Supporting Global Knowledge Sharing using Cross-Language Information Retrieval

Petra Galuščáková and Douglas W. Oard

University of Maryland

NASA has over the years made substantial investment in information systems to provide access to, for example, technical reports [3], oral histories [2], and knowledge repositories [1]. One characteristic of all of these systems is that they have been designed principally for English. That makes sense; English is the language of international science and engineering, and NASA users are invariably fluent in English. But the dominance of English does not mean that English is the only language we need access to. All of the oral histories in the Shuttle-MIR oral history collection were conducted in English; people who could speak only Russian simply weren't interviewed. Chinese technical publications might be of value to NASA engineers who working on similar problems. And understanding the global reaction to NASA's activities would benefit from systems that could process Hindi and French as easily as English. These are three of several scenarios in which Cross-Language Information Retrieval (CLIR) capabilities could be useful.

At its core, CLIR involves the use of queries in one language (e.g., English) to find content (which might be written or spoken) in another language. CLIR systems can be combined with speech recognition, machine translation, and summarization to build complete information systems that allow people to find and use the content that they need, even if they are not fluent in the language in which that content was created.

We are presently engaged in a multi-year research project in which we are developing deployable CLIR techniques. In our project we can flexibly configure systems to accommodate different needs (e.g., simple or complex query languages, narrow or broad searches) and we leverage system combination for both search and summarization, using multiple machine translation systems and complementary ways of doing content processing in order to optimize search quality. We are able to rank retrieved documents for interactive use, or to select sets of retrieved content for further processing (e.g., for topic modeling or entity linking, although both tasks are outside the scope of our present project). Our current systems work with both digital text and digitized speech, and in prior work we have extended similar systems to scanned text (using OCR, with appropriate error models). Prior CLIR systems have been extensively optimized for specific language pairs, but our approach is designed to be extended to new

languages with minimal effort.

Our present work is supported by the IARPA MATERIAL program (MAchine Translation for English Retrieval of Information in Any Language), and it has been tested on content sampled from multiple genres. We are interested in participating in the workshop to learn about NASA information management applications beyond those we have already thought of that could benefit from CLIR, and to better understand any unique characteristics of such applications that could help to shape our future work.

References

[1] Edward J. Hoffman and Jon Boyle. Tapping agency culture to advance knowledge services at NASA. *Public Manager*, 42(3):23, 2013.

[2] Roger D. Launius. We can lick gravity, but sometimes the paperwork is overwhelming: NASA, oral history, and the contemporary past. *Oral History Review*, 30(2):111–128, 2003.

[3] Michael L. Nelson, Gretchen L. Gottlich, David J. Bianco, Sharon S. Paulson, Robert L. Binkley, Yvonne D. Kellogg, Chris J. Beaumont, Robert B. Schmunk, Michael J. Kurtz, Alberto Accomazzi, and Omar Syed. The NASA Technical Report Server. *Internet Research*, 5:25–36, 1995.