[From: Makoto Nagao (ed.) *Machine translation summit*. Tokyo: Ohmsha, 1989]

7 • 4 Concluding Remarks 3

Makoto Nagao
Kyoto University, Japan

There were several discussions about the development of machine translation, but no one referred to the cost of the system's construction. There were some talks about the cost of operation, but here I would like to stress the very heavy cost for the construction of a system for the newcomers to this field. Huge amount of money is necessary, and many years, many people. No one presented the numerical values for these factors. I dare to try it. For example, I believe that at least six or seven competent grammar writers, six or seven competent software technologists, and many people for the dictionary construction are necessary, all of whom are to be trained and well organized to the purpose. At least three or four years are necessary for the construction of the basic framework of a machine translation system. After the development period there will be a very long, hard time for the improvement of the system by the interaction with users. This is a terrible task to do. I believe many manufacturers of machine translation systems here calculated their development costs. They cannot disclose the cost, of course, but they will agree to my view, or they may emphasize the cost much more. But anyway, this is a very important factor for the newcomers to this field to consider seriously.

Next one is how to develop a machine translation system from scratch. You have to be very careful about the model design. If the model of machine translation is poor, then you cannot improve the system at a certain advanced stage. You come to a certain barrier, or a deadlock, which you cannot break through, no matter how much money, time and people you invest. I mean the quality of translation cannot be improved by the continuous efforts by the limitation of the translation model. So you have to be very careful about system design, linguistic and software models of machine translation. The system must be open-ended, I believe. This is quite hard for the system's design, but this is an important factor. The system should accept new linguistic components and new processing mechanisms, when they appear, because the development of machine translation system requires a long time, for example five years or seven years. During this development period there will be an advancement in linguistic theories, some parts of which can be brought into practical machine translation system. For example, we can introduce some good grammatical rules concerning modalities and text coherence, which were not included in the system's framework at the starting point. During the development you may encounter such situations as to introduce such new linguistic factors into your systems, which must be done easily without changing drastically the other parts which have been constructed already. Another important thing which we have to recognize is that, although a core part of the language is regulated by solid grammatical rules, or basic linguistic theories, there are many exceptional linguistic phenomena. There are a huge amount of special expressions which should be translated into some specific expressions in other languages without any theoretical reasons. You have to be able to handle these irregularities in your model, either by grammatical rules or by lexicon. You have to have such mechanisms in your system. Another factor is to keep the consistency in a variety of meanings in a machine translation system, which is really a huge system constructed by very many people, and which discloses inconsistencies very often. A single person cannot construct a system. Many people must cooperate with each other in the construction of grammar, dictionaries and software. So you have to construct the grammatical part, the dictionary part or the software part in a proper way by independent developers and you have to be able to combine these results into a beautiful

total. This is a kind of management problem. This should be done very carefully, indeed, and of course you have to go gradually from the core part to the peripheral parts. In the grammar constructions, we recommend the sub-grammar network idea. And even in that case you have to write core grammatical rules first. Then you write additional rules, and finally you write the domain specific rules.

In the software, there are the core translation program, peripheral programs and varieties of supporting programs. For example, there are many important supporting programs, such as dictionary correction, error location in the translation process, and text data handling for reference in the dictionary construction and so on.

Next point is that, due to many participating people in the construction of the system, we have to prepare a working manual for the people to follow in the construction of a system. Typically in the dictionary construction we have to keep 20 or 30 people at the same time, and have to keep the same quality of the dictionary contents. For that you have to carefully prepare a working manual which clarifies what kind of tasks they have to do and in which way. You have to prepare not only working manuals, but also multilingual texts, various kinds of ordinary dictionaries and some grammar books and so on for their reference during the dictionary construction. How to consult these reference materials should also be specified in a manual to keep the quality of the dictionary as equal as possible by different persons. Of course in the construction of grammatical rules, the same preparation should be done.

Now let me say something about the future direction of machine translation. Of course, we have to aim at AI oriented system. But this is just for research curiosity, not for a practical system at this moment. The translators' know-how is important, but it is quite difficult to incorporate their knowledge into practical machine translation system. Expert system of the translators' knowledge must be wonderful probably, but the construction of this expert system and the combination to a machine translation system are very hard to achieve. We have to have a mechanism of best first search for disambiguation of sentential analysis, and also a mechanism of parallel passing. These are current research topics in artificial intelligence. Knowledge use in machine translation is another important research topic in the future. I would like to classify the knowledge into two. One is linguistic knowledge and the other is world knowledge. We have to reflect on ourselves about whether we are utilizing 100 percent of the linguistic knowledge in machine translation. This is a very difficult question, but I don't think that we are. I feel that 30 to 40 percent of the linguistic knowledge is currently being utilized. We must get more linguistic knowledge in the form that can be utilized by machine before relying on world knowledge. This is a healthy approach. We have to examine much more in detail about the sentential structures and meaning structures. Also we have to study much more the relationship between sentences, that is, contextual relations. In this way the examination of the relationships must be expanded from adjacent words and phrases, to sentences and paragraphs. World knowledge is very popular in artificial intelligence, but it is quite difficult to prepare the world knowledge in every field in a very precise way for use in machine translation.

Next I would like to talk about a few interesting activities in Japan. There are some good researches for parallel parsing algorithm. I don't know how many years are necessary for the algorithm to be practically usable, but it looks quite soon. We must push the activity forward until it is realized. It is a very important topic. It must be a very powerful algorithm that can give us not only a great speed, but also the facilities to specify semantic constraints and preference factors. Another element is the automatic tuning of a translation system to different text areas.

Next point is to study and clarify what information is necessary as the content of machine translation dictionary and how it is obtained from where. Also important is what kind of grammatical rules are necessary and effective. These have never been discussed so far. Grammatical rules have always been written in personal subjective judgments, and have never been disclosed in detail. I believe that in the future objective representation and explanation are necessary for each

grammatical rule as for its effectiveness to the analysis of sentences.

Next one is the design of controlled language, that is sentential styles which a machine can process in reliable way. For the design and construction of a system in general, there must be a specification of input and output. This is an essential part in engineering. In machine translation, which is an engineering problem, we must have an well-formed specification for input sentences for the design of a system. We have to begin from how we can handle this design. We have to decide the domain of texts. Think about a car, for example. You will not imagine a car that would climb Mt. Fuji, or an automobile that can travel rice field safely. In the design of ordinary automobiles there is no consideration about rice field. In this way the engineering design always has a domain or a range of applicability. In the case of machine translation, what kind of sentential styles, or what kind of expressions are within the scope of the design, or outside, must be decided. This must be clearly delineated. Once this is delineated, an acceptable machine translation system can be designed, and it will function as it is designed. This is a basic idea for the control of input sentences to machine translation system. If this controlled language is appropriately designed, the system will become economical and labor saving. However, it is quite difficult to clarify the specifications of a controlled language. I don't think that can be achieved in a year or two. Of course, there are quite a few people who are opposed to this idea. They think that natural languages are intrinsically natural so that they cannot be put in the framework of a controlled language. This may be true. Expressions in anger or in arguments contain varieties of implicit meanings, which are difficult to convey to another language. Translation of such expressions are difficult even to a professional human translator. Here as an engineering problem, we want to have an efficient machine translation system for normal documents. What is required for us is an efficient system that attains a reasonable quality level. We must make careful decisions to achieve this goal.

Preparation of reference text materials are very important in addition to the ordinary dictionaries for the construction of machine translation dictionaries. We must prepare multilingual texts and must do quick sorting for easy reference. There are many things that are necessary for the construction of a system. There is a need for a software to extract unknown words in a text material and enter them into a dictionary. Construction of terminology data banks is also important, especially for the developing countries where there are limited specialized terminologies in their own languages. We have to collect and expand systematically the range of specialized terminologies.

Finally I want to say a little bit about the ALPAC report. There still exists the ghost of 1965 ALPAC in the world of machine translation. We have to beat down it by showing various advancements in machine translation systems at present. What kind of advancements can we see now? We have several new parsing technologies that have appeared since 1965. We have many labor saving devices for dictionary construction. We have more advanced linguistic theories, and comparative study of tense, aspect and modality between English and Japanese, for example. More effective use of semantic markers in the analysis stage, more accurate selection mechanism of words in the transfer stage and sophisticated mechanism in the generation stage. Even more difficult intersentenial processing is being tried to a certain degree. Structural transformations to bridge the gap between different language families such as Japanese and Indoeuropean languages have been developed to get more natural translation. Lots of such mechanisms are incorporated in Japanese - English machine translation systems nowadays. In this way various components of a machine translation system are becoming increasingly clearer. Real advancement was made in this area since ALPAC. Let me give you a different example. Think about FORTRAN compiler. The internal structure of FORTRAN compiler was not clearly presented and not so much understood in the 1960's and the construction of a compiler took a long time. Nowadays however, through progressive advancement the time for FORTRAN compiler construction has substantially been reduced. Similarly, in 1965 the internal structure of machine translation system was not clarified enough, and was extremely difficult to construct. Today many linguistic elements and software components of a machine translation system have been made

clearer and many newcomers can construct a machine translation system in an extremely short period of time. Of course there are still very many areas that need further linguistic clarification and more progress in software. But the situations are continuously improving, and I would say that many of the linguistic elements recently developed will be incorporated into practical machine translation systems shortly. There are a few practical systems which give better cost-effectiveness to human translation. In this way there has been substantial progress since the ALPAC report. We cannot wait the construction of a machine translation system until the completion of the theory of language. We can construct a good engineering product without knowing every precise detail of the science for the product.

I want to make a short comment about the topic of pivot and transfer. People are always curious about whether pivot method is good or bad. I think the problem is best illustrated by Dr. Tsujii's paper in this conference. He pointed out that it depends upon where to use a system. Translation of technical papers, essays, business letters, and dialogues requires quite different technologies. For example, the transfer method may be better for technical papers. If dialogues which are fragmental, incomplete, and ungrammatical sometimes, are to be translated then the pivot approach will be quite right. It will be also applicable if you can restrict your task domain into a very narrow area, where you can describe the knowledge structure of the field in great detail. If you have to treat a very wide area with a large volume of texts, then the transfer method is more appropriate. I think this will continue to be true for the foreseeable future.