중심어간의 공기 정보와 구문 규칙을 기반으로 한 확률적 한국어 구문 분석*

이 공주○ 김 재훈* 김 길창

한국과학기술원 전산학과

* 한국 해양 대학교

Probabilistic Parsing of Korean Sentences Based on Lexical Co-occurrence and Syntactic Rules

Kong Joo Lee Jae-Hoon Kim* Gil Chang Kim

Department of Computer Science, KAIST

* Korea Martime University

요 약

어휘 정보는 구문 구조의 중의성을 해결하는데 중요한 정보원으로서 작용할 수 있다. 본 논문에서는 입력 문장에 대한 구조적 중의성을 해결하는데 확률 구문 규칙뿐만 아니라, 어휘간에 발생할 수 있는 공기 정보를 사용할 수 있는 확률 모델을 제안한다. 제안된 확률 모델에 대하여 실험 데이타에 대해 평가한 결과 약 84% 정도의 구문 분석 정확도를 얻을 수 있었다.

1. 서론

구문 분석은 입력 문장에 해당하는 구문 구조를 주어진 구문 규칙하에서 찾아내는 작업이다. 잘 알려져 있는 바와 같이, 이와 같은 과정에서 한 개 이상의 구문 구조가 가능할 수 있다. 그렇기 때문에 구문 분석기는 여러 개의 가능한 구문 구조 중, 입력 문장에 가장 적합한 한 개의 구문 구조를 결정할 수 있어야 한다. 즉, 구조적 중의성을 해소할 수 있어야 한다. 최근에는 이와 같은 구조적 중의성 해소에 통계적 접근 방법이 널리 사용되고 있다. 가장 간단한 통계적 접근 방법은 확률 문법을 사용하는 것이다. 확률 문법을 사용한 구조적 중의성 해결은 매우 간단하다. 파스 트리의 확률값은 그 파스에서 사용되어진 확률 문법의 확률값의 곱이며, 이와 같이 얻어진 각 파스 트리의 확률값 중, 가장 높은 값을 지닌 파스 트리가 입력 문장에 대해 가장 적절한 결과 트리로써 선택되어진다. 이와 같은 방법은 그 언어에서 매우 일반적인 구조와 그렇지 않은 구조 사이의 구분만을 해 줄 수 있다.

어휘 정보는 구문 분석의 중의성 해소에서 중요한 정보원으로 작용할 수 있다. 이미 영어권에서는 전치사구 부착과 같은 문제에서 어휘 정보를 사용하고자 하는 시도들이 진행되었다 [8]. 가장 최근에 몇몇의 연구자들이 전치사구 부착 문제뿐 아니라, 구문 분석 전 과정에서 어휘 정보를 사용하고자 하는 시도들을 해오고 있다 [4, 5, 6, 9].

본 논문에서는 어휘 간의 공기 정보를 이용하는 한국어의 확률적 구문 분석기를 제안하고자 한다. 이와 같은 어휘 간의 공기 정보는 동사의 격틀 정보나 하위범주화와 유사한 형태의 정보이다. 다음의 예제를 살펴보자.

(예 1) "여기에 스케이트가 활주하기 쉽다는 이유가 있다."

(예 1)에서, '여기에'를 성분으로 갖을 수 있는 용언은 '활주하다', '쉽다', 또는 '있다'가 가능하다. 즉, '여기에'가 어떠한 용언과 결합하느냐에 따라 구조적 중의성이 발생한다. 이 때, 이와 같은 구조적 중의성은 각각의 용언 '활주하다', '쉽다', '있다'의 하위범주

*본 연구는 과학재단 특정 기초 연구 과제에 의해 지원되었음.

표 1: 기능어의 분류

기준		
구절 간 관계 명시	격조사	주격조사(jcs), 목적격조사(jco), 보격조사(jcc),부사격조사(jca), 관형격조사(jcm), 공동격조사(jct), 인용격조사(jcr), 접속격조사(jcj), 통용보조사(jxc)
	어미	대등적 연결어미(ecc), 종속적 연결어미(ecs), 관형사형 어미(etm)
구절 내 관계 명시	격조사	서술격조사(jp), 호격조사(jcv), 종결보조사(jxf)
	어미	명사형 어미(etn), 선어말 어미(ep), 종결어미(ef)
	접사	명사파생접사(xsn), 동사파생접사(xsv), 형용사파생접사(xsm), 부사파생접사(xsa)

화(subcategorization) 정보로서 '여기에'가 나올 수 있는 정도를 이용하면 쉽게 해결할 수 있다. 본 논문에서는 이와 같은 하위범주화 정보를 단순화하여, 중심어와 중심어간의 공기 정보(head-head co-occurrence)로서 표현하고, 이를 구조적 중의성을 해결하는 기본 정보로서 사용하고자 한다.

본 논문은 2절에서 한국어 구문 분석을 위한 구구조 규칙을 간략히 소개하고, 이를 기반으로 각 규칙에 따르는 중심어(head)와 중심어간의 공기 정보를 추출하는 방법을 3절에서 소개하고자 한다. 또한, 이를 기반으로 확률 규칙과 중심어간의 공기 정보를 사용하는 한국어 구문 분석을 위한 확률 모델을 소개하고, 이를 기반으로 실험 결과를 제시하고 논문의 결론을 맺고자 한다.

2. 한국어 구문 분석을 위한 구구조 문법

한국어 구문 분석을 위한 구구조 문법은 다음과 같은 형태를 취하고 있다. 한국어 기능어를 표 1에 제시된 바와 같이 크게 두 부류로 나누고 이에 따라 규칙의 형태를 다음과 같은 세가지 형태로 정의할 수 있다. 이에 관한 좀더 자세한 설명은 [1]을 참조하기 바랍니다.

- 구절 내 관계 명시 :
  구절의 문법적 성분의 변화를 유발시키거나 속성을 결정하는 기능어들이다. 이와 같은 기능어가 사용된 구절에 대한 구구조 규칙의 형태는 다음과 같다.

  TYPE I : $A \rightarrow B+\tau$

  여기서 τ는 구절 내 관계를 명시해 주는 기능어들과 보조 용언 구절($AUXP$), 그리고 ϵ이 가능하다. 이 규칙이 의미하는 바는 구절 B가 τ에 의해서 그 속성이 결정되거나 기능이 변화함을 의미한다. 이에 속하는 규칙으로는 다음과 같은 것들이 가능하다.

  $$VP \rightarrow NP+jp$$
  $$NP \rightarrow VP+etn$$
  $$NP \rightarrow ncn+xsn$$

- 구절 간 관계 명시 :
  이에 속하는 기능어들은 문법적 성분의 변화와 더불어 두 구성 성분 간의 문법적 관계를 명시하는 역할을 담당한다. 구절 간 관계를 명시하는 기능어가 사용된 구절에 대한 구구조 규칙의 형태는 다음과 같은 두 가지가 가능하다.

  TYPE II : $A \rightarrow B+\gamma\ C$

  여기서 γ는 구절 간 관계를 명시해 주는 기능어들 중에서 병렬을 표현하는 접속격 조사(jcj)와 대등적 연결어미(ecc)를 제외한 기능어들이 가능하다. 또한, 격조사의 생략이 빈번하므로 ϵ도 가능하다. 이 규칙이 의미하는 바는 구절 B가 구절 C와 γ의 문법적 관계를 형성함을 의미한다. 이에 속하는 규칙으로는 다음과 같은 것들이 가능하다.

  $$VP \rightarrow NP+jcs\ \ VP$$
  $$NP \rightarrow VP+etm\ \ NP$$
  $$VP \rightarrow VP+ecs\ \ VP$$

  TYPE III : $A \rightarrow A_1+\gamma'\ \ A_2+\gamma' \ldots A_n$

  여기서 γ'는 병렬 구조를 표현하는 접속격 조사와 대등적 연결어미, 그리고 나열을 표현하는 쉼표(sp)와 단어 접속 부사(maj)가 가능하다. 구절 간 관계를 명시하는 기능어들 중에서 병렬 형

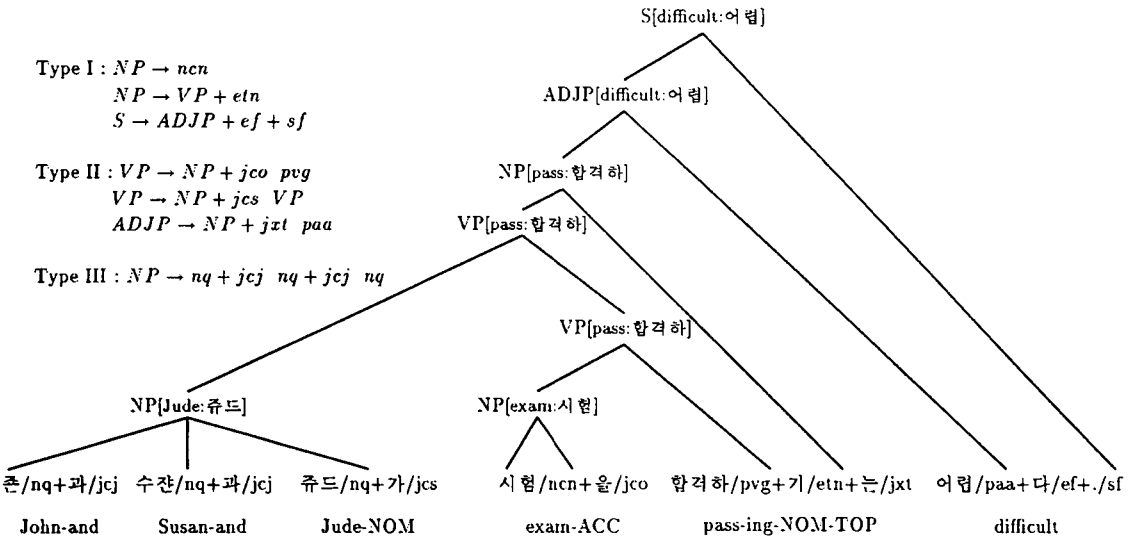

그림 1: 예제 문장과 그에 대한 구문 트리(트리에서 각 구절 옆의 괄호 안의 어휘는 그 구절의 중심어를 의미한다.)

태의 규칙을 따로이 나누어 설정한 것은 TYPE II와는 달리 이에 속하는 규칙의 오른편(RHS)에는 여러 개의 구절이 올 수 있기 때문이다. 이 규칙 형태는 주로 병렬 형태의 구조를 표현하며, 이에 속하는 규칙으로는 다음과 같은 것들이 가능하다.

$$VP \rightarrow VP{+}ecc\ \ VP{+}ecc\ \ VP$$
$$NP \rightarrow NP{+}jcj\ \ NP{+}jcj\ \ NP$$
$$NP \rightarrow ncn{+}sp\ \ ncn{+}sp\ \ ncn$$

그림 1은 예제 문장에 대하여, 앞에서 언급한 형태의 규칙들로 분석한 결과 구문 트리이다.

3. 중심어간의 공기 정보를 이용한 한국어 구문 분석

본 절에서는 용언의 격틀이나 하위범주화 정보와 유사한 중심어간의 공기 정보를 소개하고, 이를 이용한 한국어 확률적 구문 분석을 제안한다.

3.1 중심어간의 공기 정보

임의의 구절(phrase)의 중심어(head)는 그 구절을 대표할 수 있는 단어이다 [7]. 다시 말하면, 구절의 중심어는 그 구절에서 가장 중요한 단어라고 할 수 있다. 예를 들어, 명사구절(NP)의 중심어는 가장 중심 명사(noun)가 될 것이며, 동사구절(VP)의 중심어는 주요 동사가 그 구절의 중심어가 될 것이다.

그림 2는 동일한 문장에 대한 구조적 중의성을 보여주고 있다. 그림에서 괄호 안의 단어는 그 구절의 중심어를 의미한다. 명사구절 $NP_{1,2}$의 중심어는 '존'이며, 동사구절 $VP_{1,3}$의 중심어는 '합격하다'가 된다. 구문 트리 (A)에서, $VP_{0,3}$의 중심어는 $VP_{1,3}$의 중심어와 동일하다. 이는, 구문 규칙 '$VP_{0,3} \rightarrow NP_{0,1} + jco\ \ VP_{1,3}$'에 의해 중심어가 규칙의 RHS로부터 LHS로 전달되기 때문이다. 중심어에 해당하는 어휘 정보가 구조적 중의성 해소에 도움이 되는 과정을 살펴보도록 하자.

그림 2에서 트리 (A)의 경우, 구절 $NP_{0,1}$와 $VP_{1,3}$가 결합하여 구절 $VP_{0,3}$을 형성한다. 이와 같은 구절 결합으로부터 목적어 '시험'의 용언은 '합격하다'임을 알 수 있다. 반면에, 트리 (B)의 경우에는 구절 $NP_{0,1}$와 $VP_{1,4}$의 결합으로부터, 목적어 '시험'의 용언은 '졸업하다'가 됨을 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)

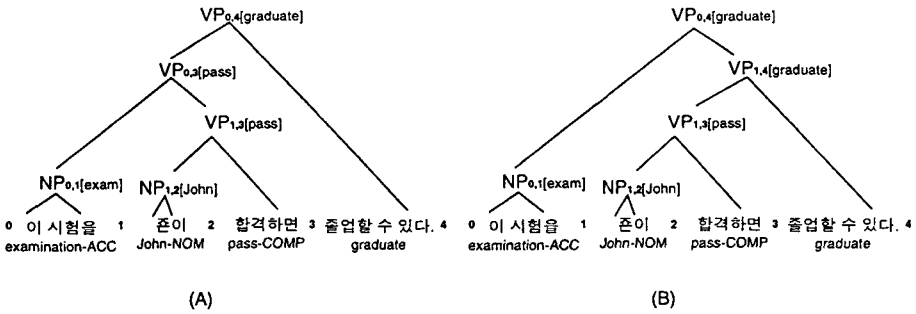

그림 2: 예제 문장에 대한 두 개의 가능한 구문 트리; 논터미널 옆의 숫자는 설명의 편의를 위하여 첨가한 것이다.

보다 더 적절한 결과임을 선택할 수 있게 된다. 이와 같은 적절한 어휘의 조합에 대한 대표적인 예제가 단순화된 용언의 격틀이나 하위범주화 정보와 유사할 것이다. 본 논문에서 이와 같은 어휘 조합을 **중심어간의 공기 정보**라고 한다. 이를 좀더 자세히 기술해보면, 중심어간의 공기 정보는 꾸며주는 중심어(modifier-head)와 꾸밈을 받는 중심어(modifiee-head), 그리고, 두 중심어간의 문법적 관계의 세가지 정보로써 표현된다. 즉, (modifier-head, syntactic-relationship, modifiee-head)의 형태를 취하게 되며, 그림 2에서 트리 (A)의 경우에는 중심어간의 공기 정보를 ('시험', 목적격, '합격하다')의 형태로 표현할 수 있다.

본 논문에서 사용하는 구문 규칙은 2절에서 언급한 바와 같이 세가지 형태의 제한된 형태를 취하게 된다. 이와 같은 제한된 형태의 규칙을 사용하기 때문에 각 규칙의 형태에 따라, 규칙의 중심어와 그 규칙 내에서 발생하는 중심어간의 공기 정보를 자동으로 정의할 수 있다. 표 2는 각 규칙의 형태에 따라 중심어가 되는 구절과 중심어간의 공기 정보를 정의하고 있다. 구절 A의 중심어를 A^h로 표현하였다. TYPE I의 규칙에서는 구절 B의 중심어 자체가 구절 A의 중심어가 되며, 이러한 형태의 규칙에서는 중심어간의 공기 정보가 발생하지 않는다. 이는 TYPE I의 규칙은 다른 구절과의 관계를 명시하는 것이 아니고 단지 구절 자체의 문법적 성질이나 속성의 변화만을 유발하기 때문이다. TYPE II 규칙의 경우에는 구절 C의 중심어가 구절 A의 중심어가 되며 이때에는 $(B^h,\ \gamma,\ C^h)$와 같은 중심어간의 공기 정보가 발생한다. 이는 중심어 B^h와 중심어 C^h가 문법적 관계 γ에 의해서 발생함을 의미한다. TYPE III의 규칙은 규칙의 RHS에 두개 이상의 구절이 존재한다는 점만 제외하고는 TYPE II의 규칙과 유사하다. 그렇기 때문에, TYPE III의 규칙에서는 한개 이상의 중심어간의 공기 정보가 발생할 수 있다. TYPE III의 규칙에서 중심어가 되는 구절은 RHS의 가장 오른쪽 구절인 A_n이 되며, RHS의 나머지 n-1개의 구절들과 A_n구절과의 공기 정보가 발생하게 된다. 표 3은 그림 1에서 발생 가능한 구문 규칙과 이에 해당하는 구절의 중심어, 그리고 중심어간의 공기 정보를 보여주고 있다.

3.2 중심어간의 공기 정보를 이용한 확률적 구문 분석 모델

본 절에서는 구조적 중의성 해결의 성능을 향상시키기 위하여 중심어간의 공기 정보를 포함할 수 있는 확률적 구문 분석 모델을 소개하고자 한다.

우선, $P(W_1^n, T)$를 문장 W_1^n와 그에 해당하는 구문 트리 T의 발생 확률이라고 할 때, 문장 확률 $P(W_1^n)$는 $P(W_1^n) = \sum_{T \in \mathcal{P}(W_1^n)} P(W_1^n, T)$와 같이 계산되어질 수 있다. 여기서, $\mathcal{P}(W_1^n)$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, 다음과 같은 식으로 표현할 수 있다.

$$P(W_1^n, T) = \prod_{rule \in T} P(rule). \qquad (1)$$

식 1의 가장 기본적인 확률 모델은 우선, 좌우 문맥 정보를 규칙의 조건부에 첨가시킴으로써 확장되어질 수 있

표 2: 각 규칙의 형태에 따르는 중심어 구절과 중심어간의 공기 정보

규칙의 형태	중심어 구절	중심어간의 공기 정보
TYPE I $A \rightarrow B+\tau$	B	-
TYPE II $A \rightarrow B+\gamma\ C$	C	(B^h, γ, C^h)
TYPE III $A \rightarrow A_1+\gamma'\ A_2+\gamma' \cdots A_n$	A_n	(A_1^h, γ', A_n^h) (A_2^h, γ', A_n^h) ..... $(A_{n-1}^h, \gamma', A_n^h)$

다. 우리는 선행 연구 결과로부터 한국어 구문 분석 시에 구문 규칙의 좌우 문맥 정보가 구조적 중의성 해소의 정확도를 향상시키는 데 많은 기여를 할 수 있음을 이미 입증한 바 있다 [2]. 좌우 문맥 정보를 첨가시킨 확률 모델은 다음과 같은 수식으로 표현할 수 있다.

$$P(W_1^n, T) = \prod_{rule \in T} P(rule | t_l, t_r) \qquad (2)$$

$$= \prod_{rule \in T} \begin{cases} P(A \rightarrow B+\tau | t_l, t_r) & \text{if } rule \in \text{Type I} \\ P(A \rightarrow B+\gamma\ C | t_l, t_r) & \text{if } rule \in \text{Type II} \\ P(A \rightarrow A_1+\gamma'\ A_2+\gamma' \cdots A_n | t_l, t_r) & \text{if } rule \in \text{Type III} \end{cases}$$

여기서, t_l과 t_r은 구절 A가 차지하는 문장 범위의 왼쪽과 오른쪽의 품사 정보가 된다. 이제, 이와 같은 기본 모델로부터 중심어간의 공기 정보를 포함할 수 있도록 모델을 확장해 보도록 한다. 우선, 입력 문장에 대한 구문 트리는 구문 규칙의 조합뿐만 아니라, 그 구문 트리 내에서 발생되는 중심어간의 공기 정보로써 표현될 수 있다고 하자. 그렇게 되면, 구문 트리의 확률값은 그 구문 트리에서 발생하는 각 규칙의 확률값과 또한, 그 구문 트리 내에서 발생하는 중심어간의 공기 정보의 선호도의 곱으로 표현될 수 있을 것이다. 규칙 형태 '$A \rightarrow B+\gamma\ C$'에서 발생하는 중심어간의 공기 정보에 대한 확률적 중요도는 조건 확률 $P(B^h|\gamma, C^h)$에 의해서 측정할 수 있으며, 이는 다음과 같이 추정한다.

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

여기서, $F(\cdot)$는 학습코퍼스에서 발생하는 빈도수를 의미한다. 이와 같은 중심어간의 공기 정보를 추가한 최종적인 확률 모델은 다음과 같다.

$$P(W_1^n, T) = \qquad (3)$$

$$\prod_{rule \in T} \begin{cases} P(A \rightarrow B+\tau | t_l, t_r) & \text{if } rule \in \text{Type I} \\ P(A \rightarrow B+\gamma\ C | t_l, t_r) \cdot P(B^h|\gamma, C^h) & \text{if } rule \in \text{Type II} \\ P(A \rightarrow A_1+\gamma' \cdots A_n | t_l, t_r) \cdot \prod_{i=1}^{n-1} P(A_i^h | \gamma', A_n^h) & \text{if } rule \in \text{Type III} \end{cases}$$

수식 3을 이용하면 그림 1에 대한 구문 트리 확률값[1]은 다음과 같이 구할 수 있다.

$$\begin{aligned} P(W_1^n, T) = {} & P(S \rightarrow ADJP+ef+sf | bos, eos) \\ & \cdot P(ADJP \rightarrow NP+jxt\ \ paa | bos, ef) \\ & \cdot P(\text{합격하} | jxt, \text{어렵}) \\ & \cdot P(NP \rightarrow VP+etn | bos, jxt) \\ & \cdot P(VP \rightarrow NP+jcs\ \ VP | bos, etn) \\ & \cdot P(\text{쥬드} | jcs, \text{합격하}) \\ & \cdot P(NP \rightarrow nq+jcj\ \ nq+jcj\ \ nq | bos, jcs) \\ & \cdot P(\text{존} | jcj, \text{쥬드}) \cdot P(\text{수잔} | jcj, \text{쥬드}) \\ & \cdot P(VP \rightarrow NP+jco\ \ pvg | jcs, etn) \\ & \cdot P(\text{시험} | jco, \text{합격하}) \\ & \cdot P(NP \rightarrow ncn | jcs, jco) \end{aligned}$$

[1] bos는 문장의 시작을 의미하며, eos는 문장의 끝을 의미한다.

표 3: 그림 1의 구문 트리에서 발생 가능한 구문 규칙과 이에 따르는 중심어와 공기 정보; 각 규칙의 중심어가 되는 구절을 강조하여 표시하였다.

규칙 형태	규칙	중심어	중심어간의 공기 정보
Type I	$NP \rightarrow$ **ncn**	'시험'	
	$NP \rightarrow$ **VP**$+etn$	'합격하'	
	$S \rightarrow$ **ADJP**$+ef+sf$	'어렵'	
Type II	$VP \rightarrow NP+jco$ **pvg**	'합격하'	('시험', jco, '합격하')
	$VP \rightarrow NP+jcs$ **VP**	'합격하'	('쥬드', jcs, '합격하')
	$ADJP \rightarrow NP+jxt$ **paa**	'어렵'	('합격하', jxt, '어렵')
Type III	$NP \rightarrow nq+jcj$ $nq+jcj$ **nq**	'쥬드'	('존', jcj, '쥬드')
			('수잔', jcj, '쥬드')

4. 예비 실험 및 결과

본 논문에서 수행한 모든 실험은 Sun Ultra1 (UltraSPARC 167MHz)에서 수행되었다. 학습 코퍼스는 수동 구문 트리 태깅된 30,000문장(796,449형태소)으로 구성되어 있으며 학습 문장의 길이는 평균 25.6개의 형태소로 구성되어 있다. 실험 코퍼스는 학습 코퍼스와는 별개의 문장으로, 모두 1,000 문장으로 구성되었으며, 평균 25.3개의 형태소로 구성되었다. 학습 코퍼스로부터 추출한 구문 규칙의 갯수는 2,614개이며 좌우 문맥 정보까지 포함하여 추출한 구문 규칙의 갯수는 약 36,600개 정도이다. 또한, 224,883개의 서로 다른 중심어간의 공기 정보가 학습 코퍼스로부터 추출되었다. 구문 규칙의 확률값과 중심어간의 공기 정보의 확률값을 추론하는 데는 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 [10]을 이용하여 다음과 같이 문맥 정보를 지니고 있는 구문 규칙에 대한 평탄화(smoothing) 작업을 수행하였다.

If $F(A \rightarrow \alpha, t_l, t_r) > K$

$$\tilde{P}(A \rightarrow \alpha \mid t_l, t_r) = \frac{F(A \rightarrow \alpha, t_l, t_r)}{\sum_i F(A \rightarrow \alpha_i, t_l, t_r)}$$

Else If $0 < F(A \rightarrow \alpha, t_l, t_r) \leq K$

$$\tilde{P}(A \rightarrow \alpha \mid t_l, t_r) = d_c \times \frac{F(A \rightarrow \alpha, t_l, t_r)}{\sum_i F(A \rightarrow \alpha_i, t_l, t_r)}$$

Else

$$\tilde{P}(A \rightarrow \alpha \mid t_l, t_r) =$$
$$Q_{(t_l,t_r)} \times (\lambda_1 \cdot P(A \rightarrow \alpha \mid t_l, \_) + \lambda_1 \cdot P(A \rightarrow \alpha \mid \_, t_r) + \lambda_2 \cdot P(A \rightarrow \alpha))$$

공기 정보는 어휘정보로 구성되어 있기 때문에 추정해야 할 파라미터의 갯수가 매우 많다. 그러므로, 학습 데이타에서 발생하지 않는 공기 정보가 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. 공기 정보 B^h는 어휘 정보 B^h_w와 품사 정보 B^h_t의 쌍인 (B^h_w/B^h_t)로 표현한다. 다음은 공기 정보의 확률값에 대한 평탄화 작업이다.

If $F(B^h_w/B^h_t, \gamma, C^h_w/C^h_t) > K$

$$\tilde{P}(B^h|\gamma, C^h) = \frac{F(B^h_w/B^h_t, \gamma, C^h_w/C^h_t)}{F(\gamma, C^h_w/C^h_t)}$$

Else If $0 < F(B^h_w/B^h_t, \gamma, C^h_w/C^h_t) \leq K$

$$\tilde{P}(B^h|\gamma, C^h) = d_c \times \frac{F(B^h_w/B^h_t, \gamma, C^h_w/C^h_t)}{F(\gamma, C^h_w/C^h_t)}$$

Else If $F(B^h_t, \gamma, C^h_w/C^h_t) + F(B^h_w/B^h_t, \gamma, C^h_t) > 0$

$$\tilde{P}(B^h|\gamma, C^h) = Q_1 \cdot (\lambda_1 \cdot \frac{F(B^h_t, \gamma, C^h_w/C^h_t)}{F(\gamma, C^h_w/C^h_t)} + \lambda_2 \cdot \frac{F(B^h_w/B^h_t, \gamma, C^h_t)}{F(\gamma, C^h_t)})$$

Else

$$\tilde{P}(B^h|\gamma, C^h) = Q_2 \cdot \frac{F(B^h_t, \gamma, C^h_t)}{F(\gamma, C^h_t)}$$

구문 분석의 정확도에 대한 평가는 PARSEVAL [3] 평가 기준을 사용하였다.

표 4는 각 확률 모델에 대한 실험 결과를 제시하고 있다. 좌우 문맥 정보를 사용한 실험 결과는 기본 모델에 비해 매우 좋은 성능을 발휘함을 볼 수 있다. 공기 정보를

표 4: 예비 실험과 그에 대한 결과 비교; LP는 동일한 구절 이름과 동일한 범위를 차지하는 구절에 대한 정확률을 의미하며, LR은 동일한 구절 이름과 동일한 범위를 차지하는 구절에 대한 재현율을 의미한다.

		기본 모델 수식 1	수식 2를 사용하는 모델	수식 3을 사용하는 모델
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**

사용한 실험 결과의 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.

5. 결론

본 논문에서는 한국어 구문 분석에 중심어간의 공기 정보를 사용하는 새로운 형태의 구문 분석기를 소개하였다. 잘 알려져 있는 바와 같이, 어휘 정보는 구조적 중의성 해소에 매우 중요한 정보로서 작용할 수 있다. 본 논문에서 사용하는 구문 규칙은 제한된 형태의 구문 형식을 지니고 있기 때문에 각 규칙마다 규칙의 중심어와 공기 정보를 자동으로 정의할 수 있었다. 공기 정보는 (중심어, 구문적관계, 중심어)의 세 가지 구성 요소로 구성되어 있으며 이는 용언의 하위범주화 정보나 격틀 정보를 간단한 형태로 표현해 놓은 것으로 볼 수 있다. 중심어 정보로서 어휘 정보를 그대로 사용하고 있기 때문에 심각한 데이타 부족 문제가 발생할 수 있으며 이에 따르는 신중한 파라미터 평탄화 작업이 수행되어야 할 것이다.

참고문헌

[1] 이공주, 김재훈, 최기선, 김길창. 구문 트리 부착 코퍼스 구축을 위한 한국어 구문 태그. 인지과학, 7(4):7-24, 1996.

[2] 이공주, 김재훈, 김길창. 한국어 구구조 문법을 기반으로 하는 확률적 구문 분석. 한국정보과학회 가을 학술발표논문집, pages 557-560, 1996.

[3] 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. In *Proceedings of Fourth DARPA Speech and Natural Language Workshop*, pages 306-311, 1991.

[4] Ezra Black, Fred Jelinek, John Lafferty, David M. Magerman, Robert Mercer, and Salim Roukos. Towards history-based grammars : Using richer models for probabilistic parsing. In *Proc. of the 31st Annual Meeting of the Assoc. for Computational Linguistics(ACL-93)*, pages 31-37, 1993.

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

[6] Michael John Collins. A new statistical parser based on bigram lexical dependencies. In *Proc. of the 34th Annual Meeting of the Assoc. for Computational Linguistics(ACL-96)*, pages 184-191, 1996.

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

[8] D. Hindle and M. Rooth. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103-120, 1993.

[9] David M. Magerman. Statistical decision-tree models for parsing. In *Proc. of the 33rd Annual Meeting of the Assoc. for Computational Linguistics(ACL-95)*, pages 276-283, 1995.

[10] 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, pages 400-401, 1987.