한국어 뉴스 분석 성능 향상을 위한 번역 전처리 기법

이지민, 정다운, 구영현, *유성준

세종대학교

leejeemin0608@gmail.com, chris410@naver.com, yhgu@sejong.ac.kr,
*sjyoo@sejong.ac.kr

Translation Pre-processing Technique for Improving Analysis Performance of Korean News

Ji-Min Lee, Da-Woon Jeong, Yeong-Hyeon Gu, *Seong-Joon Yoo

Computer Science and Engineering, Sejong University

요 약

한국어는 교착어로 1개 이상의 형태소가 단어를 이루고 있기 때문에 텍스트 분석 시 형태소를 분리하는 작업이 필요하다. 자연어를 처리하는 대부분의 알고리즘은 영미권에서 만들어졌고 영어는 굴절어로 특정 경우를 제외하고 일반적으로 하나의 형태소가 단어를 구성하는 구조이다. 그리고 영문은 주로 띄어쓰기 위주로 토큰화가 진행되기 때문에 텍스트 분석이 한국어에 비해 복잡함이 떨어지는 편이다. 이러한 이유들로 인해 한국어 텍스트 분석은 영문 텍스트 분석에 비해 한계점이 있다고 알려져 있다. 한국어 텍스트 분석의 성능 향상을 위해 본 논문에서는 번역 전처리 기법을 제안한다. 번역 전처리 기법이란 원본인 한국어 텍스트를 영문으로 번역하고 전처리를 거친 뒤 분석된 결과를 재번역하는 것이다. 본 논문에서는 한국어 뉴스 기사 데이터와 번역 전처리 기법이 적용된 영문 뉴스 텍스트 데이터를 사용했다. 그리고 주제어 역할을 하는 키워드를 단어 간의 유사도를 계산하는 알고리즘인 Word2Vec(Word to Vector)을 통해 유사 단어를 추출했다. 이렇게 도출된 유사 단어를 텍스트 분석 전문가 대상으로 성능 비교 투표를 진행했을 때, 한국어 뉴스보다 번역 전처리 기법이 적용된 영문 뉴스가 약 3배의 득표 차이로 의미있는 결과를 도출했다.

키워드: Lexrank, Word2Vec, 뉴스 데이터 분석, 번역 전처리 기법, 빈도수 분석

1. 서론

자연어는 컴퓨터에 최적화된 언어가 아닌 사람들이 쓰는 일상적인 언어이다. 텍스트 분석을 위해 이러한 자연어를 컴퓨터가 이해할 수 있는 인공어로 처리하는 과정은 필수적이다.

한국어는 교착어로 1개 이상의 형태소들이 결합되어 단어를 이루고 있기 때문에 텍스트 분석 시 완벽하게 형태소를 분리하는 것은 어렵다. 뿐만 아니라, 자연어를 처리하는 알고리즘은 대부분 외국에서 만들어졌기 때문에 한국어에 바로 적용하는 것은 무리가 있다. 영문은 She's와 같은 be 동사의 줄임 말이 단어에 붙는 경우를 제외하면 주로 띄어쓰기를 위주로 토큰화하기

*교신 저자

때문에 한국어에 비하면 자연어 처리가 용이한 편이다[1]. 뉴스의 자연어 처리에 관한 연구에서 이러한 한국어 분석에 대한 성능 문제와 적절한 분석이 이루어지는지에 대해 문제점을 제기했다[2].

뉴스 데이터에 대한 의미분석, 감성분석 등의 텍스트 분석 이 활발해지면서 비정형 데이터인 뉴스 데이터에 대한 높은 수준의 자연어 처리 또한 필요하다. 본 논문에서는 한국어보다 영문에서 전처리 성능이 더 좋다고 알려진 점[3]에서 착안해 한국어 뉴스 데이터를 영문으로 번역한 후 전처리를 진행했다. 그리고 키워드를 추출한 뒤 Word2Vec을 통해 도출된 결과를 한국어로 재번역시켰다. 본 논문에서는 이러한 기법을 번역 전처리 기법이라고 명명했다.

본 논문에서 사용하는 데이터는 2013년 7월부터 2018년 6월 사이에 작성된 '싱크홀 원인'에 관한 총 394개의 기사와 2008년 1월부터 2020년 1월 1일까지의 '당뇨병 증상'에 관한 총 211개의 기사이다. 비교군은 '싱크홀 원인'과 '당뇨병 증상'에 관한 한국어 기사와 번역 전처리 기법이 적용된 기사이다. 한국어 기사는 KoNLPy(Korean Natural Language Processing in Python)[4]를 이용해 수집한 기사에 대해 토큰화를 거쳐 명사를 추출한 후 불용어를 제거했다. 영문 기사는 한국어 전처리와 동일한 과정으로 전처리를 진행했고 이 과정에서 NLTK(Natural Language Toolkit)[5]를 이용했다.

전처리된 결과를 Lexrank[6]와 빈도수 분석에 적용해 추출된 공통 키워드를 Word2Vec에 적용했다. Word2Vec을 통해 출력된 유사 단어 결과를 텍스트 분석 전문가를 대상으로 성능 비교 투표를 진행했다.

2. 관련 연구

2.1 뉴스 데이터 분석

모바일 인터넷 이용률이 증가하면서 뉴스 데이터를 분석하는 연구가 증가하고 있다. 연구 [7]에서는 뉴스 데이터 분석을 통해 교통 사고에 대한 키워드를 추출하고 교통 사고 빈도, 사망자, 부상자 수 등의 통계를 예측했다. 정보 추출을 위해 뉴스 데이터에서 토큰화를 진행한 후 날짜, 차량 번호는 정규 표현식을 통해서, 사망, 부상에 대한 정보는 의미역 결정(Semantic Role Labeling)을 통해 추출했다. 연구 [8]에서는 뉴스에 대한 어휘 기반 감성분석과 선호도분석을 진행했다. 단어 빈도-역문서 빈도(Term Frequency - Inverse Document Frequency, TF-IDF)를 통해 중요한 키워드를 추출한 후 WordNet 사전을 사용해 감성 점수를 계산했다. 이를 통해 비즈니스, 스포츠 관련 뉴스는 긍정적인 기사가 많고 연예, 기술 관련 뉴스는 부정적인 기사가 많다는 결론을 도출했다.

2.2 키워드 추출에 관한 연구

키워드는 문서에서 중심이 되는 단어이자 문서의 내용을 압축해 보여줄 수 있는 요약어이다. 키워드 추출이란 문서의 전반적인 내용을 요약하는 것으로 의미분석(Semantic Analysis)과 같은 텍스트 분석에 용이하게 쓰인다. NLTK는 영문 자연어를 처리하는 패키지로 의미분석, 토큰화(Tokenization), 품사태깅(POS tagging) 등의 여러 기능을 제공한다.

연구 [9]에서는 소설 분석을 위해 NLTK와 제안한 의미 분석 모델을 이용해 중요한 키워드를 추출했다. 해당 모델은 품사 태깅을 기반으로 의미 관계를 구하고 의문사에 해당하는 단어가 포함된 문장을 중요한 문장으로 간주한 후 키워드 추출을 진행한다. 이런 방식으로 제안된 모델과 빈도수 분석, 위치 가중치 기반 알고리즘과 정확도를 비교해 보았을 때 제안된 모델이 가장 높은 정확도를 보였다.

연구 [10]에서는 Lexrank 알고리즘을 사용한 키워드를 추출하는 연구를 진행했다. Lexrank는 문장들 간의 유사도를 이용해 문서를 요약하는 알고리즘이다. 또한 문장의 유사도가 임계치를 넘어가면 랭크 값이 낮은 문장을 삭제해 중복된 부분이 결과에 포함되는 것을 최대한 방지한다[11,12]. Amazon, Trip Advisor 등의 사이트에서 수집한 리뷰 데이터를 사용해 실험을 진행한 결과, Lexrank 알고리즘은 비록 속도가 느리지만 우수한 성능을 보인다는 결과가 도출됐다.

키워드 추출에 자주 사용되는 알고리즘에는 TF-IDF와 Word2Vec이 있다. TF-IDF는 빈도수를 이용해 문서 내 특정 단어의 중요도를 나타내는 알고리즘이고 Word2Vec은 특정 단어와의 유사도가 높은 단어들을 추출하는 알고리즘으로 모두 단어를 벡터화 한다는 공통점이 있다.

TF-IDF는 특정한 단어가 등장한 문서의 빈도 수인 TF와 문서에서 등장한 단어 빈도를 나타내는 값인 DF의 역수를 이용해 계산한다. 모든 문서에서 자주 등장하는 단어는 중요도가 낮아지고 특정한 문서에서 많이 나오는 단어는 중요도가 높게 측정이 되는 특징이 있다.

연구 [13]에서는 TF-IDF를 이용해 문서 내 단어 간의 연관성을 측정했다. 연구에서 언급한 TF-IDF는 2가지 단점이 있다. 첫 번째로 단어의 시제가 달라지면 같은 원형을 지니더라도 다르게 구별된다. 그리고 두 번째는 단어와 문서의 빈도수를 이용해 계산하는 방식이기 때문에 단어 간의 관계나 의미를 분석하지 못한다는 것이다[14].

Word2Vec은 단어의 의미를 보존해 벡터화하는 알고리즘으로 TF-IDF의 단점을 보완할 수 있다. 연구 [15]에서는 의미분석에 대한 TF-IDF의 한계점을 개선하기 위해 TF-IDF와 Word2Vec을 결합한 벡터 모델을 제안했다. 이 연구에서는 정밀도와 재현율을

사용해 TF-IDF와 결합 벡터의 성능을 비교했다. 비교한 결과, 제안한 결합 벡터 모델이 TF-IDF보다 약 20% 더 높은 성능을 기록했다.

Word2Vec은 CBOW(Continuous Bag of Words Model)와 Skip-Gram 두 가지 방식으로 나뉜다. CBOW는 전체적인 맥락에서 단어를 예상하고 반대로 Skip-Gram은 단어로부터 주변 단어를 유추하는 방식이다. 연구 [16]에서는 의학 정보를 이용해 Word2Vec 모델을 평가했다. 의학 정보에 대한 텍스트 데이터를 기반으로 CBOW와 Skip-Gram 방식을 이용해 정확도를 비교했다. 그 결과, Skip-Gram이 CBOW보다 약 27% 더 성능이 좋다고 평가했다.

3. 번역 전처리 기법의 설계 및 구현

본 논문에서 진행한 뉴스 데이터 분석의 흐름도는 그림 1과 같다.'싱크홀 원인'과'당뇨병 증상'에 관한 한국어 뉴스 기사를 수집한 후, 한국어 뉴스에 번역 전처리 기법이 적용시켜 비교군으로 한국어 뉴스와 번역된 영문 뉴스로 만들어준다. 그리고 기사에서 명사를 추출하고 불용어를 제거한다. 이렇게 추출한 키워드를 Word2Vec 알고리즘에 적용시켜 출력된 유사 단어에 대해 성능 비교를 진행했다.

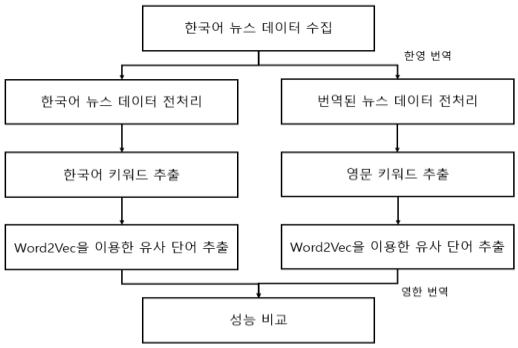

그림 1. 뉴스 데이터 분석의 흐름도

3.1 한국어 뉴스 데이터 수집

뉴스 데이터 분석을 위해'싱크홀 원인'과'당뇨병 증상'에 관한 한국어 뉴스를 수집했다.'싱크홀 원인'에 대해 2013년 7월부터 2018년 6월 까지의 총 394건의 기사를 수집했고 그 중에서 가장 긴 기사의 길이는 912자, 평균 길이는 110자이다. '당뇨병 증상'에 대해 수집한 뉴스 데이터는 2008년 1월부터 2020년 1월 1일까지 총 211개의 기사로 그 중에서 가장 긴 기사의 길이는 1,566자, 평균 길이는 255자이다.

3.2 번역 전처리

번역 전처리 기법이란 한국어 텍스트를 영문으로 번역을 시킨 후, 토큰화, 품사태깅, 명사 추출 등 과정을 거치고 그 결과를 한국어로 재번역을 시키는 것이다.

비정형 데이터인 뉴스 데이터 분석을 진행할 때 수집한 한국어 뉴스 데이터의 전처리를 위해 KoNLPy를 사용했다. 토큰화, 품사 태깅, 명사 추출, 불용어 제거 등 전처리를 진행했다. 번역 전처리 기법이 적용된 데이터는 영문이기 때문에 NLTK 라이브러리를 사용해 같은 전처리 과정을 진행했다.

3.3 키워드 추출

유사 단어를 도출하기 위해 '싱크홀 원인', '당뇨병 증상'과 관련된 키워드를 추출했다. 키워드 추출을 위해 사용한 방법은 2가지이다. 첫 번째는 Lexrank를 이용해 요약된 문서에 토큰화를 거쳐 명사를 추출했다. 두 번째는 단어의 빈도수를 계산했다. '싱크홀 원인'의 기사 원문에 '원인'이라는 단어가 나올 때 마다 해당 단어 뒤에 이어지는 100 글자를 자른 후 명사를 추출해 내림차순으로 정렬했다. 요약된 문서에서 추출한 명사와 빈도수 분석을 이용해 추출한 명사에서 서로 공통되는 단어 중 빈도수가 가장 높은 세 단어를 키워드로 간주했다. '당뇨병 증상'에 관해서도 똑같이 진행했다. 추출한 세 가지의 키워드는 각각 표 2, 3에서의 키워드 항목과 같다.

3.4 Word2Vec을 이용한 유사 단어 추출

한국어와 번역 전처리 기법이 적용된 결과를 분석하기 위해 Word2Vec 모델에서 Skip-Gram 방식을 사용했다. 표 1은 사용한 Word2Vec 모델의 파라미터 값에 대한 표이다.

표 1. Word2Vec 모델 파라미터

	차원	주변 단어	단어 빈도	학습 횟수
파라미터	100	5	2	300

그림 2는 키워드 추출에서 선정한 단어 중 빈도수가 가장 높은 단어를 Word2Vec에 적용한 결과 유사도 상위 7개의 단어에 대한 그림이다. '싱크홀 원인'에서 선정한 키워드는 '하수'이고 '당뇨병 증상'에서 선정한 키워드는 '증상'이다.

	싱크홀 원인 '하수'				당뇨병 증상 '증상'			
	한국어 전처리		번역 전처리		한국어 전처리		번역 전처리	
	단어	유사도	단어	유사도	단어	유사도	단어	유사도
0	노후	0.70	하수도	0.58	당뇨병	0.68	몸무게	0.45
1	불량	0.56	관	0.50	실신	0.47	과식증	0.45
2	박스	0.56	인내	0.49	자각	0.47	목마름	0.42
3	접합	0.55	수돗물	0.47	각막	0.45	무력감	0.40
4	부설	0.50	기대	0.47	친척	0.45	땀	0.39
5	관내	0.50	하수관	0.45	혈당	0.42	거식증	0.39
6	오염원	0.49	오염	0.43	고혈압증	0.42	발한	0.38

그림 2. 한국어, 번역 전처리 기법 적용 기사 Word2Vec 결과

3.5 성능 비교

제안한 번역 전처리 기법의 성능을 측정하기 위해 실험을 진행했다. 수집한 한국어 뉴스 기사 데이터와 그에 번역 전처리 기법을 적용한 영문 데이터에 같은 전처리 과정을 진행한 후 키워드를 추출했다. 추출한 3개의 키워드를 Word2Vec에 적용한 결과는 표 2, 3과 같다. 표 2, 3은 싱크홀의 원인과 당뇨병의 증상에 대해 키워드 추출 시 선정한 키워드를 Word2Vec 모델에 적용시켜 출력된 유사도가 높은 5개의 단어에 대한 결과이다.

표 2. 싱크홀 원인에 관한 한국어 전처리, 번역 전처리 결과 비교

키워드	한국어 전처리	번역 전처리
하수	노후, 불량, 박스, 접합, 부설	하수도, 관, 대체, 수돗물, 관로
공사	원인, 터널, 싱크홀, 롯데, 강제	관, 사이트, 지하철, 지반, 원인
누수	오수, 도래, 안양시, 송진, 수돗물	지하수, 감소, 수돗물, 인내, (물의) 흐름

표 3. 당뇨병 증상에 관한 한국어 전처리, 번역 전처리 결과 비교

키워드	한국어 전처리	번역 전처리
증상	당뇨병, 실신, 자각, 각막, 친척	몸무게, 과식증, 무력감, 목마름, 땀
당뇨	당뇨병, 환자, 생기지, 치료, 합병증	연관성, 목소리, 사탕, 진단서, 광기
혈당	당뇨병, 인슐린, 용량, 저녁 식사, 경구	피, 인슐린, 기관, 포도당, 수치

제안한 번역 전처리 기법의 성능을 측정하기 위해 표 2, 3에 대해 텍스트 분석 전문가 9명을 대상으로 투표를 진행했고 투표 결과는 그림 3과 같다. '싱크홀 원인'에서 번역 전처리 기법이 적용된 결과가 한국어에 비해서 약 3배의 득표수가 나왔다. 추출된 키워드인 '하수', '공사', '누수'에 대해 '하수'는 한국어 전처리가 더 높은 표를 얻었지만 '공사', '누수'는 번역 전처리의 득표수가 더 높았다. '당뇨병 증상'에서는 약 3.12배의 득표수를 기록했고 '증상', '당뇨', '혈당'에 대해 세 키워드 모두 번역 전처리의 득표수가 높았다.

본 논문의 한계점은 크게 두 가지가 있다. 첫 번째로 번역 과정에서 언어간의 차이로 인한 단어 변환이나 유실이 존재한다. 한국어'수도관'의 경우 영문으로 번역하면'water pipe'가 된다.'수도관'은 한국어에서 한 개의 단어로 처리되기에 Word2Vec 모델에 적용할 수 있지만'water pipe'의 경우에는 두 개의 단어이고 토큰화를 거쳐 각각의 단어로 되기 때문에 단어의 유실이 발생한다. 이런 식으로 번역 과정을 거치면서 단어가 유실되거나 형태가 변하는 등 각 언어의 특성을 고려하기 힘들다는 문제점이 있다. 두 번째로 Word2Vec 모델의 파라미터이다. 여러 파라미터가 존재하는 모델의 특성 상 최적화된 파라미터 값을 찾는 과정이 필요하다.

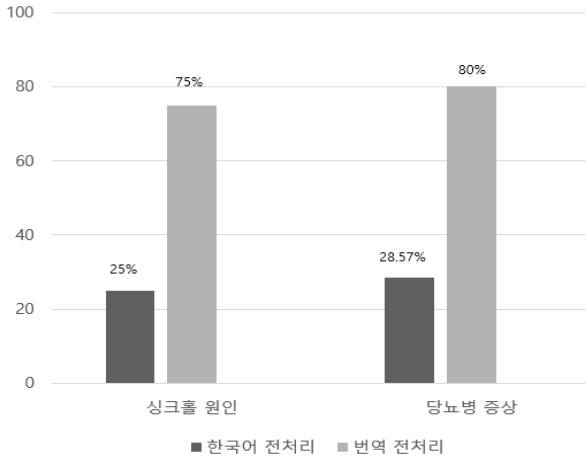

그림 3. 한국어 전처리, 번역 전처리 성능 비교 그래프

4. 결론

본 논문에서는 한국어 텍스트 분석 시 영문에 비해 분석이 상대적으로 어려운 문제를 해결하기 위해 번역 전처리 기법을 제안했다. 제안한 번역 전처리 기법의 성능 측정을 위해 한국어 뉴스 기사 데이터와 번역 전처리 기법이 적용된 영문 뉴스 텍스트 데이터를 사용했다. 실험 진행 후, 출력된 결과를 텍스트 분석 전문가들을 대상으로 투표를 진행했다. 그 결과, 약 3배의 득표 차이로 번역 전처리 기법을 적용한 결과가 한국어보다 더 성능이 좋다는 결론을 도출했다. 이를 통해 한국어 뉴스 분석

시 영문으로 번역 후 분석하는 것이 한국어 뉴스를 그대로 분석하는 것보다 더 의미 있는 결과를 도출할 수 있다는 것을 증명했다.

감사의 글

이 논문은 2019년도 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임

(2019-0-00136, 스마트시티 산업 생산성 혁신을 위한 AI융합

기술 개발)

참고 문헌 (References)

[1] 문혜정, 손원, 원중호. (2017). 텍스트 마이닝 기법을 이용한 경제심리 관련 문서 분류. 국민계정리뷰, 2017(4)

[2] 박대민. (2016). 뉴스 기사의 자연어처리:<뉴스소스 베타>를 중심으로. *커뮤니케이션 이론*, *12*(1), 4-52.

[3] 강형석, & 양장훈. (2018). 한국어 단어 임베딩 모델의 평가에 적합한 유추 검사 세트. *한국디지털콘텐츠학회 논문지*], *19*(10), 1999-2008.

[4] KoNLPy :https://konlpy-ko.readthedocs.io/ko/v0.4.3/#

[5] NLTK : https://www.nltk.org/

[6] Lexrank : https://pypi.org/project/lexrank/

[7] Chaulagain, Basanta. (2018). Casualty Information Extraction from News Article and Its Analysis.

[8] Taj, Soonh & Meghji, Areej & Shaikh, Baby. (2019). Sentiment Analysis of News Articles: A Lexicon based Approach.

[9] Hasan, H M & Sanyal, Falguni & Chaki, Dipankar. (2018). A Novel Approach to Extract Important Keywords from Documents Applying Latent Semantic Analysis. 10.1109/KST.2018.8426144.

[10] A. Kumar, A. Sharma, S. Sharma and S. Kashyap, "Performance analysis of keyword extraction algorithms assessing extractive text summarization," *2017 International Conference on Computer, Communications and Electronics (Comptelix)*, Jaipur, 2017, pp. 408-414, doi: 10.1109/COMPTELIX.2017.8004004.

[11] 설진석, 이상구. “lexrankr: LexRank 기반 한국어 다중 문서 요약”. 한국정보과학회 학술발표논문집, 458-460, 2016

[12] 김홍지, 김호준, & 이기훈. (2018, November). 다중문서 요약 고속화. In *Proceedings of KIIT Conference* (pp. 139-140).

[13] Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of ComputerApplications. 181. 10.5120/ijca2018917395.

[14] 김남훈, 주종민, 양형정, 박혁로. "단어의 의미와 문맥을 고려한 순환신경망 기반의 문서 분류." 정보처리학회논문지. 소프트웨어 및 데이터 공학, 7.7 (2018): 259-266.

[15] 박대서, 김화종. 20182TF-IDF 기반 키워드 추출에서의 의미적 요소 반영을 위한 결합벡터 제안.

[16] Miñarro Giménez, Jose Antonio & Marin Alonso, Oscar & Samwald, Matthias. (2015). Applying deep learning techniques on medical corpora from the World Wide Web: a prototypical system and evaluation.