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

Diana Inkpen, Muath Alzghool
School of Info. 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 Info. Studies/UMIACS
University of Maryland
College Park, MD 20742, USA
oard@umd.edu

Abstract

Cross-language retrieval of spontaneous speech combines the challenges of working with noisy automated transcription and language translation. The CLEF 2005 Cross-Language Speech Retrieval (CL-SR) task provides a standard test collection to investigate these challenges. We show that we can improve retrieval performance: by careful selection of the term weighting scheme; by decomposing automated transcripts into phonetic substrings to help ameliorate transcription errors; and by combining automatic transcriptions with manually-assigned metadata. We further show that topic translation with online machine translation resources yields effective CL-SR.

1 Introduction

The emergence of large collections of digitized spoken data has encouraged research in speech retrieval. 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 Cross-Language Evaluation Forum (CLEF) 2005 Cross-Language Speech Retrieval (CL-SR) track (White et al, 2005). The document collection for the CL-SR task is a part of the oral testimonies collected by the USC Shoah Foundation Institute for Visual History and Education (VHI) for which some Automatic Speech Recognition (ASR) transcriptions are available (Oard et al., 2004). The data is conversional spontaneous speech lacking clear topic boundaries; it is thus a more challenging speech retrieval task than those explored previously. The CLEF data is also annotated with a range of automatic and manually generated sets of metadata. While the complete VHI dataset contains interviews in many languages, the CLEF 2005 CL-SR task focuses on English speech. Cross-language searching is evaluated by making the topic statements (from which queries are automatically formed) available in several languages. This task raises many interesting research questions; 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 collection includes 8,104 manually-determined topically-coherent segments 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 transcripts provided by IBM Research in 2004 for which a mean word error rate of 38% was computed on held out data. Additional, metadata fields for each segment include: two sets of 20 automatically assigned thesaurus terms from different kNN classifiers (AK1 and AK2), an average of 5 manually-assigned thesaurus terms (MK), and a 3-sentence summary written by a subject matter expert. A set of 38 training topics and 25 test topics were generated in English from actual user requests. Topics were structured as Title, Description and Narrative fields, which correspond roughly to a 2-3 word Web query, what someone might first say to a librarian, and what that librarian might ultimately understand after a brief reference interview. To support CL-SR experiments the topics were re-expressed in Czech, German, French, and Spanish by native speakers in a manner reflecting

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 61–64,
New York, June 2006.

the way questions would be posed in those languages. Relevance judgments were manually generated using by augmenting an interactive search-guided procedure and purposive sampling designed to identify additional relevant segments. See (Oard et al, 2004) and (White et al, 2005) for details.

3 System Overview

Our Information Retrieval (IR) system 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. Their output was merged in order to allow for variety in lexical choices. All the translations of a topic Title field were combined in a 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 for this language pair. 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).

One scheme in particular (mpc.ntn) proved to have much better performance than other combinations. For weighting document terms we used term frequency normalized by the maximum value (m) and probabilistic collection frequency weighting (p) with cosine normalization (c). For topics we used non-normalized term frequency (n) and inverse document frequency weighting (t) without vector normalization (n). This combination worked very well when all the fields of the query were used; it also worked well with Title plus Description, but slightly less well with the Title field alone.

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 Mean uninterpolated Average Precision (MAP) computed using the *trec_eval* script. The topic fields used are indicated as: T for title only, TD for title + description, TDN for title + description + narrative. The first experiment shows results for different term weighting schemes; we then give cross-language retrieval results. For both sets of experiments, "documents" are represented by combining the ASR transcription with the AK1 and AK2 fields. Thus each document representation is generated completely automatically. Later experiments explore two alternative indexing strategies.

4.1 Comparison of Term Weighting Schemes

The CLEF 2005 CL-SR collection is quite small by IR standards, and it is well known that collection size matters when selecting term weighting schemes (Salton and Buckley, 1988). Moreover, the documents in this case are relatively short, averaging about 500 words (about 4 minutes of speech), and that factor may affect the optimal choice of weighting schemes as well. We therefore used the training topics to explore the space of available SMART term weighting schemes. Table 1 presents results for various weighting schemes with English topics. There are 3,600 possible combinations of weighting schemes available: 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 the results for 15 combinations (the best ones, plus some others to illustate the diversity of the results). mpc.ntn is still the best for the test topic set; but, as shown, a few other weighting schemes achieve similar performance. Some of the weighting schemes perform better when indexing all the topic fields (TDN), some on TD, and some on title only (T). npn.ntn was best for TD and lsn.ntn and lsn.atn are best for T. The mpc.ntn weighting scheme is used for all other experiments in this section. We are investigating the reasons for the effectiveness of this weighting scheme in our experiments.

	Weighting scheme	TDN	TD	T
		Map	Map	Map
1	Mpc.mts	0.2175	0.1651	0.1175
2	Mpc.nts	0.2175	0.1651	0.1175
3	Mpc.ntn	**0.2176**	0.1653	0.1174
4	npc.ntn	**0.2176**	0.1653	0.1174
5	Mpc.mtc	**0.2176**	0.1653	0.1174
6	Mpc.ntc	**0.2176**	0.1653	0.1174
7	Mpc.mtn	**0.2176**	0.1653	0.1174
8	Npn.ntn	0.2116	**0.1681**	0.1181
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. MAP, 25 English test topics. Bold=best scores.

4.2 Cross-Language Experiments

Table 2 shows our results for the merged ASR, AK1 and AK2 documents with multi-system topic translations for French, German and Spanish, and single-system Czech translation. We can see that Spanish topics perform well compared to monolingual English. However, results for German and Czech are much poorer. This is perhaps not surprising for the Czech topics where only a single translation is available. For German, the quality of translation was sometimes low and some German words were retained untranslated. For French, only TD topic fields were available. In this case we can see that cross-language retrieval effectiveness is almost identical to monolingual English. Every research team participating in the CLEF 2005 CL-SR task submitted at least one TD English run, and among those our mpc.ntn system yielded the best MAP (Wilcoxon signed rank test for paired samples, $p<0.05$). However, as we show in Table 4, manual metadata can yield better retrieval effectiveness than automatic description.

Topic Language	System	Map	Fields
English	Our system	0.1653	TD
English	Our system	0.2176	TDN
Spanish	Our system	0.1863	TDN
French	Our system	0.1685	TD
German	Our system	0.1281	TDN
Czech	Our system	0.1166	TDN

Table 2. MAP, cross-language, 25 test topics

Language	Map	Fields	Description
English	0.1276	T	Phonetic
English	0.2550	TD	Phonetic
English	0.1245	T	Phonetic+Text
English	0.2590	TD	Phonetic+Text
Spanish	0.1395	T	Phonetic
Spanish	0.2653	TD	Phonetic
Spanish	0.1443	T	Phonetic+Text
Spanish	0.2669	TD	Phonetic+Text
French	0.1251	T	Phonetic
French	0.2726	TD	Phonetic
French	0.1254	T	Phonetic+Text
French	0.2833	TD	Phonetic+Text
German	0.1163	T	Phonetic
German	0.2356	TD	Phonetic
German	0.1187	T	Phonetic+Text
German	0.2324	TD	Phonetic+Text
Czech	0.0776	T	Phonetic
Czech	0.1647	TD	Phonetic
Czech	0.0805	T	Phonetic+Text
Czech	0.1695	TD	Phonetic+Text

Table 3. MAP, phonetic 4-grams, 25 test topics.

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.

The phonetic form helps compensate for the speech recognition errors. With TD queries, the results improve substantially compared with the text form of the documents and queries (9% relative). Combining phonetic and text forms (by simply indexing both phonetic n-grams and text) yields little additional improvement.

4.4 Manual summaries and keywords

Manually prepared transcripts are not available for this test collection, so we chose to use manually assigned metadata as a reference condition. To explore the effect of merging automatic and manual fields, Table 4 presents the results combining man-

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

ual keywords and manual summaries with ASR transcripts, AK1, and AK2. Retrieval effectiveness increased substantially for all topic languages. The MAP score improved with 25% relative when adding the manual metadata for English TDN.

Table 4 also shows comparative results between and our results and results reported by the University of Maryland at CLEF 2005 using a widely used IR system (InQuery) that has a standard term weighting algorithm optimized for large collections. For English TD, our system is 6% (relative) better and for French TD 10% (relative) better. The University of Maryland results with only automated fields are also lower than the results we report in Table 2 for the same fields.

Table 4. MAP, indexing all fields (MK, summaries, ASR transcripts, AK1 and AK2), 25 test topics.

Language	System	Map	Fields
English	Our system	0.4647	TDN
English	Our system	0.3689	TD
English	InQuery	0.3129	TD
English	Our system	0.2861	T
Spanish	Our system	0.3811	TDN
French	Our system	0.3496	TD
French	InQuery	0.2480	TD
French	Our system	0.3496	TD
German	Our system	0.2513	TDN
Czech	Our system	0.2338	TDN

5 Conclusions and Further Investigation

The system described in this paper obtained the best results among the seven teams that participated in the CLEF 2005 CL-SR track. We believe that this results from our use of the 38 training topics to find a term weighting scheme that is particularly suitable for this collection. Relevance judgments are typically not available for training until the second year of an IR evaluation; using a search-guided process that does not require system results to be available before judgments can be performed made it possible to accelerate that timetable in this case. Table 2 shows that performance varies markedly with the choice of weighting scheme. Indeed, some of the classic weighting schemes yielded much poorer results than the one we ultimately selected. In this paper we presented results on the test queries, but we observed similar effects on the training queries.

On combined manual and automatic data, the best MAP score we obtained for English topics is 0.4647. On automatic data, the best MAP is 0.2176. This difference could result from ASR errors or from terms added by human indexers that were not available to the ASR system to be recognized. 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.

In ongoing further work we are exploring the relationship between properties of the collection and the weighting schemes in order to better understand the underlying reasons for the demonstrated effectiveness of the mpc.ntn weighting scheme.

The challenges of CLEF CL-SR task will continue to expand in subsequent years as new collections are introduced (e.g., Czech interviews in 2006). Because manually assigned segment boundaries are available only for English interviews, this will yield an unknown topic boundary condition that is similar to previous experiments with automatically transcribed broadcast news the Text Retrieval Conference (Garafolo et al, 2000), but with the additional caveat that topic boundaries are not known for the ground truth relevance judgments.

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 and 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, pages 1-20.

Douglas W. Oard, Dagobert Soergel, David Doermann, Xiaoli Huang, G. Craig Murray, Jianqiang Wang, Bhuvana Ramabhadran, Martin Franz and Samuel Gustman. 2004. Building an Information Retrieval Test Collection for Spontaneous Conversational Speech, in Proceedings of SIGIR, pages 41-48.

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 and Xiaoli Huang. 2005. Overview of the CLEF-2005 Cross-Language Speech Retrieval Track, in Working Notes for the CLEF 2005 Workshop, Vienna, Austria