EXTRACTION OF SIMPLE SENTENCES FROM MIXED SENTENCES FOR BUILDING KOREAN CASE FRAMES*

Dan-Hee Yang, Ik-Hwan Lee**, Mansuk Song**

* Department of Computer Science, ** Department of English,
Yonsei University, Seoul 120-749, Korea.
E-mail: {dhyang, mssong}@december.yonsei.ac.kr, ihlee@bubble.yonsei.ac.kr

ABSTRACT

A large number of simple sentences are needed to construct practical Case frames automatically. Until now, most studies have assumed that there are already extensive training data (especially here simple sentences) and linguistic information for their work. However, this is not true at least of Korean. Furthermore, Korean syntactic structures are significantly different from those of English. So, this paper first of all, compares Korean with English in relation to extracting simple sentences from mixed ones. Second, we suggest fundamental and detailed principles. For convenience and practicality, however, we deliberately exclude some linguistic phenomena. Finally, we attempt to develop a reliable algorithm to extract simple sentences with the ultimate goal of building Case frames.

1. INTRODUCTION

In NLP, the Case frames of a language are very important for a correct syntactic and semantic analysis of the language. The term *Case frames* is originated from the Case grammar of Fillmore. However, the term may currently refer to the syntactic part of a lexical entry in grammars such as HPSG, LFG, and the like, or in other place it sometimes includes the semantic part, too.

To confirm the common deficiency of the recent approaches to the acquisition of Case frames, let us review some of the related works. Chae-Deug Park studied on learning the Case frames of English without any consideration of preparing sufficient simple sentences as training data [7]. Chae-Kwan Song tried to automatically extract sentence patterns and the information of semantic attributes from the corpus manually tagged with parts-of-speech [8]. Tanaka used sentences analyzed by means of a full parser as training data [9].

Most of such researches so far have used the corpus tagged either by hand or by a parser. Experimental studies in a small scale manage to prepare training data manually. Such a manual arrangement, however, always results in a barrier to doing practical researches on the entire language. On the other hand, the use of any full parser, as it is without any additional processing, brings about a contradiction because the training data are obtained from the unreliable parser. Notice that Case frames are the very information for a further reliable syntactic analysis. Furthermore, currently available parsers for Korean are not even as good as those for English.

The training data needed to construct the Case frames for the entire Korean verbs are nothing but a large amount of simple sentences. Unfortunately, however ordinary sentences are not in the form of simple sentences but mixed ones. If we might extract only originally simple sentences from a given corpus, hence a corpus of tremendous size would be required, which is not expected to be available in the near future [11]. This implies that we have to extract simple sentences from mixed ones. Concerning this, assuming that the Case structures and argument structures for all Korean verbs are available, Kwang-Jin Kim extracted simple sentences from embedded ones, though the ultimate goal of his study was a machine

* This research was funded by the Ministry of Information and Communication of Korea under contract 98-86.

translation [4]. However, such linguistic information might not be available for practical NLP until sufficient simple sentences are available.

Summing up, we do not rely on such unrealistic assumptions in this study. We just use the output of NM-KTS morphological analyzer, whose rate of accuracy is 96% and probability of guessing unregistered words is 0.75, and hence it is comparatively reliable. As already discussed, the use of a full parser results in consistent reflection of the internal algorithm and Case frames of the parser. Therefore, this study proposes a partial parsing algorithm. Also, to increase the accuracy of analysis we exclude sentences that might bring about fallacy in actual analyses. The approach of partial parsing enables a large amount of incompletely (but not incorrect) analyzed sentences to be used for machine learning. This implies that we may adopt a quite different approach from full parsers.

2. PROBLEMS IN EXTRACTING SIMPLE SENTENCES

In comparison with English, Korean requires a variety of considerations in developing morphological analyzers. So does it in working on extracting simple sentences from mixed sentences. Therefore, we should, first of all, know the information status in current Korean dictionaries and the linguistic features of Korean.

A sentence of Korean may be compound, complex or mixed. A compound sentence consists of two or more coordinate clauses. A complex sentence consists of one main clause and one subordinate clause, which is a constituent of the main clause. The subordinate clause has the adverbial, adnominal or nominal functions. By combining compound and complex sentences we get a mixed sentence, which is structurally complex and compound. Adnouns are a non-inflectional word class that modifies the following nominals. Verbs and adjectives can function as adnominals when used in construction with adnominalizer endings. Adnoun clauses are made up of verbal or adjectival sentences with an adnominalizing ending (*-(n)un*, *-ten*, or *-(u)l*).

Korean dictionaries clearly show whether a verb is transitive or intransitive, but there is no information about its complements. In other words, they do not include the information on argument structures. Notice that arguments in this study are the participants (but not necessarily *minimally*) involved in the activity or state expressed by the predicate. In contrast, most of English dictionaries such as Hornby English Dictionary have the information in the form of parts of verb patterns. Currently, manual work for Korean is being done merely on restricted predicates [1,2].

Korean is relatively free to omit and invert the constituents of a sentence, which is a salient syntactic trait compared with English and thus makes it difficult to pick out the governing domain of each predicate. Furthermore, Korean is an S-O-V language, which means that a verb (or adjective) is a sentence final constituent. Other constituents are relatively free in positional ordering. There is of course a preferred ordering of constituents when no one particular constituent is highlighted for focus or contrast in a discourse. This makes the connection of predicates complicated when the omission and inversion are involved together. To elucidate this phenomenon, let us examine the following example. Hereafter, TOP stands for topic marker, OM for objective one, SM for subjective one, QU for quotative one, and ADNZ for adnominalizer.

(1) *Ku-nun chayk-ul kunye-lo-pwute pata kalochayssta.*
He-TOP book-OM her-from received intercepted
'He received a book from her and intercepted it.'

The first verb *pata* 'received' takes *chaky-ul* 'a book' (for referring to a constituent or argument in a Korean sentence, the combination of the nominal and its Case particle will be described like this) and *kunye-lo-pwute* 'from her' as its arguments while the second verb *kalochayssta* 'intercepted' takes only *chaky-ul* 'a book' as its argument (see [3] for more detail). In English, the object 'it' cannot generally be omitted. In contrast, Korean frequently omits the object as in the above case.

Dong-Young Lee proposed an algorithm of deciding which nominal functions as the subject in a sentence, which has multiple embedded clauses and which contains scrambling or pro-drop phenomenon [5]. The algorithm is summarized as follows: If a predicate is found, its subject is the noun to which the subjective

Case particle is attached, satisfying the following three conditions: (a) It is on the left side of the predicate. (b) It is closest to the predicate. (c) It was never before corresponded with other predicates. However, if condition (c) cannot be satisfied, the predicate shares the same subject with the predicate closest to the left of it.

To prove the algorithm, the study considered sentence (2) containing only quotative clauses. What is significantly problematic in NLP, however, mainly related to sentences containing relative adnoun clauses rather than to sentences such as (2). Examples (3)-(4) illustrate the flaw of the algorithm.

(2) *Chelhuy-ka Swunok-i Yengsu-ka ku yenghwa-lul poassta-ko malhayssessta-ko sayngkakhanta.*
-SM -SM -SM the movie-OM seen.had-QU said-QU thinks
'Chelhuy thinks that Swunok said that Yengsu had seen the movie.'

(3) *Ayin-i ttena sulphehanu-n ku-lul poassta.*
sweetheart-SM left sad.feeling-ADNZ him-OM saw
'We saw him feeling sad because his sweetheart had left him.'

(4) *Chelswu-ka Yenghey-lul ttaylinu-n kes-ul poassta.*
-SM -OM hit-ADNZ fact-OM saw
'Someone saw that Chelswu hit Yenghey.' or 'Chelswu saw that someone hit Yenghey.'

With the algorithm applied to sentence (3), the subject of *ttena* 'left', *sulphehanun* 'feeling sad', and *poassta* 'saw' is construed as *Ayin-i* 'a sweetheart'. Also, a sentence containing a noun clause as in (4) may have two readings. If the subject of *ttayli-nun* 'hit' were construed as Chelswu, the subject of *poassta* 'saw' would be omitted. On the contrary, if the subject of *poassta* 'saw' is construed as Chelswu, the subject of *ttayli-nun* 'hit' would be omitted. As we see in these counterexamples, the present analysis fails to account for the following two facts: One, a subject can appear on the right side of its predicate if the sentence contains an adnoun clause. The other, a subject may usually be omitted in Korean as shown below.

(5) *Moluntayyo.*
not.know.say
'He/She said that he/she did not know.'

In ordinary English sentences, only the elements of an utterance that may be recovered readily from the syntactic structure can be omitted. In Korean, however, there is a zero anaphor as in (5), which is an unmarked discourse reference, whereas the pronominal anaphor is an unmarked one in English.

For inversion or scrambling, let's consider sentence (6). With only this syntactic structure it is hard to say whether *hakkyo-lo* 'to school' is in the governing domain of *poassta* 'saw' without referring to any semantic information or context. In English, the governing domain is made clear by using the pronoun 'it' when the sentence has a long subject or object phrase, thus making inversion necessary. There is also a case of the inversion for emphasis, although it is not a frequent linguistic phenomenon.

(6) *Wuli-nun hakkyo-lo Chelswu-ka kanu-n kes-ul poassta.*
We-TOP school-to -SM going-ADNZ fact-OM saw
'We saw that Chelswu was going to school.'

Peculiarly, there are no relative pronouns and relative adverbs in Korean. In case of English, the word order itself marks Case (i.e., implicit Case marking) whereas pronouns including relative pronouns explicitly represent Case by declension (i.e., explicit Case marking). Even when the relative pronouns such as 'that' and 'what' are used, the following word can tell whether the relative pronoun is the subject or object of the sentence. The relative adverbs also indicate that the antecedent is an adverb (or complement) implying place, time, cause, and the like.

Astonishingly enough, however, the opposite is true in Korean. The Case particle attached to a nominal explicitly marks the nominal as the subject, object, or complement of the sentence. In a complex sentence containing an adnoun clause, however, the Case particle attached to the postcedent (in contrast to the term 'antecedent' of the English) of the adnoun clause disappears, only with the Case particle for the superordinate clauses (or main clauses) left. The Case particle is essential to reconstructing the clause into a complete simple sentence. To give an example,

(7) a. *Ku-nun ku-ka kongpwu-lul haysste-n hakkyo-lo tomangchyessta.*
He-TOP he-SM study-OM did-ADNZ school-to ran.away
'He ran away to the school at which he had studied.'
b. *Ku-nun hakkyo-lo tomangchyessta.*
He-TOP school-to ran.away
'He ran away to the school.'
c. *Ku-ka kongpwu-lul hakkyo-eyse hayssta.*
He-SM study-OM school-at did
'He had studied at the school.'

(7a) consists of a superordinate clause (7b) and a subordinate clause (7c). In (7a), the Case particle *-eyse* 'at' of the phrase *hakkyo-eyse* 'at the school' in the subordinate clause (7c) disappears, while the Case particle *-lo* 'to' of the phrase *hakkyo-lo* 'to school' in the superordinate clause (7b) survives. This implies that it is not possible to recover the Case particle *eyse* 'at' for the noun *hakkyo* 'school' in the subordinate clause only with the syntactic structure. Never does this phenomenon occur in English. What is required in this case is to pick out the Case through a semantic analysis. Also it is not always easy to decide whether the postcedent is a complement mainly because of the inherent absence of relative adverbs. This means that the adnominalizers of Korean adnoun clauses behave somewhat similar to both the English relative pronouns 'that, which, and who' and relative adverbs 'when, where, how'.

Finally, there may be double nominatives (subjects) or accusatives (objects) within a sentence. When an adnoun clause should be separated from the main clause, this phenomenon becomes problematic. For instance, when there are two objective Case particles within a sentence, the syntactic structure cannot give us any clue on whether each of them belongs to a different predicate or they constitute double objectives for the same predicate.

3. HOW TO APPROACH

To simplify the problems and enhance the accuracy in partial parsing, this study sets up the following fundamental and detailed principles, and puts asides some linguistic phenomena that need to be further clarified in the field of linguistics.

3.1 Fundamental Principles

① The processing priority, which reflects the degree of difficulty in partial parsing, is based on Table 1. Necessary information is taken from the corpus only by processing complete sentences (i.e., sentences with priority 1-3) in order of priority. After that, for the priority 4, if a certain event occurs over a given frequency, we credit the information. This means that we take a probabilistic approach.

Table 1. Processing priority

Priority	Type of sentences
1	Simple sentence
2	Compound sentences (conjunctive and disjunctive coordination)
3	Complex sentences containing noun clauses, predicative clause, adverb clauses, quotative clauses, long adnoun clause
4	Complex sentences containing short adnoun clauses

② Long adnoun clauses take *-(n)un* as an adnominalizing ending and modify the head noun, which takes no part in it and is appositional to the whole clause. Short adnoun clauses take *-l* or *-n* as an adnominalizing ending. There are two types of adnominal modification, depending on the structural relation between the short adnoun clause and the head noun: One, the head noun is a constituent of the adnominal clause. The other, the head noun is not its constituent. To distinguish these two types,

the former is called relative adnoun clauses, and the latter a type of appositive clause.

③ Comparing Korean with English, we interpret Korean grammatical phenomena within the paradigm of the English grammar. This is useful for NLP.

④ We exclude all the pragmatic features that are not inherent features of predicates such as occurring double nominatives or accusatives within a sentence.

⑤ We treat only the constituents to which subjective, objective, and adverbial Case particles are attached.

3.2 Detailed Principles

① Adnoun clauses are either relative clauses or appositive clauses. A relative clause is an incomplete sentence. Therefore, a subordinate clause should be considered after the predicate of a superordinate clause takes its obligatory arguments.

② There are many, so called, phrasal particles, including *-ey tayhayse* 'about', *-ey kwanhay* 'concerning' and *-lul wihay* 'for (the sake of)' as in (8). Such English preposition equivalents are treated as a single particle.

(8) *Ku-nun cengiy-lul wihay ssawessta.*
He-TOP justice-OM for fought
'He fought for justice.'

③ The information on a complement requirement is obtained from the processing outcome from priority 1 to priority 3 of Table 1. The sentences having predicates whose a complement requirement is not clear are excluded in this step.

④ The sentences containing appositive clauses like (9) are treated as predicative clauses.

(9) a. *Cohu-n cem-un ku-ka kongpwu-lul cal hanta-nun kesita.*
good-ADNZ what-TOP he-TOP study-OM well do-ADNZ fact
'What is good is he does well in school.'
b. *Sasil-un ku-ka sikyey-lul ilhepelyessta.*
fact-TOP he-TOP watch-OM lost.has
'In fact, he has lost his watch.'

⑤ In case of an object inverted in a complete or incomplete sentence, it is possible to restore the inversion according as the predicate is intransitive or not. But in case of an inverted complement, it is hard to tell whether the complement belongs to superordinate or subordinate clauses. In this case, it can be decided on the basis of the behaviors of the other sentences containing the predicate.

⑥ If a single adjective, intransitive verb, or a noun plus the adnoun form of a predicative Case particle is used as a premodifier (i.e., like *alymtawun* 'beautiful' in (10a), *yehaynghanun* 'travelling' in (10b), and *hakca-in* 'which was a scholar' in (10c)), we do not treat it as an adnoun clause because these simple adnoun clauses are not important for the purpose of this study.

(10) a. *Wuli-nun alumtawu-n kkoch-ul cohahanta.*
We-TOP beatuiful-ADNZ flower-OM like
'We like beautiful flowers.'
b. *Wuli-nun yehaynghanu-n salam-ul poassta.*
We-TOP travelling-ADNZ man-OM saw
'We saw a travelling man.'
c. *Hakca-in Socrates-nun pwulhaynghayssta.*
scholar-ADNZ -TOP unhappy.was
'Socrates, which was a scholar, was unhappy.'

⑦ If two nouns are combined by *wa / kwa* 'with *or* and' as in (11), the preceding noun and its Case particle are eliminated.

(11) a. *Chelsu-wa Yenghuy-nun kongpwuhanta.*
-and -TOP studying
'Chelsu and Yenghuy are studying.'
b. *Chelsu-wa Yenghuy-ka ssawessta.*
-with -SM fought
'Chelsu fought with Yenghuy.'

⑧ As in (12), the phrase *kunye-uy* 'her' in which the possessive Case particle '*-uy*' occurs is excluded because it is not an argument of the predicate.

(12) a. *Na-nun kunye-uy son-ul capassta.*
I-TOP she-POSS hand-OM took
'I took her by the hand.'

3.3 Outside of This Study

We remove all the constituents to which no Case particle is attached from a sentence except a predicate. For instance, in (13), *kwiyepkey* 'pretty' and *kippese* 'for joy' are removed from the sentence for a further processing even if they are virtual arguments, for they are adverbials without any Case particle. As in (14), the sentences containing multiple predicates occurring in succession are excluded because it is difficult to pick out the governing domain of each only in terms of the syntactic structures. Notice that a Korean adjective needs no copula or linking verb to make a sentence well formed. The adjective can function as a predicate by itself.

(13) a. *Kunye-nun kwiyepkey sayngkyessta.*
pretty looks
'She looks pretty.'
b. *Ku-nun kippese nalttwiessta.*
joy-for jumped
'He jumped for joy.'
(14) a. *Ku-nun entek-ul neme kako issta.*
hill-OM over go being
'He is going over a hill.'
b. *Kukes-un talla pwuthe sseke pelyessta.*'
stick bad went
'It stuck and went bad.'

The phenomena and approaches mentioned so far do not cover all linguistic phenomena of Korean. In fact, we deliberately disregarded minor or exceptional phenomena because they do not frequently occur in a real corpus and thus do little affect the amount of training data that we can obtain from a given corpus.

4. ALGORITHM FOR EXTRACTING SIMPLE SENTENCES

In partial parsing, ambiguity occurs mostly in the sentences containing relative adnoun clauses. Therefore, we focus on those types of sentences. In this study, incomplete verbs refer to verbs that take complements. Notice that this study considers only the constituents to which adverbial Case particles are attached as a complement. In Table 2 and 3, a superordinate clause is indicated by S_0 and its predicate P_0;; a subordinate clause S_1 and its predicate P_1. When S_1 is an adnoun clause, the postcedent of the clause is referred to as *M*. The searching orientation *'forward'* refers to a scan S from left to right. *'backward'* is the reverse orientation.

To begin with, we analyze sentences in the corpus morphologically by the morphological analyzer. The following shows the general form after a compound sentence is morphologically analyzed. Here, *N* refers to a nominal plus a Case particle.

$$S = N_1\ N_2\ N_3\ N_4\ P_1\ M\ N_5\ N_6\ P_0$$

Table 2. Case processing algorithm

```
1  Case processing(Input: verb of a sentence)
2  {
3    if (verb of a sentence == P0) { search start = N1; search end = N4;
4        search orientation = forward; }
5    else { search end = search start; search start = N4;
6       search orientation = backward; }
7    if ((verb of a sentence == transitive verb) and (objective was not found))
8       Case searching(objective);
9    else if ((verb of a sentence == incomplete verb) and (adverbial was not found))
10      Case searching(adverbial);
11   else if (subjective was not found)   Case searching(subjective);
12 }
```

Then, the splitting results of a sentence can be described as follows:

$S_0 = N_1\ N_2\ M\ N_4\ N_6\ P_0$
$S_1 = N_3\ M\ P_1$: in case of a relative adnoun clause
$S_1 = N_3\ P_1$: otherwise

Table 3. Case searching algorithm

```
1  Case searching(Input: Case type)
2  {
3    for (from search start to search end toward search orientation) {
4      mark which sentence it belongs to;
5      Case = Searched Case;
6      if (Case == Case type) {
7         search start = the location which it is found;
8         return(OK);
9      }
10   }
11   if ((sentence type == short adnoun clause) and (M is used == NO))
12     Take the M as the Case;
13     MOE is used = YES;
14   }
15   else if (Case type != objective)
16     return(ERROR);
17   else return(OK);
18 }
```

For simplicity and understandability, we simply explain the algorithm with the exemplar (15) by using the general form. However, notice that our algorithm can adequately treat all the types of sentences mentioned so far as well as sentences (3)-(4) given early as counterexamples.

(15) *Chelswu-ka kuli-n phwungkyenghwa-ka cenlamhoy-eyse thuksen-ulo ppophyessta.*
-SM drawn-ADNZ landscape-SM exhibition-in Special.choice-to was.selected
'The landscape that Chulsu had drawn was selected to be Special choice in an exhibition.'

The result of the morphological analysis of (15) by NM-KTS is:

'Chulsu-ka [subjective] *kulin* [P_1] *pwungkyenghwa-ka* [M/subjective] *cenlamhoy-eyse* [adverbial Case] *thuksen-ulo* [adverbial Case] *ppophyessta* [P_0].'

To begin with, mark that *M*, N_5, and N_6 between P_1 and P_0 belong to S_0. The Case processing algorithm of Table 2 will be first applied to P_0. Next, P_1. By marking each position, all *'pwungkyenghwa-ka* [ME/subjective/S_0] *cenlamhoy-eyse* [adverbial Case/S_0] *thuksen-ulo* [adverbial Case/S_0]' get to belong to S_0. Here, we do not need to find an objective Case because P_0 is an intransitive verb. When we already obtain the information that P_0 is an incomplete verb as the result of analyzing the sentences of the processing priority 1-3 in Table 1 (the first fundamental principle), we do not have to try to find an adverbial Case because it has already found. The subjective Case also has already found because *M* here takes a subjective Case particle.

For P_1, we try to find an objective Case particle in the direction of 'backward', but we cannot find it. Since *M* is not yet used by S_1, we can assume that *M* takes the objective Case. As a result, we get *'phwungkyenghwa* [objective Case/ M/S_1]'. We do not need to try to find an adverbial Case because P_1 is a complete verb (refer to the first fundamental principle). Finally, we find the subjective Case for P_1 in the direction of 'backward'. The result is *'Chulsu-ka* [subjective Case/S_0]'.

5. CONCLUSION AND FUTURE WORK

A large volume of simple sentences is a valuable resource in NLP. The collection of simple sentences that results from this study is critical to constructing argument structures and Case structures automatically. In addition, it can be used for building training data for a computer to pick out the thematic roles of arguments within a sentence. Also, in measuring the word similarity for words clustering, the rate of accuracy can be significantly enhanced because the distance between words can be calculated within simple sentences.

This study did not assume that currently non-existent information and knowledge exist. In other words, we set up the realistic experimental resources. Then we tried to construct the information necessary to develop Case frames extensively. However, the algorithm presented here has passed through a simple test. This means that an extensive test and modification have been left for future work.

6. REFERENCES

[1] Hong, Chae-Seong et al. "The Lexicon of Verbal Syntax in the Modern Korean Language," Dusan Dong-A Press, 1996.

[2] Kang, Eun-Kug, "A Study on Korean Sentence Pattern," Seokwang Academic Data Press, 1993.

[3] Kang, Hyeon-Hwa, "A Study on the Overlapping Structure of Verb Linking Constructions," Ph.D. Dissertation, Department of Korean Language & Literature. Yonsei University, 1995.

[4] Kim, Kwang-Jin et al. "Implementation of the System Dividing Simple Sentences from Embedded Sentence in Korean," In Proceedings of Hangul and Korean Language Information Processing (HKIP), 1994.

[5] Lee, Dong-Young, "A Computational Search for a Verb and its Corresponding Subject in the Korean Sentence Containing Embedded Clauses," In Proceedings of the Pacific Rim International Conference on AI., Vol. 2, pp. 219-225, 1992.

[6] Manning, "Automatic Acquisition of a Large Subcategorization Dictionary from Corpora," In Proceedings of ACL, 1992.

[7] Park, Chae-Deug, "Incremental Probabilistic Learning of Schema and Case Role Assignment," Ph.D. Dissertation, Department of Computer Science, Korea Advanced Institute of Science and Technology, 1993.

[8] Song, Chae-Kwan, Seong-Ung Hong, and Chan-Kon Park, "A Study on the Sentence Pattern of the Korean Language for Machine Translation," In Proceedings of HKIP, 1996.

[9] Tanaka, Hideki, "Verbal Case Frame Acquisition from a Bilingual Corpus: Gradual Knowledge Acquisition," In Proceedings of COLING, 1994.

[10] Yang, Dan-Hee and Mansuk Song, "Extraction of the Training Data for building Case Frames from a Corpus," In Proceedings of HKIP, 1998.

[11] Yang, Dan-Hee and Mansuk Song, "Machine Learning and Corpus Building of the Korean Language," In Proceedings of the Spring Conference of the Korea Information Science Society, 1998.