Capítulo 8

O sistema CaGE no HAREM - reconhecimento de entidades geográficas em textos em língua portuguesa

Bruno Martins, Mário J. Silva e Marcirio Silveira Chaves

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 8, p. 97–112, 2007.

Os documentos textuais (por exemplo os artigos publicados em jornais ou páginas *web*) são muitas vezes ricos em informação geográfica, e principalmente relevantes a uma dada comunidade local (como textos noticiosos sobre eventos num local específico, ou um página *web* sobre um comerciante local). A utilização de técnicas de prospecção de texto para extracção desta informação, por forma a oferecer capacidades de raciocínio geográfico a sistemas de recuperação de informação, é um problema interessante que tem vindo a ganhar notoriedade (Amitay et al., 2004; Gey et al., 2006; Jones et al., 2004; Kornai e Sundheim, 2003; Purves e Jones, 2004).

Ao contrário dos sistemas de informação geográfica (SIGs) tradicionais, que lidam com dados estruturados e geo-referenciados, a área da recuperação de informação geográfica foca o tratamento de informação não estruturada (documentos textuais, por exemplo). O reconhecimento e desambiguação de nomes de locais em texto torna-se portanto uma tarefa crucial na geo-referenciação destes recursos de informação (por exemplo, a anotação dos documentos com os âmbitos geográficos que lhes correspondem) (Amitay et al., 2004; Densham e Reid, 2003). Foram já vários os projectos de investigação que abordaram os problemas relacionados com a interpretação de terminologia geográfica em texto (Kornai e Sundheim, 2003; 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). Contudo, um problema na área é a não existência de corpora apropriados para a avaliação destes sistemas (Leidner, 2004; Martins et al., 2005), contendo as referências geográficas devidamente anotadas com coordenadas geodésicas ou com os conceitos correspondentes numa ontologia.

Embora o problema geral do REM seja uma tarefa conhecida em extracção de informação (EI), o caso particular do tratamento de referências geográficas apresenta ainda novos desafios (Sang e Meulder, 2003; Kornai e Sundheim, 2003). Mais do que anotar uma expressão de texto como uma localização, pretende-se que seja feita a anotação de forma a que a expressão geográfica seja inequivocamente descrita (Kornai e Sundheim, 2003; Leidner et al., 2003). A desambiguação completa requer que as referências geográficas sejam classificadas de acordo com o tipo (por exemplo, cidade ou país) e associadas explicitamente a conceitos numa ontologia geográfica. Esta informação (a ontologia mais os documentos anotados) pode então ser utilizada noutras tarefas, tais como a indexação e recuperação de documentos de acordo com os seus âmbitos geográficos (Jones et al., 2004).

No âmbito do desenvolvimento de um motor de busca geográfico para a *web* portuguesa, resultante da extensão do já existente `www.tumba.pt`, foi desenvolvido o CaGE (CaGE é acrónimo de *Capturing Geographic Entities*). Por desambiguação, entendemos o processo de fazer a associação entre as referências geográficas que são reconhecidas nos textos com conceitos numa ontologia geográfica.

A metodologia proposta no nosso sistema REM assenta na existência de uma ontologia geográfica contendo os nomes de locais e outros tipos de informação associados (por

exemplo, relações topológicas entre eles). Faz ainda uso de "regras de contexto" (as quais combinam pistas internas e externas, através da utilização dos nomes de locais, expressões com uma conotação geográfica, e presença de maiúsculas no texto) por forma a fazer o reconhecimento destas EM nos documentos. A abordagem tem a vantagem de ser relativamente simples (e como tal rápida, adaptando-se ao processamento de grandes volumes de texto da *web*) e de não requerer quaisquer dados de treino, os quais podem ser difíceis de obter para línguas como o português. Posteriormente, a desambiguação dos nomes geográficos reconhecidos é baseada em heurísticas adicionais, tais como a hipótese do "um referente por discurso", semelhante à proposta por Gale et al. (1992).

Estudos anteriores demonstraram que transformar ontologias ou dicionários existentes em sistemas REM úteis, ou por outro lado pegar num sistema REM e incorporar informação de uma ontologia, são ambos problemas não triviais (Cohen e Sarawagi, 2004). Esta foi a principal razão que nos levou a não adoptar à partida por um dos sistemas REM *open-source* existentes, tais como o GATE (Cunningham et al., 2002). Embora tomando como ponto de partida os trabalhos anteriores e as melhores práticas da área do REM, escolhemos abordar o problema através da construção de um novo sistema de raiz, focando nos aspectos particulares do tratamento das referências geográficas e do desempenho computacional. Este último é um aspecto crucial no processamento de colecções de documentos do tamanho da *web*.

Neste capítulo é descrito a participação do sistema CaGE no HAREM. Embora o HAREM não seja apropriado para a avaliação da totalidade um sistema como o CaGE (como argumentado no capítulo 6), considerámos ser interessante a participação num cenário selectivo, que nos permitisse medir a eficácia do sistema no reconhecimento simples (sem qualquer classificação semântica ou desambiguação dos locais reconhecidos) de referências geográficas em textos na língua portuguesa. São aqui apresentados os resultados obtidos, discutindo-se as adaptações feitas no sistema por forma a cumprir os requisitos da tarefa de avaliação.

8.1 Conceitos e trabalhos relacionados

Como descrito no capítulo 6, a extração de referências geográficas em páginas *web* portuguesas levanta algumas considerações adicionais. Os sistemas de REM tradicionais combinam recursos lexicais com uma cadeia de operações de processamento de complexidade variável (alguns sistemas utilizam etapas de anotação de morfossintáctica ou de desambiguação do sentido das palavras), consistindo de pelo menos um atomizador, listas de nomes de entidades, e regras de extracção. A atomização parte o texto em segmentos (tais como palavras, números e pontuação). As regras para o reconhecimentos de EM são a parte central do sistema, combinando os nomes presentes nos léxicos com elementos tais como a presença de maiúsculas na palavra e o contexto em que as entidades ocorrem.

Estas regras podem ser geradas à mão (a abordagem baseada em conhecimento) ou automaticamente (aprendizagem automática). O primeiro método requer um perito humano, enquanto que o último visa a obtenção automática de regras, através da análise de corpora anotados.

Os melhores métodos de aprendizagem automática para reconhecer entidades mencionadas são usualmente testados em textos jornalísticos, tendo sido reportados resultados acima dos 90% em termos da medida F na tarefa partilhada do CoNLL (Sang e Meulder, 2003). Contudo, estas abordagens requerem dados de treino balanceados e representativos, sendo que um problema ocorre quando estes dados não estão disponíveis ou são difíceis de obter. Este é geralmente o caso com línguas diferentes do inglês, ou em tarefas bastante específicas, tais como a do reconhecimento de referências geográficas.

O grau em que os léxicos ou ontologias ajudam na tarefa de REM também parece variar. Por exemplo, Malouf (2002) reportou que os léxicos não melhoraram o desempenho, enquanto que outros estudos reportam ganhos significativos usando recursos lexicais e expressões simples para o reconhecimento (Carreras et al., 2002). Mikheev et al. (1999) mostraram que um sistema de REM sem um léxico podia até comportar-se bem em muitos tipos de entidades, mas este não é caso quando se trata de entidades geográficas. 11 das 16 equipas que participaram na tarefa de REM do CoNLL-2003 integraram recursos lexicais nos seus sistemas, e todos reportaram ganhos de desempenho (Sang e Meulder, 2003). Uma conclusão importante da tarefa partilhada do CoNLL-2003 foi a de que a ambiguidade em referências geográficas é bi-direccional. O mesmo nome pode ser usado para mais do que um local (ambiguidade no referente), e o mesmo local pode ser referenciado por vários nomes (ambiguidade na referência). Este último tipo tem ainda a variante do mesmo nome poder ser usado como uma referência quer a um local, quer a outro tipo de entidades tais como pessoas ou empresas (ambiguidade na classe da referência).

8.2 Os recursos lexicais usados pelo sistema CaGE

Ao contrário de uma tarefa de REM convencional, onde a utilização de padrões de reconhecimento é muitas vezes suficiente, para reconhecer e desambiguar referências geográficas temos normalmente de nos basear num recurso de informação externo (como um léxico ou uma ontologia geográfica). Ao lidarmos com referências geográficas em texto, o nosso verdadeiro objectivo é a utilização das referências geográficas noutras tarefas de recuperação de informação, sendo que as referências devem obrigatoriamente estar associadas a uma representação única para o conceito geográfico subjacente.

No contexto dos sistemas de prospecção de texto, as ontologias são uma boa alternativa em relação aos léxicos simples, uma vez que estas modelam não só o vocabulário como também as relações entre conceitos geográficos. Estas relações podem fornecer pistas úteis para heurísticas de desambiguação.

Ontologia de Portugal		Ontologia Mundial	
Componente	Valor	Componente	Valor
Conceitos	418,743	Conceitos	12,654
Nomes	419,138	Nomes	15,405
Adjectivos	0	Adjectivos	400
Relações	419,072	Relações	24,570
Tipos de conceitos	58	Tipos de conceitos	14
Relações parte-de	419,115	Relações parte-de	13,268
Relações de adjacência	1,132	Relações de adjacência	11,302
Conceitos do tipo NUT1	3	Conceitos do tipo ISO-3166-1	239
Conceitos do tipo NUT2	7	Conceitos do tipo ISO-3166-2	3,976
Conceitos do tipo NUT3	30	Aglomerações Populacionais	751
Províncias	11	Locais	4,014
Distritos	18	Divisões Administrativas	3,111
Ilhas	11	Cidades Capitais	233
Municípios	308	Continentes	7
Freguesias	4,260	Oceanos	2
Zonas	3,594	Mares	3
Localidades	44,386		
Arruamentos	146,422		
Códigos Postais	219,691		
Conceitos com coordenadas	9,254	Conceitos com coordenadas	4,204
Conceitos com caixas limitadoras	0	Conceitos com caixas limitadoras	2,083
Conceitos com dados demográficos	308	Conceitos com dados demográficos	8,206
Conceitos com frequência do nome	0	Conceitos com frequência do nome	10,067

Tabela 8.1: Caracterização estatística das ontologias usadas no sistema CaGE.

No contexto do CaGE e do desenvolvimento de um motor de busca geográfico, duas ontologias foram criadas, para tal consolidando-se informação de diversas fontes de dados públicas. Uma das ontologias considera informação geográfica de âmbito global, enquanto que a outra foca o território português, a um maior nível de detalhe. Estes dois recursos influenciam claramente as experiências com o sistema, e deve portanto ser feita a sua caracterização. A informação considerada nas ontologias inclui nomes de locais e outros conceitos geográficos, adjectivos de local, tipos de locais (por exemplo, distrito, cidade ou rua), relações entre os conceitos geográficos (por exemplo, adjacente ou parte-de), dados demográficos, frequência em textos *web*, e coordenadas geográficas sob a forma de centróides e caixas limitadoras ("*bounding boxes*"). A Tabela 8.1 apresenta algumas estatísticas, sendo que em Chaves et al. (2005) é apresentada informação mais detalhada.

Cada conceito geográfico pode ser descrito por vários nomes. A Figura 8.1 ilustra a repetição de nomes geográficos nas duas ontologias. Para cada nome, são contados o número de conceitos diferentes que lhe correspondem. No caso da ontologia de Portugal,

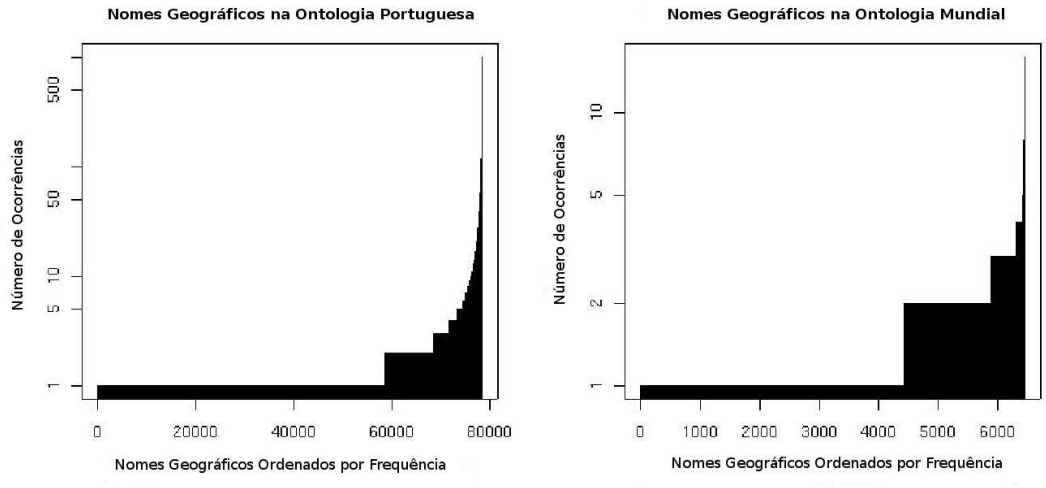

Figura 8.1: Frequência de repetição dos nomes geográficos nas ontologias.

os conceitos correspondentes a códigos postais não são apresentados, uma vez que eles já são por definição únicos e sem ambiguidade, e iriam confundir a interpretação do gráfico. As curvas apresentadas seguem a lei de Zipf (Zipf, 1949), como já notado em Li (1992), no sentido em que existe um número pequeno de nomes frequentes e uma longa lista de nomes pouco frequentes. Contudo, a Figura 8.1 também mostra que o número de nomes com múltiplas ocorrências (como a ambiguidade no referente) não é apenas um problema teórico, uma vez que eles correspondem a uma parte significativa dos nomes nas ontologias. A Tabela 8.2 apresenta exemplos de nomes geográficos comuns, correspondendo a vários conceitos.

A Figura 8.2 reforça as dificuldades associadas à utilização de nomes geográficos, desta feita mostrando a necessidade de considerar nomes compostos por múltiplas palavras. A figura separa a terminologia simples (ou seja, nomes geográficos compostos de apenas uma palavra), os nomes compostos (ou seja, nomes com várias palavras) e os casos difíceis (ou seja, nomes com hífen, abreviaturas e caracteres não alfa-numéricos). Mais uma vez, os códigos postais não são contabilizados, facilitando a interpretação do gráfico. Facilmente se pode observar que uma parte significativa dos nomes geográficos são compostos por mais do que uma palavra. As diferenças entre as duas ontologias advêm do facto da ontologia mundial conter apenas locais importantes (tais como países e cidades capitais), tendo portando um número maior de nomes simples.

Mesmo nos casos dos nomes simples podemos encontrar ambiguidade, visto que estes nomes também podem ser usados noutros contextos. Exemplos de palavras muito fre-

Ontologia de Portugal		Ontologia Mundial	
Nome do local	Número de locais	Nome do local	Número de locais
1 de Maio	618	Central	16
25 de Abril	881	Granada	10
Almada	15	Madrid	5
Bairro Alto	28	Portugal	4
Braga	11	Rio de Janeiro	4
Campo Grande	20	Roma	4
Lisboa	41	Taiwan	4
Seixal	42	Venezuela	4
Vila Franca	16	Washington	6

Tabela 8.2: Exemplos de nomes geográficos e o número de conceitos correspondentes nas ontologias portuguesa e mundial.

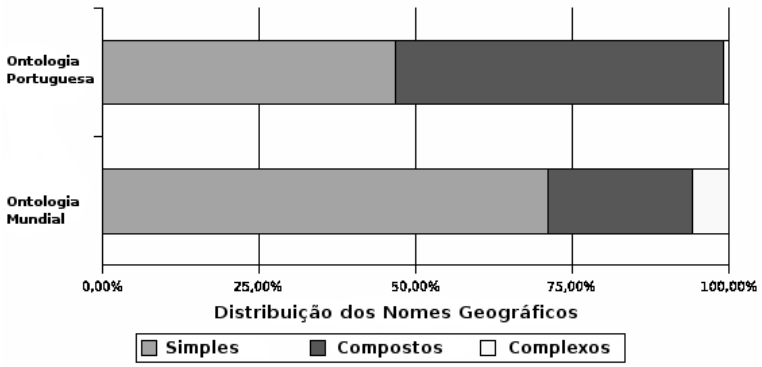

Figura 8.2: Distribuição dos nomes geográficos nas ontologias considerando a sua complexidade.

quentes que são também nomes geográficos são apresentados na Tabela 8.3. A mesma tabela mostra ainda que os nomes geográficos são muitas vezes homónimos com outros tipos de entidades, tais como pessoas (ou seja, ambiguidade na classe da referência). Por forma a lidar com este último tipo de ambiguidade, gerámos uma lista de excepções, com nomes que embora possam ter uma conotação geográfica, são muito mais frequentemente usados noutros contextos. Esta lista foi compilada através das nossas experiências (nomes que eram incorrectamente anotados foram colocados na lista), e através de um procedimento simples baseado em estatísticas num corpus da *web* (por exemplo, nomes que aparecem mais frequentemente escritos só em minúsculas do que com maiúsculas presentes foram adicionados à lista, seguindo a ideia que a detecção de letras maiúsculas pode distinguir entidades mencionadas).

Além da ontologia geográfica e da lista de excepções, a nossa técnica requer ainda

Palavras frequentes	Nomes de pessoas	
Homónimas com locais	Nome próprio	Nome de local
Central	Camilo Castelo Branco	Castelo Branco
Cruz	Cesária Évora	Évora
Direita	Teófilo Braga	Braga
Sol	Vergílio Ferreira	Ferreira
Nova	Irene Lisboa	Lisboa
Paz	Faria Guimarães	Guimarães
Casal	Almada Negreiros	Almada
Esta	Salgueiro Maia	Maia
Meio	Leonardo Coimbra	Coimbra

Tabela 8.3: Palavras frequentes e nomes de pessoas que incluem nomes de locais.

Tipo de expressão	Expressão
Identificadores	cidade, município, distrito, rua, avenida, rio, ilha, montanha, vale, país, continente, zona, região, condado, freguesia, deserto, província, povoado, aldeia, monte, vila, república, península
Localização	fora de, nos arredores de, dentro de, entre, em acima, ao longo, atrás, acima, ao lado, à esquerda, à direita
Distância Relativa	adjacente, longe de, perto de, próximo de
Orientação	este, norte, sul, oeste, oriente, ocidente, sudeste, sudoeste, nordeste, noroeste
Outras Expressões	"cidades como", "e outras cidades", "cidades, incluindo", "cidades, especialmente", "uma das cidades", "cidades tais como", padrões semelhantes para outros identificadores

Tabela 8.4: Expressões de contexto associadas a referências geográficas.

regras para efectuar o reconhecimento e desambiguação. Estas regras combinam pistas internas e externas, disparando quando um nome candidato está perto de uma expressão de contexto sugestiva. Estudos anteriores mostraram que as referências geográficas contêm muitas vezes informação sobre o tipo de locais a que se referem (por exemplo, *cidade de Lisboa*), sendo portanto passíveis de ser reconhecidas desta forma. As referências geográficas podem também conter expressões que denotem relações de distância ou de posicionamento relativo. A Tabela 8.4 exemplifica as expressões consideradas no desenvolvimento do CaGE, tendo essa lista sido baseada em trabalhos anteriores (Delboni, 2005; Kohler, 2003).

8.3 Reconhecimento e desambiguação de referências geográficas

A Figura 8.3 ilustra o procedimento utilizado pelo CaGE para identificar e desambiguar referências geográficas em texto, reflectindo os seus quatro estágios principais: pré-processamento, identificação, desambiguação e geração de anotações. O resto desta secção descreve cada um destes estágios em detalhe.

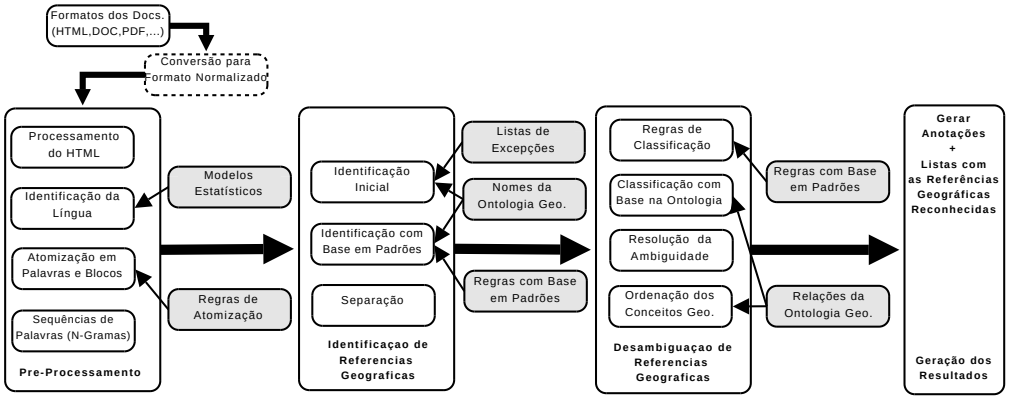

Figura 8.3: Arquitectura geral do sistema CaGE.

8.3.1 Operações de pré-processamento

A etapa de pré-processamento envolve as seguintes sub-etapas: conversão de formatos, processamento do HTML, classificação de língua, atomização e emparelhamento de *n*-gramas. As três primeiras são específicas do tratamento de textos provenientes da *web* no contexto do motor de busca geográfico. Estas foram desactivadas no contexto da produção de saídas para o HAREM, uma vez que apenas estávamos na presença de ficheiros de texto simples escritos na língua portuguesa.

A atomização das palavras e reconhecimento de frases é baseada numa tabela com os "pares de contexto" formados pelos caracteres que ocorrem antes e depois de uma dada posição no texto. Por exemplo, uma tabela para o reconhecimento de palavras coloca uma interrupção entre caracteres de pontuação e letras, mas não entre letras consecutivas ou entre caracteres de espaçamento consecutivos. As regras consideradas baseiam-se nas propostas pela Linguateca para o tratamento de corpora no projecto AC/DC (Santos e Sarmento, 2003), e descritas em `http://acdc.linguateca.pt/acesso/atomizacao.html`. Esta técnica lida com a grande maioria dos problemas de ambiguidade que ocorrem na atomização. É também simples de implementar, uma vez que a tabela de "pares de contexto" é simplesmente uma matriz de valores booleanos, em que cada linha e coluna correspondem

a um caracter ou grupo de caracteres. Um eixo representa o contexto anterior à posição, e o outro o contexto depois.

Depois do texto atomizado, as frases são divididas nos seus *n*-gramas constituintes. Isto é conseguido movendo uma janela sobre o texto de cada frase, tomando-se todas as possíveis sequências de *n* palavras consecutivas.

8.3.2 Identificação de referências geográficas

A etapa de identificação envolve a detecção de todas as sequências de *n*-gramas passíveis de constituir uma referência geográfica. Esta consiste de três sub-etapas, nomeadamente identificação inicial, identificação baseada em padrões e separação.

A identificação inicial envolve a aplicação de regras que combinam os nomes de locais na ontologia, expressões de contexto, e termos com a primeira letra em maiúsculas. As sequências de *n* palavras consecutivas identificadas na primeira etapa são inicialmente mapeadas nos nomes existentes na ontologia. Esta abordagem simples é suficiente para fazer a detecção de muitas referências, mas a ambiguidade pode conduzir a muitos erros. Por esta razão, apenas permitimos a detecção desta forma para certos tipos de conceitos geográficos na ontologia, particularmente os tipos que correspondem a regiões grandes e importantes (por exemplo, países e cidades om mais de 100.000 habitantes). Descartam-se ainda nesta fase de detecção simples os nomes geográficos presentes numa lista de excepções. Esta lista de exclusão tenta lidar com o problema de nomes muito frequentes que são usados noutros contextos que não o geográfico.

Dadas as limitações da identificação inicial, a sub-etapa seguinte usa regras para combinar os nomes geográficos com expressões de contexto e termos em maiúsculas. A Tabela 8.4 ilustra as expressões de contexto que foram consideradas. Algumas destas regras são relativamente complexas, combinando diferentes referências (por exemplo, *cidades tais como A, B ou C*) ou qualificando referências geográficas de acordo com critérios espaciais ou de posicionamento (por exemplo, *perto da cidade de X*). Contudo, o algoritmo de aplicação de regras, implementado por um autómato finito, é rápido. As regras são especificadas num ficheiro de texto, encontrando-se codificadas numa linguagem semelhante à das expressões regulares (as diferenças prendem-se com a utilização da informação de maiúsculas e dos nomes na ontologia).

É de notar que as regras consideradas na a geração de saídas para o HAREM têm algumas diferenças em relação às regras consideradas para a utilização normal do sistema. Em particular, fazemos para o HAREM um uso diferente do termos em maiúsculas, no sentido em que as directivas de anotação indicam que todas as entidades devem obrigatoriamente ter a primeira letra maiúscula[1], enquanto que no contexto das páginas *web* consideramos que os locais ocorrem muitas vezes em minúsculas. Têm-se ainda que no contexto do

[1] **Nota dos editores:** Com algumas pequenas excepções, documentadas na secção 16.1.4.

HAREM estamos interessados em reconhecer locais que não se encontrem descritos na ontologia (ou seja, reconhecidos apenas pela aplicação de regras), enquanto que nas aplicações normais do CaGE estamos apenas interessados em locais que possam ser mapeados em identificadores na ontologia, por forma a serem posteriormente usados noutras tarefas.

Finalmente, na sub-etapa de separação, os *n*-gramas passíveis de constituírem mais do que uma referência geográfica são detectados e os problemas de separação são resolvidos. Se um *n*-grama constitui uma referência, então todos os seus *n*-gramas constituintes são descartados, mantendo-se apenas a referência para o mais geral. As expressões complexas (por exemplo, *cidades tais como A, B, C*) são, neste caso, tratadas como uma excepção, mantendo-se cada referência independentemente.

8.3.3 Desambiguação de referências geográficas

Depois das referências geográficas terem sido identificadas, segue-se uma etapa de desambiguação. Esta envolve quatro sub-etapas, nomeadamente aplicação de regras de classificação, classificação baseada na ontologia, comparação das referências ambíguas com as que já se encontram desambiguadas e ordenação dos conceitos geográficos correspondentes. As regras de classificação são baseadas nas expressões de identificação usadas na etapa anterior, uma vez que muitas referências contêm palavras que podem ser usadas para inferir o tipo implícito ao conceito geográfico referenciado (por exemplo, em *cidade de Lisboa*, sabemos que a referência diz respeito à cidade e não a outro conceito).

A classificação baseada na ontologia usa as relações semânticas presentes na mesma para determinar o tipo correcto das referências. Pode-se dar o caso simples da uma referência, contendo ou não o tipo geográfico correspondente, poder ser mapeada num único conceito. Contudo, quando mais do que um conceito da ontologia está potencialmente a ser referenciado, usamos a hipótese de "um referente por discurso" para tentar a desambiguação. A hipótese diz que uma referência geográfica feita na mesma unidade de texto (ou seja, no mesmo parágrafo) refere-se ao mesmo local, ou a locais relacionados. Hipóteses semelhantes já foram usadas no passado no problema da desambiguação do sentido das palavras (Gale et al., 1992). A existência de uma relação entre dois conceitos é dada pela ontologia, sendo que consideramos os casos em que o nome ambíguo é um nome alternativo, uma região mais geral, uma região equivalente, ou uma região adjacente a um outro nome que já se encontre desambiguado.

O último estágio faz a comparação das referências ainda não desambiguadas com outras que já o tenham sido. Esta comparação é feita usando variações dos nomes das referências ambíguas, por forma a lidar com o problema de nomes truncados ou erros ortográficos. A comparação entre dois nomes é feita de acordo com as seguintes regras:

- Ambos os nomes devem ter o mesmo número de palavras.
- Maiúsculas, acentos e hífens são todos ignorados ao fazer a comparação.

- Palavras abreviadas são equivalentes (por exemplo, *Lis.* é dito equivalente a *Lisboa*).
- Palavras não abreviadas devem divergir no máximo em um caracter diferente, um caracter extra, ou um caracter a menos (por exemplo, *Lisboa* é dito equivalente a *Lusboa*).

Finalmente, nos casos não cobertos pelas heurísticas acima, mantemos a associação com todos os conceitos possíveis da ontologia. No entanto, ordenamos os conceitos possíveis de acordo com a importância do conceito geográfico referenciado, de acordo com as seguintes heurísticas:

- Regiões maiores (conceitos de topo na ontologia) são preferidas, uma vez que é mais provável que sejam mencionadas.
- Regiões com maior população são preferidas, pela mesma razão.

Em aplicações que requeiram a associação de cada referência a um único conceito, podemos usar estas heurísticas para escolher qual a referência mais provável, em lugar de manter a associação a todos os conceitos (Leidner et al., 2003).

8.3.4 Geração de anotações para a ontologia

A última etapa prende-se com a geração das saídas, mantendo-se cada referência geográfica associada com os conceitos correspondentes na ontologia. O formato usado pelo CaGE facilita o desenvolvimento de outras ferramentas de recuperação de informação, as quais usem as referências geográficas extraídas dos textos.

Sistemas anteriores optaram por associar a cada referência as coordenadas geodésicas correspondentes (Leidner et al., 2003), mas no CaGE optamos por associar as referências aos identificadores dos conceitos na ontologia. Isto traz algumas vantagens, nomeadamente ao permitir lidar com regiões imprecisas, ou no facto de não precisarmos de lidar com questões de precisão numérica associadas às coordenadas. Além de anotar cada referência com os conceitos na ontologia, mantemos ainda a associação com o tipo de conceito geográfico. O texto é anotado com etiquetas SGML correspondendo aos locais reconhecidos, tal como no seguinte exemplo:

```
O tempo de viagem entre a <PLACE type=administrative
subtype="city" geoid="GEO_146">cidade de Lisboa</PLACE> e a
<PLACE type=administrative subtype="city" geoid="GEO_238">cidade
do Porto</PLACE> é de duas horas e meia.
```

Além das anotações SGML, há ainda a possibilidade de gerar uma lista com todos os identificadores da ontologia reconhecidos no texto, assim como a frequência de ocorrência correspondente. Esta lista será a preferencialmente usada por outras ferramentas de recuperação de informação que façam uso das referências geográficas.

Para o HAREM foi necessário converter o formato SGML do nosso sistema no formato aceite pelo evento (ver capítulo 16). Para o mesmo exemplo fornecido acima, a anotação HAREM é a seguinte:

```
O tempo de viagem entre a cidade de <LOCAL>Lisboa</LOCAL> e
a cidade do <LOCAL>Porto</LOCAL> é de duas horas e meia.
```

Note-se que os tipos considerados pelo HAREM para a classificação semântica dos locais não se mapeavam directamente na nossa ontologia. Não foi tentado nenhum mapeamento dos nossos tipos de classificação para os considerados pelo HAREM, pelo que apenas participamos num cenário selectivo de identificação de EM de categoria LOCAL, sem qualquer classificação semântica. Outra das adaptações necessárias prende-se com o facto de as directivas para a anotação do HAREM especificarem que não se deve incluir os prefixos em minúsculas (tal como *cidade de*) como parte das anotações HAREM.

8.4 Experiências de avaliação no Mini-HAREM

Tal como descrito anteriormente, a nossa participação no HAREM limitou-se a num cenário selectivo de identificação de EM de categoria LOCAL, visto a colecção dourada e as directivas de anotação não considerarem a classificação semântica das entidades geográficas de acordo com os tipos geográficos usados no nosso sistema, nem muito menos a associação das mesmas com os conceitos geográficos da nossa ontologia.

Participámos na primeira edição do HAREM com uma versão inicial do sistema, mas neste capítulo apenas descrevemos os resultados obtidos na segunda edição do evento (o Mini-HAREM), onde os resultados obtidos com uma versão do sistema significativamente melhorada foram consistentemente melhores.

Para o Mini-HAREM foram geradas duas saídas. Uma delas corresponde à utilização da ontologia portuguesa, tal como descrita na secção 8.2, e a outra corresponde à utilização de uma ontologia conjugando as ontologias portuguesa e mundial. Aquando da primeira edição no HAREM, e por inspecção da colecção dourada usada como recurso de avaliação, verificámos que muitos dos locais anotados correspondiam a países e cidades internacionais importantes. Como o nosso sistema está fortemente dependente da ontologia, pensamos que a ontologia portuguesa seria insuficiente para um bom desempenho do sistema. Nas Tabela 8.5 e 8.6 é feito um resumo dos resultados obtidos por cada uma das saídas. A Tabela 8.6 apresenta ainda os melhores resultados obtidos no evento de acordo com as várias medidas de avaliação consideradas.

Da análise das tabelas ressalta que os resultados obtidos são aceitáveis em termos de precisão e abrangência no reconhecimento simples de EM de categoria LOCAL. Observa-se ainda que a segunda saída, gerada com uma ontologia com nomes de locais estrangeiros, é consistentemente melhor.

	Total	Identificados	Correctos	Parcialmente Correctos	Espúrias	Em Falta
Saída 1	893	686	469 (52,5%)	50 (5,6%)	169 (18,9%)	379 (42,4%)
Saída 2	893	696	486 (54,4%)	49 (5,5%)	163 (18,2%)	363 (40,6%)

Tabela 8.5: Número de EM de categoria *LOCAL* reconhecidos nas saídas para o Mini-HAREM.

	Precisão	Abrangência	Medida F	Erro Combinado	Sobre-geração	Sub-geração
Saída 1	69,78%	53,61%	0,6063	0,5514	0,2464	0,4244
Saída 2	71,17%	55,47%	0,6235	0,5331	0,2342	0,4065
Melhor resultado	92,07%	73,91%	0,7085	0,4398	0	0,2290

Tabela 8.6: Resultados obtidos no Mini-HAREM.

No que diz respeito ao desempenho computacional, e usando um PC Intel Pentium 4 com o sistema operativo Linux e 2 GB de RAM, o CaGE procedeu à anotação do texto a um débito de sensivelmente 50 KB de texto por segundo.

Embora o sistema CaGE tenha ficado ligeiramente aquém dos melhores resultados, importa frisar que a tarefa proposta pelo HAREM é ligeiramente diferente da tarefa de anotação executada pelo CaGE[2]. Em primeiro lugar, as EM na colecção dourada anotadas como `<LOCAL TIPO="CORREIO">` e correspondentes a moradas completas (por exemplo, a morada *Rua 25 de Abril, 77 R/C ESQ - Cruz de Pau - 2840 Seixal*) eram apenas parcialmente reconhecidos pelo nosso sistema (ou seja, este reconhece as entidades *Rua 25 de Abril*, *Cruz de Pau* e *Seixal* separadamente). A tarefa de reconhecimento de moradas completas não foi considerada durante o desenvolvimento do CaGE. Existe muita variabilidade nas expressões deste tipo, levando a um elevado custo computacional para a execução da tarefa.

Em segundo lugar, as EM anotadas na colecção dourada como `<LOCAL TIPO="VIRTUAL">` não eram reconhecidos pelo nosso sistema, visto estes muitas vezes não corresponderem a qualquer localização física. Os locais de tipo virtual podem dizer respeito a endereços electrónicos ou a sítios abstractos com função de alojamento de conteúdos, tais como jornais ou programas de televisão. Uma vez que estes locais não têm interesse no contexto da utilização num motor de busca geográfico, o sistema CaGE nunca foi concebido para reconhecer este tipo de entidades.

Em terceiro lugar, as EM anotadas na colecção dourada como `<LOCAL TIPO="ALARGADO">` também não eram reconhecidos pelo nosso sistema. De acordo com as directivas de anotação, estes locais correspondem a edificações ou pontos de referência tais como bares, hotéis ou centros de congressos. Este caso particular, e visto

[2] **Nota dos editores:** O facto de três subtipos de `LOCAL` contemplados no HAREM não interessarem ao CaGE teria sido razão para que este concorresse ao HAREM apenas no cenário selectivo `LOCAL (ADMINSTRATIVO; GEOGRAFICO)`.

que estes locais têm uma correspondência física, trata-se de uma limitação do nosso sistema, sendo que numa versão futura pretendemos também fazer o reconhecimento e desambiguação destes casos.

Num cenário selectivo correspondente apenas à anotação de entidades do tipo <LOCAL TIPO="ADMINISTRATIVO"> e <LOCAL TIPO="GEOGRAFICO">, a melhor saída do CaGE teria obtido uma precisão e abrangência de 67,1% e 66,5%, respectivamente. É ainda de salientar que o CaGE teria detectado um total de 27 ocorrências apenas parcialmente correctas, apesar de neste cenário não estarem a ser considerados locais do tipo ALARGADO ou CORREIO. Num mesmo cenário, o melhor sistema a concurso no HAREM teria obtido uma precisão e abrangência de 82,8% e 61,6%, respectivamente. Estas diferenças entre os dois sistemas estão relacionadas quer com limitações do sistema CaGE no reconhecimento de algumas entidades, quer com o facto de as directivas de anotação do HAREM diferenciarem os nomes de locais que assumem no texto um papel semântico diferente.

Pelas razões apresentadas, 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 precisas e que melhor reflictam a temática geográfica. Este tema foi já desenvolvido no capítulo 6, por isso não o repetiremos aqui.

8.5 Conclusões

Este capítulo descreveu o sistema CaGE para o reconhecimento, classificação e desambiguação de referências geográficas em textos na língua portuguesa. O mesmo foi desenhado segundo métodos rápidos e simples, por forma lidar de forma robusta com grandes quantidades de documentos. O reconhecimento de referências geográficas é apenas um meio para outras utilizações em ferramentas de recuperação de informação conscientes da geografia. A abordagem aqui descrita é parte de um projecto de âmbito mais largo, visando a construção de um motor de busca geográfico para a *web* portuguesa, baseado na atribuição de âmbitos geográficos aos documentos. Este motor de busca, e consequentemente a abordagem descrita neste capítulo, foi usado no contexto das edições de 2005 e 2006 do GeoCLEF, uma avaliação conjunta semelhante ao TREC dedicada aos sistemas de recuperação de informação geográficos (Gey et al., 2006; Martins et al., 2007).

Para o evento de avaliação HAREM foram feitas algumas adaptações ao sistema, por forma a testar o desempenho do mesmo num cenário selectivo de reconhecimento simples de EM de categoria LOCAL. Neste capítulo apresentamos os resultados obtidos pelo nosso sistema no Mini-HAREM, sendo ainda discutidas as limitações no evento no que diz respeito à avaliação de sistemas focados no tratamento de referências geográficas. Em futuras edições do HAREM, gostaríamos de ver o cenário das referências geográficas tratado

em maior profundidade, nomeadamente através da anotação da colecção dourada de uma forma mais precisa.

A nossa participação no HAREM indicou resultados aceitáveis em termos de precisão e abrangência no reconhecimento de referências geográficas, embora exista ainda lugar para diversos melhoramentos. Estudos adicionais com outras colecções de documentos, maiores e devidamente anotadas com referências geográficas, são quanto a nós necessários para se tirarem mais conclusões.

Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação para a Ciência e Tecnologia, através do projecto com referência POSI/SRI/40193/2001 e da bolsa de doutoramento com referência SFRH/BD/10757/2002.