Lessons from the MALACH Project:
Applying new technologies to improve intellectual access to large oral history collections

Douglas W. Oard, University of Maryland, USA

Abstract:

In this talk I will describe the goals of the MALACH project (Multilingual Access to Large Spoken Archives) and our research results. I'll begin by describing the unique characteristics of the oral history collection that we used, in which Holocaust survivors, witnesses and rescuers were interviewed in several languages. Each interview has been digitized and extensively catalogued by subject matter experts, thus producing a remarkably rich collection for the application of machine learning techniques. Automatic speech recognition techniques originally developed for the domain of conversational telephone speech were adapted to process these materials with word error rates that are adequate to provide useful features to support interactive search and automated clustering, boundary detection, and topic classification tasks. As I describe our results, I will focus particularly on the evaluation methods that that we have used to assess the potential utility of this technology. I'll conclude with some remarks about possible future directions for research on applying new technologies to improve intellectual access to oral history and other spoken word collections. This is joint work with Charles University (Prague), IBM Research (T.J. Watson), the Johns Hopkins University (Baltimore), the University of Southern California (Los Angeles), an the University of West Bohemia (Pilsen),

About the Speaker:

Douglas Oard is Associate Dean for Research at the College of Information Studies of the University of Maryland, College Park, where he holds joint appointments as Associate Professor in the College of Information Studies and in the Institute for Advanced Computer Studies. He earned his Ph.D. in Electrical Engineering from the University of Maryland, and his research interests center around the use of emerging technologies to support information seeking by end users. Dr. Oard's recent work has focused on interactive techniques for cross-language information retrieval, searching conversational media, and leveraging observable behavior to improve user modeling. Additional information is available at http://www.glue.umd.edu/~oard/.

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), page 89,
Prague, 28 June 2007.