명사의 의미 정보를 이용한 복합명사 분석의 중의성 해소

+강유환0 *정천영 ++서영훈
+, ++충 북 대 학 교 컴 퓨 터 공 학 과
*혜천대학 컴퓨터통신계열 & 컴퓨터정보통신연구소
+pingkey@dcenlp.chungbuk.ac.kr, *cyjung@hcc.ac.kr, ++yhseo@cbucc.chungbuk.ac.kr

Analysis Disambiguation of Compound Nouns by Using the Semantic Information of Nouns in Korean

+Yu-Hwan Kang0 *Cheon-Young Jung ++Young-Hoon Seo
+, ++Dept. of Computer Engineering, Chungbuk National University
*School of Computer Science & Telecommunications, Hyechon College
& Research Institute for Computer and Information Communication

요 약

접사 처리는 복합명사 분석에서 중요한 문제인데 접사가 복합명사에 포함되어 있을 경우 여러 중의적 형태로의 분석이 가능하고 또한 미등록어 문제를 발생시킬 수 있기 때문이다. 단순한 접사 사전 정보만으로는 효율적인 분석을 수행할 수 없으므로 추가적인 정보가 필요하다. 본 논문에서는 접사로 인한 복합명사의 분석 중의성을 해소하기 위하여 명사의 의미 정보를 이용하는 방법에 대해 제안한다. 명사 의미 정보는 시소러스의 의미 계층 정보로 최상위 계층 정보와 하위 4계층의 정보로 구성된다. 명사+접미사 형태의 의미 결합 정보를 구한 후, 접미사를 포함하는 복합명사의 단위 명사들 간의 의미 결합 정보를 구한다. 이렇게 구해진 명사들 간의 의미 결합 정보는 사전 정보에 추가되며, 접사로 인한 중의적 분석 문제가 발생할 경우 명사들 간의 결합 정보를 이용하여 올바른 분석 후보를 선택한다.

1. 서론

한국어에서 명사와 명사는 띄어 쓰는 것을 원칙으로 하나 붙여 써도 무방하기 때문에 한국어에서의 복합명사는 매우 다양한 형태로 나타난다. 표기 방법이 다양하고 제약이 거의 없기 때문에 사람들 사이에서는 별로 문제가 되지 않지만 복합명사를 단위 명사로 분해하는 작업이 어렵고 중의적 분석이 가능하기 때문에 한국어 처리 시스템에서 복합명사의 처리는 매우 중요하다. 이러한 복합명사 분석의 어려움은 정보 검색 시스템이나 기계 번역 시스템과 같은 한국어 정보 처리 시스템에서 심각한 문제를 야기시킨다[1].

복합명사 분해 문제는 한국어뿐만 아니라 독일어나 중국어, 일본어 등 다른 언어에서도 발생한다. 하지만 독일어의 경우는 복합명사를 단위명사로 쉽게 분해할 수 있고, 중국어와 일본어는 단어의 경계가 없기 때문에 단어 분할 과정에서 복합명사 분해 문제가 발생한다[2,3,4].

복합명사 분해는 기계 번역 분야, 정보 검색 분야 그리고 맞춤법 검사 분야에서 필수 불가결한 작업이다. 한영 기계 번역 분야에서 복합명사 분해는 복합명사에 대한 올바른 대역어를 찾는데 필요하고, 정보 검색 분야에서는 시스템의 재현율을 높이기 위한 색인어를 추출하기 위해 필요하다. 또한 맞춤법 검사 분야에서 맞춤법 오류의 대부분을 차지하는 것이 띄어 쓰기 오류이기 때문에 복합명사에 대한 띄어 쓰기 방법으로 복합명사 분해가 사용될 수 있다. 근래에는 정보 검색 분야에서 처리하는 인터넷 문서에서 복합명사가 많이 나타나기 때문에 복합명사를 자동색인하기 위한 연구가 활발히 진행되고 있다[5,6,7,8].

한국어 복합명사 분해를 위한 기존 연구는 크게 복합명사의 음절 길이에 따라 복합명사 분해 패턴을 순서적으로 적용하여 분해하는 연구[9], 통계적 정보를 이용하여 복합명사의 중의적 분해

를 해결하는 연구[10] 등이 있다.

최재혁(1996)은 복합명사의 길이에 따른 분해 패턴을 미리 정의한 다음 이를 순서대로 적용함으로써 복합명사를 분해하는 방법을 제안하였다[9]. 예를 들어 5음절 복합명사의 경우 전체를 단위 명사로 보고 분해를 시도해 본 다음 2_3 음절 패턴을 적용하여 복합명사를 2음절과 3음절 단위 명사로 분해한다. 이 경우에서도 분해가 이루어지지 않은 경우 다시 3_2 음절 패턴을 적용하여 복합명사 분해를 시도한다. 하지만 이 연구는 복합명사의 중의적 분해 문제를 해결하지 못하는 단점이 있다.

윤보현(1995)은 복합명사를 분해할 때 중의적 분해 문제가 발생하는 경우 중심어 빈도와 통계적 선호 규칙을 적용함으로써 중의적 분해 문제를 해결하는 방법을 제안하고 있다[10].

강승식(1998)은 형태소 분석 결과로 추정된 복합명사를 단위 명사들로 분해하기 위해 네 개의 분해규칙과 두 가지 예외 규칙을 사용하여 가능한 분해 후보들을 생성하고, 분해 후보들에 대해 가중치를 부여함으로써 최적 후보를 선택하는 복합명사 분해 알고리즘을 제안하였다[11].

한국어 복합명사 분석의 정확도를 높이기 위해서는 복합명사의 중의적 분석 문제가 발생하는 경우와 복합명사에 미등록어가 포함되어 있는 경우의 처리가 잘 이루어져야 한다. 복합명사에 접사가 포함되어 있는 경우 복합명사의 중의적 분석 문제와 미등록어 처리 문제가 발생할 수 있는데, 단순한 접사 사전 정보만으로는 이를 효율적으로 해결할 수 없다. 따라서 본 논문에서는 접사로 인해 발생하는 중의적 분석 문제와 미등록어 발생 문제를 해결 할 수 있도록 명사의 의미 정보를 이용하는 방법에 대해 제안한다.

2. 복합명사 분석 방법

2.1 복합명사 분석 문제

복합명사 분석 문제는 중의적 분석의 발생과 복합명사에 미등록어가 포함된 경우로 인해 어려움이 발생한다. 5음절 복합명사 '부정합격자' 의 경우 다음과 같이 분석된다.

부정/nc+합격자/nc … (1)
부정합/nc+격자/nc … (2)

'부정합격자' 의 경우는 시스템 사전에 단위 명사 '부정', '부정합', '합격자', '격자' 가 존재함으로써 두 가지 형태의 분석 후보가 생성된다. 이 경우 올바른 후보 선택을 하기 위해 최장일치법을 이용하거나 통계 정보를 이용할 수 있지만, 최장일치의 경우 두 가지 형태가 동일하기 때문에 후보를 선택할 수 없고, 분할 패턴과 같은 통계 정보를 이용하여야 한다. 6음절 복합명사 '대학생선교회' 의 경우 시스템 사전만을 이용하여 복합명사를 분석하면 아래와 같이 잘못된 분석이 일어난다.

대학/nc+생선/nc+교회/nc … (3)

'대학생선교회' 의 올바른 분석은

대학생/nc+선교회/nc … (4)

이다. 하지만 '선교회' 가 사전에 등록되어 있지 않다면 (3)과 같은 형태로만 분석되는 문제가 발생한다. 미등록어인 '선교회' 를 사전에 등록하면 간단히 문제를 처리할 수 있지만 명사에 접미사가 결합된 미등록어 형태가 너무 많기 때문에 이들 모두를 사전에 등록하는 것은 어렵다. 접미사 정보를 이용하여 (4) 형태의 분석을 만들어 내도, (3)과 (4) 중 올바른 후보 선택의 문제가 발생하게 된다. 접미사가 결합된 미등록어 처리를 위해서는 접미사 정보를 이용하여 분석을 시도하여야 한다. '서울대교무처장' 의 경우 접사사전을 사용하여 복합명사 분석을 시도할 경우 아래와 같은 세 가지 형태의 분석이 가능하다.

서울대/N_+교무/nc+처장/nc … (5)
서울/nc+대교무/N_+처장/nc … (6)
서울/nc+대교/nc+무처장/N_ … (7)

하지만 이 중에서 최적 후보를 선택하는 것은 어려운 문제이다. 위의 예들을 바탕으로 복합명사 분석 시 발생하는 문제를 정리하면

① 복합명사가 두 가지 이상의 형태로 분석 가능한 경우 … (1), (2)
② 올바른 분석 후보가 없는 경우 … (3)
③ 복합명사에 미등록어가 포함되어 후보 선택이 어려운 경우 … (5)-(7)

의 세 가지로 요약할 수 있다.

①과 같이 분석 후보가 여러 개 생성된 경우 최장일치나 통계 정보를 이용하여 후보를 선택할 수 있지만 정확률이 낮아질 수 있다. ②의 경우 접사 처리를 고려하지 않고 시스템 사전만을 이용하여 복합명사 분석을 수행하면 잘못된 분석 결과를 얻을 수 있으므로, 복합명사에 접사가 포함되어 있는 경우 사전 검색이 완료되었어도 접사 처리를 추가로 해주어야 한다. 접사 처리가 이루어진 후에는 다시 올바른 후보 선택의 문제가 남게 된다. ③의 경우 접사 정보만으론 올바른 후보 선택이 어려우므로 후보 선택을 위한 추가적인 방법이 필요하다. 따라

서 올바른 분석 후보를 선택하고 분석 결과가 올바른지를 검증하기 위한 작업이 필요한데, 본 논문에서는 분석 후보의 선택과 검증을 위하여 명사간의 결합 가능성을 판단할 수 있는 명사의 의미 정보를 구축하고 이를 이용하여 복합명사 분석을 시도하는 방법을 제안한다.

2.2 명사 의미 정보의 필요성

2.1에서 분석 후보 (2)와 (3)은 사람이 직접 분석할 경우 쉽게 잘못된 분석으로 판단할 수 있다. 왜냐하면 사람은 이미 (2)와 (3)의 명사들이 서로 결합하기에 어울리지 않다는 정보를 알고 있기 때문이다. 하지만 컴퓨터는 단지 사전 정보만을 이용함으로써 (2)와 (3)이 잘못된 분석인지를 판단할 수 없다. 따라서 컴퓨터가 이해할 수 있는 형태의 정보를 사전에 추가하여야 하는데 본 시스템에서는 시소러스로부터 추출한 명사의 의미 정보를 이용하여 명사들의 의미 결합 정보 사전을 만드는 방법을 이용하였다. 참고적으로 본 시스템에서 사용한 시소러스는 약 10만 어절에 대해 구축되어 있으며 약 3000여 개의 의미 부류로 나누어져 있다. 명사 의미 정보는 시소러스로부터 추출한 명사의 의미 계층 정보로 최상위 계층 정보와 하위 4계층의 정보를 추출하여 구성하였다. 명사의 의미 결합 정보로부터 '대학' 과 '생선' 그리고 '생선' 과 '교회' 가 서로 결합하기에 부적합한 단어라는 지식을 얻음으로써 잘못된 분석 후보를 제거하고 올바른 분석 후보인지를 검증하게 된다. 예를 들어 '대학' 의 경우 시소러스로부터 조직, 단체, 기관 등의 의미 계층 정보를 얻을 수 있고, '생선' 의 경우 구체물, 무생물, 음식 등의 의미 계층 정보를 얻을 수 있다. 또한 명사의 의미 결합 정보에 조직(단체, 기관)이 구체물(무생물, 음식)과 서로 결합하기에 부적절하다는 의미 결합 규칙이 있다면 분석 후보 (3)이 잘못된 분석임을 알 수 있을 것이다. 하지만 한국어의 경우 명사와 명사간의 결합이 자유롭고 복합명사의 형태가 매우 다양하게 나타나기 때문에 모든 단위 명사들에 대해 의미 결합 관계를 조사하고 규칙을 작성하는 것은 불가능한 일이다. 따라서 본 논문에서는 명사간의 의미 결합 정보를 구축하는 대신에 명사에 결합된 접사를 바탕으로 명사간의 의미 결합 정보를 구축하였다.

2.3 접사와 결합되는 명사의 의미 정보 구축

먼저 접사와 명사간의 의미 정보를 구축하기 위하여 태그된 말뭉치로부터 접두사, 접미사 리스트를 추출하였으며, 상위 빈도를 보이는 접미사 목록은 아래 표와 같다.

표 1 상위 빈도를 보이는 접미사 목록

순위	접미사	빈도수
1	자	16488
2	회	13215
3	장	12656
4	사	8586
5	원	7140
.	.	.
.	.	.
21	가	3208
.	.	.
.	.	.

접미사는 그 의미에 따라 더 세분화할 수 있는데, 예를 들어 접미사 '가' 는 다음과 같은 다섯 가지 형태로 세분화된다. 먼저 어떤 방면의 전문인 등을 나타내는 '-家' (예: 정치가, 건축가, 양심가)와 성에 붙어서 그 성임을 나타내는 '-哥' (예: 홍가, 최가), 큰 도시를 작게 나눈 구획이나 특수한 성격의 거리를 나타내는 '-街' (예: 을지로 3가, 대학가, 금융가), 노래의 이름이나 종류를 나타내는 '-歌' (예: 흥부가, 이별가), 그리고 값이라는 뜻이나 원자가를 나타내는 '-價' (예: 최고가, 적정가, 2가 알코올)로 세분화할 수 있다. 또한 일정한 표면이나 끝나는 한계선을 나타내는 의미의 명사 가(예: 개울가, 난로가)와 옳거나 좋음을 나타내는 명사 가(예: 18세 이상 관람가)도 존재한다. 명사 '가' 는 띄어 쓰는 것이 원칙이나 말뭉치에서는 대부분 붙여 쓴 형태로 나타났기 때문에 접미사 '가' 의 범주에 포함시켰다. 접미사 정보를 추출함에 있어서 말뭉치에는 세분화된 접미사 정보가 들어 있지 않으므로 접미사 '가' 로 끝나는 명사를 세분화하여 추출하려면 모두 수작업에 의존해야 한다. 표 2는 말뭉치에서 접미사 '가' 로 끝나는 단일 명사의 출현 빈도를 보여주며, 표 3은 접미사 '가' 와 함께 나타난 명사의 의미 정보를 보여준다.

표 2 접미사 '가' 로 끝나는 단일 명사의 출현 비율

접미사 '가' 세분류	비율
-家	40.3%
-哥	2.7%
-街	15.1%
-歌	4.7%
-價	33.9%
명사	3.3%

표 3은 접미사 '가'로 끝나는 단일 명사의 의미 정보

접미사 '가' 세분류	최상위 의미	비율
-家	활동	49.1%
	지식	19.6%
	기타	31.3%
-哥	이름	100%
-街	곳	78.1%
	기타	21.9%
-歌	활동	23.1%
	지식	15.4%
	고유명사	61.5%
-價	활동	53%
	구체물	24%
	기타	23%
명사	곳	55.6%
	기타	44.4%

여기서 명사 의미 정보란 접미사를 떼고 남은 단일 명사의 의미를 뜻한다. 표 2와 표 3으로부터 접미사 '가' 로 끝나는 명사는 대부분 접미사 '-家, -街, -價' 와 함께 쓰이며, '활동' , '곳' 의 의미를 담고 있는 명사임을 알 수 있다. 따라서 접미사의 세부 의미와 상관없이 대체적으로 접미사 '가' 와 함께 나오는 명사는 활동' , '곳' , '지식' 등의 의미를 갖는 명사라고 생각할 수 있다. 다음에 나오는 표 4는 접미사 '가' 로 끝나는 2어절 복합명사(예: 가요평론가, 공급예정가, 고급주택가)의 명사간 의미 결합 패턴 정보를 보여준다.

표 3과 표 4의 정보로부터 접미사 '가' 에 대해 다음과 같은 의미 결합 정보를 작성할 수 있다.

```
SFX "가" {
    Noun-Sem : [활동,곳,지식,추상물,…]
    Pattern-Prev : [구체물,활동,조직,지식,…]
    Pattern-Next : BLANK
}
```

즉, 접미사 '가' 와 함께 결합되어 나타나는 명사의 의미 정보는 '활동, 곳, 지식, 추상물' 이며, 복합명사 형태일 경우 이전 단일 명사의 의미 정보는 '구체물, 활동, 조직, 지식' 이 될 수 있음을 나타낸다('Pattern-Next' 값은 접미사 다음 오른쪽에 나오는 명사의 의미 정보를 뜻하는데, 'BLANK' 라는 것은 아직 작성된 우측 의미 결합 정보가 없음을 뜻한다).

표 4 접미사 '가' 로 끝나는 2어절 복합명사의 의미 결합 정보

상위 의미 결합 패턴 목록
구체물+활동
활동+추상물
활동+활동
조직+활동
구체물+추상물
지식+활동

3. 시스템 설계 및 실험

한국어 복합명사의 경우 중의적 분석 문제와 접사를 포함하는 미등록어의 처리가 어려운 문제였다. 또한 사전을 통하여 분해된 복합명사라고 하더라도 복합명사 분석의 정확률을 높이기 위해 분석 결과가 올바른지 검증하는 단계가 필요하다. 복합명사 분석을 위한 시스템의 구성은 그림 1과 같다.

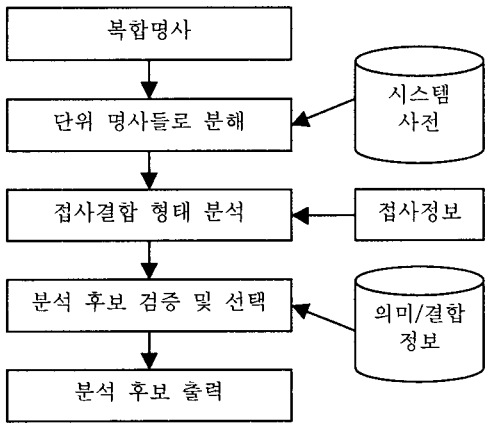

그림 1 복합명사 분석 시스템 구성도

먼저 복합명사가 입력되면 사전을 참조하여 복합명사로부터 가능한 모든 단위 명사들을 추출하고, 추출된 단위 명사들을 조합하여 복합명사 분석 후보를 만든다. 두 번째 단계에서는 단위 명사+접사 형태의 분석 후보를 만든다. 두 번째 단계를 수행하는 이유는 '대학+생선+교회' 와 같이 사전 검색을 통해 분석이 이루어 졌어도 미등록어가 복합명사에 포함되어 있어 잘못된 분석 결과를 내 줄 수 있기 때문이다. 마지막 단계에서는 분석된 후보들을 접사와의 결합 정보와 명사 의미 결합 정보를 이용하여 분석 후보를 검증한다. 만일 두 번째 단계에서 생성된 후보가 분석 후보 검증 단계에서 올

바른 후보로 판단될 경우에 후자의 것을 분석 후보로 선택해 준다. 결합 정보에 위배되는 후보는 제거되고 분석 후보들 중 최적의 후보를 선택한 후 이를 복합명사 분석 결과로 출력한다.

예를 들어 5음절 복합명사 '사회사상가' 가 입력으로 들어오면 복합명사 분석의 두 단계를 거쳐 '사회/nc+사상/nc+가/xsn' 와 '사회/nc+사/xsn+상가/nc' 의 분석 후보가 생성된다. 그 다음 분석 후보 검증 단계에서 '사상' 과 접미사 '가' 의 결합 형태를 분석하고 앞에 나오는 명사 '사회' 와의 결합 관계가 올바른 결합 형태인지를 조사한다. 같은 방법으로 두 번째 분석 후보에 대해서도 명사의 의미 결합 형태를 분석한다. 분석 후보들 중 의미 결합 정보가 올바른 후보가 있으면 이를 분석 후보로 선택하고, 만일 두 후보 모두 의미 결합 정보가 올바른 형태라고 판단이 될 경우에는 의미 결합 정보의 우선순위가 높은 후보를 선택한다.

실험에서는 접미사 '가' 에 대한 명사의 의미 결합 정보만이 구축되었기 때문에 접미사 '가' 로 끝나는 2어절 복합명사 중 중의적으로 분석되는 어절에 대하여 분석 후보 선택을 시도하였다. 명사의 의미 결합 정보를 이용하여 분석 후보를 선택한 경우의 약 96%가 올바른 후보를 선택한 것으로 나왔다. 접미사 '가' 에 대한 명사의 의미 결합 형태 중 말뭉치에서 낮은 빈도를 갖는 경우의 결합 형태는 제외 하였기 때문에 명사의 의미 결합 형태가 존재하지 않는 경우가 있었다.

4. 결론 및 향후 연구 과제

본 논문에서는 복합명사 분석을 위해 명사의 의미 정보를 이용하는 방법에 대하여 제안하였다. 한국어 복합명사는 형태가 다양하게 나타나므로 중의적 분석 문제가 발생하고, 미등록어를 포함하는 복합명사의 경우 분석 정확률이 떨어지는 문제가 있다. 또한 사전 정보만을 이용하여 복합명사를 분석할 경우 올바른 분석 후보의 선택과 미등록어 처리에 어려움이 있다. 따라서 복합명사 분석의 정확도를 높이기 위하여 접사와 결합하는 명사의 의미 정보와 결합 정보 등을 이용하여 복합명사 분석을 시도함으로써 올바른 분석 후보가 선택될 수 있도록 하였다. 그러나 아직 추출된 정보가 부족하기 때문에 충분할 실험을 수행하지는 못하였다. 앞으로 계속 명사 의미 정보를 구축하고 시스템에 적용해 봄으로써 복합명사 분해의 정확률이 높아질 수 있음을 검증하고자 한다. 향후 연구 과제로 시스템을 계속 확장해 나가고, 보다 체계적인 정보 구축을 위하여 다양한 형태로 분석 실험하고자 한다.

5. 참고 문헌

[1] Hyouk R. Park, Young S. Han, Kang H. Lee, Key-Sun Choi, "A Probabilistic Approach to Compound Noun Indexing in Korean Texts", Proceedings of the 16th International Conference on Computational Linguistics, vol.1, pp.514-518, 1996.

[2] T. Pachunke, O. Metineit, K. Wothke and R. Schmidt, "Broad Coverage Automatic Morphological Segmentation of German Words", Proceedings of the 14th conference on Computational Linguistics, pp.1218-1222, 1992

[3] Shiho Nobesawa, et al, "Segmenting a Sentence into Morphemes Using Statistic Information Between Words", Proceedings of the 15th International Conference on computational Linguistics, pp.227-233, 1994

[4] P. Wong and C. Chan, "Chinese Word Segmentation based on Maximum Matching and Word Binding Force", Proceedings of the 16th International conference on Computational Linguistics, pp.200-203, 1996

[5] 강승식, "한국어 형태소 분석을 위한 복합 명사의 인식 방법", 인지과학회 춘계 학술발표 논문집, pp.175-189, 1993

[6] 윤보현, 임희석, 임해창, "통계정보를 이용한 한국어 복합명사의 분석 방법", 한국정보과학회 봄 학술발표 논문집, pp.925-928, 1995

[7] 장동현, 맹성현, "효율적인 색인어 추출을 위한 복합명사 분석 방법", 제8회 한글 및 한국어 정보처리 학술발표논문집, pp.32-35, 1996

[8] 윤보현, 조민정, 임해창, "통계 정보와 선호 규칙을 이용한 한국어 복합 명사의 분해", 한국정보과학회 논문지(B), 24권 8호, pp.900-909, 1997

[9] 최재혁, "음절수에 따른 한국어 복합명사 분리 방안", 제8회 한글 및 한국어정보처리 학술발표논문집, pp.262-267, 1996

[10] 윤보현, 임희석, 임해창, "통계정보를 이용한 한국어 복합명사의 분석 방법", 한국정보과학회 봄 학술발표 논문집, pp.925-928, 1995

[11] 강승식, "한국어 복합명사 분해 알고리즘", 한국정보과학회 논문지(B), 25권 1호, pp.172-182, 1998