The place of place in geographical IR

Diana Santos
Linguateca, SINTEF ICT
Pb 124, N-0314 Oslo, Norway
Diana.Santos@sintef.no

Marcirio Silveira Chaves
Linguateca, XLDB
University of Lisboa, Faculty of Sciences
1749-016 Lisboa, Portugal
mchaves@di.fc.ul.pt

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Design, Measurement, Experimentation

1. INTRODUCTION

Let us define geographical IR (GIR) as the activity whose purpose is to retrieve information in a geographically-aware way. In other words, considering the geographical dimension as special. GIR presupposes two things:

- the possibility to associate to (possibly retrieve from) the collection geographical information
- the existence (or the possibility of creation) of semantic repositories that allow geographical reasoning, henceforth called geo-ontologies.

The most common kind of collections for GIR so far are the Web and other document collections, which are mainly textual. This paper is concerned with the non-trivial relationship between reference to place in natural language (NL) and common GIR assumptions. There are two main ways in which NL texts and GIR meet: in the attempt to derive or populate geo-ontologies from text itself, and in the attempt to label Web pages with what is called geo-scopes, deriving these from clues in the pages themselves.

We will survey briefly the two, noting in passing that both approaches are bottom-up in the sense that they look at the texts, but the second makes use of a prior information source, a geo-ontology, which is typically top-down (see GeoNet-PT01 in Table 1).

1.1 Ontology extraction from texts

Automatic ontology extraction (AOE) from texts attempts, given a collection, to arrive at a partial structure of the knowledge in these texts. It seems that, given the way AOE works, one should apply these techniques to texts dealing with geography, as Bilhaut et. al. do [4].

However, GIR is most often than not concerned with other subjects, especially on the Web: most papers deal with services, shopping, tourism or news.

A first remark is that the kind of *location* presupposed by these kinds of Webpages or documents is not necessarily the one that is implicitly structured in geographical texts. If this intuition proves true, there are two ways out of this dilemma:

- create different "ontologies" for each kind of text; or
- pick whatever ontology you want, and project it on your texts.

1.2 Assigning scopes for geo-indexation

Geo-scoping, or grounding, is another task, concerned with indexing Web resources, not with the locations they are *about*, but with the locations they are *in*. To see the difference, consider a page about Switzerland in a Mexican site, or a page about shoe sales in (a shop in) Switzerland.

To retrieve information about Switzerland, a user would like to retrieve the first page, which should be marked as being about Switzerland. But to get information about places where to buy cheap shoes, a geographic context-aware search engine should have marked the second page as referring to a location in Switzerland (most probably a city or even a city quarter), in order to satisfy a user located in Switzerland.

It is not clear that these two different tasks have been sufficiently discriminated in the GIR literature. Often they are presented as tasks in sequence [14], or at least closely related tasks [2, 6].

The second task is, however, separate from the first, and the kind of reasoning required is different in kind, as well as the resources required: for the second task, the geo-ontology required is clearly map-based, in the sense that it can be conceived as a grid and, depending on the specification of the user, reasoning can be done to order hits according to geographical proximity. Note that in this case it does not make sense that all pages are indexed according to this kind of information structure.

The first task, on the other hand, may be conceived as a topic distillation task about a geographically defined subject.

Before we discuss in more detail the consequences of this dichotomy for GIR and its evaluation, let us look at place in natural language.

GIR'06, August 10, 2006, Seattle, Washington.

Table 1: Quantitative description of Geo-Net-PT01 for 11 location types: MW stands for multi-word, NUT for "Nomenclature of Territorial Units for statistical purposes", T(otal) for perfect and P for partial matching

Location type	# of distinct terms	number of words in terms					# of MW terms	Total ambiguity	1 gram ambiguity	
		1	2	3	4	$\sum$ >4			T	P
NUT1	3	1	0	0	2	0	2	3	0	0
NUT2	7	5	0	0	2	0	2	7	5	0
NUT3	30	8	11	8	3	0	22	6	2	4
regiao	2	0	1	0	1	0	2	0	-	-
provincia	11	4	6	0	1	0	7	5	2	1
distrito	18	15	2	1	0	0	3	18	15	0
concelho	323	203	27	68	22	3	121	301	193	1
ilha	11	0	1	6	4	0	11	1	-	-
freguesia	3,597	2,133	336	764	287	77	1,462	2,799	1884	51
localidade	26,924	10,851	4,098	9,661	1,783	531	16,073	3.655	2388	607
zona	3,593	1,201	540	1,233	456	163	2,392	1,241	804	55
Total	34,519	14,421	5,022	11,741	2,561	774	-	-	-	-

2. THE ROLE OF PLACE IN LANGUAGE

A culture- and language-dependent concept An influential paradigm in linguistics has claimed that conceptualization of place is basic in the way human language is structured [11]. Nonetheless (or maybe consequently), place is one of the most culture- and language-dependent concepts there is, as substantiated by a large amount of contrastive (and translation) data [23, 24, 16].

Vagueness A key concept to understand NL is vagueness [15, 17]. In a geographical context, this means that speakers are very rarely precise enough to distinguish between a city and its downtown, between a city and a state the city is capital of, or between a different time-slice of the city and the present one. Likewise, no matter the language, *near* is always dependent on its argument – in addition to several other context factors. Furthermore, it is generally acknowledged in Named entity Recognition (NER) that geo-political "places" (country names, capital cities, city names, etc.) are used frequently to denote organizations, or a group of people (that lives there, or that governs those places), or even a more abstract reality. (This is a different issue from the often mentioned geo/non-geo ambiguity [2].)

Context-dependence Context (or co-text) ranges from the time and *place* in which the utterance was created, to all sorts of political, cultural, sociological and interpersonal details. In fact, there is always a lot of knowledge shared by speakers/writers and addressees, that prevents saying the obvious (and shared) and *adds* knowledge (hence only what is peripheral gets mentioned or defined). See [5] for a similar point. Associations to place are very context dependent: Greece and Hungary are members of the *North Atlantic* Treaty Organization; Israel plays in the *European* football cup. This casts doubt on the suggestion of separating thematic from geographical relevance, as suggested by [6, 14]. Fonseca et. al. [9] describe other cases involving different categorizations of the "same" object (bodies of water), distinguishing between vertical and horizontal navigation in geo-ontologies.

Borders are not natural Geography is a science developed to make war (as in the name of Lacoste's famous book[1]): so, limits to states and regions, such as borders, are not naturally or neutrally determined and may continue to differ. Also, they are very much time-dependent. According to Leidner, the information in a large world-wide geo-ontology suffers 20,000 changes a month [12].

Location is dependent on the style of the text, in terms of what has been called genre, or user need [1]. Back to our example of Web pages, selling shoes or providing an overview about a country is different, and so is the kind of geographic references made and the "ontology" presupposed.

Every concept is a function of all other concepts A simple analysis of GeoCLEF topics (plus their logical extensions) shows many ways in which geographically-aware systems could be called to play a role. As hinted in Santos [17], in the limiting case every concept is a function of – or related to – all other concepts. It should therefore be noted that the model of Bucher et al. [6], also employed in GeoCLEF 2005 [10], namely (theme, relationship, location) is a too simplified model for GIR.[2]

Context explicit its use as a location The less prominent a place name is (for example because it is homonymous with another common word, or with a more well-known place) the more probable it is that the context explicitly makes clear its use as a location. On the contrary, a clear place name won't be preceded or followed by its kind. So "similar" place names may be employed differently in natural language, depicting the full range of internal vs. external evidence determined NE's. In other words, and as usual in natural language, Gricean considerations play a prominent role in the expression of place [3].

[1] *La géographie ça sert d'abord à faire la guerre*, 1976.

[2] In pseudo-logical terms, a concept is a function of time and location. E.g. concept(time,location). time(concept, location). location(concept, time).

Considering the user needs Finally, although it is possible to agree on the meaning of whatever complex place denoting expression, as was the case with *former Eastern bloc countries* in topic 35 of GeoCLEF 2006, it ultimately depends on the purpose of the information seeker which region is most relevant. So it is worth defining real user needs when evaluating GIR.

Summing up, there are many differences between a map-centered view of GIR and the way place is conveyed in natural language. Probably the most general conclusion is that they are complementary, and that considerably more study should be devoted to how to integrate both views.

3. SOME EMPIRICAL DATA

We have done some measures to assess the quantitative import of some of the points discussed above, using:

WPT 03 The first publically available snapshot of the Portuguese Web, based on a crawl by tumba! search engine (`tumba.pt`) in 2003 and available as a mySQL database [21].

Geo-Net-PT01 A large geo-ontology of administrative locations in Portugal, by integrating several authoritative sources [8].

SIEMÊS A broad-coverage NER system for Portuguese, SIEMÊS [19].

HAREM Golden Collection (GC) A manually revised and annotated NER collection for Portuguese, deployed for the HAREM evaluation contest [18].

Table 1 measures the overlap between a map-based geo-ontology (Geo-Net-PT01) and the texts we wanted to ground. The location types, in the first column, represent the administrative division of Portugal. These are culture-dependent, since the granularity and nomenclature often vary when dealing with different countries.

We also extracted some statistics about geographical ambiguity in Geo-Net-PT01 (Table 2), and about reference to geographical concepts in Portuguese texts, using the NE-annotated GC from HAREM (excluding the Brazilian texts, since Geo-Net-PT01 only contains data about Portugal). This collection contains 127 documents and 68,336 words.

Table 2: Distribution and ambiguity of the terms in Geo-Net-PT01 by size in words

Term size	Distinct	Ambiguous (%)	Token ambiguous (%)
One	11,561	2,433 (21.04)	5,293 (45.78)
Two	4,569	381 (8.34)	834 (18.25)
Three	10,984	705 (6.42)	1,462 (13.31)
Four	2,351	194 (8.25)	404 (17.18)
Five	589	20 (3.39)	42 (7.13)
Six	109	0	0
Seven	42	0	0
Eight	6	0	0
Nine	6	0	0
Total	30,217	3,733 (12.35)	8,035 (26.59)

Table 3 presents the match of the LOCAL category in Geo-Net-PT01 and GKB-ML - a multilingual ontology with 15,005 names in four languages (751 (4.87%) in Portuguese): only ca. 33% (22%) of the administrative place names appear in these geo-ontologies. Although neither postal addresses (correio) nor virtual locations (such as TV or newspapers) would ever be present in Geo-Net-PT01, it is relevant to note that they are seen as locations in natural language.

Note that SIEMÊS has been developed independently of WPT 03 and tumba!. In fact, the lexical knowledge it employs is encoded in the REPENTINO gazetteer [20], most of which was created from the Web in Portuguese (all Web, not specifically the one indexed by tumba!). Note also that the categories of SIEMÊS, in addition to be shaped by HAREM guidelines, were based on practical considerations of which names were possible to amass in large quantities. Neither of these considerations relate in any way to the specific problem of assigning scopes to Web resources. Rather, both HAREM and SIEMÊS belong to the information extraction tradition in NLP.

In the full paper, we present briefly the procedure followed and the (preliminary) results, discussed in Chaves and Santos [7]:

- How many documents (on a random subset of the Portuguese Web, 32,000 documents) mention geographical locations by name[3] at all? 24,468 (76.46%).
- Whenever locations are mentioned, how many and how repeatedly? On average, 11.31 locations and 7.34 distinct per document with location.
- What kind of locations are we talking about? Most of them (70%) correspond to city, town or village names, followed by full address (7.31%), socio-cultural places (7.24%) and countries (4.14%).
- How often are these locations present in Geo-Net-PT01? Only 10% of the types can be found in Geo-Net-PT01.
- What about the ambiguity with people's and organization's names? We found that 31.21% of the person distinct named entities (NEs) and 23.43% of the organization distinct NEs contain a geographic name included in Geo-Net-PT-01.

4. FINAL REMARKS

We presented the distinct roles of place names in NL. These roles include culture- and language-dependency, vagueness and context-dependence among others. These different ways place is conveyed in natural language can complement the map-centered view of GIR.

Further works include deepening the comparison between the places in geo-ontologies built from authoritative data sources and places mentioned in NL texts, as well as the building of geo-ontologies from these texts.

Acknowledgments

Linguateca is financed by grant POSI/PLP/43931/2001 from the Portuguese Fundação para a Ciência e Tecnologia, co-financed by POSI.

5. REFERENCES

[1] R. Aires, S. Aluísio, and D. Santos. User-aware page classification in a search engine. In *Stylistic Analysis Of Text For Information Access, SIGIR'05 Workshop*, Salvador, Bahia, Brazil, 19 August 2005.

[3] As recognized by SIEMÊS, whose precision and recall in HAREM Web texts is respectively 70% and 75%.

Table 3: Location types in the PT part of the manually NE-annotated HAREM collection

Type HAREM GC	# Tokens	# Distinct (%)	Geo-Net-PT01		GKB-ML	
			# Tokens (%)	# Distinct (%)	# Tokens (%)	# Distinct (%)
Administrativo	754	275 (36.47)	248 (32.89)	105 (38.18)	169 (22.41)	108 (39.27)
Alargado	110	97 (88.18)	34 (30.91)	18 (18.56)	9 (8.18)	3 (3.09)
Correio	9	9 (100)	–	–	–	–
Geográfico	65	52 (80)	30 (46.15)	16 (30.77)	18 (27.69)	8 (15.38)
Virtual	42	29 (69.05)	–	–	–	–
Total	980	462 (47.14)	312 (31.84)	139 (30.09)	196 (20)	119 (25.76)

[2] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: Geotagging web content. In *ACM SIGIR'04*, 2004.

[3] M. Aurnague. A unified processing of orientation for internal and external localization. In M. Aurnague, A. Borillo, M. Borillo, and M. Bras, editors, *4th European Workshop on Semantics of Time, Space, and Movement and Spatio-Temporal Reasoning* (Toulouse, France, 4-8 September, 1992), Groupe "Langue, Raisonnement, Calcul", pages 39-52.

[4] F. Bilhaut, T. Charnois, P. Enjalbert, and Y. Mathet. Geographic reference analysis for geographic document querying. In *Workshop on the Analysis of Geographic References - NAACL-HLT'03*, (Edmonton, Alberta, Canada, 31 May 2003), page s/p.

[5] C. Brewster, F. Ciravegna, and Y. Wilks. Background and Foreground Knowledge in Dynamic Ontology Construction: Viewing Text as Knowledge Maintenance. In Y. Ding and K. van Rijsbergen and I. Ounis and J. Jose, editors, *Semantic Web, Workshop held the 26th Annual International ACM SIGIR Conference* (Toronto, Canada, July 28-August 1, 2003).

[6] B. Bucher, P. Clough, H. Joho, R. Purves, and A. K. Syed. Geographic IR systems: Requirements and evaluation. In *22nd International Cartographic Conference ICC'05* (A Coruña, Spain, 11-16 July 2005). CD-ROM.

[7] M. S. Chaves and D. Santos. What kinds of geographical information are there in the Portuguese Web? In R. Vieira, P. Quaresma, M. das Graças Volpes Nunes, N. Mamede, C. Oliveira, and M. C. Dias (editors), *7th Workshop on Computational Processing of Written and Spoken Language - PROPOR'06* (Itatiaia, RJ, Brazil, 13-17 May 2006), Springer, pages 264-267.

[8] M. S. Chaves, M. J. Silva, and B. Martins. A Geographic Knowledge Base for Semantic Web Applications. In C. A. Heuser, editor, *20th Brazilian Symposium on Databases* (Uberlândia, Minas Gerais, Brazil, 3-7 October 2005), pages 40-54.

[9] F. T. Fonseca, M. J. Egenhofer, C. A. Davis, and G. Câmara. Semantic granularity in ontology-driven geographic information systems. *Annals of Mathematics and Artificial Intelligence*, 36(1-2):121–151, 2002.

[10] F. Gey, R. Larson, M. Sanderson, H. Joho, P. Clough, and V. Petras. GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track overview. In *6th Workshop of the Cross-Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop (CLEF'05)*, Vienna, Austria, 21-23 September, 2005.

[11] G. Lakoff and M. Johnson. *Metaphors We Live By*. University of Chicago Press, Chicago and London, 1980.

[12] J. Leidner. Towards a reference corpus for automatic toponym resolution evaluation. In *Workshop on Geographic Information Retrieval held at the 27th Annual International ACM SIGIR Conference (SIGIR'04)*, Sheffield, UK, 2004.

[13] J. Leidner, G. Sinclair, and B. Webber. Grounding spatial named entities for information extraction and question answering. In *Workshop on the Analysis of Geographic References held at the Joint Conference for Human Language Technology and the Annual Meeting of the Noth American Chapter of the Association for Computational Linguistics (HLT/NAACL'03)*, pages 31-38, Edmonton, Alberta, Canada, 2003.

[14] B. Martins, M. J. Silva, and M. S. Chaves. Challenges and resources for evaluating geographical IR. In *Workshop on Geographic Information Retrieval at CIKM'05*, pages 65–69, 2005.

[15] D. Santos. The importance of vagueness in translation: Examples from English to Portuguese. *Romansk Forum* 5 (1997), June 1997, pages 43-69.

[16] D. Santos. *Translation-based corpus studies: Contrasting Portuguese and English tense and aspect systems*. Amsterdam/New York, NY: Rodopi, 2004.

[17] D. Santos. What is natural language? Differences compared to artificial languages, and consequences for natural language processing. Invited lecture at SBLP'06 and PROPOR'06 (Itatiaia, RJ, Brazil, 15 May 2006).

[18] D. Santos, N. Seco, N. Cardoso, and R. Vilela. HAREM: an Advanced NER Evaluation Contest for Portuguese. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik and D. Tapias (editors), *5th International Conference on Language Resources and Evaluation (LREC'06)* (Genoa, Italy, 22-28 May 2006), pages 1986–1991.

[19] L. Sarmento. SIEMÊS - a named entity recognizer for Portuguese relying on similarity rules. In R. Vieira, P. Quaresma, M. das Graças Volpes Nunes, N. Mamede, C. Oliveira, and M. C. Dias (editors), *7th Workshop on Computational Processing of Written and Spoken Language - PROPOR'06* (Itatiaia, RJ, Brazil, 13-17 May 2006), Springer, pages 90-99.

[20] L. Sarmento, A. S. Pinto and L. Cabral. REPENTINO - A wide-scope gazetteer for Entity Recognition in Portuguese. In R. Vieira, P. Quaresma, M. das Graças Volpes Nunes, N. Mamede, C. Oliveira, and M. C. Dias (editors), *7th Workshop on Computational Processing of Written and Spoken Language - PROPOR'06* (Itatiaia, RJ, Brazil, 13-17 May 2006), Springer, pages 31-40.

[21] L. Sarmento. *BACO - A large database of text and co-occurrences*. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik and D. Tapias (editors), *5th International Conference on Language Resources and Evaluation (LREC'06)* (Genoa, Italy, 22-28 May 2006), pages 1787-1790.

[22] M. J. Silva, B. Martins, M. S. Chaves, N. Cardoso, and A. P. Afonso. Adding Geographic Scopes to Web Resources. *CEUS - Computers, Environment and Urban Systems, Elsevier Science*, 2006 (in press).

[23] D. I. Slobin. From "thought and language" to "thinking for speaking". In J. Gumperz and S. C. Levinson, editors, *Rethinking linguistic relativity*, pages 70–96. Cambridge University Press, Cambridge, 1996.

[24] J.-P. Vinay and J. Darbelnet. *Stylistique Comparée du Français et de l'Anglais: Méthode de traduction*. Nouvelle édition révue et corrigée, Paris: Didier, 1997. First edition: 1958.