The Grammatical Function Analysis between Korean Adnoun Clause and Noun Phrase by Using Support Vector Machines

Songwook Lee
Dept. of Computer Science,
Sogang University
1 Sinsu-dong, Mapo-gu
Seoul, Korea 121-742
gospelo@nlprep.sogang.ac.kr

Tae-Yeoub Jang
Dept. of English,
Hankuk University of Foreign Studies
270, Imun-dong,
Dongdaemun-gu, Seoul,
Korea 130-791
tae@hufs.ac.kr

Jungyun Seo
Dept. of Computer Science,
Sogang University
1 Sinsu-dong, Mapo-gu
Seoul, Korea 121-742
seojy@ccs.sogang.ac.kr

Abstract

This study aims to improve the performance of identifying grammatical functions between an adnoun clause and a noun phrase in Korean. The key task is to determine the relation between the two constituents in terms of such functional categories as subject, object, adverbial, and appositive. The problem is mainly caused by the fact that functional morphemes, which are considered to be crucial for identifying the relation, are frequently omitted in the noun phrases. To tackle this problem, we propose to employ the Support Vector Machines(SVM) in determining the grammatical functions. Through an experiment with a tagged corpus for training SVMs, the proposed model is found to be useful.

1 Introduction

Many structural ambiguities in Korean sentences are one of the major problems in Korean syntactic analyses. Most of those ambiguities can be classified into either of two categories known as "noun phrase (NP) attachment problem" and "verb phrase (VP) attachment problem". The NP attachment problem refers to finding the VP which is the head of an NP. On the other hand, the VP attachment problem refers to finding the VP which is the head of a VP.

In resolving the NP attachment problem, functional morphemes play an important role as they are the crucial elements in characterizing the grammatical function between an NP and its related VP. However, the problem is that there are many NPs that do not have such functional morphemes explicitly attached to each of them. This omission makes it difficult to identify the relation between constituents and subsequently to solve the NP attachment problem. Moreover, most Korean sentences are complex sentences, which also makes the problem more complicated.

In this research, we make an attempt to solve this problem. The focus is on the analysis of the grammatical function between an NP and an embedded adnoun clause with a functional morpheme omitted.

We adopt Support Vector Machines(SVM) as the device by which a given adnoun clause is analyzed as one of three relative functions (subject, object, or adverbial) or an appositive. Later in this paper (section 3), a brief description of SVM will be given.

2 Korean Adnoun Clauses and their analysis problems

Adnoun clauses are very frequent in Korean sentences. In a corpus, for example, they appear as often as 18,264 times in 11,932 sentences (see section 4, for details). It means that effective analyses of adnoun clauses will directly lead to improved performance of lexical, morphological and syntactic processing by machine.

In order to indicate the difficulties of the adnoun clause analysis, we need to have some basic knowledge on the structure of Korean adnoun clause formation. Thus, we will briefly illustrate the types of Korean adnoun clauses. Then, what makes the analysis tricky will be made clear.

2.1 Two types of adnoun clauses

There are two types of adnoun clauses in Korean : *relative adnoun clause* and *appositive adnoun clause*. The former is a more general form of adnoun clause and its formation can be exemplified as follows :

1.a Igeos-eun(this) geu-ga(he) sseu-n(wrote) chaeg-ida(book-is).
(This is the book which he wrote.)

1.b Igeos-eun(this) chaeg-ida(book-is).
(This is a book.)

1.c Geu-ga(he) chaeg-eul(book) sseoss-da(wrote).
(He worte the book.)

1.a is a complex sentence composed of two simple sentences *1.b* and *1.c* in terms of adnoun clause formation. The functional morpheme '*eul*', which represents the object relation between '*chaeg*' and '*sseoss-da*' in *1.c*, does not appear in *1.a* but 'chaeg' is the functional *object* of '*sseu-n*' in *1.a*. This adnoun clause is called a relative adnoun clause whose complement moves to the NP modified by the adnoun clause and the NP modified by a relative adnoun clause is called a head NP. In *1.a* '*geu-ga sseun*' is a relative adnoun clause and 'chaeg' is its head noun (or NP).

Let us consider another example of an adnoun clause.

2. Geu-ga(he) jeongjigha-n(be honest) sasil-eun(fact) modeun(every) saram-i(body) an-da(know).
(Everybody knows the fact that he is honest.)

The adnoun clause in *2* is a complete sentence which has all necessary syntactic constituents in itself. This type of adnoun clause is called an appositive adnoun clause. And the head NP modified by the appositive adnoun clause is called a complement noun (Lee, 1986; Chang 1995). In *2*, '*geu-ga jeongjig-han*' is an appositive adnoun clause and 'sasil' is a complement noun. Generally, such words as "*iyu*(reason), *gyeong-u*(case), *jangmyeon*(scene), *il(work)*, *cheoji*(condition), *anghwang*(situation), *saggeon*(happening), *naemsae*(smell), *somun*(rumor) *and geos*(thing)" are typical examples of the complement noun (Chang, 1995; Lee, 1986).

2.2 The problems

The first problem we are faced with when analyzing grammatical functions of Korean adnoun clauses is obviously the disappearance of the functional morphemes which carry important information, as shown in the previous subsection (2.1).

Apart from the morpheme-ommission problem, there is another reason for the difficulty. As it is directly related to a language particular syntactic characteristic of Korean, we need first to understand a unique procedure of Korean relativization. Unlike English, in which relative pronouns (e.g., *who, whom, whose, which* and *that*) are used for relativization and they themselves bear crucial information for identifying grammatical function of the head noun in relative clauses (see example 1.a, in section 1), there is no such relative pronouns in Korean. Instead, an adnominal verb ending is attached to the verb stem and plays a grammatical role of modifying its head noun. However, the problem is that these verb ending morphemes do not provide any information about the grammatical function associated with the relevant head noun.

Take 3.a-c for examples.

3.a *Sigdang-eseo(restaurant) bab-eul(rice) meog-eun(ate) geu(he).*
(He who ate a rice in a restaurant.)

3.b *Sigdang-eseo geu-ga meog-eun bab.*
(the rice which he ate in a restaurant.)

3.c *Geu-ga bab-eul meog-eun sigdang.*
(the restaurant in which he ate a rice.)

Despite all three sentences above have the same adnominal ending '*eun*', the grammatical function of each relative noun is different. The grammatical function of the head noun in 3.a is subject, in 3.b, object and in 3.c, adverbial.

The word order gives little information because Korean is a partly free word-order language and some complements of a verb may be frequently omitted. For example, in sentence 4, the verb of relative clause '*sigdang-eseo meog-eun(who ate in the restaurant or which one ate in the restaurant)*' have two omitted complements which are subject and object. So '*bab*' can be identified as either of subject or object in the relative clause.

4. *Sigdang-eseo(restaurant) meog-eun(ate) bab-eul(rice) na-neun(I) boass-da(saw).*
(I saw the rice which (one) ate in a restaurant.)

Korean appositive adnoun clauses have the same syntactic structure of relative adnoun clauses as in example 2 in section 2.

Yoon et al. (1997) classified adnoun clauses into relative adnoun clauses and appositive adnoun clauses based on a complement noun dictionary which was manually constructed, and then tries to find the grammatical function of a relative noun using lexical co-occurrence information. But as shown in example 5, a complement noun can be used as a relative noun, so Yoon et al. (1997)'s method using the dictionary has some limits.

5. Geu-ga(he) balgyeonha-n(discover) sasil-eul(truth) mal-haess-da(talk).
(He talked about the truth which he discovered.)

Li et al. (1998) described a method using conceptual co-occurrence patterns and syntactic role distribution of relative nouns. Linguistic information is extracted from corpus and thesaurus. However, he did not take into account appositive adnoun clauses but only considered relative adnoun clauses.

Lee et al. (2001) classified adnoun clauses into appositive clauses and one of relative clauses. He proposed a stochastic method based on a maximum likelihood estimation and adopted the backed-off model in estimating the probability $P(r|v,e,n)$ to handle sparse data problem (the symbols r, v, e and n represent the grammatical relation, the verb of the adnoun clause, the adnominal verb ending, and the head noun modified by an adnoun clause, respectively). The backed-off model handles unknown words effectively but it may not be used with all the backed-off stages in real field problems where higher accuracy is needed.

3 Support Vector Machines

The technique of Support Vector Machines(SVM) is a learning approach for solving two-class pattern recognition problems introduced by Vapnik (1995). It is based on the Structural Risk Minimization principle for which error-bound analysis has been theoretically motivated (Vapnik, 1995). The problem is to find a decision surface that separates the data points in two classes optimally. A decision surface by SVM for linearly separable space is a hyperplane H : $y = \boldsymbol{w} \cdot \boldsymbol{x} - b = 0$ and two hyperplanes parallel to it and with equal distances to it,

$$H_1 : y = \boldsymbol{w} \cdot \boldsymbol{x} - b = +1,$$
$$H_2 : y = \boldsymbol{w} \cdot \boldsymbol{x} - b = -1,$$

with the condition that there are no data points between H_1 and H_2, and the distance between H_1 and H_2 is maximized.

We want to maximize the distance between H_1 and H_2. So there will be some positive examples on H_1 and some negative examples on H_2. These examples are called support vectors because they only participate in the definition of the separating hyperplane, and other examples can be removed and/or moved around as long as they do not cross the planes H_1 and H_2. In order to maximize the distance, we should minimize $||\mathbf{w}||$ with the condition that there are no data points between H_1 and H_2,

$$\boldsymbol{w} \cdot \boldsymbol{x} - b \geq +1 \text{ for } y_i = +1,$$
$$\boldsymbol{w} \cdot \boldsymbol{x} - b \leq -1 \text{ for } y_i = -1.$$

The SVM problem is to find such **w** and b that satisfy the above constraints. It can be solved using quadratic programming techniques(Vapnik, 1995). The algorithms for solving linearly separable cases can be extended so that they can solve linearly non-separable cases as well by either introducing soft margin hyperplanes, or by mapping the original data vectors to a higher dimensional space where the new features contain interaction terms of the original features, and the data points in the new space become linearly separable (Vapnik, 1995). We use

SVM^{light}[1] system for our experiment (Joachimes, 1998).

SVM performance is governed by the features. We use the verb of each adnoun clause, the adnominal verb ending and the head noun of the noun phrase. To reflect context of sentence, we use the previous noun phrase, which is located right before the verb, and its functional morpheme. The previous noun phrase is the surface level word list not the previous argument for the verb in adnoun clause. Part of speech(POS) tags of all lexical item are also used as feature. For example, in sentence '*Igeos-eun geu-ga sseu-n chaeg-ida.*', '*geu*' is a previos noun pharse feature, '*ga*' is its functional morpheme feature, '*sseu*' is a verb feature, '*n*' is a verb ending feature, '*chaeg*' is a head noun feature and all POS tags of lexical items are features.

Because we found that the kernel of SVM does not strongly affect the performance of our problem through many experiments, we concluded that our problem is linearly separable. Thus we will use the linear kernel only.

As the SVMs is a binary class classifier, we construct four classifiers, one for each class. Each classifier constructs a hyperplane between one class and other classes. We select the classifier which has the maximal distance from the margin for each test data point.

4 Experimental Results

We use the tree tagged corpus of Korean Information Base which is annotated as a form of phrase structured tree (Lee, 1996). It consists of 11,932 sentences, which corresponds to 145,630 eojeols. Eojeol is a syntactic unit composed of one lexical morpheme with multiple functional morphemes optionally attached to it. We extract the verb of an adnoun clause and the noun phrase which is modified by the adnoun clause. We regard an eojeol consisting of a main verb and auxiliary-verbs as a single main-verb eojeol. In case of a complex verb, we only take into account the first part of it. Every verb which has adnominal morphemes and the head word of a noun phrase which is modified by adnoun clause, were extracted. Because Korean is head-fiinal language, we regard the last noun of a noun phrase as the head word of the noun phrase.

The total number of extracted pairs of verb and noun is 18,264. The grammatical function of each pair is manually tagged. To experiment, the data was subdivided into a learning data set from 10,739 sentences and a test data set from 1,193 sentences. We use 16,413 training data points and 1,851 test data points in all experiments.

Table 1 shows an accuracy at each of the grammatical categories between an adnoun clause and a noun phrase with SVMs, compared with the backed-off method which is proposed by (Lee, 2001).

Table 1. the acuracy of SVM and Backed-off model at each of the grammatical categories

	subj	obj	adv	app	total
SVM	84.4	62.9	92.0	97.5	88.7
SVM with context feature	88.8	75.6	89.6	96.1	90.8
Backed-off	86.2	42.0	62.0	91.7	83.5
proportion in the training data(%)	52.8	4.5	6.7	36.0	100

It should be noted that SVM outperforms Backed-off model in Table 1. By using context information we acquire an improvement of overall 2.1%.

Table 2 represents the accuracies of the proposed model compared with the Li's model. The category 'appositive' is not taken into account for fair comparison. It should be noted that Li et al. (1998)'s results are drawn from most frequent 100 verbs while ours, from 4,684 verbs all of which are in the training corpus.

Table 2. the accuracy of SVM without considering appositive clauses

	subj	obj	adv	total
SVM with context feature	94.1	87.8	85.7	93.3
Li et al. (1998)	90	92	89.2	90.4

[1] The SVMlight system is available at http://ais.gmd.de/~thorsten/svm_light/.

It is shown that our proposed model shows the better overall result in determining the grammatical function between an adnoun clause and its modifying head noun.

Most errors are caued by lack of lexical information. Actually, lexical information in 19% of the test data has not occurred in the training data. The other errors are caused by the characteristics that some verbs in adnoun clauses can have dual subjects which we did not consider in the problem. Take 6 for an example.

6. Nun-i(eyes) <u>*keu-n*</u>*(be big) Cheolsu*
(Cheolsu who has big eyes)

In example 6, the context NP is '*nun*' and its functional word is '*i*' which may represent that it is subject of '*keu-da*', thus system may wrongly determine that '*Cheolsu*' is not a subject of '*keu-da*' because the subject of 'keu-da' has been made with '*nun*'. However, both '*Cheolsu*' and '*nun*' are the subjects of 'keu-da'.

5 Conclusion and Future works

Adnoun clause is a typical complex sentence structure of Korean. There are various types of grammatical relations between an adnoun clause and its relevant noun phrase. Unlike in between general content words and modifying clauses where their grammatical relations can be easily extrated in terms of various grammatical characteristics by the functional morphemes, the functional morphemes are omitted in a noun phrase when it is modified by an adnoun clause. This omission makes it difficult to characterize their grammatical relation.

In this paper, we used SVM to take care of this problem and analyze the relation between noun phrase and adnoun clause. We reflected context information by using the previous word of the verb in adnoun clauses as feature. Context information helped the grammatical function analysis between adnoun clause and the head noun. The SVM can also handle the sparse data problem as the backed-off model does. We acquired overall accuracy of 90.8%, which is obviously an improvement from the previous works.

In the future, we plan to compare with other machine learning methods and to enhance our system by using a publicly available Korean thesaurus to increases general accuracy. More data needs to be collected for further performance improvement. We will also work on utilizing the proposed model in some partial parsing problem.

References

Chang, Suk-Jin, 1995. *Information-based Korean Grammar*, Hanshin Publishing Co.

Yoon, J., 1997. Syntactic Analysis for Korean Sentences Using Lexical Association Based on Co-occurrence Relation, Ph.D. Dissertation, Yonsei University.

Katz, S., 1987. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recogniser. *IEEE Transactions on Acoustics, Speech, and Signal processing*, Vol. ASSP-35, No. 3.

Lee, Ik-Sop, Hong-Pin Im, 1986, *Korean Grammar Theory*, Hagyeonsa.

Lee, Kong Joo, Jae-Hoon Kim, Key-Sun Choi, and Gil Chang Kim. 1996, Korean syntactic tagset for building a tree annotated corpus. *Korean Journal of Cognitive Science*, 7(4):7-24.

Lee, Songwook, Tae-Yeoub Jang, Jungyun Seo. 2001, The Grammatical Function Analysis between Adnoun Clause and Noun Phrase in Korean, In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, pp709-713.

Li, Hui-Feng, Jong-Hyeok Lee, Geunbae Lee, 1998. Identifying Syntactic Role of Antecedent in Korean Relative Clause Using Corpus and Thesaurus Information. In *Proceeding of COLING-ACL*, pp.756-762.

Vapnik, Vladimir N. 1995, *The Nature of Statistical Learning Theory*. Springer, New York.

Joachims, Thorsten. 1998, Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *European Conference on Machine Learning*, pp. 137-142.