Dublin City University at CLEF 2004: Experiments in Monolingual, Bilingual and Multilingual Retrieval

Gareth J. F. Jones, Michael Burke, John Judge, Anna Khasin,
Adenike Lam-Adesina and Joachim Wagner

School of Computing, Dublin City University, Dublin 9, Ireland
email: {gjones,mburke,jjudge,akhasin,adenike,jwagner}@computing.dcu.ie

Abstract. The Dublin City University group participated in the monolingual, bilingual and multilingual retrieval tasks. The main focus of our investigation for CLEF 2004 was extending our information retrieval system to document languages other than English, and completing the multilingual task comprising four languages: English, French, Russian and Finnish. Our retrieval system is based on the City University Okapi BM25 system with document preprocessing using the Snowball stemming software and stopword lists. Our French monolingual experiments compare retrieval using French documents and topics, and documents and topics translated into English. Our results indicate that working directly in French is more effective for retrieval than adopting document and topic translation. A breakdown of our multilingual retrieval results by the individual languages shows that similar overall average precision can be achieved when there is significant underlying variation in performance for individual languages.

1 Introduction

Dublin City University's (DCU) participation in the CLEF 2004 monolingual, bilingual and multilingual track builds on our existing work at the University of Exeter [1]. This previous work was limited to English language retrieval. For non-English retrieval, documents and topics were translated into English using machine translation. Thus English was used as a "pivot" language for all tasks. Retrieval was based on the City University distribution of the Okapi system augmented with a summary-based pseudo-relevance feedback system. Our work for CLEF 2004 concentrated on extending our retrieval system to work directly in the document language with topic translation when needed. Our strategy is to extend our existing Okapi based retrieval system to make use of the Snowball stemmers and stop word lists [2]. Using these tools we completed runs for monolingual French, Russian and Finnish documents, official bilingual runs for French and Russian, and the multilingual task consisting of English, French, Russian and Finnish, together with the additional monolingual and bilingual runs needed for the multilingual task.

This paper is organised as follows: Section 2 outlines the details of our retrieval system and describes its extension to non-English retrieval, Section 3 reports our experimental results, and finally Section 4 concludes the paper.

2 Retrieval System

2.1 Summary of Okapi System

The basis of our experimental retrieval system is the City University research distribution version of the Okapi system, as used in our previous CLEF participation [1]. The standard Okapi environment includes tools for English language preprocessing. These preprocessing tools, including stopword removal and stemming, are coded directly into the software and cannot be readily modified or replaced in the distributed software. A further limitation is that it can only handle ASCII English characters and punctuation symbols. In order to extend the system to other languages we moved the preprocessing outside Okapi itself and encode the text using English language characters, as described in the next section, prior to entering the data into Okapi. Search terms are weighted using the standard BM25 weighting scheme and we use our summary-based pseudo relevance feedback (PRF) method [4].

For English language runs we continued to use the standard Okapi system system. The documents and search topics are processed to remove stopwords from a list of about 260 words; suffix stripped using the Okapi implementation of Porter stemming [3], and terms are further indexed using a small set of synonyms.

2.2 Language Independent use of the Okapi System

By carrying out data preprocessing and then encoding the text into English language ASCII characters prior to entering the data into the Okapi system, it can be used as a language independent retrieval system. This section describes the preprocessing method we used for non-English documents for our CLEF 2004 experiments.

The documents and topics are prepared using a pipeline of pre-processing components. Firstly, the data is tokenised to isolate the text body from the SGML/XML markup tags. Then, all punctuation characters are deleted from the text body, with the following exceptions: full stops, commas, semi-colons, colons, exclamation marks and question marks. Whitespace is inserted to separate these punctuation characters from word tokens. The characters are then converted to lower case. Distinct mappings must be used for the character set of each language. The Russian characters were converted to KOI-8 character encoding as required by the Snowball tools, while the Finnish and French documents use ISO Latin 1. Conversion of the Russian data loses some data, for example the degree sign prevalent in weather forecasts is lost, further some corruption of the original data to "boxdrawing" symbols was observed. We made two different conversions: one that just replaces every character outside the KOI-8 set with whitespace, and one in which we tried to do optimal/most frequently correct substitutions.

At the next stage stop words are removed. The stop word lists provided by Snowball are used for French, Russian and Finnish. The Russian stopword list used here consists only of the simple first part of the Snowball list. The words are then passed to the Snowball stemmer. The only alteration to the default stemmer functionality is the conversion of the Russian character encoding from ISO to KOI-8. Finally, the whitespace preceding the maintained punctuation characters is removed[1].

Since the Okapi system does not accept the special characters outside English used in French, Russian and Finnish, all character strings in these languages were encoded using the 26 lowercase letters a to z. The encoding guarantees that different input words are discriminably represented and that the reverse operation (decoding) can be performed easily if required. The encoded form is not readable by humans and string similarities do not stay intact. However, neither of these is a problem, since no one will be reading the encoded documents, and fuzzy matching is not used in our query-document matching. For example, the three French words "pécheur", "pêcheur" and "pêcheurs" are encoded as *gropmdpbtfui*, *cbppmdpbtfui* and *klcgrwruwanejd*. Encoded strings are then passed into the Okapi system for indexing. When used in the this manner no stopword removal, stemming or other processing is performed within the Okapi system itself.

Topic statements are similarly processed to remove stopwords, apply stemming, and apply the character encoding, prior to being applied to the Okapi retrieval system.

3 Experimental Results

This section presents results and analysis of our experimental runs. Full details of the retrieval tasks are given in the track overview paper [5]. All runs use the Title and Description CLEF topic fields. For our experiments, we report precision at ranks 5,10, 15 and 20, average precision and total number of relevant documents retrieved.

System parameters were selected using CLEF 2003 test collections. In all cases Okapi parameters were set as follows: $k_1 = 1.0$ and $b = 0.75$. The summary generation method combines Luhn's keyword cluster method, a title terms frequency method, a location/header method and a query-bias method to form an overall significance score for each sentence. For PRF we explored four sentence selection criteria for document summary generation as follows: L = Luhn method, T = title method, Q = query-bias method, and A = linear sum of all methods. The L, T and Q methods in each case use only this single measure of sentence significance. The 20 top ranked PRF expansion terms were selected from the summaries of the top 5 ranked documents, with the top 20 ranked documents used to rank potential expansion terms for selection, unless otherwise specified for individual tasks. The original topic terms were upweighted by a fac-

[1] The punctuation symbols must be maintained in the document to facilitate summarization for PRF.

tor of 3.5 relative to terms introduced by PRF. Full details of the summary-based PRF method are given in [4].

3.1 Monolingual Retrieval

This section presents results for our monolingual retrieval experiments. Official runs were carried out for French, Russian and Finnish document collections. Monolingual English document results are also included here for use in comparative analysis of the multilingual retrieval results later in this section.

French Runs Table 1 shows results for French monolingual retrieval. Separate results are shown for documents and topics in French, and documents and topics translated into English using Systran MT. For French language retrieval experiments, the PRF summary length was set to 4 sentences, and for translated documents and topics to 6 sentences. It can be seen that working in French produces superior retrieval performance with respect to both precision and recall metrics. This document and topic translation approach was used in our previous work [1]. The result here indicates that extending our retrieval system to the document language is immediately beneficial.

Russian Runs Table 2 shows results for Russian monolingual retrieval. The PRF summary length is 6 sentences here. This is a small document collection and the lack of variation in recall for the different summary methods is perhaps not surprising. The Snowball preprocessing of Russian is rather limited, and further development of our Russian language preprocessing is planned, but these results are generally encouraging.

Finnish Runs Table 3 shows results for Finnish monolingual retrieval. Summary length is 4 sentences with 30 documents this time used for expansion term selection. Our preprocessing of Finnish here again employs the Snowball stemming. This does not fully address the complex structure of Finnish word compounds, and again further work is planned to extend word decompounding. While average precision appears reasonable here, recall is poor in some cases, probably resulting from the failure to properly address the decompounding issues.

English Runs Table 4 shows English monolingual results. Our retrieval system appears to be performing fairly well on this dataset.

3.2 Bilingual Runs

This section gives results for our bilingual retrieval experiments. Results are shown for our official runs for German and Dutch topics to French documents, and English topics to Russian documents, together with additional unofficial results for English topics to French and Finnish document sets also reported for later comparison with multilingual retrieval results.

Table 1. Monolingual French retrieval results. (Relevant: 915)

Documents		Original French				Translated to English			
		L	T	Q	A	L	T	Q	A
Prec.	5 docs	0.445	0.429	0.437	0.429	0.400	0.396	0.404	0.400
	10 docs	0.369	0.361	0.365	0.363	0.349	0.349	0.341	0.347
	15 docs	0.333	0.327	0.339	0.335	0.320	0.317	0.316	0.317
	20 docs	0.307	0.298	0.305	0.295	0.287	0.288	0.290	0.286
Av. Precision		0.420	0.410	0.414	0.424	0.394	0.400	0.397	0.393
Rel. Ret.		839	844	849	843	781	774	772	774

Table 2. Monolingual Russian retrieval results. (Relevant: 123)

		L	T	Q	A
Prec.	5 docs	0.177	0.200	0.200	0.177
	10 docs	0.136	0.129	0.138	0.132
	15 docs	0.104	0.102	0.106	0.102
	20 docs	0.084	0.088	0.087	0.085
Av Precision		0.363	0.379	0.372	0.350
Rel. Ret.		101	101	101	101

Table 3. Monolingual Finnish retrieval results. (Relevant: 413)

		L	T	Q	A
Prec.	5 docs	0.382	0.382	0.391	0.369
	10 docs	0.311	0.309	0.307	0.298
	15 docs	0.253	0.258	0.250	0.242
	20 docs	0.206	0.211	0.212	0.199
Av Precision		0.432	0.448	0.449	0.425
Rel. Ret.		311	333	327	304

Table 4. Monolingual English retrieval results. (Relevant: 375)

		L	T	Q	A
Prec.	5 docs	0.362	0.367	0.366	0.367
	10 docs	0.281	0.286	0.281	0.286
	15 docs	0.238	0.237	0.230	0.233
	20 docs	0.202	0.204	0.201	0.201
Av Precision		0.482	0.498	0.487	0.491
Rel. Ret.		356	348	343	359

Table 5. Bilingual retrieval results German topics to retrieve French documents. Topics translated into French using Systran MT. (Relevant: 915)

		L	T	Q	A
Prec.	5 docs	0.314	0.318	0.310	0.327
	10 docs	0.263	0.263	0.265	0.265
	15 docs	0.248	0.241	0.241	0.250
	20 docs	0.227	0.219	0.222	0.235
Av Precision		0.296	0.295	0.296	0.299
% mono.		70.5%	72.0%	71.5%	70.5%
Rel. Ret.		710	727	713	704
chg. Rel. Ret.		-129	-117	-136	-139

Table 6. Bilingual retrieval results Dutch topics to retrieve French documents. Topics translated into French using Systran MT. (Relevant: 915)

		L	T	Q	A
Prec.	5 docs	0.342	0.339	0.355	0.347
	10 docs	0.302	0.286	0.296	0.296
	15 docs	0.274	0.267	0.269	0.268
	20 docs	0.251	0.245	0.248	0.251
Av Precision		0.339	0.331	0.333	0.334
% mono.		80.7 %	80.7%	80.4%	78.8%
Rel. Ret.		768	777	770	778
chg. Rel. Ret.		-76	-67	-79	-65

Table 7. Bilingual retrieval results English topics to retrieve Russian documents. Topics translated into Russian using PROMT and a Merged combination of MT systems. (Relevant: 123)

		PROMT				Merged			
		L	T	Q	A	L	T	Q	A
Prec.	5 docs	0.177	0.182	0.177	0.182	0.177	0.171	0.159	0.177
	10 docs	0.109	0.106	0.109	0.106	0.118	0.106	0.100	0.109
	15 docs	0.077	0.080	0.078	0.077	0.086	0.078	0.075	0.078
	20 docs	0.063	0.068	0.065	0.063	0.074	0.068	0.0068	0.068
Av Precision		0.296	0.321	0.305	0.320	0.317	0.310	0.281	0.313
% mono.		81.5%	84.7%	82.0%	91.4%	87.3%	81.8%	75.5%	89.4%
Rel. Ret.		95	96	95	96	94	95	95	95
chg. Rel. Ret.		-6	-5	-6	-5	-7	-6	-6	-6

Table 8. Bilingual retrieval results English topics to retrieve French documents. Topics translated into French and documents translated into English using Systran. (Relevant: 915)

Documents		Original French				Translated to English			
		L	T	Q	A	L	T	Q	A
Prec.	5 docs	0.314	0.322	0.310	0.310	0.331	0.331	0.343	0.318
	10 docs	0.274	0.282	0.276	0.278	0.280	0.274	0.267	0.276
	15 docs	0.246	0.261	0.260	0.259	0.259	0.254	0.250	0.252
	20 docs	0.231	0.239	0.236	0.237	0.236	0.232	0.225	0.228
Av Precision		0.322	0.335	0.328	0.323	0.318	0.321	0.302	0.298
% mono.		76.7%	81.7%	79.2%	76.2%	80.7%	80.3%	76.1%	75.8%
Rel. Ret.		745	757	754	745	727	716	715	715
chg. Rel. Ret.		-94	-87	-95	-89	-51	-58	-57	-59

Table 9. Bilingual retrieval results English topics to retrieve Finnish documents. Topics translated into Finnish using InterTrans. (Relevant: 413)

		L	T	Q	A
Prec.	5 docs	0.191	0.182	0.187	0.187
	10 docs	0.171	0.160	0.167	0.167
	15 docs	0.147	0.141	0.150	0.145
	20 docs	0.124	0.124	0.126	0.120
Av Precision		0.200	0.200	0.202	0.203
% mono.		46.3%	44.6%	45.0%	47.8%
Rel. Ret.		212	201	218	192
chg. Rel. Ret.		-99	-121	-109	-112

German to French Runs Table 5 shows results for German to French bilingual retrieval. PRF summary length is 4 sentences. Topics were translated directly from German to French using Systran via the Babelfish (`http://www.babelfish.altavista.com`) website. Comparing these results to the monolingual French retrieval results in Table 1, we observe about a 30% reduction in average precision, accompanied by an average reduction in relevant documents retrieved of around 120.

Dutch to French Runs Table 6 shows results for Dutch to French bilingual retrieval. PRF parameters are the same as German to French retrieval, topics again being translated directly using Babelfish. In this case, we see that average precision is reduced by only 20% relative to the monolingual results in Table 1, with a smaller decrease in relevant retrieved relative to monolingual retrieval averaging around 70.

English to Russian Runs Table 7 shows results for English to Russian bilingual retrieval. PRF summary length is 6 sentences with only 6 documents used for expansion term selection. Topics were translated using three online MT systems:

Systran (`http://www.systranbox.com/systran/box`), PROMT (`http://www.online-translator.com/default.asp?lang=en`) and LogoMedia (`http://www.logomedia.net/`). Results are shown for PROMT topic translation, and a union merge of the three translations. The merged results show a marginal relative reduction in performance metrics, this is perhaps a little surprising with respect to the number of relevant retrieved, where the greater range of terms in the merged translated topics might be expected to locate more relevant documents. The bilingual average precision varies between 75% and 90% of the monolingual performance shown in Table 2, with only a small number of relevant documents not retrieved. However, the very small number of relevant documents available means that these results must be treated with caution.

English to French Runs Table 8 shows unofficial results for English topic to French documents. Results are shown for both topic and document translation, using the same retrieval and PRF parameters used for the monolingual results in Table 1. Using topic translation average precision is between 75% and 80% of monolingual performance, with an average reduction in relevant documents retrieved of around 90. Using document translation there is a similar percentage reduction in average precision, but the reduction in relevant documents retrieved averages only 55 in this cases. Overall topic translation still outperforms document translation, as observed in Table 1, but the difference is smaller for bilingual than monolingual retrieval.

English to Finnish Runs Table 9 shows unofficial runs for English topic to Finnish documents. Topic translation was carried out using InterTrans[2]. Results here compared to the monolingual results in Table 3 are relatively poor. Average precision is only about 45% of monolingual, with a reduction of around 100 in the number of relevant documents retrieved. This latter figure represents a reduction of more than 30% in the number of relevant documents retrieved relative to the monolingual results. The impact of this comparatively low performance on the multilingual retrieval task is examined in the next section.

3.3 Multilingual Runs

This section gives out multilingual retrieval results. These experiments investigate a number of different scenarios of document and topic translation, and merging to form a multilingual output list. Use of these alternative scenarios is intended both to better understand the behaviour of list merging under different circumstances for multilingual IR, and to simulate alternative operational conditions.

Results are reported for existing data fusion methods. The initial results show overall multilingual performance. These are then broken down by language to examine the retrieval behaviour for each separate language within the multilingual output and to compare the effect of the different merging strategies.

[2] translations kindly provided by Jacques Savoy.

The data fusion methods used were designated s and u in our submission to CLEF 2003 [1]. For s data fusion each document is scored as follows,

$$sms_x(j) = \frac{ms_x(j)}{gms}$$

where $sms_x(j)$ is the revised matching score for document j in list x, $ms_x(j)$ is the original matching score of j in x, and gms is the global maximum matching score across the lists to be merged. For u data fusion each document is scored as follows,

$$ums_x(j) = \frac{ms_x(j)}{gms} \times rank_x$$

where $ums_x(j)$ is the revised score of j in x, $ms_x(j)$ and gms have the same definitions as before, and $rank_x$ is the anticipated likelihood of finding a relevant document in list x. This merging scheme is related to the Collection Size-Based Interleaving method proposed in [6]. In this case retrieved documents were interleaved into a merged list based only on collection size. This strategy was based on the observation that CLEF topics often have a distribution of relevant documents across the different languages in proportion to collection size. In our case we combine this concept with the matching score. The principle of linear list weighting using $rank_x$ can be used more generally to take account of the variable effectiveness of retrieval for different collections. Notably for our experiments, based on our training results and those observed for the test topics in Table 9, we would anticipate performance for Finnish retrieval in the multilingual task to be weaker than that for the other languages, and hence we can choose to allocate it a low value of $rank_x$.

Multilingual with s data fusion Table 10 shows results for our official multilingual retrieval experiments created using s merging. The topic language used in all cases is English. All runs were carried out using PRF with A type summaries. A number of different sets of document lists were formed as follows:

1. data fusion of monolingual English results and separate bilingual French, Russian and Finnish runs reported in Tables 4,7,8,9. For Russian the PROMT translated topics were used;
2. English and translated French documents merged into a single collection, retrieval run output fused with Russian and Finnish bilingual runs as in 1;
3. as 2, but a collection of *The Times* UK 1995 was combined with the merged English and translated French collection;
4. separate monolingual English and translated French document runs were data fused with the bilingual Russian and Finnish runs;
5. as 4, except that the English monolingual and translated French document retrieval runs used PRF expansion terms taken from merged collection used in 2.

Table 10. Multilingual retrieval results with fused lists as described in the text using *s* data fusion. (Relevant: 1826)

		1	2	3	4	5
Prec.	5 docs	0.388	0.360	0.372	0.380	0.364
	10 docs	0.354	0.330	0.350	0.352	0.356
	15 docs	0.316	0.311	0.328	0.327	0.331
	20 docs	0.302	0.292	0.316	0.306	0.315
Av Precision		0.263	0.248	0.272	0.273	0.274
Rel. Ret.		1244	1119	1232	1244	1216

Table 11. Breakdown of multilingual retrieval results by language for the various merging schemes using *s* data fusion.

Merging			English	French	Finnish	Russian
Scheme	Relevant		375	915	413	123
1	Prec.	5 docs	0.119	0.225	0.062	0.018
		10 docs	0.117	0.204	0.047	0.021
		15 docs	0.103	0.189	0.037	0.016
		20 docs	0.102	0.184	0.031	0.012
	Av Precision		0.166	0.232	0.058	0.057
	Rel. Ret.		310	714	145	75
2	Prec.	5 docs	0.176	0.118	0.076	0.041
		10 docs	0.164	0.118	0.058	0.035
		15 docs	0.148	0.125	0.049	0.029
		20 docs	0.133	0.126	0.046	0.023
	Av Precision		0.228	0.134	0.077	0.075
	Rel. Ret.		330	557	154	78
3	Prec.	5 docs	0.181	0.131	0.071	0.041
		10 docs	0.159	0.137	0.062	0.038
		15 docs	0.143	0.140	0.056	0.029
		20 docs	0.134	0.141	0.054	0.024
	Av Precision		0.230	0.165	0.077	0.108
	Rel. Ret.		319	680	155	78
4	Prec.	5 docs	0.195	0.122	0.071	0.047
		10 docs	0.171	0.127	0.067	0.035
		15 docs	0.149	0.129	0.061	0.029
		20 docs	0.135	0.128	0.057	0.024
	Av Precision		0.240	0.159	0.082	0.110
	Rel. Ret.		323	692	154	75
5	Prec.	5 docs	0.181	0.114	0.076	0.047
		10 docs	0.181	0.131	0.060	0.032
		15 docs	0.156	0.136	0.053	0.028
		20 docs	0.142	0.138	0.050	0.024
	Av Precision		0.228	0.153	0.074	0.106
	Rel. Ret.		328	663	151	74

Table 12. Results for merged English and translated French collections, and for separate English and translated French collections with PRF from merged collection.

		English	French	English	French
Prec.	5 docs	0.233	0.196	0.381	0.351
	10 docs	0.214	0.196	0.288	0.289
	15 docs	0.182	0.197	0.241	0.283
	20 docs	0.162	0.191	0.212	0.254
Av Precision		0.267	0.206	0.492	0.321
Rel. Ret.		342	703	352	754

Table 13. Results for merged English and translated French collections combined with UK Times 1995, and for separate English and translated French collections with PRF from merged collection.

		English	French	English	French
Prec.	5 docs	0.233	0.196	0.348	0.363
	10 docs	0.219	0.188	0.288	0.300
	15 docs	0.194	0.189	0.227	0.275
	20 docs	0.174	0.187	0.202	0.250
Av Precision		0.297	0.209	0.482	0.335
Rel. Ret.		344	713	351	774

From the results in Table 10, average precision is best for methods 3, 4 and 5, with best recall for methods 1 and 4. Overall method 4 is the most effective for this experiment.

Table 11 shows results for merging schemes 1 to 5 broken down by the individual languages in the merged lists. It can be seen that the overall dramatic reduction in performance between schemes 1 and 2 shown in Table 10 results entirely from loss in performance for the French documents. There is a significant reduction in all precision measures, and an average loss of more than 3 relevant documents per topic. Interestingly the combination with the *The Times* UK data in scheme 3 appears to overcome this problem to a significant extent with regard to recall, with a small improvement in the precision measures also being observed. By contrast while the precision for French is reduced in schemes 2 and 3 compared to scheme 1, it is much improved for English while the recall remains largely unchanged. There is also an improvement in the precision measures for Finnish and Russian in schemes 2 and 3 compared to scheme 1. The overall effect of improved precision for English, Finnish and Russian, with reasonable performance for French, mean that the multilingual result for scheme 3 is the overall best of these schemes with respect to precision, although the difference in recall between schemes 1 and 3 is marginal. Merging four separate lists in schemes 4 and 5 produces better average precision results than scheme 1. Looking again at Table 11, it can be seen that retrieval for English, Finnish and Russian is more effective for schemes 4 and 5, whereas French retrieval is more effective with the untranslated documents in scheme 1. The average French matching scores appear to introduce bias in scheme 1. The errors introduced by document translation may help to reduce this effect when merging four lists in schemes 4 and 5, but this issue needs to be investigated further.

Combined English and French Collections Columns 1 and 2 of Table 12 show separate English and French retrieval within the combined collection used for merging scheme 2 in Table 10 prior to fusion with Russian and Finnish document lists. Comparing these results with those for scheme 2 in Table 11, it can be seen that loss in effectiveness in the multilingual retrieval results is caused mainly by the behaviour of the French documents. There is a large loss in average precision and the number of relevant documents retrieved, presumably because of low matching scores arising from document translation errors causing these documents to be dropped from the bottom of the merged list in Table 12. By contrast Table 13 shows corresponding results for the English and translated French collections merged with *The Times* UK 1995 as used with merging scheme 3 in Table 10. This shows an improvement for English document retrieval, which is also reflected in the results in Table 12. While there is not a significant difference between the French results in Tables 12 and 13 prior to multilingual fusion, scheme 3 shows a good improvement over scheme 2 in Table 11. The additional information from *The Times* collection may produce more robust matching scores for the translated French documents based on selection of expansion terms or term weights.

Columns 3 and 4 of Tables 12 and 13 show results for the English and translated French documents with PRF using the respective merged collections. Column 3 can be compared with column A in Table 4, and column 4 with translated documents column A in Table 8. While there is little change to the effectiveness of English document retrieval from using merged collection PRF, there is an observable improvement in both precision and recall for the translated French documents. It is then a little surprising to see in Table 11 that the French retrieval performance is actually lower for scheme 5 than for scheme 4.

Multilingual with u data fusion Table 14 shows multilingual IR results using the u data fusion scheme with $rank_x$ values set as follows: English: 1.5, French: 1.3, Russian: 1.2 and Finnish: 0.8. For the merged English and translated French documents the $rank_x$ value was set to 1.5. These values were set intuitively based on collection size and anticipated likelihood of retrieving relevant documents. The results in Table 14 show similar trends to those already observed in Table 10, although there is a general trend to slightly higher precision values and a very small reduction in relevant documents retrieved.

Table 15 again shows the breakdown in retrieval performance for the different languages. Comparing these results with those in Table 11, it can be seen that the $rank_x$ bias improves both the precision and recall for English in all conditions. Interestingly it makes no difference for French merging in scheme 1, suggesting that there is already a considerable bias towards French documents in this case.

The smaller $rank_x$ for Finnish and Russian leads to a reduction in both precision and recall for both of these languages. There are relatively few relevant

Table 14. Multilingual retrieval results with fused lists as described in the text using u data fusion. (Relevant: 1826)

		1	2	3	4	5
Prec.	5 docs	0.392	0.368	0.372	0.400	0.404
	10 docs	0.350	0.352	0.379	0.354	0.396
	15 docs	0.324	0.319	0.345	0.331	0.357
	20 docs	0.310	0.294	0.311	0.301	0.330
Av Precision		0.268	0.250	0.275	0.275	0.278
Rel. Ret.		1236	1106	1219	1221	1212

Table 15. Breakdown of multilingual retrieval results by language for the various merging schemes using u data fusion.

Merging			English	French	Finnish	Russian
Scheme	Rel. Avail.		375	915	413	123
1	Prec.	5 docs	0.191	0.225	0.009	0.006
		10 docs	0.167	0.202	0.007	0.009
		15 docs	0.151	0.185	0.007	0.014
		20 docs	0.137	0.184	0.007	0.013
	Av Precision		0.238	0.230	0.022	0.046
	Rel. Ret.		338	716	115	67
2	Prec.	5 docs	0.286	0.110	0.013	0.012
		10 docs	0.224	0.139	0.013	0.024
		15 docs	0.184	0.140	0.009	0.024
		20 docs	0.163	0.138	0.008	0.022
	Av Precision		0.309	0.144	0.028	0.055
	Rel. Ret.		354	555	129	68
3	Prec.	5 docs	0.276	0.098	0.027	0.029
		10 docs	0.219	0.153	0.022	0.029
		15 docs	0.197	0.151	0.016	0.026
		20 docs	0.167	0.146	0.014	0.022
	Av Precision		0.363	0.158	0.033	0.104
	Rel. Ret.		336	682	128	73
4	Prec.	5 docs	0.291	0.106	0.031	0.035
		10 docs	0.226	0.118	0.024	0.038
		15 docs	0.191	0.133	0.022	0.029
		20 docs	0.166	0.132	0.0519	0.024
	Av Precision		0.366	0.159	0.035	0.102
	Rel. Ret.		337	689	125	70
5	Prec.	5 docs	0.224	0.184	0.013	0.035
		10 docs	0.202	0.194	0.018	0.029
		15 docs	0.178	0.182	0.015	0.024
		20 docs	0.154	0.179	0.012	0.022
	Av Precision		0.261	0.195	0.019	0.104
	Rel. Ret.		335	684	122	71

documents available for Russian (123) compared to English (375), and, as noted earlier, bilingual performance for Finnish using our simple retrieval scheme is poor, with only about 30% of the available 413 relevant documents appearing in the data fused list. This compares to relevant document retrieved proportions in the data fused list of more than 80% for English, 70% for French and 60% for Russian. Hence biasing against Russian and Finnish has little impact on the overall multilingual result.

Thus, while these simple merging schemes can be biased towards larger collections containing more relevant documents to improve overall average precision multilingual, this is likely to be at the cost of retrieval effectiveness for the collections suspected of containing small numbers of relevant documents or for which the retrieval effectiveness is expected to be poor.

4 Conclusions and Further Work

Our work for CLEF 2004 has produced a retrieval framework based on BM25 that can be easily adapted to new document languages. While our experiments have demonstrated that this approach can be effective, further work is needed to improve preprocessing for specific languages. Our multilingual experiments reveal interesting behaviour for individual language components of merged retrieval lists. While these results help us understand the merged multilingual retrieval results, they do not solve the problem of achieving truly effective reliable merging.

References

[1] A. M. Lam-Adesina and G. J. F. Jones. Exeter at CLEF 2003: Experiments with Machine Translation for Monolingual, Bilingual and Multilingual Retrieval. In *Proceedings of Fourth Workshop of the Cross-Language Evaluation Forum (CLEF 2003)*, C. Peters, J. Gonzalo, M. Braschler and M. Kluck (Eds.), Lecture Notes in Computer Science, Springer, Heidelberg, Germany, pages 271-285, 2004.

[2] *Snowball* toolkit `http://snowball.tartarus.org/`

[3] Porter, M. F.: An algorithm for suffix stripping. *Program* 14:10-137, 1980.

[4] Lam-Adesina, A. M. and Jones, G. J. F.: Applying Summarization Techniques for Term Selection in Relevance Feedback. In *Proceedings of the 24th Annual International ACM SIGIR Conference*, pages 1-9, New Orleans, ACM, 2001.

[5] Peters, C., Braschler, M., Di Nunzio, G., and Ferro, N.: CLEF 2004: Ah Hoc Track Overview and Results Analysis. In *Proceedings of Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004)*, Peters, C., Clough, P, Gonzalo, J., Jones, G., Kluck, M., and Magnini, B. (Eds.), Lecture Notes in Computer Science, Springer, Heidelberg, Germany (in print), 2005.

[6] Braschler, M., Göhring, A., and Schäuble, P.: Eurospider at CLEF 2002. In *Proceedings of Third Workshop of the Cross-Language Evaluation Forum (CLEF 2002)*, Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds.), Lecture Notes in Computer Science (LNCS 2785), Springer, Heidelberg, Germany, pages 164-174, 2003.