WordNet-based Index Terms Expansion for Geographical Information Retrieval

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

August 20, 2006

Abstract

This paper presents the results obtained by our group at the GeoCLEF 2006. Our system used a method based on the expansion of index terms, which exploits WordNet synonyms and holonyms. This may help in finding implicit geographic information from text, particularly in the cases in which the indication of the containing geographical entity is omitted. The system is based on the Lucene search engine. We submitted two kind of runs, one using WordNet to expand the index terms, the other without any expansion. 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.

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, Algorithms, Performance, Experimentation

Keywords

Geographical Information Retrieval, Index Term Expansion, WordNet

1 Introduction

The application of Natural Language Processing techniques to geographical names presents various issues. Although most of the information available in electronic format involves some kind of spatial awareness, correctly identifying 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.

In our 2005 participation to the GeoCLEF, the use of automatic query expansion did not obtain good results [5]. Although currently are available some effective query expansion techniques [4] 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 [6] for the expansion of index terms by means of synonyms and holonyms, a technique we described last year even if we were not able to send runs due to the time needed to index the collection [3]. We used the subset of the WordNet ontology related to the geographical domain. It is quite difficult to calculate the number of geographical entities stored in WordNet, due to the lack of an explicit annotation of the synsets, however we retrieved some figures by means the *has_instance* relationship, resulting in 654 cities, 280 towns, 184 capitals and national capitals, 196 rivers, 44 lakes, 68 mountains. Geographical resources like gazetteers usually contains a much greater quantity of information. For instance, the Geonet Names Server[1] (GNS) contains more than 5 million of place names.

In the following section we describe in detail our technique for the expansion of index terms; in section 3 we present and discuss the results obtained.

2 Expansion of Index Terms with WordNet

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 Lucene[2] search engine, an open source project freely available from Apache Jakarta. A Porter stemmer was used during the indexing phase, particularly the Snowball[3] implementation. The indexing process is performed by means of the Lucene search engine, generating two index for each text: 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[4], 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
```

[1] http://earth-info.nga.mil/gns/html/index.html
[2] http://lucene.jakarta.org
[3] http://snowball.tartarus.org/
[4] Freely available from the OpenNLP project: http://opennlp.sourceforge.net

```
-> western hemisphere, occident, New World
-> America
```

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* }. The result of the expansion of index terms is that the above text will be indexed also by terms like *Michigan*, *North America* that were not explicitly mentioned in it.

3 Experimental Results

This year we submitted four runs, two generated using the WordNet-based system and two with the system without the index terms expansion. 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 both systems the topic fields are analyzed in search of collocations (e.g. pairs "noun-noun" or "adjective-noun"). In 2005 we observed that this lead to worse results [3], however we consider this step necessary in order to identify correctly geographical entities such as *North America*, or *northern hemisphere* which are compound.

In Figure 1 we show the interpolated precision/recall graphs obtained for the "title and description" runs with the system that did not use the WordNet-based index term expansion (*rfiaUPV01*) and the WordNet-enhanced one (*rfiaUPV03*). Figure 2 contains the precision/recall graphs for the runs which included also the topic narrative. In this case *rfiaUPV04* is the run obtained with the WordNet-based system.

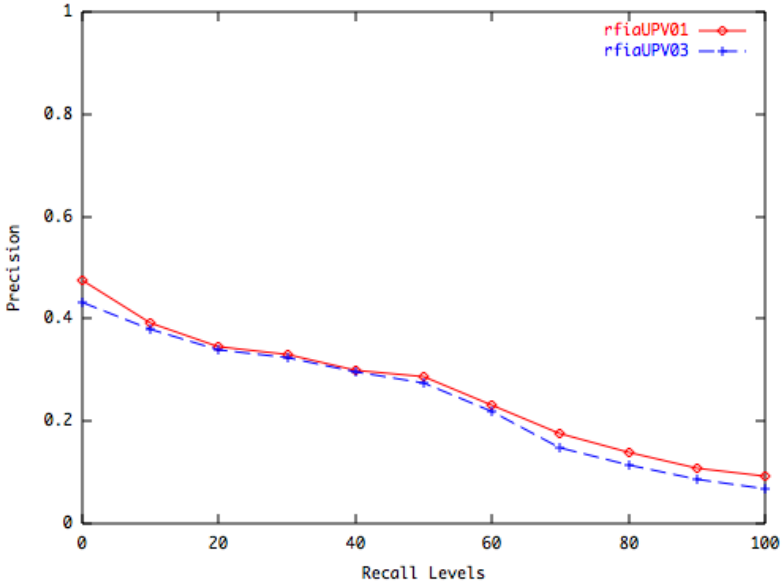

Figure 1: Interpolated precision/recall graph for the two "title and description" runs: *rfiaUPV01*, obtained using the system without WordNet, and *rfiaUPV03*, obtained using the system with WordNet

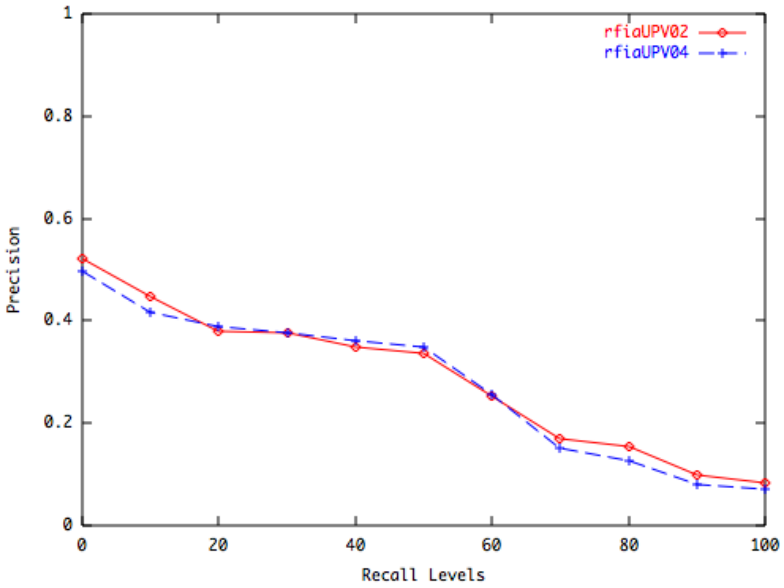

Figure 2: Interpolated precision/recall graph for the two "title, description and narrative" runs: *rfiaUPV02*, obtained using the system without WordNet, and *rfiaUPV04*, obtained using the system with WordNet.

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. Topic 40 does not contain any name of geographical place:

```
<EN-title> Cities near active volcanoes </EN-title>
<EN-desc> Cities, towns or villages threatened by the eruption of a volcano
</EN-desc>
```

Topic 48 contains references to places (*Greenland* and *Newfoundland*) for which WordNet does not provide many informations.

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. It was interesting to note that in topic 44 the difference between "title-desc" runs was os 6 documents retrieved in favour of the WordNet-based run, whereas the runs using all the topic fields obtained the same recall. This can be explained with the fact that the narrative of this topic contains a list of states that are meronyms of *Yugoslavia* (therefore they were indexed together with the holonym *Yugoslavia*).

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 an improvement in precision. Our next work directions will be the implementation of the same method with a richer (in terms of coverage of geographical places) resource such as the Getty Thesaurus of Geographical Names, or an ontology we are currently developing using the GNS and GNIS gazetteers together with WordNet itself and Wikipedia [2].

Acknowledgments

We would like to thank R2D2 CICYT (TIC2003-07158-C04-03) and ICT EU-India (ALA/95/23/2003/077-054) research projects for partially supporting this work.

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.