RESEARCH AND DEVELOPMENT OF COOPERATION PROJECT ON A MACHINE TRANSLATION SYSTEM FOR JAPAN AND ITS NEIGHBORING COUNTRIES

Hozumi TANAKA, Shun ISHIZAKI, Akira UEHARA, Hiroshi UCHIDA

Machine Translation System Laboratory
Center of the International Cooperation for Computerization
30-9, Shiba 5-chome, Minato-ku Tokyo, 108 Japan

1. BRIEF DESCRIPTION OF THE PROJECT

This project has been promoted by the Japanese Ministry of International Trade and Industry as the Research and Development Cooperation Project on a Machine Translation System for Japan and its neighboring countries.

The objective of this project is to verify the feasibility of commercializing machine translation systems. The project particularly aims at translating multiple languages through the Interlingua approach. At present, five languages, Chinese, Indonesian, Malaysian, Thai and Japanese, have been selected for machine translation under this project.

The task to develop a machine translation system has been contracted with the Center of the International Cooperation for Computerization (CICC). The Machine Translation System Laboratory in the CICC will develop the system in cooperation with the various organizations of the concerned countries. The laboratory has been conducting joint research with the official research organizations of the People's Republic of China, Republic of Indonesia, Malaysia and Kingdom of Thailand. This is an international research cooperation project involving five countries.

The period of the project is scheduled to be six years, beginning in 1987. The first two years have been devoted to basic research. The third and subsequent years are for full-scale development. Approximately two years in the second half of the development period will be devoted for system improvements.

2. DESCRIPTION OF THE PROJECT

(1) Goals of the Project

The evolution of the international economic society is supported by harmony and cooperation by the various countries of the world. International interchanges in the industrial sector must become more active and deep through cooperation activities and joint research and development projects.

The language problem must be addressed in undertaking this project. The use of English as a common language has limitations. As interchanges deepen and the number of projects increases, the languages of the other countries need to be understood more directly. The demands on translation increase as the scales of the respective projects grow. Much of the translation work requires technically specialized knowledge, and specialist translators will be needed. Can these large volumes of information be translated efficiently by using computer technology?

The purpose of this project is to verify the feasibility of commercializing a machine translation system covering the languages of five Asian countries. The languages included are Chinese, Thai, Indonesian, Malaysian and Japanese.

As a technical goal, translation among these languages is aimed at. To achieve this, the languages to be translated are temporarily translated into a common language, called the Interlingua language, from which the target languages are generated.

The basic system aims at translating 5,000 words per hour with an accuracy of more than 90%, provided that the original texts are grammatically correct and all the words in the original texts are contained in the system dictionary, using Unix workstations. To further improve quality of translation, the system will be constructed to allow translators to participate in editing at any time.

The basic system will have a basic dictionary of 50,000 words and 25,000 technical term dictionary for information processing for each language. The grammatical rules will initially be designed for accepting approximately 6,000 sample sentences. Grammar will be improved in the system by inputting a large volume of sample sentences for evaluation.

If these goals are reached, the technical and economical prospects of developing a practical translation system for this language are considered to have been accomplished.

(2) Development Schedule

The research and development period is scheduled for six years, which is broken down roughly into three stages.

In the first stage(1987 to 1988), a basic dictionary of 5,000 words will be created and grammar will be test produced based on about ten sample sentences. In the demonstration in November, 1988, translation tests were made using about ten limited sample sentences.

In the second stage (1989 to 1990), the basic dictionary will be expanded to 50,000 words, and technical term dictionary of 12,500 words will be incorporated. Grammar will be developed based on slightly more than 3,000 sample sentences. The input and output technology and translation-support technology will be developed. In this stage, translation experiments of selected sample natural sentences can be started. A verification test will be conducted by the end of March, 1991.

In the third stage(1991 to 1992), the technical term dictionary will be expanded to 30,000 words. Using the system developed in the previous stage, a large-scale translation experiment will be repeated to improve the dictionaries, the grammar, and the overall system. The system viability will be verified in 1992.

3. DEVELOPMENT ORGANIZATION

This program is jointly studied by five countries; the People's Republic of China, Republic of Indonesia, Malaysia and Kingdom of Thailand, and Japan to develop machine translation systems.

The Machine Translation System Laboratory of the Center of the International Cooperation for Computerization, which is entrusted by the Japanese Ministry of International Trade and Industry (MITI), conducts R & D under the guidance of the Electrotechnical Laboratory of MITI. The laboratory is conducting research in cooperation with the Japan Electronic Dictionary Research Institute in developing electronic dictionary technology. The computer are also cooperating in this project.

In the four Asian countries directly participating in this project, the following research organizations are working under the guidance of the governments in these countries. :

China - CSTC; China Software Technique Corporation
Thailand - NECTEC; National Electronics and Computer Technology Center
Malaysia - MOE; Ministry of Education
Indonesia - BPPT; Agency for the Assessment and Application of Technology

4. DESCRIPTION OF RESEARCH AND DEVELOPMENT

(1) Organization of the Machine Translation System

The machine translation system developed by this project consists of the following elements:

1. Input system, translation support system
2. Sentence analysis system

3. Interlingua (intermediate language)
4. Electronic dictionary system
5. Sentence generation system
6. Output system, translation support system

The functions of these elements are as follows:

1) The input system inputs texts including tables, list and graphics. The word processor and OCR technology have already been developed. Input texts are checked on the screen by operators who are capable of translating and who edit and modify them if necessary.

2) The text analysis system analyzes input texts by such processes as morpheme analysis, syntactic analysis and semantic analysis and it converts them into Interlingua. During this process, the data of the electronic dictionary system and grammar rules will be used.

3) Interlingua will be the kernel for translation among many languages. The texts to be translated will be temporarily translated into Interlingua and will then be translated into the desired language.

4) The electronic dictionary system will describe the correspondence between the various languages and Interlingua. Information such as grammar and semantic information needed for machine translation will be managed. The basic dictionaries for the four languages will contain 50,000 words each; the technical term dictionaries for the information field will contain 25,000 words each.

5) The sentence generation system will generate various target languages using grammar and morpheme generation processes based on the results of analysis, namely Interlingua.

6) The translation support system outputs translation results. Output texts will be checked on the screen by operators who are capable of translating and editing them.

7) The system integrates all of the functions mentioned above, as well as the document file management system for the various languages, to work as one networked machine translation system.

(2) Progress of Research and Development

The status of research and development for 1988 is outlined below.

1) The development of Basic System

The basic system was developed in 1988 based on the accomplishments made in the previous year. The analysis, dictionary, generation, input and output, and translation support systems were tentatively manufactured and they were integrated into a tentative total system. The verification tests were made using a limited number of sample sentence.

This work has been conducted jointly by inviting a large number of researchers from the other four countries to Japan. The sample sentences and Interlingua used in the verification tests followed the guidance of the Electrotechnical Laboratory.

2) Generating Dictionary Data

The number of words of the basic dictionaries for each language is approximately 5,000 words. Japan Electronic Dictionary Research Institute (EDR) is providing cooperation to build the basic dictionaries.

3) Study of Interlingua

Overview of our Interlingua will be shown in section 5.

5. INTERLINGUA

Interlingua has to represent all information expressed in sentences. The most important information in a sentence is so-called semantic content based on deep case relations that represent binary relations between two concepts. Interlingua consists of following four kinds of information.

(1) Events and facts

Events and facts are expressed in the same representation form as the concept dictionary developed at EDR. Five relations are used to describe events and facts:

1) Case relation (agent, object, manner, implement, material, time, time-from, time-to, duration, location, place, source, goal, scope)
2) Event relation (condition, co-occurrence, sequence)
3) Semantic relation (part-of, element-of)
4) Constraint relation (frequency, unit, quantity, number, modify, standard)
5) Quantifier (all, some, each, ……)
6) Other relations (possessor, purpose, degree, and, or, cause)

(2) Speakers View (tense, aspect, ……)

(3) Intention (imperative, question, exclamation, emphasis, topic, focus, .…)

(4) Structure of sentences

We expect our Interlingua will be refined by our project through which we will have many experiences of Interlingua.

6. CONCLUSION

A multi-lingual machine translation has many difficult problems. Especially, "Interlingua" approach seems to be a difficult research theme, and we should do more research on the theme. We should not abandon the Interlingua approach because of the difficulties. Through conducting our project, many problems will be clear and we will recognize what is the most important problem. Finally, we are going to open the Translation Centers in the participating countries to use them as bases for system expansions in the future and to provide machine translation service actually as public service.