The Use of WordNet in Information Retrieval

Rila Mandala, Tokunaga Takenobu, and Tanaka Hozumi
Department of Computer Science
Tokyo Institute of Technology
{rila,take,tanaka}@cs.titech.ac.jp

Abstract

WordNet has been used in information retrieval research by many researchers, but failed to improve the performance of their retrieval system. Thereby in this paper we investigate why the use of WordNet has not been successful. Based on this analysis we propose a method of making WordNet more useful in information retrieval applications. Experiments using several standard information retrieval test collections show that our method results in a significant improvement of information retrieval performance.

1 Introduction

Development of WordNet began in 1985 at Princeton University (Miller, 1990). A team lead by Prof. George Miller aimed to create a source of lexical knowledge whose organization would reflect some of the recent findings of psycholinguistic research into the human lexicon. WordNet has been used in numerous natural language processing, such as part of speech tagging (Segond et al., 97), word sense disambiguation (Resnik, 1995), text categorization (Gomez-Hidalgo and Rodriguez, 1997), information extraction (Chai and Biermann, 1997), and so on with considerable success. However the usefulness of WordNet in information retrieval applications has been debatable.

Information retrieval is concerned with locating documents relevant to a user's information needs from a collection of documents. The user describes his/her information needs with a query which consists of a number of words. The information retrieval system compares the query with documents in the collection and returns the documents that are likely to satisfy the user's information requirements. A fundamental weakness of current information retrieval methods is that the vocabulary that searchers use is often not the same as the one by which the information has been indexed. Query expansion is one method to solve this problem. The query is expanded using terms which have similar meaning or bear some relation to those in the query, increasing the chances of matching words in relevant documents. Expanded terms are generally taken from a thesaurus.

Obviously, given a query, the information retrieval system must present all useful articles to the user. This objective is measured by recall, i.e. the proportion of relevant articles retrieved by the system. Conversely, the information retrieval system must not present any useless article to the user. This criteria is measured by precision, i.e. the proportion of retrieved articles that are relevant.

Voorhees used WordNet as a tool for query expansion (Voorhees, 1994). She conducted experiments using the TREC collection (Voorhees and Harman, 1997) in which all terms in the queries were expanded using a combination of synonyms, hypernyms, and hyponyms. She set the weights of the words contained in the original query to 1, and used a combination of 0.1, 0.3, 0.5, 1, and 2 for the expansion terms. She then used the SMART Information Retrieval System Engine (Salton, 1971) to retrieve the documents. Through this method, Voorhees only succeeded in improving the performance on short queries and a little with no significant improvement for long queries. She further tried to use WordNet as a tool for word sense disambiguation (Voorhees, 1993) and applied it to text retrieval, but the performance of retrieval was degraded.

Stairmand (Stairmand, 1997) used WordNet to compute lexical cohesion according to the method suggested by Morris (Morris and Hirst, 1991), and applied this to information retrieval.

He concluded that his method could not be applied to a fully-functional information retrieval system.

Smeaton (Smeaton and Berrut, 1995) tried to expand the queries of the TREC-4 collection with various strategies of weighting expansion terms, along with manual and automatic word sense disambiguation techniques. Unfortunately all strategies degraded the retrieval performance.

Instead of matching terms in queries and documents, Richardson (Richardson and Smeaton, 1995) used WordNet to compute the semantic distance between concepts or words and then used this term distance to compute the similarity between a query and a document. Although he proposed two methods to compute semantic distances, neither of them increased the retrieval performance.

2 What's wrong with WordNet?

In this section we analyze why WordNet has failed to improve information retrieval performance. We run exact-match retrieval against 9 small standard test collections in order to observe this phenomenon. An information retrieval test collection consists of a collection of documents along with a set of test queries. The set of relevant documents for each test query is also given, so that the performance of the information retrieval system can be measured. We expand queries using a combination of synonyms, hypernyms, and hyponyms in WordNet. The results are shown in Table 1.

In Table 1 we show the name of the test collection (Collection), the total number of documents (#Doc) and queries (#Query), and all relevant documents for all queries (#Rel) in that collection. For each document collection, we indicate the total number of relevant documents retrieved (Rel-ret), the recall ($\frac{\text{Rel-ret}}{\text{\#Rel}}$), the total number of documents retrieved (Ret-docs), and the precision ($\frac{\text{Rel-ret}}{\text{Ret-docs}}$) for each of no expansion (Base), expansion with synonyms (Exp. I), expansion with synonyms and hypernyms (Exp. II), expansion with synonyms and hyponyms (Exp. III), and expansion with synonyms, hypernyms, and hyponyms (Exp. IV).

From the results in Table 1, we can conclude that query expansion can increase recall performance but unfortunately degrades precision performance. We thus turned to investigation of why all the relevant documents could not be retrieved with the query expansion method above. Some of the reasons are stated below :

- Two terms that seem to be interrelated have different parts of speech in WordNet. This is the case between *stochastic* (adjective) and *statistic* (noun). Since words in WordNet are grouped on the basis of part of speech in WordNet, it is not possible to find a relationship between terms with different parts of speech.
- Most of relationships between two terms are not found in WordNet. For example how do we know that Sumitomo Bank is a Japanese company ?
- Some terms are not included in WordNet (proper name, etc).

To overcome all the above problems, we propose a method to enrich WordNet with an automatically constructed thesaurus. The idea underlying this method is that an automatically constructed thesaurus could complement the drawbacks of WordNet. For example, as we stated earlier, proper names and their interrelations among them are not found in WordNet, but if proper names and other terms have some strong relationship, they often cooccur in the document, so that their relationship may be modelled by an automatically constructed thesaurus.

Polysemous words degrade the precision of information retrieval since all senses of the original query term are considered for expansion. To overcome the problem of polysemous words, we apply a restriction in that queries are expanded by adding those terms that are most similar to the entirety of query terms, rather than selecting terms that are similar to a single term in the query.

In the next section we describe the details of our method

3 Method

3.1 Co-occurrence-based Thesaurus

The general idea underlying the use of term co-occurrence data for thesaurus construction is that words that tend to occur together in documents are likely to have similar, or related,

Table 1: Term Expansion Experiment Results using WordNet

Collection	#Doc	#Query	#Rel		Base	Exp. I	Exp. II	Exp. III	Exp. IV
ADI	82	35	170	Rel-ret	157	159	166	169	169
				Recall	0.9235	0.9353	0.9765	0.9941	0.9941
				Ret-docs	2,063	2,295	2,542	2,737	2,782
				Precision	0.0761	0.0693	0.0653	0.0617	0.0607
CACM	3204	64	796	Rel-ret	738	756	766	773	773
				Recall	0.9271	0.9497	0.9623	0.9711	0.9711
				Ret-docs	67,950	86,552	101,154	109,391	116,001
				Precision	0.0109	0.0087	0.0076	0.0070	0.0067
CISI	1460	112	3114	Rel-ret	2,952	3015	3,076	3,104	3,106
				Recall	0.9479	0.9682	0.9878	0.9968	0.9974
				Ret-docs	87,895	98,844	106,275	108,970	109,674
				Precision	0.0336	0.0305	0.0289	0.0284	0.0283
CRAN	1398	225	1838	Rel-ret	1,769	1,801	1,823	1,815	1,827
				Recall	0.9625	0.9799	0.9918	0.9875	0.9940
				Ret-docs	199,469	247,212	284,026	287,028	301,314
				Precision	0.0089	0.0073	0.0064	0.0063	0.0060
INSPEC	12684	84	2543	Rel-ret	2,508	2,531	2,538	2,536	2,542
				Recall	0.9862	0.9953	0.9980	0.9972	0.9996
				Ret-docs	564,809	735,931	852,056	869,364	912,810
				Precision	0.0044	0.0034	0.0030	0.0029	0.0028
LISA	6004	35	339	Rel-ret	339	339	339	339	339
				Recall	1.0000	1.0000	1.0000	1.0000	1.0000
				Ret-docs	148,547	171,808	184,101	188,289	189,784
				Precision	0.0023	0.0020	0.0018	0.0018	0.0018
MED	1033	30	696	Rel-ret	639	662	670	671	673
				Recall	0.9181	0.9511	0.9626	0.9640	0.9670
				Ret-docs	12,021	16,758	22,316	22,866	25,250
				Precision	0.0532	0.0395	0.0300	0.0293	0.0267
NPL	11429	100	2083	Rel-ret	2,061	2,071	2,073	2,072	2,074
				Recall	0.9894	0.9942	0.9952	0.9942	0.9957
				Ret-docs	267,158	395,280	539,048	577,033	678,828
				Precision	0.0077	0.0052	0.0038	0.0036	0.0031
TIME	423	24	324	Rel-ret	324	324	324	324	324
				Recall	1.000	1.000	1.000	1.000	1.000
				Ret-docs	23,014	29,912	33.650	32,696	34,443
				Precision	0.0141	0.0108	0.0096	0.0095	0.0094

meanings. Co-occurrence data thus provides a statistical method for automatically identifying semantic relationships that are normally contained in a hand-made thesaurus. Suppose two words (A and B) occur f_a and f_b times, respectively, and cooccur f_c times, then the similarity between A and B can be calculated using a similarity coefficient such as the Dice Coefficient :

$$\frac{2 \times f_c}{f_a + f_b}$$

3.2 Predicate-Argument-based Thesaurus

In contrast with the previous section, this method attempts to construct a thesaurus according to predicate-argument structures. The use of this method for thesaurus construction is based on the idea that there are restrictions on what words can appear in certain environments, and in particular, what words can be arguments of a certain predicate. For example, a *cat* may *walk*, *bite*, but can not *fly*. Each noun may therefore be characterized according to the verbs or adjectives that it occurs with. Nouns may then be grouped according to the extent to which they appear in similar constructions.

First, all the documents are parsed using the Apple Pie Parser, which is a probabilistic chart parser developed by Satoshi Sekine (Sekine and Grishman, 1995). Then the following syntactic structures are extracted :

- Subject-Verb
- Verb-Object
- Adjective-Noun

Each noun has a set of verbs and adjective that it occurs with, and for each such relationship, a dice coefficient value is calculated.

- $C_{sub}(v_i, n_j) = \frac{2 \times f_{sub}(v_i, n_j)}{f(v_i) + f_{sub}(n_j)}$, where $f_{sub}(v_i, n_j)$ is the frequency of noun n_j occurring as the subject of verb v_i, $f_{sub}(n_j)$ is the frequency of the noun n_j occurring as subject of any verb, and $f(v_i)$ is the frequency of the verb v_i
- $C_{obj}(v_i, n_j) = \frac{2 \times f_{obj}(v_i, n_j)}{f(v_i) + f_{obj}(n_j)}$, where $f_{obj}(v_i, n_j)$ is the frequency of noun n_j occurring as the object of verb v_i, $f_{obj}(n_j)$ is the frequency of the noun n_j occurring as object of any verb, and $f(v_i)$ is the frequency of the verb v_i
- $C_{adj}(a_i, n_j) = \frac{2 \times f_{adj}(a_i, n_j)}{f(a_i) + f_{adj}(n_j)}$, where $f(a_i, n_j)$ is the frequency of noun n_j occurring as argument of adjective a_i, $f_{adj}(n_j)$ is the frequency of the noun n_j occurring as argument of any adjective, and $f(a_i)$ is the frequency of the adjective a_i

We define the object similarity of two nouns with respect to one predicate, as the minimum of each dice coefficient with respect to that predicate, i.e.
$SIM_{sub}(v_i, n_j, n_k) = min\{C_{sub}(v_i, n_j), C_{sub}(v_i, n_k)\}$
$SIM_{obj}(v_i, n_j, n_k) = min\{C_{obj}(v_i, n_j), C_{obj}(v_i, n_k)\}$
$SIM_{adj}(a_i, n_j, n_k) = min\{C_{adj}(a_i, n_j), C_{adj}(a_i, n_k)\}$

Finally the overall similarity between two nouns is defined as the average of all the similarities between those two nouns for all predicate-argument structures.

3.3 Expansion Term Weighting Method

A query q is represented by a vector $\overrightarrow{q} = (q_1, q_2, ..., q_n)$, where the q_i's are the weights of the search terms t_i contained in query q.

The similarity between a query q and a term t_j can be defined as belows :

$$simqt(q, t_j) = \sum_{t_i \in q} q_i * sim(t_i, t_j)$$

Where the value of $sim(t_i, t_j)$ can be defined as the average of the similarity values in the three types of thesaurus. Since in WordNet there are no similarity weights, when there is a relation between two terms in WordNet, their similarity is taken from the average of the similarity between those two terms in the co-occurrence-based and in predicate-argument-based thesauri.

With respect to the query q, all the terms in the collection can now be ranked according to their *simqt*. Expansion terms are terms t_j with high $simqt(q, t_j)$.

The $weight(q, t_j)$ of an expansion term t_j is defined as a function of $simqt(q, t_j)$:

$$weight(q, t_j) = \frac{simqt(q, t_j)}{\sum_{t_i \in q} q_i}$$

where $0 \leq weight(q, t_j) \leq 1$.

An expansion term gets a weight of 1 if its similarity to all the terms in the query is 1. Expansion terms with similarity 0 to all the terms in the query get a weight of 0. The weight of an expansion term depends both on the entire retrieval query and on the similarity between the terms. The weight of an expansion term can be interpreted mathematically as the weighted mean of the similarities between the term t_j and all the query terms. The weight of the original query terms are the weighting factors of those similarities.

Therefore the query q is expanded by adding the following query :

$$\overrightarrow{q_e} = (a_1, a_2, ..., a_r)$$

where a_j is equal to $weight(q, t_j)$ if t_j belongs to the top r ranked terms. Otherwise a_j is equal to 0.

The resulting expanded query is :

$$\overrightarrow{q}_{expanded} = \overrightarrow{q} \circ \overrightarrow{q_e}$$

where the $\circ$ is defined as the concatenation operator.

The method above can accommodate the polysemous word problem, because an expansion term which is taken from a different sense to the original query term is given very low weight.

4 Experimental Results

In order to evaluate the effectiveness of the proposed method in the previous section we conducted experiments using the WSJ, CACM, INSPEC, CISI, Cranfield, NPL, and LISA test collections. The WSJ collection comprises part of the TREC collection (Voorhees and Harman, 1997). As a baseline we used SMART (Salton, 1971) without expansion. SMART is an information retrieval engine based on the vector space model in which term weights are calculated based on term frequency, inverse document frequency and document length normalization. The results are shown in Table 2. This table shows the average of 11 point uninterpolated recall-precision for each of baseline, expansion using only WordNet, expansion using only predicate-argument-based thesaurus, expansion using only cooccurrence-based thesaurus, and expansion using all of them. For each method we give the percentage of improvement over the baseline. It is shown that the performance using the combined thesauri for query expansion is better than both SMART and using just one type of thesaurus.

Table 2: Experiment Result using Combined Thesauri

Coll	Base	Expanded with			
		WordNet only	Pred-arg only	Cooccur only	Combined
WSJ	0.245	0.251 (+2.6%)	0.258 (+5.2%)	0.294 (+19.8%)	0.384 (+56.7%)
CACM	0.269	0.281 (+4.5%)	0.291 (+8.3%)	0.297 (+10.6%)	0.533 (+98.2%)
INSPEC	0.273	0.283 (+3.7%)	0.284 (+4.3%)	0.328 (+20.4%)	0.472 (+73.1%)
CISI	0.216	0.231 (+7.2%)	0.236 (+9.4%)	0.262 (+21.6%)	0.391 (+81.3%)
Cran	0.412	0.421 (+2.3%)	0.441 (+7.0%)	0.487 (+18.3%)	0.667 (+62.1%)
NPL	0.201	0.210 (+4.2%)	0.217 (+8.1%)	0.238 (+17.5%)	0.333 (+65.5%)
LISA	0.304	0.313 (+3.1%)	0.327 (+7.6%)	0.369 (+21.4%)	0.485 (+59.7%)

5 Discussions

In this section we discuss why our method of using WordNet is able to improve the performance of information retrieval. The important points of our method are :

- the coverage of WordNet is broadened
- weighting method

The three types of thesaurus we used have different characteristics. Automatically constructed thesauri add not only new terms but also new relationships not found in WordNet. If two terms often cooccur together in a document then those two terms are likely bear some relationship. Why not only use the automatically constructed thesauri ? The answer to this is that some relationships may be missing in the automatically constructed thesauri. For example, consider the words *tumor* and *tumour*. These words certainly share the same context, but would never appear in the same document, at least not with a frequency recognized by a cooccurrence-based method. In general, different words used to describe similar concepts may never be used in the same document, and are thus missed by the cooccurrence methods. However their relationship may be found in the WordNet thesaurus.

The second point is our weighting method. As already mentioned before, most attempts at automatically expanding queries by means of WordNet have failed to improve retrieval effectiveness. The opposite has often been true: expanded queries were less effective than the original queries. Beside the "incomplete" nature of WordNet, we believe that a further problem, the weighting of expansion terms, has not been solved. All weighting methods described in the past researches of query expansion using WordNet have been based on "trial and error" or adhoc methods. That is, they have no underlying justification.

The advantages of our weighting method are:

- the weight of each expansion term considers the similarity of that term with all terms in the original query, rather than to just one or some query terms.
- the weight of the expansion term accommodates the polysemous word problem.

This method can accommodate the polysemous word problem, because an expansion term taken from a different sense to the original query term sense is given very low weight. The reason for this is that, the weighting method depends on all query terms and all of the thesauri. For example, the word *bank* has many senses in WordNet. Two such senses are the financial institution and the river edge senses. In a document collection relating to financial banks, the river sense of *bank* will generally not be found in the cooccurrence-based thesaurus because of a lack of articles talking about rivers. Even though (with small possibility) there may be some documents in the collection talking about rivers, if the query contained the finance sense of *bank* then the other terms in the query would also concerned with finance and not rivers. Thus rivers would only have a relationship with the *bank* term and there would be no relationships with other terms in the original query, resulting in a low weight. Since our weighting method depends on both query in its entirety and similarity in the three thesauri, the wrong sense expansion terms are given very low weight.

6 Related Research

Smeaton (Smeaton and Berrut, 1995) and Voorhees (Voorhees, 1994) have proposed an expansion method using WordNet. Our method differs from theirs in that we enrich the coverage of WordNet using two methods of automatic thesaurus construction, and we weight the expansion term appropriately so that it can accommodate the polysemous word problem.

Although Stairmand (Stairmand, 1997) and Richardson (Richardson and Smeaton, 1995) have proposed the use of WordNet in information retrieval, they did not used WordNet in the query expansion framework.

Our predicate-argument structure-based thesaurus is based on the method proposed by Hindle (Hindle, 1990), although Hindle did not apply it to information retrieval. Instead, he used mutual information statistics as a similarity coefficient, wheras we used the Dice coefficient for normalization purposes. Hindle only extracted the subject-verb and the object-verb predicate-arguments, while we also extract adjective-noun predicate-arguments.

Our weighting method follows the Qiu method (Qiu and Frei, 1993), except that Qiu used it to expand terms only from a single automatically constructed thesarus and did not consider the use of more than one thesaurus.

7 Conclusions

This paper analyzed why the use of WordNet has failed to improve the retrieval effectiveness in information retrieval applications. We found that the main reason is that most relationships between terms are not found in WordNet, and some terms, such as proper names, are not included in WordNet. To overcome this problem we proposed a method to enrich the WordNet with automatically constructed thesauri.

Another problem in query expansion is that of polysemous words. Instead of using a word sense disambiguation method to select the apropriate sense of each word, we overcame this problem with a weighting method. Experiments proved that our method of using WordNet in query expansion could improve information retrieval effectiveness.

Future work will include experiments on larger test collections, and the use of WordNet in methods other than query expansion in information retrieval.

8 Acknowledgements

The authors would like to thank Mr. Timothy Baldwin (TIT, Japan) for his comments on the earlier version of this paper, Dr. Chris Buckley (Cornell Univesity) for the SMART support, and Mr. Satoshi Sekine (New York University) for the Apple Pie Parser support.

References

J.Y. Chai and A. Biermann. 1997. The use of lexical semantics in information extraction. In *Proceedings of the Workshop in Automatic Information Extraction and Building of Lexical Semantic Resources*, pages 61–70.

J.M. Gomez-Hidalgo and M.B. Rodriguez. 1997. Integrating a lexical database and a training collection for text categorization. In *Proceedings of the Workshop in Automatic Information Extraction and Building of Lexical Semantic Resources*, pages 39–44.

D. Hindle. 1990. Noun classification from predicate-argument structures. In *Proceed-*

ings of 28th Annual Meeting of the ACL, pages 268–275.

G.A Miller. 1990. Special issue, wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4).

J. Morris and G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. In *Proceedings of ACL Conference*, pages 21–45.

Qiu and H.P. Frei. 1993. Concept based query expansion. In *Proceedings of the 16th ACM SIGIR Conference*, pages 160–169.

P Resnik. 1995. Disambiguating noun grouping with respect to wordnet senses. In *Proceedings of 3rd Workshop on Very Large Corpora.*

R. Richardson and A.F. Smeaton. 1995. Using wordnet in a knowledge-based approach to information retrieval. Technical Report CA-0395, School of Computer Applications, Dublin City University.

G. Salton. 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Prentice-Hall.

F. Segond, A. Schiller, G. Grefenstette, and J. Chanod. 97. An experiment in semantic tagging using hidden markov model tagging. In *Proceedings of the Workshop in Automatic Information Extraction and Building of Lexical Semantic Resources*, pages 78–81.

S. Sekine and R. Grishman. 1995. A corpus-based probabilistic grammar with only two non-terminals. In *Proceedings of the International Workshop on Parsing Technologies.*

A.F. Smeaton and C. Berrut. 1995. Running trec-4 experiments: A chronological report of query expansion experiments carried out as part of trec-4. Technical Report CA-2095, School of Comp. Science, Dublin City University.

M.A. Stairmand. 1997. Textual context analysis for information retrieval. In *Proceedings of the 20th ACM-SIGIR Conference*, pages 140–147.

E.M. Voorhees and D. Harman. 1997. Overview of the fifth text retrieval conference (trec-5). In *Proceedings of the Fifth Text REtrieval Conference*, pages 1–28. NIST Special Publication 500-238.

E.M. Voorhees. 1993. Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of the 16th ACM-SIGIR Conference*, pages 171–180.

E.M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th ACM-SIGIR Conference*, pages 61–69.