분석 배제 정보와 후절어를 이용한 한국어 명사추출

이도길[U] 류원호 임해창
고려대학교 컴퓨터학과
{dglee, whryu, rim}@nlp.korea.ac.kr

Korean Noun Extraction Using Exclusive Segmentation Information and Post-noun morpheme sequences

Do-Gil Lee[U] Won-Ho Ryu Hae-Chang Rim
Dept. of Computer Science & Engineering. Korea University

요 약

명사 추출기는 정보검색, 문서분류, 문서요약, 정보추출 등의 분야에서 사용되고 있으며, 정확한 명사 추출과 빠른 색인 속도는 이들 시스템 성능과 밀접한 관계가 있다. 한국어에서 명사를 추출하기 위해서는 형태소 분석이 필요한데, 본 논문에서는 대량의 품사부착된 말뭉치로부터 추출한 분석배제 정보와 후절어를 이용함으로써 형태소 분석을 생략하거나 보다 단순한 처리에 의해 명사를 추출하는 방법을 제안한다. 또한 형태소 분석시 복잡한 음운 현상을 처리하기 위해 많은 음운 규칙을 적용하는 대신 음운 복원 정보를 사용하여 음운 현상을 처리하는 방법을 제안한다. 실험결과에 의하면, 제안된 방법에 의한 명사추출기는 비교적 높은 정확률과 재현율을 나타내며, 빠른 속도를 보였다.

1. 서론

최근 인터넷의 발전으로 인해 다양한 정보가 문서화되어 공유되고 있으며, 그 양 또한 매우 빠른 속도로 증가하고 있다. 사용자가 원하는 정보를 빠르고 정확하게 찾고자 하는 요구가 높아짐에 따라 자연어를 처리하고자 하는 연구가 활발히 진행되고 있다. 특히, 자연어 처리의 응용 분야 중에서 정보 검색 연구에 대한 요구가 높아지고 있는 실정이다.

이러한 정보는 문서로 저장되어 있는데 문서는 하나 이상의 문장들로 이루어져 있다. 이 문장들의 내용을 대표하는 것이 그 속에 있는 명사라고 간주하고 문장으로부터 추출된 명사를 색인어로 사용하고 있다. 엄밀하게는 의미분석과 같은 상위단계의 자연어처리 과정을 거쳐야 완벽하게 문장을 이해할 수 있지만, 정보검색과 같이 방대한 양의 문서를 빠르게 처리해야 하는 분야에서는 현재까지의 기술수준과 효율성을 감안해 볼 때 적용하기 어렵다.

명사 추출기는 정보검색의 자동색인 시스템과 질의처리기의 필수적인 요소로서 정확한 색인이 추출과 빠른 색인 속도는 정보검색 시스템의 성능과 밀접한 관계가 있다. 또한 명사추출기는 문서분류, 문서요약, 정보추출 등의 분야에서도 사용되고 있다.

기존의 한국어 명사추출 시스템은 크게 세 가지로 나눌 수 있다.

- 형태소 분석기를 이용하는 방법[1][2]
- 형태소 분석기와 품사태거를 이용하는 방법[3]
- 언어분석 도구를 사용하지 않는 방법[4][5]

형태소 분석기를 이용하는 방법은 형태소 분석 결과로 얻어진 모든 명사를 추출하므로 재현율은 높으나, 분석의 중의성으로 인해 정확하지 않은 결과가 포함될 수 있다. 형태소 분석기와 품사태거를 이용하는 방법은 품사태깅을 수행하여 명사로 결정된 단어만을 추출한다. 중의성을 해결하므로 보다 정확한 결과를 얻을 수 있으나, 형태소 분석 및 태깅 단계를 거쳐야하므로 분석시간이

오래 걸린다는 단점이 있다. 언어분석 도구를 사용하지 않는 방법은 형태소 분석을 하지 않고 사전에 저장된 어휘 정보만을 사용하여 명사 여부를 판단한다. 시스템이 단순하고 구현이 쉬우며 분석 속도가 빠른 반면, 형태소 분석을 하는 방법에 비해 정확률이 낮다.

한국어에서 명사를 추출할 때 고려해야 할 사항은 다음과 같다.

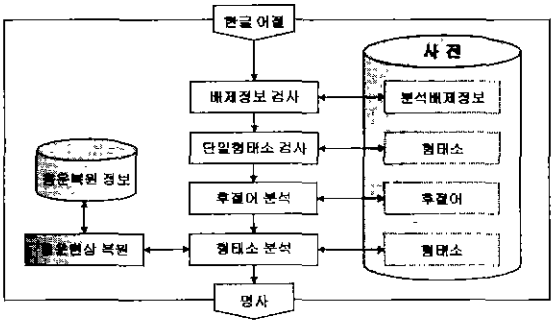
- 빠른 속도
- 한국어의 특성
- 미등록어 인식 문제
- 복합명사 인식 및 분리 문제

본 논문의 궁극적인 목표는 문서에 존재하는 명사를 비교적 정확하면서도 빠르게 추출하는 데에 있다. 대체로 명사 추출을 위해서는 형태소 분석을 하지만, 한국어 형태소 분석도 형태소 분리와 결합제약 검사, 형태소 결합 등 많은 연산을 필요로 한다. 본 논문에서는 분석배제 정보와 후절어를 이용함으로써 명사추출을 위해 최대한 형태소 분석을 거쳐야하는 어절을 줄여서 속도를 향상시키는 방법을 제안한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 제안하는 한국어 명사추출 방법에 대해서 설명하고, 3장에서 실험 및 평가를 하고, 마지막으로 4장에서 결론 및 향후연구를 기술하여 끝을 맺는다.

2. 한국어 명사추출기

본 논문의 명사추출기의 구성은 다음 [그림 1]과 같다.



[그림 1] 명사추출기 구성도

명사추출 과정은 크게 3단계의 과정을 거친다. 각각의 입력된 한글 어절에 대해서, 분석배제 정보를 이용하여 명사로 분석될 가능성이 없는 어절을 제거하고, 단일 형태소 및 후절어 분석을 통하여 명사를 추출하거나 미등록어를 추정한다. 마지막으로 위의 두 과정을 거치는 동안 제거되지 않거나 명사가 추출되지 않은 어절은 형태소 분석 과정을 거친다. 이 때 음운현상으로 인해 원형이 변형되어 분석되지 않은 어절에 대해서는 음운정보를 이용하여 원형을 복원한 후 다시 형태소 분석을 한다. 각 과정에 대한 자세한 설명은 다음 절들에서 기술한다.

2.1. 분석배제 정보를 이용한 형태소 분석 배제

품사부착된 말뭉치로부터 얻어낸 여러 가지 정보들을 이용하여 명사가 없는 어절에 대해서는 분석과정을 생략함으로써 형태소 분석에 필요한 탐색공간을 줄이고자 하였다. 한국어 어절 중에는 첫음절의 종성이나 처음 2음절 또는 3음절을 살펴보았을 때 또는 어절 가운데에 특정한 문자열이 나타나는 경우, 전체 어절 내에 명사가 거의 나타나지 않는 경우가 있다. 또한 빈도가 높은 어절 중에서 명사가 존재하지 않고 중의성이 거의 없는 어절이 많다.

다음은 본 논문에서 사용한 분석배제 정보를 종류별로 분류한 것이다.

- 음소 단위 배제 정보
  첫음절에 종성 "ㅆ", "ㄶ", "ㅎ", "ㄾ", "ㅥ"이 존재하는 어절

- 음절 단위 배제 정보
  "같", "보였" 등으로 시작되는 어절
  "다른", "륽"이 어절 내에 존재하는 어절

- 어절 단위 배제 정보
  학습 말뭉치에 나타난 어절 중에서 명사가 추출되지 않는 고빈도 어절

이러한 분석배제 정보를 학습 말뭉치 내에서 추출하여 사전에 추가하고, 형태소 분석 전에 사전을 탐색하여 주어진 어절과 일치하는 것이 명사가 나타나지 않는다고 간주하여 더 이상의 분석을 하지 않는다. 그러나 이러한 정보를 찾기가 쉽지 않으므로 본 논문에서는 학습 말뭉치에 나타나는 고빈도 어절에 대해서만 추출하였다. 분

석베제 정보를 이용하면 많은 이절에 대한 형태소 분석을 배제할 수 있다. 본 논문에서 사용한 배제 정보의 예는 [그림 2], [그림 3]에 있다.

가볍	김었	같	거쳐	걸려	걸어
걸쳐	것	그걸	그것	그녀	그들
그런	나는	다시	대해	되는	되이
되었	들	들어	따라	때문	물었
보았	보였	싶	아니	어떤	어떻
없	위해	의해	이러	알어났	알이
주었	하고	하는	하면	하였	하지
한다	합니	해서			

[그림 2] 음절 단위 배제 정보의 예

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

[그림 3] 어절 단위 배제 정보의 예

2.2. 후절어를 이용한 명사추출

후절어란 체언 뒤에 붙을 수 있는 형식형태소의 열(sequence)로서 정의한다. 후절어는 일종의 기능어휘집으로서, 어절에서 후절어가 발견되면 바로 후절어의 앞에 위치한 체언을 검사함으로써 모든 형태소 분석 과정을 거칠 필요가 없이 명사를 추출할 수 있으므로 분석속도를 높일 수 있다[6]. 후절어를 사용하여 명사를 추출하는 기존의 방법으로는 [4]가 있다. 그러나 여기서는 미등록어 추정에 대한 언급이 없다.

후절어를 이용하여 명사를 추출하는 과정을 예를 들어 설명하면, "사랑합니다"는 형태소 분석시 "사랑/명사+하/용언화접미사+ㅂ니다/어미"로 분석이 되나. 후절어를 사용하면 "사랑/명사+합니다/후절어"로 분석이 되므로 명사 "사랑"을 바로 추출할 수 있다. 따라서, 후절어를 사용하여 체인으로 분석된 어절에 대해서는 더 이상의 복잡한 형태소 분석과정을 수행하지 않는다. 또한 미등록어를 추정할 때에도 후절어를 사용한다. 후절어가 발견되면 일단 후절어의 앞부분을 체언이라고 간주하여 저장한 후, 형태소 분석을 수행한 결과가 전혀 없을 때에는 저장된 미등록어를 명사로 출력한다.

후절어의 유형은 조사 및 "하다", "되다", "시키다" 등과 같은 용언화접미사의 활용형과 "밥먹다"와 같이 흔히 명사와 결합되어 복합용언으로 사용되는 용언의 활용형이 있다. 마지막 유형의 경우는 엄밀하게는 띄어쓰기 오류라고 볼 수 있으나 실세계의 언어현상에서는 자주 발생하므로 이러한 유형도 후절어로 추출하였다. 그러나 말뭉치로부터 추출되는 모든 후절어를 사용하면 사전의 크기가 커지고, 과분석이 발생하는 문제와 말뭉치 자체의 오류로 인해 잘못된 후절어가 추출되는 문제가 있으므로 여기서는 빈도가 2이상이며, 서로 다른 둘 이상의 체언과 결합하는 후절어만 추출하였다. 본 논문에서는 품사부착된 말뭉치로부터 2574개의 후절어를 추출하여 사전에 저장하여 사용한다.

[그림 4]는 용언화 접미사 "시키"의 활용형인 후절어의 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	시키려면
시키려면	시키리라는	시키며	시키면	시키면서
시키면서도	시키므로	시키어	시키어야	시키자는
시키자마자	시키자면	시키지	시키지는	

[그림 4] 후절어의 예

순수하게 후절어만을 이용해서 미등록어를 추정하는 경우 여러 가지 문제점이 있다. 문자열 "으로"의 경우 "으로"와 "로"가 후절어이다. 예를 들어, "어플리케이션으로"는 "어플리케이션"+"으로"와 "어플리케이션으"+"로"의 두 가지로 분석된다. 또한 후절어가 어미로도 사용되는 경우에는 용언을 잘못 추정하여 미등록어로 인식할 수 있다. 예를 들어, "버리는"은 "버리"+"는"으로 분석되어 용언 "버리"를 미등록어로 잘못 추정한다. 이와 같은 문제를 해결하려면 주어진 어절이 체언부인지 용언부인지를 가려야 한다. 용언을 분석하려면 일반적인 형태소 분석기의 기능을 모두 가지고 있어야 한다.

후절어가 발견된 앞부분이 체언이 아닌 경우 일단 미등록어로 간주한다. 후절어를 이용한 명사추출에서 체언 또는 용언으로 분석되지 않으면 형태소 분석 단계로 넘어가는데, 이 때에도 체언 또는 용언으로 분석되지 않을 경우 미등록어로 추정하게 된다.

후절어의 앞부분이 다음의 경우에는 미등록어로 추정하지 않는다.

- 후절어가 조사이고 후절어의 앞부분과 결합할 수 없을 때

- 종성 '씨'을 포함하는 경우
- 끝음절이 "으", "느", "에", "니" 중의 하나일 경우

전체 어절이 다음의 경우에는 미등록어로 추정하지 않는다.

- 2음절 이상의 후절어가 결합된 경우
- 종성 '씨'을 포함하는 경우
- 끝음절이 "은", "는", "을", "를", "에" 중의 하나일 경우

2.3. 형태소 분석과 음운 현상 복원

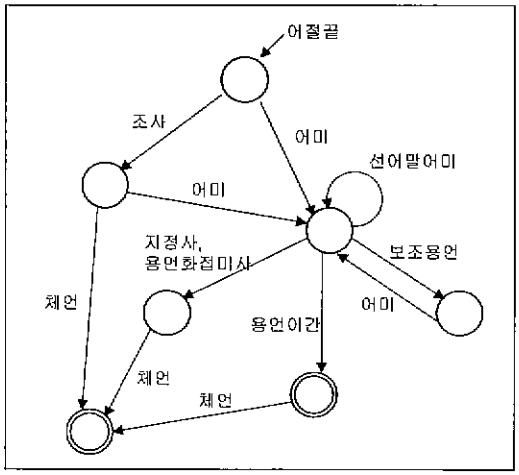
분석배제 정보에 의해 걸러지거나 후절어 분석을 통하여 사용하여 명사로 분석되지 않은 어절은 최종적으로 형태소 분석과정을 거친다.

본 논문에서 사용하고 있는 품사체계는 고려대 사전[7]을 따르는데 이는 범용 형태소 분석을 위한 것이다. 일반적으로 범용 형태소 분석기는 상위 단계의 자연어처리의 전처리 과정으로서 주로 사용되기 때문에 채택하고 있는 품사체계는 상당히 세분화되어 있다. [7]에서 사용하는 전자사전의 품사집합의 수도 총 52개로 많은 편이다. 이렇게 세분화된 품사 집합을 단순화하여 내부적으로 12개의 품사만 사용한다. 중의성으로 인해 형태소간 결합제약을 검사해야 하는 경우와 같이 반드시 필요한 경우에만 세분화된 품사 집합을 사용한다.

한국어 어절의 유형을 단일어, 실질 형태소와 형식 형태소의 결합형, 그리고 복합명사로 나눈다. 단일어는 명사, 관형어, 부사, 감탄사 등과 같이 한 번의 사전탐색만으로 분석이 가능하므로 여기서는 실질형태소와 형식형태소의 결합형, 복합명사에 대해서만 살펴본다.

본 논문의 형태소 분석은 예측기반 우좌분석을 기본으로 하기 때문에 실질형태소와 형식형태소의 결합형에 대한 유형을 어절의 끝에서부터 살펴보면 [그림 5]와 같은 오토마타로 구성할 수가 있다. 어절의 끝에서부터 형태소를 분리하면서 해당되는 다음 상태로 전이하여 더 이상 분리할 형태소가 없을 때 즉, 어절의 시작위치에 이르렀을 때 종결 상태에 있다면 올바른 어절로 간주한다.

[그림 5]에서 조사(또는 어미)의 경우는 최장 조사나 그보다 한 음절 짧은 조사에 대해서만 상태전이가 일어난다. 체언은 단일명사 뿐만 아니라 복합명사도 포함한다.



[그림 5] 어절 생성 전이도

한국어 복합명사를 정규식으로 표현하면 다음과 같다.

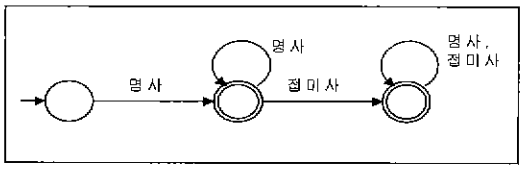
$$\text{복합명사} = ([\text{접두사}]\text{명사}[\text{접미사}]^*)^+$$

본 논문에서는 접두사가 붙는 형태는 사전에 등록되어 있는 걸로 간주한다. 따라서 복합명사에 대한 정규식을

$$\text{복합명사} = (\text{명사}[\text{접미사}]^*)^+$$

와 같이 수정할 수 있다.

이 정규식을 오토마타로 표현한 것은 다음 [그림 6]과 같다.



[그림 6] 복합명사 생성 전이도

본 논문에서 복합명사 분석은 좌우 최장일치 방법을 사용하고 있다.

사전에는 등록되어 있지만 실제로는 거의 쓰이지 않는 단어가 명사로 등록되어 있어서 복합 명사로 잘못 인식되어 명사로 추출되는 경우가 있다. 예를 들어 "하고"의 경우, 사전에 명사로 등록되어 있기 때문에 "공부하고"라는 어절을 분석하면 "공부/명사+하고/조사", "공부/명사+하/용언화접미사/+고/어미", "공부/명사+하고/명사"와 같이 분석되어 단일명사 "공부"와 복합명사 "공부하

고"를 추출한다. 이러한 경우를 방지하기 위해서 미리 리스트를 만들어 불필요한 분석을 줄이고자 하였다. 본 논문에서는 이러한 정보를 명사합성 배제 정보라 부르고, 수동으로 선택된 92개의 명사를 사용하며, 예는 [그림 7]과 같다.

기계	기서	가하	개로	계기	관하
구기	구나	군요	기가	기와	기의
나서	나이	내고	다면	다운	당히
대하	도가	도록	도와	등에	러가
리라	만이	인과	메시	명하	..

[그림 7] 명사합성 배제 정보의 예

한국어에는 용언과 어미의 활용(불규칙), 준말(축약, 탈락)과 같은 많은 음운현상이 있다. 이러한 음운현상들은 자소단위로 나타난다. 따라서 한국어를 음절단위로 분석하는 데에는 한계가 있다. 음운현상을 처리하기 위해서는 각 어절마다 많은 규칙을 적용해야 한다. 본 논문에서는 이러한 음운현상을 처리하기 위해 규칙을 적용하지 않고 후절어 분석과 형태소 분석시 실패한 어절에 대해 품사부착된 말뭉치로부터 추출한 음운복원정보를 이용하여 문자열을 복원한 후 다시 형태소분석을 한다.

음운복원정보는 품사부착된 말뭉치에서 원시어절과 품사가 부착된 어절에서 품사를 제외한 복원된 어절이 일치하지 않을 경우, 불일치가 발생한 음절부터 끝음절까지의 한글 부분만을 저장한다. 예를 들어 "사랑했다."는 "사랑/NNG+하/XSV+았/EP+다/EF+./SF"로 품사부착되는데, 원시어절 "사랑했다."와 복원된 어절 "사랑하았다."는 서로 같지 않으므로, "했다"와 "하았다"를 저장한다. 여기서는 후절어와 마찬가지로 빈도가 2이상인 경우에만 저장하여 4692개의 정보를 사용한다. 복원할 문자열이 여러 개가 있는 경우, 예를 들어, "갔다"의 경우 빈도 1인 경우를 제외하면 [표 1]과 같이 4개로 변형이 가능한데, 음운복원정보를 이용하여 문자열을 교체할 때는 빈도가 높은 것부터 적용한다.

[표 1] "갔다"의 음운복원정보

773 갔다	가았다
4 갔다	그았다
3 갔다	가아았다
2 갔다	어기았다

다음은 음운복원정보를 이용하여 원시어절을 복원하는 예를 보여준다.

[표 2] 음운복원정보를 이용한 예

자소단위 분리 방지	공부인 → 공부이ㄴ
축약된 형태 복원	공부해 → 공부하아
생략된 형태 복원	공부다 → 공부이다
불규칙 현상 복원	구워 → 굽어

3. 실험 및 평가

일반적으로 형태소 분석의 성능은 등록된 품사 표제어에 따라 많은 영향을 받는다. 같은 알고리즘을 사용하더라도 등록된 품사 표제어에 따라 성능이 좌우된다. 일반적으로 사전에 등록된 표제어가 많을수록 미분석이 발생할 가능성은 낮으나 과분석이 발생할 가능성이 높고, 표제어가 적을수록 과분석이 발생할 가능성은 낮으나 미분석이 발생할 가능성이 높다. 타 연구 결과와의 정확한 비교를 하려면 동일한 사전과 실험 말뭉치에서 비교가 이루어져야하나 동일한 사전에서 실험을 하는 것이 실제로는 어렵기 때문에 서로 다른 사전에서 정확도를 측정할 수밖에 없다.

사전의 구조는 유한 상태 변환기(Finite State Transducer)에 기반하여 표제어를 색인한 사전[8]을 이용한다. FST를 이용한 사전은 대량의 사전 표제어에 대해서도 사전의 크기가 많이 커지지 않고, 탐색 속도는 표제어 수에 영향을 받지 않는 장점이 있다.

여기서는 본 논문에서 구현한 명사 추출기의 정확률과 재현율, 시스템의 속도, 분석배제 정보와 후절어를 이용했을 때 형태소 분석 생략 가능한 어절의 비율을 측정하였다.

3.1. 학습 및 실험 말뭉치

분석배제 정보와 음운 복원 정보의 추출에 사용된 말뭉치는 세종계획 150만 품사부착 말뭉치이고, 실험을 위해서는 ETRI 28만 품사부착 말뭉치[9]를 사용하였다.

실험 말뭉치의 특성은 아래 [표 3]과 같다.

[표 3] 실험 말뭉치 특성

	문서 수	총 어절 수	문서당 평균 어절 수
소설	26	167,061	6425
비소설	44	108,559	2467
뉴스	41	12,649	308
전체	111	288,269	2597

3.2. 분석배제 정보와 후절어를 이용한 형태소 분석 배제

실험 말뭉치의 총 어절 수는 288,269 어절이고, 분석배제 정보로 걸러낸 어절 수는 79,209 어절로서 전체의 약 27%를 차지한다. 또한 후절어를 통해 걸러낸 어절 수는 전체의 약 40%를 차지하는 114,678 어절이다. 결론적으로 약 67%의 어절이 형태소 분석을 거치지 않고 명사 추출이 가능함을 알 수 있다.

3.3. 정확률, 재현율 평가

명사 추출의 정확률과 재현율 평가는 [10]를 따른다. 각각의 값은 다음과 같이 계산된다.

$$\text{정확률} = \frac{\text{정답과 일치하는 응답 명사의 개수}}{\text{응답 명사의 개수}}$$

$$\text{재현율} = \frac{\text{정답과 일치하는 응답 명사의 개수}}{\text{정답 명사의 개수}}$$

[표 4]에 말뭉치의 분야별로 각각의 문서에 대한 정확률과 재현율의 평균치를 구하였다.

[표 4] 정확률, 재현율 평가 결과

	정확률	재현율
소설	0.757526	0.919255
비소설	0.840953	0.923654
뉴스	0.874288	0.908311
전체	0.833725	0.916943

위의 실험은 명사의 빈도를 전혀 고려하지 않은 것이다. [표 5]에서는 빈도를 고려하여 정확률과 재현율을 측정하였다.

[표 5] 빈도를 고려한 정확률, 재현율 평가 결과

	정확률	재현율
소설	0.824445	0.885056
비소설	0.880991	0.918459
뉴스	0.843537	0.904053
전체	0.877333	0.901607

실험에서 보듯이 정확률과 재현율을 비교했을 때, 정확률이 재현율보다 낮게 나타났다. 이는 기호가 포함되어 있지 않은 순수한 한글 어절인 경우 정답은 한 어절당 하나의 명사를 가지고 있는 반면, 응답은 중의성이 있는 경우 하나 이상의 명사를 출력하므로 과분석이 발생하기 때문이다. 또 하나의 이유는 "감기는"과 같이 어절 자체의 중의성으로 인해 용언과 체언으로 모두 분석되는 경우 정답은 문맥에 따라 결과를 출력하지만, 응답은 중의적인 분석 결과와는 상관없이 체언으로 분석된 경우 명사를 출력하기 때문이다.

말뭉치 분야별 성능은 비소설, 뉴스, 소설 순으로 높게 나타났으며, 빈도를 고려한 경우가 고려하지 않은 경우보다 정확률은 높고, 재현율은 낮게 나타났다.

3.4. 속도 측정

평균 분석 속도는 43,131어절/초(리눅스, 펜티엄 III 450Mhz, RAM 256MB)로서 빠른 수행속도를 나타내고 있다. 분석배제 정보와 후절어를 사용하지 않았을 때의 평균 분석 속도는 34,308어절/초이다. 분석배제 정보와 후절어를 이용하는 것이 속도 향상에 도움이 됨을 알 수 있다.

타 시스템과의 비교를 수행한 결과, 본 시스템은 10,093어절/초(Solaris5.6, Sun UltraSPARC 143Mhz, RAM 256MB)인데 비해, 동일한 환경에서 [2]은 1,772어절/초, [11]은 1,756어절/초를 보임으로써 모든 어절의 형태소 분석을 수행하는 이들 방법에 비해 본 시스템이 속도면에서 우위를 나타낸다.

4. 결론 및 향후 연구

본 논문은 한국어 명사를 추출하는 방법에 관해서 논하였다. 한 어절에 하나 이상의 형태소가 결합 가능한 한국어에서 정확하게 명사를 추출하기 위해서는 형태소 분석이 필수적이나, 한국어의 특성을 말뭉치로부터 추출하여 분석배제 정보와 후절어를 이용하여 형태소 분석을 생략하거나 보다 단순한 방법으로 명사를 추출하는 방법을 제안하였다.

또한 형태소 분석시 복잡한 음운 현상을 처리하기 위해 많은 음운 규칙을 적용하는 대신 음운 복원 정보를 사용하여 음운 현상을 처리하는 방법을 알아보았다.

실험결과 분석배제 정보와 후절어를 이용함으로써 약 67%의 어절이 형태소 분석을 거치지 않고 명사 추출이 가능했고, 이로 인해 속도도 향상됨을 알 수 있었다.

앞으로 보다 견고한 복합명사 분석과 미등록어 인식을 위한 연구가 이루어져야 하겠고, 품사 태깅처럼 복잡도가 높지 않으면서도 부분적인 어절 문맥을 고려하여 중의성을 해소하는 방안에 대한 연구를 수행할 예정이다.

5. 참고 문헌

[1] 최재혁, "형태소 분석을 통한 한·영 자동 색인어 추출 시스템", 정보과학회논문지 제23권 제12호, pp.1279-1288, 1996.

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

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

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

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

[6] 강승식, 권혁일, 김동렬, "한국어 자동 색인을 위한 형태소 분석의 기능", 한국정보과학회 춘계 학술발표 논문집, 제22권 제1호, pp.929-932, 1995.

[7] 이상주, "형태소분석을 위한 한국어 전자사전: 현황 및 고려사항", 고려대학교, 자연어처리연구실, 기술문서 KU-NLP-TR-1998-01, 1998.

[8] 백대호, 이호, 임해창, "Finite State Transducer를 이용한 한국어 전자 사전의 구조", 제7회 한글 및 한국어정보처리 학술발표 논문집, pp.181-187, 1995.

[9] 한국전자통신연구원, "품사 부착 말뭉치 구축 지침서", 1999. http://aladin.etri.re.kr/~nlu/STANDARD/.

[10] 김진동, 임해창, 박재득, 이재성, "한국어 형태소 분석 시스템에 대한 평가 방법 및 적용 사례 분석", 제1회 형태소 분석기 및 품사태거 평가 워크숍 논문집, pp.44-49, 1999.

[11] 강승식, 이하규, "한국어 형태소 분석기 HAM의 형태소 분석 및 철자 검사 기능", 제8회 한글 및 한국어 정보처리 학술발표 논문집, pp.246-252, 1996.