Using GeoWordNet for Geographical Information Retrieval[⋆]

Davide Buscaldi and Paolo Rosso

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

Abstract. We present a method that uses GeoWordNet for Geographical Information Retrieval. During the indexing phase, all places are disambiguated and assigned their coordinates on the world map. Documents are first searched for by means of a term-based search method, and then re-ranked according to the geographical information. The results show that map-based re-ranking allows to improve the results obtained by the base system, which relies only on textual information.

1 Introduction

One of the main issues in Geographical Information Retrieval (GIR) consists in finding the perfect balance between the thematic part and the geographical part in queries [1,2]. Currently available GIR systems are not able to perform significantly better than standard keyword-based IR systems. In our past participations at GeoCLEF we attempted to integrate geographical knowledge at keyword level in the Lucene[1] search engine, focusing on the use of the WordNet [3] ontology for both query reformulation and index term expansion.

Ferres and Rodríguez [4] obtained good results at GeoCLEF 2007 by combining textual retrieval with map-based filtering and ranking. This kind of integration between geographical knowledge and term-based ranking was previously introduced by [5] in 2006, but it did not demonstrate useful. However, we attempted 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 [6]. By combining this resource with the WordNet-based toponym disambiguation algorithm presented in [7], we were able to assign to the place names in the collection their actual geographical coordinates and to perform some geographical reasoning. We named the resulting system GeoWorSE (an acronym for *Geographical Wordnet Search Engine*). This is the first time that GeoWordNet is used for IR.

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

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

[2] http://www.dsic.upv.es/grupos/nle/resources/geo-wn/download.html

2 The GeoWorSE GIR System

During the indexing phase the documents are examined in order to find location names (*toponyms*) by means of the Stanford NER system [8]. When a toponym is found, the disambiguator determines the correct reference for the toponym. Then, the system adds the toponym coordinates (retrieved from GeoWordNet) to the *geo* index and stores in the *wn* index the toponym together with its holonyms and synonyms. All document terms are stored in the *text* index.

The topic text is split into "content" terms, which are searched in the *text* index, and the "geo" part, constituted by toponyms extracted by the Stanford NER. The "geo" terms are searched for 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 Lucene's weighting scheme. At the same time, the toponyms are analyzed in order to find a geographical constraint that can be of the following two types:

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

The nature of the constraint is determined automatically, on the basis of the data contained in GeoWordNet (that is, whether the toponym can be expanded to an area by means of its meronyms or not). For instance, topic $10.2452/58-GC$ contains 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 expands *South America* 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 [9].

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 compared the results obtained with the system using three configurations:

- The Lucene system, without WordNet expansion neither the map-based reranking (label: Luc)
- The system with WordNet expansion but without the map-based reranking (label: L+WN)
- The system with WordNet expansion and map-based reranking (label: GWN)

The results were calculated over all the topics of the GeoCLEF since 2005.

In Table 1 we show the obtained results.

Table 1. Mean Average Precision (MAP) and R-Precision obtained for all topics, using TD (topic and description) and TDN (topic, description and narrative) fields.

		TD		TDN	
system	year	MAP	R-Prec	MAP	R-Prec
Luc	2005	0.311	0.340	0.321	0.333
	2006	**0.251**	**0.242**	**0.274**	**0.265**
	2007	0.228	0.245	0.249	0.268
	2008	0.224	0.248	0.210	0.223
average		0.253	0.269	0.263	0.272
L+WN	2005	**0.328**	**0.362**	0.324	0.339
	2006	0.245	0.236	0.261	0.252
	2007	0.242	**0.252**	**0.264**	**0.272**
	2008	**0.269**	**0.277**	**0.216**	**0.226**
average		**0.271**	**0.282**	**0.266**	0.272
GWN	2005	0.320	0.352	**0.326**	**0.347**
	2006	0.247	0.239	0.263	0.261
	2007	0.242	0.247	0.253	0.263
	2008	0.264	0.267	0.204	0.211
average		0.268	0.276	0.262	0.271

The results show that there is no significant difference between the use of the map-based re-ranking and the use of the WordNet-enhanced method. We believe that there are two reasons for this behaviour: the first one is the presence of errors in toponym disambiguation, the second one the fact that in the re-ranking phase the rank of documents is not taken into account. In both cases further work is needed in order to estimate how these features may affect the results.

4 Conclusions and Further Work

We introduced a map-based filtering method in our WordNet-based GIR system. The obtained results do not show any significant improvement over the previous method. We will carry out a study of the weights and the formulae that are used to re-rank documents. As a future work, we would like to implement a dynamical ranking scheme, such as the one proposed by [10], based on *geographic specificity*.

References

1. Kornai, A.: Evaluating Geographic Information Retrieval. In Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., eds.: Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers. Lecture Notes in Computer Science. Springer (2006) pp. 928–938
2. Buscaldi, D., Rosso, P.: On the Relative Importance of Toponyms in GeoCLEF. In Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D., eds.: 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. Lecture Notes in Computer Science. Springer (2008) pp. 815–822
3. Miller, G.A.: WordNet: A Lexical Database for English. In: Communications of the ACM. Volume 38. (1995) pp. 39–41
4. Ferrés, D., Rodríguez, H.: TALP at GeoCLEF 2007: Results of a Geographical Knowledge Filtering Approach with Terrier. In Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D., eds.: 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. Lecture Notes in Computer Science, Springer (2008) pp. 830–833
5. Martins, B., Cardoso, N., Silveira Chaves, M., Andrade, L., Silva, M.J.: The University of Lisbon at GeoCLEF 2006. In Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M., eds.: Evaluation of Multilingual and Multi-modal Information Retrieval 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers. Lecture Notes in Computer Science. Springer (2007) pp. 986–994
6. Buscaldi, D., Rosso, P.: Geo-WordNet: Automatic Georeferencing of WordNet. In: Proc. 5th Int. Conf. on Language Resources and Evaluation, LREC-2008, Marrakech, Morocco (2008)
7. Buscaldi, D., Rosso, P.: A Conceptual Density-based Approach for the Disambiguation of Toponyms. International Journal of Geographical Information Systems **22**(3) (2008) pp. 301–313
8. Finkel, J.R., Grenager, T., Manning, C.: 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), U. of Michigan - Ann Arbor, ACL (2005) pp. 363–370
9. Graham, R.L.: An Efficient Algorithm for Determining the Convex Hull of a Finite Planar Set. Information Processing Letters **1**(4) (1972) pp. 132–133
10. Yu, B., Cai, G.: A Query-aware Document Ranking Method for Geographic Information Retrieval. In Purves, R., Jones, C., eds.: Proceedings of the 4th ACM Workshop On Geographic Information Retrieval, GIR 2007, Lisbon, Portugal, November 9, 2007. ACM (2007) pp. 49–54