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, 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 this year. The main focus of our investigation this year was extending our retrieval system to document languages other than English, and completing the multilingual task comprising four languages: English, French, Russian and Finnish. Results from our French monolingual experiments indicate that working in French is more effective for retrieval than adopting document and topic translation to English. However, comparison of our multilingual retrieval results using different topic and document translation reveals that this result does not extend to retrieved list merging for the multilingual task in a simple predictable way.

1 Introduction

Dublin City University's (DCU) participation in CLEF 2004 Monolingual, Bilingual and Multilingual track builds on previous work at the University of Exeter [1]. Previously had focussed on document translation using machine translation using English as a "pivot" language for all tasks. Our work for CLEF 2004 concentrated on extending our retrieval system to work in the document language with topic translation when needed. Our strategy is to use our existing Okapi retrieval system and make use of the Snowball stemmers and stop word lists [2]. Using these tools we completed runs for monolingual French, Russian and Finnish, official bilingual French and Russian, and the multilingual track consisting of English, French, Russian and Finnish, together with the additional monolingual and bilingual runs needed for the multilingual task.

This paper briefly describes our retrieval system and reports results from our runs with analysis of the observed performance levels.

2 Methodology

In this section we first give a brief overview of our retrieval system, and then describe how we adapted it to be language independent and the specific detail of preprocessing for our CLEF 2004 runs.

2.1 Retrieval System

The basis of our experimental retrieval system remains the City University research distribution version of the Okapi system, as used in our previous CLEF participation [1]. The basic Okapi system includes tools for English language preprocessing and can only handle ASCII characters for English characters and punctuation symbols.

For English language runs we continued to use this 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 [3] and terms were indexed using a small set of synonyms.

Terms are weighted using the standard BM25 weighting scheme and all runs use our summary-based pseudo relevance feedback (PRF) method [4]. The summary generation method combines Luhn's keyword cluster method, a title terms frequency method, a location/header method and a query-bias method from to form an overall significance score for each sentence.

2.2 Data Preprocessing

In order to use the Okapi software with non-English document we carried out the language preprocessing outside Okapi and then encoded the resulting character strings into ASCII as follows.

Language Preprocessing The documents and topic are prepared using a pipeline of preprocessing 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 third step is the conversion of characters to lower case. A finite set of upper case characters are mapped to lower case equivalents. Distinct mappings must be used for each character set. 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 corruptiion of the original data to "boxdrawing" symbols was observed. The Russian stopword list used here consists only of the simple first part of the Snowball list.

At the next stage stop words are removed. The stop word lists provided by Snowball are used for French and Finnish stop word removal. 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].

Text Encoding The OKAPI system does not accept special characters that are used in Finnish, French and Russian. All character strings were encoded with just 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 easily performed. However, the encoded form is not readable by humans and string similarities do not stay intact. The latter is not a problem, since we do not want to retrieve fuzzy matches to our queries with OKAPI. Example: for the three words pécheur, pêcheur and pêcheurs are encoded as gropmdpbtfui, cbppmdpbtfui and klcgrwruwanejd.

3 Results

This section presentation results and analysis of our experimental runs. We report precision at rank 10, average precsion and total number of relevant documents retrieved. System parameters were selected using CLEF 2003 test collections for each language. All runs use the Title and Description CLEF topic fields. In all cases Okapi $K1 = 1.0$ and $b = 0.75$. The following PRF summary sentence selection: T = title method, Q = query-bias method, A = linear summ of all methods, L = Luhn method. The 20 top ranked PRF expansion terms were selected from the summaries of the top 5 ranked documents. The original topic terms were upweighted by a factor of 3.5 relative to terms introduced by PRF.

3.1 Monolingual Retrieval

French Runs Table 1 shows results for French monolingual retrieval. Separate results are shown for documents and topics in French and translated into English into Systran MT. For French document PRF summary length is 4 sentences and for translated 6 sentences, with 20

[1] Punctuation must be maintain to facilitate document summarization for PRF.

	French			English		
	T	Q	A	Q	A	L
Prec. 10 docs	0.361	0.365	0.363	0.341	0.347	0.349
Av Precision	0.410	0.414	0.424	0.397	0.393	0.394
Rel. Ret.	844	849	843	772	774	781

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

	T	Q	A	L
Prec. 10 docs	0.129	0.138	0.132	0.136
Av Precision	0.379	0.372	0.350	0.363
Rel. Ret.	101	101	101	101

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

	T	Q	A	L
Prec. 10 docs	0.309	0.307	0.298	0.311
Av Precision	0.448	0.449	0.425	0.432
Rel. Ret.	333	327	304	311

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

	T	Q	A	L
Prec. 10 docs	0.286	0.281	0.286	0.281
Av Precision	0.498	0.487	0.491	0.482
Rel. Ret.	348	343	359	356

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

documents used for expansion term selection in both cases. It can be seen that working in French produces superior retrieval performance with respect to both precision and recall metrics.

Russian Runs Table 2 shows results for Russian monolingual retrieval. The PRF summary length is 6 sentences here with 20 documents used for expansion term selection. This is a small document collection and the lack of variation in recall for the different summary methods is perhaps not surprising. 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 used for expansion term selection. Our preprocessing of Finnish here only employs the Snowball stemming. This does not fully address the complex structure of Finnish word compounds, and further work is planned to extend word decompounding. While average precision appears reasonable here, recall appears poor in some cases, probably resulting from the failure to properly address decompounding.

English Runs Table 4 shows English monolingual results. Our retrieval system appears to be performing fairly well on this dataset. Results are included here for analysis of the multilingual retrieval runs.

3.2 Bilingual Runs

German to French Runs Table 5 shows results for German to French bilingual retrieval. PRF summary length is 4 sentences with 20 documents used for expansion term selection. Topics were translated directly from German to French using Systran via the Babelfish (`http://www.babelfish.altavista.com`) website. We observed about 30% reduction in average precision relative to monolingual our French retrieval accompanied by a large reduction in relevant documents retrieved.

	T	Q	A	L
Prec. 10 docs	0.263	0.265	0.265	0.263
Av Precision	0.295	0.296	0.299	0.296
% mono.	72.0%	71.5%	70.5%	—
Rel. Ret.	727	713	704	710

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

	T	A	L
Prec. 10 docs	0.286	0.296	0.302
Av Precision	0.331	0.334	0.339
% mono.	80.7%	78.8%	—
Rel. Ret.	777	778	768

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

					Merged
	T	Q	A	L	A
Prec. 10 docs	0.106	0.109	0.106	0.109	0.109
Av Precision	0.321	0.305	0.320	0.296	0.313
% mono.	64.5%	62.6%	65.2%	61.4%	63.7%
Rel. Ret.	96	95	96	95	95

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

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% with a matching smaller decrease in relevant retrieved relative to monolingual retrieval.

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 are translated using 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 Systran and a union merge of the three translations. The merged results show a marginal reduction in performance metrics, this is perhaps a little surprising with respect to the number of relevant retrieved.

English to Finnish Runs and English to French Runs Table 8 and Table 9 show results for English to Finnish and English to French bilingual retrieval respectively. English to Finnish topic translation was carried out using InterTrans[2] and topics translated from English to French using Systran. These results are included here for analysis of the multilingual retrieval runs.

3.3 Multilingual Runs

Table 10 shows results for our multilingual runs. In all cases PRF used A type summaries. Multilingual output was generated by merging separate lists using data fusion, each document being assigned a score $w = ms(j)/gmax\_ms$, where $ms(j)$ is the original document score and $gmax\_ms$ is the global maximum $ms(j)$ across the lists being merged.

The fused lists shown in Table 10 are as follows: 1: monolingual results merged, 2: English and translated French documents merged into a single collection, combined retrieval run fused with Russian and Finnish monolingual, 3: as 2, but *The Times* UK 1995 merged with combined

[2]Kindly provided by Jacques Savoy.

	T	Q	A	L
Prec. 10 docs	0.160	0.167	0.167	0.171
Av Precision	0.200	0.202	0.203	0.200
% mono.	44.6%	45.0%	47.8%	46.3%
Rel. Ret.	201	218	192	212

Table 8: Bilingual retrieval results English topics to retrieve Finnish documents. Topics translated into Finnish using InterTrans. (Relevant: 413)

	French			English		
	T	Q	A	T	Q	A
Prec. 10 docs	0.282	0.276	0.278	0.274	0.267	0.276
Av Precision	0.335	0.328	0.323	0.321	0.302	0.298
% mono.	81.7%	79.2%	76.2%	—	76.1%	75.8%
Rel. Ret.	757	754	745	716	715	715

Table 9: Bilingual retrieval results English topics to retrieve French documents. Topics translated into French using Systran. (Relevant: 915)

	1	2	3	4	5
Prec. 10 docs	0.354	0.330	0.350	0.352	0.356
Av Precision	0.263	0.248	0.272	0.273	0.274
Rel. Ret.	1244	1119	1232	1244	1216

Table 10: Multilingual retrieval results with fused lists as described in the text. (Relevant: 1826)

collection for retrieval, 4: separate English, translated French, Russian and Finnish fused, 5: as 4, expect English and French PRF expansion taken from merged collection from 2.

Table 11 shows results for these runs broken down by the individual languages in the merged lists. It can be seen that the dramatic reduction in performance between schemes 1 and 2 shown Table 10 results entirely from loss in performance for the French documents. Interestingly the combination with the *The Times* UK data in scheme 3 appears to overcome this problem. Similarly working with the 4 separate lists in schemes 4 and 5 produces better overall results than scheme 1 with the untranslated documents. The dominance of French in scheme 1 needs to be further investigated. The French collection is by far the largest here, which may lead to it dominating the scores, the errors introduced by document translation may help to ameliorate this effect, but this issue needs to be investigated properly.

Merged English and French Collections Table 12 shows English and French retrieval within the merged collection list used for scheme 2 in Table 10 prior to fusion with Russian and Finnish. Comparing these results with those for scheme in Table 11 it can be seen that loss in retrieval in the multilingual fusion is caused mainly by the behaviour of the French documents, presumably because of low matching scores arising from document translation errors. By contrast Table 13 shows corresponding results for the collection merged with *The Times* UK 1995. While there is no significant change in the results prior to multilingual fusion, scheme 3 shows a good improvement of scheme 2 in Table 11, the additional information from *The Times* collection may produce more robust matching scores for the translated French documents.

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 9. While there is little change to the effectiveness of English document retrieval from using merged PRF, there is an observable improvement in both precision and recall for the translated French documents. Looking at the behaviour of these lists schemes 4 and 5 of Table 11, there is a little effect on the average performance in the merged lists.

Merging Scheme		English	French	Finnish	Russian
	Relevant	375	915	413	123
1	Prec. 10 docs	0.117	0.204	0.047	0.021
	Av Precision	0.166	0.232	0.058	0.057
	Rel. Ret.	310	714	145	75
2	Prec. 10 docs	0.164	0.118	0.058	0.035
	Av Precision	0.228	0.134	0.077	0.075
	Rel. Ret.	330	557	154	78
3	Prec. 10 docs	0.159	0.137	0.062	0.038
	Av Precision	0.230	0.165	0.077	0.108
	Rel. Ret.	319	680	155	78
4	Prec. 10 docs	0.171	0.127	0.067	0.035
	Av Precision	0.240	0.159	0.082	0.110
	Rel. Ret.	323	692	154	75
5	Prec. 10 docs	0.181	0.131	0.060	0.032
	Av Precision	0.228	0.153	0.074	0.106
	Rel. Ret.	328	663	151	74

Table 11: Breakdown of multilingual retrieval results by language for the various merging schemes.

	English	French	English	French
Prec. 10 docs	0.214	0.196	0.288	0.289
Av Precision	0.267	0.206	0.492	0.321
Rel. Ret.	342	703	352	754

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

	English	French	English	French
Prec. 10 docs	0.219	0.188	0.288	0.300
Av Precision	0.297	0.209	0.482	0.335
Rel. Ret.	344	713	351	774

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.

4 Conclusions and Further Work

Our work for CLEF 2004 has produced a system that can be easily adapted to different document languages. Further work is needed to improved preprocessing for specific languages. While our multilingual experiments show interesting behaviour for individual language components of merged retrieva lists, further investigation is needed to better understand the reasons for these results.

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 Workshop of the Cross-Language Evaluation Forum (CLEF 2003)*, Trondheim, Norway, C. Peters et al. editors, Springer-Verlag, 2004.

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

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

[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*, pages 1-9, New Orleans, ACM, 2001.