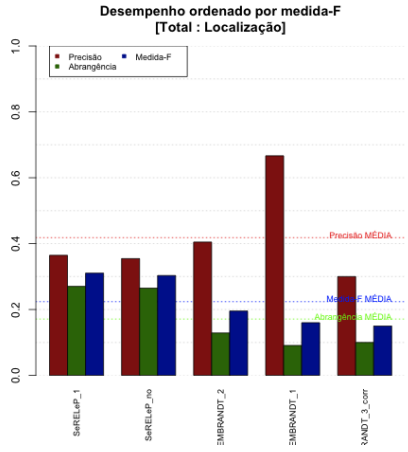
(a) Todos os cenários do ReRelEM



(b) Cenário do ReRelEM: Identidade



(c) Cenário do ReRelEM: Inclusão



(d) Cenário do ReRelEM: Localização

Figura 4.4: Avaliação nos cenários selectivos do ReRelEM

Como seria de esperar, quando não se considera a relação `outra`, os sistemas REMBRANDT e SeRELeP aumentam a sua abrangência (repare-se no deslocamento para a direita dos valores de abrangência desses sistemas, sem que a precisão seja afectada), porque excluindo as relações `outra` o número de relações que o sistema tem de reconhecer é menor. Já o desempenho do sistema SEI-Geo, pelo contrário, não se altera. Essa manutenção nos resultados do SEI-Geo é um efeito do processo de avaliação do ReRelEM: como todas as relações que contêm EM espúrias ou mal classificadas são desconsideradas da avaliação, e o SEI-Geo só identificou EM classificadas como `LOCAL`, são selecionados apenas os alinhamentos que envolvem entidades que sejam `LOCAL` e que estejam bem classificadas, o que acaba por, naturalmente, excluir as relações `outra`. Com isso, para o SEI-Geo, a alteração nos cenários de avaliação não faz diferença.

Outro factor que se destaca no mesmo gráfico é o desempenho dos sistemas ser significativamente melhor no reconhecimento da relação de identidade do que no das outras duas relações: o sistema REMBRANDT obteve valores de abrangência entre 0,65 e 0,69, para uma precisão de cerca de 0,77, e o sistema SeRELeP obteve 0,55 e 0,89 para as mesmas métricas, no reconhecimento da identidade.

No caso das outras relações, os resultados foram mais baixos e também mais variáveis, e em média os sistemas obtiveram um pior desempenho no reconhecimento da relação de localização (0,17 de abrangência média e 0,42 de precisão média), do que no da relação de inclusão (0,24 de abrangência média e 0,51 de precisão média).

Os gráficos 4.4(b), 4.4(c) e 4.4(d) mostram outra perspectiva dos valores de precisão e de abrangência dos cenários identidade, inclusão e localização que se encontram no gráfico 4.4(a), juntamente com os valores de medida F. Destaca-se que:

- o sistema REMBRANDT obteve o melhor desempenho em termos de medida F de todas as relações incluindo ou não a relação `outra`, e, em particular, no reconhecimento da relação de identidade com um valor de cerca de 0,73;
- o sistema SEI-Geo foi o melhor sistema a reconhecer relações de inclusão, com uma medida F ligeiramente abaixo de 0,45;
- o sistema SeRELeP foi o melhor a reconhecer relações de localização, com uma medida F perto de 0,31.

Embora seja naturalmente cedo para tirar conclusões, estes valores sugerem que as relações mais difíceis de identificar parecem ser as de localização.

4.5 Considerações finais

Apresentamos aqui o ReRelEM, uma pista piloto criada no Segundo HAREM cujo objetivo é a identificação de relações semânticas entre entidades mencionadas. Assim como no HAREM, a escolha das relações semânticas foi feita a partir da análise de textos, e como bem observou a Cláudia Oliveira, mesmo sem partir de relações pré-definidas, algumas categorias tradicionais, como sinonímia, hiperonímia e meronímia, são capturadas pelas relações de identidade e algumas ocorrências das relações de inclusão. Nesse sentido, um desdobramento interessante seria a comparação entre relações lexicais entre sintagmas nominais e entre EM.

De fato, como pista piloto, temos a sensação de que muito mais estaria por fazer: analisar mais textos, o que certamente leva a relações mais equilibradas ou generalizáveis (quanto mais textos, mais relações e, quanto mais relações, mais possibilidades de generalização) e, principalmente, possibilita validar as opções tomadas; investigar outras formas de avaliação; anotar com ainda mais precisão e segurança, visto que uma versão final das directivas de anotação só se concretizou com o fim do processo de anotação.

Com o ReRelEM, damos mais um passo no sentido não apenas de alavancar a área de REM para a língua portuguesa, mas talvez de REM em qualquer língua, visto ser essa uma tarefa, ao que saibamos, inovadora na forma como foi definida. Além disso, como resultado final, este piloto já oferece um material de grande valor: a própria CD do ReRelEM, disponível, anotada por linguistas, bem como os programas de avaliação, especificamente desenvolvidos para este efeito, e que esperamos que sejam úteis em muitas outras tarefas relacionadas com a detecção e estudo de relações semânticas em texto em português.

Agradecimentos

Agradecemos a Cláudia Oliveira, Renata Vieira e Violeta Quental pelos valiosos comentários e sugestões.

Capítulo 5

Avaliação à medida no Segundo HAREM

Hugo Gonçalo Oliveira, Cristina Mota, Cláudia Freitas, Diana Santos e Paula Carvalho

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 5, p. 97–129.

Um dos desafios que se coloca na realização de uma avaliação conjunta é o de definir uma forma de avaliar cada participação, por um lado, e o de comparar o seu desempenho com o desempenho das demais participações, pelo outro.

Esse desafio é ainda maior se, por um lado, a avaliação incluir várias pistas com objectivos distintos, e, por outro, existirem sistemas muito diferentes entre si, cada um com uma finalidade específica, por vezes até peculiar, como aconteceu no Segundo HAREM. Basta referir que, em dez participantes, não houve sequer dois que executassem exactamente a mesma tarefa.

A característica que mais se evidencia na avaliação do Segundo HAREM é então a possibilidade de poder comparar as várias participações com base em diferentes vistas sobre a mesma colecção dourada (CD), a que chamamos cenários (selectivos). Essa flexibilidade vem aliás do Primeiro HAREM, tendo sido aperfeiçoada na actual avaliação e aplicada às várias pistas do Segundo HAREM. De facto, como veremos, a avaliação do Primeiro HAREM foi o ponto de partida para a avaliação do Segundo HAREM, muito particularmente no que respeita ao HAREM clássico. Naturalmente, as novas pistas obrigaram ao desenvolvimento de novos programas de avaliação, e alguns programas que adoptámos do Primeiro HAREM também sofreram várias alterações, em função das mudanças introduzidas nesta edição.

5.1 Avaliação do HAREM clássico

5.1.1 Pontuações

No Segundo HAREM, cada entidade mencionada (EM) pode receber uma de três pontuações, no que respeita à sua identificação: `Correcta`, `Em falta` e `Espúria`. Estas pontuações são igualmente utilizadas para pontuar a classificação de cada atributo das EM.

5.1.2 Uma única medida

No Primeiro HAREM existiam apenas dois níveis de classificação: categorias e tipos. De acordo com esses níveis, foi definido um conjunto de quatro medidas (plana, só tipos, só categorias e classificação semântica combinada - CSC) para a avaliação do reconhecimento de entidades mencionadas (ver Santos et al. (2007)). Essas medidas eram baseadas no pressuposto de que todas as EM estavam classificadas obrigatoriamente com categoria e tipo.

No Segundo HAREM, o esquema de anotação adoptado define uma hierarquia de quatro níveis (se tivermos em conta somente o HAREM clássico): identificação da entidade (delimitando-a simplesmente com as etiquetas EM), e preenchimento dos respectivos atributos categoria, tipo e subtipo. Além disso, não é obrigatória a marcação dos níveis mais baixos.

Assim, para esta avaliação, foi definida uma única medida, mais robusta e abrangente, que é aplicada a cada entidade correctamente identificada. A medida engloba os quatro níveis da hierarquia, possibilitando a atribuição de diferentes pesos a cada um desses níveis e ainda a penalização por classificações erradas.

Basta identificar correctamente a EM para se receber o valor 1. O valor total da medida é obtido somando esse valor às parcelas relativas aos níveis da classificação. A medida tem em conta várias questões:

$$1+\sum_{i=1}^{N}((1-\frac{1}{n_{cats}})\cdot cat_{certa_i}\cdot\alpha+(1-\frac{1}{n_{tipos}})\cdot tipo_{certo_i}\cdot\beta+(1-\frac{1}{n_{sub}})\cdot sub_{certo_i}\cdot\gamma)$$
$$-\sum_{i=0}^{M}(\frac{1}{n_{cats}}\cdot cat_{esp_i}\cdot\alpha+cat_{certa_i}\cdot\frac{1}{n_{tipos}}\cdot tipo_{esp_i}\cdot\beta+tipo_{certo_i}\cdot\frac{1}{n_{sub}}\cdot sub_{esp_i}\cdot\gamma) \quad (5.1)$$

$K_{certo_i} = \begin{cases} 1 & \text{se o atributo } K_i \text{ estiver correcto,} \\ 0 & \text{se } K_i \text{ estiver incorrecto ou omisso} \end{cases}$

$K_{esp_i} = \begin{cases} 1-K_{certo_i} & \text{se o atributo } K_i \text{ estiver preenchido} \\ 0 & \text{se } K_i \text{ estiver omisso} \end{cases}$

$K \in \{cat, tipo, sub\}$

N = número de diferentes classificações vagas na CD, de acordo com o cenário selectivo.

M = número de classificações espúrias na participação, de acordo com o cenário selectivo.

α, β, γ = parâmetros correspondentes aos pesos das categorias, tipos e subtipos.

Figura 5.1: Fórmula de avaliação no Segundo HAREM

- O preenchimento de cada atributo por parte de um sistema é sempre opcional.
- O peso dado a cada atributo é tanto maior quanto mais difícil for acertar o valor do atributo, ou seja, é proporcional ao número de possibilidades que existem para o preencher. Por outro lado, quanto à penalização aplica-se o critério inverso. Isto é, quanto mais difícil for acertar o preenchimento de um atributo, menor é a penalização por ter o atributo espúrio.
- O número de categorias (n_{cats}) pode variar de acordo com o cenário selectivo que se está a avaliar, sendo no cenário total do Segundo HAREM igual a 10[1] e na avaliação de apenas uma categoria igual a 1 (não existindo, neste caso, valorização adicional pela classificação). O número de tipos (n_{tipos}) e de subtipos (n_{sub}) pode variar, respectivamente, de acordo com o valor da categoria ou categoria e tipo a que se referem, por um lado, e ainda com o cenário selectivo escolhido, pelo outro.
- Cada classificação vaga de uma EM é pontuada de forma independente. Isto leva a que a única forma possível de obter o valor máximo seja o sistema ter classificado a EM vaga exactamente da mesma forma que esta se encontra na CD (isto é, sem classificações espúrias e com todas as diferentes classificações possíveis correctas).
- É feita a distinção entre não conseguir atribuir o valor a um atributo (quando esse atributo é omisso) e assumir que o valor desse atributo não está contemplado no conjunto de categorias, tipos e/ou subtipos propostos nas directivas (classificando-o como `OUTRO`). Considera-se assim que todos os tipos podem ter um `SUBTIPO="OUTRO"`. Desta forma, quando a EM que se está a avaliar pertence a uma categoria e tipo sem subtipo definido, a última parcela toma o valor 0.

[1] Recorda-se o elenco das dez categorias do Segundo HAREM: `ABSTRACCAO`, `ACONTECIMENTO`, `COISA`, `LOCAL`, `OBRA`, `ORGANIZACAO`, `PESSOA`, `TEMPO`, `VALOR` e `OUTRO`.

- A penalização resultante do preenchimento espúrio de um atributo nunca contém mais de uma parcela. Ou seja, para calcular a penalização é apenas contabilizada a parcela do atributo espúrio no nível mais alto da hierarquia. Por exemplo, se uma classificação tiver a categoria correcta e o tipo espúrio, o subtipo será obrigatoriamente também espúrio. No entanto, a penalização será dada apenas pelo tipo espúrio e não pelo subtipo.

Para a avaliação oficial do Segundo HAREM optou-se por dar maior importância à classificação da categoria do que à classificação do tipo, que por sua vez deveria ter mais importância do que a classificação do subtipo.

Assim, aos atributos categoria, tipo e subtipo foram atribuídos os pesos $\alpha = 1$, $\beta = 0{,}5$ e $\gamma = 0{,}25$, respectivamente. O módulo responsável pelo cálculo da medida permite contudo parametrizar não só o valor dos pesos mas também o valor dado à identificação (ver secção 5.6.6). Note-se, aliás, que nada impede que os valores α, β e γ sejam maiores que um[2]

5.1.3 Cenários selectivos

À semelhança do que aconteceu no Primeiro HAREM, os sistemas puderam participar num cenário selectivo, ou seja, puderam identificar e classificar apenas um subconjunto das categorias, tipos e subtipos que haviam sido definidos como alvo da avaliação. Assim, além de ser avaliado no cenário total, cada sistema foi avaliado no cenário selectivo que indicou.

O cenário selectivo, apesar de ter a vantagem de permitir avaliar um sistema no cenário que se propôs realizar, tem a limitação de não permitir comparar directamente sistemas que participem em cenários diferentes. Por isso, no Segundo HAREM, fomos mais longe e avaliámos cada participação não só no cenário total e no cenário que o participante propôs, mas também em função dos cenários selectivos propostos pelos outros participantes. Além disso, os sistemas foram ainda avaliados em cenários constituídos apenas por cada uma das categorias previstas nas directivas, como foi feito no Primeiro HAREM. Esta opção permitiu, portanto, uma comparação mais fina de cada sistema com os restantes.

Definimos, assim, os seguintes tipos de cenário:

- **Cenário de avaliação**: conjunto de categorias, tipos e subtipos em que se pretende avaliar uma participação.
- **Cenário de participação**: conjunto de categorias, tipos e subtipos que o sistema se propõe classificar.

Na figura 5.2, ilustra-se de forma muito genérica e simplificada os dois tipos de cenário, e no que consiste a avaliação nesses cenários. Como se pode ver, as categorias reconhecidas por uma dada participação constituem um cenário de participação (não se encontra ilustrado, mas o cenário de participação pode naturalmente ser coincidente com o cenário total). Cada um dos cenários de participação é então usado como cenário de avaliação, levando a que haja um processo de adaptação tanto da CD como das várias participações a esse cenário. Veja-se que na avaliação no cenário selectivo C1, a CD passa a ter menos entidades (todas as entidades cujas categorias não façam parte do cenário são removidas),

[2] A fórmula proposta não permite anular completamente o peso da identificação. Por isso, e por uma questão de gereralidade, estamos agora convencidos de que teria sido melhor também associar um peso independente à identificação.

e as entidades da categoria LOCAL deixam de ter o atributo TIPO; como a participação B só tem entidades da categoria LOCAL, a única adaptação consiste em fazer desaparecer o atributo TIPO. A mesma figura ilustra também que se pode ter como cenário de avaliação, um cenário selectivo (C3) diferente dos cenários de participação.

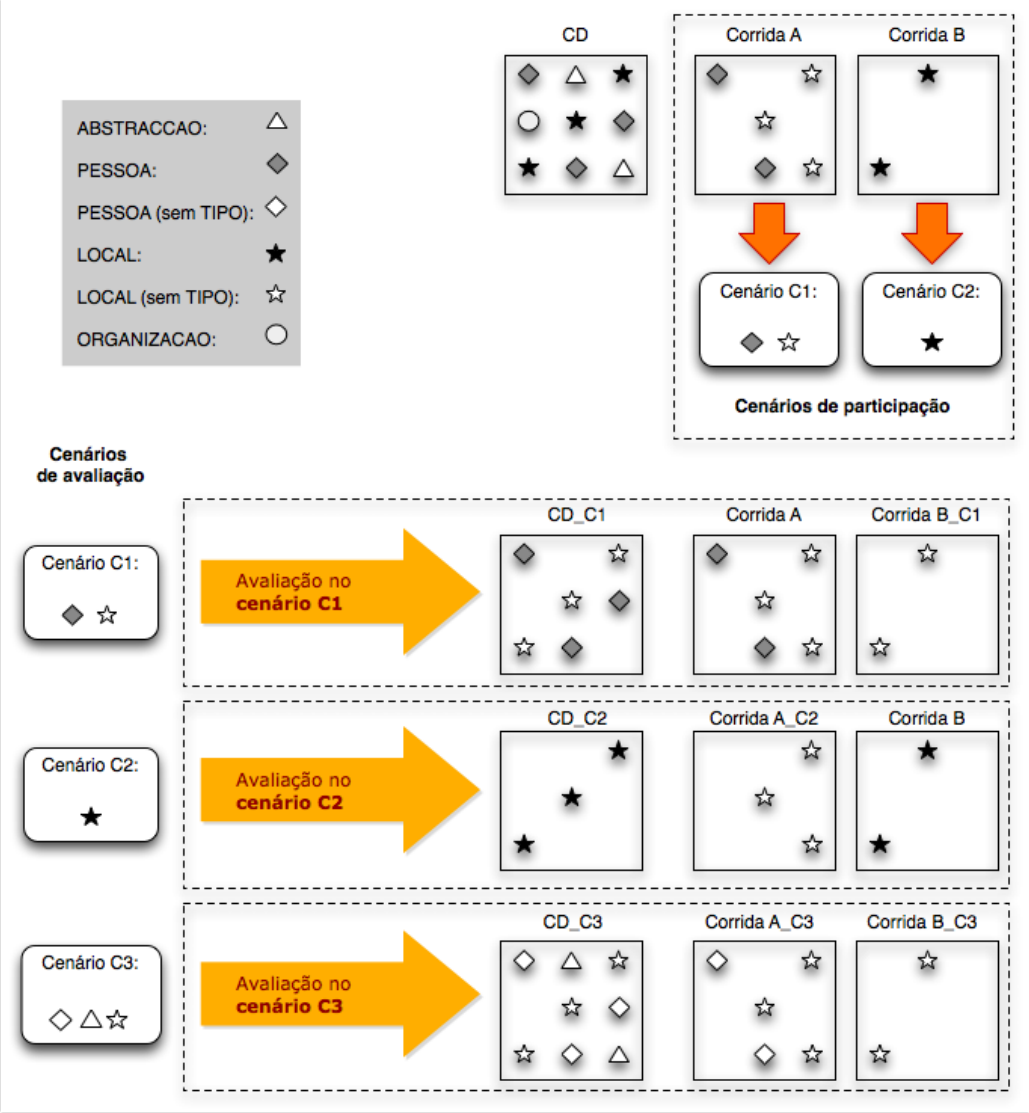


Figura 5.2: Exemplos de cenários de participação e avaliação

Note-se, a este respeito, que, na avaliação num dado cenário selectivo, por oposição ao que aconteceria se o cenário de avaliação fosse o cenário total, a diminuição do número de categorias, tipos ou subtipos irá reflectir-se na medida utilizada (ver secção 5.1.2).

Um problema que se nos pôs relativamente à avaliação por cenários selectivos foi a existência de várias participações com EM que apenas foram identificadas, ou seja, sem qualquer tipo de classificação. Depois de reflexão aturada, decidimos utilizar também o cenário de participação para, nestes casos, tomar a decisão de manter ou não a EM. Quando a EM não tem qualquer classificação assume-se que poderia ter qualquer uma das classificações possíveis dentro do cenário de participação. Sendo assim, a EM é removida apenas se o cenário de participação não tiver classificações comuns com o cenário de avaliação, ou seja, se a intersecção dos cenários for nula. Se isto não se verificar, a EM mantém-se. Alguns exemplos da utilização de ambos os tipos de cenário selectivo encontram-se mais à frente neste capítulo, na figura 5.9.

5.1.4 Avaliação de ALT

A etiqueta ALT foi utilizada para anotar todas as segmentações possíveis de EM (veja-se a secção respectiva do capítulo 1 para uma motivação linguística desta opção, assim como as regras sistemáticas de segmentação de ALT, no apêndice D).

Os participantes foram incentivados a utilizar esta etiqueta e a sua avaliação foi feita de duas formas:

- **Avaliação estrita de ALT**, onde são contabilizadas todas as alternativas possíveis para um segmento de texto, tendo cada alternativa um peso igual ao inverso do número de alternativas dentro desse segmento. Por exemplo, a todos os valores atribuídos a EM dentro de um ALT com três elementos será associado um peso de 1/3. O sistema só terá, assim, o valor máximo possível se tiver apresentado as três alternativas no seu resultado.

- **Avaliação relaxada de ALT**, onde é seleccionado o elemento do ALT que maximiza a classificação do sistema, tal como foi feito no Primeiro HAREM.

5.2 Avaliação da pista do TEMPO

Tal como referimos no capítulo 3, a avaliação da pista do TEMPO foi feita de forma integrada com o HAREM clássico. No entanto, foi necessário ter em conta que as entidades da categoria TEMPO têm atributos específicos, além dos atributos do HAREM clássico.

Na avaliação do TEMPO estendido tivemos por objectivo dar crédito adicional aos sistemas pelo correcto preenchimento dos atributos estendidos da categoria TEMPO, ou seja, TEMPO_REF, SENTIDO, VAL_NORM e VAL_DELTA (cf. capítulo 2), sem os penalizar pela atribuição de valores espúrios. Assim, a fórmula completa usada na avaliação das entidades com a categoria TEMPO encontra-se na figura 5.3. Repare-se que a parcela inicial 5.2 é a medida de avaliação do HAREM clássico apresentada na figura 5.1, sendo as restantes parcelas usadas para avaliar os atributos estendidos de TEMPO.

Destaca-se ainda que, enquanto os atributos VAL_NORM para o tipo DURACAO e VAL_DELTA são avaliados como um todo, não sendo valorizado o correcto preenchimento de cada um dos campos que compõe o valor desses atributos, a situação é diferente para o atributo VAL_NORM quando o tipo é DATA ou HORA. Neste caso, cada um dos campos individuais do atributo contribui separadamente para o valor da medida de classificação da EM. Exemplificando, no

$$1+\sum_{i=1}^{N}((1-\frac{1}{n_{cats}})\cdot cat_{certa_i}\cdot\alpha+(1-\frac{1}{n_{tipos}})\cdot tipo_{certo_i}\cdot\beta+(1-\frac{1}{n_{sub}})\cdot sub_{certo_i}\cdot\gamma)$$

$$-\sum_{i=0}^{M}(\frac{1}{n_{cats}}\cdot cat_{esp_i}\cdot\alpha+cat_{certa_i}\cdot\frac{1}{n_{tipos}}\cdot tipo_{esp_i}\cdot\beta+tipo_{certo_i}\cdot\frac{1}{n_{sub}}\cdot sub_{esp_i}\cdot\gamma) \quad (5.2)$$

$$+tr_{certo}\cdot\delta+s_{certo}\cdot\lambda \quad (5.3)$$

$$+\begin{cases} vd_{certo}\cdot\varepsilon \\ vn_{certo}\cdot\varepsilon \\ (E_{certa}+A_{certo}+D_{certo}+H_{certo}+M_{certo}+ES_{certa}+lim_{certo})\cdot\xi \\ (H_{certo}+M_{certo}+lim_{certo})\cdot\eta \end{cases} \quad (5.4)$$

$$K_{certo_i}=\begin{cases} 1 & \text{se o atributo } K_i \text{ estiver correcto,} \\ 0 & \text{se } K_i \text{ estiver incorrecto ou omisso} \end{cases}$$

$$K_{esp_i}=\begin{cases} 1-K_{certo_i} & \text{se o atributo } K_i \text{ estiver preenchido} \\ 0 & \text{se } K_i \text{ estiver omisso} \end{cases}$$

$K \in \{cat, tipo, sub\}$

N = número de diferentes classificações vagas na CD, de acordo com o cenário selectivo.

M = número de classificações espúrias na participação, de acordo com o cenário selectivo.

tr, s, vd, vn: classificação referente, respectivamente, aos atributos `TEMPO_REF`, `SENTIDO`, `VAL_DELTA`, e `VAL_NORM` (quando `TIPO="DURACAO"`).

E, A, D, H, M, ES, lim: classificação referente, respectivamente, aos campos era, ano, dia, mês, hora, minutos, estação e limite do atributo `VAL_NORM` (quando `TIPO="DATA"` e `TEMPO_REF="ABSOLUTO"`).

H, M, lim: classificação referente, respectivamente aos campos hora, mês e limite do atributo `VAL_NORM` (quando `TIPO="HORA"`).

α, β, γ = parâmetros correspondentes aos pesos das categorias, tipos e subtipos.

δ, λ, ε, ξ, η = parâmetros correspondentes aos pesos dos atributos estendidos.

Figura 5.3: Fórmula da medida de avaliação do TEMPO estendido

primeiro caso, se o campo `H` (da hora) estiver mal preenchido, todo o atributo é considerado incorrecto, mas, no segundo caso, o facto de esse campo estar mal preenchido não impede que os restantes sejam considerados certos e pontuados como tal.

Esta distinção pareceu-nos, por um lado, modelar melhor o que se espera com o procedimento de normalização, e, pelo outro, nivelar de forma mais apropriada os sistemas. Por exemplo, para a frase 5.1, qual dos seguintes valores de `VAL_NORM` propostos por um sistema participante deverá ter uma medida mais elevada: `A0M0S20D2H0M0S0` ou `A0M0S3D2H0M0S0`?

(5.1) O evento durou <EM ID="tp-1" CATEG="TEMPO" TIPO="DURACAO" VAL_NORM="A0M0S23D2H0M0S0">**vinte e três semanas e 2 dias**</EM>

De facto, `A0M0S20D2H0M0S0` (correspondente a vinte semanas e dois dias) está mais próximo do valor pretendido do que `A0M0S3D2H0M0S0` (três semanas e dois dias), mas não tratámos estes dois casos de forma diferente porque muito possivelmente uma penalização realista teria de depender da aplicação prática do sistema, e podemos facilmente conceber casos em que "mais próximo" é igualmente deficiente.

No entanto, dado que a uma mesma expressão temporal com valor durativo pode corresponder mais do que uma representação de `VAL_NORM` (ou de `VAL_DELTA`, no caso de uma data com valor referencial), o valor desses atributos é considerado como certo se, ao ser convertido para uma unidade mínima, o valor convertido for igual ao valor da CD após conversão para a mesma unidade. Entende-se por unidade mínima a menor unidade usada para especificar este atributo, para a entidade em causa, na CD e na participação. As conversões são feitas de acordo com a tabela que se encontra nas directivas do TEMPO (cf. tabela B.6, no apêndice B).

Continuando com o mesmo exemplo, o valor `A0M5S0D13H0M0S0` (correspondente a cinco meses e treze dias) representa a mesma expressão e seria deste modo considerado como certo, caso fosse fornecido por um sistema em alternativa à resposta `A0M0S23D2H0M0S0`, que estaria na CD. De acordo com a tabela de conversões, que convenciona que um mês são trinta dias, e que uma semana corresponde a sete dias, ambos os valores seriam convertidos para 163 dias antes de serem comparados.

De acordo com a fórmula que apresentámos, os sistemas podem ser avaliados de quatro modos distintos no que diz respeito às entidades da categoria `TEMPO`:

Clássico, ignorando as parcelas referentes aos atributos estendidos, correspondendo portanto à avaliação dos atributos `CATEG`, `TIPO` e `SUBTIPO`;

Estendido completo, usando todas as parcelas da fórmula;

Estendido sem normalização, usando as parcelas (5.2) e (5.3);

Estendido só com normalização, usando as parcelas (5.2) e (5.4).

De certa forma, estes modos de avaliação na pista do TEMPO são semelhantes à avaliação por cenários selectivos, em que apenas um subconjunto dos atributos é tido em conta.

5.3 Avaliação do ReRelEM

5.3.1 Pontuações e medidas

A avaliação do ReRelEM é feita através da comparação entre as relações que estão na participação e as relações que estão na CD do ReRelEM.

De acordo com essa comparação, cada relação pode receber as pontuações: `Correcta`, `Espúria` ou `Em falta`, sendo que a medida utilizada é igual a 1 se a relação for correcta ou 0 caso contrário.

5.3.2 Expansão de relações

A fim de facilitar a tarefa dos sistemas mas, principalmente, de facilitar o processo de anotação humana, decidimos que não seria necessário anotar **todas** as relações entre EM que existem num documento.

Porém, para poder fazer a avaliação é necessário explicitar tanto na CD como na participação todas as relações que se encontram implícitas. Para tanto, essas relações são expandidas de forma automática pelo módulo `Expandidor` (ver em mais detalhe a sua descrição na secção 5.8.2).

A expansão consiste na explicitação de relações inversas (no caso de existirem), seguida pela aplicação de regras de expansão a pares de relações (como já mencionado no capítulo 4), o que pode dar origem a novas relações, antes apenas implícitas no texto.

5.3.3 Selecção de alinhamentos

A avaliação do ReRelEM centra-se apenas na avaliação de relações, e não na de EM, que é da responsabilidade do HAREM clássico. Assim, a avaliação de cada participação no ReRelEM é feita com base no subconjunto de anotações que são comuns à CD, obtido através do seguinte procedimento:

- remoção, na CD, de todas as EM em falta e respectivas relações;
- remoção, na participação, de todas as EM espúrias e respectivas relações;
- remoção de todas as classificações de EM diferentes do lado da CD e da participação, mantendo apenas as classificações comuns;
- remoção das relações entre facetas que tiverem sido removidas no ponto anterior (ver capítulo 4 para a noção de faceta).

5.4 Métricas

Nas secções anteriores foram apresentados os princípios, as pontuações e as medidas utilizadas na avaliação das três pistas integradas no Segundo HAREM. Quanto às métricas utilizadas, estas foram sempre:

- **Precisão:** afere a qualidade da participação, em termos da proporção de respostas correctas dentro do total de respostas dadas.
- **Abrangência:** afere a qualidade da participação, em termos da proporção de respostas correctas no universo de respostas possíveis.
- **Medida F:** combina a precisão e a abrangência para cada tarefa, de acordo com a seguinte fórmula:

$$Medida\ F = \frac{2 \cdot Precisão \cdot Abrangência}{Precisão + Abrangência} \tag{5.5}$$

A noção de resposta correcta varia naturalmente em função da medida utilizada. Concretizando e resumindo, para o HAREM clássico a unidade é o valor da medida já apresentada, para o TEMPO estendido é o valor das medidas anteriormente apresentadas e para o ReRelEM é o número de relações correctas.

5.5 Vista geral da arquitectura

Esta secção apresenta a arquitectura da plataforma utilizada na avaliação do Segundo HAREM: uma arquitectura modular, fortemente apoiada na arquitectura de avaliação do Primeiro HAREM (Seco et al. (2007)), onde se procurou que existissem vários módulos, cada um com a função de executar uma tarefa simples e específica.

5.5.1 Formato das colecções

No Segundo HAREM, seguindo as sugestões de Martins e Silva (2007) e de Almeida (2007), optou-se por utilizar colecções de documentos na notação XML, tendo como principal objectivo facilitar o processamento e a validação do material.

Uma colecção é então constituída por vários documentos delimitados pela etiqueta `DOC`. Essa etiqueta contém um atributo, `DOCID`, que é preenchido com um identificador único do documento na colecção. O documento pode conter outras etiquetas XML, como acontece com os textos anotados:

- Todas as entidades mencionadas são anotadas com a etiqueta `EM`, tendo ainda atributos para a sua classificação (`CATEG`, `TIPO` e `SUBTIPO`), atributos para identificar as suas relações com outras EM (`COREL` e `TIPOREL`) e atributos específicos do `TEMPO` (`TEMPO_REF`, `SENTIDO`, `VAL_NORM`, e `VAL_DELTA`).
- O elemento `ALT` representa um conjunto de análises alternativas de identificação (onde podem estar incluídas EM), separadas pelo caracter "|".

A correcção do formato dos vários ficheiros é verificado por meio de um validador, criado para o efeito.[3]

5.5.2 Os módulos

Todos os módulos foram implementados em Java, com a excepção do `Gerador de relatórios individuais` (implementado em R) e do `Avaliador do TEMPO estendido` (implementado em Awk). A biblioteca Java JDOM[4] foi utilizada para a manipulação de XML.

Cada módulo tem como entrada um (ou mais) ficheiros de texto, que são processados de forma a produzir um resultado pronto a ser tratado pelo módulo seguinte. A figura 5.4 representa todos os módulos de avaliação do Segundo HAREM, na ordem pela qual devem ser utilizados para se chegar aos resultados da avaliação. Podem ver-se não só os módulos para a avaliação do HAREM clássico mas também a indicação de onde entram os módulos para a avaliação do TEMPO estendido e do ReRelEM. A parte do diagrama com o fundo verde corresponde à avaliação individual de uma participação num determinado cenário selectivo; os módulos que se encontram fora desse fundo verde (ou seja, o `Gerador de resultados HAREM` e o `Gerador de relatórios individuais`), utilizam como entrada o conjunto de todas as avaliações.

Há ainda que referir o ficheiro `harem.conf`, que contém toda a informação relativa às possibilidades válidas para o preenchimento dos atributos das EM:

- Árvore de categorias, tipos e subtipos;
- Atributos específicos do `TEMPO`;
- Tipos de relação e respectivas inversas, se existirem.

Nas próximas secções encontra-se uma descrição mais detalhada dos vários módulos.

[3] Desenvolvido inicialmente por David Cruz e melhorado por Luís Miguel Cabral e mais tarde integrado no SAHARA (ver apêndice G).

[4] http://www.jdom.org

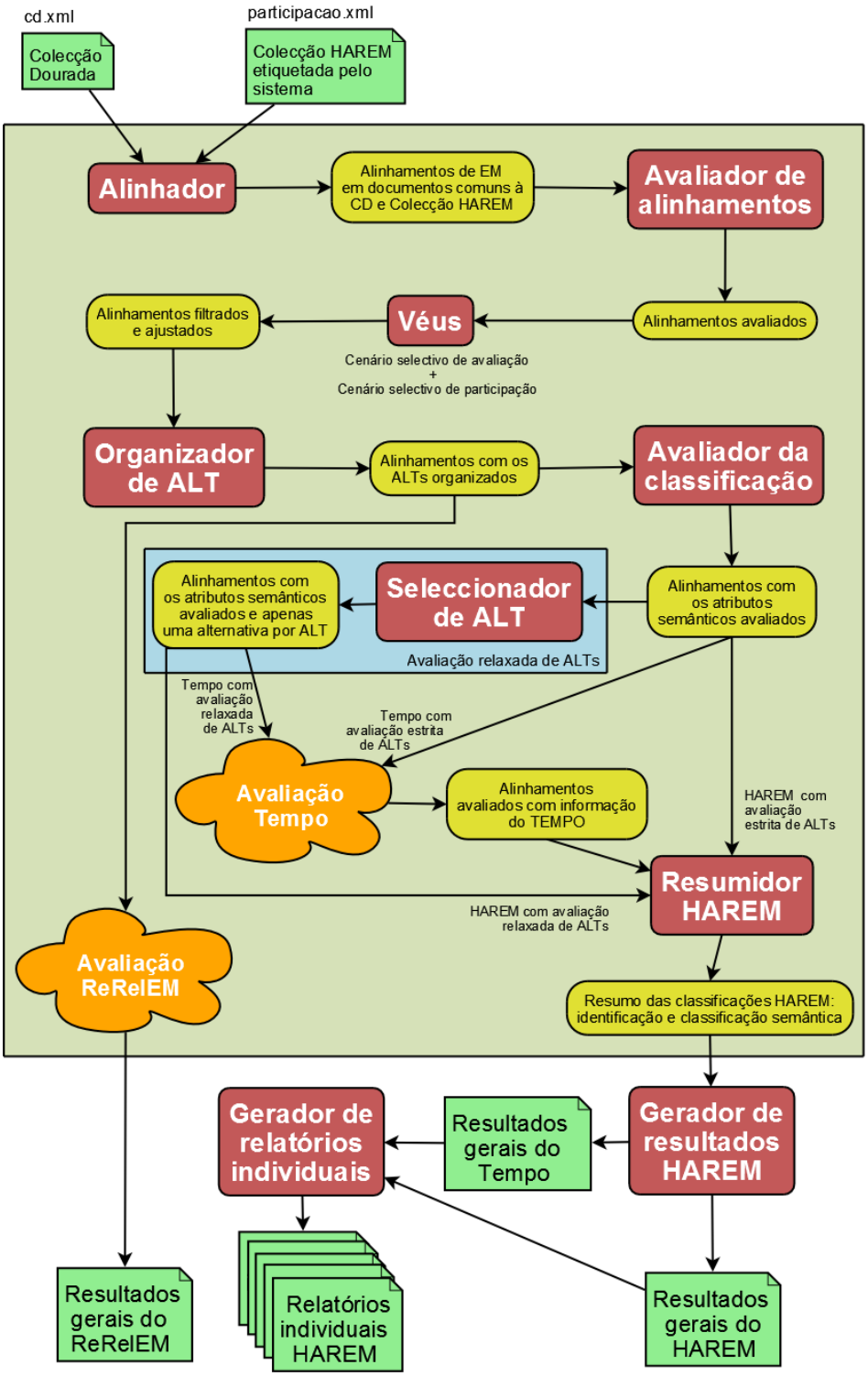


Figura 5.4: Arquitectura dos programas de avaliação no Segundo HAREM

```
#comentários

DOC DOCID\_1
<VERIFICACAO\_MANUAL>Informação para o juiz humano</VERIFICACAO\_MANUAL>
Alinhamento 1
Alinhamento 2
(...)
Alinhamento n

DOC DOCID_2
(...)

DOC DOCID_N
(...)
```

Figura 5.5: Formato da saída do `Alinhador`

5.6 Módulos de avaliação do HAREM clássico

Como já foi dito, grande parte dos módulos utilizados na avaliação do HAREM clássico foi construída a partir de módulos já criados para o Primeiro HAREM (Seco et al. (2007)), mais propriamente o `Alinhador`, o `Avaliador de alinhamentos`, o `Véus`, o `Avaliador da classificação`, o `Resumidor das classificações` e o `Gerador de resultados`[5]. Em alguns casos, acabaram contudo por existir alterações substanciais devido a todas as novidades do Segundo HAREM.

5.6.1 Alinhador

Uma participação é alinhada com a CD utilizando o `Alinhador`.

Este módulo alinha EM de documentos que estão tanto na CD como na participação, ignorando os outros documentos da Colecção do Segundo HAREM que estejam presentes na participação. Desta forma não é necessário utilizar um programa anterior para extrair da participação os documentos que fazem parte da CD, como acontecia no Primeiro HAREM.

5.6.1.1 Formato da saída

A saída do `Alinhador` é exemplificada na figura 5.5. As primeiras linhas, iniciadas por #, podem conter comentários, como por exemplo o nome da CD e participação a que se refere o alinhamento. De seguida, o início dos alinhamentos de cada documento é identificado através de uma linha iniciada por `DOC`, seguida pelo identificador do documento (valor do elemento `DOCID`).

Os alinhamentos podem ser de um dos cinco tipos que existiam no Primeiro HAREM conforme o exemplo adaptado de Seco et al. (2007):

- **um para um:** uma EM da CD alinha exactamente a uma EM na participação. `<EM>17:00<EM> --> [<EM>17:00</EM>]`
- **um para muitos:** uma EM da CD alinha a mais do que uma EM na participação. `<EM>17:00<EM> --> [<EM>17</EM>, <EM>00</EM>]`

[5] Estes módulos eram designados no Primeiro HAREM AlinhEM, AvalIDa, Véus, Emir, Ida e Sultão, respectivamente.

```
<EM>17:00<EM> ---> [<EM>17:00</EM>]::[Correcto]}
<EM>17:00<EM> ---> [<EM>17</EM>, <EM>00</EM>]::[Parcialmente\_Correcto\_por\_Defeito(0.25;
    0.75), Parcialmente\_Correcto\_por\_Defeito(0.25; 0.75)]}
```

Figura 5.6: Avaliação de alinhamentos

- **muitos para um:** mais do que uma EM da CD alinham a uma EM na participação.
  `<EM>17<EM> --> [<EM>17:00</EM>]`
  `<EM>00<EM> --> [<EM>17:00</EM>]`
- **nenhum para um:** uma EM é identificada na participação mas não há uma EM correspondente na CD. `<EM CATEG="ESPURIO">Ontem</EM> --> [<EM>Ontem</EM>]`
- **um para nenhum:** uma EM da CD não foi marcada como tal na participação. `<EM>Departamento de Informática</EM> --> [null]`

5.6.1.2 Etiquetas ALT

O tratamento das etiquetas `ALT` sofreu uma alteração considerável no Segundo HAREM: agora, estas etiquetas podem surgir tanto na CD como nas saídas dos participantes, que foram incentivados a utilizá-las para marcar conjuntos de análises alternativas de identificação. A sua utilização foi, portanto, também alvo de avaliação, o que levou a que o `Alinhador` tivesse de estar preparado para lidar com participações que enviassem mais do que uma segmentação para os mesmos fragmentos de texto.

5.6.1.3 Etiquetas OMITIDO

As etiquetas `OMITIDO` foram utilizadas pelas mesmas razões que o foram no Primeiro HAREM, nomeadamente para permitir algum controlo sobre partes do texto sem relevância linguística para a avaliação conjunta Santos e Cardoso (2007b). Devido à adopção do XML, os programas reconhecem agora `OMITIDO` como um elemento válido (embora apenas fazendo sentido do lado da CD). Internamente, sempre que um bloco de texto ocorre dentro de etiquetas `OMITIDO`, esse bloco é tratado como se fosse uma EM, com um atributo particular, que indica que é para ser omitida (`OMITIDO`). Ao escrever os alinhamentos na saída, o `Alinhador` ignora todos aqueles que do lado da CD tiverem uma EM com o atributo `OMITIDO`.

5.6.2 Avaliador de alinhamentos

O `Avaliador de alinhamentos` é exactamente igual ao utilizado no Primeiro HAREM. Os alinhamentos produzidos pelo `Alinhador` são avaliados, comparando a delimitação das EM do lado da CD com a delimitação das EM do lado da participação. A classificação é colocada à frente de cada alinhamento, como ilustrado na figura 5.6:

As possíveis pontuações são: `Correcto`, `Em Falta`, `Espúrio`, `Parcialmente_Correcto_por_Defeito` e `Parcialmente_Correcto_por_Excesso`. Apesar de no Segundo HAREM não serem contabilizadas EM parcialmente identificadas, o `Avaliador` mantém estes alinhamentos, que são tratados na fase seguinte.

```
Antes:
<EM>17:00<EM> ---> [<EM>17</EM>, <EM>00</EM>]::[Parcialmente\_Correcto\_por\_Defeito(0.25;
    0.75), Parcialmente\_Correcto\_por\_Defeito(0.25; 0.75)]

Depois:
<EM>17:00<EM> ---> [null]::[Em Falta]
<EM CATEG="ESPURIO">17</EM> ---> [<EM>17</EM>]::[Espurio]
<EM CATEG="ESPURIO">00</EM> ---> [<EM>00</EM>]::[Espurio]
```

Figura 5.7: Passagem de entidades parcialmente correcta a espúrias

5.6.3 Véus

O módulo `Véus` aplica os filtros utilizados para adaptar um conjunto de alinhamentos a um cenário selectivo definido por um conjunto de categorias, tipos e subtipos. Como já explicado na secção 5.1.3, é assim possível avaliar uma participação no cenário que for pretendido (o chamado cenário de avaliação), independentemente de ter sido o cenário em que o sistema participou, um subconjunto desse cenário ou mesmo um cenário que inclua algumas categorias em que o sistema não participou.

Ao contrário do Véus do Primeiro HAREM, o `Véus` no Segundo HAREM não filtra os documentos por género textual ou variante (PT, BR, etc.), visto que essa informação deixou de estar incluída no cabeçalho dos próprios documentos, passando antes a estar compreendida num ficheiro separado, que designámos como `Meta`.

O `Véus` do Primeiro HAREM permitia além disso parameterizar a avaliação de acordo com três estilos diferentes, sendo um deles o estilo `muc`, em que os casos parcialmente correctos não eram contabilizados. Esse estilo foi o único utilizado no Segundo HAREM, levando a que, a partir desta fase, deixassem de existir alinhamentos do tipo "*um para muitos*". As EM parcialmente correctas passaram assim a ser consideradas espúrias e as EM alinhadas com mais de uma EM passaram a contar como estando em falta. Na figura 5.7 ilustra-se a transformação de uma entidade parcialmente correcta em duas entidades espúrias.

5.6.3.1 Representação dos cenários selectivos

O `Véus` ajusta as EM nos alinhamentos de acordo com ambos os cenários selectivos – o de participação (opção `-sistema`) e o de avaliação (opção `-avaliacao`) –, evitando assim que tenha de ser criada uma CD e uma participação para cada cenário que se pretenda avaliar. Tal já foi ilustrado na anterior figura 5.2.

Os cenários são representados por uma lista com as categorias, tipos e subtipos que estão incluídos nesse cenário. A figura 5.8 ilustra a representação de alguns cenários.

5.6.3.2 Formato da saída

A saída do `Véus` continua a ser um conjunto de alinhamentos, agora conforme o estilo `muc` e de acordo com os cenários de avaliação e de participação. A única diferença no formato é que na primeira linha da saída é colocada a representação do cenário de avaliação, iniciada pelo caracter `#`.

Filtro	Descrição
"*"(ou sem a opção do cenário)	Cenário total.
"PESSOA(*):LOCAL(*):ORGANIZACAO(*)"	Apenas as categorias PESSOA, LOCAL e ORGANIZACAO e o seu conjunto normal de tipos e subtipos.
"LOCAL(FISICO{*};HUMANO{*})"	A categoria LOCAL apenas com os tipos FISICO e HUMANO e o seu conjunto normal de subtipos.
"LOCAL(FISICO{*};HUMANO{RUA,PAIS,DIVISAO,REGIAO})"	A categoria LOCAL apenas com o tipo FISICO como seu conjunto normal de subtipos e com o tipo HUMANO apenas com os subtipos RUA, PAIS, DIVISAO e REGIAO.

Figura 5.8: Exemplos de representação de cenários.

5.6.3.3 Exemplo de aplicação de filtros pelo Véus

A figura 5.9 exemplifica a aplicação de um filtro para avaliar apenas a categoria PESSOA numa participação que concorreu exclusivamente nas categorias PESSOA e ORGANIZACAO. O cenário de participação é utilizado apenas na decisão de manter a situação da última linha da figura, em que existe uma EM não classificada que é espúria, mas que não é eliminada porque tanto o cenário de avaliação como o cenário de participação contêm a categoria PESSOA.

5.6.4 Organizador de ALT

Como já foi dito na secção 5.1.4, no Segundo HAREM existem dois tipos de avaliação relacionados com a utilização de ALT, sendo o Organizador de ALT o módulo que possibilita a avaliação estrita. Esta avaliação obedece aos seguintes três passos:

1. O Alinhador associa as EM da participação às EM na CD, estando estas dentro ou fora de ALT. A saída fica organizada de acordo com os ALT na CD.

2. O Véus remove todos os alinhamentos parcialmente correctos. A partir deste passo, só não existem EM espúrias ou em falta dentro de um ALT se a participação contiver um ALT exactamente igual ao que está na CD. (A aplicação de um filtro para avaliar um cenário selectivo pode também dar origem a alternativas vazias ou duplicadas.)

3. Finalmente, o Organizador de ALT garante que:

   - não haja análises iguais, que possam ter surgido após a aplicação do Véus, por meio da remoção de duplicados;
   - não haja alternativas que deixaram de conter EM após a aplicação do Véus;

Antes	Depois
<EM CATEG="PESSOA"> -> <EM CATEG="PESSOA">	<EM CATEG="PESSOA"> -> <EM CATEG="PESSOA">
<EM CATEG="PESSOA|ORGANIZACAO"> -> <EM CATEG="PESSOA">	<EM CATEG="PESSOA"> -> <EM CATEG="PESSOA">
<EM CATEG="PESSOA"> -> <EM CATEG="PESSOA|ORGANIZACAO">	<EM CATEG="PESSOA"> -> <EM CATEG="PESSOA">
<EM CATEG="PESSOA"> -> <EM CATEG="ORGANIZACAO">	<EM CATEG="PESSOA"> -> [null]
<EM CATEG="ORGANIZACAO"> -> <EM CATEG="ORGANIZACAO">	nada
<EM CATEG="ORGANIZACAO"> -> [null]	nada
<EM CATEG="ESPURIO"> -> <EM CATEG="ORGANIZACAO">	nada
<EM CATEG="PESSOA"> -> [null]	<EM CATEG="PESSOA"> -> [null]
<EM CATEG="ESPURIO"> -> <EM>	<EM CATEG="ESPURIO"> -> <EM>

Figura 5.9: Exemplos da aplicação do Véus, com cenário de participação constituído por PESSOA e ORGANIZACAO e cenário de avaliação PESSOA.

- o peso 1/(total de elementos) seja atribuído a cada EM dentro de um ALT (esse peso é colocado no fim de cada alinhamento, depois do caracter '^').

A figura 5.10 exemplifica a evolução de processamento de um ALT ao longo dos passos descritos acima.

5.6.5 Listador de espúrios

O Listador de espúrios é um módulo que pode ser aplicado sobre a saída do Organizador de ALT no caso de se pretender listar todas as EM espúrias, para efeitos de depuração.

5.6.6 Avaliador da classificação

O Avaliador da classificação recebe a saída do Organizador de ALT e calcula a pontuação de cada atributo, comparando as categorias, tipos e subtipos do lado da CD com os mesmos atributos do lado da participação.

A medida da classificação é também calculada aplicando aos atributos de cada EM e respectivas pontuações a fórmula que se encontra na figura 5.1.

5.6.6.1 Formato da saída

A saída do Avaliador da classificação continua a conter os alinhamentos, cuja avaliação (anexada ao final da linha correspondente a cada um) é constituída por:

- Pontuação dos atributos: pontuação do preenchimento de cada atributo nas EM da participação. O resultado da comparação com os atributos das EM na CD pode ser

```
CD:
O <ALT>
<EM ID="Xyz-60" CATEG="ACONTECIMENTO" TIPO="ORGANIZADO">Tour de França de 2009</EM> |
<EM ID="Xyz-60-aa" CATEG="ACONTECIMENTO" TIPO="ORGANIZADO">Tour de França</EM> <EM ID="Xyz-5"
    CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">de 2009</EM> | <EM ID="Xyz-60-aaa" CATEG
    ="ACONTECIMENTO" TIPO="ORGANIZADO">Tour</EM> de <EM ID="Xyz-61" CATEG="LOCAL" TIPO="
    HUMANO" SUBTIPO="PAIS">França</EM> <EM ID="Xyz-5-aa" CATEG="TEMPO" TIPO="TEMPO_CALEND"
    SUBTIPO="DATA">de 2009</EM>
</ALT> ...

Participação:
O Tour de <EM ID="Xyz_1" CATEG="LOCAL" TIPO="FISICO" SUBTIPO="REGIAO">França</EM> <EM ID="
    Xyz_2" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">de 2009</EM> ...

Saída do alinhador:
<ALT>
<ALT1>
<EM ID="Xyz-60" CATEG="ACONTECIMENTO" TIPO="ORGANIZADO">Tour de França de 2009</EM> ---> [<EM
    ID="Xyz_1" CATEG="LOCAL" TIPO="FISICO" SUBTIPO="REGIAO">França</EM>, <EM ID="Xyz_2" CATEG
    ="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">de 2009</EM>]
</ALT1>
<ALT2>
<EM ID="Xyz-60-aa" CATEG="ACONTECIMENTO" TIPO="ORGANIZADO">Tour de França</EM> ---> [<EM ID="
    Xyz_1" CATEG="LOCAL" TIPO="FISICO" SUBTIPO="REGIAO">França</EM>]
<EM ID="Xyz-5" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">de 2009</EM> ---> [<EM ID="
    Xyz_2" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">de 2009</EM>]
</ALT2>
<ALT3>
<EM ID="Xyz-60-aaa" CATEG="ACONTECIMENTO" TIPO="ORGANIZADO">Tour</EM> ---> [null]
<EM ID="Xyz-61" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="PAIS">França</EM> ---> [<EM ID="Xyz_1"
    CATEG="LOCAL" TIPO="FISICO" SUBTIPO="REGIAO">França</EM>]
<EM ID="Xyz-5-aa" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">de 2009</EM> ---> [<EM ID="
    Xyz_2" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">de 2009</EM>]
</ALT3>
</ALT>

Saída do Véus (com cenário de avaliação "apenas LOCAL"):
<ALT>
<ALT1>
<EM CATEG="ESPURIO">França</EM> ---> [<EM ID="Xyz_1" CATEG="LOCAL" TIPO="FISICO" SUBTIPO="
    REGIAO">França</EM>]::[Espurio]
</ALT1>
<ALT2>
<EM CATEG="ESPURIO">França</EM> ---> [<EM ID="Xyz_1" CATEG="LOCAL" TIPO="FISICO" SUBTIPO="
    REGIAO">França</EM>]::[Espurio]
</ALT2>
<ALT3>
<EM ID="Xyz-61" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="PAIS">França</EM> ---> [<EM ID="Xyz_1"
    CATEG="LOCAL" TIPO="FISICO" SUBTIPO="REGIAO">França</EM>]::[Correcto]
</ALT3>
</ALT>

Saída do Organizador de ALT:
<ALT>
<ALT1>
<EM ID="Xyz-61" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="PAIS">França</EM> ---> [<EM ID="Xyz_1"
    CATEG="LOCAL" TIPO="FISICO" SUBTIPO="REGIAO">França</EM>]::[Correcto]^1
</ALT1>
</ALT>
```

Figura 5.10: Evolução de um ALT

```
<EM ID="551" CATEG="ACONTECIMENTO" TIPO="EVENTO">Proclamação da Carta dos Direitos
    Fundamentais</EM> ---> [<EM ID="133" CATEG="ACONTECIMENTO" TIPO="ORGANIZADO">Proclamação
    da Carta dos Direitos Fundamentais</EM>]::[{Categoria(Correcto:[ACONTECIMENTO] Espurio:[]
     Em_Falta:[]) Tipo(Correcto:[] Espurio:[ORGANIZADO] Em_Falta:[EVENTO]) Subtipo(Correcto:
    [] Espurio:[] Em_Falta:[]) MaxCSC_CD(2.275) MaxCSC_S(2.275) CSC(1.775) Peso(1.0)}]

<ALT>
<ALT1>
<EM ID="113" CATEG="PESSOA" TIPO="INDIVIDUAL">Filipe II de Espanha</EM> ---> [null]::[{
    Categoria(Correcto:[] Espurio:[] Em_Falta:[]) Tipo(Correcto:[] Espurio:[] Em_Falta:[])
    Subtipo(Correcto:[] Espurio:[] Em_Falta:[]) MaxCSC_CD(2.3375) MaxCSC_S(0.0) CSC(0.0) Peso
    (0.0) PALT(0.5)}]
</ALT1>
<ALT2>
<EM ID="113aa" CATEG="PESSOA" TIPO="INDIVIDUAL">Filipe II</EM> ---> [null]::[{Categoria(
    Correcto:[] Espurio:[] Em_Falta:[]) Tipo(Correcto:[] Espurio:[] Em_Falta:[]) Subtipo(
    Correcto:[] Espurio:[] Em_Falta:[]) MaxCSC_CD(2.3375) MaxCSC_S(0.0) CSC(0.0) Peso(0.0)
    PALT(0.5)}]
<EM ID="2" CATEG="LOCAL|ORGANIZACAO" TIPO="HUMANO|ADMINISTRACAO" SUBTIPO="PAIS|">Espanha</EM>
    ---> [<EM ID="1629" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="DIVISAO">Espanha</EM>]::[{
    Categoria(Correcto:[LOCAL] Espurio:[] Em_Falta:[ORGANIZACAO]) Tipo(Correcto:[HUMANO]
    Espurio:[] Em_Falta:[]) Subtipo(Correcto:[] Espurio:[DIVISAO] Em_Falta:[PAIS]) MaxCSC_CD
    (3.7583333333333333) MaxCSC_S(2.4833333333333334) CSC(2.2333333333333334) Peso(1.0) PALT
    (0.5)}]
</ALT2>
</ALT>

<EM ID=185" CATEG="PESSOA" TIPO="INDIVIDUAL">Sócrates </EM> ---> [null]::[{Categoria(Correcto:
    [] Espurio:[] Em_Falta:[]) Tipo(Correcto:[] Espurio:[] Em_Falta:[]) Subtipo(Correcto:[]
    Espurio:[] Em_Falta:[]) MaxCSC_CD(2.3375) MaxCSC_S(0.0) CSC(0.0) Peso(0.0)}]

<EM CATEG="ESPURIO">Escola </EM> ---> [<EM ID="111" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="
    DIVISAO">Escola </EM>]::[{Categoria(Correcto:[] Espurio:[] Em_Falta:[]) Tipo(Correcto:[]
    Espurio:[] Em_Falta:[]) Subtipo(Correcto:[] Espurio:[] Em_Falta:[]) MaxCSC_CD(0.0)
    MaxCSC_S(2.4833333333333334) CSC(0.0) Peso(0.0)}]
```

Figura 5.11: Exemplos de alinhamentos após a aplicação do `Avaliador da classificação`.

pontuado como: `Correcto`, `Espúrio` ou `Em Falta`. Se uma categoria não estiver correcta, o tipo e o subtipo não são pontuados e, se um tipo não estiver correcto, o subtipo também não é pontuado.

- `CSC`: classificação efectiva da anotação do sistema, comparada com a anotação da CD;
- `Peso`: peso do alinhamento (0 se for `Espúrio` ou `Em Falta`, 1 se estiver `Correcto`). No Primeiro HAREM poderia ter valores entre 0 e 1 para entidades parcialmente identificadas.
- `PALT`: peso do alinhamento dentro de um `ALT` (por omissão é igual a 1);
- `MaxCSC_CD`: classificação máxima da anotação na CD;
- `MaxCSC_S`: classificação máxima possível com a anotação presente na participação.

Alguns exemplos de alinhamentos após a aplicação do `Avaliador da classificação` no cenário total estão na figura 5.11.

```
Avaliação Global – Classificação
Valor máximo possível para a Classificação na CD: 15971.775396825311
Valor máximo possível para a Classificação do sistema: 11681.552480158529
Valor da Classificação do sistema: 7424.880753968094
Precisão Máxima do Sistema: 0.6356073618279316
Abrangência Máxima na CD: 0.4648751043321037
Medida F: 0.5369972675258301
```

Figura 5.12: Fragmento de saída do resumidor de classificações

5.6.7 Seleccionador de ALT

O `Seleccionador de ALT` é utilizado apenas na avaliação relaxada de `ALT`. É aplicado sobre a saída do `Avaliador da classificação` e tem como objectivo seleccionar apenas o melhor elemento dentro de cada `ALT`.

Existem dois critérios para a selecção do melhor elemento, empregues em sequência (ou seja, se o primeiro não permitir decidir, recorre-se ao segundo):

1. Elemento com a melhor medida F.
2. Elemento com com o maior valor da medida de classificação.

5.6.8 Resumidor das classificações

O `Resumidor das classificações` processa a saída do `Avaliador da classificação`, utilizando os valores que este último calcula para cada alinhamento para obter as medidas totais na participação e aferir o desempenho do sistema. São apresentadas as medidas tanto para a tarefa de classificação como para identificação apenas (onde é considerado um ponto por EM correctamente identificada – por outras palavras, na fórmula de avaliação apresentada na figura 5.1, os pesos α, β e γ são zero).

A figura 5.12 mostra um fragmento da saída do `Resumidor das classificações`, onde se pode ver o valor da precisão, da abrangência e da medida F obtidos por um sistema. Há ainda a referir que no cabeçalho da saída do `Resumidor das classificações` se encontra a representação do cenário selectivo em questão.

5.6.9 Gerador de resultados

O `Gerador de resultados` tem como objectivo juntar os resultados de todas as saídas do `Resumidor de classificações`, comparando-os de forma a criar uma tabela (em HTML) com as participações ordenadas pela melhor medida F, mostrando ainda outros valores, como a precisão e a abrangência.

O `Gerador de resultados` é invocado uma vez para cada cenário. Dessa forma, é gerada uma tabela para cada um dos cenários avaliados, cuja consulta facilita a visualização e comparação do desempenho dos vários sistemas.

5.6.10 Gerador de relatórios individuais

O `Gerador de relatórios individuais` gera um relatório de desempenho detalhado por cada sistema participante. Este módulo processa os ficheiros em HTML produzidos

pelo Gerador de resultados para a avaliação do HAREM clássico e para a avaliação do TEMPO estendido, mas não para a avaliação do ReRelEM.

Cada relatório inclui tabelas e gráficos individuais de um sistema (ou seja, contendo apenas os resultados das suas corridas). Por exemplo, na figura 5.13(a) ilustra-se o gráfico comparativo dos cenários de avaliação com o desempenho das corridas do sistema Cage2.

O relatório inclui igualmente gráficos comparativos com todas as corridas avaliadas no Segundo HAREM, de forma a que as corridas possam ser comparadas como se fossem de sistemas diferentes (embora as corridas do sistema a que o relatório diz respeito se encontrem destacadas). Veja-se, por exemplo, a figura 5.13(b), que mostra o gráfico precisão/abrangência gerado para o relatório do sistema da Priberam, e a figura 5.13(c), que mostra o gráfico das três métricas de avaliação gerado para o relatório do sistema REMBRANDT.

Além disso, foram também criados gráficos que mostram o desempenho em termos de medida F das várias corridas agrupadas pelo sistema a que pertencem. Este último gráfico permite comparar de forma mais imediata os sistemas (cf. figura 5.13(d), que destaca o desempenho do sistema XIP-L2F/Xerox).

Os gráficos e tabelas são criados para todos os cenários de avaliação, modos de avaliação do TEMPO estendido e avaliação por categoria.

5.7 Módulo de avaliação da pista do TEMPO

Se as entidades da categoria TEMPO apenas tivessem os atributos do HAREM clássico CATEG, TIPO e SUBTIPO, os sistemas teriam sido avaliados tendo por base apenas a CD do Segundo HAREM e o encadeamento de avaliação do HAREM clássico teria sido suficiente para fazer a avaliação dos sistemas na pista do TEMPO, de acordo com o que foi descrito na secção anterior. Como tal não foi o caso, foi necessário incluir na sequência de processamento um módulo adicional para avaliar as entidades da categoria TEMPO quanto aos atributos estendidos na CD do TEMPO, o que é ilustrado na figura 5.4 com a nuvem *Avaliação TEMPO*.

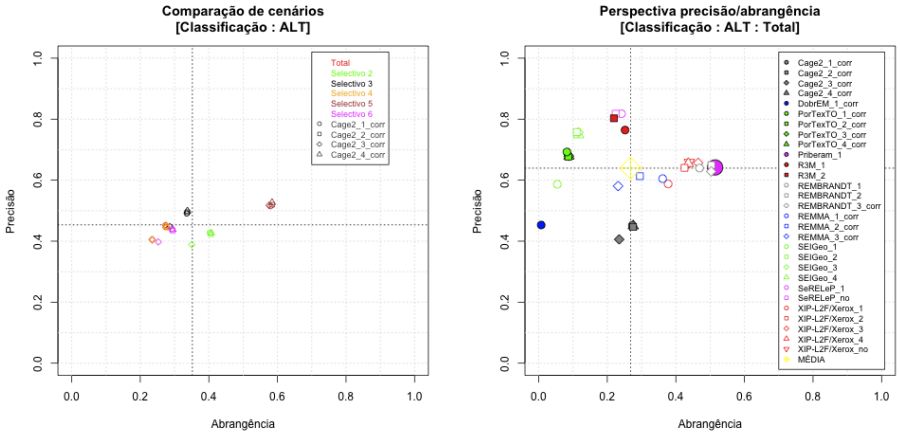
A avaliação dos sistemas na pista do TEMPO usando a CD do TEMPO pressupõe então que a avaliação da classificação, ou seja, o HAREM clássico, tenha sido previamente realizada. Ou seja, o procedimento de avaliação segue o encadeamento esquematizado na figura 5.4, em que *cd.xml*, a entrada do processo, corresponde à CD do TEMPO.

A nuvem de avaliação do TEMPO estendido contém apenas um módulo que analisa os alinhamentos após terem sido processados pelo módulo Avaliador da classificação e produz um novo ficheiro de alinhamentos. Cada linha nesse ficheiro corresponde a uma linha do ficheiro original, mas modificada para conter a avaliação dos atributos estendidos de TEMPO. Essa avaliação é feita de acordo com a fórmula descrita na secção 5.2.

A figura 5.14 mostra uma linha no ficheiro de entrada, que contém o valor da medida de classificação para a entidade em análise, e a linha correspondente no ficheiro de saída, com a informação específica da avaliação dos atributos adicionais de TEMPO.

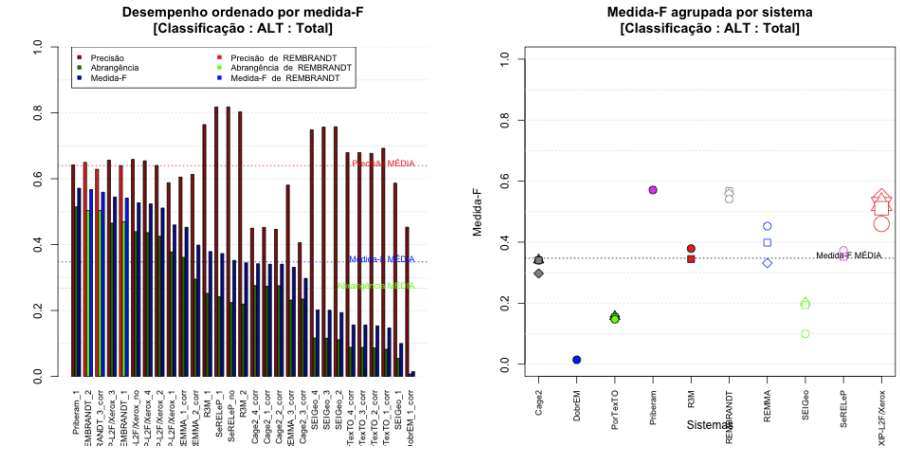
O módulo de avaliação do TEMPO estendido analisa cada alinhamento do ficheiro de entrada, produzindo um novo, em que os valores MaxCSC_CD, MaxCSC_S e CSC são incrementados com o valor adicional máximo do lado da CD, do lado do sistema e o valor acrescentado que o sistema tem ao acertar nos atributos estendidos, respectivamente.

O alinhamento passa também a incluir três campos auxiliares que indicam a decomposição dos valores totais. No exemplo, estes campos estão preenchidos com os seguintes



(a) Gráfico de precisão/abrangência comparativo dos vários cenários para um único participante

(b) Gráfico de precisão/abrangência comparativo das várias corridas participantes, com um dos sistemas destacados



(c) Gráfico de precisão, abrangência e medida F comparativo das várias corridas participantes, com um dos sistemas destacados

(d) Gráfico de medida F comparativo das várias corridas participantes agrupadas por sistema, com um dos sistemas destacados

Figura 5.13: Exemplos de gráficos presentes nos relatórios individuais

valores:

```
MaxCSC_CD=MaxCSC_CD_class(2.4875)+MaxCSC_CD_TEMPO(1)
MaxCSC_S=MaxCSC_S_class(2.4875)+MaxCSC_S_TEMPO(1)
CSC=CSC_class(2.4875)+CSC_TEMPO(1).
```

Além disso, são igualmente adicionados campos que indicam se os atributos estão correctos ou em falta.

```
Entrada:
<EM ID="H2-Ren_2003_6465-168" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="HORA" VAL_NORM="
    +---------T1806E---LM-">às 18h06</EM> ---> [<EM ID="251-26003-26011" CATEG="TEMPO" TIPO="
    TEMPO_CALEND" SUBTIPO="HORA" VAL_NORM="+---------T1806E---LM-">às 18h06</EM>]::[{Categoria(
    Correcto:[TEMPO] Espurio:[] Em_Falta:[]) Tipo(Correcto:[TEMPO_CALEND] Espurio:[]
    Em_Falta:[]) Subtipo(Correcto:[HORA] Espurio:[] Em_Falta:[]) MaxCSC_CD(2.4875) MaxCSC_S
    (2.4875) CSC(2.4875) Peso(1.0)}]

Saída:
<EM ID="H2-Ren_2003_6465-168" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="HORA" VAL_NORM="
    +---------T1806E---LM-">às 18h06</EM> ---> [<EM ID="251-26003-26011" CATEG="TEMPO" TIPO="
    TEMPO_CALEND" SUBTIPO="HORA" VAL_NORM="+---------T1806E---LM-">às 18h06</EM>]::[{Categoria(
    Correcto:[TEMPO] Espurio:[] Em_Falta:[]) Tipo(Correcto:[TEMPO_CALEND] Espurio:[]
    Em_Falta:[]) Subtipo(Correcto:[HORA] Espurio:[] Em_Falta:[]) MaxCSC_CD(3.4875) MaxCSC_S
    (3.4875) CSC(3.4875) Peso(1.0)}]::[{ValNorm(Correcto:[+---------T1806E---LM-] Em_Falta:[])
    MaxCSC_CD=MaxCSC_CD_class(2.4875)+MaxCSC_CD_TEMPO(1) MaxCSC_S=MaxCSC_S_class(2.4875)+
    MaxCSC_S_TEMPO(1) CSC=CSC_class(2.4875)+CSC_TEMPO(1)}]}
```

Figura 5.14: Avaliação dos atributos estendidos de entidades da categoria `TEMPO`

Após esta actualização dos alinhamentos, o ficheiro será processado pelo módulo `Resumidor de classificações`, seguindo o processamento normal que seguiria a avaliação do HAREM clássico, de modo a produzir as medidas de precisão, abrangência e medida F no preenchimento dos atributos do `TEMPO`.

5.8 Módulos de avaliação do ReRelEM

Esta secção apresenta a avaliação da pista do ReRelEM, que consiste em reconhecer relações entre EM (ver capítulo 4), e descreve também uma aplicação criada para visualizar grafos onde estejam representadas as relações anotadas num documento com o esquema de anotação do ReRelEM (ver secção 5.8.9).

Como se pode verificar na figura 5.15, a primeira etapa da avaliação é a conversão da notação da CD do ReRelEM e das participações para uma notação compatível com a esperada pelos módulos (ver secção 5.8.1). A fase seguinte utiliza alguns dos módulos desenvolvidos para o HAREM clássico e, por fim, acontece a avaliação de relações, que compreende a expansão das relações, a selecção dos alinhamentos sujeitos à avaliação, a normalização dos identificadores das EM da participação, a alteração da representação utilizada nos ficheiros intermédios para focar as relações, a filtragem de alguns tipos de relação e a avaliação propriamente dita, com a geração de tabelas com os resultados.

5.8.1 Conversão de notação

A conversão de notação é a primeira etapa da avaliação do ReRelEM, realizada antes mesmo da CD do ReRelEM e da participação serem processadas pelos módulos do HAREM clássico. Esta conversão é necessária uma vez que foi criada uma nova notação para possibilitar a identificação e facilitar a leitura de relações entre diferentes facetas de EM vagas. A notação, a que chamamos tipo 2, diverge da notação até aqui utilizada (tipo 1) na quantidade de informação colocada no atributo `TIPOREL`: (i) a faceta da própria EM (`CATORIGEM`) que entra na relação, (ii) o tipo da relação (`TIPOREL`), (iii) o ID da EM relacionada (`COREL`) e (iv) a faceta da EM relacionada (`CATALVO`), separados por `**`.

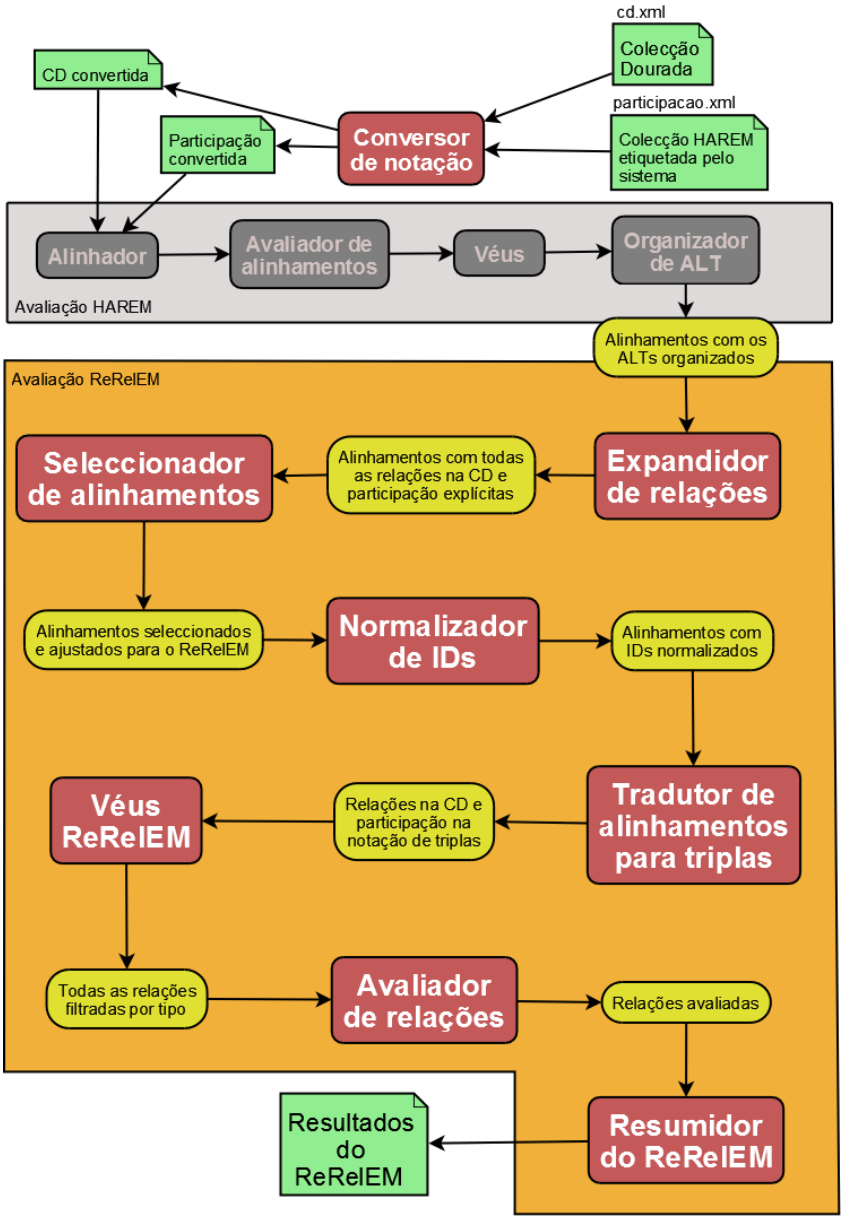


Figura 5.15: Arquitectura dos programas de avaliação para o ReRelEM.

```
Tipo 1: <EM ID="X" CATEG="CATORIGEM1|CATORIGEM2" COREL="ID1 ID1 ID2" TIPOREL="rel1 rel2 rel3">

Tipo 1 (com especificação de facetas relacionadas): <EM ID="X" CATEG="CATORIGEM1|CATORIGEM2"
    COREL="ID1 ID2 ID2" TIPOREL="rel1 rel2 rel3" FAC\_ORIGEM="CATORIGEM1 CATORIGEM2
    CATORIGEM1" FAC\_ALVO="CATALVO1 CATALVO1 CATALVO1">

Tipo 2: <EM ID="X" CATEG="CATORIGEM1|CATORIGEM2" TIPOREL="CATORIGEM1**rel1**ID1**CATALVO1
    CATORIGEM2**rel2**ID1**CATALVO1 CATORIGEM1**rel3**ID2**CATALVO1">
```

Figura 5.16: Formatos da CD do ReRelEM

```
Notação de tipo 1:

<EM ID="X" CATEG="LOCAL">Europa</EM>
<EM ID="Y" CATEG="LOCAL|ORGANIZACAO">UE</EM>
<EM ID="Z" CATEG="LOCAL|ORGANIZACAO" COREL= "X Y Y" TIPOREL="incluido incluido incluido" FAC\
    _ORIGEM="LOCAL LOCAL ORGANIZACAO" FAC\_ALVO="LOCAL LOCAL ORGANIZACAO">Portugal</EM>

Notação de tipo 2:

<EM ID="X" CATEG="LOCAL">Europa</EM>
<EM ID="Y" CATEG="LOCAL|ORGANIZACAO">UE</EM>
<EM ID="Z" CATEG="LOCAL|ORGANIZACAO" COREL= "X Y Y" TIPOREL= "LOCAL**incluido**X**LOCAL LOCAL
    **incluido**Y**LOCAL ORGANIZACAO**incluido**Y**ORGANIZACAO>Portugal </EM>
```

Figura 5.17: Entidades anotadas com notações de tipo 1 e 2

Apesar da notação de tipo 2 facilitar a anotação e leitura das relações por humanos, o seu processamento coloca alguns problemas técnicos[6]. Foi por isso decidido que, embora mantida na CD do ReRelEM, esta notação seria transformada durante o processo de avaliação, a fim de manter a compatibilidade com os demais programas. Assim, o atributo `TIPOREL` volta a ter apenas o tipo da relação, e cada EM recebe dois novos atributos para representar as facetas nas relações (`FAC_ORIGEM` e `FAC_ALVO`). A essa notação chamamos tipo 1 com especificação de facetas, porque o preenchimento do atributo `TIPOREL` é igual ao tipo 1, e a diferença encontra-se nos dois atributos adicionais.

A figura 5.16 ilustra genericamente as notações de tipo 1 (com e sem facetas especificadas) e tipo 2.

Para cada relação que dá entrada nos programas é obrigatória uma entrada nos atributos `FAC_ORIGEM` e `FAC_ALVO`. Quando se trata de uma relação entre EM vagas e existe a informação das facetas relacionadas (no tipo 2), as facetas são obtidas através da informação no atributo `TIPOREL`. Quando essa informação não existe (EM simples ou EM numa participação onde essa informação não existe), considera-se que as categorias das EM relacionadas são também as facetas na relação. Na figura 5.17, ilustramos várias entidades anotadas tanto com o tipo 1 como 2.

[6] Dificuldades relativas às várias informações que passam a estar contidas no atributo `TIPOREL` e a replicação, também no `TIPOREL`, do ID, que já se encontra no atributo `COREL`.

5.8.2 Expandidor de relações

O `Expandidor de relações` é um módulo chave no ReRelEM, utilizado para maximizar todas as relações implícitas no conjunto de relações anotadas na CD ou nas participações.

O conjunto de relações anotadas é aumentado por meio da explicitação de todas as relações inversas ou simétricas (se existirem) que não se encontram marcadas. O módulo testa ainda se a combinação de cada par de relações existente pode originar uma nova relação através das regras de expansão. A informação relativa aos tipos de relação e respectivas inversas encontra-se no ficheiro `harem.conf` de forma a poder ser facilmente parametrizada.

5.8.2.1 Expansão

As regras de expansão utilizadas neste primeiro ReRelEM resumem-se às seguintes quatro (já mencionadas no capítulo 4):

1. A `ident` B $\wedge$ B `ident` C $\Rightarrow$ A `ident` C
2. A `inclui` B $\wedge$ B `inclui` C $\Rightarrow$ A `inclui` C
3. A `inclui` B $\wedge$ B `sede_de` C $\Rightarrow$ A `sede_de` C
4. A `ident` B $\wedge$ B `qualquer_relação` C $\Rightarrow$ A `qualquer_relação` C

O `Expandidor` procura normalizar cada par de relações de forma a verificar se algum segue uma das regras anteriores. A normalização passa essencialmente pelos passos descritos a seguir:

1. verificar se as duas relações do par têm um argumento comum (B) e outro diferente;
2. transformar todas as relações que têm inversa no tipo convencionado directo (`inclui`, `sede_de`, ...);[7]
3. colocar como primeiro elemento do par a relação que tiver B no segundo argumento, e como segundo elemento do par, a relação que tiver B no primeiro argumento;
4. se o par obtido seguir uma das regras de expansão, ele dá origem a uma nova relação que, se ainda não existir, será adicionada ao conjunto de relações existentes.

Se algum destes passos não for possível, o par não dará origem a uma nova relação. Na figura 5.18 estão alguns exemplos do procedimento de expansão.

É possível optar pela expansão apenas das relações do lado da CD (`-exptudo nao`), embora por omissão a expansão seja efectuada dos dois lados. Esta opção foi desenvolvida para verificar as consequências da expansão de eventuais relações erradas.

[7] Como a relação de identidade é simétrica, esta transformação é realizada se necessário.

Par	**Passo 1**	**Passo 2**	**Passo 3**	**Resultado**
B ident C ∧ A ident B	Sim	B ident C ∧ A ident B	A ident B ∧ B ident C	A ident C
A incluido B ∧ C inclui A	Sim	B inclui A ∧ C inclui A	Impossível	Nada
A incluido B ∧ A inclui C	Sim	B inclui A ∧ A inclui C	B inclui A ∧ A inclui C	B inclui C
A ocorre_em B ∧ B incluido C	Sim	B sede_de A ∧ C inclui B	C inclui B ∧ B sede_de A	C sede_de A
A ocorre_em B ∧ C incluido D	Não	Nada	Nada	Nada
B ident A ∧ B natural_de C	Sim	B ident A ∧ B natural_de C	A ident B ∧ B natural_de C	A natural_de C

Figura 5.18: Exemplos de aplicação das regras de transitividade

5.8.2.2 Compatibilidade de facetas

Consideramos as facetas intervenientes em cada relação do lado da participação as próprias categorias das EM, sejam elas simples ou vagas.[8]. Definiu-se por isso que uma relação é igual a outra se tiver o mesmo tipo, os mesmos argumentos (em termos de EM) e facetas compatíveis, ou seja, facetas iguais ou facetas passíveis de unificação, considerando que uma categoria não existente (se por acaso o sistema apenas marcou `EM` e não especificou a categoria) é compatível com qualquer conjunto de categorias. Por exemplo:

- `LOCAL` é compatível com `LOCAL`;
- `LOCAL|ORGANIZACAO` é compatível com `LOCAL`;
- `EM` é compatível com `LOCAL`.

5.8.3 Seleccionador de alinhamentos

O `Seleccionador de alinhamentos` é aplicado ao conjunto de alinhamentos produzidos pelo `Organizador de ALT`, de onde vai seleccionar apenas os alinhamentos (ou parte) relativos a respostas correctas dadas pelos sistemas no HAREM clássico, de acordo com o descrito na secção 5.3.3.

5.8.4 Normalizador de identificadores (ID)

Para que seja possível comparar as relações do lado da CD com as relações do lado da participação é necessário normalizar os ID das EM de cada lado, utilizando um identificador comum. O `Normalizador de identificadores` vai substituir o valor dos atributos ID das EM do lado da participação pelo valor dos atributos ID das EM correspondentes na CD. As entradas dos atributos `COREL`, `TIPOREL`, `FAC_ORIGEM` e `FAC_ALVO` são também normalizadas: os valores dos ID em `COREL` são adaptados aos novos ID e as relações que envolvam pelo menos um ID ou um conjunto *ID+faceta* inexistente são removidas.

5.8.5 Tradutor de alinhamentos para triplas

Como o foco do ReRelEM é a identificação e avaliação de relações, para facilitar a depuração e o processamento pelos módulos seguintes, o módulo `Tradutor de alinhamentos para triplas` transforma os vários alinhamentos em listas de relações, representadas por triplas. Para cada documento são listadas as relações do lado da CD e as relações do lado da participação. As triplas têm o formato: `arg1 tiporel arg2`, onde `arg1` e `arg2` são os argumentos da relação e podem ser representados por um ID (de uma EM) ou por um ID seguido de uma categoria (faceta).

Ilustramos a conversão da notação de alinhamentos na notação de triplas exemplificando a entrada e a saída do `Tradutor` nas figuras 5.19 e 5.20, respectivamente.

Neste caso a entrada do módulo é um ficheiro de alinhamentos com ID normalizados. Como se verifica, as relações representadas nos alinhamentos através dos atributos `COREL`, `TIPOREL`, `FAC_ORIGEM` e `FAC_ALVO` são transformadas em duas listas: as relações do lado da CD e as relações do lado da participação.

[8] Note-se que, nas participações, as relações nunca podiam estar definidas entre facetas, visto que isto foi algo que a organização marcou após a ocorrência da própria avaliação conjunta no sentido estrito.

```
DOC hub-66526
<EM ID="hub-66526-532" CATEG="VALOR" TIPO="QUANTIDADE">11</EM> ---> [<EM ID="hub-66526-532"
    CATEG="VALOR" TIPO="QUANTIDADE">11</EM>]::[Correcto]
<EM ID="hub-66526-538" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="CONSTRUCAO" COREL="hub-66526-556"
    TIPOREL="incluido" FACS_ORIGEM="LOCAL" FACS_ALVO="LOCAL">Biblioteca Pública Municipal do
    Porto</EM> ---> [<EM ID="hub-66526-538" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="CONSTRUCAO"
    COREL="hub-66526-540" TIPOREL="sede_de" FACS_ORIGEM="LOCAL" FACS_ALVO="LOCAL">Biblioteca
    Pública Municipal do Porto</EM>]::[Correcto]^0.5
<EM ID="hub-66526-540" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="DIVISAO" COREL="hub-66526-556"
    TIPOREL="incluido" FACS_ORIGEM="LOCAL" FACS_ALVO="LOCAL">Porto</EM> ---> [<EM ID="hub
    -66526-540" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="DIVISAO" COREL="hub-66526-538" TIPOREL="
    ocorre_em" FACS_ORIGEM="LOCAL" FACS_ALVO="LOCAL">Porto</EM>]::[Correcto]^0.5
<EM ID="hub-66526-25" CATEG="PESSOA" TIPO="INDIVIDUAL" COREL="hub-66526-560" TIPOREL="ident"
    FACS_ORIGEM="PESSOA" FACS_ALVO="PESSOA">D. Afonso Henriques</EM> ---> [<EM ID="hub
    -66526-25" CATEG="PESSOA" TIPO="INDIVIDUAL" COREL="hub-66526-560" TIPOREL="ident"
    FACS_ORIGEM="PESSOA" FACS_ALVO="PESSOA">D. Afonso Henriques</EM>]::[Correcto]
<EM ID="hub-66526-542" CATEG="ORGANIZACAO|LOCAL" TIPO="INSTITUICAO|HUMANO" SUBTIPO="|
    CONSTRUCAO" COREL="hub-66526-564␣hub-66526-542␣hub-66526-556␣hub-66526-568␣hub-66526-564␣
    hub-66526-568␣hub-66526-556␣hub-66526-564␣hub-66526-564␣hub-66526-542" TIPOREL="ident␣
    ocorre_em␣ocorre_em␣ocorre_em␣ocorre_em␣incluido␣incluido␣sede_de␣ident␣sede_de"
    FACS_ORIGEM="ORGANIZACAO␣ORGANIZACAO␣ORGANIZACAO␣ORGANIZACAO␣ORGANIZACAO␣LOCAL␣LOCAL␣
    LOCAL␣LOCAL␣LOCAL" FACS_ALVO="ORGANIZACAO␣LOCAL␣LOCAL␣LOCAL␣LOCAL␣LOCAL␣LOCAL␣ORGANIZACAO
    ␣LOCAL␣ORGANIZACAO">Santa Cruz</EM> ---> [<EM ID="hub-66526-542" CATEG="LOCAL|ORGANIZACAO
    |LOCAL" TIPO="HUMANO|INSTITUICAO|HUMANO" SUBTIPO="CONSTRUCAO||DIVISAO" COREL="hub
    -66526-564" TIPOREL="ident" FACS_ORIGEM="LOCAL|ORGANIZACAO" FACS_ALVO="LOCAL|ORGANIZACAO"
    >Santa Cruz</EM>]::[Correcto]
<EM ID="hub-66526-556" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="PAIS" COREL="hub-66526-568␣hub
    -66526-542␣hub-66526-538␣hub-66526-564␣hub-66526-540␣hub-66526-542␣hub-66526-564" TIPOREL
    ="inclui␣sede_de␣inclui␣inclui␣inclui␣inclui␣sede_de" FACS_ORIGEM="LOCAL␣LOCAL␣LOCAL␣
    LOCAL␣LOCAL␣LOCAL␣LOCAL" FACS_ALVO="LOCAL␣ORGANIZACAO␣LOCAL␣LOCAL␣LOCAL␣LOCAL␣ORGANIZACAO
    ">Portugal</EM> ---> [<EM ID="hub-66526-556" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="PAIS">
    Portugal</EM>]::[Correcto]
<EM ID="hub-66526-560" CATEG="PESSOA" TIPO="INDIVIDUAL" COREL="hub-66526-25" TIPOREL="ident"
    FACS_ORIGEM="PESSOA" FACS_ALVO="PESSOA">Afonso Henriques</EM> ---> [<EM ID="hub-66526-560
    " CATEG="PESSOA" TIPO="INDIVIDUAL" COREL="hub-66526-25" TIPOREL="ident" FACS_ORIGEM="
    PESSOA" FACS_ALVO="PESSOA">Afonso Henriques</EM>]::[Correcto]
<EM ID="hub-66526-563" CATEG="PESSOA" TIPO="INDIVIDUAL">Luís de Camões</EM> ---> [<EM ID="hub
    -66526-563" CATEG="PESSOA" TIPO="INDIVIDUAL">Luís de Camões</EM>]::[Correcto]
<EM ID="hub-66526-564" CATEG="ORGANIZACAO|LOCAL" TIPO="INSTITUICAO|HUMANO" SUBTIPO="|
    CONSTRUCAO" COREL="hub-66526-564␣hub-66526-556␣hub-66526-542␣hub-66526-568␣hub-66526-542␣
    hub-66526-542␣hub-66526-542␣hub-66526-556␣hub-66526-568␣hub-66526-564" TIPOREL="sede_de␣
    incluido␣sede_de␣incluido␣ident␣ocorre_em␣ident␣ocorre_em␣ocorre_em␣ocorre_em"
    FACS_ORIGEM="LOCAL␣LOCAL␣LOCAL␣LOCAL␣LOCAL␣ORGANIZACAO␣ORGANIZACAO␣ORGANIZACAO␣
    ORGANIZACAO␣ORGANIZACAO" FACS_ALVO="ORGANIZACAO␣LOCAL␣ORGANIZACAO␣LOCAL␣LOCAL␣LOCAL␣
    ORGANIZACAO␣LOCAL␣LOCAL␣LOCAL">Santa Cruz</EM> ---> [<EM ID="hub-66526-564" CATEG="LOCAL|
    ORGANIZACAO|LOCAL" TIPO="HUMANO|INSTITUICAO|HUMANO" SUBTIPO="CONSTRUCAO||DIVISAO" COREL="
    hub-66526-542" TIPOREL="ident" FACS_ORIGEM="LOCAL|ORGANIZACAO" FACS_ALVO="LOCAL|
    ORGANIZACAO">Santa Cruz</EM>]::[Correcto]
<EM ID="hub-66526-568" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="DIVISAO" COREL="hub-66526-564␣hub
    -66526-556␣hub-66526-542␣hub-66526-542␣hub-66526-564" TIPOREL="inclui␣incluido␣inclui␣
    sede_de␣sede_de" FACS_ORIGEM="LOCAL␣LOCAL␣LOCAL␣LOCAL␣LOCAL" FACS_ALVO="LOCAL␣LOCAL␣LOCAL
    ␣ORGANIZACAO␣ORGANIZACAO">Coimbra</EM> ---> [<EM ID="hub-66526-568" CATEG="LOCAL" TIPO="
    HUMANO" SUBTIPO="DIVISAO">Coimbra</EM>]::[Correcto]
EOD
```

Figura 5.19: Entrada do `Tradutor de alinhamentos para triplas`.

```
DOC hub-66526
[CD]
hub-66526-556 LOCAL inclui hub-66526-568 LOCAL
hub-66526-556 LOCAL sede_de hub-66526-542 ORGANIZACAO
hub-66526-556 LOCAL inclui hub-66526-538 LOCAL
hub-66526-556 LOCAL inclui hub-66526-540 LOCAL
hub-66526-556 LOCAL inclui hub-66526-564 LOCAL
hub-66526-556 LOCAL inclui hub-66526-542 LOCAL
hub-66526-556 LOCAL sede_de hub-66526-564 ORGANIZACAO
hub-66526-542 LOCAL sede_de hub-66526-564 ORGANIZACAO
hub-66526-542 LOCAL incluido hub-66526-568 LOCAL
hub-66526-542 LOCAL sede_de hub-66526-542 ORGANIZACAO
hub-66526-542 LOCAL ident hub-66526-564 LOCAL
hub-66526-542 LOCAL incluido hub-66526-556 LOCAL
hub-66526-542 ORGANIZACAO ident hub-66526-564 ORGANIZACAO
hub-66526-542 ORGANIZACAO ocorre_em hub-66526-556 LOCAL
hub-66526-542 ORGANIZACAO ocorre_em hub-66526-568 LOCAL
hub-66526-542 ORGANIZACAO ocorre_em hub-66526-564 LOCAL
hub-66526-542 ORGANIZACAO ocorre_em hub-66526-542 LOCAL
hub-66526-25 PESSOA ident hub-66526-560 PESSOA
hub-66526-564 LOCAL incluido hub-66526-568 LOCAL
hub-66526-564 LOCAL sede_de hub-66526-542 ORGANIZACAO
hub-66526-564 LOCAL ident hub-66526-542 LOCAL
hub-66526-564 LOCAL incluido hub-66526-556 LOCAL
hub-66526-564 LOCAL sede_de hub-66526-564 ORGANIZACAO
hub-66526-560 PESSOA ident hub-66526-25 PESSOA
hub-66526-540 LOCAL incluido hub-66526-556 LOCAL
hub-66526-568 LOCAL inclui hub-66526-542 LOCAL
hub-66526-568 LOCAL inclui hub-66526-564 LOCAL
hub-66526-568 LOCAL incluido hub-66526-556 LOCAL
hub-66526-568 LOCAL sede_de hub-66526-542 ORGANIZACAO
hub-66526-568 LOCAL sede_de hub-66526-564 ORGANIZACAO
hub-66526-564 ORGANIZACAO ocorre_em hub-66526-556 LOCAL
hub-66526-564 ORGANIZACAO ocorre_em hub-66526-568 LOCAL
hub-66526-564 ORGANIZACAO ocorre_em hub-66526-564 LOCAL
hub-66526-564 ORGANIZACAO ident hub-66526-542 ORGANIZACAO
hub-66526-564 ORGANIZACAO ocorre_em hub-66526-542 LOCAL
hub-66526-538 LOCAL incluido hub-66526-556 LOCAL
[Part]
hub-66526-25 PESSOA ident hub-66526-560 PESSOA
hub-66526-564 LOCAL|ORGANIZACAO ident hub-66526-542 LOCAL|ORGANIZACAO
hub-66526-560 PESSOA ident hub-66526-25 PESSOA
hub-66526-540 LOCAL ocorre_em hub-66526-538 LOCAL
hub-66526-542 LOCAL|ORGANIZACAO ident hub-66526-564 LOCAL|ORGANIZACAO
hub-66526-538 LOCAL sede_de hub-66526-540 LOCAL
EOD
```

Figura 5.20: Saída do `Tradutor de alinhamentos para triplas`.

5.8.6 Véus para o ReRelEM

Este módulo é semelhante ao `Véus` do HAREM clássico, mas aplicado aos tipos de relação: tem como entrada um filtro constituído por um conjunto de tipos de relação, separadas pelo caracter ';'. As relações de tipos que não se encontrem no filtro são removidas tanto do lado da CD como do lado da participação. A saída deste módulo terá no cabeçalho não apenas a descrição do cenário selectivo do HAREM em que é feita a avaliação, mas também a descrição do cenário do ReRelEM, sendo, de resto, em tudo semelhante à saída do `Alinhamentos para triplas`.

5.8.7 Avaliador de relações

O `Avaliador de relações` compara as relações do lado da participação com as relações do lado da CD.

Antes da comparação ser feita, todas as relações da CD que especializam/explicitam a relação `outra` são convertidas outra vez nesta mais geral (ver secção 4.1.4), de acordo com a definição original da tarefa.

Em seguida, são consideradas correctas todas as relações que estão na participação e para as quais existe na CD uma relação compatível. Considera-se que duas relações são compatíveis se os seus argumentos forem iguais, as suas facetas forem compatíveis (ver secção 5.8.2.2) e se o seu tipo for igual.

O resultado da aplicação do `Avaliador` é um conjunto de relações correctas, espúrias e em falta, ao qual é anexado o número total de relações na CD (`Rels_CD`) e o número total de relações na participação (`Rels_Part`).

A figura 5.21 mostra um excerto da saída do avaliador de relações, com todas as relações de determinado documento avaliadas.

5.8.7.1 Resumidor das classificações do ReRelEM

O `Resumidor das classificações do ReRelEM` processa a saída do `Avaliador de relações` agregando os resultados para obter os resultados globais da avaliação do ReRelEM na participação. Na figura 5.22 encontra-se a título de exemplo uma saída deste módulo, onde se podem ver as métricas e medidas calculadas.

5.8.8 Gerador de resultados do ReRelEM

O `Gerador de resultados do ReRelEM` tem uma função equivalente à do `Gerador de resultados do HAREM` (secção 5.6.9), agrupando todos os resultados do ReRelEM num ficheiro em HTML com os sistemas participantes ordenados por medida F.

5.8.9 Visualizador de relações

O `Visualizador de relações` é um programa que desenha grafos de relações a partir de ficheiros anotados segundo as directivas do ReRelEM. Tem como entrada (se nada for especificado) um ficheiro de alinhamentos, como por exemplo a saída do `Organizador de ALT`, do `Expandidor de relações`, do `Seleccionador de alinhamentos` ou do `Normalizador de ID`. É também possível ter uma entrada sob a forma de ficheiro de relações representadas através de triplas (utilizando a opção `-entrada triplas`).

```
DOC hub-94570
[Rels_CD(33) Rels_Part(10)]
hub-94570-125 LOCAL ident hub-94570-128 LOCAL :: Aval(Relação Correcta)
hub-94570-128 LOCAL ident hub-94570-125 LOCAL :: Aval(Relação Correcta)
hub-94570-114 PESSOA ident hub-94570-115 PESSOA :: Aval(Relação Correcta)
hub-94570-115 PESSOA ident hub-94570-114 PESSOA :: Aval(Relação Correcta)
hub-94570-112 PESSOA incluido hub-94570-114 PESSOA :: Aval(Espuria)
hub-94570-112 PESSOA incluido hub-94570-115 PESSOA :: Aval(Espuria)
hub-94570-118 ORGANIZACAO ocorre_em hub-94570-131 LOCAL :: Aval(Espuria)
hub-94570-131 LOCAL sede_de hub-94570-118 ORGANIZACAO :: Aval(Espuria)
hub-94570-114 PESSOA inclui hub-94570-112 PESSOA :: Aval(Espuria)
hub-94570-115 PESSOA inclui hub-94570-112 PESSOA :: Aval(Espuria)
hub-94570-110 TEMPO incluido hub-94570-117 TEMPO :: Aval(Em Falta)
hub-94570-112 PESSOA ident hub-94570-114 PESSOA :: Aval(Em Falta)
hub-94570-112 PESSOA ident hub-94570-115 PESSOA :: Aval(Em Falta)
hub-94570-112 PESSOA outra hub-94570-113 PESSOA :: Aval(Em Falta)
hub-94570-112 PESSOA outra hub-94570-118 ORGANIZACAO :: Aval(Em Falta)
hub-94570-118 ORGANIZACAO outra hub-94570-112 PESSOA :: Aval(Em Falta)
hub-94570-118 ORGANIZACAO outra hub-94570-113 PESSOA :: Aval(Em Falta)
hub-94570-118 ORGANIZACAO outra hub-94570-114 PESSOA :: Aval(Em Falta)
hub-94570-118 ORGANIZACAO outra hub-94570-115 PESSOA :: Aval(Em Falta)
hub-94570-125 LOCAL outra hub-94570-116 PESSOA :: Aval(Em Falta)
hub-94570-113 PESSOA outra hub-94570-114 PESSOA :: Aval(Em Falta)
hub-94570-113 PESSOA outra hub-94570-115 PESSOA :: Aval(Em Falta)
hub-94570-113 PESSOA outra hub-94570-112 PESSOA :: Aval(Em Falta)
hub-94570-113 PESSOA outra hub-94570-118 ORGANIZACAO :: Aval(Em Falta)
hub-94570-128 LOCAL outra hub-94570-116 PESSOA :: Aval(Em Falta)
hub-94570-116 PESSOA outra hub-94570-128 LOCAL :: Aval(Em Falta)
hub-94570-116 PESSOA outra hub-94570-125 LOCAL :: Aval(Em Falta)
hub-94570-117 TEMPO outra hub-94570-114 PESSOA :: Aval(Em Falta)
hub-94570-117 TEMPO inclui hub-94570-111 TEMPO :: Aval(Em Falta)
hub-94570-117 TEMPO inclui hub-94570-110 TEMPO :: Aval(Em Falta)
hub-94570-117 TEMPO outra hub-94570-115 PESSOA :: Aval(Em Falta)
hub-94570-117 TEMPO outra hub-94570-112 PESSOA :: Aval(Em Falta)
hub-94570-114 PESSOA outra hub-94570-118 ORGANIZACAO :: Aval(Em Falta)
hub-94570-114 PESSOA ident hub-94570-115 PESSOA :: Aval(Em Falta)
hub-94570-114 PESSOA outra hub-94570-113 PESSOA :: Aval(Em Falta)
hub-94570-111 TEMPO incluido hub-94570-117 TEMPO :: Aval(Em Falta)
hub-94570-115 PESSOA outra hub-94570-118 ORGANIZACAO :: Aval(Em Falta)
hub-94570-115 PESSOA outra hub-94570-113 PESSOA :: Aval(Em Falta)
hub-94570-115 PESSOA ident hub-94570-114 PESSOA :: Aval(Em Falta)
EOD
```

Figura 5.21: Fragmento da saída do `Avaliador de relações`

```
Total na CD: 1626
Total identificadas: 722
Total correctamente identificadas: 409
Espúrios: 313
Em Falta: 1217
Precisão: 0.5664819944598338
Abrangência: 0.25153751537515373
Medida F: 0.34838160136286195
```

Figura 5.22: Fragmento da saída do `Resumidor das classificações do ReRelEM`

Como a entrada pode consistir num conjunto de relações ou alinhamentos para vários documentos, é também possível escolher a origem do grafo a visualizar, indicando o ID do documento e o lado (CD ou participação).

No grafo, as relações podem ser identificadas pela cor dos arcos: `ident` (verde), `inclui` (vermelho), `incluido` (violeta), `ocorre_em` (azul), `sede_de` (azul claro) e `outra` (preto).

As figuras 5.23 e 5.24 correspondem a um grafo do mesmo documento, antes da expansão (5.23) e depois de ser expandido (5.24).

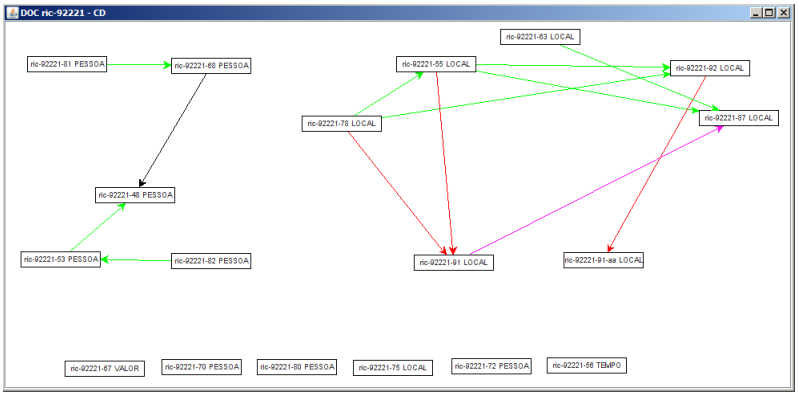


Figura 5.23: Grafo obtido através do `Visualizador de relações` sobre um conjunto de relações antes da expansão (saída do `Normalizador de ID`).

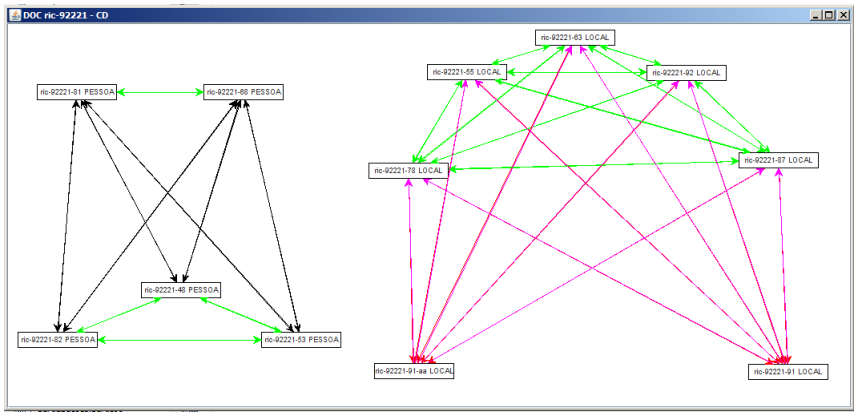


Figura 5.24: Grafo obtido através do `Visualizador de relações` sobre um conjunto de relações expandidas (saída do `Expandidor de relações`).

5.9 Observações finais

Todos os programas descritos podem ser invocados através de um serviço na rede, o SAHARA (descrito no apêndice G). Além disso, e como é norma da Linguateca, o có-

digo foi disponibilizado ao público. Todos os programas se encontram assim incluídos no pacote de recursos do Segundo HAREM, a LÂMPADA (disponível de http://www.linguateca.pt/HAREM/PacoteRecursosSegundoHAREM.zip), sob licença BSD.

Pensamos com esta postura aumentar o impacto e a utilidade da nossa actividade de organização de avaliações conjuntas, e contribuir para a melhoria do parque informático em termos de processamento do português.

Agradecimentos

Agradecemos ao Nuno Cardoso e ao Alberto Simões as valiosas sugestões de enriquecimento e clarificação do presente capítulo.

Capítulo 6

Segundo HAREM: Balanço e perspectivas de futuro

Diana Santos, Cláudia Freitas, Hugo Gonçalo Oliveira, Paula Carvalho e Cristina Mota

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 6, p. 131–146.

Neste capítulo fazemos o balanço do Segundo HAREM, tentando documentar tanto os pontos fortes desta avaliação como as questões que, na nossa perspectiva, não foram completa ou adequadamente resolvidas. Este exercício de reflexão é feito numa perspectiva de tentar projectar o que aprendemos para o futuro, no caso de vir a existir a oportunidade de organizar uma terceira edição desta avaliação conjunta.

Sempre que tal for pertinente, referiremos o balanço do Primeiro HAREM, documentado em Santos e Cardoso (2007b), para dar uma dimensão histórica em relação ao que foi conseguido e ao que ainda ficou por fazer.

Visto que as três pistas mereceram no livro capítulos distintos e trouxeram questões e problemas diversos, resolvemos apresentar um balanço por pista, ao contrário do que fizemos no encontro do Segundo HAREM (Santos et al., 2008e). Contudo, as questões que se refiram ao HAREM como um todo serão discutidas na primeira oportunidade.

Começamos por apresentar as questões que nos parecem dever ser melhoradas ou que não correram como esperado, relatando, em seguida, as que tiveram um desfecho a (nosso) contento. Terminamos com algumas perguntas e sugestões para o futuro, após uma caracterização crítica da participação neste Segundo HAREM.

6.1 HAREM clássico: balanço geral

Em relação ao HAREM clássico, não nos podemos esquivar à seguinte autocrítica: assumimos que todo o trabalho árduo de desbravamento do texto, com a criação das directivas e estabelecimento de categorias, já havia sido feito, e que portanto o tempo que levaria o processo de anotação da colecção dourada de uma segunda edição seria muito menor.

Contudo, novos textos (e uma nova equipe) tendem, sempre, a levantar novas dúvidas de anotação, e portanto levam a refinamentos e alterações nas directivas, e, consequentemente, a novas possibilidades de anotação. Com isso, lamentamos não ter havido tempo de produzir um documento único com as directivas do Segundo HAREM, onde seria possível encontrar tanto as características e categorias adoptadas do Primeiro HAREM (e, portanto, descritas no seu âmbito) como as introduzidas no Segundo HAREM, descritas no sítio do Segundo HAREM.

Outra questão que tem sido recorrente em todas as avaliações conjuntas que a Linguateca já organizou prende-se com a dificuldade de arranjar um esquema de classificação válido e consensual para o tipo e género de textos utilizados. Com efeito, ainda não foi desta que ficámos plenamente satisfeitos com o resultado obtido (e divulgado na LÂMPADA, o pacote de recursos do Segundo HAREM).

De qualquer forma, identificámos como especialmente problemáticas as seguintes questões, que discutimos separadamente em seguida:

- Peso da identificação em relação à classificação
- Delimitação das entidades mencionadas
- Dois modelos filosóficos opostos incluídos no HAREM

Iremos também aflorar a questão da utilização do XML, apresentando depois os aspectos positivos, tais como:

- Progresso na definição da tarefa

- Recursos mais ricos e mais bem documentados
- Véus mais bem aproveitados
- Relação explícita com outras tarefas
- Desenvolvimento de ferramentas para facilitar a avaliação conjunta

6.1.1 Identificação vs. classificação

Uma das questões que tentámos resolver e melhorar em relação ao Primeiro HAREM foi evitar a separação absoluta da identificação em relação à classificação, tornando a medida única e entrando em conta com esses dois aspectos como duas faces da mesma moeda (seguindo a proposta de Santos e Cardoso (2007b)). De um ponto de vista conceptual, também quisemos promover a "identificação" como a "classificação o mais vaga possível", ou seja, uma supercategoria das dez categorias usadas no Segundo HAREM, como explicado detalhadamente no capítulo anterior).

Contudo, e malgrado essa mudança, parece-nos que a identificação ainda teve um peso demasiado grande em relação à classificação, fazendo com que sistemas que se cingiram a identificar EM fossem superiores aos que também tentaram classificá-las. Por exemplo, o vencedor em termos de precisão na classificação para o cenário total, o SeRELeP, só fez identificação, o que é, no mínimo, pouco natural.

Além disso, logo que os participantes seleccionem um subconjunto de categorias (ou seja, concorram num cenário selectivo) estão implicitamente a classificar. Isso é de sobremaneira flagrante nos casos em que concorrem apenas numa única categoria, mas também se aplica quando competem num conjunto de categorias (e não em todas).

Isto leva-nos a concluir que, a não remover o premiar da identificação simples, deveríamos garantir que esse prémio fosse ínfimo comparado com a classificação. E, como trabalho futuro, aqui fica deixado o repto de estudar várias combinações possíveis de pesos para ver qual a combinação mais adequada.

6.1.2 Delimitação das entidades mencionadas

Para simultaneamente reduzir a importância das diferentes estratégias de identificação e garantir que as EM na colecção dourada estivessem bem delimitadas – ao contrário da CD do Primeiro HAREM, em que estavam a ser reconhecidas inadequadamente como EM sequências como, por exemplo, *de Planck*, como notado em Santos e Cardoso (2007b) – usámos a seguinte estratégia:

1. Procurámos todos os casos da CD em que havia palavras em minúsculas que achámos que deviam fazer parte da entidade (que formalmente corresponde a uma entidade complexa ou multipalavra);
2. Produzimos uma lista exaustiva dessas palavras (a lista completa encontra-se no apêndice A, secção A.6) e tornámo-la pública, de modo a que todos os participantes pudessem tê-las em conta no desenvolvimento dos seus sistemas;
3. Declarámos que quaisquer outros casos não deviam ser marcados, no sentido de restringir ao máximo a identificação de minúsculas, no âmbito da tarefa específica de REM, tal como é definida no modelo do HAREM.

Esta última declaração, embora tornasse a tarefa igual para todos, provocou muita confusão e descontentamento por parte dos participantes, sobretudo porque não podíamos explicar o verdadeiro motivo da escolha de tais elementos (que, reiterando, era a garantia de uma delimitação perfeita na CD), em detrimento de outros lexical e semanticamente próximos ou equivalentes.

Na verdade, do ponto de vista da avaliação, seria irrelevante o modo como os sistemas tratassem todos os outros casos não contemplados na CD. Assim, é fácil perceber, agora, que o terceiro passo não só era desnecessário como até seria prejudicial caso quiséssemos estender a colecção dourada (o que não veio a acontecer), pois isso implicaria uma actualização da dita lista.

Contudo, a nossa acção também não é totalmente indefensável: de facto, uma análise mais apurada indica que temos aqui uma tensão irredutível entre a) especificar a 100% uma tarefa para todos os participantes e b) produzir directivas linguisticamente apropriadas (ou mesmo simplesmente consensuais).

De facto, não existe uma descrição formal (no sentido de rigorosa, explícita e completa) do que é uma EM em português[1], e como organização do HAREM (tanto no Primeiro como neste Segundo), nós propusemos uma descrição (quase) meramente gráfica: *Uma EM deve conter pelo menos uma letra em maiúsculas, e/ou algarismos*, ver página 214 do capítulo 16 de Santos e Cardoso (2007a))

A outra alternativa, nomeadamente a de aceitar qualquer que fosse a delimitação que os (autores dos) sistemas achassem correcta, teria a desvantagem de deixar a tarefa mal definida, e implicar que, nessas condições, um sistema seria melhor ou pior pontuado dependendo da maior ou menor proximidade que tivesse relativamente à posição dos anotadores da CD.

6.1.3 Modelos de avaliação conjunta incongruentes entre si

Talvez a maior fraqueza deste HAREM tenha sido a coexistência, no seu seio, de tarefas concebidas em termos de filosofias diferentes de avaliação conjunta, dentro de uma mesma tarefa, o HAREM clássico (no que respeita à categoria TEMPO).

Com efeito, no modelo do Primeiro HAREM e que pretendemos continuar neste Segundo (ver Santos (2007d); Santos et al. (2008d)), é o contexto que decide a análise a atribuir a uma dada expressão e a análise é a da pessoa (ou do grupo) que anota a colecção dourada (sem quaisquer limitações ou simplificações), que tenta reproduzir fielmente a compreensão humana dos textos em questão, resultando assim num tecto, ou melhor, no alvo daquilo que idealmente se pretende obter, mas que pode ser quase impossível de automatizar.

No modelo de avaliação conjunta proposto para a categoria TEMPO, por outro lado, a tarefa apresenta-se como "capaz de ser executável em seis meses de desenvolvimento" (ver página 36 do capítulo 2), ou seja, não é para representar já toda a informação que um ser humano consiga identificar[2]. No capítulo respectivo, os autores mencionam, quer como guias iniciais quer como escolhas posteriores, critérios de "minimizar interacções complexas", deixando claramente questões mais espinhosas para mais tarde. Contudo, no

[1] Estamos a referir-nos à delimitação do texto que aponta/menciona a entidade mencionada, não à entidade "real" em si (ver figura 4.1 de Santos (2007d)).

[2] Os autores até falam em progressões em pequenos passos, ou seja, a sua abordagem foi definir primeiro tarefas que lhes pareciam mais fáceis.

nosso entender, ao simplificar a tarefa em alguns casos (como a introdução da preposição na expressão temporal independentemente do seu sentido) transformam-na em algo que não é consistentemente semântico, mas uma mistura de certa forma arbitrária entre vários tipos de considerações (sintácticas, semânticas, ...).

Não nos ficaria bem estar aqui a argumentar outra vez a favor de um modelo contra o outro (visto que o modelo do HAREM já foi defendido em Santos (2007d)). O que nos interessa sublinhar é aquilo que nos parece uma situação infeliz desta "incongruência" de modelos: como resultado desta situação, os recursos de avaliação – em particular, a colecção dourada do HAREM clássico, mas também a do ReRelEM – exibem categorias anotadas segundo filosofias diferentes. Uma que se pretende "derradeira" em termos de interpretação humana; outra que representa um primeiro passo numa sequência definida pelos autores da proposta, e, na nossa opinião, de forma relativamente arbitrária: basta pensar que o que pode ser simples para um sistema pode ser complicado de alcançar por outro, ou que essa sequência impõe restrições à forma como os sistemas são construídos.

Seria assim útil que uma nova anotação "derradeira" do TEMPO fosse levada a cabo (de acordo com a filosofia do HAREM), assim como seria também interessante proceder a uma anotação mais "fácil" e preliminar das outras categorias, aliás proposta também no encontro do Segundo HAREM (assim como no do Primeiro), em que os países fossem sempre considerados LOCAL, etc.

Em conclusão, esta é apenas uma observação sobre modelos de anotação incongruentes misturados nos mesmos recursos, sem tentar sequer escolher um deles.

Embora entrando no terreno da especulação, é possível que tenha sido aliás isso que levou a que as iniciativas para o inglês de anotação temporal (veja-se por exemplo Wilson et al. (2001); Pustejovsky et al. (2003)) fossem separadas do ACE, ao contrário do que aconteceu para o português no HAREM. O tempo o dirá se a interligação das duas comunidades (em vez da sua separação) trará vantagens para a nossa língua, ou se ambas as tarefas acabarão finalmente por divergir.

6.1.4 Novo formato XML

Uma das questões apontadas por vários participantes no Primeiro HAREM e que nos comprometemos a melhorar neste Segundo estava relacionada com o uso de XML, veja-se Martins e Silva (2007) e Almeida (2007). Mas o maravilhoso mundo da padronização parece melhor ao longe... De perto, descobrimos que há várias versões dos padrões, incompatíveis entre si, isto mesmo ao nível da visualização na rede.

De facto, tal "solução" acabou por criar mais problemas do que resolveu: não só foi preciso reformatar as antigas CD (que disponibilizámos para treino), como levou a que parte significativa dos programas tivesse de ser reescrita.

É no entanto preciso reconhecer que podemos ter sido demasiado cautelosos na migração para XML, tendo dado lugar a um híbrido ainda mais difícil de caracterizar e processar. Em particular, parece-nos agora que a nossa notação dos ALT deveria ter sido generalizada à vagueza da classificação, no sentido de que A|B (sintaxe do Primeiro HAREM), passada agora para CATEG="A|B" no Segundo HAREM, deveria sim ter sido transformada em algo como uma das seguintes alternativas:

1. <ALT ID="x">
   <ALTN><EM CATEG="OBRA">...</EM></ALTN>

```
<ALTN><EM CATEG="LOCAL">...</EM></ALTN>
</ALT>
```

2.
```
<ALT>
<ALTN><EM ID="x" CATEG="OBRA"></ALTN>
<ALTN><EM ID="y" CATEG="LOCAL"></ALTN>
</ALT>
```

No entanto, a primeira alternativa não permitia atribuir diferentes identificações (ID) a alternativas com diferentes delimitações, enquanto a segunda parecia impor diferentes identificações ao que nós reputamos uma **única** EM.

Outra questão que terá de ser melhor equacionada é o uso de espaços para permitir mais de um valor no mesmo atributo (usado no ReRelEM), e que não é boa prática.

Confessamos, assim, a necessidade de mais uma ronda para definir, em XML, SGML ou ainda outro formalismo, uma representação que seja simultaneamente adequada e fácil de processar, abarcando a marcação da vagueza tanto na classificação como na identificação, assim como a problemática do encaixe, ou seja, da recursividade na definição das EM.

6.1.5 Progresso na definição da tarefa e nos desafios

Contudo, de uma forma geral, pensamos que é indiscutível que o Segundo HAREM representou um claro progresso em relação ao Primeiro, sob várias perspectivas.

Em primeiro lugar, algumas das limitações detectadas foram colmatadas, levando a que a medida de avaliação fosse melhor motivada (ver Santos et al. (2008d) e sobretudo o capítulo 5), aliás em paralelo com a especificação da tarefa: `CATEG` e `TIPO` passaram a ser opcionais e EM uma marcação a nível mais elevado, como já mencionado.

Outra questão que facilita a referência posterior e a discussão de exemplos concretos é o facto de cada EM ter um identificador único.

Também nos orgulhamos, do ponto de vista linguístico, de termos avançado significativamente na descrição das EM em português, em particular na especificação das combinações de `ALT` sistemáticas, apresentadas no apêndice D. Estabelecemos assim uma primeira lista, com base em corpos, de combinações entre EM de diversos tipos que nos pareceram produtivas em português.

Isto relaciona-se com o facto de termos alargado a interpretação dos `ALT`, que passaram a identificar consistentemente todas as EM possíveis e não apenas a maior. Além disso, a avaliação dos `ALT` deixou de ser feita por critérios quantitativos cegos em termos de fracção do número de palavras coincidentes, como acontecia no Primeiro HAREM (Santos et al., 2007), para passar a sê-lo em termos do conteúdo previamente anotado.

Finalmente, o termos facultado material de treino para as três tarefas também permitiu uma definição mais clara do que se esperava dos sistemas.

Contudo, não queremos de forma alguma dar a ideia de termos coberto todos os pormenores ou questões levantadas pela análise do material. Em particular, há duas áreas em que estamos conscientes de que é preciso mais trabalho, nomeadamente o que respeita a coordenação de EM, e a classificação de entidades que se refiram a meios de comunicação social.

Tabela 6.1: Contabilização da informação adicional no Segundo HAREM, nas três pistas

Número de subtipos marcados	2480 [7769]
Número de ALT	255
Número de EM de TEMPO só em minúsculas	426
Número de EM do TEMPO com atributos do tempo estendido	192
Número de atributos do TEMPO estendido	372
Número de relações semânticas (entre facetas)	614

6.1.6 Recursos mais ricos, mais bem revistos e documentados

Talvez o resultado com maior impacto deste Segundo HAREM sejam os recursos linguísticos criados, que passaram por um crivo apertado quer da equipe da organização quer, em alguns casos, dos próprios participantes.

Em primeiro lugar, houve uma grande preocupação de fundamentação da tarefa e de melhoria e correcção dos problemas identificados nas CD do Primeiro HAREM.

Depois, podemos afirmar que as CD foram muito bem revistas (sob vários ângulos e por várias pessoas), e que as opções tomadas (após a definição das tarefas) foram bem documentadas. De facto, houve muita revisão e consideração das divergências, linguísticas e de interpretação, que os textos suscitaram, tendo-se procedido aliás à compilação de diversa informação para estudos futuros, relativa a dúvidas e discordâncias, e que foi na sua maioria tornada pública na LÂMPADA. A maior parte dos casos problemáticos, aliás mais uma vez seguindo as recomendações do Primeiro HAREM (Santos e Cardoso, 2007b), foi marcada (em 68 casos) como OMITIDO, de forma a não basearmos a comparação dos sistemas em casos desviantes.

Outra melhoria relativamente evidente, mas que nos parece útil não passar em branco neste balanço, é o facto de termos marcado os recursos com bastante mais informação do que no Primeiro HAREM.

Estamos a referir-nos não só ao facto de as categorias LOCAL e TEMPO terem subtipos que foi necessário preencher, como ao facto de que, devido às novas directivas do TEMPO terem um critério muito mais abrangente (não exigirem algarismos ou nomes próprios), ter havido um número muito maior de EM a classificar. Na tabela 6.1 apresentamos uma contabilização da informação adicional presente no conjunto das três CD como um todo.[3]

6.1.7 Cenários selectivos melhor aproveitados

Outra melhoria a que já fizemos menção no capítulo anterior é termos levado até às últimas consequências a questão dos cenários selectivos neste Segundo HAREM, ou seja, foi finalmente implementada a visão dos cenários selectivos (realizados pelo Véus) como constituindo ontologias distintas.

Assim, além de ser possível, como no Primeiro HAREM, comparar cada sistema segundo as suas próprias condições, os chamados cenários selectivos de avaliação possibilitaram a comparação dos vários sistemas entre si, ao fornecer a possibilidade de ver **todos** os sistemas segundo todos os ângulos individuais (representados pelos cenários selectivos de participação de cada sistema).

[3] No caso dos subtipos, o número dentro de parêntesis rectos indica o número de subtipos se se tiver em conta a vagueza.

A Figura 6.1 ilustra o caso de um sistema que participa num determinado cenário (o seu cenário selectivo de participação, à direita) a ser avaliado noutro cenário de avaliação (proposto por outro sistema, por exemplo, ou considerado de interesse por outra razão – indicado pelas caixas de contorno sólido na ontologia que se encontra do lado esquerdo). De acordo com esse cenário de avaliação, o sistema seria apenas avaliado relativamente aos elementos da ontologia que têm igualmente contorno sólido.

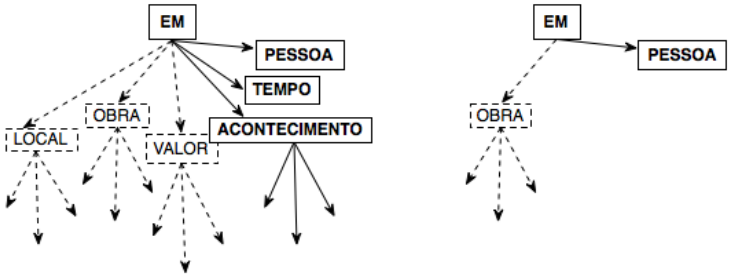


Figura 6.1: Cenários selectivos de participação e de avaliação vistos como alinhamento de ontologias

6.1.8 Potencialidades de investigação do valor do REM noutras áreas

Além de um recurso valioso para REM, a constituição da nova colecção do HAREM e dos resultados dos sistemas – disponibilizados na LÂMPADA – permite fazer investigação com base num recurso comum em várias áreas afins, nomeadamente:

- recolha de informação geográfica (RIG)
- resposta automática a perguntas (RAP)

Como brevemente referido no capítulo 1, a maior parte dos textos que constituem a colecção foram retirados da colecção CHAVE (Santos e Rocha, 2005), melhor dizendo do monte de documentos atribuído aos tópicos do GeoCLEF 2007 (Mandl et al., 2008).[4] Criámos assim uma colecção em que, para cada tópico, os sistemas anotaram com EM (e talvez também com relações) todos os documentos relevantes, assim como alguns dos irrelevantes que pertenciam ao monte. Tal colecção permite medir, de forma rigorosa, a (ir)relevância do REM para RIG, se se contabilizar as ocorrências relevantes para a tarefa em jogo.

Uma questão semelhante põe-se em relação à influência ou necessidade de identificar EM nas perguntas, no âmbito da resposta automática a perguntas. Com efeito, é costume assumir que o REM aplicado ao texto das perguntas é uma vantagem óbvia em RAP, muitas vezes trocando a compreensão da pergunta pela própria resposta, já conhecida do sistema.

[4] *Monte* em RI é um conjunto de documentos que foram considerados relevantes para responder a um determinado tópico por um grupo de sistemas participantes numa avaliação conjunta, conjunto esse que é verificado por juízes humanos, que atribuem o julgamento relevante/não relevante a cada documento do monte. Veja-se Rocha e Santos (2007a) e Gonzalez et al. (2007).

Quisemos neste HAREM investigar a possibilidade de anotar as perguntas sem contexto (assumindo ignorância total em relação à resposta), não só para ver até onde conseguíamos ir, como para fornecer exemplos mais diversificados de casos onde os sistemas REM são chamados a dar uma classificação, como é o caso do QA@CLEF (Giampiccolo et al., 2008). Os resultados da anotação humana estão na colecção dourada e podem ser investigados por todos os interessados nesta problemática. (Trabalho relacionado no âmbito da RAP é por exemplo Roberts e Hickl (2008), que definem uma hierarquia complexa de tipos de respostas.)

6.1.9 Ferramentas para auxiliar o HAREM

Pensamos também ser necessário salientar que o trabalho de organização do Segundo HAREM deu origem a uma maior e mais sofisticada panóplia de ferramentas, sistemas e serviços que foram disponibilizados aos participantes e ao público em geral.

Em primeiro lugar, foi posto à disposição de todos os participantes (e do público em geral) um validador (como serviço na rede) que permitia o teste atempado da sintaxe das saídas dos sistemas e das suas consequentes participações no HAREM.[5]

Em segundo lugar, todo o processo de envio de participações para o Segundo HAREM também foi monitorizado e apoiado por um sistema automático.[6]

Em terceiro lugar, após essa ideia ter surgido durante o Encontro do Segundo HAREM, foi desenvolvido um sistema na rede que permite fazer o teste e avaliação posterior (de acordo com as CD do Segundo HAREM) de novas participações (não oficiais).[7]

Finalmente, e como já referido, foi tornada pública uma ferramenta para ajudar à edição e criação de colecções douradas, o Etiquet(H)AREM, que foi, além disso, usada no âmbito da criação das próprias CD.[8]

6.2 Pista do TEMPO: algumas observações

Apesar de no capítulo 2 ter sido feito um balanço desta pista pelos proponentes, julgamos ser necessário chamar a atenção para outros aspectos ligados à vertente organizativa que nos coube em mãos.

Já foi identificado como um aspecto problemático neste Segundo HAREM (cf. capítulo 3) o facto de os proponentes da pista do TEMPO não terem levado a cabo a própria criação dos recursos e dos programas de avaliação, o que trouxe à organização do HAREM não só bastante trabalho adicional, mas sobretudo bastante insegurança sobre a própria anotação realizada.

Não vamos portanto referir outra vez esta questão, a não ser para indicar que tal situação deverá ser evitada de futuro. De facto, por muita boa vontade que um determinado grupo, neste caso a Linguateca, possa ter em prestar um bom serviço à comunidade, parece-nos pouco apropriado (digamos até mesmo desmotivante) criar recursos de avaliação e fazer escolhas que vão contra as nossas próprias opiniões, como aconteceu com esta pista.

Consideramos, pois, que os proponentes de pistas independentes da pista geral (em termos de filosofia e objectivos a atingir) deverão: ou deixar a organização ter a última

[5] Este validador foi desenvolvido por David Cruz e Luís Miguel Cabral.
[6] O sistema foi desenvolvido por Luís Miguel Cabral.
[7] Este sistema foi desenvolvido por Nuno Cardoso.
[8] Essa ferramenta foi desenvolvida por Hugo Gonçalo Oliveira.

palavra na decisão dos critérios, ou serem eles a tomar em ombros toda a sua organização (nomeadamente no que se refere à criação e anotação de recursos linguísticos e criação e implementação dos próprios programas de avaliação).

Dito isto, gostávamos de mais uma vez salientar que tal não é uma crítica aos três proponentes do grupo do TEMPO – que desde o início, aliás, nos disseram claramente que seriam participantes – e aos quais estamos gratos por esta colaboração. Esta conclusão é simplesmente fruto de uma experiência que não poderíamos ter, nem eles, antes de a levar a cabo.

Pensamos ser contudo inegável que a pista do TEMPO enriqueceu o HAREM e que levou a um progresso notável na descrição das expressões temporais e sua normalização em português.

6.3 ReRelEM: primeiro balanço

Se apreciarmos agora a pista do ReRelEM, temos de concordar a posteriori com a Renata Vieira de que enveredámos por uma tarefa demasiado ambiciosa no âmbito de um piloto.

Com efeito, carregámos com as complexidades inerentes ao HAREM, sem tentar proceder a qualquer tipo de simplificação: o `ALT`, a vagueza, os cenários selectivos do HAREM, etc. e definimos sobre essa tarefa, já de si complexa, uma outra completamente nova.

Além disso, tivemos de comparar participantes muito diversos. Diríamos mesmo completamente distintos (ver capítulo 4). Enquanto o REMBRANDT seguiu à risca o que esperávamos, os outros dois sistemas participantes divergiram substancialmente, e de forma completamente inesperada para a organização. Por um lado, o SeRELeP não enviou a classificação (apenas tentou identificar as relações); por outro, o SEI-Geo escolheu apenas uma relação (`inclui`), competindo portanto num cenário selectivo do ReRelEM, e além disso sem identidade.[9]

Com esta variedade de participação, cedo nos demos conta de que a forma de avaliação que tínhamos inicialmente proposto era demasiado ingénua e simples, e que muitas outras questões tinham de ser equacionadas. Revemos assim aqui as principais questões discutidas entre os membros da organização[10], algumas das quais ainda sem resposta.

6.3.1 A expansão das participações

Ao aplicarmos as regras associadas a cada relação à participação de um dado sistema, podemos estar a desdobrar os seus erros em muito mais relações erradas. Será que vale a pena fazer a avaliação também sem expansão, assumindo que apenas o que é explicitamente marcado pelo sistema deve ser pontuado, ou esperando que os sistemas tenham, eles próprios, o seu mecanismo de expansão? Após acesa discussão, acabámos por considerar que, de acordo com os pressupostos do HAREM, em que o que interessa é a semântica, teríamos de expandir (levar às últimas consequências em termos de compreensão do texto) tanto as corridas dos sistemas como a CD.

[9] De facto, ao conceber o ReRelEM como uma extensão de detecção de co-referência, ou seja, marcação de identidade entre EM, não tínhamos considerado sequer a possibilidade de haver sistemas que não tratassem primeiro da identidade.

[10] Convém ressalvar que esta discussão se deu internamente e não conjuntamente com os participantes, como seria desejável, uma vez que decorreu durante o processo de anotação da CD e de desenvolvimento dos programas de avaliação.

6.3.2 Relação com a vagueza

Embora inicialmente tivéssemos falado em relações entre EM, demo-nos conta de que diferentes facetas de uma EM vaga poderiam entrar em relações distintas com outra EM, e que a única forma de ter tudo convenientemente anotado era, no caso das EM vagas, descer ao nível da faceta durante a anotação da CD do ReRelEM.

Mencionamos novamente esta problemática, já discutida no capítulo 4, porque deu origem à repetição total do trabalho de anotação da CD do ReRelEM, contrastando com o tratamento dos `ALT`, referido nas próximas linhas.

Além disso, pôs de certa forma em causa precisamente a nossa escolha inicial de ter um único ID por EM e não por faceta, o que é algo que terá de ser mais bem pensado em futuras edições, além de ter exigido mais um conjunto de ferramentas para lidar com esta situação.

6.3.3 O que fazer aos ALT?

Embora tenhamos conseguido um processo satisfatório, ainda que trabalhoso, de lidar com a vagueza da classificação no ReRelEM, o mesmo não aconteceu em relação aos `ALT`. Ou seja, não respondemos ainda de forma conclusiva à pergunta: como é que a formulação de alternativas de identificação (que muitas vezes também redundam em mudanças de classificação) interage com a especificação de relações?

A única coisa que nos pareceu sempre clara é que não fazia sentido a declaração de relações entre alternativas de um mesmo `ALT`. Mas a interacção entre um texto marcado com `ALT` e a formulação de relações entre essas EM e o resto do texto não foi ainda considerada seriamente de um ponto de vista linguístico, e constitui naturalmente trabalho de reflexão futuro.

De facto, neste primeiro ReRelEM limitámo-nos a aceitar como certas todas as relações, independentemente de estarem "repetidas" dentro de diferentes alternativas ou não. Ou seja, se *Universidade de Lisboa* aparecesse como EM duas vezes em alternativas diferentes dentro de um mesmo `ALT`, e se em ambas as vezes estivesse relacionada com outra EM (por exemplo *Universidade*, algumas frases mais tarde) essa relação seria contada como certa, ou como errada, duas vezes.

6.3.4 O que fazer a participações inconsistentes?

Outra questão que se nos pôs foi como lidar com a marcação de relações que, depois de expandidas, levassem a uma contradição.

Neste primeiro ReRelEM, para o cálculo dos resultados ignorámos completamente esse aspecto (simplesmente fazendo a expansão enquanto for possível), mas estamos plenamente conscientes de que ainda falta especificar o que fazer, e como pontuar, nesses casos.

De momento apenas conseguimos postular que as EM em questão deviam ser consideradas negativamente para atribuir a classificação à tarefa do ReRelEM, mas, naturalmente, é necessário precisar a forma como isso deve ser feito, e desenvolver mais uma ferramenta auxiliar de detecção de inconsistências entre relações marcadas[11].

[11] Uma funcionalidade embrionária já se encontra no `Expandidor`, que detecta alguns tipos de inconsistências se invocado com a opção `-ver_inconsistencias`.

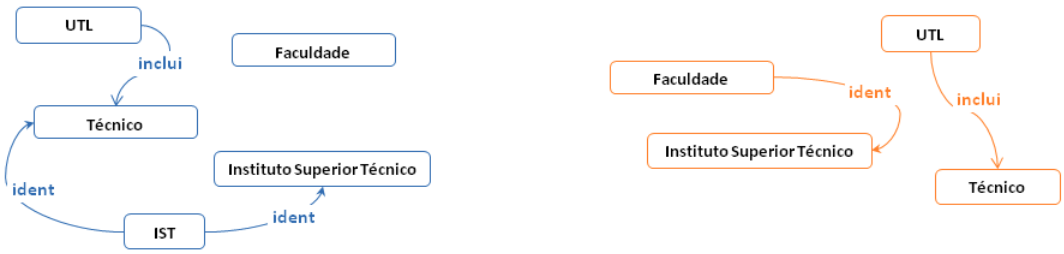


Figura 6.2: Exemplo de relações na CD e numa participação

6.3.5 Que sentido faz a comparação?

Embora tal seja relativamente evidente, pôs-se-nos com mais acuidade a pergunta fundamental de como comparar o incomparável: se um sistema só procura reconhecer a identidade e outro só reconhece a localização, o que têm os dois em comum? Ou melhor, o que têm as duas relações em comum para poderem ser comparadas?

O problema aqui é ainda mais agudo do que no HAREM em geral, porque as próprias relações podem implicar graus de complexidade muito diferentes.

6.3.6 A identidade é diferente?

Finalmente, uma das considerações que nos tomou mais tempo, e à qual acabámos por não dar seguimento, foi a intuição de que a relação de identidade era diferente e devia ser separadamente analisada, antes de pontuar as outras relações. Com base nessa ideia, aplicámos uma medida de avaliação do agrupamento[12] obtido a partir das relações de identidade, para medir o desempenho de um sistema quanto ao reconhecimento da identidade (por outras palavras, para medir os grupos obtidos a partir das relações de identidade propostas pelo sistema comparando-os com os grupos obtidos a partir da colecção dourada).

O problema é que, depois desse primeiro agrupamento (em que substituíamos as EM pelo grupo a que pertenciam), não conseguimos fazer sentido das relações (que não a identidade) propostas pelos sistemas, e compará-las com as relações na CD.

Veja-se a figura 6.2, com um exemplo de relações marcadas na CD e numa participação fictícia.

A figura 6.3 apresenta o resultado do agrupamento para o caso da figura 6.2. Se substituirmos as EM por um representante do agrupamento, como representar, e pontuar, por exemplo, a relação entre *UTL* e o *Técnico* proposta pelo sistema? De facto, como ilustra a figura 6.4, o agrupamento contituído por *Técnico* na participação faz parte do agrupamento que está envolvido na relação de inclusão na CD, o que poderia ser suficiente para considerar a relação como correcta. No entanto, em vez de um agrupamento estar incluído no outro, poderíamos ter uma sobreposição parcial entre agrupamentos. Continuando com a nossa participação fictícia, imagine-se que o sistema tinha estabelecido a relação *UTL* `inclui` *Instituto Superio Técnico* em vez de *UTL* `inclui` *Técnico*. Será que nesse caso a relação também deve ser considerada correcta?

[12] Agrupamento é a nossa tradução para o termo inglês *clustering*.

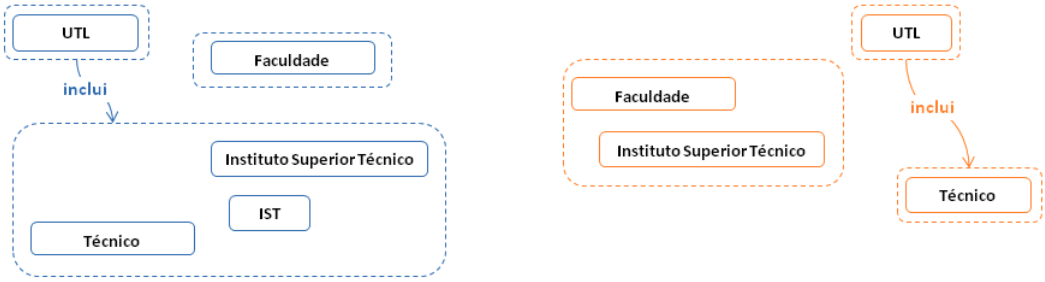


Figura 6.3: Exemplo de agrupamento na CD e numa participação

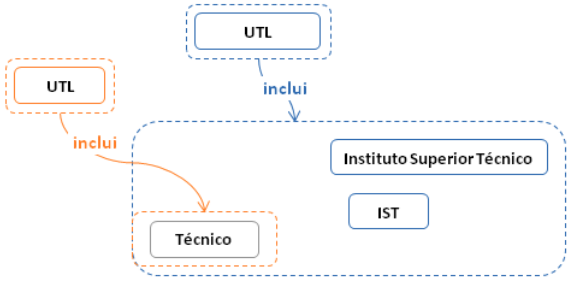


Figura 6.4: Tentativa de avaliação de relações (que não a identidade) com base em agrupamentos

Não tendo conseguido, de facto, arranjar uma solução satisfatória para o emparelhamento dos grupos e para a consequente avaliação das relações (não identidade) entre eles, acabámos por desistir de tratar de forma diferente a identidade.

Contudo, se a única relação do ReRelEM fosse esta (encontrando-nos portanto em presença de uma avaliação de co-referência simples), as medidas do agrupamento ainda nos pareceriam uma boa alternativa à opção tomada, e colocamo-las aqui na figura 6.5 para delas dar conta aos leitores.

6.3.7 Progresso na área da semântica computacional

Embora tendo tropeçado em várias dificuldades imprevistas, o que se reflectiu no atraso da publicação dos resultados (pode dizer-se que durante alguns meses fomos aumentando a sofisticação do tratamento das corridas quase de semana para semana), não podemos deixar de nos orgulhar por termos proposto, e avaliado, uma tarefa mais complicada do que qualquer outra por nós conhecida em termos de avaliação conjunta de qualquer língua.

Estamos convencidos de que a tarefa exploratória do ReRelEM que levou à explicitação dos tipos de relações entre EM é, do ponto de vista linguístico e computacional, inovadora (veja-se o capítulo 4), assim como as decisões de avaliação, embora preliminares.

Também produzimos material que é interessante estudar em profundidade, e até cruzar informação entre as várias pistas ou tarefas, embora o tamanho do recurso dourado para o ReRelEM seja inquestionavelmente pequeno.

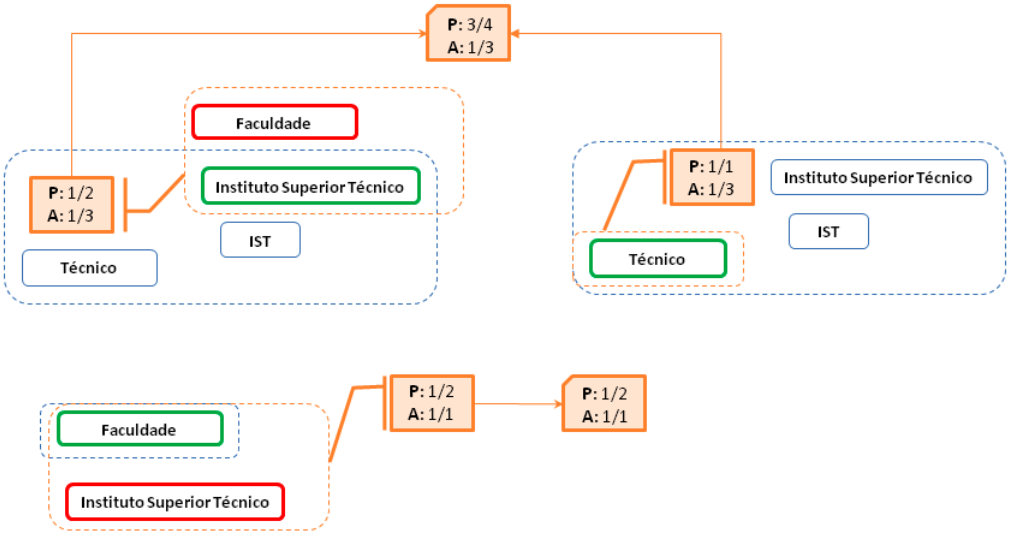


Figura 6.5: Classificação da medida de agrupamento (avaliando só identidade)

6.4 O HAREM tem futuro?

Nesta secção, um pouco distinta das anteriores, porque olhamos para o futuro e não para o passado, tentamos equacionar as vantagens de um terceiro HAREM e as recomendações que podemos deixar para os futuros organizadores, como foi feito em Santos e Cardoso (2007b).

A primeira pergunta a formular é: existe uma comunidade de REM em português que pretende continuar? Ou de cada vez que uma nova avaliação conjunta ocorre aparecem maioritariamente novos grupos, porque os anteriores já progrediram para outras tarefas?[13] Dado que não houve quase nenhuma sobreposição entre o Primeiro e o Segundo HAREM em termos da massa dos participantes, a pergunta parece-nos pertinente. De certa forma, o facto de termos tido vários novos participantes (e até bastante interesse internacional, que infelizmente depois não se concretizou em termos de participação), é cancelado pelo facto de que muitos antigos participantes não responderam à chamada.

Contudo, e dada a conclusão (em ambos os Encontros) de que a maioria dos grupos estaria interessada em continuar, seria necessário organizar uma terceira edição de avaliação para tirar a prova.

Muito brevemente, façamos contudo uma panorâmica da comunidade que respondeu presente neste HAREM, para indagar, pelo menos parcialmente, quais os objectivos científicos de cada participante:[14]

- a comunidade de RI(G) parece ser a mais estável, com três participantes (Cage, SEI-Geo e REMBRANDT, embora só o primeiro tenha participado no Primeiro HAREM)

[13] De acordo com a formulação do Marcirio Chaves, porque consideram que a área do REM já se encontra resolvida para o português.

[14] Não estamos naturalmente a dizer – o que seria trivialmente falso, dado o HAREM – que estas comunidades não se intersectam! Pelo contrário, a maior parte dos sistemas pode dizer-se que pertence a mais do que uma.

- a comunidade de resposta automática a perguntas (RAP) teve grande protagonismo e dos melhores resultados (Priberam e XIP-L2F/XEROX)
- a comunidade do processamento do tempo juntou-se ao HAREM (XIP-L2F/XEROX e PorTexto)
- a comunidade de extracção de informação (EI) também se mostrou interessada, com quatro participantes (REMMA, R3M, DobREM e SEI-Geo)
- a comunidade da co-referência também participou com bons resultados, embora só com um participante (SeRELeP)

Primaram pela ausência, contudo, alguns dos principais actores na semântica computacional do português, e em particular ambos os vencedores da primeira edição, o PALAVRAS-NER e o Cortex. Embora ambos tenham invocado falta de tempo, ou melhor, outras prioridades, é significativo que não achassem que valia a pena tornar a participar. De facto, apenas um sistema repetiu a participação, o CaGe, embora dois participantes antigos aparecessem com novos sistemas (SEI-Geo e R3M), o que é naturalmente positivo.

Outra questão que nos preocupa é a cada vez menor participação de grupos brasileiros nas actividades da Linguateca. Mais uma vez só tivemos um grupo do Brasil[15], apesar de termos sempre o maior cuidado em organizar as avaliações de forma a que as duas variantes estivessem igualmente bem representadas. Isto pode dever-se ao calendário limitado com que fomos forçados a organizar este Segundo HAREM, mas é algo que devemos considerar com mais cuidado em futuras avaliações.

Além disso, é preciso também reconhecer que a participação nas discussões de preparação ficou muito aquém das nossas expectativas. A maior parte dos participantes não quis fazer jus à caracterização "conjunta", que devia ser parte integrante de uma avaliação conjunta, preferindo aceitar as regras (de qualquer das tarefas) sem debate. Muito provavelmente seria necessária uma reunião presencial (como foi o caso nas Morfolimpíadas (Costa et al., 2007)) para pôr todas as pessoas à volta de uma mesa a discutir casos concretos.

Obviamente que outra pergunta associada a uma eventual continuação do HAREM é a de identificar diferentes alternativas de continuidade. Visto que todos os recursos são finitos, porque não organizar (ou participar, conforme o ponto de vista) uma avaliação noutra área? Por exemplo integrando como parte ou constituinte o próprio REM, mas não o fazendo o objecto principal... Ou seja, porque não fazer o processo inverso do MUC, que começou com extracção de informação em geral e especializou para tarefas mais delimitadas? O HAREM poderia ter começado com essas tarefas mais delimitadas e desenvolver no sentido de extrair mais informação, em forma de gabaritos (as "templates" do MUC).

Ainda outra forma de evoluir/mudar o enquadramento do HAREM seria acoplá-lo ou associá-lo a uma avaliação internacional que contivesse mais línguas. Nesse caso a comunidade óbvia seria o CLEF (Rocha e Santos, 2007a; Braschler e Peters, 2004).

Vamos contudo nas linhas que se seguem assumir que irá existir um Terceiro HAREM – ainda e só para o português – e cujo foco seja ainda a classificação de EM e de relações entre elas, para podermos fornecer algumas recomendações para a futura edição, com base na nossa experiência:

[15] No Primeiro HAREM, tivemos apenas um grupo brasileiro, o CorTex, que foi aliás o vencedor do Mini-HAREM, embora inicialmente mais grupos tenham indicado interesse, aliás como na presente edição.

- Parece-nos aconselhável manter as tarefas do HAREM clássico apenas com modificações pontuais (marcando-as claramente nas directivas anteriores, mas refazendo-as e publicitando-as com tempo para haver uma discussão até presencial das mesmas) para permitir uma comparação de progresso do Segundo para o Terceiro HAREM.
- Uma reunião presencial de discussão, ou mesmo várias, parece obviamente importante para obter um consenso inicial, assim como esclarecer muitas coisas que podem não ser óbvias a participantes pela primeira vez.
- Sugerimos que os participantes sejam envolvidos na escolha dos textos que pertencem à colecção do Terceiro HAREM (embora a escolha dos textos da CD tenha de ser secreta e feita pela organização), para permitir que essa colecção responda aos interesses de investigação da comunidade.
- Interessaria obter uma "garantia" de participação, ou um prémio de participação, que diminuísse o grau de desistência dos participantes inscritos. Uma hipótese de "garantia" poderia ser o enviarem-nos uma versão do seu sistema que seria usado se não conseguissem participar à última hora.

Seja como for, não nos parece necessário nem apropriado começar desde já a organizar uma nova iniciativa neste campo.

De facto, estamos conscientes de que os próprios resultados postos à disposição de toda a comunidade permitem, ou mesmo exigem, estudos aprofundados, que vão desde validação estatística a comparação entre as duas edições do HAREM, antes de ser apropriada a organização de uma nova edição.

Esperamos por isso que muitos investigadores possam beneficiar do trabalho já feito e identificar questões interessantes em relação ao processamento semântico da nossa língua, além do mero treino e desenvolvimento de melhores sistemas para esta tarefa específica. Tanto do lado mais linguístico da descrição da língua, como do lado mais computacional do desenvolvimento de ferramentas para explorar os recursos complexos criados pela anotação humana, como do lado da metodologia de avaliação e da reflexão sobre as conclusões estatitisticamente válidas sobre o desempenho dos sistemas, muito ainda há para fazer.

Agradecimentos

Agradecemos a Jorge Baptista, Marcirio Chaves e Mírian Bruckschen os comentários a versões anteriores deste capítulo, que nos ajudaram a melhorá-lo significativamente.

Parte II

O HAREM pelos participantes

Capítulo 7

O sistema CaGE no Segundo HAREM

Bruno Martins

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 7, p. 149–158.

Os documentos textuais (por exemplo, artigos publicados em jornais ou páginas na rede) são muitas vezes ricos em informação geográfica. A utilização de técnicas de prospecção de texto para a extracção desta informação, por forma a introduzir capacidades de raciocínio geográfico em sistemas de recuperação de informação, é um problema interessante que tem vindo a ganhar notoriedade (Purves e Jones, 2007).

As técnicas de reconhecimento de entidades mencionadas (EM) encontram na recuperação de informação geograficamente contextualizada uma natural área de aplicação. Contudo, mais do que anotar uma expressão de texto como uma localização, esta área de aplicação requer que as anotações produzidas contenham uma desambiguação completa dos nomes de locais. Por outras palavras, as referências geográficas devem ser associadas explicitamente a coordenadas de latitude e longitude, ou a identificadores de conceitos num almanaque geográfico. Esta informação (ou seja, as coordenadas na superfície terrestre, ou o almanaque geográfico em conjunto com os documentos anotados) pode então ser utilizada noutras tarefas de recuperação de informação, tais como a pesquisa de documentos de acordo com os seus âmbitos geográficos.

O sistema CaGE surgiu no contexto de um trabalho de doutoramento que aborda o problema do reconhecimento e desambiguação de nomes de locais, argumentando que esta é uma tarefa crucial na geo-codificação de documentos textuais (Martins, 2009). O principal objectivo do sistema CaGE é atribuir âmbitos geográficos (ou seja, a área geográfica que o documento descreve como um todo) a documentos textuais, tendo-se que numa versão mais recente do sistema, este objectivo estende-se também aos âmbitos temporais. O sistema foi já usado como módulo independente em vários projectos relacionados com recuperação de informação geograficamente contextualizada, nomeadamente no GREASE (Silva et al., 2006) e no DIGMAP (Borbinha et al., 2007; Martins et al., 2008). O CaGE também participou na primeira edição da avaliação conjunta HAREM (ou seja, no Primeiro HAREM e no Mini-HAREM) com o objectivo de avaliar o seu desempenho em cenários selectivos focados no reconhecimento de EM com categoria LOCAL (Martins e Silva, 2007). Embora os objectivos e os critérios associados ao sistema CaGE se distanciem consideravelmente daqueles que foram considerados no HAREM, um correcto tratamento das referências geográficas depende, em grande medida, da capacidade do sistema em reconhecer as EM da categoria LOCAL, tal como estas foram definidas no contexto das regras semânticas utilizadas na avaliação conjunta.

No Segundo HAREM, o sistema CaGE participou num cenário de avaliação mais abrangente, o qual correspondeu ao reconhecimento de entidades das categorias PESSOA, ORGANIZACAO e TEMPO, e ao reconhecimento e classificação em tipos e subtipos de entidades da categoria LOCAL. Visto que as referências geográficas são muitas vezes ambíguas em relação a entidades de outras categorias (por exemplo, nomes próprios de pessoas que correspondem também a nomes de locais), temos que a consideração destas outras categorias por parte do sistema CaGE pode ajudar naquele que é o seu objectivo principal.

Este capítulo descreve a participação do sistema CaGE na segunda edição da avaliação conjunta HAREM. É feita uma descrição detalhada do sistema e é apresentado o contexto no qual o mesmo foi desenvolvido. São discutidos os resultados obtidos e são ainda listadas as principais melhorias que se pretendem introduzir em desenvolvimentos futuros do sistema.

7.1 Descrição do sistema

O CaGE é um sistema híbrido apoiado por dicionários e regras de desambiguação. As subsecções que se seguem detalham o seu funcionamento, apresentando ainda os dicionários usados pelo sistema.

Estudos anteriores indicam que o problema do reconhecimento de EM pode ser abordado de forma bastante eficaz através do uso de métodos de aprendizagem automática (McCallum e Li, 2003). Contudo, para o caso específico do reconhecimento e desambiguação completa de entidades geográficas, é necessária a utilização de um recurso de informação externo (ou seja, um almanaque geográfico), visto que as referências devem ser inequivocamente associadas a uma representação única para o conceito geográfico que lhes está subjacente (ou seja, coordenadas de latitude e longitude ou identificadores no almanaque geográfico).

7.1.1 Os dicionários e o almanaque usados pelo sistema CaGE

No que diz respeito aos dicionários usados no reconhecimento das EM correspondentes às categorias PESSOA e ORGANIZACAO, assim como das entidades do tipo *"período temporal"* correspondentes à categoria TEMPO, foram usados os seguintes recursos lexicais para a sua construção:

- A base de dados de nomes de entidades denominada REPENTINO, um acrónimo de *REPositório para o reconhecimento de ENTIdades com NOme*[1] (Sarmento et al., 2006).
- Nomes de pessoas listadas na Internet Movie Database (IMDB[2]).
- Nomes de autores listados na base de dados de autoridades bibliográficas PORBASE[3].
- Listas de períodos temporais e de nomes próprios comuns extraídas da Wikipédia.
- Traduções para português dos nomes de períodos temporais definidos no contexto do ECAI Time Period Directory (Petras et al., 2006).
- Dicionários distribuídos com um sistema de reconhecimento de entidades de código aberto (em inglês, *open source*) para a língua inglesa denominado BALIE, acrónimo de *BAseLine Information Extraction* (Nadeau, 2007).

É de salientar que alguns dos recursos lexicais que foram considerados apresentam uma percentagem elevada de nomes em outras línguas que não o português, particularmente nomes na língua inglesa. No entanto, apesar do HAREM apenas utilizar textos em português, nada impede que neles sejam mencionados nomes próprios provenientes de outras línguas (tal como por exemplo nomes próprios de actores de cinema norte-americanos).

No que diz respeito aos dicionários usados no reconhecimento de entidades da categoria LOCAL, foram usados os seguintes recursos lexicais para a sua construção:

[1] http://www.linguateca.pt/REPENTINO/
[2] http://www.imdb.com/interfaces/
[3] http://www.porbase.org/

- As versões portuguesa e multilingue do almanaque geográfico GeoNET-PT01 (Chaves et al., 2005b).
- A base de dados de nomes de locais disponibilizada pelo serviço GeoNames[4].
- A lista de nomes do almanaque geográfico usado no projecto DIGMAP. Uma descrição detalhada deste almanaque é dada por Manguinhas et al. (2008).

O CaGE faz ainda uso de um dicionário de excepções para entidades do tipo local, o qual foi construído manualmente com base em ensaios com o sistema. Este dicionário inclui nomes de entidades que, embora tenham uma conotação geográfica, são maioritariamente usados noutros contextos.

Para a desambiguação completa das EM que correspondem a locais ou a períodos temporais, é ainda utilizado um almanaque mais específico para este tipo de informação, o qual foi desenvolvido no contexto do projecto DIGMAP. A figura 7.1 ilustra os conceitos principais que lhe estão subjacentes.

O almanaque DIGMAP associa os nomes de locais aos conceitos geográficos que lhes estão subjacentes (ou seja, os mesmos locais podem ser associados a vários nomes), definindo ainda a cobertura geo-espacial (por exemplo, coordenadas de latitude e longitude) para cada conceito geográfico, assim como uma hierarquia de relações de inclusão entre os conceitos geográficos que traduz uma organização administrativa da superfície terrestre. O mesmo almanaque define ainda conceitos temporais correspondentes a períodos históricos, listando para cada um deles os nomes que lhes estão associados.

7.1.2 Funcionamento geral do sistema

De um ponto de vista algorítmico, o sistema CaGE assenta numa sequência de operações de processamento com quatro etapas principais:

Etapa 1 : Identificação inicial das entidades mencionadas

a. Os textos são inicialmente atomizados através do algoritmo que é fornecido com as bibliotecas da linguagem Java para o processamento de texto, mais concretamente na classe `java.text.BreakIterator`. Este algoritmo funciona com base numa tabela contextual de pares de caracteres (Gillam, 1999). A separação nos diferentes átomos é determinada com base nos caracteres que ocorrem em ambos os lados de uma dada posição no texto (por exemplo, a tabela para atomização em palavras indica uma separação entre caracteres de pontuação e letras, mas não entre letras consecutivas).

b. Os átomos identificados no texto são percorridos em sequência com um algoritmo do tipo *"janela de análise deslizante"*, por forma a extrair o conjunto de sequências de palavras que ocorrem no texto. Como resultado deste passo, são identificadas todas as sequências de palavras contendo um máximo de seis elementos (ou seja, n-gramas de palavras com $1 \leq n \leq 6$). Uma consequência deste passo é que as entidades mencionadas no texto que tenham um comprimento superior a seis palavras serão ignoradas pelo sistema CaGE.

[4] http://www.geonames.org

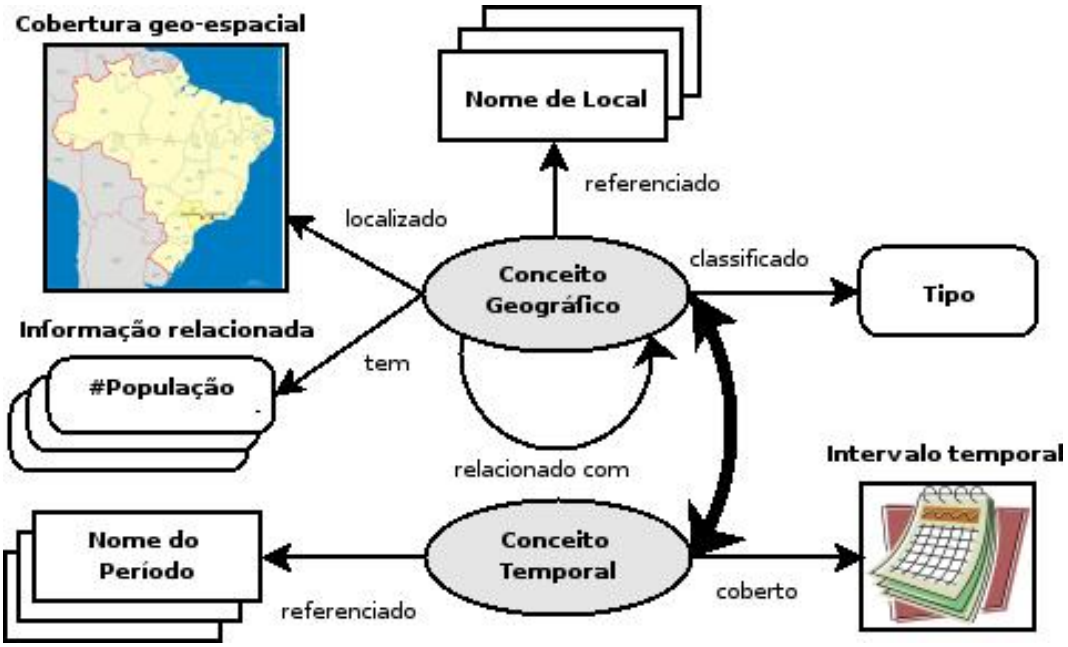


Figura 7.1: Os principais elementos de informação no almanaque DIGMAP

c. As sequências de palavras identificadas no passo anterior são filtradas, por forma a registar as sequências de palavras que começam com letras maiúsculas e que não ocorrem desta forma apenas no início de frases. Este passo implementa uma heurística que diz que as entidades são normalmente mencionadas com a primeira letra em maiúscula.

d. É feita uma pesquisa nos dicionários usados pelo sistema, por forma a mapear as sequências de palavras resultantes do passo anterior com as entidades que lhes correspondem. No caso de existirem mapeamentos diferentes para uma dada sequência e para as suas subsequências, é apenas registado o mapeamento mais geral, ou seja aquele que corresponde à sequência de maior tamanho. Uma vez que para cada sequência de n palavras são também testados mapeamentos para as duas subsequências de $n-1$ palavras, podem, por exemplo, ocorrer dois mapeamentos na expressão textual *Pedrógão Pequeno* (ou seja, um para a sequência *Pedrógão* e outro para a sequência *Pedrógão Pequeno*). No caso do exemplo apresentado, seria apenas considerado o mapeamento mais geral, ou seja *Pedrógão Pequeno*.

e. No caso de sequências de palavras mapeadas com uma entidade da categoria `LOCAL`, é feito um mapeamento adicional usando o dicionário de excepções. No caso de ser registada a ocorrência de um caso de excepção, é removido o mapeamento entre a sequência de palavras e a entidade de categoria `LOCAL`.

f. São usadas expressões regulares para identificar entidades da categoria `TEMPO` que não se encontram definidas nos dicionários (por exemplo, várias formas de

expressar datas de calendário).

Etapa 2 : Classificação das entidades mencionadas e tratamento da ambiguidade

a. No caso das entidades identificadas nos dicionários, para as quais foram registados vários mapeamentos possíveis, são usadas regras de classificação desenvolvidas manualmente para encontrar a categoria e tipo de entidade correspondentes. Estas regras são baseadas na ocorrência de palavras-chave no contexto textual da entidade (ou seja, os dois átomos que ocorrem no texto, antes ou depois da entidade em questão). Por exemplo, uma das regras usadas no CaGE corresponde ao padrão `cidade de [EM] -> LOCAL-CIDADE`, indicando que todas as EM que são precedidas das palavras *cidade de* devem ser classificadas com a categoria `LOCAL` e o tipo `CIDADE`.

b. Para as entidades que permanecem ambíguas após a execução do passo anterior, é feita uma classificação por escolha circular (em inglês, *round-robin classification*) entre as várias categorias e tipos possíveis para a entidade em questão (Fürnkranz, 2002). O argumento por detrás desta estratégia é o de que, escolhendo uma entidade diferente em cada situação ambígua e ir sequencialmente percorrendo o conjunto de atribuições possíveis, minimiza-se o número de erros introduzidos pelo sistema.

Etapa 3 : Desambiguação completa de entidades geográficas e temporais

a. Para cada entidade da categoria `LOCAL` identificada na segunda etapa, é feita uma pesquisa no almanaque DIGMAP, por forma a associar as entidades aos conceitos geográficos subjacentes. Esta pesquisa combina o nome mencionado no texto com o tipo associado à entidade, caso este tenha sido já resolvido com base num mapeamento com um dicionário.

b. No caso de a pesquisa ao almanaque, efectuada no passo anterior, retornar vários conceitos geográficos possíveis, estes são ordenados de acordo com uma heurística do tipo "*um sentido por omisso*" a qual diz que, na maior parte dos casos, uma dada referência geográfica encontra-se associada a um único tipo em concreto (o nome *Lisboa* é mais frequentemente usado como uma referência à cidade capital do país do que a uma pequena vila). A utilização da heurística "*um sentido por omisso*" encontra-se descrita em maior detalhe em Martins et al. (2008).

c. Para as entidades que permanecem ambíguas após o passo anterior (ou seja, as entidades que correspondem a diferentes conceitos no almanaque geográfico), é ainda usada uma heurística do tipo "*referentes relacionados por cada unidade de discurso*" por forma a melhorar a ordenação dos conceitos geográficos correspondentes à entidade. Caso exista uma relação hierárquica entre um dos conceitos possíveis para a entidade em questão, e os outros conceitos mencionados no documento, então é dada uma maior importância a esse conceito aquando da sua ordenação. Mais uma vez, o mecanismo subjacente à heurística "*referentes relacionados por cada unidade de discurso*" encontra-se descrito em maior detalhe em Martins et al. (2008).

d. Para as entidades correspondentes a períodos temporais identificadas na primeira etapa, é feita uma pesquisa no almanaque DIGMAP por forma a associar a entidade ao conceito temporal que lhe está subjacente.

Etapa 4 : Atribuição de âmbitos geográficos e temporais aos documentos

a. É atribuído um âmbito geográfico à totalidade do documento com base na combinação de todas as referências geográficas identificadas no texto. Esta atribuição é feita com base no algoritmo originalmente proposto por Amitay et al. (2004), o qual assenta na utilização de uma hierarquia de relações de inclusão entre os conceitos geográficos reconhecidos no texto. O almanaque DIGMAP é usado como fonte de dados para estas relações.

b. É atribuído um âmbito temporal ao documento com base no intervalo mínimo que cobre todas as referências temporais identificadas no texto.

7.1.3 Aplicações práticas do sistema CaGE

Como resultado das quatro etapas de processamento apresentadas atrás, tem-se que o sistema CaGE permite não só reconhecer e classificar entidades mencionadas em textos, como também desambiguar as entidades correspondentes a referências geográficas ou temporais. Finalmente, o sistema CaGE suporta ainda a atribuição de âmbitos geográficos e temporais aos documentos como um todo, combinando a diferente informação extraída do texto. Mais detalhes sobre os aspectos relacionados com a desambiguação completa de referências geográficas e temporais, assim como sobre a atribuição de âmbitos, podem ser consultados em Martins et al. (2008).

Um serviço na rede (em inglês, *Web service*) com base no sistema CaGE encontra-se disponível para utilização online, mais concretamente no URL `http://geoparser.digmap.eu`. Este serviço oferece uma interface XML através da qual se pode invocar o reconhecimento e desambiguação das EM num dado documento textual. A interface XML segue, em linhas gerais, uma proposta do Open Geospatial Consortium para a implementação de serviços na rede para o geo-processamento de recursos textuais (Lansing, 2001). Uma vez que a saída do serviço é um documento XML bem formado, torna-se relativamente simples desenvolver outros serviços que explorem a informação extraída dos documentos. A figura 7.2 mostra o ecrã principal de uma aplicação na rede que utiliza o serviço do CaGE por forma a reconhecer e desambiguar referências geográficas em canais de notícias RSS, possibilitando a exploração das notícias sobre um mapa dinâmico.

O almanaque geográfico desenvolvido no contexto do projecto DIGMAP, o qual é usado na desambiguação completa de entidades da categoria `LOCAL`, encontra-se também acessível através de um serviço na rede, mais concretamente através do URL `http://gaz.digmap.eu`. Este serviço permite a realização de consultas sobre o almanaque, as quais podem combinar aspectos tais como o nome dos locais, os seus tipos, e a sua localização sobre a superfície terrestre. A interface deste serviço segue o formato XML proposto no contexto do Alexandria Digital Library Gazetteer (Hill et al., 1999).

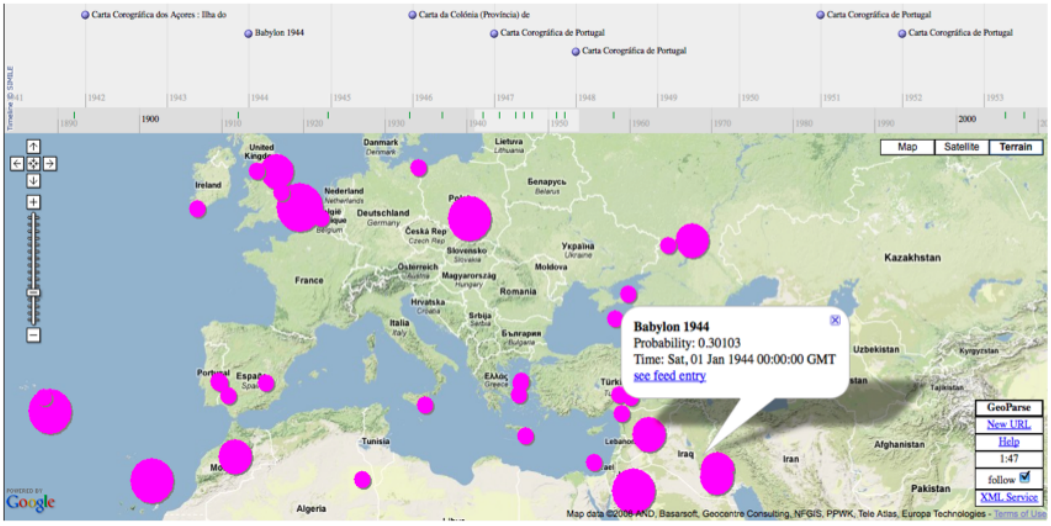


Figura 7.2: Um serviço na rede para exploração de feeds RSS, construído com base no CaGE

7.2 Experiências no HAREM e análise dos resultados

Tal como mencionado anteriormente, o sistema CaGE participou no Segundo HAREM apenas num cenário selectivo, o qual corresponde ao reconhecimento de entidades da categoria PESSOA, ORGANIZACAO e TEMPO, e ao reconhecimento e classificação em tipos e subtipos de entidades da categoria LOCAL (exceptuando-se o subtipo VIRTUAL). Para as entidades da categoria TEMPO, apenas se trataram os casos de datas absolutas (*15 de Abril de 1979*) e nomes de períodos temporais (por exemplo, *idade média*).

O HAREM aborda apenas o reconhecimento e classificação das EM, não considerando o problema da desambiguação completa das entidades correspondentes a locais ou a períodos temporais. A etapa 4 do algoritmo usado pelo CaGE, assim como os passos b) e c) da etapa número 3, não foram portanto utilizados no contexto da participação no HAREM.

Foram enviadas quatro corridas ao Segundo HAREM, as quais correspondem aos cenários experimentais que se encontram descritos abaixo:

1. Utilização dos vários dicionários, assim como do almanaque DIGMAP, para desambiguação de entidades da categoria LOCAL. A classificação em tipos e subtipos das entidades da categoria LOCAL foi feita com base no almanaque DIGMAP (ou seja, as entidades do tipo LOCAL tinham de estar forçosamente definidas no almanaque DIGMAP).

2. Utilização dos vários dicionários, exceptuando-se o dicionário de casos de excepção.

3. Utilização dos dicionários contendo nomes de locais e períodos temporais, excluindo-se os restantes dicionários. Nesta corrida, eram apenas reconhecidas as entidades das categorias TEMPO e LOCAL.

Tabela 7.1: Resultados obtidos na classificação de entidades mencionadas no cenário selectivo 2

Corrida	Posição	Precisão	Abrangência	Medida F
4	14	0,4264	0,4070	0,4164
1	16	0,4277	0,4025	0,4148
2	17	0,4226	0,4059	0,4141
3	20	0,3883	0,3500	0,3682
Melhor corrida	1	0,7347	0,5893	0,6325

Tabela 7.2: Resultados obtidos na identificação de entidades mencionadas no cenário selectivo 2

Corrida	Posição	Precisão	Abrangência	Medida F	TotalEMCD	TotalEMSis
4	16	0,4615	0,4553	0,4584	5538,3	5463,5
1	17	0,4643	0,4520	0,4581	5538,3	5391,5
2	18	0,4576	0,4547	0,4562	5538,3	5503,5
3	20	0,4225	0,3929	0,4072	5538,3	5151,2
Melhor corrida	1	0,8561	0,7127	0,6813		

4. Utilização de todos os dicionários sem quaisquer restrições adicionais (ou seja, as entidades do tipo LOCAL podiam estar definidas no almanaque DIGMAP ou num dos restantes dicionários).

As tabelas 1 e 2 apresentam, respectivamente, os resultados obtidos na classificação e identificação de entidades para o cenário selectivo 2 do Segundo HAREM, o qual considerava várias categorias de entidades (ou seja, LOCAL, TEMPO, ORGANIZACAO e PESSOA) assim como a classificação em tipos para as entidades da categoria LOCAL. Na tabela 2, as colunas TotalEMCM e TotalEMSis representam, respectivamente, o número total de entidades existente na colecção dourada e o número total de entidades retornado pelo sistema.

A corrida número 4 foi a que obteve melhores resultados, correspondendo a uma diferença de cerca de 0,2 em termos da medida F para com a melhor corrida neste mesmo cenário.

As tabelas 3 e 4 apresentam, respectivamente, os resultados obtidos na classificação e identificação de entidades para o cenário selectivo 5 do Segundo HAREM, o qual considera apenas entidades da categoria LOCAL, exceptuando-se ainda as entidades do tipo VIRTUAL. Estes resultados são ligeiramente superiores aos obtidos no cenário selectivo 2. Mais uma vez, a corrida número 4 foi a que obteve o melhor resultado, o qual corresponde a uma diferença de aproximadamente 0,1 em termos da medida F para com a melhor corrida neste mesmo cenário.

Comparando com os resultados obtidos pelo sistema CaGE na anterior edição do Mini-HAREM, em condições semelhantes àquelas que são usadas no cenário selectivo 5 do Segundo HAREM, temos que os resultados obtidos pela corrida número 4 são ligeiramente inferiores (ou seja, uma diferença de aproximadamente 0,1 em termos da medida F).

Em suma, o sistema CaGE obteve resultados modestos na sua participação no Segundo HAREM. Embora os dicionários utilizados pelo sistema apresentem uma cobertura adequada (por exemplo, para o caso das entidades da categoria LOCAL, tem-se que os dicionários utilizados pelo CaGE listam cerca de dois milhões de nomes diferentes), as regras e heurísticas utilizadas pelo sistema carecem ainda de alguma optimização.

Tabela 7.3: Resultados obtidos na classificação de entidades mencionadas no cenário selectivo 5.

Corrida	Posição	Precisão	Abrangência	Medida F
4	11	0,5267	0,5844	0,5540
2	12	0,5196	0,5851	0,5504
1	13	0,5147	0,5802	0,5455
3	14	0,5178	0,5754	0,5451
Melhor corrida	1	0,7080	0,70236	0,6246

Tabela 7.4: Resultados obtidos na identificação de entidades mencionadas no cenário selectivo 5.

Corrida	Posição	Precisão	Abrangência	Medida F	TotalEMCD	TotalEMSis
4	11	0,5198	0,6788	0,5888	1418	1851,5
2	12	0,5091	0,6802	0,5823	1418	1894,5
1	13	0,5049	0,6781	0,5788	1418	1904,5
3	14	0,5084	0,6689	0,5777	1418	1865,5
Melhor corrida	1	0,7186	0,7856	0,6572		

7.3 Conclusões

Neste capítulo foi descrito o sistema CaGE para reconhecimento de entidades geográficas, assim como a sua adaptação para a participação no Segundo HAREM e os respectivos resultados obtidos nesta avaliação conjunta. Muito embora se tenham obtido resultados relativamente modestos, a participação no HAREM foi bastante útil, tendo permitido já detectar e corrigir diversas falhas existentes no sistema.

Tal como na primeira edição do HAREM, existe um desfasamento considerável entre os critérios e os objectivos estabelecidos para o sistema CaGE e aqueles que são contemplados pela avaliação conjunta. Embora o sistema CaGE tente fazer reconhecimento de entidades no contexto em que elas são mencionadas, este não tem como objectivo o reconhecimento da função das entidades no texto (por exemplo, os casos de metonímia, tais como no exemplo *Portugal pronunciou-se...*, são sempre marcados como entidades de uma mesma categoria, neste caso `LOCAL` e não `PESSOA`). Este desfasamento explica, em parte, os modestos valores alcançados pelo sistema em termos das diversas métricas de avaliação consideradas.

Como principais desafios de trabalho futuro, há a considerar a melhoria das regras de classificação de entidades, assim como um tratamento mais profundo das referências temporais, segundo as directivas genéricas da pista do TEMPO tal como definidas para esta edição do HAREM. Este último ponto é particularmente interessante para o caso de aplicações de recuperação de informação, uma vez que a extracção da dimensão temporal dos documentos, assim como a sua combinação com a dimensão geográfica, pode suportar mecanismos de recuperação de informação mais adequados a alguns domínios.

Capítulo 8

PorTexTO: sistema de anotação/extracção de expressões temporais

Olga Craveiro, Joaquim Macedo e Henrique Madeira

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 8, p. 159–170.

As técnicas de recolha de informação assumiram um papel de grande relevo nos últimos anos, em virtude da importância assumida pelos motores de busca na Internet. No entanto, a utilização de informação temporal para melhorar os resultados das pesquisas tem sido pouco explorada, apesar de existir um grande potencial para conseguir esse melhoramento. De facto, a noção de tempo é essencial para muitas das pesquisas efectuadas num sistema de recolha de informação, como por exemplo, na área da saúde onde será pertinente reconstruir o historial clínico dos pacientes com a capacidade de encontrar eventos e apresentá-los num espaço temporal, permitindo, desta forma, dar maior exactidão ao relatório (Alonso et al., 2007). Outro dos exemplos da aplicação da dimensão temporal nos sistemas de recolha de informação é dado pelo trabalho desenvolvido por Uehara e Sato (2005) na implementação de arquivos www.

No entanto, nem sempre o tempo surge de forma explícita nos documentos, mas as referências temporais podem ajudar a identificar a relevância dos documentos encontrados. Beigbeder (2004) apresenta um estudo de como os diferentes aspectos temporais podem intervir nas diversas etapas da recolha de informação. O interesse no processamento de informação temporal tem crescido nos últimos anos e tem-se intensificado nas mais diversas áreas de investigação (Mani et al., 2004; Pustejovsky et al., 2005).

O objectivo do sistema que desenvolvemos é o de identificar informação temporal existente em documentos, para posteriormente ser utilizada, como papel importante na ordenação da lista de resultados obtida pelas pesquisas efectuadas em sistemas de recolha de informação.

Como existe ainda pouco trabalho desenvolvido no processamento de informação temporal da língua portuguesa, decidimos criar um sistema de raiz que seguisse um algoritmo simples e rápido. O processo de anotação/extracção de informação num sistema de recolha de informação terá de ser rápido para que não comprometa todo o sistema.

Pretende-se que o sistema PorTexTO, designado por *PORtuguese Temporal EXpressions Tool*, seja um sistema simples e com baixo tempo de processamento. Para que o desempenho não seja comprometido, o sistema poderá não encontrar todas as expressões temporais existentes nos documentos que processar, mas deverá encontrar as que ocorram mais vezes nos documentos em português e que são definidas através de estudo estatístico.

O PorTexTO faz um processamento frase a frase dos documentos, e utiliza padrões de expressões temporais para o reconhecimento das entidades mencionadas, ao contrário de outros sistemas de processamento de linguagem natural onde o processamento é feito termo a termo, identificando cada termo segundo as suas características linguísticas (Mani, 2004).

Na detecção de expressões temporais na língua inglesa existem diversas abordagens, embora não seja do nosso conhecimento que alguma utilize padrões de expressões criados com recurso às co-ocorrências das palavras temporais. Uma das abordagens apresentada por Mani e Wilson (2000) utiliza uma anotação manual num conjunto de teste e um conjunto de regras obtidas por aprendizagem automática. Makkonen e Ahonen-myka (2003) fazem uma divisão dos termos temporais em categorias e utilizam autómatos de estados finitos no reconhecimento das expressões. Outra abordagem que também utiliza autómatos de estados finitos foi apresentada por Schilder e Habel (2001), mas que introduziram preposições nos seus autómatos. Esta abordagem tem por base o trabalho de Allen (1983) na detecção de intervalos temporais.

A implementação do sistema PorTexTO seguiu as directivas gerais e as directivas do TEMPO (Hagège et al., 2008) publicadas para o Segundo HAREM.

Neste capítulo é descrito o sistema PorTexTO, sendo apresentadas em pormenor as várias etapas de processamento dos dois módulos do sistema (`Anotador` e `Processador de co-ocorrências`), a sua participação no Segundo HAREM e é ainda efectuada uma análise aos resultados obtidos. Por fim, são apresentadas as conclusões e o trabalho futuro.

8.1 Descrição do sistema

O sistema PorTexTO é um sistema de reconhecimento de entidades mencionadas temporais em textos na língua portuguesa. Neste sistema, o processamento dos documentos é efectuado frase a frase, ou seja, o texto é previamente dividido em frases, com a ajuda do atomizador da Linguateca (o módulo de Perl `Lingua::PT::PLNbase`[1]) passando em seguida cada frase pelas diversas etapas de processamento.

A identificação das expressões temporais é feita usando padrões de expressões, criados a partir de co-ocorrências existentes em referências temporais. Os padrões encontram-se armazenados num ficheiro para que facilmente possam ser acrescentados novos padrões ou alterados os existentes. Este ficheiro foi designado por `REGEX` e encontra-se representado na figura 8.1.

O PorTexTO permite o processamento de documentos tanto em formato de texto simples não estruturado como em formato estruturado em XML. O resultado produzido pelo sistema pode ser um ficheiro no seu formato original, mas com as devidas anotações nas expressões temporais encontradas ou então um ficheiro com todas as expressões temporais encontradas e a sua posição relativamente ao texto original. Como o formato pretendido pelo Segundo HAREM era o do ficheiro original devidamente anotado, só será abordado neste capítulo o módulo que produz este resultado e que foi designado por módulo `Anotador`. A figura 8.1 apresenta a arquitectura deste módulo.

Este módulo tem como entrada o texto original e os padrões de expressões que são previamente criados por outro dos módulos do PorTexTO, o módulo `Processador de co-ocorrências` (ver descrição detalhada na secção 8.1.2).

Além da colecção e dos padrões de expressões, o sistema tem ainda como entrada uma lista de palavras-chave temporais usadas para unicamente excluir do processamento frases que não contenham expressões temporais e assim conseguir diminuir o tempo final de processamento dos documentos. Esta lista funciona no sistema como um filtro das frases a serem processadas. As frases excluídas são todas as que não têm nem datas, nem nenhuma das palavras-chave temporais. As palavras-chave temporais são definidas consoante a lista de expressões temporais que o sistema deverá identificar e classificar, de modo a atingir os objectivos de uma determinada tarefa. Por exemplo, se existirem padrões de expressões temporais com a palavra temporal *ano*, então *ano* deverá existir na lista de palavras-chave temporais.

De seguida, são apresentados com maior detalhe os módulos `Anotador` e `Processador de co-ocorrências`.

8.1.1 Módulo `Anotador`

O módulo `Anotador` é responsável por identificar as expressões temporais, mediante os padrões definidos pelo módulo `Processador de co-ocorrências`, fazer a sua classificação

[1] Disponível em http://search.cpan.org/~ambs/Lingua-PT-PLNbase-0.20.

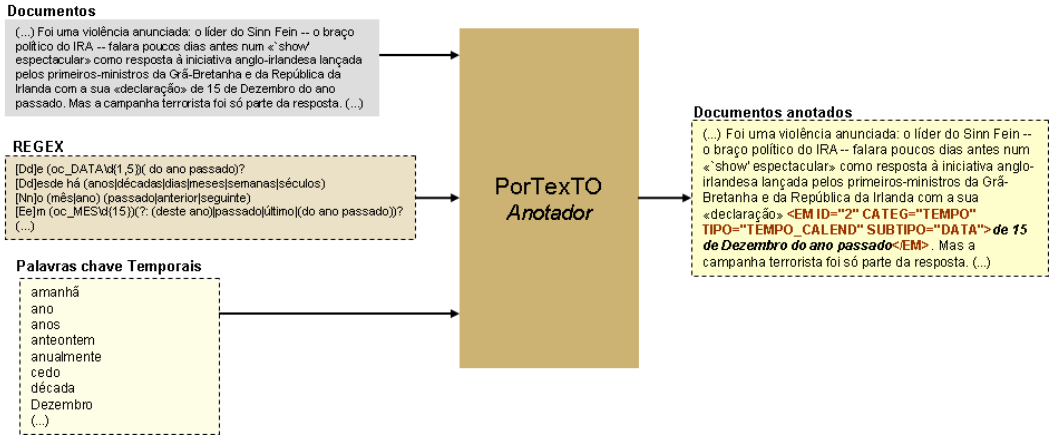


Figura 8.1: Arquitectura do módulo `Anotador` do sistema PorTexTO.

e, posteriormente, proceder à anotação no texto original.

O funcionamento deste módulo é apresentado na figura 8.2. Os documentos de entrada são trabalhados um de cada vez e o processamento de cada documento é feito numa frase de cada vez. Cada frase será submetida às quatro etapas de processamento do `Anotador` do PorTexTO. A frase só será dividida nos seus termos caso haja necessidade de reconhecer datas com o mês por extenso, ou datas que tenham também o dia da semana. Por exemplo, a frase *Domingo, 7 de Setembro de 2008* é dividida nos seguintes termos: *Domingo*, *7*, *de*, *Setembro*, *de*, *2008*. Com estes termos, é verificado se os que estão à esquerda e à direita do mês podem fazer ou não parte de uma data. No caso desses termos poderem constar de uma data então serão incluídos na expressão que vai ser marcada como `DATA`. No caso contrário, só o mês é que será marcado, mas com o marcador `MES`. No exemplo apresentado, a frase inicial terá a marcação `DATA`.

Na primeira etapa é decidido se a frase vai ser ou não processada. Uma frase só será processada caso possa conter, pelo menos, uma expressão temporal. Isto é, a frase deverá ter termos numéricos ou, pelo menos, uma das palavras definidas na lista de palavras-chave temporais (ver figura 8.1). As frases que não têm nenhuma expressão temporal ficam excluídas do processamento.

Vamos considerar a frase (8.1), extraída da colecção do Segundo HAREM, para exemplificar as próximas etapas do processamento.

(8.1) A missão científica da nave foi concluída *em 30 de abril de 2002*.

Na segunda etapa são geradas expressões candidatas a entidades mencionadas temporais. No caso da frase conter dígitos, o sistema aplica regras para reconhecer horas, datas completas ou incompletas (datas constituídas só por dia e mês ou mês e ano) e anos.

De seguida, as frases são analisadas quanto à existência de meses (palavras por extenso ou abreviaturas) e dias da semana, sendo posteriormente aplicadas regras para reconhecimento de datas constituídas por termos numéricos e palavras. Depois de identificadas, as expressões são marcadas e passam a ser expressões candidatas. Nesta etapa, a frase do exemplo passaria a:

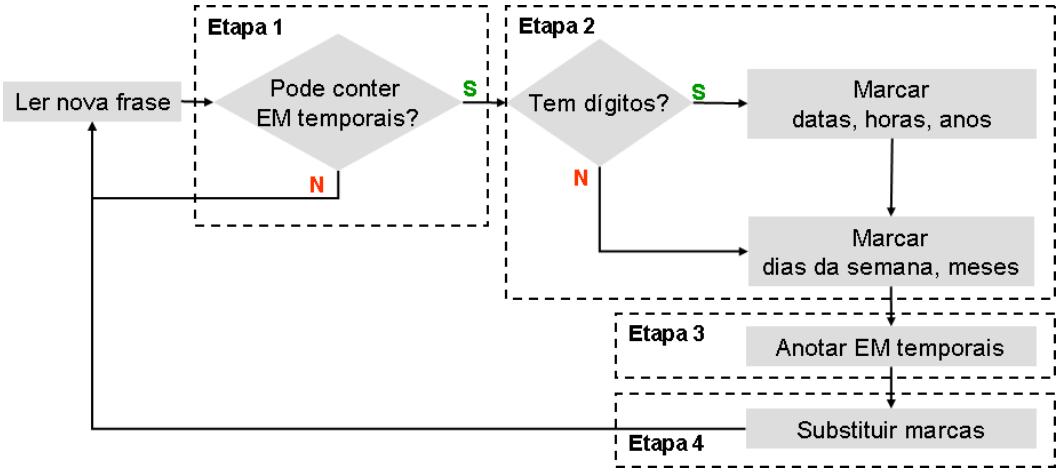


Figura 8.2: Funcionamento do módulo `Anotador` do sistema PorTexTO.

(8.2) A missão científica da nave foi concluída *em* `oc_DATA12`.

Na terceira etapa, verifica-se a correspondência entre a frase resultante das anteriores etapas (frase (8.2)) e os padrões definidos para o actual processamento e que foram anteriormente criados pelo módulo `Processador de co-ocorrências` (ver secção 8.1.2). A frase (8.2) tem correspondência com o padrão, definido através da seguinte expressão regular:

```
[Ee]m (oc_DATA\d{1,5})(?:  (deste ano)|passado|último)?
```

A anotação da expressão será efectuada caso ocorra correspondência e a classificação atribuída está associada ao padrão responsável pela correspondência. A frase (8.2) resultaria na frase (8.3).

(8.3) A missão científica da nave foi concluída <EM ID="41" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**em oc_DATA12**</EM>.

A quarta e última etapa é responsável por substituir pelo texto original, todas as marcas que foram colocadas na frase durante a execução da segunda etapa de processamento. A frase de exemplo depois de terminado o processamento ficará como em (8.4).

(8.4) A missão científica da nave foi concluída <EM ID="41" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**em 30 de Abril de 2002**</EM>.

8.1.2 Módulo `Processador de co-ocorrências`

O módulo `Processador de co-ocorrências` só é executado quando ainda não existem padrões de expressões temporais ou os que existem são insuficientes para a tarefa a desempenhar pelo PorTexTO. O objectivo principal deste módulo é determinar as expressões temporais mais utilizadas numa determinada colecção segundo uma abordagem estatística e

com estas expressões criar os padrões que vão ser posteriormente utilizados no módulo `Anotador`. Este módulo produz como resultado um ficheiro com os padrões definidos através de expressões regulares e a respectiva classificação. A arquitectura deste módulo é apresentada na figura 8.3.

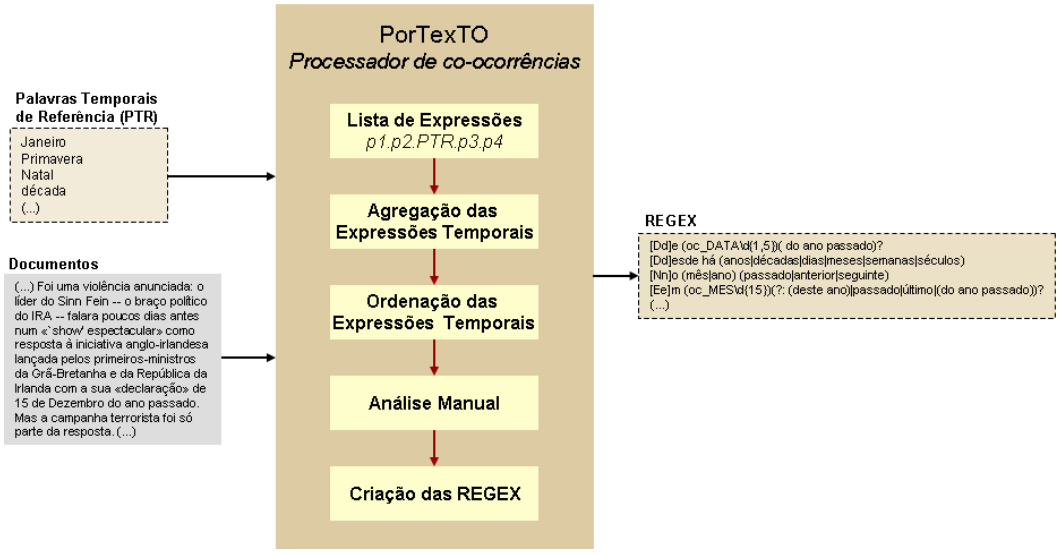


Figura 8.3: Arquitectura do módulo `Processador de co-ocorrências` do sistema PorTexTO.

A determinação das co-ocorrências é realizada utilizando um conjunto de palavras, palavras essas que são referidas neste capítulo como palavras temporais de referência (PTR). A lista das PTR deve ser constituída por todas as palavras temporais que apareçam em expressões com no mínimo duas palavras, como por exemplo, meses do ano, estações do ano, dias da semana, festividades e unidades de medida temporal, para que realmente haja a necessidade de determinar as palavras que ocorrem conjuntamente. Assim, as palavras que sozinhas formam uma expressão temporal devem ser excluídas desta lista, como por exemplo, advérbios temporais com terminação *-mente* (e.g., *diariamente*).

O funcionamento deste módulo está dividido em cinco etapas de processamento e tal como no módulo `Anotador` os documentos de entrada também são processados frase a frase.

O objectivo da primeira etapa é obter uma lista com as expressões encontradas onde as co-ocorrências detectadas têm uma distância máxima de *n* palavras antes e/ou *n* palavras depois da palavra temporal de referência. Por exemplo, considerando a palavra temporal de referência *ano* podemos obter as seguintes expressões *No* ***ano*** *passado*, *No último* ***ano***, *No próximo* ***ano*** *de 2009*. A lista obtida, para além das expressões encontradas, tem também o respectivo número de ocorrências.

Na segunda etapa é feita a agregação das expressões temporais contidas na lista criada na etapa anterior. As expressões são agregadas quando têm mais do que uma palavra que ocorre com uma determinada palavra temporal de referência, na mesma posição. Quando ocorre a agregação de expressões o número total de ocorrências passa a ser a soma das ocorrências de cada uma das expressões agregadas. Por exemplo, as expres-

sões *No **ano** passado* e *No **ano** seguinte* são agregadas e passamos a ter o padrão `No` **`ano`** `passado|seguinte`.

A agregação também é feita para as referências temporais consideradas como datas, os meses do ano e as horas. Por exemplo, as expressões *Em **Janeiro*** e *Em **Fevereiro*** são agregadas no padrão `Em` **`oc_MES`**. No caso da expressão *No dia 25 de Janeiro* o padrão criado seria `No dia` **`oc_DATA`**.

A terceira etapa ordena a lista de expressões, já com expressões que foram agregadas, por ordem decrescente do número total de ocorrências.

Na quarta etapa é efectuada uma análise manual necessária para excluir todas as expressões que embora tenham uma unidade lexical que representa um elemento temporal, não sejam na realidade uma expressão temporal e outras que não façam sentido, pelo senso comum de quem tem domínio da língua. Por exemplo, as expressões *Feliz Natal* e *Bom Ano Novo* são excluídas da lista, embora de acordo com as directivas do Segundo HAREM estas expressões temporais devessem ter como classificação `GENERICO` no atributo `TIPO` (Hagège et al., 2008). No entanto o PorTexTO não faz esta classificação.

Esta é a etapa mais complexa do módulo `Processador de co-ocorrências`, pois mesmo com o conhecimento da língua portuguesa e seguindo as directivas do TEMPO definidas para o Segundo HAREM aparecem sempre situações dúbias tornando-se difícil decidir se realmente uma determinada expressão deve ser ou não considerada uma expressão temporal.

A quinta etapa cria as expressões regulares que definem os padrões que foram considerados após a análise manual e associa a respectiva classificação, segundo as directivas do Segundo HAREM (Hagège et al., 2008). Como exemplo, apresentamos de seguida um dos padrões criados:

```
[Nn]o (mês|ano) (passado|anterior|seguinte)
```

8.2 Participação no Segundo HAREM

A participação do PorTexTO no Segundo HAREM foi efectuada com a sua versão 1.0 (PorTexTO 1.0). Este sistema só foi submetido à avaliação nas tarefas de identificação e classificação de entidades mencionadas no cenário `TEMPO`, uma vez que só processa expressões temporais.

Na classificação das entidades mencionadas temporais o sistema PorTexTO só utilizou o `TIPO` e o `SUBTIPO` da categoria `TEMPO`, não tendo utilizado nem a classificação de `TEMPO` estendido, nem a normalização. Além disso, ficaram excluídos da classificação o subtipo `INTERVALO` do tipo `TEMPO_CALEND` e o tipo `GENERICO`.

Os padrões utilizados no sistema PorTexTO 1.0 só representam expressões temporais simples que se iniciam com os termos *a*, *às*, *de*, *em*, *há*, *durante*, *desde*, *pelas*, e os termos *no*, *naquele*, *neste*, *nesse*, *este*, *esse* também no género feminino e no plural. Por expressões temporais simples entendem-se as expressões linguísticas compostas por uma só unidade lexical que denote um elemento temporal.

As expressões temporais compostas, como por exemplo, *no **dia** 10 do **mês** passado*, no PorTexTO 1.0 são tratadas como duas expressões simples, não obtendo a correcta classificação segundo as directivas do TEMPO do Segundo HAREM (Hagège et al., 2008).

Os padrões das expressões temporais foram criados pelo módulo `Processador de co-ocorrências` utilizando a lista das palavras temporais de referência composta pelos

meses do ano, estações do ano, dias da semana, festividades (*Natal*, *Páscoa*, *Carnaval* e *Entrudo*), unidades de medida temporal (*década*, *período*, *século*, *ano*, *mês*, etc.) e outras (*altura*, *instante*, *momento*, *tempo*). Na determinação das co-ocorrências a distância máxima considerada foi de n=2 palavras antes e/ou n=2 palavras depois da palavra temporal de referência.

O módulo `Processador de co-ocorrências` foi aplicado à colecção do HAREM versão 2.0[2] utilizada nos dois eventos de avaliação do Primeiro HAREM. Esta colecção apresenta cerca de 520 mil palavras distribuídas por cerca de 40 mil linhas e provenientes de 1202 documentos de diferentes géneros textuais (textos técnicos, políticos, literários, expositivos, jornalísticos, entrevistas, mensagens de correio electrónico e páginas web) (Santos e Cardoso, 2007a).

Os padrões de expressões utilizados no Segundo HAREM só foram criados depois da colecção do Segundo HAREM também ter sido submetida ao módulo `Processador de co-ocorrências`. Seguindo uma abordagem estatística, os resultados obtidos neste módulo foram posteriormente incluídos na lista de expressões que já possuíamos aquando do processamento dos outros textos e verificou-se o aparecimento de novas expressões temporais e alteração da ordem de outras expressões na lista de frequência.

O PorTexTO 1.0 participou com quatro corridas. A corrida PorTexTO_1 serviu só para validar o envio de resultados ao Segundo HAREM e verificar se existiam algumas incoerências na anotação das entidades mencionadas. Após o envio desta corrida verificámos que estava tudo correcto.

As outras corridas (PorTexTO_2, PorTexTO_3 e PorTexTO_4) têm pequenas diferenças ao nível da definição das expressões regulares utilizadas – mais precisas e menos abrangentes ou menos precisas e mais abrangentes, como por exemplo, `[Nn]o (passado|último) ano` e `[Nn]o \w+ ano`, respectivamente.

O envio destas três corridas teve por objectivo podermos avaliar qual a penalização da precisão quando se aumentou a abrangência na definição das expressões regulares usadas nos padrões.

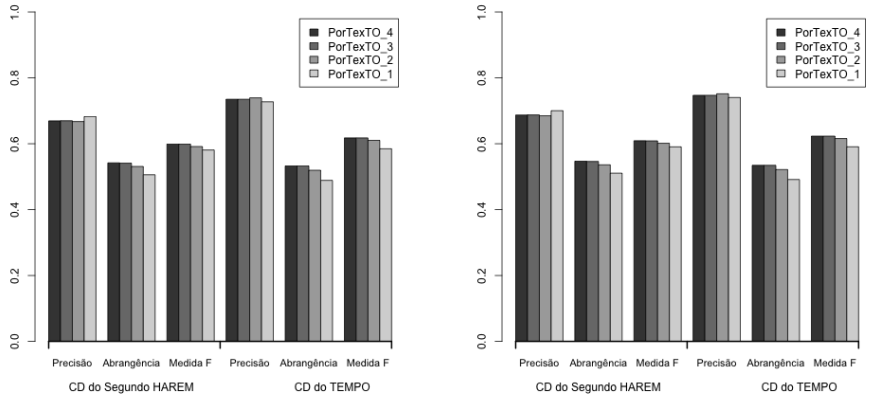
8.3 Resultados da participação no Segundo HAREM

Os resultados obtidos pelo sistema PorTexTO 1.0 na sua primeira participação em avaliações conjuntas excederam as nossas expectativas porque esta versão apresentava muitas limitações, tal como foi referido na secção 8.2. O sistema foi criado de raiz para a participação no Segundo HAREM e não houve tempo suficiente até à avaliação para tudo estar implementado e devidamente testado.

O sistema utilizou para o processamento um computador pessoal com 1GB de memória RAM, processador Intel Core 2 E6600 a 2.4 GHz e sistema operativo Microsoft Windows XP Professional, versão 2002, SP2. Em termos do desempenho computacional, o sistema PorTexTO 1.0 anotou a colecção do Segundo HAREM a um débito de aproximadamente 22KB por segundo, tendo processado as 33 mil linhas com cerca de 675 mil palavras da colecção do Segundo HAREM em cerca de três minutos e vinte segundos. O tempo de processamento conseguido é bastante aceitável para os objectivos traçados para o sistema, como por exemplo a sua incorporação numa aplicação de recolha de informação (em inglês, *ad-hoc retrieval)*.

[2] Disponível no sítio do HAREM (http://www.linguateca.pt/HAREM), na secção dedicada ao Primeiro HAREM.

Na apresentação dos resultados obtidos iremos focar-nos no único cenário que se adequa ao nosso sistema, o cenário constituído pela categoria TEMPO, uma vez que nos restantes cenários existem outras categorias que estão também a ser avaliadas e que o PorTexTO não tentou reconhecer. Faremos a análise tanto na CD do Segundo HAREM, como na CD do TEMPO. Neste último caso, apenas estamos interessados no modo de avaliação do HAREM clássico, ou seja, sem ter em conta os atributos específicos da categoria TEMPO, pois não tentámos atribuí-los. Não fizemos distinção entre os dois tipos de avaliação de ALT, já que nas directivas do TEMPO não foi considerada a etiqueta ALT e, consequentemente, no cenário da categoria TEMPO a avaliação estrita ou relaxada de ALT produz os mesmos valores.



(a) Classificação (b) Identificação

Figura 8.4: Resultados das quatro corridas do PorTexTO obtidos na pista do TEMPO na CD do Segundo HAREM e na CD do TEMPO.

Como se pode constatar pelos gráficos da figura 8.4, as quatro corridas apresentam resultados bastante similares com diferenças pouco significativas. No entanto, a corrida PorTexTO_1 teve uma maior diferença relativamente às outra corridas, como era esperado, já que quando esta corrida foi submetida a avaliação o sistema ainda não estava devidamente configurado.

Nos resultados obtidos nas corridas PorTexTO_2, PorTexTO_3 e PorTexTO_4 verificámos que a definição dos padrões de forma mais abrangente não foi muito penalizadora para a precisão, em ambas as colecções douradas e tanto na identificação como na classificação.

Na CD do TEMPO, as corridas PorTexTO_3 e PorTexTO_4 chegam a ter os mesmos resultados.

Os resultados obtidos pelo PorTexTO na categoria TEMPO na CD do Segundo HAREM, relativamente às métricas abrangência e precisão para a classificação e para a identificação não têm diferenças significativas (ver tabela 8.1), ou seja, as entidades temporais identificadas pelo PorTexTO como TEMPO estão a ser bem classificadas quanto ao seu tipo e subtipo. Aliás, este comportamento é seguido na CD do TEMPO, como se pode constatar pelos resultados apresentados na tabela 8.2. Como as três corridas (PorTexTO_2, PorTexTO_3 e

Tabela 8.1: Resultados do PorTexTO na pista do TEMPO na CD do Segundo HAREM.

Corrida	Classificação			Identificação		
	Precisão	Abrangência	Medida F	Precisão	Abrangência	Medida F
PorTexTO_4	0,6694	**0,5419**	**0,5990**	0,6871	**0,5470**	**0,6091**
PorTexTO_3	0,6698	0,5410	0,5986	0,6875	0,5462	0,6087
PorTexTO_2	0,6674	0,5310	0,5915	0,6849	0,5360	0,6014
PorTexTO_1	**0,6825**	0,5058	0,5810	**0,7002**	0,5106	0,5905

Tabela 8.2: Resultados da corrida PorTexTO_4 na pista do TEMPO na CD do Segundo HAREM (CDSH) e na CD do TEMPO (CDT).

CD	Classificação			Identificação		
	Precisão	Abrangência	Medida F	Precisão	Abrangência	Medida F
CDSH	0,6694	**0,5419**	0,5990	0,6871	**0,5470**	0,6091
CDT	**0,7350**	0,5327	**0,6177**	**0,7470**	0,5345	**0,6231**

PorTexTO_4) tiveram resultados bastante idênticos, nesta tabela só apresentamos a corrida PorTexTO_4.

O PorTexTO conseguiu uma pequena melhoria na precisão quando a avaliação utilizou o subconjunto da CD – CD do TEMPO, mas que não é significativa. A diferença foi de aproximadamente 6%.

Nesta CD, o sistema PorTexTO na corrida PorTexTO_4 obteve a 5.ª posição, mas a diferença relativamente à corrida XIP-L2F/Xerox_3 do sistema que obteve a 1.ª posição verificou-se somente na abrangência. O sistema XIP-L2F/Xerox foi o melhor sistema com uma precisão de aproximadamente 75% para uma abrangência de 78%. A figura 8.5 faz a comparação entre os resultados obtidos pelos dois sistemas.

A precisão poderá ser melhorada quando o sistema passar a identificar expressões temporais compostas. Mas a sua maior limitação deve-se à abrangência dos padrões utilizados, sendo necessário acrescentar mais padrões para conseguir reconhecer um maior número de entidades mencionadas temporais. Por exemplo, padrões para identificar expressões com classificação de subtipo `INTERVALO` do tipo `TEMPO_CALEND` (*entre 1990 e 2000*, *de 2 a 6 meses*).

8.4 Conclusões e trabalho futuro

O sistema PorTexTO foi desenvolvido com o objectivo de utilizar um algoritmo simples para conseguir obter um bom desempenho computacional, mesmo que não identifique e classifique todas as expressões temporais, conforme foi referido na secção introdutória deste capítulo.

A participação do PorTexTO no Segundo HAREM foi bastante importante, pois para além de permitir saber qual o desempenho do sistema nas suas tarefas de reconhecimento de entidade mencionadas temporais, também facultou o acompanhamento na criação das directivas de classificação e normalização destas entidades.

Os resultados obtidos pelo sistema PorTexTO na sua versão 1.0 são bastante motivadores. Os resultados vieram demonstrar que o algoritmo seguido pelo sistema permite obter

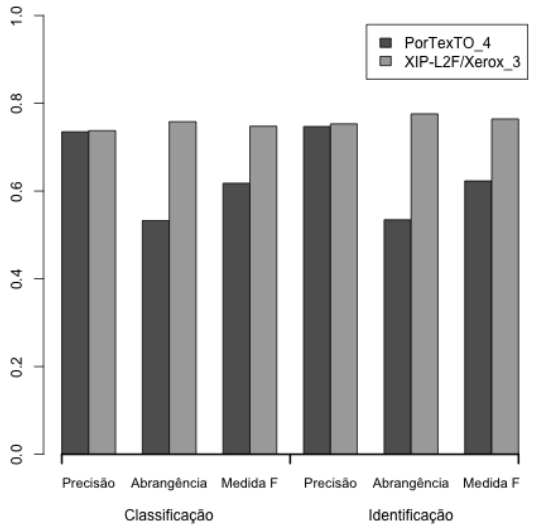


Figura 8.5: Resultados da corrida PorTexTO_4 e da corrida XIP-L2F/Xerox_3 para a pista do TEMPO na CD do TEMPO.

um bom desempenho, embora obviamente ainda precise de várias afinações. O sistema ainda se encontrava numa versão inicial e os padrões de expressões utilizados no reconhecimento das entidades mencionadas temporais foram muito limitados, mas, mesmo assim, o sistema conseguiu uma abrangência de aproximadamente 55%.

Os objectivos foram atingidos nesta participação, pois apesar do sistema não ter conseguido reconhecer todas as entidades mencionadas temporais, cerca de 75% das que o sistema reconheceu estão correctas.

No entanto, o sistema necessita de ultrapassar as limitações verificadas na versão 1.0. Um dos melhoramentos a fazer será criar padrões que representem expressões com um maior número de palavras, especificamente ultrapassar o limite de n=2 na determinação de co-ocorrências. O processamento de expressões também deveria ser alargado a expressões temporais complexas, isto é, criar padrões para expressões com mais de uma unidade temporal. Por exemplo, reconhecer a seguinte expressão, como uma única entidade mencionada: *no **dia** 10 do **mês** passado*.

A tarefa da criação de padrões de expressões poderá ser mais automatizada, para que com mais facilidade se possa acrescentar novos padrões.

Num trabalho futuro será interessante utilizar o sistema PorTexTO já melhorado, no reconhecimento de entidades mencionadas temporais em outras línguas. A adaptação do sistema ao processamento da língua, para além da língua portuguesa, não será uma tarefa difícil, porque os módulos funcionam de forma independente da língua. No entanto será sempre necessário ter algum conhecimento da língua em que se pretende aplicar o sistema. Resumidamente, o PorTexTO necessita de ter a parte importante do reconhecimento de entidades mencionadas temporais que é a lista de PTR na língua em questão, e as pa-

lavras-chave temporais dessa língua (Pustejovsky et al., 2005). Com esta informação, o módulo `Processador de co-ocorrências` cria os padrões de expressões que são necessários para que o módulo `Anotador` possa fazer o reconhecimento de entidades mencionadas temporais nos textos dessa língua.

Capítulo 9

Adaptação do sistema de reconhecimento de entidades mencionadas da Priberam ao HAREM

Carlos Amaral, Helena Figueira, Afonso Mendes, Pedro Mendes, Cláudia Pinto e Tiago Veiga

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 9, p. 171–179.

A Priberam tem vindo a desenvolver um sistema de REM na sua plataforma de desenvolvimento linguístico (Amaral et al., 2004a), cuja adaptação para a participação no segundo HAREM se descreve no secção 9.1.1. Este sistema está já em uso como módulo independente em vários produtos concebidos pela empresa, nomeadamente no FliP[1], no sistema de resposta automática a perguntas (Amaral et al., 2005), nos sistemas de extracção de informação em motores de pesquisa (Amaral et al., 2004b) como os utilizados nos sítios da TSF[2] e do JN[3] e numa ferramenta de tratamento dos acórdãos do Supremo Tribunal de Justiça[4], o IncogniX, que é usada para remover as referências às entidades envolvidas nos processos.

O presente capítulo pretende fazer a descrição geral do sistema de reconhecimento de entidades mencionadas (REM) da Priberam e do trabalho que foi necessário realizar para a sua primeira participação no HAREM. Na primeira secção, descreve-se o funcionamento do sistema e a adaptação realizada para o reconhecimento das categorias, tipos e subtipos propostos no Segundo HAREM. A detecção de entidades mencionadas (EM) para a pista do TEMPO foi a que maiores problemas levantou nesta adaptação, visto que os critérios estabelecidos para a criação de regras no sistema da Priberam diferiam bastante daqueles propostos pela pista do TEMPO, quer a nível da detecção e construção das EM, quer a nível da sua classificação. Na segunda secção, procede-se à análise dos resultados obtidos pelo sistema da Priberam nesta segunda edição do HAREM. Por fim, lista-se o trabalho e as melhorias que se pretendem futuramente realizar.

9.1 Descrição do sistema

O sistema de REM da Priberam tem por base um léxico com classificação morfossintáctica e semântica, correspondendo a cada entrada no léxico uma ligação a um ou mais níveis de uma ontologia multilingue (Amaral et al., 2004a), que está estruturada através de relações de proximidade conceptual. A cada entrada lexical pode corresponder um ou mais sentidos, que, por sua vez, contêm diferentes valores morfológicos e semânticos (ver exemplo (9.1)).

(9.1) *árvore*
```
s1 [planta lenhosa]
N (SING|, FEM|, VEGETAL)
s2 [estrutura de representação]
N(SING|, FEM|, ABSTR|CONCR)
s3 [eixo, veio]
N(SING|, FEM|, CONCR|, Pde|)
```

A identificação das EM passa, numa primeira fase, pela simples herança dos valores semânticos e morfológicos previamente estabelecidos nesse léxico. No entanto, e como é dito em (Santos, 2007d, p. 53), esta abordagem "ingénua" tem de ser complementada com uma outra que explore também o contexto em que a EM está inserida, pelo que a análise

[1] *Ferramentas para a Língua Portuguesa*. O FLiP inclui um corrector sintáctico, um corrector ortográfico, um dicionário de sinónimos e um hifenizador. Está disponível na rede uma versão de demonstração em http://www.flip.pt/.
[2] Ver http://www.tsf.pt/.
[3] Ver http://www.jn.pt/.
[4] Os acórdãos tratados estão disponíveis em http://www.dgsi.pt/.

e herança dos valores dos nomes classificados no léxico tem de ser complementada com a análise contextual sintáctico-semântica.

O sistema é construído com recurso ao uso de regras contextuais (Amaral et al., 2004a) que permitem a atribuição ou alteração de valores morfológicos e semânticos a unidades isoladas ou a sequências de unidades. Este tipo de regras permite, por exemplo, a criação de locuções através da combinação estrita de sequências de palavras, como em (9.2), de categorias gramaticais com palavras, como em (9.3), apenas de categorias gramaticais, como em (9.4), e ainda combinações de listas de palavras, às quais chamamos "*constantes*", com categorias ou palavras únicas, como em (9.5).

(9.2) `Pal(secretaria) Pal(de) Pal(estado) = N`

(9.3) `Pal(às) Pal(primeiras) Pal(horas) Pal(de) Cat(N(DIASEMANA)) = ADV`

(9.4) `Cat(ADV) Cat(CARD) = CARD`

(9.5) `Constante Extensaodeagua = Pal(mar, oceano, rio, lago)`
`Extensaodeagua Pal(de) Cat(Nprop) = EM`

As regras contextuais do REM são, na sua maior parte, dependentes da língua do léxico que alimenta o sistema, apesar de este ter também a capacidade de detectar EM cujos elementos não fazem parte desse léxico (por exemplo, *Red Label*, *Armory Show*); nestes casos a identificação da EM ignora frequentemente a sua classificação se o contexto não permitir que lhe sejam atribuídos valores semânticos.

As constantes desempenham um papel crucial na detecção e classificação de EM. Para além de permitirem agrupar palavras, como as preposições ou outras palavras gramaticais, que repetidamente fazem parte de EM (ver (9.6)), levando a uma poupança de tempo na escrita das regras, permitem ainda, em muitos casos, detectar e classificar as EM através da aglomeração paradigmática de palavras com determinadas afinidades semânticas e morfológicas, o que possibilita, com um grau de certeza relativamente elevado, classificar a EM (ver (9.7)). As palavras contidas nessas constantes podem, sobretudo se forem escritas com inicial maiúscula, fazer parte da entidade ou ainda permitir a identificação e classificação da entidade se estiverem escritas com inicial minúscula, especialmente quando acompanhadas de informação contextual (ver (9.8)). Neste último caso é necessário cuidado adicional, pois as ambiguidades morfológicas e semânticas são maiores quando se trata de nomes comuns (por exemplo, *serra da Estrela* vs. *serra do Manuel*[5]).

(9.6) `Constante PreposicaoDe = Lema(de)`

(9.7) `Constante Listadeorganizacoes = Pal(instituto, instituição, organização, associação)`

(9.8) `Cat(NPROP) PreposicaoDe Cat(NPROP) = ENT(ORGANIZACAO) if before $$ is Listadeorganizacoes`

[5] A existência da palavra *serra* a anteceder um nome próprio não é, neste caso, suficiente só por si para identificar e classificar a EM como topónimo.

Para além de listas de palavras ou de lemas, as constantes podem ainda conter categorias com ou sem restrições morfológicas e semânticas, que poderão ser usadas repetidamente nas regras de detecção de EM, facilitando a sua escrita (ver exemplo (9.9)).

(9.9)
```
Constante Antroponimo = NPROP(PESSOA)
Constante Nprop = NPROP

Antroponimo Nprop = EM(PESSOA)
Antroponimo Nprop Antroponimo = EM(PESSOA)
```

A lista de palavras em minúsculas permitidas pelo Segundo HAREM na construção das entidades está inserida nas constantes. Neste caso, como a lista é bastante limitada (parece, por exemplo, pouco coerente que permita a inclusão da palavra *tio* e não a inclusão de outras relações de parentesco, como *primo*) e o nosso sistema permite uma lista mais abrangente de palavras em minúsculas nas EM, criámos constantes apenas para a participação no Segundo HAREM.

As regras de REM levam não só em conta as sequências de nomes próprios, separadas ou não por determinadas preposições, assim como o contexto em que são detectadas. Deste modo, uma EM, que sem contexto poderia ser classificada como antropónimo, poderá ser classificada como organização se imediatamente antes tiver algo que a identifique como tal (ver exemplo (9.10)). Por exemplo, uma EM como *Ricardo Jorge*, tipicamente marcada com a etiqueta `PESSOA`, será classificada como `ORGANIZACAO` se for antecedida por uma expressão como *instituto*.

(9.10)
```
Cat(ENT(PESSOA)) = ENT(ORGANIZACAO)
if before $$ is Listadeorganizacoes
```

9.1.1 Adaptação do sistema ao Segundo HAREM

O HAREM vai na sua segunda edição, mas foi a primeira vez que a Priberam participou nesta avaliação conjunta, apesar de já não ser uma estreante em iniciativas deste género, uma vez que vem participando no CLEF (Cross-Language Evaluation Forum)[6] desde 2005, sempre com resultados acima da média (Amaral et al., 2005; Cassan et al., 2006; Amaral et al., 2007).

Nesta edição do HAREM, a Priberam participou no cenário total, assim como os sistemas REMBRANDT e SeRELeP (este apenas para a identificação e não para a classificação).

Como a Priberam possui uma plataforma única para o desenvolvimento dos seus produtos linguísticos (Amaral et al., 2004a) (cujos módulos podem ser usados individualmente para fins distintos), a introdução de novas categorias e de novos tipos e subtipos nas regras pode ser realizada com relativa facilidade.

Apesar de o sistema contemplar já grande parte das categorias de EM estabelecidas pelo Segundo HAREM (`PESSOA`, `LOCAL`, `ORGANIZACAO`, `VALOR`, `TEMPO`), foi necessário criar regras para classificação de EM com as categorias `COISA` (por exemplo, *Oseltamivir*, *Leucotomia*), `ABSTRACCAO` (por exemplo, *Medicina*, *Psiquiatria*), `ACONTECIMENTO` (por exemplo, *Conferência de Dadores*, *Dia de Reis*) e `OBRA` (por exemplo, *The Streets of Paris*, *Lei Rouanet*). Estas entidades já

[6] Ver http://www.clef-campaign.org/.

eram identificadas pelo sistema automático de resposta a perguntas antes da participação no HAREM, mas as EM eram extraídas com valores semânticos indefinidos.

Para todas as categorias, no entanto, foi necessário afinar as regras do detector para que reconhecesse subtipos de EM, nomeadamente topónimos dos subtipos AGUACURSO (por exemplo, *Tejo*, *Eufrates*), AGUAMASSA (por exemplo, *Oceano Pacífico*, *Lago dos Cisnes*), RELEVO (por exemplo, *Monte Rosa*, *Evereste*) e ILHA (por exemplo, *Martinica*, *Ilha de Moçambique*) e antropónimos do tipo GRUPOMEMBRO (por exemplo, *Povos Indígenas*), pois anteriormente apenas eram reconhecidas como entidades gerais de lugar e de pessoa sem etiquetas restritivas.

Para o reconhecimento e classificação destes subtipos, foi necessário acrescentar novas etiquetas semânticas no léxico, visto que a classificação das EM no sistema se baseia, numa primeira fase, na herança dos traços atribuídos no léxico, como ficou descrito no primeiro ponto desta secção. Foram também criadas novas constantes para a classificação contextual das EM. Para tal, a ontologia usada pela Priberam foi bastante útil, porque permitiu uma extracção mais exaustiva de nomes relacionados com os tipos e subtipos que se pretendiam implementar.

Visto que os valores semânticos calculados pelo sistema não são os mesmos que foram estabelecidos pelo HAREM, foi necessário criar um filtro que estabelecesse as equivalências entre as categorias e valores originais do sistema e os do HAREM. Este filtro recorre a um ficheiro de configuração em XML (ver exemplo (9.11)), que facilmente se consegue modificar para permitir a etiquetação do texto com novas categorias e valores semânticos.

(9.11)
```
<TIPO NOME="EM">
    <TRACO NOME="TipoEM">
        <VALORES>ANTROP_IND</VALORES>
    </TRACO>
</TIPO>
<SUBSTRING>
    <EM ID="{0}" CATEG="PESSOA" TIPO="INDIVIDUAL">{1}</EM>
</SUBSTRING>
```

No exemplo (9.11) o nó <TIPO> indica a categoria gramatical, o nó <TRACO> o nome do traço cujos valores são indicados em <VALORES>. Finalmente, no nó <SUBSTRING>, <EM> atribui os valores correspondentes no HAREM.

9.2 Análise dos resultados da participação no Segundo HAREM

9.2.1 Resultados do HAREM clássico

Em termos absolutos, tendo em conta a medida de abrangência, isto é, apenas estabelecendo a comparação de entidades detectadas pelo sistema da Priberam e aquelas marcadas na colecção dourada (CD) do Segundo HAREM, os resultados são bastante animadores, uma vez que o sistema da Priberam identifica correctamente 72,29% das EM (ver tabela 9.1). Considerando também a medida de abrangência, a percentagem das EM classificadas correctamente é menor (51,46%), apesar de no cenário total ter tido o valor mais elevado entre todos os sistemas.

Tabela 9.1: Resultados do sistema de REM da Priberam na classificação e na identificação com avaliação estrita de ALT

	Classificação			Identificação		
Cenário	**Precisão**	**Abrangência**	**Medida F**	**Precisão**	**Abrangência**	**Medida F**
Total	0,6417	**0,5146**	**0,5711**	0,6994	**0,7229**	**0,7109**
Selectivo 2	0,5920	**0,5893**	0,5906	0,5830	**0,7127**	0,6414
Selectivo 3	0,7263	**0,5641**	**0,6350**	0,7643	**0,8158**	**0,7892**
Selectivo 4	0,6441	**0,5175**	**0,5739**	0,6958	**0,7222**	**0,7088**
Selectivo 5	0,3287	0,7000	0,4473	0,2863	**0,7856**	0,4197
Selectivo 6	0,6110	**0,5343**	0,5701	0,2863	**0,7144**	0,6746

Tabela 9.2: Posição do sistema da Priberam nos vários cenários possíveis no HAREM clássico

	Classificação		Identificação	
Cenário	**Avaliação estrita de ALT**	**Avaliação relaxada de ALT**	**Avaliação estrita de ALT**	**Avaliação relaxada de ALT**
Total	1	1	1	1
Selectivo 2	8	7	8	7
Selectivo 3	1	1	1	1
Selectivo 4	1	1	1	1
Selectivo 5	19	19	19	19
Selectivo 6	4	3	2	1

No cenário total em que participou, o sistema da Priberam obteve os melhores resultados na medida F[7] entre todos os participantes, quer na classificação quer na identificação das EM, para além da primeira posição na medida de abrangência, apesar de em nenhum dos cenários ter alcançado a melhor marca na medida de precisão. Através da comparação das tabelas 9.2 e 9.3, pode verificar-se como o sistema tem resultados bastante mais elevados na identificação do que na classificação.

Considerando os valores da medida F, o sistema da Priberam alcançou a primeira posição em 13 dos 24 cenários no total das avaliações possíveis no HAREM clássico[8] (ver tabela 9.2). No entanto, como se pode verificar na tabela 9.3, o sistema não apresenta resultados tão satisfatórios quando a avaliação é feita por categoria, o que indica que necessita de melhorias na vertente da classificação semântica das EM.

Fazendo a avaliação do sistema por categorias de EM (ver tabela 9.3), constatamos que ele se comporta melhor nas categorias `ABSTRACCAO` e `COISA`, quer em identificação quer em classificação, assim como em abrangência retira os melhores resultados na categoria `PESSOA`, mas apenas na classificação. O sistema tem resultados mais baixos na identificação e classificação de EM com as categorias `LOCAL`, `TEMPO` e `VALOR`.

No que diz respeito especificamente à categoria `TEMPO`, os resultados devem-se em larga medida ao facto de os critérios estabelecidos para a detecção de entidades e locuções temporais no Segundo HAREM não serem em grande parte compatíveis com as regras existentes no sistema da Priberam. Optou-se então por criar relações entre os valores semânticos

[7] A medida F é uma medida geral que combina os valores da precisão e da abrangência. Ver secção 5.4.

[8] De acordo com os relatórios individuais disponíveis no sítio do HAREM (http://www.linguateca.pt/HAREM).

Tabela 9.3: Posição do sistema da Priberam na avaliação por categorias.

Categoria	Classificação	Identificação	N.º de participantes
ABSTRACCAO	1	1	10
ACONTECIMENTO	11	11	16
COISA	1	1	13
LOCAL	18	19	24
OBRA	9	9	15
ORGANIZACAO	8	10	20
PESSOA	8	8	21
TEMPO	16	22	22
VALOR	12	12	14

Tabela 9.4: Posição do sistema da Priberam no HAREM clássico na CD do TEMPO.

	Classificação		Identificação	
Cenário	Avaliação estrita de ALT	Avaliação relaxada de ALT	Avaliação estrita de ALT	Avaliação relaxada de ALT
Total	2	-	1	-
TEMPO	16	-	16	-
Selectivo 2	8	-	8	-
Selectivo 4	2	-	1	-
Selectivo 6	4	-	1	-

calculados pelo nosso sistema e os propostos pela pista do TEMPO do Segundo HAREM, ainda que em variados casos não tenha sido possível a construção das entidades de acordo com esses critérios (por exemplo, *no domingo, dia 28 de Janeiro* (CD) / *no domingo* (Priberam) / *28 de Janeiro* (Priberam), *em 1996* (CD) / *1996* (Priberam), *do século 21* (CD) / *século 21* (Priberam)). Grande parte das EM consideradas em falta na avaliação da pista do TEMPO deve-se à exclusão das preposições e contracções nas EM pelo nosso sistema.

9.2.2 Resultados da pista do TEMPO

Apesar de os resultados da pista do TEMPO terem sido menos satisfatórios do que os do HAREM clássico, sobretudo pelas razões apontadas no ponto anterior, o sistema posicionou-se no primeiro lugar da identificação na CD do TEMPO, no cenário total, com 0,6939 de EM correctamente identificadas, e no segundo lugar na classificação no mesmo cenário, com 0,5004 de EM correctamente classificadas; nos cenários selectivos 4 e 6 na CD do TEMPO, o sistema da Priberam colocou-se também na primeira posição (ver tabela 9.4).

Na pista do TEMPO, os resultados, tal como no HAREM clássico, são mais elevados em identificação (primeira posição, entre todos os participantes da pista do TEMPO, no cenário total e nos cenários selectivos 4 e 6) e inferiores em classificação, sendo a melhor marca alcançada no cenário selectivo 4 (ver tabela 9.5).

Tabela 9.5: Posição do sistema da Priberam na pista do TEMPO, no modos de avaliação: estendido completo (EC), estendido sem normalização (ESN) e estendido só com normalização (ESCN).

	Classificação			**Identificação**		
Cenário	**EC**	**ESN**	**ESCN**	**EC**	**ESN**	**ESCN**
Total	5	5	5	1	1	1
TEMPO	16	16	16	16	16	16
Selectivo 2	7	7	7	8	8	8
Selectivo 4	2	2	2	1	1	1
Selectivo 6	5	6	7	1	1	1

9.3 Conclusões e trabalho futuro

Nas secções anteriores, descrevemos sucintamente o funcionamento do sistema de REM da Priberam e a sua adaptação ao Segundo HAREM, assim como os respectivos resultados na avaliação.

No âmbito do trabalho desenvolvido pela Priberam, quer a nível de correcção sintáctica, quer a nível de sistemas de resposta automática a perguntas ou ainda em motores de pesquisa, o desenvolvimento e aperfeiçoamento do REM é de grande importância. No corrector sintáctico do FLiP, o REM é crucial para se entenderem determinadas sequências de palavras como unidades morfossintácticas únicas, que irão permitir a correcção de erros de concordância com a unidade completa e não com um dos seus elementos em particular. Permite ainda que a construção da árvore sintáctica das frases seja mais precisa e evite a sobregeração, nomeadamente em casos de entidades que contêm preposições e que poderiam levar à criação de árvores com múltiplos sintagmas preposicionais. No sistema de resposta automática a perguntas, o REM tem também papel relevante, porque permite fazer a equivalência exacta entre os pivôs da pergunta e os textos indexados dos *corpora*. Por exemplo, na pergunta *Quem é Robert Redford?*, Robert Redford é reconhecido como uma EM e a equivalência irá ser feita com a entidade completa, passando a dar-se menos importância aos elementos da locução se aparecerem isolados ou fizerem parte de outras entidades.

Para além da importância da identificação das EM, a sua classificação tem também um papel relevante, uma vez que permite, no caso dos sistemas de resposta automática a perguntas, estabelecer a categoria das perguntas, restringindo assim o leque de respostas possíveis.

O REM é também importante em casos em que é útil realizar listagens para restrição de pesquisas em motores de busca.[9]

Eventos como o HAREM permitem-nos testar em maior escala o sistema e detectar muitas das suas falhas. Há, no entanto, casos em que os nossos objectivos e critérios se distanciam bastante daqueles preconizados em avaliações deste tipo, nomeadamente, no caso do Segundo HAREM, na pista do TEMPO. Se se definir uma EM como uma "entidade com nome próprio" (Santos e Cardoso, 2007c, pp. 3), grande parte das locuções temporais e numéricas não se enquadraria nesta definição. O sistema da Priberam não detectava este tipo de expressões como EM, pelo que os resultados mais baixos nestas categorias podem

[9] O Jornal de Notícias (http://www.jn.pt) e a TSF (www.tsf.pt) usam nos seus sítios um sistema de restrição de pesquisa desenvolvido pela Priberam com o seu módulo de REM.

ser explicados pelo pouco tempo que tivemos para o trabalho de adaptação aos critérios do HAREM.

De qualquer modo, a uniformização das categorias de EM é uma questão problemática e de difícil consenso, assim como o são outro tipo de categorizações semânticas como as ontologias ou outras bases de dados lexicais com relações conceptuais e semânticas. No limite, cada sistema manterá as categorizações que mais lhe convêm, especialmente se estivermos a falar de produtos comerciais que respondem a determinadas necessidades dos utilizadores, sendo inevitável que, para efeitos de avaliação conjunta, os sistemas se adaptem aos critérios estabelecidos pela organização deste tipo de eventos. No caso do Segundo HAREM, conseguimos, apesar de ainda não termos avaliado todos os casos de EM que o sistema não consegue identificar ou classificar, chegar já a algumas conclusões do trabalho que é necessário realizar. Estas conclusões serão complementadas num futuro próximo com a análise detalhada dos ficheiros de avaliação produzidos pelos programas disponibilizados no sítio da Linguateca.

A maior parte das ocorrências de metonímia (Santos, 2007d, pp. 46-49) não é ainda detectada pelo sistema, pelo que casos como *Palácio de Belém* em *o* ***Palácio de Belém*** *pronunciou-se* são marcados como `LOCAL` e não como `PESSOA`. Para tal, poderiam contribuir em larga medida as restrições de selecção dos verbos, isto é, a marcação de valores nos verbos que indiquem o tipo semântico dos seus argumentos (ver (9.12)).

(9.12) Palácio de Belém $_{[sujeito]}$ pronunciou-se $_{[sujeito\ humano|grupo\ humano]}$

Para além da marcação do tipo de argumentos do verbo, também a marcação semântica dos nomes pode revelar-se útil, pois permite, em casos em que certos adjectivos apenas qualificam nomes com determinado campo semântico (ver (9.13)) ou em que determinados nomes apenas se podem relacionar com EM de determinado tipo semântico (ver (9.14)), identificar a categoria certa da EM em causa.

(9.13) Palácio de Belém satisfeito $_{[qualificador\ de\ nome\ humano]}$ pronunciou-se

(9.14) A queixa $_{[+complemento\ de+nome\ humano|grupo\ humano]}$ do Palácio de Belém

Como já foi visto acima, o sistema da Priberam não detectou ou errou sistematicamente a classificação de várias categorias de EM, nomeadamente `ABSTRACCAO/IDEIA`, `ACONTECIMENTO/EVENTO`, `COISA/CLASSE`, `COISA/MEMBROCLASSE`, `COISA/OBJECTO`, `COISA/SUBSTANCIA`, `PESSOA/GRUPOCARGO`, `PESSOA/GRUPOIND`, `PESSOA/MEMBRO`, `PESSOA/POVO`, pelo que terão de ser melhoradas as regras para que sejam detectadas e classificadas EM com estas categorias.

Há ainda outras questões que também terão de ser resolvidas no futuro, nomeadamente o reconhecimento de palavras em início de frase ou após travessão como nomes próprios, que o sistema ainda continua a reconhecer como nomes comuns (por exemplo, *STN – Sistema de Transmissão do Nordeste*).

A ontologia da Priberam, cuja utilidade no desenvolvimento actual do sistema de REM já ficou descrita acima (ver secção 9.1), sendo construída com base em relações semânticas e conceptuais entre palavras e expressões, terá também um papel importante na evolução e melhoramento do sistema, pois poderá auxiliar na extracção de EM através da análise do contexto em que se encontram.

Capítulo 10

R3M, uma participação minimalista no Segundo HAREM

Cristina Mota

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 10, p. 181–193.

O sistema R3M apresentou-se no Segundo HAREM como um sistema de reconhecimento de pessoas, organizações e locais. Optámos por nos cingir a esta categorias, dado que, de uma forma geral, têm sido mais extensivamente estudadas na área de extracção de informação, e não tínhamos disponibilidade de dedicar mais tempo ao desenvolvimento do nosso sistema.

No entanto, o sistema R3M foi desenhado de modo a que fosse flexível, permitindo no futuro estender facilmente o reconhecimento a outras categorias, assim como incluir o reconhecimento de relações entre entidades mencionadas. Além de ser flexível, o sistema caracteriza-se também por fazer um uso mínimo de recursos linguísticos construídos manualmente, sejam estes regras ou textos anotados. Este último critério resulta do facto de tanto regras como textos anotados criados manualmente serem dispendiosos e morosos de obter, como já argumentado por diversos autores que, normalmente, optam por métodos de aprendizagem semi-supervisionados (Ji e Grishman, 2006; Collins e Singer, 1999; Miller et al., 2004) e não supervisionados (Etzioni et al., 2005).

Assim, o nosso sistema assenta numa estratégia de aprendizagem semi-supervisionada que recorre a um algoritmo de co-treino para inferir regras de classificação (Collins e Singer, 1999). O algoritmo de co-treino que Collins e Singer (1999) apresentam tem a grande vantagem de obter bons resultados de classificação que rondam os 80% de correcção (em inglês, *accuracy*) usando apenas um número muito reduzido de exemplos previamente anotados.

Salientamos desde já que a estratégia proposta por estes autores foi aplicada com sucesso ao problema de REM em textos escritos em português por Mota (2009)[1], que introduziu diversas modificações com vista a obter um anotador de entidades em texto e não apenas um classificador de listas de entidades[2]. O sistema R3M é pois uma reimplementação do sistema criado por Mota (2009), apresentando em relação a este diversas melhorias.

Neste capítulo começamos por descrever o sistema R3M, destacando as melhorias que fomos introduzindo (secção 10.1) relativamente ao sistema em que nos inspirámos. Em seguida, na secção 10.2, mostraremos e analisaremos os resultados da nossa participação. Concluímos o capítulo (secção 10.3) mencionando aspectos positivos e negativos da nossa participação no Segundo HAREM.

10.1 Descrição do sistema R3M

A arquitectura geral do sistema R3M, ilustrada na figura 10.1, é idêntica à do sistema implementado por Mota (2009), a qual foi inspirada, como já referimos, na proposta de Collins e Singer (1999).

Muito sucintamente, como se pode ver na figura, trata-se de um sistema modular sequencial, que separa a fase de identificação de entidades mencionadas da sua classificação. Pode ver-se, igualmente, que o sistema envolve uma fase de treino, em que aprende regras de classificação com base num algoritmo de co-treino, e uma fase de teste que usa as regras

[1] Tal como discutido pela autora, a tarefa de REM que realizou era mais semelhante à tarefa proposta na MUC do que no HAREM. Contudo, esse factor tem pouca relevância na arquitectura do sistema, pois são os exemplos anotados usados para treino que condicionam o tipo de regras aprendidas pelo algoritmo de co-treino.

[2] A diferença entre anotador e classificador reside sobretudo no facto de que no primeiro caso o sistema tem por tarefa delimitar e classificar num texto as entidades que encontra, sendo avaliado de acordo com medidas de precisão e de abrangência, enquanto um classificador tem por objectivo classificar uma lista previamente identificada de entidades, sendo avaliado de acordo com uma medida de correcção.

aprendidas para classificar entidades em novos textos. Ambas as fases partilham os módulos de identificação (de entidades e contextos envolventes) e extracção de características (em inglês, *features*). A fase de teste contém ainda um módulo de propagação que produz um texto final anotado.

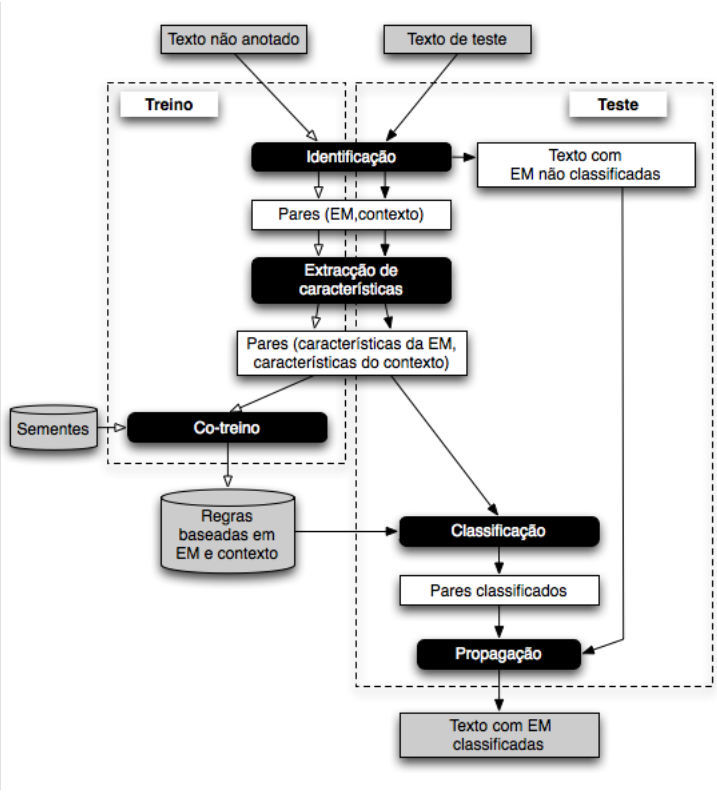


Figura 10.1: Arquitectura geral do sistema R3M

O sistema R3M distingue-se do sistema de Mota (2009) sobretudo ao nível da implementação.

A principal diferença da nossa implementação para a do sistema no qual nos inspirámos é o facto de termos substituído o sistema NooJ (Silberztein, 2004), o qual era usado pelo módulo de identificação, pelo conjunto de ferramentas JET (Grishman, 1999-2006). Este último conjunto de ferramentas apresenta as seguintes vantagens:

- foi concebido a pensar em extracção de informação, mas inclui os vários módulos típicos de processamento de linguagem natural (consulte-se a tabela 10.1, a qual lista os vários módulos e apresenta em destaque os que foram usados nesta fase de desenvolvimento);
- permite criação manual de regras, assim como a sua aprendizagem automática;
- é facilmente portável para outros sistemas operativos, por estar implementado em Java;

Tabela 10.1: Módulos do Jet; as ferramentas marcadas com "X" foram usadas pelo R3M

Módulos do Jet	**Módulos usados pelo R3M**
Atomizador	X
Segmentador de frases	X
Consultador de dicionário	X
Etiquetador morfossintáctico (HMM^3)	X
Etiquetador de EM	-
Analisador de grupos nominais	-
Analisador sintáctico	-
Analisador sintáctico estatístico	-
Reconhecedor de padrões	X
Resolvedor de referências	-

- permite parametrização e criação de pequenos programas de invocação.

Apesar do conjunto de ferramentas Jet ter sido concebido tendo em vista o processamento de textos escritos em inglês, a arquitectura era suficientemente genérica para podermos utilizar os vários módulos no processamento de textos em português. Para isso, seria naturalmente necessário ter dados em português para treinar os módulos que tivessem sido treinados com dados em inglês, ou então criar novas regras no caso em que as regras fossem específicas de inglês.

Quanto aos módulos de classificação e co-treino, originalmente implementados em Lush (Bottou e LeCun, 2003), estes módulos foram reimplementado em R (R Development Core Team, 2008), com o objectivo de no futuro poder vir a tirar partido dos vários módulos de análise estatística existentes no R.

Em seguida descreveremos sucintamente cada módulo e os recursos envolvidos.

10.1.1 Identificação

Quer estejamos numa fase de treino ou de teste, o módulo de identificação é responsável por identificar, em textos não anotados, candidatos a entidades e o contexto em que se encontram. O resultado produzido por este módulo é uma lista de pares constituídos por entidade e contexto.

Este módulo tem duas fases principais: detecção de candidatos a EM e detecção dos contextos em que as entidades ocorrem.

10.1.1.1 Detecção de candidatos a EM

O objectivo da primeira fase é detectar os candidatos a EM. A fase de detecção é composta pelos seguintes passos: atomização, delimitação de frases, consulta de dicionários e aplicação de regras que identificam ou excluem candidatos.

As regras de exclusão tentam evitar que certas sequências de palavras iniciadas em maiúscula sejam marcadas como candidatas. Por exemplo, uma vez que não estamos a fazer reconhecimento de expressões temporais, criámos uma regra de exclusão para os nomes dos meses e de estações do ano. Também excluímos palavras que se iniciassem por maiúscula e que reunissem uma das seguintes condições: (i) fossem palavras vazias (em

inglês, *stopwords*) ou (ii) estivessem ligadas à palavra seguinte por um hífen. A lista de palavras vazias, que constituiu um dos poucos dicionários que usámos[4], foi obtida seleccionando todas as palavras gramaticais (cerca de 3500) contidas no Port4NooJ, o módulo de português para o NooJ descrito em Barreiro (2008).

As regras de identificação de candidatos limitam-se a delimitar sequências de palavras iniciadas por maiúscula. Estas sequências podem também incluir um conjunto limitado de elementos de ligação[5], desde que a palavra inicial e final sejam iniciadas por maiúscula. Dado que foi fornecida pela organização do Segundo HAREM uma lista de palavras em minúscula que podiam iniciar uma entidade mencionada (a lista pode ser consultada no apêndice A, secção A.6), essas palavras se existissem no texto também foram incluídas como fazendo parte do candidato a EM. Esta lista constituiu o outro dicionário que usámos e que contém cerca de 170 entradas[6].

10.1.1.2 Detecção do contexto da EM

Nesta fase, candidatos a EM que se encontrem em determinados contextos definidos por um pequeno conjunto de regras são identificados juntamente com o respectivo contexto. Por contexto deve entender-se uma sequência de palavras que ocorra antes ou depois do candidato a EM. Os pares de candidato a EM e contexto serão fornecidos ao módulo de classificação.

Salientamos que simplificámos os contextos de Mota (2009), de forma a que não necessitássemos de um analisador sintáctico. Precisámos mesmo assim de informação morfossintáctica e por esse motivo treinámos o etiquetador morfossintáctico do Jet com base nos textos da Floresta Sintá(c)tica (Afonso et al., 2002).

Os contextos que considerámos podem não corresponder a um constituinte sintáctico, pois não impusemos nenhuma estrutura sintáctica em particular à sequência de palavras que constituem o contexto. Apenas definimos as seguintes restrições:

- o limite à esquerda, quando existe, de um contexto à esquerda do candidato a EM deve corresponder a: artigo, palavra vazia, preposição ou sequência de dois atómos separados por hífen ou "/";
- o limite à direita de um contexto à esquerda do candidato a EM deve corresponder a: nome, adjectivo ou forma verbal, seguido ou não de vírgula ou qualquer das palavras permitidas como limite à esquerda de um contexto à esquerda;
- o limite à esquerda de um contexto à direita do candidato a EM deve corresponder a: nome, adjectivo, forma verbal ou *que*, antecedido ou não de vírgula;
- o limite à direita, quando existe, de um contexto à direita do candidato a EM deve corresponder a: palavra vazia, preposição ou artigo.

Estas restrições, muito genéricas, foram obtidas por observação de vários exemplos, e ainda precisam de mais experimentação e refinamento.

[4] Cada entrada do dicionário é constituída por uma palavra vazia associada à etiqueta `stw`.

[5] Como elementos de ligação, considerámos as preposições *de*, *em*, *por* e *para* contraídas (excepto no último caso) ou não com o artigo definido, e também os caracteres "-" e "/".

[6] Neste dicionário, as entradas relativas a cargos têm a etiqueta `crg` e as entradas correspondentes a formas de tratamento têm a etiqueta `ft`.

No caso dos contextos à esquerda, estávamos sobretudo a tentar captar (i) contextos em que a entidade estivesse integrada num grupo nominal ou preposicional, (ii) contextos em que a entidade estivesse aposta a um grupo nominal e (iii) contextos verbais de que a entidade pudesse ser o complemento do verbo. A tabela 10.2 ilustra exemplos de entidades e respectivos contextos à esquerda que são detectados por este módulo.

Tabela 10.2: Pares de entidades e contexto à esquerda

Contexto à esquerda	Entidade
Dividir o	IRA
o aeroporto de	Londres
o sector mais violento do	IRA
O seu fundador,	Michael Collins
o segundo mais sobrecarregado com barracas da	Área Metropolitana de Lisboa
As imagens emocionaram o	País
o momento mais incrível do	Mundial
viajava para	Lisboa

Com os contextos à direita tentámos encontrar (i) grupos nominais apostos às entidades ou (ii) construções verbais em que o sujeito poderia ser a entidade. Na tabela 10.3 ilustram-se algumas entidades com o seu contexto à direita.

Tabela 10.3: Pares de entidades e contexto à direita

Entidade	Contexto à direita
Karl Wendlinger	, piloto da
Paikou	, ficou quase completamente submerso pelas
Hamas	está interessado
Aung San Suu Kyi	continua presa
David Bernardino	, prestigiado médico

Criámos ainda regras que associam informação contextual às restantes entidades envolvidas numa estrutura de coordenação de entidades, quando a primeira ou a última das entidades coordenadas tenham sido previamente associadas a informação contextual. No caso de ser a primeira entidade, associa-se o seu contexto à esquerda e, de forma simétrica, no caso de ser a última entidade associa-se o seu contexto à direita. No exemplo 10.1, o contexto à esquerda de *Guiné* vai ser igualmente o contexto de *Angola* e *Moçambique*, enquanto no exemplo 10.2, o contexto à direita de *Foca* vai ser também o contexto de *FIA*.

(10.1) Os novos governos da *Guiné*, de *Angola* e de *Moçambique*

(10.2) *FIA* e a *Foca* omitiram essa informação

Além disso, nos casos em que entidades sejam seguidas de outras dentro de parêntesis, a informação de contexto de uma é associada à outra. No exemplo 10.3, é a informação do contexto à direita de *AR-Santana* que é associada a *Administração Regional de Santana*; no exemplo 10.4 é o contexto à esquerda de *IML* que é associado à entidade dentro de parêntesis.

(10.3) *Administração Regional de Santana* (*AR-Santana*) culpam o

(10.4) o laudo do *IML* (*Instituto Médico Legal*)

Uma vez que era nosso objectivo minimizar a dependência de textos manualmente anotados (neste caso, a Floresta não foi manualmente anotada, mas foi manualmente revista), uma das nossas ideias futuras era limitar o contexto à esquerda e à direita de outro modo, por exemplo, considerar como relevante uma janela de n palavras, ou até que o próximo canditato a EM seja encontrado, em vez de obrigar as palavras limite a serem de uma determinada categoria morfossintáctica.

10.1.2 Extracção de características

O módulo de extracção de características analisa a lista de pares entidade-contexto e cria uma nova lista constituída por pares de vectores de características. Um dos vectores tem as características próprias da entidade, e o outro vector tem as características referentes ao contexto.

Como características da entidade considerámos: a entidade em si, cada constituinte individualmente (excepto elementos de ligação), se a entidade só tem letras maiúsculas e o comprimento da entidade (as entidades com mais de cinco constituintes ficam todas com o mesmo comprimento, seis); como características do contexto usámos: o contexto completo, cada constituinte do contexto e o tipo de contexto (se é à esquerda ou à direita). Em ambos os casos, as palavras vazias não são consideradas constituintes individuais.

Por exemplo, considerando a entidade *Paikou* cujo contexto à direita é *, ficou quase completamente submerso pelas* (ver tabela 10.3), obtém-se o seguinte par de vectores:

```
((entidade=Paikou, inclui=Paikou, sigla=falso, comprimento=1),
 (contexto=, ficou quase completamente submerso pelas,
 inclui=ficou, inclui=submerso, tipo=direito))
```

10.1.3 Classificação

Este módulo determina a classificação dos pares de vectores de características obtidos pelo módulo de extracção de características. Para tal, o módulo usa um conjunto de regras (de classificação) que são inferidas por um algoritmo de co-treino, como explicado na secção 10.1.4.

Uma regra (de classificação) corresponde a um triplo (x,y,z) em que z, designada *precisão* da regra, corresponde a uma estimativa da probabilidade condicional $p(y|x)$ de observar a categoria y quando a entidade tem a característica x. As características tanto podem ser referentes à entidade em si como ao seu contexto.

A classificação de uma entidade (representada por um par de vectores de características) é escolhida usando a regra que tiver maior valor de precisão de entre o conjunto de regras aplicáveis a essa entidade. O conjunto de regras aplicáveis é constituído por todas as regras cuja característica x faça parte do vector de características da entidade.

Como já referimos, cingimos o leque de categorias possíveis às categorias `PESSOA`, `ORGANIZACAO` e `LOCAL`. Adicionalmente, usámos uma categoria extra, `OUTRA`. Esta categoria

não existe no conjunto de categorias da avaliação, e é utilizada para dar conta de entidades que não pertencem a nenhuma das categorias que nos interessavam, mas que podiam ter sido extraídas pelos módulos anteriores.

10.1.4 Co-treino

Tal como descrevemos na secção anterior, as regras de classificação são triplos (*x*,*y*,*z*), em que *x* é uma característica, *y* a categoria associada à característica e *z* a precisão da regra. Estas regras são inferidas incrementalmente de forma semi-supervisionada, usando um algoritmo de co-treino. Este algoritmo parte de um pequeno conjunto de regras e aprende novas regras a partir de pares entidade-contexto não classificados.

O primeiro algoritmo de co-treino foi proposto por Blum e Mitchell (1998) para classificar páginas da rede. A ideia central é aprender alternadamente regras sobre duas vistas diferentes, mas complementares, que se tem sobre um determinado problema a partir de um pequeno conjunto de exemplos classificados e de uma grande quantidade de dados não classificados. No caso da classificação de entidadade mencionadas, tal como proposto por Collins e Singer (1999), uma das vistas é a própria entidade e a outra vista é o contexto em que ela se encontra (o algoritmo 10.1 descreve os passos envolvidos na aprendizagem baseada em co-treino).

Algoritmo 10.1: Algoritmo de co-treino implementado no sistema R3M

Require: S /* Sementes constituídas por características internas classificadas /*
Require: N /* Pares não classificados, (em_i, c_i), em que $em_i = (em_{i1}, \ldots, em_{im})$ é o vector de características extraídas da EM i e $c_i = (c_{i1}, \ldots, c_{1n})$ é o vector de características extraídas do contexto da EM i /*

C /* Pares classificados, (em_i, c_i) /*
regras_EM /* Regras baseadas em características extraídas da EM /*
regras_contexto /* Regras baseadas em características extraídas do contexto da EM /*
$n \leftarrow 5$
$p \leftarrow 0.95$
$\alpha \leftarrow 0.1$
regras_EM $\leftarrow$ S
while $n < 2500$ **do**
  $C \leftarrow$ `Classificar`(N, *regras_EM*)
  regras_EM $\leftarrow$ `Aprender`(*entidades*(C), α, n, p)
  $C \leftarrow$ `Classificar`(N, *regras_contexto*)
  regras_EM $\leftarrow S \cup$ `Aprender`(*contextos*(C), α, n, p)
  $n \leftarrow n+5$
end while
$C \leftarrow$ `Classificar`(N, *regras_EM* $\cup$ *regras_contexto*)
regras_finais $\leftarrow$ `Aprender`(C, α)

A primeira vista que o algoritmo usa é a das entidades, ou seja, o primeiro conjunto de regras a ser aplicado, designadas "*sementes*", contém regras de classificação que dizem respeito a características extraídas de entidades. Estas regras são aplicadas aos pares entidade-contexto não classificados que foram extraídos nos passos anteriores. Por exemplo, se

o conjunto de sementes fosse constituído apenas pela regra (entidade=Paikou, LOCAL, 0,95), todos os pares entidade-contexto cuja entidade fosse *Paikou* seriam classificados como LOCAL.

Em seguida, o algoritmo infere regras de contexto, com base nos contextos dos pares que forem classificados nesse primeiro passo (ou seja, usa a vista do contexto para obter novas regras). Por exemplo, se as características do contexto de um dos pares classificados fosse (contexto= ficou quase completamente submerso pelas,inclui=ficou,inclui=submerso,tipo=direito), seria possível gerar uma regra por cada característica de contexto desta entidade, em que *y* seria a categoria com que o par foi classificado (no caso, LOCAL) e *z* seria a precisão estimada dados todos os pares classificados pelo algoritmo.

As regras de contexto inferidas são usadas para classificar novamente os pares entidades-contexto. A partir dos novos pares classificados, o algoritmo pode agora inferir regras baseadas nas características das entidades.

O passo de classificação usa o mesmo método de classificação descrito na secção 10.1.3. Em cada passo de inferência de regras, apenas as *n* regras mais frequentes por categoria e que tenham uma precisão acima de um certo limiar são adicionadas ao novo conjunto de regras.

As sementes usadas pelo algoritmo de co-treino foram obtidas a partir da colecção dourada do Primeiro HAREM. Para cada entidade classificada como PESSOA, ORGANIZACAO ou LOCAL criámos uma regra (*x*,*y*,*z*) em que *x* é a característica entidade= preenchida com a entidade que ocorre na colecção dourada, *y* é a categoria mais frequente para essa entidade na colecção dourada e *z* é a probabilidade da entidade ter essa categoria estimada a partir da colecção dourada.

Como só estávamos interessados em pessoas, organizações e locais, todas as entidades da colecção dourada que não pertencessem a essa categoria foram passadas para a categoria OUTRA, excepto entidades TEMPO e VALOR que foram ignoradas. Desta forma poderíamos treinar o sistema com quatro categorias, em que uma delas representa exemplos negativos, em vez de treinar só com as três em que estávamos interessados.

Os pares entidade-contexto não classificados utilizados pelo algoritmo foram extraídos da colecção do Primeiro HAREM, de acordo com os passos ilustrados na fase de treino da figura 10.1 (ver secção 10.1); a colecção dourada do Mini-HAREM foi usada como colecção de teste durante a fase de desenvolvimento do sistema.

10.1.5 Propagação

Este módulo só é aplicado se estivermos numa fase de teste, de modo a produzir a anotação final do texto. A sua função é reconhecer entidades que não se encontram nos contextos representados nas regras descritas na secção 10.1.1.2, mas que podem ser idênticas a entidades que já foram reconhecidas nas fases anteriores e que têm uma classificação associada.

Tomemos como exemplo a EM *Portugal* nas frases 10.5 e 10.6.

(10.5) De regresso ao reino de *Portugal*, «mais cheio de glórias que de despojos», foi bem acolhido por D. Manuel

(10.6) Em *Portugal*, o Instituto Nacional de Saúde elaborou cenários de uma eventual pandemia de gripe humana de origem em aves

No primeiro caso (10.5), *Portugal* encontra-se num contexto que é capturado pelas regras de contexto: *De regresso ao reino de*, formando-se assim um par entidade-contexto que será fornecido ao algoritmo de classificação; porém, no segundo caso, *Portugal* não se encontra num contexto previsto pelas regras e, portanto, essa ocorrência não vai ser analisada pelo modo de classificação.

O módulo de propagação vai então analisar essa ocorrência como uma entidade cuja classificação será a classificação que mais vezes é produzida para *Portugal* na fase de classificação.

Caso *Portugal* ocorra integrado noutro candidato a entidade, essa ocorrência será ignorada pelo módulo de propagação. Por exemplo, na frase 10.7, *Portugal* encontra-se integrado numa entidade maior, *Artes Tradicionais de Portugal*, que também não foi reconhecida num contexto previsto nas regras. De forma a não segmentar essa entidade (e dado que optámos por não produzir anotações com ALT), essa ocorrência de *Portugal* não é tida em conta.

(10.7) foi aberta a exposição internacional «*Artes Tradicionais de Portugal*»

Essencialmente, este módulo é utilizado para aumentar a abrangência do sistema, uma vez que permite a classificação de entidades que não foram classificadas pelo módulo de classificação, por falta de contexto. No entanto, como o módulo de propagação se limita a escolher a classificação mais frequente, não tendo em conta mais nenhuma informação, a classificação pode não ser a correcta, o que poderá fazer diminuir a precisão.

Veja-se, por exemplo, que *Portugal* em 10.8 deverá ser classificado de forma diferente (`PESSOA|ORGANIZACAO`) do que em 10.6 (`LOCAL`). No entanto, o módulo de propagação atribui a ambos a mesma classificação.

(10.8) Quando Granada caiu e a reconquista cristã se impôs então a toda a Península, os dois reinos católicos, *Portugal* e Espanha

As entidades classificadas pelo módulo de classificação como `OUTRA` são ou ignoradas por este módulo ou anotadas apenas como entidades, sem conter os atributos de classificação. Neste último caso, a ausência de classificação tem o significado de que o sistema identifica a sequência delimitada como uma entidade e que a sua classificação não é nenhuma das três que queríamos analisar.

10.2 Resultados

Devido a problemas no módulo de aprendizagem das regras de classificação que não foram resolvidos atempadamente, os resultados da nossa participação no Segundo HAREM acabaram por ficar reduzidos à identificação de entidades mencionadas.

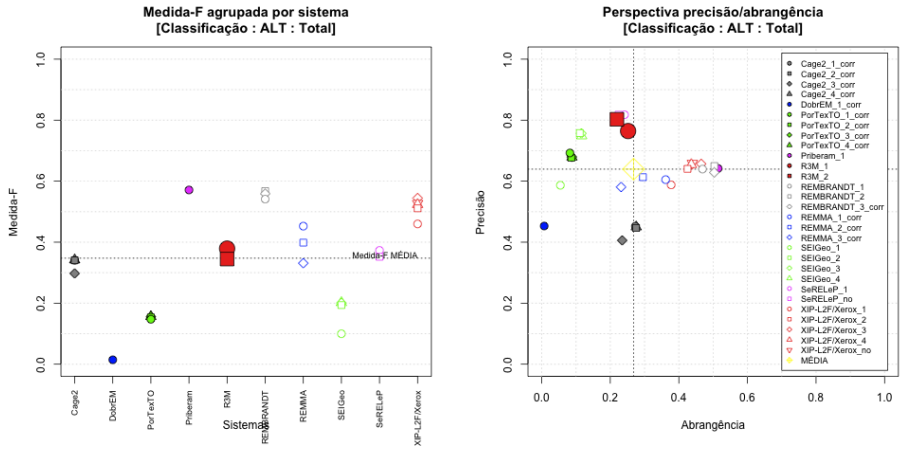
A nossa ideia inicial era participar com duas corridas. Uma corrida incluiria todas as entidades classificadas como `PESSOA`, `ORGANIZACAO` e `LOCAL`, e também as entidades classificadas como `OUTRA` (neste último caso, a anotação não incluiria classificação, de modo a indicar que o sistema as identificou como entidades, mas não as conseguiu classificar como pertencendo a uma das três classes que pretendia reconhecer). A outra corrida incluiria apenas as entidades classificadas como `PESSOA`, `ORGANIZACAO` e `LOCAL`, o que quer dizer que as entidades classificadas como `OUTRA` seriam descartadas antes de produzir o resultado final.

Pretendíamos deste modo verificar se era preferível manter as entidades no resultado final, mesmo que não se soubesse a sua classificação, participando num cenário mais ambicioso (todas as categorias menos `VALOR` e `TEMPO`), ou anotar apenas `PESSOA`, `ORGANIZACAO` e `LOCAL`, participando apenas num cenário selectivo com essas três categorias.

Como os problemas que ocorreram levaram a que não pudéssemos distinguir as entidades que seriam classificadas pelo algoritmo como `OUTRA` das restantes (pessoas, organizações e locais), pois o algoritmo começou a associar a todas as entidades a mesma categoria, acabámos por participar com duas corridas num cenário selectivo com todas as categorias menos `VALOR` e `TEMPO` (cenário selectivo 3):

- R3M_1, que inclui todas as entidades que são identificadas, mesmo as que não se encontram em contextos previstos pelas regras de detecção de contextos;
- R3M_2, que inclui apenas as entidades que são identificadas em contextos previstos pelas regras de detecção de contextos, e ainda as entidades que são reconhecidas pelo módulo de propagação.

Apesar de não termos feito classificação, começamos por mostrar na figura 10.2 o desempenho obtido pelas nossas corridas na classificação das entidades no cenário total com avaliação estrita de ALT, que corresponde ao cenário ideal que os sistemas deveriam alcançar.



(a) Medida F (b) Precisão e abrangência

Figura 10.2: Resultados de classificação no cenário total com avaliação estrita de ALT

Embora os resultados não sejam naturalmente os desejáveis (a melhor corrida, R3M_1, obteve 0,3790 de medida F, ficando em 12º lugar), são equiparáveis aos de outros sistemas que fizeram classificação, como o REMMA, ou ainda melhores, como em comparação com o Cage2. Em particular, na figura 10.2(b) vê-se que o ponto mais forte do sistema R3M em relação a estes dois sistemas é ter um dos melhores valores de precisão que ronda os 0,8, para valores relativamente semelhantes de abrangência (apenas uma das corridas

Tabela 10.4: Identificação no cenário selectivo 3 (todas as categorias excepto TEMPO e VALOR)

Saída	Posição (em 25)	Precisão	Abrangência	Medida F
R3M_1	3	0.7768	0.8134	0.7947
R3M_2	5	0.8116	0.7064	0.7553

do REMMA se destaca com um valor mais elevado). Tal como discutido no capítulo 6, isso mostra que a identificação tem um peso talvez demasiado grande em relação ao da classificação na medida de avaliação.

Centrar-nos-emos, agora, na avaliação da identificação, pois interessa-nos sobretudo perceber o desempenho do sistema na detecção de entidades. Se as entidades não estiverem a ser bem identificadas e delimitadas o algoritmo de aprendizagem estará a treinar sobre dados com mais ruído. Como não utilizámos a etiqueta ALT, mostraremos apenas resultados obtidos com avaliação relaxada de ALT.

Como se pode ver na tabela 10.4, que mostra os resultados obtidos no cenário selectivo 3, a corrida R3M_1 ficou em terceiro lugar na identificação com uma medida F de 0,7947 enquanto a corrida R3M_2 ficou em quinto lugar com 0,7553 de medida F. Estes valores confirmam que apesar de o sistema R3M não ter feito classificação, teve um bom desempenho na identificação das entidades que se propôs reconhecer.

Também se pode ver, e como seria de esperar, que a corrida R3M_1 tem maior abrangência (cerca de 0,11 a mais) do que a corrida R3M_2, pois inclui todas as entidades identificadas na fase de detecção, independentemente do contexto em que ocorrem. Mesmo assim, a precisão dessa corrida é apenas ligeiramente menor (cerca de 0,04) do que a da corrida R3M_2, o que mostra que pode não haver grande vantagem em descartar entidades no caso de não se saber a sua classificação (que é o que a corrida R3M_2 pretende simular: entidades que tenham sido identificadas na fase de detecção de entidades são eliminadas quando se verifica que não ocorrem em pelo menos um contexto previsto pelas regras). Por exemplo, na frase 10.9, *Hugo Estenssoro* e *Londres* são inicialmente identificadas como entidades, cujos contextos não estão previstos nas regras de contexto. Como não existe nenhuma ocorrência de *Hugo Estenssoro* num contexto que possa ser usado para a classificar, esta entidade não fez parte da corrida R3M_2, apesar de fazer parte da corrida R3M_1; *Londres*, como ocorre noutros contextos previstos nas regras, por aplicação do módulo de propagação acabaria por ser reconhecida (e está então anotada em ambas as corridas).

(10.9) *Hugo Estenssoro*, em *Londres*

10.3 Comentários finais

Quando participámos no Primeiro HAREM com o sistema Stencil/NooJ (Mota e Silberztein, 2007), adaptámos um sistema que estávamos na altura a desenvolver para anotar semi-manualmente o CETEMPúblico (Rocha e Santos, 2000) com entidades mencionadas (Mota, 2006). Contudo, essa adaptação não foi total. Em particular, não seguimos o modelo semântico do HAREM, não tentando anotar de forma distinta, por exemplo, *Portugal* nas frases (1.1) a (1.5), ilustradas no capítulo 1: em todos os casos tentámos atribuir a categoria LOCAL.

No Segundo HAREM, ao contrário do que fizemos no Primeiro, que reconhecemos como incorrecto, optámos por seguir mais fielmente as "regras do jogo". Queremos com isto dizer que tentámos desenvolver um sistema que estivesse conforme ao modelo semântico do HAREM e às directivas de anotação. Se assim não fosse, acreditamos que estariamos a enfraquecer a validade do objectivo principal de uma avaliação conjunta que é comparar o desempenho dos sistemas numa tarefa que é comum a todos os participantes.

O maior sucesso da nossa participação foi termos reutilizado um conjunto de ferramentas genéricas que tinham sido desenvolvidas com vista ao processamento de textos escritos em inglês, e aplicado essas ferramentas conjuntamente com recursos portugueses que existiam ou que tivemos de criar.

Sem contar com o facto de a reimplementação só por si constituir uma melhoria do ponto de vista técnico em relação ao sistema em que nos inspirámos, durante o desenvolvimento do sistema de base, fomos melhorando alguns aspectos em relação a esse sistema. Em particular:

- simplificámos as regras de detecção do contexto de EM, o que passou por dispensar um módulo de análise sintáctica;
- incluímos exemplos negativos na fase de aprendizagem, o que evitou criar regras manuais na fase de detecção para excluir entidades de categorias que não queremos reconhecer (o que de certa forma obrigaria a ter praticamente regras para as reconhecer de forma a ter um elevado grau de sucesso na sua exclusão).

Como trabalho futuro gostaríamos de explorar três questões que acábamos por não ter oportunidade de implementar:

- Integrar um módulo de selecção de textos antes da fase de treino, cujo objectivo seria seleccionar textos que pudessem potenciar o resultado do classificador num conjunto de teste. Por exemplo, Ji e Grishman (2006) mostraram que seleccionando frases anotadas mais relevantes é mais importante do que aumentar simplesmente o número de frases do conjunto de treino.
- Detectar o contexto sem necessitar de ter informações morfossintácticas.
- Usar contextos (anotados) também como sementes. Dado que o modelo semântico do HAREM depende fortemente do contexto, estamos em crer que seria mais importante usar contexto como sementes do que entidades classificadas.

Agradecimentos

Agradeço os comentários valiosos e construtivos de Diana Santos, Luís Costa, Bruno Martins, Cláudia Freitas, Hugo Oliveira e Paula Carvalho, em diversas fases de redacção do capítulo, que contribuíram para a sua melhoria substancial.

Capítulo 11

REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto

Nuno Cardoso

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 11, p. 195–211.

O REMBRANDT (**R**econhecimento de **E**ntidades **M**encionadas **B**aseado em **R**elações e **AN**álise **D**etalhada do **T**exto) é um sistema de reconhecimento de entidades mencionadas (REM) e de detecção de relações entre entidades (DRE), projectado para reconhecer todo o tipo de entidades mencionadas (EM) em textos escritos em português. O REMBRANDT explora intensamente a Wikipédia como fonte de conhecimento, e aplica um conjunto de regras gramaticais que aproveitam os vários indícios internos e externos das EM para extrair o seu significado (McDonald, 1996).

O REMBRANDT foi desenvolvido no âmbito do meu doutoramento, que foca o problema de reformulação automática de consultas realizadas a um motor de busca de âmbito geográfico, de uma forma semântica (Cardoso, 2008), e faz parte da linha de investigação seguida no projecto GReaSE (http://xldb.di.fc.ul.pt/wiki/Grease), que procura dotar os motores de busca com capacidade de raciocínio geográfico (Silva et al., 2006). O REMBRANDT nasce da necessidade de desenvolver uma ferramenta de anotação de texto capaz de reconhecer EM que possuam uma forte ligação a locais geográficos, como é o caso de nomes de países, cidades, rios, universidades, monumentos ou sedes de organizações. As aplicações do REMBRANDT envolvem a detecção de âmbitos geográficos nas consultas dos utilizadores, e a geração de "*assinaturas geográficas*" dos documentos, ou seja, listas de EM geográficas que traduzem o âmbito geográfico de cada um dos documentos, e que são usadas na recuperação e ordenação de documentos segundo critérios geográficos.

A tarefa de REM inclui vários desafios, e em relação às pretensões do REMBRANDT o problema de desambiguação de sentidos merece particular destaque; os nomes geográficos podem ser usados em diversos contextos, como é o caso de nomes de pessoas (por exemplo, *Camilo Castelo Branco*) ou de organizações (por exemplo, *France Press*), podem ser usados de forma metonímica (por exemplo, *Varsóvia* para citar o pacto) e até podem designar entidades geográficas bem diferentes (por exemplo, *Cuba* designa um país e uma cidade portuguesa). Santos e Chaves (2006) fazem uma análise sobre os contextos das EM geográficas. Assim sendo, os objectivos do REMBRANDT não se limitam ao reconhecimento de EM geográficas, abrangendo portanto todas as EM relevantes no texto precisamente para facilitar o processo de desambiguação de EM geográficas e para poder situar melhor o contexto da EM.

O REMBRANDT está disponível a todos de forma gratuita, incluindo o código fonte, sob a licença GPL em http://xldb.di.fc.ul.pt/Rembrandt.

11.1 Inspiração para o REMBRANDT

A estratégia dependente da língua do REMBRANDT foi inspirada em parte pelo sistema de REM criado por Bick (2007), o PALAVRAS_NER, e que obteve os melhores resultados nas tarefas de identificação e de classificação de EM do primeiro evento de avaliação do primeiro HAREM, em Abril de 2006.

O sistema PALAVRAS_NER baseia-se no analisador morfossintáctico PALAVRAS (Bick, 2003), que usa uma sintaxe própria para criar regras manuais que exploram os indícios das EM no texto. Contudo, as semelhanças entre o PALAVRAS_NER e o REMBRANDT acabam por aqui, uma vez que o REMBRANDT usa um sistema próprio de criação e aplicação de regras gramaticais, e utiliza a Wikipédia como base de conhecimento para a classificação de EM.

A ideia de usar a Wikipédia em vez de um almanaque para assistir o REMBRANDT na sua tarefa surge através dos trabalhos recentes em torno da Wikipédia, e que evidenciam

as potencialidades que este recurso possui para as tarefas de extracção de informação (Wu e Weld, 2007). Um exemplo disso é o trabalho de Auer e Lehmann (2007), que explora as caixas de informação (em inglês, *infoboxes*) das páginas da Wikipédia para gerar conhecimento em forma de factos representados por triplas em RDF. O projecto associado ao seu trabalho, o `DBpedia.org`, contava em 2008 com cerca de 100 milhões de triplas RDF extraídas automaticamente a partir da Wikipédia (Auer et al., 2007).

O REMBRANDT foi inicialmente concebido para usar a Wikipédia como se se tratasse de um simples almanaque, mais actualizado e vasto do que os almanaques usados por outros sistemas de REM. No entanto, o funcionamento do REMBRANDT rapidamente evoluiu para tirar partido da riqueza da informação e estrutura da Wikipédia, permitindo inclusive a prospecção de informação adicional sobre cada EM, como acontece no caso de extracção de informação geográfica implícita (Cardoso et al., 2008b).

O REMBRANDT possui uma interface própria para interagir com a Wikipédia, a SASKIA, com o objectivo de facilitar as tarefas de navegação na estrutura de categorias, ligações e redireccionamentos da Wikipédia com vista à extracção de conhecimento. Já existe, por exemplo, o RENOIR (Santos et al., 2008a), que é uma ferramenta de construção de consultas semânticas que usa a API da SASKIA para executar consultas específicas à colecção da Wikipédia.

11.2 Anatomia do REMBRANDT

O REMBRANDT suporta vários formatos de ficheiros, na leitura e na escrita (texto simples, HTML ou XML), e pode ser executado em qualquer plataforma que possua uma máquina virtual de Java. Para processamento de grandes quantidades de texto, o REMBRANDT pode funcionar em regime de mapeamento e redução (em inglês, *MapReduce*) (Dean e Ghemawat, 2008) através da sua extensão para o Apache Hadoop (`http://hadoop.apache.org`), permitindo a distribuição das tarefas de REM por vários computadores disponíveis.

A figura 11.1 resume o funcionamento do REMBRANDT. Os documentos são tratados, um de cada vez, numa linha de processos de anotação sucessivos até à sua versão final e definitiva, tal como por exemplo em Gruhl et al. (2004). Ao longo da linha de processos, as EM entretanto reconhecidas vão guardando um historial de alterações desde que são detectadas pela primeira vez até à sua última modificação. Este sistema de rastreio de EM facilita a depuração das acções de cada processo, permitindo a afinação do sistema de regras e leis a aplicar a cada EM, bem como vigiar as suas aplicações ao longo da vida de cada EM. O funcionamento do REMBRANDT pode ser subdividido em três etapas principais:

i. Reconhecimento de expressões numéricas e geração de candidatas a EM

Os textos são previamente divididos em frases e em unidades, com a ajuda do atomizador da Linguateca (disponível através do módulo de Perl `Lingua::PT::PLN`, em `http://search.cpan.org/~ambs/Lingua-PT-PLN-0.17`). Um primeiro conjunto de regras identifica expressões numéricas no texto, tais como unidades compostas só por algarismos, ou números por extenso, ordinais e cardinais. De seguida são aplicadas regras para reconhecer expressões temporais e valores, tirando proveito dos números já reconhecidos no passo anterior.

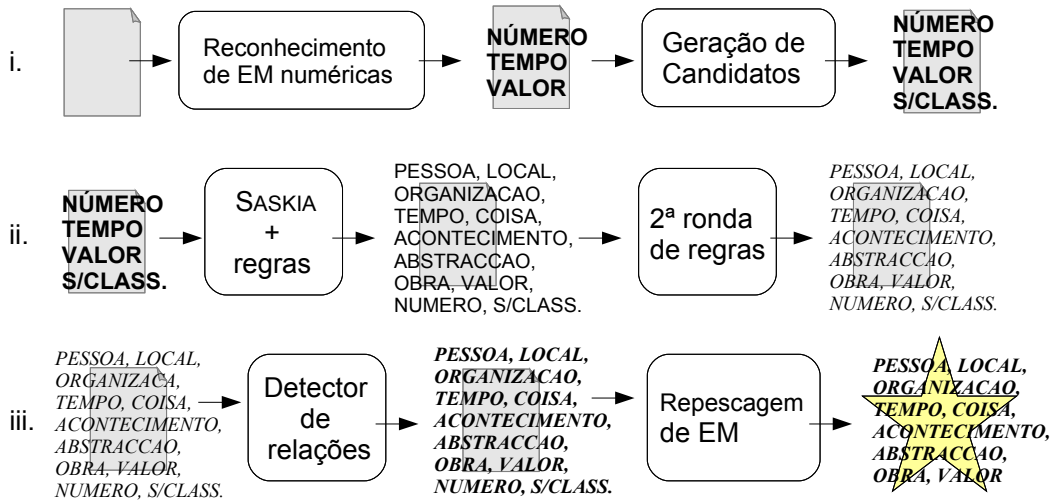


Figura 11.1: O funcionamento do REMBRANDT, dividido em três etapas.

A geração de candidatas a EM é feita através da identificação de sequências de unidades contendo pelo menos uma letra maiúscula e/ou um algarismo, podendo existir uma das seguintes unidades, desde que não comece ou termine a EM: *de*, *da*, *do*, *das*, *dos*, *e* (doravante designados por unidades *daeose*, devido à sua expressão regular `'d[aeo]s?|e'`).

ii. Classificação de EM

Na segunda etapa, cada uma das candidatas a EM é classificada primeiro pela SASKIA, e depois é novamente classificada através de regras gramaticais. Esta estratégia de "dupla classificação" das EM tem um conjunto de vantagens: primeiro, a SASKIA realiza uma classificação de acordo com vários significados que a EM pode ter, e que são reunidos nas páginas típicas de desambiguação da Wikipédia. Desta forma, cria-se um ponto de partida a partir do qual o processo de desambiguação pode trabalhar com vista à selecção do significado correcto da EM. Segundo, as regras gramaticais englobam indícios externos e internos das EM, o que de certa forma supervisiona as classificações da SASKIA segundo o contexto da EM.

Para ilustrar o funcionamento das regras, considere o exemplo dado pela frase (11.1), onde a SASKIA classificou previamente a EM *Angola* como sendo `LOCAL/HUMANO/PAIS`. A aplicação de uma regra gramatical dedicada à captura de ruas vai mudar a classificação da EM *Angola* para `LOCAL/HUMANO/RUA`, devido à presença da expressão *Rua da* antes da EM.

(11.1) Eu moro na Rua de Angola.

A terminar a etapa de classificação, é aplicada uma segunda ronda de regras gramaticais, que aproveita as classificações existentes para detectar EM com uma morfologia mais elaborada. É nesta fase que, por exemplo, as EM que possuem um termo *daeose* são analisadas na sua eligibilidade para serem representadas através de uma etiqueta `<ALT>`,

ou se porventura é melhor serem divididas em EM mais pequenas, que serão novamente classificadas pela SASKIA e pelas regras gramaticais.

iii. Repescagem de EM sem classificação

Na última etapa, realiza-se a detecção de relações entre EM através de um conjunto de regras específicas para a tarefa. As relações entretando detectadas são usadas para repescar algumas EM sem classificação, mas que estão relacionadas com EM devidamente classificadas.

Após a tarefa de DRE, é feita uma última repescagem de EM com nomes de pessoas, através de uma comparação com uma lista de nomes comuns. Por último, as EM que persistem sem classificação são eliminadas, bem como números por extenso sem uma letra maiúscula (uma vez que não são considerados EM segundo as directivas em vigor, excepto no caso de expressões temporais). As EM de categoria `NUMERO` são convertidas em `VALOR/QUANTIDADE`.

11.3 SASKIA

A SASKIA é a interface responsável por pré-processar as colecções da Wikipédia, e por realizar uma classificação inicial às EM, com base na informação extraída da Wikipédia. A API da SASKIA permite realizar operações simples de interacção com a colecção da Wikipédia, como por exemplo a navegação nas páginas, a extracção de categorias, a recolha e filtragem de âncoras ou a normalização de títulos das páginas.

11.3.1 Pré-processamento da Wikipédia

A Wikipédia gera periodicamente imagens estáticas dos conteúdos relativos a cada língua, disponibilizadas em http://download.wikipedia.org, onde podem ser acedidas pelo público em geral. Estas imagens são compostas por vários ficheiros em XML e ficheiros em SQL, consoante o nível de informação associada (por exemplo, a inclusão ou não das páginas de discussão, das páginas dos utilizadpres, ou do histórico de alterações das páginas).

Os ficheiros em XML contêm o texto das páginas no seu formato MediaWiki original (http://meta.wikimedia.org/wiki/Help:Editing), enquanto os ficheiros em SQL incluem o código para a criação das tabelas (metadados das páginas, ligações entre páginas, informação das categorias e tabela de redireccionamentos) e os respectivos dados.

A SASKIA foi desenvolvida inicialmente em torno do ficheiro em XML das páginas portuguesas da imagem estática da Wikipédia em português. O pré-processamento do ficheiro era feito através de uma versão modificada do programa Wikipedia Preprocessor (http://sourceforge.net/projects/wikiprep), extraindo as ligações, o texto das âncoras, as categorias, os títulos, os subtítulos, as listas ordenadas, as páginas relacionadas, os URL externos e o texto filtrado de cada documento. A informação extraída era armazenada e indexada através do Lucene (http://lucene.apache.org).

No entanto, esta estratégia serve perfeitamente para imagens da Wikipédia semelhantes à portuguesa, cujo ficheiro em XML (de cerca de 1,4 GB de tamanho, em Março de 2008) é pré-processado em poucas horas; para imagens como a versão inglesa da Wikipédia, com cerca de 28 GB de tamanho, em Fevereiro de 2008, o tempo de pré-processamento é proibitivamente longo. Assim sendo, após a participação no segundo HAREM, a SASKIA

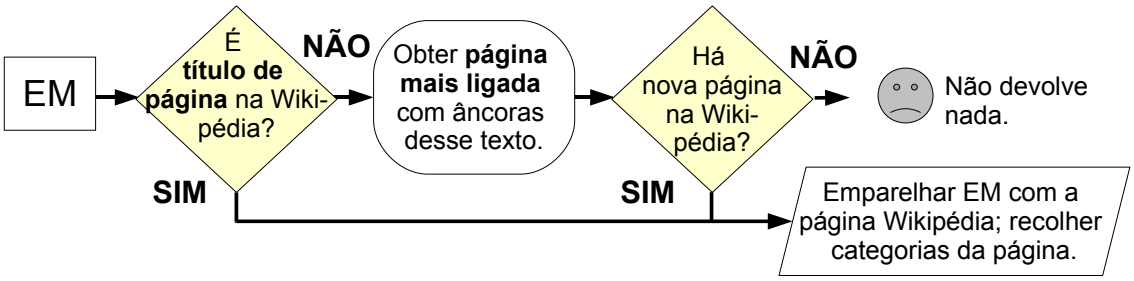


Figura 11.2: Emparelhamento de EM realizado pela SASKIA.

foi melhorada de forma a explorar as imagens da Wikipédia em formato SQL, podendo ser pré-configurada de forma a escolher a fonte de informação (XML ou SQL) a usar para cada uma das acções da sua API (a partir da versão 0.8 do REMBRANDT).

11.3.2 Estratégia de classificação

O procedimento de classificação da SASKIA usado no Segundo HAREM está dividido em três etapas: (i) associação da EM a uma página da Wikipédia, (ii) recolha de categorias associadas à EM, e (iii) mapeamento das categorias da Wikipédia às classificações do HAREM. Para evitar consultas repetidas à SASKIA, esta possui uma memória temporária (em inglês, *cache*) interna que guarda os resultados das classificações anteriores, acelerando significativamente o tempo de resposta para EM já analisadas previamente.

i. Emparelhamento de EM

A SASKIA começa por procurar uma página da Wikipédia com o título exactamente igual ao texto da EM. Se for encontrada, passa-se para a etapa seguinte; se não for encontrada, o texto da EM é usado para encontrar a página mais ligada através de âncoras cujo texto é idêntico ao texto da EM (ver figura 11.2).

O uso das ligações entre páginas da Wikipédia permite à SASKIA lidar com as várias formas de designação de uma mesma entidade. Um exemplo é as diversas formas de designar a entidade *Estados Unidos da América* (país): *EUA*, *USA*, *Estados Unidos*, *E.U.A.* ou *América do Norte*, que são tudo formas vulgares e abreviadas de referir a mesma entidade. Uma vez que a grande maioria das ligações da Wikipédia com o texto da âncora contendo estas variantes aponta para a página (http://pt.wikipedia.org/wiki/Estados_Unidos_da_América), o emparelhamento das EM faz-se de uma forma simples e robusta.

Esta estratégia de análise do texto das âncoras só pode ser usada se a SASKIA tiver ao seu dispor a versão pré-processada em XML da Wikipédia. Caso se opte por usar a versão em SQL da Wikipédia para realizar o emparelhamento das EM, a SASKIA recorre à tabela de redireccionamentos para encontrar a página respectiva. A principal desvantagem desta opção é que o emparelhamento fica dependente da existência de uma entrada de redireccionamento explícito na tabela SQL.

ii. Recolha de categorias

Para cada uma das categorias da página da Wikipédia emparelhada na etapa anterior, a SASKIA analisa o seu tipo e, caso seja necessário, visita mais páginas relacionadas e extrai as suas categorias, adicionando-as à lista. A SASKIA adopta uma estratégia de profundidade-primeiro na sua navegação entre páginas, limitada até ao quarto nível de profundidade (ver figura 11.3). Os tipos de categoria da Wikipédia que a SASKIA reconhece são os seguintes:

Categoria normal. Estas categorias são simplesmente adicionadas à lista de categorias.

Autocategoria. Designo por autocategoria toda a categoria que possui o mesmo nome do título da página que a contém. Por exemplo, a página da Wikipédia da cidade do Porto (`http://pt.wikipedia.org/wiki/Porto`) possui a autocategoria `Categoria:Porto`. Nestes casos, a SASKIA analisa a página da categoria (`http://pt.wikipedia.org/wiki/Categoria:Porto`) e adiciona as suas categorias à lista.

Categoria de desambiguação. A `Categoria:Desambiguação` é usada nas páginas da Wikipédia que esclarecem os diversos significados da EM, e que reúnem ligações para as respectivas páginas desambiguadas. Nestes casos, a SASKIA extrai as ligações da página que possuam o texto da EM na âncora (com base no XML), ou as ligações para páginas da Wikipédia cujo título contém o texto da EM (com base no SQL), visita as páginas referenciadas e recolhe as suas categorias.

Categoria de acrónimo. A `Categoria:Acrónimos` é usada nas páginas da Wikipédia cujo título é um acrónimo. A SASKIA procura a presença desta categoria nas páginas de desambiguação, para que a extracção de ligações para outras páginas se faça sem usar o texto da EM (que é um acrónimo). Exceptuando este caso, esta categoria não é utilizada para nenhum fim, e não é adicionada à lista de categorias.

Um exemplo ilustrativo de uma página com as categorias `Acrónimos` e `Desambiguação` é a página da Wikipédia da PSP (`http://pt.wikipedia.org/wiki/PSP`). Sendo uma página de desambiguação, as suas ligações apontam para páginas sobre entidades com significados distintos, como a Polícia de Segurança Pública, a PSP PlayStation Portable ou o Paint Shop Pro.

Contudo, como a sigla *PSP* é um acrónimo, o texto da EM não pode ser directamente usado para filtrar as ligações. Nesses casos, a SASKIA compara o acrónimo e a ligação, e determina se o texto da âncora (com base no XML) ou o título da página-alvo (com base no SQL) representa uma expansão do acrónimo. Se tal se verificar, a nova página é então visitada e as suas categorias são recolhidas.

A utilização do texto das EM para filtrar as ligações da página de desambiguação é essencial, uma vez que nem todas as suas ligações são relevantes (por exemplo, pode haver uma ligação para a página de Portugal na frase de descrição sumária do sentido da Polícia de Segurança Pública).

iii. Classificação das categorias

Finalmente, a SASKIA aplica uma lista de regras gramaticais específicas sobre cada uma das categorias, com o objectivo de extrair o seu significado e a sua referência geográfica,

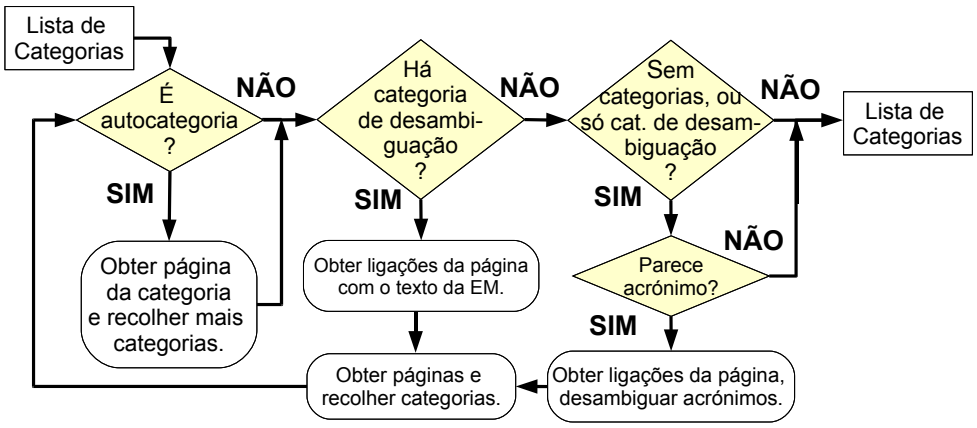


Figura 11.3: Recolha de categorias pela SASKIA.

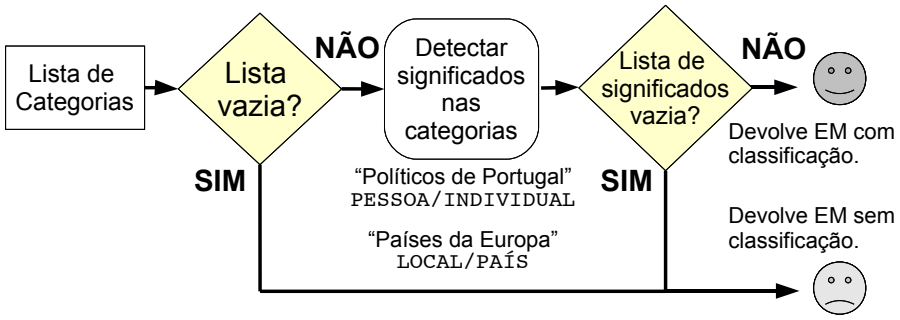


Figura 11.4: Classificação das categorias pela SASKIA.

caso exista (ver figura 11.4). Por exemplo, a categoria *Cantores de Portugal* possui uma morfologia frequentemente usada na Wikipédia portuguesa, para o qual foi criada uma regra que procura uma definição, seguida de um termo *daeose* opcional e de uma entidade geográfica (tanto como nome, como na sua variante em adjectivo, como acontece com a categoria *antigas Províncias portuguesas*).

Após a aplicação válida desta regra, a definição *cantores* é mapeada à respectiva classificação do HAREM (`PESSOA/INDIVIDUAL`) com a ajuda de um almanaque de definições interno. A entidade geográfica *Portugal* é recolhida, desambiguada e associada à EM como se tratasse de informação geográfica implícita sobre o âmbito geográfico da EM, conforme descrito por Cardoso et al. (2008b).

11.4 Regras gramaticais

As regras gramaticais representam padrões nas frases que indiciam a presença de EM com determinadas propriedades semânticas, e definem as acções a tomar quando estas são apli-

cadas com sucesso. As regras são compostas por uma ou mais cláusulas, ou seja, unidades de padrões mais simples.

As cláusulas são aplicadas ordenadamente, uma de cada vez, a uma parte seleccionada da frase. Cada cláusula retorna verdade se as unidades alinhadas com esta corresponderem ao seu padrão, ou retorna falso no caso oposto. Se todas as cláusulas retornarem verdade, a regra diz-se bem sucedida e retorna por sua vez um valor verdadeiro. Em contraste, se pelo menos uma das cláusulas retornar falso, ou se a frase terminar e ainda houver cláusulas obrigatórias para aplicar, a regra falha e retorna um valor falso.

Quando a regra é bem sucedida e retorna verdade, segue-se a execução de uma acção pré-determinada, definida pelo campo **Acção**. Desta forma, é possível definir regras com actuações diferentes, como é o caso de regras de DRE, regras de geração de `<ALT>` ou regras de geração de novas EM.

As regras gramaticais usadas na etapa de classificação de EM (regras de detecção de indícios internos e externos) possuem o campo **Acção:GerarEM** para que a sua aplicação resulte em novas EM com novas classificações definidas na regra através dos campos **categoria**, **tipo** e **subtipo**. Adicionalmente, o campo **PolíticaDaRegra** define a política de escolha das unidades que irão fazer parte da nova regra, e pode tomar dois valores: i) **Regra**, o que inclui todas as unidades capturados por todas as cláusulas da regra (normalmente usado para regras de índicio interno) e ii) **Cláusula**, onde cada cláusula especifica se as unidades capturadas por ela vão ser incluídas ou não na nova EM (normalmente usado para regras de indício externo).

11.4.1 Propriedades das cláusulas

As cláusulas também possuem propriedades próprias que descrevem a sua forma de actuação. As propriedades mais importantes de uma cláusula são as seguintes:

Cardinalidade, que define se a cláusula é obrigatória ou opcional, e determina o número de vezes que pode ser aplicada. A cardinalidade pode tomar os seguintes valores: i) **Zero ou um**, semelhante à semântica de '.?' das expressões regulares. A cláusula é opcional, e devolve verdadeiro quer tenha sido correspondida ou não a um conjunto de unidades. ii) **Zero ou mais**, semelhante à semântica de '.*?' das expressões regulares. A cláusula é opcional, e devolve verdadeiro independentemente das vezes que conseguir ser correspondida. iii) **Um**, semelhante à semântica de '.' das expressões regulares. A cláusula é obrigatória e é executada uma única vez, devolvendo verdadeiro se for correspondida. iv) **Um ou mais**, semelhante à semântica '.+' das expressões regulares, onde é obrigatório haver pelo menos uma correspondência verdadeira. As cláusulas adoptam uma estratégia gananciosa (em inglês, *greedy*), procurando a maior sequência de unidades possível, antes de passar para a cláusula seguinte (se existir).

Critério, que define o tipo de correspondência, e pode tomar os seguintes valores: i) **Simples**, que faz uma comparação simples entre unidades, ii) **Expressão**, que aplica uma expressão regular a um termo, iii) **EM**, que procura a presença de uma EM com um determinado leque de classificações, e iv) **Conceito**, que usa uma lista de conceitos, comparando cada elemento dessa lista, um por um, até encontrar uma expressão que corresponda às unidades.

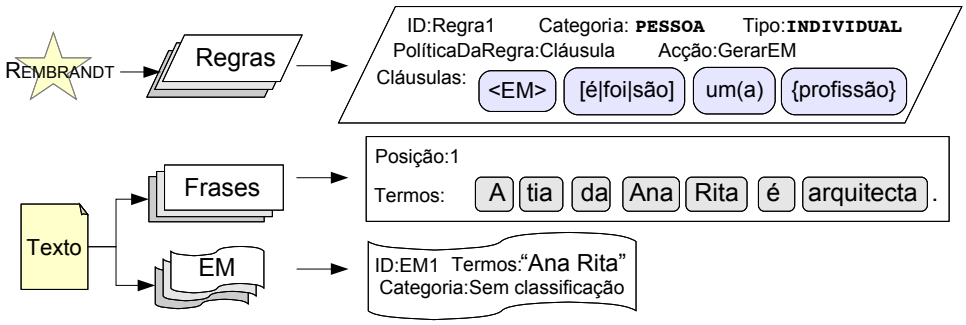


Figura 11.5: Selecção de regras gramaticais, frases e de EM.

Padrão, que instancia o padrão a ser aplicado na comparação, de acordo com o critério. Assim sendo, o padrão pode incluir um termo, uma expressão regular, uma lista de classificações, ou uma lista de conceitos (ou seja, uma lista de listas de expressões regulares).

Inclusão, que define se as unidades correspondidas pela cláusula irão fazer parte da EM, no caso da regra ter o campo **Acção:GerarEM**. Este campo é lido só se a respectiva regra definir o campo **PolíticaDaRegra:Cláusula**.

11.4.2 Aplicação das regras

A aplicação das regras ao texto é feita de uma forma sequencial, uma regra de cada vez, a todas as frases do texto (também de forma sequencial, da primeira para a última frase). Para cada frase do texto, a regra activa começa pelo primeiro termo da frase, e invoca sucessivamente cada uma das cláusulas. Após esse passo, a regra muda o seu posicionamento para um termo à direita, até serem esgotadas todas as combinações possíveis de alinhamento da regra com a frase. As regras bem sucedidas são logo executadas, e no caso das regras de geração de EM, as novas EM ficam imediatamente disponíveis para serem usadas na aplicação das regras seguintes.

Esta forma de encadeamento de regras de acordo com a sua ordem inicial permite a elaboração de regras sequenciais. Por exemplo, para capturar a EM *entre Abril e Maio*, é usada uma primeira regra que reconhece os meses, e depois é aplicada uma segunda regra, que procura um padrão `entre <EM> e <EM>` para então juntar as unidades todas numa nova EM.

A figura 11.5 ilustra a aplicação da regra gramatical com o identificador `Regra_1` à frase (11.2). Note-se que a EM *Ana Rita*, previamente reconhecida como candidata a EM e que se encontra de momento sem classificação, foi invocada para esta aplicação pois faz parte da frase. As propriedades da `Regra_1` determinam que, caso seja bem sucedida, irá gerar uma nova EM que terá a classificação de `PESSOA/INDIVIDUAL`, e que a escolha das unidades da nova EM será feita pelas cláusulas.

(11.2) A tia da Ana Rita é arquitecta.

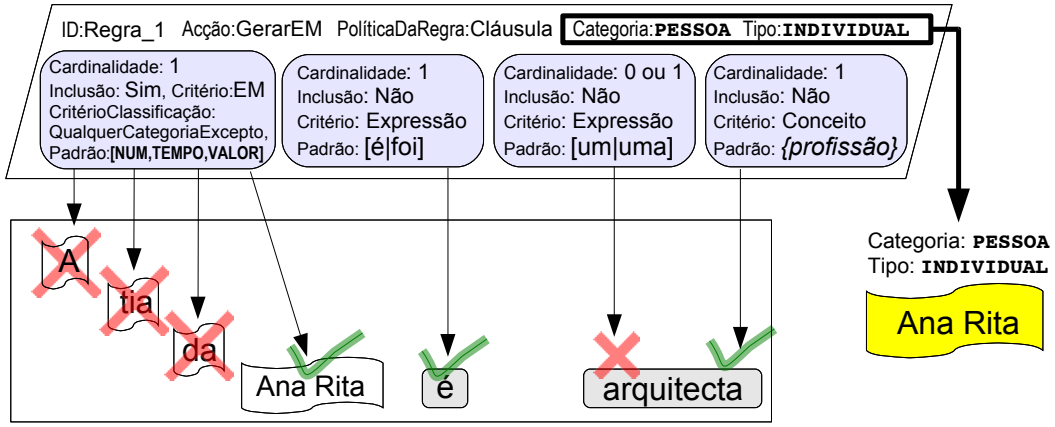


Figura 11.6: Aplicação de regras gramaticais.

A aplicação da regra começa com o alinhamento da primeira cláusula com o início da frase. Esta cláusula procura encontrar uma EM (**Critério:EM**) que não tenha nenhuma das seguintes classificações: `NÚMERO`, `TEMPO` e `VALOR` (**CritérioClassificação:QualquerCategoriaExcepto**). Como tal, a cláusula irá falhar sucessivamente quando alinhada às unidades *A*, *tia* e *da*, uma vez que não fazem parte de nenhuma EM (sempre que uma cláusula falha, a regra é então repetida com um novo alinhamento à direita, geralmente de um termo). Finalmente, a EM *Ana Rita* é então correspondida pela primeira cláusula, e os seus duas unidades são guardadas devido ao campo **Inclusão:Sim**.

De seguida, a regra passa para a cláusula seguinte, que é alinhada ao termo seguinte. Esta segunda cláusula procura um padrão no termo que corresponda a *é* ou *foi*, e é obrigatório encontrar esse termo para que a regra seja bem sucedida. Como o termo *é* é encontrado, é a vez da terceira cláusula, que é opcional visto que possui o campo **Cardinalidade:Zero ou Um**. Este tipo de cláusulas são úteis para representar pequenas variações nas morfologias das frases que se procura detectar, como é o caso de *(...) é uma arquitecta* ou *(...) é arquitecta*. As cláusulas opcionais retornam sempre positivas, independentemente de conseguirem ser correspondidas às unidades.

Finalmente, a quarta e última cláusula, de **Critério:Conceito**, procura a definição de uma profissão/ocupação através de uma lista de expressões regulares que podem ser simples (por exemplo, `['[Aa]rquitec?t[oa]s?']`) ou compostas (por exemplo, `['[Tt][êê]cnic[oa]s?','[Oo]ficia[li]s?', 'de', '[Cc]ontas']`). Quando uma das expressões regulares é correspondida, a regra verifica que não há mais cláusulas a satisfazer e retorna com sucesso, gerando então a nova EM `<EM CATEG=¨PESSOA¨ TIPO=¨INDIVIDUAL¨>Ana Rita</EM>` (ver figura 11.6).

11.4.3 Tribunal de EM

Quando as regras geram novas EM que se sobrepõem a EM já existentes, diz-se que há um "*conflito*" entre EM. Assim sendo, o REMBRANDT possui um tribunal de EM, onde os

conflitos entre EM são resolvidos. No tribunal, a EM "*ré*", que já existia na lista de EM do documento, e a EM "*acusadora*", recém-gerada pela regra, esgrimem argumentos para que se possa decidir o seu destino. O tribunal dispõe de um conjunto de leis de resolução de conflito, de onde é seleccionada a lei mais adequada para o conflito em questão, e do qual sai um veredicto que pode incluir desde a eliminação de uma das EM, da geração de alternativas `<ALT>`, até à fusão das EM numa única.

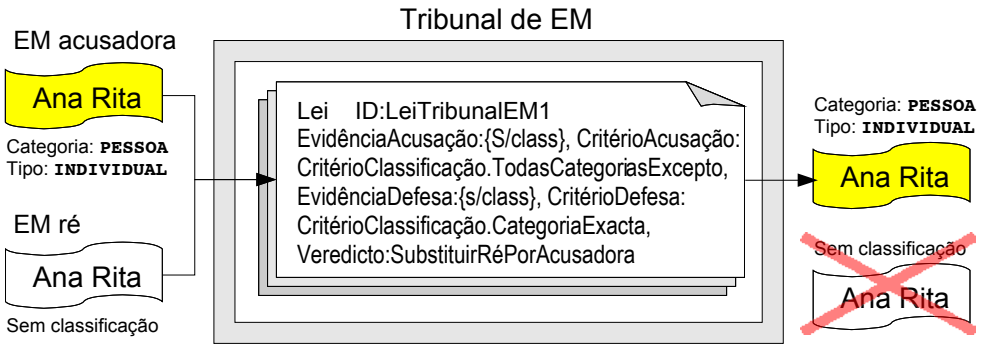


Figura 11.7: Exemplo de aplicações de uma lei no tribunal de EM.

A figura 11.7 ilustra a aplicação das leis para a situação retratada na figura 11.6. A lei aplicada diz que uma EM sem classificação deverá sempre ser substituída por uma EM equivalente com uma classificação válida. Os indícios das EM consideradas nas leis incluem leques de classificações e formas de sobreposição entre EM em conflito (ou seja, se uma EM está contida noutra, se sobreposta, se está contida e ajustada à esquerda, entre outros).

O tribunal permite uma certa organização e prioritização das EM geradas pelas regras. No entanto, a passagem pelo tribunal é um comportamento por omissão, feito se não houver indicação em contrário na regra; se for necessário, é possível definir o campo **PolíticaConflito** na regra, para definir previamente o veredicto a tomar em caso de conflitos. Este campo serve, por exemplo, para aplicar em regras de captura de `<ALT>`, onde não há propriamente um conflito entre EM, mas sim uma interpretação alternativa do sentido das EM envolvidas.

11.5 Detecção de relações entre EM

O sistema de detecção de relações do REMBRANDT usa heurísticas básicas de relacionamento entre EM com base nas suas unidades, nas suas categorias e nas ligações das respectivas páginas da Wikipédia. As heurísticas são aplicadas a EM não-numéricas (isto é, sem classificação `VALOR`, `NUMERO` ou `TEMPO`) e seguem o seguinte procedimento:

1. EM com o mesmo texto são rotuladas como sendo idênticas (`ident`). As EM que foram emparelhadas à mesma página da Wikipédia também são rotuladas como sendo idênticas; desta forma, EM como por exemplo *Cavaco Silva* e *Aníbal Cavaco Silva* são associadas com a etiqueta `ident`.

2. EM que se sobrepõem a outras EM (no caso de `<ALT>`) ou que são separadas por um termo *daeose* são analisadas nas suas classificações, que determinam o tipo de relação entre elas. Por exemplo, a relação `ocorre_em` é usada quando uma EM com categoria `ACONTECIMENTO` se sobrepõe ou é vizinha de uma EM de categoria `LOCAL`, como acontece em *Jogos Olímpicos de Pequim*; a relação `sede_em` é usada na mesma situação, mas com uma EM com categoria `CONSTRUCAO`, como acontece em *Museu Militar do Porto*. No final são repescadas relações `ident` a EM que possuem texto em comum e alinhado a um extremo, como acontece com nomes de pessoas, por exemplo, *José Sócrates* e *Sócrates*.

3. EM que estejam emparelhadas a páginas da Wikipédia são analisadas de forma a encontrar relacionamentos com EM vizinhas na mesma frase, através das ligações da página. Os textos das âncoras da página (usando a Wikipédia em XML) ou os títulos das páginas-alvo (usando a Wikipédia em SQL) podem indiciar uma relação entre as EM, como é ilustrado pelo exemplo das EM *Neil Armstrong* e *NASA*; uma vez que a página da Wikipédia do astronauta contém uma ligação para a página da NASA, é adicionada uma relação `outra` entre estas duas EM.

4. Finalmente, é aplicada uma série de regras gramaticais vocacionadas para detectar relações entre EM numa mesma frase e que ainda não possuem relações entre elas. Essas regras gramaticais definem o campo **Acção:GerarRelação**, e possuem cláusulas com o campo **Papel**, que identificam o papel de cada uma das EM visadas, e o tipo de relacionamento detectado.

O mecanismo de detecção de relações entre EM do REMBRANDT ainda está nos seus passos iniciais, no entanto o seu papel será determinante para a desambiguação de sentidos de EM com várias classificações. Por exemplo, a EM *Armstrong* é classificada pela SASKIA como sendo um local e uma pessoa ao mesmo tempo; contudo, se na sua vizinhança existir a EM *NASA*, a detecção de relações pode assinalar que afinal o seu sentido é de um nome de pessoa.

11.6 Resultados no Segundo HAREM

O REMBRANDT enviou um total de três corridas para o Segundo HAREM. A fonte de informação usada pela SASKIA foi o ficheiro em XML relativo à imagem estática da Wikipédia de 2 de Março de 2008, que conta com 405.752 páginas e 5.010.715 ligações. A geração de corridas foi realizada no regime de mapeamento e redução do Hadoop v0.15, num grupo (em inglês, *cluster*) de 7 máquinas Linux com 19 processos de mapeamento e 7 processos de redução, tendo demorado uma média de 100 minutos para etiquetar a colecção HAREM.

Na altura do Segundo HAREM, a etapa de DRE foi a mais pesada em termos de processamento do REMBRANDT chegando a contribuir com mais de metade do tempo de processamento. Para este facto contribui a falta de optimização do sistema de DRE, que procura relações para todas as combinações de pares de EM possíveis que ainda não possuem uma relação, fazendo com que o tempo de operação evolua de forma exponencial em relação ao número de EM do documento, e chegando a vários minutos num documento longo.

11.6.1 Corridas

As três corridas foram geradas por diferentes versões 0.7 do REMBRANDT. As diferenças entre versões limitaram-se à rectificação de alguns erros no funcionamento do REMBRANDT e ao melhoramento da SASKIA e das regras gramaticais, após uma análise das saídas geradas. Em resumo:

11.6.1.0.1 Corrida REMBRANDT_1. Gerada pelo REMBRANDT 0.7.1, possui uma versão não testada de regras gramaticais focadas na detecção de <ALT>, uma nova interface de escrita de <ALT>, e o sistema de detecção de relações parcialmente programado, mas ainda não afinado quanto à sua estratégia adoptada e nas regras gramaticais usadas.

11.6.1.0.2 Corrida REMBRANDT_2. Gerada pelo REMBRANDT 0.7.2, possui melhoramentos da SASKIA a nível da estratégia em páginas de desambiguação e de acrónimos, e foram afinadas as regras próprias para os <ALT>, abrangendo mais casos elegíveis. O sistema de relações foi aperfeiçoado de maneira a propagar a eliminação de relações, o que acontecia frequentemente devido à eliminação de EM sem classificação na última etapa, deixando órfãs muitas das relações encontradas.

11.6.1.0.3 Corrida REMBRANDT_3. Gerada pelo REMBRANDT 0.7.3, possui um sistema de detecção de relações afinado com um leque final de regras. Foram rectificados vários problemas a nível da escrita da saída. Ao nível da SASKIA, a navegação entre páginas para a recolha de categorias realiza-se agora com profundidade quatro em vez de três, e possui melhorias a nível de resolução de acrónimos.

11.6.2 Resultados na tarefa de REM

Corrida	Avaliação estrita de ALT			Avaliação relaxada de ALT		
	Precisão	Abrangência	Medida F	Precisão	Abrangência	Medida F
REMBRANDT_2	0,6497	0,5036	0,5674	0,6622	0,5173	0,5808
REMBRANDT_3	0,6286	0,5032	0,5590	0,6424	0,5163	0,5725
REMBRANDT_1	0,6396	0,4690	0,5412	0,6505	0,4809	0,5530

Tabela 11.1: Resultados do REMBRANDT no HAREM clássico, no cenário total.

A tabela 11.1 apresenta os resultados globais do REMBRANDT, no HAREM clássico. A corrida REMBRANDT_2 obteve os melhores valores, o que é um resultado curioso, uma vez que a corrida REMBRANDT_3 foi gerada com o propósito de rectificar vários problemas observados na corrida REMBRANDT_2.

A tabela 11.2 apresenta os resultados do REMBRANDT discriminados por categoria. Salienta-se o facto de o REMBRANDT ter tido um bom desempenho na categoria LOCAL e no respectivo cenário selectivo que incluía somente EM de classificação LOCAL/HUMANO e LOCAL/FISICO, o que é particularmente relevante do ponto de vista da sua aplicação em sistemas de recuperação de informação geográfica.

Os resultados para as EM de categoria PESSOA beneficiaram bastante dos passos de repescagem de EM, quer pela detecção de relações, quer pelo uso de um almanaque de

Categoria	Melhor corrida	Classificação			Identificação		
		P	A	F	P	A	F
PESSOA	REMBRANDT_3	0,7683	0,5368	0,6320	0,7747	0,5411	0.6371
LOCAL	REMBRANDT_1, 3	0,5484	0,6607	0,5993	0,5553	0,7241	0,6286
VALOR	REMBRANDT_3	0,4127	0,7176	0,5241	0,4161	0,7247	0,5287
TEMPO	REMBRANDT_3	0,5904	0,4030	0,4790	0,6098	0,4093	0,4899
ORGANIZACAO	REMBRANDT_3	0,5350	0,3231	0,4029	0,6035	0,3624	0,4529
OBRA	REMBRANDT_3	0,5251	0,2171	0,3072	0,5276	0,2188	0,3093
ACONTECIMENTO	REMBRANDT_3	0,5630	0,2026	0,2980	0,6312	0,2242	0,3308
ABSTRACCAO	REMBRANDT_3	0,1956	0,1433	0,1655	0,2085	0,1534	0,1768
COISA	REMBRANDT_2	0,0451	0,0566	0,0502	0,0425	0,0724	0,0536

Tabela 11.2: Resultados do REMBRANDT discriminados por categoria, e ordenados pela medida F da tarefa de classificação.

nomes, uma vez que a SASKIA está mais vocacionada para reconhecer nomes de celebridades, e as regras gramaticais revelaram-se insuficientes para abranger os variados indícios externos de nomes de pessoas.

No caso oposto, os resultados para as EM de categoria VALOR saíram algo prejudicados pela conversão automática de EM de categoria NUMERO para VALOR/QUANTIDADE. Este problema também se reflecte na abrangência das EM de categoria TEMPO, muito por culpa da dificuldade em detectar o verdadeiro significado dos números que representam anos. O quinteto de categorias com melhores resultados é fechado pela categoria ORGANIZACAO, que não teve o desempenho que se esperava devido à estratégia algo simplista de geração de candidatas a EM, que precisa de ser revista e adaptar-se para outras morfologias de EM.

11.6.3 Resultados na tarefa de DRE

Os resultados da pista do ReRelEM são analisados de uma forma global no capítulo 4. A tabela 11.3 apresenta os resultados obtidos pelo REMBRANDT na tarefa de detecção de relações entre EM, o ReRelEM. O cenário total diz respeito a todas as EM presentes na CD, enquanto que o cenário selectivo 5 diz respeito às EM de categoria LOCAL e de tipo HUMANO ou FISICO, ou seja, EM de cariz geográfico.

Cenário	Medida	Melhor corrida	Total			Melhor corrida	Selectivo 5		
			P	A	F		P	A	F
Todas	Relações	REMB. 1	0,5822	0,3669	0,4502	REMB. 1	0,9178	0,6204	0,7403
Identidade	Relações	REMB. 1	0,7723	0,6934	0,7307	REMB. 3	0,9184	0,9000	0,9091
Inclusão	Relações	REMB. 1	0,3236	0,3261	0,3243	REMB. 1	0,9615	0,4098	0,5747
Localização	Relações	REMB. 2	0,4048	0,1288	0,1954	REMB. 1	-	-	-

Tabela 11.3: Resultados do REMBRANDT na pista do ReRelEM.

No geral, a corrida REMBRANDT_1 obteve os melhores resultados em DRE se considerarmos a medida F. As outras duas corridas, apesar de obterem valores de medida F próximos, nota-se que sacrificaram a precisão para aumentar de forma ténue a abrangência, o que in-

dica que as alterações introduzidas nas corridas REMBRANDT_2 e REMBRANDT_3 também introduziram muito ruído.

Em mais detalhe, nota-se que o REMBRANDT é eficaz na detecção de relações de identidade para todo o tipo de EM (cerca de 0,73 de medida F), mas não na detecção de relações de inclusão (0,32) e de localização (0,19). Contudo, para as EM de cariz geográfico os valores são mais elevados, com cerca de 0,9 na identificação de relações entre entidades geográficas, e 0,57 na detecção de inclusões.

Em resumo, o REMBRANDT destaca-se pela positiva na tarefa de detecção de relações para entidades geográficas, o que é encorajador dado os propósitos do REMBRANDT de extracção de pistas geográficas do texto. O desempenho do REMBRANDT neste capítulo ainda possui uma margem de progressão considerável, face ao conjunto simples de regras de DRE usadas.

11.7 Conclusões e trabalho futuro

O REMBRANDT é um sistema de REM muito ambicioso, propondo-se etiquetar todo o tipo de EM existentes no texto, e detectar o tipo de relacionamento entre elas, a partir de uma estratégia de regras gramaticais manuais co-adjuvadas por um sistema de extracção de conhecimento automático a partir da Wikipédia, a SASKIA. Assim sendo, é essencial acompanhar a evolução do seu desempenho ao longo do seu desenvolvimento, de forma a afinar adequadamente todos os diversos passos que compõem a linha de processamento do REMBRANDT.

A participação do REMBRANDT no Segundo HAREM reveste-se de particular importância, pois permitiu ter uma primeira noção do desempenho do REMBRANDT nas suas tarefas, em particular do seu nível de eficiência em relação ao reconhecimento de entidades geográficas, e à detecção de relações entre elas. Os resultados obtidos são animadores e mostram que a estratégia adoptada pelo REMBRANDT permite obter desempenhos satisfatórios em REM.

Após a participação no Segundo HAREM, o REMBRANDT já foi melhorado em diversos aspectos, e há uma lista de melhoramentos a realizar no REMBRANDT a curto e médio prazo, nomeadamente:

Abstracção da camada de classificação. O REMBRANDT foi desenvolvido em torno da hierarquia de classificação adoptadas pelo Segundo HAREM, estando esta codificada de raiz no funcionamento do REMBRANDT. Para trabalho futuro, o REMBRANDT irá suportar diferentes leques de categorização de EM, permitindo a sua adaptação a domínios mais específicos e o aumento da resolução da classificação.

Adaptação para várias línguas. O REMBRANDT v0.8 já suporta várias línguas na sua tarefa de REM, embora não possua ainda forma de detectar automaticamente a língua do documento. A adaptação a outras línguas requer a re-escrita das regras e das leis do REMBRANDT e a readaptação das definições da Wikipédia às características da imagem respectiva. Actualmente, o REMBRANDT já processa textos na língua inglesa, embora o seu desempenho esteja ainda aquém do desempenho observado para o português.

Detecção de contextos. O REMBRANDT precisa de uma "3ª ronda" de regras gramaticais específicas para capturar o contexto mais genérico das EM, de forma a adaptar-se

melhor à metodologia de REM sugerida pelo HAREM. Um exemplo típico é a utilização de entidades geográficas em contextos abstractos, como é patente na frase *A honra da França estava em jogo*, que o HAREM reconhece como sendo uma ABSTRACCAO/IDEIA. Uma vez que a EM *França* não é referida num papel geográfico, tal facto tem de transparecer na forma de actuação do REMBRANDT. Em estudo está a possibilidade de esta 3ª ronda de regras ser realizada através de métodos de aprendizagem automática, uma vez que é difícil representar o contexto de forma objectiva e contundente através de regras gramaticais.

Pré-processamento da Wikipédia. A SASKIA precisa de optimizar o seu processo de pré-processamento dos ficheiros de XML, de forma a lidar convenientemente com ficheiros de grandes dimensões. Para tal, está prevista a implementação de um pré-processador próprio, em vez de usar uma ferramenta externa, que consiga capturar e organizar a informação contida nas caixas de informação. A SASKIA também está a ser melhorada de forma a usar uma camada abstracta de representação dos documentos, de forma a que o seu funcionamento não dependa do tipo de ficheiro (XML ou SQL) usado no pré-processamento, e permitindo a exploração de outras fontes de informação (a DBpedia, por exemplo).

Melhorias na API da SASKIA. A SASKIA deverá ser capaz de explorar mais informação a partir das páginas da Wikipédia, como é o caso de coordenadas geográficas, os primeiros parágrafos do texto, ou as caixas de informação. Adicionalmente, a SASKIA terá de se adaptar às novas tendências de organização de categorias da Wikipédia, onde começa a ser comum encontrar categorias que são constituídas por subcategorias, numa hierarquia de dois níveis que é algo semelhante à categorização usada no HAREM.

Re-utilização de conhecimento adquirido durante a anotação. O âmbito de acção do REMBRANDT está restringido ao nível do documento, ou seja, não há transposição de conhecimento adquirido entre documentos. O REMBRANDT poderá tirar partido de um centro de armazenamento de conhecimento, onde poderá guardar informação importante sobre EM normalmente usadas, e desta forma agilizar a anotação de novos documentos. Por exemplo, a EM HAREM poderá estar explicitamente descrita num determinado documento, mas noutro documento o seu significado poderá ser muito difícil de extrair, devido à falta de indícios no texto.

Desambiguação de sentidos a partir da DRE. A fraca abrangência obtida pelo REMBRANDT na tarefa ReRelEM indicia que existe uma margem de progresso considerável nesta fase. O REMBRANDT irá depender em grande parte da detecção de relações para a desambiguação de sentidos das EM, em particular das EM geográficas.

Capítulo 12

REMMA - Reconhecimento de Entidades Mencionadas do MedAlert

Liliana Ferreira, António Teixeira e João Paulo da Silva Cunha

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 12, p. 213–229.

Este capítulo descreve o sistema REMMA (Reconhecimento de Entidades Mencionadas do MedAlert), um reconhecedor de entidades mencionadas que usa a Wikipédia como fonte de conhecimento externo. O REMMA foi desenvolvido no âmbito do projecto MedAlert – Sistema de Processamento de Linguagem Médica (http://www.ieeta.pt/sias/medalert). O MedAlert usa a informação disponibilizada pela Rede Telemática de Saúde (RTS) (Cunha et al., 2006), em utilização no Hospital Infante D. Pedro e na região de Aveiro, e tem como principal objectivo a utilização de técnicas de extracção automática de informação de textos médicos, de modo a inferir, de uma forma automática, irregularidades/dúvidas suscitadas pelas decisões tomadas pelos profissionais de saúde. O MedAlert, que deverá tomar a forma dum módulo escalável e adaptável a diferentes configurações de sistemas de informação hospitalares, pretende usar técnicas de processamento de linguagem natural (PLN) para extrair informação de um amplo conjunto de textos médicos, particularmente cartas de alta e textos contendo directivas médicas. Esta informação, bem como a proveniente de recursos externos como ontologias e outras fontes de conhecimento médico, deverá ser utilizada no apoio e validação de decisões, melhorando, assim, o cuidado médico, com a redução de erros, melhoria de segurança e aumento da satisfação. O REM é considerado como uma subtarefa importante da maioria das aplicações de engenharia de linguagem e um primeiro passo para a extracção de informação. Deste modo, tornou-se essencial o desenvolvimento de um módulo capaz de identificar e classificar entidades que respondam a um conjunto de perguntas relevantes, reduzindo, consequentemente, a complexidade da extracção de factos.

O REMMA é apresentado neste capítulo no âmbito da sua participação no Segundo HAREM e, consequentemente, como um sistema de REM em textos não especializados. Esta participação teve como objectivo principal explorar diferentes abordagens e perceber qual a utilidade da utilização de fontes de conhecimento externo na tarefa de REM, para uma posterior adaptação à área em que nos concentramos, a medicina.

O REM foi definido nas conferências MUC (Hirschman, 1998) como sendo a tarefa de detectar e classificar expressões em texto que pertençam a diferentes classes (por exemplo, pessoa, local, organização, data, tempo). Desde que o REM apareceu, duas principais aproximações foram adoptadas para lidar com a tarefa. Uma é referida como baseada em conhecimento e usa explicitamente recursos tais como regras e almanaques construídos e mantidos, de uma forma geral, manualmente. A outra segue o paradigma da aprendizagem automática e usa normalmente como colecção de treino um corpo anotado que é usado para o treino de um algoritmo de aprendizagem supervisionada. Inicialmente, e principalmente para as conferências MUC, a maior parte dos sistemas REM usavam uma aproximação baseada em conhecimento. Este método provou obter bons resultados, tendo, o melhor sistema obtido, na medida F, uma classificação de 0,9339 (Mikheev et al., 1998). No entanto, esta aproximação apresenta um problema relevante: os almanaques e as regras são difíceis de construir e manter, sendo, em particular, difícil evitar a sobreposição entre almanaques.

O REMMA tenta contornar esta questão através da utilização da Wikipédia como fonte de conhecimento externo, em particular, através da extracção de categorias semânticas a partir da primeira frase de uma página da Wikipédia.

12.1 A Wikipédia como fonte de conhecimento para REM

Recentemente, tem-se vindo a assistir a um crescimento rápido e bem-sucedido da Wikipédia (http://www.wikipedia.org), uma enciclopédia electrónica livre e que está a ser construída por milhares de colaboradores em todo o mundo. A Wikipédia tinha em Outubro de 2008 mais de 2 561 000 artigos na versão inglesa e cerca de 428 000 artigos na sua versão portuguesa. Uma vez que a Wikipédia pretende ser uma enciclopédia, a maior parte dos artigos são sobre entidades mencionadas e mais estruturados do que texto livre. A Wikipédia é actualizada diariamente, ou seja, novas entidades são adicionadas e revistas constantemente (Voss, 2005). Deste modo, a extracção de conhecimento a partir da Wikipédia para o PLN é uma forma promissora de permitir a criação de aplicações em grande escala, aplicáveis em situações da vida real. De facto, vários estudos surgiram recentemente em que a Wikipédia é explorada como fonte de conhecimento (Auer et al., 2007; Ruiz-Casado et al., 2006; Santos et al., 2008a; Wu e Weld, 2007; Zesch et al., 2008). A maior parte destes estudos concentram-se na extracção automática de almanaques da Wikipédia (Toral e Muñoz, 2006) e na utilização da estrutura interna da Wikipédia para a desambigação de entidades mencionadas (Bunescu e Pasca, 2006). O estudo com mais relevância para o trabalho apresentado neste capítulo é o de Kazama e Torisawa (2007), onde se utiliza o sintagma nominal da primeira frase de um artigo da Wikipédia para a extracção da categoria semântica. No REMMA, optou-se por identificar na primeira frase do artigo um conjunto de palavras indicativas da categoria e tipo de uma dada entidade. Com este trabalho pretende-se determinar até que ponto as classificações semânticas extraídas a partir de um artigo da Wikipédia, em particular da primeira frase do artigo, podem ser consideradas como *definições* da entidade descrita no artigo. Por exemplo, o artigo da Wikipédia sobre a Universidade de Aveiro começa com a frase (12.1).

(12.1) *A Universidade de Aveiro (UA) é uma universidade pública portuguesa localizada em Aveiro.*

A extracção da palavra *universidade* desta frase permite inferir a classificação a atribuir à entidade *Universidade de Aveiro*. O método utilizado na obtenção destas classificações é descrito em detalhe na secção 12.2.

Várias razões determinaram a escolha da Wikipédia para utilização como fonte de informação no REMMA. A principal foi a necessidade de desenvolver um sistema capaz de reconhecer entidades de domínio geral e a impossibilidade de construir ou aceder a um almanaque de grande dimensão. Outras motivações baseiam-se em diversas características da Wikipédia, como por exemplo:

- É um recurso de informação de grandes dimensões. Em Outubro de 2008 continha mais de 7 milhões de artigos em cerca de 200 línguas e aproxidamente 428 mil entradas na versão portuguesa.
- O seu conteúdo tem uma licença livre, estando sempre disponível para a investigação sem restrições e sem a necessidade da aquisição de direitos.
- É um recurso de domínio geral, podendo, desta forma, ser usado na tarefa de extracção de informação de sistemas de domínio aberto.

- Os dados apresentados têm algum grau de formalidade e de estruturação (por exemplo, categorias) o que ajuda no seu processamento.
- É actualizada e revista continuamente através da colaboração de diversas pessoas.

A Wikipédia disponibiliza todo o conteúdo para cada uma das diferentes línguas, em formato XML, bem como as ferramentas necessárias para a sua conversão para SQL. O REMMA utiliza a informação disponibilizada em formato SQL e fez uso da estrutura interna desta base de dados. As secções seguintes descrevem a estrutura básica da Wikipédia no seu contexto da sua utilização no REMMA. O esquema completo da base de dados pode ser consultado em http://www.mediawiki.org/wiki/Manual:Database_layout.

12.1.1 Estrutura básica

Uma página da Wikipédia é identificada por um nome único, que pode ser obtido através da concatenação das palavras existentes no título com "_", mantendo a primeira letra da primeira palavra maiúscula. Seguindo o exemplo anterior, o nome único para a página *Universidade de Aveiro* é `Universidade_de_Aveiro`.

Usualmente, o título da página é o nome mais comum para a entidade descrita neste. Quando o nome é ambíguo, o título é também qualificado com uma expressão em parênteses, como no caso da página referente à flor *Cravo*, que é descrita, na Wikipédia, na página intitulada `Cravo_(flor)`.

De uma forma geral, existe um relacionamento de correspondência de muitos-para-muitos entre os nomes e as entidades. Este relacionamento é definido na Wikipédia através das *páginas de redirecção* e das *páginas de desambiguação*. Estes dois conceitos são explorados em mais detalhe nas secções 12.1.2 e 12.1.3.

No entanto, as páginas da Wikipédia têm outras estruturas úteis para a extracção de conhecimento, tais como as categorias e as ligações internas. Estes dois conceitos são descritos nas secções 12.1.4 e 12.1.5.

12.1.2 Redirecção

Existe uma *página de redirecção* para cada nome alternativo que possa ser usado para referir uma entidade na Wikipédia. As redirecções são marcadas como `#REDIRECT [[A B C]]` nos ficheiros fonte, onde "`[[...]]`" é a sintaxe que indica uma ligação a outro artigo na Wikipédia. As redirecções são usadas por várias razões relacionadas com a ambiguidade. Por exemplo, são usadas para expansão de abreviaturas tal como de `UA` para `Universidade de Aveiro`. Também são usadas no contexto de desambiguações mais difíceis, como as descritas na secção seguinte.

12.1.3 Páginas de desambiguação

Alguns autores criam uma *página de desambiguação* para um nome de entidade ambíguo[1]. Estas páginas são usadas para nomes que podem ter vários significados e possuem referências a outras páginas que dizem respeito a diferentes entidades que partilham o mesmo nome, enumerando todos os artigos possíveis para esse nome. Por exemplo, a página de

[1] *Ambíguo* refere-se ao caso em que o nome pode ser usado para referir várias entidades (i.e., artigos da Wikipédia)

desambiguação para o nome *Madeira* lista doze entidades associadas, isto é, para além dos nomes não ambíguos originados pelas páginas de redirecção, pode encontrar-se nestas páginas outros homónimos de uma dada entidade. A tabela 12.1 apresenta alguns exemplos das páginas que são apresentadas através da desambiguação do nome *Madeira*.

12.1.4 Categorias

Toda a página da Wikipédia deve ter pelo menos uma categoria. A categoria é uma página especial gerada automaticamente a partir das ligações que a esta vão dar. Regra geral, e para fins de organização, toda e qualquer página da Wikipédia deve ser categorizada por quem a criou de modo a garantir a geração automática da página da categoria e uma correcta catalogação das páginas da Wikipédia.

A tabela 12.1 apresenta alguns exemplos que exploram a organização interna da Wikipédia. Por exemplo, a página sobre o arquipélago da Madeira, com o título *Madeira (arquipélago)*, está associado a um conjunto de categorias, entre as quais *Região Autónoma da Madeira* e *Regiões vitivinícolas*.

Tabela 12.1: Exemplos de títulos e categorias de artigos relativos à desambiguação da palavra *Madeira*.

Título	**Redirecção**	**Categorias**
Madeira (material)	Madeira	Madeira
Madeira (arquipélago)	Região Autónoma da Madeira	Região Autónoma da Madeira, NUTS III portuguesas, ... Regiões vitivinícolas
Madeira Beach	—	Cidades da Flórida
Jamila Madeira	—	Loulé, Políticos Portugal
Vinho Madeira	—	Vinhos de Portugal

12.1.5 Ligações internas

Os artigos da Wikipédia contêm frequentemente menções a entidades já definidas. Estas ligações devem ser feitas através da utilização de ligações internas. Dois exemplos de ligações internas estão representados no exemplo (12.2) retirado da página sobre a Região Autónoma da Madeira.

(12.2)
```
A Madeira, oficialmente designada por Região Autónoma da
Madeira, é um território [[Portugal|português]] dotado de
autonomia política e administrativa através do [[Estatuto
Político Administrativo da Região Autónoma da Madeira]],
previsto na [[Constituição da República Portuguesa]].
```

A expressão da segunda ligação (*Estatuto Político Administrativo da Região Autónoma da Madeira*) corresponde ao título do artigo a que se refere. A mesma expressão é usada na

versão apresentada ao utilizador. Se o autor quiser que seja apresentada uma expressão diferente (por exemplo, *português* em vez de *Portugal*) então a expressão alternativa é incluída numa ligação com outro nome (em inglês, *piped link*), após o titulo. O exemplo (12.3) ilustra a expressão apresentada para o exemplo anterior.

(12.3) `A Madeira, oficialmente designada por Região Autónoma da Madeira, é um território português dotado de autonomia política e administrativa através do Estatuto Político Administrativo da Região Autónoma da Madeira, previsto na Constituição da República Portuguesa.`

12.2 O sistema REMMA

Nesta secção é descrito em mais detalhe o sistema REMMA e a sua arquitectura. A secção 12.2.1 apresenta a plataforma base usada pelo REMMA. A secção 12.2.2 foca a arquitectura do sistema, descrevendo os métodos utilizados para a classificação das entidades.

12.2.1 A plataforma base - UIMA

Uma característica do sistema é a sua integração na plataforma UIMA. O UIMA, *Unstructured Information Management Architecture* (Ferrucci e Lally, 2004), é uma plataforma livre, escalável e extensível, para a criação, integração e desenvolvimento de sistemas de gestão de informação não estruturada. Embora seja uma arquitectura com um certo grau de complexidade, tem diversas vantagens, como por exemplo:

- Disponibiliza algumas ferramentas de pré-processamento, tais como leitores e finalizadores genéricos, atomizador, separador em frases e outros anotadores simples;
- Uniformiza a estrutura dos resultados;
- Foca a modelação em vez de na programação.

O UIMA usa uma Estrutura de Análise Comum (em inglês, *Common Analysis Structure*, CAS) que permite aos anotadores acesso de leitura ao objecto a ser processado (por exemplo, um documento) e acesso de leitura/escrita aos resultados da análise ou às anotações associadas às diferentes regiões dos objectos. Estas regiões podem corresponder a palavras, frases ou parágrafos no texto. A CAS é partilhada entre os diversos anotadores que processam a colecção de objectos, passando de um anotador para o seguinte no processo.

12.2.2 A arquitectura

A arquitectura do REMMA está apresentada na figura 12.1.

O REMMA começa por ler os documentos, um por um, e guardar os respectivos metadados. No caso da colecção do Segundo HAREM é guardada a identificação do documento em análise. Os textos são posteriormente divididos em frases e átomos com a ajuda das ferramentas de pré-processamento disponíveis no UIMA. O analisador TreeTagger (Schmid, 1995) foi usado na obtenção das categorias morfossintácticas. Este analisador morfossintáctico foi utilizado exclusivamente para eliminar algumas preposições e advérbios dos candidatos a EM, justificando-se desta forma o uso de um analisador estatístico.

Figura 12.1: Arquitectura do REMMA

As anotações geradas por estas ferramentas são armazenadas na CAS e usadas nos diversos anotadores que constituem o módulo de REM. A figura 12.2 apresenta a sequência de anotadores utilizados na identificação e classificação das entidades.

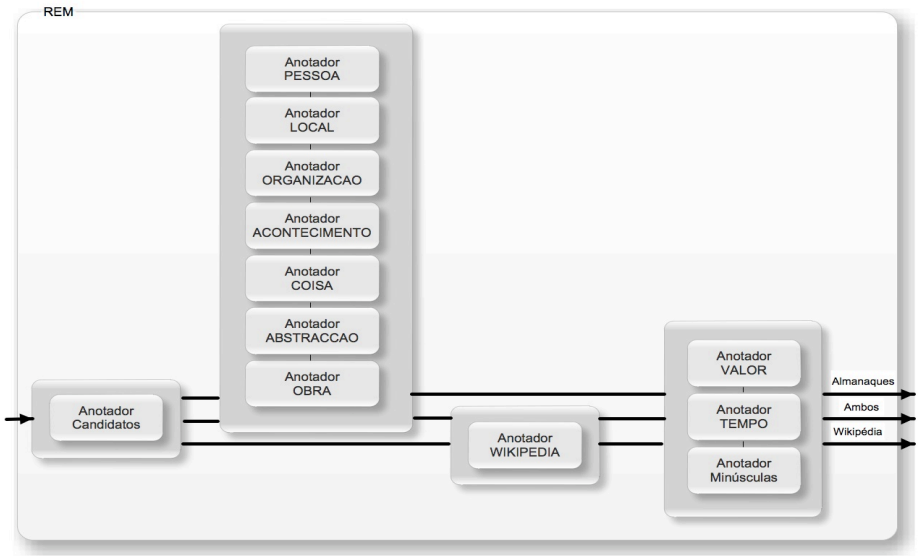


Figura 12.2: Anotadores do REMMA

O primeiro anotador a ser invocado é o `Anotador de Candidatos` que identifica todas as expressões candidatas a entidades mencionadas. As entidades candidatas são todos os conjuntos de termos iniciados por letra maiúscula. Na geração da expressão candidata foi também considerada a presença de termos de ligação com um comprimento inferior a 5 caracteres (por exemplo, *e*, *em*, *de*, *da*, *do*, *dos*, *das*, *para*, etc.). Não foram contempladas expressões contendo algarismos. Estas expressões candidatas foram posteriormente analisadas pelos anotadores de classificação.

O REMMA foi desenvolvido de modo a contemplar duas abordagens de classificação distintas. A primeira baseia-se em almanaques e regras muito simples e é descrita na secção seguinte. A classificação com base na informação extraída da Wikipédia pode ser realizada em conjunto ou separadamente do método anterior e é apresentada em detalhe na secção 12.2.2.2 . A utilização de duas abordagens distintas justifica-se pela necessidade de perceber quais as vantagens e desvantagens inerentes a cada método, em particular, de que forma a utilização da Wikipédia permite melhorar os resultados.

Na tarefa de classificação com base na informação extraída da Wikipédia foi utilizado um subconjunto de todo o conteúdo da Wikipédia, que é disponibilizado em XML para cada uma das diferentes línguas. Foi utilizada a Wikipédia portuguesa de Fevereiro de 2008, que inclui 1 290 836 páginas. Os dados foram posteriormente exportados para uma base de dados SQL, de modo a poderem ser usados neste sistema. Optou-se por não usar a informação existente nas páginas de desambiguação, mas apenas a redirecção que a comunidade Wikipédia entende ser a que mais utilizadores estão à procura. Relativamente às categorias associadas a cada página da Wikipédia, observou-se que uma página pode ter mais do que uma categoria, e que muitas vezes estas categorias não são claros hiperónimos da entidade a ser analisada. Assim, esta informação não foi usada, uma vez que a sua utilização implicaria a necessidade de seleccionar uma categoria apropriada nas categorias listadas, ficando esta tarefa para trabalho futuro.

Os anotadores relativos às categorias `TEMPO` e `VALOR` e o anotador `Minúsculas`, desenvolvido para a inclusão nas entidades da informação relativa às palavras começadas por minúsculas contempladas nas directivas do Segundo HAREM, são descritos separadamente na secção 12.2.2.3.

12.2.2.1 Classificação com base em regras e almanaques

Esta primeira abordagem baseou-se numa utilização combinada de um conjunto de regras de análise de contexto com a consulta de diversos almanaques de pequena dimensão. Os almanaques utilizados tinham sido já criados manualmente no âmbito de projectos desenvolvidos anteriormente na área da extracção de informação de relatórios médicos e contêm nomes de entidades de diversas classes semânticas, como por exemplo, listas de nomes de pessoas, listas de cidades portuguesas, listas de doenças e sintomas clínicos, etc.

As regras utilizadas foram criadas manualmente e baseiam-se no contexto em que a expressão é referida. Estas regras exploram certas classes semânticas de palavras, como por exemplo as relativas a cargos, tipos de locais, tipos de organização e outros. A tabela 12.2 lista alguns exemplos de palavras utilizadas para anotar as classes semânticas `PESSOA`, `LOCAL`, `ORGANIZACAO` e `ACONTECIMENTO`. Note-se que a tabela apenas apresenta informação relativa à anotação da *categoria* da entidade. Estas listas foram posteriormente subdivididas de modo a fornecerem informação relativa à anotação *tipo* da entidade mencionada, caso esta exista.

Os anotadores que usam a informação contida nestes almanaques e regras começam por dividir a expressão candidata nos seus vários termos e atribuem uma categoria semântica caso algum dos termos da expressão exista nas listas usadas. Quando esta anotação não é conseguida, procuram na expressão candidata palavras pertencentes à classe semântica em análise. Dois exemplos ilustrativos do tipo de cobertura deste módulo são apresentados em (12.4), onde *Paulo* é um nome existente no almanaque relativo a nomes

Tabela 12.2: Exemplos e quantidade de palavras usadas para a definição de regras contextuais das entidades PESSOA, LOCAL, ORGANIZACAO e ACONTECIMENTO.

PESSOA (N=110)	**LOCAL** (N=58)	**ORGANIZACAO** (N= 50)	**ACONTECIMENTO** (N=28)
Ministro	Praça	Museu	Campeonato
Chefe	Avenida	Faculdade	Concerto
Princesa	Rua	Sindicato	Congresso
Reitora	Cidade	Prefeitura	Exposição
...	...	...	...

de pessoas, e (12.5), onde *Provedor* é uma das palavras usadas nas regras contextuais da entidade PESSOA.

(12.4) **Entrada:** Ao que parece, Paulo Pinto Mascarenhas tem a convicção firme
Saída: Ao que parece, <EM ID="xxx" CATEG="PESSOA" TIPO="INDIVIDUAL">**Paulo Pinto Mascarenhas**</EM> tem a convicção firme

(12.5) **Entrada:** do Provedor do Espectador, não veio qualquer espécie de pressão.
Saída: do <EM ID="xxx" CATEG="PESSOA" TIPO="CARGO">**Provedor do Espectador**</EM>, não veio qualquer espécie de pressão.

12.2.2.2 Classificação com recurso à Wikipédia

A classificação com base na informação extraída da Wikipédia pode ser realizada em conjunto com a descrita na secção anterior, analisando neste caso apenas as entidades candidatas não anotadas anteriormente, ou individualmente, procurando uma classificação para todas as entidades candidatas identificadas.

Aquilo que se pretende é que este anotador seja capaz de encontrar uma entidade na Wikipédia correspondente à identificada nos textos em análise. Por exemplo, na frase (12.6), a expressão candidata a EM *Universidade de Harvard* é identificada pelo Anotador de Candidatos (ver secção 12.2.2). O objectivo passa por perceber de que forma é que esta entidade é descrita na Wikipédia e, consequentemente extrair a respectiva classificação do artigo.

(12.6) Concluiu os seus estudos de medicina, em 1870, na Universidade de Harvard, onde iniciou a sua carreira como professor de fisiologia em 1872.

Deste modo, cada uma das entidades candidatas identificadas é convertida num identificador da Wikipédia através da concatenação dos vários termos da expressão com o caracter "_". Por exemplo, a expressão *Universidade de Harvard* é convertida em Universidade_de_Harvard e o artigo correspondente recuperado, seguindo a redirecção, caso esta exista, até obter uma página de não-redireccionamento[2].

[2] Existem na Wikipédia algumas páginas para outros conteúdos que não os usuais artigos. Estes são distinguidos por um atributo *namespace*. Para a recuperação dos artigos que precisamos foi apenas analisado o *namespace* 0, que é o mais comum para estes artigos.

Embora não exista uma regra de formatação estrita, é normal que os artigos da Wikipédia comecem com uma pequena frase que define a entidade descrita no artigo. Por exemplo, o artigo com o título `Universidade_de_Harvard` começa com a frase (12.7).

(12.7) A Universidade Harvard (em inglês Harvard University) é uma das instituições educacionais mais prestigiadas do mundo, bem como a instituição de ensino superior mais antiga dos Estados Unidos da América.

Tal como neste exemplo, a primeira frase da maioria dos artigos contém uma expressão que indica a categoria semântica da entidade em análise. Neste caso, é a palavra *instituição*.

O método seguido concentra-se assim na extracção de tais nomes, a partir da primeira frase do artigo. Para tal foi necessário começar por remover etiquetas desnecessárias, tais como itálicos, negritos e ligações internas. As ligações internas foram convertidas para a expressão adequada (por exemplo, `[[língua inglesa | inglês]]` para `inglês`, ver secção 12.1.5). O artigo foi posteriormente dividido em frases de acordo com os padrões `\n`, `<br>` e regras simples de segmentação para o ponto final (.).

Após obtenção da primeira frase foram aplicadas regras simples, semelhantes às utilizadas no método anterior, ou seja, procuram na primeira frase do artigo da Wikipédia palavras-chave indicativas da classe semântica do artigo. Alguns exemplos, bem como a quantidade de palavras utilizadas por este anotador, são listados na tabela 12.3.

Tabela 12.3: Exemplos e quantidade de palavras-chave usadas na extracção de uma categoria semântica da primeira frase de um artigo.

PESSOA (N=15)	**LOCAL** (N=15)	**ORGANIZACAO** (N=12)	**ACONTECIMENTO** (N=2)
imperador	planeta	partido	acordo
engenheiro	cidade	movimento	competição
professor	ilha	universidade	
piloto	continente	...	
...	...	...	

Um exemplo de aplicação deste anotador, retirado da colecção usada no Segundo HAREM, é ilustrado em (12.8).

(12.8) **Entrada:** A popularidade do piloto Ayrton Senna na França era comparável à de seu maior rival, Alain Prost, quatro vezes campeão mundial.
Saída: A popularidade do piloto <EM ID="xxx" CATEG="PESSOA" TIPO="INDIVIDUAL">**Ayrton Senna**</EM> na França era comparável à de seu maior rival, <EM ID="yyy" CATEG="PESSOA" TIPO="INDIVIDUAL">**Alain Prost**</EM>, quatro vezes campeão mundial.

De notar que os nomes das entidades anotadas na frase, *Ayrton Senna* e *Alain Prost*, não existiam em nenhum dos almanaques utilizados no método anterior e também que, no caso da entidade *Alain Prost*, não existia qualquer regra contextual que a reconhecesse.

12.2.2.3 Anotadores VALOR, TEMPO e Minúsculas

Os anotadores desenvolvidos para as categorias semânticas TEMPO, TEMPO e Minúsculas são apresentados separadamente pois são independentes dos descritos anteriormente e utilizados na produção de todas as corridas do REMMA.

Os anotadores VALOR e TEMPO começam por identificar conjuntos de termos contendo pelo menos um algarismo ou que pertençam a uma lista de palavras pré-definida. No caso do anotador tempo as listas contêm, por exemplo, nomes de épocas festivas e estações do ano (*Páscoa*, *Carnaval*, *Primavera*, *Verão*, ...), dias da semana, meses, advérbios de frequência (*diariamente*, *todos os anos*, ...), etc. Para o anotador valor foram utilizadas listas contendo nomes de várias unidades (*metro*, *Kg*, *Gb*, etc.) e de nomes de moedas (*euros*, *dólares*, *contos*, etc.).

Estes anotadores incluem expressões regulares para identificar expressões como *em 25 de Abril [de 1974]*.

O anotador de minúsculas expande a anotação efectuada a uma dada entidade caso esta seja precedida por uma palavra começada por minúscula, que esteja incluída nas respectivas directivas. Caso a entidade não tenha ainda sido anotada, a palavra em minúscula precedente é analisada, de modo a inferir qual a anotação que deverá ser adicionada.

Um exemplo de aplicação, para cada um dos anotadores TEMPO e Minúsculas, encontra-se ilustrado em (12.9) e (12.10), respectivamente. Em (12.10), *(ex-)presidente* pertence à lista de palavras em minúsculas definida pelas directivas (a lista completa encontra-se no apêndice A, secção A.6).

(12.9) **Entrada:** As fortes chuvas que atingiram ontem
Saída: As fortes chuvas que atingiram <EM ID="xxx" CATEG="TEMPO" TIPO="TEMPO_CALEND">**ontem**</EM>

(12.10) **Entrada:**quando o ex-presidente José Sarney disse que sua maior missão era conduzir o país até as eleições.
Saída:quando o <EM ID="xxx" CATEG="PESSOA" TIPO="CARGO">**ex-presidente José Sarney**</EM> disse que sua maior missão era conduzir o país até as eleições.

Após a anotação das entidades identificadas pelos vários métodos descritos, um último anotador é chamado, o Finalizador. Este anotador analisa a CAS e cria o(s) documento(s) de saída. É este anotador que produz o documento XML final, através da análise das anotações associadas às diferentes regiões do(s) documento(s). É também neste passo que é efectuado o processamento de anotações alternativas <ALT>. O Finalizador determina a existência de duas ou mais anotações referentes a regiões encerradas noutra(s). Neste caso, as entidades são etiquetadas com duas ou mais anotações distintas separadas por "|". Um exemplo da saída gerada por este anotador é apresentada em (12.11).

(12.11) <ALT>
<EM ID="xxx" CATEG="PESSOA" TIPO="CARGO">**líder do Sinn Fein**</EM>
| líder do <EM ID="yyy" CATEG="ORGANIZACAO" TIPO="INSTITUICAO">**Sinn Fein**</EM>
</ALT>

12.3 Resultados no Segundo HAREM

Tal como referido, a participação do REMMA no Segundo HAREM pretendia avaliar de que forma a extracção de conhecimento a partir da Wikipédia para o REM permite criar aplicações úteis e substituir a utilização, e consequentemente a criação e manutenção, de listas e almanaques de grande dimensão. Deste modo, e tendo em consideração a arquitectura do sistema REMMA, foram geradas três corridas distintas. O nome de cada corrida é relativo à fonte de conhecimento principal usada para a obter (dentro de parêntesis encontra-se o nome que lhes foi atribuído pela organização).

- **Corrida Almanaques (REMMA_2_corr):** Corrida criada com a utilização dos anotadores de regras contextuais e almanaques. Em particular foram utilizados anotadores para as classes semânticas `LOCAL`, `PESSOA`, `ORGANIZACAO`, `ACONTECIMENTO`, `OBRA`, `ABSTRACCAO` e `COISA`. A estes anotadores seguiram-se os anotadores `VALOR`, `TEMPO` e `Minúsculas`.
- **Corrida Wiki (REMMA_3_corr):** Corrida gerada pela utilização isolada do anotador `Wikipédia`, seguido dos anotadores `VALOR`, `TEMPO` e `Minúsculas`.
- **Corrida Ambos (REMMA_1_corr):** Corrida gerada pela utilização sequencial de todos os anotadores desenvolvidos.

Relembramos da secção 12.2.2 que a figura 12.2 apresenta a sequência de anotadores utilizados na produção das diversas corridas.

As secções seguintes apresentam os resultados obtidos no Segundo HAREM na tarefa de classificação.

12.3.1 Usar a Wikipédia tem potencial para melhor desempenho?

Comparando os resultados obtidos pelas diferentes corridas, apresentados na tabela 12.4, e relativamente à medida F, podemos observar que a melhor corrida é a *Ambos*. Esta observação é verdadeira em ambas as avaliações de ALT, estrita e relaxada. No entanto, é de notar, que este resultado é obtido à custa de uma maior abrangência em relação às restantes corridas (mais $\sim$0,13 em relação à corrida *Wiki* e mais $\sim$0,07 em relação à corrida *Almanaques*, em ambas as avaliações (estrita e relaxada)) e de uma ligeira perda de precisão em relação à corrida *Almanaques* (menos $\sim$0,01).

A corrida *Wiki* é a que obtém piores resultados em todas as métricas e em ambas as avaliações. Observa-se uma diminuição dos valores da medida F em cerca de 0,12 (avaliações estrita e relaxada) em relação à corrida *Ambos*.

A utilização de almanaques gera resultados superiores aos obtidos com a utilização isolada da Wikipédia, sendo, no entanto, ainda inferiores aos obtidos com a utilização de todos os anotadores. Relativamente à corrida *Ambos*, observa-se ainda um decréscimo de aproximadamente 0,05 na medida F (ambas as avaliações).

12.3.2 Para a Wikipédia todas as categorias nascem iguais?

A tabela 12.5 apresenta os resultados obtidos para cada uma das categorias, na tarefa de classificação. Sobressai da análise da tabela o facto de as categorias `LOCAL`, `PESSOA`,

Tabela 12.4: Resultados do REMMA no HAREM clássico para a tarefa de classificação.

Versão	Avaliação estrita de ALT			Avaliação relaxada de ALT		
	Precisão	**Abrangência**	**Medida F**	**Precisão**	**Abrangência**	**Medida F**
Almanaques	**0,6132**	0,2952	0,3985	**0,6340**	0,3084	0,4150
Wiki	0,5808	0,2316	0,3312	0,5950	0,2409	0,3429
Ambos	0,6050	**0,3615**	**0,4526**	0,6226	**0,3750**	**0,4681**

ORGANIZACAO, ACONTECIMENTO e OBRA obterem claramente melhores resultados pela utilização da Wikipédia. As categorias ABSTRACCAO e COISA parecem ser imunes à utilização da Wikipédia.

Observa-se também que os resultados obtidos para as categorias VALOR e TEMPO são independentes da utilização da Wikipédia quando usada em conjunto com almanaques, tendo, no entanto, curiosamente, sofrido um ligeiro decréscimo na medida F (menos de 0,01) com a utilização isolada da Wikipédia.

Tabela 12.5: Resultados do REMMA para cada uma das categorias na tarefa de classificação.

Categoria	Melhor Versão	Classificação		
		Precisão	**Abrangência**	**Medida F**
LOCAL	Ambos	0,5700	0,5089	**0,5377**
PESSOA	Ambos	**0,6666**	0,3677	0,4740
VALOR	Ambos = Almanaques	0,3589	**0,5202**	0,4247
ORGANIZACAO	Ambos	0,5829	0,2397	0,3397
TEMPO	Ambos = Almanaques	0,4744	0,2538	0,3307
ACONTECIMENTO	Ambos	0,4044	0,1473	0,3159
OBRA	Ambos	0,5146	0,1212	0,1962
ABSTRACCAO	Ambos = Almanaques	0,2231	0,0392	0,0667
COISA	Ambos = Almanaques	0,2227	0,0318	0,0557

Na figura 12.3 observa-se a distribuição dos valores da precisão e abrangência discriminados por categoria e corrida, onde se nota uma maior precisão do REMMA relativamente à abrangência. Mais uma vez se observam os piores resultados obtidos nas categorias semânticas ABSTRACCAO e COISA.

12.3.3 Esta abordagem é competitiva?

Nesta secção tenta-se perceber até que ponto o sistema REMMA é competitivo, analisando os resultados de uma forma comparativa com os obtidos pelos melhores e piores sistemas para cada uma das categorias.

A figura 12.4 apresenta os valores obtidos em cada uma das métricas, discriminados em termos de categoria, e após uma normalização relativa aos valores máximos e mínimos obtidos, isto é, relativa aos resultados do melhor e do pior sistema em cada uma das categorias. Por exemplo, é possível perceber que o REMMA obteve a melhor precisão para

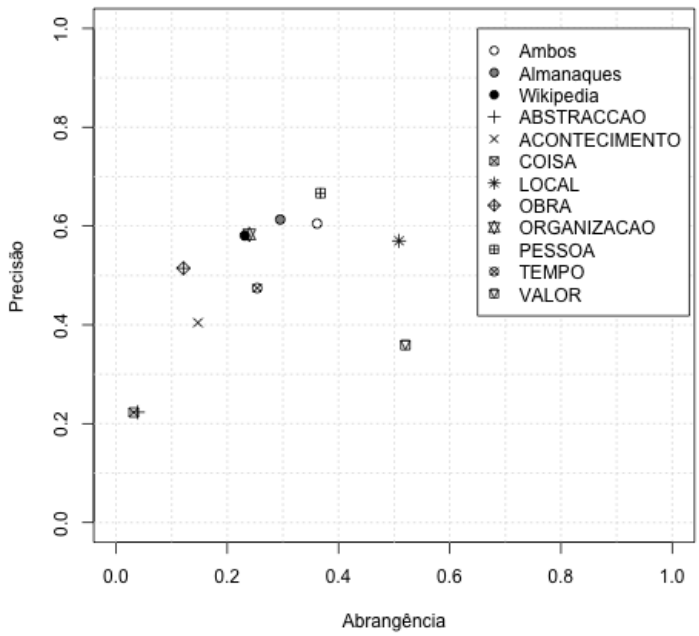


Figura 12.3: Distribuição de precisão e abrangência na tarefa de classificação das diversas categorias e corridas.

a categoria ABSTRACCAO (valor 1 no gráfico 12.4), tendo no entanto a sua abrangência e medida F sido a menor obtida por todos os sistemas participantes (igual a 0 no gráfico).

O gráfico permite observar que o REMMA obteve, de uma forma geral, resultados bastante precisos, estando sempre próximo do melhor sistema para cada uma das categorias. O contrário pode ser observado relativamente à abrangência.

A melhor classificação do REMMA, no cenário total, é relativa à corrida *Ambos*, tendo ficado na posição 10 em 29.

12.3.4 Comparação com o REMBRANDT

Após a realização do encontro do Segundo HAREM (Setembro de 2008) apercebemo-nos da existência de um sistema participante com uma abordagem bastante semelhante à do REMMA, o REMBRANDT (ver capítulo 11). Este sistema usa a informação relativa às categorias existentes na Wikipédia para a obtenção da classificação adequada a cada categoria do Segundo HAREM.

Relativamente a este sistema, comparando os resultados obtidos no cenário total e disponibilizados no capítulo 11, o REMMA obteve valores para a medida F inferiores aos ob-

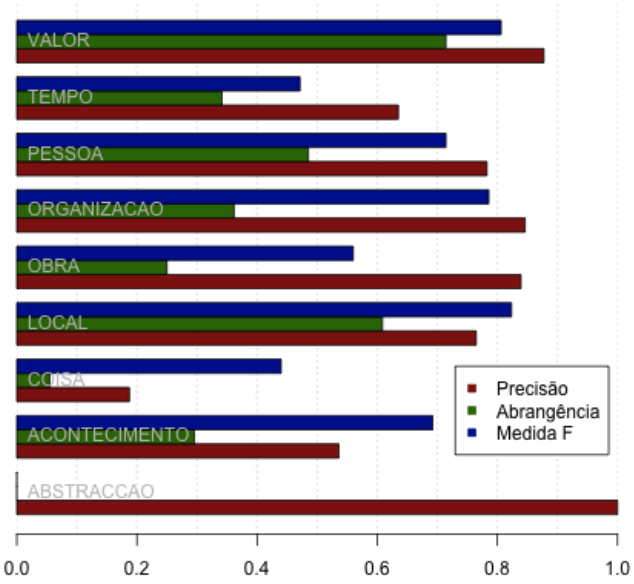


Figura 12.4: Comparação normalizada relativamente aos valores máximos e mínimos

tidos pelo REMBRANDT (cerca de 0,11 inferior, nas melhores corridas de ambos), tendo os sistemas, no entanto, precisões comparáveis.

No que diz respeito à análise individual de cada uma das categorias, observam-se melhores resultados do REMMA na classificação das entidades pertencentes à categoria ACONTECIMENTO (aproximadamente 0,02) e valores muito semelhantes para a categoria COISA, tendo para as restantes sido obtida pelo REMMA uma medida F inferior à do REMBRANDT.

12.4 Discussão

Os resultados obtidos pelo REMMA sugerem a utilidade da extracção de categorias semânticas da Wikipédia para a tarefa de REM, mesmo quando efectuada através de um método tão simples como o apresentado.

No entanto, estes indicam também que a utilização isolada da informação contida na Wikipédia, sem recurso a qualquer almanaque ou regra contextual, é uma solução com piores resultados, isto é, sugerem a existência de espaço para várias melhorias, como por exemplo, a necessidade de uma utilização mais abrangente das várias estruturas internas da Wikipédia.

Os resultados relativos às diferentes categorias avaliadas indicam uma imunidade das

categorias ABSTRACCAO e COISA ao uso da Wikipédia. Este resultado pode estar relacionado com uma maior ambiguidade, e consequente dificuldade, na categorização destas entidades. Outro factor que pode estar na origem destes resultados é o uso de um conjunto insuficiente de palavras na extração da classificação a partir da primeira frase do artigo. As categorias VALOR e TEMPO também não apresentam melhorias pelo uso da Wikipédia como fonte de conhecimento. Isto pode dever-se à introdução de categorias erradas aquando da utilização do anotador Wikipédia isoladamente, como por exemplo na análise de palavras relativas a estações do ano, meses, dias da semana ou unidades.

De uma forma geral, podemos afirmar que o REMMA é um sistema bastante preciso, tendo apresentado precisões competitivas relativamente aos demais sistemas. A comparação com o REMBRANDT em particular, indica precisões comparáveis, faltando ao REMMA abrangência. Isto deve-se, entre outros factores, à necessidade de criar mais regras e regras mais complexas para a identificação de EM candidatas, bem como métodos de resolução de conflitos e ambiguidades. Note-se, no entanto, que, no contexto da extracção de informação na área da medicina, importa a existência de um sistema preciso, capaz de anotar correctamente a informação existente, em oposição a um sistema que extraia muita informação com ruído.

Os resultados obtidos mostram, assim, existir espaço para várias melhorias. A tarefa de identificação do REMMA é realizada actualmente através de expressões regulares bastante simples, não contemplando expressões com algarismos. Uma melhoria nesta tarefa implicaria com certeza melhor abrangência. O facto de o anotador Wikipédia recolher informação de uma página apenas e só caso esta não seja um página de desambiguação pode evitar a introdução de ruído nos resultados, no entanto, a criação de métodos de resolução de conflitos entre entidades ou de desambiguação, bem como a utilização de outras estruturas internas disponíveis na Wikipédia, como é o caso das categorias, implicaria certamente melhorias significativas no REMMA.

De uma forma geral, a participação do REMMA no HAREM, embora direccionada a textos não especializados, permitiu perceber a utilidade de diferentes abordagens, acabando por indicar que a utilização de recursos e soluções semelhantes para a área em que nos concentramos, a medicina, é uma abordagem promissora, mesmo com a utilização de abordagens simples como a apresentada.

12.5 Conclusão e trabalho futuro

Para a participação no Segundo HAREM foi desenvolvido um sistema capaz de explorar fontes de conhecimento externas, como a Wikipédia, de modo a evitar a criação e a manutenção de almanaques de domínio geral de grande dimensão. A principal motivação para esta abordagem foi o carácter dispendioso desta tarefa, quer em termos de tempo, quer em termos dos recursos necessários. Foi desenvolvido um sistema composto por um conjunto de anotadores UIMA, capaz de usufruir de vários tipos de recursos, sejam estes almanaques simples de domínio geral, ou, categorias semânticas extraídas a partir da análise da primeira frase de um artigo da Wikipédia.

A utilização da Wikipédia demonstrou ser útil para a melhoria da classificação das entidades mencionadas, dando uma indicação clara da utilidade deste tipo de fontes de conhecimento e abrindo portas à procura e aplicação de soluções semelhantes a textos da área da medicina. Existem actualmente diversas wikis públicas e relativas a vários domí-

nios. O futuro do sistema REMMA passará, assim, pela utilização de recursos semelhantes relativos à área da medicina, de modo a melhorar a tarefa de extracção de informação que nos propomos realizar no âmbito do projecto MedAlert.

No entanto, ficou também claro neste trabalho a necessidade usar técnicas de desambiguação e de explorar outras estruturas internas disponibilizadas nas wikis públicas, como é o caso das categorias e das ligações internas na Wikipédia. Relativamente às páginas de desambiguação essa necessidade é mais evidente pelo facto de a Wikipédia ser uma enciclopédia em constante crescimento, o que implica um aumento constante do número de artigos e assuntos definidos e consequentemente, um aumento da ambiguidade das suas páginas. Uma interessante tarefa a realizar futuramente é o desenvolvimento de uma técnica de recuperação do título do artigo mais adequado ao contexto em questão, a partir de uma página de desambiguação.

A utilização do conteúdo da Wikipédia para a extracção de relações semânticas entre entidades é também uma interessante tarefa a realizar futuramente e uma área de grande interesse no âmbito do projecto MedAlert (Ferreira et al., 2008).

Agradecimentos

O projecto RTS foi financiado pelo programa "Aveiro Digital" da iniciativa "Portugal Digital" e pelo programa POSI do Governo Português. O projecto GERESmed é financiado pela Fundação para a Ciência e Tecnologia (GRID/GRI/81819/2006).

Capítulo 13

Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM

Marcirio Silveira Chaves

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 13, p. 231–245.

A maior quantidade de conhecimento existente atualmente está disponível em textos na rede. De acordo com Wilks (2008), 85% da informação disponível para ciência, empresas e aquela encontrada de modo informal na rede estão no formato não estruturado (texto, a maior parte). Contudo, de modo a tornar-se verdadeiramente útil para sistemas que fazem algum tipo de processamento inteligente, esse conhecimento precisa ser identificado e classificado corretamente.

Para tentar aproveitar melhor esse conhecimento, é necessário reconhecer inicialmente as entidades mencionadas presentes nos documentos. A tarefa de reconhecimento de entidades mencionadas (REM) em textos em português tem ganho mais atenção desde 2005, quando ocorreu a primeira avaliação desses sistemas no Primeiro HAREM (Santos e Cardoso, 2007a). Os resultados alcançados pelos sistemas participantes evidenciaram a necessidade de mais pesquisa na área, motivando dois eventos subsequentes: o Mini-HAREM (integrado no Primeiro HAREM) e o Segundo HAREM. Conforme descrito no capítulo 4, o Segundo HAREM também desafiou os sistemas a reconhecerem as relações existentes entre entidades mencionadas.

Nesse capítulo eu concentro a atenção no reconhecimento de entidades mencionadas da categoria local e de suas relações. Para isso, eu desenvolvi o SEI-Geo, um Sistema de Extração, Anotação e Integração de Conhecimento Geográfico baseado essencialmente no uso de padrões e de geo-ontologias. O SEI-Geo está inserido no contexto de minha tese de doutoramento e tem como um de seus objetivos expandir o conhecimento existente em bases de conhecimento geográfico com informação textual. Nesse capítulo é descrito apenas o módulo de extração e anotação de informação geográfica, o qual foi utilizado no processo de anotação de locais e relações em textos.

A participação do SEI-Geo no Segundo HAREM é motivada pela necessidade de se mensurar a qualidade da parte do sistema SEI-Geo que trata do reconhecimento de locais e suas relações em textos.

13.1 Trabalhos relacionados

O uso de padrões tem sido aplicado para extração de informação de textos em diversos trabalhos (Agichtein e Gravano, 2000; Etzioni et al., 2005; Cafarella et al., 2005; McDowell e Cafarella, 2008). Todos esses trabalhos processam textos em língua inglesa. O interesse sobre a informação geográfica presente em textos em português tem ganho atenção apenas nos últimos anos. Especificamente no tratamento de informação geográfica em texto em língua portuguesa, dois trabalhos têm sido reportados na literatura como mais relevantes (Vasconcelos Borges, 2006; Martins et al., 2007b). O primeiro eu descrevo brevemente a seguir e o segundo, o sistema CaGE (*Capturing Geographic Entities*), participou em todas as edições do HAREM, foi descrito em Martins et al. (2007b) e está presente nesse livro no capítulo 7.

Delboni (2005) e Vasconcelos Borges (2006) foram pioneiros ao processar textos na variante brasileira da língua portuguesa. Vasconcelos Borges propôs uma ontologia de lugar (OnLocus) que possui conceitos geográficos norteados pela divisão administrativa do Brasil. No trabalho dela, OnLocus foi mais explorada na parte de endereços postais e telefones.

Numa amostra de 75.413 páginas da rede brasileira, Vasconcelos Borges encontrou 57% delas com presença de endereços que foram detectados de acordo com

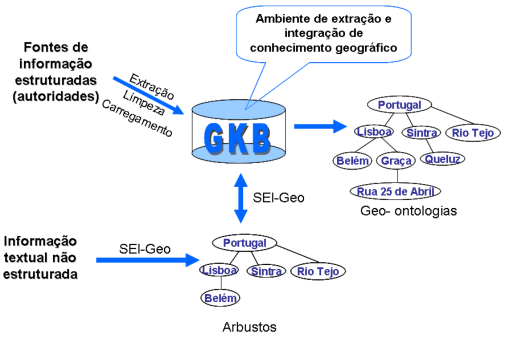


Figura 13.1: Arquitetura global do sistema de gestão de conhecimento geográfico.

um conjunto de padrões pré-definidos. Os seis principais tipos de padrões utilizados são: `Telefone`, `EndereçoBásico+CidadeEstado+CEP`, `EndereçoBásico+Telefone`, `EndereçoBásico+CidadeEstado`, `EndereçoBásico+CEP` e `CEP`. Esses padrões foram aplicados à coleção WBR05 (Modesto et al., 2005), extraindo 2.137.601 endereços de 603.798 páginas, o que representa 14,77% do total de páginas dessa coleção.

Delboni também apresenta um conjunto de expressões de posicionamento desenvolvidas para detectar nomes geográficos em textos em português. Essas expressões são baseadas em quatro tipos de relações espaciais: difusas[1] (por exemplo, *perto*, *depois* e *acima*), direcionais (por exemplo, *em frente*, *ao lado*, *atrás*), métricas (por exemplo, *quilômetros*, *minutos*, *quadras*) e topológicas (por exemplo, *dentro de*, *no coração de*, *na praça de alimentação*). Os experimentos realizados por Delboni indicam que as relações direcionais e, principalmente, as métricas são predominantemente utilizadas no contexto de uma expressão de posicionamento (informação geográfica), enquanto os demais tipos de relações são empregados em outros contextos.

13.2 O SEI-Geo

O SEI-Geo é um sistema que está integrado numa arquitetura global do sistema de gestão de conhecimento geográfico desenvolvido na minha tese de doutoramento, a GKB – *Geographic Knowledge Base* (Chaves et al., 2005b), e que está representada na figura 13.1.

A GKB é um ambiente de extração e integração de conhecimento geográfico que contém informações provenientes de fontes de dados administrativas semi-estruturadas de autoridades junto com um conjunto de regras para integração de informação. A expansão do conhecimento contido na GKB ocorre com informação proveniente de textos. Esses textos são a entrada de informação para o SEI-Geo, que é o responsável por gerar uma representação estruturada (em forma de arbustos) do conhecimento geográfico extraído e integrá-lo no repositório da GKB. Programas simples para geração de ontologias exportam o conhecimento armazenado nesse repositório.

Esta seção descreve o SEI-Geo utilizado no Segundo HAREM. O sistema é composto por dois módulos principais: extrator e anotador de informação geográfica e integrador de

[1] Em inglês, *fuzzy*.

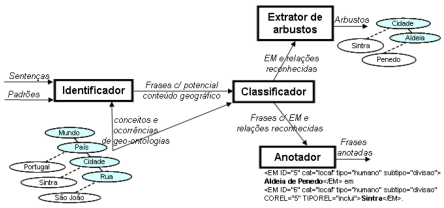


Figura 13.2: Arquitetura do módulo de extração e anotação de informação geográfica do SEI-Geo.

conhecimento geográfico. O primeiro tem como objetivo identificar, classificar, extrair arbustos e anotar o conhecimento geográfico disponível em textos representando-o de forma estruturada. A figura 13.2 apresenta a sua arquitetura. A seguir são descritas as funções de cada sub-módulo.

`Identificador`: recebe como entrada uma coleção de textos previamente segmentados em sentenças, mais um conjunto de padrões e de conceitos e ocorrências de geo-ontologias. As sentenças com potencial conteúdo geográfico são a entrada do módulo `Classificador`.

`Classificador`: recebe as sentenças e consulta as geo-ontologias para fazer a desambiguação e identificar relações semânticas.

`Extrator de arbustos`: recebe os locais reconhecidos e constrói os arbustos. Um arbusto é composto por pelo menos duas entidades mencionadas e uma relação. Não há número máximo de entidades mencionadas e relações pré-definido. Um exemplo de arbusto é `<Aldeia de Penedo,parte da,cidade de Sintra>`. Os arbustos são formalizados no formato de triplas *Resource Description Framework* (RDF), o qual está disponível em `http://www.w3.org/RDF`.

`Anotador`: recebe a sentença com locais e relações reconhecidos e faz a anotação no formato solicitado por qualquer aplicação. No caso do HAREM, anota a sentença de acordo com as diretivas da avaliação (Santos et al., 2008c).

Os padrões usados como entrada no algoritmo são descritos a seguir juntamente com o procedimento executado pelo algoritmo. Exceto os padrões do tipo Hearst, todos os demais são imediatamente sucedidos por preposição antes do nome de local. A lista completa das preposições utilizadas é como segue: *a*, *de*, *da*, *das*, *do*, *dos*, *entre*, *na*, *nas*, *no*, *nos*, *em*, *à*, *para*, *pra*, *ao*.

Conceitos geográficos: Conceitos de uma geo-ontologia existente mais conceitos complementaresàqueles inseridos no SEI-Geo, mas ausentes nas geo-ontologias. O algoritmo extrai todos os conceitos definidos na geo-ontologia que estão presentes na sentença. Sempre que o algoritmo encontra um conceito, ele verifica se esse conceito é sucedido por um nome de local. Em caso positivo, anota o nome como um local.

Padrões do tipo Hearst traduzidos para o português e estendidos: são aqueles definidos em Hearst (1992) acrescentados de algumas variantes adaptadas ao português (por exemplo, *é o distrito*, *é um concelho* e *é uma das cidades*). O algoritmo utiliza os

padrões do tipo Hearst do tipo `[Nome de local] é um (d[eao]s?)? [Conceito geográfico]` e `[Conceito geográfico] tal(is) como [Nome de local]`. Para cada padrão encontrado na frase, o algoritmo anota os locais presentes. Um `Nome de local` é todo o nome próprio que refere-se a um local geográfico.

Relações métricas, direcionais, difusas e de orientação: relações métricas descrevem proximidade usando unidades de medida, relações direcionais indicam posicionamento em relação a determinado local, construção ou objeto, entre outros, relações difusas descrevem proximidade através da utilização de termos qualitativos e imprecisos e relações de orientação, que foram inicialmente definidas como direcionais no trabalho de Güting (1994), são expressos através de pontos cardeais. A seguir é apresentada a lista completa dos termos utilizados desses padrões:

Métrico: *distante(s)*, *distância*, *km(s)*, *kilómetro(s)*, *quilómetro(s)*, *minuto(s)*, *minuto(s)*, *metro(s)*

Direcional: *ao lado*, *atrás*, *em frente*, e *defronte*

Difuso: *antes*, *depois*, *acima*, *abaixo*, *próxima*, *próximo*, *perto* e *proximidades*

Orientação: *norte*, *sul*, *leste*, *oeste*, *nordeste*, *noroeste*, *sudeste*, *sudoeste*

Todos esses padrões são descritos em (Delboni, 2005) como aqueles que geram melhores resultados quando uma pessoa deseja expressar posicionamento.

Substantivos: *água(s)*, *afogada(s)*, *afogado(s)*, *beira(s)*, *cabo(s)*, *capital(ais)*, *eleição(ões)*, *favela(s)*, *herdade(s)*, *guerra(s)*, *litoral(ais)*, *margem(ns)*, *natural(ais)*, *penitenciária(s)*, *periferia(s)*, *prefeito(s)*, *procedente(s)*, ex-*prefeito(s)* e *praia(s)*. Nomes próprios que ocorrem após esses substantivos são anotados como entidades candidatas (ou seja, topônimos) a locais. Alguns desses substantivos (por exemplo, *praia* e *cabo*) são candidatos a conceitos para expandir geo-ontologias.

Advérbios *aqui*, *cá*, *lá* e *longe*. Nomes próprios que ocorrem após esses advérbios são anotados como entidades candidatas a locais.

Verbos: *chegar*, *falecer*, *localizar*, *morar*, *morrer*, *mudar*, *nascer*, *ser*, *situar*, *sediar*, *realizar*, *viver*, *voltar*, *ir*, *vir*. Esse tipo de padrão inclui variações de tempo, gênero e número. Nomes próprios que ocorrem após esses verbos são anotados como entidades candidatas a locais.

Nomes de Entidades: Ocorrências das geo-ontologias.

13.2.1 Geo-ontologias utilizadas pelo SEI-Geo

Ontologias podem desempenhar um papel importante na tarefa de REM, especificamente para locais, assim como dicionários e almanaques. Para Malouf (2002) sua utilização não auxilia a melhoria dos resultados, enquanto Mikheev et al. (1999) encontraram bons resultados utilizando almanaques. Carreras et al. (2003) apresentaram resultados melhores com o uso de almanaques. Mikheev et al. também comprovaram que a utilização de almanaques é necessária para identificar nomes de locais. Em português, Martins et al. (2007b) obtiveram resultados satisfatórios com o uso de geo-ontologias para reconhecer locais.

Tabela 13.1: Estatística sobre as geo-ontologias utilizadas pelo SEI-Geo.

Estatística	Geo-Net-PT 10	WGO
# de entidades	4.651	13.124
# de nomes distintos	3.749	10.442
# de relações	6.304	24.712
# de relações parte-de	4.956	13.341
# de relações de adjacência	1.348	11.371

As geo-ontologias na abordagem do SEI-Geo fornecem listas de nomes e conceitos. Uma das vantagens das geo-ontologias é que elas permitem que se explore as relações existentes entre locais reconhecidos em textos com base nas relações nelas definidas.

A tabela 13.1 apresenta as estatísticas das geo-ontologias utilizadas pelo SEI-Geo. Os valores da Geo-Net-PT 10 incluem os dez principais conceitos da geo-ontologia, até o nível de freguesia. A geo-ontologia completa de Portugal (Geo-Net-PT) contém mais de 400.000 entidades, é um recurso público que foi desenvolvido no Pólo XLDB da Linguateca em colaboração com o projeto GREASE e está disponível em http://xldb.fc.ul.pt/geonetpt.

Além da Geo-Net-PT, o SEI-Geo utiliza a *World Geographic Ontology* (WGO) (Chaves et al., 2005a; Martins et al., 2007b). Essa geo-ontologia contém nomes, conceitos e relações sobre as principais divisões administrativas do mundo desde países e territórios até cidades com mais de 100.000 habitantes, além de entidades geográficas no domínio físico, tais como oceanos, mares e montanhas. Ambas as geo-ontologias foram utilizadas para suportar a participação do sistema de recuperação de informação geográfica (RIG) da Universidade de Lisboa nas quatro edições do GeoCLEF[2], de 2005 a 2008 (Cardoso et al., 2006; Martins et al., 2007a; Cardoso et al., 2008a,c). Esse sistema de RIG obteve o primeiro lugar na avaliação em 2006 nas tarefas monolíngue inglês e português.

As fontes de informação da Geo-Net-PT são provenientes de autoridades administrativas de Portugal (por exemplo, Instituto Nacional de Estatística (INE), Correios, Telégrafos e Telefones (CTT) e Associação Nacional de Municípios Portugueses (ANMP)), enquanto a WGO é formada na sua maioria por dados do *World Gazetteer*, da Wikipedia e do Instituto Geográfico Português.

13.2.2 Algoritmos de identificação e classificação de locais

O algoritmo 13.1 formaliza a fase de identificação de locais no módulo extrator e anotador de informação geográfica do SEI-Geo. Antes de descrever os algoritmos é necessário definir os seguintes termos:

Entidade candidata (EC): é um topônimo, um nome próprio (composto por pelo menos uma palavra). Exemplos de entidades candidatas incluem *Grécia*, *Brasília* e *concelho de Braga*.

Entidade Geográfica (EG): é um objeto com significado no domínio do discurso (correspondente à *feature* na ISO 19109 (ISO19109, 2006)). No domínio geográfico, a *província do Algarve*, o *concelho de Évora* e a *freguesia de Santa Isabel* são exemplos de tais

[2] O GeoCLEF foi um fórum internacional de avaliação de sistemas de RIG. Mais detalhes em http://www.uni-hildesheim.de/geoclef.

entidades numa ontologia. Essas entidades geográficas devem ter uma referência numa ontologia, e essa referência fornece o significado no domínio geográfico. Formalmente, uma entidade geográfica é uma `EC` que refere apenas uma referência na ontologia (por exemplo, `<[<concelho,Évora>],[GEO_346]>)`.

Além desses termos, a sintaxe de `w[+1]` significa a palavra sucessora daquela que está sendo comparada no ciclo (*for*). Por exemplo, no seguinte excerto de uma sentença *... perto de Aveiro ...*, se *perto* é o padrão sendo comparado, `w[+1]` é igual a *de*.

Algoritmo 13.1: Algoritmo para identificação de locais implementado no SEI-Geo.

```
1:  WGO_adm = {ocorrências do domínio administrativo da WGO}
2:  WGO_fis = {ocorrências do domínio físico da WGO}
3:  WGO = WGO_adm ⋃ WGO_fis
4:  GN = {ocorrências da Geo-Net-PT}
5:  P = {Adjetivo ∪ Adverbio ∪ Conceito geográfico ∪ Fuzzy ∪ Hearst ∪ Metrico ∪ Orientacao ∪ Substantivo
    ∪ Verbo}
6:  S = {frases do texto}
7:  Prep = {a,de,da,das,do,dos,entre,na,nas,no,nos,em,à,para,pra,ao}
8:  for all s ∈ S do
9:      for all w ∈ s do
10:         if w ∈ P then
11:             EC = identificaEC(w[+1],s)
12:             if EC != null then
13:                 Algoritmo 13.2 (EC)
14:             end if
15:         else if w ∈ {WGO ⋃ GN} then
16:             EC = identificaEC(w,s)
17:             if EC != null then
18:                 Algoritmo 13.2 (EC)
19:             end if
20:         end if
21:     end for
22: end for
23:
24: sub identificaEC(w,s) {
25: for all w ∈ s do
26:     if w ∈ {Prep ∪ ^ ([0-9]|[A-Z]) ∪ (length(w) ≥ 2)} then
27:         EC += w
28:     end if
29:     if EC[0] ∈ Prep then
30:         EC = EC[1,length(EC)]
31:     end if
32:     if EC[-1] ∈ Prep then
33:         EC = EC[0,length(EC)-1]
34:     end if
35: end for
36: return EC
37: }
```

A fase de identificação de locais recebe como entrada ocorrências de geo-ontologias, padrões que incluem termos frequentemente utilizados ao redor de nomes de locais em textos e preposições que ocorrem em nomes de locais. Toda vez que um padrão é encontrado numa frase, o algoritmo invoca a função `identificaEC` que identifica e retorna uma EC ou `null`, caso não seja um nome candidato a local. Essa função encontra os delimitadores da EC, ou seja, o início e o fim da mesma através de preposições e termos cuja primeira letra é maiúscula e seu comprimento é maior ou igual a dois. Após encontrar uma EC o algoritmo invoca a função de classificação de locais, descrita no algoritmo 13.2. Caso a palavra que está sendo comparada com os padrões não seja um padrão e sim um nome que está presente em geo-ontologias, o algoritmo verifica se a próxima palavra da sentença faz parte do nome. Se fizer, invoca a função `identificaEC`. Senão, assume a palavra como nome de local e invoca a função de classificação de locais, descrita no algoritmo 13.2.

Algoritmo 13.2: Algoritmo para classificação de locais implementado no SEI-Geo.

1: EC = {nome extraído do texto = entidade candidata}
2: WGO_{adm} = {ocorrências do domínio administrativo da WGO}
3: WGO_{fis} = {ocorrências do domínio físico da WGO}
4: GN = {ocorrências da Geo-Net-PT}
5: **if** EC $\in$ WGO_{adm} **then**
6: EG = {id do pai mais acima na hierarquia da WGO_{adm}}
7: **else if** EC $\in$ GN **then**
8: EG = {id do pai mais acima na hierarquia da GN}
9: **else if** EC $\in$ WGO_{fis} **then**
10: EG = {id do pai mais acima na hierarquia da WGO_{fis}}
11: **else**
12: EG = {id tipo Humano}
13: **end if**
14: AnotaEG(EG);

A partir de uma EC o algoritmo consulta a WGO e, caso encontre o nome nessa geo-ontologia, verifica se esse nome está no domínio administrativo da WGO. Se encontra, atribui o identificador da entidade geográfica com conceito mais alto na hierarquia da WGO. Por exemplo, se encontra a EC *França*, atribui o conceito *país* e não *cidade* ou *vila*. Se não encontra, tenta atribuir o identificador da entidade geográfica do domínio físico da WGO. Caso o nome não esteja na WGO, o algoritmo procura na Geo-Net-PT. Se encontra, atribui o identificador da entidade geográfica com conceito mais alto na hierarquia da Geo-Net-PT, critério adotado para desambiguação também. A função `AnotaEG` anota as entidades geográficas conforme o domínio das geo-ontologias nos quais elas foram reconhecidas. Essa função é um simples conversor dos tipos das geo-ontologias para os tipos definidos no Segundo HAREM. Caso o nome não esteja em nenhuma das geo-ontologias, ele é anotado como um local com o tipo humano.

Nos casos de ambiguidade entre nomes do domínio administrativo e físico, o algoritmo 13.2 usa uma heurística que prioriza o domínio administrativo. Por exemplo, se um mesmo nome se refere a uma cidade e a um lago e não possui nenhum discriminador (conceito geográfico) no texto, o algoritmo assume que o nome se refere à cidade. Essa abordagem também foi usada em Volz et al. (2007).

Uma das vantagens de utilizar geo-ontologias na fase de classificação de locais é o fato de se reconhecer um local com um nível mais específico de granularidade. Ao invés de classificar o local como administrativo ou físico, é possível classificá-lo como uma freguesia, localidade ou lago, por exemplo.

13.2.3 Reconhecimento de relações semânticas entre EM – ReRelEM

Uma das funcionalidades do SEI-Geo é a extração de relações semânticas entre entidades geográficas. Uma das pistas do Segundo HAREM propõe o desafio de reconhecer relações entre EM. A participação do SEI-Geo nessa tarefa restringiu-se ao reconhecimento de relações entre entidades pertencentes à categoria local, que é um dos problemas tratados pelo SEI-Geo.

A abordagem utilizada pelo SEI-Geo na tarefa de reconhecimento de relações foi baseada em geo-ontologias. Todos os locais encontrados num documento são projetados sobre ontologias com o objetivo de encontrar relações de inclusão (`inclui`/`incluido`) entre eles. Caso encontre alguma relação, o SEI-Geo anota a mesma no documento. Essa abordagem permite testar até que ponto um algoritmo de reconhecimento de relações geográficas consegue ser preciso e abrangente só com o uso de ontologias.

Um fator importante a destacar é o âmbito no qual uma relação pode ocorrer. O SEI-Geo foi desenvolvido originalmente para relacionar locais dentro de uma mesma sentença. Entretanto, de acordo com as diretivas da tarefa de ReRelEM, relações devem ser identificadas ao nível do documento. As corridas contemplando essas duas variantes são descritas a seguir.

13.3 Descrição das corridas

Apesar de o Segundo HAREM promover o reconhecimento de diversas categorias (por exemplo, `PESSOA`, `ORGANIZACAO` e `TEMPO`), o SEI-Geo participou apenas no reconhecimento da categoria `LOCAL`, dos tipos `HUMANO` e `FISICO` e de todos os sub-tipos desses tipos. Dentro do mesmo evento também foi promovida a tarefa de reconhecimento de relações semânticas entre entidades mencionadas (ReRelEM). O SEI-Geo participou nessa tarefa anotando relações de inclusão entre locais.

O SEI-Geo participou no Segundo HAREM com quatro corridas. As variações realizadas nas quatro corridas do SEI-Geo, são as geo-ontologias de entrada do sistema e o âmbito das relações a serem reconhecidos.

Corrida 1 (Geo-Net-PT): utilizou somente a Geo-Net-PT até o nível de localidade, ou seja, conceitos e entidades geográficas acima do conceito de localidade inclusive.

Corrida 2 (WGO - Relação com âmbito no documento): utilizou apenas a WGO com nomes geográficos de todo o mundo, incluindo nomes de países, cidades capital, principais regiões administrativas e cidades com mais de 100.000 habitantes. Na tarefa de ReRelEM essa corrida anota relações existentes entre locais ao longo de todo o documento.

Corrida 3 (Duas ontologias - Relação com âmbito na sentença): o âmbito das relações foi restrito ao nível de sentença, conforme o SEI-Geo foi projetado originalmente. Na

tarefa de ReRelEM a corrida 3 anota relações existentes somente para locais que estejam na mesma sentença.

Corrida 4 (Duas ontologias - Relação com âmbito no documento): o âmbito das relações foi o documento completo, o que caracteriza a proposta original da tarefa ReRelEM.

As corridas 3 e 4 utilizaram as ontologias WGO e Geo-Net-PT, essa mutilada no nível de localidades, ou seja, os nomes de localidades não foram incluídos nessas corridas.

Sempre que o algoritmo encontra um mesmo nome em ambas, a opção é feita pela geo-ontologia WGO, uma vez que a mesma possui conceitos que estão na parte superior da hierarquia das ontologias. Por exemplo, *França* é uma freguesia do *concelho de Bragança* e um país, como país está acima na hierarquia das ontologias, o SEI-Geo assume que o nome *França* num texto refere-se ao país e não à freguesia, a não ser que esteja precedido pelo conceito *freguesia*.

13.4 Análise dos resultados

Essa seção descreve os resultados da participação do SEI-Geo no Segundo HAREM. A avaliação dos sistemas participantes nesse evento foi realizada através de seis cenários seletivos, nos quais a categoria `LOCAL` estava presente nos cenários 2, 3, 4, 5 e 6. O cenário seletivo 5 é formado exclusivamente pela categoria `LOCAL` com os tipos `FISICO` e `HUMANO` e todos seus subtipos, o que corresponde ao cenário de participação do SEI-Geo.

A tabela 13.2 apresenta os resultados alcançados no cenário seletivo 5 considerando a classificação com avaliação relaxada de ALT, uma vez que o SEI-Geo não usa a opção de marcação com a etiqueta ALT. A última linha da tabela 13.2 apresenta os resultados dos melhores sistemas para cada medida.

Tabela 13.2: Resultados cenário seletivo 5 considerando a classificação com avaliação relaxada de ALT.

Corrida	Classificação			Identificação		
	P	A	F	P	A	F
2	**0,6821**	0,5182	0,5890	0,7109	0,5346	0,6102
3	0,6801	0,5377	**0,6006**	**0,7075**	0,5552	0,6222
4	0,6726	**0,5413**	0,5999	0,7009	**0,5595**	**0,6223**
Melhor sistema	0,7105	0,7126	0,6325	0,7212	0,8017	0,6651

As figuras 13.3 e 13.4 apresentam um comparativo dos resultados do SEI-Geo comparados com os demais sistemas participantes. A figura 13.3, referente à classificação, mostra que o SEI-Geo, nas duas melhores corridas (3 e 4), conseguiu atingir resultados acima da média dos sistemas em todas as medidas: precisão, abrangência e medida F. No que diz respeito à identificação, a figura 13.4 apresenta o SEI-Geo com valores de precisão e medida F acima da média dos sistemas, mas com abrangência inferior, o que evidencia uma das limitações do SEI-Geo.

Os resultados da classificação, na tabela 13.2, evidenciam a qualidade do SEI-Geo no reconhecimento de locais para o português.

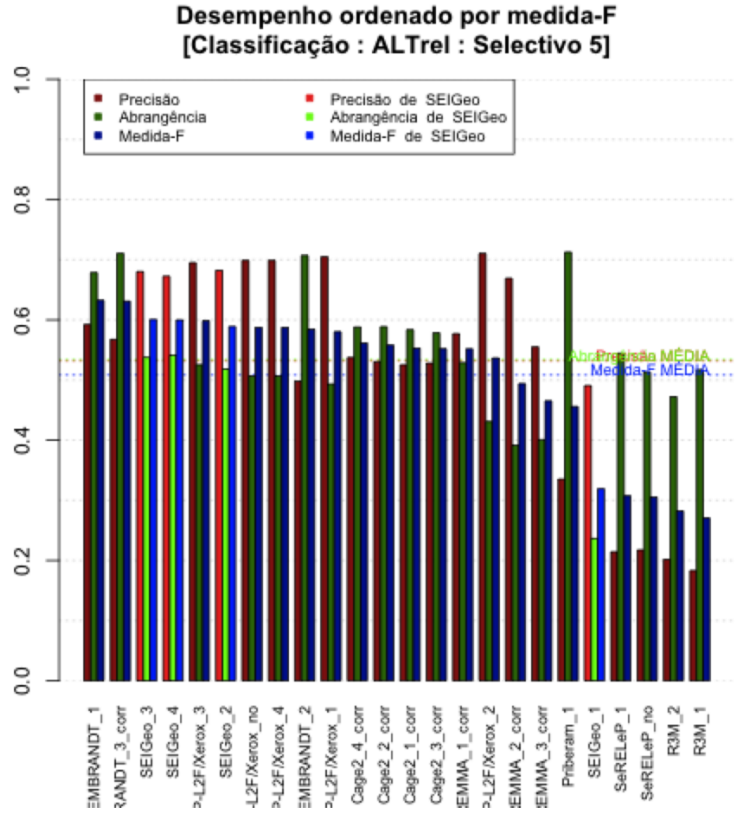


Figura 13.3: Cenário Seletivo 5 - Resultado da classificação ordenado pela medida F.

Além desses resultados, o SEI-Geo também alcançou o primeiro lugar na medida de precisão nos cenários total, 2, 3, 4 e 6 para a tarefa de classificação e identificação com avaliação relaxada de ALT. Os valores da precisão nesses cenários variam de 0,86 a 0,91.

Nos resultados por categoria, que tem em conta todos os tipos e subtipos da categoria LOCAL, a tabela 13.3 indica que o SEI-Geo aproxima-se bastante do melhor sistema nas medidas de precisão e medida F na tarefa de classificação. No que diz respeito à identificação, o SEI-Geo é o sistema mais preciso entre os concorrentes e alcançou um valor próximo ao melhor sistema na medida F.

A tabela 13.4 apresenta os resultados do SEI-Geo no HAREM clássico distribuídos por subtipos da categoria LOCAL. O SEI-Geo apresenta os melhores resultados para os subtipos PAIS, AGUACURSO e AGUAMASSA. Embora carecendo de informação sobre a geografia física na geo-ontologia, o SEI-Geo ainda é capaz de alcançar resultados competitivos por meio do uso de padrões para os subtipos físicos.

A tabela 13.5 apresenta os resultados das principais corridas do SEI-Geo na tarefa de ReRelEM. Apesar do sistema identificar corretamente as relações que se propõe a identificar, sua abrangência ainda é muito baixa, comparada ao melhor sistema nessa medida. Conforme já descrito no capítulo 4 e de acordo com a tabela 13.5, o SEI-Geo foi o melhor

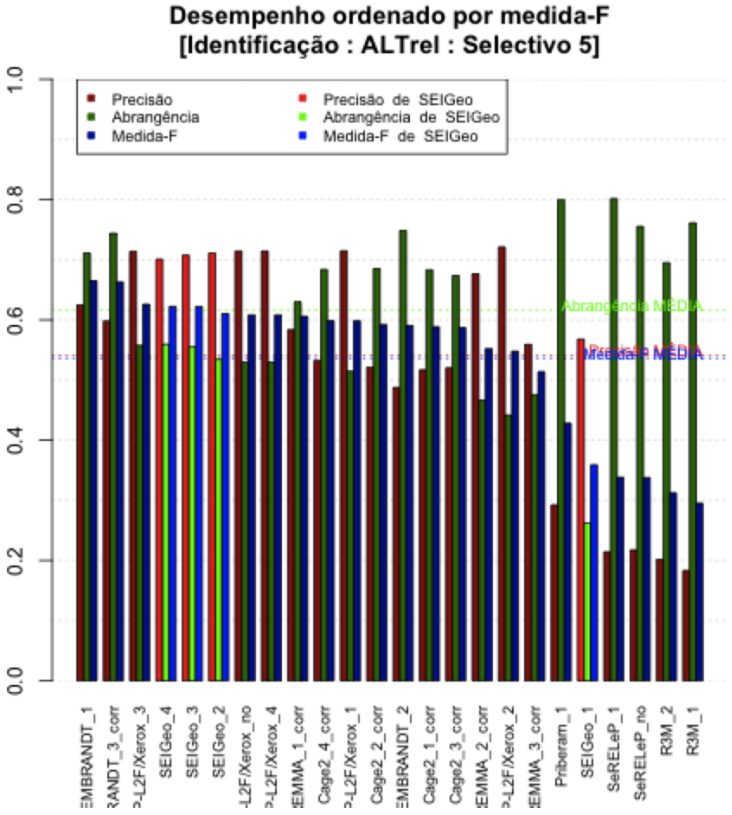


Figura 13.4: Cenário Seletivo 5 - Resultado da identificação ordenado pela medida F.

Tabela 13.3: Resultados da categoria LOCAL considerando a classificação com avaliação relaxada de ALT.

Corrida	Classificação			Identificação		
	P	A	F	P	A	F
2	**0,6830**	0,5029	0,5793	**0,7121**	0,5175	0,5994
3	0,6810	0,5215	**0,5906**	0,7087	0,5375	0,6113
4	0,6736	**0,5252**	0,5902	0,7020	**0,5416**	**0,6115**
Melhor sistema	0,6928	0,7015	0,6078	**0,7121**	0,7982	0,6376

sistema no reconhecimento de relações de inclusão para a categoria LOCAL. O SEI-Geo não reconheceu relações de identidade (por exemplo, USA=EUA) nos textos. O reconhecimento desse tipo de relação foi deixado para trabalho futuro.

Tabela 13.4: Avaliação dos subtipos da categoria LOCAL.

	Precisão	Abrangência	Medida F
PAIS	0,8488	0,6518	**0,7503**
DIVISAO	0,6384	0,3818	0,5101
REGIAO	1,0000	0,0448	0,5224
CONSTRUCAO	0,3636	0,0220	0,1928
RUA	0,4615	0,1818	0,3216
OUTRO	0,0408	0,0625	0,0517
AGUACURSO	0,7143	0,6250	**0,6697**
AGUAMASSA	0,8889	0,4444	**0,6666**
RELEVO	0,5714	0,4000	0,4857
PLANETA	0,3333	0,3333	0,3333
ILHA	0,3333	0,1111	0,2222

Tabela 13.5: Resultado da participação do SEI-Geo na tarefa de ReRelEM do Segundo HAREM - Avaliação de relações - Cenário total - Inclusão.

Corrida	P	A	F	Espúrios	Falta	Tot. CD	Tot. id.	Tot. correc. id.
3	1,0	0,0769	0,1428	0	72	78	6	6
2	**0,9166**	**0,2973**	**0,4490**	2	52	74	24	22
4	0,9166	0,2820	0,4314	2	56	78	24	22
Melhor sistema	1,0	0,4231	**0,4490**	0	52	74	24	22

13.5 Discussão

Após a análise dos resultados da participação do SEI-Geo no Segundo HAREM é possível concluir que a combinação das geo-ontologias WGO e Geo-Net-PT produziu os melhores resultados. A contribuição da Geo-Net-PT ainda é mínima, mas o suficiente para ser um diferencial quando os resultados são comparados com os outros sistemas participantes.

É importante notar que a Geo-Net-PT foi mutilada no nível de localidade. Os nomes de localidade inserem muitos falsos positivos no processo de reconhecimento de EM. O uso de nomes de localidade da Geo-Net-PT (por exemplo, *Caracol*, *Namorados* e *Nabo*) implica numa sobre-geração de EM reconhecidas. Por exemplo, nas sentenças (13.1) a (13.3) os nomes poderiam ser reconhecidos pelo SEI-Geo como locais por estarem nas geo-ontologias.

(13.1) *Caracol* é barato em Aveiro.

(13.2) *Namorados* são sempre felizes.

(13.3) *Nabo* de qualidade encontra-se na feira do Manuel.

A Corrida 1 alcançou o pior resultado entre as quatro corridas submetidas. A restrição à Geo-Net-PT implicou numa grande perda de abrangência e precisão do SEI-Geo. Esse resultado pode ser surpreendente quando comparado à participação do sistema CaGE no Mini-HAREM. Ao utilizar somente a Geo-Net-PT, o CaGE teve uma perda de precisão, abrangência e, consequentemente, medida F[3] que atingiu apenas 2 pontos percentuais em

[3] A medida F usando a Geo-Net-PT foi de 0,6063, enquanto usando a Geo-Net-PT+WGO o CaGE alcançou 0,6235.

relação à saída que utilizou a Geo-Net-PT e a WGO. Tal fato é um indício de que o SEI-Geo é bastante dependente do conteúdo das geo-ontologias, ao contrário do CaGE.

A principal limitação do SEI-Geo está na medida de abrangência. Tal fato pode ser justificado pela simplicidade do sistema, uma vez que não há análise sintática do texto, o conjunto de padrões é limitado e as ontologias são desprovidas de locais físicos (apenas a WGO contém locais físicos customizados para as participações do sistema da Universidade de Lisboa nas quatro edições do GeoCLEF).

Por outro lado, o SEI-Geo apresentou resultados satisfatórios para a medida de precisão nas corridas 3 e 4, obtendo o melhor resultado no cenário seletivo na tarefa de identificação de locais.

Os valores de precisão do SEI-Geo são prejudicados pelo fato de o sistema não discernir entidades mencionadas em contexto. Por exemplo, na frase (13.4) o SEI-Geo reconhece *Aveiro* como uma entidade mencionada da categoria `LOCAL` ao invés de `PESSOA` do tipo `POVO`, conforme a diretiva do Segundo HAREM.

(13.4) *Aveiro* estava em festa durante o Segundo HAREM.

Quanto a participação do SEI-Geo na tarefa de ReRelEM, os resultados indicam que a abordagem e as geo-ontologias utilizadas auxiliam bastante, mas não são suficientes para reconhecer relações entre locais em textos, apesar de o SEI-Geo ter sido o melhor sistema no reconhecimento de relações de inclusão.

Cabe ainda destacar que a tarefa de ReRelEM é avaliada sobre uma coleção de 12 documentos com 579 entidades mencionadas e 603 relações, que após expansão totalizam 5.716 relações.

Finalmente, uma nota sobre o custo computacional do SEI-Geo. A tabela 13.6 apresenta os tempos de processamento das corridas submetidas ao Segundo HAREM. Esses tempos foram obtidos em um servidor com sistema operacional Linux, processador Intel(R) Xeon(TM) CPU 3.20GHz e 8GB de memória.

Tabela 13.6: Tempos de processamento das corridas submetidas ao Segundo HAREM.

Corrida	1	2	3	4
Minutos	27	76	30	101

13.6 Conclusões

Esse capítulo descreveu a participação do SEI-Geo no Segundo HAREM, evidenciando os pontos positivos e negativos do sistema. A participação no HAREM clássico foi bem sucedida e o sistema atingiu resultados próximos aos sistemas que representam o estado da arte no REM em português. Na tarefa de ReRelEM, ainda há muito que melhorar no sistema dada a limitação do reconhecimento de relações baseado somente em geo-ontologias. Contudo, os melhores sistemas nessa tarefa alcançaram resultados que estão bastante distantes do expectável, considerando que em tarefas de REM e reconhecimento de relações são esperados valores mais elevados de medida F. Os resultados dos três sistemas participantes nessa tarefa apontam para a necessidade de novas abordagens para se tratar o problema.

Trabalhos futuros com o SEI-Geo incluem: melhor tratamento na identificação e reconhecimento de endereços, locais da geografia física e relações entre os locais. Especificamente nesse último item, a expansão das geo-ontologias com locais históricos, nomes alternativos e locais da geografia física é fundamental. Além disso, como o modelo-base da base de conhecimento onde as geo-ontologias estão armazenadas suporta a inserção de novos domínios de conhecimento, o domínio de organizações pode auxiliar na identificação e reconhecimento de relações, uma vez que locais frequentemente são referenciados próximos a organizações em textos. Finalmente, a exploração do uso de locativos para relacionar locais presentes na mesma sentença pode auxiliar o SEI-Geo a melhorar seu desempenho na tarefa de ReRelEM.

Agradecimentos

Este trabalho foi financiado pela FCT através do Projeto Linguateca (POSC 339/1.3/C/NAC), do Projeto GREASE II (PTDC/EIA/73614/2006) e pelo Programa de Financiamento Plurianual (LaSIGE).

Capítulo 14

Sistema SeRELeP para o reconhecimento de relações entre entidades mencionadas

Mírian Bruckschen, José Guilherme Camargo de Souza, Renata Vieira e Sandro Rigo

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 14, p. 247–260.

Neste capítulo, é apresentado um sistema focado no reconhecimento de relações entre EM. As subtarefas de identificação e classificação das EM são realizadas pelo analisador sintático PALAVRAS (Bick, 2000), deixando como tarefa do sistema aqui apresentado somente o reconhecimento das relações entre estas EM.

O sistema faz inferência das relações a partir de regras heurísticas simples, que consideram apenas informações presentes no próprio texto e informações adicionais providas pelo PALAVRAS. Assim, o sistema não faz uso de bases de conhecimento adicionais para o reconhecimento das relações, resolvendo a tarefa através de regras linguísticas e de posicionamento das EM em cada texto analisado.

O restante do documento está organizado da seguinte forma: a seção 14.1 relata brevemente a experiência anterior do grupo em análise de correferência, que foi parte da motivação para participação do Segundo HAREM; a seção 14.2 faz uma descrição detalhada do sistema projetado e desenvolvido, assim como uma discussão dos resultados; e a seção 14.3 finaliza o documento com considerações finais e futuras direções do trabalho.

14.1 Trabalhos relacionados e motivação

Uma das motivações que nos levou a participar do processo do Segundo HAREM foi a experiência anterior do nosso grupo na tarefa de resolução de correferência – que pode ser facilmente associada à relação de identidade proposta pelo Segundo HAREM.

Apesar de a proposta do HAREM ser uma tarefa diferenciada da nossa experiência com resolução de correferência, tratando exclusivamente nomes próprios mas também outras relações além de identidade, reconhecemos a importância da avaliação conjunta, que nunca ocorreu no contexto de resolução de correferência para a língua portuguesa antes do Segundo HAREM.

A abordagem para resolução de correferência, tal como adotada em Souza et al. (2008), leva em consideração não apenas as entidades mencionadas mas todos os tipos de sintagmas nominais referenciais presentes em um texto: indefinidos, definidos, pronomes e nomes próprios, mas sem dar enfoque à distinção de categorias (pessoa, local, organização). Assim, dado um conjunto de sintagmas nominais de um texto, o sistema tem por objetivo agrupar os sintagmas em cadeias que evocam a mesma entidade. O processo de resolução de correferência desenvolvido é formado por três momentos: (i) geração de pares de sintagmas nominais, (ii) classificação dos pares quanto a sua anaforicidade e (iii) agrupamento dos pares anafóricos em cadeias.

Junto com a geração dos pares de sintagmas, são verificadas características que são consideradas no processo de classificação automática por aprendizado. Essas características são informações morfossintáticas, posicionais e semânticas. O classificador é induzido por aprendizado de máquina supervisionado com base em um corpo anotado, o Summ-it[1] (Collovini et al., 2007).

O classificador indica quais pares de sintagmas são relacionados por anaforicidade. A partir dos pares identificados, os conjuntos de sintagmas correferentes são formados.

Existem outras abordagens bastante conhecidas e bem-sucedidas, mas geralmente aplicadas a outras línguas (o inglês, em particular) (Soon et al., 2001).

A participação no HAREM foi uma experiência complementar e inspiradora. Acreditamos que uma revisão e união das duas abordagens seja produtiva e importante para o

[1] Disponível em: http://www.inf.pucrs.br/~linatural/procacosa.htm.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<DOC DOCID="2ght33"> (...)
A relação do tecno com o binômio homem-máquina, no diálogo com <EM ID="2ght33-EM_2">Jeff Mills
    </EM>, o legendário produtor de tecno de <EM ID="2ght33-EM_3">Detroit</EM> nos leva a
    impressionantes insights do impacto que a terceira onda tem causado na paisagem
    contemporânea. <EM ID="2ght33-EM_4" COREL="2ght33-EM_3" TIPOREL="ident">Detroit</EM>,
    essa "cidade portátil", virtualizada na minimalista batida de um sequenciador automático,
     profetiza em sua música - que já nos deu a <EM ID="2ght33-EM_5">Motown</EM>, <EM ID="2
    ght33-EM_6">Stooges</EM>, e <EM ID="2ght33-EM_7">MC5</EM> - o zeitgeist deste início de
    milênio. (...)
</DOC>
```

Figura 14.1: Trecho de arquivo de saída do SeRELeP

entendimento do problema e aperfeiçoamento da tarefa. Focar nos diferentes tipos de categorias de cadeias, por exemplo, pode ser uma maneira de organizar os sistemas, com isso melhorar os resultados e refinar e sua avaliação.

14.2 SeRELeP: Sistema de reconhecimento de RElações em textos de Língua Portuguesa

SeRELeP é um **S**istema de reconhecimento de **RE**lações em textos de **L**íngua **P**ortuguesa. Foi desenvolvido visando a participação na pista de reconhecimento de relações entre EM (ReRelEM) do Segundo HAREM (consulte-se o capítulo 4 para uma apresentação da pista). A ferramenta, sua metodologia de desenvolvimento e resultados são detalhados nesta seção.

14.2.1 Visão geral

O SeRELeP propõe-se a demarcar as relações de identidade (`ident`), ocorrência (`ocorre_em`) e inclusão (`inclui`) entre EM conforme as diretrizes do Segundo HAREM, que se encontram no apêndice C. Partindo do reconhecimento e classificação de EM efetuados pelo analisador PALAVRAS, o sistema processa a coleção de textos do HAREM e retorna a mesma coleção com a anotação das relações entre EM.

O sistema tem como entrada o arquivo de texto da coleção do HAREM (em formato XML[2]) e seus respectivos arquivos em formato XCES[3]. Para obtenção dos arquivos neste formato, é necessário o processamento do corpo anotado em TigerXML (König e Lezius, 2003) pelo conversor Tiger2XCES (Bruckschen et al., 2008b). Como saída, o SeRELeP devolve um arquivo com o texto já marcado com as EM e suas relações, também em formato XML. A figura 14.1 traz um trecho de um arquivo de saída como exemplo. Na figura 14.2 é ilustrado todo o processo de anotação automática.

Nessa figura, SeRELeP é o sistema identificador de relações entre as EM, e o SeRELeP Tools é um conjunto de pequenos programas auxiliares, necessários à etapa de pré-processamento. A entrada do processo é um arquivo XML no formato do HAREM, fornecido no início da participação da avaliação conjunta. Este arquivo contém diversos textos individuais.

[2] *eXtensible Markup Language*
[3] XML CES: *Corpus Encoding Standard for XML*, conforme http://www.xces.org/

Os textos são extraídos pelo SeRELeP Tools em dois formatos: texto plano, que é a entrada para o PALAVRAS, e XML do HAREM, que é entrada para o SeRELeP efetivamente. Além do XML do HAREM, o SeRELeP ainda precisa de outra entrada, que são os textos anotados pelo PALAVRAS no formato XCES.

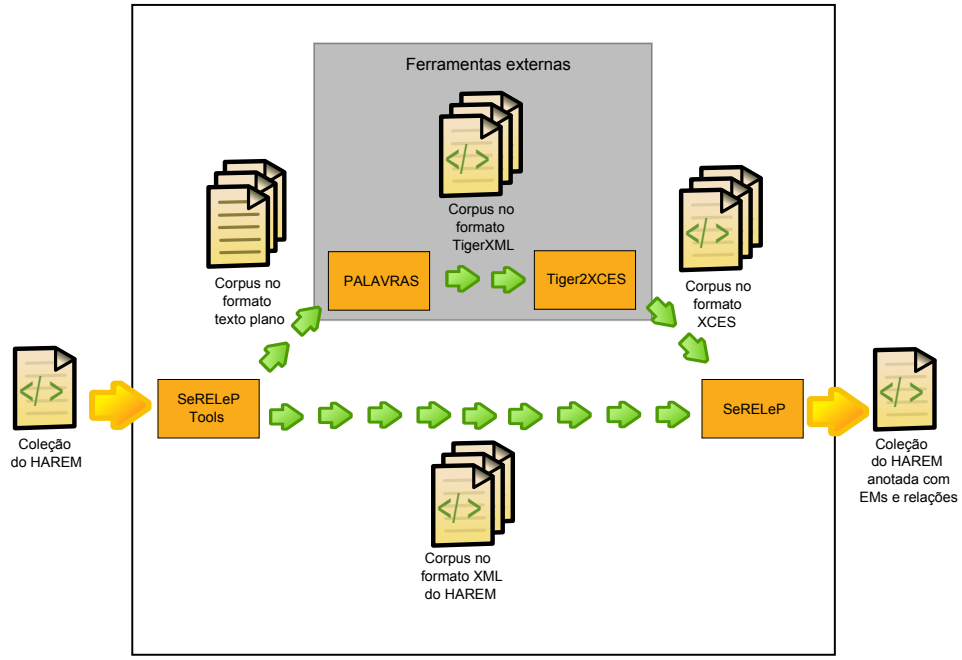


Figura 14.2: Processo de anotação automática de EM e relações da coleção do HAREM

A utilização do formato XCES é devida à diminuição de complexidade de interpretação por analisadores automáticos, pois é um formato proposto a fim de atender a vários critérios, levantados como necessários a um formato padrão de representação de informação linguística. Alguns destes critérios são expressividade, independência de mídia, adequação semântica, simplicidade (e legibilidade por humanos, tanto quanto possível), incrementabilidade e extensibilidade (Ide e Romary, 2004). O formato adotado baseia-se nessas diretivas e é apresentado em detalhe no relatório do projeto PLN-BR (Bruckschen et al., 2008a). Seguindo este formato, a anotação do PALAVRAS é codificada em três arquivos XCES a partir de cada texto original: `token`, `pos` e `phrase`. Cada um destes arquivos representa um nível de anotação linguística. Os itens de informação são delimitados pelo elemento `struct`, e todas as suas características por elementos `feat`. A estrutura do documento anotado é predominantemente vertical ao invés de horizontal; existem muitos elementos no arquivo, mas cada um destes possui poucos atributos. Esta característica favorece a criação de analisadores para este tipo de arquivo, e de igual forma torna a informação mais clara para leitura por seres humanos (Bruckschen et al., 2008a).

O arquivo `token` identifica as unidades lexicais (ou átomos). A cada unidade corresponde um elemento XML ao qual é dado um identificador, e nos campos `from` e `to` são informados o início e o fim deste elemento no texto, de acordo com sua posição em carac-

```
<?xml version="1.0" standalone="yes" ?>
<cesAna version="1.0.4" xmlns="http://www.xces.org/schema/2003">
 <struct to="5" type="token" from="0">
  <feat name="id" value="t1" />
  <feat name="base" value="Mills" />
 </struct>
 <struct to="7" type="token" from="6">
  <feat name="id" value="t2" />
  <feat name="base" value="é" />
 </struct>
 <struct to="10" type="token" from="8">
  <feat name="id" value="t3" />
  <feat name="base" value="um" />
 </struct>
 <struct to="16" type="token" from="11">
  <feat name="id" value="t4" />
  <feat name="base" value="homem" />
 </struct>
 <struct to="22" type="token" from="17">
  <feat name="id" value="t5" />
  <feat name="base" value="calmo" />
 </struct>
 <struct to="23" type="token" from="22">
  <feat name="id" value="t6" />
  <feat name="base" value="." />
 </struct>
</cesAna>
```

Figura 14.3: Trecho de arquivo de *token* no formato XCES

teres. No exemplo 14.1, o átomo *um* é identificado e vai da posição 9 à 10.

(14.1) Mills é um homem calmo

O arquivo `POS` (*part-of-speech*) representa a informação de nível morfossintático (as etiquetas semânticas também são representadas neste arquivo). Os elementos de `token` são referenciados em cada elemento de `POS`.

Finalmente, o arquivo `phrase` descreve a informação de nível sintático: sentenças e sintagmas, identificação de sujeitos, predicados e objetos. Os grupos sintagmáticos são identificados como intervalos de elementos `token`.

A seguir, são ilustrados os trechos de arquivos de `token`, `POS` e `phrase` do formato XCES referenciados. Estes são as figuras 14.3, 14.4 e 14.5, respectivamente. Todos eles ilustram a sentença 14.1.

O SeRELeP e seu módulo de programas auxiliares SeRELeP Tools foram desenvolvidos em Python[4], utilizando a biblioteca SAX[5] para processamento de arquivos XML.

A definição das classes para representação de informação linguística baseou-se no conversor Tiger2XCES, por sua vez, escrito na linguagem de programação Java.

O arquivo de programa principal é o `SeRELeP/serelep.py`, que processa os arquivos XCES (`token`, `pos` e `phrase`) e executa os dois principais métodos do processo de reconhecimento automático de relações entre as EM: i) a procura por nomes próprios e sua classificação, e ii) a inferência das relações a partir de critérios especificados nos métodos

[4] http://python.org/

[5] *Simple API for XML*, conforme disponível em http://www.python.org/doc/lib/module-xml.sax.html.

```
<?xml version="1.0" standalone="yes" ?>
<cesAna version="1.0.4" xmlns="http://www.xces.org/schema/2003">
 <struct type="pos">
  <feat name="id" value="pos1" />
  <feat name="class" value="prop" />
  <feat name="tokenref" value="t1" />
  <feat name="canon" value="Mills" />
  <feat name="gender" value="M" />
  <feat name="number" value="S" />
  <feat name="complement" value="hum" />
 </struct>
 <struct type="pos">
  <feat name="id" value="pos2" />
  <feat name="class" value="v-fin" />
  <feat name="tokenref" value="t2" />
  <feat name="canon" value="ser" />
  <feat name="complement" value="fmc" />
  <feat name="complement" value="mv" />
  <feat name="tense" value="PR" />
  <feat name="person" value="3S" />
  <feat name="n_form" value="VFIN" />
  <feat name="mode" value="IND" />
 </struct>
 <struct type="pos">
  <feat name="id" value="pos3" />
  <feat name="class" value="art" />
  <feat name="tokenref" value="t3" />
  <feat name="canon" value="um" />
  <feat name="gender" value="M" />
  <feat name="number" value="S" />
 </struct>
 <struct type="pos">
  <feat name="id" value="pos4" />
  <feat name="class" value="n" />
  <feat name="tokenref" value="t4" />
  <feat name="canon" value="homem" />
  <feat name="gender" value="M" />
  <feat name="number" value="S" />
  <feat name="semantic" value="Hattr" />
 </struct>
 <struct type="pos">
  <feat name="id" value="pos5" />
  <feat name="class" value="adj" />
  <feat name="tokenref" value="t5" />
  <feat name="canon" value="calmo" />
  <feat name="gender" value="M" />
  <feat name="number" value="S" />
  <feat name="complement" value="np-close" />
 </struct>
 <struct type="pos">
  <feat name="id" value="pos6" />
  <feat name="class" value="pu" />
  <feat name="tokenref" value="t6" />
 </struct>
</cesAna>
```

Figura 14.4: Trecho de arquivo de POS no formato XCES

```
<?xml version="1.0" standalone="yes" ?>
<cesAna version="1.0.4" xmlns="http://www.xces.org/schema/2003">
 <struct to="t5" type="phrase" from="t1">
  <feat name="id" value="phr1" />
  <feat name="cat" value="s" />
  <feat name="function" value="" />
 </struct>
 <struct to="t5" type="phrase" from="t1">
  <feat name="id" value="phr2" />
  <feat name="cat" value="fcl" />
  <feat name="function" value="STA" />
 </struct>
 <struct to="t1" type="phrase" from="t1">
  <feat name="id" value="phr3" />
  <feat name="cat" value="prop" />
  <feat name="function" value="S" />
 </struct>
 <struct to="t2" type="phrase" from="t2">
  <feat name="id" value="phr4" />
  <feat name="cat" value="v-fin" />
  <feat name="function" value="P" />
 </struct>
 <struct to="t5" type="phrase" from="t3">
  <feat name="id" value="phr5" />
  <feat name="cat" value="np" />
  <feat name="function" value="Cs" />
  <feat name="head" value="t4" />
 </struct>
</cesAna>
```

Figura 14.5: Trecho de arquivo de *phrase* no formato XCES

apropriados. O processo de reconhecimento de relações é descrito em detalhe na subseção 14.2.2.

Além das classes de representação da informação linguística do texto e analisadores para cada um dos tipos de entrada (`token`, `POS` e `phrase` em XCES, e XML do HAREM), foram desenvolvidos, de forma individual, os métodos para cada uma das técnicas para reconhecimento de relações (descritos na subseção 14.2.2). A figura 14.6 representa as relações entre principais arquivos de código-fonte e métodos do sistema.

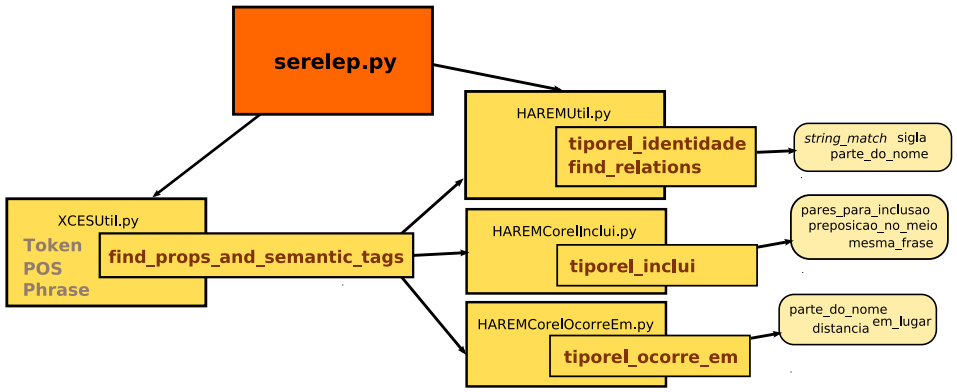


Figura 14.6: Visão geral do código-fonte e métodos principais do SeRELeP

Todos os métodos auxiliares usam primariamente dicionários e listas, que são estruturas básicas do Python, e bastante otimizadas. As informações utilizadas (como as etiquetas do PALAVRAS e sua correspondência com as categorias do HAREM) são também definidas na forma de dicionários. Esta decisão por certo influenciou positivamente no desempenho computacional do sistema, cuja única etapa mais demorada é o pré-processamento.

Além da utilização de estruturas básicas da linguagem, é preciso considerar o fato de que as regras utilizadas são simples e facilmente otimizáveis. Depois do pré-processamento, o processo todo não ultrapassa 10 minutos para o reconhecimento de relações e anotação de toda a coleção do Segundo HAREM – 1048 textos (tempo verificado num portátil Core Duo 1.6 GHz, 1GB DDR, executando a distribuição de Ubuntu GNU/Linux 7.10, com núcleo (em inglês, *kernel*) de Linux 2.6).

Com relação ao sistema, convém observar que foi desenvolvido de forma modularizada para facilitar a inclusão de novas técnicas e regras como, por exemplo, o uso de outras técnicas de reconhecimento de relações através da utilização de bases de dados externas.

14.2.2 Reconhecimento de relações entre entidades mencionadas

A marcação `prop` (nome próprio) do analisador PALAVRAS é utilizada para identificação e delimitação das EM no texto, e as suas etiquetas semânticas usadas para a classificação.

As etiquetas (e sua correspondência com as categorias do HAREM) foram as utilizadas na primeira edição do HAREM (Bick, 2007). A figura 14.7 ilustra estas etiquetas e sua correspondência com as categorias. O significado de cada etiqueta está disponível na página de documentação do PALAVRAS-VISL[6].

As associações realizadas entre etiquetas do PALAVRAS e classes do HAREM são diretas, não houve um tratamento adicional para a vagueza, a etiqueta `civ` sempre será referente a `LOCAL`, por exemplo.

Figura 14.7: Etiquetas semânticas do PALAVRAS e as categorias do HAREM

Conforme já comentado, o HAREM propõe quatro relações entre EM: `ident` (identidade), `inclui`, `ocorre_em` e `outra`. Destas, o SeRELeP trata as três primeiras. O tratamento de cada uma destas relações atualmente é detalhada mais adiante nesta seção.

[6] http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html

Convém notar que as heurísticas descritas abaixo fazem uso da etiquetação semântica do pré-processamento no que refere-se à categorização das EM. A figura 14.8 expressa na forma de um grafo dirigido as relações `ident`, `inclui` e `ocorre_em` entre as EM pertencentes a estas classes conforme tratado pelo SeRELeP. Salientamos que o grafo representado é uma simplificação, e que o nó `ORGANIZACAO/LOCAL` são de facto dois e `PESSOA/.../TEMPO` representa cinco nós distintos cada um deles referente a um tipo de EM (de outro modo, seria possível existir uma relação `ident`, por exemplo, entre `ORGANIZACAO` e `LOCAL`, o que como veremos mais adiante não acontece).

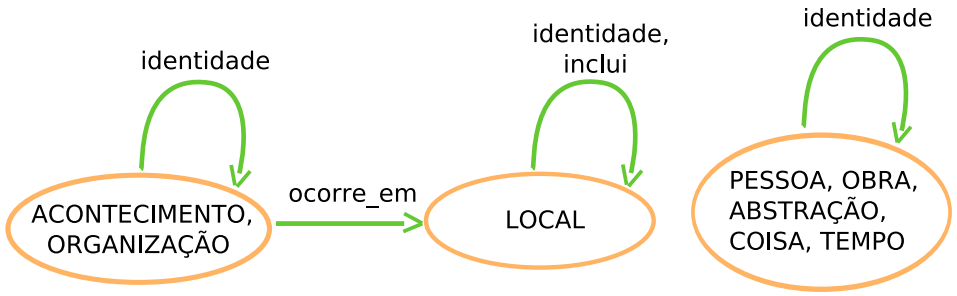


Figura 14.8: Relações e classes semânticas das EM conforme tratado pelo SeRELeP

As relações são reconhecidas numa determinada ordem, começando a partir da relação `ident`, já que as demais dependem dessa. Depois desta, são avaliadas as relações `inclui` e `ocorre_em`.

A relação `ident` é atribuída a EM que se referem a uma mesma entidade no mundo. A atribuição desta relação dá-se através das seguintes regras:

i) comparação direta de cadeias de carateres, isto é, se as EM possuem exatamente o mesmo nome;

ii) se uma é sigla da outra, isto é, se uma das EM retoma as iniciais da outra e possui mais de um caractere (a EM *DA* poderia ser relacionada com *Departamento de Artes*, por exemplo, mas *A*, isoladamente, não seria relacionada a *Artes*);

iii) se as EM comparadas forem da classe `PESSOA` e parte do sintagma de uma for igual ao sintagma da outra (como *Carmem* e *Carmem Miranda*, por exemplo). Além disso, as EM devem pertencer à mesma categoria semântica (uma EM de `ACONTECIMENTO` só pode estabelecer relação de identidade com outra EM de `ACONTECIMENTO`, por exemplo).

O exemplo 14.2 traz um trecho de texto com referências a uma mesma entidade que devem ser marcadas como possuindo a relação identidade entre si: *São Leopoldo* e *SL*.

(14.2) *São Leopoldo* é uma cidade localizada na região metropolitana de Porto Alegre, no Rio Grande do Sul. Dentre as diversas atrações da cidade, localiza-se em SL o Museu do Trem, um museu ferroviário, e um teatro recém inaugurado pela prefeitura junto à biblioteca municipal.

Já a relação `inclui`, tratada entre EM de `LOCAL` e simétrica da relação `incluído`, é estabelecida mediante as seguintes regras:

i) as duas EM não podem ter relação `ident` entre si;

ii) devem estar na mesma sentença;

iii) deve haver uma preposição que denote inclusão, como *em*, *no* e *na*.

Pode-se observar que são regras simples. Desta forma, não foi uma surpresa a baixa abrangência dos resultados desta relação. Os únicos casos de relações que seriam encontrados seriam os realmente explícitos, onde numa mesma sentença fosse referenciada a entidade incluída e a que a inclui.

O exemplo 14.3 traz um trecho de texto com entidades que estão ligadas pela relação de inclusão. No exemplo temos três entidades, e três relações de inclusão (uma delas implícita): *Rio Grande do Sul* `inclui` *São Leopoldo*, *Brasil* `inclui` *Rio Grande do Sul* e, consequentemente, *Brasil* `inclui` *São Leopoldo*.

(14.3) *São Leopoldo* é localizado no Rio Grande do Sul, no **Brasil**.

Finalmente, a relação `ocorre_em` é tratada entre EM de `ACONTECIMENTO` e `LOCAL` ou de `ORGANIZACAO` e `LOCAL`. As regras obedecidas por ela são verificadas na seguinte ordem:

i) se houver uma EM de `LOCAL`, cujo sintagma seja parte do sintagma da EM de `ACONTECIMENTO` ou `ORGANIZACAO` verificada, essa EM inserida é relacionada à EM de `LOCAL` em questão (como em *Brigada Militar de Porto Alegre* `ocorre_em` *Porto Alegre*);

ii) se isso não acontecer, é verificada a existência de uma EM de `LOCAL` na mesma sentença da EM de `ACONTECIMENTO/ORGANIZACAO` analisada. Se existir, esta EM de `ACONTECIMENTO/ORGANIZACAO` será relacionada a esta EM de `LOCAL` através da relação `ocorre_em`;

iii) se não, busca a EM de `LOCAL` mais próxima dentro do texto (se houver) para relacionar com a EM de `ACONTECIMENTO/ORGANIZACAO` analisada.

Com este conjunto de heurísticas, o SeRELeP obteve, entre os outros sistemas, o melhor resultado no reconhecimento da relação `ocorre_em`.

São ilustrados dois casos de ocorrência (de EM de `ACONTECIMENTO`) no exemplo 14.4.

(14.4) Ocorre em *São Leopoldo* a São Leopoldo Fest, festa que se dá em homenagem à chegada dos imigrantes alemães fundadores da cidade, e que reúne participantes de todo o estado. Além disso, a cidade é palco anualmente da sua Feira do Livro, com a participação de escritores, seus fãs, e diversas personalidade do mundo literário.

14.2.3 Resultados

Na coleção final anotada pelo SeRELeP, não classificamos as EM explicitamente, somente as suas relações, e por este motivo os resultados da classificação de EM (HAREM clássico) não são aqui exibidos. O principal motivo para não ser feita esta anotação foi o foco na tarefa de reconhecimento de relações. Sendo esta classificação resultado do processamento de uma ferramenta externa à desenvolvida e aqui descrita, não incluimos essa informação

na coleção anotada devolvida para avaliação – apesar de ter sido usada na inferência das relações.

Ainda assim, mostramos na tabela 14.1 os resultados da tarefa de identificação de EM no cenário total com avaliação estrita de ALT obtidos pela corrida SeRELeP_1. Esses são resultados obtidos pelo PALAVRAS, que mantém-se entre os melhores sistemas que participaram do Segundo HAREM na tarefa de identificação de entidades.

Tabela 14.1: Resultados oficiais do HAREM clássico (PALAVRAS)

	Precisão	**Abrangência**	**Medida F**
Identificação	0,82	0,60	0,69

Tabela 14.2: Comparativo dos resultados oficiais da pista do ReRelEM

	Precisão	**Abrangência**	**Medida F**
REMBRANDT_1	0,58	0,44	0,50
SeRELeP_1	0,58	0,31	0,40
SeRELeP_no	0,57	0,30	0,39
REMBRANDT_2	0,27	0,48	0,35
REMBRANDT_3_corr	0,25	0,48	0,32

Tabela 14.3: Comparativo dos resultados oficiais da pista do ReRelEM por relação

		Precisão	**Abrangência**	**Medida F**
REMBRANDT	ident	0,77	0,69	0,73
	inclui	0,32	0,33	0,33
	ocorre_em	0,40	0,13	0,20
SEI-Geo	ident	-	-	-
	inclui	0,92	0,30	0,45
	ocorre_em	-	-	-
SeRELeP	ident	0,89	0,55	0,68
	inclui	0,54	0,11	0,18
	ocorre_em	0,36	0,27	0,31

Os resultados da tarefa de reconhecimento de relações são ilustrados na tabela 14.2, que traz uma comparação da avaliação do reconhecimento de relações, no cenário seletivo do ReRelEM que inclui todas as relações menos `outra`, das cinco corridas melhores posicionadas. Diferentes corridas significam diferentes anotações, potencialmente por sistemas com ajustes e parâmetros diferentes. No caso particular do SeRELeP a corrida SeRELeP_1 distingue-se da corrida SeRELeP_no, não oficial, pela correção de um erro de delimitação das entidades identificado e reportado pela organização. A tabela 14.3 compara os resultados do SeRELeP com os dos restantes sistemas nos cenários seletivos do ReRelEM constituídos por cada uma das relações tratadas.

A relação com melhores resultados do SeRELeP é claramente a `ident`. Atribui-se este desempenho ao fato de que regras simples, como as utilizadas e descritas neste documento, já abrangem boa parte das relações entre EM. Ainda assim, o melhor classificado

nesta relação é o sistema REMBRANDT, enquanto o SeRELeP lidera no reconhecimento da relação `ocorre_em` e o SEI-Geo, a relação `inclui`.

Quanto ao reconhecimento de relações pelo SeRELeP, algumas questões ainda devem ser tratadas em maior detalhe, como apelidos não relacionados ao nome original (um exemplo seria *Pequena Notável* e *Carmem Miranda*, que têm relação de identidade não-detectada pelo sistema) e o mesmo nome com pequenas diferenças de grafia, comumente causados por erros de digitação (*Maria de Costa* e *Maria da Costa*).

As relações `inclui` e `ocorre_em` possuem resultados inferiores à relação `ident`.

O reconhecimento das relações, sobretudo nesses casos, pode ter sido afetado por problemas de classificação das EM[7]. Um exemplo disso é a marcação de EM de `LOCAL` tais como *Biblioteca Victor Civita* como `ORGANIZACAO`, no texto cujo trecho é ilustrado no exemplo 14.5. Em 14.6, por sua vez, a EM *Broadway* não é indicada na coleção dourada como um lugar, mas sim um grupo de pessoas, pertencendo à categoria semântica `PESSOA`, e não `LOCAL`. Já de acordo com o analisador sintático, a classificação resultante é `LOCAL`.

(14.5) Ele, que fez consultoria dos textos presentes na mostra, vai realizar uma palestra gratuita na sexta-feira, às 19h30, sobre a vida e a obra da artista na *Biblioteca Victor Civita*, no Memorial.

(14.6) *Carmen Miranda* conquistou a Broadway

Por outro lado, algumas EM corretamente classificadas tiveram relações identificadas incorretamente. No exemplo 14.7, temos algumas EM marcadas, e as seguintes relações identificadas: *África do Sul* `inclui` *Durban* (correta) e *Durban* `inclui` *Portugal* (incorreta). É possível perceber facilmente que o filtro simples aplicado a esta sentença (entidade de `LOCAL`, seguida pela preposição *em* e por outra entidade de `LOCAL`) ocasionou o problema, muito embora as entidades tivessem sido corretamente classificadas.

(14.7) Sua mãe casa-se pela segunda vez em 1895 por procuração, na *Igreja de São Mamede* em Lisboa, com o Comandante João Miguel Rosa cônsul de *Portugal* em Durban (África do Sul), o qual havia conhecido um ano antes.

Este é um dos exemplos que ilustra o motivo pelo qual entendemos que a utilização de informação externa ao texto (como bases de dados e ontologias de domínio – neste caso, geográfico) seriam acréscimos interessantes ao sistema apresentado.

14.3 Considerações finais

Motivadas por um fator crítico bastante rígido (tempo), as principais técnicas para o reconhecimento de relações foram baseadas em heurísticas simples, resultantes da análise dos exemplos no corpo.

Entendemos nossa participação como uma experiência modesta, uma vez que utilizamos ferramentas já existentes para o REM. A escolha do formato XCES deve-se à experiência na participação no projeto PLN-BR.

[7] Apesar de não ser anotada na coleção a classificação das EM, esta informação foi utilizada para a inferência das relações.

No entanto, o desenvolvimento do sistema demandou um tempo e esforço razoáveis, e conseguimos enviar a coleção do Segundo HAREM para participar da avaliação conjunta somente nos últimos dias do prazo final.

Tivemos uma surpresa muito positiva ao receber os resultados no final de agosto de 2008, e ver que o SeRELeP era bastante competitivo (com precisão melhor na maioria dos casos) com os outros dois sistemas concorrrentes na pista do ReRelEM. Ambos os sistemas, comparativamente, eram trabalhos mais maduros.

Atribuímos os bons resultados, em parte, à tarefa de REM muito bem-executada pelo PALAVRAS. Apesar dos problemas percebidos por nós no pré-processamento – e nossa queixa do quanto isso influenciou os nossos resultados, considerando que muitos dos erros provêm dessa etapa – também temos que mencionar que o PALAVRAS ainda é o melhor analisador morfossintático para a língua portuguesa.

Ainda, é importante lembrar que algumas técnicas para o reconhecimento das relações eram mais difíceis, e implementadas até em caráter experimental, mas que a maioria delas foi criada tendo com base a análise subjetiva de textos, procurando por padrões. As regras lingüísticas para reconhecimento da relação de identidade, por outro lado, ficaram óbvias tão logo nos debruçamos sobre as diretrizes da tarefa e vimos o quanto cada relação deveria abranger.

Com certeza, apesar de já positivos, estes resultados poderiam ser bastante aprimorados. Como trabalho futuro, pretendemos utilizar outras informações morfossintáticas adicionais como aposto e predicativo, que podem auxiliar principalmente na relação `ident`. Além disso, pretende-se usar algoritmos de distância de edição, como o utilizado para a correção ortográfica, a fim de tratar os casos onde há pequenas diferenças de grafia nos nomes das entidades (Navarro, 2001).

Também pretendemos explorar bases de conhecimento externas tais como ontologias de domínio, corpos ou conteúdos disponíveis na rede (inicialmente a partir da Wikipedia[8]), tal como realizado pelo sistema REMBRANDT (ver capítulo 11). Acreditamos que isso deverá melhorar substancialmente os resultados das relações `inclui` e `ocorre_em`.

Outra frente de pesquisa é a utilização do SeRELeP no auxílio à geração de ontologias a partir de processamento de textos. Para isso, consideramos importante abranger outras entidades do texto na composição dos relacionamentos, tais como substantivos comuns, a exemplo da resolução de correferência clássica.

Gostaríamos ainda de realizar uma avaliação do SeRELeP desconsiderando informação de vagueza, a fim de verificar como seriam os resultados se a avaliação considerasse por exemplo a categoria país englobando seus aspectos de localidade, admistrativos e de população.

Pode-se observar aplicações bastante interessantes para sistemas de reconhecimento de relações. Grande parte destas aplicações pode ser relacionada à extração e classificação de informações, como a procura de notícias sobre alguma pessoa ou lugar, posição (em inglês, *ranking*) de entidades mais evidentes, sistemas de identificação de assuntos ou notícias relacionadas, sistemas de respostas automáticas baseados em consultas à rede e análise semântica dos resultados destas consultas.

Como extensão deste trabalho, está sendo desenvolvido o sistema SeRELeP-Olympics (Bruckschen et al., 2008c), que trata do uso do reconhecimento da relação de identidade, inicialmente, para a lista de tópicos mais frequentes (em inglês, *hot topics*) num portal de

[8] http://pt.wikipedia.org/

notícias sobre as Olimpíadas. Nesse sistema, todas as notícias que referenciem a mesma entidade (como *Cielo*, *Cesar Cielo* e *Cesar Cielo Filho*, que nomeiam o nadador brasileiro ganhador da medalha de ouro) são marcadas como relacionadas àquela entidade, independente da forma com que ela foi referenciada. E, nesse caso, são consideradas as referências feitas em diferentes documentos. No futuro, pretende-se incluir nessa aplicação outras relações, como a de inclusão (por exemplo, ocorrências de *Pequim* deveriam aumentar a posição de *China*).

Agradecimentos

Agradecemos ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e à FAPERGS (Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul) pelo apoio no desenvolvimento deste trabalho.

Capítulo 15

Reconhecimento de entidades mencionadas com o XIP: Uma colaboração entre a Xerox e o L2F do INESC-ID Lisboa

Caroline Hagège, Jorge Baptista e Nuno Mamede

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 15, p. 261–274.

Apresentamos neste capítulo um sistema de reconhecimento de entidades mencionadas (REM) desenvolvido numa colaboração entre o L^2F (INESC-ID Lisboa) e o XRCE (Xerox Research Centre Europe, Grenoble, França). Trata-se de um sistema baseado em regras (no que diz respeito à parte do reconhecimento das entidades mencionadas) e integrado numa ferramenta mais geral de análise sintáctica do português, XIP (Xerox Incremental Parser). Uma das características da nossa abordagem é que a parte de REM está completamente integrada numa cadeia mais geral do processamento do português que vai da segmentação à análise sintáctica.

O capítulo desenvolve-se do seguinte modo: Começaremos por apresentar brevemente a ferramenta que usamos (secção 15.1), descrevendo a estratégia adoptada para enriquecer o analisador sintáctico com um módulo de REM. Numa segunda parte, descreveremos em pormenor as várias etapas de processamento e os recursos empregues, léxicos (secção 15.2) e regras (secção 15.3), no reconhecimento das diversas categorias de EM. Procuraremos, sobretudo, dar uma panorâmica geral do funcionamento sistema e explicar a metodologia seguida no seu desenvolvimento. Apresentaremos também o mecanismo de propagação (secção 15.4.4), disponível no XIP, que permite inferir novas EM a partir de EM previamente reconhecidas. Finalmente (secção 15.5), comentaremos os resultados obtidos e faremos um balanço da nossa participação na avaliação conjunta do Segundo HAREM.

15.1 XIP: Uma ferramenta para o processamento lexical, sintáctico e semântico

Começamos por fazer uma breve apresentação do sistema XIP (Xerox Incremental Parser) que utilizamos para a tarefa de reconhecimento de entidades mencionadas.

O XIP (Ait-Mokhtar et al., 2002) é um analisador cujo primeiro objectivo é a extracção de dependências sintácticas. O analisador processa documentos em formato de texto ou XML e produz como saída uma representação sintáctica do conteúdo do documento.

O XIP oferece um formalismo rico, que permite expressar um leque importante de regras, que vão da desambigação das categorias das palavras, até à construção de dependências, passando pela delimitação de sintagmas nucleares[1].

Importa, no entanto, frisar desde já que o conceito de dependência que adoptámos no XIP não corresponde à noção de dependência considerada, por exemplo, em Tesnières (1959) mas constitui uma perspectiva muito mais abrangente, nomeadamente por não respeitar o princípio de projectividade. Também difere da noção de dependência apresentada em Tapanainen e Järvinen (1997) por se aplicar a relações não necessariamente binárias e por poder permitir relações não só entre unidades lexicais mas também entre sintagmas nucleares.

O XIP é actualmente utilizado no processamento de várias línguas (inglês, francês, japonês, italiano, espanhol e português), estando as gramáticas dessas línguas em diferentes estádios de desenvolvimento.

15.1.1 Ilustração

Para ilustração, veja-se nas figuras 15.1 e 15.2 a análise feita pelo XIP da frase (15.1).

[1] Traduzimos a palavra do inglês "chunk" por "sintagma nuclear". Mais precisamente, por sintagma nuclear consideramos um grupo sintáctico, não recursivo, cujo limite direito corresponde à cabeça sintáctica dum sintagma tradicional.

```
TOP{
 PP{Na visão}
 PP{do ministro}
 NP{o seguro}
 AP{agrícola}
 VF{desempenhará}
 NP{importante papel}
 PP{no projeto}
 PP{do Governo}
 VINF{de estimular}
 NP{a agricultura}
 PP{através do NOUN{programa Brasil Empreendedor Rural}} .
}
```

Figura 15.1: Construção de sintagmas nucleares

```
MAIN(desempenhará)
DETD(visão,a)
DETD(ministro,o)
DETD(seguro,o)
DETD(projeto,o)
DETD(Governo,o)
DETD(agricultura,a)
DETD(programa Brasil Empreendedor Rural,o)
PREPD(visão,Na)
PREPD(ministro,do)
PREPD(projeto,no)
PREPD(Governo,do)
PREPD(programa Brasil Empreendedor Rural,através do)
MOD_PRE(papel,importante)
MOD_POST(seguro,agrícola)
MOD_POST(visão,ministro)
MOD_POST(projeto,Governo)
MOD_POST(estimular,programa Brasil Empreendedor Rural)
SUBJ_PRE(desempenhará,seguro)
CDIR_POST(desempenhará,papel)
CDIR_POST(estimular,agricultura)
```

Figura 15.2: Dependências (principais) extraídas

(15.1) Na visão do ministro, o seguro agrícola desempenhará importante papel no projeto do Governo de estimular a agricultura, através do programa Brasil Empreendedor Rural.

Nesta saída do sistema, pode-se observar que, além da delimitação dos sintagmas nucleares (NP, PP, AP, etc.), o XIP permite também extrair relações gramaticais entre constituintes, tais como sujeito (SUBJ) ou complemento directo (CDIR), além de identificar o núcleo da frase (MAIN) e várias outras relações sintácticas, tais como a relação entre o determinante e o núcleo nominal por ele determinado (DETD) ou entre preposição e núcleo nominal (PREPD). Finalmente, são também extraídas algumas relações genéricas de modificação (MOD) entre um núcleo (seja ele verbal, nominal ou de outra categoria) e um argumento ou modificador deste núcleo[2] .

[2] Nesta fase do desenvolvimento da gramática do português, por falta de informação lexical sistemática, ainda não fazemos a distinção entre complementos (argumentos de um operador) e adjuntos.

15.1.2 Desenvolvimento do módulo de REM

O desenvolvimento do módulo de REM[3] seguiu uma metodologia que obedece a duas orientações gerais:

- integração do módulo de REM no âmbito mais abrangente do processamento morfossintáctico do português.
- tratamento incremental da informação linguística.

15.1.2.1 Integração do REM no processamento geral do português

A integração do módulo de REM na cadeia de processamento é motivada por vários factores: em particular, o reconhecimento das EM permite melhorar os resultados dos outros módulos de processamento linguístico. Com efeito, as entidades mencionadas constituem superficialmente uma estrutura sintáctica por vezes complexa. No entanto, enquanto EM, elas correspondem muitas vezes a um nome. Por exemplo, a expressão *E tudo o vento levou*, que corresponde ao título de uma obra, tem a estrutura superficial de uma frase. Contudo, para proceder a uma correcta análise sintáctica da frase (15.2), é necessário determinar que *E tudo o vento levou* corresponde a um nome, para, por exemplo, não se interpretar *E* como uma coordenação e se estabelecer adequadamente a relação de dependência entre *ontem* e *fomos rever* e não entre o advérbio e o verbo *levou*.

(15.2) Fomos rever *E tudo o vento levou ontem*

O facto de se ter acesso à estrutura sintáctica permite ir mais longe na tarefa de REM. De facto, é possível, graças à informação sintáctica, resolver certos casos de uso metonímico de EM. Por exemplo, pode-se determinar o emprego metonímico de *Portugal* como `PESSOA` (`GRUPOIND`) em vez de `LOCAL` em exemplos como *Portugal respondeu...* sabendo que, aqui, *Portugal* - por defeito um nome de lugar - é o sujeito de um *verbum dicendi*.

Sobre as vantagens de integrar o REM na análise sintáctica, veja-se por exemplo Brun e Hagège (2004).

15.1.2.2 Tratamento incremental da informação linguística

Apresentamos na figura 15.3 a arquitectura geral, ilustrando a forma como o módulo de REM está integrado no sistema desenvolvido para o português (XIP-PT).

15.2 Léxico e pré-processamento

15.2.1 O que é uma entrada lexical no XIP?

Uma entrada lexical no XIP corresponde a um conjunto de traços (atributos-valores). Todos os traços e todos os valores possíveis têm de ser declarados explicitamente, com excepção de alguns traços geridos pelo sistema que são os traços `lemma`, `surface`, `maj` e `toutmaj`.

Os traços `lemma` e `surface`, cujo valor é uma cadeia de caracteres, correspondem, respectivamente, ao lema da unidade linguística e à forma de superfície da unidade linguística;

[3] Este desenvolvimento foi iniciado pelos trabalhos de Loureiro (2007) e Silva Romão (2007).

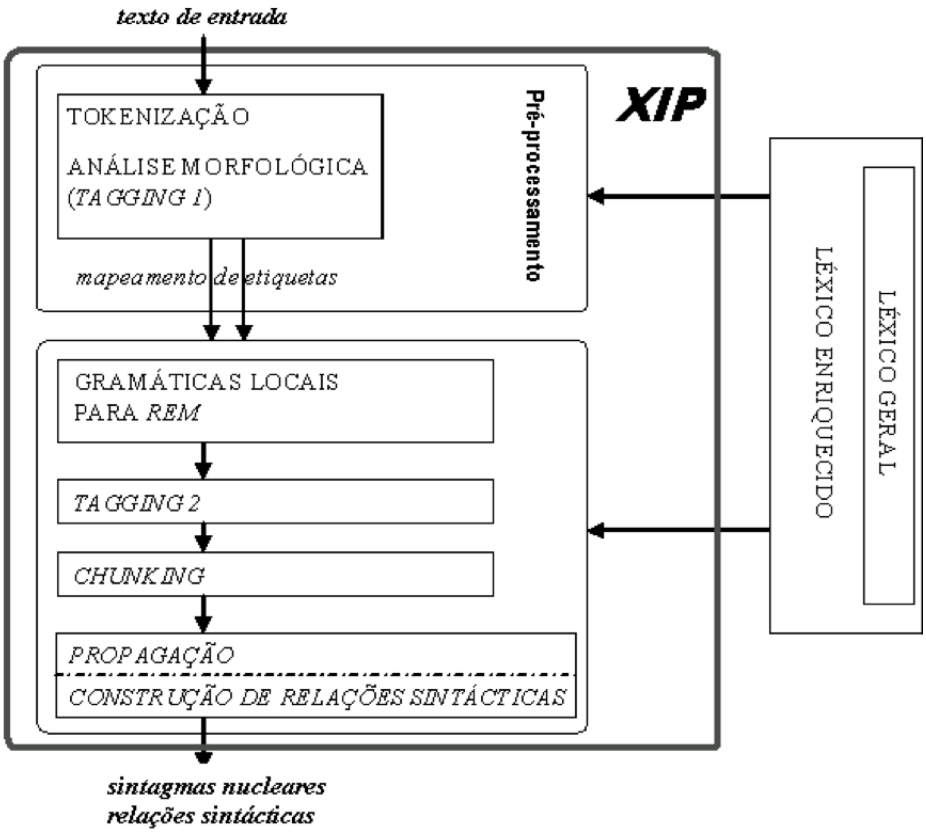


Figura 15.3: Arquitectura geral do sistema XIP-PT

`maj` e `toutmaj` são traços booleanos que indicam, respectivamente, se a forma de superfície começa por uma maiúscula, ou se a forma de superfície é totalmente em maiúscula. Todos os outros traços associados ao léxico são traços sintácticos ou semânticos e são definidos e declarados na gramática.

15.2.2 Dois tipos de léxicos

Consideramos dois tipos de léxico no XIP:

- léxico pré-existente
- léxico definido no XIP

15.2.2.1 Léxico pré-existente

Chamamos léxico pré-existente ao léxico oriundo da ferramenta de análise morfológica (e, possivelmente, do processo de anotação morfossintáctica (em inglês, *POS tagging*), a que chamamos pré-processamento sintáctico. Para integrar este léxico no XIP, é necessário

definir o mapeamente entre as categorias e traços do pré-processamento sintáctico (ver figura 15.3) e as categorias e traços que vão ser manipulados dentro do XIP[4].

15.2.2.2 Léxico definido no XIP

No XIP, também é possível definir directamente entradas lexicais ou, então, modificar entradas lexicais pré-existentes. Para o Segundo HAREM, privilegiámos esta última abordagem. Sendo assim, os dados lexicais necessários para a tarefa de REM foram acrescentados, sob a forma de léxico XIP, aos léxicos gerais utilizados pelo analisador.

A construção de novos recursos lexicais, definidos no XIP especificamente para a tarefa de REM, consistiu, pois, essencialmente nos seguintes passos:

- introdução de novas entradas que constituem EM:

  ```
  Herbert += [people:+,individual:+,firstname:+].
  ```

  Aqui, definiu-se uma nova entrada `Herbert`, à qual se associaram novos traços booleanos, a saber: `people:+, individual:+, firstname:+` . São estes traços que permitem marcar esta entrada como o primeiro nome de uma pessoa.

- enriquecimento de entradas do léxico pré-existente com novos traços:

  Este processo consistiu basicamente na marcação de elementos linguísticos que funcionam como pistas contextuais, isto é, que servirão em seguida para a identificação e classificação de EM (ver mais adiante a apresentação das regras locais). Exemplo:

  ```
  arcebispo:  noun += [cargo:+].
  ```

  Aqui, acrescentámos à entrada (lema) `arcebispo`, já existente nos léxicos, o traço `cargo:+`. Este traço será depois utilizado por diversas regras (ver adiante).

É de notar que tanto a parte de enriquecimento como a parte de novas entradas foi feita com base em listas de palavras já existentes ou criadas manualmente para o efeito. Por outras palavras, não utilizámos processos automáticos para enriquecimento lexical.

15.2.3 Adaptação do pré-processamento

Além do enriquecimento do léxico, as regras de desambiguação categorial da gramática geral foram adaptadas especificamente para a tarefa de REM. Para desambiguação categorial, adoptámos também uma abordagem incremental que pode ser resumida da maneira seguinte:

- uma primeira fase de desambiguação por regras (a que chamamos `Tagging1` na figura 15.3);
- uma segunda fase de desambiguação por regras (integrada no módulo a que chamamos `Tagging2`, na figura 15.3);

[4] Nesta participação no HAREM, trabalhámos com duas ferramentas de pré-processamento distintas (pertencentes a cada uma das instituições – INESC-ID e XEROX – que colaboraram neste trabalho). O resultado de cada mapeamento permitiu que, mesmo sendo dois léxicos distintos, fosse possível utilizar uma gramática e módulo de REM comuns no XIP.

- uma escolha por defeito realizada por um HMM (*Hidden Markov Model*) igualmente integrado em `Tagging2`.

A primeira fase de desambiguação é muito específica e corresponde à desambiguação particular de certas formas linguísticas. Por exemplo, a regra:

```
5> verb<lemma:podar>, verb<lemma:poder> = verb<lemma:poder> |
verb[inf:+] | .
```

pode ser descrita por: antes de uma forma verbal no infinitivo não flexionado, a unidade lexical *pode* é uma forma do verbo modal *poder*, e não a forma do presente do conjuntivo do verbo *podar*.

Para a tarefa de REM, além das regras existentes, foram acrescentadas novas regras que dizem respeito explicitamente aos acréscimos lexicais que foram feitos para o REM. Por exemplo, a regra seguinte permite desambiguar a entrada `Natal` (quadra festiva ou estado do Brasil):

```
20> noun[maj:+,surface:Natal] %= | noun[denot_time:+],
prep[lemma:de], art | noun[one_day=+,maj=+,proper=+] .
```

Esta regra determina que, depois de uma palavra como *altura* ou *tempo* seguida pela preposição *de* e um artigo, a palavra *Natal* corresponde à quadra festiva (a interpretação como estado do Brasil é então excluída).

15.3 Gramáticas locais para o REM

15.3.1 Expressão de gramáticas locais em XIP

O XIP oferece um formalismo que permite, entre outras coisas, exprimir regras de reescrita tomando em consideração, facultativamente, os contextos à esquerda e à direita da expressão regular a reescrever:

```
LHS = | <reg_expr_esq> | <reg_expr> | <reg_expr_dir> |
```

`LHS` vai corresponder a um novo nó resultante do emparelhamento de `reg_expr` (expressão regular de categorias, aumentadas pela possibilidade de fazer restrições sobre os traços associados a estas categorias), com, eventualmente, a verificação dos contextos à esquerda e à direita (`reg_expr_esq` e `reg_expr_dir` respectivamente) que correspondem também a expressões regulares de categorias.

A este novo nó, representado por `LHS`, podem ser acrescentados novos traços (na medida em que eles forem previamente declarados).

Este formalismo é usado para definir regras de gramáticas locais para as EM em dois tipos de situações:

- para a delimitação de EM constituídas por mais de uma unidade lexical;
- na utilização do contexto imediato para delimitar e classificar EM.

15.3.2 Delimitação de EM complexas

Algumas das EM a reconhecer são constituídas por mais de uma palavra gráfica (unidades), possibilidade que, naturalmente, está contemplada nas directivas do Segundo HAREM. É o caso de *Oceano Atlântico*, *senhor Pedro da Conceição* e muitos outros.

Contudo, nas fases preliminares de pré-processamento, as várias unidades que constituem estas expressões produtivas apenas foram considerados individualmente, sendo função das gramáticas locais juntar agora esses elementos numa única EM.

Por exemplo, a regra a seguir constrói um nome complexo ao qual se acrescentam os traços `cargo:+` e `people:+`, para uma sequência que começa por um elemento lexical que tem o traço `cargo:+`, seguido ou pelo adjectivo *honorário* ou pelo adjectivo *mor* em maiúscula.

```
1> noun[cargo=+,mwe=+,people=+] @=
?[cargo,maj], (punct[hifen]),
adj[lemma:"honorário", maj]; adj[lemma:mor].
```

Assim, sequências como *Cônsul Honorário* ou *Sargento-mor* vão, graças a esta regra, ser consideradas como nomes de cargo.

15.3.3 Utilização de contexto imediato

Para algumas unidades lexicais, é o contexto imediato que permite reconhecer ou classificar uma EM. As regras locais utilizando o contexto adaptam-se perfeitamente a esta tarefa. No exemplo seguinte, apresentamos uma dessas regras:

```
1> NOUN[org=+, institution=+] @= |?[lemma:governo, maj: ],
prep[lemma:de], (art)| ?[location].
```

Esta regra faz com que, num contexto à direita constituído por *governo* seguido da preposição *de* e eventualmente seguido por um artigo, um elemento lexical marcado com o traço `location` passe a ser classificado como uma organização institucional.

Note-se aqui a restrição (`maj:`$\sim$, isto é, a palavra não deverá começar por maiúscula) associada ao lema *governo* no contexto à direita: esta restrição é devida às actuais directivas do Segundo HAREM que estipulam que os nomes de organizações devem sempre começar por maiúscula. Assim, numa expressão como *o governo de Lisboa*, só a palavra *Lisboa* será marcada como uma organização.

É de salientar que estas regras se podem aplicar a sequências de categorias ambíguas. Relembramos que, na arquitectura que definimos (v. figura 15.3), a aplicação das regras locais para EM se faz depois da aplicação de um primeiro módulo de desambiguação, mais específico, e que a maior parte das ambiguidades categoriais ainda não foram inteiramente resolvidas. Estas gramáticas locais procedem, pois, a uma desambiguação suplementar, na medida em que, se houver emparelhamento com uma regra, serão seleccionadas as categorias com que essas regras emparelharem.

A seguir a estas regras locais será, então, aplicado o módulo de desambiguação (misto, isto é, combinando HMM e regras), que permitirá resolver as ambiguidades restantes (`Tagging2`, na figura 15.3).

15.4 Últimas fases de processamento das EM

A possibilidade de utilizar expressões regulares contextuais (ver secção acima) permite atingir um certo grau de generalização na formulação e representação desse contexto. No entanto, para um grau de generalização ainda maior, utilizam-se outras técnicas que necessitam de uma análise linguística mais complexa: no caso da tarefa de REM, são necessárias a análise do texto em sintagmas nucleares (particionamento, em inglês *chunking*) e o cálculo de relações sintácticas (dependências) entre potenciais entidades mencionadas e outros constituintes da frase (Brun e Hagège, 2004). As últimas fases do processamento das EM consistem, assim, no aproveitamento dos módulos de particionamento, de construção de dependências e de propagação de traços. Todos estes módulos se encontram integrados no XIP (v. figura 15.3). É deles que falaremos a seguir.

15.4.1 Particionamento

O módulo de particionamento do XIP permite fazer uma análise sintáctica preliminar do texto, construindo para cada frase uma sequência de sintagmas nucleares. As entidades mencionadas reconhecidas nas fases anteriores (por codificação lexical ou através das gramáticas locais) recebem, de um modo geral, a etiqueta `NOUN`, o que permite que elas se integrem naturalmente nas regras gerais de construção dos sintagmas nucleares (em inglês, *chunks*) nominais da gramática. Por outras palavras, EM delimitadas (simples ou complexas) vão ser núcleos nominais de sintagmas nominais nucleares.

15.4.2 Dependências

A construção das dependências permite exprimir relações sintácticas, como as de sujeito, objecto directo, etc., entre os diversos constituintes das frases. Além das relações gramaticais clássicas, construídas pela gramática, cria-se para as EM uma nova relação unária (`NE`), cujo argumento consiste na EM reconhecida. A esta relação são associados os traços que permitem classificar o tipo de EM.

Ilustramos a análise em sintagmas nucleares e em dependências com o exemplo `ex:xip:joaninha`.

(15.3) Joaninha Sampaio vivia na Lourinhã

A figura 15.4 apresenta a interface gráfica do XIP, na qual se mostra a análise em sintagmas nucleares (NP, VF, PP). A primeira parte da saída representa a árvore de sintagmas nucleares (i.e., os sintagmas nucleares estão todos ligados a um nó `TOP`).

A sequência *Joaninha Sampaio* foi correctamente delimitada e etiquetada como um único nome (`NOUN`), tendo sido também classificada como um nome de pessoa (`NE_INDIVIDUAL_PEOPLE(Joaninha Sampaio)`). Este nome complexo constitui o núcleo do sintagma nuclear nominal (`NP`) *A Joaninha Sampaio*. Verificamos ainda que foram construídas várias dependências gramaticais, como, por exemplo, a relação de sujeito (`SUBJ`) entre *viver* e *Joaninha Sampaio*.

As dependências unárias `NE*` correspondem às EM que foram encontradas. O nome genérico da relação unária é `NE`, a que se juntam, ligados por caracteres de sublinhado, os traços associados à dependência e que, neste caso, consistem na classificação destas EM.

Estas dependências são criadas graças aos traços que foram previamente associados aos nomes *Joaninha Sampaio* e *Lourinhã*.

As outras relações correspondem à relações sintácticas calculadas entre diversos constituintes. É o aproveitamente destas relações que permite generalizar contextos para o cálculo de novas EM (ver ponto seguinte).

15.4.3 Generalizando o contexto para classificar EM

Um dos problemas com que se defronta a tarefa de REM consiste na resolução dos casos de metonímia, aspecto que, naturalmente, também está contemplado nas directivas do Segundo HAREM. Um exemplo típico desta situação consiste no uso de nomes de países para se referir ou ao conjunto dos habitantes/o povo ou às instituições da respectiva organização política. Muitos destes fenómenos de transferência metonímica seguem padrões regulares (cf. avaliação conjunta de detecção de metonímia de SemEval 2007 (Markert e Nissim, 2007)). Contudo, a fim de capturar estes fenómenos de metonímia, é necessário levar em consideração um contexto relativamente alargado. As regras contextuais das gramáticas locais não seriam, então, o formalismo mais adequado ou mais eficiente para representar esse tipo de contexto.

Para ilustrar o que dizemos, tomemos o exemplo (15.4).

(15.4) Portugal ratificou o tratado.

O nome *Portugal* está marcado no léxico como nome de país. A priori, no fim da cadeia de processamento, esta unidade lexical seria classificada como uma EM de tipo `LOCAL`. No entanto, neste contexto sintáctico, isto é, como sujeito de um verbo como ratificou, *Portugal* não designa aqui o espaço geográfico de um país mas sim a entidade que representa a sua organização político-administrativa (ou eventualmente, mas noutros contextos, um grupo de pessoas). Ora, é graças às dependências previamente calculadas, nomeadamente à relação de sujeito (ou de agente da passiva) entre *Portugal* e o verbo *ratificar* que é possível que o sistema corrija a classificação cega do nome de país como `LOCAL` e, tal como se indica nas directivas gerais de classificação do Segundo HAREM, passe a tratá-lo, pois, como `ORGANIZACAO`.

A grande vantagem de levar a cabo esta correcção ao nível das dependências resulta da possibilidade de, com apenas uma regra (ver adiante) dar conta de casos como: `ex:xip:naoratificou`, `ex:xip:foiratificado`, `ex:xip:queratificou`, etc. pois, para todos estes casos, a palavra *Portugal* é analisada como estando numa relação sujeito ou agente com o verbo ratificar.

(15.5) Portugal ainda não ratificou o tratado

(15.6) O tratado foi ratificado por Portugal

(15.7) Portugal, que ratificou este tratado

Eis um exemplo de uma regra que permite transformar uma EM de tipo geográfico em EM de tipo organização quando ela é sujeito (ou agente) de *ratificar*.

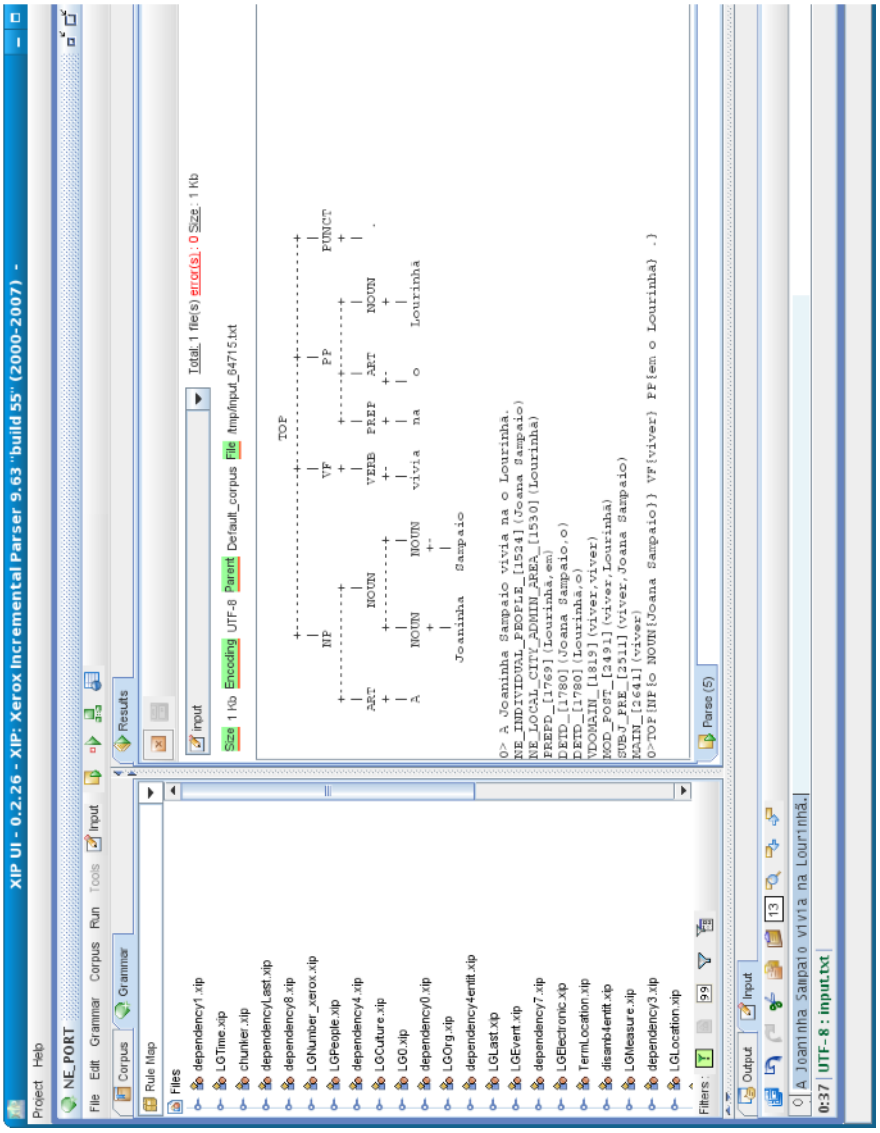


Figura 15.4: Exemplo de análise das EM, dos sintagmas nucleares das dependências pelo XIP

```
if ( ^NE[local:+,admin_area:+](#1) &
            ( SUBJ(?[lemma:ratificar],#1)
            | AGENT(?[lemma:ratificar],#1)
            )
   )
NE[features=\~,org=+,administration=+](#1)
```

Para esta avaliação conjunta do HAREM, limitámo-nos a uma primeira abordagem, ainda muito preliminar, do tratamento da metonímia. A maior parte dos casos de metonímia previstos nas directivas não estão ainda contemplados no nosso sistema, seja por falta de tempo seja por discordância com algumas das escolhas da organização. Com efeito, consideramos que o emprego metonímico de uma EM deveria passar a ser explicitamente anotado, de modo a ter em paralelo a classificação semântica básica (literal ou por defeito) e a classificação (ou interpretação semântica) que resulta do fenómeno de transferência metonímica de acordo com Brun et al. (2007).

Note-se que as nossas regras de correcção para dar conta dos empregos metonímicos foram feitas manualmente. No entanto, tendo a possibilidade de dispor de um corpus anotado, a aquisição automática deste tipo de regra pode ser explorada.

15.4.4 Propagação

A propagação é um mecanismo que permite conservar a informação sobre EM previamente calculadas e propagar essa informação ao resto da análise de um texto. Este processo parte do pressuposto de que, num mesmo documento, novas EM são introduzidas num contexto suficientemente rico para que possam ser classificadas de forma não ambígua; no entanto, muitas vezes, essas EM são retomadas nesse mesmo texto mas já sem se apresentarem nessa distribuição característica. Trata-se, tipicamente, do caso de nomes de pessoas mas pode também acontecer com outro tipo de entidades.

O XIP oferece, além das operações habituais para o processamento linguístico, uma linguagem dedicada (em inglês, *scripting language*), que pode ser usada para algumas tarefas simples (contagens de ocorrências ou de relações por exemplo). A propagação é realizada graças a esta linguagem dedicada e processa-se em dois passos:

1. Marcação da EM
   Se houver uma EM atestada (marcada no léxico ou calculada por regras locais), então a totalidade ou parte da sequência correspondente a esta EM será marcada graças às variáveis particulares geridas pelo XIP.

2. Restituição da EM
   Se, no decorrer da análise, for encontrada uma sequência de caracteres previamente marcada, pode-se então associar-lhe uma operação qualquer (neste caso a construção de uma dependência `NE` unária associada a esta sequência).

O exemplo (15.8) (extraído da colecção do Segundo HAREM) ilustra o processo de propagação:

(15.8) Um capitão norueguês chamado *Trygve Petersen* conduziu o Mira de novo a Portugal... <frase intermédia>... *Petersen* não trazia carga nenhuma.

Nem *Trygve* nem *Petersen* estavam previamente codificados no léxico. No entanto, graças ao contexto, no primeiro caso, e ao mecanismo de propagação, no segundo caso, é possível obter os seguintes nomes de entidades:

```
NE_INDIVIDUAL_PEOPLE(Trygve Petersen)
NE_INDIVIDUAL_PEOPLE_PROPAG(Petersen)
```

Num primeiro passo, *Trygve Petersen* é classificado como uma entidade de tipo `PESSOA`, por se tratar de um complemento de *chamado*. A cadeia *Trygve Petersen* recebe a categoria `NOUN`, à qual vão estar associados os traços `individual:+` e `people:+`.

Num segundo passo, graças à linguagem dedicada, integrada no XIP, variáveis numéricas, que são indexadas aos lemas *Trygve* e *Petersen* (representados respectivamente por `PERSON##2` e `PERSON##3`) e que foram anteriormente inicializadas a 0, vão passar a ter o valor 1.

```
Script:
noun#1[people,individual]{?*,noun#2[title:~,location:~,org:~,initial:~,maj:+],
  ?*,noun#3[last,title:~,location:~,initial:~,maj:+]} |
if (NE[people](#1) )
{ PERSON##2=1; PERSON##3=1; }
```

Esta instrução faz com que cada lema correspondente a sub-sequências de um nome complexo de tipo `PESSOA` (traços `people:+` e `individual:+`) venha a ser marcado de igual modo graças a estas variáveis. Neste exemplo concreto, os lemmas `Trygve` e `Petersen` são, pois, associados às variáveis `PERSON`, cujo valor será igual a 1 (as variáveis de lemas são inicializadas a 0 por defeito).

A regra seguinte exemplifica a propagação destes traços:

```
DependencyRules:
| noun#1[toutmaj:~,maj:+] |
    if ( PERSON##1:1 & ~NE[people](#1) )
NE[people=+,individual=+,propag=+](#1)
```

Esta regra determina que: (i) se a um nome começando por maiúscula (traço `maj:+`) mas não totalmente em maiúsculas (traço `toutmaj` negado) for associada uma variável `PERSON` cujo valor é 1; e (ii) se este nome ainda não estiver marcado como sendo `NE` de tipo `PERSON` deve, então, ser criada uma relação unária `NE` com traços `people:+` e `individual:+` à qual se acrescenta a informação de que se trata do resultado de uma propagação (`propag=+`) desses traços.

Assim, na medida em que a variável `PERSON` foi inicializada a 1 durante o processamento da primeira frase, a ocorrência isolada de *Petersen* na segunda frase do exemplo será também considerada como um nome de pessoa e será marcada com uma dependência unária que classifica os nomes de pessoas individuais.

As variáveis são associadas de forma consistente aos lemas durante toda a fase de análise (i.e. todo o documento que estiver a ser processado). No entanto, é possível durante o processamento reinicializar o valor destas variáveis. Tipicamente, no caso do Segundo HAREM, consideramos que não é desejável propagar variáveis além do âmbito de um documento.

Uma vez que os documentos estão delimitados por balizas </DOC>, produzimos a regra seguinte:

```
Script:
| #1[lemma:"</DOC>"]; #1[lemma:"</doc>"] |
{ CleanAllLemmas; }
```

Esta regra determina que, cada vez que o segmento </DOC> ou </doc> for encontrado num texto, todas as variáveis de lemas sejam reinicializadas a 0.

A propagação é um mecanismo extremamente poderoso, que permite aumentar a abrangência (em inglês, *recall*) de um sistema de REM. No entanto, pode também ter efeitos perversos, sobretudo se a entidade inicial não tiver sido correctamente classificada.

Para o Segundo HAREM, utilizámos o mecanismo de propagação de uma maneira limitada e exclusivamente para os nomes de pessoa[5]. Deixamos para uma próxima avaliação conjunta do HAREM a complexa tarefa de desenvolver e estender a outros casos as regras de propagação, contando levar, então, em linha de conta o grau de confiança associado à classificação de uma EM reconhecida na propagação dos seus traços a outras EM.

15.5 Resultados e perspectivas

Os resultados que obtivemos no Segundo HAREM foram bastante encorajadores. Nesta primeira participação numa avaliação conjunta de REM, fomos o terceiro sistema em termos de medida F (obtendo até os melhores resultados em medida F para o cenário selectivo 2), considerando que não levámos em conta algumas das categorias previstas para o Segundo HAREM como ABSTRACCAO e COISA. Investimos bastante energia na tarefa de reconhecimento das expressões temporais e, tanto para o TEMPO clássico, como para a tarefa específica do TEMPO de acordo com a proposta por nós apresentada (Hagège et al., 2008), obtivemos os melhores resultados.

No nosso trabalho, favorecemos claramente a precisão em relação à abrangência. Achamos, no entanto, que a abrangência poderá ainda ser bastante aumentada e com relativa facilidade uma vez que, por falta de tempo, deixámos por codificar muito léxico (já identificado) e apenas utilizámos de forma incipiente e exploratória o mecanismo de propagação de traços.

A experiência desta nossa participação no HAREM mostrou-nos que negligenciámos alguns aspectos, como o da formatação dos resultados finais, tarefa que nos ocupou muito mais tempo do que esperávamos e que terá prejudicado duas das nossas corridas. Certamente levaremos isto em consideração numa próxima avaliação conjunta.

Temos consciência de que muito ficou ainda por fazer, tanto do ponto de vista da codificação do léxico, como do ponto de vista do desenvolvimento das regras das gramáticas locais e de propagação de traços. Estamos confiantes de que o sistema ainda tem bastante margem para melhoramento.

[5] Ainda não fizemos a avaliação quantitativa do benefício da propagação para o sistema mas, a título indicativo, foram detectados 1700 nomes de entidades da categoria PESSOA na colecção do Segundo HAREM graças a este mecanismo sobre um total de 10017 entidades com esta categoria.

Apêndices

Apêndice A

Segundo HAREM: Directivas de anotação

Diana Santos, Paula Carvalho, Cláudia Freitas e Hugo Gonçalo Oliveira

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Apêndice A, p. 277–286.

Nota das editoras: Este apêndice reproduz a versão 4.1 das directivas do HAREM clássico, que foi actualizada pela última vez no dia 12 de Março de 2008. Incluímos também o elenco de categorias e a descrição da sintaxe, que se encontravam em páginas distintas na rede, nas secções A.4 e A.5, respectivamente, do presente documento. Incluímos igualmente na secção A.6 a lista de minúsculas disponibilizada e actualizada pela última vez no dia 7 de Abril de 2008.

Este texto descreve a tarefa objecto do Segundo HAREM, concentrando-se nas modificações em relação ao Primeiro, já bem documentado em Cardoso e Santos (2007).

Como seria de esperar, o Segundo HAREM vai ser mais abrangente que o anterior, não só ao corrigir e melhorar algumas arestas em relação ao Primeiro (muitas delas já discutidas no livro (Santos e Cardoso, 2007a)), mas por incluir duas novas tarefas/pistas, nomeadamente a normalização de expressões temporais (apêndice B) e a detecção de relações semânticas entre EM, o ReRelEM (apêndice C).

De forma a compatibilizar todas estas alterações num único formato, tornámos a sintaxe mais flexível, combinando numa mesma caracterização de saída a identificação de (i) apenas categorias, (ii) categorias e tipos, e (iii) categorias, tipos e subtipos, sendo todas estas classificações opcionais.

Todas as EM começam com <EM ID="xxx"> e acabam com </EM>. O único atributo obrigatório é o ID; que, para facilidade de processamento, restringimos a uma combinação de apenas letras não acentuadas (maiúsculas ou minúsculas), algarismos, e os caracteres "-" e "_". Veja a secção A.5 para mais pormenores.

Note-se também que, visto que CATEGs, TIPOs e SUBTIPOs são opcionais, passa a haver uma maior clarificação no significado de OUTRO, que não significará ignorância, visto que esta será marcada pela **falta** de valor desse atributo. OUTRO indica assim explicitamente uma classificação distinta do elenco sugerido (seja a nível das CATEGorias, dos TIPOs ou dos SUBTIPOs).

Da mesma forma, considera-se opcional a identificação da relação entre duas EM (COREL) e o tipo de relação (TIPOREL), assim como os vários atributos associados a uma análise mais fina de expressões temporais.

A.1 Motivação para as presentes directivas

Embora a nova organização tenha naturalmente algumas opiniões divergentes em relação à anterior (em particular em relação à elegância de algumas distinções), tentámos, excepto nos casos mais problemáticos, manter aquilo que já tinha sido feito na edição anterior, para poupar trabalho aos antigos participantes e garantir alguma continuidade.

Descrevemos, em seguida, as modificações que as diferentes categorias sofreram. Excepto em relação aos subtipos, todas essas modificações se encontram reflectidas na nova versão das colecções douradas do Primeiro HAREM.

A.2 Questões de delimitação

Mudámos ligeiramente a definição operacional de EM, de três formas.

A.2.1 Desaparecimento de entidades complexas

No Primeiro HAREM, tínhamos algumas categorias que poderiam ser designadas como entidades complexas, ou semi-estruturadas, cuja identificação – embora extremamente relevante num contexto de extracção de informação – era difícil de conceber como REM, como era o caso de

- moradas (anterior `CATEGORIA LOCAL` e `TIPO CORREIO`)
- referências bibliográficas (anterior `CATEGORIA OBRA` e `TIPO PUBLICACAO`)

que pensamos agora fazer mais sentido analisar em termos dos mais pequenos constituintes, aliás em termos semelhantes ao que já tinha sido feito para outro tipo de "entidades complexas", como, por exemplo

- informações sobre direitos de autor (copyright notices) em páginas Web

A.2.2 Tratamento mais convencional de expressões com várias palavras

Além disso, e visto que a sugestão de ter EM iniciadas por "de" não foi considerada satisfatória, passámos a considerar que

- algumas das expressões que haviam sido classificadas como EM não o eram de facto [Por exemplo, "de Belém" em *pastéis de Belém*.]
- outras deveriam continuar a ser identificadas como constituintes de uma EM maior, a qual compreenderia todos os termos (eventualmente grafados com inicial minúscula) que designam a classe ou o objecto que essa EM representa. [Ex: "gaiola de Faraday" e não apenas "de Faraday"]

Resumindo, o critério formal da obrigatoriedade de maiúscula na identificação de EM mantém-se (ou seja, "médio oriente" não é considerado EM), excepto para o `TEMPO`, em que as regras são diferentes.

Contudo, quando outras expressões que fazem claramente parte da EM se encontram grafadas em minúsculas devem ser igualmente identificadas, pois a incorrecta identificação das EM põe em causa a sua própria classificação.

- correcto: [ministro da Administração Interna] — `PESSOA/CARGO`
- incorrecto: ministro da [Administração Interna] — `DISCIPLINA/ABSTRACCAO`
- correcto: [relógio de Sol] — `COISA/CLASSE`
- incorrecto: relógios de [Sol] — `FISICO/PLANETA`

Note-se contudo que isto é apenas válido para casos em que se pode defender que estamos em presença de uma expressão com várias palavras, e não é para ser generalizado à detecção de sintagmas nominais. Assim, em *a casa do João* apenas *João* como pessoa deve ser marcado.

A.2.3 Introdução de intervalos de valores como EM

Também inspirados pela proposta de reclassificação das entidades temporais, decidimos que intervalos de valores, assim como a especificação mais fina desses valores passava a fazer parte integrante da EM de VALOR.

Por exemplo, veja-se a frase

> Ele saltou <EM ID="" CATEG="VALOR" TIPO="QUANTIDADE" SUBTIPO="n">**entre 7 a 10 metros**</EM>a sua fuga.

em que uma EM substitui as duas que seriam esperadas no Primeiro HAREM.

A.3 Mudanças por categoria

Passamos agora a fazer um apanhado das mudanças nas categorias.

A.3.1 VALOR

Mantemos a classificação anterior, com os tipos CLASSIFICACAO, MOEDA e QUANTIDADE.

A única diferença é que intervalos de valores, como *entre 3 e 4%* ou *de 5 a 10 kg*, passam a ser uma única EM, assim como as EM também incluem as preposições ou quantificadores relacionados com outras formas de descrever uma quantidade, tal como *cerca de 200 gramas*, *menos de 10%* ou *aproximadamente 15 euros*.

A.3.2 VARIADO

Deixa de haver a categoria VARIADO, passando a haver também a categoria OUTRO, com a mesma interpretação do OUTRO nos tipos ou subtipos.

A.3.3 PESSOA

Foi adicionado mais um tipo, o de POVO, para cobrir casos como *Não há música como a do Brasil*, *A House Music conquistou Inglaterra, Holanda, Alemanha e Ibiza* ou *Lisboa ficou horrorizada com essa notícia*.

Além disso, não sofreu modificações, excepto na lista de formas de tratamento, que foi actualizada, entre outras coisas ao tentar incluir-se mais sistematicamente as usadas no Brasil.

A.3.4 ORGANIZACAO

Passou-se SUB para SUBTIPO do tipo de organização em questão, ou seja, passam a ser possíveis os casos

- CATEG="ORGANIZACAO" TIPO="INSTITUICAO" SUBTIPO="SUB"
- CATEG="ORGANIZACAO" TIPO="EMPRESA" SUBTIPO="SUB"
- CATEG="ORGANIZACAO" TIPO="ADMINISTRACAO" SUBTIPO="SUB"

O subtipo SUB não será contudo alvo de análise, anotação ou comparação no Segundo HAREM, mas foi mantido nas CD do Primeiro HAREM por uma questão de consistência.

A.3.5 LOCAL

Como indicado acima, deixámos de considerar o tipo CORREIO como uma EM, preferindo a marcação separada de ruas, estados e países dentro de moradas.

Além disso a informação marcada como LOCAL ALARGADO no Primeiro HAREM passou a ser considerada como informação adicional em relação aos tipos ADMINISTRATIVO ou GEOGRAFICO (agora rebaptizados de HUMANO ou FISICO). Assim, as EM anteriormente marcadas como LOCAL tipo ALARGADO passam a ter um SUBTIPO.

Passa pois a existir apenas uma tripartição da categoria LOCAL em FISICO, HUMANO e VIRTUAL, em que FISICO substitui o anterior termo GEOGRAFICO, e HUMANO o anterior termo ADMINISTRATIVO.

Além da categoria TEMPO, esta foi a única categoria onde os participantes desejaram uma classificação mais fina de subtipos.

Esta lista é o resultado da discussão pelos participantes envolvidos (mencionados na secção dos agradecimentos), a qual não pretende de forma alguma ser uma descrição exaustiva de todos os tipos conceptuais de lugares em português, mas apenas a soma das várias sensibilidades, experiências e opiniões da organização e dos já mencionados participantes.

Para locais de tipo HUMANO (note-se que os nomes são indicativos, não exaustivos)

- PAIS: inclui países, principados, e uniões de países, como é, por exemplo, o caso da União Europeia
- DIVISAO: inclui agregados populacionais como metrópoles, cidades, aldeias, vilas ou freguesias, assim como outras divisões administrativas tais como estados (Brasil), concelhos, distritos, províncias (Portugal), continentes, ou bairros fiscais
- REGIAO: localização cultural ou tradicional, sem valor administrativo, tal como a Baixa, o Grande Porto, o Médio-Oriente, o Terceiro Mundo ou o Nordeste (brasileiro)
- CONSTRUCAO: inclui todo o tipo de construções, desde edifícios, aglomerados de edifícios ou zonas específicas de um edifício (por exemplo, sala, galeria, jardim ou piscina), a pontes, barragens, portos, etc.
- RUA: inclui todo o tipo de arruamentos, como como ruas, avenidas, estradas, travessas, praças, pracetas, becos, largos, etc.
- OUTRO

Para locais de tipo FISICO

- AGUACURSO: inclui rios, ribeiros, riachos, afluentes, quedas de água, etc.
- AGUAMASSA: inclui lagos, mares, oceanos, golfos, estreitos, canais, bacias, barragens, etc.
- RELEVO: inclui montanhas, cordilheiras, montes, serras, planícies, planaltos, vales, etc.
- PLANETA: inclui todos os corpos celestes
- ILHA: inclui ilhas e arquipélagos

- REGIAO: designa uma região geográfica/natural, tal como o Bósforo, ou os Bálcãs, a Meseta Ibérica, a região do Amazonas, o Deserto do Sahara, ou os continentes vistos como região da geografia física
- OUTRO

Quanto a locais de tipo VIRTUAL, que indica localização abstracta, propomos os seguintes SUBTIPOS

- COMSOCIAL: inclui todos os meios de comunicação social, como jornais, televisão, rádio
- SITIO: inclui todos os locais virtuais no sentido electrónico: Web, WAP, ftp etc.
- OBRA: referência a uma obra impressa
- OUTRO

ilustrados respectivamente pelos seguintes exemplos:

- Essa afirmação saiu no *Diário de Notícias* ontem.
- Vai ao *Público on-line* ou à *Linguateca* e vê os anúncios que lá estão.
- No último *Harry Potter* vem a explicação da morte do Dumbledore.

Note-se que se a comunicação social é explicitamente na internet, então é SITIO. De resto, se nada for dito sobre isso (ou seja, não estiver a indicação *online* ou outra forma de o indicar, por exemplo através de um URL), então considera-se COMSOC.

Note-se também que, além de deixar de considerar URL como EM, também deixámos fora números de telefone e de fax.

A.3.6 ACONTECIMENTO

Não sofreu alteração, ou seja, mantém inalterados os tipos EFEMERIDE, EVENTO e ORGANIZADO.

A.3.7 OBRA

Em relação às EM da categoria OBRA, reduzimos os tipos de OBRA aos seguintes:

- REPRODUZIDA, da qual há muitas cópias/exemplares
- ARTE, que significa peça única
- PLANO, que se distingue das outras OBRAs pelo seu carácter contingente e circunstancial (note-se que estava anteriormente em ABSTRACCAO)

Na mesma linha que retirámos a categoria de CORREIO dos LOCAL, deixámos de entrar em conta com PUBLICACAO, que deixa de ser considerada uma EM de todo.

A.3.8 ABSTRACCAO

Esta categoria foi consideravelmente simplificada, retendo apenas os tipos DISCIPLINA, ESTADO, IDEIA e NOME. Por um lado, foram retirados desta categoria os tipos MARCA (convertido para categoria COISA de tipo CLASSE ou IDEIA) e PLANO (passado para categoria OBRA tipo PLANO). Por outro lado, DISCIPLINA, ESCOLA e OBRA foram todas juntas em DISCIPLINA. Muito resumidamente, então:

- DISCIPLINA: passou a referir quer uma disciplina ou área, quer uma escola literária, científica, artística, religiosa ou ideológica, ou mesmo um estilo musical.
- ESTADO representa, como anteriormente, sobretudo doenças.
- IDEIA é a mais abstracta das abstracções
- NOME, como anteriormente, representa um objecto linguístico e não a entidade que designa.

A.3.9 COISA

Esta categoria foi a que sofreu mais alterações, e por isso decidimos reescrever completamente as directivas no que lhe diz respeito.

Em primeiro lugar e como já mencionado, mudámos os critérios de identificação, de forma a que casos em que apenas por questões de convenção se grafam com letra maiúscula deixem de ser abrangidas pela noção de EM: ou seja, *pastéis de Belém*, *flauta de Bisel*, visto que estão em variação livre com *pastéis de feijão* ou *guitarra acústica*, e apenas se grafam em maiúscula por os seus nomes derivarem de locais ou pessoas.

Noutros casos sobretudo de terminologia científica, mantivemos a classificação de COISA tipo CLASSE mas identificando o conceito todo, ou seja, as EM passam a ser *constante de Planck* e *aparelho de Golgi*.

Basicamente a principal questão associada à categoria COISA é que debaixo desta designação estão "coisas" ontologicamente muito diferentes mas que a linguagem natural e em particular o português não distingue formalmente, como classe/membro, classe/subclasse e exemplo/classe. Para tentar produzir sobretudo critérios mais claros de anotação, sem querer forçar distinções que não estão lá (ou que os anotadores humanos têm dificuldade em fazer), redefinimos o seguinte elenco de tipos de COISA objecto de classificação no Segundo HAREM:

- OBJECTO que tem um nome individualizado, e que inclui desde animais (vivos, individuais) a planetas, passando por meios de locomoção tal como barcos ou foguetões e ursos de peluche.
- SUBSTANCIA nome de uma substância que, por ser massiva, não permite em geral distinções entre individuos ou espécies. É no entanto concreta e por isso não pode ser classificada como abstracção
- CLASSE passa pois a ter apenas as classes que são designadas por nomes próprios, tais como marcas ou modelos, assim como raças de animais ou programas de computador

- MEMBROCLASSE designa elementos que não têm nome individual mas que são designados pelo nome da classe a que pertencem, tal como *Ford* ou *iPod* em *o meu Ford* ou *o Ipod dela*, ou mesmo *Coca Cola* em *Quem me roubou a minha Coca Cola?* ou *Fox Terrier* em *Viste o meu Fox Terrier?*

A.4 Elenco de categorias do Segundo HAREM

Na tabela A.1 encontra-se o elenco de categorias, tipo e subtipos do HAREM clássico. Entre parênteses encontra-se: i) à frente das categorias o número de tipos; ii) à frente dos tipos, o número de subtipos.

A.5 Segundo HAREM: sintaxe

Uma EM é identificada pela etiqueta <EM> com atributos e terminada por </EM>.

Por exemplo,

<EM ID="xxx" CATEG="A" TIPO="B" SUBTIPO="C" COREL="corel" TIPOREL="tiporel">**Qualquer Coisa**</EM>

Os atributos possíveis

- têm de aparecer em maiúsculas;
- só podem ser ID, CATEG, TIPO, SUBTIPO, COREL, TIPOREL, TEMPO_REF, SENTIDO, VAL_NORM, VAL_DELTA, COMENT;
- o seu valor tem de ser incluído entre aspas, a seguir ao sinal de igual.

O único atributo obrigatório é o ID, que tem de ser uma combinação de apenas letras não acentuadas (maiúsculas ou minúsculas), algarismos, e os caracteres "-" e "_". A cada EM corresponde um ID único.

Os valores dos atributos COREL e TIPOREL estão descritos nas directivas do ReRelEM (Santos et al., 2008b).

Os valores dos atributos TEMPO_REF, SENTIDO, VAL_NORM e VAL_DELTA estão descritos nas directivas do TEMPO (Hagège et al., 2008).

Se várias possibilidades de identificar uma expressão correspondem a segmentações diferentes, usa-se <ALT>, separando as várias alternativas pelo caracter |.

Por exemplo, <ALT> alt1 | alt2 | alt3 </ALT>, em que alt1, alt2, alt3 são texto eventualmente marcado com <EM>. Para cada alternativa alt1, alt2, alt3 deve corresponder um ID diferente.

Para a tarefa de classificação, para todas as EM excepto as do TEMPO, uma EM no máximo terá a forma

<EM ID="xxx" CATEG="A" TIPO="B" SUBTIPO="C">**Entidade**</EM>.

Os valores possíveis para CATEG, TIPO e SUBTIPO:

- podem ser omitidos

- o TIPO só pode ser especificado se a CATEG também o for, e tem de pertencer a essa categoria
- o SUBTIPO só pode ser especificado se o TIPO também o for, e tem de pertencer a esse tipo
- o SUBTIPO só está definido para os tipos FISICO, HUMANO, VIRTUAL da CATEG LOCAL e para os TIPOS TEMPO_CALEND da CATEG TEMPO
- podem ser simples (veja-se a tabela na secção A.4), ou complexos
- valores complexos (correspondendo a vagueza) criam-se através da concatenação de vários valores através do caracter |.
- se um dado valor é omitido, usa-se o vazio
- a ordem dos valores complexos tem de ser idêntica nos três atributos, ou seja a ordem dos tipos tem de ser igual à ordem das categorias a que correspondem
- é necessário repetir a categoria se se quiser especificar alternativas entre tipos dessa mesma categoria
- é necessário repetir o tipo se se quiser especificar alternativas entre subtipos desse mesmo tipo

É possível incluir o que se quiser dentro do atributo COMENT, excepto caracteres especiais do XML como & < > ou aspas.

A.6 Lista de minúsculas

Nesta página, encontra-se a lista de palavras ou expressões em minúsculas que devem fazer parte da EM, no âmbito do Segundo HAREM. Relembramos que as regras para a identificação e classificação das EMs temporais se encontram separadamente definidas em Hagège et al. (2008).

Esta lista não é exaustiva, nem pretende descrever a língua portuguesa, tendo sido criada apenas com o objectivo de fornecer a todos os participantes no HAREM os mesmos critérios de identificação, neste caso, por extenso.

Note-se que as (sequências de) palavras listadas em seguida apenas devem ser tidas em consideração se surgirem imediatamente acompanhadas por uma outra palavra ou expressão iniciadas por maiúscula, as quais podem ser eventualmente antecedidas da preposição *de* e respectivas contracções.

Por exemplo:

<EM ID="Ex1" CATEG="PESSOA" TIPO="INDIVIDUAL">**presidente Lula**</EM>

presidente italiano <EM ID="Ex2" CATEG="PESSOA" TIPO="INDIVIDUAL">**Romano Prodi**</EM>

<EM ID="Ex3" CATEG="PESSOA" TIPO="CARGO">**duque de Bragança**</EM>

Os elementos das listas encontram-se organizados alfabeticamente por CATEG/TIPO (isto é, podem surgir no âmbito de uma EM classificada com os atributos a seguir especificados):
PESSOA/CARGO ou PESSOA/GRUPOCARGO ou PESSOA/INDIVIDUAL ou PESSOA/GRUPOIND

alta-comissária, altas-comissárias, alto-comissário, altos-comissários; bispo, bispos; chanceler, chanceleres; chefe, chefes; condessa, condessas, conde, condes; cônsul, cônsules, consulesa, consulesas; czar, czares, czarina, czarinas; dire(c)tor, dire(c)tora, dire(c)toras, dire(c)tores; dire(c)tor-geral, dire(c)tora-geral, dire(c)toras-gerais, dire(c)tores-gerais; duque, duques, duquesa, duquesas; embaixador, embaixadores, embaixatriz, embaixatrizes; infanta, infantas, infante, infantes; governador, governadora, governadoras, governadores; líder, líderes; ministra, ministras, ministro, ministros; padre, padres; patrão, patroa, patroas, patrões; porta-voz, porta-vozes; presidente, presidentes; primeira-ministra, primeiras-ministras, primeiro-ministro, primeiros-ministros; princesa, princesas, príncipe, príncipes; rabi(no), rabi(no)s; rainha, rainhas, rei, reis; reitor, reitora, reitoras, reitores; secretária de Estado, secretárias de Estado, secretário de Estado, secretários de Estado; secretária-geral, secretárias-gerais, secretário-geral, secretários-gerais; sultão, sultãos, sultões; visconde, viscondes, viscondessa, viscondessas

As palavras listadas podem ser, eventualmente, precedidas de "ex-", "vice", "co" ou "sub".

PESSOA/INDIVIDUAL ou PESSOA GRUPOIND

arquite(c)ta, arquite(c)tas, arquite(c)to, arquite(c)tos, avó, avós, avô, avôs; bispo, bispos; dom, dona, donas; doutor, doutora, doutoras, doutores; engenheira, engenheiras, engenheiro, engenheiros; irmã, irmão, irmãos, irmãs; madre, madres; mestre, mestres; padre, padres; professor, professora, professoras, professores; rabi(no), rabi(no)s; senhor, senhora, senhoras, senhores; seu; sir, sô; tia, tias, tio, tios; vovó, vovós, vovô, vovôs.

Todas as combinações de *senhor* seguido de cargo ou de título, tal como *senhor ministro* ou *senhor padre*, são também aceites. Estes casos podem ser antecedidos por *excelentíssimo* (ou respectiva abreviatura).

Aceitam-se igualmente as abreviaturas convencionalmente associadas aos elementos destas listas, sempre que estas existam.

ABSTRACCAO/ESTADO

doença; mal; sindroma; síndroma; síndrome

COISA/SUBSTANCIA

vitamina, vitaminas.

As palavras compreendidas nas listas são consideradas da mesma forma pela avaliação do Segundo HAREM quer estejam em maiúsculas ou minúsculas.

Agradecimentos

Agradecemos ao Marcirio Chaves, Nuno Cardoso, Caroline Hagège, Nuno Mamede, Bruno Martins e Mário Silva os comentários, sugestões e dúvidas formulados na discussão dos subtipos de LOCAL, cujo resultado final é, contudo, da nossa responsabilidade. Agradecemos também ao Nuno Cardoso e à Cristina Mota a correcção de muitos problemas em versões anteriores.

Tabela A.1: Elenco de categorias do Segundo HAREM

Categorias	Tipos	Subtipos
ABSTRACCAO (5)	DISCIPLINA	
	ESTADO	
	IDEIA	
	NOME	
	OUTRO	
ACONTECIMENTO (4)	EFEMERIDE	
	EVENTO	
	ORGANIZADO	
	OUTRO	
COISA (5)	CLASSE	
	MEMBROCLASSE	
	OBJECTO	
	SUBSTANCIA	
	OUTRO	
LOCAL (4)	FISICO (7)	ILHA, AGUACURSO, PLANETA, REGIAO, RELEVO, AGUAMASSA, OUTRO
	HUMANO (6)	RUA, PAIS, DIVISAO, REGIAO, CONSTRUCAO, OUTRO
	VIRTUAL (4)	COMSOCIAL, SITIO, OBRA, OUTRO
	OUTRO	
OBRA (4)	ARTE	
	PLANO	
	REPRODUZIDA	
	OUTRO	
ORGANIZACAO (4)	ADMINISTRACAO	
	EMPRESA	
	INSTITUICAO	
	OUTRO	
PESSOA (8)	CARGO	
	GRUPOCARGO	
	GRUPOIND	
	GRUPOMEMBRO	
	INDIVIDUAL	
	MEMBRO	
	POVO	
	OUTRO	
TEMPO (5)	DURACAO	
	FREQUENCIA	
	GENERICO	
	TEMPO_CALEND (4)	HORA, INTERVALO, DATA, OUTRO
OUTRO		
VALOR (4)	CLASSIFICACAO	
	MOEDA	
	QUANTIDADE	
	OUTRO	
OUTRO (1)		

Apêndice B

Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o Segundo HAREM

Caroline Hagège, Jorge Baptista e Nuno Mamede

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Apêndice B, p. 289–308.

Nota das editoras: Este apêndice inclui a versão das directivas do TEMPO disponibilizada no dia 18 de Fevereiro de 2008 e corrigida pela última vez no dia 13 de Abril de 2008. Notamos que o título original do documento usava o termo *HAREM II*, que foi substituído aquando da edição do livro por *Segundo HAREM*. Notamos ainda que a última versão da proposta inclui uma adenda que se encontra também reproduzida na secção B.6.

B.1 Preâmbulo

No âmbito do Reconhecimento de Entidades Mencionadas (REM), uma das tarefas de reconhecida importância consiste no reconhecimento de expressões temporais (entidades mencionadas da categoria TEMPO). Esta proposta tem por finalidade acrescentar uma nova faceta a esta tarefa já na próxima campanha do HAREM: a normalização das entidades mencionadas de tipo TEMPO.

Para levar a cabo esta tarefa, é necessário, por um lado, completar e enriquecer a actual definição da categoria TEMPO, tal como se encontra em Cardoso e Santos (2007). Por outro lado, a noção de entidade mencionada da categoria TEMPO tem de ser alargada à noção mais geral de expressão temporal.

B.2 Motivação da proposta

As motivações para esta proposta são as seguintes:

1) levar em conta os avanços e as direcções gerais de trabalhos recentes no âmbito do processamento de expressões temporais em textos (ver, por exemplo, TimeML em Saurí et al. (2006) e a campanha TempEval em Verhagen et al. (2007)).

   Mais precisamente, considera-se que a tarefa de REM de expressões temporais pode e deve ser vista como um primeiro passo para um mais rico processamento do sistema de referências temporais em textos. Neste sentido, tem-se a convicção que a comunidade interessada em REM em língua portuguesa poderia beneficiar bastante se passasse a considerar desde já os trabalhos de PLN efectuados no domínio do processamento das expressões que denotam e estruturam as referências temporais em textos; por outro lado, parece necessário assegurar que a segmentação/delimitação e a classificação de expressões temporais preconizadas pelo HAREM sejam compatíveis com linhas de investigação já existentes neste domínio e internacionalmente estabelecidas.

2) enriquecer a actual categorização proposta em Cardoso e Santos (2007).

   Se se aceitar a ideia de prolongar a identificação de EM temporais de modo a chegar-se à sua normalização, será necessário alargar o conceito actual de entidade temporal. Por exemplo, expressões temporais de tipo FREQUENCIA (i.e. repetição de eventos no tempo), cuja definição será explicitada mais adiante, não parecem estar contempladas nas directivas de anotação do primeiro HAREM.

B.3 Proposta

A proposta que se segue é largamente inspirada nos trabalhos recentes do TimeML (cf. http://www.timeml.org.site).

B.3.1 Categoria TEMPO

Na categoria TEMPO, considera-se uma grande parte das expressões que, semanticamente, denotam: (i) um **momento** no calendário (que pode ser concebido como um ponto ou como um intervalo); (ii) uma expressão de quantificação temporal que exprime uma **duração**; ou (iii) uma **repetição** de eventos no tempo; considera-se ainda (iv) o emprego **genérico** de algumas dessas expressões, geralmente associadas à noção de tempo. Nesta secção, apresenta-se, em primeiro lugar, uma definição geral das entidades mencionadas da categoria TEMPO. Definem-se, de seguida, os critérios que permitem determinar se uma expressão linguística pertence ou não à categoria TEMPO. Apresentam-se, ainda, os critérios que permitem delimitar uma expressão temporal complexa.

B.3.1.1 Definição da entidade de tipo TEMPO

B.3.1.1.1 Critérios para a identificação

Uma expressão temporal é qualquer expressão que responde ao critério 1 e a pelo menos um dos subcritérios de 2 **ou**, então, poderá ser uma expressão temporal genérica, que responde ao critério 3. As expressões temporais poderão **não conter algarismos ou palavras em maiúsculas**. Consideram-se necessários e suficientes para uma definição de expressão temporal os seguintes critérios, ordenados como acima se referiu:

critério 1 – uma expressão temporal em contexto pode responder adequadamente a uma das interrogativas "(<*prep*>) *quando?*", "(<*prep*>) *quanto tempo*?", "(<*haver*>) *quanto tempo?*" ou "*com que frequência?*".

critério 2 – uma expressão temporal contém pelo menos uma unidade lexical que corresponda a um dos seguintes tipos:

- **2-1** - uma data numérica (por exemplo, *29-10-2008*);
- **2-2** - uma unidade de medida temporal (*dia*, *mês*, *trimestre*, *ano*, *século*, etc.) ou um advérbio terminado em "*–mente*" derivado destas expressões (*diariamente*, *semanalmente*, *mensalmente*, etc.);
- **2-3** - um nome correspondente à designação de uma destas unidades de medida de tempo. Isto é: nome de meses (*Setembro*, *Dezembro*, etc.), nome de dia (*segunda-feira*, *domingo*, etc.);
- **2-4** - um nome de festividade, religiosa ou não (*Natal*, *Páscoa*, *Quaresma*, *Entrudo*); nomes de estações do ano (*Primavera*, *Inverno*); nomes de festividades, que podem incluir o nome *dia* (*dia de Santo António*, *dia de Nossa Senhora da Conceição*, *dia de São Valentim*, *dia dos namorados*, *no São Martinho*, etc.);
- **2-5** - alguns advérbios de tempo (simples, não derivados e semanticamente não ambíguos), tais como: *hoje*, *ontem*, *amanhã*, *outrora*; algumas locuções adverbiais de tempo (ou advérbios compostos), semanticamente não ambíguas, como, por

exemplo: *antes de ontem*, *depois de amanhã*; excluem-se da actual campanha de avaliação os seguintes advérbios simples, sintáctica ou semanticamente ambíguos, apesar de poderem representar expressões temporais: *agora*, *ainda*, *já*, *sempre*.

2-6 - um sintagma preposicional cujo núcleo seja uma das palavras *altura*, *tempo*, *momento*, *período*, *era*, etc., quando estas palavras forem determinadas por um demonstrativo (por exemplo: *nesse tempo*), ou especificados por uma relativa (por exemplo: ***na altura** em que ela adoeceu*), um possessivo (por exemplo: *durante a nossa era*) ou modificado por outro sintagma preposicional introduzido por *de* (por exemplo: *durante a era dos dinossauros*) ou então por um adjectivo capitalizado (por exemplo: *durante o período Barroco*, *Cretáceo*, etc.);

2-7 - Os complementos determinativos com a forma de *Num Ntmp* de nomes predicativos, que não respondem adequadamente ao critério (1) mas que são indubitavelmente EM a anotar (e.g. *uma viagem de 5 dias*); a preposição *de* deve ser incluída na EM;

2-8 - expressões de frequência, como as seguintes: *de vez em quando*, *às vezes*, *de quando em quando*, *frequentemente*, etc.

2-9 - expressões da forma *Prep* + <unidade de medida temporal> + *que* + verbo *vir* ou verbo *passar* (por exemplo, *no ano que passou*, *para o mês que vem*)

2-10 - expressões com os verbos *fazer* ou *haver* e <unidade de medida temporal> (e.g. *há três anos*, *faz duas semanas*).

Notas:

(a) Excluem-se, no critério 1, as expressões de tipo genérico como o emprego de o *inverno* em frases como *Adoro o inverno*, que serão retomadas de forma autónoma, no critério 3, abaixo.

(b) Excluem-se também, com o critério 1, as locuções que, embora contendo expressões do conjunto identificado no critério 2, não respondem adequadamente às interrogativas de tempo. Trata-se de locuções como *de dia para dia*, que encontramos em frases como *A situação agrava-se de dia para dia*, que funciona como um circunstancial de modo; repare-se na inaceitabilidade do par pergunta resposta: P:(*quando*, *com que frequência*, *em quanto tempo*) *é que a situação se agrava*? R: *de dia para dia*.

(c) Repare-se também que, para qualquer dos pontos 2-2 a 2-9 do critério 2, se pode fazer uma definição *em extensão* dos elementos em questão. Assegura-se, assim, o problema de intersubjectividade das anotações.

critério 3 – uma expressão temporal que contém uma unidade lexical do tipo das que foram definidas no critério 2 mas para a qual o critério 1 não se aplica.

Trata-se de expressões temporais genéricas como *o mês de Julho* em exemplos como *Adoro o mês de Julho* onde *o mês de Julho* não responde à pergunta *quando?* embora contenha elementos lexicais como os que foram definidos no critério 2.

Os critérios apresentados permitem contemplar, nos exemplos que se seguem, as expressões representadas em negrito:

- ***Em 2008*** *haverá mais confiança no futuro*, que, em contexto, responde à pergunta *quando?* (critério 1) e responde ao critério 2-1;
- *Chegou* ***no dia 5 de Junho de 2006*** que, em contexto, responde à pergunta *quando?* (critério 1) e responde aos critérios 2-1 e 2-3;
- *Viveu em Lisboa* ***entre 2000 e 2003*** que responde à pergunta *quando?* e responde ao critério 2-1;
- ***De um dia para o outro*** *o restaurante mudou completamente o seu menu* que, em contexto, responde à pergunta *quando?* (critério 1) e contém a palavra *dia* (critério 2-2);
- *Trabalhei* ***durante dois meses*** que responde à pergunta (*durante*) *quanto tempo?* (critério 1) e que contém a palavra *mês* (critério 2-2);
- *O padeiro vem* ***duas vezes por semana*** que, em contexto, responde à pergunta *com que frequência?* (critério 1) e contém a palavra *semana* (critério 2-2);
- *Vou visitar os meus pais* ***semanalmente***, que, em contexto, responde à pergunta *com que frequência?* (critério 1) e contém um advérbio terminado em *–mente* derivado de um nome de tempo *semana* (critério 2-2);
- *Chegou* ***no dia de Natal*** que, em contexto, responde à pergunta *quando?* (critério 1) e responde aos critérios 2-2 e 2-4;
- ***De hoje em diante*** *vou trabalhar* que, em contexto, responde à pergunta (*a partir de*) *quando?* (critério 1), e contém a palavra *hoje* (critério 2-5);
- *Esteve em Lisboa* ***há dois anos*** que, em contexto, responde à pergunta *há quanto tempo?* (critério 1) e que responde ao critério 2-2;
- *As vindimas fazem-se* ***nesta altura do ano***, que em contexto, responde à pergunta *quando?* (critério 1) e que responde ao critério 2-6;
- *Aconteceu* ***durante a era dos dinossauros*** que, em contexto, responde à pergunta *quando?* (critério 1) e que responde ao critério 2-6;
- *Vou à pesca* ***de vez em quando***, que em contexto, responde à pergunta *com que frequência?* e que responde ao critério 2-8;
- *Ficou doente* ***dois anos mais tarde***, que em contexto, responde à pergunta *quando?* e que responde ao critério 2-2.

Consideram-se também expressões genéricas como:

- ***A Primavera*** *é a mais bela estação do ano*, em virtude do critério 3 (expressões temporais genéricas).

Pontos importantes:

Embora obedeçam sempre ao critério 1 e possam às vezes obedecer a algum dos subcritérios de 2, excluem-se as orações subordinadas de tempo (por exemplo, `quando o meu pai chegar` que responde ao critério 1).

Do mesmo modo, as expressões temporais introduzidas pelas locuções prepositivas (ou preposiçoes compostas) *antes de* e *depois de* só serão marcadas como entidades mencionadas de tempo quando introduzirem um sintagma/grupo nominal (*antes de Domingo*, *depois de 2008*) ou ou sintagma adverbial (*antes de amanhã*, *depois de ontem*), em que ocorre um dos elementos lexicais especificados nos subcritérios de 2. Pelo contrário, quando estas locuções funcionam como conjunções e introduzem orações subordinadas temporais (*antes de o meu pai chegar*, *depois de ter acabado o trabalho*), todo o complemento adverbial será ignorado.

Também se excluem expressões fixas/idiomáticas com valor temporal, tais como *Quando as galinhas tiverem dentes* ou *daqui para frente*, apesar de responderem adequadamente ao critério 1. *Quando as galinhas tiverem dentes*, é, superficialmente, uma oração subordinada, conquanto seja uma expressão fixa, e, *daqui para frente* não obedece a nenhum dos subcritérios de 2.

Note-se que do ponto de vista linguístico, não há qualquer razão para excluir estas expressões (compostas, idiomáticas, subordinadas) do processamento de expressões temporais. A exclusão deste tipo de expressões, por ora, prende-se apenas com os limites que se pretendem delinear para a tarefa de reconhecimento de entidades mencionadas.

B.3.1.1.2 Critérios para a delimitação das EM da categoria TEMPO

A fim de se poder anotar de maneira unívoca as entidades da categoria `TEMPO`, convém ainda definir rigorosamente critérios sintáctica e semanticamente motivados que deverão ser seguidos a fim de se delimitar com precisão as fronteiras das entidades a anotar.

Nesta proposta, considera-se que a totalidade da expressão temporal deverá ser delimitada entre as balizas `<EM ID=... CATEG="TEMPO">` e `</EM>`, isto é, incluindo **a preposição** que a introduzir, no caso da expressão temporal ser um sintagma preposicional (e.g. *no ano passado*), **ou o determinante** no caso de ser um sintagma nominal (e.g. *dois dias depois*).

No caso de expressões complexas como *dois dias depois do Natal*, a questão que se coloca é a de se saber se esta expressão deverá ser considerada como uma só EM ou, então, segmentada em duas subexpressões *dois dias* + *depois do Natal* (obedecendo tanto a expressão mais longa como ambas as subexpressões aos critérios definitórios mencionados acima).

Os critérios adoptados para a segmentação são os definidos em Hagège e Tannier (2007) e que aqui foram reproduzidos:

> Uma expressão temporal complexa **deverá ser dividida** em unidades menores se se verificarem **simultaneamente** os critérios seguintes:
>
> 1 - cada expressão componente é sintacticamente válida quando combinada independentemente com o evento que modifica.
>
> 2 - cada expressão componente, combinada com o evento que modifica, está logicamente implicada na expressão complexa. Ou seja, cada combinação "evento + expressão_temporal_mínima" deve ser logicamente implicada pela combinação "evento + expressão_temporal_complexa". Em outras palavras, o valor de verdade de todas a combinações

"evento+expressão_temporal_mínima" deve poder ser deduzido do valor de verdade da combinação "evento + expressão_temporal_complexa" (ver exemplos abaixo para ilustração).

Exemplos:

Na frase:

Visitei-o ***dois dias nesta semana***,

a expressão *dois dias nesta semana* deverá ser considerada como constituída por duas entidades, pois cada subexpressão, *dois dias* (DURACAO) e *nesta semana* (DATA), pode combinar-se separadamente com o evento (*visitei-o*) e, se considerarmos que o valor de verdade da frase é verdadeiro, ambos os valores de verdade de *visitei-o dois dias* e de *visitei-o nesta semana* são verdadeiros.

Na frase:

Visitei-o ***dois dias depois***,

a expressão *dois dias depois* deverá ser considerada como uma só entidade (DATA). Com efeito, apesar de cada uma das subexpressões (*dois dias* (DURACAO) e *depois* (DATA)) poder combinar-se individualmente com o evento (*visitei-o*), verifica-se uma diferença de significado relativamente à interpretação da expressão complexa, mais concretamente, surge um novo adverbial, de DURACAO. O critério 2 não se verifica, pois: se se supuser que a asserção *visitei-o dois dias depois* é verdadeira, nada garante que o valor de verdade *visitei-o dois dias* seja verdadeiro.

Na frase:

Isso aconteceu ***dois dias depois do Natal***,

a expressão *dois dias depois do Natal* deve ser considerada como uma só entidade (DATA). Com efeito, ao considerar as duas sub expressões *Isso aconteceu dois dias* e *Isso aconteceu depois do Natal*, pode-se verificar que nem só a primeira subexpressão é duvidosa do ponto de vista da aceitabilidade sintáctica, mas também que, mesmo que fosse aceitável, o valor de verdade desta sub expressão (FREQUENCIA) não é logicamente implicado pelo valor de verdade da frase inicial (DATA).

NB: Casos ambíguos como:

Vimo-nos *<EM ID="..." CATEG="TEMPO">****dois dias depois do Natal****</EM>*,

Esta frase é ambígua e pode ser interpretada como:

1) *Vimo-nos do dia 27 de Dezembro*;

2) *Vimo-nos durante dois dias, a seguir ao 25 de Dezembro*.

Neste caso, embora a presença do segundo membro tenha tendência em 'forçar' a leitura complexa da expressão temporal (DATA), a ambiguidade será expressa na anotação (ver exemplos finais do ponto B.5).

Outros exemplos de delimitação de expressões temporais complexas

Emigrou há 23 anos depois do 25 de Abril.

Neste exemplo devem ser consideradas duas expressões temporais separadas (*há 23 anos* e *depois do 25 de Abril*. Com efeito, além das duas sub-expressões *emigrou há 23 anos* e *emigrou depois do 25 de Abril* serem sintacticamente válidas, se se considerar que o valor de verdade da frase é verdadeiro, também o valor de verdade das duas sub-expressões é verdadeiro.

Pelas mesmas razões, expressões como:

durante um fim de semana em Abril,
depois das férias do Natal em 2003

devem ser consideradas como pares de expressões temporais separadas:

durante um fim de semana (DATA) *em Abril* (DATA),
depois das férias do Natal (DATA) *em 2003* (DATA).

No entanto, uma expressão como *dois anos mais tarde* em:

Ficou doente ***dois anos mais tarde***

deve ser considerada como uma única expressão. Com efeito embora as sub-expressões *Ficou doente dois anos* e *Ficou doente mais tarde* sejam sintacticamente válidas, o valor de verdade de *Ficou doente dois anos* (DURACAO) não pode ser deduzido do valor de verdade da totalidade da expressão (DATA). (Nada se pode dizer sobre a duração da doença se se considerar que *Ficou doente dois anos mais tarde* for verdadeiro).

Qualquer expressão temporal deverá ser anotado por <EM ID=... CATEG="TEMPO"> e possuir o atributo obrigatório TIPO.

O atributo TIPO é o único atributo obrigatório do elemento EM de categoria TEMPO.

Os diferentes valores do atributo TIPO são:

TEMPO_CALEND (tempo calendário),

DURACAO (duração),

FREQUENCIA (frequência)

GENERICO (genérico).

Cada um deste tipo é detalhado nas secções seguintes.

B.3.2 TIPO = "TEMPO_CALEND"

As entidades de tipo TEMPO_CALEND são expressões que permitem inserir o predicado que elas modificam numa linha temporal (como um ponto ou um intervalo).

Correspondem aos seguintes subtipos:

- **datas** sejam elas **absolutas** (fórmulas contendo três campos ANO-MES-DIA, na qual até dois campos no máximo podem ser omitidos) ou **referenciais** (expressões temporais cuja resolução implica conhecer ou o momento da enunciação, ou outra data de um evento que funciona como referência).

- **intervalos** (expressões denotando uma duração no tempo e que têm explicitamente dois limites)

- **horas** (expressão temporais com valor de DATA mas com granularidade inferior à unidade *dia*).

B.3.2.1 SUBTIPO = "DATA"

As expressões deste subtipo podem representar *datas absolutas* ou *datas relativas* (que são referências). No primeiro caso, a expressão contém a informação necessária para localizar essa data num calendário (e.g. na expressão *em 23 de Outubro de 2007*, a informação está totalmente especificada em relação aos 3 campos; pelo contrário, nas expressões *em 23 de Outubro* e *em 2007*, a informação está parcialmente especificada em relação aos 3 campos).

Também são consideradas como abrangidas pelo subtipo DATA as expressões que exprimem *datas relativas*, isto é, para as quais é necessário determinar um ponto de referência para poder localizá-las na linha temporal (e.g. *dois dias mais tarde*, *na quinta-feira passada*, *ontem*, *na próxima terça feira*, etc.).

Apresentam-se de seguida alguns exemplos de expressões temporais do tipo TEMPO_CALEND e subtipo DATA:

- *Vou viajar* <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">***no dia 19 de Outubro de 2007***</EM>. Data absoluta completa (campos dia, mês e ano preenchidos);

- *Trabalhei em Londres* <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">***em 1998***</EM>. Data absoluta incompleta (campos dia e mês não preenchidos);

- *Vou a Lisboa* <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">***no próximo dia 22***</EM>. Data relativa;

- *Vai haver uma festa* <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">***na próxima terça feira***</EM>. Data relativa;

- *Fui a Lisboa* <EM ID="..." CATEG="TEMPO" TIPO="DATA">***na semana passada***</EM>. Data relativa;

- *A Joana nasceu* <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">***no Inverno***</EM> (exemplo do guia de anotação do Mini-HAREM). Data relativa;

- *Vou a Londres* <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">***no próximo Inverno***</EM> (exemplo do guia de anotação do Mini-HAREM). Data relativa.

NOTA IMPORTANTE: No último exemplo, existe uma diferença relativamente às directivas do guia de anotação (Cardoso e Santos, 2007). Com efeito, no Mini-HAREM, a expressão *Inverno* era considerada como uma duração (PERIODO segundo a terminologia utilizada no guia). É importante sublinhar que se considera que entidades de subtipo DATA não implicam um valor aspectual pontual mas que podem ser representados por intervalos com uma granularidade variável.

Assim, a expressão *no próximo inverno*, referida acima, é de subtipo DATA, pois mesmo que seja uma expressão que subentende uma certa duração, a ida a Londres mencionada na frase pode ser ancorada num calendário mediante o conhecimento da data da enunciação, que permite resolver se se trata do Inverno do ano de 2007/2008 ou de outro inverno

qualquer (se a data de enunciação for o presente, tratar-se-á do tempo calendário entre 21 de Dezembro de 2007 e 20 de Março de 2008).

B.3.2.2 Expressões de datas relativas: dois tipos de referências considerados

Fez-se no ponto anterior a distinção entre expressões temporais correspondente a uma data absoluta (isto é, que permite, sem recurso a nenhum contexto, localizar na linha do tempo o evento ao qual a data está associada) e expressões temporais relativas que são referenciais.

Explicitam-se agora os dois tipos de expressões temporais relativas consideradas: expressões temporais relativas que fazem referência ao tempo da enunciação e expressões temporais relativas cuja referência está introduzida no discurso.

Um exemplo típico desta distinção pode ser dado através do exemplo seguinte:

*Chegou **ontem***

*Chegou **no dia anterior**.*

Nestes dois exemplos está-se na presença de expressões temporais que podem permitir localizar no calendário o evento associado (`TIPO="TEMPO_CALEND"`). Na medida em que não se trata de um intervalo de tempo com limites explícitos, nem da expressão de uma hora, pode-se associar a estas expressões o valor `SUBTIPO="DATA"`. Mas não se trata aqui de expressões correspondente a uma data absoluta, mas sim a datas relativas. Como datas relativas, estas expressões são referenciais. Com efeito, será necessário se se quiser localizar o evento *Chegou* na linha do tempo, ter em conta uma referência.

No primeiro exemplo, esta referência é o momento da enunciação.

Com efeito, se a asserção *Chegou ontem* for produzida no dia 4/12/2007, pode-se inferir que o evento *Chegou* ocorreu no dia 3/12/2007. O tempo no qual ocorre o evento neste exemplo é função do tempo do momento da enunciação (tempo_enunciação – 1 dia).

Fala-se neste caso de expressão temporal referencial relativa ao momento da enunciação.

No segundo exemplo, embora também se trate de uma data referencial, a referência não é o momento da enunciação. A localização temporal de chegou é independente do momento em que for produzida a asserção. Neste caso, a referência é outra data/evento que aparece no contexto textual ou discursivo.

Por exemplo:

> *O barco só devia chegar ao porto no dia 25 de Novembro, no entanto chegou no dia anterior*

Vê-se, contextualizando o exemplo, que a referência da expressão *no dia anterior* é o evento da chegada do barco ao porto que ocorreu no dia 25/11. Conhecendo esta referência pode-se então deduzir que o evento *chegou* ocorreu no dia 24/11. Assim, neste caso está-se em presença de uma expressão referencial textual.

Esta distinção entre data absoluta, data referencial relativa ao momento de enunciação e data relativa com referência textual é formalizada, na próxima sub-secção, através do atributo `TEMPO_REF`.

B.3.2.3 Atributo TEMPO_REF

O atributo TEMPO_REF, diz apenas respeito às expressões temporais de TIPO="TEMPO_CALEND" SUBTIPO="DATA" (ver sub-secção B.3.2.1).

No caso de datas absolutas, o valor do atributo TEMPO_REF é ABSOLUTO.

No caso de datas referenciais, conforme o tipo da referência (ver sub-secção B.3.2.2) o valor do atributo TEMPO_REF é respectivamente ENUNCIACAO ou TEXTUAL.

Por exemplo, na frase *Partiu* ***no dia 3 de Novembro de 2007***, a expressão temporal de tipo DATA *no dia 3 de Novembro de 2007* permite determinar sem ambiguidade que o evento da partida ocorreu no intervalo entre 3/11/2007 00:00 e 3/11/2007 24:00.

No caso das expressões com valor temporal relativo (e.g. *dois dias depois*, *na próxima sexta-feira*), a expressão temporal por si só, não é suficiente para poder situar o evento num calendário. No primeiro exemplo trata-se de um caso de referência textual (TEMPO_REF="TEXTUAL"), no segundo caso de uma referência ao momento da enunciação (TEMPO_REF="ENUNCIACAO").

Os seguintes exemplos ilustram o uso do atributo TEMPO_REF e dos seus possíveis valores:

- *Nasceu* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO">* ***a 3 de Janeiro de 1986*** *</EM>*.
- *Nasceu* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">* ***dois dias depois do Natal*** *</EM>*.

No exemplo acima, note-se primeiro que, conforme os critérios de segmentação que definimos (ver ponto B.3.1.1.2) , esta expressão complexa tem de ser considerada como um todo. Para poder situar no calendário o evento (o nascimento) que a expressão localiza temporalmente, é necessário conhecer um tempo de referência correspondente ao *Natal*. Sendo esta referência textual, o valor de TEMPO_REF é TEXTUAL

- *Nasceu* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ENUNCIACAO">* ***na sexta-feira passada*** *</EM>*.

Para poder calendarizar o evento *Nasceu* do exemplo acima, é necessário conhecer a data na qual foi enunciada a frase. A partir desta data de enunciação é que se poderá calcular o dia que corresponde à sexta-feira anterior à esta data de enunciação. O valor de TEMPO_REF neste caso é ENUNCIACAO.

- *Nasceu* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">* ***dois dias depois*** *</EM>*.

No exemplo acima, para poder localizar o evento no tempo, é necessário conhecer um tempo de referência que terá sido introduzido previamente no discurso. Neste caso, o valor de TEMPO_REF é TEXTUAL (mesmo que a referência não apareça explicitamente na frase).
NOTA IMPORTANTE:

Em caso de expressões como ***dois dias*** *depois de o meu pai chegar*, em conformidade com os critérios explicitados nos pontos anteriores, só a subexpressão *dois dias* será anotada entre as balizas TEMPO. Esta expressão é de tipo TEMPO_CALEND SUBTIPO="DATA" e o valor de TEMPO_REF é TEXTUAL, sendo o tempo de referência neste caso a data a que se refere a oração subordinada temporal (i.e. *depois de o meu pai chegar*).

Outros exemplos:

- *O Pedro chegou* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ENUNCIACAO">**ontem**</EM>*.
- *O Pedro partiu* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**na semana seguinte**</EM>*.
- *O Pedro chegou a Paris* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ENUNCIACAO">**no domingo**</EM>*. *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**Dois dias depois**</EM> foi para Grenoble*.

B.3.2.4 Atributos SENTIDO e VAL_DELTA

No caso de expressões temporais referenciais (`TEMPO_REF` tem o valor `TEXTUAL` ou `ENUNCIACAO`, e dois novos atributos podem estar presentes na anotação: o atributo `SENTIDO` e o atributo `VAL_DELTA`.

O atributo `SENTIDO` permite dar uma informação complementar que tem por finalidade a normalização de expressões temporais referenciais. Mais precisamente, vai indicar se o seu valor temporal se situa cronologicamente *antes* ou *depois* do *tempo de referência*. Os possíveis valores do atributo `SENTIDO` são, pois, `ANTERIOR`, `POSTERIOR`, `SIMULT`, `ANTERIOR_OU_SIMULT`, `POSTERIOR_OU_SIMULT`. Estes valores correspondem respectivamente aos casos em que o valor temporal denotado pela expressão de data relativa referencial se situa antes, depois, ao mesmo tempo, antes ou ao mesmo tempo, ao mesmo tempo ao depois do valor temporal da referência.

Quanto ao atributo `VAL_DELTA`, ele tem por valor uma expressão que indica a distância temporal entre o tempo do evento denotado pela expressão temporal e o momento de referência (seja este o tempo da enunciação ou outro) quando esta distância temporal aparece explicitamente no texto. No caso desta distância temporal não ser explícita, o valor de `VAL_DELTA` é omitido.

No caso da distância temporal ser explícita, o valor de `VAL_DELTA` corresponde ao valor temporal que se deve incrementar ou subtrair a partir do tempo de referência para obter o valor temporal do evento associado à expressão temporal a anotar.

Os valores possíveis de `VAL_DELTA` são representados da maneira seguinte:

A<digitos>**M**<digitos>**S**<digitos>**D**<digitos>**H**<digitos>**M**<digitos>**S**<digitos>

Onde:

- as letras `A`, `M`, `S`, `D`, `H`, `M`, `S` são constantes que devem aparecer nesta ordem e que correspondem respectivamente ao valores de Anos, Meses, Semanas, Dias, Horas, Minutos e Segundos.
- os <digitos> à direita das letras constantes correspondem ao número de Anos, Meses, Semanas, Dias, Horas, Minutos e Segundos que se devem adicionar ou diminuir à data de referência para obter o valor temporal da expressão anotada.

Por exemplo:

- *Apareceu* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="POSTERIOR" VAL_DELTA="A0M0S2D0H0M0S0">**duas semanas**</EM> depois da festa*.

Para proceder à normalização da expressão *duas semanas* é necessário conhecer um tempo de referência (TEMPO_REF="TEXTUAL") que corresponde aqui à data da *festa*. O valor do atributo SENTIDO é POSTERIOR na medida em que a data do evento (*apareceu*) teve lugar após esta data de referência (festa); o valor de VAL_DELTA indica que esta distância temporal entre a data do evento e a data de referência corresponde a duas semanas: os valores de todos os campos são 0 excepto para o campo S (semana), em que se indica 2.

Outros exemplos:

- *Veio* <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ENUNCIACAO" SENTIDO="ANTERIOR" VAL_DELTA="A0M0S0D1H0M0S0">***ontem***</EM>;
- *O Pedro partiu* <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="POSTERIOR" VAL_DELTA="A0M0S1D0H0M0S0">***na semana seguinte*** </EM>;
- *Nasceu* <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="POSTERIOR" VAL_DELTA="A0M0S0D2H0M0S0">***dois dias depois do Natal***</EM>;

Nota-se que em falta de informação explícita no texto, VAL_DELTA poderá ser omitido ter um valor indefinido que será representado pela cadeia vazia "".

Por exemplo:

- *O Pedro chegou* <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="POSTERIOR">***depois***</EM>.

No exemplo acima, trata-se de uma expressão de tipo DATA referencial. O ponto de referência é um evento ou uma data que não está presente na própria frase mas que foi introduzido anteriormente no discurso (TEMPO_REF="TEXTUAL"). O evento da chegada do Pedro ocorre a seguir a este momento de referência (SENTIDO="POSTERIOR"). No entanto, não se tem explicitamente a distância temporal entre o evento da chegada do Pedro e o momento de referência. Por esta razão, o valor de VAL_DELTA será reduzido à cadeia vazia (VAL_DELTA=""). Poderá também ser omitido.

B.3.2.5 SUBTIPO = "HORA"

Trata-se de expressão temporais com valor de DATA mas com granularidade inferior à unidade dia. Do ponto de vista da sua definição, mantem-se a proposta do primeiro HAREM.

Exemplo:

- *O Pedro está disponível* <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="HORA" VAL_NORM="+——T15–E–LMA">***às 15:00*** </EM>.

B.3.2.6 SUBTIPO = "INTERVALO"

Corresponde a uma expressão complexa, isto é, composta por duas expressões temporais elementares/simples, mas que, semanticamente, forma um única entidade mencionada e que tem **explicitamente** dois *limites temporais* (*limite inicial* e limite final).

Exemplos:

- *Trabalhei em Londres* <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO">***entre 2000 e 2003***</EM>.

- *Trabalhei em Londres* <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO">***de Outubro a Dezembro de 2007***</EM>.
- *Vai demorar* <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO">***de 3 a 6 meses***</EM>.

NOTA 1: Embora certas expressões do SUBTIPO="DATA" (ver B.3.2.1) ou de TIPO="DURACAO" (ver B.3.3 a seguir) definam implicitamente um intervalo de tempo (e.g. *todo o inverno* define implicitamente um intervalo entre 21 de Dezembro e 20 de Março), no tipo INTERVALO, os limites **têm de ser explícitos** para se poder considerar o subtipo INTERVALO.

B.3.3 TIPO = "DURACAO"

Corresponde a uma expressão TEMPO que refere uma duração de tempo contínuo. Mas, ao contrário das datas, trata-se de expressões que não exprimem propriamente a localização de um evento (calendarização do evento), mas sim quantificação temporal, sendo constituídas por nomes de unidades de medida de tempo e determinantes com função de quantificadores (e.g.. numerais). Podem, por vezes, ser introduzidas, facultativamente, pela preposição *durante* e respondem adequadamente à interrogativa (*prep*) *quanto tempo?*.

Exemplos:

- *Fiquei* <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">***dois meses***</EM> *em Lisboa*.
- *O urso fica* <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">***todo o inverno***</EM> *na toca*.

B.3.4 TIPO = "FREQUENCIA"

O tipo FREQUENCIA corresponde às expressões TEMPO que exprimem uma repetição no tempo. Estas expressões respondem às interrogativas do tipo *com que frequência?*.

Exemplos:

- *Vou ver os meus pais* <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">***diariamente*** </EM>.
- *Vou ver os meus pais* <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">***todos os dias***</EM>;
- *Vou ver os meus pais* <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">***duas vezes por semana***</EM>;
- *Vou ver os meus pais* <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">***dia sim dia não***</EM>.

B.3.5 TIPO = "GENERICO"

Trata-se de expressões TEMPO que não se referem um data específica embora a expressão linguística seja composta por unidades lexicais que denotam elementos temporais. Estas expressões obedecem ao critério 3 definido em B.3.1.1.

Exemplos:

- *Adoro* <EM ID="..." CATEG="TEMPO" TIPO="GENERICO">***o verão***</EM>.
- <EM ID="..." CATEG="TEMPO" TIPO="GENERICO">***Fevereiro***</EM>*é o mês mais curto do ano*.

B.3.6 Atributo VAL_NORM

O atributo `VAL_NORM` será apenas atribuído à algumas entidades `TEMPO`. Pretende ser um primeiro passo para a normalização de expressões temporais. Este atributo vai estar presente exclusivamente para as seguintes entidades temporais: `<EM ID=... CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" /> <EM ID=... CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="HORA" /> <EM ID=... CATEG="TEMPO" TIPO="DURACAO" />`

B.3.6.1 Atributo VAL_NORM para expressões de subtipo DATA absoluta

Recorde-se que as expressões de subtipo `DATA` têm um atributo `TEMPO_REF` que poderá ter um dos seguintes valores: `ABSOLUTO`. Só é neste caso que se vai calcular a data absoluta correspondente a expressões temporais. (por outras palavras, não se vai tentar normalizar as datas referenciais no âmbito desta proposta).

O valor de `VAL_NORM` obedece ao seguinte formato:

`<Era><Ano><Mes><Dia>`**`T`**`<Hora><Minuto>`**`E`**`<ESTACAO>`**`LM`**`<limite_aberto>`

Onde:

<**Era**> corresponde a 1 carácter que é + ou – conforme a data seja depois ou antes da nossa era;

<**Ano**> corresponde a 4 caracteres de tipo dígito que representam o valor do ano ou então a subsequência "—-";

<**Mes**> corresponde a 2 caracteres de tipo dígito que representam o valor do mês ou então a subsequência "–";

<**Dia**> corresponde a 2 caracteres de tipo dígito que representam o valor do dia ou então a subsequência "–";

<**Hora**> corresponde a 2 caracteres de tipo dígito que representam o valor da hora ou então a subsequência "–";

<**Minuto**> corresponde a 2 caracteres de tipo dígito que representam o valor dos minutos ou então a subsequência "–";

<**ESTACAO**> corresponde a duas letras capitalizadas correspondente às estações do ano. `IN` para Inverno, `PR` para Primavera, `VE` para Verão e `OU` para Outono. No caso da data absoluta não ser expressa em termos de estação do ano, este campo terá por valor a subsequência "–";

<**limite_aberto**> indica se a expressão normalizada de data absoluta introduz um intervalo de tempo com limite anterior ou limite posterior não determinado (em aberto). Os valores respectivos são `A` (no caso de limite anterior em aberto; este caso a expressão temporal corresponde ao limite posterior); no caso de limite posterior em aberto '`P`; neste caso, a expressão temporal corresponde ao limite anterior) e "-" quando a data absoluta não corresponde a um intervalo com um dos limites aberto.

Exemplos:

- *Nasceu* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+19860103T—E–LM-">**a 3 de Janeiro de1986**</EM>*;
- *A Lia nasceu* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+1996—T—EPRLM-">**na Primavera de 1996**</EM>*;
- *A Inês vai à escola* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+200709–T—E–LMP">**desde Setembro de 2007**</EM>*.

NOTA: Se a expressão temporal data absoluta referir uma data anterior ao ano 9999 a.C. ou posterior ao ano 9999 d.C. , então o valor de VAL_NORM não é calculado atribuindo-se-lhe convencionalmente os valores:
-99999999T9999E-LM- ou +99999999T9999E-LM-

B.3.6.2 Atributo VAL_NORM para expressões de tipo HORA

No caso de expressões do subtipo HORA, também é utilizado o formato:

<Era><Ano><Mes><Dia>**T**<Hora><Minuto>**E**<ESTACAO>**LM**<limite_aberto>

que foi empregue para as expressões de subtipo DATA, TEMPO_REF="ABSOLUTO". Neste caso, no entanto, os campos correspondente a <ERA><Ano><Mes><Dia> correspondem necessariamente à subsequência "+———" e o campo <ESTACAO> corresponde a "–". O campo <limite_aberto> pode corresponder a D ou E se a expressão de tipo HORA corresponder a uma expressão que introduz um intervalo aberto anterior ou posterior.

Exemplos:

- *A reunião durou* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="HORA" VAL_NORM="+———T0220E–LM-">**2 horas e 20 minutos**</EM>*.
- *Está disponível* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="HORA" VAL_NORM="+———T15–E–LMA">**antes das 3:00 da tarde**</EM>*.
- *O discurso de* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="HORA" VAL_NORM="+———T–50E–LM-">**50 minutos**</EM> foi muito maçador*.

B.3.6.3 Atributo VAL_NORM para expressões de tipo DURACAO

Finalmente, para expressões de tipo DURACAO, o valor do atributo VAL_NORM corresponde ao valor utilizado para VAL_DELTA e exprime uma distância temporal.

Relembra-se aqui o formato deste valor:

A<digitos>**M**<digitos>**S**<digitos>**D**<digitos>**H**<digitos>**M**<digitos>**S**<digitos>

Onde:

- as letras A, M, S, D, H, M, S são constantes que devem aparecer nesta ordem e que correspondem respectivamente ao valores de Anos, Meses, Semanas, Dias, Horas, Minutos e Segundos;

- os <digitos> à direita das letras constantes correspondem ao número de Anos, Meses, Semanas, Dias, Horas, Minutos e Segundos que se devem adicionar ou diminuir à data de referência para obter o valor temporal da expressão anotada.

Exemplos:

- *Fiquei* <EM ID="..." CATEG="TEMPO" TIPO="DURACAO" VAL_NORM="A0M2S0D0H0M0S0">***dois meses***</EM> *em Lisboa*.

Finalmente, para certas expressões de tipo DURACAO, que apresentam uma forma quantificação indefinida, tais como *vários anos*, *durante muitos dias*, *durante séculos*, etc. e na presença de quantificadores como *todo*, *inteiro*, etc. em *todo o inverno*, *o ano todo*, etc. não se especifica o atributo VAL_NORM.

- *Fiquei* <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">***durante muitos meses***</EM> *em Lisboa*.
- *O urso fica* <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">***todo o inverno***</EM> *na toca*.

B.4 Resumo das principais modificações

A principais modificações feitas em relação à proposta do Mini-HAREM são as seguintes:

- O tipo CÍCLICO, definido no guia de anotação do Mini-HAREM, desaparece;
- Expressões como *Natal* e *Páscoa* são consideradas como TIPO="DATA" mesmo que sejam cíclicas, na medida em que integram uma entidade que poderá ser ancorada num calendário;
- Propõe-se também não anotar entidades mencionadas que correspondam a períodos implícitos tais como os exemplificados na página 10 do guia de anotação:

  Depois da IBM fui trabalhar para a Sun
  Depois de trabalhar na IBM em 1993, fui trabalhar para a Sun

  Estes exemplos contêm uma elipse e uma anotação como a que é proposta no guia de anotação faz com que, no primeiro exemplo, *IBM* seja considerada como uma entidade TEMPO enquanto que, no segundo, seja considerada como uma organização, o que parece pouco coerente e não corresponde aos objectivos gerais da tarefa de REM (ou pelo menos desta subtarefa de reconhecimento de entidades temporais).

B.5 Alguns exemplos de anotação

Nesta secção ilustram-se com alguns exemplos comentados a proposta de anotação das entidades TEMPO apresentada neste documento. Apresentam-se alguns exemplos que parecem colocar algumas dificuldades.

Aconteceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="-99999999T9999E99LM-">***na era dos dinossauros***</EM>

A expressão *na era dos dinossauros* deve ser considerada como uma entidade a anotar (critério 1 e 2-6 da caracterização. Trata-se de uma expressão de subtipo DATA cuja granularidade é de centenas de bilhões de anos. (a era dos dinossauros começou há cerca de 240 bilhões de anos e eles povoaram a terra cerca de 165 bilhões de anos). Como se ultrapassa o ano -9999 o valor de VAL_NORM corresponde a "-99999999T9999E99LM-"> (ver nota no ponto B.3.6.1)

<EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">***De vez em quando***</EM> *vou passear na montanha*.

Trata-se de uma expressão de tipo frequência (responde à pergunta *com que frequência?*). que deverá ser anotada, em virtude do critério definitório 2-9 do ponto B.3.1.1.1)

Fui ver o meu pai <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">***na semana seguinte ao Natal passado***</EM>

Trata-se de um exemplo um pouco artificial mas que vai ilustrar como considerar a referência em caso de expressões temporais encaixadas.

Primeiro, no que diz respeito à delimitação da expressão temporal tem de se considerar a totalidade da expressão (ver ponto B.3.1.1.2). Trata-se obviamente de uma expressão de data referencial. Para determinar qual é a referência, considera-se o primeiro nível de encaixe da expressão (i.e. *na semana seguinte a X* (sendo *X* a referência, mesmo que esta referência também seja ela própria de tipo referencial).

Vimo-nos ***dois dias depois do Natal***.

<ALT> <EM ID="1" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="POSTERIOR" VAL_DELTA="A0M0S0D2H0M0S0">***dois dias depois do Natal***</EM>

|

<EM ID="2" CATEG="TEMPO" TIPO="DURACAO" VAL_NORM="A0M0S0D2H0M0S0">***dois dias***</EM>

<EM ID="3" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="POSTERIOR">***depois do Natal***</EM> </ALT>

No exemplo acima, tem-se uma ambiguidade que permite duas interpretações:

Vimo-nos no dia 27 de Dezembro

Vimo-nos durante dois dias depois do dia 25 de Dezembro

Esta ambiguidade está representada através do elemento ALT que introduz as várias alternativas, e do elemento '|' , que as separa.

<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">***Naquela semana***</EM> *vimo nos* <EM ID="..." CATEG="TEMPO" TIPO="DURACAO" VAL_NORM="A0M0S0D3H0M0S0">***três manhãs***</EM>.

O exemplo acima tem semelhanças com uma das interpretações do exemplo anterior. Neste caso há dois elementos de categoria `TEMPO` que deverão ser anotados. Um é do subtipo `DATA` relativa com referência textual, e com atributo `SENTIDO` e `VAL_DELTA` indefinidos. O segundo é do tipo `DURACAO` (responde à pergunta *quanto tempo?*).

Nota-se que esta representação não distingue partes de dias como *manhã* ou *tarde*. O valor normalizado de *três manhãs* vai corresponder aqui ao valor normalizado de *três dias*.

> *Vou a Madrid* *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ENUNCIACAO" SENTIDO="POSTERIOR" VAL_DELTA="A1M0S0D0H0M0S0">**para o ano que vem**</EM>*

Este exemplo permite relembrar qual a delimitação duma expressão temporal de tipo PREP+unidade_medida_temporal+que+verbo vir (ou verbo passar).

Trata-se de uma expressão temporal relativa, cuja referência é o tempo da enunciação. A distância temporal entre a referência e a expressão corresponde a 1 ano.

B.6 Adenda

ALGUMAS PRECISÕES SOBRE O TEMPO

O estatuto da vírgula nas expressões de tipo `TEMPO`

A vírgula, poderá ou não ser um separador nas expressões temporais.

Caso 1) a vírgula faz parte integrante da expressão temporal

Por exemplo:

> *Quarta-feira, 13 de Fevereiro de 2008*
> *Fevereiro 13, 2008*

Nestes exemplos a vírgula está integrada dentro da expressão:

> *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" SENTIDO="+20080213T—E–LM-" VAL_DELTA="">**Quarta-feira, 13 de Fevereiro de 2008**</EM>*
>
> *<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" SENTIDO="+20080213T—E–LM-" VAL_DELTA="">**Fevereiro 13, 2008**</EM>*

Caso 2) assinala a coordenação de expressões temporais.

Por exemplo:

> *Nos dias 25, 26 e 27 de Janeiro*

Neste exemplo, há 3 expressões temporais a considerar:

<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" SENTIDO="+—0125T—E–LM-" VAL_DELTA=",">***Nos dias 25***</EM> <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" SENTIDO="+—0126T—E–LM-" VAL_DELTA="e">***26***</EM>

<EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" SENTIDO="+—0127T—E–LM-" VAL_DELTA="">***27 de Janeiro***</EM>.

Nota-se que, na normalização, «*Nos dias 25*» e «*26*» foram normalizados como «*25 de Janeiro*» e «*26 de Janeiro*» respectivamente.

Fracções de unidades de medidas temporais na normalização

Propomos, para a normalização de fracções de unidades de medidas temporais, considerar as seguintes equivalências:

1 milénio = 1000 anos
1 século = 100 anos
1 ano = 12 meses
1 mês = 30 dias (aproximação)
1 quinzena = 14 dias
1 semana = 7 dias
1 dia = 24 horas
1 h = 60 minutos
1 minuto = 60 segundos (utilizado nas durações)

Assim, se encontramos uma expressão como: "*Durante meio ano*" esta expressão deverá ser normalizada da mesma forma que "*durante 6 meses*", isto é `VAL_NORM="A0M6S0D0H0M0S0"`

Propomos que, na normalização das expressões temporais, só se levem em conta valores inteiros.

Assim, quando a divisão da unidade de medida temporal convertida tem um resto diferente de 0, consideramos apenas o valor inteiro ignorando o resto da divisão.

Isto vai acontecer nas fracções de segundos (para durações) e nas fracções de minutos para horas. Poderá também acontecer em casos como "*durante 2/3 da semana*" (se este caso se produzir) onde a divisão (2*7)/3 tem um resultado de 4, ignorando o resto de 0.666.

Por exemplo:

Durante 1/5 do ano

terá o valor de `VAL_NORM="A0M2S0D0H0M0S0"` que corresponde a $12/5 \sim 2$ meses

Durante 2/3 do século

terá o valor de `VAL_NORM="A66M0S0D0H0M0S0"` que corresponde a $(2*100)/3 \sim 66$ anos

Agradecimentos

Agradece-se à Diana Santos e à Cláudia Freitas pela releitura da versão inicial e pelas muitas sugestões que permitiram melhorar esta proposta.

Apêndice C

ReRelEM - Reconhecimento de Relações entre Entidades Mencionadas. Segundo HAREM: proposta de nova pista

Cláudia Freitas, Diana Santos, Paula Carvalho e Hugo Gonçalo Oliveira

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Apêndice C, p. 309–317.

Nota das editoras: Este apêndice reproduz a versão 2.2 das directivas para reconhecimento de relações semânticas entre EM, pista do ReRelEM, que foi disponibilizada e actualizada pela última vez no dia 10 de Abril de 2008. Por uma questão de uniformização, colocámos os agradecimentos separadamente numa secção final. O assunto iniciado pelas palavras "A DISCUTIR AINDA", com relação à vagueza, foi apenas resolvido durante a própria anotação da colecção dourada, visto que nenhum participante do ReRelEM se exprimiu sobre essa questão. A decisão tomada está documentada no capítulo 4, secção 4.2.3.

Neste documento preliminar descrevemos uma tarefa piloto que propomos para o Segundo HAREM, e que pretende identificar se existem relações entre as diversas EM de um texto. A inspiração para esta proposta tem várias fontes:

- a existência da tarefa de co-referência no MUC e de "entity-link-tracking" no ACE
- a existência de trabalho sólido sobre a co-referência em português (Collovini et al., 2007)
- a emergência da área "extracção de relações" na extracção de informação (Chu-Carroll e Prager, 2007; Culotta e Sorensen, 2004; Roth e tau Yih, 2004; Zhao e Grishman, 2005)

Por estas razões, e embora tivéssemos de limitar a tarefa para ser realizável, pensamos ser o momento certo para tentar desafiar sistemas de processamento do português para esta tarefa, no âmbito do Segundo HAREM, à imagem da proposta do `TEMPO`.

Como a definição de quais as relações relevantes entre EM é um trabalho altamente subjetivo, e como trata-se de uma tarefa-piloto, prevemos inicialmente a identificação de apenas quatro (ou seis) tipos de relação entre EM, detalhados a seguir:

- identidade (sem `TIPOREL` ou `TIPOREL="ident"`)
- inclusão (`TIPOREL="inclui"` ou `TIPOREL="incluido"`)
- ocorre_em (`TIPOREL="ocorre_em"` ou `TIPOREL="sede_de"`)
- outra (`TIPOREL="outra"`)

Esta proposta foi obtida após marcação exaustiva de alguns textos e discussão alargada sobre a capacidade de consenso generalizado sobre o elenco (mais extenso) das relações originalmente propostas ao grupo mencionado acima.

C.1 Directivas para anotação das relações entre EM

Nos exemplos a seguir, só estarão marcadas as EM que estão ligadas por alguma relação. Para facilitar a leitura, omitimos também a anotação das categorias, tipos e subtipos das EM.

C.1.1 Regras gerais de integração da pista no HAREM

Cada EM recebe uma identificação única (ID), obrigatória no Segundo HAREM.

Para cada EM que apresentar uma relação de co-referência (ou outra relação) com uma outra EM já anotada, deve-se indicar, no campo `COREL`, a ID da EM relacionada e, em seguida, o tipo de relação entre as EM, no campo `TIPOREL`.

> Um dos telescópios já está pronto e em funcionamento no <EM ID="L111">**Havaí**</EM>, <EM ID="L165" **COREL="L111" TIPOREL="inclui"**>**EUA**</EM>.

É importante lembrar que só consideramos as relações **entre EM**, isto é relações que envolvem uma EM e pronomes, ou outros tipos de sintagmas nominais, por exemplo, não devem ser anotadas. Ou seja, em

> Batizado de <EM ID="AB60">**Santanaraptor placidus**</EM>, **o fóssil** é o único a ser encontrado no país com restos de tecido mole, como fibras musculares, vasos sanguíneos e pele. Para os paleontólogos, achar esse tipo de evidência equivale a acertar na loteria. "é como se **o dinossauro** tivesse sido enterrado ontem? (...) O nome é uma alusão à região onde **ele** viveu (...)

nem "o fóssil", nem "o dinossauro" nem "ele" devem receber qualquer marca.

Além disso, apenas consideramos relações entre EM **em um mesmo texto**, ou seja, a pista não se refere a relações entre textos diferentes da coleção.

É possível, por outro lado, que uma `COREL` tenha um ID de uma entidade que ainda não foi mencionada no texto, desde que essa entidade exista. Isso permite que um dado sistema primeiro avalie o texto completo para, em seguida, marcar as relações existentes entre as EM segundo qualquer tipo de algoritmo.

C.1.2 Relações múltiplas de uma dada EM

É naturalmente possível que uma dada EM possua relações diferentes com mais de uma EM. Nesses casos, marcamos as diferentes relações em uma estrutura de lista:

> A actual administração dos <EM ID="471">**Hipermercados Extra**</EM>, presidida por <EM ID="471" **COREL="470" TIPOREL="outra"**>**Abílio Diniz**</EM>, líder do grupo <EM ID="472" **COREL="470 471" TIPOREL="inclui outra"**>**Pão de Açúcar**</EM>,...

C.1.3 Equivalência entre relações

Note-se que não é preciso identificar exaustivamente todas as relações entre todas as EM de um texto. Pelo contrário, se existirem quatro EM com o mesmo referente, basta especificar três relações, e não doze (veja-se Villain et al., 1995).

Da mesma forma, a marcação de uma relação implica a sua inversa. Não é portanto preciso marcar `inclui` e `incluido` no mesmo par, ou "outra"duas vezes.

Veja-se um exemplo de duas maneiras equivalentes de anotar a mesma frase:

> Em 9 de Setembro de 1895, foi organizado em <EM ID="15">**New York**</EM> o <EM ID="16" COREL="15" TIPOREL="ocorre_em">**Congresso Americano de Bowling**</EM> ("<EM

> ID="17" **COREL="16 15" TIPOREL="ident ocorre_em"**>**ABC**</EM> - <EM ID="18" **COREL="16 15" TIPOREL="ident ocorre_em"**>**American Bowling Congress**</EM>"), sediado em <EM ID="19" **COREL="15 16 17 18" TIPOREL="incluido sede_de sede_de sede_de"**>**Milwaukee**</EM>, com o objectivo de aplicar medidas correctivas contra os excessos de jogatina e aperfeiçoar ainda mais as regras.
>
> Em 9 de Setembro de 1895, foi organizado em <EM ID="15" **COREL="19" TIPOREL="inclui"**>**New York**</EM> o <EM ID="16" **COREL="15" TIPOREL="ocorre_em"**>**Congresso Americano de Bowling**</EM> ("<EM ID="17" **COREL="16"**>**ABC**</EM> - <EM ID="18" **COREL="16"**>**American Bowling Congress**</EM>"), sediado em <EM ID="19" **COREL="16" TIPOREL="sede_de"**>**Milwaukee**</EM>, com o objectivo de aplicar medidas correctivas contra os excessos de jogatina e aperfeiçoar ainda mais as regras.

C.1.4 Opcionalidade de marcação de `TIPOREL` no caso de identidade

Para facilitar a tarefa, a relação de identidade é considerada a relação padrão e, por isso, não precisa estar marcada (ou seja, numa EM pode existir `COREL` e não `TIPOREL`):

> O presidente em exercício, <EM ID="567">**Mascarenhas Ferreira**</EM>, havia já confirmado(...). (...), mas sim por desinteligências quanto à forma de actuar de <EM ID="867" **COREL="567"**>**Mascarenhas Ferreira**</EM>, a quem alguns dirigentes acusam....
>
> <EM ID="FG51">**João Steiner**</EM>, astrofísico da USP, durante a (...), explicou <EM ID="FG560" **COREL="FG51"**>**Steiner**</EM>

A exceção é para os casos em que há mais de uma relação para uma dada EM, e uma delas é de identidade. Nessas situações a identidade precisa ser marcada através de `TIPOREL="ident"`, para evitar confusão de etiquetas.

Opcionalidade não implica, naturalmente, proibição, o que significa que os sistemas podem marcar sempre `ident` explicitamente se o preferirem.

C.2 Tipos de relações a marcar

A seguir descrevemos os tipos de relação que devem ser anotados. Lembramos novamente que a relação de identidade, por ser considerada padrão, não precisa receber o atributo `TIPOREL`.

C.2.1 Relação de identidade

A relação de identidade ocorre entre EM que pertencem à mesma categoria. Além de marcar como idênticas as EM que têm o mesmo referente, esta relação, porque se refere às EM e não às expressões textuais, vincula também EM relacionadas por abreviaturas, acrônimos, traduções e "nomes alternativos":

> <EM ID="1220">**Universidade de Trás-os-Montes e Alto Douro**</EM> (<EM ID="282" **COREL="1220"**>**UTAD**</EM>)

C.2.2 Relação de inclusão

A relação de inclusão é bastante genérica e abrangente, e compreende desde relações entre EM do tipo LOCAL a relações entre EM do tipo ORGANIZACAO e ABSTRACCAO. Quando a entidade descrita por uma EM inclui a entidade descrita por outra, a relação entre as duas EM é marcada como TIPOREL="inclui".

> O <EM ID="119">**Centro de Convenções de Curitiba**</EM>, endereço presente há muitos anos na cidade, escondido na <EM ID="120" **COREL="119" TIPOREL="inclui"**>**Rua Barão do Rio Branco**</EM>.

Quando a relação é inversa, é marcada como TIPOREL="incluido":

> Chama-se "<EM ID="11">**Feira Nova de Outubro**</EM>", é organizada pela Câmara Municipal, e é bem antiga, pois remonta aos finais do século XIV. (...) Complementarmente, para além desta vertente tradicional, a <EM ID="13" COREL="11">**Feira Nova de Outubro**</EM> inclui também um <EM ID="14" **COREL="11" TIPOREL="incluido"**>**Pavilhão de Actividades Económicas**</EM>, onde qualquer empresa pode comercializar e/ou fazer divulgação dos seus produtos.

A relação de inclusão também vincula EM que, embora expressas pela mesma palavra, não apresentam uma relação de identidade, mas antes uma relação entre um elemento de uma classe e a generalidade de uma classe.

> Astrônomos brasileiros esperam fotografar os primeiros planetas fora do Sistema Solar com a ajuda do maior telescópio do mundo, o <EM ID="aa89">**Gemini**</EM> (...). (...) os telescópios <EM ID="AAFG56" **COREL="aa89" TIPOREL="inclui"**>**Gemini**</EM> têm capacidade científica...

A seguir, diversos exemplos em que a relação de inclusão está presente:

- entre ORGANIZACAO e ORGANIZACAO

  > <EM ID="123">**PSD/Vila Real**</EM> O deputado social-democrata Fernando Pereira anunciou ontem a sua candidatura à presidência da <EM ID="435" **COREL="123" TIPOREL="incluido"**>**Comissão Política Distrital de Vila Real do PSD**</EM>. (...) Ao contrário do que seria legítimo pensar, a candidatura de Fernando Pereira não aparece como resposta aos maus resultados obtidos pelo <EM ID="222" **COREL="123" TIPOREL="inclui"**>**PSD**</EM> nas eleições autárquicas.

- entre ABSTRACCAO e ABSTRACCAO

  > (...) as funções que venho ocupando no <EM ID="119">**European Script Fund**</EM> do <EM ID="1690" **COREL="119" TIPOREL="inclui"**>**Programa Media das Comunidades Europeias**</EM> fazem com que a única relação institucional...

- entre LOCAL e LOCAL

(...) havia perdido as grandes batalhas da guerra civil no <EM ID="ff203">**Ribatejo**</EM> (<EM ID="ff204" **COREL="ff203" TIPOREL="incluido"**>**Pernes**</EM>, <EM ID="ff205" **COREL="203" TIPOREL="incluido"**>**Almoster**</EM> e <EM ID="ff206" **COREL="ff203" TIPOREL="incluido"**>**Asseiceira**</EM>)

(...) e refugiara-se com o seu quartel-general no <EM ID="ff210">**Alentejo**</EM>, a única região do <EM ID="ff212" **COREL="ff210" TIPOREL="inclui"**>**Reino**</EM>

Em estudos no <EM ID="DS58">**Terceiro Mundo**</EM>, os cientistas se desobrigariam de fornecer aos doentes o melhor tratamento médico conhecido. (...) O debate surgiu após estudos em <EM ID="QQ87" **COREL="DS58" TIPOREL="incluido"**>**Ruanda**</EM> e na <EM ID="QQ90" **COREL="DS58" TIPOREL="incluido"**>**Tailândia**</EM>

(...) encontrou o fóssil na região da <EM ID="AB78">**Chapada do Araripe**</EM>, <EM ID="AB79" **COREL="AB78" TIPOREL="inclui"**>**Ceará**</EM>

- entre TEMPO e TEMPO (note-se que aqui, devido às especificidades da pista do TEMPO, ao contrário do resto do HAREM, *era dos grandes répteis*, embora completamente em minúsculas, deve ser marcado como EM)

  (...), no <EM ID="AB59">**período Cretáceo**</EM> (o último da <EM ID="AB659" **COREL="AB59" TIPOREL="inclui"**>**era dos grandes répteis**</EM>

- entre COISA e COISA

  Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o <EM ID="AB80">**Santanaraptor**</EM> ocuparia uma posição no grupo <EM ID="AB90" **COREL="AB80" TIPOREL="inclui"**>**Tyrannoraptora**</EM>, o mesmo do <EM ID="AB850" **COREL="AB90" TIPOREL="incluido"**>**Tyrannossaurus rex**</EM>

C.2.3 Relação de localização, ou de ocorrência em

Esta relação ocorre freqüentemente entre ORGANIZAçõES ou ACONTECIMENTOS, por um lado e LOCAIS, por outro, indica a localização de um evento ou de uma organização em um determinado local. é expressa por TIPOREL="ocorre_em" ou, de maneira inversa, TIPOREL="sede_de".

Embora a designação ocorre_em seja mais apropriada em português para acontecimentos, optámos por ter apenas um nome de relação, visto que a diferença é visível através da categoria a que pertence a entidade relacionada. Leia-se portanto localizada_em quando a relação é entre uma ORGANIZACAO e um LOCAL.

Alguns exemplos desta relação:

- Entre ACONTECIMENTO e LOCAL

  Em 9 de Setembro de 1895, foi organizado em <EM ID="15">**New York**</EM> o <EM ID="16" **COREL="15" TIPOREL="ocorre_em"**>**Congresso Americano de Bowling**</EM> (“<EM ID="17" **COREL="16 15" TIPOREL="ident ocorre_em"**>**ABC**</EM> - <EM ID="18" **COREL="16 15" TIPOREL="ident ocorre_em"**>**American Bowling Congress**</EM>”), sediado em <EM ID="19"

COREL="15 16 17 18" TIPOREL="incluido sede_de sede_de sede_de">**Milwaukee**</EM>, com o objectivo de aplicar medidas correctivas contra os excessos de jogatina e aperfeiçoar ainda mais as regras.

A <EM ID="ff001">**Concessão de évora Monte**</EM> ou <EM ID="ff002" COREL="ff001">**Capitulação de évora Monte**</EM> (depois impropriamente chamada de Convenção de évora Monte) foi um acordo assinado entre liberais e miguelistas na pacata vila alentejana de <EM ID="ff005" **COREL="ff001" TIPOREL="sede_de"**>**évora Monte**</EM> (hoje concelho de <EM ID="ff006" **COREL="ff005" TIPOREL="inclui"**>**Estremoz**</EM>), em 26 de Maio de 1834

- Entre `ORGANIZACAO` e `LOCAL`

  A <EM ID="hg65">**IBM Research**</EM>, com o seu quartel general em <EM ID="hgoi76" **COREL="hg65" TIPOREL="sede_de"**>**Yorktown Heights**</EM>, lidera o ranking das publicações americanas na indústria.

C.2.4 Outras relações

Esta relação, marcada por `TIPOREL="outra"`, compreende outros tipos de relação ainda não previstos ou que, até o momento, não nos pareceram relevantes para merecerem uma especificação mais detalhada.

De notar que é importante que, se houver sistemas que marquem explicitamente outras relações que não se incluam nas três anteriores, estamos dispostos a mapeá-las automaticamente nesta categoria, para minimizar o trabalho dos participantes.

É contudo conveniente salientar que a relação "ocorre no mesmo texto que" não se encontra abrangida pela relação `outra`, e que existem portanto casos em que não esperamos que seja natural estabelecer relações entre EM do mesmo texto, ou do mesmo parágrafo.

Exemplos já cobertos pelo nosso trabalho preliminar anteriormente citado (e posto à disposição dos participantes em Dezembro de 2007), mas que classificamos aqui apenas como `outra`, são:

- relação de cargo entre uma `PESSOA` e uma `ORGANIZACAO`

  <EM ID="ex1-39">**Miguel Rodrigues**</EM>, que trabalha nos <EM ID="ex1-40" **COREL="ex1-39" TIPOREL="outra"**>**Serviços Administrativos**</EM>

  <EM ID="115">**Vale Abraão**</EM>, de <EM ID="112" COREL="115" TIPOREL="outra">**Manoel de Oliveira**</EM> teve início com a nomeação de <EM ID="115a">**Pedro Santana Lopes**</EM> para <EM ID="116" **COREL="115a" TIPOREL="outra"**>**secretário de Estado da Cultura**</EM>

- relação de parentesco entre duas `PESSOA`s.

  <EM ID="ex3-12">**D. Miguel I**</EM>, considerado soberano usurpador do trono de sua sobrinha <EM ID="ex3-13" **COREL="ex3-12" TIPOREL="outra"**>**D. Maria da Glória**</EM>

- relação entre um projecto e a `COISA` ou `LOCAL` a que diz respeito

> O <EM ID="119">**Centro de Convenções de Curitiba**</EM>, endereço presente há muitos anos na cidade, escondido na <EM ID="120" COREL="119" TIPOREL="inclui">**Rua Barão do Rio Branco**</EM> agora sendo revitalizada dentro do projeto <EM ID="121" **COREL="120" TIPOREL="outra"**>**Cores da Cidade**</EM> pode ter um impulso que o coloque como ponto de convergência das iniciativas de negócios...
>
> Astrônomos brasileiros esperam fotografar os primeiros planetas fora do Sistema Solar com a ajuda do maior telescópio do mundo, o <EM ID="aa89">**Gemini**</EM> (...). O projeto <EM ID="FG56Oy" **COREL="aa89" TIPOREL="outra"**>**Gemini**</EM>, resultado de um consórcio de sete países, envolve

- relação de baptismo entre um nome (ABSTRACCAO) e o que lhe deu origem

  > O exemplar de <EM ID="AB880">**Santanaraptor**</EM> encontrado (...) O nome é uma alusão à região onde ele viveu (a <EM ID="RR56" **COREL="AB880" TIPOREL="outra"**>**Formação Santana**</EM>).

De notar que também não consideramos como relações de identidade aquelas que ligam a menção do nome (ABSTRACCAO) à referência da entidade, ou seja a relação de identidade pressupõe o mesmo tipo semântico. Por exemplo:

> A <EM ID="ff001">**Concessão de évora Monte**</EM> ou <EM ID="ff002" COREL="ff001">**Capitulação de évora Monte**</EM> (depois impropriamente chamada de <EM ID="ff003" **COREL="ff001" TIPOREL="outra"**>**Convenção de évora Monte**</EM>)
>
> Batizado de <EM ID="AB60">**Santanaraptor placidus**</EM>, o fóssil é o único a ser encontrado no país (...). Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o <EM ID="AB80" **COREL="AB60" TIPOREL="outra"**>**Santanaraptor**</EM> ocuparia uma posição ...

O fato de o HAREM considerar as EM no contexto implica que entidades relacionadas por metonímia, como *Espanha* no exemplo abaixo, que ora pode ser referida como LOCAL, ora como ORGANIZACAO do tipo ADMINISTRACAO, também deverão estar relacionadas – neste caso, a relação é do tipo outra. Ou, visto de outra maneira, é a relacão do tipo outra que nos garante a presença de um vínculo entre entidades relacionadas por metonímia.

> a retirada para <EM ID="ex3-28" CATEG="LOCAL" TIPO="ADMINISTRATIVO">**Espanha**</EM> para auxiliar a causa carlista de seu primo Don Carlos (pretendente absolutista ao trono da <EM ID="ex3-31" CATEG="ORGANIZACAO" TIPO="ADMINISTRACAO" **COREL="ex3-28" TIPOREL="outra"**>**Espanha**</EM>

C.2.5 Relações entre EM vagas

Outro ponto importante diz respeito a relações entre EM que, embora expressas pelo mesmo mesmo item lexical, seu uso, no contexto, salienta diferentes facetas de seu significado, não sendo possível, ou melhor, necessário, escolher entre elas.

Nos termos do HAREM, trata-se de relações entre EM que são vagas, e que por isso apresentam mais de uma classificacao semântica. Nestes casos, é possível que apenas uma

das facetas participe de uma determinada relação (e, teoricamente, nada impede ainda que outra faceta seja evidenciada em outra relação).

No exemplo abaixo, a EM descrita pelo termo *Concessão de évora Monte* pode ser interpretada como ACONTECIMENTO ou como OBRA (do tipo PLANO). Porém, apenas a faceta ACONTECIMENTO pode apresentar uma relação do tipo SEDE_DE com a EM descrita por *Évora Monte*.

A DISCUTIR AINDA: Neste caso, marcam-se duas relações: aquela identificada recebe o nome da relação, e a outra relação, decorrência da classificação vaga, é anotada como outra:

> A <EM ID="ex3-4" CATEG="OBRA|ACONTECIMENTO" TIPO="PLANO|EVENTO">**Concessão de évora Monte**</EM> ou <EM ID="ex3-5" CATEG="OBRA|ACONTECIMENTO" TIPO="PLANO|EVENTO" **COREL="ex3-4"**>**Capitulação de évora Monte**</EM> (depois impropriamente chamada de <EM ID="ex3-6" CATEG="ABSTRACCAO" TIPO="NOME" **COREL="ex3-4"**>**Convenção de évora Monte**</EM>outra) foi um acordo assinado entre liberais e miguelistas na pacata vila alentejana de <EM ID="ex3-7" CATEG="LOCAL" TIPO="ADMINISTRATIVO" **COREL="ex3-4" TIPOREL="outra sede_de"**>**évora Monte**</EM> (hoje concelho de <EM ID="ex3-9" CATEG="LOCAL" TIPO="ADMINISTRATIVO" **COREL="ex3-7"**>**inclui**</EM>)

Da mesma forma, isso pode acontecer quando a vagueza implica diferenças na delimitação, o que no HAREM é indicado pelas etiquetas ALT.

> Boa parte do poderio militar americano atual foi desenvolvido <ALT><EM ID="oiu65">**durante a era Reagan**</EM>|durante a era <EM ID="xf5">**Reagan**</EM></ALT>. <EM ID="o065pre" **COREL="xf5"**>**Reagan**</EM> gostava particularmente de ...

De notar que ainda não foi tomada nenhuma decisão específica em relação às categorias do TEMPO, visto que a própria definição da tarefa engloba alguns conceitos de co-referência, e porque a sua definição não foi feita por nós.

C.2.6 Quadro-resumo das categorias por tipo de relações a marcar

Tabela C.1: Quadro-resumo das categorias por tipo de relações a marcar

Relação	**Categorias a que se aplica**
Identidade	Todas, exigindo igualdade de CATEG, TIPO e SUBTIPO
Inclusão	Todas menos VALOR, exigindo igualdade de CATEG
Ocorrência ou localização	ACONTECIMENTO e LOCAL; ORGANIZACAO e LOCAL
Outras	Todas, sem qualquer restrição

Agradecimentos

Agradecemos as muitas sugestões e comentários da Renata Vieira, cuja proposta inicial de fazer uma pista de co-referência nos pôs nesta pista, e do David Cruz, Nuno Cardoso e Cristina Mota sobre versões preliminares deste documento. **Aproveitamos para pedir a todos os participantes do HAREM que se pronunciem.**

Apêndice D

Critérios de ALT no Segundo HAREM

Paula Carvalho, Cláudia Freitas e Diana Santos

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Apêndice D, p. 319–321.

Nota das editoras: Este apêndice destaca a secção *Critérios de ALT*, disponibilizados aos participantes, com as opções de anotação da colecção dourada. O documento com as opções tomadas foi modificado pela última vez no dia 5 de Junho de 2008. Não incluímos esse documento na íntegra aqui, uma vez que essas opções se encontram documentadas e motivadas nos capítulos 1, 3 e 4, respectivamente.

Apresentam-se primeiro as regras sistemáticas usadas na anotação dos ALT, por CATEGoria. As regras mais finas encontram-se a seguir.

1. PESSOA

   * <PESSOA> | < PESSOA> de <LOCAL>
     <PESSOA INDIVIDUAL> | <PESSOA INDIVIDUAL> de <LOCAL>
     <PESSOA CARGO> | <PESSOA CARGO> de <LOCAL>
     <PESSOA GRUPOCARGO> | <PESSOA GRUPOCARGO> de <LOCAL>
     <PESSOA GRUPOIND> | <PESSOA GRUPOIND> de <LOCAL>

   * <PESSOA> |< PESSOA> de <ORGANIZACAO> <PESSOA CARGO> | <PESSOA CARGO> de <ORGANIZACAO>
     <PESSOA GRUPOMEMBRO> | <PESSOA GRUPOMEMBRO> de <ORGANIZACAO>
   * <PESSOA> |< PESSOA> de <ORGANIZACAO|LOCAL>
     <PESSOA CARGO>| <PESSOA CARGO> de <ORGANIZACAO|LOCAL>
     <PESSOA GRUPOCARGO>| <PESSOA GRUPOCARGO> de <ORGANIZACAO|LOCAL>
   * <PESSOA> |< PESSOA> de <PESSOA>
     <PESSOA GRUPOMEMBRO> | <PESSOA GRUPOMEMBRO> de <PESSOA>

2. ORGANIZACAO

   * <ORGANIZACAO>| <ORGANIZACAO> de <ORGANIZACAO>
   * <ORGANIZACAO>| <ORGANIZACAO> de <LOCAL>

3. LOCAL

   * <LOCAL>| <LOCAL> de <LOCAL>
     <LOCAL HUMANO CONSTRUCAO> | <LOCAL HUMANO CONSTRUCAO> de <LOCAL>
     <LOCAL HUMANO REGIAO> | <LOCAL HUMANO REGIAO> de <LOCAL>
     <LOCAL HUMANO OUTRO>| <LOCAL HUMANO OUTRO> de <LOCAL>
   * <LOCAL>| <LOCAL> de <ORGANIZACAO>
     <LOCAL HUMANO CONSTRUCAO> | <LOCAL HUMANO CONSTRUCAO> de <ORGANIZACAO>

4. OBRA

   * <OBRA> | <OBRA> de <PESSOA>
   * <OBRA ARTE|LOCAL> | <OBRA ARTE|LOCAL> de <LOCAL>

5. ACONTECIMENTO

    * <ACONTECIMENTO> | <ACONTECIMENTO> de <ORGANIZACAO>
    * <ACONTECIMENTO> | <ACONTECIMENTO> de <LOCAL>
    * <ACONTECIMENTO> | <ACONTECIMENTO> de <TEMPO>

6. ABSTRACCAO

    * <ABSTRACCAO DISCIPLINA> | <ABSTRACCAO DISCIPLINA> de <LOCAL>
    * <ABSTRACCAO DISCIPLINA> | <ABSTRACCAO DISCIPLINA> de <PESSOA>

7. COISA

    * <COISA CLASSE> | <COISA CLASSE> de <ORGANIZACAO>

Regras que refinam e por isso contradizem estas regras gerais:

- A regra PESSOA de LOCAL não foi empregue nos casos em que a pessoa, referida pelo seu título nobiliário (que marcámos como CARGO) nos casos: conde, duque e marquês. Esta opção deve-se ao facto de termos considerado como demasiado remota a relação que se estabelece entre a menção ao título e ao nome do local (*Conde de Ourém, Duque de Bragança, Marquês de Pombal*).
- As regras de segmentação de OBRA não se aplicam quando a OBRA está entre aspas ou plicas. Nesse caso considerámos que apenas a OBRA maior está a ser mencionada.
- A regra LOCAL de LOCAL não se aplica quando consideramos que foi o LOCAL pequeno que deu o nome, como LOCAL, ao LOCAL mais pequeno, ou seja: *Mosteiro dos Jerónimos* deu origem ao nome de *Jerónimos* como LOCAL (freguesia de Lisboa, ou zona próxima do Mosteiro), e por isso apenas o LOCAL Mosteiro dos Jerónimos deve ser marcado, quando nos encontramos em presença da expressão completa.

Note-se que as regras referentes à justaposição de várias palavras encontram-se na secção correspondente, assim como as regras relativas a ALT com nada encontram-se descritas na secção associada à delimitação de EM (veja-se a secção 1.3.2).

Apêndice E

Exemplário do Segundo HAREM

Paula Carvalho, Cláudia Freitas, Diana Santos e Hugo Gonçalo Oliveira

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Apêndice E, p. 323–338.

Nota das editoras: Neste apêndice reproduzimos a versão 1.0 do exemplário, a qual foi actualizada e disponibilizada na sua versão final em 19 de Março de 2008. Os exemplos encontram-se organizados por CATEGORIA/TIPO/SUBTIPO.

E.1 PESSOA

E.1.1 INDIVIDUAL

1. A cerimónia foi presidida pelo Primeiro Ministro, **Engenheiro António Guterres** e contou com a presença do Ministro da Defesa Nacional, **Dr**. **Castro Caldas**
2. A **rainha Isabel II** surpreeendeu a Inglaterra (...)
3. Quando o **Papa João Paulo II** visitou Fátima (...)
4. **Sua Santidade o Papa Bento XVI** é o atual Papa da Igreja Católica
5. Carta aberta a Sua Santidade, o **Papa Bento XVI**.
6. O deputado social-democrata **Fernando Pereira** anunciou (...)
7. **Tia Maria** e **Tio Manel**, dois simpáticos e alegres residentes de uma aldeia serrana da Beira Alta, estão casados há 43 anos.
8. O **Primeiro-Ministro José Sócrates** anunciou o aumento do complemento solidário para idosos de 323,5 para 400 euros.
9. **D. Catarina de Áustria** (ou **Catarina de Habsburgo**, ou mais raramente, **Catarina de Espanha**) foi arquiduquesa da Áustria, princesa de Espanha e rainha de Portugal (da casa dos Habsburgos).
10. O bisavô de **Catarina**, **Miguel** e **Paulo Portas** era "um pequeno proprietário rural de Tomiño, na baixa Galiza, junto ao Rio Minho", explica o arquitecto **Nuno Portas**, pai do trio.

E.1.2 CARGO

1. Intervenção do **Ministro da Presidência** na tomada de posse do **Alto Comissário para a Imigração e Minorias Étnicas**, cerimónia que foi presidida pelo **Primeiro-Ministro**
2. O actual **alto-comissário para a Imigração e Diálogo Intercultural**, Rui Marques, foi substituído no cargo, a seu pedido, por Rosário Farmhouse.
3. A reunião mensal do **ministro da Administração Interna** com os Governadores Civis teve lugar em Bragança, no sábado, dia 2 de Fevereiro.
4. O **Presidente da Câmara de Lisboa**, António Costa, não excluiu hoje a hipótese de se demitir do cargo se o PSD inviabilizar na Assembleia Municipal (...)

5. Hoje é também aniversário da Princesa Victoria Ingrid Alice Desirée (foto ao lado), que um dia será **Rainha da Suécia**

6. O primeiro-ministro britânico, Gordon Brown, afirmou hoje que o seu antecessor Tony Blair seria um grande candidatoa **presidente da UE**

7. Intervenção de Encerramento de **Sua Excelência o Primeiro Ministro**

8. O **presidente da República** é, de uma forma geral, o **chefe de Estado**

9. Entre as altas entidades presentes destacam-se os senhores Almirantes Fuzeta da Ponte e Vieira Matias, **Presidente da Academia de Marinha**, **Inspector-Geral das Forças Armadas**, **Reitores da Universidade de Lisboa** e **da Universidade Nova de Lisboa**, **Director do Instituto de Altos Estudos da Força Aérea**, ...

Nota 1: No exemplo 9, embora Reitores esteja no plural, o que poderia dar a indicação de que se trataria de um GRUPOCARGO, no contexto em questão, a EM refere apenas um indivíduo (CARGO). A marca de plural deve-se ao facto de a referida EM se encontrar numa estrutura de coordenação.

E.1.3 GRUPOCARGO

1. Rui Pereira exortou os **Governadores Civis** a avançarem, este ano, com estruturas de coordenação distritais de segurança rodoviária

2. Nos dias 04 e 05 de maio, a **Presidência** reuniu-se com lideranças dos movimentos da IECLB.

3. O cortejo tem a seguinte ordenação: os doutores, dois a dois, por ordem inversa de precedência das unidades e, nestas, por categoria e antiguidade, os mais recentes à frente; **Reitores**, **Vice-Reitores** e Professores de outras Universidades, **Presidentes dos Centros Regionais** (desde que sejam professores), **Vice-Reitores da UCP**, Reitor Honorário (quando houver), Reitor dando a direita ao Ministro da Educação, sempre que este esteja presente, Magno Chanceler.

4. O **Conselho de Ministros** tomou nota, com satisfação, dos esforços do Secretariado

5. Em todos os períodos, os **presidentes da República** intervieram e marcaram a história do país.

E.1.4 GRUPOMEMBRO

1. **Cristãos** e **Muçulmanos** não estavam sempre em guerra.

2. Ontem, passaram 40 anos sobre a edição do primeiro álbum dos **Pink Floyd**

3. Sábado, 31 de Março, pelas 21h30, o palco do Teatro José Lúcio da Silva recebe a **Companhia Nacional de Bailado** (**CNB**)

4. O **Sporting** vai defrontar hoje a **Udinese**, na segunda mão da terceira pré-eliminatória da Liga dos Campeões.

5. A **Inglaterra** derrotou nesta quarta-feira a **Suíça** por 2 a 1, em amistoso internacional disputado no Estádio Wembley, em Londres.
6. A **Selecção** foi automaticamente apurada para a fase de grupos.
7. Após a etapa de Montreal, a **Ferrari** ocupa a quinta posição no Mundial de Construtores, com 45 pontos, 31 a menos que a líder **Renault**.
8. O **INEM** conduziu a senhora ao serviço de urgências mais próximo do local do acidente.
9. Eu assisti ao show do **U2**.

E.1.5 MEMBRO

1. O **GNR** apenas me disse que eu estava autuado e pediu-me a identificação.
2. George Harrison era conhecido como "o **Beatle** discreto e quieto".
3. A posição do **Hindu** relativamente a essa questão é compreensível.
4. **Diário de Notícias** - Confirma cortes nos benefícios fiscais, bem como nos limites de deduções e abatimentos?

   Sousa Franco - Vamos promover a revisão profunda do Estatuto dos Benefícios Fiscais, introduzindo maior justiça.

E.1.6 GRUPOIND

1. O **Governo** deslocou-se ao fim da tarde da passada segunda-feira ao Palácio de Belém para apresentar cumprimentos de Ano Novo ao Presidente da República
2. Já a **Família Espírito Santo**, que detém cerca de um terço do Banco Espírito Santo, tinha um património avaliado em quase 1,3 milhões de euros.
3. São parentes de certeza, pois, os **Fonsecas**, **Coutinhos**, **Freires** e **Andrades** dessa zona são todos parentes uns dos outros.

E.1.7 POVO

1. **Portugal** consome muito peixe.
2. A esperança média de vida do **Terceiro Mundo** é absurdamente baixa.
3. **Odivelas** precisa de quem as saiba ouvir, de quem conheça os seus problemas e de quem tenha soluções para esses problemas
4. Ronaldo conquistou a **Inglaterra**

E.2 ABSTRACCAO

E.2.1 DISCIPLINA

1. A controvérsia a respeito da personalidade de Jesus (S.W.S) é a principal diferença entre o **Islamismo** e o **Cristianismo**.

2. Doutorou-se em **Engenharia de Sistemas e Computadores**, especialidade de **Informática**, pela Universidade da Madeira em 2001.

3. O **Judo** é uma actividade física e desportiva, composta por técnicas de projecção, imobilização, estrangulamento e luxação (braços)

4. A música **Gospel** é uma fascinante herança da escravatura negra que procurou na fé Cristã amenizar o seu quotidiano, os seus males e consolar a sua dor, recorrendo naturalmente à música (misturando vários estilos musicais como o **Jazz**, **Blues**, rock, etc.) criando um estilo próprio.

5. O **Socialismo** é um sistema sócio-político caracterizado pela apropriação dos meios de produção pela coletividade.

E.2.2 ESTADO

1. A **Síndrome de Estocolmo** (**Síndrome Stockholm**) é um estado psicológico particular desenvolvido por pessoas que são vítimas de seqüestro.

2. A **síndrome de Alström** é uma doença hereditária muito rara.

3. Uma **DTS** não tratada aumenta a transmissão de HIV em dez vezes

4. Que vantagens terá um doente com **Alzheimer** saber o que tem?

5. Os portadores da [**HIV**]/[**Aids**] | [**HIV/AIDS**] ainda sofrem preconceito.

Nota 2: No exemplo (5), o símbolo "|" marca as diferentes possibilidades de segmentação, consideradas correctas, da(s) EM em análise.

E.2.3 IDEIA

1. Neste blogue praticam-se a **Liberdade** e o **Direito de Expressão** próprios das sociedades avançadas.

2. Todos os anos, muitos licenciados portugueses fazem as malas e atravessam o Atlântico em busca do "**Sonho Americano**".

3. Qualquer dia já ninguém acredita na **República** e na **Democracia**.

4. A administração Bush identifica-se com a **Justiça Divina**.

5. O senhor acredita na **Ressurreição**?

E.2.4 NOME

1. Novo Papa tem o nome de **Bento XVI**.
2. Conhece-se o facto da povoação ter nascido numa zona denominada "**Lugar Velho**", uma zona baixa que impedia o avanço dos assaltantes vindos do mar.
3. Don Tapscott, outro autor canadiano, baptizou-a de "**Geração Net**", e escreveu um quase manifesto sobre o que eles são e o que querem, a que deu o nome de **Growing Up Digital - The Rise of the Net Generation**.
4. O que significa a sigla **JCB**?
5. Conhecida como a "**Cidade do Sol**", Natal tem belas praias, esquibunda e passeios de dromedário.

E.3 ACONTECIMENTO

E.3.1 EFEMERIDE

1. A **Batalha de Aljubarrota** foi uma das batalhas mais importantes na História de Portugal.
2. O **25 de Abril** abriu os horizontes ao país.
3. Cinco anos depois do **11 de Setembro**, Bin Laden continua a monte.
4. A revelação da utilização de urânio empobrecido em munições aquando da **Guerra do Golfo** surgiu, pois, como chocante "inovação" militar.
5. Apesar de viver nos EUA, nunca comemora o **Thanksgiving**.
6. Dr Alkatiri tomou posse como Primeiro-Ministro a 20 de Maio de 2002, o dia da **Restauração da Independência da República Democrática de Timor-Leste**.

Nota 3: De acordo com as actuais directivas do TEMPO, dia da Restauração da Independência da República Democrática de Timor-Leste deverá ser igualmente classificado como uma expressão temporal.

E.3.2 ORGANIZADO

1. O **Campeonato da Liga de Andebol** arrancou esta quarta-feira.
2. A Daikin patrocina os pavilhões da Bélgica e da Nova Zelândia na [**Exposição Mundial de 2005**]
3. O Eurosport preparou uma cobertura inovadora para a edição deste ano da **Volta a França**, um dos mais importantes eventos desportivos a nível mundial.
4. A organização do **Rock in Rio 2008** acabou de confirmar mais um grande nome para o seu cartaz.

5. A cerimónia dos **Óscares 2008** será no dia 24 de Fevereiro.
6. Participou no **2º Simpósio Doutoral da Linguateca**
7. Esta dimensão do futuro da justiça europeia será debatida durante a **Presidência Portuguesa**, em particular na conferência no âmbito da visita de uma delegação de juízes e advogados-gerais do Tribunal de Justiça das Comunidades Europeias a Lisboa, em 12 e 13 de Julho.

E.3.3 EVENTO

1. O **Benfica-Sporting** vai realizar-se às 19:15. Duas horas depois decorrerá, no Dragão, o **FC Porto-Boavista**.
2. O "**Concerto de encerramento da Presidência portuguesa da União Europeia**" terá lugar no dia 19 de Dezembro, pelas 21h, no Centro Cultural de Belém.
3. Foi cancelado **André Sardet**, em Cascais, dia 7 de Junho

E.4 COISA

E.4.1 CLASSE

1. Tem **WAP**, **MMS**, **GPRS**, **Bluetooth**, **IRDA** e todas as macacadas dos aparelhos de última geração.
2. O **Doberman** é, de maneira geral, um cão muito ativo, enérgico e determinado, extremamente ligado à família a que pertence.
3. Se a **Declaração de IRS** não foi entregue pela Internet, o comprovativo da entrega desta declaração será o duplicado do **Modelo 3** com o respectivo carimbo do Serviço de Finanças onde a declaração foi entregue.
4. A Santogal, empresa fundada há 60 anos, representa a **Ferrari** e a **Maserati** em Portugal há oito anos, nas cidades de Lisboa e Porto.
5. Uma das refeições preferidas das crianças é o **HappyMeal** da McDonald's.

E.4.2 MEMBROCLASSE

1. Matilde levantou-se, pegou no **Diário de Notícias** e deitou-o no lixo.
2. O meu **Toshiba** está novamente avariado.
3. O **Pastor Alemão** foi o primeiro classificado no concurso.
4. Ainda não entreguei a minha **Declaração de IRS** nas Finanças.
5. O seu **Ferrari** nunca o deixa ficar mal.
6. A criança deliciou-se com o **HappyMeal** que acabou de comer.

E.4.3 OBJECTO

1. Infelizmente, o **Titanic** afundou-se e milhares de vidas perderam-se com ele.
2. Divide o seu apartamento em Lisboa com o cão, **Bobi**.
3. Os cientistas deram conta desta enorme onda expansiva cujo tamanho é comparável à órbita de **Saturno** em volta do **Sol**.
4. A **Exposição Permanente** está organizada cronologicamente.

E.4.4 SUBSTANCIA

1. Superko e outros especialistas crêem que, embora a quantidade de **LDL**, ou colesterol mau, no sangue possa ser perigosa, o tamanho das suas partículas é mais preocupante. Os doentes de padrão B têm tendência para problemas associados a este, incluindo baixos níveis de **HDL** e níveis elevados de triglicéridos.
2. Cientistas espanhóis descobrem nova forma de mapear **ADN**.
3. Esse fármaco tem uma percentagem elevada de **Betametasona**.
4. Não há sintomas registados de excesso de **vitamina B12**.
5. A presidência portuguesa da União Europeia (UE) propõe assim que a meta definida para 2020, que é a de reduzir as emissões de dióxido de carbono (**C02**) em 20%, até 2050.

E.5 LOCAL

E.5.1 HUMANO

E.5.1.1 PAIS

1. Qualquer cidadão da **União Europeia** pode agora escrever ao Parlamento Europeu.
2. Quem vive no **Mónaco**, o principado mais badalado do planeta, não sabe o que é imposto de renda.
3. Nosso café faz sucesso na **Terra do Sol Nascente**.
4. Desde que eclodiu o conflito na [**ex-Jugoslávia**] | ex-[**Jugoslávia**], em 1991, a **República Federal da Jugoslávia** (**RFJ**), compreendendo a **Sérvia** e **Montenegro**, concedeu asilo a cerca de 650.000 refugiados, dos quais 480.000 são da **Bósnia-Herzegovina** (**BH**) e 170.000 dos sectores de **Krajina**, na **Croácia**.

E.5.1.2 DIVISAO

1. Segundo dados do INE de 2006, o concelho de **Sintra**, apesar de ter menos residentes do que o concelho de **Lisboa**, é o que mais crianças tem.
2. Victor Gil nasceu em 1939, em **Santana** (**Figueira da Foz**).

3. **Vila de Rei** volta a ser palco de um evento gastronómico, desta feita dedicado a dois produtos tipicamente portugueses.
4. O governador Jon Corzine promulgou segunda-feira a lei que abole a pena de morte no estado de **New Jersey**
5. A fiscalização aconteceu em **Mato Grosso do Sul**.
6. Desde o início da semana, moradores da **Rocinha** e do **Vidigal** aguardam uma nova guerra entre traficantes das duas favelas...

E.5.1.3 REGIAO

1. As alterações climáticas vão provocar um aumento da disponibilidade de água no **Norte da Europa**, em zonas pouco povoadas, e uma redução nos países do sul
2. As transnacionais expulsam os negócios locais no **Terceiro Mundo** e apoderam-se dos seus mercados.
3. Os Estados Unidos não pretendem construir novas bases militares em **África**, apesar da criação do novo comando militar africano (AFRICOM)
4. Que este processo de revitalização da **Baixa** e **Terreiro do Paço** carece, no entanto, de estudos sobre a estrutura do edificado, sobre o estado do subsolo, as condições para a instalação de novas actividades e a sustentabilidade dos processos de reabilitação
5. Pelo menos uma vez por mês vai ser possível andar nas compras até à meia-noite, na [**Baixa do Porto**] | [**Baixa**] do [Porto]

E.5.1.4 CONSTRUCAO

1. O **Aeroporto da Madeira** e o **Aeroporto de Porto Santo** são ponto de partida e de chegada de várias companhias aéreas internacionais.
2. Localizado no espaço adjacente à **Piscina Oceânica**, o **Porto de Recreio** será, em breve, uma extensão do **Passeio Marítimo**, de livre acesso e usufruto.
3. O **Circuito Idade Maior**, no **Jardim da Estrela**, é um espaço de manutenção física vocacionado para a prática de desporto sénior situado no centro de Lisboa.
4. A exposição de pintura "Múltiplos Falantes", de João Moniz", está patente na **Galeria Trem**, em Faro, até dia 30 de Junho.
5. O **Pólo Gomes Teixeira** da **FCUP** está situado na Praça Gomes Teixeira, no centro histórico da cidade do Porto, junto à **Torre dos Clérigos**.

E.5.1.5 RUA

1. O Pólo Gomes Teixeira da FCUP está situado na **Praça Gomes Teixeira**, no centro histórico da cidade do Porto, junto à Torre dos Clérigos.
2. Virar à direita no cruzamento da **Av. Lusíada** com a **Av. dos Combatentes**

3. Já todos ouvimos falar dos incomparáveis museus de Nova Iorque, das luxuosas lojas da **5ª Avenida** e dos clássicos musicais da Broadway.

4. Seguimos pela **Rua 5 de Outubro**, antiga **Rua da Selaria**, por onde somos conduzidos à **Praça Grande** ou **de Geraldo**, que marcou o centro da Segunda área de desenvolvimento da cidade, a partir da Porta da Selaria: as arcadas de raiz medieval, com sucessivos arranjos, são característica marcante, prolongando-se pelas ruas da **República**, **João de Deus**, **Largo Luís de Camões** e **Rua José Elias Garcia**.

5. Jantaram no Planet, nos **Champs Elysées**.

Nota 4: No exemplo 4, [de Geraldo] representa a EM [Praça de Geraldo]; como se encontra numa estrutura de coordenação, o nome "Praça", deste segundo membro da estrutura, encontra-se localmente omitido.

E.5.1.6 OUTRO

1. Assistiu, entusiasmado, ao concerto, na **1.ª PLATEIA**.

2. Corria sempre, pela manhã, na **Praia de Sta**. **Cruz**, em Torres Vedras.

3. Ficámos de nos encontrar na estação da **Baixa-Chiado**.

E.5.2 FISICO

E.5.2.1 AGUAMASSA

1. O **Estreito de Gibraltar** é um estreito que separa o **Golfo de Cádis** (no **Oceano Atlântico**) do **Mar de Alborão** (parte ocidental do **Mar Mediterrâneo**).

2. Uma equipa de 14 remadores britânicos e irlandeses estabeleceu novo recorde da travessia do **Atlântico**, ao chegar das Canárias a Barbados em 33 dias, 7 horas e 30 minutos.

3. A **Barragem do Alqueva** tem uma extensão de 1160 quilómetros...

4. Pode encontrar tudo isto e muito mais no **Estuário do Sado**, a terceira zona húmida mais importante do País.

5. A **Lagoa de Araruama** é uma lagoa brasileira que tem um grande corpo d'água com saída para o mar, na Região dos Lagos do Estado do Rio de Janeiro.

E.5.2.2 AGUACURSO

1. Os principais rios que desaguam no Mar do Norte são o **Elba** (em Cuxhaven), o **Weser** (em Bremerhaven), o **Ems** em Emden, o **Reno** e o **Mosa** (em Rotterdam, ou Roterdã), o **Schelde** (em Flushing), o **Tâmisa** e o **Humber** (em Hull).

2. A seca que tem atingido o país fez com que os problemas de poluição no rio **Trancão** se tornassem mais visíveis

3. Tirou fotografias fantásticas às **Cataratas do Niagara**

4. As buscas para encontrar a mulher desaparecida desde segunda-feira na ribeira do **Jamor**, em Belas, devido às cheias, foram suspensas às 18h30 de hoje.

5. O rio **Sado** é o maior rio totalmente português a sul do **Tejo**. Corre de sul para noroeste. Os seus principais afluentes são: **Rio Roxo**, **Rio Figueira**, **Rio Odivelas**, **Rio Xarrama**, **Rio Alcáçovas**, **Rio São Martinho**, **Rio Marateca**, **Rio Campilhos**, **Rio Alvalade**, **Rio Corona** e **Rio Arcão**.

E.5.2.3 RELEVO

1. Grupo viajava em avião que bateu contra montanha na **Cordilheira dos Andes**

2. Em 1997, João Garcia fez a primeira tentativa de escalar o **Everest**, face norte, mas apenas conseguiu atingir os 8.200m

3. A **Serra da Estrela** é a maior elevação de Portugal Continental, e a segunda maior em território da República Portuguesa (apenas o **Pico**, nos Açores, a supera).

4. A **Fossa de Porto Rico**, no Oceano Atlântico, tem uma profundidade de 8.648 metros (28.374 pés)

5. A **Meseta Central**, a unidade de relevo mais antiga da Península Ibérica, tem origem no **Maciço Hespérico**

E.5.2.4 PLANETA

1. A ISS é um esboço do que poderá ser o futuro da a humanidade no espaço, através de sucessivas bases cada vez mais longe do nosso berço (primeiro a **Lua**, depois possivelmente os pontos lagrangeanos em que as forças da **Terra** e do **Sol** ou da **Terra** e da **Lua** se equilibram, e mais tarde **Marte**).

2. Mas apesar de a **Via Láctea** ter um grande tamanho, comparada com determinadas galáxias do universo ela é relativamente uma anã, tome em consideração por exemplo a colossal **Markarian 348** que tem uma impressionante dimensão de 13 vezes superior à **Via Láctea**

E.5.2.5 REGIAO

1. Os recursos naturais de **África** são hoje desejados por todas as grandes economias mundiais.

2. Quem descobriu a **Índia** e em que reinado?

3. Uma viagem por 13 países ligando o calor das areias do **Deserto do Sahara**, à neve do frio da **Escandinávia**

4. Os países dos **Balcãs Ocidentais** estão ainda a passar por uma importante reestruturação económica, política e social, bem como por processos de reforma.

5. O sul da Califórnia, como a nossa **Península Ibérica**, é um exemplo de contrastes geográficos e climáticos muito abusados.

E.6 VIRTUAL

E.6.0.6 COMSOCIAL

1. O anúncio a que me referia estava no **Diário de Notícias** de ontem.
2. Li essa notícia no **BlueBus**
3. Essa música passa constantemente na **Rádio Comercial**
4. Esse Senhor que, de vez em quando, aparece nos **Donos da Bola** não percebe nada de futebol.

E.6.0.7 SITIO

1. Li este post e adorei saber que se podia aceder à **Torre de Tombo** via net.
2. Podes tentar fazer uma pesquisa simples no **GOOGLE** ou no **YAHOO**
3. Faço quase todas as minhas compras no **Continente Online**
4. Essas informações são permanentemente actualizadas na **Bolsa de Emprego Público**
5. O número de visitantes do **Público** online aumentou significativamente no último ano

E.6.0.8 OBRA

1. Para mais informações sobre este concurso, consultar **Regulamento** (.pdf 57Kb /.doc 34Kb).
2. Esta constatação não significa, contudo, que todos os dados apresentados no "**Código Da Vinci**" tenham o mesmo valor.
3. Mesmo assim, estou satisfeita, pois através da historia do Thiago, que foi publicada no site "doeação" e na Revista "**Pais e Filhos**", 06 crianças já puderam ser encaminhadas a tratamento e estão vivas!!

E.7 OBRA

E.7.1 ARTE

1. Leonardo Da Vinci criou muitos quadros. Dois deles são muito conhecidos: **Mona Lisa** e a **Última Ceia**.
2. O **Mosteiro de Alcobaça** foi a primeira das "7 maravilhas de Portugal" a ser anunciada.
3. A **Torre de Belém** é um dos monumentos mais expressivos da cidade de Lisboa.
4. **Nossa Senhora do Pranto** foi roubada de sua capela de Vale do Grou no Verão de 1972.

E.7.2 PLANO

1. O Estado subordina-se à **Constituição** e funda-se na legalidade.
2. O **Tratado de Tordesilhas**, assim denominado por ter sido celebrado na povoação castelhana de Tordesillas, foi assinado em 7 de Junho de 1494, entre Portugal e Castela.
3. Em princípio, não temos nada contra a revisão da **Lei de Bases da Prevenção e da Reabilitação e Integração das Pessoas com Deficiência**
4. **Lei n.º 67/98** de 26 de Outubro – **LEI DA PROTECÇÃO DE DADOS PESSOAIS**
5. O **Projecto de Urbanização Falagueira / Venda-Nova**, é um mega projecto imobiliário que prevê a ocupação dos poucos terrenos livres do Concelho da Amadora...

E.7.3 REPRODUZIDA

1. O **Código Da Vinci** foi editado pela Bertrand
2. Na actual edição de **Equador** estão indicadas, na bibliografia, 29 livros, além de jornais e revistas do início do século.
3. O filme **Call girl**, de António-Pedro Vasconcelos, em exibição desde 27 de Dezembro em 40 salas de cinema, foi já visto por mais de 63.000 espectadores
4. **Voo Nocturno**, o mais recente álbum de Jorge Palma acaba de conquistar o galardão de dupla platina. É a primeira vez que um disco do músico português atinge uma fasquia tão alta. O registo, fortemente impulsionado pelo single **Encosta-te a Mim**, continua entre os mais vendidos.

E.8 ORGANIZACAO

E.8.1 ADMINISTRACAO

1. Em Portugal, a **Presidência do Conselho de Ministros** é o departamento governativo que apoia directamente o Primeiro-Ministro na sua função de Presidente do Conselho de Ministros, ou seja de chefe do Governo.
2. O **Conselho de Ministros** discute e aprova Propostas de Lei e pedidos de autorização legislativa (autorização para fazer leis) à **Assembleia da República**
3. A **Administração Bush** não pretende alterar as suas previsões económicas para o orçamento que será apresentado no próximo mês no Congresso.
4. O **Parlamento** iraniano votou hoje uma lei que obriga o **Governo** a "acelerar" o programa nuclear e a rever a cooperação com a Agência Internacional da Energia.
5. O **Ministério da Saúde** contratou 30 médicos uruguaios para trabalharem no 112.

E.8.2 EMPRESA

1. A **RTP** despediu vários trabalhadores.
2. Desde 2005 que trabalha para a **Agência Lusa**.
3. O **Sporting** contratou Dionísio, um jogador de 19 anos que actuava no Piedense.
4. A **Ferrari** lançou, neste domingo, em Maranello, o 53º carro para a disputa de um Mundial de Fórmula 1.
5. O vice-ministro do Ensino da Guiné-Bissau, Joaquim Baldé, disse à **Agência Lusa** que o Acordo Ortográfico da Língua Portuguesa deve ser assinado no próximo ano.

E.8.3 INSTITUICAO

1. A **Companhia Nacional de Bailado** foi criada em 1977 por despacho de David Mourão Ferreira, então Secretário de Estado da Cultura
2. O **Secretariado da Associação de Estudos Europeus de Coimbra** (**AEEC**) funciona em instalações da **Faculdade de Direito da Universidade de Coimbra** | **Faculdade de Direito** da **Universidade de Coimbra**
3. A **Igreja Católica** sempre se viu, portanto, como uma união ou comunhão na diversidade.
4. O incidente aconteceu na noite de 3 de Janeiro e está a ser investigado pela **Polícia Judiciária**
5. As cirurgias programadas para hoje foram adiadas, por falta do pessoal da enfermagem, apurou a Agência Lusa, que cita Abel Rebeca, do **Sindicato dos Enfermeiros Portugueses** (**SEPI**).

E.9 VALOR

E.9.1 CLASSIFICACAO

1. Os candidatos que ficarem classificados [entre o 4º e o 10º] lugar receberão uma Menção Honrosa, bem como um prémio em Software Educativo IMAGINA, Dragões & Companhia, no valor de 60€.
2. Pela primeira vez desde 2002, o cartaz desta **15ª** edição traduz um investimento exclusivamente europeu, privilegiando, além da contínua presença de músicos portugueses (quinteto de Nelson Cascais), duas das mais prestigiadas cenas jazzísticas do continente...

E.9.2 MOEDA

1. Estado já gastou **160 Milhões de Euros** em consultoria para privatizar a Galp

2. Quando praticadas por pessoas singulares, aplicar-se-ão coimas que oscilam **entre 16 000 a 22 500 euros** em caso de dolo.

3. Pagamento trimestral sem iva: **431.37 USD** (Dolares)* (Poupa **3,63 USD**) - Pagamento semestral sem iva: **797.40 USD** (Dolares)* (Poupa **72,60 USD**) ...

4. Hoje custa **menos de 5 reais**. O salário mínimo era **R$ 200**!

5. As moedas eram cunhadas em ouro e prata, sendo que as de ouro valiam **1**, **2**, e **4 mil réis**.

6. Em comunicado, a CGTP sustenta que apenas as pensões até **1,5 IAS** (**611,12 euros**) tem aumentos de 2,4 por cento, em linha com a inflação prevista para 2007.

E.9.3 QUANTIDADE

1. Paciente de **67 anos** é esquecida em aparelho de tomografia.

2. O carro faz **26,6 km/l**.

3. O lago Bosumtwi, hoje com **mais de 10 km** de largura, sofreu ainda mais com a seca, ficando totalmente sem água.

4. Essa perturbação neurocomportamental afecta **entre 5 a 10%** das crianças em idade escolar, sendo mais comum no sexo masculino.

5. Vendo casa em bom estado e terreno com **aproximadamente 1500m2**.

E.10 Exemplos de vagueza

1. **Portugal** está horrorizado com os acontecimentos em Madrid [PESSOA POVO | ORGANIZACAO ADMINISTRACAO]

2. Bento XVI acusou ainda a **Humanidade** de estar demasiado "preocupada consigo própria". [PESSOA GRUPOIND | PESSOA GRUPOMEMBRO | PESSOA POVO]

3. A **Administração Bush** identifica-se com a Justiça Divina. [PESSOA GRUPOMEMBRO | PESSOA GRUPOCARGO | ORGANIZACAO ADMINISTRACAO]

4. O leitor compra o **DN** ou o **JN** à sexta feira e recebe um cartão para trocar por um DVD ao comprar o jornal de Sábado. [COISA CLASSE | COISA MEMBROCLASSE]

5. Serão disponíveis no site as 8.000 fotos recuperadas entre os destroços das **Torres Gémeas** [COISA OBJECTO | LUGAR HUMANO CONSTRUCAO]

6. Convém lembrar a que a **Colecção Berardo** vale 316 milhões de euros [COISA OBJECTO | OBRA ARTE]

7. Brown visita a região mais castigada pelas inundações na **Inglaterra**. [LOCAL HUMANO PAIS | LOCAL FISICO REGIAO]

8. De acordo com o **Diário de Notícias** de ontem, domingo, os pais de filhos recém-nascidos que recorreram à vacina foram informados de que os lotes existentes ... [LOCAL VIRTUAL COMSOCIAL | ORGANIZACAO EMPRESA | PESSOA GRUPOMEMBRO]
9. Hugo Chavez afirmou que os **Estados Unidos** deveriam ser os primeiros a ser incluídos na lista internacional de terroristas [ORGANIZACAO ADMINISTRACAO | PESSOA GRUPOMEMBRO | PESSOA GRUPOIND]
10. Violou vários artigos referidos na Constituição... [OBRA PLANO | LOCAL VIRTUAL OBRA]

Apêndice F

Manual do Etiquet(H)AREM

Paula Carvalho e Hugo Gonçalo Oliveira

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Apêndice F, p. 339–346.

Nota das editoras: Este apêndice reproduz a versão do dia 29 de Abril de 2008 do manual de utilização do Etiquet(H)AREM, publicado electronicamente, em pdf, como relatório da Linguateca separado (Carvalho e Gonçalo Oliveira, 2008).

O **Etiquet(H)AREM** é uma ferramenta de auxílio à anotação de corpora, concebida por Hugo Oliveira, para a etiquetagem de Entidades Mencionadas (EMs) e de relações entre EMs, no âmbito do HAREM (http://www.linguateca.pt/HAREM/).

F.1 Requisitos básicos na utilização do programa

(i) A utilização desta ferramenta pressupõe a instalação de uma máquina de JAVA - Java Runtime Environment (JRE) 1.6 ou mais recente (http://www.java.com/en/download/manual.jsp).

(ii) O ficheiro a ser anotado tem de estar em formato xml, caso contrário o programa não o abre.

(iii) Só são suportados ficheiros XML com DTDs, se estas forem externas. Nesse caso, o ficheiro .dtd terá de se encontrar na mesma directoria para onde o DOCTYPE estiver a apontar. No caso de o ficheiro ter uma DTD interna, não há garantias de bom funcionamento do programa.

(iv) Os valores possíveis para os atributos das EMs estão compreendidas no ficheiro `harem3.conf` (cf. tabela F.1).

F.2 Lista de notações a utilizar

O ficheiro `harem3.conf` corresponde à listagem das Categorias (C), e respectivos tipos (T) e/ou subtipos (S), previstos no âmbito das Directivas do Segundo Harem. O referido ficheiro pode incluir ainda outros atributos igualmente tidos em consideração na anotação (caso de (X) e (Y), como abaixo referido), bem como as relações (R) previstas entre as EMs.

Para adicionar uma nova categoria, tipo ou subtipo, basta introduzir o respectivo nome (em maiúsculas), antecedido de C: , T: ou S: , respectivamente. No que respeita à categoria `TEMPO`, é ainda possível adicionar os atributos X: (`TEMPO_REF`) e Y: (`SENTIDO`), ambos relativos ao subtipo `DATA`.

Os atributos categoria, tipo e subtipo (e eventuais 'subsubtipos') encontram-se organizados hierarquicamente, por esta ordem. Assim, sempre que se insere uma entrada do tipo T:xxx, a categoria a que esse tipo pertence corresponderá à entrada C:yyy mais próxima e imediatamente acima de T:xxx. Os subtipos funcionam de forma idêntica.

Para especificar os tipos de Relações (R) entre EMs, basta declará-las a seguir a R:. Neste caso, convencionou-se que as relações seriam grafadas em minúsculas, ao contrário das categorias, dos tipos e dos subtipos, que são grafados em maiúsculas.

F.3 Manuseamento do programa propriamente dito

Iniciar o Etiquet(H)arem:

Tabela F.1: harem3.conf

```
#Categorias (C), Tipos (T) e Subtipos (S)
#TEMPO_REF (X), SENTIDO (Y)
#Tipos de referencia (R)

C:PESSOA         T:EVENTO
T:INDIVIDUAL     T:OUTRO
T:CARGO          C:ABSTRACCAO
T:GRUPOCARGO     T:DISCIPLINA
T:GRUPOMEMBRO    T:ESTADO
T:MEMBRO         T:IDEIA
T:GRUPOIND       T:NOME
T:POVO           T:OUTRO
T:OUTRO          C:COISA
C:ORGANIZACAO    T:CLASSE
T:ADMINISTRACAO  T:SUBSTANCIA
T:EMPRESA        T:OBJECTO
T:INSTITUICAO    T:MEMBROCLASSE
T:OUTRO          T:OUTRO
C:LOCAL          C:VALOR
T:HUMANO         T:CLASSIFICACAO
S:PAIS           T:QUANTIDADE
S:DIVISAO        T:MOEDA
S:REGIAO         T:OUTRO
S:CONSTRUCAO     C:OUTRO
S:RUA            T:OUTRO
S:OUTRO          C:TEMPO
T:FISICO         T:CALENDARIO
S:AGUACURSO      S:DATA
S:AGUAMASSA      X:ABSOLUTO
S:RELEVO         X:TEXTUAL
S:PLANETA        X:ENUNCIACAO
S:ILHA           Y:ANTERIOR
S:REGIAO         Y:POSTERIOR
S:OUTRO          Y:ANTERIOR_OU_SIMULT
T:VIRTUAL        Y:POSTERIOR_OU_SIMULT
S:COMSOCIAL      S:INTERVALO
S:SITIO          S:HORA
S:OBRA           T:DURACAO
S:OUTRO          T:FREQUENCIA
C:OBRA           T:GENERICO
T:REPRODUZIDA    #Tipos de referencia(R)
T:ARTE           R:ident
T:PLANO          R:incluido
T:OUTRO          R:inclui
C:ACONTECIMENTO  R:ocorre_em
T:EFEMERIDE      R:sede_de
T:ORGANIZADO     R:outro
```

a) Clicar duas vezes sobre a aplicação etiquet(h)arem.jar, ou, em alternativa,

b) Abrir explicitamente o programa numa consola: java –jar etiquetharem.jar

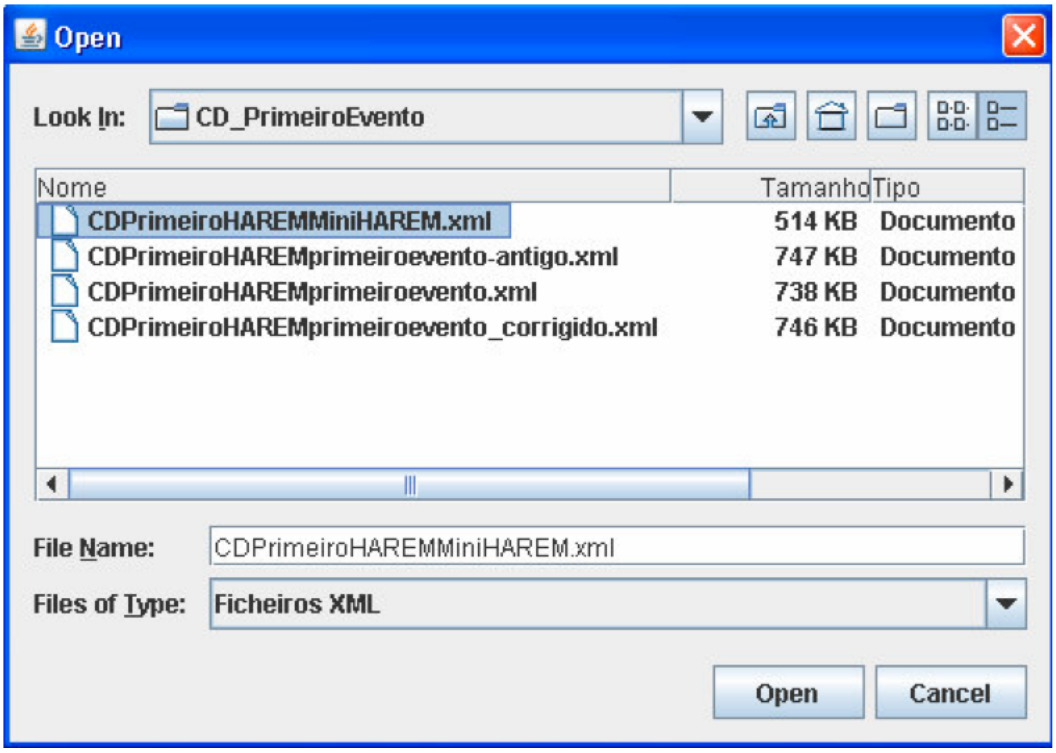


Figura F.1:

Obs: Ao abrir a aplicação, é imediatamente pedido para seleccionar o ficheiro a anotar (cf. figura F.1).

F.4 Menus do Etiquet(H)arem

(i) **Ficheiro**

ABRIR – Permite abrir um (novo) ficheiro.
GUARDAR – Permite gravar o ficheiro de trabalho.
GUARDAR COMO – Permite atribuir um novo nome ao ficheiro de trabalho.
TERMINAR – Permite sair da aplicação.

Obs: Sempre que um dado ficheiro é aberto, à frente de Documentos aparece uma lista que é preenchida com o `DOCID` de todos os documentos (DOC) do ficheiro. Inicialmente é mostrado o primeiro documento, mas é possível visualizar qualquer documento dessa listagem, através da selecção do `DOCID` correspondente (cf. figura F.2).

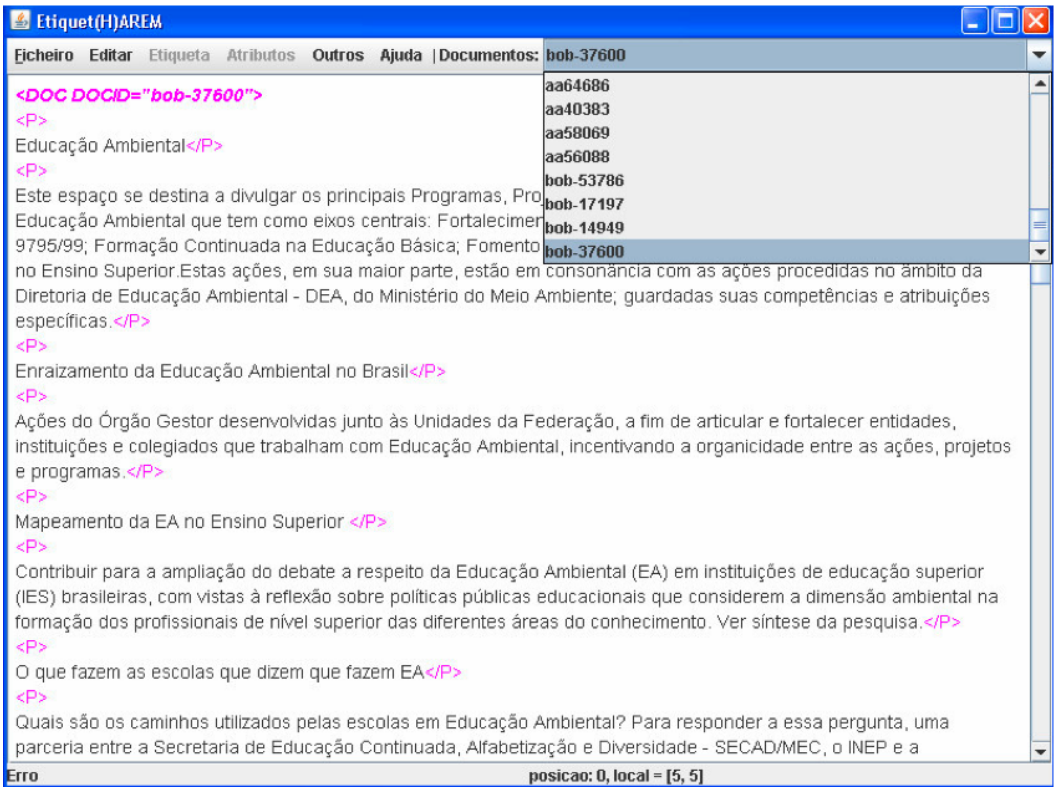


Figura F.2:

(ii) **Editar**

Os comandos compreendidos neste menu são idênticos aos utilizados na generalidade das aplicações.

ANULAR – permite anular uma operação. No entanto, uma operação para o programa pode não ser o mesmo do que uma operação para o utilizador; por exemplo, a anulação de uma etiqueta inserida implica a repetição do comando.

REPETIR - permite repetir a operação anulada pelo comando anterior.

CUT-TO-CLIPBOARD, COPY-TO-CLIPBOARD e PASTE-FROM-CLIPBOARD , comandos que permitem, respectivamente, cortar um fragmento do texto, copiar um fragmento do texto ou adicionar um fragmento ao texto.

(iii) **Etiqueta**

EM – Serve para atribuir uma etiqueta a uma palavra ou sequência de palavras previamente seleccionadas no texto (cf. figura F.3).

EM VAGA – Deve ser utilizado para etiquetar EMs que possam ser vagas entre 2 ou mais categorias, tipos e/ou subtipos. A vagueza é representada através do

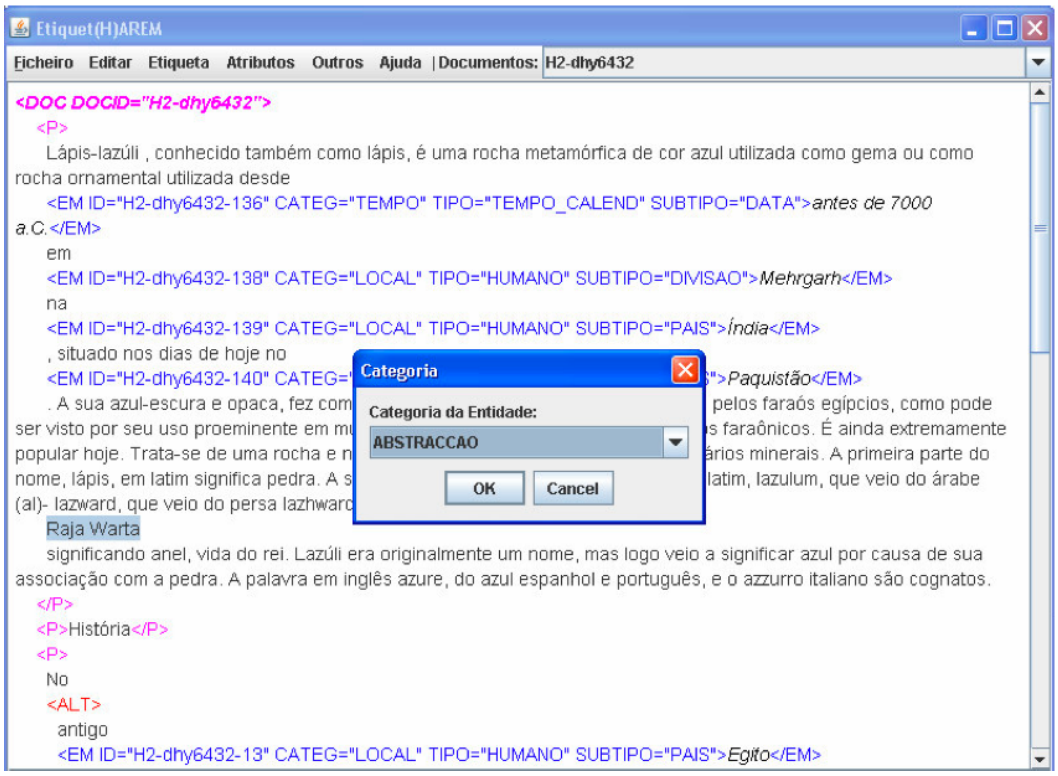


Figura F.3:

caracter "|". Ao seleccionar esta funcionalidade, o programa pede para escolher o índice de vagueza, ie., o número de etiquetas (ou interpretações) diferentes que a referida EM poderá receber (2, 3, 4, 5, 6).

EM ALTERNATIVA – Esta funcionalidade permite atribuir duas ou mais análises alternativas a uma mesma sequência de palavras previamente seleccionadas no texto. Neste caso, o programa repetirá o fragmento do texto seleccionado tantas vezes quanto o número de análises alternativas seleccionadas (2, 3, 4, 5, 6). As diferentes análises encontram-se separadas através do caracter "|", e o fragmento do texto onde existem análises alternativas está delimitado, à esquerda e à direita, pelas etiquetas `<ALT>` e `</ALT>`, respectivamente.

REPETIR, REMOVER e ALTERAR – Estes comandos permitem, respectivamente, repetir, remover ou alterar uma etiqueta previamente atribuída a uma dada EM. Para isso, basta seleccionar toda a etiqueta e proceder às alterações desejadas.

AUMENTAR VAGUEZA – Esta funcionalidade permite atribuir uma nova análise a uma EM previamente etiquetada no texto. Para isso, basta seleccionar toda a etiqueta associada a essa EM e introduzir os novos atributos desejados.

NOVA ALTERNATIVA – Esta funcionalidade permite introduzir uma nova análise alternativa a uma EM previamente etiquetada no texto com duas ou mais aná-

lises alternativas. Neste caso, o programa apenas reproduz um novo bloco de texto, sem qualquer notação, para posterior etiquetagem.

OMITIR – Esta funcionalidade permite marcar um fragmento de texto como "omitido", colocando-o entre as etiquetas `<OMITIDO>` `</OMITIDO>`. O texto omitido não será alvo de avaliação.

(iv) **Atributos**

Este menu serve fundamentalmente para adicionar nova informação a uma dada EM que já tenha sido anteriormente etiquetada.

CORRELAÇÃO – Permite inserir o tipo de relação que uma dada EM mantém com uma outra EM. Para isso é necessário seleccionar antes a EM e respectiva etiqueta. Será depois mostrada uma lista com todas as EMs já anotadas dentro do mesmo documento, de forma a que o utilizador possa escolher aquela com que existe a relação. Depois disso, será pedido que se seleccione o tipo de relação (cf. figura F.4).

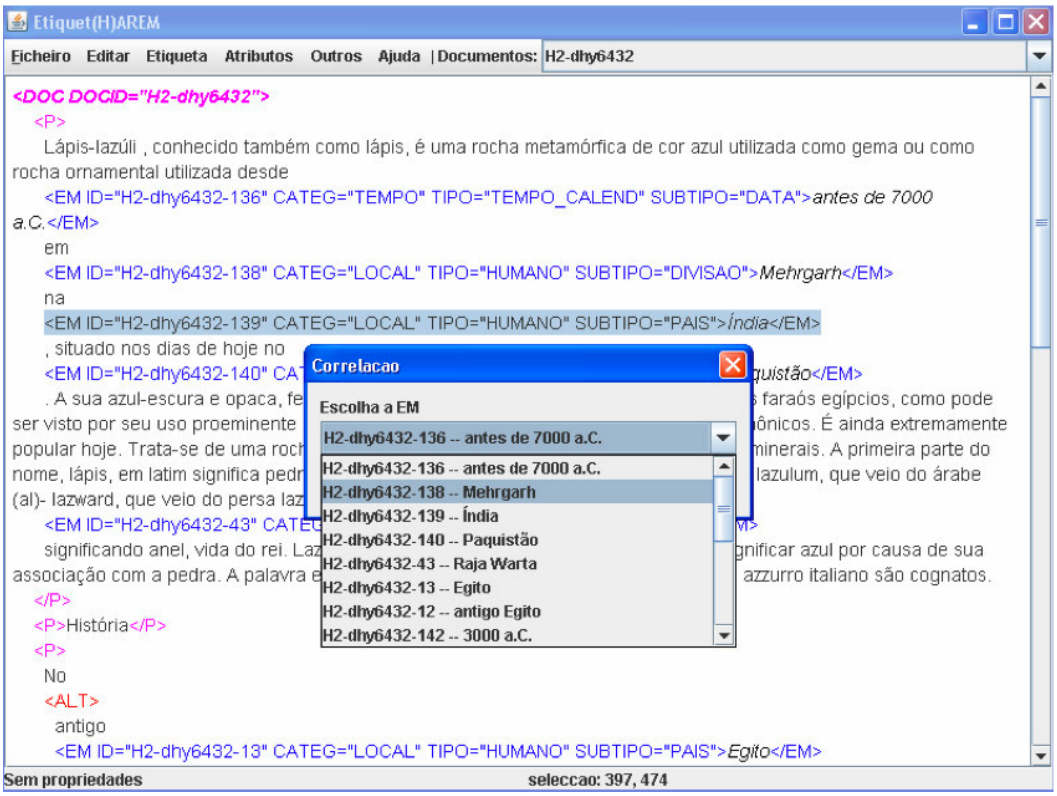


Figura F.4:

`TIPO` e `SUBTIPO` – Estas funcionalidades permitem inserir o `TIPO` e/ou `SUBTIPO` a uma EM do texto cuja etiqueta não contenha esses atributos. Para isso, é necessário

seleccionar antes a EM e respectiva etiqueta. Será depois mostrada uma lista com as possibilidades que estes campos podem ter.

TEMPO – Permite inserir os atributos TEMPO_REF e SENTIDO, relativos à categoria TEMPO tipo TEMPO_CALEND subtipo DATA.

COMENTÁRIO – Permite inserir o atributo comentário (COMENT) na EM. É necessário ter algum cuidado na sua utilização já que este atributo se pode inserir em qualquer parte do texto, sendo, no entanto, válido apenas quando se encontra dentro de uma etiqueta de EM. O atributo comentário pode ser utilizado pelo anotador para acrescentar algo à sua anotação, por exemplo, a indicação de que não tem a certeza se a mesma foi bem feita.

META ERRO – Trata-se de uma funcionalidade que, por enquanto, não está a ser usada. Foi implementada sobretudo para dar conta de (cf. Cardoso e Santos (2007)):

> casos em que há enganos de ortografia ou grafia no texto, em particular quando uma palvara tem uma maiúscula a mais ou a menos e tal é notório, escolhemos corrigir mentalmente a grafia (maiúscula /minúscula) de forma a poder classificar correctamente. Além disso, estamos a pensar em marcar estes casos, na colecção dourada, com uma classificação META="ERRO".
>
> Certo : O grupo terrorista <PESSOA TIPO="GRUPO" META="ERRO">Setembro negro</PESSOA>

(v) **Outros**

Este menu compreende uma série de comandos que envolvem a manipulação e apresentação do próprio texto: LOCALIZAR – Permite identificar uma palavra ou sequência de palavras no texto do documento que se está a visualizar.

MOSTRA ETIQUETAS e ESCONDE ETIQUETAS – permitem a visualização do texto com ou sem etiquetas, respectivamente.

VALIDAR XML – permite fazer uma validação do XML, tendo em conta (se existir) a DTD.

TAMANHO DA LETRA – permite aumentar ou diminuir o tamanho da letra do texto visualizado.

(vi) **Ajuda**

COMO ETIQUETAR – Explica os diferentes modos de atribuição de uma etiqueta ou atributo a uma dada EM no texto.

ACERCA – Dá a indicação do programa e respectiva versão que se está a utilizar.

Agradecimentos

Queremos agradecer à Diana Santos e à Cláudia Freitas as importantes sugestões a versões preliminares deste documento. Este trabalho foi desenvolvido no âmbito do projecto Linguateca contrato nº 339/1.3/C/NAC, financiado pelo governo português e pela União Europeia.

Apêndice G

SAHARA - Serviço de Avaliação HAREM Automático

Nuno Cardoso

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Apêndice G, p. 347–354.

Nota das editoras: Este apêndice foi preparado especialmente para o presente livro, depois de ter surgido a ideia no Encontro do Segundo HAREM de implementar um serviço de avaliação na rede. Este serviço, criado por Nuno Cardoso, ficou patente ao público desde 4 de Outubro de 2008.

O SAHARA, disponível em www.linguateca.pt/HAREM → Avaliador, é um serviço na rede que permite a avaliação imediata de saídas de sistemas de REM de acordo com o ambiente de avaliação usado no Segundo HAREM. O SAHARA facilita consideravelmente a avaliação de sistemas de REM, visto que elimina a necessidade de executar uma sucessão de comandos específicos de cada programa de avaliação para obter um conjunto de valores de desempenho. O SAHARA permite ainda a comparação imediata com os resultados oficiais do Segundo HAREM, bem como o acesso aos resultados de cada programa de avaliação, para depuração mais detalhada.

Uma avaliação no SAHARA decorre em três passos:

1. Validação da corrida enviada pelo utilizador, de acordo com o formato do Segundo HAREM
2. Configuração da avaliação pretendida, ou seja, a definição dos cenários, modos de avaliação, e colecções a serem utilizadas
3. Apresentação dos resultados, com um conjunto de tabelas e gráficos que resumem o desempenho do sistema.

G.1 Primeiro passo: validação

O SAHARA começa por pedir o ficheiro criado pelo sistema de REM (no formato específico de XML do Segundo HAREM), que pode encontrar-se no computador do utilizador, ou num dado URL (ver figura G.1). O avaliador aceita entradas comprimidas por ZIP ou GZIP, desde que incluam o ficheiro XML com o mesmo nome do ficheiro enviado. O formato desse ficheiro é então validado, invocando, entre outros testes, o validador do HAREM que recorre a regras de RelaxNG (van der Vlist, 2003). Só se não forem encontrados erros se avança para o passo da configuração. Caso contrário, o avaliador volta ao passo inicial, mostrando adicionalmente uma lista de erros devidamente localizados no XML com a indicação do número da linha e da coluna, para facilitar a depuração.

G.2 Configuração da avaliação

O formulário de configuração está dividido em quatro secções: i) Selecção das pistas, ii) Escolha dos cenários, iii) Selecção do modo de avaliação e iv) Escolha da colecção dourada, que serão descritas em seguida.

G.2.1 Selecção das pistas

A pista clássica do HAREM, descrita no capítulo 1, é obrigatória e está incluída em todas as opções da caixa de selecção (ver figura G.2). Contudo, o utilizador pode optar por

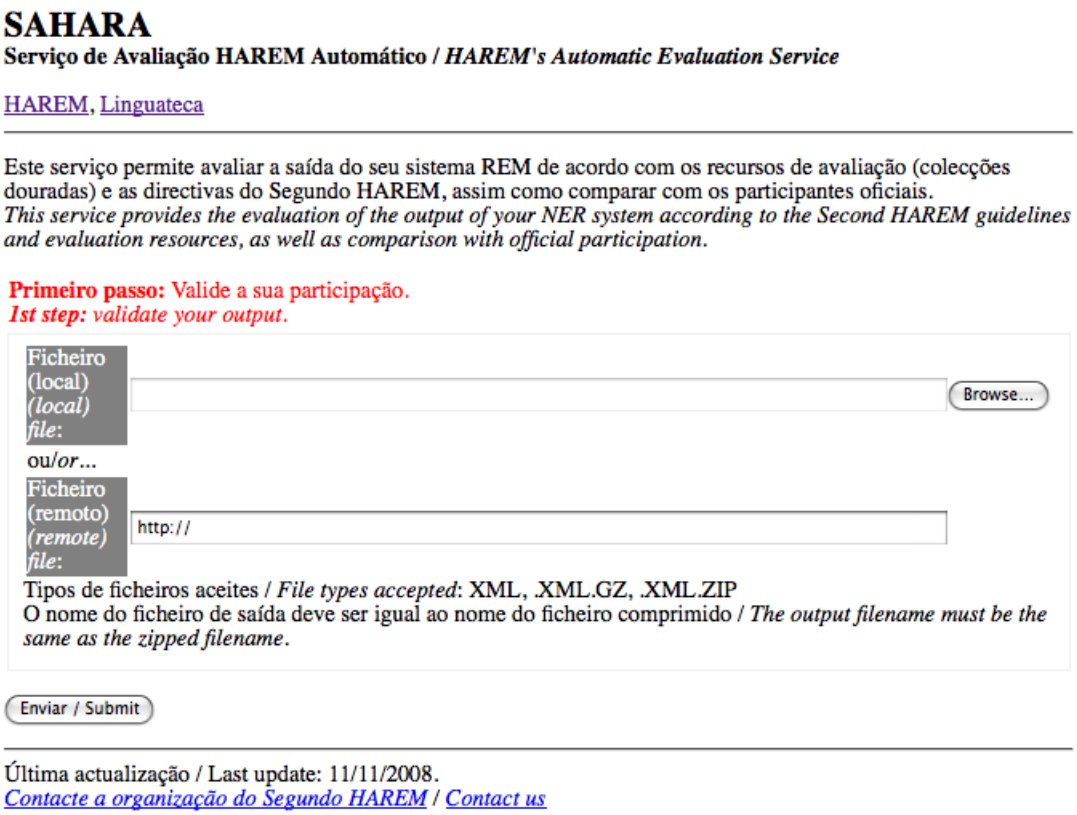


Figura G.1: Página de entrada do SAHARA.

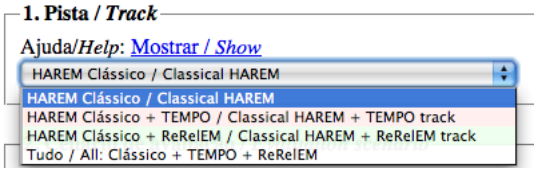


Figura G.2: Configuração da avaliação: selecção das pistas.

incluir também na avaliação as outras pistas: a pista do TEMPO e a pista do ReRelEM, descritas nos capítulos 2 e 4, respectivamente. Para o conjunto de pistas escolhidas, só é possível escolher uma única colecção dourada; para usar diferentes colecções douradas, é preferível realizar uma avaliação de cada vez.

A cada pista está associada uma cor para que seja mais fácil fazer corresponder as várias opções no formulário à pista respectiva.

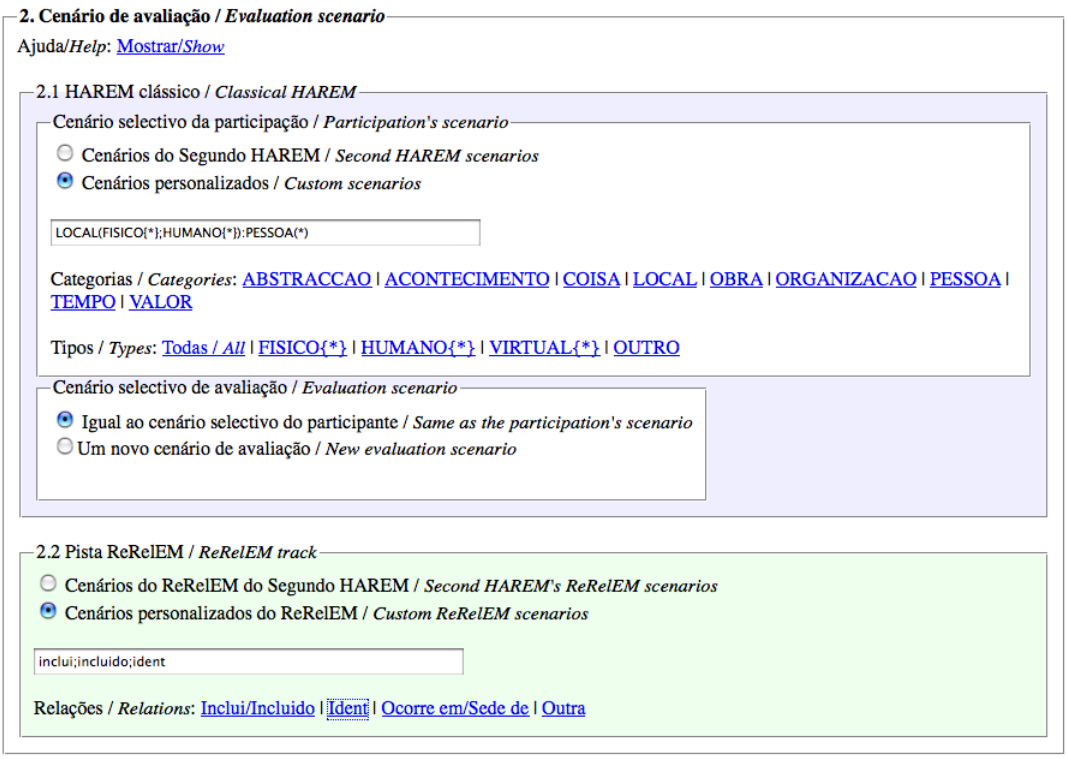


Figura G.3: Configuração da avaliação: Escolha dos cenários.

G.2.2 Escolha dos cenários

Os cenários de avaliação permitem escolher o leque de categorias a usar, tanto do lado do participante (Cenário selectivo da participação), como no lado da avaliação (Cenário selectivo de avaliação) – ver figura G.3. Do lado do participante, o formulário permite escolher entre cenários pré-definidos (usados para a geração de resultados oficiais do Segundo HAREM) e cenários personalizados. O utilizador pode especificar o leque de categorias, tipos e subtipos que quer ver avaliados. Do lado da avaliação, é possível escolher entre usar o mesmo cenário especificado no lado do participante, ou então outro cenário personalizado. A avaliação por cenários foi descrita nas secções 5.1.3 e 5.6.3.

Caso se opte por avaliar na pista do ReRelEM, aparece uma caixa para a escolha de cenários específicos para o ReRelEM, descritos na secção 5.8.6. O funcionamento é semelhante à escolha dos cenários para a pista clássica. O utilizador pode escolher cenários para o ReRelEM já usados no Segundo HAREM, ou então criar o seu próprio cenário personalizado.

G.2.3 Selecção do modo de avaliação

Os modos de avaliação são parâmetros avançados de configuração da avaliação (ver figura G.4). Para a pista clássica, o utilizador pode escolher a forma de tratamento das alternati-

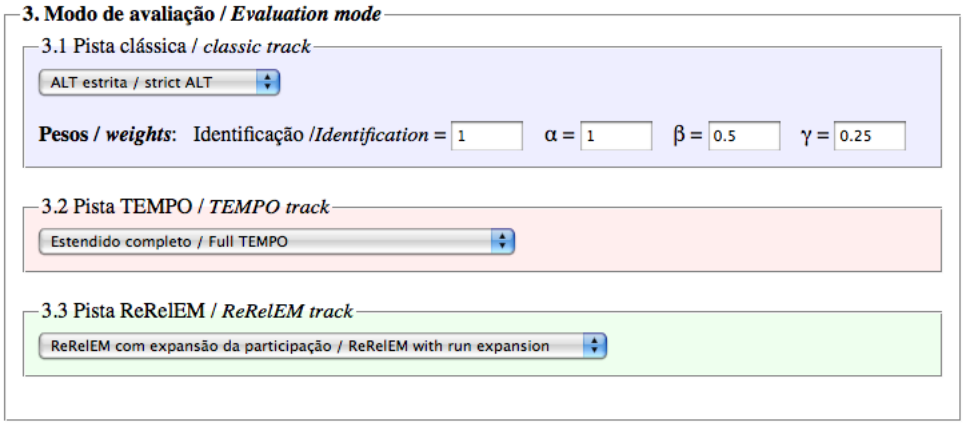


Figura G.4: Configuração da avaliação: selecção do modo de avaliação.

vas <ALT>, apresentada nas secções 5.1.4 e 5.6.7, entre a forma restrita e a forma relaxada, e também especificar os pesos na fórmula da medida de classificação.

Para a pista do TEMPO, o utilizador pode optar por usar o leque completo de categorização TEMPO (Estendido completo), escolhido por omissão, ou então optar pelo modo sem normalização ou pelo modo só com normalização. A avaliação da pista do TEMPO foi descrita nas secções 5.2 e 5.7.

Para a pista do ReRelEM, cuja avaliação foi apresentada nas secções 5.3 e 5.8, o utilizador pode escolher uma avaliação onde as relações são expandidas automaticamente (escolha por omissão), ou então realizar uma avaliação sem expansão de relações, do lado da corrida do participante.

G.2.4 Escolha da colecção dourada

A última secção do formulário de configuração diz respeito à escolha da colecção dourada para a avaliação. As escolhas da colecção estão restritas (por enquanto) às colecções douradas usadas no Segundo HAREM e, consoante as pistas de avaliação escolhidas, ao pedaço da colecção dourada que possui anotações para as pistas seleccionadas.

A colecção dourada do Segundo HAREM possui 129 documentos, onde um subconjunto de 30 documentos foi anotado com todos os atributos da categoria TEMPO e, dentro desse subconjunto, 12 documentos foram anotados com relações para a pista do ReRelEM. Como tal, um utilizador que pretenda avaliar nas pistas clássica + TEMPO, pode escolher entre o subconjunto TEMPO ou o subconjunto ReRelEM, mas não pode escolher a colecção dourada completa.

G.3 Apresentação dos resultados

De acordo com as pistas seleccionadas e a configuração escolhida, o SAHARA desenha um plano de execução para a avaliação pretendida, e executa os diferentes programas em sequência, que pode ser visualizada (através de uma série de caixas) enquanto o utili-

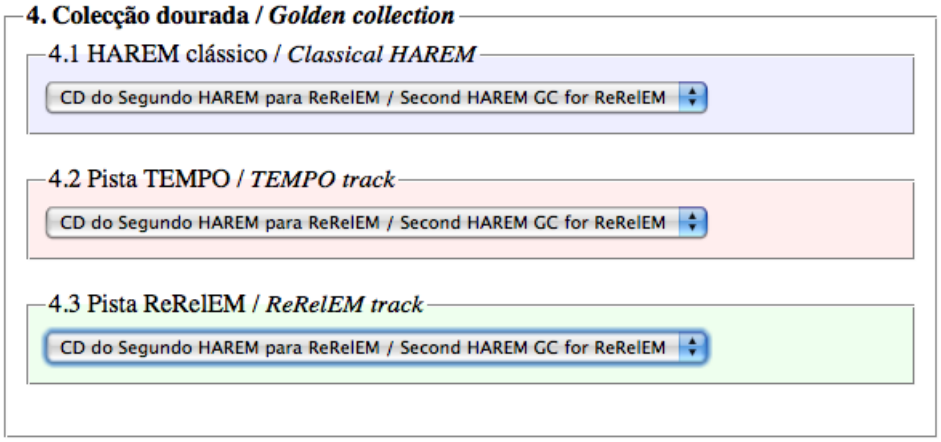


Figura G.5: Configuração da avaliação: escolha da colecção dourada.

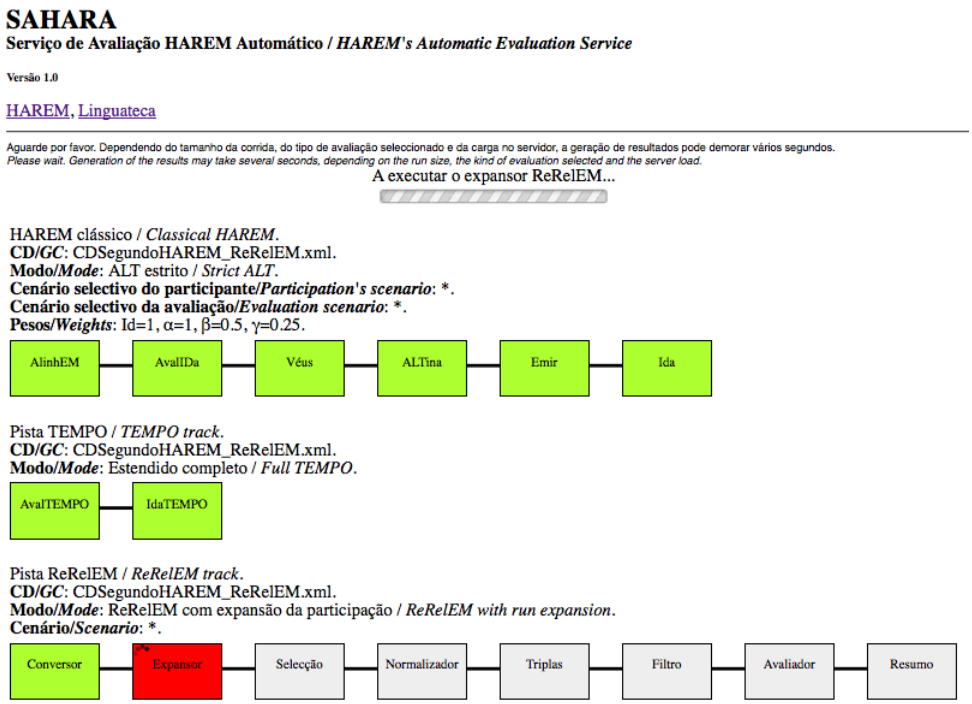


Figura G.6: Execução da avaliação.

zador espera pelos resultados, como ilustrado na figura G.6. Os parâmetros escolhidos na configuração são novamente apresentados, para que se possa rever o tipo de avaliação associada aos resultados a gerar.

Uma vez terminada a geração de resultados, o SAHARA verifica se há resultados oficiais que sejam comparáveis à avaliação realizada, isto é, se existem sistemas participantes no Segundo HAREM que tenham sido avaliados conforme o mesmo cenário de avaliação, o mesmo modo de avaliação e a mesma colecção dourada. (De qualquer maneira, note-se que os resultados oficiais foram gerados com os pesos por omissão, e que a alteração dos pesos desvirtua a comparação.) Se existirem resultados oficiais, o gráfico-resumo dos resultados para cada pista apresenta também os resultados para os três melhores sistemas (considerando apenas a melhor saída de cada sistema), como apresentado na figura G.7.

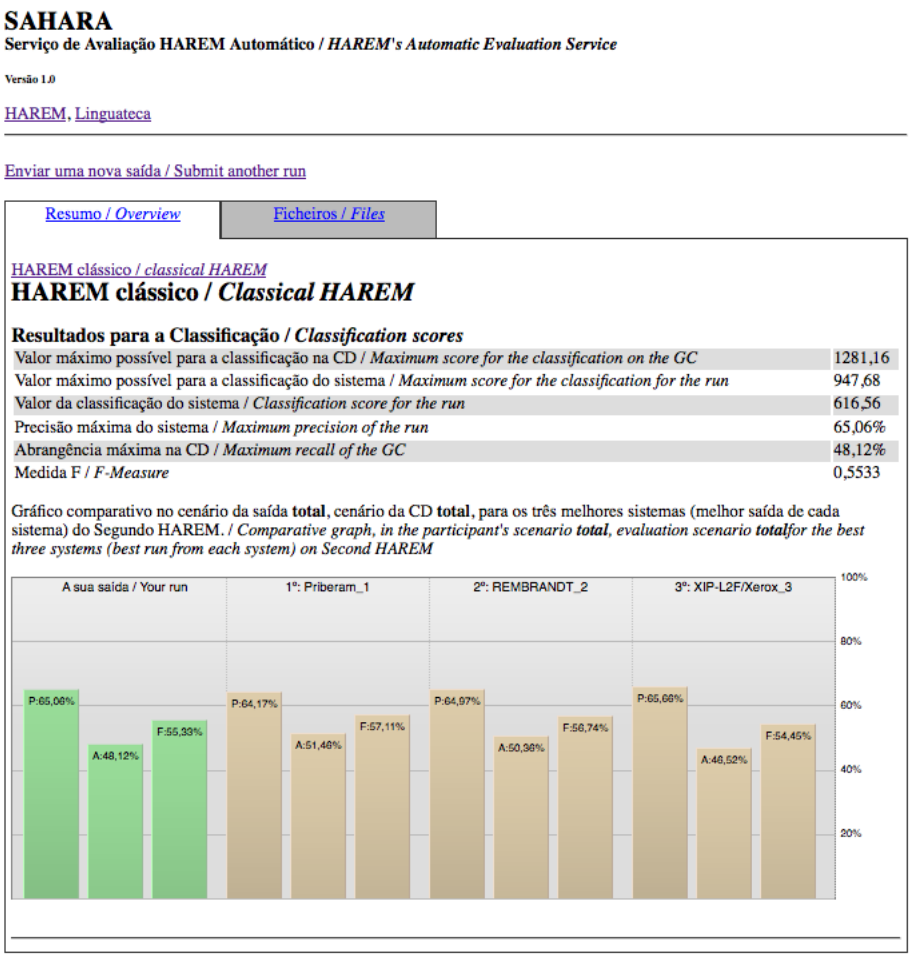


Figura G.7: Apresentação de resultados: tabela de resultados e gráfico comparativo.

Para uma análise mais detalhada dos resultados da avaliação, o utilizador pode escolher o separador *'Ficheiros / Files'* para ter acesso à saída de cada programa de avaliação invocado pelo SAHARA (ver figura G.8). Adicionalmente, é possível clicar em 'Depuração dos programas de avaliação' para voltar a ver a sequência de execução dos programas. Ao clicar em cada caixa do diagrama, o SAHARA mostra a linha de comandos executada, o que

é útil para quem pretenda ir buscar os programas de avaliação e executá-los numa linha de comandos.

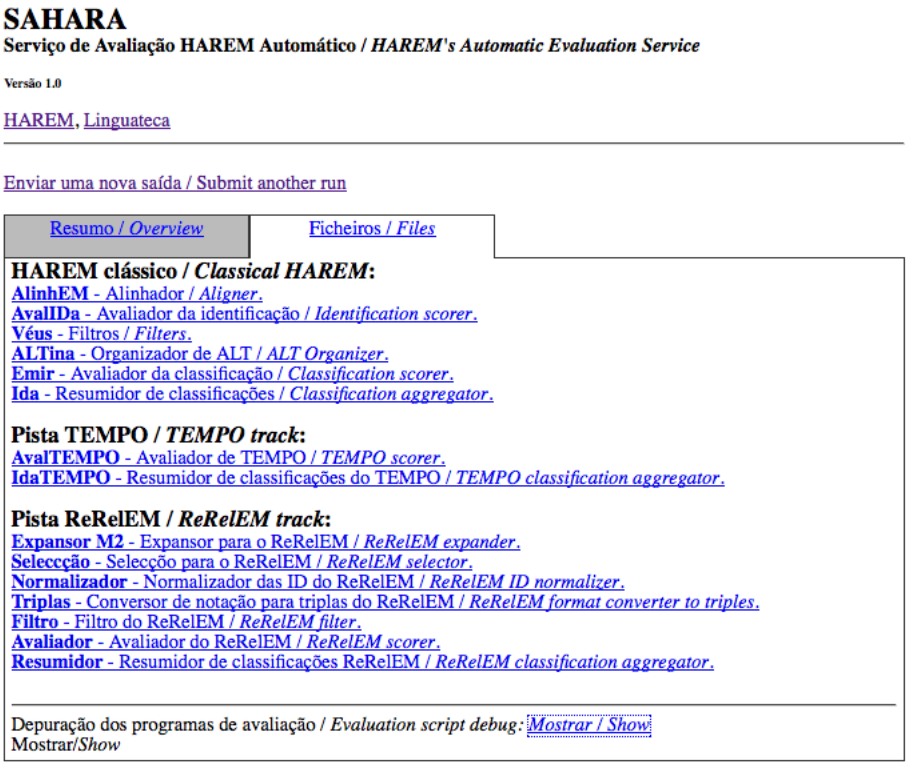


Figura G.8: Lista de ficheiros de saída dos programas de avaliação.

Apêndice H

Apresentação detalhada das colecções do Segundo HAREM

Cristina Mota, Diana Santos, Paula Carvalho, Cláudia Freitas e Hugo Gonçalo Oliveira

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Apêndice H, p. 355–377.

Nota das editoras: Este apêndice foi criado especialmente para o presente livro, de modo a complementar as informações sobre as colecções do Segundo HAREM fornecidas nos capítulos 1, 3 e 4.

Neste apêndice apresentamos em mais pormenor a constituição das várias colecções usadas no Segundo HAREM, segundo as seguintes vertentes: variante do português, tipo de texto e fonte do texto.

Para os três casos, apresentamos, na forma gráfica e por tabela, tanto a distribuição dos documentos, como a distribuição das palavras..

Começamos por detalhar a distribuição dos documentos e de palavras entre português do Brasil e de Portugal, nas figuras H.1 e H.2, respectivamente.

Convém no entanto recordar a forma de constituição da colecção do HAREM para explicar a razão da tripartição das figuras ao longo deste apêndice, nomeadamente distinguindo entre colecção HAREM total, colecção HAREM sem documentos da colecção CHAVE e colecção dourada. É que a colecção do HAREM foi construída como a soma da colecção dourada + os exemplos já apresentados e disponibilizados ao público + a CD do Primeiro HAREM + documentos obtidos da colecção CHAVE.

As duas parcelas intermédias foram incluídas para permitir mais tarde comparações entre o desempenho dos sistemas neste e no anterior HAREM, ou para comparar o desempenho entre material de treino e material novo.

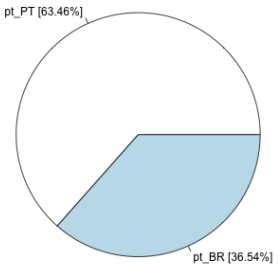
Mas do ponto de vista de distribuição, a colecção CHAVE é homogeneamente composta por documentos jornalísticos e faz sentido retirá-la por exemplo quando se está interessado nos diferentes géneros.

Depois apresentamos a distribuição de documentos e de palavras por tipo de texto, nas figuras H.3 e H.4, respectivamente. Notamos que os documentos podiam ser classificados com mais do que um tipo de texto. Assim, no caso dos documentos com mais de um tipo de texto, esse documento contribuiu com peso $1/n$, sendo n o número de classificações diferentes desse documento.

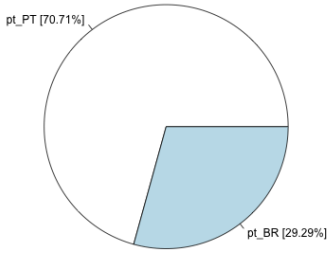
Salientamos que, relativamente à classificação por tipo de texto, não há um conjunto de géneros (ou tipos de texto) consensuais ou sequer amplamente utilizados em português (e mesmo em qualquer outra língua). Por isso de cada vez que é preciso caracterizar uma coleção de textos heterogénea deparamo-nos com problemas e com discordâncias sobre quer a grelha quer a granularidade. Isto pode aliás ver-se nas três avaliações conjuntas organizadas pela Linguateca, que usaram três bitolas diferentes (a das Morfolimpíadas, descrita em Costa et al. (2007), a do Primeiro HAREM, descrita em Rocha e Santos (2007b) e esta que apresentamos aqui). Convém também indicar que, pelo menos do nosso conhecimento, também existem as do Corpus NILC e do Lácio-Web (descrito em Pinheiro e Aluísio (2003); Aluisio et al. (2004) e usado em Aires (2005)) e as do Corpus do Português de Referência do CLUL (Bacelar do Nascimento et al., 2000).

Por último, apresentamos nas figuras H.5 e H.6 a origem dos textos.

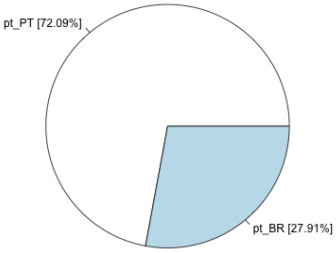
Finalmente, destacamos que todos os valores usados para criar estas tabelas constam do ficheiro meta distribuído na LÂMPADA, o pacote de recursos do Segundo HAREM.



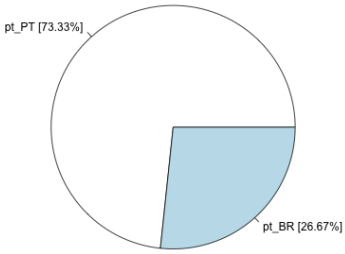
(a) Colecção do Segundo HAREM incluindo os documentos da colecção CHAVE



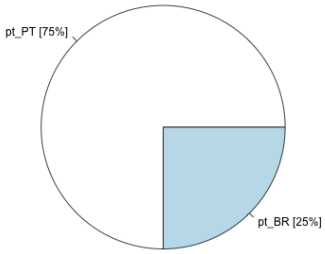
(b) Colecção do Segundo HAREM excluindo os documentos da colecção CHAVE



(c) Colecção dourada



(d) CD do TEMPO



(e) CD do ReRelEM

Figura H.1: Distribuição de documentos por variante de português

Tabela H.1: Colecção do Segundo HAREM: distribuição de documentos por variante de português, incluindo documentos da colecção CHAVE

Variante de português	Total	%
pt_PT	660	63,46%
pt_BR	380	36,54%

Tabela H.2: Colecção do Segundo HAREM: distribuição de documentos por variante de português, excluindo documentos da colecção CHAVE

Variante de português	Total	%
pt_PT	99	70,71%
pt_BR	41	29,29%

Tabela H.3: Colecção dourada: distribuição de documentos por variante de português

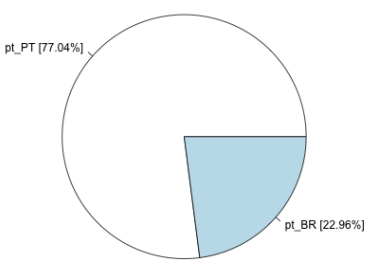
variante de português	Total	%
pt_PT	93	72,09%
pt_BR	36	27,91%

Tabela H.4: CD do TEMPO: distribuição de documentos por variante de português

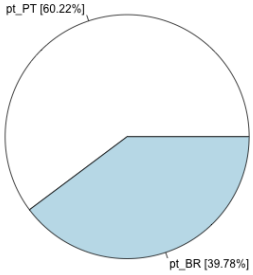
variante de português	Total	%
pt_PT	22	73,33%
pt_BR	8	26,67%

Tabela H.5: CD do ReRelEM: distribuição de documentos por variante de português

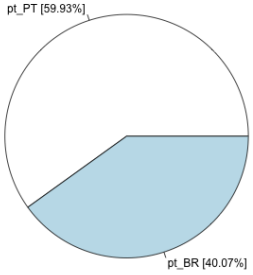
variante de português	Total	%
pt_PT	9	75%
pt_BR	3	25%



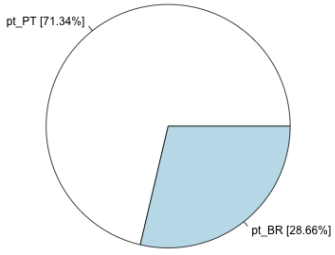
(a) Colecção do Segundo HAREM incluindo os documentos da colecção CHAVE



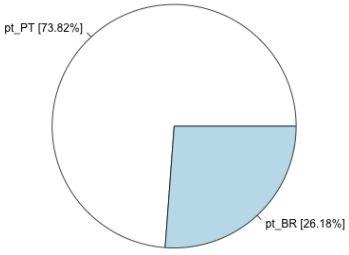
(b) Colecção do Segundo HAREM excluindo os documentos da colecção CHAVE



(c) Colecção dourada



(d) CD do TEMPO



(e) CD do ReRelEM

Figura H.2: Distribuição de palavras por variante de português

Tabela H.6: Colecção do Segundo HAREM: distribuição de palavras por variante de português, incluindo documentos da colecção CHAVE

Variante de português	Total	%
pt_PT	515264	77,04%
pt_BR	153553	22,96%

Tabela H.7: Colecção do Segundo HAREM: distribuição de palavras por variante de português, excluindo documentos da colecção CHAVE

Variante de português	Total	%
pt_PT	47615	60,22%
pt_BR	31448	39,78%

Tabela H.8: Colecção dourada: distribuição de palavras por variante de português

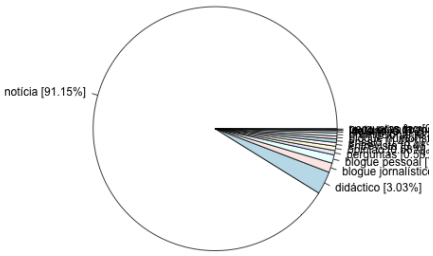
variante de português	Total	%
pt_PT	44555	59,93%
pt_BR	29795	40,07%

Tabela H.9: CD do TEMPO: distribuição de palavras por variante de português

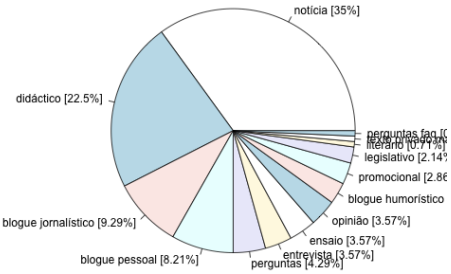
variante de português	Total	%
pt_PT	9268	71,34%
pt_BR	3724	28,66%

Tabela H.10: CD do ReRelEM: distribuição de palavras por variante de português

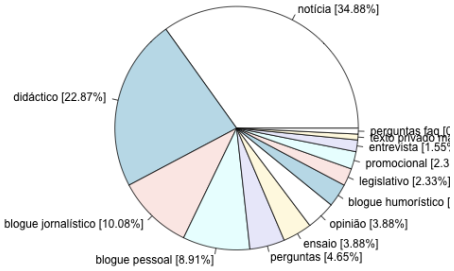
variante de português	Total	%
pt_PT	3271	73,82%
pt_BR	1160	26,18%



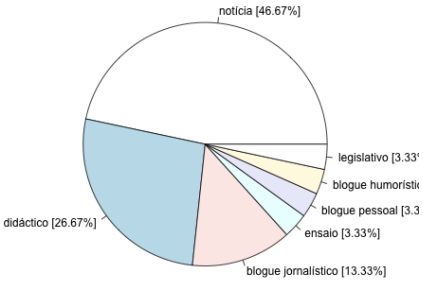
(a) Colecção do Segundo HAREM incluindo os documentos da colecção CHAVE



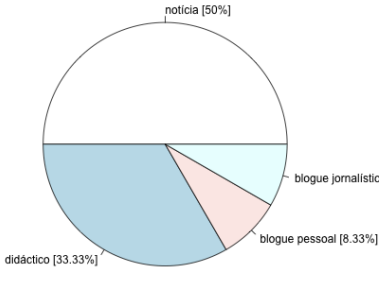
(b) Colecção do Segundo HAREM excluindo os documentos da colecção CHAVE



(c) Colecção dourada



(d) CD do TEMPO



(e) CD do ReRelEM

Figura H.3: Distribuição de documentos por tipo de texto

Tabela H.11: Colecção do Segundo HAREM: distribuição de documentos por tipo de texto, incluindo documentos da colecção CHAVE

Tipo de texto	Total	%
notícia	948	91,15%
didáctico	31,5	3,03%
blogue jornalístico	13	1,25%
blogue pessoal	11,5	1,11%
perguntas	6	0,58%
opinião	6	0,58%
entrevista	5	0,48%
ensaio	5	0,48%
blogue humorístico	4	0,38%
promocional	4	0,38%
legislativo	3	0,29%
literário	1	0,1%
texto privado manuscrito	1	0,1%
perguntas faq	1	0,1%

Tabela H.12: Colecção do Segundo HAREM: distribuição de documentos por tipo de texto, excluindo documentos da colecção CHAVE

Tipo de texto	Total	%
notícia	49	35%
didáctico	31,5	22,5%
blogue jornalístico	13	9,29%
blogue pessoal	11,5	8,21%
perguntas	6	4,29%
entrevista	5	3,57%
ensaio	5	3,57%
opinião	5	3,57%
blogue humorístico	4	2,86%
promocional	4	2,86%
legislativo	3	2,14%
literário	1	0,71%
texto privado manuscrito	1	0,71%
perguntas faq	1	0,71%

Tabela H.13: Colecção dourada: distribuição de documentos por tipo de texto

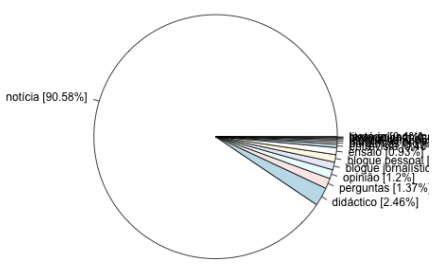
tipo de texto	Total	%
notícia	45	34,88%
didáctico	29,5	22,87%
blogue jornalístico	13	10,08%
blogue pessoal	11,5	8,91%
perguntas	6	4,65%
ensaio	5	3,88%
opinião	5	3,88%
blogue humorístico	4	3,1%
legislativo	3	2,33%
promocional	3	2,33%
entrevista	2	1,55%
texto privado manuscrito	1	0,78%
perguntas faq	1	0,78%

Tabela H.14: CD do TEMPO: distribuição de documentos por tipo de texto

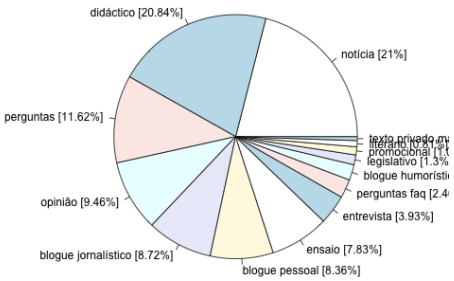
tipo de texto	Total	%
notícia	14	46,67%
didáctico	8	26,67%
blogue jornalístico	4	13,33%
ensaio	1	3,33%
blogue pessoal	1	3,33%
blogue humorístico	1	3,33%
legislativo	1	3,33%

Tabela H.15: CD do ReRelEM: distribuição de documentos por tipo de texto

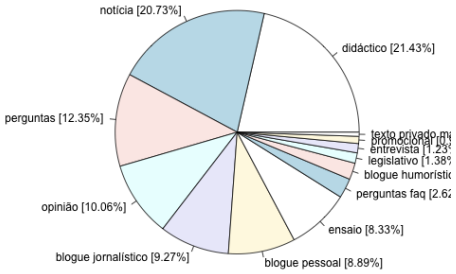
tipo de texto	Total	%
notícia	6	50%
didáctico	4	33,33%
blogue pessoal	1	8,33%
blogue jornalístico	1	8,33%



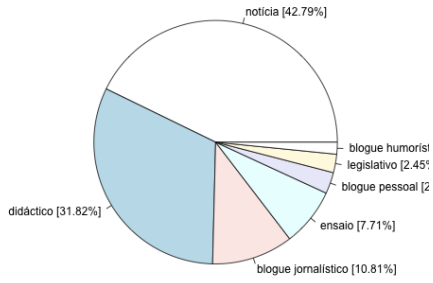
(a) Colecção do Segundo HAREM incluindo os documentos da colecção CHAVE



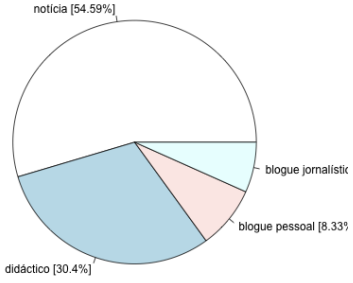
(b) Colecção do Segundo HAREM excluindo os documentos da colecção CHAVE



(c) Colecção dourada



(d) CD do TEMPO



(e) CD do ReRelEM

Figura H.4: Distribuição de palavras por tipo de texto

Tabela H.16: Colecção do Segundo HAREM: distribuição de palavras por tipo de texto, incluindo documentos da colecção CHAVE

Tipo de texto	Total	%
notícia	605815	90,58%
didáctico	16479,5	2,46%
perguntas	9184	1,37%
opinião	8023	1,2%
blogue jornalístico	6893	1,03%
blogue pessoal	6610,5	0,99%
ensaio	6193	0,93%
entrevista	3106	0,46%
perguntas faq	1945	0,29%
blogue humorístico	1639	0,25%
legislativo	1028	0,15%
promocional	850	0,13%
literário	644	0,1%
texto privado manuscrito	407	0,06%

Tabela H.17: Colecção do Segundo HAREM: distribuição de palavras por tipo de texto, excluindo documentos da colecção CHAVE

Tipo de texto	Total	%
notícia	16604	21%
didáctico	16479,5	20,84%
perguntas	9184	11,62%
opinião	7480	9,46%
blogue jornalístico	6893	8,72%
blogue pessoal	6610,5	8,36%
ensaio	6193	7,83%
entrevista	3106	3,93%
perguntas faq	1945	2,46%
blogue humorístico	1639	2,07%
legislativo	1028	1,3%
promocional	850	1,08%
literário	644	0,81%
texto privado manuscrito	407	0,51%

Tabela H.18: Colecção dourada: distribuição de palavras por tipo de texto

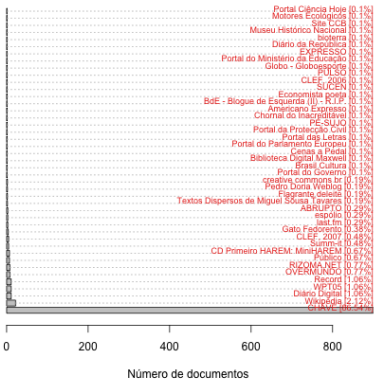
tipo de texto	Total	%
didáctico	15933,5	21,43%
notícia	15416	20,73%
perguntas	9184	12,35%
opinião	7480	10,06%
blogue jornalístico	6893	9,27%
blogue pessoal	6610,5	8,89%
ensaio	6193	8,33%
perguntas faq	1945	2,62%
blogue humorístico	1639	2,2%
legislativo	1028	1,38%
entrevista	916	1,23%
promocional	705	0,95%
texto privado manuscrito	407	0,55%

Tabela H.19: CD do TEMPO: distribuição de palavras por tipo de texto

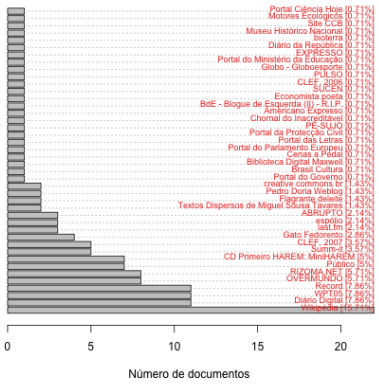
tipo de texto	Total	%
notícia	5559	42,79%
didáctico	4134	31,82%
blogue jornalístico	1404	10,81%
ensaio	1002	7,71%
blogue pessoal	369	2,84%
legislativo	318	2,45%
blogue humorístico	206	1,59%

Tabela H.20: CD do ReRelEM: distribuição de palavras por tipo de texto

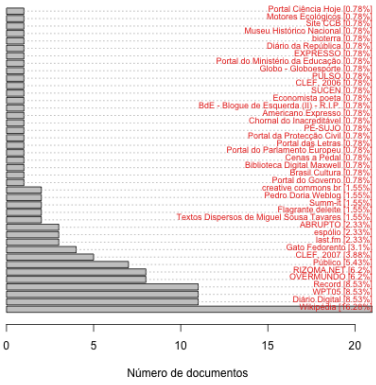
tipo de texto	Total	%
notícia	2419	54,59%
didáctico	1347	30,4%
blogue pessoal	369	8,33%
blogue jornalístico	296	6,68%



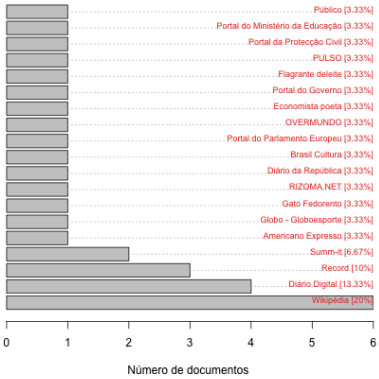
(a) Colecção do Segundo HAREM incluindo os documentos da colecção CHAVE



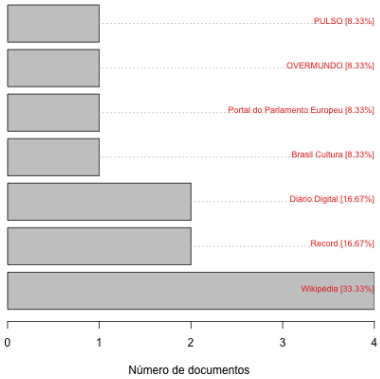
(b) Colecção do Segundo HAREM excluindo os documentos da colecção CHAVE



(c) Colecção dourada



(d) CD do TEMPO



(e) CD do ReRelEM

Figura H.5: Distribuição de documentos por fonte

Tabela H.21: Colecção do Segundo HAREM: distribuição de documentos por fonte, incluindo documentos da colecção CHAVE

Fonte	Total	%
CHAVE	900	86,54%
Wikipédia	22	2,12%
Diário Digital	11	1,06%
WPT05	11	1,06%
Record	11	1,06%
OVERMUNDO	8	0,77%
RIZOMA.NET	8	0,77%
Público	7	0,67%
CD Primeiro HAREM: MiniHAREM	7	0,67%
Summ-it	5	0,48%
CLEF, 2007	5	0,48%
Gato Fedorento	4	0,38%
last.fm	3	0,29%
espólio	3	0,29%
ABRUPTO	3	0,29%
Textos Dispersos de Miguel Sousa Tavares	2	0,19%
Flagrante deleite	2	0,19%
Pedro Doria Weblog	2	0,19%
creative commons br	2	0,19%
Portal do Governo	1	0,1%
Brasil Cultura	1	0,1%
Biblioteca Digital Maxwell	1	0,1%
Cenas a Pedal	1	0,1%
Portal do Parlamento Europeu	1	0,1%
Portal das Letras	1	0,1%
Portal da Protecção Civil	1	0,1%
PÉ-SUJO	1	0,1%
Chornal do Inacreditável	1	0,1%
Americano Expresso	1	0,1%
BdE - Blogue de Esquerda (II) - R.I.P.	1	0,1%
Economista poeta	1	0,1%
SUCEN	1	0,1%
CLEF, 2006	1	0,1%
PULSO	1	0,1%
Globo - Globoesporte	1	0,1%
Portal do Ministério da Educação	1	0,1%
EXPRESSO	1	0,1%
Diário da República	1	0,1%
bioterra	1	0,1%
Museu Histórico Nacional	1	0,1%
Site CCB	1	0,1%
Motores Ecológicos	1	0,1%
Portal Ciência Hoje	1	0,1%

Tabela H.22: Colecção do Segundo HAREM: distribuição de documentos por fonte, excluindo documentos da colecção CHAVE

Fonte	Total	%
Wikipédia	22	15,71%
Diário Digital	11	7,86%
WPT05	11	7,86%
Record	11	7,86%
OVERMUNDO	8	5,71%
RIZOMA.NET	8	5,71%
Público	7	5%
CD Primeiro HAREM: MiniHAREM	7	5%
Summ-it	5	3,57%
CLEF, 2007	5	3,57%
Gato Fedorento	4	2,86%
last.fm	3	2,14%
espólio	3	2,14%
ABRUPTO	3	2,14%
Textos Dispersos de Miguel Sousa Tavares	2	1,43%
Flagrante deleite	2	1,43%
Pedro Doria Weblog	2	1,43%
creative commons br	2	1,43%
Portal do Governo	1	0,71%
Brasil Cultura	1	0,71%
Biblioteca Digital Maxwell	1	0,71%
Cenas a Pedal	1	0,71%
Portal do Parlamento Europeu	1	0,71%
Portal das Letras	1	0,71%
Portal da Protecção Civil	1	0,71%
PÉ-SUJO	1	0,71%
Chornal do Inacreditável	1	0,71%
Americano Expresso	1	0,71%
BdE - Blogue de Esquerda (II) - R.I.P.	1	0,71%
Economista poeta	1	0,71%
SUCEN	1	0,71%
CLEF, 2006	1	0,71%
PULSO	1	0,71%
Globo - Globoesporte	1	0,71%
Portal do Ministério da Educação	1	0,71%
EXPRESSO	1	0,71%
Diário da República	1	0,71%
bioterra	1	0,71%
Museu Histórico Nacional	1	0,71%
Site CCB	1	0,71%
Motores Ecológicos	1	0,71%
Portal Ciência Hoje	1	0,71%

Tabela H.23: Colecção dourada: distribuição de documentos por fonte

fonte	Total	%
Wikipédia	21	16,28%
Diário Digital	11	8,53%
WPT05	11	8,53%
Record	11	8,53%
OVERMUNDO	8	6,2%
RIZOMA.NET	8	6,2%
Público	7	5,43%
CLEF, 2007	5	3,88%
Gato Fedorento	4	3,1%
last.fm	3	2,33%
espólio	3	2,33%
ABRUPTO	3	2,33%
Textos Dispersos de Miguel Sousa Tavares	2	1,55%
Flagrante deleite	2	1,55%
Summ-it	2	1,55%
Pedro Doria Weblog	2	1,55%
creative commons br	2	1,55%
Portal do Governo	1	0,78%
Brasil Cultura	1	0,78%
Biblioteca Digital Maxwell	1	0,78%
Cenas a Pedal	1	0,78%
Portal do Parlamento Europeu	1	0,78%
Portal das Letras	1	0,78%
Portal da Protecção Civil	1	0,78%
PÉ-SUJO	1	0,78%
Chornal do Inacreditável	1	0,78%
Americano Expresso	1	0,78%
BdE - Blogue de Esquerda (II) - R.I.P.	1	0,78%
Economista poeta	1	0,78%
SUCEN	1	0,78%
CLEF, 2006	1	0,78%
PULSO	1	0,78%
Globo - Globoesporte	1	0,78%
Portal do Ministério da Educação	1	0,78%
EXPRESSO	1	0,78%
Diário da República	1	0,78%
bioterra	1	0,78%
Museu Histórico Nacional	1	0,78%
Site CCB	1	0,78%
Motores Ecológicos	1	0,78%
Portal Ciência Hoje	1	0,78%

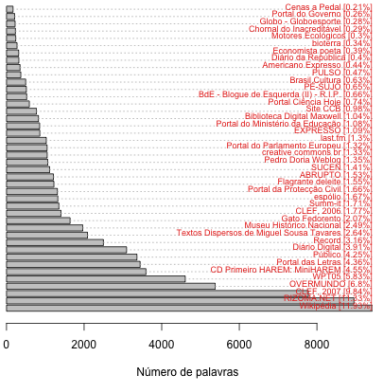
Tabela H.24: CD do TEMPO: distribuição de documentos por fonte

fonte	Total	%
Wikipédia	6	20%
Diário Digital	4	13,33%
Record	3	10%
Summ-it	2	6,67%
Americano Expresso	1	3,33%
Globo - Globoesporte	1	3,33%
Gato Fedorento	1	3,33%
RIZOMA.NET	1	3,33%
Diário da República	1	3,33%
Brasil Cultura	1	3,33%
Portal do Parlamento Europeu	1	3,33%
OVERMUNDO	1	3,33%
Economista poeta	1	3,33%
Portal do Governo	1	3,33%
Flagrante deleite	1	3,33%
PULSO	1	3,33%
Portal da Protecção Civil	1	3,33%
Portal do Ministério da Educação	1	3,33%
Público	1	3,33%

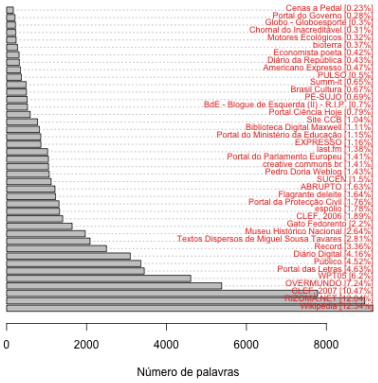
Tabela H.25: CD do ReRelEM: distribuição de documentos por fonte

fonte	Total	%
Wikipédia	4	33,33%
Record	2	16,67%
Diário Digital	2	16,67%
Brasil Cultura	1	8,33%
Portal do Parlamento Europeu	1	8,33%
OVERMUNDO	1	8,33%
PULSO	1	8,33%

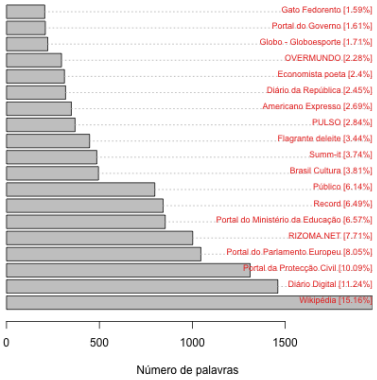
(a) Colecção do Segundo HAREM incluindo os documentos da colecção CHAVE



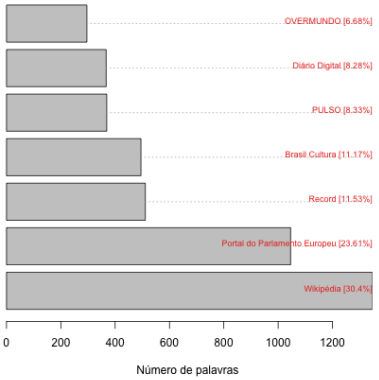
(b) Colecção do Segundo HAREM excluindo os documentos da colecção CHAVE



(c) Colecção dourada



(d) CD do TEMPO



(e) CD do ReRelEM

Figura H.6: Distribuição de palavras por fonte

Tabela H.26: Colecção do Segundo HAREM: distribuição de palavras por fonte, incluindo documentos da colecção CHAVE

Fonte	Total	%
CHAVE	589754	88,18%
Wikipédia	9430	1,41%
RIZOMA.NET	8955	1,34%
CLEF, 2007	7781	1,16%
OVERMUNDO	5380	0,8%
WPT05	4607	0,69%
CD Primeiro HAREM: MiniHAREM	3594	0,54%
Portal das Letras	3444	0,51%
Público	3361	0,5%
Diário Digital	3094	0,46%
Record	2498	0,37%
Textos Dispersos de Miguel Sousa Tavares	2087	0,31%
Museu Histórico Nacional	1966	0,29%
Gato Fedorento	1639	0,25%
CLEF, 2006	1403	0,21%
Summ-it	1350	0,2%
espólio	1323	0,2%
Portal da Protecção Civil	1311	0,2%
Flagrante deleite	1222	0,18%
ABRUPTO	1209	0,18%
SUCEN	1113	0,17%
Pedro Doria Weblog	1064	0,16%
creative commons br	1049	0,16%
Portal do Parlamento Europeu	1046	0,16%
last.fm	1026	0,15%
EXPRESSO	859	0,13%
Portal do Ministério da Educação	854	0,13%
Biblioteca Digital Maxwell	823	0,12%
Site CCB	772	0,12%
Portal Ciência Hoje	589	0,09%
BdE - Blogue de Esquerda (II) - R.I.P.	521	0,08%
PÉ-SUJO	512	0,08%
Brasil Cultura	495	0,07%
PULSO	369	0,06%
Americano Expresso	349	0,05%
Diário da República	318	0,05%
Economista poeta	312	0,05%
bioterra	272	0,04%
Motores Ecológicos	237	0,04%
Chornal do Inacreditável	230	0,03%
Globo - Globoesporte	222	0,03%
Portal do Governo	209	0,03%
Cenas a Pedal	168	0,03%

Tabela H.27: Colecção do Segundo HAREM : distribuição de palavras por fonte, excluindo documentos da colecção CHAVE

Fonte	Total	%
Wikipédia	9430	11,93%
RIZOMA.NET	8955	11,33%
CLEF, 2007	7781	9,84%
OVERMUNDO	5380	6,8%
WPT05	4607	5,83%
CD Primeiro HAREM: MiniHAREM	3594	4,55%
Portal das Letras	3444	4,36%
Público	3361	4,25%
Diário Digital	3094	3,91%
Record	2498	3,16%
Textos Dispersos de Miguel Sousa Tavares	2087	2,64%
Museu Histórico Nacional	1966	2,49%
Gato Fedorento	1639	2,07%
CLEF, 2006	1403	1,77%
Summ-it	1350	1,71%
espólio	1323	1,67%
Portal da Protecção Civil	1311	1,66%
Flagrante deleite	1222	1,55%
ABRUPTO	1209	1,53%
SUCEN	1113	1,41%
Pedro Doria Weblog	1064	1,35%
creative commons br	1049	1,33%
Portal do Parlamento Europeu	1046	1,32%
last.fm	1026	1,3%
EXPRESSO	859	1,09%
Portal do Ministério da Educação	854	1,08%
Biblioteca Digital Maxwell	823	1,04%
Site CCB	772	0,98%
Portal Ciência Hoje	589	0,74%
BdE - Blogue de Esquerda (II) - R.I.P.	521	0,66%
PÉ-SUJO	512	0,65%
Brasil Cultura	495	0,63%
PULSO	369	0,47%
Americano Expresso	349	0,44%
Diário da República	318	0,4%
Economista poeta	312	0,39%
bioterra	272	0,34%
Motores Ecológicos	237	0,3%
Chornal do Inacreditável	230	0,29%
Globo - Globoesporte	222	0,28%
Portal do Governo	209	0,26%
Cenas a Pedal	168	0,21%

Tabela H.28: Colecção dourada: distribuição de palavras por fonte

fonte	Total	%
Wikipédia	9175	12,34%
RIZOMA.NET	8955	12,04%
CLEF, 2007	7781	10,47%
OVERMUNDO	5380	7,24%
WPT05	4607	6,2%
Portal das Letras	3444	4,63%
Público	3361	4,52%
Diário Digital	3094	4,16%
Record	2498	3,36%
Textos Dispersos de Miguel Sousa Tavares	2087	2,81%
Museu Histórico Nacional	1966	2,64%
Gato Fedorento	1639	2,2%
CLEF, 2006	1403	1,89%
espólio	1323	1,78%
Portal da Protecção Civil	1311	1,76%
Flagrante deleite	1222	1,64%
ABRUPTO	1209	1,63%
SUCEN	1113	1,5%
Pedro Doria Weblog	1064	1,43%
creative commons br	1049	1,41%
Portal do Parlamento Europeu	1046	1,41%
last.fm	1026	1,38%
EXPRESSO	859	1,16%
Portal do Ministério da Educação	854	1,15%
Biblioteca Digital Maxwell	823	1,11%
Site CCB	772	1,04%
Portal Ciência Hoje	589	0,79%
BdE - Blogue de Esquerda (II) - R.I.P.	521	0,7%
PÉ-SUJO	512	0,69%
Brasil Cultura	495	0,67%
Summ-it	486	0,65%
PULSO	369	0,5%
Americano Expresso	349	0,47%
Diário da República	318	0,43%
Economista poeta	312	0,42%
bioterra	272	0,37%
Motores Ecológicos	237	0,32%
Chornal do Inacreditável	230	0,31%
Globo - Globoesporte	222	0,3%
Portal do Governo	209	0,28%
Cenas a Pedal	168	0,23%

Tabela H.29: CD do TEMPO: distribuição de palavras por fonte

fonte	Total	%
Wikipédia	1969	15,16%
Diário Digital	1460	11,24%
Portal da Protecção Civil	1311	10,09%
Portal do Parlamento Europeu	1046	8,05%
RIZOMA.NET	1002	7,71%
Portal do Ministério da Educação	854	6,57%
Record	843	6,49%
Público	798	6,14%
Brasil Cultura	495	3,81%
Summ-it	486	3,74%
Flagrante deleite	447	3,44%
PULSO	369	2,84%
Americano Expresso	349	2,69%
Diário da República	318	2,45%
Economista poeta	312	2,4%
OVERMUNDO	296	2,28%
Globo - Globoesporte	222	1,71%
Portal do Governo	209	1,61%
Gato Fedorento	206	1,59%

Tabela H.30: CD do ReRelEM: distribuição de palavras por fonte

fonte	Total	%
Wikipédia	1347	30,4%
Portal do Parlamento Europeu	1046	23,61%
Record	511	11,53%
Brasil Cultura	495	11,17%
PULSO	369	8,33%
Diário Digital	367	8,28%
OVERMUNDO	296	6,68%

Apêndice I

Resumo de resultados do Segundo HAREM

Cristina Mota, Hugo Gonçalo Oliveira, Diana Santos, Paula Carvalho e Cláudia Freitas

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Apêndice I, p. 379–403.

Nota das editoras: Este capítulo destaca os principais resultados do Segundo HAREM, actualizados pela última vez no dia 24 de Junho de 2008, no caso do HAREM clássico e da pista do TEMPO, e no dia 4 de Setembro de 2008, no caso da pista do ReRelEM.

I.1 Resultados do HAREM clássico

I.1.1 Avaliação estrita de ALT

Tabela I.1: Classificação no cenário total

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
Priberam_1	0,6417	**0,5146**	**0,5711**	17767,0257	14246,3355
REMBRANDT_2	0,6497	0,5036	0,5674	17767,0257	13772,0168
REMBRANDT_3_corr	0,6286	0,5032	0,5590	17767,0257	14223,9639
XIP-L2F/Xerox_3	0,6566	0,4652	0,5445	17767,0257	12587,5043
REMBRANDT_1	0,6396	0,4690	0,5412	17767,0257	13026,8739
XIP-L2F/Xerox_no	0,6586	0,4393	0,5270	17767,0257	11851,1561
XIP-L2F/Xerox_4	0,6544	0,4363	0,5236	17767,0257	11847,6498
XIP-L2F/Xerox_2	0,6404	0,4252	0,5111	17767,0257	11798,3965
XIP-L2F/Xerox_1	0,5877	0,3776	0,4598	17767,0257	11415,1748
REMMA_1_corr	0,6050	0,3615	0,4526	17767,0257	10615,7687
REMMA_2_corr	0,6132	0,2952	0,3985	17767,0257	8552,5854
R3M_1	0,7644	0,2520	0,3790	17767,0257	5857,0000
SeRELeP_1	0,8178	0,2415	0,3729	17767,0257	5247,5000
SeRELeP_no	**0,8180**	0,2244	0,3521	17767,0257	4873,1667
R3M_2	0,8029	0,2194	0,3447	17767,0257	4855,5000
Cage2_4_corr	0,4499	0,2757	0,3419	17767,0257	10885,9950
Cage2_1_corr	0,4523	0,2732	0,3406	17767,0257	10731,0998
Cage2_2_corr	0,4466	0,2753	0,3406	17767,0257	10952,0784
REMMA_3_corr	0,5808	0,2316	0,3312	17767,0257	7085,7250
Cage2_3_corr	0,4059	0,2348	0,2975	17767,0257	10277,5325
SEIGeo_4	0,7485	0,1166	0,2017	17767,0257	2766,6482
SEIGeo_3	0,7567	0,1157	0,2007	17767,0257	2717,0839
SEIGeo_2	0,7576	0,1111	0,1938	17767,0257	2605,7577
PorTexTO_4_corr	0,6790	0,0882	0,1562	17767,0257	2309,0729
PorTexTO_3_corr	0,6794	0,0881	0,1560	17767,0257	2304,0979
PorTexTO_2_corr	0,6770	0,0865	0,1533	17767,0257	2269,2729
PorTexTO_1_corr	0,6925	0,0823	0,1472	17767,0257	2112,5604
SEIGeo_1	0,5866	0,0545	0,0998	17767,0257	1651,3804
DobrEM_1_corr	0,4530	0,0073	0,0144	17767,0257	287

Tabela I.2: Identificação no cenário total

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
Priberam_1	0,6994	**0,7229**	**0,2771**	7255,6667	7499,1667
SeRELeP_1	0,8178	0,5915	0,4085	7255,6667	5247,5000
R3M_1	0,7644	0,6170	0,3830	7255,6667	5857,0000
REMBRANDT_2	0,7577	0,6214	0,3786	7255,6667	5951,0000
REMBRANDT_3_corr	0,7457	0,5962	0,4038	7255,6667	5801,1667
SeRELeP_no	0,8180	0,5494	0,4506	7255,6667	4873,1667
R3M_2	0,8029	0,5373	0,4627	7255,6667	4855,5000
REMBRANDT_1	0,7545	0,5527	0,4473	7255,6667	5315,1667
XIP-L2F/Xerox_3	0,7214	0,5315	0,4685	7255,6667	5345,3333
XIP-L2F/Xerox_no	0,7226	0,5026	0,4974	7255,6667	5046,8333
XIP-L2F/Xerox_4	0,7182	0,4994	0,5006	7255,6667	5045,8333
XIP-L2F/Xerox_2	0,7023	0,4841	0,5159	7255,6667	5001,5000
REMMA_1_corr	0,7083	0,4516	0,5484	7255,6667	4625,3333
XIP-L2F/Xerox_1	0,6601	0,4414	0,5586	7255,6667	4851,8333
REMMA_2_corr	0,6756	0,3467	0,6533	7255,6667	3723,5000
Cage2_1_corr	0,5108	0,3773	0,6227	7255,6667	5359,6667
Cage2_4_corr	0,5059	0,3787	0,6213	7255,6667	5431,6667
Cage2_2_corr	0,5027	0,3791	0,6209	7255,6667	5471,6667
REMMA_3_corr	0,7176	0,3061	0,6939	7255,6667	3094,8333
Cage2_3_corr	0,4529	0,3195	0,6805	7255,6667	5118,1667
SEIGeo_4	0,8963	0,1358	0,8642	7255,6667	1099,5000
SEIGeo_3	**0,9056**	0,1349	0,8651	7255,6667	1080,5000
SEIGeo_2	0,9043	0,1289	0,8711	7255,6667	1034
PorTexTO_4_corr	0,7003	0,0898	0,9102	7255,6667	930,8333
PorTexTO_3_corr	0,7007	0,0897	0,9103	7255,6667	928,8333
PorTexTO_2_corr	0,6983	0,0880	0,9120	7255,6667	914,8333
PorTexTO_1_corr	0,7147	0,0839	0,9161	7255,6667	851,8333
SEIGeo_1	0,7623	0,0672	0,9328	7255,6667	639,5000
DobrEM_1_corr	0,4530	0,0179	0,9821	7255,6667	287

Tabela I.3: Classificação no cenário 2 (LOCAL Físico e Humano + PESSOA, ORGANIZACAO e TEMPO sem tipos)

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
XIP-L2F/Xerox_3	0,7159	0,5666	**0,6326**	10489,2103	8301,3512
REMBRANDT_2	0,6821	0,5675	0,6195	10489,2103	8727,1012
REMBRANDT_3_corr	0,6758	0,5672	0,6168	10489,2103	8804,4286
XIP-L2F/Xerox_no	0,7205	0,5236	0,6065	10489,2103	7623,1905
XIP-L2F/Xerox_4	0,7202	0,5231	0,6060	10489,2103	7619,6905
XIP-L2F/Xerox_2	0,7113	0,5189	0,6001	10489,2103	7651,9226
REMBRANDT_1	0,6881	0,5288	0,5980	10489,2103	8061,0208
Priberam_1	0,5920	**0,5893**	0,5907	10489,2103	10441,8720
XIP-L2F/Xerox_1	0,6511	0,4529	0,5342	10489,2103	7295,8988
REMMA_1_corr	0,6344	0,4243	0,5085	10489,2103	7016,0833
REMMA_2_corr	0,6625	0,3418	0,4510	10489,2103	5412,5833
SeRELeP_1	0,6493	0,3273	0,4352	10489,2103	5287,1667
R3M_1	0,6020	0,3384	0,4333	10489,2103	5897,0000
Cage2_4_corr	0,4264	0,4070	0,4164	10489,2103	10012,6786
SeRELeP_no	0,6531	0,3055	0,4163	10489,2103	4906,8333
Cage2_1_corr	0,4277	0,4025	0,4148	10489,2103	9871,3095
Cage2_2_corr	0,4226	0,4059	0,4141	10489,2103	10074,4286
R3M_2	0,6424	0,2995	0,4085	10489,2103	4890,5000
REMMA_3_corr	0,6146	0,2739	0,3789	10489,2103	4674,0417
Cage2_3_corr	0,3883	0,3500	0,3682	10489,2103	9455,5744
SEIGeo_4	0,7260	0,1707	0,2763	10489,2103	2465,7262
SEIGeo_3	0,7339	0,1695	0,2753	10489,2103	2421,8869
SEIGeo_2	**0,7348**	0,1627	0,2664	10489,2103	2322,4732
PorTexTO_4_corr	0,6871	0,1077	0,1862	10489,2103	1644,1250
PorTexTO_3_corr	0,6875	0,1075	0,1860	10489,2103	1640,6250
PorTexTO_2_corr	0,6849	0,1055	0,1829	10489,2103	1616,1250
PorTexTO_1_corr	0,7002	0,1005	0,1758	10489,2103	1505,8750
SEIGeo_1	0,5568	0,0780	0,1369	10489,2103	1470,1042
DobrEM_1_corr	0,4425	0,0121	0,0236	10489,2103	287

Tabela I.4: Classificação no cenário 3 (Só identificação sem TEMPO e VALOR)

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
Priberam_1	0,7264	**0,5641**	**0,6351**	13961,3903	10843,3876
REMBRANDT_2	0,6845	0,5068	0,5824	13961,3903	10336,4543
REMBRANDT_3_corr	0,6554	0,5047	0,5703	13961,3903	10751,5389
REMBRANDT_1	0,6750	0,4635	0,5496	13961,3903	9587,3739
XIP-L2F/Xerox_3	0,7137	0,3945	0,5081	13961,3903	7716,3313
XIP-L2F/Xerox_no	0,7062	0,3945	0,5063	13961,3903	7799,4144
XIP-L2F/Xerox_4	0,7051	0,3937	0,5053	13961,3903	7795,9456
REMMA_1_corr	0,6645	0,3699	0,4753	13961,3903	7773,0458
XIP-L2F/Xerox_2	0,7063	0,3551	0,4726	13961,3903	7019,4965
XIP-L2F/Xerox_1	0,6457	0,3715	0,4717	13961,3903	8033,3102
R3M_1	0,7596	0,3188	0,4491	13961,3903	5860,5000
SeRELeP_1	0,8103	0,3048	0,4429	13961,3903	5251,0000
SeRELeP_no	**0,8109**	0,2833	0,4199	13961,3903	4876,6667
R3M_2	0,7971	0,2774	0,4116	13961,3903	4859
REMMA_2_corr	0,6990	0,2870	0,4069	13961,3903	5732,4208
Cage2_4_corr	0,4989	0,3375	0,4026	13961,3903	9444,5575
Cage2_2_corr	0,4950	0,3370	0,4010	13961,3903	9505,8909
Cage2_1_corr	0,4905	0,3357	0,3986	13961,3903	9553,9623
Cage2_3_corr	0,4479	0,2864	0,3494	13961,3903	8925,4950
REMMA_3_corr	0,6775	0,2074	0,3176	13961,3903	4274,6458
SEIGeo_4	0,7462	0,1464	0,2448	13961,3903	2739,1607
SEIGeo_3	0,7543	0,1453	0,2437	13961,3903	2690,0714
SEIGeo_2	0,7553	0,1396	0,2356	13961,3903	2579,9077
SEIGeo_1	0,5843	0,0685	0,1226	13961,3903	1636,6220
DobrEM_1_corr	0,4530	0,0093	0,0182	13961,3903	287

Tabela I.5: Classificação no cenário 4 (cenário total sem subtipos)

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
Priberam_1	0,6442	**0,5175**	**0,5739**	17197,0285	13815,9264
REMBRANDT_2	0,6558	0,5026	0,5691	17197,0285	13178,2514
REMBRANDT_3_corr	0,6346	0,5017	0,5604	17197,0285	13594,7646
XIP-L2F/Xerox_3	0,6561	0,4639	0,5435	17197,0285	12158,7285
REMBRANDT_1	0,6459	0,4670	0,5421	17197,0285	12434,2306
XIP-L2F/Xerox_no	0,6582	0,4381	0,5261	17197,0285	11444,9368
XIP-L2F/Xerox_4	0,6539	0,4351	0,5225	17197,0285	11441,4306
XIP-L2F/Xerox_2	0,6383	0,4233	0,5090	17197,0285	11403,5528
REMMA_1_corr	0,6044	0,3710	0,4598	17197,0285	10555,5729
XIP-L2F/Xerox_1	0,5861	0,3776	0,4593	17197,0285	11079,7056
REMMA_2_corr	0,6120	0,3031	0,4054	17197,0285	8515,0701
R3M_1	0,7593	0,2586	0,3858	17197,0285	5857,5000
SeRELeP_1	0,8125	0,2479	0,3800	17197,0285	5248,0000
SeRELeP_no	**0,8131**	0,2304	0,3591	17197,0285	4873,1667
R3M_2	0,7967	0,2250	0,3509	17197,0285	4856
Cage2_4_corr	0,4478	0,2772	0,3424	17197,0285	10643,4444
Cage2_1_corr	0,4504	0,2749	0,3414	17197,0285	10496,3111
Cage2_2_corr	0,4442	0,2768	0,3411	17197,0285	10713,7778
REMMA_3_corr	0,5804	0,2375	0,3371	17197,0285	7037,9215
Cage2_3_corr	0,4027	0,2351	0,2969	17197,0285	10040,4611
SEIGeo_4	0,7640	0,1100	0,1924	17197,0285	2477,0000
SEIGeo_3	0,7721	0,1092	0,1914	17197,0285	2433,2333
SEIGeo_2	0,7727	0,1048	0,1845	17197,0285	2331,7167
PorTexTO_4_corr	0,6838	0,0851	0,1514	17197,0285	2140,9167
PorTexTO_3_corr	0,6842	0,0850	0,1512	17197,0285	2136,3167
PorTexTO_2_corr	0,6818	0,0834	0,1486	17197,0285	2104,1167
PorTexTO_1_corr	0,6975	0,0795	0,1427	17197,0285	1959,2167
SEIGeo_1	0,6151	0,0524	0,0966	17197,0285	1465,0000
DobrEM_1_corr	0,4530	0,0076	0,0149	17197,0285	287

Tabela I.6: Classificação no cenário 5 (LOCAL com tipos FISICO e HUMANO)

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
REMBRANDT_1	0,5848	0,6704	**0,6247**	2087,8214	2393,6548
REMBRANDT_3_corr	0,5582	**0,7024**	0,6221	2087,8214	2626,8542
SEIGeo_3	0,6785	0,5303	0,5953	2087,8214	1631,8452
SEIGeo_4	0,6709	0,5339	0,5946	2087,8214	1661,4345
XIP-L2F/Xerox_3	0,6903	0,5177	0,5917	2087,8214	1565,8720
SEIGeo_2	0,6810	0,5110	0,5839	2087,8214	1566,5774
XIP-L2F/Xerox_no	0,6931	0,4999	0,5808	2087,8214	1505,8988
XIP-L2F/Xerox_4	0,6931	0,4999	0,5808	2087,8214	1505,8988
XIP-L2F/Xerox_1	0,6977	0,4867	0,5734	2087,8214	1456,4196
REMBRANDT_2	0,4856	0,6981	0,5727	2087,8214	3001,5476
Cage2_4_corr	0,5267	0,5844	0,5540	2087,8214	2316,5774
Cage2_2_corr	0,5196	0,5851	0,5504	2087,8214	2351,3274
Cage2_1_corr	0,5147	0,5802	0,5455	2087,8214	2353,7470
Cage2_3_corr	0,5178	0,5754	0,5451	2087,8214	2319,8274
REMMA_1_corr	0,5683	0,5196	0,5428	2087,8214	1908,7500
XIP-L2F/Xerox_2	**0,7080**	0,4254	0,5315	2087,8214	1254,4018
REMMA_2_corr	0,6599	0,3838	0,4854	2087,8214	1214,3750
REMMA_3_corr	0,5493	0,3947	0,4593	2087,8214	1500
Priberam_1	0,3287	0,7001	0,4474	2087,8214	4446,2262
SEIGeo_1	0,4878	0,2323	0,3148	2087,8214	994,4167
SeRELeP_1	0,2072	0,5329	0,2984	2087,8214	5369,5000
SeRELeP_no	0,2104	0,5020	0,2965	2087,8214	4981
R3M_2	0,1943	0,4610	0,2734	2087,8214	4952,5000
R3M_1	0,1760	0,5036	0,2608	2087,8214	5975,5000

Tabela I.7: Classificação no cenário 6 (cenário total sem ABSTRACCAO e COISA)

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
REMBRANDT_2	0,6521	0,5296	**0,5845**	15972,7754	12971,3489
REMBRANDT_3_corr	0,6375	0,5304	0,5790	15972,7754	13290,5853
XIP-L2F/Xerox_3	0,6516	0,5083	0,5711	15972,7754	12460,5665
Priberam_1	0,6110	**0,5343**	0,5701	15972,7754	13968,2822
REMBRANDT_1	0,6457	0,4953	0,5605	15972,7754	12251,9811
XIP-L2F/Xerox_no	0,6529	0,4795	0,5529	15972,7754	11731,6453
XIP-L2F/Xerox_4	0,6486	0,4763	0,5493	15972,7754	11728,1766
XIP-L2F/Xerox_2	0,6357	0,4648	0,5370	15972,7754	11680,0379
XIP-L2F/Xerox_1	0,5821	0,4117	0,4823	15972,7754	11298,4433
REMMA_1_corr	0,6028	0,3859	0,4706	15972,7754	10226,1708
REMMA_2_corr	0,6174	0,3165	0,4185	15972,7754	8187,7958
R3M_1	0,6857	0,2516	0,3681	15972,7754	5861,0000
SeRELeP_1	0,7408	0,2435	0,3665	15972,7754	5250,5000
Cage2_4_corr	0,4380	0,2950	0,3525	15972,7754	10755,8284
Cage2_1_corr	0,4399	0,2920	0,3510	15972,7754	10602,7331
Cage2_2_corr	0,4345	0,2944	0,3510	15972,7754	10820,9117
REMMA_3_corr	0,5689	0,2496	0,3470	15972,7754	7009,4792
SeRELeP_no	0,7373	0,2251	0,3449	15972,7754	4876,1667
R3M_2	0,7240	0,2202	0,3377	15972,7754	4858,5000
Cage2_3_corr	0,3977	0,2529	0,3092	15972,7754	10155,2034
SEIGeo_4	0,7415	0,1272	0,2171	15972,7754	2739,1607
SEIGeo_3	0,7496	0,1262	0,2161	15972,7754	2690,0714
SEIGeo_2	**0,7510**	0,1213	0,2089	15972,7754	2579,9077
PorTexTO_4_corr	0,6784	0,0971	0,1699	15972,7754	2285,8021
PorTexTO_3_corr	0,6788	0,0969	0,1696	15972,7754	2280,8771
PorTexTO_2_corr	0,6764	0,0951	0,1668	15972,7754	2246,4021
PorTexTO_1_corr	0,6919	0,0906	0,1602	15972,7754	2091,2646
SEIGeo_1	0,5778	0,0592	0,1073	15972,7754	1635,3929
DobrEM_1_corr	0,4495	0,0081	0,0159	15972,7754	287

Tabela I.8: Classificação de ABSTRACCAO

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
Priberam_1	0,1099	0,5132	**0,1810**	581,8000	2716,8000
REMBRANDT_3_corr	0,1956	0,1433	0,1655	581,8000	426,3000
REMBRANDT_1	0,2067	0,1268	0,1572	581,8000	357,0000
REMBRANDT_2	0,1272	0,2045	0,1568	581,8000	935,7000
R3M_1	0,0548	**0,5638**	0,0998	581,8000	5989
SeRELeP_no	0,0541	0,4641	0,0968	581,8000	4994
SeRELeP_1	0,0529	0,4899	0,0955	581,8000	5384
R3M_2	0,0507	0,4323	0,0907	581,8000	4962,5000
REMMA_2_corr	**0,2231**	0,0392	0,0667	581,8000	102,2000
REMMA_1_corr	**0,2231**	0,0392	0,0667	581,8000	102,2000

Tabela I.9: Classificação de ACONTECIMENTO

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
XIP-L2F/Xerox_no	**0,7250**	0,1897	**0,3007**	424,7500	111,1458
REMBRANDT_3_corr	0,5630	0,2026	0,2980	424,7500	152,8542
XIP-L2F/Xerox_4	0,6632	0,1735	0,2751	424,7500	111,1458
REMBRANDT_1	0,6053	0,1760	0,2728	424,7500	123,5208
XIP-L2F/Xerox_1	0,5806	0,1735	0,2672	424,7500	126,9583
XIP-L2F/Xerox_3	0,6146	0,1638	0,2587	424,7500	113,2083
XIP-L2F/Xerox_2	0,6146	0,1638	0,2587	424,7500	113,2083
REMMA_1_corr	0,4044	0,1473	0,2159	424,7500	154,6875
REMMA_2_corr	0,4949	0,1370	0,2146	424,7500	117,5625
REMBRANDT_2	0,1381	0,2276	0,1719	424,7500	699,9167
Priberam_1	0,0668	0,4327	0,1158	424,7500	2749,9167
R3M_2	0,0338	0,3936	0,0622	424,7500	4950,1667
R3M_1	0,0331	**0,4654**	0,0618	424,7500	5975,6667
SeRELeP_1	0,0333	0,4206	0,0617	424,7500	5369,6667
SeRELeP_no	0,0303	0,3559	0,0559	424,7500	4982,1667
REMMA_3_corr	0,1818	0,0206	0,0370	424,7500	48,1250

Tabela I.10: Classificação de COISA

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
Priberam_1	0,0762	0,3899	**0,1275**	512,4000	2622,8000
R3M_2	0,0479	0,4635	0,0868	512,4000	4962,5000
SeRELeP_no	0,0449	0,4381	0,0815	512,4000	4994,5000
SeRELeP_1	0,0443	0,4655	0,0809	512,4000	5383,5000
R3M_1	0,0435	**0,5084**	0,0801	512,4000	5989,5000
REMMA_2_corr	0,2227	0,0318	0,0557	512,4000	73,2000
REMMA_1_corr	0,2227	0,0318	0,0557	512,4000	73,2000
REMBRANDT_2	0,0451	0,0566	0,0502	512,4000	643,5000
REMBRANDT_1	0,1982	0,0217	0,0391	512,4000	56,0000
REMBRANDT_3_corr	0,1404	0,0226	0,0390	512,4000	82,6000
XIP-L2F/Xerox_no	**1**	0,0027	0,0054	512,4000	1,4000
XIP-L2F/Xerox_4	**1**	0,0027	0,0054	512,4000	1,4000
XIP-L2F/Xerox_1	**1**	0,0027	0,0054	512,4000	1,4000

Tabela I.11: Classificação de LOCAL

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
REMBRANDT_1	0,5484	0,6607	**0,5993**	2350,1964	2831,0714
REMBRANDT_3_corr	0,5274	**0,6930**	0,5989	2350,1964	3088,1667
SEIGeo_3	0,6800	0,5138	0,5853	2350,1964	1775,8036
SEIGeo_4	0,6726	0,5175	0,5849	2350,1964	1808,2679
XIP-L2F/Xerox_3	0,6770	0,5047	0,5783	2350,1964	1752,0387
SEIGeo_2	0,6826	0,4954	0,5741	2350,1964	1705,5357
XIP-L2F/Xerox_no	0,6770	0,4869	0,5664	2350,1964	1690,3780
XIP-L2F/Xerox_4	0,6770	0,4869	0,5664	2350,1964	1690,3780
XIP-L2F/Xerox_1	0,6804	0,4740	0,5587	2350,1964	1637,0863
REMBRANDT_2	0,4702	0,6874	0,5584	2350,1964	3435,8185
Cage2_4_corr	0,5319	0,5527	0,5421	2350,1964	2442,3899
Cage2_2_corr	0,5254	0,5533	0,5390	2350,1964	2474,8899
REMMA_1_corr	0,5700	0,5089	0,5377	2350,1964	2098,2500
Cage2_3_corr	0,5235	0,5441	0,5336	2350,1964	2442,7024
Cage2_1_corr	0,5201	0,5478	0,5336	2350,1964	2475,2470
XIP-L2F/Xerox_2	**0,6901**	0,4118	0,5158	2350,1964	1402,5060
REMMA_2_corr	0,6624	0,3761	0,4798	2350,1964	1334,4375
Priberam_1	0,3471	0,6816	0,4599	2350,1964	4614,9345
REMMA_3_corr	0,5497	0,3859	0,4535	2350,1964	1650
SEIGeo_1	0,4961	0,2297	0,3140	2350,1964	1088,0000
SeRELeP_1	0,2135	0,4876	0,2970	2350,1964	5368
SeRELeP_no	0,2153	0,4563	0,2926	2350,1964	4980,5000
R3M_2	0,1996	0,4206	0,2708	2350,1964	4951,5000
R3M_1	0,1823	0,4634	0,2616	2350,1964	5974

Tabela I.12: Classificação de OBRA

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
REMBRANDT_3_corr	0,5251	0,2171	**0,3072**	662,6250	274
REMBRANDT_1	0,5261	0,1728	0,2602	662,6250	217,6250
REMBRANDT_2	0,2274	0,2794	0,2507	662,6250	814,1250
REMMA_1_corr	0,5146	0,1212	0,1962	662,6250	156,0625
REMMA_2_corr	0,5876	0,0908	0,1573	662,6250	102,4375
XIP-L2F/Xerox_no	0,4670	0,0847	0,1434	662,6250	120,1875
XIP-L2F/Xerox_4	0,4670	0,0847	0,1434	662,6250	120,1875
XIP-L2F/Xerox_1	0,4674	0,0826	0,1404	662,6250	117,1250
Priberam_1	0,0798	0,3323	0,1287	662,6250	2758,0625
SeRELeP_1	0,0471	0,3818	0,0838	662,6250	5377
SeRELeP_no	0,0451	0,3396	0,0796	662,6250	4990
R3M_1	0,0431	**0,3894**	0,0777	662,6250	5980
R3M_2	0,0395	0,2950	0,0696	662,6250	4954,5000
XIP-L2F/Xerox_2	**0,6038**	0,0332	0,0629	662,6250	36,4375
REMMA_3_corr	0,3066	0,0324	0,0587	662,6250	70,1250

Tabela I.13: Classificação de ORGANIZACAO

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
REMBRANDT_3_corr	0,5350	0,3231	**0,4029**	1570,4375	948,3750
REMBRANDT_1	0,5752	0,2803	0,3769	1570,4375	765,2917
REMBRANDT_2	0,3706	0,3432	0,3564	1570,4375	1454,5000
XIP-L2F/Xerox_no	0,5136	0,2625	0,3475	1570,4375	802,7500
XIP-L2F/Xerox_4	0,5136	0,2625	0,3475	1570,4375	802,7500
XIP-L2F/Xerox_1	0,5036	0,2625	0,3451	1570,4375	818,4167
REMMA_1_corr	0,5829	0,2397	0,3397	1570,4375	645,7917
Priberam_1	0,2277	0,5010	0,3131	1570,4375	3456,0417
REMMA_2_corr	**0,6603**	0,2033	0,3109	1570,4375	483,5417
XIP-L2F/Xerox_3	0,4116	0,1998	0,2690	1570,4375	762,5000
XIP-L2F/Xerox_2	0,4185	0,1947	0,2658	1570,4375	730,8125
Cage2_4_corr	0,3425	0,1851	0,2403	1570,4375	848,6667
Cage2_2_corr	0,3425	0,1851	0,2403	1570,4375	848,6667
Cage2_1_corr	0,3409	0,1844	0,2394	1570,4375	849,6667
R3M_2	0,1570	0,4946	0,2384	1570,4375	4945,6667
SeRELeP_1	0,1514	0,5168	0,2342	1570,4375	5360,6667
Cage2_3_corr	0,3623	0,1693	0,2307	1570,4375	733,8333
SeRELeP_no	0,1498	0,4745	0,2277	1570,4375	4974,1667
R3M_1	0,1420	**0,5398**	0,2249	1570,4375	5968,6667
REMMA_3_corr	0,4413	0,0714	0,1229	1570,4375	254,1458

Tabela I.14: Classificação de PESSOA

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
REMBRANDT_3_corr	**0,7683**	0,5368	**0,6320**	3079,6771	2151,5312
REMBRANDT_2	0,6850	0,5500	0,6101	3079,6771	2472,8333
XIP-L2F/Xerox_3	0,7122	0,5332	0,6099	3079,6771	2305,7500
REMBRANDT_1	0,7640	0,4917	0,5983	3079,6771	1981,9062
XIP-L2F/Xerox_2	0,6952	0,4999	0,5816	3079,6771	2214,4688
XIP-L2F/Xerox_4	0,6811	0,4994	0,5762	3079,6771	2258,0729
XIP-L2F/Xerox_no	0,6802	0,4994	0,5759	3079,6771	2260,9479
Priberam_1	0,4712	**0,7157**	0,5682	3079,6771	4677,9167
XIP-L2F/Xerox_1	0,5660	0,4430	0,4970	3079,6771	2410,4479
REMMA_1_corr	0,6666	0,3677	0,4740	3079,6771	1698,8854
REMMA_2_corr	0,7064	0,3323	0,4520	3079,6771	1448,7604
SeRELeP_1	0,3208	0,5586	0,4076	3079,6771	5362,3333
R3M_1	0,3077	0,5961	0,4059	3079,6771	5966,8333
SeRELeP_no	0,3235	0,5226	0,3996	3079,6771	4974,3333
R3M_2	0,3188	0,5122	0,3930	3079,6771	4948,3333
Cage2_4_corr	0,4136	0,2851	0,3376	3079,6771	2123,1667
Cage2_2_corr	0,4094	0,2816	0,3337	3079,6771	2118,1667
Cage2_1_corr	0,4039	0,2803	0,3309	3079,6771	2137,1667
Cage2_3_corr	0,2968	0,1890	0,2309	3079,6771	1961
REMMA_3_corr	0,6337	0,1399	0,2292	3079,6771	679,9375
DobrEM_1_corr	0,4330	0,0404	0,0739	3079,6771	287,5000

Tabela I.15: Classificação de TEMPO

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
XIP-L2F/Xerox_3	0,6812	**0,7314**	**0,7054**	1834,5875	1969,7750
XIP-L2F/Xerox_2	0,6768	0,7224	0,6989	1834,5875	1958,1875
XIP-L2F/Xerox_no	**0,7420**	0,6031	0,6654	1834,5875	1491,2500
XIP-L2F/Xerox_4	0,7388	0,6005	0,6625	1834,5875	1491,2500
PorTexTO_4_corr	0,6694	0,5419	0,5990	1834,5875	1485,0813
PorTexTO_3_corr	0,6698	0,5410	0,5986	1834,5875	1481,9063
PorTexTO_2_corr	0,6674	0,5310	0,5915	1834,5875	1459,6813
PorTexTO_1_corr	0,6825	0,5058	0,5810	1834,5875	1359,6688
REMBRANDT_3_corr	0,5904	0,4030	0,4790	1834,5875	1252,4250
REMBRANDT_1	0,5887	0,3998	0,4762	1834,5875	1246,0750
XIP-L2F/Xerox_1	0,6160	0,3540	0,4496	1834,5875	1054,2875
REMBRANDT_2	0,4058	0,4007	0,4032	1834,5875	1811,5500
REMMA_2_corr	0,4744	0,2538	0,3307	1834,5875	981,4000
REMMA_1_corr	0,4744	0,2538	0,3307	1834,5875	981,4000
REMMA_3_corr	0,4723	0,2496	0,3266	1834,5875	969,5000
Priberam_1	0,0832	0,1826	0,1143	1834,5875	4028,1125
Cage2_3_corr	0,0823	0,0294	0,0434	1834,5875	656
Cage2_4_corr	0,0771	0,0294	0,0426	1834,5875	700
Cage2_2_corr	0,0769	0,0294	0,0426	1834,5875	702
Cage2_1_corr	0,0804	0,0245	0,0376	1834,5875	560
SeRELeP_no	0,0006	0,0016	0,0009	1834,5875	4996
SeRELeP_1	0,0006	0,0016	0,0008	1834,5875	5386

Tabela I.16: Classificação de VALOR

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
REMBRANDT_3_corr	**0,4127**	**0,7176**	**0,5241**	489,5000	851
REMBRANDT_1	0,4008	0,6979	0,5092	489,5000	852,3750
REMMA_2_corr	0,3589	0,5202	0,4247	489,5000	709,5000
REMMA_1_corr	0,3589	0,5202	0,4247	489,5000	709,5000
XIP-L2F/Xerox_3	0,3100	0,6558	0,4209	489,5000	1035,6250
REMMA_3_corr	0,3527	0,5202	0,4204	489,5000	721,8750
REMBRANDT_2	0,2426	0,7061	0,3611	489,5000	1424,6250
XIP-L2F/Xerox_no	0,2648	0,5444	0,3563	489,5000	1006,5000
XIP-L2F/Xerox_2	0,2468	0,5026	0,3310	489,5000	996,8750
XIP-L2F/Xerox_4	0,2448	0,5033	0,3294	489,5000	1006,5000
XIP-L2F/Xerox_1	0,2364	0,4801	0,3168	489,5000	994,1250
Priberam_1	0,1055	0,7099	0,1836	489,5000	3295,1250
SeRELeP_1	0,0024	0,0266	0,0044	489,5000	5386
SeRELeP_no	0,0022	0,0225	0,0040	489,5000	4996

I.1.2 Avaliação relaxada de ALT

Tabela I.17: Classificação no cenário total com avaliação relaxada de ALT

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
Priberam_1	0,6592	**0,5352**	**0,5908**	17695,7702	14368,0113
REMBRANDT_2	0,6622	0,5173	0,5808	17561,6024	13718,3113
REMBRANDT_3_corr	0,6424	0,5163	0,5725	17622,6643	14163,5054
XIP-L2F/Xerox_3	0,6664	0,4840	0,5607	17625,3786	12799,5750
REMBRANDT_1	0,6505	0,4809	0,5530	17542,0690	12966,5185
XIP-L2F/Xerox_no	0,6701	0,4584	0,5444	17620,0411	12054,2643
XIP-L2F/Xerox_4	0,6660	0,4552	0,5408	17620,0411	12042,8893
XIP-L2F/Xerox_2	0,6510	0,4438	0,5278	17605,1060	12001,8798
XIP-L2F/Xerox_1	0,6062	0,3956	0,4788	17617,4036	11496,2685
REMMA_1_corr	0,6226	0,3750	0,4681	17594,5042	10597,5875
REMMA_2_corr	0,6340	0,3084	0,4150	17596,5708	8559,8125
R3M_1	0,7820	0,2633	0,3940	17502,1690	5894
SeRELeP_1	**0,8343**	0,2512	0,3861	17622,1643	5305
SeRELeP_no	0,8340	0,2331	0,3644	17613,4667	4923
R3M_2	0,8178	0,2289	0,3577	17489,8649	4896
Cage2_4_corr	0,4751	0,2865	0,3574	17843,0054	10759,6845
Cage2_2_corr	0,4717	0,2861	0,3562	17843,0054	10823,2845
Cage2_1_corr	0,4775	0,2839	0,3561	17838,0387	10606,2500
REMMA_3_corr	0,5950	0,2409	0,3429	17481,4708	7078,3000
Cage2_3_corr	0,4289	0,2465	0,3131	17858,9720	10261,7917
SEIGeo_4	0,7558	0,1214	0,2092	17506,3774	2812,0036
SEIGeo_3	0,7640	0,1206	0,2082	17506,3774	2762,4393
SEIGeo_2	0,7644	0,1158	0,2011	17506,2315	2652,3548
PorTexTO_4_corr	0,6899	0,0909	0,1607	17481,9274	2304,5125
PorTexTO_3_corr	0,6903	0,0908	0,1605	17481,9274	2299,5375
PorTexTO_2_corr	0,6880	0,0891	0,1578	17481,9274	2264,7125
PorTexTO_1_corr	0,7044	0,0849	0,1516	17481,9274	2108,0000
SEIGeo_1	0,5960	0,0566	0,1034	17462,7315	1659,4857
DobrEM_1_corr	0,4596	0,0075	0,0148	17444,4607	285

Tabela I.18: Identificação no cenário total com avaliação relaxada de ALT

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
Priberam_1	0,7188	**0,7485**	**0,2515**	7232	7531
SeRELeP_1	0,8343	0,6153	0,3847	7193	5305
R3M_1	0,7820	0,6436	0,3564	7161	5894
REMBRANDT_2	0,7726	0,6384	0,3616	7185	5937
SeRELeP_no	0,8340	0,5710	0,4290	7191	4923
REMBRANDT_3_corr	0,7609	0,6103	0,3897	7208	5781
R3M_2	0,8178	0,5595	0,4405	7156	4896
REMBRANDT_1	0,7660	0,5651	0,4349	7177	5295
XIP-L2F/Xerox_3	0,7365	0,5527	0,4473	7204	5407
XIP-L2F/Xerox_no	0,7403	0,5242	0,4758	7202	5099
XIP-L2F/Xerox_4	0,7362	0,5207	0,4793	7202	5094
XIP-L2F/Xerox_2	0,7180	0,5048	0,4952	7197	5060
REMMA_1_corr	0,7264	0,4660	0,5340	7197	4617
XIP-L2F/Xerox_1	0,6809	0,4620	0,5380	7201	4886
REMMA_2_corr	0,6967	0,3607	0,6393	7198	3726
Cage2_1_corr	0,5370	0,3901	0,6099	7275	5285
Cage2_4_corr	0,5322	0,3916	0,6084	7277	5355
Cage2_2_corr	0,5289	0,3921	0,6079	7277	5394
REMMA_3_corr	0,7321	0,3164	0,6836	7152	3091
Cage2_3_corr	0,4771	0,3341	0,6659	7283	5100
SEIGeo_4	0,9016	0,1410	0,8590	7149	1118
SEIGeo_3	**0,9108**	0,1400	0,8600	7149	1099
SEIGeo_2	0,9088	0,1339	0,8661	7149	1053
PorTexTO_4_corr	0,7115	0,0924	0,9076	7153	929
PorTexTO_3_corr	0,7120	0,0923	0,9077	7153	927
PorTexTO_2_corr	0,7097	0,0906	0,9094	7153	913
PorTexTO_1_corr	0,7271	0,0864	0,9136	7153	850
SEIGeo_1	0,7729	0,0696	0,9304	7142	643
DobrEM_1_corr	0,4596	0,0184	0,9816	7138	285

I.2 Resultados da pista do TEMPO

I.2.1 HAREM clássico na CD do TEMPO

Tabela I.19: Classificação de TEMPO na CD do TEMPO

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
XIP-L2F/Xerox_3	0,7376	**0,7580**	**0,7477**	361,5500	371,5375
XIP-L2F/Xerox_2	0,7235	0,7506	0,7368	361,5500	375,0875
XIP-L2F/Xerox_no	**0,7689**	0,6419	0,6997	361,5500	301,8000
XIP-L2F/Xerox_4	0,7637	0,6375	0,6949	361,5500	301,8000
PorTexTO_4_corr	0,7350	0,5327	0,6177	361,5500	262,0250
PorTexTO_3_corr	0,7350	0,5327	0,6177	361,5500	262,0250
PorTexTO_2_corr	0,7393	0,5195	0,6102	361,5500	254,0875
PorTexTO_1_corr	0,7273	0,4888	0,5847	361,5500	242,9750
REMBRANDT_3_corr	0,6028	0,4481	0,5140	361,5500	268,7500
XIP-L2F/Xerox_1	0,6785	0,4101	0,5112	361,5500	218,5000
REMBRANDT_1	0,5825	0,4407	0,5018	361,5500	273,5125
REMBRANDT_2	0,4077	0,4407	0,4236	361,5500	390,7500
REMMA_2_corr	0,4565	0,2581	0,3297	361,5500	204,4000
REMMA_1_corr	0,4565	0,2581	0,3297	361,5500	204,4000
REMMA_3_corr	0,4585	0,2503	0,3238	361,5500	197,4000
Priberam_1	0,0950	0,1946	0,1277	361,5500	740,3625
Cage2_1_corr	0,0667	0,0166	0,0266	361,5500	90
Cage2_4_corr	0,0508	0,0166	0,0250	361,5500	118
Cage2_3_corr	0,0504	0,0166	0,0250	361,5500	119
Cage2_2_corr	0,0504	0,0166	0,0250	361,5500	119
SeRELeP_no	0,0011	0,0028	0,0016	361,5500	909
SeRELeP_1	0,0010	0,0028	0,0015	361,5500	953
R3M_2	0	0	0	361,5500	877
R3M_1	0	0	0	361,5500	1074
DobrEM_1_corr	0	0	0	361,5500	40

I.2.2 TEMPO completo

Tabela I.20: Classificação de TEMPO completo na CD do TEMPO

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
XIP-L2F/Xerox_3	0,8087	**0,7024**	**0,7518**	711,3505	617,8713
XIP-L2F/Xerox_2	0,8036	0,6987	0,7475	711,3505	618,4880
XIP-L2F/Xerox_no	**0,8267**	0,5838	0,6843	711,3505	502,3337
XIP-L2F/Xerox_4	0,8247	0,5816	0,6821	711,3505	501,6671
XIP-L2F/Xerox_1	0,7326	0,3468	0,4707	711,3505	336,7003
PorTexTO_4_corr	0,7350	0,2708	0,3957	711,3505	262,0250
PorTexTO_3_corr	0,7350	0,2708	0,3957	711,3505	262,0250
PorTexTO_2_corr	0,7393	0,2641	0,3891	711,3505	254,0875
PorTexTO_1_corr	0,7273	0,2484	0,3704	711,3505	242,9750
REMBRANDT_3_corr	0,6028	0,2277	0,3306	711,3505	268,7500
REMBRANDT_1	0,5825	0,2240	0,3235	711,3505	273,5125
REMBRANDT_2	0,4077	0,2240	0,2891	711,3505	390,7500
REMMA_2_corr	0,4565	0,1312	0,2038	711,3505	204,4000
REMMA_1_corr	0,4565	0,1312	0,2038	711,3505	204,4000
REMMA_3_corr	0,4585	0,1272	0,1992	711,3505	197,4000
Priberam_1	0,1115	0,1186	0,1150	711,3505	756,3626
Cage2_1_corr	0,0667	0,0084	0,0150	711,3505	90
Cage2_4_corr	0,0508	0,0084	0,0145	711,3505	118
Cage2_3_corr	0,0504	0,0084	0,0145	711,3505	119
Cage2_2_corr	0,0504	0,0084	0,0145	711,3505	119
SeRELeP_no	0,0011	0,0014	0,0012	711,3505	909
SeRELeP_1	0,0010	0,0014	0,0012	711,3505	953
R3M_2	0	0	0	711,3505	877
R3M_1	0	0	0	711,3505	1074
DobrEM_1_corr	0	0	0	711,3505	40

I.2.3 TEMPO sem normalização

Tabela I.21: Classificação de TEMPO estendido sem normalização na CD do TEMPO

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
XIP-L2F/Xerox_3	0,7733	**0,7472**	**0,7600**	516,3505	498,8713
XIP-L2F/Xerox_2	0,7670	0,7420	0,7543	516,3505	499,4880
XIP-L2F/Xerox_no	**0,7955**	0,6230	0,6987	516,3505	404,3337
XIP-L2F/Xerox_4	0,7929	0,6199	0,6958	516,3505	403,6671
XIP-L2F/Xerox_1	0,7005	0,3930	0,5035	516,3505	289,7003
PorTexTO_4_corr	0,7350	0,3730	0,4949	516,3505	262,0250
PorTexTO_3_corr	0,7350	0,3730	0,4949	516,3505	262,0250
PorTexTO_2_corr	0,7393	0,3638	0,4876	516,3505	254,0875
PorTexTO_1_corr	0,7273	0,3423	0,4655	516,3505	242,9750
REMBRANDT_3_corr	0,6028	0,3137	0,4127	516,3505	268,7500
REMBRANDT_1	0,5825	0,3086	0,4034	516,3505	273,5125
REMBRANDT_2	0,4077	0,3086	0,3513	516,3505	390,7500
REMMA_2_corr	0,4565	0,1807	0,2589	516,3505	204,4000
REMMA_1_corr	0,4565	0,1807	0,2589	516,3505	204,4000
REMMA_3_corr	0,4585	0,1753	0,2536	516,3505	197,4000
Priberam_1	0,1115	0,1634	0,1326	516,3505	756,3626
Cage2_1_corr	0,0667	0,0116	0,0198	516,3505	90
Cage2_4_corr	0,0508	0,0116	0,0189	516,3505	118
Cage2_3_corr	0,0504	0,0116	0,0189	516,3505	119
Cage2_2_corr	0,0504	0,0116	0,0189	516,3505	119
SeRELeP_no	0,0011	0,0019	0,0014	516,3505	909
SeRELeP_1	0,0010	0,0019	0,0014	516,3505	953
R3M_2	0	0	0	516,3505	877
R3M_1	0	0	0	516,3505	1074
DobrEM_1_corr	0	0	0	516,3505	40

I.2.4 TEMPO só normalização

Tabela I.22: Classificação de TEMPO estendido só com normalização na CD do TEMPO

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
XIP-L2F/Xerox_3	0,7908	**0,6970**	**0,7410**	556,5500	490,5375
XIP-L2F/Xerox_2	0,7797	0,6922	0,7334	556,5500	494,0875
XIP-L2F/Xerox_no	**0,8146**	0,5852	0,6811	556,5500	399,8000
XIP-L2F/Xerox_4	0,8107	0,5823	0,6778	556,5500	399,8000
PorTexTO_4_corr	0,7350	0,3461	0,4706	556,5500	262,0250
PorTexTO_3_corr	0,7350	0,3461	0,4706	556,5500	262,0250
XIP-L2F/Xerox_1	0,7232	0,3450	0,4672	556,5500	265,5000
PorTexTO_2_corr	0,7393	0,3375	0,4634	556,5500	254,0875
PorTexTO_1_corr	0,7273	0,3175	0,4421	556,5500	242,9750
REMBRANDT_3_corr	0,6028	0,2911	0,3926	556,5500	268,7500
REMBRANDT_1	0,5825	0,2863	0,3839	556,5500	273,5125
REMBRANDT_2	0,4077	0,2863	0,3364	556,5500	390,7500
REMMA_2_corr	0,4565	0,1676	0,2452	556,5500	204,4000
REMMA_1_corr	0,4565	0,1676	0,2452	556,5500	204,4000
REMMA_3_corr	0,4585	0,1626	0,2401	556,5500	197,4000
Priberam_1	0,0950	0,1264	0,1085	556,5500	740,3625
Cage2_1_corr	0,0667	0,0108	0,0186	556,5500	90
Cage2_4_corr	0,0508	0,0108	0,0178	556,5500	118
Cage2_3_corr	0,0504	0,0108	0,0178	556,5500	119
Cage2_2_corr	0,0504	0,0108	0,0178	556,5500	119
SeRELeP_no	0,0011	0,0018	0,0014	556,5500	909
SeRELeP_1	0,0010	0,0018	0,0013	556,5500	953
R3M_2	0	0	0	556,5500	877
R3M_1	0	0	0	556,5500	1074
DobrEM_1_corr	0	0	0	556,5500	40

I.3 Resultados do ReRelEM

I.3.1 HAREM clássico na CD do ReRelEM

Tabela I.23: Classificação no cenário total na CD do ReRelEM

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
XIP-L2F/Xerox_3	0,7254	0,5016	**0,5931**	1283,6524	887,6586
REMBRANDT_3_corr	0,6458	**0,5156**	0,5734	1283,6524	1024,8920
REMBRANDT_2	0,6569	0,4966	0,5656	1283,6524	970,5426
XIP-L2F/Xerox_2	0,7001	0,4571	0,5531	1283,6524	838,1586
XIP-L2F/Xerox_no	0,6859	0,4573	0,5487	1283,6524	855,8628
XIP-L2F/Xerox_4	0,6859	0,4573	0,5487	1283,6524	855,8628
REMBRANDT_1	0,6396	0,4733	0,5440	1283,6524	950,0217
Priberam_1	0,6143	0,4858	0,5425	1283,6524	1015,0717
REMMA_1_corr	0,6615	0,4017	0,4999	1283,6524	779,5437
XIP-L2F/Xerox_1	0,6191	0,3895	0,4782	1283,6524	807,5774
REMMA_2_corr	0,6575	0,3247	0,4347	1283,6524	633,9562
REMMA_3_corr	0,6635	0,2791	0,3929	1283,6524	539,9312
R3M_1	0,8514	0,2411	0,3758	1283,6524	363,5000
SeRELeP_1	0,8468	0,2197	0,3489	1283,6524	333
Cage2_1_corr	0,4577	0,2664	0,3368	1283,6524	747,1488
Cage2_4_corr	0,4511	0,2658	0,3345	1283,6524	756,4988
Cage2_2_corr	0,4503	0,2658	0,3343	1283,6524	757,8155
R3M_2	**0,8735**	0,2045	0,3314	1283,6524	300,5000
SeRELeP_no	0,8506	0,2041	0,3292	1283,6524	308
Cage2_3_corr	0,4239	0,2397	0,3062	1283,6524	725,8101
SEIGeo_4	0,7783	0,1283	0,2203	1283,6524	211,5673
SEIGeo_3	0,7757	0,1263	0,2173	1283,6524	209,0839
SEIGeo_2	0,7795	0,1209	0,2094	1283,6524	199,1506
PorTexTO_4_corr	0,8335	0,0805	0,1468	1283,6524	124,0000
PorTexTO_3_corr	0,8335	0,0805	0,1468	1283,6524	124,0000
PorTexTO_2_corr	0,8335	0,0805	0,1468	1283,6524	124,0000
SEIGeo_1	0,7137	0,0779	0,1405	1283,6524	140,1399
PorTexTO_1_corr	0,8265	0,0766	0,1403	1283,6524	119,0250
DobrEM_1_corr	0,3333	0,0066	0,0130	1283,6524	25,5000

Tabela I.24: Classificação no cenário 5 (LOCAL com tipos FISICO e HUMANO) na CD do ReRelEM

Corrida	Precisão	Abrangência	Medida F	Máx. CD	Máx. Sistema
REMBRANDT_3_corr	0,5341	**0,6410**	**0,5827**	181,2024	217,4970
SEIGeo_4	0,6961	0,4908	0,5757	181,2024	127,7589
XIP-L2F/Xerox_3	0,7274	0,4750	0,5747	181,2024	118,3274
SEIGeo_3	0,6926	0,4827	0,5689	181,2024	126,3006
REMBRANDT_1	0,5320	0,5836	0,5566	181,2024	198,7768
SEIGeo_2	0,6939	0,4614	0,5542	181,2024	120,4673
XIP-L2F/Xerox_no	0,6612	0,4663	0,5469	181,2024	127,7857
XIP-L2F/Xerox_4	0,6612	0,4663	0,5469	181,2024	127,7857
REMBRANDT_2	0,4771	0,6312	0,5434	181,2024	239,7351
XIP-L2F/Xerox_2	**0,7332**	0,4279	0,5404	181,2024	105,7440
Cage2_4_corr	0,5521	0,5290	0,5403	181,2024	173,6280
Cage2_2_corr	0,5503	0,5290	0,5395	181,2024	174,1696
Cage2_1_corr	0,5480	0,5239	0,5357	181,2024	173,2530
XIP-L2F/Xerox_1	0,6691	0,4465	0,5356	181,2024	120,9107
Cage2_3_corr	0,5345	0,5027	0,5181	181,2024	170,4018
REMMA_1_corr	0,5963	0,4401	0,5064	181,2024	133,7500
Priberam_1	0,3808	0,6331	0,4756	181,2024	301,2530
REMMA_3_corr	0,6095	0,3553	0,4489	181,2024	105,6250
SEIGeo_1	0,6236	0,2907	0,3966	181,2024	84,4732
REMMA_2_corr	0,5829	0,2815	0,3796	181,2024	87,5000
SeRELeP_no	0,2750	0,4856	0,3512	181,2024	320
SeRELeP_1	0,2601	0,4967	0,3414	181,2024	346
R3M_2	0,2576	0,4443	0,3261	181,2024	312,5000
R3M_1	0,2312	0,4829	0,3127	181,2024	378,5000
DobrEM_1_corr	0,0196	0,0028	0,0048	181,2024	25,5000

I.3.2 ReRelEM no cenário total

Tabela I.25: Avaliação de todas as relações no cenário total

				Valor da medida de classificação		
Corrida	Precisão	Abrangência	Medida F	CD	Máx. Sistema	Sistema
REMBRANDT_1	0,5822	0,3669	**0,4502**	714	450	262
SeRELeP_1	0,5831	0,2672	0,3665	1654	758	442
SeRELeP_no	0,5665	0,2694	0,3652	1518	722	409
REMBRANDT_2	0,2711	**0,4043**	0,3246	841	1254	340
REMBRANDT_3_corr	0,2450	0,3929	0,3018	817	1310	321
SEIGeo_2	0,9167	0,1618	0,2750	136	24	22
SEIGeo_4	0,9167	0,1549	0,2651	142	24	22
SEIGeo_1	0,2500	0,0952	0,1379	42	16	4
SEIGeo_3	**1**	0,0429	0,0822	140	6	6

Tabela I.26: Avaliação de relação de identidade no cenário total

				Valor da medida de classificação		
Corrida	Precisão	Abrangência	Medida F	CD	Máx. Sistema	Sistema
REMBRANDT_1	0,7724	**0,6934**	**0,7308**	274	246	190
REMBRANDT_2	0,7657	0,6718	0,7157	326	286	219
REMBRANDT_3_corr	0,7674	0,6462	0,7016	342	288	221
SeRELeP_1	**0,8897**	0,5489	0,6789	470	290	258
SeRELeP_no	0,8846	0,5476	0,6765	420	260	230

Tabela I.27: Avaliação da relação de inclusão no cenário total

				Valor da medida de classificação		
Corrida	Precisão	Abrangência	Medida F	CD	Máx. Sistema	Sistema
SEIGeo_2	0,9167	0,2973	**0,4490**	74	24	22
SEIGeo_4	0,9167	0,2821	0,4314	78	24	22
REMBRANDT_1	0,3226	0,3261	0,3243	184	186	60
SEIGeo_1	0,2500	0,1538	0,1905	26	16	4
SeRELeP_no	0,5357	0,1061	0,1772	424	84	45
REMBRANDT_2	0,1123	0,4094	0,1763	254	926	104
SeRELeP_1	0,5238	0,0961	0,1624	458	84	44
REMBRANDT_3_corr	0,0896	**0,4231**	0,1479	208	982	88
SEIGeo_3	**1**	0,0769	0,1429	78	6	6

Tabela I.28: Avaliação da relação de localização no cenário total

				Valor da medida de classificação		
Corrida	Precisão	Abrangência	Medida F	CD	Máx. Sistema	Sistema
SeRELeP_1	0,3646	**0,2703**	**0,3104**	518	384	140
SeRELeP_no	0,3545	0,2648	0,3032	506	378	134
REMBRANDT_2	0,4048	0,1288	0,1954	132	42	17
REMBRANDT_1	**0,6667**	0,0909	0,1600	132	18	12
REMBRANDT_3_corr	0,3000	0,1000	0,1500	120	40	12

Tabela I.29: Avaliação de relações sem `outra` no cenário total

				Valor da medida de classificação		
Corrida	Precisão	Abrangência	Medida F	CD	Máx. Sistema	Sistema
REMBRANDT_1	0,5822	0,4441	**0,5038**	590	450	262
SeRELeP_1	0,5831	0,3057	0,4011	1446	758	442
SeRELeP_no	0,5665	0,3030	0,3948	1350	722	409
REMBRANDT_2	0,2711	0,4775	0,3459	712	1254	340
REMBRANDT_3_corr	0,2450	**0,4791**	0,3242	670	1310	321
SEIGeo_2	0,9167	0,1618	0,2750	136	24	22
SEIGeo_4	0,9167	0,1549	0,2651	142	24	22
SEIGeo_1	0,2500	0,0952	0,1379	42	16	4
SEIGeo_3	**1**	0,0429	0,0822	140	6	6

I.3.3 ReRelEM no cenário 5

Tabela I.30: Avaliação de todas as relações no cenário 5 (LOCAL com tipos FISICO e HUMANO)

				Valor da medida de classificação		
Corrida	Precisão	Abrangência	Medida F	CD	Máx. Sistema	Sistema
REMBRANDT_1	0,9178	**0,6204**	**0,7403**	216	146	134
REMBRANDT_3_corr	0,7123	0,6089	0,6565	248	212	151
REMBRANDT_2	0,7594	0,5367	0,6289	300	212	161
SeRELeP_no	0,6064	0,3333	0,4302	342	188	114
SeRELeP_1	0,6000	0,3276	0,4238	348	190	114
SEIGeo_2	0,9167	0,1618	0,2750	136	24	22
SEIGeo_4	0,9167	0,1549	0,2651	142	24	22
SEIGeo_1	0,2500	0,0952	0,1379	42	16	4
SEIGeo_3	**1**	0,0429	0,0822	140	6	6

Tabela I.31: Avaliação de relação de identidade no cenário 5 (LOCAL com tipos FISICO e HUMANO)

				Valor da medida de classificação		
Corrida	Precisão	Abrangência	Medida F	CD	Máx. Sistema	Sistema
REMBRANDT_3_corr	**0,9184**	**0,9000**	**0,9091**	100	98	90
REMBRANDT_1	0,9130	0,8936	0,9032	94	92	84
REMBRANDT_2	**0,9184**	0,8654	0,8911	104	98	90
SeRELeP_no	0,9130	0,6774	0,7778	124	92	84
SeRELeP_1	0,9130	0,6774	0,7778	124	92	84

Tabela I.32: Avaliação de relação de inclusão no cenário 5 (LOCAL com tipos FISICO e HUMANO)

				Valor da medida de classificação		
Corrida	Precisão	Abrangência	Medida F	CD	Máx. Sistema	Sistema
REMBRANDT_1	0,9615	0,4098	**0,5747**	122	52	50
REMBRANDT_3_corr	0,5446	**0,4122**	0,4692	148	112	61
REMBRANDT_2	0,6339	0,3622	0,4610	196	112	71
SEIGeo_2	0,9167	0,2973	0,4490	74	24	22
SEIGeo_4	0,9167	0,2821	0,4314	78	24	22
SeRELeP_no	0,5556	0,1376	0,2206	218	54	30
SeRELeP_1	0,5556	0,1339	0,2158	224	54	30
SEIGeo_1	0,2500	0,1538	0,1905	26	16	4
SEIGeo_3	**1**	0,0769	0,1429	78	6	6

Tabela I.33: Avaliação de relações sem `outra` no cenário 5 (`LOCAL` com tipos `FISICO` e `HUMANO`)

				Valor da medida de classificação		
Corrida	Precisão	Abrangência	Medida F	CD	Máx. Sistema	Sistema
REMBRANDT_1	0,9178	**0,6204**	**0,7403**	216	146	134
REMBRANDT_3_corr	0,7123	0,6089	0,6565	248	212	151
REMBRANDT_2	0,7594	0,5367	0,6289	300	212	161
SeRELeP_no	0,6064	0,3333	0,4302	342	188	114
SeRELeP_1	0,6000	0,3276	0,4238	348	190	114
SEIGeo_2	0,9167	0,1618	0,2750	136	24	22
SEIGeo_4	0,9167	0,1549	0,2651	142	24	22
SEIGeo_1	0,2500	0,0952	0,1379	42	16	4
SEIGeo_3	**1**	0,0429	0,0822	140	6	6

Bibliografia

(Afonso et al., 2002) Susana Afonso, Eckhard Bick, Renato Haber e Diana Santos. Floresta sintá(c)tica: um treebank para o português. Em Anabela Gonçalves e Clara Nunes Correia, editoras, *Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001)*, Lisboa, Portugal, 2-4 de Outubro de 2002, p. 533–545. APL.

(Agichtein e Gravano, 2000) Eugene Agichtein e Luis Gravano. Snowball: Extracting relations from large plain-text collections. Em Peter J. Nürnberg, David L. Hicks e Richard Furuta, editores, *Proceedings of the Fifth ACM Conference on Digital Libraries*, San Antonio, TX, EUA, 2-7 de Junho de 2000, p. 85–94.

(Aires, 2005) Rachel Virgínia Xavier Aires. Uso de marcadores estilísticos para a busca na Web em português. Tese de doutoramento, ICMC - USP - São Carlos, Agosto de 2005.

(Ait-Mokhtar et al., 2002) Salah Ait-Mokhtar, Jean-Pierre Chanod e Claude Roux. Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering*, 8(2–3):121–144, 2002.

(Allen, 1983) James F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983. (doi=http://doi.acm.org/10.1145/182.358434).

(Almeida, 2007) José João Dias de Almeida. RENA - Reconhecedor de Entidades. Em Santos e Cardoso (2007a), p. 157–172. http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap13-SantosCardoso2007-Almeida.pdf.

(Alonso et al., 2007) Omar Alonso, Michael Gertz e Ricardo Baeza-Yates. On the value of temporal information in information retrieval. *SIGIR Forum*, 41(2):35–41, 2007.

(Aluisio et al., 2004) Sandra Aluisio, Gisele Montilha Pinheiro, Aline M. P. Manfrin, Leandro H. M. de Oliveira, Luiz C. Genoves Jr. e Stella E. O. Tagnin. The Lácio-Web: Corpora and tools to advance Brazilian Portuguese language investigations and computational linguistic tools. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editoras, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisboa, Portugal, Maio de 2004, p. 1779–1782.

(Amaral et al., 2004a) Carlos Amaral, Helena Figueira, Afonso Mendes, Pedro Mendes e Cláudia Pinto. A workbench for developing natural language processing tools. Em *Pre-proceedings of the 1st Workshop on International Proofing Tools and Language Technologies*, Patras, Grécia, Julho de 2004.

(Amaral et al., 2004b) Carlos Amaral, Dominique Laurent, André Martins, Afonso Mendes e Cláudia Pinto. Design and implementation of a semantic search engine for Portuguese. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editoras, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisboa, Portugal, Maio de 2004, p. 247–250.

(Amaral et al., 2005) Carlos Amaral, Helena Figueira, André Martins Afonso Mendes, Pedro Mendes e Cláudia Pinto. Priberam's question answering system for Portuguese. Em *Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop (CLEF 2005)*, Viena, Áustria, Setembro de 2005.

(Amaral et al., 2007) Carlos Amaral, Adán Cassan, Helena Figueira, Afonso Mendes, Pedro Mendes, Cláudia Pinto e Daniel Vidal. Priberam's question answering system in QA@CLEF 2007. Em Alessandro Nardi e Carol Peters, editores, *Cross Language Evaluation Forum: Working notes for the CLEF 2007 workshop (CLEF 2007)*, Budapeste, Hungria, Setembro de 2007.

(Amitay et al., 2004) Einat Amitay, Nadav Har'El, Ron Sivan e Aya Soffer. Web-a-Where: Geotagging Web content. Em Mark Sanderson, Kalervo Järvelin, James Allan e Peter Bruza, editores, *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*, Sheffield, Reino Unido, 25-29 de Julho de 2004, p. 273–280. ACM Press.

(Auer e Lehmann, 2007) Sören Auer e Jens Lehmann. What have Innsbruck and Leipzig in common? Extracting semantics from wiki content. Em Enrico Franconi, Michael Kifer e Wolfgang May, editores, *The Semantic Web: Research and applications, 4th European Semantic Web Conference, ESWC 2007, Innsbruck, Austria, June 3-7, Proceedings*. Springer, 2007, p. 503–517.

(Auer et al., 2007) Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak e Zachary Ives. DBpedia: A nucleus for a Web of open data. Em Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Diana Maynard Peter Mika, Riichiro Mizoguchi, Guus Schreiber e Philippe Cudré-Mauroux, editores, *6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007, Proceedings*. Springer, 2007, p. 722–735.

(Bacelar do Nascimento et al., 2000) Maria Fernanda Bacelar do Nascimento, Luísa Pereira e João Saramago. Portuguese corpora at CLUL. Em Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis e Gregory Stainhauer, editores, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, Atenas, Grécia, 31 de Maio-2 de Junho de 2000, p. 1603–1608.

(Baptista, 2003) Jorge Baptista. Some families of compound temporal adverbs in Portuguese. Em *Workshop on Finite-State Methods for Natural Language Processing, International Conference of the European Chapter of the Association for Computational Linguistics*, Budapeste, Hungria, 13-14 de Abril de 2003, p. 97–104.

(Barreiro, 2008) Anabela Barreiro. Port4NooJ: Portuguese linguistic module and bilingual resources for machine translation. Em Xavier Blanco e Max Silberztein, editores, *Proceedings of the 2007 International NooJ Conference*, Barcelona, Espanha, 7-9 de Junho de 2008, p. 19–47. Cambridge Scholars Publishing.

(Battistelli et al., 2008) Delphine Battistelli, Javier Couto, Jean-Luc Minel e Sylviane R. Schwer. Représentation algébrique des expressions calendaires et vue calendaire d'un texte. Em *Actes de TALN 2008*, Avignon, Junho de 2008, p. 365–373.

(Beigbeder, 2004) Michel Beigbeder. Les temps du document et la recherche d'information. *Document numérique*, 8(4):55–64, 2004.

(Bick, 2000) Eckhard Bick. *The parsing system "Palavras": Automatic grammatical analysis of Portuguese in a constraint grammar framework*. Aarhus, Dinamarca, Aarhus University Press, Novembro de 2000. Tese de doutoramento, Aarhus University.

(Bick, 2003) Eckhard Bick. Multi-level NER for Portuguese in a CG framework. Em Nuno J. Mamede, Jorge Baptista, Isabel Trancoso e Maria das Graças Volpe Nunes, editores, *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003. Faro, Portugal, June 2003 (PROPOR 2003)*. Springer Verlag, Berlim/Heidelberg, 26-27 de Junho de 2003, p. 118–125.

(Bick, 2007) Eckhard Bick. Functional aspects on Portuguese NER. Em Santos e Cardoso (2007a), p. 145–155. (Este artigo foi previamente publicado pela Springer, na série LNAI, vol. 3960, ISBN-10: 3-540-34045-9). http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap12-SantosCardoso2007-Bick.pdf.

(Blum e Mitchell, 1998) Avrim Blum e Tom Mitchell. Combining labeled and unlabeled data with co-training. Em *COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory*, Madison, Wisconsin, EUA, 1998, p. 92–100. ACM Press.

(Boguraev et al., 2005) Branimir Boguraev, Jose Castaño, Rob Gaizauskas, Bob Ingria, Graham Katz, Bob Knippen, Jessica Littman, Inderjeet Mani, James Pustejovsky, Antonio Sanfilippo, Andrew See, Andrea Setzer, Roser Saurí, Amber Stubbs, Beth Sundheim, Svetlana Symonenko e Marc Verhagen. TimeML 1.2.1 - A formal specification language for events and temporal expressions, 2005. http://timeml.org/site/publications/timeMLdocs/timeml_1.2.1.html.

(Borbinha et al., 2007) José Luís Borbinha, Gilberto Pedrosa, Diogo Reis, João Luzio, Bruno Martins, João Gil e Nuno Freire. DIGMAP - Discovering our past world with digitised maps. Em László Kovács, Norbert Fuhr e Carlo Meghini, editores, *Research and advanced technology for digital libraries, 11th European Conference, ECDL 2007, Budapest, Hungary, September 16-21, 2007, Proceedings*. Springer Verlag, Berlim, Heidelberg, Setembro de 2007, p. 563–566.

(Bottou e LeCun, 2003) Léon Bottou e Yann LeCun. Lush: Reference manual, 2003. http://lush.sourceforge.net/.

(Braschler e Peters, 2004) Martin Braschler e Carol Peters. Cross-Language Evaluation Forum: Objectives, results, achievements. *Information Retrieval*, 7(1-2):7–31, Janeiro/Abril de 2004.

(Bruckschen et al., 2008a) Mírian Bruckschen, Fernando Muniz, José Guilherme Camargo de Souza, Juliana Thiesen Fuchs, Kleber Infante, Marcelo Muniz, Patrícia Nunes Gonçalves, Renata Vieira e Sandra Aluísio. Anotação lingüística em XML do corpus PLN-BR. Relatório Técnico NILC-TR-09-08, NILC, 2008.

(Bruckschen et al., 2008b) Mírian Bruckschen, José Guilherme Camargo de Souza e Renata Vieira. Tiger2XCES. Relatório técnico, Laboratório PLN – FACIN – PUCRS, 2008.

(Bruckschen et al., 2008c) Mírian Bruckschen, Renata Vieira e Sandro Rigo. SeRELeP--Olympics: hot topics for a news portal based on semantic types and named entities. Em *Proceedings of WebMedia 2008*, Vila Velha, Brasil, 2008.

(Brun e Hagège, 2004) Caroline Brun e Caroline Hagège. Intertwining deep syntactic processing and named entities detection. Em *Proceedings of the ESTal Conference*, Alicante, Espanha, Setembro de 2004, p. 195–206.

(Brun et al., 2007) Caroline Brun, Maud Ehrmann e Guillaume Jacquet. A hybrid system for named entity metonimy resolution. Em *Proceedings of the 4th International Workshop on Semantic Evaluations*, Praga, República Checa, Junho de 2007.

(Bunescu e Pasca, 2006) Razvan Bunescu e Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. Em *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, Trento, Itália, Abril de 2006, p. 9–16.

(Cabral, 2007) Luís Miguel Cabral. SUPeRB - Sistema uniformizado de pesquisa de referências bibliográficas. Tese de mestrado, Faculdade de Engenharia da Universidade do Porto, Março de 2007.

(Cabral et al., 2008) Luís Miguel Cabral, Diana Santos e Luís Fernando Costa. SUPeRB - Gerindo referências de autores de língua portuguesa. Em *VI Workshop Information and Human Language Technology (TIL'08)*, Vila Velha, ES, Brasil, 28-29 de Outubro de 2008.

(Cafarella et al., 2005) Michael J. Cafarella, Doug Downey, Stephen Soderland e Oren Etzioni. KnowItNow: Fast, scalable information extraction from the Web. Em *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canadá, Outubro de 2005, p. 563–570. ACL.

(Cardoso, 2006) Nuno Cardoso. Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Tese de mestrado, Faculdade de Engenharia da Universidade do Porto, Outubro de 2006. (Republicado como DI/FCUL TR-06-26, Departamento de Informática, Universidade de Lisboa, Novembro 2006).

(Cardoso, 2008) Nuno Cardoso. Novos rumos para a recuperação de informação geográfica em português. Em Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 11 de Setembro de 2008, p. 71–85. http://www.linguateca.pt/LivroL10/Cap11-Costaetal2008-Cardoso.pdf.

(Cardoso e Santos, 2007) Nuno Cardoso e Diana Santos. Directivas para a identificação e classificação semântica na colecção dourada do HAREM. Em Santos e Cardoso (2007a), p. 211–238. http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap16-SantosCardoso2007-CardosoSantos.pdf.

(Cardoso et al., 2006) Nuno Cardoso, Bruno Martins, Leonardo Andrade, Marcirio Silveira Chaves e Mário J. Silva. The XLDB group at GeoCLEF 2005. Em Carol Peters, Frederic Gey, Julio Gonzalo, Henning Müeller, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini e Maarten de Rijke, editores, *Accessing Multilingual information Repositories:6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Vienna, Austria, September 2005. Revised Selected papers (CLEF'2005)*. Springer, Berlim/Heidelberg, 2006, p. 997–1006.

(Cardoso et al., 2008a) Nuno Cardoso, David Cruz, Marcirio Silveira Chaves e Mário J. Silva. Using geographic signatures as query and document scopes in geographic IR. Em Peters et al. (2008), p. 802–810.

(Cardoso et al., 2008b) Nuno Cardoso, Mário J. Silva e Diana Santos. Handling implicit geographic evidence for geographic IR. Em *ACM 17th Conference on Information and Knowledge Management (CIKM 2008)*, Napa Valley, CA, EUA, 26-30 de Outubro de 2008, p. 1383–1384.

(Cardoso et al., 2008c) Nuno Cardoso, Patrícia Sousa e Mário J. Silva. The University of Lisbon at GeoCLEF 2008. Em Francesca Borri, Alessandro Nardi e Carol Peters, editores, *Cross Language Evaluation Forum: Working notes for the CLEF 2008 workshop*, Aarhus, Dinamarca, 17-19 de Setembro de 2008.

(Carreras et al., 2003) Xavier Carreras, Lluís Màrquez e Lluís Padró. Simple named entity extractor using AdaBoost. Em Walter Daelemans e Miles Osborne, editores, *Proceedings of Seventh Conference on Computational Natural Language Learning (CoNLL-2003)*, Edmonton, Canadá, 31 de Maio e 1 de Junho de 2003, p. 152–155. ACL.

(Carvalho e Gonçalo Oliveira, 2008) Paula Carvalho e Hugo Gonçalo Oliveira. Manual de utilização do Etiquet(H)AREM, 29 de Abril de 2008. http://www.linguateca.pt/aval_conjunta/HAREM/ManualUtilEtiquetHAREM.pdf.

(Carvalho e Mota, 2009) Paula Carvalho e Cristina Mota. Análise contrastiva do tratamento do TEMPO na primeira e segunda edição do HAREM. Em preparação, 2009.

(Carvalho, 2007) Paula Cristina Quaresma da Fonseca Carvalho. Análise e representação de construções adjectivais para processamento automático de texto. Adjectivos intransitivos humanos. Tese de doutoramento, Universidade de Lisboa, 2007.

(Cassan et al., 2006) Adán Cassan, Helena Figueira, André Martins, Afonso Mendes, Pedro Mendes e Cláudia Pinto. Priberam's question answering system in a cross-language enviroment. Em Alessandro Nardi, Carol Peters e José Luís Vicedo, editores, *Cross Language Evaluation Forum: Working notes for the CLEF 2006 workshop (CLEF 2006)*, Alicante, Espanha, 20-22 de Setembro de 2006.

(Chaves et al., 2005a) Marcirio Silveira Chaves, Bruno Martins e Mário J. Silva. GKB - Geographic Knowledge Base. Relatório Técnico 05–12, Departamento de Informática, Universidade de Lisboa, Julho de 2005. http://www.di.fc.ul.pt/tech-reports/05-12.pdf.

(Chaves et al., 2005b) Marcirio Silveira Chaves, Mário J. Silva e Bruno Martins. A geographic knowledge base for Semantic Web applications. Em Carlos Alberto Heuser, editor, *Proceedings do 20º Simpósio Brasileiro de Banco de Dados (SBBD)*, Uberlândia, MG, Brasil, 3-7 de Outubro de 2005, p. 40–54.

(Chinchor, 1998) Nancy Chinchor. MUC-7 named entity task definition (version 3.5). Em *Proceedings of the 7th Message Understanding Conference (MUC-7)*, Fairfax, VA, EUA, 29 de Abril-1 de Maio de 1998. Morgan Kaufmann.

(Chu-Carroll e Prager, 2007) Jennifer Chu-Carroll e John Prager. An experimental study of the impact of information extraction accuracy on semantic search performance. Em *Proceedings of the ACM Sixteenth Conference on Information and Knowledge Management, CIKM'07*, Lisboa, Portugal, Novembro de 2007, p. 505–514.

(Collins e Singer, 1999) Michael Collins e Yoram Singer. Unsupervised models for named entity classification. Em *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, MD, EUA, 1999, p. 100–110.

(Collovini et al., 2007) Sandra Collovini, Thiago I. Carbonel, Juliana Thiesen Fuchs, Jorge César Coelho, Lucia Helena Machado Rino e Renata Vieira. Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática. Em *TIL, V Workshop em Tecnologia da Informação e da Linguagem Humana*, Rio de Janeiro, RJ, Brasil, Julho de 2007, p. 1605–1614.

(Costa et al., 2007) Luís Costa, Paulo Rocha e Diana Santos. Organização e resultados morfolímpicos. Em Santos (2007a), p. 15–33.

(Cruse, 1986) D. A. Cruse. *Lexical semantics*. Cambridge University Press, 1986.

(Culotta e Sorensen, 2004) Aron Culotta e Jeffrey Sorensen. Dependency tree kernels for relation extraction. Em *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, Barcelona, Espanha, Julho de 2004, p. 423–429.

(Cunha et al., 2006) João Paulo Silva Cunha, Isabel Cruz, Ilídio Oliveira, António Sousa Pereira, César Telmo Costa, Ana Margarida Oliveira e Amândio Pereira. The RTS project: Promoting secure and effective clinical telematic communication within the Aveiro region. Em *eHealth 2006 High Level Conference*, Málaga, Espanha, Maio de 2006, p. 1–10.

(Dahl, 1973) Östen Dahl. On generics. Technical Report 6, Department of Linguistics, University of Göteborg, Abril de 1973.

(Dean e Ghemawat, 2008) Jeffrey Dean e Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

(Delboni, 2005) Tiago Marques Delboni. Expressões de posicionamento como fonte de contexto geográfico na Web. Tese de mestrado, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais de Belo Horizonte - UFMG, 26 de Agosto de 2005.

(Etzioni et al., 2005) Oren Etzioni, Michael J. Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld e Alexander Yates. Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, 165 (1):91–134, 2005.

(Fellbaum, 1998) Christiane Fellbaum, editora. *WordNet: An electronic lexical database*. The MIT Press, 1998.

(Ferreira et al., 2008) Liliana Ferreira, António Teixeira e João Paulo da Silva Neto. Ontology-driven vaccination information extraction. Em *5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008)*, Barcelona, Espanha, 12-13 de Junho de 2008.

(Ferrucci e Lally, 2004) David Ferrucci e Adam Lally. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

(Freitas, 2007) Maria Cláudia de Freitas. Elaboração automática de ontologias de domínio: discussão e resultados. Tese de doutoramento, Pontifícia Universidade Católica do Rio de Janeiro, Janeiro de 2007.

(Fürnkranz, 2002) Johannes Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2:721–747, 2002.

(Giampiccolo et al., 2008) Danilo Giampiccolo, Pamela Forner, Anselmo Peñas, Christelle Ayache, Dan Cristea, Valentin Jijkoun, Petya Osenova, Paulo Rocha, Bogdan Sacaleanu e Richard Sutcliffe. Overview of the CLEF 2007 multilingual question answering track. Em Peters et al. (2008), p. 200–236.

(Gillam, 1999) Richard Gillam. Finding text boundaries in Java : Overcoming differences in international style, 1999. http://www.ibm.com/developerworks/java/library/j-boundaries/boundaries.html.

(Gonzalez et al., 2007) Marco Gonzalez, Leonardo Cavalheiro Langie e Vera Lúcia Strube de Lima. Avaliação de sistemas de recuperação e categorização de textos: métodos e aplicações. Em Santos (2007a), p. 231–245.

(Gonçalo Oliveira et al., 2008) Hugo Gonçalo Oliveira, Paulo Gomes e Diana Santos. PAPEL: a dictionary-based lexical ontology for Portuguese. Em António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira e Paulo Quaresma, editores, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*. Springer Verlag, 8-10 de Setembro de 2008, p. 31–40.

(Grishman, 1999-2006) Ralph Grishman. Jet (Java Extraction Toolkit), 1999-2006. http://cs.nyu.edu/cs/faculty/grishman/jet/doc/Jet.html.

(Gross, 1986) Maurice Gross. *Grammaire transformationnelle du français - III - Syntaxe de l'adverbe*. ASSTRIL, 1986.

(Gruhl et al., 2004) D. Gruhl, L. Chavet, D. Gibson, J. Meyer, P. Pattanayak, A. Tomkins e J. Zien. How to build a WebFountain: An architecture for very large-scale text analytics. *IBM Systems Journal*, 43(1):64–77, 2004.

(Güting, 1994) Ralf Hartmut Güting. An introduction to spatial database systems. *The VLDB Journal*, 3(4):357–399, 1994.

(Hagège e Tannier, 2007) Caroline Hagège e Xavier Tannier. XRCE-T: XIP temporal module for TempEval campaign. Em *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Praga, República Checa, Junho de 2007, p. 492–495. ACL.

(Hagège et al., 2008) Caroline Hagège, Jorge Baptista e Nuno Mamede. Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o HAREM II, 2008. (Republicado neste volume como apêndice B). http://www.linguateca.pt/aval_conjunta/HAREM/2008_04_13_Tempo.pdf.

(Hearst, 1992) Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. Em *Proceedings of the Fourteenth Conference on Computational Linguistics*, Nantes, França, 23-28 de Julho de 1992, p. 539–545. ACL.

(Hill et al., 1999) Linda L. Hill, James Frew e Qi Zheng. Geographic names: the implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*, 5(1), Janeiro de 1999. http://www.dlib.org/dlib/january99/hill/01hill.html.

(Hirschman, 1998) Lynette Hirschman. The evolution of evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language*, 12(4):281–305, 1998.

(Ide e Romary, 2004) Nancy Ide e Laurent Romary. International standard for a linguistic annotation framework. *Natural Language Engineering*, 10(3-4):211–225, 2004.

(ISO19109, 2006) ISO19109. ISO 19109, Acessado em Novembro de 2006. https://www.seegrid.csiro.au/twiki/pub/Xmml/FeatureModel/19109_DIS2002.pdf.

(Ji e Grishman, 2006) Heng Ji e Ralph Grishman. Data selection in semi-supervised learning for name tagging. Em *Proceedings of the Workshop on Information Extraction Beyond The Document*, Sydney, Austrália, Julho de 2006, p. 48–55. ACL.

(Kazama e Torisawa, 2007) Jun'ichi Kazama e Kentaro Torisawa. Exploiting Wikipedia as external knowledge for named entity recognition. Em *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Praga, República Checa, Junho de 2007, p. 698–707.

(Krifka et al., 1995) Manfred Krifka, Francis Jeffrey Pelletier, Gregory N. Carlson, Alice ter Meulen, Gennaro Chierchia e Godehard Link. Genericity: An introduction. Em Gregory N. Carlson e Francis Jeffry Pelletier, editores, *The generic book*, 1995, p. 1–124. The University of Chicago Press.

(König e Lezius, 2003) Esther König e Wolfgang Lezius. The TIGER language - A description language for syntax graphs, formal definition. Relatório técnico, IMS, University of Stuttgart, 2003.

(Lansing, 2001) Jeff Lansing. Geoparser service draft candidate implementation specification 0.7.1, 2001. (OGC Paper 01-035). http://feature.opengis.org/members/archive/arch01/01-035.pdf.

(Lopes e Santos, 1993) Ana Cristina Macário Lopes e Pedro Santos. A condicionalidade das frases genéricas. *Cadernos de Semântica*, 17, 1993.

(Loureiro, 2007) João Miguel Sanches Loureiro. Reconhecimento de entidades mencionadas (obra, valor, relações de parentesco e tempo) e normalização de expressões temporais. Tese de mestrado, Instituto Superior Técnico, Universidade Técnica de Lisboa, Novembro de 2007.

(Makkonen e Ahonen-myka, 2003) Juha Makkonen e Helena Ahonen-myka. Utilizing temporal information in topic detection and tracking. Em Traugott Koch e Ingeborg Torvik Solvberg, editores, *Research and Advanced Technology for Digital Libraries, 7th European Conference, ECDL 2003, Trondheim, Norway, August 2003, Proceedings*. Springer-Verlag, 2003, p. 393–404.

(Malouf, 2002) Robert Malouf. Markov models for language-independent named entity recognition. Em Dan Roth e Antal van den Bosch, editores, *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, Taipei, Formosa, 31 de Agosto-1 de Setembro de 2002, p. 187–190.

(Mandl et al., 2008) Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker e Xie Xing. GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval track overview. Em Peters et al. (2008), p. 745–772.

(Manguinhas et al., 2008) Hugo Miguel Álvaro Manguinhas, Bruno Emanuel da Graca Martins e José Borbinha. A geo-temporal Web gazetteer service integrating data from multiple sources. Em *3rd IEEE International Conference on Digital Information Management*, Londres, Reino Unido, Novembro de 2008. IEEE.

(Mani, 2004) Inderjeet Mani. Recent developments in temporal information extraction. Em *Proceedings of RANLP'03*, Borovets, Bulgaria, 2004, p. 45–60. John Benjamins.

(Mani e Wilson, 2000) Inderjeet Mani e George Wilson. Robust temporal processing of news. Em *ACL '00: Proceedings of the 38th Annual Meeting of Association for Computational Linguistics*, Hong Kong, China, 2000, p. 69–76. ACL.

(Mani et al., 2004) Inderjeet Mani, James Pustejovsky e Beth Sundheim. Introduction to the special issue on temporal information processing. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(1):1–10, 2004.

(Markert e Nissim, 2007) Katja Markert e Malvina Nissim. SemEval-2007 task 08: Metonymy resolution at SemEval-2007. Em *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Praga, República Checa, Junho de 2007, p. 36–41. ACL.

(Martins, 2009) Bruno Martins. Geographically aware Web text mining. Tese de doutoramento, Faculdade de Ciências, Universidade de Lisboa, Março de 2009.

(Martins e Silva, 2007) Bruno Martins e Mário J. Silva. O HAREM e a avaliação de sistemas para o reconhecimento de entidades geográficas em textos em língua portuguesa. Em Santos e Cardoso (2007a), p. 79–86. http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap06-SantosCardoso2007-MartinsSilva.pdf.

(Martins et al., 2007a) Bruno Martins, Nuno Cardoso, Marcirio Silveira Chaves, Leonardo Andrade e Mário J. Silva. The University of Lisbon at GeoCLEF 2006. Em Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke e Maximilian Stempfhuber, editores, *Evaluation of multilingual and multi-modal information retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers*. Springer, Berlim/Heidelberg, 2007, p. 986–994.

(Martins et al., 2007b) Bruno Martins, Mário Silva e Marcirio Silveira Chaves. O sistema CaGE no HAREM - reconhecimento de entidades geográficas em textos em língua portuguesa. Em Santos e Cardoso (2007a), p. 97–112. http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap08-SantosCardoso2007-Martinsetal.pdf.

(Martins et al., 2008) Bruno Martins, Hugo Manguinhas e José Luis Borbinha. Extracting and exploring the geo-temporal semantics of textual resources. Em *Proceedings of the 2th IEEE International Conference on Semantic Computing (ICSC 2008), August 4-7, 2008, Santa Clara, California, USA*, 2008, p. 1–9. IEEE Computer Society.

(McCallum e Li, 2003) Andrew McCallum e Wei Li. Early results for named entity recognition with conditional random fields, feature induction and Web-enhanced lexicons. Em Walter Daelemans e Miles Osborne, editores, *Proceedings of Seventh Conference on Computational Natural Language Learning (CoNLL-2003)*, Edmonton, Canadá, 31 de Maio e 1 de Junho de 2003, p. 188–191. ACL.

(McDonald, 1996) David D. McDonald. Internal and external evidence in the identification and semantic categorization of proper names. Em Branimir Boguraev e James Pustejovsky, editores, *Corpus processing for lexical acquisition*. The MIT Press, Cambridge, MA, EUA & Londres, Inglaterra, 1996, p. 21–39.

(McDowell e Cafarella, 2008) Luke K. McDowell e Michael Cafarella. Ontology-driven, unsupervised instance population. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):218–236, 2008.

(Mikheev et al., 1998) Andrei Mikheev, Claire Grover e Marc Moens. Description of the LTG system used for MUC-7CHV. Em *Proceedings of the 7th Message Understanding Conference (MUC-7)*, Fairfax, VA, EUA, 29 de Abril-1 de Maio de 1998. Morgan Kaufmann.

(Mikheev et al., 1999) Andrei Mikheev, Marc Moens e Claire Grover. Named entity recognition without gazetteers. Em *Proceedings of EACL'99: Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, Bergen, Noruega, 8-12 de Junho de 1999, p. 1–8.

(Miller et al., 2004) Scott Miller, Jethran Guinness e Alex Zamanian. Name tagging with word clusters and discriminative training. Em Susan Dumais, Daniel Marcu e Salim Roukos, editores, *HLT-NAACL 2004: Main Proceedings*, Boston, Massachusetts, EUA, 2-7 de Maio de 2004, p. 337–342. ACL.

(Modesto et al., 2005) Marco Modesto, Álvaro R. Pereira Jr., Nívio Ziviani, Carlos Castillo e Ricardo Baeza-Yates. Um novo retrato da Web brasileira. Em *Proceedings of the XXII Seminário Integrado de Software e Hardware - SEMISH*, São Leopoldo, Brasil, 2005, p. 2005–2017.

(Molinier e Levrier, 2000) Christian Molinier e Françoise Levrier. *Grammaire des adverbes, Description des formes en -ment*. Droz, 2000.

(Mota, 2003) Cristina Mota. Avaliação conjunta de sistemas de reconhecimento de entidades mencionadas, 28 de Junho de 2003. (Apresentação no Encontro AvalON). http://www.linguateca.pt/aval_conjunta/acetatosAvalon/avalon-srem3.ppt.

(Mota, 2006) Cristina Mota. NooJ as a corpus annotator of named entities, 1-3 de Junho de 2006. http://nooj.matf.bg.ac.yu/pptpdf/13Cristina%20Mota-NooJasacorpusannotatorofnamedentities.pdf.

(Mota, 2009) Cristina Mota. How to keep up with language dynamics: A case study on named entity recognition. Tese de doutoramento, Instituto Superior Técnico, Universidade Técnica de Lisboa, Maio de 2009.

(Mota e Silberztein, 2007) Cristina Mota e Max Silberztein. Em busca da máxima precisão sem almanaques. O Stencil/NooJ no HAREM. Em Santos e Cardoso (2007a), p. 191–208. http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap15-SantosCardoso2007-MotaSilberztein.pdf.

(Móia, 2000) Telmo Móia. Identifying and computing temporal locating adverbials with a particular focus on Portuguese and English. Tese de doutoramento, Faculdade de Letras da Universidade de Lisboa, Fevereiro de 2000.

(Nadeau, 2007) David Nadeau. Semi-supervised named entity recognition: Learning to recognize 100 entity types with little supervision. Tese de doutoramento, University of Ottawa, Novembro de 2007.

(Navarro, 2001) Gonzalo Navarro. A guided tour to approximate string matching. *ACM Computer Surveys*, 33(1):31–88, 2001.

(Peeters, 2000) Bert Peeters. Setting the scene. Recent milestones in the lexicon-encyclopedia debate. Em Bert Peeters, editor, *The lexicon - encyclopedia interface*. Elsevier Science, Oxford, 2000, p. 1–52.

(Peters et al., 2008) Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Doug W. Oard, Anselmo Peñas, Vivien Petras e Diana Santos, editores. Springer, 2008.

(Petras et al., 2006) Vivien Petras, Ray R. Larson e Michael Buckland. Time period directories: A metadata infrastructure for placing events in temporal and geographic context. Em *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, Chapel Hill, NC, EUA, 2006, p. 151–160. ACM.

(Pinheiro e Aluísio, 2003) Gisele Montilha Pinheiro e Sandra Maria Aluísio. Córpus NILC: descrição e análise crítica com vistas ao projeto Lacio-Web. Relatório Técnico NILC-TR-03-03, NILC, Fevereiro de 2003. http://www.nilc.icmc.usp.br/lacioweb/downloads/NILC-TR-03-03.zip.

(Purves e Jones, 2007) Ross Purves e Chris Jones, editores. *GIR '07: Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, Lisboa, Portugal, 2007. ACM. ISBN 978-1-59593-828-2.

(Pustejovsky et al., 2003) James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer e Graham Katz. TimeML: Robust specification of event and temporal expressions in text. Em *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*, Tilburg, Holanda, 15-17 de Janeiro de 2003.

(Pustejovsky et al., 2005) James Pustejovsky, Robert Knippen, Jessica Littman e Roser Sauri. Temporal and event information in natural language text. *Computers and the Humanities*, 39(2-3):123–164, Maio de 2005.

(R Development Core Team, 2008) R Development Core Team. R: A language and environment for statistical computing, 2008. (ISBN 3-900051-07-0). http://www.R-project.org.

(Roberts e Hickl, 2008) Kirk Roberts e Andrew Hickl. Scaling answer type detection to large hierarchies. Em *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)LREC 2008*, Marraquexe, Marrocos, 28-30 de Maio de 2008, p. 1505–1510. European Language Resources Association (ELRA).

(Rocha e Santos, 2007a) Paulo Rocha e Diana Santos. CLEF: Abrindo a porta à participação internacional em avaliação de RI do português. Em Santos (2007a), p. 143–158.

(Rocha e Santos, 2007b) Paulo Rocha e Diana Santos. Disponibilizando a <OBRA>Colecção Dourada</OBRA> do <ACONTECIMENTO>HAREM</ACONTECIMENTO> através do projecto <LOCAL | ORGANIZACAO | ABSTRACCAO>AC/DC</LOCAL | ORGANIZACAO | ABSTRACCAO>. Em Santos e Cardoso (2007a), p. 307–326. http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap20-SantosCardoso2007-RochaSantos.pdf.

(Rocha e Santos, 2000) Paulo Alexandre Rocha e Diana Santos. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. Em Maria das Graças Volpe Nunes, editora, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, Atibaia, SP, Brasil, 19-22 de Novembro de 2000, p. 131–140. ICMC/USP.

(Roth e tau Yih, 2004) Dan Roth e Wen tau Yih. A linear programming formulation for global inference in natural language tasks. Em Hwee Tou Ng e Ellen Riloff, editores, *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, Boston, MA, EUA, 6 e 7 de Maio de 2004, p. 1–8.

(Ruiz-Casado et al., 2006) Maria Ruiz-Casado, Enrique Alfonseca e Pablo Castells. From Wikipedia to semantic relationships: A semi-automated annotation approach. Em *1st Workshop on Semantic Wikis: From Wiki to Semantics, at the 3rd European Semantic Web Conference (ESWC 2006)*, Budva, Montenegro, Junho de 2006.

(Santos, 1997) Diana Santos. The importance of vagueness in translation: Examples from English to Portuguese. *Romansk Forum*, 5:43–69, Junho de 1997. (Republicado como "A relevância da vagueza para a tradução, ilustrada com exemplos de inglês para português / The relevance of vagueness for translation: Examples from English to Portuguese". TradTerm 5 (1998), p. 41-70, 71-78.).

(Santos, 2006) Diana Santos. What is natural language? Differences compared to artificial languages, and consequences for natural language processing, 15 de Maio de 2006. (Palestra convidada no SBLP2006 e no PROPOR'2006). http://www.linguateca.pt/Diana/download/SantosPalestraSBLPPropor2006.pdf.

(Santos, 2007a) Diana Santos, editora. *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, 2007.

(Santos, 2007b) Diana Santos. Avaliação conjunta. Em Santos (2007a), p. 1–12.

(Santos, 2007c) Diana Santos. Evaluation in natural language processing, 6-17 de Agosto de 2007. (Curso na ESSLI 2007, Dublin, Irlanda.). http://www.linguateca.pt/Diana/download/EvaluationESSLLI07.pdf.

(Santos, 2007d) Diana Santos. O modelo semântico usado no Primeiro HAREM. Em Santos e Cardoso (2007a), p. 43–57. http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap04-SantosCardoso2007-Santos.pdf.

(Santos e Cardoso, 2007a) Diana Santos e Nuno Cardoso, editores. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. 2007. ISBN 978-989-20-0731-1. http://www.linguateca.pt/aval_conjunta/LivroHAREM/Livro-SantosCardoso2007.pdf.

(Santos e Cardoso, 2007b) Diana Santos e Nuno Cardoso. Balanço do primeiro HAREM e perspectivas de trabalho futuro. Em Santos e Cardoso (2007a), p. 87–94. http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap07-SantosCardoso2007-SantosCardoso.pdf.

(Santos e Cardoso, 2007c) Diana Santos e Nuno Cardoso. Breve introdução ao HAREM. Em Santos e Cardoso (2007a), p. 1–16. http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap01-SantosCardoso2007-SantosCardoso.pdf.

(Santos e Chaves, 2006) Diana Santos e Marcirio Silveira Chaves. The place of place in geographical IR. Em *Proceedings of GIR06, the 3rd Workshop on Geographic Information Retrieval (GIR 2006)*, Seattle, EUA, 10 de Agosto de 2006, p. 5–8.

(Santos e Rocha, 2005) Diana Santos e Paulo Rocha. The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. Em Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck e Bernardo Magnini, editores, *Multilingual information access for text, speech and images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*. Springer, Berlim/Heidelberg, 15-17 de Setembro de 2004 de 2005, p. 821–832. (Republicação de "CHAVE: topics and questions on the Portuguese participation in CLEF". In Carol Peters & Francesca Borri (eds.), Cross Language Evaluation Forum: Working Notes for the

CLEF 2004 Workshop (CLEF 2004) (Bath, UK, 15-17 September 2004), Pisa, Italy: IST-CNR, p. 639-648.).

(Santos et al., 2007) Diana Santos, Nuno Cardoso e Nuno Seco. Avaliação no HAREM: Métodos e medidas. Em Santos e Cardoso (2007a), p. 245–282. (Republicação de Relatório técnico DI-FCUL TR-06-17 : Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa. Novembro de 2006). http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap18-SantosCardoso2007-Santosetal.pdf.

(Santos et al., 2008a) Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling e Yvonne Skalban. Getting geographical answers from Wikipedia: the GikiP pilot at CLEF. Em Francesca Borri, Alessandro Nardi e Carol Peters, editores, *Cross Language Evaluation Forum: Working notes for the CLEF 2008 workshop*, Aarhus, Dinamarca, 17-19 de Setembro de 2008.

(Santos et al., 2008b) Diana Santos, Paula Carvalho, Cláudia Freitas e Hugo Gonçalo Oliveira. ReRelEM – Reconhecimento de Relações entre Entidades Mencionadas, 2008. (Republicado neste volume como apêndice C).

(Santos et al., 2008c) Diana Santos, Paula Carvalho, Cláudia Freitas e Hugo Gonçalo Oliveira. Segundo HAREM: Directivas de anotação, 2008. (Republicado neste volume como apêndice A). http:/www.linguateca.pt/HAREM/.

(Santos et al., 2008d) Diana Santos, Cláudia Freitas, Hugo Gonçalo Oliveira e Paula Carvalho. Second HAREM: new challenges and old wisdom. Em António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira e Paulo Quaresma, editores, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, Aveiro, Portugal, 8-10 de Setembro de 2008, p. 212–215. Springer Verlag.

(Santos et al., 2008e) Diana Santos, Hugo Gonçalo Oliveira, Cláudia Freitas, Cristina Mota e Paula Carvalho. Segundo HAREM: Balanço e perspectivas de futuro, 7 de Setembro de 2008. (Apresentação no Encontro do Segundo HAREM). http://linguateca.dei.uc.pt/harem/encontro/Santosetal2008SegundoHAREM.ppt.

(Sarmento et al., 2006) Luís Sarmento, Ana Sofia Pinto e Luís Cabral. REPENTINO - A wide-scope gazetteer for entity recognition in Portuguese. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006 (PROPOR'2006)*. Springer Verlag, Berlim/Heidelberg, 13-17 de Maio de 2006, p. 31–40.

(Saurí et al., 2006) Roser Saurí, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer e James Pustejovsky. TimeML annotation guidelines (2006), 2006. http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf.

(Schilder e Habel, 2001) Frank Schilder e Christopher Habel. From temporal expressions to temporal information: Semantic tagging of news messages. Em *Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing*, Toulouse, França, 2001, p. 1–8. ACL.

(Schmid, 1995) Helmut Schmid. TreeTagger, a language independent part-of-speech tagger. Relatório técnico, Institut fur Maschinelle Sprachverarbeitung, Universidade de Estugarda, 1995.

(Seco, 2007) Nuno Seco. MUC vs HAREM: A contrastive perspective. Em Santos e Cardoso (2007a), p. 35–41. http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap03-SantosCardoso2007-Seco.pdf.

(Seco et al., 2007) Nuno Seco, Nuno Cardoso, Rui Vilela e Diana Santos. A arquitectura dos programas de avaliação do HAREM. Em Santos e Cardoso (2007a), p. 283–306. http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap19-SantosCardoso2007-Secoetal.pdf.

(Silberztein, 2004) Max Silberztein. NooJ: A cooperative, object-oriented architecture for NLP. Em Claude Muller, Jean Royauté e Max Silberztein, editores, *INTEX pour la linguistique et le traitement automatique des langues*. Presses Universitaires de Franche-Comté, Besançon, França, 2004, p. 359–370.

(Silva et al., 2006) Mário J. Silva, Bruno Martins, Marcirio Silveira Chaves, Ana Paula Afonso e Nuno Cardoso. Adding geographic scopes to Web resources. *CEUS - Computers Enviroment and Urban Systems*, 30(4):378–399, 2006.

(Silva Romão, 2007) Luís Carlos da Silva Romão. Reconhecimento de entidades mencionadas em língua portuguesa: Locais, pessoas, organizações e acontecimentos. Tese de mestrado, Instituto Superior Técnico, Universidade Técnica de Lisboa, Novembro de 2007.

(Soon et al., 2001) Wee Meng Soon, Hwee Tou Ng e Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.

(Souza et al., 2008) José Guilherme C. de Souza, Patricia Nunes Gonçalves e Renata Vieira. Learning coreference resolution for portuguese texts. Em António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira e Paulo Quaresma, editores, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*. Springer Verlag, 8-10 de Setembro de 2008, p. 153–162.

(Tapanainen e Järvinen, 1997) Pasi Tapanainen e Timo Järvinen. A non-projective dependency parser. Em *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, DC, EUA, 1997, p. 64–71. ACL.

(Tesnières, 1959) Lucien Tesnières. *Eléments de syntaxe structurale*. Klincksieck, 1959.

(Toral e Muñoz, 2006) Antonio Toral e Rafael Muñoz. A proposal to automatically build and maintain gazetteers for named entity recognition by using Wikipedia. Em *Proceedings of the workshop on New Text Wikis and blogs and other dynamic text sources, 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Itália, Abril de 2006, p. 56–61.

(Uehara e Sato, 2005) Minoru Uehara e Nobuyoshi Sato. Information retrieval based on temporal attributes in WWW archives. Em *ICPADS '05: Proceedings of the 11th International Conference on Parallel and Distributed Systems (ICPADS'05)*, Fuduoka, Japan, 2005, p. 756–761. IEEE Computer Society.

(van der Vlist, 2003) Eric van der Vlist. *RELAX NG*. O'Reilly Media, Inc, 2003. http://books.xmlschemata.org/relaxng/page2.html.

(Vasconcelos Borges, 2006) Karla Albuquerque de Vasconcelos Borges. Uso de uma ontologia de lugar urbano para reconhecimento e extração de evidências geo-espaciais na Web. Tese de doutoramento, Universidade Federal de Minas Gerais, 2006.

(Verhagen et al., 2007) Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz e James Pustejovsky. SemEval-2007 task 15: TempEval temporal relation identification. Em *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Praga, República Checa, Junho de 2007, p. 75–80. ACL.

(Vilain et al., 1995) Marc Vilain, John Burger, John Aberdeen, Dennis Connolly e Lynette Hirschman. A model-theoretic coreference scoring scheme. Em *Proceedings of the 6th Message Understanding Conference (MUC-6)*, Columbia, Maryland, EUA, 6-8 de Novembro de 1995, p. 45–52. Morgan Kaufmann.

(Volz et al., 2007) Raphael Volz, Joachim Kleb e Wolfgang Mueller. Towards ontology-based disambiguation of geographical identifiers. Em *I3*, 2007, p. 19–22. http://ceur-ws.org/Vol-249/submission_132.pdf.

(Voss, 2005) Jakob Voss. Measuring Wikipedia. Em *10th International Conference of the International Society for Scientometrics and Informatics*, Julho de 2005, p. 221–231.

(Wilks, 2008) Yorick Wilks. The Semantic Web: Apotheosis of annotation, but what are its semantics? *IEEE Intelligent Systems*, 23:41–49, 2008.

(Wilson et al., 2001) George Wilson, Inderjeet Mani, Beth Sundheim e Lisa Ferro. A multilingual approach to annotating and extracting temporal information. Em *Proceedings of the Workshop on Temporal and Spatial Information Processing*, Toulouse, França, 7 de Julho de 2001, p. 81–87. ACL.

(Wu e Weld, 2007) Fei Wu e Daniel S. Weld. Autonomously semantifying Wikipedia. Em *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM'07)*, Lisboa, Portugal, 7-10 de Novembro de 2007, p. 41–50. ACM.

(Zesch et al., 2008) Torsten Zesch, Christof Müller e Iryna Gurevych. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. Em *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marraquexe, Marrocos, Maio de 2008, p. 1646–1652. European Language Resources Association (ELRA).

(Zhao e Grishman, 2005) Shubin Zhao e Ralph Grishman. Extracting relations with integrated information using kernel methods. Em *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 2005)*, Ann Arbor, Michigan, EUA, Junho de 2005, p. 419–426. ACL.

Esta bibliografia foi criada com a ajuda do SUPeRB (Cabral, 2007; Cabral et al., 2008), desenvolvido no âmbito da Linguateca por Luís Miguel Cabral a partir de um trabalho inicial de Paulo Rocha.

Conteúdo