부사 정보를 이용한 구문 구조 선택

†신 승 은, *정 천 영, ††서 영 훈
†, ††충북대학교 컴퓨터공학과
*혜천대학 컴퓨터통신계열
& 컴퓨터정보통신연구소
†seshin@dcenlp.chungbuk.ac.kr, *cyjung@hcc.ac.kr, ††yhseo@cbucc.chungbuk.ac.kr

Parse Tree Selection using Adverb Information

†Seung-Eun Shin, *Cheon-Young Jung, ††Young-Hoon Seo
†, ††Dept. of Computer Engineering, Chungbuk National University
*School of Computer Science & Telecommunications, Hyechon College
& Research Institute for Computer and Information Communication

요 약

자연 언어 처리의 구문 구조 분석에서는 수식 관계의 중의성에 의한 많은 구문 구조가 생성된다. 이러한 중의성을 해소하는데 어휘 정보가 유용하다는 것은 잘 알려져 있다.

본 논문은 한국어의 구문 구조 분석 시 중의성을 해소하기 위해 어휘 정보로 부사 수식 정보와 부사 확률 정보를 사용한다. 부사들의 사용과 수식 패턴들을 대량의 말뭉치로부터 조사하고, 수식 패턴들 중 비교적 규칙적인 것들을 부사 수식 정보로, 피수식어의 상대적 위치와 피수식어의 품사에 대한 확률을 부사 확률 정보로 구성하였다. 구문 구조들 중 가장 옳은 구문 구조를 선택하기 위해 부사 수식 정보와 부사 확률 정보를 이용하였고, 구문 분석에서 부사에 의한 중의성을 해소하였다.

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].

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. 가장 널리 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 구조의 확률값은 그 구문 구조에서 사용되어진 확률 규칙의 확률값의 곱이며, 이와 같이 얻어진 각 구문 구조의 확률값 중 가장 높은 값을 지닌 구문 구조가 입력 문장에 대해 가장 적절한 결과 구조로써 선택되어진다[5].

구조적 중의성의 해소를 위하여 사용되는 통계적 접근 방식 중, 또 다른 방법에는 술어 하위 범주 정보와 격틀 정보, 어휘 공기 정보와 같은 어휘 정보(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 구조에 높은 확률값을 부여하기 때문에, 정확한 구문 분석을 하는데 부족함이 있다.

다음의 예문을 보자.

예문 1.

- 나는 망원경으로 할머니가 가는 것을 보았다.
- 나는 집으로 할머니가 가는 것을 보았다.

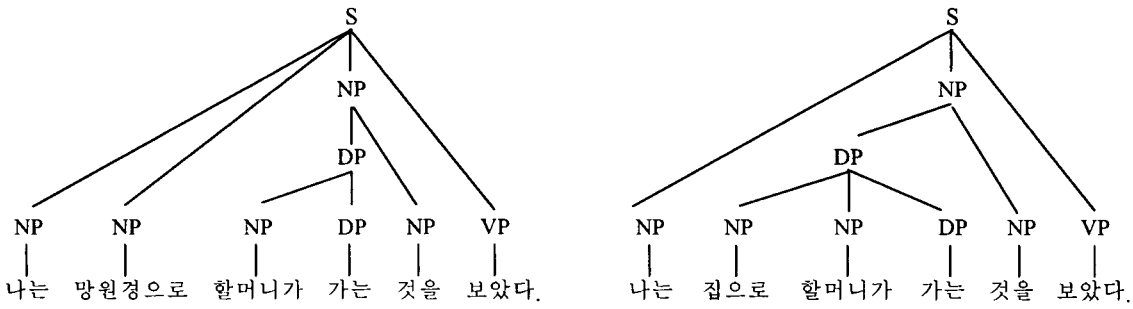
비록 위 예문 1 의 두 문장이 동일한 품사열로 구성되어 있지만, 두 문장의 구문 구조는 다르다(그림 1). 일반적으로 확률 문법은 어휘는 고려하지 않고, 품사 태그만으로 규칙들을 표현하기 때문이다. 따라서 확률 문법을 보완할 추가의 정보가 필요한데, 확률 어휘 정보는 올바른 구문 분석을 하는데 도움이 된다.

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[6].

본 논문에서는 구문 구조 분석에서 수식어의 하나인 부사에 의한 중의성을 기술하고, 이를 해소하기 위해 대량의 말뭉치로부터 부사의 사용과 수식 패턴을 추출하고 추출된 통계 정보와 특징들을 이용하여 부사 수식 정보(Modificatory Adverb Information)와 부사 확률 정보(Probabilistic Adverb Information)를 구성하고 이를 적용하여 구문 분석 결과 중 옳은 구문 구조를 선택함으로써 부사에 의한 한국어 구문 구조 분석의 중의성 해소 방법을 제시한다.

2. 부사에 의한 구조적 중의성

구문 분석 과정에서 중의성 해소를 위해 일반적으로 PCFG(Probabilistic Context-Free Grammar)나 PDG(Probabilistic Dependency Grammar)와 같은 확률 문법을 사용하고 있으나 이러한 확률 문법의 구문 규칙은 대개 구문 태그와 품사 태그로만 표현된다. 따라서 PCFG 나 PDG 로도 해결하지 못하는 구조적 중의성 문제가 여전히 존재하게 된다. 이러한 구조적 중의성 문제들 중 하나가 비교적 쓰임이 자유로운 부사에 의한 중의성이다. 한국어에서 부사는 동사, 형용사, 명사 그리고 문장을 수식한다. 부사의 명사 수식에 대한 세 가지 견해가 있다. 첫 번째는 관형사와 부사의 두 범주로 보는 방식이고, 두 번째는 이 어휘들이 부사인데, 이들이 용언을 수식할 때에는 부사어이고, 체언을 수식할 때에는 관형어로 쓰인다는 입장과 세 번째 이들이 항상 부사이며 체언을 수식하는 경우도 전형적인 관형어와는 다른 수식 양상을 보이므로 관형어로 사용되는 것으로 보기 어렵다는 입장이 있다[6]. 본 논문에서는 부사의 명사 수식을 세 번째 견해를 따른다. 따라서 많은 구조적 중의성이 부사에 의해 발생된다.



S: Sentence, NP: Noun Phrase, VP: Verb Phrase, DP: Definitive Phrase

그림 1: 다른 구문 구조를 가지는 동일한 품사열의 문장 (예문 1)

```
1: ((SUBJ (MODT DT "어느") NN "조가")
    (OBJE NN "이야기를")
    (MOAD AD "가장")
    (MOAD AD "잘")
    (MOVV VV "하였는지")
    VV "생각하여 봅시다")

2: ((SUBJ (MODT DT "어느") NN "조가")
    (OBJE NN "이야기를")
    (MOAD (MOAD AD "가장") AD "잘")
    (MOVV VV "하였는지")
    VV "생각하여 봅시다")

3: ((SUBJ (MODT DT "어느") NN "조가")
    (OBJE NN "이야기를")
    (MOAD AD "가장")
    (MOVV (MOAD AD "잘") VV "하였는지")
    VV "생각하여 봅시다")

4: ((SUBJ (MODT DT "어느") NN "조가")
    (OBJE NN "이야기를")
    (MOVV (MOAD (MOAD AD "가장") AD "잘")
          VV "하였는지")
    VV "생각하여 봅시다")

5: ((SUBJ (MODT DT "어느") NN "조가")
    (OBJE NN "이야기를")
    (MOVV (MOAD AD "가장")
          (MOAD AD "잘")
          VV "하였는지")
    VV "생각하여 봅시다")

6: ((SUBJ (MODT DT "어느") NN "조가")
    (MOVV (OBJE NN "이야기를")
    (MOAD (MOAD AD "가장") AD "잘")
          VV "하였는지")
    VV "생각하여 봅시다")

7: ((SUBJ (MODT DT "어느") NN "조가")
    (MOVV (OBJE NN "이야기를")
          (MOAD AD "가장")
          (MOAD AD "잘")
          VV "하였는지")
    VV "생각하여 봅시다")

8: ((MOVV (SUBJ (MODT DT "어느") NN "조가")
          (OBJE NN "이야기를")
          (MOAD (MOAD AD "가장") AD "잘")
          VV "하였는지")
    VV "생각하여 봅시다")

9: ((MOVV (SUBJ (MODT DT "어느") NN "조가")
          (OBJE NN "이야기를")
          (MOAD AD "가장")
          (MOAD AD "잘")
          VV "하였는지")
    VV "생각하여 봅시다")
```

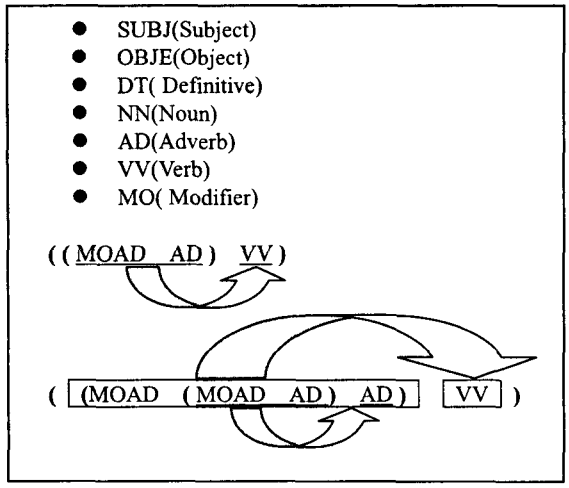


그림 2. 예문 2의 구문 구조 (Parse Trees)

다음의 예문을 보자.

예문 2.

- 어느 조가 이야기를 가장 잘 하였는지 생각하여 봅시다.

그림 2 는 구문분석기를 통해 생성된 예문 1 의 Parse Tree 이다. 이 Parse Tree 는 부사 '가장'과 부사 '잘'에 의한 구조적 중의성을 포함하고 있다. 9 개의 구문 분석 결과들 중 5 개의 구문 분석 결과들[1, 2, 3, 5, 7, 9]은 부사 '가장'과 부사 '잘'에 의한 구조적 중의성을 포함한다. 그러나 이러한 구조적 중의성은 부사 '가장'이 다음 어절이 부사일 경우 부사를 수식한다는 어휘 정보와 부사 '잘'이 바로 뒤의 용언을 수식한다는 어휘 정보들로 구성된 부사 수식 정보를 적용함으로써 해소할 수 있다.

한국어 구조 중의성 해소를 위한 수식어 사전은 국어 정보 베이스[11]를 대상 말뭉치(684372 어절)로 하여, 각각의 부사에 대한 통계 정보와 특징들로 구성되어있다[7]. 표 1 은 대상 말뭉치에서 추출한 부사의 결과이며, 표 2 는 추출한 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 부사에 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%
나머지1321개 부사	58.17%

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 (상위 10 개)

3. 부사 수식 정보

부사 수식 정보는 잘못된 구문 분석 결과를 제거하기 위해 사용하며, 수식어 사전의 통계 정보와 말뭉치에서 추출된 부사의 특징들을 이용하여 구성한다. 통계 정보에 근거하여 문장 속에서 각 부사의 중의성 해소에 이용할 정보를 찾고, 이것으로부터 부사 수식 정보를 구성한다.

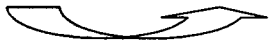
그림 3 은 부사 '가장'의 통계 정보와 그것으로부터 구성한 부사 수식 정보를 보여주고 있다. 그림 3.1 에서 '가장'의 통계 정보는 수식어 사전의 내용이며, '가장'의 부사 수식 정보는 수식어 사전과 대상 말뭉치로부터 만들어진다. 부사 수식 정보를 살펴보면 JJ 와 NJJ 를 볼 수 있는데, JJ 는 '가장'(부사) 다음에 J(형용사)가 오면 '가장'은 바로 뒤의 형용사를 수식함을 의미하며, NJJ 는 '가장'(부사) 다음에 N(명사),

1. ' 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 %)

2. 부사 수식 정보
   - JJ
   - VV
   - DD
   - NJJ
   - NVV

그림 3. 부사 '가장'의 통계 정보와 부사 수식 정보

1. 부사 A 의 부사 수식 정보 : NVV
   NVV : A + noun + verb



2. 부사 B 의 부사 수식 정보 표기법 : XY*Z
   X : B 다음 어절의 품사
   Y : X 다음 어절의 품사
   Z : B 가 수식하는 어절의 품사
   Y* : Y 가 0 또는 한번 이상 나타날 수 있음

그림 4. 부사 수식 정보

J(형용사)의 순서로 나타날 때 '가장'(부사)은 명사 다음의 형용사를 수식함을 의미한다. 즉, 부사 수식 정보에서 마지막 문자는 부사가 수식하는 피수식어이며, 그 앞의 문자들은 부사와 피수식어 사이에 나타나는 어휘들을 의미한다. 그림 4 를 보면, A 는 부사이며, N 은 명사 그리고 V 는 동사이다. 여기서 A 의 부사 수식 정보 NVV 는 A 가 뒤 어절이 명사 + 동사일 경우 동사를 수식함을 의미한다. 그림 4.2 는 부사 수식 정보의 표기법을 나타낸다.

한국어에서 수식어는 피수식어의 앞에서 수식을 하므로 수식어의 뒤에 오는 어휘들을 검사하여 피수식어를 찾아낼 수 있다. 그러므로 부사 수식 정보는 각 부사의 수식어 사전 통계 정보와 대상 말뭉치의 예문에서 부사의 뒤에 오는 어휘들을 검사함으로써 구성할 수 있다. 이렇게 구성한 부사 수식 정보를 구문 분석 결과에 적용함으로써 구문 분석에서 생성된 부사에 의한 중의성 있는 구문 분석 구조를 제거할 수 있다. 부사 수식 정보는 각각의 부사에 대해 구성하며, 어떤 한 문장의 구조적 중의성을 해소할 경우 그 문장의 모든 부사의 부사 수식 정보를 적용한다.

4. 부사 확률 정보 (Probabilistic Adverb Information)

부사 수식 정보는 잘못된 구문 분석 결과를 제거하기 위해 사용하는 반면, 부사 확률 정보는 가장 옳은 구문 분석 결과로 예측되어지는 분석 결과를 선택하기 위해 사용한다. 부사 확률 정보는 수식어와 피수식어의 상대적 위치와 피수식어의 품사 확률에 의해 만든다. 표 3 은 부사 '더'와 '잘'의 부사 확률 정보이다. 각 부사의 통계 정보를 이용하여 피수식어의 품사 확률과 상대적 위치 확률을 구해 부사 확률 정보를 구성한다.

이러한 부사 확률 정보는 그림 5 와 같이 구문 분석 결과에 적용할 수 있다. 그림 5 는 구문 분석 결과에 부사 확률 정보를 적용한 Scoring 의 예이다. Score 는 각각의 부사가 수식하는 피수식어의 품사 확률값과 피수식어의 위치 확률값을 곱한 값이며, 가장 높은 Score 를 갖는 구문 분석 결과를 부사에 대한 가장 옳은 구문 분석 결과로 선택한다. 실제로 그림 5 에서도 분석 결과 1 이 부사에 대한 옳은 분석 결과이다. 이렇게 계산된 Score 는 부사에 대한 가장 옳은 구문 분석 결과를 선택하기 위한 기준이 된다.

피수식어와의 상대적 위치	더	잘
1	0.9443	0.9941
2	0.0383	0.0059
3	0.0098	
4	0.0076	
피수식어의 품사	확 률	
V (Verb)	0.3147	0.9879
J (adjective)	0.4754	0.0121
N (Noun)	0.1180	
D (adverb)	0.0831	

표 3. '더'와 '잘'의 부사 확률 정보

Sentence : "나는 노래를 더 잘 부를 수 있다."

[[Parse Tree]]

Score 1 : 0.077064
1: ((SUBE NN "나는")
(OBJE NN "노래를")
(MOAD (MOAD AD "더") AD "잘")
VV "부를 수 있다")
0.077064 = 0.9443 × 0.0831 × 0.9941 × 0.9879

Score 2 : 0.011837
2: ((SUBE NN "나는")
(OBJE NN "노래를")
(MOAD AD "더")
(MOAD AD "잘")
VV "부를 수 있다")
0.011837 = 0.0383 × 0.3147 × 0.9941 × 0.9879

그림 5. 구문 분석 결과에 부사 확률 정보를 적용한 Scoring 의 예

5. 실험 결과

부사 수식 정보는 '가장', '매우', '안', '잘', '못'에 대해 구축하였으며, 부사 확률 정보는 '거의', '다시', '더', '더욱', '많이', '보다', '서로', '이렇게'에 대해 구축하였다. 구축한 부사 수식 정보와 부사 확률 정보를 실험용 말뭉치에 적용하여 중의성 해소 실험을 하였다. 실험용 말뭉치는 두 가지를 사용하였다. 하나는 수식어 사전을 구축하기 위해 사용하였던 국어 정보 베이스이고, 다른 하나는 ETRI 품사 태그 부착 말뭉치이다. 이 두 가지의 말뭉치에서 부사 수식 정보와 부사 확률 정보를 구축한 부사를 포함하는 문장을 100 문장씩 추출하여 중의성 해소 실험(실험 1)을 하고, 모든 부사를 포함하는 문장을 100 문장씩 추출하여 중의성 해소 실험(실험 2)을 하였다.

표 4 와 표 5 는 각각 실험 1 과 실험 2 에 대한 실험 결과이다. 실험 1 에서는 부사 수식 정보와 부사 확률 정보를 구축한 부사를 포함하는 문장들을 실험용 말뭉치로 사용하여 50.4%의 중의성 해소율을 보였다. 이것은 부사 이외의 원인으로 인한 중의성 때문이다. 실험 2 에서는 모든 부사를 포함하는 문장들을 실험용 말뭉치로 사용하였기 때문에 다소 낮은 평균 중의성 해소율을 보인다. 또한 두 개의 말뭉치에서 중의성이 차이가 있는 것은 국어 정보 베이스에서 추출된 실험용 말뭉치에 '그러나'와 '그리고'와 같은 문두에 사용된 접속 부사를 포함하는 문장들이 비교적 많이 포함되었기 때문이다.

사용 말뭉치	평균 중의성 해소율
국어 정보 베이스	48.17 %
ETRI 품사 태그 부착 말뭉치	52.63 %
평 균	50.40 %

표 4. 실험 1 의 실험 결과

사용 말뭉치	평균 중의성 해소율
국어 정보 베이스	7.07 %
ETRI 품사 태그 부착 말뭉치	8.93 %
평 균	8.00 %

표 5. 실험 2 의 실험 결과

$$\text{중의성 해소율} = \frac{\text{제거된 구문 구조의 수}}{\text{전체 구문 구조의 수}} \times 100\%$$

이 실험은 13 개의 부사에 대한 부사 수식 정보와 부사 확률 정보를 적용한 것이므로 부사 수식 정보와 부사 확률 정보를 확장한다면 구문 분석에서의 중의성 해소에 큰 역할을 할 것이다.

6. 결 론

본 논문에서는 부사의 통계 정보와 문장에서의 특징을 이용하여 부사의 부사 수식 정보와 부사 확률 정보를 구축하고, 구문 분석 결과에 적용함으로써 부사에 의한 구조적 중의성 해소 방법을 제안하였다. 부사 수식 정보는 구문 분석 결과 중 잘못된 구문 분석을 제거하기 위해 사용하고, 부사 확률 정보는 부사에 대해 가장 옳은 구문 분석 결과로 예측되어지는 분석 결과를 선택하기 위해 사용한다. 부사 수식 정보와 부사 확률 정보를 구축한 부사를 포함하는 말뭉치에 대한 중의성 해소 실험은 50.4%, 모든 부사를 포함하는 말뭉치에 대한 중의성 해소 실험은 8%의 중의성 해소율을 보였다. 이것은 13 개의 부사에 대한 부사 수식 정보와 부사 확률 정보를 작성한 것이므로 정보를 확장한다면 더욱 향상된 중의성 해소율을 보일 것이다. 이 실험으로 부사의 통계 정보와 문장에서의 특징을 이용하여 부사에 의한 중의성 해소의 가능성을 보였다.

향후 연구 과제로는 부사 수식 정보와 부사 확률 정보에 의해 해소되지 않는 중의성에 대한 연구와 부사 수식 정보와 부사 확률 정보의 확장이 필요하며, 이를 위한 어휘 정보의 자동 구축에 대한 연구도 필요하다. 더 나아가 부사 뿐만 아니라 다른 수식어들에 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 어휘정보

의 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] 임유종, "한국어 부사 연구", 한국문화사, 1999.

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

[8] 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.

[9] 손남익, "국어부사연구", 박이정, 1995.

[10] 남기심, 고영근, "표준 국어문법론", 탑출판사, 1996