New Developments in Machine Translation Technology

Hozumi Tanaka
Graduate school of Information Science and Engineering
Tokyo Institute of Technology

1. Introduction

The project organized by the CICC and neighboring Asian countries to develop a multilingual machine translation system is approaching its final stage. When it started in 1987, Japan was most active in the field of machine translation research and ranked among the leaders in technological attainments. Its Asian neighbors had been on the course of rapid economic development, and mutual exchange in the areas of technology, economics, and culture had become increasingly important. To facilitate exchange, the communication gap caused by language must be filled. Machine translation is the key to resolve this problem.

As economies around the world grew in scale, the importance of machine translation has become recognized globally. This led to the first international conference called Machine Translation Summit in Hakone in the fall of 1987, with the support of the Ministry of International Trade and Industry. Subsequent Summits were held in Germany, the U.S., and again in Japan. In 1995, the fifth summit is scheduled to be held in Luxembourg. It has also been recognized that the development of a large, high-quality machine translation dictionary is for developers of machine translation systems a vital matter requiring their urgent attention. The Electronic Dictionary Research Laboratory was established and a large-scale electronic dictionary (EDR dictionary) was developed under the auspices of MITI.

For the future economic development of our Asian neighbors, it is evident that information processing technology will play an important role, and Japan has been asked to transfer its technologies to these countries. It was under such conditions that the effort to transplant machine translation technology was planned by MITI.

Needless to say, machine translation contains leading-edge technologies in the area of information processing. Companies that had announced their cooperation in the MITI project competed to develop the necessary high technology. Yet none of the machine translation technologies have been completed, and there are still unsolved problems. The Machine Translation Project for Asian Languages under the CICC differs from past ODA programs in that it is not a mere transfer of existing technology, but directs attention to new machine translation technology: the multilingual machine translation technology based on Interlingua. Although founded on the existing framework of technology transfer, it is a new attempt to develop innovative technology together with Japan's neighbors.

2. Interlingua

It is globally believed that Interlingua is a technology of the distant future. There was no other active experiment with Interlingua than that in Japan. Because of the choice of system, the project was viewed with suspicion not only by Asian countries but also by the rest of the world. Notwithstanding such resistance, the apparently difficult translation system was adopted because of the following reasons.

(1) The project is the world's first attempt at machine translation between many languages with dissimilar structures, and comprises a prime opportunity to obtain new and further knowledge of Interlingua.

(2) A bold research theme is necessary to achieve dramatic progress in technology. A challenging theme instead of mere transfer of existing technology will foster and contribute to Asian countries the growth of researchers who can lead the world in leading-edge machine translation systems.

(3) There are possibilities of making accomplishments in knowledge representation required for Interlingua. Knowledge representation is the nucleus of artificial intelligence research. The project may lead to the new AI technology believed to be necessary for future information processing.

(4) The use of Interlingua necessitates attention to the problem of semantics. Computer analysis of semantics is directly applicable to research on interactive systems that communicate with computers in natural language.

What has machine translation based on Interlingua become today? Compared to what it was in 1987, this approach has won recognition. Overseas, Carnegie-Mellon University is presently involved in a project that clearly aspires to develop an Interlingua system. A growing number of researchers have shown serious interest in this system as a desired form of multilingual machine translation, rather than a castle in the air. There is no doubt that our machine translation project influenced such new developments.

This writer believes that Interlingua as designed in the project should be documented for the world to know after sorting out its design concept and logical foundation. We have such an obligation. Naturally, Interlingua is still in the process of research and further development. There are numerous shortfalls in the Interlingua developed under the project. There may have been researchers who felt they had been coerced to compromise in design at times and had to design a language distant from their ideal. Notwithstanding, there is no need to fear shortcomings. It should be noted that the project's Interlingua was an invaluable experiment never before attempted. Efforts should be made to disseminate this Interlingua internationally for development of a better Interlingua in the future.

Inasmuch as machine translation is still an incomplete technology, there are technologies that were made immediately effective by employing surface-level processing. One such technology is example-based machine translation. Although such technology is important in view of the infancy of machine translation technology, it should be noted that it does not offer the true essence of machine translation. Machine translation requires that it ultimately establish an extensive natural language processing that extends into semantics. Technology that focuses only on machine translation is unlikely to achieve what had been described in (2) through (4) above.

3. Future Issues

The nucleus of machine translation technology is natural language processing. For greater progress, natural language processing must be developed further. Past technologies innovated for natural language processing should be reviewed, particularly in the area of analysis

technology.

In the conventional classification of natural language analysis technology, it can be divided into morphological analysis, syntactic analysis, and semantic/context analysis. Because these three are intricately linked, research on how to integrate them is essential.

In Japanese, Thai, and Chinese, morphological analysis on spacing between words is necessary. Technology in this area is believed to have been completed for the Japanese language but problems remain. For instance, not all complex nouns can be registered in the dictionary. Hence, such nouns must be broken up and analysed as a sequence of morphological elements in the dictionary. This segmentation is called a morphological analysis of complex nouns. If this fails, it is evident that accurate machine translation will not be achieved. Technology that upgrades the accuracy of morphological analysis must be developed.

An excellent algorithm has been developed and employed as a tool for syntactic analysis. In this author's opinion, attention should be paid to generalized LR algorithm for syntactic analysis. Personal experience indicates that this algorithm is the fastest. In relation to this, research should also be done on how to write grammar. Theory of grammar should not be limited to what is within human understanding. Linguists have worked in the past to establish grammatical theories that man can understand. Research should be directed to development of grammar theory that can work on the computer, and grammar based on the theory in cooperation with linguists and computational linguists.

The third area of semantic/context analysis poses many difficult problems. The concept dictionary developed in the project may help in semantic analysis. For the organization of concept system, it should be developed not only for superordinate/subordinate relationships but for part/whole and homonym/antonym relationships, as well. WordNet under development in the United States may provide ideas in this respect.

Context analysis is not technology restricted to one sentence alone. It must examine the relationship between sentences. In the past, research focused on morphology and syntax, and on structure limited to a single sentence. For instance, the structures of noun phrases and simple sentences have been the focus of attention. In reality, however, text is more complex, with a number of sentences and phrases built into a single text. This leads to the philosophical question of whether close and careful analysis of partial structures will lead to assessment of structure as a whole. In spite of the difficulty, research into context analysis may be necessary in the future.

Research should also be made into ways of making use of the present incomplete machine translation systems. This is linked to the problem of human interface. Evaluation standards for the project should be refined for steady progress in machine translation technology.

For greater sophistication, gathering an electronic corpus in huge volume is necessary. The collection of corpus had been done partially in this project, as well. In the future, corpus of larger quantity and greater quality will be obtained by forming a consortium and sharing it together. Inasmuch as such corpuses will be useful directly for example-based machine translation and disambiguation of gathering and analyzing statistical data. A corpus of several tens of million examples is believed essential.

In connection with corpus, it is important to develop a corpus with a dependence structure such as a bracketed corpus. In this writer's research group, research on extracting grammar automatically from bracketed corpus given by EDR has begun. The results have been quite beyond expectations. If the volume of bracketed corpus increases, grammatical rules will be extracted automatically on a larger scale, and it may become possible to use them in actual machine translation systems. Machine translation requires a corpus that is not only bilingual but also multilingual. Despite its diminutive size, a multilingual corpus has been gathered under the CICC project. It is nevertheless necessary to develop it on a larger scale. This writer wishes to stress the importance because it will provide the energy for the development of a better system in the future.

In addition to use of statistical data drawn from corpus, energy must also be directed to analysis of language data. It is also necessary to extract heuristics from the analytic findings and generalize them. Unlike the top-down approach of conventional linguistics, corpus is believed vital in the bottom-up approach to language research.

4. Conclusion

It must be noted that technology developed in the Machine Translation Project for Asian Languages can be applied to systems other than machine translation. These include high-precision word processing, dialogue systems, text-proofreading support, data retrieval, and text summarization systems. Nations worldwide must work toward the development of applications of natural language processing in systems other than machine translation.

As computer technology advances to become massively parallel and fault-tolerant, a large-scale knowledge base and knowledge system will be built. There will be information superhighways built to link such systems. Subsequently, an intelligent system that integrates them will be developed. At that stage, the interface between the system and human will play an important role. Then sophisticated software will become more important than hardware. This writer believes that the problem of building an information superhighway can be resolved to some extent with time and money. Whether the multimedia industry founded on this superhighway project will grow is determined by software: specifically, the development of software on a human-machine interface and excellent software for natural language processing. Machine translation technology will be vital in this respect.

In the APEC Conference held in Jakarta recently, the importance was emphasized of a multimedia information communication network linking the United States with the Asia/Pacific region. The establishment of such an infrastructure will undoubtedly accelerate economic globalization. In such an environment, smooth communication between countries speaking different languages will become all the more important. An information superhighway linking nations will increase the importance of machine translation.

Finally, the author would like to present his personal wish that the dictionaries and corpuses developed under the project for each nation will be announced publicly for academic use as soon as possible because he believes they should not be buried unseen.