The Role of Multi-word Units in Interactive Information Retrieval

Olga Vechtomova

Department of Management Sciences, University of Waterloo,
Waterloo, Canada
ovechtom@engmail.uwaterloo.ca

Abstract. The paper presents several techniques for selecting noun phrases for interactive query expansion following pseudo-relevance feedback and a new phrase search method. A combined syntactico-statistical method was used for the selection of phrases. First, noun phrases were selected using a part-of-speech tagger and a noun-phrase chunker, and secondly, different statistical measures were applied to select phrases for query expansion. Experiments were also conducted studying the effectiveness of noun phrases in document ranking. We analyse the problems of phrase weighting and suggest new ways of addressing them. A new method of phrase matching and weighting was developed, which specifically addresses the problem of weighting overlapping and non-contiguous word sequences in documents.

1 Introduction

Multiword units (MWUs), also commonly referred to in IR literature as 'phrases'[1], received much attention in information retrieval research throughout its more than 30-year old history. This interest can be partially attributed to the fact that phrases typically have a higher information content and specificity than single words, and therefore represent the concepts expressed in text more accurately than single terms. Ideally document and query representations should be mapped directly and unambiguously to the underlying concepts conveyed in text. However, at present, this still remains a difficult goal to reach. Most of the leading statistical IR models, such as probabilistic [1,2] and vector-space [3] rely on the use of single terms and are based on strong term independence assumptions to make them computationally tractable. Experimentally these models have consistently demonstrated high performance results with a variety of large test collections in the evaluation exercises such as TREC [4]. Nevertheless, many attempts have been made to introduce phrases into the retrieval process, but so far with mixed results.

MWUs comprise a wide variety of lexical associations with various degrees of idiomaticity or compositionality, such as named entities ('Tony Blair', 'United Nations'), nominal compounds ('amusement park', 'free kick') and phrasal verbs

[1] We will use these terms interchangeably throughout the paper.

D.E. Losada and J.M. Fernández-Luna (Eds.): ECIR 2005, LNCS 3408, pp. 403 – 420, 2005.

('reach out', 'kick the bucket'). Although MWUs can belong to different lexical categories, our focus is on nominal MWUs, primarily because nouns and noun phrases are considered to be the most content-bearing syntactic category. Also, there is some evidence from previous research that noun phrases hold more promise for query expansion in IR [5].

Query expansion is a widely used technique in IR. In automatic query expansion (AQE) additional terms or phrases are added to the original query by the system, whereas in interactive query expansion (IQE) users select terms or phrases manually. Terms and phrases for query expansion can be retrieved using statistical or linguistic methods from a variety of sources, the most common being top-ranked documents in the retrieved set (blind or pseudo-relevance feedback) and documents judged relevant by the user in the retrieved set (relevance feedback). Single-term interactive query expansion techniques were extensively evaluated in the past [24, 25]. Some researchers investigated the use of phrases in IQE (see section 2.3), however no systematic comparison of different types of phrases in IQE has been conducted so far. In this work we are interested in studying how different types of phrases can help users to interactively enhance their initial search formulation.

This paper has two foci:

1. To investigate the utility of multiword units (MWUs) in interactive query expansion;
2. To study the effectiveness of MWUs in the document ranking.

The main goal of the first focus of this study was to investigate the following hypotheses:

Hypothesis 1: Nominal MWUs are better candidates for interactive query expansion than single terms.

Hypothesis 2: Nominal MWUs which exhibit strong degree of stability in the corpus are better candidates for interactive query expansion than noun phrases selected by the frequency parameters of the individual terms they contain.

We used a combined syntactico-statistical approach for selecting nominal MWUs for interactive query expansion. In the first selection pass, noun phrases were obtained using a part-of-speech (POS) tagger and a noun phrase chunker. In the second pass, statistical measures were applied to select strongly bound MWUs. In particular, we have experimented with two statistical measures to select MWUs from text: the C-value [6] and the Log-Likelihood [7]. Selected MWUs were then suggested to the user for interactive query expansion. Techniques developed for the selection of MWUs are presented in section 3. Experiments investigating the above hypotheses and evaluation results are described in section 5.

The goal of the second focus of this work is to study the effectiveness of noun phrases in document ranking. We contribute to the previous findings in the field by further analysing the problems of phrase matching and weighting and suggesting new ways of addressing them. The following hypothesis was investigated:

Hypothesis 3: Ranking documents using phrases leads to better performance than ranking documents by single terms.

We have developed a new method of phrase-based document ranking, which specifically addresses the problem of weighting overlapping phrases in documents, which in statistical IR models like probabilistic ones [2] leads to the problem of the artificial over-inflation of the document score. The method is described in detail in section 4.

2 Previous Research

2.1 Statistical Versus Syntactical Phrases

Hypotheses claiming that phrases are better contents discriminators than single terms have been studied since the beginning of research on automated IR in the 60s. Simple statistical co-occurrence based techniques for identification of phrases have always been rivalled by NLP-based techniques. The main considerations in favour of NLP were: (1) it may have better tools to uncover meaningful linguistic phrases and (2) it can capture the syntactical relationships between words.

Statistical phrases are typically short-span collocations extracted from text using different modulations of their frequency parameters. Syntactical phrases are identified using a variety of NLP methods ranging from low-level techniques such as part-of-speech tagging, aimed at identifying word-sequences of a certain syntactic pattern like adjective + noun, to more complex methods like extended N-grams and shallow syntactic parsing, attempting to discover uniform semantic units underlying various forms of expression.

At the early stages the motivation for research on automatic phrase generation came from the determination to emulate human indexing. The belief was that complex normalising descriptions of the kind assigned to documents by human indexers are more useful than simple terms. One of the early experiments on phrase indexing was carried out by Bely [8], who used very elaborate NLP techniques to identify instantiations of thesaurus concepts and their semantic relationships in documents. Despite the fact that no retrieval evaluation was conducted, the research suggested that the relational structure of the descriptions was not flexible enough for sufficient matching. Another historically important piece of research was undertaken by Salton [11], whose technique consisted in identification of thesaurus terms in text supported by syntactic analysis. The comparison of performance results for syntactic phrases and for statistical phrases, defined as within-sentence co-occurrences of thesaurus descriptor constituents, showed that there was no performance improvement in using syntactical phrases over simple statistical phrases.

One of the most comprehensive early evaluations of phrases in IR was undertaken by Fagan [9,10]. The main focus of his experiments was systematic evaluation of statistical phrases under different parameter settings, such as distance between their constituents and their frequency values. The evaluation results showed that performance for statistical phrases was in general better than for single terms. He then

compared performance for statistical phrases with performance for syntactical phrases, which he obtained using syntactic parsing, stemming and normalisation to head-modifier pairs. The evaluation showed that linguistically-derived phrases gave results similar to or worse than statistically extracted phrases. When he analysed earlier work taking into account his findings, he concluded that the same pattern, statistical phrases ≥ syntactical phrases ≥ single terms, was evident in all the experiments. The performance gains from the use of statistical phrases obtained in his experiment were in the range of 17% to 39%. He concluded that syntactical phrases gave poor performance because queries and documents did not share exactly the same phrases. Among the reasons for the systems' inability to match documents and queries by syntactic phrases, Fagan pointed out the low collection frequency of the best phrases and the fact that the documents involved were abstracts. Stzalkowski et al. [12] pointed to another main reason for this, namely, the limited amount of information about the user's information need conveyed by the queries.

It is worthwhile to note that the above earlier studies of phrases in IR were undertaken on rather small collections (for example Fagan used a 10MB CACM collection of abstract-length documents). The last decade in IR research saw two major changes: (1) statistical models using single term weighting have been refined to achieve very high and robust performances; (2) the size of test collections has grown dramatically. Some of the phrase indexing and search techniques which used to work well with the old retrieval techniques on small collections, no longer give positive results.

More recent study of syntactic and statistical phrases was undertaken by Mitra et al. [13]. By statistical phrases they understood contiguous bigrams of non-stopwords which occur in at least 25 documents. Syntactical phrases were defined in their experiments as specific POS-tag sequences (e.g. Noun-Noun, Adjective-Noun). Their studies demonstrate that overall both statistical and syntactical phrases have very little effect on performance. Syntactical phrases showed marginally better performance than statistical phrases when used on their own (i.e. without single terms) in retrieval. An interesting finding, which emerged from their study, is that phrases tend to improve precision at higher recall levels, and have little or no effect on precision at lower recall levels. This suggests that phrase search may not prove to be a "precision-enhancing technique", but rather a "recall-enhancing technique".

2.2 Phrase Weighting

We consider that one of the major and yet unsolved problems of phrase-based techniques is weighting. Phrases like single terms vary in their contents-discriminating ability, so it may be possible to treat a phrase in the same way as a single term, and calculate, for example, its inverse document frequency (idf) in the same manner. However phrases also have other characteristics, which single terms do not have, and which may need to be reflected in their weighting. One of the most prominent characteristics of phrases is the degree of the stability in the corpus. We distinguish the following types of phrases by their stability in the corpus:

1. Combinations of terms which occur only with each other in many document collections, for example “Burkina Faso”.
2. Combinations of terms which frequently occur together as a phrase and whose syntactic structure does not permit any changes (i.e. intervening words, change of word order), for example “amusement park”, “stainless steel”, “acrylic paint”. Typically, one or all terms in such phrases may form lexical-syntactic constructions with other terms as well. If the expression has some degree of idiomaticity (i.e. the phrase as a whole has a different meaning than the combination of individual meanings of its parts), for example “Mad Cow Disease”, we may not be able to substitute all or some of the words with related or synonymous words without the radical change of meaning. For example we cannot substitute “mad” with “crazy” in the above example.
3. And finally combinations of terms which have strong degree of flexibility, namely allow intervening words, change of word order, substitution of phrase components with synonyms, hypernyms or hyponyms. For example the exact meaning underlying the phrase “animal protection” can be also represented in text as “protection of animals”. The word “animal” can be substituted with hyponyms, such as “reptile” or “mammal”.

The above categorisation of phrases has the following implications for IR:

- If the search on one term is highly likely to match on the entire phrase (what is typically the case with the phrases of the first category and some phrases of the second category above), then applying phrase search techniques will not be useful.
- If we search by a phrase belonging to the third category, it may be beneficial to relax search constraints to accommodate possible lexical-syntactic variations of the phrase. With this category of phrases, it may even be useful to relax search constraints to allow match on terms separated by longer distances, in order to capture within-topic relations between terms, rather than only phrasal relations.

The integration of phrase-search into the IR models, which were designed for single-term indexing and searching, is problematic. For example a probabilistic model of IR [2] calculates the document score by non-linearly combining weights of query term occurrences in the document. Phrases may be treated by the model in the same way as single terms, however Robertson et al. [14] pointed at the following problem: considering that a query may contain both single terms and phrases, and that some of the single terms may also be part of phrases, then the document matching on the phrase will also match on the single term. As a result both the weight of the phrase occurrence and the weight of the term occurrence will contribute to the document score, artificially inflating it. The solution suggested in [14] was to subtract the weight of the single term occurring in the query from the weight of the phrase, containing that term.

In this paper we examine phrase-weighting further and point at another problem that needs to be addressed, namely when the query contains two or more phrases which share one/more terms. In particular this situation can happen following query expansion, where the user or the system selects a number of phrases to be added to the original query. An example of such phrases is: “stainless steel” and “steel manufacturing”. If these phrases match the contiguous string “stainless steel manufacturing” in text, then we face a similar problem of over-inflating the document

score as pointed at in [14]. This problem, however, cannot be solved using their technique. We propose a new method of phrase matching and weighting in the document, which attempts to address this problem. The technique is presented in section 4.

2.3 Use of Phrases in Interactive Query Expansion

Phrases can play a useful role in interactive query expansion by helping the users to formulate their information need, in particular when the information need is vague, and the users do not know what exactly they are trying to find. Marchionini [15] and Smeaton and Kelledy [16] have argued that the process of formulating the query is more cognitively demanding on the part of the user than the process of selecting terms and phrases from the list, as the former involves recall, while the latter – recognition. According to cognitive psychology findings, recall is more demanding than recognition. Therefore in real-world search applications users prefer to formulate terse search statements, which tend to produce poor results, and then browse through the retrieved documents, finding more words and phrases and manually reformulating their queries. Extracting related terms/phrases from the documents retrieved by the original query and showing them to the user facilitates this process as the user does not have to go through large amounts of text.

Smeaton and Kelledy [16] have experimentally studied the usefulness of statistically-selected phrases in interactive query expansion. In particular they compared the effectiveness of user-selected phrases in search with the user-selected single terms and their combinations. They also looked at the differences between these techniques when used by novice and expert searchers. The best results are obtained when phrases are used in combination with single terms. Also phrase-based query expansion tends to be less effective with the novice searcher than the expert searcher.

The contribution of our study to the field of interactive query expansion is that we systematically evaluated the effect of different types of phrases and single terms on retrieval performance in the large-scale TREC experimentation settings.

3 Query Expansion Methods

In this section we describe the developed techniques for interactive query expansion using MWUs following blind feedback. The idea of blind (pseudo-relevance) feedback is to take top-ranked documents, retrieved using the original user's query and extract query expansion terms/phrases from them. Our approach is to extract query expansion phrases from query-biased summaries of the *n* top-ranked documents. We used a method proposed in [17] of building query-biased summaries which are composed of *m* sentences selected using two main factors: (1) the *idf* weights of the original query terms present in the sentence, and (2) information value of the sentence, i.e. the combined *tf.idf* value of its words.

In our experiments we used 2-sentence summaries of the 25 top-retrieved documents[2]. We then apply Brill's rule-based tagger [18] and the BaseNP noun

[2] These parameters showed good performance in the past experiments [17].

phrase (NP) chunker [19] to extract noun phrases from the document summaries. Multi-word units are then selected from the list of obtained noun phrases using the C-value and the Log-Likelihood. The two subsections below describe these techniques.

3.1 Selection of Query Expansion Phrases Using the C-Value

MWUs are characterised foremost by relative stability in the corpus. Some of the noun phrases output by the NP chunker are chance word groupings, and not stable MWUs. We were interested in exploring the value of MWUs compared to all noun-phrases in representing useful query expansion concepts to the user. The method of selecting stable MWUs from noun phrases using C-value is outlined below.

Noun phrases output by the NP chunker are ranked by the average *idf* of their constituent terms. For each phrase we generate the list of all phrases that it subsumes, i.e. contiguous or non-contiguous combinations of words in forward order, including the original complete phrase. For each subphrase, the C-value is calculated. The C-value is a measure of stability of an n-gram in the corpus [6]. The C-value formula we used is as follows [26]:

$$C - value\ (a) = (length\ (a) - 1)(\ freq\ (a) - \frac{t(a)}{c(a)}) \qquad (1)$$

Where:

t(a) – frequency of the phrase *a* in longer phrases;
c(a) – number of longer phrases including *a*;
freq(a) – frequency of the phrase *a* in the corpus;
length(a) – number of words in the phrase *a.*

All subphrases for a given phrase are ranked by the C-value. The top-ranked subphrase is then used to replace the original phrase in the list of candidate query expansion terms. The original complete phrase may get a higher C-value than any of its subphrases, in which case it is kept without changes.

For example, in our experiment, the bigram "World Cup" received the highest C-value out of all its subphrases generated from the phrase "grueling IAU 100-kilometer World Cup" and as a consequence was selected for the phrase list.

Some of the original noun phrases may contain intervening modifiers which are too specific. The reason why we considered non-contiguous word combinations is to eliminate such modifiers and to obtain the most stable and recurrent word combinations. The problem, however, is that some of the resulting phrases are too general (e.g. original phrase: *freak training accident*, selected sub-phrase: *freak accident*), or may have weak or no semantic relatedness to the original phrase (e.g., original phrase: *Moroccan born American runner Khalid Khannouchi*; selected sub-phrase: *born American*). As a result we may have strong topic drift and precision loss at the expense of having linguistically correct MWUs. We did not experiment with using only contiguous word combinations, which might help avoid some of the above problems, but remain for future work.

The obtained phrases are then ranked by their C-value, top *n* of which are shown to the user for interactive query expansion. Table 1 shows the 15 top-ranked phrases selected for the topic 404 "Marathon Training".

Table 1. Top 15 subphrases ranked by C-value and the original phrases from which they were derived (topic "Marathon Training")

Selected sub-phrase	Original phrase
World Cup	grueling IAU 100-kilometer World Cup
web site	marathon's web site
San Diego	San Diego Rock Roll Marathon
York City	York City Marathon
Olympic Games	Athens Olympic Games
training camp	training camp
world title	world half marathon title Paula Radcliffe
Athens Olympics	Athens Olympics
Medical Association	International Marathon Medical Directors Association
World Athletics	World Masters Athletics
Training Center	Duoba National Plateau Training Center
Olympic team	Olympic marathon team Athletics Kenya
training base	altitude training base
world's fastest	world's fastest
Road Race	25-kilometer 10-kilometer Road Race

3.2 Selection of Query Expansion Phrases Using the Log-Likelihood Ratio

The Log-Likelihood [20] has been extensively used for the identification of statistically significant word collocations in text and has shown good results for English.

$$Loglike\ (a,b) = 2 \times (\log\ \theta_1^{s1}(1-\theta_1)^{n1-s1} + \log\ \theta_2^{s2}(1-\theta_2)^{n2-s2} - \log\ \theta^{s1}(1-\theta)^{n1-s1} - \log\ \theta^{s2}(1-\theta)^{n2-s2}) \quad (2)$$

Where:

$s1 = f(a, b)$ $\quad s2 = f(b) - f(a, b)$

$$\theta_1 = \frac{s1}{n1} \qquad \theta_2 = \frac{s2}{n2} \qquad n1 = f(a) \qquad n2 = N - f(a)$$

$$\theta = \frac{f(b)}{N}$$

N – number of words in the corpus;
$f(a,b)$ – frequency of words a and b appearing together in text;
$f(a)$ – frequency of a; $f(b)$ – frequency of b.

The phrase weighting is done as follows: first, from each phrase output by the NP chunker all contiguous bigrams are derived. For each bigram, its Log-Likelihood score is calculated using the Ngram Statistics Package [21]. The highest Log-Likelihood score of any bigram derived from the phrase is taken as the phrase weight. Top n phrases ranked using this weighting scheme are shown to the user for

interactive query expansion. This is a rather crude phrase weighting method, but it does reward phrases which contain a strongly bound collocation which stands as a focus of our experiment.

4 Phrase-Based Document Retrieval

Following the interactive query expansion stage where the users select query expansion phrases, the next step is to use them in search. Intuitively using them as phrases in search should lead to better precision than if we split them into single words. One problem associated with the use of phrases in a statistical IR model, such as probabilistic [2] is that some terms may occur in multiple phrases. For example, we assume there are two phrases in the expanded query: "*air traffic*" and "*traffic control*", and two documents: the first containing one phrase "*air traffic control*", and the second – two phrases "*air traffic*" and "*traffic control*". How should they be weighted? If we calculate weights of each phrase in the document separately and then add them up to get the document score, as is currently done in the probabilistic model for single terms, then both documents would get equal scores. That obviously should not be the case. But then how should the phrase weight be calculated for the first document? The situation gets more complex if we allow for non-contiguous word combinations, i.e. matching the following: "1 *air* 2 *traffic* 10 *control*" (where numbers denote positions of the words in text). Allowing match on non-contiguous word combinations is good for recall as it relaxes search constraints, but the distance between the phrase elements should be inversely related to the phrase weight. Therefore, the two main issues to be addressed by the phrase search algorithm are:

- remove the problem of overlapping phrases;
- reflect the distance between the phrase elements in the phrase weight.

We have developed the following phrase search algorithm, which attempts to address the above problems:

The first step is to retrieve a set of documents using a best-match document retrieval function[3] and a query which consists of all single terms extracted from the query expansion phrases. The next step is to re-rank these documents by using phrase information. We take the top 1000 documents per topic in the retrieved set, stem the terms in each document and create a document representation, consisting only of the stemmed occurrences of terms from the query in their original order and their sequential position number in text.

For each query phrase, all possible subphrases (i.e. contiguous and non-contiguous, ordered and non-ordered combinations of words) are generated and recorded in the list ranked in the descending order of their length. For each subphrase in the list we use *cgrep* – a pattern matching program for extracting minimal matching strings [22] to extract the minimal spans of text in the document containing the subphrase. Each time cgrep returns matching strings, they are removed from the document representation and the procedure is repeated with the same phrase. If no matching

[3] We used the Okapi BM25 search function [2].

string is found, the program attempts to match the next phrase in the list, and so on. In this way we can match progressively longer spans containing the phrase or its subphrases. An example of extracted windows for the phrase "practical implementation" is given in figure 1 (the number preceded by the '#' sign is the sequential position of the following word in the original document text).

```
# 106 implementation # 120 practical
# 120 practical # 186 implementation
# 4 implementation
# 21 implementation
# 43 implementation
# 59 implementation
```

Fig. 1. An example of windows extracted from a document

As we can see, windows extracted using the above method might overlap. Our approach to eliminating overlaps in windows is a two-step process: (1) rank the windows by their weight and (2) remove overlapping words from the lower ranked windows.

4.1 Window Weighting

In this approach the window weight is calculated from the combination of *idf* weights of individual terms occurring in it. The following formula was used:

$$WindowWeight(w) = \sum_{i=1}^{n} idf_i \times \frac{n}{(span+1)^p} \quad (3)$$

Where:

i – word in the window w;
n – number of words in the window w;
$span = pos(n) - pos(1)$
where: $pos(i)$ – position number of the i[th] word in the window w;
p – tuning parameter[4].

So, the more informative the words in the window are, the shorter the span is, and the more words there are in the window, the higher is the weight of the window.

4.2 Removing Duplicate Windows

After the windows are ranked, we remove overlapping words by doing pairwise comparison of all windows. If two windows have overlapping word(s), these words are removed from the lower ranked window. The windows shown in figure 1 after the removal of overlapping words are illustrated in figure 2.

[4] Experiments showed that 0.2 gives the best performance on the HARD track 2003 corpus.

```
# 106 implementation # 120 practical
# 4 implementation
# 21 implementation
# 43 implementation
# 59 implementation
# 186 implementation
```

Fig. 2. An example of windows after the removal of overlapping words

All windows extracted for every phrase from the document are then added to the same list, weighted using the formula (3) above and have the overlapping words removed. For each window we also keep the index of the phrase which was used to extract it.

4.3 Calculating Document Scores

The next step is to calculate document scores. First, for each phrase in the query we calculate its weight in the document as follows:

$$PhraseWeight(a) = \frac{(k+1) \times \sum_{w=1}^{n} WindowWeight(w)}{k \times NF + n} \quad (4)$$

Where:

w – window, extracted for the query phrase a;
n – number of windows extracted for the phrase a;
NF – document length normalisation factor (see equation 5 below).
k – phrase frequency normalisation factor[5].

The document length normalisation factor was calculated in the same way as in the BM25 document ranking function [2]:

$$NF = (1-b) + b \times \frac{Doclen}{AveDocLen} \quad (5)$$

Where:

Doclen – document length (word count);
AveDoclen – average document length in the corpus;
b – tuning constant[6].

Document score is then calculated as the sum of *PhraseWeight* values for all query phrases that occur in the document:

$$DocumentScore(d) = \sum_{a=1}^{n} PhraseWeight(a) \quad (6)$$

[5] Experiments showed that k=1.2 gives the best performance on the HARD track 2003 corpus.
[6] Spärck-Jones et al. have experimentally determined that 0.75 gives best results on TREC data [2].

Where: a – the query phrase occurring in the document d;
n – number of query phrases occurring in the document d.

Finally the top 1000 documents in the originally retrieved set are re-ranked by the new document scores.

5 Evaluation

The testbed for our experiments is the Okapi IR system based on the Robertson/Spärck Jones probabilistic model of retrieval [2]. The evaluations of the developed techniques were conducted within the framework of the HARD (High Accuracy Retrieval from Documents) track of TREC 2004 [23, 27]. The HARD track evaluation framework includes an interactive component, which allowed us to test interactive query expansion techniques. The interactive evaluation experiment consists of the following steps:

1. TREC organisers release the search statements (topics) formulated by the annotators (users) in the traditional TREC format (Title, Description and Narrative) to the participating sites.
2. Participating sites use any information from the topics to produce the initial (baseline) document sets and compose clarification forms for the user to fill in. The purpose of clarification forms is to clarify or refine the annotator's search statement.
3. The annotator fills out clarification forms (with a 3-minute time limit per form).
4. Participating sites use the annotator's feedback to the clarification forms to improve the search (for example by query expansion). The end result is a new document set.
5. The annotator performs relevance judgements of the retrieved sets[7].

The HARD track test collection includes the document corpus (635,650 documents from eight newswire collections) and 50 topics. In addition to the traditional TREC topic fields of Title, Description and Narrative, the topics also contained several Metadata fields, describing various additional search criteria, such as "genre", "retrieval element" and "familiarity". We did not use any of the metadata except "retrieval element" in the runs reported here. In all expansion runs for topics with the retrieval element "Document" we used the Okapi document retrieval function BM25, and for topics with the retrieval element "Passage" we used the Okapi passage retrieval function BM250.

We conducted two baseline runs using only the information available in the TREC topics: in the run *baseTD*, we used all non-stopword terms extracted from the Title and Description fields of the topic and in *baseT*, we used all terms from the Title field only. For both runs we applied Okapi BM25 search function.

[7] Top 75 documents from two runs per site were added to the relevance judgement pool. Each document in the pool was assigned a binary relevance judgement. The same annotator who formulated the topic provided feedback to all clarification forms for that topic and performed relevance judgements.

Four clarification forms were generated for each topic. Phrases for each clarification form were extracted from 2-sentence query-biased summaries [17] of the top 25 documents retrieved in the run *baseTD*, as Title+Description gave higher performance than Title on HARD 2003 data.

- *1st clarification form:* top *n* phrases selected using the C-value method (section 3.1 above);
- *2nd clarification form:* single terms from the phrases displayed in the 1st clarification form;
- *3rd clarification form:* top *n* phrases output by the NP chunker and ranked by the average *idf* of their constituent terms;
- *4th clarification form:* top *n* phrases selected using the Log-Likelihood ratio (section 3.2 above).

The 2nd clarification form was introduced in order to investigate Hypothesis 1 (section 1), which suggests that users select better terms when they are shown to them in the context of phrases (in the 1st clarification form), than separately. By comparing the phrases selected from the 3rd clarification form with the 1st and 4th we aim to investigate Hypothesis 2, which suggests that the application of the measures of phrase stability in the corpus leads to better phrases for query expansion.

Five query expansion runs were conducted. Runs 1, 2, 3 and 4 used the feedback provided by the users to the 1st, 2nd, 3rd and 4th sets of clarification forms accordingly. In each run the query was constructed by splitting the phrases selected by the user from the corresponding clarification form into single terms and adding them to the original query terms. Each term in the expanded query was weighted in Okapi using pseudo-relevance data[8]. The BM25/BM250 search function was then used to search the query against the database. Run 5 was conducted using the developed phrase search algorithm. Here for each topic we take the top 1000 documents retrieved in the run 1 (i.e. using single terms from the user-selected phrases from the 1st clarification form) and re-rank them using the method presented in section 4.

6 Results

The results of the evaluation are presented in table 1. All expanded runs significantly improve the performance over the baseline run BaseTD (t-test at .05 significance level).

Retrieval performance of the expanded queries created from the user feedback to clarification forms 1 and 2 is very similar. This suggests that users tend to select similarly good terms whether they are shown to them in the context of phrases or on their own. Hypothesis 1, formulated in the beginning of the paper, is therefore not supported. On average users selected 21 phrases from the 1st clarification form and 27 single terms from the 2nd form. There were 675 phrase-terms selected only from the 1st form, 384 terms selected only from the 2nd form and 921 terms selected from both forms.

[8] The number of documents used in the blind feedback was used as the number of known relevant documents.

Table 2. Results of the runs, averaged over all topics

Run	Precision at 10 documents	Average Precision
Baseline, Title terms (**BaseT**)	0.3089	0.2196
Baseline, Title + Description (**BaseTD**)	0.42	0.2693
Single-term search, Query expansion with phrases from clarification form 1 (**ExpRun1**)	0.4889	0.3176
Single-term search, Query expansion with terms from clarification form 2 (**ExpRun2**)	0.48	0.3026
Single-term search, Query expansion with phrases from clarification form 3 (**ExpRun3**)	0.4911	0.3191
Single-term search, Query expansion with phrases from clarification form 4 (**ExpRun4**)	0.4689	0.3019
ExpRun1 reranked using the phrase-search algorithm (**ExpRun5**)	0.4422	0.3233

There is also negligible difference between the performance of the queries from phrases selected using the average *idf* of their terms (ExpRun3) and queries from phrases selected using the measures of phrase stability in the corpus: the C-value (ExpRun1) and the Log-Likelihood ratio (ExpRun4). This suggests that the statistical component of phrase selection does not play an important role when it is combined with syntactical phrase selection techniques, such as POS-tagging and NP-chunking. Hypothesis 2 is, therefore, not supported.

Table 3. Precision at various recall levels of the single-term search method (ExpRun1) and the phrase search method (ExpRun5)

Recall level	ExpRun1	ExpRun5
at 0.00	0.6606	0.6259
at 0.10	0.5713	0.5182
at 0.20	0.4852	0.4614
at 0.30	0.4263	0.4316
at 0.40	0.3782	0.3882
at 0.50	0.3392	0.3553
at 0.60	0.2749	0.3027
at 0.70	0.2222	0.25
at 0.80	0.1671	0.188
at 0.90	0.096	0.1241
at 1.00	0.0456	0.0774

The phrase search algorithm (ExpRun5) did not demonstrate improvement in the average precision or precision at 10 documents over the performance of the single-term search method (ExpRun1). While average precision increased slightly (1.8%), precision at 10 documents dropped by 9%. The use of phrases improved average precision in 17 topics and degraded precision in 28 topics. The average gain was 56%, while the average loss was 24%. More interesting results, however, emerge from the analysis of precision at various recall levels (table 3).

At low recall levels, precision of the single-term run is higher, but beginning from 30% recall, the precision of the phrase-based run starts to exceed the precision of the single-term run. These results are consistent with the results evidenced in the earlier studies [13]. The likely explanation of this pattern, suggested in [13], is that in the single-term retrieval documents at high ranks tend to contain a large number of different single terms with high *idf*, therefore the likelihood is high that they cover the topic of the query. However, at lower ranks the number of single term matches is much lower and, therefore, there are more possibilities for topic drift. Phrases usually have much higher weight than single terms, therefore they tend to dominate the document match. At higher ranks this may have a negative effect of over-emphasising a single aspect of the query, whereas at lower ranks phrase-match helps to promote documents with few good matches on phrases and demote documents with matches on single terms which can be peripheral to the query topic.

The results of the phrase-based search experiments partially support Hypothesis 3: precision at high recall levels is better than in the single-term search, whereas precision at low recall levels is inferior.

We performed a detailed analysis of phrase search in one topic (429) “Biodynamic and organic farming”. The user has selected 38 phrases with an average length of 2 words. The single-term search retrieved 31 relevant documents with the average precision of 0.46. Re-ranking the retrieved document set by phrases improved average precision to 0.56. Upon detailed examination of the results it was observed that 17 relevant documents were promoted on average 70 ranks higher in the ranked set, whereas 14 documents were demoted on average 127 ranks lower. The phrase search method tends to rank higher those documents which match few phrases completely and ranks lower the documents which match more phrases, but mostly by one term. The rationale of this approach to ranking is that in the latter case we have less supporting evidence that the matching single term is related to the concept expressed in the query phrase. In some documents, however, this approach fails. For example, one of the relevant documents retrieved for the topic 429 was demoted from rank 53 to 349 because it matched predominantly one term per query phrase. For example the phrase “sustainable development” matched only instances of “sustainable”, which however was used in related context in phrases such as “sustainable growing” and “sustainable production”. Another document, however, was promoted from rank 542 to 62 because it matched many complete phrases either in contiguous positions or separated by a few words.

We are currently experimenting with various parameters of the phrase-search algorithm in order to understand its behaviour better and possibly to obtain better results. One of the parameters is the maximum span for phrase match. In the reported experiments we did not set any span limit. The rationale for this was to capture not only phrasal, but also within-topic relations between terms. So, a document which contains two terms from the same phrase in one paragraph is possibly more likely to

be relevant than a document which contains these terms in different sections. However, this approach may be more useful with long multi-topic documents, rather than short documents. Since HARD track collection consisted mainly of short news articles, this aspect of the phrase search method is unlikely to help distinguish between relevant and non-relevant documents more than single-term match would do. So, setting the span limit to only capture phrasal relations between terms may be sufficient.

7 Conclusions

In this paper we presented a comparative evaluation of different phrase selection techniques in interactive query expansion and a phrase-based document ranking method. A combined syntactico-statistical method was used for the selection of phrases. First, noun phrases were selected using a part-of-speech tagger and a noun-phrase chunker, and secondly, different statistical measures were applied to select phrases for query expansion. Three selection methods were used: C-value, Log-Likelihood ratio and the average *idf* of phrase terms to select phrases, which were then shown to the user for interactive query expansion. Evaluation experiments did not demonstrate substantial difference between these statistical methods in their effect on retrieval performance.

We also studied whether users select better terms when they are shown in the context of phrases, than separately. The users were asked to select query expansion items from two clarification forms: one with the complete phrases selected by the C-value, and the other with the single terms from these phrases. The two query expansion runs gave very similar results, which suggests that presenting terms in the context of phrases does not provide more help to the users in selecting good query expansion terms.

The phrase-based document ranking method demonstrated high precision gains at higher recall levels and losses in precision at lower recall levels as compared to single-term document ranking. We are currently working on improving our phrase-weighting formulae. As discussed earlier in the paper, phrases differ by their stability in the corpus, therefore they should not be treated uniformly in search. For example, a document which has a partial match on a non-compositional or idiomatic phrase (e.g. "Salt Lake City") is more likely to be non-relevant, than a document that has a partial match on a non-idiomatic expression (e.g. "organic product"). Therefore the weight of the partially matching phrase should be reduced more in the first case than in the second. One of the extensions of this work will be to use measures of phrase stability to estimate phrase weight in the documents.

References

1. Robertson S.E., Spärck Jones K. Relevance Weighting of Search Terms. Journal of the American Society for Information Science, 27 (1976) 129–146
2. Spärck Jones, K., Walker, S. and Robertson, S.E. A probabilistic model of information retrieval: development and comparative experiments. Information Processing and Management, Vol. 36, n. 6, (2000) 779–808 (Part 1); 809–840 (Part 2)

3. Salton, G; Wong, A.; Yang, C. S. A vector space model for information retrieval. Communications of the ACM, Vol. 18, n.11 (1975) 613–620
4. Voorhees E. and Buckland, L. (Eds.) Proceedings of the Twelfth Text Retrieval Conference, NIST, Gaithersburg, MD, 2004
5. Xu J. and Croft B. Query expansion using local and global document analysis. In Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR '96), Zurich, Switzerland (1996) 4–11
6. Frantzi, K.T. and Ananiadou, S. Extracting nested collocations. In Proceedings of the 16th Conference on Computational Linguistics, COLING (1996) 41–46
7. Dunning, T. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, Vol. 19, n. 1 (1993) 61–74
8. Bely, N., Borillo, A., Virbel, J. and Siot-Decauville, N. Procédures d'analyse sémantique appliquée à la documentation scientifique. Paris: Gauthier (1970)
9. Fagan J.L. Automatic Phrase Indexing For Document Retrieval: An Examination Of Syntactic And Non-Syntactic Methods. In Proceedings of the Tenth ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans (1987) 91–108
10. Fagan J.L., The effectiveness of a nonsyntatic approach to automatic phrase indexing for document retrieval, Journal of the American Society for Information Science, Vol. 40 n. 2 (1989) 115–132
11. Salton, G. and Lesk M.E. Computer Evaluation of Indexing and Text Processing Journal of the ACM (JACM), Vol. 15 , n. 1 (1968) 8–36
12. Strzalkowski T. Perez-Carballo J. Evaluating natural language processing techniques in information retrieval. In: Strzalkowski, T. (Ed.) Natural language information retrieval. Kluwer Academic Publishers (1999) 113–145
13. Mitra, M., Buckley, C., Singhal, A. and Cardie, C. An Analysis of Statistical And Syntactic Phrases. In Proceedings of RIAO97, Computer-Assisted Information Searching on the Internet, Montreal, Canada (1997) 200–214
14. Robertson, S.E., Zaragoza, H. and Taylor, M. Microsoft Cambridge at TREC-12: HARD track. In Proceedings of the Twelfth Text Retrieval Conference, Voorhees, E. and Buckland, L., (Eds.), NIST, Gaithersburg, MD, (2004) 418–425
15. Marchionini, G. Interfaces for End-User Information Seeking. Journal of the ASIS Vol. 43, n. 2 (1992) 156–163
16. Smeaton, A. F. and Kelledy, F. User-Chosen Phrases in Interactive Query Formulation for Information Retrieval, In Proceedings of the 20th BCS-IRSG Colloquium, Grenoble, France, Springer-Verlag Workshops in Computing (1998)
17. Vechtomova, O., Karamuftuoglu, M., Lam, E. Interactive Search Refinement Techniques for HARD Tasks. In Proceedings of the Twelfth Text Retrieval Conference, Voorhees, E. and Buckland, L., (Eds.), NIST, Gaithersburg, MD, (2004) 820–827
18. Brill, E. Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging. Computational Linguistics, Vol. 21, n. 4 (1995) 543–565
19. Ramshaw, L. and Marcus, M. Text Chunking Using Transformation-Based Learning. In Proceedings of the Third ACL Workshop on Very Large Corpora, MIT (1995)
20. Manning, C.D., Schütze, H. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts (1999)
21. Banerjee, S. and Pedersen, T. The Design, Implementation and Use of the Ngram Statistics Package. In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City (2003)

22. Clarke, C.L.A. and Cormack, G.V. On the use of Regular Expressions for Searching Text. University of Waterloo Computer Science Department Technical Report number CS-95-07, University of Waterloo, Canada (1995)
23. Allan, J. HARD Track Overview in TREC 2003 High Accuracy Retrieval from Documents. In Proceedings of the Twelfth Text Retrieval Conference, Voorhees, E. and Buckland, L., (Eds.), NIST, Gaithersburg, MD, (2004) 24-37
24. Beaulieu, M. and Jones S. Interactive searching and interface issues in the Okapi best match probabilistic retrieval system. Interacting with Computers. Vol. 10, n. 3 (1998) 237–248
25. Ruthven I. Re-examining the potential effectiveness of interactive query expansion. Proceedings of the 26th ACM-SIGIR conference, Toronto, Canada (2003) 213–220.
26. Vintar Š. (2004) Comparative Evaluation of C-Value in the Treatment of Nested Terms. In Proceedings of MEMURA 2004 Workshop (Methodologies and Evaluation of Multiword Units in Real-world Applications), Language Resources and Evaluation Conference (LREC), Lisbon, Portugal (2004) 54-57.
27. Vechtomova O., Karamuftuoglu M., Approaches to High Accuracy Document Retrieval in HARD Track. To appear in Proceedings of the Thirteenth Text Retrieval Conference, Voorhees, E. and Buckland, L., (Eds.), NIST, Gaithersburg, MD (2005).