X-바 이론을 변형한 자질기반의 한국어 구구조 문법

박소영 황영숙 정후중 곽용재 임해창

고려대학교 컴퓨터학과 자연어처리연구실
(136-701)서울특별시 성북구 안암동 5가 1
{ssoya, yshwang, hjchung, yjkwak, rim}@nlp.korea.ac.kr

Feature-based Korean Phrase Structure Grammar adjusting X-bar Theory

So-Young Park, Young-Sook Hwang, Hoojung Chung, Yong-Jae Kwak, Hae-Chang Rim
NLP. Lab., Dept. of Computer Science & Engineering, Korea University

요 약

본 논문에서는 X-바 이론을 한국어에 적용하여 서로 다른 범주들간의 구조적 일반성을 파악하고, 한국어에 가능한 규칙만을 허용하여 불가능한 규칙을 배제시킬 수 있는 틀을 제시하고자 한다. 한국어가 비중심어간 어순이 자유롭고 기능어가 발달했다는 점을 고려하여, 중심어와 보충어 관계 중심의 기존 X-바 이론을 통사적 파생과 의미적 파생, 수식 및 하위범주의 관계를 중심으로 변형한다. 또한, 한국어의 빈번한 생략현상과 부분 자유 어순에 효과적으로 대응할 수 있도록 이진결합 중심의 CNF(Chomsky Normal Form)를 따른다. 제안하는 자질기반의 한국어 구구조 문법은 직관적이고 간단하면서도 대부분의 문장을 처리할 수 있을 만큼 표현력이 뛰어나다는 장점이 있다. 신문기사에서 454문장을 추출하여 실험한 결과, 약 97%의 문장에 대해 올바른 구문 분석 결과를 생성할 수 있음을 보였다.

1. 서론

한국어는 하나의 형태소가 하나의 문법적 기능을 수행하며, 의미를 나타내는 실질형태소에 기능을 나타내는 형식형태소가 첨가되어 하나의 어절을 구성한다는 특징이 있다[20]. 또한, 한국어에서 문장성분은 위치정보가 아니라 기능어에 의해 결정되므로, 기능어가 발달하고, 어순의 제약이 적으며, 문장성분의 생략현상이 빈번하게 일어난다.

이러한 특성을 고려하여 한국어에 적합한 문법을 개발하고자 하는 많은 연구들이 진행되어 왔다[8,15,16,17,21]. 박성숙[8]과 윤덕호[16]는 문법범주와 자질로 구성된 복합토큰을 이진 형식의 규칙에 적용하는 방법을 제시하였다. 복합토큰에서 문법범주는 실질형태소에 의존하고 자질은 형식형태소에 의존하여 결정되도록 하였다. 이러한 방법에 대해서 윤덕호[16]가 의존문법으로 접근한 반면, 박성숙[8]은 일반 구구조 문법으로 접근하였다. 그러나, 하나의 토큰이 여러 개의 형태소에 의해 복합적으로 구성되므로, 하나의 규칙에서 여러 기능을 함축적으로 처리하도록 하기 위한 부담이 따른다.

이공주[17]는 형태소를 구문 최소 단위로 이용하는 구구조 문법을 제안하면서 하나 이상의 어절에 영향을 끼칠 수 있는 기능어의 역할을 강조하였다. 구구조 문법을 이용하면 범주간의 관계와 위상을 효과적으로 설명할 수 있을 뿐만 아니라[10], 구구조 규칙만을 이용하므로 모델에 필요한 인자수가 많지 않다[2]는 장점이 있다. 그러나, 영어권에서 사용하는 구구조 규칙의 형태를 어순의 제약이 적은 한국어에 그대로 적용할 경우, 규칙의 길이가 길어지고 규칙의 수가 기하급수적으로 증가할 뿐만 아니라 적용될 규칙을 선택하기 위한 부가적인 처리가 필요하다는 문제점이 있다.

한편, Briscoe와 Waegner[3]는 X-바 이론의 중심어 개념을 도입하여, 직접적 지배관계로 표현되는 자질값을 계승·통합하여 구문분석하는 방법을 제시하였다. 즉, 문법에 대한 언어학적 제약을 도입하여 문법학습시 문법의 생성능력은 제한하지 않으면서 언어학적으로 이해할 수 없는 규칙

을 제거할 수 있게 하였다.

본 논문에서 제안하는 자질기반의 한국어 구구조 문법은 기능어의 역할을 충분히 살릴 수 있도록 형태소를 기본 단위로 하며, 한국어의 빈번한 생략현상과 부분 자유 어순에 효과적으로 대응할 수 있도록 규칙을 이진 형식으로 제한한다. 또한, X-바 이론을 한국어에 적용하여 서로 다른 범주들간의 구조적 일반성을 파악하고 한국어에 가능한 규칙만을 허용하고 불가능한 규칙을 배제시킬 수 있는 틀을 제시한다. 이는 규칙의 기하급수적인 증가를 방지하고, 하나의 규칙에서 하나의 문법적 기능을 표현하며, 한국어의 빈번한 생략현상과 부분 자유 어순의 특징을 반영한 것이다.

2. X-바 이론을 변형한 자질기반의 한국어 구구조 문법

Chomsky가 동사구와 명사구의 유사성을 포착하기 위해서 변항 X를 고안하면서부터 출발한 X-바 이론은 중심어와 보충어 관계를 효과적으로 규명하여 문장 구조 형성 원리를 가장 간결하고 명료한 형태로 제시하여 준다[9,10]. 세부적인 내부구조에 대해 여러 이론들의 견해가 서로 다르지만, 중심어인 X를 기준으로 그 중심어를 최대로 투사하여 결합한다는 원리는 거의 모든 문법이 일치한다[11].

본 논문은 한국어의 특성에 적합하도록 통사적 파생, 의미적 파생, 수식 및 하위범주 관계를 중심으로 X-바 이론을 변형시킨 방법을 제안하고자 한다.

2.1 기본 자질의 구성

기존의 X-바 이론은 구문분석의 기본 범주인 명사와 동사, 형용사, 전치사를 명사성 자질 ±N과 동사성 자질 ±V의 복합체[1] 표현함으로써 범주간의 통사적 일반성을 파악하여 자연어의 보편적인 범주들을 분류한다[11].

한국어 구문범주의 일반성을 파악하기 위해서는 한국어에 적합한 자질을 구하는 것이 무엇보다 중요하다고 할 수 있다. 그러므로 한국어 문장 구성시 핵심적인 역할을 하는 체언과 용언을 중심으로 기능어의 역할까지 고려될 수 있어야 한다. 따라서 본 논문에서 제안하는 자질기반의 한국어 구구조 문법의 구문범주는 기본적으로 체언자질(N), 용언자질(V), 바자질(BAR)에 의해 구성된다.

V \ N	n-	n	n+
v-	IC	_NC	DC
v	VC		
v+	CAS	NC	AC

<표1>구문범주의 체언·용언 자질

본 논문에서는 체언자질과 용언자질을 바탕으로 하여, 한국어의 구문범주를 체언상당어구(NC)와 체언접속어(_NC), 서술어(VC), 부사어(AC), 관형어(DC), 필수격(CAS), 독립어(IC)로 구분한다. 체언접속어의 경우, 일반적인 체언상당어구와 달리 서술어와 결합하지 않으므로 체언상당어구와 체언접속어를 구분한다. 그리고 체언상당어구와 체언상당어구가 결합하여 새로운 체언상당어구를 생성하는데 비해, 필수격은 오로지 서술어와 결합한다는 점을 감안하여 체언상당어구와 필수격도 분리한다.

이러한 기본 범주를 자질로 표현하기 위해서 <표1>과 같이 체언자질 n-, n, n+와 용언자질 v-, v, v+을 도입한다. 체언자질은 해당 범주가 체언상당어구이면 n 자질을 갖고, 체언상당어구는 아니지만 뒤따르는 체언상당어구를 수식하거나 한정하면 n+ 자질을 가지며, 그렇지 않으면 n- 자질을 갖는다. 용언자질도 체언자질가 마찬가지로, 해당범주가 용언상당어구이면 v 자질을 갖고, 용언상당어구는 아니지만 뒤따르는 용언상당어구를 수식하거나 한정하면 v+ 자질을 가지며, 그렇지 않으면 v- 자질을 갖는다.

한편, 본 논문에서는 구문범주의 설정을 위해 X-바 이론의 투사과정에서 이용되는 바 개념을 도입하였는데, 품사태그에 대한 구문범주와 바 자질은 <표2>와 같이 표현된다. 기존의 X-바 이론에서는 일반적으로 0, 1, 2 바를 이용하여 문장을 분석하는데, 기능어가 발달하였다는 한국어의 특징을 반영하기 위해 -1 바[2] 자질을 추가한다. 내용어의 경우 기본적으로 0 자질을 갖고, 기능어의 경우 -1 자질을 갖는다. 기능어에 의해 통사적으로 파생하는 경우 파생된 결과를 기본형으로 취급하여 0 자질을 부여하고, 의미적으로 파생한 경우 1 자질을 부여한다. 수식을 받거나 하위범주 관계로 결합하면 2 자질을 갖는다.

1) ±N자질과 ±V자질을 이용하여 명사[+N-V], 형용사[+N+V], 동사[-N+V], 전치사[-N-V]로 구분한다. 따라서, +N자질로 명사와 형용사를 모두 가리킬 수 있다.

2) Hoekstra(1980)는 0과 -1 바를 이용하여 단어의 파생어 규칙에 X-바 이론을 적용하고자 하였다[13].

IC^{-1}	호격조사(pv)
IC^{0}	감탄사(c)
VC^{-1}	동사화접미사(xsvv),형용사화접미사(xsvj), 서술격조사(i), 선어말어미(ep), 종결어미(eff), 연결어미(efc)
VC^{0}	동사(vv),형용사(vj)
CAS^{-1}	주격조사(ps),목적격조사(po),보격조사(pc), 여격조사(pt),부사격조사(pa),보조사(px)
$\_NC^{-1}$	접속격조사(pn)
NC^{-1}	복수형접미사(xsnpl),명사형접미사(xsn), 명사형전성어미(efn)
NC^{0}	보통명사(nnc),비단위성의존명사(nnb), 단위성의존명사(nnbu),고유명사(nnp),인 칭대명사(npp),지시대명사(npi),수사(nu)
DC^{-1}	관형형전성어미(efd),관형격조사(pd)
DC^{0}	성상관형사(da), 지시관형사(di), 수관형사(du)
AC^{-1}	부사화접미사(xsa),부사형전성어미(efa)
AC^{0}	성상부사(aa),지시부사(ai),서술부사(ap)
S^{-1}	마침표(ss.),느낌표(ss!),물음표(ss?), 종결기호($)

<표2> 품사태그와 구문범주의 관계

2.2 문법 규칙의 형식

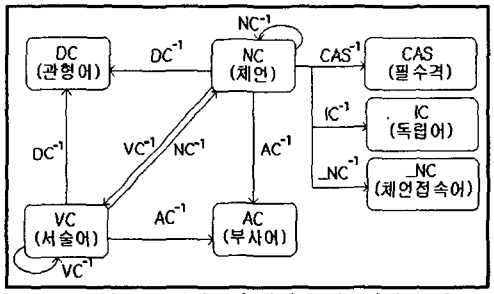
Chomsky는 '조건을 만족하는 두 개의 정보가 합쳐져 둘을 포괄하는 상위의 정보구조를 생성한다'는 이진 결합 관계 개념[5]을 바탕으로 구구조 문법 규칙의 형식을 CNF(Chomsky Normal Form)으로 제약하여 규칙의 수를 감소시켰다. 반면, 길이가 긴 규칙에 대해 하나이상의 CNF 규칙을 이용하여 표현함으로써 표현 능력은 그대로 유지시켰다. N^p와 N^q, N^r이 임의의 비단말노드이고 w^k가 임의의 단말노드라고 할 때, CNF는 규칙에 대해 $N^p \rightarrow w^k$ 형식과 $N^p \rightarrow N^q\ N^r$ 형식만을 허용한다[2].

한국어의 경우, 필수성분과 기능어의 생략현상이 빈번하게 발생한다. 일반적인 구구조 문법에서는 동사구의 중심어인 동사에 의해 하위범주화와 문장 유형이 결정되는데 비해, CNF는 단순히 두 개의 비단말노드만을 고려하는 $N^p \rightarrow N^q\ N^r$ 형식이므로 동사의 하위범주화 문제가 심각하게 고려되지 않는다는 장점이 있다. CNF의 규칙 형식은 생략현상이 빈번한 한국어 분석에 적합하므로[12], 제안하는 자질기반의 한국어 구구조 문법의 규칙 형식은 CNF를 따른다.

2.3 문법의 구성 원리

X-바 이론에서는 중심어인 X를 기준으로 그 중심어가 최대로 투사하여 결합한다는 개념을 바탕으로, 중심어 X^0이 보충어(complement)와 통합하여 X^1로 중간 투사되고 X^1이 X^1의 명시어(specifier)와 통합하여 X^2로 최대 투사된다고 본다. 보충어나 명시어가 존재하지 않을 경우, 공범주(∅)와의 결합을 허용한다. 이 때 바의 수를 "$X^n \rightarrow X^{n-1}$또는n"으로 제약하여 X-바 이론이 모든 범주 X에 적용 가능하도록 한다[6,9,11,12,14].

한국어의 경우, 동사구 분석시 보충어에 해당하는 주성분과 명시어에 해당하는 부속성분간의 어순이 자유로우므로, X-바 이론을 그대로 한국어에 적용하면 X^2가 다시 X^1로 투사될 수 있고, 어미나 조사 등의 기능어의 역할이 충분히 고려되지 않는다는 문제점이 발생한다[12]. 따라서, X-바 이론을 한국어에 적용하기 위해서는 기능어의 역할과 부분적으로 자유로운 어순을 고려하여 수정해야 한다.



<그림1> 구문범주의 파생과 기능어의 관계

<그림1>과 같이, 기능어는 구문범주의 파생에 관여하는데, 그 성격에 따라 통사적 파생 기능어와 의미적 파생 기능어로 나누어 볼 수 있다. 통사적 파생 기능어란 명사형 전성어미 '-음'이나 서술격조사 '-이-'와 같이 구문 범주를 변경시키는 기능어를 의미한다. 이에 반해 의미적 파생 기능어란 복수형 접미사 '-들'처럼 구문 범주는 변경시키지 않고 의미를 첨가시키는 역할을 하는 기능어를 말한다. 이와 같이 통사적 파생 기능어와 의미적 파생 기능어의 역할이 서로 다르므로 이에 해당하는 규칙은 분리되어 처리되어야 한다.

한편, 한국어에서 비중심어간의 어순은 자유롭지만 핵이 되는 중심어는 후행한다는 특성을 살리기 위해서는 수식과 하위범주에 관련된 규칙을 통합하여 처리하여야 한다. 이렇게 함으로써, 중심어 X가 후행한다는 특성을 유지시키면서 수식어와 하위범주간의 어순이 크게 영향을 받지 않을 수 있다.

이러한 점들을 반영하여 본 논문에서는 통사적

파생, 의미적 파생, 수식 및 하위범주 관계를 중심으로 <그림2>와 같이 X-바 이론을 변형한다. 또한, 처리의 효율성을 위해 공범주(∅)를 허용하지 않으며, 체언자질과 용언자질을 나타내는 변수 F로써 중심어와 수식 및 하위범주의 결합 관계를 표현한다.

예를 들어, 문장 "그녀는 예뻤다."를 제안하는 자질기반의 한국어 구구조 문법의 구성 원리에 적용하면 다음과 같다. 먼저, 형태소 분석결과인 품사태그의 열은 <표2>와 같은 "$X^{0또는-1}$→품사태그" 규칙에 의해 구문범주를 부여받게 된다. 구문범주의 열은 통사적 파생과 의미적 파생, 수식 및 하위범주 관계에 의해 결합하면서 구문트리를 구성하게 된다.

"그녀는"의 경우, 인칭대명사가 보조사에 의해 필수격으로 파생되므로, 통사적 파생 규칙 "CAS^0 →NC^0 CAS^{-1}"에 적용된다.

이와 달리 "예뻤다"의 경우는 형용사가 선어말어미와 종결어미와 결합하여도 기본자질은 그대로 VC가 되므로, 의미적 파생 규칙 "VC^1→VC^0 VC^{-1}"과 "VC^1→VC^1 VC^{-1}"이 적용된다.

"그녀는 예뻤다"에서 필수격과 서술어는 하위범주 관계에 의해 결합되는데, 필수격의 체언자질과 용언자질이 [n-v+]이고 서술어의 체언자질과 용언자질이 [n-v]이므로, 수식 및 하위범주에 관한 규칙 "VC^2→$CAS^0_{F=v+}$ $VC^1_{F=v}$"이 적용된다.

"그녀는 예뻤다"와 마침표"."는 문장에 관련된 규칙 S→VC^2 S^{-1}이 적용되어 완성된 문장을 구성하게 된다.

여기서 주의할 점은 의미적 파생의 경우 바의 제약이 ≤1이라는 것이다. 즉, 수식 및 하위범주 규칙 "VC^2 → $CAS^0_{F=v+}$ $VC^0_{F=v}$"에 의해 생성된 "그녀는 예쁘"는 선어말어미 "었"과 결합할 수 없다. 왜냐하면 "그녀는 예쁘"가 VC^2로 의미적 파생 규칙의 적용 조건 ≤1에 만족하지 못 하기 때문이다. 관형어가 체언상당어구만을 수식하는데 비해서, 부사어의 경우는 어간과 어미가 결합된 서술어를 수식한다[20]는 점을 반영하여, 본 논문에서는 수식 및 하위범주의 결합관계보다 의미적 파생 관계에 우선순위를 부여한다.

2.4 부가자질의 구성

앞에서 제시한 구성 원리를 바탕으로 문법을 생성하면, 규칙이 직관적이고 간단하면서도 대부분의 문장을 처리할 수 있을 만큼 표현력이 뛰어난 반면, 분석결과의 수가 급격히 증가하여 나타난다. 그러므로, 구성 원리를 바탕으로, 좀더 세분화된 제약 정보를 표현할 수 있는 부가적인 규칙이 필요하다. X가 상위 범주로 투사될 때 많은 정보들이 상호 전달되어 공유되는데, 이러한 정보들을 이용하면 분석결과의 급증을 완화시킬 수 있을 것이다.

예를 들어, 기본자질 정보만을 이용하면 문장 "너 때문이다."의 구문 분석결과는 다음과 같이 두 가지 경우가 발생한다.

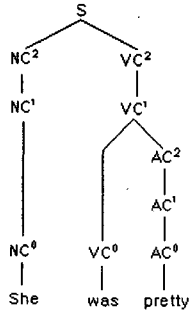
< 기존의 X-바 이론 >

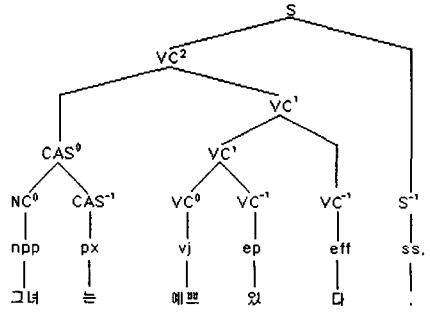
X^0	→	어휘	
X^1	→	X^0	COMP
X^2	→	[SPEC,X^1]	X^1
S	→	NC^2	VC^2

< 제안하는 X-바 이론의 변형 >

$X^{0또는-1}$	→	품사태그		
X^0	→	$Y^{\leq 2}$	X^1	(통사적 파생)
X^1	→	$X^{\leq 1}$	X^1	(의미적 파생)
X^2	→	$Z^{\leq 2}_{F=f+또는f}$	$X^{\leq 2}_{F=f}$	(수식 · 하위범주)
S	→	$X^{\leq 2}$	S^{-1}	(문장)

※ "≤k"는 "0≤바의 수≤k"의 의미





<그림2> 한국어의 특성을 반영한 X-바 이론의 변형

(1) ((((너때문)이))다).)

(2) ((너((때문이)다))).)

즉, 주어가 생략된 경우로 처리한 (1)의 분석결과 뿐만 아니라, 주격조사가 생략된 경우로 처리한 (2)의 분석결과까지 나오게 된다. 이 때 '때문'의 품사가 의존명사라는 정보와 의존명사는 선행하는 관형어나 체언상당어구의 도움을 받아야만 문장성분으로 쓰일 수 있다는 정보를 규칙체계에 반영한다면, (2)의 분석결과는 제거될 수 있을 것이다.

따라서, 품사와 격, 보조용언구와 관련된 정보를 다음과 같은 부가자질을 도입하여 표현하고자 한다.

$$ADD = \{ x \mid x \subseteq \{ pos_1, pos_2, \ldots pos_n, no\_pos, sub, obj, dat, aux, pas, cau \} \}$$

pos_1, pos_2, ... pos_n는 n개의 품사태그들로서 해당 구문범주의 품사 정보를 나타내며, 구문범주가 수식관계에 의해 생성된 경우는 no_pos를 이용하여 표현한다. 이러한 품사정보는 구문범주간의 결합여부를 결정하는데 영향을 끼칠 수 있다.

주격(subject) 조사나 목적격(object) 조사, 여격(dative) 조사에 의해 명확하게 격에 제시된 필수격이 서술어의 하위범주가 되는 경우는 sub와 obj 및 dat 자질을 부여한다. 이를 이용하여 동사와 필수격의 중복결합을 방지할 수 있다.

시제나 양상, 피동(passive), 사동(causative)을 표현하는 보조용언구(auxiliary)는 문장의 의미나 통사구조에 영향을 끼치므로 aux와 pas, cau 자질을 부여하여 처리한다.

2.5 문법의 구성

자질기반의 한국어 구구조 문법 G는 G = (NT, T, P, S)와 같이 4개의 순서쌍으로 정의된다. T는 단말노드(Terminal)의 유한 집합으로 형태소 분석 결과인 품사 태그로 구성된다. S는 문장(Sentence)을 나타낸다. P는 생성규칙(Production rule)의 유한 집합으로 <그림3>을 바탕으로 하여 구성된다. NT는 구문범주의 유한집합을 나타내는 비단말(NonTerminal)기호로 다음과 같이 정의된다.

$$NT = \{ \langle base, add \rangle \mid base \in BASE, add \in ADD \}$$

$$BASE = \{ \langle N, V, BAR \rangle \mid N \in \{ n+, n, n- \}, V \in \{ v+, v, v- \}, BAR \in \{ -1, 0, 1, 2 \} \}$$

$$ADD = \{ x \mid x \subseteq \{ pos_1, pos_2, \ldots pos_n, no\_pos, sub, obj, dat, aux, pas, cau \} \}$$

구문범주 집합 NT는 기본자질(base)과 부가자질(add)의 쌍을 원소로 하여 구성되며, 임의의 구문범주에 대해 $X^{bar}[add]$로 간단히 표현할 수 있다. <그림2>에서 제시된 문법에 부가자질연산을 추가하여 <그림3>과 같이 규칙을 확장한다.

$X^{0 또는 -1}[\alpha]$ → 형태소

$X^0[\beta]$ → $Y^{\leq 2}[\alpha]$ $X^1[\beta]$

$X^1[\alpha \cup \beta]$ → $X^{\leq 1}[\alpha]$ $X^1[\beta]$

$X^2[\alpha \# \beta]$ → $Z_{F=f+또는f}^{\leq 2}[\alpha]$ $X_{F=f}^{\leq 2}[\beta]$

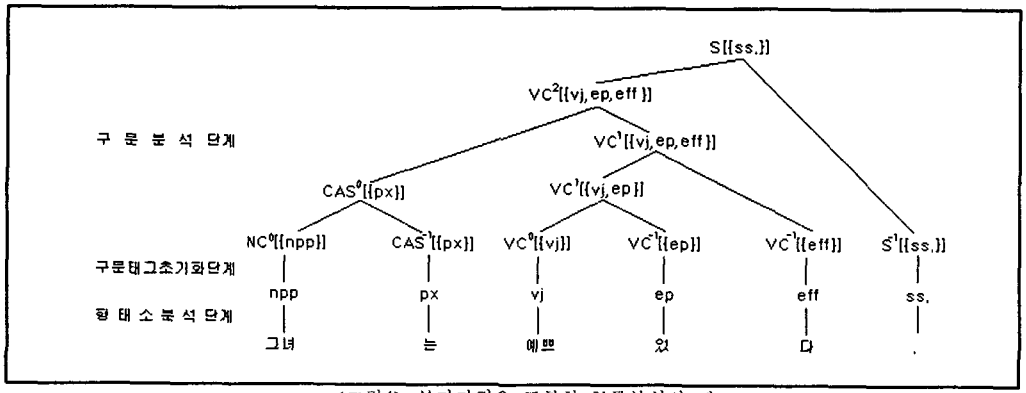
$S[\beta]$ → $X^{\leq 2}[\alpha]$ $S^1[\beta]$

※ POS = {nnc, nnb, ... pos_n}

※ $\alpha \# \beta$ =
- β - POS ∪ {no_pos} if X ≠ VC
- $\beta \cup$ {sub} else if ps ∈ α
- $\beta \cup$ {obj} else if po ∈ α
- $\beta \cup$ {dat} else if pt ∈ α
- β otherwise

<그림3> 부가자질을 추가하여 확장한 생성규칙

앞에서 제시된 예문인 "그녀는 예뻤다."를 확장된 규칙에 적용하면, <그림4>과 같이 분석된다. 부가자질 연산을 중심으로 살펴보면 다음과 같다.



<그림4> 부가자질을 포함한 구문분석의 예

품사태그에 대해 부가자질 α를 포함하는 "$X^{(0또는1)}$ [α]"형의 구문범주를 부여받아 초기화한 후, 구문범주에 대해 규칙을 적용하여 결합관계를 결정한다.

"그녀는"의 경우, 기능어 "는"에 의해서 미지격이 결정되므로 보조사를 나타내는 px 정보는 상위 구문범주에 계승된다. 즉, 통사적 파생의 구문범주는 통사적 파생 기능어에 의해 결정되는 것이다.

"예뻤다"의 경우, 형용사의 어간 "예쁘"가 선어말어미와 어말어미와 결합하여도 형용사의 성질은 유지되므로 형용사를 나타내는 vj 정보는 그대로 상위 구문범주로 계승된다. 또한, 선어말어미와 종결형 어말어미 정보도 $\alpha \cup \beta$연산에 의해 포함된다. 이는 의미적 파생 기능어의 의미 정보가 제약정보로써 활용될 수 있도록 고려한 것이다.

"그녀는 예뻤다"를 처리할 때, 부가자질에 대해 $\alpha \# \beta$연산을 하는데, X=VC이지만 보조사(px)에 의한 필수격으로 그 격을 명확하게 판단할 수 없으므로, sub, obj, dat를 포함하지 않고 β를 그대로 유지시킨다.

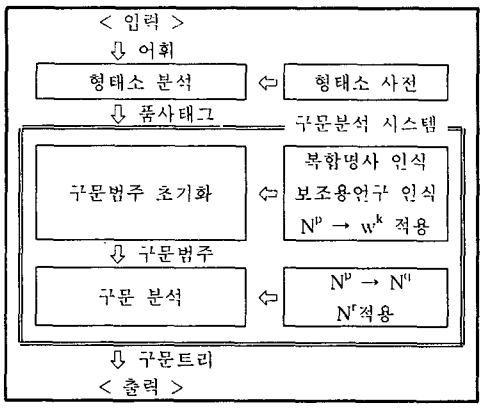
3. 구문분석 시스템의 구현 및 평가

제안하는 구문 분석 시스템에서 입력된 품사태그열은 기본자질과 부가자질로 구성된 구문범주로 초기화된 후, 규칙에 적용되어 구문분석되도록 구성된다. 즉, 구문범주 초기화단계에서 각 품사태그가 "$N^p \rightarrow w^k$" 형식의 규칙에 의해 구문범주를 부여받고, 구문분석 단계에서 "$N^p \rightarrow N^q\ N^r$" 형식의 규칙을 적용하여 이를 처리한다. 구문분석 단계에서의 부담을 감소시키기 위해, 구문범주의 초기화 단계에서 복합명사와 보조용언구를 인식한다.

3.1 구문분석 시스템의 구성

제안하는 자질기반의 한국어 구구조 문법의 성능을 검증하기 위해 <그림5>과 같은 구문분석 시스템을 구현하여 실험하였다. 제안하는 구문분석 시스템은 구문범주 초기화 단계와 구문분석단계로 구성되어 있으며, 구문분석 시스템에서 사용된 형태소 태그는 본 연구실에서 개발한 형태소 분석[19] 및 태깅 시스템[7]의 결과를 사용한다.

구문범주 초기화 단계에서는 품사 태그를 바탕으로 기본자질 및 부가자질로 생성하여 구문범주를 초기화한다. 이 때 복합명사 및 보조용언구를 인식하여 구문분석 단계의 부담을 줄인다. 구문분석 단계에서는 비단말노드간의 이진 결합과 관련된 규칙을 이용하여 구문분석한다.



<그림5> 구문분석 시스템구성도

구문범주 초기화 단계에서 논문 [1]의 명사쌍 관련 테이블을 바탕으로 복합명사를 인식하여 처리한다. 나열된 여러 명사에 대한 명사간의 결합순서까지 제안하는 문법에서 고려할 경우, 구문분석기의 부담이 가중되므로 이를 제안하는 문법의 고려 대상에서 배제한다. 또한, 시제나 양상, 피동, 사동 등을 나타내는 보조용언구는 문장의 의미나 통사 구조에 영향을 끼칠 수 있으므로[4,18], 구문분석 초기화단계에서 이를 보조용언구로 인식하고 aux, pas, cau 자질을 부가자질로 부여하여 구문분석 단계에서 의미적 파생 기능어로 처리한다.

3.2. 실험 및 평가

본 논문에서 제안하는 자질기반의 한국어 구구조 문법을 실험하기 위해, 품사 태그가 부착된 1993년 동아일보 사설 말뭉치에서 454문장을 추출하였다. 실험한 문장에 대해 살펴보면 <표3>과 같다.

형태소길이	(문장수)	어절수		
		평균	최대	최소
1 - 10	(100)	3.47	6	1
11 - 15	(100)	5.94	9	3
16 - 20	(100)	8.04	11	5
21 - 25	(100)	9.95	14	7
26 - 30	(54)	11.91	15	8

<표3> 실험문장의 길이

제안하는 자질기반의 한국어 구구조 문법에 대해 추출한 454문장을 대상으로 실험하였는데, <표4>와 같이 약 97%에 해당하는 441문장을 제대로 분석하였다. 정확하게 분석하지 못한 13문장 중 7문장은 구문범주의 초기화 단계에서 이용하

구 분 / 형태소길이	생성된 구문 트리 수								
	이진			바제약			부가자질		
	최대	최소	평균	최대	최소	평균	최대	최소	평균
1 - 10	1651	15	396.27	94	10	36.24	5	1	1.36
11 - 15	9980	16	3283.87	9718	15	2441.48	99	1	20.06
16 - 20	980765	1854	406155.45	9738	1024	3880.69	964	10	164.20
21 - 25	980325	55725	370540.50	9735	172	4083.24	984	4	416.93
26 - 30	973245	31235	406628.96	93170	1168	11955.23	785	15	336.87

<표5> 생성된 구문 트리의 수

는 복합명사인식의 오류로 인해 제대로 분석하지 못 하였으며[3], 5개의 문장은 선행하는 목적어를 하위범주로 갖는 동작성 보통명사를 일반 보통명사로 처리하여 분석에 실패하였다[4]. 품사와 격 정보를 부가자질에 부여하였는데, 형용사의 하위범주로 목적어가 나타난 경우는 예외적인 경우로 정확하게 처리하지 못 하였다.[5]

구문범주 초기화 단계에서는 복합명사를 인식하도록 처리하였는데, 391개의 복합명사중 약 98%에 해당하는 384개의 복합명사가 제대로 인식되었다. 반면, 부사적 성질을 갖는 보통명사에 대해 제대로 처리하지 못하였다.

구 분 / 분석결과		이진		바제약		부가자질	
		문장수	적용율(%)	문장수	적용율(%)	문장수	적용율(%)
올바른 분석을 포함하는 경우		442	97.3	442	97.3	441	97.1
분석에 실패한 경우	복합명사 인식오류	7	1.6	7	1.6	7	1.6
	동작성 보통명사의 처리 미흡	5	1.1	5	1.1	5	1.1
	하위범주화에 대한 예외발생	0	0.0	0	0.0	1	0.2

<표4> 구문분석시스템 실험결과

참고적으로, <표4>에서 이진문법은 구문범주간의 결합관계만을 고려하여 표현한 것이고, 바 제약은 이진문법에 바 자질의 제약을 고려하여 분석한 것이다. 부가자질이란 좀 더 세분화된 정보를 이용하기 위해 이진문법과 바 제약뿐만 아니라 부가자질까지 포함하여 처리한 것이다.

실험문장에 대해 제안된 자질기반의 한국어 구구조 문법으로 분석할 때 구문 트리는 <표5>과 같이 생성된다. 이는 부가자질의 정보를 이용하면, 생성되는 구문 트리의 수가 급격히 증가되는 것이 다소 완화될 수 있음을 보여준다. 그럼에도 불구하고, 동사 하위범주화 정보의 부족과 빈번한 기능어 생략 현상, 부사절이나 관형절의 처리 등으로 인하여 여전히 구문 분석 결과의 수가 지나치게 많이 생성되었다.

제안된 자질 기반의 한국어 구구조 문법으로 구문 분석을 하면, <표6>같이 간선(edge)이 생성된다. 전체 생성된 간선에 대해 정확하게 생성된 간선을 다음과 같이 계산하여 비교하였다.

$$\text{정확율}(\%) = 100 \times \frac{\text{정확하게 생성된 간선수}}{\text{전체 생성된 간선수}}$$

구분 / 형태소길이	정확한 간선수 평균	이진문법		바제약		부가자질	
		생성된 간선수 평균	정확율(%)	생성된 간선수 평균	정확율(%)	생성된 간선수 평균	정확율(%)
1-10	18.58	344.46	5.39	321.10	5.78	24.78	74.97
11-15	24.98	3802.57	0.65	3634.88	0.68	355.09	7.03
16-20	33.02	36536.62	0.09	12705.22	0.25	1693.56	1.94
21-25	351.65	39589.40	0.88	38244.26	0.91	3393.25	10.36
26-30	479.22	298954.31	0.16	39950.91	1.19	3998.38	11.98

<표6> 생성된 간선수와 정확율

이와 같이 간단한 원리만 적용하여도 대부분의 문장을 올바르게 분석할 수 있었지만, 분석결과의 수가 지나치게 많이 생성된다는 문제점이 나타났다. 그러므로, 앞으로 부가자질에 포함되는 제약정보를 보완하는 한편 통계정보를 이용하는 연구가 요구된다.

4. 결론 및 추후 연구

본 논문에서는 X-바 이론을 한국어에 적용하여 서로 다른 범주들간의 구조적 일반성을 파악하고, 한국어에 가능한 규칙만을 허용하여 불가능한 규칙을 배제시킬 수 있는 틀을 제시하였다. 한국어

3) 예문 : 이런 때 과감한 금융계의 수술 없이 한국경제가 살아나기는 지극히 어렵다.

4) 예문 : 신임 최회장은 이를 의식, 취임사에서 전경련의 몇 가지 새 운영계획을 밝혔다.

5) 예문 : IAEA와의 싸움은 궁극적으로 유엔안보리의 결행을 가능토록 국제 여론조성에 도움을 주기 때문이다.

다는 점을 고려하여, 중심어와 보충어 관계 중심의 기존 X-바 이론을 통사적 파생과 의미적 파생, 수식 및 하위범주의 관계를 중심으로 변형하였다. 또한, 한국어의 빈번한 생략현상에 효과적으로 대응할 수 있도록 이진결합 중심의 CNF를 따르도록 하였다.

자질 기반의 한국어 구구조 문법은 간단하면서도 대부분의 문장을 처리할 수 있을 만큼 표현력이 뛰어난 반면, 제약정보의 부족으로 분석결과가 지나치게 많이 생성된다는 문제점이 제기되었다.

구문분석기의 성능향상을 위해서는 복합명사의 인식능력을 개선시키고 부가자질에 포함되는 제약 정보를 보완하는 한편 구문 태그 부착 말뭉치에서 통계정보를 추출하고 이를 활용하는 작업이 진행되어야 할 것이다.

참고문헌

[1]Bo-Hyun Yun, Yong-Jae Kwak, Hae-Chang Rim, "Alleviating Syntactic Term Mismatches in Korean Text Retrieval", International Journal of Information Processing and Management. (To be published)

[2]Eugene Charniak. 『Statistical Language Learning』, The MIT press, 1993.

[3]Ted Briscoe, Nick Waegner, "Robust Stochastic Parsing Using the Inside-Outside Algorithm", In Workshop Notes, Statistically-Based Natural Language Programming Techniques, AAAI, 1992.

[4]강호관, 이종혁, 이근배, "새로운 어절 해석에 기반한 한국어 의존관계 파서", 제9회 한글 및 한국어 정보처리 학술발표 논문집, pp.327~331, 1997.

[5]김영택, 『자연 언어 처리』, 교학사, 1994.

[6]김운태, "X-bar 이론에 대한 연구", 대구대학교 영어영문학과 석사학위논문, 1988.

[7]김진동, 임희석, 임해창, "Twoply HMM : 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델", 정보과학회 논문지(B), 24권 12호, pp. 1502~1512, 1997.

[8]박성숙, 심영섭, 한성국, 최운천, 지민제, 이용주, "이진결합중심의 한국어 chart parser", 제5회 한글 및 한국어 정보처리 학술발표 논문집, pp.15~24, 1993.

[9]박승혁, 『최소주의 문법론』, 한국문화사, 1997.

[10]박영순, 『현대 한국어 통사론』, 집문당, 1997.

[11]박한기, "X-bar이론에 의한 영어의 구구조 연구", 전남대학교 영어영문학과 박사학위 논문, 1989.

[12]서정목, 『국어통사구조연구I』, 서강대학교 출판부, 1994.

[13]시정곤, 『국어의 단어형성 원리』, 국학자료원, 1994.

[14]신재정, "X-bar 제약분석", 경북대학교 영어영문학과 석사학위 논문, 1985.

[15]우승균, "구문관계를 이용한 한국어 구문분석", 한국과학기술원 전산학과 석사학위 논문, 1991.

[16]윤덕호, 김영택, "다단계 여과 및 탐색을 이용한 의존문법에 기반을 둔 한국어 분석알고리즘", 한국정보과학회 논문지, 19권 6호, pp.614~623,1992.

[17]이공주, 김재훈, 김길창, "제한된 형태의 구구조 문법에 기반한 한국어 구문분석", 정보과학회논문지(B), 25권 4호, pp.722-732, 1998

[18]이공주, 김재훈, 장병규, 최기선, 김길창, "한국어 구문트리태깅 코퍼스 작성을 위한 한국어 구문 태그", 기술보고서, KAIST, CS/TR-96-102, 1996.

[19]임해창, "한글자동색인을 위한 기초도구 구축 연구", 최종연구보고서, 한국과학기술원 연구개발정보센터, 1995.

[20]조규빈, 『하이라이트 고교문법자습서』, 지학사, 1993.

[21]홍영국, 이종혁, "한국어 의존 해석을 위한 형태-통사적 품사 분류 체계", 정보과학회 논문지(B), 22권 9호, pp. 1375~1383, 1995.