Capítulo 6

O HAREM e a avaliação de sistemas para o reconhecimento de entidades geográficas em textos em língua portuguesa

Bruno Martins e Mário J. Silva

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 6, p. 79–86, 2007.

O HAREM focou uma tarefa genérica de REM em textos na língua portuguesa (Santos et al., 2006), sendo que os tipos de entidades considerados foram mais genéricos do que apenas locais (por exemplo, pessoas, organizações, valores ou abstracções). Tem-se ainda que no caso específico dos locais, não foi feita qualquer atribuição dos mesmos a coordenadas ou a conceitos numa ontologia, e portanto a tarefa de desambiguação não foi considerada. A classificação semântica atribuída às entidades era também bastante genérica (ver capítulo 16), dividindo-se em `CORREIO`, `ADMINISTRATIVO`, `GEOGRAFICO`, `VIRTUAL` e `ALARGADO`. Note-se que muitos destes tipos de "locais" não correspondem a entidades físicas (ou seja, locais com correspondência no mundo real), e portanto um sistema como o CaGE, especialmente desenhado para a tarefa do reconhecimento e desambiguação de referências geográficas em páginas *web* (descrito no capítulo 8), não estaria à partida interessado na extracção destas entidades.

Estas características levam-nos a considerar que a tarefa de avaliação no HAREM, tal como foi definida, não é adequada para a avaliação da totalidade um sistema como o CaGE. Sistemas de extracção de informação focados no problema de extracção de referências geográficas apenas podem fazer uso do HAREM num cenário selectivo bastante restrito, por forma a medir a eficácia no reconhecimento simples e sem classificação geográfica ou desambiguação dos locais reconhecidos. Parece-nos importante que uma futura edição do HAREM considere o caso das referências geográficas de uma forma diferente, através da utilização de anotações na colecção dourada que sejam mais abrangentes e que melhor reflictam a temática geográfica. Nesse sentido, este capítulo apresenta algumas propostas para futuras edições do HAREM, as quais assentam sobretudo em alterações às directivas de anotação (ver capítulo 16).

6.1 Conceitos e trabalhos relacionados

A extracção de diferentes tipos de entidades mencionadas em texto é uma tarefa básica em processamento da linguagem natural, e um dos componentes chave da MUC (Chinchor, 1998b). O problema foi automatizado com sucesso, sendo frequente obter-se um desempenho semelhante ao de um ser humano. No entanto, o caso específico das referências geográficas levanta algumas considerações adicionais:

- As nossas entidades (referências geográficas e a sua classificação em tipos tais como ruas, cidades ou países) são mais específicas do que os tipos básicos considerados no MUC (pessoas, organizações, locais).
- A especificação completa de uma localização geográfica pode necessitar de relações espaciais (por exemplo, distância, direcção, ou topologia). Expressões contendo este tipo de relações devem ser consideradas como referências geográficas.

- Mais que reconhecer referências geográficas, é necessário fazer também a correspondência com os conceitos numa ontologia, uma vez que o reconhecimento só por si não atribui um sentido às referências reconhecidas. O REM é estendido com classificação semântica por tipo geográfico e com a associação a conceitos numa ontologia, ambos problemas mais complexos do que o simples reconhecimento (Kornai e Sundheim, 2003).

- Por forma a processar grandes quantidades de texto em tempo útil, os documentos individuais devem ser processados num tempo razoável. Esta restrição afecta seriamente a escolha de heurísticas a considerar pelo sistema. Infelizmente, tem-se que as questões de desempenho tendem a ser ignoradas em estudos de avaliação de REM, e o evento HAREM não foi uma excepção.

A investigação na especialização geográfica da tarefa genérica do REM está agora apenas a começar, mas existem já resultados publicados sobre este problema em concreto (Li et al., 2002; Olligschlaeger e Hauptmann, 1999; Smith e Mann, 2003; Smith e Crane, 2001; Schilder et al., 2004; Manov et al., 2003; Nissim et al., 2004; Leidner et al., 2003; Rauch et al., 2003). Por exemplo, a *Workshop on the Analysis of Geographical References* focou tarefas mais complexas que o simples reconhecimento de entidades geográficas em texto (Kornai e Sundheim, 2003). Alguns dos sistemas apresentados lidavam com a classificação e o mapeamento das referências geográficas nas coordenadas geodésicas correspondentes, embora apenas tenham sido reportadas experiências iniciais. Várias heurísticas foram já testadas, mas os sistemas variam muito nos tipos de classificação e desambiguação que efectuam, sendo que os recursos usados para avaliação também não se encontram normalizados. Não existe até hoje uma solução geral para o problema, e não existe ainda nenhum recurso de avaliação do tipo "colecção dourada" para a avaliação de sistemas de REM focados em referências geográficas.

Pensamos que o HAREM pode ter um papel importante no desenvolvimento desta área, possibilitando a avaliação de sistemas de extracção de informação que tratem o problema das referências geográficas em texto de uma forma mais abrangente do que apenas limitando-os a uma tarefa de reconhecimento simples.

6.2 Proposta para futuras edições do HAREM

Tal como exposto atrás, a colecção dourada e as directivas de anotação utilizadas pelo HAREM não se adequam à avaliação de sistemas que lidem explicitamente com o problema das referências geográficas. No entanto, pensamos ser possível fazer uma re-anotação da colecção dourada por forma a torná-la mais útil a este problema, não sendo para isso necessário um grande dispêndio de esforço. A nossa proposta para futuras edições do

HAREM vai essencialmente no sentido de considerar a sub-tarefa do reconhecimento das referências geográficas a um maior nível de detalhe.

No que resta desta secção abordamos três aspectos que nos parecem de especial importância, nomeadamente a existência de uma classificação semântica refinada para as entidades de categoria LOCAL, a existência de anotações para ontologias geográficas padrão, e a possibilidade dos sistemas considerarem sub-anotações e anotações alternativas para uma entidade. É ainda descrito outro aspecto que, embora de menor importância, deveriam ser também levado em conta numa futura edição do HAREM, nomeadamente a consideração do desempenho computacional como uma métrica de avaliação.

6.2.1 Classificação semântica refinada para as EM de categoria LOCAL

Em primeiro lugar, achamos que os tipos considerados para a classificação semântica das EM de categoria LOCAL deveriam ser estendidos por forma a melhor reflectir a temática geográfica. As etiquetas propostas no HAREM tiveram por base necessidades genéricas em processamento de linguagem natural. Como tal, pensamos que as etiquetas recomendadas para anotação da referências geográficas estão distantes das necessidades deste domínio específico, e carecem de uma revisão para futuras edições. Os tipos GEOGRAFICO e ADMINISTRATIVO, tal como se encontram definidos nas directivas de anotação, poderiam ser estendidos com sub-tipos mais específicos, tais como *rio*, *montanha* no primeiro caso, e *país*, *cidade*, *município* ou *freguesia* no segundo.

A hierarquia de tipos a considerar poderia, por exemplo, ser baseada num almanaque ou ontologia geográfica já existente (vários encontram-se amplamente divulgados, tais como o GeoNET (Chaves et al., 2005), o Getty TGN (Harpring, 1997), a *geonames ontology* (Vatant, 2006) ou o almanaque do projecto Alexandria Digital Library (Hill et al., 1999; Hill, 2000). Desta forma, teríamos uma classificação semântica para as EM de categoria LOCAL inspirada em trabalhos conhecidos na área do processamento de informação geográfica. Sistemas de anotação que, no seu funcionamento interno, utilizem uma hierarquia de tipos geográficos diferente, devem à partida conseguir traduzir os tipos geográficos por eles considerados para os tipos definidos nestes recursos. Estas próprias ontologias e almanaques incluem uma definição precisa de quais os tipos geográficos que consideram (Hill, 2000).

6.2.2 Geração de anotações para ontologias geográficas padrão

Além de uma classificação semântica mais refinada para as EM de categoria LOCAL, pensamos que a colecção dourada deveria conter as referências geográficas associadas a alguma forma de identificação única, por forma a se poder também testar uma tarefa de desambiguação completa. Poder-se-ia, mais uma vez, recorrer a almanaques ou ontologias geográficas padrão listados anteriormente. Exceptuando a GeoNetPT, todos os restantes recursos

são de âmbito global, contendo na sua maioria nomes geográficos em inglês. Contudo, a associação de uma referência geográfica em texto com o conceito correspondente na ontologia não depende obrigatoriamente do nome, mas sim do conceito que se encontra referenciado. Todos os recursos anteriormente listados descrevem conceitos geográficos relativos a Portugal, apresentando ainda alguns nomes em português (por exemplo, nomes alternativos para regiões geográficas importantes).

A anotação de cada local na colecção dourada seria estendida por forma a incluir uma referência para os identificadores correspondentes a esse conceito geográfico numa das ontologias. Este campo poderia incluir vários identificadores, no caso do local subjacente se encontrar definido por vários conceitos na ontologia, ou mesmo ser deixado em branco caso o local não se encontre definido.

Embora a anotação da colecção dourada com identificadores numa qualquer ontologia levasse à necessidade de que todos os sistemas que desejem fazer anotações desta forma partilhem esse mesmo recurso de informação externo, poder-se-ia considerar um cenário em que as referências geográficas fossem anotadas com as coordenadas geodésicas correspondentes, em lugar de se fazer as anotações com os conceitos na ontologia. Desta forma, a avaliação da tarefa de desambiguação podia ser feita com base nas coordenadas físicas reais associadas ao local, em lugar de depender de informação externa, sendo que cada sistema ficava livre de usar diferentes recursos para fazer a anotação. Ontologias padrão como as mencionadas anteriormente contêm coordenadas geodésicas, ou mesmo informação poligonal, para a maioria dos conceitos que definem, sendo que fazer a anotação da colecção dourada desta forma não nos parece problemático. Note-se no entanto que caso se usem coordenadas, a tarefa de avaliação necessita de contabilizar questões de imprecisão nas coordenadas (por exemplo, definindo uma distância mínima), visto que diferentes sistemas podem associar coordenadas diferentes ao mesmo conceito (devido, por exemplo, a factores de precisão numérica).

6.2.3 Possibilidade de considerar sub-anotações e anotações alternativas

As directivas de anotação do HAREM, tal como se encontram definidas, consideram que os nomes de locais que são dados como parte do nome de uma entidade de outro tipo (por exemplo, uma organização) não devem ser reconhecidos como tal. Por exemplo em *Câmara Municipal de Braga*, a totalidade da expressão deveria ser anotada como uma organização, sem que *Braga* fosse anotado como um local (ver secção 16.7.2). Para mais, o HAREM considerou o facto de os nomes dos locais muitas vezes assumirem um papel semântico diferente, não devendo nestes casos ser anotados como locais. Por exemplo, na frase *Portugal apoia envio de missão da ONU*, o nome *Portugal* deverá ser anotado como uma organização. Ainda que o papel semântico das entidades seja nestes casos claramente diferente do de uma referência explícita a uma localização, é também claro que estas entidades

continuam a ter uma forte conotação geográfica.

No sentido de resolver as questões colocadas acima, pensamos que as regras de anotação deveriam ser estendidas de forma a considerar sub-anotações e anotações alternativas. Nos casos como *Panificadora de Lisboa*, a expressão completa poderia ser anotada como uma organização e a palavra *Lisboa* nela contida poderia ser anotada como um local. Em casos como o da frase *Portugal apoia envio de missão da ONU*, deveria ser possível anotar *Portugal* de acordo com o seu papel semântico de local e o seu papel semântico de organização, mantendo-se desta forma os vários papéis semânticos possíveis para a palavra. Pretendemos assim que o HAREM continue a potenciar o desenvolvimento de sistemas que lidem com tarefas de desambiguação semântica das entidades, sem no entanto penalizar os sistemas que se focam numa tarefa de reconhecimento mais simples à semelhante do MUC, ou mais especializada num determinado tipo de entidades.

O HAREM poderia, por exemplo, considerar um formato de anotação que permitisse associar várias propriedades (possivelmente até de ontologias ou hierarquias de classificação diferentes) ao mesmo conteúdo textual. Em lugar de se providenciarem as anotações juntamente com o texto, poderíamos ter um esquema semelhante ao que se apresenta de seguida, no qual as anotações são feitas independentemente do texto, desta forma possibilitando que várias anotações possam facilmente ser feitas ao mesmo bloco do texto, ou até mesmo que as anotações sejam estendidas ao longo do tempo com novas classes de informação.

```
<DOCUMENTO>
<TEXTO>
Portugal envia missão de apoio.
</TEXTO>
<ANOTACOES>
<EM morf="m,s" palavra_inicio="1" palavra_fim="1" />
<EM classe="local" tipo="administrativo" subtipo="país"
geoid="GEO_1" palavra_inicio="1" palavra_fim="1" />
<EM classe="organização" tipo="administração"
palavra_inicio="1" palavra_fim="1" />
</ANOTACOES>
</DOCUMENTO>
```

Este esquema é bastante semelhante ao usado na proposta inicial do Open Geospatial Consortium para um serviço de anotação de referências geográficas em textos (Lansing, 2001). No entanto, um esquema desta natureza pressupõe a existência de uma atomização comum (isto é, partilhada por todos os sistemas participantes), visto que cada anotação é feita com base num átomo de início e fim para a mesma. Anteriores eventos de avaliação conjunta, focados no problema do REM, foram já baseados em colecções douradas

previamente atomizadas (Sang e Meulder, 2003). Contudo, uma conclusão importante do HAREM foi que a tarefa da atomização de textos em português é relativamente complexa, sendo que diferentes sistemas podem optar por fazer a atomização de diferentes formas (ver capítulos 18 e 19). Idealmente, a tarefa de avaliação deverá ser tanto quanto possível independente da atomização usada pelos sistemas, pelo que o esquema de anotação anterior poderá não ser o mais indicado.

Note-se ainda que o esquema de anotações alternativas em que cada entidade pode ter mais do que um tipo semântico associado deverá ser diferente do considerado nas directivas do HAREM para o caso da vagueza na classificação semântica. Em vez da anotação típica do HAREM, a qual não obedece aos requisitos da linguagem XML, e que se encontra exemplificada em baixo:

```
<LOCAL|ORGANIZACAO tipo="ADMINISTRATIVO|ADMINISTRACAO"
MORF="M,S">Portugal</LOCAL|ORGANIZACAO> envia missão de apoio.
```

Fazemos duas propostas de melhoria da representação de anotações das entidades mencionadas. A primeira seria de uma forma semelhante ao seguinte exemplo:

```
<EM classe="local|organizacao" tipo="local:administrativo"
subtipo="local:administrativo:pais" tipo="organização:administração"
morf="m,s" geoid="GEO_1"> Portugal </EM> envia missão de apoio.
```

Embora o exemplo anterior já obedeça aos requisitos da linguagem XML, a interpretação dos valores associados aos atributos das anotações `<EM>` pode ainda obrigar à criação de código adicional para processamento dos valores dos atributos. A segunda proposta teria um formato de anotação que define diferentes atributos XML para cada um dos tipos de entidades e classificações possíveis:

```
<EM local organizacao masculino singular tipo-local="administrativo"
subtipo-local="pais" tipo-organizacao="administracao" geoid="GEO_1">
Portugal</EM> envia missão de apoio.
```

6.2.4 Desempenho computacional

Além dos pontos referidos atrás, que essencialmente se relacionam com a anotação da colecção dourada de uma forma mais abrangente, há dois outros pontos que achamos importante rever, nomeadamente a consideração do desempenho computacional como uma métrica de avaliação. Esta é, quanto a nós, uma variável importante que afecta o desenvolvimento de qualquer sistema de REM, sendo que muitas vezes os sistemas optam por usar heurísticas mais simples em troca de ganhos significativos em desempenho. Juntamente com o envio das saídas dos sistemas, os participantes deveriam ser encorajados a partilhar

com os restantes o tempo que os seus sistemas demoraram a proceder à anotação dos textos, assim como a plataforma de *hardware* onde a anotação foi executada. Embora de uma forma algo informal, estes dados já permitiriam efectuar uma comparação dos diferentes sistemas participantes no que diz respeito à variável desempenho.

6.3 Conclusões

Neste capítulo discutimos as limitações do HAREM no que diz respeito aos sistemas focados no tratamento de referências geográficas. Em futuras edições, gostaríamos de ver o cenário das referências geográficas tratado em maior detalhe, nomeadamente através da anotação da colecção dourada de uma forma mais abrangente.