Capítulo 7

O sistema CaGE no Segundo HAREM

Bruno Martins

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 7, p. 149–158.

Os documentos textuais (por exemplo, artigos publicados em jornais ou páginas na rede) são muitas vezes ricos em informação geográfica. A utilização de técnicas de prospecção de texto para a extracção desta informação, por forma a introduzir capacidades de raciocínio geográfico em sistemas de recuperação de informação, é um problema interessante que tem vindo a ganhar notoriedade (Purves e Jones, 2007).

As técnicas de reconhecimento de entidades mencionadas (EM) encontram na recuperação de informação geograficamente contextualizada uma natural área de aplicação. Contudo, mais do que anotar uma expressão de texto como uma localização, esta área de aplicação requer que as anotações produzidas contenham uma desambiguação completa dos nomes de locais. Por outras palavras, as referências geográficas devem ser associadas explicitamente a coordenadas de latitude e longitude, ou a identificadores de conceitos num almanaque geográfico. Esta informação (ou seja, as coordenadas na superfície terrestre, ou o almanaque geográfico em conjunto com os documentos anotados) pode então ser utilizada noutras tarefas de recuperação de informação, tais como a pesquisa de documentos de acordo com os seus âmbitos geográficos.

O sistema CaGE surgiu no contexto de um trabalho de doutoramento que aborda o problema do reconhecimento e desambiguação de nomes de locais, argumentando que esta é uma tarefa crucial na geo-codificação de documentos textuais (Martins, 2009). O principal objectivo do sistema CaGE é atribuir âmbitos geográficos (ou seja, a área geográfica que o documento descreve como um todo) a documentos textuais, tendo-se que numa versão mais recente do sistema, este objectivo estende-se também aos âmbitos temporais. O sistema foi já usado como módulo independente em vários projectos relacionados com recuperação de informação geograficamente contextualizada, nomeadamente no GREASE (Silva et al., 2006) e no DIGMAP (Borbinha et al., 2007; Martins et al., 2008). O CaGE também participou na primeira edição da avaliação conjunta HAREM (ou seja, no Primeiro HAREM e no Mini-HAREM) com o objectivo de avaliar o seu desempenho em cenários selectivos focados no reconhecimento de EM com categoria LOCAL (Martins e Silva, 2007). Embora os objectivos e os critérios associados ao sistema CaGE se distanciem consideravelmente daqueles que foram considerados no HAREM, um correcto tratamento das referências geográficas depende, em grande medida, da capacidade do sistema em reconhecer as EM da categoria LOCAL, tal como estas foram definidas no contexto das regras semânticas utilizadas na avaliação conjunta.

No Segundo HAREM, o sistema CaGE participou num cenário de avaliação mais abrangente, o qual correspondeu ao reconhecimento de entidades das categorias PESSOA, ORGANIZACAO e TEMPO, e ao reconhecimento e classificação em tipos e subtipos de entidades da categoria LOCAL. Visto que as referências geográficas são muitas vezes ambíguas em relação a entidades de outras categorias (por exemplo, nomes próprios de pessoas que correspondem também a nomes de locais), temos que a consideração destas outras categorias por parte do sistema CaGE pode ajudar naquele que é o seu objectivo principal.

Este capítulo descreve a participação do sistema CaGE na segunda edição da avaliação conjunta HAREM. É feita uma descrição detalhada do sistema e é apresentado o contexto no qual o mesmo foi desenvolvido. São discutidos os resultados obtidos e são ainda listadas as principais melhorias que se pretendem introduzir em desenvolvimentos futuros do sistema.

7.1 Descrição do sistema

O CaGE é um sistema híbrido apoiado por dicionários e regras de desambiguação. As subsecções que se seguem detalham o seu funcionamento, apresentando ainda os dicionários usados pelo sistema.

Estudos anteriores indicam que o problema do reconhecimento de EM pode ser abordado de forma bastante eficaz através do uso de métodos de aprendizagem automática (McCallum e Li, 2003). Contudo, para o caso específico do reconhecimento e desambiguação completa de entidades geográficas, é necessária a utilização de um recurso de informação externo (ou seja, um almanaque geográfico), visto que as referências devem ser inequivocamente associadas a uma representação única para o conceito geográfico que lhes está subjacente (ou seja, coordenadas de latitude e longitude ou identificadores no almanaque geográfico).

7.1.1 Os dicionários e o almanaque usados pelo sistema CaGE

No que diz respeito aos dicionários usados no reconhecimento das EM correspondentes às categorias PESSOA e ORGANIZACAO, assim como das entidades do tipo *"período temporal"* correspondentes à categoria TEMPO, foram usados os seguintes recursos lexicais para a sua construção:

- A base de dados de nomes de entidades denominada REPENTINO, um acrónimo de *REPositório para o reconhecimento de ENTIdades com NOme*[1] (Sarmento et al., 2006).
- Nomes de pessoas listadas na Internet Movie Database (IMDB[2]).
- Nomes de autores listados na base de dados de autoridades bibliográficas PORBASE[3].
- Listas de períodos temporais e de nomes próprios comuns extraídas da Wikipédia.
- Traduções para português dos nomes de períodos temporais definidos no contexto do ECAI Time Period Directory (Petras et al., 2006).
- Dicionários distribuídos com um sistema de reconhecimento de entidades de código aberto (em inglês, *open source*) para a língua inglesa denominado BALIE, acrónimo de *BAseLine Information Extraction* (Nadeau, 2007).

É de salientar que alguns dos recursos lexicais que foram considerados apresentam uma percentagem elevada de nomes em outras línguas que não o português, particularmente nomes na língua inglesa. No entanto, apesar do HAREM apenas utilizar textos em português, nada impede que neles sejam mencionados nomes próprios provenientes de outras línguas (tal como por exemplo nomes próprios de actores de cinema norte-americanos).

No que diz respeito aos dicionários usados no reconhecimento de entidades da categoria LOCAL, foram usados os seguintes recursos lexicais para a sua construção:

[1] http://www.linguateca.pt/REPENTINO/
[2] http://www.imdb.com/interfaces/
[3] http://www.porbase.org/

- As versões portuguesa e multilingue do almanaque geográfico GeoNET-PT01 (Chaves et al., 2005b).
- A base de dados de nomes de locais disponibilizada pelo serviço GeoNames[4].
- A lista de nomes do almanaque geográfico usado no projecto DIGMAP. Uma descrição detalhada deste almanaque é dada por Manguinhas et al. (2008).

O CaGE faz ainda uso de um dicionário de excepções para entidades do tipo local, o qual foi construído manualmente com base em ensaios com o sistema. Este dicionário inclui nomes de entidades que, embora tenham uma conotação geográfica, são maioritariamente usados noutros contextos.

Para a desambiguação completa das EM que correspondem a locais ou a períodos temporais, é ainda utilizado um almanaque mais específico para este tipo de informação, o qual foi desenvolvido no contexto do projecto DIGMAP. A figura 7.1 ilustra os conceitos principais que lhe estão subjacentes.

O almanaque DIGMAP associa os nomes de locais aos conceitos geográficos que lhes estão subjacentes (ou seja, os mesmos locais podem ser associados a vários nomes), definindo ainda a cobertura geo-espacial (por exemplo, coordenadas de latitude e longitude) para cada conceito geográfico, assim como uma hierarquia de relações de inclusão entre os conceitos geográficos que traduz uma organização administrativa da superfície terrestre. O mesmo almanaque define ainda conceitos temporais correspondentes a períodos históricos, listando para cada um deles os nomes que lhes estão associados.

7.1.2 Funcionamento geral do sistema

De um ponto de vista algorítmico, o sistema CaGE assenta numa sequência de operações de processamento com quatro etapas principais:

Etapa 1 : Identificação inicial das entidades mencionadas

a. Os textos são inicialmente atomizados através do algoritmo que é fornecido com as bibliotecas da linguagem Java para o processamento de texto, mais concretamente na classe `java.text.BreakIterator`. Este algoritmo funciona com base numa tabela contextual de pares de caracteres (Gillam, 1999). A separação nos diferentes átomos é determinada com base nos caracteres que ocorrem em ambos os lados de uma dada posição no texto (por exemplo, a tabela para atomização em palavras indica uma separação entre caracteres de pontuação e letras, mas não entre letras consecutivas).

b. Os átomos identificados no texto são percorridos em sequência com um algoritmo do tipo *"janela de análise deslizante"*, por forma a extrair o conjunto de sequências de palavras que ocorrem no texto. Como resultado deste passo, são identificadas todas as sequências de palavras contendo um máximo de seis elementos (ou seja, n-gramas de palavras com $1 \leq n \leq 6$). Uma consequência deste passo é que as entidades mencionadas no texto que tenham um comprimento superior a seis palavras serão ignoradas pelo sistema CaGE.

[4] http://www.geonames.org

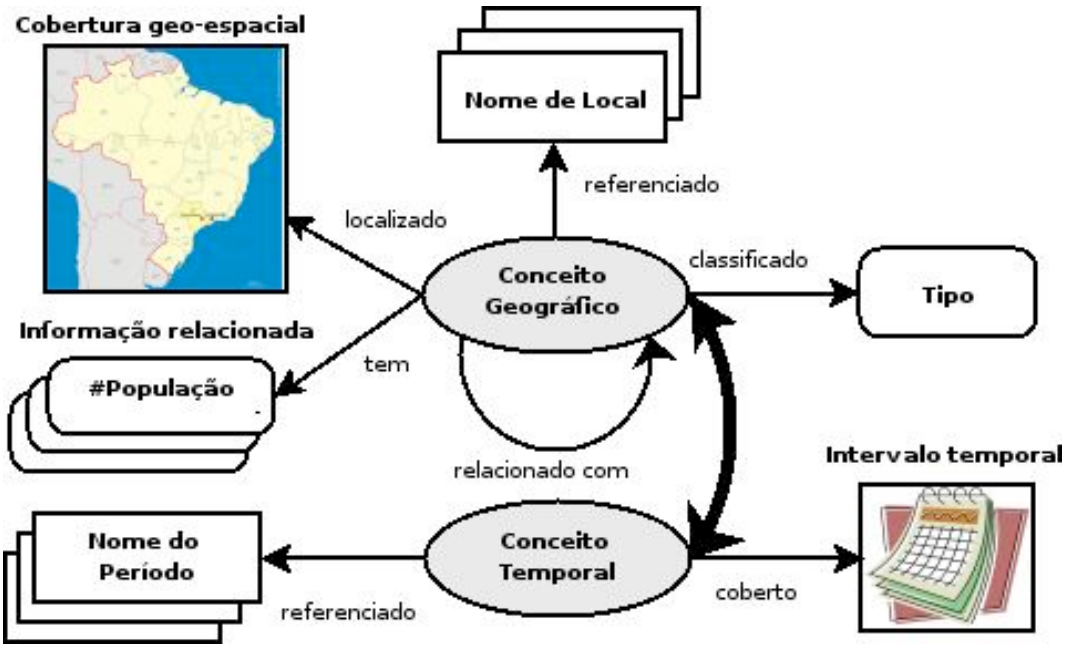

Figura 7.1: Os principais elementos de informação no almanaque DIGMAP

c. As sequências de palavras identificadas no passo anterior são filtradas, por forma a registar as sequências de palavras que começam com letras maiúsculas e que não ocorrem desta forma apenas no início de frases. Este passo implementa uma heurística que diz que as entidades são normalmente mencionadas com a primeira letra em maiúscula.

d. É feita uma pesquisa nos dicionários usados pelo sistema, por forma a mapear as sequências de palavras resultantes do passo anterior com as entidades que lhes correspondem. No caso de existirem mapeamentos diferentes para uma dada sequência e para as suas subsequências, é apenas registado o mapeamento mais geral, ou seja aquele que corresponde à sequência de maior tamanho. Uma vez que para cada sequência de n palavras são também testados mapeamentos para as duas subsequências de $n-1$ palavras, podem, por exemplo, ocorrer dois mapeamentos na expressão textual *Pedrógão Pequeno* (ou seja, um para a sequência *Pedrógão* e outro para a sequência *Pedrógão Pequeno*). No caso do exemplo apresentado, seria apenas considerado o mapeamento mais geral, ou seja *Pedrógão Pequeno*.

e. No caso de sequências de palavras mapeadas com uma entidade da categoria `LOCAL`, é feito um mapeamento adicional usando o dicionário de excepções. No caso de ser registada a ocorrência de um caso de excepção, é removido o mapeamento entre a sequência de palavras e a entidade de categoria `LOCAL`.

f. São usadas expressões regulares para identificar entidades da categoria `TEMPO` que não se encontram definidas nos dicionários (por exemplo, várias formas de

expressar datas de calendário).

Etapa 2 : Classificação das entidades mencionadas e tratamento da ambiguidade

a. No caso das entidades identificadas nos dicionários, para as quais foram registados vários mapeamentos possíveis, são usadas regras de classificação desenvolvidas manualmente para encontrar a categoria e tipo de entidade correspondentes. Estas regras são baseadas na ocorrência de palavras-chave no contexto textual da entidade (ou seja, os dois átomos que ocorrem no texto, antes ou depois da entidade em questão). Por exemplo, uma das regras usadas no CaGE corresponde ao padrão `cidade de [EM] -> LOCAL-CIDADE`, indicando que todas as EM que são precedidas das palavras *cidade de* devem ser classificadas com a categoria `LOCAL` e o tipo `CIDADE`.

b. Para as entidades que permanecem ambíguas após a execução do passo anterior, é feita uma classificação por escolha circular (em inglês, *round-robin classification*) entre as várias categorias e tipos possíveis para a entidade em questão (Fürnkranz, 2002). O argumento por detrás desta estratégia é o de que, escolhendo uma entidade diferente em cada situação ambígua e ir sequencialmente percorrendo o conjunto de atribuições possíveis, minimiza-se o número de erros introduzidos pelo sistema.

Etapa 3 : Desambiguação completa de entidades geográficas e temporais

a. Para cada entidade da categoria `LOCAL` identificada na segunda etapa, é feita uma pesquisa no almanaque DIGMAP, por forma a associar as entidades aos conceitos geográficos subjacentes. Esta pesquisa combina o nome mencionado no texto com o tipo associado à entidade, caso este tenha sido já resolvido com base num mapeamento com um dicionário.

b. No caso de a pesquisa ao almanaque, efectuada no passo anterior, retornar vários conceitos geográficos possíveis, estes são ordenados de acordo com uma heurística do tipo "*um sentido por omisso*" a qual diz que, na maior parte dos casos, uma dada referência geográfica encontra-se associada a um único tipo em concreto (o nome *Lisboa* é mais frequentemente usado como uma referência à cidade capital do país do que a uma pequena vila). A utilização da heurística "*um sentido por omisso*" encontra-se descrita em maior detalhe em Martins et al. (2008).

c. Para as entidades que permanecem ambíguas após o passo anterior (ou seja, as entidades que correspondem a diferentes conceitos no almanaque geográfico), é ainda usada uma heurística do tipo "*referentes relacionados por cada unidade de discurso*" por forma a melhorar a ordenação dos conceitos geográficos correspondentes à entidade. Caso exista uma relação hierárquica entre um dos conceitos possíveis para a entidade em questão, e os outros conceitos mencionados no documento, então é dada uma maior importância a esse conceito aquando da sua ordenação. Mais uma vez, o mecanismo subjacente à heurística "*referentes relacionados por cada unidade de discurso*" encontra-se descrito em maior detalhe em Martins et al. (2008).

d. Para as entidades correspondentes a períodos temporais identificadas na primeira etapa, é feita uma pesquisa no almanaque DIGMAP por forma a associar a entidade ao conceito temporal que lhe está subjacente.

Etapa 4 : Atribuição de âmbitos geográficos e temporais aos documentos

a. É atribuído um âmbito geográfico à totalidade do documento com base na combinação de todas as referências geográficas identificadas no texto. Esta atribuição é feita com base no algoritmo originalmente proposto por Amitay et al. (2004), o qual assenta na utilização de uma hierarquia de relações de inclusão entre os conceitos geográficos reconhecidos no texto. O almanaque DIGMAP é usado como fonte de dados para estas relações.

b. É atribuído um âmbito temporal ao documento com base no intervalo mínimo que cobre todas as referências temporais identificadas no texto.

7.1.3 Aplicações práticas do sistema CaGE

Como resultado das quatro etapas de processamento apresentadas atrás, tem-se que o sistema CaGE permite não só reconhecer e classificar entidades mencionadas em textos, como também desambiguar as entidades correspondentes a referências geográficas ou temporais. Finalmente, o sistema CaGE suporta ainda a atribuição de âmbitos geográficos e temporais aos documentos como um todo, combinando a diferente informação extraída do texto. Mais detalhes sobre os aspectos relacionados com a desambiguação completa de referências geográficas e temporais, assim como sobre a atribuição de âmbitos, podem ser consultados em Martins et al. (2008).

Um serviço na rede (em inglês, *Web service*) com base no sistema CaGE encontra-se disponível para utilização online, mais concretamente no URL `http://geoparser.digmap.eu`. Este serviço oferece uma interface XML através da qual se pode invocar o reconhecimento e desambiguação das EM num dado documento textual. A interface XML segue, em linhas gerais, uma proposta do Open Geospatial Consortium para a implementação de serviços na rede para o geo-processamento de recursos textuais (Lansing, 2001). Uma vez que a saída do serviço é um documento XML bem formado, torna-se relativamente simples desenvolver outros serviços que explorem a informação extraída dos documentos. A figura 7.2 mostra o ecrã principal de uma aplicação na rede que utiliza o serviço do CaGE por forma a reconhecer e desambiguar referências geográficas em canais de notícias RSS, possibilitando a exploração das notícias sobre um mapa dinâmico.

O almanaque geográfico desenvolvido no contexto do projecto DIGMAP, o qual é usado na desambiguação completa de entidades da categoria `LOCAL`, encontra-se também acessível através de um serviço na rede, mais concretamente através do URL `http://gaz.digmap.eu`. Este serviço permite a realização de consultas sobre o almanaque, as quais podem combinar aspectos tais como o nome dos locais, os seus tipos, e a sua localização sobre a superfície terrestre. A interface deste serviço segue o formato XML proposto no contexto do Alexandria Digital Library Gazetteer (Hill et al., 1999).

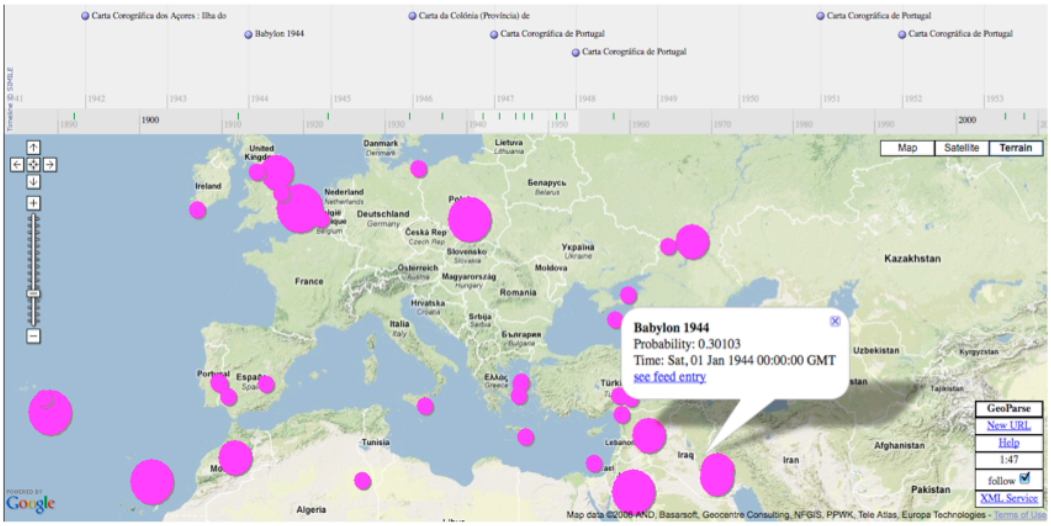

Figura 7.2: Um serviço na rede para exploração de feeds RSS, construído com base no CaGE

7.2 Experiências no HAREM e análise dos resultados

Tal como mencionado anteriormente, o sistema CaGE participou no Segundo HAREM apenas num cenário selectivo, o qual corresponde ao reconhecimento de entidades da categoria PESSOA, ORGANIZACAO e TEMPO, e ao reconhecimento e classificação em tipos e subtipos de entidades da categoria LOCAL (exceptuando-se o subtipo VIRTUAL). Para as entidades da categoria TEMPO, apenas se trataram os casos de datas absolutas (*15 de Abril de 1979*) e nomes de períodos temporais (por exemplo, *idade média*).

O HAREM aborda apenas o reconhecimento e classificação das EM, não considerando o problema da desambiguação completa das entidades correspondentes a locais ou a períodos temporais. A etapa 4 do algoritmo usado pelo CaGE, assim como os passos b) e c) da etapa número 3, não foram portanto utilizados no contexto da participação no HAREM.

Foram enviadas quatro corridas ao Segundo HAREM, as quais correspondem aos cenários experimentais que se encontram descritos abaixo:

1. Utilização dos vários dicionários, assim como do almanaque DIGMAP, para desambiguação de entidades da categoria LOCAL. A classificação em tipos e subtipos das entidades da categoria LOCAL foi feita com base no almanaque DIGMAP (ou seja, as entidades do tipo LOCAL tinham de estar forçosamente definidas no almanaque DIGMAP).

2. Utilização dos vários dicionários, exceptuando-se o dicionário de casos de excepção.

3. Utilização dos dicionários contendo nomes de locais e períodos temporais, excluindo-se os restantes dicionários. Nesta corrida, eram apenas reconhecidas as entidades das categorias TEMPO e LOCAL.

Tabela 7.1: Resultados obtidos na classificação de entidades mencionadas no cenário selectivo 2

Corrida	Posição	Precisão	Abrangência	Medida F
4	14	0,4264	0,4070	0,4164
1	16	0,4277	0,4025	0,4148
2	17	0,4226	0,4059	0,4141
3	20	0,3883	0,3500	0,3682
Melhor corrida	1	0,7347	0,5893	0,6325

Tabela 7.2: Resultados obtidos na identificação de entidades mencionadas no cenário selectivo 2

Corrida	Posição	Precisão	Abrangência	Medida F	TotalEMCD	TotalEMSis
4	16	0,4615	0,4553	0,4584	5538,3	5463,5
1	17	0,4643	0,4520	0,4581	5538,3	5391,5
2	18	0,4576	0,4547	0,4562	5538,3	5503,5
3	20	0,4225	0,3929	0,4072	5538,3	5151,2
Melhor corrida	1	0,8561	0,7127	0,6813		

4. Utilização de todos os dicionários sem quaisquer restrições adicionais (ou seja, as entidades do tipo LOCAL podiam estar definidas no almanaque DIGMAP ou num dos restantes dicionários).

As tabelas 1 e 2 apresentam, respectivamente, os resultados obtidos na classificação e identificação de entidades para o cenário selectivo 2 do Segundo HAREM, o qual considerava várias categorias de entidades (ou seja, LOCAL, TEMPO, ORGANIZACAO e PESSOA) assim como a classificação em tipos para as entidades da categoria LOCAL. Na tabela 2, as colunas TotalEMCM e TotalEMSis representam, respectivamente, o número total de entidades existente na colecção dourada e o número total de entidades retornado pelo sistema.

A corrida número 4 foi a que obteve melhores resultados, correspondendo a uma diferença de cerca de 0,2 em termos da medida F para com a melhor corrida neste mesmo cenário.

As tabelas 3 e 4 apresentam, respectivamente, os resultados obtidos na classificação e identificação de entidades para o cenário selectivo 5 do Segundo HAREM, o qual considera apenas entidades da categoria LOCAL, exceptuando-se ainda as entidades do tipo VIRTUAL. Estes resultados são ligeiramente superiores aos obtidos no cenário selectivo 2. Mais uma vez, a corrida número 4 foi a que obteve o melhor resultado, o qual corresponde a uma diferença de aproximadamente 0,1 em termos da medida F para com a melhor corrida neste mesmo cenário.

Comparando com os resultados obtidos pelo sistema CaGE na anterior edição do Mini-HAREM, em condições semelhantes àquelas que são usadas no cenário selectivo 5 do Segundo HAREM, temos que os resultados obtidos pela corrida número 4 são ligeiramente inferiores (ou seja, uma diferença de aproximadamente 0,1 em termos da medida F).

Em suma, o sistema CaGE obteve resultados modestos na sua participação no Segundo HAREM. Embora os dicionários utilizados pelo sistema apresentem uma cobertura adequada (por exemplo, para o caso das entidades da categoria LOCAL, tem-se que os dicionários utilizados pelo CaGE listam cerca de dois milhões de nomes diferentes), as regras e heurísticas utilizadas pelo sistema carecem ainda de alguma optimização.

Tabela 7.3: Resultados obtidos na classificação de entidades mencionadas no cenário selectivo 5.

Corrida	Posição	Precisão	Abrangência	Medida F
4	11	0,5267	0,5844	0,5540
2	12	0,5196	0,5851	0,5504
1	13	0,5147	0,5802	0,5455
3	14	0,5178	0,5754	0,5451
Melhor corrida	1	0,7080	0,70236	0,6246

Tabela 7.4: Resultados obtidos na identificação de entidades mencionadas no cenário selectivo 5.

Corrida	Posição	Precisão	Abrangência	Medida F	TotalEMCD	TotalEMSis
4	11	0,5198	0,6788	0,5888	1418	1851,5
2	12	0,5091	0,6802	0,5823	1418	1894,5
1	13	0,5049	0,6781	0,5788	1418	1904,5
3	14	0,5084	0,6689	0,5777	1418	1865,5
Melhor corrida	1	0,7186	0,7856	0,6572		

7.3 Conclusões

Neste capítulo foi descrito o sistema CaGE para reconhecimento de entidades geográficas, assim como a sua adaptação para a participação no Segundo HAREM e os respectivos resultados obtidos nesta avaliação conjunta. Muito embora se tenham obtido resultados relativamente modestos, a participação no HAREM foi bastante útil, tendo permitido já detectar e corrigir diversas falhas existentes no sistema.

Tal como na primeira edição do HAREM, existe um desfasamento considerável entre os critérios e os objectivos estabelecidos para o sistema CaGE e aqueles que são contemplados pela avaliação conjunta. Embora o sistema CaGE tente fazer reconhecimento de entidades no contexto em que elas são mencionadas, este não tem como objectivo o reconhecimento da função das entidades no texto (por exemplo, os casos de metonímia, tais como no exemplo *Portugal pronunciou-se...*, são sempre marcados como entidades de uma mesma categoria, neste caso `LOCAL` e não `PESSOA`). Este desfasamento explica, em parte, os modestos valores alcançados pelo sistema em termos das diversas métricas de avaliação consideradas.

Como principais desafios de trabalho futuro, há a considerar a melhoria das regras de classificação de entidades, assim como um tratamento mais profundo das referências temporais, segundo as directivas genéricas da pista do TEMPO tal como definidas para esta edição do HAREM. Este último ponto é particularmente interessante para o caso de aplicações de recuperação de informação, uma vez que a extracção da dimensão temporal dos documentos, assim como a sua combinação com a dimensão geográfica, pode suportar mecanismos de recuperação de informação mais adequados a alguns domínios.