US-Korea Joint Workshop on Digital Libraries
San Diego, August 10-11, 2000

Multilingual Natural Language Processing

Ulf Hermjakob

USC Information Sciences Institute
ulf@isi.edu · http://www.isi.edu/~ulf/

Multilingual digital libraries face particular challenges, including character sets and their encoding, machine translation, and cross language information retrieval. To meet these challenges, research in computational linguistics has moved more and more towards empirical approaches, including statistical models and machine learning. While these approaches are often relatively language-independent in principle, they typically share the need for language specific training data for applications in specific languages. Empirical parsing for example relies on treebanks, and machine translation on large parallel corpora (and often even larger monolingual corpora for the target language). Other critical resources include monolingual lexicons and semantic ontologies, as well as extensive dictionaries for multilingual applications.

As an example of adapting machine-learning based natural language techniques that were initially developed for English, we ran a three-month three-person research project in 1999, in which we annotated 1187 sentences from the Korean newspaper Chosun and used the resulting treebank to train a Korean parser, achieving word level parse accuracy of 89.8% recall and 91.0% precision. For initial pre-processing, the parser uses the KMA segmenter and morphological analyzer and KTAG tagger provided by Prof. Rim of Korea University.

More generally, at the Natural Language Processing research group at the USC Information Sciences Institute, we have pursued various applications of human language, including question answering from the web, machine translation, automated text summarization, information retrieval (from web and text collections), and large ontologies of semantic (meaning) symbols. Most of the research had a multilingual focus, with various projects covering English, Spanish, Japanese, Mandarin Chinese, Korean, Arabic, Bahasa Indonesian and French. ISI's current Webclopedia project, part of DARPA's TIDES program, uses our empirical NLP approaches to answer natural language questions with short (<= 50 byte) answers from web documents, at first in English, and then with cross-language answer retrieval at a later project stage.

Dr. Ulf Hermjakob is a senior research scientist at the Information Sciences Institute of the University of Southern California. He has worked on machine-learning based deterministic parsing for English, Japanese and Korean, on machine translation, and is currently also on ISI's Webclopedia project on question answering. Dr. Hermjakob obtained his Master's in Computer Science from the University of Karlsruhe in Germany and his Ph.D. (1997) from the University of Texas at Austin with a thesis on "Learning Parse and Translation Decisions From Examples With Rich Context".