Investigating Cross-Language Speech Retrieval for a Spontaneous Conversational Speech Collection

Diana Inkpen and **Muath Alzghool**
School of Information Technology and Eng.
University of Ottawa
Ottawa, Ontario, Canada, K1N 6N5
{diana,alzghool}@site.uottawa.ca

Gareth J.F. Jones
School of Computing
Dublin City University
Dublin 9, Ireland
Gareth.Jones@computing.dcu.ie

Douglas W. Oard
College of Information Studies & IACS
University of Maryland
College Park, MD 20742, USA
oard@glue.umd.edu

Abstract

Cross-language retrieval of spontaneous speech combines the challenges of working with noisy automated document transcripts and language translation. The CLEF 2005 Cross-Language Speech Retrieval (CL-SR) task provides a standard test collection to investigate these challenges. In our experimental investigation we show that we can improve retrieval performance by careful selection of the term weighting scheme and by combining the automatic transcripts with manually-assigned metadata. We further show that online machine translation resources can be used for topic translation to give effective CL-SR.

1 Introduction

The emergence of large collections of digitized spoken data has encouraged research in Spoken Document Retrieval (SDR). Previous studies, notably those at TREC (Garafolo et al, 2000), have focused mainly on well-structured news documents. In this paper we report on work carried out for the CLEF 2005 Cross-Language Speech Retrieval (CL-SR) track. The document collection for the CL-SR task is a part of the oral testimonies collected by the Shoah Visual History Foundation (VHF) for which automatic speech recognition (ASR) output has been generated within the MALACH project (Oard et al., 2004). This data is conversional spontaneous speech lacking clear topic boundaries. This is thus a more challenging SDR task than those explored previously. The data used for the CL-SR is also annotated with a range of automatic and manually generated sets of metadata. While the complete MALACH dataset is multilingual (ML), the current CL-SR task works only with English documents. However, as a first move towards the longer term objective of ML speech retrieval, the CLEF 2005 CL-SR explores cross-lingual searching, by making the experimental search queries (topics) available in several languages. This task raises many interesting research questions about the use of the multiple data fields in retrieval and cross-lingual translation. In this paper we explore alternative term weighting methods and content indexing strategies.

The remainder of this paper is structured as follows: Section 2 briefly reviews details of the CLEF 2005 CL-SR task; Section 3 describes the system we used to investigate this task; Section 4 reports our experimental results; and Section 5 gives conclusions and details for our ongoing work.

2 Task description

The CLEF-2005 CL-SR document set comprises 8104 “documents” which are manually-determined topically-coherent segments taken from 272 interviews with Holocaust survivors, witnesses and rescuers, totaling 589 hours of speech. Two ASR transcripts are available for this data, in this work we use the ASRTEXT2004A field provided by IBM research with a word error rate of 38%. Additional, metadata fields for each document include: two sets of 20 automatically assigned keywords determined using two different kNN classifiers (AK1 and AK2), a set of a varying number of manually-assigned keywords (MK), and a manual 3-sentence summary written by an expert in the field. A set of 38 training topics and 25 test topics were generated for this task. The topics provided with the collection were created in English from actual user requests. Topics were structured using the standard TREC format of Title, Description and Narrative fields. To enable CL-SR experiments the topics were translated into Czech, German, French, and Spanish by native speakers. Relevance judgments were generated using a search-guided procedure and standard pooling methods. See (Oard et al, 2004) and (White et al, 2005) for full details of the collection design.

3 System Overview

Our system for investigating the CLEF 2005 CL-SR task was built with off-the-shelf components. Topics were translated from French, Spanish, and German into English using seven free online machine translation (MT) tools were used. Their output was merged in order to allow for variety in lexical choices. All the translations of a topic Title field were combined in merged Title field of the translated topics; the same procedure was adopted for the Description and Narrative fields. Czech language topics were translated using InterTrans the only web-based MT system available to us.

Retrieval was carried out using the SMART IR system (Buckley et al, 1993) applying its standard stop word list and stemming algorithm.

In system development using the training topics we tested SMART with many different term weighting schemes combining collection frequency, document frequency and length normalization for the indexed collection and topics (Salton and Buckley, 1988). In this paper we employ the notation used in SMART to describe the combined schemes: xxx.xxx. The first three characters refer to the weighting scheme used to index the document collection and the last three characters refer to the weighting scheme used to index the topic fields. For example, lpc.atc means that lpc was used for documents and atc for queries. lpc would apply log term frequency weighting (l) and probabilistic collection frequency weighting (p) with cosine normalization to the document collection (c). atc would apply augmented normalized term frequency (a), inverse document frequency weight (t) with cosine normalization (c).

We experimented with many weighting schemes. The one that seemed to work best on the training data was lnn.ntn. For weighting document terms is uses logarithm of term frequency (l), no collection frequency factor (n), without normalization (n). For topics it uses non-normalized term frequency (n) and inverse document frequency weighting (t) without vector normalization (n). This combination worked well when all the fields of the query were used and with Title plus Description, but less well with the Title field only.

4 Experimental Investigation

In this section we report results from our experimental investigation of the CLEF 2005 CL-SR task. For each set of experiments we report standard TREC mean average precision (Map) computed using the *trec_eval* script. The topics fields used are indicated as: T for title only, TD for title + description, TDN for title + description + narrative. The first experiments shows results on the test topics for different term weighting schemes, and we then give cross-language retrieval results. For both sets of experiments documents are represented by the combination of the ASR transcription and the AK1 and AK2 fields. Thus each document representation is generated completely automatically. In further sets of experiments we give results of using two alternative indexing strategies.

4.1 Comparison of Term Weighting Schemes

Table 1 presents results for various weighting schemes for document and topics. There are 3600 possible combinations of weighting schemes: 60 schemes (5 x 4 x 3) for documents and 60 for queries. We tested a total of 240 combinations. In Table 1 we present in the table the results for 15 combinations (the best ones, plus some other ones to show the diversity of the results). lnn.ntn is still the best for the test topic set; and there might be a few other weighting schemes that achieve similar performance. Some of the weighting schemes perform best when indexing all the topic fields (TDN), some on TD, and some on title only (T). lnn.ntn was best for TDN and TD and lsn.ntn and lsn.atn were best for T. The lnn.ntn weighting scheme is used for all other experiments in this section.

Note that for mpc.ntn and other schemes that contain the probabilistic term “p”, due to a minor bug in Smart, some documents were returned as answer to the same query more than once. Because of the duplicates, the Map scores initially looked very high. Then we preprocessed the results to eliminate the duplicates and kept the first 1000 distinct results for each query, to retrieve the same number of documents per query as in the other experiments. The Map scores became a bit lower than for lnn.ntn.

	Weighting scheme	**TDN**	**TD**	**T**
		Map	**Map**	**Map**
1	lnn.ntn	**0.1366**	**0.1313**	0.1207
2	lnc.ntn	0.1362	0.1214	0.1094
3	mpc.ntn	0.1283	0.1219	0.1107
4	npc.ntn	0.1283	0.1219	0.1107
5	mpc.mtc	0.1283	0.1219	0.1107
6	mpc.mts	0.1282	0.1218	0.1108
7	mpc.nts	0.1282	0.1218	0.1108
8	npn.ntn	0.1258	0.1247	0.1118
9	lsn.ntn	0.1195	0.1233	**0.1227**
10	lsn.atn	0.0919	0.1115	**0.1227**
11	asn.ntn	0.0912	0.0923	0.1062
12	snn.ntn	0.0693	0.0592	0.0729
13	sps.ntn	0.0349	0.0377	0.0383
14	nps.ntn	0.0517	0.0416	0.0474
15	mtc.atc	0.1138	0.1151	0.1108

Table 1. Results of the various weighting schemes, for English. In bold are the best scores for TDN, TD, and T.

4.2 Cross-Language Experiments

Table 2 shows our results for the merged ASR, AK1 and AK2 documents with merged topic translations for French, German and Spanish, and single Czech translation. Examining these results we can see that Spanish topics perform well compared to monolingual English. However, results for German and Czech are much reduced. This is perhaps not surprising for the Czech topics were only a single translation is used. For German, we noticed that the quality of translation was sometimes low; some words were kept in German. For French, only TD topic fields were available, and so this condition is examined separately. In this case we can see that cross-language performance is almost identical to monolingual English. Use of the automatically-generated document fields was a required condition of the CL-SR task. Our results were the best submitted in this required submission mode (English topics, TD). The next-best system, from the University of Maryland, came in close. This difference was not statistically significant, but the difference to the other five systems was significant.

The required run provides a desirable condition to explore since generation of manual metadata such as manually-assigned keywords or expert written summaries is very expensive. However, as we show later in Table 4, manual metadata fields can produce significantly better retrieval performance than the automatically derived descriptions.

Topic Language	**System**	**Map**	**Fields**
English	Our system	0.1366	TDN
English	Our system	0.1313	TD
English	U Maryland	0.1288	TD
Spanish	Our system	0.1156	TDN
French	Our system	0.1275	TD
German	Our system	0.0936	TDN
Czech	Our system	0.0822	TDN

Table 2.Results for topics in all the languages (lnn.ntn). Comparison with another system.

Language	**Map**	**Fields**	**Description**
English	0.0986	T	Phonetic
English	0.1019	TD	Phonetic
English	0.0981	T	Phonetic+Text
English	0.1066	TD	Phonetic+Text
Spanish	0.0898	T	Phonetic
Spanish	0.0972	TD	Phonetic
Spanish	0.0948	T	Phonetic+Text
Spanish	0.1009	TD	Phonetic+Text
French	0.0931	T	Phonetic
French	0.1052	TD	Phonetic
French	0.0929	T	Phonetic+Text
French	0.1072	TD	Phonetic+Text
German	0.0744	T	Phonetic
German	0.0782	TD	Phonetic
German	0.0746	T	Phonetic+Text
German	0.0789	TD	Phonetic+Text
Czech	0.0479	T	Phonetic
Czech	0.0583	TD	Phonetic
Czech	0.0510	T	Phonetic+Text
Czech	0.0614	TD	Phonetic+Text

Table 3. Results on phonetic n-grams, and combination of text and phonetic transcriptions (lnn.ntn).

4.3 Results on Phonetic Transcriptions

In Table 3 we present results for an experiment where the text of the collection and topics, without stemming, is transformed into a phonetic transcription. Consecutive phones are then grouped into overlapping n-gram sequences (groups of n sounds, n = 4 in our case) that we used for indexing. The phonetic n-grams were provided by Clarke (2005), using NIST's text-to-phone tool[1]. For example, the phonetic form for the query fragment *child survivors* is: ch_ay_l_d s_ax_r_v ax_r_v_ay r_v_ay_v v_ay_v_ax ay_v_ax_r v_ax_r_z.

[1] http://www.nist.gov/speech/tools/

We wanted to test the hypothesis that the phonetic form could help compensate for the speech recognition errors made when the collection was produced. When the fields TD were indexed, the results are better than when only T is indexed. When combining phonetic and text forms (by simply indexing both phonetic n-grams and text), the result improved compared to using only the phonetic forms. But the Map scores are lower than the results on the text form for documents and queries.

4.4 Manual summaries and keywords

Table 4 explores the effect using the manual fields: manual keywords and 3-line summaries (full manual transcripts were not available). The retrieval performance increased significantly for all topic languages (more than double). The Map score increased from 13.66% to 32.56% when using the manual metadata, for English TDN. Table 4 also shows comparative results between and our results and the results from University of Maryland (they used manually-generated lists of names appearing the document, the manual keywords, and the summary fields). For English TDN and French TD our results are better. When combining manual keywords and manual summaries with ASR transcripts, AK1, and AK2, the results are lower than the ones in the first line of Table 4 (the Map score is 27.71% for English TDN).

Table 4.Results of indexing all the fields of the collections: MK and summaries (lnn.ntn). Comparison with another system.

Language	System	Map	Fields
English	Our system	0.3256	TDN
English	UMaryland	0.3129	TD
English	Our system	0.2989	TD
English	Our system	0.2754	T
Spanish	Our system	0.2548	TDN
French	Our system	0.2608	TD
French	UMaryland	0.2480	TD
German	Our system	0.2275	TDN
Czech	Our system	0.1667	TDN

5 Conclusions and Further Investigation

The system described in this paper obtained the best retrieval results on the required run among the seven teams that participated in this track. We believe that the choice of the weighting scheme used for indexing the terms is important. Table 2 shows that performance varies with the weighting scheme; it can be lower for the some of the classic indexing schemes.

In this paper we presented the results on the test queries, but our conclusions also applied on the training queries.

On the manual data, the best Map score we obtained is 32.56%, for English topics. On automatic data the best result is 13.66% Map score. This difference shows that the poor quality of the ASR transcripts severely hurts the performance of IR systems on this collection. In future work we plan to investigate methods of removing or correcting some of the speech recognition errors in the ASR transcripts, using semantic coherence measures.

The challenges of CLEF CL-SR task will continue to expand in subsequent years as new document and topic languages are introduced. This will also introduce new tasks of seeking relevant segments from within interviews where no manual segmentation has been carried out. This is related to previous TREC SDR experiments in an unknown-story boundary condition, but the topic boundaries will be less well-defined (Garafolo et al, 2000).

References

Chris Buckley, Gerard Salton, and James Allan. 1993. Automatic retrieval with locality information using SMART. In Proceedings of the First Text REtrieval Conference (TREC-1), pages 59–72.

Charles L. A. Clarke. 2005. Waterloo Experiments for the CLEF05 SDR Track, in Working Notes for the CLEF 2005 Workshop, Vienna, Austria

John S. Garofolo, Cedric G.P. Auzanne, Ellen M. Voorhees. 2000. The TREC Spoken Document Retrieval Track: A Success Story. In Proceedings of the RIAO Conference: Content-Based Multimedia Information Access, Paris, France, pages1-20.

Douglas W. Oard, Dagobert Soergel, David Doermann, Xiaoli Huang, G. Craig Murray, Jianqiang Wang, Bhuvana Ramabhadran, Martin Franz, Samuel Gustman. 2004. Building an Information Retrieval Test Collection for Spontaneous Conversational Speech, in Proceedings of SIGIR.

Gerard Salton and Chris Buckley. 1988. Term-weighting approaches in automatic retrieval. Information Processing and Management, 24(5):513-523.

Ryen W. White, Douglas W. Oard, Gareth J. F. Jones, Dagobert Soergel, Xiaoli Huang. 2005. Overview of the CLEF-2005 Cross-Language Speech Retrieval Track, in Working Notes for the CLEF 2005 Workshop, Vienna, Austria