A WordNet-based Query Expansion method for Geographical Information Retrieval

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

August 31, 2005

Abstract

This report describes a query expansion method based on the expansion of geographical terms by means of WordNet synonyms and meronyms. We used this method for our participation to the GeoCLEF 2005 English monolingual task, while using the well-known Lucene search engine for indexing and retrieval. The obtained results show that the proposed method was not suitable for the GeoCLEF track, while WordNet can be used in a more effective way during the indexing phase, by adding synonyms and holonyms to the index terms.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; I.2 [**Artificial Intelligence**]: I.2.7 Natural Language Processing

General Terms

Measurement, Performance, Experimentation

Keywords

Query Expansion, WordNet, Geographical Information Retrieval

1 Introduction

Geographical entities can appear in very different forms in text collections, such as when a foreign name is used instead of the English one, or when the citation of some region or place omits the name of a larger geographical entity containing them. This is a well-known problem in the field of Information Retrieval. The use of semantic knowledge may help to solve this problem, even if no strong experimental results are yet available in support of this hypothesis. Some results [7] show improvements by the use of semantic knowledge, others do not [3]. 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 [6], while other methods using thesauri with synonyms obtained less promising results [5]. 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, the resolution of ambiguity is usually less difficult and therefore better results can be obtained by the use of effective query expansion techniques based on ontologies, as demonstrated by the query expansion techniques developed for the SPIRIT project [2].

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 for the use with the TREC-8[1] adhoc task.

2 Query Reformulation

There can be many different ways to refer to a geographical entity. This may occur specially for foreign names (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 lumière*, 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 ({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 that exploits 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 stop-words 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.

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. "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)*, and the result is that *Canberra* is 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.*. Two meronyms contains 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)*

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

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 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[2] search engine, an open source project available for free download from Apache Jakarta. The Porter stemmer [1] was used during the indexing phase, and for this reason the expanded queries are also stemmed by the *SnowballAnalyzer* (provided with the Lucene API) before being submitted to the search engine itself.

3 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 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 the expansion. For both runs we plotted the precision/recall graph (see Figure 1) which displays the precision values obtained at each of the 10 recall levels.

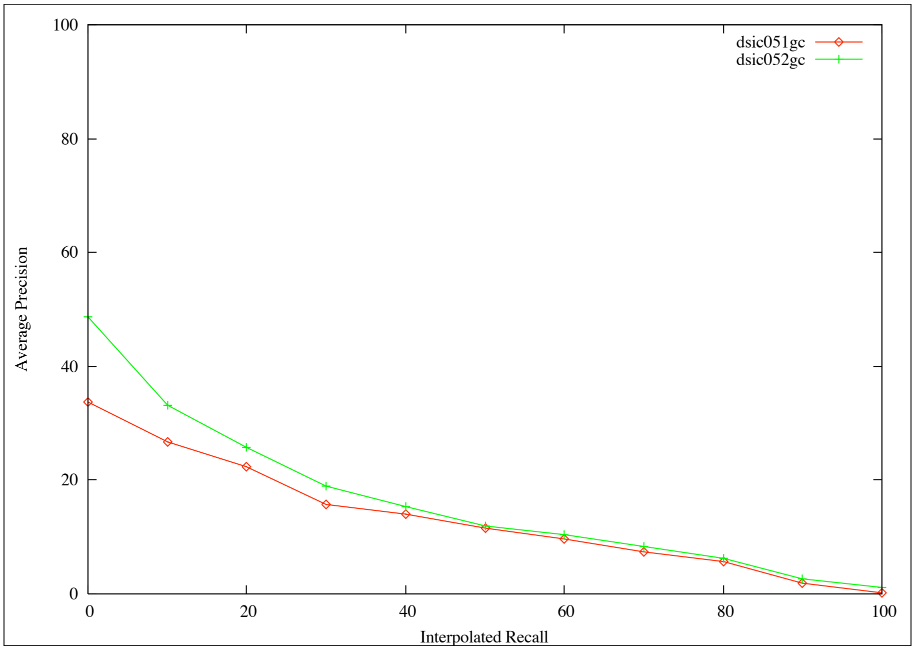


Figure 1: Precision/Recall graphs obtained for the two submitted runs, the *dsic051gc*, using only topic title and description, and the *dsic052gc*, which uses also the "concept" and "location" fields.

It is clear that the results obtained by taking into account the "concept" and "location" fields are generally better than those obtained working only on the topic title and definition, as can be observed also by the results obtained over the single topics in Table 1.

The obtained results show that the performance of our system stands around the average of the participants to the exercise. From these results the advantage of the query expansion method is not clear, even if it proved effective in some topics (particularly 16 and 7). We suppose this is due to the fact 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

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

Topic	dsic051gc	dsic052gc	Topic	dsic051gc	dsic052gc
001	9.23%	34.61%	014	5.71%	5.12%
002	5.24%	5.38%	015	65.08%	70.03%
003	14.71%	4.11%	016	60.82%	56.76%
004	0.03%	0.04%	017	20.06%	20.72%
005	0.31%	12.47%	018	0.52%	1.62%
006	11.17%	2.10%	019	0.45%	0.81%
007	4.45%	19.82%	020	1.26%	13.26%
008	0.14%	0.27%	021	9.83%	10.44%
009	30.68%	32.42%	022	15.76%	17.06%
010	7.94%	10.52%	023	1.13%	0.80%
011	0.00%	0.01%	024	31.82%	22.00%
012	0.18%	3.81%	025	0.00%	0.00%
013	12.54%	21.85%			

Table 1: Average precision obtained for each topic for the two submitted runs.

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, thus resulting in achieving a smaller precision.

In order to evaluate in a more precise way the query expansion, we compared the results obtained 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.

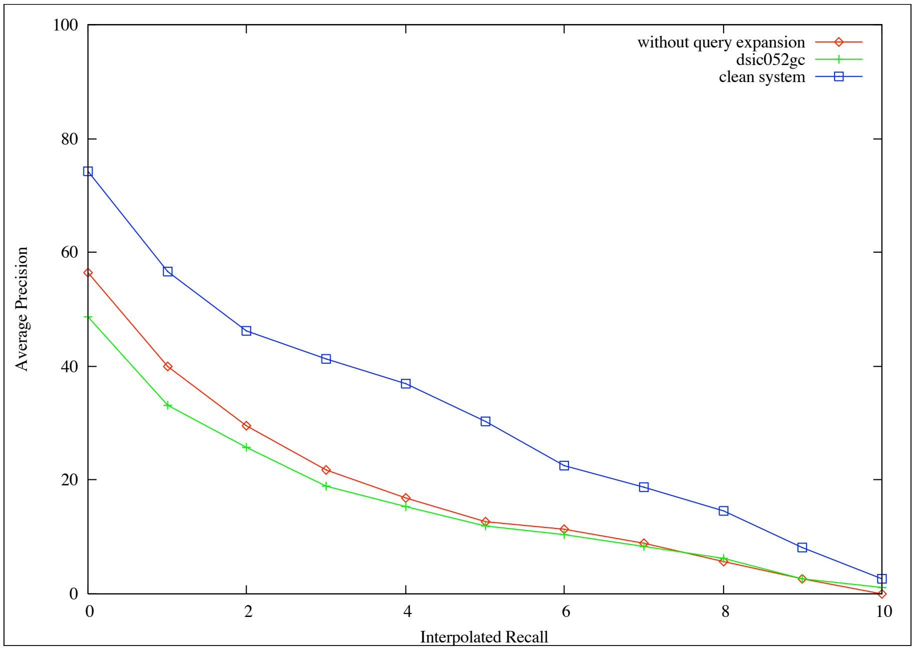
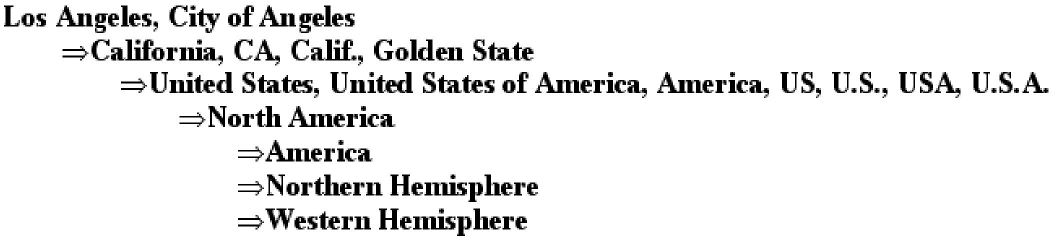


Figure 2: Comparison of the results obtained for the run *dsic052gc* with two baselines: the unexpanded query (*without query expansion*) and the terms without any kind of preprocessing but the removal of stopwords (*clean system*).

Figure 2 shows that the query expansion not only did not prove useful, but was even worst than the simplest search strategy, which however was not used for the GeoCLEF experiments.

However, we still investigated another method to use WordNet in the geographical retrieval task. We performed a late experiment by using WordNet synonyms and holonyms (holonymy is the inverse relationship of meronymy) during the indexing phase. In detail, an additional indexing field (*geo*) was used together with the standard text indexing. The Named Entities detector based on maximum entropy from the *openNLP* project [3] was used in order to individuate location names during the indexing phase (this seemed to be faster and more reliable to us than looking into WordNet to find if one of the word senses has the *location* synset among its hypernyms). For every location name l, the synonyms of l and all its holonyms (even the inherithed ones) are added to the *geo* index. For instance, if *Los Angeles* is found in the text, the synonym "City of the Angels" is added to the *geo* index, together with the holonyms: {*California, Golden State, CA, Calif.*} ,{*United States, United States of America, America, US, U.S., USA, U.S.A.*}, {*North America*} , {*America*, {*Northern Emisphere*} and {*Western Hemisphere*. A more clear view of the holonym tree obtained from WordNet is shown in Figure 3.



Los Angeles, City of Angeles
⇒California, CA, Calif., Golden State
⇒United States, United States of America, America, US, U.S., USA, U.S.A.
⇒North America
⇒America
⇒Northern Hemisphere
⇒Western Hemisphere

Figure 3: The holonym tree obtained from WordNet for *Los Angeles*.

Due to the slowness of the process, we completed only the indexing of the *Glasgow Herald 1995* collection in time for this paper. 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 Figure 4.

Even in this case we compared the obtained results with the standard search (i.e., no term was searched in the *geo* index), as for the baseline obtained by using only the tokenized fields from the topic. In order to clarify the difference, the following string was submitted to Lucene for the WordNet-enhanced search: "text:shark text:attacks geo:california geo:australia", while for the standard search the submitted string was: "text:shark text:attacks text:california text:australia". It is clear that this method gives much better results than the query expansion, and even better than the baseline.

4 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 [4]. The results obtained for our participation at the GeoCLEF do not confirm the previous results. We believe that this is due to the different nature of the searches in the two exercises; more precisely, in the TREC-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 user's information needs. In such a context the use WordNet during the indexing phase proved to be more effective, by adding the synonyms and the holonyms of the encountered geographical entities to each document's index terms. We plan

[3] http://opennlp.sourceforge.net

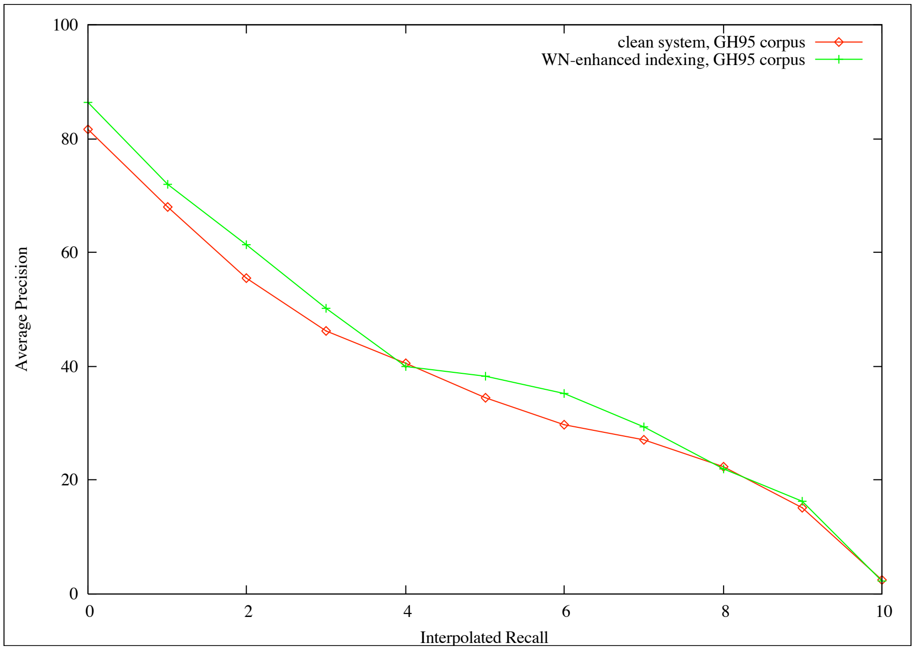


Figure 4: Results obtained over the GH95 corpus with the clean system and the WordNet-enhanced indexing based on synonyms and holonyms.

to perform more experiments on both the whole GeoCLEF and TREC-8 collections in order to verify the effectiveness of this indexing method.

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] Baeza-Yates, R., Ribeiro-Neto, B., “Modern Information Retrieval”, Addison-Wesley, 1999.

[2] Fu, G., Jones, C.B., Abdelmoty, A.I., “Ontology-based Spatial Query Expansion in Information Retrieval”, ODBASE 2005, accepted.

[3] Rosso, P., Ferretti, E., Jiménez, D., Vidal, V., “Text Categorization and Information Retrieval Using WordNet Senses”, CICLing 2004, Lecture Notes in Computer Science, Vol. 2945. Springer-Verlag, 2004

[4] Calcagno L., Buscaldi D., Rosso P., Gómez J. M., Masulli F., Rovetta S., “Comparison of Indexing Techniques based on Stems, Synsets, Lemmas and Term Frequency”. In: Workshop “Red Temática en Tecnología del Habla, Valencia, Spain (2004), pp. 171-176.

[5] Voorhees, E.M., “Query Expansion using lexical-semantic relations”, ACM SIGIR 1994: pp.61-70, ACM, 1994

[6] Xu, J., W.B. Croft, “Query Expansion using Local and Global Document Analysis”, ACM SIGIR 1996: pp.4-11, ACM, 1996

[7] Bo-Yeong , K., Hae-Jung, K., Sang-Jo, L., "Performance Analysis of Semantic Indexing in Text Retrieval", CICLing 2004, Lecture Notes in Computer Science, Vol. 2945. Springer-Verlag, 2004