[From: IFTT'89, April 26-28, Oiso Prince Hotel, Japan]

Towards Greater Sophistication of the Machine Translation System

Hozumi Tanaka
Tokyo Institute of Technology

1. Current Status of Machine Translation

With help from humans during the pre-editing and post-editing, the current machine translation system is becoming practical when used for translating documents with a relatively clear style. This includes documents, such as, scientific and technological papers, and newspaper articles. Current machine translation system is not used for translating literature, poems, and artistic works. The very fact that support from humans such as post-editing is needed shows that machine translation is not a perfect technology. By controlling the range of the documents to be translated and by bypassing the most difficult problems in machine translation by getting help from the humans, recent research has shown that machine translation systems can be put to practical use.

Recently, the users of machine translation systems have realized that it is more important to learn how to utilize the current machine translation systems than demand perfect translations. Indeed, it is a little too early to demand perfect machine translation given the current state of machine translation technology. Even if we can only hope for incomplete translations, machine translation is beginning to be put to practical use in controlled domains. It is insufficient, however, because we have to limit the range of translation to a narrow area. Needless to say, some clients for translation services want a perfect translation. For these people, it cannot be helped that there are some problems in using the machine translation system.

2. The Tasks of Machine Translation Technology

By controlling the range of documents to be translated by the machine translation system and by reconsidering the machine translation system as a human-machine system, we can realize a machine translation system that can be put to practical use. But, this is a path to the improvement of the capability of machine translation systems and not a path to a complete machine translation system. Natural language processing technology is the core of machine translation. To drastically improve upon the capabilities of the current machine translation systems, we need qualitative improvements in the natural language processing technology. Let me expand upon this problem, in search of the future direction in research in this area.

(1) Natural language processing technology

In natural language processing technology different levels are considered: morphological analysis, syntactic analysis, semantic analysis, context analysis, etc. It was believed that the analyses should be done in this order and many machine translation systems were designed on this basis.

It was assumed that the difficulties of analysis increased when we analyze language in this order. As far as the English language is concerned, the morphological analysis can be done without any problems in a comparative sense. In cases of agglutinative languages, however, such as, Japanese, Chinese, and Thai, the morphological analysis is difficult. When we go deep to the roots of these languages, we find that in the morphological analysis of the agglutinative languages, we need results from syntactic, semantic, and contextual analysis. For instance, in Chinese, we need results from semantic analysis. In Thai, there are no signs to indicate the end of the sentence. Therefore, before we extract morphemes elements, we need to know where a sentence starts and ends. Therefore, we need to have results from contextual analysis. Whatever the case is, we need a computational model for performing all these analyses at the same time. One of the most promising models in this area is a language analysis mode using the augmented context-free grammar.

Fortunately, in the syntactic analysis, a high-speed algorithm has been developed. There is the Tomita method (Tomita 85), which is based on LR(k) method. The SAX, which has been developed by Matsumoto (Matsumoto 86), has bee realized as a logical program, which achieved results almost equivalent to the Bottom-up Chart Method (Kay 80). Tomita method and SAX incorporate a breadth-first search strategy. I think the most useful way is to integrate morphological, semantic, and contextual analysis into excellent syntactic analysis computational models mentioned above.

There are many problems with semantic analysis. The major research task of the natural language processing system in the field of artificial intelligence has been semantic analysis. How to express the semantic structure related to the problem of meaning continues to be a very hot research theme. The research in the area of artificial intelligence put a focus on the importance of knowledge for semantic analysis. Since this problem is related to the dictionary, I will take it up in (3).

As we read through sentences, the ambiguity is gradually clarified and the meaning becomes clear. We can call this the process of incremental disambiguation, and we need a computational model in this area (Mellish 85). This relates to the problem of expression for the tokens with ambiguity (Sowa 84; Barwise 83; Mukai 85). The elimination of ambiguity is related to the problem of the selection of translation words, and this is definitely important for machine translation. As a framework for computation, we might be able to use the idea of constraint programming.

For contextual analysis, we need research in the area of anaphoric analysis. One way to do this would be the "focus model", but we need a refinement of the "focus model" itself. (Sidner 83; Carter 87). The recent theories on context, such as situation semantics and discourse representation theory (DRT), do not give direct answers to this problem.

Sentence generation has less urgency than the analysis. In analysis, we need to eliminate ambiguity on several levels. Sentence generation starts when the ambiguity has been dissolved. Because, however, sentence generation is directly connected to the qualitative evaluation of machine translation systems, we need to do research in the area when style is considered.

3) Dictionaries

The problem of knowledge is also the problem of dictionaries. What has been made clear by the research on machine translation systems so far is that the quality of the dictionary exerts a great effect on the results of machine translation. The Electronic Dictionary Project in Japan has been started with this idea in mind. In artificial intelligence research, people have been interested in the problem of the quality of knowledge rather than the quantity of knowledge. A machine translation system must have a huge amount of knowledge in the dictionary. Therefore, we must solve the problems of quantity and quality at the same time.

The above problem is whether to use manpower for this work or to introduce methods of learning and obtaining knowledge. At the research level, however, learning and knowledge acquisition is very poor. Therefore, we need to depend on manpower as in the EDR and CYS projects. We also know, however, that there are limits to these methods. We need to proceed with research for converting the knowledge in the existing dictionaries and encyclopedias into forms that can be computed. There is a possibility for a new learning theory in this area.

Dictionaries cannot be made overnight. I think we need to involve certain organizations, for the developing and maintaining of dictionaries from an international point of view.

3) Linguistic theories

Some of the recent grammar theories, such as, GPSG, LFG, HPSG, etc., consider plausibilities as computational models. The paradigm of these linguistic theories from a computational model point of view is unification, constraint, and the quantity of information to be included in the dictionary items. (Bresnan 82; Gazdar 86; Sells 86; Shieber 86). In HPSG, the quantity and form of information are considered most important, and the information included in the dictionary items will be increased. Only a few grammar rules are included. Reduction in the number of grammar rules may be effective for the free word-order languages where word order can be flexible, such as, in the cases of Japanese and Thai. The reduction of grammar rules may render the syntactic analysis algorithm unnecessary. We must note, however, that the amount of information processing concerning the items in the dictionary is increased and the amount of computation necessary will increase. There is a trade-off between the number of grammar rules and amount of computation necessary for information processing of the dictionary items. Therefore, the grammar rules must be developed with this point in mind.

There is little contribution from the linguists regarding the problem of semantics. It might be a good idea to have the linguists there when a machine translation system is constructed. This will certainly lead to new discoveries.

4) Computational model

When we consider performing complete language analysis the sequential computation model currently used is not sufficient. With the progress of VLSI technology in the future, we will be able to have a very small, cheap, and powerful computer. It is about time that we used parallel processing in language analysis. Fortunately, the SAX and Tomita method use such a syntactic algorithm based on breadth-first strategy. Both these have good compatibility with the parallel processing model.

A dictionary with a large amount of knowledge is a huge knowledge base. We should be thinking about parallel processing in conjunction with the search of dictionaries. In the fifth-generation computer project promoted in

Japan, parallel processing is a major task to be tackled. In this sense, it would be meaningful to think of a machine translation system using the fifth-generation computer.

The machine translation systems currently used are often based on the C language. As machine translation systems become huge and complex, the computer language used will also become a problem. I think we need a more sophisticated language than C. In this direction, there will be contributions from artificial intelligence and software engineering.