[From: *Georgetown University Round Table on Languages and Linguistics 1989*, James E.Alatis, ed.]

New directions of machine translation

Makoto Nagao
Kyoto University

1 Introduction. There are growing demands for language translation in every country because the exchange of persons, information and commercial products among different countries is increasing year by year. Especially strong demands exist in Japan for translation between Japanese and English in the business world. Human translation is slow and cannot fill the demand. Moreover, the quality of human translation is not necessarily good.

We have been conducting research in Japanese-English machine translation at Kyoto University since 1965, and developed a sentence-parsing program, dictionary handling systems, and so on during 1965-75. Then we developed a machine translation system (MT), TITRAN, which translated titles of research papers in science and engineering. We constructed TITRAN-EJ (English to Japanese), TITRAN-JE (Japanese to English), and TITRAN-JF (Japanese to French) until the end of the seventies. TITRAN-EJ was used by many researchers and received favorable acceptance. It translated *He is a boy* into *Helium is a boy* in Japanese! The system was so much specialized in the scientific field.

Then in 1982 we started a national MT project, which was supported by the Agency of Science and Technology of the Japanese Government, and which aimed at the translation of abstracts of scientific and technical papers. In this four-year project we developed two MT systems called Mu-JE (Mu is an acronym of our system) and Mu-EJ. The systems were quite large, and included 70,000 words and 1500 grammatical rules in each language. The system is now under reconstruction for the purpose of connecting it to a large database of the Japan Information Center of Science and Technology.

This research and development activity gave great encouragement to computer companies in Japan. We transferred technology to these companies, and they started the construction of commercial MT systems. At present there are about ten commercial MT systems (half JE and half EJ) and several others under development.

The merit of introducing MT systems for the translation of documents is gradually being realized in Japanese companies, and it is widely recognized that the systems will have increasing acceptance in society in two or three years. This is because the users have become wise enough to use a system optimally in a given environment by understanding the possibilities and the limitations of the present-day MT systems, and by not having excessive expectations as to the translation quality.

The research results so far in academic circles have been almost completely transferred to industry, and basic research is moving toward

completely new models, which will give new impetus to higher level MT systems in the next century.

It is quite important to clarify different components related to machine translation in order to evaluate the state-of-the-art of present-day MT systems and to consider future directions. So far, MT systems have treated one sentence at a time for translation without any reference to previous sentences in a text. However, there are varieties of interconnecting information among adjacent sentences, essential for the interpretation of the individual sentence. Several linguistic and computational linguistic research projects are going on in this area, and useful results will be included in MT in the near future. This will become a basis for more advanced MT systems. Several different factors related to MT are pointed out here, and some future perspectives are given.

2 Classification of sentential styles for MT. Present-day MT systems cannot accept arbitrary sentential styles. They do not accept literary works. They have a lot of trouble with precisely translating patent documents, contract documents, and so on. The systems have great difficulty in analysing long sentences of 50 or more words. Thus we have to be very careful about sentence categories and different sentential styles. There are many categorizations of input sentences for MT systems. The followings are some typical classifications:

A Kind of text.

(1) Newspaper headlines, titles of articles. In this category of sentences the most important factor is the speed of information dissemination. The delay will be a few hours at most. Low quality of translation may be permissible. The level of quality at which the readers can understand the meaning from past information and environmental context will be allowed.

(2) Technical information, technical news, economic and social news, and so on. The speed of information transfer is most important. The delay will be a few days at most. Readers of the translated materials in categories (1) and (2) will be professionals in the text area, not the general public. Therefore they are accustomed to the quality of MT and can properly interpret awkward translations.

(3) Operation and maintenance manuals for devices and systems. The volume is great, and translation will require a longer time. The translation quality may not be high, but must keep a certain level, because the translated materials will be printed and distributed to many customers of the devices and systems. Postediting will be required for machine translated sentences.

(4) Business letters which are prototypical. They can be translated by identifying customary expressions. Postediting will be required so that recipients can read the letters readily and easily.

(5) Scientific and technological documents and research papers. First rate translation quality will not be required because specialists in these areas will read and understand the contents.

(6) Conference documents, contract documents, law documents, etc. These are to be translated very carefully. Not only professional translators but also specialists in the document area are needed for postediting.

(7) Patent documents. A huge amount of patent documents has been accumulated and the number is increasing daily. There is a severe shortage of translators in this area, and MT is expected. However, the sentences in patent documents are complicated and difficult to interpret even for human translators. MT will be difficult without heavy professional pre- and postediting.

(8) Articles in newspapers and journals, which are not too literary.

(9) Literary works.

(10) Dialogues in a restricted task domain. These include man-machine, man-man; written sentential conversation by on-line typewriters.

(11) Free dialogue (written sentential conversation).

(12) Speeches such as lectures (one person).

(13) Spoken dialogue (two or more persons).

B Classification of sentential styles (1).

(1) Sentences expressing facts only.

(2) Sentences which include time relations, expectations, assumptions, and conditions.

(3) Sentences which include the speaker's intention, mental state, and so on.

(4) Dialogue sentences which presuppose hearer's knowledge.

(5) Sentences which include metaphors, culture-specific expressions.

C Classification of sentential styles (2).

(1) Length of a sentence (in Japanese, less than 20 characters, 20-40 characters, 40-70 characters, 70 or more).

(2) Number of predicates in a sentence.

(3) Kind and number of conjunctive phrases and clauses.

(4) Sublanguage (use of specific styles in a particular domain) and other complex embedded structure.

(5) Broken and fragmentary sentences.

3 Linguistic components for MT. There are many linguistic theories to be considered in the interpretation of sentences and texts. Some of these theories are already incorporated in MT, but others are still at the research level and cannot be brought into MT systems. Generally speaking, linguistic theories provide us with the basic philosophy of language. Actual language data (sentences) are so complex that the theories cannot explain the reasons for many of them. Linguistic theories are usually based on a very small set of linguistic phenomena, and explain only a small part of language phenomena. MT must treat all the varieties of linguistic expressions, and there must be consistency among the theories taken into an MT system. There are many varieties of linguistic systems which might be considered for MT systems.

D Linguistic components.

(1) Kind of grammar formalism (context-free, annotated context-free grammar, transformational grammar, case grammar, unification grammar, lexical functional grammar, generalized phrase structure grammar, etc.).

(2) Treatment of linguistic semantics (semantic primitives, thesaurus,...).

(3) Dictionary content (part of speech, pronunciation, semantic information, case information, modality, volition,...).

(4) Kind of dictionary words (common word, compound word, terminological word).

(5) Kind of dictionary (monolingual, bilingual, multilingual, concept dictionary).

(6) Knowledge representation and its use in MT.

(7) Ambiguity resolution by contextual information.

(8) Treatment of anaphora, ellipsis, focus, speaker's intention and paraphrase.

4 Aspects of MT systems. MT systems can be classified and considered from many different points of view and aspects. One well-known classification is: fully automatic MT, human-aided MT, and machine-aided human translation. There are many other aspects, some of which are listed here:

E Translation method.

(1) Transfer method.
(2) Pivot method.
(3) Direct method.
(4) Paraphrasing and MT.
(5) Learning function in MT.

F Pair of languages.

(1) From a specific language to another specific one.
(2) Bidirectional between two languages.
(3) Multilingual.

G Human intervention.

(1) Preediting.
(2) Postediting.
(3) Preediting and postediting.
(4) Interactive help at arbitrary stages.
(5) No human intervention.

H Analysis of translation errors.

(1) Ambiguity resolution (one candidate, all possible candidates, several candidates in the sequence of plausibility).

(2) Errors in the analysis phase (conjunctive noun phrase, conjunctive sentential clause, determination of prepositional modifiers, etc.).
(3) Errors in the semantic interpretation.
(4) Errors in the selection of a target language word.
(5) Errors in the generation of a target language sentence.
(6) Errors caused by lack of contextual information.
(7) Vagueness of the original sentence.

5 MT Use and market. Construction of reliable MT systems is difficult at the present stage of computational linguistics technology. Users of MT should have some special skills in order to utilize the systems. They are asked to augment the dictionary contents and sometimes grammar rules to tune the system to their own text translation style. Therefore user companies of MT systems must have professional persons to maintain and improve their MT system. We can categorize user classes in the following way:

I User classes of MT systems.

(1) A few user companies for a system (less than ten). User companies are expected to keep professional persons for their systems, and they must cooperate closely with the manufacturers of their MT systems in order to improve those systems.
(2) A small number of users for a system (less than 100). Users must know which details of their system to improve. Dictionary enhancement and improvement are typical examples. Such users maintain only slight contacts with manufacturers for the improvement of their system.
(3) A reasonable number of users for a system (around 1,000). A system must be compact and reliable enough for users to handle it without any help from the manufacturer of the system. Users must be familiar with the system's weaknesses and must have a variety of skills to avoid them and to insure getting reasonable results.
(4) A large number of users for a system (popularized stage). System must have a learning capability and must be adapted to customer's text domain automatically. A system must be very cheap, and people must be able to use it without any trouble in ordinary homes.

6 Present status and future perspectives. MT systems at present are far from satisfactory as a commercial product. However, there are various possibilities for utilizing them in specialized situations, such as quick stock-market reports, quick survey of newspaper articles, and so forth. The most important thing for users is to recognize the capabilities and incapacities of present-day MT systems, and to have the ability to discover suitable application areas.

One immediate possibility is to find a very restricted small area where stylized sentences and a small vocabulary are used. TAUM-METEO is a good example. Authors must learn to write source sentences carefully so that no postediting is necessary. This type of system can be found in categories A-1, 2, B-1, 2, C-1 (less than 40 characters), D-1, 4, 5. Another application area

is that of translation of operation and maintenance manuals for machinery, electronics and so on. The market in this category is growing rapidly in Japan.

Systems in this category presuppose that the input sentences are short and do not have carefully nuanced implications. Sometimes preediting is required as well as postediting, which is essential. These systems will fall in the categories A-3, 4, B-1, 2, C-1 (less than 40 characters and at most 70 characters), D-1, 2, 3, 4, 5, I-1. There are other applications such as business letter translation for quick reading at receiver side, and technical translation combined with information retrieval and database systems. The Japan Information Center of Science and Technology is now constructing a Japanese-English MT system to be attached to its databases. When it is completed, anybody in the world can send a retrieval request to the Center in English, and get retrieved results in English translation by machine.

Several future directions are conceivable for MT systems. One is to improve the present MT systems by giving more accurate information to grammar rules and dictionaries. Readability of translated materials must be greatly improved. The document categories must be enlarged from technical reports to much wider text areas such as A-4 to A-8.

Another direction is a multilingual MT. The Eurotra project is typical. There is a similar project in Japan, a Japanese Government project for translating technical documents between the Japanese and Chinese, Indonesian, Malaysian, and Thai languages. It began two years ago, will continue another four or five years. More futuristic is a Japanese electronic dictionary project supported by our government, which aims at representing all the word meanings in individual concepts (millions of concepts); via these concepts, the correspondence of word usages in different languages will be achieved. This is called a concept dictionary. The project is now in its third year. Japanese and English are the languages at present. The project will continue another several years.

Dialogue translation is another futuristic research topic. We established a research institute called the Automatic Translation Telephone Research (ATR) Institute three years ago in Kansai Research Park. The aim is to construct a prototype of automatic speech translation between Japanese and English in another twelve years. The institute has a variety of research projects from speech analysis/generation, language analysis/generation, knowledge representation in the task domain of a dialogue, and so on. Linguistic theories such as D-3 to 8 are intensively studied for incorporation into the system. There is a growing number of linguistic research projects for dialogue sentences in the United States and in Japan, and we can expect that this research will succeed.

To improve translation quality we must study more intensively the mechanism for understanding language. Proper translation expressions, including word selection and sentential styles, can be obtained by means of the factors which control the coherence of a text, from the reader's knowledge, and from the factors which control his focus of attention. The old vs. new information distinction influences the text styles of the reader, and so on. Japanese researchers are of the opinion that MT technology, taken in a wide sense, will be a common base-technology for the coming information-based society.

Appendix: Commercial MT systems in Japan.

Manufacture	System name	Language	Sets sold
Fujitsu	ATLAS-1	E-J	170
Fujitsu	ATLAS-2	J-E	95
Sharp	DUET	E-J	
Toshiba	AS-TRANSAC	E-J	
Hitachi	HICATS/EJ	E-J	30
Hitachi	HICATS/JE	J-E	70
Sanyo	SWP-7800	J-E	500
Oki-Osaka Gas	PANSEE	J-E	150
NEC	PIVOT	J-E/E-J	30

From: *Nikkan Kougyou Newspaper* (1989.1.30)

References

Nagao, M., and J. Tsujii. 1976. Analysis of Japanese sentences by using semantic and contextual information. American Journal of Computational Linguistics, microfiche 41.1.

Nagao, M., J. Tsujii, and J. Nakamura. 1985. The Japanese Government project for machine translation. Computational Linguistics 11.2-3.

Nagao, M., J. Tsujii, and J. Nakamura. 1986. Machine translation from Japanese to English. Proceedings of the IEEE 74.7:7.

Nagao, M., J. Tsujii, K. Mitamura, T. Hirakawa, and M. Kume. 1980. A machine translation system from Japanese into English: Another perspective of MT systems. COLING 80.10.

Nagao, M., J. Tsujii, Y. Ueda, and M. Takiyama. 1980. An attempt to computerized dictionary data bases. COLING 80.10.

Nakamura, J., J. Tsujii, and M. Nagao. 1985. Grammar writing system (GRADE) of Mu-machine translation project and its characteristics. Journal of Information Processing 8. 2:10.