WordNet as a Geographical Information Resource

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

Abstract

Geographical entities often appears 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 known problem in the field of Information Retrieval. The use of an ontology like WordNet can help in addressing this issue. In this paper we propose an automatic method to expand the geographical terms in queries by using the WordNet ontology and another method that expands the terms during the indexing phase. The proposed methods exploits the synonymy, meronymy and holonymy relationships provided by WordNet, together with some information extracted from the gloss.

1 Introduction

One of the major problems in Information Retrieval (IR) is that the terms used in the query by the user may be different from those terms that describe the same concept in the database. The research community is constantly increasing its effort to prove that semantic knowledge may helps to solve this problem. Nowadays, no strong experimental results are yet available in support of this hypothesis. Some results [Kang et al., 2004] show improvements by the use of semantic knowledge, others do not [Rosso et al., 2003]. The most important Information Retrieval conferences (SIGIR, TREC) show the predominance 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 users query. In the field of IR, the expansion techniques based on statistically derived associations have proven useful [Xu and Croft, 1996], while other methods using thesauri with synonyms obtained less promising results [Voorhees, 1994]. 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 easier 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 [Gaihua Fu and Abdelmoty, 2005]. Moreover, the retrieval of information involving some kind of spatial awareness is obtaining an increasing interest by the IR researchers, as testified by the creation of the GeoCLEF[1] evaluation exercise at the CLEF 2005.

In our work we investigated the use of the WordNet ontology only in this specific domain, by applying both a query expansion method, based on the synonymy and meronymy relationships, and a term expansion method based on holonymy relationship to geographical terms.

2 Query Reformulation

Many geographical entities are usually referred to in different ways. This is true specially for foreign names (e.g. *Rome* can be indicated also with its original italian name, *Roma*), acronyms (e.g. *U.S.A.* or *U.S.* are usually preferred to the extended form *United States of America*), or even some historical names (again, Rome is also known as *the eternal city*). 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 by means of the relationship "part of".

Taking into account these observations, we developed a query expansion method that exploits these relationships. First of all, the query is POS-tagged using the SVMTool POS tagger [Giménez and Márquez, 2004]. After this step, the query expansion is performed as follows:

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

[1] http://ir.shef.ac.uk/geoclef2005/

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 37–42.

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: *foreign minorities, Germany* is POS-tagged as follows: JJ/foreign, NNS/minorities, NNP/Germany. Therefore, *Germany* is selected as w. The corresponding WordNet synset is *Germany, Federal Republic of Germany, Deutschland, FRG* and, since its hypernyms include the *country, state, land* synset, the following synonyms of Germany are added to the expanded query: *Federal Republic, Deutschland, FRG*. The following meronyms contains the word "capital" in synset or gloss:

- `Berlin, German capital (capital of Germany located in eastern Germany)`
- `Bonn (a city in western Germany on the Rhine River; was the capital of West Germany between 1949 and 1989)`
- `Munich, Muenchen (the capital and largest city of Bavaria in south eastern Germany)`
- `Aachen, Aken, Aix-la-Chapelle (a city in western Germany near the Dutch and Belgian borders; formerly it was Charlemagnes northern capital)`

Therefore, the resulting expanded query is: *foreign minorities, Germany Federal Republic Deutschland FRG Berlin German Bonn Munich Muenchen Aachen Aken Aix la-Chapelle*.

3 Index Terms Expansion

We investigated also a method based on the expansion of the geographical terms in the documents, that uses WordNet's synonymy and holonymy relationships. This method can be considered the "inverse" of the previous exposed one, since it is based on the inverse relationships of the ones used for the query expansion (synonymy is simmetric, while holonymy is the inverse of meronymy).

In this method, an additional indexing field (*geo*) is used together with the standard text indexing. The named entities detector based on maximum entropy from the *openNLP* project[2] was used in order to individuate names of geographical entities during the indexing phase. Since the tool does not perform a classification of the named entities, we developed the following heuristics in order to identify the geographical ones: when a Named Entity is detected, we look in WordNet if one of the word senses has the "location" synset among its hypernyms. If this is true, then the entity is considered a geographical one.

For every geographical location 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*}. The obtained holonyms tree is:

```
Los Angeles, City of Angels
=>California, Golden State, CA, Calif.
 =>United States, US, U.S., U.S.A.
  =>North America
   =>America
   =>Northern Hemisphere
   =>Western Hemisphere
```

The advantage of this method is that knowledge about the enclosing, broader, geographical entities is stored together with the index term. Therefore, searches addressing, for instance, *California*, will match with documents where the names *San Francisco*, *Sacramento*, *Los Angeles*, *Pasadena*, etc. appear, even if "California" is not mentioned explicitly in the documents.

4 IR Systems and Test Collections

We used two IR systems for the experiments: the Furbo IR system, a Python[3] implementation of Smart [Salton, 1971], and the well-known Lucene[4] open source search engine. For the indexing process Furbo uses a vectorial space model [Baeza-Yates and Ribeiro-Neto, 1999] with the usual TF-IDF (Term Frequency - Inverse Document Frequency) weighting. Lucene was used with the default settings. Both the engines use stems. The indexed terms were passed through a stemming process, done by replacing each non-stopword term with its stemmed form, obtained by using the Snowball implementation of Porter's algorithm [Baeza-Yates and Ribeiro-Neto, 1999].

Three experiments with different set-ups have been carried out (see Table 2 for details); for the first one we used Furbo as IR system and a subset of the TREC-8[5] adhoc task collection. Table 1 resumes the document collection statistics. The selected portion was the FBIS collection, containing 130471 documents.

Fifty queries, extracted from the TREC-8 adhoc task topics numbered from 401 to 450, have been used to evaluate the query expansion method against the FBIS collection. Each query is constituted by the title of the related topic. For each topic there is a list of relevant documents, which have been extracted from the TREC-8 relevance judgments by excluding the judgments of the documents not indexed by Furbo. The unjudged documents are assumed to be not relevant. The provided relevance judgments give 1667 relevant documents in the FBIS collection for the 50 topics used for the experiments.

[2] `http://opennlp.sourceforge.net`

[3] `http://www.python.org`

[4] `http://lucene.jakarta.org`

[5] `http://trec.nist.gov`

Table 1: TREC-8 adhoc task document collection statistics.

Source	# of Docs
The Financial Times (FT), 1991-1994	210158
Federal Register (FR), 1994	55630
Foreign Broadcast Information Service (FBIS)	130471
Los Angeles Times (LA94), 1994	131896
Total	528155

The query expansion method was also tested on the GeoCLEF 2005 test collections, which include the *Los Angeles Times*, year 1994, and the *Glasgow Herald* (GH95), 1995. The query set was the entire GeoCLEF 2005 set of queries, the *topic* and *description* fields were used. The experiment with the index terms expansion was performed only on the Glasgow Herald collection. In both the GeoCLEF experiments we used the Lucene search engine as IR system.

5 Results

In Table 2 we show a resume of the configurations (queries, search engine, collections) used for the experiments. In all three experiments, the top 1000 ranked documents have been returned by the system.

Table 2: Experiments configuration. `TREC-AdHoc-QE`: experiment with Query Expansion, using queries from the TREC 8 AdHoc task; `GeoCLEF-QE`: experiment with Query Expansion, using queries from the GeoCLEF 2005 task; `GeoCLEF-ITE`: experiment with Index Terms Expansion, queries from the GeoCLEF 2005.

Experiment	IR Syst.	Collection
`TREC-AdHoc-QE`	Furbo	FBIS
`GeoCLEF-QE`	Lucene	LA94, GH95
`GeoCLEF-ITE`	Lucene	GH95

In the `TREC-AdHoc-QE` experiment, we performed two runs, one with the unexpanded queries, the other one with expansion. For both runs we plotted the precision/recall graph (see Figure 1) which displays the precision values obtained at each of the standard recall levels. Recall indicates the number of relevant documents returned over the total number of relevant documents, while precision can be viewed as a measure of the quantity of relevant documents that can be found in the top positions of the result list. In Table 3 we show the average recall and precision values for the two runs (with and without Query Expansion).

Table 3: Average recall and precision for the runs of the TREC-AdHoc-QE experiment, with Query Expansion (QE) and without (¬QE).

	QE	¬QE
Avg. Recall	67.79%	65.63%
Avg. Precision	14.29%	15.24%

The obtained results show that the overall recall improved by $\sim$ 2% when using the query expansion method, although precision was lower than the one obtained without the query expansion. This is due to the fact that the expansion may introduce unnecessary information, resulting in a worst ranking of the really relevant documents. For example, if the user is asking about "shark attacks in California", since Sacramento is the capital of California, it 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.

The second experiment (`GeoCLEF-QE`) was part of our participation in the GeoCLEF 2005 task. In this case, during the POS tagging phase, the system looks for word pairs of the kind "adjective noun" or "noun noun". The aim of the inclusion of additional step was both to imitate the search strategy that a human would attempt, and to handle longer queries than those of the TREC-8 AdHoc task. Stopwords are also removed from the query during this phase. Therefore, given the query #1 of the GeoCLEF: "Shark Attacks off Australia and California" the terms (without expansion) handed over to the search engine are: "shark attacks", "Australia" and "California". In order to evaluate in a more precise way the query expansion, we compared the results of this experiment 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. The obtained graph is showed in Figure 2. In this case it is quite obvious that the query expansion not only did not prove useful, but was even worst than the simplest search strategy, which however was not used in our GeoCLEF participation. The results of the task [Gey et al., 2005] prove that our system was one of the worst ones.

We suppose that the reason of such bad results was both the use of the "description" field of the topics, that is longer that the "title" field that was used for the TREC-AdHoc-QE experiment, and the different nature of the queries in the TREC and GeoCLEF. For instance, "Foreign minorities in Germany", "Ireland peace talks." are TREC-8 topics, where *Germany* and *Ireland* can be considered more *political* entities rather than geographical ones: in these cases looking for the capital of a country is more effective, since it is often used as a reference to its government. In fact, the user is interested in the behaviour of the government of a certain country with regards to a specific matter. On the other hand, in tipical GeoCLEF queries such "Shark attacks off California and Australia", "Walking holidays in Scotland", the geographical names are actually related to a region, and there is no supposed correlation between the regions and their capital (as observed before in the case of Sacramento) –

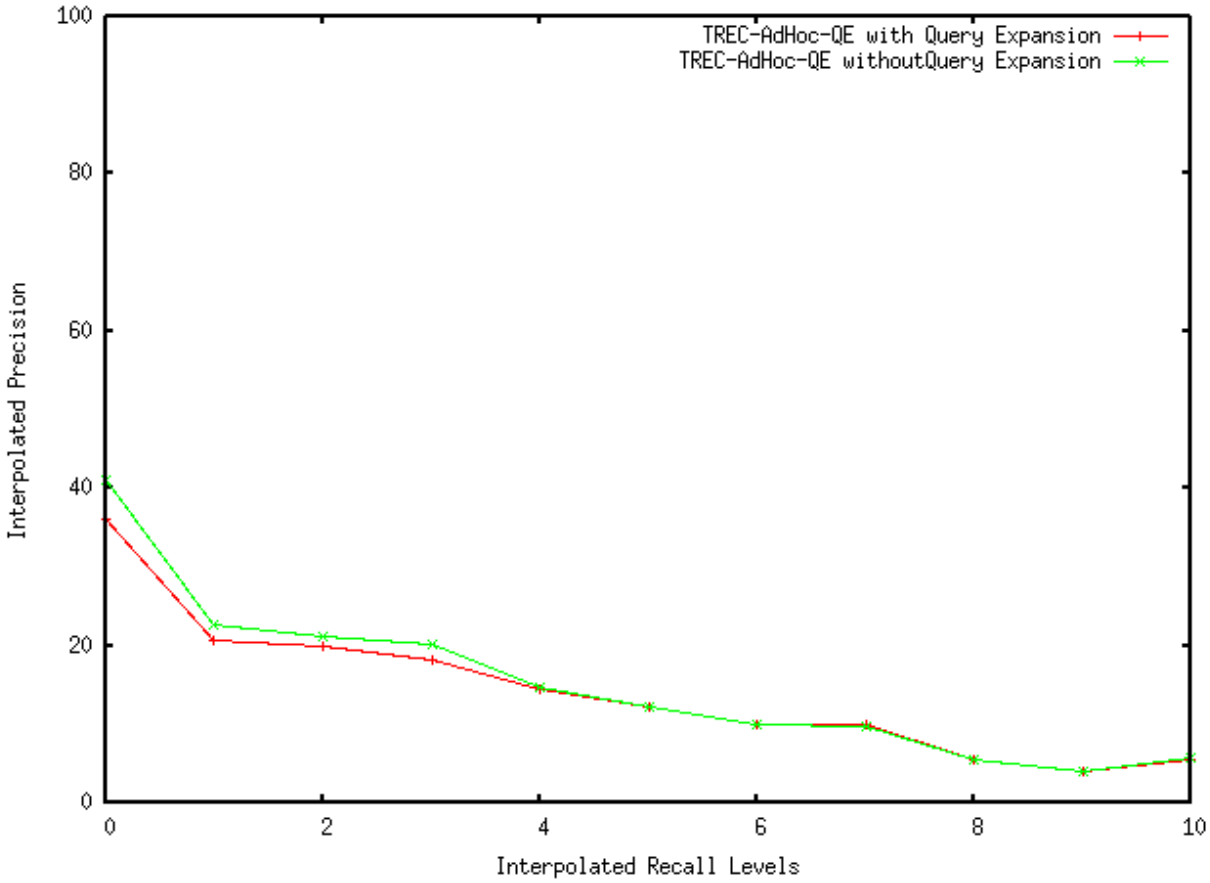

Figure 1: Precision/Recall graphs obtained with query expansion and without, for the two runs of the TREC-AdHoc-QE experiment.

the user in this case is interested in finding information with geographical restrictions.

This observation constituted the starting point for our last experiment `GeoCLEF-ITE`, in which WordNet holonyms were used during the indexing phase. The method has been described in section 3. 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 were submitted to Lucene as for the simplest search strategy (tokenization of "'title" and "description" fields and removal of stopwords), using the usual Lucene syntax for multi-field queries (i.e., all the geographical terms were labelled with the index label "geo:"). 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". The obtained results are showed in Figure 3. It can be observed that, although only a part of the collection was used, this method gives much better results than the query expansion, and even better than the baseline.

6 Conclusions

Our experiments demonstrate that WordNet can be used effectively as a resource for the Geographical Information Retrieval task. The results obtained with the TREC-8 AdHoc set of topics were not fully satisfactory, and the bad performance in the GeoCLEF task seems to confirm that the effectiveness of query expansion techniques depends on domain and on the task. However, the results obtained with the expansion of the index terms by means of holonyms and synonyms are interesting and worth further investigation. Future work will include an evaluation of the index term expansion method using the complete GeoCLEF collection and the TREC-8 topics. It will be also interesting to evaluate WordNet against specialized ontologies like the Getty Thesaurus of Geographical Names[6] (TGN).

[6] `http://www.getty.edu/research/tools/vocabulary/tgn/`

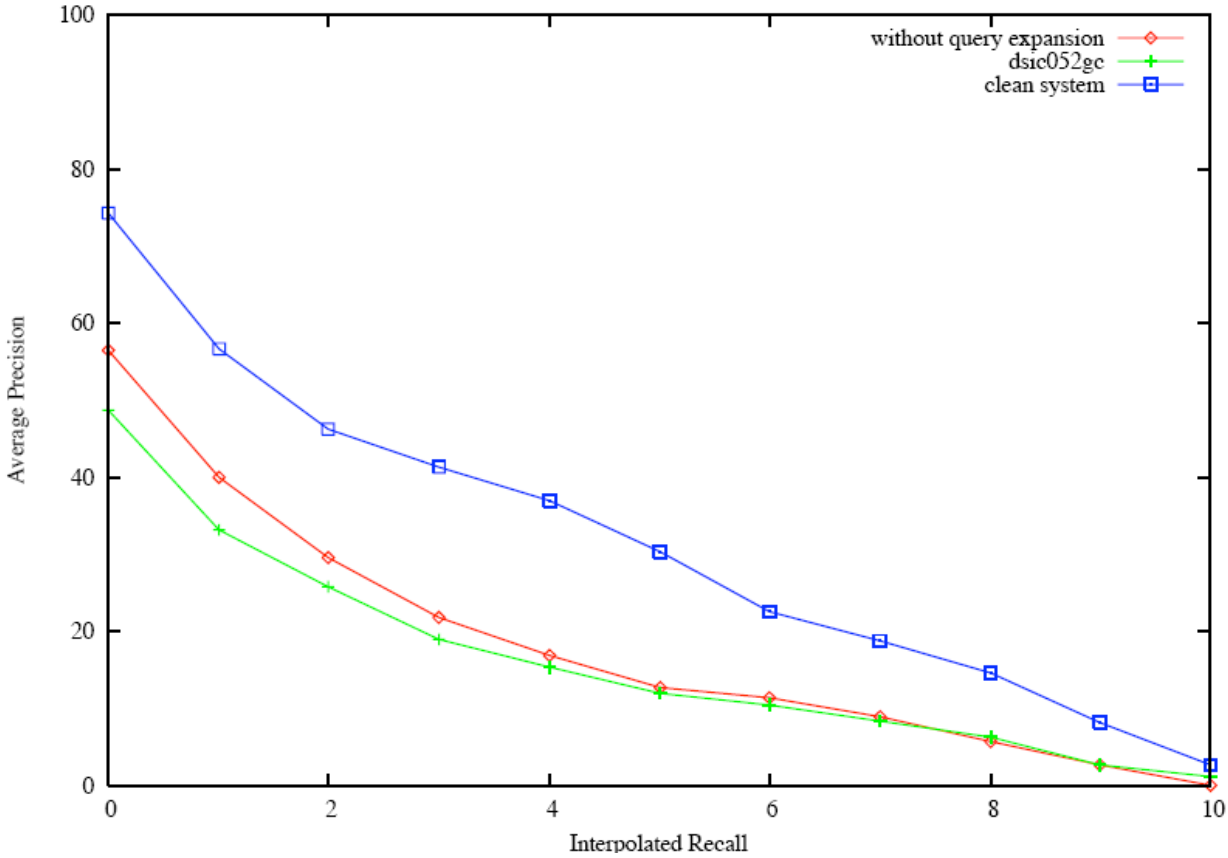

Figure 2: Precision/Recall graphs obtained with query expansion (dsic052gc), without query expansion and without query text analysis (clean system) for the GeoCLEF-QE experiment.

Acknowledgements

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. Special thanks to Matteo DellAmico for implementing the Furbo IR system.

References

Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press / Addison-Wesley.

Christopher B. Jones Gaihua Fu and Alia I. Abdelmoty. 2005. Ontology-based spatial query expansion in information retrieval. In Robert Meersman and Zahir Tari, editors, *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*, volume 3761 of *Lecture Notes in Computer Science*, Agia Napa, Cyprus, October. Springer Verlag.

Fredric Gey, Ray Larson, Mark Sanderson, Hideo Joho, and Paul Clough. 2005. Geoclef: the clef 2005 cross-language geographic information retrieval track. In *Proceedings of the CLEF2005 workshop*.

Jesús Giménez and Lluís Márquez. 2004. Svmtool: A general pos tagger generator based on support vector machines. In *Proceedings of the 4th LREC*.

Bo-Yeong Kang, Hae Jung Kim, and Sang-Jo Lee. 2004. Performance analysis of semantic indexing in text retrieval. In Alexander F. Gelbukh, editor, *CICLing*, volume 2945 of *Lecture Notes in Computer Science*, pages 433–436. Springer.

Paolo Rosso, Edgardo Ferretti, Daniel Jiménez, and Vicente Vidal. 2003. Text categorization and information retrieval using wordnet senses. pages 299–304, Brno, Czech Republic, December. Masaryk University Brno, Czech Republic.

Gerard Salton. 1971. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice hall Inc., Englewood Cliffs, NJ.

Ellen M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69, New York, NY, USA. Springer-Verlag New York, Inc.

Jinxi Xu and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR*

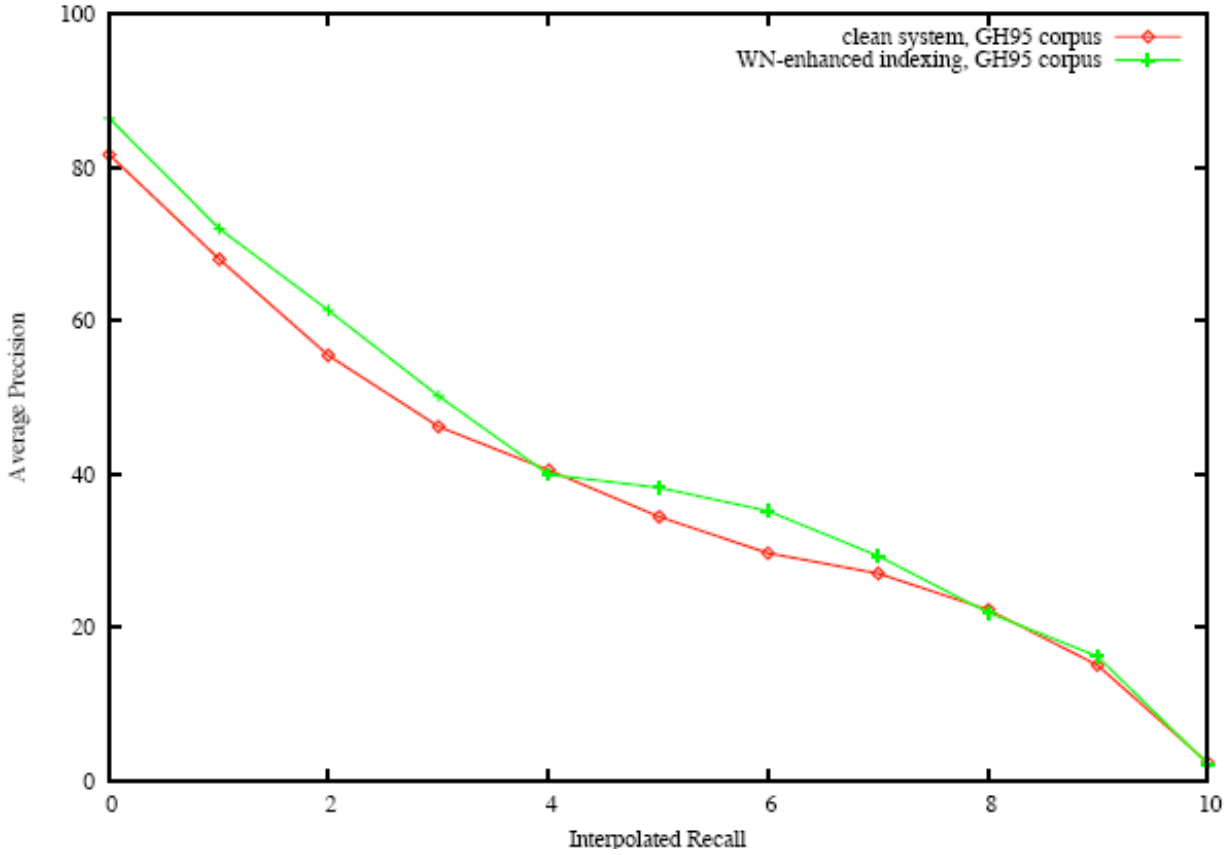

Figure 3: Comparison of the results obtained over the GH95 corpus with the clean system and the Index Term Expansion method based on synonyms and holonyms.

conference on Research and development in information retrieval, pages 4–11, New York, NY, USA. ACM Press.