(제 10회 한글 및 한국어 정보처리 학술대회)

한국어 구문 분석을 위한 LTAG 시스템

정 의석 *윤 준태 김 선호 송 만석
연세대학교 컴퓨터과학과, 서울시 서대문구 신촌동 134, 120-749
*한국과학기술원 전산학과, 대전 광역시 유성구 구성동 371-1, 301-701
{bishop, pobi, mssong}@december.yonsei.ac.kr, *jtyoon@world.kaist.ac.kr

The LTAG System for Korean Syntactic Analysis

Euisok Chung, *Juntae Yoon, Seonho Kim, and Mansuk Song
Dept. of Computer Science, Yonsei University, *Dept. of Computer Science, KAIST

요 약

한국어 구문 분석에 적용되어 왔던 의존 문법이나 구구조 문법들은 각각의 장점들만큼 구문 중의성 해결의 어려움과 구구조 기술 한계성등의 문제점들을 내포하고 있다. 따라서 본 연구는 LTAG(lexicalized tree adjoining grammar)을 기반으로 기존 문법들의 장점들을 수용하는 새로운 구구조 문법 시스템을 제안한다. 이는 기본 트리 프레임, 기본 트리 명시 규칙, 기본 연산 제약 규칙으로 구성되어 있으며 역방향 구문 분석 기법을 이용한다. 끝으로 실험을 통하여 제안하는 시스템의 한국어 구문 분석에 대한 타당성을 보이고자 한다.

1 서 론

자연어 처리 분야에서 구문 분석은 대상 언어의 일반적 구문 체계, 즉 문법 체계를 기반으로 한다. 한국어의 구문 분석은 핵심어구의 생략과 부분 자유 어순을 잘 처리할 수 있는 의존 문법이 주류를 이루고 있다. 그러나 어절 사이의 의존 관계만을 파악하는 의존 문법은 문장의 구조적인 정보를 이용하지 못하므로 구구조 문법을 이용한 경우보다 구문 중의성의 처리가 어렵게 된다[12]. 따라서 문장의 구조적인 정보를 제시할 수 있으며 핵심어구의 생략과 부분 자유 어순의 문제를 해결할 수 있는 문법 체계가 한국어 구문 분석에 타당하다 할 수 있다.

본 연구는 LTAG[5]을 기반으로 한국어 구문 분석에 적합한 구구조 문법 체계의 제시를 목적으로 한다. LTAG은 초기 트리(initial tree)와 보조 트리(auxiliary tree)로 구성되어 구구조를 표현하는 기본 트리(elementary tree)와, 대치(substitution)와 결합(adjunction) 연산으로 구문 분석을 수행하는 기본 트리 연산으로 구성된다[6]. LTAG의 장점은 어휘 자체가 문법에 반영되어 어휘의 지역적 구문 구조를 효과적으로 표현할 수 있다는 점으로 이는 강력한 통계 모델로의 확장을 가능하게 한다[9].

이러한 LTAG의 장점들을 계승하며 한국어에 적합한 LTAG 체계의 구축에 대한 연구가 지금까지 진행 되어 왔다[3, 4, 14]. [4, 14]는 MC-TAG을 적용하여 한국어의 부분 자유 어순 문제에 접근하였다. 특히 [14]에서는 트리 변형 규칙을 제시하여 LTAG 자체의 문제인 기본 트리가 과다하게 발생되는 문제점을 해결하고자 하였다. 과다한 기본 트리 발생 현상은 어휘의 지역적 구문 속성에 대한 LTAG의 표현력으로 인한 것으로 해당 어휘와 결합되는 기본 트리 결정의 모호성을 가중시킨다는 문제점을 야기한다[1].

본 연구에서는 어휘의 하위구조화 정보를 어휘 속성(lexical feature)으로 가정하고 LTAG의 기본 트리에서 제외하여 어휘 구문 속성의 일반화를 가능하게 하였을 뿐만 아니라 기본 트리수를 감소시켰으며, 문장의 구문 구조 해석의 관점에서 한국어의 자유 어순, 핵심어 생략등의 문제에 접근할 수 있게 하였다. 제안하는 LTAG 체계는 기본 트리 프레임, 기본 트리 명시 규칙, 기본 연산 제약

규칙으로 구성되어 있다. 기본 트리 프레임은 네 가지 형태의 프레임으로 한국어 어휘의 지역적 구문 구조를 일반화하며, 기본 트리 명시 규칙에 의해 어휘 정보를 지닌 기본 트리를 생성한다. 기본 연산 제약 규칙은 구문 분석시 과도한 모호성의 발생을 막기 위한 경험적 규칙(heuristic)으로 구성된다.

2 한국어에 적합한 LTAG 체계

본 절에서는 LTAG에 기반한 한국어 구문 구조의 형식화를 위한 기본 트리 프레임과 기본 트리 명시 규칙을 소개한다.

2.1 한국어 구문 구조

한국어 문장의 짜임새에 있어 형식 형태소의 역할은 단일 실질 형태소와의 관계로만 파악되기보다는 구절과의 관계로 보아야 된다[7]. 따라서 구문 구조에서 한국어 어절의 문법적 성격을 표현하기 위한 형식 형태소의 처리가 비중있게 다루어져 왔다[15].

예문 1) 철수가 그 학교에 갔다.

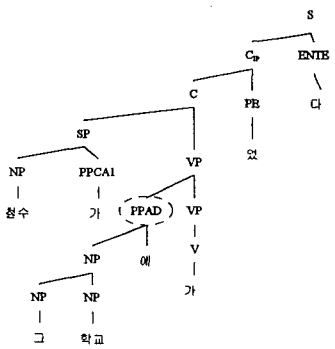

그림 1. 형태소 태그의 구문 태그화

먼저 본 연구에서는 구문 범주로서 S, C, SP, NP, AD 5가지와 명사절 NP_C, 부사절 AD_C, 시제나 인칭의 표현을 위한 C_{IP}를 설정하였으며[1], 형식 형태소의 품사 태그[2] 를 구문 범주로 확장하였다.

이러한 형식 형태소의 구문 범주화는 형식 형태소와 관계된 실질 형태소간의 구문 관계를 명시적으로 나타낼 수 있으며 기능어의 문법적 성분 표현을 가능하게 한다는 장점이 있다. 그림 1은 형식 형태소의 구문 관계 표현의 일례를 보여준다.

그림 1의 구문 트리에서 부사격 조사인 PPAD는 명사구와 서술어와의 관계를 자명하게 표현하고 있다. 또하나 본 연구에서 특이한 점은 모든 서술어는 하나의 주어를 취함으로써 절이 된다고 가정하여 SP를 동사구를 만드는 일반 NP로부터 분리하고 주격 조사가 SP에 내포되는 것으로 보았다.

2.2 기본 트리 연산

기본 트리 연산에는 대치 연산과 결합 연산이 있으며 그림 2는 이들 연산이 적용되는 예를 보여준다. LTAG은 그림 3과 같이 해당 어휘에 적합한 기본 트리를 추출한 후 기본 연산인 대치 연산과 결합 연산을 이용하여 구문 분석을 진행 한다.

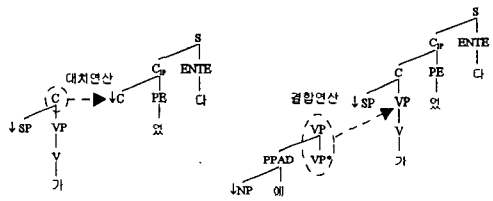

그림 2. 대치 연산과 결합 연산

2.3 기본 트리 프레임(Elementary Tree Frame : ETF)

그림 1의 구문 트리는 4가지 형태의 LTAG 기본 트리로 구성되어 있다. 그림 3[3]의 *1)*은 서술어 원형의 기본 트리를 보여주며, *2)*는 종결어미, 선어말 어미와 주격 조사의 기본 트리를 나타내고, *3)*은 부사격 조사의 기본 트리를, *4)*는 관형어의 기본 트리를 나타낸다. 여기서 초기 트리는 *1)*과 *2)*가 되며 *3)*과 *4)*는 보조 트리가 된다.

본 연구에서는 기본 트리를 초기 트리 α의 두형태, 보조 트리 β의 두형태로 일반화하였다. α형태는 서술어 기본 트리 표현을 위한 α1과 하나의 대치 노드와 형태소를 결합하여 새로운 구문 범주를 이루는 α2로 구성된다. 그리고 β형태

1) S : sentence, C : clause, C_{IP} : inflective phrase, SP : subject phrase

2) 본 논문의 형태소 태그는 [13]의 태그를 따른다.

3) ↓는 대치 노드이고, *는 결합 노드를 의미한다.

는 조사에 의한 성분의 서술어 결합을 표현하는 β1과 명사나 부사, 관형사등의 단일 결합형태를 표현하는 β2로 구성되어 있다.

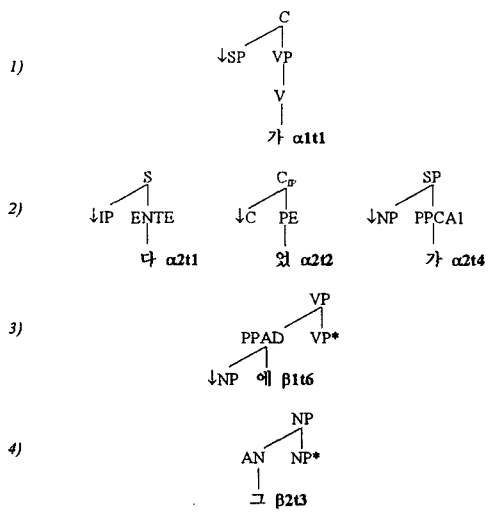

그림 3. 기본 트리의 구성

표 1은 α형태와 β형태의 네가지 ETF로 기본 트리를 일반화한 것이다. 이러한 기본 트리 프레임을 이용하여 기준 어휘(anchor)를 일반화한 형태소 태그 *tw*를 기준으로 ETF의 *[d]*와 *[u]*를 결정하면 기본 트리를 생성할 수 있다. 생성된 기본 트리는 index를 가지고 있어 구문 분석 후 최종 완성된 구문 트리로부터 구성된 기본 트리들을 명시적으로 추출할 수 있다.[4]

2.4 기본 트리 명시화

표 1의 ETF로부터 기본 트리를 생성할 때 하나의 *tw*에 하나 이상의 기본 트리가 생성될 수 있다. 이는 어휘의 구문 중의성으로 인해 발생하는데 본 연구에서는 이의 해결을 위해 표 2의 기본 트리 명시화 규칙을 제시한다. 규칙의 적용은 해당 어휘의 좌우 문맥과 문장 규칙(sentence rule: SR)에 따라 진행된다. 좌우 문맥은 *tw*의 좌우 연접 형태소로써 조사, 명사, 부사, 동사, 관형사가 속성이 된다. 문장 규칙은 해당 어휘의 문장내 위치 정보와 성분 중복 여부로 구성된다.

4) 이는 Synchronous TAG을 적용한 기계 변역을 가능하게 한다[3].

문장내 위치 정보는 H(+|-)로 형태소의 위치가 문두에 출연했을 경우 VP보다는 VP의 상위절인 C에 결합하는 것으로 처리한다. 해당되는 형태소로는 PPAU, PPCA2, PPAD, AD로 보조사의 주제화 현상, 격조사의 자유 어순 현상과 문장 부사의 처리를 필요로 한다.

표 1. 기본 트리 프레임의 일부

Type	ETF	{d}	{u}	tw[5]	Index
α	C ↓[u] VP ↓[d] tw◇ w	ε	SP	V	α1t1
		:	:	:	:
	[u] ↓[d] tw◇ w	C	S	ENTE	α2t1
		C	C_{IP}	PE	α2t2
		C	NP_C	ENTR2, ENTE	α2t3
		NP	SP	PPCA1	α2t4
		ε	ε	NP	α2t5
		:	:	:	:
β	[u] tw◇ [u]* ↓[d] w	NP	C	PPAU, PPAD, PPCA2	β1t1
		:	:	:	:
		NP	VP	PPCA2, PPAD, PPAU, PPCA1 PPCO	β1t6
		:	:	:	:
	[u] tw◇ [u]* w		C	NP, AD, CJ	β2t1
			VP	NP, AD, CJ	β2t2
			NP	AN, NP, CJ	β2t3
			:	:	:

예문 2) 밥이 떡이 된다.

성분 중복 여부를 판단하는 D(+|-)는 중출문의 처리를 위해 필요하다. 예문 2)는 '밥이'가 주어이고 '떡이'가 보어이다. 그러나 둘 모두 주격 조사로 해석되므로 둘 중 하나만을 α형태인 SP로 가정해야 하고, 다른 하나는 β형태인 VP결합으로

5) V 용언, NP 체언, AX 보조 용언, CO 지정사, AD 부사, AN 관형어, CJ 접속어, PPCO 보문소/ ENTE 종결, PE 선어말 어미/ PPCA1 주격, PPCA2 목적격, PPCA3 관형격, PPAU 보조사, PPAD 부사격, PPCJ 접속조사/ ENTR1 관형사형, ENTR2 명사형, ENTR3 부사형 전성 어미 / ENCO1 대등적, ENCO2 종속적, ENCO3 보조적 연결어미

간주해야 한다. 따라서 기준이 되는 서술어와 주격 조사의 중복 여부의 판단이 필요한 것이다.

표 2. 기본 트리 명시화 규칙의 일부

Tw	Index	SR	조사	명사	부사	동사	관형사
AD	β2t1 β2t2 β2t5	H+ H-			R+		
ENTR1	β1t3 β1t7			R+	R+		
ENTR2	α2t3 β1t3 β1t5		R+	R+		R+	
ENCO2	β1t2 β1t5					R- R+	
PPCA1	α2t4 β1t6	D- D+					
PPCA2	β1t1 β1t6	H+ H-					
:	:	:	:	:	:	:	:

좌우 문맥은 '좌측 형태소-기준어휘-우측 형태소'의 삼항으로 표현된다. 좌측 문맥의 발생 여부는 L(+|-)로 나타내고 우측 문맥의 발생 여부는 R(+|-)로 나타낸다. 즉 현재 기본 트리 할당을 요구하는 형태소의 태그가 ENTR1일 경우 표 1에 의해 연접된 우측 형태소가 명사나 부사의 여부를 판단한 후 기본 트리가 결정되어 진다.

3 구문 분석

본 절에서는 2절에서 기술된 기본 트리 프레임과 기본 트리 명시화 규칙을 기반으로 한 역방향 구문 분석 기법에 대해 기술한다.

3.1 역방향 구문 분석

역방향 구문 분석(backward parsing)은 한국어와 같은 중심어 후행 언어에서 유용한 구문 분석 방법이다[11]. 따라서 본 연구의 구문 분석은 입력 문장의 각 형태소에 기본 트리가 할당된 후 LTAG 연산에 따라 역방향으로 진행된다. 그림 4는 예문 1)의 역방향 구문 분석 과정을 보여주고 있다. 이의 최종 결과는 그림 1의 구문 트리가 된다.

LTAG은 구문 분석의 결과로부터 문장 구성 요소의 표면적 구조인 구문 분석 트리 이외에 구문 분석의 유도 과정을 기술한 유도 트리(derivation tree)를 추출할 수 있다[2]. 그림 5에서 보여지는 유도 트리를 의존 문법의 관점에서 본다면 기본 트리간 지배 의존 관계를 표현하는 의존 그래프로 볼 수 있다. 이는 제안하는 문법 체계가 의존 문법의 장점들을 수용하고 있다는 것을 보여 주고 있다.

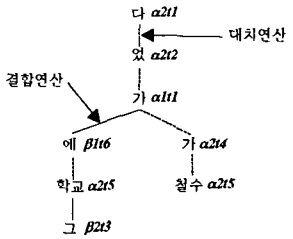

그림 5. 유도 트리

3.2 기본 연산 제약 규칙

구문 분석시 수행되는 기본 트리 연산에서 결합이나 대치 연산의 대상 노드들의 수가 구문 구조 모호성 증감에 영향을 미친다는 것은 분명하다. 이러한 경우 연산 대상 노드들에 일정한 제약을 가한다면 과도한 구문 모호성 발생을 어느 정도 제어할 수 있으리라 본다. 따라서 본 연구는 기본 연산 제약 규칙으로써 구문 모호성 발생 문제에 접근하였다.

기본 연산 제약 규칙은 인접 규칙, 인접 제약 규칙, 결합 연산 제약 규칙의 경험적 규칙으로 구성되어 있다. 인접 규칙은 기본 연산 적용시 두 연산 대상 노드를 포함하고 있는 기본 트리의 기준 어휘간 거리가 1일때 적용되는 규칙이다.

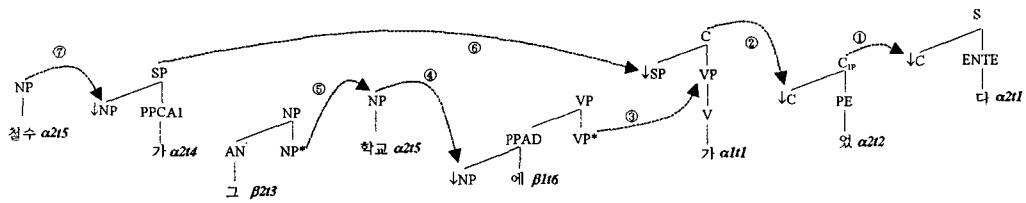

그림 4. 역방향 구문분석의 진행과정

그림 4 에서 ⑤, ⑥번[6]을 제외한 모든 연산에 인접 규칙이 적용되고 있다. 인접 제약 규칙은 인접 규칙의 예외 사항으로써 SP, PPAD, AN, PPAU, ENTR1, PPCA3, PPCJ, CJ를 포함하고 있는 기본 트리의 결합시 형태소의 인접 규칙을 무시하게 한다. 결합 연산 제약 규칙은 보조 동사로의 결합을 제약하는 규칙과 명사구에서 수식언이나 연속된 명사가 발생하는 경우 결합 대상 노드를 단일화하기 위한 방법이다.

4 실험 및 평가

본 연구는 저수준 구문 분석(shallow level parsing) 방법으로 문장 구조 분석에 접근 한다. 구문 분석 결과는 구문 중의성을 내포한 트리 군(tree forest) 형태로 제시되어 모호성 발생 지점을 자명하게 제시하여 준다. 최종 트리 군의 형태는 그림 6과 같다.

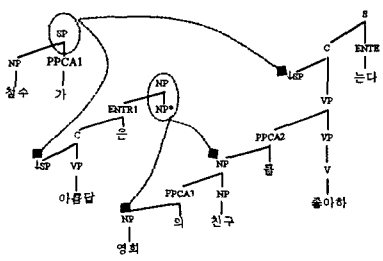

그림 6. 트리 군

본 연구의 실험은 두가지로 진행되었다. 첫번째 실험은 [10]의 SERI 구문 분석기 성능 평가용 문장 모음 중 중복된 유형의 문장을 제외한 196문장들을 실험 데이터로 하여 한국어의 특수 문형들에 대한 실험을 하였고, 두번째 실험은 [15]의 구문 분석 말뭉치 중 50여 문장을 무작위로 추출하여 생성 트리 수에 대한 실험을 하였다.

평가 방법은 [10]의 방법을 이용하였다. 구문 분석 결과는 트리 군 형태로 제시되므로 트리 군이 생성할 수 있는 모든 구문 트리를 고려 하여 평가하였다. 그리고 구문 분석 정확도의 기준은 생성된 구문 트리의 구조적 타당성 여부에 따라 판단하였다.

표 3. 문법 현상 정확도

어절수	문장 수	정확도
자유어순	8	0.85
중출문[7]	6	0.93
생략	9	0.86
내포문[8]	51	0.67
접속문	13	0.73

첫번째 실험의 분석 결과 정확도는 0.86으로 산출되었다. 그리고 표 3은 분석결과 중 일부 문장의 문법 현상별 정확도를 평가한 것으로 문장 수가 많지는 않지만 이는 비교적 자유 어순, 중출문, 생략 현상등의 한국어 특수 구문의 처리에 무리가 없음을 보여준다.

표 4. 실험결과 II

어절수	개수	생성트리수의 평균	정확도
1~5	2	1	1
6~10	15	4.6	0.76
11~15	21	33.6	0.56
16~20	12	350.8	0.6
21~25	1	1463	0.6
계	51		0.64

두번째 실험은 비교적 장문들로 구성되어 있으며, 평균 생성 트리수를 포함하여 분석하였다. 분석 결과 본 논문의 기본 트리 프레임이 기존의 구문 구조 규칙보다 생성되는 구문 트리의 수를 현격히 감소시켰음을 알 수 있다.

구문 분석이 실패한 문형으로는 '어제부터 오늘까지가~'와 '80점에서 90점까지를~'등의 어휘별 의미 유사도를 필요로 하는 유형과 '~와 함께'등과 같이 연어 현상으로 다루어야 되는 유형이 있었다. 그리고 '제출시 고려해야~'에서 '-시'와 같은 접미사는 기본 트리 프레임에서 고려되지 않아 구문 분석에 실패하였다.

5 결 론

본 논문에서는 한국어 구문 분석을 위한 LTAG 시스템을 제시하였다. 한국어의 구문 구조를 일반화한 기본 트리 프레임이 기존의 구구조 문법으로는 접근하기 힘들었던 자유 어순이나 생략 현상등

6) ⑤는 관형사 수식의 모호성, ⑥은 주어 결정의 모호성을 야기한다.

7) 주격 중출(이중주어)과 목적격 중출(이중목적어)

8) 명사절, 관형절, 부사절등의 안김문을 내포한 문장[8].

을 처리할 수 있으며, LTAG 자체의 문제점인 기본 트리의 과도한 생성을 피할 수 있음을 보였다. 그리고 구문 분석 결과를 트리 군의 형태로 제시하여 구문 분석이후 단계의 접근을 용이하게 하였다.

실험 결과 어휘의 의미 정보와 구문 분석 이전 단계에서 연어 현상 처리의 필요성을 확인할 수 있었다. 본 연구의 결과를 기반으로 어휘 정보나 통계적 기법을 추가한다면 보다 좋은 결과를 얻을 수 있을 것이다.

참고문헌

[1] Aravind K. Joshi and B. Srinvas, "Disambiguation of Super Parts of Speech(or Supertags): Almost Parsing," *Proceedings of the 15th International Conference on Computa-tional Linguistics,* Kyoto, Japan, pp. 154-160, 1994

[2] Christy Doran, Dania Egeni, Beth Ann Hockey, B. Srinvas and Martin Zaidel, "XTAG System – A Wide Coverage Grammar for English," *Proceedings of the 15th International Conference on Computational Linguistics,* Kyoto, Japan, pp. 922-928, 1994

[3] Hyun S. Park, "Mapping Scrambling Korean Sentences into English Using Synchronous TAGs", *Proceedings of ACL,* 1995

[4] Young-Suk Lee, *Scrambling as a Case-Driven Obligatory Movement*, Ph.D. Dissertation, University of Pensylvania, 1993

[5] Schabes, Y., Abeille, A., and Joshi, A.K. "Parsing strategies with 'lexicalized' grammars: Application to tree adjoining grammars," *Proceedings of the 12th International Conference on Computational Linguistics,* Budapest, Hungary, 1988

[6] The XTAG Research Group, "A Lexicalized Tree Adjoining Grammar for English," *IRCS Report 95-03*, University of Pensylvania, 1995

[7] 김기혁, "국어문법연구", 도서출판 박이정, 1995

[8] 남기심, 고영근, "표준 국어문법론", 탑출판사, 1985

[9] 서정연, 김창현, "통계적 방법을 이용한 구문 분석", 정보과학회지 제14권 제7호, pp. 58-70, 1996

[10] 성원경, 장명길, 박재득, 류법모, 이현아, 박동인, "SERI Test Suites'97 : 한국어 구문 분석기 성능 평가용 문장 모음", 제9회 한글 및 한국어 정보 처리 학술대회, pp. 320-326, 1997

[11] 양성일, 확장 문맥 자유 문법과 패턴-액션 규칙을 이용한 한국어 구문 분석에 관한 연구, 연세대학교 대학원 전산과학과 석사 학위 논문, 1995

[12] 엄미현, 한국어의 구조적 중의성 해소에 관한 연구, 연세대학교 대학원 전산과학과 석사 학위 논문, 1996

[13] 윤준태, 공기 관계 기반 어휘 연관도를 이용한 한국어 구문 분석, 연세대학교 대학원 컴퓨터과학과 박사학위 논문, 1998

[14] 이공주, 김길창, "TAG을 기반으로 한 한국어 구문 분석기에서 트리 변형 규칙", 제1회 지능기술 공동학술회의, pp. 100-105, 1995

[15] 이공주, 김재훈, 김길창, 제한된 형태의 구구조 문법에 기반한 한국어 구문 분석, 정보과학회논문지(B), 24권 4호, pp. 722-732, 1998