DCU at CLEF 2006: ROBUST CROSS LANGUAGE TRACK

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

Abstract

The main focus of the DCU group's participation in the CLEF 2006 Robust Track in CLEF 2006 was not to identify and handle difficult topics in the topic set per se, but rather to explore a new method of re-ranking a retrieved document set. The initial query is used to re-rank documents retrieved using a query expansion method. The intention is to ensure that the query drift that might occur as a result of the addition of expansion terms chosen from irrelevant documents in pseudo relevance feedback (PRF) is minimised. By re-ranking using the initial query, the relevant set is forced to mimic the initial query more closely while not removing the benefits of PRF. Our results show that although our PRF is consistently effective for this task, the application of our re-ranking method generally has little effect on the ranked output.

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

Robust cross-language information retrieval, Pseudo relevance feedback, Document reranking

1 Introduction

This paper describes the DCU experiments for the CLEF 2006 Robust Track. Our official submission included monolingual runs for English and for Spanish, Italian and French where topics and documents had been translated into English and a bilingual run for Spanish using English topics. Unfortunately due to errors in our system we were unable to submit result for monolingual and bilingual German.

Our general approach was to translate non-English documents and topics into English for use as a pivot language. Collections and topics were translated into English using the Systran Version: 3.0 Machine Translator (Sys). Pseudo Relevance Feedback (PRF) which aims to expand query by selecting potential useful terms from the top retrieved documents to improve retrieval has been shown to be effective in our previous submissions to CLEF 2001-2005, and also in our other research work outside of CLEF. Therefore, we again use this method with our extended PRF method of term selection from document summaries rather than full documents that has been thoroughly tested in our past research work. In addition, for this task we explored the application of a new post-retrieval re-ranking method that we are developing.

The remainder of this paper is structured as follows: Section 2 covers background to robust information retrieval tasks, Section 3 describes our system setup and the information retrieval (IR) methods used, Section 4 presents our experimental results and section 5 concludes the paper with a discussion of our findings.

2 Background

The robust track was first introduced in the Text Retrieval Conference (TREC) in 2003. The aim was to explore methods of improving retrieval effectiveness for topics that performed poorly using standard generally high performing IR methods, i.e. hard topics. For these topics it is usually the case that although relevant documents exist in the target collection, the topic is not discriminatory enough to find the relevant documents or bring them into the retrieved set of potentially relevant documents. Several approaches have been taken by TREC participants which aim to tackle these hard topics and improve IR effectiveness. This work falls into two main categories: either using a contemporaneous collection (e.g. the web) for query expansion or re-ordering the original ranking of the retrieved relevant documents.

Kwok. et al. used the web as a contemporaneous collection from which terms were selected for query expansion [1]. They argued that the reason why PRF is not effective for hard topics is because assumed relevant documents where the expansion are taken from, are usually irrelevant and thus would cause a query drift for hard topics. Therefore they expand the initial query from the web and use the expanded query for retrieval. The list from the initial retrieval step and the expanded query list are then combined into a new list. Results for this approach showed an improvement in IR performance for both normal and hard topics. Interestingly results for runs using short queries were found to be better than those for long queries.

Amati et al. also found that query expansion from the web resulted in better retrieval for hard topics as long as the queries are short [2]; for longer queries PRF should be limited to the target collection.

Piatko et al. used a re-ranking method that aimed to improve the initial ranking of retrieved relevant documents using a method called the minimal matching span [3]. This method aims to improve the ranking of relevant documents by estimating the minimal length of consecutive sets of document terms containing at least one occurrence of each query term in the set. Documents with high scores have their ranking improved. Results for this method showed an improvement in average precision results compared to not re-ranking. The benefits of this re-ranking method were more visible at the top ranks of the retrieved document set.

3 System Setup

For our experiments we used the City University research distribution version of the Okapi system retrieval system. Stopwords were removed from both the documents and search topics, and the Okapi implementation of Porter stemming algorithm [4] was applied to both the document and search terms.

3.1 Term Weighting

The okapi system is based on the BM25 weighting [5] scheme where document terms are weighted as follows,

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

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 2003 ad hoc retrieval task data.

3.2 Pseudo-Relevance Feedback

Short and imprecise queries can affect IR effectiveness. To curtail this negative impact, relevance feedback (RF) via query expansion (QE) is often employed. QE aims to improve initial query statements by addition of terms from user assessed relevant documents. These terms are selected using document statistics and usually describe the information request better. Pseudo-Relevance Feedback (PRF) whereby relevant documents are assumed and used for QE is on average found to give improvement in retrieval performance, although this is usually smaller than that observed for true user-based RF.

PRF can result in a query drift if expansion terms are selected from assumed relevant document which are in fact not relevant. In our past research work [6] we discovered that although a top-ranked document might not be relevant, it often contains information that is pertinent to the query. Thus, we developed a new method that select appropriate terms from document summaries. These summaries are constructed in such a way that they contain only sentences that are closely related to the initial query. Our QE method selects terms from summaries of the top 5 ranked documents. The summaries are generated using the method described in [6]. 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) [5] 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) is the standard Robertson/Sparck Jones relevance weight [5] reproduced below

$$rw(i) = \log \frac{(r(i)+0.5)(N - n(i) - R + r(i) + 0.5)}{(n(i) - r(i) + 0.5)(R - r(i) + 0.5)} \tag{3}$$

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.

3.3 Re-ranking Methodology

As part of our investigation for the CLEF 2006 robust track we explored the application of a further novel re-ordering of the retrieved document list obtained from our PRF process. This reordering method attempts to ensure that retrieved documents with more matching query terms have their ranking improved, while not discarding the effect of document weighting scheme used. To this end we devised a document re-ranking formula as follows:

$$\frac{doc\_wgt}{(1-b)+(b*nmt/mmt)} \tag{4}$$

where doc_wgt = the original document matching score
b = an empirical value ranging between 0.1 and 0.5
nmt = the number of original topic terms that occur in the document
mmt = the mean of the value nmt for a given query over all retrieved documents

4 Experimental results

In this section we describe our parameter selection and present our experimental results for the CLEF 2006 Robust track. Results are given for baseline retrieval without feedback, after the application of our PRF method and after the further application of our re-ranking procedure.

The CLEF 2006 topics consist of three fields: Title, Description and Narrative. We conducted experiments used the Title and Description (TD) or Title, Description and Narrative (TDN) fields. For all runs we present the precision at both 10 and 30 documents cutoff (P10 and P30), standard TREC average precision results (AvP), the number of relevant documents retrieved out of the total number of relevant in the collection (RelRet), and the change in number of RelRet compared to Baseline runs.

4.1 Selection of System Parameters

To set appropriate parameters for our runs development runs were carried out using the training topics provided. The topics provided were taken from the CLEF 2003 The Okapi parameters were set as follows k1=1.2 b=0.75. 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 were 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.

4.2 Experimental Results

Table 1 summarises the results of our experiments. Results are shown for the following runs:

Baseline – baseline results without PRF using Title, Description and Narrative topic fields (TDN)
f20narr – feedback results using the Title, Description and Narrative topic fields. 20 terms are added to the initial query.
f20re-ranked - same as F20narr, but documents are re-ranked using the formula (4) above.
f20desc – feedback results using the Title and Description sections of query. 20 terms are added to the initial query.

Comparing the Baseline and f20narr runs it can be seen that application of PRF improves all the performance measures for all runs with the exception of the RelRet for Spanish monolingual where there is a small reduction. By contrast for the Spanish bilingual run there is a much larger improvement in RelRet than is observed for any of the other runs.

Application of the re-ranking method to the f20narr list produces little change in the ranked output. The only notable change is a further improvement in the RelRet for the Spanish bilingual task. Varying the value of the b factor in equation 4 made only a small difference to the results. We are currently investigating the reasons for this results, and exploring approaches to the re-ranking method which will have a greater impact on the output ranked lists.

5 Conclusions

This paper has presented a summary of our results for the CLEF 2006 Robust Track. The results show that our summary-based PRF method is consistently effective across this topic set. We also explored the use of a novel post-retrieval re-ranking method. Application of this procedure led to very modification in the ranked lists, and we are currently exploring alternative variations on this method.

Run-ID	English	French	Spanish	Italian	Spanish bi
Baseline (TDN)					
P10	422	395	485	382	357
P30	265	269	351	262	266
Av.P	544	470	445	388	314
RelRet	1496	2065	4468	1736	3702
f20narr (TDN)					
P10	436	425	507	434	413
P30	276	294	375	296	300
Av.P	558	504	478	459	357
RelRet	1508	2091	4413	1779	3856
Chg RelRet	+12	+26	-55	+43	+154
f20re-ranked (TDN)					
P10	433	424	509	434	407
P30	276	295	377	296	298
Av.P	558	508	480	459	358
RelRet	1507	2092	4426	1783	3900
Chg RelRet	+11	+27	-42	+47	+198
f20desc (TD)					
P10	396	370	450	398	386
P30	261	272	358	279	288
Avep	494	452	435	419	343
RelRet	1493	2074	4474	1778	3759
Chg RelRet	+3	+9	+6	+42	+57

Table 1: Retrieval results for Baseline, PRF and re-ranked results for the CLEF 2006 Robust track for monolingual English, monolingual French, Spanish and Italian with document and topic translation to English, and Spanish bilingual with document translation to English.

References

1. K.L. Kwok, L. Grunfeld, H.L. Sun and P. Deng. TREC2004 Robust Track Experiments using PIRCS, Proceedings of TREC 2004, NIST, 2004.
2. G. Amati, C. Carpineto, and G. Romano. Fondazione Ugo Bordoni at TREC 2004, Proceedings of TREC 2004, NIST, 2004.
3. Christine Piatko, James Mayfield, Paul McNamee, and Scott Cost JHU/APL at TREC 2004: Robust and Terabyte Tracks, Proceedings of TREC 2004, NIST, 2004.
4. M.F. Porter. An algorithm for suffix stripping. Program, 14:10-137, 1980.
5. 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.
6. 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.