Use of Noun Phrases in Interactive Search Refinement

Olga Vechtomova
Department of Management Sciences
University of Waterloo
200 University Avenue West, Waterloo, Canada
ovechtom@engmail.uwaterloo.ca

Murat Karamuftuoglu
Department of Computer Engineering
Bilkent University
06800 Bilkent Ankara, Turkey
hmk@cs.bilkent.edu.tr

Abstract

The paper presents an approach to interactively refining user search formulations and its evaluation in the new High Accuracy Retrieval from Documents (HARD) track of TREC-12. The method consists of showing to the user a list of noun phrases, extracted from the initial document set, and then expanding the query with the terms taken from the phrases selected by the user. The results show that the method yielded significant gains in retrieval performance. The paper also discusses post-TREC experiments conducted to explore the use of Pointwise Mutual Information measure in selecting multiword units for query expansion and the use of n-grams in the search process.

1 Introduction

Query expansion following relevance feedback is a well-established technique in information retrieval, which aims at improving user search performance. It combines user and system effort towards selecting and adding extra terms to the original query. The traditional model of query expansion following relevance feedback is as follows: the user reads a representation of a retrieved document, typically its full-text or abstract, and provides the system with a binary relevance judgement. After that the system extracts query expansion terms from the document, which are added to the query either manually by the searcher – interactive query expansion, or automatically – automatic query expansion. Intuitively interactive query expansion should produce better results than automatic, however this is not consistently the case (Beaulieu 1997, Koenemann and Belkin 1996, Ruthven 2003).

In this paper we present a new approach to interactive query expansion, which we developed and tested within the framework of the High Accuracy Retrieval from Documents (HARD) track of TREC (Text Retrieval Conference).

1.1 HARD track

The Text REtrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology and U.S. Department of Defense, was started in 1992 to support research into large-scale evaluation of text retrieval methodologies.

The main goal of the new HARD track in TREC-12 is to explore what techniques could be used to improve search results by using two types of information:

1. Extra-linguistic contextual information about the user and the information need, which was provided by track organisers in the form of metadata. It specifies the following:

- *Genre* – the type of documents that the searcher is looking for. It has the following values:
  - Overview (general news related to the topic);
  - Reaction (news commentary on the topic);
  - I-Reaction (as above, but about non-US commentary)
  - Any.
- *Purpose* of the user's search, which has one of the following values:
  - Background (the searcher is interested in the background information for the topic);
  - Details (the searcher is interested in the details of the topic);
  - Answer (the searcher wants to know the answer to a specific question);
  - Any.
- *Familiarity* of the user with the topic on a five-point scale.
- *Granularity* – the amount of text the user is expecting in response to the query. It has the following values: Document, Passage, Sentence, Phrase, Any.
- *Related text* – sample relevant text found by the users from any source, except the evaluation corpus.

2. Relevance feedback given by the user in response to topic clarification questions. This information was elicited by each site by means of a (manually or automatically) composed set of clarification forms per topic. The forms are filled in by the annotators (users), and provide additional search criteria.

In more detail the HARD track evaluation scenario consists of the following steps:

1) The track organisers invite annotators, each of whom formulates one or more topics. An example of a typical HARD topic is given below:

Title: Red Cross activities

Description: What has been the Red Cross's international role in the last year?

Narrative: Articles concerning the Red Cross's activities around the globe are on topic. Has the RC's role changed? Information restricted to international relief efforts that do not include the RC are off-topic.

Purpose: Details

Genre: Overview

Granularity: Sentence

Familiarity: 2

2) Participants receive Title, Description and Narrative sections of the topics, and use any information from them to produce one or more baseline runs.

3) Participants produce zero or more clarification forms with the purpose of obtaining feedback from the annotators. Only two forms were guaranteed to be filled out and returned. According to the HARD track specifications, a clarification form for each topic must fit into a screen with 1152 x 900 pixels resolution, and the user may spend no more than 3 minutes filling out each form.

4) All clarification forms from different sites for a topic are filled out by the annotator, who has composed that topic.

5) Participants receive the topic metadata and the annotators' responses to clarification forms, and use any data from them to produce one or more final runs.

6) Two runs per site (baseline and final) are judged by the annotators. Top 75 documents, retrieved for each topic in each of these runs, are assigned binary relevance judgement by the annotator – author of the topic.

7) The annotators' relevance judgements are then used to calculate the performance metrics (see section 3).

The evaluation corpus used in the HARD track consists of 372,219 documents, and includes three newswire corpora (New York Times, Associated Press Worldstream and Xinghua English) and two governmental corpora (The Congressional Record and Federal Register). The overall size of the corpus is 1.7Gb.

The primary goal of our participation in the track was to investigate how to achieve high retrieval accuracy through relevance feedback. The secondary goal was to study ways of reducing the amount of time and effort the user spends on making a correct relevance judgement.

Traditionally in bibliographical and library IR systems the hitlist of retrieved documents is represented in the form of the titles and/or the first few sentences of each document. Based on this information the user has to make initial implicit relevance judgements: whether to refer to the full text document or not. Explicit relevance feedback is typically requested by IR systems after the user has seen the full-text document, an example of such IR system is Okapi (Robertson et al. 2000, Beaulieu 1997). Reference to full text documents is obviously time-consuming, therefore it is important to represent documents in the hitlist in such a form, that would enable the users to reliably judge their relevance without referring to the full text. Arguably, the title and the first few sentences of the document are frequently not sufficient to make correct relevance judgement. Query-biased summaries, usually constructed through the extraction of sentences that contain higher proportion of query terms than the rest of the text – may contain more relevance clues than generic document representations. Tombros and Sanderson (1998) compared query-biased summaries with the titles plus the first few sentences of the documents by how many times the users have to request full-text documents to verify their relevance/non-relevance. They discovered that subjects using query-biased summaries refer to the full text of only 1.32% documents, while subjects using titles and first few sentences refer to 23.7% of documents. This suggests that query-biased representations are likely to contain more relevance clues than generic document representations.

We have experimented with a similar approach in HARD track. Given the restrictions on the available space for relevance feedback, we created micro-summaries that consisted of single sentences for each of the top ranked documents in the baseline run. The sentences were selected according to concentration of the content-bearing and query-related words in them. The users were asked to select those sentences that might indicate relevant documents. Contrary to our preliminary experiments, the official HARD track runs for this method was not successful (Vechtomova 2004). In this paper we, therefore, describe the more successful method of directly eliciting query expansion terms from the users and evaluation of its effectiveness in HARD track of TREC 2003.

The method extracts noun phrases from top-ranked documents retrieved in the baseline run and asks the user to select those, which might be useful in retrieving relevant documents. The selected phrases are then used in constructing an expanded query, which retrieves a new set of documents. This approach aims to minimise the amount of text the user has to read in relevance feedback, and to focus the user's attention on the key information clues from the documents.

The remainder of this paper is organised as follows: section 2 presents the query expansion method we developed, section 3 discusses its evaluation, sections 4 and 5 describe post-TREC experiments we have conducted with phrases. Section 6 concludes the paper and outlines future research directions.

2 Query Expansion Method

The user feedback mechanism that we evaluated consists of automatically selecting noun phrases from the top-ranked documents retrieved in the baseline run, and asking the users to select all phrases that contain possibly useful query expansion terms.

The research question explored here is whether noun phrases provide sufficient context for the user to select potentially useful terms for query expansion.

An important question in query expansion is which part of the document should be used in extracting expansion

terms/phrases. Two common approaches in IR are: (1) to extract candidate terms from the entire document; (2) to extract them only from the best matching passages. The rationale for the second approach is that documents may be about multiple topics, not all of which are relevant to the user's query, therefore we would reduce the amount of noise by extracting terms/phrases only from those parts of the documents, which are likely to be related to the user's query.

We developed a method of selecting sentences in the documents, which are (1) most likely to be related to the query, and (2) have high information density. The best *n* sentences are then used for extracting noun phrases. In more detail the sentence selection algorithm is outlined below.

In all our experiments we used an experimental IR system *Okapi* (Robertson et al. 2000), and its best-match search function *BM25*.

2.1 Sentence selection

The sentence selection algorithm consists of the following steps:

We take N top-ranked documents, retrieved in response to query terms from the topic title. The full-text of each of the documents is then split into sentences. For every sentence that contains one or more query terms, i.e. any term from the title field of the topic statement, two scores are calculated: S1 and S2.

Sentence selection score 1 (S1) is the sum of *idf* of all query terms present in the sentence.

$$S1 = \sum idf_q \qquad (1)$$

Sentence selection score 2 (S2):

$$S2 = \frac{\sum W_i}{f_s} \qquad (2)$$

Where: W_i – Weight of the term *i*, see (3);
f_s – length normalisation factor for sentence *s*, see (4).

The weight of each term in the sentence, except stopwords, is calculated as follows:

$$W_i = idf_i(0.5 + (0.5 * \frac{tf_i}{t\max})) \qquad (3)$$

Where: idf_i – inverse document frequency of term *i* in the corpus; tf_i – frequency of term *i* in the document; *tmax* – *tf* of the term with the highest frequency in the document.

To normalise the length of the sentence we introduced the sentence length normalisation factor *f*:

$$f_s = \frac{s\max}{slen_s} \qquad (4)$$

Where: *smax* – the length of the longest sentence in the document, measured as the number of non-stopwords it contains; *slen* – the length of the current sentence.

All sentences in the document were ranked by S1 as the primary score and S2 as the secondary score. Thus, we first select the sentences that contain more query terms, and therefore are more likely to be related to the user's query, and secondarily, from this pool of sentences select the one which is more content-bearing, i.e. containing a higher proportion of terms with high *tf***idf* weights.

2.2 Noun phrase selection

We take top 25 documents from the baseline run, and select 2 sentences per document using the algorithm described above. We have not experimented with alternative values for these two parameters.

We then apply Brill's rule-based tagger (Brill 1995) and BaseNP noun phrase chunker (Ramshaw and Marcus 1995) to extract noun phrases from these sentences.

The phrases are then parsed in Okapi to obtain their term weights, removing all stopwords and phrases consisting entirely of the original query terms. The remaining phrases are ranked by the sum of weights of their constituent terms. Top 78 phrases are then included in the clarification form for the user to select. This is the maximum number of phrases that could fit into the clarification form.

All user-selected phrases were split into single terms, which were then used to expand the original user query. The expanded query was then searched against the HARD track database using Okapi BM25 search function. The official TREC evaluation results are discussed in section 3.

Following TREC 2003 we have also experimented with:

1) the use of phrases in searching instead of single terms;
2) the use of an association measure (pointwise mutual information) in selecting noun phrases for query expansion.

These experiments and their results are discussed in sections 4 and 5.

3 Evaluation

The runs submitted to the HARD track were evaluated in three different ways. The first two evaluations are done at the document level only, whereas the last one takes into account the granularity metadata.

1. SOFT-DOC – document-level evaluation, where only the traditional TREC topic formulations (title, description, narrative) are used as relevance criteria.

2. HARD-DOC – the same as the above, plus 'purpose', 'genre' and 'familiarity' metadata are used as additional relevance criteria.

3. HARD-PSG – passage-level evaluation, which in addition to all criteria in HARD-DOC also requires that retrieved items satisfy the granularity metadata (Allan 2004).

Document-level evaluation was done by the traditional IR metrics of mean average precision and precision at various document cutoff points. In this paper we focus on document-level evaluation. Passage-level evaluation is discussed elsewhere (Vechtomova et al. 2004).

3.1 Document-level evaluation

For all of our runs we used Okapi BSS (Basic Search System). For the baseline run we used keywords from the title field only, as these proved to be most effective in our preliminary experiments. Topic titles were parsed in Okapi, weighted and searched using BM25 function against the HARD track corpus.

Document-level results of the submitted runs are given in table 1. UWAThard1 is the baseline run using original query terms from the topic titles. UWAThard3 is an experimental run using the query expansion method described earlier. Query expansion resulted in 18% increase in average precision (SOFT-DOC evaluation) and 26.4% increase in average precision (HARD-DOC evaluation). Both improvements are statistically significant (using t-test at .05 significance level). On average 19 phrases were selected by users per topic.

Run	SOFT-DOC Evaluation		HARD-DOC evaluation	
	P@ 10	AveP	P@ 10	AveP
UWAThard1 (baseline run)	0.4875	0.3134	0.3875	0.2638
UWAThard3 (experimental run)	0.5958	0.3719	0.4854	0.3335

Table 1: Document-level evaluation results

In total 88 runs were submitted by participants to the HARD track. All our submitted runs are above the median in all evaluation measures shown in table 1. The only participating site, whose expansion runs performed better than our UWAThard3 run, was the Queen's college group (Kwok et al. 2004). Their best baseline system achieved 32.7% AveP (HARD-DOC) and their best result after clarification forms was 36%, which gives 10% increase over the baseline. We have achieved 26% improvement over the baseline (HARD-DOC), which is the highest increase over baseline among the top 50% highest-scoring baseline runs.

3.2. Analysis of performance by topic

We have conducted a topic-by-topic analysis of its performance in comparison with the baseline. Figure 1 shows the average precision (HARD-DOC) of these two runs by topic. It is not surprising, that performance of query expansion following blind feedback tends to depend on performance of the original query. The fewer relevant documents are retrieved at the top of the ranked list by the original query, the fewer good candidate query expansion terms are extracted, and hence the lower is the performance of the expanded run. This tendency is evident from Figure 1.

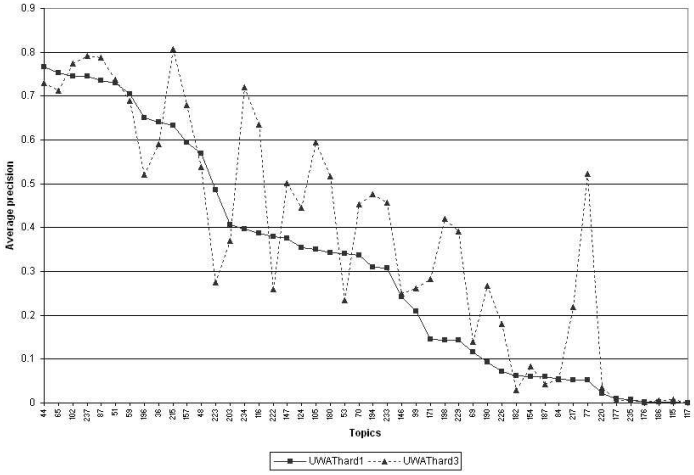

Figure 1: Results by topic of the baseline (UWAThard1) and the query expansion run (UWAThard3)

We have analysed two groups of topics: (1) topics, which yielded significantly worse results in runs with the expanded query (UWAThard3) than runs with the original query terms (baseline); and (2) topics, which had low performance both with the original and the expanded queries. Some examples of topic titles in the first group are: "Corporate mergers" (topic 222), "Sports scandals" (223), "Oscars" (53) and "IPO activity" (196). One factor that all of these topics have in common is that query expansion phrases selected by the users from the candidate phrases shown to them contain a large number of proper names.

Generally, proper names are considered to be good candidates for query expansion, as they usually have relatively low collection frequency. However, in our current model, we break user-selected multi-term phrases into their constituent terms and use them in the search process. For example, a proper name "Dan Leonard" selected by users for query expansion in topic 223 ("Sports scandals") was decomposed into single terms, each of which could match references to unrelated individuals. This results in many false matches. The situation is also aggravated in many cases by high *idf* values of some of the proper name components, which dominate the search results.

Examples of topic titles in the second group are: "National leadership transitions" (187), "School development" (182), "Virtual defense" (115), "Rewriting Indian history" (177) and "Restricting the Internet" (186). The majority of terms in these queries have very high number of postings, which suggests that they are either topic-neutral words (e.g., restrict, rewrite, transition), or they represent ideas or entities that were popular in

newswire and governmental publications at the time (e.g., Internet, Indian). Moreover, these queries do not represent fixed phrases, i.e., that co-occur frequently in English language. Compare queries in this group to the query "Mad cow disease" (65), which performed very well. Although, the number of postings of individual terms is very high, the query represents a fixed expression, which occurs as a phrase in 213 documents.

Another reason of failure, which applies to both groups above, is over-stemming. We used Porter's stemmer with the strong stemming function in our searches. This function reduces various derivatives of the same lexeme to a common stem. For example, topic "Product customization" failed, because stems 'product' and 'custom' matched such words as 'production', 'productivity', 'customer', 'customs'. Strong stemming is seen as a recall-enhancing technique. Weak stemming is likely to be more appropriate to the HARD task, as we are more interested in achieving high precision, rather than recall. Weak stemming keeps suffixes, and removes only endings, such as plural forms of nouns and past tense forms of verbs.

Another common reason for failure is that, some topic titles simply have insufficient information, for example in topic 186 ("Restricting the Internet"), the Description and Narrative sections narrow down the relevance criteria to the documents related to governmental restrictions of the Internet use in China.

4 Use of Statistical Association Measures for Noun Phrase Selection

In this and the following section we describe post-TREC experiments that we have conducted with the goal of better understanding the effect of noun phrases on retrieval performance.

In the phrase selection method, described in section 2, noun phrases, output by the noun phrase chunker, were ranked using the average idf of their constituent terms. This method is suitable for determining the informativeness of the individual words in the phrase, however it does not tell us whether the n-gram is a fixed multiword unit or a chance co-occurrence of words. In our HARD track experiments we noted that some of the phrases output by the chunker were, in fact, unsatisfactory chance word combinations.

We decided to explore the question whether the use of multiword units selected by Pointwise Mutual Information (Church et al. 1991) rather than any n-grams extracted by the phrase chunker in selecting terms for query expansion would result in better retrieval performance. We have conducted several experimental runs to address the above question, which are outlined below:

Run 1: All n-grams (n >= 2), output by the noun phrase chunker, were ranked by the average *idf* of their constituent terms. Single terms from the top *m* phrases were added to the original query terms from the topic title and searched using BM25 function.

Run 2: From each n-gram (n >= 2), output by the noun phrase chunker, we extracted all bigrams of adjacent words. For each bigram, *Pointwise Mutual Information* (*PMI*) was calculated as follows:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} = \frac{f(x,y)N}{f(x)f(y)} \qquad (5)$$

Where:
f(x,y) - the number of documents in the corpus containing words *x* and *y* adjacent to each other and in the same order of occurrence;
f(x) and *f(y)* - the numbers of documents that contain words *x* and *y* respectively;
N – the number of documents in the corpus.

The reasons for using numbers of documents instead of word frequencies are pragmatic: numbers of documents are easily obtainable from the IR system Okapi, whereas calculating actual bigram frequencies is computationally expensive. We recognise, however, that the *PMI* score is more accurate when word frequencies are used.

N-grams were ranked by the highest *PMI* of their constituent bigrams. This is a rather crude selection method, but given the fact that we apply it to syntactically selected noun phrases, it is likely to produce satisfactory results.

Single terms from the top *m* n-grams were used in retrieval in the same manner as in run 1.

Run 3: *PMI* has limitations as a tool for selecting strongly associated bigrams, one of which is that it is biased towards low frequency words. Following Manning and Schuetze (1999), we used *I(x,y)*f(x,y)* for ranking bigrams in this run. The other parameters in this run are the same as in run 2.

Run 4: The same as run 1; all n-grams with n >= 1 are used.

Run 5: The same as run 4; n-grams containing no bigrams with PMI > 0 are removed.

In all runs the number of query expansion terms/phrases (*m*) was set to 30. The results are presented in table 2.

Run	Precision @ 10	Average Precision
Run 1	0.5220	0.2837
Run 2	0.5180	0.2730
Run 3	0.5220	0.2782
Run 4	0.4980	0.2805
Run 5	0.5200	0.2776

Table 2: Evaluation results

The results do not provide any evidence that *PMI* is more useful than *idf* in selecting phrases for automatic query expansion. We have not tried other association measures, which may produce different results than *PMI*. Nevertheless, *PMI* or *I(x,y)*f(x,y)* can still be useful in selecting candidate query expansion phrases to be shown to the user in interactive query expansion. Table 3 shows top phrases ranked by *idf* , and *I(x,y)*f(x,y)* for the topic (180) "Euro Introduced".

Top phrases ranked by average idf of their constituent terms	Top phrases ranked by the highest I(x,y)*f(x,y) of their constituent bigrams
VietCombank Ho Chi Minh City	Japanese yen
emotive topic	U.S dollar
15-nation bloc	VietCombank Ho Chi Minh City
VND-euro exchange rate	central bank
rollercoaster day	monetary policy
unified currency	exchange rate risks
HKFE's other Rolling Forex futures contracts	dollar euro exchange rate
Euro Transaction	exchange rate stability
Shorten Euro Transition Period Brussels	VND-euro exchange rate
third pillar	percentage point
BSS	15-nation bloc
Rolling Forex Euro Futures Contract	euro trades
tighter controls	oil pricing
euro zone	neutral stance
Own Single Currency	euro zone nations
euro bonds	euro zone
Monday's local newspaper De Morgen	currency traders
euro's launch	Monday's local newspaper De Morgen

Table 3: Top-ranked phrases (phrases in the shaded cells are selected by both methods).

5 Use of Phrases vs. Single Words in Search

Intuitively, the use of phrases, such as compound terms and proper names in search is expected to result in higher precision than the use of their constituent words separately.

We hypothesise that adjacent pairs of words, which have strong degree of association, will result in higher search precision when used in search as a phrase, as opposed to when used as single words.

To test this hypothesis we conducted an experiment, comparing two experimental retrieval runs against the baseline. The runs are described in more detail below.

Baseline run: All terms from TREC titles were used in search as single terms. BM25 search function was used to perform the search. This was exactly the same way we searched in the baseline run of HARD track.

Experimental run 1: All bigrams of adjacent words in TREC titles were extracted. For each bigram, Pointwise Mutual Information was calculated.

All bigrams with PMI > 0 are used in search as a phrase, i.e. using *Adjacency*[1] operator. Bigrams with PMI < 0 are split into individual words.

[1] *Adjacency* is a pseudo-Boolean operator, which retrieves an unranked set of all documents, which contain the specified terms in adjacent positions in the same order as they were entered in the search statement.

For example, in the topic title "Amusement Park Safety"

I (amusement, park) = 1.66

I (park, safety) = -7.6

The logical representation of the final query will be:

(amusement ***Adjacency*** park) ***BM25*** safety

Experimental run 2: The same as the experimental run 1 above, but all terms in the title are also added to the query as single terms, for example:

(amusement **Adjacency** park) **BM25** safety **BM25** amusement **BM25** park

The rationale behind including all terms into the query as single terms, is to relax the search criteria: if the phrase is rare, and retrieves only few documents, use of single terms will ensure that other documents which contain parts of the phrase will also be retrieved. Typically, phrases have quite high idf, therefore top retrieved documents are very likely to contain the phrases, used in the query.

Only 16 topic titles out of 50 had any bigrams with positive PMI. Both experimental runs had worse overall average precision than the baseline (see table 4). Only 2 topics in the experimental run 1 had better AveP than the baseline, whereas 7 topics in the experimental run 2 had better AveP than the baseline (see figure 2).

Run	Precision @ 10	Average precision
Single terms (baseline)	0.5240	0.3087
Bigrams + remaining single terms (experimental run 1)	0.5000	0.2819
Bigrams + all single terms (experimental run 2)	0.5180	0.3065

Table 4: Phrase search evaluation results

One of the reasons for this counter-intuitive result could be the fact that the bigrams may contain terms that are themselves in the query; either as single terms or as part of other bigrams. Robertson and his colleagues suggest that search term weighting should take into account the case of bigrams that have as their constituent terms single query terms, and propose a weighting scheme that adjusts their weights (Robertson et al. 2004). However, their method does not take into account the case of two or more bigrams that share a common term. We need further research to understand and deal with the complex case of bigrams containing other query terms, either, those which are part of other bigrams or exist as single terms in the query.

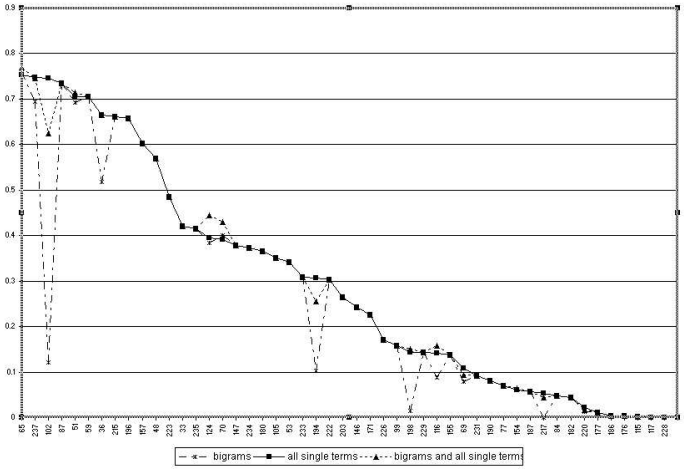

Figure 2: Results in average precision by topic of the two experimental runs and the baseline run.

6 Conclusions and Future Work

In this paper we presented a user-assisted search refinement technique, which consisted in showing to the user a list of noun phrases, extracted from the initial document set, and then expanding the query with the terms from the user-selected phrases.

The focus of our experiments in the HARD track of TREC-12 was on developing effective methods of gathering and utilising the user's relevance feedback. The evaluation results suggest that the expansion method overall is promising, demonstrating statistically significant performance improvement over the baseline run. More analysis needs to be done to determine the key factors influencing the performance of individual topics.

Post-TREC experiments conducted suggest that the use of PMI as a means of selecting n-grams for the purpose of term selection for query expansion is not promising. However, we should note that, there was a number of simplifying assumptions in the use of PMI for the above purpose, which might have had a negative impact in its usefulness. It seems, however, possible that the use of PMI multiplied by the joint term frequency in selecting candidate query expansion phrases would result in a better selection of phrases to be shown to the user in interactive query expansion. We intend to experiment with alternative association measures to draw more strong conclusions about the usefulness of multiword units in IR.

After the official TREC results, we also conducted experiments to explore the use of phrases in searching. The results were not positive and confirmed the conclusions of the previous experiments reported in IR literature. We noted the difficulty in adjusting the weights of bigrams in searching when their constituent terms include other search terms.

Acknowledgements

This material is based on work supported in part by Natural Sciences and Engineering Research Council of Canada.

References

Allan, J. (2004). HARD Track Overview. Proceedings of the Twelfth Text Retrieval Conference, November 18-21, 2003, Gaithersburg, MD.

Beaulieu, M. (1997). Experiments with interfaces to support Query Expansion. Journal of Documentation, 53(1), pp. 8-19

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging. Computational Linguistics, 21(4), pp. 543-565.

Church, K., Gale, W., Hanks, P. and Hindle D. (1991). Using statistics in lexical analysis. *In Lexical Acquisition: Using On-line Resources to Build a Lexicon*, ed. U. Zernik, Englewood Cliffs, NJ: Lawrence Elbraum Associates, pp. 115-164.

Koenemann, J. and Belkin, N. J. (1996). A case for interaction: a study of interactive information retrieval behavior and effectiveness. Proceedings of the Human Factors in Computing Systems Conference, Zurich, pp. 205-215.

Kwok, L. et al. (2004). TREC2003 Robust, HARD and QA track experiments using PIRCS. Proceedings of the Twelfth Text Retrieval Conference, November 18-21, 2003, Gaithersburg, MD.

Manning, C.D. and Schuetze, H. (1999). Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, Massachusetts, 1999.

Ramshaw, L. and Marcus, M. (1995). Text Chunking Using Transformation-Based Learning. Proceedings of the Third ACL Workshop on Very Large Corpora, MIT.

Robertson, S.E., Zaragoza, H. and Taylor, M. (2004). Microsoft Cambridge at TREC–12: HARD track. Proceedings of the Twelfth Text Retrieval Conference, November 18-21, 2003, Gaithersburg, MD.

Robertson, S.E., Walker, S. and Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. Information Processing and Management, 36, pp. 95-108.

Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. Proceedings of the 26th ACM-SIGIR conference, Toronto, Canada, pp. 213-220.

Sparck Jones, K., Walker, S. and Robertson, S.E. (2000). A probabilistic model of information retrieval: development and comparative experiments. Information Processing and Management, 36(6), pp. 779-808 (Part 1); pp. 809-840 (Part 2).

Tombros, A. and Sanderson, M. (1998). Advantages of Query Biased Summaries in Information Retrieval. Proceedings of the 21st ACM SIGIR conference, Melbourne, Australia, pp. 2-10.

Vechtomova, O., Karamuftuoglu, M. and Lam, E. (2004). Interactive Search Refinement Techniques for HARD Tasks. Proceedings of the Twelfth Text Retrieval Conference, November 18-21, 2003, Gaithersburg, MD.