https://doi.org/10.5392/JKCA.2017.17.04.576

명사후문자열을 이용한 미등록어 인식

Korean Unknown-noun Recognition using Strings Following Nouns in Words

박기탁, 서영훈
충북대학교 컴퓨터공학과

Ki-Tak Park(kitakpark@chungbuk.ac.kr), Young-Hoon Seo(yhseo@chungbuk.ac.kr)

요약

사전에 등록되지 않은 미등록어는 형태소분석에서 뿐만 아니라 자연언어처리의 모든 분야에서 문제를 발생시킨다. 본 논문에서는 명사후문자열을 이용하여 미등록어를 인식하는 방법을 제안한다. 명사후문자열이란 명사를 포함하고 있는 어절에서 명사 뒤에 나오는 문자열을 의미하며, 조사, 접미사+조사, 동사화접미사+어미 등이 이에 속한다. 문서에 출현한 미등록어 포함 어절들을 모아 정렬한 다음, 동일한 앞부분을 가지는 어절이 두 개 이상일 경우에 한하여 미등록어 인식을 시도한다. 이 어절들에서 동일한 앞부분을 미등록 명사로, 그 다음 음절부터 끝 음절까지를 명사후문자열로 추정한다. 그리고 세종말뭉치에서 추출한 명사후문자열 정보를 이용하여 미등록 명사를 결정한다. 포털사이트 기사를 이용하여 실험한 결과, 2가지 형태 이상으로 출현한 미등록어에 대해 정확률 99.64%, 재현율 99.46%의 높은 인식 성능을 보였다.

■ 중심어 : | 미등록어 인식 | 명사후문자열 | 한국어 언어처리 | 세종 말뭉치 | 유사어절 비교 |

Abstract

Unknown nouns which are not in a dictionary make problems not only morphological analysis but also almost all natural language processing area. This paper describes a recognition method for Korean unknown nouns using strings following nouns such as postposition, suffix and postposition, suffix and eomi, etc. We collect and sort words including nouns from documents and divide a word including unknown noun into two parts, candidate noun and string following the noun, by finding same prefix morphemes from more than two unknown words. We use information of strings following nouns extracted from Sejong corpus and decide unknown noun finally. We obtain 99.64% precision and 99.46% recall for unknown nouns occurred more than two forms in news of two portal sites.

■ keyword : | Unknown-noun Recognition | Noun-following Strings | Korean Processing | Sejong Corpus | Similarity Eojeol Comparison |

I. 서 론

자연언어이해, 기계번역, 정보검색, 질의응답시스템 등 다양한 자연언어처리 분야에서 미등록어는 많은 문제를 유발시킨다. 미등록어는 사전에 등록되어 있지 않은 단어를 의미하며, 일반적으로 신조어, 고유명사, 복합명사, 합성어, 전문용어, 외래어, 약어와 같은 명사가 거의 대부분이다. 이들 미등록어는 자연언어이해의 첫

접수일자 : 2016년 12월 23일
수정일자 : 2017년 01월 31일
심사완료일 : 2017년 02월 06일
교신저자 : 서영훈, e-mail : yhseo@chungbuk.ac.kr

단계인 형태소분석을 어렵게 할 뿐만 아니라, 구문분석, 의미분석 등 자연언어이해 모든 단계에서 처리를 어렵게 하거나 불가능하게 만든다. 또한 기계번역에서는 번역 단어를 찾지 못하게 하고, 정보검색에서는 관련 정보의 검색을 어렵게 한다. 실제로 인터넷 뉴스 기사를 대상으로 조사한 결과, 조사 대상 어절 51,435개에서 중복을 제외한 미등록어의 개수가 3,320개로 약 6.5%에 이르렀다. 상대적으로 미등록어가 많지 않은 뉴스 기사에서도 이와 같이 무시할 수 없을 정도의 미등록어가 출현하기 때문에 미등록어의 처리는 자연언어처리의 성능 향상을 위해 매우 중요하다.

미등록어 인식 및 수집에 대한 기존 연구들은 형태소 분석을 이용하거나 형태소의 품사 패턴, 명사의 출현 특성 등을 사용한다. 형태소 분석이나 명사 출현 특성 등을 이용하면 비교적 정확하게 명사를 추출할 수 있지만, 어휘 사전에 존재하지 않는 단어를 정확하게 추출할 수 없는 경우가 많다. 예를 들어, '세리자와', '닌텐도', '화웨이'등과 같이 명사가 명사+조사로 분석될 가능성이 있는 경우, '해찬들', '바빌론'등과 같이 명사가 명사+접미사로 분석될 가능성이 있는 경우에는 미등록명사를 정확하게 결정하기 어렵다. 따라서 조사, 접미사 등을 포함한 미등록어나 형태소 분석이 어려운 미등록어를 정확히 인식할 수 있는 방법이 필요하다.

본 논문에서는 대용량의 말뭉치로부터 구성한 명사후문자열 사전을 이용하여 미등록어를 인식한다. 제안하는 방법은 형태소 분석을 이용하지 않고, 문서에 있는 전체 어절 중에서 하나의 미등록명사를 포함하면서 서로 다른 두 가지 이상의 형태로 출현한 단어를 대상으로 미등록어를 인식한다. 본 논문은 형태소분석이 목표가 아니고, 미등록어의 수집이 목표이기 때문에 높은 정확률을 가지는 미등록어 수집 방법을 제시한다.

본 논문의 구성은 다음과 같다. 먼저, 2장에서는 기존에 국내외에서 이루어진 미등록어 인식과 관련된 연구에 대해 소개하고, 3장에서는 명사후문자열을 이용하기 위한 사전의 구성 및 미등록어 인식을 위한 방법을 기술한다. 4장에서는 실험 결과와 평가를 진행하고, 5장에서는 결론과 향후 연구에 대해서 기술한다.

II. 관련 연구

영어의 미등록어 탐색의 경우, 한 단어는 하나의 품사로만 구성되므로 형태소 분석이 필요 없다. 문장에서 미등록어가 나오면 해당 단어 주변 문맥을 파악하고, 명사로 여겨지는 단어의 원형을 파악해 미등록어인지 판단한다. Aastha Gupta[1]는 미등록어가 등장하는 문장(S)과 미등록어(U)에 태그를 부착하고, 웹으로부터 앞에서 부착한 태그와 유사한 도메인을 수집하여(다중 도메인 웹 기반 알고리즘), 이 수집된 도메인들로 미등록어(U)에 대한 새로운 태그 할당 가능성을 계산하여 미등록어를 인식하였다. 김형철[2]은 영어의 접두사, 접미사, 품사 자질을 이용하여 미등록어의 품사를 추정하였다.

그리스어는 고도의 굴절 언어로, 소리나는대로 표기하고, 풍부한 어휘의 양, 문법의 특이성 때문에, 영어처럼 미등록어의 주변 문맥을 통해서는 인식하기 힘들다. Giorgos S. Orphanos[3]는 넓은 범위의 사전을 바탕으로 모든 품사의 모호성 관계를 보여주는 형태소 분석 말뭉치를 준비하고 미등록어의 특성을 준비하여, 품사 추측이 필요한 단어에 대해 준비한 말뭉치를 이용해 의사 결정 트리의 유도 과정을 진행하여 적합한 패턴을 찾아 품사를 부착하는 강력한 품사 태거를 만들었다.

한국어는 하나의 어절이 여러 가지 형태소로 이루어지는 교착어이기 때문에 미등록어 인식을 포함한 형태소분석이 활발하게 연구되었다.

강승식[4]은 조사사전을 이용하여 형태소 분석에 실패한 어절로부터 최장조사를 떼어나는 최장조사일치법으로 미등록 명사를 추정하였다. 박봉래[5]는 형태소 분석에 실패한 어절을 모아 그룹화한 후, 어근이 되는 미등록어를 추출하는 유사어절비교방식으로 미등록 명사를 추정하였다. 양장모[6]는 미등록어의 결합 관계 정보, 미등록어 뒤에 나오는 형태소의 의미 정보, 좌.우 어절과의 연관 관계, 전체 문헌에서 미등록어가 사용된 유형 등의 언어 정보를 이용하여 미등록어의 위치 추정 및 고유 명사의 범주를 추정하였다. 박봉래[7]는 각 어절을 그들의 부분 문자열을 이용하여 색인하고 그 색인 리스트를 이용하여 특정 길이의 윈도우 안에서 동일 음

절열을 가지고 있는 유사 어절들을 구해서 미등록어가 포함된 어절을 인식하였다. 차정원[8]은 형태소 패턴 사전을 이용하여 미등록어를 해당 형태소들로 분리하였다. 미등록어의 음절 패턴에 따라 패턴 사전으로부터 가능한 품사 태그들을 모두 추출하고 태깅 단계에서 가장 가능성 있는 후보를 선택하였다. 박봉래[9]는 동일한 미등록어가 사용된 용례 어절 또는 용례 구의 비교 분석을 통한 미등록어를 인식하였다. 김선호[10]는 접미사 배열(suffix array)과 최장 공통 앞 문자열(longest common prefix)을 이용해 명사를 분리하고, 명사가 추출되는 경우 로컬 사전에 추가하였다. 이도길[11]은 명사 출현 특성을 이용해 한국어 명사를 추출하는 방법을 제안하였다. 명사 출현 특성은 명사가 나타나는 특성에 대한 정보인 명사 접미 음절열과 명사가 나타나지 않는 특성에 대한 정보인 배제 정보가 있다. 박소영[12]은 문서에서 여러 번 나타난 미등록어에 대해 전문분석(유사어절비교)을 통해 미등록명사를 인식하는 단계, 문서에 한번 나타난 미등록어에 대해서도 웹문서를 바탕으로 광범위하게 인식하는 단계, 기본형이 어절에 그대로 나타나는 미등록명사뿐만 아니라 기본형이 변형하여 나타날 수 있는 미등록용언을 인식하는 단계, 이 3가지 단계로 미등록어를 인식하였다. 박용현[13]은 명사 출현 특성과 후절어 사전을 이용하여 명사를 추출했다. 최맹식[14]은 미등록어절 패턴과 SVM(Support Vector Machine)을 단계적으로 사용하여 미등록어로 인해 형태소 분석이 잘못된 어절을 인식하였다. 김보겸[15]은 형태소 분석 말뭉치에서 자동으로 접미사 패턴을 학습하고, 고유명사나 일반명사와 같은 등록어 뒤에 붙는 다양한 접미사를 분석하여 미등록어를 추정하였다.

본 논문은 기존 논문 [5][7][9][12]와 같은 유사어절비교를 이용한 미등록어 탐색 기법과 유사한 방법이지만 음절의 크기를 1부터 N까지 차례로 키워가며 유사어절 그룹화를 하는 점과 유사어절 그룹에서 명사가 추출되는데 사용된 어절은 다음 실험 대상에서 제외하는 점이 다르다. 또, 본 논문에서 사용한 명사후문자열은 [4][10][13]과 같은 명사 뒤에 나오는 조사 사전 또는 접미사 배열 정보, 후절어 사전을 이용한 미등록어 탐색 기법과 유사하지만 명사후문자열 추출 과정에서 정확도를 높이기 위해 4번 이상 출현한 명사후문자열만 추출해 사용하는 점이 다르다. 이외에도, 형태소의 품사 패턴을 이용한 미등록어 탐색[8][14][15], 명사의 출현 특성을 이용한 미등록어 탐색 방법 [6][11][13] 등이 있었다.

III. 본 론

3. 명사후문자열을 이용한 미등록어 인식

3.1 미등록어 인식을 위한 사전

본 논문에서는 형태소 분석을 하지 않으면서 높은 정확률로 미등록어를 수집하기 위해 세가지 사전, 즉, 어절 사전, 명사후문자열사전, 명사사전을 이용한다. 이 사전들은 세종 말뭉치[16] 중 약 6천만 어절 규모의 문어원시 말뭉치와 약 1천 5백만 어절 규모의 형태분석말뭉치로부터 추출하였다. 각 사전과 사전 규모는 [표 1]에 있다.

어절 사전은 세종 현대문어원시말뭉치에 등장한 모든 어절들을 모아 놓은 사전이다. 이 사전은 문서에 출현한 어절들 중에서 미등록어가 포함된 단어를 추출하기 위한 사전이다. 즉, 문서에서 추출한 단어중 이 사전에 없는 단어들을 대상으로 미등록어를 추출한다.

표 1. 사전의 구성

사전 내용	사전 크기
어절 사전	2,579,965
명사후문자열 사전	5,104
명사 사전	130,478

명사후문자열 사전은 세종 형태분석말뭉치에서 명사가 포함된 모든 단어를 추출한 다음, 명사 다음에 나오는 문자열들을 모은 사전이다. 즉, 조사, 접미사, 접미사+조사, 용언화접미사+어미 등을 포함한 다양한 형태소 조합들로 명사후문자열이 구성된다. 이 명사후문자열 사전에는 말뭉치에 등장한 명자후문자열의 횟수까지 저장되어 있는데, 이 횟수는 중의성을 가지는 미등록어 대상중 하나를 선택하는 용도로 사용된다. 말뭉치의 태깅 오류 등을 고려하여 말뭉치에서 4번 이상 출현하는 명사후문자열 5,104개를 미등록어 수집에 이용한다.

명사 사전은 미등록어로 인식된 명사가 이미 등록되어 있는 명사인지 확인하기 위해 사용하는 사전이다. 즉, 세종 형태분석말뭉치에서 명사후문자열 앞에 나타난 명사들을 수집해 놓은 사전으로, 사전에 등록되어 있는 명사 사전으로 대체할 수 있다.

3.2 미등록어의 인식

본 논문에서 제안하는 미등록어를 인식하는 단계는 [그림 1]과 같다.

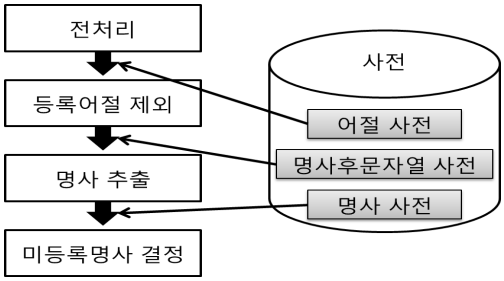

그림 1. 미등록어 인식기

3.2.1 전처리

전처리 단계에서는 인터넷에서 기사를 수집하고 한글만 처리하기 위하여 기호, 영어, 한자, 숫자를 제거한 후, 어절 단위로 정렬하여 저장한다. 예를 들어, “겜블릭(대표 박정준)은 자사가 개발하는 전략 모바일 RPG <삼국지천하대전>이 사전예약을 진행한다고 5일 밝혔다.”라는 문장이 입력될 때 ‘겜블릭’, ‘대표’, ‘박정준’, ‘은’, ‘자사가’, ‘개발하는’, ‘전략’, ‘모바일’, ‘삼국지천하대전’, ‘이’, ‘사전예약을’, ‘진행한다고’, ‘일’, ‘밝혔다’라는 어절로 나뉜다.

3.2.2 등록어절 제외

등록어절 제외 단계에서는 전처리 단계에서 추출한 어절들을 어절사전과 비교하여, 어절사전에 존재하지 않는 어절만을 미등록어 후보 어절로 추출한다.

3.2.3 명사 추출

명사를 추출하는 단계에서 이용되는 방법은 정렬된 미등록어절에서 앞부분이 유사한 어절끼리 그룹을 만들고, 그룹 안에서 유사부분 뒤에 나오는 명사후문자열이 명사후문자열 사전에 등록되어 있는지 비교하여 설정한 임계값 이상으로 일치하면 명사로 추출한다.

정확률을 높이기 위해 음절 증가에 관련된 규칙 2가지, 임계값에 관한 규칙 1가지, 명사후문자열에 관한 규칙 2가지를 적용한다. 적용한 규칙은 다음과 같다.

1. 크기 1부터 N까지 차례로 유사어절 그룹화
2. 명사후문자열이 임계값 이상 일치하면 명사 추출
3. 명사가 추출된 어절은 다음 실험 대상에서 제외
4. 2개 이상 어절의 N+1음절이 명사 파생 접미사인 경우, 이를 포함한 명사후문자열 출현 빈도 합을 이용해 명사 추출 범위 결정
5. 모든 어절에 N+1음절이 같을 경우, 명사후문자열 출현 빈도를 이용해 명사 추출 범위 결정

‘정보’, ‘정보를’, ‘정보통신은’, ‘정보통신에서’, ‘정보통신에는’, ‘정보통신기술진흥센터’, ‘정보통신기술진흥센터의’

그림 2. 규칙 1,2,3 적용 어절

규칙 1은 1 음절부터 최장 길이의 N 음절까지 순차적으로 음절의 크기를 키워가며 유사어절 그룹화를 진행한다. 예를 들어, [그림 2]에서 ‘정보’, ‘정보통신’, ‘정보통신기술진흥센터’ 순으로 유사어절 그룹화가 진행된다.

규칙 2는 명사를 추출할 때 임계값은 공백을 포함한 2개 이상의 명사후문자열 일치로 정한다. 예를 들어, [그림 2]에서 ‘정보’, ‘정보를’라는 어절이 있으면 유사부분은 ‘정보’가 되고, 각각 명사후문자열은 공백과 ‘를’이 된다. 명사후문자열이 공백이 되는 것은 해당 어절이 명사 그 자체라고 판단하기 때문에, ‘를’이라는 명사후문자열이 명사후문자열 사전에 존재하기 때문에 ‘정보’는 명사로 추출된다. 2개 이상의 명사후문자열 일치로 임계값을 정한 이유는 제일 많은 명사가 인식되기 때문이다. 즉, 임계값을 높일 경우 정확률은 더욱 높아지지만, 재현율이 현저히 낮아진다.

규칙 3은 명사 추출에 사용된 어절은 실험 대상에서 제외하는 것이다. 예를 들어, [그림 2]에서 4음절 크기의 명사 ‘정보통신’ 추출에 사용된 어절 ‘정보통신에서’, ‘정보통신에는’을 제외시키지 않으면 ‘정보통신에’까지

명사로 추출하게 된다.

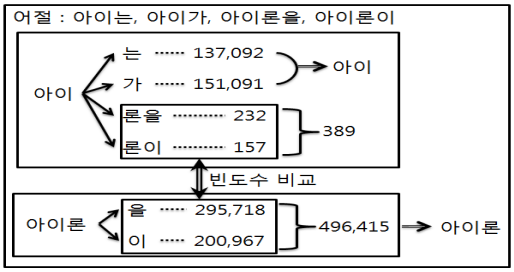

그림 3. 규칙 4 예제

규칙 4는 크기 N의 유사어절 그룹에서 2개 이상 어절의 N+1음절이 명사 파생 접미사인 경우에 명사 파생 접미사를 포함한 명사후문자열 출현 빈도 합과 명사 파생 접미사를 포함하지 않은 명사후문자열 출현 빈도 합을 비교하여 빈도수가 높은 쪽의 명사 부분을 미등록명사로 결정한다. 예를 들어, [그림 3]에서 2음절 크기의 유사어절 '아이'의 명사후문자열('는', '가', '론을', '론이')은 명사후문자열 사전에 모두 존재하기 때문에, 규칙 4를 적용하지 않으면 명사 '아이'는 추출이 되지만 '아이론'은 추출되지 못한다. '아이' 다음에 나오는 '론'이 명사 파생 접미사이기 때문에 규칙 4를 적용하면, '론을', '론이' 명사후문자열 출현 빈도와 '을', '이' 명사후문자열 출현 빈도를 비교하고, 출현 빈도가 높은 명사후문자열을 선택한다. 비교 결과 명사후문자열 '을', '이' 출현 빈도 합이 더 높기 때문에 3음절 크기의 유사어절 '아이론'을 미등록명사로 추출한다. 여기서 '아이론'은 머리를 펴거나 웨이브를 넣어 모양을 낼 때 쓰이는 전자제품으로 주로 '고데기'로 알려져 있다.

명사 파생 접미사의 종류는 '제', '상', '계', '형', '용', '론', '치', '자', '기', '질', '대', '성', '가', '당' 등이 있다. '들'의 경우에는 '해찬들'같은 상표나 '아들'같은 명사를 제외하고 대부분 '사람들'과 같이 명사의 복수형으로 사용되기 때문에 제외한다.

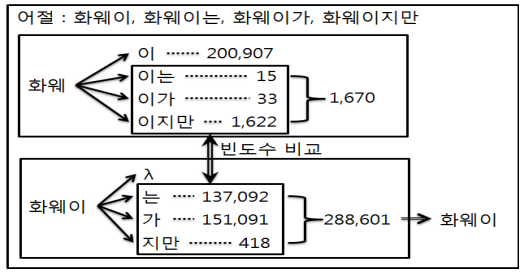

그림 4. 규칙 5 예제

규칙 5는 크기 N의 유사어절 그룹에서 모든 어절에 N+1음절이 같은 경우에 크기 N+1의 유사어절 그룹에서도 명사 추출이 가능할 때, 크기 N의 명사후문자열 출현 빈도 합과 크기 N+1의 명사후문자열 출현 빈도 합을 비교하여 합이 높은 단어를 선택한다. 예를 들어, [그림 4]에서 '화웨이'가 선택된다. 어절 '화웨이'의 명사후문자열 출현 빈도를 비교에 사용하지 않은 이유는 '화웨'의 명사후문자열 '이'는 존재하지만 '화웨이'의 명사후문자열은 존재하지 않기 때문이다.

3.2.4 미등록명사 결정

미등록명사 결정 단계는 명사 추출 단계에서 추출된 명사를 명사 사전과 비교하여 등록되어있지 않은 명사를 최종 미등록어로 결정하는 단계이다. 만일, 미등록명사로 추출된 단어가 '드론'과 '유저'인데, '유저'가 이미 명사 사전에 존재한다면, '드론'만을 미등록어로 결정한다.

IV. 실험 및 평가

제안하는 명사후문자열을 이용한 미등록어 인식의 성능을 평가하기 위해 포털사이트 Daum, Naver IT 관련 뉴스 기사를 추출하여 실험 데이터로 사용한다. 실험 데이터를 뉴스 기사로 선정한 이유는 다음과 같다. 뉴스 기사는 새로운 정보가 많이 올라오고, 오탈자가 거의 없다. 따라서, 오탈자를 고려하지 않고 미등록어 인식 성능을 평가하기 위해 뉴스 기사를 실험 데이터로 사용한다. 또, 한 주제와 관련된 어절이 많을수록 미등록어 인식이 수월해지고, 매일 비중 있는 주제가 달라

지기 때문에 포털사이트, 날짜별로 실험 데이터를 모아 실험을 진행한다. 실험은 포털사이트마다 5번씩 총 10번 진행하였다. 포탈사이트마다 날짜별로 뉴스기사를 추출하고 3장에서 설명한 미등록명사 인식기의 과정으로 미등록 명사를 추출한다. 시험 데이터로 사용된 데이터의 특성은 [표 2]와 같다. 각 단계를 진행했을 때 추출된 어절과 미등록어 개수는 [표 2]와 같고, 이 시험데이터를 이용한 실험 결과는 [표 3]과 같다.

제안한 미등록어 추출 방법의 성능 평가를 위해 다음과 같은 정확률 및 재현율 식을 이용하였다. 이중 재현율2는 본 논문에서 대상으로 하는 2형태 이상으로 출현한 미등록명사만을 대상으로 한 재현율을 의미한다.

표 2. 테스트 데이터

Portal	전체 단어	미등록 명사	2형태 이상 출현한 미등록어
Daum	20,365	1,332	221
Naver	31,070	1,988	334
합계	51,435	3,320	555

표 3. 시험 결과

2형태 이상 출현한 미등록명사	추출된 미등록명사	옳게 추출된 미등록명사
555	554	552

$$\text{정확률} = \frac{\text{올바르게 추출된 미등록명사수}}{\text{추출된 미등록명사수}} \quad (1)$$

$$\text{재현율1} = \frac{\text{올바르게 추출된 미등록명사수}}{\text{전체 미등록명사수}} \quad (2)$$

$$\text{재현율2} = \frac{\text{올바르게 추출된 미등록명사수}}{\text{2형태 이상으로 출현한 미등록명사수}} \quad (3)$$

위와 같은 수식으로 계산 하였을 경우 본 논문에서 제안한 시스템의 정확률과 재현율은 [표 4]와 같다.

표 4. 정확률 및 재현율

정확률	재현율2	재현율1
99.64%	99.46%	16.63%

본 논문은 문서에서 2가지 형태 이상으로 나타나는 미등록명사의 인식을 목표로 하고 있기 때문에 정확률1은 사실 별 의미가 없으나, 참고를 위해 기술하였다. 만일 형태소분석을 이용하거나 미등록명사를 포함하는 다른 형태의 단어를 인터넷에서 추가로 수집하는 등의 방법을 이용하면 재현율1은 다른 논문에서 제안한 정도의 성능으로 쉽게 향상시킬 수 있다.

제안한 시스템에서 미등록명사로 잘못 결정한 단어가 3개이다. 각 단어에 대해 살펴보면 다음과 같다.

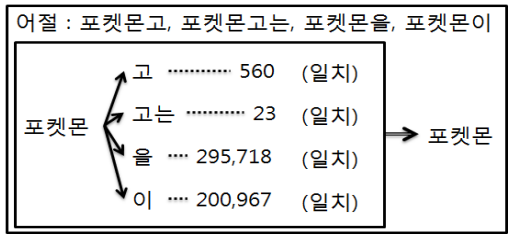

그림 5. 오류 추출 단어 1

옳게 추출되지 않은 첫 번째 단어는 '포켓몬고'였다. 이 단어가 추출되지 못한 이유는 [그림 5]와 같다. 3음절 크기의 유사어절 '포켓몬'에서 용언화접사+어미 '이고'의 축약형인 '고', 용언화접사+어미인 '고는', '을', '이'가 명사후문자열 사전에 모두 존재하여 '포켓몬'이 명사로 추출되고

모든 어절이 다음 실험대상에서 제외되기 때문에 '포켓몬고'는 추출하지 못한다. 하지만, 7월 18일 Naver와 Daum 뉴스 기사에서는 '포켓몬', '포켓몬고' 2가지 모두 추출되었는데, 이는 '포켓몬', '포켓몬고' 관련 유사어절이 충분했기 때문이다. 만약 음절 결합 정보를 이용한다면 '몬' 뒤에 '고'가 올 수 없기 때문에 '포켓몬고'가 올바르게 추출될 수 있겠지만, 본 논문에서는 명사후문자열만을 이용한 미등록명사 추출 방법을 제안하기 때문에 이런 방법을 적용하지 않았다.

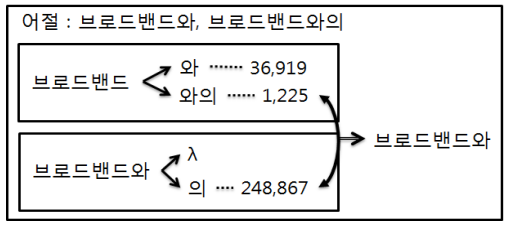

그림 6. 오류 추출 단어 2

시스템이 잘못 추출한 두 번째 단어는 '브로드밴드와'이다. 이 단어가 올바르게 추출되지 못한 이유는 [그림 6]과 같다. '브로드밴드' 명사후문자열 '와의' 출현 빈도 1,225보다 '브로드밴드와' 명사후문자열 '의' 출현 빈도 248,867가 더 높기 때문에 '브로드밴드와'가 추출된다. '브로드밴드'도 관련 유사어절이 충분하다면 올바르게 추출될 수 있다.

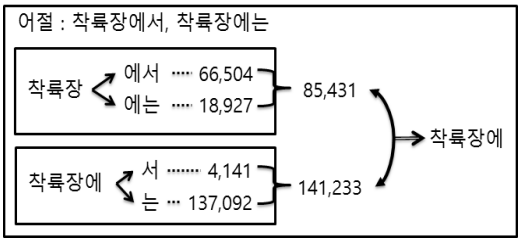

그림 7. 오류 추출 단어 3

잘못 추출된 세 번째 단어는 "착륙장에'이다. 이 단어가 올바르게 추출되지 못한 이유는 [그림 7]과 같고 두 번째 오류 추출 단어와 같다. 즉, '착륙장' 명사후문자열 '에서', '에는' 출현 빈도 합 85,431보다 '착륙장에' 명사후문자열 '는' 출현 빈도 137,092가 더 높기 때문에 '착륙장에'가 추출된다.

두 번째와 세 번째 오류 단어는 모두 규칙 5와 관련이 있는데, 전체적으로 규칙 5를 적용할 때의 성능이 더 좋았다. 또한 위 미등록명사가 들어 있는 어절의 수가 많다면 실험에서 오류로 추출된 단어 모두 올바르게 추출될 수 있다.

본 논문의 연구를 기존 세 가지 연구와 비교했을 때 정확률은 매우 높았다. 본 논문의 연구가 두 가지 이상의 형태로 출현한 미등록어만을 대상으로 하고 있으므로 재현율 비교는 적절하지 않지만, 참고를 위해 [표 4]에 기술하였다.

표 5. 성능 비교

	재현율	정확률
김선호	68.00%	93.00%
최맹식	8.77%	83.70$
박소영	93.52%	93.52%
제안하는 방법	16.63(99.46)%	99.64%

김선호[10]는 2번 이상 출현한 미등록어를 대상으로 68.00%의 재현율, 93.00%의 정확률을 보였다. 최맹식[14]은 어절 단위로 미등록어 인식을 하였는데 1번 출현한 미등록어까지 인식하여 재현율은 8.77%가 나왔고 정확률은 83.70%를 보였다. 박소영[12]은 세 가지 단계를 합한 미등록어 인식 성능은 93.52%의 재현율과 93.52% 정확률이 나왔지만, 2번 이상 출현한 단어가 대상이 되는 전문분석기반 미등록명사 인식 단계를 사용하였을 때는 97.01%의 정확률과 52.63%의 재현율을 보였다. 비교를 통해 1번 출현한 미등록어의 재현율이 낮은 점을 제외하고, 2번 출현한 미등록어 인식 성능은 타 논문보다 높게 나타난 것을 볼 수 있다.

V. 결론 및 향후 연구

본 논문에서는 명사후문자열을 이용하여 미등록명사를 인식하는 방법을 제안하였다. 미등록명사의 인식을 위해 어절 사전, 명사후문자열 사전, 명사 사전이 이용된다. 어절 사전은 세종 문어원시말뭉치로부터 수집한 어절들의 사전이며, 다른 두 사전은 세종 형태분석말뭉치에서 명사가 포함된 모든 어절들을, 명사와 명사 다음에 나오는 문자열로 이등분하여 각각 명사 사전과 명사후문자열 사전을 구축하였다.

본 논문에서는 미등록명사가 포함된 어절이 서로 다른 두가지 이상의 형태로 나타날 경우만을 대상으로 미등록명사를 인식한다. 미등록명사를 인식하기 위해 먼저 대상 문서에 출현하는 모든 어절들을 추출하여 어절 사전과 비교하여, 어절 사전에 존재하지 않는 어절들을 미등록어 후보로 선정한다. 이 미등록어 후보들을 정렬하여 앞부분이 같은 단어들을 그룹으로 만들고, 동일한 앞부분을 명사로 그 뒷부분을 명사후문자열로 추정하고 명사후문자열 사전을 이용하여 미등록명사 후보를 추출한다. 그리고 이 미등록명사 후보를 명사 사전과 비교하여 최종 미등록명사를 결정한다.

포털사이트의 인터넷 뉴스를 이용하여 실험한 결과, 정확률은 99.64%, 두 가지 이상의 형태로 출현한 미등록어에 대한 재현율은 약 99.46%의 높은 성능을 보였

다. 반면, 한 가지 형태로 출현한 미등록어 인식에 대한 재현율은 상당히 낮았는데, 이것은 이러한 경우의 미등록어 인식을 논문 범위에서 제외했기 때문이다. 만일 기존 연구들처럼 형태소분석을 이용하거나, 인터넷에서 동일 미등록명사를 포함하는 단어들을 추가로 수집하여 인식한다면, 문서에 포함된 모든 미등록어들에 대한 인식을 대상 범위로 하더라도 높은 재현율을 보일 수 있을 것으로 예상된다.

참 고 문 헌

[1] Aastha Gupta, Rachna Rajput, Richa Gupta, and Monika Arora, “Improved POS Tagging for Unknown Words,” International Journal of Soft Computing and Engineering (IJSCE), Vol.4, Issue.ICCIN-2K14, pp.2231-2307, 2014(3).

[2] 김형철, 서형원, 김재훈, “접사 정보를 이용한 영어 미등록어의 품사부착 성능개선,” 한국마린엔지니어링학회 공동학술대회 논문집, pp.375-376, 2009.

[3] Giorgos S. Orphanos and Dimitris N. Christodoulalds, “POS Disambiguation and Unknown Word Guessing with Decision Trees,” Proceedings of EACL '99, 1999.

[4] 강승식, *음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석*, 서울대학교 컴퓨터공학과, 박사학위논문, 1993.

[5] 박봉래, 황영숙, 임해창, “유사어절의 TAIL 패턴 분석에 기반한 미등록 명사 추정,” 한국정보과학회 봄 학술발표논문집, pp.907-910, 1996.

[6] 양장모, 김민정, 권혁철, “언어정보를 이용한 한국어 미등록어 추정,” 한국정보과학회 봄 학술발표 논문집, pp.957-960, 1996.

[7] 박봉래, 황영숙, 임해창, “확장 정의된 유사어절의 분석에 근거한 실시간 미등록어 인식,” 한글 및 한국어 정보 처리 학술 발표 논문집, pp.222-228, 1996.

[8] 차정원, 이원일, 이근배, 이종혁, “미등록어 추정을 이용한 TAKTAG의 개선,” 제8회 한글 및 한국어 정보처리 학술대회, pp.203-206, 1996.

[9] 박봉래, 황영숙, 임해창, “용례 분석에 기반한 미등록어의 인식,” 정보과학회논문지(B), 제25권, 제2호, pp.397-407, 1998.

[10] 김선호, 윤준태, 송만석, “한국어 문서 처리를 위한 동적 생성 로컬 사전 기반 미등록어 분석,” 정보과학회논문지 : 소프트웨어 및 응용, 제29권, 제6호, pp.407-416, 2002.

[11] 이도길, 이상주, 임해창, “명사 출현 특성을 이용한 효율적인 한국어 명사 추출 방법,” 정보과학회논문지 : 소프트웨어 및 응용, 제30권, 제2호, pp.173-183, 2003.

[12] 박소영, “웹문서를 이용한 단계별 한국어 미등록어 인식 모델,” 한국해양정보통신학회논문지, 제13권, 제9호, pp.1898-1904, 2009.

[13] 박용현, 황재원, 고영중, “한국어 명사 출현 특성과 후절어를 이용한 명사 추출기,” 정보과학회논문지: 소프트웨어 및 응용, 제37권, 제12호, pp.919-927, 2010.

[14] 최맹식, 김학수, “기계학습에 기반한 한국어 미등록 형태소 인식 및 품사 태깅,” 정보과학회논문지(B), 제18권, 제1호, pp.45-49, 2011.

[15] 김보겸, 이재성, “확률 기반 미등록 단어 분리 및 태깅,” 정보과학회논문지, 제43권, 제4호, pp.430-436, 2016.

[16] https://ithub.korean.go.kr/

저 자 소 개

박 기 탁(Ki-Tak Park) 준회원

- 2016년 2월 : 충북대학교 컴퓨터공학과 졸업(학사)
- 2016년 3월 ~ 현재 : 충북대학교 컴퓨터공학과(석사 과정)

<관심분야> : 자연언어처리, 정보검색

서 영 훈(Young-Hoon Seo) 종신회원

- 1983년 : 서울대학교 컴퓨터공학과 졸업(학사)
- 1985년 : 서울대학교 컴퓨터공학과 졸업(석사)
- 1991년 : 서울대학교 컴퓨터공학과 졸업(박사)
- 1994년 ~ 1995년 : 미국 Carnegie Mellon 대학 기계번역센터 객원교수
- 1988년 ~ 현재 : 충북대학교 전자정보대학 컴퓨터공학과(교수)

<관심분야> : 자연언어처리, 한영기계번역, 정보검색, 질의응답시스템