Position Statement

Implementable Lexical Interlingua for Translingual Information Access and Its implications on Resource Sharing

Young-Suk Lee
Information Systems Technology Group, MIT Lincoln Laboratory
YSL@LL.MIT.EDU

In developing an interoperable US/Korea joint cross-lingual digital libraries, the most critical areas of research include translingual information detection, extraction and summarization as well as machine translation. This is because the users of each language need to access information in other languages (translingual information detection) and have the information presented in the user's own language, most preferably in a summarized format (information extraction, summarization and translation).

As is well-documented in the literature, however, cross-lingual information retrieval using the traditional term translation method performs only at most 60% as well as mono-lingual information retrieval in terms of precision measures. And to have a truly interoperable efficient translingual information retrieval system, an alternative approach needs to be sought, which avoids the problems of incorrect translation and lack of translation terms. My position on the alternative method is concept-based approach to translingual information retrieval for which both the query terms and the database to be searched for are "*translated into*" language transparent concepts and the search is based on concept matches rather than word string matches. Advantages of the concept-based approach over the traditional query term translation method is that it becomes possible to acquire rules from annotated corpus, which can predict the category of new words (avoiding the problem of missing translation terms) and the rules which disambiguate the categories of ambiguous words (mitigating the problem of incorrect translations). In addition, our experimental results indicate that concept-tagging improves the performance of extraction (i.e. summarization) system when adequately combined with term frequency methods.

The current view on concept-based approach to translingual information detection, extraction and summarization implies that the most critical resource to share in the US-Korea joint effort on digital libraries is "***bilingual lexicons***" from which we can derive other resources needed for translingual information access including concept lexicons and translation grammars/lexicons. Once we identify large scale bilingual lexicons, and have them available for collaborative digital library R&D purposes, the subsequent hurdles for successful implementations of translingual digital library become easily overcome by utilizing the currently available translingual information technology such as **CCLINC** (Common Coalition Language System at Lincoln Laboratory).

Technical Biography

Young-Suk Lee received a B.A. degree in English from Seoul National University, Korea. She came to the U.S. in 1987 to pursue her higher education and received an M.S.E. degree in Computer and Information Science, and a Ph.D. degree in Linguistics from the University of Pennsylvania. From 1994 to 1995, she was a lecturer in linguistics at Yale University. Since she joined Information Systems Technology Group (http://www.ll.mit.edu/) at MIT Lincoln Laboratory as a technical staff member in 1995, she has been conducting research on machine traslation of text and speech, translingual information retrieval, information extraction and summarizaiton, and speech recognition. She has developed and implemented the algorithms for automatic acquisition of lexicons/grammars for machine translation, which expedites the porting of the system to new domains and languages. She has also been developing algorithms which combine statistical modeling and linguistic understanding to produce high precision language/domain independent information extraction and summarizaiton systems as well as information retrieval system. She has also implemented an integrated CCLINC (Common Coalition Language System at Lincoln Laboratory) translingual information system which combines human/human interactive system via speech-to-speech translation and human/machine system for translingual information detection, extraction and summarization. She is a member of Association for Computation Linguistics, and is currently serving on the IRAL (Information Retrieval with Asian Languages) program committee.