언어정보 DB 구축을 위한 문법적 주석 상의 몇 문제*

- 기존 국어사전의 어휘 정보 수용과 관련된 문제를 중심으로

신 선 경°, 한 영 균

(울산대 국문과)

Problems in Syntactic Annotation for Building a LDB in Korean

Shin, Sun-kyung, Han, Young-Gyun

(Dept. of Korean Language and Literature, University of Ulsan)

요 약

한 언어에 대한 포괄적인 언어정보 데이타베이스의 구축에 있어서는 수집된 텍스트에 대한 상세한 문법정보의 주석이 일차적 작업 대상이 된다. 이는 통사적 정보가 단순히 구문 분석상의 문제들을 해결하기 위한 정보를 제공해주는 것일뿐 아니라 형태소 해석 및 문장 의미의 파악등 자연언어 이해시스템 전반의 성능을 향상시키는 데에 중요한 몫을 차지하기 때문이다.

각개 단어의 문법적 기능에 대한 주석은 사전적 정의에 따른다면 '품사'로 표현할 수 있을 것이다. 그런데 품사는 각개 단어가 지니는 고유한 어휘의미적 정보이기보다는 구문구조에 의존적인 양상을 보인다. 이는 사전에 따라서 각개 단어에 대한 품사 정보가 달리 나타나는 점에서도 간취할 수 있는데, 한편으로 한국어 언어정보 데이타베이스 구축을 위한 문법적 주석에 있어서는 기존 사전의 품사정보에만 의존할 수는 없다는 문제점이 제기된다. 따라서 각 어휘들의 구문정보(혹은 품사정보)를 어떻게 기술할 것인가가 해결되어야 하는 것이다.

본 연구에서는 일차적으로 각 어휘들의 문장 안에서의 기능을 바탕으로 한 주석 체계를 설정하고 그에 따라서 약 12만개의 문장에 대한 일차적 형식화를 수작업으로 처리하였다. 이는 향후 자동적으로 문법적 주석이 가능하도록 해주는 시스템의 개발을 지원하기 위한 언어정보의 수집에 목적을 둔 것인데, 이를 통해서 기존 국어사전에서의 언어정보상의 미비점을 수정·보완할 몇 가지 근거를 마련할 수 있었다.

* 본 연구는 한국통신의 장기기초연구과제의 하나인 '정보검색용 전자사전의 제작을 위한 연구'의 일부임.

Ⅰ. 서 론

1.0. 본 연구는 최종적으로 자연언어처리의 여러 영역에서 범용성을 지니는 언어정보 데이타베이스를 구축하고, 자연언어처리시스템에서 활용할 수 있는 전자사전을 구현하는 데에 필요한 언어학적 기반을 제공하는 데에 목표를 둔다. 이는 최근의 자연언어처리가 단일화문법(Unified Grammar)에 입각한 사전 정보의 극대화를 요구하고 있는 데에 따른 것이다.

1.1. 자연언어시스템이 요구하는 사전적 정보의 획득은 크게 두 가지 방법에 의존하고 있다. 그 하나는 이미 만들어져 있는 일반언어사전의 문법정보를 이용하는 방법이고[3, 4, 5, 6], 다른 하나는 텍스트자료의 분석을 통해서 추출해내는 것이다[2, 3]. 그러나 한국어 처리 시스템을 위한 사전의 개발에 있어서는 전자의 방법에 의존하기는 어렵다. 무엇보다도 기계가독형사전(Machine Readable Dictionary)이 존재하지 않는다는 점을 들 수 있다. 일반언어사전을 이용하는 가장 큰 장점이 단시일 안에 적은 비용으로 많은 정보를 추출해 낼 수 있다는 것이다. 그러나 이는 한국어 MRD가 존재하지 않는 상황에서는 기대하기 어려운 일인 것이다. 둘째로는 기존 국어사전의 문법정보가 자연언어처리시스템이 요구하는 언어정보의 양을 충족할만큼 충분치 않다는 점을 들 수 있다. 특히 구문정보에 관한한 기존의 국어사전은 자연언어처리 시스템에서 요구되는 정보를 전혀 충족시키지 못하는 것이다[1]. 이러한 점을 감안하여 본 연구에서는, 가능한한 기존 국어사전의 정보를 충분히 활용하되, 다른 한편으로 텍스트 자료를 직접 분석·정리함으로써 필요한 언어정보를 추출하고 추출된 언어정보를 형식화하여 언어정보 데이타베이스를 구축함으로써 MTD(Machine Tractable Dictionary) 구현의 자료를 축적하기로 하였다.

1.2. 언어정보 데이타베이스의 구축을 위한 텍스트의 분석은 몇 가지 단계를 거치게 되는데, 그 중의 하나가 문법적 주석(Syntactic Annotation)이라고 할 수 있다. 이는 각 어휘들이 갖는 구문론적 정보를 제공하는 데에 그 목표를 둔다. 이와 같은 어휘의 구문정보는 각 어휘 개개의 어휘정보를 더 충실히 하는 것은 물론 서술어를 핵(head)으로 하는 문장의 분석 및 생성에 유용한 정보를 제공한다.

1.3. 본 연구에서 기본단위가 되는 것은 문장이다. 따라서 각 어휘들은 문장의 핵인 서술어를 중심으로 그 것의 하위범주로서 자신의 구문정보를 부여받게 된다. 한

어휘의 구문정보를 구체화하여 주는 규칙들로는 동사의 하위범주화 규칙(subcategorization rule), 그리고 그 하위범주관계를을 실제 문장에서 구체적인 문법소(격조사나 어미 등)와의 결합 등으로 실현시켜 주는 격부여규칙(case assingment rule) 등이 있다.

II. Notation System

2.0. 텍스트에 대한 문법적 주석 작업은 크게 두 단계로 나뉜다. 그 첫번째 단계는 각 어휘들을 문법적 기호로 전환하여 표면형에 충실하게 선조적으로 표기하여 주는 것이다. 이 작업을 통해 구축된 자료는 문장 내의 각 어휘들에 관한 형태, 통사론적 정보를 제공하게 된다. 그러나 이상의 작업은 문장의 구문정보나 기타 의미해석에 필요한 문장의 계층구조까지 밝히기에는 불충분하여 각 성분 간의 계층적 관계를 보충하여 밝혀 주는 이차 주석작업이 불가피하다. 이차 주석 단계를 통해서 일차단계에서 기호화된 각 성분들을 그들 사이의 계층 관계를 중심으로 묶어 주는 일과 그 묶인 뭉치들에 고유의 꼬리표를 달아 주는 일이 행해진다.

2.1. 일차적 주석 작업의 기본이 되는 기호체계는 각 성분들 간의 관계 즉 서술어와 그것의 보어, 첨어관계를 좀더 간명하게 보기 위하여 현행 품사 분류체계를 단순화하여 VP(동사구-- 동사와 형용사 등 문장의 서술어가 될 수 있는 어휘들), NP(명사구), ADP(부사구), IT(감탄사 등과 같이 문장 내에서 독립어로 쓰이는 어휘들), S(보문이나 관계절 등 문장 안에 안기는 절 이상의 단위) 등을 기본단위로 설정하였고 각 성분 사이의 문법관계를 구체화하는 문법형태소들(조사, 혹은 동사의 어미 등)은 그 실현형에 충실하게 밝혀 적어 주었다.

기호체계에 있어서 어휘소(NP,VP,ADP 등)과 형태소(어미, 조사 등)를 구별하는 이원적 체계를 택한 것은 NP나 VP, ADP등의 어휘소는 서술어(predicate)와 논항(argument, complement) 등의 서술구조의 기본성분으로, 형태소는 이 성분들 간의 논리적 관계를 규정해 주는 운용소(operator)로 인식하고자 하는 서술논리학적 관점과 영향과 실제 문장에서 각 성분 간의 관계가 표면문장에서 어떤 형태로 실현되느냐는 것을 구체적으로 기술하고 각 어휘의 형태, 통사적 정보를 제공한다는 본 작업의 목적에 충실하고자하는 실질적 이유에 바탕을 두고 있다.

2.2. 본 주석작업은 문장을 기본 단위로 하여 이루어졌다. 원칙적으로 단일문(하나의 서술어를 갖는 문장)을 기본단위로 하고 있으나 중문이나 내포문 등의 복합문일 경우 이를 구태여 단문으로 재분석하지 않고 주절의 서술어만을 고려하여 단일문과 같이 처리하였다. 이때 명사구를 수식하는 관계절이나 보문절은 S로 묶어 명사구에 얹혀 있는 명사구 내의 한 수식 성분으로 처리하였고, 소위 접속구문으로 분류되는 중문의 경우는 모두 풀어 선조적(linear)으로 표시하였다.

예를 들면, 다음의 예 (1)과 같다.

(1) a. 마음씨가 착한 영희는 그 거지에게 밥을 주었다.

--> NP(s-n)는 DNP에게 NP를 VP었다.

b. 동생은 춤을 추고 나는 노래를 불렀다.

--> NP은 NP을 VP고 NP는 NP를 VP었다.

여기에서 S로 표시된 성분들은 다시 검색, 축출되어 선조적으로 재분석된다. 문장의 일차 주석 단계에서는 서술어와 그의 자매항 사이의 의미 해석에 영향을 미치지 않는 한, 각 성분들은 선조적으로 분석하여 표기하는 것을 원칙으로 하였다. 따라서 명사구 간의 계층적 수식관계나 동사구 사이의 계층적 관계는 고려하지 않는다.

2.3. 이상의 처리에서 우리의 입장은 매우 공시적(synchronic)이다. 즉, 각 어휘들을 처리함에 있어 그것의 형태론적 내력을 밝히는 일, 각 어휘 간의 파생이나 굴절 등의 형태론적 규칙의 적용 등에 관심을 두기 보다는 그 어휘가 문장 안에서 어떠한 형태로 나타나며, 어떤 품사로 실현되어 어떠한 기능을 수행하고, 다른 성분들과는 어떠한 관계를 맺는가 하는 것에 더 많은 관심을 둔다. 따라서 한 어휘의 품사를 결정하는데 있어서도 그 어휘의 의미나 형태론적 내력보다는 문장 내에서 주어지는 통사적 정보가 더 큰 영향을 미치는 것이다

III. 기존 사전의 어휘정보 수용 상의 몇 가지 문제

3.1. 본 연구가 동사를 중심으로 하는 문장의 통사구조뿐만 아니라 그 속에서 하나의 성분으로 기능하는 각 어휘의 형태, 통사적 정보를 제공하는 데 목적을 두고 있

는 것이므로, 각 어휘의 문법적 기능을 어떻게 정의하여 기호화하느냐는 것이 일차적으로 제기되는 문제라고 할 수 있다. 그런데 기존의 사전에 담겨있는 문법정보는 품사 설정에 있어서조차 그 기준이 확실하지 않아 자연언어 처리 시스템을 지원하기 위한 언어정보 데이타 베이스 구축을 목적으로 하는 본 연구의 바탕이 되기에는 적지 않은 문제를 안고 있다. 본 작업에서 각 어휘를 기존 사전의 언어정보에 따라 처리하려고 했을 때 생긴 문제점들로는 다음과 같은 것들을 들 수 있다.

1) 소위 품사통용어로 분류되는 어휘들의 문법적 기능 정의.

예) 그들은 각자 자기 길로 갔다. : 부사
각자의 일은 각자가 책임을 져야 한다.: 명사
김씨, 이씨, 박씨 들이 모여 산다. : 의존명사
이 사람들은 누구입니까? : 접사
사람들이 모두들 잘들 하는구먼? : 특수조사

2) 통사적 파생 어구의 처리 방안

예) 틈틈이 : 부사 일하는 틈틈이 : VPㄴ ADP
가듯이 : 부사 구름에 달 가듯이 : NP에 NP# ADP
셈치다 : 동사 먹은 셈치고 가네 : VPㄴ VP고 VP네

3) 사전에 등재되지 않은 명사구의 처리.

예) terminology의 경우 : 손짚고 옆돌기를 하여 봅시다.
동사 +'음 /기'를 갖는 전성명사의 경우: 평강공주의 보살핌과 일깨움으로

4) 동사구의 재구조화로 서술성을 잃은 동사들의 처리.

- 수 있다 : 태양의 고도는 간이 고도 측정기로도 측정할 수 있다.

- 것 이다 : 독도는 푸른 섬으로 변하게 될 것이다.
- 것 같다 : 오늘은 반가운 손님이 오실 것 같습니다.
- 말이다 : 너 말이야 그러면 못 쓴단 말이야
  난 네가 좋단 말이야.

5) 품사 설정에 반영된 띄어쓰기 규칙의 불규칙성 문제

예) 가을쯤에 너희들과 연극을 한 번 꾸며 보려 하고 있지.
아냐, 아냐. 그냥 한번 해 본 말이야.

3.2. 1)의 품사통용어 문제는 이미 국어사전에 표제어로는 등재되어 있으나, 실제 용법상의 기능을 정확히 기술해주지 못한 부분이 있다는 점이 문제로 지적될 수 있고, 2)는 파생규칙을 거쳐 품사전성이 일어났으나 전성 전의 품사의 성격을 그대로 가져 문제가 되었다. 3), 4)의 문제는 아예 표제어로도 등재되지 않은 어구들이지만, 실제로는 하나의 표제어항으로 등재되어야 할 것이다. 2), 3), 4)의 문제는 기존의 국어사전들이 주로 어휘형태소를 중심으로 표제어항을 선정하고 있는 데에서 발생된 문제라고 할 수 있지만, 실제 이러한 어구들이 간여하고 있는 경우에 기존 사전의 정보에만 의존한다면 문장구조의 해석상의 오류가 야기되는 것이다.

이상을 통해 볼 때, 기존 국어사전의 언어정보를 보완할 수 있는 방안이 강구되어야 할 것이다. 여기서는 간단히 품사통용어로 분류되는 어휘들의 처리와 관련된 문제를 중심으로 살펴보기로 한다.

품사통용어란 '한 단어가 둘 이상의 품사적 기능을 동시에 가지고 있는 것'을 이르는 말로 여기에는 다음과 같은 예를 들 수 있다.

명사 ⇔ 부사 : 각각, 가까이, 거의, 오늘, 지금, 모두, 한번, 다섯, 합리적, ...
대명사 ⇔ 관형사 : 이, 그, 저, 아무 ...
의존명사 ⇔ 특수조사 : 뿐, 대로, 만큼 .....
의존명사 ⇔ 접사 ⇔ 특수조사 : 들 ....

주석화 작업에서 각 어휘를 그 어휘의 품사정보에 따라 기호화해 줄 때, 그 어휘가 어떠한 품사로 정의될 것인가 하는 것을 결정해 주는 가장 중요한 기준은 그 어휘가 문장 내에서 갖는 구문론적 기능이다. 대부분의 품사통용어들은 그들이 어떠한 품사로 정의되느냐와는 무관하게 자신의 고유의 어휘적 의미는 그대로 유지하고 있는 경우가 많으므로 결국 그것의 품사를 결정하여 주는 일은 그 어휘가 어절 또는 문장에서 다른 어휘들과 갖는 대강의 결합관계(syntagmatic relation) 및 대치관계(paradigmatic relation)에 철저히 의존하게 되는 것이다. 그러나 기존의 사전들에서의 품사통용어에 대한 처리는 그 기준이 확실하지 않고 대부분의 경우 통사적 기준과 화용론적(pragmatic) 의미 차이 등을 혼동하여 정의하고 있어 상당한 규칙성이 요구되는 자연언어처리시스템에 이용하기가 어렵게 되어 있다. 또한, 사전의 종류에 따라 각 어휘에 대한 품사적 정보를 다르게 정의하는 경우가 많고, 정의가 일치한다 하더라도 그 정의를 따를 경우 전체 체계 속에서 설명될 수 없는 구조를 생성하게 되어 통사원칙에 위배되는 경우가 많다. 따라서, 구조를 통해 나타나는 각 어휘의 구문적 정보에 따라 각 어휘의 품사 정보를 재조정하는 것이 불가피하다.

예를 들어, '멀리'와 같은 어휘는 사전에는 예 (4)와 같이 부사로만 등재되어 있다. 그러나 예 (5 a,b)에서 보는 것같이 '멀리'는 조사와의 연결이 자연스럽고 심지어 (4)c에서와 같이 관형사의 수식을 받는 등 명사와 같이 쓰여 이를 부사로 처리할 경우, 부사가 관형사의 수식을 받는 문법적 일탈 형태가 생기므로 이를 부사와 명사의 자격을 동시에 갖는 통용어로 처리함이 옳은 것이다.

(4) a. 멀리 : (부) [멀+리(<-이)] 멀게. P 멀리 던지다/멀리서 들리는 포소리
(국어대사전(1991), 금성출판사)

b. : (부) 멀게. 예문없음 (새국어사전(1988), 교학사)

c. : (부) 시간적으로나 공간적으로 사이가 아주 떨어지게.
P 멀리서 들리는 새소리 (새우리말큰사전(1989), 삼성출판사)

d. : (부) 시간적으로나 공간적으로 사이가 아주 떨어지게.
P 산너머 저 멀리 (국어대사전,이희승 편(1982), 민중서림)

(5) a. 멀리에서 북소리가 들려 왔습니다.

b. 저 기차는 멀리로 가는 기차란다.

c. 저 멀리 산넘어 행복이 있다 해서

d. 저 멀리가 네 고향이라오.

이때, 부사와 명사의 변별 기준이 되는 것은 첫째, 격조사 동반 여부와 둘째, 관형어에 의한 수식 여부 등으로 부사는 원칙적으로 격조사를 동반할 수 없고 관형절의 수식을 받을 수 없다.

마찬가지로 '오늘' 같은 단어도 사전의 정의는 달라질 수 밖에 없다.

(5) a. 오늘이 개학날입니다.

b. 오늘의 일정는 어떻게 되나요?

(6) a. 할머니께서는 오늘# 시골로 가십니다.

b.* 할머니께서는 오늘에 시골로 가십니다.

위의 예 (5)에서 보듯이 ' 오늘'은 격조사와 자유롭게 결합하는 등 명사임을 의심할 수 없으며, 실재로 모든 사전에서 명사로만 정의하고 있다. 그러나 예(6)을 보면 (6a)의 '오늘'은 조사와의 결합을 허용하지 않으며 (6b)와 같이 조사를 복원시켜 놓았을 때 비문이 된다. 이를 통해 (6a)는 조사가 생략된 명사구가 아니라는 것을 알 수 있고 위에서 세웠던 우리의 변별 기준에 의하여 예(6)에서 '오늘'은 부사로 쓰였다는 것을 알 수 있다.이와 같은 '오늘'의 부사성은 다음의 예 (7)의 대비를 통해 더욱 확실해 진다.

(7) a. 나는 오늘# 이 곳을 떠난다.

*나는 오늘에 이곳을 떠난다.

b. 나는 일찍# 이곳을 떠난다.

*나는 일찍에 이곳을 떠난다.

c.*나는 아침# 이곳을 떠난다.

나는 아침에 이곳을 떠난다.

(7)의 예를 통해 보면 '오늘'은 명사인 '아침'과는 다르고, 부사인 '일찍'과는 같은 계열관계(paradigmatic relation)를 보이며 부사로서 기능하고 있음을 알 수 있다. 따라서 '오늘'은 명사와 부사의 두 기능을 담당하는 것으로 파악할 수 있다.

이상에서 우리는 품사통용어처리의 두 가지 예를 통해 기존사전의 문제점을 엿볼 수 있었고, 한국어 처리 시스템을 위하여서는 기존사전의 문제점을 수정·보완할 필요가 있음을 알 수 있었다. 본 연구가 수행하고 있는 문법적 주석작업이 각 표제어들의 구문정보를 제공하는 것 뿐만 아니라, 표제어 항목의 결정에도 중요한 역할을 할 수 있다는 가능성을 확인할 수 있었다. 이는 텍스트에 대한 문법적 주석과정이 단순히 통사적 정보의 파악에서 그치는 것이 아니라, 표제어항의 추가 및 어휘정보의 보완에 적극적으로 기여할 수 있음을 의미하는 것이다.

참 고 문 헌

[1] 남기심·이상섭·김슬옹·이기황, "기존 국어사전의 언어정보적 평가", 우리말 정보화잔치 '91논문집 별쇄, 1991.

[2] 이상섭·최윤철·정영미·나동열, 「전자사전의 구현을 위한 어휘 데이타베이스의 설계와 국어정보 획득 도구의 개발」, '91 통신학술연구과제보고서, 1992.

[3] Atkins, S., N. Calzorari and E. Picchi, *Computational Lexicography*, material for Pre-Euralex Tutorial, 1992.

[4] Carroll, J. and C. Grover, "The Derivation of a Large computational lexicon for English fron LDOCE," in Boguraev and Briscoe(eds) 1989.

[5] Martin, Willy, "On the Parsing of Definitions," in Euralex '92 Proceedings pp. 247-256.

[6] Neff, Mary S. and Branimir K Boguraev, *From Machine-Readable Dictionary to Lexical Data Base*, IBM Research Division, T. J. Watson Research Center, 1990.