명사 출현 특성을 이용한 효율적인 한국어 명사 추출 방법

An Efficient Method for Korean Noun Extraction Using Noun Patterns

저자
(Authors)

이도길, 이상주, 임해창
Do-Gil Lee, Sang-Zoo Lee, Hae-Chang Rim

출처
(Source)

정보과학회논문지 : 소프트웨어 및 응용 30(1·2), 2003.2, 173-183 (11 pages)
Journal of KISS : Software and Applications 30(1·2), 2003.2, 173-183 (11 pages)

발행처
(Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/Article/NODE00615660

APA Style

이도길, 이상주, 임해창 (2003). 명사 출현 특성을 이용한 효율적인 한국어 명사 추출 방법. 정보과학회논문지 : 소프트웨어 및 응용, 30(1·2), 173-183.

명사 출현 특성을 이용한 효율적인 한국어 명사 추출 방법

(An Efficient Method for Korean Noun Extraction Using Noun Patterns)

이 도 길 † 이 상 주 †† 임 해 창 †††

(Do-Gil Lee) (Sang-Zoo Lee) (Hae-Chang Rim)

요 약 형태소 분석을 한 후 명사를 추출하는 방법은 모든 어절에 대해 빈번한 사전 참조와 음운 복원을 위한 규칙 적용을 수행하므로 많은 연산을 필요로 하고 중의성이 있는 어절에 대해 모든 가능한 분석결과를 생성하므로 명사 추출의 관점에서는 비효율적이다. 본 논문에서는 명사 추출의 관점에서 형태소 분석시 불필요한 연산을 줄이기 위해 명사 출현 특성을 고려하는 명사 추출 방법을 제안한다. 명사 출현 특성은 명사의 존재에 대한 긍정적 또는 부정적인 단서를 표현하는 한국어의 특성으로, 배제 정보와 명사 접미 음절열이 있다. 배제 정보는 명사가 없는 어절을 미리 배제하여 형태소 분석에 요구되는 탐색 공간을 줄이고, 명사 접미 음절열은 바로 앞에 있는 명사를 검사함으로써 단순한 방법으로 명사를 추출하거나 미등록어를 인식하는 데에 사용한다. 또한 본 논문에서는 형태소 분석시 복잡한 음운 현상을 처리하기 위해 많은 음운 규칙을 적용하는 대신 음운 복원 정보를 사용하여 음운 현상을 처리한다. 실험 결과에 의하면 본 방법은 기존의 형태소 분석 방법에 의한 명사 추출에 비해 정확도는 떨어지지 않으면서 수행 속도 면에서 매우 효율적임을 알 수 있다.

키워드 : 명사 추출, 형태소 분석, 미등록어 추정

Abstract Morphological analysis is the most widely used method for extracting nouns from Korean texts. For every Eojeol, in order to extract nouns from it, a morphological analyzer performs frequent dictionary lookup and applies many morphonological rules, therefore it requires many operations. Moreover, a morphological analyzer generates all the possible morphological interpretations (sequences of morphemes) of a given Eojeol, which may be unnecessary from the noun extraction's point of view. To reduce unnecessary computation of morphological analysis from the noun extraction's point of view, this paper proposes a method for Korean noun extraction considering noun occurrence characteristics. Noun patterns denote conditions on which nouns are included in an Eojeol or not, which are positive cues or negative cues, respectively. When using the exclusive information as the negative cues, it is possible to reduce the search space of morphological analysis by ignoring Eojeols not including nouns. Post-noun syllable sequences(PNSS) as the positive cues can simply extract nouns by checking the part of the Eojeol preceding the PNSS and can guess unknown nouns. In addition, morphonological information is used instead of many morphonological rules in order to recover the lexical form from its altered surface form. Experimental results show that the proposed method can speed up without losing accuracy compared with other systems based on morphological analysis.

Keyword : noun extraction, morphological analysis, unknown word guessing

† 학생회원 : 고려대학교 컴퓨터학과
dglee@nlp.korea.ac.kr
†† 종신회원 : (주) 엔엘피솔루션 대표이사
zoo@nlpsolution.com
††† 종신회원 : 고려대학교 컴퓨터학과 교수
rim@nlp.korea.ac.kr
논문접수 : 2001년 10월 31일
심사완료 : 2002년 11월 18일

1. 서 론

명사 추출은 문서에 존재하는 모든 명사를 찾는 작업이다. 일반적으로 한국어 정보검색에서는 문서를 대표하기 위한 색인어로서 명사를 사용한다. 따라서 명사 추출기의 성능은 정보검색 시스템의 성능과 밀접한 관계에

있다. 명사 추출기는 정보검색의 자동색인 시스템의 필수적인 요소이고 문서분류, 문서요약, 정보추출 등의 자연어처리 응용 분야에서도 사용되고 있다.

한국어에서 명사를 추출하기 위해 가장 많이 사용되는 방법은 형태소 분석이다. 형태소 분석은 자연어처리의 가장 기본적인 단계로서 지금까지 오랜 기간 동안 연구되어져 왔고 다양한 방법을 이용한 여러 시스템들이 개발되었다. 이들 중에는 자연어처리 분야의 기술을 필요로 하는 여러 분야에서 실제로 사용되고 있는 것도 있다. 그러나 인터넷을 통해 공유되는 텍스트 정보가 기하급수적으로 증가하고 있는 현실을 감안하면 형태소 분석 속도는 여기에 미치지 못하고 있고, 아직 개선의 여지가 있다.

영어를 비롯한 굴절어(inflectional language)와는 달리 한국어는 터키어, 핀란드어, 일본어 등과 같은 교착어(agglutinative language)로서 한 어절이 하나 이상의 형태소로 이루어져 있다. 또한 다양한 음운 현상으로 인해 원형이 변형된 경우가 매우 많다. 명사를 추출하기 위해서는 어절을 이루고 있는 형태소들을 올바르게 인식해야 한다. 이러한 작업을 위해서는 빈번한 사전 참조와 여러 규칙을 적용해야 하므로 많은 연산이 필요하다. 뿐만 아니라, 형태소 분석은 어절에서 모든 가능한 분석 결과를 생성한다. 명사 추출의 관점에서는 어절 내에 있는 명사를 인식하여 추출하는 것이 목적이므로 모든 가능한 분석 결과를 생성하는 것은 불필요할 수 있다.

미등록어 문제는 자연어처리 응용 시스템에서 가장 큰 문제 중의 하나이다. 미등록어란 사전에 등록되어 있지 않은 단어를 말하는데, 미등록어의 대부분은 고유 명사, 신조어, 전문 용어, 약어, 외래어와 같은 명사이다. 기존의 연구들은 미등록어 문제를 완화하려고 명사를 획득하기 위한 노력을 해왔다. 그러나 명사는 개방 범주(open category)에 속하기 때문에 수시로 만들어지고 없어지므로 모든 명사를 사전에 추가하는 것이 불가능하고, 이러한 사전에 대한 구축과 유지 및 보수가 매우 어려우며, 비용에 비해 효과는 그다지 크지 않다. 게다가 미등록어는 문서의 내용을 표현하는 데 있어서 중요한 단어로서 사용되는 경우가 많다. 따라서 미등록어를 정확히 인식하는 것이 매우 중요하다.

본 논문에서는 한국어 문서로부터 효율적으로 명사를 추출하는 방법에 대하여 기술한다. 본 논문은 다음과 같이 구성되어 있다. 2장에서는 관련 연구에 대해 논하고, 3장에서는 명사 출현 특성에 대해서, 4장에서는 효율적으로 명사를 추출하는 방법에 대해서 기술한다. 5장에서는 실험 및 평가에 대해서 논한다. 마지막으로 6장에서 결론을 맺는다.

2. 관련연구

1999년에 열린 "형태소 분석기 및 품사 태거 평가 대회(이하 MATEC99)"에서는 형태소 분석기, 품사 태거, 명사 추출에 대한 평가를 했다[1]. 이 중에서 명사 추출에 대한 참가자들의 방법은 다음과 같이 크게 세 가지로 나눌 수 있다.

형태소 분석기를 이용하는 방법[2-4]

명사를 추출하기 위한 가장 보편적인 방법인 형태소 분석기를 이용하는 방법은 형태소 분석 결과로 얻어진 모든 명사를 추출하는 방법으로서 명사 추출의 재현율은 높으나, 어절의 중의성으로 인해 정확하지 않은 결과가 포함되어 정확률이 대체로 낮다.

형태소 분석기와 품사 태거를 이용하는 방법[5,6]

이 방법은 품사 태거에 의해 명사로 결정된 단어만을 추출한다. 중의성을 해결하기 때문에 가장 정확한 결과를 얻을 수 있으나 형태소 분석과 품사 부착 과정을 거쳐야 하므로 긴 수행 시간이 필요하다. 또한 품사 태거는 입력 단위가 문장이므로, 명확하게 문장 분리가 되어 있지 않은 문서에 대해서는 전처리 작업으로서 문장 분리 과정을 거쳐야 한다.

언어분석 도구를 사용하지 않는 방법[7,8]

이 방법은 형태소 분석과 같은 언어분석 도구를 사용하지 않고 사전에 저장된 어휘 정보만을 사용하여 명사 여부를 판단한다. 이운재(1999)는 먼저 불용어 사전을 검색하여 해당 어절이 불용어인 경우에 제거하고, 최장일치 방법에 의해 명사 후보를 추출한다. 이 때 명사 후보가 이미 사전에 있는 명사이면 해당 명사를 추출하고 그렇지 않으면 조사/어미 사전을 이용하여 명사를 추정한다. 장동현(1999)은 말뭉치에서 추출한 명사 리스트로부터 트라이 구조의 정방향 역방향 사전을 구성하고, 이 두 가지 사전과 조사/어미 사전에 기반하여 최장 일치 방법과 분석 순서를 정의해 놓은 몇 가지 규칙을 이용하여 명사를 추출한다.

이 방법은 시스템이 단순하고 구현이 쉬우며 분석 속도가 빠른 반면, 다른 방법들에 비해 잘못된 분석을 하는 경우가 많다. 특히 형태소 분석을 하지 않으므로 다양한 활용형이 존재하는 용언을 인식하는 것이 어렵다. 경우에 따라서, 미등록 명사를 추정할 때 '버려'나 '구워'와 같은 용언 어절 전체를 미등록 명사로 추정하거나, '은' 또는 '는'과 같이 조사와 어미로 동시에 사용되는 형태소의 경우에 '위해서는'이나 '깊은'으로부터 '위해서'나 '깊'이라는 잘못된 명사를 추정할 가능성이 있다.

정확한 분석을 얻기 위해서는 형태소 분석기와 품사 태거를 함께 사용하는 것이, 빠른 분석을 위해서는 비교적 간단한 방법을 사용하는 것이 유리하다. 분석할 자료의 양과 시스템의 사용 목적에 따라 이러한 분석의 정도를 결정하는 것이 바람직하다. 분석 시간과 정확도는 이율배반적(trade-off)인 관계에 있다. 정확한 분석을 하기 위해서는 그만큼 많은 연산과정이 필요하고 그에 따라 많은 시간이 소모된다. 이 논문에서는 효율성의 기준을 불필요한 연산을 줄임으로써 수행 시간을 줄이는 데 두고 있다. 이와 관련해 대용량의 문서에서 정확도에 크게 영향을 미치지 않으면서도 빠르게 명사를 추출하는 방법에 대해 논의하고자 한다.

3. 명사 출현 특성

명사 출현 특성이란 어절 내에 명사가 나타나는 조건과 나타나지 않는 조건에 대한 한국어의 특성을 말한다. 어절을 이루는 음소나 음절이 특정한 조건이 될 때는 명사가 나타나지 않는다거나, 특정 조건에서는 명사가 나타날 수 있다는 특성은 빠르고 정확하게 명사를 추출하는 데에 유용한 단서가 될 수 있다. 본 논문에서는 명사가 나타나는 특성에 대한 정보는 명사 접미 음절열로, 명사가 나타나지 않는 특성에 대한 정보는 배제 정보로 표현한다.

3.1 배제 정보

한국어 어절 중에는 첫 음절의 종성이나 처음 2음절 또는 3음절을 살펴보았을 때 또는 어절 가운데에 특정한 문자열이 나타나는 경우, 전체 어절 내에 명사가 거의 나타나지 않는 경우가 있다.

음소 단위 배제 정보

어절의 첫음절에 존재하는 특정 종성의 집합

"ㅆ", "ㄵ", "ㅎ", "ㄸ", "ㄼ"

부분 어절 단위 배제 정보

어절의 처음에 나타나는 특정 부분 어절의 집합

예) "같", "보였", "하였"

어절의 어느 위치나 존재하는 특정 부분 어절의 집합

예) "다른", "렇"

어절 단위 배제 정보

명사가 존재하지 않는 고빈도 어절의 집합

예) "가까운", "갖고"

그림 1 배제 정보의 분류

또한 빈도가 높은 어절 중에서 명사가 존재하지 않고 중의성이 거의 없는 어절이 많다. 한국어 어절에서 명사가 나타나지 않는 특성에 대한 정보를 본 논문에서는 배제 정보라고 부른다. 배제 정보를 이용하면 명사가 없는 어절에 대한 분석과정을 생략함으로써 형태소 분석 과정에 필요한 탐색공간을 줄일 수 있다.

그림 1은 본 논문에서 사용한 배제 정보를 종류별로 분류한 것이다.

부분 어절 단위 배제 정보는 45개, 어절 단위 배제 정보는 778개를 사용하고 있다. 본 논문에서 사용한 배제 정보의 예는 그림 2와 3에 있다.

가볍	감았	같	거쳐	걸려	걸어
걸쳐	것	그걸	그것	그녀	그들
그런	나는	다시	대해	되는	되어
되었	될	들어	따라	때문	몰았
보았	보였	싶	아니	어떤	어떻
없	위해	의해	이러	일어났	일이
주었	하고	하는	하면	하였	하지
한다	합니	해서			

그림 2 부분 어절 단위 배제 정보

가장	가지	가지고	거야	건	걸
게	그	그가	그는	그들은	그를
그의	나를	난	내	내가	다
대한	더	데	된	두	듯
듯한	마치	먼저	몇	모두	못지

그림 3 어절 단위 배제 정보의 예

형태소 분석에서 분석 효율을 높이고자 하는 시도가 있었다[9,10]. 강승식(1995)은 한국어의 음절 특성을 이용하여 과다한 용언 분석 후보를 줄였고, 임희석(1995)은 형태소 분석시 불필요한 후보의 생성을 줄이기 위해 배제 정보를 사용하였다. 임희석(1995)과 본 논문에서 사용한 배제 정보와의 차이점은 임희석(1995)은 특정 음소, 음절, 문자열은 특정 형태소 내에 사용될 수 없다는 가정에 기반하고, 본 논문에서는 특정 음소나 음절로 시작되거나 특정 음절열이 존재하는 어절 또는 특정한 어절 자체는 명사를 포함하고 있는 어절에 나타나지 않는다는 가정을 사용한다. 또한 임희석(1995)은 형태소 분석 과정 중에 불필요한 후보의 생성을 줄이기 위해 배제 정보를 사용했으나 본 논문에서는 어절에 대한 분석 자체를 배제하기 위해 이러한 정보를 사용한다.

3.2 명사 접미 음절열

명사 접미 음절열은 체언 뒤에 결합되는 음절의 열로서 정의하는데, 명사의 출현에 대한 좋은 단서가 된다.

어절에서 명사 접미 음절열이 발견되면 바로 그 앞에 위치한 체언을 검사함으로써 복잡한 형태소 분석 과정을 거칠 필요없이 명사를 추출할 수 있으므로 분석속도를 높일 수 있다.

한국어는 교착어이므로 명사와 결합될 수 있는 형식 형태소의 조합(결합)은 이론적으로는 매우 많다. 그러나 실제 언어 현상에서는 일정한 수가 반복되어 사용되므로 명사 접미 음절열을 이용하는 것은 의미있다. 명사 접미 음절열의 유형은 조사 및 “하다”, “되다”, “시키다” 등과 같은 용언화 접미사의 활용형과 “밥먹다”와 같이 흔히 명사와 결합되어 복합용언으로 사용되는 용언의 활용형이 있다.

본 논문에서 명사 접미 음절열은 품사부착된 말뭉치로부터 빈도가 2이상이고, 두 종류 이상의 체언과 결합한 것만 추출하였다. 말뭉치로부터 모든 명사 접미 음절열을 추출하게 되면 사전의 크기가 커지고 과분석이 발생할 수 있으며, 말뭉치 자체의 오류로 인해 잘못된 명사 접미 음절열이 추출될 수 있기 때문이다. 본 논문에서 사용한 명사 접미 음절열의 수는 2,576개이다. 그림 4는 용언화 접미사 “시키다”의 활용형으로 사용된 명사 접미 음절열의 예이다.

시키거나	시키게	시키겠다고	시키겠다는	시키고
시키고는	시키고서야	시키고야	시키고자	시키구
시키기	시키기가	시키기까지	시키기는	시키기는커녕
시키기도	시키기라도	시키기로	시키기를	시키기만
시키기에	시키기에는	시키기위해	시키길	시키느냐
시키느냐에	시키느라고	시키는	시키는가	시키는가를
시키는가에	시키는데	시키더니	시키더라도	시키던
시키도록	시키듯이	시키라고	시키라는	시키려
시키려고	시키려고만	시키려는	시키려는데	시키려던
시키려면	시키리라는	시키며	시키면	시키면서
시키면서도	시키므로	시키어	시키어야	시키자는
시키자마자	시키자면	시키지	시키지는	

그림 4 명사 접미 음절열의 예

[11]에서는 본 논문의 명사 접미 음절열과 같은 개념인 ‘기능 어휘집’에 대해 언급함으로써 색인어 추출 시 형태소 분석의 필요성을 주장하고 있다. 그러나 기능 어휘집의 사용에 대한 가능성과 이에 따른 문제점만을 언급하였을 뿐 구체적인 실험을 통해 밝히고 있지 않으므로, 본 논문에서는 이를 명사 추출에 적용하여 이러한 정보가 효율성을 높이는 데에 유용하게 사용될 수 있음을 실험을 통해 입증하고자 한다.

4. 한국어 명사 추출기

본 명사 추출기의 전체적인 구성은 그림5에 있다. 명사 추출 과정은 5단계의 과정을 거친다.

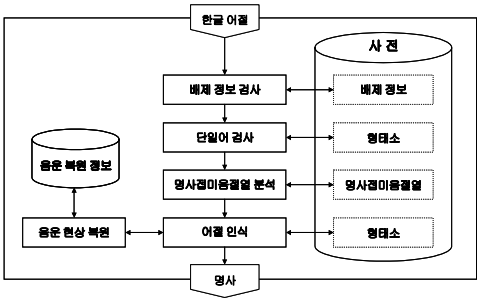

그림 5 명사 추출기 구성도

각각의 입력된 한글 어절에 대해서, 먼저 배제 정보를 이용하여 명사로 분석될 가능성이 없는 어절을 제거하고, 형태소 사전을 참조하여 단일어 검사를 한다. 하나 이상의 형태소가 결합된 어절은 명사 접미 음절열 분석을 통하여 명사를 추출하거나 미등록어를 추정한다. 위의 과정들을 거치는 동안 제거되지 않거나 명사가 추출되지 않은 어절은 어절 인식 과정을 거친다. 이 때 음운 현상으로 인해 원형이 변형되어 분석되지 않은 어절에 대해서는 음운 복원 정보를 이용하여 원형을 복원한 후 다시 어절 인식 과정을 수행한다. 각 과정에 대한 자세한 설명은 다음 절들에서 기술한다.

4.1 단계 1 : 배제 정보 검사

앞에서 설명한 바와 같이 복잡한 분석을 수행하기 전에 명사가 존재하지 않는 어절을 제거하기 위해 배제 정보를 사용한다. 먼저, 어절 내에 배제 정보와 일치하는 부분이 있는지를 검사하고, 일치하는 부분이 있다면 명사가 존재하지 않으므로 해당 어절의 분석을 마친다. 이 때, 각 어절과 배제 정보에 대한 검사는 음소, 부분 어절, 어절 단위 배제 정보의 순으로 이루어진다. 배제 정보 검사 알고리즘은 그림 6에 제시되어 있다.

```
배제정보검사(EJ) {
        EJ = 입력 어절
        phf = EJ의 첫음절의 종성

        /* 음소 단위 배제 정보 검사 */
        if (phf 가 "ㅆ", "ㅎ", "ㅊ", "ㄸ", "ㅃ" 중 하나인가?) return TRUE

        /* 부분 어절 단위 배제 정보 검사 */
        if (EJ가 부분 어절 단위 배제 정보로 시작하는가?) return TRUE

        /* "-다른-", "-렇-" 을 포함하는 어절인지 검사 */
        if (EJ에 "다른"이나 "렇"이 존재하는가?) return TRUE

        /* 어절 단위 배제 정보 검사 */
        if (EJ가 어절 단위 배제 정보인가?) return TRUE

        /* 지금까지 return 되지 않은 경우는 명사분석이 가능한 어절임 */
        return FALSE
}
```

그림 6 배제 정보 검사 알고리즘

4.2 단계 2 : 단일어 검사

단일어는 한국어에서 명사, 관형어, 부사, 감탄사 등과 같이 하나의 형태소가 하나의 어절을 이루는 단어를 말한다. 이러한 경우는 한 번의 사전탐색만으로 분석이 가능하므로 분석이 매우 단순하다. 그러나 여기서 주의할 점은 둘 이상의 품사를 가질 수 있는 단어들이 존재한다는 것이다. 단일어가 명사와 다른 품사를 모두 가질 수 있는 경우는 그림 7과 같은 우선 순위에 따라 분석한다.

1. 명사와 부사인 경우는 부사로 결정한다.
2. 단음절 명사와 다른 품사인 경우는 다른 품사로 결정한다.
3. 2음절 이상 명사와 다른 품사인 경우는 명사로 결정한다.

그림 7 단일어의 중의성 해소 규칙

4.3 단계 3 : 명사 접미 음절열 분석

이 단계는 주어진 어절에서 확실히 명사인 경우를 가려내거나 미등록어 후보를 생성하는 역할을 한다. 명사 접미 음절열을 이용한 명사추출은 형태소 분석에 의한 방법에 비해 매우 단순하다. 예를 들어, "사랑합니다"의 올바른 형태소 분석 결과는 "사랑/NNCV+하/XSVV+ㅂ니다/EFF"로서 어절이 3개의 형태소로 분리가 이루어져야 하나, 명사 접미 음절열을 사용하면 "사랑/NNCV+합니다/PNSS"로 분석이 되므로 명사 "사랑"을 바로 추출할 수 있다.[1] 미등록어를 추정할 때에도 명사 접미 음절열을 사용한다. 어절에서 명사 접미 음절열이 발견되면 그 앞부분을 미등록어 후보로 간주하여 저장한 후, 최종 단계까지 수행했을 때에도 다른 분석 결과가 없을 때에는 저장된 미등록어를 명사로 추출한다.

명사 접미 음절열을 이용하여 명사와 미등록어 후보를 추출하는 방법은 먼저, 입력 어절에서 최장 명사 접미 음절열을 찾고 그 앞부분이 사전에 등록된 명사이거나 복합명사[2]인 경우는 해당 부분을 명사로 추출하고 더 이상의 분석을 하지 않는다. 만일 명사 접미 음절열의 앞부분이 체언이 아닌 경우 명사 접미 음절열의 앞부분과 어절 전체를 일단 미등록어 후보로 간주하고 다음 단계로 넘긴다. 만약 어절에서 명사 접미 음절열이 발견되지 않은 경우에는 어절 전체가 복합명사이거나 등록되지 않은 명사 접미 음절열, 용언 혹은 미등록어일 가능성이 있다. 이 때에는 먼저 복합명사인지 검사한 후에 복합명사가 아닐 경우에는 어절 전체를 미등록어 후보로 간주한다. "연극과는"과 같이 중의적인 분석이 가능한 경우가 존재할 수 있으므로 한 음절 짧은 명사 접미 음절열이 존재하는 경우는 위의 과정을 반복한다.

앞에서 언급한 언어분석 도구를 사용하지 않는 방법과 같이 명사 접미 음절열만을 이용하여 미등록어를 추정하게 되면 여러 가지 오분석이 발생할 수 있다. 가령 문자열 "으로"의 경우는 "으로"와 "로"가 모두 명사 접미 음절열이다. "어플리케이션으로"는 "어플리케이션+으로"와 "어플리케이션으+로"의 두 가지 추정결과를 생성한다. 또한 명사 접미 음절열이 어미로도 쓰이는 경우에는 용언을 미등록어로 추정할 수 있다. 예를 들어, "버리는"은 "버리+는"으로 분석될 수 있다. 이와 같은 문제를 해결하려면 주어진 어절이 체언부인지 용언부인지를 가려야 하고 용언을 분석하려면 일반적인 형태소 분석기의 기능을 모두 가지고 있어야 한다. 본 논문에서는 그와 같은 기능을 단계 4와 5에서 수행하고 있다.

미등록어의 과생성을 방지하기 위한 한가지 방법으로 다음과 같은 휴리스틱 규칙으로 제약을 둔다.

규칙 1 : 미등록어 추정시 명사 접미 음절열의 앞부분이 다음의 경우에는 미등록어로 추정하지 않는다.

- 명사 접미 음절열이 조사이고 그 앞부분과 결합할 수 없을 때
- 종성 "ㅆ"을 포함하는 경우
- 끝음절이 "으", "느", "에", "니" 중 하나일 경우

규칙 2 : 어절 전체를 미등록어로 추정시 다음의 경우에는 미등록어로 추정하지 않는다.

- 2음절 이상의 명사 접미 음절열이 결합된 경우
- 종성 "ㅆ"을 포함하는 경우
- 끝음절이 "은", "는", "을", "를", "에" 중 하나일 경우

여기서, 명사 접미 음절열이 그 앞부분과 결합할 수 없다는 것은 조사와 선행어의 끝음절이 서로 결합할 수 없다는 것을 말한다. 모든 조사는 결합할 수 있는 선행어의 끝음절에 따라 유종성 결합 조사, 무종성 결합 조사, 'ㄹ'종성 결합 조사로 나눌 수 있는데, 순서대로 선행어의 끝음절이 유종성, 무종성, 'ㄹ'종성일 때 결합할 수 있다. 예를 들어 조사"는"은 무종성 결합 조사이므로 "철수+는"은 결합 가능하지만 "상범+는"은 결합할 수 없다.

4.4 단계 4 : 어절 인식

대부분의 어절은 앞의 세 단계를 통해 단순하게 처리할 수 있으나, 이것만으로는 모든 어형에 대한 완벽한 분석은 불가능하다. 이 단계에서는 어절에 대해 보다 자

1) 여기서 사용된 품사 태그 NNCV, XSVV, EFF, PNSS는 각각 보통 명사, 용언화 접미사, 종결어미, 명사 접미 음절열이다.
2) 명사추출기 내부에서는 최장일치에 의한 복합명사 분해를 수행하지만, 본 논문에서는 설명하지 않는다.

세한 분석을 함으로써 이전 단계의 분석 결과를 확정한다. 이 단계로 입력된 어절은 명사가 존재하지 않거나 3단계에서 미등록어로 추정된 어절이다. 분석과정은 어절을 이루는 단위 형태소들을 분리한다는 점에서 형태소 분석과 유사하나 다음과 같은 점에서 일반적인 형태소 분석과 다르다. 첫째, 형태소 분석기는 어절 내의 모든 가능한 형태소-품사 쌍에 대한 결과를 출력해야 하나 어절 인식은 어절 내에 명사가 존재하는지 그렇지 않은지를 결정하기 위해 즉, 올바른 어절인지의 여부만을 판단하는 데 주안점을 두고 있다. 둘째, 일반적으로 형태소 분석기는 세분화된 품사집합을 사용하나[3], 어절 인식을 위해서는 대분류 수준의 품사 11개만을 사용한다.

지금부터 어절 인식에 대해 살펴보자. 실질 형태소와 형식 형태소가 결합된 복합어에 대한 유형을 그림 8과 같이 오토마타(finite state automata)로 구성할 수가 있다. 탐색은 하향식, 깊이 우선, 우좌분석에 기반한다. 하향식은 어절로부터 각 형태소를 인식하는 방법으로서 예측적(predictive)인 처리가 가능하므로 불필요한 분석을 줄일 수 있다. 우좌 분석은 문자열의 우측으로부터 좌측으로 분리하는 것을 말한다. 어절의 끝에서부터 음절 단위로 형태소를 분리하면서 해당되는 다음 상태로 전이하여 더 이상 분리할 형태소가 없을 때 즉 어절의 시작위치에 이르렀을 때 종결 상태(final state)에 있다면 올바른 어절로 간주한다. 올바른 어절이라고 인식되면 미등록어 추정 결과는 무시하고, 그렇지 않으면 추정된 미등록어 후보를 명사로 추출한다.

이 과정에서 중의성이 발생하면 원칙적으로 모든 가능한 명사를 추출하되 다음과 같은 원칙에 따라 우선순위를 둔다.

1. 미등록어 추정의 우선 순위는 낮춘다.
2. 단음절 명사의 우선 순위는 낮춘다.
3. 비자립명사와 자립명사가 모두 가능한 경우는 비자립명사의 우선 순위를 높인다.

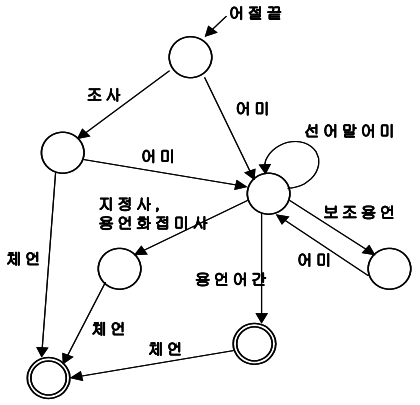

그림 8 어절 생성 전이도

3) 세종계획 말뭉치, ETRI 말뭉치, 고려대 말뭉치의 경우 각각 44개, 27개, 72개로 이루어진 품사집합을 사용한다.

4.5. 단계 5 : 음운 현상 복원

앞 단계인 어절 인식에서는 음운 현상으로 인해 원형이 변형된 경우는 처리하지 않는다. 그러나 한국어에는 용언과 어미의 불규칙 활용, 축약, 탈락과 같은 많은 음운 현상이 있다. 음운 현상을 처리하기 위한 기존의 연구는 각 어절마다 많은 규칙을 적용해 왔다[12]. 본 논문에서는 이러한 음운 규칙을 사용하지 않고 품사부착된 말뭉치로부터 추출한 음운 복원 정보를 이용하여 문자열을 복원한 후 다시 앞 단계인 어절 인식 단계를 수행한다.

어절 인식 단계와 이 단계를 합하여 기존의 형태소 분석 모듈로 대체하여도 되나 분석의 효율을 위해 본 논문에서는 이와 같은 방법을 사용한다.

음운 복원 정보는 품사부착된 말뭉치에서 원시어절과 품사가 부착된 어절에서 품사를 제외한 복원된 어절이 일치하지 않을 경우, 불일치가 발생한 음절로부터 끝음절까지의 한글 부분만을 저장한다. 예를 들어 "사랑했다."는 "사랑/NNG+하/XSV+았/EP+다/EF+./SF"로 품사 부착되는데, 원시 어절 "사랑했다."와 복원된 어절 "사랑하았다."는 서로 같지 않으므로, "했다"와 "하았다"를 저장한다. 여기서 전자는 "복원대상 문자열(source)"이고, 후자는 "복원할 문자열(target)"이다. 본 논문에서는 이러한 정보를 빈도가 2이상인 경우만 저장하여 4692개를 사용한다.

음운 복원 정보를 이용하여 원형을 복원할 때 고려해야 될 사항은 주어진 어절의 부분문자열 중에서 "복원대상 문자열"이 둘 이상 존재할 수 있고, "복원대상 문자열"에 대한 "복원할 문자열"도 둘 이상 존재할 수 있다는 점이다. "복원대상 문자열"은 문자열이 긴 것을 먼저 적용하고, "복원할 문자열"은 빈도가 높은 것을 먼저 적용한다. 표 1은 "달려갔다"에 적용 가능한 음운 복원 정보를 적용할 순서대로 나열한 것이다.

"복원대상 문자열" 중에서 가장 긴 문자열 "려갔다"를 "리어가았다"로 먼저 복원한다. 만약 어절 인식이 실패하면 "복원대상 문자열" "갔다"에 대해서 빈도가 가장 높은 "복원할 문자열"인 "가았다"를 적용한다. 이러한 과정은 어절인식이 하나라도 성공할 때까지 반복된다.

음운 복원 정보를 이용한 음운 현상 복원은 일종의

표 1 “달려갔다”의 음운 복원 정보

빈도	복원대상 문자열	복원할 문자열
46	려갔다	리어가았다
773	갔다	가았다
4	갔다	그았다
3	갔다	가아았다
2	갔다	어가았다
1348	다	이다
3	다	하다
2	다	아다

기분석 방법으로서 음운 규칙을 적용하기 위해 모든 규칙에 대해 매번 조건을 검사하고 적용하는 대신 문자열에 대한 비교와 교체만으로써 간단하고 빠르게 음운 현상을 복원할 수 있다. 또한 자소 단위의 처리를 전혀 하지 않으므로 각 어절에 대한 코드변환 작업이 필요없다.

5. 실험 및 평가

형태소 분석이나 명사 추출은 등록된 사전 표제어에 따라 많은 영향을 받는다. 같은 알고리즘을 사용하더라도 등록된 사전 표제어에 따라 다른 결과를 나타낸다. 일반적으로 사전에 등록된 표제어가 많을수록 미분석이 발생할 가능성은 낮으나 과분석이 발생할 가능성이 높고, 표제어가 적을수록 그 반대의 현상이 나타난다.

타 연구와의 공정한 비교를 하려면 동일한 사전과 실험 환경, 실험 말뭉치에서 비교가 이루어져야하나 현실적으로 이와 같은 동일한 조건에서 실험을 하는 것이 어렵기 때문에 알고리즘에 대한 평가는 자체 실험에 의존할 수밖에 없다. 그러나 전체 시스템에 대한 평가도 나름대로 의미가 있으므로 다른 시스템과의 비교 실험도 수행하였다. 본 논문의 실험은 리눅스, 펜티엄 Ⅲ 700MHz, 2Gbyte RAM에서 수행되었다.

5.1 평가 척도

본 논문에서 사용한 배제 정보와 명사 접미 음절열 음운 복원 정보는 세종계획 150만 품사부착 말뭉치로부터 추출하였고, 본 논문에서는 이를 ‘추출 말뭉치’로 부르기로 한다. 실험을 위해서는 ETRI 28만 품사부착 말뭉치를 사용하였고 실험 말뭉치의 특성은 표 2에 있다.

표 2 실험 말뭉치 특성

분야	문서 수	총 어절 수	문서당 평균 어절 수
소설	26	167,065	6,426
비소설	44	108,570	2,468
뉴스	41	12,656	309
전체	111	288,291	2,597

본 논문에서는 MATEC1999의 평가 방식과 평가 척도를 따른다. 이 평가 방식의 특징은 각 어절 내에서 자립 명사만을 추출하는 것이 목표이며, 복합 명사는 분리된 단위 명사들을 추출하지 않고 단일형만을 찾는 것을 원칙으로 한다. 예를 들어, “사과나무”는 단위 명사인 “사과”와 “나무”로 분리되나 “사과나무”만을 출력해야 한다. 명사 추출의 성능 평가를 위한 척도로는 정확률(P), 재현율(R), F-measure(F)를 사용하고 각각의 값은 다음과 같이 계산된다.

$$P=\frac{\text{정답과 일치하는 응답 명사의 개수}}{\text{응답 명사의 개수}}\times 100(\%)$$

$$R=\frac{\text{정답과 일치하는 응답 명사의 개수}}{\text{정답 명사의 개수}}\times 100(\%)$$

$$F=\frac{2PR}{P+R}$$

이 때, 명사의 개수는 중복을 허용하지 않는 즉, 여러 번 나타난 명사도 한 번 나타난 것으로 가정하여 빈도를 무시한 계산 방식이다.

5.2 자체 실험

각 단계가 성능에 영향을 미치는 정도와 문서의 분야별 성능을 알아보기 위해 그림 9와 같이 시스템을 나누었다. “유형1”은 배제 정보 검사와 명사 접미 음절열 검사를 하지 않은 것으로서 일반적인 형태소 분석의 과정과 유사하고, “유형2”와 “유형3”은 각각 명사 접미 음절열 검사와 배제 정보 검사를 하지 않는 것이다. 마지막으로 “유형4”는 모든 단계를 수행한 것으로서 명사 출현 특성을 고려한 시스템이다.

유형 1: 단계 2 + 단계 4 + 단계 5
유형 2: 단계 1 + 단계 2 + 단계 4 + 단계 5
유형 3: 단계 2 + 단계 3 + 단계 4 + 단계 5
유형 4: 단계 1 + 단계 2 + 단계 3 + 단계 4 + 단계 5

그림 9 각 단계별 시스템 분류

표 3에 의하면, 배제 정보와 명사 접미 음절열을 모두 사용한 “유형4”의 성능이 가장 좋았다. 주목할 만한 것은 배제 정보를 이용하는 “유형2”와 “유형4”의 정확률이 그렇지 않은 “유형1”과 “유형3”보다 높은데 그 이유는 중의성이 있는 어절을 배제 정보가 어느 정도 제거해 주기 때문으로 보인다. 문서의 분야별로는 소설의 성능이 가장 낮게 나타났는데 이는 구어체 문장이 많기 때문으로 생각된다.

앞의 실험은 5.1절에서 언급하였듯이 문서에 나타난 명사의 빈도를 전혀 고려하지 않은 것이다.

표 3 각 단계별/문서 분야별 성능 평가

분야		유형 1	유형 2	유형 3	유형 4
소설	재현율	80.5	80.4	91.9	91.6
	정확률	78.4	83.4	72.2	77.1
	F-measure	79.4	81.9	80.8	83.8
비소설	재현율	80.0	79.9	92.3	92.1
	정확률	82.6	87.7	79.3	84.9
	F-measure	81.3	83.6	85.3	88.4
뉴스	재현율	76.3	76.1	90.6	90.3
	정확률	81.3	87.4	81.2	87.6
	F-measure	78.7	81.3	85.6	88.9
전체	재현율	78.7	78.6	91.6	91.3
	정확률	81.1	86.6	78.3	84.1
	F-measure	79.9	82.4	84.4	87.6

빈도를 고려하지 않는다는 것은 한 번 나타난 것과 여러 번 나타난 것이 동일하게 취급되는 것으로서 정보 검색의 관점에서 볼 때 문서 내에서 한번만 색인어로 추출되어도 검색된다는 점을 감안한 것이다. 문서 내에서의 명사의 빈도를 고려했을 때의 실험 결과는 표4에 있다.

표 4 빈도를 고려한 실험

장르		유형 1	유형 2	유형 3	유형 4
소설	재현율	72.8	72.7	87.9	87.6
	정확률	78.9	87.0	73.6	83.2
	F-measure	75.7	79.2	80.1	85.3
비소설	재현율	79.3	79.2	91.6	91.4
	정확률	83.8	89.9	80.7	88.6
	F-measure	81.5	84.2	85.8	90.0
뉴스	재현율	75.4	75.3	89.8	89.6
	정확률	82.2	87.9	81.8	87.7
	F-measure	78.7	81.1	85.6	88.7
전체	재현율	76.3	76.2	90.1	89.9
	정확률	82.1	88.5	79.5	87.0
	F-measure	79.1	81.9	84.4	88.4

빈도를 고려했을 때와 고려하지 않았을 때의 결과는 거의 유사하다.

표 5는 배제 정보와 명사 접미 음절열이 형태소 분석을 해야 하는 어절을 얼마나 감소시킬 수 있는지와 그에 따른 분석속도에 대한 실험이다. 제안한 방법에서 일반적인 형태소 분석 과정과 유사한 단계는 "단계 4"이다. 전체 어절 중에서 "단계 4"를 수행한 어절의 비율을 계산한 것이 어절 인식 비율이다.

표 5 단계 4의 실행 비율 및 분석 속도

	유형 1	유형 2	유형 3	유형 4
어절 인식 비율 (%)	81.6	61.1	17.8	10.5
분석속도 (어절/초)	86,091	100,401	88,057	102,313

배제 정보와 명사 접미 음절열 검사를 사용하지 않는 경우에는 전체 어절의 약 82%가 어절 인식을 수행하고, 배제 정보와 명사 접미 음절열을 모두 사용하는 경우는 단지 전체의 약 11%만이 어절 인식 단계를 수행하고 있다. 어절 인식을 수행한 어절이 줄어든 비율만큼 복잡한 분석 과정을 생략했다는 의미이다. 분석 속도에 대한 결과에서는 "유형 1"에 대해, "유형 2"는 배제정보를 사용한 것이고, "유형 3"은 명사 접미 음절열을 사용한 것인데 모두 "유형 1"보다 빠른 분석 속도를 보이고 있고, 두 가지를 모두 사용하는 "유형 4"는 가장 빠른 분석 속도를 나타냈다. 결론적으로 어절 인식을 하는 어절의 비율이 적을수록 분석 속도가 높아짐을 보이고 있다. 기존 형태소 사전의 2% 정도의 추가 부담으로 효율성을 상당히 높였다.[4)]

5.3 비교 실험

본 논문에서 제안한 시스템인 NE2001과 고려대학교의 형태소 분석기 KOMA[13], 고려대학교의 품사태거 HanTag[14], 그리고 HAM 5.0a[15]에 대한 실험 결과가 표 6에 있다[5)].

표 6 타 시스템과의 비교 결과

시스템명		NE2001	KOMA	HanTag	HAM 5.0a
빈도 고려 ×	재현율	91.34	93.12	88.68	91.02
	정확률	84.08	60.10	90.54	77.23
	F-measure	87.56	73.06	89.60	83.56
빈도 고려 ○	재현율	89.86	93.67	88.58	90.67
	정확률	87.02	58.07	91.77	76.46
	F-measure	88.42	71.70	90.15	82.96
분석속도(어절/초)		103,940	8,944	5,281	21,279

4) 기존 형태소 사전 표제어는 161,438개이고, 배제 정보와 명사 접미 음절열이 포함된 사전 표제어는 164,794개이다.

5) HAM5.0a에는 여러 가지 선택 사항이 있으며, 이 중에서 1음절 명사를 추출하고 복합명사를 이루는 단위 명사들을 추출하지 않도록 하는 "-1c"를 사용한 결과임

KOMA와 HAM 5.0a는 형태소 분석에 기반한 방법이고[6], HanTag은 형태소 분석[7]과 품사 태깅에 기반한 방법이다. NE2001과 KOMA는 명사 출현 특성을 제외하면 동일한 사전 표제어와 구조를 사용한다.

실험 결과에 따르면, 많은 명사 후보를 생성하기 때문에 KOMA는 재현율은 가장 높으나 정확률은 매우 낮다. 반면에 HanTag은 품사 중의성을 해소하기 때문에 정확률은 가장 높으나 재현율은 비교적 낮다. HAM은 불용어 사전을 이용하여 불필요한 분석 후보를 어느 정도 제거하기 때문에 중간 정도의 결과를 보이고 있다. NE2001은 F-measure에 있어서 HanTag에 이어 두 번째로 좋은 성능을 보이고 있다.

속도 면에서는 HanTag은 형태소 분석과 품사 태깅을 거치기 때문에 가장 느리다. 상대적으로 가장 적은 분석을 수행하는 NE2001의 평균 분석 속도는 103,940 어절/초로서, 다른 시스템들에 비해서 약 5배에서 20배에 이르는 빠른 속도를 나타냈다. 전체 시스템의 비교를 통해 제안한 방법은 기존의 방법과 유사한 성능을 보이면서도 속도 면에서 큰 장점이 있음을 알 수 있다.

2장에서 언급했던 언어분석 도구를 사용하지 않는 방법들[7][8]의 성능을 살펴보면, 이운재(1999)에서는 약 43,000 어절/초의 분석속도[8]와 86%의 재현율과 88%의 정확률을, 장동현(1999)에서는 91%의 재현율과 77%의 정확률을 보고하고 있다.[9] 2장에서 언어분석 도구를 사용하지 않는 방법은 잘못된 분석을 할 경우가 많다고 하였는데, [7]에서는 88%의 비교적 높은 정확률을 보이고 있다. 이에 대해서는 다음과 같은 사항들이 원인이 될 수 있다. 첫째, 실험에 사용된 말뭉치가 다르다는 점이다. [7]과 [8]의 실험은 MATEC1999 대회에서 사용한 약 3만 어절 정도의 실험 말뭉치에 대한 실험 결과이고, 본 논문에서 수행한 실험은 ETRI 말뭉치 전체를 학습 말뭉치로 사용한 것이므로 직접적인 비교는 불가능하다. 또한 MATEC1999 대회의 학습 말뭉치는 실험 말뭉치와 동일한 분야의 문서로 이루어져 있으므로 학습 말뭉치로부터 추출한 명사를 사용했다면 비교적 성능이 높게 나올 가능성이 있다. 둘째, [7]의 분석 단계 중에는 불용 어절을 제거하는 부분이 있으며, 이는 본 논문에서 사용하고 있는 배제 정보 중에서 어절 단위 배제 정보에 해당한다. [7]에서 사용한 불용 어절의 종류와 걸러진 어절의 비율은 알 수 없으나, 표 3에서의 실험을 통해 배제 정보가 중의적인 어절을 미리 제거하는 효과가 있다는 것을 알 수 있듯이, 불용 어절의 제거가 정확률을 높이는 데에 기여했을 것이라고 생각된다.

5.4. 사전에 등록된 명사의 역할

미등록어 문제와 관련하여 사전에 등록된 명사의 역할을 조사하기 위해 사전에 등록된 명사의 크기에 따른 실험을 하였다(표 7 참조). "형태소사전+추출말뭉치"는 기존 형태소 사전에 등록된 명사와 추출 말뭉치로부터 추출한 명사를 모두 사용한 것이고, "형태소 사전"은 형태소 사전에 등록된 명사만을, "추출 말뭉치"는 추출 말뭉치로부터 추출한 명사만을 사용하는 것이고, "형태소 사전+명사추정×"은 형태소 사전에 등록된 명사만을 사용하되 미등록어 추정을 하지 않는 것이다. 마지막으로 "명사제거"는 형태소 사전에서 모든 명사를 제거한 것으로서 문서에 나타나는 모든 명사가 미등록어인 상태이며, 이 때 명사 추출기가 추출하는 모든 명사는 미등록어로 추정된 것이다.

표 7 사전에 등록된 명사에 따른 실험

		재현율	정확률	F-measure	명사의 수
빈도 고려 ×	명사제거	70.3	59.1	64.2	0
	추출말뭉치	90.0	82.9	86.3	50,772
	형태소사전+명사추정×	87.6	87.5	87.5	140,340
	형태소사전	91.3	84.1	87.6	140,340
	형태소사전+추출말뭉치	91.8	83.4	87.4	150,936
빈도 고려 ○	명사제거	66.1	66.4	66.3	0
	추출말뭉치	88.9	85.7	87.3	50,772
	형태소사전+명사추정×	86.1	89.4	87.7	140,340
	형태소사전	89.9	87.0	88.4	140,340
	형태소사전+추출말뭉치	90.2	86.3	88.2	150,936

실험 결과에 따르면, 사전에 등록된 명사가 많아질수록 재현율은 높아지지만 정확률에서는 형태소 사전에 있는 명사만 사용한 경우보다 오히려 다소 낮아졌다. 명사의 수에 비해 추출 말뭉치로부터 추출한 적은 수의 명사로도 상대적으로 높은 정확도를 보이는 것도 주목할 만하다. 이러한 현상은 형태소 사전에 포함된 명사의 수는 많지만 자주 사용되지 않는 명사가 많기 때문으로 생각된다. 명사의 수가 증가하더라도 일정한 정도 이상의 성능 향상에는 한계가 있음을 알 수 있다. 또한 단순히 명사의 수보다는 자주 나오는 명사가 중요함을 알 수 있다. 형태소 사전을 이용하되 미등록어 추정을 하지

6) KOMA는 범용 형태소 분석기이고, HAM 5.0a는 형태소 분석을 이용한 색인어 추출기이다.
7) HanTag은 KOMA의 형태소 분석 결과를 입력으로 받는다.
8) 리눅스, 펜티엄 II, 400MHz에서 실험한 결과임
9) 빈도를 고려하지 않은 실험이다.

않은 경우는 추정을 한 경우보다 재현율은 낮게 정확률은 높게 나타났다. 미등록어 추정 오류의 영향을 받지 않기 때문에 정확률은 높으나 사전에 등록되지 않은 명사는 추출하지 못하므로 재현율은 낮다.

5.5. 오류 분석

명사 추출과 관련된 오류는 크게 추출되지 않아야 할 명사가 추출되는 경우(과분석)와 추출되어야 할 명사가 추출되지 않은 경우(미분석)로 나눌 수 있다. 과분석은 정확률에, 미분석은 재현율에 영향을 미친다. 모든 경우에 있어서 가장 큰 원인은 분석의 중의성이다. 과분석은 둘 이상의 분석 결과 중에서 옳지 않은 경우를 포함하여 추출했을 때, 미분석은 둘 이상의 분석 결과 중에서 옳지 않은 경우만을 선택했을 때 주로 발생한다. 오류의 원인을 좀 더 구체적으로 살펴보면, "가지는"이나 "배우는"과 같이 중의성이 있을 때, 문장에서는 명사로 사용되지 않았지만 명사로 추출될 때는 과분석이 되고, "입을"이나 "말을"과 같이 단음절 명사이면서 용언으로도 분석이 가능한 경우는 용언 분석의 우선 순위가 높기 때문에 미분석이 된다.

결론적으로, 이러한 문제를 확실히 해결하는 방법은 품사 태깅과 같이 중의성을 해소하는 것뿐인데, 품사 태거는 형태소 분석기에 의존적이고 품사 태거 역시 오분석을 할 가능성이 있다. 5.3절의 실험에서 보듯이, 품사 태거를 이용한 방법이 가장 정확한 결과를 보였지만 성능의 향상에 비해 속도 면에서 효율성이 떨어진다.

6. 결론 및 향후연구

지금까지 명사 출현 특성을 고려한 효율적인 명사 추출 방법에 대해서 알아보았다. 본 논문은 2장에서 분류한 세 가지 방법 중에서 언어분석 도구를 사용하지 않는 방법과 형태소 분석기를 이용하는 방법을 함께 사용하는 방법이라고 볼 수 있다. 한 어절에 하나 이상의 형태소가 결합 가능하고, 다양한 음운 현상이 발생하는 한국어에서 명사를 추출하기 위해서는 형태소 분석이 필수적이지만 복잡한 연산 과정과 많은 분석 결과를 생성함으로 인해 수행 시간 면에서 비효율적이다. 이러한 단점을 해소하기 위해 명사 출현 특성을 이용한다. 명사 출현 특성은 명사의 존재에 대한 긍정적 또는 부정적인 단서를 표현하는 한국어의 특성으로서, 배제 정보와 명사 접미 음절열이 있다. 배제 정보는 명사가 없는 어절을 미리 배제함으로써 분석 시간을 줄이고, 중의적인 어절도 제거되는 효과가 있으므로 정확성도 높일 수 있다. 명사 접미 음절열은 명사를 단순한 방법으로 추출할 수 있게 하고, 미등록어를 인식하는 데에 효과적이다. 또한 형태소 분석시 복잡한 음운 현상을 처리하기 위해 많은 음운 규칙을 적용하는 대신 음운 복원 정보를 사용하여 음운 현상을 처리하는 방법을 알아보았다.

대량의 품사부착 말뭉치의 이용이 용이해지면서 말뭉치로부터 유용한 정보를 추출할 수 있다. 본 논문에서 사용한 배제 정보, 명사 접미 음절열, 음운 복원 정보는 품사부착 말뭉치로부터 획득하였다. 실험결과 명사 추출의 정확률과 재현율은 기존 연구와 유사한 정도의 결과를 보이면서도, 명사 출현 특성을 고려함으로써 전체적인 시스템의 속도를 매우 향상시킬 수 있었다. 따라서, 제안한 방법은 대량의 문서를 빠르게 처리해야 하는 정보 검색과 같은 분야에 유용하게 쓰일 수 있다.

현재 본 논문에서 사용한 배제 정보는 수작업으로 추출하였으나 말뭉치로부터 자동으로 획득하는 방법에 대해서도 연구할 계획이다. 명사가 나타나지 않는 어절에 대한 분석을 배제하기 위해서 배제 정보 중에서 고빈도 어절을 사용했으나 이 중에서 항상 명사가 존재하는 어절에 대한 기분석 결과를 저장함으로써 분석 속도를 향상시키는 방안도 고려할 만하다.

참 고 문 헌

[1] 이재성, 박재득, 차건회, 박세영, "형태소분석기 및 품사 태거 평가대회(MATEC99) 개요", 제1회 형태소 분석기 및 품사태거 평가 워크숍 논문집, pp.13-22, 1999.

[2] 김남철, 서영훈, "형태소 분석기 CBKMA와 색인어 추출기 CBKMA/IX", 제1회 형태소 분석기 및 품사태거 평가 워크숍 논문집, pp.50-59, 1999.

[3] 이중영, 신병훈, 이공주, 김지은, 안상규, "COM 기반의 다목적 형태소 분석기를 이용한 명사 추출기", 제1회 형태소 분석기 및 품사태거 평가 워크숍 논문집, pp.167-172, 1999.

[4] 안동언, "좌우접속정보를 이용한 명사추출기", 제1회 형태소 분석기 및 품사태거 평가 워크숍 논문집, pp.173-178, 1999.

[5] 심준혁, 김준석, 이근배, "통계와 규칙을 이용한 강인한 품사태거", 제1회 형태소 분석기 및 품사태거 평가 워크숍 논문집, pp.60-75, 1999.

[6] 권오욱, 정유진, 김미영, 류동원, 이문기, 이종혁, "음절단위 CYK 알고리즘에 기반한 형태소 분석기 및 품사태거", 제1회 형태소 분석기 및 품사태거 평가 워크숍 논문집, pp.76-88, 1999

[7] 이운재, 김선배, 김길연, 최기선, "모듈화된 형태소 분석기의 구현", 제1회 형태소 분석기 및 품사태거 평가 워크숍 논문집, pp.123-136, 1999.

[8] 장동현, 맹성현, "학습데이타를 이용하여 생성한 규칙과 사전을 이용한 명사 추출기", 제1회 형태소 분석기 및 품사태거 평가 워크숍 논문집, pp.151-156, 1999.

[9] 강승식, “음절 특성을 이용한 한국어 불규칙 용언의 형태소 분석”, 한국정보과학회 논문지, 제22권 제10호, pp.1480-1487, 1995.
[10] 임희석, 윤보현, 임해창, “배제 정보를 이용한 효율적인 한국어 형태소 분석기”, 한국정보과학회 논문지, 제22권 제6호, pp.957-964, 1995.
[11] 강승식, 권혁일, 김동렬, “한국어 자동 색인을 위한 형태소 분석의 기능”, 한국정보과학회 춘계 학술발표 논문집, 제22권 제1호, pp.929-932, 1995.
[12] 강승식, “한국어 형태소 분석기에서 불규칙 용언의 분석 모형”, 한국정보과학회 논문지, 제19권 제2호, pp.151-164, 1992.
[13] 이상주, 박봉래, 김진동, 류원호, 이도길, 임해창, “예측 기반 형태소 분석기와 결합 독립 모형 기반 품사 태거 및 고속 명사 추출기”, 제1회 형태소 분석기 및 품사 태거 평가 워크숍 논문집, pp.145-150, 1999.
[14] 김진동, 임희석, 임해창, “Twoply HMM : 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델”, 한국정보과학회 논문지(B), 제24권, 제12호, pp.1502-1512, 1997.
[15] 강승식, 이하규, “한국어 형태소 분석기 HAM의 형태소 분석 및 철자 검사 기능”, 제8회 한글 및 한국어 정보처리 학술발표 논문집, pp.929-932, 1995.

이 도 길
1999년 2월 고려대학교 컴퓨터학과 학사. 2001년 2월 고려대학교 컴퓨터학과 석사. 2001년 3월~현재 고려대학교 컴퓨터학과 박사 과정. 관심분야는 한국어 정보처리, 기계학습, 정보검색, 생물정보학

이 상 주
1992년 2월 고려대학교 컴퓨터학과 학사. 1995년 2월 고려대학교 컴퓨터학과 석사. 1999년 8월 고려대학교 컴퓨터학과 박사 1999년 11월~2001년 3월 일본 동경대학교 정보과학과 연구원(일본학술진흥회 지원). 1997년 3월~2002년 2월 고려대학교 기초과학연구소 연구원. 2002년 3월~현재 (주)엔엘피솔루션 대표이사. 관심분야는 자연어처리, HCI, 기계학습, 정보검색

임 해 창
1991년 ~ 현재 고려대학교 컴퓨터학과 교수. 1993년 인지 과학회 이사. 1994년-998년 한국 정보과학회 편집위원. 1998년 5월 ~ 2000년 5월 한국정보과학회 한국어정보처리연구회 운영위원장 1999년 3월 ~ 2000년 8월 고려대학교 컴퓨터과학기술연구소 연구소장. 관심분야는 자연어처리, 구문분석, 정보검색, 기계학습