MATES/CK: A Chinese-Korean Machine Translator

Jin-Xia Huang[*,**]
[*]KORTERM, AITrc, Computer Science Department, Korea Advanced Institute of Science and Technology
373-1 Yusong-gu, Gusong-dong, Daejon 305-701, Republic of Korea
[**]Yanbian University of Science & Technology
Yanji City, Jilin Province, 133001, P.R.China
korterm@korterm.kaist.ac.kr

Hyi-Jeong Song, Yeong-Mi Song, Ji-Hyoun Kim, Won-Sek Kang, Chung-Won Seo, Young-Souk Chae, Key-Sun Choi
KORTERM, AITrc, Computer Science Department, Korea Advanced Institute of Science and Technology
373-1 Yusong-gu, Gusong-dong, Daejon 305-701, Republic of Korea
korterm@korterm.kaist.ac.kr

Abstract

A Chinese-Korean Machine Translation System MATES/CK has been developed by KAIST (Korea Advanced Institute of Science and Technology) since 1996. A prototype system of MATES/CK was originally implemented by the end of 1997 (Li and Choi, 1997), in which the example-based MT methodology was adopted. The second version of MATES/CK was demonstrated in the exhibition of MT Summit' 99 (Singapore). After that, we have made a constant effort to improve the performance of the system in some aspects, including sentence/phrase pattern based structural transfer, statistical and rule based Chinese POS tagging, statistical pruning with attribute knowledge.

MATES/CK system follows the typical three-phase scheme (analysis/transfer /synthesis) of a conventional transfer-based system. Some distinctive features are proposed and employed to improve the system performance, including "pattern-based and statistics-oriented" hybrid design (Zhang and Choi, 1999), a method for quantifying the attribute knowledge by the classification between "weakly-restricted" and "strongly-restricted" attribute knowledge, a flexible pattern match algorithm and a transfer model based on sentence and phrase patterns, a statistical information based lexical selection algorithm, and a rule and phonetic based morphology selection and generation algorithm.

The analysis module of MATES/CK is composed of a Chinese segmentation module, POS tagger and a Chinese parser. Segmentation and POS tagging model is based on the quantitative statistical analysis algorithm with the rule-based disambiguation method. FB (Forward-Backward) algorithm with the confidence interval in the parameter estimation is employed in the statistical analysis. A rule-based method is adopted in the disambiguation of Chinese word segmentation and POS tagging.

Because a complete parsing is not always necessary for MT, in MATES/CK, the output of analysis module is either a syntactic tree or a chunk list with a predicate-argument structure. To construct the candidate set of Chinese syntactic trees by means of GLR algorithm with "weakly-restricted" attribute pruning technique. The "statistics-oriented" technique is proposed to select the most preferred candidate.

Transfer module consists of a lexical selection component and a pattern based structural transfer component. Two kinds of patterns are employed in structural transfer module, one is parameterised flexible sentence pattern that is both for the source sentence failed in parsing (so the input will be POS tagged sentence or Chinese pruning list) and syntactic tree gotten through successful parsing, and the other one is called phrase pattern that is only for the syntactic tree gotten through successful parsing. All of the sentence patterns will be parameterized and be given priority by the lexical/POS tag/phrase tag information that they contain. Instead of sentence patterns, phrase patterns are prepared for the purpose of matching with syntactic trees that fails in sentence pattern matching. This means that the transfer module gets fragments of Chinese syntactic tree and transfers to their corresponding Korean syntactic tree. This process assumes that the syntactic parsing is successful. In pattern matching, we employed flexible pattern match algorithm.

Lexical selection is implemented by using Viterbi algorithm and a bilingual dictionary with probabilistic information getting from word-aligned bilingual corpus.

Synthesis module consists of a generator and a Korean morphological table. A rule based morphology selection and phonetics based morphology generation method are adopted.

As a prototype machine translation system, MATES/CK provides some useful tools and functions: Chinese-analysis dictionary editor, Chinese-Korean bilingual dictionary editor, Chinese POS tagger, Chinese parser, Chinese example search (by word or CFG rule), LR table generator, and so on.

Under the framework, several related researches have been undertaken, including the constructions of the Chinese/Korean/English trilingual aligned corpus and Chinese Tree bank, Chinese-Korean word alignment (Huang and Choi, 2000), auto-construction of phrase pattern through word/phrase alignment (Huang, 2000), and some promising results have been obtained.

In this demo, we will show the MATES/CK system and some key techniques employed in it.

Acknowledgements:

This work was supported by the Korea Science and Engineering Foundation (KOSEF) and the Advanced Information Technology Research Center (AITrc).

Reference

Li, Junjie and Key-Sun Choi (1997). Corpus-Based Chinese-Korean Abstracting Translation System. *Proceedings of IJCAI-97, Nagoya, Japan*

Zhang, Min and Key-Sun Choi (1999). Pattern-based and statistics-oriented Chinese-Korean Machine Translation. *Proceedings of MT Summit 99, Singapore*

Huang, Jin-Xia and Key-Sun Choi (2000), Chinese-Korean Word Alignment Based on Linguistic Comparison, *Proceedings of ACL-2000, HongKong*

Huang, Jin-Xia (2000), Auto-Construction of Phrase pattern through Word/Phrase Alignment, *Masters thesis*