Resolution of Governor Selection Ambiguity For Korean Noun Phrase Using Automatically Constructed Lexical Information

Hoojung Chung and **Young-Sook Hwang** and **Yong-Jae Kwak** and
So-Young Park and **Hae-Chang Rim**
Department of Computer Science & Engineering, Korea University
136-701, Seoul, KOREA
{hjchung, yshwang, yjkwak, ssoya, rim}@nlp.korea.ac.kr

Abstract

A natural language parsing system using dependency grammar analyzes a sentence by identifying governor for each linguistic constituent in the sentence. One of the difficult problems in parsing Korean language is to select a correct governor for a noun phrase. To solve this problem, we propose an automatic method to generalize cooccurrence data into conceptual-level lexical information using a thesaurus and raw corpus. And we also present a method to resolve governor selection ambiguity for Korean noun phrase using those lexical information. Experimental result shows that the parser using conceptual-level lexical information as well as cooccurrence information resolves governor selection ambiguity of noun phrase with 92.3% of accuracy.

1 Introduction

Dependency grammar has been widely used for Korean language parsing because it's ability to handle variable-ordered word and discontinuous constituents is compatible with Korean(Na, 1994). It is very simple to analyze a Korean sentence using the dependency grammar. The parsing is performed through identifying a governor for each linguistic constituent.

However, resolving syntactic ambiguity is not so easy. Usually, there are multiple governor candidates for a constituent in a given sentence and it is not easy to select a correct governor for the constituent. We call it "governor selection ambiguity resolution" to select a correct governor for the constituent. To be used practically, the parser should return a single result and, therefore, the governor selection ambiguity should be resolved.

For the purpose of resolving governor selection ambiguity, many approaches have been proposed and one of those approaches is using probabilities. Being able to reflect the tendency to human being's parsing ability easily, a probabilistic approach is widely used for the governor selection ambiguity resolution. Another advantage of the probabilistic approach is that it offers an easy way of learning. A typical way of using probability in governor selection is to use probabilistic dependency grammar (Kim, 1994).

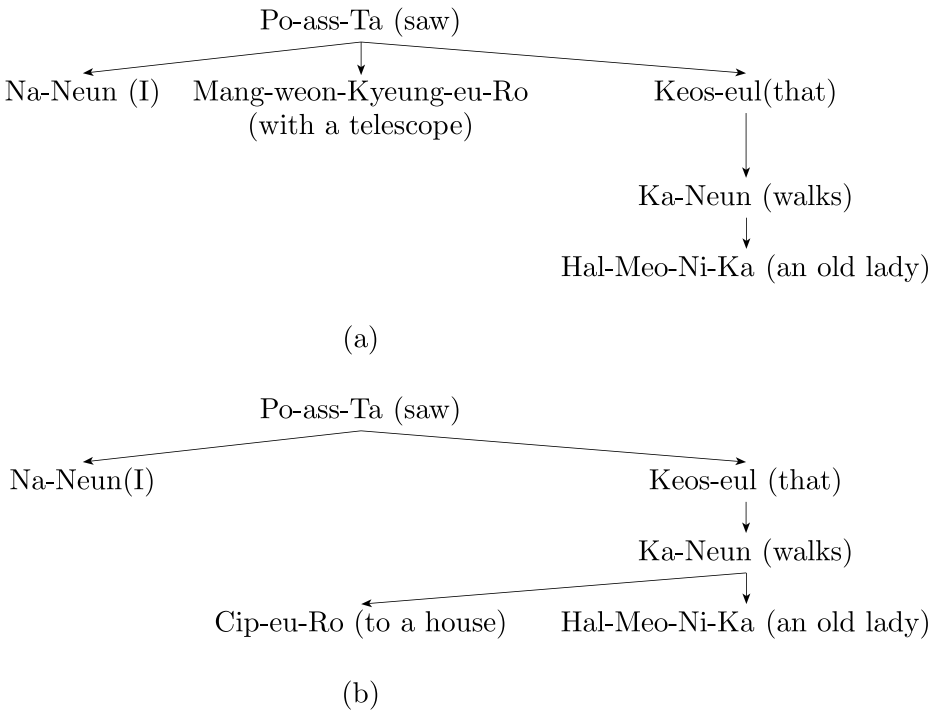
Though this approach can select a generally preferable governor, it is short for parsing precisely. Let's consider the following example.

- Na-Neun Mang-weon-Kyeong-eu-Ro Hal-Meo-Ni-Ka Ka-Neun Keos-eul Po-ass-Ta.
  (I saw (that) an old lady walked with a telescope)
- Na-Neun Cip-eu-Ro Hal-Meo-Ni-Ka Ka-Neun Keos-eul Po-ass-Ta.
  (I saw (that) an old lady walked to a house)

Although the two example sentences consist of identical part-of-speech tags[1], their syntactic structures are different(Figure 1). It is impossible to analyze both of the sentences correctly using probabilistic dependency grammar in this case, since the dependency rules consider part-of-speech tags only. (It is similar to prepositional phrase attachment problem in parsing of English) To overcome this limitation, lexical information has been utilized.

Cooccurrence lexicons, verb subcategorization frames or similar lexical information can be extracted automatically from a large corpus and used as clue to solve this kind of am-

[1] <pronoun + auxilary particle> <common noun + adverbial case particle> <common noun + subjective case particle> <verb + adnominal case particle> <bound noun + objective case particle> <verb + prefinal ending + fianl ending>



(a)

(b)

Figure 1: Sentences produce different syntactic structure with same POS tag sequences

biguity for English (Brent, 1991), (Hindle and Rooth, 1993), (Manning, 1993), (Resnik, 1993), (Li and Abe, 1995). Generally, these kinds of lexical information are in the form of concept or class-based lexicons to avoid data sparseness problem. For Korean, however, there was not any research on constructing verb subcategorization automatically for resolving syntactic ambiguity and it is not easy to get manually constructed lexical information for computational usage. The only way to consider the lexical relation in parsing Korean is to use cooccurrence lexicons which may cause data sparseness problem(Lee et al., 1997),(Yoon, 1997).

This paper describes a method for resolving governor selection ambiguity for Korean noun phrases[2] using conceptual level lexical information as well as lexical-level cooccurrence information. Selecting a correct governor for a noun phrase is one of the difficult and important problem in Korean syntactic analysis. We will also present a way to construct the conceptual level lexical information automatically using raw corpus and a thesaurus. The conceptual level lexical information here means the information of argument(noun and case-particle) required by each predicate. Since a role of predicate is very important in Korean, this information may be useful in various fields of Korean language processing as well.

2 Construction of Conceptual level Lexical Information

First, extract all cooccurrence <predicate, case-particle, noun> from a large raw corpus. A heuristic-based partial parser is used for extracting such data because the raw corpus does not contain any syntactic information in it. Though the result of the partial parser is not precise as that of a full-parser, it can provide useful cooccurrence information. This information is much similar to cooccurrence lexicons which are used to resolve the ambiguity in other researches. Then the nouns in the cooccurrence information are used to general-

[2] *Noun phrase* here indicates the *eojeol* (the spacing unit in Korean like a *word* in English) consists of one or more nouns and a case-particle. For example, a noun phrase *Hal-Meo-Ni-Ka* is madeup with a noun *Hal-Meo-Ni(old lady)* and a subjective case-particle *ka* .

<Meok(eat), obj, Pap(rice)>
<Meok(eat), obj, eum-Sik(food)> ⇒ <Meok(eat), obj, eum-Sik(food)>
<Meok(eat), obj, Kwa-Ca(cookie)>

Figure 2: Generating conceptual level lexical information using cooccurrence information

ize the cooccurrence information into conceptual level lexical information in the form of <predicate, case-particle, concept> (Figure 2). However, we cannot use the cooccurrence information itself for constructing conceptual level information because it includes unappropriate <predicate, case-particle, noun> ternary lexicons(which means there is no dependency relation among predicate, case-particle, noun)[3]. To avoid making erroneous conceptual-level information, we exclude cooccurrence ternary lexicons from cooccurence information whose mutual information value $MI(pred; noun, case-particle)$ is smaller than a certain threshold and use the rest for constructing conceptual level information.

There are two difficult problems in generalizing a noun into a concept. The first is to disambiguate the sense of the noun in cooccurrence information. The second problem arises in clustering nouns for generalizing them into a specific concept. We use a thesaurus which stores hierarchical structure for noun senses to solve both problems.

2.1 Noun Sense Disambiguation

We make following assumption to disambiguate the sense of the noun in coocurrence information.

- Nouns cooccurred with identical predicate and case-particle have similar senses(or concepts).

And we assume that if any two nouns are similar then they have a MSCA (Most Specific Common Abstraction) in a thesaurus and the distance between two nouns in the thesaurus is short. For example, two nouns *Pap(rice)* and *eum-Sik(food)* which has cooccurred with the <predicate, case-particle> pair, <*Meok (eat), obj* >, have a MSCA and their distance to each other is short. We have to consider the depth of the MSCA which indicates how specific the two nouns are. Therefore, we can define the relational distance between two nouns n and m in the thesaurus as follows.

$$d(n,m) = \frac{dist\_in\_thesaurus(n,m)}{depth(MSCA(n,m))} \quad (1)$$

We select the sense of a noun n by using following equation.

$$sense(n) = \mathrm{argmin}_i d(n_i, m_j^x) \quad (m^x \in M) \quad (2)$$

M in the above equation is the set of nouns m^x which are cooccurred with identical <predicate, case-particle> pair. n_i is the ith sense of the noun n. Using this equation, we can disambiguate the sense of the noun in cooccurrence information.

2.2 Clustering Nouns For Generalization

To generalize nouns into a specific concept, we use MSCA of the nouns. First, make every pair of nouns, which are coocurred with certain predicate and case-particle, into a cluster. Then merge any two clusters if there is ancestor-descendant relation between MSCA of the two clusters. In this way, conceptual level lexical information in the form of <predicate, case-particle, concept> can be constructed automatically.

3 Resolving Governor Selection Ambiguity for Korean Noun Phrase

3.1 Association Score

We use an association score function $Assoc(p, c, n)$ to determine an appropriate governor for a Korean noun phrase. $Assoc(p, c, n)$ reflects how a given noun phrase(noun n + case-particle c) is statistically relative to a certain governor(predicate p). The function is defined as follows :

[3] If we used syntactically analyzed corpus for extracting cooccurence information, this errors might be removed

$$Assoc(p,c,n) = \alpha \times \overline{Assoc}(p,c,n) + (1-\alpha) \times \overline{Assoc}(p,c)$$
$$(0.5 \leq \alpha \leq 1)$$

We used $\overline{Assoc}(p,c)$ – the association score of a case for a predicate – to back off the $\overline{Assoc}(p,c,n)$. A constant α is set up by experiments. $\overline{Assoc}(p,c,n)$ and $\overline{Assoc}(p,c)$ are :

$$\begin{aligned}\overline{Assoc}(p,c,n) &= \max(P(n,c|p), \frac{P(concept(n),c|p)}{N}) \\ &\approx \frac{\max(f(p,c,n), \frac{P(p,concept(n),c)}{N})}{f(p)} \\ &\quad (when f(p) \neq 0)\end{aligned}$$

$$\begin{aligned}\overline{Assoc}(p,c) &= P(c|p) \\ &\approx \frac{f(c,p)}{f(p)} \qquad (when f(p) \neq 0)\end{aligned}$$

where $class(b)$ is the concepts that the noun n is subordinated and N is the number of nouns that are subordinated in the concept $concept(n)$.

As mentioned before, $\overline{Assoc}(p,c,n)$ reflects how a given noun phrase is statistically relative to a certain predicate. The conditional probability, $P(n,c|p)$, measures the strength of the statistical association between the given predicate p and the noun n, with the given case-particle p. Similar to the $P(n,c|p)$, the $P(class(n),c|p)$ is the conditional probability estimates the strength of the statistical association between a predicate and a concept $class(n)$, with the given case-particle. The higher score between $P(n,c|p)$ and $\frac{P(concept(n),c|p)}{N}$ is assigned to the association score $\overline{Assoc}(p,c,n)$.

$\overline{Assoc}(p,c)$ indicates how much the predicate requires the given case-predicate. It can be measured similarly.

3.2 Guessing a Governor of Korean Noun Phrase Using Association Score

To select an appropriate governor for a noun phrase, a parser analyzes all possible syntactic structures and assigns a probability for each of them. The probability of a parse tree is calculated by the following equation :

$$P(T) = \prod_i P(relation_i) \qquad (3)$$

where $P(relation_i)$ is the probability of the dependency $relation_i$. Select a parse tree which has the highest probability as the correct parse tree. Governors that are selected by noun phrases in the best parse tree are considered as the appropriate governor for each noun phrase.

4 Experiment Results

We extracted coocurrence information from 8 million word Korean raw corpus using a partial parser after annotating part-of-speech tags to the raw corpus with an automatic part-of-speech tagger(Kim et al., 1997). And then we generalized the noun part of the coocurrence information using a thesaurus containing 12,833 nouns(Cho and Ok, 1997).

We used a simple parser based on dependecy grammar to evaluate our method for resolving governor selection ambiguity. The parser used the association score to disambiguate syntactic structure of sentences. We tested on 100 sentences, which were extracted from training and testing corpus. There were 142 noun phrases which are difficult to select an appropriate governor.

We have performed two experiments to evaluate our approach. In the experiment A, we used only cooccurrence information for resolving government selection ambiguity. And then, we used conceptual level lexical information as well as cooccurrence information in the experiment B to compare our propsed method with other methods. The experimental results are represented in Table 1. In this table, a degree of ambiguity indicates an average number of governor candidates for each noun phrase. The table shows that the result of using conceptual-level lexical information together with cooccurrence information is better than the result of using cooccurrence information alone, especially in the testing set. This means our proposed method is better than others' methods which use cooccurrence information only.

5 Conclusion

This paper describes an approach for resolving governor selection ambiguity for Korean noun

Table 1: Experimental results of a governor selection for a noun phrase

	degree of ambiguity	Exp. A	Exp. B(our method)
Training Set	2.52	94.3%	94.3%
Testing Set	2.49	88.7%	90.1%
Overall	2.50	91.5%	92.3%

phrases using automatically constructed lexical information. This method consists of two phases. In the first phase, we constructed cooccurrence information from raw corpus and generalized it to conceptual level lexical information using a thesaurus. In the second phase, we use an association score, which reflects lexical information constructed in the first phase, to resolve governor selection ambiguity. The approach shows the accuracy of 94.3% and 90.1% in determining correct governors for noun phrases of training and testing sentences respectively.

We found out our assumption that *nouns cooccurred with identical predicate and case-particle have similar senses* may produce incorrect result when the predicate is polysemous word. We will consider those polysemous predicates in our future work.

Since the role of a predicate is very important in parsing Korean, the proposed approch of constructing lexical information can be utilized in other fields of Korean information processing such as resolving ambiguous syntactic role of antecedents in Korean relative clauses.

References

Michael R. Brent. 1991. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th Annual Meeting of the Association for Computiotional Linguistics.*

Pyeong-Ok Cho and Cheol-Yung Ok. 1997. Construction of a semantic hierarchy of Korean nouns. In *Proceedings of the 9th Hangul and Korean Information Processing Conference.*

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1).

Jin-Dong Kim, Heui-Seok Lim, and Hae-Chang Rim. 1997. Twoply Hidden Markov Model : A Korean POS tagging model based on morpheme-unit with eojeol-unit context. In *Proceedings of the 1997 Inter. Conf. on Computer Processing of Oriental Languages.*

Hiongun Kim. 1994. Korean syntactic analysis with probabilistic dependency grammar. Master's thesis, Dept. of Computer Science. KAIST.

Kong Joo Lee, Jae-Hoon Kim, and Gil Chang Kim. 1997. Probabilistic parsing of Korean sentence based on lexical co-occurrence and syntactic rules. In *Proceedings of the 9th Hangul and Korean Information Processing Conference.*

Hang Li and Naoki Abe. 1995. Generalizing case frames using a thesaurus and the MDL principle. In *Proceedings of Recent Advances in Natural Language Processing.*

Christopher D. Manning. 1993. Automatic acquisition of a large subcategorisation dictionary from corpora. In *Proceedings of the 31st Annaual Meeting of the Association for Computational Linguistics.*

Dong Ryul Na. 1994. Investigation on parsing Korean. *Korea Information Science Society Review*, 12(8).

Philip Stuart Resnik. 1993. *Selection and Information: A Class-based Approach to Lexical Relationships.* Ph.D. thesis, University of Pennsylvannia.

Juntae Yoon. 1997. *Syntactic Analysis for Korean Sentences Using Lexical Association Based on Co-occurence Relation.* Ph.D. thesis, Dept. of Computer Science. Yonsei Univ.