# 한국어 통사 및 의미 정보를 활용한 명사구 인식[1]

한규열○, 안광모, 서영훈
충북대학교 컴퓨터공학과
sept102○@nlp.chungbuk.ac.kr, ahnmo@chungbuk.ac.kr, yhseo@chungbuk.ac.kr

# Shallow parser using Korean information[1]

Kyou-youl Han○, Kwang-mo Ahn, Young-Hoon Seo
Dept. of Computer Engineering, Chungbuk National University

**요 약**

본 논문에서는 한국어 문장의 통사적 특성과 제한된 통계정보를 이용한 명사구의 패턴에 의한 명사구 인식에 대해 기술한다. 본 논문의 명사구 인식기는 관형사와 관형격 조사, 관형형 어미에 관련된 패턴의 명사구 인식을 수행하고, 시간과 장소를 나타내는 특정한 명사에 의해 유도되는 명사구를 인식한다. 또한 복합명사 결합의 문제를 의미쌍 간의 결합도의 문제로 분류하고 해결방법을 제시한다. 실험 결과는 본 논문에서 제안하는 통사적으로 확실한 정보와 제한된 통계정보를 이용한 명사구 인식기가 높은 수준의 명사구 인식을 수행한다는 것을 보여준다.

## 1. 서 론

한국어의 일반적인 자연언어처리 단계는 형태소 분석, 구문 분석, 의미 분석, 화용 분석의 단계로 세분화된다. 이중에서 구문 분석의 단계는 형태소 분석된 문장을 입력으로 하여 각 형태소들이 문장 내에서 가지는 역할, 혹은 상호관계를 분석하는 것이다. 그러나 각 문장에 대해 하나의 완전한 분석 구조를 얻는 전통적인 전체 파싱(full parsing) 방법에 근거한 자연언어 파서는, 제약된 영역을 제외하고는 아직도 만족할 만한 성능과 강건성(robustness)을 보이지 못하고 있다. 이와 같은 문제점이 발생하는 이유는 다음과 같다. 첫 번째로, 자연언어는 본질적으로 중의성(重義性, ambiguity)을 갖는다. 두 번째로, 일반적인 구문 분석의 복잡도는 문장의 길이가 증가할수록 그 길이에 지수적으로 증가하게 된다. 이러한 어려움을 극복하기 위해서, 전체 파싱을 수행하기 전에 명사구 인식의 단계를 거치게 된다.

명사구 인식 단계의 목적은 문장 내에서 같은 역할을 수행하는 몇 개의 연속된 단어들이 모인 연속체(chunk)를 인식하여 하나의 명사구로 통합시키는 것이다. 이것은 문장을 서로 겹치지 않는 최소한의 문장 성분으로 쪼개는 것과 같은 의미이다.

(S0) 철수는 그 모르는 세계가 아름답기도 했다.

예를 들어, S0에서 '모르는[타동사(VV+ETM)]'과 '그[관형사(MM)]'는 모두 '세계[명사(NNG)]'의 앞에 위치하고 있다. '모르는[타동사(VV+ETM)]'은 용언으로서 하위범주화 성분을 요구하지만 그 범위가 '그[관형사(MM)]'를 넘어갈 수 없고 명사를 제외한 두 어절 모두 수식어 역할을 한다는 한국어의 통사적 특징에 의해 [그 모르는 세계]가 하나의 명사구로 인식될 수 있다. 이와 같이 명사구 인식을 통하여 인식된 성분들을 하나의 명사와 같이 취급하여 전체 파싱을 수행하는 경우와 명사구 인식의 단계를 거치지 않은 경우의 복잡도와 효율의 비교는 논의 대상이 될 수 없다. 또한 명사구 인식은 그 자체로도 정보 추출의 분야에서 유용하게 사용될 수 있으며, 문장 전체의 구조를 필요로 하지 않는 자연언어처리 시스템에서 활용될 수 있다.

이와 같은 구 단위화(chunking)의 방법론으로는 통계적인 방법과 비통계적인 방법으로 나눌 수 있다. 비통계적인 접근은 문장 내에서 구 단위화가 가능한 통사적으로 타당한 정보를 찾아내는 것이고, 통계적인 접근은 대량의 말뭉치로부터 추출된 통계 정보를 이용하는 방법이다. 비통계적인 접근은 통사적으로 확실한 정보를 이용하여 구의 인식을 수행하기 때문에 신뢰도가 높지만, 넓은 언어현상에 모두 적용될 수 없다는 한계가 있다. 통계적인 접근은 비통계적인 접근으로 적용되지 않는 언어현상에도 적용할 수 있다는 장점이 있지만, 말뭉치가 편중되어 있거나 불충분한 경우에 심각한 자료 부족 문제(sparseness)로 인해 신뢰도가 떨어지게 된다는 단점이 있다.

단위화의 개념은 [1]에서 처음 제안되었다. 여기에서 구 단위를 특정한 통사적 기능을 갖는 단어들의 연속체로 보고 이를 "하나의 중심어를 포함한 겹쳐지지 않는 단어들의 묶음"으로 표현하였다. [2,3]에서는 여러 단계의 유한 상태 오토마타(finite state automata)를 구성하여 부분적으로 파싱을 수행하는데 앞 단계에서는 명칭이나 간단한 명사구를 인식하고 뒷 단계에서는 복잡한 구들이 인식된다. 통계 정보를 사용한 연구로 [4]는 품사 태그가 부착된 단어들에 부분적 구문 구조를 할당하기 위해 최대 엔트로피 모형에 기반을 둔 통계적 기법을 적용하였다. [5]는 변형기반 학습기법을 적용하여 학습 말뭉치로부터 생성된 초기 시스템의 오류를 교정할 수 있

1) 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (IITA-2008-(C1090-0801-0046))

는 규칙을 적용하였다. [6]는 메모리 기반 학습(memory-based learning) 방법에 의한 단위화를 제안하며 학습용 예제들을 트라이 형태로 저장하여 두었다가 입력에 따라 트라이에 저장된 정보를 활용하여 구 단위화의 수행 여부를 결정하였다. 국내의 연구로 동사구 장벽 알고리즘을 이용한 연구[7,8], 구인식과 의존문법에 기반한 구문분석[9], 세 단계에 걸친 한국어 구단위화 인식에 대한 연구[10], 규칙 기반 학습에 의한 명사구 인식[11], 결정트리 학습기법과 최대 엔트로피를 이용한 구인식[12], 기계학습을 통한 한국어 기본구 인식의 성능 향상에 관한 연구[13], 구인식을 이용한 문장의 의존관계 추출 기법[14], 의존명사와 관련된 구인식[15]등이 있다.

본 논문에서 제안하는 명사구 인식은 전체 파싱의 입력으로 사용된다. 그러므로 명사구의 옳지 않은 인식 결과는 전체적인 구문분석의 성능을 저하시키게 된다. 이에 본 논문에서는 통사적으로 확실한 정보를 명사구 인식에 활용하고 제한적으로 통계정보를 적용한다. 형태소 분석 단계에서는 복합명사를 분해하는데 관심이 많지만, 구문분석에서는 이렇게 분해된 복합명사를 다시 결합시킬 필요가 있다. 일반적으로 지금까지의 연구는 연속된 명사들을 하나의 복합명사로 처리하였다. 그러나 한국어에서는 조사의 생략이 빈번하고, 조사의 생략을 감안하지 않은 구문 분석은 옳지 않은 결과를 유도할 가능성이 높다. 연속된 일반 명사는 각각의 의미를 갖고 있으며 그 의미에 따라 복합명사를 구성할 가능성이 높은 쌍과 그렇지 않은 쌍으로 구분될 수 있다.

## 2. 구문분석을 위한 명사구 인식

본 논문에서 제안하는 명사구 인식기는 전체 파싱을 수행하기 이전의 단계에 명사구를 인식한다. 명사구 인식의 결과가 전체 파싱의 입력으로 사용되기 때문에 명사구 인식은 통사적으로 확실한 정보를 기초로 하여 수행된다. 또한 이 장에서는 명사구 인식의 범위에 해당하는 복합명사 결합문제를 의미쌍의 문제로 분류하고 해결 방법에 대하여 기술한다.

한국어는 문장의 중심어가 후행하고 부분적으로 어순이 자유로운 특징이 있다. 이러한 특징은 서영훈[16]이 제안한 한국어 구문 분석 방법과 같이 문장의 중심어가 필요로 하는 하위범주화 성분을 결합시킴으로써 문장의 분석이 가능하도록 한다. 그리고 S1의 '그(관형사)' 와 같이 한국어의 경우 특정 형태소는 용언이 하위범주화 성분을 취하는 범위를 제한시키는 것을 알 수 있다. 본 논문에서는 이와 같이 한국어에서 발생하는 보편적인 통사적 성질을 적용하여 정확한 명사구 인식을 수행한다.

본 논문에서는 명사구 인식의 범위를 크게 세 가지 범주로 분류한다.

(S1) 오늘날 우리가 알고 있는 [그 주옥같은 작품]을 쓴 사람은 에드워드 드레베라고 못박았다.
(S2) [그 땅]에 나는 금은 질이 좋았다.
(S3) 증인이 [맨 먼저] [돌]로 치고, 그 다음에 모든 백성이 뒤따라서 돌로 치게 하여라.

첫 번째로, 어휘에 구분 없이 모든 한국어의 특성에 적용할 수 있는 부분이다. S1과 S2의 문장에 해당하는 명사구 관련 패턴은 한국어의 어느 문장에서나 예외 없이 하나의 명사구로 결합이 가능하다. 그리고 S3은 "MAG(부사) N_(명사) N_(명사)"의 패턴에 의해 명사구의 범위를 결정할 수 있다. 이 문장에서는 "부사+명사"가 결합하여 문장 전체를 수식하는 부사구의 역할을 수행하는데, 이런 특수한 경우에는 문장내의 명사의 역할을 수행하는 명사구의 범위뿐만 아니라, 부사구의 범위까지 추가적으로 인식할 수 있다. 본 논문에서 제안하는 명사구 인식기는 관형사와 관련된 패턴 17개와 관형격 조사와 관련된 패턴 21개를 사용하여 명사구 인식을 수행하며, 다음은 구축된 패턴의 예이다.

표 1. 명사구 인식에 사용된 패턴의 예

| 관련 품사 | 패턴 |
|---|---|
| 관형사 | MM MAG V_ ETM N_ JX |
| 관형사 | MM MM V_ ETM N_ JK |
| 관형사 | MM MM N_ JX |
| ... | |
| 관형격 조사 | N_ JKG N_ |
| 관형격 조사 | S N_ XSV ETM N_ JKG N_ |
| .... | |

(MM:관형사, MAG:부사, V_:용언, JX:보조사, JK:격조사, ETM:관형형 어미, N_: 명사, JKG: 관형격 조사, XSV:용언화 접미사, S: 문장기호)

표1에서 관련 품사는 명사구 인식의 단서가 되는 품사이며 패턴은 단서로부터 시작되는 패턴이 일치하는 경우 패턴에 해당하는 명사구 결합 규칙에 의해 결합이 가능한 품사 열을 나타낸다. 패턴 'MM N_ JX'에서 명사구로 인식할 수 있는 단서 'MM(관형사)'로부터 그 명사구의 끝을 의미하는 'JX(보조사)'에 의해 'MM N_'이 하나의 명사구로서 그에 해당하는 결합 규칙에 의해 단위화가 이루어진다. 여기에서 패턴에 해당하는 결합 규칙은 'MM V_ ETM N_ JK'의 패턴에서는 'MM(관형사)'와 'V_(용언)+ETM(관형형 어미)'가 모두 후행하는 명사를 수식하는 구조로 명사구 단위화가 이루어져야 하지만 'MM MAG V_ ETM N_ JX'과 같은 패턴에서는 'MAG(부사)'가 우선적으로 'V_(용언)'을 수식하고 부사와 용언이 결합된 단위가 명사를 수식해야 하는 것과 같이 각 패턴에 해당하는 결합 규칙이 존재해야 한다. 또한 패턴 'MM N_ JX'와 'N_ JKG N_'은 수식어가 피수식어의 앞에 위치한다는 한국어의 특징을 표현하고 있으면서 명사구를 이룬 단위가 문장에서 또 다른 패턴에 의해 더욱 넓은 범위의 명사구로 인식될 수 있음을 내포하고 있다.

(S4) 그 때에 제사장은 [제단 앞]에서 입은 그 옷을 벗고 다른 옷으로 갈아입어야 한다.
(S5) 가업을 이어 농사를 짓기 위해 농업전문학교에

진학했던 권씨는 [전역 후] 꿈을 접었다.

두 번째로, S4와 S5와 같이 시간과 장소에 관련된 명사구이다. 시간과 장소에 관련된 어휘는 “앞, 뒤, 옆, 전, 후” 들과 같이 특정한 시간이나 장소의 상태, 혹은 그 위치를 나타내는 어휘이며 이러한 어휘들은 앞이나 뒤에 특정 명사를 두어 특정한 시점이나 위치를 나타낸다. 하지만 S5와 같은 문장에서는 결합되는 범위가 좌측에 위치한 명사로 제한되어 특정 시점을 나타낼 수 있어야 한다. 이와 같은 패턴은 2007년 초에 발표된 세종 21계획의 최종 성과 발표에서 분류하고 있는 관계장소·관계시간의 의미를 갖는 명사에 해당하는 291개의 어휘에 적용된다.

(S6) 맨 끝자리에 앉아 있던 동색 댁이 [장미 바구니] 쪽으로 나가더니 길게 늘어뜨린 분홍색 리본을 이리저리 둘러본다.

세 번째로, 복합명사의 결합에 대한 의미망을 활용한 통계적 접근이다. 복합명사는 시대가 변화하고 시간이 흐름이 존재하는 이상 그 생성과 소멸이 빈번한 것은 자명한 사실이다. 그러므로 복합명사를 기록한 기분석 사전을 이용한 어휘 접근 방식은 명사사전에 새롭게 생성된 명사를 추가한다고 해서 연속된 명사간의 복합명사 결합 여부를 결정지을 수 없는 한계가 있다. 이에 본 논문에서는 각 명사가 지니고 있는 의미를 사용하여 통계적 접근을 시도한다. 통계정보는 세종 21계획의 결과로 제공받은 세종 천만어절 말뭉치 원문에서 추출한 연속하는 일반명사 473,535쌍에 해당하는 2,334,726개의 의미쌍에서 중복 제거된 129,218개의 의미쌍 통계정보가 구축되었으며, 이 중에서 출현빈도 1에 해당하는 의미쌍 26,568개는 통계정보에서 제외하였다. 여기에서 일반명사에서 얻은 의미쌍의 수가 많은 것은 각 명사가 복수의 의미를 갖을 수 있기 때문이다. 이렇게 구축된 통계 정보에 임계값을 설정하면 연속한 명사가 복합명사를 구성할 가능성이 높은 쌍과 그렇지 않은 쌍으로 구분될 수 있다. 예를 들어 S6에서 연속된 명사 ‘장미’와 ‘바구니’는 각각 ‘나무’와 ‘용기’의 의미를 갖는다. 여기에서 ‘나무’와 ‘용기’의 의미쌍이 갖는 확률보다 임계값을 낮게 하면 두 명사는 복합명사로 인식될 수 있다. 연속하는 각 명사는 여러 개의 의미를 가질 수 있기 때문에 가능한 모든 의미쌍에 대한 확률 값을 살펴보고 타당한 값을 선택해야 한다. 본 논문에서는 연속하는 명사의 결합도를 두 명사가 가질 수 있는 모든 의미쌍의 확률 값에서 최상위 확률 값과 최하위 확률 값을 제외한 나머지 값들의 평균을 확률 값으로 선택하였다. 임계값은 테스트 집합 $\boldsymbol{X}$로부터 정답 집합 $\boldsymbol{Y}$에 대하여 가장 높은 정확률을 보이는 임계값 $\theta$를 구하는 과정으로 얻을 수 있다. 그림 1은 집합 $\boldsymbol{X}$를 정답인 집합과 정답이 아닌 집합으로 구분하는 임계논리 유닛의 모형이다[17]. 본 논문에서 사용한 임계값 $\theta$는 연속하는 명사가 갖는 의미쌍이 갖는 확률을 입력 집합 $\boldsymbol{X}$에 대응시키고, 모든 의미쌍에 대한 가중치 $\boldsymbol{W}$를 같게 하여 구하였다. 구축된 통계정보와 복합명사의 결합 여부를 결정하는 임계값은 복합명사 기분석 사전과 의미망에 인식되지 않은 연속하는 명사에 대하여 제한되어 적용된다.

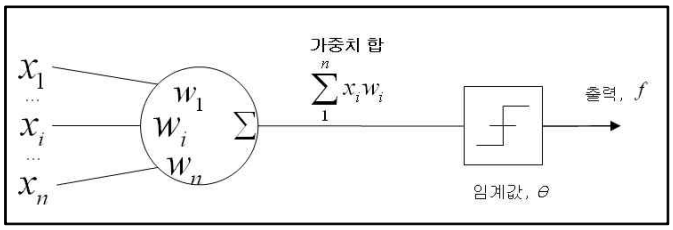


그림 162 . 임계논리 유닛

또한, 복합명사 기분석 사전의 경우 의미망을 사용하면 보다 넓은 범위의 복합명사를 인식할 수 있다. 다음은 세종 21계획의 최종 성과 발표의 명사 사전에 포함된 의미 정보 646개를 바탕으로 의미망 사전을 재구성한 내용의 예이다. 그림 2의 단계3에 해당하는 “음식”의 의미 범주는 4단계에 존재하는 “음료”, “떡”, “빵”의 상위 의미이며 “알콜음료”는 “음료”의 하위 의미임을 나타낸다.

| … | 단계3 | 단계4 | 단계5 | … |
|---|---|---|---|---|
| | 음식 | | | |
| | | 음료 | | |
| | | | 알콜음료 | … |
| | | 떡 | | |
| | | 빵 | | |
| … | | | | |

그림 2. 의미망 사전의 예

그리고 그림 3은 세종 21계획 최종 발표의 성과물로 제공받은 복합명사상세 사전의 ‘가게’에 해당하는 내용의 일부이다. 그림 3의 내용은 ‘가게’의 앞에 기구나 악기등과 같은 의미를 갖는 명사가 오면 복합명사로 구성이 될 수 있다는 것이다.

```
<표제어>~가게</표제어>
<선택제약  논항="X"  의미역="대상">기구|악기|음식|자재|
화장품</선택제약>
```

그림 3. 세종전자사전 복합명사 상세사전의 일부

의미망 사전 적용의 예를 들어보면, “술 가게”라는 연속하는 명사가 복합명사로 구성되는 과정이다. 그림 3과 같이 “가게”라는 일반명사와 결합되어 복합명사를 이룰 수 있는 의미 범주에 “음식”이 존재한다. 그러나 “술”이 갖는 의미는 “알콜음료”로 복합명사의 구성이 실패하게 된다. 이러한 경우에, 통계정보에 각각의 의미를 나타내는 “알콜음료 상업건물”의 의미쌍이 없을지라도 명사구 인식기는 “음식”이 갖는 의미의 하위 의미를 검색하여 “알콜음료 상업건물”의 의미쌍이 복합명사로 구성될 수 있음을 인식할 수 있다.

## 3. 명사구 인식기의 설계 및 구현

### 3.1 시스템 설계

본 논문에서 제안하는 명사구 인식기는 본 연구실이 보유하고 있는 형태소 분석 시스템 CBKMA V3.1을 이용하였다. CBKMA V3.1은 시스템 사전, 어미 사전, 조사 사전, 기분석 사전 등으로 구성되어 있으며, 품사 태그 집합은 26개로 이루어져 있다. 품사 태거는 태그 부착 말뭉치로부터 학습된 어절별 중의성 해소 규칙과 trigram 통계 정보를 이용한 복합적 접근법(hybrid approach)에 기반을 둔 한국어 품사 태깅 시스템을 사용하였다.

제안한 시스템은 전체 파싱 이전의 단계에서 명사구 인식을 수행하며, 2장에서 제안한 명사구 관련 패턴 38개와 관계 장소 및 관계시간에 해당하는 패턴, 그리고 의미망과 통계 정보를 이용한 복합 명사의 결합에 대하여 명사구 인식을 수행한다.

명사구 인식의 과정을 살펴보면, 우선 전처리 단계에서 형태소 분석 단계에서 분리된 “명사+접사”, “조사+조사”, “수사+수사”, “수사+의존명사”와 같이 형태소 분석된 결과를 구문 분석을 위한 입력 단위로 변환한다. 또한, 구축된 복합 명사의 통계 정보와 의미망 사전을 사용하여 복합명사의 결합을 인식한다. 구문 분석을 위한 입력단위로 변환된 결과는 명사구 인식과정에서 관계 장소 및 관계시간, 그리고 구축된 명사구 관련 패턴에 의해 명사구 인식이 수행된다.

### 3.2 시스템 구성도

3.1절의 설계는 다음과 같은 구성도로 표현될 수 있다.

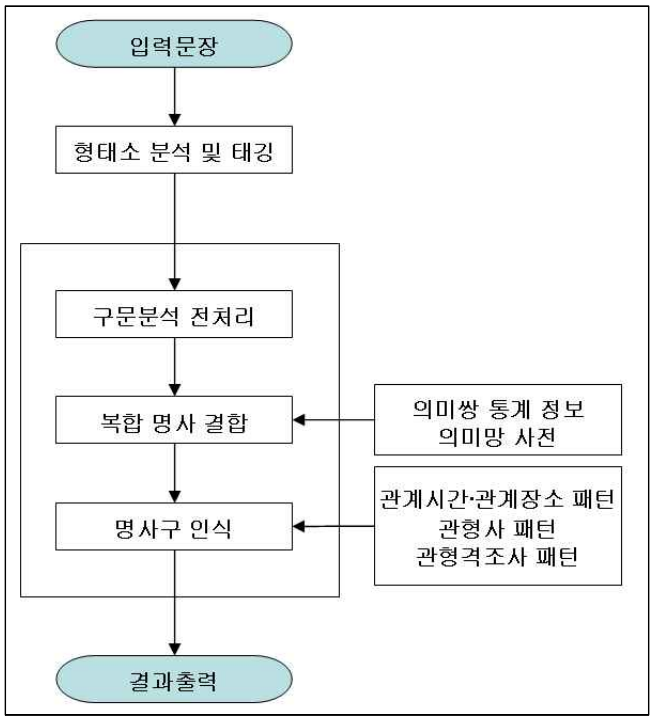


그림 4. 명사구 인식기 구상도

## 4. 실험 및 평가

### 4.1 실험 대상

제안한 시스템에 필요한 명사구 인식 패턴을 구축하기 위해서 세종 21계획의 결과로 제공받은 말뭉치의 원문에서 추출한 10,000개의 문장을 중에서 9,500문장을 사용하였고 나머지 500문장에 대해 명사구 인식을 실험하였다. 복합 명사의 결합도에 대한 통계정보의 구축은 제공받은 말뭉치 원문 전체를 대상으로 하였으며 통계정보에 의한 복합명사 가중치 실험은 명사구 인식 패턴에서 사용한 실험 문장 500개를 대상으로 하였다.

### 4.2 실험 결과

실험에 사용되기 위해 수집된 500개 문장을 명사구 인식 시스템에 적용하여 명사구 인식을 수행한 결과는 표1과 같다.

명사구 인식 시스템의 인식 오류는 인식된 결과 중에서 수동으로 구축한 정답 문장과 일치하지 않는 것으로 판단하였다. 실험에 사용한 본 연구실의 형태소 분석기 CBKMA V3.1은 99% 이상의 정확률을 보이고 있으며, 품사 태거의 경우 형태소 분석기의 분석 오류를 제외하고 98.88%의 정확률을 보인다.

표1. 제안한 명사구 인식 시스템의 실험 결과

| 실험 말뭉치 (500문장) | 올바른 명사구 인식 | 명사구 인식 오류 |
|---|---|---|
| 4,211(어절) | 219(93%) | 16(7%) |

명사구 인식의 오류 16개는 수량 호응에 관한 오류 4개, 문장내의 접속을 유도하는 문장성분에 의한 오류 2개, 복합 명사 7개, 그리고 형태소 분석 오류 3개가 발생하였다. S7-S10의 대괄호는 실험에 사용한 시스템이 출력한 명사구 인식의 결과이고 밑줄은 각 문장의 명사구 인식의 정답에 해당하는 명사구의 범위를 나타낸다.

(S7) 김수영씨가 내개 술 [한 잔]을 내지 않아야 경우가 옳은가?
(S8) [그 다음] [순간, 기천의 눈]에서는 번갯불이 번쩍하였다.
(S9). [모든 국민]은 [거주, 이전의 자유]를 가진다.
(S10) 기타 지방 자치 [단체의 조직과 운영]에 관한 사항은 법률로 정한다.

명사구 인식의 오류는 첫 번째로, 수량 호응 정보의 부족으로 발생하였다. S7에서 수사에 관한 표현을 고려한다면 [술 한 잔]이 하나의 명사구로 인식되어야 올바른 구문 분석이 이루어질 수 있다. 그러나 실험에서는 ‘한(수사)’과 ‘잔(의존명사)’이 명사구 인식기에서 구인식에 성공하였지만 호응 정보에 ‘술’의 수량 호응 정보가 누락되어 ‘술’과 ‘한 잔’은 하나의 명사구로 통합될 수 없었

다. 한국어의 경우 명사마다 수량을 세는 단위가 다르고 그에 따른 수량 호응 정보에 따른 정보가 필요하다는 것을 알 수 있다.

두 번째로, 같은 품사를 같은 형태소가 다른 문장에서 서로 다른 구성성분의 역할을 수행하는 경우에 발생하였다. S8의 '순간, 기천의 눈'과 S9의 '거주, 이전의 자유'는 모두 '명사 문장기호 명사+관형격조사 명사'의 동일한 형태소 열을 갖지만 S8에서의 ',' 는 문장의 좌우를 분리하여 '그 다음 순간' 이 문장에서 시점을 나타내는 성분이 되지만 S9에서는 좌우의 성분을 결합시키는 접속사의 역할을 하고 있다. 이와 같은 경우는 문장 좌우의 문맥이 파악되어야 정확한 구문 분석이 이루어질 수 있다. 일반적으로 이러한 문장은 통사적으로 중의적인 문장으로 분류하여 구문 분석의 단계에서 가능한 모든 파스트리를 제시하기도 하지만 본 논문에서는 이러한 분석을 오류에 포함시켰다.

세 번째로, 복합 명사의 결합이 정확히 이루어지지 않은 분석오류이다. S10에서는 '기타 지방 자치 단체'가 우선 복합 명사로 인식이 되어야 올바른 구문 구조가 만들어질 수 있지만 '자치'와 '단체'에 대한 복합 명사의 정보가 기분석 사전과 의미망 사전에 존재하지 않고 의미쌍의 확률이 임계값보다 작기 때문에 '자치'와 '단체'가 결합되지 못한 경우이다. 이와 같은 맥락에서 살펴보면 연속하는 두 명사에 대한 의미쌍 통계정보에 대한 임계치가 적절하게 조정되어 있지 않다면 복합 명사에 대한 결합이 올바르게 수행될 수 없다는 것을 알 수 있다.

## 5. 결론

본 논문에서는 통사적으로 확실한 패턴, 그리고 제한된 통계 정보와 의미망을 적용한 명사구 인식 시스템을 제안하였다. 세종 21계획 최종 성과 발표로 제공 받은 말뭉치와 사전으로부터 통계 정보와 의미망, 명사구 인식 패턴을 구축하고 명사구 인식에 적용하였다. 실험에서 명사구 인식의 오류는 형태소 분석에 대한 오류와 중의성으로 발생하는 인식오류 및 복합명사결합의 오류를 포함하고 있기 때문에 통사적인 정보와 제한된 통계정보로도 높은 수준의 명사구 인식이 이루어질 수 있음을 보인다.

향후 연구로 더욱 견고한 명사구 인식의 정보 구축과 복합 명사 결합의 가중치의 설정을 들 수 있다. 명사구 인식 패턴의 신뢰성을 높이고 수량 호응정보가 늘어나고, 복합명사결합에 관한 가중치 설정이 더욱 견고할수록, 제안된 시스템의 성능은 높아질 것이다.

## 5. 참고문헌

[1] S. Abney, "Parsing by Chunks", in R. Berwick, S. Abney, C. Tynny, eds, Principle-Based Parsing, Kluwer, pp.257-78, 1991

[2] C. Cardie, S. Mardis, D. Pierce, "Combining Error-Driven Pruning and Classification for Partial Parsing", PROC ICML-99 (International Conference on Machine Learning), 1999.

[3] J. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, M. Tyson, "FASTUS: A Cascaded Finite-State Transducer for Extracting Informaion from Natural-Language Text", in Roche, Schabes, eds., Finite-State Language Processing, MIT, pp. 383-406,1997

[4] W. Skut, T. Brants, "A Maximum-Entropy Partial Parser for Unrestricted Text", Proc 6th Workshop On Very Large Corpora, 1998.

[5] L.A Ramshaw, M.P. Marcus, "Text chunking using Transformation-Based Learning", In Proc. of the 3rd ACL workshop on Very Large Corpora, 1995.

[6] S. Argamon, I. Dagan, Y. Krymolowski, "A Memory-Based Approach to Learning Shallow Natural Language Patterns," Proc ACL/Coling pp.67-73, 1998.

[7] 신효필, "최소자원 최대효과의 구문분석", 제 11회 한글 및 한국어 정보처리 학술대회, pp.242-248, 1999.

[8] 김미영, 강신재, 이종혁, "규칙과 어휘정보를 이용한한국어 문장의 구묶음(Chunking)", 제12 회 한글 및 한국어 정보처리 학술대회 논문집, pp.11-17, 2000.

[9] 김미영, 간신재, 이종혁, "단위(Chunk)분석과 의존문법에 기반한 한국어 구문분석", 한국정보과학회 2002 봄 학술발표 논문집, pp.327-329

[10] Juntae Yoon, et. al. "Three Types of Chunking in Korean and Dependency Analysis based on Lexical Association," In Proc. of the 18th International Conference on Computer Processing Languages(ICCPOL'99), pp. 59-65, 1999.

[11] 양재형, "규칙기반 학습에 의한 한국어의 기반 명사구 인식", 정보과학회 논문지: 소프트웨어 및 응용, 제 27권 제 10호, pp. 1062-1071, 2000.

[12] 박성배, 장병탁, "최대 엔트로피 모델을 이용한 텍스트 단위화 학습", 제 13회 한글 및 한국어 정보처리학술대회, pp. 130-137, 2001.

[13] 황영숙, 정후중, 박소영, 곽용재, 임해창, "자질집합선택 기반의 기계학습을 통한 한국어 기본구 인식의 성능향상", 정보과학회논문지: 소프트웨어 및 응용 제 2

9권 제 9호, 2002.
[14] 박의규, 조민희, 김성원, 나동열, “구묶음과 구간분할을 이용한 의존 관계 추출 기법”, 제 16회 한글 및 한국어 정보처리 학술대회, pp.131~137, 2004
[15] 박의규, 나동렬, “한국어 구문 분석을 위한 구묶음 기반 의존명사 처리”, 인지과학, 제17권 제2호, pp119~138, 2006.
[16] 서영훈, “의미 정보를 이용하는 중심어 주도의 한국어 파싱”, 서울대학교 컴퓨터공학과 공학박사 학위논문, 1991.
[17] Nilsson, N., "Teleo Reactive Programs for Agent Control", Jornal of Artificail Intelligence Research, 1, pp. 139-158, 1994.