# 확률적 CFG 파싱을 활용한 한국어 복합명사 구조 분석의 중의성 해소

김동성
고려대학교
dsk202@korea.ac.kr

# Disambiguation on the Analysis of Korean Complex Nominals, Using Probabilistic CFG Parsing

Dong-Sung, Kim
Korea University

**요 약**

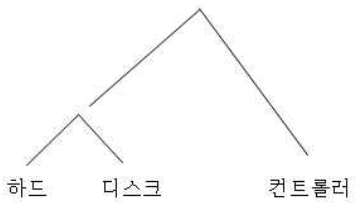
본 논문은 한국어 복합명사 구조의 분석을 목적으로 한다. 연구는 이론 언어학뿐만이 아니라 정보처리, 정보검색과 같은 언어의 전산적 처리에서도 중요한다. 복합명사 구조는 크게 외심구조와 내심구조로 나뉘며 내심구조의 경우에 좌분지나 우분지 구조로 분석이 되어야 하는 중의성이 있다. 기존의 Lauer 모델은 사전적 정보에서 발견되는 확률 정보를 구조 정보에 연결하기 위한 모델로 의존모델과 인접모델을 제시하였다. 본 연구에서는 구조에 기반을 둔 확률정보를 결합하기 위한 확률적 CFG 파싱 방법을 활용하고자 하였다. 이를 위해서 실제 코퍼스상에서 발견되는 복합명사 패턴을 대상으로 구조적 분석을 화자 직관을 통해서 진행하고, 이를 다시 Lauer 모델과 확률적 CFG 파싱 방법 응용과 비교해 보았다. 결과적으로 화자 직관에 가장 일치한 예측을 하였으며, 구조에 대한 정보 해석이 가능하였다.

주제어: 형태론, 코퍼스, 전산형태론, 복합 명사구 분석

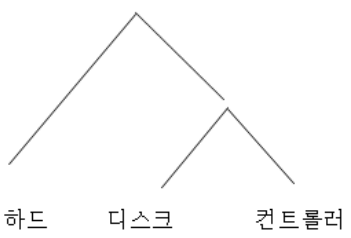
## 1. 서론

구조적으로 분석이 가능한 복합명사는 [1]에 따르면 이분지 제약(Binary Branching Constraints)에 따라 두 개의 서로 다른 구조적 분석이 가능하다. 예를 들어서 '하드 디스크 컨트롤러'라는 복합명사구는 (1a, 1b)와 같은 괄호매김의 중의성(Bracketing Ambiguity)이 가능하다. (1a, b) 중 일반적인 지식에 의해서 (1a)의 구조만을 선택한다.

(1) a.




b.




기존 연구는 시소러스(Thesaurus)나 워드넷(WordNet)과 같은 언어 자원을 활용하는 연구[2, 3, 4]와 코퍼스에서 발견되는 특정 통계 정보를 활용하는 연구[5, 6, 7, 8, 9, 10]들이 있는데, 언어 자원에서 발견되는 통계량을 활용한다.

통사적으로 (1a, b)는 각각 좌분지(left branch), 우분지(right branch) 구조이므로 두 개의 서로 다른 규칙에 대한 파싱 작업으로 분석할 수 있다. 본 연구에서는 중의성 해소를 위해서 확률적 CFG 방식에 따라서 분석하고자 한다. 확률적 CFG 파싱 방식을 위해서 핵이 되는 어휘를 상위 마디 비단말 어휘정보로 활용해서 확률 정보를 활용하였다. 이 방식은 구조적 정보를 활용한 중의성 해소 방안이다.

논문의 구성은 다음과 같다. 2절은 관련연구에 대한 소개이다. 3절은 복합명사 구조 분석에 대한 언어학적 논의이다. 4절은 확률적 CFG 파싱 방식을 소개하고 5절에서는 이를 활용한 실험과 결과이다. 5절은 이 논문의 결론이다.

## 2. 관련연구

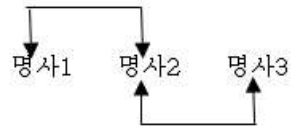
복합 명사구조 분석은 언어학적, 정보검색, 음성합성이나 음운연구에서 중요하다. 언어학적으로 명사구 합성은 생산성이 매우 높은 형태론적 과정으로 새로운 구조에 대한 언어학적 예측이 필요하다[11]. 정보검색의 측면에서 색인어 목록의 작성에서 복합명사 분석은 중요하다[5]. 또한 음성합성이나 음운론 연구의 경우에 띄어

읽기는 운율의 단위를 결정하는 주요 요소인데, 복합명사구의 구조적 차이가 운율 결정의 매우 중요한 요소이다[12].

서론에서 논의한 바와 같이 중의성 해소를 위해서 확률 및 통계량을 활용한 여러 모델들이 제시되었다. 이 중 [4]는 구조적 중의성을 해결할 두 개의 다른 모델을 제시하였다. 하나는 인접모델(Adjacency Model)이고 다른 하나는 의존모델(Dependency Model)이다. 각각은 (2a, b)와 같다. 두 개의 다른 모델을 활용해서 중의성을 해결하였는데, 영어의 경우에 인접모델이 의존모델보다 더 정확하게 예측한다고 주장하였다.

(2) a. 인접모델


b. 의존모델




확률 및 통계량을 활용한 연구 방식들은 구조에 대한 분석이 아닌 어휘 정보에 대한 연산을 활용하였다. 서론에서 논의한 바와 같이 구조적으로 좌분지 구조와 우분지 구조에 대한 선택이 중의성 해소 방식이다. 따라서 구조적 방식에 대한 연구가 필요하다.

본 연구에서는 이러한 점을 고려해서 구조적 중의성 해소 방식을 도입한다. 중의성 해소 방식을 통사구조를 통해서 해소하는데, 확률적 CFG 방식을 활용한다. 이 방식은 각 분지의 확률정보를 적용해서 어느 확률이 더 높은지를 연산한다[13].

## 3. 복합명사 구조 분석

복합 명사구조에 대한 분석은 크게 핵이 내부에 있는 내심구조와 내부에 없는 외심구조를 구분한다. 핵이 내부에 없는 외심구조의 경우는 고유 명사화된 경우가 많으므로, 구조적 중의성이 없다. 예를 들어서 'England-France'나 'mother-child'와 같은 외심구조는 내부에 핵이 없으므로 서로 동격 관계이며, 'Mr. Big'이나 'President Clinton' 또는 '오바마 대통령'과 같은 복합구조도 핵이 없는 구조이다. 또한 '남녀 고용 평등법'과 같은 구조는 일반 명사의 연쇄이지만 고유명사화된 것으로 어휘화된 항목으로 외심구조이다.

내심구조의 경우에는 핵-수식 관계, 술어-논항 관계와 같은 통사적 관계와 유사성을 분석한다. 내심구조는 이분지 제약에 따라서 두 개씩 짝지어지는 구조를 만들게 되지만, 핵이 내부에 없는 외심구조는 이분지 제약을 따르지 않는다. 내심구조는 괄호매김의 모호성으로 인해서 구조적 관계를 설정해야 하지만, 외심구조는 고유 명사화된 것처럼 행동하며 구조적 중의성이 없다. 반면에 내심구조는 핵이 내부에 있는 구조로 구조적 중의성이 발견된다. 내심구조의 구조적 관계는 핵-수식어 관계로 구성되었는지, 아니면 논항-술어 관계로 구성되었는지에 따라서 구분될 수 있다. 언어학적으로 [명사1 명사2 명사3]의 복합명사는 표 1과 같이 8가지의 유형이 있다.

표 1 복합명사 구조 유형

| 유형 | 형태 | 예 |
|---|---|---|
| 1 | [[외심 N1 N2]수식 N3핵어] | 협동 조합 중앙회 |
| 2 | [[외심 N1 N2]논항 N3핵어] | 정치 자금법 위반 |
| 3 | [[N1수식 N2핵어]수식 N3핵어] | 무역 역조 시정 |
| 4 | [[N1논항 N2술어] N3술어] | 상품 불매 운동 |
| 5 | [N1수식 [외심 N2 N2]핵어] | 개인 고향 방문단 |
| 6 | [N1수식 [N2수식 N3핵어]] | 거대 대국적 기업 |
| 7 | [N1수식 [N1논항 N3술어]] | 국내 제품 불매 |
| 8 | [N1논항 [N2논항 N3술어]] | 검찰 참고인 조사 |

술어-논항 관계에 대한 분석은 술어의 논항구조에 달려 있으므로, 술어의 논항구조 정의를 참조해야 한다. 술어화가 가능한 명사는 불가능한 명사와 다르다. 술어화가 가능한 명사는 서술성 명사로 '-하다'와의 형태론적 조어법이 가능하다. 세종전자사전을 활용하면 서술성명사를 활용할 수 있는데, 서술성 명사인 '-하다'동사에 대한 논항구조가 수록되어 있다. '조사'의 경우에 세종전자사전에는 <Agent(인간 또는 인간집단)-이, Theme(전체)-에대해>와 같이 정의되어 있다. 따라서 '검찰 참고인 조사'의 경우에는 '검찰'이 Agent로, '참고인'이 Theme으로 분석된다.

반면에 핵-수식 관계 분석은 수식어가 소유격화가 되어 꾸며주는 관계가 성립한다. 예를 들어서 '주부 살림살이'는 '주부의 살림살이'와 같이 소유격 접미사 '-의'가 수식어에 접사화 과정을 겪을 수 있다. 또한 한자어의 경우에는 '-적'의 접미사 부착이 가능하다. 핵-수식어 과정은 생산성이 매우 높은 형태론적 과정이다. 따라서 언어 자료인 사전이나 코퍼스에서 발견되지 않는 예가 더 많다. 따라서 많은 경우에 핵-수식 관계를 가능하다고 가정해야 할 경우가 있다. 따라서, 예를 들어서 '폭력 교실'과 같은 구조가 '폭력적 교실'로 해석이 가능하더라도 언어 자료에는 나타나지 않는다. 이러한 경우에도 핵-수식 관계를 설정하는 것은 핵-수식 관계의 생산성의 자료에서 발견되는 것보다 더 높기 때문에, 직관적으로 관계성을 포착하는 것이 더 적절하기 때문이다.

그 외에 외심구조로 등위적 합성구조가 있는데, 예를 들어서 '전자 정보 통신'과 '문화 체육 예술'과 같이 명사를 나열한 구조이다. 이러한 구조는 내부 구조에 괄호매

김의 중의성이 없다.

## 4. 확률적 CFG 파싱

주어-동사가 있는 문장에서 서술어가 각 논항에게 부여하는 논항 구조가 합성어의 경우에도 가능하다고 [14]는 주장하였다. 이러한 주장은 명사 합성 구조에 있어서 통사적 규칙의 적용이 가능하다는 것을 의미한다. CFG 규칙을 통사적 파싱에 적용하였듯이 복합명사 구조 분석에도 적용할 수 있다는 것을 의미한다.

일반적으로 확률적 파싱 방식에 의한 방식은 파스 트리 T에 대한 문장 S의 확률의 결합 확률(Joint Probability)인 P(T, S)는 (3)과 같이 연산한다.

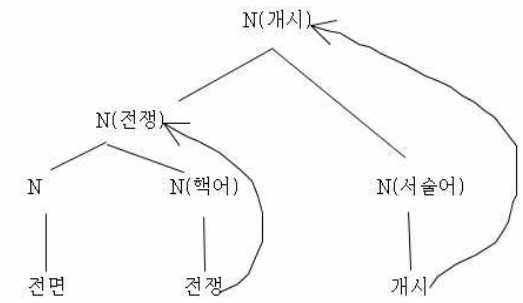
(3) $P(S, T) = P(T)P(S|T)$

(3)의 모델은 서로 다른 규칙들의 확률 연산 중 최고가 되는 규칙들을 (4)와 같이 찾아내게 된다.

(4) T(S) = agrmax P(S, T)
$\approx$ agrmax $\prod P(RHS|LHS)$

명사구 합성의 경우에 핵이나 서술어가 어휘가 분지 구조의 핵으로 작동한다. 문법 이론적으로도 이러한 성분들이 구조의 핵이 되므로, 이러한 어휘들을 대상으로 상위 구성성분으로 대체해서 어휘화된(lexicalized) CFG를 구성하면 좌분지 구조의 경우에 (5)와 같은 정보를 얻을 수 있다.

(5)




(5)에서 '전쟁'에 해당하는 요소는 핵어로 '전면'의 수식을 받는다. 또, '개시'의 경우에 이 요소는 서술어로 [전면 전쟁]을 술어로 취하는데, 핵어인 '전쟁'을 Theme으로 취한다. 이 경우에 각 구성 성분들 중 핵어나 서술어가 되는 요소들의 정보를 위쪽 마디 어휘 정보로 활용하면 다음과 같은 비단말 CFG 규칙이 만들어진다.

(6) R1: LHS1 (N(개시) →
RHS1 (N(전쟁) N(개시))
R2: LHS2 (N(전쟁)) →
RHS2 (N(전면) N(전쟁))

(6)을 활용하면 (7)과 같은 확률 CFG 연산이 가능하다.

(7) P(RHS1|LHS1) × P(RHS2|LHS2)
$= \frac{P(RHS1 \sqcap LHS1)}{P(LHS1)} \times \frac{P(RHS2 \sqcap LHS2)}{P(LHS2)}$
=P(전쟁|개시)×P(전면|전쟁)

다시 정리하면, 좌분지 구조인 (8a)는 (8b)로 우분지 구분조인 (9a)는 (9b)로 연산된다.

(8) a. [[N1 N2] N3]
b. P(N1|N2)×P(N2|N3)
(9) a. [N1 [N2 N3]]
b. P(N1|N3)×P(N2|N3)

핵심어나 중심어가 되는 단어를 상위 노드나 최상위 노드로 바꾸어주면, 좌분지 구조와 우분지 구조마다 서로 다른 유형의 확률적 CFG 파싱 방식이 생성이 된다. (2)에서 [4]는 인접 모델이나 확률 모델 중 하나가 중의성 해소 작업에 활용될 수 있다고 주장하였으나, 확률적 CFG 파싱 방식을 활용하면 좌분지 및 우분지 구조마다 서로 다른 모델을 적용되는 것이 구조적으로 산출된다. [4]의 논의를 활용하면 좌분지 모델의 경우에는 인접모델이 우분지 모델의 경우에는 의존모델을 활용하는 것이 적절한 것으로 보인다.

## 5. 실험 및 결과

연구를 위해서 추출한 자료는 형태소 분석이 된 세종 천만코퍼스를 대상으로 [명사1 명사2 명사3]의 어휘 군집을 모두 추출하였다. 이를 대상으로 외심구조가 될 가능성이 있는 고유명사가 포함된 경우를 제외하였다. 또한 생산성이 너무 낮은 경우는 고려하지 않기 위해서 출현 빈도가 모두 5회 이상이 경우로 제한을 하였다. 이후에 모두 1,000여개가 수집이 되었는데, 화자 직관을 활용해서 구조 분석을 한 자료를 얻기 위해서 2명의 화자에게 동일한 목록의 [명사1 명사2 명사3] 어휘 군집을 보여주고 어느 괄호매김이 타당한지를 표시하게 하였다. 실험에서는 외심구조인 것과 좌분지, 우분지 구조인지를 화자 직관으로 구분하게 하였다. 결과적으로 화자 직관 실험을 통해서 얻어진 자료 중 화자간 불일치 중 문맥적 모호성을 지닌 경우, 특이한 문제가 없이 예측이 되지 않는 경우를 제외하고 자료를 정리해서 270개의 자료를 최종적인 연구 자료로 선택하였다.

실험은 화자 직관을 활용한 실험 결과와 비교해서 어떠한 처리 결과가 화자 직관과 가장 일치하는 지를 찾아보았다. 먼저, [4]의 제안에 따라서 (2)의 모델을 적용해서 어느 모델이 가장 일치하는지를 조사하였다. 세종 천만코퍼스 상에서 발견되는 빈도 및 확률 정보를 대상으로 [N1 N2 N3] 괄호매김이 [[N1 N2] N3]이고, [N1 N2]를 예측하는 규칙이 R1이 't1→t2'이고, 상위 구조 [[N1 N2] N3]를 예측하는 규칙 R2가 't2→t3'라면 의존모델은 (10a)를 인접모델은 (10b)를 적용하였다. 여기서

확률정보는 $P(t1 \rightarrow t2) = \frac{count(w1w2)}{count(w1)count(w2)}$ 와 같이 추출한다.

(10) a. $R_{\text{인접모델}} = \frac{P(t1 \rightarrow t2)}{P(t2 \rightarrow t3)}$

b. $R_{\text{의존모델}} = \frac{P(t1 \rightarrow t2)}{P(t1 \rightarrow t3)}$

인접모델과 의존모델은 각각 75.7%와 80.1% 정도로 화자 직관의 결과와 일치하였다. (8, 9)에서 제시한 모델을 활용하면 화자 직관과 전체적으로 90.8% 정도 일치한다. 표 1과 관련되어서 각각의 유형에 대한 일치를 보면 그림 1과 같다.

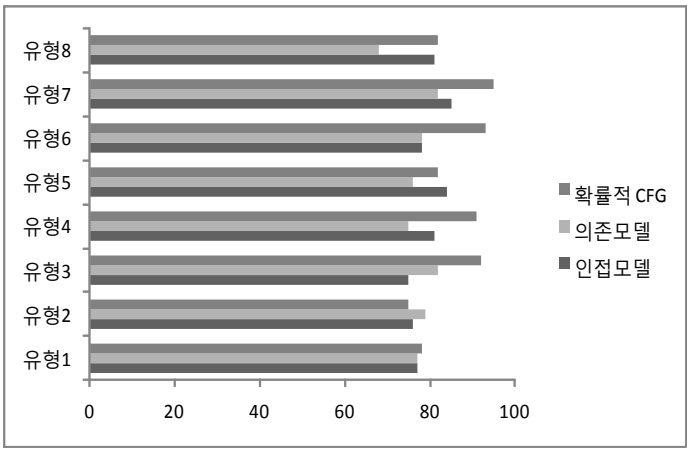



그림 1 유형별 일치도 (단위 %)

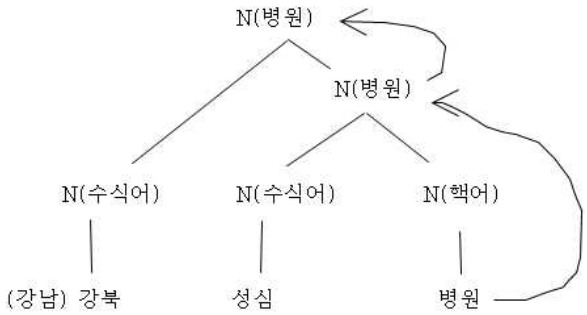
그림 1에서 보여준 것처럼, 전반적으로 확률적 CFG 모델이 [4]의 모형보다 더 정확하게 예측한다. 또한 구조적으로 외심구조의 경우보다 내심구조를 더 정확하게 예측한다. 외심구조는 어휘화된 항목으로 하나의 단어처럼 행동한다. 이에 비해서 내심구조는 구조적으로 예측이 되어야 할 부분이다. 내심구조의 경우에 구조적으로 핵-술어, 술어-논항구조와 같은 구조를 더 잘 예측하며, 전체 논항 구조가 발견되는 구조인 [Agent [Theme 술어]]와 같은 구조의 경우에는 정확도가 가장 낮다. 이를 통해서 보면 외심구조는 내심구조와 다르게 처리되어야 하며, 다중 어휘인지를 먼저 검사해야 한다는 것을 보여준다.

[4]에서 제시된 모델은 확률정보를 통사구조 모델과 결합한 것으로 좌분지 및 우분지 구조를 결정하기 위한 모델이다. 따라서 구조적 핵어 성분에 대한 분석이 없고, 괄호매김의 정보가 좌분지인지 아니면 우분지 인지만을 결정한다.

인접모델이 의존모델보다 더 정확하다면 괄호매김안에서의 결합력이 구조적 결합력보다 더 높다는 것을 말한다. 예를 들어서 [검찰 [참고인 조사]]와 같은 합성구조에서 [참고인 조사]의 결합력이 [검찰 참고인]보다 더 높다는 것으로 나타난다는 것을 의미한다. 반면에 의존모델이 인접모델보다 더 정확하다는 것은 구조적으로 어떠한 구조인지에 대한 정보가 더 중요하다는 것을 의미한다. 즉 [참고인 조사]의 구조에서 '조사'라는 서술성 명사가 구조적으로 '검찰'과 어떠한 관계성을 갖는지가 더 중요하다는 것을 의미한다.

구조에 대한 정보는 의미적 정보를 포함한다. 예를 들어서 [강남 [성심 병원]]과 [강북 [성심 병원]]은 [성심 병원]의 부류로 [성심 병원]이라는 범주에 포함되어져야 한다. 이 경우에 핵어는 '병원'이고 '강남'이나 '강북'은 [성심 병원]의 수식어로 작용한다. 이와 같이 핵어에 대한 정보가 필요하며, 이러한 정보는 (11)과 같이 상위 분지 마디로 전달되어야 한다.

(11)




인접모델이나 의존모델의 경우에는 구조적 해석을 통한 의미해석을 시도하지 못한다. 유형별 일치도를 나타내는 그림 1을 살펴보면 유형 2의 경우에는 의존모델이 다른 모델보다 더 나은 예측을 하지만 다른 유형의 경우에는 두 모델이 확률적 CFG 모델보다 더 나은 예측을 하지 못한다. 자세히 결과를 분석하면, 확률적 CFG의 경우에는 외심구조가 포함된 구조 해석은 전반적으로 아닌 경우보다 예측력이 떨어진다. 그러나 의존모델이나 인접모델의 경우에는 예측력이 일관성 있게 나타나지 못한다. 따라서, 일관성의 측면에서 구조해석을 통한 의미해석이 두 모델에서는 가능하지 않는 것으로 보인다.

외심구조는 연이은 단어들이 하나의 의미로 사용된 경우가 많은데, 특정한 경우에는 고유명사화한 경우가 있다. 예를 들어서 '협동 조합, 고향 방문단, 하드 디스크'는 고유명사화된 어휘들이다. 또한 연어(Collocation)와 같이 다중어 표현(Multiword Expression) 형식으로 여러 개의 단어들이 연결되어 활용한 경우이다. 이러한 외심구조는 앞에서 기술한 바와 같이 구조적 특성을 발견할 수 없는 구조이다. 따라서 구조적 특성을 활용하는 확률적 CFG 모형에서 제안한 모형에서는 포착하기 어렵다. 특히 본 연구에서 제안한 모형에서는 일관적으로 외심구조를 포착하는데 어려운데, [4]에서 제안한 모형은 비일관적으로 예측한다.

외심구조에 대한 예측 모형은 구조모형이 아닌 연어측정 모형이 더 정확하게 예측한다. 특히 코퍼스 통계를 활용한 모형이 더 적절하다. [17]에서 논의한 여러 모형 중 예를 들어서 t-test, z-test, ch-squared test와 같이 통계적 검증 모형이 적절하다. 통계적 모형은 통계적 유의 수준에 따라서 검정을 하므로, 가설에 대한 입증 및 유의 수준의 조절을 통한 적절한 수준을 추출하는 장

점이 있으므로, 더 객관적인 추출방식을 제공한다.

본 연구에서 진행한 의미 중의성 해소 작업에서 어려운 점은 두 개의 다른 구조가 주위 문맥에 의존한 경우이다. 예를 들어서 (12a, b) 구조가 모두 가능한데, 주변 문맥에 의존되어서 해석되어야 한다.

(12) a. [[자동 지급] 금지]
b. [자동 [지급 금지]]

여기서 문맥이란 해당 문서나 해당되는 지역 문맥(local context)을 말하는데, 전체 코퍼스에 해당되는 전역 문맥(global context)으로 해석하면 전혀 다른 해석이 된다. 연구에서는 화자 직관 실험을 통해서 중의성이나 지역 문맥에 의존한 경우를 제거하여서 전역 문맥으로 해석되는 경우를 고려하였다. 이는 일반적으로 해석될 수 있는 경우만을 실험의 대상으로 하기 위함이다.

## 6. 결론

본 연구에서는 [N1 N2 N3]의 경우에 [[N1 N2] N3], [N1 [N2 N3]]의 두가지 다른 구조적 선택에 대한 중의성을 해결하고자 하였다. 기존의 연구에서 Lauer는 의존모델과 인접모델이라는 두 가지 유형의 모델을 제시하였다. 이 모델은 전반적으로 확률모형을 구조모형에 적용한 것이므로, 전체적으로 구조적 유형에 따른 예측이 아니다. 따라서 구조적 정보보다는 실질적으로 단어의 확률 정보를 활용해서 이를 구조적 유형 예측 모델과 결합하였다. 구조에 대한 정보는 통사적 정보인 논항구조 및 의미 정보까지 포함하기 때문에 이에 대한 접근이 필요하다.

본 연구에서는 확률적 CFG 모델을 적용하였다. 제시한 모델이 인접모델과 의존모델보다 더 정확하게 예측을 하였으며, 구조에 대한 해석 정보를 확률로 표현하였다. 따라서 [4]의 모델보다 더 향상된 방식이라고 할 수 있다.

실제 언어현상에서 [N1 N2 N3]보다 더 많은 수의 명사들이 연쇄를 이룰 수 있으므로, 본 연구에서 진행한 연구보다 더 복잡한 패턴에 대한 연구가 향후 필요하다. 또한 정제된 사전을 통한 논항 구조의 분석, 복합명사 구조의 유형에 대한 예측과 같이 더 복합적이고 복잡한 부분에 대한 연구가 필요하다.

또한 어휘화된(lexicalized) 확률적 CFG 방식이나[15] 기타 쪼개기(split) 방식과[16] 같이 다양한 유형의 향상된 확률적 CFG 방식을 활용한 연구도 향후 진행해야 할 것이다. 또한 연구에서 화자 직관을 활용해서 정제된 유형의 적은 데이터만을 활용하였는데, 실제 언어 현상을 대표할 많은 양의 데이터를 처리하는 것도 필요하다. 또한 5절에서 지적한 바와 같이 문맥 정보에 대한 다양한 연구도 필요하다.

## 참고문헌


[1] Spencer, A. (1993) Mophology. Blackwell Publishing.

[2] Baldwin, T., Bannard, C., Tanaka, T., and Widdow, D., (2003) An empirical model of multiword expression decomposability. In Proceedings of ACL-2003, 89-96.

[3] Jiang, J., and Conrath, C. (1999) Multi-Word Complex Concept Retrieval via Lexical Semantic Similarity. International Conference on Information Intelligence and Systems (ICIIS'99).

[4] Lauer, M. (1995) Corpus statistics meet the compound noun: Some empirical results. In Proceedngs of ACL-1995.

[5] 원형석, 박미화, 이근배.(2000) 복합명사 분할과 명사구 합성을 이용한 통합 색인 기법. 정보과학회논문지, 27(1), 84-94.

[6] 윤보현, 조민정, 임해창. (1997). 통계 정보와 선호 규칙을 이용한 한국어 복합명사의 분해. 정보과학회논문지, 24(8), 900-909.

[7] Pereira J,, and Lopes, G. (1999) A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. Sixth Meeting on Mathematics of Language, 369-381.

[8] Nicholson, J. and Baldwin, T. (2006) Interpretation of Compound Nominalisations using Corpus and Web Statistics. In Proceedings of the COLING/ACL 2006. 54-61.

[9] Park, H., Han, Y., Lee, K., and Choi, K. (1996) A probabilistic approach to compound noun indexing in Korean texts. In Proceedings of the 16th conference on computational linguistics, 514-518.

[10] Yoon, J., Choi, K., and Song, M. (2001) A corpus-based approach for Korean nominal compound analysis based on linguistic and statistical information. Natural Language Engineering 7, 251-270.

[11] Levi, J. (1978) The syntax and semantics of complex nominals. New York: Academic Press.

[12] Lieber, R. (1983) Argument linking and compinding in English. Linguistic Inquiry 14. pp. 251-286. MIT Press.

[13] Charniak, E. (1993) Statistical Language Learning. MIT Press.

[14] Di Scullo, A. and E., Williams (1987) On the Definition of Word. MIT Press.
[15] Collins, M. (1999) Head-Driven Statistical Models for Natural Language Parsing. Ph.D. Dissertation, Univ. of Pennsylvania.
[16] Klein, D. and C. Manning (2003) Accurate unlexicalized parsing. In HLT-NAACL03.
[17] 김동성. (2011 예정) 언어 자료를 활용한 한국어 복합명사 구조 분석. 언어학, 19(3).