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

Multi-Lingual Machine Translation (MMT) Project

Susumu Funaki
General Manager, R&D Division
Machine Translation System Laboratory
Center of the International Cooperation for Computerization (CICC)

1. Preface

The Center of the International Cooperation for Computerization (CICC) was established on June 1, 1983, with the support of Ministry of International Trade and Industry (MITI) and the Japanese computer industry. Its purposes are promotion of computerization cooperation with developing countries, enlightenment of developing countries regarding computerization, guidance and promotion of international exchange related to international cooperation for computerization, etc. These activities of CICC are supported by MITI and seventy private companies.

As a part of these activities, CICC has in the past conducted study and research on the word processor system for the mother language, and study and research on a machine translation system for Asian languages. On the basis of know-how and information accumulated as a result of these activities, CICC has since 1987 put into practice a project called "The Research and Development Cooperation Project on a Machine Translation System for Japan and its Neighbouring Countries". This project has been entrusted to CICC from the Ministry of International Trade and Industry.

In this project, five languages of Asian countries are the subject of translation: Chinese, Indonesian, Malaysian, Thai, and Japanese. Moreover, an "interlingua" method suitable for mutli-lingual machine translation has been adopted for the development of the machine translation system.

The Machine Translation System Laboratory established in CICC has been implementing the joint research with official research institutes of the five participating countries. The development was begun in 1987. The first two years were spent on basic study and research, full-scale development was commenced in the third year, and the last two years will be devoted to system improvement.

2. Necessity and Significance of the Project

The economic activities of Japan's neighbouring countries - China, Southeast Asia, etc. - have become remarkably active in the recent years. As a result, economic, technological, and cultural exchanges between these countries and Japan have escalated with astonishing speed. For the further development of these neighbouring countries, technology transfer from Japan and other developed countries is essential. Also necessary are the smooth and extensive introduction of technological information that can constitute a base for technology transfer. Under such circumstances, there is a great demand for efficient translation at low cost.

Under the present conditions, however, few translators are engaged in translating Japanese into the languages of its neighbours; even fewer translators are equipped with technical knowledge. Therefore, information exchange has relied on translation through English, a foreign language for both

sides. In the future, international exchange will involve more extensive _elds and levels. Accordingly, a translation method superior to the present in speed, cost, and accuracy, will be required to overcome the difficulties involved in learning English as a foreign language.

The present Machine Translation System project was launched with the following expectations:

(1) It will be possible for a computer to translate large quantities of material at high speed and low cost,

(2) Translated words and style will be more uniform compared to manual translation,

(3) Personal translation experience and skill can be combined with the system, thus allowing them to be jointly possessed by society.

Although this project is a research and development in a high-technology field of natural language processing by computer, it is centered around the development of dictionary and grammatical rules for each language and joint research on interlingua. Researchers of the participating countries, therefore, assume a vital role in the realization of this project.

By promoting this project, we can expect that exchange and technology transfer from Japan to neighbouring countries will be implemented in the following aspects:

(1) A more active exchange of technology and culture will be enabled between Japanese and languages of neighbouring countries through the use of machine translation system, which is the fruit of our research and development

(2) Information processing technology will be transferred through the joint development of an advanced information processing system

(3) Information processing technology platform will be established extensively in the countries involved, centering around the technology of information processing in the mother language as a result of our research and development.

3. The Machine Translation System Developed in this Project

The technical target of this project is to realize a mutli-lingual machine translation system for five languages based on interlingua. The project will prove that the system has the ability to translate documents at practical accuracy and speed. The project has selected the field of information processing as a model to prove the system's ability. The translation speed must be five thousand words per hour using a workstation. Its accuracy must be 80 to 90 per cent on condition that the original sentences are grammatically correct and all the words used in the sentences are included in the system's dictionary. Moreover, the system shall be composed so that a translator can participate in editing at any time to improve the quality of the output.

The functions of each subsystem composing the machine translation system developed in this project are as follows:

(1) Input system deals with the input of sentences, tables and figures. For this purpose, word pro-

cessor and OCR are developed. The input sentences will be displayed on the CRT and will be checked by an operator skilled in translation. Depending on the situation, pre-editing and modification, etc. will be done on the sentences so that they can be treated mechanically.

2) The sentence analysis system has the function of converting the input sentences into interlingua through morphological analysis, syntax analysis, and semantic analysis. During such processes, an electronic dictionary system and sentence analysis rules are used.

3) The interlingua is a mechanism for conducting mutli-lingual translation. Input texts will be converted into interlingua and subsequently translated into the target language.

4) The electronic dictionary system stores descriptions on the correspondence between each language and the interlingua, and controls such necessary data for machine translation as grammatical and semantic information. Fifty thousand terms are to be registered in the basic dictionary of each language along with twenty-five thousand words associated with information processing in the technical term dictionary.

5) The sentence generation system will generate the target language from the results of analysis written in interlingua. The processes employed for this are sentence style generation, syntax generation and morphological generation.

6) Output system and the translation support system will output finished translated sentences. These sentences should be edited with modifications, etc. by the operator as required.

7) The integration system will unite the functions enumerated above and the text file management functions of each language by a network, realizing one translation system.

4. Interlingua Method and Transfer Method

Compared to transfer method, interlingua method has the advantage of conducting mutli-lingual machine translation efficiently. However, transfer method has the ability to translate one pair of languages at high quality and low cost; especially when the two languages have similar linguistic characteristics.

<Advantages of interlingua method>

(1) Efficient system development

When developing a mutli-lingual machine translation system, the number of analysis/generation systems required will be fewer, therefore the development will be more simple and efficient. For example, in the case of transfer method, ten systems will be necessary for translating
five languages; whereas in the case of interlingua method, five systems will be sufficient for the same purpose.

(2) Easy expansion of subject language

In order to add a new language to be translated, five systems will have to be developed in transfer method. On the other hand, in the case of interlingua method, a system between the

new language and interlingua will be necessary.

(3) Localized development

In the case of transfer method, the two countries will have to work jointly for development. In the case of interlingua method, each country can develop their corresponding systems separately. Furthermore, the developer is not required to have a thorough knowledge of the other language.

5. R&D Organization

The research and development have been done by the Machine Translation System Laboratory of CICC, under the direction of the Ministry of International Trade and Industry. Guidance and advice has been provided by the Electrotechnical Laboratory belonging to the same ministry. This project is the subject of joint research and development between Japan and four neighbouring countries. In these partner countries, the following institutes have taken charge of development.

(1) In China: China National Computer Software and Technology Service Corporation (CS&S)

(2) In Indonesia: Agency for the Assessment and Application of Technology (BPPT)

(3) In Malaysia: Ministry of Education (MOE)

(4) In Thailand: National Electronics and Computer Technology Center (NECTEC)

The Machine Translation System Laboratory also has a research partnership with the Japan Electronics Dictionary Research Institute, Ltd. for the development of dictionaries. The MTSL also enjoys the participation and support of companies related to computers for system development in general. Moreover, since 1990, the MTSL has entrusted the integration system development concerned with the interfacing of computers to the Interoperability Technology Association for Information Processing.

6. Development Schedule and Results

The term of research was originally planned as six years. However, owing to delay in the establishment of R&D organization at the initial stages of the project and shortage of budget, the project term has been extended to eight years, which covers three principal steps. The first phase examines the system's basic elements (1987 to 1988), the second phase handles system development (1989 to 1992), and the third phase is for system improvement and proving (1993 to 1994).

Performed in the first phase were the formation of a dictionary with five thousand basic terms, trial compilation of the grammar based on a few model sentences, and a verification test using these sentences to illustrate that the basic system can function. The first verification test was conducted in November 1988.

Being performed in the second phase are expansion of the basic dictionary to fifty thousand terms, preparation of the technical term dictionary with twenty-five thousand terms, analysis/generation rules based on three thousand sample sentences, compilation of the grammar, and the development

of input/output system and its translation support system. In this phase, translation test using unspecified natural sentences can be started. A verification test to prove the system's validity was conducted in 1990, and the third verification test was conducted in December 1992.

To be accomplished in the third phase are improvement of dictionaries and grammar, improvement of the system, and testing of translation using the system developed in the previous steps. The fourth verification test is scheduled in 1994.

The target for the electronic dictionary system is to register fifty thousand basic terms and twenty-five thousand technical terms in the field of information processing. The development of about forty thousand basic terms and about twenty thousand technical terms were finished in 1991. By the end of 1992, following the same line of development, fifty thousand basic terms and twenty-five thousand technical terms will be registered. Development has begun for the concept system which describes the vertical relationships between concept types included in these dictionaries, and a concept dictionary containing the concept description which explains the relation between concepts.

For the sentence analysis/generation systems, development of analysis and generation rules each based on about three thousand sample sentences were conducted. These rules were inserted into the system, then evaluated and improved with a mass of evaluation sentences. The compilation of these analysis/generation systems was started in 1989. In 1991 through the actual and full-scale assembling of these grammatical rules, the formation of analysis/generation rules each based on about one thousand eight hundred sentences in each language were completed and inserted into the system. A machine translation system with rules based on six thousand sentences is said to have the ability to translate natural natural sentences.

The input/output system, the translation support system, and the integration system, have been developed experimentally and improved since 1989.

From 1992, full-scale OSI development has been conducted as an integration system development link to construct an international network that will make possible data exchange between different types of machines.

Executed in the last phase will be the improvement of dictionaries and grammatical rules through a mass of evaluation sentences. Also to be developed are supporting systems for dictionary development and for grammatical development, with abundant functions. These developments can be implemented individually in each participating country.

This project is designed to transfer technology through joint research and development process. For this purpose, the research staff of the participating countries are invited to Japan and Japanese researchers are also dispatched to national laboratories of these countries. Each year about one hundred researchers are invited to Japan, and the same number of Japanese are dispatched to these countries for joint research. This latter includes the establishment of annual planning, and arrangement of research work with the overseas researchers.

Equipment such as workstations and personal computers have been installed in the CICC laboratory for the development of the present system. Equipment of the same type has been transferred to

the partner countries so that the R&D environment can be equal among participating countries.

7. Future Goals

This project, started in 1987, has finished the scheduled two-year development stage at the end of 1990, and has commenced the last step of improvement this year.

In the improvement phase, the following emendations are considered as future targets:

(1) An improved interlingua more suitable for mutli-lingual machine translation

(2) An improved electronic dictionary system that facilitates more apropos expressions

(3) An improved sentence analysis/generation system that can translate longer and more complex sentences

(4) An improved, more combinable translation system to operate the electronic dictionary system, sentence analysis system, and sentence generation system

(5) An improved supporting system for electronic dictionary development - one that can facilitate system improvement and expansion

(6) An improved, easier-to-use translation support system and input/output system

(7) An improved integration system for the construction of an international network that implements data exchange between disparate machines

Notwithstanding future budget considerations, by the end of 1994 we envision translation from Japanese to Chinese, Indonesian, Malaysian and Thai (and vice versa) on nearly the same level as those of presently marketed translation systems for Japanese and English when they first appeared on market.

The target of this project is to hurdle the language barrier that inhibits international technological exchange by developing a viable machine translation system. After accomplishing this it will be important to diffuse widely the results of the project and to adopt it efficiently. The countries participating in the R&D have proceeded with vigor. Yearly promotional discussions have been held. International bodies such as UNESCO are also interested in this project. Thus, an important sequel is to draft and to settle upon specific plans for efficient use of the product.

Taking every opportunity, we must obtain a consensus about the practical use of the system, devoting careful attention to the opinions of representatives of a wide spectrum of fields.