Expansion-Based Technologies in Finding Relevant and New Information:

THU TREC2002 Novelty Track Experiments*

Min Zhang, Ruihua Song, Chuan Lin, Shaoping Ma, Zhe Jiang, Yijiang Jin, Yiqun Liu, Le Zhao
State Key Lab of Intelligent Tech. & Sys., CST Dept, Tsinghua University, Beijing 100084, China
zhangmin99@mails.tsinghua.edu.cn

1 Introduction

This is the first time that Tsinghua University took part in TREC. In this year's novelty track, our basic idea is to find the key factor that help people find relevant and new information on a set of documents with noise. We paid attention to three points: 1. how to get full information from a short sentence; 2. how to complement hidden well-known knowledge to the sentences; 3. how to make the determination of duplication.

Accordingly, expansion-based technologies are the key points. Studies of expansion technologies have been performed on three levels: efficient query expansion based on thesaurus and statistics, replacement-based document expansion, and term-expansion-related duplication elimination strategy based on overlapping measurement.

Besides, two issues have been studied: finding key information in topics, and dynamic result selection. A new IR system has been developed for the task. In the system, four weighting strategies have been implemented: ltn.lnu[1], BM2500[2], FUB1[3], FUB2[3]. It provides both similarity and overlapping measurements, based on term expansion. Comparisons can be made on sentence-to-sentence or sentence-to-pool level.

2 Query Expansion

In the task, it is most possible that a relevant sentence is mismatched to the query if we only use the original topic words. Therefore proper query expansion (*QE*) technology is necessary and helpful. Besides thesaurus based *QE* described in section 1 and 2, we proposed a new statistical expansion approach called *local co-occurrence based query expansion*, shown in section 3.

2.1 Using WordNet

Firstly Wordnet[4] is used as the thesaurus to expand query words. Totally three kinds of information were observed in our experiments: hyponyms (descendants), synonyms and coordinated words.

Figure 2.1 shows the effects of *QE* using WordNet hyponyms. Effects of using WordNet synonyms and coordinated words are shown in Table 2.1. In the figure, *hpyo* means to expand all hyponyms and sub-hyponyms of each topic word. And *hypo_1*, *hypo_2* and *hypo_3* refer to expanding words in the direct one or two or three levels of hyponyms respectively. *Hypo_leaf* is to expand hyponyms in leaf nodes of WordNet. Baseline result used long query.

Results show that the more words expanded, the worse the retrieval performance is. All kinds of hyponyms expansion did not help retrieval. Expanding first level hyponyms (average P*R=0.066) makes trivial improvement to the baseline (average P*R = 0.064). Shown in the table, expansion based on synonyms achieves a little improvement in terms of average P*R while it does not help in terms of F-measure.

* Supported by the Chinese National Key Foundation Research & Development Plan (Grant G1998030509), Natural Science Foundation No.60223004, and National 863 High Technology Project No. 2001AA114082.

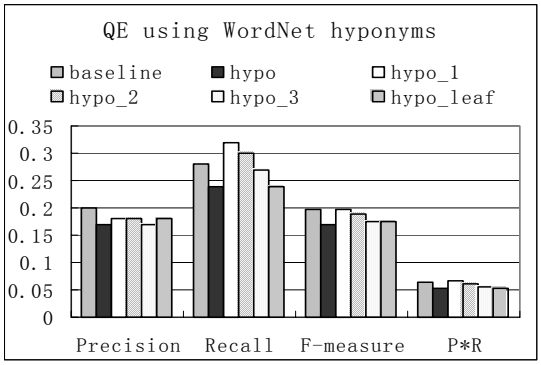


Figure 2.1 Effects of QE with WordNet hyponyms.

Table 2.1 Effects of QE using WordNet synonyms and coordinate words

	P	R	F	P*R
Baseline	0.2	0.28	0.197	0.064
Hypo_1	0.18	0.32	0.197	0.066
Synset	0.17	0.32	0.195	0.068
Coordinate	0.18	0.29	0.189	0.061

P: Average precision R: Average Recall

F: F-measure

P*R: Average Precision*Recall

2.2 Using Dr. Lin Dekang's synonyms dictionary

We also observed the performance by Dr. Lin Dekang's synonyms dictionary[5]. It provides two kinds of synonym dictionaries, based on dependency and mutual information respectively. This *QE* approach works better than the baseline in training set, while makes trivial improvement in test data (see Table 2.2).

Table 2.2 Effects of QE by Dr. Lin Dekang's synonyms dictionary

Ave. Precision	Ave. Recall	F-measure	Ave. P*R
0.18	0.31	0.196	0.067

2.3 QE based on local co-occurrence

We proposed a new statistical expansion approach, which expands terms highly co-occurred in a fixed window size with any of headwords in the relevant document set, called *local co-occurrence expansion (LCE)*. The results are extremely good. Other than most expansion techniques, *LCE* made consistent great progress in terms of both recall and precision. Experimental results are shown in Table 2.3. By using *LCE*, we got 15% and 28% improvement in terms of F-measure and average P*R respectively.

Figure 2.2 gives the overview of query expansion technologies used in our novelty experiments.

Table 2.3 Effects of *QE* by local co-occurrence expansion

	Baseline	LCE
Ave. Precision	0.20	0.21
Ave. Recall	0.28	0.34
F-measure	0.197	0.227
Ave. P*R	0.064	0.081

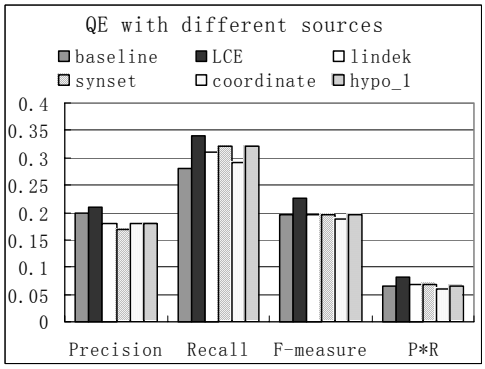


Figure 2.2 Overview of QE experiments

3 Document Expansion

Sometimes, the query mentions a general topic while some relevant documents describe detailed information. For example, the concept of "vehicle" in query is expressed by specific words such as "car", "truck" and "aircraft" in documents. In this case, (1) *QE* may take too many useless words because of aimless of expansion; (2) Setting weights for original and expanded terms is one of the main difficulties in *QE*. Therefore we proposed term expansion in documents (referred as *DE*) to solve the problem.

Other than *QE*, the concept network in WordNet is definitely helpful. We used three levels of hypernyms

(ancestor) and their synonyms, referred as *hype_3* in our experiments. The algorithm of document expansion (*DE*) is as following. For each noun in a relevant document, if its 3-level hypernyms include any keyword in query, then replace the noun with the keyword. By doing this, the documents evolve into expanded documents while the query takes no change. Experimental results in Table 3.1 show that *DE* got higher performance than *Q*E under the same circumstances. The key point of *DE* is replacement. The keyword and its hyponyms were represented by an identical word, while the keyword and its hyponym were treated as different words in *QE*. Essentially *DE* used the concept space instead of the term space.

Table 3.1 Comparisons between *QE* and *DE*

Method	Ave. Precision	Ave. Recall	F-measure	Ave. P*R
QE (*hypo_3*)	0.14	0.25	0.179	0.057
DE (*hype_3*)	0.18	0.40	0.248	0.079

4 Combination of QE and DE

4.1 Topic Classification by *QE* and *DE*

QE and *DE* are oriented from two aspects of retrieval problem and may work well for different topics. Therefore we classified the topics into two classes according to topic or document characteristics to perform QE or DE respectively, which lead to better performance than either approach.

One intuitive method of classification is topic-oriented. Define fields' similarities in topic: FS_{td} (<title> and <desc>), FS_{tn} (<title> and <narr>) and FS_{dn}(<desc> and <narr>). In our experiments we use the following rules: if $FS_{dn} < \theta_1$ and $(FS_{td}+FS_{dn}-2FS_{tn}) < \theta_2$, then the topic should use *DE* on the topic, otherwise *QE* is performed. The thresholds θ_1 and θ_2 are set according to 0.07 and 0.035.

The other one is document-oriented. Compute the value of: *(# words expanded)/(# words in docs)* for each topic. Only when the value is greater than θ, use *DE*. In our experiments, $\theta = 0.058$.

All the parameters were set according to TREC2002 training examples. It got better performance although the thresholds are not fit for testing data completely. The effects of two approaches are shown in Table 4.1, where *TOTC* and *DOTC* means topic similarity and *DE* oriented topic classification, respectively.

Table 4.1 Effects of topic classification

Method	Ave. Precision	Ave. Recall	Ave. P*R
QE (*LCE*)	0.21	0.34	0.081
DE	0.22	0.28	0.066
TOTC	0.23	0.34	0.087
DOTC	0.23	0.374	0.086

4.2 Result Combination

We've tried several different combination strategies. Here are two that work pretty well. One is called re-ranking (Eq4.1), and another one is called combining inversed rank (Eq4.2). We used Eq2.7 in the experiments. The combined approaches are *QE*(*LCE*) and *DE*. $\lambda \leq 0.3$.

If $Doc_i \in$ result list1 & $Doc_i \in$ list2, then $Sim_i'=\lambda S_{1i}$, ($\lambda>1$) else $S'=S_{1i}$	4.1
if $Doc_i \in$ result list1 or $Doc_i \in$ list2, $Sim_i'=\lambda*1/Rank_{1i} + (1-\lambda)*1/Rank_{2i}$, ($\lambda<1$)	4.2

5 Overlap Measurement Strategy Based on Term Expansion

On eliminating repetitive information, rather than concept of similarity, we used the concept of sentence overlapping. It represents the extent of the information taken by one sentence overlapped by another one.

This overlapping measure is unsymmetrical to the compared two sentences. Our experimental results show it is better than the symmetrical measure of similarity. Eq5.1 shows the overlapping of document B by document A, where A is the document preceding B.

$$Overlap\,B_A = \frac{A \cap B}{B} \qquad 5.1$$

Then the overlapping factor of B is $max\{Overlap\ B_i|$ document i preceding $B\}$.

In repetitive information elimination, term expansion was performed. Suppose the two sentences that should be compared are D_1 and D_2, the expanded parts of the original sentences are E_1 and E_2 respectively. Then the basic idea of elimination with term expansion (*TE*) is shown as Eq5.2.

$$OverlapTE(D_1, D_2) = Overlap(D_1, D_2) + \Delta\, Overlap(E_1, D_2) + \Delta\, Overlap(E_2, D_1) \qquad 5.2$$

Table 5.1 shows the result of eliminating repetitive information by using standard qrels of relevant information as the input of the second step. It seems that the dataset used in TREC2002 is not redundant enough for testing the system ability of finding new information.

Table 5.1 Effects of repetition elimination by using qrels of relevant

	Ave precision	Ave recall	Ave P*R
Qrels of relevant info, no elimination	0.91	0.99	0.905
Elimination without TE	0.92	0.99	0.904
Elimination with TE	0.92	0.98	0.900

6 Special Issues

6.1 Finding Keyword in Topics

In Novelty track, all the four domains of the topic can be used to retrieval, while the most useful information is taken by only several keywords. Therefore, finding key information from the topic is an important issue. We classified words in the topic into three classes by statistical learning and rule-based learning: *useful keywords* that contain the most useful words and were used to perform retrieval, *general describing words* that contain little information and were discarded directly and *negative words* that were applied to refine retrieval results.

To remove the topic-free words that contain no more information on describing the topic, two statistical learning methods were performed. Suppose the impact factor of the term is IF_i, terms with impact factor lower than a threshold were general description words. IF_i can be calculated by the two approaches:

$IF_i = qtf_i / sum_i$ 6.1	$IF_i = tf_i / n_i$ 6.2

Where qtf_i is the term frequency for t_i in the topic, sum_i is the summation of qtf_i in past TREC queries, tf_i is the term frequency in relevant documents and n_i is the number of documents that the term occurs.

6.2 Dynamic Result Selection

In general information retrieval experiments, the system returns fixed number of results to all the topics. In most cases, however, different topic has different number of relevant documents. Therefore, h*ow many is enough* is an important issue. We give the algorithm to select the documents whose similarity and rank fit in with the thresholds. Figure 6.1 and 6.2 show the effects of dynamic result selection.

7 Runs Submitted

Table 7.1 show the runs we submitted in novelty experiments, where *DOTC, TOTC* and *QE*(*LCE*) have the same definition of Table 4.1. *Comb_QE_DE* is the combining inversed rank of *QE* and *DE.* The first step

results of above four results are got by Okapi system. And the last result is got by our new system with short query. All the second step results were got by the new system.

Figure 6.1 Result number deduction

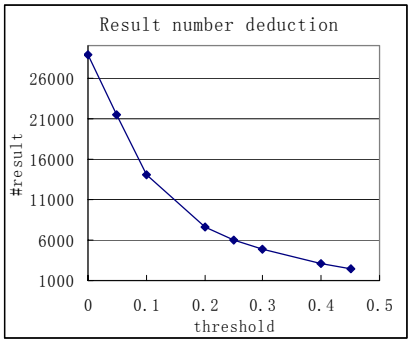
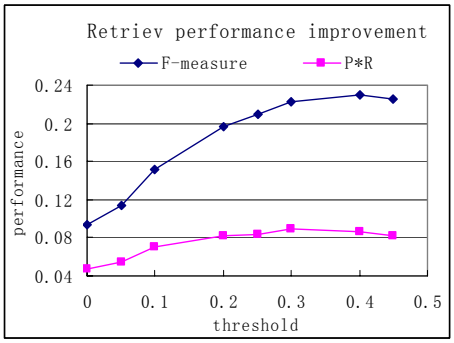


Figure 6.2 Retrieval performance improvement



	Finding relevant information			Elimination repetitive information		
	Ave P	Ave R	Ave P*R	Ave P	Ave R	Ave P*R
Thunv1. DOTC	0.23	0.34	0.086	0.22	0.30	0.073
Thunv2. TOTC	0.23	0.34	0.087	0.23	0.29	0.074
Thunv3. Comb_QE_DE	0.20	0.41	0.088	0.20	0.35	0.073
Thunv4. QE(LCE)	0.21	0.34	0.081	0.21	0.28	0.067
Thunv5. New System	0.19	0.35	0.066	0.18	0.31	0.060

Table 7.1 Submitted runs and evaluation results of Tsinghua University in TREC2002 novelty Track

8 Conclusion and Discussion

In this year's TREC experiments, we mainly focused on the expansion-related technologies. Besides thesaurus based QE, which made only a little progress, we studied a new statistical expansion approach, called *local co-occurrence expansion*. The results are extremely good. It made consistent great progress not only in recall but also in precision. Furthermore, we proposed a novel document term expansion (*DE*) approach. Experimental results proofed encouraging effect of DE. Combinations of *QE* and *DE* by topic classification lead to better performance than either approach. On eliminating repetitive information, rather than concept of similarity, we used the concept of overlap with term expansion. Unfortunately however, it did not take improvement in the experiments.

However, it seems that the dataset used in TREC2002 is not redundant enough for testing the system ability of finding new information, which may influence the conclusion of effectiveness of different approaches. We still take an optimistic view of redundancy elimination technology based on term expansion and overlap measurement.

References

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1999

[2] Robertson, S. E., and Walker, S. (1999). Okapi/Keenbow at TREC-8. In *TREC-8*.

[3] Gianni Amati, Claudio Carpineto and Giovanni Romano, FUB at TREC-10 Web Track: A probabilistic framework for topic relevance term weighting. In TREC-10.

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

[5] Dekang Lin. MiniParser. http://www.cs.ualberta.ca/~lindek/minipar.htm.