Capítulo 13

Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM

Marcirio Silveira Chaves

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 13, p. 231–245.

A maior quantidade de conhecimento existente atualmente está disponível em textos na rede. De acordo com Wilks (2008), 85% da informação disponível para ciência, empresas e aquela encontrada de modo informal na rede estão no formato não estruturado (texto, a maior parte). Contudo, de modo a tornar-se verdadeiramente útil para sistemas que fazem algum tipo de processamento inteligente, esse conhecimento precisa ser identificado e classificado corretamente.

Para tentar aproveitar melhor esse conhecimento, é necessário reconhecer inicialmente as entidades mencionadas presentes nos documentos. A tarefa de reconhecimento de entidades mencionadas (REM) em textos em português tem ganho mais atenção desde 2005, quando ocorreu a primeira avaliação desses sistemas no Primeiro HAREM (Santos e Cardoso, 2007a). Os resultados alcançados pelos sistemas participantes evidenciaram a necessidade de mais pesquisa na área, motivando dois eventos subsequentes: o Mini-HAREM (integrado no Primeiro HAREM) e o Segundo HAREM. Conforme descrito no capítulo 4, o Segundo HAREM também desafiou os sistemas a reconhecerem as relações existentes entre entidades mencionadas.

Nesse capítulo eu concentro a atenção no reconhecimento de entidades mencionadas da categoria local e de suas relações. Para isso, eu desenvolvi o SEI-Geo, um Sistema de Extração, Anotação e Integração de Conhecimento Geográfico baseado essencialmente no uso de padrões e de geo-ontologias. O SEI-Geo está inserido no contexto de minha tese de doutoramento e tem como um de seus objetivos expandir o conhecimento existente em bases de conhecimento geográfico com informação textual. Nesse capítulo é descrito apenas o módulo de extração e anotação de informação geográfica, o qual foi utilizado no processo de anotação de locais e relações em textos.

A participação do SEI-Geo no Segundo HAREM é motivada pela necessidade de se mensurar a qualidade da parte do sistema SEI-Geo que trata do reconhecimento de locais e suas relações em textos.

13.1 Trabalhos relacionados

O uso de padrões tem sido aplicado para extração de informação de textos em diversos trabalhos (Agichtein e Gravano, 2000; Etzioni et al., 2005; Cafarella et al., 2005; McDowell e Cafarella, 2008). Todos esses trabalhos processam textos em língua inglesa. O interesse sobre a informação geográfica presente em textos em português tem ganho atenção apenas nos últimos anos. Especificamente no tratamento de informação geográfica em texto em língua portuguesa, dois trabalhos têm sido reportados na literatura como mais relevantes (Vasconcelos Borges, 2006; Martins et al., 2007b). O primeiro eu descrevo brevemente a seguir e o segundo, o sistema CaGE (*Capturing Geographic Entities*), participou em todas as edições do HAREM, foi descrito em Martins et al. (2007b) e está presente nesse livro no capítulo 7.

Delboni (2005) e Vasconcelos Borges (2006) foram pioneiros ao processar textos na variante brasileira da língua portuguesa. Vasconcelos Borges propôs uma ontologia de lugar (OnLocus) que possui conceitos geográficos norteados pela divisão administrativa do Brasil. No trabalho dela, OnLocus foi mais explorada na parte de endereços postais e telefones.

Numa amostra de 75.413 páginas da rede brasileira, Vasconcelos Borges encontrou 57% delas com presença de endereços que foram detectados de acordo com

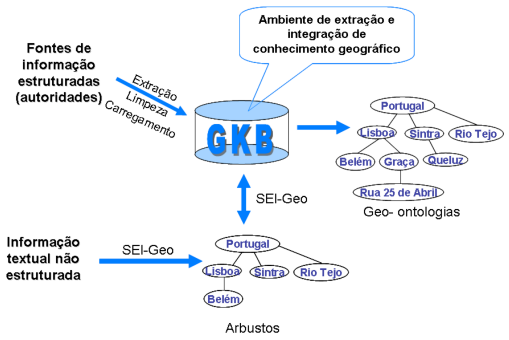

Figura 13.1: Arquitetura global do sistema de gestão de conhecimento geográfico.

um conjunto de padrões pré-definidos. Os seis principais tipos de padrões utilizados são: `Telefone`, `EndereçoBásico+CidadeEstado+CEP`, `EndereçoBásico+Telefone`, `EndereçoBásico+CidadeEstado`, `EndereçoBásico+CEP` e `CEP`. Esses padrões foram aplicados à coleção WBR05 (Modesto et al., 2005), extraindo 2.137.601 endereços de 603.798 páginas, o que representa 14,77% do total de páginas dessa coleção.

Delboni também apresenta um conjunto de expressões de posicionamento desenvolvidas para detectar nomes geográficos em textos em português. Essas expressões são baseadas em quatro tipos de relações espaciais: difusas[1] (por exemplo, *perto*, *depois* e *acima*), direcionais (por exemplo, *em frente*, *ao lado*, *atrás*), métricas (por exemplo, *quilômetros*, *minutos*, *quadras*) e topológicas (por exemplo, *dentro de*, *no coração de*, *na praça de alimentação*). Os experimentos realizados por Delboni indicam que as relações direcionais e, principalmente, as métricas são predominantemente utilizadas no contexto de uma expressão de posicionamento (informação geográfica), enquanto os demais tipos de relações são empregados em outros contextos.

13.2 O SEI-Geo

O SEI-Geo é um sistema que está integrado numa arquitetura global do sistema de gestão de conhecimento geográfico desenvolvido na minha tese de doutoramento, a GKB – *Geographic Knowledge Base* (Chaves et al., 2005b), e que está representada na figura 13.1.

A GKB é um ambiente de extração e integração de conhecimento geográfico que contém informações provenientes de fontes de dados administrativas semi-estruturadas de autoridades junto com um conjunto de regras para integração de informação. A expansão do conhecimento contido na GKB ocorre com informação proveniente de textos. Esses textos são a entrada de informação para o SEI-Geo, que é o responsável por gerar uma representação estruturada (em forma de arbustos) do conhecimento geográfico extraído e integrá-lo no repositório da GKB. Programas simples para geração de ontologias exportam o conhecimento armazenado nesse repositório.

Esta seção descreve o SEI-Geo utilizado no Segundo HAREM. O sistema é composto por dois módulos principais: extrator e anotador de informação geográfica e integrador de

[1] Em inglês, *fuzzy*.

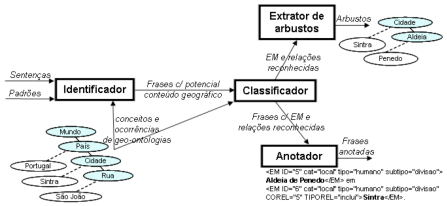

Figura 13.2: Arquitetura do módulo de extração e anotação de informação geográfica do SEI-Geo.

conhecimento geográfico. O primeiro tem como objetivo identificar, classificar, extrair arbustos e anotar o conhecimento geográfico disponível em textos representando-o de forma estruturada. A figura 13.2 apresenta a sua arquitetura. A seguir são descritas as funções de cada sub-módulo.

`Identificador`: recebe como entrada uma coleção de textos previamente segmentados em sentenças, mais um conjunto de padrões e de conceitos e ocorrências de geo-ontologias. As sentenças com potencial conteúdo geográfico são a entrada do módulo `Classificador`.

`Classificador`: recebe as sentenças e consulta as geo-ontologias para fazer a desambiguação e identificar relações semânticas.

`Extrator de arbustos`: recebe os locais reconhecidos e constrói os arbustos. Um arbusto é composto por pelo menos duas entidades mencionadas e uma relação. Não há número máximo de entidades mencionadas e relações pré-definido. Um exemplo de arbusto é `<Aldeia de Penedo,parte da,cidade de Sintra>`. Os arbustos são formalizados no formato de triplas *Resource Description Framework* (RDF), o qual está disponível em `http://www.w3.org/RDF`.

`Anotador`: recebe a sentença com locais e relações reconhecidos e faz a anotação no formato solicitado por qualquer aplicação. No caso do HAREM, anota a sentença de acordo com as diretivas da avaliação (Santos et al., 2008c).

Os padrões usados como entrada no algoritmo são descritos a seguir juntamente com o procedimento executado pelo algoritmo. Exceto os padrões do tipo Hearst, todos os demais são imediatamente sucedidos por preposição antes do nome de local. A lista completa das preposições utilizadas é como segue: *a*, *de*, *da*, *das*, *do*, *dos*, *entre*, *na*, *nas*, *no*, *nos*, *em*, *à*, *para*, *pra*, *ao*.

Conceitos geográficos: Conceitos de uma geo-ontologia existente mais conceitos complementaresàqueles inseridos no SEI-Geo, mas ausentes nas geo-ontologias. O algoritmo extrai todos os conceitos definidos na geo-ontologia que estão presentes na sentença. Sempre que o algoritmo encontra um conceito, ele verifica se esse conceito é sucedido por um nome de local. Em caso positivo, anota o nome como um local.

Padrões do tipo Hearst traduzidos para o português e estendidos: são aqueles definidos em Hearst (1992) acrescentados de algumas variantes adaptadas ao português (por exemplo, *é o distrito*, *é um concelho* e *é uma das cidades*). O algoritmo utiliza os

padrões do tipo Hearst do tipo `[Nome de local] é um (d[eao]s?)? [Conceito geográfico]` e `[Conceito geográfico] tal(is) como [Nome de local]`. Para cada padrão encontrado na frase, o algoritmo anota os locais presentes. Um `Nome de local` é todo o nome próprio que refere-se a um local geográfico.

Relações métricas, direcionais, difusas e de orientação: relações métricas descrevem proximidade usando unidades de medida, relações direcionais indicam posicionamento em relação a determinado local, construção ou objeto, entre outros, relações difusas descrevem proximidade através da utilização de termos qualitativos e imprecisos e relações de orientação, que foram inicialmente definidas como direcionais no trabalho de Güting (1994), são expressos através de pontos cardeais. A seguir é apresentada a lista completa dos termos utilizados desses padrões:

Métrico: *distante(s)*, *distância*, *km(s)*, *kilómetro(s)*, *quilómetro(s)*, *minuto(s)*, *minuto(s)*, *metro(s)*

Direcional: *ao lado*, *atrás*, *em frente*, e *defronte*

Difuso: *antes*, *depois*, *acima*, *abaixo*, *próxima*, *próximo*, *perto* e *proximidades*

Orientação: *norte*, *sul*, *leste*, *oeste*, *nordeste*, *noroeste*, *sudeste*, *sudoeste*

Todos esses padrões são descritos em (Delboni, 2005) como aqueles que geram melhores resultados quando uma pessoa deseja expressar posicionamento.

Substantivos: *água(s)*, *afogada(s)*, *afogado(s)*, *beira(s)*, *cabo(s)*, *capital(ais)*, *eleição(ões)*, *favela(s)*, *herdade(s)*, *guerra(s)*, *litoral(ais)*, *margem(ns)*, *natural(ais)*, *penitenciária(s)*, *periferia(s)*, *prefeito(s)*, *procedente(s)*, ex-*prefeito(s)* e *praia(s)*. Nomes próprios que ocorrem após esses substantivos são anotados como entidades candidatas (ou seja, topônimos) a locais. Alguns desses substantivos (por exemplo, *praia* e *cabo*) são candidatos a conceitos para expandir geo-ontologias.

Advérbios *aqui*, *cá*, *lá* e *longe*. Nomes próprios que ocorrem após esses advérbios são anotados como entidades candidatas a locais.

Verbos: *chegar*, *falecer*, *localizar*, *morar*, *morrer*, *mudar*, *nascer*, *ser*, *situar*, *sediar*, *realizar*, *viver*, *voltar*, *ir*, *vir*. Esse tipo de padrão inclui variações de tempo, gênero e número. Nomes próprios que ocorrem após esses verbos são anotados como entidades candidatas a locais.

Nomes de Entidades: Ocorrências das geo-ontologias.

13.2.1 Geo-ontologias utilizadas pelo SEI-Geo

Ontologias podem desempenhar um papel importante na tarefa de REM, especificamente para locais, assim como dicionários e almanaques. Para Malouf (2002) sua utilização não auxilia a melhoria dos resultados, enquanto Mikheev et al. (1999) encontraram bons resultados utilizando almanaques. Carreras et al. (2003) apresentaram resultados melhores com o uso de almanaques. Mikheev et al. também comprovaram que a utilização de almanaques é necessária para identificar nomes de locais. Em português, Martins et al. (2007b) obtiveram resultados satisfatórios com o uso de geo-ontologias para reconhecer locais.

Tabela 13.1: Estatística sobre as geo-ontologias utilizadas pelo SEI-Geo.

Estatística	Geo-Net-PT 10	WGO
# de entidades	4.651	13.124
# de nomes distintos	3.749	10.442
# de relações	6.304	24.712
# de relações parte-de	4.956	13.341
# de relações de adjacência	1.348	11.371

As geo-ontologias na abordagem do SEI-Geo fornecem listas de nomes e conceitos. Uma das vantagens das geo-ontologias é que elas permitem que se explore as relações existentes entre locais reconhecidos em textos com base nas relações nelas definidas.

A tabela 13.1 apresenta as estatísticas das geo-ontologias utilizadas pelo SEI-Geo. Os valores da Geo-Net-PT 10 incluem os dez principais conceitos da geo-ontologia, até o nível de freguesia. A geo-ontologia completa de Portugal (Geo-Net-PT) contém mais de 400.000 entidades, é um recurso público que foi desenvolvido no Pólo XLDB da Linguateca em colaboração com o projeto GREASE e está disponível em http://xldb.fc.ul.pt/geonetpt.

Além da Geo-Net-PT, o SEI-Geo utiliza a *World Geographic Ontology* (WGO) (Chaves et al., 2005a; Martins et al., 2007b). Essa geo-ontologia contém nomes, conceitos e relações sobre as principais divisões administrativas do mundo desde países e territórios até cidades com mais de 100.000 habitantes, além de entidades geográficas no domínio físico, tais como oceanos, mares e montanhas. Ambas as geo-ontologias foram utilizadas para suportar a participação do sistema de recuperação de informação geográfica (RIG) da Universidade de Lisboa nas quatro edições do GeoCLEF[2], de 2005 a 2008 (Cardoso et al., 2006; Martins et al., 2007a; Cardoso et al., 2008a,c). Esse sistema de RIG obteve o primeiro lugar na avaliação em 2006 nas tarefas monolíngue inglês e português.

As fontes de informação da Geo-Net-PT são provenientes de autoridades administrativas de Portugal (por exemplo, Instituto Nacional de Estatística (INE), Correios, Telégrafos e Telefones (CTT) e Associação Nacional de Municípios Portugueses (ANMP)), enquanto a WGO é formada na sua maioria por dados do *World Gazetteer*, da Wikipedia e do Instituto Geográfico Português.

13.2.2 Algoritmos de identificação e classificação de locais

O algoritmo 13.1 formaliza a fase de identificação de locais no módulo extrator e anotador de informação geográfica do SEI-Geo. Antes de descrever os algoritmos é necessário definir os seguintes termos:

Entidade candidata (EC): é um topônimo, um nome próprio (composto por pelo menos uma palavra). Exemplos de entidades candidatas incluem *Grécia*, *Brasília* e *concelho de Braga*.

Entidade Geográfica (EG): é um objeto com significado no domínio do discurso (correspondente à *feature* na ISO 19109 (ISO19109, 2006)). No domínio geográfico, a *província do Algarve*, o *concelho de Évora* e a *freguesia de Santa Isabel* são exemplos de tais

[2] O GeoCLEF foi um fórum internacional de avaliação de sistemas de RIG. Mais detalhes em http://www.uni-hildesheim.de/geoclef.

entidades numa ontologia. Essas entidades geográficas devem ter uma referência numa ontologia, e essa referência fornece o significado no domínio geográfico. Formalmente, uma entidade geográfica é uma `EC` que refere apenas uma referência na ontologia (por exemplo, `<[<concelho,Évora>],[GEO_346]>)`.

Além desses termos, a sintaxe de `w[+1]` significa a palavra sucessora daquela que está sendo comparada no ciclo (*for*). Por exemplo, no seguinte excerto de uma sentença *... perto de Aveiro ...*, se *perto* é o padrão sendo comparado, `w[+1]` é igual a *de*.

Algoritmo 13.1: Algoritmo para identificação de locais implementado no SEI-Geo.

```
1: WGO_adm = {ocorrências do domínio administrativo da WGO}
2: WGO_fis = {ocorrências do domínio físico da WGO}
3: WGO = WGO_adm ⋃ WGO_fis
4: GN = {ocorrências da Geo-Net-PT}
5: P = {Adjetivo ∪ Adverbio ∪ Conceito geográfico ∪ Fuzzy ∪ Hearst ∪ Metrico ∪ Orientacao ∪ Substantivo
   ∪ Verbo}
6: S = {frases do texto}
7: Prep = {a,de,da,das,do,dos,entre,na,nas,no,nos,em,à,para,pra,ao}
8: for all s ∈ S do
9:     for all w ∈ s do
10:        if w ∈ P then
11:            EC = identificaEC(w[+1],s)
12:            if EC != null then
13:                Algoritmo 13.2 (EC)
14:            end if
15:        else if w ∈ {WGO ⋃ GN} then
16:            EC = identificaEC(w,s)
17:            if EC != null then
18:                Algoritmo 13.2 (EC)
19:            end if
20:        end if
21:    end for
22: end for
23:
24: sub identificaEC(w,s) {
25: for all w ∈ s do
26:    if w ∈ {Prep ∪ ^([0-9]|[A-Z]) ∪ (length(w) ≥ 2)} then
27:        EC += w
28:    end if
29:    if EC[0] ∈ Prep then
30:        EC = EC[1,length(EC)]
31:    end if
32:    if EC[-1] ∈ Prep then
33:        EC = EC[0,length(EC)-1]
34:    end if
35: end for
36: return EC
37: }
```

A fase de identificação de locais recebe como entrada ocorrências de geo-ontologias, padrões que incluem termos frequentemente utilizados ao redor de nomes de locais em textos e preposições que ocorrem em nomes de locais. Toda vez que um padrão é encontrado numa frase, o algoritmo invoca a função `identificaEC` que identifica e retorna uma EC ou `null`, caso não seja um nome candidato a local. Essa função encontra os delimitadores da EC, ou seja, o início e o fim da mesma através de preposições e termos cuja primeira letra é maiúscula e seu comprimento é maior ou igual a dois. Após encontrar uma EC o algoritmo invoca a função de classificação de locais, descrita no algoritmo 13.2. Caso a palavra que está sendo comparada com os padrões não seja um padrão e sim um nome que está presente em geo-ontologias, o algoritmo verifica se a próxima palavra da sentença faz parte do nome. Se fizer, invoca a função `identificaEC`. Senão, assume a palavra como nome de local e invoca a função de classificação de locais, descrita no algoritmo 13.2.

Algoritmo 13.2: Algoritmo para classificação de locais implementado no SEI-Geo.

1: EC = {nome extraído do texto = entidade candidata}
2: WGO_{adm} = {ocorrências do domínio administrativo da WGO}
3: WGO_{fis} = {ocorrências do domínio físico da WGO}
4: GN = {ocorrências da Geo-Net-PT}
5: **if** EC $\in$ WGO_{adm} **then**
6: EG = {id do pai mais acima na hierarquia da WGO_{adm}}
7: **else if** EC $\in$ GN **then**
8: EG = {id do pai mais acima na hierarquia da GN}
9: **else if** EC $\in$ WGO_{fis} **then**
10: EG = {id do pai mais acima na hierarquia da WGO_{fis}}
11: **else**
12: EG = {id tipo Humano}
13: **end if**
14: AnotaEG(EG);

A partir de uma EC o algoritmo consulta a WGO e, caso encontre o nome nessa geo-ontologia, verifica se esse nome está no domínio administrativo da WGO. Se encontra, atribui o identificador da entidade geográfica com conceito mais alto na hierarquia da WGO. Por exemplo, se encontra a EC *França*, atribui o conceito *país* e não *cidade* ou *vila*. Se não encontra, tenta atribuir o identificador da entidade geográfica do domínio físico da WGO. Caso o nome não esteja na WGO, o algoritmo procura na Geo-Net-PT. Se encontra, atribui o identificador da entidade geográfica com conceito mais alto na hierarquia da Geo-Net-PT, critério adotado para desambiguação também. A função `AnotaEG` anota as entidades geográficas conforme o domínio das geo-ontologias nos quais elas foram reconhecidas. Essa função é um simples conversor dos tipos das geo-ontologias para os tipos definidos no Segundo HAREM. Caso o nome não esteja em nenhuma das geo-ontologias, ele é anotado como um local com o tipo humano.

Nos casos de ambiguidade entre nomes do domínio administrativo e físico, o algoritmo 13.2 usa uma heurística que prioriza o domínio administrativo. Por exemplo, se um mesmo nome se refere a uma cidade e a um lago e não possui nenhum discriminador (conceito geográfico) no texto, o algoritmo assume que o nome se refere à cidade. Essa abordagem também foi usada em Volz et al. (2007).

Uma das vantagens de utilizar geo-ontologias na fase de classificação de locais é o fato de se reconhecer um local com um nível mais específico de granularidade. Ao invés de classificar o local como administrativo ou físico, é possível classificá-lo como uma freguesia, localidade ou lago, por exemplo.

13.2.3 Reconhecimento de relações semânticas entre EM – ReRelEM

Uma das funcionalidades do SEI-Geo é a extração de relações semânticas entre entidades geográficas. Uma das pistas do Segundo HAREM propõe o desafio de reconhecer relações entre EM. A participação do SEI-Geo nessa tarefa restringiu-se ao reconhecimento de relações entre entidades pertencentes à categoria local, que é um dos problemas tratados pelo SEI-Geo.

A abordagem utilizada pelo SEI-Geo na tarefa de reconhecimento de relações foi baseada em geo-ontologias. Todos os locais encontrados num documento são projetados sobre ontologias com o objetivo de encontrar relações de inclusão (`inclui`/`incluido`) entre eles. Caso encontre alguma relação, o SEI-Geo anota a mesma no documento. Essa abordagem permite testar até que ponto um algoritmo de reconhecimento de relações geográficas consegue ser preciso e abrangente só com o uso de ontologias.

Um fator importante a destacar é o âmbito no qual uma relação pode ocorrer. O SEI-Geo foi desenvolvido originalmente para relacionar locais dentro de uma mesma sentença. Entretanto, de acordo com as diretivas da tarefa de ReRelEM, relações devem ser identificadas ao nível do documento. As corridas contemplando essas duas variantes são descritas a seguir.

13.3 Descrição das corridas

Apesar de o Segundo HAREM promover o reconhecimento de diversas categorias (por exemplo, `PESSOA`, `ORGANIZACAO` e `TEMPO`), o SEI-Geo participou apenas no reconhecimento da categoria `LOCAL`, dos tipos `HUMANO` e `FISICO` e de todos os sub-tipos desses tipos. Dentro do mesmo evento também foi promovida a tarefa de reconhecimento de relações semânticas entre entidades mencionadas (ReRelEM). O SEI-Geo participou nessa tarefa anotando relações de inclusão entre locais.

O SEI-Geo participou no Segundo HAREM com quatro corridas. As variações realizadas nas quatro corridas do SEI-Geo, são as geo-ontologias de entrada do sistema e o âmbito das relações a serem reconhecidos.

Corrida 1 (Geo-Net-PT): utilizou somente a Geo-Net-PT até o nível de localidade, ou seja, conceitos e entidades geográficas acima do conceito de localidade inclusive.

Corrida 2 (WGO - Relação com âmbito no documento): utilizou apenas a WGO com nomes geográficos de todo o mundo, incluindo nomes de países, cidades capital, principais regiões administrativas e cidades com mais de 100.000 habitantes. Na tarefa de ReRelEM essa corrida anota relações existentes entre locais ao longo de todo o documento.

Corrida 3 (Duas ontologias - Relação com âmbito na sentença): o âmbito das relações foi restrito ao nível de sentença, conforme o SEI-Geo foi projetado originalmente. Na

tarefa de ReRelEM a corrida 3 anota relações existentes somente para locais que estejam na mesma sentença.

Corrida 4 (Duas ontologias - Relação com âmbito no documento): o âmbito das relações foi o documento completo, o que caracteriza a proposta original da tarefa ReRelEM.

As corridas 3 e 4 utilizaram as ontologias WGO e Geo-Net-PT, essa mutilada no nível de localidades, ou seja, os nomes de localidades não foram incluídos nessas corridas.

Sempre que o algoritmo encontra um mesmo nome em ambas, a opção é feita pela geo-ontologia WGO, uma vez que a mesma possui conceitos que estão na parte superior da hierarquia das ontologias. Por exemplo, *França* é uma freguesia do *concelho de Bragança* e um país, como país está acima na hierarquia das ontologias, o SEI-Geo assume que o nome *França* num texto refere-se ao país e não à freguesia, a não ser que esteja precedido pelo conceito *freguesia*.

13.4 Análise dos resultados

Essa seção descreve os resultados da participação do SEI-Geo no Segundo HAREM. A avaliação dos sistemas participantes nesse evento foi realizada através de seis cenários seletivos, nos quais a categoria `LOCAL` estava presente nos cenários 2, 3, 4, 5 e 6. O cenário seletivo 5 é formado exclusivamente pela categoria `LOCAL` com os tipos `FISICO` e `HUMANO` e todos seus subtipos, o que corresponde ao cenário de participação do SEI-Geo.

A tabela 13.2 apresenta os resultados alcançados no cenário seletivo 5 considerando a classificação com avaliação relaxada de ALT, uma vez que o SEI-Geo não usa a opção de marcação com a etiqueta ALT. A última linha da tabela 13.2 apresenta os resultados dos melhores sistemas para cada medida.

Tabela 13.2: Resultados cenário seletivo 5 considerando a classificação com avaliação relaxada de ALT.

Corrida	Classificação			Identificação		
	P	A	F	P	A	F
2	**0,6821**	0,5182	0,5890	0,7109	0,5346	0,6102
3	0,6801	0,5377	**0,6006**	**0,7075**	0,5552	0,6222
4	0,6726	**0,5413**	0,5999	0,7009	**0,5595**	**0,6223**
Melhor sistema	0,7105	0,7126	0,6325	0,7212	0,8017	0,6651

As figuras 13.3 e 13.4 apresentam um comparativo dos resultados do SEI-Geo comparados com os demais sistemas participantes. A figura 13.3, referente à classificação, mostra que o SEI-Geo, nas duas melhores corridas (3 e 4), conseguiu atingir resultados acima da média dos sistemas em todas as medidas: precisão, abrangência e medida F. No que diz respeito à identificação, a figura 13.4 apresenta o SEI-Geo com valores de precisão e medida F acima da média dos sistemas, mas com abrangência inferior, o que evidencia uma das limitações do SEI-Geo.

Os resultados da classificação, na tabela 13.2, evidenciam a qualidade do SEI-Geo no reconhecimento de locais para o português.

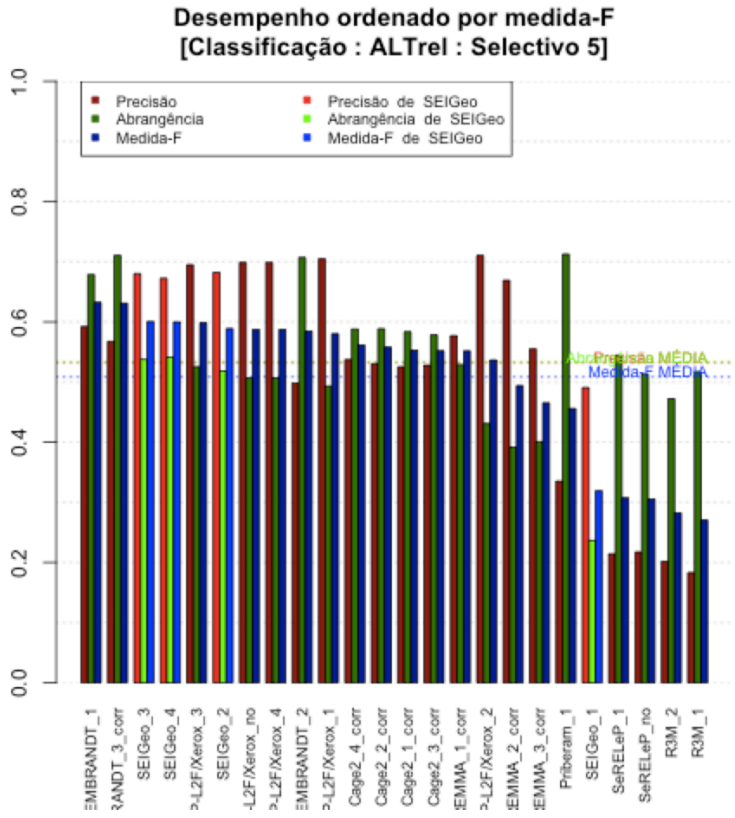

Figura 13.3: Cenário Seletivo 5 - Resultado da classificação ordenado pela medida F.

Além desses resultados, o SEI-Geo também alcançou o primeiro lugar na medida de precisão nos cenários total, 2, 3, 4 e 6 para a tarefa de classificação e identificação com avaliação relaxada de ALT. Os valores da precisão nesses cenários variam de 0,86 a 0,91.

Nos resultados por categoria, que tem em conta todos os tipos e subtipos da categoria LOCAL, a tabela 13.3 indica que o SEI-Geo aproxima-se bastante do melhor sistema nas medidas de precisão e medida F na tarefa de classificação. No que diz respeito à identificação, o SEI-Geo é o sistema mais preciso entre os concorrentes e alcançou um valor próximo ao melhor sistema na medida F.

A tabela 13.4 apresenta os resultados do SEI-Geo no HAREM clássico distribuídos por subtipos da categoria LOCAL. O SEI-Geo apresenta os melhores resultados para os subtipos PAIS, AGUACURSO e AGUAMASSA. Embora carecendo de informação sobre a geografia física na geo-ontologia, o SEI-Geo ainda é capaz de alcançar resultados competitivos por meio do uso de padrões para os subtipos físicos.

A tabela 13.5 apresenta os resultados das principais corridas do SEI-Geo na tarefa de ReRelEM. Apesar do sistema identificar corretamente as relações que se propõe a identificar, sua abrangência ainda é muito baixa, comparada ao melhor sistema nessa medida. Conforme já descrito no capítulo 4 e de acordo com a tabela 13.5, o SEI-Geo foi o melhor

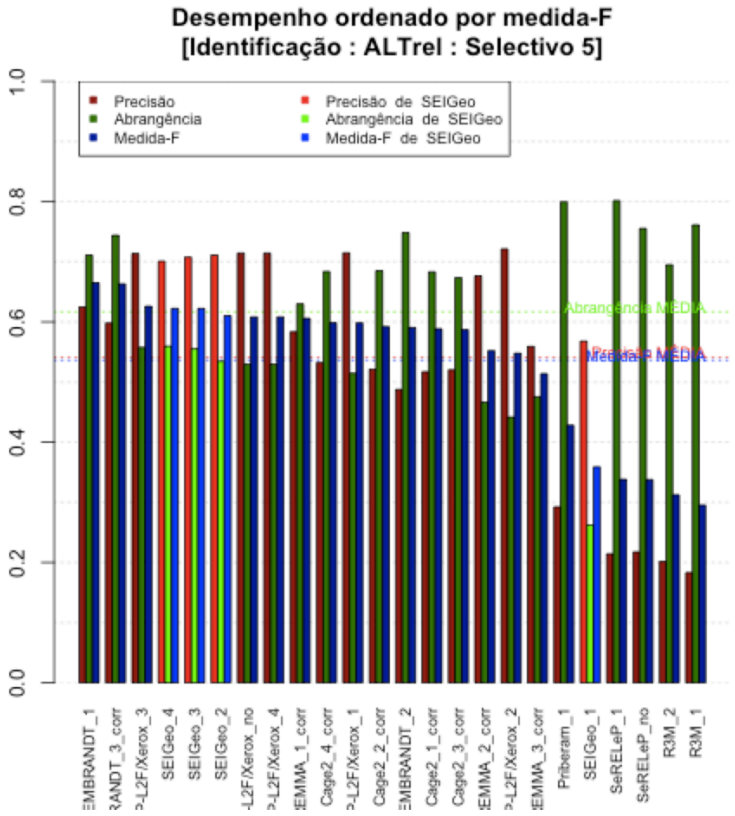

Figura 13.4: Cenário Seletivo 5 - Resultado da identificação ordenado pela medida F.

Tabela 13.3: Resultados da categoria LOCAL considerando a classificação com avaliação relaxada de ALT.

Corrida	Classificação			Identificação		
	P	A	F	P	A	F
2	**0,6830**	0,5029	0,5793	**0,7121**	0,5175	0,5994
3	0,6810	0,5215	**0,5906**	0,7087	0,5375	0,6113
4	0,6736	**0,5252**	0,5902	0,7020	**0,5416**	**0,6115**
Melhor sistema	0,6928	0,7015	0,6078	**0,7121**	0,7982	0,6376

sistema no reconhecimento de relações de inclusão para a categoria LOCAL. O SEI-Geo não reconheceu relações de identidade (por exemplo, USA=EUA) nos textos. O reconhecimento desse tipo de relação foi deixado para trabalho futuro.

Tabela 13.4: Avaliação dos subtipos da categoria LOCAL.

	Precisão	Abrangência	Medida F
PAIS	0,8488	0,6518	**0,7503**
DIVISAO	0,6384	0,3818	0,5101
REGIAO	1,0000	0,0448	0,5224
CONSTRUCAO	0,3636	0,0220	0,1928
RUA	0,4615	0,1818	0,3216
OUTRO	0,0408	0,0625	0,0517
AGUACURSO	0,7143	0,6250	**0,6697**
AGUAMASSA	0,8889	0,4444	**0,6666**
RELEVO	0,5714	0,4000	0,4857
PLANETA	0,3333	0,3333	0,3333
ILHA	0,3333	0,1111	0,2222

Tabela 13.5: Resultado da participação do SEI-Geo na tarefa de ReRelEM do Segundo HAREM - Avaliação de relações - Cenário total - Inclusão.

Corrida	P	A	F	Espúrios	Falta	Tot. CD	Tot. id.	Tot. correc. id.
3	1,0	0,0769	0,1428	0	72	78	6	6
2	**0,9166**	**0,2973**	**0,4490**	2	52	74	24	22
4	0,9166	0,2820	0,4314	2	56	78	24	22
Melhor sistema	1,0	0,4231	**0,4490**	0	52	74	24	22

13.5 Discussão

Após a análise dos resultados da participação do SEI-Geo no Segundo HAREM é possível concluir que a combinação das geo-ontologias WGO e Geo-Net-PT produziu os melhores resultados. A contribuição da Geo-Net-PT ainda é mínima, mas o suficiente para ser um diferencial quando os resultados são comparados com os outros sistemas participantes.

É importante notar que a Geo-Net-PT foi mutilada no nível de localidade. Os nomes de localidade inserem muitos falsos positivos no processo de reconhecimento de EM. O uso de nomes de localidade da Geo-Net-PT (por exemplo, *Caracol*, *Namorados* e *Nabo*) implica numa sobre-geração de EM reconhecidas. Por exemplo, nas sentenças (13.1) a (13.3) os nomes poderiam ser reconhecidos pelo SEI-Geo como locais por estarem nas geo-ontologias.

(13.1) *Caracol* é barato em Aveiro.

(13.2) *Namorados* são sempre felizes.

(13.3) *Nabo* de qualidade encontra-se na feira do Manuel.

A Corrida 1 alcançou o pior resultado entre as quatro corridas submetidas. A restrição à Geo-Net-PT implicou numa grande perda de abrangência e precisão do SEI-Geo. Esse resultado pode ser surpreendente quando comparado à participação do sistema CaGE no Mini-HAREM. Ao utilizar somente a Geo-Net-PT, o CaGE teve uma perda de precisão, abrangência e, consequentemente, medida F[3] que atingiu apenas 2 pontos percentuais em

[3] A medida F usando a Geo-Net-PT foi de 0,6063, enquanto usando a Geo-Net-PT+WGO o CaGE alcançou 0,6235.

relação à saída que utilizou a Geo-Net-PT e a WGO. Tal fato é um indício de que o SEI-Geo é bastante dependente do conteúdo das geo-ontologias, ao contrário do CaGE.

A principal limitação do SEI-Geo está na medida de abrangência. Tal fato pode ser justificado pela simplicidade do sistema, uma vez que não há análise sintática do texto, o conjunto de padrões é limitado e as ontologias são desprovidas de locais físicos (apenas a WGO contém locais físicos customizados para as participações do sistema da Universidade de Lisboa nas quatro edições do GeoCLEF).

Por outro lado, o SEI-Geo apresentou resultados satisfatórios para a medida de precisão nas corridas 3 e 4, obtendo o melhor resultado no cenário seletivo na tarefa de identificação de locais.

Os valores de precisão do SEI-Geo são prejudicados pelo fato de o sistema não discernir entidades mencionadas em contexto. Por exemplo, na frase (13.4) o SEI-Geo reconhece *Aveiro* como uma entidade mencionada da categoria `LOCAL` ao invés de `PESSOA` do tipo `POVO`, conforme a diretiva do Segundo HAREM.

(13.4) *Aveiro* estava em festa durante o Segundo HAREM.

Quanto a participação do SEI-Geo na tarefa de ReRelEM, os resultados indicam que a abordagem e as geo-ontologias utilizadas auxiliam bastante, mas não são suficientes para reconhecer relações entre locais em textos, apesar de o SEI-Geo ter sido o melhor sistema no reconhecimento de relações de inclusão.

Cabe ainda destacar que a tarefa de ReRelEM é avaliada sobre uma coleção de 12 documentos com 579 entidades mencionadas e 603 relações, que após expansão totalizam 5.716 relações.

Finalmente, uma nota sobre o custo computacional do SEI-Geo. A tabela 13.6 apresenta os tempos de processamento das corridas submetidas ao Segundo HAREM. Esses tempos foram obtidos em um servidor com sistema operacional Linux, processador Intel(R) Xeon(TM) CPU 3.20GHz e 8GB de memória.

Tabela 13.6: Tempos de processamento das corridas submetidas ao Segundo HAREM.

Corrida	1	2	3	4
Minutos	27	76	30	101

13.6 Conclusões

Esse capítulo descreveu a participação do SEI-Geo no Segundo HAREM, evidenciando os pontos positivos e negativos do sistema. A participação no HAREM clássico foi bem sucedida e o sistema atingiu resultados próximos aos sistemas que representam o estado da arte no REM em português. Na tarefa de ReRelEM, ainda há muito que melhorar no sistema dada a limitação do reconhecimento de relações baseado somente em geo-ontologias. Contudo, os melhores sistemas nessa tarefa alcançaram resultados que estão bastante distantes do expectável, considerando que em tarefas de REM e reconhecimento de relações são esperados valores mais elevados de medida F. Os resultados dos três sistemas participantes nessa tarefa apontam para a necessidade de novas abordagens para se tratar o problema.

Trabalhos futuros com o SEI-Geo incluem: melhor tratamento na identificação e reconhecimento de endereços, locais da geografia física e relações entre os locais. Especificamente nesse último item, a expansão das geo-ontologias com locais históricos, nomes alternativos e locais da geografia física é fundamental. Além disso, como o modelo-base da base de conhecimento onde as geo-ontologias estão armazenadas suporta a inserção de novos domínios de conhecimento, o domínio de organizações pode auxiliar na identificação e reconhecimento de relações, uma vez que locais frequentemente são referenciados próximos a organizações em textos. Finalmente, a exploração do uso de locativos para relacionar locais presentes na mesma sentença pode auxiliar o SEI-Geo a melhorar seu desempenho na tarefa de ReRelEM.

Agradecimentos

Este trabalho foi financiado pela FCT através do Projeto Linguateca (POSC 339/1.3/C/NAC), do Projeto GREASE II (PTDC/EIA/73614/2006) e pelo Programa de Financiamento Plurianual (LaSIGE).