한국어 문법관계에 대한 부분구문 분석

(Shallow Parsing on Grammatical Relations in Korean Sentences)

이 성 욱 † 서 정 연 ††

(Songwook Lee) (Jungyun Seo)

요 약 본 연구의 목적은 한국어 문장의 문법관계를 분석하는 데 있다. 주된 문제는 문장의 주어, 목적어, 부사어를 문장에서 찾아내는 것이다. 이 문제를 해결하기 위해서 한국어 구문 분석에서 발생하는 여러 중의성을 고려해야 한다. 우리는 문법관계의 중의성을 먼저 해결하고 그 다음에 주어진 명사구와 용언구의 문법관계 확률을 이용하여 용언구의 술어-논항 관계 중의성을 해소하는 통계적 방법을 제안한다. 제안된 방법은 어절간의 거리, 교차구조 금지, 일문일격의 원칙 등의 한국어 언어 특성을 반영하였다. 용언구와 명사구 사이의 문법관계에 대한 확률은 지지벡터 분류기를 이용하여 추정하였다. 제안된 방법은 문법관계 및 구문구조 부착 말뭉치를 이용하여 자동으로 문법관계를 학습하였고 주어, 목적어, 부사 각각의 문법관계분석에 대해 각각 84.8%, 94.1%, 84.8%의 성능을 얻었다.

키워드 : 부분구문분석, 문법관계, 지지벡터기계

Abstract This study aims to identify grammatical relations (GRs) in Korean sentences. The key task is to find the GRs in sentences in terms of such GR categories as subject, object, and adverbial. To overcome this problem, we are faced with the many ambiguities. We propose a statistical model, which resolves the grammatical relational ambiguity first, and then finds correct noun phrases (NPs) arguments of given verb phrases (VP) by using the probabilities of the GRs given NPs and VPs in sentences. The proposed model uses the characteristics of the Korean language such as distance, *no-crossing* and *case property*. We attempt to estimate the probabilities of GR given an NP and a VP with Support Vector Machines (SVM) classifiers. Through an experiment with a tree and GR tagged corpus for training the model, we achieved an overall accuracy of 84.8%, 94.1%, and 84.8% in identifying subject, object, and adverbial relations in sentences, respectively.

Key words : Shallow parsing, grammatical relation, Support Vector Machine

1. 서 론

문법관계 정보는 정보 검색, 정보 추출, 문서 요약 및 질의응답 시스템 등에 유용하게 이용된다[1-3]. 문장의 문법관계를 분석하는 연구가 그 동안 많이 수행되어왔다. [4]와 [5]는 단어와 품사 정보로 유한상태기계를 구성하여 문법관계를 분석하였다. 그러나 규칙 패턴과 문법이 수동으로 작성되어 다른 응용분야나 다른 언어로의 확장 및 통합에 어려움이 있는 단점이 있다. [6]에서는 계단형 은닉마르코프 모형을 이용하여 문법관계를 결정하였는데 품사 결정에 사용되어온 방법을 문법관계 결정에 적용하였다. 문법관계를 위한 태거는 어휘확률과 문맥확률을 이용하여 동작한다. [7]과 [8]은 지역순차패턴을 기억기반 학습법[9]을 이용하여 인식하는데 [7]은 단어의 품사정보를 가지고 명사구 단위화(chunking)와 주어, 목적어를 결정하였고 [8]은 [7]을 확장하여 문법관계 결정단계를 덧붙였다. [8]은 먼저 문장을 여러 개의 구로 단위화를 하고 그 다음에 각 구 사이의 문법관계를 부착하였다. 그들은 품사 및 어휘 정보를 이용하였고 명사구 및 용언구의 단위화로 더 나은 성능을 얻었다. 그러나 기억기반 학습 알고리즘은 학습할 때보다 실행할 때 느린 단점이 있고 모든 학습 말뭉치를 저장할 대용량의 공간이 요구되는 단점이 있다. [10]은 파서의 결과물에서 임계값을 조절하여 문법관계를 추출하였는데 재현율은 낮지만 높은 정확률을 가지는 문법관계를 추

· 본 연구는 2005년도 동서대학교 학술연구조성비의 지원을 받았습니다.

† 정 회 원 : 동서대학교 컴퓨터공학과 교수
leesw@dongseo.ac.kr
†† 종신회원 : 서강대학교 컴퓨터학과/바이오융합기술협동과정 교수
seojy@ccs.sogang.ac.kr

논문접수 : 2004년 10월 13일
심사완료 : 2005년 8월 19일

출하였다. 한국어의 문법관계 분석에 있어서 [11]에서는 격의 분포를 이용하는 통계적 방법을 이용하여 명사구의 격 중의성 문제의 해결을 시도했었으나 그 적용률(coverage)이 70% 정도로 낮았다. [12]는 한국어의 문법관계 분석에 있어서 보조사로 인해 발생하는 중의성을 용언의 하위범주화 사전, 용언과 체언간의 선택 제약 정보와 체언의 의미정보를 제공하는 시소러스 등의 지식과 몇 가지 휴리스틱 규칙을 이용하여 문제의 해결을 시도했었으나 그 성능이 구축된 지식 정보의 질에 영향을 받는 단점이 있다.

대부분의 이전 연구는 주요 구문요소와 그 구문요소의 머리말을 찾은 다음에 용언과 각 구문요소의 머리말 사이의 문법관계를 주어, 목적어 및 기타 수식어 등으로 결정하였다. 다시 말해 먼저 술어-논항 관계 중의성을 해결하고 문법관계의 중의성을 해결하는 것이다. 그러나 우리는 미리 문법관계를 결정하고 그 문법관계 정보를 이용하여 술어-논항 관계 중의성을 해소하는 새로운 방법을 제안한다.

한국어 문장의 문법관계에 있어서 주된 문제는 술어-논항 관계 중의성과 문법관계의 중의성의 해결이라 할 수 있다. 술어-논항 관계 중의성은 접속문제의 일종이고 문법관계의 문제는 한국어의 문법관계를 명시적으로 나타내는 조사의 생략과 보조사의 사용으로 나타난다고 할 수 있다. 또한 대부분 한국어 문장이 복합문 형태로 많이 나타나는 것 역시 문제를 더 어렵게 만든다.

본 연구에서 우리는 술어-논항 관계 중의성과 문법관계의 중의성을 해소하고자 한다. 주어-술어, 목적어-술어, 부사어-술어 등의 관계를 문장에서 올바로 찾아주는 부분 구문 분석기를 제안한다. 구문구조 부착 말뭉치로부터 문법관계를 부착하여 문법관계 확률 추정에 이용하였고 제안된 방법은 한국어의 언어특성 즉 어절간 거리, 교차구조 제한, 격 제한의 원칙 등을 반영하였다. 또한 지지벡터기계 분류기를 제안된 방법에서 사용되는 확률값의 추정에 사용하였고 실험결과, 제안된 방법은 좋은 성능을 얻었다.

2. 한국어 문장의 중의성

한국어 문장의 문법관계를 파악하기 위해서 술어-논항 관계 중의성과 문법관계의 중의성을 해소해야 한다. 이 장에서 문법관계 분석의 어려움을 몇 가지 예를 들어 설명한다.

2.1 술어-논항 관계 중의성

한국어는 부분적으로 자유 어순이고 복문의 쓰임이 단문보다 비교적 많다. 실제 실험에 사용된 말뭉치에서 한 문장이 평균 3.6개의 용언을 가지고 있었다. 이는 한국어 문장이 대체로 복문구조를 가지는 것을 반영하고 있다. 단문의 경우, 모든 명사구는 한 용언과 관계를 가지므로 구문 구조의 분석이 간단하다. 그러나 복문의 경우, 모든 명사구는 여러 용언들 중 하나의 용언과 관계를 가지므로 구문 구조의 분석이 어렵고 많은 중의성을 발생시킨다. 더군다나 생략된 성분이 있는 복문의 경우 분석의 어려움은 가중된다고 할 수 있다. 아래는 주어 관계에서 중의성이 나타나는 예문이다.

(1.a) 이 **책이** 신문 기사를 최대한 ①**활용하**는 비법을 ②**기술하**고 있다.

(1.b) **사람들이** 신문기사를 최대한 ①**활용하**는 비법을 ②**기술하**고 있다.

외형적으로 위 예문은 상당히 유사해 보이지만 구조적으로 완전히 다른 문장이다. 예1.a에서 '책이'는 '기술하다'의 주어이고 1.b에서 '사람들이'는 '활용하다'의 주어이다.

위와 같이 한국어의 술어-논항 관계 중의성은 영어의 전치사구 접속문제와 같은 접속문제의 일종이다. 이러한 접속문제를 다루는 여러 연구는 말뭉치로부터 통계적 정보를 자동으로 추출하는 통계적 방법이 주류를 이루고 있다[13-18]. 이들 연구에서 어휘 정보가 이러한 관계 중의성을 해소하는데 가장 유용한 정보 중 하나이다. 그러나 어휘 정보만으로는 술어-논항 관계 중의성의 해소에 불충분하다. 다음 예문을 살펴보자.

(1.c) 단군 **신화는** 원시공동체 사회가 ①무너지고 고조선이 ②세워지는 사실을 ③담고있다.

예문(1.c)에서 '신화는'의 술어는 ①'무너지다', ②'세우다', ③'담다' 중 하나가 될 수 있다. 어휘 공기빈도를중의성 해소에 사용했을 때, (신화는, 무너지다) 쌍의 빈도수가 (신화는, 세우다) 쌍이나 (신화는, 담다) 쌍 보다 더 많다. 따라서 어휘정보를 사용했음에도 '신화는'의 술어로 '무너지다'로 잘못 선택되게 된다.

그런데, 이런 오류는 격관계를 살펴봄으로 해결할 수 있다. 즉 하나의 용언이 하나의 주어 또는 하나의 목적어만을 가지게 하는 것으로 올바른 관계를 찾게 할 수 있다. 예문 (1.c)에서 ①'무너지다'와 ②'세우다'가 이미 주어를 가지고 있으면면 '신화는'의 술어는 '담다'로만 제한 할 수 있게 된다. 이런 문법관계를 사용하기 위해서는 술어-논항 관계 중의성을 해소하기 이전에 모든 문법관계를 미리 분석해야만 한다. 따라서 술어-논항 관계 중의성의 해결에 어휘 정보와 문법관계 정보가 중요한 단서가 된다.

2.2 문법관계의 중의성

문장 구조가 제대로 분석되었다고 하더라도 문법관계 중의성은 여전히 존재한다. 한국어 문법관계 분석에서 조사가 중요한 단서를 제공하며, 대부분의 명사구의 조사로 그 문법관계를 파악할 수 있다. 예를 들어 '-이/-

가'는 주격 조사이고, '-을/-를'은 목적격 조사이다. 이러한 조사들은 그 문법관계를 명시적으로 나타낸다. 그런데 보조사 '-은/-는, -만, -도, -부터' 등은 대부분의 문법관계에 사용될 수 있다. 다음 예문을 보자.

(2.a) 철수는 학교에 갔다. - 주어

(2.b) 철수는 선생님이 때렸다. - 목적어

(2.c) 철수가 아파도 학교는 갔다. - 부사어

위 예문에서 밑줄 친 명사구는 비록 같은 '-는'을 조사로 가지나, 문법관계는 각각 주어, 목적어, 부사어로 서로 다르다.

또한, 보조사의 쓰임뿐만 아니라, 조사가 생략되었을 경우에도 문법관계 분석에 어려움이 있다. 따라서 문법관계의 중의성은 보조사가 사용되었거나 조사가 생략된 명사구에서 발생하며 문법관계의 분석에 이를 고려해야 한다.

3. 문법관계분석을 위한 통계 모형

이 장에서 문법관계를 분석하는 방법을 설명한다. 우리는 주어, 목적어, 부사어 중에 하나의 문법관계만을 고려하며, 각각의 문법관계를 각각 독립적으로 분석하는 방법을 취한다. 즉, 주어 관계를 분석할 때 목적어나 부사어 관계는 전혀 고려하지 않는다. 현재 분석하려는 문법관계를 대상 문법관계라고 정의하자. 대상 문법관계가 정해지면 문장에서 가능한 모든 명사구와 용언구 쌍으로부터 대상 문법관계를 가지고 있는 쌍을 추출한다. 최종적으로 명사구와 용언구 쌍 중에서 최대확률을 가지는 쌍을 결정하여 대상 문법관계를 찾게 된다. 자세한 내용은 다음 절에서 설명한다.

3.1 대상 문법관계를 위한 후보 집합 선택

주어진 문장 $W_1,...,W_n$에서 모든 용언 $V_1, ..., V_t$을 품사 태그 정보로 추출한다. 용언열 중 관형절을 이끌며 대상 문법관계를 피수식 명사구와 이미 가지고 있는 용언은 [13]에서 제안한 방법을 이용하여 제외한다. 용언 열 $V_1, ..., V_t$ 중의 하나 이상의 용언과 대상문법관계 tgr을 가질 수 있는 모든 후보 명사구를 선택한다. 즉, 어절 W_i가 하나 이상의 용언과 식 (1)을 만족하는 대상 문법관계를 가질 수 있는 경우에만 후보 명사구가 된다. tgr에 대한 후보 명사구 열을 $C_1,...,C_m$라 하자. 식 (1)을 만족하는 모든 후보 명사구를 찾으면 우리는 문법관계의 중의성을 줄일 수 있다. W_i와 V_j의 문법관계를 r이라 하면 $r \in \{subj, obj, adv\}$ 이다.

$$\arg\max_r P(r \mid W_i, V_j) = tgr \quad (1)$$

주어진 $C_1,...,C_m$과 $V_1, ..., V_t$에 대한 대상 문법관계의 집합 R_{tgr}의 조건확률은 식 (2)와 같이 쓸 수 있다. 단, 후보 명사구와 어떤 용언의 관계는 다른 후보의 관계와 독립이라고 가정한다. V_k는 용언열 $V_1, ..., V_t$ 중 k번째 후보명사구 C_k와 관계를 가지는 용언을 의미한다.

$$P(R_{tgr} \mid C_1,...,C_m, V_1,...,V_t) \approx \prod_{k=1}^{m} P(r = tgr \mid C_k, V_k) \quad (2)$$

식 (2)를 최대화하는 (C_k, V_k)쌍의 열을 구하게 되면 우리는 대상 문법관계를 모두 찾을 수 있다. 모든 쌍은 한국어의 특성인 중심어 후위의 원칙 $loc(C_k) < loc(V_k)$을 만족해야 한다(함수 $loc(\cdot)$는 문장에서 인자가 나타난 위치를 뜻한다).

그림 1은 주어진 예문 "아버지는 철수가 학교에서 나오는 것을 보았다"의 대상 문법관계 tgr 관계들에 대한 초기 상태를 나타낸다. 각 상태는 주어진 W_i, V_j에 대한 tgr 관계를 나타내고 상태 출력 확률은 $P(r \mid W_i, V_j)$이다.

그림 1 가능한 모든 문법관계들에 대한 초기 상태들

가장 높은 확률값을 가지는 문법관계가 대상 문법관계가 아닌 상태들은 식 (1)에 의해 배제시킨다. 그림 2는 대상문법관계가 주어일 때 가능한 상태들을 나타낸다. 식 (1)에 의해 주어 관계가 아닌 상태들은 배제되었다. 우리의 목적은 식 (2)를 최대화하는 최적 경로를 찾는 것이다.

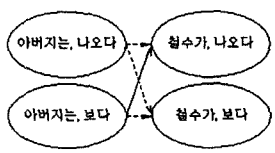


그림 2 확률 $\arg\max_r P(r \mid W_i, V_j) = subj$ 에 의해 남은 관계

3.2 교차구조 제한 및 격 제한 원칙에 의한 경로 제한

그림 2의 점선으로 된 경로는 교차구조이므로 제거되어야 한다. 따라서 우리는 식 (3)과 같이 이러한 경로를 제거하는 함수를 추가하였다. 경로 제한 함수를 추가하므로 제안된 방법은 더 이상 확률 모델이 아니라 주어진 $C_1,...,C_m$와 $V_1, ..., V_t$에 대한 tgr의 가중치를 나타내는 가중치 함수 형태가 된다. 식 (4)에서 $loc(V_k) = loc(V_{k-1})$은 같은 용언이 이웃하는 명사구와 동일한 문법관계를 가지는 것을 뜻하며, 이 때 경로 제한 함수가 0인 것은 격 제한

원칙을 반영하기 위한 부분이다. $loc(C_k) < loc(V_{k-1}) < loc(V_k)$은 이웃하는 구조와 교차구조가 만들어진 것을 나타내며, 이 때 경로 제한 함수가 0인 것은 교차구조 제한을 위한 부분이다.

$$f_{tgr}(C_1,...,C_m,V_1,...,V_t) \approx \prod_{k=1}^{m} P(r=tgr \mid C_k,V_k) \cdot f(C_k,V_k,C_{k-1},V_{k-1}) \quad (3)$$

$$f(C_k,V_k,C_{k-1},V_{k-1}) = \begin{cases} 0, & \text{iff } loc(V_k)=loc(V_{k-1}), \\ & or\ (loc(C_k)<loc(V_{k-1})<loc(V_k)) \\ 1, & \text{otherwise} \end{cases} \quad (4)$$

그림 3은 교차구조 제한과 격 제한을 통해 남아있는 상태와 경로를 나타낸다. 우리는 최적 경로를 제한된 경로들 중에서 찾기만 하면 된다.

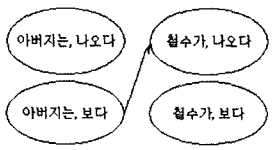


그림 3 교차구조 제한과 격 제한을 통한 경로 제한 예

비터비 알고리즘[19]을 이용하면 식 (3)을 최대화하는 최적경로를 쉽게 찾을 수 있다. 그림 3에는 단지 하나의 경로만 있으므로 '아버지는'이 '보다'의 주어이고 '철수가'가 '나오다'의 주어임을 쉽게 알 수 있다.

명사구와 용언구 사이의 거리 정보도 둘 사이에 관계를 가지는 데 큰 영향을 끼친다. 거리 정보는 두 단어의 관계에 있어서 중요한 정보임이 밝혀져 왔다. 우리는 어절간 거리 정보 d를 특정 문법관계의 거리 분포 $P(d|r=tgr)$로 반영하였다. 이 경우에 원거리에서 확률값이 0이 되는 경우가 발생할 수 있는데, 이 때에는 0이 아닌 확률값을 갖는 가장 가까운 거리의 값으로 근사해서 사용한다. 식 (3)에 거리 특성을 반영하면 식 (5)와 같다.

$$f_{tgr}(C_1,...,C_m,V_1,...,V_t) \approx \prod_{k=1}^{m} P(r_k=tgr \mid C_k,V_k) \cdot f(C_k,V_k,C_{k-1},V_{k-1}) \cdot P(d \mid r_k=tgr) \quad (5)$$

우리는 $P(r=tgr \mid C_k,V_k)$을 지지벡터기계를 이용하여 추정하는데 이는 다음 장에서 설명한다.

4. 지지벡터 기계를 이용한 문법관계 학습

$P(r|NP, VP)$는 MLE (maximum likelihood estimation) 등의 방법을 사용하면 추정할 수 있으나 어휘 정보의 사용에 따른 자료부족 문제가 발생한다. 우리는 문법관계 확률 $P(r|NP,VP)$을 지지벡터분류기[20]의 출력값으로 대체하여 사용한다. 명사구, 명사구의 조사, 용언 등의 어휘 자질과 각 어휘의 품사 자질을 사용하여 지지벡터분류기를 학습한다. 자질 벡터의 차원은 각 자질의 어휘의 개수의 총합이 되며 각각의 자질은 자질의 유무에 따라 이진으로 표현되었다. 학습데이타에서 출현하는 각 명사구와 용언 사이의 문법관계가 그 문법관계를 위한 지지벡터분류기의 양의 자질로 사용되었고 동시에 다른 문법관계를 위한 분류기의 학습에는 음의 자질로 사용되었다.

여러 실험 결과, SVM의 커널은 시스템 성능에 큰 영향을 끼치지 않아 선형 커널을 사용한다. SVM은 이진 분류기이므로 각 문법관계에 대한 분류기를 각각 학습하였고 실험에 SVM^{light}[1]를 이용하였다.

어떤 문법관계 r에 대한 SVM 분류기의 출력값을 $SVMr(NP, VP)$라고 하자. 3장에서 기술한 것과 유사하게 대상 후보 열 $C_1, ..., C_m$을 주어진 문장 $W_1,...,W_n$으로부터 $\arg_r \max SVM_r(W_i,V_j)=tgr$을 이용하여 구할 수 있다. 대부분의 경우 $\max SVM_r(W_i,V_j)$은 양수이지만 만약 이 값이 음수일 때에는 대상후보 열에서 제외한다. 후보 집합을 찾은 후에 $SVM_{tgr}(C_k,V_k)$을 식 (6)과 같이 사용하여 대상문법관계 결정 모델에 사용한다. 식 (10)은 식 (5)의 문법관계 확률을 SVM분류기의 출력값으로 바꾼 것이다.

$$f_{tgr}(C_1,...,C_m,V_1,...,V_t) \approx \prod_{k=1}^{m} SVM_{tgr}(C_k,V_k) \cdot f(C_k,V_k,C_{k-1},V_{k-1}) \cdot P(d \mid r=tgr) \quad (6)$$

식 (10)을 최대화하는 주어-용언과 같은 대상 문법관계를 찾는 게 목적이므로 분류기의 출력값을 0과 1사이로 사상하는 것은 불필요하다.

5. 실험 및 토의

우리는 구문구조가 부착된 한국어정보베이스 말뭉치[21]를 실험에 사용했다. 실험에는 145,630어절의 11,932문장을 사용했다. 이 말뭉치로부터 용언 및 수식 명사구 쌍 69,135개에 수동으로 문법관계를 부착하고 문법관계 학습에 사용하였다.

학습에서 사용되지 않은 5,056어절의 475문장을 평가에 사용하였고, 정확률과 재현율로 평가하고 F1 평가 -2*P*R/(P+R) [22]로도 나타냈다.

1) SVM^{light} 시스템은 http://svmlight.joachims.org에서 얻을 수 있다.

표 1은 제안된 방법의 평가 말뭉치에서의 성능을 나타낸다.

표 1 문법관계 분석 결과

	Subj	Obj	Adv	Total
P	83.9	97.0	86.3	88.3
R	85.7	91.2	83.3	86.4
F1	**84.8**	**94.1**	**84.8**	**87.4**
Proportions %	36.2	28.4	35.4	100

표 1에서와 같이 목적어 관계가 주어나 부사어보다 좋은 성능을 나타냈다. 학습 말뭉치에서 평균 술어의 개수는 주어, 목적어, 부사어 각각의 경우에 3.8개, 2.0개, 2.3개였다.

그림 [4]는 제안한 방법에서 두 동사의 목적어를 찾아낸 예를 나타낸다. 주어, 목적어, 부사어에 해당하지 않는 어절의 관계는 null로 나타내며, 해당 어절이 어떤 용언의 논항인지는 세번째 열의 어절 번호로 표시하고 그 문법관계는 subj, obj, adv 등으로 첫 번째 열에 표시를 한다. 두번째 열은 해당 어절의 어절 번호이며, 네번째 열은 품사가 부착된 해당 어절이다[23].

```
null 1    -1 넷째/nno+ ,/sp
null 2    -1 채택/ncpa+ 되/xsv+ ㄴ/etm
obj 3      7 전략/ncn+ 을/jco
obj 4      5 끈기/ncn+ 를/jco
null 5    -1 가지/pvg+ 고/ecc
null 6    -1 강력히/mag
null 7    -1 실천/ncpa+ 하/xsv+ 는/etm
null 8    -1 것/nbn+ ./sf
```

그림 4 문법관계 분석의 예

대부분의 오류 중 가장 심각한 오류는 문법관계 후보열이 잘못된 후보를 포함하고 있는 경우이다. 이 오류는 주어진 명사구와 용언 사이의 가장 높은 확률을 가지는 대상 문법관계를 위한 후보 명사구 결정 단계에서 발생하는 오류이므로 다른 언어 특성을 반영해도 제거 되지 않는 오류이다. 문법관계 후보열 결정에서의 정확률은 평가데이타에 대해 97.7%였다. 올바른 후보 중 3.1%가 후보열에 포함되지 못했고 포함된 후보 중 약 3.5%가 오류 후보였다. 그 외의 대부분의 오류는 자료부족문제에 기인한다. 실제 데이타에서 평가 데이타의 약 15%만이 학습 데이타에서 발견되었고 평가 데이타의 19%에서는 하나의 어휘도 발견되지 않았다. 이런 오류들은 학습 데이타를 늘이면 어느 정도 감소시킬 수 있을 것이다. 격 제한 원칙의 예외로 나타나는 이중 주어나 이중 목적어를 갖는 문장이 네 개 있었다. 이러한 이중 주어나 이중 목적어의 경우에는 이들을 문장 성분으로 갖는 동사에 대한 사전 구축으로 해결해 나갈 수 있으리라 생각된다.

표 2에서는 [15]의 구문분석기의 결과와 비교를 했다. [15]의 결과에는 문법관계가 명시되어 있지 않기 때문에 구조가 올바르게 분석되었으면 문법관계도 올바르다고 가정하였다. 1,743어절의 195개의 문장으로 제안된 방법과 [15]의 결과와 비교했다.

표 2 [15]의 구문분석기와 문법관계 분석 결과 비교

		195개 문장			
		Subj	Obj	Adv	Total
제안된 방법	P	83.8	96.2	90.8	89.5
	R	86.1	89.4	88.1	87.7
	F1	84.9	92.8	89.4	88.6
[15]의 방법	P	81.5	84.4	66.1	77.5
	R	81.9	83.0	70.7	78.9
	F1	81.1	83.7	68.4	78.2
Proportions (%)	%	36.1	33.8	30.1	100

표 2에서와 같이 제안된 방법이 한국어에서 최상의 결과를 보이는 구문분석기 중 하나인 [15]의 구문분석기보다 문법관계 분석에 있어서 좋은 성능을 보였다. 결과적으로, 제안된 방법의 실험결과를 통해 술어-논항 관계 중의성을 문법관계 정보를 이용하여 해소할 수 있다고 할 수 있다. 영어권의 경우, [8]은 F1평가로 주어, 목적어 각각 81.8%와 81.0%의 성능을 얻었으나 본 논문에서 제안한 방법과 직접적 비교는 어렵지만 [8]에서 사용된 말뭉치의 양이 본 연구에서 사용된 양보다 상당히 많은 것을 고려할 때, 우리가 더 적은 말뭉치로 더 좋은 성능을 얻었다고 할 수 있다.

6. 결론 및 향후 과제

본 논문에서 한국어 문법관계를 결정하는 부분 구문 분석기를 제안하고 구문구조 말뭉치를 이용하여 구현하였다. 제안된 부분 구문 분석기는 주어-용언, 목적어-용언, 부사어-용언 관계를 각각 독립적으로 분석하며 문법관계에 대한 통계 정보는 구문구조와 문법관계 부착 말뭉치에서 자동적으로 추출하였다. 실험을 통해 주어, 목적어, 부사어 관계의 결정에 각각 84.8%, 94.1%, 84.8%의 정확도를 얻었다. 제안된 방법은 명사구와 용언구 사이의 문법관계에 대한 정보를 지지벡터 분류기를 이용하여 학습하였고 학습된 분류기를 이용하여 문법관계의 중의성과 술어-논항 관계 중의성 해소를 시도하였고 좋은 결과를 얻었다. 한국어의 언어특성인 교차구조 제한, 격제한 원칙 및 어절간의 거리 등을 제안된 방법에 반영하였다.

더 신뢰할 만한 결과와 더 나은 성능을 위해서 좀더 많은 데이타가 필요하며, 현재 제안된 방법을 전체 구문 분석기로 확장할 수 있는 방법을 연구하고 있다.

참 고 문 헌

[1] Grenfenstette, G. (1997). SQLET: Short query linguistic expansion techniques, palliating one-word queries by providing intermediate structure to text. *In Proc. of the RIAO'97*, 500-509.

[2] Palmer, M., Passonneau, R., Weir, C. & Finin, T. (1993). The KERNEL text understanding system. *Artificial Intelligence*, 63, 17-68.

[3] Yeh, A. (2000). Using existing systems to supplement small amounts of annotated GRs training data. *Proc. of the ACL2000*, 126-132. Hong Kong.

[4] Grenfenstette, G. (1996). Light parsing as finite-state filtering. *Workshop on Extended Finite State Models of Language, ECAI'96*, Budapest, Hungary.

[5] Aït-Mokhtar, S. & Chanod, J-P. (1997). Subject and object dependency extraction using finite-state transducers. In *Proceedings of the ACL/EACL'97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources*, 71-77. Madrid, Spain.

[6] Brants, T., Skut, W. & Krenn, B. (1997). Tagging grammatical functions. In Proceedings of the 2nd Conference on EMNLP, 64-74. Providence, RI.

[7] Argamon, S., Dagan, I. & Krymolowski, Y. (1998). A memory-based approach to learning shallow natural language patterns. In *Proceedings of the 36th Annual Meeting of the ACL*, 67-73. Montreal, Canada.

[8] Buchholz, S., Veenstra, J. & Daelemans, W. (1999). Cascaded GR assignment. In *Proceedings of the Joint Conference on EMNLP and Very Large Corpora*, 239-246.

[9] Stanfill, C. & Waltz, D. (1986). Toward memory-based reasoning. *Communications of the ACM*, 29:1213-1228.

[10] Carroll, J. & E. Briscoe (2002). High precision extraction of GRs. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, Taipei, Taiwan.

[11] 양재형, 김영택, "통계정보를 활용한 한국어 미지격 명사구의 문법기능 결정", *정보과학회논문지*, Vol. 21, No. 5, pp. 808-15, 1994. 5.

[12] 양재형, 심광섭, "시소러스와 하위범주화 사전을 이용한 격모호성 해결", 정보과학회논문지(B) 제26권 제9호, 1999. 9.

[13] Lee, S., Seo, J. & Jang, T. Y. (2003). Analysis of the grammatical functions between adnoun and NPs in Korean using Support Vector Machines. *Natural Language Engineering*, Cambridge University Press, Vol. 9, No. 3, pp. 269-280, Sept.

[14] Hindle, D. and Rooth, M. (1993). "Structural ambiguity and lexical relations," *Computational Linguistics*, 19:103-120.

[15] Lee, K. J., Kim, J. H., & Kim, G. C. (1997). An Efficient Parsing of Korean Sentence Using Restricted Phrase Structure Grammar, *Computer Processing of Oriental Languages*, Vol. 12, No. 1, pp. 49-62.

[16] Collins, Michael. (1996). A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of ACL-96*, Sant Cruz, CA, USA.

[17] Charniak, E. (2001). Immediate-head parsing for language models. *Proceedings of ACL 2001*, 116-123.

[18] Srinivas, B. (2000). A lightweight dependency analyzer for partial parsing. *Natural Language Engineering*, 6(2), 113-138.

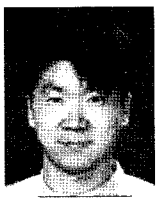
[19] Viterbi, A. J. (1967). Error bounds for convolution codes and an asymptotically optimal decoding algorithm. *IEEE trans. on Information Theory*, 12:260-269.

[20] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.

[21] Lee, K. J., KIM, J. H., Choi, K. S. & Kim, G. C. (1996). Korean syntactic tagset for building a tree annotated corpus. *Korean Journal of Cognitive Science*, 7(4):7-24.

[22] Rijsbergen, C.J.van. (1979). *Information Retrieval*. Buttersworth, London.

[23] 김길창, 임해창, 서정연, 나동렬, "한국어 이해에 나타나는 중의성 문제 처리 모델에 관한 연구", 연구결과 보고서, 한국과학재단, 1997.10.



이 성 욱

1996년 서강대학교 컴퓨터학과 학사 1998년 서강대학교 컴퓨터학과 석사 2003년 서강대학교 컴퓨터학과 박사 2003년~2004년 서강대학교 산업기술연구소 연구원. 2003년~2005년 서강대학교 정보통신대학원 대우교수. 2004년~2005년 LG전자 기술원 선임연구원. 2005년~현재 동서대학교 컴퓨터정보공학부 전임강사. 관심분야는 형태소 및 구문분석, 단어의미분별, 대화 언어처리

서 정 연

정보과학회논문지 : 소프트웨어 및 응용 제 32 권 제 9 호 참조