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.

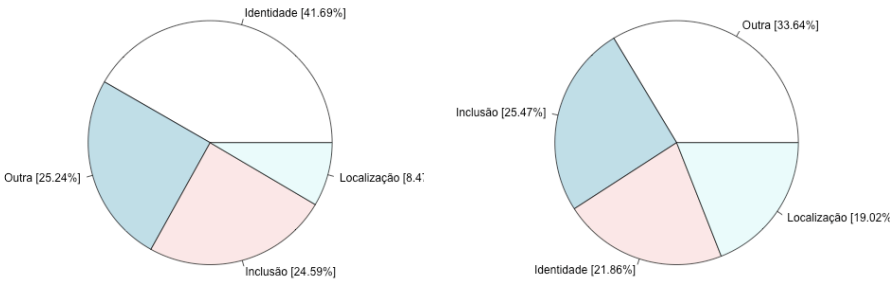
A CD do ReRelEM é composta por doze textos, 4417 palavras, 573 entidades mencionadas e 614 relações manualmente anotadas. Após a expansão das relações, tal como mencionado na secção anterior, a CD do ReRelEM passa a ter 6477 relações. A tabela 4.2 apresenta a distribuição das relações, antes e depois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

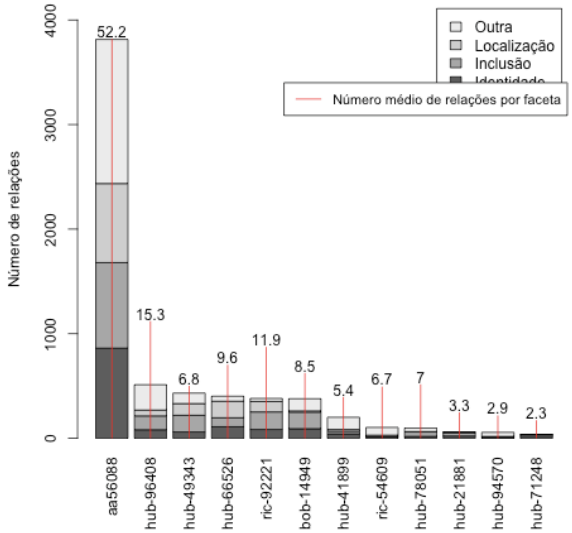


Figura 4.2: Distribuição de relações pelos documentos da CD do ReRelEM

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).

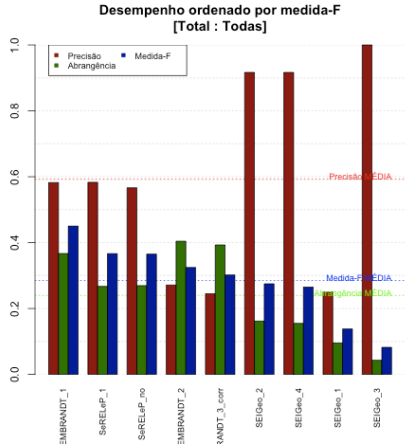
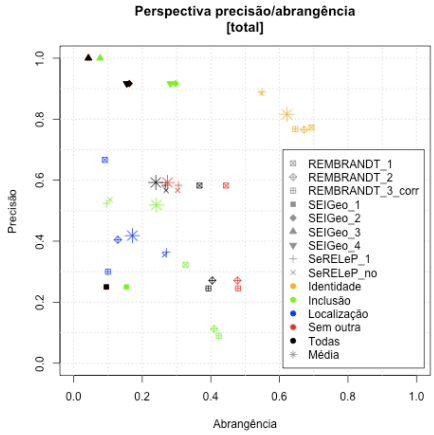


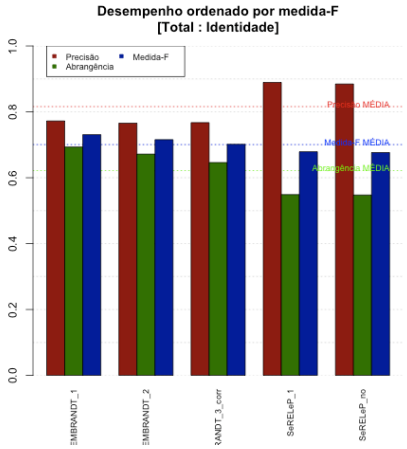
Figura 4.3: Avaliação de todas as relações no cenário total

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.

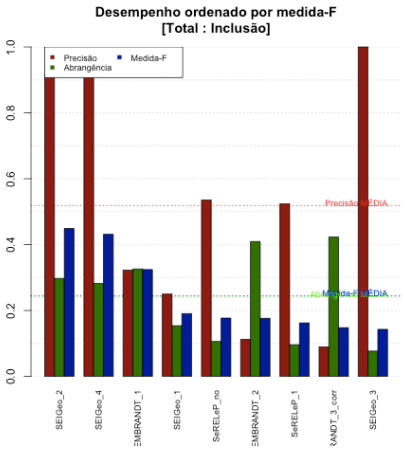
No primeiro gráfico dessa figura (4.4(a)) são comparados os vários cenários do ReRelEM em termos de precisão e abrangência: todas as relações (cenário *todas*), todas as relações menos a relação `outra` (cenário *Sem outra*), só relações de identidade (cenário *Identidade*), só relações de inclusão (cenário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



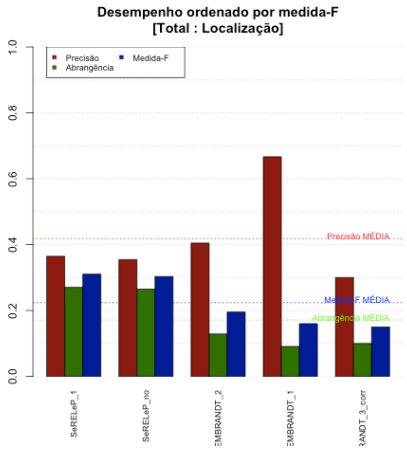
(a) Todos os cenários do ReRelEM



(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.