데이터 기반 형태소 분석기의 신규명사 추출 경향 분석

김노은[01], 정상근[1]*

[1]충남대학교

noeunkim511@gmail.com, hugmanskj@gmail.com

New Nouns Extraction Analysis of Data-based Korean Morphological Analyzer

Noeun Kim[01], Sangkeun Jung[1]

[1]School of Computer Science and Engineering, Chungnam National University

요 약

데이터 기반 한국어 형태소 분석기 5 개에 대한 신규명사 추출 경향 분석을 위하여 2000 년도부터 2019 년도까지 각 년도를 대표하는 신조어 포함 문장을 수집해 분석을 시행하고, 각 분석기의 명사 추출 결과를 비교한다. 대표 문장의 추출에는 당대 트렌드를 반영하는 미디어 차트 키워드를 활용하였다. 양적분석을 통해 데이터 기반 형태소 분석 기법이 신규 발생한 명사들에 대해 꾸준한 성능의 하락세를 나타낸다는 사실을 밝히고 새로운 분석 방향 도입의 필요성을 제시한다.

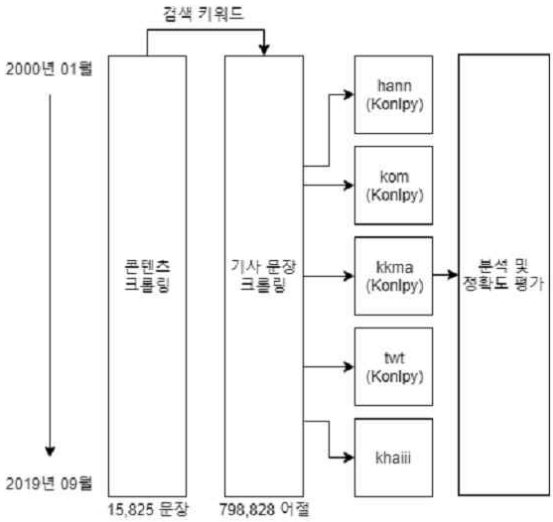
1. 서 론

신규 어휘는 당대 시대상을 투영한다. 특정 시기에 정립된 개념을 설명하기 위해 기존에 존재하지 않던 새로운 단어가 형성되는가 하면, 대체어의 탄생으로 퇴화하는 단어들 역시 존재하기 때문이다. 이것은 문자 어휘가 언어학적인 요소보다 문화와 사회적인 환경 요인의 영향을 크게 받으며 변화하는 것을 의미한다. 키워드 색인이 수행된 후에 텍스트 간 관계 분석이 가능하기 때에 변동성이 큰 명사의 정확한 추출은 데이터 마이닝 및 데이터 과학 분야의 키워드 도출 에 있어 중요한 역할을 한다.

한글 텍스트 데이터 내 명사 품사를 분류하고자 사용하는 분석도구가 '형태소 분석기'이다. 자연어처리 작업 중 가장 기반이 되는 단계이며 입력받은 한국어 텍스트 데이터를 한글의 최소 단위인 형태소로 구분한 결과값을 도출한다. 그러나 현존하는 대부분의 데이터 기반 형태소 분석기는 2007 년의 세종 1차 말뭉치에 의존하여 작동하기 때문에 최신 구어, 신규명사 등의 신규 어휘를 효과적으로 검출하지 못한다는 한계가 있다.

이에 본 연구는 과거 자원에 의존하며 수동적인 업데이트를 수반하는 현존 데이터 기반 형태소 분석기의 한계를 밝힘으로써 새로운 분석 방향 도입의 필요성을 제시하고자 한다.

표 1. 전체 분석 과정 다이어그램



2. 신규명사 분석을 위한 말뭉치

2.1 데이터 수집

시대별 신규명사 추출 양상을 확인하기 위해서는 미등록 신조어를 포함할 것으로 예상되는 문장을 무작위로 수집하여 형태소분석을 시행해야 한다. 그러나 신조어를 규정하는 판단 기준은 개인에 따라 상이하다는 한계가 존재하기 때문에 신규명사를 포함한 대표 문장 수집을 위해 시대상 반영 키워드 수집이 먼저 선행되어야 한다. 키워드 수집과 이를 통한 기사 문장 수집의 두 단계를 거친다면 수집된 문장을 시대상 반영과 신규명사 포함의 두 가지 조건을 모두 만족하는 대표 문장으로서 규정할 수 있다.

* 교신저자(Corresponding Author)

표 2. 분석기 별 분석 대상 및 해당하는 태그 범주 표

	kkma	kom	hann	khaiii	twt
Description	**Tag**				
보통명사	NNG	NNG	NC	NNG	
고유명사	NNP	NNP	NQ	NNP	
일반 의존 명사	NNB	NNB	NB	NNB	
단위 의존 명사		NNM		NP	
수사	NR	NR	NN	NP	
대명사	NP	NP	NP	NR	
선어말 어미				EP	Noun
종결 어미				EF	
연결 어미				EC	
명사형 전성 어미				ETN	
관형형 전성 어미				ETM	

분석 이전 문장 수집 단계에서 어휘 변화 양상을 담고 있는 당대 대표 문장을 수집하고자 당대 이슈 키워드를 활용해 기사 수집 입력 데이터를 사전에 구축하는 과정은 다음과 같다.

(1) 2000.01-2019.09 간 멜론 순위 15,825 개 수집
(2) 문장 검색 키워드 15,825 개 정의
(3) 키워드 별 약 10 개, 총 131,944 개 기사 수집
(4) 외국어 필터링 후 798,828 개 어절 수집

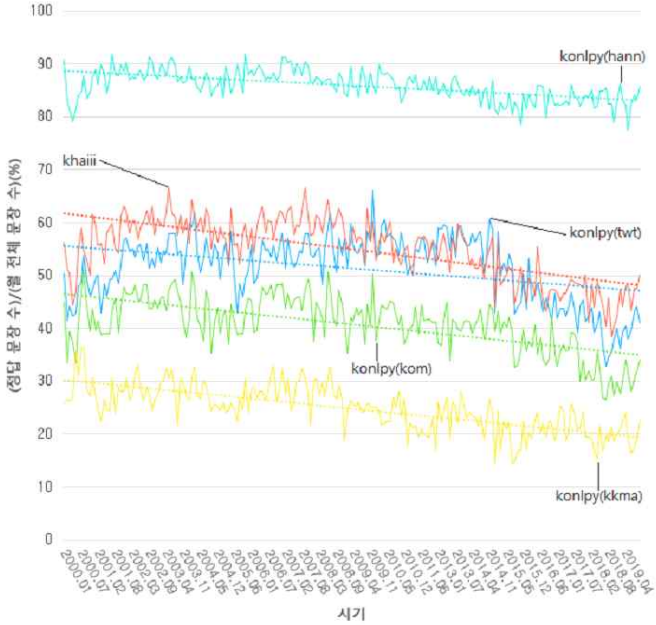
키워드 자원을 입력값으로써 약 20여년 간의 기사 제목 말뭉치를 수집 하였다. 문장의 종류와 신규어휘의 다양성을 보장하고자 언론사의 제한은 두지 않았으며, 기사 제목 말뭉치 검색 시기 설정은 키워드 수집 년/월과 일치하도록 하였다.

2.2 신규명사 포함 문장 형태소 분석

분석에 사용한 형태소 분석기 및 라이브러리는 [1]twt, [2]hann, [3]kom, [4]kkma, [5]Khaiii의 다섯 종류이다. 데이터 기반 형태소 분석기 Khaiii와 Konlpy는 각각 한국 법률 말뭉치(kolaw) 사전 및 대한민국 국회 의안 말뭉치(kobill), KAIST 말뭉치를 이용해 생성된 Hannanum 시스템 사전, 세종 말뭉치를 이용해 생성된 Kkma 시스템 사전, Twitter 분석기 사전을 기반으로 동작한다.

각 분석 도구에 따라 같은 한글 품사를 지칭하는 태그 명칭이 모두 상이하다. 표2 에 분석 대상 태그의 범주를 정의하였다. 직접 수집한 말뭉치자원을 분석하여 신규명사로 추정되는 단어를 추출하는 과정은 다음과 같다.

표 3. 분석기 별 신규명사 추출 정확도 비교 시계열 그래프



체언어절 수집 : 품사 구분을 통해 신규명사를 추출하는 것이 본 연구의 목적이기 때문에 분석 결과 체언 태그(Noun)로 구분된 어절 수집' 단계에서 불필요한 '동사', '형용사', '어근', '접미사' 등의 항목을 제외시킨다.

오류 및 예외처리 : 분석기가 사용하는 말뭉치 종류에 따라 체언을 구분하는 세부 태그명세가 다르기 때문에 오류 및 특수문자를 포함한 예외 처리' 단계에서 탐지를 원하는 태그를 각기 다르게 처리해 주어야 한다. 가장 상위 태그인 'N' 값을 이용해 단순분류하면 신규명사 추출과 무관한 종결어미, 수사, 조사 등이 포함되어 탐지되는 오류가 발생할 위험이 있다.

신규명사 분석성능 판별: 주요 시대상을 반영한 키워드를 입력값으로 설정하여 수집한 문장에 신조어가 포함되었다고 가정한다. 해당 신조어가 데이터 기반 형태서 분석기를 통해 신규명사로 정확히 분류 되었는가 판별하기 위해 다음의 평가 과정을 적용한다.

원문 텍스트와 비교하여 다섯 개 분석기 별 체언으로 인식된 어절이 형태 변화가 생긴 경우에 패널티 점수를 부과한다. 점수 부과 기준은 다음과 같다. (1)누락된 문자가 있거나 (2)명사의 분해가 발생한 어절은 신규명사 추출에 실패한 문장으로 인식한다. 또한, 분류 오류로 인해 태깅 이후에도 (3)중복 출현한 단어에 대해서도 실패값을 부여한다.

표 4. 분석기 별 신규명사 추출 정확도 감소율

	2006년	20013년	20019년	누적 감소율
hann	87.5	85.36	84.42	3.52%
khaiii	59.09	52.91	47.37	19.83%
twt	49.32	48.90	36.16	26.68%
kom	39.09	37.10	27.67	29.21%
kkma	24.32	23.96	20.89	14.10%

(누적 감소율) = (2006 감소율 - 2019 감소율)/(2006 감소율)

전체 텍스트 데이터 처리를 마친 뒤 (1)누락, (2)분해, (3)중복 패널티가 부여되지 않은 문장은 총점 0 점으로 신규명사 추출에 성공한 것으로 인지한다. 최종적으로 시간축 데이터의 흐름에 따라서 신규명사 등장 빈도를 통계화 하기 위해 기간 별 신규명사 추출 성공 횟수를 누적하여 기록한다.

각 분석기는 보유하고 있는 말뭉치의 한계로 인해 신조어를 포함한 문장을 분석한 결과, 신규명사의 어절 전체 또는 일부가 누락되거나 어미가 변형된 형태를 띄기 때문에 이와 같은 접근 방법을 사용한다.

3. 신규명사 분포 양상 분석

앞절에서 제시한 방법에 따라 결과 분석을 수행한다. 신규명사 추출 성능의 한계를 밝히고자 현존 분석기 간 신규명사 분류 정확도를 통계화 하는 것이 본 연구의 궁극적인 목적이기 때문에 분석 속도 및 기타 품사 분류 정확도 등의 요인을 제외하고 시간(DATE; X) 축과 추출 성공/실패(SCORE; Y)축의 두 가지 변수를 이용해 시계열 그래프를 그려 모델을 단순화 한다.

결과는 **표 3**과 같다. 시간(x)축의 시점 t와 t+1 간 간격은 한 달(month)을 주기로 인덱싱 하여 204 개 행으로 압축 하였으며, y축 데이터는 주기 별 전체 문장 중 정답을 맞춘 문장의 수를 백분율 값으로 나타낸 것이다. 분석기 별 탐지 분류 태그를 다르게 설계하였기 때문에 같은 문장에 대한 판별 점수가 각기 다르게 나타난다.

표 4에서 각 분석기의 성능 순위와 연간 감소율을 확인할 수 있다. 지속적으로 높은 정답 확률을 보인 분석기는 hann(Konlpy)이며, 최하위 성능을 기록한 분석기는 kmma(Konlpy)이다. 가장 가파른 감소 추세선을 보인 kom(Konlpy)의 감소율은 29.21%이며, 반대로 가장 적은 감소율을 보인 hann(Konlpy)의 감소율은 3.52%이다.

4. 결론

최근 데이터 기반 형태소 분석기는 최신 말뭉치 라이브러리를 업데이트 하는 수동적인 대처를 통해 신규명사의 지속적인 출현에 대응하고 있다. 그러나 데이터 기반의 형태소 분석기는 훈련의 대상이 되는 말뭉치에 의존적일 수 밖에 없기 때문에, 업데이트 없이는 신규 발생한 명사들에 대해서 분석의 한계를 가지고 있다.

본 연구 결과로 말미암아 한국어 자연어처리 성능을 높이고자 최신 말뭉치로 설계된 형태소 분석기를 적용하는 현안만으로는 지속적으로 형성되는 신조어에 즉각 대응할 수 없다는 한계를 증명 하였다. 정보화 속도가 점차 가속될수록 데이터 마이닝의 기본 도구가 되는 형태소 분석기의 명사 추출의 한계는 더욱 커질 것이다. 이에 신규 명사에 효율적이고 관리 용이한 분석 방법을 고민해야 할 것이다.

감사의 글

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2019-0-0004, 준지도학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발)

이 논문은 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2019R1F1A1060601)

참고 문헌

[1] 이동주, 연종흠, 황인범, 이상구, 꼬꼬마: 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구, 2010, 정보과학회논문지: 컴퓨팅의 실제 및 레터, Volume 16, No.11.

1) konlpy.org/en/v0.4.4/api/konlpy.tag/#module-konlpy.tag._twitter
2) konlpy.org/en/v0.4.4/api/konlpy.tag/#module-konlpy.tag._hannanum
3) konlpy.org/en/v0.4.4/api/konlpy.tag/#module-konlpy.tag._komoran
4) konlpy.org/en/v0.4.4/api/konlpy.tag/#module-konlpy.tag._kkma
5) https://github.com/kakao/khaiii