Geographically Aware Web Text Mining

Simpósio Doutoral da Linguateca
4 de Outubro de 2006

Bruno Emanuel Martins

Orientador: Mário J. Silva

Motivation

- **Human information needs often relate to specific places**
- **Web information often contains a geographical context**
- **Current Web-IR ignores geographical semantics**

Clear need for Geo-IR technology

- **Multidisciplinary problem combining IR, GIS, NLP, ...**
- **Commercial systems like local.google and metacarta**
- **Many research questions still open**

Thesis Statement

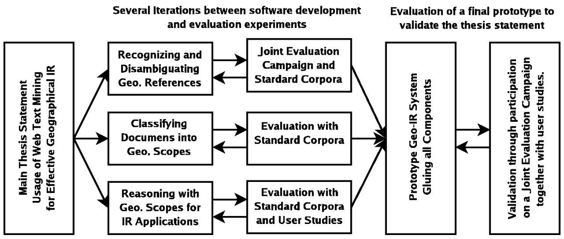
Text mining can be applied to extract geographic context information, leading to better information retrieval technology that outperforms standard approaches in geographically aware relevance.

Assumptions

- **Geo-IR problem can be decomposed in three sub-tasks**
  - •Recognizing and disambiguating Geographic Expressions
  - •Assigning documents to Geographic Scopes
  - •Building IR applications that account for Geographic Scopes
- **Geographic information is pervasive on the Web**
  - •Previous work in the SPIRIT project
  - •Work by Marcirio Chaves, Janet Kohler, Vivian Zhang et al, …
- **Docs and queries can be assigned to encompassing geo. scopes**
  - •One sense per discourse assumption from NLP

Validation Methodology

Experimental validation methodology

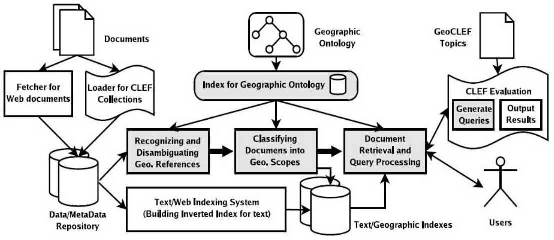

Several Iterations between software development and evaluation experiments
Evaluation of a final prototype to validate the thesis statement
Main Thesis Statement Usage of Web Text Mining for Effective Geographical IR
Recognizing and Disambiguating Geo. References
Joint Evaluation Campaign and Stardard Corpora
Classifying Documens into Geo. Scopes
Evaluation with Standard Corpora
Reasoning with Geo. Scopes for IR Applications
Evaluation with Standard Corpora and User Studies
Prototype Geo-IR System Gluing all Components
Validation through participation on a Joint Evaluation Campaign together with user studies.

Geo-IR System Components

- **Gazetteers and Geographic Ontologies**
- **Recognizer for Geographical References in Text**
- **Assigner of Geographic Scopes to the Documents**
- **Handler for Geographic Queries**
- **Geo-IR Systems using Document Scopes**

Prototype System

Software from tumba! + Specific Geo-IR components



Documents

Fetcher for Web documents

Loader for CLEF Collections

Data/MetaData Repository

Geographic Ontology

Index for Geographic Ontology

Recognizing and Disambiguating Geo. References

Classifying Documens into Geo. Scopes

Document Retrieval and Query Processing

Text/Web Indexing System (Building Inverted Index for text)

Text/Geographic Indexes

GeoCLEF Topics

CLEF Evaluation

Generate Queries

Output Results

Users

Gazetteers and Geographic Ontologies

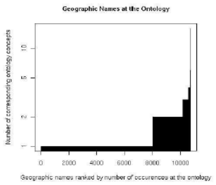
Important component of Geo-IR

- Reference status together with the test corpus
- Getty Thesaurus of Geographical Names (TGN)
  - About 1,000,000 places around the globe
  - Hierarchical
  - Spatial information in the form of coordinates and MBRs

Widely used resource!

Our Geographical Ontologies

OWL ontologies for PT and the world



Ontology statistic	Value
Ontology concepts	12,654
Geographic names	15,405
Unique geographic names	11,347
Concept relationships	24,570
Concept types	14
Part-of relationships	13,268
Adjacency relationships	11,302
Concepts with spatial coordinates	4,204 (33.2%)
Concepts with bounding boxes	2,083 (16.5%)
Concepts with demographics	8,206 (64.8%)
Concepts with corpus frequency	10,057 (79.5%)

http://xldb.di.fc.ul.pt/geonetpt/

Geo-IR System Components

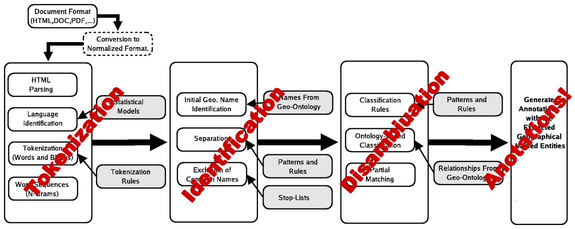
- Gazetteers and Geographic Ontologies
- Recognizer for Geographical References in Text
- Assigner of Geographic Scopes to the Documents
- Handler for Geographic Queries
- Geo-IR Systems using Document Scopes

Finding Geographic References in Text

- Named entity recognition (NER) is familiar within IE
  - Evaluation methodology, annotated corpora, ...
  - Existing results (e.g. importance of gazetteers)
  - We can build on previous NER efforts (e.g. extend annotations)
- Our problem is more complex
  - Disambiguating references with respect to their type
  - Grounding references to the ontology (or coordinates)
  - Web environment, address the Portuguese language, …
- Associated text-processing tasks
  - Language classification, tokenization, ...

Finding Geographic References in Text

4-Step Approach for Recognizing Geographic References



Document Format (HTML,DOC,PDF,...)

Conversion to Normalized Format.

HTML Parsing

Language Identification

Tokenization (Words and Blocks)

Word Sequences (N-Grams)

Statistical Models

Tokenization Rules

Initial Geo. Name Identification

Separation

Exclusion of Common Names

Names From Geo-Ontology

Patterns and Rules

Stop-Lists

Classification Rules

Ontology-Based Classification

Partial Matching

Patterns and Rules

Relationships From Geo-Ontology

Generate Annotations with the Extracted Geographical Named Entities

Tokenization

Identification

Disambiuation

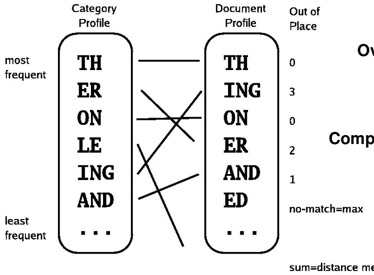
Anotations!

Step 1 : Shallow Processing

- **HTML Parsing**
  - Conversion of other file formats to HTML
  - Fault tolerant parser written by hand
- **Tokenization**
  - Tightly coupled with HTML parsing
  - Context-pairs table (context given by surrounding characters)
  - Words, sentences, n-grams
- **Language classification**
  - Character N-Grams used for classification

Language Classification

Similarity to N-gram profiles: $sim(x,y) = \frac{2*\sum_{t \in ng(x) \cap ng(y)} \log P(t)}{\sum_{t \in ng(x)} \log P(t) + \sum_{t \in ng(y)} \log P(t)}$



Over 90% performance on Web data

Comparable to state-of-the-art over newswire text

Problem: PT!=BR

Finding Geographic References in Text

Existing systems for handling place references

System	Classify	Ground	Evaluation Results
InfoXtract [24]	✓	✓	93.8% accuracy
Informedia DVL [35]	✓	✓	75% accuracy
Web-a-Where [2]	✓	✓	63.1-81.7% accuracy
Smith and Mann [44]	✓		21.82-87.38% accuracy
Schilder et al. [40]	✓	✓	74 % $f1$-score
KIM system [26]	✓		88.1% $f1$-score
Nissim et al. [34]	✓		$f1$-score around 75%
Leidner et al. [23]	✓	✓	-
Metacarta [37]	✓	✓	-

Newswire Corpus	Words	Entities	Precision	Recall
Portuguese (HAREM)	89,241	1,276	86.63%	87.22%
English (CoNLL-2003)	301,418	10,645	96.59%	95.65%
German (CoNLL-2003)	310,318	6,579	83.19%	72.90%
Spanish (CoNLL-2002)	380,923	6,981	85.76%	79.43%
Dutch (CoNLL-2002)	333,582	4,461	78.54%	80.67%

Corpora used in NER evaluation experiments

Finding Geographic References in Text

Our results in handling geo-references in text

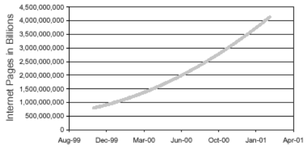
	Recognition			Grounding		
Corpus	Pre.	Rec.	F_1	Pre.	Rec.	F_1
Portuguese (HAREM)	89%	68%	77%	-	-	-
English (CoNLL-03)	85%	79%	81%	-	-	-
Spanish (CoNLL-02)	83%	76%	79%	-	-	-
Portuguese HTML	90%	76%	82%	89%	76%	81%
English HTML	91%	75%	82%	90%	73%	80%
German HTML	79%	72%	91%	77%	70%	73%
Spanish HTML	86%	75%	80%	83%	72%	77%

- **Rule-based approach for recognizing references in text**
  - names from ontology + context patterns + capitalization
- **Heuristics for disambiguating+grounding references**
  - e.g. one reference per discourse

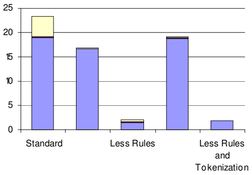
Computational Aspects

- **Simple algorithms and heuristics should be preferred**
- **Millions of documents on the Web**
- **Additional experiments currently underway**

Web growth [SearchEngineWatch]



NERC in different settings



Geo-IR System Components

- Gazetteers and Geographic Ontologies
- Recognizer for Geographical References in Text
- **Assigner of Geographic Scopes to the Documents**
- **Handler for Geographic Queries**
- **Geo-IR Systems using Document Scopes**

Assigning Geographic Scopes

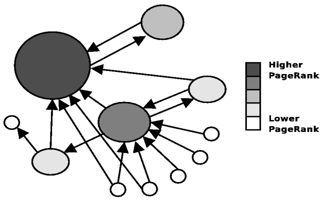
- **Hard document classification task**
  - Place references in text are very sparse and ambiguous
  - Need to explore relationships between place references
- **Previously reported results**
  - Web-a-Where system from Amitay et al.
    - 38% accuracy in finding correct "focus" of a Web page
    - Much better if we consider partial matches
  - Ding et al., Yamada et al., Gravano et al.
- **Existing corpora for evaluation**
  - Web pages from ODP under Top:Regional
  - Reuters collections (although only broad categories -- countries)

Assigning Geographic Scopes

We proposed a Graph-Ranking method

PageRank

Weighted Graph from Ontology



$$S(V_i) = (1-d)s_i + d * \sum_{V_j \varepsilon In(V_i)} \frac{w_{ij}}{\sum_{v_k \varepsilon Out(V_j)} w_{jk}} S(V_j)$$

Assigning Geographic Scopes

Results for our document geo-referencing approach on ODP pages

Multilingual global ontology : ODP `Top:Regional`		
	Measured Accuracy	
Granularity Level	Most Frequent Reference	Graph-Ranking
Continent	91%	92%
Country	76%	85%
Exact Scope Matches	67%	72%

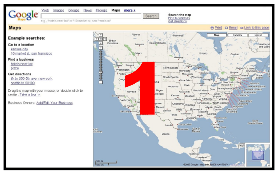
Portuguese ontology : ODP `Top:Regional:Europe:Portugal`		
	Measured Accuracy	
Granularity Level	Most Frequent Reference	Graph-Ranking
NUT 1	84%	86%
NUT 2	58%	65%
NUT 3	44%	59%
Municipalities	28%	31%
Exact Scope Matches	34%	53%

- Based on a graph ranking algorithm to select most "important" scope
  - References from text + Ontology + PageRank on weighted graph

Geo-IR System Components

- Gazetteers and Geographic Ontologies
- Recognizer for Geographical References in Text
- Assigner of Geographic Scopes to the Documents
- Handler for Geographic Queries
- **Geo-IR Systems using Document Scopes**

Query formulation in Geo-IR



1. Map interface
   - Spatial coordinates
2. Form interface
   - Multiple fields
3. Text input field
   - Single query string

Your web browser is not fully supported by Google Maps. More information »

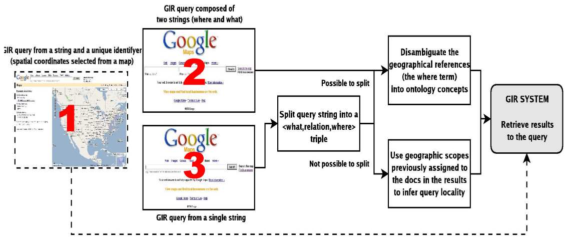
View maps and find local businesses on the web.

Your web browser is not fully supported by Google Maps. More information »

View maps and find local businesses on the web.

Processing geographical queries

- Queries are <what,relationship,where> triples
  - ***INPUT: "hotels in Seattle" or "hotels" + "in" + "Seattle"***
  - ***OUTPUT:*** **<hotels,IN,*Seattle*> + match *Seattle* to ontology concepts**


GIR query composed of two strings (where and what)
GIR query from a string and a unique identifyer (spatial coordinates selected from a map)
1
2
3
GIR query from a single string
Possible to split
Not possible to split
Split query string into a <what,relation,where> triple
Disambiguate the geographical references (the where term) into ontology concepts
Use geographic scopes previously assigned to the docs in the results to infer query locality
GIR SYSTEM
Retrieve results to the query

Results with CLEF topics

Dataset	Number of Queries	Correct Triples		Time per Query	
		ML	TGN	ML	TGN
GeoCLEF05 EN	25	19	20		
GeoCLEF05 PT	25	20	18	288.1 msec	334.5 msec
GeoCLEF06 EN	32	28	19		
GeoCLEF06 PT	25	23	11		
ImgCLEF06 EN	24	16	18		

•Most CLEF topics are adequately handled

•Over 80% accuracy with ML ontology

•Results with TGN were worst

•Comparable performance with commercial geocoders

Geo-IR System Components

- Gazetteers and Geographic Ontologies
- Recognizer for Geographical References in Text
- Assigner of Geographic Scopes to the Documents
- Handler for Geographic Queries
- Geo-IR Systems using Document Scopes

Geo-IR Systems Using Scopes

- **IR making use of the geo-scopes for the documents**
- **Combination of thematic and geographic relevance**
  - **How to define, compute and evaluate geographic relevance?**
- **Methodology from TREC and CLEF (GeoCLEF2005-2006)**
  - **Standard collection, queries, relevance judgments**
  - **Test functionalities that are not available on standard systems**
- **Compare text mining (i.e. scopes) approach with:**
  - **Standard IR approach**
  - **Query expansion using the geographical ontology**
- **Integration with the Tumba! Web search engine**

Geo-IR Relevance

- **Relevance=Textual Relevance + Geographic Relevance**
- **Textual Relevance=State-of-the-art IR**

•Okapi BM25 ranking formula, using extension for weighted fields

•Query expansion through blind feedback

- **Geographic Relevance=Set of heuristics**

•Spatial proximity (normalized according to the area of the query)

•Ontological relatedness (Lin's similarity measure)

•Shared population (approximation for the area of overlap)

•Spatial adjacency

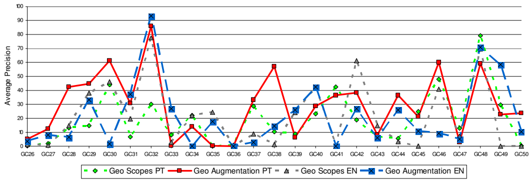
Geo-CLEF 2006 Results

- **Both Geo-IR approaches are better than standard IR**
- **Geo. Query expansion performed better than text mining… why?**
- **Problems when assigning scopes (particularly for PT)**

Measure	Run 1		Run 2		Run 3		Run 4	
	PT	EN	PT	EN	PT	EN	PT	EN
num-q	25	25	25	25	25	25	25	25
num-ret	5232	3324	23350	22483	22617	21228	10483	10652
num-rel	1060	378	1060	378	1060	378	1060	378
num-rel-ret	607	192	828	300	519	240	624	260
map	0,301	0,303	0,257	0,158	0,193	0,208	0,293	0,215
R-prec	0,359	0,336	0,281	0,153	0,239	0,215	0,346	0,220
bpref	0,321	0,314	0,254	0,140	0,208	0,191	0,306	0,199
gm-ap	0,203	0,065	0,110	0,027	0,074	0,024	0,121	0,047
ircl-prn.0.50	0,347	0,304	0,256	0,162	0,163	0,221	0,305	0,215
ircl-prn.1.00	0,002	0,116	0,012	0,056	0,000	0,025	0,003	0,094
P5	0,488	0,384	0,416	0,208	0,432	0,240	0,536	0,288
P10	0,496	0,296	0,392	0,180	0,372	0,228	0,480	0,240
P15	0,472	0,243	0,360	0,171	0,341	0,195	0,440	0,224
P20	0,442	0,224	0,350	0,156	0,318	0,170	0,424	0,212

Results for individual queries

- **Geo. query expansion is better for most queries**
- **Are some queries more "geographical" than others?**
- **Still analysing the results**



Conclusions

- **Geo-IR techniques achieve improvements over baseline**
- **One scope per document seems to be to restrictive**
  - •Ongoing experiments to test with multiple scopes
  - •Scalability issues in computing relevance
- **No definitive conclusion on if text mining is a good approach for Geo-IR**
  - •Set parameters differently for each query?
  - •Just use query expansion?

Future of Geo-IR

- **User interface aspects**
  - •Deep integration with mapping functionalities
  - •Collaborative annotation of documents (e.g. del.icio.us)
  - •Clustered and faceted interfaces (explore different dimensions in data)
- **Improving performance and scalability**
  - •OK for GeoCLEF collections but how about the Web?
- **Other types of documents (e.g. pictures) and other kinds of tasks (e.g. question answering)**
- **Continuing with evaluation forums like GeoCLEF**
  - •Also addressing the subtasks (e.g. NER) and related tasks

tumba!

Pesquisa na Web Portuguesa

Thanks for your attention

bmartins@xldb.di.fc.ul.pt