AMBIGUITY RESOLUTION USING LEXICAL ASSOCIATION

Juntae Yoon and **Seonho Kim** and **Mansuk Song**
{queen, pobi, mssong}@december.yonsei.ac.kr
Department of Computer Science
Yonsei University, Seoul, Korea

Abstract

Lexical information has been shown to be crucial for decisions on ambiguities. Many statistical parsers is based on probabilities of this dependencies. Our system tried to conjoin lexical information to the best first parsing method and to show that every nodes can be determined using GAT(global association table) ,which is a new data structure to manage the lexical associations. In Korean, the structual ambiguity and the grammatical case ambiguity influence the accuracy of the parser. Lexical information between pairs of words is computed by co-occurrence data extracted from the corpus and to be extended to the conceptual association with thesaurus ,which it attempts to reduce parameter space.

1. Introduction

In Korean, to parse a sentence is to analyze the dependency relation among eojeols. Lexical association between eojeols can be applied in analyzing the dependency realtions in the agglutinative language such as Korean. Therefore, it is necessary to measure the lexical association to choose the correct parse tree. Besides, the grammatical cases of noun phrases are unknown in Korean when the NP has the auxiliary postposition, the postpostion of the NP is omitted, or the NP is moved by the relativization. To identify the unknown grammatical case, the lexical association between the verb and the noun phrase with a postposition is required because the grammatical case is determined by the postposition in Korean.

In this paper, we suggest the global association table(GAT) where lexical associations are globally controlled. The GAT provides the parser with the useful information required for parsing such as the lexical association. The lexical association between the predicate and the NP, is estimated by the cooccurrence relations extracted from corpus. On the basis of the associations presented by the GAT, the actions of the parser are directed and the unknown grammatical cases are identified. We extracted verb and noun co-occurrence data by the partial parser from 30 million eojeol corpus. To reduce parameter space the thesaurus was used on the assumption that the words in the same group behave similarly. Thus the associations of the predicate and the noun were estimated by the co-occurrence of verbs and noun classes. The system was shown to be efficient and precise by experiments.

2. Two kinds of ambiguities

Two types of ambiguities can be appeared in noun phrases and verbs in Korean sentences. The first is the structural ambiguities that are common in most languages. As the head follows its complement in Korean, the ambiguities are inevitable. In (Table 1), the nominal eojeol , '컴퓨터를(computer)', has the possibility to be dependent on '이용해서(using)' and '찾는다(seek)' in the parsing process. Second, the role of the noun phrase is decided by the postpositions for the most part but undecided sometimes. Several types of postpositions serves noun phrases as case markers. For instance, the postposition '을/를' makes nouns *objects*. However, The grammatical case cannot be turned out until parsing process under certain circumstances. For example, auxiliary postpositions add some meaning to NPs instead of marking grammatical cases. The postpositions of NPs can be sometimes omitted. The grammatical cases are veiled in these NPs, so they are uncovered in parsing process. The movement by relativization is also another

Structural ambiguity
많은(many) 사람들이(people) 컴퓨터를(computer) 이용해서(using) 자료를(data) 찾는다(seek). → Many people seek data using the computer.
Ambiguities of grammatical cases caused by the auxiliary postpostion
The first noun phrase is object and the second is subject though their postpositions are the same. . . . 책(book)-도 좋아했다(liked) . . . 나(I)-도 책을(book) 좋아했다(liked)
Ambiguities of grammatical cases caused by the relativization
. . . 메리가(Mary) 만난(meet) 친구(friend) . . . → . . . the friend whom Mary met . . . . . . 학교에(school) 간(go) 친구(friend) . . . → . . . the friend who went to school . . .

Table. 1: The examples of the ambiguities

example. The NP of the clause is moved out of relative clauses. In the relativization, two ambiguities should be resolved. The parser detects the moved noun and then catches what its grammatical case is. In (Table 1), '도' is the auxiliary postposition. The grammatical cases of the nominal eojeols, '책-도' and '나-도', can be identified in the parsing process dynamically. In the third example of the table, '친구' was moved from the clause. It is the object in the former sentence and the subject in the latter one.

3. Defining Global Association Table

We define the global association table as the data structure to record the association between eojeols. In order that the parser obtains information for disambiguations, it looks up the GAT in parsing. Our parser should resolve two ambiguities - structure and grammatical cases. Therefore, the GAT provides two kinds of information. One is for the comparison of associations and the other is for the identification of the grammatical cases.

The row and the column of the table represent eojeols occurring to the left-hand side and to the right-hand side in the parsing process, respectively. The left-hand side eojeol is the complement, and the right-hand side, the candidate for its head. That is, the $GAT(i,j)$ describes the degree of association in case the ith eojeol has a dependency relation to the jth eojeol. Because the head follows its complement in Korean, and the table is a triangular matrix.

To evaluate the association, we extracted co-occurrence data between predicates and nouns by the partial parser. The number of the pairs is 2,000,000 , but the number of the pairs whose the frequency is more than 2, is only 450,000. Considering the number of words in the word dictionary, we can't get enough co-occurrence data for analysis from the corpus. We use the thesaurus (Lim, 1992) to compute the association between groups of words. The parameter space for verb-noun co-occurrences can be reduced in to the co-occurrences of verbs and noun classes. This follows the assumption that words within a group behave similarly. In addition, the requirement ratio for the postposition of the verb is defined. That is, the parameter space was built in terms of the groups of nouns and the grammatical case that the verb demands.

3.1. Lexical Association

We use lexical associations for disambiguation. The lexical association of a nominal eojeol and a predicative eojeol is based on the frequency of co-occurrence. The association of modifier-head relations such as an adverb and a verb, or an adnominal and a noun, is estimated by distance.

First, the co-occurrence data of verbs and nouns were collected. The co-occurrence pairs of nominal eojeols and predicative eojeols were extracted by the partial parser from a corpus of 30 million eojeols. This approach explored by (Hindle, 1993) was shown to be effective for disambiguation aof the preposition attachment.

Second, the selectional restrictions were extracted from the co-occurrence data. Since about 15 percent of the words in our thesaurus have more than two categories, there are few words to have multiple categories. We assigned the thesaurus classes of the words which has a single category.

Third, we builded up the co-occurrence data of verbs and functional words, which was made with the data described above.

We use the data to define the association between the verb and the noun phrase as follows. Let

$$V = \{v_1, \ldots, v_l\},\ N = \{n_1, \ldots, n_m\},$$
$$C = \{c_1, \ldots, c_n\},\ S = \{\phi, \text{가}\ , \text{를}, \text{에}, \ldots\}$$

V, N, C, S be the sets of predicates, nouns, noun classes, and syntactic relations respectively. The postposition is given ϕ, in case the grammatical case is unknown. Given $v \in V, s \in S, c \in C, n \in N$, association score, *Assoc*, between v and n with syntactic relation s is defined to be

$$Assoc_{VN}(n, s, v) = \lambda_1 P(c, s|v) + \lambda_2 P(s|v) \qquad (1)$$

The conditional probability, $P(c, s|v)$ measures the strength of the statistical association between the given verb, v and the class of the noun, n with the given syntactic relation, s. That is, it favors those that have more co-occurrences of the classes of nouns and syntactic relations for verbs. As mentioned before, we use the verb-postposition collocation to back off the $P(n, s|v)$. $P(s|v)$ that means how much the verb requires the given syntactic relation. The number of the pairs is about 240,000 which reach to 12% of the number of verb-noun-postposition triples. The λ_1 and λ_2 are set up by experiments.

3.2. Making GAT

The association value of two eojeols is recorded in the GAT only when the eojeols have a dependency relation. The association is represented by a pair, $\langle$*association-value*, *syntactic-relation*$\rangle$. The association value is calculated by the formula (2) and (3) described in the previous section. The parser uses the value to resolve the structural ambiguities of verbs and nouns. If the unknown syntactic case occurred in noun and verb relations, the candidates are recorded in the GAT with the possible syntactic relation. It is easy to find out the syntactic relation from the formula. Several candidates for the grammatical case is written in the GAT[i,j] when the unknown grammatical case occurred. Three candidates are good for Korean because the maximum three complements can be subcategorized by the head in general cases. The GAT is sorted by the association to look up the most probable phrase in the parsing process. Thus, the global association table is implemented by the global association list.

The following example is represented by (Table 2),

ex 1) (0) 공룡에(dinosaur) (1) 대한(for) (2) 자료를(data) (3) 가진(to have) (4) 화일을(file) (5) 지금(now)
(6) 찾아라(find)
→ Find the files that have the data for dinosaur now

In (Table 2), the cells which are marked with '-' mean that two eojeols don't have any dependency relation. It is most likely that the first eojeol has the possibility to have the dependency relation to the second eojeol. The GAT gives the association to the parser while parsing. The unknown case occurred in the relative clause caused by the fourth eojeol. The omitted postposition is presumed by the GAT. The table indicates that it is most probable that '화일(file)' was moved out of the object of the former clause. That is, the eojeol, '화일(file)' can be the object of the eojeol, '가지다(have)'. In addition, it may be moved out of the subject of the relative clause by the GAT. Therefore, the parser checks both possibilities. If it's all

	0	1	2	3	4	5	6
0	-	(0.08,에)	-	(0.01,에)	-	-	(0.01,에)
1	-	-	(0.08,ϕ)	-	-	-	-
2	-	-	-	(0.07,를)	-	-	(0.04,를)
3	-	-	-	-	(0.06,를) (0.02,가)	-	- -
4	-	-	-	-	-	-	(0.06,를)
5	-	-	-	-	-	-	(1, ϕ)
6	-	-	-	-	-	-	-

Table. 2: The global association table(GAT) for the example sentence, ex 1

right that the nominal eojeol is the object of the predicative eojeol, two eojoels are merged in the objective relation. However, the alternative(subject movement) is checked if it makes an erroneous result.

4. Parsing Algorithm

4.1. Parsing Algorithm

The parsing is directed by the following three operation in the stack and input buffer. Basic operations are CREATE, ATTACH, and DROP. However, its operation is conditioned not by rule matching but by the lexical association of the GAT as shown in the following description.

CREATE If the most probable candidate for the head of the eojeol, e_i, is e_j, that is, $j = index(max(G(i)))$, then merge e_i or the phrase including e_i with e_j or the phrase including e_j, and generate a new phrase.

ATTACH If the e_j is not the most probable candidate for the head of the eojeol e_i, that is, $j \neq index(max(G(i)))$ then wait until e_i meets the most probable candidate indicated by the GAT.

DROP DROP operation is accompanied with CREATE operation in our system because binary grammar is constructed for Korean and thus the binary relation between words is considered.

The GAT gives the parser the prediction of the best candidate, here expressed by the function, $index(max(G(i)))$, which returns the eojeol index of the most probable candidate for the head of the ith eojeol, e_i. When the new node is generated, the unknown grammatical case is recovered, if any. In case that a nominal eojeol has the unknown case caused by the auxiliary postposition or the omission of the postposition, the parser tries to identify the grammatical case. When the noun phrase is moved out of the relative clause, both the moved noun phrase and its grammatical case have to be identified from the predicative eojeol of the clause. The parser turns out the postpostition of the moved NP with the item given by the GAT.

4.2. Parsing

(Figure 1) represents the analysis steps of the sentence in (ex 1). In the fifth row of the figure, the ATTACH operation is executed by the GAT in (Table 2), because the lookahead is not the candidate for the head of the complement on the stack top. Thus the eojeol, '자료를', has to wait until it meets its best candidate. The eojeol, '찾아라' is the best candidate for the eojeol, '자료를', which was estimated by the GAT.

The CREATE operation executed because it is most probable that the eojeol, '가진' is dependent on the next eojeol, '자료를'. The unknown grammatical case is identified in the fourth row because the predicative eojeol is relativized by virtue of the adnominal ending. First, the moved constituent is assumed as the object of the clause by virtue of (Table 2). However, the parser recognizes that the object has been already governed by the predicatve eojeol. Thus it tries for the alternative, that is, the second item of GAT(3,4). The grammatical case is subjective by the given association.

	OP	Stack Top		First Lookahead	
		Constituents	Head	Constituents	Head
1	A		공룡에(dinosaur)		대한(for)
2	A	공룡에 대한	대한(for)		자료를(data)
3	A	공룡에 대한 자료를	자료를(data)		가진(to have)
4	A,C	공룡에 대한 자료를 가진	가진(to have)		화일을(file)
5	B	공룡에 대한 자료를 가진 화일을	화일을(file)		지금(now)
6	A	지금	지금(now)		찾아라(find)
7	A	공룡에 대한 자료를 가진 화일을	화일을(file)	지금 찾아라	찾아라(find)

Fig. 1: an example of analyzing the sentence in (ex 1). OPs are A: Create & Drop operation, B: Attach operation, C: Identification of the unknown case

	Brackets of noun and verb	Correct Brackets	%
result	1727	1595	92.4

Table. 3: experimental results for structural ambiguity resolution

5. Experiment Results

We report the result of analyzing 408 sentences, which were separated from the training corpus. Two kinds of tests have been executed to estimate the resolution of ambiguities. First, a complement has several candidates for its head. To ensure that our method is effiective, the experiment was conducted for the case that the complement is the nominal eojeol and the candidate for the head, the predicative eojeol. (Table 3) shows the accuracy of the structural ambiguity resolution.

Second, the identification of the unknown cases was checked. The results are represented in (Table 4). To improve the accuracy of the system, the parser has to consider linguistic knowledge. The movement of the NP in the relativization is the linguistic phenomenon and all NP cannot be moved.

References

Allen, J. 1995. *Natural Language Understanding.* Benjamin Cummings.

Collins, M. J. 1996. *A New Statistical Parser Based on Bigram Lexical Dependencies* In *Proceedings of 34th Annual Meeting of Association for Computational Linguistics.*

Hindle, D. and Rooth, M. 1993. *Structural Ambiguity and Lexical Relations* Computational Linguistics

Kobayasi Y., Tokunaga T., and Tanaka H. 1994. *Analysis of Japanese Compound Nouns using Collocational Information* In *Proceedings of COLING-94.*

Lim, H. 1992. The Research for Classification of the Korean, (*in Korean*). National Language Research Institute., 1992

Marcus, M. 1980. *A Theory of Syntactic Recognition for Natural Language.* Cambridge, MA: MIT Press.

Total number of unknown cases	Success		Failure	
378	302	80%	76	20%

Table. 4: The results of the identification of unknown cases