Cross-Language Access to Recorded Speech in the MALACH Project

Douglas W. Oard[1], Dina Demner-Fushman[1], Jan Hajic[2], Bhuvana Ramabhadran[3], Samuel Gustman[4], William J. Byrne[5], Dagobert Soergel[1], Bonnie Dorr[1], Philip Resnik[1], and Michael Picheny[3]

[1] University of Maryland, College Park, MD 20742 USA,
(oard,demner,bonnie,resnik)@umiacs.umd.edu, ds52@umail.umd.edu
[2] Charles University, CZ-11800 Praha 1, Czech Republic
hajic@ufal.ms.mff.cuni.cz
[3] IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA
(bhuvana,picheny)@us.ibm.com
[4] Survivors of the Shoah Visual History Foundation, Los Angeles, CA 90078 USA
sam@vhf.org
[5] Johns Hopkins University, Baltimore, MD 21218 USA,
byrne@jhu.edu

Abstract. The MALACH project seeks to help users find information in a vast multilingual collections of untranscribed oral history interviews. This paper introduces the goals of the project and focuses on supporting access by users who are unfamiliar with the interview language. It begins with a review of the state of the art in cross-language speech retrieval; approaches that will be investigated in the project are then described. Czech was selected as the first non-English language to be supported, so results of an initial experiment with Czech/English cross-language retrieval are reported.

1 Introduction

Digital archives of recorded speech are emerging as an important way of capturing the human experience. Before such archives can be used efficiently, however, their contents must be described in a way that supports access to the information that they contain. Our ability to collect and store digitized speech now greatly outstrips our ability to manually describe what we have collected, but present automated technologies for search and exploration in spoken materials still have sharply limited capabilities. The MALACH (Multilingual Access to Large Spoken Archives) project is working with what we believe is the world's largest coherent archive of video oral histories to apply emerging speech recognition and natural language processing technologies to this important problem. In this paper we identify the research issues raised by the project, with particular emphasis on those related to handling Eastern European languages, describe the approach that we plan to take to explore these issues, and present the results of an initial cross-language retrieval experiment between Czech and English.

2 The MALACH Project

MALACH is a five-year project that brings together the Survivors of the Shoah Visual History Foundation (VHF), the University of Maryland, the IBM T.J. Watson Research Center, Johns Hopkins University, Charles University and the University of West Bohemia to (1) advance the state of the art in speech recognition to handle spontaneous, emotional, heavily accented, and elderly speech in multiple languages with uncued language switching, (2) develop automated techniques for the generation of metadata in multiple languages to support information access, and (3) advance the state of the art in information retrieval to provide efficient search, indexing and retrieval of recorded interviews.

In 1994, after releasing the film *Schindler's List*, Steven Spielberg was approached by many survivors to listen to their stories of the Holocaust. Spielberg decided to start the VHF so that as many survivors as possible could tell their stories so that they could be used to teach about the horrors of intolerance. Today, the VHF has collected almost 52,000 testimonies (116,000 hours of video) in 32 languages. Five copies of each testimony exist, including an MPEG-1 3 Mb/s version for digital distribution. The entire digitized archive occupies 180 terabytes of storage. During the collection of each testimony, a forty page survey in the survivor or witnesses native language was taken. These surveys have been entered into the digital library and form an initial catalog for searching through the subject matter in each interview. Working with information scientists and historians, a cataloging system for indexing the content of each testimony has been created. Today, thirty catalogers work with this system to manually assign metadata from a thesaurus tailored for this purpose to portions of each video at a rate of about 1,000 hours per month.

Our preliminary experiments with three different speaker-independent English speech recognition systems trained with very different material (broadcast news, dictation and telephone conversations) resulted in remarkably similar word error rates (nearly 60%) on the VHF collection. Conventional adaptation techniques improved the performance significantly, bringing the word error rate down to around 33% for fluent speakers and 46% for heavily accented and disfluent speakers (which are common in the collection). Moreover, approximately 15% of the words were outside the vocabulary on which the recognizer was trained, so these domain-specific terms, many of which are named entities, will require special handling. Although the ultimate goal of speech recognition research is to produce readable transcripts, our more limited immediate goal in this project is to produce transcripts that are adequate to support metadata creation and information retrieval (see [8] for a description of the issues involved). The huge size of the collection makes it a unique resource for exploring the effect of corpus size on the word error rate reduction that can be achieved through long-term adaptation, and the presence of so many languages makes it an unmatched resource for exploring the potential of bootstrapping speech recognition systems in less frequently spoken languages for which sharply limited quantities of annotated training data might be available.

The linear nature of recorded speech poses unique problems for information access that we plan to investigate. The linguistic diversity of the collection adds additional challenges, both because we want to support queries in the same language as the materials (and must therefore support many languages) and because we think it will be valuable to provide access to interviews that were not conducted in the language in which the query was posed. We have described the other issues in detail elsewhere (see [5]), so for the remainder of this paper we focus on supporting cross-language access to spoken word collections.

3 Cross-Language Access to Recorded Speech

Cross-language access to recorded speech poses three challenges: (1) automatically searching the collection using queries that may be expressed in a language different from that used by the speaker, (2) manually selecting portions of the top-ranked recordings that satisfy the searcher's information need, and (3) making use of the selected parts of those recordings. We believe that the first two problems are tractable within the scope of this project, and we ultimately hope that our results will be applied synergistically with ongoing research efforts in speech-to-speech translation (see [12] for an overview of recent research on this topic). Present speech-to-speech translation systems achieve robust performance only in limited domains, however, so we expect that in the near term cross-language use of materials that are found using the technology that we are building will be achieved though collaboration between a subject matter expert (who may lack needed language skills) and an assistant who is fluent in the spoken language. In the remainder of this section we therefore address each of the first two challenges in turn.

Cross-language text retrieval and monolingual speech retrieval are both well-researched problems (see [10] for a survey of the first and [1] for a survey of the second), and these two challenges have been explored together using the Topic Detection and Tracking (TDT) collections. All of the work to date on speech retrieval—both monolingual and cross-language—has focused on broadcast news, one of the more tractable applications for speech recognition technology. The conversational, emotional, and accented speech in the VHF collection pose substantial challenges to presently available speech recognition techniques, and addressing those challenges is a principal focus of the MALACH project. In a related project, we are developing inexpensive techniques for extending cross-language retrieval capabilities to new language pairs, and as speech recognition techniques that are tuned to the characteristics of the VHF collection are developed we will begin to explore the potential of automated cross-language search techniques for this application. We expect that techniques we have used with the TDT collection such as document expansion based on blind relevance feedback [7] and phonetic transliteration rules learned from examples [9] will be useful, and what we learn about the unique characteristics of oral history interviews in our monolingual experiments may help us to see additional opportunities to improve cross-language search capabilities.

The availability of manually-assigned thesaurus descriptors offers a complementary basis for cross-language searching. Thesaurus-based searching can be quite effective if the searcher is familiar with the structure of the thesaurus, so one option is to arrange for the assistance of a specially trained search intermediary. Automated text classification can also be used to help users find thesaurus terms that are associated with natural language (free text) queries (see [4] for an example of how this can be done). The VHF collection provides an unmatched source of data for training automatic text classification algorithms, with nearly 200,000 training examples in which a three-sentence English summary of a topically-coherent interview segment is associated with one or more thesaurus descriptors. In another project, we are exploring the extension of similar capabilities to new languages by annotating the English half of a parallel (translation-equivalent) corpus, projecting the annotations to the other language, and then training a new classifier using the projected annotations (see [13] for a description of how this idea has been applied to related problems). In order to apply this idea to our classification task, we will need to assemble a large collection of topically appropriate parallel texts and we will need to develop classification algorithms that are relatively insensitive to the types of divergences that we observe between languages (e.g., head switching). Translation of documents about the Holocaust is a common practice, so we expect to be able to meet the first requirement for at least some language pairs, and we expect that the second task will benefit from the results of an investigation of translation divergence that we are presently undertaking (see [3] for some early results from that work). We therefore expect to be able to provide some useful degree of mapping between free text search terms and controlled vocabulary thesaurus descriptors, even before thesaurus term translations are available.

Cross-language document selection has only recently received attention from researchers. In the first multi-site evaluation effort of interactive cross-language retrieval from a collection of text, the Cross-Language Evaluation Forum's interactive track (iCLEF) [11], Lòpez-Ostenero, et al. found that translated key phrases could be used as a basis for selection as effectively as full (but sometimes disfluent) machine translations. Merlino and Maybury found that similar phrase-based summaries were also helpful for interactive monolingual retrieval from multimedia (audio and video) collections. The manually assigned thesaurus descriptors in the VHF collection provide an excellent starting point for extending these techniques to support interactive cross-language selection of recorded speech. The present thesaurus contains only English vocabulary, but the relatively compact and specialized vocabulary used in the thesaurus facilitates translation, and the concept relationships that the thesaurus encodes offer an additional source of evidence to guide automatic or semiautomatic disambiguation algorithms (see [2] for an example of how we have previously exploited similar constraints). We also plan to leverage the human-assigned descriptors in the VHF collection to explore the degree to which we can provide similar support for document selection using automatic text classification techniques based on annotation projection through parallel corpora.

4 Searching Czech Documents with English Queries

Czech will be the first non-English language for which we plan to develop speech recognition techniques that are tuned to the VHF collection. We are not aware of any prior work on Czech/English cross-language information retrieval, so we have conducted a preliminary experiment to begin to explore the issues involved in supporting automatic search between Czech and English.

For our experiments, we used an information retrieval test collection from the Cross-Language Evaluation Forum (CLEF 2000). The collection contains 113,000 English news stories from the Los Angeles Times (about 435 MB of text), 33 English topic descriptions,[1] and binary (yes-no) relevance judgments for topic-document pairs.[2] The title and description fields of the 33 topic descriptions were translated from English into Czech by a native speaker of Czech in a way that they felt represented a natural style of expression for a statement of an information need in that language. In CLEF topic descriptions, the title field is typically crafted in a manner similar to typical Web queries (2-3 content-bearing terms), while the description fields are typical of what a searcher might initially say to an intermediary such as a librarian that might help with their search (1-2 sentences).

We obtained translation knowledge for our automated cross-language search system from two sources:

- We submitted each word in the Czech queries (and their lemmas, obtained as described below) to the PC Translator V98 (for MS Windows) Czech-English machine readable bilingual dictionary (http://www.langsoft.cz/) and aligned the results to create a bilingual term list. The dictionary contained only Czech lemmas, and translations were found in this way for 213 of the 291 unique words that appear in the queries.
- We obtained 800 additional English-Czech lemma pairs from the freely available Ergane translation tool (http://download.travlang.com/Ergane). These term pairs are not query-specific—we used every term pair that Ergane provides.

We merged these two resources to form a single bilingual term list for use in subsequent processing.

We used a simple translation process for automatic word-by-word query-translation, in which the following processing stages were tried in order: (1) Look up the lemma in the Czech side of the bilingual term list (we used a morphological analyzer distributed with the Prague Dependency Treebank to lemmatize Czech words [6]), and (2) if the lemma is not found, strip diacritic marks from the characters in the word to obtain a 7-bit ASCII representation of the word that might be a correct transliteration. We also tried a variant of the third stage in

[1] The CLEF 2000 collection contains 40 topics, but no relevant English documents are known for topics 2, 6, 8, 23, 25, 27, and 35, so they were excluded.

[2] The set of relevance judgments is incomplete, but the pooled assessment methodology generally results in reliable comparisons between alternative conditions.

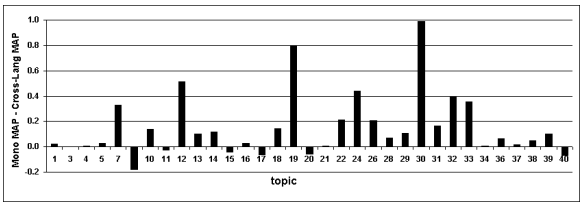
which certain obvious transliteration corrections were made manually (afrika to africa, rusku to russia) as a way of exploring the potential effect on retrieval if we were to develop a more sophisticated automatic transliteration algorithm.

We used the InQuery text retrieval system with the default English stemmer (kstem) enabled for both document and query processing. Structured queries were formed by including alternate translations for a single query term in InQuery's #syn operator, which has the effect of computing a single term weight for the aggregate set of translations by first aggregating the term frequency (aboutness) and document frequency (term specificity) evidence separately and then computing the term weight on the aggregated statistics. This approach has been shown to outperform techniques based on aggregating term weights in a broad range of language pairs (when translation probabilities are not known). We used the widely used trec_eval program to compute the uninterpolated average precision for each query, and report the mean uninterpolated average precision over 33 queries as a single-figure measure of retrieval effectiveness.

Figure 1 (a) shows the results for title-only and title+description queries. The "Monolingual" results establish an upper baseline, obtained by forming queries using the English topic descriptions that had served as the basis for the Czech translations that are used in the remainder of the runs. The Dictionary-based Query Translation (DQT) runs were created as described above, with the "DQT + Names" run showing the effect of manually correcting the transliteration of some names (also described above). The "No Translation" runs establish a lower baseline, obtained by using the Czech queries to search the English document collection without benefit of any translation.

	Title	Title+Desc
Monolingual	0.3279	0.4008
DQT	0.1826	0.1705
DQT + Names	0.1882	0.1996
No Translation	0.0285	0.0258

(a)



(b)

Fig. 1. Uninterpolated average precision, (a) mean, (b) by topic, title queries.

These results are typical of initial experiments in new language pairs, for which it is common to obtain between 40% and 60% of monolingual retrieval effectiveness when simple word-by-word dictionary-based techniques are used. Better results can typically be obtained if pre-translation and/or post-translation query expansion is performed using well-tuned blind relevance feedback, and incorporation of translation probabilities and making greater use of context (e.g., by translating phrases in preference to single words) can also be helpful. These

early results are, however, sufficient to serve as a basis for some further investigation into the unique characteristics of Czech. We therefore performed a query-by-query failure analysis, using the plots in Figures 1 (b) to identify the topics for which the spread between the DQT and Monolingual runs was relatively large. We then examined the topic descriptions and the translated queries in an effort to identify a plausible cause for this effect. Through this process, we made the following observations:

Topic 7. All terms were translated, but soccer was expanded to soccer and football. This could have an adverse effect on precision when searching newspaper stories from the United States.

Topic 12. The query term "Solar Temple" (a named entity) had solar replaced by sun, sunshine, etc., and English stemming failed to discover the relationship between these words and solar.

Topic 19. The query term "Gulf War Syndrome" resulted in many synonyms for gulf and syndrome, and war was lost in the translation process.

Topic 24. The query term "World Trade Organization" resulted in loss of world and replacement of trade with business, commercial, mercantile, etc., none of which were conflated by the English stemmer with trade.

Topic 30. The named entity Nice (a city in France) was transliterated as niche (an English word with an unrelated meaning).

Topic 32. Several query terms were lost in the translation process.

The same analysis produced similar results on title+description queries. Word-by-word dictionary-based cross-language retrieval techniques are known to be vulnerable to mishandling phrases and named entities, so these results do not point to any unusual characteristics that are unique to the Czech/English language pair. We therefore expect that techniques that we have used previously in other language pairs to boost cross-language retrieval effectiveness and to integrate translation with speech recognition are likely to achieve similar results in Czech.

5 Conclusion

The collection that we are working with contains tens of thousands of hours of speech in Eastern European languages, so we have a keen interest in collaboration with members of the research communities that have come together for the Text, Speech and Dialog conference. We see significant opportunities for synergy with those working on speech recognition, spoken language identification, natural language processing, text classification, machine translation and information retrieval, with the MALACH project providing an unmatched environment in which to demonstrate how these technologies can be integrated to produce compelling applications. The potential impact of such joint efforts could extend far beyond this one project, however. It is our hope that the technologies we develop will also be used in the service of additional efforts to preserve our memory, and to give an enduring voice to those who have built the world in which our children and grandchildren will live.

Acknowledgments

The authors would like to thank Ivona Kucerova for translating the queries and Jan Curin for help with PC Translator. This work has been supported in part by NSF grant IIS-0122466 and DARPA cooperative agreement N660010028910.

References

1. James Allan. Perspectives on information retrieval and speech. In Anni R. Coden, Eric W. Brown, and Savitha Srinivasan, editors, *Information Retrieval Techniques for Speech Applications*, pages 1–10. Springer, 2002. Lecture Notes in Computer Science 2273.
2. Bonnie J. Dorr, Gina Anne Levow, and Dekang Lin. Construction of a Chinese-English verb lexicon for embedded machine translation in cross-language information retrieval. *Machine Translation*, 2002. To appear.
3. Bonnie J. Dorr, Lisa Pearl, Rebecca Hwa, and Nizar Habash. Improved word-level alignment: Injecting knowledge about MT divergences. Technical Report CS-TR-4333, University of Maryland, Institute for Advanced Computer Studies, 2002.
4. Frederic C. Gey, Michael Buckland, Aitao Chen, and Ray Larson. Entry vocabulary—a technology to enhance digital search. In *First International Conference on Human Language Technologies*, 2001.
5. Samuel Gustman, Dagobert Soergel, Douglas Oard, William Byrne, Michael Picheny, Bhuvana Ramadhadran, and Douglas Greenberg. Supporting access to large digital oral history archives. In *The Second Joint Digital Libraries*, June 2002. to appear.
6. Jan Hajic, Eva Hajicova, Petr Pajas, Jarmila Panevova, Petr Sgall, and Barbora Vidova-Hladka. Prague dependency treebank 1.0, 2001. LDC2001T10.
7. Gina-Anne Levow and Douglas W. Oard. Signal boosting for translingual topic tracking. In James Allan, editor, *Topic Detection and Tracking: Event-based Information Organization*, pages 175–195. Kluwer Academic Publishers, Boston, 2002.
8. J. Scott McCarley and Martin Franz. Influence of speech recognition errors on topic detection. In *Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 342–344, 2000.
9. Helen Meng, Berlin Chen, Erika Grams, Sanjeev Khudanpur, Gina-Anne Levow, Wai-Kit Lo, Douglas Oard, Patrick Schone, Karen Tang, Hsin-Min Wang, and Jianqiang Wang. Mandarin-English information (MEI): Investigating translingual speech retrieval. In *First International Conference on Human Language Technologies*, San Diego, March 2001.
10. Douglas W. Oard and Anne R. Diekema. Cross-language information retrieval. In *Annual Review of Information Science and Technology*, volume 33. American Society for Information Science, 1998.
11. Douglas W. Oard and Julio Gonzalo. The CLEF 2001 interactive track. In Carol Peters, editor, *Proceedings of the Second Cross-Language Evaluation Forum*. 2002.
12. Wolfgang Wahlster, editor. *Verbmobil: Foundations of speech-to-speech translation*. Springer, Berlin, 2000.
13. D. Yarowsky, G. Nagi, and R. Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *First International Conference on Human Language Technologies*, 2001.