# 한국어 명사 출현 특성과 후절어를 이용한 명사 추출기

## (Korean Noun Extractor using Occurrence Patterns of Nouns and Post-noun Morpheme Sequences)

박 용 현 † 황 재 원 † 고 영 중 ††
(Yonghyun Park) (Jaewon Hwang) (Youngjoong Ko)

**요 약** 최근 모바일 기기의 발전으로 인하여, PC뿐만 아니라 모바일 기기에서의 정보검색의 요구가 증가하고 있다. 모바일 기기에서 명사를 추출하기 위하여 기존의 언어분석도구를 사용하게 되면, 상대적으로 적은 메모리를 가지고 있는 모바일 기기에는 부담이 되게 된다. 따라서, 모바일 기기에 적합한 언어분석도구의 필요성이 증가하고 있다. 본 논문에서는 대량의 말뭉치로부터 추출한 명사 출현 특성과 후절어를 이용하여 명사를 추출하는 방법을 제안한다. 제안된 명사 추출기는 형태소 분석기를 사용한 기존 명사 추출기의 명사 사전의 약 4% 용량인 146KB의 명사 사전만을 사용함에도 불구하고, 최종적으로 $F_1$-measure 0.86라는 좋은 성능을 얻었다. 또한, 명사 사전에 대한 의존도가 낮으므로, 미등록 명사 추출에 대한 성능이 높을 것으로 예상된다.

**키워드** : 모바일 기기, 명사 추출, 명사 출현 특성, 미등록어 추출

***Abstract*** Since the performance of mobile devices is recently improved, the requirement of information retrieval is increased in the mobile devices as well as PCs. If a mobile device with small memory uses a tradition language analysis tool to extract nouns from korean texts, it will impose a burden of analysing language. As a result, the need for the language analysis tools adequate to the mobile devices is increasing. Therefore, this paper proposes a new method for noun extraction using post-noun morpheme sequences and noun patterns from a large corpus. The proposed noun extractor has only the dictionary capacity of 146KB and its performance shows 0.86 $F_1$-measure; the capacity of noun dictionary corresponds to only the 4% capacity of the existing noun extractor with a POS tagger. In addition, it easily extract nouns for unknown word because its dependence for noun dictionaries is low.

**Key words** : Mobile device, Noun extraction, Noun pattern, Unknown word extraction

## 1. 서 론

현대 사회에서의 모바일 기기는 예전과 달리 단순한 통화와 메시지 송수신 기능뿐만 아니라 다양한 어플리케이션을 통하여 사용자에게 많은 기능을 제공하는 중요한 기기가 되었다. 이로 인해, 모바일을 이용한 서비스가 증가하고 있으며, 그 중 하나가 모바일 정보검색이다. 정보검색의 전처리 과정은 많은 계산량을 요구하기 때문에 웹을 이용한 정보검색의 경우에는 컴퓨팅파워가 높은 서버에서 처리하게 된다. 그러나, 모바일 기기에 존재하는 데이터에 대하여 전처리 작업을 실시하는 경우, 웹 정보검색처럼 서버에서 처리하여 결과를 얻어오는 것은 매우 비효율적이다. 따라서, 모바일 기기에서 활용 가능한 언어분석도구의 개발이 필요하다.

· 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2010-0016994)

† 학생회원 : 동아대학교 컴퓨터공학과
ra2kstar@gmail.com
stfcap@gmail.com
†† 종신회원 : 동아대학교 컴퓨터공학과 교수
yjko@dau.ac.kr

논문접수 : 2010년 3월 5일
심사완료 : 2010년 10월 7일



정보과학회논문지: 소프트웨어 및 응용 제37권 제12호(2010.12)

본 논문에서는 한국어 명사 출현 특성과 후절어를 이용하여 모바일 기기에 적합한 명사 추출기를 제안한다. 제안한 명사 추출기의 특징은 다음과 같다. 품사별로 분류하여 사전을 구성하고, 각 사전마다 다른 규칙을 적용하여 명사가 존재하지 않는 어절을 언어분석도구를 사용하지 않고도 정확하게 제거 할 수 있었다. 또한, 후절어를 적용함에 있어 중의성을 가지는 단어만을 이용하여 명사사전을 구성함으로써 명사사전의 수를 획기적으로 줄일 수 있었다. 개발된 명사추출기는 형태소분석 과정을 거치지 않고, 명사사전의 수가 적으므로 낮은 컴퓨팅파워와 적은 메모리를 사용하는 모바일 기기에 적합하다는 장점을 가진다.

제안한 시스템은 1999년에 전자통신연구원(ETRI)에서 개최한 “제 1회 형태소분석기 및 품사태거 평가대회(MATEC′99)”의 기준을 따랐으며, 대회에 참여한 시스템들과 비교 실험을 진행하였다. 실험결과, MATEC′99에 제안된 시스템들의 평균 $F_1$-measure보다 높은 성능을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대하여 논하고 3장에서는 명사 출현 특성을 이용하여 구성한 사전과 그 적용 규칙에 대해서 설명한다. 4장에서는 본 논문에서 제안한 명사 추출기의 구성에 대해서 기술하며, 5장에서는 타 연구와의 비교 실험 및 평가에 대해서 논한다. 마지막으로, 6장에서는 결론 및 향후 연구에 대해서 기술한다.

## 2. 관련 연구

색인어 중에서 가장 중요한 품사로 인식되고 있는 명사를 효율적으로 추출하기 위하여 많은 연구가 이루어져 왔으며, 1999년에는 ETRI에서 MATEC′99를 개최하여 명사 추출에 대한 활발한 연구를 도왔다. 이 대회에 제출된 명사 추출기를 ‘이도길 외(2003)’[4]에서는 3가지 유형으로 분류하였는데, 그 유형은 다음과 같다.

*<유형 1> 형태소분석기를 이용하는 방법*
*<유형 2> 형태소분석기 및 품사태거를 이용하는 방법*
*<유형 3> 언어분석도구를 사용하지 않는 방법*

[1,2,3]의 경우 <유형 1>의 방식으로써 언어분석도구인 형태소분석기를 사용하여 명사를 추출하는 방식이다. 하지만, 분석의 중의성 문제가 시스템 성능에 큰 영향을 미쳐 언어분석도구를 사용한 것에 비해 높은 성능을 보여주고 있지는 못한다.

[4,5]의 경우 <유형 2>의 방식으로써 형태소분석기로 인해 일어나는 분석의 중의성 문제를 품사태거를 이용하여 해결하였다. 그 결과, 3가지 유형 중 가장 높은 성능을 보여주고 있다.

[6,7]의 경우는 <유형 3>의 방식으로써 언어분석도구를 사용하지 않고 사전에 저장된 어휘 정보를 이용하여 명사를 추출한다.

<유형 1,2>의 경우, 모두 언어분석도구를 사용하는 방법으로써 비교적 높은 성능을 나타내고 있지만 분석 속도가 느리고 프로그램과 사전의 용량이 크다는 단점이 있다. 이 문제를 해결하기 위해 [4]에서는 분석 배제 정보와 후절어를 이용하여 명사 미 출현 어절에 대한 분석을 제외시키고, 음운 현상 복원 단계를 통하여 분석 속도를 향상시키려는 연구가 있었다. 그 결과 <유형 2> 방식임에도 불구하고, 매우 빠르고 높은 성능을 보였다. 또한, <유형 3>의 경우는 상대적으로 분석 속도가 빠르며 구현이 쉬운 반면, 형태소 분석 단계를 거치치 않으므로 오분석이 많으며, 명사 사전에 대한 의존도가 높아 미등록어에 대한 처리가 힘들다. 따라서, 정확한 분석 결과를 얻기 위해서는 <유형 1,2>의 방식을, 빠른 분석과 가벼운 시스템을 위해서는 <유형 3>의 방식을 사용하는 것이 더 좋다. 이것은 결국, 명사 추출기를 사용할 목적에 따라 결정되어 지는 것이므로, 모바일 기기의 경우에는 <유형 3>의 방법이 더욱 적당하다 할 수 있다. 따라서, 본 논문에서는 모바일 기기에 적합한 시스템을 구축하기 위하여 언어분석도구를 사용하지 않는 <유형 3>의 방법을 사용하였다.

## 3. 사전 구성

정확한 명사를 추출하기 위한 많은 연구에서는 언어분석도구를 사용하거나, 명사 사전에 의존해 왔다. 명사 사전에 대한 의존도가 높으면 분석에 대한 정확성은 높아질 수 있으나, 미등록어에 대한 추정이 어려워지며, 시스템의 용량이 커지게 된다. 따라서, 언어분석도구와 명사 사전에 의존하지 않고 명사를 추출하려는 연구가 많이 이루어져 왔으며 그 대표적인 방법이 명사 출현 특성과 후절어를 이용하는 것이다.

한국어 명사 출현 특성이란 어절 내에서 특정 음소나 음절이 일정 조건을 만족할 경우 그 어절 내에서 명사의 존재 여부를 판단할 수 있는 특성을 말한다. 그 특성을 이용하여 명사가 출현하지 않는 어절에 대한 분석을 제외시켜 분석 속도를 향상시키고 오분석을 줄일 수 있다. 또한, 후절어를 사용함으로써 형태소 분석 과정을 거치지 않고도 효율적으로 명사를 추출할 수 있다.

본 논문에서 사용하는 사전은 대량의 말뭉치로부터 명사가 출현하지 않는 어절의 특성을 가진 패턴을 추출하여 분석대상에서 제외시키는 제거사전, 명사 이후에 출현하는 특정 음절열을 추출한 후절어사전, 그리고 추출된 후보에 대한 최종 판정을 하는 명사사전으로 구성되어 있다.

### 3.1 제거사전

'배제 정보'란 명사가 존재하지 않는 어절의 특징을 품사부착 말뭉치로부터 수집한 정보를 말한다. [4]에서는 음소, 부분 어절, 어절단위 배제 정보의 순으로 검사가 이루어지며, 이 과정을 통하여 명사가 없는 어절에 대해서는 분석과정을 생략하여 형태소 분석에 필요한 탐색공간을 줄이려는 연구가 있었다. 이 연구에서 제안한 '배제 정보'는 한국어 명사 출현 특성을 이용하여 효율적으로 명사를 추출하는 대표적인 방법이라 할 수 있다.

본 논문에서는 품사부착 말뭉치로부터 음소, 어절 부분뿐만 아니라 동사/형용사, 부사/관형사 등의 품사별로 세분화된 '배제 정보'를 추출하였으며, 각 품사에 적용가능한 규칙을 이용하여 '품사별 제거사전'을 구성하였다. 품사별 제거사전을 이용한 시스템은 명사가 출현하지 않는 어절을 각 단계를 거치면서 정확하게 제거함으로써 속도뿐만 아니라 정확성까지도 향상시켰다.

3.1.1 동사/형용사 제거사전

'동사/형용사 제거사전'의 경우 대량의 말뭉치로부터 명사가 출현하지 않은 어절을 대상으로 동사와 형용사가 포함된 어절을 추출한 뒤, 추출된 어절로부터 변형이 일어나지 않는 어근을 추출하여 사전을 구성하였다. 이 사전의 경우 1음절부터 최대 2음절까지의 음절 매칭 결과와 그 후에 변형이 일어나는 음절에 대한 정보를 이용하여 동사/형용사의 여부를 판정한다. 사전에 저장된 예는 표 1, 표 2에 있다.

표 1 1음절 동사/형용사 제거사전의 예

| 어근 | 변형 가능 음절 정보 |
|---|---|
| 합 | 쳐, 쳤, 치, 친, 한, 해, 했 |
| 해 | 쌌, 쳤, 친, 칠 |
| 행 | 할, 함, 했 |
| 향 | 긋, 한, 할, 했 |
| 헐 | 뜯, 렸, 린, 벗, 었 |
| 험 | 한, 했 |
| 해 | 댄, 댔, 졌, 쳐, 친, 펐 |
| 헷 | 갈 |
| 혼 | 난, 날, 났, 쭐 |

표 2 2음절 동사/형용사 제거사전의 예

| 어근 | 변형 가능 음절 정보 |
|---|---|
| 가기 | 가, 까, 는, 도, 란, 로, … |
| 가노 | 라 |
| 가누 | 고, 구, 기, 는, 던, 러, … |
| 가느 | 냐, 니, 다, 라 |
| 가더 | 나, 냐, 니, 라, 란, 랍 |
| 가도 | 록 |
| 가르 | 쳤, 치, 칠, 켜, 켰, 킬 |
| 가리 | 울, 워, 웠, 쳤, 치, 칠, … |
| 가버 | 려, 렸, 리, 린, 릴 |

예를 들어, "가리웠다"의 경우 '가리'가 사전에 어근으로 존재하므로 그 후의 변형 가능 음절에 대해 확인을 하게 되며, '웠'이 어근 '가리'의 변형 가능 음절 정보에 존재하므로 해당 어절은 동사/형용사로 판단하여 분석대상에서 제외시킨다. 만일 "가리비다"라는 어절의 경우 어근인 '가리'가 존재하지만, '비'에 대한 변형 가능 음절 정보가 존재하지 않으므로, 본 사전에서는 분석대상에서 제외시키지 않게 된다. 이와 같이 저장된 '동사/형용사 제거 사전'의 수는 1음절 정보 504개와 2음절 정보 2,127개로 총 2,631개이다.

3.1.2 부사/관형사 제거사전

명사 미 출현 어절에 대한 분석 제거 단계의 마지막인 '부사/관형사 제거사전'은 부사/관형사만 존재하는 것이 아니라, '동사/형용사 제거사전'의 규칙을 적용시키기에 모호성이 존재하여 그 사전에 포함시키지 못한 '동사/형용사' 정보도 포함되어 있다. '부사/관형사 제거사전'에 적용된 규칙은 다음과 같으며, 사전은 1,874개로 이루어져 있다.

규칙 1. 분석 어절과 사전의 정보가 첫 음절부터 매칭이 되는 경우
규칙 2. 분석 어절과 사전의 정보가 완전히 동일한 경우
규칙 3. 분석 어절과 사전의 정보의 매칭 후 다음 음절의 초성이 'ㅎ'으로 시작되는 경우

표 3 부사/관형사 제거사전의 예

| 사전 내용 | 적용규칙 | 사전 내용 | 적용 규칙 |
|---|---|---|---|
| 가가호호 | 1 | 가히 | 2 |
| 가급적 | 1 | 각기 | 2 |
| 가까 | 1 | 갑자기 | 2 |
| 가까스로 | 1 | 든든 | 3 |
| 가끔 | 1 | 듣다못 | 3 |
| 가득 | 1 | 듯 | 3 |
| 가져 | 2 | 따끈 | 3 |
| 가지고 | 2 | 떳떳 | 3 |
| 가질 | 2 | 똑똑 | 3 |

입력된 어절에 대하여 해당 어절이 표 3과 같이 구성된 '부사/관형사 제거사전'에 존재할 경우 사전의 정보에 저장 되어 있는 규칙에 따라 제거 여부를 판정하게 된다. 예를 들어, "듣다못해"라는 어절이 입력되었을 때, '듣다못'이 '부사/관형사 제거사전'에 존재하여 '듣다못'의 적용 규칙을 확인한다. '듣다못'의 경우, 규칙 3이 적용규칙으로 되어 있으므로 입력어절의 '듣다못'다음 음절을 확인하게 되고, 다음 음절의 초성이 'ㅎ'으로 시작되므로 "듣다못해"어절은 분석대상에서 제외시킨다.

또한, 단일어의 경우 일반적으로 명사, 관형사, 부사, 감탄사와 같이 하나의 형태소가 하나의 어절을 이루지만, 둘 이상의 품사를 가지는 단어가 존재한다. 본 논문에서는 입력 어절이 표 3의 '각기'와 같이 단일어로 명사와 부사 두 가지 품사를 가질 수 있는 경우 부사로 결정하여 명사후보에서 제외시킨다.

3.1.3 음소/음절 단위 제거사전

'음소/음절 단위 제거사전'은 [4]의 '음소 단위 배제 정보' 및 '음절 단위 배제 정보'에서 첫 음절에 대하여 적용 가능한 정보를 추출하여 구성하였다. 각 어절의 첫 음절에 대하여 추출한 특성이 존재하면 그 어절에 대하여 명사가 출현하지 않는다고 판단하여 분석 대상에서 제외시킨다.

### 3.2 후절어사전

후절어란 체언 이후에 나타나는 일련의 음절열을 의미한다. 앞선 많은 연구에서 명사를 추출하기 위해 후절어를 사용하였다. 최근에도 '홍진표 외(2008)'[8]에서는 명사 이후에 나타나는 어절 패턴을 이용하여 사전을 구성하였고, 이 결과를 형태소 분석에 사용하였다. 이렇듯 '후절어'를 이용하는 것은 명사 추출을 효율적으로 하기 위한 대표적인 방법이라 할 수 있다. 본 논문에서도 명사를 추출하기 위해 후절어를 사용하며, 대량의 말뭉치로부터 2회 이상 출현한 후절어를 추출하여 사전을 구성하였다. '후절어사전'의 수는 2,907개이며 그림 1은 조사 "에서"에 대한 후절어사전의 예이다.

| | | | | |
|---|---|---|---|---|
| ~에서 | ~에서라도 | ~에서부터는 | ~에서와 | ~에서조차도 |
| ~에서가 | ~에서라면 | ~에서부터도 | ~에서의 | ~에서지만 |
| ~에서건 | ~에서라야 | ~에서부터의 | ~에서이고 | ~에서처럼 |
| ~에서고 | ~에서를 | ~에서야 | ~에서이기는 | |
| ~에서까지 | ~에서마저 | ~에서였는데 | ~에서이든 | |
| ~에서까지도 | ~에서마저도 | ~에서였는지 | ~에서인 | |
| ~에서나 | ~에서만 | ~에서였는지도 | ~에서인가를 | |
| ~에서나마 | ~에서만도 | ~에서였다는 | ~에서인들 | |
| ~에서는 | ~에서만은 | ~에서였다면 | ~에서인지 | |
| ~에서던가 | ~에서만이 | ~에서였던 | ~에서인지는 | |
| ~에서도 | ~에서만이라도 | ~에서였던지 | ~에서일 | |
| ~에서든 | ~에서밖에 | ~에서였으나 | ~에서일는지 | |
| ~에서든지 | ~에서보다 | ~에서였으며 | ~에서일망정 | |
| ~에서라고 | ~에서부터 | ~에서였을 | ~에서임을 | |
| ~에서라는 | ~에서부터가 | ~에서였지만 | ~에서조차 | |

그림 1 후절어사전의 예

제거사전을 통해 분석에서 제외되지 않은 어절에 대하여 최장일치법을 적용하여 명사를 추출하게 된다. 후절어를 통하여 명사를 추출하는 과정을 예로 들면, '집에서조차' 라는 어절에 대하여 후절어인 '~에서조차'와 '~조차'가 후절어 적용 후보가 되지만, 최장일치법에 의해 '~에서조차'가 적용이 되어 '집'이라는 명사가 추출되게 된다. 최장일치법에 의한 명사 추출방법은 분석의 중의성으로 인하여 오분석을 가져 올 수도 있지만, 본 논문에서는 명사사전을 이용하여 이러한 오분석을 교정할 수 있도록 하였다.

### 3.3 명사사전

명사사전은 한국어 형태소 분석을 위해서 많은 언어분석도구에서 사용되고 있다. 하지만, 한국어 명사는 그 수가 매우 많으며, 신조어 또한 계속 늘어나기에 모두 사전으로 구성하기는 힘들다. 따라서, 모바일 기기에 적용시키기 위한 언어분석도구에는 적합하지가 않다. 또한, 명사사전에 대한 의존도가 높아질수록 미등록어에 대한 추출이 어려워진다. 따라서, 본 논문에서는 명사를 추출하는데 후절어를 이용하였다.

후절어사전을 사용하면, 언어분석도구를 사용하지 않고도 효율적으로 명사를 추출할 수 있지만, 분석의 중의성으로 인하여 많은 오분석이 발생하게 된다. 예를 들어, "상대로"를 형태소 분석기를 이용하여 명사를 추출하면, "상대/일반명사+로/부사격조사"로 분석이 되어 명사인 '상대'를 추출하게 된다. 그러나 후절어를 이용하면 후절어사전에 '~대로'와 '~로'가 존재하여, '상대'와 '상'이 명사후보로 추출되지만 최장일치법에 의해 '상'이라는 잘못된 명사를 추출하게 된다. 이런 분석의 중의성을 해소하기 위해서 본 논문에서는 명사사전을 사용하게 된다. 본 논문에서 사용한 명사사전은 하나의 명사에 대하여 다음과 같은 정보가 같이 포함되어 저장된다.

1. 음절 정보
2. 명사
3. 적용 규칙

음절 정보는 후절어가 가진 첫 음절들 중 오분석이 일어날 가능성이 있는 음절을 추출한 것이며, 명사 정보는 그 음절을 마지막 음절로 가지고 있는 명사들이다. 그리고 적용 규칙은 최종 명사 판정을 위하여 해당 명사에 적용되는 규칙을 말한다. 이런 음절 정보의 개수는 31개이며, 음절 정보에 해당되는 명사의 개수는 2,379개이다. 표 4는 분석의 중의성이 있는 '대'음절에 대한 명사사전의 예이다.

적용된 후절어의 첫 음절이 명사사전의 음절 정보에 존재하는 음절일 경우, 추출된 명사후보에 후절어의 첫 음절을 포함하여 명사사전을 검색한다. 앞의 예에서, "상대로"의 경우 최장일치법에 의하여 적용된 후절어 '~대로'의 첫 음절인 '대'가 명사사전의 음절 정보에 존재하는 음절이므로, '대'를 추출된 명사후보 '상'과 합쳐서 '상대'라는 명사후보를 만들게 된다. 만약 생성된 명사후보 '상대'가 명사사전에 존재한다면 최종 명사 판정은 해당 명사의 적용 규칙을 따른다. 적용 규칙은 다음과 같다.

표 4 명사사전의 예

| 음절정보 | 명사 | 적용 규칙 | 음절정보 | 명사 | 적용 규칙 |
|---|---|---|---|---|---|
| … | … | … | 대 | 서울대 | 2 |
| 대 | 사립대 | 1 | 대 | 선발대 | 1 |
| 대 | 사열대 | 2 | 대 | 선봉대 | 1 |
| 대 | 삼각대 | 2 | 대 | 성대 | 2 |
| 대 | 상대 | 2 | 대 | 성삼대 | 2 |
| 대 | 상지대 | 2 | 대 | 세대 | 1 |
| 대 | 생리대 | 2 | 대 | 세면대 | 2 |
| 대 | 서강대 | 2 | 대 | 수사대 | 2 |
| 대 | 서산대 | 2 | … | … | … |

규칙 1. 생성된 명사 후보와 명사 사전의 명사와 일치하지 않아도, 생성된 명사 후보를 최종 명사로 추출

규칙 2. 생성된 명사 후보가 사전의 명사와 완전히 일치할 경우에만 생성된 명사 후보를 최종 명사로 추출

규칙 1은 사전에 존재하는 명사의 앞에 다른 명사가 붙어 복합명사로 사용 될 수 있는 명사들이다. 표 4의 '사립대'의 경우, '지방사립대'나 '서울사립대'와 같이 '사립대'에 '지방'이나 '서울' 등의 명사가 붙어 복합명사를 이루는 경우가 많다. 그러나, 이런 모든 명사를 사전에 포함 시키는 것은 명사사전의 수를 늘리기만 할 뿐, 효율적이지 못하다. 따라서, '사립대'와 같이 복합명사로의 사용이 빈번한 명사에 대해서는 1번 규칙을 적용하여 저장함으로써, 명사사전의 수를 줄였다. 반면, 복합명사로의 사용이 적은 경우는 입력된 어절과 완전히 일치하는 명사후보만 명사로 추출하여 오분석을 줄일 수 있도록 규칙 2를 적용하였다. 앞선 예의 '지방사립대로'에 대한 분석은 다음과 같은 순서로 진행된다.


그림 2 명사 사전 규칙 적용의 예

1. 최장일치법에 의하여 '~대로'가 후절어로 적용하여 명사 후보 '지방사립'을 생성.
2. 적용된 후절어 '~대로'의 첫음절인 '대'가 명사사전의 음절 정보에 존재.
3. 추출된 명사 후보인 '지방사립'과 적용된 후절어 '~대로'의 첫 음절인 '대'를 결합하여 새로운 명사 후보 '지방사립대'를 생성.
4. 명사사전에 '사립대'가 규칙 1로 존재.
5. '사립대'의 적용 규칙이 1이므로 '지방사립대'의 복합명사 사용이 허용.
6. 생성된 명사 후보 '지방사립대'를 최종 명사로 추출.

## 4. 명사의 추출

본 논문에서 제안한 명사추출기의 전체적인 구성은 그림 3과 같다. 본 논문에서는 한글 명사에 대한 추출이 목적이므로 외래어는 분석에서 제외시키도록 하였다. 전처리 과정은 기호와 외국어를 제거하고 분석을 위한 정보를 저장한다. 전처리 과정이 끝난 어절은 제거사전의 정보를 통해 명사의 존재 유무를 판단하여 명사가 존재하지 않는 어절은 분석대상에서 제외한다. 제외되지 않은 어절은 후절어를 적용하여 명사를 추출하게 되는데, 제안한 시스템은 제거사전을 통해서 제거되지 않았음에도 불구하고, 적용되는 후절어가 없는 경우, 단일어라고 판단하고 해당 어절을 명사로 추출하게 된다. 마지막으로 명사사전을 통하여 중의성이 있는 후절어가 적용되었는지 판단한 후, 최종적으로 명사를 추출하게 된다.

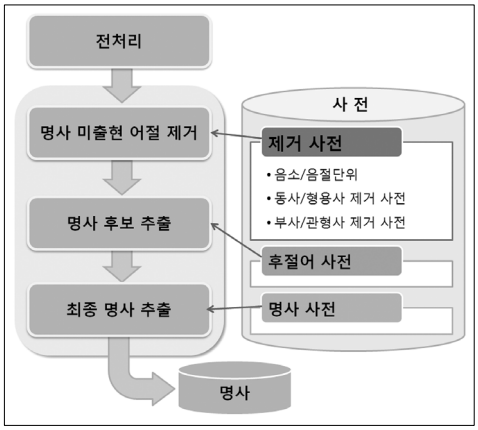


그림 3 명사 추출기의 구성

각 단계의 자세한 진행 과정은 다음과 같다.

### 4.1 전처리 단계

전처리 단계에서는 분석을 용이하게 하기 위하여 외래어와 기호를 제거한 후, 어절 단위로 저장한다. 각 어절에 대한 분석 정보는 첫 음절부터 하나씩 제거된 음절열이 저장되게 된다. 예를 들어, "지방사립대로 편입합니다"라는 문장이 입력되었을 경우, "지방사립대로"와 "편입합니다"라는 어절로 나누고, 해당 어절에 대하여

그림 4와 같은 음절열 정보가 저장된다.


그림 4 음절열 정보의 예

### 4.2 제거사전을 이용한 명사 미 출현 어절 제거

전처리 과정을 거친 각 어절에 대하여 제거사전을 적용하여 명사 출현 가능성에 대한 판정을 내리게 된다. 이 단계는 언어분석도구를 사용하지 않고 명사를 추출하기 위하여 명사가 출현하지 않는 어절을 제거함으로써 분석 속도를 높이고, 오분석을 줄이기 위한 단계이다. 따라서, 본 논문에서는 이 단계의 정확성을 높이기 위해서 사전을 품사별로 나누고, 각 사전 마다 다른 적용 규칙을 사용하였다. 분석의 순서는 다음과 같다.

1. 음소/음절 단위 검사
2. 동사/형용사 제거사전 검사
3. 부사/관형사 제거사전 검사

각 어절에 대하여 제거사전을 검사하여 각 사전의 적용 규칙과 일치하는 어절이 존재할 경우, 그 어절 내에서는 명사가 존재하지 않으므로 분석에서 제외시킨다. 입력된 어절인 “지방사립대로”와 “편입합니다”의 경우, 제거사전의 적용 규칙과 일치하는 내용이 없으므로 명사가 존재한다고 판단하여 다음 단계로 진행된다.

### 4.3 후절어를 이용한 명사 후보 추출

앞서 말한 것과 같이 본 논문에서 제안한 시스템에서는 제거 단계를 거치고도 후절어가 적용되지 않는 어절에 대해서는 단일어 명사라 판단하고 명사로 추출하게 된다. 그 외의 어절에 대해서는 최장일치법을 통하여 후절어를 적용, 명사후보를 추출하게 된다. 그림 5는 “지방사립대로”와 “편입합니다”라는 어절에 대하여 후절어를 적용한 예이다.

그림 5의 예에서 “지방사립대로”에 대하여 적용된 후절어는 ‘~대로’와 ‘~로’ 이다. 그리고, “편입합니다” 어절에 대하여 적용된 후절어는 ‘~합니다’와 ‘~다’이다.

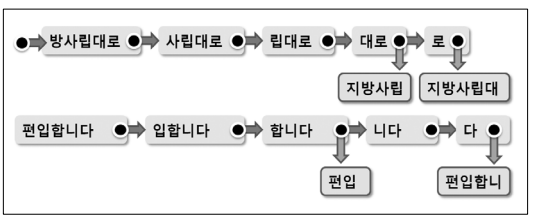


그림 5 각 어절의 후절어 적용 예

두 어절 “지방사립대로”와 “편입합니다”에 대하여 각각 두 개의 후절어가 적용이 되었지만 시스템은 최장일치법에 의해 “지방사립대로”어절은 ‘~대로’를, “편입합니다” 어절은 ‘~합니다’를 적용한 후, 명사후보로 ‘지방사립’과 ‘편입’을 추출하게 된다.

### 4.4 명사사전을 이용한 중의성 해소 및 최종 명사 추출

중의성이 있는 후절어를 통한 명사 추출은 잘못된 명사를 추출할 가능성이 있다. 따라서, 본 논문에서는 중의성이 있는 후절어의 첫 음절에 대한 명사사전을 구성하였으며 이 단계에서 후절어의 오적용 여부를 판단하게 된다. 만약 추출된 명사후보에 대하여 적용된 후절어의 첫 음절이 명사사전 리스트에 있다면, 사전을 검색한 후 적용 규칙에 의하여 최종 명사를 추출하게 된다.

그림 5의 예에서, “편입합니다”라는 어절에 대해 최종적으로 적용된 후절어 ‘~합니다’ 의 경우, 후절어 첫 음절인 ‘합’은 명사사전 리스트에 존재하지 않으므로 명사후보인 “편입”을 최종 명사로 판정한다. 그러나, “지방사립대로”에 적용된 ‘~대로’의 경우, 첫 음절인 ‘대’가 명사사전의 음절 정보에 존재한다. 따라서, “지방사립대로”에 적용된 ‘~대로’는 중의성이 있는 후절어이므로 그림 2에 설명한 순서대로 명사사전을 이용하여 중의성을 해소한 후, “지방사립대”를 최종 명사로 판정한다.

## 5. 실험 및 성능 평가

본 논문에서 제안한 명사 추출기와 타 연구와의 공정한 비교를 위해서는 동일한 실험 말뭉치와 실험 환경이 필요하다. 그러나 동일한 실험 환경에서 실험을 한다는 것은 사실상 불가능하기 때문에 분석 속도에 대한 비교 평가는 하지 않았다. 실험 말뭉치는 MATEC′99에서 사용한 ETRI 28만 품사부착 말뭉치를 사용하였으며, 학습 말뭉치는 21세기 세종계획 말뭉치를 사용하였다. 이에 대한 실험 말뭉치의 특성은 표 5와 같다.

표 5 실험 말뭉치의 특성

| 분야 | 문서 수(개) | 어절 수(개) |
|---|---|---|
| 소설 | 26 | 165,223 |
| 비소설 | 44 | 108,386 |
| 뉴스 | 41 | 12,655 |
| 전체 | 111 | 286,264 |

### 5.1 평가 척도

명사 추출기의 성능 평가를 위해서 정확률(Precision), 재현율(Recall), $F_1$-measure를 사용하였으며, 각각의 값은 다음과 같이 계산된다.

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

$$재현율(R) = \frac{정답과\ 일치하는\ 응답\ 명사의\ 개수}{정답\ 명사의\ 개수}$$

$$F_1 - measure = \frac{2PR}{P+R}$$

본 논문에서는 MATEC'99에서 사용한 평가 방식과 평가 척도를 따른다[9].

### 5.2 성능 비교 실험

명사 추출기의 성능을 정확률과 재현율, 그리고 $F_1$-measure로 나타내었다. 실험 대상으로는 본 논문에서 제안한 시스템과 <유형 2> 방식의[4], 그리고 <유형 3> 방식의 '장동현 외(1999)'[6]에 대하여 성능을 비교하였다.

표 6 성능 비교 실험 결과

| 분야 | | 제안 시스템 | 유형 2 [4] | 유형 3 [6] |
|---|---|---|---|---|
| 뉴스 | 재현율 | ***0.91*** | 0.90 | 0.90 |
| | 정확률 | ***0.89*** | 0.87 | 0.81 |
| 비소설 | 재현율 | ***0.92*** | 0.92 | 0.92 |
| | 정확률 | ***0.86*** | 0.84 | 0.81 |
| 소설 | 재현율 | ***0.88*** | 0.91 | 0.90 |
| | 정확률 | ***0.78*** | 0.84 | 0.65 |
| 전체 | 재현율 | ***0.89*** | 0.91 | 0.91 |
| | 정확률 | ***0.82*** | 0.84 | 0.77 |
| $F_1$-measure | | ***0.86*** | 0.87 | 0.83 |
| 명사사전의 수 | | ***2,379*** | 50.772 | 5,854 |

[4]의 시스템은 <유형 2>의 방식으로써 언어분석도구를 사용한 방식이다. 따라서, 가장 높은 $F_1$-measure 값을 나타내고 있지만, 언어분석도구를 사용함으로 인하여 시스템과 사전의 용량이 크고, 계산량이 많아서 모바일 기기에 직접적으로 적용시키기에는 무리가 있다. 또한 [6]의 시스템은 <유형 3>의 방식으로써 본 논문의 유형과 일치한다. 그러나, 많은 오분석으로 인하여 상대적으로 성능이 낮은 것을 확인할 수 있다. 본 논문에서 제안한 시스템은 [6]과 같은 <유형 3>의 방식을 사용하였지만, 한국어 명사 출현 특성을 이용함으로써 [6]의 시스템 보다 약 3%의 높은 성능을 보였다.

### 5.3 명사사전 비교

그림 6은 제안된 시스템들에서 사용한 명사사전의 수를 비교하고 있다.

본 논문에서 제안한 시스템은 한국어 명사 출현 특성과 적용 규칙을 이용하여 2,379개의 명사만을 사용하여 명사사전을 구성하였으며, 사전의 용량은 146KB이다. 제안된 시스템의 명사사전은 최소 50,772개에서 최대 150,936개의 명사를 사용한 [4]의 시스템의 명사사전에 비하여 약 96%를 줄이고도, 성능 차이는 약 1% 밖에 보이지 않으며, MATEC'99의 명사 추출 부분의 평균 $F_1$-measure인 0.824보다 높은 성능을 보였다.

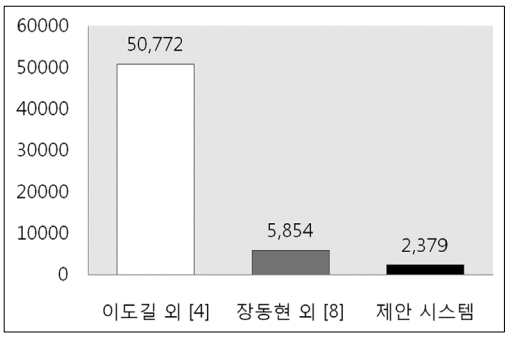


그림 6 명사사전의 수

### 5.4 사전의 역할 분석

본 논문이 제안한 세분화 된 사전의 적용 규칙과 명사 사전의 역할을 분석하기 위하여 세분화된 사전의 적용 여부를 달리하여 성능을 평가해 보았다. 여기서 나타나는 성능은 ETRI 28만 품사부착 말뭉치의 분야를 나누지 않은 전체 말뭉치에서 재현율과 정확률 그리고 $F_1$-measure 값으로 나타내었으며, 각 사전이 제거한 어절의 수를 확인해 보았다.

표 7 사전 역할 분석 실험

| 척도 | (1) | (2) | (3) | 모두 사용 |
|---|---|---|---|---|
| 재현율 | 0.91 | 0.91 | 0.88 | 0.89 |
| 정확률 | 0.71 | 0.73 | 0.81 | 0.82 |
| $F_1$-measure | 0.80 | 0.81 | 0.84 | 0.86 |
| 제거 어절 수 | 58,245 | 48,959 | 84.405 | 84,405 |

(1)의 경우는 '동사/형용사 제거사전'을 사용하지 않은 경우이고, (2)의 경우는 '부사/관형사 제거사전'을 사용하지 않은 경우이다. 또한, (3)의 경우는 '명사사전'을 사용하지 않은 결과를 나타낸다. 마지막의 경우는 모든 사전을 사용한 결과이다. 실험 결과를 살펴보면, (1)과 (2)의 경우에는 모든 사전을 사용한 경우보다 분석에서 제외되지 않은 어절의 수가 많아서 재현율은 높게 나타나지만, 오분석으로 인하여 정확률이 매우 떨어지는 것을 확인할 수 있다. 그리고, (3)의 경우보다 사전을 모두 사용한 경우가 성능이 약 2% 가량 높은 이유는 명사 사전의 사용으로 인하여 오분석의 비율이 낮아졌기 때문으로 분석된다.

### 5.5 세분화 된 사전의 효과

본 논문에서 사용한 품사별로 세분화된 제거사전의 효과를 확인하기 위하여 제거사전의 세분화 적용 유무를 달리하여 실험을 실시하였다. 세분화를 적용한 사전은 본 논문에서 제안한 방법으로, 품사부착 말뭉치로부터 추출한 배제 정보를 품사별로 나누어 제거사전을 구

표 8 규칙 적용 효과 실험

| 분류 | 세분화 미적용 | 세분화 적용 |
|---|---|---|
| 제거 어절 수 | 58,796 | 84,405 |
| 재현율 | 0.78 | 0.88 |
| 정확률 | 0.80 | 0.81 |
| $F_1$-measure | 0.79 | 0.84 |

성하고, 각 사전마다 다른 규칙을 적용한 것이다. 세분화를 적용하지 않은 사전은 품사별로 나누지 않고 제거 사전을 구성하였다. 세분화를 적용하지 않은 사전은 품사별로 다른 규칙을 적용할 수 없으므로, 음절 매칭을 이용하여 어절의 제거 유무를 결정하였다. 또한, 동등한 비교를 위하여 본 논문에서 제안한 '명사사전'을 사용하지 않고 성능을 평가하였다.

표 8을 살펴보면 세분화 된 사전을 적용시킨 시스템이 적용시키지 않은 시스템보다 약 26,000어절 정도 더 많은 어절을 분석에서 제외시켰으며, 재현율과 정확률에서 모두 성능이 향상 되었다는 것을 알 수 있다. 이것은 품사별로 세분화 된 제거사전이 명사가 출현하지 않는 어절을 더욱 정확하게 분류해 내어 오분석과 과분석을 줄인 결과이다. 따라서, 제거사전을 세분화 하여 각 품사마다 다른 규칙을 적용시키는 것이 효과가 있다는 것을 보여준다.

## 6. 결론 및 향후 연구

본 논문에서는 모바일 기기에 적합한 명사 추출기를 제안하였으며 타 연구와의 비교 실험을 실시하였다. 본 논문에서 제안한 시스템은 기존의 언어분석도구를 사용하여 명사를 추출하는 방식이 아닌, 명사를 포함하지 않는 어절을 제거하고 후절어를 이용하여 명사를 추출하는 방법이다. 이를 위해, 세분화된 제거사전을 이용하여 분석을 정확성을 높이고 후절어를 이용한 명사 추출 기법이 가지고 있는 문제점인 중의성을 해소하기 위하여 중의성 해소를 위한 명사사전을 구축하고 사용하였다. 실험을 통하여 제안한 시스템이 작은 수의 명사사전을 가지고도 높은 성능을 나타낼 수 있다는 것을 알 수 있었다. 따라서, 제안한 시스템은 모바일 기기와 같이 낮은 컴퓨팅 파워와 작은 메모리만을 사용할 수 있는 환경에 적합하다 할 수 있다. 제안한 시스템이 사용한 사전이 한국어의 모든 특성을 반영하지는 못하므로 포함되지 못한 패턴이 더 추가가 된다면 지금의 성능보다 향상 될 수 있다. 하지만, 그만큼 사전의 용량이 커지게 되므로 성능과 프로그램 크기의 관계는 상충적이라고 할 수 있다.

향후 연구로, 본 논문에서 사용한 사전은 21세기 세종계획 말뭉치로부터 추출한 패턴으로부터 수작업을 통하여 사전을 구성하였으나, 수작업의 양을 줄이면서 패턴을 추출하는 방법을 연구할 계획이다. 또한, '후절어사전'의 경우 명사 이후에 나타나는 모든 음절열을 후절어로 구성하였으나, 후절어 역시 일련의 패턴이 존재하는 것을 알 수 있었다. 따라서, 후절어도 패턴을 적용하여 사전의 크기를 줄일 수 있는 방안도 연구해 볼 것이다.

## 참 고 문 헌

[1] D. An, "A Noun Extractor using Connectivity Information," *Proc. Morphological Analyzer and Tagger Evaluation Contest (MATEC'99)*, pp.173-178, Oct. 1999. (in Korean)

[2] N. Kim, Y. Seo, "A Korean Morphological Analyzer CBKMA and A Index Extractor CBKMA/IX," *Proc. Morphological Analyzer and Tagger Evaluation Contest (MATEC'99)*, pp.50-59, Oct. 1999. (in Korean)

[3] J. Lee, B. Shin, K. Lee, J Kim, S. Ahn, "Noun Extractor based on a multi-purpose Korean morphological engine implemented with COM," *Proc. Morphological Analyzer and Tagger Evaluation Contest (MATEC'99)*, pp.167-172, Oct. 1999. (in Korean)

[4] D. Lee, S. Lee, H. Rim, "An Efficient Method for Korean Noun Extraction Using Noun Patterns," *Journal of KIISE : Software and Applications*, vol.30, no.1-2, pp.173-183, Feb. 2003. (in Korean)

[5] J. Shim, J. Kim, J. Cha, G. Lee, "Robust Part-of-Speech Tagger using Statistical and Rule-based Approach," *Proc. Morphological Analyzer and Tagger Evaluation Contest (MATEC'99)*, pp.60-75, Oct. 1999. (in Korean)

[6] D. Jang, S. Myaeng, "A Noun Extractor based on Dictionaries and Heuristic Rules Obtained from Training Data," *Proc. Morphological Analyzer and Tagger Evaluation Contest (MATEC'99)*, pp.151-156, Oct. 1999. (in Korean)

[7] W. Lee, S. Kim, G. Kim, K. Choi, "Implementation of Modularized Morphological Analyzer," *Proc. Morphological Analyzer and Tagger Evaluation Contest (MATEC'99)*, pp.123-136, Oct. 1999. (in Korean)

[8] J. Hong, J. Cha, "A New Korean Morphological Analyzer using Eojeol Pattern Dictionary," *Proc. of the KCC-2008*, vol.35, no.1, pp.279-284, June. 2008. (in Korean)

[9] J. Lee, J. Park, K. Cha, S. Park, "Morphological Analyzer and Tagger Evaluation Contest(MATEC99) Overview," *Proc. Morphological Analyzer and Tagger Evaluation Contest (MATEC'99)*, pp.13-22, Oct. 1999. (in Korean)

박 용 현

2010년 동아대학교 컴퓨터공학과(학사)
2010년~현재 동아대학교 컴퓨터공학과 석사과정. 관심분야는 자연어처리, 형태소 분석, 대화형 시스템 등

황 재 원

정보과학회논문지 : 소프트웨어 및 응용
제 37 권 제 4 호 참조

고 영 중

정보과학회논문지 : 소프트웨어 및 응용
제 37 권 제 4 호 참조