EXETER at CLEF 2003: Cross-Language Spoken Document Retrieval Experiments

Adenike Lam-Adesina, Gareth J. F. Jones*
Department of Computer Science, University of Exeter, EX4 4QF, U.K.
Email: {A.M.Lam-Adesina,G.J.F.Jones}@ex.ac.uk

Abstract

Cross-Language Spoken Document Retrieval (CLSDR) combines both the complexities of retrieval from collections characterized by speech transcription errors and language translation issues between search requests and documents. Thus achieving effective retrieval in this domain is potentially very challenging. For the CLEF 2003 SDR task we adopted a standard query translation strategy using commercial machine translation tools.

1 Introduction

Both Cross-Language Information Retrieval (CLIR) and Spoken Document Retrieval (SDR) are affected by limitations in language processing technologies. In the case of the former this relates to translation between the languages of the document collection and in the latter to the difficulties encountered in transcription of spoken data. These issues are analyzed in more details in [1]. Spoken Document Retrieval (CLSDR) combines both the difficulties of both CLIR and SDR. Thus retrieval in this domain is very challenging.
For CLEF 2003 CLSDR task we adopted a query translation strategy and investigated the use of a large text collection to augmented the spoken document test set. All query statements were translated from the source language into English using two machine translation tools: Systran Version:3.0 (SYS) and Globalink Power Translation Pro Version 6.4 (PRO) Machine Translator (MT) systems.
The remainder of this paper summarizes are retrieval system and gives results and initial analysis of our experimental results.

2 System Setup

The basis of the experimental system was the same as that used for our submissions to the monolingual, bilingual and multilingual tasks for CLEF 2003. The system combines Okapi BM25 term weighting with pseudo relevance feedback (PRF), and standard procedures of stop word removal and Porter stemming. Full details are given in [2]. The parameters of the PRF system were set identically to those for the text retrieval system given in [2]. The Okapi parameters *K1* and *b* were optimized for the SDR test collection.

3 Merged collections

In our experiments for the CLSDR pilot track held at CLEF 2002 we experimented with the combination of the test collection with a small contemporaneous text document collection for term weight estimation [3]. This method aims to improve retrieval performance for the test set by better estimated of term weights. Our results for CLEF 2002 indicated that the method can give improvements in retrieval performance even when using only a small number of additional documents. Results for ITC-irst however showed that large improvements can be realized if a much larger number of contemporaneous documents is used [4]. However, this large collection of truly contemporaneous documents was not available to us. This led us to investigate the use of an alternative large text document collection. In this case we used the document set from the TREC-8 and TREC-9 ad hoc retrieval tasks. This consists of around 500,000 text documents. In addition, we used again used the two small collections of truly contemporaneous text documents. These sources are taken from New York Times Newswire Service (excluding non-NYT sources) and Associated Press Worldstream Service (English content only), totaling about 20,000 news stories, and are taken from exactly the same period as the spoken document test collection. These three text collections were merged collection into a single collection which was used as the pilot collection from which initial query statements are expanded in experiments reported in the next section.

4 Experimental Results

* now at School of Computing, Dublin City University, Ireland email: Gareth.Jones@computing.dcu.ie

This section describes our results for the CLEF 2003 SDR Tasks. We report baseline and feedback results for five topic languages, English, French, German, Italian and Spanish. Our results include runs for topic translations using both SYS and PRO MT systems. Results are presented for the following methods:

1. Baseline run without feedback (exebase)
2. Feedback runs using expanded query from the test collection (exepl)
3. Feedback runs using expanded query from the pilot collection and term weight estimated from the test collection. Initial query terms are upweighted by multiply by 1.5 (exeprn1.5)
4. Same as 3 but initial query terms are upweighted by 3.5 (exeprn3.5)

SYS MT	English	French	German	Italian	Spanish
exebase	311	227	203	231	250
exepl	382	281	270	279	292
Rel_ret	1795	1558	1498	1638	1641
% chg	22.8%	23.8%	33.0%	20.7%	16.8%
exeprn1.5	364	283	274	299	304
Rel_ret	1824	1618	1541	1684	1720
% chg	17.0%	24.7%	34.9%	29.4%	21.6%
exeprn3.5	371	276	268	296	307
Rel_ret	1789	1577	1524	1653	1707
% chg	19.3%	21.6%	32.0%	28.1%	22.8%

Table 1: Average precision retrieval results for topic translation using SYS MT before and after application of different feedback methods

PRO MT	English	French	German	Italian	Spanish
Exebase	311	235	188	234	235
exeprn1.5	364	262	242	301	315
Rel_ret	1824	1589	1431	1624	1710
% chg	17.0%	11.5%	28.7%	28.6%	34.0%
exeprn3.5	371	256	229	293	308
Rel_ret	1789	1574	1420	1602	1682
% chg	19.3%	8.9%	21.8%	25.2%	31.1%

Table 2: Average precision retrieval results for topic translation using PRO MT before and after application of different feedback methods

Results for out CLSDR runs are shown in Tables 1 and 2 for Systran and Power Translator Pro MT respectively. It can be seen that as expected the monolingual English result is the best in all cases with respect to both average precision and number of relevant documents retrieved. CLSDR performance is comparable for the French, Italian and Spanish topic statements with lower results for the German topics. This result is a little surprising for Systran French topic translation which has previously been shown to be more effective than other topic translations in our CLEF bilingual text retrieval experiments [5]. PRF using only the test collection is observed to be effective for query expansion in all cases. Results for query expansion using the pilot strategy are more mixed. In the case of Italian and Spanish topics this approach clearly outperforms test collection only query expansion. However, there is little difference between the results for these methods when using French and German topics.

5 Conclusions and Further Work

The results for the CLEF 2003 CLSDR task reported in this paper establish baseline performance figures against which the exploration of techniques for CLSDR can be measured. The experiments reported here show that PRF is effective for this task, as would be expected since it is generally a useful techniques for text CLIR and SDR. The use of large additional test collections for parameter estimation for query expansion can produce improvements in performance over test collection only based expansion, but cannot be relied upon to do so. While there is clearly scope to develop a more detailed investigation of the interaction of translation and indexing errors, an initial further set of experiments is planned

using the pilot collection weights in the final retrieval phase. This technique was observed to be effective for our small pilot collection in previous experiments [3].

References

[1] G.J.F. Jones and M.Federico, CLEF 2002 Cross-Language Spoken Document Retrieval Pilot Track Report, In *Proceedings of the CLEF 2002: Workshop on Cross-Language Information Retrieval and Evaluation*, Rome, September 2002. Springer Verlag.

[2] 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 the CLEF 2003: Workshop on Cross-Language Information Retrieval and Evaluation*, Trondheim, August 2003.

[3] G.J.F. Jones and A.M. Lam-Adesina, Exeter at CLEF 2002: Cross-Language Spoken Document Retrieval Experiments, In *Proceedings of the CLEF 2002: Workshop on Cross-Language Information Retrieval and Evaluation*, Rome, September 2002. Springer Verlag.

[4] N.Bertoldi and M.Federico, Cross-Language Spoken Document Retrieval on the TREC SDR Collection. In *Proceedings of the CLEF 2002: Workshop on Cross-Language Information Retrieval and Evaluation*, Rome, September 2002. Springer Verlag.

[5] G.J.F. Jones and A.M. Lam-Adesina. Exeter at CLEF 2001: Experiments with Machine Translation for Bilingual Retrieval. In Proceedings of the CLEF 2001: Workshop on Cross-Language Information Retrieval and Evaluation, pages 59-77, Darmstadt, Germany, 2001.