지배가능 경로 문맥을 이용한 의존 구문 분석의 수식 거리 확률 모델*

우연문[0]*, 송영인*, 박소영*, 임해창*, 정후중**
* 고려대학교 컴퓨터학과 자연어처리연구실
** 야후 코리아
* { wooym[0], song, ssoya, rim }@nlp.korea.ac.kr, **hoojung.chung@gmail.com

Modification Distance Model for Korean Dependency Parsing Using Headible Path Contexts

Yeon-Moon Woo[0]* Young-In Song* So-Young Park* Hae-Chang Rim* Hoo-Jung Chung**
* Dept. of Computer Science and Engineering, Korea University
** Yahoo Korea Corp.

요 약

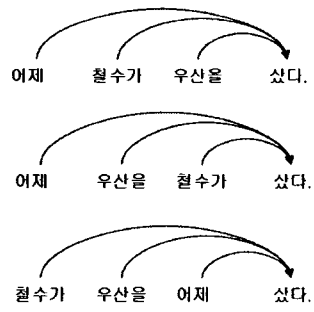
본 논문에서는 한국어 의존 구문 분석을 위한 새로운 확률 모델을 제안한다. 한국어가 자유 어순 언어라 할지라도 지역적 어순은 존재하기 때문에 의존관계를 결정하기 위해 의존하는 두 어절인 의존소와 지배소 사이의 수식 거리가 유용하다는 것은 이미 많은 연구를 통해 밝혀졌다. 본 연구에서는 수식 거리의 정확한 수식 거리의 추정을 위해 지배가능 경로 문맥을 이용한 수식 거리 확률 모델을 제안한다. 제안하는 모델의 구문 분석 성능은 86.9%이며, 기존에 제안된 구문 분석 모델과 비교하여 높은 구문 분석 결과를 보이며, 특히 원거리 의존관계에 대하여 더욱 향상된 성능을 보인다.

1. 서론

구문 분석이란 하나의 입력 문장이 주어졌을 때, 그에 적합한 구문 트리 구조를 만드는 문제이다. 구문 분석은 기계번역, 정보 추출, 질의 응답 등의 많은 자연어 처리 문제에서 중요한 역할을 한다.

의존 구문 분석은 문장을 성분의 조합으로 간주하는 구 구조 구문 분석과는 달리, 문장의 구조를 문장 내의 두 어절 간의 의존관계인 의존문법으로 문장의 구조를 정의하며, 이러한 의존관계에 의해 문장을 분석한다[3]. 의존관계란 지배소와 의존소 사이의 비대칭 관계이다. [그림 1.1]은 의존문법을 사용한 예이다. 여기서 '어제', '철수가', '우산을'은 의존소이고, '샀다'가 지배소이다. 주어, 목적어, 그리고 부사어는 용언을 수식한다. 이 외에도 관형어가 체언을 수식하는 경우, 복합 명사의 의존관계 등 지배소와 의존소간의 다양한 의존관계가 존재한다.

앞서 설명한 바와 같이 의존문법은 어절 사이의 관계만을 고려하기 때문에 한국어와 같은 자유 어순 언어에 유리하다. 영어와 달리 한국어는 조사, 어미와 같은 기능 형태소가 성분을 결정하는 문법적 기능을 하기 때문

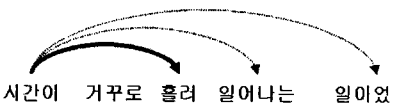


[그림 1.1] 동일한 의미를 가진 문장에 대한 의존 구조

에, 주어와 목적어의 위치가 바뀐다 할지라도 문장의 의미는 동일한 경우가 많다. 이러한 한국어의 성질 때문에 의존문법은 한국어 구문 분석을 하는데 널리 사용된다.

의존문법을 이용한 구문 분석에 대한 많은 연구가 진행되어 왔다. 하지만 만족할 만한 성능을 보여주지 못하고 있다. 그 이유는 구조의 중의성이 존재하기 때문이다. [그림 1.2]은 의존 구문 분석의 중의성을 보여준다.

* 본 연구는 한국과학재단 특정기초연구 (과제번호 R01-2006-11162-0) 지원으로 수행되었음



[그림 1.2] 의존문법의 중의성

용언을 수식할 수 있는 주어 '시간이'는 문장에 존재하는 용언인 '흘러', '일어나는', '일이었다' 등 3개의 지배소 후보를 가지게 된다. 이렇듯 자연어 문장들은 많은 구조적 중의성을 갖고 있기 때문에 이를 해결하기 위한 많은 구문 분석 연구들이 진행되어 왔다.

의존 구문 분석의 구조적 중의성을 해소하기 위해서 많은 통계 기반 방법들이 제안되었다. 통계 기반 방법이란 말뭉치로부터 학습한 통계 값을 분석 결과에 부여함으로써 구문 분석 구조의 중의성을 해소하는 방법이다. 한국어 의존 구문 분석을 위해 원시 말뭉치로부터 추출한 공기정보를 이용하는 방법[9], 단어 또는 어휘의 의존성을 고려한 방법[7] 등이 제안되었다.

통계 기반 의존 구문 분석에서는 파스 트리가 두 어절 간의 의존관계로 구성되고, 각 의존소가 갖는 의존관계에 대한 독립 가정을 사용하면 트리 T의 확률을 (1.1)와 같이 표현할 수 있다.

$$\Pr(T \mid S) \approx \prod_i \Pr(dep_i \mid S) \qquad (1.1)$$

i는 어절 번호를 의미하고, dep_i는 i번째 어절이 갖는 의존관계를 의미한다. 결국 통계적인 의존 구문 분석 문제는 확률 $\Pr(dep_i \mid S)$를 어떻게 정의하느냐에 대한 문제로 해석할 수 있다.

가장 기본적인 방법은 어휘 의존(lexical bigram dependency) 확률, 즉 단어간의 어휘 의존성에 기반하여 확률을 계산한 방법이다. 하지만 두 어절이 유사 어휘 연관성을 가지는 경우에 중의성이 발생하기 때문에, 의존소와 지배소간의 수식 거리를 이용하는 연구들이 제안되었다[2][7].

수식 거리를 자질로 사용하는 것에 대한 자료 부족 문제를 피하기 위해, 최근에는 의존소의 표층 문맥을 자질로 사용하여 수식 거리에 대한 확률을 추정하는 모델이 제안되었다[1]. 하지만 단순히 의존소의 표층 문맥을 사용하기 때문에 주변 문맥을 효과적으로 고려하지 못한다. 이는 의존소가 원거리를 갖는 의존관계일 경우에 정확한 수식 거리를 추정하는 것이 어려운 이유이고, 그 결과 원거리 의존관계의 오류를 발생시킨다.

본 연구에서는 의존소의 수식 거리를 효과적으로 고려하기 위해, 의존소의 '지배가능 경로 문맥'을 사용하여 의존소의 수식 거리를 추정하는 확률 모델을 제안한다. 지배가능 경로(Headible Path)란 하나의 의존소 w_i가 수식할 수 있는 지배소, 즉 w_{i+1}에서 root까지의 경로를 의미한다[4]. 지배가능 경로 문맥은 이러한 지배가능 경로를 의존소의 주변 문맥으로 간주한다.

한국어 의존문법의 경우, 대다수의 문장이 지배소 후위 제약과 투영 제약(No crossing)을 따르기 때문에 지배가능 경로 문맥을 사용함으로써 긴 문맥을 보다 효과적으로 고려하는 것이 가능하다. 제안하는 모델은 어휘를 사용한 경우에 86.9%, 어휘를 사용하지 않은 경우에 85.8%의 정확률을 보이며, 특히 수식 거리가 3이상인 원거리 의존관계에 대해서 [1]보다 각각 0.8%, 1.3%의 성능 향상을 보인다.

본 논문의 2장에서는 제안하는 방법에 대한 관련 연구를 설명한다. 3장에서는 제안하는 방법을 설명하고, 4장에서는 구문 분석 모델을 제안한다. 5장에서는 실험 및 평가, 6장에서는 실험 결과에 대하여 분석한다. 마지막 장에서는 향후 연구 및 본 연구의 결론을 설명한다.

2. 관련 연구

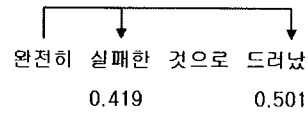
의존 구조의 중의성을 해결하기 위해 그 동안 어휘 의존성을 고려하는 방법, 어휘 의존성과 거리 자질을 고려하는 방법[2][7], 수식 거리 확률을 이용한 방법[1] 등이 제안되었다. 이 장은 제안하는 방법과 관련된 다양한 기존 모델들과 그 문제점에 대해서 설명한다.

가장 기본적인 방법은 어휘 의존(lexical bigram dependency) 확률, 즉 단어간의 어휘 의존성에 기반하여 확률을 계산한 방법이다.

$$\Pr(dep_i \mid S) = \Pr(w_i \rightarrow w_j \mid w_i w_j) \qquad (2.1)$$

어휘 의존 확률은 의존소 w_i와 지배소 w_j가 발생하였을 때 w_i의 지배소가 w_j일 확률을 의미한다. 단어간의 어휘 의존성을 사용함으로써 [그림 1.2]과 같은 구조의 중의성 해결이 가능하다. 의존소 '시간이'는 '흘러'와 의존관계가 발생할 확률이 다른 지배소 후보들과 의존관계가 발생할 확률보다 높기 때문에 올바른 구문 분석이 가능하다.

하지만 이 방법은 의존소와 지배소의 공기 빈도를 이용하기 때문에 자료 부족 문제가 발생한다. 그리고 유사한 어휘 연관성을 가지는 경우 중의성을 해결하지 못하는 경우가 발생한다.



[그림 2.1] 유사한 어휘 연관성의 예

[그림 2.1]의 경우, 의존소 '완전히'의 지배소는 '실패한'이지만, 어휘 의존 확률이 더 높은 '드러났다.'를

지배소로 선택하게 된다. 또한, 어휘 연관성 외의 문맥과 같은 다른 정보를 고려하지 못하는 단점이 있다.

둘째, 어휘 의존성과 수식 거리 자질을 고려한 방법이다. 어휘 의존성만을 고려한 방법의 단점을 해결하기 위해 단어간의 거리를 자질로 사용한 모델을 제안하였다. 두 단어간의 거리가 다른 의존관계를 분리하여 확률을 계산함으로써 유사한 어휘 연관성을 가지는 문제를 해결하려 하였다.

$$\Pr(dep_i \mid S) = \Pr(w_i \rightarrow w_j \mid w_i, w_j, d(w_i, w_j)) \quad (2.2)$$

$$where\ d(w_i, w_j) = \begin{cases} j-i & if\ j-i < K \\ long & else \end{cases}$$

의존소 w_i와 지배소 w_j의 수식 거리 d를 자질로 사용한 결과, [2]는 Penn Wall Street Journal 말뭉치에 대해 85.5%의 F-Score를 보인다.

하지만 수식 거리를 자질로 사용하기 때문에 심각한 자료 부족 문제가 발생한다. 또한 [그림 1.1]에서처럼 성분의 위치를 자유롭게 사용하는 한국어와 같은 자유 어순 언어를 처리하는 데 문제점이 발생한다.

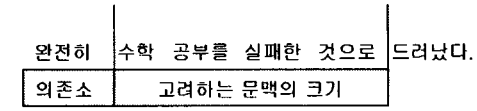
셋째, 수식 거리 확률을 이용한 방법이다. 수식 거리를 자질로 사용하는 것에 대한 자료 부족 문제를 피하기 위해, 수식 거리에 대한 확률 모델이 제안되었다[1]. 두 단어간의 어휘 의존 확률과 의존소의 수식 거리 확률을 분리하여 확률을 계산하였는데, 수식 거리 확률 모델은 의존소의 수식 거리를 의존소 주변의 표층 문맥(surface context) 패턴을 이용하여 수식 거리를 추정한다.

$$\Pr(dep_i \mid S) = \Pr(w_i \xrightarrow{d(w_i,w_j)} w_j \mid S) = \Pr(w_j \rightarrow w_i \mid w_i, w_j, {}_f t_{i-1}, {}_c t_{j+1}) \cdot \Pr(d(w_i, w_j) \mid w_i, t_{i+1}, t_{i+2}, \mathrm{L}) \quad (2.3)$$

$$where\ d(w_i, w_j) = \begin{cases} j-i & if\ j-i < K \\ long & else \end{cases}$$

의존소 w_i의 수식 거리가 d일 확률을 의존소와 의존소의 t_{i+1}, t_{i+2}(w_{i+1}, w_{i+2}의 품사)와 같은 표층 문맥을 이용하여 추정하였다.

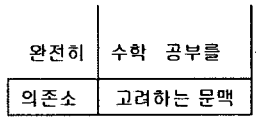
자유 어순 언어라 할지라도 어순을 완전히 무시할 수 없고 지역적 어순은 존재하기 때문에[1][6][8], 의존소의 지역문맥(local context)의 어순을 고려하였고, 수식 거리에 대한 확률을 계산하기 때문에 자료 부족 문제를 완화하였다. 그렇지만 어휘가 아닌 패턴(품사열)을 사용한다 할지라도, 이 역시 긴 문맥을 고려할 경우 자료 부족 문제가 발생한다. [그림 2.2]는 큰 문맥을 자질로 사용하였을 경우의 자료 부족 문제를 보여준다.



Pr(length(완전히) =3 | 완전히 tag(수학) tag(공부를) tag(실패한) tag(것으로)) = ?

[그림 2.2] 문맥의 크기가 클 경우 자료 부족 발생

또한, 단순히 의존소의 표층 문맥을 사용하기 때문에 주변 문맥을 효과적으로 고려하지 못한다. 이는 의존소가 원거리를 갖는 의존관계일 경우 정확한 수식 거리를 추정하는 것이 어려운 이유이고, 그 결과 원거리 의존관계의 오류를 발생시킨다.



Pr(length(완전히) =3 | 완전히 tag(수학) tag(공부를))
= Pr(length(완전히) =5 | 완전히 tag(수학) tag(공부를))

[그림 2.3] 원거리 의존관계 고려 시 동일 확률 부여

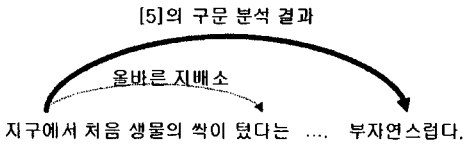
[그림 2.3]와 같이 2개의 지역 문맥을 자질로 사용하였을 경우, 지역 문맥의 크기보다 큰 거리에 대해서 동일 확률을 부여하게 되어 올바른 구문 분석에 실패하게 된다. [그림 2.3]의 어절 '수학'의 유무는 의존소 '완전히'의 의존관계를 파악하는데 영향을 미치지 않는다. 그렇기에 수식 거리를 결정하는 데에 영향을 미치는 효과적인 문맥의 선택이 필요하다.

3. 제안하는 방법

이 장에서는 본 연구에서 제안하는 방법인 지배가능 경로 문맥을 고려한 수식거리 결정에 대해 설명한다. 3.1절에서는 연구 동기를 설명하고, 3.2절에서는 제안하는 방법의 기본 아이디어와 장점에 대해서 설명한다.

3.1 연구 동기

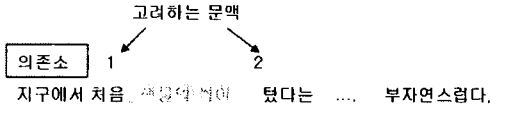
수식 거리를 고려할 때, 의존소의 표층 문맥을 이용한 [1]은 고려하는 지역 문맥이 최대 3이기 때문에 이보다 긴 거리에 대해서 [그림 3.1]와 같은 구문 분석 오류가 발생한다.



[그림 3.1] 표층 문맥을 이용한 수식 거리 모델의 구문 분석 오류

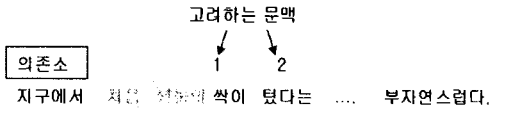
이는 고려하는 지역 문맥의 크기보다 긴 수식 거리의 확률이 동일하기 때문에, 어휘 의존 확률이 큰 '부자연스럽다.'를 지배소로 선택하기 때문이다.

[그림 3.2], [그림 3.3]는 수식 거리 확률을 추정할 때 고려할 문맥을 임의적으로 선택했을 때의 결과이다.



$argmax_d$ Pr(Length(지구에서) = d | 지구에서 tag(처음) tag(텄다는))
= 2

[그림 3.2] « 생물의 싹이 »를 삭제한 문맥을 고려한 수식 거리 확률 모델의 적용



$argmax_d$ Pr(Length(지구에서) = d | 지구에서 tag(싹이) tag(텄다는))
= 3

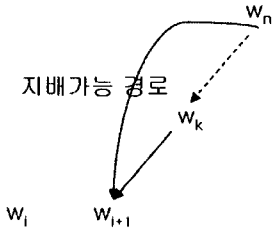
[그림 3.3] « 처음 »을 삭제한 문맥을 고려한 수식 거리 확률 모델의 적용

[그림 3.2], [그림 3.3]와 같이 어떤 문맥을 고려하느냐에 따라 수식 거리 확률이 달라진다. 따라서, 수식 거리를 고려할 때 필요한 문맥을 효과적으로 선택하게 되면, 구문 분석 결과의 성능 향상, 특히 원거리 의존 관계의 성능 향상이 기대 된다. 3.2절에서는 문맥 선택의 방법에 대해서 설명한다.

3.2 지배가능 경로 문맥을 고려한 수식 거리 결정

본 연구에서 제안하는 방법은 의존소의 수식 거리를 결정하기 위해 지배가능 경로 문맥을 자질로 사용함으로써 기존의 구문 분석기에서 해결하지 못한 원거리 의존관계를 해결하자는 것이다.

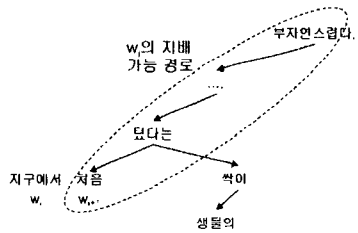
지배가능 경로는 부분 트리에서 의존소가 수식할 수 있는 지배소들의 경로를 의미한다[4]. [그림 3.4]를 통해 w_i의 지배가능 경로는 w_i의 오른쪽 인접 어절 w_{i+1}



[그림 3.4] w_i의 지배가능 경로

에서부터 조상 노드들로 연결되어 문장의 중심어인 w_n까지의 경로인 것을 알 수 있다.

의존소와 그 의존소의 지배가능 경로를 문맥으로써 사용하여 수식 거리를 결정한다. 한국어에서는 의존소의 어휘, 품사가 의존소의 수식 거리에 영향을 미치고, 의존소의 왼쪽 문맥은 수식 거리를 결정하는데 도움을 주지 못한다[1]. 또한, 지배소 후위의 원칙으로 인해 오른쪽 우선 구문 분석을 하기 때문에 의존소의 지배가능 경로를 쉽게 얻어낼 수 있는 장점이 있다[11].



[그림 3.5] « 지구에서 »의 지배가능 경로

[그림 3.5]는 수식 거리를 결정하기 위해 고려하는 지배가능 경로 문맥이다.

지배가능 경로 문맥을 이용하여 수식 거리 확률을 추정하는 것은 다음과 같은 장점이 있다.

첫째, 3.1절에서 언급한 바와 같이, 문맥 제거의 효과가 있다. 지배가능 경로 문맥 외의 어절들은 투영 제약에 따라 실제로 의존소의 지배소가 될 수 없으므로, 의존소의 수식 거리에 대해 영향을 미치지 않는 어절의 제거 효과가 있다.

둘째, 지배가능 경로 문맥은 한국어의 어순을 제약하는데 도움이 된다. 한국어가 비교적 자유 어순 언어에 가깝다 할지라도 지역적 어순 제약은 존재한다[8]. 또한, 의존 파스 트리의 조상(ascendant) 노드들은 어순을 제약하는데 도움이 된다[6].

셋째, 큰 표면 문맥을 고려할 수 있다는 장점이 있다. 큰 지배가능 경로 문맥은 큰 표현 문맥을 압축한 결과로 간주할 수 있다. 이를 통해 원거리를 갖는 의존관계에 대해 분별력 증가를 기대할 수 있다.

다음 장에서는 지배가능경로 문맥을 고려한 수식 거리 확률 모델에 대해 설명한다.

4. 구문 분석 모델

이 장에서는 본 연구에서 제안하는 확률 모델을 설명한다. 4.1절에서는 제안하는 확률 모델을 설명한다. 4.2절에서는 본 연구에서 사용하는 어휘 의존 확률 모델, 4.3절에서는 지배가능 경로 문맥을 이용한 수식 거리 확률 모델을 설명한다.

4.1 확률 모형화

통계 기반의 구문 분석은 주어진 문장 S에 대해서, 모든 가능한 파스 트리 집합 V(T)중 가장 큰 확률을 갖는 파스 트리 T를 찾는 문제이다.

$$T_{best} = \arg\max_{T \in V(T)} \Pr(T \mid S) \tag{4.1}$$

의존 파스 트리는 의존관계로 이루어진 집합이므로, 문장 S는 트리에 있는 모든 의존관계들의 곱으로 나타낼 수 있다.

$$\begin{aligned}
&\Pr(T \mid S) \\
&\quad = \Pr(dep_1 dep_2 .. dep_{|S|-1} \mid S) \\
&\quad = \Pr(dep_{|S|-1} \mid S)\Pr(dep_{|S|-2} \mid dep_{|S|-1} S) \\
&\quad \text{L } \Pr(dep_1 \mid dep_2 dep_3 \text{L } dep_{|S|-1} S) \\
&\quad \approx \prod_{i<|S|} \Pr(dep_i \mid S)
\end{aligned} \tag{4.2}$$

본 연구에서는 의존관계 dep_i가 i번째 어절 w_i, w_i의 왼쪽 인접 어절의 기능 형태소 ${}_f t_{i-1}$, 지배가능 경로 문맥 Φ_i, w_i의 지배소 w_j, w_j의 오른쪽 인접 어절의 내용 형태소 ${}_c t_{j+1}$에 의해서 결정된다고 가정한다.

$$\begin{aligned}
&\prod_{i<|S|} \Pr(dep_i \mid S) \\
&\approx \prod_{i<|S|} \Pr(dep_i \mid w_i, \Phi_i, w_j, {}_f t_{i-1}, {}_c t_{j+1})
\end{aligned} \tag{4.3}$$

본 연구에서 제안하는 파싱 모델은 (4.4)와 같다. 의존관계 dep_i는 어휘 의존 확률을 위한 두 단어 사이의 의존관계와 지배가능 경로를 이용한 수식 거리 확률을 위한 의존소의 수식 거리이다. 이 때 각 조건부 확률에 영향을 미치지 않는 자질들을 제거한다. 이처럼 어휘 의존 확률 모델과 수식 거리 확률 모델로 분리하는 것은 확률 추정 시에 자료 부족 문제를 완화시킨다[1].

$$\begin{aligned}
&\Pr(dep_i \mid S) \\
&= \Pr(w_i \xrightarrow{d(w_i, w_j)} w_j \mid w_i, \Phi_i, w_j, {}_f t_{i-1}, {}_c t_{j+1}) \\
&= \Pr(w_i \rightarrow w_j \mid w_i, w_j, {}_f t_{i-1}, {}_c t_{j+1}) \\
&\quad \Pr(d(w_i, w_j) \mid w_i, \Phi_i)
\end{aligned} \tag{4.4}$$

w_i는 의존소, w_j는 지배소이다. $d(w_i, w_j)$는 지배가능 경로 문맥상의 거리, w_i는 의존소 Φ_i의 지배가능 경로 문맥이다.

4.2 외부 문맥을 이용한 어휘 의존 확률 모델

어휘 의존(lexical bigram dependency) 확률 모델은 단어간의 어휘 의존성에 기반하여 확률을 계산하는 방법이다.

$$\Pr(w_i \rightarrow w_j \mid w_i, w_j, {}_f t_{i-1}, {}_c t_{j+1}) \tag{4.5}$$

어절 w_i와 w_j가 문장에서 공기하였을 때 의존관계가 발생할 확률을 의미한다. ${}_f t_{i-1}$와 ${}_c t_{j+1}$는 두 어절의 외부 문맥의 품사이며, 외부 문맥을 자질로써 사용하는 것은 구문 분석 성능을 올리는데 도움을 준다[10].

4.3 지배가능 경로 문맥을 이용한 수식 거리 확률 모델

지배가능 경로 문맥을 이용한 수식 거리 확률 모델은 수식 거리를 결정하고자 하는 의존소와 해당하는 문맥(지배가능 경로 문맥)이 주어졌을 때 의존소가 어떤 지역을 수식할 것인지의 확률을 계산한다.

$$\begin{aligned}
&\Pr(d(w_i, w_j) \mid w_i, \Phi_i) \\
&= \Pr(d(w_i, w_j) \mid w_i, t_{i+1}, t_{head(i+1)}, \ldots)
\end{aligned} \tag{4.6}$$

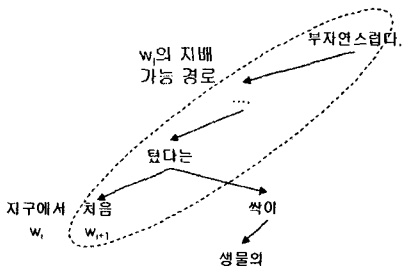
$d(w_i, w_j)$는 의존소 w_i의 지배가능 경로상의 수식 거리를 의미하며, 해당하는 의존소 w_i와 w_i의 지배가능 경로 문맥 $tag_{i+1}, tag_{head(i+1)}, \ldots$에 의해서 결정된다.

$$d(w_i, w_j) = \begin{cases} w_{i+1}\text{부터 } w_j\text{사이에 있는 지배가능 후보의 개수} \\ \quad if\ w_{i+1}\text{부터 } w_j\text{사이에 있는 지배가능 후보의 개수} < K \\ long\ \ else \end{cases} \tag{4.7}$$

K는 고려하는 지역의 개수를 의미하는데, 수식 거리 $d(w_i, w_j)$가 K이상의 원거리 수식 거리일 경우, '원거리 수식 거리'를 의미하는 long을 추정하게 된다. K=2이면 $d(w_i, w_j)$= 1, $d(w_i, w_j)$= long의 2가지 경우가

가능하며, K=3이면 $d(w_i, w_j)$ = 1, $d(w_i, w_j)$ = 2, $d(w_i, w_j)$ = long 이 가능하다.

[그림 4.1]은 제안하는 확률 모델을 적용한 예제이다.



[그림 4.1] 지배가능 경로 문맥을 이용한 수식 거리 확률

의존소 '지구에서'는 실제 수식 거리가 4인 원거리 의존관계임에도 불구하고, 지배가능 경로 문맥에서의 수식 거리는 2이다. 따라서 어절 '텄다는'의 수식 거리 확률 값이 '부자연스럽다.'보다 높아져 올바른 구문 분석 결과를 보인다.

5. 실험 및 평가

이 장에서는 본 연구에서 제안하는 모델의 실험과 성능에 대해 설명한다. 5.1절에서는 실험 환경을 설명한다. 5.2절에서는 구문 분석 실험, 5.3절에서는 기존 모델과의 성능 비교 평가를 보여준다.

5.1 실험 환경

본 연구의 실험은 KAIST 언어자원 구구조 구문 분석 말뭉치[12]로부터 변환한 의존 구문 분석 말뭉치를 사용하였다. 총 31,080 문장으로 구성되어 있으며, 27,694문장을 학습 집합으로, 3,386개를 실험 집합으로 사용하였다. 문장의 평균 어절 수는 각각 12.2개, 12.45개이다.

구문 분석기의 입력 문장은 품사 부착 후, 보조용언 구 청킹을 한 문장이다. 보조용언 구의 청킹은 규칙 기반의 보조용언 구의 전처리기를 사용하였고, 약 97.5%의 정확률을 가진다. 보조용언 구를 전처리한 문장의 평균 어절 수는 학습, 실험 집합에서 각각 11.13, 11.31개이다. 본 실험에서는 후방 빔 탐색 알고리즘[5]를 사용하여 구문 분석을 한다.

5.2 구문 분석 실험

제안한 구문 분석 모델의 평가를 위해 아크-정확률과 아크-재현율을 결합한 F_1-measure와 문장 정확도(Exact-Matching)가 평가 척도로 사용되었다. 평가 척도에 대한 설명은 [그림 5.1]와 같다.

아크 정확률 (Arc Precision, AP)

$$= \frac{\text{구문 분석 파스트리에서 올바른 아크의 수}}{\text{구문 분석 파스트리에서 모든 아크의 수}}$$

아크 재현율 (Arc Recall, AR)

$$= \frac{\text{구문 분석 파스트리에서 올바른 아크의 수}}{\text{정답 파스 트리에서 모든 아크의 수}}$$

$$F_1\text{-measure} = \frac{2 \cdot AP \cdot AR}{AP + AR}$$

$$\text{Exact-Matching} = \frac{\text{정확히 분석된 문장의 수}}{\text{문장의 수}}$$

[그림 5.1] 평가 척도

[표 5.1]은 자질로 사용한 지배가능 경로 문맥의 크기 N과 지역의 개수 K에 따른 구문 분석 모델의 성능이다. N이 3일 때의 F_1-measure가 가장 좋은 것을 볼 수 있다. N이 2일 때 K가 고려하는 문맥보다 클 경우(K=4인 경우), F_1-measure는 올라가나, 문장 정확도는 떨어지는 것을 볼 수 있다.

N	K	평가 척도	
		F_1-measure	Exact Matching
2	3 ({1,2,long})	86.6%	34.6%
2	4({1,2,3,long})	86.7%	34.2%
3	4({1,2,3,long})	86.9%	34.3%

[표 5.1] 문맥의 크기(N), 지역의 개수(K)에 따른 어휘를 사용한 구문 분석기의 성능

[표 5.2]는 어휘를 사용하지 않고 품사만을 사용한 구문 분석 모델의 성능이다. 어휘를 사용한 구문 분석기의 성능과 비슷한 경향을 보인다.

N	K	평가 척도	
		F_1-measure	Exact Matching
2	3	85.4%	32.2%
2	4	85.7%	32.3%
3	4	85.8%	32.3%

[표 5.2] 문맥의 크기(N), 지역의 개수(K)에 따른 어휘를 사용하지 않은 구문 분석기의 성능

어휘를 사용하지 않은 구문 분석기의 경우, 더 적은 확률 추정 데이터가 필요하며, 구문 분석 속도도 월등히 빠르고, 성능 또한 크게 떨어지지 않는 것을 볼 수

있다.

5.3 성능 비교 평가

본 연구에서 제안하는 모델과 관련 연구에서 설명한 구문 분석 모델의 성능 비교를 하였다. 성능 비교에 쓰인 모델은 다음과 같다.

Model1 어휘 의존 확률 모델에 외부 문맥(outer context)를 사용한 모델이다. 제안하는 모델의 베이스 라인(baseline)이 된다.

$$\Pr(dep_i \mid S) = \Pr(w_i \rightarrow w_j \mid w_i, w_j, {}_f t_{i-1}, {}_c t_{j+1}) \quad (5.1)$$

Model2 [7][2]가 제안한 의존하는 단어 사이의 거리를 자질로 사용한 모델이다.

$$\Pr(dep_i \mid S) = \Pr(w_i \rightarrow w_j \mid w_i, w_j, d(w_i, w_j)) \quad (5.2)$$

Model3 [1]에서 제안된 표층 문맥을 사용한 수식 거리 확률을 사용한 모델이다.

$$\Pr(dep_i \mid S) = \Pr(w_i \rightarrow w_j \mid w_i w_j ftag_{i-1} ctag_{j+1}) \cdot \Pr(d(w_i, w_j) \mid w_i t_{i+1} t_{i+2} \mathrm{L}) \quad (5.3)$$

5.3.1 구문 분석 성능 비교

앞서 설명한 비교 모델들은 제안한 모델과 같은 실험 환경을 갖는다. 결과는 [표 5.3]와 같다.

평가 척도	MODEL1	Model2	Model3	Proposed
F_1-measure	79.7%	83.2%	86.7%	86.9%
Exact Matching	26.2%	28.2%	33.9%	34.3%

[표 5.3] 어휘를 사용한 구문 분석기의 성능 비교

수식 거리를 사용한 모든 모델들이 베이스 라인 모델인 MODEL1보다 나은 성능을 보이고 있다. 이를 통해 수식 거리가 의존 구문 분석을 하는데 유용한 정보임을 알 수 있다.

본 연구에서 제안한 수식 거리 모델을 사용한 구문 분석기는 베이스 라인인 MODEL1보다 7%이상의 성능 향상을 보인다. 지배가능 경로 문맥을 사용한 수식 거리 모델이 구문 분석 성능 향상에 큰 영향을 미치는 것을 보인다. 또한 제안하는 모델의 성능은 Model2 보다 3.7%가 높은 것을 보여준다. Model3과는 성능이 비등하지만 역시 미세한 성능 향상을 보인다.

[표 5.4]는 어휘를 사용하지 않은 구문 분석 모델에 대한 성능 비교이다. 앞선 어휘를 사용한 모델들의 성능 비교와 마찬가지로 다른 모델보다 제안하는 모델이 높은 성능을 보임을 알 수 있다.

평가 척도	MODEL1	Model2	Model3	Proposed Model
F_1-measure	76.3%	80.1%	85.6%	85.8%
Exact Matching	22.4%	20.0%	32.2%	32.3%

[표 5.4] 어휘를 사용하지 않은 구문 분석기의 성능 비교

5.3.2 수식 거리 별 성능 비교

제안하는 모델과 유사한 Model3은 구문 분석 결과 역시 제안하는 모델과 크게 차이가 나지 않았다. 하지만 수식 거리를 추정하고자 고려하는 관점이 다르기 때문에 수식 거리 별 성능을 통해 두 모델을 비교하였다.

수식 거리	Model3	Proposed
1	0.972(20932/21538)	0.963(20747/21538)
2	0.759(3303/4351)	0.799(3475/4351)
>2	0.670(6059/9043)	0.678(6121/9043)
Total	0.867(30294/34932)	0.879(30343/34932)

[표 5.4] 수식 거리 별 성능 비교

[표 5.4]를 통해 수식 거리가 1일 때의 정확률은 Model3보다 떨어지지만 수식 거리 2이상의 의존관계의 정확률은 크게 향상된 것을 볼 수 있었다.

6. 분석

이 장에서는 제안하는 모델이 Model3에 비해 성능이 향상되는 이유와 수식거리 별 성능 차이에 대해서 설명한다.

Model3는 표층 문맥을 이용하여 수식 거리를 결정하였고, 제안하는 모델은 지배가능 경로 문맥을 이용하여 수식 거리를 결정하였지만 두 모델간의 성능 차이는 그리 크지 않았다. 이를 밝히기 위해 학습과 실험에 쓰였던 말뭉치를 가지고 다음을 조사 하였다.

한국어의 특징상 수식 거리가 1인 의존소가 많이 존재함을 알 수 있었다. KAIST 언어자원 구문 분석 트리뱅크 말뭉치에서 의존관계 중, 총 64.1%가 수식 거리가 1임을 볼 수 있었다. 또한, 자질의 중복성을 살펴보았다. 고려하는 문맥의 크기를 각각 2로 동일하게 설정한 뒤, 동일한 자질이 전체 자질 중 57.9%를 차지하였다. 둘 중 하나라도 해당하는 경우, 즉 수식 거리가 1이거나 고려한 자질이 동일한 경우는 전체 의존소 중 90.1%를 차지하였다.

이는 두 모델이 자질의 관점은 틀리지만, 한국어의 특성상 의존소의 표층 문맥(surface context)는 본 연구에서 제안하는 지배가능 경로 문맥과 일치할 가능 성이 높음을 알 수 있었고, 대다수의 경우 동일한 구문 분석 결과를 보여줌을 의미한다. 그 결과 제안하는 모

델은 Model3와 마찬가지로 베이스 라인으로 제시된 MODEL1에 비해 큰 성능 향상을 보인다.

하지만 의존소는 지배가능 경로 문맥에서 수식 거리가 결정되므로 의존소의 표층 문맥을 이용하는 것보다 지배가능 문맥을 고려하는 것이 바람직하다. 이는 앞선 성능 비교에서 살펴본 바와 같이 구문 분석 성능의 미세한 향상과 원거리 의존관계의 성능 향상을 통해 알 수 있다.

수식 거리 별 성능에서 수식 거리가 1,2일 때 성능 차이가 나는 이유는 두 모델의 수식 거리의 분포가 다르기 때문이다. 이는 수식 거리의 단위가 다름에서 이유를 찾을 수 있다. Model3는 문장의 실제 수식 거리를 사용하는 반면에, 제안하는 모델의 수식 거리는 지배가능 경로 상의 수식 거리를 사용한다.

수식 거리가 1일 경우, Model3와 제안하는 모델은 의존소의 우측 인접 어절을 수식하는 경우이다. 하지만 수식 거리가 2이상인 경우, 표층 문맥과 지배가능 경로 문맥의 차이가 발생한다. 다음은 두 모델이 수식 거리의 분포이다.

수식 거리	1	2	3	<3
Proposed -지배가능 경로 문맥의 수식 거리	62.2%	18.5%	9.3%	10.0%
Model3 -표층 문맥(실제 문장)의 수식 거리	62.2%	12.6%	6.8%	18.3%

[표 6.1] 각 모델의 수식 거리 분포

자료 부족 문제가 발생한 경우 모델을 backed-off 하기 때문에, 수식 거리 2를 선호할 가능성이 더 상대적으로 더 많다. 이러한 경향이 반영된 결과, [표 5.4]와 같이 문장의 수식 거리 1, 2일 때 구문 분석 성능 차이를 보인다.

7. 결론 및 향후 연구

7.1 향후 연구

본 연구에서 제안하는 수식 거리 모델은 인접 거리 의존관계에 대한 성능에 상대적으로 오류가 많았다. 대부분의 의존관계를 차지하는 인접 거리의 의존관계에 대한 정확도 향상이 필요하다.

또한, 의존소와 지배소의 수식 거리를 정확히 추정하기 위해 지배소의 주변 문맥 사용 방법에 대한 연구가 필요하다.

7.2 결론

본 논문에서는 지배가능 경로 문맥을 이용한 수식 거리확률 모델을 제안하였다. 실험을 통해 제안한 모델의 성능이 베이스라인 모델의 성능을 크게 앞서는 것을 보았으며, 표층 문맥을 이용한 모델과의 비교에서도 미세한 성능 향상을 볼 수 있었다. 특히 원거리 의존관계의 정확도 크게 향상됨을 알 수 있었다.

한국어의 경우 수식 거리가 1인 경우가 많기 때문에 수식 거리 모델 추정 시, 표층 문맥과 지배가능 경로 문맥이 일치하는 경우가 많았다. 이를 6장을 통해서 알아 보았다. 하지만, 의존소가 의존관계를 갖는 지배소는 지배가능 경로 문맥에서 결정되므로, 수식 거리를 결정할 때 지배가능 경로 문맥을 고려하는 것이 바람직하다.

참고 문헌

[1] H.Chung, 'Statistical Korean Dependency Parsing Model based on Surface Contextual Information', 고려대학교 박사학위 논문, 2004

[2] M.Collins, 'A new statistical parser based on bigram lexical dependencies', In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, 1996

[3] M. Covington, 'A dependency parser for variable-word-order languages', Research Report AI-1990-01, University of Georgia, 1990

[4] C.Kim, et al. 'A Right-to-Left Chart Parser for Dependency Grammar using Headible Paths', Proceeding of the 1994 International Conference on Computer Processing of Oriental Language, 1994

[5] S.Sekine,et al.'Backward beam search algorithm for dependency analysis of Japanese', In Proceedings of the 18th International Conference on Computational Linguistics, page 745~760, 2000

[6] K.Seo, et al. 'A Probabilistic model of the dependency parse for the variable-word-order language by using ascending dependency', Computer Processing of Oriental Languages, page 309~322, 1999

[7] 김학수, 서정연, '어휘 의존 정보에 기반한 한국어 통계적 구문분석기', 97년도 정보과학회 인공지능 연구회 춘계 발표 논문집, page 74~90, 1997

[8] 류법모 외 2인, '한국어 파서에서의 지역 의존관계의 이용', 제 8회 한글 및 한국어 정보처리 학술대회, page 464~468, 1996

[9] 윤준태, '공기 관계 기반 어휘 연관도를 이용한 한국어 구문 분석', 연세대학교 박사학위 논문, 1997

[10] 이공주, '언어특성에 기반한 한국어의 확률적 구문 분석', KAIST 박사학위 논문, 1998

[11] 장두성, 최기선, '내부 및 외부 확률을 이용한 의존문법의 비통제 학습', 제 12회 한글 및 한국어 정보처리 학술회의 논문집, 2000

[12] 최기선, 'KAIST 언어자원 v.2001', 2001