MT SUMIT
Sep. 1987, at Hakone

Technical Prospects of Machine Translation

Hirosato Nomura

NTT Basic Research Laboratories
Musashino-shi, Tokyo 180, Japan

During the last three decades, many challenges on machine translation have been conducted around the world, and we have successfully reached the first step of practical machine translation systems. I would like to stress in this panel discussion the first step of practical machine translation technology, the next step, and the steps further in the future.

The theoretical and methodological bases of machine translation relate to computational linguistics theories and computer technologies for natural language processing. Application of these theories and methodologies involves many issues such as dictionaries, term-banks, grammars, analysis of source sentences, transfer of intermediate representations, generation of target sentences, computer environments for developing and examining machine translation systems, operational environments, pre-editing of source sentences, post-editing of raw machine-translated sentences, controlled languages, and so on.

We have accumulated a large amount of experience concerning these fundamental problems during last ten years. In the United States, there is much experience with commercial machine translation systems such as Systran, Weidner, Logos, Alps, and Smart. The Canadian government developed Taum Meteo, which has been used to translate sentences concerning weather information for a long time. The Commission of the European Communities has long-term experience on Systran. Grenoble University is engaged in development of a machine translation system called Ariane, and Saarbruecken University is engaged in developing a machine translation system called Susy.

There are some recent ambitious national level projects for developing machine translation systems such as the Eurotra project of the Commission of the European Communities, the French governmental project, the United Kingdam Alvey project, German projects including Stuttgart and Saarbruecken, and the Japanese Mu project. Also, New Mexico State University and Carnegie Mellon University have started machine translation projects.

Japan has also recently started some new government level projects related to machine translation. They involve a dictionary project by EDR (Elecronic Dictionary Research Institute), a telephony interpretation project by ATR (Advanced Technology Research Institute), and a machine translation system development project by CICC (Center of the International Cooperation for

Computerization of Japan). The CICC project concerns translation between Japanese and languages of neighboring countries including China, Indonesia, Malaysia, and Thailand. Also, there are other activities in machine translation in other countries.

I believe that we can say that we have made a reasonable amount of progress in designing or developing practical machine translation systems through the work just mentioned. It has already been proved that machine translation is feasible and can be used effectively in some restricted areas and for some restricted purposes. Thus, we do not need a second version of the ALPAC report; instead we need a new report concerning how to use machine translation effectively assuming the current state of technology. The report should also relate to what technologies we have established and what technologies we have to develop in the future.

Although we have much experience in machine translation, we might not have a sufficient amount of evaluation or comparison of methodologies with respect to the problems I mentioned earlier.

We might now be in the stage of examining what level of translation could be accomplished by extending the size of dictionaries and grammars. Fundamental information has already been described in dictionaries and grammars, and thus most effort might have to be put into incorporating extra idiomatic information and domain-specific information. Also much effort should be put on how to produce natural or acceptable sentence expressions. The most successful use of current machine translation systems might be achieved through insight into what the machine translation systems can do and what they can not do. Such affirmative or sympathetic cooperation will lead machine translation technologies to higher and wider possibilities in the practical use of machine translation systems.

We must suppose some technological levels of machine translation systems which will be accomplished step by step. The situation is the same as for other artificial equipment available, since we cannot expect to achieve the perfect system we have in mind. Therefore, we have to make effort to find the best way to apply the imperfect system which has been designed by applying less than fully developed technologies.

We no doubt will remain for a while in the first step, where we have been attempting feasibility studies of practical machine translation systems. I mean by feasibility studies the process in which we have learned the real problems through the trial and error procedures of our experience. We are now in a stage in the first step in which we have to evaluate it so that we can summarize which are the well-established technologies and thus so clarify what problems remaining. This will give us some concrete perspectives for the succeeding steps.

The dictionary is one of the most important components in a machine translation system. Basic level dictionaries includes information concerning morphology and word-level syntax. On the other hand, advanced dictionaries might have to contain a wider variety of information such as semantics or descriptions of the concept each word relates to, and so on. What kind of information lexical items should include depends on the linguistic basis adopted in the design of the natural language processing components of a machine translation system. Lexical items should also be related to a thesaurus structure which reflects hierarchies of words in terms of semantic relationships such as upper-lower, whole-part, etc.

The term-bank is another kind of dictionary that concerns domain-specific words. How many terms should be incorporated into a term-bank? What is the best strategy for developing and maintaining such a huge data-base? Is a term-bank only a collection of domain-specific words accompanying their translation equivalents? It is clear that a term represents a domain-specific concept or defines a domain-specific fragment of knowledge. Need we not utilize the meaning of the term to produce high-quality translations?

How about the computational linguistics theories and their applications to designing the machine translation systems? For a long time, the generative phrase-structure grammar theory seemed to be the main stream in the computational linguistics area. However, it also seems that most machine translation systems did not adopt it directly as a basis for their designs, although some syntax-oriented systems adopted it with modification. Is it sufficient for developing a really practical and useful machine translation system? I guess that most people will disagree with the matter. If not, what shall be the theoretical basis of machine translation? How about the recent activities concerning computational linguistics theories or new frameworks for representing and processing linguistic information? They include Functional Unification Grammar, Lexical Functional Grammar, Definite Clause Grammar, Generalized Phrase Structure Grammar, PATR-II, Categorial Unification Grammar, and so on, and possibly also Government and Binding theory. Can they contribute to the design of more powerful machine translation systems in the near future? If so, what must we do? Some or most of them seem to be approaching the so called universal grammar, which means they may provide a language independent framework for representing and processing linguistic information. Can such a universal grammar resolve the problems coming from differences of language expressions which might be based on culture differences established through a long history? There is also a problem concerning application of the Situation Semantics theory for interpreting the meaning of sentences uttered in a given environment or situation.

Strategies of source sentence analysis are also closely related to the linguistics theories we adopt. The method for syntactic analysis seems to have been well established. However, how can we incorporate semantic, context and

discourse analyses into syntactic analysis? It is true that most current machine translation systems are applying not only syntactic analysis but also semantic analysis to some extent. Do we need more different or effective approaches for these purposes? In Japanese sentence analysis, for example, case analysis or similar strategies have been adopted in developing machine translation systems in Japan. These strategies motivate semantic-oriented approaches in designing machine translation systems between Japanese and Indo-European languages. It might be a reason why the commercial machine translation systems in Japan came into the market so early or after such a short period of development. Is it a specific problem for Japanese processing?

How about the two well-known approaches, transfer and pivot? The merits and demerits of both have been discussed for a long time. However, it might be true that semantic approaches will reveal the benefit of the pivotal approach for the simple construction of practical machine translation systems in a short period. Transfer of a pure syntactic structure from one language into another language is not a good choice of strategy in translation between langauges having very different syntactic structures such as Japanese and English. Nevertheless, this fact does not mean to push out the transfer approaches in machine traslation since there might be possibilities to realize so-called syntax-preserving translation.

Generation is another important issue in machine translation. Without producing acceptable translations, machine translation systems are of no use at all. This relates to the problem of what to say and how to say it. The second problem, namely how to express an idea, will be resolved by providing enough information for sentence represention while some problems will remain in utilizing contextual information. What do we have to do concerning the first problem, namely what to say? There is no reason why each source sentence has to be translated into a corresponding isolated sentence in the target language. Rather, a series of source sentences should be rearranged into a different series of target sentences, since the way of expression or segmentation of a paragraph into sentences is crucially different in differentf languages.

Since the development of dictionaries and term-banks will continue for a long time and their revision will be rather frequent, we need to have a good computer environment. We will also need the same kind of environment for extending and maintaining the grammars and translation systems.

From the point of view of user-friendly computer systems, a good man-machine interface is needed. This is also needed for pre-editing of source sentences and post-editing of machine-translated sentences.

What is a controlled language? There are some difficulties in developing a fully automatic translation system which needs no pre-editing or post-editing. Controlled language or sub-language provides an answer by which we can develop a really usable machine translation system in some restricted areas and

for some restricted purposes. However, the regulation put on controlled language will gradually become looser with advances in machine translation. Thus, we must be careful in designing the controlled language.

From the practical point of view, there is a problem of which strategy we have to adopt for successful use of current and near-future machine translation systems. There are many possibilities for adapting current machine translation systems to daily-life translation. Wise use of machine translation will reduce cost; however, bad use of machine translation will waste both time and money. What is the most efficient strategy for using current machine translation systems?

I have touched on many problems concerning machine translation, from which, I hope, we can establish some perspectives for future machine translation technology through careful study. Since the range of problems is so wide, varying from the theoretical aspect of computational linguistics to the technical aspects of writing programs, I do not expect each panelist to give full answers to all problems. Rather I expect that each panelist will suggest more concrete problems and insights based on his experiences in machine translation, from which we can imagine the real perspectives.