Query expansion with long-span collocates

Olga Vechtomova[a] Stephen Robertson[bc] Susan Jones[b]

[a]Department of Management Sciences
University of Waterloo, Canada
ovechtom@engmail.uwaterloo.ca[1]
[b] Centre for Interactive Systems Research
City University, London, UK
{ser, sa386}@soi.city.ac.uk
[c] Microsoft Research Cambridge, UK
ser@microsoft.com

Abstract

The paper presents two novel approaches to query expansion with long-span collocates – words, significantly co-occurring in topic-size windows with query terms. In the first approach – global collocation analysis – collocates of query terms are extracted from the entire collection, in the second – local collocation analysis – from a subset of retrieved documents. The significance of association between collocates was estimated using modified Mutual Information and Z score. The techniques were tested using the Okapi IR system. The effect of different parameters on performance was evaluated: window size, number of expansion terms, measures of collocation significance and types of expansion terms. We present performance results of these techniques and provide comparison with related approaches.

1 Introduction

1.1 Definition of collocation

Words in natural language texts have individual patterns of co-occurrence with each other: every word has a tendency of more frequent occurrence near some words, and less frequent – near others. Words which co-occur near each other with more than random probability are known as *collocates*. There are two major groups of factors that cause word collocation:

- Lexical-grammatical and habitual;
- Lexical-semantic.

Lexical-grammatical or habitual restrictions limit the choice of words that can be used in the same grammatical structures with the word in question. These factors operate within short distances, i.e. within the same lexical or grammatical constructions. The examples of short-span collocations motivated by lexical-grammatical factors are terminological expressions (*fission reactor fuel*), light verbs (*make a decision, do a favour*) and phrasal verbs (*check in, cut down*). Short-span collocation can also be caused by referential association of the words' meanings (*deep sea, bright day*) and by habitual or customary patterns which evolved in the language (*rancid butter, sour milk*).

[1] This work was done while O.Vechtomova was in the Centre for Interactive Systems Research, City University, London.

Lexical-semantic associations exist between some words which are used to describe the same topic. These relations are not confined to the same lexical or grammatical structure, but can span over long distances in text. Words which are related semantically and belong to the same semantic domain tend to be used together to describe the same topics, thus creating the phenomenon of long-span collocation.

Researchers understand the phenomenon of collocation differently. Some only recognise short-span collocations. Palmer (1981), for example, understands by collocation adjacent word combinations resulting from referential associations of their meanings or habitual patterns of word use. Manning and Schuetze (1999) define collocation as grammatically bound elements occurring in a certain order which are characterised by limited compositionality. They admit the existence of word associations across larger expanses of text, but they suggest calling such associations 'co-occurrences' and to reserve the term 'collocation' only for grammatically bound combinations.

Other researchers recognise the existence of long-span collocation motivated by lexical-semantic relations. Halliday and Hasan (1976) point out that collocation is a realisation of lexical cohesion in text. They argue that words co-occur because they are in some kind of lexical-semantic relation. Hoey (1991) also emphasises the role of long-distance relations in text in creating the phenomenon of collocation.

In the light of the approaches mentioned above, we recognise two major types of collocation:

1. Short-span collocation (due to lexical-grammatical or habitual restrictions);
2. Long-span collocation (due to the existence of certain lexical-semantic relations between words).

Both lexical-grammatical/habitual restrictions and lexical-semantic relations are the linguistic factors that cause the phenomenon of word collocation. The nature of these linguistic factors is quite complex and the development of a general method to identify collocations in text through the analysis of these factors is not an easy task. Instead it is possible to isolate the phenomenon of collocation in text empirically, not by looking for what causes it, but by identifying what characterises it most typically - namely, more than random probability of its occurrence.

There have been developed a number of empirical approaches to identification of collocates from text within the branch of linguistics based on the analysis of empirical data – *corpus linguistics*. Such approaches are based on the statistical analysis of frequency characteristics of words. Two common measures of estimating the degree of association between collocates are Mutual Information and Z score. In our experiments we used modified versions of these statistics, which will be described later in this paper.

In our research we focused on long-span collocates and their use as query expansion terms. We hypothesised that by expanding the original user's query with the terms which tend to be used in the same topics as the query terms, and, hence, holding certain semantic relations with them, we may enhance our chances of matching the query against relevant topics in the documents.

1.2 Use of collocation in IR

There have been a wide range of approaches towards using word collocation or co-occurrence in Information Retrieval. The types of collocation they address are:

- short-span collocation, used mainly for phrase identification;
- document-wide co-occurrences of terms;
- long-span collocation, used to obtain some context information.

Statistically defined short-span collocations are commonly used by *non-linguistic indexing* (NLI) approaches to generate composite indexing units – joined terms or statistical phrases, which are statistical surrogates of the genuine linguistic phrases (e.g. Fagan 1989, Croft 1991).

Document-level term co-occurrence has been rather intensively explored over the last 30 years. The earliest co-occurrence research in IR dates back to the work done by Sparck Jones (1970), who studied the use of document-wide co-occurrence in automatic index term classification. Van Rijsbergen (1977) and Harper (1978) and later Smeaton (1983) experimented with identifying significantly associated document-level co-occurrences using Expected Mutual Information (EMIM) and using them in building dependence trees – maximum spanning trees (MST). Closely related terms from the MST were then used for query expansion.

The general motivation behind research on document-wide co-occurrence was to understand the effect that information about the presence of more than one term in a document can have on retrieval performance. Approaches to document-wide co-occurrence consider only presence or absence of two or more terms in documents. With the arrival of full-text collections containing long multi-topic documents, considering document-level term dependencies no longer seems adequate. Instead, exploitation of term dependencies within more homogeneous subdocument semantic units – topics – may lead to improvements in retrieval performance. Approaches using long-span collocation focus on studying term dependencies within *limited* spans of text, and attempt to capture statistical evidence of relations pertaining to a topic in a document.

1.3 Global vs. Local analysis

In IR there can be distinguished two approaches to analysing contextual environments around query terms: analysis of contexts around every occurrence of the query term in the collection – *global analysis methods*, and analysis of contexts of all query term occurrences in a subset of documents for which some relevance information (known or assumed) is available – *local analysis methods*.

We studied both global and local approaches to query expansion, and in this paper we will describe two query expansion techniques which we developed: *global collocation analysis* and *local collocation analysis*.

A number of research groups developed somewhat related techniques. Qiu and Frei (1993) developed a global query expansion method where query expansion terms are selected from an automatically constructed co-occurrence based term-term similarity thesaurus on the basis of the degree of their similarity to all terms in the query. *PhraseFinder* technique developed by Jing and Croft (1994) is used for automatic construction of a co-occurrence thesaurus. Each indexing unit, defined through a set of phrase rules, is recorded in the thesaurus with a list of its most strongly associated collocates. They define collocates as index units co-occurring in windows of 3-10 sentences, which approximate the size of an average paragraph.

Among local techniques the most widely known is Xu and Croft's (1996) *Local Context Analysis* (LCA). LCA is a type of local feedback: collocates of query terms, defined as noun groups and taken

from the retrieved *N* top ranked passages of fixed size of 300 words, are ranked by the significance of their association with all query terms using a variant *tf*idf* measure, and used for query expansion. Over time LCA has shown rather consistent performance improvements over the baseline (INQUERY) on TREC collections. Some other approaches, focusing on the use of delimited document parts (e.g., best passages, windows) following local feedback, are those by Buckley (2000), Cormack (2000), Hawking (1998), Ishikawa (1998) and Strzalkowski (2000).

In the following sections we will describe the global and local collocation analysis techniques that we developed and query expansion experiments using these techniques.

2 Selection of collocates

2.1 Windowing technique

We define collocates of a single instance of the term in question as all words that occur within a fixed-length *window* surrounding this term. Each window is centred around a *node* term. Since in our methods we identify collocates of query terms, a window is defined for each instance of each query term in a set of relevant documents (local analysis) or in the entire collection (global analysis).

Prior to applying the windowing technique, we stem all the words in the document using Porter stemming algorithm and remove stopwords. A window is defined as a fixed number of words to the left and right of the node. Ideally left and right sides of the window are of equal lengths, but, as will be described later, in practice it is not always the case. The choice of defining windows by counting the number of words, instead of using natural language constructs like sentences and paragraphs was made, first, because the lengths of the latter are highly irregular, which will complicate the estimation of association strength between collocates. The second reason is that NL structures require more computational effort to delimit in text. For some corpus linguistics applications the order in which collocates occur in text is important, for example in machine translation, speech recognition, optical character recognition. Order-sensitive collocation statistics in such applications is necessary for making correct lexical choices in short-span lexical-syntactic structures (Church 1990, 1991, 1994). In our research the order in which collocates occur together is unimportant, since we are interested in semantically related collocates that can occur anywhere within the large area surrounding the node. In other words, collocates of a term instance are all words that have either backward co-occurrence with it, i.e. occur within *S* word span to its left, or forward co-occurrence, i.e. occur within *S* word span to the right (figure 1).

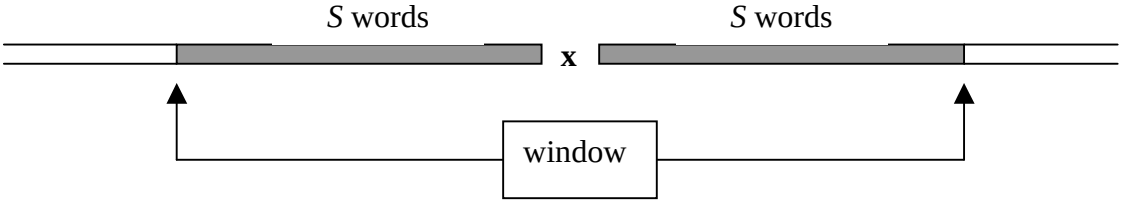

Figure 1. Window around node *x*, defined as spans of *S* words to the left/right of *x*

However, for two reasons, the windows actually used are often smaller than suggested by this distance. A window around term *x* may be truncated if either (a) it hits a document boundary (figure 2), or (b) it hits another occurrence of term *x* (figures 3 and 4). The latter truncation of the window is necessary to avoid duplicate extraction of the same word as a collocate of two instances of *x* when they occur near each other, i.e. when the distance between them is less than *S* words. If another instance of *x* is found after the node *x*, we truncate the window at this point (figure 3), if another *x* is found before the node *x* we ignore the left-hand half of the window altogether (figure 4).

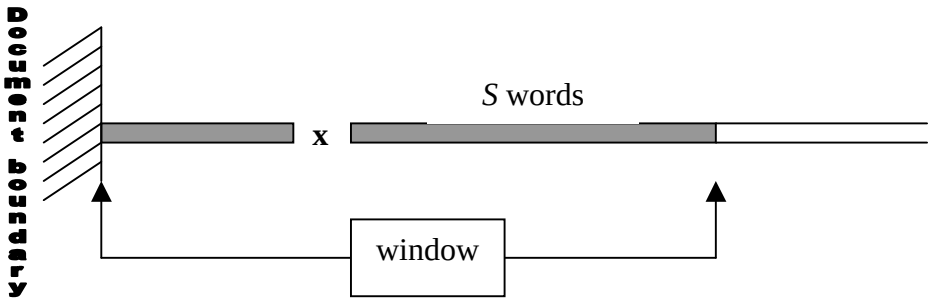

Figure 2. Window truncated by hitting the document boundary

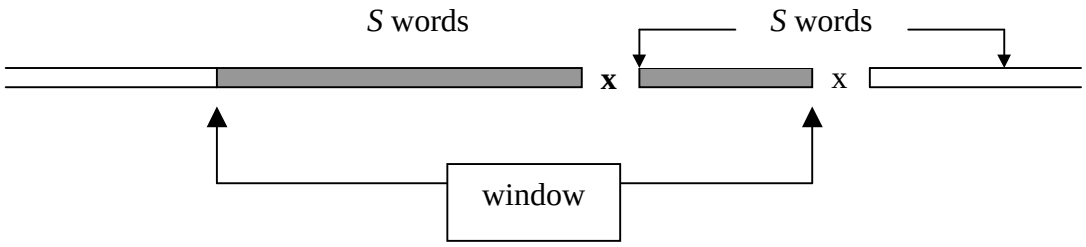

Figure 3. Right-hand half of the window truncated by hitting another occurrence of x after the node

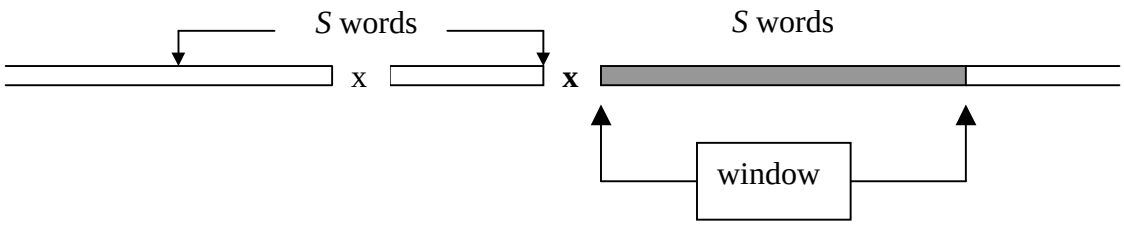

Figure 4. Left-hand half of the window ignored when another occurrence of *x* is found before the node

A decision must be made in the windowing technique regarding the size of the span (*S)* to the left/right of the node. The span can be measured either in syntactic units, such as phrases, sentences, paragraphs or even entire texts (Harper 1978), or by the number of words to the left and right of the

node, for example, 4 words (Sinclair 1974), 5 words (Church 1990), 400 words (Beeferman 1997), 4, 10, 50 words (Edmonds 1997). The choice of the span size is usually determined by which syntactical or semantic constructs of the text are under analysis, e.g. phrases, sentences, paragraphs, topics. Since we are interested in topical relations between words, the span size must be of the scale of a topic in text. A topic is a rather nebulous entity, which often cannot be indisputably delimited even by humans. Moreover it is not necessarily present in the form of an uninterrupted stretch of text, but can re-surface throughout the text, being interwoven with other topics. There have been proposed more complex approaches for topic detection by using, for example, lexical chains (Hearst 1994). Such techniques can be rather computationally demanding, and hence may not always be suitable for search-time use. We prefer to use a more crude and fast technique of collocate extraction from fixed-length windows, which is complemented by the second stage – selection of significant collocates via statistical measures.

The initial span chosen for the global collocation analysis experiments was 100 words. This decision was motivated by the research of Beeferman et al. (1997), who established that a word's influence on its environment stretches as far as several hundred words, levelling off at 400 words. The factors affecting this distribution at such big distances from the node cannot be lexical-syntactical, but semantic and topical. In our experiments on local collocation analysis other smaller span sizes were tried: 50, 25, 15 and 10. We did not experiment with span sizes smaller than these for the reason that in short-span environments lexical-syntactic factors dominate word relations, whereas we were specifically interested in topical relations. The decision of not using span sizes larger than 100 was made, first, because the monotonic decay of the word distribution curve in Beeferman's experiments suggests that the influence of the node weakens with the distance, therefore increasing the chance of noise terms. Secondly, the spans used would still yield a sufficient number of terms needed for query expansion.

Since an ideal window is symmetrical, its size is $S + S$, where S is the span size. However as the observed window sizes around instances of a given term in a document/corpus are variable (figures 2, 3 and 4), we calculate the average window size – v_x – around the term x. Average window sizes, as will be described in the next section, are needed to calculate collocation significance scores – MI and Z. Also, as will be described later, we defined two variants of Z and MI statistics: for the global collocation analysis, where collocates are found for every occurrence of x in the collection, and for the local collocation analysis, where collocates are identified for the instances of x in relevant documents.

In the global method, v_x is estimated by summing the observed window sizes around all instances of x in the corpus and dividing them by the frequency of occurrence of x in the corpus $f(x)$:

$$v_x = \frac{\sum_{i=1}^{f(x)} W_i}{f(x)} \qquad (1)$$

where W_i is the observed window around ith instance of x in the corpus;
$f(x)$ is the frequency of x in the corpus.

In the local method, v_x is estimated similarly by dividing the sum of observed windows of x in relevant documents by $f_r(x)$ – the frequency of x in relevant documents.

2.2 Selection of significant collocates

There exist several statistical measures for collocation selection, e.g., frequency counts, likelihood ratio, chi-square test, mutual information and Z score (Manning 1999). Comparison of the effectiveness of various statistics in collocation selection was not the aim of our research. Our main focus was to explore whether long-span collocation can be used effectively in query expansion. We have chosen two statistics which are most commonly used in corpus linguistics (McEnery 1996) – mutual information (MI) and Z score.

Mutual information originated in the field of information theory (Fano 1961), and since then has been used extensively in a wide variety of applications, e.g., speech recognition (Jelinek 1990), information retrieval (Van Rijsbergen 1977) and various uses of corpus linguistics like lexicography, lexical analysis (Church 1991, 1994), word sense disambiguation (Yarowsky 1992), and analysis of aligned corpora (McEnery 1996).

Z score is a statistic for hypothesis testing, i.e. for assessing whether a certain event is due to chance or not. When used for collocation selection, Z score tests whether the co-occurrence of two words is due to other factors than chance. It is very similar to a *t* score, the difference lying in the fact that Z is used with the data distributed normally. As will be described in more detail later, the large size of the corpus used in this project warrants normal distribution, hence *Z* score instead of *t*.

Church et al. (1991, 1994) used MI and *t* score in their study of synonymy and lexical substitutability. Specifically they were interested in the possibilities of identifying differences between near-synonyms from the patterns of their use in text, i.e. from the regularities of their co-occurrence with other words. They argued that MI is a better tool for finding associations between words, while *t* is good for identifying dissimilarities in the use of near synonyms.

Church et al. (1994) pointed that mutual information and *t* score tend to bring to the top different kinds of collocations: *t* score tends to pick high frequency word combinations, and may have a drawback of showing syntactical collocations with functional words, while mutual information highlights less frequent word combinations that are specific to both words, e.g. fixed phrases, some compound terms and proper names. The drawback of MI is that it can reward very low-frequency corpus-specific collocates, that are not easily generalised across corpora.

Our experiments also showed that there is usually a marked difference between the type of words selected by mutual information and Z score (see examples in section 3.1.1). Our query expansion experiments, however, did not demonstrate that one of these statistics is better suited for the query expansion task than the other.

The following two subsections will describe MI and Z measures and our modified formulas.

2.2.1 Mutual Information

The mutual information score between a pair of words or any other linguistic units "compares the probability that the two words are used as a joint event with the probability that they occur

individually and that their co-occurrences are simply a result of chance" (McEnery 1996, p.71). The mutual information score grows with the increase in frequency of word co-occurrence. If two words co-occur mainly due to chance their mutual information score will be close to zero. If they occur predominantly individually, then mutual information will be a negative number.

The standard formula for calculating mutual information score is:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \qquad (2)$$

where $P(x,y)$ is the probability that words x and y occur together;
$P(x)$ and $P(y)$ are the probabilities that x and y occur individually.

MI statistic is usually applied where term x immediately follows term y in text, e.g. as used in (Church 1991, 1994). Church et al. estimate the probability that two words occur as a joint event $P(x,y)$ as $f(x,y)/N$, where joint frequency – $f(x,y)$ denotes the number of times that y appears immediately after x. Our interpretation of $f(x,y)$ is different – as the frequency with which y occurs either sides of x within the maximum distance of S words (where S is the span size). Therefore the standard MI formula was modified to provide for unordered co-occurrence within a distance more than one word. But the most important difference from the standard MI is the asymmetry of our approach. Standard MI is a symmetrical measure, i.e. $I(x,y) = I(y,x)$ as joint probabilities are also symmetrical: $P(x,y) = P(y,x)$. The asymmetry of our approach arises due to the use of average window sizes. As described in the previous section the actual window sizes around instances of a term x are often smaller than the ideal window size of $(S + S)$. For this reason we use the average of all windows around term $x – v_x$ to estimate the probability of occurrence of y in the windows around $x – P_v(x,y)$. However, if we were to start with y and consider the occurrences of x in the windows around y, we would replace v_x in the formula with v_y. In general these two are different.

The modified MI formula for the global method is:

$$I_v(x,y) = \log_2 \frac{P_v(x,y)}{P(x)P(y)} = \log_2 \frac{\frac{f(x,y)}{Nv_x}}{\frac{f(x)f(y)}{N^2}} \qquad (3)$$

where $f(x,y)$ – joint frequency of x and y in the corpus;
$f(x)$ and $f(y)$ – frequencies of independent occurrence of x and y in the corpus;
v_x – average window size around x in the corpus;
N – corpus size.

The modified MI formula for the local method is:

$$Local\ I_v(x,y) = \log_2 \frac{\frac{f_r(x,y)}{Rv_x(R)}}{\frac{f_r(x)}{R}\frac{f_c(y)}{N}} \qquad (4)$$

where $f_r(x,y)$ – joint frequency of x and y in the relevant documents;
$f_c(y)$ – frequency of y in the corpus;
$f_r(x)$ – frequency of x in the relevant documents;
$v_x(R)$ – average window size around x in the relevant documents;
N – corpus size;
R – size of the relevant set (in tokens).

The use of corpus information for y in the locally-defined version of MI (and of Z, below) deserves some comment. If a term y occurs significantly more frequently in the relevant documents than in the collection, our local formula might select it, even if (looked at purely from the point of view of relevant documents) its window-based collocation with x is no more than random. In other words, the local collocation measurement includes a global document-level component. It might be argued that a pure collocation-based method should exclude such a document-level component; this would suggest using local frequency information about y in the formula. Our choice was to go the other way.

While mutual information is useful in filtering out pairs of words whose joint probability of occurrence is greater than chance, it gives very limited information as to how far joint probability differs from chance. Very high mutual information scores generally indicate strong bond between two words, whereas lower scores can be misleading, especially with low frequencies. Therefore it is not safe to make assumptions about the strength of words' association without knowing how much of that association is due to chance.

2.2.2 Z score

Z score is a more reliable statistic: it gives us an indication with varying degrees of confidence that an association is genuine by measuring the distance in standard deviations between the observed frequency of occurrence of y around x, and its expected frequency of occurrence given the null hypothesis. For a chance pair of words in the conditions of low word frequencies we may misleadingly get a high mutual information score, whereas their Z score will not be high since the variances of probabilities will be large.

Our approach to measuring the significance of collocations with Z score is somewhat similar to Church's et al. use of a *t* statistic (Church 1991, 1994). However, there are three main differences:

(a) We are interested in collocations within a substantial window around the starting node;
(b) The argument on which the measure is based is asymmetric: it considers, given a word x, the probability that word y will occur within the window. (The resulting formula is also asymmetric, for the same reasons as those leading to the asymmetry of the MI formula, discussed above);
(c) Because we are dealing with collocations over a large corpus, the small-sample characteristics which lead to the choice of the *t* statistic do not apply – we use the *Z* statistic instead.

We take as null hypothesis that the presence of x does not predict the presence or absence of y in the windows – that any location in these windows is exactly as likely to contain y as any other location in the corpus.

In the global method the total number of locations which might contain term y collocated with x is v_x $f(x)$. Under the null hypothesis, the probability that any given one of these locations contains y is $f(y)/N$. Thus the expected number of occurrences of y in these locations is the mean of a binomial distribution, v_x $f(x)$ $f(y)/N$. Also, because the probability $f(y)/N$ is in general very small, the mean

square error of this expected value (the variance of the binomial distribution) is approximately also $v_x\, f(x)\, f(y)/N$.

But we actually observe $f(x,y)$ occurrences of y within these windows around x. Therefore we can calculate a normal deviate (Z score) as

$$Z = \frac{f(x,y) - \frac{v_x\, f(x) f(y)}{N}}{\sqrt{\frac{v_x\, f(x) f(y)}{N}}} \qquad (5)$$

This score can be compared with normal distribution tables in the usual way.

Under the null hypothesis as formulated above, for small samples this could be interpreted as a t score with $v_x\, f(x)$-1 degrees of freedom. In our case this will always be large enough to warrant the normal approximation. However, it does flag up one problem: suppose $f(x)$ were only one. It would then appear that we artificially inflated our sample by considering a window of size 100 words (say) either side, when the locations we are considering all relate to a single instance of x. Our response to this problem is simply to avoid using the method on terms with very small frequencies ($f(x) < 30$).

Church and Hanks use a different estimate for the variance, involving the co-occurrence frequency $f(x,y)$. Our asymmetric argument and explicit formulation of the null hypothesis suggest the variance based on the individual frequencies.

For the local method, we modified the above global Z function as:

$$Local\ Z = \frac{f_r(x,y) - \frac{f_c(y)}{N} f_r(x) v_x(R)}{\sqrt{\frac{f_c(y)}{N} f_r(x) v_x(R)}} \qquad (6)$$

where $f_r(x,y)$ – joint frequency of x and y in the relevant documents;
$f_c(y)$ – frequency of y in the corpus;
$f_r(x)$ – frequency of x in the relevant documents;
$v_x(R)$ – average window size around x in the relevant documents;
N – corpus size.

The reason for using corpus information for y is the same as in the local MI formula (section 2.2.1).

According to Church et al. (1991) the threshold of significance of association between two collocates measured by t score should be no less than 1.65 standard deviations. In our experiments we adopted the same threshold for filtering out insignificant associations in both global and local analyses.

3 Query expansion experiments

The experimental platform used for the research described in this paper is *Okapi* – an experimental IR system, which implements the Robertson & Sparck Jones probabilistic model in term weighting, document ranking and relevance feedback mechanisms (Robertson 1976, Sparck Jones 2000). Okapi BM25 function was used for all weighted searches.

All query expansion experiments were run on FT 96 collection from TREC Disk 4 and 50 TREC topics (251-300). Porter stemming algorithm and a stopword list with some high frequency terms were used for both indexing the database and processing queries. We used short queries created from the contents of title fields of the topics. The motivation for using short queries is twofold: first, they correspond to the type of briefly formulated queries frequently submitted by real users in practice, secondly, they are good candidates for query expansion, as there is more scope for expansion, than with long queries.

3.1 Global collocation analysis

The aim of global collocation analysis is to identify all terms that co-occur significantly in the same topics as query terms throughout the collection and use them in query expansion. The idea was not to identify exclusively terms from relevant topics, but terms from a diverse range of contexts, which would represent different aspects of a query term.

Query expansion with global collocates requires a pre-processing stage, during which we build a collocation resource – a database of global collocates of query terms. There is one record in the database for each term in the 50 queries we used for the experiments. In each record the query term is associated with a list of its significant global collocates.

3.1.1 Methodology

The construction of a collocation database consists of the following stages: first, collocates around all instances of each query term in the collection are extracted using the windowing technique described above. Secondly, collocates of each query term are ranked by the significance of their association using either Z or MI score. Then, top *N* ranked collocates of each query term are written into the record of the query term in question. We decided to use a fixed number of top-ranked collocates in MI- and Z-ranked lists for query expansion, instead of selecting collocates with scores above a certain threshold. The problem with the latter approach is that both MI and Z scores are highly variable for collocates of different query terms, what would result in a large number of collocates selected for some terms, and none for others. For this reason we experimented with selecting a fixed number (8 and 16) of top ranked MI and Z collocates. A very large margin of significant collocates in both MI- and Z-ranked lists will still exceed these fixed numbers of top collocates. In other words, all collocates we choose either from Z-, or MI-ranked lists will have significant values of the corresponding statistic.

Examples of top ranked MI and Z collocates are given in tables 1 and 2. As demonstrated in table 1, for general terms MI usually tends to reward very rare terms, most of them – proper names, while Z top-ranks less rare terms of the type one would expect to find in manually constructed thesauri. With

more specific query terms (table 2), the divergence between the two statistics is less obvious. For a more detailed lexical-semantic analysis of MI and Z collocates see (Vechtomova 2000).

MI	Z
Ofsted (The Office for Standards in Education) 5.00	**school** 351.41
GNVQ (General National Vocational Qualifications) 4.98	**teacher** 220.43
Blatch (Lady Blatch, schools minister) 4.91	**student** 169.62
Natfhe (The University and College Lecturers' Union) 4.85	**pupil** 155.44
Gruchy (Nigel de Gruchy, general secretary of the National Association of Schoolmasters/Union of Women Teachers) 4.79	**curriculum** 155.22
GCSE 4.79	**university** 147.05
educationalist/educationally 4.78	**college** 144.69
truancy 4.76	**A-level** 141.42
A-level 4.72	**vocation** 130.10
Nasuwt (National Association of Schoolmasters/Union of Women Teachers) 4.71	**Patten** (Christopher Patten, chairman, Conservative party) 120.88

Table 1. Lists of top collocates for the term *education* sorted by MI and Z statistics

MI	Z
NOx* (nitrogen oxides) 9.26	**oxide*** 200.95
monoxide* 9.24	**emission** 181.22
particulate* 9.20	**monoxide*** 168.74
superconductor* 9.18	**particulate*** 162.55
legume 8.68	**dioxide** 154.99
gasification 8.65	**NOx*** 154.90
urea 8.56	**sulphur** 152.57
ammonia* 8.24	**superconductor*** 115.37
autocatalyst 8.17	**carbon** 111.17
superconduct(ing/ivity) 8.11	**diesel** 102.24
soot 8.05	**pollution** 102.00
oxide* 7.99	**fertiliser** 94.18
co-generation 7.90	**hydrocarbon** 89.79
phosphor 7.90	**pollute** 89.76
nitrate* 7.85	**nitrate*** 84.35
lolly 7.81	**ammonia*** 81.37

Table 2. Lists of top collocates for the term *nitrogen* sorted by MI and Z statistics
(* terms are top ranked in both lists)

For our experiments we constructed four databases, different by the type and number of collocates associated with each query term:

- top 8 Z-ranked collocates;
- top 8 MI-ranked collocates;
- top 16 Z-ranked collocates;
- top 16 MI-ranked collocates.

A collocation database serves as an intermediate layer in the existing searching technique. When the original query is submitted, it is searched first against one of the collocation databases. Each query term matches a single record. The contents of this record – the query term's collocates – are added to the original query (figure 5).

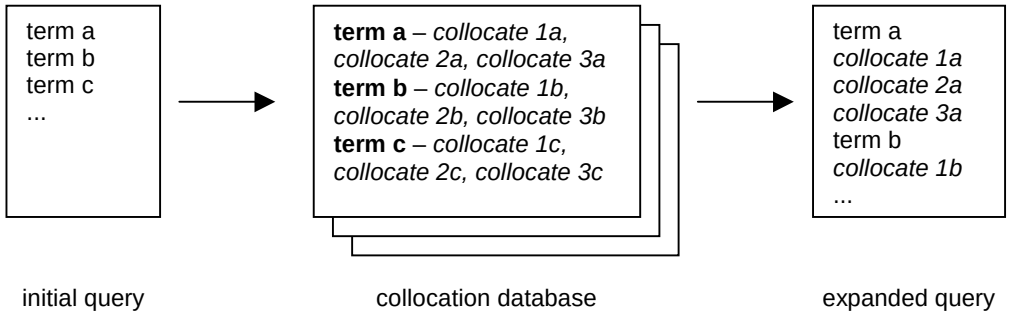

Figure 5. Query expansion with global collocates

The collocation database was searched using a simple unweighted search, since there is only one record corresponding to each query term. The expanded query was searched against FT 96 using Okapi weighted function BM25.

3.1.2 Results

All query expansion runs with global collocates showed worse performance than the run with unexpanded queries (table 3).

	no expansion	top 8 MI collocates	top 8 Z collocates	Top 16 MI collocates	top 16 Z collocates
Retrieved	42686	43534	44000	44000	44000
Relevant	1583	1583	1583	1583	1583
Relevant retrieved at 1000	632	573	520	526	504
Average precision (non-interpolated) for all rel docs (averaged over queries)	0.1310	0.0432	0.0375	0.0344	0.0340

Table 3. Summary of retrieval results for query expansion with global collocates

The analysis of runs top 8 MI and top 8 Z by query showed that top 8 Z improved 7 queries, did not affect 3 and hurt 34 queries; top 8 MI improved 3, did not affect 2 and hurt 39 queries.

The rather disappointing performance of global collocates as query expansion terms led us to the conclusion that information gathered in the form of single words occurring in the environments of *all* instances of a query term in the corpus, does not have substantial relevance-discriminating power, even though the terms are ranked by significance of their co-occurrence with the query terms. The main reason for this is thought to be the fact that many query terms are words from the general lexicon, that can occur in a broad range of contexts. Even occurrences of the same sense of a word can be used in a wide range of topics. As the number of relevant documents in which a query term occurs is usually much smaller than the number of non-relevant documents with this term, only a small proportion of collocates come from contexts which have any relatedness to the topic of user's interest. It was therefore considered that using collocates from the contexts of query terms in documents for which there exists some evidence of relevance to the user's need, might result in a better performance. This led us to the development of another method of query expansion with query terms' collocates following local feedback – *local collocation analysis*, which will be described in the next section.

3.2 Local collocation analysis

Local collocation analysis is a form of local (relevance or pseudo-relevance) feedback, whereby significant collocates of all query term occurrences in a subset of retrieved documents (i.e. documents judged or assumed as relevant) are selected for query expansion. Our evaluation of query expansion with local collocates was conducted using relevance feedback. We simulated user's relevance judgements by using first 5 relevant documents taken from the top 1000 documents retrieved by bm2500. Information about the relevance of the documents was taken from TREC relevance judgements file. The reason for not using pseudo-relevance (blind) feedback for evaluation is due to the fact that relevance feedback generally gives better results than pseudo-relevance feedback. If local collocation analysis showed significant improvements over the existing relevance feedback technique, it would be evaluated with blind feedback as well.

3.2.1 Methodology

The query expansion process consists of the following steps:

1. Collocates are extracted from the fixed size windows around all instances of each query term in the subset of retrieved relevant documents.
2. Collocates of each query term are ranked by either MI, or Z score.
3. Top *N* ranked collocates of each query term are added to the original query. Duplicate terms are removed.
4. The expanded query is run against the text collection using Okapi BM25 function.

One of the questions associated with query expansion following relevance feedback is whether original query terms should be given extra weight. We decided not to give extra weight to the original query terms. Our decision was motivated by the findings of earlier experiments with the probabilistic retrieval model, which showed that artificially increased weights of the original query terms did not improve performance (Sparck Jones 2000). It is however unclear what effect increased query term weights will have on query expansion with collocates, and further experimentation in this area is needed.

We evaluated local collocation analysis with a range of values for the following key parameters:

- Window size (200, 100, 50, 30, 20);
- Measure of collocation significance for ranking collocates (local variants of MI and Z statistics);
- Number of top ranked collocates in the expanded query (8, 12, 16, 21);
- Number of Okapi relevance feedback terms in the expanded query (20 and 10).

A large number of combinations of different values for the above variables is possible. We did not replicate runs for every possible combination of variable values. Instead a more selective approach has been adopted: those runs which showed best results with a certain value for one variable, were replicated with a range of values for other variables.

Some of our preliminary experiments on a single topic suggested that it might improve performance, if we merge in the expanded query two types of terms: local collocates and terms selected using existing Okapi relevance feedback (RF) technique. Okapi relevance feedback method selects terms from the known relevant documents on the basis of their probability of occurrence in relevant documents. It is an entirely different method of term selection, and as we ascertained later from our experiments, a reasonable proportion of terms, selected by our method and Okapi relevance feedback, do not overlap. Therefore, Okapi RF terms and collocates may retrieve different sets of relevant documents, and hence we may be able to retrieve both sets of relevant documents if we include both sets of terms into the expanded query.

The methodology for building expanded queries from both local collocates and Okapi RF terms consists of the following stages:

1. Collocates are extracted from the relevant documents (as in the previously described method of query expansion with local collocates only) and ranked by local versions of Z or MI.
2. Okapi RF terms are selected from the relevant documents using the standard Okapi relevance feedback algorithm (Sparck Jones 2000).
3. Top *N* collocates per query term are added to the expanded query.
4. Top *I* Okapi RF terms are added to the expanded query. Duplicate terms are removed.
5. The expanded query is run against the text collection using Okapi BM25 function.

All above experiments were run within two types of experimental searching scenarios:

- *Retrospective searching*. 5 relevant documents for local collocation analysis were taken from the same document collection, which was then searched with the expanded queries;
- *Predictive searching based on half collections*. 5 relevant documents in the even half of the collection (all documents with even record numbers) were used for the extraction of local collocates. Searching with the expanded queries was done on the odd half of the collection.

The advantage of using predictive searching is that it allows us to see whether expansion terms have any predictive value. In retrospective searching very rare collocates, occurring, say, only in the known relevant documents, could perform well simply because they retrieve the same relevant documents from which they are derived. In predictive searching it would be evident that such terms have no or little predictive value to retrieve new relevant documents.

3.2.2 Results

Performance of local collocation analysis runs (tables A.2 and A.4 in the Appendix) was evaluated against Okapi relevance feedback runs (tables A.1 and A.3) with the same sets of relevant documents and using comparable numbers of Okapi RF terms for query expansion.

All retrospective and predictive collocation runs were substantially better than corresponding unexpanded runs. Queries expanded with both collocates and Okapi RF terms performed better than queries expanded with collocates only.

Runs with MI or Z collocates taken from large windows (200 and 100) demonstrated some precision growth with the increase in the number of collocates used. This tendency was not evident in runs with collocates from smaller window sizes. Another tendency, observed in MI and Z runs with a small number of collocates per query term (8, 12), is that precision grows with the decrease in the window sizes. It was less evident in the runs with a larger number of collocates per query term.

An interesting fact is that precision of the lowest performing runs '8 MI COL (200 window size)' and '8 Z COL (200 window size)' can be improved in two ways: either by decreasing the size of windows, or by increasing the number of collocates per query term. The upper limits of both ways of precision improvement are, however, very similar.

The best performance among both MI and Z predictive runs with collocates only was achieved by using 12 collocates from windows of size 20.

Combining of local Z collocates with Okapi RF terms for query expansion showed to perform reasonably well both retrospectively and predictively. Often addition of 20 Okapi RF terms results in precision gains over the corresponding runs with collocates only. Combined runs suggested some improvement over Okapi relevance feedback: run 'PRED 8 Z COL + 20 OK (100 window size)' is 4.3% better than 'PRED OK 35' and run 'PRED 16 MI COL + 20 OK (100 window size)' is 5.7% better than 'PRED OK 35'.

Window size has no consistent effect on the performance of combined runs. Combined runs did not demonstrate the pattern of precision growth with the increase in the number of collocates either.

Z score can be considered a somewhat better statistic than MI for the selection of query expansion terms, however the difference between the performance results of MI- and Z-ranked collocates was very narrow.

3.2.3 Influence of different categories of terms on performance

Though queries expanded with a combination of collocates and Okapi terms did not demonstrate large performance gains over Okapi RF, they suggested some tendencies towards improving performance. Because combined queries consisted of several categories of terms, we were interested to see what impact on performance each category of terms has.

For the analysis we chose one of the best runs in the predictive experiments – 'PRED 8 Z COL + 20 OK (100 window size)'. We identified 10 categories of terms that may be present in the expanded queries in combined runs (see table 4).

The influence of each term in the query on performance was identified by taking out this term from the query, running the query and recording its performance in average precision. Average precision of the query without the term in question was compared to average precision of the complete query. If, for example, average precision of the former is greater, then the term is considered to degrade the performance. We categorised the influence of each term on performance as:

- Improving;
- Indifferent;
- Degrading.

After performance data for each query term was obtained, we summed the number of query terms in each term category for each performance group. The results are summarised in table 4 and presented as a chart in figure 6.

Category	Improve	Indifferent	Degrade	Total
1. Collocate	83 (20%)	181 (44%)	151 (36%)	415
2. Collocate of 2 or more query terms	21 (17%)	62 (51%)	39 (32%)	122
3. Collocate of 1 query term	62 (21%)	119 (41%)	112 (38%)	293
4. Okapi RF term	187 (28%)	170 (26%)	302 (46%)	659
5. Original query term	84 (52%)	51 (32%)	25 (16%)	160
6. Collocate of 2 or more query terms and an Okapi RF term	11 (22%)	23 (47%)	15 (31%)	49
7. Collocate of 1 query term and an Okapi RF term	22 (24%)	30 (33%)	40 (43%)	92
8. Collocate and an Okapi RF term	33 (23%)	53 (37%)	55 (40%)	141
9. Collocate and an original term	11 (79%)	1 (7%)	2 (14%)	14
10. Okapi RF term and an original query term	43 (86%)	2 (4%)	5 (10%)	50

Table 4. Influence of categories of terms in the expanded queries 'PRED 8 Z COL + 20 OK (100 window size)' on average precision

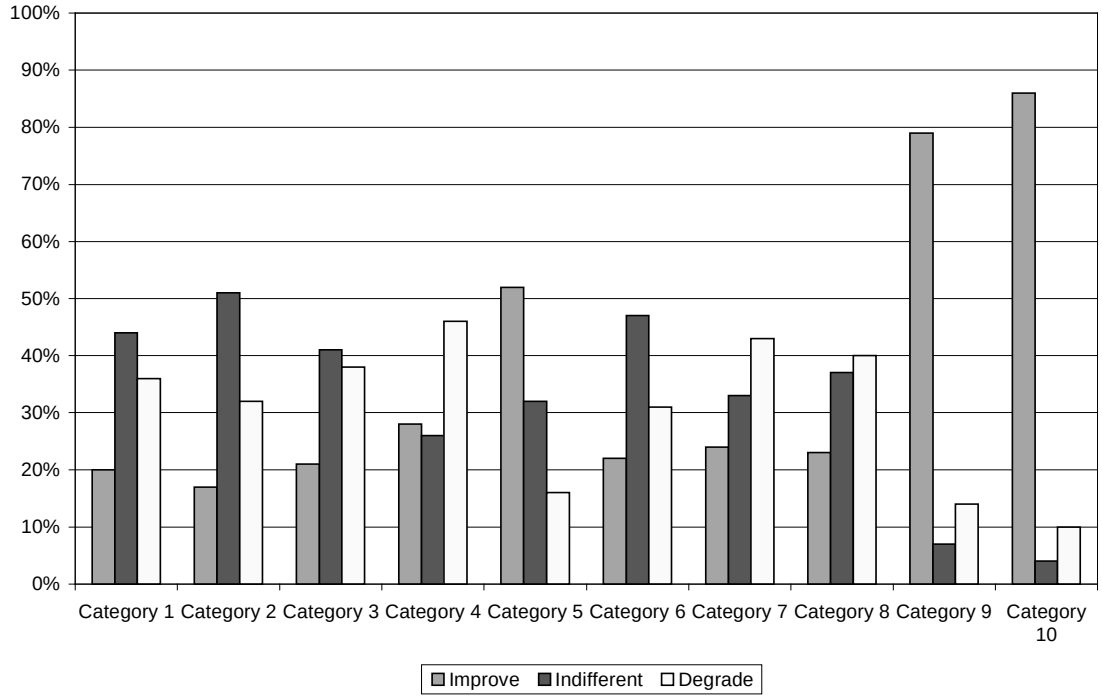

Figure 6. Influence of categories of terms in the expanded queries 'PRED 16 Z COL + 20 OK (200 window size)' on average precision

The results, rather surprisingly, showed that either collocates, or Okapi RF terms hurt precision in a larger number of cases, than improve it. More Okapi RF terms (category 4) than collocates (category 1) improve precision, however even more of them hurt precision.

It was expected that collocates of 2 or more query terms would be better relevance discriminators than collocates of 1 query term only. The results showed that collocates of 2 or more query terms (category 2) improve precision in fewer cases than any collocates (category 1) or collocates of 1 query term (category 3). On the other hand, collocates of 2 or more query terms hurt precision less often than categories 1 or 3.

Terms which are both collocates and Okapi RF terms (categories 6, 7 and 8) improve precision in a larger number of cases than terms of categories 1, 2 and 3. Although, except category 6 (a collocate of 2/more query terms and an Okapi RF term), they also hurt precision in a larger number of cases.

The category that suggested greater improvement than degradation of performance is category 5: *Original query terms*. The term's status as an original query term plus either a collocate (category 9), or an Okapi RF term (category 10), indicates a much higher relevance discriminating ability. There is, however, a rather low number of terms in these categories: 14 in category 9 and 50 in category 10.

4 Conclusions

In this paper we presented two novel methods of query expansion using two types of long-span collocates: global collocates – co-occurring significantly with query terms in the entire collection, and local collocates – co-occurring with query terms in a sub-set of retrieved documents. Query expansion with global collocates of each query term demonstrated worse performance than unexpanded queries. Our approach was to identify collocates of each query term independently of the other terms in the query. Since the majority of query terms are words with rather general lexical meanings, they tend to occur in a wide range of topics, many of them unrelated to the user's topic. Use of only those collocates which are associated strongly with *all* terms in a query may be a better approach, which will reduce the amount of noise terms from unrelated topics. There have been developed a number of collection-wide query expansion techniques which select terms on the basis of their association with the entire query (Qiu 1993, Jing 1994), and which demonstrated certain performance gains.

In other words, the reason why global collocates of individual query terms are poor expansion terms is that query terms have a very high dimensionality of contexts of occurrence. To achieve any performance gains, we need to reduce this dimensionality. Selection of collocates associated with all query terms is one way to do it. Our approach was to reduce the number of unrelated contexts by looking only at contexts of query term occurrences in a subset of documents judged relevant. Expansion with local collocates from such contexts of query terms demonstrated substantial performance improvements over unexpanded queries. Merging in the expanded queries local collocates and Okapi relevance feedback terms improved performance over both initial queries and queries expanded with local collocates alone. However, improvement over Okapi relevance feedback terms was marginal.

Our experiments with local collocates showed that the window size for collocate selection and collocation ranking measures are not the most critical factors affecting performance. Again, as in the expansion with global collocates, we might benefit from selecting collocates by their association with all terms in a query. There is evidence from past research that use of terms strongly associated with the entire query in local feedback contributes to performance gains: Xu and Croft's LCA (Xu 1996).

Another factor which might have strong effect on performance is the type of lexical units, used as collocates. We used any single terms as collocates. Terms belonging to certain parts of speech or composite lexical units conforming to specific phrase rules may be more effective than single terms. Jing and Croft (1994) conducted a rather extensive research of the effect expansion with different syntactic categories has on performance. Their experiments showed that noun phrases of the form N, NN, NNN result in better performance than any other types of single or compound terms.

To conclude, we believe that collocation is a rich source of information about contextual environments of terms, and further experimentation with collocates as query expansion terms is needed. The focus should be on selecting terms co-occurring significantly with all query terms. Also, despite the fact that in our experiments local collocates performed much better than global, we believe that global collocation methods deserve further study. Local feedback approaches may have one limitation: if we consider only few documents from the top of the retrieved set, which are likely to be biased towards the initial query, is the information we get from them always enough to retrieve other documents which might cover different aspects of the topics relevant to the user's need? In other words, do we get enough information to diversify the query formulation if we consider only a limited number of documents biased towards the underspecified short query formulation? It might be

that global techniques, provided that they can handle the task of selecting features related to the entire query, could give us a richer query expansion material than local techniques.

Acknowledgements

The work reported in this paper was conducted at City University, supported in part by a grant from Microsoft Research Ltd.

References

Beeferman D., Berger A., Lafferty J. (1997) A model of lexical attraction and repulsion. In: 35th Annual Meeting of the Association for Computational Linguistics [(proceedings of a conference in Madrid in 1997)], pp. 373-380.

Buckley C., Waltz J. (2000) SMART in TREC 8. In: Voorhees E.M. and Harman D.K., eds. The Eighth Text REtrieval Conference (TREC-8) [(proceedings of a conference in Gaithersburg in 1999)], Gaithersburg, MD, NIST, 2000. pp. 577-582.

Church K., Hanks P. (1990) Word association norms, mutual information and lexicography. American Journal of Computational Linguistics, 16(1):22-29.

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

Church K., Gale W., Hanks P., Hindle D. (1994) Lexical substitutability. In: Atkins B.T.S. and Zampoli A., eds. Computational Approaches to the Lexicon. Oxford University Press, 1994. pp. 153-177.

Cormack G.V., Clarke C.L.A., Palmer C.R. and Kisman D.I.E. (2000) Fast Automatic Passage Ranking (MultiText Experiments for TREC-8). In: Voorhees E.M. and Harman D.K., eds. The Eighth Text REtrieval Conference (TREC-8) [(proceedings of a conference in Gaithersburg in 1999)], Gaithersburg, MD, NIST, 2000. pp. 735-742.

Croft W.B., Turtle H.R. and Lewis D.D (1991) The use of phrases and structured queries in information retrieval. In: the 14th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR '91) [(proceedings of a conference in Chicago in 1991)], ACM Press, New York, 1991. pp. 32-45.

Edmonds P. (1997) Choosing the word most typical in context using a lexical co-occurrence network. In: 35th Annual Meeting of the Association for Computational Linguistics [(proceedings of a conference in Madrid in 1997)], pp. 507-509.

Fagan J.L. (1989) The effectiveness of a non-syntactic approach to automatic phrase indexing for document retrieval. Journal of the American Society for Information Science, 40(2):115-132.

Fano R. (1961) Transmission of information. Cambridge, Massachusetts, MIT Press, 1961.

Halliday M.A.K., Hasan R. (1976) Cohesion in English. Longman, 1976.

Harper D. J., Van Rijsbergen C. J. (1978) An evaluation of feedback in document retrieval using co-occurrence data. Journal of Documentation, 34(3):189-216.

Hawking D., Thistlewaite P., Craswell N. (1998) ANU/ACSys TREC-6 Experiments. In: Voorhees E.M. and Harman D.K., eds. The Sixth Text REtrieval Conference (TREC-6) [(proceedings of a conference in Gaithersburg in 1997)], Gaithersburg, MD, NIST, 1998, pp. 275-291.

Hearst, M. (1994) Multi-paragraph segmentation of expository text. In: 32nd Annual Meeting of the Association for Computational Linguistics [(proceedings of a conference in New Mexico in 1994)], pp. 9-16.

Hoey M. (1991) Patterns of Lexis in Text. Oxford University Press, 1991.

Ishikawa K., Satoh K., Okumura A. (1998) Query Term Expansion based on Paragraphs of the Relevant Documents. In: Voorhees E.M. and Harman D.K., eds. The Sixth Text REtrieval Conference (TREC-6) [(proceedings of a conference in Gaithersburg in 1997)], Gaithersburg, MD, NIST, 1998. pp. 577-584.

Jelinek F. (1990) Self-organised language modelling for speech recognition. In: Waibel A. and Lee K., eds. Readings in Speech Recognition. San Mateo, California, Morgan Kaufmann Publishers, 1990.

Jing Y. and Croft B. (1994) An association thesaurus for information retrieval. In: RIAO'94, 4th International Conference [(proceedings of a conference in New York, 1994)] pp.146-160.

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

McEnery T. and Wilson A. (1996) Corpus Linguistics, Edinburgh, 1996.

Qiu Y. and Frei H.P. (1993) Concept based query expansion. In: the 16th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR '93) [(proceedings of a conference in Pittsburgh in 1993)], ACM Press, New York, 1993. pp. 160-169.

Palmer F.R. (1981) Semantics. Second edition. Cambridge University Press, 1981.

Robertson S.E., Sparck Jones K. (1976) Relevance Weighting of Search Terms. Journal of the American Society for Information Science, 27:129-146.

Sinclair J. M., Jones S. (1974) English lexical collocations: A study in computational linguistics, 1974. Reprinted as chapter 2 of Foley J.A. ed., J.M. Sinclair on Lexis and Lexicography. Singapore: UniPress, 1996.

Smeaton A.F. and Van Rijsbergen C.J. (1983) The retrieval effects of query expansion on a feedback document retrieval system. The Computer Journal, 26(3):239-246.

Sparck Jones K. and Jackson D.M. (1970) The use of automatically obtained keyword classifications for information retrieval. Information Processing and Management, 5:175-201.

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

Strzalkowski T. et al. (2000) Natural Language Information Retrieval: TREC-8 Report. In: Voorhees E.M. and Harman D.K., eds. The Eighth Text REtrieval Conference (TREC-8) [(proceedings of a conference in Gaithersburg in 1999)], Gaithersburg, MD, NIST, 2000. pp. 381-390.

Van Rijsbergen C. J. (1977) A theoretical basis for the use of co-occurrence data in information retrieval. Journal of Documentation: 33(2):106-119.

Vechtomova O. and Robertson S. (2000) Integration of collocation statistics into the probabilistic retrieval model. In: the 22nd BCS-IRSG [(proceedings of a conference in Cambridge, England, 2000)], pp. 165-177.

Xu J. and Croft B. (1996) Query expansion using local and global document analysis. In: the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR '96) [(proceedings of a conference in Zurich in 1996)], ACM Press, New York, 1996. pp. 4-11.

Yarowsky D. (1992) Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In: COLING-92 [(proceedings of a conference in Nantes in 1992)], 1992. pp. 454-460.

Appendix

Run name	Query description	Average precision
RETRO UNEXPANDED	Original query terms	**0.1310**
RETRO OK 20	20 Okapi RF terms	**0.4945**
RETRO OK 25	25 Okapi RF terms	**0.5096**
RETRO OK 30	30 Okapi RF terms	**0.5184**
RETRO OK 35	35 Okapi RF terms	**0.5259**

Table A.1. Retrospective performance results of Okapi runs

Run name	Window size / Query description	200	100	50	30	20
RETRO 8 Z COL	8 Z collocates/query term	**0.4758**	**0.4720**	**0.4733**	**0.4857**	**0.4810**

RETRO 12 Z COL	12 Z collocates/query term	**0.4896**				
RETRO 16 Z COL	16 Z collocates/query term	**0.5034**				
RETRO 21 Z COL	21 Z collocates/query term	**0.5029**				
RETRO 8 Z COL + 20 OK	8 Z collocates/query term + 20 Okapi RF terms	**0.5194**	**0.5230**	**0.5245**	**0.5258**	**0.5263**
RETRO 16 Z COL + 10 OK	16 Z collocates/query term + 10 Okapi RF terms	**0.5257**				
RETRO 16 Z COL + 20 OK	16 Z collocates/query term + 20 Okapi RF terms	**0.5171**	**0.5264**	**0.5271**	**0.5313**	**0.5316**
RETRO 21 Z COL + 10 OK	21 Z collocates/query term + 10 Okapi RF terms	**0.5219**				
RETRO 8 MI COL	8 MI collocates/query term	**0.4458**	**0.4610**	**0.4551**	**0.4690**	**0.4733**
RETRO 12 MI COL	12 MI collocates/query term	**0.4688**				
RETRO 16 MI COL	16 MI collocates/query term	**0.4877**				
RETRO 21 MI COL	21 MI collocates/query term	**0.4991**				
RETRO 8 MI COL + 20 OK	8 MI collocates/query term + 20 Okapi RF terms	**0.5227**	**0.5251**	**0.5240**	**0.5249**	**0.5274**
RETRO 16 MI COL + 10 OK	16 MI collocates/query term + 10 Okapi RF terms	**0.5274**	**0.5260**	**0.5292**	**0.5291**	**0.5270**
RETRO 16 MI COL + 20 OK	16 MI collocates/query term + 20 Okapi RF terms	**0.5264**	**0.5267**	**0.5282**	**0.5301**	**0.5290**
RETRO 21 MI COL + 10 OK	21 MI collocates/query term + 10 Okapi RF terms	**0.5272**				

Table A.2. Retrospective performance results (in average precision) of query expansion runs with local collocates ranked by local Z and MI

Run name	Query description	Average precision
PRED UNEXPANDED	Original query terms	**0.0799**
PRED OK 10	10 Okapi RF terms	**0.1343**
PRED OK 20	20 Okapi RF terms	**0.1400**
PRED OK 25	25 Okapi RF terms	**0.1520**
PRED OK 30	30 Okapi RF terms	**0.1483**
PRED OK 35	35 Okapi RF terms	**0.1533**

Table A.3. Predictive performance results of Okapi runs

Run name	Window size / Query description	200	100	50	30	20
PRED 8 Z COL	8 Z collocates/query term	**0.1268**	**0.1294**	**0.1302**	**0.1376**	**0.1433**
PRED 12 Z COL	12 Z collocates/query term	**0.1346**	**0.1383**	**0.1459**	**0.1401**	**0.1518**
PRED 16 Z COL	16 Z collocates/query term	**0.1386**	**0.1456**	**0.1362**	**0.1367**	**0.1432**

PRED 21 Z COL	21 Z collocates/query term	**0.1391**	**0.1480**	**0.1325**	**0.1396**	**0.1384**
PRED 8 Z COL + 20 OK	8 Z collocates/query term + 20 Okapi RF terms	**0.1536**	**0.1602**	**0.1549**	**0.1568**	**0.1553**
PRED 16 Z COL + 10 OK	16 Z collocates/query term + 10 Okapi RF terms	**0.1388**				
PRED 16 Z COL + 20 OK	16 Z collocates/query term + 20 Okapi RF terms	**0.1527**	**0.1561**	**0.1495**	**0.1485**	**0.1477**
PRED 21 Z COL + 10 OK	21 Z collocates/query term + 10 Okapi RF terms	**0.1413**				
PRED 8 MI COL	8 MI collocates/query term	**0.1006**	**0.0971**	**0.1158**	**0.1399**	**0.1371**
PRED 12 MI COL	12 MI collocates/query term	**0.1222**	**0.1202**	**0.1365**	**0.1435**	**0.1524**
PRED 16 MI COL	16 MI collocates/query term	**0.1302**	**0.1437**	**0.1341**	**0.1372**	**0.1380**
PRED 21 MI COL	21 MI collocates/query term	**0.1390**	**0.1440**	**0.1295**	**0.1355**	**0.1359**
PRED 8 MI COL + 20 OK	8 MI collocates/query term + 20 Okapi RF terms	**0.1455**	**0.1426**	**0.1501**	**0.1565**	**0.1528**
PRED 16 MI COL + 10 OK	16 MI collocates/query term + 10 Okapi RF terms	**0.1398**				
PRED 16 MI COL + 20 OK	16 MI collocates/query term + 20 Okapi RF terms	**0.1515**	**0.1626**	**0.1522**	**0.1486**	**0.1487**
PRED 21 MI COL + 10 OK	21 MI collocates/query term + 10 Okapi RF terms	**0.1477**				

Table A.4. Predictive performance results (in average precision) of query expansion runs with local collocates ranked by local Z and MI