Workshop on Geographic Information Retrieval, SIGIR 2006

The place of place in geographical IR

Diana Santos
Marcirio Silveira Chaves

Linguateca - www.linguateca.pt

Seattle, August 10th

Structure of the presentation

- Purpose of this paper
- Discussion of language matters concerning place
- Present work on the measurement of the above issues
  - Preliminary work done
  - Further goals

Purpose of the paper

- argue for natural language knowledge to inform GIR applications
  - challenging assumptions routinely made about geographical information in texts
  - discussing the notion of geographical ontology
  - presenting several problems with the current modelling of geographical information
- provide preliminary data that measures or assesses the several points described, implicitly providing support for those claims

What is natural language (processing)?

- Natural language is the oldest and most successful knowledge representation language
- Used for comunication, negotiation, and reason (->logic)
- Main features
  - vagueness
  - context-dependent
  - implicit knowledge
  - evolves/dynamic/creative
- Different natural languages
  - different world view
  - different glue/implicit

What is NL processing?

- Using computers to do things with natural language
- To be useful for humans
- Most intelligent human tasks involve language
  - as center (communicating, teaching, converting)
  - as periphery (mathematics papers, medical diagnosis)
- Daily tasks
  - writing (and creating or conveying information or affection)
  - reading (and finding information)
  - translating (and mediating)
  - teaching and learning and documenting
- Enormous political impact

GIR is NLP where place is important

- GIR is mainly retrieving of texts giving special attention to location
- The semantics of place is
  - therefore important for GIR
  - often provided as a map
- What are places? how are they called? how are they related? what reasoning they allow? is what people are after when formalizing the geographical part of IR
  - formalization in NLP is often
    - done resorting to ontologies -> *geo-ontologies*
    - actualized by information extraction mechanisms (such as NER, ontology population) -> *extracting geo info from texts*
    - made by assigning "semantic/extralinguistic" values to linguistic expressions -> *assigning coordinates to names*

Start: What is known about (place in) NL?

- Why reinvent the wheel when all these subjects have already been looked at?

OR

- If one starts where the others stopped progress in the field should be quicker!

- GIR is just another instantiation of the eternal problems of NL understanding, why proceeed as if it were a new discipline?

OR

- why proceed ignoring how NL works?

Place in NL is ...

- dependent on language and culture and time and politics
  - Malvinas or Falkland
  - Spain or Iberian Peninsula
  - Gaza strip: Israel or Palestina
  - UK or England
- vague
  - in Lisbon, in Portugal, in Europe
  - *near* depends on its argument(s)
  - used metonymically to refer to a people, a group of representatives of a state, its government, etc.

Place in NL is ... (2)

- context-dependent
  - of the style/genre of the text
  - of the purpose of the text
  - of the intended readers of the text
  - of the purpose of the reader/information seeker
  - of whether the full location has already been mentioned

  Reguengos de Monsaraz *é ... Quando estive em* ***Reguengos****, ...*
- ambiguous
  - the same place name can denote different places
  - the same name can denote places and non-places (*Porto, Faro*)

A geo-ontology in Web IR

- Formalizes knowledge about geography that is valid in a large collection of texts... OK, but texts have
  - different purposes
  - different readers
  - different languages and cultures
- Are we talking about a shared consensus ???
- Or are we attempting an information structure that encompasses, in its generality, all the inconsistencies and differences recognized above?
- Or, we are in fact only interested in a extremely tiny bit of place information, namely the one related with physical navigation for consumers? (*restaurants in Switzerland, shoes in Genebra*)

Let us try to measure these phenomena

1. Given an administrative geo-ontology
   - how often it is ambiguous (geo/geo and geo/non-geo)
   - how often it is present in NL texts
2. Given a set of place names found in a sizeable text collection
   - how often they are present in an administrative geo-ontology
   - how often they refer to places
   - what kinds of texts include locations
3. Given an NE annotated collection for evaluation purposes
   - how often place names are used metonymically
   - overlap with the geo-ontology
4. Given a set of geographical topics
   - how often place depends on other issues

Let us try to measure these phenomena (2)

5. Given a set of user logs of a Web search engine
   - which are place names and why they were used
   - its overlap with a geo-ontology
6. Given a set of annotated texts
   - study the distribution of place names
   - study the correlation with kind of text
   - study the co-occurrence with other place names and with other names

- Work in progress on Portuguese... in several directions, some of these will be presented here.

First results: projection of Geo-Net on the Web

- flatten Geo-Net (extract all name-class pairs: *Lisboa-província)*
- compute their overlap (total and partial) **(Table 1)**
  - Total: *Lisboa* (city) and *Lisboa* (province); *Castelo Branco* (district and city)
  - Partial: *Monsaraz* and *Reguengos de Monsaraz*; *Castelo* and *Castelo Branco*
- study their occurrence in WPT03 – BACO (a snapshot of the Portuguese Web in 2003 as indexed by *tumba!*) **(Chaves & Santos 2005)**
  - using frequency lists (with capitalization distinctions) for one-word names
  - using n-gram searches for MW-names
  - checking co-occurrence with non-capitalized words

First results: study of locations in Web text

- Using SIEMÊS to NE-annotate a subset of WPT03 **(Chaves & Santos05)**
  - presence of the locations in Geo-NET-PT-01: only 10% of types were there
  - presence of the locations in GKB-ML (also covering the Portuguese names of cities outside Portugal): only 3.18% of the types were there
  - what kinds of locations appear on the Web
  - how often are location names included in PEOPLE or ORGANIZATION names already automatically disambiguated in context :
    - 31% of the types of PEOPLE names contain a word in Geo-Net-PT01;
    - 23% of the types of ORGANIZATION names contain a word in Geo-Net-PT01

First results: study of locations in text

- study their occurrence in HAREM golden colection **(Table 3)**
  - presence of locations in
    - Geo-NET-PT01: 30% of types were there
    - GKB-ML: 25% of types were there
  - what kind of locations appear in text
    - 76% ADMINISTRATIVE, 11% LATU SENSU
    - 6.6% PHYSICAL, 4.3% VIRTUAL, 0.01% ADDRESS
  - how often are location names included in PEOPLE or ORGANIZATION names already manually disambiguated in context:
    - 1.85% of the types of PEOPLE names contain a word in Geo-Net-PT01
    - 3.5% of the types of ORGANIZATION names contain a word in Geo-Net-PT01

First results: theoretical analysis of GeoCLEF topics and their extensions

1. non geographic subject restricted to a place (*music festivals in Germany*) GeoCLEF 2005 and GeoCLEF 2006
2. geographic subject with non-geographic restriction (*rivers with vineyards*) GeoCLEF 2006
3. geographic subject restricted to a place (*cities in Germany*) GeoCLEF 2006
4. non-geographic subject associated to a place (independence, concern, economic handlings to favour/harm that region, etc.) (*independence of Quebec, love of Peru*)

First results: theoretical analysis of GeoCLEF topics and their extensions (cont.)

5. non-geographic subject that is a complex function of place (place is a function of topic) (*European football cup matches, winners of Eurovision Song Contest*)
6. geographical relations among places (*how is the Himalayan related to Nepal*?)
7. geographical relations among events (*Did Waterloo occur more north than the battle of X? Were the findings of Lucy more to the south than those of the Cromagnon in Spain*?)
8. relations between events which require their precise localization (*was it the same river that flooded last year and in which killings occurred in the XVth century*?)

First results: use of place names in (con)text, in the HAREM NER annotated golden collection

- *Portugal*
- *... is a big consumer of football shows* (its population)
- *... sent peace-keeping troops to East Timor* (its government)
- *... has emmigrants all over the world* (socio-ideological-legal being)
- *... lost the match against Germany* (its football team)

VERSUS really PLACE

- *I was born in Portugal*
- *Portugal has the best beaches of the world*
- *These prehistoric relics were found in Portugal*

Concluding remarks

- Many of the properties of NL have been recognized or rediscovered in GIR before, we do not claim to be the first to point them!
- Still, it seems relevant to show that **they derive naturally from the way NL is** and are not at all specific to GIR
- The measurement of many of these issues still remains to be done. This paper simply argues for its need, presents some preliminary results for Portuguese, and hopes that studies for other languages will soon appear.

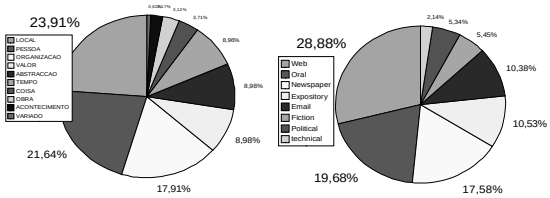
Background on the resources used

if there are related questions...

The golden collection of HAREM (PT)

- 257 (127) documents, with 155,291 (68,386) words and 9,128 (4,101) manually annotated NEs
- Places: 2157 (980) tokens, 979 (462) types

NE category distribution on the GCs Text Genre Distribution on the GCs



GeoNet e GKB-ML

Statistic	Geo-Net-PT01	GKB-ML
# of features	**418,065**	**12,293**
# of relationships	**419,867**	**12,258**
# of part-of relationships	**418,340 (99.83%)**	**12,245 (99,89%)**
# of equivalence relationships	395 (0.09%)	2,501(20,40%)
# of adjacency relationships	1,132 (0.27%)	13 (0.10%)
Avg. broader features per feature	1.0016	1.07
Avg. narrower features per feature	10.56	475.44
Avg. equivalent features per feature with equivalent	1.99	3.82
Avg. adjacent features per feature with adjacent	3.54	6.5
# of features without ancestors	3 (0.00%)	1(0.00%)
# of features without descendants	374,349 (89.54%)	12,045 (97,98%)
# of features without equivalent	417,867 (99.95%)	11,819 (96,14%)
# of features without adjacent	417,739 (99.92%)	12,291 (99,99%)

WPT 03 and BACO

- WPT03 a snapshot of the Portuguese Web in May-June 2003 as indexed by *tumba!*
- 3,6 million documents (3,4 in HTML)
- 854,936,550 tokens, 4,243,875 types
- 1,150,000 log entries (searches for six months in 2003)
- BACO: conversion to mySQL format of the text after duplicate removal

SIEMÊS

- NER for Portuguese (Sarmento 2006, Sarmento et al. 2006)
- Based on a large gazetteer, REPENTINO (450,000 instances, ~150,000 locations)
- and on similarity rules
- 64.09% precision and 69.83% recall in HAREM for LOCAL
- 61.3% precision and 56.7% recall in miniHAREM