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

Multilingual Natural Language Processing

Ulf Hermjakob

USC Information Sciences Institute
http://www.isi.edu/natural-language
ulf@isi.edu

Multilingual digital libraries face particular challenges, including character sets and their encoding, machine translation, and cross language information retrieval. To meet the challenges and complexity of machine translation, research in computational linguistics has moved more and more towards empirical approaches, following a trend set by speech recognition, where empirical approaches are clearly dominant, as well as in parsing, which is increasingly based on statistical (Collins, 1997; Charniak, 2000) and machine learning (Hermjakob, 2000) approaches. Building on initial research by (Brown, 1990; Berger, 1994), two examples of current statistical machine translation research are (Knight, 1999) and (Ney, 2000).

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.

This shows that Digital Libraries and Natural Language Processing can *mutually* benefit each other. Digital Libraries can clearly profit from machine translation, while various subtasks of Natural Language Processing can greatly benefit from resources such as bilingual lexicons and parallel corpora, for which Digital Libraries are a natural repository. The size of parallel corpora needs to be quite large, hundreds of thousands or even millions of sentences. It is also desirable that these resources mirror application domains.

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 (WEBCLOPEDIA project), machine translation (GAZELLE, QuTE, EGYPT projects), automated text summarization (SUMMARIST), information retrieval from web and text collections (MuST/C*ST*RD), and large ontologies of semantic (meaning) symbols (SENSUS). 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.

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.

———————————————————

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”.

References

A. Berger, P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. R. Gillett, J. D. Lafferty, R. L. Mercer, H. Printz, L. Ures 1994. The Candide system for machine translation. In *Proceedings of ARPA Workshop on Human Language Technologies.*

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, P. Roossi 1990. A Statistical Approach to Machine Translation. In *Computational Linguistics 12 (2)*, pages 79-85

E. Charniak 2000. A Maximum-Entropy-Inspired Parser In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 132-139.

M. J. Collins 1997. Three Generative, Lexicalised Models for Statistical Parsing. In *35th Proceedings of the ACL*, pages 16–23.

U. Hermjakob 2000. Rapid Parser Development: A Machine Learning Approach for Korean. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 118-123.

K. Knight 1999. Decoding Complexity in Word-Replacement Translation Models. In *Journal of Computational Linguistics, 25(4).*

C.-Y. Lin 1998. Assembly of Topic Extraction Modules in SUMMARIST. In *Working Notes of the AAAI 1998 Spring Symposium on Intelligent Text Summarization.*

H. Ney, F.-J. Och, C. Tillmann, S. Vogel Oct. 2000. Statistical Translation of Spoken Dialogue in the Verbmobil System. In *Proc. of Workshop Multilingual Speech Communication.*