Dublin City University at CLEF 2006: Cross-Language Speech Retrieval (CL-SR) Experiments

Gareth J. F. Jones, Ke Zhang and Adenike M. Lam-Adesina
Centre for Digital Video Processing & School of Computing
Dublin City University, Dublin 9, Ireland
{gjones,kzhang,adenike}@computing.dcu.ie

Abstract

The Dublin City University participation in the CLEF 2006 CL-SR task concentrated on exploring the combination of the multiple fields associated with the documents. This was based on use of the extended BM25F field combination model originally developed for multi-field text documents. Additionally, we again conducted runs with our existing information retrieval methods based on the Okapi model. This latter method required an approach to determining approximate sentence boundaries within the free-flowing automatic transcription provided to enable us to use our summary-based pseudo relevance feedback (PRF). Experiments were conducted only for the English document collection. Topics were translated into English using Systran V3.0 machine translation.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Cross-language spoken document retrieval, Multi Field Document Retrieval, Pseudo relevance feedback

1 Introduction

The Dublin City University participation in the CLEF 2006 CL-SR task concentrated on exploring the combination of the multiple fields associated with the speech documents. It is not immediately clear how best to combine the diverse fields of this document set most effectively in ad hoc information retrieval tasks, such as the CLEF 2006 CL-SR task. Our study is based on using the document field combination extended version of BM25 termed BM25F introduced in [1]. In addition, we carried out runs using our existing information retrieval methods based on the Okapi model to this data set [2]. Our official submissions included both English monolingual and French bilingual tasks using automatic only and combined automatic and manual fields. Topics

were translated into English using the Systran V3.0 machine translation system. The resulting translated English topics were applied to the English document collection.

The remainder of this paper is structured as follows: Section 2 summarises the motivation and implementation of the BM25F retrieval model, Section 3 overviews our basic retrieval system and describes our sentence boundary creation technique, Section 4 presents the results of our experimental investigations, and Section 5 concludes the paper with a discussion of our results.

2 Field Combination

The "documents" of the speech collection are based on sections of extended interviews which are segmented into topically related section. The spoken documents are provided with a rich set of data fields, full details of these are given in [3]. In summary the fields comprise:

- a transcription of the spoken content of the document generated using an automatic speech recognition (ASR) system,
- two assigned sets of keywords generated automatically (AKW1,AKW2),
- one assigned set of manually generated keywords (MKW1),
- a short three sentence manually written summary of each document,
- and a manually determined list of the names of all the individuals appearing in the interview.

Two standard methods of combining multiple document fields for tasks such as this are:

- to simply merge all the fields into a single document representation and apply standard single document field information retrieval methods,
- to index the fields separately, perform individual retrieval runs for each field and then to merge the resulting ranked lists by summing in a process of data fusion.

The topic of field combination for this type of task with ranked information retrieval schemes is explored in [1]. This paper demonstrated the weaknesses of the simple standard combination methods and proposed an extended version of the standard BM25 term weighting scheme referred to as BM25F which combines multiple fields in a more well founded way.

The BM25F combination approach uses a simple weighted summation of the multiple fields of the documents to form a single field for each document in the usual way. The importance of each document field for retrieval can be determined empirically in separate runs for each field, the terms appearing in each field are multiplied by a scalar constant representing this importance, and the components of all fields summed to form the overall single field document representation for indexing.

Once the fields have been combined in a weighted sum standard single field information retrieval methods can be applied.

3 System Setup

The basis of our experimental system is the City University research distribution version of the Okapi system [4]. The documents and search topics are processed to remove stopwords from a standard list of about 260 words, suffix stripped using the Okapi implementation of Porter stemming [5] and terms are indexed using a small standard set of synonyms. None of these procedures were adapted for the CLEF 2006 CL-SR test collection.

Our experiments augmented the standard Okapi retrieval system with two variations of pseudo relevance feedback (PRF) based on extensions of the Robertson selection value (rsv) for expansion term selection. One method is a novel field-based PRF which we are currently developing [6], and the other a summary-based method used extensively in our earlier CLEF submissions [7].

3.1 Term Weighting

Document terms were weighted using the Okapi BM25 weighting scheme developed in [4] calculated as follows,

$$cw(i,j) = cfw(i) \times \frac{tf(i,j) \times (k_1 + 1)}{k_1 \times ((1-b) + (b \times ndl(j))) + tf(i,j)}$$

where $cw(i,j)$ represents the weight of term i in document j, $cfw(i) = log((N - n(i) + 0.5)/(n(i) + 0.5))$, $n(i)$ is the total number of documents containing term i, and N is the total number of documents in the collection, $tf(i,j)$ is the within document term frequency, and $ndl(j) = dl(j)/\text{Av}.dl$ is the normalized document length where $dl(j)$ is the length of j. k_1 and b are empirically selected tuning constants for a particular collection. The matching score for each document is computed by summing the weights of terms appearing in the query and the document. The values used for our submitted runs were tuned using the CLEF 2005 training and test topics.

3.2 Pseudo-Relevance Feedback

The main challenge for query expansion is the selection of appropriate terms from the assumed relevant documents. For the CL-SR task our query expansion method operates as follows.

3.2.1 Field-Based PRF

Query expansion based on the standard Okapi relevance feedback model makes no use of the field structure of multi-field documents. We are currently exploring possible methods of making use of field structure to improvement the quality of expansion term selection. For this current investigation we adopted the following method.

The fields are merged as described in the previous section and retrieved performed using the initial query. The rsv is then calculated separately for each field of the original document, but where the document position in the ranked retrieval list has been determined using the combined document. The ranked rsv lists for each field are then normalised with respect to the highest scoring term in each list, and then summed to form a single merged rsv list from the expansion terms are selected. The objective of this process is to favour the selection of expansion terms which are ranked highly by multiple fields, rather than those which may obtain a high rsv value based on their association with a minority of the fields.

3.2.2 Summary-Based PRF

The method used here is based on our work originally described in [7], and modified for the CLEF 2005 CL-SR task [2]. A summary is made of the ASR transcription of each of the top ranked documents, which are assumed to be relevant for each PRF. Each document summary is then expanded to include all terms in the other metadata fields used in this document index. All non-stopwords in these augmented summaries are then ranked using a slightly modified version of the rsv [4]. In our modified version of rsv(i), potential expansion terms are selected from the augmented summaries of the top ranked documents, but ranked using statistics from a larger number of assumed relevant ranked documents from the initial run.

Sentence Selection The summary-based PRF method operates by selecting topic expansion terms from document summaries. However, since the transcriptions of the conversational speech documents generated using automatic speech recognition (ASR) do not contain punctuation, we developed a method of selecting significant document segments to identify documents "summaries". This uses a method derived from Luhn's word cluster hypothesis. Luhn's hypothesis states that significant words separated by not more than 5 non-significant words are likely to be strongly related. Clusters of these strongly related word were identified in the running document transcription by searching for word groups separated by not more than 5 insignificant words.

Words appearing between clusters are not included in clusters, but can be ignored for the purposes of query expansion since they are by definition stop words.

The clusters were then awarded a significance score based on two measures:

Luhn's Keyword Cluster Method: Luhn's method assigns a sentence score for the highest scoring cluster within a sentence [8]. We adapted this method to assign a cluster score as follows:

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

where SS1 = the sentence score
SW = the number of bracketed significant words
TW = the total number of bracketed words

Query-Bias Method This method assigns a score to each sentence based on the number of query terms in the sentence as follows:

$$SS2 = \frac{TQ^2}{NQ}$$

where SS2 = the sentence score
. TQ = the number of query terms present in the sentence
NQ = the number of terms in a query

The overall score for each sentence (cluster) was then formed by summing these two measures for each sentence. The sentences were then ranked by score with the highest scoring sentences selected as the document summary.

4 Experimental Investigation

This section gives results of our experimental investigations for the CLEF 2006 CL-SR task. We first present results for our field combination experiments and then those for experiments using our summary-based PRF method.

For our formal submitted runs the system parameters were selected by optimising results for the CLEF 2005 CL-SR training and test collection. Our submitted runs for the CLEF 2006 CL-SR task are indicated using a * in the results table.

4.1 Field Combination Experiments

Two sets of experiments were carried out using the field combination method. The first uses all the document fields combining manual and automatically generated fields, and the other only the automatically generated fields. We report results for our formal submitted runs using our field-based PRF method and also baseline results without feedback. We also give further results obtained by optimising performance for our systems using the CLEF 2006 CL-SR test set.

4.1.1 All Field Experiments

Submitted Runs Based on development runs with the CLEF 2005 data the Okapi parameters were set empirically as follows: $k_1 = 6.2$ and $b = 0.4$, and the document fields were weighted as follows:

Name field × 1;
Manualkeyword field × 10;
Summary field × 10;
ASR2006B × 2;
Autokeyword1 × 1;
Autokeyword2 × 1.

TD	Recall	MAP	P5	P10	P30
Baseline:	1844	0.223	0.366	0.293	0.255
PRF*	1864	0.202	0.321	0.288	0.252

Table 1: Results for monolingual English with all document fields with parameters trained on CLEF 2005 data.

TD	Recall	MAP	P5	P10	P30
Baseline	1491	0.158	0.306	0.256	0.204
PRF*	1567	0.160	0.291	0.252	0.199

Table 2: Results for French-English bilingual with all document fields with parameters trained on CLEF 2005 data.

TD	Recall	MAP	P5	P10	P30
Baseline:	1908	0.234	0.364	0.342	0.303
PRF	1929	0.243	0.364	0.370	0.305

Table 3: Results for monolingual English with all document fields with parameters optimised for the CLEF 2006 topics.

TD	Recall	MAP	P5	P10	P30
Baseline	1560	0.172	0.315	0.267	0.231
PRF	1601	0.173	0.315	0.267	0.225

Table 4: Results for French-English bilingual with all document fields with parameters optimised for the CLEF 2006 topics.

The contents of each field were multiplied by the appropriate factor and summed to form the single field document for indexing. Note the unusually high value of k_1 arises due to the change in the $tf(i,j)$ profile resulting from the summation of the document fields [1].

We conducted monolingual English and bilingual French-English runs. The French topics were translated into English using Systran Version 3.0.

The results of English monolingual runs are shown in Table 1, and those for French bilingual in Table 2. For both topic sets the top 20 ranked terms were added to the topic for the PRF run with the original topic terms upweighted by 3.0.

It can be seen from these results that, as is usually the case for cross-language information retrieval, performance for monolingual English is better than bilingual French-English for all measures. A little more surprising is that while the application of the field-based PRF gives a small improvement in the number of relevant documents retrieved, there is little effect on MAP, and precision at high ranked cut off points is generally degraded. PRF methods for this task are the subject of ongoing research, and we will be exploring these results further.

Further Runs Subsequent to the release of the relevance set for the CLEF 2006 topic set further experiments were conducted to explore the potential for improvement in retrieval performance when the system parameters are optimised. We next show our best results achieved so far using the field combination method. For these runs the fields were weighted as follows:

Name field × 1;
Manualkeyword field × 5;
Summary field × 5;
ASR2006B × 1;
Autokeyword1 × 1;
Autokeyword2 × 1,

and the Okapi parameters set empirically as follows: $k_1 = 10.5$ and $b = 0.35$.

TD	Recall	MAP	P5	P10	P30
Baseline	1290	0.071	0.163	0.163	0.149
PRF*	1361	0.073	0.152	0.142	0.146

Table 5: Results for monolingual English with only auto document fields with parameters trained on CLEF 2005 data.

TD	Recall	MAP	P5	P10	P30
Baseline	1070	0.047	0.119	0.113	0.106
PRF*	1097	0.047	0.106	0.094	0.102

Table 6: Results for French-English bilingual with only auto document fields with parameters trained on CLEF 2005 data.

TD	Recall	MAP	P5	P10	P30
Baseline	1335	0.080	0.224	0.215	0.169
PRF	1379	0.094	0.188	0.206	0.184

Table 7: Results for monolingual English with only auto document fields with parameters optimised for the CLEF 2006 topics.

TD	Recall	MAP	P5	P10	P30
Baseline	1110	0.050	0.121	0.127	0.123
PRF	1167	0.055	0.127	0.142	0.124

Table 8: Results for French-English bilingual with only auto document fields with parameters optimised for the CLEF 2006 topics.

The results of English monolingual runs are shown in Table 3, and those for French bilingual in Table 4. For monolingual English the top 20 terms were added to the topic for PRF run with the original topic terms upweighted by 33.0 For the French bilingual runs the top 60 terms were added to the topic with the original terms upweighted by 20.0. For these additional runs all test topics were included in all cases.

Looking at these additional results it can be seen that parameter optimisation gives a good improvement in all measures. It is not immediately clear whether this arises due to the instability of the parameters of our system, or a difference in some feature of the topics between CLEF 2005 and CLEF 2006. We will be investigating this issue further. Performance between monolingual English and bilingual French-English is similar to that observed for the submitted runs. PRF is generally more effective or neutral with the revised parameters, again we plan to conduct further exploration of PRF for multi-field documents is planned to better understand these results.

4.1.2 Automatic Only Field Experiments

Submitted Runs Based on development runs with the CLEF 2005 CL-SR data the system parameters for the submitted automatic only field experiments were set empirically as follows: $k_1 = 5.2$ and $b = 0.2$ and the document fields were weighted as follows:

ASR2006B $\times$ 2;
Autokeyword1 $\times$ 1;
Autokeyword2 $\times$ 1.

These were again summed to form the single field document for indexing. The same French-English topic translations were used for the automatic only field experiments as for the all field experiments.

The results of English monolingual runs are shown in Table 5, and those for French bilingual in Table 6. The top 30 terms were added to the topic for PRF run with the original topic terms

TDN	Recall	MAP	P10	P30
Baseline	1832	0.246	0.391	0.321
PRF*	1895	0.277	0.439	0.357

Table 9: Results for monolingual English with all document fields.

TDN	Recall	MAP	P10	P30
Baseline	633	0.029	0.069	0.068
PRF	993	0.047	0.118	0.107

Table 10: Results for monolingual English with only auto document fields.

TD	Recall	MAP	P10	P30
Baseline	627	0.025	0.069	0.061
PRF	900	0.039	0.091	0.089

Table 11: Results for monolingual English with only auto document fields.

upweighted by 3.0.

From these results it can again be seen that there is the expected reduction in performance between monolingual English and bilingual French-English. The field-based PRF is once again shown generally not to be effective with this dataset. There is a small improvement in the number of relevant documents retrieved, but no there is little positive impact for precision.

Further Runs Subsequent to the release of the relevance set for the CLEF 2006 topic set further experiments were conducted to explore the potential for improvement in retrieval performance when the system parameters are optimised. We next show our best results achieved so far using the field combination method. For these runs the field weights were identical to those above for the submitted runs, and the Okapi parameters were modified as follows: $k_1 = 40.0$ and $b = 0.3$.

The results of English monolingual runs are shown in Table 7, and those for French bilingual in Table 8. For monolingual English the top 40 terms were added to the topic for PRF run with the original topic terms upweighted by 3.0 For the French bilingual runs the top 60 terms were added to the topic with the original terms upweighted by 3.5. For these additional runs again all test topics were included in all cases.

These results show an improvement in all metrics relative to the submitted runs. Once again optimising the system parameters results in an improvement effectiveness of the PRF method. Further experiments are planned to explore these results further.

4.2 Summary-Based PRF Experiments

For these experiments the document fields were combined into a single field for indexing without application of the field weighting method. The Okapi parameters were again selected using the CLEF 2005 CL-SR training and test collections. The values were set as follows k_1=1.4 b=0.6. For all our PRF runs, 3 documents were assumed relevant for term selection and document summaries comprised the best scoring 6 clusters. The rsv values to rank the potential expansion terms were estimated based on the top 20 ranked assumed relevant documents. The top 40 ranked expansion terms taken from the clusters were added to the original query in each case. Based on results from our previous experiments in CLEF, the original topic terms are up-weighted by a factor of 3.5 relative to terms introduced by PRF.

All Field Experiments Table 9 shows results for monolingual English retrieval based on all document fields using TDN field topics. Our sumitted run using this method is denoted by the *. Rather surprisingly the baseline result for this system is better than either of those using the field-weighted method shown in Table 1 and Table 3. The runs in Tables 1 and Table 3 are based

on only TD field topics, while those in Table 9 use TDN fields. However, the difference is probably to be too big to be explained by this factor alone. We will be investigating the reason for these differences. The summary-based PRF is shown to be effective here, as it has in many previous submissions to CLEF tracks in previous years.

Automatic Only Field Experiments Tables 10 and 9 show results for monolingual English retrieval based on only auto document fields using TDN and TD topic fields respectively. Results using TD topics are lower than those for TDN field topics. The summary-based PRF method is again effective for this document index. By contrast to the all field experiments, the results here are rather lower than those in Table 5 and Table 7 using the field weighted method. We will again be exploring the reasons for the differences.

5 Conclusions

This paper has described results for our participation in the CLEF 2006 CL-SR track. Experiments explored use of a field combination method for multi-field documents, and two methods of PRF. Results indicate that further exploration is required of the field combination approach and our new field-based PRF method. Our existing summary-based PRF method is shown to be effective for this task.

References

[1] Robertson, S. E., Zaragoza, H., and Taylor, M.: Simple BM25 Extension to Multiple Weighted Fields, Proceedings of the 13th ACM International Conference on Information and Knowledge Management, pages 42-49, 2004.

[2] Lam-Adesina, A. M., and Jones, G. J. F.: Dublin City University at CLEF 2005: Cross-Language Speech Retrieval (CL-SR) Experiments, Proceedings of the CLEF 2005: Workshop on Cross-Language Information Retrieval and Evaluation, Vienna, Austria, 2005.

[3] White, R. W., Oard, D. W., Jones, G. J. F., Soergel, D., and Huang, X.: Overview of the CLEF-2005 Cross-Language Speech Retrieval Track, Proceedings of the CLEF 2005: Workshop on Cross-Language Information Retrieval and Evaluation, Vienna, Austria, 2005.

[4] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu,M. M. and Gatford, M.:Okapi at TREC-3, Proceedings of the Third Text REtrieval Conference (TREC-3), pages 109-126. NIST, 1995.

[5] Porter, M. F.: An Algorithm for Suffix Stripping, Program, 14:10-137, 1980.

[6] Zhang, K.: Cross-Language Spoken Documenr Retrieval from Oral History Archives, MSc Dissertation, School of Computing, Dublin City University, 2006.

[7] Lam-Adesina, A. M., and Jones, G. J. F.: Applying Summarization Techniques for Term Selection in Relevance Feedback, Proceedings of the Twenty-Fourth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1-9, New Orleans, 2001. ACM.

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