Exeter at CLEF 2003: Experiments with Machine Translation for Monolingual, Bilingual and Multilingual Retrieval

Adenike M. Lam-Adesina and Gareth J. F.Jones*
Department of Computer Science
University of Exeter EX4 4QF
United Kingdom
{A.M.Lam-Adesina,G.J.F.Jones}@exeter.ac.uk

Abstract. The University of Exeter group participated in the monolingual, bilingual and multilingual-4 retrieval tasks this year. The main focus of our investigation this year was the small multilingual task comprising four languages, French, German, Spanish and English. We adopted a document translation strategy and tested four merging techniques to combine results from the separate document collections, as well as a merged collection strategy. For both the monolingual and bilingual tasks we explored the use of a parallel collection for query expansion and term weighting, and also experimented with extending synonym information to conflate British and American English word spellings.

1 Introduction

This paper describes our experiments for CLEF 2003. This year we participated in the monolingual, bilingual and multilingual-4 retrieval tasks. The main focus of our participation this year was the multilingual-4 task (being our first participation in this task), our submissions for the other two tasks build directly on our work from our past CLEF experiments [5][6]. Our official submissions included monolingual runs for Italian, German, French and Spanish, bilingual German to Italian and Italian to Spanish, and the multilingual-4 task comprising English, French, German and Spanish collections.

Our general approach was to use translation of both document collections and search topics into a common language. Thus the document collections were translated into English using Systran Version:3.0 Machine Translator (Sys MT), and all topics translated into English using either Systran Version:3.0 or Globalink Power Translation Pro Version 6.4 (Pro MT) machine translation (MT) systems.

Following from our successful use of Pseudo-Relevance Feedback methods in past CLEF exercises [5][6] and supported by past research work in text retrieval exercises [1][2][3], we continued to use this method with success for improved retrieval. In our previous experimental work [4][5] we demonstrated the effectiveness of a new PRF method of term selection from document summaries, and found it to be more reliable than query expansion from full documents, this method is again used in the results reported here.

* now at School of Computing, Dublin City University, Ireland
email: Gareth.Jones@computing.dcu.ie

Following from last year, we again investigated the effectiveness of query expansion and term weight estimation from a parallel (pilot) collection [7], and found that caution needs to be exercised when using the collections to achieve improved retrieval performance for translated documents.

The remainder of this paper is structured as follows: in Section 2 we present our system setup and the information retrieval methods used, Section 3 describes the pilot search strategy, Section 4 presents and discusses experimental results and Section 5 concludes the paper with a discussion of our findings

2 System Setup

The basis of our experimental system is the City University research distribution version of the Okapi system. The documents and search topics were processed to remove stopwords from a list of about 260 words; suffix stripped using the Okapi implementation of Porter stemming [8] and terms were indexed using a small set of synonyms. Since the English document collection for CLEF 2003 incorporates both British and American documents, the synonym table was expanded this year to include some common British words that have different American spelling.

2.1 Term Weighting

Document terms were weighted using the Okapi BM25 weighting scheme developed in [9] and further elaborated in [10] and calculated as follows,

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

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) = dl(j)/avdl* where *dl(j)* is the length of *j and avdl* 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. In our experiments values of k1 and b are estimated based on the CLEF 2002 data.

2.2 Pseudo-Relevance Feedback

Retrieval of relevant documents is usually affected by short or imprecise queries. Relevance Feedback (RF) via query expansion aims to improve initial query statements by addition of terms from user assessed relevant documents. Expansion terms are selected using document statistics and aim to describe the information request better. Pseudo-Relevance Feedback (PRF) whereby relevant documents are assumed and used for query expansion is on average found to give improvement in retrieval performance although this is usually smaller than that observed for true user based RF.

The main implementation issue for PRF is the selection of appropriate expansion terms. In PRF problems can arise if assumed relevant documents are actually non-relevant thus leading to selection of inappropriate terms. However, the selection of such documents might suggest partial relevance, thus term selection from a relevant section or at least a related one might prove more beneficial.

Our query expansion method selects terms from summaries of the top 5 ranked documents. The summaries are generated using the method described in [4]. The summary generation method combines Luhn's keyword cluster method [11], a title terms frequency method [4], a location/header method [12] and the query-bias method from [13] to form an overall significance score for each sentence. For all our experiments we used the top 6 ranked sentences as the summary of each document. From this summary we collected all non-stopwords and ranked them using a slightly modified version of the Robertson selection value (rsv) [14] reproduced below. The top 20 terms were then selected in all our experiments.

$$rsv(i) = r(i) \times rw(i) \tag{2}$$

where r(i) = number of relevant documents containing term i
rw(i) = the standard Robertson/Sparck Jones relevance weight [14] reproduced below

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

where n(i) = the total number of documents containing term i
r(i) = the total number of relevant documents term i occurs in
R = the total number of relevant documents for this query
N = the total number of documents

In our modified version, potential expansion terms are selected from the summaries of the top 5 ranked documents, and ranked using statistics from assuming that the top 20 ranked documents from the initial run are relevant.

3 Pilot Searching

Query expansion is aimed at improving initial search topics in order to make them a better expression of the user's information need. This is normally achieved by adding terms selected from relevant or assumed relevant documents retrieved from the test collection to the initial query. However, it has been shown [15] that if additional documents are available these can be used in a pilot set for improved selection of expansion terms. The underlying assumption in this method is that a bigger collection than the test collection can help to achieve better term expansion and/or more accurate parameter estimation, and hopefully better retrieval and document ranking. Based on this assumption we explore the idea of pilot searching in our CLEF experiments.

The Okapi submissions for the TREC-7 [7] and TREC-8 [15] ad hoc tasks used the TREC disks 1-5, of which the document test set is a subset, for parameter estimation and query expansion. The method was found to be very effective. In order to explore the utility of pilot searching for our experiments, we used the TREC-7 and TREC-8 ad hoc document test collection itself for our pilot runs. This collection was used on its own for pilot searching without combination with the current CLEF test collections. The TREC and CLEF document collections are taken from the same time period, and the U.S. English CLEF 2003 English documents also appear within the TREC collection. The pilot searching procedure is carried out as follows:

1. Run the unexpanded initial query on the pilot collection using BM25 without feedback.
2. Extract terms from the summaries of the top R assumed relevant documents.
3. Select top ranked terms using (2) based on their distribution in the pilot collection.
4. Add desired number of selected terms to initial query.
5. Store equivalent pilot cfw(i) of search terms.
6. Either apply expanded query to the test collection and estimate term weights based on test collection, or
   apply expanded query with term weights estimated from pilot collection from the test collection.

4 Experimental results

This section describes the establishment of the parameters of our experimental system and gives results from our investigations for the CLEF 2003 monolingual, bilingual and multilingual-4 tasks. We report procedures for system parameters selection, baseline retrieval results for all languages and translation systems without the application of feedback, and corresponding results after the application of different methods of feedback including results for term weight estimation from pilot collections. The CLEF 2003 topics consist of three fields: Title, Description and Narrative. All our experiments use the Title and the Description fields of the topics. For all runs we present the average precision results (Av.P), the % change from results for baseline no feedback runs (% chg.), and the number of relevant documents retrieved out of the total number of relevant in collection (Rel-Ret).

4.1 Selection of System Parameters

To set appropriate parameters for our runs development runs were carried out using the CLEF 2002 collections. For CLEF 2003 more documents were added to all individual collections, and thus we are assuming that these parameters are suitable for these modified collections as well. The Okapi parameters were set as follows k1=1.4 b=0.6. For all our PRF runs, 5 documents were assumed relevant for term selection and document summaries comprised the best scoring 6 sentences in each case. Where the length of sentence was less than 6, half of the total number of sentences was chosen. The rsv values to rank the potential expansion terms were estimated based on the top 20 ranked assumed relevant documents. The top 20 ranked expansion terms taken from these summaries were added to the original query in each case. Based on results from our previous experiments, the original topic terms are upweighted by a factor of 3.5 relative to terms introduced by PRF. Since the English document collection for CLEF 2003 includes documents taken from both American and Britich English sources in our development runs we experimented with updated synonym information to conflate British and American English word spellings. This method resulted in a further 4% improvement in average precision compared to the baseline no feedback results for our English monolingual unofficial run for CLEF 2002[1]. We anticipate this being a useful technique for CLEF 2003 as well, and the updated synonym list is again used for all our experiments reported here.

In the following tables of results the following labeling conventions are adopted for the selection of topic expansion terms and cfw(i) of the test collection:

TCow(i): topic expansion using only the test collection.
PCow(i): topic expansion using the TREC document pilot collection.
CCow(i): topic expansion using the combined multilingual-4 collection.
TopCow(i): topic expansion using a translated collection in the topic language.
TCcfw(i): cfw(i) values taken from the test collection in the final retrieval run.
PCcfw(i): cfw(i) values taken from the TREC document pilot collection in the final retrieval run.
CCcfw(i): cfw(i) values taken from the combined multilingual-4 collection in the final retrieval run.
TopCcfw(i): cfw(i) values taken from a translated collection in the topic language.

4.2 Monolingual runs

We submitted runs for four languages (German, French, Italian and Spanish) in the monolingual task. Official runs are marked with a * and additional unofficial runs are presented for all languages. In this section we include results for the native English document collection as well for comparison. In all cases results are presented for the following:

1. Baseline run without feedback.
2. Feedback runs using expanded query and term weights from the test collection.
3. Feedback runs using expanded query from pilot collection and term weights from test collection.
4. Feedback runs using expanded query and term weights from pilot collection.

[1] Given that the CLEF 2002 English collection contains only American English documents, we found this improvement in performance from spelling conflation a little surprising for the CLEF 2002 task, and we intend to carry our further investigation into the specific sources of this improvement in performance.

5. An additional Feedback run is presented where query is expanded using a pilot run on a merged collection of all four text collection comprising the small multilingual collections. with the terms weights being taken from the test collection.
6. As 5, but with the term weights taken from the combined small multilingual pilot collection.

Results are presented for both Sys and Pro MT systems

4.2.1 German Monolingual runs

Table 1. Retrieval results for topic translation for German monolingual runs for both Sys MT and Pro MT topic translation

	Sys MT			Pro MT		
Run-ID	Av.P	% chg.	Rel_Ret	Av.P	% chg.	Rel_Ret
1. Baseline	0.488	-	1706	0.441	-	1580
2. TCow(i), TCcfw(i)	**0.568***	+16.4%	1747	0.511*	+15.9%	1657
3. PCow(i), TCcfw(i)	0.512*	+4.9%	1727	0.457	+3.6%	1616
4. PCow(i), PCcfw(i)	0.458	-6.1%	1665	0.431	-2.3%	1575
5. CCow(i), TCcfw(i)	0.550	+12.7%	1751	0.494	+12.0%	1663
6. CCow(i), CCcfw(i)	0.551	+12.9%	1750	0.512	+16.1%	1672

4.2.2 French Monolingual runs

Table 2. Retrieval results for topic translation for French monolingual runs for both Sys MT and Pro MT topic translation

	Sys MT			Pro MT		
Run-ID	Av.P.	% chg.	Rel_Ret	Av.P	% chg.	Rel_Ret
1. Baseline	0.487	-	918	0.422	-	885
2. TCow(i), TCcfw(i)	0.521*	+6.9%	933	0.457*	+8.3%	897
3. PCow(i), TCcfw(i)	0.491*	+0.8%	921	0.403	-4.5%	890
4. PCow(i), PCcfw(i)	0.489	+0.4%	920	0.426	+0.9%	885
5. CCow(i), TCcfw(i)	0.519	+6.6%	931	0.446	+5.7%	893
6. CCow(i), CCcfw(i)	**0.553**	+13.6%	931	0.467	+10.7%	891

Examination of Tables 1 to 4 reveals a number of consistent trends. Considering first the baseline runs. In all cases Sys MT translation of the topics produces better results than use of Pro MT. This is not too surprising since the documents were also translated with Sys MT, and the result indicates that consistency (and perhaps quality) of translation is important. All results show that our PRF method results in improvement in performance over the baseline in cases.

The variations in PRF results for query expansion for the different methods explored are very consistent. The best performance is observed in all cases, except Pro MT Spanish, using only the test collection for expansion term selection and collection weighting. Thus although query expansion from pilot collections has been

4.2.3 Italian Monolingual runs

Table 3. Retrieval results for topic translation for Italian monolingual runs for both Sys MT and Pro MT topic translation

	Sys MT			Pro MT		
Run-ID	Av.P	% chg.	Rel_Ret	Av.P	% chg.	Rel_Ret
1. Baseline	0.419	-	761	0.387	-	742
2. TCow(i), TCcfw(i*)*	**0.494***	+17.9%	787	0.449*	+16.0%	759
3. PCow(i), TCcfw(i)	0.432*	+3.1%	762	0.402	+3.89%	745
4. PCow(i), PCcfw(i)	0.393	-6.2%	754	0.387	0%	735
5. CCow(i), TCcfw(i)	0.456	+8.8%	771	0.452	+16.8%	759
6. CCow(i), CCcfw(i)	0.454	+8.4%	770	0.481	+24.3%	761

4.2.4 Spanish Monolingual runs

Table 4. Retrieval results for topic translation for Spanish monolingual runs for both Sys MT and Pro MT topic translation

	Sys MT			Pro MT		
Run-ID	Av.P	% chg.	Rel_Ret	Av.P	% chg.	Rel_Ret
1. Baseline	0.422	-	2163	0.393	-	2111
2. TCow(i), TCcfw(i*)*	**0.470***	+11.3%	2195	0.452*	+15.0%	2145
3. PCow(i), TCcfw(i)	0.426*	+0.9%	2114	0.415	+5.6%	2081
4. PCow(i), PCcfw(i)	0.372	-11.8%	1973	0.397	+1.0%	2039
5. CCow(i), TCcfw(i)	0.462	+9.5%	2200	0.466	+18.6%	2148
6. CCow(i), CCcfw(i)	**0.470**	+11.3%	2167	0.462	+17.6%	2142

shown to be very effective in other retrieval tasks [6], the method did not work very well for CLEF 2003 documents and topics. Perhaps more surprising is the observation that term weight estimation from the pilot collection actually resulted in average precision in most cases lower than that of the baseline no feedback run. This result is very unexpected particularly since the method has been shown to be very effective and has been used with success in our past research work for CLEF 2001 [5] and 2002 [6].

Query expansion from the merged document collection (used for the multilingual task) of Spanish, English, French, and German also resulted in improvement in retrieval performance, in general slightly less than that achieved in the best results for French, German and Spanish using only the test collection. The result for this method is lower for Italian run, this is probably arises due to the absence of the Italian document collection from the merged collection. The use of the combined collection cfw(i) has mixed impact on performance.

4.2.5 English Monolingual runs

Table 5. Retrieval results English monolingual runs

Run-ID	Av.P	% chg.	Rel_Ret
1. Baseline	0.456	-	982
2. TCow(i), TCcfw(*i)*	**0.483**	+5.9%	998
3. PCow(i), TCcfw(i)	0.425	-6.8%	994
4. PCow(i), PCcfw(i)	0.472	+3.5%	995
5. CCow(i), TCcfw(i)	0.477	+4.6%	992
6. CCow(i), CCcfw(i)	0.434	-4.8%	986

Table 5 shows the results for runs 1-6 for the native English document collection with native English topic statements. Again in this case the best performance is achieved using test collection expansion and term weighting. Expansion from the pilot collection is again unsuccessful with corresponding term weighting giving improved results for these expanded topic statements. Expansion from the combined collection is successful, but using the corresponding term weights degrades performance below the baseline. These results for the pilot collection are again surprising, particularly in the case of the use of the TREC document collection. The pilot collection and the test collection are both original English documents. Thus, based on previous results we might expect this to be more reliable than the earlier results for the translated documents in Tables 1-4.

4.2.6 Native English Topic Runs

Table 6. Baseline retrieval results for translated documents with native English topics

			Sys MT		Pro MT	
Original Document Language	Av.P	Rel_Ret	Av. P. % chg.	Rel_Ret chg.	Av.P. % chg.	Rel_Ret chg.
French	0.469	868	-3.7%	-17	+11.1%	-17
German	0.465	1619	-4.7%	-87	+5.4%	+39
Italian	0.400	751	-4.5%	-10	+3.4%	+9
Spanish	0.480	2045	+13.7%	-118	+22.1%	-66

Table 6 shows an additional set of baseline results for the different translated language collections with the untranslated native English language topic statements. These runs were carried out without any feedback or alternative test collection weights to explore the impact of topic translation without interfering effects from these additional techniques.

The results show that in general retrieval performance is best for Sys MT topics rather than for the original English topics. This is perhaps surprising since the original English statements will be more “accurate” readings of the topics in English, however the vocabulary match between the documents and topics into English using the same resources is more effective for retrieval. By contrast the original English topics perform better then the topic translations using Pro MT. Overall the trends here are consistent with our monolingual and bilingual retrieval results submitted to CLEF 2002 [6].

4.3 Bilingual runs

For the Bilingual task we submitted runs for both the German-Italian and Italian-Spanish tasks. Official runs are again marked with a * and additional unofficial runs are presented. In all cases, results are presented for the following experimental conditions:

7. Baseline run without feedback.
8. Feedback runs using expanded query and term weights from the target collection.
9. Feedback runs using expanded query from pilot collection and term weights from test collection.
10. Feedback runs using expanded query and term weights from pilot collection.
11. We investigated further the effectiveness of pilot collection and the impact of vocabulary differences for different languages. This is done by expanding initial query statement from the topic collection and then applying the expanded query on the target collection (i.e. for German-Italian bilingual runs initial German query statement is expanded from the German collection and applied on the test collection).
12. Additionally both the expanded query and the corresponding term weights are estimated from the topic collection.
13. The topics were expanded using method 11 and then further expanded using method 8. The term weights in the target language are estimated from the test collection.
14. As 13 with the term weights estimated from the topic collection.

Results are again presented for both Sys MT and Pro MT topic translations.

4.3.1 Bilingual German to Italian

Table 7. Retrieval results for topic translation for Italian bilingual runs for Sys MT and Pro MT

	Sys MT			Pro MT		
Run-ID	Av.P	% chg	Rel_Ret	Av.P	% chg	Rel_Ret
7. Baseline	0.311	-	725	0.314	-	668
8. TCow(i),TCcfw(i)	0.370	+18.9%	748	0.359	+14.3%	701
9. PCow(i),TCcfw(i)	0.339	+9.0%	724	0.334	+6.4%	671
10. PCow(i),PCcfw(i)	0.327	+5.1%	715	0.335*	+6.7%	659
11. TopCow(i),TCcfw(i)	0.365	+17.4%	743	0.355*	+13.1%	691
12. TopCow(i),TopCcfw(i)	0.415*	+33.4%	750	0.397*	+26.4%	702
13.TopC>TCow(i), TCcfw(i)	0.433	+39.2%	750	0.418	+33.1%	735
14 TopC->TCow(i), TopCcfw(i)	**0.441**	+41.8%	749	0.421	+34.1%	733

4.3.2 Bilingual Italian to Spanish

Table 8. Retrieval results for topic translation for Spanish bilingual runs for Sys MT and Pro MT

	Sys MT			Pro MT		
Run-ID	Av.P	% chg	Rel_Ret	Av.P	% chg	Rel_Ret
7. Baseline	0.327	-	1938	0.349	-	1923
8. TCow(i),TCcfw(i)	0.376	+14.9%	2042	**0.417**	+19.5%	2064
9. PCow(i),TCcfw(i)	0.331	+1.2%	1915	0.365	+4.6%	1940
10. PCow(i),PCcfw(i)	0.339	+3.7%	1870	0.364*	+4.3%	1872
11. TopCow(i),TCcfw(i)	0.389	+18.9%	2071	**0.417***	+19.5%	2011
12. TopCow(i),TopCcfw(i)	**0.391***	+19.6%	2051	0.385*	+10.3%	2004
13. TopC->TCow(i), TCcfw(i)	0.389	+19.0%	2064	0.379	+8.6%	1968
14.TopC->TCow(i), TopCcfw(i)	0.382	+16.8%	2059	0.367	+5.2%	1932

Tables 7 and 8 show results for our bilingual runs. For the bilingual runs topic expansion and weighting using the test collection is shown to be better than using the TREC pilot collection for both tasks. Query expansion and term weight estimation from pilot collection resulted in improvement in average precision ranging from 1.2% to 9% for both results, although it failed to achieve comparable performance to other methods, which is again surprising but consistent with the monolingual results.

For our bilingual runs we also tried a new method of query expansion and term weight estimation from the topic language collection. For this condition the topic was first applied on the translated test collection associated with the topic language, i.e. translated German topics were applied to the translated German documents. We experimented with cfw(i) values taken from the test collection and from the topic collection. Interestingly using the topic collection cfw(i) improves results, dramatically so in the case of the German to Italian task. For the German to Italian task this method resulted a +33% improvement in average precision over the baseline when using test collection cfw(i). It also worked well for the Spanish bilingual run giving about 19% improvement in average precision. The use of term weights from the topic collection gives a large improvement over the result using test collection weights in the case of the German-Italian task, but for the Italian-Spanish task this has a negligible effect in the case of Systran MT and makes performance worse for Globalink MT. It is not immediately clear why these collections should behave differently, but it may relate to the size of the document collections, the Italian collection being much smaller than either of the German or Spanish collections.

We also explored a further strategy of double topic expansion. The topic is first expanded using the topic collection and then further expanded using the test collection. For the German to Italian task the result is further improved, resulting in a +41.8% in average precision for the Sys MT topics when the topic collection weights were used. However, this strategy is not effective for the Italian to Spanish task, but it can be noted that, unlike the German to Italian task, using test collection cfw(i) is still more effective than using the topic collection cfw(i).

4.4 Multilingual Retrieval

Multilingual information retrieval presents a more challenging task in cross-lingual retrieval experiments. A user submits a request in a single language (e.g. English) in order to retrieve relevant documents in different languages e.g. English, Spanish, Italian, German, etc. We approached the multilingual-4 task in two ways. First, we retrieved relevant documents using the English topics individually from the four different collections and then explored merging the results together using different techniques (described below). Secondly we merged the translated document collections with the English collection to form a single collection and performed retrieval directly from this collection without using a separate merging stage.

Different techniques for merging separate result lists to form a single list have been proffered and tested. All of the techniques suggest that making assumptions that the distribution of relevant documents in the results set for retrieval from individual collection is similar is not true [16]. Hence, straight merging of relevant documents from the sources will result in poor combination.

Based on these assumptions we examined four merging techniques for combining the retrieved results from the four collections to form a single result list as follows:

$$u = \frac{doc\_wgt}{g\max\_wt * rank} \tag{3}$$

$$p = doc\_wgt \tag{4}$$

$$s = \frac{doc\_wgt}{g\max\_wt} \tag{5}$$

$$d = \frac{doc\_wgt - \min\_wt}{\max\_wt - \min\_wt} \tag{6}$$

where *u*, *p*, *s* and *d* are the new document weight for all document in all collections and corresponding results are labeled mult4* where * can be *u*, *p, s* or *d* depending on the merging scheme used. The variables in (3)-(6) are defined as follows:

doc_wgt = the initial document matching score
gmax_wt = the global maximum matching score i.e the highest document from all collections for a given query
max_wt = the individual collection maximum matching score for a given query
min_wt = the individual collection minimum matching score for a given query
rank = a parameter to control the effect of size of collection - a collection with more documents gets a higher rank (value ranges between 1.5 and 1).

To test the effectiveness of the merging schemes, we merged the 4 text collection into a single large combined collection. Expanded queries from this combined test collection (CCow(i),CCcfw(i)) and from the TREC data pilot collection (PCow(i),CCcfw(i)) were then applied on the resultant merged collection. For all official runs (mult4*) English queries are expanded from the TREC-7 and 8 pilot collections and then applied on the test collection.

Table 9. Retrieval results for small Multilingual task before and after applications of different merging strategies

Run-ID	Av.P	P10	P30	%chg.	Rel_Ret
Baseline	0.383	0.593	0.476	-	4613
PCow(i),CCcfw(i)	**0.438***	**0.623**	**0.524**	**+14.3%**	**4828**
CCow(i),CCcfw(i)	0.425	0.617	0.517	+10.9%	4853
PCow(i),TCcfw(i)mult4u	0.351*	0.520	0.434	-8.4%	4574
PCow(i)TCcfw(i)mult4p	0.356*	0.532	0.438	-7.0%	4457
PCow(i)TCcfw(i)mult4s	0.356*	0.518	0.438	-7.0%	4428
PCow(i)TCcfw(i)mult4d	0.331*	0.525	0.433	-13.5%	4609
CCow(i),TCcfw(i)mult4s	0.400	0.593	0.486	+4.4%	4675

An additional run CCow(i),CCcfw(i)mult4s was conducted whereby the expanded query was estimated from the merged document collection and applied on the individual collection before being merged using equation 5 above.

The baseline result for our multilingual run (Baseline) perhaps might not present a realistic platform for comparison with the feedback runs using the different merging strategies (PCow(i),TCcfw(i)mult4*). This is because it was achieved from a no feedback run from the merged multilingual collection.

The multilingual-4 results show that the different merging strategies for combining the retrieved lists from the separate collections provide similar retrieval performance. The result for merging strategy using (6) (which has been shown to be effective in past retrieval tasks) however resulted in about 14% loss in average precision compared to the baseline run. The more sophisticated merging strategies failed to show any improvement over raw score merging (4), although the merging strategy using (6), gave the highest number of relevant document retrieved for all the merging strategies.

Both our bilingual and monolingual runs show that retrieval results using query expansion and term weight estimation from pilot collection resulted in loss in average precision compared to baseline no feedback run in most cases. This might have contributed to the poor result from the different merging techniques for the multilingual runs (PCow(i),TCcfw(i)mult4*) where the expanded topic statement was calculated from the TREC pilot collection. For the multilingual results using the merging techniques (PCow(i),TCcfw(i)mult4*), we expanded the initial English queries and then applied these to the individual collections, the term weights were estimated from the individual test collections. However, results from our monolingual runs using this query expansion method were not very encouraging, and this might perhaps have contributed to the poor results after the application of the different merging techniques compared to the method whereby all the collections are merged to form one big collection.

To test this hypothesis, we conducted an additional run whereby we used the merged collection as the pilot collection and expanded the initial query from it (CCow(i),TCcfw(i)mult4s). The expanded topic was then applied on the individual collections and resultant result file merged using (5). The result showed an improvement of about 4% compared to that achieved from the baseline no feedback run from the merged collection (Baseline). It also resulted in about 11% increase in average precision over result from query expansion from the pilot collection (PCow(i),TCcfw(i)mult4s).

The best result for the multilingual task was achieved by expanding the initial query from the pilot collection and applying it on the merged collection. Query expansion from the merged collection (CCow(i),CCcfw(i)) also resulted in about 10% improvement in average precision. These results suggest that merging a collection in a multilingual task can be more beneficial than merging the result lists taken from the retrieval from individual collections. This result is presumably due to the more robust and consistent parameter estimation in the combined document collection. In many operational situations combining collections in this way will not be practical either due to the physical separation of the collections or the lack of opportunity to translate them into a common pivot language. From this perspective Multilingual IR can be viewed as distributed information retrieval task where there may be varying degrees of cooperation between the various collections. In this environment list merging is an essential component of the multilingual retrieval process. Results for the combined collection illustrate that better retrieval results than achieved using the currently proposed merging strategies is easily available using these documents, and further research is clearly required to develop distributed merging strategies that can approach combined collection retrieval performance.

5 Conclusions

For our participation in the CLEF 2003 retrieval tasks we updated our synonym information to include common British and American English words. We explored the idea of query expansion from pilot collection and got some disappointing results which are contrary to past retrieval work utilizing the use of expanded queries and term weight estimation from pilot collections. This result may be caused by vocabulary and term distribution mismatch between our translated test collection and the native English pilot collection, however this trend was also observed for the native English document collection, and further investigation is needed to ascertain whether this or other reasons underlie this negative result.

For the bilingual task we explored the idea of query expansion from a pilot collection in the topic language. This method resulted in better retrieval performance. Although we are working in English as our search language throughout, this result is related to the ideas of pre-translation and post-translation feedback explored in earlier work on CLIR [2], and the effectiveness of combining pre- and post-translation feedback appears to be related to the properties of the document collections.

The different merging strategies used for combining our results for the multilingual task failed to perform better than raw score merging. Further investigation is needed to test these methods, particularly as some of them have been shown to be effective in past research. Merging the document collections resulted in better average precision than merging separate retrieval result lists. However, it will often not be possible to merge the various collections together, in this case an effective method of merging the result list is needed. Further investigation will be conducted to examine the possibility of improving the results achieved from merging result lists.

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 Francisco, 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.
3. G.Salton and C. Buckley. Improving Retrieval performance by Relevance Feedback. Journal of the American Society for Information Science, pages 288-297, 1990.
4. 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, pages 1-9, New Orleans, 2001. ACM.
5. G.J.F. Jones and A.M. Lam-Adesina. Exeter at CLEF 2001: Experiments with Machine Translation for Bilingual Retrieval. In Proceedings of the CLEF 2001 Workshop on Cross-Language Information Retrieval and Evaluation, pages 59-77, Darmstadt, Germany, 2001.
6. A.M.Lam-Adesina and G.J.F.Jones. Exeter at CLEF 2002: Experiments with Machine Tranlsation for Monolingual and Bilingual Retrieval, Proceedings of the CLEF 2002 Workshop on Cross-Language Information Retrieval and Evaluation, Rome, Italy, 2002.
7. 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.
8. M.F. Porter. An algorithm for suffix stripping. Program, 14:10-137, 1980.
9. S.E Robertson, 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.
10. S.E Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu and M. Gatford, Okapi at TREC-3. In D.K. Harman, editor, Proceedings of the Third Text REtrieval Conference (TREC-3), pages 109-126. NIST, 1995.
11. H.P. Luhn. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 2(2):159-165, 1958.
12. H.P. Edmundson. New Methods in Automatic Abstracting. Journal of the ACM, 16(2):264-285, 1969
13. A. Tombros and M. Sanderson. The Advantages of Query-Biased Summaries in Information Retrieval. In proceedings of the 21st Annual International ACM SIGIR Conference Research and Development in Information Retrieval, pages 2-10, Melbourne, 1998. ACM.
14. S.E. Robertson. On term selection for query expansion. Journal of Documentation, 46:359-364, 1990.
15. S.E. Robertson, 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. Jacques Savoy. Report on CLEF-2002 Experiments: Combining Multiple Sources of Evidence. In Proceedings of the CLEF 2002: Workshop on Cross-Language Information Retrieval and Evaluation, pages 31-46, Rome Italy, September 2002.