두 개의 명사쌍으로 이루어진 복합명사사전을 이용한 연속된 명사열의 구묶음

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

Chunking of Contiguous Nouns using Compound Noun Dictionary of Length Two

Kwangmo Ahn○, Young-Hoon Seo
Dept. of Computer Engineering, Chungbuk National University

요 약

구문분석에서 구문중의성(syntactic ambiguation)은 구문분석의 성능에 많은 영향을 미친다. 구문중의성을 일으키는 많은 요인들이 있지만, 특히 조사가 발달된 한국어의 구문분석에서 조사가 생략된 명사들은 구문중의성을 증가시키는 큰 요인 중 하나이다. 그렇기 때문에 조사가 없거나 생략된 연속된 명사열(contiguous nouns)의 길이가 길어질수록 구문중의성은 지수적으로 증가하게 된다. 따라서 현재까지의 연구에서는 이런 명사열들을 마치 하나의 명사처럼 구묶음을 하여 처리하는 경우가 많았다. 하지만, 조사가 없는 명사열들을 모두 하나의 명사구처럼 처리하여 구문분석을 수행할 경우, 주요 문장성분들이 잃어버리게 되는 경우가 발생한다. 따라서 본 논문에서는 하나의 명사처럼 쓰일 수 있는 조사가 없는 연속된 명사열을 복합명사구라고 정의하고, 두 개의 명사쌍으로 구축된 복합명사사전만을 이용하여 세 개 이상의 명사로 구성된 복합명사구들을 사전에 등록하지 않고도 복합명사구를 구묶음하는 방법에 대하여 기술한다.

실험을 위해 세종사전 150,546개의 예문에서 두 개 이상의 조사가 생략된 21,482개의 명사쌍을 추출하여 복합명사사전으로 변환하였으며, 총 6,316개의 사전 데이터가 구축되었다. 복합명사 구묶음 모듈은 조사가 생략된 명사열을 입력으로 받아서 우에서 좌로 검색하며 구묶음이 가능한 명사들을 연결하고, 연결된 명사들끼리 하나의 복합명사로 구묶음을 한다. 실험은 사전을 구축할 때 쓰였던 말뭉치와 사전을 구축할 때 쓰이지 않은 말뭉치를 이용하여 수행하였다. 결과는 사전을 구축할 때 쓰인 말뭉치를 이용하였을 때는 96.76%의 정확도를 보였으며, 사전을 구축할 때 쓰이지 않은 말뭉치를 이용하였을 경우는 12.23%의 정확도를 보였다.

1. 서 론

현재, 구문분석에 대한 연구 결과들은 응용분야(기계번역, 정보검색, Q&A 시스템 등)에 쓰이기에는 성능 상의 많은 한계를 가지고 있다. 특히 문장이 가진 구문중의성(syntactic ambiguation)은 구문분석의 성능에 가장 큰 영향을 주는 요소 중 하나라고 할 수 있다. 구조적 중의성이 많은 문장은 많은 분석 결과를 가지므로 분석 시간 및 분석 대상을 증가시켜 구문분석의 성능이 떨어지게 된다. 그리고 응용시스템의 입력으로 쓰이는 구문분석의 결과가 많으면, 응용시스템의 성능 또한 저하시키게 된다.

한국어는 조사가 발달한 언어이며 조사를 통하여 문장 내에서 격정보를 파악할 수 있다. 하지만 한국어는 조사가 생략되는 경우가 자주 있으며, 조사가 없거나 생략된 (앞으로는 "조사가 없는"이란 용어를 사용한다) 명사는 문장 내에서 그 격정보를 상실하게 된다. 즉, 조사가 없는 명사는 구문중의성을 일으키는 큰 요인 중 하나이다. 그런데 한국어 문장에서는 조사가 없는 연속된 명사열(contiguous nouns)이 나타날 경우가 종종 있으며, 이러한 경우 명사열의 길이에 따라 지수적으로 구문중의성이 발생하게 된다. 따라서 이러한 경우, 기존의 구문분석에서는 조사가 없는 명사열을 하나의 명사구처럼 구묶음을 하여 처리하는 경우가 많다. 하지만 이것은 주요 문장성분들을 잃어버리게 되는 경우가 발생하게 된다. 예를 들어, "스키가 한국에서도 대중적인 겨울 스포츠로 자리를 잡았다." 라는 문장에서 '겨울' 과 '스포츠' 는 둘을 묶어 하나의 명사처럼 쓰일 수 있지만, "지난 겨울 아파트 수도관이 동결되어 주민들을 불편을 겪었다." 라는 문장의 '겨울' 은 문장에서 부사적인 역할을 하며, 따라서 '아파트' 와는 같이 구묶음이 되어서는 안 된다. 이를 해결하기 위해 사전을 구축하여 처리하는 경우를 생각해 볼 수 있는데, 하나의 명사처럼 쓰일 수 있는 연속된 명사열들의 조합이 많은데다가 하나의 명사처럼 쓰일 수 있는 명사열들이 두 개 이상인 경우도 많이 존재하기 때문에 사실상 이들을 모두 사전으로 구축하는 것은 무리가 있다.

따라서 본 논문에서는 복합명사를 이룰 수 있는 두 개의 명사쌍만을 추출하여 복합명사사전을 구축하고, 이렇게 구축된 복합명사사전을 이용하여 조사가 없는 세 개 이상의 연속된 명사열들도 구묶음을 할 수 있는 방법에 대하여 기술한다. 문장 내에서 조사가 없는 연속적인 명사열을 추출하여 우에서 좌로 분석해 나가고, 이 때 복

합명사사전을 이용하여 우측의 명사와 좌측의 명사가 서로 복합명사구를 이룰 수 있게 되면 그 둘을 서로 연결해 나가게 된다. 이렇게 우에서 좌로 분석을 모두 마쳤을 때, 서로 연결된 모든 명사열들끼리 구묶음을 수행하여 복합명사구를 인식하게 된다. 자세한 것은 3장에서 다루도록 하겠다.

본 논문의 구성은 다음과 같다. 2장에서는 구묶음에 대한 기존의 연구에 대하여 기술한다. 3장에서는 하나의 명사처럼 쓰일 수 있는 두 개의 명사쌍을 이용하여 구축된 복합명사사전으로 세 개 이상의 연속된 명사열을 구묶음 하는 방법에 대하여 설명한다. 4장에서 실험을 통하여 본 논문의 방법에 대한 성능을 평가해 보며, 5장의 결론을 통하여 논문을 마치도록 한다.

2. 기존 연구

구묶음(chunking)을 하는 방법은 Abney에 의해서 처음으로 제안되었다[1]. Abney는 구문분석을 구묶음(chunking)과 붙임(attaching)의 단계로 나누었다. 사람이 문장을 읽을 때 끊어 읽는 운율적 휴지를 경계로 한 단어의 경계를 기준으로 문장 성분을 묶는 것을 구묶음으로 보았다. 이렇게 구묶음이 된 덩어리(chunk)는 그 다음 단계의 붙임모듈(attacher)이 분석해야 할 후보를 줄여 중의성 줄이고 성능을 증가시킬 수 있다. 예를 들어, "저 아름다운 여자는 나의 아내이고, 그 옆의 귀여운 아기는 나의 아들이다." 라는 문장에서 분석해야 할 후보의 수는 11개이다. 현재 구문분석의 복잡도는 $O(n^3)$이므로 위의 문장은 $11^3 = 1331$만큼의 복잡도를 갖게 된다. 하지만 "[저 아름다운 여자]는 [나의 아내]이고, [그 옆의 귀여운 아기]는 [나의 아들]이다." 와 같이 구묶음을 수행하였을 경우, 복잡도는 $4^3 = 64$로 감소하게 된다. 따라서 구묶음이 구문분석의 성능을 향상시킴에는 반박의 여지가 없다.

한국어의 경우도 많은 구묶음에 대한 연구들이 있어 왔다. [2,3,4]의 경우는 기반 명사구 인식에 대한 연구를 하였다. 기반 명사구(base NP)란 명사구 내부에 다른 명사를 포함하지 않는 명사구를 말하며, I(inside), O(out), B(begin) 등의 태그를 이용하여 구묶음을 표현한다. [2]는 형태소 태그 부탁 말뭉치를 규칙 기반 학습 알고리즘을 이용하여 규칙을 학습하며, 학습된 규칙을 이용하여 구묶음을 수행한다. [3]은 기본구를 인식하기 위한 자질들을 학습 알고리즘을 이용하여 선택하고, 선택된 자질집합을 이용하여 기본구 인식을 학습한다. [4]에서는 tri-gram HMM과 어절 문맥 정보를 이용하여 기반 명사구 인식의 성능을 높였다. [2,3,4]와 같은 기반 명사구 인식은 명사를 수식하는 관형사 또는 관형형 전성어미와 결합한 용언과의 구묶음을 처리할 수 있다. 명사구 인식에 관한 연구 중 등위접속명사구를 인식하는 연구도 있었다[5]. 이 연구에서는 병렬명사구의 대칭성과 교환정렬 모델 및 수식관계 정보를 이용하여 등위접속명사구를 인식하였다. 또한, [6]에서는 의존명사와 관련된 구묶음에 대한 연구를 수행하였다. 여기서는 의존명사를 단위 명사와 비단위 명사로 구분하였으며 이를 규칙화하여 의존명사 관련 명사구를 인식한다.

이밖에도 구묶음에 대한 많은 연구들이 있지만, 조사가 없는 연속된 명사열을 구묶음하는 연구들은 찾아보기가 힘들다.

3. 복합명사구의 구묶음

3.1 복합명사구의 정의

복합명사(compound noun)란, 본디 '눈물(눈+물)' 이나 '늦더위(늦+더위)' 와 같이 둘 이상의 말이 결합된 명사를 뜻하며, 복합명사구라 하면 이런 복합명사를 포함하는 명사구라 할 수 있다. 하지만 본 논문에서는 복합명사구를 다음과 같이 정의한다.

- 본 논문에서 복합명사구의 정의
  둘 이상의 조사가 없는 명사열에서 하나의 명사처럼 쓰일 수 있는 명사열의 묶음(단, 명사열의 마지막 명사는 조사를 포함할 수 있다)

그리고 다음은 본 논문에서 정의한 복합명사구의 예들이다.

- 복합명사의 예
(1) 가게 간판
(2) 관리 팀장
(3) 일급 자동차 정비사
(4) 방송 심의 규정 위반

위의 예에서 (1)과 (2)와 같이 두 개의 명사가 하나의 복합명사를 이루는 경우도 있지만, (3)과 (4)와 같이 세 개 이상의 명사들이 복합명사를 이루는 경우도 있다.

3.2 복합명사사전

조사가 없는 명사열들에 대해서 살펴보면, 표1과 같은 유형으로 나누어 볼 수 있다.

표1. 조사가 없는 명사열의 분류와 예

	분류	예시
1	**부사 관련 명사열**	올해 신입사원
2	**등위 접속 관계 명사열**	호두 밤 잣
3	**단위의존명사 관련 명사열**	남자 서너 명
4	**수식 관련 명사열**	장교 육성
5	**하나의 명사와 같은 명사열**	꼬마 전구

유형1의 경우는 '올해' 와 '신입사원' 이 조사가 없는 명사열을 이루고 있다. 하지만 '올해' 라는 명사는 문장 내에서 부사적인 역할을 하며, 따라서 신입사원과는 하나의 명사처럼 쓰일 수 없다. 유형2의 경우는 사실 ',(쉼표)' 가 생략되어 문법적으로 잘못된 경우이지만, 직관적으로 보았을 때 각 명사들이 등위접속관계라는 것을 알 수 있다. 유형3은

‘남자’의 수를 나타내는 ‘서너’라는 수를 나타내는 관형사, 그리고 단위를 나타내는 의존명사 ‘명’이 결합되어 하나의 명사구를 이루는 경우이다. 3번 유형과 관련된 연구는 또한 [6]에서 자세히 다루고 있다. 4번과 같은 유형은 좌측의 명사가 우측의 명사를 수식하는 형태이라고 볼 수 있다. 4번의 예시를 보게 되면 “장교 육성”은 명사 ‘장교’ 다음에 관형격조사 ‘-의’가 생략된 형태라고 볼 수 있다(관형격조사를 붙이게 되어도 명사열 본래의 의미에서 달라지지 않는다). 이런 유형은 좌측의 명사에 조사 ‘-의’ 뿐만 아니라 ‘-하는(또는 한)’, ‘-적(인)’ 등을 붙이게 되었을 때 수식관계를 이루게 되는 경우도 있다. 다음은 4번 유형의 명사열에 대한 몇 가지 예이다.

- 수식 관련 명사열의 예
  (1) 환경부(의) 장관
  (2) 엘리트(적인) 장교
  (3) 관리(하는) 담당

유형5의 경우는 좌측의 명사에 조사 없이 우측의 명사와 결합하여 하나의 명사처럼 볼 수 있는 경우이다. 정확히 말하자면 유형5의 예시는 띄어쓰기오류이지만 이런 형태의 명사열은 한국어의 문장에서 종종 나타나는 경우이기 때문에 유형5로 분류하였다.

본 논문에서 정의한 복합명사구는 표1의 유형4와 유형5의 경우이며, 복합명사사전을 구축하기 위하여 코퍼스로부터 조사가 없는 두 개의 명사열(마지막 명사는 조사가 붙을 수 있다.)을 추출하였다. 이 중 표1의 유형4와 유형5와 명사열을 선별하고 이것을 이용하여 복합명사사전을 구축하였다. 그림1은 추출된 명사열을 복합명사사전으로 변환하는 것을 보여준다.

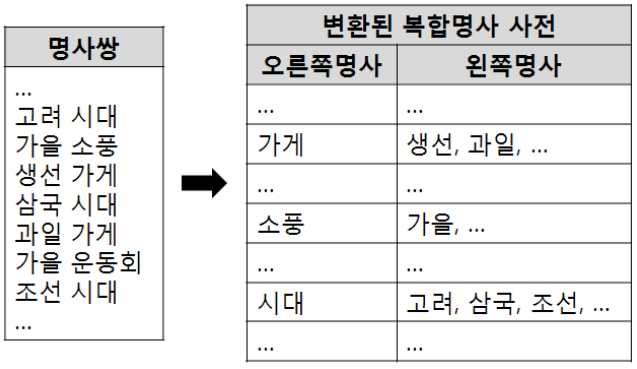

그림1. 두 개의 명사쌍으로부터 복합명사사전의 구축

그림1에서 오른쪽명사는 왼쪽명사의 다음에 올 수 있는 명사들을 의미한다. 예를 들어, ‘시대’의 왼쪽에는 ‘고려’, ‘삼국’, ‘조선’과 같은 명사들이 결합하여 “고려 시대”, “삼국 시대”, “조선 시대”와 같은 단어가 될 수 있음을 의미한다.

3.3 복합명사구의 구묶음

이번 절에서는 본 논문에서 구축한 복합명사사전을 바탕으로 복합명사와 같이 쓰일 수 있는 세 개 이상의 연속된 명사열에 대하여 구묶음을 수행하는 방법을 설명한다. 그림1과 같이 구축된 사전을 이용하여 구묶음을 수행할 경우 문장 내에서 발견된 명사열의 오른쪽부터 왼쪽으로 구묶음을 시도한다. 즉, 오른쪽에 나온 명사를 복합명사 사전에서 찾은 후 왼쪽에 나온 명사가 사전에서 찾아진 오른쪽명사의 왼쪽명사가 된다면 구묶음이 일어나게 된다.

하지만 세 단어 이상으로 구성된 복합명사구의 경우는 단순히 두 개의 명사열로 이루어진 복합명사구 사전을 이용하여서 구묶음이 제대로 일어나지 않게 된다. 예를 들어, “프로 야구 정기 시즌”이라는 명사열이 복합명사구를 이루고 있고 사전에 “프로 야구”, “야구 시즌”, “정기 시즌”이 등록되어 있다면, “[프로 야구] [정기 시즌]”으로 복합명사구가 불완전하게 인식이 될 것이다. 따라서 본 논문에서는 두 개의 단어로 이루어진 복합명사사전을 이용하여 세 단어 이상으로 이루어진 명사열을 인식하기 위해 다음과 같은 방법을 제안한다. $n_1\ n_2 \cdots n_k$로 이루어진 복합명사구가 입력으로 들어왔을 경우, 명사열의 우측에서 좌측으로 분석을 시작한다. 이때 “$n_{k-1}\ n_k$”이 복합명사사전에 등록이 되어 있다면 이 두 명사를 연결(link)한다. 연결에 성공하면 “$n_{k-2}\ n_k$”가 또한 복합명사사전에 등록이 되어 있는지 확인하며 등록이 되어 있다면 연결을 하게 된다. 이러한 과정을 연결이 실패하거나(즉, 복합명사사전에 등록이 되어 있지 않거나) “$n_1\ n_k$”까지 수행하고, 명사 n_{k-1}에 대해서도 같은 과정을 반복한다. 이런한 방법으로 “$n_1\ n_2$”까지 분석을 마친 후 연결된 명사들끼리 구묶음을 수행하게 된다. 다음은 본 논문에서 제시한 복합명사구의 구묶음 알고리즘이다.

명사열(noun list) $n_1\ n_2\ n_3 \cdots n_k$가 주어졌을 경우

```
CNL(noun list)
{
   i = k;
   j = i-1;
   LN(nj, ni);
   Combine();
}

LN(nj, ni)
{
   if(i == 1) return;
   if(nj is a left noun of ni){
      link nj and  ni;
      j--;
      LN(nj, ni);
   }
   else{
      i--;
      j = i-1;
      LN(nj, ni);
   }
}
```

복합명사의 구묶음 알고리즘

CNL(chunking of noun list) 함수의 입력은 조사가 없

는 명사열이며 입력된 명사열은 LN(link nouns) 함수를 통하여 오른쪽명사와 구묶음이 가능한 왼쪽명사들을 연결한다. 마지막으로 Combine 함수를 통하여 연결된 명사들끼리 구묶음을 한다.

그림2는 위의 과정을 명사열 "강남 초등학교 가을 소풍 주가 폭락"로 분석하였을 경우의 과정을 나타낸다.

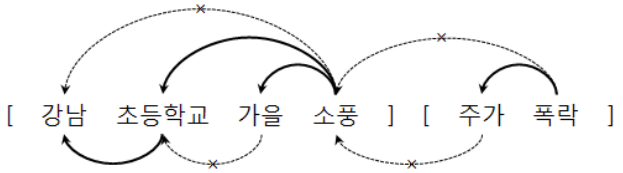

그림2. 복합명사가 인식되는 과정

이러한 방법으로 복합명사를 인식하는 이유는 왼쪽의 명사가 오른쪽의 명사의 수식어이기 때문이다. 예를 들어 "저 아기의 예쁜 인형"이란 명사구는 '예쁜', '아기의', '저'가 '인형'의 수식어가 될 수 있으며, '저'도 '아기의'의 수식어가 될 수 있어 하나의 명사구로 구묶음을 할 수 있는 것과 같은 원리이다.

4. 실 험

4.1 복합명사사전의 구축

복합명사 사전을 구축하기 위하여 세종 사전 예문 150,546개 문장에서 두 개 이상의 조사가 없는 명사열(마지막 명사는 조사가 붙을 수 있음)을 포함하는 문장을 추출하였다. 여기서 다시 두 개의 명사가 하나의 명사처럼 쓰일 수 있는 명사쌍을 추출한 후 이를 다시 사전으로 변환하였다. 복합명사사전의 구축 과정에서 '김 박사'와 같이 성이나 이름과 합쳐져서 복합명사를 이루는 경우는 제외하였다. 추출된 명사쌍은 총 21,482개이며, 이것으로부터 변환된 복합명사사전의 데이터 수는 총 6,316개이다.

4.2 실험 결과 및 분석

본 논문에서는 두 개의 명사쌍으로 구축된 복합명사사전을 이용하여 세 개 이상의 복합명사를 구묶음하는 방법에 대하여 설명하였다. 실험 결과는 표2와 같다. 실험데이터는 말뭉치로부터 세 개 이상의 연속된 명사로 구성된 명사열을 추출하였으며, 표2에서 실험데이터A는 복합명사사전을 구축하기 위해 사용했던 세종전자사전예문 말뭉치에서 추출한 명사열이며, 실험데이터B는 복합명사사전을 구축했을 때 사용하지 않은 세종현대원시말뭉치에서 추출한 명사열이다. 정확도는 다음과 같은 식으로 계산하였다.

$$\text{정확도} = \frac{b}{a} \times 100$$

표2. 복합명사 구묶음 실험 결과

실험데이터 종류		A	B
a	실험 복합명사의 개수	648	417
b	구묶음이 바르게 된 경우	627	51
c	구묶음이 잘못된 경우	21	366
정확도(%)		96.76	12.23

실험 결과 복합명사 구묶음의 정확도는 실험데이터A에 대한 정확도는 96.76%이며, 실험데이터B에 대한 정확도는 12.23%이였다. 이 중 구묶음이 잘못된 경우는 다음과 같은 두 가지 유형이 있었다.

(1) [예비역 장성 출신] [국방부 간부]가 조사를 받고 있다.
(2) 김 과장이 [해외 지사 설립] [건]에 대해 발안했다.

(1)의 경우, "예비역 장성 출신"의 명사들과 "국방부 간부"에 있는 명사들 중에서 서로 어울려 하나의 복합명사처럼 쓰이는 명사가 없기 때문이다. 이런 경우는 본 논문에서 제시한 방법만으로 해결하기 힘든 문제이다. 그런데 여기서 "출신"이라는 명사는 "~ 출신 ~" 형태로 양쪽의 명사를 이어주는 역할을 하는 단어이다. 이런 종류의 명사들은 따로 수집을 하여 양쪽 명사구를 이어주면 어느 정도 해결이 가능할 것이다. (2)의 경우는 사전에 "설립 건"이라는 명사쌍을 등록하지 않아 불완전하게 구묶음이 되었다. (2)의 경우는 사전에 데이터를 추가함으로써 쉽게 해결이 가능하다.

실험데이터A는 복합명사사전을 구축할 때 썼던 데이터이기 때문에 높은 정확도를 보였다. 하지만 실험데이터B의 결과는 정확도가 10%가 겨우 넘는 수준인데, 이것은 사전에 등록된 사전데이터만으로 실험데이터B에 존재하는 명사쌍을 모두 처리할 수 없었기 때문이다. 물론, 이것을 위해 매번 사전데이터를 추가하는 것 또한 쉽지 않은 작업임에 분명하다. 그래서 이를 위해 명사들의 의미쌍을 이용하여 본 논문의 방법으로 복합명사구를 처리하는 방법에 대하여 연구가 진행 중이며, 실험데이터B와 같이 사전에 등록이 되지 않은 명사들이 많아도 복합명사구가 인식될 것으로 기대된다.

5. 결론 및 향후 연구

본 논문에서 정의한 복합명사는 명사들 간의 조합에 따라 생성되며, 그 명사열의 길이가 정해지지 않았기 때문에 이런 복합명사를 처리하기 위한 사전을 구축하기는 매우 힘이 든다. 반면, 본 논문에서 제안하는 방법으로 복합명사를 구묶음 할 경우 복합명사사전은 단 두 개의 명사로 이루어진 복합명사를 이용하여 그보다 더 긴 복합명사들을 구묶음할 수 있으므로 사전의 구축이 용이하고, 사전데이터의 양도 훨씬 줄어들게 된다. 하지만, 단순히 단어들만을 사전에 등록하여 모든 복합명사를 처리하기에는 여전히 무리가 따른다. 또한, 명사쌍을 추출하여 이것으로부터 복합명사를 이룰 수 있는 것들을 선별하는 작업 또한 만만한 문제가 아니다. 따라서 향후 연구로는 복합명사사전에 의미정보를 이용하여 복합명사의 구묶음의 성능을 높이고, 기계학습기법 등을 활용하여

추출된 명사쌍 중 복합명사를 이룰 수 있는 것들을 선별하는 방법들에 대한 연구도 필요할 것이다

참고문헌

[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] 양재형, “규칙 기반 학습에 의한 한국어의 기반 명사구 인식”, 정보과학회 논문지, 제27권, 제10호, pp. 1062~1071, 2000.

[3] 황영숙, 정후중, 박소영, 곽용재, 임해창, “자질집합 선택 기반의 기계학습을 통한 한국어 기본구 인식의 성능향상”, 정보과학회논문지:소프트웨어 및 응용, 제29권, 제9호, pp.654~668, 2002.

[4] 서충원, 오종훈, 최기선, “어절의 중심어 정보를 이용한 한국어 기반 명사구 인식”, 제15회 한글 및 한국어 정보처리 학술대회, pp.145~151, 2003.

[5] 최용석, 신지애, 최기선, “확률모형과 수식정보를 이용한 와/과 병렬명사구 범위결정”, 정보과학회논문지:소프트웨어 및 응용, 제35권, 제2호, pp.128~136, 2008.

[6] 박의규, 나동열, “한국어 구문분석을 위한 구묶음 기반 의존명사 처리”, 인지과학, 제17권, 제2호, pp.119~138, 2006.