Classification of Relative Clauses Using Easily Obtainable Features

MI-YOUNG KIM
School of Computer Science and Engineering, Sungshin Women’s University, Dongseon-dong 3-ga, Seongbuk-gu, Seoul, 136-742, Korea
miykim@sungshin.ac.kr

The detection of a gap in relative clauses is essential in syntactic and semantic analysis of natural language processing. However, it is difficult to recognize whether a relative clause has a gap or it is gapless. Previous work related to relative clauses has been focusing mostly on theoretic linguistics without practically automatic classification, or classifying relative clauses using deep-level knowledge only available for a specific language. So, this paper proposes automatic classification method of a relative clause — whether it has a gap or gapless —, using easily obtainable features from any language. Features are extracted from the lexical forms and POS-tags in a relative clause, its headnoun, and contexts around a relative clause. Based on Support Vector Machines learning algorithm, our proposed method outperformed the baseline system by 25.11 percent. We also analyze the contribution rate of each simple feature to the classification, and the effect of contexts around a relative clause on the classification performance.

Keywords: Relative clause; Machine learning; SVM; Classification; Gap; Korean language processing.

1. Introduction

1.1. Why is the detection of relative clause types needed

In natural language processing, to construct the syntactic structure of an input sentence is essential. One crucial task of the syntactic analyzer concerns the efficient recovery of syntactic functions and their correct interpretation. Because Korean is an agglutinative language, one word consists of ‘content word + function word’. So, we can recognize the syntactic function of a word by its function word. However, in relative clause constructions, the syntactic relation between a relative clause and its following headnoun is hidden. In other words,
the function word of a headnoun does not indicate the syntactic function related to the preceding relative clause.

Let me show one example sentence. In Figure 1, ‘geunyeo-ga kkaetteuri-n’ is a relative clause, and its headnoun is the following word ‘changmun-eul’. In the word ‘geunyeo-ga’, ‘-ga’ is a function word which indicates that ‘geunyeo-ga’ functions as a subject in the clause that ‘geunyeo-ga’ belongs to. ‘Changmun-eul’ — the headnoun of the relative clause — has the function word ‘eul’ which indicates that ‘changmun-eul’ functions as an object of the clause that ‘changmun-eul’ is included in. The function word indicates the syntactic function of the clause that it belongs to. The headnoun of a relative clause does not belong to the relative clause, but it belongs to another following clause. So, the syntactic function of the headnoun to the relative clause is hidden. Therefore, we must identify the hidden relation between a relative clause and its headnoun.

Figure 2 shows the semantic analysis result of the example sentence in Figure 1. In the surface sentence, ‘kkaetteuri-n’ is the main predicate of a relative clause headnoun na-neun  geunyeo-ga   kkaetteuri-n      changmun-eul   surihae-ss-da.

Figure 1. Example sentence of a ‘gap relative clause’.

Figure 2. Semantic analysis graph of Figure 1 sentence.
Classification of Relative Clauses Using Easily Obtainable Features

A relative clause, 'kkaetteuri-n', requires its headnoun — 'changmun-eul' — as a theme. For a theme node of 'kkaetteuri-n', the semantic graph made a node 'Trace' which indicates the already appearing node 'changmun-eul'. To figure out the theme relation between a relative clause and its headnoun, we must first recognize the relative clause has a gap.

However, all relative clauses do not have a gap. Figure 4 is the semantic analysis result of the sentence in Figure 3. In Figure 4, the relative clause does not require its headnoun as its complement/adjunct. The headnoun is in opposition with the preceding relative clause. In other words, the relative clause is gapless. So, in the semantic graph, there is no trace node between 'saenggak-eul' and 'al-ass-da-neun'. As shown in Figure 2 and Figure 4, according to the type of relative clause, the analysis result is different. So, we first need to identify the type of a relative clause: whether it has a gap or is gapless.

![Figure 3. Example sentence of a 'gapless relative clause'.](image)

![Figure 4. Semantic analysis of a Figure 2 sentence.](image)
1.2. Types of relative clauses

Relative clauses can be classified into two types. One is that the headnoun of a relative clause functions as a complement/adjunct of the relative clause. We call this type of relative clause ‘gap relative clause’. When processing a gap relative clause, the syntactic/semantic structure must indicate the syntactic/semantic relation between the gap relative clause and its headnoun. The other type is that a relative clause does not require its headnoun as its complement/adjunct. We call this type of relative clause ‘gapless relative clause’.

Let us show the following 4 types of relative clauses.

(1) complement-gap relative clause
   *na-neun geunyeo-ga kkaetteuri-n changmun-eul gochyeo-ss-da*
   (I) (she) (broke) (window) (fixed)
   (I fixed the window she broke.)

(2) adjunct-gap relative clause
   *geu-ga namu-leul jalla-ss-deon kal-eul iyongha-e ra*
   (he) (tree) (cut) (knife) (use)
   (Use the knife which he cut trees with.)

(3) gapless relative clause (apposition)
   *geu-ga geu munje-leul pul geos-ira-neun saenggak-eun mal-ara*
   (he) (the) (problem) (solve) (thought) (do not)
   (do not think that he will solve the problem)

(4) gapless relative clause (lapse of time)
   *geu-ga tteona-n jikhu-e geunyeo-neun ul-eoss-da*
   (he) (left) (next) (she) (cried)
   (She cried after he left.)

In (1), the relative clause ‘geunyeo-ga kkaetteuri-n’ requires its headnoun ‘changmun-eul’ as its object(complement). We call this type of relative clause ‘complement-gap relative clause’. In (2), the relative clause ‘geu-ga namu-leul jalla-ss-deon’ requires its headnoun ‘kal-eul’ as its adverbial(adjunct). We call this type of relative clause ‘adjunct-gap relative clause’. We call types of (1) and (2) as ‘gap relative clause’. In (3), the relative clause ‘geu-ga geu munje-leul pul geos-ira-neun’ does not require its headnoun ‘saenggak-eun’ as its complement/adjunct. The headnoun is in apposition with the preceding relative clause. In (4), the headnoun ‘jikhu-e’ only indicates the lapse of time of the relative clause. We call this type of relative clause such as (3), (4) ‘gapless relative clause’. Our purpose is to classify a relative clause into one of these two types of relative clauses.
Previous work about relative clauses has mainly focused on the linguistically theoretic researches, or classified relative clauses using deep-level knowledge only available for a specific language. The practical classification method of relative clauses in natural language processing has not been deeply researched. So, this paper proposes the method to automatically classify the type of relative clause: whether a relative clause has a gap or is gapless. We experiment with the classification performance using Support Vector Machines. To apply SVM algorithm, we simply get features obtainable from any language, not requiring deep-level knowledge.

We analyze which features are effective or non-effective. We also analyze the change of the performance by adding context features around a relative clause.

This paper is organized as follows. Section 2 presents previous work about relative clauses. In Section 3, the features for machine learning will be described. In Section 4, some experimental results will show that the proposed machine learning-based method is effective in classifying relative clauses. Finally the conclusion is given in Section 5.

2. Previous Work

Several researches about relative clauses have been performed for the Japanese language. Yoshida et al. [14] presented three experiments which show that Japanese speakers are able to use cues from numeral classifiers to anticipate an upcoming relative clause structure, and hence avoid the processing difficulty normally associated with relative clauses. If a classifier mismatches with the following noun, it helps the speakers to predict upcoming relative clauses. This focuses on linguistic study about speaking Japanese, but it is not an important issue to predict ahead the existence of relative clauses in Korean language processing. Moreover, the appearance of a classifier together with a relative clause seldom occurs. Ishizuka [6] researched the subject and object gap of relative clauses in Japanese. The experiments about the reading time of participants showed that object-gap relative clauses are more complex than subject-gap relative clauses in Japanese. Citko [4] researched the situation that the head undergoes raising from the relative clause internal position in a syntactic tree. He insisted that this head promotion could be explained by Deletion Under Identity principle.

Lin et al. [11] researched on Chinese possessor relative clauses (PRC). Chinese PRCs are prenominal, and a parser is not aware of any gaps until reaching a relativizer. When the parser reaches the relativizer, it incorporates the prior region as part of a relative clause. This is different from typical Chinese
RCs that are postnominal. This paper suggests that the gap-searching mechanism (top-down, structure-based parsing mechanism), and subject-gap preference are universal for both prenominal and postnominal relative clauses.


From the practical point of view, Baldwin et al. [1] automatically classified Japanese relative clause constructions for computational processing. He used various, and deep-level knowledge, case frame from Goi-Taikei pattern-based valency dictionary, verb classes, inflectional analysis information, basic noun semantics, and trigger patterns. He employed C4.5 decision trees program for the classification of types of relative clauses. However, much time is needed to construct a similar knowledge for Korean, and we do not have open knowledge resources that can be obtained easily.

As mentioned before, many researches about relative clauses have been performed in the linguistic point of view. However, practically applicable classification methods of relative clauses have not been researched deeply. In the next section, we explain the automatic classification method of relative clauses using easily obtained features.

### 3. Easily Obtainable Features

In Korean, the head word of a relative clause is a predicate, and the type of predicate of a relative clause is always ‘adnominal’. The headnoun of relative clause is usually located next to the adnominal predicate. The type of relative clause is determined according to the relations between the adnominal predicate of a relative clause and its headnoun. Gap relative clause means the headnoun functions as a complement/adjunct of the adnominal predicate. Gapless relative clause means the adnominal predicate does not require the headnoun as its complement/adjunct. So, we extract features from these two words, — adnominal predicate and headnoun.

As mentioned before, a Korean word consists of ‘content word + function word’. A lot of function words in Korean exist. Several function words can be used together in one Korean word. In other words, the Korean word is ‘content
word + several function words’. As a feature, we use the content word information of an adnominal predicate. Among the function words, we use only the first function word as a feature, not the last function word. Because the last function word of an adnominal predicate is always ‘adnominal ending’, it does not give information. In the same way, from a headnoun, we also use the content word and the first function word information as features. The reason that the last function word is not used in a headnoun is as follow: the last function word of a headnoun indicates the syntactic function of the following predicate. So, it is not related to the preceding pronominal predicate. As shown in Table 1, we extract two kinds of features from each of the content word and the first function word. One is surface form, and the other is part-of-speech tag information.

When a relative clause has a gap, its headnoun usually functions as a subject/object of the relative clause. According to the ‘one argument per clause’ principle, we assume the following: if an adnominal predicate already has a subject/object in a sentence, then its headnoun cannot function as its subject/object again. Then, the relative clause would be gapless. So, we detect the information whether a subject/object already exists of an adnominal predicate in a sentence. Because Korean is an SOV language, the subject/object of an adnominal predicate is located preceding the adnominal predicate. We detect the subject/object of an adnominal predicate among the words preceding the adnominal predicate. We restrict the left boundary of the subject/object position of an adnominal predicate as the nearest predicate preceding the adnominal predicate, and the right boundary as the adnominal predicate.

In summary, we obtain the surface form and POS-tag information from the content word and the first function word of an adnominal predicate and its headnoun. As context information, the subject/object existence information of an adnominal predicate is used. As described in Table 1, 10 kinds of features are used for machine learning, and these features are easily obtainable from any language without deep-level knowledge. The 8 features from an adnominal predicate and its headnoun are called ‘basic features’, and the 9th and 10th features from contexts around a relative clause are called ‘context features’. High dimensional data is constructed because a number of Korean surface forms and POS-tags are used, and all features are binary.

As a machine learning algorithm, we employ Support Vector Machines. SVM has been used in many NLP applications [8, 9, 13]. SVMs have good characteristics to cope with the data sparseness problem and achieve high generalization even with training data of a very high dimension. As an SVM program, SVMlight [7] was used.
4. Experimental Evaluation

4.1. Performance analysis

For experiments, KIBS96 (Korea Information Base System) corpus about history (05, 006, 007, 008) and health (001, 002, 005, 006, 007) were used. The dataset consists of 40,379 sentences, and 43,977 relative clauses. We experimented with SVM machine learning algorithm using 10-fold cross validation.

Experiments were performed focusing on the following three things.

1. Using basic features vs. using basic + context features
2. Performance change when each kind of feature is removed
3. Baseline performance vs. our performance

In the experiments, we obtained the following results.

1. The precision using basic features outperformed that of basic + context features by 1.19 percent (see Table 2).
2. The most significant feature is the surface form of the content word in a headnoun. However, the POS-tag of the first function word in a headnoun makes the performance worse (see Table 3).
3. SVM-based method outperformed the baseline system by 23.92~25.11 percent (see Table 2).

We constructed a baseline system to compare its performance with our performance. The baseline system classifies the type of relative clause into the major class–gap relative clause. As shown in Table 2, our proposed method performed better than the baseline system by 23.92~25.11 percent. We tried to compare our performance with those of other systems, such as Baldwin [1]’s.

---

Table 1. Linguistic feature types used for learning.

<table>
<thead>
<tr>
<th>Basic Features</th>
<th>Adnominal predicate</th>
<th>Headnoun</th>
<th>Context Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>2nd</td>
<td>3rd</td>
<td>4th</td>
</tr>
<tr>
<td>1st</td>
<td>2nd</td>
<td>3rd</td>
<td>4th</td>
</tr>
<tr>
<td>Surface form of a content word</td>
<td>POS-tag of a content word</td>
<td>Surface form of the first function word</td>
<td>POS-tag of the first function word</td>
</tr>
<tr>
<td>5th</td>
<td>6th</td>
<td>7th</td>
<td>8th</td>
</tr>
<tr>
<td>Surface form of a content word</td>
<td>POS-tag of a content word</td>
<td>Surface form of the first function word</td>
<td>POS-tag of the first function word</td>
</tr>
<tr>
<td>10th</td>
<td>11th</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Existence of Object of an adnominal predicate</td>
<td>Existence of Subject of an adnominal predicate</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Classification of Relative Clauses Using Easily Obtainable Features

However, it is meaningless because of different knowledge, features, language, and training data.

4.2. Analysis of contribution of features to the performance

The most significant feature is the surface form of the content word in a headnoun. In a gapless relative clause, the headnoun is usually a bound-noun, time-related noun, or a specific word that is in apposition to the relative clause, not a complement/adjunct. This means that we have much information only from the surface form of the content word in a headnoun. So, it is natural that the most significant feature is the surface form of the content word in a headnoun.

As shown in Table 2, the context features make the performance worse. We guess the reason as follow: Because double subject construction often appears in

Table 2. Precision of classification of relative clauses.

<table>
<thead>
<tr>
<th>Features</th>
<th>Our SVM Performance</th>
<th>Baseline System Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision of basic features</td>
<td>86.31%</td>
<td>61.20%</td>
</tr>
<tr>
<td>Precision of basic + context features</td>
<td>85.12%</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Precision change when one kind of feature is removed.

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using all basic features</td>
<td>86.31%</td>
</tr>
<tr>
<td>From adnominal predicate</td>
<td></td>
</tr>
<tr>
<td>Without surface form of the content word</td>
<td>83.82% (−2.49%)</td>
</tr>
<tr>
<td>(remove 1st feature)</td>
<td></td>
</tr>
<tr>
<td>Without POS-tag of the content word</td>
<td>85.12% (−1.19%)</td>
</tr>
<tr>
<td>(remove 2nd feature)</td>
<td></td>
</tr>
<tr>
<td>Without Surface form of the first function</td>
<td>85.33% (−0.98%)</td>
</tr>
<tr>
<td>word (remove 3rd feature)</td>
<td></td>
</tr>
<tr>
<td>Without POS-tag of the first function word</td>
<td>85.59% (−0.72%)</td>
</tr>
<tr>
<td>(remove 4th feature)</td>
<td></td>
</tr>
<tr>
<td>From headnoun</td>
<td></td>
</tr>
<tr>
<td>Without surface form of the content word</td>
<td>82.37% (−3.94%)</td>
</tr>
<tr>
<td>(remove 5th feature)</td>
<td></td>
</tr>
<tr>
<td>Without POS-tag of the content word</td>
<td>84.47% (−1.84%)</td>
</tr>
<tr>
<td>(remove 6th feature)</td>
<td></td>
</tr>
<tr>
<td>Without surface form of the first function</td>
<td>85.81% (−0.50%)</td>
</tr>
<tr>
<td>word (remove 7th feature)</td>
<td></td>
</tr>
<tr>
<td>Without POS-tag of the first function word</td>
<td>86.73% (+0.42%)</td>
</tr>
<tr>
<td>(remove 8th feature)</td>
<td></td>
</tr>
</tbody>
</table>
a Korean sentence, ‘one argument per clause’ principle is not applied well. So, the context features about subject/object existence are not effective in classifying the type of a relative clause.

As shown in Table 3, after removing one kind of feature, the precision becomes worse except for one case. It means that all the features except one are effective on the classification performance. Only the POS-tag of the first function word from a headnoun makes the performance worse, and the surface form of the first function word in a headnoun is the least significant feature. The function word in a headnoun does not indicate the syntactic function related with the preceding adnominal predicate. It indicates the syntactic function related with another following predicate. So, it is understandable that both the function word-related features in a headnoun are not much effective.

5. Conclusion

We propose a machine-learning based method to classify a relative clause into ‘gap relative clause’ and ‘gapless relative clause’. In previous work, little attention has been paid to automatically classify relative clauses. So, we apply SVM-based machine learning method using easily obtainable features from an adnominal predicate, headnoun, and contexts around a relative clause. In the experiments, our system outperformed the baseline system. Because of the common double subject constructions in Korean sentences, context features about subject/object existence did not contribute to the performance. The most significant feature was the surface form of the content word in a headnoun, and the bad feature was the POS-tag of the function word in a headnoun.

We plan to continue our research to classify a gap relative clause into two classes — ‘subject gap relative clause’ and ‘object gap relative clause’.

Acknowledgments

This work was supported by the Sungshin Women’s University Research Grant of 2006.

References


Classification of Relative Clauses Using Easily Obtainable Features


