자질 기반 LTAG 파서를 위한 언어자원 구축

구상옥[0] 강보영 김미진 이상조
경북대학교 컴퓨터공학과
(tomato[0], boyoung,jean321)@sejong.knu.ac.kr, sjlee@knu.ac.kr

Building Resources for Feature-based Tree Adjoining Grammar

Sang-Ok Koo[0] Bo-Young Kang Mi-Jin Kim Sang-Joo Lee
Dept. of Computer Engineering, Kyungpook National University

요 약

Feature-based Lexicalized Tree Adjoining Grammar(FB-LTAG)는 한국어와 같이 부분자유어순이나 중요문장성분생략과 같은 현상이 빈번한 언어를 처리하기에 적합한 문법이다. FB-LTAG 구문분석기를 구현하기 위해서는 우선 한국어의 형태적, 통사적 특징이 체계적으로 표현된 언어자원이 구축되어야 한다. 본 연구는 FB-LTAG를 이용한 한국어 문장분석 연구의 기초 단계의 연구로서, FB-LTAG에 필요한 언어자원들의 효율적 구축 방법을 제안한다. 먼저 FB-LTAG 문법이론에 대해 간략하게 정리하고, FB-LTAG를 이용한 한국어 문장분석시스템의 전체구성 및 시스템에 필요한 여러 언어자원들을 소개한다. 마지막으로 XML을 이용한 언어자원의 표기 방법과 기존 전자사전을 이용한 언어자원 구축 방법을 제안한다.

1. 서 론

자연어처리에서 구문분석은 대상 언어에서 허용되는 문장의 구조를 형식적으로 정의하는 문법(grammar)과 컴퓨터에 의한 절차적 처리 방법인 파싱 알고리즘에 기반을 두고 있다. 지금까지 자연어의 구문분석에 적용되어 왔던 문법으로는 문장의 계층적 구조를 명확하게 제시할 수 있는 구구조문법과 한국어와 같은 자유어순언어를 처리하는 능력이 우수하다고 알려진 의존문법, 그리고 범주문법, 확률문맥자유문법, 트리결합문법(TAG) 등이 있다[1]. 이 중에서 구구조문법과 의존문법이 한국어 구문분석을 위한 문법체계의 주류를 형성하여 왔다. 그러나 구구조문법은 한국어의 자유어순과 중요문장성분 생략 현상을 처리하는데 한계가 있고, 의존문법은 문장의 구조적 정보를 제시할 수 없기 때문에 구문중의성 해결에 있어서 어려움이 있다.

본 논문은 자질기반 어휘화 트리결합문법 (Feature-based Lexicalized Tree Adjoining Grammar) 에 기반한 새로운 한국어 문장분석 시스템을 제안한다. FB-LTAG는 어휘기반 문법으로서 문맥자유문법을 완전하게 어휘화할 수 있으며, 문장의 구조적 정보를 나타내는 구문트리(derived tree) 뿐 아니라, 의존트리(dependency tree)와 같이 단어나 구문간의 의존관계를 나타내는 유도트리(derivation tree)도 얻을 수 있는 장점이 있다. 또한 언어적 특성을 트리구조나 자질구조에 투영함으로써 한국어의 부분자유어순이나 중요문장성분생략과 같은 현상을 처리하기에 적합하다.

본 논문의 구성은 다음과 같다. 먼저 FB-LTAG의 기본이론에 대해 간략하게 소개한 뒤, 본 논문에서 제안하는 FB-LTAG에 기반 한 구문분석 시스템의 전체 구성을 보여준다. 다음으로 구문분석에 필요한 각종 언어자원들의 표기 방법에 대해 말하고, 언어자원의 효율적 구축 방법을 제안한다.

2. Feature-Based Tree Adjoining Grammar

FB-LTAG는 TAG 형식[2]의 바탕 위에 어휘화[3]와 통합기반 자질구조[4] 이론이 포함된 문법체계이다. 그래서 TAG, LTAG, FB-LTAG는 형식면에서는 차이가 없다. 일반 문법과는 다르게, TAG에서 다루는 기본 단위는 문자열(strings)이 아니라, 트리라는 구조체(objects)이다. TAG는 5개의 부분 (Σ, NT, I, A, S) 으로 이루어져 있는데, Σ 는 terminal symbol의 유한집합, NT 는 non-terminal symbol의 유한집합($\Sigma \cap NT = \varnothing$), S 는 특별한(distinguished) non-terminal symbol의 집합($S \in NT$), 그리고 I 와 A 는 각각 초기트리(initial tree)와 보조트리(auxiliary tree)를 나타내며 이들을 통틀어 기본트리 (elementary tree)라고 한다. 초기트리는 순환구조를 제외한 구문분석 트리의 기본 구조를 나타내고, 보조트리는 순환구조를 표현하는 기능을 한다. 초기트리는 말단(경계부분, frontier)에 ↓ (down arrow) 표식이 붙는 대치노드(substitution node)를 하나 이상 가질 수 있고, 보조트리는 말단에 * (asterisk) 표식이 붙는 non-terminal symbol을 반드시 하나 가지는데 이를 풋 노드(foot node)라고 한다.

TAGs의 여러 변형들 중 Lexicalized TAGs는 초기트리와 보조트리의 말단에 적어도 하나의 terminal symbol이 있어야 하며, Feature Structure Based TAGs에서는 기본트리의 각 노드마다 자질구조가 연결되어 있다. 자질구조에는 그 노드가 다른 노드와 어떻게 상호작용 하는지에 대한 정보 즉 중요한 언어적 장치 및 제약조건이 들어있으며, HPSG, GPSG, LFG와

같은 문법들과 마찬가지로, 구문분석시에 자질통합이 일어나면서, 언어적 특성 및 제약을 실현한다.

FB-LTAG에서는 기본트리들이 대치(substitution)와 결합(adjoining)이라는 두 가지 연산에 의해서 통합되면서 문장전체의 구조를 나타내는 구문트리로 발전한다. 이 두 연산이 행해질 때, 자질통합도 함께 이루어지며 이러한 과정이 FB-LTAG에서의 구문분석 과정이다. 그림 1과 2는 FB-LTAG에서의 대치연산과 결합연산 및 각 연산에서의 자질통합[1]을 보여준다.

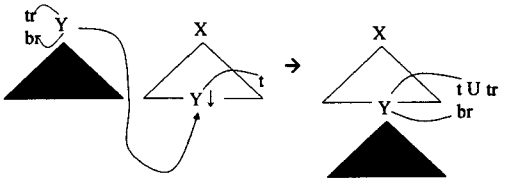

그림 1. 대치연산 (Substitution)

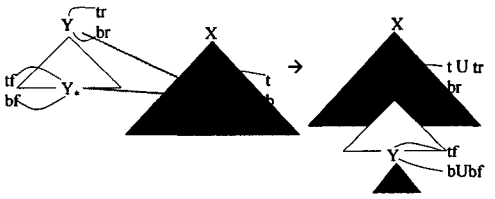

그림 2. 결합연산 (Adjoining)

다음은 LTAG문법에 의해" 민이는 꽃을 매우 좋아한다" 라는 문장을 분석한 예이다. 그림 3은 문장의 각 어절에 대한 기본트리를 보여주고, 그림 4는 대치연산과 결합연산에 의한 위 문장의 전체 구문트리를 보여준다. 그리고 그림 5는 연산과정을 저장하여 트리로 표현한 유도트리이다. 유도트리의 노드에서 트리이름 오른쪽의 숫자는 연산이 일어난 노드의 주소를 나타낸다. 유도트리는 의존트리와 비슷하게 어휘들간의 의존관계를 나타낸다.

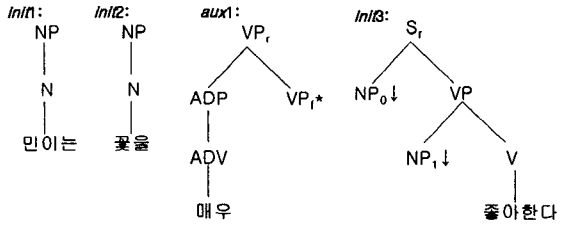

그림 3. 기본트리들 (Elementary Trees)

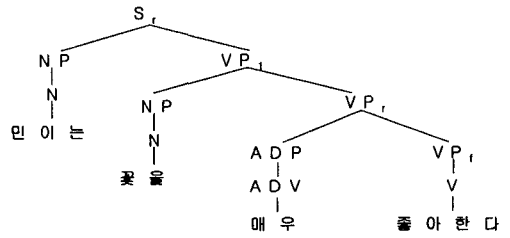

그림 4. 구문트리 (Derived Tree)

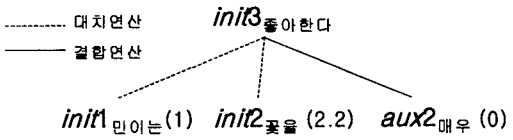

그림 5. 유도트리 (Derivation Tree)

3. FB-LTAG기반의 문장분석 시스템

기존의 FB-LTAG 구문분석기로는 펜실바니아 대학교(University of Pensylvania)의 인지과학연구소에서 진행중인 XTAG 프로젝트의 결과물인 XTAG 시스템이 있다[5][6].[2] 그리고, 정의석[7]은 LTAG를 이용한 역방향 한국어 구문분석기를 구현하였다. 본 논문에서 제안하는 문장분석 시스템의 전체 구성은 다음과 같다.

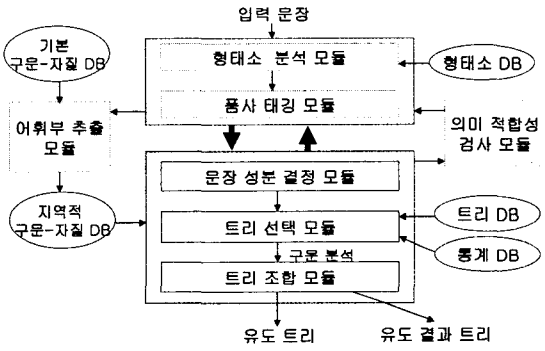

그림 6. FB-LTAG기반의 문장분석 시스템

제안된 시스템은 펜실바니아 대학의 Korean XTAG을 확장한 것으로, 문장성분결정모듈(constituent decision module)과 의미적합성검사모듈(concept adapter)이라는 새로운 모듈을 가진다.

4. 언어자원(linguistic resources)

이 장에서는 FB-LTAG를 위한 언어자원들을 소개하고, 그것들의 표기방법 및 구축방법을 제안한다.

[1] 자질구조는 상위노드와 관련된 정보를 포함하는 top 부분과 하위노드의 정보를 포함하는 bottom부분으로 구성된다.(단, 대치노드는 top 노드만을 가진다) t는 top자질구조를, b는 bottom자질구조를 나타내며, tr과 br은 각각 root 노드의 top과 bottom자질구조를, tf와 bf는 foot 노드의 자질구조를 나타낸다. U는 자질통합연산(unification)을 나타낸다.

[2] 영어, 한국어, 중국어에 대한 각각의 XTAG시스템이 존재하며, 현재 연구단계에 있다. Korean XTAG 시스템의 경우, 이를 이용한 한영 기계번역을 위한 연구가 진행중이다.

4.1 FB-LTAG를 위한 언어자원

본 논문에서 제안하는 시스템에서 필요로 하는 언어자원은 형태소DB, 구문-자질DB, 그리고 트리DB이다. 형태소DB는 어휘의 원형에 대한 정보를 가지는 언어자원으로서, 해당 어휘를 닻노드(anchor node, 기본트리에서 어휘아이템을 가지는 노드)로 가지는 기본트리로의 사상(mapping)을 위한 링크를 가진다. 구문-자질DB는 한국어 구문의 기본 단위인 어절에 대한 자질정보를 저장한다. 용언의 활용과 조사나 접사에 의한 파생 및 조어현상이 두드러진 한국어의 문법 특성상, 모든 어절에 대한 자질정보를 저장하는 것은 비효율적이다. 이를 해결하기 위해서 본 시스템에서는 구문-자질DB를 두 가지로 나누어 효율성을 높이고자 하였다. 먼저 기본 구문-자질DB는 한국어 어절 유형 52가지와 고빈도어절에 대한 자질정보를 수록하고 있다[3]. 고빈도어절 외에 활용이나 파생에 의해 어형변화가 일어난 어절에 대한 자질은 어휘부 추출모듈을 통해 자질정보를 추출하여 지역적 구문-자질DB에 저장된다. 마지막으로 트리DB에는 어휘화된 기본트리가 저장되며, 기본트리들은 같은 하위범주(subcategory)를 가지는 트리들끼리 트리패밀리(tree family)로 묶여진다. 트리패밀리는 어휘아이템이 나타날 수 있는 모든 통사적 환경을 표현한다.

4.2 XML을 사용한 언어자원 표기(Encoding)

TagML은 FB-LTAG 시스템을 위한 언어자원을 표기하고 활용하기 위해 XML[4] DTD로 정의된 마크업언어이다[8]. 본 연구에서는 TagML을 수정하여, 특별히 한국어처리를 위해 최적화된 KtagML을 XML스키마로 정의하였다. KtagML에는 한국어의 풍부한 어형변화를 처리하기 위해 구문-자질DB에 대한 정의가 첨가되었고, 문장성분자질과 내포문자질, 대등문자질, 그리고 각종 의미자질 등이 첨가되었다.

4.3 언어자원 구축

언어자원의 구축은 대상 언어에 대한 언어학적 판단 능력이 있는 사람에 의해 정확하게 구축되어야 한다. 잘못된 언어자원을 이용하여 분석을 하면 잘못된 결과를 얻을 수 있기 때문이다. 본 연구에서는 국립국어연구원의 세종전자사전[9]으로부터 어휘 및 그 어휘의 자질정보, 또 용언의 하위범주정보를 추출하여 보여주고 이를 사용자가 선택 또는 수정할 수 있게 하는 언어자원구축도구를 구현하였다. 그림 7은 언어자원구축도구의 인터페이스이다. 이를 이용하여 사전구축자는 세종전자사전으로부터 추출된 정보를 참조하여 반자동으로 언어자원을 구축할 수 있다. 세가지 언어자원은 각각 XML 문서 형태로 저장되기 때문에 데이터의 변환이 용이하다. 본 연구에서는 현재까지 세종전자사전에 수록된 단어 중 명사 1000개와 용언 800개에 대한 형태소 DB를 구축하였고, 한국어 어절유형 52가지와 고빈도어절 10000개에 대한 기본 구문-자질DB를 구축하였다. 그리고 세종용언사전의 격틀정보를 참고하여 트리패밀리를 규정하고 트리DB를 구축하였다.

[3] http://nlp.kookmin.ac.kr/
[4] http://www.w3.org/

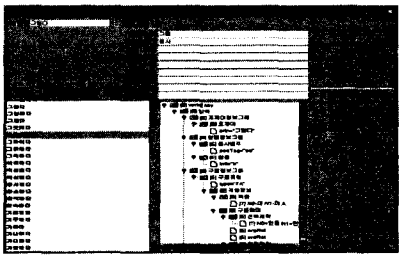

그림 7. 언어자원 추출 및 편집기

5. 결론 및 향후과제

본 논문은 FB-LTAG를 이용한 한국어 문장분석 시스템을 제안하고, 이에 필요한 언어자원을 표기 및 구축하는 방법을 제시하였다. 한국어의 효율적인 문장분석을 위해 기존의 XTAG 시스템에 문장성분결정모듈과 의미적합성검사모듈을 첨가하였다. 그리고 언어자원을 표기하기 위해 KtagML이라는 마크업언어를 정의하고, 이 정의에 따라 세종전자사전으로부터 어휘 및 자질, 그리고 하위범주정보를 추출하여 언어자원을 구축하였다.

이 연구는 FB-LTAG를 이용한 한국어 문장분석 연구의 기초단계의 연구이다. 앞으로 이 연구에서 구축된 언어자원을 활용하여 구문분석기를 실제로 구현하고 그 성능을 평가하는 연구를 진행하고자 한다. 언어자원은 앞으로 지속적인 연구와 경험에 의해 많은 수정과 보완이 이루어져야 할 것이다. 그리고 이를 이용한 정보추출 및 생성에 대한 연구도 향후과제로 남겨두고자 한다.

참고문헌

[1] 김영택, 자연어처리, 생능출판사, 2001.
[2] A. K. Joshi, L. Levy, and M. Takahashi, Tree Adjoining Grammars, Journal of Computer and System Sciences, 1975.
[3] Y. Schabes, A. Abeillé, and A. K. Joshi, Parsing Strategies with ' lexicalized' grammars: Application to Tree Adjoining Grammars, in Procedings of the 12th International Conference on Computational Linguistics (COLING ' 88), Budapest, Hungry, 1988.
[4] K. Vijay-Shanker, and A. K. Joshi, Unification Based Tree Adjoining Grammars, in J.Wedekind, ed., Unification-Based Grammars, MIT Press, Cambridge, Massachusetts, 1991.
[5] The XTAG Research Group, Lexicalized Tree Adjoining Grammar for English, Institute for Research in Cognitive Science, February 26, 2001.
[6] Chung-hye Han, Juntae Yoon, Nari Kim and Martha Palmer, A Feature-Based Lexicalized Tree Adjoining Grammar for Korean, September, 2000.
[7] 정의석, LTAG를 이용한 한국어 구문분석에 관한 연구, 연세대학교 석사학위논문, 1998.
[8] Patrice Bonhomme and P. Lopez, TagML: XML encoding of Resources for Lexicalized Tree Adjoining Grammars, Proceedings of LREC 2000, Athens, May 2000.
[9] 홍재성 외, 21세기 세종계획 전자사전 개발분과 연구보고서, 문화관광부 국립국어연구원, 2001.