중심어 간의 공기정보를 이용한 한국어 확률 구문분석 모델

이 공 주† · 김 재 훈††

요 약

구문 분석에서 가장 큰 문제점 중 하나는 구문 구조의 중의성을 어떻게 해결하느냐에 달려있다. 확률 구문 규칙은 구문 구조의 중의성 해결에 한 방법이 될 수 있다. 본 논문에서는 중심어 간의 공기정보를 이용하여 한국어 구문 구조의 중의성을 해결하는 확률 모델을 제안하고자 한다. 중심어는 어휘를 이용하기 때문에 자료 부족 문제를 야기시킬 수 있다. 이 때문에 자료부족 문제를 어떻게 해결하느냐에 따라 어휘 정보 사용의 성공이 결정될 수 있다. 본 논문에서는 구문규칙을 단순화하고 Back-off 방법을 이용해서 이 문제를 완화한다. 제안된 모델은 실험 데이터에 대해 약 84%의 정확도를 보였다.

Korean Probabilistic Syntactic Model using Head Co-occurrence

Kong Joo Lee† · Jae-Hoon Kim††

ABSTRACT

Since a natural language has inherently structural ambiguities, one of the difficulties of parsing is resolving the structural ambiguities. Recently, a probabilistic approach to tackle this disambiguation problem has received considerable attention because it has some attractions such as automatic learning, wide-coverage, and robustness. In this paper, we focus on Korean probabilistic parsing model using head co-occurrence. We are apt to meet the data sparseness problem when we're using head co-occurrence because it is lexical. Therefore, how to handle this problem is more important than others. To lighten the problem, we have used the restricted and simplified phrase-structure grammar and back-off model as smoothing. The proposed model has showed that the accuracy is about 84%.

키워드 : 어휘 공기정보(lexical cooccurrence), 확률 구문 분석(probabilistic parsing)

1. 서 론

대부분의 자연언어 처리에서와 마찬가지로 구문 분석에서도 중의성 해결이 가장 큰 문제이다. 그러므로 구문 분석기는 기본적으로 중의성 해결 방법을 제공해야 한다. 최근에 가장 많은 관심을 가지는 중의성 해결 방법은 통계적 접근 방법이다. 이 방법은 대량의 구문 트리가 태깅된 코퍼스로부터 확률 정보를 추출하고, 이 확률 정보를 이용하여 구문 분석의 중의성을 해소할 수 있다[13].

확률 문법을 이용한 중의성 해결은 매우 간단하다. 최종적인 구문 트리의 확률값은 그 구문 트리에서 사용된 규칙의 확률값의 곱으로 계산된다. 그렇게 함으로써 가장 높은 확률값을 갖는 최종 구문 트리가 가장 적절한 결과로서 결정된다. 이와 같은 방법은 자주 사용된 규칙들이 많이 적용된 구문 트리가 최종 결과로써 선호된다. 그러나 쉽게 추측할 수 있듯이 이와 같은 단순한 방법은 매우 낮은 정확도를 보인다. 이런 이유로 최근에는 확률 문법의 문맥으로 여러 형태의 어휘 정보를 사용하며, 이와 같은 방법은 중의성 해결의 정확도를 쉽게 높일 수 있다[20, 22]. 그러나 어휘 정보를 추출하기 위해서는 대량의 코퍼스가 필요하며, 일반적으로 코퍼스를 구축하는 데는 많은 시간과 인력이 소요된다. 또한 충분한 양의 코퍼스가 존재한다고 해도 여전히 자료 부족(data sparseness) 문제를 피할 수 없다[24]. 이 때문에 자료 부족 문제를 어떻게 해결하느냐가 어휘 정보 사용의 성공 여부를 결정한다고 할 수 있다.

본 논문에서는 어휘 정보로서 중심어 간의 공기정보와 확률 문법을 사용하는 한국어의 확률 구문분석 모델을 제안한다. 공기정보는 용언의 격틀 정보나 하위범주화(subcategorization)와 유사한 형태의 정보이다. 예문 (1)을 통해서 중

※ 이 논문은 한국과학재단의 해외 post-doc 연구지원비와 2002년도 두뇌한국(BK)21 사업에 의해서 지원되었음.
† 정 회 원 : (주)한국마이크로소프트 연구원
†† 정 회 원 : 한국해양대학교 컴퓨터공학과 교수
논문접수 : 2002년 8월 30일, 심사완료 : 2002년 9월 27일

심어 간의 공기정보가 확률 구문분석에 미치는 영향을 살펴보자.

(1) *산이 높으면 온도는 심하게 차이가 나고 날씨 변화도 큽니다.*

예문 (1)에서 '산이'와 결합될 수 있는 용언으로는 '높다', '심하다', '나다', 또는 '크다'가 가능하다. 다시 말해서, '산이'와 어떠한 용언이 결합하느냐에 따라 구조적 중의성이 발생할 수 있다. 두 어휘간의 공기정보를 사용한다는 것은 '산이'와 함께 쓰일 가능성이 가장 높은 용언 정보를 이용하여 이를 구문 분석에 이용하는 것이다. 이와 같은 공기정보는 구문 트리가 태깅된 코퍼스로부터 자동으로 추출할 수 있다.

이미 영어권에서는 여러 형태의 어휘 정보를 확률 구문분석의 성능을 향상시키기 위해 널리 사용하고 있다[18, 20, 22]. 그러나 한국어의 경우, 충분한 양의 코퍼스가 부족하기 때문에 확률 구문 분석시 어휘 정보를 충분히 사용할 수 없다. 본 논문에서는 이와 같은 문제를 완화시키기 위해 구구조 문법 규칙의 형식을 제한하였으며, 또한 자료부족 문제를 해결하기 위한 방법으로 널리 사용되는 back-off 평탄화 모델을 사용하였다.

본 논문의 구성은 다음과 같다. 2절에서는 어휘 정보를 이용한 구문 분석기에 대한 관련연구를 간략히 살펴본다. 3절에서 자료부족 문제를 다루기 위해 한국어 구구조 문법 규칙의 간략화에 대해서 소개하고, 4절에서 중심어와 중심어 간의 공기정보를 정의하도록 한다. 5절에서 공기정보를 이용한 확률 모델을 기술하고 6절에서 매개변수 평탄화 과정에 대해서 기술한다. 7절에서 실험 및 성능 평가에 대해서 기술하고 끝으로 8절에서 논문의 결론을 맺도록 한다.

2. 관련 연구

본 절에서는 식 (1)과 같은 일반적인 확률 구문분석 모델에 어휘 정보를 추가하는 구문분석 방법에 대해서 간략히 살펴보고자 한다.

$$P(T \mid S) = \prod_{r_i \in T} P(r_i \mid S) \quad (1)$$

여기서 T와 S는 각각 파스 트리와 입력문장이고 r_i는 파스 트리 T에 포함된 구문규칙이다. 입력 문장 S에 대한 파스 트리 T의 확률 값은 T에 포함된 모든 구문규칙 r_i의 확률 값의 곱으로 정의된다.

이와 같은 기본 모델에 어휘 정보를 첨가하는 다양한 모델들이 소개되었다. [16]에서는 어휘 정보가 많이 포함된 결정 트리를 이용한 확률 구문분석 방법을 제안했으며, 그 확률 모델은 식 (2)와 같다.

$$P(T \mid S) = \prod_{d_i \in T} P(d_i \mid d_{i-1}\, d_{i-2} \cdots, S) \quad (2)$$

여기서 d_i는 각각의 결정(decision)이다. 파스 트리의 확률은 각 결정 단계의 확률 값 d_i의 곱으로 표현되었다. d_i는 구구조 문법의 비단말 기호와 품사뿐만 아니라 어휘도 포함한다. 이 방법은 결정트리와 같은 규칙 기반 접근 방법과 확률 문법을 결합하여 좋은 결과를 얻었다. [18]에서는 동사의 격틀 정보에 해당하는 어휘 정보를 이용한 확률 구문분석 방법을 제안했으며 그 확률 모델은 식 (3)과 같다.

$$P(T \mid S) = \prod_{c \in T} P(h(c) \mid t(c), h\rho(c), h\rho^2(c), a(c)) \times p(r(c) \mid h(c)) \quad (3)$$

여기서 c는 구문성분(constituent)이고, $h(c)$는 c의 중심어(head)이고, $\rho(c)$는 c의 부모성분을 의미하며, $r(c)$는 c를 확장하는데 사용한 규칙이다. 즉, 구문 트리를 구성하는 각각의 구성성분에 대해, 그 구성성분 자체의 중심어와, 부모의 중심어, 그리고 그 부모의 부모의 중심어에 대한 확률 정보를 동시에 이용하였다. 그렇게 함으로써, 예를 들어 동사 'give'의 경우에 'give'의 직접목적어와 간접목적어의 어휘 관계를 동시에 하나의 확률식에 표현할 수 있었다. [22]에서는 어휘 의존 관계를 이용한 통계 구문분석 방법을 제안하였으며, 두 어휘간의 의존관계를 중의성 해소에 사용하였다. 우선, 입력 문장을 하나의 중심어로 표현될 수 있는 단위(baseNP)로 다시 재편성한 후, baseNP와 그 중심어의 의존관계만을 이용하여 구문 확률을 계산하였다. 이 이외에도 영어에 대해서 [20]에서는 어휘들간의 의존관계를 이용하고 있으며, 한국어의 경우에도 의존관계의 지배어를 중심으로 부모 중심어들의 정보를 이용한 통계 구문분석 방법이 제안되었다[21]. 또한 [10]에서는 한국어에 대해 상호정보(mutual information)를 이용하여 확률 의존 구문 분석 방법을 제안하였다.

이들 연구에서 보는 바와 같이 구문구조의 중의성을 해소하기 위한 방법으로 구문 규칙 혹은 의존관계에 있는 어휘 정보를 많이 사용하고 있다. 한국어에 대해서는 앞에서 몇 가지 방법을 언급했지만 구문 트리가 태깅된 코퍼스가 부족하기 때문에 확률적 구문 분석에 대한 연구가 많이 진행되지 못한 상태이다. 그러나 최근에는 전체 문장에 대한 분석보다는 동사구절이나 명사구절과 같이 조금 작은 부분에 대한 연구들이 진행되고 있다[1, 2, 4, 8, 11].

[4]에서는 주로 부사에 의해서 발생하는 구조적 중의성만을 다루고 있다. 코퍼스에서 발생 빈도가 높은 부사들을 추출하고 이것들에 대하여 수식어 사전을 구축하였다. 수식어 사전은 문장에서의 수식어와 피수식어의 위치 정보, 이들간의 공기 관계, 문장에서의 패턴 등의 통계적인 정보를 포함하고 있으며, 이를 이용하여 부사 수식어에 대한 구조적 중

의성을 해결코자 하였다. [11]에서는 결합범주문법(combinatory categorial grammar)을 기반으로 주로 병렬 명사구절의 처리 방법을 다루고자 하였다. 구문 트리가 태깅된 코퍼스로부터 명사간의 공기 유사 정보를 추출하고, 구문 분석 단계에서 이를 이용하여 중심 명사간의 공기 유사도를 비교하여 구조적 중의성을 해결하였다. [8]에서는 국어 사전의 뜻풀이말을 대상으로 한 구문 분석기를 구축하였다. 우선, 국어 사전의 뜻풀이말에 대해 수동으로 구문 트리 태깅을 수행하고, 이로부터 411개의 구문 규칙과 그 확률을 추출하였다. 각 문장의 스코어는 그 문장의 트리에 쓰인 규칙들의 확률값과 각 어절의 태그 확률값의 곱으로써 표현하고, 가장 높은 스코어를 갖는 파스 트리가 정답으로 추출되도록 하였다.

구문 분석의 기본 정보가 될 수 있는 하위범주화 사전 구축[9], 격틀 자동 구축[12], 구문의미사전 구축[3]에 관한 지속적인 연구들이 진행되고 있다. 이러한 정보들은 구문 분석의 성능 향상에 많은 기여를 할 것이다.

3. 제한된 형식의 한국어 구구조 문법

확률 구문분석 모델에서는 여러 형태의 확률 정보가 필요하다. 일반적으로 이와같은 확률 정보는 대량의 구문 트리 코퍼스를 이용해서 구해진다. 구문 트리 코퍼스를 구축하는 일은 많은 인력, 노력과 시간이 소요된다. 저자는 이 비용을 줄이고 확률 구문분석 모델의 자료부족 문제를 다소 완화하기 위해 제한된 형태를 가지는 구구조 문법을 제안했다[5]. 본 절에서는 [5]에서 제안된 한국어 구구조 문법에 대한 간략한 설명을 하고자 한다. 이 구구조 문법은 기능어의 특성에 따라 다음과 같이 크게 세 가지로 나눠 볼 수 있다[1].

3.1 구절 내 관계

구절의 문법적 성분의 변화를 유발시키거나 속성을 결정하는 기능어들이다. 이에 속하는 기능어들을 <표 1>에서 볼 수 있다.

〈표 1〉 구절 내 관계 명시 기능어들

구절 내 관계 명시	격조사	서술격조사(jp), 호격조사(jcv), 종결보조사(jxf)
	어 미	명사형어미(etn), 선어말어미(ep), 종결어미(ef)
	접 사	명사파생접사(xsn), 동사파생접사(xsv), 형용사파생접사(xsm), 부사파생접사(xsa)

대표적으로 서술격조사의 경우에는 명사구절을 용언구절로 바꾸는 역할을 수행한다. 또한, 명사형어미의 경우에는 동사구절이나 형용사구절을 명사구절로 바꾸는 역할을 수행한다. 이와 같은 기능어가 사용된 구절에 대한 구구조 규칙의 형태와 간단한 예제는 <표 2>와 같다.

1) 이에 관한 좀더 자세한 설명이나 예는 [6]을 참조하기 바랍니다.

〈표 2〉 TYPE 1의 규칙과 그 예제

TYPE 1 : $A \rightarrow B + \tau$	
$VP \rightarrow NP + jp$	명사구절이 서술격조사와 결합하여 동사구절로 변함
$NP \rightarrow VP + etn$	동사구절이 명사형어미와 결합하여 명사구절로 변함
$NP \rightarrow ncn + xsn$	명사와 접사가 결합하여 명사구절을 이룸

여기서 τ는 구절내 관계를 명시해주는 기능어들과 보조용언구절(AUXP), 그리고 ε이 가능하다. 이 규칙이 의미하는 바는 구절 B가 τ에 의해서 그 속성이 결정되거나 기능이 변화함을 의미한다.

3.2 구절간 관계

이에 속하는 기능어들은 문법적 성분의 변화와 더불어 두 구성 성분간의 문법적 관계를 명시하는 역할을 담당한다. 이에 속하는 기능어들을 <표 3>에서 볼 수 있다.

〈표 3〉 구절 간 관계 명시 기능어들

구절 간 관계 명시	격조사	주격조사(jcs), 목적격조사(jco), 보격조사(jcc), 부사격조사(jca), 관형격조사(jcm), 공동격조사(jcj), 인용격조사(jcr), 접속격조사(jcj), 통용보조사(jxc)
	어 미	대등적연결어미(ecc), 종속적연결어미(ecs), 관형사형어미(etm)

구절간 관계 명시에 해당하는 기능어가 사용된 구구조 규칙의 형태는 TYPE 2와 TYPE 3의 두 가지가 가능하다.

<표 4>의 TYPE 2 규칙에서 γ는 접속격조사와 대등적 연결어미를 제외한 구절간 관계 명시의 기능어들이다. 또한 격조사의 생략이 가능하므로 γ는 ε이 가능하다. TYPE 2의 규칙이 의미하는 바는 구절 B가 구절 C와 γ의 관계로 결합하여 새로운 구절 A를 형성함을 의미한다.

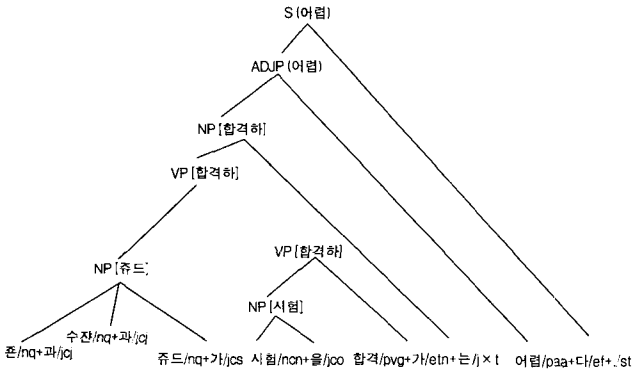
〈표 4〉 TYPE 2의 규칙과 그 예제

TYPE 2 : $A \rightarrow B + \gamma \quad C$	
$VP \rightarrow NP + jcs \quad VP$	명사구절과 동사구절이 주격관계로 결합하여 새로운 동사구절을 형성
$NP \rightarrow VP + etm \quad NP$	동사구절이 명사구절을 관형어로 한정
$VP \rightarrow VP + ecs \quad VP$	동사구절과 동사구절이 종속적으로 연결되어 새로운 동사구절을 형성

<표 5>의 TYPE 3 규칙은 병렬구조를 나타낸다. 여기서 γ'는 대등적 연결어미와 접속격조사, 그리고 나열을 표현하는 쉼표(sp)와 단어 접속 부사(maj) 등이 가능하다. TYPE 3의 규칙을 TYPE 2와 다르게 정의한 것은 이에 속하는 규칙의 오른편(RHS)에는 여러 개의 구절이 올 수 있기 때문이다.

〈표 5〉 TYPE 3의 규칙과 그 예제

TYPE 3 : $A \rightarrow A_1 + \gamma\ A_2 + \gamma \cdots A_n$	
$VP \rightarrow VP + ecc\ VP + ecc\ VP$	병렬 동사 구절
$NP \rightarrow NP + jcc\ NP + jcc\ NP$	병렬 명사 구절
$NP \rightarrow ncn + sp\ ncn + sp\ ncn + sp\ ncn$	병렬 명사 구절



(그림 1) 예제 문장과 그에 대한 구문 트리

(그림 1)은 예제 문장에 대하여, 앞에서 언급한 형태의 규칙들로 분석한 결과 구문 트리이고, 그에 해당하는 규칙은 <표 6>에서 볼 수 있다.

〈표 6〉 (그림 1)에서 사용된 구문 규칙들

TYPE 1	$NP \rightarrow ncn$	(시험을)
	$NP \rightarrow VP + etn$	(시험을 합격하기)
	$S \rightarrow ADJP + ef + sf$	(존과 수잔과 … 어렵다.)
TYPE 2	$VP \rightarrow NP + jco\ \ pvg$	(시험을 합격하)
	$VP \rightarrow NP + jcs\ \ VP$	(존과 수잔과 … 합격하)
	$ADJP \rightarrow NP + jxt\ \ paa$	(존과 수잔과 … 어렵다)
TYPE 3	$NP \rightarrow nq + jcj\ nq + jcj\ nq$	(존과 수잔과 쥬드)

4. 한국어 확률 구문 분석을 위한 중심어 간의 공기정보

2절에서 언급했듯이 어휘간의 관계 정보는 구문분석의 성능을 개선하기 위해 널리 사용된다. 본 논문에서는 중심어 간의 공기정보를 이용하여 성능을 개선하고자 한다. 중심어는 임의의 구절(phrase)에서 그 구절을 대표할 수 있는 단어이다[15]. 예를 들어, 명사구절의 중심어는 그 구절의 가장 중심이 되는 명사가 될 것이고, 동사구절의 중심어는 가장 의미있는 동사가 될 것이다. 한국어는 잘 알려져 있는 바와 같이 중심어 후위 언어(head-final language)이다. 그렇기 때문에 몇몇 예외를 제외하고는 대부분의 중심어는 그 구절의 가장 마지막 단어가 될 것이다. 이와 같은 중심어 정보는 앞 절에서 정의한 구문 규칙 타입에 따라 자동으로 정의할 수 있다(<표 7>).

〈표 7〉 각 규칙 타입에 따른 중심어

		Head(A)
TYPE 1	$A \rightarrow B + \tau$	Head(B)
TYPE 2	$A \rightarrow B + \gamma\ \ C$	Head(C)
TYPE 3	$A \rightarrow A_1 + \gamma\ A_2 + \gamma \cdots A_n$	Head(A_n)

<표 8>은 각 규칙의 형태에 따라 중심어가 되는 구절과 중심어 간의 공기정보를 정의하고 있다. 우선, 구절 A의 중심어를 A^h라고 표시하기로 하자. TYPE 1의 경우, 구절 B의 중심어 자체가 구절 A의 중심어가 되며, 이러한 형태의 규칙에서는 중심어 간의 공기정보가 발생하지 않는다. 이는 TYPE 1의 규칙은 다른 구절과의 관계를 명시하는 것이 아니고 단지 구절 자체의 문법적 성질이나 속성의 변화만을 유발하기 때문이다. TYPE 2 규칙의 경우에는 구절 C의 중심어가 구절 A의 중심어가 되며 이때에는 (B^h, γ, C^h)와 같은 중심어 간의 공기정보가 발생한다. 이는 중심어 B^h와 중심어 C^h가 문법적 관계 γ에 의해서 발생함을 의미한다. TYPE 3의 규칙은 규칙의 RHS에 두개 이상의 구절이 존재한다는 점만 제외하고는 TYPE 2의 규칙과 유사하다. 그렇기 때문에, TYPE 3의 규칙에서는 한 개 이상의 중심어 간의 공기정보가 발생할 수 있다. TYPE 3의 규칙에서 중심어가 되는 구절은 RHS의 가장 오른쪽 구절인 A_n이 되며, RHS의 나머지 n-1개의 구절들과 A_n구절과의 공기정보가 발생하게 된다.

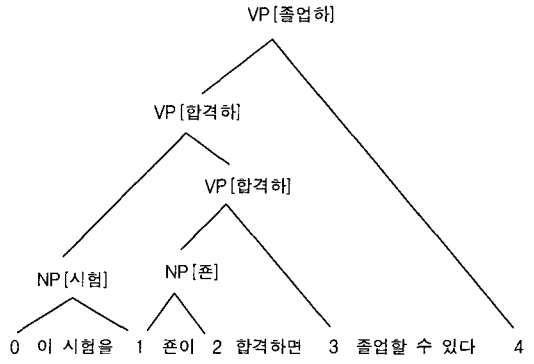
〈표 8〉 각 규칙 형태에 따른 중심어 간의 공기정보

		중심어 간의 공기정보
TYPE 1	$A \rightarrow B + \tau$	N/A
TYPE 2	$A \rightarrow B + \gamma\ \ C$	(B^h, γ, C^h)
TYPE 3	$A \rightarrow A_1 + \gamma\ A_2 + \gamma \cdots A_n$	(A_1^h, γ, A_n^h) (A_2^h, γ, A_n^h) … $(A_{n-1}^h, \gamma, A_n^h)$

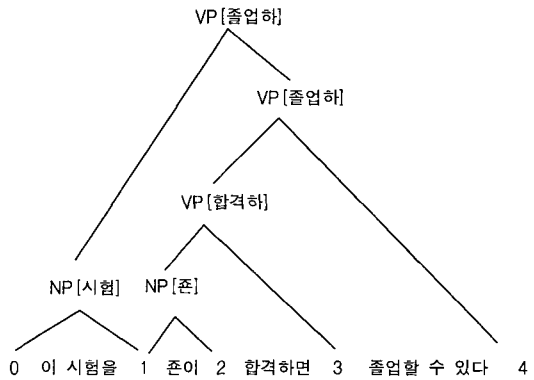
(그림 2)에서는 동일한 문장에 대한 구조적 중의성을 보여주고 있다. 명사구절 $NP_{1,2}$의 중심어는 '쥔'이며, 동사구절 $VP_{1,3}$의 중심어는 '합격하다'가 된다. ($NP_{i,j}$는 i번째 어절부터 j번째 어절까지를 포함하고 있는 명사구절을 의미한다.) 구문 트리(A)에서 $VP_{0,3}$의 중심어는 $VP_{1,3}$의 중심어와 동일하다. 이는 구문 규칙 $VP \rightarrow NP + jco\ VP$에서 LHS의 VP의 중심어는 RHS의 마지막 VP의 중심어로부터 전달되어지기 때문이다. 중심어 간의 공기정보가 구조적 중의

성 해소에 도움이 되는 과정은 다음과 같다.

(그림 2)의 트리(A)의 경우, 구절 $NP_{0,1}$와 $VP_{1,3}$이 결합하여 구절 $VP_{0,3}$을 형성한다. 이와 같은 구절 결합으로부터 목적어 '시험'의 용언은 '합격하다'임을 알 수 있다. 반면에 트리(B)의 경우, 구절 $NP_{0,1}$과 $VP_{1,4}$의 결합으로부터 목적어 '시험'의 용언은 '졸업하다'임을 알 수 있다. 일반적으로 '시험을'이라는 단어는 '졸업하다'라는 용언보다는 '합격하다'라는 용언과 더 잘 어울린다. 이와 같이 구문 구조의 중의성이 발생될 때, 더 잘 어울리는 어휘의 조합을 갖는 구문 트리가 더 적절한 결과임을 알 수 있다. 즉, 어휘의 조합 ('시험을', '졸업하다')보다는 ('시험을', '합격하다')가 더 적절한 조합이므로 이러한 정보를 사용하여 구문 분석기는 구문 트리 (A)가 (B)보다 더 적절한 결과임을 선택할 수 있게 된다. 이와 같은 적절한 어휘의 조합에 대한 대표적인 예제가 용언의 격틀이나 하위범주화 정보일 것이다. 본 논문에서는 이와 같은 어휘 조합을 중심어 간의 공기정보라고 한다. 이를 좀더 자세히 기술해 보면, 중심어 간의 공기정보는 꾸며주는 중심어(modifier-head)와 꾸밈을 받는 중심어(modifiee-head), 그리고, 두 중심어 간의 문법적 관계의 세 가지 정보로써 표현된다. 즉, (modifier-head, syntactic-relationship, modifiee-head)의 형태를 취하게 되며, (그림 2)에서 트리 (A)의 경우에는 중심어 간의 공기정보를 ('시험', jco(목적격), '합격하다')의 형태로 표현할 수 있다.



(A)



(B)

(그림 2) 예제 문장에 대한 구조적 중의성

5. 중심어 간의 공기정보를 이용하는 확률 모델

본 절에서는 구조적 중의성을 해결하기 위한 확률 모델을 제안하고자 한다.

우선 $P(W_1^n, T)$를 문장 W_1^n과 그에 해당하는 구문 트리 T의 발생 확률이라고 할때, 문장확률은 식 (4)와 같이 계산될 수 있다.

$$P(W_1^n) = \sum_{T \in P(W_1^n)} P(W_1^n, T) \tag{4}$$

여기서 $P(W_1^n)$는 문장 W_1^n에 대한 가능한 모든 구문 트리의 집합이다. 구문 분석의 확률 모델은 $P(W_1^n, T)$값을 최대화시키는 구문 트리 T를 찾음으로써 구문 분석의 중의성을 해결할 수 있다. 가장 간단한 확률적 모델은 단순한 확률 구문 규칙(PCFG)을 사용하는 것이다. 여기서 구문 트리의 확률값은 그 구문 트리를 형성하는 모든 구문 규칙의 확률값의 곱으로써 표현된다. 즉, 식 (5)로 표현할 수 있다.

$$P(W_1^n, T) = \prod_{rule \in T} P(rule) \tag{5}$$

식 (5)의 가장 기본적인 확률 모델은 우선, 좌우 문맥 정보를 규칙의 조건부에 첨가시킴으로써 확장되어질 수 있다. 우리는 선행 연구 결과로부터 한국어 구문 분석시에 구문 규칙의 좌우 문맥 정보가 구조적 중의성 해소의 정확도를 향상시키는 데 많은 기여를 할 수 있음을 이미 입증한 바 있다 [5, 7]. 좌우 문맥 정보를 첨가시킨 확률 모델은 식 (6)으로 표현할 수 있다.

$$P(W_1^n, T) = \prod_{rule \in T} P(rule \mid t_l, t_r) \tag{6}$$

식 (6)은 각각의 규칙 타입에 따라 식 (7)과 같이 다시 기술될 수 있다.

$$P(W_1^n, T) = \prod_{rule \in T} \begin{cases} P(A \rightarrow B + \tau \mid t_l, t_r) & \text{if rule in TYPE 1} \\ P(A \rightarrow B + \gamma\, C \mid t_l, t_r) & \text{if rule in TYPE 2} \\ P(A \rightarrow A_1 + \gamma' A_2 + \gamma' \cdots A_n \mid t_l, t_r) & \text{if rule in TYPE 3} \end{cases} \tag{7}$$

여기서 t_l과 t_r은 구절 A가 차지하는 문장 범위의 왼쪽과 오른쪽의 품사 정보가 된다. 이제, 이와 같은 기본 모델로부터 중심어 간의 공기정보를 포함할 수 있도록 모델을 확장해 보도록 한다. 우선, 입력 문장에 대한 구문 트리는 구문 규칙의 조합뿐만 아니라, 그 구문 트리내에서 발생되는

중심어 간의 공기정보로써 표현될 수 있다고 하자. 그렇게 되면, 구문 트리의 확률값은 그 구문 트리에서 발생하는 각 규칙의 확률값과 또한, 그 구문 트리내에서 발생하는 중심어 간의 공기정보의 선호도의 곱으로 표현될 수 있을 것이다. 규칙 형태 $A \rightarrow B + \gamma\ C$에서 발생하는 중심어 간의 공기정보에 대한 확률적 중요도는 조건 확률 $P(B^h \mid \gamma, C^h)$에 의해서 측정할 수 있으며, 이는 식 (8)과 같이 추정한다.

$$P(B^h \mid \gamma, C^h) = \frac{F(B^h, \gamma, C^h)}{F(\gamma, C^h)} \tag{8}$$

여기서 $F(\cdot)$는 학습코퍼스에서 발생하는 빈도수를 의미한다. 이와같은 중심어 간의 공기정보를 추가한 최종적인 모델은 식 (9)와 같다.

$$P(W_1^n, T) = \prod_{rule \in T} \begin{cases} P(A \rightarrow B + \tau \mid t_l, t_r) & \text{if rule in TYPE 1} \\ P(A \rightarrow B + \gamma\ C \mid t_l, t_r) \cdot P(B^h \mid \gamma, C^h) & \text{if rule in TYPE 2} \\ P(A \rightarrow A_1 + \gamma' A_2 + \gamma'' \cdots A_n \mid t_l, t_r) \cdot \Pi_{i=1}^{n-1} P(A_i^h \mid \gamma', A_n^h) & \text{if rule in TYPE 3} \end{cases} \tag{9}$$

식 (9)를 이용하여 (그림 1)에 대한 구문 트리의 확률값을 다음과 같이 구할 수 있다.

$$P(W_1^n, T) = P(S \rightarrow ADJP + ef + sf \mid bos, eos)$$

- $P(ADJP \rightarrow NP + jxt\ paa \mid bos, ef)$
- $P(\text{합격하} \mid jxt, \text{어렵})$
- $P(NP \rightarrow VP + etn \mid bos, jxt)$
- $P(VP \rightarrow NP + jcs\ VP \mid bos, etn)$
- $P(\text{쥬드} \mid jcs, \text{합격하})$
- $P(NP \rightarrow nq + jcj\ nq + jcj\ nq \mid bos, jcs)$
- $P(\text{쥰} \mid jcj, \text{쥬드}) \cdot P(\text{수잔} \mid jcj, \text{쥬드})$
- $P(VP \rightarrow NP + jco\ pvg \mid jcs, etn)$
- $P(\text{시험} \mid jco, \text{합격하})$
- $P(NP \rightarrow ncn \mid jcs, jco)$

6. 확률 추정 및 평탄화

구문 규칙의 확률값과 중심어 간의 공기정보의 확률값을 추론하는데는 MLE(Maximum Likelihood Estimation) 방법이 사용되었다. 문맥 정보를 지니고 있는 문법 규칙 확률은 자료 부족(data sparseness)으로 인하여 정확한 확률값 추정이 어렵다. 본 논문에서는 back-off[24]을 이용하여 식 (10)과 같이 문맥 정보를 지니고 있는 구문 규칙에 대한 평탄화(smoothing) 작업을 수행하였다.

$$\begin{aligned} &IF\ F(A \rightarrow \alpha, t_l, t_2) > K \\ &P(A \rightarrow \alpha \mid t_l, t_r) = \frac{F(A \rightarrow \alpha, t_l, t_r)}{\sum_i F(A \rightarrow \alpha_i, t_l, r_r)} \\ &Else\ IF\ 0 < F(A \rightarrow \alpha, t_1, t_2) \le K \\ &P(A \rightarrow \alpha \mid t_l, t_r) = d_c \times \frac{F(A \rightarrow \alpha, t_l, t_r)}{\sum_i F(A \rightarrow \alpha_i, t_l, r_r)} \\ &Else \\ &P(A \rightarrow \alpha \mid t_l, t_r) = Q(t_l, t_r) \times (\lambda_1 \cdot P(A \rightarrow \alpha \mid t_l,) \\ &+ \lambda_1 \cdot P(A \rightarrow \alpha \mid , t_r) + \lambda_2 \cdot P(A \rightarrow \alpha)) \end{aligned} \tag{10}$$

공기정보는 어휘정보로 구성되어 있기 때문에 추정해야 할 파라미터의 개수가 매우 많다. 그러므로, 학습 데이타에서 발생하지 않는 공기정보가 상당히 많게 되고, 결과적으로 데이타 부족으로 인한 심각한 문제에 봉착하게 된다. 본 연구에서는 중심어의 정보를 각각 어휘와 품사 정보(part-of-speech)로 표현하고, 어휘 정보가 부족할 경우에는 품사 정보를 이용하여 평탄화를 수행하였다. 공기정보 B^h는 어휘 정보 B_w^h와 품사 정보 B_t^h의 쌍인 (B_w^h / B_t^h)로 표현한다. 다음의 식 (11)은 공기정보의 확률값에 대한 평탄화 작업이다.

$$\begin{aligned} &IF\ F(B_w^h / B_t^h, \gamma, C_w^h / C_t^h) > K \\ &P(B^h \mid \gamma, C^h) = \frac{F(B_w^h / B_t^h, \gamma, C_w^h / C_t^h)}{(\gamma, C_w^h / C_t^h)} \\ &Else\ If\ 0 < F(B_w^h / B_t^h, \gamma, C_w^h / C_t^h) \le K \\ &P(B^h \mid \gamma, C^h) = d_c \times \frac{F(B_w^h / B_t^h, \gamma, C_w^h / C_t^h)}{(\gamma, C_w^h / C_t^h)} \\ &ElseIf\ F(B_t^h, \gamma, C_w^h / C_t^h) + F(B_w^h / B_t^h, \gamma, C_t^h) > 0 \\ &P(B^h \mid \gamma, C^h) = Q_1 \cdot \left(\frac{F(B_t^h, \gamma, C_w^h / C_t^h)}{F(\gamma, C_w^h / C_t^h)} \right) \\ &\qquad + \frac{F(B_w^h / B_t^h, \gamma, C_t^h)}{F(\gamma, C_t^h)} \\ &Else \\ &P(B^h \mid \gamma, C^h) = Q_2 \cdot \frac{F(B_t^h, \gamma, C_t^h)}{F(\gamma, C_t^h)} \end{aligned} \tag{11}$$

7. 실험 및 결과 고찰

7.1 실험 환경

학습 코퍼스는 수동 구문 트리 태깅된 30,000문장(796,449 형태소)으로 구성되어 있으며 학습 문장의 길이는 평균 25.6개의 형태소로 구성되어 있다[6]. 실험 코퍼스는 학습 코퍼스와는 별개의 문장으로, 모두 1,000 문장으로 구성되었으며,

평균 25.3개의 형태소로 구성되었다. 학습 코퍼스로부터 추출한 구문 규칙의 개수는 2,614개이며 좌우 문맥 정보까지 포함하여 추출한 구문 규칙의 개수는 약 36,600개 정도이다. 또한, 224,883개의 서로 다른 중심어 간의 공기정보가 학습 코퍼스로부터 추출되었다.

7.2 성능 평가

구문 분석의 정확도에 대한 평가는 PARSEVAL[17] 평가 기준을 사용하였다. <표 9>는 각 확률 모델에 대한 실험 결과를 제시하고 있다. 여기서 LP는 동일한 구절이름과 동일한 범위를 차지하는 구절에 대한 정확률을 의미하며, LR은 동일한 구절이름과 동일한 범위를 차지하는 구절에 대한 재현율을 의미한다. 좌우 문맥 정보를 사용한 실험 결과 식 (7)은 기본 모델 식 (5)에 비해 매우 좋은 성능을 발휘함을 볼 수 있다. 공기정보를 사용한 실험 결과 식 (9)의 경우, 학습 데이타에 대해서는 괄목할 만한 향상을 보이는 데 비해, 실험 데이타에 대해서는 약 1% 정도의 향상만을 보이고 있다. 이는 어휘 정보를 그대로 사용하기 때문에 발생하게 되는 데이타 부족 문제가 현재 사용하는 매개변수 평탄화 방법에 의해서 많이 해결되지 못하기 때문으로 인식되어진다. 이와 같은 실험 결과로부터 확률 모델의 중요도만큼이나 평탄화 방법도 중요함을 인지할 수 있었고, 학습 데이타의 크기를 증가시키는 작업 또한 지속적으로 진행되어야 할 것이다.

〈표 9〉 구문 분석의 정확도에 대한 실험 결과

모 델		식 (5)	식 (7)	식 (9)
학습 코퍼스	LP	78.42	87.92	96.29
	LR	76.09	87.23	95.53
실험 코퍼스	LP	78.12	83.25	84.46
	LR	75.97	82.86	83.85

8. 결 론

본 논문에서는 어휘 정보를 이용한 한국어의 확률 구문 분석기를 소개하였다. 각 구절마다 중심어를 정의하였으며, 사용된 구문 규칙의 형태에 따라 중심어 간의 공기정보를 정의하였다. 중심어 간의 공기정보는 구문 트리가 태깅된 코퍼스로부터 자동으로 추출할 수 있었다. 구문 분석기가 사용하는 확률 모델은 확률 구문 규칙과 중심어 간의 공기정보를 같이 사용하는 형태이다. 중심어 간의 공기정보는 어휘 정보이기 때문에 심각한 자료 부족 현상이 발생하였다. 자료 부족 현상은 확률 모델의 성능을 저하시키는 가장 중요한 원인이었으며, 이를 어떻게 해결하느냐가 가장 중요한 문제중의 하나였다. 어휘 정보는 구조적 중의성 해소에 매우 유용한 정보임은 확실하다. 그러나 어휘 정보가 갖고 있는 근본적인 문제-자료 부족 현상-는 어휘 정보 사용을 어렵게 만든다. 어휘 정보가 실제 구조적 중의성을 해결하는데 실질적인 도움이 되기 위해서는 자료 부족 문제에 대한 연구가 선행되어야 할 것이다.

참 고 문 헌

[1] 김재훈, "부분 구문분석 방법론", 정보처리학회지, 제7권 제6호, pp.83-96, 2000.

[2] 박성배, 장병탁, "최대 엔트로피 모델을 이용한 텍스트 단위화 학습", 제13회 한글 및 한국어 정보처리학술대회논문집, pp. 130-138, 2001.

[3] 송영빈, 채영숙, 박용일, 이정민, 설가영, 황혜리, 한나리, 최기선. "동사의 애매성 해소를 위한 구문의미사전의 구축", 한글 및 한국어 정보처리학술대회, pp.280-287, 1999.

[4] 신승은, 서영훈, "부사 정보를 이용한 한국어 구조 중의성 해소", 한글 및 한국어 정보처리학술대회, pp.110-115, 2000.

[5] 이공주, 김재훈, 김길창, "제한된 형태의 구구조 문법에 기반한 한국어 구문 분석", 정보과학회논문지(B), 제25권 제4호, pp. 722-732, 1998.

[6] 이공주, 김재훈, 최기선, 김길창, "구문 트리 부착 코퍼스 구축을 위한 한국어 구문 태그", 인지과학, 제7권 제4호, pp.7-24, 1996.

[7] 이공주, 김재훈, 김길창. "한국어 구구조 문법을 기반으로 하는 확률적 구문 분석", 한국정보과학회 가을학술발표논문집, pp. 557-560, 1996.

[8] 이수광, 옥철영, "확률적 문법규칙에 기반한 국어사전의 뜻풀이말 구문분석기", 정보과학회논문지, 제28권 제5호, pp.448-460, 2001.

[9] 이수선, 박현재, 우요섭, "한국어 분석의 중의성 해소를 위한 하위범주화 사전 구축", 한글 및 한국어 정보처리학술대회, pp. 257-264, 1999.

[10] 정석원, 박의규, 나동렬, 윤준태, "격관계와 상호정보를 이용한 한국어 의존 파서", 제13회 한글 및 한국어 정보처리학술대회논문집, pp.450-456, 2001.

[11] 조형준, 박종철, "결합범주문법과 구문분석", 한글 및 한국어 정보처리학술대회, pp.223-230, 1999.

[12] 최용석, 이주호, 최기선, "격틀 자동구축과 격틀평가 방법에 관한 연구", 한글 및 한국어 정보처리학술대회, pp.272-279, 1999.

[13] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.

[14] D. Hindle and M. Rooth, "Structural ambiguity and lexical relations," *Computational Linguistics*, Vol.19, No.1, pp. 103-120, 1993.

[15] David Crystal. A Dictionary of Linguistics and Phonetics. Basil Blackwell, 1985.

[16] D. M. Magerman, "Statistical decision-tree models for pars-

ing," *Proc. of the 33rd Annual Meeting of the Assoc. for Computational Linguistics (ACL-95)*, pp.276-283, 1995.

[17] E. Black, S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. "A procedure for quantitatively comparing the syntactic coverage of English grammars," *Proceedings of Fourth DARPA Speech and Natural Language Workshop*, pp.306-311, 1991.

[18] Eugene Charniak. Parsing with context-free grammar and word statistics. Technical Report CS-95-28, Dept. of Computer Science, Brown Univ., 1995.

[19] E. Black, F. Jelinek, J. Lafferty, D. M. Magerman, R. Mercer, and S. Roukos. "Towards history-based grammars : Using richer models for probabilistic parsing," *Proc. of the 31st Annual Meeting of the Assoc. for Computational Linguistics (ACL-93)*, pp.31-37, 1993.

[20] J. Eisner, "Bilexical grammars and a cubic-time probabilistic parser," *Proceedings workshop on Parsing Technologies*, pp.54-56, 1997.

[21] K. J. Seo, K. C. Nam, and K. S. Choi, "A probabilistic model of the dependency parse for the variable-word-order languages by using ascending dependency," *Computer Processing of Oriental langauge*, Vol.12, No.3. pp.309-323, 1999.

[22] M. J. Collins. "A new statistical parser based on bigram lexical dependencies," *Proc. of the 34th Annual Meeting of the Assoc. for Computational Linguistics (ACL-96)*, pp.184-191, 1996.

[23] S. F. Chen and J. Goodman, An Empirical Study of Smoothing Techniques for Language Modeling, TR-10-98, Computer Science Group Harvard University Cambridge, Massachusetts, 1998.

[24] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-35, pp.400-401, 1987.

이 공 주

e-mail : kjoolee@microsoft.com

1992년 서강대학교 전자계산학과(학사)

1994년 한국과학기술원 전산학과(공학석사)

1998년 한국과학기술원 전산학과(공학박사)

1998년~현재 (주)한국마이크로소프트 연구원

관심분야 : 자연언어처리, 자연어인터페이스, 기계번역, 정보검색

김 재 훈

e-mail : jhoon@mail.hhu.ac.kr

1986년 계명대학교 전자계산학과(학사)

1988년 한국과학기술원 전산학과(공학석사)

1996년 한국과학기술원 전산학과(공학박사)

1988년~1997년 한국전자통신연구원, 선임연구원

1997년~1999년 한국해양대학교, 컴퓨터공학과, 전임강사

2000년~2002년 한국과학기술원 첨단정보기술연구센터 연구원

2001년~2002년 USC, Information Sciences Institute 방문연구원

1999년~현재 한국해양대학교 컴퓨터공학과 조교수

관심분야 : 자연언어처리, 한국어 정보처리, 정보검색, 정보추출