[MT Summit IV, July 20-22, 1993, Kobe, Japan]

The R&D Activities of MT in China

Yang Tianxing
Director of the Department of Computers,
Ministry of Electronics Industry, P.R.C.

The machine translation is one of the practical technologies. In order to meet the needs of opening to the outside world and promoting international technical exchange and cooperation, the R&D of MT systems have been considered as a key link in developing China's computer application technologies.

In recent years, we have made some new progress in the R&D of MT and commercialization of MT. The TRANSTAR, the first commercialized system in China, has been improved greatly in both the speed and the accuracy of translation, through upgraded and revised several times. The R&D of the intelligent MT system is one of items for the National High-Tech Development Programs. As the achievement in the first stage, the pocket translator has been released in market. Some new MT prototypes, such as CEMT-3 Chinese-English MT system used for translating the documents on space technology, have been appraised by experts.

In the period of the Eighth Five-Year Plan (1991-1995) , we are engaged in developing Chinese Information Processing platforms in order to support the R&D of various application systems, such as the machine translation,the natural language interface, the intelligent full-text retrieval system and etc. The project is organized by the Research Center of Computer and Microelectronics Industrial Development (CCID), MEI. The members of research institutions include some outstanding universities, e.g. Qinghua University, Beijing University, Beijing Languages Institute, Northeast University, and so on. Prof. Chen Liwei, the President of Chinese Information Processing Society of China, is the senior adviser of the project. The purpose of the project is the R&D of the fundamental and key technologies on Chinese Information Processing, so as to set up an environment to develop the NLP application systems.

The Goal of the Project:

a. To develop a large-scale Chinese lexicon used for natural language processing.

b. To develop a Chinese syntactic and semantic rule base which covers the various language phenomena.

c. To establish a Chinese corpus which supports the natural language processing.

d. To develop the basic software of natural language processing.

We are trying to make some breakthrough both theoretically and technically. In order to find an effective approach of Chinese information processing, we combine the two presently active approaches, i.e., the rationalism based on linguistic theory and the empiricism based on analysis of large corpora, so as to make use of the strength of both.

In this project, the establishment of large corpora and the processing technique of corpora are emphasized. At present, we have made considerable progress on the part-of-speech tagging system, the phrase tagging system, the example-based sense tagging system, and so on. For example, the part-of-speech tagging system, using bigram model and a tagset of 113 grammatical categories, can perform an overall accuracy of 97% for tagging the Chinese running text. It is estimated that, by the end of the Eighth Five-Year Plan, a grammatically tagged corpus with 5 million Chinese characters will be completed.

International Exchange and Cooperation

Unlike Indo-European languages, Chinese is a language which has no inflexion and emphasizes the meaning. In order to enhance the capability of Chinese analysis, many institutions in China are engaged in studying in Chinese semantics to find an effective solution. For example, Qinghua University in Beijing is developing an example-based sense tagging system and CCID is developing a large scale Chinese semantic electronic dictionary.

I hope the above-mentioned efforts may contribute to the semantic dictionary research work as well as the R&D of the large scale knowledge base in other countries.

I also agree with Prof. Tanaka who proposed a few year ago the establishment of the International Research Center which is devoted to the R&D of the large scale electronic dictionary and knowledge base. I quite agree with his view point, and hope it will be carried out soon.

Since 1987, China have taken part in the R&D of Multilingual Machine Translation Project initiated by MITI of Japan and organized by CICC. This is the first time for our country to join in a large international project in MT fields. Thanks to the Interlingua method of the project, we can devote ourselves to developing MT technologies related to our own native language and can cooperate with other countries to carry out the project. At present, the project is running very well. Eight institutions and several tens of researchers in China participate in the project, which is under the guidance and support from Japanese experts and scholars of some large institutions in Japan. The CICC project has brought MT and NLP in China up to a new level.

In conclusion, I hope through this symposium, the international exchange and cooperation in the field of MT between China and Japan as well as other countries will be further promoted.