Using the WordNet Ontology in the GeoCLEF Geographical Information Retrieval Task

Davide Buscaldi, Paolo Rosso, and Emilio Sanchis Arnal

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

Abstract. This paper describes how we managed to use the WordNet ontology for the GeoCLEF 2005 English monolingual task. Both a query expansion method, based on the expansion of geographical terms by means of WordNet synonyms and meronyms, and a method based on the expansion of index terms, which exploits WordNet synonyms and holonyms. The obtained results show that the query expansion method was not suitable for the GeoCLEF track, while WordNet could be used in a more effective way during the indexing phase.

1 Introduction

Geographical entities can appear in very different forms in text collections. The problems of using text strings in order to identify a geographical entity are well-known and are related mostly to ambiguity, synonymy and names changing over time. Moreover, since in this case we are not using spatial databases, explicit information of regions including the cited geographical entities is usually missing from texts. 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 [2]. 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.

Automatic query expansion is used to add terms to the user's query. In the field of IR, the expansion techniques based on statistically derived associations have proven useful [3], while other methods using thesauri with synonyms obtained less promising results [4]. This is due to the ambiguity of the query terms and its propagation to their synonyms. The resolution of term ambiguity (Word Sense Disambiguation) is still an open problem in Natural Language Processing. Nevertheless, in the case of geographical terms, ambiguity is not as frequent as in the general domain (even if it still represents a major problem: for instance, 16 places named "*Genoa*" can be found in various locations all over the world: one in Italy, another in Australia and the remaining ones in the United States); therefore, better results can be obtained by the use of effective query

C. Peters et al. (Eds.): CLEF 2005, LNCS 4022, pp. 939–946, 2006.

expansion techniques based on ontologies, as demonstrated by the query expansion techniques developed for the SPIRIT project [5].

In our work we used the WordNet ontology only in the geographical domain, by applying a query expansion method, based on the synonymy and meronymy relationships, to geographical terms. The method is based on a similar one we previously developed using queries from the TREC-8[1] adhoc task [6]. 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. As a comparison, a specialized resource like the Getty Thesaurus of Geographic Names (TGN)[2] contains 3094 entities of type "city".

2 Query Reformulation

There can be many different ways to refer to a geographical entity. This may occur particularly for foreign names, where spelling variations are frequent (e.g. *Rome* can be indicated also with its original italian name, *Roma*), acronyms (e.g. *U.K.* or *G.B.* used instead of the extended form *United Kingdom of Great Britain and Northern Ireland*), or even some popular names (for instance, *Paris* is also known as the *ville lumiére*, i.e., the city of light). Each one of these cases can be reduced to the *synonymy* problem. Moreover, sometimes the rhetoric figure of *metonymy* (i.e., the substitution of one word for another with which it is associated) is used to indicate a greater geographical entity (e.g. *Washington* for *U.S.A.*), or the indication of the including entity is omitted because it is supposed to be well-known to the readers (e.g. *Paris* and *France*).

WordNet can help in solving these problems. In fact, WordNet provides synonyms (for instance, {*U.S., U.S.A., United States of America, America, United States, US, USA* } is the synset corresponding to the "*North American republic containing 50 states*"), and meronyms (e.g. *France* has *Paris* among its meronyms), i.e., concepts associated through the "part of" relationship.

Taking into account these observations, we developed a query expansion method in order to take advantage from these relationships. First of all, the query is tagged with POS labels. After this step, the query expansion is done in accordance to the following algorithm:

1. Select from the query the next word (w) tagged as proper noun.
2. Check in WordNet if w has the {*country, state, land*} synset among its hypernyms; if not, return to 1, else add to the query all the synonyms, with the exception of stopwords and the word w, if present; then go to 3.
3. Retrieve the meronyms of w and add to the query all the words in the synset containing the word *capital* in its gloss or synset, except the word *capital* itself. If there are more words in the query, return to 1, else end.

[1] http://trec.nist.gov

[2] http://www.getty.edu/research/conducting_research/vocabularies/tgn/

For example, the query: *Shark Attacks off Australia and California* is POS-tagged as follows: NN/shark, NNS/attacks, PRP/off, NNP/Australia CC/and NNP/California. Since "Shark" and "Attacks" do not have the {*country, state, land*} synset among their hypernyms, therefore Australia is selected as the next w. The corresponding WordNet synset is {*Australia, Commonwealth of Australia*}, with the result of adding "*Commonwealth of Australia*" to the expanded query. Moreover, the following meronym contains the word "capital" in synset or gloss: "*Canberra, Australian capital, capital of Australia - (the capital of Australia; located in southeastern Australia)*", therefore *Canberra* is also included in the expanded query. The next w is *California*. In this case the WordNet synset is {*California, Golden State, CA, Calif.*}, and the words added to the query are "*Golden State*", "*CA*" and "*Calif.*". The following two meronyms contain the word "capital":

- *Los Angeles, City of the Angels - (a city in southern California; motion picture capital of the world; most populous city of California and second largest in the United States)*
- *Sacramento, capital of California - (a city in north central California 75 miles northeast of San Francisco on the Sacramento River; capital of California)*

Moreover, during the POS tagging phase, the system looks for word pairs of the kind "adjective noun" or "noun noun". The aim of this step was to imitate the search strategy that a human would attempt. Stopwords are also removed from the query during this phase. Therefore, the expanded query that is handed over to the search engine is: *"shark attacks" Australia California "Commonwealth of Australia" Canberra "Golden State" CA Calif. "Los Angeles" "City of the Angels" Sacramento.*

For this work we used the Lucene[3] search engine, an open source project freely available from Apache Jakarta. The Porter stemmer [7] was used during the indexing phase, and for this reason the expanded queries are also stemmed by Snowball[4] before being submitted to the search engine itself.

3 Expansion of Index Terms

The expansion of index terms is a method that exploits the WordNet ontology in a somehow opposite way with respect to the query expansion. It is based on *holonyms* instead of meronyms, and uses synonyms too. 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

[3] http://lucene.jakarta.org
[4] http://snowball.tartarus.org/

the query contains "Houston", the city in Texas. The adopted weighting scheme is the usual *tf-idf*. The geographical terms in the text are identified by means of a Named Entity (NE) recognizer based on maximum entropy[5], and put into the *geo* index, together with all its synonyms and holonyms obtained from WordNet.

For instance, consider the following text:

"On Sunday mornings, the covered market opposite the station in the leafy suburb of Aulnay-sous-Bois - barely half an hour's drive from central Paris - spills opulently on to the streets and boulevards."

The NE recognizer identifies *Paris* as a geographical entity. A search for Paris synonyms in WordNet returns {*Paris, City of Light, French capital, capital of France*}, while its holonyms are:

```
-> France, French Republic
   -> Europe
       -> Eurasia
           -> northern hemisphere
           -> eastern hemisphere, orient.
```

Therefore, the following index terms are put into the *geo* index: {Paris, City of Light, French capital, capital of France, France, French Republic, Europe, Eurasia, northern hemisphere, eastern hemisphere, orient}. The result of the expansion of index terms is that the above text will be indexed also by words like *France*, *Europe* that were not explicitly mentioned in it.

4 Experimental Results

We submitted only the two mandatory runs, one using the topic title and description fields, and the second including the "concept" and "location" fields. For both runs only the query expansion method was used. For every query the top 1000 ranked documents have been returned by the system. We performed two runs, one with the unexpanded queries, the other one with expansion. For both runs we plotted the precision/recall graph (see Fig. 1) which displays the precision values obtained at each of the 10 standard recall levels.

The obtained results show that our system was the worst among the participants to the exercise [8]. The query expansion technique proved effective only in a few topics (particularly the topic number 16: "Oil prospecting and ecological problems in Siberia and the Caspian Sea"). The worst results were obtained for topic number 5 ("Japanese Rice Imports").

We suppose there are two main explanations for the obtained results: the first is that the keyword grouping heuristic was too simple: for instance, in topic number 5 the words are grouped as: "Japanese Rice" and "Imports", even if the topic description says: "Find documents discussing reasons for and consequences

[5] Freely available from the OpenNLP project: http://opennlp.sourceforge.net

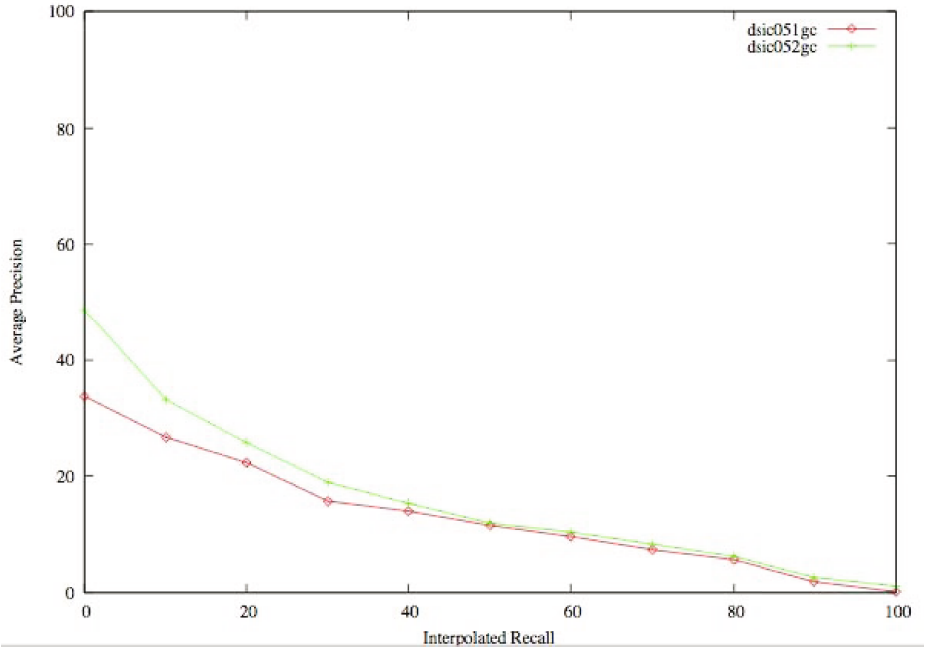

Fig. 1. Interpolated precision/recall graph for the two system runs: *dsic051gc*, using only the topic title and description fields, and *dsic052gc*, using also the "concept" and "location" fields

of the first imported rice in Japan". Therefore, in this case a better grouping should be "Japanese" and "Rice Imports".

Another reason could be that the expansion may introduce unnecessary information. For example, if the user is asking about "shark attacks in California", we have seen that *Sacramento* is added to the query. Therefore, documents containing "shark attacks" and "Sacramento" will obtain an higher rank, with the result that documents that contain "shark attacks" but not "Sacramento" are placed lower in the ranking. Since it is unlikely to observe a shark attack in Sacramento, the result is that the number of documents in the top positions will be reduced with respect to the one obtained with the unexpanded query, with the consequence of achieving a smaller precision.

In order to better understand the obtained results, we compared them with two baselines, the first obtained by submitting to the Lucene search engine the query without the synonyms and meronyms, and the latter by using only the tokenized fields from the topic. For instance, the query *"shark attacks" Australia California "Commonwealth of Australia" Canberra "Golden State" CA Calif. "Los Angeles" "City of the Angels" Sacramento* would be *"shark attacks" Australia California* for the first baseline (without WN) and *shark attacks Australia California* in the second case.

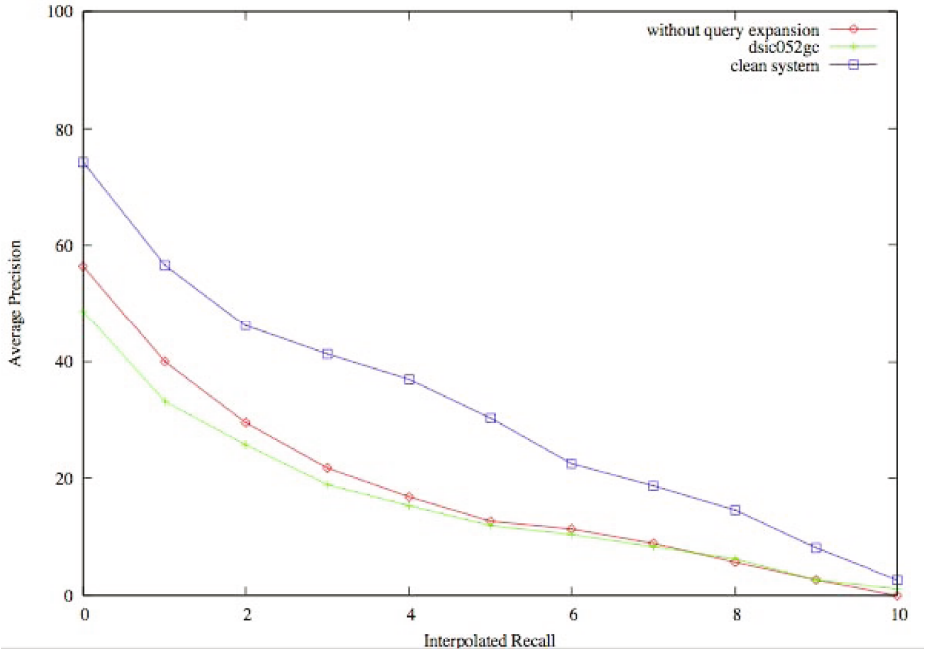

Fig. 2. Comparison of our best run (dsic052gc) with the "without query expansion" baseline and the clean system (neither query expansion nor keyword grouping)

The interpolated precision/recall graph in Fig. 2 demonstrates that both of our explanations for the obtained results are correct: in fact, the system using keyword grouping but not query expansion performs better than the system that uses both; however, this system is still worse than the one that do not use neither the query expansion nor keyword grouping.

The experiments carried out using the expansion of index terms method gave significantly better results than the query expansion, even if, due to the slowness of the indexing process (due principally to the Named Entity recognition), we were not able to send these runs for evaluation to the GeoCLEF; moreover, we were able to complete the indexing of the Glasgow Herald 1995 collection only. The topics (all-fields) were submitted to Lucene as for the simplest search strategy, but using the usual Lucene syntax for multi-field queries (e.g. all the geographical terms were labelled with "geo:"). The obtained results are displayed in Fig. 3.

We compared the results obtained with the standard search (i.e., no term was searched in the geo index). In order to make the difference between the two systems more comprehensible, the following string was submitted to Lucene for topic 1 when using the WordNet-enhanced search based on index term expansion: "text:shark text:attacks geo:california geo:australia", whereas in the case of the standard search method the submitted string was: "text:shark text:attacks text:california text:australia". It can be observed than the results obtained by

means of the expansion of index terms method are considerably better than those obtained using query expansion; however, a more detailed study needs to be carried out in order to verify if the results are also better than those obtained with the standard system.

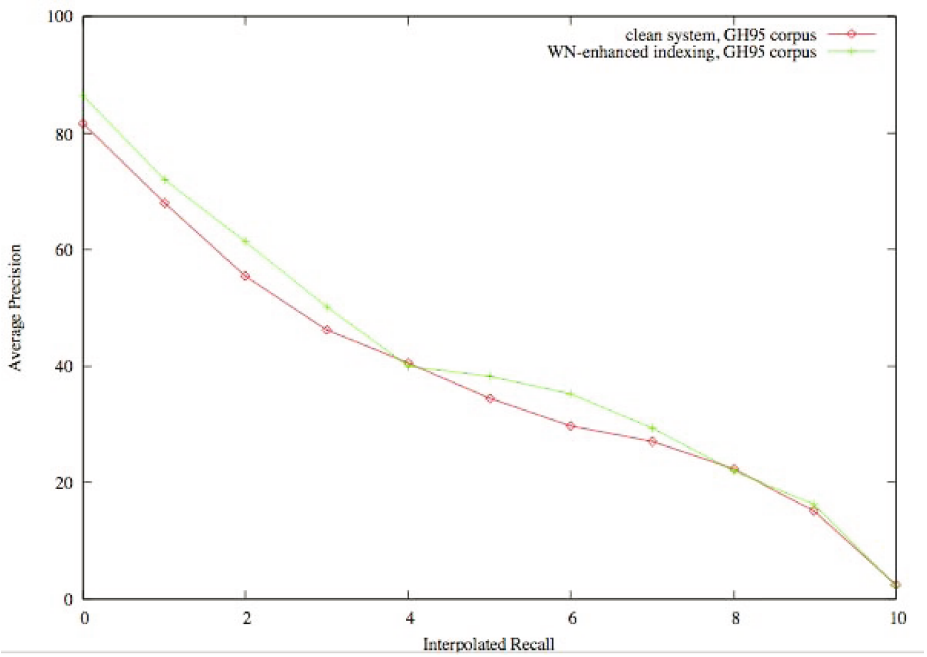

Fig. 3. Results obtained with the expansion of index terms method (*WN-enhanced indexing*), compared with the clean system baseline (indexing restricted to the Glasgow Herald 1995 collection)

5 Conclusions and Further Work

Our query expansion method was tested before only on a set of topics from the TREC-8 collection, demonstrating that a small improvement could be obtained in recall, but with a deterioration of the average precision. However, the results obtained in our participation at the GeoCLEF 2005 did not confirm the previous results. We believe that this was due to the different nature of the searches in the two exercises; more precisely, in theTREC-8 queries the geographical names usually represent political entities: "U.S.A.", "Germany", "Israel", for instance, are used to indicate the American, German or Israeli government (therefore the proposed query expansion method, which added to the query Washington, Berlin or Jerusalem, proved effective), while in GeoCLEF the geographical names just represent a location constraint for the users information needs. In such a context the use of WordNet during the indexing phase proved to be more effective, by

adding the synonyms and the holonyms of the encountered geographical entities to each documents index terms. Further work will include experiments over the whole collection with the expansion of index terms method, and a comparison of WordNet with a geographically specialized resource such as the Getty Thesaurus of Geographical Names.

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. We would like to thank also the CLEF 2005 organizing committee (this work is a revised version of the paper "A WordNet-based Query Expansion method for Geographical Information Retrieval").

References

1. Bo-Yeong, K., Hae-Jung, K., Sang-Lo, L.: Performance analysis of semantic indexing in text retrieval. In: CICLing 2004, Lecture Notes in Computer Science, Vol. 2945, Mexico City, Mexico (2004)
2. Rosso, P., Ferretti, E., Jiménez, D., Vidal, V.: Text categorization and information retrieval using wordnet senses. In: CICLing 2004, Lecture Notes in Computer Science, Vol. 2945, Mexico City, Mexico (2004)
3. Xu, J., Croft, W.: Query expansion using local and global document analysis. In: Proceedings of the ACM SIGIR 1996, New York, USA (1996)
4. Voorhees, E.: Query expansion using lexical-semantic relations. In: Proceedings of the ACM SIGIR 1994. (1994)
5. Fu, G., Jones, C., Abdelmoty, A.: Ontology-based spatial query expansion in information retrieval. In: Proceedings of the ODBASE 2005 conference. (2005)
6. Calcagno, L., Buscaldi, D., Rosso, P., Gomez, J., Masulli, F., Rovetta, S.: Comparison of indexing techniques based on stems, synsets, lemmas and term frequency distribution. In: III Jornadas en Tecnología del Habla, Valencia, Spain (2004)
7. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Wokingham,UK (1999)
8. Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P.: Geoclef: the clef 2005 cross-language geographic information retrieval track. In: Working notes for the CLEF 2005 Workshop (C.Peters Ed.), Vienna, Austria (2005)