[MT Summit IV, July 20-22, 1993, Kobe, Japan]

Machine Translation: What Have We To Do?

Makoto Nagao
President of IAMT

1. State of the Art of MT Users

In AAMT there are two investigation committees, one the study group on efficient use of MT systems, and the other the study group on controlled language for MT. There is another similar group in Japan Electronic Industries Development Association (JEIDA), which investigates the proper use of MT systems including the establishment of MT evaluation methodology. The foliowings are mainly based on the reports of these investigation group.

(l) MT users who translate large volume of documents are, first of all, professional translation companies, and then manual production sections and translation sections of manufacturing companies. They translate several hundreds to one thousand pages per month by one MT terminal. They make efforts to improve the efficiency and the quality of translation by limiting the translation documents to very specific topics.

The average time for pre-editing in such cases are 5-15 minutes for a page of A4 size, and the average time for post-editing is much longer, about 10-30 minutes. Major portions of pre-editings for Japanese texts are to cut and change a long sentence into several shorter ones and the recovery of omissions such as subject and object in a sentence. Major actions in post-editings are to check and correct translation words and sometimes to rewrite a whole sentence for un-understandable or no translations.

There are varieties of pre/post-editing operations and it is somewhat difficult to establish a kind of standard pre-editing operations common to different MT systems. Each MT system has its own authoring and pre-editing softwares to help both people and an MT system. In experienced customers improvements in system use are usually done by establishing the editing know-hows of their own and user dictionary enrichment. Very few users have customized their own user interface themselves.

(2) Kind of works on MT systems can be typically categorized as pre-editing, post-editing/rewriting, system operation, and dictionary management. In a small user companies all these works are done by a person. In large MT users these are performed by different specialists to increase the overall productivity. Training of people on these MT operations is generally done in a translation company itself. People engaged in these jobs are not necessarily good at English. Some have fairly high level of English, but many others are rather poor. They have almost no knowledge in a specific field of document to be translated. So some miss-translations may be found in ambiguous expressions and in technically very complex descriptions. But when a system is provided with a good terminology dictionary of a document field, the translated materials have a certain quality level. This is a profitable point of using an MT system.

(3) Major complaints and comments for MT systems are about the quality of translation. In these more complaints are on J-E MT systems than E-J systems, probably because the former is more difficult to construct. There are many other complaints from long experienced users, such as bad user interface, and incompatibility of an MT system with other systems such as word processors, formatters and printing facilities. These demands come from the customer's wishes to improve the overall throughput cost and time. All of these indicate that there are many peripheral problems to solve and improve besides the MT quality.

(4) Many MT users are developing their own dictionaries in varieties of specific areas such as information, telecommunication, electronics, mechanics and aeronautics. The size of these dictionaries are generally 2000-6000 words for both E-J and J-E systems. These dictionaries have such information as parts of speech and simple semantic markers, but usually they don't include definitions and example phrases. Some advanced user dictionaries have synonyms and other information to distinguish different meanings of a word.

The priority sequence of translation word selection is usually based on the frequency of word usage. But it is rather difficult to obtain a reliable frequency data because the frequency depends on a particular field and a particular document. Therefore proper selection of translation words must be done by human in the post-editing phase.

(5) In big translation companies where several MT systems are introduced, dictionaries are commonly used via network. This is to be extended to outside networks. There are many demands for softwares for good interfaces for adding new words by customers, for finding unknown words, typical specific expressions in a document which need specific translation, and for keeping the history of dictionary change. Some customers have intentions to open their user dictionaries to other users by mutual exchange basis or by proper prices. Therefore it is quite important to establish an exchange format and an exchange market of user dictionaries.

(6) MT users are doing varieties of individual efforts to reduce translation cost. One of the typical examples is to utilize a world-wide computer network in such a way that English texts translated from Japanese by MT are sent to US for post-editing by native speakers of English, or that pre-editing of English texts for E-J MT is done in Singapore by English-speaking people and sent back to an MT system in Japan for translation into Japanese.

(7) Documents in electronic form are steadily increasing but there are still many printed documents for translation. OCR is nowadays used cost-effectively as an input for an MT system particularly in big translation companies. Error rate of OCR for Japanese characters is less than 2% for ordinary printed pages.

(8) Some MT users estimate that more than ten thousand pages are to be translated by machine per year to produce profit, but some others estimate a few thousand pages to make a balance. The estimation of cost effectiveness of using MT systems is very difficult because it depends on many factors such as document sources, quality of original documents, quality of pre/post editors and so on even if the system is the same.

(9) Some on-line networks in Japan provide MT services, but they just give back raw translations in a few hours depending on the translation quantity. The user, if he/she wishes, can forward the

result to a post-editing company on the same computer network and can get back the post-edited results next day. A company running on-line service says that it makes profit because it has no person for pre/post-editing and other MT operations. The cost is just for the maintenance of a computer and softwares. Customers may not be satisfied with translation results, but they generally think that the poor quality is due to their poor input texts and do not ask the service company for the improvement. The users are increasing by 50% annually and more than 90% of them are small companies. The translation cost is about 800yen per page, which is about 1/3 of the ordinary MT translation cost.

(10) For a successful use of a present-day MT system we can point out the following factors.

(a) Long experiences of MT use and rich knowledge about the goodness and badness of an MT system are important.

(b) Effective management of people engaged in MT production and MT processes is to be established for a customer's specific use characteristics of an MT system.

(c) Effective use of an MT system must be designed within its capacity.

(d) Specialists of MT are to be educated and kept in customer company.

(e) Customization of an MT system is important, particularly in user dictionary and special post-editing functions.

(f) Effective use of pre/post editing facilities must be considered by customers.

(g) MT systems and input/output devices such as word processor, on-line network connections, softwares for insertion of figures and tables, printing formatter of final output texts etc. are to be linked together effectively so that the total throughput will become very efficient.

(11) Nowadays there are increasing number of individuals who use MT systems through on-line service or by installing an MT software on their personal computers. Some of the personal users of MT are very positive to MT systems and make lots of individual efforts for the improvements of MT systems and dictionary contents. They do not translate a big volume of documents but just translate their personal letters or documents which they are interested in reading. Very often their document translation is just for quick grasp of the contents and therefore no post-editing is performed in such cases. This type of MT use by individual will become very common at the time when MT systems are widely accepted in the society. We have to investigate more exactly and thoroughly the behavior of the users in this category for the future widespread circulation of MT systems.

2. Future Steps

2.1 MT Users

(1) Terminological words in specific areas are to be collected in a large scale. They are enormous and the task of collection for MT use is very expensive. The best way is to ask academic

societies of these specific areas to cooperate with us in the collection of the terminological words in electronic form. These terminological dictionaries are hopefully to be open to the public.

(2) Clearing house function about the information of individual user dictionaries is very important for mutual exchange of special terminology dictionaries. Some guidelines must be established to realize easy exchange of user dictionaries, such as exchange format of dictionary contents.

(3) Experiences of MT system use must be exchanged among users. Particularly the experiences which realized better performances of a system use are valuable to other interested people. Site visit to such users will be valuable.

(4) Users are always interested in the evaluations of translation quality, user interface conveniences and cost-effectiveness of a system. We have to establish guidelines for the evaluation of these factors. They are particularly important to those who are going to introduce a new MT system.

(5) Accumulation of typical letter samples with translation in varieties of situations will be valuable for future individual users who want to compose foreign language letters on their personal computers. The same is true for companies which have to write lots of business letters and reports of fixed types everyday. Exchange of these materials is also recommended.

These are just few examples which IAMT or each regional association for MT can take initiative for the users of MT systems.

2.2 MT Developers

(1) Descriptions of MT systems in a PR pamphlet and on yellow pages of MT are not equal in their description items and levels. They are not comparable each other, so that users or potential users cannot judge whether they will fit their purposes. MT developers must make efforts to clarify what kind of information is to be given to general public. This is particularly important when there appear lots of MT softwares in the market and when people cannot use and test all of them.

(2) A standard evaluation method of MT must be established not only from the standpoint of MT users but also from the standpoint of MT developers. Every commercial product is tested from varieties of points before it is put into the market. These tests are usually objective and very severe. MT systems must be tested by the same idea. JEIDA MT group published a set of sample sentences which will reveal what kind of grammatical abilities an MT system has or has not.

(3) Every industrial product has its own specification or condition of usage. MT systems should have the same specification. This means that they have their explicit specifications about what kind of sentences are acceptable and what are not. This leads to the design of a controlled language or the clarification of what kind of pre-editing must be performed. In AAMT there is a group to discuss about controlled language. Fruitful results are to be expected.

(4) Another important problem related to pre-editing or controlled language is the design and in-

troduction of SGML into the documents to be translated by machine. Nowadays documents are transferred by a network all over the world, and they have chances to be edited and translated by different systems. If we can define a common SGML for MT document and if it is adopted by both document production systems and MT systems, the quality of MT will become significantly improved, the speed of MT and pre/post editing will become very high, and we will be able to achieve remarkable total efficiency. It is of course tedious, time-consuming and impractical to put on full SGML markers to a document. We have to develop a simplified marking method, good man-machine interface program to do marking semi-automatically, and a conversion program from these simplified markings to a standard SGML for MT.

(5) There are many other things to do, such as, (i) cooperative development of MT dictionaries, particularly phrasal dictionaries, (ii) an integrated system of machine translation and word processing, particularly functions of handling tables and figures, (iii) networking of MT systems for common use of specialized dictionaries for multi-lingual translation, for pre/post editing in different places, and so on, (iv) development of a program for a cooperative translation work among several MT users (CSCW: computer supported cooperative work), (v) easy customization/learning function to different document fields, (vi) extension of MT systems to multi-languages, and so on.

2.3 MT Researchers

(1) Some new MT methods are being developed such as probabilistic MT and example-based MT. These must be studied more intensively and more serious considerations must be given for the development of practical systems in these principles.

(2) Multi-lingual MT system was studied in Europe, and is being developed by CICC among five Asian languages. This is a very important topic in MT research and development activities, and must be intensified by international cooperation.

(3) Basic researches in natural language processing must be continued particularly in such topics as discourse analysis, dialogue analysis, syntactic and semantic disambiguation, constructing good MT dictionaries and knowledge base for use in language understanding, introduction of learning functions in MT, and so on.

2.4 Government Sponsors for R&D in MT

Clarification of natural language mechanism is equally difficult to the clarification of human brain function, and therefore R&D in MT must be continued for a long period. International cooperation at various levels are to be encouraged. Technology transfer of MT must be done to developing countries where scientific and technical information in advanced countries must be translated into their own languages. The idea of establishing an MT Center in each country sponsored by the government will be very valuable for the promotion of technology transfer in such countries. The role of the centers will be to develop MT systems, to translate scientific and technical documents, to become an information center of a country, and to act as a connecting point of domestic and international information flow in the area of science and technology through translation. Cooperative development of MT systems of different languages and multi-lingual terminology dictionaries by the countries where these languages are spoken must be coordinated by the leadership of concerned governments.