Geo-ontologias e padrões para reconhecimento de locais em textos: a participação do SEI-Geo no Segundo HAREM

Marcirio Silveira Chaves
Pólo XLDB da Linguateca
LaSIGE - Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa

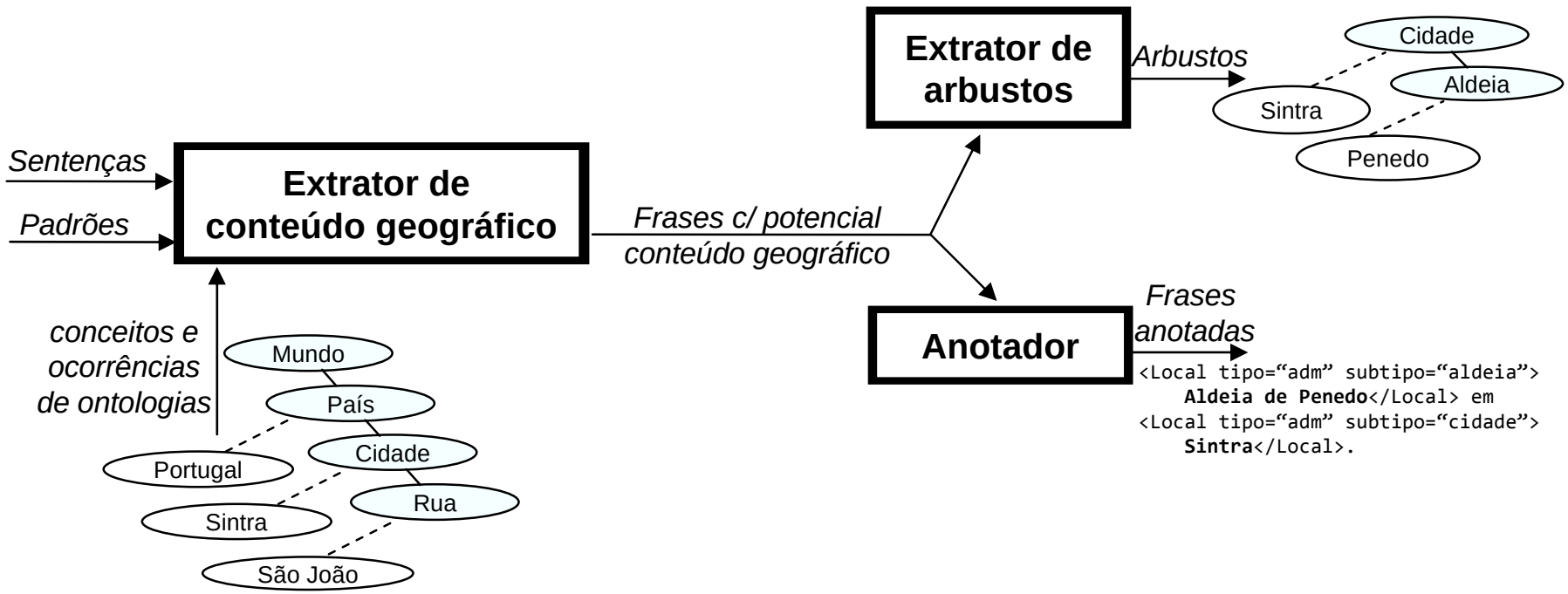
Encontro do Segundo HAREM
Aveiro - Portugal
07 de setembro de 2008

Motivação

- Validação da hipótese:
  - Padrões e geo-ontologias são suficientes para reconhecimento de locais
- Comparação com sistemas do estado da arte
- Avaliação de parte do Sistema de Extração, anotação e Integração de Conhecimento Geográfico (SEI-Geo)

Estrutura da apresentação

SEI-Geo: Arquitetura - Módulo de Extração e Anotação



Padrões

- **Conceitos geográficos**: Conceitos de uma geo-ontologia existente mais conceitos complementares
- **Padrões do tipo Hearst** traduzidos para o português e estendidos
  - e.g.: 'é o distrito', 'é um concelho' e 'é uma das cidades'
  - '[Nome de local] é um (d[eao]s?)? [Conceito]' e '[Conceito] tal(is) como [Nome de local]'

Padrões

- Relacionamentos métricos, direcionais, fuzzy e orientação:
  - **Métricos**: descrevem proximidade (e.g. 'km', 'minutos' e 'cerca de')
  - **Direcionais**: 'ao lado', 'atrás' e 'em frente'
  - **Fuzzy**: proximidade através da utilização de termos qualitativos e imprecisos (e.g. 'próximo', 'perto' e 'acima')
  - **Orientação**: expressos através de cardinais (e.g. 'norte', 'sul', 'leste' e 'oeste')

Padrões

- **Adjetivos**: capital(ais), litoral(ais), longe, natural(ais) e procedente(s).
- **Advérbios**: 'cá', 'aqui' e 'lá'.
- **Verbos**: chegar, era, falecer, morar, etc...
- **Nomes de Entidades**: ocorrências das geo-ontologias.

Geo-Ontologias

- A geo-ontologia completa de Portugal (Geo-Net-PT)
  - + de 400.000 entidades
  - recurso público e gratuito desenvolvido no Pólo XLDB da Linguateca em colaboração com o projecto GREASE
  - disponível em **http://xldb.fc.ul.pt/geonetpt**
- Os valores da Geo-Net-PT incluem os top 10 conceitos da geo-ontologia, até o nível de freguesia.
  - **PAI,NT1,NT2,NT3,REG,PRO,DST,ILH,CON,FRG**

Geo-Ontologias

- *World Geographic Ontology* (WGO)
  - Contém nomes, conceitos e relacionamentos sobre as principais divisões administrativas do mundo
  - países, territórios e cidades $\geq$ 100.000 habitantes
  - entidades geográficas no domínio físico (e.g. oceanos, mares e montanhas)

Estatística sobre as geo-ontologias utilizadas pelo SEI-Geo

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

Descrição das Corridas

- C1: Geo-Net-PT até o nível de localidade, ou seja, conceitos e entidades geográficas acima do conceito de localidade inclusive
- C2: **WGO - Relacionamento com Âmbito no Documento**
- C3: **Duas Ontologias - Relacionamento com Âmbito na Sentença**
- C4: **Duas Ontologias - Relacionamento com Âmbito no Documento**

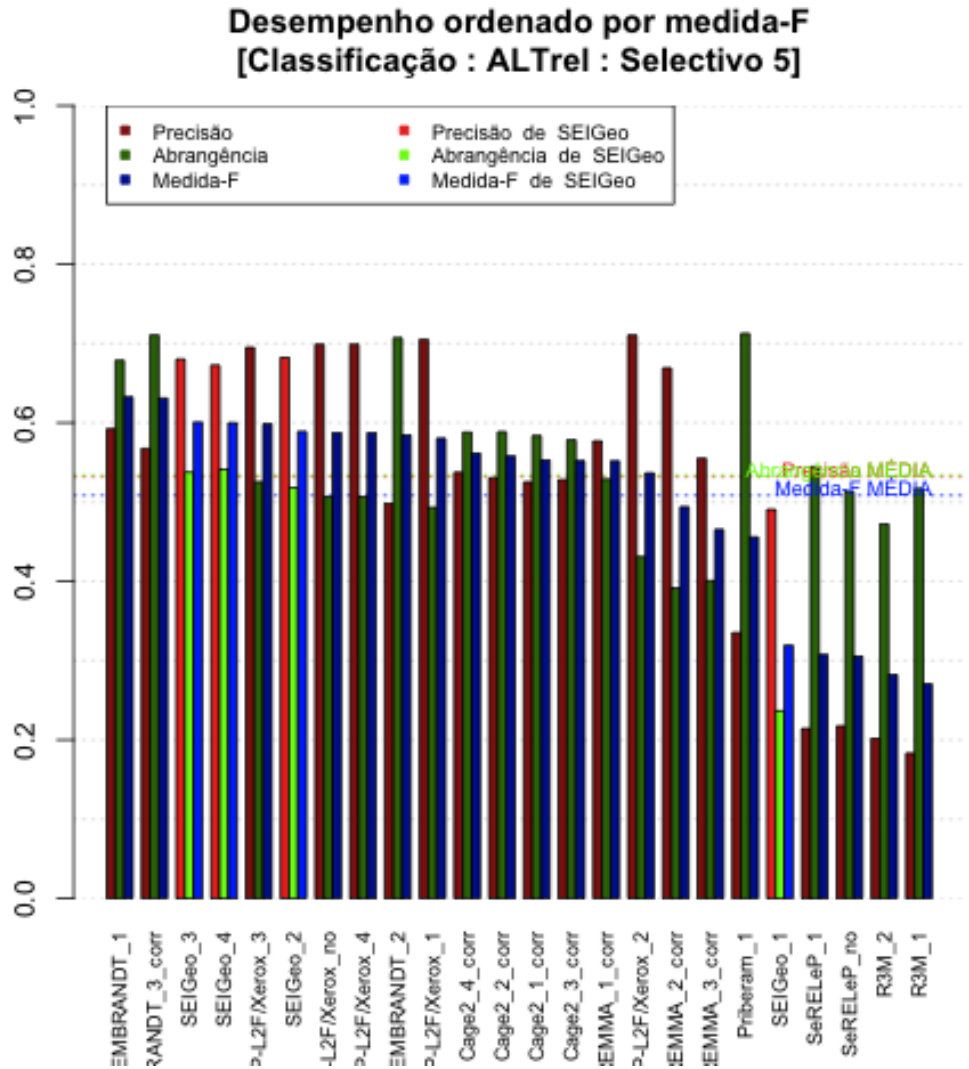
Contexto da Participação do SEI-Geo

- A avaliação dos sistemas participantes no Segundo HAREM:
  - 6 cenários seletivos
  - Categoria Local estava presente nos cenários 2, 3, 4, 5 e 6
- Cenário seletivo 5
  - Categoria Local com os tipos Físico e Humano e todos seus sub-tipos

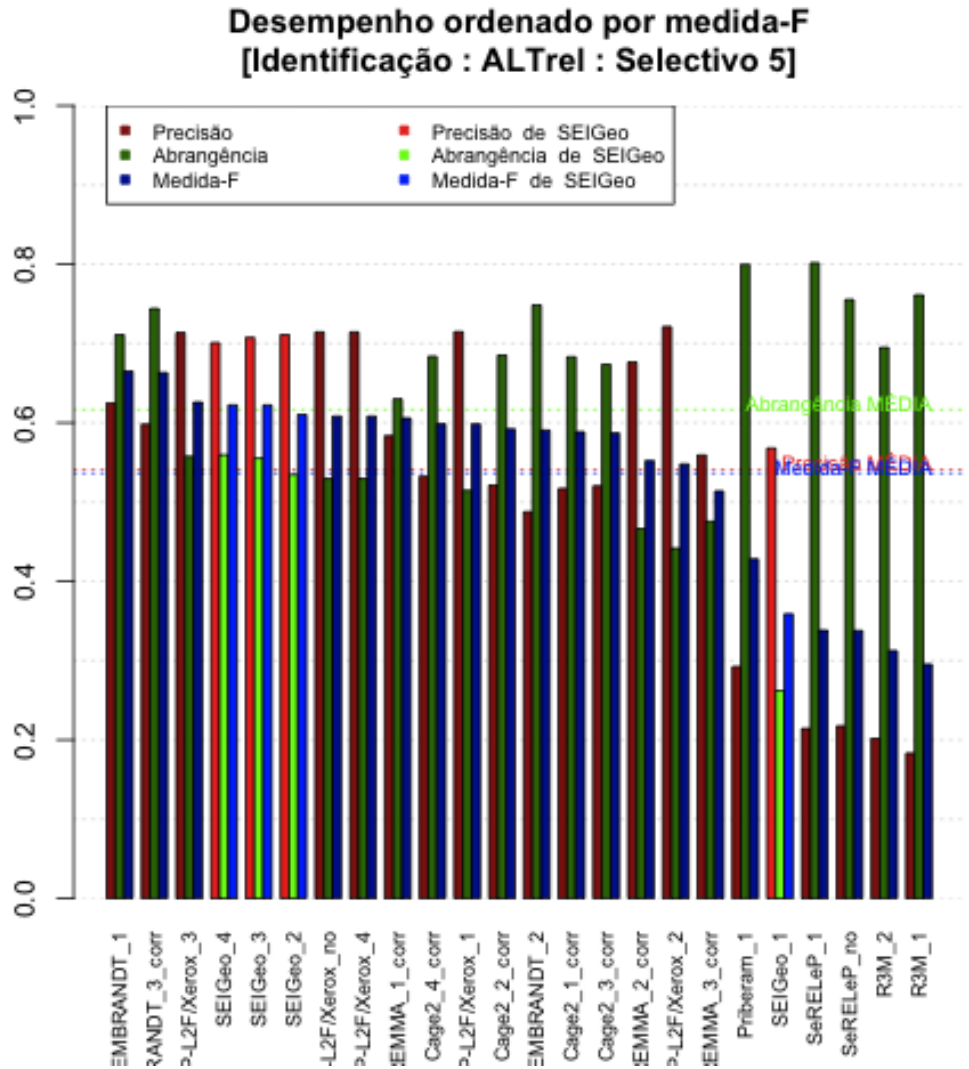
Resultados Cenário Seletivo 5 - Avaliação relaxada de ALT

	Classificação Semântica Combinada			Identificação		
Corrida	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	0,8332	0,8134	0,7939

Cenário Seletivo 5 - Resultado da **classificação** ordenado pela medida F



Cenário Seletivo 5 - Resultado da **identificação** ordenado pela medida F



Análise

- SEI-Geo pode alcançar os melhores sistemas no reconhecimento de locais para o português
- SEI-Geo alcançou o primeiro lugar na medida de precisão nos cenários Total, 2, 3, 4 e 6 para a tarefa de identificação e identificação com avaliação relaxada de ALT
  - Os valores da precisão nesses cenários variam de 0,86 à 0,91

Resultado por Categoria: Local

Tabela: Resultados da categoria Local - Avaliação relaxada de ALT.

	Classificação Semântica Combinada			Identificação		
Corrida	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

- Classificação:
  - SEI-Geo aproxima-se bastante do melhor sistema nas medidas de precisão e medida F
- Identificação:
  - SEI-Geo: **sistema mais preciso** entre os concorrentes
  - Segundo lugar na medida F

Resultado da participação do SEI-Geo na tarefa de ReRelEM do Segundo HAREM - Avaliação de Relações

Corrida	P	A	F	Espúrios	Falta	Tot. CD	Tot. id.	Tot. correc. id.
3	1,0	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	0,4122	0,5747	2	72	122	52	50

- + Precisão
- - Abrangência
- - Medida F

Considerações: a Participação do SEI-Geo no Segundo HAREM

- Combinação das ontologias WGO e Geo-Net-PT produziu os melhores resultados
- Contribuição da Geo-Net-PT é mínima, mas o suficiente para ser um diferencial quando os resultados são comparados com os outros sistemas participantes
- Geo-Net-PT foi mutilada no nível de localidade
  - Os nomes de localidade inserem muitos falsos positivos no processo de reconhecimento de EMs
  - (e.g. 'Caracol', 'Namorados' e 'Nabo') implica numa sobre-geração de EMs reconhecidas

Considerações: a Participação do SEI-Geo no Segundo HAREM

- Principal limitação do SEI-Geo: Abrangência
- Resultados satisfatórios para a medida de precisão nas corridas 3 e 4 - melhor resultado no cenário total na tarefa de identificação de locais
- No cenário seletivo 5 - classificação semântica - atingiu o segundo melhor resultado com medida F de 0,6006

Considerações: a Participação do SEI-Geo no Segundo HAREM

- Desempenho na tarefa de ReRelEM indica que a abordagem e as geo-ontologias utilizadas não são suficientes para reconhecer relações entre locais em textos
- A partir dos resultados do HAREM Clássico já eram esperados resultados menos positivos na tarefa de ReRelEM

Conclusões

- Participação no HAREM Clássico bem sucedida
- SEI-Geo atingiu resultados próximos os 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 mínimo esperado

Trabalhos Futuros

- Melhor tratamento na identificação e reconhecimento de endereços, locais da geografia física e relações entre os locais
- Relações: a expansão das geo-ontologias com locais históricos, nomes alternativos e locais da geografia física é fundamental
- 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
  - Locais frequentemente são referenciados próximos a organizações em textos