SUPER-FUNCTION BASED JAPANESE-ENGLISH MACHINE TRANSLATION

Manabu Sasayama Fuji Ren Shingo Kuroiwa

Department of Information Science & Intelligent Systems
Faculty of Engineering, Tokushima University
E-mail: {sinoyama, ren, kuroiwa}@is.tokushima-u.ac.jp

ABSTRACT

We have presented a new method called Super-Function Based Machine Translation (SFBMT). SFBMT uses Super-Function (SF) to translate without syntactic and semantic analysis as most conventional MT systems do. SF represents correspondence with a source language sentence structure and a target language sentence structure directory. According to this feature, the system realizes very fast translation. Moreover we do not feel the sense of incompatibility in the translation result because the SFs are automatically generate from a large corpus. In this paper, we dscribe (1) SF based machine translation, (2) extracting SF, (3) the evaluation experiment. As the experimental results, we obtain 84% of effective rates of translation.

Keywords: Machine Translation, Super-Function, Corpus base

1. INTRODUCTION

Almost no user who is satisfied with a machine translation system. They express dissatisfaction with quality of translation or accuracy of translation. Especially generation of an unnatural translation loses their reliance.

The method of translation used with many commercial machine translation system is a rule based machine translation. The rule based machine translation must describe grammar or a rule of various restrictions by manual. Therfore, as a problem, the rule which covers all of a various language phenomenon is described by manual, which is difficult. Moreover, even if the rule which covers all of a various language phenomenon is described, a natural translation is not necessarily generated. In order to avoid those problems of the rule base translation, the translation method using a corpus is proposed [Uthino, Sirai, Yokoo and Oyama, 2001, McTAIT, 2001, Izuha, 2001, Echizenya, Araki, Momouchi and Tochinai, 2001, Aramaki, Kurohashi, Sato and Watanabe, 2001]. The advantage of corpus based machine translation can generate a natural translation. Although flexibility is low because the corpus is used directly. In addition, In order to acquire high translation accuracy, a huge quantity of a corpus is required. So many of researches of the method of corpus based machine translation, performing processing like the rule base avoid these problem. But by this method, the problem of the rule base will also be included. It will be contrary to the view of the method of corpus based machine translation.

In the present condition of such machine translation, we propose a method using SF ([Ren, 1997, Ren, 1999]), which is a corpus base. This is called Super-Function Based Machkine Translation (SFBMT). The SF is a function that shows the correspondence between source language sentence patterns and target language sentence patterns. Currently, we consider the SF only for nouns and other sentences. Our experiences show that SFBMT has following advantages;

1. Because the detailed syntactic analysis and semantic analysis have not been used, the translation speed is very fast.
2. Because SF is a corpus base, a fluent translation sentence is generated.
3. Because a corpus is held as a function, the quantity of a corpus required for translation becomes less.

In this paper, first, we describe SF based machine translation, extracting SF from a bilingual corpus automatically and the evaluation experiment. In this study, we need morphological analysis. Therefore, we use morphological analysis tool in japanese (Chasen)[Nara Institute of Science and Technology].

SF is defined in Chapter 2. The translation procces is described in Chapter 3. Chapter 4 describes the extraction method of SF. The translation experiment is conducted in Chapter 5.

2. SF DEFINITION AND FORMAT

2.1. SF DEFINITION

A Super-Function (SF) is a function that represents the correspondence between source and target language sentence patterns. SF is defined a structure to remove a noun from a sentence. A removed part-of-speech is not only a noun. But in this paper, we pay attention only a noun. A SF can be represented as formula (1) using a formal description.

$$[O\_STRing] << O\_Noun > + < O\_STRing > * > + [O\_STRing] \rightarrow SF(T\_STRing, T\_Noun) \quad (1)$$

Notation: [] means optional (i.e., 0 or 1); + means 1 or more; * means 0 or more; O means source language; T means target language. STRing means a natural language (source language or target language) character string expect nouns (that is SF); O_STRing means a source language character string; T_STRing means a target language character string; → means the correspondence.

Moreover we sometimes rewrite SF (1) as formula (2). This formula is more general and more comprehensible than formula (1).

$$SF\_O(O\_STRing, O\_Noun) \rightarrow SF\_T(T\_STRing, T\_Noun) \quad (2)$$

This formula shows that SF consists of string and nouns. We show some SF as follows.

f1:< J_Noun > → < E_Noun >
ex. Eikoku → the United Nations

f2:<J_Noun> 1 NOMINARAZU <J_Noun>2
→ not only <E_Noun>1 but also <E_Noun>2
ex. YUTAKA NOMINARAZU KENKOU HA JYUYOUDEARU
→ Not only wealth but also health is important.

f3: <J_Noun>1 HA <J_Noun>2 MADE <J_Noun>3 NI NOTTA
→ <E_Noun>1 took <E_Noun>3 to <E_Noun>2
ex. KARE HA EKIMADE TAXI NI NOTTA → HE TOOK A TAXI TO THE STATION

Notation: In the examples, J means Japanese, E means English.

f1 is a bilingual dictionary. The SF, such as f1 will not be discussed in this paper for brevity. f2 and f3 are basic formula. The number of nouns corresponds in the same sentence.

2.2. SF FORMAT

We use an example to describe the SF format. We use the following notation when describing a SF. First, variable (nouns) is represented with X, constant is represented with S. Lowercase o and t shows Original language and Target language. The symbol $\prod$ means string connection. The following is a SF according to above example(3).

$$S_o 0 \prod X_o i S_o i = S_t 0 \prod X_t j S_t j \quad (3)$$

$S_o i$:constant of source language
$S_t j$:constant of target language
$X_o i$:variable of source language
$X_t j$:variable of target language

i and j represent each number. Further, Table (VLCT) is a table which represents the correspondence location relationship between the variable source and target language. The following is example. VLCT of the example is shown Table 1.

Japanese:
Kare ha Eki made Taxi ni notta.
English:
He took a taxi to the station.
Japanese:
$S_o 0 = \phi$ $X_o 1$=Kare $S_o 1$=ha $X_o 2$=Eki
$S_o 2$=made $X_o 3$=Taxi $S_o 3$=ninotta.
English:
$S_t 0 = \phi$ $X_t 1$=He $S_t 1$=took $X_t 2$=a taxi
$S_t 2$=to $X_t 3$=the station $S_t 3$=.

VLCT_XY means that the VLCT is constructed based on X language. VLCT_YX means that the VLCT is constructed based on Y language. X and Y are Japanese or English in this table. VLCT_JE (i) =j represent Japanese variable (i) correspond to English variable (j). In the same way, VLCT_EJ (i) =j represent English variable (i) correspond to Japanese variable (j). It is clear that the VLCT_YX can be easily inferred from VLCT_XY.

Table 1. example VLCT

Japanese-English	English-Japanese
VLCT_JE(1)=1	VLCT_EJ(1)=1
VLCT_JE(2)=3	VLCT_EJ(2)=3
VLCT_JE(3)=2	VLCT_EJ(3)=2

Table 2. NTB(left) and ETB(right)

J	E	J	E	condition
ϕ	ϕ	1	1	1p
ha	took	2	3	a
made	to	3	2	the
ninotta.	.	-	-	-

3. SF BASED MACHINE TRANSLATION

3.1. SF ARCHITECTURE

Two architectures could be considered representing the SF. One is a Directional Graph (DG). Another one is a Transformation Table (TTB).

3.1.1. DIRECTIONAL GRAPH

Circle node and edge composes Directional Graph. In SF, Circle nodes represents constants and edges represents variables. This is proved using the example of the fig.(2,1). ϕ in value of the first circle is null character in this example fig.(2,1). Null character is used when there is no word which corresponds to the first character. And use when there is no node which corresponds to target language.

3.1.2. TRANSFORMATION TABLE

SF is Transformation Table (TTB). TTB consists of a Node Table and a Edge table. NTB is also called Constant Table (CTB). ETB is called Variable Table (VTB). NTB is correspondence table of SF of source language and target language. ETB is correspondence

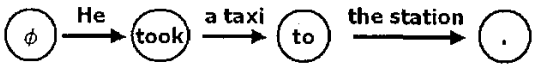

Figure 1. English sentence DG

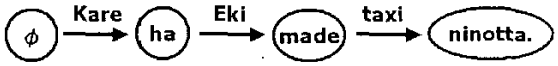

Figure 2. Japanese sentence DG

Table 3. Unite table

J	E	edge	condition
ϕ	ϕ	1	1p
ha	eat	3	a
de	in	2	the
wotaberu.	.	-	-

table of variable. ETB represent the order of variable of source language in target language. ETB describe condition of noun to distinguish inanimate and living thing. Table 2 shows the example of NTB and ETB. Left is NTB and right is ETB in Table 2. Condition in Table 2 represent that condition "1p" is the code of declinable pronouns, condition "a" and "the" are article. However, the condition of an article is used only when strict translation is required. The purpose of the role of this NTB is to prevent "I takes...". English is because a verbal singular and the verbal plural exist.

Theoretically, NTB and ETB are required. It is used in fact, unifying NTB and ETB. Unified table is called unite table. Table 3 shows Unite table. Unite table allocates special edges in individual SF. This table 3 shows matching a SF and the order of nouns. Unite table is simpler than NTB and ETB.

3.2. TRANSLATION PROCESS

This section explains SF based translation process. SF required for translation assumes that sufficient quantity is obtained. The next chapter describes the extraction

Table 4. Unite table in actual system

Japanese node	Z:ha:made:taxi:ninotta.
Order · Fixed noun	Z · Z
Order · Noun	1:3:2 · 1p:a:the
English node	Z:took:to:.

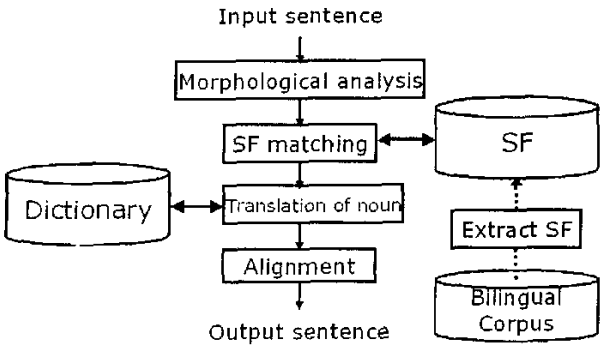

Figure 3. Outline of the system

method of SF.

A flow below shows the algorithm of SF based translation process.

1. Original language sentence is analyzed in the morpheme.
2. Searching matched SF in NTB.
3. Take out condition and order of noun in ETB.
4. The noun translates in the dictionary.
5. Align the SF and nouns.

3.2.1. A PRACTICAL EXAMPLE OF TRANSLATION

Fig.3 shows the outline of SF based translation system. This system reads SF file(data base) and Japanese-English word dictionary at the time of starting, and holds them to hash. The advantage of using hash is that search speed is very quick. Table.4 shows the example of SF read into the system. The hash key of SF is a Japanese node. The record is a English node, a Fixed noun and a noun. A fixed noun is a Japanese noun without an English noun, when SF is extracted. Chapter 4 describes in details. When there are conditions of a fixed noun, for example, in the case of AA, if the 3rd noun of the inputted Japanese sentence is not surely DD, it does not match. ":" is delimiter. "Z" is null character which is "ϕ" in the unite table.

An input sentence "Kare ha Eki made Taxi ni notta.(He took a taxi to the station.)" is actually translated. Fig.3 shows a Input sentence can translate at four steps. The 1st step is a morphological analysis. The 2nd step is a SF matching. The 3rd step is a translation of a noun. The final step is a alignment.

First step:Morphological Analysis

Kare ha Eki made Taxi ni notta.
⇓Morphological Analysis
/Kare/ha/Eki/mada/Taxi/ninotta/./

Nouns are $[Kare][Eki][Taxi]$ and others are $[Z][ha][made][ninotta]$.

Second step:Matching SF in Table 4

Japanese node matched $[Z][ha][made][ninotta.]$. A Fixed noun is nothing. As a result, english node is "Z:took:to:.".

Third step:Translation of a noun

$[Kare][Eki][Taxi]$
⇓Align
$[Kare], a[Taxi], the[Eki]$
⇓Dictionary
$[He][a\ taxi][the\ station]$

Final step:Alignment

$[He][a\ taxi][the\ station]$ "Z:took:to:"
⇓Align
"He took a taxi to the station."

Translation result is "He took a taxi to the station.". This way, TTB is necessary for the translation. When saying oppositely, Sentences which can be translated increases in proportion to kind of SF in SF based machine translation.

4. EXTRACTING SF

SF can be automatically extracted from a bilingual corpus. This chapter explains the technique of extracting SF from a bilingual corpus automatically. Next, the number of duplications of extracted SF and the number according to node are investigated and considered.

4.1. EXTRACTION METHOD

Extracting SF from bilingual corpus is the same as creating a node table and an edge table. That is, a node table and an edge table are created from a bilingual corpus. First, a morphological analysis divides a Japanese sentence into a morpheme, part-of-speech information is tagged. Next, a sentence is divided into a noun and others using the information of the tagger. Others constitutes a node table. In a noun, the difference between

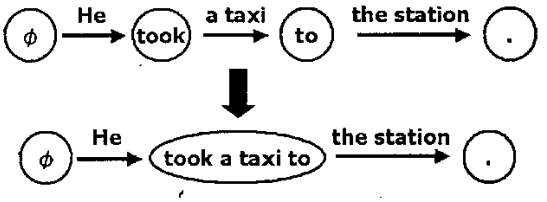

Figure 4. Admission of a noun

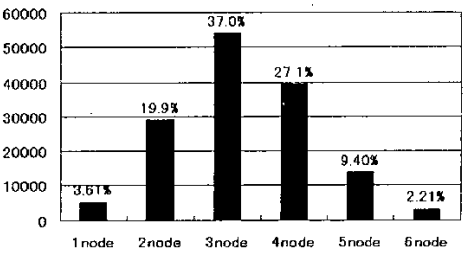

Figure 5. Total SF

the order of the noun in a Japanese sentence and the order of the noun in English is checked using a dictionary. It is the order of a noun. In addition, It is confirmed that each noun is a pronoun or a general noun, existence of a article. In this general way, because the order and conditions of a noun became clear, an edge table is constituted.

Incidentally, SF cannot extract when the number of Japanese nodes differs from the number of English nodes. Example, When there is an English noun but there is not a corresponding Japanese noun. In such a case, Number of English node is more than number of Japanese node. Then, it is necessary to make the same number of English node and Japanese nodes. For that reason, English noun without a Japanese corresponding noun is made into a node. This processing is defined as admission of a noun. Fig 4. shows admission of a noun. SF can be extracted as a result of adjusting the number of nodes by this processing. However, this processing is applied only to English sentence and cannot be applied to Japanese sentence. Because if this processing is applied to a Japanese sentence,the morphological-analysis result of an input sentence is not matched at the time of translation. Thus the information about the noun and the order is held instead of applying this processing to a Japanese noun without a corresponding English noun. This information on a noun and an order are the fixed noun in Table 4.

4.2. THE EXTRACTION RESULT OF SF

SF was extracted from the about 200,000 pairs of Japanese-English bilingual-corpora sentences of EDR. About 140,000 SF has been extracted except for overlapping SF. Duplicate SF is an about 50,000-pair sentence. The bilingual sentence which failed in extraction is about 10,000. Fig. 5 shows the extraction result classified by node. 1node is "thanks" , "sorry" , "・・・". According to Fig. 5, 99% of the total is formed by less than six nodes. In contrast, there were extremely few seven or more nodes.

4.3. DISCUSSION

Almost all SF is concentrating on less than six nodes because the object which SF extracts is only a simple sentence. The number of extracted SF was 140,000 (except for duplication), peculiar sentences are almost all. However, Japanese has two or more sentence end expression of the same meaning. If expression is unified, duplication will increase. As a result, many sentences can be translated in one SF.

5. TRANSLATION EXPERIMENT

The translation experiment was conducted using the system of Chapter 3, and SF extracted in Chapter 4.

5.1. EXPERIMENTAL CONDITION

SF used for the experiment is 140,000 which was extracted from the EDR corpus in Chapter 4. The Japanese sentences for evaluation are 256 Japanese sentences which are high school English reference book Japanese-English bilingual examples. The success conditions of translation are that coincidence of a translation result and English of an evaluation sentence is conditions. However, SF with two or more translation candidates, if the candidate who is in agreement with English of an evaluation sentence is also one, it would succeed in translation. Moreover, the translation result of a noun is not included in translation success conditions. It is because the translation result of the noun in a dictionary has the possibility of synonymous words as English of a translation result. A manual judges that the

Table 5. Translation experiment result

	Translation success	SF matched
256 sentences	216 sentences (84%)	226 sentences (88%)

translation result is in agreement with English of an evaluation sentence.

5.2. EXPERIMENTAL RESULT

Table 5. shows the translation result. According to it, 216 sentences succeeded in translation among 256 sentences. However, other candidates existed in 86 sentences among 216 sentences which succeeded in translation. Next, in coincidence of only a Japanese node and conditions (SF matched), it increased by ten sentences and they were 226 sentences among 256 sentences. Then, the example which succeeded in translation is as follows.

Input sentence:Kyonen ha *yuki* ga ookatta.
Translation result:We had a lot of *snow* last year.

5.3. DISCUSSION

The system was able to output the fluent translation result as a result of the experiment. Especially the system was able to be translated also in the required sentence of a formal subject. As a problem, how is a candidate reduced when there are two or more translation candidates. Besides, in English, "ride" and "take" are properly used by the noun. But in Japanese, it is the same "noru".

6. CONCLUSIONS

In this paper, we define the SF for the construction of Super-Function based translation system. We constructed an experimental system and verified effectiveness of SF.

In the chapter 2 and 3, we define SF and confirm flow of SF based translation, because we examine a necessary one for the system. As a result, the following are necessary.

1. Morphological analysis tool
2. Transformation Table and SF
3. Bilingual dictionary

In these parts, we use existing parts in 1 and 3. We make Transformation Table. There are Node Table and Edge Table in transformation table. Node Table makes Original language SF correspond to Target language SF. Edge Table keeps the condition and the order of nouns. To keep simple in this time, Conditions of nouns are not thought. Moreover we unite Node Table with Edge Table. In chapter 4, we extracted SF from the bilingual corpus. Admission processing of a node can extract SF from almost all sentences. In chapter 5, the result of the evaluation experiment, the rate of translation of the system using this technique was 84%. In addition to all translation sentences were fluent sentences.

The future works, we need to narrow the translation candidate of SF using a concept dictionary. Besides, we need to improve the quality of SF by adjusting the sentence end of a Japanese sentence.

7. ACKNOWLEDGMENT

This work has been partly supported by the Education Ministry of Japan under Grant-in-Aid for Scientific Research on Priority Areas and Grant-in-Aid for Scientific Research B. The authors wish to thank all our colleagues participating in this project.

REFERENCES

[1] Utino,Sirai, Yokoo,Oyama, Furuse, 2001, the Institute of Electronics,Information and Communication Engineers, vol.J84-D-2 No.6, pp.1167-1174.

[2] McTAIT, K. Memory-based translation using translation Patterns, 2001, UMIST

[3] Izuha, T. Machine Translation Using Bilingual Term Entres, 2001, the Institute of Electronics Information and Communication Engineers, vol.101,No.189, July, pp.1-7.

[4] Echizenya, H. Araki, K. Momouchi, Y. Tochinai, K. Recursive Acquisition Method of Translation Rules by Focusing on Local Parts:GA-ILMT2,2001, the Institute of Electronics Information and Communication Engineers, vol.101, No.189, July, pp.9-16.

[5] Aramaki, E. Kurohashi, S. Sato, S. Watanabe, H. Finding Translation Correspondences from Parallel Parsed Corpus for Example-based Translation, 2001, the Institute of Electronics Information and Communication Engineers, vol.101, No.189, July, pp.1-7.

[6] Ren, F. Super-function based machine translation. 1997, Language Engineering, Proceedings of JSCL and TsingHua University Press, pp.305-312.

[7] Ren, F. Super-function based machine translation. 1999, Communications of COLIPS, pp.83-100.

[8] ChaSen version 1.0 is officially released on 19 February 1997 by Computational Linguistics Laboratory, Graduate School of Information Science, Nara Institute of Science and Technology