Multilingual Access to Large Spoken Archives

Douglas W. Oard
College of Information Studies and
Institute for Advanced Computer Studies
University of Maryland, College Park, MD, USA

Abstract

Spoken word collections promise access to unique and compelling content, and most of the technology needed to realize that promise is now in place. Decreasing storage costs, increasing network capacity, and the availability of software to encode and exchange digital audio make possible physical access to spoken word collections at a previously unimaginable scale. Effective support for intellectual access – the problem of finding what you are looking for – is much more challenging, however. In this talk I will briefly describe work that has been done on this problem at the Text Retrieval Conferences, the Topic Detection and Tracking evaluations, and in individual research projects around the world. I will then describe a unique resource, a collection of 116,000 hours of oral history interviews recorded in 32 languages in 57 countries that has been assembled by the Survivors of the Shoah Visual History Foundation. Nearly 10,000 hours of this audio has been manually segmented, summarized and indexed, making this an unrivaled resource with which we can explore a broad array of data-driven techniques. My main focus will be to explain how we are leveraging this exceptional resource to develop the ability to index similar materials automatically.

The project we call MALACH (Multilingual Access to Large spoken ArCHives) builds on a long heritage of increasingly demanding applications for speech recognition technology. The accented, emotional and elderly speech in the Shoah Foundation's collection are so challenging that state-of-the-art systems initially yielded a 90% word error rate! We now have speech recognition systems that achieve better than half that error rate for two languages, English and Czech. That's nowhere near good enough to produce readable transcripts, but it is approaching a point where other language technologies can begin to make headway. I'll illustrate that point with our latest results from across the project on speech recognition, natural language processing components, and information retrieval system design.

The scope of this one project is breathtaking, directly involving nine research teams from six institutions on two continents (Charles University, IBM T.J. Watson Research Lab, Johns Hopkins University, the Shoah Foundation, the University of Maryland, and the University of West Bohemia), with interests that range from the information needs of historians to the modeling of Czech colloquial pronunciation. Virtually every topic in computational linguistics finds expression in that range. We plan to ultimately build speech recognition systems in at least five languages (adding Russian, Polish and Slovak to what we have now), so morphology and language modeling are critical issues. The diverse range of languages in the collection make

translation and translingual search essential capabilities. The sheer size of the collection and the strict linearity of the audio medium call for effective summarization. References to named entities are important hooks for many information seeking strategies, so named entity detection and co-reference resolution techniques that are robust in the face of pronunciation variations are needed. An interview is a dialog, and these interviews contain a rich discourse structure, thus effective discourse and dialog analysis could lead to new ways of supporting access. And, of course, progress on all of this depends fundamentally on evaluation.

As with any collection, we must respect the wishes of its creators when using it in our research. In this case, more than 50,000 people contributed their life stories. The stories speak of some of the greatest inhumanity ever witnessed, and many of those who told those stories still walk among us. Much as we might wish that ELRA or the LDC could obtain and release the entire collection, that is not likely to happen any time soon. But the Shoah Foundation does hope to begin the process of clearing subsets of the collection for research use over the next year or so, and we are gearing our annotation and test collection development efforts to maximize the overlap with what they will ultimately release. Now is therefore a propitious time to begin to think about how these unique materials might be used in your own research. Since the dawn of language, the oral tradition has been the dominant form in which we have told our stories and passed on our culture. Over the past few thousand years, the written form has moved to the fore, principally because access to the written word has been more easily supported by the available technology. We now stand on the verge of restoring the balance and building an oral tradition that gives lasting voice to those who choose not to write their stories. I invite you to join us in that quest!

This material is based on work supported by the National Science Foundation under grant IIS-0122466. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.