부사 정보를 이용한 한국어 구조 중의성 해소

신 승 은†, 서 영 훈††
충북대학교 컴퓨터공학과
† seshin@dcenlp.chungbuk.ac.kr ††yhseo@cbucc.chungbuk.ac.kr

Korean Structural Disambiguation using Adverb Information

Seung-Eun Shin, Young-Hoon Seo
Dept. of Computer Engineering, Chungbuk National University

요 약

자연 언어 처리의 구문 구조 분석에서는 중의성 있는 결과가 많이 생성된다. 이러한 중의성을 해소하는데 어휘정보가 유용하다는 것은 잘 알려져 있으며, 이러한 어휘정보와 이를 이용한 중의성 해소에 관한 연구가 많이 이루어지고 있다.

본 논문은 한국어의 구문 구조 분석 시 부사에 의해 발생되는 중의성을 해소하기 위해 수식어 사전을 이용하여 구문 분석에서의 구조 중의성을 해소하였다. 수식어 사전의 어휘정보와 대상 말뭉치를 통해 각각의 부사에 대한 문법을 구성하고, 이를 이용하여 한국어 구문 구조 분석에서 부사에 의해 발생되는 중의성을 줄일 수 있다.

1. 서 론

자연 언어 처리에서 구문 분석이란 주어진 문장의 구조를 구문 규칙에 따라 분석하는 작업을 말한다. 구문 분석 과정에서는 일반적으로 하나 이상의 구문 구조가 생성되며, 이들 중 올바른 구문 구조를 선택하는 작업을 구조적 중의성 해소(structural disambiguation) 작업이라고 한다[1, 2, 3].

한국어를 분석하는데 있어서 다른 자연 언어와 마찬가지로 중의성 해소는 매우 중요한 문제이며, 이러한 문제를 제대로 해결하지 않고서는 실용적인 한국어 처리 시스템을 개발한다는 것이 거의 불가능하다[4].

한국어가 가지는 중의성 중, 구조적 중의성을 해소하기 위하여 많은 연구들이 있어왔다. 구조적 중의성을 해소하는 방법에는 규칙을 이용한 접근 방법과 통계를 이용한 접근 방법이 있는데, 최근에는 통계적 접근 방법이 널리 사용되고 있다. 통계적 접근 방법은 대량의 말뭉치로부터 구조적 중의성 해소에 필요한 확률 정보를 추출하기 때문에, 실제 사람들이 문장을 사용하는 경향을 쉽게 반영할 수 있으며, 지식 획득이 용이하다는 장점을 갖는다. 흔히 사용되는 방법에는 확률 문맥 자유 문법(Probabilistic Context-Free Grammar)이나 확률 의존 문법(Probabilistic Dependency Grammar)과 같은 확률 문법을 이용하는 방법이 있다. 확률 문법을 사용한 구조적 중의성 해결은 매우 간단하다. 구문 트리의 확률값은 그 구문 구조에서 사용되어진 확률 규칙의 확률값의 곱이며, 이와 같이 얻어진 각 구문 트리의 확률값 중 가장 높은 값을 지닌 구문 구조가 입력 문장에 대해 가장 적절한 결과 트리로써 선택되어진다[5].

구조적 중의성의 해소를 위하여 사용되는 통계적 접근 방식 중, 또 다른 방법에는 술어 하위 범주 정보와 격틀 정보, 어휘 공기 정보와 같은 어휘정보(Lexical Information)에 확률을 부여하여 사용하는 방법이 있다. 어휘 정보란 어휘 자체가 가지는 특징들을 기술해 놓은 정보를 말한다. 두 방법 중 확률 문법을 이용한 방법은, 일반적으로 선호되는 구문 구

조에 높은 확률값을 부여하기 때문에, 정확한 구문 분석을 하는데 부족함이 있다. 일반적으로 확률 문법은 어휘는 고려하지 않고, 품사 태그만으로 규칙들을 표현하기 때문이다. 따라서 확률 문법을 보완할 추가의 정보가 필요한데, 확률 어휘정보는 올바른 구문 분석을 하는데 도움이 된다.

한국어에서는 술어의 역할이 매우 중요하기 때문에 술어-인자에 대한 어휘정보에 대한 많은 연구와 이를 이용해 구조적 중의성을 해결하는 연구가 많이 이루어지고 있다. 그러나 한국어와 같이 수식어의 쓰임이 비교적 자유로운 언어에서 수식어에 의한 구조적 중의성 문제는 큰 문제라 할 수 있으나, 수식어에 의한 중의성 해소에 대한 연구는 아직까지 미흡하다. 이러한 연구 중의 하나가 수식어 사전이다. 수식어 사전이란 대상 말뭉치를 형태소 분석기를 통해 분석한 결과 중, 각 수식어에 대한 통계적인 정보와 문장에서 사용되는 수식어의 특징들을 찾은 것으로써 구문 분석 시 이용할 수 있는 수식어에 대한 어휘정보이다 [6].

본 논문에서는 구문 구조 분석에서 수식어의 하나인 부사에 의한 중의성을 기술하고, 이를 해소하기 위해 수식어 사전의 부사 정보를 이용하여 문법 파일을 구성하고 이를 적용함으로써 부사에 의한 한국어 구문 구조 분석의 중의성 해소 방법을 제시한다.

2. 부사에 의한 중의성과 부사 정보

일반적으로 PCFG 나 PDG 와 같은 확률 문법을 사용하여 중의성 문제를 해결하고 있으나 이러한 확률 문법의 구문 규칙은 대개 구문 태그와 품사 태그로만 표

```
[[Parse Tree]]

 1: ((SUBJ (MODT DT +SB (baseform "어느"))
           NN (baseform "조") (caseform "이"))
     (OBJE NN (baseform "이야기") (caseform "을"))
     (MOAD AD (baseform "가장"))
     (MOAD AD (baseform "잘"))
     (MOVV VV +connec +senend (baseform "하") (finform "는지"))
     VV +SE +senend (auxvform " 보") (puncform ".") (baseform "생각하") (finform "읍시다"))

 2: ((SUBJ (MODT DT +SB (baseform "어느"))
           NN (baseform "조") (caseform "이"))
     (OBJE NN (baseform "이야기") (caseform "을"))
     (MOAD (MOAD AD (baseform "가장"))
           AD (baseform "잘"))
     (MOVV VV +connec +senend (baseform "하") (finform "는지"))
     VV +SE +senend (auxvform " 보") (puncform ".") (baseform "생각하") (finform "읍시다"))

 3: ((SUBJ (MODT DT +SB (baseform "어느"))
           NN (baseform "조") (caseform "이"))
     (OBJE NN (baseform "이야기") (caseform "을"))
     (MOAD AD (baseform "가장"))
     (MOVV (MOAD AD (baseform "잘"))
           VV +connec +senend (baseform "하") (finform "는지"))
     VV +SE +senend (auxvform " 보") (puncform ".") (baseform "생각하") (finform "읍시다"))

 4: ((SUBJ (MODT DT +SB (baseform "어느"))
           NN (baseform "조") (caseform "이"))
     (OBJE NN (baseform "이야기") (caseform "을"))
     (MOVV (MOAD (MOAD AD (baseform "가장"))
                 AD (baseform "잘"))
           VV +connec +senend (baseform "하") (finform "는지"))
     VV +SE +senend (auxvform " 보") (puncform ".") (baseform "생각하") (finform "읍시다"))
```

```
5: ((SUBJ (MODT DT +SB (baseform "어느"))
          NN (baseform "조") (caseform "이"))
     (OBJE NN (baseform "이야기") (caseform "을"))
     (MOVV (MOAD AD (baseform "가장"))
          (MOAD AD (baseform "잘"))
          VV +connec +senend (baseform "하") (finform "는지"))
     VV +SE +senend (auxvform "  보") (puncform ".") (baseform "생각하") (finform "읍시다"))

  6: ((SUBJ (MODT DT +SB (baseform "어느"))
          NN (baseform "조") (caseform "이"))
     (MOVV (OBJE NN (baseform "이야기") (caseform "을"))
          (MOAD (MOAD AD (baseform "가장"))
               AD (baseform "잘"))
          VV +connec +senend (baseform "하") (finform "는지"))
     VV +SE +senend (auxvform "  보") (puncform ".") (baseform "생각하") (finform "읍시다"))

  7: ((SUBJ (MODT DT +SB (baseform "어느"))
          NN (baseform "조") (caseform "이"))
     (MOVV (OBJE NN (baseform "이야기") (caseform "을"))
          (MOAD AD (baseform "가장"))
          (MOAD AD (baseform "잘"))
          VV +connec +senend (baseform "하") (finform "는지"))
     VV +SE +senend (auxvform "  보") (puncform ".") (baseform "생각하") (finform "읍시다"))

  8: ((MOVV (SUBJ (MODT DT +SB (baseform "어느"))
               NN (baseform "조") (caseform "이"))
          (OBJE NN (baseform "이야기") (caseform "을"))
          (MOAD (MOAD AD (baseform "가장"))
               AD (baseform "잘"))
          VV +connec +senend (baseform "하") (finform "는지"))
     VV +SE +senend (auxvform "  보") (puncform ".") (baseform "생각하") (finform "읍시다"))

  9: ((MOVV (SUBJ (MODT DT +SB (baseform "어느"))
               NN (baseform "조") (caseform "이"))
          (OBJE NN (baseform "이야기") (caseform "을"))
          (MOAD AD (baseform "가장"))
          (MOAD AD (baseform "잘"))
          VV +connec +senend (baseform "하") (finform "는지"))
     VV +SE +senend (auxvform "  보") (puncform ".") (baseform "생각하") (finform "읍시다"))
```

그림 1. 구문분석기를 통해 생성된 예문 1 의 Parse Tree

현된다. 따라서 PCFG 나 PDG 로도 해결하지 못하는 구조적 중의성 문제가 여전히 존재하게 된다. 이러한 구조적 중의성 문제들 중 하나가 비교적 쓰임이 자유로운 수식어에 의한 중의성이다. 다음의 예문을 보자.

예문 1.
어느 조가 이야기를 가장 잘 하였는지 생각하여 봅시다.

그림 1 은 구문분석기를 통해 생성된 예문 1 Parse Tree 이다. 이 Parse Tree 는 부사 '가장'과 부사 '잘'에 의한 중의성을 포함하고 있다.

수식어 중 비교적 쓰임이 자유로운 부사에 대하여 수식어 사전이 구축되었다. 수식어 사전은 국어 정보베이스를 대상 말뭉치(684372 어절)로 하여, 형태소 분석기를 통해 부사를 추출하고 각 부사에 대한 통계적인 정보와 문장에서 사용되는 부사의 특징들을 정리함으로써 구축되었다.

표 1 은 대상 말뭉치에서 추출된 부사의 결과이며,

표 2 는 추출된 부사 정보와 대상 말뭉치로부터 구축된 수식어 사전 구축 현황을 나타낸다.

수식어 사전은 문장에서의 수식어와 피수식어의 위치 정보와 이들간의 공기 관계, 문장에서의 패턴 등의 통계적인 정보를 포함하고 있으며, 이러한 통계적인 정보들은 앞에서 설명한 수식어에 의한 중의성 해소를 위해 사용된다.

- 대상 말뭉치: 국어 정보 데이터 베이스 (684372 어절)
- 사용 형태소 분석기 : CBKMA

부사 종류 수	1351
전체 부사 수	47792

상위 30개 부사	41.83%
나머지1295개 부사 (출현빈도수 200 이하)	43.62%

표 1. 부사 추출 결과

	빈도순위	빈도수	부사	비율
1	1	2089	그러나	4.37%
2	2	1442	그리고	3.02%
3	4	967	가장	2.02%
4	6	921	더	1.93%
5	8	783	따라서	1.64%
6	9	723	바로	1.51%
7	10	707	다시	1.48%
8	12	615	같이	1.29%
9	13	605	이미	1.27%
10	14	589	잘	1.23%

표 2. 수식어 사전 구축 현황

그림 1 에서 부사 '가장'과 부사 '잘'에 의한 중의성을 포함하는 Parse Tree(1, 2, 3, 5, 7, 9)는 부사 '가장'의 바로 뒤에 부사가 나올 경우, '가장'은 뒤의 부사를 수식하고, '잘'은 바로 뒤의 용언을 수식한다는 어휘 정보를 이용하여 제거할 수 있다. 이것은 수식어 사전의 어휘 정보를 이용하여 수식어에 의한 구조적 중의성을 해소함을 보인다.

3. 문법 파일의 구성

문법 파일은 수식어 사전의 통계적인 정보들을 이용하여 구성되어진다. 통계 정보에 근거하여 문장 속에서 각 부사의 중의성 해소에 이용할 정보를 찾고, 이것으로부터 문법 파일을 구성하게 된다.

- ' 가장' 통계 정보
  - 피수식어의 품사
    - 형용사 수식 : 642 (68.59 %)
    - 동사 수식 : 190 (20.30 %)
    - 부사 수식 : 74 (7.91 %)
    - 명사 수식 : 19 (2.03 %)
    - 관형사 수식 : 8 (0.85 %)
    - 예외 : 3 (0.32 %)
      (명사 ' 가장' 으로 쓰인 경우)
  - ' 가장' 과 피수식어의 위치 관계
    - 피수식어가 1 어절 뒤에 있는 경우 : 876 (93.59 %)
    - 피수식어가 2 어절 뒤에 있는 경우 : 58 (6.20 %)
    - 피수식어가 3 어절 뒤에 있는 경우 : 2 (0.21 %)
- 문법 파일
  - JJ
  - VV
  - DD
  - NJJ
  - NVV

그림 2. 부사 '가장'의 통계정보와 문법파일

한국어에서 수식어는 피수식어의 앞에서 수식을 하므로 수식어의 뒤에 오는 어휘들을 검사하여 피수식어를 찾아낼 수 있다. 그러므로 문법 파일은 각 수식어의 수식어 사전의 통계 정보와 대상 말뭉치의 예문에서 수식어의 뒤에 오는 어휘들을 검사함으로써 구성되어질 수 있다. 이렇게 구성된 문법 파일을 구문 분석 결과에 적용함으로써 구문 분석에서 생성된 수식어에 의한 중의성 있는 구문 분석 구조를 제거할 수 있다.

그림 2 는 '가장'의 통계 정보와 그것으로부터 구성된 문법 파일을 보여주고 있다. 그림 2 에서 '가장'의 통계 정보는 수식어 사전의 내용이며, '가장'의 문법

파일은 수식어 사전과 대상 말뭉치로부터 만들어진다. 문법 파일을 살펴보면 JJ 와 NJJ 를 볼 수 있는데, JJ 는 '가장'(부사) 다음에 J(형용사)가 오면 '가장'은 바로 뒤의 형용사를 수식함을 의미하며, NJJ 는 '가장'(부사) 다음에 N(명사), J(형용사)의 순서로 나타날 때 '가장'(부사)은 명사 다음의 형용사를 수식함을 의미한다. 즉, 문법에서 마지막 문자는 부사가 수식하는 피수식어이며, 그 앞의 문자들은 부사와 피수식어 사이에 나타나는 어휘들을 의미한다.

문법 파일은 각각의 부사에 대해 구성되어지며, 어떤 한 문장의 구조적 중의성을 해소할 경우 그 문장의 모든 부사의 문법 파일이 적용되어진다.

4. 실험 결과

실험용 말뭉치는 두 가지를 사용하였다. 하나는 수식어 사전을 구축하기 위해 사용되었던 국어 정보 베이스이고, 다른 하나는 ETRI 품사 태그 부착 말뭉치이다. 이 두 가지의 말뭉치에서 부사를 포함하는 문장을 추출하여 중의성 해소 실험에 사용하였다. 먼저, 추출된 문장들을 구문분석기를 통해 Parse Tree 를 생성하고, 다시 문법 파일을 적용함으로써 실험하였다.

사용 말뭉치	평균 중의성 해소율
국어 정보 베이스	42.59 %
ETRI 품사 태그 부착 말뭉치	39.47 %
평 균	41.03 %

표 3. 부사 정보에 의한 중의성 해소 실험 결과

표 3 은 두 개의 실험용 말뭉치에 대한 중의성 해소 실험 결과이다. 실험 결과, 국어 정보 베이스 말뭉치가 ETRI 품사 태그 부착 말뭉치보다 약간의 좋은 결과가 나왔다. 이것은 국어 정보 베이스가 수식어 사전의 말뭉치로 사용되었기 때문이며, 실험용 말뭉치가 커질수록 이러한 차이는 줄어들 것이다.

위의 실험 결과가 부사에 의한 중의성을 모두 해소한 것은 아니다. 그 원인은 두 가지로 볼 수 있다. 첫 번째는 어휘들의 오분석으로 인해 문법 파일의 적용이 잘못된 것이다. 이러한 경우는 문법 파일의 적용이 의미가 없어진다. 즉, 부사 '안'은 장소를 가리키는 명사로도 분석이 가능한 경우가 있다. '안'이 명사로 분석된 경우, 부사 '안'의 문법 파일의 적용은 의미가 없어지게 된다. 또 다른 하나는 문법 파일이다. 문법 파일은 각각의 수식어에 대한 통계정보와 문장에서 수식어의 사용에 따라 구성되어지므로 문법 파일이 수식어가 사용된 모든 경우를 표현하지 못하기 때문이다. 따라서 이러한 문제는 대량의 말뭉치로부터 어휘정보를 추출하고, 이를 적용하여 문법 파일을 구성함으로써 해결되어질 수 있다.

5. 결 론

본 논문에서는 수식어 사전의 부사 정보를 이용하여 부사의 문법 파일을 작성하고, 구문분석 결과에 적용함으로써 부사에 의한 구조적 중의성 해소 방법을 제안하였다. 부사에 의한 중의성 해소 실험을 통해 41.03 %의 중의성 해소율을 보였다. 부사에 의한 중의성 해소는 문법 파일에 따라 이루어지며, 중의성 해소 결과는 문법 파일의 질에 따라 결정되어진다. 이것은 수식어에 의한 중의성 해소를 통해 더욱 향상된 구문분석 결과를 얻을 수 있음을 보인다.

이 실험으로 수식어 사전의 통계 정보를 이용하여 부사에 의한 중의성 해소의 가능성을 보였으며, 향후 다른 수식어들의 통계 정보를 이용하여 구조적 중의성 해소에 더 향상된 결과를 얻을 수 있다. 이를 위해 수식어 사전의 확장과 더 정확한 문법 파일의 작성을 위한 연구가 필요하다. 지금까지는 성분부사에 대한 연구만 이루어지고 있으나 향후 접속부사와 문장부사에 의한 중의성 해소 방안에 대한 연구도 이루어져야 하며, 또한 수식어 어휘정보의 자동 구축 방법에 대한 연구도 이루어져야 할 것이다.

참고 문헌

[1] 김영택, "자연언어처리", 교학사, 1994

[2] Makoto Nagao, "자연언어처리", 홍릉과학출판사, 1998

[3] 정후중, 황영숙, 곽용재, 박소영, 임해창, "구문 분석에서의 중의성 해소를 위한 일반화된 어휘정보의 자동 구축 및 적용", 제 10 회 한글 및 한국어 정보처리 학술 발표 논문집, pp.269~275, 1998.

[4] 심광섭, 김영택, "통계 정보를 이용한 구조적 중의성 해소", 한국정보과학회 논문지 제 21 권 제 2 호, 1994.2

[5] 이공주, 김재훈, 김길창, "중심어간의 공기정보와 구문 규칙을 기반으로 한 확률적 한국어 구문 분석", 제 9 회 한글 및 한국어 정보처리 학술발표 논문집, pp.332~338, 1997.

[6] 신승은, 서영훈, "한국어 구조 중의성 해소를 위한 수식어 사전", 한국정보과학회 충청지부 추계 학술발표논문집 제 11 권 1 호, pp.73~76, 1999.

[7] 이수선, 박현재, 우요섭, "한국어 분석의 중의성 해소를 위한 하위범주화 사전 구축", 제 11 회 한글 및 한국어 정보처리 학술 발표 논문집, pp.257~264, 1999.

[8] 송영빈, 채영숙, 박용일, 이정민, 설가영, 황혜리, 한나리, 최기선, "동사의 애매성 해소를 위한 구문 의미사전의 구축", 제 11 회 한글 및 한국어 정보처리 학술 발표 논문집, pp.280~287, 1999.

[9] 남기심, 고영근, "표준 국어문법론", 탑출판사, 1996

[10] 엄미현, 신대규, 나동렬, "한국어의 구조적인 애매성", 한국정보과학회 봄 학술발표논문집 제 23 권 1 호, pp.911~914, 1996.