A WordNet-based indexing technique for Geographical Information Retrieval

Davide Buscaldi, Paolo Rosso, and Emilio Sanchis

Dpto. de Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia, Spain,
{dbuscaldi,prosso,esanchis}@dsic.upv.es

Abstract. This paper presents an indexing technique based on WordNet synonyms and holonyms. This technique has been developed for the Geographical Information Retrieval task. It may help in finding implicit geographic information contained in texts, particularly if the indication of the containing geographical entity is omitted. Our experiments were carried out with the Lucene search engine over the GeoCLEF 2006 set of topics. Results show that expansion can improve recall in some cases, although a specific ranking function is needed in order to obtain better results in terms of precision.

1 Introduction

Nowadays, many documents in the web or in digital libraries contain some kind of geographical information. News stories often contain a reference that indicates the place where an event took place. Nevertheless, the correct identification of the locations to which a document refers to is not a trivial task. Explicit information about areas including the cited geographical entities is usually missing from texts (e.g. usually *France* is not named in a news related to *Paris*). Moreover, using text strings in order to identify a geographical entity creates problems related to ambiguity, synonymy and names changing over time.

Ambiguity and synonymy are well-known problems in the field of Information Retrieval. The use of semantic knowledge may help to solve these problems, even if no strong experimental results are yet available in support of this hypothesis. Some results [1] show improvements by the use of semantic knowledge; others do not [7]. The most common approaches make use of standard keyword-based techniques, improved through the use of additional mechanisms such as document structure analysis and automatic query expansion.

We investigated the use of automatic query expansion by means of WordNet [6] meronyms and synonyms in our previous participation to the GeoCLEF, but the proposed technique did not obtain good results [5,3]. Although there are some effective query expansion techniques [4] that can be applied to the geographical domain, we think that the expansion of the queries with synonyms and meronyms does not fit with the characteristics of the GeoCLEF task. Other methods using thesauri with synonyms for general domain IR also did not achieve promising results [8].

In our work for GeoCLEF 2006 we focused on the use of WordNet in the indexing phase, specifically for the expansion of index terms by means of synonyms and holonyms. We used the subset of the WordNet ontology related to the geographical domain.

2 WordNet-based Index Terms Expansion

The expansion of index terms is a method that exploits the *holonymy* relationship of the WordNet ontology. A concept A is *holonym* of another concept B if A contains B, or viceversa B is part of A (B is also said to be *meronym* of A). Therefore, our idea is to 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.

We used the well-known Lucene[1] search engine. Two indices are generated for each text during the indexing phase: a *geo* index, containing all the geographical terms included in the text and also those obtained through WordNet, and a *text* index, containing the stems of text words that are not related to geographical entities. Thanks to the separation of the indices, a document containing "John Houston" will not be retrieved if the query contains "Houston", the city in Texas. The adopted weighting scheme is the usual $tf \cdot idf$. The geographical terms in the text are identified by means of a Named Entity (NE) recognizer based on maximum entropy[2], and put into the *geo* index, together with all its synonyms and holonyms obtained from WordNet.

For instance, consider the following text:

"*A federal judge in Detroit struck down the National Security Agency's domestic surveillance program yesterday, calling it unconstitutional and an illegal abuse of presidential power.*"

The NE recognizer identifies *Detroit* as a geographical entity. A search for Detroit synonyms in WordNet returns {*Detroit, Motor city, Motown*}, while its holonyms are:

```
-> Michigan, Wolverine State, Great Lakes State, MI
   -> Midwest, middle west, midwestern United States
       -> United States, United States of America, U.S.A., USA,
            U.S., America
           -> North America
             -> northern hemisphere
             -> western hemisphere, occident, New World
             -> America
```

[1] http://lucene.jakarta.org
[2] Freely available from the OpenNLP project: http://opennlp.sourceforge.net

Therefore, the following index terms are put into the *geo* index: { *Michigan, Wolverine State, Great Lakes State, MI, Midwest, middle west, midwestern United States, United States, United States of America, U.S.A., USA, U.S., America, North America, northern hemisphere, western hemisphere, occident, New World* }.

3 Experimental Results

The runs submitted at GeoCLEF 2006 were four, two with the WordNet-based system and the other ones with the "clean" system (i.e. without the expansion of index terms). The runs were the mandatory "title-description" and "title-description-narrative" for each of the two systems. For every query the top 1000 ranked documents have been returned.

In table 1 we show the recall and average precision values obtained. Recall has been calculated for each run as the number of relevant documents retrieved divided by the number of relevant documents in the collection (378). The average precision is the non-interpolated average precision calculated for all relevant documents, averaged over queries.

The results obtained in term of precision show that non-WordNet runs are better than the other ones, particularly for the all-fields run *rfiaUPV02*. However, as we expected, we obtained an improvement in recall for the WordNet-based system, although the improvement was not so significant as we hoped (about 1%).

Table 1. Average precision and recall values obtained for the four runs. WN: tells whether the run uses WordNet or not.

run	WN	avg. precision	recall
rfiaUPV01	no	25.07%	78.83%
rfiaUPV02	no	**27.35%**	80.15%
rfiaUPV03	yes	23.35%	79.89%
rfiaUPV04	yes	26.60%	**81.21%**

In order to better understand the obtained results, we analyzed the topics in which the two systems differ more (in terms of recall). Topics 40 and 48 resulted the worst ones for the WordNet based system. The explication is that topic 40 does not contain any name of geographical place ("*Cities near active volcanoes*"); topic 48 contains references to places (*Greenland* and *Newfoundland*) for which WordNet provides little information.

On the other hand, the system based on index term expansion performed particularly well for topics 27, 37 and 44. These topics contain references to countries and regions (*Western Germany* for topic 27, *Middle East* in the case of 37 and *Yugoslavia* for 44) for which WordNet provides a rich information in terms of meronyms.

4 Conclusions and Further Work

The obtained results show that the expansion of index terms by means of WordNet holonyms can improve slightly the recall. However, a better ranking function needs to be implemented in order to obtain also an improvement in precision. Further investigation directions will include the implementation of the same method with a richer (in terms of coverage of geographical places) resource, an ontology we are currently developing using the GNS and GNIS gazetteers together with WordNet itself and Wikipedia [2], as the experimentation of various ranking functions that weight differently the geographical terms with respect to the non-geographical ones.

Acknowledgments

We would like to thank the TIN2006-15265-C06-04 research project for partially supporting this work. This paper is a revised version of the work titled "WordNet-based Index Terms Expansion for Geographical Information Retrieval" included in the CLEF 2006 Working Notes.

References

1. K. Bo-Yeong, K. Hae-Jung, and L. Sang-Lo. Performance analysis of semantic indexing in text retrieval. In *CICLing 2004, Lecture Notes in Computer Science, Vol. 2945*, Mexico City, Mexico, 2004.
2. D. Buscaldi, P. Rosso, and P. Peris. Inferring geographical ontologies from multiple resources for geographical information retrieval. In *Proceedings of the 3rd GIR Workshop, SIGIR 2006*, Seattle, WA, 2006.
3. D. Buscaldi, P. Rosso, and E. Sanchis. Using the wordnet ontology in the geoclef geographical information retrieval task. In *Proceedings of the CLEF 2005 Workshop*, Vienna, Austria, 2005.
4. G. Fu, C.B. Jones, and A.I. Abdelmoty. Ontology-based spatial query expansion in information retrieval. In *Proceedings of the ODBASE 2005 conference*, 2005.
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. Paolo Rosso, Edgardo Ferretti, D. Jiménez, and Vicente Vidal. Text categorization and information retrieval using wordnet senses. In *CICLing 2004, Lecture Notes in Computer Science, Vol. 2945*, Mexico City, Mexico, 2004.
8. Ellen Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the ACM SIGIR 1994*, 1994.