On Expanding Query Vectors with Lexically Related Words

Ellen M. Voorhees
Siemens Corporate Research, Inc.
755 College Road East
Princeton, NJ 08540
ellen@learning.scr.siemens.com

Abstract

Experiments performed on small collections suggest that expanding query vectors with words that are lexically related to the original query words can improve retrieval effectiveness. Prior experiments using WordNet to automatically expand vectors in the large TREC-1 collection were inconclusive regarding effectiveness gains from lexically related words since any such effects were dominated by the choice of words to expand. This paper specifically investigates the effect of expansion by selecting query concepts to be expanded by hand. Concepts are represented by WordNet synonym sets and are expanded by following the typed links included in WordNet. Experimental results suggest that this query expansion technique makes little difference in retrieval effectiveness within the TREC environment, presumably because the TREC topic statements provide such a rich description of the information being sought.

1 Introduction

The IR group at Siemens Corporate Research is investigating how *concept spaces* — data structures that define semantic relationships among ideas — can be used to improve retrieval effectiveness in systems designed to satisfy large-scale information needs. As part of this research, we expanded document and query vectors automatically using selected synonyms of original text words for TREC-1 [5]. The retrieval results indicated that this expansion technique improved the performance of some queries, but degraded the performance of other queries. We concluded that improving the consistency of the method would require both a better method for determining the *important* concepts of a text and a better method for determining the correct sense of an ambiguous word.

We took TREC-2 as an opportunity to investigate the effectiveness of vector expansion when good concepts are chosen to be expanded. As in TREC-1, query vectors were expanded using WordNet synonym sets. However, the synonym sets associated with each query were selected manually (by the author). These results therefore represent an upper-bound on the effectiveness to be expected from a completely automatic expansion process.

The results of the TREC-2 evaluation indicate that the query expansion procedure used does not significantly affect retrieval performance even when important concepts are identified by hand. Some expanded queries are more effective than their unexpanded counterparts, but for other queries the unexpanded version is more effective. In either case, the effectiveness difference between the two versions is seldom large. Further testing suggests that more extreme expansion procedures can cause larger differences in retrieval performance, but the net effect over a set of queries is degraded performance compared to no expansion at all.

The remainder of the paper discusses the experiments in detail. The next section describes the retrieval environment, including a description of WordNet. Section 3 provides evaluation results for both the official TREC-2 runs and some additional supporting runs. The final section explores the issue of why the expansion fails to improve retrieval performance.

2 The Retrieval Environment

The expansion procedure used in this work relies heavily on the information recorded in WordNet, a manually-constructed lexical system developed by George Miller and his colleagues at the Cognitive Science Laboratory at Princeton University [4]. WordNet's basic object is a set of strict synonyms, called a *synset*. Synsets are organized by the lexical relations defined on them, which differ depending on part of speech. For nouns, the only part of WordNet used in this study, the lexical relations include antonymy, hypernymy/hyponymy (*is-a* relation) and three different meronym/holonym (*part-of*) relations. The *is-a* relation is the dominant relationship, and organizes the synsets into a set of approximately ten hierarchies[1]. Figure 1 shows a piece of WordNet. The figure contains all the ancestors and descendents as defined by the *is-a* relation for the six senses of the noun *swing*. Also shown is that one of the senses, a child's toy, is *part-of* a playground.

Given a synset, there is a wide choice of words to add to a query vector — one can add only the synonyms within the synset, or all descendents in the *is-a* hierarchy, or all words in synsets one link away from the original synset regardless of link type, etc. One of the goals of this work is to discover which such strategies are effective. Wang et al. found expanding vectors from relational thesauri to be effective [6], but based those conclusions on experiments performed on one small collection. Experiments we performed as part of our TREC-1 work showed showed serious degradation when anything other than synonyms were used in the expansion — but the TREC-1 results were dominated by the problem of finding good synsets to expand. This work examines the effectiveness of the different relation types assuming good synsets are used as the basis.

Siemens' official TREC-2 runs consist of one routing run (topics 51–100 against the documents on disk three) and two ad hoc runs (topics 101–150 against the documents on the first two disks). All of the runs are manual since the input text of the topics was modified by hand. There are two types of modifications: parts of the topic statement that explicitly list things that are not relevant were removed, and synsets containing nouns germane to the topic statement were added as a new section of the topic text. Document text was indexed completely automatically (once the errors were fixed[2]) using the standard SMART indexing routines [1] (i.e., tokenization, stop word removal, and stemming). In general, only the "text" fields of the documents were indexed. For example, only the title, abstract, detailed claims, claims, and design claims sections were indexed for the patent subcollection. The manually assigned keywords included in some of the Ziff documents were not used, nor were the photograph captions of the *San Jose Mercury* collection.

The goal in selecting synsets to be included in a topic statement was to pick synsets that emphasized important concepts of the topic. One aspect of the problem is sense resolution: selecting the synset that contains the correct sense of an ambiguous original topic word. However, since one purpose of the experiments is to investigate how effective lexical relations are in expanding queries assuming good starting concepts, I did not restrict myself to adding only synsets that contain some original topic word. For example, topic 93 asks for information about the support of the NRA and never mentions the word *gun*. Nevertheless, I believed *gun* to be an important concept of the topic and added the synset containing *gun* meaning "a weapon that discharges a missile from a metal tube or barrel" to the topic. (*Rifle*, a word that does appear in the topic statement, is a grandchild of this synset in WordNet, with the intervening synset being {*firearm, piece, small-arm*}). Synset selection was also influenced by the fact that these synsets would be used to expand the query. Early experiments demonstrated that expansion worked poorly when synsets with very many children in the *is-a* hierarchy (e.g. *country*) were used, so those synsets were avoided. Furthermore, when selecting one sense among the different senses in WordNet was difficult, I frequently used the words related to the synsets as a way of making a decision. Figure 2 shows the original text of topic 93 and the synsets that were added to it.

Some topics contained important concepts that had no corresponding synset. Occasionally, the missing synset was a gap in WordNet; for example, *toxic waste*, *genetic engineering*, and *sanctions* meaning economic disciplinary measures are not in version 1.3 of WordNet. More often, the important concept was a proper noun or highly technical term that one wouldn't expect to be in WordNet. *NRA* or *National Rifle Association*, for example, is an important concept for topic 93 but does not occur in WordNet. Nothing was added to the topic texts for concepts that lacked corresponding synsets in these experiments, although making some provision for them would improve retrieval performance.

Once the text of the topics is annotated with synsets, the remainder of the processing is automatic. Selected fields of the topic statements (the title, nationality, narrative, factors, description, and concept fields) are indexed using the standard SMART routines. The terms derived from these sections are "original query terms". The expansion procedure is invoked when the synonym set section is reached. The procedure is controlled by a set of parameters that specifies for each relation type included in WordNet the maximum length of a chain of that type of link that may be followed. A chain begins at each synset listed in the synset section of the topic text and may contain only links of a single type. All synonyms contained within a synset of the chain are added to the query. Collocations such as *change_of_location* in Figure 1 are broken into their component words, stop words such as *of* are removed, and the remaining words are stemmed. The word stems

[1] The actual structure is not quite a hierarchy since a few synsets have more than one parent.

[2] There were seven errors total in files patn_014 and patn_051 that were not on the official list, but caused the files to not conform to the patent collection's readme file. These errors — missing '/TEXT' tags , 'TEXT' tags preceding 'OREF' tags, and the like — were also fixed manually.

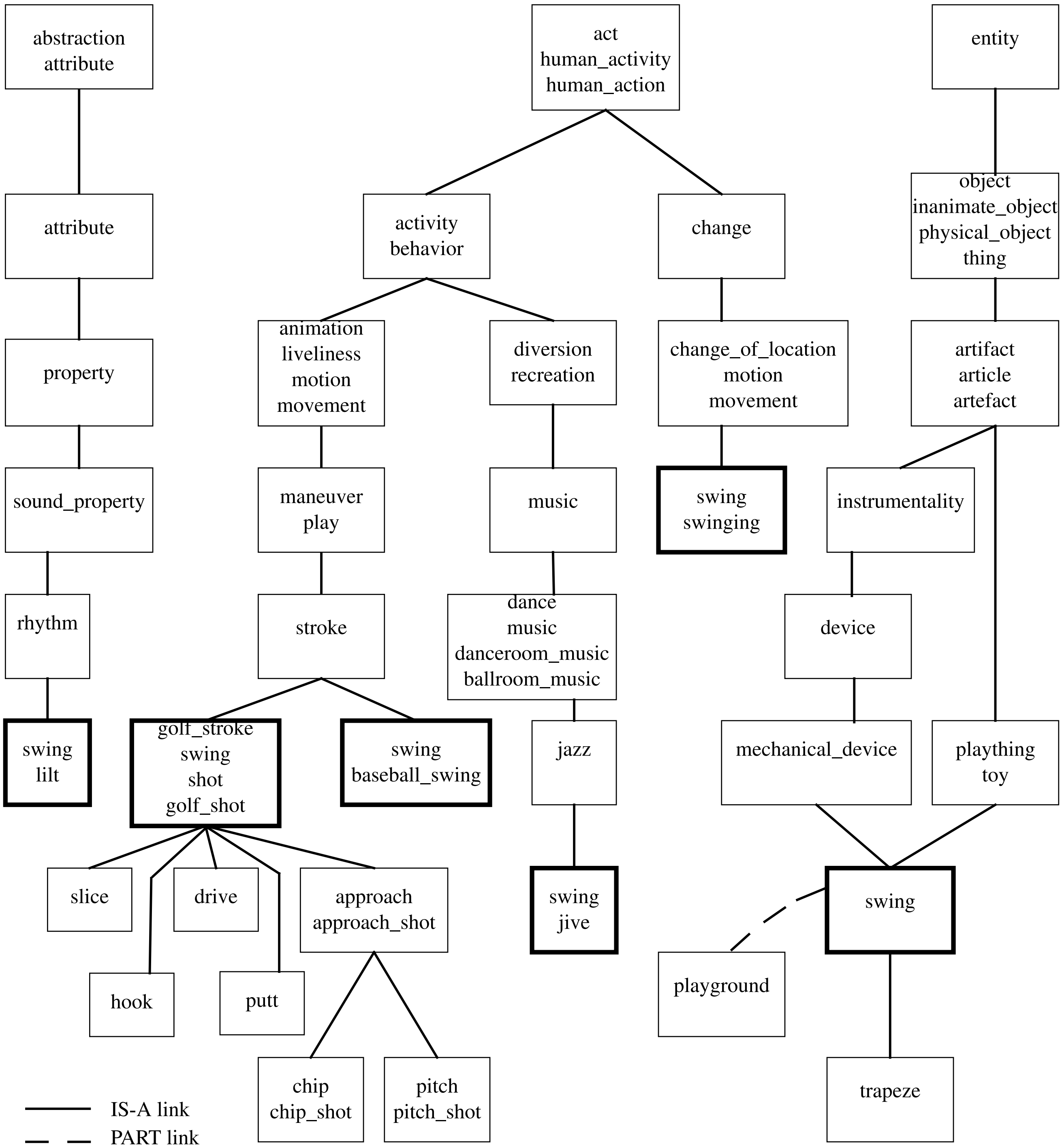

Figure 1: Relations defined for the six senses of the noun *swing* in WordNet.

```
<title> Topic: What Backing Does the National Rifle Association Have?

<desc> Description:
Document must describe or identify supporters of the National Rifle
Association (NRA), or its assets.

<narr> Narrative:
To be relevant, a document must describe or name individuals or organizations who are
members of the NRA, or who contribute money to it.  A document is also relevant
if it quantifies the NRA's financial assets or identifies any other NRA holdings.

<con> Concept(s):
1. National Rifle Association, NRA
2. contributor, member, supporter
3. holdings, assets, finances

<syn>
{funds, finance, monetary_resource, cash_in_hand, pecuniary_resource}
{supporter, protagonist, champion, admirer, booster}
{gun}
```

Figure 2: Topic 093 and the synonym sets selected for it.

plus a tag indicating the lexical relation through which the stems are related to the original synset are then appended to the original query terms.

As an example of the expansion process, consider the synsets for *swing* shown in Figure 1. If the synset added to the topic is the synset containing *golf_stroke*, and any number of hyponym (child) links may be traversed, then the stems of *golf*, *stroke*, *swing*, *shot*, *slice*, *hook*, *drive*, *putt*, *approach*, *chip*, and *pitch* would be added to the query vector. If hyponym chains are limited to length one, then *chip* and *pitch* would not be added. If the synset added to the topic is the one containing swing meaning plaything and any link type may be followed for one link, then the stems of *swing*, *mechanical*, *device*, *plaything*, *toy*, *playground*, and *trapeze* would be added to the query.

Stems added through different lexical relations are kept separate using the extended vector space model introduced by Fox [3]. Each query vector is comprised of subvectors of different concept types (called *ctypes*) where each ctype corresponds to a different lexical relation. A query vector potentially has eleven ctypes: one for original query terms, one for synonyms, and one each for the other relation types contained within the noun portion of WordNet (each half of a symmetric relation has its own ctype). An original query term that is a member of a synset selected for that query appears in both of the respective ctypes. Similarly, a word that is related to a synset through two different relations appears in both ctypes.

The similarity between a document vector D and an extended query vector Q is computed as the weighted sum of the similarities between D and each of the query's subvectors:

$$sim(D,Q) = \sum_{\text{ctype } i} \alpha_i (D \cdot Q_i)$$

where $\cdot$ denotes the inner product of two vectors, Q_i is the ith subvector of Q, and α_i, a real number, reflects the importance of ctype i relative to the other ctypes. Terms in documents vectors are weighted using the *lnc* weights suggested by Buckley et al. [2]; that is, the weight of a term is set to $1.0 + \ln(tf)$ where tf is the number of times the term occurs in the document and is then normalized by the square root of the sum of the squares of the weights in the vector (cosine normalization). Query terms are weighted using *ltN*: the log term frequency factor above is multiplied by the term's inverse document frequency, and the weights in the ctype representing original query terms are normalized by the cosine factor. Weights in additional ctypes are normalized using the length computed for the original terms' ctype. This normalization strategy allows the original query term weights to be unaffected by the expansion process and keeps the weights in each ctype comparable with one another.

3 Experiments

The training data for the routing queries was used both to refine the synsets that were included in the topic text and to select the type of relations used to expand the query vectors. In some cases a synset that appears to be a logical choice for a query is nonetheless detrimental. For example, adding the synset for *death* to topic 59 (weather fatalities) causes the query to retrieve far too many articles reporting on deaths that have no relation to the weather. I produced five different versions of synset-annotated topic texts, although the differences between versions are not very large. The version used in the official routing run added an average of 2.9 synsets to a topic statement, with a minimum of 0 synsets added and a maximum of 6 synsets added.

Of course, the utility of a synset depends in part on how that synset is expanded and the relative weights given to the different link types (the α's in the similarity function above). Table 1 lists the various combinations that were evaluated using the training data. Four different expansion strategies were tried: expansion by synonyms only, expansion by synonyms plus all descendents in the *is-a* hierarchy, expansion by synonyms plus parents and all descendents in the *is-a* hierarchy, and expansion by synonyms plus any synset directly related to the given synset (i.e., a chain of length 1 for all link types). Different α values were also investigated. Assuming original query terms are more important than added terms, the α for the original terms subvector was set to one and the α for other subvectors varied between zero and one as shown in Table 1.

The most effective run was the one that expanded a query synset by any synset directly related to it and had $\alpha = .5$ for all added subvectors. Therefore, this strategy was used to produce the official routing queries from the final version of the annotated text. The scheme added an average of 24.7 words to a query vector (minimum 0, maximum 70), and an average of 20.2 (0, 66) words that are not part of the original text.

The average number of relevant documents retrieved at rank 100 for this run is 40.7 and at rank 1000 is 133.3; the mean "average precision" is .2984. In general, the individual query results are at or slightly above the median, with a few queries significantly below the median. Of more interest to this study is how the expanded queries compare to unexpanded queries. A plot of average recall versus average precision for these two runs is given in Figure 3. As can be seen, the effectiveness of the two query sets is very similar.

Since there was no way to evaluate the relative effectiveness of different expansion schemes for the ad hoc queries, the same same expansion scheme as was used for the official routing run — chains of length one for any relation type and all α's = .5 — was used for the ad hoc run. Furthermore, there could be no refining of which synsets to add, so only one version of synset-annotated text was produced. An average of 2.7 (minimum 0, maximum 6) synonyms was added to an ad hoc topic text. The expansion process added an average of 17.2 (0, 66) terms and 12.8 (0, 55) terms that are not part of the original text.

Siemens actually submitted two ad hoc runs. The first was the expanded queries with α's set to 0, a run that is equivalent to no expansion and is used as a base case. The second Siemens ad hoc run used the .5 α values. A plot of the effectiveness of the two ad hoc runs is given in Figure 4. The differences in effectiveness between unexpanded and expanded queries is even smaller for the ad hoc queries than it is for the routing queries. The average number of relevant documents retrieved at rank 100 is 46.9 for both the unexpanded and expanded queries. The average number of relevant documents retrieved at rank 1000 is 161.4 for the unexpanded queries and 161.3 for the expanded queries. The mean "average precision" is .3408 and .3397 respectively.

A possible explanation for the little difference made by expanding the queries is that the expansion parameters used were too conservative. To test this hypothesis, additional runs were made using the same set of synsets but allowing longer chains of links and/or using greater relative link weights (the α's). Table 2 lists the additional combinations tested using both the ad hoc queries versus the documents on disks one and two, and the routing queries versus the documents on disk 3. As was the case for the routing training runs, the strategy used for the official TREC-2 runs (all links of length one, α's = .5) was the most effective expansion strategy. The more aggressive expansion strategies did make larger differences in retrieval effectiveness compared to the unexpanded queries, but across the set of queries the aggregate difference was negative. Hence it is unlikely that the conservative expansion strategy is the reason for the lack of improvement.

4 Conclusion

The experimental evidence clearly shows this query expansion technique provides little benefit in the TREC environment. The most likely reason for why this should be so is the completeness of the TREC topic descriptions. Query expansion is a recall-enhancing technique and TREC topic descriptions are already large compared to queries found in traditional IR collections. Although most of the expanded queries did have some new terms added to them, the most important terms frequently appeared in both the original term set and the set of expanded terms. This had an effect on the relative weight of those terms in the overall similarity

Expansion by synonyms only

orig terms α	synonyms α
1	.1
1	.3
1	.5
1	.8

Expansion by synonyms plus all descendents

orig terms α	synonyms α	descendents α
1	.1	.1
1	.3	.1
1	.3	.3
1	.5	.1
1	.5	.3
1	.5	.5
1	.8	.1
1	.8	.3
1	.8	.5

Expansion by synonyms plus parents and all descendents

orig terms α	synonyms α	descendents α	parents α
1	.1	.1	.1
1	.3	.1	.1
1	.3	.3	.3
1	.5	.1	.1
1	.5	.3	.1
1	.5	.3	.3
1	.5	.5	.1
1	.5	.5	.3
1	.5	.5	.5
1	.8	.1	.1
1	.8	.3	.1
1	.8	.3	.3
1	.8	.5	.1
1	.8	.5	.3
1	.8	.5	.5

Expansion by synonyms plus any directly related synset

orig terms α	synonyms α	other α
1	.3	.1
1	.3	.3
1	.5	.1
1	.5	.3
1	.5	.5
1	.3	.5

Table 1: Combinations of expansion strategies and relation weights tested.

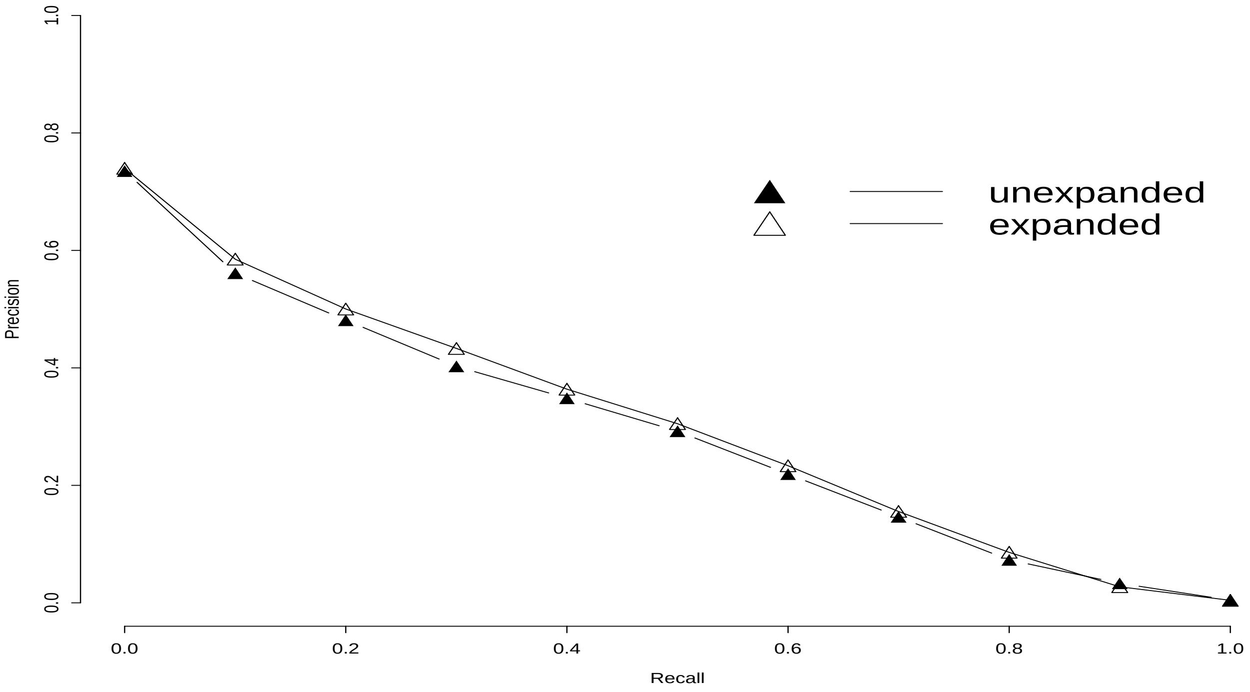

Figure 3: Effectiveness of expanded versus unexpanded routing queries on test documents (disk 3).

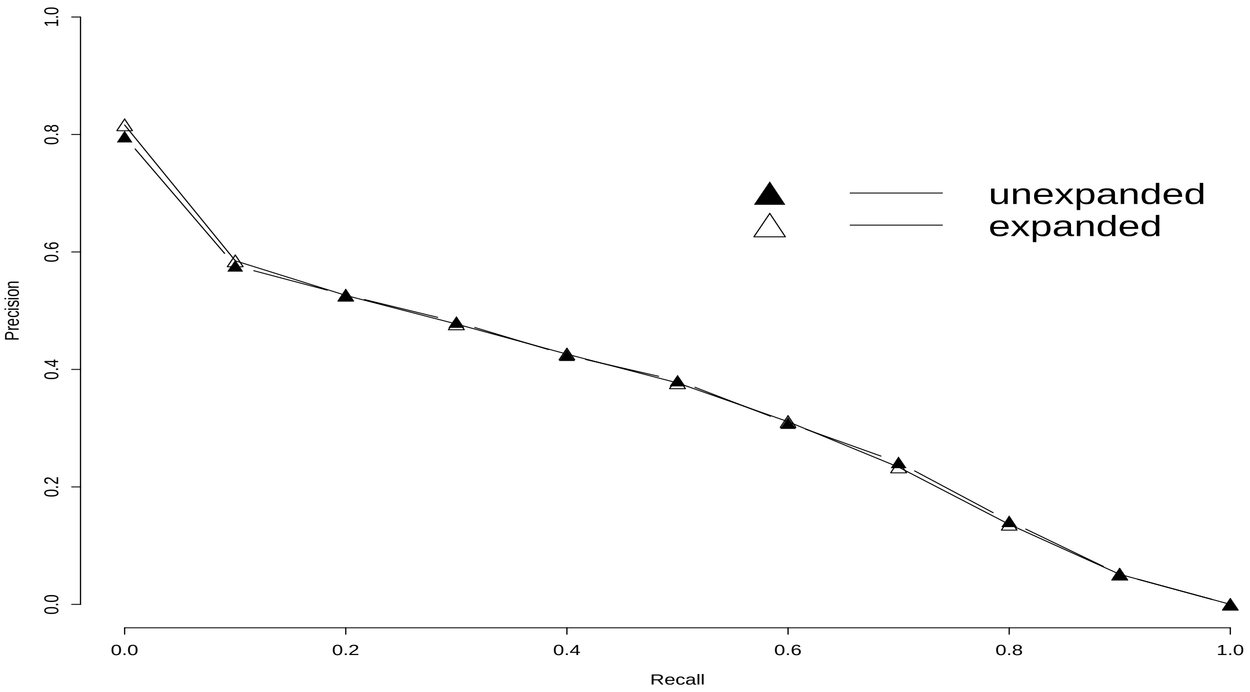

Figure 4: Effectiveness of expanded versus unexpanded ad hoc queries.

Expansion by synonyms plus parents and all descendents			
orig terms α	synonyms α	descendents α	parents α
1	.5	.5	.5
1	1	.5	.5
1	1	1	1
1	2	1	1

Expansion by synonyms plus any directly related synset		
orig terms α	synonyms α	other α
1	0	0
1	.5	.5
1	1	1

Table 2: Additional combinations of expansion strategies and relation weights tested.

computed for a document, especially when some important terms had no corresponding synset. Topic 112 provides an example of this effect. The topic is concerned about the world-wide investment in biotechnology. I added synonym sets for *investment* and *capital* to the topic. WordNet does not contain *biotechnology*, although it does contains *biomedical_science*. Thus, I also added the *biomedical_science* synset and a synset containing *gene*. The resulting expanded query performed significantly worse than the unexpanded version (33 relevant retrieved in the first 100 versus 52 relevant retrieved). The problem is that the expanded query places too much emphasis on money and not enough on biotechnology. Thus these results indicate that simply recognizing which are the important concepts in a query statement is not sufficient to ensure improved retrieval performance. An expansion procedure must also preserve the relative weights of those concepts.

Another possible explanation is that WordNet is not suited for this task — it was not designed to be used in this manner and it may not contain the necessary links. Even if this is true, however, it is unlikely that any other broad-coverage knowledge base would be better suited. The success of relevance feedback and other routing techniques suggests that the most useful relations are specific and idiosyncratic.

A second goal of this work was to characterize the effectiveness of different types of lexical relations when used to expand a query. Assuming the set of words to be expanded is well chosen, any closely related word — regardless of the type of relation — may be a good additional word. Wang et al. reached a similar conclusion [6]. Nevertheless, an added word should be weighted less than the original word that caused it to be included. Runs in which added words were equally or more heavily weighted than original words were consistently less effective than the more conservatively weighted runs. Similarly, runs that added words that were loosely related to original words (i.e., when long paths of links were followed) were consistently less effective than runs that used only near relatives.

Acknowledgements

Geoff Towell carefully read a draft of this paper and suggested changes that improved its presentation and clarity.

References

[1] Chris Buckley. Implementation of the SMART information retrieval system. Technical Report 85-686, Computer Science Department, Cornell University, Ithaca, New York, May 1985.

[2] Chris Buckley, Gerard Salton, and James Allan. Automatic retrieval with locality information using SMART. In D. K. Harman, editor, *Proceedings of the First Text REtrieval Conference (TREC-1)*, pages 59–72. NIST Special Publication 500-207, March 1993.

[3] Edward A. Fox. *Extending the Boolean and Vector Space Models of Information Retrieval with P-norm Queries and Multiple Concept Types.* PhD thesis, Cornell University, 1983. University Microfilms, Ann Arbor, MI.

[4] George Miller. Special Issue, WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 1990.

[5] Ellen M. Voorhees and Yuan-Wang Hou. Vector expansion in a large collection. In D. K. Harman, editor, *Proceedings of the First Text REtrieval Conference (TREC-1)*, pages 343–351. NIST Special Publication 500-207, March 1993.

[6] Yih-Chen Wang, James Vandendorpe, and Martha Evens. Relational thesauri in information retrieval. *Journal of the American Society for Information Science*, 36(1):15–27, January 1985.