Extracting Geographic Entities with Conditional Random Fields

David Batista

XLDB, Faculty of Sciences, University of Lisbon

Priberam - Machine Learning Seminars, Lisbon
13th April 2010

Research Lines

- XLDB Research Team
- Addresses topics in:
  - geographic information retrieval, text mining and natural language processing
  - web archiving and search
  - information visualization
  - biomedical informatics

Outline

GREASE project

Geographic Reasoning for Search Engines

- information access methods to collections of documents
- geographically rich text and meta-data
- emphasis on the web
- some useful resources available (later)

(minimal) Introduction

Information Retrieval System

- 1) Crawling: downloads documents from the web
- 2) Storage: pre-processes and stores documents
- 3) Indexing: generates term indexes and weights documents
- 4) Interface: processes queries and presents results to the user

Classical Information Retrieval

- Document $\rightarrow$ Set of words
- Semantic content is ignored
- A document can only be retrieved by matching the words it contains

Geographic Information Retrieval

- Augmentation of Information Retrieval with geographic metadata
- Requires semantic data to be present (e.g: location)
- A document is only retrieved if the location name is present

How to overcame this limitation

- Semantic association of words to place names needs to be extracted
- Mapping of words to geographic concepts
- **Geo-Parsing:** Identification of place names in texts
- **Geo-Coding:** Association of a place name to a unique identifier (Geographic Knowledge Base)

How to overcame this limitation

- Semantic association of words to place names needs to be extracted
- Mapping of words to geographic concepts
- **Geo-Parsing:** Identification of place names in texts
- **Geo-Coding:** Association of a place name to a unique identifier (Geographic Knowledge Base)

How to overcame this limitation

- Semantic association of words to place names needs to be extracted
- Mapping of words to geographic concepts
- **Geo-Parsing:** Identification of place names in texts
- **Geo-Coding:** Association of a place name to a unique identifier (Geographic Knowledge Base)

Geo-Parsing

Rules based systems (developed at XLDB):

- SEI-Geo (developed by Marcirio Chaves):
  - expressions: list of verbs, prepositions, adjectives
  - geographic feature types: district, civil parish, municipality
  - list of place names from an ontology
  - e.g: *".. he lives 2 kms north from the city of Lisbon."*

Geo-Parsing

Rules based systems (developed at XLDB):

- SEI-Geo (developed by Marcirio Chaves):
  - expressions: list of verbs, prepositions, adjectives
  - geographic feature types: district, civil parish, municipality
  - list of place names from an ontology
  - e.g: *".. he lives 2 kms north from the city of Lisbon."*

Geo-Parsing

Rules based systems (developed at XLDB):

- SEI-Geo (developed by Marcirio Chaves):
  - expressions: list of verbs, prepositions, adjectives
  - geographic feature types: district, civil parish, municipality
  - list of place names from an ontology
  - e.g: *".. he lives 2 kms north from the city of Lisbon."*

Geo-Parsing

Rules based systems (developed at XLDB):

- SEI-Geo (developed by Marcirio Chaves):
  - expressions: list of verbs, prepositions, adjectives
  - geographic feature types: district, civil parish, municipality
  - list of place names from an ontology
  - e.g: *".. he lives 2 kms north from the city of Lisbon."*

Geo-Parsing

- REMBRANDT (developed by Nuno Cardoso)
  - Named-Entity Recognition (NER) system (not only place names)
  - manually crafted rules for capturing internal and external evidence of named entities
  - explores the Wikipedia document structure to classify all kinds of named entities
  - works both for Portuguese and English documents

Geo-Parsing

- REMBRANDT (developed by Nuno Cardoso)
  - Named-Entity Recognition (NER) system (not only place names)
  - manually crafted rules for capturing internal and external evidence of named entities
  - explores the Wikipedia document structure to classify all kinds of named entities
  - works both for Portuguese and English documents

Geo-Parsing

- REMBRANDT (developed by Nuno Cardoso)
  - Named-Entity Recognition (NER) system (not only place names)
  - manually crafted rules for capturing internal and external evidence of named entities
  - explores the Wikipedia document structure to classify all kinds of named entities
  - works both for Portuguese and English documents

Geo-Parsing

- REMBRANDT (developed by Nuno Cardoso)
  - Named-Entity Recognition (NER) system (not only place names)
  - manually crafted rules for capturing internal and external evidence of named entities
  - explores the Wikipedia document structure to classify all kinds of named entities
  - works both for Portuguese and English documents

Geo-Coding

- Geographic Knowledge Base: Ontology, Gazetteer
- Ambiguity Problem
- **Referent Ambiguity:** "Souto"
  - 1 Village (*aldeia*)
  - 6 Civil Parishes (*freguesias*)
- **Reference Ambiguity:**
  - "Praça do Comércio"
  - "Terreiro do Paço"
- **Referent Class Ambiguity:**
  - "Souto": forest of chestnut (*mata de castanheiros*)

Geo-Coding

- Geographic Knowledge Base: Ontology, Gazetteer
- Ambiguity Problem
- **Referent Ambiguity:** "Souto"
  - 1 Village (*aldeia*)
  - 6 Civil Parishes (*freguesias*)
- **Reference Ambiguity:**
  - "Praça do Comércio"
  - "Terreiro do Paço"
- **Referent Class Ambiguity:**
  - "Souto": forest of chestnut (*mata de castanheiros*)

Geo-Coding

- Geographic Knowledge Base: Ontology, Gazetteer
- Ambiguity Problem
- **Referent Ambiguity:** "Souto"
  - 1 Village (*aldeia*)
  - 6 Civil Parishes (*freguesias*)
- **Reference Ambiguity:**
  - "Praça do Comércio"
  - "Terreiro do Paço"
- **Referent Class Ambiguity:**
  - "Souto": forest of chestnut (*mata de castanheiros*)

Geo-Coding

- Geographic Knowledge Base: Ontology, Gazetteer
- Ambiguity Problem
- **Referent Ambiguity:** "Souto"
  - 1 Village (*aldeia*)
  - 6 Civil Parishes (*freguesias*)
- **Reference Ambiguity:**
  - "Praça do Comércio"
  - "Terreiro do Paço"
- **Referent Class Ambiguity:**
  - "Souto": forest of chestnut (*mata de castanheiros*)

Geo-Coding

- **Referent Ambiguity:**
  - Based on hierarchy levels: population, administrative divisions
  - One sense per discourse
  - Minimize bounding polygon
- **Reference Ambiguity:** load GKB with more data, e.g historical names, alternative names
- **Referent Class Ambiguity:** difficult to handle, better treated in geo-parsing phase

Geo-Coding

- **Referent Ambiguity:**
  - Based on hierarchy levels: population, administrative divisions
  - One sense per discourse
  - Minimize bounding polygon
- **Reference Ambiguity:** load GKB with more data, e.g historical names, alternative names
- **Referent Class Ambiguity:** difficult to handle, better treated in geo-parsing phase

Geo-Coding

- **Referent Ambiguity:**
  - Based on hierarchy levels: population, administrative divisions
  - One sense per discourse
  - Minimize bounding polygon
- **Reference Ambiguity:** load GKB with more data, e.g historical names, alternative names
- **Referent Class Ambiguity:** difficult to handle, better treated in geo-parsing phase

Conditional Random Fields for Geo-Parsing

Conditional Random Fields (CRF)

- Probabilistic model often used for labeling sequential data
- Probability of a given word to belong to a particular category: $p(\vec{y}|\vec{x})$
- A CRF on (X, Y) specified by:
  - a vector $f = (f_1, f_2, ..., f_m)$ of features
  - a weight vector $\lambda = (\lambda_1, \lambda_2, ..., \lambda_m)$.
- Trained automatically from annotated Corpora

Annotated Data in Portuguese for training?

- HAREM - Evaluation Contest for Named Entity Recognizers in Portuguese
- 3 Editions: 2005, 2006, 2008
- Golden Collections: hand annotated documents
- 9 categories: PERSON, ORGANIZATION, PLACE, DATETIME, MASTERPIECE, VALUE, ABSTRACTION, EVENT, THING

Training of the CRF model: corpus + features

- Machine Learning software package: Minorthird
- Training: HAREM's Golden Collections (2005 + 2006)
- Test: HAREM's Golden Collections (2008)

Properties	2005	2006	2008
Document Size	731 Kb	512 Kb	1098 Kb
Unique PLACE names	488	371	612
Total PLACE names	1099	759	1200

Trainning of the CRF model: corpus + features

- Labels: *BEGIN*, *INSIDE*, *END*, *UNIQUE*, *NEG*
- Minorthird default features:
  - *charTypePattern.9+* token is composed by numbers only;
  - *charTypePattern.X+x+* token is capitalized;
  - *eq.lc.avenida* the value of token itself;

Precision	Recall	F-Measure
0,64	0,45	0,53

Training of the CRF model: corpus + features

Features based on dictionaries (taken from SEI-Geo)

- isPreposition: *da, de, do, entre, à, ao, em, no, na, etc.*
- isFeatureType: *rua, avenida, largo, concelho, distrito, etc.*
- isLocalPrefix:
  - isAdjectiv: *capital, litoral, longe, natural, etc.*
  - isAdverb: *cá, aqui, lá, etc.*
  - isVerb: *chegar, localizar, morar, habitar, viver, etc.*
- isGeoName: place names from the ontology

Training of the CRF model: corpus + features

- Labels: *BEGIN*, *INSIDE*, *END*, *UNIQUE*, *NEG*
- Generated features:
  - *charTypePattern.9+* token is composed by numbers only;
  - *charTypePattern.X+x+* token is capitalized;
  - *eq.lc.avenida* the value of token itself;
- Dictionary based:
  - *isFeatureType*, *isGeoName*
  - *isPreposition*, *isLocalPrefix*

Precision	Recall	F-Measure
0,69	0,47	0,56

Trainning of the CRF model: Results

- Comparing with other systems

System	Precision	Recall	F-1
REMBRANDT	0.56	0.73	0.63
SEIGeo	0.71	0.51	0.59
Minorthird	0.69	0.47	0.56
SeRELeP	0.22	0.79	0.34

- Recall is low, overfitting?
- Generated features not good enough to capture all the evidences of places?
- Size of training corpus 1 243 Kb is enough?

Training of the CRF model: features for **BEGIN**

previousLabel.1.NEG	8.42
previousLabel.1.null	-1.36
right.token_0.isPreposicao.true	3.02
tokenNeg_1.eq.lc.av	2.92
tokens.eq.lc.av	2.92
eq.charTypePattern.X+ÃX+	0.73
eq.charTypePattern.X+êx+	0.60
eq.lc.bairro	-0.06
eq.lc.concelho	0.81
eq.lc.condado	0.91
eq.lc.estrada	0.41

Trainning of the CRF model: features for **CONTINUE**

previousLabel.1.localContinue	19.19
previousLabel.1.localBegin	18.21
token_0.eq.lc.da	0.38
token_0.eq.lc.das	0.44
token_0.eq.lc.de	0.66
token_0.eq.lc.do	0.60
token_0.eq.lc.dos	0.85
tokens.isFeatureType.true	-0.65
tokens.isGeoName.true	0.52
tokens.isLocalPrefix.true	-0.60
tokens.isPreposicao.true	0.74

Training of the CRF model: features for **END**

previousLabel.1.localContinue	19.61
previousLabel.1.localBegin	18.92
left.tokenNeg_1.isPreposicao.true	2.50
left.tokenNeg_1.isGeoName.true	2.48
right.token_0.isGeoName.true	-3.54
right.token_0.isPreposicao.true	-4.52

Training of the CRF model: features for **UNIQUE**

previousLabel.1.NEG	8.76
left.tokenNeg_1.eq.lc.para	5.10
left.tokenNeg_1.isFeatureType.true	4.40
tokenNeg_1.eq.lc.europa	3.90
left.tokenNeg_1.eq.lc.em	3.55
left.tokenNeg_1.eq.charTypePattern.9+	3.21
left.tokenNeg_1.eq.charTypes.AA	3.08

Conclusions

- Recall for trained CRF model is still relatively low
- Tunning of selected the features for training might increase results
- Use REMBRANDT tags as features
- BIG limitation: lack of large Portuguese labeled corpus for CRF training
- Other software packages: MALLET, LingPipe

Conclusions

- Recall for trained CRF model is still relatively low
- Tunning of selected the features for training might increase results
- Use REMBRANDT tags as features
- BIG limitation: lack of large Portuguese labeled corpus for CRF training
- Other software packages: MALLET, LingPipe

Conclusions

- Recall for trained CRF model is still relatively low
- Tunning of selected the features for training might increase results
- Use REMBRANDT tags as features
- BIG limitation: lack of large Portuguese labeled corpus for CRF training
- Other software packages: MALLET, LingPipe

Conclusions

- Recall for trained CRF model is still relatively low
- Tunning of selected the features for training might increase results
- Use REMBRANDT tags as features
- BIG limitation: lack of large Portuguese labeled corpus for CRF training
- Other software packages: MALLET, LingPipe

Conclusions

- Get more labeled data
- Use Active Learning to label more data
- Larger corpus: CHAVE - newspapers articles (2,2 GBytes)
    - automatically labeled by REMBRANDT
    - F-Measure: 56.7% for the full NER task (HAREM II)
    - F-Measure: 62.5% for the PLACE (HAREM II)

Available Resources

Geo-Net-PT02: Geographic Ontology of Portugal

Geo-Net-PT02

- public Geographic Ontology of Portugal
- contains all the geographic administrative data of Portugal (distritos, concelhos, ruas, etc)
- divided into two domains: administrative, physical
- published in the Web Ontology Language (OWL) (other formats available)
- licensed under a Creative Commons Attribution 3.0 License
- `http://xldb.fc.ul.pt/wiki/Geo-Net-PT_02_SPARQL_endpoint`

Geo-Net-PT02

Feature Type	Nº Features	(%)
Postal Code	187 014	48.44
Street Segments	146 422	37.93
Settlement	44 386	11.50
Civil Parishes	42 60	0.93
Zone	3 594	0.08
Municipality	308	0.01
NUT	40	0.01
Districts	18	0.00
Province	11	0.00
Island	11	0.00
Region	2	0.00
Country	1	0.00
Total	386 067	100.00

(a) Statistics of the Administrative Domain

Feature Type	Nº Features	(%)
Stream	2 421	42.65
Beach	588	9.83
Museum	507	8.93
Archaeological Site	414	7.29
Hotel	381	6.71
Natural Region	304	5.36
Castle	256	4.51
Spring	220	3.88
Historic Hamlet	217	3.82
Reservoir	90	1.59
Touristic Resource	84	1.48
Other	224	3.95
Total	5 676	100.00

(b) Statistics of the Physical Domain

Geo-Net-PT02

Names	Administrative	Physical
N° Names	77 748	5 209
Ambiguous	19 647 (25%)	329 (6%)
Non-Ambiguous	58 101 (75%)	4 880 (94%)

(a) Referent ambiguity in Geo-Net-PT02 names

Feature Type	Total N° Features	N° Features with a non unique name
Street	91 310	58 770 (64.36%)
Travessa	18 150	10 613 (58.47%)
Town square	7 284	4 095 (56.22%)
Avenue	3 630	1 905 (52.48%)

(b) The most ambiguous feature types in Geo-Net-PT02

WPT05: Crawl of the "portuguese" Web

WPT05

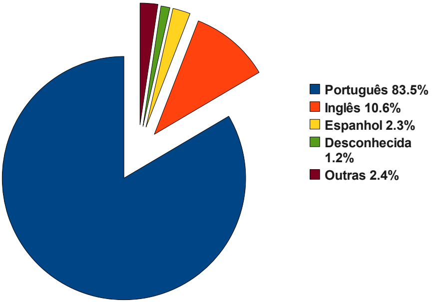
- over 10 million documents from the Portuguese web
- contents were crawled in 2005 according to the following criteria:
  - hosted in a .pt domain
  - hosted in a .com, .org, .net or .tv domain, and referenced by a hyperlink from a .pt domain.

WPT05

- Available in different versions and formats:
- WPT05 Metadata:
  - contains the attributes of each of the collected contents
  - the automatically extracted text and identified language
  - the RDF/XML format
- WPT05 Contents:
  - contains the harvested contents in raw form, as they have been archived,
  - Internet Archive ARC format

WPT05

- WPT05 identified languages:



WPT05

Geographic Entities in WPT05

- 7 millions of documents in Portuguese: 26 GBytes
- Extraction using a MapReduce cluster, 10 cores:
  - 4 x Intel(R) Xeon(R) CPU @ 2.50GHz
  - 6 x Quad-Core AMD Opteron(tm) Processor 2350 @ 1GHz
- 78 326 unique geographic entities extracted
- 18 586 (23.7%) correspond to geographic concepts

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

WPT05 N-Grams Collection: n-grams from the portuguese web crawl

WPT 05 Portuguese N-Grams

- n-grams extracted from the crawled texts
- only texts identified as Portuguese were used
- word grams from 1 to 5
    - unigrams: 9 058 689
    - bigrams: 129 248 724
    - trigrams: 501 610 788
    - tetragrams: 985 212 499
    - pentagrams: 1 323 408 463

WPT 05 Portuguese N-Grams

Example for tetragrams:

A	detenção	de	Carlos	3
A	detenção	de	certas	1
A	detenção	de	Cães	1
A	detenção	de	cidadão	1
A	detenção	de	cidadãos	2
A	detenção	de	cinco	2
A	detenção	de	clérigos	1
A	detenção	de	Davoudi	4
A	detenção	de	equipamentos	1

REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto

REMBRANDT

- powerfull evergrowing external knowledge base, Wikipedia
- http://xldb.fc.ul.pt/wiki/Rembrandt
- download and use it, but need Wikipedia dump
- Web API (soon!)
- used to tag CHAVE collection (`http://www.linguateca.pt/CHAVE/`)

The End

Thank you for your attention!

Hope you have enjoyed your meal!