Experiments for the Cross Language Speech Retrieval Task at CLEF 2006

Muath Alzghool and Diana Inkpen

School of Information Technology and Engineering
University of Ottawa
{alzghool,diana}@site.uottawa.ca

Abstract. This paper presents the second participation of the University of Ottawa group in the Cross-Language Speech Retrieval (CL-SR) task at CLEF 2006. We present the results of the submitted runs for the English collection and very briefly for the Czech collection, followed by many additional experiments. We have used two Information Retrieval systems in our experiments: SMART and Terrier, with several query expansion techniques (including a new method based on log-likelihood scores for collocations). Our experiments showed that query expansion methods do not help much for this collection. We tested different Automatic Speech Recognition transcripts and combinations. The retrieval results did not improve, probably because the speech recognition errors happened for the words that are important in retrieval. We present cross-language experiments, where the queries are automatically translated by combining the results of several online machine translation tools. Our experiments showed that high quality automatic translations (for French) led to results comparable with monolingual English, while the performance decreased for the other languages. Experiments on indexing the manual summaries and keywords gave the best retrieval results.

1 Introduction

This paper presents the second participation of the University of Ottawa group in the Cross-Language Speech Retrieval (CL-SR) track at CLEF 2006. We briefly describe the task [8]. Then, we present our systems, followed by results for the submitted runs for the English collection and very briefly for the Czech collection. We present results for many additional runs for the English collection. We experimented with many possible weighting schemes for indexing the documents and the queries, and with several query expansion techniques. We tested with different speech recognition transcripts to see if the word error rate has an impact on the retrieval performance. We describe cross-language experiments, where the queries are automatically translated from French, Spanish, German and Czech into English, by combining the results of several online machine translation (MT) tools. At the end we present the best results, when summaries and manual keywords were indexed.

2 System Description

The University of Ottawa Cross-Language Information Retrieval (IR) systems were built with off-the-shelf components. For translating the queries from French, Spanish,

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 778 – 785, 2007.

German, and Czech into English, several free online machine translation tools were used. The idea behind using multiple translations is that they might provide more variety of words and phrases, therefore improving the retrieval performance. Seven online MT systems [4] were used for translating from Spanish, French, and German. We combined the outputs of the MT systems by simply concatenating all the translations. All seven translations of a title made the title of the translated query; the same was done for the description and narrative fields. We used the combined topics for all the cross-language experiments reported in this paper. For translation of the Czech language topics into English we were able to find only one online MT system.

For the retrieval part, the SMART [2,11] IR system and the Terrier [1,9] IR system were tested with many different weighting schemes for indexing the collection and the queries.

SMART was originally developed at Cornell University in the 1960s. SMART is based on the vector space model of information retrieval. we used mainly the lnn.ntn weighting scheme [2 ,11] which performs very well in CLEF-CLSR 2005 [4] .

We have also used a query expansion mechanism with SMART, which follows the idea of extracting related words for each word in the topics using the Ngram Statistics Package (NSP) [10]. We extracted the top 6412 pairs of related words based on log likelihood ratios (high collocation scores in the corpus of ASR transcripts), using a window size of 10 words. We chose log-likelihood scores because they are known to work well even when the text corpus is small. For each word in the topics, we added the related words from this list of pairs. We call this approach SMARTnsp.

Terrier was originally developed at University of Glasgow. It is based on Divergence from Randomness models (DFR) where IR is seen as a probabilistic process [1, 9]. We experimented with the In(exp)C2 weighting model, one of Terrier's DFR-based document weighting models.

We have also used a query expansion mechanism in Terrier, which follows the idea of measuring divergence from randomness. In our experiments, we applied the Kullback-Leibler (KL) model for query expansion [3, 9].

3 Experimental Results

3.1 Submitted Runs

Table 1 shows the results of the submitted results on the test data (33 queries). The evaluation measure we report is the standard measure computed with the trec_eval script (version 8): MAP (Mean Average Precision). The information about what fields of the topic were indexed is given in the column named Fields: T for title only, TD for title + description, TDN for title + description + narrative. For each run we include an additional description of the experimental settings and which document fields were indexed. For the uoEnTDt04A06A and uoEnTDNtMan runs we used the indexing scheme In(exp)C2 from Terrier; and for uoEnTDNsQEx04, uoFrTDNs, and uoSpTDNs we used the indexing scheme lnn.ntn from SMART. We used SMARTnsp query expansion for the uoEnTDNsQEx04 run, KL query expansion for

uoEnTDNtMan and uoEnTDt04A06A, and we didn't use any query expansion techniques for uoFrTDNs and uoSpTDNs.

Our required run, English TD (0.0565), obtained a lower result than our automatic English TDN run (0.0768), mainly due to different system settings, not due to the additional field N. Comparing the result of our required run to the best required run, submitted by Dublin City University (dcuEgTDauto, 0.0733) [6], their result was better with relative improvement 30%; but we obtained a comparable result using SMART with blind relevance feedback (SMARTnsp, 0.0754 – more details are given in section 3.2); our result was better with relative improvement 2%.

Table 1. Results of the five submitted runs, for topics in English, French, and Spanish. The required run (English, title + description) is in bold.

Runs for English	**MAP**	**Fields**	**Description**
uoEnTDNtMan	0.2902	TDN	Terrier: MANUALKEYWORD + SUMMARY
uoEnTDNsQEx04	0.0768	TDN	SMART: NSP query expansion ASRTEXT2004A + AUTOKEYWORD2004A1, A2
uoFrTDNs	0.0637	TDN	SMART: ASRTEXT2004A + AUTOKEY-WORD2004A1, A2
uoSpTDNs	0.0619	TDN	SMART: ASRTEXT2004A + AUTOKEY-WORD2004A1, A2
uoEnTDt04A06A	**0.0565**	TD	Terrier: ASRTEXT2004A + ASRTEXT2006A + AUTOKEYWORD2004A1, A2

We also participated in the task for Czech language. We indexed the Czech topics and ASR transcripts. Table 2 shows the results of the submitted runs on the test data (29 topics) for the Czech collection. The evaluation measure we report is the mean Generalized Average Precision (mGAP), which rewards retrieval of the right time-stamps in the collection. MAP scores could not be used because the speech transcripts were not segmented. We used the quickstart collection provided: each document contains 4-minute passages that start each minute (overlapping passages). From our results, we note:

- The mGAP is substantially low for all submitted runs.
- There is a small improvement when we indexed the field ENGLISH-MANUKEYWORD relative to the case when we indexed CZECHMANUKEY-WORD.
- We got small improvements if CZECHMANUKEYWORD was added to the ASR field.
- Terrier's results are slightly better than SMART's for the required run.

Comparing our best run (0.0235) to the best run by University of West Bohemia team (0.0456) [5], our results were lower because we did not use any Czech-specific processing in our system (such as removing stop words or stemming), while in [5] this was done. In the rest of the paper we focus only on the Eglish CL-SR collection.

Table 2. Results of the five submitted runs for Czech collection. The required run (title + description) is in bold.

Runs for Czech	mGAP	Fields	Description
uoC-zEnTDNsMan	0.0235	TDN	SMART: ASRTEXT, CZECHAUTOKEYWORD, CZECHMANUKEYWORD, ENGLISH MANUKEYWORD, ENGLISHAUTOKEYWORD
uoCzTDNsMan	0.0200	TDN	SMART: ASRTEXT, CZECHAUTOKEYWORD, CZECHMANUKEYWORD
uoCzTDNs	0.0182	TDN	SMART: ASRTEXT, CZECHAUTOKEYWORD
uoCzTDs	**0.0211**	TD	SMART: ASRTEXT, CZECHAUTOKEYWORD
uoCzEnTDt	**0.0218**	TD	Terrier: ASRTEXT, CZECHAUTOKEYWORD

3.2 Comparison of Systems and Query Expansion Methods

Table 3 presents results for the best weighting schemes: for SMART we chose lnn.ntn and for Terrier we chose the In(exp)C2 weighting model, because they achieved the best results on the training data. We present results with and without relevance feedback. According to Table 3, we note that:

- Blind relevance feedback helps to improve the retrieval results in Terrier for TDN, TD, and T for the training data; the improvement was high for TD and T, but not for TDN. For the test data there is a small improvement.
- NSP relevance feedback with SMART does not help to improve the retrieval for the training data (except for TDN); the improvement on the test data was small.
- SMART results are better than Terrier results for the test data, but not for the training data.

Table 3. Results (MAP scores) for Terrier and SMART, with or without relevance feedback, for English topics. In bold are the best scores for TDN, TD, and T.

	System	**Training**			**Test**		
		TDN	**TD**	**T**	**TDN**	**TD**	**T**
1	SMART	**0.0954**	0.0906	0.0873	0.0766	0.0725	0.0759
	SMARTnsp	0.0923	0.0901	0.0870	**0.0768**	**0.0754**	**0.0769**
2	Terrier	0.0913	0.0834	0.0760	0.0651	0.0560	0.0656
	TerrierKL	0.0915	**0.0952**	**0.0906**	0.0654	0.0565	0.0685

3.3 Comparison of Retrieval Using Various ASR Transcripts

In order to find the best ASR transcripts to use for indexing the segments, we compared the retrieval results when using the ASR transcripts from the years 2003, 2004, and 2006 or combinations. We also wanted to find out if adding the automatic keywords helps to improve the retrieval results. The results of the experiments using Terrier and SMART are shown in Table 4 and Table 5, respectively. We note from the experimental results that:

Table 4. Results (MAP scores) for Terrier, with various ASR transcript combinations. In bold are the best scores for TDN, TD, and T.

Segment fields	Terrier					
	Training			**Test**		
	TDN	TD	T	TDN	TD	T
ASRTEXT 2003A	0.0733	0.0658	0.0684	0.0560	0.0473	0.0526
ASRTEXT 2004A	0.0794	0.0742	0.0722	**0.0670**	0.0569	0.0604
ASRTEXT 2006A	0.0799	0.0731	0.0741	0.0656	0.0575	0.0576
ASRTEXT 2006B	0.0840	0.0770	0.0776	0.0665	**0.0576**	0.0591
ASRTEXT 2003A+2004A	0.0759	0.0722	0.0705	0.0596	0.0472	0.0542
ASRTEXT 2004A+2006A	0.0811	0.0743	0.0730	0.0638	0.0492	0.0559
ASRTEXT 2004A+2006B	0.0804	0.0735	0.0732	0.0628	0.0494	0.0558
ASRTEXT 2003A+ AUTOKEYWORD2004A1,A2	0.0873	0.0859	0.0789	0.0657	0.0570	0.0671
ASRTEXT 2004A+ AUTOKEYWORD2004A1, A2	0.0915	**0.0952**	0.0906	0.0654	0.0565	0.0685
ASRTEXT 2006B+ AUTOKEYWORD2004A1,A2	**0.0926**	0.0932	0.0909	0.0717	0.0608	0.0661
ASRTEXT 2004A+2006A+ AUTOKEYWORD2004A1, A2	0.0915	**0.0952**	**0.0925**	0.0654	0.0565	**0.0715**
ASRTEXT 2004A+2006B+ AUTOKEYWORD2004A1,A2	0.0899	0.0909	0.0890	0.0640	0.0556	0.0692

Table 5. Results (MAP scores) for Terrier, with various ASR transcript combinations. In bold are the best scores for TDN, TD, and T.

Segment fields	SMART					
	Training			**Test**		
	TDN	TD	T	TDN	TD	T
ASRTEXT 2003A	0.0625	0.0586	0.0585	0.0508	0.0418	0.0457
ASRTEXT 2004A	0.0701	0.0657	0.0637	0.0614	0.0546	0.0540
ASRTEXT 2006A	0.0537	0.0594	0.0608	0.0455	0.0434	0.0491
ASRTEXT 2006B	0.0582	0.0635	0.0642	0.0484	0.0459	0.0505
ASRTEXT 2003A+2004A	0.0685	0.0646	0.0636	0.0533	0.0442	0.0503
ASRTEXT 2004A+2006A	0.0686	0.0699	0.0696	0.0543	0.0490	0.0555
ASRTEXT 2004A+2006B	0.0686	0.0713	0.0702	0.0542	0.0494	0.0553
ASRTEXT 2003A + AUTOKEYWORD2004A1,A2	0.0923	0.0847	0.0839	0.0674	0.0616	0.0690
ASRTEXT 2004A+ AUTOKEYWORD2004A1,A2	**0.0954**	0.0906	0.0873	**0.0766**	**0.0725**	**0.0759**
ASRTEXT 2006B+ AUTOKEYWORD2004A1,A2	0.0869	0.0892	0.0895	0.0650	0.0659	0.0734
ASRTEXT 2004A+ 2006A + AUTOKEYWORD2004A1,A2	0.0903	**0.0932**	0.0915	0.0654	0.0654	0.0777
ASRTEXT 2004A +2006B + AUTOKEYWORD2004A1,A2	0.0895	0.0931	**0.0919**	0.0652	0.0655	0.0742

- Using Terrier, the best field is ASRTEXT2006B which contains 7377 transcripts produced by the ASR system on 2006 and 727 transcripts produced by the ASR system in 2004, this improvement over using only the ASRTEXT2004A field is

very small. On the other hand, the best ASR field using SMART is ASRTEXT2004A.

- Any combination between two ASRTEXT fields does not help.
- Using Terrier and adding the automatic keywords to ASRTEXT2004A improved the retrieval for the training data but not for the test data. For SMART it helps for both the training and the test data.
- In general, adding the automatic keywords helps. Adding them to ASRTEXT2003A or ASRTEXT2006B improved the retrieval results for the training and test data.
- For the required submission run English TD, the maximum MAP score was obtained by the combination of ASRTEXT 2004A and 2006A plus autokeywords using Terrier (**0.0952)** or SMART (**0.0932**) on the training data; on the test data the combination of ASRTEXT 2004A and autokeywords using SMART obtained the highest value, **0.0725**, higher than the value we report in Table 1 for the submitted run.

3.4 Cross-Language Experiments

Table 6 presents results for the combined translation produced by the seven online MT tools, from French, Spanish, and German into English, for comparison with monolingual English experiments (the first line in the table). All the results in the table are from SMART using the lnn.ntn weighting scheme.

Since the result of combined translation for each language was better than when using individual translations from each MT tool on the CLEF 2005 CL-SR data [4], we used combined translations in our experiments.

The retrieval results for French translations were very close to the monolingual English results, especially on the training data. On the test data, the results were much worse when using only the titles of the topics, probably because the translations of the short titles were less precise. For translations from the other languages, the retrieval results deteriorate rapidly in comparison with the monolingual results. We believe that the quality of the French-English translations produced by online MT tools was very good, while the quality was lower for Spanish, German and Czech, successively.

Table 6. Results of the cross-language experiments, where the indexed fields are ASRTEXT2004A, and AUTOKEYWORD2004A1, A2 using SMART (lnn.ntn).

Language	**Training**			**Test**		
	TDN	**TD**	**T**	**TDN**	**TD**	**T**
English	0.0954	0.0906	0.0873	0.0766	0.0725	0.0759
French	0.0950	0.0904	0.0814	0.0637	0.0566	0.0483
Spanish	0.0773	0.0702	0.0656	0.0619	0.0589	0.0488
German	0.0653	0.0622	0.0611	0.0674	0.0605	0.0618
Czech	0.0585	0.0506	0.0421	0.0400	0.0309	0.0385

3.5 Manual Summaries and Keywords

Table 7 presents the results when only the manual keywords and the manual summaries were used. The retrieval performance improved a lot, for topics in all the languages. The MAP score jumped from 0.0654 to 0.2902 for English test data, TDN, with the In(exp)C2 weighting model in Terrier. The results of cross-language experiments on the manual data show that the retrieval results for combined translation for French and Spanish language were very close to the monolingual English results on training data and test data. For all the experiments on manual summaries and keywords, Terrier's results are better than SMART's.

Our results on manual summaries and keywords for TD and TDN (0.2902, 0.2710) were better than the submitted runs by Dublin City University (0.2765, 0.2015) [6], the relative improvements for TD and TDN were 5% and 34% respectively.

Table 7. Results of indexing the manual keywords and summaries, using SMART with weighting scheme lnn.ntn, and Terrier with (In(exp)C2).

Language and System	**Training**			**Test**		
	TDN	**TD**	**T**	**TDN**	**TD**	**T**
English SMART	0.3097	0.2829	0.2564	0.2654	0.2344	0.2258
English Terrier	**0.3242**	**0.3227**	**0.2944**	**0.2902**	**0.2710**	**0.2489**
French SMART	0.2920	0.2731	0.2465	0.1861	0.1582	0.1495
French Terrier	0.3043	0.3066	0.2896	0.1977	0.1909	0.1651
Spanish SMART	0.2502	0.2324	0.2108	0.2204	0.1779	0.1513
Spanish Terrier	0.2899	0.2711	0.2834	0.2444	0.2165	0.1740
German SMART	0.2232	0.2182	0.1831	0.2059	0.1811	0.1868
German Terrier	0.2356	0.2317	0.2055	0.2294	0.2116	0.2179
Czech SMART	0.1766	0.1687	0.1416	0.1275	0.1014	0.1177
Czech Terrier	0.1822	0.1765	0.1480	0.1411	0.1092	0.1201

4 Conclusion

We experimented with two different systems: Terrier and SMART, with various weighting schemes for indexing the document and query terms. We proposed a new approach for query expansion that uses collocations with high log-likelihood ratio. Used with SMART, the method obtained a small improvement on test data (not statistically significant according to a Wilcoxon signed test). The KL blind relevance feedback method produced only small improvements with Terrier on test data. So, query expansion methods do not seem to help for this collection.

The improvements of mean word error rates in the ASR transcripts (of ASRTEXT2006A relative to ASRTEXT2004A) did not improve the retrieval results. Also, combining different ASR transcripts (with different error rates) did not help.

For some experiments, Terrier was better than SMART, for others it was not; therefore we cannot clearly choose one or another IR system for this collection.

The idea of using multiple translations proved to be good. More variety in the translations would be beneficial. The online MT systems that we used are rule-based

systems. Adding translations by statistical MT tools might help, since they could produce radically different translations.

On the manual data, the best MAP score we obtained is 0.2902, for the English test topics. On automatically-transcribed data the best result is 0.0766 MAP score. Since the improvement in the ASR word error rate does not improve the retrieval results, as shown from the experiments in section 3.3, we think that the justification for the difference to the manual summaries is due to the fact that summaries contain different words to represent the content of the segments. In future work we plan to investigate methods of removing or correcting some of the speech recognition errors in the ASR contents and to use speech lattices for indexing.

References

1. Amati, G., van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems 20(4), 357–389 (2002)
2. Buckley, C., Salton, G., Allan, J.: Automatic retrieval with locality information using SMART. In: Text REtrieval Conference (TREC-1), pp. 59–72 (1993)
3. Carpineto, C., de Mori, R., Romano, G., Bigi, B.: An information-theoretic approach to automatic query expansion. ACM Transactions on Information Systems (TOIS) 19(1), 1–27 (2001)
4. Inkpen, D., Alzghool, M., Islam, A.: Using various indexing schemes and multiple translations in the CL-SR task at CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
5. Ircing, P., Müller, L.: The University of West Bohemia at CLEF 2006, the CL-SR track. In: Evaluation of Multilingual and Multi-modal Information Retrieval Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 19-21 (2006)
6. Jones, G.J.F., Zhang, K., Lam-Adesina, A.M.: Dublin City University at CLEF 2006: Cross-Language Speech Retrieval (CL-SR) Experiments. In: Evaluation of Multilingual and Multi-modal Information Retrieval Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006, September 19-21, Alicante, Spain (2006)
7. Oard, D.W., Soergel, D., Doermann, D., Huang, X., Murray, G.C., Wang, J., Ramabhadran, B., Franz, M., Gustman, S.: Building an Information Retrieval Test Collection for Spontaneous Conversational Speech. In: Proceedings of SIGIR (2004)
8. Oard, D.W., Wang, J., J.,, Jones, G.J.F., White, R.W., Pecina, P., Soergel, D., Huang, X., Shafran, I.: Overview of the CLEF 2006 cross-language speech retrieval track. In: Working Notes of the CLEF- 2006 Evaluation, Alicante, Spain, p. 12 (2006)
9. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier Information Retrieval Platform. In: Losada, D.E., Fernández-Luna, J.M. (eds.) ECIR 2005. LNCS, vol. 3408, Springer, Heidelberg (2005), http://ir.dcs.gla.ac.uk/wiki/Terrier
10. Pedersen, T., Banerjee, S.: The design, implementation and use of the Ngram Statistics Package. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico (2003)
11. Salton, G., Buckley, C.: Term-weighting approaches in automatic retrieval. Information Processing and Management 24(5), 513–523 (1988)