개념패턴과 통계정보를 이용한 한국어 미지격의 구문관계 결정 방법

이휘봉, 강인수, 이종혁
포항공과대학교 전자계산학과

Resolution of Ambiguous Grammatical Functions of Korean Using Conceptual Patterns and Statistical Information

Hui-Feng Lee, Insu Kang, Jong-Hyeok Lee
Dept. of Computer Science and Engineering, POSTECH
{hflee, dbaisk, jhlee}@madonna.postech.ac.kr

요약

본 논문은 보조사로 인해 야기되는 한국어 미지격의 구문관계 중의성 해소를 위한 새로운 기법을 제안한다. 기존의 연구는 수작업으로 얻어진 동사의 의미적 선택 제약을 사용하는 방식과 단어 간의 공기패턴과 빈도를 어휘 레벨에서 추출하여 중의성을 해소하는 방식으로 나뉠 수 있다. 본 논문은 말뭉치에서 어휘 레벨이 아닌 개념패턴과 격의 분포 값을 자동으로 추출하여 미지격의 구문관계를 결정한다. 개념패턴과 용언의 격 분포 정보를 적용하여 구문분석 단계에서 실험한 결과, 본 논문이 제안한 방법은 92%의 미지격 결정 정확율을 보였다. 개념패턴은 지식의 저장공간을 줄이고 격 결정 범위를 확장할 수 있기에 범용 구문분석 시스템으로의 확장을 가능하게 한다.

1 서론

한국어의 보조사 '은/는', '만', '도' 등은 명사구의 격 결정 애매성을 가져온다. 보조사의 사용은 아래와 같은 실제의 문장들을 예문으로 보인다.

(1) 이 비평은 옥편의 필요성도 강조하고 있다.
(2) 이것을 통하여 민족적 단결도 이루어 왔다.
(3) 영희의 특정한 전문성은 전체성의 특징도 지니고 있다.
(4) 영희는 하루종일 책만 읽는다.

보조사에 의해 야기된 명사의 문법기능 중의성을 해결하기 위하여 기존 연구들은 일차적으로 동사의 하위범주화 정보 (subcategorization information)를 사용하고, 여기에 의미 표지인 의미적 선택 제약을 사용하였다[2, 5, 6]. 이러한 방법에서 의미표지와 하위범주화 정보는 실제 말뭉치에서의 사용을 고려한 것이 아니고, 또한 지식베이스나 의미표지체계의 구축에 수작업으로 인한 엄청난 노력 및 시간이 필요하며, 적용영역의 확장이나 이전에 따른 지식베이스의 재구축, 지식베이스 코딩 과정에서의 일관성 유지의 어려움으로 인하여 실제 적용이 어려워 진다. [4]에서는 의미정보를 이용하는 대신 대량의 말뭉치로부터 통계적 관련성 정보(명사와 동사 간의 공기정보, 동사의 하위 범주화 정보)를 사용한 방법을 제안하였다. 이 방법은 지식을 자동으로 추출할 수 있어서 지식베이스의 구축을 용이하게 하는 반면, 문제점으로는 주로 모든 통계정보를 미리 처리하여 저장하는데 거대한 저장공간이 소요된다는 것이다. 즉 [명사, 동사, 문법관계]로 표현되는 어휘 레벨(lexical level)의 통계 정보를 저장하여야 하는데, 이 트리플의 개수는 전체 단어 수 즉 사전의 엔트리 수보다 훨씬 크게 된다. 따라서 넓은 적용영역으로의 확장이 어려워 진다.

본 논문에서는 기존의 수작업으로 의미표지를 작성하는 어려움과 실제 어휘들 간의 공기정보를 이용하는 통계적 방법의 제한점들을 극복하는 방법으로 개념패턴과 통계적 정보를 이용하

는 미지격 문법관계의 2 단계 결정방법을 제안한다. 첫번째 단계는 개념패턴과 통계정보를 추출하는 학습단계이다. 두번째 단계는 습득된 지식을 적용하여 실제 구문분석 과정에서 사용하는 적용단계이다.

다음으로 2 장에서는 개념패턴과 동사의 격분포값의 추출 방법을 소개하고, 3 장은 추출된 지식을 이용한 미지격 결정 방법을 기술하고, 4 장에서 실험과 평가, 그리고 5 장에서는 결론과 앞으로의 연구 방향을 기술한다.

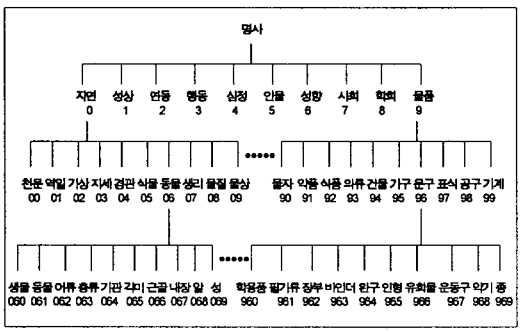
2 개념패턴과 통계정보의 추출

2.1 개념패턴의 추출

2.1.1 어휘 공기 패턴의 추출

미지격 명사가 동사와 어떤 구문적 역할을 하는가는 그 명사가 격조사와 함께 동사와 실제 문장에서 사용될 때 어떤 구문적 역할을 하는가를 관찰하면 판단이 가능하다. "철수가 책을 읽고 있다"의 문장에서 "*철수*"는 동사 "*읽다*"와 결합하여 주격으로, 그리고 "*책*"은 "*읽다*" 결합하여 목적격으로 사용되고 있다. 말뭉치에서 이러한 격조사가 탈락되지 않은 문장들에서 동사와 특정한 구문관계로 사용되는 공기단어들을 수집하여, 명사들을 추상된 개념으로 표현하면 실제 미지격 중의성을 해소할 수 있다. 본 논문에서는 이러한 추상화된 (명사의 개념, 구문관계, 동사) 패턴을 개념패턴으로 간주한다.

명사는 동사와 주격, 목적격, 부사격으로 쓰일 수 있는데, 부사격은 보조사를 사용하여 강조의 뜻을 표현하는 경우가 적으므로 해결 대상에서 제외하였다. 본 논문에서는 개념패턴의 자동습득을 위하여 부분 파싱 기법을 도입하여 문장을 분석하고 말뭉치에서 빈도수가 높은 84 개의 동사와 명사간의 구문적 패턴(syntactic relational pattern, SRP) (명사, 구문관계, 동사)을 먼저 추출한다. SRP 의 구문관계는 주격조사 (이/가/께서), 목적격조사(을/를)들을 이용하여 판단한다. 개념패턴의 학습 단계에서 사용된 말뭉치는 국어 정보 베이스(Korean National Language Information Base, KNLIB)의 600 만 어절의 품사 태깅되지 않은 말뭉치를 이용하였다. 말뭉치에서 빈도수가 높은 84 개의 동사에 대하여 추출된 2 종류(주어, 목적어)의 SRP 패턴은 5,138,000 개에 달하였다.



[그림 1] NSD 개념 계층 구조

2.1.2 개념패턴으로의 일반화 원칙

개념패턴으로의 일반화(generalization)를 위하여 類語新辭典 (New Synonym dictionary, NSD)[3]의 개념 계층 구조를 사용한다. NSD 구조는 [그림 1]와 같은데, 본 논문에서는 NSD 의 계층을 상위(top) 레벨로부터 L_1, L_{10}, L_{100}, L_{1000}로 표기하고, 레벨 L_{1000}, L_{100}과 L_{10}에서 명사와 동사간의 개념패턴을 추출한다. NSD 의 레벨 L_{1000}에는 1000 개의 개념 코드가 존재하고, L_{100}에는 100 개의 코드가 존재한다. 한국어에 대한 NSD 코드는 본 실험실에서 개발한 일-한 기계번역 시스템 COBALT-J/K 의 어휘사전을 이용한다. 한국어 단어의 이와 같은 코드 표시는 구문관계 패턴(syntactic relational pattern, SRP)의 일반화를 자동화할 수 있게 한다.

일반화 처리를 시작하기 전에 지정된 동사의 각 SRP 의 명사들에 대하여 자동으로 NSD 코드를 표기하고, 같은 패턴에 동일한 코드가 여러 번 나타나면 출현 빈도를 더하여 개념 빈도 패턴(conceptual frequency pattern, CFP)을 생성한다. 하나의 명사가 시소러스 상에서 복수 개의 개념을 나타내는 경우는 명사가 지니고 있는 개념의 수를 나눈 빈도수 값을 증가한다. 즉, 특정 명사가 개념 코드 C_1, C_2, ..., C_n을 가지면 각각의 코드 빈도에 $\frac{1}{n}$를 증가시킨다. 이것은 다의어의 잡음(noise) 생성을 방지하기 위함이다. 이로서 L_{1000}에서의 초기의 의미 코드의 집합이 얻어진다. CFP_j는 ({<C_1, f_1>, <C_2, f_2>, ..., <C_n, f_n>}, SR_j, V_k)와 같은 양식을 가지며, C_i는 개념 코드를, f_i는 개념 코드 C_i의 출현 빈도를, 그리고 SR_j는 C_i가 동사 V_k와 문장에서 가지는 구문관계를 가리킨다.

일반화 처리는 이러한 CFP 들을 분석하여, 바닥 레벨, 즉 L_{1000} 에서부터 시작한다. 개념 코드가 바닥 레벨 L_{1000}에 가까울수록 협의의 개념을 나타내고, L_1에 가까울 수록 광의의 개념을

나타낸다. 따라서 일반화에서 가장 중요한 원칙은 바닥 레벨에서 일반화가 가능하면 그것이 더욱 선호된다는 것이다. 선택된 개념은 단어의

[표1] 일반화를 위한 필터의 임계치

Level	표준편차 $\sigma_{0,\ell}$의 임계치		$k_{0,\ell}$의 임계치
	subj	obj	
L_{1000}	2.0	8.0	$k_{0,1000} = 4.0$
L_{100}	6.0	16.0	$k_{0,100} = 1.0$
L_{10}	30.0	50.0	$k_{0,10} = -0.6$

협의의 사용법을 최대한 표현하는 것이다. 이 원칙에 의해서 L_{1000} 에서 일반화를 우선 시도하고, 일반화가 불가능한 것들은 L_{100}과 L_{10}에서 보다 광의의 개념을 추출한다. 두번째 원칙은 각 레벨에서의 공기 패턴으로 선택될 의미 코드 후보는 그 레벨에서의 다른 의미코드들과의 경쟁을 통해서 결정된다는 것이다. 따라서 절대적인 빈도수보다는 일반화의 후보로 되는 개념의 빈도수의 상대적인 분포 모양에 따라서 추출 여부가 결정된다.

한 동사의 CFP는 구문관계에 따라서 서로 다른 분포의 모양을 나타낸다. 예를 들어, 동사 "*떠나다*"(leave)의 말뭉치에서 추출한 주격과 목적격의 CFP을 이용하여 NSD의 개념 코드를 X 축으로 하고, 개념 코드의 출현 빈도수를 Y축으로 하는 히스토그램을 작성하여 살펴보면, "*떠나다*"의 주격으로는 사람에 관한 코드(500~599)가 많이 나타나고, 목적격으로는 위치(100~109), 장소(700~709), 건물(940~949) 등에 관한 코드가 많이 나타난다.

개념패턴으로의 일반화를 위하여 동사 V_k에 대하여 레벨 L_ℓ에서 먼저 공기하는 명사 N의 개념 C_i의 빈도수 f_i의 분포를 분석하고, 평균 빈도수 $f_{ave,l}$와 $f_{ave,l}$를 중심으로 한 표준 편차 σ_ℓ를 계산한다. 다음으로 f_i를 관련된 z-score인 $k_{f,\ell}$로 대체한다. $k_{f,\ell}$ 은 코드 빈도 f_i의 레벨 L_ℓ에서의 상대적 강도(strength)이고, 코드 빈도 f_i가 개념 레벨 L_ℓ에서 평균 빈도수 $f_{ave,l}$ 보다 몇 배나 더 큰가를 표시한다. 본 논문에서는 [1, 9]의 정의를 참조하여 L_ℓ에서 σ_ℓ과 $k_{f,\ell}$을 [수식 1]과 [수식 2]로 정의한다.

$$\sigma_\ell = \sqrt{\frac{\sum_{n=1}^{n_\ell} \left(f_{n,\ell} - f_{ave,\ell}\right)^2}{n_\ell - 1}} \qquad \text{[수식1]}$$

$f_{n,\ell}$:NSD의 개념코드 n이 L_ℓ의 빈도수
$f_{ave,\ell}$:개념코드가 L_ℓ에서의 평균 빈도수
n_ℓ :L_ℓ에서의 개념코드의 개수

$$k_{f_{n,\ell},\ell} = \frac{f_{n,\ell} - f_{ave,\ell}}{\sigma_\ell} \qquad \text{[수식2]}$$

빈도수에 대한 표준 편차 σ_ℓ은 L_ℓ에서 코드 빈도의 분포 모양을 나타낸다. σ_ℓ이 작으면 코드 빈도 분포의 히스토그램은 평평한 모양을 나타내는데, 이것은 L_ℓ에서의 모든 개념들이 동사와 개념 관계 SR_i를 가질 때 선호도(preference)에 큰 차이가 없다는 것을 의미한다. 반면, σ_ℓ가 크면 L_ℓ의 히스토그램에 하나 혹은 몇 개의 높은 빈도의 개념 코드가 존재한다는 것을 의미한다. 이러한 개념을 가진 명사는 동사와 자주 공기하기 때문에 개념 코드를 추출하여 저장해 둘 필요가 있다. 일반화는 이러한 개념들을 각 레벨에서 추출하는 과정이라 할 수 있다.

2.1.3 개념 코드의 일반화 과정

정확한 개념 코드를 찾아 내기 위하여 실험을 통하여 표준편차의 임계치 (threshold) $\sigma_{0,\ell}$와 코드 출현 빈도의 strength의 임계치 $k_{0,\ell}$를 [표 1]과 같이 지정한다. $k_{0,\ell}$의 값을 작게 지정하면 L_ℓ에서 더 많은 개념 코드가 추출된다. $\sigma_{0,\ell}$와 $k_{0,\ell}$의 값은 낮은(low) 레벨 개념 계층에서 뽑은 개념들의 수와 좀 더 높은 개념 계층에서 뽑는 개념의 수 사이의 균형을 기준으로 한다. 낮은 계층에 있는 개념 코드의 추출은 많이 사용되는 특정한 개념들의 공기 패턴을 중시하고, 높은 계층의 개념들은 전체 시스템의 구문관계 결정 성능에 영향을 줄 수 있다. 임계치를 지정할 때, 구문관계가 다르고 일반화하는 개념 레벨이 다르면 값이 서로 다를 수 있다.

[표 2]에는 말뭉치에서 추출한 동사 "떠나다"의 개념 빈도 패턴(CFP)의 값들이다. [표 2]에 등록된 앤트리(entry)들은 해당 코드가 8번 이상 나타난 개념 코드와 출현 빈도이다. 출현 빈도가 8보다 적은 코드들의 빈도는 other에 지정하였다. [표 2]를 근거하여 $k_{n,\ell}$를 계산하면 다음과 같다.

$$k_{1,1000} = \frac{1 - 0.932}{2.82513} = 0.024$$

... ...

$$k_{12,1000} = \frac{12 - 0.932}{2.82513} = 3.9176$$

$$k_{14,1000} = \frac{14 - 0.932}{2.82513} = \mathbf{4.626}$$

[표2] "떠나다"의 주격 개념 빈도 패턴(CFP)

code(f)	code(f)	code(f)	code(f)	code(f)
061(10)	410(12)	**411(14)**	**430(16)**	481(8)
482(9)	**500(23)**	**501(31)**	**503(31)**	**507(35)**
508(30)	511(11)	513(8)	514(8)	**521(15)**
522(19)	523(10)	**540(15)**	572(8)	576(9)
590(8)	595(12)	814(9)	other(571)	

개념 코드의 총빈도수: 932
평균 빈도 : $f_{ave,\ 1000} = 932/1000 = .932$
표준 편차: $\sigma_t = 2.825130$
* 'other' 는 코드의 출현 빈도가 8 보다 적은 것들의 합
* 괄호 내의 수치는 코드의 출현 빈도

[표 1]에서 임계치 $k_{0,1000}$를 4.0 으로 지정하였기에, $k_{14,1000}$의 값이 이 임계치(threshold)를 만족한다. 따라서 개념 코드의 출현 빈도가 14 이상이면 L_{1000}에서 공기하는 개념 코드로 뽑힌다. L_{1000}에서 조건을 만족하는 개념 코드를 선택해 낸 후에 그 코드와 빈도를 해당 그룹(같은 레벨 코드의 마감 자리만 다르고 기타는 같은 것들. 예: 411, 412, 413,...)에서 제외시킨다. 예를 들면 위의 [표 2]에서 코드 411 을 선택한 후 코드 411 과 코드의 빈도 14 를 그룹 {410(12), 411(14), 412(3), 413(0), 414(0), 415(0), 416(1), 417(0), 418(0), 419(0)}에서 제외시킨다. 다음으로 이 그룹들을 상위 레벨 개념과 빈도로 추상화 한다. 위의 개념 410 으로부터 419 까지의 개념들은 상위 그룹 41 로 추상화되고, 하위 개념들의 빈도의 합인 16 을 개념 코드 41 의 빈도로 부여한다. L_{1000} 에서의 작업이 끝난후 L_{100} 에서 $k_{n,t}$의 값이 $k_{0,100}$ 값보다 큰 개념 코드를 찾아냄으로서 높은 레벨에서 일반화를 시도한다. 이렇게 찾아낸 개념 C_i는 구문관계 SR_j, V_k와 같이 개념패턴(conceptual pattern, CP)으로 저장해서 구문관계를 계산할 때 사용한다. 이러한 과정을 통해 얻어진 동사 "*떠나다*"의 주격에 대한 개념패턴은 ({411, 430, 500,..., 06, 11,..., 99, 1}, subj, 떠나다)이다. 여기에서 표시된 개념 코드들은 L_{1000}, L_{100}, L_{10} 레벨의 개념들로 구성된다.

2.2 동사의 격 분포 통계정보의 추출

명사의 구문관계를 결정하기 위하여 앞절에서 추출한 개념패턴을 이용할 뿐만 아니라, 동사 V_k의 말뭉치에서 추출한 격의 분포 값을 이용할 필요가 있다. 어떤 동사들은 주격과 목적격에 대한 제약이 모두 미약함으로, 개념패턴으로는 미지격 결정이 어렵다. 이러한 동사들의 예문으로 "철수가 영희를 보았다"를 들 수 있다. 이러한 경우는 말뭉치에서 동사의 격 분포 값 ([4]에서는 동사의 하위범주화 정보)을 이용하여 결정한다. 격의 분포(case distribution, CD) 값 $CD_{gr}(V_k)$의 정의는 [4]의 정의를 따른다. 즉,

$$CD_{gr}(V_k) = \frac{freq_{gr}(V_k)}{freq(V_k)},\ \text{gr} \in \{\text{subj, obj}\}$$ [수식3]

[수식 3]에서 $freq_{gr}(V_k)$는 동사 V_k가 말뭉치내에서 구문관계 gr를 보어로 가지는 빈도이고, $freq(V_k)$는 동사 V_k가 말뭉치에서 사용된 빈도이다. CD 값은 동사가 주어 혹은 목적어를 가질 가능성이 얼마나 큰가를 보여준다.

3 미지격의 결정 방법

문장에서 미지격 명사 구문관계를 결정할 때, 먼저 문장내의 동사와 결합할 수 있는 격 구조를 분석하고 동사의 결합가 정보를 참조하여 비어있는 매개 변수(argument)를 기록해 둔다. 이런 빈 변수(empty argument)는 미지격 명사가 가질 수 있는 격들이다.

명사는 때로는 여러 의미를 가질 수 있는데, 명사의 구문관계를 결정할 때 고려하여야 한다. 명사 N_p의 의미를 $C_1, C_2, C_3, ..., C_n$이라 하고, CP_i는 구문관계 SR_i를 결정할 수 있는 개념패턴 ($\{P_1, P_2, ..., P_m\}$, SR_i, V_k)을 가르킨다. 개념패턴과 구문 의존 트리에 기반한 명사 N_p와 동사 V_k사이에 구문관계 SR_i를 결정하는 평가치 $SIM_i(N_p, V_k)$는 [수식 4]로 정의한다. 그리고 개념 C_w와 CP_i의 제약을 표현한 개념 P_j간의 개념 유사도는 [수식 5]에서 정의한 $Csim(C_w, P_j)$으로 계산한다.

$$SIM_i(N_p, V_k) = \max(Csim(C_w, P_j)),\quad 1 \le w \le n,\ i \in \{\text{subj, obj}\}$$ [수식 4]

$$Csim(C_w, P_j) = \frac{2 * Level(MSCA(C_w, P_j))}{Level(C_w) + Level(P_j)} * penalty$$

[수식 5]

[수식 5]에서 *MSCA*(most specific common ancestor)는 NSD 개념 계층 구조에서 두 개념간의 가장 가까운 상속 노드를 가르키고, *Level*(C_w)는 개념 C_w가 NSD 개념 계층 구조에서의 위치를 가리킨다. *penalty*는 개념의 가중치를 표시함으로, [그림 2]에서 표시하듯이 명사의 개념 C_w가 개념 P_j의 하위 개념이면 *penalty*를 1로(penalty가 없음) 지정하고, 그렇지 않으면 0.5의 값을 지정함으로서 유사도 값을 감소시킨다.

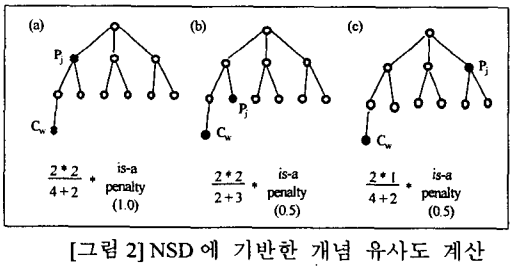
위의 정의를 따라 명사 N_p의 V_k와의 구문관계 결정 알고리즘은 다음과 같다.

[S1] 초기화: $R = \{SR_i \mid SR_i$는 문장에서 동사 V_k의 비어(empty) 있는 격$\}$

[S2] 개념 유사도 계산: 동사 V_k에 대하여 집합 R의 SR_i를 위한 개념패턴인 CP_i, 그리고 CP_i에 있는 임의의 개념 코드 P_j에 대하여, $SIM_i(N_p, V_k)$를 계산한다.

[S3] 명사의 구문관계 결정: 계산된 모든 유사도 $\{SIM_i(N_p, V_k) \mid 1 \le i \le n\}$중에서 가장 큰 $SIM_i(N_p, V_k)$를 가지는 구문관계 SR_i를 명사 N_p가 동사 V_k와의 구문관계로 결정한다. 만약 두개의 $SIM_i(N_p, V_k)$가 같으면, 동사 V_k의 구문관계 분포 값 $CD_{gr}(SR_i)$를 참조하여, $CD_{gr}(SR_i)$ 값이 큰 구문관계 gr를 선택한다.

여기서 명사가 가질 수 있는 구문관계는 주격, 목적격 중의 하나이다.



[그림 2] NSD에 기반한 개념 유사도 계산

4 실험 및 평가

실험을 위하여 울산대 100만 어절 말뭉치에서 3절에 사용된 84개의 동사에 대하여 보조사를 포함한 284 문장을 임의적으로 추출하였다. "교과서가 말하다"와 같이 명사들이 비정상적으로 사용된 예문은 실험에서 제외시켰다. 실험용 문장에는 각 동사마다 미지격이 포함된 3~4개의 문장으로 구성되었다. 실험문장에 대하여, 본 논문이 제안한 방법을 구문관계 결정에 적용한 결과 92%의 높은 미지격 결정 정확율을 보였다. 실험과정에서 "타다", "묻다", "그리다"와 같이 중의성이 있는 동사들과 결합되는 미지격 결정의 정확도가 낮았다. 이러한 동형이의어에 대하여 동사의 의미를 구분하여 개념패턴을 작성하면 높은 정확율을 기대할 수 있는데, 현재 의미태깅된 한국어 말뭉치가 없는 상황에서 이러한 작업은 어렵다고 본다.

5 결론

본 논문은 보조사로 인해 발생하는 미지격 명사의 구문관계를 결정하기 위하여 개념패턴과 통계정보를 이용한 방법을 제안하였다. 이 방법은 개념패턴을 말뭉치에서 자동으로 추출함으로서 기존의 의미제약 작성의 비효율성을 극복하고, 어휘 레벨의 공기정보를 추출하는 통계적 방법의 저장공간 문제점을 효율적으로 극복할 수 있다. 개념패턴을 이용함으로서 넓은 범위에서의 구조적 모호성 해소도 가능하게 하였다. 이러한 기법은 대량의 말뭉치를 요구하고, 동사가 각 매개 변수에 대한 개념의 요구성이 달라져야만 효율성이 높아질 수 있다. 그리고 한국어 문장에서 의존명사가 미지격으로 사용되는 경우 격결정의 해소가 더욱 어려워 진다. 이러한 문제는 향후의 연구 과제이다.

참고문헌

[1] 곽종근, "일본어 코퍼스로부터 동사-명사 연어패턴의 자동 추출," 포항공대 석사학위논문, 1996.

[2] 나동렬, "한국어 파싱에 대한 고찰," 정보과학회지, Vol. 12, No. 8, pp.33-46, 1994.

[3] 大野晋十, 浜西正人, *類語新辭典*, 角川書店, 東京, 1981(일본어).

[4] 양재형, 김영택, "통계 정보를 활용한 한국어 미지격 명사구의 구문관계 결정," 정보과학회 논문지, Vol. 21, No.5, pp. 808-815, 1994.

[5] 윤덕호, 김영택, "미지문법관계 속성을 이용한 LFG에서의 한국어 문장분석 연구," 정보과학회 논문지, Vol. 16, No. 9, pp.434-444, 1989.

[6] 조혁규, 권혁철, "단일화와 챠트를 이용한 한국어 구문분석시스템의 구현," 정보과학회 논문지, Vol. 17, No. 4, 1990.

[7] Jong-Hyeok Lee, Geunbae Lee, "A Dependency Parser of Korean Based on Conectionist/Symbolic Techniques," Lecture Notes on Artificial Intelligence 990, Springer-Verlag, Berlin, pp. 95-106, 1995.

[8] Hui-Feng Lee, Jong-Hyeok Lee, Geunbae Lee, "Identifying Syntactic Role of Antecedent in Korean Relative Clause Using Corpus and Thesaurus Information," in proc. of COLING-ACL'98, University of Montreal, pp. 756-762, 1998.

[9] Frank Smadja, "Retrieving Collocations from Text: Xtract," Computational Linguistics, Vol. 19, No. 1, pp. 143-177, 1993.