Ambiguous Place Names on the Web*

Davide Buscaldi

Natural Language Engineering Lab., ELiRF Research Group,
Dpto. de Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia, Spain,
dbuscaldi@dsic.upv.es

Abstract. Geographical information is achieving an increasing importance in the World Wide Web. Everyday, the number of users looking for geographically constrained information is growing. Map-based services, such as Google or Yahoo Maps provide users with a graphical interface, visualizing results on maps. However, most of the geographical information contained in web documents is represented by means of toponyms, which in many cases are ambiguous. Therefore, it is important to properly disambiguate toponyms in order to improve the accuracy of web searches. The advent of the semantic web will allow to overcame this issue by labelling documents with geographical IDs. In this paper we discuss the problems of using toponyms in web documents instead of identifying places using tools such as Geonames RDF, focusing on the errors that affect a prototype geographical web search engine, Geooreka!, currently under development.

1 Introduction

The interest of users for geographically constrained information in the Web has increased over the past years, boosted by the availability of services such as Google Maps[1]. Sanderson and Kohler [1] showed that 18.6% of the queries submitted to the Excite search engine contained at least a geographic term, while Gan et al. [2] estimated that 12.94% of queries submitted to the AOL search engine expressed a geographically constrained information need. Most of the geographical information contained in the Web and unstructured text is composed by *toponyms*, or place names. There are two main problems that derive from using toponyms to represent geographical information. The first one is the polysemy of toponyms, or toponym ambiguity: a toponym may be used to represent more than one place, such as "Puebla" which may be used to indicate the city at $19^o3'$N, $98^o12'$W, the state in which it is contained, a suburb of Mexicali in the state of Baja California, or three more small towns in Mexico. The second problem is that the mere inclusion of a toponym in a document does not always mean that the document is geographically relevant with respect to the region or

* We would like to thank the TIN2009-13391-C04-03 research project for partially supporting this work.

[1] http://maps.google.com

area represented by the toponym. In the first case, the solution is constituted by the *Toponym Disambiguation* (TD) task, also named toponym grounding or resolution; in the second case, the solution is to carry out *Geographic Scope Resolution*, which is also affected by the problem of toponym ambiguity [3].

The Geonames ontology[2] provide users with RDF description of more than 6 million places. The use of this ontology would allow to include geospatial semantic information in the Web, eliminating the need of toponym disambiguation. Unfortunately, as noted by [4], in the Web "references to geographical locations remain unstructured and typically implicit in nature", determining a "lack of explicit spatial knowledge within the Web" which "makes it difficult to service user needs for location-specific information". In this paper, with the help of the Geooreka![3] system [5], a prototype web search engine developed at the Universidad Politécnica of Valencia in Spain, we will the problems that users interested in geographically constrained information may found because of the ambiguity of toponyms in the web.

2 Geooreka!: a Geographical Web Search Engine

Geooreka! is a search engine developed on the basis of our experiences at GeoCLEF[4] [6,7], which suggested us that the use of term-based queries could not be the optimal method to express a geographically constrained information need. For instance, it is common for users to employ vernacular names that have vague spatial extent and which do not correspond to the official administrative place name terminology. Another issue is the use of vague geographical constraints that are difficult to automatically translate from the natural language to a precise query. For instance, the query "Cultivos de tabaco al este de Puebla" ("Tobacco plantations East of Puebla") presents a double problem because of the ambiguity of the place name and the fact that the geographical constraint "East of" is vague (for instance, it does not specify if the search should be constrained within Mexico or extend to other countries).

These issues are addressed in Geooreka! by allowing the user to specify his geographical information needs using a map-based interface. The user writes a natural language query in order to represent the query theme (e.g., "Cultivos de tabaco") and selects a rectangular map in a box (Figure 1), representing the query geographical footprint. All toponyms in the box are retrieved using a PostGIS database, and then the Web is queried in order to check the maximum Mutual Information (MI) between the thematic part of the query and all the places retrieved. The complete architecture of the system can be observed in Figure 2. Web counts and MI are used in order to determine which combinations theme-toponym are most relevant with respect to the information need expressed by the user (*Selection of Relevant Queries*). In order to speed-up the process,

[2] http://www.geonames.org/ontology/
[3] http://www.geooreka.eu
[4] http://ir.shef.ac.uk/geoclef/

Fig. 1. Main page of Geooreka!

web counts are calculated using the static Google 1T Web database[5], indexed using the jWeb1T interface [8], whereas Yahoo! Search is used to retrieve the results of the queries composed by the combination of a theme and a toponym.

2.1 Model of Theme-Place Relevance

The key issue in the selection of the relevant queries is to obtain a relevance model that is able to select pairs theme-toponym that are most promising to satisfy the user's information need. On the basis of the theory of probability, we assume that the two component parts of a query, theme T and a place G, are independent if their conditional probabilities are independent, i.e., $p(T|G) = p(T)$ and $p(G|T) = p(G)$, or, equivalently, their joint probability is the product of their probabilities:

$$\hat{p}(T \cap G) = p(G)p(T) \tag{1}$$

If probabilities are calculated using page counts, that is, as the number of pages in which the term (or phrase) representing the theme or toponym appears, divided by $F_{max} = 2,147,436,244$ which is the maximum term frequency contained in the Google Web 1T database, then $\hat{p}(T \cap G)$ is the *expected* probability of co-occurrence of T and G in the same web page. It is clear that this represents a rough estimation of the fact that T occurred in G, since the mere inclusion of G in a page where T is mentioned does not guarantee the semantic relation between G and T.

Considering this model for the independence of theme and place, we can measure the divergence of the expected probability $\hat{p}(T \cap G)$ from the observed probability $p(T \cap G)$: the more the divergence, the more informative is the result

[5] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

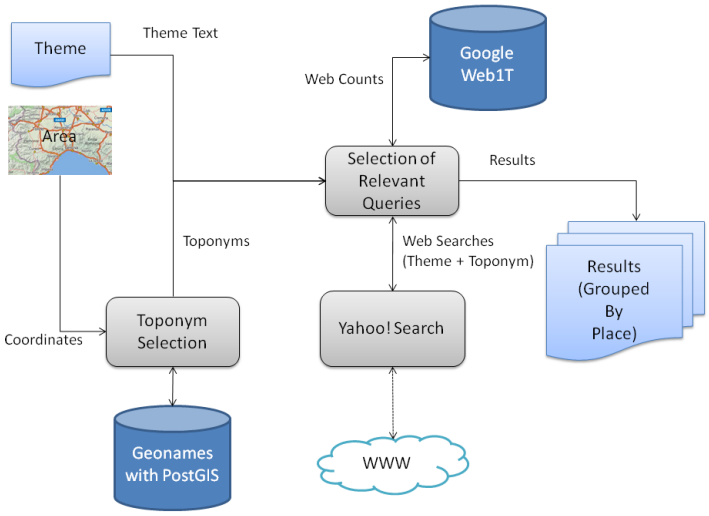

Fig. 2. Architecture of Geooreka!

of the query. The Kullback-Leibler measure [9] is commonly used in order to determine the divergence of two probability distributions.

$$D_{KL}(p(T \cap G)||\hat{p}(T \cap G)) = p(T \cap G) \log \frac{p(T \cap G)}{p(T)p(G)} \quad (2)$$

This formula is exactly one of the formulations of the *Mutual Information* (MI) of T and G, usually denoted as $(I(T;G))$.

3 Evaluation

Geooreka! has been evaluated over the GeoCLEF 2005 test set, in order to compare the results that could be obtained by specifying the geographic footprint by means of keywords and those that could be obtained using a map-based interface to define the geographic footprint of the query. With this setup, topic title only was used as input for the Geooreka! thematic part, while the area corresponding to the geographic scope of the topic was manually selected. Probabilities were calculated using the number of occurrences in the GeoCLEF collection. Occurrences for toponyms were calculated by taking into account only the *geo* index. The results were calculated over the 25 topics of GeoCLEF-2005, minus the queries in which the geographic footprint was composed of disjoint areas (for instance, "Europe" and "USA" or "California" and "Australia"), which could not be processed by Geooreka!. Mean Reciprocal Rank (MRR) was used as a measure of accuracy. The GIR system GeoWorSE, where queries are specified by text, was used as a baseline [10]. Table 1 displays the obtained results.

Table 1. MRR obtained with Geooreka!, using GeoCLEF or the WWW as target collection, compared to the MRR obtained using the GeoWorSE system, Topic Only runs.

topic	GeoWorSE	Geooreka! (GeoCLEF collection)	Geooreka! (Web)
GC-002	0.250	1.000	0.083
GC-003	0.013	1.000	1.000
GC-005	1.000	1.000	0.000
GC-006	0.143	0.000	0.500
GC-007	1.000	1.000	0.125
GC-008	0.143	1.000	0.000
GC-009	1.000	1.000	0.067
GC-010	1.000	0.333	0.250
GC-012	0.500	1.000	0.000
GC-013	1.000	0.000	0.000
GC-014	1.000	0.500	0.091
GC-015	1.000	1.000	1.000
GC-016	0.000	0.000	1.000
GC-017	1.000	1.000	0.143
GC-018	1.000	0.333	0.500
GC-019	0.200	1.000	0.045
GC-020	0.500	1.000	0.090
GC-021	1.000	1.000	0.000
GC-022	0.333	1.000	0.076
GC-023	0.019	0.200	0.125
GC-024	0.250	1.000	1.000
GC-025	0.500	0.000	0.000
average	0.584	0.698	0.280

The results show that the web-based results are sensibly worse than those obtained on the static collection. This is due primarily to two reasons: in the first place, because topics were tailored on the GeoCLEF collection. Therefore, some topics refer explicitly to events that are particularly relevant in the collection and are easier to retrieve. For instance, query `GC-005` "Japanese Rice Imports" targets documents regarding the opening of the Japanese rice market for the first time to other countries; "Japan" and "Rice" in the document collection appear together only in such documents, therefore it is easier to retrieve the relevant documents when searching the GeoCLEF collection.

The second factor affecting the results for the Web-based system is the ambiguity of toponyms, which does not allow to correctly estimate the probabilities for places. For instance, in the results obtained for topic `GC-008` ("Milk Consumption in Europe"), the MI obtained for "Turkey" was abnormally high with respect to the expected value for this country. The reason is that in most documents, the name "turkey" was referring to the animal and not to the country. This kind of ambiguity represents one of the most important issue at the time of estimating the probability of occurrence of places. Ambiguity (or, better, the polysemy of toponyms) grows together with the size and the scope of the collection being searched. The GeoCLEF collection was also semantically tagged using WordNet and Geonames IDs to identify the places referenced by toponyms, while Web content is rarely tagged using precise IDs, therefore increasing the chance of error in the estimation of probabilities for places which share the same name.

There are three kind of toponym ambiguity that can be recognised (after the two main types identified by [11]:

- Geo / Non-Geo ambiguity: in this case, a toponym is ambiguous with respect to another class of name (such as "Turkey" which may be the animal or the country);
- Geo / Geo ambiguity of different class: for instance, "Puebla" the city or the state;
- Same class Geo / Geo ambiguity.

The solution in all cases would be to use an ontology to precisely identify places in documents; the only difference is the amount of information that the ontology should include. For the first type of ambiguity, the only information needed is whether the name represents a place or not. In the second case, we would also need to know the class of the place. Finally, in the Geo / Geo ambiguity, we may differentiates places using their coordinates or by knowing the including entity, or both. The Geonames ontology contains all these information and represents the best option at the time of geographically tag place names.

4 Conclusions

The results obtained with Geooreka! over a static, semantically-labelled (at least from a geographical viewpoint) collection compared to the results obtained in

the Web showed that the imprecise identification of places is a problem for search engines destined to users who are interested in searching for geographically constrained information. The use of precise semantically tagging schemes for toponyms, such as Geonames RDF, would allow these search engines to produce more reliable results. Spreading the use of geographical tagging for the Semantic Web would also allow users to mine information using geographical constraints in a more effective way. In this sense, we would like to encourage the use of Geonamen in order to produce accurate geographically tagged Web content.

References

1. Sanderson, M., Kohler, J.: Analyzing geographic queries. In: Proceedings of Workshop on Geographic Information Retrieval (GIR04). (2004)
2. Gan, Q., Attenberg, J., Markowetz, A., Suel, T.: Analysis of geographic queries in a search engine log. In: LOCWEB '08: Proceedings of the first international workshop on Location and the web, New York, NY, USA, ACM (2008) 49–56
3. Andogah, G.: Geographically Constrained Information Retrieval. PhD thesis, University of Groningen (2010)
4. Boll, S., Jones, C., Kansa, E., Kishor, P., Naaman, M., Purves, R., Scharl, A., Wilde, E.: Location and the web (locweb 2008). In: Proceeding of the 17th international conference on World Wide Web. WWW '08, New York, NY, USA, ACM (2008) 1261–1262
5. Buscaldi, D., Rosso, P.: Geooreka: Enhancing Web Searches with Geographical Information. In: Proc. Italian Symposium on Advanced Database Systems SEBD-2009, Camogli, Italy (2009) 205–212
6. Buscaldi, D., Rosso, P., Sanchis, E.: Using the WordNet Ontology in the GeoCLEF Geographical Information Retrieval Task. In Peters, C., Gey, F.C., Gonzalo, J., Mller, H., Jones, G.J., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D., eds.: Accessing Multilingual Information Repositories. Volume 4022 of Lecture Notes in Computer Science. Springer, Berlin (2006) 939–946
7. Buscaldi, D., Rosso, P.: On the relative importance of toponyms in geoclef. In: Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers, Springer (2007) 815–822
8. Giuliano, C.: jWeb1T: a library for searching the Web 1T 5-gram corpus. (2007) Software available at http://tcc.itc.it/research/textec/tools-resources/jweb1t.html.
9. Kullback, S., Leibler, R.A.: On Information and Sufficiency. Annals of Mathematical Statistics **22**(1) (1951) pp. 79–86
10. Buscaldi, D., Rosso, P.: Using GeoWordNet for Geographical Information Retrieval. In: Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. (2009) 863–866
11. Amitay, E., Harel, N., Sivan, R., Soffer, A.: Web-a-where: Geotagging web content. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK (2004) 273–280