Identifying Syntactic Role of Antecedent in Korean Relative Clause Using Corpus and Thesaurus Information

Hui-Feng Li, Jong-Hyeok Lee, Geunbae Lee
Department of Computer Science and Engineering
Pohang University of Science and Technology
San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Republic of Korea
hflee@madonna.postech.ac.kr, {jhlee, gblee}@postech.ac.kr

Abstract

This paper describes an approach to identifying the syntactic role of an antecedent in a Korean relative clause, which is essential to structural disambiguation and semantic analysis. In a learning phase, linguistic knowledge such as conceptual co-occurrence patterns and syntactic role distribution of antecedents is extracted from a large-scale corpus. Then, in an application phase, the extracted knowledge is applied in determining the correct syntactic role of an antecedent in relative clauses. Unlike previous research based on co-occurrence patterns at the lexical level, we represent co-occurrence patterns with concept types in a thesaurus. In an experiment, the proposed method showed a high accuracy rate of 90.4% in resolving ambiguities of syntactic role determination of antecedents.

1 Introduction

A relative clause is the one that modifies an antecedent in a sentence. To determine the syntactic role of the antecedent in a verb argument structure of relative clause is important in parsing and structural disambiguation(Li et al., 1998). While applying case frames of a verb for structural disambiguation, identifying the role of antecedent will affect the correctness of structural disambiguation impressively.

In this paper, we will describe a method of identifying the syntactic role of antecedents, which consists of two phases. First, in the learning phase, conceptual patterns (CPs) and syntactic role distribution of antecedents are extracted from a corpus of 6 million words, the Korean Language Information Base (KLIB). The conceptual patterns reflect the possible case restriction of a verb with concept types, while the syntactic role distribution shows the preference of syntactic role of antecedents of a verb. Second, in the application phase, the syntactic role of an antecedent is decided using CPs and the syntactic role distribution.

In regards to the rest of this paper, Section 2 will review the problems and related work. Section 3 will describe a statistical approach of conceptual pattern extraction from a large corpus as knowledge for determining syntactic roles. Section 4 will describe how to identify syntactic roles using conceptual patterns and syntactic role distribution of antecedents in the corpus. Section 5 will then present an experimental evaluation of the method. The last section makes a conclusion with some discussion. The Yale Romanization is used to represent Korean expressions.

2 Problems and Related Work

In English, it is possible to recognize the syntactic role of antecedents by their position (trace) in relative clauses and the valency information of verbs. For example, the syntactic role of an antecedent *man* can be recognized as subject of the relative clause in a sentence "He is the *man* who lives next door" and as object in a sentence "He is the *man* whom I met." The relative pronouns such as *who, whom, that, whose*, and *which* can also be used in identifying the role of antecedents in relative clauses.

However, it is not a trivial work to identify the syntactic role of antecedents in Korean relative clauses. Korean is such a head final language that the antecedent comes after the relative clause. The rest of this section will describe three main characteristics of Korean relative clauses that make it difficult to determine the syntactic role of their antecedents. The **first characteristic** is that unlike English, Korean lacks relative words corresponding to English

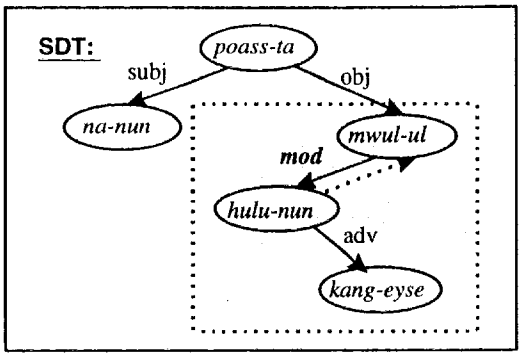

Figure 1: Syntactic dependency tree for (1)

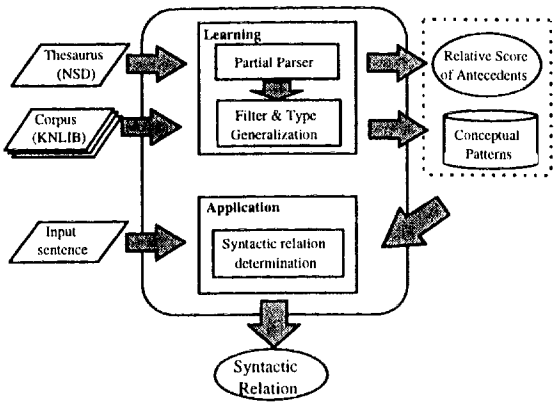

Figure 2: System architecture

relative pronouns. Instead, an adnominal verb ending follows its verb stem of a relative clause modifying an antecedent. The adnominal verb ending does not provide any information about the syntactic role of antecedent. For example, the relative clause *kang-eyse hulu-* (flow in a river) in sentence (1) modifies the antecedent *mwul-* (water), while adnominal verb ending *-nun* provides no clue about the syntactic role of the antecedent *mwul* (water). Figure 1 shows the syntactic dependency tree (SDT) of sentence (1). We need to decide the syntactic role of the antecedent *mwul-* (water) in the argument structure of the verb *hulu-* (flow) when applying case frames of the verb for structural disambiguation. The dependency parser (Lee, 1995) only gives the syntactic relation *mod* between them, which should be regarded as *subject* in the relative clause.

(1) *nanun kang-eyse hulu-***nun** *mwul-lul poatta.*
(I saw water that flowed in a river.)

As the **second characteristic**, the syntactic role of an antecedent cannot be determined by word order. This is because Korean is a relatively free word-order language like Japanese, Russian, or Finnish, and also because some arguments of a verb may be frequently omitted. In sentence (2), for example, the verb of relative clause *nolay-lul pwulless-ten* (where [I] sang a song [at the place]) have two arguments [I] and [place] omitted. Thus, the antecedent *kos-* (place) might be identified as *subject* or *adverbial* in the relative clause.

(2) *nolay-lul pwulless-ten kos-ey na-nun kassta.*
(I went to the place where [I] sang a song [at the place].)

The **third characteristic** of Korean relative clauses is that the case particle of an antecedent, that indicates the syntactic role in the relative clause, is omitted during relativization. In fact, in a relatively free-word order language, the case particles are very important to the syntactic role determination.

Due to lack of syntactic clues, it is very difficult to construct general rules for identifying the syntactic role of antecendents. Thus, the corpus-based method has been prefered to the rule-based one in solving the problem of syntactic role determination in Korean relative clauses. Yang and Kim (1993) proposed a corpus-based method, where, for each noun/verb pair, its word co-occurrence and subcategorization scores are extracted at lexical level. Park and Kim (1997) described a method of semantic role determination of antecedents using verbal patterns and statistic information from a corpus. These word co-occurrence patterns are all at lexical-level, so we have to construct a large amount of word co-occurrence patterns and statistical information before applying to a real large-scale problem. Actually, the system performance mainly relies on the domain of application, the number of word co-occurrence patterns extracted, and the size of corpus.

In the following sections, we will describe an approach to acquiring statistical information at conceptual level rather than at lexical level from a corpus using conceptual hierarchy in the Kadokawa thesaurus titled *New Synonym Dictionary* (Ohno and Hamanishi, 1981), and also describe a method of syntactic role determination using the extracted knowledge. The system architecture is shown in Figure 2.

3 Extraction of Statistic Information from Corpus

First, for each of 100 verbs selected by order of frequency in the KLIB (Korean Language Information Base) corpus of 6 million words, its syntactic relational patterns (SRPs) of the form (*Noun, Syntactic relation, Verb*) are extracted from the corpus. Then, the nominal words in the SRPs are substituted with their corresponding concept codes at level 4 of the Kadokawa thesaurus. A nominal word may have multiple meanings such as $C_1, C_2, ..., C_n$. However, since we cannot determine which meaning of the nominal word is used in a SRP, we uniformly add $\frac{1}{n}$ to the frequency of each concept code. Through this processing, the syntactic relational pattern (SRP) changes into the conceptual frequency pattern (CFP), $(\{< C_1, f_1 > , < C_2, f_2 >, ..., < C_m, f_m >\}, SR_j, V_k)$, where C_i represents a concept code at level four of the Kadokawa thesaurus, f_i indicates the frequency of the code C_i, and SR_j shows a syntactic relation between these concept codes and verb V_k. These patterns are then generalized by a concept type filter into more abstract conceptual patterns (CPs), $\{(\{C_1, C_2, ..., C_n\}, SR_j, V_k) | 1 \leq j \leq 5, 1 \leq k \leq 100\}$. Unlike in CFPs, the concept code in the more generalized CPs may be not only at level four (denoted as L_4), but also at level three (L_3) and two (L_2). In addition to the CPs, we also extract the syntactic role distributiion of antecedents.

3.1 Retrieving Syntactic Relational Patterns from Corpus

Unlike the conventional parsing problem whose main goal is to completely analyze a whole sentence, the extraction of syntactic relational patterns (SRPs) aims to partially analyze sentences and thus to get the syntactic relations between nominals and verbs. For this, we designed a partial parser, the analysis result of which is obviously not as precise as that of a full-parser. However, it can provide much useful information. For the set of 100 verbs, a total of 282,216 syntactic relational patterns (SRPs) was extracted from the KLIB corpus. During the generalization step, the problematic patterns are filtered out.

In Korean, the syntactic relation of nominal words toward a verb is mainly determined by case particles. During the extraction of SRPs (N_i, SR_j, V_k), we only consider the syntactic relation SR_js determined by 5 types of case particles: nominative (*-i/ka/kkeyse*), accusative (*-ul/lul*), and three adverbial (*-ey/eynun, -se/eyse/eysenun, -lo/ulo/ulonun*).

3.2 Conceptual Pattern Extraction

3.2.1 Thesaurus Hierarchy

For the purpose of type generalization of nominal words in SRPs, the Kadokawa thesaurus titled *New Synonym Dictionary* (Ohno and Hamanishi, 1981) is used, which has a four-level hierarchy with about 1,000 semantic classes. Each class of upper three levels is further divided into 10 subclasses, and is encoded with a unique number. For example, the class *'stationary'* at level three is encoded with the number *96* and classified into ten subclasses, Figure 3 shows the structure of the Kadokawa thesaurus.

To assign the concept code of Kadokawa thesaurus to Korean words, we take advantage of the existing Japanese-Korean bilingual dictionary (JKBD) that was developed for a Japanese-Korean MT system called COBALT-J/K. The bilingual dictionary contains more than 120,000 words, the meaning of which is encoded with the concept codes that are at level four in the Kadokawa thesaurus. Thus, Korean words in the SRPs are automatically assigned their corresponding concept codes of level four through JKBD.

3.2.2 Principle of Generalization

We encoded the nouns in SRPs extracted by the parser with concept codes from the Kadokawa thesaurus, and examined histograms of the frequency of concept codes. We observed that the frequency of codes for different syntactic relations of a verb showed very different distribution shapes. This means that we could use the distribution of concept codes, together with their frequencies as clues for conceptual pattern ex-

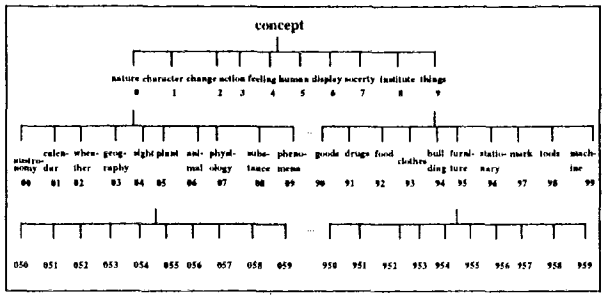

Figure 3: Concept hierarchy of Kadokawa thesaurus

traction. From the histograms of codes of both subject and object relational patterns for the verb *ttena-ta* (leave), we observed that concept codes about human (codes from 500 to 599) appear most frequently in the role of subject, and codes of position (from 100 to 109), codes of place (from 700 to 709) and codes of building (from 940 to 949) appear most often in the role of object.

For each verb V_k, we first analyzed the co-occurrence frequencies f_i of concept codes C_i of noun N, and then computed an average frequency $f_{ave,\ell}$ and standard deviation σ_ℓ around $f_{ave,\ell}$, at level ℓ (denoted as L_ℓ) of the concept hierarchy. We then replaced f_i with its associated z-score $k_{f,\ell}$. $k_{f,\ell}$ is the strength of code frequency f at L_ℓ, and represents the standard deviation above the average of frequency $f_{ave,\ell}$. Referring to Smadja's definition (Smadja, 1993), the standard deviation σ_ℓ at L_ℓ and strength $k_{f,\ell}$ of the code frequencies are defined as shown in formulas 1 and 2.

$$\sigma_\ell = \sqrt{\frac{\sum_{i=1}^{n_\ell}(f_{i,\ell} - f_{ave,\ell})^2}{n_\ell - 1}} \quad (1)$$

$$k_{f_{i,\ell},\ell} = \frac{f_{i,\ell} - f_{ave,\ell}}{\sigma_\ell} \quad (2)$$

where $f_{i,\ell}$ is the frequency of concept code C_i at L_ℓ of Kadokawa thesaurus, $f_{ave,\ell}$ is the average frequency of codes at L_ℓ, n_ℓ is the number of concept codes at L_ℓ.

3.2.3 Code Generalization

The standard deviation σ_ℓ at L_ℓ characterizes the shape of the distribution of code frequencies. If σ_ℓ is small, then the shape of the histogram will tend to be flat, which means that each concept code can be used equally as an argument of a verb with syntactic role SR_i. If σ_ℓ is large, it means that there is one or more codes that tend to be peaks in the histogram, and the corresponding nouns for these concept codes are likely to be used as arguments of a verb. The filter in our system selects the patterns that have a variation larger than threshold $\sigma_{0,\ell}$, and pulls out the concept codes that have a strength of frequency larger than threshold $k_{0,\ell}$. If the value of the variation is small, than we can assume there is no peak frequency for the nouns. The patterns that are produced by the filter should represent the concept types of extracted words that appear most frequently as syntactic role SR_i with verb V_k.

Level	Threshold of standard deviation $\sigma_{0,\ell}$					Threshold of Strength $k_{0,\ell}$
	subj	obj	adv_1	adv_2	adv_3	
L_4	2.0	8.0	0.5	0.1	0.9	$k_{0,4}$=4.0
L_3	6.0	16.0	1.5	2.0	2.0	$k_{0,3}$=1.0
L_2	30.0	50.0	15.0	4.0	10.0	$k_{0,2}$=-0.60

Table 1: Thresholds of the filter

We later analyzed the distribution of frequency f_i in CFP_js to produce an average frequency $f_{ave,\ell}$ and standard deviation σ_ℓ. Through experimentation, we decided the threshold of standard deviation $\sigma_{0,\ell}$ and strength of frequency $k_{0,\ell}$ as shown in Table 1. The lower the value of threshold $k_{0,\ell}$ is assigned, the more concept codes can be extracted as conceptual patterns from the CFPs. We maintained a balance between extracting conceptual codes at low levels of the conceptual hierarchy for the specific usage of concept type and extracting general concept types for enhancing overall system performance. These values may be variable in different application.

In Table 2, we enlist the concept types that have more than 5 appearances in the CFP of verb *ttena-ta* (leave). The strength of frequencies for generalization is calculated with formula 2.

$$k_{1,4} = \frac{1 - 0.932}{2.82513} = 0.024$$

code (freq.)	code (freq.)	code (freq.)	code (freq.)	code (freq.)	code (freq.)
061(10)	086(7)	117(5)	118(7)	158(5)	160(5)
179(5)	324(5)	410(12)	411(14)	430(16)	436(5)
480(7)	481(8)	482(9)	500(23)	501(31)	503(31)
507(35	508(30)	511(11)	513(8)	514(8)	515(5)
516(5)	519(6)	521(15)	522(19)	523(10)	525(7)
530(5)	535(6)	540(15)	550(7)	572(8)	576(9)
580(7)	581(7)	590(8)	591(5)	595(12)	814(9)
822(5)	828(5)	830(5)	833(7)	941(8)	997(7)
998(6)	other(427)				

* No. of codes: $n_4 = 932$
* Average freq.: $f_{ave,4} = 932/1000 = 0.932$
* Standard deviation: $\sigma_\ell = 2.821530$
* 'other' in the table means the total freq. of nouns less than 5
* The numbers in brackets are the frequencies of code appearance

Table 2: Concept types and frequencies in CFP ($\{< C_i, f_i >\}$,subj,*ttena-ta*)

... ...

$$k_{12,4} = \frac{12 - 0.932}{2.82513} = 3.9176$$

$$k_{14,4} = \frac{14 - 0.932}{2.82513} = \mathbf{4.626}$$

Since the value of $k_{0,4}$ is set at 4.0, as shown in Table 1, the concept codes with frequencies of more than 13, as the equation for $k_{14,4}$ shows, are selected as generalized concept types at L_4. After abstraction at L_4, the system performs generalization at L_3. It removes selected frequencies, such as frequency 14 of code 411 in Table 2, and sums up the frequencies of the remaining concept codes to form the frequency of higher level group. For example, the system removes the frequency for code 411 from the group {410(12), 411(14), 412(3), 413(0), 414(0), 415(0), 416(1), 417(0), 418(0), 419(0)}, then sums up the frequencies of the remaining codes for a more abstract code of 41. The frequency of code 41 then becomes 16. Through this process, the system performs a generalization at L_3 for the more abstract types of the concept. The system calculates σ_ℓ and strength $K_{f,\ell}$, selects the most promising codes, and stores conceptual patterns $(\{C_1, C_2, C_3, ...\}, SR_j, V_k)$ as the knowledge source for syntactic role determination in real texts, where concept type C_i is created by the generalization procedure. After generalization of the CFP patterns for the subject role of the verb *ttena-ta* (leave), the produced conceptual patterns are: ({411, 430, 500, ..., 06, 11, ..., 99, 1}, *subj, ttena-ta*).

3.3 Syntactic Role Distribution of Antecedents

In (Yang et al., 1993), they defined subcategorization score (SS) of a verb considering the verb argument structure in a corpus. They asserted that the SS of a verb represents how likely a verb might have a specific grammatical complement.

We observed from analyzing the corpus that we cannot infer the syntactic roles of antecedents from subcategorization scores since the syntactic role distribution of verb arguments in a corpus is so different from the syntactic role distribution of antecedents due to the property of free word language. In Korean, an argument of a verb could be omitted, and so the subcategorization score don't provide possible trend of the role of antecedent in many cases. For example, 26.8% of arguments of the verb *ttena-ta* (leave) are used as subjects, and 54.4% are used as objects, but 74.41% of antecedents of the verb are of subject role, and 6.9% are of object role.

Although the distribution of antecedents is necessary to our task, we cannot automatically retrieve the syntactic role distribution of them from the corpus. We extracted relative clauses for specific verbs from the corpus, and then counted the number of syntactic roles of the antecedents manually by language trained people. Since there are about 200 to 500 relative clauses for each verb in the corpus, it is possible to check this information. This information is represented by relative score $RS_k(SR_i)$ of syntactic role SR_i for antecedents of verb V_k as is shown bellow and is used in syntactic role determination as described in section 4:

$$RS_k(SR_i) = \frac{freq_k(SR_i)}{freq(V_k)} \qquad (3)$$

where $freq(V_k)$ are the frequency of verb V_k of relative clauses, and $freq_k(SR_i)$ is the frequency of syntactic role SR_i of antecedents in relative clauses including verb V_k in the corpus.

4 Identifying Deep Syntactic Relation

While determining syntactic relation for antecedents of relative clauses, the system checks the argument structure of the verb in a relative clause first, and then records the *empty* (or omitted) arguments of the verb in relative

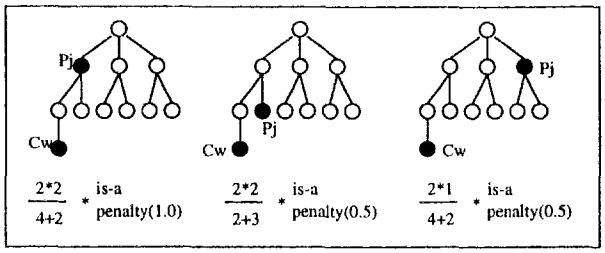

Figure 4: Conceptual similarity computation

Syntactic relation	No. of appearances	Percentage (%)	Accuracy (%)
subject	1,087	61.34%	90%
object	431	24.32%	92%
adverb(*-ey*)	121	6.82%	89%
adverb(*-eyse*)	19	1.08%	92%
adverb(*-lo*)	114	6.44%	89%
total	1,772	100%	90.4%

Table 3: The test results of syntactic role determination for antecedents

clause referring to the verb valency information. The antecedent that the verb phrase is modifying can be one of these empty arguments.

An antecedent (a noun) usually has one or more meanings, which causes ambiguity in determining the correct syntactic relation between the antecedent and a verb. We assume that an antecedent has meanings $C_1, C_2, C_3, ..., C_n$, and that CP_i is a conceptual pattern $(\{P_1, P_2, ..., P_m\}, SR_i, V_k)$ corresponding to syntactic relation SR_i of verb V_k. The evaluation score $SIM_i(N_p, V_k)$ of an antecedent N_p that can be syntactic role SR_i with verb V_k is defined as formula 4, and conceptual similarity $Csim(C_w, P_j)$ between concept C_w and P_j as formula 5.

$$SIM_i(N_p, V_k) = max(Csim(C_w, P_j)) \;\; 1 \leq w \leq n, \;\; 1 \leq j \leq m \tag{4}$$

$$Csim(C_w, P_j) = \frac{2 * level(MSCA(C_w, P_j))}{level(C_w) + level(P_j)} * is_penalty \tag{5}$$

where $MSCA(C_w, P_j)$ in $Csim(C_w, P_j)$ represents the most specific common ancestor (MSCA) of concepts C_w and P_j in the Kadokawa concept hierarchy. $Level(C_w)$ refers to the depth of concept C_w from the root node in the concept hierarchy. Is_a Penalty is a weight factor reflecting that C_w as a descendant of P_j is preferable to other cases. Conceptual similarity computation with formula 5 is shown in Figure 4.

Based on these definitions, the syntactic relation SR_j between antecedent N_p and verb V_k can be calculated as follows:

1. Let R = $\{SR_i | SR_i$ is a syntactic relation of an *empty* (or omitted) argument in the relative clause of V_k, $1 \leq i \leq 5\}$.
2. For each conceptual pattern CP_i of verb V_k of which SR_i is in R, and for each concept code P_i in CP_i, compute $SIM_i(N_p, V_k)$.
3. Determine the syntactic relation of antecedent N_p to SR_j on the condition that $SIM_j(N_p, V_k)$ has the largest value in $\{SIM_i(N_p, V_k) | 1 \leq i \leq 5\}$ and SR_j in R. If two or more $SIM_i(N_p, V_k)$ have the same value, decide syntactic role referring to the higher relative score $RS_k(SR_i)$ of the syntactic role of the verb V_k.

Here, syntactic relation can be one of *subj, obj, adv1, adv2,* and *adv3*. The symbols *adv1, adv2,* and *adv3* represent adverbs with case particles *-ey, -eyse,* and *-lo*, respectively.

5 Experimental Evaluation

An informal way to evaluate the correctness of syntactic relation determination is to have an expert examine the test patterns and source sentences that the patterns appears, and give his/her judgment about the correctness of the results produced by the system. In our experiment, the correctness of syntactic and conceptual relation determination was evaluated manually by humans who were well trained in dependency syntax.

As a test set, we extracted 1,772 sentences that included relative clauses for the 100 verbs from 1.5 million word corpora of integrated Korean information base and test books of primary school. The distribution of syntactic relation of antecedents among them and the test results were shown in Table 3. There were 1,087 antecedents (61.34%) that were of subject role. The baseline accuracy of the problem is 61.34%. That is, if we always select subject role for antecedents, the accuracy will reach 61.34%.

Our system showed 90.4% of accuracy on average in syntactic relation identification, which shows that the conceptual patterns and relative score of syntactic relation produced in the first phase can be a good source for determining the syntactic relation of an antecedent.

Through experiment, we observed several factors that affect the performance of the system. First, the multiple meanings of a noun will affect the frequency distribution of concept codes. In our system, we cope with this problem by adjusting the threshold of standard deviation and strength value. The second problem is the sparseness of corpus domain. If the corpus for learning is specified as a certain domain, it will greatly increase the validity of conceptual patterns. If we use a sense tagged corpus in the learning stage, we can achieve high accuracy in syntactic relation determination.

6 Concluding Remarks

This paper describes an approach for syntactic role determination between an antecedent and a verb in relative clause for semantic analysis. This method consists of two phases. In the first phase, the system extracts conceptual patterns and syntactic role distribution of antecedents from a large corpus. In the second phase, the system applies the extracted conceptual patterns as knowledge in determining correct syntactic relations for structural disambiguation and semantic analysis in MT system for CG generation.

Unlike previous research that calculates statistical information at a lexical level for every pair of words, which may require a lot of space to store resulting patterns, we represent those co-occurrence patterns with concept types of Kadokawa thesaurus. The problematic concept types are filtered out by the type generalization procedure. We used a corpus of 6 million words for conceptual pattern extraction. Our method can cope with the general scope of texts. In the experiment evaluation, the proposed method showed a high accuracy rate of 90.4% in identifying the syntactic role of antecedents.

The method described in this paper can be used in resolving syntactic role of antecedents in relative clauses of other free word order languages, and can also be used in generating selectional restrictions of case frames of verbs.

References

Lee, J. H. and G. Lee. 1995. A Dependency Parser of Korean based on Connectionist/Symbolic Techniques. *Lecture Notes on Artificial Intelligence 990*, pages 95-106. Springer-Verlag, Berlin.

Li, H. F., J. H. Lee and G. Lee. 1998. Conceptual Graph Generation from Syntactic Dependency Structures in an MT Environment. (to be published by *Computer Processing of Oriental Languages* in 1998).

Ohno, S. and M. Hamanishi. 1981. *New Synonym Dictionary, Kadokawa Shoten*, Tokyo (written in Japanese).

Park, S. B. and Y. T. Kim. 1997. Semantic Role Determination in Korean Relative Clauses Using Idiomatic Patterns. In *Proceedings of 17th International Conference on Computer Processing of Oriental Languages*, pages 1-6. Hong Kong.

Smadja, F. 1993. Retrieving Collocations from Text: Xtract, *Computational Linguistics*, 19(1):143-177.

Yang, J. and Y. T. Kim. 1993. Identifying Deep Grammatical Relations in Korean Relative Clauses Using Corpus Information. In *Proceedings of Natural Language Processing Pacific Rim Symposium '93*, pages 337-344. Tae-Jon, Korea.