Research on Processing a Multiple Nominative Case Construction in Korean-English Machine Translation by Using WordNet

Donghyeok Lee
Korea University,
Institute of Korean Culture
5ga, Anam-dong, Seongbuk-gu
Seoul, Korea
metacog@korea.ac.kr

Janggeun Oh
Korea University,
Institute of Korean Culture
5ga, Anam-dong, Seongbuk-gu
Seoul, Korea
domplatz@korea.ac.kr

Hochol Choe
Korea University,
Institute of Korean Culture
5ga, Anam-dong, Seongbuk-gu
Seoul, Korea
hocherl@korea.ac.kr

Junghye Choi
Korea University,
Institute of Korean Culture
5ga, Anam-dong, Seongbuk-gu
Seoul, Korea
einsiris@chol.com

Abstract

In this paper, we focus on explaining the process of the Korean multiple nominative case construction through using WordNet. Multiple nominative case constructions, in which the nominative case marker *-i/-ga* is allocated not only to subject but to other syntactic functions, generate much confusion and make the Korean-English Machine Translation, a process mainly depending on the case marker with words, the more difficult process. Also, the predicates in the multiple nominative case constructions can sometimes be ambiguous. To overcome those difficulties, we classified the multiple nominative case constructions into 4 types by argument analysis and then processed them in Korean-English Machine Translation by using WordNet.

1 Introduction

In this paper, we focus on explaining the process of the Korean multiple nominative case constructions through using WordNet. A multiple nominative case construction is a sentence in which the nominative case marker *-i/-ga* is allocated not only to subject but to other syntactic functions. The following types can be the examples of multiple nominative case constructions.

(1) a. Type 1: Nae-ga horangi-ga museop-da. (I am afraid of tiger.)

b. Type 2: Kkoch-i jangmi-ga yeppeu-da. (Rose is the most beautiful flower.)

c. Type 3: Kokkiri-ga ko-ga gil-da. (Elephant has a long nose.)

d. Type 4: Naengjanggo-ga meonji-ga man-ta. (Refrigerator has lots of dusts.)

However, it is difficult to translate the Korean multiple nominative case construction into English automatically with the following matters:

Korean language is a Marker Language which depends on the marker in order to define a case. For instance, the word with *-i/-ga* is the subject in a sentence. However, there can be more than two words with the nominative case marker, *-i/-ga*, in one sentence, although subject cannot be more than one in a sentence. This matter should be solved in system of Korean-English Machine Translation.

If we cannot decide the syntactic functions by case marker, then we should use the syntactic/semantic information of predicate. But in the case of the ambiguous word like *yeppeu-da* (pretty), it is hard to decide the syntactic function of the word with *-i/-ga* by simply referring to the syntactic/semantic information of predicate.

(2) a. Kkoch-i jangmi-ga yeppeu-da. (Rose is the most beautiful flower.)

b. Yeong-hui-ga nun-i yeppeu-da. (Yeong-hui has beautiful eyes.)

So, it would be necessary to formulate the meaning of predicate systematically and utilize the semantic relation of the words with nominative case marker, *-i/-ga*, to overcome those problems. Thus, we suggest the way of processing the multiple nominative case constructions using Korean WordNet.

2 Types of the Korean Multiple Nominative Case Constructions

Korean multiple nominative case constructions can be modeled as follows:

Type 1

a. Nuna-ga ki-ga aju keu-da. (My sister is very tall.)

b. Agi-ga bae-ga gopat-da. (The baby was hungry.)

Type 2

a. Nae-ga horangi-ga museop-da. (I am afraid of a tiger.)

b. Agi-ga sarang-i pilyoha-da. (Babies need love.)

c. Nae-ga chingu-ga mip-da. (I hate the friend.)

d. Mul-i eoleum-i doin-da. (Water transforms to ice.)

e. Geu chingu-ga haksaeng-i ani-da. (The friend is not a student.)

Type 3

a. Kkoch-i jangmi-ga yeppeu-da. (Rose is the most beautiful flower.)

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 207–210.

b. Saengseon-i domi-ga masi-ta. (Sea bream is the most delicious fish.)

c. Sagwa-ga neunggeum-i masi-ta. (A crab apple is the most delicious apple.)

Type 4

a. Wonsungi-ga pal-i gil-da. (Monkey has long arms.)

b. Kokkiri-ga ko-ga gil-da. (Elephant has a long nose.)

c. Tokki-ga apbal-i chal-ta. (Rabbit has short forefeet.)

d. Geu yeoja ai-ga nun-i yeppeu-da. (The girl has beautiful eyes.)

f. Bihaegi-ga sokdo-ga ppareu-da. (The speed of plane is high.)

In type 1, the second noun and predicate are in collocation. The predicate of type 2 needs two arguments, while the predicate of type 3 and 4 demands one argument. Type 3 and 4 need one argument: the arguments in type 4 constitute single class, while those in type 3 don't.

(3) a. *Kkotjangmi-ga yeppeu-da. (Rose, the flower, is beautiful.)

b. Wonsungipal-i gil-da. (Arms of the monkey are long.)

3 The Structure of the Korean-English Machine Translation

In Korean-English Machine Translation system, Korean is the source language and English is the target language. The Korean-English Machine Translation system of this article has a structure, shown in Fig. 1.

In the Figure 1, a dictionary and WordNet can participate in the morphological and syntactic analysis and generation process. In the translation process, syntactic analysis follows morphological analysis in dealing with the source language, and then, the generation process of the target language is carried out.

4 Processing of the Multiple Nominative Case Constructions

4.1 Processing type 1: collocation

In type 1, the second noun with *–i/-ga* and predicate are in collocation. For example, in "Agi-ga bae-ga gopa-ta.", "bae-ga gopeu-da" composes a collocation. Collocation is the close combination of the two specific words by convention. The motivation to constitute the collocation is diverse, and as for type 1, it is due to the lexical divergence between the target language and the source language: in Korean, the source language, "bae-ga gopeu-da", corresponding to a single word "hungry" in English, the target language, forms a phrase. So, "bae-ga gopeu-da" can be processed as the collocation. However, in English WordNet (Fellbaum 1998), collocation is not stored in the vocabulary section as the semantic information. For example, although blond hair is the typical example of the English collocation, it is only listed as a gloss in the "blond" section, as shown below:

The adj blond has 1 sense (first 1 from tagged texts)

1. (2) blond, blonde, light-haired – (being or having light colored skin and hair and usually blue or grey eyes; "blond Scandinavians"; "a house full of light-haired children")

However, our Korean-English Machine Translation system stores collocations, like "bae-ga gopeu-da", along with idioms and indicates their meaning information.

Therefore, the collocation of "bae-ga gopeu-da" became one unit like idioms, and behaves as a whole in the sentence. To do so, $[[\text{bae-ga}]_{NP}\ \text{gopeu-da}]_{VP}$ should be identified as a collocation, and then, corrected as $[\text{bae-ga gopeu-da}]_{VP}$ before the syntactic analysis. After the process, although "Agi-ga bae-ga gopeu-da" is still apparently a multiple nominative case construction, *-ga* with "agi" is the only nominative case in the sentence and defines "agi" as the subject.

4.2 Processing type 2: meaning class

Type 2 can be identified by the meaning class of the predicate and some specific words. If the predicate is the psychological verb like "museop-da" or "mip-da", then the sentence can be classified as type 2. In addition to the psychological verbs, some specific predicates like "pilyoha-da", "doi-da", "ani-da" also form sentence of type 2. In the WordNet, meaning domain of a word can be found by using unique beginner. If the predicate of a sentence belongs to [psychol.feature, feeling] or specific predicates like "pilyoha-da", "doi-da", "ani-da", then, the first noun with *–i/-ga* is treated as the subject and the second noun with *–i/-ga*, the object.

4.3 Processing type 3 and 4: semantic relation

The predicates of type 3 and 4 are the state verbs [state], not the collocation or [psychol.feature, feeling]: for example, "jjal-ta", "gil-da", "yeppeu-da", etc. However, they are not confined to one type only, but can be used in both types and, as a result, ambiguous in nature. So, they cannot be processed by simply defining the meaning feature of the predicate like type 2. To discriminate type 3 and 4, the semantic relation of the words with *–i/-ga* should be looked up in the WordNet. In type 3, the words with *–i/-ga* are in hypernym-hyponym relation.

(4) Koch-i jangmi-ga yeppeu-da. (Rose, the flower, is beautiful.)

Sense 1
'jangmi' (a rose)
⇒ rose
⇒ Rosa
⇒ a shrub
⇒ a flower
⇒ a plant
⇒ a living thing
⇒ an object
⇒ an individual

For example, "kkot (flower)" and "jangmi (rose)" are in hypernym-hyponym relation, as shown in (4). By the relation, the first word with *–i/-ga* becomes the topic, and the second word the subject.

In the meantime, the characteristics of type 4 is the holonym-meronym relation between the words with *–i/-ga*. For example, in the sentence "Kokkiri-ga ko-ga gil-da. (Elephant has a long nose.)", animal is the holonym of "ko (nose)" and "kokkiri (elephant)" is the hyponym of animal.

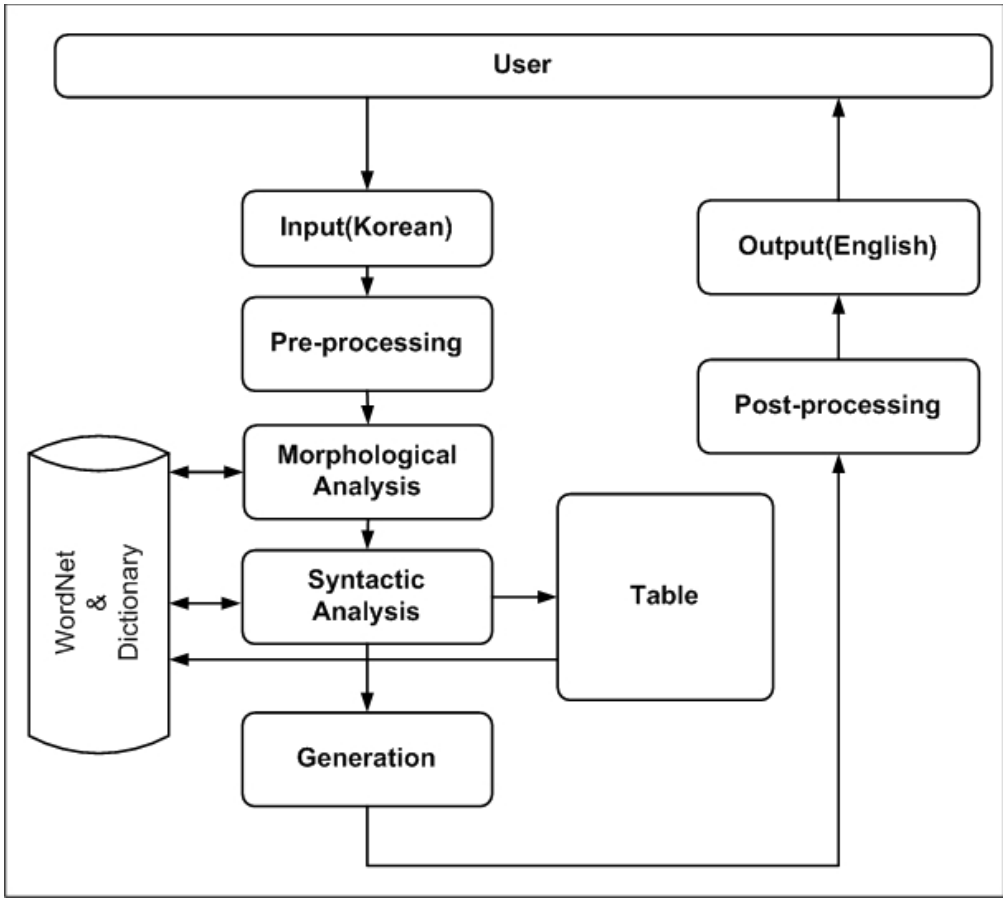

Figure 1: The structure of the Korean-English Machine Translation system.

동사 "배가 고프다"는 1개의 센스를 갖는다.
1. "배가 고프다" – (배 속이 비어 음식을 먹고 싶다; "아침을 굶었는데도 이상하게 배가 고프지 않았다", "배가 고프니 꼼짝도 하기 싫다")

(5) "Kokkiri-ga ko-ga gil-da. (Elephant has a long nose.)"

Sense 1

nose, olfactory organ – (the organ of smell and entrance to the respiratory tract; the prominent part of the face of man or other mammals; "he has a cold in the nose")

PART OF: face, human face – (the front of the human head from the forehead to the chin and ear to ear; "he washed his face"; "I wish I had seen the look on his face when he got the news")

PART OF: head, caput – (the upper part of the human body or the front part of the body in animals; contains the face and brains; "he stuck his head out the window")

PART OF: body, organic structure, physical structure – (the entire structure of an organism (especially an animal or human being); "he felt as if his whole body were on fire")

PART OF: animal, animate being, beast, brute, creature, fauna – (a living organism characterized by voluntary movement)

PART OF: homo, man, human being, human – (any living or extinct member of the family Hominidae characterized by superior intelligence, articulate speech, and erect carriage)

PART OF: upper respiratory tract – (the nose and throat and trachea)

Sense 1

elephant – (five-toed pachyderm)

=> proboscidean, proboscidian – (massive herbivorous mammals having tusks and a long trunk)

=> placental, placental mammal, eutherian, eutherian mammal – (mammals having a placenta; all mammals except monotremes and marsupials)

=> mammal, mammalian – (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)

=> vertebrate, craniate – (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)

=> chordate – (any animal of the phylum Chordata having a notochord or spinal column)

=> animal, animate being, beast, brute, creature, fauna – (a living organism characterized by voluntary movement)

=> organism, being – (a living thing that has (or can develop) the ability to act or function independently)

=> living thing, animate thing – (a living (or once living) entity)

=> object, physical object – (a tangible and visible entity; an entity that can cast a shadow; "it was full of rackets, balls and other objects")

=> physical entity – (an entity that has physical existence)

=> entity – (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

So, the second word with *–i/-ga* becomes the subject of the sentence, and the first one, the modifier.

4.4 The sequence of multiple nominative case constructions processing

By far, we have classified the multiple nominative case constructions into 4 types and explained the method to process them in the Korean-English Machine Translation system. To realize the multiple nominative case constructions classification method explained in this investigation, the sequence of process of each type should be determined. The Korean-English Machine Translation system of this investigation follows the processing sequence listed below:

Morphological analysis → Syntactic analysis (NP analysis → Clause definition → Predicate identification → Argument analysis → Tense, aspect, modal analysis → Negation analysis → Adjunct analysis) → Generation

Among the 4 types of the multiple nominative case constructions, type 1 is the first to be processed, because collocation should be processed between morphological analysis and syntactic analysis. The next is type 2, which is characterized by the meaning class of the predicate or some specific words as their predicates. Finally, type 3 and 4 would be processed, because there is no specific feature in their predicates.

5 Conclusion

In this work, the method for processing multiple nominative case constructions using WordNet in the Korean-English Machine Translation has been explained. Multiple nominative case constructions, in which the nominative case marker *-i/-ga* is allocated not only to subject but to other syntactic functions, generate much confusion and make the Korean-English Machine Translation, a process mainly depending on the case marker with words, the more difficult process. Also, the predicates in the multiple nominative case constructions can sometimes be ambiguous.

To overcome those difficulties, we first classified the multiple nominative case constructions into 4 types by argument analysis. Type 1 contains a collocation of the second noun with *–i/ga* and the predicate, and type 2 has predicates that need 2 arguments. Type 3 and 4 need one argument: the arguments in type 4 constitute single class, while those in type 3 don't. And then, we showed the processing method of each type in the actual Korean-English Machine Translation system. The processing method of type 1 is similar to those of general collocation. Type 2 could successfully be processed by considering the meaning class and some specific words. Type 3 and 4 could be processed by the semantic relation between the 2 words with *–i/-ga*: if they are in hypernym or hyponym relation, they can be classified into type 3, and if in holonym or meronym relation, then into type 4. Also, in the linear sequential process of the Korean-English Machine Translation system, type 1 is the first to be processed, and then type 2, and finally type 3 and 4 are treated.

References

Fellbaum, C. 1998. *WordNet*. Cambridge and London: The MIT Press.

Lee, Dong-hyeok. 2004. "Processing Methods of *Issta* Constructions for Korean-English Machine Translation." *Language Research* 40-1, 253–277.

Yu, Hyeong-seon. 1999. "A Study on Argument Structure of the multiple nominative case construction." *Case and Case Marker of Korean Language*, 717–731.