The UPV at GeoCLEF 2007

Davide Buscaldi and Paolo Rosso
Dpto. de Sistemas Informticos y Computación (DSIC),
Universidad Politcnica de Valencia, Spain
{dbuscaldi, prosso}@dsic.upv.es

Abstract

In this work we attempted to determine the relative importance of the geographical and WordNet-extracted terms with respect to the remainder of the query. Our system is based on Lucene and uses LingPipe for Named Entity recognition. Geographical terms are expanded with WordNet holonyms and synonyms and indexed separately. We checked the relative importance of the terms by boosting them with reduction factors (0.75, 0.5 and 0.25). The comparison to the clean system (using only Lucene) shows that it is possible to improve the mean average precision if the importance of geographical terms is equal or less than the half with respect to the content words in the query. We also observed that WordNet holonyms may help in improving the recall but the term expansion is sensible to ambigue place names. As a further work, we will need to implement a toponym disambiguation method in order to reduce the impact of this kind of ambiguity.

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

General Terms

Measurement, Performance, Experimentation

Keywords

Geographical Information Retrieval, Index Term Expansion

1 Introduction

Since our first participation at the GeoCLEF we have been developing a method that can use the information contained in the WordNet [6] ontology for the Geographical Information Retrieval task. In our first attempt [2, 5] we simply used synonyms (alternate names) and meronyms of locations that appeared in the query in order to expand the query itself. This method performed poor, due to the noise introduced by the expansion. Subsequently, we introduced a method that exploits the inverse of the meronymy relationship - *holonymy* (a concept A is *holonym* of another concept B if A contains B). We named this method *Index Term Expansion* [3]. With this method we add to the geographical index terms the informations about their holonyms, such that a user looking information about *Spain* will find documents containing *Valencia*, *Madrid* or *Barcelona* even if the document itself does not contain any reference to Spain. The results obtained with this method showed that the inclusion of WordNet holonyms allowed to obtain an improvement in recall, although it was not so significant as we hoped (about 1%). Moreover, we noticed that the

use of the Index Term Expansion method did not allow to obtain the same precision of the clean system. We individuated the reason of this behaviour in the fact that the geographical terms were assigned the same importance of the other terms of the query. Therefore, in this participation we attempted to determine the relative importance of geographical and WordNet-extracted terms with respect to the remainder of the terms of the query. This has been done by means of the separation of the index of geographical terms from the general index and the creation of another index that contains only WordNet-extracted terms.

In the following section, we describe the system and how index term expansion works. In section 3 we describe the characteristics of our submissions and show a resume of the obtained results.

2 Our System

The core of the system is constituted by the Lucene[1] open source search engine, version 2.1. The engine is supported by a module that uses LingPipe[2] for HMM-based Named Entity recognition (this module performs the task of recognizing geographical names in text), and another one that is based on the MIT Java WordNet Interface [3] in order access the WordNet ontology and find synonyms and holonyms of the geographical names.

2.1 Indexing

During the indexing phase, the documents are examined in order to find location names (*toponym*) by means of LingPipe. When a toponym is found, then two actions are performed: first of all, the toponym is added to a separate index (*geo* index) that contains only the toponyms. In the second place, WordNet is examined in order to find holonyms (recursively) and synonyms of the toponym. The retrieved holonyms and synonyms are put in another separate index (*wn* index), containing only wordnet-related information.

For instance, consider the following text from the document `GH950630-000000` in the Glasgow Herald 95 collection:

> ...The British captain may be seen only once more here, at next month's world championship trials in Birmingham, where all athletes must compete to win selection for Gothenburg...

The following toponyms are added to the *geo* index: "Birmingham", "Gothenburg". Birmingham is found in WordNet both as *Birmingham, Pittsburgh of the South*, in the United States and *Birmingham, Brummagem*, an important city in England. The holonyms in the first case are *Alabama*, *Gulf States*, *South*, *United States of America* and their synonyms. In the second case, we obtain *England*, *United Kingdom*, *Europe* and their synonyms. All these words are added to the *wn* index for Birmingham, since we did not use any method in order to disambiguate the toponym. For Gothenburg we obtain *Sweden* and *Europe* again, together with the original Swedish name of Gothenburg (*Goteborg*). These words are also added to the *wn* index.

2.2 Searching

For each topic, LingPipe is run again in order to find the geographical terms. In the search phase, we do not use WordNet. However, the toponyms individuated by LingPipe are searched in the geographical and/or WordNet indices.

[1] http://lucene.apache.org/
[2] http://www.alias-i.com/lingpipe/
[3] http://www.mit.edu/~markaf/projects/wordnet/

3 Experiments

We submitted a total of 12 runs at GeoCLEF 2007. 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. The remaining runs used the *geo* index or *wn* index or both, with different weightings that were submitted using the Lucene "Boost" operator. This operator allows to assign relative importance to terms. This means that a term with, for instance, a boost factor of 4 will be four times more important than the other terms in the query. We used 0.75, 0.5 and 0.25 as boost factor for geographical and WordNet terms, in order to study their importance in the retrieval process.

In the following tables we show the results obtained in terms of Mean Average Precision and Recall for all the submitted runs.

Table 1: Mean Average Precision (MAP) and Recall obtained for all the "Title+Description only" runs.

run ID	geo boost	wn boost	MAP	Recall
rfiaUPV01	0	0	0.226	0.886
rfiaUPV03	0.5	0.0	0.227	0.869
rfiaUPV05	0.5	0.25	0.238	0.881
rfiaUPV07	0.75	0.0	0.224	0.860
rfiaUPV08	0.75	0.25	0.224	0.860
rfiaUPV09	0.25	0.25	0.239	0.888
rfiaUPV10	0.25	0.0	0.236	0.891
rfiaUPV11	0.5	0.5	0.239	0.886
rfiaUPV12	0.75	0.75	0.231	0.877

Table 2: Mean Average Precision and Recall obtained for the "All fields" runs.

run ID	geo boost	wn boost	MAP	Recall
rfiaUPV02	0	0	0.247	0.903
rfiaUPV04	0.5	0.0	0.256	0.915
rfiaUPV06	0.5	0.25	0.263	0.926

The results obtained with the topic title and description (Table 1) show that by considering geographical terms less important is possible to obtain a better MAP. The integration of WordNet terms allows to improve further the MAP, although it has almost no effect over recall. This may be due to the noise introduced by the ambiguity of some toponyms. However, if we consider all the topic fields (Table 2) we can observe that the introduction of WordNet allowed to improve also the recall. We need to carry out further study of the data in order to fully understand these results.

4 Conclusions and Further Work

The obtained results show that geographical terms are less important than content words in the topics. Reducing the importance of geographical terms allowed to improve the mean average precision. The impact of WordNet is not clear. We suppose that the effects of the introduction of WordNet synonyms and holonyms are conditioned by the ambiguity of some toponyms, such as "Birmingham" that can be a city in Alabama or in England. The ambiguity of toponyms is a common problem in news text [4], and currently various approaches are being developed [7, 1].

We plan to carry out more experiments in order to understand better the impact of toponym ambiguity over geographical information retrieval.

Acknowledgements

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

References

[1] D. Buscaldi and P. Rosso. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Systems*, 2008. accepted, to be published.

[2] D. Buscaldi, P. Rosso, and E. Sanchis. Using the wordnet ontology in the geoclef geographical information retrieval task. 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, *Accessing Multilingual Information Repositories*, volume 4022 of *Lecture Notes in Computer Science*, pages 939–946. Springer, Berlin, 2006.

[3] D. Buscaldi, P. Rosso, and E. 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] Eric Garbin and Inderjeet Mani. Disambiguating toponyms in news. In *conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT05)*, pages 363–370, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[5] Fredric Gey, Ray Larson, Mark Sanderson, Hideo Joho, and Paul Clough. Geoclef: the clef 2005 cross-language geographic information retrieval track. In *Working notes for the CLEF 2005 Workshop (C.Peters Ed.)*, Vienna, Austria, 2005.

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

[7] Simon Overell, Joao Magalhaes, and Stefan Rüger. Place disambiguation with co-occurrence models. In Carol Peters, editor, *GeoCLEF 2006 Workshop*, Alicante, Spain, 2006.