Corpus-Based Approach for Nominal Compound Analysis for Korean Based on Linguistic and Statistical Information

Juntae Yoon*
jtyoon@linc.cis.upenn.edu
IRCS
Univ. of Pennsylvania
Philladelphia, PA 19104, USA

Key-Sun Choi
kschoi@world.kaist.ac.kr
KORTERM
Dept. of Computer Science
KAIST, Taejon 305-701, Korea

Mansuk Song
mssong@december.yonsei.ac.kr
Dept. of Computer Science
Yonsei Univ.
Seoul 120-749, Korea

Abstract

Accurate nominal compound analysis is crucial for in application of natural language processing such as information retrieval and extraction as well as nominal compound interpretation. In the nominal compound analysis area, some corpus-based approaches have reported successful results by using statistal cooccurrences of nouns. But a nominal compound often has the similar structure to a simple sentence, e.g. the complement-predicate structure, as well as representing compound meaning with several nouns combined. Due to the grammatical characteristics of nominal compounds, the framework based only on statistcal association between nouns often fails to analyze their structures accurately, especially in Korean. This paper presents a new model for Korean nominal compound analysis on the basis of linguistic and statistical knowledge. The syntactic relations often have an effect on determining the structure of nominal compounds, and we analyzed 40 million word corpus in order to acquire syntactic and statistical knowledge. The structure of a nominal compound is analyzed based on the linguistic lexical information extracted. By experiments, it is shown that our method is effective for accurate analysis of Korean nominal compounds.

1 Introduction

Nominal compound analysis is one of crucial issues that have been continuously studied by computational and theoretical linguists. Many linguists have dealt with nominal compounds in view of semantic interpretation, and tried to explain how nominal compounds are semantically interpreted (Levi, 1978; Selkirk, 1982). In the field of natural language processing, various computational models have been established for syntactic analysis and semantic interpretation of nominal compounds (Finin, 1980; McDonald, 1982; Arens *et al.* , 1987; Pustejovsky *et al.* , 1993; Kobayasi *et al.* , 1994; Vanderwerde, 1994; Lauer, 1995). Recently it has been shown that noun phrase analysis is effective for the improvement of the application of natural language processing such as information retrieval (Zhai, 1997).

Parsing nominal compound is a basic step for all problems related to it. From a bracketing point of view, structural ambiguity is also a main problem in nominal compound analysis like in other parsing problems. Recent works have shown that the corpus-based approach for nominal compound analysis makes a good result to resolve the ambiguities (Pustejovsky *et al.* , 1993; Kobayasi *et al.* , 1994; Lauer, 1995; Zhai, 1997).

Lauer (1995) has compared two different models of corpus-based approaches for nominal compound analysis. One was called as the *adjacency model* which was inspired by (Pustejovsky *et al.* , 1993), and the other was referred to as the *dependency model* which was presented by Kobayasi *et al.* (1994)[2] and Lauer (1995). Given a nominal compound of three nouns $n_1 n_2 n_3$, let As be a metric used to evaluate the association of two nouns. In the adjacency model, if $As(n_1, n_2) \geq As(n_2, n_3)$, then the structure is determined as $((n_1\ n_2)\ n_3)$. Otherwise, $(n_1\ (n_2\ n_3))$. On the other hand, in

* This work was partially supported by a KOSEF's post-doctoral fellowship grant.

[2] In their work, the structure is determined by comparing the multiplication of the associations between all two nouns, that is, by comparing $As(n_1, n_2)As(n_2, n_3)$ and $As(n_1, n_3)As(n_2, n_3)$. It makes similar results to the dependency model.

the dependency model, the decision is dependent on the association strength of n_1 for n_2 and n_3. That is, the left branching tree $((n_1\ n_2)\ n_3)$ is constructed if $As(n_1, n_2) \geq As(n_1, n_3)$, and the right branching tree $(n_1\ (n_2\ n_3))$ is made, otherwise. Lauer (1995) has claimed that the dependency model makes intuitive sense and produces better results.

In this paper, we propose a new model for nominal compound analysis on the basis of word co-occurrences and grammatical relationships immanent in nominal compounds. The grammatical relation can sometimes make the disambiguation more precise as well as it gives a clue of the nominal interpretation. For example, in the nominal compound "KYEONGJAENG(competition) YUBAL[3](bringing about) CHEJE(system)" which means *system to bring about competition*, the nominal compound "KYEONGJAENG CHEJE(competition system)" co-occurs much more frequently than "KYEONGJAENG YUBAL(bringing about competition)". However, its structure is selected to be [[KYEONGJAENG YUBAL] CHEJE]. Why it is analyzed in such a way can be shown easily by transforming the nominal compound to the clause. Because "YUBAL(bringing about)" is the predicative noun that derives the verb with the predicative suffix attached, the modifying noun phrase can be transformed to the corresponding VP which has the meaning of "to bring about competition" (Figure 1). The verb "YUBAL-HA-NEUN(to bring about)" in VP takes the "KYEONGJAENG(competition)" as the object. The predicative noun "YUBAL(bringing about)" also subcategorizes a noun phrase "KYEONGJAENG(competition)" in the same manner as the verb. In the right syntactic tree of Figure 1, it should be noted that the object of a verb does not have the dependency relation to the noun outside the maximal projection of its head, VP. Likewise, the object "KYEONGJAENG(competition)" does not have any dependency with the other noun over the predicative noun "YUBAL(bringing about)".

2 Structure of Nominal Compound

There is not any adjective derivation in Korean. Rather, a noun itself plays an adverbial or adjective role in a nominal compound, or modifies other noun with possessive postposition attached. Table 1 shows various relations occurred in nominal compounds.

As shown in the example, there is a relationship between two nouns which have dependency relation in a nominal compound. For instance, the first nominal compound in the example expresses compound meaning of individual nouns, i.e. *the attribute that a file has*. On the other hand, in the example (c) of the example, the noun "GAENYEOM(concept)" is the object of the predicative noun "GUBUN(discrimination)". A nominal compound, as such, often has the similar structure to a simple sentence, e.g. complement-predicate structure, as well as representing compound meaning with several nouns combined.

Many researchers have tried to explain constraints given in the process of word combination and the principle of semantic composition. Levi (1978) has tried to find the semantic constraints which govern the combination of each noun in a nominal compound. Sproat (1985) has taken into consideration the predicate-argument relation of nominals on the basis of generative syntax. He explained that the nominalization suffix nominalizes the syntactic category of a verb, but θ role of the verb is percolated into its parent node.

We claim that the nominalization is the phenomenon occurred at the syntactic level, and hence the syntactic relations should be reflected in nominal parsing. Namely, for accurate nominal compound parsing, we need syntactic knowledge about nominal compound in addition to lexical information about lexical selection. We propose a nominal parsing model based on two relations, which can be immediately applied to nominal interpretation. We classify the syntactic relations in a nominal compound as follows:

modifier-head relation One noun (adnominal, adjective) adds a certain meaning to the other noun (head) producing a compound meaning (1, 2 in Table 1).

complement-predicate relation One is the

[3] YUBAL is a noun in Korean which means to cause to bring about something

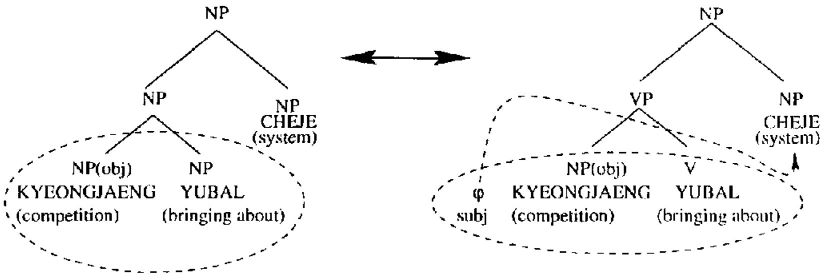

Figure 1: Example shows that syntactic relations have influence on determining the structure of a nominal compound

nominal compound	meaning
PA'IL(file) SOGSEONG(attribule)	file attribute
GIBON(basis) GAENYEOM(concept)	basic concept
GAENYEOM(concept) GUBUN(discrimination)	discrimination of concept
DAETONGRYEONG(president) DANGSEON(being elected)	being elected to president
GONGDONG(working together) BEONYEOG(translation)	to translate together

Table 1: Role of modifying noun in nominal compound

complement (subject, object, adverb) of the other noun (predicative noun) in a nominal compound (3, 4, 5 in Table 1).

When considering the complement-predicate relation, we can figure out some syntactic constraints imposed on nominal compounds. For example, in "PA'IL(file) SOGSEONG(attribute) BYEONKYEONG(change)", "SOGSEONG(attribute)" is the object of the predicative noun "BYEONKYEONG(change)". It can be expanded to a sentence like "*X changes the file attribute*". In other words, the syntactic levels of two phrases "PA'IL SOGSEONG(file attribute)" and "BYEONKYEONG(change)" in the compound noun are different, where one is NP and the other is VP. That the syntactic lèvels (i.e. syntactic categories) of nominal compounds are different means that the different method is required for the proper analysis of their structures.

Next, a predicative noun does not subcategorize more than two nominals with the same grammatical cases. For instance, a predicative noun in a nominal compound governs either a subject or an object at most. The situation is very similar to that occurred in a sentence. In this paper, this is called *one case per sentence*, which means that a predicative noun cannot subcategorize two nouns of the same grammatical cases when the relations of nominals can be expanded to a sentence.

3 Acquiring Lexical Knowledge

We collect lexical co-occurrence instances from corpus in order to get knowledge for nominal compound analysis. The text material is composed of 40 million eojeols of Yonsei Lexicographical Center corpus and KAIST corpus (330M bytes). The Korean morphological analyzer, the POS tagger and the partial parser are used to obtain co-occurrences.

In order to construct linguistic lexical data for nominals, we first extracted verb-noun co-occurrence data from corpus using the partial parser. A noun is connected to a verb with a syntactic relation, and the co-occurrences are represented by triples $\langle verb, noun, syntactic\ relation\rangle$. The postpositions are reposited in the *syntactic relation* field in order to represent the syntactic relations which might occur between two nouns. Nominal pairs with complement-predicate relation are derived from the data extracted.

Predicative nouns become verbs with the verbalization suffix such as '-HA-' attached. For example, the predicative noun 'KEOMSAEK(retrieval)' is verbalized to 'KEOMSAEK-HA(retrieve)' by adding the suffix '-HA-'. Therefore, we can get complement-predicate relations by reducing verbs to predicative nouns with cutting, if any, the verbalization suffix. Table 2 shows some noun-noun co-occurrence examples of complement-predicate relation derived in that way.

Second, co-occurrences composed of only two nouns (complete nominal compound) were obtained. In Korean, complete nominal compounds are extracted in the following way. Let us suppose that N, NA, NP be the set of nouns, the set of nouns with the possessive postposition, and the set of nouns with a postposition except the possessive postposition, respectively.

- For eojeols e_1, e_2, e_3, where $e_1 \notin N \cup NA, e_2 \in N \cup NA, e_3 \in NP$, count (n_2, n_3), where n_2 and n_3 are the nouns that belong to e_2 and e_3 respectively.

The data could contain two relations e.g. modifier-head relation and complement-head relation. Therefore, we manually divide them into two classes by hand according to the relation. Many erroneous pairs could be removed by the manual process. Furthermore, we manually assign to each nominal pair syntactic relations such as SUBJ, OBJ and ADV since the syntactic relation does not explicitly appear from pairs obtained in the second (Table 3), Actually, there is no immanent syntactic relation between two nouns of modifier-head relation. On the other hand, some syntactic relation such as case marker and adverbial relation can be given to two nouns with complement-predicate relation. Some examples are given in Table 3. The data of complement-head relation are merged with those established with the partial parser, which are complement-head co-occurrences. The rest of the data have modifier-head co-occurrences.

Consequently, the complement-predicate co-occurrence is represented with a triple ⟨*comp-noun*, *pred-noun*, *syn-rel*⟩ as shown in Table 2. Syntactic relation is described with postposition for case mark or ADV in Korean. The syntactic relation is not given to the modifier-head co-occurrence.

In the corpus based approach for natural language processing, we should take into consideration the data sparseness problem because the data do not contain whole phenomena of the language in most cases. Many researchers have proposed conceptual association to back off the lexical association on the assumption that words within a class behave similarly (Resnik, 1993; Kobayasi *et al.* , 1994; Lauer, 1995). Namely, word classes were stored instead of word co-occurrences.

Here, we must note that predicates does not act according to their semantic category. Predicates tend to have wholly different case frames from each other. Thus, we stored individual predicative nouns and semantic classes of their arguments instead of each semantic class for two nouns. In effect, given a word co-occurrence pair (n_1, n_2) and, if any, a syntactic relation s, it is transformed and counted in the following way.

```
Let $c_i$ be the thesaurus class which $n_i$ belongs to.
If $(n_1, n_2)$ are a pair in co-occurrences of complement-predicate relation
Then
  For each $c_i$ which $n_1$ belongs to,
    Increase the frequency of $(c_i, n_2, s)$ with the count of $(n_1, n_2)$.
    (Here, $s$ is an immanent syntactic relation)
Else
  For each class $c_i$ and $c_j$ to which $n_1$ and $n_2$ belongs respectively,
    Increase the frequency of $(c_i, c_j)$ with the count of $(n_1, n_2)$
```

Consequently, we built two knowledge sources with different properties, so that we needed to make the method to deal with them. In the next section, we will explain the effective method of analysis based on that different lexical knowledge.

4 Nominal Compound Analysis

In order to make the process efficient, the analyzer identifies the relations in a nominal compound, if any, which can be the guideline of phrase structuring, and then analyzes the structures based on the relations.

Figure 2 shows an example of the phrase structure of a nominal compound to include the complement-predicate relation. We showed that the nominal compound with the complement-predicate relation can be expanded to a simple sentence which contains NPs and VP. This means again that the nominal compound with

argument	predicative noun	syntactic relation
GAENYEOM(concept)	YEONGU(study)	OBJ
GYEONJEHAG(economics)	YEONGU(study)	OBJ
GWAHAGJA(scientist)	YEONGU(study)	SUBJ

Table 2: Noun-noun co-occurrence examples derived from lexical data of predicate YEONGU-HA(research)

first noun	second noun	immanet syntactic relation (meaning)
DAMBAE(tobacco)	GAGE(store)	
CHARYANG(car)	GAGYEOG(price)	
GEUMSOG(metal)	GAGONG(process)	OBJ(process metal)
WANJEON(wholeness)	GADONG(operation)	ADV(operate wholly)

Table 3: Examples of two nouns analyzed

the complement-predicate relation can be divided into one or more phrasal units which we call *inside phrase*.

The nominal compound in Figure 2 has three inside phrases - NP$_{SUBJ}$, NP$_{OBJ}$ and V. Some nominal compounds may not have any inside phrase. Besides, the structure in each inside phrase can be determined by the word co-occurrence based method presented by Lauer (1995) and (Kobayasi *et al.* , 1994), i.e. only statistical association.

4.1 Association between nouns

Inside phrases can be detected based on the association, since two nouns associated with the complement-predicate relation indicate existence of an inside phrase. We distinguish the association relation by discriminating knowledge source. Thus the associations are calculated in a different way as follows. Here, $ambi(n)$ is the number of thesaurus classes in which n appears, and N_{CP} and N_{MH} are the total number of the complement-predicate and the modifier-head co-occurrences respectively.

1. Complement-Predicate
The association can be computed based on the complement-predicate relations obtained from complement-predicate co-occurrence data. It measures the strength of statistical association between a noun, n_1, and a predicative noun, n_2, with a given syntactic relation s which is the syntactic relation like subject, object, adverb. Let c_i be categories to which n_1 belongs. Then, the degree that n_1 is associated with n_2 as the complement of n_2 is defined as follows:

$$Assoc_{CP}(n_1, n_2) = \frac{1}{N_{CP}} \times \sum_i \frac{freq(c_i, n_2)}{ambi(n_1)} \quad (1)$$

2. Modifier-Head
The association of two nouns is estimated by the co-occurrences which were collected for the modifier-head relation. In the similar way to the above, let c_i and c_j be the categories to which n_1 and n_2 belongs respectively. Then, the association degree of n_1 and n_2 is defined as follows:

$$Assoc_{MH}(n_1, n_2) = \frac{1}{N_{MH}} \times \sum_{i,j} \frac{freq(c_i, c_j)}{ambi(n_1)ambi(n_2)} \quad (2)$$

The syntactic relation is determined by the association. If the association between two nouns can be computed by the formula 1, the complement-predicate relation is given to the nouns. If not, the relation of two nouns is simply concluded with the modifier-head relation. We can recognize the syntactic relation inside a nominal compound by the association involved. In order to distinguish the associations in accordance with the relations, the association is expressed by a triple $\langle relation, (syn\text{-}rel, value)\rangle$. The *relation* is chosen with *CP* or *MH* according to the formula used to estimate the association. If *relation* is *CP*, the *syn-rel* has as its value SUBJ, OBJ, ADV etc., which are given by co-occurrence data acquired. Otherwise, ϕ is assigned. Lastly, the *value* is computed by the formula. The association is estimated in the following way,

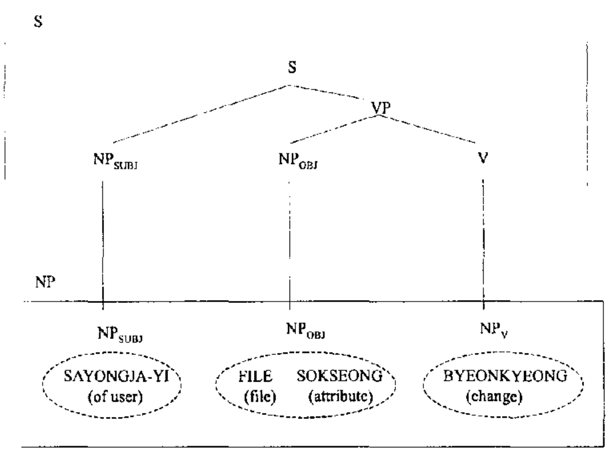

Figure 2: Example of the phrase structure of a nominal compound

therefore:

If $Assoc_{CP}(n_1,n_2) > 0$
$Assoc(n_1,n_2) = \langle CP,(syn\text{-}rel, Assoc_{CP}(n_1,n_2))\rangle$
else
$Assoc(n_1,n_2) = \langle MH,(\phi, Assoc_{MH}(n_1,n_2))\rangle$

If no co-occurrence data for a nominal compound are found in both databases, the modifier-head relations is assumed and the left association is favored for unseen data. The preference of left association is reasonable for bracketing of nominal compounds since the left associations occupy the bracketing patterns much more than the right associations as shown in Table 6.

4.2 Parsing

Since the head always follows its complement in Korean, the ith noun in the nominal compound consisting of n nouns has head candidates of $n-i$ that it might be depend on, and the parser selects the most probable one from them. The parser determines the head of a complement by an association degree of head candidates for the complement.

The easiest way is to have the head candidate list sorted on the association, and select most strongly associative one. In the process of selection, the following constraints are imposed if the relation of two nouns is complement-predicate(CP). Given a nominal compound of three nouns (n_1, n_2, n_3),

- If (n_2, n_3) are related with CP and the syntactic relation of (n_2, n_3) is the same as that of (n_1, n_3), then n_1 is not dependent on n_3. This is called *one case per sentence* constraint.
- If n_1 has an association with n_2 by CP relation, it does not have dependency relation with n_3. See Figure 1
- If n_2 plays an adverbial role for n_3, then n_1 is not linked with n_2.
- Cross dependency is not allowed. It means that dependent-head relations do not cross each other.

As an example, given the nominal compound "$_1$DAEJUNG(public) $_2$MUNHWA(culture) $_3$BIPAN(criticism)", we can get the association table as shown in Table 4. According to the table, the first and second noun can be linked with the modifier-head relation and the association degree of 0.00021. The second noun can depend on the third noun with the complement-predicate relation, and the association degree is 0.00018. Furthermore, the argument is inferred to the object of the predicate, which can be easily recognized by the co-occurrence data extracted.

The table is sorted on the association so that the parser can easily search for the probable candidate for head. In order to effectively detect inside phrases and check the constraints, the syntactic relation should be checked prior to the comparison of the association value. That is, the first key is the *relation* and the second, *association value*. Thus, *CP* > *MH*, and the

	1	2	3
1	-	$(MH, \langle\phi, 0.00021\rangle)$	$(CP, \langle OBJ, 0.00014\rangle)$
2	-	-	$(CP, \langle OBJ, 0.00018\rangle)$
3	-	-	-

Table 4: Association table(AT) for the example nominal compound "DAEJUNG MUNHWA BIPAN"

association values are compared in case of the same *relation* value.

As a consequence, the association table is actually implemented to the association list as follows:

[**DAEJUNG**(public)] - (3,OBJ,$\langle$*CP*,0.00014$\rangle$)
→ (2,ϕ,$\langle$*MH*,0.00021$\rangle$)
[**MUNHWA**(culture)] - (3,OBJ,$\langle$*CP*,0.00018$\rangle$)

From the list we know it is probable that the noun "DAEJUNG(public)" is dependent on "BIPAN(criticism)" with OBJ relation. On the other hand, two words "DAEJUNG(public)" and "MUNHWA(culture)" are found in modifier-head co-occurrences and thus associated with the modifier-head relation. Then, the parsing process can be defined as follows:

$$head(n_i) = n_l \qquad (3)$$
$$l = index(\max_{j=i+1,\ldots,k}(Assoc(n_i, n_j)))$$

Here *index* returns the index of noun n_l whose association with n_i is the maximum. Namely, the parser tries to find the following candidate for the head of each noun n_i in a nominal compound consisting of k nouns, and make a link between them. If constraints are violated while parsing, the next candidate of the list is considered by the parser. According to the algorithm, the given example is parsed as follows:

1. There is only one candidate for "MUNHWA". "MUNHWA(culture)" has the dependency on "BIPAN(criticism)" with object relation. The fact that there is the complement-predicate relation between two nouns indicates that those are the elements of inside phrases, where one belongs to NP and the other has the property of VP. The inside phrases are detected by the syntactic relation.
2. The most probable candidate of "DAEJUNG(public)" is also "BIPAN(criticism)", but it violates *one case per sentence* since the predicative noun already took the object. Thus, another candidate is taken.
3. The next head candidate "MUNHWA(culture)" is satisfactory to the constraints as the modifier-head relation, and "DAEJUNG(public)" is linked to "MUNHWA(culture)" with the relation.

5 Experimental Results

For experiments, we collected 387 nominal compounds from a million word corpus. Nominal compounds composed of more than four nouns (a series of 5 nouns or more) are excluded because the number of them is too small to evaluate our system.

Some examples of analysis are shown in Table 5. In the table, the modifier-head relation is represented with MH, and the complement-predicate is described with OBJ and SUBJ that means object and subject respectively. No depedency between nouns is marked with '-'. For instance, the modifier-head relation is assigned to "MUSOG SINANG" which have the meaning of *the religion of private society that is traditional and superstitious*. However, we don't know about the semantic relation hidden in the results analyzed. In addition, the nominal compound "JISIK'IN-YI(intellectual's) CHAEK'IM(responsibility) HOIPI(evasion)" means *that the intellectual evades his responsibility*. Actually, its structure is determined as [JISIK'IN-YI$_{SUBJ}$ [CHAEK'IM$_{OBJ}$ HOIPI$_{V}$]] which can be expanded to a simple sentence.

Bracketed patterns of the example nominal compounds are shown in Table 6. According to the table, the baseline accuracy of the default system is at least 73.6%. As shown in Table 7, the precision for analysis of nominal compounds

nominal compounds(n1, n2, n3)	structure	$R(n_1, n_2)$	$R(n_1, n_3)$	$R(n_2, n_3)$
MUSOG SIN'ANG JEONTONG (private society,religion,tradition)	((n1 n2) n3)	MH	-	MH
DAEJUNG MUNHWA BIPAN (public, culture, criticism)	((n1 n2) n3)	MH	-	OBJ
FRANCE KEUNDAE MUNHAG (France, modern, literature)	(n1 (n2 n3))	-	MH	MH
JISIK'IN-YI CHAEK'IM HOIPI (intellectual's, responsibility, evasion)	(n1 (n2 n3))	-	SUBJ	OBJ

Table 5: Examples of some nominal compound analyses, $R(n_i, n_j)$ is the syntactic relation between n_i and n_j identified

# of nouns in NP	pattern	freq
3	(n1-YI (n2 n3))	54
	((n1-YI n2) n3)	31
	((n1 n2) n3)	189
	(n1 (n2 n3))	41
4	(n1-YI (n2 (n3 n4)))	2
	((n1-YI (n2 n3)) n4)	10
	(((n1-YI n2) n3) n4)	4
	(n1-YI ((n2 n3) n4)	6
	((n1 n2) (n3 n4))	9
	(((n1 n2) n3) n4)	32
	(n1 ((n2 n3) n4))	2
	((n1 (n2 n3)) n4)	6
	(n1 (n2 (n3 n4)))	1

Table 6: the patterns of nominal compound structures

of the length three and four is about 88.3% and 66.3%. The result is fairly good in that nominal compounds of length three occur much more frequently than those of length four. Overall precision of analysis is about 84.2%.

In addition, we compared three different models to evaluate our system - default model by the dominant pattern, dependency model presented by Kobayasi *et al.* (1994) and Lauer (1995), and our model. In the default analysis, nominal compounds were bracketed by the dominant patterns shown in Table 6. For the dependency model, we used the method presented by Lauer (1995).

Table 8 shows the comparison of the results produced by our algorithm and the other two methods. Our system made a better result in the disambiguation process. The results show that the syntactic information in nominal phrases plays an important role in deciding their structures.

However, there are still errors produced. Some nouns has the word sense ambiguity, and are used as both predicative noun and common noun. Because of the sense ambiguity, some modifier-head relations are misrecognized to complement-predicate. Other errors contain the same kind of results as (Lauer, 1995). To overcome the errors, we think that semantic relations immanent in two nouns are considered.

6 Conclusion

Many statistical parsers have not taken care of analysis of nominal compounds. Furthermore, many researches which dealt with nominal compound parsing seemed not to have computational approaches for linguistic phenomenon in nominal compounds.

We proposed Korean nominal compound analysis based on linguistic statstical knowledge. Actually, immanent syntactic relations like subject and object as well as structures of nominal compounds are identified using our nominal compound analyzer and knowledge acquisition method. Syntactic relations identified can be effectively used in semantic interpretation of nominal compound. Moreover, the parser was more accurate by using linguistic knowledge such as structural information and syntactic relation immanent in nouns.

It is expected that our parsing results including identification of syntactic relations are useful for the application system such as information extraction because many nominal compounds are contained in Korean document bodies and titles, which often represent some events.

However, the system still has some difficul-

	# of nominal compounds	# of success	precision
3 nouns	315	278	88.3
4 nouns	72	48	66.3
total	387	326	84.2

Table 7: Overall results of nominal compound analysis

	total		3 nouns	4 nouns
	# of success	precision	precision	precision
(1)	285	73.6	77.1	58.3
(2)	315	81.4	85.4	63.9
(3)	326	84.2	88.3	66.3

Table 8: Results of nominal compound analysis (1) default analysis by pattern (2) results using the dependency model (3) results using our algorithm

ties, which caused erroneous results. In the future work, we feel it is necessary that lexical parameters be transformed into conceptual parameters with large size of semantic knowledge, and further studies on linguistic properties of nominals be made.

References

Arens, Y., Granacki, J. J., and Parker, A. C. 1987. Phrasal Analysis of Long Noun Sequences In *Proceedings of the 25th Annual Meeting of ACL*

Choi, K. S., Han, Y. S., Han, Y. G., and Kwon, O. W. 1994. KAIST Tree Bank Project for Korean: Present and Future Development. In *Proceedings of the International Workshop on Sharable Natural Language Resources.*

Finin, T. W. 1980. The semantic interpretation of compound nominals. University of Illinois at Urbana-Champaign. University Microfilms International.

Hindle, D., and Rooth, M. 1993. Structural Ambiguity and Lexical Relations. In Computational Linguistics Vol. 19(1).

Isabelle, P. 1984. Another Look at Nominal Compounds In *Proceedings of COLING 84*

Kobayasi, Y., Takenobu, T., and Hozumi, T., 1994. Analysis of Japanese Compound Nouns Using Collocational Information. In *Proceedings of COLING 94*

Lauer, M. 1995. Corpus Statistics Meet the Noun Compound: Some Empirical Results. In *Proceedings of the 33rd Annual Meeting of ACL*

Levi, J. 1978. The Syntax and Semantics of Complex Nominals. Academic

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

McDonald, D. B. 1982. Understanding Noun Compounds. Carnegie-Mellon University.

Pustejovsky, J. and Anick, P. G. 1988. On the Semantic Interpretation of Nominals In *Proceedings of COLING 88*

Pustejovsky, J., Bergler, S., and Anick, P. 1993. Lexical Semantic Techniques for Corpus Analysis. In Computational Linguistics Vol. 19(2).

Resnik, P. 1993. Selection and Information: A Class-Based Approach to Lexical Relationships. Ph.D. dissertation. University of Pennsylvania, Philadelphia, PA.

Selkirk, E. 1982. The Syntax of Words. MIT Press

Sproat, R. W. 1985. On Deriving the Lexicon. Doctoral Dissertation, MIT.

Sproat, R. W. and Liberman M. Y. 1987. Toward Treating English Nominals Correctly. In *Proceedings of the 25th Annual Meeting of ACL*

Vanderwerde, L. 1994. Algorithm for Automatic Interpretation of Noun Sequences In *Proceedings of COLING 94*

Zhai, C. 1997. Fast Statistical Parsing of Noun Phrases for Documenting Indexing. In *Proceedings of the 5th Conference on Applied Natural Language Processing*