Geographic Signatures for Semantic Retrieval

David S Batista, Mário J Silva, Francisco M Couto, Bibek Behera

University of Lisbon, Faculty of Sciences, LaSIGE

6th Workshop on Geographic Information Retrieval, Zürich
18th-19th February 2010

Outline

1. Motivation
   - Why Geographic Signatures?
2. Related Work
   - Geographic Named Entities Recognition
   - Geographic Information Representation
   - Semantic Similarities
3. Results
   - Implementation
   - Results
4. Conclusions and Future Work

Motivation

Previous Work

Capture geographic semantics as a single geographic scope:

- Text $\longrightarrow$ Geographic References
- Geographic References $\longrightarrow$ Geographic Concepts
- Geographic Concepts $\longrightarrow$ Encompassing Concept (Scope)

Geographic Signatures

- Insted of one single scope
- List of maximally disambiguated geographic references
- Coordinates, bounding boxes, populations counts

How are the signatures generated?

- Geo-parsing
  - manually coded rules: too restrictive, very specific.
  - machine learning:
    - extract features from text (sorrounding words, words properties)
    - use features to infer rules (probabilistically)
- Geo-coding
  - Need an external knowledge base (ontologies, gazetteers, encyclopedias)
  - Ambiguity

How are the signatures generated?

- Geo-parsing
  - manually coded rules: too restrictive, very specific.
  - machine learning:
    - extract features from text (sorrounding words, words properties)
    - use features to infer rules (probabilistically)
- Geo-coding
  - Need an external knowledge base (ontologies, gazetteers, encyclopedias)
  - Ambiguity

Related Work

Conditional Random Fields (CRF)

- 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
- Achieved very good results in gene and protein recognition

Geo-Net-PT02: Geographic Ontology of Portugal

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
Nº 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 Nº Features	Nº 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

Semantic Similarity Measures

Semantic Similarities

- Given two ontology concepts, return a numerical value, reflecting the closeness in meaning between them.
- Can be applyed to DAG structures using the *Information Content* (IC).
- $IC(c) = -\log p(c)$.
- $p(c)$ is the probability of occurrence of c in a specific corpus.
- IC is cumulative, that is, the IC of a concept c depends on its descendants in its subtree.

Semantic Similarities: Example

- "I was born close to **Santa Catarina** in **Lisboa**"
  - Lisboa, municipality (ID_1)
  - Lisboa, small locality in the municipality of Monção (ID_2)
  - Santa Catarina, civil parish in the municipality of Lisboa (ID_3)
  - Santa Catarina, small locality in the municipality in Caldas da Rainha (ID_4)

$$SSM\left(ID_1, ID_3\right) = 0.584$$
$$SSM\left(ID_1, ID_4\right) = 0.065$$
$$SSM\left(ID_2, ID_3\right) = 0.063$$
$$SSM\left(ID_2, ID_4\right) = 0.041$$

Implementation and Results

Trainning of the CRF model: corpus + features

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

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

- Additional features:
  - *charTypePattern.9+* token is composed by numbers only;
  - *charTypePattern.X+x+* token is capitalized;
  - *eq.lc.avenida* the value of token itself;
  - *isFeatureType* Geo-Net-PT02 feature types;
  - *isGeoName* districts, municipalities and civil parishes;
  - *isLocalPrefix* list of verbs and adjectives close to geographic references;
  - *isPreposition* a list of prepositions;

Trainning of the CRF model: Results

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

- Test on the GC of HAREM's 2008 event
- Recall is low, overfitting?
- Size of trainning corpus: 1 243 Kb

Semantic Similarity Measures in Geo-Net-PT02

- Calculated the *IC* for each concept in Geo-Net-PT02
- Ocurrences of concept's name in Google N-Grams corpus
- Use as SSM function *Jiang and Conrath (1997)*

Disambiguation algorithm

Pairwise desambiguation following the order of extraction:

*"...he went through **Avenida da República** to **Marquês de Pombal**, there he took the subway to **Rossio** ..."*

- X = concepts for "Avenida da República"
- Y = concepts for "Marquês de Pombal"
- Z = concepts for "Rossio"

- $SSM((\forall\ x \in X), (\forall\ y \in Y))$, select the pair of concepts (x,y) that gives the best similiraty.
- $SSM(y, (\forall\ z \in Z))$, select the z that gives the best similartiy.

Evaluation

- Manually annotated Wikipedia articles for the 18 districts of Portugal
- Extraction using the generated CRF Model
  - Precison: 0.69
  - Recall: 0.47
  - F-1: 0.56
- Disambiguation using the described pairwise algorithm

Evaluation: Geographic Entities Extraction

Page of	Entities	Precision	Recall	F-1
Aveiro	22	0,80	0,58	0,67
Beja	24	0,69	0,37	0,48
Braga	190	0,37	0,51	0,43
Bragança	11	0,56	0,39	0,46
Castelo Branco	23	0,71	0,46	0,56
Coimbra	85	0,52	0,38	0,44
Évora	11	0,90	0,37	0,52
Faro	58	0,68	0,53	0,60
Guarda	46	0,60	0,48	0,53
Leiria	98	0,70	0,44	0,54
Lisboa	225	0,66	0,50	0,57
Portalegre	79	0,41	0,56	0,48
Porto	101	0,40	0,53	0,45
Santarém	22	0,83	0,42	0,55
Setúbal	38	0,73	0,53	0,62
Viana do Castelo	12	0,84	0,48	0,62
Vila Real	51	0,52	0,62	0,57
Viseu	80	0,46	0,59	0,52

Evaluation: Desambiguation

Page of	Correctly Extracted	Correctly Disambiguated
Aveiro	100%	70%
Beja	88%	87%
Braga	71%	67%
Bragança	100%	75%
Castelo Branco	38%	54%
Coimbra	70%	82%
Évora	100%	100%
Faro	80%	68%
Guarda	93%	76%
Leiria	90%	85%
Lisboa	96%	92%
Portalegre	90%	68%
Porto	87%	68%
Santarém	100%	81%
Setúbal	81%	70%
Viana do Castelo	100%	62%
Vila Real	77%	83%
Viseu	92%	89%

Conclusions and Future Work

Conclusions

- Initial version of a geographic signatures prototype:
- Extraction
  - Recall for trained CRF model is still relatively low
  - Tunning of selected the features for trainning might increase results
  - BIG limitation: lack of large Portuguese labelled corpus for CRF trainning
- Disambiguation
  - IC generation: *Lisboa* in a given corpus can represent the city of Lisbon or just a street
  - Frequency of a concept in the web may cause inconsistency in IC estimation

Conclusions

- Alternatively, calculate the $p(c)$ by measuring the geographical content described by a concept
- $geospace(c) = \bigcup_{d \leq c} geospace(d)$ where $d \leq c$
- Calculate the value of a spatial or social feature for a given *geospace*: area, specificity, population
- $p(c) = \sum_{i=1} \lambda_i \frac{f_i(geospace(c))}{f_i(geospace(root))}$
- More complex disambiguation: comparing names in a setence vicinity of a concept

Conclusions

- Generate geographic signatures for WPT05, crawl of the "portuguese" web
- Evalute the effectiveness of geographic sigantures in GIR

The End
Questions?