# 구문 분석에서의 어휘간 공기 정보의 활용[1]

윤 준 태
한국과학기술원 인공지능연구센터
대전시 유성구 구성동 373-1
우: 305-701
jtyoon@december.yonsei.ac.kr

최 기 선
한국과학기술원 인공지능연구센터
대전시 유성구 구성동 373-1
우: 307-701
kschoi@world.kaist.ac.kr

김 선 호
연세대학교 공과대학 컴퓨터과학과
서울 서대문구 신촌동
우: 120-749
pobi@december.yonsei.ac.kr

송만석
연세대학교 공과대학 컴퓨터과학과
서울 서대문구 신촌동
우: 120-749
mssong@december.yonsei.ac.kr

# Using Lexical Co-occurrence Information in Syntactic Analysis

Yoon, Juntae
*Department of Computer Science, CAIR*
KAIST

Choi, Key-Sun
*Department of Computer Science, CAIR*
KAIST

Kim, Seonho
Department of Computer Science, Engineering Science
Yonsei University

Song, Mansuk
Department of Computer Science, Engineering Science
Yonsei University

## 요약

구문 분석에 있어서 어휘 정보는 구문적 중의성을 해결하는 데 매우 중요한 역할을 한다. 본 논문에서는 대량의 말뭉치로부터 추출된 공기 정보가 구문 분석에서 효과적으로 이용될 수 있음을 보인다. 첫째, 공기 정보로부터 보다 의미있는 연어를 추출하고 이를 구문 분석에 이용함으로써 보다 효율적인 파서의 구축이 가능함을 밝힌다. 둘째로는 대량의 말뭉치로부터 추출한 공기 정보가 구문 분석시 보조사나 조사 생략에 의한 격 중의성 혹은 관계관형절에서 발생하는 명사구 이동에 따른 격 중의성의 해결에 적용될 수 있음을 보인다. 이를 위해 본 연구에서는 연세대학교 한국어 사전 편찬실의 연세 말뭉치 3,000만 어절과 KAIST 말뭉치 중 1,000만 어절로부터 <서술어,명사,격관계> 공기 정보를 추출하였다.

## 1 서론

최근의 자연 언어 처리의 많은 분야에서 말뭉치가 이용되어 왔다. 품사 태깅을 위해서 품사 부착 말뭉치가 이용되었으며, 구문 분석에서 발생하는 구조적 중의성의 해소를 위해 구문구조 부착 말뭉치가 이용되기도 하였다. 특히 최근의 말뭉치를 이용한 구문 분석에서는 어휘 정보가 구문적 중의성 해소에 있어서 매우 중요한 역할을 함을 보였다 [2,4,5,8,10]. 또 말뭉치로부터 구한 통계값의 이용은 중의성 해소와 함께 가중치에 의한 우선 탐색 기법으로 정확성을 크게 잃지 않으면서도 효율적인 분석을 가능하게 해 자연 언어 처리의 응용 분야에서도 유용하게 이용될 수 있다.

말뭉치를 이용한 연구는 정보가 부착된 말뭉치를 이용하는 통제 학습에 의한 연구와 정보가 부착되지 않은 원시 말뭉치로부터 정보를 추출하는 비통제 학습에 의한 연구로

[1] 본 연구는 정보통신부[지능형 멀티미디어 통합 정보베이스]와 과학기술부[대용량 국어정보 심층처리 및 품질관리 기술개발]의 지원을 받아 수행된 연구의 일환으로 이루어졌다.

크게 나눌 수 있다. 통제 학습을 이용하는 경우 수작업에 의해 구축된 비교적 정확한 자료로부터 학습할 수 있으므로 구축된 언어 지식 역시 매우 정확하다는 특징이 있다. 그러나 여기에는 수작업에 의한 정보원의 구축 자체가 매우 어렵다는 문제가 있는데, 특히 말뭉치에 구문 정보를 부착하는 일은 품사를 부착하는 일보다 훨씬 어려우며 오류를 포함할 가능성도 더 높다고 할 수 있다. 따라서 현재까지 구축된 구구조 부착 말뭉치는 충분한 양의 어휘 정보를 구축하기에는 어렵다는 점을 가지고 있다.

본 연구에서는 정보 부착 말뭉치로부터 추출하는 것보다는 다소 정확성이 떨어지나 대량의 자료로부터 어휘 정보를 얻기 위해 비통제 학습을 이용해 서술어와 명사구간의 공기 정보를 추출하며 이들이 구문 분석에서 어떻게 이용될 수 있는지를 살펴본다. 우선 공기(co-occurrence) 정보를 추출한 후 보다 의미 있는 연어 (collocation) 정보를 추출하여 구문 분석에 이용하였다. 자연 언어 처리 시스템이 응용 시스템에서 보다 효과적으로 이용되기 위해서는 효율적인 처리가 가능해야 한다. 본 연구에서 추출된 연어 정보는 구문 분석 단계에서 어절간의 그룹화를 가능하게 하여 효율성을 높일 수 있음을 보인다. 다음으로는 한국어 구문 분석에서의 또 하나의 문제인 보조사 및 조사 생략에 의한 격 중의성과 관계 관형절에서의 이동에 의한 이동 명사구의 복원에 있어서 추출된 공기 정보를 이용한다. 본 논문에서는 이를 위해 연세대학교 한국어 사전 편찬실 말뭉치 중 3,000만 어절과 KAIST 말뭉치 중 1,000만 어절로부터 <서술어,명사,조사> 공기 정보를 추출해 실험에 이용하였다.

## 2 공기 정보

본 연구에서 공기 정보는 원시 말뭉치를 형태소 분석기와 자동 태거를 이용하여 분석한 후 휴리스틱에 의해 <서술어, 명사, 조사> 관계를 추출하되 오류를 최소한으로 하기 위해 최대한 정확한 쌍들만을 추출하였다. 추출된 공기 데이터는 약 300만 쌍 가량이다. 이들의 특징을 간단히 기술하면 다음과 같다.

1. 지정사도 고려의 대상이 되었다. 지정사가 결합된 명사는 '명사+지정사'를 서술어로 간주한다.
2. 형용사의 관형형이나 지정사의 관형형이 결합된 어절에 대해서는 공기 정보 추출에서 배제하였다.
3. '-어 지다'와 같은 몇몇 보조용언은 하위범주화 틀을 변화시키는 특징을 가지고 있는데 이들은 하나로 묶어 복합 동사로 처리하였다.
4. '돌아오다'와 같이 'V1+어/E+V2' 형태의 복합동사중에는 하나의 의미 단위처럼 쓰기도 하고 어떤 동사가 의미역을 할당하는지에 대한 판단이 필요한 것들이 많은데 이들은 묶어서 하나의 동사로 처리하였다.[12].

표 1은 추출된 공기 데이터의 예를 보여 준다.

표 1 동사 '오다'와 관련된 공기 정보의 예

| 동사 | 명사 | 조사 | 빈도수 | 연어 |
|---|---|---|---|---|
| 오 | 가게 | 로 | 9 | O |
| 오 | 가게 | 를 | 2 | X |
| 오 | 가게 | 에 | 10 | O |
| 오 | 가격변동 | 가 | 1 | X |
| 오 | 가구점 | 에 | 1 | X |
| 오 | 가기 | 를 | 1 | X |
| 오 | 가마니 | 에 | 1 | X |
| 오 | 단계 | 에 | 36 | O |
| 오 | 당국 | 에 | 2 | X |
| 오 | 때 | 가 | 327 | O |
| 오 | 멤버 | 가 | 1 | X |
| 오 | 멱 | 를 | 1 | X |
| 오 | 면회객 | 가 | 1 | X |
| 오 | 몸종 | 가 | 1 | X |
| 오 | 몸짓 | 가 | 1 | X |
| 오 | 묘지 | 에 | 3 | X |
| 오 | 생활 | 가 | 3 | X |
| 오 | 서울 | 로 | 121 | O |
| 오 | 서울 | 에 | 211 | O |
| 오 | 식당 | 에서 | 2 | X |
| 오 | 어머니 | 가 | 113 | O |
| 오 | 전쟁 | 가 | 1 | X |
| 오 | 전쟁 | 로 | 2 | X |
| 오 | 지점 | 로 | 1 | X |

## 3 연어의 추출

### 3.1 연어 추출

자연어의 단어들 중에는 흔히 반복적으로 군을 이루면서 발생하는 특징을 가지는 것들이 있다. 즉, 이들에 속하는 단어들은 임의적이지만 우연에 의해 기대되는 것 이상으로 자주 발생하는 특징을 가지고 있는데 이들을 연어(collocation)라 한다. 연어는 사전 편찬학적 관점에서 그리고 자연어 처리 관점에서 모두 매우 중요하다. 따라서 이에 관한 많은 연구가 이루어져 왔는데, 이들은 주로 말뭉치로부터 구한 데이터를 바탕으로 연어에 대해 수학적인 정의를 내리고 이들을 자동적으로 추출하는 방법론을 포함하고 있다[1,3,6,9]

이들의 연구는 대개 상호 정보나 log-likelihood, z-score등을 이용하고 있는데 이들 테스트는 어떤 단어들이 우연의 결과로 발생할 확률과 그 단어들이 함께 나타날 확률을 비교하여 우연 이상의 반복적 발생을 찾아내는 특성을 가지고 있다.

본 연구에서는 각 서술어에 대해 나타나는 각 명사구 공기 정보에 대해 정규 분포를 가정하고 이들로부터 연어 정보를 추출하여 이를 구문 분석에 응용하는 방법론에 대해 논한다. 예를 들어 표 1에서 동사 '오다'에 대해 다양한 명사구가 말뭉치에서 발생하는데 이들의 많은 부분이 실제로는 매우 낮은 빈도 즉, 우연의 결과로서 발생함을 볼 수 있다. z-test는 이러한 두 가지의 발생을 매우 간단한 수식을 통해 구분할 수 있는데 다음의 식에 의해 표현된다.

$$strength = \frac{freq_{vnp_i} - \overline{f_{vnp}}}{\sigma_{vnp}} \geq k$$

여기서 $freq_{vnpi}$는 어떤 동사 $v$와 명사구 $np_i$간의 빈도이고 $\overline{f_{vnp}}$는 그 동사 $v$와 명사구 $np$들간의 평균 빈도수이며 또 $\sigma_{vnp}$는 이들의 표준 편차이다. 즉 이들이 발생하는 평균에 얼마 이상 발생하는 것들은 우연으로 보이는 것보다 훨씬 자주 발생하는 것으로 볼 수 있으며 이들을 연어로 간주한다. $k$값은 실험에 의해 정해질 수 있는데, 구문 분석에의 응용을 목적으로 이들을 추출한 본 실험에서는 자연어 발화의 지역성을 감안하여 비교적 낮은 값으로 정하였다. 표 1에서 'O'로 표시된 부분이 연어로서 선택된 부분이다. 또한 이러한 처리의 결과로 총 240,548개의 연어가 추출되었으며 이들의 총 빈도는 3,402,881이다. 한편 앞서 언급한 바와 같이 추출된 <서술어,명사,조사> 쌍의 총 수가 약 300만 쌍이며 이들의 총 빈도수가 7,244,729에 이르는데, 이는 공기 데이터로부터 추출된 약 8%의 연어가 추출되었음을 의미한다. 또 이들 연어의 총 빈도가 전체 공기 데이터 빈도에서 약 47%의 비율을 차지하고 있다. 이는 자연어의 발화에서 특정 단어들의 공기가 두드러짐을 보여준다.

이와 같이 추출된 연어는 구문 분석 뿐만 아니라 자연어 처리의 여러 분야에서 응용 가능한데 우선 기계 번역 분야를 생각할 수 있다. 개별적으로 많이 발생하는 연어는 기계 번역에서 지역적인 구의 번역에 큰 기여를 할 수 있어 번역의 질을 높이는데 유용할 것으로 생각된다. 본 연구의 결과는 술어 논항이라는 구문 요소를 포함하고 있으므로 기계 번역과 같은 분야에 직접 응용이 가능할 것으로 기대된다. 특히 이들이 전체 동사구에서 차지하는 비율을 감안하더라도 이들의 유용성을 생각할 수 있다. 또한 자연스러운 자연어 생성에도 그 효용 가치가 높다. 서로 어울려 많이 발생하는 단어들을 선택하는 것이 자연스러운 생성을 위해 필수적임은 말할 나위 없을 것이다.

## 3.2 구문 분석에서의 연어 정보 이용

구문 분석이 실제 응용 시스템에 이용되기 위해서는 정확성을 잃지 않으면서도 효율성을 유지할 수 있어야 한다. 만일 실제 파싱에 참여되는 단어를 그룹화함으로써 전체 노드의 수를 줄여 전체 가능 탐색 공간을 줄일 수 있다면 분석의 부담을 상당히 줄일 수 있을 것이다. 이를 위해 본 논문에서는 연어 정보를 이용하여 파싱의 전 단계에 어절과 어절에 대해 연어 결합을 주고 이에 대해 실제 파싱 단계에서 다른 결합에 대한 가능성을 검사하지 않을 수 있도록 하였다. 이때 연어 관계가 할당되는 명사구와 동사 사이에는 다른 동사가 존재하지 않는 지역성을 가지고 있는 것들에 한한다. 본 논문에서는 이들을 통사적 연접이라 정의한다.

**[통사적 연접]** 통사적 연접이란 하나의 보어와 머리어 사이에 그 머리어와 동일한 범주를 가지는 머리어가 존재하지 않는 것을 말한다.

이에 의하면 문장 '학교에 가면서 노래를 부른다'에서 명사구 '학교에'와 서술어 '가면서'는 통사적으로 연접하나 '학교에'와 '부른다' 사이에는 동사 '가면서'가 존재하므로 통사적으로 연접하지 않는다. 따라서 실험에서는 '학교에'와 '가면서' 그리고 '노래를'과 '부른다'에 대한 연어 가능성을 검사한다.

## 4 격 중의성 해결

한국어 처리에서 격 중의성은 의미 분석 뿐만 아니라 기계 번역시 번역어 선택 등 거의 모든 부분에서 나타나는 중요하게 다루어져야 하는 현상이다. 한국어에서 격 중의성은 세 가지 경우에 나타난다. 첫째, 보조사로 인한 명사구의 격 중의성이 있고 둘째, 조사의 생략으로 인해 격 중의성이 발생하며, 마지막으로 관계 관형절에서 명사구의 이동으로 인해 발생되는 명사구 복원시의 격 중의성이 있다. [13,14]에서는 주로 보조사나 조사 생략에 의한 격 중의성 해소를 시도하였다.

본 연구에서는 이러한 연구를 확장하여 관계 관형절의 이동된 격의 복원시 발생하는 격 중의성을 해소하고자 하며 또한 중의성을 발생시키는 격에 대해 주어와 목적어 뿐만 아니라 부사어도 가정하였다. 이를 위해 2장에서 언급된 서술어와 명사구의 공기 정보를 기반으로 다음과 같이 연관도를 정의한다.

$$Assoc(v,n,p) = \lambda_1 P(n,p \mid v) + \lambda_2 P(p \mid v) P(p \mid n)$$

식에서 $v$, $n$, $p$는 각각 동사, 명사 및 관계된 조사를 의미하며 $P(n,p|v)$는 하나의 동사에 대해 명사구가 얼마나 자주 공기하는지를 표현하며 $P(p|v)$와 $P(p|n)$은 $P(n,p|v)$를 보완하기 위한 보완식으로서 각각 동사가 격 관계를 얼마나 요구하는지 그리고 주어진 명사가 가정될 수 있는 조사를 취할 가능성이 얼마나 되는지 표현한다.

## 5 실험

본 연구에서는 실험을 위해 학습 데이터로 이용된 말뭉치로부터 분리된 텍스트로부터 실험 문장을 추출하였다. 먼저 연어를 구문 분석에 이용한 결과를 살펴 본다.

본 실험에서는 먼저 실험 문장으로부터 서술어와 명사구의 공기쌍 250개에 대해 통사적으로 연접한 데이터들이 얼마나 연어로 인식되어 하나로 묶일 수 있는지 살펴보았다. 단, 주제화 조사 '은/는'을 포함하는 명사구는 실험 대상에서 제외하였다. 표 2는 그 결과를 보여 준다.

표 2 연어를 구문 그룹화에 적용한 결과

| 공기쌍 | 연어수 | 정확도 |
|---|---|---|
| 250 | 101 | 98.0% |

표에서 보는 바와 같이 실험 공기쌍들로부터 연어로 인식될 수 있는 것들의 수는 101개로 약 40.1%였다. 따라서 40%의 쌍은 실제 파싱에 앞서 묶어 줌으로써 전체 파싱 복잡도를 줄일 수 있다는 결론이다. 이들의 첫째 조건은 연어에 의해 인식된 어절쌍을 묶어도 오류가 발생하지 않는다는 것이다. 그러나 실제로는 2개의 오류가 포함되어 있는데 이들은 모두 조사 '에서'와 관련되어 있다. 조사 '에서'를 취한 명사구들은 다른 조사를 취한 명사구들과는 달리 거리가 멀리 떨어져 있는 서술어와 결합하는 경향이 두드러지다는 특징을 가지고 있었다. 이런 개별 형태소와 관련된 통사적 특징은 분석의 질을 보다 향상시키기 위해 향후 연구되어야 할 과제이다.

둘째로 주어진 실험 문장으로부터 격 중의성 해결에 대한 실험을 하였다. 여기서 형용사의 관형형의 경우에는 모두 주어의 이동으로 볼 수 있으므로 제외하였다. 이들을 제외하면 256개의 격 중의성이 발생하였는데, 이들에 대한 실험 결과는 표 3과 같다.

표 3 격 중의성 해결에 대한 실험 결과

| 격중의성 발생 회수 | 성공 | |
|---|---|---|
| 256 | 수 | 백분율 |
| | 223 | 87.1 |

표에서 보는 바와 같이 평균적으로 87.1%의 정확성을 보였다. 오류의 가장 큰 원인은 보조사 '은/는'으로 인한 것인데, 이는 시간 관계의 부사가 포함되기도 하지만, 실제로 대부분의 경우에 주어로 쓰이기 때문이었다. 특히 분야별로 사용 환경이 다른데, 분야에 따라 선택 제약이 깨지는 현상이 발생하기 때문이다. 따라서 이의 해결을 위해서는 언어학적 고찰 뿐만 아니라 분야별 말뭉치의 수집을 통한 공기 정보 구축 등이 필요할 것으로 생각된다.

## 6 결론

본 논문에서는 말뭉치로부터 추출한 공기 데이터가 구문 분석에서 어떻게 이용될 수 있는지를 보였다. 실험에서 볼 수 있는 바와 같이 대량의 원시 말뭉치로부터 구축된 공기 데이터는 다양한 방법으로 구문 분석에 응용될 수 있으며 또한 효과적이고 효율적인 분석을 가능하게 함을 볼 수 있었다.

이는 향후 분야에 따른 지식을 구축할 경우 보다 좋은 결과를 가져 올 수 있으리라 기대된다. 분야마다 선택되는 용어는 매우 다른 측면을 보이며 따라서 분야별 말뭉치의 수집이 요구되며 그에 맞는 공기 데이터를 구축할 경우 보다 나은 구문 분석 결과를 기대할 수 있으리라 생각된다.

또한 문장의 구조 분석이라는 측면에서 볼 때 이와 같은 연구를 바탕으로 구문 구조 말뭉치로부터 통계 학습에 의해 얻어진 언어 지식을 함께 이용한다면 보다 정확하고 효율적인 파서의 구축이 가능하리라 생각된다.

## 참고문헌


[1] Church, K., W. and Hanks, P., Word Association Norms, Mutual Information, and Lexicography, In *Proceedings of 27th Annual Meeting of ACL*, 1989

[2] Collins, M. J., A New Statistical Parser Based on Bigram Lexical Dependencies, In *Proceedings of the 34th Annual Meeting of ACL*, 1996

[3] Dunning, T., Accurate Methods for the Statistics of Surprise and Coincidence, Computational Linguisitics, Vol. 19(2), 1993

[4] Eisner, J. M., Three New Probabilistic Models for Dependency Parsing: An Exploration, In *Proceedings of COLING '96*, 1996

[5] Hindle, D. and Rooth, M., Structural Ambiguity and Lexical Relations, Computational Linguistics, Vol. 19(1), 1993

[6] Kjellmer, G., Patternsof Collocability, Theory and Practice in Corpus Linguistics, edited by Aarts, J. & Meijs, W., Rodopo, B. V., Amsterdam, 1990

[7] Lauer, M., Corpus Statistics Meet the Noun Compound: Some Empirical Results, In *Proceedigs of 33rd Meeting of ACL*, 1995

[8] Magerman, D., Statistical Decision-Tree Models for Parsing, In *Proceedings of the 34th Annual Meeting of ACL*, 1995

[9] Smadja, F., Retrieving Collocations from Text: Xtracts. Computational Linguistics, 1993

[10] Yoon, J., Kim, S. and Song, M., New Parsing Method Using Global Association Table, In *Proceedings of 5th International Workshop on Parsing Technology*, 1997

[11] Yoon, J. and Song, M., Yet Another Compound Noun Analysis, In *Proceedings of NLPRS '97*, 1997

[12] 강현화, 동사 연결 구성의 다단계성에 대한 연구, 'V어-V' 구조를 중심으로, 연세대학교 국어국문학과, 박사학위 논문, 1996

[13] 김선호, 윤준태, 송만석, 통계를 기반으로 한 어휘 관계 학습, 한국정보과학회 봄 학술발표 논문집, 1996

[14] 양재형, 김영택, 통계 정보를 활용한 한국어 미지격 명사구의 문법 기능 결정, 한국정보과학회 논문지, 제21권, 제5호, 1994