# What kind of information is necessary for NLP and MT?

Makoto Nagao
Department of Electrical Engineering
Kyoto University, Kyoto

## 1 Japanese efforts for electronic dictionaries

Researchers in natural language processing (NLP) and machine translation (MT) in the past were mainly interested in linguistic theories, parsing and generation. They discussed specific types of language expressions such as garden path sentences, and did not pay much attention to problems which may arise when huge volume of real existing sentences are handled, such as newspaper articles, patent documents, and so on. Once we get into this area of processing large text corpus, we confront with several problems such as: we have to write a complete set of grammatical rules, we have to prepare a comprehensive dictionary, and so on. A major problem here is not in the linguistically interesting but seldom-appearing linguistic phenomena, but in the average success rate of parsing, generation etc. for a large text corpus, which seldom includes such sophisticated sentential structures as garden path sentences. It includes different types of difficult problems, for example, parsing of long sentences such as sentences composed of more than thirty words, and building a good lexicon.

In the university researches where new idea is most important, building a good complete lexicon has not been a central issue. At companies on the contrary where commercial machine translation systems are commercialized, people are forced to construct a complete dictionary which includes all the common words and sufficient number of terminology words in several specific fields. Dictionary construction takes a long time and a big money. It has to have a consistency in the whole dictionary, and the quality of the contents must be uniform for all the words. Japanese companies spent a lot of money for the dictionaries for machine translation. From this bitter experience they got together to construct a basic electronic dictionary which they can share themselves by the help of the Japanese Government.

Before getting into some more details of this EDR project, let me describe a brief history of electronic dictionary efforts in Japan. The first attempt to computerize an ordinary Japanese dictionary was at Electrotechnical Laboratory in the middle of 1970's where an ordinary Japanese dictionary: Shin Meikai Kokugo Dictionary (about 60,000 words) was put into a computer. Soon after this attempt we computerized an English-Japanese dictionary (Concise English-Japanese dictionary). Bunrui-Goi-Hyo (Japanese thesaurus) was computerized in the middle of 1980's, which was similar to Roget's thesaurus and included about 60,000 words.

A Japanese word processor was first commercialized in 1978. It was accepted widely in the Japanese society, and in the middle of 1980's there were more than twenty companies which produced word processors. All these word processors adopted the method to convert phonetic input (Kana) to Chinese character words. One of the earliest large scale dictionaries for machine translation was the dictionary system built by ourselves for Mu machine translation project during 1982–1986. It included a Japanese dictionary, an English dictionary, a Japanese-to-English transfer dictionary and an English-to-Japanese transfer dictionary, each of which had approximately sixty thousand headwords.

EDR dictionary project started in 1986 and will be terminated in 1994. It includes about 400,000 concepts (different meanings), and the corresponding Japanese and English terms (about 200,000 words each).

IPA organized a group of young Japanese linguists to construct a precise verb dictionary of Japanese. The dictionary is called IPAL dictionary, which has case frames for 861 typical Japanese

verbs. For each Japanese verb surface cases indicated by Japanese postpositions (ga, o, ni, de, etc.) are given with some semantic markers in each case slot. They differentiated 18 semantic markers. For each case frame several typical example sentences are given for the reference.

Nowadays several Japanese dictionaries and English-Japanese dictionaries are used at Kyushuu University and several other research institutes for the purpose of automatic acquisition of lexical knowledge for machine use.

# 2 Information which must be included in an electronic dictionary

When we discuss about what kind of information must be included in a dictionary, we must first discuss about what kind of linguistic processing we are supported to perform. There may exist an opinion that the dictionary contents depend heavily on what kind of linguistic theories or frameworks we take, and therefore we cannot construct a kind of universal neutral dictionary. It is true, but we can imagine a kind of neutral dictionary descriptions the major parts of which can easily be converted for use in a specific linguistic frameworks.

The dictionary information which was included in Mu machine translation system will be still the richest as a Japanese electronic dictionary [1][2]. The followings are the information included in the dictionary of Mu machine translation system.

## 2.1 Mono-lingual dictionary

lexical item, word length, word stem, conjugation, pronunciation

part of speech, subcategorization of part of speech

derivations (noun-, verb-, adjectival-, adverbial-), related words

subject code, semantic code, (thesaurus code), synonym, antonym

aspectual features

volition

case frame

ideomatic expressions

(1) Derivational information is important from the standpoint of sentence generation. For example, we can say

It is a matter of politics.
It is a political problem.
He is not a careful person.
He is a careless person.

(2) We set up the following distinctions for Japanese verbs which showed varieties of aspects and modalities when combined with aspectual and tense- related postpositions.

stative verb,
semi-stative verb,
durational verb,
effective verb,
inchoative verb,
iterative verb,
point-action verb

Verbs are also classified from the standpoint of volition by the agent.

(3) There are a few problems in the case frame description. One is how many cases we distinguish. Another is how many semantic markers we introduce to specify the nouns which can get into a particular case slot. The third one is how many different case frames we can distinguish for a particular verb. We distinguished 32 different cases and used about sixty semantic markers.

Here the number of semantic markers is a particularly important factor in achieving precise analysis of a language. A research group at NTT Research Lab. established a semantic marker system which included about three thousand semantic markers, and constructed a very precise machine translation dictionary. Their MT system could distinguish very sophisticated differences between expressions by their semantic marker system. This result showed that the granurality of semantics must be at least two or three thousand. Recently I discussed that a semantic marker system of that number of semantic markers are almost equivalent to an example-based system where many different examples are accumulated and used as a kind of standards to input sentences to be analyzed, and clarified that a semantic marker system less than that, for example few hundred or less, are far inferior to an example-based system [3].

(4) A case frame which is valid in a language must sometimes be divided into several when a translation into another language is considered. For example, in English

wear a suit,
wear shoes,
wear a pair of glasses,
wear a wristwatch

may be handled by the same case frame whose object slot has a semantic marker "things which can be put on human body". However, when we consider the Japanese translation to these expressions, we have to have the following distinction.

wear a suit : kiru(object(suit))
wear shoes : haku(object(shoes))
wear a pair of glasses :
kakeru(object(glasses))
wear a wristwatch :
tsukeru(object(wristwatch))

That is, the case frame for "wear" must be separated at least into these four different frames where each objective case has its own subcategorized semantic marker.

## 2.2 Bi-lingual dictionary

Bi-lingual dictionary connects corresponding expressions in one language and another via a concept which is supposed to be common or neutral to these two languages. EDR distinguished about 400,000 different concepts to which words or phrases of Japanese and English are corresponded. Sometimes a concept which can be expressed by a word in a language cannot be expressed by a word in another language and therefore be expressed by a long phrase. For example "Menkui" in Japanese has no word in English, and must be expressed as "a person who puts too much store by good looks". When it comes to a bi-lingual verb dictionary, the situation becomes much more complex. A tree-to-tree transformation must be introduced such as

SHISAKU-SURU(agent(x), object(y))
⟶X make Y on an experimental bases.

# 3 Other information in an electronic dictionary for machine translation

## 3.1 Example phrases

We had an interesting experiment by using IPAL verb dictionary [3]. As is already mentioned IPAL dictionary has corresponding example sentences to each case frame of a verb. When a sentence is given the selection of a proper case frame of a verb in a given sentence is usually done by referring to cases and their semantic markers. However there are only 18 semantic markers and as a result a noun which is specified by one or several semantic markers tends to cover a broader semantic domain than a necessary and sufficient domain. Therefore it happens very often that several case frames are corresponded to a given sentence. But the situation is totally different when a given sentence is compared with example sentences attached to each case frame. In this case a certain similarity calculation is done between the input sentence and example sentences by using a thesaurus, and a unique case frame is chosen properly for an input sentence. We tested these two approaches and obtained the result that the case frame selection by semantic marker was only less than 30% successful, while the case frame selection by example matching was more than 65% successful. This showed us the importance of example phrases in a dictionary.

## 3.2 Metaphor information

Any simple sentences require metaphorical interpretation to some degree. Therefore a dictionary must have information about metaphorical uses of a word. The information is usually given by examples, or by a proper explanation of a metaphorical use. Therefore a sentence interpretation program must have a mechanism which can paraphrase metaphorical expressions. For example "eat fuel up" must be paraphrased as "consume fuel in great quantities" in the process of interpretation.

## 3.3 Knowledge for a proper interpretation or understanding

Extra-linguistic information is required for a proper interpretation and understanding of an expression [4]. Man has a lot of knowledge or a common sense, but he or she does not notice the impor-

tant role of it in the understanding of expressions. For example the following two sentences

> He left the hall through the back door.
> He left the parliament through the back door.

are easily understandable, but a very similar sentence

> He left the government through the back door.

is not interpretable because the government is not a room or a construction and has nothing to do with a door. The hall and the parliament are understood as constructions and we can imagine that they have back doors. If a man is forced to interpret the third sentence then he or she may guess that

> through the back door $\longrightarrow$ secretly, stealthily
> left $\longrightarrow$ resigned

Extra-linguistic knowledge is very often required for the disambiguation of polysemic word meaning. For example

> I went to a bank to open an account.
> I went to a bank to take a walk.

We have a kind of knowledge that we can open an account at a bank, and that there is usually a path on the bank of a river. "The box is in the pen" is another famous example. Therefore we have to have such knowledge in a computer. That is, we have to store a kind of encyclopedic knowledge represented in a certain way which allows machines to use easily. This will be one of the most important problems in natural language processing in the near future.

# References


[1] M. Nagao, J. Tsujii, J. Nakamura: Machine translation from Japanese into English, *Proc. of the IEEE*, Vol. 74, No. 7, July 1986.

[2] M. Nagao: Semantic elements in machine translation, in M. Stamenov ed., *Current Advances in Semantic Theory*, John Benjamins, 1992.

[3] M. Nagao: Some rationales and methodologies for example-based approach, *Proc. Future Generation Natural Language Processing*, July 30–31, 1992, Manchester, pp. 82–94.

[4] M. Nagao: Some dictionary information for machine translation, in Cheng-wing Guo ed., *Machine Tractable Dictionaries—Design and Construction*, Ablex Pub. (forthcoming).