Directivas para identificação e classificação morfológica na colecção dourada do HAREM

Nuno Cardoso, Diana Santos e Rui Vilela

DI–FCUL TR–06–19

Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Campo Grande, 1749–016 Lisboa
Portugal

Technical reports are available at `http://www.di.fc.ul.pt/tech-reports`. The files are stored in PDF, with the report number as filename. Alternatively, reports are available by post from the above address.

Directivas para identificação e classificação morfológica na colecção dourada do HAREM

Nuno Cardoso[†], Diana Santos[‡] e Rui Vilela[*]
[†]Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa
[*]Departamento de Informática, Universidade do Minho
[‡]SINTEF ICT, Oslo
[†]ncardoso@xldb.di.fc.ul.pt, [‡]diana.santos@sintef.no,
[*]ruivilela@di.uminho.pt

Resumo

Neste relatório técnico apresentam-se as directivas usadas na compilação da colecção dourada do HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas (REM) em português, organizada pela Linguateca. A colecção dourada (dois conjuntos, de 129 e 128 textos de vários géneros literários) foi manualmente anotada com a identificação de nomes próprios e a sua classificação morfológica. Para o fazer, foi preciso primeiro chegar a um consenso e depois estudar os vários casos problemáticos que surgiram da análise dos próprios textos. O resultado desse processo de refinamento das directivas e de resolução (e subsequente documentação) dos casos que foram surgindo, encontra-se assim no presente documento.

O relatório está dividido em duas partes: a que relata as decisões feitas quanto à identificação das EM, e a que trata da sua classificação morfológica. Visto que a motivação primordial do HAREM era uma análise semântica leve, implicando pois o reconhecimento de várias categorias distintas (na tarefa de classificação semântica), a questão da morfologia é descrita em relação a essas categorias de EM.

Nota Preliminar

O HAREM constituiu a primeira avaliação (conjunta) para sistemas de reconhecimento de entidades mencionadas (REM) em textos em português [1, 5, 4, 10, 2, 8, 13, 6, 12, 3], no âmbito da responsabilidade da Linguateca de organizar avaliações conjuntas para a comunidade científica interessada no processamento computacional do português [7]. O HAREM foca a tarefa de identificação e classificação de entidades mencionadas (EM, ou seja, nomes próprios) no texto, uma tarefa relevante para várias áreas de processamento de linguagem natural (PLN) como a resposta automática a perguntas, a extracção de informação e a tradução automática, entre outras.

No HAREM procurámos desenvolver uma nova metodologia de avaliação em REM que abrangesse as especificidades da tarefa, tendo contemplado questões que não tinham sido ainda abordadas com profundidade suficiente pelos anteriores eventos de avaliação internacionais. Exemplos são a vagueza das EM, a caracterização morfológica das mesmas, e a avaliação de expressões apenas parcialmente identificadas.

O HAREM culminou na organização de dois eventos de avaliação, o principal em Fevereiro de 2005 e a sua sequela, o MiniHAREM, em Abril de 2006 [11], e contou com 10 sistemas participantes oriundos de 6 países (Brasil, Dinamarca, Espanha, França, México e Portugal). Os participantes enviaram um total de 38 saídas, ou seja, anotações automáticas da colecção de textos utilizando os seus sistemas REM,

que foram avaliadas. Os resultados foram publicados, e os relatórios detalhados do desempenho de cada sistema foram entregues aos respectivos participantes.

A 15 de Julho de 2006, a Linguateca organizou o Encontro do HAREM na Universidade do Porto, logo a seguir à Primeira Escola de Verão da Linguateca, que reuniu os participantes, organizadores e outros interessados no HAREM. No encontro, os participantes apresentaram os seus sistemas, a organização apresentou detalhadamente o seu trabalho, e todos debateram várias questões sobre o futuro do HAREM, numa sessão final, que demonstrou bem o interesse da comunidade em prosseguir com mais eventos de avaliação em REM. Está assim em curso a organização de um livro (electrónico) sobre o HAREM [9] englobando as comunicações nesse encontro e outra documentação relevante.

Os participantes no HAREM tiveram um papel activo no desenvolvimento da metodologia e na anotação das colecções de texto segundo a metodologia aprovada. Adicionalmente, a organização do HAREM desenvolveu uma plataforma de avaliação de sistemas de REM, para aferir o desempenho dos sistemas participantes, que se encontra publicamente disponível no sítio do HAREM, `http://www.linguateca.pt/HAREM`. Outra das contribuições do HAREM que reputamos valiosa foi a criação manual de uma colecção dourada para avaliação, juntamente com documentação extensa das directivas usadas, constituindo esses textos a primeira obra dedicada ao tratamento exaustivo em corpora da semântica dos nomes próprios em português.

É essa documentação que pretendemos tornar mais acessível agora, na forma de um par de relatórios técnicos, este e o seu gémeo [4], cuja forma final foi publicada na Web em finais de Março de 2006 por ocasião do MiniHAREM, cristalizando definitivamente as normas usadas no primeiro HAREM.

O trabalho de organização do HAREM enquadra-se no projecto da Linguateca, financiada pela Fundação para a Ciência e Tecnologia (FCT) através do projecto POSI/PLP/43931/2001, e co-financiada pelo POSI.

1 Introdução

Neste documento, apresentamos as directivas usadas na etiquetagem da colecção dourada do HAREM e, consequentemente, qual o comportamento esperado pelos sistemas que nele participem.

Começamos por descrever o formato do que consideramos um texto anotado com entidades mencionadas (EM), e qual a definição operacional da classificação morfológica destas.

Depois indicamos quais os critérios usados na anotação morfológica da colecção dourada. Noutro texto [4]será indicada a metodologia seguida na classificação semântica.

2 Regras gerais de etiquetagem

Cada EM rotula-se com uma etiqueta de abertura e uma etiqueta de fecho, cujo formato é semelhante ao das etiquetas usadas em XML. A etiqueta de abertura contém a categoria atribuída, e possui atributos como o tipo (`TIPO`) ou a classificação morfológica (`MORF`). Na etiqueta de fecho, coloca-se a categoria usada na etiqueta de abertura. Um exemplo de uma EM etiquetada é:

```
Certo: os <PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Beatles</PESSOA>.
```

Os nomes das categorias e dos tipos não incluem caracteres com acentos ou cedilhas, nem caracteres em minúsculas.

```
Certo:  <ABSTRACCAO TIPO="DISCIPLINA">Matemática</ABSTRACCAO>.
Errado: <ABSTRACÇÃO TIPO="DISCIPLINA">Matemática</ABSTRACÇÃO>.
Errado: <Abstraccao TIPO="Disciplina">Matemática</Abstraccao>.
```

Os valores dos atributos TIPO e MORF são rodeados por aspas.

Não pode haver nenhum espaço imediatamente a seguir à etiqueta de abertura e antes da etiqueta de fecho.

```
Certo:  O <PESSOA TIPO="INDIVIDUAL">João</PESSOA> é um professor.
Errado: O<PESSOA TIPO="INDIVIDUAL"> João</PESSOA> é um professor.
Errado: O <PESSOA TIPO="INDIVIDUAL">João </PESSOA>é um professor.
```

Se a EM contém espaços, esses devem manter-se inalterados.

```
Certo:  O <PESSOA TIPO="INDIVIDUAL">João Mendes</PESSOA> é um professor.
Errado: O <PESSOA TIPO="INDIVIDUAL">JoãoMendes</PESSOA> é um professor.
```

As aspas, parênteses, pelicas ou travessões não são para incluir na EM, se a englobarem como um todo (ver caso 1). No entanto, se fizerem parte integrante da mesma, são para incluir (caso 2).

Caso 1

```
Certo:  A "<OBRA TIPO="ARTE">Mona Lisa</OBRA>"
Errado: A <OBRA TIPO="ARTE">"Mona Lisa"</OBRA>
```

Caso 2

```
Certo: O <PESSOA TIPO="INDIVIDUAL">Mike "Iron" Tyson</PESSOA>
Certo: <PESSOA TIPO="INDIVIDUAL">John (Jack) Reagan</PESSOA>
```

2.1 Recursividade das etiquetas

Não pode haver etiquetas dentro de etiquetas, como se ilustra nos exemplos (errados) seguintes:

```
Errado: <PESSOA TIPO="GRUPO"><ORGANIZACAO TIPO="SUB">Bombeiros
        </ORGANIZACAO></PESSOA>
Errado: <ORGANIZACAO TIPO="INSTITUICAO">Departamento de
        <ABSTRACCAO TIPO="DISCIPLINA">Informática</ABSTRACCAO>
        do IST</ORGANIZACAO>
```

2.2 Vagueza na classificação semântica

No caso de haver dúvidas entre várias categorias ou tipos, deve utilizar-se o operador '|'. Por exemplo, em *"Ajudem os Bombeiros!"*, se se considerar que não existe razão para preferir uma das duas seguintes classificações para *Bombeiros*, nomeadamente entre <PESSOA TIPO=¨GRUPO¨> e <ORGANIZACAO TIPO=¨INSTITUICAO¨>, devem-se colocar ambas:

```
Certo: Ajudem os <PESSOA|ORGANIZACAO TIPO="GRUPO|INSTITUICAO">
       Bombeiros</PESSOA|ORGANIZACAO>!
```

Podem ser especificados mais do que uma categoria ou tipo, ou seja, <A|B|C|...>.

Caso a dúvida seja em diferentes atributos TIPO da mesma categoria, deve-se tambm repetir a respectiva categoria na etiqueta. Por exemplo, se a dúvida for relativa ao tipo de organização na frase "O ISR trata dessa papelada" (entre EMPRESA ou INSTITUICAO), deve-se repetir a categoria ORGANIZACAO tantas vezes quantos os tipos escolhidos:

```
Certo: O <ORGANIZACAO|ORGANIZACAO TIPO="EMPRESA|INSTITUICAO">ISR
</ORGANIZACAO|ORGANIZACAO> trata dessa papelada.
```

2.3 Vagueza na identificação

Se houver dúvidas (ou análises alternativas) de qual a identificação da(s) EM(s) que deverá ser considerada correcta, as várias alternativas são marcadas entre as etiquetas <ALT> e </ALT>, que delimitam e juntam as várias alternativas, que são separadas pelo caracter '|'. O exemplo abaixo mostra a etiquetagem a usar, quando não se consegue decidir por uma única identificação:

```
O <ALT><PESSOA TIPO="GRUPOMEMBRO">Governo de Cavaco Silva</PESSOA> |
Governo de <PESSOA TIPO="INDIVIDUAL">Cavaco Silva</PESSOA></ALT>
```

2.4 Critérios de identificação de uma EM

Uma EM deve conter pelo menos uma letra em maiúsculas e/ou algarismos.

```
Certo:  <TEMPO TIPO="DATA">Agosto</TEMPO>
Errado: <TEMPO TIPO="DATA">ontem de manhã</TEMPO>
```

A única excepção a esta regra abrange os nomes dos meses, que devem ser considerados EM (ou parte de EM), mesmo se grafados com minúscula. Esta excepção deve-se ao facto de haver grafia maiúscula em Portugal e minúscula no Brasil nesse caso.

```
Certo: <TEMPO TIPO="DATA">agosto de 2001</TEMPO>
```

Existe também um conjunto de palavras relativas a certos domínios que também são excepções a esta regra, e que serão descritas em detalhe abaixo.

Se uma determinada EM, etiquetada como tal, aparecer depois sem maiúsculas no mesmo texto ou noutro, não deve ser outra vez etiquetada, ou seja, uma EM tem de conter obrigatoriamente pelo menos uma letra maiúscula e/ou algarismos.

No entanto, o inverso não é verdade, isto é, uma palavra com pelo menos uma letra maiúscula ou um número pode não ser uma EM. Um caso clássico são as palavras que iniciam as frases, mas também há que considerar o uso excessivo de maiúsculas em certos géneros de textos, como a *web*, onde casos como *Contactos*, *História*, *Página Inicial*, *Voltar*, *Menu*, *E-mail*, entre outros, não devem ser por regra identificados como EM.

Aplicando o mesmo raciocínio, as frases totalmente escritas em maiúsculas (como acontece em títulos de destaque) deverão ser analisadas cuidadosamente, e só deverão conter etiquetas as EM claras. Por exemplo, se uma linha rezar "CLIQUE AQUI PARA VER A EDUCAÇÃO EM 1993", "EDUCAÇÃO" não deve ser considerada uma EM, uma vez que, naquele contexto, a palavra não deveria conter nenhuma maiúscula. No entanto, o ano deve ser marcado como TEMPO, de tipo DATA ou PERIODO.

Outro exemplo: "ABALO EM LISBOA SEM VÍTIMAS". Neste caso, consideramos correcto marcar "LISBOA" como EM, visto que assumimos que manteria a maiúscula se a frase não fosse exclusivamente grafada em maiúsculas. Note-se, de qualquer maneira, que estes casos caem um pouco fora do âmbito do HAREM, em que se utilizou um critério predominantemente gráfico, baseado nas convenções da língua escrita.

Palavras que foram incorrectamente grafadas apenas com minúsculas não são classificadas pelo HAREM como EM em caso nenhum.

2.5 Relação entre a classificação e a identificação

Embora a classificação deva ter em conta o significado da EM no texto, a identificação (ou seja a sua delimitação) deve restringir-se às regras das maiúsculas enunciadas acima. Ou seja, apenas a parte associada ao nome próprio deve ser identificada, embora classificada, se for caso disso, a entidade maior em que se enquadra. Vejam-se os seguintes exemplos:

```
Certo: a filha de <PESSOA TIPO="INDIVIDUAL">Giuteyte</PESSOA>
Certo: o tratado de <ACONTECIMENTO TIPO="EFEMERIDE">Tordesilhas
       </ACONTECIMENTO> dividiu o mundo
```

Embora apenas *Tordesilhas* tenha sido identificado, é *o tratado de Tordesilhas* que é classificado como um ACONTECIMENTO.

Isso também se aplica aos casos em que no texto um fragmento ou parte da EM é compreendida como relatando anaforicamente a uma entidade não expressa na sua totalidade. Por exemplo, na frase "A Revolução de 1930 foi sangrenta, e a de 1932 ainda mais", deve marcar-se *1932* como <ACONTECIMENTO TIPO="EFEMERIDE"> e não como <TEMPO TIPO="DATA">.

Nos casos em que há enganos de ortografia ou grafia no texto, em particular quando uma palavra 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>
```

Outras excepções, mais sistematicamente apresentadas, são as seguintes:

Para poder distinguir mais facilmente os casos de classes de objectos cujo nome inclui um nome próprio (geralmente de uma pessoa), adicionámos a seguinte regra de identificação para a categoria COISA: a preposição anterior também deve fazer parte da EM em *constante de Planck*, *bola de Berlim* ou *porcelana de Limoges*.

Por outro lado, consideramos que as EM de categoria VALOR e do tipo QUANTIDADE ou MOEDA devem incluir a unidade, independentemente de esta ser grafada em maiúscula ou minúscula.

Finalmente, no caso de doenças, formas de tratamento e certo tipo de acontecimentos consideramos aceitáveis um conjunto finito de nomes comuns precedendo a própria EM, cuja lista exaustiva se encontrará num apêndice futuro.

2.6 Escolha da EM máxima

Para evitar uma excessiva proliferação de EM com identificações alternativas, os sistemas e CD são construídos de forma a escolher a EM máxima, ou seja, aquela que contém, numa única interpretação possível, o maior número de palavras. Assim, e muito embora fosse possível ter tomado a decisão inversa e pedir, por exemplo, o máximo número de EM com uma interpretação possível separada, a escolha recaiu em preferir a EM maior.

Por exemplo:

```
Certo: O <PESSOA TIPO="CARGO">ministro dos Negócios Estrangeiros da
       Governo Sócrates</PESSOA>
Certo: <ORGANIZACAO TIPO="INSTITUICAO">Comissão de Trabalhadores da
       IBM Portugal</ORGANIZACAO>
Certo: <ACONTECIMENTO TIPO="EFEMERIDE">Jogos Olímpicos de Inverno de
       2006</ACONTECIMENTO>
```

As únicas excepções a esta regra são períodos descritos por duas datas, e intervalos de valores descritos por duas quantidades.

3 Regras gerais da tarefa de classificação morfológica

Considerámos como passíveis de ser classificadas morfologicamente (isto é, EM que devem ter o atributo MORF):

- As categorias PESSOA, ORGANIZACAO, COISA, ABSTRACCAO, ACONTECIMENTO, OBRA, e VARIADO na sua totalidade.
- Na categoria LOCAL, os tipos ADMINISTRATIVO e GEOGRAFICO.
- Na categoria TEMPO, o tipo CICLICO.

As seguintes EM não têm atributo MORF:

- A categoria VALOR na sua totalidade.
- Na categoria LOCAL, os tipos CORREIO.
- Na categoria TEMPO, o tipo HORA.

E finalmente, nos seguintes casos as EM podem ou não ter o atributo MORF:

- Na categoria LOCAL, o tipo VIRTUAL.
- Na categoria TEMPO, os tipos DATA e PERIODO.

Uma série de exemplos de aplicação são apresentados posteriormente para clarificar em que situações ocorrem estas excepções.

3.1 Género (morfológico)

Consideramos que o género de uma EM pode ter três valores:

M: EM com género masculino.

F: EM com género feminino.

?: Para os casos em que o género é indefinido.

3.2 Número

Consideramos que o número de uma EM pode ter três valores:

S: EM no singular.

P: EM no plural.

?: Para os casos em que o número é indefinido.

3.3 Exemplos de não atribuição de MORF na categoria LOCAL

Em alguns casos particulares da sub-categoria VIRTUAL, o atributo MORF foi omitido, devido ao facto de não ser possível avaliar morfologicamente números de telefone.

```
Certo: <LOCAL TIPO="VIRTUAL">(48) 281 9595</LOCAL>
```

Os casos que possuam a etiqueta MORF são, pelo contrário, geralmente casos em que a entidade é de outro tipo básico, mas é empregue no contexto na acepção de LOCAL.

```
Certo: Como capturar da <LOCAL TIPO="VIRTUAL" MORF="F,S">Internet</LOCAL>...
Certo: uma ordem do governo local publicada na "<LOCAL TIPO="VIRTUAL"
       MORF="F,S">Gazeta de Macau</LOCAL>" ordenava...
Certo: E só depois da publicação no '<LOCAL TIPO="VIRTUAL" MORF="M,S">
       Diário da República</LOCAL>' é que tomou-se conhecimento do traçado.
```

3.4 Exemplos de não atribuição de MORF na categoria TEMPO

Nos tipos PERIODO e DATA há casos distintos em que são aplicados o atributo MORF.

As datas especificadas em termos de anos ou de dias não possuem nunca a etiqueta MORF.

```
Certo: Este ano de <TEMPO TIPO="PERIODO">1982</TEMPO> deve...
Certo: <TEMPO TIPO="PERIODO">1914-1918</TEMPO>...
Certo: ia ser a <TEMPO TIPO="DATA">17 de Dezembro</TEMPO> porque saiu...
Certo: Em <TEMPO|TEMPO TIPO="DATA|PERIODO">91</TEMPO>, foram angariados...
```

As classificações que possuem atributo MORF são meses, séculos, e períodos históricos.

```
Certo: Cinema para o mês de <TEMPO TIPO="PERIODO" MORF="M,S">
       Maio</TEMPO>.
Certo: Mas já vem do <TEMPO TIPO="DATA" MORF="M,S">século XVI
       </TEMPO> o feriado.
Certo: os povoadores cristãos da <TEMPO|ACONTECIMENTO TIPO="PERIODO
       |EFEMERIDE" MORF="F,S">Reconquista</TEMPO|ACONTECIMENTO>.
Certo: Nesta <TEMPO TIPO="PERIODO" MORF="F,S">Primavera</TEMPO>,
       encontrei-me com os meus amigos.
Certo: está agora previsto para <TEMPO TIPO="DATA" MORF="M,S">Outubro
       </TEMPO> ou <TEMPO TIPO="DATA" MORF="M,S">Novembro</TEMPO>
```

4 Regras de atribuição de classificação morfológica

Considera-se o contexto e o texto adjacente para determinar o género e o número de uma dada EM, que à partida pode não ter género ou número definido.

Quando nem esse contexto nem o conhecimento lexical dos anotadores permite atribuir valores definidos, usa-se o valor '?', não especificado.

Exemplos:

```
Certo: O <PESSOA TIPO="INDIVIDUAL" MORF="M,S">João</PESSOA>
       é um professor.
Certo: A <PESSOA TIPO="INDIVIDUAL" MORF="F,S">João</PESSOA>
```

```
       não veio.
Certo: O apelido <ABSTRACCAO TIPO="NOME" MORF="?,S">João
       </ABSTRACCAO> é muito raro.
```

Ou seja, o nome *João* tem diferentes interpretações da sua classificação morfológica, consoante o contexto em que se encontra inserido.

4.1 Exemplos na categoria LOCAL

Algumas localidades administrativas são precedidas por artigo, determinando assim o género e número da entidade que designam (*o Porto*, *a Madeira*, *o Brasil*, *a Guarda*, *o Minho*, *o Rio Grande do Sul*, *os Estados Unidos*). Contudo, muitas outras não levam artigo e torna-se mais difícil de atribuir uma classificação morfológica.

Pareceu-nos em alguns casos haver consenso, tal como para *Portugal* (M,S), *Lisboa* (F,S), *Bragança* (F,S), *Brasília* (F,S), *Nova Iorque* (F,S) e *Colónia* (F,S), mas noutros casos apenas pudemos usar '?' no género, tal como em *Chaves*, *São Paulo* (estado ou cidade), *Castelo Branco*, *Braga* ou *Madrid*, excepto quando tal é especificado no contexto.

```
Certo: <LOCAL TIPO=ADMINISTRATIVO MORF="F,S">Leiria</LOCAL> é linda.
Certo: do concelho de <LOCAL TIPO="ADMINISTRATIVO" MORF="?,S">
       Aregos</LOCAL>.
Certo: todo o noroeste(de <LOCAL TIPO="ADMINISTRATIVO" MORF="?,S">
       Resende</LOCAL> ao...
Certo: em <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Portugal</LOCAL>
       seria...
Certo: ...aqui em <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">São Paulo
       </LOCAL>.
Certo: ...em <LOCAL TIPO="ADMINISTRATIVO" MORF="F,S">Nova Iorque
       </LOCAL> e saímos...
Certo: ...polícia de <LOCAL TIPO="ADMINISTRATIVO" MORF="F,S">Colónia
       </LOCAL> foram suspensos...
```

4.2 Exemplos na categoria ORGANIZACAO

Geralmente o número e género de uma organização são definidos pelo número e género da primeira palavra do nome, *Charcutaria Brasil* (F,S), *Armazéns do Chiado* (M,P), *Banco X* (M,S) ou *Caixa Y* (F,S), enquanto empresas internacionais têm geralmente associado o género feminino: A *Coca-Cola*, *a Benetton*, *a IBM*, *a Microsoft*, *a Sun*, *a Lotus*, *a Ferrari*, etc.

```
Certo: junto do <ORGANIZACAO TIPO="EMPRESA" MORF="M,S">Banco Sotto
       Mayor</ORGANIZACAO>.
Certo: Uma acção da <ORGANIZACAO TIPO="EMPRESA" MORF="F,S">Cartier
       </ORGANIZACAO>.
Certo: A acção da <ORGANIZACAO TIPO="EMPRESA" MORF="F,S">Portugal
       Telecom</ORGANIZACAO> resultou...
Certo: Esta página tem o apoio da <ORGANIZACAO TIPO="EMPRESA"
       MORF="F,S">IP</ORGANIZACAO>.
```

4.3 Exemplos na categoria PESSOA

No caso de GRUPOMEMBRO, ou seja, grupos de pessoas, o número é geralmente plural, e o género depende do sexo dos membros. *As Doce*, *os ABBA*, *os Xutos e Pontapés*, *os Beatles*, *as Spice Girls*, *os GNR*...

```
Certo: os <PESSOA TIPO="GRUPOMEMBRO" MORF="M,P">Stones</PESSOA>
Certo: e antes dos <PESSOA TIPO="GRUPOMEMBRO" MORF="M,P">R.E.M.</PESSOA>
Certo: <PESSOA TIPO="GRUPOMEMBRO" MORF="M,P">Peruanos</PESSOA>
       com diamantes falsos.
Certo: depois os <PESSOA TIPO="GRUPOMEMBRO" MORF="M,P">Mouros</PESSOA> que
       lhe deram o nome...
Certo: ...dez minutos o <PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Bastia
       </PESSOA>assegurou a presença na final...
```

4.4 Exemplos na categoria ACONTECIMENTO

No caso de EVENTO, os acontecimentos desportivos que tenham duas equipas, o número é singular, e o género é masculino, visto que correspondem a um jogo.

```
Certo: seguintes jogos: <ACONTECIMENTO TIPO="EVENTO" MORF="M,S">
       Penafiel-Rio Ave</ACONTECIMENTO>
Certo: e o <ACONTECIMENTO TIPO="EVENTO" MORF="M,S">
       Nacional-Académica</ACONTECIMENTO>
```

4.5 Exemplos na categoria ABSTRACCAO

No caso do tipo DISCIPLINA, a maior parte das EM que se refiram a disciplinas na área da educação tem género feminino, o número pode variar consoante o primeiro átomo.

```
Certo: e <ABSTRACCAO TIPO="DISCIPLINA" MORF="F,S">Filosofia</ABSTRACCAO>
       em todas as universidades.
Certo: <ABSTRACCAO TIPO= DISCIPLINA MORF= F,S >Ciência da Informação
       </ABSTRACCAO>.
Certo: futuros professores de <ABSTRACCAO TIPO="DISCIPLINA" MORF="F,S">
       Educação Física</ABSTRACCAO>.
Certo: As <ABSTRACCAO TIPO="DISCIPLINA" MORF="F,P">TI</ABSTRACCAO> são
       uma ferramenta...
```

Já em relação a desportos, o género é em geral masculino, embora haja alguns que, por serem originários de palavras portuguesas femininas, mantêm o género, tal como *Vela* ou *Luta livre*.

```
Certo: Página do time de <ABSTRACCAO TIPO="DISCIPLINA" MORF="M,S">
       Handebol</ABSTRACCAO>
```

Referências

[1] Nuno Cardoso. Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Proposta de tese de mestrado. Faculdade de Engenharia da Universidade do Porto. Janeiro de 2006. `http://www.linguateca.pt/documentos/NCardosoPropostaTese.pdf`.

[2] Nuno Cardoso. Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Apresentação no 2º Simpósio Doutoral da Linguateca, FCUL, Lisboa, Portugal, 10–11 de Abril de 2006. `http://www.linguateca.pt/documentos/CardosoSDL2006.pdf`.

[3] Nuno Cardoso. Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Tese de Mestrado, Faculdade de Engenharia da Universidade do Porto, Outubro 2006. Também disponível como Relatório Técnico DI-FCUL TR–06–26.

[4] Nuno Cardoso e Diana Santos. Directivas para identificação e classificação semântica na colecção dourada do HAREM. 29 de Março de 2006. Republicado como Relatório técnico DI-FCUL TR–06–18.

[5] Diana Santos. HAREM: the first evaluation contest for Named Entity Recognition in Portuguese. IST, Lisboa, Portugal. 24 de Fevereiro de 2006. `http://www.linguateca.pt/documentos/SantosISTFev2006.pdf`.

[6] Diana Santos. Reconhecimento de entidades mencionadas. Palestra convidada na PUC, Rio de Janeiro, Brasil, 18 de Maio de 2006. `http://www.linguateca.pt/documentos/SantosPalestraPUCRio2006.pdf`.

[7] Diana Santos, editora. *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press. 2007.

[8] Diana Santos e Nuno Cardoso. "A Golden Resource for Named Entity Recognition in Portuguese". Em Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, *Proceedings of the 7th International Workshop on Computational Processing of the Portuguese Language, PROPOR 2006*, volume 3960 de *Lecture Notes in Computer Science*, págs. 69–79, Itatiaia, Brasil, 13-17 Maio 2006. Springer.

[9] Diana Santos e Nuno Cardoso, editores. *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentação e actas do encontro*. Linguateca. 2007. Em preparação.

[10] Diana Santos, Nuno Cardoso e Nuno Seco. "Avaliação no HAREM: Métodos e Medidas". Relatório Técnico DI-FCUL TR–06–17, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa, Outubro 2006.

[11] Diana Santos, Nuno Cardoso, Nuno Seco e Rui Vilela. "Breve introdução ao HAREM". Em Diana Santos e Nuno Cardoso, editores, *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentação e actas do encontro*, Linguateca, 2007.

[12] Diana Santos, Nuno Seco, Nuno Cardoso e Rui Vilela. "HAREM: An Advanced NER Evaluation Contest for Portuguese". Em Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik e Daniel Tapias, editores, *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, págs. 1986–1991, Génova, Itália, 22–28 Maio 2006. ELRA.

[13] Nuno Seco, Diana Santos, Nuno Cardoso e Rui Vilela. "A Complex Evaluation Architecture for HAREM". Em Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, *Proceedings of the 7th International Workshop on Computational Processing of the Portuguese Language, PROPOR 2006*, volume 3960 de *Lecture Notes in Computer Science*, págs. 260–263, Itatiaia, Brasil, 13–17 Maio 2006. Springer.

Índice