

Prospecção de Conceitos Geográficos na Web

David Soares Batista

Orientado por:

Prof. Mário J. Gaspar da Silva

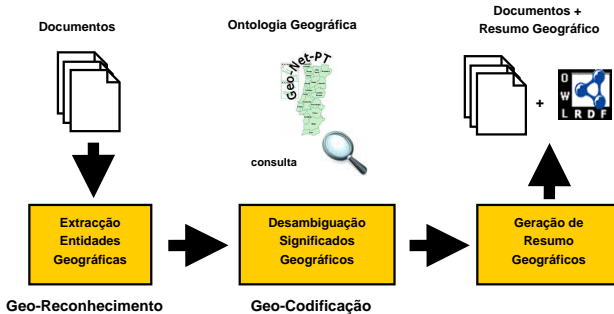
XLDB, LaSIGE, Faculdade de Ciências

November 13, 2009

Motivação

- Análise da recolha da Web portuguesa de 2003 mostra que cerca de 76% dos documentos contêm referências geográficas, *Chaves and Santos (2004)*
- Análise aos *logs* de consultas a um motor de pesquisa: 18.6% contêm um termo geográfico, 14.8% o nome de um lugar, *Sanderson and Kohler (2004)*

Objectivos



- Resumo geográfico: referências geográficas desambiguadas

Extracção de Informação Geográfica

- **Geo-codificação:** associação de entidades a conceitos geográficos
 - ① *Ambiguidade referente:* mesmo nome é usado para mencionar mais do que um local
 - "Souto"
 - 1 aldeia
 - 6 freguesias
 - ② *Ambiguidade referênciã:* o mesmo local tem vários nomes
 - Praça do Comércio
 - Terreiro do Paço
 - ③ *Ambiguidade na classe do referente:* o nome do local com outros significados
 - "Souto": mata de castanheiros

Conditional Random Fields

- probabilidade de um dado elemento (ex: palavra) pertencer a uma determinada categoria: $p(\vec{y}|\vec{x})$
- dependência condicional de cada y_i em \vec{x} : $f(i, y_{i-1}, y_i, \vec{x})$
- "Rolling Stones tocam na cidade de **Coimbra**"

palavra	categoria	funções geradas
Rolling	NEG	tokens.eq.charTypePattern.X+=1; previousLabel.1.null=1
Stones	NEG	tokens.eq.charTypePattern.X+=1; previousLabel.NEG=1
tocam	NEG	tokens.lenght=5; previousLabel.NEG=1
na	NEG	right.tokenNeg_1.eq.charTypePattern.X+x=0
cidade	NEG	tokens.eq.charTypePattern.X+=0; previousLabel.1.END=1
de	NEG	tokens.eq.charTypePattern.X+=0;left.tokenNeg_1.eq.lc.cidade
Coimbra	POS	tokens.eq.charTypePattern.X+=1;left.tokenNeg_1.eq.lc.de left.tokenNeg_2.eq.lc.cidade

Conditional Random Fields: vantagens

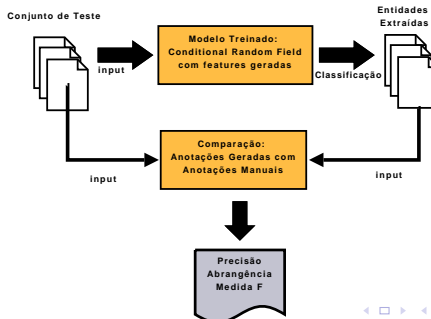
- número arbitrário de funções de característica
- avaliação de toda a sequência de entrada
- redução de pressupostos de independência entre as variáveis de observação

Conditional Random Fields

Fase de Treino



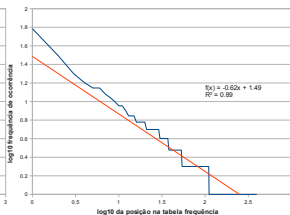
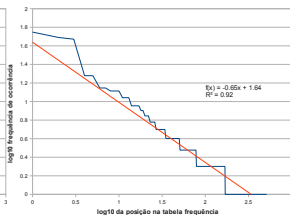
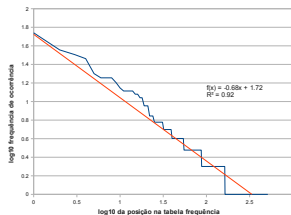
Fase de Classificação



HAREM - Coleções Douradas

- 3 corpus anotados para 10 tipos de entidades: **abstracção, acontecimento, coisa, local, obra, organização, pessoa, tempo, valor, outro**
- Aplicação de um *script* XSLT mantém apenas <local>

	MiniHAREM	HAREM I	HAREM II
Tamanho	514 Kbytes	734 Kbytes	1.1 Mbytes
Nº Entidades Únicas	397	514	612
Total	792	1146	1200



HAREM - Colecções Douradas

- Tabelas de Frequências Acumuladas

HAREM I		Mini-Harem		HAREM II	
Entidade	Freq. Acumulada	Entidade	Freq. Acumulada	Entidade	Freq. Acumulada
Brasil	4.80%	Brasil	7.70%	Lisboa	4.67%
São Paulo	7.94%	São Paulo	11.62%	Portugal	8.75%
Portugal	10.73%	Itália	14.14%	Brasil	12.67%
Braga	13.26%	Angola	16.16%	Coimbra	14.25%
Lisboa	15.01%	Braga	17.93%	EUA	15.83%
Europa	16.58%	Egito	19.70%	Europa	17.00%
Porto	18.15%	Portugal	21.21%	Porto	18.17%
Espanha	19.72%	Santos	22.60%	França	19.25%
Guimarães	21.12%	São Vicente	23.86%	Detroit	20.33%
Marília	22.34%	Europa	25.00%	São Paulo	21.42%

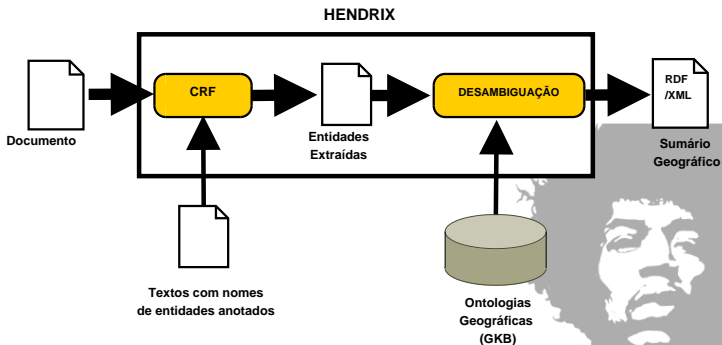
Ontologias Geográficas

- Geo-Net-PT
- WGO - World Geographic Ontology
- Wiki WGO 2009 (Wikipédia 2009)

	Geo-Net-PT		WGO		Wiki WGO
	Adm	Fis	Adm	Fis	-
Conceitos Geográficos	388 049	5 662	12 982	721	136 347
Nomes	265 044	8 250	12 102	750	499 820
Relações parte-de	386 431	390	12 562	513	-
Relações de adjacência	33 051	2 404	11 170	12	-

HENDRIX

Hendrix is an **Entity Name Desambiguator and Recognizer** for **Information eXtraction**



Modelo Conditional Random Fields

MinorThird: *"a collection of Java classes for storing text, annotating text, and learning to extract entities and categorize text"* W. Cohen, Carnegie Mellon University (suporte para Conditional Random Fields)

- Treino: CD de Mini-Harem + HAREM I
- Teste: CD HAREM II

Precisão	Abrangência	Medida-F
0,64	0,45	0,53

PAREDES

Paredes **A**dvocates **R**ecognized **E**ntities for **D**esambiguation and **E**xtraction of **S**ummaries

- Emparelhamento de nomes de entidades extraídas com nomes de conceitos geográficos
- Desambiguação significados geográficos
 - Redução de referências
 - Relações entre entidades
 - Medidas de Semelhança Semântica
- Geração de resumos geográficos



Processo de Emparelhamento

- 2 tipos de consultas:
 - 1) caracteres minúsculos e expansão abreviaturas
 - ex: "Sta. Iria da Azóia" → "santa iria da azóia"

Processo de Emparelhamento

- 2 tipos de consultas:
 - 1) caracteres minúsculos e expansão abreviaturas
 - ex: "Sta. Iria da Azóia" → "santa iria da azóia"
 - 2) detectar tipo conceito geográfico (ex: rua, concelho, avenida, etc)
 - se detecta: usa tipo de conceito geográfica na consulta
 - ex: "Avenida da Liberdade" → todas avenidas com nome "liberdade"

Desambiguação

- Reduzir a **número mínimo** os **conceitos geográficos** associados a um nome
- **Âmbito geográfico** de um documento: reduzir a **um único conceito geográfico**

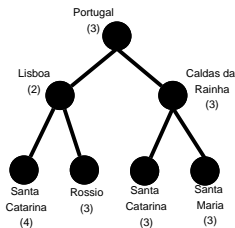
Desambiguação

- Reduzir a **número mínimo** os **conceitos geográficos** associados a um nome
- **Âmbito geográfico** de um documento: reduzir a **um único conceito geográfico**
- Redução de número referências a conceitos geográficos
 - Apenas referências encontradas na Geo-Net-PT
 - Entidades sem tipo de conceito associado, apenas subdivisões, excluem-se os arruamentos:
 - ex: "Beja" poderá corresponder a: 1 Distrito, 1 Concelho, 3 Freguesias ou 15 arruamentos.
 - ex: "Londres" corresponde a 8 conceitos geográficos do tipo arruamento
 - Entidades com tipo de conceito associado, todas as referências são utilizadas

Desambiguação

Medidas de Similaridade Semântica

- Âmbito
outra tese mestrado
- API permite saber
a proximidade semântica
- $SSM(IC_1, IC_2) \in [0, 1]$



Desambiguação

Medidas de Similaridade Semântica

- "Eu nasci e vivi perto de **Santa Catarina em Lisboa**"
 - Lisboa, Concelho (#146)
 - Lisboa, uma localidade no Concelho de Monção (#379800)
 - Santa Catarina, Freguesia no Concelho de Lisboa (#418458)
 - Santa Catarina, Freguesia nas Caldas da Rainha (#295404)

$$SSM(146, 418458) = 0.584$$

$$SSM(146, 295404) = 0.065$$

$$SSM(379800, 418458) = 0.063$$

$$SSM(379800, 295404) = 0.041$$

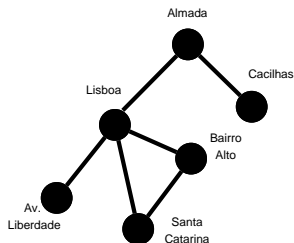
Desambiguação

- **Heurística I:**
 - Nomes de entidades associados a referências a conceitos geográficos
 - Extrair todas as relações possíveis: grafo
 - Âmbito geográfico documento: referência que mais relações.

Desambiguação

● Heurística I:

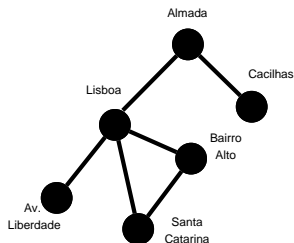
- Nomes de entidades associados a referências a conceitos geográficos
- Extrair todas as relações possíveis: grafo
- Âmbito geográfico documento: referência que mais relações.
- **Lisboa; Av. Liberdade; Almada; Cacilhas; Bairro Alto; Santa Catarina;**



Desambiguação

● Heurística I:

- Nomes de entidades associados a referências a conceitos geográficos
- Extrair todas as relações possíveis: grafo
- Âmbito geográfico documento: referência que mais relações.
- **Lisboa; Av. Liberdade; Almada; Cacilhas; Bairro Alto; Santa Catarina;**



- Âmbito geográfico: Lisboa

Desambiguação

- **Heurística II:** aplica medidas de similaridade semântica a pares de entidades pela ordem de ocorrência no documento.
 - ex: "...deslocou-se pela **Avenida da República** em direcção ao **Marquês de Pombal**, aí apanhou o metro em direcção ao **Rossio**"
 - Âmbito geográfico do documento: antecessor comum mais próximo.

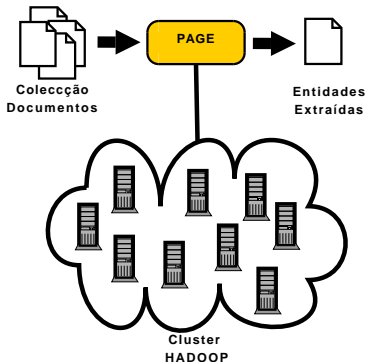
Desambiguação

- **Heurística II:** aplica medidas de similaridade semântica a pares de entidades pela ordem de ocorrência no documento.
 - ex: "...deslocou-se pela **Avenida da República** em direcção ao **Marquês de Pombal**, aí apanhou o metro em direcção ao **Rossio**"
 - Âmbito geográfico do documento: antecessor comum mais próximo.
- **Heurística III:** semelhante à heurística II: em vez de se calcular o antecessor comum, extrai as relações entre as entidades.
- Âmbito geográfico documento: referência com mais relações.

PAGE

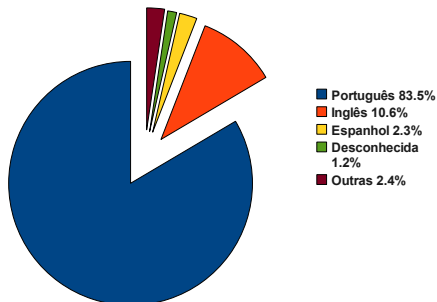
Page **A**cquires **G**eographic **E**ntities

Hadoop (MapReduce) + Modelo CRF Treinado



WPT05

- Recolha da Web Portuguesa de 2005:
 - 12 Milhões Documentos
 - Versão pré-processada: 40 GBytes texto
 - Identificação Linguística (n -gramas)



WPT05

- 7 milhões documentos em português
- 26 GBytes texto.
- Extracção sob o PAGE, cerca 16 dias:
 - 4 x Intel(R) Xeon(R) CPU @ 2.50GHz
 - 6 x Quad-Core AMD Opteron(tm) Processor 2350 @ 1GHz
- Extraídas 78 326 entidades únicas
- 18 586 (23.7%) correspondentes nas ontologias

Ontologia	Nº Entidades	Percentagem
Geo-Net-PT02	13 097	70.47%
World Geographic Ontology	2 191	11.79%
Wiki WGO 2009	8 742	47.04%

Avaliação Heurísticas

- 18 páginas da Wikipédia para os distritos de Portugal
- Âmbitos geográficos gerados pelo HENDRIX

Artigo	Heurística I	Heurística II	Heurística III
Aveiro	Aveiro (Distrito)	Portugal (PAI)	Aveiro (Distrito)
Beja	Beja (Distrito)	Continente (NT1)	Beja (Distrito)
Braga	Norte (NT2)	Continente (NT1)	Norte (NT2)
Bragança	Norte (NT2)	Norte (NT2)	Norte (NT2)
Castelo Branco	Beira Baixa (Província)	Continente (NT1)	Beira Baixa (Província)
Coimbra	Porto (Distrito)	Continente (NT1)	Coimbra (Distrito)
Évora	Alentejo Central (NT3)	Continente (NT1)	Alentejo Central (NT3)
Faro	Algarve (NT2)	Continente (NT1)	Algarve (Província)
Guarda	Guarda (Distrito)	Continente (NT1)	Guarda (Distrito)
Leiria	Leiria (Distrito)	Continente (NT1)	Beira Litoral (Província)
Lisboa	Lisboa (NT2)	Continente (NT1)	Grande Lisboa (NT3)
Portalegre	Norte (NT2)	Continente (NT1)	Norte (NT2)
Porto	Braga (Distrito)	Continente (NT1)	Porto (Distrito)
Santarém	Santarém (Distrito)	Continente (NT1)	Alentejo (NT2)
Setúbal	Lisboa (NT2)	Continente (NT1)	Lisboa (NT2)
Viana do Castelo	Viana do Castelo (Distrito)	Norte (NT2)	Viana do castelo (Distrito)
Vila Real	Minho (Província)	Continente (NT1)	Vila Real (Distrito)
Viseu	Norte (NT2)	Continente (NT1)	Viseu (Distrito)

GikiCLEF2009

- resposta a perguntas:
 - envolve raciocínio geográfico
 - respostas: artigos na Wikipedia
- 1 modelo CRF treinado para 4 tipos entidades
 - PESSOA, LOCAL, EVENTO, ORGANIZAÇÃO
- Problema detectado: 1 modelo CRF não é suficiente

Entidade	Precisão	Abrangência	Medida-F
PESSOA	0.5915	0.4095	0.4840
LOCAL	0.4590	0.5006	0.4789
EVENTO	0.3281	0.2515	0.2847
ORGANIZAÇÃO	0.4464	0.4783	0.4618

- 8 sistemas participantes, XLDB obteve 2º lugar

Conclusões

- Valores Precisão/Abrangência baixos comparando com sistemas de regras manuais

	Precisão	Abrangência	Medida-F
REMBRANDT	0,56	0,73	0,63
SEIGeo_2	0,71	0,51	0,59
HENDRIX (M3rd CRF)	0,64	0,45	0,53
SeRELeP_1	0,22	0,79	0,34

- Modelo de CRF falha em captar expressões de localização:
*"O município é limitado a norte pelos municípios de **Cuba** e **Vidigueira**, a leste por **Serpa**, a sul por **Mértola** e **Castro Verde** e a oeste por **Aljustrel** e **Ferreira do Alentejo**."*
- apenas "Cuba" é extraído

Conclusões

- Funções características geradas são pouco expressivas
- Funções característica foram geradas automaticamente
- Corpus anotado é suficiente para um treino/avaliação eficaz ?
Leidner (2006)
- HAREM - anotação:
 - " ..das lojas Modelo de <LOCAL>Eiras</LOCAL>, no distrito de <LOCAL>Coimbra</LOCAL> e de <LOCAL>Lagoa</LOCAL>, no concelho de <LOCAL>Portimao</LOCAL>"
- Medidas de Semelhança Semântica para desambiguação

Trabalho Futuro

- *Baseline* com artigos da Wikipedia anotados
- Melhorar Modelo
 - Reanotação CD
 - Codificar expressões de localização em funções característica
- Minorthird é bom? Mais pacotes de *software* com CRF disponíveis!
- Erros ortográficos na WPT05 (Web) → emparelhamento de *strings* por aproximação

FIM