한국어 구문분석의 효율성을 개선하기 위한 구문제약규칙의 학습

(Learning Syntactic Constraints for Improving the Efficiency of Korean Parsing)

박 소 영 † 곽 용 재 † 정 후 중 †† 황 영 숙 † 임 해 창 †††

(So-Young Park) (Yong-Jae Kwak) (Hoo-Jung Chung) (Young-Sook Hwang) (Hae-Chang Rim)

요 약 본 논문에서는 한국어 구문분석에 적합한 다양한 구문정보에 대해 살펴보고, 이를 바탕으로 학습한 제약규칙을 이용하여 구문분석모델의 효율성을 개선시키는 방법을 제안한다. 제안하는 방법의 특징은 다음과 같다. 첫째, 제약규칙을 이용하여 불필요한 중간결과물의 생성을 제약하므로, 구문분석의 효율성이 향상된다. 둘째, 제약규칙의 학습에 이용되는 구문정보가 한국어의 특성을 적절히 반영하고 있으므로, 한국어 문장에 대해 비교적 견고하게 분석할 수 있다. 셋째, 제약규칙은 결정트리 학습알고리즘에 의해 말뭉치에서 자동으로 학습되므로, 제약규칙의 획득이 용이하다. 제약규칙을 이용하여 실험한 결과 구문분석모델의 과생성이 1/2~1/3로 줄고 처리속도가 2~3배 빨라졌다.

키워드 : 한국어 구문분석, 제약규칙, 결정트리 학습 알고리즘

Abstract In this paper, we observe various syntactic information for Korean parsing and propose a method to learn constraints and improve the efficiency of a parsing model by using the constraints. The proposed method has the following three characteristics. First, it improves the parsing efficiency since we use constraints that can prevent the parser from generating unsuitable candidates. Second, it is robust on a given Korean sentence because the attributes for the constraints are selected based on the syntactic and lexical idiosyncrasy of Korean. Third, it is easy to acquire constraints automatically from a treebank by using a decision tree learning algorithm. The experimental results show that the parser using acquired constraints can reduce the number of overgenerated candidates up to 1/2~1/3 of candidates and it runs 2~3 times faster than the one without any constraints.

Key words : Korean Parsing, Constraints, Decision Tree Learning Algorithm

1. 서 론

문장성분은 서로 유기적으로 결합하여 문장을 구성하는데, 구문분석은 문법을 바탕으로 이를 분석하는 것을 목적으로 한다. 그러므로, 구문분석은 다양한 언어현상을 포함하는 문장에 대해 견고하게 분석하고, 올바른 분석결과만을 허용하여 효율적으로 분석해야 한다. 그런데, 구문분석에서 견고성을 강조하면 불필요한 중간결과물이 많이 생성되어 효율성이 떨어지고, 효율성을 강조하면 필요한 중간결과물이 생성되지 않아 견고한 분석이 어려워진다[1]. 따라서, 효과적인 구문분석을 위해서는 견고성과 효율성의 균형을 고려하여 구문분석모델을 구성해야 한다.

이러한 점을 고려하여 그동안 다양한 구문분석모델이 제안되었는데, 이들은 언어이론기반 접근방법, 확률기반 접근방법, 제약기반 접근방법으로 분류될 수 있다[2]. 먼저, 언어이론기반 접근방법은 언어학적 이론이나 가설을 바탕으로 구문분석모델을 구성하여 문장을 분석하는 방법이다. 그러나, 이 방법은 언어학적으로 정교하게 접

· 이 논문은 2000년도 한국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2000-V01452-E00305).

† 학생회원 : 고려대학교 컴퓨터학과
ssoya@nlp.korea.ac.kr
yjkwak@nlp.korea.ac.kr
yshwang@nlp.korea.ac.kr
†† 비 회 원 : 고려대학교 컴퓨터학과
hjchung@nlp.korea.ac.kr
††† 종신회원 : 고려대학교 컴퓨터학과 교수
rim@nlp.korea.ac.kr

논문접수 : 2001년 3월 8일
심사완료 : 2002년 7월 16일

근해야하므로 모델이 복잡하여 구현이 어렵고, 중간결과물이 많이 생성될 수 있다는 문제점이 있다[2]. 이를 개선하기 위해서 확률기반 접근방법과 제약기반 접근방법이 등장하였다. 확률기반 접근방법은 여러 구문구조에서 가장 확률이 높은 구문구조를 문장의 구문구조로 선택한다. 이 방법은 필요한 정보를 말뭉치에서 학습하므로 모델의 구현이 용이하고, 다양한 문장에 대해 비교적 견고하게 분석할 수 있다[3]. 그러나 이 방법은 일반적으로 문법이 불필요한 중간결과물을 생성하는 것을 허용한다는 한계가 있다[2]. 한편, 제약기반 접근방법은 구문분석의 효율성을 고려하여 불필요한 중간결과물의 생성에 제약을 가한다[4]. 하지만, 제약규칙을 획득하는데 수작업을 필요로 하므로, 구문분석모델의 개발과 확장이 어렵다는 단점이 있다.

따라서, 본 논문에서는 한국어 구문분석을 위한 기존의 확률모델에 제약규칙을 도입하여 기존모델의 효율성을 개선시킬 수 있는 방법을 제안한다. 제안하는 방법은 중간결과물의 생성여부에 대해 제약규칙을 이용하여 점검하므로, 불필요한 중간결과물의 생성을 미리 방지할 수 있다. 이러한 제약규칙은 결정트리 학습알고리즘에 의해 말뭉치에서 자동으로 학습되는데, 한국어 언어현상을 고려한 구문정보를 이용하므로 한국어 문장에 대해 비교적 견고하게 분석할 수 있다.

앞으로 2장에서는 한국어의 특성과 함께 구문분석모델과 관련된 기존연구를 알아보고, 3장에서는 구문분석 방법에 대해 간단히 살펴본다. 그리고, 4장에서는 제안하는 구문제약규칙의 학습방법에 대해 자세히 설명하고, 5장에서는 제안하는 방법이 구문분석모델의 성능을 얼마나 개선시켰는가를 실험을 통해 평가한다. 마지막으로 6장에서는 결론을 맺는다.

2. 관련연구

효과적인 구문분석을 위해서는 한국어 문장에서 나타나는 다양한 언어현상의 특성을 반영하여 구문분석모델을 구성해야한다. 이때, 구문분석모델은 견고성과 효율성의 사이에서 균형을 이루어야한다. 이를 고려하여 한국어에 적합한 구문분석모델을 개발하고자 하는 많은 연구가 있었다.

2.1 한국어의 특성

한국어 문장을 분석하는 가장 기본적인 방법은 주어, 목적어, 서술어와 같은 문장성분의 결합관계를 CFG규칙으로 간단히 표현하고, 이를 이용하여 구문을 분석하는 것이다. 그런데, 한국어에서는 어순이 비교적 자유롭고 생략현상이 자주 나타난다[5]. 따라서, 이러한 방법으로 구문분석을 하면 정보가 불충분하여 중간결과물이 지나치게 많이 생성되고, 이로 인해 구문분석의 효율성이 떨어지게 된다[1].

이를 보완하기 위해서 구문정보와 함께 다양한 정보를 이용하여 한국어 문장성분의 특성을 나타내야 한다[5]. 즉, 의미, 기능, 길이 등의 다양한 관점에서 문장성분의 특성을 표현하여 문장을 분석하여야 한다. 예를 들어, "영희는 예쁜 인형을 좋아한다."에서 "예쁜 인형을"은 두 개의 어절로 구성되어 있으며, 명사 "인형"이 그 의미를 대표하고, 목적격 조사 "-을"에 의해 목적어 기능을 수행한다고 분석할 수 있다. 이러한 정보는 구문분석의 과정에서 매우 유용하게 이용될 수 있다.

한편, 한국어 문장에서 문장성분은 수식관계, 대등관계, 하위범주관계와 같은 다양한 형태로 결합하는데, 결합관계에 따라 고려되는 정보가 다르다. 예를 들어, "예쁜 영희"는 관형어 "예쁜"이 명사 "영희"를 수식하는 수식관계이며, 이러한 수식관계는 문장성분의 어휘정보나 품사정보에 영향을 받는다. 그리고, "철수가 달린다"는 주어 "철수가"가 서술어 "달린다"의 논항이 되는 하위범주관계이며, 이러한 하위범주관계에서는 서술어가 주어나 목적어를 필요로 하는지 아닌지에 대한 정보가 유용하다. 또한, "예쁜 영희와 멋진 철수"는 명사구 "예쁜 영희"와 "멋진 철수"가 서로 대등하게 결합하는 대등관계인데, 이러한 대등관계에서는 두 문장성분의 유사성을 나타내는 정보가 중요한 역할을 한다. 이와 같이, 각 언어현상은 서로 다른 특성을 가지고 있으므로, 서로 구분하여 처리할 필요가 있다.

이러한 점을 고려하여 본 논문에서는 각 언어현상과 구문정보의 관계를 적절히 고려하여 제약규칙을 학습한다. 즉, 견고하게 구문분석을 하면서 불필요한 중간결과물이 생성되지 않도록, 결합후보의 구문범주 뿐만 아니라 품사나 길이와 같은 다양한 구문정보를 점검하여 결합여부를 판단한다.

2.2 기존의 구문분석 모델

견고성과 효율성을 고려하여 그동안 다양한 구문분석모델이 제시되었는데, 대표적으로 언어이론기반 접근방법, 확률기반 접근방법, 제약기반 접근방법이 있다. 첫째, 전통적인 방법인 언어이론기반 접근방법은 언어학적 이론이나 가설을 바탕으로 구문분석모델을 구성하고, 이를 이용하여 문장을 분석한다. 그런데, 이론이나 가설 자체가 언어학자의 주관적인 문법적 판단에 영향을 받을 수 있다[2]. 그리고, 다양한 문장에 대해 견고하게 구문분석을 하기 위해서는 구문분석모델을 언어학적으로 매우 정교하게 구성하여야 하므로, 구문분석모델을

개발하고 확장하는데 많은 어려움이 따른다[2,6]. 일반적으로 이러한 방법은 분석과정이 복잡할 뿐만 아니라 중간결과물이 과생성되어 효율성이 떨어진다는 문제점이 있다[6].

둘째, 확률기반 접근방법은 생성된 여러 구문구조 후보에서 가장 확률이 높은 구문구조를 문장의 구문구조로 선택하여 중의성을 해결하는 방법이다[3]. 최근에는 어휘정보를 고려한 모델[7,8,9], 문맥정보를 고려한 모델[10,11], 구문구조의 생성과정을 고려한 모델[12]과 같이 다양한 정보를 반영한 구문분석모델이 제안되고 있다. 이러한 확률기반 접근방법은 객관적인 언어정보를 포함하는 말뭉치에서 언어정보와 확률정보를 학습하므로, 구문분석모델의 개발과 확장이 용이하다. 그리고, 구문구조 하나만을 선택하므로 최종결과물의 중의성문제가 자연스럽게 해결된다[3]. 그러나, 문법이 불필요한 중간결과물을 생성하도록 허용하며[2], 자료부족문제를 완화하기위해 도입된 평탄화방법이 중간결과물의 과생성을 더욱 가중시킨다.

셋째, 제약기반 접근방법은 여러 구문구조 후보에서 부적절한 구문구조를 제약하는 방법이다[4]. 즉, 중간결과물의 과생성에 대해 생성전 회피하는 방법[13], 생성후 제거하는 방법[14], 생성시 벌점을 부여하는 방법[4,15] 등이 있다. 이를 위해, 사람의 직관력에 의존하여 제약규칙을 만들거나[4,15,16], HPSG와 같은 이미 구축된 정교한 문법에서 제약규칙을 자동으로 추출할 수 있다[13,14]. 이러한 제약규칙이 중간결과물의 과생성을 제약하여, 구문분석모델의 효율성을 향상시킨다[6]. 그러나, 제약규칙이 사람의 직관력이나 문법에 의존하므로, 다양한 문장에 대해 적용율이 높은 제약규칙을 획득하는 것이 쉽지 않다. 이를 보완하여, 말뭉치를 바탕으로 서로 조합가능한 하나이상의 CFG 규칙을 하나의 규칙으로 학습하여, 문법을 특성화하는 방법[2]이 제안되고 있다. 그러나, 이러한 방법은 어순이 비교적 고정적인 영어나 불어에 적용가능하지만, 어순이 자유롭고 생략현상도 빈번하게 나타나는 한국어에 그대로 적용하기에는 많은 어려움이 따른다. 게다가, 구문분석에 유용한 여러 정보를 효과적으로 활용하지 않았다는 한계가 있다.

따라서, 본 논문에서는 확률기반 접근방법에 제약기반 접근방법을 도입하여 효율성을 개선시키는 방법을 제안한다. 즉, 제안하는 방법은 한국어 구문분석을 위한 기존의 확률모델을 크게 수정하지 않고, 제약규칙을 활용하여 불필요한 과생성을 줄인다. 또한, 제안하는 방법은 사람의 직관력이나 기존문법에 의존하는 제약기반 접근방법을 보완하여, 말뭉치를 바탕으로 제약규칙을 자동으로 학습하므로 제약규칙의 획득 및 확장이 매우 용이하다. 그리고, 하나의 규칙에 좀더 많은 문장성분의 특성을 표현하려는 문법 특성화 방법과 달리, 두 문장성분이 포함하고 있는 다양한 자질정보를 중심으로 제약규칙을 구성하므로, 어순이 자유롭고 생략현상이 빈번하게 일어나는 한국어에 대해 견고하게 분석할 수 있다.

3. 제약규칙을 이용한 구문분석

구문분석에 이용되는 문법은 한국어 자질기반 문법[17]이며, 구문분석 방법에 대해 간단히 소개하면 다음과 같다. 먼저 문법은 G=(T,FS,FO,S)로 정의된다. 이때 T는 품사부착결과인 단말노드집합이고, FS는 자질구조 집합이며, FO는 자질연산집합으로서 { $\alpha \rightarrow \beta$ | $\alpha \in$ {S}∪FS, $\beta \in$T } ∪ { $\alpha \rightarrow \beta\ \gamma$ | $\alpha \in$ {S}∪FS, $\beta, \gamma \in$FS, β나 γ는 중심어 }의 부분집합이고, S는 문장시작 자질구조이다. 우선, 각 형태소-품사 쌍에 대해 자질구조를 할당하고, 인접한 두 자질구조에 대한 자질연산을 수행하여 부모 자질구조를 생성하며, 이러한 과정을 통해 문장을 분석한다. 이때 부모자질구조는 중심자질구조를 바탕으로 결정되는데, 각 자질에 따라 중심자질은 다를 수 있다.

예를 들어, "철수가 TV를 보면서 밥을 먹는다"라는 문장을 분석해보자, 먼저, "철수가", "TV를", "보면서", "밥을", "먹는다"에 대한 자질구조가 [그림 1]과 같이 간단히 구성된다고 가정하자. 이를 "서술어 → 논항 서술어" 규칙과 "서술어 → 서술어 서술어" 규칙에 해당하는 자질연산을 수행하면, [그림 3]과 같이 올바른 분석 "(철수가 ((TV를 보면서) (밥을 먹는다)))"와 틀린 분석 "((철수가 (TV를 보면서)) (밥을 먹는다))"가 나타날 수 있다.

그런데, 대등구조는 유사한 성격의 하위성분을 선호한다는 특성을 반영하여 제약하면, 틀린 분석은 제거되

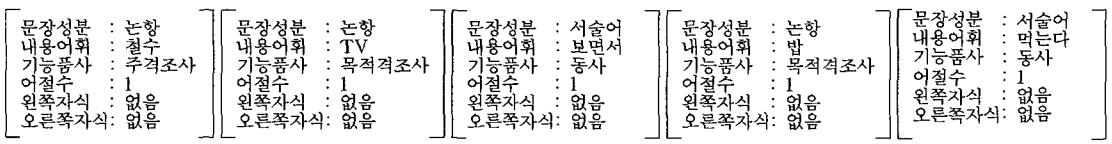

그림 1 초기화된 자질구조

```
IF (좌기능품사=동사) & (우기능품사=동사) & (좌우어절수차=0)   THEN 자질연산 허용
ELSE IF (좌기능품사=동사) & (우기능품사=동사) & (좌우어절수차>0)   THEN 자질연산 불허
...
```

그림 2 "서술어 → 서술어 서술어"를 위한 제약규칙의 예

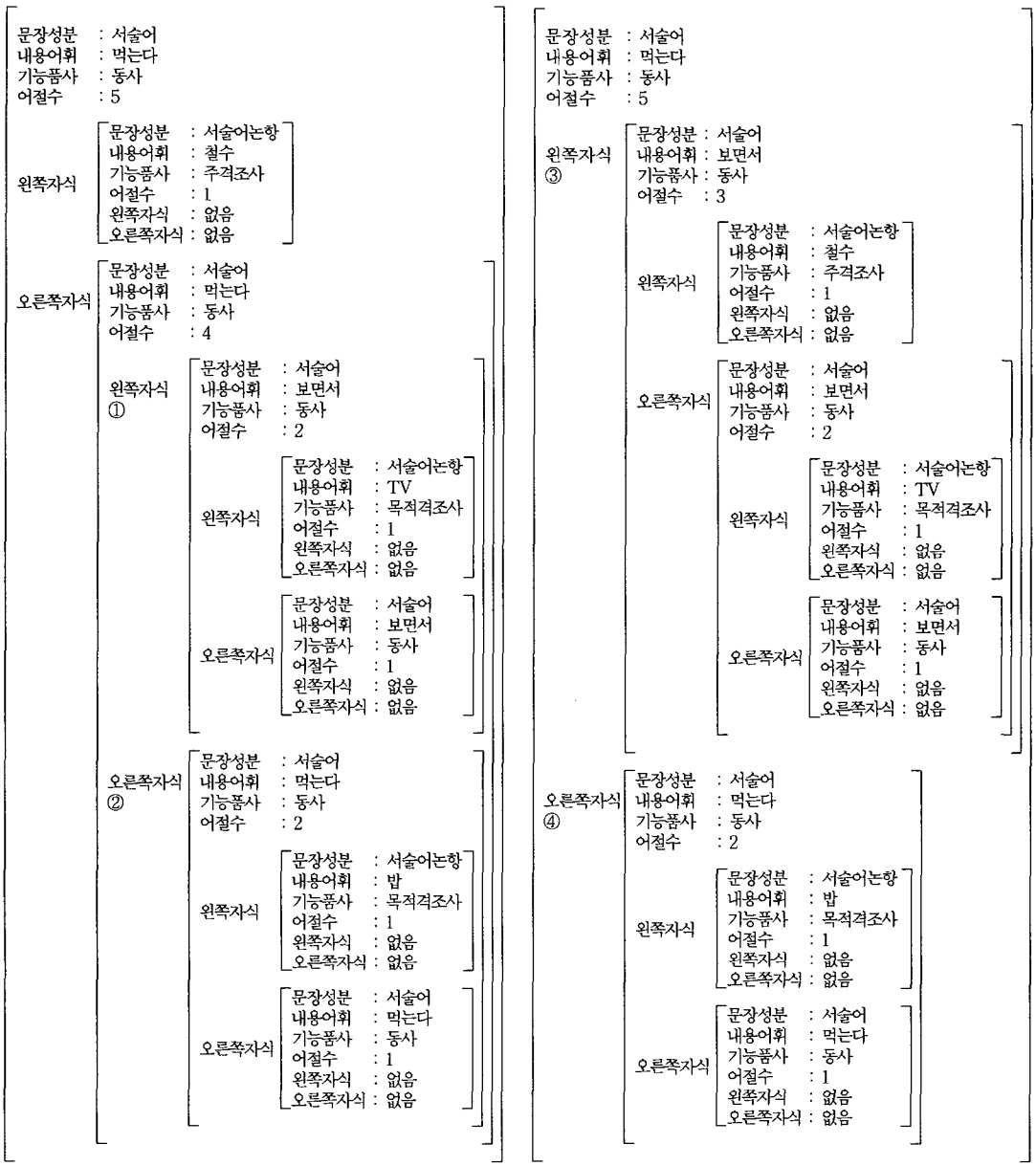

그림 3 "철수가 TV를 보면서 밥을 먹는다"의 구문분석결과

고 올바른 분석만 허용된다. 다시 말해서, "서술어 → 서술어 서술어"에 대한 제약규칙이 [그림 2]와 같이 주어진다고 가정하자.

[그림 3]에서 왼쪽자질구조①과 오른쪽자질구조②에 대해 제약규칙을 점검하면, 좌기능품사(보면서)=동사, 우기능품사(먹는다)=동사, 좌우어절수차=2-2=0이므로, 자질연산을 허용하여 부모자질구조가 생성된다. 하지만, 왼쪽자질구조③과 오른쪽자질구조④에 대해 제약규칙을 점검하면, 좌우어절수차=3-2=1이므로 자질연산을 불허한다. 이와 같이, 주어진 문장성분의 구문적 특성이 반영된 자질값을 미리 점검하여 자질연산의 적용을 회피하면, 중간결과물의 과생성을 줄일 수 있다. 다음장에서는 이러한 제약규칙을 어떻게 획득하는지를 구체적으로 살펴본다.

4. 구문제약규칙

문장성분의 기본정보만을 이용하여 구문분석을 수행하면, 중간결과물이 지나치게 많이 생성된다는 문제점이 발생한다. 따라서, 불필요한 자질연산 수행을 억제할 수 있는 다양한 구문정보를 자질구조에 포함하고, 자질연산을 수행할 때 이를 점검하여 제약할 필요가 있다. 이 장에서는 불필요한 중간결과물을 줄이는데 유용한 구문정보의 종류와 이를 자동으로 획득하는 방법에 대해 살펴본다.

4.1 구문제약정보

구문분석과정에서 발생하는 중간결과물의 과생성을 줄이기 위해서는 문장성분이 포함하고 있는 다양한 정보를 활용해야 한다. 제안하는 방법은 구문, 내용, 기능, 길이, 형태, 격, 일치에 관한 정보를 제약정보로 이용하여 문장성분을 표현한다. 그리고 이러한 구문정보를 제약규칙에서 다룰 수 있도록 자질구조는 [그림 4]와 같이 구성된다. 각 구문제약정보에 대해 자세히 살펴보면 다음과 같다.

첫째, 구문자질(SYNC)은 문장성분의 기본적인 구문범주정보를 나타내며, 문장(S), 명사구(NP), 논항(PP), 서술어(VP), 관형어(DP), 부사어(AP), 독립어(IP)로 문장성분을 구분한다. 그리고, 문장성분의 기능을 나타내는 형식형태소는 P 대신 -로 표현된다. 예를 들어, 관형어 "철수의=철수+의"에서 고유명사 "철수"는 실질형태소이므로 구문범주가 NP이고, 관형격조사 "+의"는 형식형태소이므로 구문범주가 D-가 된다.

둘째, 내용자질(CONT)은 문장성분의 의미에 해당하는 정보를 나타내고, 내용중심어의 정보를 포함한다. 예를 들어, "철수의=철수+의"에서 명사 "철수"가 그 의미를 대표하므로, 내용자질 값은 명사 "철수"이다. 이와 같이, 내용자질 값으로 내용중심어의 형태소 어휘를 활용할 수 있다. 그러나, 어휘를 그대로 이용하면 자료부족문제가 심각해지고 제약규칙의 수가 급격히 증가할 수 있으므로, 제안하는 방법에서는 어휘 대신 품사를 이용하여 내용자질을 표현한다.

셋째, 기능자질(FUNC)은 문장성분의 역할을 나타내는 정보로 기능중심어의 정보를 담고 있다. 예를 들어, "영희의 예쁜 인형"에서 "영희의"와 "인형"은 둘 다 명사를 포함하지만, 전자는 관형격조사 "+의"에 의해 관형어 기능을 수행한다. 그리고 "영희의"와 "예쁜"은 둘 다 관형어이지만, 전자는 관형격조사 "+의"에 의해 관형어 역할을 수행하고 후자는 관형형어미 "+ㄴ"에 의해 관형어 역할을 하므로, 이를 구분할 필요가 있다. 이러한 기능자질은 지나치게 단순화되어 있는 구문자질을 보완하기 위한 것이며, 내용자질과 마찬가지로 어휘 대신 품사를 이용한다.

넷째, 길이자질(LENG)은 문장성분을 구성하는 구성요소의 수를 나타내는 정보로 "길이=⌞ 2loge(어절수)+1 ⌟ =2loge(어절수)+1의정수값"의 식을 이용하여 구한다. 이러한 길이정보를 이용하면, 문장성분이 일반적으로 간단하게 구성된다는 특징을 반영할 수 있다. 이때, 길이는 길이변화에 비교적 둔감하도록, 어절수를 로그함수에 적용하고 형태소수 대신 어절수를 바탕으로 결정한다. 예를 들어, "영희는 신문을 본다"는 3개의 어절로 구성되어 있으므로 ⌞ 2loge(3)+1 ⌟=3이 된다.

다섯째, 형태자질(FORM)은 문장성분의 상태를 나타내는 정보로서 base(기본형), fin(종결형), con(연결형),

fs = [
SYNC(구문) : 문장성분1 | 문장성분2 | ... | 문장성분m
CONT(내용) : 품사1 | 품사2 | ... | 품사n
FUNC(기능) : 품사1 | 품사2 | ... | 품사n
LENG(길이) : ⌞2loge(어절수)+1⌟
FORM(형태) : base(기본형) | con(연결형) | fin(종결형) | comma(쉼표형)
CASE(격) : 0000000000000 | 1000000000000 | ..... | 1111111111111
LEFT(좌자식) : NULL | fsL
RIGHT(우자식) : NULL | fsR
]

그림 4 제안하는 자질구조

comma(쉼표형)로 구분한다. 쉼표는 문장성분을 열거하는 기능뿐만 아니라 문맥상 끊어 읽는 기능도 있으므로, con과 comma를 구분한다. 기본형태는 base이고, 종결형 어말어미를 포함한 문장성분의 형태는 fin이며, 연결형 어말어미나 접속격조사를 포함한 문장성분의 형태는 con이고, 쉼표를 포함하는 문장성분의 형태는 comma이다. 예를 들어, 관형형 어말어미를 포함하는 "본=보+ㄴ"는 기본적으로 base이고, 연결형 어말어미를 포함하는 "보고=보+고"는 con이며, 종결형 어말어미를 포함하는 "본다=보+ㄴ다"는 fin이다. 이러한 정보를 이용하면, 보조사는 종결형 서술어와 자주 결합하며 목적어나 주어는 연결형 서술어나 종결형 서술어와 잘 결합한다는 경향을 반영하여 중간결과물의 과생성을 제약할 수 있다.

여섯째, 격자질(CASE)은 서술어와 논항의 하위범주관계를 나타내는 정보로 서술어가 이중주어나 이중목적어를 취하지 않도록 유도한다. 격은 주어, 목적어, 보어 등의 대표적인 13개 격이 비트열로 표현되는데, 각 비트는 1(참)이나 0(거짓)의 값을 가지며 기본값은 0이다. 주어진 문장성분이 필수격이면, 해당격은 참이 되고 나머지는 거짓이 된다. 그리고, 주어진 문장성분이 서술어이면, 서술어가 이미 포함한 격은 참이 되고 나머지는 거짓이 된다. 예를 들어, 주격조사가 포함된 "철수+가"는 10000000000000와 같이 주어 비트만 참이 된다. 또한 주어와 목적어를 포함하는 "철수가 그림을 그린다"는 11000000000000와 같이 주어와 목적어 비트가 참이고 나머지는 거짓이 된다.

일곱째, 일치자질(SAME)은 결합하려는 좌우 자질구조에 대해 각 자질값의 일치여부를 나타내는 것으로 대등구조를 제약할 때 유용하게 이용된다. 그런데, 좌우자질구조의 모든 자질에 대한 일치여부를 이용하면 자료부족문제가 심각하게 나타날 수 있으므로, 제안하는 방법에서는 좌우 자질구조에서 가능자질의 일치여부와 길이차만을 이용한다. 즉, 좌우 자질구조에서 기능자질 값의 일치여부에 대해서 좌우 기능자질 값이 같으면 "일치"값을 부여하고, 좌우 기능자질 값이 다르면 "불일치" 값을 부여한다. 그리고, 길이자질에 대해서는 좌우자질구조의 길이차를 이용한다.

4.2 구문제약정보와 언어현상

한국어 문장에는 다양한 언어현상이 나타나는데, 각 언어현상마다 고려해야할 특성이 서로 다르다. 그런데, 많은 구문제약정보를 이용하여 제약규칙을 학습하면, 학습비용의 부담이 커지고 부적절한 제약규칙이 학습될 수 있다. 그러므로, 효과적인 학습을 위해서는 구문분석의 과생성을 제약하는데 유용한 제약정보를 고려하는 한편, 각 언어현상에 적합한 제약정보를 선별할 필요가 있다. 이 절에서는 수식관계, 대등관계, 하위범주관계에서 발생하는 중간결과물의 과생성에 대해 분석하고, 각 언어현상에 대해 어떤 제약정보를 고려할 수 있는지를 살펴본다.

첫째, 수식관계에서는 구문, 기능, 내용, 길이, 형태의 정보를 바탕으로 제약하여 중간결과물의 과생성을 줄인다. 이는 수식관계에서는 수식어와 피수식어의 일반적인 정보 자체가 중요하고, 격정보나 일치정보는 영향을 끼치지 않는다는 점을 반영한 것이다. 예를 들어, "새 옷을 입은 영희"에서 관형사 "새"는 일반명사는 수식하지만 고유명사는 거의 수식하지 않는다는 특성을 이용하면, 아래에서 (2)는 제거될 수 있다.

(1) ((새) (옷))을 입은 영희

(2) ((새) (옷을 입은 영희))

둘째, 하위범주관계에서는 일반정보뿐만 아니라 격정보도 매우 유용하게 사용된다. 그러므로, 하위범주관계에서는 구문, 기능, 내용, 길이, 형태와 함께 격정보를 이용하여 제약한다. 예를 들어, "철수는 TV를 보고 영희는 신문을 본다"에 대해 구문분석하면 다음과 같이 다양한 분석이 나올 수 있다. (3)에서 "보고"는 주어와 목적어를 필요로 하는 서술어이고, (4)에서 "본다"는 이미 주어와 목적어를 포함하는 서술어이다. 게다가, (3)에서 오른쪽 자질구조의 길이값은 $\llcorner 2\log_e(1)+1 \lrcorner = 1$이며, (4)에서 오른쪽 자질구조의 길이값은 $\llcorner 2\log_e(4)+1 \lrcorner = 3$이다. 이와 같이 격정보나 길이정보를 이용하면, (4)은 제거될 수 있다.

(3) 철수는 ((TV를) (보고)) 영희는 신문을 본다

(4) 철수는 ((TV를) (보고 영희는 신문을 본다))

셋째, 대등관계에서는 일반정보와 함께 좌우 자질구조에서 각 자질값의 일치여부도 중요하게 사용된다. 따라서, 대등관계에서는 구문, 기능, 내용, 길이, 형태 정보뿐만 아니라, 좌우자질구조에 대한 기능일치, 길이차를 이용하여 제약한다. 예를 들어, "철수는 TV를 보고 영희는 신문을 본다"에 대해 구문분석할 때, 다음과 같은 다양한 서술어 대등구조가 나올 수 있다. 그런데, "보고"와 "본다"의 길이차이를 고려하면, 길이차가 동일한 (5)이 선택되고 (6)나 (7)의 결과는 제거될 수 있다.

(5) ((철수는 TV를 보고) (영희는 신문을 본다))

(6) 철수는 ((TV를 보고) (영희는 신문을 본다))

(7) 철수는 TV를 ((보고) (영희는 신문을 본다))

지금까지 한국어 문장에 나타나는 언어현상을 수식관계, 하위범주관계, 대등관계로 구분하고, 각 언어현상에 대해 어떤 제약정보를 활용할 수 있는지를 알아보았다.

표 1 VP → VP VP를 위한 학습집합 일부

학습속성										목적속성
좌기능	좌내용	좌길이	좌형태	우기능	우내용	우길이	우형태	좌우기능	길이차	자질연산
동사	동사	2	연결형	동사	동사	2	종결형	일치	0	허용
동사	동사	3	연결형	동사	동사	2	종결형	일치	1	불허
동사	동사	2	연결형	형용사	형용사	2	종결형	불일치	0	허용
형용사	형용사	3	연결형	형용사	형용사	1	기본형	일치	2	불허
형용사	형용사	1	연결형	동사	동사	2	종결형	불일치	1	허용
서술격조사	일반명사	2	연결형	서술격조사	일반명사	2	종결형	일치	0	허용
...	...	...	...	...	...	...	...	...	...	...

다음 절에서는 이를 바탕으로 어떻게 학습하는지에 대해 살펴본다.

4.3 구문제약규칙의 학습

이 절에서는 귀납적 추론에서 가장 널리 사용되고 실용적인 기계학습방법의 하나인 결정트리 학습방법 C4.5 [18]를 이용하여 제약규칙을 자동으로 학습하는 방법을 살펴본다. 결정트리 학습방법은 일반적으로 각 학습속성에 대해 정보이득량(Information Gain)을 계산하여, 목적속성값에 따라 학습집합을 잘 분류하는 최적속성을 선택하고, 이를 바탕으로 학습집합을 분류한다. 학습집합에 대해 일관성있는 가설을 찾을때까지 최적속성의 선택과정과 학습집합의 분류과정을 반복하여 결정트리를 확장해나간다. 이러한 결정트리 학습방법은 한정된 가설공간이 아닌 완전한 가설공간에서 가설을 탐색하며, 오류가 포함된 집합에도 견고하다. 또한, 결정트리는 if-then 형식으로 표현가능하므로, 학습된 지식이 쉽게 이해될 수 있다[18].

그런데, 결정트리 학습방법에서 사용하는 정보이득량은 속성값의 종류가 적은 속성보다는 속성값의 종류가 다양한 속성을 선호하는 경향이 있다. 이를 보완하기 위해, 속성값이 다양한 속성에 벌점을 부여하는 정보이득률(Information Gain Ratio)을 이용한다. 한편, 학습집합이 충분히 크지 않거나 학습집합에 오류가 포함되면, 학습결과가 학습집합에 과적응되어 나타날 수 있다. 이러한 과적응을 완화하기 위해, 학습된 결정트리에서 오류가능성이 높은 가지를 잘라내는 가지치기 방법을 활용한다[18]. 이러한 가지치기를 통해 일반화가 수행되므로, 오류에 견고하게 학습할 수 있으며 학습집합에서 배제된 자료에 대해서도 처리가 가능하다.

결정트리학습방법을 이용한 제약규칙의 학습은 다음과 같은 순서로 진행된다. 첫째, 구문범주로 구성된 기본규칙에서 구조적 중의성을 심각하게 유발하는 19개의 규칙에 대해 조사하고, 이들을 수식관계, 대등관계, 하위범주관계로 분류한다. 즉, 기본규칙 55개중에서 어절내 언어현상을 표현하는 36개의 규칙을 제외한 어절간 언어현상을 표현하는 19개의 규칙에 대해서 조사한다. 그리고, 각 기본규칙이 어떤 결합관계인지를 바탕으로 적절한 학습속성을 선택한다. 예를 들어, [표 1]에서 "VP → VP VP"는 대등관계이므로, 좌우 자질구조에 대한 구문, 기능, 내용, 길이, 형태의 구문제약정보와 함께 좌우기능의 일치여부, 길이차를 학습속성으로 이용한다. 제약규칙은 기본규칙별로 따로 학습되는데, 이미 규칙에 구문범주가 포함되어 있으므로 학습집합에서 구문정보는 생략한다.

둘째, 분류된 각 기본규칙에 대해 [표 1]과 같이 자질연산의 적용여부에 대해 "허용"값을 갖는 정례와 "불허"값을 갖는 반례를 수집한다. 이를 위해 제약규칙이 없는 구문분석기를 이용해서 학습말뭉치의 문장을 분석하고, 분석결과에서 학습말뭉치와 일치하는 결과는 정례로 추출하고 나머지는 반례로 추출한다. 예를 들어, [그림 3]으로부터 첫번째 학습예제와 두번째 학습예제가 추출될 수 있다. 이를 학습말뭉치와 비교하면, 첫번째 학습예제는 정답이므로 "허용"값을 갖고, 두번째 학습예제는 오답이므로 "불허"값을 갖는다.

셋째, 추출된 정례와 반례에 대해서 결정트리 학습방법을 적용하여 각 기본규칙에 대해 [그림 5]와 같은 결정트리를 획득한다. 이를 위해, 어떤 학습속성이 학습집합을 정례집합과 반례집합으로 가장 잘 분류하는지를 정보이득률로 평가한다. 그리고, 평가결과로부터 최적속성을 선택하고, 최적속성을 기준으로 학습집합을 정례집합과 반례집합으로 분류한다. 분류된 학습집합에 대해 이러한 선택과정과 분류과정을 재적용하여, 결정트리를 점차 확장시킨다. 이는 정례집합과 반례집합이 적절히 분류될 때까지 계속 반복된다[18].

넷째, 학습된 결정트리를 제약규칙으로 구문분석에 적용하여 중간결과물을 줄인다. 기본규칙에 의해 인접한

```
좌우기능 = 일치:
|   길이차 <= 1 :
|   |   좌형태 = 연결형:
|   |   |   길이차 <= 0 :
|   |   |   |   우형태 = 연결형: 불허
|   |   |   |   우형태 = 쉼표형: 불허
|   |   |   |   우형태 = 종결형:
|   |   |   |   |   좌내용 = 일반명사:  불허
|   |   |   |   |   좌내용 = 동사: 허용
|   |   |   |   |   좌내용 = 형용사: 허용
...
```

그림 5 VP → VP VP를 위한 결정트리 일부

두 자질구조가 결합후보로 결정되면, 좌우자질구조의 자질값에 대해 해당 결정트리를 점검한다. 이때, 결정트리의 단말노드 값이 "불허"이면 자질연산을 회피하고, 나머지는 자질연산을 허용한다. 예를 들어, "(철수가 (TV를 보면서)(밥을 먹는다))"를 위한 [그림3]의 자질구조 ①과 ②에 대해 [그림5]의 결정트리를 적용하면, 둘 다 동사이므로 좌우기능=일치, 길이차=2-2=0<=1, 좌형태(보면서)=연결형, 길이차<=0, 우형태(먹는다)=종결형, 좌내용(보면서)=동사이므로, ①과 ②는 자질연산에 적용될 수 있다.

5. 실험 및 평가

제안하는 제약규칙의 성능을 평가하기 위해서 [그림6]과 같은 구문분석기를 구현하였고, 1.2GHz의 Athlon CPU와 256MB의 메모리를 갖추고 Windows 2000이 설치된 시스템에서 실험하였다. 구문분석기는 자질구조 초기화 단계, 구문분석 단계, 중의성해결 단계로 구성된다. 이는 제약규칙과 중의성 해결모델의 평가를 구분하기 위해서 구문분석 단계와 중의성 해결 단계를 분리한 것이다. 먼저, 자질구조 초기화 단계에서는 품사가 부착된 문장을 입력으로 받고 형태소와 품사정보를 바탕으로 자질구조를 초기화한다. 그리고, 구문분석기의 부담을 고려하여, 복합명사와 같은 기본 명사구나 'ㄹ 수 있'과 같은 보조용언구는 하나의 복합구로 묶어서 초기화한다. 다음단계인 구문분석 단계에서는 초기화된 자질구조를 자질연산에 적용하여 구문분석을 하는데, 압축기법[19]을 도입하여 중간결과물의 중복생성을 줄인다. 이때, 결합후보가 되는 좌우 자질구조의 자질값에 대해서 제약규칙을 점검하여 자질연산의 적용여부를 결정한다. 그런데, 구문분석에 실패한 문장에 대해서는 제약규칙을 배제하고 기본규칙만으로 재분석을 시도하여, 구문분석

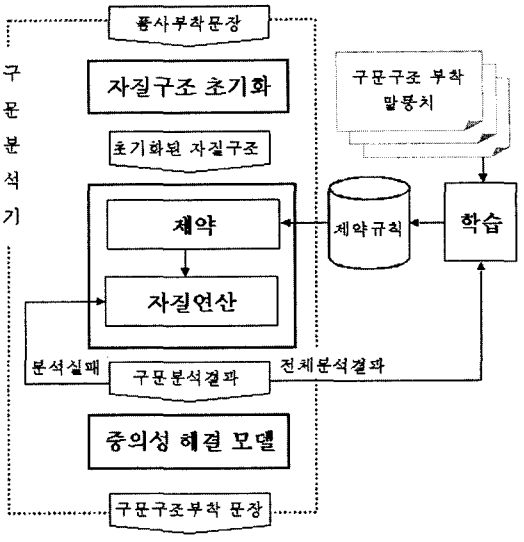

그림 6 제안하는 구문분석 모델

의 견고성을 유지한다. 마지막으로 구문중의성 해결 단계에서는 중의성을 포함하는 구문분석결과에 대해 조건부 연산 GLR 파서[20]를 이용하여 구문구조 중의성을 해결한다.

실험에 이용되는 구문구조 부착 말뭉치는 12,084문장으로 구성되어 있으며, 학습말뭉치는 문장당 평균 22.46개의 형태소를 가지는 10,906문장이고, 실험말뭉치는 문장당 평균 22.93개의 형태소를 포함하는 1,178문장이다. 학습말뭉치의 문장에 대해 제약규칙이 없는 구문분석기를 이용해서, 정례와 반례를 추출하여 학습집합을 만든다.

제안하는 제약규칙을 이용하면 구문분석기의 성능이 얼마나 개선되는가를 평가하기 위해서 [표 3]과 같이 미생성, 과생성, 처리시간, 재현율에 대해 조사하였다. 구문분석기의 견고성을 평가하는 미생성은 [표 2]의 b와 같이 정답 자질구조 중에서 생성되지 않는 자질구조의 수를 말하며, 미생성이 적을수록 견고한 것이다. 구문분석기의 공간적인 효율성을 평가하는 과생성은 [표 2]의 c와 같이 정답이 아닌 자질구조의 수를 의미하며, 과생성이 적을수록 효율적인 것이다. 구문분석기의 시간적인 효율성을 평가하는 처리시간은 입력된 문장에 대

표 2 모델평가를 위한 분할표

말뭉치 \ 모델	생성○	생성×
정답	a	b
오답	c	d

해 구문구조가 생성되는데 걸리는 시간이다. 재현율은 [표 2]에서 $\frac{a}{a+b}$에 해당하며, 필요한 자질구조를 얼마나 잘 생성했는가를 평가한다. 이를 위해 말뭉치와 분석결과를 비교하는데, 구문범주가 같고 동일한 형태소열을 지배하는 자질구조를 정답으로 인정한다[21].

구문분석단계에서는 여러 구문구조를 모두 인정하므로 c의 값이 크지만, 중의성 해결단계에서는 하나의 구문구조만을 선택하므로 c의 값이 작아진다. 따라서, 제약규칙의 불필요한 자질연산에 대한 회피능력과 중의성 해결 모델의 선택능력이 구분되도록, 제약규칙을 적용한 결과에 대한 재현율과 중의성 해결 단계 후의 최종 재현율을 구분하여 평가한다. 그리고, 제안하는 방법은 이진형식의 구문구조를 생성하는데, 이러한 구조는 중의성을 해결하면 재현율과 정확율이 동일하게 나타나므로 재현율로 통합하여 나타낸다.

제약규칙은 [표 3]과 같이 "제약없음", "동일속성", "구분속성"으로 나누어 실험하였다. 우선, "제약없음"은 제약규칙을 이용하지 않고 구문분석한 결과이다. 그리고, "동일속성"은 모든 언어현상에 대해 구문, 내용, 기능, 길이, 형태를 이용하여 학습한 결과를 제약규칙으로 적용한 결과이다. "구분속성"은 수식관계, 대등관계, 하위범주관계의 특성을 고려하여 격자질이나 일치자질을 학습속성에 추가하고, 언어현상별로 제약규칙을 학습하여 제약규칙으로 활용한 결과이다. "동일속성"에서는 28,347개의 제약규칙이 학습되었고 "구분속성"은 32,802개의 제약규칙이 학습되었다.

[표 3]은 제약규칙을 이용하면, 미생성은 약간 늘어나지만 과생성은 1/2~1/3로 줄고 처리시간은 2~3배정도 빨라진다는 것을 보여준다. 게다가, 과생성 감소로 중의성 해결 모델에서 선택해야하는 후보가 줄어들면서, 재현율도 약 2%이상 향상됨을 알 수 있다. 이와 같이, 제약규칙을 이용하면, 구문분석의 처리시간이 문장당 평균 0.26초에서 0.12초로 빨라진다. 그리고, 실험집합에서 "구분속성"을 이용하면 "동일속성"에 비해 과생성이 다소 감소함을 알 수 있다. 그런데, [표 3]에서 학습집합에서도 미생성이 발생하는데, 이는 초기화단계의 복합구 인식과정에서 오류가 포함되어 구문분석과정에 영향을 끼치기 때문이다.

6. 결 론

지금까지 한국어 구문분석에 적합한 다양한 구문정보에 대해 알아보고, 이를 바탕으로 학습한 제약규칙을 이용하여 구문분석모델의 효율성을 개선시키는 방법에 대해 살펴보았다. 본 논문에서 제안하는 방법의 특징은 다음과 같다.

첫째, 구문제약규칙을 이용하여 오분석을 제거하므로, 주어진 문장을 효율적으로 분석할 수 있다. 확률기반 접근방법은 자체적으로 오분석을 제거하지는 않으므로, 중간결과물이 많이 생성된다는 문제점이 있다. 따라서, 제안하는 방법에서는 오분석을 제거하기 위해 구문제약규칙을 이용하여 불필요한 중간결과물의 생성을 제약한다.

둘째, 구문제약규칙을 말뭉치에서 자동으로 학습할 수 있으므로, 구문제약규칙의 획득이 용이하다. 언어이론기반 접근방법이나 제약기반 접근방법은 사람의 직관력에 의존하는 경향이 있으므로, 구문분석모델의 개발, 유지, 보수에 어려움이 따른다. 그러므로, 제안하는 방법은 제약규칙을 쉽게 얻을 수 있도록, 결정트리 알고리즘을 이용하여 제약규칙을 말뭉치에서 자동으로 학습한다. 또한, 학습된 언어학적 지식을 쉽게 파악할 수 있도록, 학습된 제약규칙을 결정트리 학습알고리즘으로 학습하여 if-then 형식으로 표현한다.

셋째, 다양한 언어현상에 대해 적절한 학습속성을 이용하여 구문제약규칙을 학습하므로, 주어진 문장에 대해 비교적 견고하게 분석할 수 있다. 제안하는 방법은 구문, 내용, 기능, 길이, 형태와 같은 다양한 정보를 이용하여 문장성분을 표현하고, 수식관계, 대등관계, 하위범

표 3 실험결과

		제약규칙 적용결과					최종
		정답 (a)	미생성 (b)	과생성 (c)	처리시간 (초)	재현율 (%)	재현율 (%)
학습 10,906문장	제약없음	476,333	1,373	4,639,821	3,345.29	99.71	80.10
	동일속성	474,232	3,474	1,684,937	1,202.82	99.27	83.50
	구분속성	474,162	3,544	1,634,118	1,198.22	99.26	83.64
실험 1,178문장	제약없음	53,890	138	526,734	317.92	99.74	80.03
	동일속성	53,495	533	233,568	153.10	99.01	82.10
	구분속성	53,453	575	221,854	149.50	98.94	82.18

주관계에 대한 특성을 고려하여 학습속성을 구성한다. 이를 이용하여 중간결과물의 생성을 제약하는데, 효율성의 개선 정도에 비해 구문분석모델의 견고성이 크게 떨어지지 않았다.

이러한 특성에도 불구하고, 제안하는 방법은 여전히 오분석 결과를 포함하고 있다. 이러한 한계를 보완하기 위해서, 수식관계, 대등관계, 하위범주관계 뿐만 아니라 어절 내에서 일어나는 파생현상이나 굴절현상에 대해서도 제약규칙을 적용하는 연구가 필요하다. 그리고, 구문구조의 중의성 해결에 큰 영향을 끼치는 형태소 어휘를 적절히 활용하는 연구도 함께 진행되어야 한다.

참 고 문 헌

[1] 나동렬, "한국어 파싱에 대한 고찰", 한국정보과학회지, 제12권, 제8호, pp.33-46, 1994.

[2] Nicola Cancedda, Christer Samuelsson, "Experiments with Corpus-based LFG Specialization," In Proceedings of the Sixth Applied Natural Language Processing Conference, pp.204-209, 2000.

[3] 서정연, 김창현, "통계적 방법을 이용한 구문분석," 한국정보과학회지, 제14권, 제7호, pp.58-70, 1996.

[4] Wolfgang Menzel, Ingo Schroder, "Decision Procedures for Dependency Parsing Using Graded Constraints," In Proceedings of COLLING-ACL Workshop on Processing of Dependency Grammars, pp.78-87, 1998.

[5] 이공주, "언어특성에 기반한 한국어의 확률적 구문분석", 한국과학기술원 박사학위 논문, 1997.

[6] A. Voutilainen. "Three Studies of Grammar-Based Surface Parsing of Unrestricted English Text," PhD thesis, University of Helsinki, 1994.

[7] David M. Magerman, "Statistical Decision-Tree Models for Parsing," In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pp276-283, 1995.

[8] 김학수, 서정연, "어휘 의존 정보에 기반한 한국어 통계적 구문분석기", 한국정보과학회 인공지능 연구회 봄 학술발표 논문집, pp.61-65, 1997.

[9] Michael Collins, "Head-Driven Statistical Models for Natural Language Parsing," Ph.D. Thesis, University of Pennsylvania, 1999.

[10] David M. Magerman, Carl Weir, "Efficiency, Robustness and Accuracy in Picky Chart Parsing," In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, pp.40-47, 1992.

[11] 조정미, 서정연, 김길창, "말뭉치로부터 자동 추출된 문맥 반영 구문규칙을 이용한 영어 구문 분석", 한국정보과학회논문지, 제21권, 제9호, pp.1702-1710, 1994.

[12] Ezra Black, Fred Jelinek, John Lafferty, David M. Magerman, Robert Mercer, Salim Roukos, "Towards History-based Grammars: Using Richer Models for Probabilistic Parsing," In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pp.31-37, 1993.

[13] Bernd Kiefer, Hans-Ulrich Krieger, John Carroll, Rob Malouf, "A bag of useful techniques for efficient and robust parsing," In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp.473-480, 1999.

[14] Kentaro Torisawa, Kenji Nishida, Yusuke Miyao, Jun-Ichi Tsujii, "An HPSG Parser with CFG filtering," Natural Language Engineering, Vol 6, Part 1, pp.63-80, 2000.

[15] 권혜진, 이원일, 이근배, 이종혁, "범주문법에 기반한 한국어의 의미분석", 한국정보과학회 봄 학술발표 논문집, pp.915-918, 1996.

[16] J. T. Maxwell, R. M. Kaplan, "The interface between phrasal and functional constraints," Computational Linguistics, Vol.19, Num.4, pp.571-590, 1993.

[17] 박소영, 황영숙, 임해창, "X-바 이론의 중심어 개념을 도입한 형태소 단위의 한국어 자질기반 문법," 한국정보과학회 논문지(B), 제26권 제10호, pp.1247-1259, 1999.

[18] J. Ross Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, 1993.

[19] Masaru Tomita, "Efficient Parsing for Natural Languge: A Fast Algorithm for Practical Systems," Kluwer Academic Publishers, 1986.

[20] Yong-Jae Kwak, Young-Sook Hwang, Hoo-Jung Chung, So-Young Park, Sang-Zoo Lee, and Hae-Chang Rim, GLR Parser with Conditional Action Model(CAM), Proc. of the 6th Natural Language Processing Pacific Rim Symposium, pp.359-366, 2001.

[21] Joshua Goodman, Parsing Algorithms and Metrics, In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, pp.177-183, 1996.

박 소 영

1997년 상명대학교 전자계산학과 학사. 1999년 고려대학교 컴퓨터학과 석사. 1999년 ~ 현재 고려대학교 컴퓨터학과 박사과정. 관심분야는 자연어처리, 기계번역, 한국어 정보처리

곽 용 재

1997년 고려대학교 컴퓨터학과 학사. 1999년 고려대학교 컴퓨터학과 석사. 1999년 ~ 현재 고려대학교 컴퓨터학과 박사과정. 관심분야는 자연어처리, 구문분석, 정보검색, 기계학습

정 후 중

1997년 고려대학교 컴퓨터학과 학사. 1999년 고려대학교 컴퓨터학과 석사. 1999년 ~ 현재 고려대학교 컴퓨터학과 박사과정. 관심분야는 자연어처리, 정보검색

황 영 숙

1991년 고려대학교 전산과학과 학사. 1991년 ~ 1995년 쌍용정보통신 근무. 1998년 고려대학교 컴퓨터학과 석사. 1998년 ~ 현재 고려대학교 컴퓨터학과 박사과정. 관심분야는 자연어처리, 기계학습, 정보추출

임 해 창

1991년 ~ 현재 고려대학교 컴퓨터학과 교수. 1993년 인지 과학회 이사. 1994년 -998년 한국 정보과학회 편집위원. 1998년 5월 ~ 2000년 5월 한국정보과학회 한국어정보처리연구회 운영위원장. 1999년 3월 ~ 2000년 8월 고려대학교 컴퓨터과학기술연구소 연구소장. 관심분야는 자연어처리, 구문분석, 정보검색, 기계학습