Exeter at CLEF 2001: Experiments with Machine Translation for Bilingual Retrieval

Gareth J. F. Jones Adenike M. Lam-Adesina

Department of Computer Science,
University of Exeter,
Exeter, EX4 4PT, U.K.
email: {G.J.F.Jones,A.M.Lam-Adesina}@ex.ac.uk

Abstract

The University of Exeter participated in the CLEF 2001 bilingual task. The main objectives of our experiments were to compare retrieval performance for different topic languages with similar easily available machine translation resources and to explore the application of new pseudo relevance feedback techniques recently developed at Exeter to Cross-Language Information Retrieval (CLIR). We also report recent experimental results from our investigations of the combination of results from alternative machine translation outputs; specifically we look at the use of data fusion of the output from individual retrieval runs and merging of alternative topic translations.

1 Introduction

The CLEF 2001 bilingual task is the first standardised evaluation to enable comparison of Cross-Language Information Retrieval (CLIR) behaviour for European and Asian language pairings. The primary objective of our participation in the bilingual task was to explore retrieval behaviour with the English language document collection with an many of the translated topic sets as possible. Our official submissions covered French, German, Chinese and Japanese topics. In this paper we also report results for English, Italian and Spanish topics for comparison. In order to compare results for different language pairings fairly we wished to use similar translation resources for each pair. To this end we chose to use commercially developed machine translation (MT) resources. For our official submissions we used the online babelfish translation system (available at `http://babelfish.altavista.com` based on *SYSTRAN*. In this paper we report comparative results for French, German, Italian and Spanish using *Globalink Power Translator Pro Version: 6.4*. There is an underlying assumption with this approach that the translation resources for each language pair have been subject to an amount of development which makes such a comparison fair. In addition, we report more recent results combining the outputs of these MT resources using data fusion and query combination. Our general approach was to use a topic translation strategy for CLIR. Topic statements were submitted to the selected MT system, the output collected and then applied to the information retrieval system.

Pseudo-relevance feedback has been shown to be effective in many retrieval applications including CLIR [1] [2]. We have recently conducted experimental work with the Okapi BM25 probabilistic retrieval model and a new pseudo relevance feedback query-expansion method using document summaries [3]. This work also investigated a novel approach to term-selection that separates the choice of relevant documents from the selection of a pool of potential expansion terms. These techniques were shown to be considerably more effective than using full-document expansion on the TREC-8 ad hoc retrieval task. The result was an improvement of around 15% in average precision on short topic statements compared to a baseline without feedback. Our CLEF 2001 submission investigated the application of this technique to CLIR.

The remainder of this paper is organised as follows: Section 2 reviews the information retrieval methods used, Section 3 outlines the features of our summarisation system, Section 4 describes our combination methods, Section 5 gives experimental results, and Section 6 concludes the paper.

2 Information Retrieval Approach

The experiments were carried out using the City University research distribution version of the Okapi system. The documents and search topics were processed to remove stop words from a list of around 260 words, suffix stripped using Porter stemming [4] and terms were further indexed using a small set of synonyms.

Document terms are weighted using the Okapi *combined weight* (cw), often known as BM25, originally developed in g[5] and further elaborated in [6]. The BM25 cw for a term is calculated as follows,

$$cw(i,j) = \frac{cfw(i) \times tf(i,j) \times (K1+1)}{K1 \times ((1-b) + (b \times ndl(j))) + tf(i,j)}$$

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

$$ndl(j) = \frac{dl(j)}{\text{Average } dl \text{ for all documents}},$$

where $dl(j)$ is the length of j. $K1$ 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 multitopic.

2.1 Relevance Feedback

Assuming that relevant documents are available within a collection, the main reason that they may not be retrieved is the query-document match problem. Short and imprecise often results in relevant documents being retrieved at low rank or not being retrieved at all, and retrieval of non-relevant documents at high rank. Relevance feedback (RF) using query expansion is one method which seeks to overcome the query-document match problem. Pseudo-relevance feedback (PRF) methods in which a number of topic ranked documents are assumed to be relevant are on average found to give improvement in retrieval performance; although this is usually smaller than that observed for true RF. Post-translation PRF has been shown to be effective in CLIR in various studies including [2] [1].

The main implementational issue for PRF is the selection of appropriate expansion terms. 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. 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 can arise 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 [7] [8] has the advantage of being able to provide an overview of the distribution of the relevant pieces of information within the retrieved documents. However, this method has not been found to provide significant improvement in retrieval performance. We have developed 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 [9] [10] demonstrated 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. In [11] Tombros showed that query-biased summaries are more effective than using simple leading sentence summaries for user relevance decisions. Thus in our summaries we also make use of query-biased summaries. A related approach to the one reported here is described by Strzalkowski in [12] where a RF procedure using summaries of retrieved relevant documents is used. A weakness of

this approach is that all terms from the summaries were added to the query. We prefer to adopt the approach taken in the Okapi TREC submissions [13] [14] [15] which expand queries conservatively using only a small number of terms chosen using a statistical selection criteria [16].

The expansion terms were ranked using the Robertson selection value (rsv) [16], 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 [17]. $rw(i)$ 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 $n(i)$ is the total number of documents containing term i, R is the total number of relevant documents for this query, and N is the total number of documents.

The rsv has generally been based on taking an equal number of relevant documents for both the available expansion terms and term ranking. In our experiments we have explored the use of a more sophisticated approach which takes a smaller number of relevant documents to determine the pool of potential expansion terms than the number of documents used to determine the rsv ranking. It should be noted that the $r(i)$ value for term each i is calculated based on its occurrence in the entire document rather than on the summary alone.

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. Since we require a very robust summarizer for the different text types likely to be encountered within a retrieval system we adopt a summarisation method based on sentence extraction. Sentence extracted summaries are formed by scoring the sentences in the document using various criteria, ranking the sentences, and then taking a number of the top ranking sentences as the summary.

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 [18]. In order to determine the sentences of a document that should 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. Following the work of Tombros [11], 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 brackets ([—]), 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. In a sample study the title of TREC documents 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. 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 [19] 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.

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 follows,

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

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.

3.4 Query-Bias Method

The addition of a sentence score factor bias to score sentences containing query terms more highly can 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 is the same method used in [11].

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 summarisation system was implemented so that the relative weight of each component of SSS could be varied. 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 15document 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 Combination Methods

The combination of evidence from multiple information sources has been-shown to be useful for text retrieval in TREC [20]. In our experiments we examine two forms of index combination defined in [20]: *data fusion* and *query combination*.

4.0.1 Data Fusion

For data fusion the ranked document lists produced independently by topics translated using Babelfish and Power Translator Pro were combined by adding the corresponding query-document matching scores from the two lists and forming a new re-ranked list using the composite scores. We investigated both simple summing of the matching scores and summation of after the scores had separately been normalised with respect to the highest score in each list.

4.0.2 Query Combination

One of the important issues in query translation for CLIR is the choice of the best translation(s) of the search terms. The output of an individual MT system gives the best overall translation available for the input given its rules and dictionary. Thus different MT systems often given different translated outputs. In query combination the translated queries produced by the two MT systems were combined into a single representation to score against document archive. A set of combined queries was formed by taking the unique items from the existing query sets.

		Topic Language						
		English	French	German	Italian	Spanish	Chinese	Japanese
Prec.	5 docs	0.494	0.477	0.392	0.383	0.417	0.336	0.434
	10 docs	0.406	0.366	0.330	0.298	0.353	0.287	0.332
	15 docs	0.353	0.326	0.288	0.257	0.312	0.253	0.268
	20 docs	0.317	0.289	0.263	0.231	0.284	0.231	0.245
Av Precision		0.484	0.473	0.398	0.375	0.389	0.341	0.411
% change CLIR		—	-2.3%	-17.8%	-22.5%	-19.6%	-29.5%	-15.1%

Table 1: Baseline retrieval results for topic translation using Babelfish.

		Topic Language				
		English	French	German	Italian	Spanish
Prec.	5 docs	0.494	0.438	0.438	0.472	0.464
	10 docs	0.406	0.368	0.349	0.364	0.383
	15 docs	0.353	0.332	0.302	0.321	0.340
	20 docs	0.317	0.296	0.271	0.288	0.303
Av Precision		0.484	0.438	0.439	0.427	0.417
% change CLIR		—	-9.5%	-9.3%	-11.8%	-13.8%

Table 2: Baseline retrieval results for topic translation using Power Translator Pro.

5 Experimental Results

This section describes the establishment of the parameters of our experimental system and gives results from our CLEF 2001 investigation. We report procedures for the selection of system parameters, baseline retrieval results for different language pairs and translation systems without application of feedback, corresponding results with use of feedback, and results for our data combination experiments. In all cases the results use mandatory Title and Description fields from the search topics.

5.1 Selection of System Parameters

Various parameters had to be selected for our experimental system. In order to do this with carried out a series of development runs using the CLEF 2000 bilingual test collection. This data consisted of the English document set and topic sets in French, German, Italian and Spanish.

The Okapi parameters were set as follows: $K1 = 1.0$ and $b = 0.5$. In the pseudo relevance feedback runs 5 documents were assumed to be relevant in each case for term selection, document summaries comprised the best scoring 4 sentences in each case. Following experimentation with the sentence scoring components it was found that the best retrieval results were achieved when the Luhn and Title where given twice the relative weight compared to the Location and Query-Bias methods. The top 20 ranked expansion terms taken from these summaries were added to the original topic in each case. The rsv values to rank the potential expansion terms were selected by assuming the top 20 ranked documents were relevant. The original topic terms are upweighted by a factor of 3.5 relative to terms introduced by pseudo relevance feedback.

5.2 Baseline Results

Tables 1 and 2 show baseline retrieval results for topic translation using Babelfish and Power Translator Pro respectively. From these tables can be seen that the two MT systems give comparable performance for the European languages, although the result for French with Babelfish is particularly good. The worst overall result is achieved for the Chinese topics, although performance for Japanese is similar to that for the European languages.

		Topic Language						
		English	**French***	**German***	Italian	Spanish	**Chinese***	**Japanese***
Prec.	5 docs	0.498	0.477	0.421	0.400	0.477	0.357	0.426
	10 docs	0.400	0.366	0.336	0.320	0.394	0.296	0.362
	15 docs	0.362	0.326	0.301	0.286	0.342	0.258	0.305
	20 docs	0.329	0.289	0.275	0.252	0.299	0.224	0.266
Av Precision		0.517	0.489	0.415	0.395	0.423	0.336	0.431
% change no FB.		+6.8%	+3.4%	+4.3%	+5.3%	+8.7%	-1.5%	+4.9%
% change CLIR		—	-5.4%	-19.7%	-23.6%	-18.1%	-35.0%	-16.6%

Table 3: Retrieval results for topic translation using Babelfish with summary-based expansion term selection. * indicates official CLEF 2001 submitted run.

		Topic Language				
		English	French	German	Italian	Spanish
Prec.	5 docs	0.498	0.464	0.472	0.481	0.481
	10 docs	0.400	0.402	0.381	0.396	0.411
	15 docs	0.362	0.346	0.318	0.233	0.355
	20 docs	0.329	0.316	0.284	0.295	0.313
Av Precision		0.517	0.466	0.456	0.432	0.419
% change no FB		+6.8%	+6.4%	+3.9%	+1.2%	+0.5%
% change CLIR		—	-9.9%	-11.8%	-16.4%	-18.9%

Table 4: Retrieval results for topic translation using Power Translator Pro with summary-based expansion term selection.

5.3 Feedback Results

Tables 3 and 4 show retrieval results after the application of our summary based pseudo relevance feedback method. The results for French, German, Chinese and Japanese in Table 3 are our official submissions for the CLEF 2001 bilingual task. It can be seen that in all but case feedback improves the average precision. The one exception is for Chinese where performance is marginally worse after feedback. The reasons for this decrease in average precision have not yet been investigated. The average improvement in performance is around 5%. This is somewhat less than the 15% improvement that we observed in our previous experiments with the TREC-8 ad hoc task. Further investigation is needed to establish the reasons for this. One reason may be that in our previous work we worked only with the topic Title fields, meaning that there is often significant room for improvement in retrieval performance for individual topics by adding additional terms to the request. In the case of the CLEF runs here we are using both the Title and Description fields meaning that there may be less room for improvement from adding additional terms. Further experimental will be carried out to explore this possibility.

5.4 Combination Results

5.4.1 Data Fusion

Baseline Results Tables 5 and 6 show baseline results for Data Fusion with simple and normalised score addition respectively. Results using normalised scores are worse than the simple addition method, which overall generally gives a small improvement in performance compared to either translation on in isolation. The result for German is particularly good, giving the same value as the English baseline. This result is unusually good for CLIR, but appears from investigation to be correct.

Feedback Results Tables 7 and 8 show results for Data Fusion with summary based feedback applied. Feedback again generally results in an improvement in average precision, except for the case of German, where as noted previous the baseline Data Fusion was unusually good. Simple score addition is still superior to addition of normalised scores.

		Topic Language				
		English	French	German	Italian	Spanish
Prec.	5 docs	0.494	0.477	0.494	0.485	0.464
	10 docs	0.406	0.385	0.394	0.387	0.385
	15 docs	0.353	0.342	0.352	0.333	0.342
	20 docs	0.317	0.303	0.305	0.296	0.301
Av Precision		0.484	0.479	0.484	0.426	0.423
% change		—	-1.0%	-0.0%	-12.0%	-12.6%

Table 5: Baseline retrieval results for Data Fusion.

		Topic Language				
		English	French	German	Italian	Spanish
Prec.	5 docs	0.494	0.477	0.481	0.481	0.451
	10 docs	0.406	0.379	0.379	0.377	0.368
	15 docs	0.353	0.333	0.326	0.322	0.326
	20 docs	0.317	0.296	0.287	0.283	0.289
Av Precision		0.484	0.463	0.467	0.417	0.420
% change		—	-4.3%	-3.5%	-13.8%	-13.2%

Table 6: Baseline retrieval results for Data Fusion with score normalisation.

5.4.2 Query Combination

Tables 9 and 10 show baseline and feedback retrieval results respectively for Query Combination. It can be seen that the performance of Query Combination compared to the individual translation and the Data Fusion methods is rather unpredictable. Based on the results here Data Fusion appears to be a better strategy for combining information from multiple topic translations.

6 Concluding Remarks and Further Work

This paper has presented our results for the CLEF 2001 bilingual English language retrieval task. The results indicate that similar retrieval results are achieved using different commercial machine translation systems, but that some improvement can often be gained from applied Data Fusion methods to the output from the retrieval systems for different topic translations. Results for six different query-document pairs indicate that similar performance can be achieved for CLIR for Asian and European language topics for retrieval of English document despite the greater difference between the languages in the former case. However, the result for Chinese topics is the worst, and further investigation is required to better understand the reason for this. In addition, we intend to do query by query analysis of retrieval performance across the different languages pairs to investigate the effect of individual translation effects on retrieval behaviour. The application of our summary based pseudo relevance feedback method was generally shown to be effective, although the improvement was generally less than hoped for. This result will also be the subject of further investigation.

References

[1] G. J. F. Jones, T. Sakai, N. H. Collier, A. Kumano, and K. Sumita. A Comparison of Query Translation Methods for English-Japanese Cross-Language Information Retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 269–270, San Fransisco, 1999. ACM.

[2] L. Ballesteros and W. B. Croft. Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 84–91, Philadelphia, 1997. ACM.

		Topic Language				
		English	French	German	Italian	Spanish
Prec.	5 docs	0.498	0.489	0.502	0.481	0.506
	10 docs	0.400	0.415	0.396	0.400	0.413
	15 docs	0.362	0.352	0.338	0.335	0.345
	20 docs	0.329	0.322	0.313	0.305	0.303
Av Precision		0.517	0.489	0.476	0.451	0.426
% change		—	-5.4%	-7.9%	-12.8%	-17.6%

Table 7: Retrieval results for Data Fusion with summary-based expansion term selection.

		Topic Language				
		English	French	German	Italian	Spanish
Prec.	5 docs	0.498	0.489	0.494	0.472	0.485
	10 docs	0.400	0.411	0.377	0.389	0.400
	15 docs	0.362	0.343	0.316	0.329	0.338
	20 docs	0.329	0.317	0.292	0.292	0.297
Av Precision		0.517	0.489	0.463	0.436	0.426
% change		—	-5.4%	-10.4%	-15.7%	-17.6%

Table 8: Retrieval results for Data Fusion with score normalisation with summary-based expansion term selection.

[3] A. M. Lam-Adesina and G. J. F. Jones. Applying Summarization Techniques for Term Selection in Relevance Feedback. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, 2001. ACM.

[4] M. F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980.

[5] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, Dublin, 1994. ACM.

[6] S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In D. K. Harman, editor, *Overview of the Fourth Text REtrieval Conference (TREC-4)*, pages 73–96. NIST, 1996.

[7] J. P. Callan. Passage-Level Evidence in Document Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 302–310, Dublin, 1994. ACM.

[8] D. Knaus, E. Mittendorf, P. Schauble, and P. Sheridan. Highlighting Relevant Passages for users of the Interactive SPIDER Retrieval System. In *Proceedings of the Fourth Text REetrieval Conference (TREC-4)*, pages 233–238. NIST, 1996.

[9] J. Xu and W. B. Croft. Query Expansion Using Local and Global Document Analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–10, Zurich, 1996. ACM.

[10] J. Allan. Relevance Feedback with too much Data. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 337–343, Seattle, 1995. ACM.

[11] A. Tombros and M. Sanderson. The advantages of query-biased summaries in Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–10, Melbourne, 1998. ACM.

		Topic Language				
		English	French	German	Italian	Spanish
Prec.	5 docs	0.494	0.464	0.447	0.472	0.451
	10 docs	0.406	0.372	0.375	0.372	0.375
	15 docs	0.353	0.323	0.331	0.331	0.339
	20 docs	0.317	0.290	0.301	0.294	0.303
Av Precision		0.484	0.455	0.457	0.440	0.403
% change		—	-6.0%	-5.6%	-9.1%	-16.7%

Table 9: Baseline retrieval results for Query Combination.

		Topic Language				
		English	French	German	Italian	Spanish
Prec.	5 docs	0.498	0.543	0.447	0.472	0.477
	10 docs	0.400	0.396	0.396	0.385	0.409
	15 docs	0.362	0.335	0.335	0.338	0.346
	20 docs	0.329	0.304	0.304	0.304	0.318
Av Precision		0.517	0.482	0.407	0.469	0.421
% change		—	-6.8%	-21.3%	-9.3%	-18.6%

Table 10: Retrieval results for Query Combination with summary-based expansion term selection.

[12] T. Strzalkowski, J. Wang, and B. Wise. Summarization-based query expansion in information retrieval. In *Proceedings of 17th COLING-ACL'98*, pages 1–21, Montreal, 1998. ACL.

[13] S. E. Robertson, S. Walker, S. Jones, M. M. Hancok-Beaulieu, and M. Gatford. Okapi at TREC-3. In D. K. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3*, pages 109–126. NIST, 1995.

[14] S. E. Robertson, S. Walker, and M. M. Beaulieu. Okapi at TREC-7: automatic ad hoc, filtering, vls and interactive track. In E. Voorhees and D. K. Harman, editors, *Overview of the Seventh Text REtrieval Conference (TREC-7)*, pages 253–264. NIST, 1999.

[15] S. E. Robertson and S. Walker. Okapi/Keenbow. In E. Voorhees and D. K. Harman, editors, *Overview of the Eighth Text REtrieval Conference (TREC-8)*, pages 151–162. NIST, 2000.

[16] S. E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46:359–364, 1990.

[17] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[18] H. P. Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[19] H. P. Edmundson. New Methods in Automatic Astracting. *Journal of the ACM*, 16(2):264–285, 1969.

[20] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Information Processing and Management*, 31:431–448, 1995.