Ad Hoc Retrieval Experiments Using WordNet and Automatically Constructed Thesauri

Rila Mandala, Takenobu Tokunaga, Hozumi Tanaka,
Akitoshi Okumura and Kenji Satoh

NEC Corp. and Tokyo Institute of Technology
Japan

Abstract

This paper describe our method in automatic-adhoc task of TREC-7. We propose a method to improve the performance of information retrieval system by expanded the query using 3 diffferent types of thesaurus. The expansion terms are taken from hand-crafted thesaurus (WordNet), co-occurrence-based automatically constructed thesaurus, and syntactically predicate-argument based automatically constructed thesaurus.

1 Introduction

A critical problem in information retrieval is that the vocabulary that the searchers use is not the same as the one by which the documents have been indexed. The word synonymy is one example of this problem. If a user use a synonym of a word which document has been indexed in his/her query, then that documents could not be retrieved.

Query expansion [3] is one method in information retrieval to avoid this problem. The expansion terms can be taken from thesaurus [4, 9, 11]. There are many research of query expansion using thesaurus in the literature. Briefly there are two types of thesaurus, i.e. hand-crafted thesaurus and automatically constructed thesaurus. WordNet [10] is one example of hand-crafted thesaurus which is publically available in machine readable form.

Corpus-based thesaurus is a thesaurus which is constructed automatically from the corpus without intervention of human. There are two different method to extract thesaural relationships from corpora predicate-argument (also called head-modifier) method [6, 5, 8, 7, 13] and co-occurrence statistical method [1, 2, 12, 15]

We propose the use of WordNet, co-occurrence-based and predicate-argument-based automatically constructed thesauri for query expansion in automatic-adhoc task of TREC-7.

2 Method

2.1 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, 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

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

2.2 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 [7]. 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 bottom-up probabilistic chart parser developed by Satoshi Sekine [16]. Its grammar is a semi-context sensitive grammar and it was automatically extracted from Penn Tree Bank syntactically tagged corpus made at the University of Pennsylvania. Its performance is 0.71 of precision, 0.70 of recall and 3.03 of average crossing.

Using this parser, 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 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.

2.3 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 [12] :

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

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 in the thesauri.

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 z 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.

3 Experiments

As a retrieval engine we used SMART [14] version 11.0. SMART is an information retrieval system based on the vector space model in which term weights are calculated based on term frequency, inverse document frequency, and document length normalization. We used *lnc* for document's term weighting and *ltc* for query's term weighting.

We ran experiments in the automatic-adhoc task framework using only title, only description, and all terms of the topics. The results are shown belows

	Title	Description	All
Total number of documents over all queries			
Retrieved	50000	50000	50000
Relevant	4674	4674	4674
Rel_ret	2435	3149	3106
Interpolated Recall - Precision Averages			
at 0.00	0.6957	0.7782	0.8161
at 0.10	0.4528	0.5643	0.5783
at 0.20	0.3622	0.4377	0.4511
at 0.30	0.2864	0.3519	0.3575
at 0.40	0.2148	0.2981	0.2899
at 0.50	0.1438	0.2300	0.2177
at 0.60	0.1017	0.1786	0.1618
at 0.70	0.0530	0.1212	0.1121
at 0.80	0.0321	0.0749	0.0636
at 0.90	0.0049	0.0159	0.0306
at 1.00	0.0005	0.0067	0.0054
Average precision (non-interpolated) over all rel docs	0.1898	0.2584	0.2565
Precision			
At 5 docs	0.4720	0.5840	0.5800
At 10 docs	0.4260	0.5460	0.5480
At 15 docs	0.4080	0.5013	0.4973
At 20 docs	0.3700	0.4640	0.4700
At 30 docs	0.3320	0.4113	0.4147
At 100 docs	0.2012	0.2406	0.2484
At 200 docs	0.1395	0.1771	0.1771
At 500 docs	0.0791	0.1038	0.1023
At 1000 docs	0.0487	0.0630	0.0621
R-Precision (precision after R (= num_rel for a query) docs retrieved)			
Exact	0.2403	0.2993	0.2989

Figure 1 shows the 11 point interpolated precision graph for the retrieval result using title only, description only, and title+description+narrative.

4 Discussion of Result

As expected, the performance of retrieval using only title of topic yields a worst performance. The use of only description of topic has a higher retrieval performance than the use of all sections of topic. This can be explained that the narrative section of topic has some negation statements

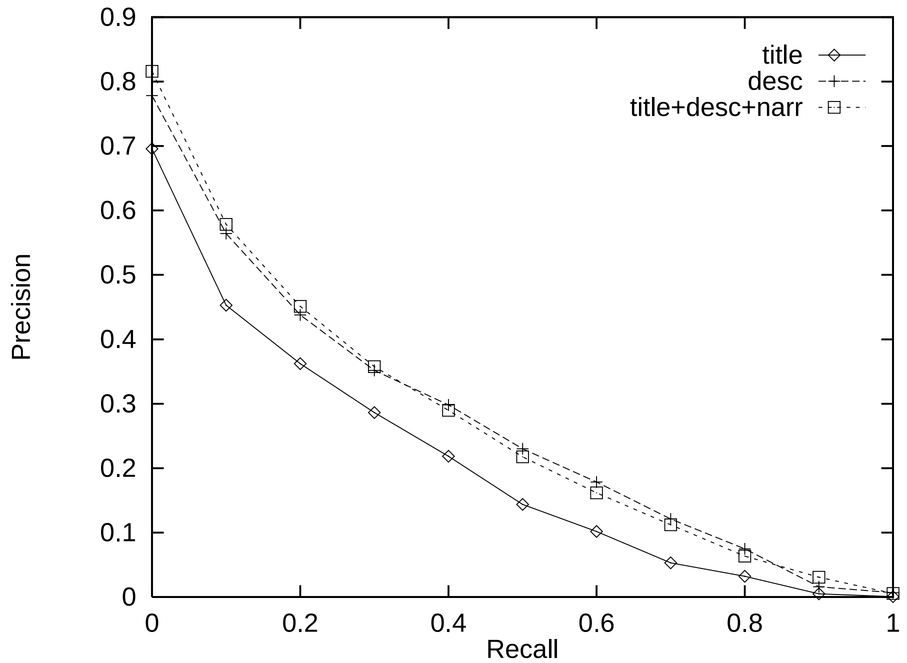

Figure 1: 11-point interpolated precision using title, description, and title+description+narrative

which could not be handled properly by our system yet. Expanding terms occur in the negation statement degraded the performance very much.

5 Conclusion

We have implemented and experimented a method for query expansion using WordNet and corpus-based thesauri. To avoid the wrong expansion terms, a weighting method is utilized whereby the weight of expansion terms depends on the similarity value of those terms in the various thesauri and on the weight of all terms in original query.

In the future, we will investigate the proper method to handle term expansion in the negation statement.

6 Acknowledgements

The authors would like to thank Mr. Timothy Baldwin (TIT, Japan) for his comments on the earlier version of this paper. We also grateful to Dr. Chris Buckley (SabIR Research) and Dr. Satoshi Sekine (New York University) for providing the SMART information retrieval engine and Apple Pie Parser respectively.

References

[1] Carolyn J. Crouch. An aproach to the automatic construction of global thesauri. *Information Processing and Management*, Vol. 26, No. 5, pp. 629–640, 1990.

[2] Yang B. Crouch, C.J. Experiments in automatic statistical thesaurus construction. In *Proceeding of the Fifteenth ACM SIGIR Conference*, pp. 77–82, 1992.

[3] Ekmekcioglu F.C. Effectiveness of query expansion in ranked-output document retrieval systems. *Journal of Information Science*, Vol. 18, pp. 139–147, 1992.

[4] Edward A. Fox. Lexical relations enhancing effectiveness of information retrieval systems. *SIGIR Forum*, Vol. 15, No. 3, pp. 6–36, 1980.

[5] Gregory. Grafenstette. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publ., 1994.

[6] G. Grefenstette. Use of syntactic context to produce term association lists for text retrieval. In *Proceedings of the 15th ACM SIGIR Conference*, pp. 89–97, 1992.

[7] D. Hindle. Noun classification from predicate-argument structures. In *Proceedings of 28th Annual Meeting of the ACL*, pp. 268–275, 1990.

[8] Y. Jing and B. Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO*, pp. 146–160, 1994.

[9] Joana Kristensen. Expanding end-users query statements for free text searching with a search-aid thesaurus. *Information Processing and Management*, Vol. 29, No. 6, pp. 733–744, 1993.

[10] G.A Miller. Special issue, wordnet an on-line lexical database. *International Journal of Lexicography*, Vol. 3, No. 4, 1990.

[11] Chris D. Paice. A thesaural model of information retrieval. *Information Processing and Management*, Vol. 27, No. 5, pp. 433–447, 1991.

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

[13] G Ruge. Experiments on linguistically-based term associations. *Information Processing and Management*, Vol. 28, No. 3, pp. 317–332, 1992.

[14] G. Salton. *The SMART Retrieval System Experiments in Automatic Document Processing.* Prentice-Hall, 1971.

[15] H. Schutze and J.O. Pederson. A cooccurrence-based thesaurus and two applications to information retrieval. In *Proceedings of RIAO Conference*, pp. 266–274, 1994.

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