A Machine Translation System for Foreign News in Satellite Broadcasting

Teruaki Aizawa, Terumasa Ehara**, Noriyoshi Uratani, Hideki Tanaka,
Naoto Kato, Sumio Nakase*, Norikazu Aruga*, and Takeo Matsuda*

NHK Science and Technical
Research Laboratories
1-10-11, Kinuta, Setagaya-ku,
Tokyo 157, Japan
**Current Address: ATR Interpreting
Telephony Research Laboratories

*Catena-Resource
Laboratories Inc.
Ichibancho-27 Bldg.
27, Ichibancho, Chiyoda-ku,
Tokyo 102, Japan

A machine translation system of English to Japanese is described, which has been used in 24-hour direct satellite broadcasting by NHK to translate "World News."

In order to treat a wide scope of news sentences, the system is provided with more than 100,000 lexical entries as well as about 3,000 grammatical rules which can robustly analyze various types of undefined words. It is also effective in translation of news sentences to preprocess proper nouns, to resolve structural ambiguities by weighting grammatical rules, and to select appropriate words using semantic markers. The operational experiments on machine translation in satellite broadcasting are briefly discussed.

1 Introduction

Since December 1986, NHK, the Japan Broadcasting Corporation, has been conducting two-channel, direct satellite broadcasting using Japan's BS-2b broadcasting satellite. The two satellite channels now have 24-hour nationwide TV broadcasting services. The core of the services on Channel 1 is "World News," in which news from across the globe is broadcast.

The languages spoken in NHK's World News are English, French, German, Italian, Russian, Korean, and Chinese. Urgent and important news has simultaneous interpretation services. In usual cases, however, services only superimpose Japanese subtitles on the TV screen. Actually, more than 50 bilingual translators prepare a manuscript by transcribing and translating the original news. All the work must be done in a limited time, even at midnight due to the time difference between Japan and other countries.

A machine translation system was introduced to make easier this daily work. As a first step, the English World News has been experimentarily broadcast, about 5 minutes a day, using the Japanese translation provided by the MT system. We think this is cultivating a new possibility of machine translation in Japan [1].

2 Satellite Broadcasting and Machine Translation

Usually the generation of the subtitles proceeds as follows:

1) a bilingual translator prepares a manuscript by transcribing and translating the original news;
2) a supervisor examines the manuscript; and
3) an operator inputs the final manuscript into the processing equipment.

Our MT system was introduced in step 1. First of all, the original news is greatly summarized in English since the length of a subtitle script is at most 30 Japanese characters per a display screen. Preediting is also carried out in this step to provide a better input for the MT system. After postediting, the final result is given to step 2.

The system is based on the STAR machine translation system [2], and works basically by a transfer method. The translation process can be divided into 4 main steps: morphological analysis, syntactic analysis, transfer, and generation. The morphological analysis identifies words as well as locally fixed sequences of words. In the syntactic analysis, all the possible surface structures for an input sentence are derived, and then the best candidates are chosen by using the "weight mechanism" described below.

At present, the size of the dictionary is about 100,000 entries, and the grammar has about 3,000 CFG-type rules. The system can translate a sentence having 11 words on average within 2 seconds using a 3 MIPS UNIX computer. Further characteristics of the system are discussed below.

3 Characteristics of the Machine Translation System for News Sentences

3.1 Characteristics of News Sentences

Examining a large body of English news consisting of more than 3.5 million words from the World News and the basic news service of AP (Associated Press),we can summarize the linguistic properties of the news sentences as follows:

1) About 75,000 different words are used, and they are difficult to classify by news fields.

2) Various types of proper nouns such as human, nation and location names appear frequently. Human names are often with related words like titles.

3) Many verbs having human subjects are used. Among others, "say," "call," "report," "talk," "ask," "think," "want," and "feel" are often found.

4) Many kinds of numeral expressions come out. Some of them are too complex to translate.

5) Colloquial expressions appear frequently.

3.2 Local Preprocessor for Proper Nouns

In order to treat the above-mentioned characteristics 1) and 2) of the news sentences, our MT system has a preprocess called "Local Context Translation"(LOCT), which constitutes the second part of the morphological analysis. Its main role is to identify and translate various types of locally fixed sequences of words such as

"U.S. President George Bush,"

"July 14th, 1789,"

"The Metropolitan Museum of Art, New York," etc.

Rules for human names with the title of the position

```
human      -->   khuman  name
khuman     -->   (FORMER) rank
rank       -->   STATE PRESIDENT
                 ;[=>"大統領"]
           -->   CHINESE PRESIDENT
                 ;[=>"国家首席"]
           -->   party PRESIDENT
                 ;[=>"委員長"]
           -->   firm PRESIDENT
                 ;[=>"社長"]
FORMER     -->   former,acting,etc.
STATE      -->   U.S.,French,and etc
                 ;nation names except "Chinese"
CHINESE    -->   Chinese
party      -->   ...
                 ;party names
firm       -->   ...
                 ;company names
```

Rules for defining the human names

```
name       -->   HNAME
                 ;for defined words
           -->   (R_NAME)+
                 ;for undefined words
HNAME      -->   ...
                 ;names defined in the lexicon
R_NAME     -->   [A-Z][a-z]+
                 ;names defined from the input
```

Figure 1 Rules for the Local Context Translation.

The LOCT can perform translation of human names with related words like titles, identification of undefined proper nouns, and selection of words. To analyze local patterns, the LOCT has a set of CFG-rules, different from the global analysis rules, as shown in Figure 1.

By these rules, "President" can be translated differently into Japanese depending on the previous word. R_NAME picks up an undefined proper noun from the input text as a sequence of one capital letter and some small letters.

3.3 Robust Processing of Undefined Words

In addition to a large dictionary, our system has a powerful processor for undefined words to cover a wide scope of news sentences. The processor estimates the lexical items of undefined words, and gives them to the syntactic analyzer.

The main processing functions are:

1) Identification of locally fixed patterns : As explained in 3.2, the LOCT can estimate the grammatical values for undefined words by identifying local patterns.

2) Estimation by the ending form of a word : Many English words have their own ending forms corresponding to grammatical values. For example, a word ending with "-tion(s)," "-ly," or "-able" can be estimated as a noun, an adverb, or an adjective, respectively.

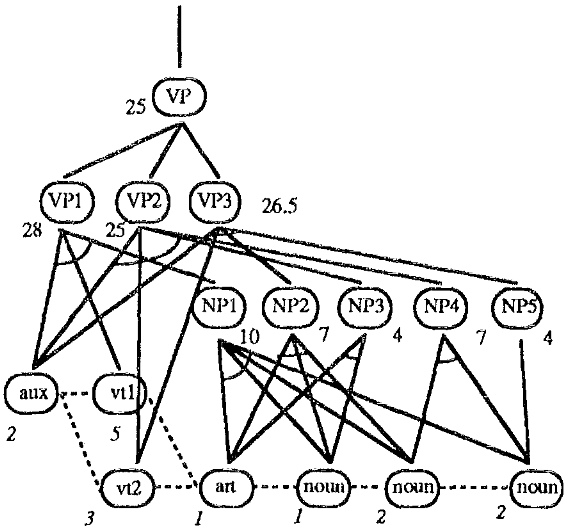
3) Special cases : The processor has some heuristic rules for a word starting with a capital letter, a short word, a sequence of numeral digits.

3.4 Structural Disambiguation by Weighting Grammatical Rules

The syntactic analysis consists of two steps.

1) All the possible surface structures for an input sentence are derived as an AND/OR graph [3].

2) The best candidates are extracted from this graph using our "weight mechanism," which can be formulated as a search problem for an AND/OR graph having nodes with costs.



(He) is teaching the girl English grammar.

: 9 VP --> (+2 aux) vt1 NP
: 5 VP --> (+2 aux) vt2 1.5 : NP NP
: 2 NP --> (art) (+1 nou) * nou

Figure 2 AND/OR graph for a VP

Figure 2 gives an AND/OR graph for a verb phrase: *"(He) is teaching the girl English grammar."* where italicized numerals show the weights of words given by the lexicon. The weight of the other node is calculated from those of the daughter nodes and the corresponding rule. For example, the weight of VP3-node is calculated by using the second rule in Figure 2 as follows:

$$2(\mathrm{aux})+3(\mathrm{vt2})+1.5*7(\mathrm{NP2})+4(\mathrm{NP5})+5+2 = 26.5$$

Among three VP candidates in this example, VP2 is chosen as the best one, since it has the smallest weight.

The weight represents some kind of "incomprehensibility," "complexity," or" rareness" of a word, a phrase, or a sentence. All the words and rules in our system have been assigned their own weights. Our experiments on machine translation in satellite broadcasting show that the best candidates are chosen for about 78% of the successfully analyzed World News sentences.

3.5 Word Selection by Semantic Markers

Semantic markers are employed for Japanese word selection. Their effectiveness has been shown particularly in the areas mentioned below.

1) Selection of a Japanese translation of "they"

The word "they" quite often appears in news sentences. It has two major Japanese translations:"*karera*" and "*sorera*" which refer to objects with will and objects without will, respectively. The confusion between the two intolerably degrades Japanese translation.

A simple strategy that uses semantic markers and verb characteristics can make a proper selection in many cases without pronoun analysis. As mentioned in 3.1, verbs having human subjects are frequently utilized in news sentences. Meanwhile, verbs like "melt" take subject nouns that have no will. If "*karera*" has a marker [HIWILL] (objects with high will) and "*sorera*" has nothing, translation control of "they" is realized by specifying the subject of a verb as [HIWILL] or nothing.

2) Basic verb's translation word selection

Verbs frequently used in news sentences are basic and thus have various meanings. To select a proper Japanese translation of a basic verb, we have set some special markers for news sentences. One of them is [CRIMINAL] which is utilized to obtain a special translation of "catch." Consider the sentence:

"*The police caught the assailant, who has a history of mental illness.*"

The word "catch" was successfully translated as "*taiho-suru* (arrest)," since "assailant" has the marker [CRIMINAL] and the translation description of "catch" defined its Japanese as "*taiho-suru*" when it took an object noun that belongs to [CRIMINAL].

4 Results, Considerations, and Problems

The following results were obtained for 1,393 World News sentences which were input to our MT system during the three months of our trials.

On a strict judgment, the number of successfully analyzed sentences was 898 (64.5%), of which 698 sentences (78%) were properly translated as first candidates by our weight mechanism. As far as these sentences were concerned, the mechanism was very effective. About 30% of the failure in analysis was due to errors in the input sentences such as misspelling or grammar mistakes. Colloquial expressions were also difficult to analyze.

<u>Example of translation</u>

"*Mrs. Nishi, with the help of a lawyer, is trying to collect workman's compensation for her husband's death.*"

==>

" 一人の弁護士の助けがある　Nishi　夫人が、
彼女の夫の死のために労働者災害補償を
集めることを試みている。"

5 Concluding Remarks

We described a practical machine translation system of English to Japanese which has been utilized in satellite broadcasting by NHK to translate the World News.

Toward the second stage of application to broadcasting from April 1990, we are now trying to improve our system specifically in treatments of numeral expressions and colloquial expressions which have not yet been fully considered though they are major characteristics of news sentences.

We also started a design of a French-Japanese machine translation system based on a similar structure to our present English-Japanese system.

References

[1] M. Nagao et al.: A Japanese view of machine translation in light of the considerations and recommendations reported by ALPAC, U.S.A., Machine Translation System Research Commitee, JEIDA (1989).

[2] S. Nakase: On syntactic analysis technique in English-Japanese machine translation, SIGNL Meeting of IPSJ, 69-7 (1988).

[3] A. Martelli and U. Montanari: Optimizing decision trees through heuristically guided search, CACM, 21(12), 1025-1039 (1978).