Machine Translation Through Language Understanding

Makoto Nagao

Dean, Faculty of Engineering,
Kyoto University, Kyoto, Japan

Abstract

In this paper is described a general framework of a next generation machine translation system which translates a text not sentence by sentence but by considering inter-sentential discourse. The method is a step closer to human translation than the present-day machine translation systems. Particularly important are a detailed discourse analysis and a flexible text generation by using information obtained from the discourse analysis.

1 Some Problems in Present-day Machine Translation Systems

Machine translation systems which were implemented on main frame computers in 80's are now all on personal computers and workstations. As the Internet has been penetrated into ordinary home computers in Japan, translation of English text on the Internet is demanded everywhere. To cope with this demand more than one million sets of E/J MT systems have been implemented on PCs for use to read the Internet information. However the quality of translation has not significantly improved over the past ten years.

Present-day machine translation systems are designed on the compositionality principle. A given sentence is analyzed syntactically and is represented as a tree whose leaves are words of a sentence. Those leaves are replaced by words of a target language and minimum tree structure adjustments are done to fit to a target language structure. Then a target sentence is produced from the tree by picking up the words on the leaves of the tree from left to right.

This translation process includes at least the following problems.

1. As the translation process is based on the compositionality principle, the tree structure of a source sentence is basically kept in a target sentence. This is often a cause of ununderstandability in the target language world.
2. Meaning of each word and phrase of a source sentence is not distinguished clearly in the analysis stage, so that a proper choice of translation expression cannot be achieved. Sometimes misleading translations are produced.
3. As the speaker's (or writer's) intension is not obtained from the analysis, proper expressions in a target language is hard to be composed.
4. As the translation is sentence by sentence, anaphoric relation of pronouns and those between nouns are not obtained, and ellipsis cannot be recovered. Without the information of this kind, high quality translation cannot be achieved.
5. As the discourse structure of a set of sequentially written sentences is not analyzed, readability of translated text is very bad.

2 Efforts to Improve Present-day Machine Translation Systems

As present-day machine translation systems have such deficiencies as described above, the systems produce acceptable translation only when each input sentence is self-closed, that is, when it has no strong connection with adjacent sentences. When a sentence has strong dependence on the previous sentences, reasonable translation is hardly achieved unless these adjacent sentences are simultaneously analyzed and mutual dependence is clarified.

We are studying these problems to improve present-day machine translation systems.

(1) We are making efforts to check wide range of a word sequence of a sentence, and to distinguish the function and meaning of a word in a sentence. For example, the meaning of "company" in

I enjoyed your company.
I visited your company.

is determined by seeing the word sequence, "enjoy your company", and "visit your company". In Japanese,

KAISHA-E YUKU.
(to company) (go)

usually means

go to work / go to the office

instead of

go to the company

which is a literal translation of "KAISHA-E YUKU".

(2) However, in a context where you have a business with the company and to see a person of the company, "KAISHA-E YUKU" has to be translated as

go to the company.

Also, when your visit to the company gave you a satisfaction, you may say

I enjoyed your company.

In this case "company" is in the meaning of "firm". In this way translation is quite dependent on circumstance and context.

(3) When we adopt the principle mentioned in (1) above, we come to the principle of analogy translation or example-based machine translation. People speak or write sentences by using analogical expressions to those which they composed or observed in the past somewhere. So, when a translation phrase is known to an expression, the translation of a similar expression may be composed as similar in a target language. For example,

live in a house : IE-NI SUMU
(house) (live in)

will produce by analogical inference,

live in a mansion : MANSHON-NI SUMU
live in an apartment house : APAATO-NI SUMU.

Similarly,

sick children : BYOUKI-NO KODOMO
(sick) (children)

will produce

healthy adults : KENKOU-NA OTONA.
(healthy) (adults)

But in the following sentence:

The medicine cured the sick children.

the translation of "sick children" is better as

KODOMO-NO BYOUKI
(children's) (disease)

than

BYOUKI-NO KODOMO.
(sick) (children)

So, the translation pair:

The medicine cured the sick children.

SONO KUSURI-GA KODOMO-NO BYOUKI-WO NAOSHITA.
(the) (medicine) (children's) (disease) (cured)

can be used in the translation of the following sentence:

The dose hurt the healthy adults.

The translation will come out as,

SONO KUSURI-GA OTONA-NO KENKOU-WO GAISHITA.
(the) (dose) (adults') (health) (hurt)

and this is a better translation than the translation:

SONO KUSURI-GA KENKOU-NA OTONA-WO GAISHITA.
(the) (dose) (healthy) (adults) (hurt)

In this way, example-based machine translation is able to choose better translation without any definite analytical reasons. The choice of this kind is very difficult in the ordinary machine translation systems because the system must have a program to choose the best one from many candidates whose algorithm cannot be described definitely. Another good point of example-based machine translation is that the system's improvement is very simple, that is, just add new examples and their translations

There are of course several weak points in example-based machine translation. The whole process of machine translation cannot be driven by examples, because too many examples are to be stored and the search space for the best combination of example phrases for an input sentence becomes huge. Anaphora resolution and discourse analysis cannot be handled, and so on. Therefore the best way is to combine the syntactic method and example-based method. There are increasing number of machine translation systems which use this compound method.

(4) The information expressed in a sentence includes not only the meaning of a sentence determined by the combination of word meanings, but also the deep meanings and implications of a sentence, and the intension and emotion of the speaker. For example,

Table 1: Determination of Referents

	demonstrative pronoun	personal pronoun	zero pronoun	average
learning sample	87%	100%	86%	87%
test sample	86%	82%	76%	78%

I eat vegetable.

can be translated as

YASAI-WO TABERU
(vegetable) (eat)
YASAI-HA TABERU.

Which of these two is a proper translation depends both on discourse and speakers intension. Besides the standard interpretation of "eat vegetable", it may imply

I don't eat meat.

The following two sentences,

I can't understand.
I don't understand.

are translated into the same expression

RIKAI DEKIMASEN.
(understand) (can not)

However, nuance is not the same for these two. The first sentence expresses definitely the impossibility of understanding, which may imply the demand for better explanation. The second sentence expresses honestly that the speaker does not understand but he has no intention to blame the bad explanation. A better translation for this mental situation will be

WAKARIMASEN
(do not understand)

Machine has to understand this difference.

As long as you work for me, you will do what I say.

is essentially an imperative sentence. These understanding processes are the basis for better translation. We have to make efforts to develop programs which can handle these problems.

(5) Proper treatment of anaphoric expressions and ellipsis is very important in translation. Ellipsis is everywhere in Japanese expressions such as subject omissions, object omissions and others. We have to estimate and recover what are omitted to produce proper English sentences. Anaphora includes not only pronoun references but also noun-noun references and indirect references. We have been studying these anaphora and ellipsis problems by utilizing the production system framework, and by writing many heuristic rules. A recent result is shown in Table 1. Indirect anaphora requires knowledge dictionary. In the example

Table 2: Kind of Relations between Adjacent Sentences. (Si is followed by Sj)

relation	
parallel	Si and Sj describe different aspects of the same topic. Si and Sj are often changeable without changing the context.
contrastive	Si and Sj are contrastive descriptions.
topic chaining	Si and Sj have the same topic.
focus-topic chaining	focus word of Si is the topic of Sj.
elaboration	Sj is an elaborating description of Si.
reasoning	Sj gives a reason for Sj.
cause-result	Sj is the result of Si.
change	the state of Si is changed to the state of Sj.
example	Sj is an example of Si.
question-answer	Sj is an answer to Si.

I saw a house. The roof was red.

the roof in the second sentence is of the house in the first sentence. This reference is possible by the knowledge that a house normally has a roof. We have to construct a knowledge dictionary of this kind as a first step. We are accumulating phrases like "A of B" from a large text corpus in Japanese to identify such a reference as "a roof of a house".

(6) Clarification of relations between adjacent sentences

Sentences are connected by several relations. Typical relations are shown in Table 2. English text does not use sentential connectives so frequently as Japanese. Therefore in Japanese translation we are required to insert proper conjunctions between sentences which do not exist in English text to increase the readability of a text. We have developed many heuristic rules to estimate what relations in Table 2 exist between sentences.

A parser of sentences in the future should be able to handle all these phenomena. The importance of a parser is not in the parsing speed but in the ability that the parser can extract as much delicate information as possible from a text.

3 Towards Machine Translation Through Language Understanding

Translation should aim at the representation of all the information in a source language text in a target language text. Translation is not sentence by sentence, but a whole paragraph or text must be mapped to a whole paragraph or text in a target language. The process may be called translation by understanding a whole paragraph or text. This translation by language understanding is a process whose details are impossible to describe.

What is language understanding? Answer is not simple. Take a simple sentence, for example. Superficial understanding of its meaning may be the same for many persons, but its deep interpretation may differ by individuals. Its implications or its associative inference may be totally different by individuals because they have their own internal knowledge and experiences which are all different. How deep is the inference, what cultural background we can suppose for the interpretation,

etc, depend on a person and a situation. In this way language understanding is quite difficult to define. Therefore I do not go into the discussion of what is language understanding.

Here I would like to define language understanding by computer in the following within the present-day natural language processing technology. Sentences are analyzed syntactically and semantically, and the case frame representations are obtained. Tense, aspect and modality are analyzed. Topic and focus are grasped in a sequence of sentences. Inter-sentential anaphoric references are clarified on the case frame representation. Elliptic elements are recovered and speakers intentions are estimated to a certain extent.

The results of these analyses are integrated into a proper representation. Here we call this representation a semantic network (in a new sense). Language understanding here means to obtain a semantic network from a text. Each sentence of a text corresponds to a portion of a semantic network. This semantic network framework is a kind of neutral representation for many languages. Analysis process from a text to a semantic network representation is still not perfect, but by increasing the contents of knowledge dictionary and by writing more precise heuristic rules for syntax, semantics and discourse, we will be able to get more and more reliable analysis results.

A real difficult problem we have to attack at this stage is therefore to generate a target language text from a semantic network, which should be logically acceptable, highly readable, sufficiently exact and reflect the author's intention and attitude. The following are what we are making effort to realize in our sentence generation study.

1. Generation of a sequence of sentences (that is, coherent text) from a semantic network. Generation must be flexible in the sense that any node of a semantic network can be taken as a starting topic of a text generation. A semantic network may be represented not by a sentence but by several sentences connected by proper conjunctive words.

2. In the generation of a text we can assume the author as adult/child, man/woman, high ranking person or not, and so on, by the choice of which sentential style and word choice may differ significantly. The same is true for the combination of who speaks to whom. The same content can be expressed in a very formal way or in a casual or conventional style. Modality is another parameter for composing a sentence. Expressions are different for the same content depending on the confidence degree of a speaker. All those parameters have effects on a generated text. We have to extract these parameters from the original text for translation during the text analysis process. Sentence generation process should be able to generate any style of text from the same semantic network to cope with varieties of input text style. For example, “have a meal” may be expressed as

   take a meal
   eat food
   dine
   sit (down) at table
   have a lunch (when at noon)

   Choice of an expression is largely dependent on the situation of the utterance as well as on the parameters mentioned above. Japanese language has a rich honorific expressions and the choice of a proper expression in a certain situation is quite complex because it is culture dependent.

3. Pronominalization, proper use of articles, topic/focus distinction which affects sentential style, axe to be considered in the process of sentence generation.

4. Different expressions can cause the same effect. For example, the following two sentences,

   Please lend me this book.
   Please permit me to read this book.

   are uttered from the same demand: "I want to read this book", and the speaker would expect the same speech act to get the book by these expressions. We have to study how we can recognize that these different expressions have the same effect, and how we generate varieties of expressions depending on circumstances and on our own taste.

5. Varieties of dictionaries should be prepared to realize the above mentioned flexible text generation, such as, thesaurus, dictionary of synonymous expressions and so on, with lots of conditional information about who, when, where and how these expression are used.

6..We have constructed a text generation program with small knowledge dictionary, which can produce different style of texts from a semantic network. Few examples are shown in the following with the corresponding semantic networks shown Fig. 1. The experiments were all done in Japanese, and the following are unsatisfactory translation into English of the generated Japanese texts which are generally satisfactory in the text coherency aspect. I hope that the reader may be able to understand what we are trying to achieve from this bad English translation which are not good in word selection and in sentential style.

Example Al (Initial topic is "the Netherlands")

(1) The Netherlands has a mild weather, and dairy farming and floriculture are extensively carried out. (2) It is located in the north-west of European continent. (3) It is an area of production of tulips whose flower is beautiful. (4) Tulip is a perennial grass belonging to Liliaceae, and come into flower in spring. (5) The bulb is planted in autumn.

Example A2 (Initial topic is "tulip")

(1) Tulip is a perennial plant belonging to Liliaceae, and the flower is beautiful. (2) The bulb is planted in autumn. (3) The flower open in spring. (4) The producing center of tulip is the Netherlands where the weather is mild. (5) The Netherlands is thriving in dairy farming and floriculture. (6) It is situated in the north-west of European continent.

Example Bl (Initial topic is "GON")

(1) GON was alone, and was a child fox. (2) He lived in a forest in a mountain. (3) He made fun at a village near the mountain day and night. (4) The mountain was at a short distance from NAKAYAMA. (5) The forest was covered with fernery.

Example B2 (Initial topic is "forest")

(1) The forest was covered with fernery. (2) It was in a mountain near a village. (3) There lived GON who was alone. (4) GON was a child fox, and he made fun at the village night and day. (5) The mountain was a little away from NAKAYAMA.

Example Cl (Initial topic is "www")

(1) WWW is an abbreviation of World Wide Web, and is accessed by a browser which displays hypertext. (2) The browser can fetch texts. (3) Hypertext has pointers to texts.

Example C2 (Initial topic is "browser")

(1) A browser makes access to WWW which is an abbreviation of World Wide Web. (2) It displays hypertext which has pointers to texts. (3) It can fetch texts.

4 Conclusion

After fifteen years of practical experience of machine translation systems, we have to consider seriously a framework of machine translation through language understanding. Text analysis method has been advanced step by step these fifteen years including anaphora and discourse analysis. Therefore our problem now is text generation which is flexible to varieties of environmental conditions and which is coherent, easily readable and understandable. Varieties of sophisticated information which are extracted from a source text should be utilized in a target language text generation. For a flexible generation of texts we have to prepare lots of culture dependent information. Particularly important is the construction of dictionaries of synonyms, antonyms, and expressions which have similar meaning in deep sense. Also we have to study more heavily the human mental behavior in connection with sentential composition. By making constant efforts in the problems which I mentioned in this paper, I guess that the machine translation framework through language understanding will be realized in another ten years.

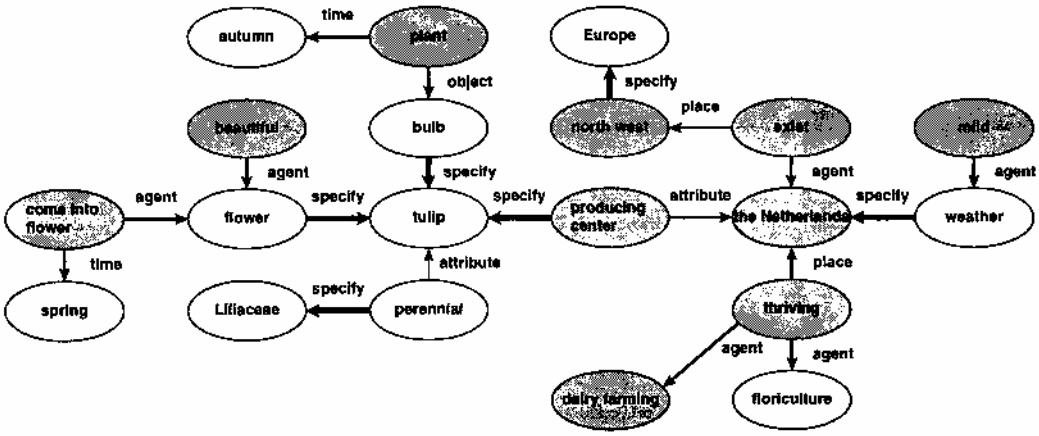

Figure 1-A: Semantic network for texts of example A.

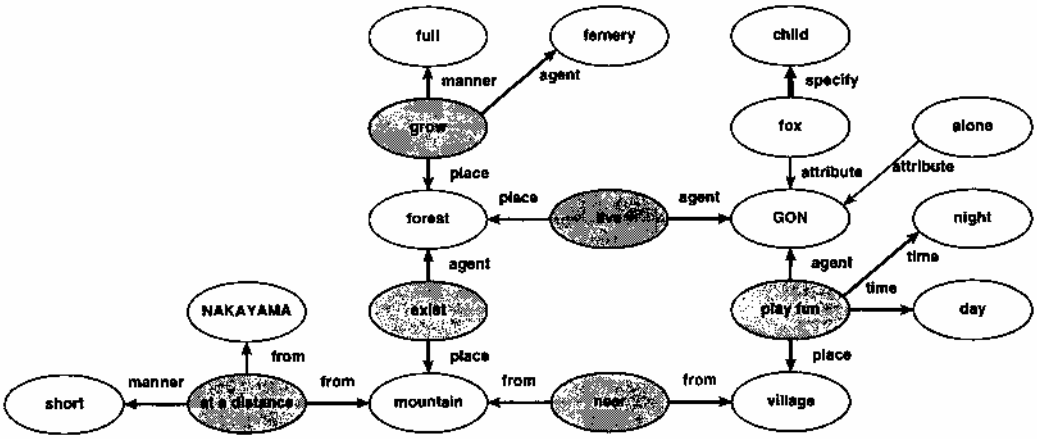

Figure 1-B: Semantic network for texts of example B.

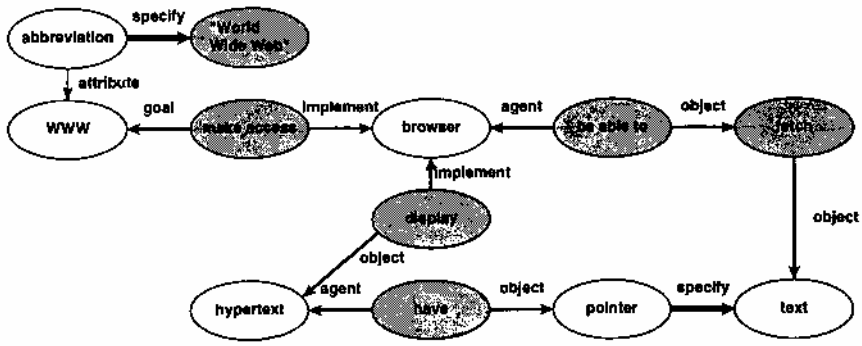

Figure 1-C: Semantic network for texts of example C.