명사 의미 부류를 이용한 연속된 명사열의 구묶음

Chunking of Contiguous Nouns using Noun Semantic Classes

안광모, 서영훈
충북대학교 전자정보대학 컴퓨터공학부

Kwang-Mo Ahn(ahnmo@nlp.cbnu.ac.kr), Young-Hoon Seo(yhseo@chungbuk.ac.kr)

요약

본 논문에서는 조사가 없는 연속된 명사열 중 하나의 명사처럼 쓰일 수 있는 명사열을 복합명사구라 정의하고, 의미 정보를 이용한 복합명사구의 구묶음 방법을 제시한다. 복합명사구의 구묶음에는 구문분석 말뭉치에서 추출한 명사쌍과 이들의 의미부류정보를 이용한다. 이러한 명사쌍과 의미부류정보는 신뢰도를 위해 세종말뭉치의 구문분석 말뭉치와 상세사전을 기반으로 구축하였다. 이들 정보를 이용한 복합명사구 구묶음 모듈은 길이(명사의 수)가 2 이상인 복합명사구에 대해서도 구묶음을 수행할 수 있다. 복합명사구 구묶음을 위해 '왼쪽명사-오른쪽명사' 쌍 38,940개, '왼쪽명사-오른쪽명사의미부류' 쌍 65,629개, '왼쪽명사의미부류-오른쪽명사' 쌍 46,094개, '왼쪽명사의미부류-오른쪽명사의미부류' 쌍 45,243개의 정보를 구축하여 이용하였다.

실험을 위하여 신문기사의 내용으로 이루어진 세종형태소분석 말뭉치로부터 길이가 3 이상인 조사와 결합하지 않은 연속된 명사열을 포함하는 1,000 문장을 임의로 선별하였으며, 실험결과는 86.89%의 정밀도와 80.48%의 재현율, 그리고 83.56%의 f-measure를 보였다.

■ 중심어 : | 구묶음 | 복합명사구 | 구문분석 | 자연언어처리 |

Abstract

This paper presents chunking strategy of a contiguous nouns sequence using semantic class. We call contiguous nouns which can be treated like a noun the compound noun phrase. We use noun pairs extracted from a syntactic tagged corpus and their semantic class pairs for chunking of the compound noun phrase. For reliability, these noun pairs and semantic classes are built from a syntactic tagged corpus and detailed dictionary in the Sejong corpus. The compound noun phrase of arbitrary length can also be chunked by these information. The 38,940 pairs of 'left noun - right noun', 65,629 pairs of 'left noun - semantic class of right noun', 46,094 pairs of 'semantic class of left noun - right noun', and 45,243 pairs of 'semantic class of left noun - semantic class of right noun' are used for compound noun phrase chunking.

The test data are untrained 1,000 sentences with contiguous nouns of length more than 2 randomly selected from Sejong morphological tagged corpus. Our experimental result is 86.89% precision, 80.48% recall, and 83.56% f-measure.

■ keyword : | Chunking | Compound Noun Phrase | Syntax Analysis | Natural Language Processing |

* 본 논문은2008년도 충북대학교 학술연구지원사업의 연구비지원에 의하여 연구되었음(This work was supported by the research grant of the Chungbuk National University in 2008)

접수번호 : #100125-001
접수일자 : 2010년 01월 25일
심사완료일 : 2010년 03월 12일
교신저자 : 서영훈, e-mail : yhseo@chungbuk.ac.kr

I. 서 론

구문분석에 대한 많은 연구들이 있어 왔고 지금도 많은 연구들이 진행되고 있지만, 아직까지 구문분석의 성능은 실용 분야에 쓰일 단계는 아니다. 따라서 구문분석의 성능을 높이기 위한 많은 방법들이 연구되고 있는데, 그 중 하나가 구묶음(chunking)을 하는 방법이다. 구묶음은 구문분석의 분석 후보를 줄여주어 구문분석의 복잡도를 감소시키고 구문분석의 중의성을 줄여준다. 예를 들어, 문장 "저 아름다운 여자는 나의 아내이고, 그 옆의 귀여운 여자 아이는 나의 딸이다."의 경우, 분석 후보는 12개이고 구문분석의 인식 복잡도는 $O(n^3)$이므로 $12^3 = 1,728$만큼의 복잡도를 갖게 되지만, "[저 아름다운 여자]는 [나의 아내]이고, [그 옆의 귀여운 여자 아이]는 [나의 딸]이다."와 같이 구묶음을 수행하고 나면 분석 후보가 4개로 줄어들어 $4^3 = 64$만큼의 복잡도로 감소하게 된다. 따라서 구묶음이 구문분석의 성능 향상에 기여한다는 것은 반박의 여지가 없다.

한국어는 조사가 발달된 언어로써, 조사정보를 이용하여 명사가 그 문장 내에서 어떠한 격을 갖는지 알 수 있다. 하지만 조사정보만을 이용하여 명사가 문장에서 가지는 격을 결정하기 어려운 경우가 있는데 보조사와 결합한 명사라든지 또는 조사가 생략되거나 없는(앞으로는 "조사가 없는"이란 용어만을 사용) 명사가 문장에 존재하는 경우가 대표적이다. 이러한 보조사와 결합한 명사나 조사가 없는 명사들은 구문중의성을 증가시켜 구문분석의 성능을 떨어뜨리게 되는 주된 요인 중 하나이다. 특히, 조사가 없는 연속된 명사열이 문장에서 나타나게 되면 그 문장이 갖는 구문중의성은 명사열의 길이에 따라 지수적으로 커지게 되며, 그렇기 때문에 기존 연구에서는 이러한 명사열을 하나의 명사처럼 구묶음을 하여 구문분석에서 처리하는 경우가 많았다. 하지만 이렇게 처리하는 방법은 올바르지 않은데, 그 이유는 조사가 없는 연속된 명사열을 하나의 명사처럼 일괄적으로 처리하게 되면 문장의 주요 성분들을 잃어버리게 되거나 구문구조가 올바르지 않게 분석되기 때문이다. 예를 들면, "지난 봄 학교에서 놀이공원으로 봄 소풍을 갔다."라는 문장을 보면 첫 번째의 [봄 학교]는 하나의 명사처럼 처리할 수 없는 경우이며, 두 번째의 [봄 소풍]은 하나의 명사처럼 처리할 수 있는 경우이다. 조사가 없는 명사열을 하나의 명사구처럼 처리할 경우 두 번째의 경우는 문제가 되지 않지만 첫 번째의 경우는 '봄'이 '지난'과 결합하여 문장 내의 부사와 같은 역할을 해야 한다. 따라서 이 경우 조사가 없는 연속된 명사열을 무조건 하나의 명사구로 구묶음하여 구문분석을 하게 되면 잘못된 결과를 내놓게 된다. 이러한 조사가 없는 연속된 명사열을 처리하기 위해서 각 명사열에 대하여 사전을 구축하여 처리할 수도 있겠지만, 명사열의 길이가 일정하지 않고 각 명사들의 조합을 생각한다면 모든 명사열을 사전으로 처리한다는 것은 불가능하다.

따라서 본 논문에서는 두 개의 명사 또는 의미쌍을 이용하여 하나의 명사처럼 쓰일 수 있는 연속된 명사열을 구묶음하는 방법에 대하여 기술한다. 명사쌍은 세종 구문분석 말뭉치로부터 자동으로 추출되었으며, 의미쌍은 추출된 명사쌍의 명사에 대한 의미부류정보를 이용하여 구축되었는데, 명사의 의미부류정보는 세종체언상세사전에서 추출하였다. 구묶음 모듈은 추출된 명사쌍과 의미쌍을 이용하여 문장을 우에서 좌로 분석해 나가고, 조사가 없는 명사열이 탐색되었을 때 추출된 명사-의미쌍 정보를 이용하여 구묶음을 수행하게 된다. 자세한 내용은 3장 이후부터 다룬다.

본 논문의 구성은 다음과 같다. 2장에서는 구묶음 관련 연구에 대하여 기술하며, 3장에서는 조사가 없는 연속된 명사열에 대하여 기술한다. 4장에서는 명사-의미쌍을 자동으로 추출하는 방법에 대하여 기술하고, 5장에서는 조사가 없는 연속된 명사열을 구묶음하는 방법에 대하여 기술한다. 그리고 6장에서 실험을 통하여 본 논문의 방법을 평가해 본다. 마지막으로 7장의 결론을 통하여 본 논문을 마친다.

II. 관련 연구

처음으로 구묶음(chunking)에 대한 방법론은 Abney에 의해서 제안되었다[1]. Abney는 구묶음을 수행하는

부분과 붙임(attaching)을 수행하는 부분으로 구문분석의 단계를 나누었다. 이 연구에서는 사람이 문장을 읽을 때 끊어 읽는 운율적 휴지를 경계로 한 단어의 경계를 기준으로 문장 성분을 묶은 것을 구묶음으로 보았다. 구묶음이 된 덩어리(chunk)는 붙음모듈(attacher)이 분석해야 할 분석 후보를 줄여주어 성능을 향상시킬 수 있게 된다.

영어권의 구묶음에 대한 연구로 [2]의 경우는 Brill이 제안한 변형기반학습(transformation-based learning) 기법을 구묶음에 적용하였다. 이 연구에서는 구묶음이 된 학습 말뭉치를 학습하여 이를 다시 다른 말뭉치의 구묶음 인식기로 활용하였다. 구묶음 인식기는 비순환 기본 명사구(non-recursive basic noun phrase)를 구묶음으로 인식하였다. 그 외 [3]과 [4]는 유한 상태 오토마타(finite state automata)를 이용하여 구묶음 기법을 개발하였으며, [5]는 구묶음 태그를 이용하여 구묶음 구조를 나타낼 수 있도록 하였다. 이들의 연구들은 90%의 성능을 넘지만 조사가 없는 영어권의 구묶음 연구들은 본 논문에서 제안하는 명사열들에 대한 구묶음은 아니다.

한국어의 경우도 구묶음에 대한 여러 연구들이 있었다. [6-8]의 경우 기반 명사구 인식에 대한 연구를 하였는데, 기반 명사구(base NP)란 명사구 내부에 다른 명사를 포함하지 않는 명사를 말한다. 이 연구들은 I(inside), O(out), B(Begin) 등의 태그를 이용하여 구묶음을 표현하는데, [6]의 경우는 형태소분석 말뭉치를 규칙 기반 알고리즘을 이용하여 학습시키고, 학습된 규칙으로 구묶음을 수행하였다. [7]은 기본구 인식을 위한 자질들을 학습 알고리즘을 이용하여 선택한 후, 선택된 자질 집합을 이용하여 기본구 인식을 학습한다. [8]에서는 tri-gram HMM과 어절 문맥 정보를 이용하여 기반 명사구 인식의 성능을 향상시켰다. 기반명사구의 인식은 관형사나 관형형어미와 결합한 용언의 수식을 받는 명사구등을 처리할 수 있다. [9]의 연구에서는 등위접속 명사구를 인식하였는데, 병렬명사구의 대칭성과 교환정렬 모델 및 수식관계 정보를 이용하여 등위접속명사구를 인식하였다. [10]의 경우는 의존명사와 관련된 구묶음에 대한 규칙을 연구하였으며, 의존명사를 단위 명사와 비단위 명사로 구분하여 의존명사 관련된 구묶음을 수행하였다. 이 외에도 많은 구묶음에 대한 연구들이 있지만 조사가 없는 연속된 명사열을 구묶음하는 연구들은 찾아보기 힘들다.

III. 복합명사구

복합명사(compound noun)는 원래 둘 이상의 단어가 합쳐져 생성된 명사를 의미한다. 예를 들어 영어의 'blackboard(black+board)', 'basketball(basket+balll)'이나 한국어의 '새해(새+해)', '솔밭(솔+밭)'와 같은 명사들이다. 그리고 복합명사구라 하면 이러한 복합명사들이 모여 구(phrase)를 이루는 것이라 할 수 있다. 하지만 본 논문에서는 복합명사구(compound noun phrase)를

"조사가 생략되거나 없는 둘 이상의 연속된 명사열에서 마치 하나의 명사처럼 쓰일 수 있는 명사들의 묶음(단, 명사열의 마지막 명사는 조사를 포함할 수 있다.)"

으로 정의한다. 조사가 없는 명사열에 대하여 살펴보면 다음 [표 1]과 같은 유형으로 나누어 볼 수 있다.

표 1. 조사가 없는 명사열의 분류와 예

	유형	예시
1	부사성명사를 포함하는 명사열	올해 학교
2	등위 접속 관계의 명사열	고양이 개
3	수식 관계의 명사열	관리 대상
4	복합명사와 같은 명사열	꼬마 전구

유형1의 경우는 조사가 없는 명사열이기는 하지만 앞의 명사가 문장의 부사 성분으로 쓰여 뒤의 명사와 구묶음을 수행할 수 없는 경우이다. 유형 2의 경우는 명사열을 이루는 각각의 명사가 등위접속관계를 이루는 것으로 하나의 명사처럼 구묶음이 될 수 없는 경우이다. 이러한 유형의 경우는 올바른 구문분석을 위해서 다른 처리 방법을 필요로 한다. 유형3의 경우는 좌측의 명사가 우측의 명사를 수식하는 형태이다. 유형3의 예시를 보면 '관리'라는 단어 다음에 관형격조사 '-의'가 생략되

어 '대상'을 수식하는 경우라고 볼 수 있다(관형격조사가 결합 되어도 본래 의미는 변하지 않음). 유형4는 마치 본래 의미의 복합명사와 같이 두 단어가 결합되어 하나의 명사처럼 쓰일 수 있는 경우이다. 유형4의 예시는 띄어쓰기오류라고 볼 수 있지만 한국어 문장에서 종종 나타나는 경우이기 때문에 따로 분류하였다. 본 논문에서 정의한 복합명사구는 [표 1]의 유형3과 유형4이며, 다음은 그 예들이다.

- 본 논문에서 정의한 복합명사구의 예
(1) 겨울 스포츠
(2) 국내 경기
(3) 수해 복구 공사
(4) 교량 복구 준공 검사 과정

(1)과 (2)는 조사가 없는 연속된 명사열에서 명사의 수가 2개인 경우이며, (3)과 (4)는 명사의 수가 2개를 넘는 명사열의 경우이다. 이러한 조사가 없는 명사는 서론에서도 기술하였듯이 구문중의성을 증가시켜 구문분석의 성능을 감소시키는 요인으로 작용하며, 이런 명사들이 연속해서 나오게 되면 구문중의성은 더욱더 커지게 된다. 세종 형태소분석 말뭉치 1,155,716문장 중 조사가 없는 연속된 명사열을 포함하는 문장을 추출한 결과 총 423,503 개의 문장이 추출되었으며(36.64%), 이중 조사가 없는 명사를 3개를 포함하는 명사열을 갖는 문장도 129,752문장(11.23%)이 나타났다. 또한 한 문장에 이러한 명사열이 여러 개를 포함하는 경우가 대부분이다. 따라서 이러한 조사가 없는 연속된 명사열에 대하여 올바르게 구묶음을 수행한다면 구문분석의 성능은 향상될 것이다.

IV. 명사-의미쌍의 추출

본 논문에서 복합명사구의 구묶음은 두 개의 명사 또는 의미부류로 구성된 명사-의미쌍을 이용하여 수행한다. 이번 장에서는 세종구문분석 말뭉치로부터 우선 두 개의 명사로 구성된 명사쌍을 자동으로 추출하고, 명사쌍으로부터 명사의미사전을 이용하여 명사-의미쌍을 추출하는 방법에 대하여 기술한다.

1. 명사쌍의 추출

본 논문에서 정의한 복합명사구는 대부분이 왼쪽의 명사가 오른쪽의 명사를 수식하는 형태로 되어 있다. 그리고 이런 복합명사구의 의미적 중심어(semantic-head)는 수식을 받는 오른쪽의 명사이다. 명사의 수가 3개 이상인 복합명사구의 경우 각 명사들을 두 개씩 쌍을 이루어보면 두 명사의 중심어가 되는 명사(오른쪽 명사)들이 있고, 또 그 중심어가 되는 명사들끼리 서로 쌍을 이루어보면 수식을 하는 명사와 그것의 수식을 받아 중심어가 되는 명사가 있게 된다. 이런 형태로 복합명사구를 이루다보면 복합명사구도 구문구조(syntactic structure)를 갖게 된다. 다음 [그림 1]은 복합명사구 "야생 동물 불법 거래"의 구문구조를 나타낸다.

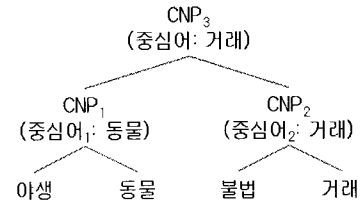
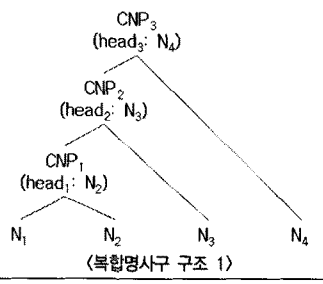
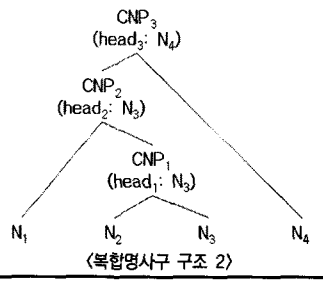


그림 1. "야생 동물 불법 거래" 의 구문구조

위 [그림 1]에서 우선 '야생'과 '동물'이 서로 수식관계를 갖게 되며 이중 중심어$_1$은 '야생'의 수식을 받는 '동물'이 된다. 또 '불법'과 '거래'가 서로 수식관계를 갖게 되므로 중심어$_2$는 '거래'가 되게 된다. 마지막으로 중심어$_1$('동물')과 중심어$_2$('거래')가 서로 수식관계를 이루며 이것의 중심어는 '거래'가 되게 된다. 즉, 위와 같은 구조에서 두 개씩 쌍을 이루어 복합명사구를 이룰 수 있는 명사쌍을 추출하면 "야생 동물", "불법 거래", 그리고 "동물 거래"가 되게 된다. 본 논문에서는 세종 구문분석 말뭉치의 구문분석된 문장에서 조사가 없는 연속된 명사열들로 이루어진 명사구를 추출하고, 추출된 명사구로부터 두 개 명사로 이루어진 명사쌍을 추출하였다. [표 2]는 세종 구문분석 말뭉치로부터 명사쌍을 추

출하는 과정의 예를 보여준다.

표 2. 세종구문분석말뭉치로부터 명사쌍 추출의 예

복합명사구의 구문구조	명사쌍
 〈복합명사구 구조 1〉	[N1 N2] [N2 N3] [N3 N4]
 〈복합명사구 구조 2〉	[N2 N3] [N1 N3] [N3 N4]

이와 같은 방법을 통하여 세종구문분석말뭉치 77,121개의 문장으로부터 총 38,940개의 명사쌍을 추출하였다.

2. 명사-의미쌍의 추출

자연언어 문장에 나타나는 복합명사구를 모두 사전에 기록한다는 것은 거의 불가능하다. 명사사전에 m개의 명사가 등록되어 있고 n개의 명사로 이루어진 복합명사구 사전을 구축할 경우, m^n개 만큼의 복합명사구 후보로부터 복합명사구를 사전을 구축해야 한다. 따라서 본 논문에서는 복합명사구의 구묶음에 의미정보를 이용하고자 의미쌍을 추출하였다. 의미쌍은 4-1절의 방법과 같이 구문분석말뭉치로부터 추출한 명사쌍과 명사의미사전, 의미사전을 이용하여 구축하였다. 명사의미사전은 세종체언상세사전 내에 각 명사별로 기술되어 있는 의미부류정보를 이용하여 구축되었으며, 의미사전은 트리(tree) 형태의 계층적인 구조로 되어있는 646개의 세종의미부류를 이용하여 구축되었다. 다음 [그림 2]는 명사 '가구'에 대한 체언상세사전의 일부 내용이다.

```
<superEntry>
  <orth>가구</orth>
  <entry n="1" pos="nng_s">
    …
    <sem_class>가구</sem_class>
    …
  </entry>
  <entry n="2" pos="nng_s">
    …
    <sem_class>인간집단</sem_class>
    …
```

그림 2. '가구' 에 대한 체언상세사전의 내용 일부

그리고 [그림 3]은 세종의미부류 계층구조이며, [표 3]은 이것을 계층구조를 이용하여 구축된 의미사전의 일부이다.

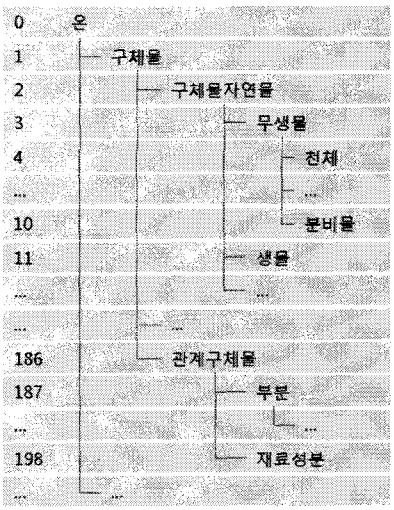


그림 3. 세종의미부류 계층구조

표 3. 의미사전의 일부

의미부류 인덱스	의미부류	부모 인덱스	마지막 자식인덱스
0	온	-1	645
1	구체물	0	198
2	구체자연물	1	89
3	무생물	2	10
4	천체	3	-1
…	…	…	…
10	분비물	3	-1
…	…	…	…

[표 3]에서 부모인덱스가 -1일 경우는 계층구조의 최상위 의미부류임을 의미하며, 마지막 자식인덱스가 -1인 경우는 말단 의미부류임을 의미한다. 부모인덱스는 현재 의미부류의 바로 상위의 부모의미부류 인덱스를 의미하며, 마지막 자식인덱스는 현재 의미부류의 자식 의미부류 중 가장 마지막에 있는 의미부류의 인덱스를 의미한다. 또한 의미부류인덱스 i인 의미부류가 '현재 의미부류인덱스 < i ≤ 마지막 자식인덱스'인 경우는 현재의미부류의 하위의미부류임을 의미한다.

아래 [표 4]는 체언상세사전과 의미사전을 이용하여 구축한 명사의미사전의 일부를 나타낸 것이다.

표 4. 명사의미사전의 일부

명사	의미부류 인덱스
...	...
가관	438 468
가교	257 481
가구	113 200
가구점	248
...	...

의미부류색인번호는 646개의 세종의미부류를 이용하여 구축된 의미사전 대한 의미부류 인덱스이며, 의미부류 인덱스를 이용하여 의미사전을 바로 탐색할 수 있다.

N_1의 의미부류가 S_{11}, S_{12}, …, S_{1m} 이고, N_2의 의미부류가 S_{21}, S_{22}, …, S_{2n}인 명사쌍 [N_1 N_2]에 대한 명사-의미쌍은 다음 [표 5]와 같이 세 가지 유형으로 구축하였다.

표 5. 명사쌍 [N_1, N_2]에 대한 명사-의미쌍의 구축

[왼쪽명사 - 오른쪽명사의미부류]조합	
N_1 → S_{21}, S_{22}, ..., S_{2n}	[N_1 S_{21}] [N_1 S_{22}] ... [N_1 S_{2n}]
[왼쪽명사의미부류 - 오른쪽명사]조합	
S_{11}, S_{12}, ..., S_{1m} → N_2	[S_{11} N_2] [S_{12} N_2] ... [S_{1m} N_2]
[왼쪽명사의미부류 - 오른쪽명사의미부류]조합	
S_{11}, S_{12}, ..., S_{1m} × S_{21}, S_{22}, ..., S_{2n}	[S_{11} S_{21}] [S_{11} S_{22}] ... [S_{1m} S_{2n}]

이와 같은 방법으로 구축된 의미쌍은 [왼쪽명사 - 오른쪽명사의미부류]쌍이 65,629개, [왼쪽명사의미부류 - 오른쪽명사]쌍이 46,094개, [왼쪽명사의미부류 - 오른쪽명사의미부류]쌍이 45,243개이다.

V. 복합명사구의 구묶음

4장에서 복합명사구를 이룰 수 있는 두 개의 명사쌍과 의미쌍을 구축하는 방법에 대하여 기술하였다(앞으로는 구축된 명사쌍과 의미쌍을 복합명사구사전이라 한다). 이번 장에서는 앞 장과 같은 방법으로 구축된 복합명사구사전을 이용하여 복합명사구를 구묶음하는 방법에 대하여 기술한다.

1. 복합명사구의 구묶음

명사의 개수가 2개인 복합명사구의 경우는 단순히 복합명사구사전에서 검색이 성공하는 경우 구묶음을 수행하면 된다. 하지만 명사의 개수가 3개 이상인 복합명사구에 대하여 구묶음을 수행할 경우는 이렇게 단순히 복합명사구사전을 비교/검색하는 방법으로는 올바른 구묶음을 수행할 수 없다. 예를 들어, "교량 복구 준공 검사 과정"이라는 복합명사구가 있고 복합명사구 사전에 "교량 복구", "교량 준공", "교량 검사", "검사 과정"이란 복합명사구가 등록되어 있을 경우, "[교량 복구] 준공 [검사 과정]"과 같이 불완전하게 복합명사구가 인식되게 될 것이다. 따라서 본 논문에서는 복합명사구사전을 이용하여 세 개의 명사 이상으로 구성된 복합명사구의 구묶음을 다음과 같은 방법으로 한다.

```
ns[k] = {N1, N2, ···, Nk};   // 명사열 N1 N2 ··· Nk
CNPChunker()
{
 right = k;        // 명사열의 마지막 명사의 인덱스
 left = right - 1; // 마지막 명사의 바로 왼쪽 명사 인덱스
 LinkNoun(left, right);
 Combine();
}
LinkNoun(left, right)
{
 if(right == 2 and left is 10) return;
 if(ns[left] is a left noun of ns[right] and
      left is not -1){
      link ns[left] and ns[right];
      left--;
      LinkNoun(left, right);
 }
 else{
      right--;
      left = right-1;
      LinkNoun(left, right);
 }
}
```

복합명사구의 구묶음 방법

CNPChunker() 함수는 복합명사구의 구묶음을 수행하는 함수이며 복합명사구를 이룰 수 있는 명사들을 연결해주는 함수 LinkNoun()함수와 연결된 함수를 하나의 복합명사구로 묶어주는 Combine()함수를 호출한다. 명사열 N_1 N_2 ··· N_k가 주어졌을 때, LinkNoun() 함수는 복합명사구사전을 이용하여 우에서 좌로 주어진 명사열을 분석한다. LinkNoun() 함수는 먼저 명사 N_k와 그것의 왼쪽에 있는 명사들을 우에서부터 좌(N_{k-1}에서 N_1까지)로 하나씩 비교해가면서 명사 N_k와 복합명사구를 이룰 수 있는지 확인한다. 이때 "N_{k-1} N_k"가 복합명사를 이룰 수 있다면(즉, 복합명사구 사전에 두 명사에 대한 명사-의미쌍이 등록되어 있다면) 이 두 명사를 서로 연결(link)한다. 두 명사의 연결에 성공했을 경우 "N_{k-2} N_k"에 대해서도 복합명사구사전을 이용하여 복합명사구를 이룰 수 있는지 확인하며, 복합명사구를 이룰 수 있다면 역시 연결을 하게 된다. 이러한 과정은 명사 N_k와 그것의 왼쪽에 있는 명사들과의 연결이 실패하거나(명사 N_k와 그것의 왼쪽 방향에 있는 비교 대상의 명사가 복합명사구를 이룰 수 없거나) 더 이상 N_k에 대한 왼쪽 방향의 명사가 없을 경우(즉, N_1 N_k까지 다 비교를 마쳤을 때) 명사 N_k와의 비교를 마치고, 명사 N_{k-1}과 그것의 왼쪽방향 명사들에 대해서도 위와 같은 과정을 반복한다. 이러한 방법으로 "N_1 N_2"까지 분석을 마친 후, Combine() 함수를 통하여 연결된 명사들끼리 구묶음을 수행하고 구묶음을 마치게 된다. 다음 [그림 4]는 이러한 방법을 이용하여 조사가 없는 연속된 명사열 " 강남 초등학교 가을 소풍 주가 폭락"을 분석하는 과정을 보여준다.

그림 4. 복합명사구가 인식되는 과정

이러한 방법으로 복합명사구를 인식하는 이유는 왼쪽의 명사가 오른쪽의 명사의 수식어이며 왼쪽의 명사가 오른쪽명사를 수식하고 나면 오른쪽 명사가 그 명사구의 중심어가 되어 또다시 왼쪽에 나오는 명사의 수식을 받을 수 있기 때문이다. 예를 들어 "저 예쁜 강아지의 주인"에서 '예쁜'이 '강아지의'를 수식하여 '강아지의'가 중심어로 하는 명사구로 묶이게 된다. 또다시 '저'가 다음 명사구의 중심어인 '강아지의'의 수식어가 되므로 하나의 명사구로 묶이게 된다. 마지막으로 앞에서 구묶음된 명사구의 중심어 '강아지의'가 '주인'을 수식하여 명사구로 묶이게 되어 "저 예쁜 강아지의 주인" 전체가 하나의 명사구로 구묶음이 되는 것과 같은 원리이다.

2. 의미쌍을 이용한 복합명사구의 구묶음

의미쌍을 이용하여 구묶음을 수행할 때는 명사의미사전과 의미사전을 이용한다. 명사열 "N_1 N_2"를 의미쌍을 이용하여 구묶음을 수행할 경우, "명사 N_1과 명사 N_2의 의미부류들의 쌍", "명사 N_1의 의미부류들과 명사 N_2의 쌍", "명사 N_1의 의미부류들과 명사 N_2의 의미부류들의 쌍"을 추출하고 이를 복합명사사전과 비교하여 구묶음을 수행하게 된다. 각 명사의 의미부류들은 명사

의미사전을 검색하여 찾는다. 세 가지의 명사-의미쌍들은 다시 복합명사사전을 이용하여 구묶음이 가능한지를 판단하게 되는데, 명사의 어휘정보만을 비교할 경우는 단순히 단어가 일치하면 되지만 의미부류를 비교할 경우는 의미부류의 상하위 관계를 비교해야 한다(의미사전은 상하위 관계의 계층적인 구조, 4-2절에 기술). 예를 들어 복합명사구사전에 (음식 가게)라는 "왼쪽명사의미부류-오른쪽명사"쌍이 있으며 입력된 명사열이 "술 가게"일 경우, '술'의 의미부류는 '알콜음료'이며 이는 '음식'이라는 의미부류가 아니다. 하지만 [그림 5]와 같이 '알콜음료'는 '음식'의 하위 의미부류이며 이는 '알콜음료'가 '음식'이라는 의미부류에 속한다는 뜻이므로 명사열 "술 가게"는 구묶음이 될 수 있도록 하여야 한다.

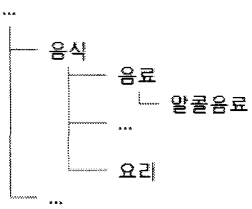


그림 5. 의미부류 '음식' 과 그것의 하위의미부류들

따라서, 본 논문의 복합명사구 구묶음 모듈은 입력된 명사의 의미부류가 복합명사구사전에 등록된 의미부류보다 하위의미부류이면 구묶음이 이루어질 수 있도록 한다.

VI. 실험 및 결과분석

1. 실 험

본 논문에서 기술한 복합명사구 구묶음 성능을 측정하기 위해 신문기사 내용으로 이루어진 세종형태소분석 말뭉치로부터 1,000 문장을 임의로 추출하였으며, 각 문장은 명사의 수가 3개 이상으로 된 조사가 없는 연속된 명사열(마지막 명사는 조사가 있을 수 있다)을 하나 이상 포함한다. 실험 말뭉치에서 연속된 명사열은 총 개이며, 이중 복합명사구로 쓰일 수 있는 연속된 명사열은 총 2,982개이다.

각 문장에 대하여 복합명사구의 구묶음은 다음의 6가지의 경우로 나누어서 수행하였다.

1. 단순히 어휘정보([왼쪽명사 - 오른쪽명사]쌍)만을 이용
2. 왼쪽명사의 의미부류와 오른쪽명사의 어휘정보([왼쪽명사 - 오른쪽명사의미부류]쌍)를 이용한 경우
3. 왼쪽명사의 어휘정보와 오른쪽명사의 의미부류([왼쪽명사의미부류-오른쪽명사]쌍)를 이용한 경우
4. 3과 4를 같이 이용한 경우
5. 왼쪽명사의 의미부류와 오른쪽명사의 의미부류([왼쪽명사의미부류-오른쪽명사의미부류]쌍)을 이용한 경우
6. 3과 4, 그리고 5를 같이 이용한 경우

그리고 [표 6]은 위의 6가지 경우에 대한 실험결과이다. 정밀도(precision)은 전체 실험결과 수에 대한 정답의 비율이며, 재현율(recall)은 전체 정답에 대한 실험결과의 정답 수의 비율이다. 그리고 f-measure는 다음의 식을 이용하여 구한다.

$$f-measure = \frac{2 \times precision \times recall}{precision + recall} \times 100$$

표 6. 실험 결과

총 문장		총 복합명사구의 수	
1,000		2,982	
	[왼쪽명사 - 오른쪽명사]쌍		
1	올바른 구묶음 수	잘못된 구묶음 수	
	0	0	
	precision	recall	f-measure
	0%	0%	0%
	[왼쪽명사 - 오른쪽명사의미부류]쌍		
2	올바른 구묶음 수	잘못된 구묶음 수	
	1,338	660	
	precision	recall	f-measure
	66.97%	44.87%	53.73%
	[왼쪽명사의미부류 - 오른쪽명사]쌍		
3	올바른 구묶음 수	잘못된 구묶음 수	
	1,260	600	
	precision	recall	f-measure
	67.74%	42.25%	52.04%

4	2+3		
	올바른 구묶음 수		잘못된 구묶음 수
	1,918		464
	precision	recall	f-measure
	80.52%	64.32%	71.51%
5	[왼쪽명사의미부류 오른쪽명사의미부류]쌍		
	올바른 구묶음 수		잘못된 구묶음 수
	2,154		848
	precision	recall	f-measure
	83.55%	72.23%	77.48%
6	2+3+5		
	올바른 구묶음 수		잘못된 구묶음 수
	2,400		362
	precision	recall	f-measure
	86.89%	80.48%	83.57%

위 표에서 올바른 구묶음 수는 구묶음이 올바르게 수행된 복합명사구의 개수이며, 잘못된 구묶음 수는 구묶음이 되지 말아야 할 명사열이 구묶음이 된 경우이다.

2. 결과 분석

복합명사구사전을 구축하기 위해 사용된 (구문분석) 말뭉치와 실험을 위해 쓰인 (형태소분석)말뭉치는 서로 다른 말뭉치이다. [표 6]의 실험결과, 단순히 어휘정보만을 이용하였을 경우 성능이 0%가 나왔는데, 이것은 아무리 많은 말뭉치를 이용하여 복합명사구 사전을 구축한다고 하더라도 복합명사구를 이룰 수 있는 명사들의 조합이 많아 복합명사구의 구묶음을 수행하는 것은 불가하다는 것을 보여준다. 즉, 어휘정보만으로 복합명사구사전을 구축하는 것은 의미가 없음을 뜻한다. 하지만 어휘정보만을 사용하지 않고 의미부류를 이용하여 구묶음을 수행할 경우 단순히 명사의 어휘정보만을 이용하였을 때보다 성능이 상당히 증가한 것을 볼 수 있다(왼쪽이나 오른쪽 한쪽의 명사의미부류만을 사용하였을 경우 50%이상의 성능을 보임). 양쪽 명사 모두의 의미부류 쌍을 이용하였을 경우(실험5) 어느 한쪽의 의미부류만을 이용하였을 때(실험2와 3)보다 20%의 성능 향상이 있었으며, 모든 명사-의미쌍 정보를 이용하였을 경우 가장 높은 성능을 보였다. 이는 명사의미사전에 등록되지 않은 명사(즉, 의미부류를 알 수 없는 명사)가 포함된 복합명사구를 구묶음할 때, 어휘정보가 반영되었기 때문이다.

구묶음이 잘못된 경우는 [표 7]과 같이 분류할 수 있다.

표 7. 잘못된 복합명사구 구묶음의 분류와 비율

부분적으로 구묶음이 된 경우	74.13%
과도하게 구묶음이 된 경우	24.61%
잘못된 구묶음	1.26%

부분적으로 구묶음이 된 경우는 정답 복합명사구의 일부만이 구묶음이 되거나 되지 않은 경우를 말하며, 과도하게 구묶음이 된 경우는 정답 복합명사구의 범위를 넘어서 구묶음이 된 경우를 말한다. 그리고 잘못된 구묶음은 결과 구묶음의 일부가 정답 복합명사구의 일부만을 포함하고 있는 경우이다. 우선 부분적으로 구묶음이 된 경우가 가장 많은 비율을 차지하고 있는데, 여기에 가장 큰 영향을 미친 요소는 명사의미사전에 등록되지 않은 미등록 명사와 고유명사, 전문용어 및 외래어이다. 이러한 문제는 사전에 단어를 등록하면서 어느 정도 해결될 수 있으나 모든 단어를 등록한다는 것은 어느 정도 한계가 있어 사실 상 근본적인 해결책은 되지 못한다. 이것은 본 연구만 국한된 문제가 아니며 많은 연구가 필요한 부분이다. 과도하게 구묶음이 된 경우는 일부 부사성 명사까지 구묶음을 했기 때문에 발생한 경우가 가장 많았다. 예를 들어 '나머지'와 같은 명사는 "나머지 학생은 뒤에 앉으세요."와 같은 문장에서처럼 뒤에 나오는 명사와 복합명사구를 이룰 수도 있지만, "그는 당황한 나머지 유리컵을 깨고 말았다."와 같이 복합명사구를 이룰 수 없는 경우도 있다. 이런 명사로는 '가운데', '때', ' 경우' 등이 있다. 잘못된 구묶음의 경우도 이와 같은 부사성 명사로 인한 오류로 발생하였다. 이런 명사들의 경우 보통 앞에 '용언+는(ㄴ)/관형형 어미'가 오는 경우 앞의 용언과 결합하여 부사적인 역할을 하게 되며, 이런 부사성 명사에 대해서는 휴리스틱을 이용하여 어느 정도 해결이 가능하다. 그 밖에 과도한 구묶음을 수행하는 원인으로 명사의 의미가 중의성을 갖기 때문이다. 본 논문에서는 의미의 중의성은 고려하지 않았으며 단순히 코퍼스로부터 추출된 명사쌍의 명사들이 가지는 모든 의미부류들의 조합으로 복합명사구사전을 구축하였으며, 복합명사구를 구묶을 할 경우도 명사가 가지는 모든 의미부류 중 복합명사구를 이룰 수 있는 의미가 있으면 구묶음을 수행하였다.

이러한 문제를 해결하기 위해서는 의미중의성을 해소할 수 있는 방법에 대한 연구가 선행되어야 할 것이다.

VII. 결론 및 향후 연구

기존의 연구들은 조사가 없는 연속된 명사열들을 단순히 하나의 명사구로 묶어 처리를 해왔다. 하지만 이러한 처리 방법은 잘못된 구문구조를 야기할 수 있는 문제가 발생한다.

본 논문에서는 조사가 없는 연속된 명사열들을 구묶음하기 위해 세종구문분석말뭉치로부터 복합명사구를 이룰 수 있는 명사 두 개로 구성된 명사쌍들을 추출하고, 그 명사쌍으로부터 의미쌍을 추출하여 복합명사구 사전을 구축하는 방법과 두 개의 명사 또는 의미부류의 쌍으로만 이루어진 복합명사구사전을 이용하여 복합명사구의 구묶음을 하는 방법에 대하여 기술하였다. 본 논문의 방법으로 구문분석말뭉치로부터 명사쌍과 의미쌍을 자동으로 추출하여 사전을 구축할 수 있으며, 단지 두 개의 명사쌍을 이용하여 사전을 구축하기 때문에 사전 데이터의 추가 및 수정 또한 용이하다. 하지만 복합명사구사전의 질이 구문분석말뭉치의 질에 의존적이라는 단점 또한 존재하며, 다의적인 명사가 갖는 의미중의성 또한 해결해야할 과제이다. 그리고 미등록 명사 및 고유명사, 전문용어 , 외래어 등은 본 논문의 방법으로 모두 처리하기에는 무리가 있다. 이를 위해 복합명사구사전의 정제작업과 사전데이터의 수집 작업은 물론이며, 명사의 의미중의성 해소 및 본 논문의 방법으로 해결이 되지 않는 명사들을 처리하기 위한 휴리스틱 규칙에 대한 연구, 그리고 미등록어 처리에 대한 연구가 필요하다.

조사가 없는 연속된 명사열을 올바르게 구묶음을 하지 않고 구문분석의 단계로 넘기면 구문중의성이 발생하거나, 올바르지 않은 구문분석 결과를 내놓게 되므로 구문분석의 성능을 떨어뜨리게 된다. 따라서 본 연구는 구문분석의 성능을 높이는데 기여를 할 것을 기대한다.

참 고 문 헌

[1] S. Abney, "Parsing by Chunks," In R.C. Berwick, S.P. Abney and C. Tenny, editors, Principle-Based Parsing: Computation and Psycholinguistics, Kluwer, pp.257-278, 1991.

[2] Ramshaw, M. Marcus, "Text chunking using transformation-based learning," In Proceedings of the Third ACL Workshop on Very Large Corpora, Association for Computational Linguistics, pp.157-176, 1995.

[3] Bourigault, "Surface grammatical analysis for the extraction of terminological noun phrase," In Proceeding of the Fifteenth International Conference on Computational Linguistics, pp.977-981, 1992.

[4] Kupiec, "An algorithm for finding noun phrase correspondences in bilingual corpora," In Proceeding of the 31st Annual Meeting of the Association for Computational Linguistics, pp.17-22, 1993.

[5] Voutilainen, "NPTool, a detector of English noun phrase," In Proceedings of the Workshop on Very Large Corpora, Association for Computational Linguistics, pp.48-57, 1993.

[6] 양재형, "규칙 기반 학습에 의한 한국어의 기반 명사구 인식", 정보과학회 논문지, 제27권, 제10호, pp. 1062-1071, 2000.

[7] 황영숙, 정후중, 박소영, 곽용재, 임해창, "자질집합선택 기반의 기계학습을 통한 한국어 기본구 인식의 성능향상", 정보과학회논문지:소프트웨어 및 응용, 제29권, 제9호, pp.654-668, 2002.

[8] 서충원, 오종훈, 최기선, "어절의 중심어 정보를 이용한 한국어 기반 명사구 인식", 제15회 한글 및 한국어 정보처리 학술대회, pp.145-151, 2003.

[9] 최용석, 신지애, 최기선, "확률모형과 수식정보를 이용한 와/과 병렬명사구 범위결정", 정보과학회 논문지:소프트웨어 및 응용, 제35권, 제2호,

pp.128-136, 2008.

[10] 박의규, 나동열, "한국어 구문분석을 위한 구묶음 기반 의존명사 처리", 인지과학, 제17권, 제2호, pp.119-138, 2006.

저 자 소 개

안 광 모(Kwang-Mo Ahn) **준회원**

- 2007년 2월 : 충북대학교 컴퓨터공학과(공학사)
- 2009년 2월 : 충북대학교 컴퓨터공학과(석사)
- 2009년 3월 ~ 현재 : 충북대학교 컴퓨터공학과 박사과정

<관심분야> : 한국어 형태소분석 및 품사 태깅, 한국어 구문분석, 질의응답시스템

서 영 훈(Young-Hoon Seo) **종신회원**

- 서울대학교 컴퓨터공학과 졸업 학사(1983), 석사(1985), 박사(1991)
- 1994년 ~ 1995년 : 미국 Carnegie Mellon 대학 기계번역 센터 객원교수
- 1988년 ~ 현재 : 충북대학교 전자정보대학 컴퓨터공학부, 컴퓨터정보통신연구소

<관심분야> : 자연언어처리, 한국어 구문분석, 한영기계번역, 정보검색, 질의응답시스템