The UPV at GeoCLEF 2008: The GeoWorSE System

Davide Buscaldi and Paolo Rosso
Natural Language Engineering Lab,
Dpto. de Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia, Spain
{dbuscaldi, prosso}@dsic.upv.es

Abstract

This year our system was complemented with a map-based filter. During the indexing phase, all places are disambiguated and assigned their coordinates on the map. These coordinates are stored in a separate index. The search process is carried out in two phases: in the first one, we search the collection with the same method applied in 2007, which exploits the expansion of index terms by means of WordNet synonyms and holonyms. The next phase consists in a re-ranking of the results of the previous phase depending on the distance of document toponyms from the toponyms in the query, or depending on the fact that the document contains toponyms that are included in an area defined by the query. The area is calculated from the toponyms in the query and their meronyms. This is the first attempt to use GeoWordNet, a resource that includes the geographical coordinates of the places listed in WordNet, for the Geographical Information Retrieval task. The results show that map-based filtering allows to improve the results obtained by the base system, based only on the textual information.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; I.2 [**Artificial Intelligence**]: I.2.7 Natural Language Processing

General Terms

Measurement, Performance, Experimentation, Text Analysis

Keywords

Geographical Information Retrieval, Index Term Expansion, Map-based Filtering

1 Introduction

Our group has been participating in the GeoCLEF task since 2005, focusing on the use of WordNet [8] ontology for the Geographical Information Retrieval Task (GIR). The method that produced the best results was introduced in 2006 and refined in 2007. It uses exclusively textual information [3, **?**], exploiting the part-of (or *holonymy*) relationship provided by WordNet. It consists in an expansion of geographical locations with their holonyms, that is, the containing entities. These are stored in an index together with their synonyms. This method would allow a user searching for information about *Spain* to find also documents containing *Valencia*, *Madrid* or *Barcelona*, although the original document does not contain the word "Spain".

For our 2008 participation, we attempted to improve the method by introducing map-based filtering. The most succesful methods in 2006 [7] and 2007 [4] both combined textual retrieval with geographical-based filtering and ranking. This observation prompted us to introduce a similar feature in our system. The main obstacle was determined by the fact that we use WordNet, which did not provide us with geographical coordinates for toponyms. Therefore, we first had to develop GeoWordNet [2], a georeferenced version of WordNet. By combining this resource with the WordNet-based toponym disambiguation algorithm in [1], we are able to assign to the place names in the collection their actual geographical coordinates and to perform some geographical reasoning. We called the resulting system GeoWorSE (an acronym for *Geographical Wordnet Search Engine*).

In the following section, we describe the GeoWorSE system. In section 3 we describe the characteristics of our submissions and the obtained results.

2 The GeoWorSE System

The core of the system is constituted by the Lucene[1] open source search engine, version 2.1. Named Entity Recognition and classification is carried out by the Stanford NER system based on Conditional Random Fields [5]. The access to WordNet is done by the MIT Java WordNet Interface [2]. The toponym disambiguator is based on the method presented in [1].

2.1 Indexing

During the indexing phase, the documents are examined in order to find location names (*toponym*) by means of the Stanford NER system. When a toponym is found, the disambiguator determines the correct reference for the toponym. Then, a modified lucene indexer adds to the *geo* index the toponym coordinates (retrieved from GeoWordNet); finally, it stores in the *wn* index the toponym together with its holonyms and synonyms. All document terms are stored in the *text* index. In Figure 1 we show the architecture of the indexing module.

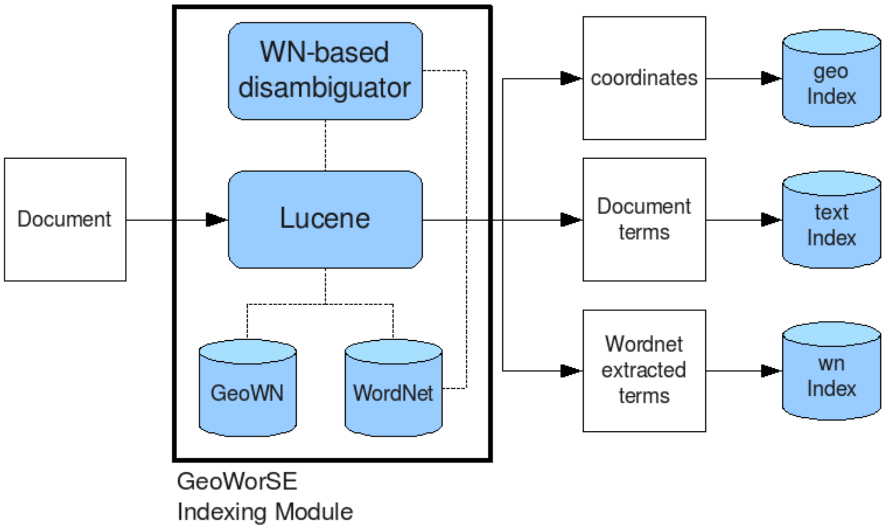

Figure 1: Diagram of the Indexing module

[1] http://lucene.apache.org/

[2] http://www.mit.edu/∼markaf/projects/wordnet/

The indices are then used in the search phase, although the *geo* index is not used for search: it is used only to retrieve the coordinates of the toponyms in the document.

2.2 Searching

The architecture of the search module is shown in Figure 2.

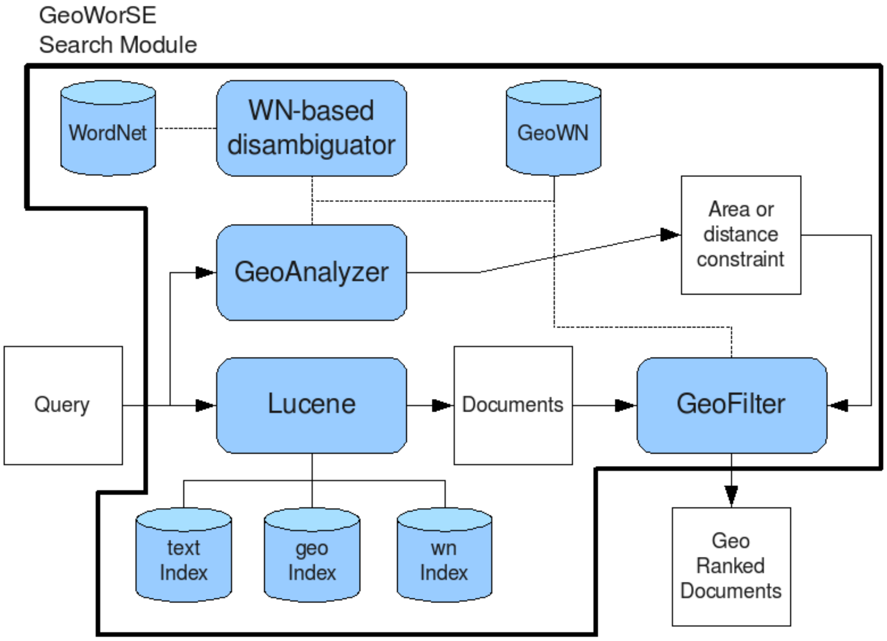

Figure 2: Diagram of the Search module

The topic text is searched by Lucene in the text index. All the toponyms are extracted by the Stanford NER and searched for by Lucene in the wn index with a weight 0.25 with respect to the content terms. The result of the search is a list of documents ranked using the Lucene's weighting scheme (basically, this is the output that the system presented in 2007 would have returned). At the same time, the toponyms are passed to the GeoAnalyzer, which creates a geographical constraint that is used to re-rank the document list. The GeoAnalyzer may return two types of geographical constraints:

- a *distance* constraint, corresponding to a point in the map: the documents that contain locations closer to this point will be ranked higher;
- an *area* constraint, correspoinding to a polygon in the map: the documents that contain locations included in the polygon will be ranked higher;

For instance, in topic $10.2452/58 - GC$ there is a distance constraint: "Travel problems at major airports near to London". Topic $10.2452/76 - GC$ contains an area constraint: "Riots in South American prisons". The GeoAnalyzer determines the area using WordNet meronyms: *South America* is expanded to its meronyms: *Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Guyana, Paraguay, Peru, Uruguay, Venezuela*. The area is obtained by calculating the convex hull of the points associated to the meronyms using the Graham algorithm [6].

The topic narrative allows to increase the precision of the considered area, since the toponyms in the narrative are also expanded to their meronyms (when possible). Figure 3 shows the convex hulls of the points corresponding to the meronyms of "South America", using only topic and description (left) or all the fields, including narrative (right).

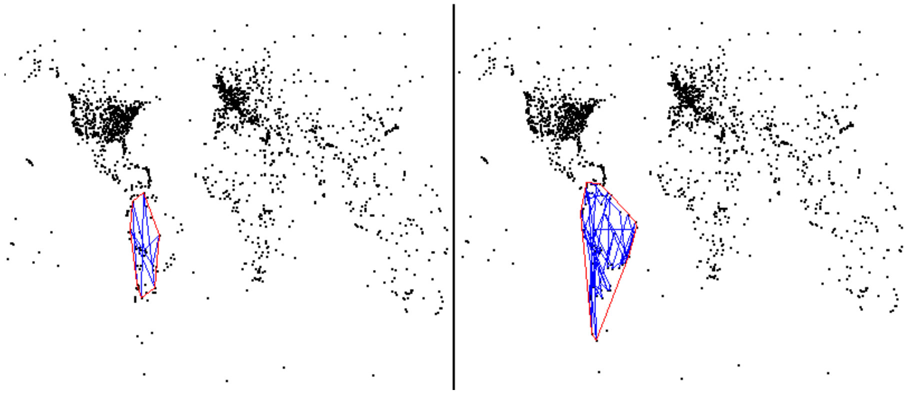

Figure 3: Areas corresponding to “South America” for topic $10.2452/76 - GC$, calculated as the convex hull (in red) of the points (connected by blue lines) extracted by means of the WordNet meronymy relationship. On the left, the result using only topic and description; on the right, also the narrative has been included. Black dots represents the locations contained in GeoWordNet.

The objective of the GeoFilter module is to re-rank the documents retrieved by Lucene, according to geographical information. If the constraint extracted from the topic is a *distance* constraint, the weights of the documents are modified according to the following formula:

$$w(doc) = w_{Lucene}(doc) * (1 + \exp(-\min_{p \in P} d(q,p))) \tag{1}$$

Where w_{Lucene} is the weight returned by Lucene for the document *doc*, P is the set of points in the document, and q is the point extracted from the topic.

If the constraint extracted from the topic is an *area* constraint, the weights of the documents are modified according to formula 2:

$$w(doc) = w_{Lucene}(doc) * \left(1 + \frac{|P_q|}{|P|}\right) \tag{2}$$

where P_q is the set of points in the document that are contained in the area extracted from the topic.

3 Experiments

We submitted a total of 6 runs at GeoCLEF 2008. Two runs were used as “benchmarks”: they were obtained by using the base Lucene system, without index term expansion, in one case considering only topic title and description, and all fields in the other case. One run was generated with the system we presented in 2007 (without the re-ranking by the geofilter module). For the three remaining submissions we used the new system with topic and description only, topic, description and narrative, and a configuration that do not use wordnet information during the search phase.

In Table 1 we show the results obtained in terms of Mean Average Precision and R-Precision for all the submitted runs.

The obtained results show that the runs that used only the information contained in the Title and Description fields were considerably better than runs that included also the narrative, inverting the trend of the past GeoCLEF exercises, where TDN runs usually were better than TD ones. We analyzed the results topic by topic and compared the performance of runs that used TD only and runs that used also narrative. The topics that present the greatest difference between the two types of runs are $10.2452/GC - 76$, $10.2452/GC - 77$ and $10.2452/GC - 91$, in which the use of narrative makes the results worse. Figure 4 shows in detail the average difference between the two types of runs.

Table 1: Mean Average Precision (MAP) and R-Precision obtained for all runs.

system	run ID	fields	MAP	R-Prec
base system	NLEL0802	TDN	0.201	0.217
	NLEL0804	TD	0.224	0.248
2007 system	NLEL0803	TDN	0.216	0.226
GeoWorSE (2008)	NLEL0505	TDN	0.204	0.211
	NLEL0806	TD	**0.254**	**0.262**
GeoWorSE (no WN)	NLEL0807	TDN	0.202	0.219

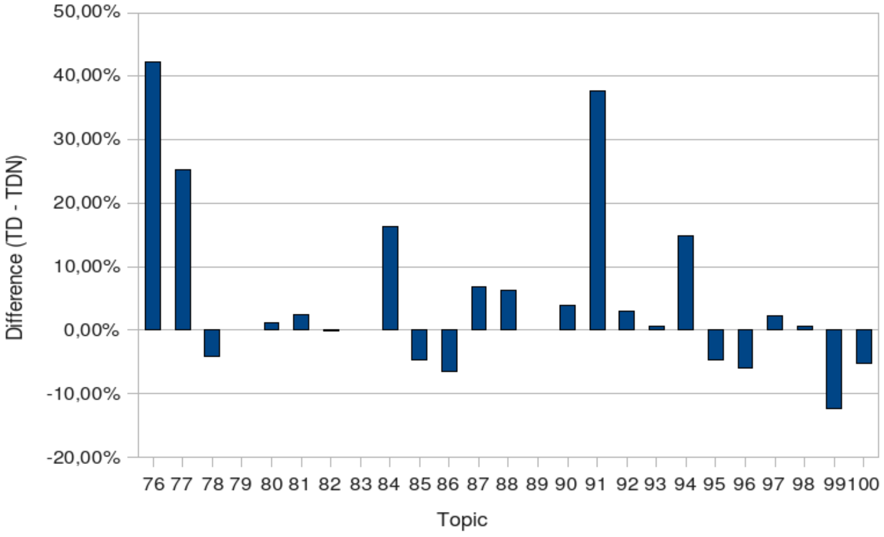

Figure 4: Average difference (in percentage) for the mean average precision obtained with runs that used only title and description (TD) and runs that used also narrative (TDN).

The analysis of the narratives of these three runs shows that they include a long list of place names. These lists alter the balance between content keywords and geographical terms of the query, with the effect of giving more importance in the query to the geographical constraint than over the words related to the information searched for.

The comparison of the new method with the baseline obtained without the map constraint shows that map filtering allowed to improve the results over the method previously used.

4 Conclusions and Further Work

We introduced a map-based filtering method that allowed us to improve the results obtained with our WordNet-based method. The best results were obtained with the map-based method, taking into account only topic title and description. We believe that topic narrative could be used more efficiently to improve the map-based filtering rather than using it directly during the search phase. We plan to carry out some experiments with this configuration in order to verify our hypothesis.

Acknowledgements

We would like to thank the TIN2006-15265-C06-04 research project for partially supporting this work.

References

[1] Davide Buscaldi and Paolo Rosso. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Systems*, 22(3):301–313, 2008.

[2] Davide Buscaldi and Paolo Rosso. Geo-wordnet: Automatic georeferencing of wordnet. In *Proc. 5th Int. Conf. on Language Resources and Evaluation, LREC-2008*, 2008.

[3] Davide Buscaldi, Paolo Rosso, and Emilio Sanchis. A wordnet-based indexing technique for geographical information retrieval. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Mller, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini, Maarten de Rijke, and Danilo Giampiccolo, editors, *Lecture Notes in Computer Sciences*, volume 4730 of *Lecture Notes in Computer Science*, pages 954–957. Springer, Berlin, 2007.

[4] Horacio Rodríguez Daniel Ferrés. TALP at GeoCLEF 2007: Using Terrier with Geographical Knowledge Filtering. In C. Peters, editor, *CLEF 2007 Working Notes*, Budapest, Hungary, 2007.

[5] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370, U. of Michigan - Ann Arbor, 2005. ACL.

[6] Ronald L. Graham. An efficient algorith for determining the convex hull of a finite planar set. *Information Processing Letters*, 1(4):132–133, 1972.

[7] Bruno Martins, Nuno Cardoso, Marcirio Silveira Chaves, Leonardo Andrade, and Mário J. Silva. The university of lisbon at geoclef 2006. In C. Peters, editor, *CLEF 2006 Working Notes*, Alicante, Spain, 2006.

[8] George. A. Miller. Wordnet: A lexical database for english. In *Communications of the ACM*, volume 38, pages 39–41, 1995.