Representation and Recognition Method for Multi-Word Translation Units in Korean-to-Japanese MT System

Kyonghi Moon
Dept. of Computer Science & Engineering
Pohang Univ. of Science and Technology
San 31 Hyoja-dong Nam-gu, Pohang 790-784
Republic of Korea
khmoon@kle.postech.ac.kr

Jong-Hyeok Lee
Dept. of Computer Science & Engineering
Pohang Univ. of Science and Technology
San 31 Hyoja-dong Nam-gu, Pohang 790-784
Republic of Korea
jhlee@postech.ac.kr

Abstract

Due to grammatical similarities, even a one-to-one mapping between Korean and Japanese words (or morphemes) can usually result in a high quality Korean-to-Japanese machine translation. However, multi-word translation units (MWTU) such as idioms, compound words, etc., need an n-to-m mapping, and their component words often do not appear adjacently, resulting in a discontinuous MWTU. During translation, the MWTU should be treated as one lexical item rather than a phrase. In this paper, we define the types of MWTUs and propose their representation and recognition method depending on their characteristics in Korean-to-Japanese MT system. In an experimental evaluation, the proposed method turned out to be very effective in handling MWTUs, showing an average recognition accuracy of 98.4% and a fast recognition time.

1 Introduction

As a transfer problem in a machine translation (MT), lexical and structural differences exist between source and target languages, which requires 1-n, m-n, or n-1 mapping strategies for machine translation system. For such mapping strategies, we need to treat several (n, or m) words (or morphemes) as a single translation unit. Although some researches (D.Santos,1990; Linden E.,1990; Yoon Sung Hee, 1992; Ha Gyu Lee, 1994; D.Arnold,1994) employ the term "idiom" for these units, we prefer MWTU (Multi-Word Translation Unit) because it is a more general and broader term for MT environment.

Up to now, some research has focused on recognition and transfer of MWTUs, although very little research has been undertaken for Korean-to-Japanese machine translation systems (Seon-Ho Kim,1997). In previous researches, some tended to simplify the problem by treating only special types of MWTUs, while others had some recognition errors and took too much recognition time because they did not restrict the recognition scope (D.Santos,1990; Yoon Sung Hee,1992; Ha Gyu Lee, 1994; Seon-Ho Kim,1997).

For a Korean-to-English MT, Lee and Kim (Ha Gyu Lee,1994) uses only weak restrictions like adjacent information for recognition scope. However, their method needs stronger restrictions to resolve recognition errors and to speed up the process. Although some differences exist depending on which kinds of source and target languages are dealt with, MWTUs in Korean-to-Japanese MT frequently have their component words close together, so that one can predict the location of their separated component words. For this reason, we can enhance the recognition accuracy and time effectively by restricting the recognition scope according to the characteristics of an MWTU rather than taking the whole sentence as the scope.

Moreover, the method by Lee and Kim (Ha Gyu Lee,1994) deals with only surface-level consistency without considering word order because Korean has almost free word order. It is obvious that the method can deal with variable

word-order MWTUs, but some incorrect recognition results are possible when meaning changes according to word order. Because MWTUs to be treated in Korean-to-Japanese MT have an almost fixed word order sequence, their meaning may vary if the word order is changed. In (1), both sentences have the same lexical words (or morphemes), but while the first sentence must be treated as an MWTU, the second, which has the different sequence from the first, does not have the meaning of an MWTU. In (1), the words surrounded with a box are an essential component morpheme for an MWTU.

keu -n ko dachi -eot -da ... . (1)
(big) (nose) (get hurt)
/*(I) had a bitter experience */
<--> *ko -reul dachi -eoseo keu -n il -ida*
(nose) (get hurt) (big)
/* It is serious (that I) got hurt in my nose */

In this paper, to solve the word order problem and thus enhance a recognition accuracy and time for MWTUs, we fix the word order in an MWTU and define the recognition scope of component words according to their characteristics. Based on it, then we propose a representation and recognition method of MWTUs for a Korean-to-Japanese MT system. In the rest of this paper, details will be presented about these proposed ideas, together with some evaluation results. For representing Korean and Japanese expressions, the 1994-SK (ROK Ministry of Education) and the Kunrei Romanization systems are used respectively.

2 Processing of MWTUs

In developing MT systems, we frequently contact with some differences in word spacing, grammar, and so on, between source and target languages. But the method and degree of difficulty of handling them highly depend upon the nature of the source and target language in the MT system. In this paper, we treat the representation and recognition methods of MWTUs according to their characteristics for only a Korean-to-Japanese MT system.

2.1 Types of MWTU

There can be 1-1, 1-m, n-1, and n-m mapping relations of morphemes between source and target language in machine translation. Due to the grammatical similarities of Korean and Japanese, Korean-to-Japanese machine translation systems have been developed under the direct MT strategy, which assumes a 1-1 mapping relation. But a uniform application of this 1-1 mapping relation will easily result in an unnatural translation.

It is not difficult to handle a 1-1 and 1-m mapping relations in Korean-to-Japanese MT system although it uses only direct MT strategy, because it is easy to recognize only one morpheme in source language, Korean. It is also due to the fact that Japanese correspondences have characteristics of non-spacing and continuity, which allows several words to be treated as a single word. In this reason, we need to consider just types with n-1 and n-m mapping relations. Table 1 shows the types of MWTUs to be handled in Korean-to-Japanese MT.

The compound words in Table 1 are the units that must be translated into one Japanese morpheme though they are compound words in Korean. For example, "*wodeu peuroseseo*" is a Korean compound word which consists of two morphemes "*wodeu*" and "*peuroseseo*", but its Japanese equivalent is only one morpheme, "*wapuro*". The Korean word "*yeojju -eo bo [-da]*" is also a compound word, made by 2 lexical morphemes "*yeojju*" and "*bo*" and 1 functional morpheme "*-eo*", but it also corresponds to only one Japanese equivalent morpheme, "*ukaga[-u]*". In these cases, the Korean compound words should be recognized as one unit to be transformed into one Japanese morpheme.

We can classify verbal nouns into 2 types according to their Japanese equivalents. Table 2 shows them. If we define a Korean verbal noun as X and its equivalent in Japanese as X', and another single word in Japanese as Y, we can describe the two types of relations between Korean and Japanese verbal nouns as below. Although the type 1 satisfies 1:1 mapping relation, the type 2 does not. So, for the type2, the verbal noun, X (e.g., "*chuka*") and "*ha[-da]*" need to be recognized as a single unit to be transformed into a Japanese equivalent, Y.

[Table 1] Types of MWTUs to be handled in Korean-to-Japanese MT

MWTU type	Examples	
Type	Korean	Japanese equivalent
1) Compound word	- Compound noun: *wodeu* *peuroseseo* (word) (processor) - Compound verb *yeojju* *-eo* *bo* *[-da]* (ask) (see)	wapuro /* word processor */ *ukaga[-u]* /* ask */
2) Verbal Noun	*chuka* *ha* *[-da]* (congratulation) (do)	*iwa[-u]* /* congratulate */
3) Collocation pattern	*soran* *piu* *[-da]* (noise) (play)	*sawa[-gu]* /* disturb */
4) Modality	*-neun* *geot* *gat* *[-da]* (thing) (equal)	*sou[-da]* /* seem */
5) Idiom	*keu* *-n* *ko* *dachi* *[-da]* (big) (nose) (get hurt)	*hido -i me -ni a [-u]* /* have a bitter experience */
6) Colloquial idiomatic phrase	*cheoeum* *boep* *-get* *-seumnida* (first) (see)	*hazime -masi -te* /* How do you do */
7) Semi-Word	*-reul* *wiha* *-n* (in favor of)	*-no tame -no* /* for */

[Table 2] Types of verbal nouns

	Korean	Japanese
Type 1	X + *ha[-da]*	X' + *suru*
Type 2	X + *ha[-da]*	Y

Collocation patterns are the units that frequently co-occurr in sentences and affect the semantics of each other. There are two kinds of collocation patterns. In one, each component morpheme is translated into different equivalents, such as "*dambae [-reul] piu[-da]*(smoke)" corresponding to "*tabako –o su[-u]*", and in the other, all component morphemes must be translated into one Japanese morpheme with an equivalent meaning, such as "*soran [-eul] piu[-da]*" corresponding to "*sawa[-gu]*". While the morphemes in the former case have a 1-to-1 mapping relation, the morphemes in the latter case have an n-to-1 mapping relation and therefore, must be treated as a single morpheme.

While some modalities consist of only one morpheme like "*-eot*" or "*-da*", there are also some modalities made up of several morphemes like "*-neun geot gat*". Accordingly, the latter must be handled as an MWTU.

An Idiom is a general idiomatic unit defined in a dictionary. Generally, since an idiom does not reflect literal meaning itself, translating their component morphemes individually results in very different meaning. In this case, it must be treated as a single unit.

A colloquial idiomatic phrase is also composed of several morphemes, but it is recognized like a single unit word. For instance, the Korean greeting "*cheoeum boep -get -seumnida*" corresponds to "*hazime –masi -te*" in Japanese. In this case, a 1-to-1 mapping transformation results in an unnatural translation. Therefore, it also should be recognized as MWTUs.

Moreover, MWTUs can be used for groups of words that can give a more natural translation when they are treated as one unit. We will call these groups of words semi-words.

2.2 The Characteristics of MWTUs

To minimize the recognition time and recognition error rate of MWTUs, we need to represent MWTUs according to their characteristics. The following shows the characteristics of MWTUs.

1) Fixed word order

All of the 7 types of MWTUs in Table 1 have a fixed word order sequence, even though Korean and Japanese are known as free word order languages. Expressions such as "*keu -n ko dachi*" and "*-neun geot gat*" must be recognized

as MWTUs, but their meaning may be changed from that of MWTUs if the word order sequence has been changed. This provides a good characteristic for simply representing MWTUs.

2) Extension by insertion of other words

For some kinds of MWTUs, it is possible to insert some grammatical morphemes or other words between their component morphemes of an MWTU. "*-do*" in (2) , "*-reul*" and "*-reul geu -ege*" in (3) are those cases.

ga [*-l*] [*su*] [*it*] *-da* ..... (2)
(go) (means) (is)
/* (I) can go */
➲ *ga* [*-l*] [*su*] ***-do*** [*it*] *-da*
/* (I) can go, too */

[*sinse*] [*ji*] *-da* ..... (3)
(a favor) (owe)
/* be obliged to */
➲ [*sinse*] ***-reul*** [*ji*] *-da*
/* be obliged to */

➲ [*sinse*] ***-reul*** ***geu*** ***-ege*** [*ji*] *-da*
(he)
/* be obliged to him */

According to this feature, the relations between immediately located two component morphemes of MWTUs can be classified as follows:

A. tightly connected : the relation that no morpheme can be inserted between them
B. loosely connected : the relation that some morphemes can be inserted between them.
   B-1. Only particles and endings of a word are allowed to be inserted between them.
   B-2. Any kinds of morphemes can be inserted between them.

[Figure 1] Relations between two adjacent component morphemes of MWTUs

3) Strong cohesion

Although some MWTUs have characteristics of extension by insertion of other words, component morphemes in an MWTU have strong cohesion, not only logically but also physically. This means that the recognition of an MWTU is possible by local comparison of its physical location. But it does not imply that the scope is limited in a simple sentence structure.

4) The predictable recognition scope of MWTUs

It is possible to predict the recognition scope between two adjacent component morphemes of MWTUs, according to the above characteristics. The scope can be predicted as follows for each type of MWTUs shown in Table 1.

Component morphemes of a compound word are contiguous to the next one, so their scopes are predictable.

Both verbal nouns and collocation patterns have the form combined with "Noun" and "Verb", where other words can be inserted between them. But in the case of "***Noun***+Verb+***Verb***", which is the form that another verb is inserted between the noun and verb, its meaning may be different in that of an MWTU. So the scope of the "Verb" can be limited up to the position of the first verb appearing after the "Noun", that is, the position where the POS(part-of-speech) appears.

Component morphemes of a modality have an especially strong cohesion. So at most, one particle is often inserted next to the bound noun. From this, we can predict the next component morpheme apart from pre component at most in distance 2.

Idioms, colloquial idiomatic phrases and semi-words consist of various component morphemes, which results in various scopes for MWTU recognition. The scopes of each component morphemes from pre-component morphemes can be determined by distance 1, distance 2, or infinity. But infinite scope can also be limited by the position which the POS of the component morpheme appears.

2.3 Representation of MWTU

The representation of an MWTU must be considered in order to enhance recognition accuracy and speed up the process. Accordingly, in this paper, we propose representation method of MWTUs according to the characteristics mentioned in section 2.2.

One basic rule for MWTU representation is that an MWTU is composed of only lexical morphemes if possible, that is, grammatical

morphemes such as particles and the endings of a word will be extracted in the representation because of the above characteristics which are freely inserted and omitted. However, grammatical morphemes affecting the meanings of MWTUs must be described.

Next, according to the characteristics described in section 2.2, we need to represent recognition scopes between adjacent component morphemes and POS of each component morpheme for the restriction of recognition scope.

$m_1(POS_1, d_{12})\ m_2(POS_2, d_{23})\ \ldots\ m_i(POS_i, d_{i,i+1})\ \ldots\ m_n(POS_n, d_{n,n+1})$

m_i : i-th component morpheme of an MWTU

POS_i : POS of m_i

$d_{i,i+1}$: maximum distance from m_i to m_{i+1}

[Figure 2] Representation of an MWTU

$d_{i,i+1}$ has 4 kinds of values according to Figure 1. For the case of A, $d_{i,i+1}$ is 1, for the case of B-1, it is 2, for the case of B-2, it is ∞, and then for the last component morpheme, it is always 0 because (n+1)-th component morpheme doesn't exist.

The examples of MWTUs described by above representation are shown in Figure 3.

- *wodeu*(N,1) *proseseo*(N,0) ➲ *wapuro*
  (word) (processor) /* word processor */
- *yeojju*(V,1) *-eo*(mC,1) *bo*(V,0) ➲ *ukaga*
  (ask) (see) /* ask */
- *keu*(ADJ,1) *-n*(mT,1) *ko*(N,2) *dachi*(V,0)
  (big) (nose) (get hurt)
  ➲ *hidoimenia* /* have a bitter experience */
- *-neun*(mT,1) *geot*(ND,2) *gat*(ADJ,0) ➲ *sou*
  (thing) (equal) /* seem */
- *chuka*(N,∞) *ha*(V,0) ➲ *iwa*
  (congratulation) (do) /* congratulation */
- *-reul*(j,1) *wiha*(V,1) *-n*(mT,0) ➲ *notameno*
  (in favor of) /* for */
- *sesang*(N, 2) *muljeong*(N, ∞) *moreu*(V,0)
  (world) (condition) (don't know)
  ➲ *seziniuto* /* be ignorant of the world */
- *jal*(B,1) *meok* (V,1) *-eot*(e,1) *-seumnida*(mT,0)
  (well) (eat)
  ➲ *gotisousamadesita*
  /* I have enjoyed my dinner very much */

[Figure 3] Examples of MWTUs

Each MWTU is entered into the dictionary as an entry word such as the general morphemes as shown in Figure 4. Additionally, for recognition, we made the first component morpheme of the MWTU have an MWTU field, which is composed of MWTUs starting from the entry word. This means that only one access to the dictionary is needed after an MWTU is confirmed. Figure 4 shows the dictionary structure for an MWTU.

(ip * nolli)
;(mouth) (use)
/* speak carelessly */
[Connection info. for Korean]
[Semantic info., Collocation pattern]
[Japanese equivalence, Connection info. for Japanese. ....]

(*ip*)
[Connection info. for Korean ,
MWTU {*ip*(N, ∞) *nolli*(V,0), *ip*(N, ∞) *bareu*(V,0),....}]
[Semantic info., Collocation pattern]
[Japanese equivalence, Connection info. for Japanese. ....]

[Figure 4] Dictionary for an MWTU

2.4 Recognition of MWTU

Some rules are required in order to recognize MWTUs represented like those in section 2.3.

First, the recognition scope of m_{i+1} after recognizing m_i is decided by POS_{i+1} and $d_{i,i+1}$. For restricting the recognition scope maximally while preventing other recognition errors, we formulated recognition scopes of each component morphemes of an MWTU as follows.

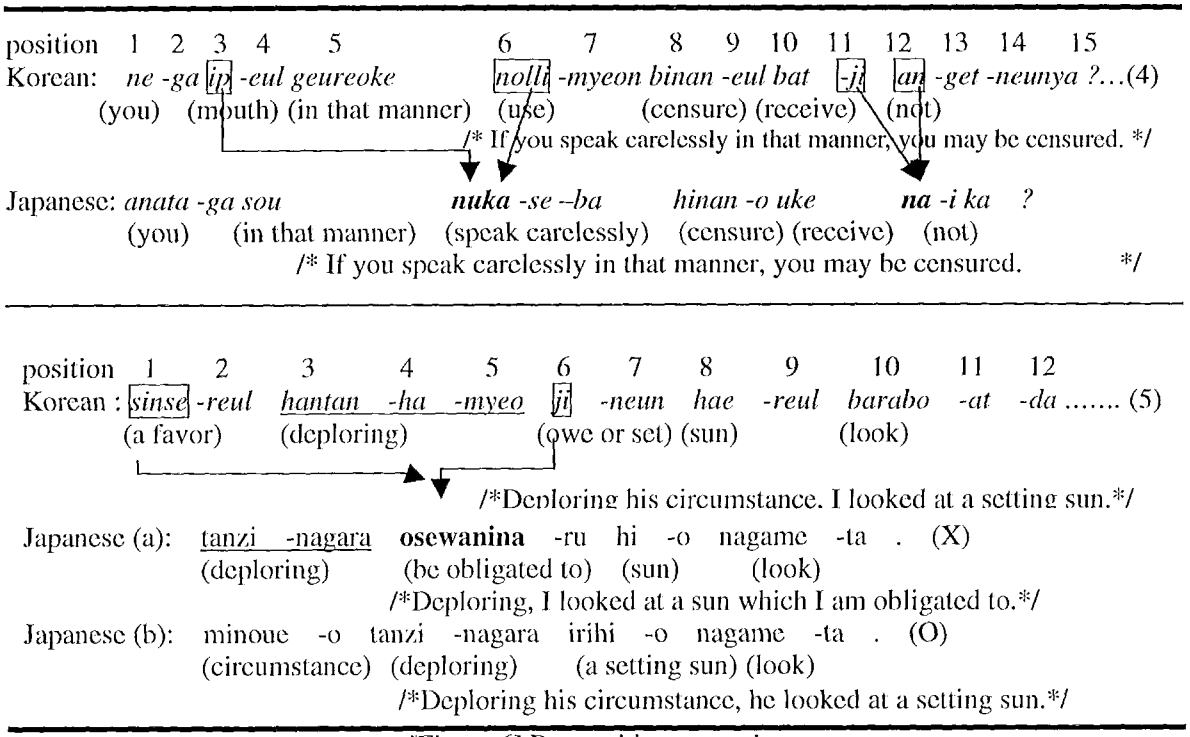
RS(Recognition Scope) = $\min[real\_dist_{i,i+1}, d_{i,i+1}]$

$real\_dist_{i,i+1}$: the distance from m_i to the point that the POS of m_{i+1} appears at first in an input sentence

$d_{i,i+1}$: maximum distance from m_i to m_{i+1}

[Figure 5] Recognition scope

In (4), for an MWTU "*ip*(N,∞) *nolli*(V,0), the recognition scope of "*nolli*" is 3 because $d_{1,2}$ is ∞ and $real\_dist_{1,2}$ is 3, which is from 6-3. For an MWTU, "*-ji*(mC,2) *an*(V,0), the recognition scope of "*an*" is 1 because $d_{1,2}$ is 2 and $real\_dist_{1,2}$ is 1, which is from 12-11. Therefore, we can recognize MWTUs by a small comparison.



[Figure 6] Recognition examples

This Recognition rule can also prohibit some recognition errors generated from unnecessary comparisons. For instance, the recognition scope of "ji" in an MWTU "sinse(N,∞)ji(V,0)" was limited by 2, which is the minimum value between $d_{1,2}(\infty)$ and $real\_dist_{1,2}(3\text{-}1=2)$. So it prohibits errors, such as Japanese (a) in (5), occurring when an MWTU is recognized in whole sentence.

The second rule states that morphemes inserted between the component morphemes of the recognized MWTU must be rearranged in the following manner:

1) If inserted morphemes are lexical morphemes, they are rearranged to the front of the MWTU. "*geureoke*(in that manner)" in (4) is such a case.

2) If they are grammatical morphemes, they are ignored when they directly follow any component of the MWTU, and they are transferred to the front of the MWTU together with the inserted lexical morphemes when they follow any inserted lexical morphemes. In (4), "*-eul*" is the former case. If any grammatical morpheme such as "*-do*" or "*-na*" is attached after "*geureoke*", it will be the latter case.

Third, if a morpheme is the common subset of the two MWTUs, we select the one such that its first component morpheme locates in the pre-position. This rule is used to reduce the recognition time by skipping morphemes which are subsets of the pre-confirmed MWTUs

Fourth, we select the superset of MWTU in case that two or more MWTUs starting from a same morpheme are recognized and one is the superset of the others. For example, let us consider two MWTUs: "*jamsi -man –yo* (wait a moment)" and "*jamsi -man*(for a little while)". If "*jamsi -man -yo*" is recognized, "*jamsi -man*" can also be recognized and "*jamsi -man -yo*" is the superset of "*jamsi -man*". In this case, we select the superset, "*jamsi -man -yo*".

3 Evaluation

To demonstrate the efficiency of our proposed method, we applied it to a Korean-to-Japanese machine translation system (COBALT-K/J), and evaluated its recognition accuracy and recognition time. COBALT-K/J consists of about 150,000 general purpose words and 7,500 MWTUs. For the test corpus, we arbitrary extracted 2,808 sentences from a 10 million word corpus, the KIBS (Korean Information Base System). MWTUs registered in the dictionary appeared 3,647 times in them.

Table 3 shows the evaluation results classified by the types of MWTUs.

[Table 3] Evaluation results on the recognition of MWTUs

Types of MWTUs	Frequ -ency	No. of success	Accur -acy	Avg. No. Of Compari son
compound word	33	32	97.0%	1
verbal noun	918	907	98.8%	1.05
Collocatio n pattern	33	29	87.9%	1.82
modality	1326	1292	97.4%	1.02
idiom	5	5	100%	1.3
colloquial idiomatic phrase	83	83	100%	1.08
semi-word	1249	1242	99.4%	1.01
total	3,647	3,590	98.4%	1.03

In Table 3, idioms, collocation patterns and compound words have a very low frequency while verbal nouns, modalities and semi-words have a relatively high frequency. Nevertheless, 98.4% of the test samples were recognized correctly. In order to recognize an MWTU, it needed only 1.03 comparisons per each component morpheme of the MWTU on the average. This shows the effectiveness and the speed of our proposed method for treating MWTUs in Korean-to-Japanese MT.

Conclusion

In this paper, we classified the different kinds of MWTUs and proposed a representation and recognition method for them in a Korean-to-Japanese MT.

MWTUs in Korean-to-Japanese MT have the characteristics of fixed word order, strong cohesion, predictable scope of its component morphemes, extension by other words, etc. Accordingly, we enhanced accuracy and recognition time by representing and recognizing MWTUs according to their characteristics.

In our experiment, 98.4% of the test samples were recognized correctly, which shows the effectiveness of our proposed method. In future work, we will research in more strict recognition restrictions and plan to extract MWTUs from a corpus automatically.

References

D. Santos(1990), *Lexical gaps and idioms in machine translation*, 13th International Conference of Computational Linguistics. Coling 90, Finland, pp. 330-335.

Linden E., Wessel K. (1990), *Ambiguity resolution and the retrieval of idioms: two approaches*, 13th International Conference of Computational Linguistics. Coling 90, Finland, pp. 245-248.

Yoon Sung Hee (1992), *Idiomatical and Collocational Approach to English-Korean Machine Translation*, Proceedings of ICCPOL '92, pp.56-60.

Ha Gyu Lee, Yung Taek Kim (1994), *Representation and Recognition of Korean Idioms for Machine Translation*, Journal of the Korean Information Science Society, Vol. 21, No. 1, pp.139-149 (written in Korean).

Seon-Ho Kim (1997), *Lexicon-Based Approach to Recognition and Transfer of Multi-Word Translation Units in Korean-Japanese Machine Translation*, MS Thesis, Pohang University of Science and Technology (written in Korean).

D.Arnold, L.Balkan, R. Lee Humphreys, S.Meijer, L.sadler (1994), *Machine Translation*, Blackwell, USA.