Noun Sense Identification of Korean Nominal Compounds Based on Sentential Form Recovery

Seong Il Yang, Young Ae Seo, Young Kil Kim, and Dongyul Ra

In a machine translation system, word sense disambiguation has an essential role in the proper translation of words when the target word can be translated differently depending on the context. Previous research on sense identification has mostly focused on adjacent words as context information. Therefore, in the case of nominal compounds, sense tagging of unit nouns mainly depended on other nouns surrounding the target word. In this paper, we present a practical method for the sense tagging of Korean unit nouns in a nominal compound. To overcome the weakness of traditional methods regarding the data sparseness problem, the proposed method adopts complement-predicate relation knowledge that was constructed for machine translation systems. Our method is based on a sentential form recovery technique, which recognizes grammatical relationships between unit nouns. This technique makes use of the characteristics of Korean predicative nouns. To show that our method is effective on text in general domains, the experiments were performed on a test set randomly extracted from article titles in various newspaper sections.

Keywords: Sense tagging, nominal compound analysis, word sense disambiguation, grammatical relation, machine translation system.

Manuscript received Mar. 12, 2010; revised July 12, 2010; accepted July 26, 2010.

This work was funded by the Ministry of Knowledge Economy of Korean government.

Seong Il Yang (phone: + 82 42 860 6335, email: siyang@etri.re.kr), Young Ae Seo (email: yaseo@etri.re.kr), and Young Kil Kim (email: kimyk@etri.re.kr) are with the Software Resesarch Laboratory, ETRI, Daejeon, Rep. of Korea.

Dongyul Ra (email: dyra@yonsei.ac.kr) is with the Computer & Telecommunications Engineering Division, Yonsei University, Wonju, Rep. of Korea.

doi:10.4218/etrij.10.1510.0083

I. Introduction

Noun phrase analyses are required, not only by information retrieval (IR), but also by almost all applications of natural language processing (NLP). Because nouns are the most frequently used part-of-speech (POS) in sentences, their meaning is always a key factor in forming the meaning of a sentence. Moreover, in the newspaper corpus, TV news caption titles, or technical documents, as well as in patent documents, we can easily find a number of compound noun phrases that contain a series of nouns. A correct semantic analysis of a compound noun, a nominal compound, is inevitable for practical NLP systems.

Based on dictionaries, co-occurrence information, and other various resources, the methods for resolving morphological or syntactic problems have achieved remarkable performance. However, the problem of word sense disambiguation (WSD) is frequently mentioned as one of the most difficult problems in NLP research [1], [2]. In spite of its importance, few WSD systems are used for commercial NLP products, unlike POS taggers and syntactic parsers. This is because most WSD studies have suffered from having little training data. For this reason, much research has focused on specific domain texts starting from a certain amount of sense-tagged corpus, and then extending the training knowledge by bootstrapping, which is an unsupervised learning scheme. Therefore, in texts from unrestricted domains, automatic WSD still remains as one of the most difficult problems in NLP.

A noun sense identification technique in nominal compounds will be used to detect the semantic relationships among unit nouns, and thereby to determine the semantic value of the nominal compound in question. For example, in a machine translation (MT) system, an accurate analysis of the

syntactic/semantic structure of noun phrases and WSD takes an essential role in selecting the proper target language word when the source language word, such as 'bank' (mound/institution) or 'crane' (bird/machine), has more than one meaning. To resolve this problem, the traditional models mainly depended on context information gathered from the training corpus.

In the case of Korean compound nouns that can be seen frequently in newspaper titles, or long-length terms within technical documents, the target word of WSD frequently occurs in the middle of a nominal compound. However, the sense tagging of unit words suffers from a lack of contextual information because the proper context can be extracted only from adjacent nouns. To avoid this weakness of traditional methods, we decided to make use of deep syntactic information that involves semantic restrictions. We found that Korean long-length nominal compounds often have more than one predicative noun and show a syntactic structure such as a whole sentence. We adopt a light verb construction[1)] in Korean [3] to recover the sentential form of a nominal compound, which allows extra resources to be applied for identifying noun senses.

In this paper, we propose a new method for the sense tagging of Korean unit nouns that compose a nominal compound. Our method is based on sentential form recovery to obtain complement-predicate relations from the syntactic structure of the recovered sentence. We argue that our method is advantageous in two respects. First, we aim to overcome the data sparseness problem by exploiting a recycled resource that was already developed in other applications of NLP. Second, if there are unit nouns in a nominal compound that has a linguistic 'predicative' feature, we will recover the syntactic/semantic relations between nouns that are in a kind of noun-verb relation.

Section II of this paper presents related works on word sense disambiguation and nominal compound analysis. Section III deals with knowledge formats and particular characteristics of Korean which are important when selecting noun senses. An overview of our method will be described in section IV. Section V describes the experimental parameters and results, and provides a discussion. In section VI, we sum up the paper and explain our future research directions.

II. Related Works

Previous works on the semantic analysis of noun phrases can be classified into two approaches according to the main resources used: a knowledge-driven approach that uses a dictionary or hierarchical lexical data [4], and a data-driven approach that uses a raw corpus or sense tagged corpus [5]. The knowledge-driven approach depends on the manually constructed knowledge sources, such as case-frame, lexicon, thesaurus, and wordnet. The data-driven approach uses co-occurrence information, such as words in a certain window-sized context surrounding a target word. The techniques introduced in these works are for all nouns appearing in a sentence not confined to those in nominal compounds. Thus, they do not take advantage of the relationships existing among the nouns in a nominal compound.

The relationships among the nouns comprising the internal structure of a nominal compound are important for determining the senses of the nouns. There are two different models for analyzing the structure of a nominal compound. Given a nominal compound of three nouns n_1, n_2, and n_3, let Rs be a metric used to measure the strength of the relation between two nouns. In the adjacency model [6], if $Rs(n_1, n_2) \geq Rs(n_2, n_3)$, then the structure is determined to be $((n_1\ n_2)\ n_3)$. Otherwise, it is analyzed as $(n_1\ (n_2\ n_3))$. On the other hand, in the dependency model [7]-[9], the decision depends on the relative strength of the relations between n_1 for n_2, n_1, and n_3. Thus, $((n_1\ n_2)\ n_3)$ will be selected if $Rs(n_1, n_2) \geq Rs(n_1, n_3)$; otherwise, $(n_1\ (n_2\ n_3))$ will be chosen.

In contrast with these models, our approach attempts to find the grammatical relationships between predicative nouns and the other nouns.

Context information collected from training data without consideration of grammatical relations contains useless information that is not related to the target word. Because, the way in which contextual information is collected from training data is very important, some other research on sense tagging on the sentence level has tried to use a parsing technique to obtain local context information using syntactic relations within a simple sentence [10]-[12]. These approaches for the sentence level are similar to our method for sense tagging on the level of nominal compounds.

Research has generally been interested in sense tagging of all words in a sentence. They do not give special attention to the nouns in nominal compounds. Authors have not employed a deliberate scheme to make use of the internal structure revealing the sentential form. This is what makes our work distinct from theirs. We summarize the major approaches (on the rows) to sense tagging, including ours, in Table 1. Each column represents the types of information used by the systems.

III. Knowledge

In this section, we describe linguistic features and knowledge formats for sense identification in a nominal compound in detail.

1) This is a method for treating predicative nouns that are Sino-Korean words or loan words mostly from English in deverbal form.

Table 1. Comparison among sense tagging approaches.

	Knowledge sources	Co-occurrence in window	Local syntactic information	Inner structure of nominal compound
Knowledge-driven appr.	O	X	X	X
Data-driven appr.	X	O	X	X
Syntactic relation appr.	O	O	O	X
Our appr.	O	O	O	O

[Simple sentence form]
Suni-ka *suhak-ul* *kongbu-rul* *ha-ta*
Suni-NOM mathematics-ACC study-ACC do-LV
(Suni does the study of mathematics.)

[Light verb construction]
Suni-ka *suhak-ul* *kongbu-ha-ta*
Suni-NOM mathematics-ACC study-VB
(Suni studies mathematics.)

Fig. 1. Example of light verb construction.

1. Light Verb Construction

Light verb construction is an often-encountered linguistic form in Korean. It consists of a light verb '*ha*' (do), '*doi*' (made), or '*siki*' (let), and a predicative noun. Having the characteristics of agglutinative languages, almost all Korean predicative nouns can be combined with a light verb for use as a verb. A Korean light verb is attached to a predicative noun as a postfix, and transforms the predicative noun to take the role of a verb. A light verb construction technique in Korean has been applied to many NLP applications [3]. The *hata*-verb in Korean is a popular form of this type of case. The *suru*-verb in Japanese is also similar.

Light verb construction means that the thematic roles cannot be decided by a light verb itself. The necessary information to decide on the thematic roles comes from the predicative noun that participates in the light verb construction. The thematic role information provided by the predicative noun in light verb construction supplies semantic constraints that are useful to disambiguate the word senses. As shown in Fig. 1, the light verb '*ha*' (do) does not presuppose any thematic roles for the arguments. However, the predicative noun '*kongbu*' (study) in front of '*ha*' has its own argument structure that assigns the thematic roles to the arguments. The symbols NOM, ACC, LV, and VB stand for a nominative case, an accusative case, a light verb, and a verb, respectively. Therefore, the predicative noun '*kongbu*' can take the role of the verb 'study' when '*ha*' is attached to it to form a *hata*-verb, that is, a light verb construction.

In Fig. 1, we can note that Korean postpositions can be treated as case markers. For example, '*ka*' and '*ul*' are attached to nouns, indicating nominative and accusative cases in the syntactic structure of '*kongbu*.'

2. Verb Pattern Resource

In order to enhance the domain adaptability of our method, we employed a large-scale resource constructed for MT systems. The verb pattern resource, which was used in the Korean-to-English MT system "FromTo-KE" [13], contains a number of semantic restrictions in deciding on the syntactic/semantic relations associated with a verb. To describe the thematic roles for a light verb, the verb pattern adopts light verb construction forms such as a *hata*-verb. At this point, we need to pay attention to the verb patterns prepared for the light verb constructions. We can use those patterns to decide the syntactic/semantic relations between nouns and a predicative noun because, as we mentioned before, almost all long-length nominal compounds have the structure of a full sentence. It was proven that it is not difficult to parse a single clause of open domain texts and select the syntactic relations between the nouns and verb in the clause. Nominal compound analysis can be improved using the same technique when a sentential form is recovered using a verb pattern.

In our MT system, we assumed that a Korean simple sentence, which has a single verb and argument nouns, is a basic meaningful source to translate. We designed the verb pattern format to represent the semantic structure of a Korean simple sentence. More than one million verb patterns were constructed manually for Korean-to-English translation in common and technical domains.

Korean-to-English verb patterns that have been used for our Korean-to-English MT system consist of two parts divided by a '>' symbol. The first part is for Korean analysis. This part consists of semantic restrictions along with postpositions for arguments and a head verb. The second part is for English generation. It consists of variable symbols and the target language words to be translated, as is shown in Fig. 2.

The first part of each pattern presents template information of a simple Korean sentence. It contains case information as well as thematic roles assigned to the verb. This part fills the argument positions with semantic codes such as 'location' and 'food.' We are currently using about 400 semantic codes to cover the semantic information of nouns for arguments of verbs. In this paper, we use only the first part of the verb patterns to deal with Korean nominal compound phrases.

sisik-ha-1:
A=location!*aeseo* B=human!*ka* C=food!*ul*
sisik-ha!*ta*
>
B taste:v C in A

Fig. 2. Example of a verb pattern.

hakkyo:
0 7 2 125
{
(SEM institution) (EROOT school)
(SEM location) (EROOT education_center)
}

Fig. 3. Lexicon entry format.

Table 2. Semantic hierarchy.

Level 1	Level 2	Level 3	Level 4	…	Examples
Concrete noun	A living thing	Animal	Mammalia		Dog, cat, tiger, lion …
	Organ of plant				Root, trunk, leaf, fruit, …
	Organ of animal				Head, eye, lungs, stomach, …
	…				…
Abstract noun	Obligation				Responsibility, task, role, …
	Achievements				New record, accomplishment, …
	…				…
Activity noun	Mental activity	Thought			Memory, recognition, inference, …
	Religious activity				Prayer, worship, sermon, …
	…				…

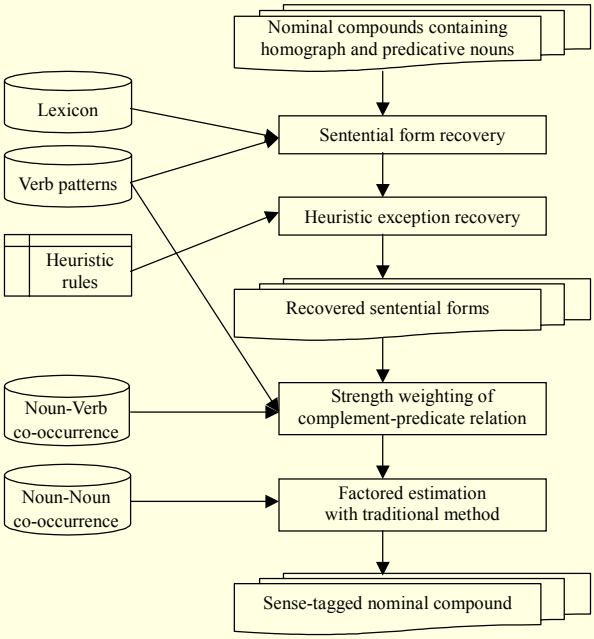


Fig. 4. System flow of noun sense identification.

In Korean, as we described before, the case information can be decided by a grammatical function of nouns that are assigned by postpositions, such as *aeseo*, *ka*, and *ul* in Fig. 2. To make a correct sentential form recovery, we generate postpositions in the argument list of the pattern. The semantic code is a selectional restriction used for deciding whether a noun could be attached to a head verb, fulfilling its case information.

3. Noun Sense Classification

Several methods to describe a semantic code have been proposed for easy and efficient descriptions of the relations between the senses of words. A typical approach for noun sense classification is to construct an ontology system according to certain conceptual criteria. Wordnet is an example of an ontology and was the first success in the classification of noun senses. However, the general ontology form is difficult to apply to our MT system.

To optimize our MT system, we classified the meaning of Korean nouns using around 400 semantic codes having 9 levels in the semantic hierarchy. The semantic codes at the last level of this hierarchy are used for assignment to each entry of the lexicon. At this point, multiple semantic codes can be assigned to a single noun. Table 2 shows a part of our semantic code hierarchy that is currently in use in our Korean-English MT system.

The root code of the semantic hierarchy begins from three types of nouns: concrete, abstract, and activity nouns. Figure 3 shows the format of lexicon entries that contain the POS code with semantic codes.

According to the grammatical functions of Korean morphemes, we use 79 POS tags for Korean morphological analysis. POS codes ranging from 0 to 78 deal with POS tags. POS codes are encoded as a number to simplify the format of the lexicon. For example, 0 indicates a proper noun for a human, 6 represents a foreign word, and 64 is a prefix for a Korean noun.

The first and third number under the lemma '*hakkyo*' describes connective information that is used for morphological analysis. The second number, 7, shows the POS code for a common noun. The forth is the frequency number.

One lemma can have more than one POS code, and one POS code can have more than one semantic code. Depending on the linguistic feature, all of the noun entries in the lexicon will be divided into two noun types (predicative or common). The predicative nouns will be used to recover the sentential form by being interpreted as a verb (via light verb construction).

IV. Proposed Method

In this section, we will describe how the system works according to our method. Figure 4 illustrates an overview of the proposed method. The following subsections will explain each step of the process in detail.

1. Recovering a Sentential Form

Nominal compound analysis for WSD can be more precise when one of the unit nouns turns out to be a predicative noun. On the basis of syntactic/semantic restrictions obtained from the MT pattern resource, we can recover the complete sentential form from the nominal compound. Using this recovered sentential form, which shows the same structure of the compound noun's internal syntactic/semantic relations, we can find complement-predicate relations existing between nouns. The predicative noun in a nominal compound can be used to generate a verb form, which allows us to find the corresponding verb pattern in the MT resource. The other unit nouns in the nominal compound preceding the predicative noun will be treated as a single noun in an argument position.

The sentential form recovery of a compound noun will be performed using the algorithm described in Fig. 5.

For each predicative noun n_j from left to right, do the next three steps.
1. Generate verb form v_j from n_j using a verb generation scheme.
2. For each pattern p_k corresponding to v_j, do the following:
   (a) For each case argument a_i in the selected pattern,
      i. find a preceding noun not marked with "matched" that satisfies the semantic restriction of a_i,
      ii. generate a proper postposition of a_i,
      iii. tag a_i and the noun with label "matched."
   (b) If all a_i's have been matched, finish the generation for v_j; otherwise, for the unit noun w_{j+1} located after v_j,
      i. find an unmatched a_i of p_k matching w_{j+1},
      ii. generate a modifier ending for v_j,
      iii. tag a_i with "matched" (but not w_{j+1}).
   (c) For a preceding noun that does not satisfy the semantic restriction of any a_i, apply the heuristic rules to treat the exceptional cases.
3. Select patterns p_k with best score corresponding to v_j.

Fig. 5. Algorithm for sentential form recovery.

If there is a unit noun w_{j+1} located after the verb form v_j, a single clause headed by the verb form v_j can be treated as a modifier clause. Thus, the unit noun w_{j+1} might be used to take one argument slot to fulfill the case restrictions of v_j. In this case, a modifier ending will be added to the verb form v_j. As a modifee, the unit noun w_{j+1} is placed on the argument position of v_j. Therefore, by satisfying the semantic restriction for w_{j+1}, an unmatched argument a_i of pattern p_k should be tagged as 'matched.'

For example, using the pattern described in Fig. 2, we can recover the sentential form of a compound noun:

baekhwajeom	*sisik-yong*	*bae*
(department store)	(taste-able)	(ship, belly, pear)

In the first step of our method, '*sisik*' (taste) can be transformed into a verb form of '*sisik-ha*' (to taste). Because the Korean postfix '*-yong*' means -able or for-, the postfix part of predicative noun can be used as a clue for LVC. In this example, the clause headed by the predicative noun '*sisik*' will be recovered as a modifier clause by using the postfix '*-yong*' as a clue. At the second step, a postposition '*aeseo*' (indicating a location) can be generated for '*baekhwajeom*' (department store) because '*baekhwajeom*' satisfies the semantic code 'location' appearing just before '*aeseo*' in the verb pattern. In the last step, because there is a noun '*bae*' (ship, belly, pear) located after the verb '*sisik-ha*' (to taste), we can create a modifier ending, '*-nun*.' The final result of sentential form recovery becomes

baekhwajeom-aeseo sisik-ha-nun bae-i-da
(It is a ship/belly/pear prepared in a department store to taste).

The noun '*bae*' is a homograph that has three senses, ship, belly, and pear. Sense identification for the noun '*bae*' can be carried out by complement-predicate relation analysis using the

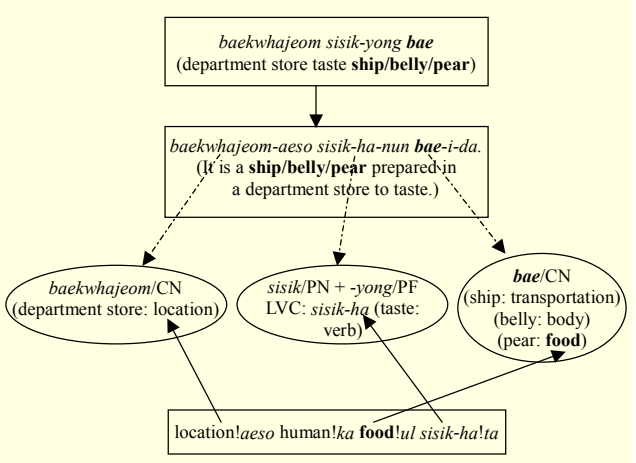


Fig. 6. Example of noun sense identification.

recovered sentence. Figure 6 shows this process in more detail.

As we depicted at the bottom of Fig. 6, the verb pattern '*sisik-ha*' was selected because the sentential form recovery of the nominal compound was performed successfully by fulfilling the case information. In Fig. 6, from the dotted lines, we can find that each word of the recovered sentence is connected to a sense list of the word. By using the semantic restrictions in the verb pattern that are pointed to by the dark-lined arrows, the target sense of the homograph noun '*bae*' can be disambiguated to be 'pear.' Note that not only we can select the target sense 'pear' from the sense list, but we can also decide that the noun '*bae*' fills the accusative case of the verb '*sisik-ha*' (indicated by the postposition '*ul*'). At the same time, '*baekwhajeom*' is used to fill the 'location' case of the verb '*sisik-ha*.' It co-occurs with '*bae*' to fill arguments of the predicative noun. CN, PN, and PF indicate a common noun, a predicative noun, and a postfix, respectively.

There can be two or more predicative nouns in a nominal compound. Each predicative noun considers the nouns preceding it and one noun just after it. For example, in "n_0 n_1 pn_2 n_3 pn_4 n_5," pn_2 tries to use n_0, n_1, and n_3, but pn_4 considers pn_2, n_3, and n_5 to fill their cases (where n denotes a common noun and pn stands for a predicative noun). Note that n_3 is used for both pn_2 and pn_4. The algorithm uses the marker 'matched' to handle this problem as in lines (a)-iii and (b)-iii of step 2.

The computational cost for the algorithm depends on two factors, the number of predicative nouns in the nominal compound and the amount of time taken to search for an appropriate pattern in the pattern database. The number of predicative nouns in a nominal compound is barely more than three, which does not make a serious increase in computational cost. The amount of time to search for a pattern in the pattern database is not problematic because efficient indexing schemes such as a direct file with a hash table or B+ tree can be used and allow the system find a pattern quickly. Actually, it takes a short time (a few milliseconds) to find patterns for a predicative noun. As long as the pattern database size is fixed, the total time for searching patterns for a nominal compound is $O(c \bullet k)$, where c is a constant and k is the number of predicative nouns.

Details on applying the heuristic rules to handle exceptional cases in step 2(c) and selecting a best pattern by estimating scores of the matched patterns in step 3 of the algorithm will be described in the next two subsections.

2. Heuristic Exception Recovery

In recovering the sentential form of a nominal compound, we found a few exceptional cases in treating unit nouns. To deal with this problem, we designed heuristic rules.

Table 3 describes a few heuristic rules of sentential form recovery. For example, as shown in the first rule of Table 3, when a common noun n_i is located before the other common noun n_{i+1} and is unable to recover the appropriate sentential form because all the cases are already filled with other nouns, we can adopt a heuristic rule to generate the Korean postposition '*ui*' (of) to represent a '*has-a*' relation between n_i and n_{i+1}.

Table 3. Heuristic rules for sentential form recovery.

Heuristic rule	Example of sentential form recovery
{ n_j[unmatched] + n_{j+1}[matched] } → { n_j + *-ui* + n_{j+1} }	*jaepum irum sogae* (product name introduction) → *jaepum-ui irum-ul sogae-ha-ta* (introducing the name of a product)
{ pn_j + pn_{j+1} } → { pn_j + *-nun* + *gut* + pn_{j+1} }	*yeonseol bulheo* (address disallowance) → *yeonseol-ha-nun gut-ul bulheo-ha-ta* (it is not allowed to address)

In the recovery attempt for the nominal compound '*jaepumirum sogae*' (product name introduction), the common noun '*jaepum*' (product) cannot fill any remaining cases of the predicative noun '*sogae*' (introduction) because the other common noun '*irum*' (name) has already filled an accusative case. In this case, a '*has-a*' relation exists between the two common nouns. Therefore, the sentential form of '*jaepum-ui irum-ul sogae-ha-ta*' (introducing the name of a product) will be generated by the heuristic rule. The suffix '*-ui*' denotes the '*has-a*' relation between '*jaepum*' and '*irum.*'

On the other hand, an '*is-a*' relation can occur between n_i and n_{i+1} under a different condition. For example, in the Korean nominal compound

gasu *Kim-Sumi* *untae* *seoneon*
(singer) (Kim-Sumi) (retirement) (announcement)

the common noun '*gasu*' (singer) cannot be assigned to any remaining case slot for the predicative noun '*seoneon*' (announcement). When the common noun '*gasu*,' which has the sense, 'occupation,' is placed in front of a proper noun for a human, as shown in the example, we can treat the relation between '*gasu*' and '*Kim-Sumi*' as an '*is-a*' relation. In Korean, we usually use '*in*' (which is) to manifest an '*is-a*' relation. Therefore, by a heuristic rule for handling '*is-a*' relations, the result of sentential form recovery becomes

gasu-in Kim-Sumi-ka untae-rul seoneon-ha-ta.
(Kim-Sumi who is the singer announces retirement.)

Another case that requires an exceptional heuristic is to perform a sentential form recovery for consecutive predicate nouns (the second rule in Table 3). To resolve this problem, we

adopt a heuristic rule that creates a dependant noun. For example, a predicative noun, pn_j, that is located before another predicative noun, pn_{j+1}, can be recovered by using a newly created dependant noun, '*gut*' (thing), and a modifier ending. The sentential form recovered from the predicative noun pn_j becomes a modifier clause for the dependant noun '*gut*' (thing). The newly created dependant noun will fill one of the case slots of the predicative noun pn_{j+1}.

As in common noun entries in the lexicon, predicative nouns have their own semantic codes, too. They can be used as a case filler. To take advantage of the semantic information of predicative nouns, the application of heuristic rules can be extended. For example, as we described above, when a new dependant noun constituent '*gut*' (thing), which has a thematic role in the sentence, is created following a predicative noun pn_j, the semantic code of the predicative noun pn_j will be copied to the dependant noun '*gut*' (thing). Then, the predicative noun pn_j can be recovered into a verb form with a modifier ending. Now, the modifier clause headed by pn_j, is represented by the modifee '*gut*' (thing) which takes a role of filling a case slot of the predicative noun pn_{j+1}. To test if it satisfies the semantic restriction of a case of pn_{j+1}, the dependant noun '*gut*' with a semantic code equal to that of pn_j will be matched against the verb patterns of pn_{j+1}.

As shown in the second rule of Table 3, the predicative noun '*yeonseol*' (address) is followed by another predicative noun '*bulheo*' (disallowance) in the Korean compound noun:

kukhae	*yeonseol*	*bulheo*
(national assembly)	(address)	(disallowance)

In this case, to make a sentential form revealing a predicate-predicate relation, the second heuristic rule creates a new dependant noun '*gut*' (thing). The semantic code 'speak' of the predicative noun '*yeonseol*' is copied to '*gut*.' Then, the LVC form of the predicative noun '*yeonseol*' becomes '*yeonseol-ha-nun*' to be the head of the modifier clause. The dependant noun '*gut*' is placed on the argument position of the other predicate '*bulheo*,' and the sense 'speak' of '*gut*' is confirmed to match a semantic restriction of the argument correctly. Therefore, we can obtain the following result as the sentential form recovery:

kukhae-aeseo yeonseol-ha-nun gut-ul bulheo-ha-ta.
(it is not allowed to address in the national assembly.)

Heuristic rules are managed under two kinds of conditions: one for applying the rule successfully and the other for avoiding the rule execution exceptionally.

3. Noun Sense Identification

As a result of sentence recovery, the syntactic/semantic restrictions for the complement-predicate relations were obtained from the pattern. Each selectional restriction for a complement-predicate relation must be satisfied by a 'noun-predicative noun' pair in the compound noun. Thus, we selected pear (which is a food) as the sense of '*bae*' in the previous example.

The problem becomes more complex because predicative nouns can have two or more verb patterns that are successful in sentential form recovery. For example, '*gaebal*' (develop) has two verb patterns. One pattern takes the semantic code 'location' as the object, but the other pattern takes 'facility' as the object. Then, these patterns should compete for selection as shown in steps 3 in the algorithm of Fig. 5.

To weigh the strength of the pattern matching result (the recovered sentence) more effectively, our algorithm tries to grasp the predicate-argument-adjunct structure for each predicate using verb patterns. As mentioned before, the verb pattern can describe not only arguments but also adjuncts. In predicate-argument-adjunct analysis, pattern matching will be performed to check whether all the case slots of a predicate were filled with nouns satisfying the correct semantic restrictions. Each predicate recovered from predicative nouns is used to pick all verb patterns. The verb patterns after pattern matching are evaluated according to the matched proportion. The matched proportion means the proportion of the cases whose semantic restrictions are fulfilled out of all cases of the predicate. For example, the verb pattern that found correct case fillers of a nominative, an adverbial, and an accusative case receives a higher score than the other verb pattern, which matched only with an accusative or a nominative case. The higher the proportion is, the higher the score of the evaluation.

Even through the pattern matching, the complement-predicate relation analysis (CPA) may not completely determine the sense identification because there can be multiple patterns with the same matched proportion. To cope with this situation where the mechanism explained so far is not enough to determine the senses of the nouns in a nominal compound, we need to extend CPA to use the statistical information explained in the next paragraphs.

In sense determination that uses statistical information, the statistical sense-tagging model, WSD_{vp}, is defined to make use of noun-verb co-occurrence information extracted from the sense tagged corpus, which was constructed manually. A nominal compound consists of n words $w_{1,n}$ and a predicative noun v.

$$WSD_{vp}(w_{1,n}, v) = \arg\max_{s_{1,n}} P(S_{1,n} = s_{1,n} \,|\, W_{1,n} = w_{1,n}, v) \quad (1)$$

$$= \arg\max_{s_{1,n}} P(s_{1,n}, w_{1,n}, v) / P(w_{1,n}, v) \quad (2)$$

$$\approx \arg\max_{s_{1,n}} P(s_{1,n}, w_{1,n}, v). \quad (3)$$

It determines the sense list of words to maximize the score of noun-verb co-occurrence probability. Equation (1) is derived by the definition of the problem, which is to determine the senses of the nouns given the nominal compound "$w_{1,n}$ v." Equation (2) is easily obtained by using Bayes' rule. Because the denominator of (2) is constant according to $s_{1,n}$, it can be omitted in (3).

When there is more than one verb pattern remaining to compete (having the same matched proportion), the pattern that has the maximum probability will be selected, and the sense list proposed by that pattern will be the final output of our process.

Nouns in argument positions have a tendency of co-occurring with a predicative noun, rather than their adjacent nouns. This is different from the Markov assumption in the morphological part of speech tagging in which only the previous tag has an effect on the determination of the current tag. Thus, we propose the WSD_{vp} function to determine the senses of single nouns in an argument position as follows. The WSD_{vp} function multiplies all of the probabilistic co-occurrence scores of a single noun and predicative noun pair.

$$WSD_{vp}(w_{1,n}, v) \approx \arg\max_{s_{1,n}} P(s_{1,n}, w_{1,n}, v) \quad (4)$$

$$= \arg\max_{s_{1,n}} \prod_{i=1}^{n} P(s_i \mid w_i) * P(v \mid s_i, w_i) \quad (5)$$

$$\approx \arg\max_{s_{1,n}} \prod_{i=1}^{n} P(s_i \mid w_i) * P(v \mid s_i). \quad (6)$$

When a lower semantic class s_i suffers from data sparseness, we can back off to more general semantic classes. The semantic hierarchy we used has nine levels.

V. Experimental Results

For experimental evaluation, we filtered out all of the erroneous results of morphological analysis that made a wrong selection out of common or predicative nouns. We conducted three experiments. The first experiment, Base, is a baseline method. In this method, the sense of each noun that is selected is the sense that is most frequently used for that word. The sense selection is made without considering other nouns in the compound. In the second experiment, sense selection is affected by the senses of other nouns. This is a traditional model that uses noun-noun co-occurrence as contextual information. We call this method the contextual co-occurrence analysis (CCA). It tries to disambiguate word senses in a compound noun w_1 w_2 … w_n without considering predicative nouns. CCA uses a word sense disambiguation function WSD_{np} for a nominal compound to choose a sense list $s_{1,n}$ of n nouns.

Table 4. Tagging accuracy of homograph unit nouns.

Total number of compound nouns	869
Average number of nouns	5.356
Total number of homographs	880
Average senses of homographs	2.767
Precision (Base)	71.25% (253 errors)
Precision (Base+CCA)	84.20% (139 errors)
Precision (Base+CCA+CPA)	87.27% (112 errors)

$$WSD_{np}(w_{1,n}) = \arg\max_{s_{1,n}} P(s_{1,n}, w_{1,n}) \quad (7)$$

$$= \arg\max_{s_{1,n}} \prod_{i=1}^{n} \sum_{j=i+1}^{n} P(s_i \mid w_i) * P(s_j \mid s_i, w_i). \quad (8)$$

Finally, the third experiment, CPA, is the proposed method for sense disambiguation with consideration of a complement-predicate analysis using the sentential form recovery technique. We recovered the sentential form of compound nouns, which reveals internal structural information.

For the experiments, a total of 3,088 homographs were selected from a Korean-English dictionary with noun sense classifications and we extracted 290,800 sentences that include the homographs from a raw newspaper corpus. We manually created a sense tagged corpus based on these sentences and used it as the training data for both the CCA and CPA methods.

For experimental evaluation, 869 nominal compounds that have more than one homograph were randomly extracted from a raw newspaper corpus that has not been used in training data. We used about 800,000 verb patterns for a common domain made in developing a Korean-to-English MT system.

We built three models and performed experimentation on them: Base, Base+CCA, and Base+CCA+CPA. CCA and CPA are added to the Base model as factored models assigning weights to each method. The models Base+CCA and Base+CCA+CPA are obtained by linearly interpolating component methods. For example, the score for Base+CCA+CPA is computed by

$$\text{Score} = \alpha \times P_{\text{Base}} + \beta \times P_{\text{CCA}} + \gamma \times P_{\text{CPA}}, \quad (9)$$

where $\alpha+\beta+\gamma = 1.0$.

The last three rows of Table 4 show the results of the experiments. CCA is data-driven and thus reflects the data-driven approach (the most common one) introduced in Table 1. The syntactic relation approach cannot be applied to nominal

compounds. It works only for sentences. The knowledge sources for a knowledge-driven approach are not available for Korean. Therefore these two approaches were not included in our experimentation.

We observe a significant increase in final accuracy (87.27%) by adding CPA. The main reason for these results is that our method, CPA, uses complement-predicate analysis patterns that are assigned to the verb form recovered from a predicative noun, enriching the context information.

The baseline approach makes errors in sense tagging because it has no way of using any contextual information inside or outside of the nominal compound. It just depends upon the relative frequency of the senses of the nouns. An example is '*kyeongbo* (alarm/walking) *daehae* (competition).' The sense 'alarm' occurs more as the sense of '*kyeongbo*' than 'walking.' Therefore, even though the context '*daehae*' prefers 'walking' to 'alarm,' the baseline system will always select 'alarm' for '*kyeongbo*.'

The proposed method, too, can produce errors in sense tagging for nouns in nominal compounds even if it uses contextual information. However, it happens only when two or more patterns are found for a predicative noun and they are almost the same in matching the nominal compound. For example, a predicative noun '*gaebal*' (develop) has two patterns 'developing a facility' and 'developing a location.' The homograph noun '*sudo*' has two possible senses, 'facility' (plumbing) and 'location' (capital city). A problem occurs to do sense tagging for the nominal compound '*sudo gaebal*,' because both verb patterns match the nominal compound to the same extent (selectional restrictions for the noun '*sudo*' is met by both patterns). In this case, the statistical information will be the main factor to make the decision. It favors 'facility' as the sense of '*sudo*' because the training data says that a noun with sense 'facility' co-occurs more with a verb with sense 'develop' than 'location.' In the text '*sudo gaebal-ul tongha-n kukka kyeongjaeng-ryeok kanghwa*' (enhancing national competitiveness through developing the capital city), the sense 'location' is the correct sense for '*sudo*.' Our system selects a wrong sense of 'facility.'

However, if there are other additional nouns in the nominal compound '*sudo gaebal*,' they can provide a clue to select a correct sense. Therefore the problem just mentioned happens very rarely and isn't a major issue to our method.

VI. Conclusion

To improve the accuracy of sense tagging, we have proposed a new sense identification method for Korean nominal compounds based on the complement-predicate relation analysis. Our method is based on the sentential form recovery technique. Adopting a recycled resource that was employed in another application of NLP, we were able to make use of it again in order to enlarge the context information as extra knowledge that can be added to traditional resources. Resulting from the additional knowledge and recovery techniques used to analyze a deep syntactic structure, we achieved prominent improvement in noun sense identification of Korean nominal compounds. Experiments were performed on 880 homographs in nominal compounds using the semantic restriction made available in consideration of complement-predicate syntactic relations. From the experiments, we found that the method using MT verb patterns for predicative nouns based on sentential form recovery was quite effective. In the experiments, we used only the words contained in nominal compounds, not the other words of the sentences in which the nominal compounds were placed. In a full sentence that contains compound nouns, we can expect even higher precision because the head noun of a nominal compound can be disambiguated by the outer context information of a co-occurring verb or adjective.

In future work, we will extend the heuristic rules to cover exceptional cases of sentential form recovery and utilize recovery algorithms to strengthen robustness for noun sense identification. Furthermore, by using the technique of sentential form recovery, we will enlarge our research area to prove the usefulness of nominal compound analysis for paraphrasing and target word selection in MT systems.

References

[1] N. Ide and J. Veronis, "Introduction to the Special Issue on Word Sense Disambiguation: The State of Art," *Computational Linguistics*, vol. 24, no. 1, 1998, pp. 2-40.

[2] S.W. McRoy, "Using Multiple Knowledge Sources for Word Sense Discrimination," *Computational Linguistics*, vol. 18, no. 1, 1992, pp. 1-30.

[3] M.P. Hong, C.H. Kim, and S.K. Park, "Treating Unknown Light Verb Construction in Korean-to-English Patent MT," *Lecture Notes Computer Science*, vol. 4139, 2006, pp. 726-737.

[4] B. Beamer et al., "UIUC: A Knowledge-Rich Approach to Identifying Semantic Relations between Nominals," *Proc. ACL SemEval Workshop*, 2007, pp. 386-389.

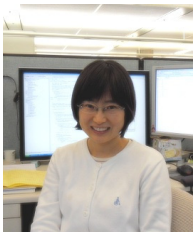
[5] D. Davidov and A. Rappoport, "Classification of Semantic Relationships between Nominals using Pattern Clusters," *Proc. ACL*, 2008, pp. 227-235.

[6] J. Pustejovsky, P. Anick, and S. Bergler "Lexical Semantic Techniques for Corpus Analysis," *Computational Linguistics*, vol. 19, no. 2, 1993, pp. 331-358.

[7] Y. Kobayasi, T. Takenobu, and T. Hozumi, "Analysis of Japanese Compound Nouns Using Collocational Information," *Proc.*

COLING, 1994, pp. 865-869.
[8] M. Lauer, "Corpus Statistics Meet the Noun Compound: Some Empirical Results," *Proc. ACL*, 1995, pp. 47-54.
[9] J. Yoon, K. Choi, and M. Song, "A Corpus-Based Approach for Korean Nominal Compound Analysis Based on Linguistic and Statistical Information," *Natural Language Engineering*, vol. 7, 2001, pp. 251-270.
[10] Y.K. Kim et al., "Word Sense Disambiguation Using Lexical and Semantic Information within Local Syntactic Relations," *Proc. 30th Annual Conf. IEEE Ind. Electron. Soc.*, 2004, pp. 3111-3114.
[11] D. Lin, "Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity," *Proc. ACL*, 1997, pp. 64-71.
[12] P. Chen et al., "A Fully Unsupervised Word Sense Disambiguation Method Using Dependency Knowledge," *Proc. HLT*, 2009, pp. 28-36.
[13] M. Hong et al., "Customizing a Korean-English MT System for Patent Translation," *Proc. 10th MT-Summit*, 2005, pp. 181-187.

Seong Il Yang received his BS and MS and finished PhD courses in computer science from Yonsei University, Korea, in 1994, 1996, and 1998, respectively. He is currently a senior member of the engineering staff of the Natural Language Processing Team at ETRI, Korea, where his research interests are broadly in the area of artificial intelligence, machine learning, natural language processing, and machine translation.



Young Ae Seo received the BS in computer science from Kyungpook National University, Korea, in 1996. She received the MS in computer science from POSTECH, Korea, in 1998. Currently, she is working on the PhD course in computer engineering at KAIST, Korea. Since 1998, she has been with ETRI, Korea, as a research member. Her research interests are natural language processing, machine translation, machine learning, and dialogue processing.

Young Kil Kim received his MS and PhD in electronics and telecommunications from Hanyang University, Korea, in 1993 and 1997, respectively. He has been a principal member of the engineering staff and team leader of the Natural Language Processing Team at ETRI, Korea. His research interests include natural language processing, dialogue understanding, and machine translation.

Dongyul Ra received his BS in electronics from Seoul National University in 1978. He received his MS and PhD in computer science from KAIST in 1980, and Michigan State University in 1989, respectively. He has been a faculty member of Yonsei University since 1991. His research interests include natural language processing, artificial intelligence, and information retrieval.