JOURNAL OF KOREA MULTIMEDIA SOCIETY, VOL. 8, NO. 6, JUNE 2005(pp. 790-797)

Resolving Multi-Translatable Verbs in Japanese-TO-Korean Machine Translation

Jung-In Kim[†], Kang-Hyuk Lee[††]

ABSTRACT

It is well-known that there are many similarities between Japanese and Korean language. For example, the order of words and the nature of the grammatical conjugation of both languages are almost the same. Another similarity is the frequent omission of the subject from a sentence. Moreover, both languages have honorific expressions and the identical concept for expressing nouns in terms of Chinese characters. Using these similarities, we have developed a word-to-word translation system which does away with any deep level analysis of syntactic and semantic structures of the two languages. If we use these similarities, the direct translation method is superior to the internal language translation method or transfer-based translation method. Although the MT system based on the direct translation method is more easily developed than the ones based on other methods, it may have a lot of difficulties when it tries to select the appropriate target word from ambiguous source verbs. In this paper, we propose a new algorithm to extract the meaning of substantives and to make use of the order of the extracted meaning. We could select 86.5% appropriate verbs in the sample sentences from IPAL-verb-dictionary. 13.5% indicates the cases in which we could not distinguish the meaning of substantives. We are convinced, however, that the succeeding rate can be increased by getting rid of the meaning of verbs thatare not used so often.

Keywords: Polysemy, verb, translation, Japanese, Korean

1. INTRODUCTION

Today, information exchange between countries is increasing enormously, and machine translation systems are also advancing rapidly. The studies of machine translation systems are improving. In the case of Korean-Japanese machine translation, there has been a great deal of research and development for years[1-4]. Machine translation systems which use intermediate languages to mutually translate several natural languages have attracted more attention than direct translation systems which performs "one to one" translations from the source language into the target language [1-4,7,8]. For Japanese and Korean only, a direct machine translation system is much more advantageous over an intermediate language system by avoiding many unnecessary passes. Moreover, Japanese and Korean have many grammatical similarities[3,4].

Since the word order is similar in Korean and Japanese, most of the syntactic and semantic analysis can be simplified for machine translation [3].

However, adirect Japanese-Korean machine translation system has problems with conjugated declinable words and multi-translatable words in Japanese. For the former, some solutions are

※ Corresponding Author : Jung-In Kim, Address : (608-711) 535 Yongdang-dong, Nam-gu, Busan, Korea
TEL : +82-51-610-8393, FAX : +82-51-610-8347
E-mail : klee@tit.ac.kr
Receipt date : April. 7, 2005, Approval date : April. 26, 2005
[†] Dept. of Game Engineering Tongmyong University of Information Technology
[††] Dept. of Game Engineering Tongmyong University of Information Technology
(E-mail : klee@tit.ac.kr)

known, but for the latter, only a multi-translatable word processing method using case forms of simple sentences has been proposed.

In this paper, we discuss a method for searching and selecting the most appropriate target word of multi-translatable Japanese verbs by using semantic analysis of indeclinable words in the source sentence.

2. POLYSEMY AND MULTI-TRANSLATABILITY OF VERBS

2.1 Polysemy of Verbs

When a verb appears in a sentence, its meaning can be often ambiguous. Different semantics are called sub-entries of a verb, and the property of having several sub-entries is called polysemous verb. For example, the Japanese verb "AOGU" has the four sub-entries, as in Illustration 1-4, whose semantics is determined by the verb-preceding objects.

Illustration 1.
彼は空を仰いだ。
그녀는 하늘을 우러러 보았다.
Illustration 2.
彼は先輩を師と仰いだ。
그는 선배를 선생으로 받들었다.
Illustration 3.
課長は部長に指示を仰いだ。
과장은 부장에게 지시를 앙청했다.
Illustration 4.
彼は毒杯を仰いだ。
그는 독배를 들이켰다.

The number of sub-entries differs from verb to verb. We computed the average to 3.9 sub-entries by using the 861 Japanese major verbs which we extracted from the IPAL verb dictionary. Considering this numerical value, Japanese-Korean translation systems cannot ignore the polysemy of verbs. The number of sub-entries is not related to the target language but depends on the source language.

2.2 Multi-Translatability of Verbs

When a polysemous verb is translated into several target words, it is called multi-translatable. Multi-translatabilitycauses polysemy, and the numbers of target words differ according to the target language. Fig. 1 show the relation of polysemy and multi-translatability. Table 1 shows the polysemy and the multi-translatability of the Japanese verb "KIRU" for a Japanese-Korean Machine Translation. 11 sub-entries and 10 target words exist in the case of Japanese-Korean machine translation.

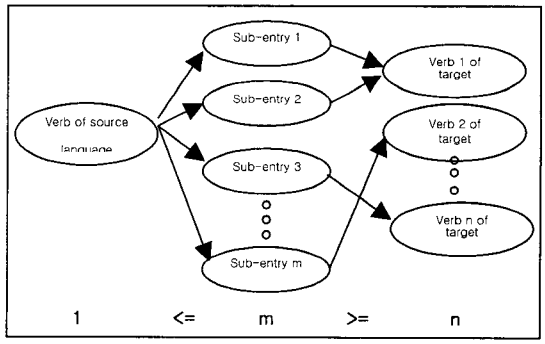


Fig. 1. Relations of Polysemy and Multi-translatability.

3. A JAPANESE-KOREAN MACHINE TRANSLATION SYSTEM CONSIDERING THE MULTI-TRANSLATABILITY OF VERBS

3.1 Overview of the System

For the processing of multi-translatable verbs, the polysemy of those verbs must be considered. After we determine the sub-entry to which a verb belongs, the sub-entry and its corresponding target word can be translated directly. In other words, if the suitable sub-entry of the verb in the source sentence can be chosen among several possibilities, the polysemy and multi-translatability of that verb are resolved. Even if we give the deep semantic analysis of the verb, the

Table 1. Polysemy and Multi-translatability of Japanese Verb "KIRU"

verb of Japanese	no	sub-entry	target word
KIRU (切る)	1	cut a part with a knife from a connected thing, or a part comes off from a true form	자르다
	2	a body is injured with a knife	베다
	3	quit relations with each other	끊다
	4	let stop a function of a machine	끄다
	5	set a limit	정하다
	6	remove water	빼다
	7	break a record	깨다
	8	Aflood breaks an embankment	부수다
	9	shuffle cards	섞다
	10	criticize something severely	비판하다
	11	do the gesture which looks like drawing in the air or water	긋다

sub-entry cannot be chosen perfectly. Moreover, it is very difficult to apply this kind of method to a direct translation method. To improve this, we propose a Japanese-Korean machine translation system which uses Case-Form-Patterns with semantics of uninflected words for processing multi-translatable verbs. This system works in 4 phases.

1) A semantic classification of indeclinable words: Assigning semantic attributes to indeclinable words.
2) The creation of a verb dictionary containing Case-Form-Patterns: Preparing Case-Form-Patterns for the verb dictionary by using the semantic attributes and target words corresponding to the Case-Form-Patterns.
3) Extraction of the semantics and Case-Form-Patterns of the uninflected words: Making patterns which match the case forms by extracting the semantics and Case-Form-Patterns of the uninflected words from the source sentence.
4) Selection algorithm for the most suitable target words: Selecting the appropriate target word by comparison of the Case-Form-Patterns in the verb dictionary with that in the source sentence.

3.2 A Semantic Classification of Uninflected Words

We adopt the classification convention of the IPAL verb dictionary as our semantic classification of uninflected words for the processing of multi-translatable verbs. A classification of semantics of uninflected words is shown in Fig. 2. The semantics of uninflected words are divided into 19 semantic attributes. The success rate of multi-translatable verb processing may be increased by classifying verbs in a more fine-grained fashion. The more the semantic classification of uninflected words becomes fine-grained, the more ambiguity arises, so it is difficult to determine to what extent a detailed description has to be made for appropriate semantic classification. Furthermore, an uninflected word might have one or more semantic attributes.

Illustration 5.
学校に 通う。
학교를 다니다 [Location: LOC]

Illustration 6.
学校を 建てる。
학교를 세우다 [Products: PRO]
Illustration 7.
学校を 休む。
학교를 쉬다 [Action: ACT]
Illustration 8.
学校が 校規を定める。
학교가 교칙을 정하다 [Organization: ORG]

In this example, only a typical attribute is used in our system.

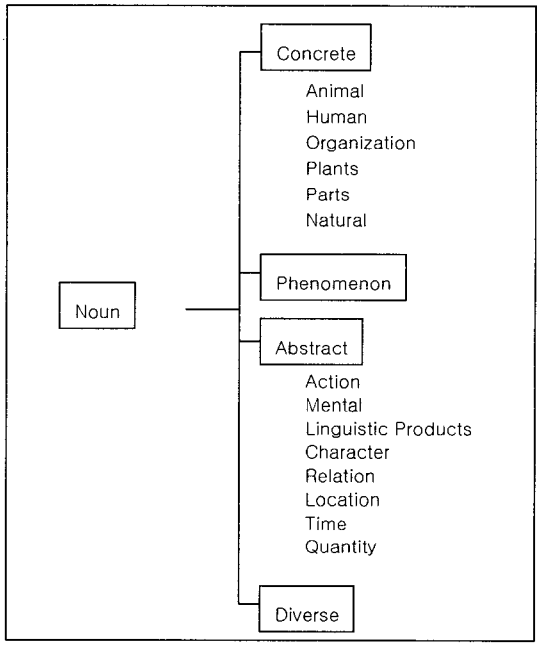


Fig. 2. The Semantic Classification of Uninflected Words.

3.3 Generation a Verb Dictionary Containing Case Forms

The most common sentence structures can be classified into three kinds of major types; the two of them are a "subject + object + verb (SOV)", a "subject + verb + object (SVO)" sentence pattern. Japanese and Korean belong to the "SOV" type and English and German to the "SVO" type. This means, the uninflected words have their positions before the verb which modifies them. The typical case forms are shown in illustration 9.

Illustration 9.
N1 [Ga/Ha] N2 [Ni/De] N3 [Kara/To/..] N4 [Wo] V

Case forms are determined by a case particles. To simplify processing, the system uses the case particles of "Ga, Wo, Ni, Kara, He, To, Yori, De". But the auxiliary particle "Ha" may replace the particle "Ga". It is a case particle following those uninflected words that do not require a case particle. Using the above mentioned case particles, the definition of the most general case forms is shown below. N1 expresses the nominative case, N2 expresses a direction/time/place/tool/etc. case N3 expresses a starting point/parallel/ etc. case and N4 expresses an objective case. Even if the word order changes, a correct pattern can be extracted from any source sentence by using the case particles.

Case-Form-Patterns = Semantics of N1
+ Senamtics of N2
+ Senamtics of N3
+ Senamtics of N4

3.4 The Extraction of the Semantics of Uninflected Words and the Case-Form-Patterns

The best way of processing multi-translatability is to use the meaning of the source sentence. But it is almost impossible to extract the meaning of the source sentence by Japanese-Korean direct machine translation systems. In this paper, we propose to extract the semantics of the uninflected words and then to use the one that fits the Case-Form-Patterns, instead of extracting the semantics from the source sentence.

In analyzing a Japanese sentence, we need to extract its case pattern in order to make its semantics clear. To process the ambiguous verb

in a sentence, we also need to prepare the information of its case pattern in advance, which is used to resolve the ambiguity of the verb. For example, the representative target word for the verb "「きる」" is "자르다", and since it is also ambiguous, its semantic information is included in the ambiguity-processing dictionary, as shown in Table 2.

3.5 A Selection Algorithm for the Most Suitable Target Word

Multi-translatability of verbs can be resolved by comparing two Case-Form-Patterns, the first pattern determines the semantics of the uninflected words that are extracted from the source sentence, and the second pattern is taken from the verb dictionary. We introduce the algorithm that chooses the appropriate target words from a verb dictionary. Algorithm 1 illustrates this process.

The number of the Japanese verbs with more than 11 ambiguities is 33. The case patterns (that is, semantics of nominal expressions) extracted from the example sentences of 「きる」 in the IPAL dictionary are given in Table 3.

Now, let us translate an example sentence with the ambiguous verb dictionary of 「きる」. Let us consider the process of selecting the target words from the third example sentence. 「私はあいつとはやく縁を切りたい」. First, we select "자르다" which is the representative target word of 「きる」, and then check if it is ambiguous. By assigning "Y" as the response to the verb ambiguity checking query, we read in the relevant information from the ambiguous verb dictionary and execute the al-

Table 2. Entry of 「きる」 in the Ambiguity-Processing Dictionary

Japanese Verb	Part Of Speech	No	Meaning 1	Meaning 2	Meaning 3	Meaning 4	Korean
きる	verb	1				PAR	베
		2				REL	끊
		3				NAT	빼
		4				ACT	긋
		5				LOC	꺽
		6				QUA	깨
		7				TIM	정하
		8	NAT			PRO	부수

Table 3. Case Forms Extracted from the Example Sentences of the Verb 「きる」

No	Example Sentences	Extracted Case Patterns			
1	花子はナイフで封筒の口を切った	HUM	PRO	NULL	PRO
2	母はうっかり包丁で指を切ってしまった	HUM	PRO	NULL	PAR
3	私はあいつとはやく縁を切りたい	HUM	NULL	HUM	REL
4	彼女は不意に電話を切った	HUM	NULL	NULL	PRO
5	彼は期間を切って仕事をした	HUM	NULL	NULL	TIM
6	彼女はさっと野菜の水を切った	HUM	NULL	NULL	NAT
7	彼は百メートル競走で十秒を切った	HUM	ACT	NULL	TIM
8	濁流が堤防を切って流れ出した	NAT	NULL	NULL	PRO
9	彼女はトランプを切った	HUM	NULL	NULL	PRO
10	批判家は彼の新作を切って捨てた	HUM	NULL	NULL	PRO
11	神父は手で十字を切った	HUM	PAR	NULL	ACT

```
1. Search a verb from source sentence
  1.1 Search a verb
     1) If the verb is not exist then goto 4.(end)
     2) If the verb have not multi-translatability then
           goto 4.(end)
2. Take out semantics from uninflection words
  2.1 Put semantics from Nmean1 to Nmean4
       matched to Case-Form-Pattern
  2.2 Make Case-Form-Pattern with
            [Nmean1] + [Nmean2]
          + [Nmean3] + [Nmean4]
     1) If Case-Form-Pattern is empty then
           goto 4.(end)
3. Compare to another Case-Form-Pattern of
    the verb dictionary
  3.1 Search a pertinent verb from the verb dictionary
  3.2 Compare semantic pattern with
       Case-Form-Pattern
     1) If two patterns are the same then get target
           word and goto 4.(success end)
     2) skip one point in the verb dictionary
     3) If Eof of the verb dictionary then goto 4.(end)
     4) Goto 3.2
4. End of Process
```

Algorithm 1. Algorithm of Appropriate Target Word Selection.

gorithm for selecting the appropriate target word. The case form pattern extracted from the example sentence is "HUM NULL HUM REL", and we compare it with the case patterns in Table 2. Starting with the comparison of "NUL NUL NUL PAR" in Table 2, the case pattern scoring the highest evaluation weight is selected. In our example, the second case pattern "NUL NUL NUL REL" gets the highest weight, and thus the Korean verb "끊다" is chosen as the target word.

4. EVALUATION

According to the IPAL verb dictionary, 33 verbs have a polysemy of more than 11. We use these 33 verbs as samples. We compare the success rate of translation of multi-translatable verbs with and without multi-translatability processing in 512 sentences, using the above mentioned 33 verbs. The translation of the verb "KIRU" in eleven sentences was successful in eight sentences by selecting the appropriate target word and unsuccessful in three sentences, which makes a success rate of 72.7%. The results of the multi-translatability processing for all 33 verbs are shown in Table 2. Table 4compares the cases of using multi-translatability processing with those of not using it, which deliver a success rate of 86.5% and 65.0% respectively. Thus, the success rate of selection of the appropriate target word has increased by 21.5% in terms of utilizing processing of multi-translatable verbs.

5. CONCLUSIONS

Multi-translatability of words is one of the most difficult problems in Japanese-Korean machine translation systems which use a direct translation method. In this paper, we proposed an algorithm for solving multi-translatability of verbs that uses the correspondence between CASE-Form- Patterns and the semantics extracted from uninflected words in the source sentence. We could show that the success rate for the suitable translation increased from 65.0% to 86.5% by using our multi-translatability processing for 33 verbs in 512 sentences. The reason for the error rate of 13.5% is that the system could not distinguish between the Case-Form-Patterns extracted from the source sentence and from another source sentence. This problem can be solved by classifying the sub-entries for uninflected words in a more fine-grained way. But there is a trade-off: the translation system may become too complex a system caused by this way of classifications.

We obtained the numerical value of the result assuming the probability of the sub-entries being equal. Normally, the frequency of the different sub-entries varies for the translation of verbs in a sentence, so that the probabilities of selecting each target word are not equally distributed. The success rate can be increased by considering the frequency of the choice of the different sub-entries. In our future research, we plan to make our multi-translatability processing system more

Table 4. The Results of a Multi-Translatability Processing

No	verbs	polysemy	Multi-transtability	Before		After	
				success	rate	success	rate
1	出る	32	5	16	50.0	28	87.5
2	掛ける	27	19	8	29.6	18	66.7
3	出す	27	4	18	66.7	27	100.0
4	掛かる	23	11	13	56.5	16	69.6
5	取る	23	14	8	34.8	15	65.2
6	入る	22	3	9	40.9	19	86.3
7	打つ	20	6	14	70.0	17	85.0
8	上がる	19	4	15	78.9	18	94.7
9	入れる	17	1	17	100.0	17	100.0
10	上げる	16	5	10	62.5	12	75.0
11	なる	16	1	16	100.0	16	100.0
12	する	15	5	11	73.3	14	93.3
13	作る	15	2	14	93.3	15	100.0
14	落とす	14	9	6	42.9	11	78.6
15	決める	14	3	12	85.7	12	85.7
16	付く	13	6	8	61.5	10	76.9
17	揃える	13	3	9	69.2	10	76.9
18	送る	13	1	13	100.0	13	100.0
19	見る	12	1	12	100.0	12	100.0
20	当たる	12	4	5	41.7	11	91.6
21	来る	12	1	12	100.0	12	100.0
22	揃う	12	3	7	58.3	9	75.0
23	張る	12	10	3	25.0	11	91.6
24	聞く	12	5	8	66.7	11	91.6
25	持つ	12	4	9	75.0	10	83.8
26	ある	12	1	12	100.0	12	100.0
27	落ちる	11	3	9	81.8	11	91.6
28	思う	11	1	11	100.0	11	100.0
29	切る	11	8	4	36.4	8	72.7
30	付ける	11	7	5	45.5	8	72.7
31	引く	11	8	3	27.3	9	81.8
32	見せる	11	1	11	100.0	11	100.0
33	やる	11	5	5	45.5	9	81.8
Total		512	164	333	65.0	443	86.5

efficiently by using frequency dependent sub-entries.

6. REFERENCES

[1] J. Kim, K. Lee, "Japanese-to-Korean Inflected Word Translation Using Connection Relations of Two Neighboring Words," *Korean Journal of Cognitive Science*, pp. 33-42, 2004.

[2] K. Moon, J. Lee, J. Kim, and G. Yang, "Resolution of Word Sense Ambiguities using Collocation Patterns in Japanese-to-Korean MT System," *Journal of Korea Information Science Society (B), Vol.25, No.8*, pp. 1270-1280, 1998.

[3] J. Kim, J. Lee, and K. Lee, "Generation of Korean Predicates Based on Modality-Feature Ordering and Lexicalizing Table in Japanese-Korean Machine Translation," *Jouof Natural Language Processing, Vol.5, No.2*, pp. 3-24, 1998.

[4] Nomura H., "Language processing and Machine Translation," *Koudansya*, 1991.

[5] Tanaka H., "Introduction to Analysis of Natural Language," *SanGyoDoSyo*, 1989.

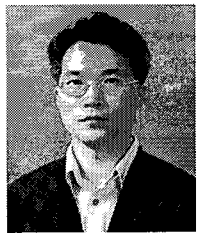
[6] Nomura H., "IPAL(Basic Verbs) Dictionary for Computers(Commented Edition)," *IPA Technical Center*, 1987.

[7] J. Kim, "Japanese-Korean Machine Translation System Using Connection Forms of Neighboring Words," *Journal of Korea Multimedia Society, Vol.7, No.7*, pp. 998-1008, 2004.

[8] J. Kim, K. Moon and J. Lee, "A Processing of Progressive Aspect "te-iru" in Japanese-Korean Machine Translation," *Journal of Korea Information Processing Society (B), Vol.8, No. 6*, pp. 685-692, 2001.

Jung-In Kim

He received his Ph.D degree in computer science from the KEIO university in 1996. After graduation, he spent 2 years as a post-doctoral researcher at POSTECH, participating in the POSCO project for Japanese-to-Korean MT system. Since 1998, he has been working for the Department of Computer Engineering at Tongmyoug University of Information Technology, BUSAN, and currently he is an associate professor. His research interests include machine translation, information retrieval, and knowledge processing. He is a member of the editorial board of Korea Information Processing Society Review, and also a member of the KISS, KMMS, IPSJ, IEICE.



Kang-Hyuk Lee

He obtained his Ph.D. degree at University of Illinois at Urbana-Champaign in 1993, majoring in natural language processing. After receiving the degree, he continued his research on NLP at KAIST as a post-doctoral fellow in 1993~1995. In 1995~1997, he joined KISTI (Korea Institute of Science & Technology Information) and served as a head research scientist of NLP division. Since 1997, he has been teaching in the department of multimedia engineering and the department of game engineering at Tongmyong University of Information Technology. His research areas are natural language parsing and generation, computational morphology and XML.