Applying Summarization Techniques for Term Selection in Relevance Feedback

Adenike M. Lam-Adesina
Department of Computer Science
University of Exeter
Exeter EX4 4PT
United Kingdom
A.M.Lam-Adesina@ex.ac.uk

Gareth J. F. Jones
Department of Computer Science
University of Exeter
Exeter EX4 4PT
United Kingdom
G.J.F.Jones@ex.ac.uk

ABSTRACT

Query-expansion is an effective Relevance Feedback technique for improving performance in Information Retrieval. In general query-expansion methods select terms from the complete contents of relevant documents. One problem with this approach is that expansion terms unrelated to document relevance can be introduced into the modified query due to their presence in the relevant documents and distribution in the document collection. Motivated by the hypothesis that query-expansion terms should only be sought from the most relevant areas of a document, this investigation explores the use of document summaries in query-expansion. The investigation explores the use of both context-independent standard summaries and query-biased summaries. Experimental results using the Okapi BM25 probabilistic retrieval model with the TREC-8 ad hoc retrieval task show that query-expansion using document summaries can be considerably more effective than using full-document expansion. The paper also presents a novel approach to term-selection that separates the choice of relevant documents from the selection of a pool of potential expansion terms. Again, this technique is shown to be more effective that standard methods.

1. INTRODUCTION

Information retrieval (IR) is the process of selecting documents from a collection in order to attempt to satisfy a user's information need. Specifically current IR systems respond to a user's search request by returning a list of potentially useful documents ranked according to a query-document relevance score. Often user search requests are very short consisting of only a few words, meaning that it is can be hard to retrieve relevant documents and rank them effectively.

Relevance feedback (RF) has been shown to be an effective method of improving retrieval performance [1][2]. In RF relevance information gathered from documents retrieved in a ranked list generated using an initial request is used to modify the search query and perform a further retrieval pass. Although RF is most effective when relevance information about retrieved documents is provided by users, it has been shown that some improvement in retrieval performance can be achieved via a pseudo-relevance feedback (PRF) method [3][4]. In PRF a number of the documents at the top of the ranked list are assumed to be relevant and then relevance feedback methods are applied as if these documents were known to be relevant. RF usually consists of two components: term reweighting and query-expansion, only the latter is considered in this investigation. Expansion terms can be selected from relevant documents (or assumed relevant in the case of PRF) according to various criteria, e.g. [2][5][6]. This approach carries the risk that terms which are unrelated to relevance, but happen to meet the selection criteria, will be added to the query with subsequent adverse effects on retrieval behaviour.

This paper describes an investigation into the use of document summarization to improve term selection in query expansion for PRF. Retrieved documents are summarized and expansion terms selected from the summaries rather than the whole document. Both context-independent summarization, where an identical summary is used regardless of the terms in the query, and query-biased summarization, where the summary depends on the terms appearing in the original query are explored. Retrieval results using the TREC-8 ad hoc task show that use of summarization can improve the effectiveness of PRF.

Further, in RF it is usually assumed that expansion terms should be selected from all available relevant documents. As part of our investigation we explore a technique that challenges this assumption. In this approach potential expansion terms are chosen only from the top ranked relevant documents, while the information used to select the best expansion terms from among these terms is based on relevance information derived from a larger set of documents. In effect the pool of available expansion terms is limited while more evidence is used to rank their potential as expansion terms. Again this approach is shown to improve retrieval performance.

The remainder of this paper is organized as follows: Section 2 reviews the use of relevance feedback in ad hoc IR systems, Section 3 outlines our text summarization methods, Section 4 describes the Okapi retrieval system used in our experimental investigation, Section 5 outlines the test data, Section 6 gives experimental results, and finally Section 7 provides conclusions and outlines directions for further work.

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA.

2. RELEVANCE FEEDBACK

In principle, information storage and retrieval is simple [26]. Assume a store of documents containing potentially relevant information. A user formulates a request for information for which the answer is contained within one or more of documents within the archive. The IR system then seeks to locate these documents and return them to the user. In practice returning relevant documents without returning non-relevant documents is usually not possible, and IR systems are configured to retrieve the largest possible number of relevant documents at high rank while retrieving the minimum number of non-relevant documents. Ideally the retrieved relevant documents should always be placed above the non-relevant ones in the ranked list.

Very often however, an IR system is unable to satisfy the user completely by the initial ranked output. Presuming that relevant documents are present in the archive, the main reason for this is often the query-document match problem. Queries are often short and imprecise, meaning that there can be problems with ambiguity and lack of terms to match with potentially relevant documents [7][8][9]. For example, documents may sometimes contain synonyms of the users' query terms. This problem often results in relevant documents being retrieved at low rank or not being retrieved at all, and retrieval of non-relevant documents at high ranks. It has been shown however that longer query statements reduce the ambiguity associated with very short queries [9] and word sense, thus it is essential to improve the quality of such queries by providing additional information. Relevance feedback (RF) using query expansion is one method which seeks to overcome the query-document match problem.

As outlined in Section 1, pseudo-relevance feedback (PRF) methods are on average found to give improvement in retrieval performance; although this is usually smaller than that observed for true RF. In PRF problems can arise when terms taken from assumed relevant documents that are actually non-relevant, are added to the query causing a drift in the focus of the query. For example, Mitra et. al. [8] report a result for TREC query number 203 "What is the economic impact of recycling tires", only 4 out of the retrieved documents were actually relevant. The rest of the documents retrieved in a first pass dealt with the recycling of other things such as plastics and glass. Adding words like "plastic" and "glass" taken from the first 20 retrieved documents to the initial query resulted in the revised query drifting away from the original focus. Thus, many more non-relevant documents were retrieved at high rank in a subsequent retrieval pass. Thus for this query PRF resulted in considerable loss of performance. However, if the initial retrieval results are good and a large proportion of the documents retrieved at high rank are relevant, feedback is likely to improve retrieval performance.

A further problem arises since many documents are multi-topic, i.e. they deal with several different topics. This means that only a portion of a document retrieved in response to a given query may actually be relevant. Nevertheless standard RF treats the whole document as relevant, the implication of this being that using terms from non-relevant sections of these documents for expansion may also cause query drift.

The exclusion of terms from non-relevant sections of documents, or those present in non-relevant documents which are not closely related to the concepts expressed in the initial query, could thus be beneficial to PRF and potentially in true RF as well. These issues have led to several attempts to develop automatic systems that can concentrate user's attention on the parts of the text that possess a high density of relevant information. This method known as passage retrieval [10][11] has the advantage of being able to provide an overview of the distribution of the relevant pieces of information within the retrieved documents. However even this approach does not alleviate the need to refer to the full text of such documents [12], nor has it being found to provide significant improvement in retrieval performance.

This paper is concerned with a novel approach to the exclusion of terms from consideration based on document summarization. In this method only terms present in the summarized documents are considered for query expansion. Earlier experiments [23][24] found out that selecting best passages from documents for query expansion is very effective in reducing the number of inappropriate possible feedback terms taken from multi-topic or non-relevant document. This submission is also partly supported by [13] which confirms that a summary helps a user to decide whether it will be worthwhile to look at the full document. In [12] Tombros shows that query-biased summaries are more effective than using simple leading sentence summaries for user relevance decisions. Thus, in this study we focus on summaries taken across the whole document. In addition to standard summaries which are independent of context, we also investigate the use of query-biased summaries. The aim of these summaries is to direct the focus of the summary onto the parts of the documents which are most likely to be relevant to the query and thus are most likely to contain potentially useful expansion terms. Section 3 reviews the topic of text summarization and describes the methods used in this investigation.

A related approach to the one reported here is described by Strzalkowski in [6] where a RF procedure using summaries of retrieved relevant documents is used. In this system expansion terms are taken from both non-relevant and relevant documents without the terms coming from the non-relevant documents having any negative effect on the retrieval performance. However, the fact that only the summaries of the documents were used as opposed to the entire text is probably responsible for removal of many terms that would otherwise have degraded retrieval performance. A potential weakness of the approach taken in this work is that *all terms* from the summaries are added to the query. We prefer to adopt the approach taken in the Okapi TREC submissions [4][14][15], which expand queries conservatively using only a small number of terms chosen using a statistical selection criteria [5]. In addition, we hope to see further improvements arising from the use of query-biased summaries.

3. SUMMARY GENERATION

Summary generation methods seek to identify document contents that convey the most "important" information within the document, where importance may depend on the use to which the summary is to be put. There are two basic approaches to summarization: information extraction with subsequent text generation, and summaries composed of extracted sentences or phrases. Since we require a very robust summarizer for the different text types likely to be encountered within a retrieval system we adopt the latter method in our work.

Sentence extracted summaries are formed by scoring the sentences in the document using some criteria, ranking the sentences, and then taking a number of the top ranking sentences as the summary. Various studies have led to the proposal of the following criteria of measuring sentence significance for effective summary generation:
(1) sentence position within the document;
(2) word frequency within the full-text;
(3) the presence or absence of certain words or phrases in the sentence;
(4) a sentence's relation to other sentences, words or paragraphs within the source document.

Each sentence score is computed as the sum of its constituent words and other scores. The following section describes the summary generation methods used in this investigation.

3.1 Luhn's Keyword Cluster Method

The first component of our summaries uses Luhn's classic cluster measure [16]. In order to determine the sentences of a document that can be used as the summary, a measure is required by which the information content of all the sentences can be analysed and graded. Luhn concluded that the frequency of a word occurrence in an article, as well as its relative position determines its significance in that article. He justified this assumption of word significance on the fact that writers usually repeat certain words as they advance in a write-up or elaborate on an aspect of the subject. He further concluded that the more often certain words are found in each other's company, the more significance can be attributed to them.

Based on this hypothesis, Luhn's method first generates a list of candidate terms that occur in the body of the documents in descending order of their term frequency within the document. Words with high frequency of occurrence within a document and those with very low frequency of occurrence in each document are classified as *insignificant* words. Words with the highest frequency within an individual document will often be standard stopwords and can be deleted. However some high frequency words may in fact be very significant, but using Luhn's basic approach would also be deleted. Take for instance TREC query 403 on the topic of "osteoporosis", the important term "bone" (which is very much related to the initial query term "osteoporosis") had a high occurrence frequency in all the documents retrieved at high rank in the first retrieval run and therefore in the documents used for feedback. In one particular document it occurred up to 40 times presenting the highest occurrence for any term in that document. However investigation of the resultant modified query showed that adding this term to the initial query for further search leads to the retrieval of additional relevant documents. Using Luhn's method might however have resulted in the classification of term "bone" as non-significant and hence to it being overlooked. Thus, a more beneficial approach might be to construct a fixed stopword list and only classify high frequency terms from this list as insignificant. This is the approach taken in this work and the process is described below.

In addition, a lower limit for significance needs to be defined. This depends on the characteristics of the documents to be summarized. Following the work of Tombros [17], which studied summarization of TREC documents, the required minimum occurrence count for significant terms in a medium-sized TREC document was taken to be 7; where a medium sized document is defined as one containing no more than 40 sentences and not less than 25 sentences. For documents outside this range, the limit for significance is computed as

$$ms = 7 + [0.1(L - NS)]$$

for documents with NS < 25 and

$$ms = 7 + [0.1(NS - L)]$$

for documents with NS > 40

where ms = the measure of significance
L = Limit (25 for NS < 25 and 40 for NS > 40)
NS = number of sentences in the document

In order to score sentences based on the number of significant words contained in them, Luhn reasoned that whatever the topic under consideration the closer certain words are, the more specifically an aspect of the subject is being treated. Hence, wherever clusters of significant words are found, the probability is very high that the information being conveyed is most representative of the article. Luhn specified that two significant words are considered *significantly related* if they are separated by not more than five insignificant words. Thus, a cluster of significant words is created whereby significant words are separated by not more than five non-significant words as illustrated below.

'The sentence [**scoring** process utilises **information** both from the **structural**] organization.'

The cluster of significant words is given by the words in the bracket ([---]), where significant words are shown in bold. The cluster significance score factor for a sentence is given by the following formula

$$SS1 = \frac{SW^2}{TW}$$

where SS1 = the sentence score
SW = the number of bracketed significant words (in this case 3)
TW = the total number of bracketed words (in this case 8)

Thus SS1 for the above sentence is 1.125. If two or more clusters of significant words appear in a given sentence, the one with the highest score is chosen as the sentence score.

3.2 Title Terms Frequency Method

The title of an article often reveals the major subject of that document. This hypothesis was examined in a sample study of TREC documents where the title of each article was found to convey the general idea of its contents. Thus, a factor in the sentence score is the presence of title words within the sentence. In order to utilise this attribute in the summary generation process, each constituent term in the title section is looked up in the body of the text. For each sentence a title score is computed as follows,

$$SS2 = \frac{TTS}{TTT}$$

where SS2 = the title score for a sentence
TTS = the total number of title terms found in a sentence
TTT = the total number of terms in a title

TTT is used as a normalization factor to ensure that this method does not have an excessive sentence score factor contribution relative to the overall sentence score.

3.3 Location / Header Method

Edmundson [18] noted that the position of a sentence within a document is often useful in determining its importance to the document. Based on this, Edmundson defined a location method for scoring each sentence based on whether it occurs at the beginning or end of a paragraph or document. The importance of the position of a sentence within a text has been confirmed more recently in [19].

To determine the effect of this sentence scoring method on the test collection a further sample study was conducted. This confirmed that the first sentences of a TREC document often provide important information about the content of the document. Thus the first two sentences of an article are assigned a location score computed as below

$$SS3 = \frac{1}{NS}$$

where SS3 = the location score for a sentence
NS = the number of sentences in the document.

Furthermore, section headings within the documents were found to provide information about the different sections discussed in the documents. Thus, marked section headings were given a similar location score.

Another criteria used in scoring a sentence is based on its position in a paragraph. However, due to the general structure of the TREC documents, where the documents are in most cases fragmented into several paragraphs with most of them consisting of less than 3 sentences and in some cases having no paragraph section at all, this criterion was felt to be inappropriate and was not used here.

3.4 Query-Bias Method

The addition of a sentence score factor bias to score sentences containing query terms more highly may reduce the query drift caused by the use of bad feedback terms. Thus, whether a relevant or non-relevant document is used, the feedback terms are taken from the most relevant section identified in the document, in relation to the submitted query.

In order to generate a query biased summary in this work, each constituent sentence of a document being processed is scored based on the number of query terms it contains. The following situation gives an example of this method.

- For a query "falkland petroleum exploration" and
- A sentence "The british minister has decided to continue the ongoing **petroleum exploration** talks in the **falkland** area"

The query score SS4 is computed as follows

$$SS4 = \frac{tq^2}{nq}$$

where tq = the number of query terms present in a sentence
nq = the number of terms in a query

Therefore the query score SS4 for the above sentence is 3. This score is assigned based on the belief that the number of query terms contained in a sentence, the more likely it is that this sentence conveys a large amount of information related to the query. This was the same method used in [12].

In order to test the effectiveness of the above equation a comparison was made between retrieval performance using the above equation and assigning an ordinal score for each occurrence of a query term in a sentence (for each occurrence of the query term in each sentence an ordinal score of 1 was ascribed to that sentence, thus for a sentence containing 10 occurrences of a query term, the query score SS4 = 10). The result of comparing the performance of the two methods of query-biased summary construction favoured the Tombros query method [12] in terms of improvement in average precision. Thus, the ordinal method was discarded in favour of the Tombros query method in this work.

3.5 Combining the Scores

The previous sections outlined the components used in scoring sentences to generate the summaries used in this work. The final score for each sentence is calculated by summing the individual score factors obtained for each method used. Thus the final score for each sentence is

$$SSS = SS1 + SS2 + SS3 + SS4$$

where SSS = Sentence Significance Score

The summarization system was implemented such that each method could be invoked independently. Thus it was possible to experiment with various combinations of the methods described above to determine the best summarization method(s) for term selection in PRF.

In order to generate an appropriate summary it is essential to place a limit on the number of sentences to be used as the summary content. To do this however it is important to take into consideration the length of the original document and the amount of information that is needed. The objective of the summary generation system is to provide terms to be used for query expansion, and not to act as a stand alone summary that can be used to replace the entire documents. Hence the optimal summary length is a compromise between maintaining terms that can be beneficial to the retrieval process, while ensuring that the length is such that non relevant terms are kept to the barest minimum if they cannot be removed totally. Experiments were performed with various maximum summary lengths to find the best one for term-selection. The lower limit of the summary length was set at 15% of the original document length because the document collection also consisted of very short documents. Thus high ranked sentences up to the maximum summary length and not less than the set minimum summary length are presented as the summary content for each document summarized. Inspection of our example summaries showed them to be reasonable representations of the original documents. However, in our case an objective measure of summary quality is their overall effect on retrieval performance.

4. INFORMATION RETRIEVAL TECHNIQUES

The experiments were carried out using the City University research distribution version of the Okapi system. The document and search topics were processed to remove stop words from a list of around 260 words, suffix stripped using Porter stemming [20] and terms were further indexed using a small set of synonyms.

4.1 Term Weighting

The BM25 probabilistic model was used for term weighting combining *standard collection frequency weighting (cfw)* (also called *inverse document frequency* weighting), a *term frequency* function, and document length normalisation. BM25 was originally developed in [21] and further elaborated in [4]. The BM25 *cw* weight for a term is calculated as follows,

$$cw(i,j) = \frac{cfw(i) \times tf(i,j) \times (k1+1)}{k1 \times ((1-b) + b(ndl(j))) + tf(i,j)}$$

where *cw(i,j)* represents the weight of term *i* in document *j*, *cfw(i)* is the standard collection frequency weight, *tf(i,j)* is the document term frequency, and *ndl(j)* is the normalized document length. *ndl(j)* is calculated as,

$$ndl(j) = \frac{dl(j)}{avdl(n)}$$

where *dl(j)* is the length of *j and avdl(n)* is the average document length for all documents. *k1* and *b* are empirically selected tuning constants for a particular collection. *k1* is designed to modify the degree of effect of *tf(i,j)*, while constant *b* modifies the effect of document length. High values of *b* imply that documents are long because they are verbose, while low values imply that they are long because they are multi-topic.

4.2 Relevance Feedback

Query expansion terms are selected by ranking all terms appearing in relevant documents according to a selection criteria. A number of terms are then added to the original query for further search. The expansion term ranking criteria used here was the Robertson selection value (*rsv*) [5] which has consistently shown itself to be the best currently available measure in many investigations. The *rsv* is defined as,

$$rsv(i) = r(i) \times rw(i)$$

where *r(i)* is again the number of relevant documents containing term *i*, and *rw(i)* is the standard Robertson/Sparck Jones relevance weight [22]. *rw(i)* is defined as,

$$rw(i) = \log \frac{(r(i)+0.5)(N-n(i)-R+r(i)+0.5)}{(n(i)-r(i)+0.5)(R-r(i)+0.5)}$$

where *n(i)* = the total number of documents containing term *i*,
R = the total number of relevant documents for this query, and
N = the total number of documents.

The *rsv* is generally based on taking an equal number of relevant documents for both the available expansion terms and term ranking. In our experiments we explore the use of an enhanced method which takes a smaller number of relevant documents to determine the pool of expansion terms than the number of documents used to determine the *rsv* ranking. Thus, as described in Section 6 we observe best PRF results by assuming 5 relevant documents for the expansion term pool set, but 20 for term ranking using the *rsv*. It should be noted however that the r(i) value for each term i is calculated based on its occurrence in the entire document rather than on the summary alone.

5. TEST DATA

The summary based term selection methods are evaluated using the TREC-8 ad hoc test set. The ad hoc task investigates the performance of systems that search a static document collection using new query statements. This is similar to the situation normally found in a library, where a user submits a query against a static set of documents. The document set consists of approximately 538,151 documents distributed on two CD-ROM disks (TREC disks 4 and 5) taken from the following sources: Federal Register (FR), Financial Times (FT), Foreign Broadcast Information Service (FBIS), LA Times (LAT). All documents are tagged using SGML with markup to indicate document number, title and section headings. Errors present in the documents are left uncorrected to provide a better simulation of a real-world task.

Search requests in the form of TREC topics consist of three parts: title, description and narrative. The title consists of individual words that best describe the information need, the description field is a one sentence description of the topic area, while the narrative gives a concise description of what makes a document relevant and what makes it irrelevant. The different parts of the TREC topic allow investigation of the effect of different query lengths on retrieval performance. For our investigation only the title field of the topics is used since this is most similar to the form of queries entered by typical users.

For proper evaluation of any retrieval system performance, a relevance assessment of each document for each topic is required. To make the relevance judgements the pooling method is used to identify a set of relevant documents from a pool of potentially relevant ones. Our experiments used the standard TREC-8 ad hoc relevance set developed using the data provided by the track participants. Each participating group returned their 1000 top-ranked documents for each topic to NIST. The top 100 documents retrieved in the top rankings generated by each participant were merged to form a pool for assessment. This pool was then shown to human assessors who performed the task of relevance judgement.

6. EXPERIMENTAL INVESTIGATION

This section describes our experimental results in a series of comparisons as follows. First, we report baseline retrieval results initially without feedback and then using standard term selection based on full documents for PRF query-expansion and context-independent summarization for query-expansion without term selection. We then give results for feedback runs exploring various methods of document summarization, first for context-independent summaries and then for query-biased summaries.

6.1 Baseline Runs

The closest point of reference for these experiments is the Okapi submission to TREC 8, we thus for consistency use the same system parameters as used in their submission [14] for our baseline (no feedback) run. Our baseline run) was performed with k1 = 1.0 and b = 0.5 respectively as used in the Okapi TREC-8 experiments.

Table 1: Baseline retrieval result and published Okapi results for TREC 8 ad hoc

Short Topic	P30	AveP
Okapi	343	239
BL-1	361	240

Table 1 shows retrieval precision at 30 document cutoff and TREC average precision for the published Okapi baseline without feedback and our baseline results without feedback (BL-1). These results are very close showing that systems achieve comparable performance prior to the application of automated feedback methods.

6.2 Baseline Pseudo-Relevance Feedback

Our experimental results for PRF use only the TREC-8 disks 4 & 5 (described above) test set for term selection. By contrast the PRF results reported for the original Okapi experiments used the complete contents of TREC disks 1-5 [15]. Experience suggests that more effective term selection results from the use of this additional data, thus the results reported here are likely to represent a lower bound on the performance achievable with these methods.

Again for direct comparison with Okapi results all our feedback runs were performed with k1=1.5 and b = 0.6, with all original query terms upweighted by multiplying the original term weights by 3.5 [14].

Table 2: Baseline Pseudo-Relevance Feedback Results

		P10	P30	AveP
BL-2	Baseline, expansion all terms from context-independent summaries, no selection	472	362	241
BL-3a1	Baseline, expansion no summary, 5 terms selected, 5 documents used in term selection	440	352	244
BL-3a2	As in BL-3a1 but 20 documents used for term selection	466	365	248
BL-3b1	As in BL-3a1 but 10 terms selected	434	344	244
BL-3b2	As in BL-3a2 but 10 terms selected	482	356	246
BL-3c1	As in BL-3a1 but 20 terms selected	428	326	238
BL-3c2	As in BL-3a2 but 20 terms selected	490	349	244

Table 2 shows a summary of baseline PRF results. In all cases the top 5 retrieved documents are assumed to be relevant. BL-2 shows the results for the baseline feedback run using all terms from context-independent document summaries of maximum length 6 sentences generated using all sentence scoring methods described in Section 3 except query bias. Runs BL-3a1, BL-3b2, BL-3b1, BL-3b2, BL-3c1, and BL-3c2 show results for term selection using *rsv* with 5, 10, and 20 expansion terms from the whole document. The BL-3*1 results indicate the standard method of assuming the same number of relevant documents for term selection and ranking, the BL-3*2 results take the top 5 documents for term selection, but top 20 for term ranking. As can be seen all methods provide a small improvement over the initial baseline, with the *rsv* selection method performing marginally better and the new term selection approach producing consistently better results. These results are the best obtained with these methods after various parameter settings had been explored.

6.3 Expansion Term Selection from Document Summaries

First runs were done using terms taken from standard summaries and secondly runs were performed using terms taken from query-biased summaries. For both sets of experiments the 5 top-ranked documents were assumed relevant and used for feedback, and the enhanced term selection method with 20 assumed relevant documents for term ranking is used throughout.

Table 3: Pseudo-Relevance Feedback Results used Context-Independent Summaries

SUMMARY LENGTH	METHOD	P10	P30	AveP
4	TITLE(T)	478	369	***264***
	LUHN(L)	468	372	263
	LOCATION(M)	478	374	***264***
	ALL(A)	460	354	258
6	TITLE(T)	498	379	***268***
	LUHN(L)	480	378	264
	LOCATION(M)	462	386	266
	ALL(A)	466	382	266
9	TITLE(T)	468	376	273
	LUHN(L)	476	374	271
	LOCATION(M)	474	372	272
	ALL(A)	486	385	***274***

Table 3 shows results using standard context-independent summaries with 20 expansion terms. It can be seen that the title summary method performs consistently well with performance improving as the maximum summary length is increased from 4 to 9 sentences, indicating that 4 sentence summaries are too short to make sufficient useful expansion terms available. These results show a large improvement over the results for full document and all-term summary expansion in Table 2. Again the parameters chosen here give the best results achieved so far using these methods.

Table 4 below shows results for experiments performed using the query-biased version of the summaries above. All parameters are the same as for the previous experiments. In this case the Luhn method is observed to generally perform best in isolation, with best performance being obtained on average for summaries of maximum length 6 sentences. Results here are in general marginally better than those for context-independent summaries shown in Table 3.

Table 4: Results for Pseudo-Relevance Feedback using term selection from Query-Biased Summaries

SUMMARY LENGTH	METHOD	P10	P30	AveP
4	QT	474	361	259
	QL	468	379	265
	QM	466	366	261
	QTM	474	361	259
	QTL	468	376	259
	QLM	468	379	265
	QA	466	379	***270***
6	QT	502	380	273
	QL	486	381	***277***
	QM	488	380	264
	QTM	500	381	273
	QTL	500	378	274
	QLM	484	382	***277***
	QA	500	378	274
9	QT	484	363	267
	QL	468	371	269
	QM	466	360	262
	QTM	474	363	267
	QTL	468	375	270
	QLM	468	371	269
	QA	466	376	***271***

6.4 Different Number of assumed Relevant Documents for Term Set Selection and Ranking

Table 5: Retrieval Results using Enhanced Term Selection Criteria

SUMMARY LENGTH & METHOD		NO OF DOCUMENTS USED IN RANKING	P10	P30	AveP
6	T	5	446	364	237
		20	498	379	***268***
	L	5	444	364	247
		20	480	378	***264***
	M	5	416	340	241
		20	462	386	***266***
	QT	5	436	353	249
		20	502	380	***273***
	QL	5	444	358	251
		20	486	381	***277***
	QM	5	442	347	246
		20	488	380	***264***
	QTM	5	436	355	249
		20	500	381	***273***
	QTL	5	446	363	256
		20	500	378	***274***
	QLM	5	444	358	251
		20	484	382	***277***
	QA	5	446	364	256
		20	500	378	***274***

Table 5 shows a summary of results comparing performance for the standard and the enhanced term selection criteria for a number of the summarization schemes. All other parameters are the same as in the previous experiments. These results illustrate that this new criteria is important in the success of the summary-based query expansion techniques. This may be attributed at least in part to the fact that terms selected from the most relevant part of the document may not occur frequently in other documents used for feedback particularly if those documents are non-relevant. However using a larger set of documents for ranking is likely to improve the classification criteria for such terms.

6.5 Variation of Relative Weight of Summary Sentence Selection Components

After analysis of the results shown above, further investigation was conducted to explore the effect of using different combinations of weighted individual summary scores on retrieval performance. This was done by upweighting the score assigned to each sentence based on the method used. Thus the *sentence significance score* (SSS) (section 3.5) was modified as shown below.

$$SSS = aSS1 + bSS2 + cSS3 + dSS4$$

where *a,b,c* and *d* = any number from 0 to 3.

The experiment was carried out by varying the values of *a, b, c* or *d* to adjust the scores assigned to sentences that satisfy the criteria of the different summary methods. Table 6 shows the results of increasing the scores of both the title and the Luhn method (T2L2), the scores of both the Luhn and the location method only (L2M2), the scores of the title and the location method (T2M2) and using unweighted scores (ORG as shown in Table 4). For runs T2L2, T2M2, L2M2 presented below the value 2 was used for either a, b, c, or d, depending on the scores to be increased.

Table 6: Effect of weighted scores

METHOD	RUN-ID			
	ORG	T2L2	L2M2	T2M2
ALL(A)	266	266	266	265
QLM	277	273	273	274
QTL	274	272	272	272
QTM	273	270	269	270
QA	274	275	274	275

The results clearly show that weighted summary scores do not have any substantial effect on average precision. However, the results show very clearly the effect of query-biased summaries on retrieval effectiveness. Combination of the three context-independent summary methods with query-bias (QA) consistently gave improvement over using only the three context-independent summary methods without query-bias (A). This support earlier results in Table 4 as shown in column ORG.

6.6 Results Summary

The results show improvement in retrieval performance using document summaries for term selection of up to 15% compared to the baseline search without feedback. Further, the use of document- summary expansion produces results up to 11% better than using standard whole document term selection. Best results using query-biased summaries are better than those for standard summaries, but overall there is little difference between them suggesting that use of document summaries may be an effective tool in interactive retrieval where it is not possible to form a query-biased summary.

Although on average there is no substantial improvement in retrieval performance for query expansion using terms selected from query-biased summaries compared to using context-independent summaries, individual cases show consistent improvement using query-biased summaries.

Retrieval performance improvement using terms selected from query-biased summaries also seems to be dependent on the summary length. The results (Table 4) show that retrieval performance improvement reaches its peak at summary length 6. Increasing the length of the summary to 9 resulted in a reduction in average precision. This suggests that more sentences which are unrelated to the initial query are finding their way into the resultant summary and thus contributing to the pool of possible expansion terms. Earlier similar experiments [23][24] show the same trend.

The enhanced term selection method was also shown to be consistently effective. On average it gave an improvement of around 10% over the standard selection method for term selection from document summaries. This means that using a larger number of documents than used for feedback in ranking terms during selection gives more discriminatory power to the term selection value.

Pseudo relevance feedback in general uses documents which cannot be guaranteed to be relevant. This situation can thus introduce bias into the process of ranking possible expansion terms based on their distribution in the set of assumed relevant documents. Using a larger number of top-ranked document from the retrieved set for ranking as used in the enhanced term selection approach is however likely to present a better criteria to measure the effectiveness of those terms in relation to relevant documents.

Retrieval improvement was also discovered not to be dependent on the relevance of feedback documents. For example, baseline results for TREC-8 title queries 404, 405, 421, 424 and 437 showed that the first five documents used for feedback were all irrelevant to the submitted query, although selected terms from the query-biased summaries of these documents resulted in improved retrieval of up to 400% compared to the baseline results. Further investigation of the same set of queries showed degradation in retrieval performance using selected terms from whole documents for queries 404, 421, 424 and 437. This confirms that summaries of retrieved documents can be effective in reducing the query drift associated with "bad" expansion terms taken from non-relevant documents.

Further investigation of the results also showed that the performance of the query-biased version of the location method is poor compared to other methods. This perhaps can be attributed to the fact that the combination of the query-bias method with the location method causes a conflict between documents whose subheadings might not be totally related to the query terms. It also shows that the first few sentences of a document might not always convey the most relevant part of the content of a document. The combination of the query-bias and the Luhn method consistently shows improvement in retrieval performance as measured by the average precision. It seems likely that the occurrence of significant terms in sentences also containing the initial query terms improves the quality of the summary generated and increases the chances of choosing beneficial terms for feedback.

In summary, the following conclusions can be drawn from the above results:

- Query expansion using selected terms from the most relevant part of retrieved document has a marked impact on retrieval effectiveness.
- Query-biased summaries can also be very efficient in cases where no relevant documents are retrieved at high rank or used for feedback.
- Retrieval effectiveness is dependent on the length of the summary generated
- Selecting terms for query expansion is more effective than using all the terms.

7. CONCLUSIONS AND FURTHER WORK

This paper has reported an investigation into the use of document summarization for term-selection in pseudo-relevance feedback. Summarization has been shown to be effective in this application with query-biased summaries potentially slightly better than context-independent summaries. Further work is required to examine whether the query biasing score factor can be modified to show greater improvement. It would also be worth further investigating whether alternative method of component weighting of summary score combination could be used to improve retrieval effectiveness. The enhanced term selection procedure has been shown empirically to be effective in all cases, and further examination of the results need to be carried out to provide a clearer explanation of this behaviour.

As stated earlier, it is observed that using larger data sets for expansion-term selection tends to produce more effective PRF results. Thus, while the experiments reported are important since they show effective PRF where only the target document set itself is available for term-selection parameter estimation, they need to be extended to the full set of TREC data (Disks 1-5) to investigate the effectiveness of term selection from document summaries under these conditions, and give a more direct comparison with existing results in [15].

The results in this paper focus only on pseudo-relevance feedback. An important area of further work is to explore the application of document summaries in relevance feedback where relevance information is provided by users rather than assumed.

Another interesting area for further work would be the application of thesauri expansion methods to selected terms from the query-biased summaries.

Finally, the effect of using different techniques of summary generation to the work described in this paper needs to be

explored. Some of these summary generation techniques are described in detail in [25].

8. ACKNOWLEDGEMENTS

We are grateful to the anonymous referees for their very useful and detailed comments on this paper.

9. REFERENCES

1. J.J.Rocchio. "Relevance Feedback in Information Retrieval" In *The smart retrieval system*, G. Salton, pages 313-323, 1971, Prentice Hall, Inc.
2. G.Salton and C.Buckley. Improving retrieval performance by relevance feedback, Journal of the American Society for Information Science, 41, pages 288-297, 1990.
3. C.Buckley, G.Salton, J.Allan, A.Singhal. Automatic Query Expansion using SMART, In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 65-80, 1995, NIST.
4. S.E.Robertson, S.Walker, S.Jones, M. M.Hancock-Beaulieu and M.Gatford, Okapi at TREC-3, In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pp109-216, 1995, NIST.
5. S.E.Robertson, On term selection for query expansion, *Journal of Documentation*, 46, pages 359-364, 1990
6. T.Strzalkowski, J.Wang and B.Wise. Summarization-Based Query Expansion in Information Retrieval, In *Proceedings of the 17th COLING*, pages 1-21, Montreal, 1998.
7. D.Harman. Relevance Feedback and other Query Modification Techniques. In : *Information Retrieval, Data structures and algorithms*, eds: W.B.Frakes and R.Baeza-Yates, pages 231-263, 1992.
8. M. Mitra, A. Singhal and C. Buckley. Improving Automatic Query Expansion, In *Proceedings of the 21st ACM SIGIR*, pages 206-214, Melbourne, 1998, ACM.
9. M.Sanderson. Word Sense Disambiguation and Information Retrieval, In *Proceedings of the 17th ACM SIGIR*, pages 142-157, Dublin, 1994, ACM.
10. J.P.Callan, Passage-Level Evidence in Document Retrieval. In *Proceedings of the 17th ACM SIGIR*, pages 302-310, Dublin, 1994, ACM.
11. D.Knaus, E.Mittendorf, P.Schauble, P.Sheridan. Highlighting Relevant Passages for users of the Interactive SPIDER Retrieval System, In *Proceedings of the Fourth Text Retrieval Conference (TREC-4)*, pages 233-238, 1995, NIST.
12. A.Tombros and M.Sanderson. Advantages of Query Biased Summaries in Information Retrieval, In *Proceedings of the 21st ACM SIGIR*, pages 2-10. Melbourne, 1998, ACM.
13. C.D.Paice. Constructing Literature Abstracts by Computer: Techniques and Prospects, *Information processing and Management*, 26(1) pages 171-186, 1989.
14. S.E.Robertson, S.Walker and M.Beaulieu, Okapi at TREC-7: automatic ad hoc, filtering, VLS and interactive track, In Proceedings of the Seventh Text Retrieval Conference (TREC-7), pages 253-264, 1998, NIST.
15. S.E.Robertson and S.Walker, Okapi/Keenbow at TREC-8, In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 151-162, 1999, NIST
16. Anastasios Tombros. Reflecting user information needs through query biased summaries. *Thesis submitted towards the award of MSc in Advanced Information systems in the University of Glasgow*. September 1997.
17. H.P.Luhn. The Automatic Creation of Literature Abstracts, *IBM Journal of Research and Development*, 2(2), pages 159-165, 1958.
18. H.P.Edmundson. New Methods in Automatic Abstracting. *Journal of the ACM*, 16(2), pages 264-285, 1969, ACM.
19. D.E.Kieras. Thematic Processes in the Comprehension of Technical Prose. In B.K Briton and J. B. Black eds, *Understanding expository text: A theoretical and practical handbook for analyzing explanatory text*, pages 89-108, 1985, Lawrence Erlbaum Associates.
20. M.F.Porter, An algorithm for suffix stripping, *Program*, 14(3), pages 130-137, 1980
21. S.E.Robertson and S.Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th ACM SIGIR*, pages232-241, Dublin, ACM.
22. S.E.Robertson and K.Sparck Jones. Relevance Weighting of Search Terms, *Journal of the American Society for Information Science*, 27(3), pages 129-146, 1976.
23. J Xu and W. Bruce Croft. Query Expansion Using Local and Global Document Analysis, *In Proceedings of the 19th ACM SIGIR*, pages 4-10, Zurich, 1996, ACM.
24. J. Allan. Relevance Feedback with too much Data, *In Proceedings of the 18th ACM SIGIR*, pages 337-343, Seattle, 1995, ACM.
25. I. Mani and M. Maybury (editors). Advances in Automatic Text Summarization, MIT Press, 1999.
26. C.J. van Rijsbergen (editor). In *Information Retrieval*, second edition, Butterworths, 1979.