Dublin City University at CLEF 2005: Cross-Language Spoken Document Retrieval (CL-SR) Experiments

Adenike M. Lam-Adesina, Gareth J. F. Jones
School of Computing, Dublin City University, Dublin 9, Ireland
{adenike,gjones}@computing.dcu.ie

Abstract

The Dublin City University participation in the CLEF CL-SR 2005 task concentrated on exploring the application of our existing information retrieval methods based on the Okapi model to the conversational speech data set. This required an approach to determining approximate sentence boundaries within the free-flowing automatic transcription provided. We also performed exploratory experiments on the use of the metadata provided with the document transcriptions. 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 - Relevance Feedback; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Cross-language spoken document retrieval, Transcription segmentation, Pseudo relevance feedback, Metadata combination

1 Introduction

The Dublin City University participation in the CLEF CL-SR 2005 task concentrated on exploring the application of our existing information retrieval methods based on the Okapi model for this data set, and exploratory experiments on the use of the provided document metadata. Our official submissions included both the English monolingual and French bilingual runs. This paper reports additional results for German and Spanish bilingual runs. Topics were translated into English using Systran V3.0 machine translation system. The resulting English topics were used for retrieving from the English document collection.

Our standard Okapi retrieval system incorporates a summary-based pseudo relevance feedback (PRF) stage. This PRF system operates by selecting topic expansion terms from document summaries, full details are described in [1]. However, since the automated transcriptions of the conversational speech documents do not contain punctuation, we needed to develop a method of selecting significant document segments. Details of this method are described in Section 2.1.

The document transcriptions are provided with a rich set of metadata. It is not immediately clear how best to exploit this most effectively in retrieval. This paper reports our initial exploratory experiments in making use of this additional information.

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

2 System Setup

The basis of our experimental system is the City University research distribution version of the Okapi system [2]. 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 standard set of synonyms. None of these procedures were adapted for the new CL-SR document set.

2.1 Term Weighting

Document terms were weighted using the Okapi BM25 weighting scheme developed in [2] calculated as follows,

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

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. Our values are tuned based on the training topics.

2.2 Pseudo-Relevance Feedback

We apply PRF for query expansion using a summary-based method described in [1] which has been shown to be effective in our previous submissions to CLEF, including [4] and elsewhere. The main challenge for query expansion is the selection of appropriate terms from the assumed relevant documents. Our query expansion method selects terms from summaries of the top ranked relevant document. All non-stopwords in the summaries are ranked using a slightly modified version of the Robertson selection value (rsv) [] shown in equation (2).

$$rsv(i) = r(i) \times rw(i) \qquad (2)$$

where r(i) = number of relevant documents containing term i, and rw(i) is the standard Robertson/Sparck Jones relevance weight [2],

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

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, and N = the total number of documents

The top ranked terms are then added to the topic. In our modified version of the *rsv(i)*, potential expansion terms are selected from the summaries of the top ranked documents, but ranked using statistics from top a large number of assumed relevant ranked documents from the initial run.

2.2.1 Sentence Selection

Our standard process for summary generation is to select representative sentences from the document. Since the document transcriptions do not contain punctuation marking we needed an alternative approach to identifying significant units in the transcription. We approached this using 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 strongly related. Strongly related word clusters were identified in the running document transcription by searching for word groups separated by not more than 5 insignificant words, as shown in Figure 1. Note that 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.

> … this chapter gives a brief description of the **[data** sets used in **evaluating** the **automatic** relevance **feedback** procedure **investigated** in this **thesis]** and also discusses the extension of …

Fig 1. Example of Sentence creation.

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 highest scoring cluster within a sentence. We adopted 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 (in this case 6)
TW = the total number of bracketed words (in this case 14)

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 form by summing these two measures for each sentence.

3 Experimental Investigation

This section describes the establishment of the parameters of our experimental system and gives results from our investigations.

3.1 Selection of System Parameters

In order to set the appropriate parameters for our feedback runs, we carried out development runs using the CLSR 2005 training topics. The Okapi parameters were set as follows k1=1.4 b=0.8. For all our PRF runs, 5 documents were assumed relevant for term selection and document summaries comprised the best scoring 4 clusters. The rsv values to rank the potential expansion terms were estimated based on the top 20 or 40 ranked assumed relevant documents. The top 20 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. For our runs we used either the Title section (dcu*tit) or the Title and Description (dcu*desc) section of each topic. Our official runs are marked *. Initial baseline monolingual results using English without query expansion are also given for comparison.

For our experiments the document fields were combined as follows,

dcua2 – combination of ASRTEXT2004A and AUTOKEYWORDA1
dcua1a2 – combination of ASRTEXT2004A, AUTOKEYWORDA1 and AUTOKEYWORDA2
dcusum – combination of ASRTEXT2004A, AUTOKEYWORDA1 and AUTOKEYWORDA2 and the SUMMARY
dcuall – combination of NAME, MANUALKEYWORD, SUMMARY and ASRTEXT2004A section of each documents.

3.2 Experimental Results

Tables 1-4 show results of our experiments using these different data combinations for the 25 test topics released for the CLEF 2005 CL-SR task. Results shown are Mean Average Precision (MAP), total relevant retrieved (Rr), and precision at cutoffs of 10 and 30 documents. Topic languages used are English, French, German and Spanish. Topics were translated into English using the Systran V3.0 machine translation system. The upper set of results in each table shows combined Title and Description topic queries and the lower set Title only topic queries.

Run-id	Topic Lang.	MAP	Rr	P10	P30
dcua2desc40f	Baseline	0.0496	536	0.1480	0.1027
	English	0.0654*	738	0.1760	0.1400
	French	0.0756	744	0.2080	0.1387
	German	0.0407	611	0.1160	0.0987
	Spanish	0.0549	727	0.1520	0.1093
dcua2tit40f	Baseline	0.0703	384	0.2280	0.1427
	English	0.0795	622	0.2520	0.1507
	French	0.0805	708	0.2520	0.1547
	German	0.0555	647	0.1840	0.1200
	Spanish	0.0681	602	0.1920	0.1293

Table 1: Results using a combination of ASRTEXT2004A and AUTOKEYWORDA1, the Title or Title and Description topic fields. Expansion terms ranked for selection using statistics of 40 top ranked documents.

Results in Table 1 show results for combination of ASRTEXT2004A with AUTOKEYWORDA1. It can be seen that the PSF method improves results for the English topics. Also that the results using Title only topics is better than using the combined Title and Description topics with respect to MAP. This result is perhaps a little surprising since the latter are generally found to be perform better and we are investigating the reasons for the results observed here. However, the number of relevant documents retrieved is generally higher when using the combined topics. Cross language results using French topics are shown to perform better than monolingual English for both MAP and relevant retrieved. This is again unusual, but not unprecedented. Results for translated German and Spanish show a reduction on the monolingual results.

Run-id	Topic Lang.	MAP	Rr	P10	P30
dcua1a2desc40f	Baseline	0.0464	500	0.1880	0.1053
	English	0.0670	784	0.1840	0.1480
	French	0.0943	773	0.2160	0.1707
	German	0.0455	611	0.0960	0.0920
	Spanish	0.0640	765	0.1640	0.1280
dcua1a2tit40f	Baseline	0.0796	472	0.2280	0.1600
	English	0.1101*	727	0.2520	0.1960
	French	0.1064*	768	0.2600	0.1907
	German	0.0740	691	0.1720	0.1493
	Spanish	0.0912	679	0.2200	0.1560

Table 2: Results using a combination of ASRTEXT2004A, AUTOKEYWORDA1 and AUTOKEYWORDA2, the Title or Title and Description topic fields. Expansion terms ranked for selection using statistics of 40 top ranked documents.

Table 2 shows results for the same experiments to those in Table 1 with the addition of AUTOKEYWORDA2 to the documents. Results here are generally show similar trends to those in Table 1 with small absolute increases in performance in most cases. In this case the performance advantage of French topics over English topics with PRF has largely disappeared.

Run-id	Topic Lang.	MAP	Rr	P10	P30
dcusumdesc40f	Baseline	0.1047	598	0.2240	0.1707
	English	0.1472	889	0.2720	0.2173
	French	0.1544	856	0.2600	0.2160
	German	0.1083	696	0.1640	0.1373
	Spanish	0.1074	860	0.1680	0.1520
Dcusumtit40f	Baseline	0.1407	618	0.2840	0.2160
	English	0.1672	770	0.2920	0.2427
	French	0.1654*	837	0.3080	0.2507
	German	0.1100	738	0.2200	0.1600
	Spanish	0.1542	736	0.2840	0.1296

Table 3: Results using a combination of ASRTEXT2004A, AUTOKEYWORDA1 and AUTOKEYWORDA2 and the SUMMARY section of each document, the Title or Title and Description topic fields. Expansion terms ranked for selection using statistics of 40 top ranked documents.

Table 3 shows results for a further set of experiments with the SUMMARY field added to the document descriptions. Results here show large increases compared to those in Table 2, indicating that the contents of the SUMMARY field are useful descriptions of the documents.

Run-id	Topic Lang.	MAP	Rr	P10	P30
Dcualldesc40f	Baseline	0.2213	1031	0.3680	0.2707
	English	0.3073	1257	0.4880	0.3773
	French	0.2762	1122	0.4960	0.3600
	German	0.2045	1001	0.3600	0.2760
	Spanish	0.2320	1160	0.3600	0.2680
Dcualltit40f	Baseline	0.2421	736	0.4120	0.3107
	English	0.2833	1009	0.4320	0.3373
	French	0.2569	1136	0.4240	0.3027
	German	0.2288	962	0.3280	0.2720
	Spanish	0.2468	908	0.3800	0.2973

Table 4: Results using a combination of NAME, MANUALKEYWORD, SUMMARY and ASRTEXT2004A section of each document, the Title or Title and Description topic fields. Expansion terms ranked for selection using statistics of 40 top ranked documents.

Table 4 shows a final set of experiments combining the NAME, MANUALKEYWORD and SUMMARY fields with ASRTEXT2004A. These results show large improvements over the results shown in previous tables. Performance for Title only and Title and Description combined topics is now similar with neither clearly showing an advantage. Monolingual English performance is clearly better than results for translated French topics, while our PRF method is still shown to be effective.

4 Conclusions and Further Work

Our initial experiments with the CLEF 2005 CL-SR task illustrate a number of points: PRF can be successfully applied to this data set, that the different fields of the document set make varying contributions to information retrieval effectiveness. In general in can be seen that manual assigned fields are more usefully than the automatically generated ones.

These experiments only represent a small subset of those that are possible with this dataset. In order to better understand the usefulness of document fields and retrieval methods more detailed analysis of these existing results and further experiments are planned.

References

[1] 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.

[2] 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.

[3] M.F. Porter. An algorithm for suffix stripping. Program, 14:10-137, 1980. H.P. Luhn. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 2(2):159-165, 1958.

[4] G.J.F.Jones, M.Burke, J.Judge, A.Khasin, A.Lam-Adesina and J.Wagner ,Dublin City University at CLEF 2004: Experiments in Monolingual, Bilingual and Multilingual Retrieval, Proceedings of the CLEF 2004: Workshop on Cross-Language Information Retrieval and Evaluation, Bath, U.K., 2004.

[5] 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.