의존문법 기반의 구간 분할법을 활용한 한국어 구문 분석기

박용욱*

Korean Parser Using Segmentation Based on Dependency Grammar

Yong-Uk Park*

요 약

최근 대부분의 한국어 구문분석기는 의존문법(Dependency Grammar)을 사용하고 있는데, 그 이유는 한국어의 특성인 어순이 비교적 자유롭고 생략이 잦은 문장에 대한 처리가 용이하기 때문이다. 하지만 의존문법을 이용한 한국어 구문분석은 많은 중의성을 도출하는 문제점이 대두되고 있다. 본 논문에서는 이러한 중의성을 보다 효과적으로 해결하기 위하여 200개가 넘는 의존문법 규칙과 문장의 구성성분을 이용한 구간분할방법, 관형형어미가 붙은 용언에 대한 처리 및 같은 용언에 중복된 구성성분 결합제한 방법을 사용한 구문분석기를 제안한다. 실험 결과 중의성 제거에 많은 효과가 있음을 보여준다.

ABSTRACT

Recently, most Korean syntactic analysis systems use Dependency Grammar, because it is quite good to analysis of Korean language structures. But Dependency Grammar makes many ambiguities during syntax analysis of Korean. We implement a system which decreases many ambiguities in syntax analysis. To decrease ambiguities we suggest several methods. First, we use about 200 dependency rules, second, we suggest a new segmentation method and third, one predicate can not have more than one subject or object. Using these methods, we can reduce many ambiguities in Korean syntactic analysis.

키워드

Korean Parser, Segmentation method, Syntactic ambiguity, syntax analysis

* 울산과학대학 컴퓨터정보학부

접수일자 2009. 05. 25
심사완료일자 2009. 07. 28

Ⅰ. 서 론

한국어 구문 분석기는 목표는 한글로 된 단어들의 선형적 나열인 문장으로부터 그 문장에 내포되어 있는 문법적 구조를 찾아내는 것이라 할 수 있다. 이러한 문법적 구조를 바탕으로 정확한 뜻을 컴퓨터가 이해할 수 있도록 하는 것이다. 자연언어 문장은 단어들의 일차원적 나열로 표기하지만 실질적으로 어떤 의미 있는 문법적 구조를 내포하고 있다[4].

한국어 구문분석에서의 어려움은 비구성적 언어로서의 한국어의 특성을 반영해야 하는 것이다. 이런 특성들에 맞게 여러 가지 연구가 있었으며, 최근에는 의존문법에 기반 한 연구가 활발하게 진행되어 왔다[1,2,3]. 한국어의 특성인 중심어 후행, 어순의 자유로움, 빈번한 생략 등에 의존문법이 비교적 잘 맞기에 이에 대한 연구가 진행되었다. 이와 같은 한국어의 특성에 적합하다고 여겨지는 의존문법은 매우 견고하고(robust) 융통성 있는 모델이지만, 이런 장점을 보장하는 모델의 단순성이 문장 분석시에 중의성을 증가시키는 문제점을 가진다[3,5].

그 동안 의존문법을 이용한 한국어 분석기에 대한 많은 연구가 있었다. 초기에는 구문분석기 구현 그 자체에 대하여 연구했으나 현재는 정확한 분석결과를 얻기 위한 연구가 활발하게 이루어지고 있다. 현재 가장 문제가 되고 있는 것 중의 하나는 대부분의 구문분석 시스템들이 가능한 모든 구문분석 결과 트리를 생성함으로서 의미분석이나 기계번역과 같은 응용분야에 적용하기 위해서는 구문 중의성을 해결해야 한다는 것이다[2].

이러한 문제를 해결하기 위한 연구가 많이 진행되어 오고 있으며, 그 방법들 중의 하나로 긴 문장을 여러 개의 구간으로 분할하여 구간단위로 구문 분석하는 방법이 있다[3,5,6]. 구간분할 방법으로는 절(phrase) 단위[3]로 분할하는 방법과 용언의 문형정보[6]를 이용한 분할 방법이 연구되었다. 본 논문에서는 또 다른 구간분할 방법으로 문장의 구성성분 단위로 구간을 분할하는 방법을 제안하다. 자세한 내용은 본론에서 자세히 설명하겠다.

본 논문에서는 문장 분석 단위로 형태소를 사용한다. 한 어절이 거의 하나의 형태소로 이루어진 영어와 달리 한국어는 하나의 어절이 여러 개의 형태소로 이루어져 있으며, 이 때문에 지역 중의성이 많이 존재한다[2]. 즉 하나의 어절에 대한 형태소 분석후보가 여러 개가 나올 수 있다. 본 논문에서 제안하는 구문분석기는 지역 중의성을 인정하고, 그에 따른 모든 가능한 구문분석구조를 결과로 출력하므로 어절단위를 사용하는 분석시스템보다 많은 분석결과트리를 낼 수 있다. 이를 해결하기 위하여 문장 구성성분 단위의 구간분할 방법을 제안한다. 또한 중의성을 많이 발생시키는 관형형전성어미가 붙은 용언에 대한 처리를 병행한 결과를 제시한다. 또한 명사구 동사구와 같은 구(phrase)에도 적용 가능한 의존규칙을 만들어서 적용했다.

Ⅱ. 관련 연구

한국어에 대한 구문분석의 결과로 나타나는 파스트리의 정확성을 높이기 위하여 많은 연구가 진행되어 왔다[2,3,7]. 구문분석 과정에서 나타나는 중의성을 해결하기 위한 방법으로는 확률적 기법을 이용하는 방법과 언어적인 특징을 이용하는 방법이 있다. 은지현[9]은 확률적 기법을 사용한 사례이다. 확률적 기법을 사용하기 위해서는 정제된 말뭉치를 기반으로 구현되는데, 아직 국내에서는 구축된 말뭉치에 대한 검증이나 보완에 대한 연구가 미흡한 실정이다[10]. 따라서 언어적인 특징을 이용한 구문분석 기법이 많이 연구되어왔다. 그 중의 하나로 규칙 기반 연구가 있는데, 이현영[5], 김창제[11], 박의규[7] 등은 구 묶음 기법(chunking)을 사용하였다. 이현영[5]는 본용언과 보조용언의 묶음 규칙을 사용했고, 김창제[11]는 서술어절과 기능어절을 묶는 규칙을 사용했다. 박의규[7]은 의존명사를 구 묶음으로 처리하여 구문분석의 결과로 나타나는 파스트리에 대한 정확도를 향상시켰다. 전은희[12]는 동사의 논항정보를 이용하여 구문분석 시스템의 성능을 향상시켰다.

또 구간분할 방법을 통하여 긴 문장의 복잡도를 낮추어 구문분석 결과의 정확도를 향상시키는 방법이 있다. 구간분할 방법은 주어진 문장을 어떤 정해진 규칙을 사용하여 몇 개의 구간으로 분할하여 구간별로 구문분석을 하는 방법이다.

김광백[3]은 구간을 나누는 기준을 세 가지 사용했는데, 관형형어미를 갖지 않은 용언 바로 다음과 "~때","~인 이유로"등 이유, 시간 등을 나타내는 구 바로 다음 그리고 "~한 김에", "~할 시에" 등에 나타난 "김,시,양,지"

등과 같은 의존 명사 어절 바로 다음에서 구간을 분할하는 방법을 사용했다. 이 방법은 구간분할 방법이 다소 복잡하다. 이헌영[6]은 용언의 문형정보를 이용하여 내포문을 분할하여 구문분석을 실시하였다.

본 논문에서 사용한 구간분할 기준은 문장의 주요 구성성분으로 주어, 목적어, 서술어 등으로 단순화 했다. 또한 앞선 연구에서는 구간분할을 먼저 실시한 후, 각 구간에 대한 구간 분석 및 구간 통합 순으로 이루어지나, 본 논문에서 제안하는 방법은 전체 문장을 분석하면서 구간분할 기준이 되는 곳을 만나면 그 지점에서 구간 분할을 실시하고, 또 용언을 만나면 앞의 분할된 구간과 구간 통합을 실시함으로 구간분할과 구간통합이 동시에 일어난다. 또한 시스템의 성능 향상을 위하여 관형형전성어미가 붙은 용언은 의미상 주어가 뒤에 오는 경우가 많으므로 앞에 오는 주어성분에 대해 결합을 제한하는 방법을 사용했다.

Ⅲ. 구문분석 시스템 및 과정

본 논문의 구문분석 시스템은 입력문장에 대하여 형태소를 추출해내는 형태소 분석과정, 분석된 형태소 리스트에 대하여 몇몇 형태소들에 대하여 결합 등을 실시하는 전처리 과정, 구간분할 및 통합과정을 통한 구문을 분석하는 과정, 최종 완성된 구문트리에 대하여 정리하는 과정으로 시스템이 구성되어져 있다.

본 논문은 입력으로 주어진 문장에 대하여 의존문법을 적용하여 가능한 모든 문장구조를 추출하는 시스템[1]을 사용하고 중의성을 줄이기 위하여 구간분할법, 관형형 전성어미가 붙은 용언에 대한 처리법 및 정교한 의존문법을 만들어 사용한다.

본 시스템에서 사용하는 구문분석기는 기본단위로 형태소를 사용한다. 이것은 어절단위를 사용하는 것보다 중의성을 증가시킨다. 그러나 언어학적 관점에서는 어절단위의 분석보다는 형태소 단위의 분석이 보다 적절하다고 볼 수 있다[2]. 본 시스템의 형태소분석기는 주어진 입력 문장으로부터 가능한 모든 형태소를 추출하고, 이것들을 구문분석기에게 넘겨준다. 구문분석기의 입장에서는 서로 다른 여러 개의 조합으로 이루어진 형태소분석 리스트를 받게 된다. 그러므로 다른 구문 분석 시스템들 보다 많은 분석 결과를 출력할 가능성을 내포하고 있고, 또한 중의성도 많이 있을 수 있다. 이에 구간분할법, 관형형 전성어미가 붙은 용언처리 및 보다 정교한 의존문법들을 통하여 중의성을 제거하도록 연구하였다.

3.1 전처리 과정

전처리 단계에서 아래와 같은 몇 가지 일들이 이루어진다.

① 이웃하는 어떤 형태소들은 하나로 합친다

② 일부 형태소들은 형태소리스트에서 제거

①에서는 선어말어미 등과 같은 통사적 의존관계를 가지지 않는 형태소들은 미리 하나로 합친다. 또 "ㄹ 수 있다"와 같은 (관형형전성어미+수+있다)등은 용언관용구로서 하나로 합친다. 또한 "눈코 뜰 새 없이", "시도때도없이"와 같이 숙어처럼 사용되는 형태소들도 하나로 합친다.

②에서는 보조용언들은 독립적인 의미를 가지지 않으므로 본용언에 추가되어도 무방하므로 제거가능하고, 또 일부 의존명사가 분석후보에 홀로 나타나는 경우에 좌우 형태소를 보고 제거할 수 있는 분석후보를 제거한다.

본 구문분석기는 품사 태거를 사용하지 않으며, 모든 형태소 분석결과에 대한 조합으로 된 형태소 리스트에 대해 분석을 시도한다. 그러므로 모든 형태소 분석결과 조합을 사용하지 않는 다른 구문분석 시스템에 비하여 입력문장에서 보다 많은 중의성이 내포되어 있을 가능성 크다고 볼 수 있다. 그러나 분석 가능한 문장의 모든 구조를 보다 정확하게 분석해 낼 수 있다.

3.2 구간 분할 및 통합 과정

한국어 구문분석에 있어서 어려운 부분 중의 한 가지는 긴 문장에 대한 분석이다. 문장이 길어지면 중의성이 증가하게 됨으로 분석결과의 정확성이 떨어지게 된다. 이를 해결하는 방법 중의 하나는 구간분할 방법이다[3]. 본 논문에서 사용하는 구간 분할 방법은 구문분석을 진행하면서 이루어지며, 구간 분할의 기준은 문장의 구성성분이다. 문장 구성성분으로는 주어, 목적어, 부사어, 관형어, 술어 등으로 나누어 볼 수 있다. 우리말의 문장구조 분석의 결과는 술어를 중심으로 나머지 구성성분들이 어떻게 술어와 관계를 맺고 있는가를 보여

주는 것이다. 이에 바탕을 두어 본 시스템에서는 술어에 직접적으로 의존관계를 맺을 수 있는 주요 문장구성성분인 주어, 목적어, 부사어를 중심으로 구간분할을 하였다. 나머지 구성성분들은 이들 구성성분에 연결되도록 분석되어진다. 다음 문장을 문장의 주요 구성성분인 주어, 목적어, 부사어 따라 분할하면 표시(/)된 것과 같이 분할된다.

이 경기는/ 마을에서/ 가까운 넓은 모래밭에서/
벌어졌다

그리고 이러한 구간은 미리 분할하지 않는다. 파싱을 진행하면서 구간을 분할하고 동시에 구간내의 구문분석을 실시한다. 분리된 각 구간 내에서 분석과정을 통하여 가능한 모든 구문분석 구조를 찾아낸다. 그리고 최종적으로 완성된 구간 파서트리만 남기고 나머지 완성되지 않은 구간 분석트리 조각들은 모두 제거한다. 그리고 구간통합은 용언을 만날 때 실시한다. 파싱을 진행하면서 용언을 만나게 되면 이전의 구간분석이 마쳐진 구간들과 결합이 가능한지를 검사하여 결합이 가능하면 구간 통합을 한다. 그러므로 구간분할과 구간통합과정이 순차적으로 일어나는 것이 아니고, 두 과정이 섞여서 발생할 수 있다. 파싱이 진행되면서 구간이 분리되기도 통합되기도 한다. 따라서 분리된 구간의 개수는 계속 변화될 수 있다. 최종적으로 마지막 술어에 대한 분석이 진행되면서 이전의 모든 구간의 분석결과들과 하나로 결합이 이루어져 구문분석이 완료된다.

구간분할 및 통합의 대략적인 알고리즘은 다음과 같다.

1) 형태소분석 결과리스트의 좌에서 우 순으로 분석을 진행한다.
2) 입력 형태소에 대하여
   2-1) 주어, 목적어 또는 부사어 성분 형태소이면, 새로운 구간을 생성해서 다음에 입력되는 형태소들에 대해서 새로운 구간에서 분석을 수행한다. 또한 이전 구간에 대하여 완성된 분석트리만 남기고 나머지 조각 트리들은 두 제거 한다
   2-2) 주어, 목적어 또는 부사어 성분 형태소 아니면 계속해서 같은 구간내에서 분석을 계속 수행 한다
   2-3) 용언에 해당하면 현재의 구간에 대하여 분석수행 후, 앞의 분리된 구간과 결합을 시도한다.

전체적으로 이루어지는 파싱 알고리즘은 [1]의 방법을 사용한다. 이 방법은 구간 내에서 가능한 모든 분석구조를 찾아낸다. 위의 알고리즘을 사용하여 "철수가 사과를 먹는 영희를 보았다"라는 문장을 분석하면 다음과 같은 과정을 거치게 된다.

입력 문장:"철수가 사과를 먹는 영희를 보았다"
형태소 분석 리스트 : 철수(명사), 가(주격조사), 사과(명사), 를(목적격조사), 먹(동사), 는(관형형 전성어미), 영희(명사), 를(목적격조사), 보다(동사), 다(종결어미)

1) "철수" 입력 -> S1 : (철수)
2) "가" 입력 -> S1 : (철수), (가), (철수, 가)
   주격조사이므로 S2 생성함
   구간 S1 정리 : (철수, 가)만 남김
3) "사과" 입력 -> S2 : (사과)
4) "를" 입력 -> S2 : (사과),(를),(사과,를)
   목적격조사이므로 S3 생성함
   구간 S2 정리 : (사과, 를)만 남김
5) "먹" 입력 -> S3 : (먹)
   용언이므로 S3와 앞의 구간 S2, S1순으로
   결합 시도함
   결합(S2, S3) -> S2 : ((사과, 를) 먹)
   S3는 제거함
6) "는" 입력 -> S2 : (((사과, 를) 먹) 는)
7) "영희" 입력 ->
   S2 : ((((사과, 를) 먹) 는) 영희)
8) "를" 입력 ->
   S2 : (((((사과, 를) 먹) 는) 영희) 를)
   목적격조사이므로 S3 생성함
9) "보다" 입력 -> S3 : (보다)
   용언이므로 S3와 앞의 구간 S2, S1순으로
   결합 시도함
   결합(S2, S3) ->
     S2 : ((((((사과,를)먹)는)영희)를)보다)
     S3는 제거함
   결합(S1, S2) ->
     S1 : ((철수, 가)

((((((사과,를)먹)는)영희)를)보다))
S2는 제거함
10) "다" 입력 ->
S1 : (((철수, 가)
((((((사과,를)먹)는)영희)를)보다))다)
* (S1, S2, S3 : 분리된 구간 번호임)

최종적으로 더 이상 입력형태소가 없으면, 구간 S1에 최종 분석결과가 남게 된다. 아래의 [그림1]는 본 시스템에서 "철수가 사과를 먹는 영희를 보았다"에 대한 분석 결과를 보여 준다

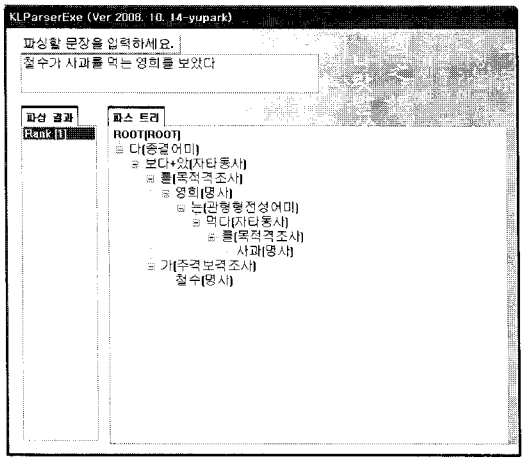
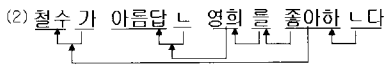


그림 1. 위 예의 분석결과 화면
Fig. 1 Analysis result of above example

3.3 관형형어미가 붙은 용언 처리

관형형어미가 붙은 용언은 일반적으로 뒤에 따라오는 명사를 의미상 주어를 취하는 경우가 많다. 그러므로 앞에 있는 주어자격을 가진 구성요소를 주어로 결합시킬 때는 주의해야 한다. 예를 들어 "철수가 아름다운 영희를 좋아한다."의 문장에서 "아름다운"은 아름답(용언)+ㄴ(관형형전성어미)로 형태소 분리된다. 또 "철수가"는 철수(명사)+가(주격조사)로 분리된다. 이때 본 논문에서 사용하는 결합규칙에 의하면 "철수+가"는 용언 "아름답"과 결합 가능하고, 또한 "아름답"은 "영희"와도 결합 가능하다. 따라서 아래의 (1),(2)와 같은 두 가지 구조가 모두 가능하다. 그러나 (1)의 분석구조는 의미적으로 잘못 분석된 것이다.

(철수가 아름답다)



(영희가 아름답다)

(1)은 "아름답"의 의미상 주어를 "철수"로 가지는 동시에 "영희"를 또한 의미적으로 주어로 취하는 이상한 구조를 가지게 됨으로 틀리게 분석된 구조이다. (2)는 "아름답"의 주체가 "영희"로 분석된 것으로서 올바르게 분석된 구조이다.

이러한 문제는 관형형 어미가 붙은 용언에서 발생하는데 그 처리가 간단하지 않다. 관형형 전성어미가 붙은 용언의 형태는 다음과 같은 3가지의 경우가 있을 수 있다[5].

1) 관계 관형형-수식받는 명사가 관형사가 된 용언의 의미상 주어 역할을 하는 경우
2) 동격 관형형-수식받는 명사가 관형사가 된 용언과 동격인 경우
3) 의존 관형형-관형형 전성어미 뒤에 '것/줄/수/데' 등이 붙은 경우

본 시스템에서는 위 1)의 경우에 대하여, 관형형어미가 붙은 용언 뒤에 꾸밈을 받은 명사가 의미적으로 주어 역할을 할 수 있는 구성성분이면 앞에 오는 주격 형태소를 결합하지 못하도록 제한함으로서 처리할 수 있었다. 앞의 예 "철수가 아름다운 영희를 좋아한다"에서 (1)의 틀리게 분석된 구조를 이 방법을 통하여 제거할 수 있었고, [그림 2]은 그 결과를 보여준다. 이러한 관형형어미가 붙은 용언에 대한 처리법을 통하여 구문분석에서 발생하는 많은 중의성을 제거할 수 있었다. 위 2),3) 경우에 대해서는 앞으로 계속 연구하여 처리할 예정이다.

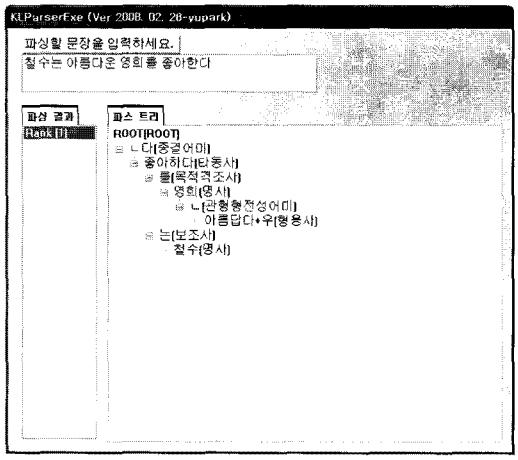


그림 2. 분석결과 화면
Fig. 2 Analysis result

3.4 같은 용언에 중복 구성성분 결합제한

단지 의존규칙에 의해서만 두 형태소간의 결합여부를 결정하면 하나의 용언에 2개 이상의 주어 또는 목적어가 결합될 수 있다. 예를 들어 "영희가 선생님께서 지나가시자 인사를 드렸다" 문장에서 "영희가"와 "선생님께서"는 "지나가다"의 주어성분으로 모두 결합 가능하다. 본 시스템의 파싱 알고리즘은 "지나가다"와 가까이 있는 "선생님께서"가 먼저 주어성분으로 결합된다. 그 다음에 "영희가"가 "지나가다"와 결합시도 할 때 의존규칙으로는 결합 가능하지만 이미 주어성분이 결합된 상태이므로 결합하지 못하게 처리한다. 따라서 "영희가"는 나중에 "드리다"와 결합하게 된다. 이를 통하여도 많은 중의성이 해결되는 결과를 가져왔다.

3.5 의존 문법 규칙

의존문법 규칙이란 입력되는 두 형태소에 대하여 어떤 관계가 가능한가를 의존문법을 사용하여 규정한 규칙을 말한다[1]. 의존문법은 두 형태소 사이의 관계에 관심을 가지는 문법으로, 두 형태소가 무슨 관계로 결합되는가를 나타낼 수 있다. 의존문법을 사용한 이러한 규칙은 입력되는 두 형태소 품사에 대하여 이진연산(binary operation)으로 이루어져 있으며, 그 결과가 다시 품사로 출력되는 닫힌 연산(closed operation)이다. [표 1]은 본 논문에서 사용하는 의존문법 규칙의 예를 보여준다.

표 1. 의존문법 규칙의 예
Table. 1 Dependcy Rules Table

지배소	의존소	결과	관계
명사	관형사	명사구	수식
명사	명사	명사구	수식
동사	동사수식부사	동사구	수식
동사	부사구	동사구	수식
형용사	형용사수식부사	형용사구	수식
형용사	부사구	형용사구	수식
관형형전성어미	형용사	관형사구	품사전성
동사	격조사구	동사구	논항
형용사	격조사구	형용사구	논항
주격보격조사	명사	격조사구	격부여
관형격조사	명사/명사구	관형사사구	격부여
종결어미	동사/동사구	종결구	종결
종결어미	형용사/형용사구	종결구	종결
온점	종결어미/종결구	문장	완성

[표 1]에서 볼 수 있는 것처럼 본 논문에서는 의존문법 규칙에 구(phrase)를 사용한다. 명사구, 동사구와 같은 구를 문법규칙에 사용하게 되면 품사 분류의 수와 규칙의 수는 늘어난다. 그러나 원래 하나의 형태소일 때와 그것이 다른 형태소와 결합한 후 성격 변화로 인하여 규칙 적용이 보다 확장될 수 있는데 이를 처리하기가 용이하다. 예를 들어, '이다'와 같은 지정사는, 단순히 지정사일때는 바로 앞의 명사 등만 지배할 수 있지만, 그것과 결합하여 지정사구가 되면 논항을 지배할 수 있게 된다. 본 구문분석기는 구를 포함하여 108개의 품사 분류와 이를 바탕으로 한 212개의 의존문법 규칙이 적용되었으며, 이러한 품사 분류와 규칙들을 계속해서 개선해 나가고 있다.

Ⅳ. 실험 및 결과

본 논문에서 사용한 구문분석의 기본단위는 어절이 아니라 형태소이다. 세종계획에서 개발한 한국어 구문트리 부착 말뭉치가 있지만 이것은 어절을 기본 단위로 사용한 것이므로 형태소를 기본단위로 분석하는 본 시스템에서 직접 활용하기는 곤란하다. 따라서 본 실험에

서는 중학교 교과서에서 추출한 임의의 문장을 통하여 실험하였다. 추출된 100에 대하여 전처리 과정을 거친 후 표1의 의존문법을 바탕으로 다음의 조건으로 실험하였다.

- 구성성분 기준의 구간 분할법 사용
- 관형형 어미가 붙은 용언에 대한 처리
- 용언에 중복된 구성성분 결합 제한

위의 3가지 방법에 대하여 사용할 때와 사용하지 않았을 때 생성되는 파스트리 결과에 대하여 아래와 같은 방식에 의해 정확률과 재현율을 평가하였다.

$$\text{정확률} = \frac{\text{결과트리 내의 올바른 의존관계 수}}{\text{결과트리 내의 의존관계 수}} \times 100$$

$$\text{재현율} = \frac{\text{결과트리 내의 올바른 의존관계 수}}{\text{정답트리 내의 의존관계 수}} \times 100$$

표 2. 정확률과 재현율 비교
Table. 2 Comparison of precision and recall

	3가지 방식 사용	사용하지 않음
정확률	97.6%	74.3%
재현율	85.4%	98.5%

본 논문에서는 품사 태거를 거치지 않고 모든 형태소 분석 조합에 대해 구문분석을 시도하기 때문에, 어절 수가 늘어날수록, 특히 용언의 숫자가 늘어날수록 중의성이 급격하게 커진다. 따라서 이러한 중의성을 해결하기 위하여 위의 3가지 방법을 사용하였는데 [표 2]와 같이 정확률이 크게 향상되는 효과를 얻었다. 반면 재현율은 다소 떨어지는 결과를 볼 수 있는데, 이는 중의성을 해결하기 위하여 각 분할된 구간에서 완성된 부분트리만 남기고 나머지는 삭제하여 남은 완성된 부분트리에 대해서만 결합한 것에서 발생하는 원인과 관형형용언에 대한 처리가 아직 온전하지 못함으로 인한 것으로 볼 수 있다. 차후 이에 대한 연구를 계속 진행할 것이다. 다음은 그림[3,4]는 예문 "이 경기는 마을에서 가까운 넓은 모래밭에서 벌어졌다"에 대하여 3가지 방법을 적용하지 않았을 때의 처리결과 화면[그림3]과 적용하여 수행한 결과 화면[그림4]을 보여준다.

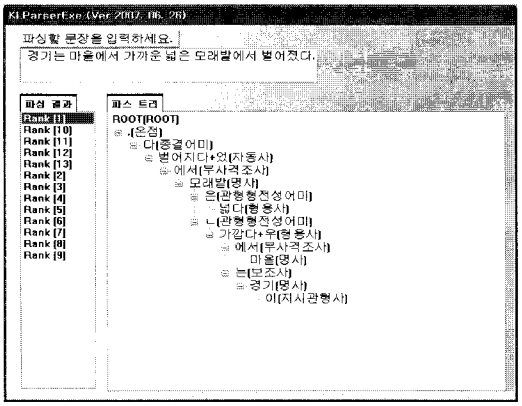


그림 3. 3가지방식 적용하지 않은 처리결과
Fig. 3 Result of not applying three methods

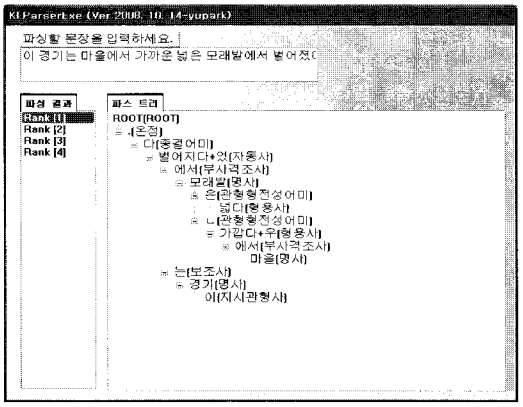


그림 4. 3가지방식 적용한 처리결과
Fig. 4 Result of applying three methods

위의 [그림3, 4]를 비교해 보면 알 수 있듯이 3가지 방식을 적용하여 구문분석을 한 경우 많은 중의성이 제거된 것을 볼 수 있다.

V. 결 론

본 논문에서는 규칙 기반 의존문법을 이용한 한국어 구문분석기에 대해 다루었다. 한 문장의 모든 형태소 분석 조합에 대해 구문분석을 시도했으며, 여러 언어적 특성에 기반을 두어 수작업으로 작성된 의존 규칙들을 적용하였다. 본 시스템은 품사 태거를 거치지 않고 모든 형태소 분석 조합에 대해 구문분석을 시도하기 때문에, 어절 수가 늘어날수록, 특히 용언의 숫자가 늘어날수록 중

의성이 급격하게 커진다. 이를 해결하기 위하여 문장 구성성분을 이용한 구간분할방법, 관형형 어미가 붙은 용언에 대한처리, 용언에 중복된 구성성분 결합 제한 방법을 사용하였다. 실험결과 중의성 해결에 많은 효과가 있음을 보여준다.

차후 연구에서는 현재 사용 중인 구성성분단위의 구간분할에 대하여 보다 많은 데이터를 가지고 실험하여 구간분할 방법을 개선하고, 관형형 어미가 붙은 용언의 여러 유형에 대한 연구와 이를 처리하는 방법에 대하여 연구할 것이다.

참고문헌

[1] 권혁철, 최준영, "단일화 기반 의존 문법을 이용한 한국어 분석기", 한국정보학회 논문지 '92.9 Vol.19, No.5, September

[2] 임경업, 정영임, 권혁철, "한국어 어휘의휘어미망에 기반한 논항 정보를 이용한 의존문법 구문분석기의 구현", 제19회 한글 및 한국어 정보처리 학술대회 pp.158-163, 2007

[3] 김광배, 박의규, 나동렬, 윤준태, "구간 분할 기반 한국어 구문분석", 제14회 한글 및 한국어 정보처리 학술대회 pp.163-168, 2002

[4] "자연언어처리", 김영택 외 공저, 생능출판사

[5] 이현영, 황이규, 이용석, "문형과 단문 분할을 이용한 한국어 구문 모호성 해결", 제12회 한글 및 한국어정보처리 학술대회 pp.116-123, 2000

[6] 이현영, 이용석, "문형을 제약조건으로 하는 단문 분할 기반 한국어 구문분석", 제18회 한글 및 한국어정보처리 학술대회 pp.140-147, 2006

[7] 박의규, 나동열, "한국어 구문분석을 위한 구묶음 기반 의존 명사 처리", 인지과학 제17권 제 2호, pp119-138, 2005

[8] 김광진, 송형훈, 이정현, "한국어 내포문을 단문으로 분리하는 시스템의 구현", 제5회 한글 및 한국어정보처리 학술대회, pp.333-352, 1993

[9] 은지현, 정민우, 이근배, "확률적 차트 파싱에 기반한 한국어 의존 구조 분석기", 제17회 한글 및 한국어정보처리 학술대회 논문집, pp.105-111, 2005

[10] 이미경, 정한민, 성원경, 박동인, "품사 표지 부착 말뭉치 검증", 제17회 한글 및 한국어 정보처리 학술대회 논문집, pp145-150, 2005

[11] 김창제, 정천영, 김영훈, 서영훈, "부분적인 어절결합을 이용한 효율적인 한국어 구문 분석기", 한국정보과학회 가을 학술발표논문집, vol. 22, No.2, 1995

[12] 전은희, 이성욱, 서정연, "한국어 동사의 격틀 정보를 이요한 구문분석 후처리기", 제13회 한글 및 한국어정보처리 학술대회 논문집, pp.445-449, 2001

[13] 윤준태, "공기 관계 기반 어휘 연관도를 이용한 한국어 구문 분석", 연세대학교 박사학위논문, 1997

[14] 조형준, "한국어 병렬구문과 결합범주문법에서의 구문분석" 한국과학기술원 석사학위논문, 1999

[15] I.A. Mel'cuk, Dependency Syntax : Theory and Practice, State Univ. of New York Press, 1988

저자소개

박용욱(Yong-Uk Park)

1991년 부산대학교 전자계산학과(이학석사)
1991년 3월 ~ 1997.2월 전자부품 연구원(KETI) 전임연구원
2000년 부산대학교 전자계산학과(박사수료)
1998년 3월~현재 울산과학대학 컴퓨터정보학부 교수
※관심분야 : 자연언어처리, 정보검색, 멀티미디어 문서처리