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.