의미 정보를 이용한 한국어 복합명사 분석

김수남*, 원상연*, 권혁철*, 주종철*, 이상기**
부산대학교 전자계산학과*, 전자통신연구원**

Analysis of Korean Compound Noun usig Semantic Information

SuNam Kim, SangYun Won, HyukChul Kwon, JongChul Ju , SangGi Lee
Dept of Computer Science, Pusan National University, ETRI

요 약

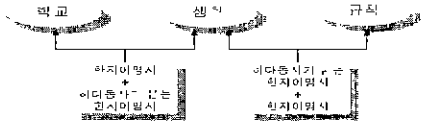
복합명사 분석은 조합이 자유롭고 길이의 제한이 없으므로 여러 가지 모호성을 발생시킨다. 이러한 모호성을 해결하는 기존 방법으로 사전을 이용하는 방법[2]과 통계적 정보를 이용하는 방법[3,4]이 있다. 본 논문에서는 하위 범주화된 어휘 정보를 가진 전자사전을 이용하여 복합명사를 분석한다. 그리고 어휘 정보만으로 처리했을 때 의미상으로 잘못된 분석이 발생할 수 있으므로 본 논문은 복합명사를 구성하는 어휘의 정보와 특성단어의 의미에 따른 복합명사 제약조건을 규칙베이스로 구축하여 분석에 이용한다. 또한 분석에 실패한 복합명사의 유형을 분석하여 각 유형에 따른 교정 방법도 제시한다. 실험 데이터는 부산일보, 교과서, 그리고 각종 문서에서 무작위로 추출한 27,945개의 복합명사를 사용하였다. 본 논문에서 제시한 의미적 제약조건을 이용하여 분석했을 때 복합명사로 잘못 쓴 어절의 검사율이 21% 향상되었다.

1 개 요

한국어에서 복합명사란 구성요소가 두 개 이상 결합하여 하나의 명사 기능을 하는 것을 말한다[1]. 오늘날 자연 언어 처리 분야인 기계번역이나 정보검색 기능을 향상시키기 위한 연구가 활발히 진행 중이다. 특히 복합명사의 분석은 자연언어 처리과정에서 차지하는 비중은 매우 크다. 신문이나 잡지 등 복합명사가 많이 사용되는 문서에서 시스템의 성능 향상을 위해 복합명사의 분석은 중요하다. 그러나 복합명사는 조합이 자유롭고 길이에 제한이 없어 그 분석에 어려움이 따른다. 더욱이 접사나 조사가 복합명사의 구성요소로 분석되는 문제는 복합명사 분석을 더욱 어렵게 한다. 이러한 복합명사 분석의 모호성을 제거하는 방법으로 사전을 이용한 방법[2]과 통계적 방법[3,4]을 들 수 있다. 통계 정보를 이용한 방법은 코퍼스에 있는 단어의 빈도를 사람이 수동 분석해야 하므로 시간적 · 경제적 손실이 크다[5]. 따라서 본 논문에서는 하위 범주화된 어휘 정보를 가진 전자사전을 이용하여 복합명사를 분석하였다. 그리고 어휘 정보만으로 분석했을 때 형태소적으로는 문제가 없지만 의미상으로 잘못된 분석이 발생할 수 있다. 본 논문에서는 복합명사 분석의 분석의 모호성과 의미 처리가 없어 틀린 복합명사를 맞게 분석하는 문제점을 극복하기 위해 복정단어의 의미 정보에 따른 제약조건을 규칙베이스로 구축하고 이를 이용하여 복합명사를 분석하였다. 그리고 분석된 결과 구성요소끼리 복합명사를 이루지 못할 때 오류의 유형을 분석하고 유형에 따라 대처어를 생성하는 방법을 이용하였다. 논문은 제2장 의미정보가 필요한 이유 제시, 제3장 의미정보에 대한 규칙베이스 구축, 제4장 의미정보를 이용한 복합명사의 분석 및 대처어 생성 과정, 제5장 실험 및 평가, 제6장 결론 및 개선 방향으로 구성된다.

2 의미 정보의 도입

일반적으로 복합명사는 하나의 문장처럼 구성요소 사이에 어휘·의미적 관계가 존재한다. 우리말 복합명사는 병렬구조를 이루며 첫번째 구성요소가 뒤에 오는 단어를 한정하는 기능을 한다. 그러므로 이 두 가지 구성요소 사이에는 서술어와 의존명사 사이의 관계처럼 어휘 의미적 연어관계가 성립한다. [그림 1]는 어휘 정보를 이용한 복합명사의 분석 예이다.



[그림 1] 어휘 정보를 이용한 복합명사 분석

[그림 1] 학교생활규칙에서는 한자어명사와 하나동사가 붙는

한자어명사가 형태소적으로 연어관계가 성립하므로 복합명사 분석이 가능하다. 그러나 형태소적 관계만 고려할 때 연어관계가 성립하나 의미상으로 맞지않는 관계가 분석되는 경우가 발생할 수 있다 [표 1]은 어휘 사전 정보만으로 분석했을 때 발생하는 잘못된 분석의 예이다

[표 1] 의미상으로 잘못 분석된 복합명사의 예

잘못된 복합명사	수정된 복합명사
건물내부	건물 내부
아기사육	아기양육
부산외각	부산외곽

[표 1]의 첫번째 예인 '건물내부'에서 '내부'는 형태소적으로는 앞 단어나 뒤 단어와 결합하여 언어관계를 이룰 수 있다. 그러나 의미상으로 볼 때 내부는 뒤에 오는 명사에 의존하며 복합명사를 이루는 단어로 언어관계가 성립하지 않는다 두번째 예인 '아기사육'에서 '사육'은 동물을 가리키는 구성요소와 복합명사를 구성할 수 있다. 그러나 사람인 '아기'는 '양육'하는 것이 의미상으로 올바르다 세번째 예인 '부산외각'에서 '외각'은 수학에서 사용되는 용어로 지역을 가리키는 '부산'과 의미상으로 연어관계가 성립되지 못한다 따라서 이러한 문제점을 해결하기 위해 특정단어의 의미적 제약조건을 규칙베이스로 구축하여 복합명사 분석을 시도하였다.

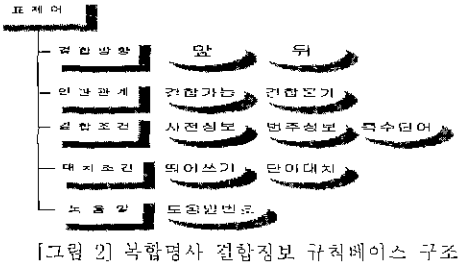
3 의미 정보에 따른 제약조건 규칙베이스 구축

복합명사는 현재 검사하고 있는 구성요소를 기준으로 앞, 뒤 명사와 언어관계를 고려해야 한다. 먼저 일반적으로 복합명사가 구성되는 조건[7]을 예를 통해 살펴보도록 한다

(1) 교통사고도 일종의 범죄라는 의식이·
(2) 아시아나기 사고추락 이후 시민들의·
(3) 출제된 문제는 사고능력과 서술능력을·
(4) 서양문화의 선구사고에 익숙한·

위의 예문[1]에 나오는 사고는 두 가지 의미를 가진다. (1)(2)는 갑작스럽게 어떤 일이 일어남을 뜻하고 (3)(4)는 생각하고 궁리함을 뜻한다. (1)(2)의 의미를 가진 사고는 능력, 서구 등과 같은 명사와 의미적 연어관계가 성립하지 않는다 이처럼 같은 명사라도 그 의미에 따라 연어관계가 성립하는 명사와 성립하지 않는 명사가 다르다 그리고 같은 의미로 사용되는 명사라도 (1)처럼 앞 단어와 언어관계가 성립하지만 (2)처럼 뒤 단어와는 언어관계가 성립하지 않는 명사도 있다

본 논문에서는 앞에서 언급한 예를 통해 살펴본 4가지 종류의 언어관계 제약조건을 특정단어의 의미 정보에 따라 규칙베이스를 구축한다 [그림 2]은 특정단어에 대한 규칙의 형태다



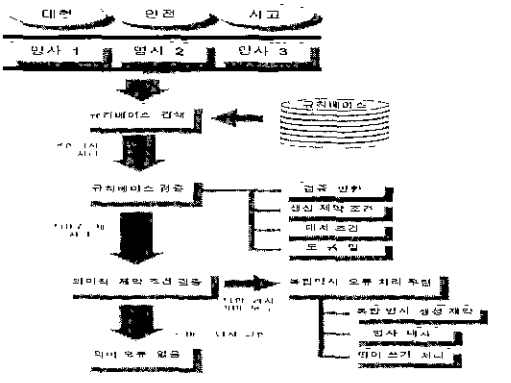
[그림 2] 복합명사 결합정보 규칙베이스 구조

하나의 특정단어는 뒤에 나오는 사전정보, 범주정보,특수단어에 대해서 결합방향, 결합가능성, 대치조건 그리고 도움말의 정보를 가진다 결합방향은 표제어를 중심으로 앞에오는 구성요소에 적용되는지 뒤에 오는 구성요소에 적용되는지를 가리킨다 연관관계는 의미적 연어관계가 성립하는지를 가리킨다 대치조건은 연어관계가 성립하지 않을 때 대치어 생성 방법을 설명한다 대치어 생성 방법에 대해서는 제4절에서 자세히 소개하겠다.

4 복합명사의 의미 분석 시스템

4.1 시스템 구성

앞 절에서 언급한 규칙베이스를 적용한 형태소 분석기의 복합명사 분석 모듈은 [그림 3]와 같다



[그림 3] 시스템 구조

입력된 어절을 전자사전에서 탐색하고 남은 미탐색어절[2]에 대해 조사, 접미사 전시를 차례로 수행한다. 만일 형태소 분석이 실패하면 복합명사 분석 모듈을 호출[8]한다. 호출된 분석 모듈은 탐색 어절 뒤부터 구성요소를 추출한다 추출된 구성요소가 특정단어의 의미적 연어관계를 검사할 필요가 없으면 어휘 정보만 이용해 연어관계를 판단한다 만일 구성요소가 의미적 연어관계 검사가 필요한 의미 검사 후보 명사[3]이면 의미적 연어관계 검사 모듈을 호출한다 호출된 모듈은 규칙베이스에서 의미 검사 후보 명사의 제약조건 규칙을 찾아 연어관계를 검사한다 이때 연어관계가 성립하면 복합명사가 가능함을 알리고 성립하지 않으면 규칙에 정의되어 있는 대치방법을 적용시켜 대치어를 생성한다

4.2 대치어 생성

복합명사를 분석한 결과가 모두 실패했을 때 대치어를 생성하는데 그 방법은 형태소적 분석과 의미적 분석 과정으로 나누어 처리한다. 형태소적 분석과정에서는 구성요소 사이를 띄어쓰기하여 대치어를 생성하고 의미적 분석과정에서는 규칙에 제시된 대치조건에 따라 생성한다. 대치조건은 3가지로 정의된다. 첫째, 단순히 두 구성요소를 띄어쓰기, 둘째, 구성요소가 후보명사로 대치, 셋째, 한 음절을 대치하는 방법이다.

(1) 띄어쓰기

1 제영숙,권석진,"빈발치도부터 추출된 통계 정보의 적용한 한국어 복합명사 분석",한국과학회논문지 1:897 pp 101~108

2 미탐색어절-검사어절 중 전자사전에서 탐색하고 남은 어절

3 의미 검사 후보 명사-의미적 연어관계를 검사하는 명사

띄어쓰기 방법은 통계 정보를 이용한 대치어 생성 방법[4]과 음절수에 따른 띄어쓰기 대치어 생성 방법[9] 등이 있다. 본 논문에서는 복합명사의 분석 때 연어관계가 성립하지 않는 구성요소 사이를 띄어 대치어를 생성한다

[표 2] 띄어쓰기의 예

복합명사로 잘못 쓴 어절	생성된 대치어
집행번제	집행 면제
가사노동절감	가사노동 절감

(2) 단어대치

대치어 생성 조건에서 제시하는 후보 명사를 의미상으로 복합명사를 이룰 수 없는 구성요소를 대신해서 삽입한다.

[표 3] 단어대치의 예

복합명사로 잘못 쓴 어절	생성된 대치어
사슴남용	사슴남획
장미사육	장미재배

(3) 음절대치

대치어 생성 조건에서 제시하는 음절 대치어를 의미상으로 복합명사를 이룰 수 없는 구성요소의 음절로 바꾸어 의미를 올바르게 한다.

[표 4] 음절대치의 예

복합명사로 잘못 쓴 어절	생성된 대치어
청구회수	청구횟수
한강도화	한강도하

5 실험 및 평가

본 논문에서 제시한 복합명사 분석 기법의 성능 평가를 위해 2가지 실험을 실시하였다 실험1은 어휘 정보만으로 처리한 복합명사 분석과 의미 정보를 도입한 분석의 성능 비교이고 실험2는 분석이 실패한 복합명사에서 대치된 생성어의 유형 분석이다 실험에 사용된 자료는 부산일보, 교과서, 그리고 각종 문서에서 무작위로 추출한 27,945개의 복합명사를 사용하였다.

(1) 어휘 정보 vs 어휘 정보 + 의미 정보

[표 5]에서 보듯이 어휘 정보만으로 복합명사를 분석했을 때 실험 어절 27,945개 중 3,682개가 분석에 실패하였다 반면 어휘 정보와 의미 정보를 동시에 이용했을 때 4,670개 어절이 복합명사 분석에 실패하였다

[표 5] 실험 1의 결과 (단위 개)

	분석 성공	분석 실패
어휘	24,263	3,682
어휘+의미	23,275	4,670

어휘 정보만 가지고 분석한 결과가 복합명사로 잘못 쓴 어절의 검사율이 낮은 이유는 의미상으로 해석이 불가능한 어절을 복합명사로 잘못 분석했기 때문이다. 본 실험에서는 전체 어절 중 복합명사로 잘못 쓴 어절의 검사율이 21% 향상되었다.

(2) 대치어 생성 분석

어휘 정보와 의미 정보를 이용한 복합명사 분석에서 복합명사로 잘못 쓴 어절이 4670개 발견되었다 이 어절에 대해 올바른 어절 생성시 각각 띄어쓰기 83%, 단어대치 3.4%, 음절대치 8.9%의 교정방법을 사용하였다

[표 6] 실험 2의 결과 (단위 개)

띄어쓰기	단어대치	음절대치	생성 실패
3881	160	416	213

여기서 띄어쓰기 대치어 생성 비율은 어휘 정보만으로 처리한 복합명사 분석과 의미 정보를 도입한 분석에서 생성한 대치어의 합이다. 그리고 미등록어에 의한 대치어 생성의 실패도 4.5% 나타났다.

6 결론

본 논문은 특정단어의 의미 정보를 이용하여 복합명사의 의미적 연어관계를 분석하였고, 분석 실패시 올바른 어절로 교정하는 방법을 제시하였다. 제시된 방법의 장점은 다음과 같다

첫째, 의미적 연어관계를 분석하여 복합명사 분석의 정확성을 높였다 실험에 사용한 27,945개의 복합명사에서 복합명사를 이룰 수 없는 단어의 검사율을 어휘 정보만 사용한 분석보다 21% 높였다

둘째, 복합명사 분석 실패시 올바른 어절로 교정하는 방법을 제시하였다 분석에 실패한 복합명사 어절에 대해 올바른 대치어를 제시하여 복합명사 사용에 도움을 주었다

셋째, 접사나 조사가 붙은 복합명사를 분석할 수 있다 하위범주화된 어휘 정보를 가진 전자사전을 이용하여 접사나 조사를 분리하여 접사나 조사에 의해 분석이 실패되는 것을 막았다

향후 연구 과제로는 특정단어의 의미 제약조건을 규칙화한 규칙베이스를 자동으로 구축하는 방법과 미등록어가 포함된 복합명사의 처리 방법을 연구하고자 한다.

참고 문헌

[1] 동아 새국어사전 이기문감수 동아출판

[2] 김지영, 권혁철, 합성명사의 의미 관계 분석 시스템을 위한 지식베이스 구축 기법, 한국정보과학회 가능 학술발표논문집, 1992. pp 985~968

[3] Chae Young-Soog, Dong-In Park, Hyuk-Chul Kwon. Analysis of Korean Compound Noun Based On Statistics Collected from Corpus, The 17th International Conference on Computer Processing of Oriental Languages (ICCPOL97), Vol.11, 1997, pp 709~714

[4] 윤보현, 임희석, 임해창, 통계정보를 이용한 한국어 복합명사의 분석방법, 한국정보과학회 봄 학술대회논문집, 제22권, 1호, 1995, pp925~928

[5] 박혁로, 신중호, 비터비 학습 알고리즘을 이용한 한글 복합명사 분석, 한국정보과학회 가을 학술대회논문집, 1997, pp219~222

[6] Makoto Nagao, [자연언어처리], 홍릉과학출판사, 1998

[7] 채영숙, 권혁철, 말뭉치로부터 추출된 통계 정보를 활용한 한국어 복합명사 분석, 한국인지과학회 논문집, Vol.8, No. 2, 1997, pp 101~108

[8] 권혁철, 채영숙, 김재훈, 김민정, 한국어 접사 검색을 위한 헝대소 분석 기법, 우리말 정보화 큰잔치, 국어정보학회, 1991, pp.179~186

[9] 최재혁(1996), 음절수에 따른 한국어 복합 명사 분리 방안, 한글및한국어정보처리, 1996, pp.262-267