KoNLPy와 Word2Vec을 활용한 한국어 자연어 처리 및 분석

Korean Natural Language Processing and Analysis Using KoNLPy and Word2Vec

저자
(Authors)

고동우, 양정진
Dong-Woo Ko, Jung-Jin Yang

출처
(Source)

한국정보과학회 학술발표논문집 , 2018.6, 2140-2142 (3 pages)

발행처
(Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/Article/NODE07503637

APA Style

고동우, 양정진 (2018). KoNLPy와 Word2Vec을 활용한 한국어 자연어 처리 및 분석. 한국정보과학회 학술발표논문집 , 2140-2142.

KoNLPy와 Word2Vec을 활용한 한국어 자연어 처리 및 분석

고동우° 양정진
가톨릭대학교 컴퓨터정보공학부
koodoowoo@naver.com, jungjin@catholic.ac.kr

Korean Natural Language Processing and Analysis Using KoNLPy and Word2Vec

Dong-Woo Ko° Jung-Jin Yang
Computer Science & Information Engineering, The Catholic University of Korea

요 약

자연어 처리란 인간의 언어를 분석해 컴퓨터가 처리하는 기술로서 빅데이터와 인공지능이 접목된 기술이다. 본 논문에서는 한국어 자연어 처리 과정을 단계별로 분석하고, KoNLPy의 Twitter 분석기와 gensim의 Word2Vec을 활용하여 특정 데이터로 학습 및 분석한 결과를 제시한다. 또한 전문용어 처리의 문제점 등 실험을 통해 얻어진 한국어 자연어 처리 과정에서 야기되는 문제점들을 분석하고 이에 대한 해결 방안을 재고한다.

1. 서 론

최근 빅데이터 분석과 인공지능을 활용한 자연어 처리(Natural Language Processing: NLP)의 요구가 증대되고 있다. 자연어 처리는 감정분석, 관련 상품 추천, 카테고리 세분화 등의 분야에서 활용 가능하다.[1,2] 자연어 처리를 하기 위해서는 텍스트를 숫자로 변환하는 Word Embedding 과정이 필요한데, 이에 쓰이는 대표적인 기술 중 하나인 'Word2Vec'을 사용할 것이다.[3] Word2Vec은 단어를 벡터 공간에 매핑하고, 벡터 값을 계산해 단어 간 유사도를 예측하는 시스템이다. 따라서 Embedding의 결과에 따라 예측의 결과가 달라진다.

Word2Vec은 단어를 벡터 공간에 매핑하기 때문에 문서를 입력했을 때 뜻을 가진 최소단위의 단어로 Token화 해주는 과정(Tokenizing)이 필요하다. 영어로 작성된 문서의 경우는 띄어쓰기를 기준으로 나누면 Tokenizing이 되기 때문에 어렵지 않게 Token화가 가능하다. 하지만 한국어 문서의 경우 어미나 조사가 붙게 되어 같은 단어도 다르게 분류되는 경우가 생긴다. 예를 들어 '아빠'에 관련된 문서를 띄어쓰기를 기준으로 Tokenizing하면 '아빠가', '아빠를', '아빠에게' 등 '아빠' 뒤에 어미나 조사가 붙어 '아빠'라는 단어는 Embedding되지 않을 것이다. 이러한 문제를 해결하기 위해서는 명사, 동사, 어미, 조사 등으로 구분하는 형태소 분류 및 전처리가 한국어 자연어 처리를 하는데 필수적으로 거쳐야 하는 과정이다. 본 논문에서는 한국어 자연어 처리에 대표적으로 사용되는 형태소 분석기를 모아서 만든 파이썬 패키지 'KoNLPy'를 활용한다.[4]

KoNLPy는 5개의 형태소 분석기가 있으며, 각 분석기는 분석기마다 가지고 있는 Dictionary를 기반으로 단어를 Tokenizing한다.

본 논문은 KoNLPy와 Word2Vec 방법을 적용하여 한국어 자연어 처리에 대하여 실험한다. 실험은 전문용어가 빈번히 사용되는 생물정보학 분야 논문과 비교적 전문용어의 사용이 적은 국문학 분야 논문의 두 분야를 데이터로 선별하여 처리한 결과를 비교 분석 함으로써 상위 방법의 한국어 자연어 처리에 대한 유용성을 알아보고자 한다.

본 논문의 구성은 1장 서론에 이어 2장에서 KoNLPy와 Word2Vec을 이용한 한국어 자연어 처리 과정을 설명하고, 3장에서는 자연어 처리 실험 결과를 분석한 뒤 실험에서 야기되는 문제점을 제시한 후 해결책을 재고한다. 끝으로, 4장에서 결론 및 향후 연구 방향을 논한다.

2. 한국어 자연어 처리 시스템

본 논문의 한국어 자연어 처리 시스템은 [그림 1]과 같다. PDF 파일을 TXT 파일로 변환 후 전처리 과정을 거쳐 Tokenizing한다. Tokenizing된 데이터를 Word2Vec 처리하는 과정으로 시스템은 마무리된다.

[그림 1] 한국어 자연어 처리 과정

2.1 Tokenizing

Tokenize란 어미, 조사 등이 붙어 구분이 불가능한 한국어를 명사, 동사, 어미, 조사 등으로 형태소를 분석해 Word2Vec에 적합한 데이터를 만드는 과정을 뜻한다.

Tokenize에 앞서 다량의 데이터를 학습시키기 위해 PDF

파일을 TXT파일로 변환하는 과정이 필요하다. PDF 파일에서 텍스트를 추출해 TXT파일로 저장하는 방법으로 파이썬 라이브러리 'PDFMiner'를 사용하였다.[5] 변환된 TXT 파일을 읽어들여 개행, 특수문자 등 학습에 불필요한 단어들을 제거하는 전처리 과정을 거치게 되면 Tokenize에 적합한 데이터가 완성된다.

Tokenize를 위해 파이썬 패키지 KoNLPy를 사용한다. KoNLPy는 C++, JAVA 등으로 구현된 형태소 분석기들을 모아 파이썬에서 사용가능하게 만든 오픈 패키지다. KoNLPy는 총 5개의 형태소 분석기(Hannanum, Kkma, Komoran, Mecab, Twitter)로 이루어져 있다. 분석기는 각각의 Dictionary를 갖고 있고, Dictionary 내용을 기반으로 형태소 분석을 한다. 따라서 Dictionary에 포함되지 않은 단어는 Token화가 불가능하다는 문제점을 안고 있다. 본 논문에서는 KoNLPy 분석기 중 대중적으로 사용되고 분석 속도가 빠른 Twitter 분석기를 사용한다.

```
tokenize('한국어 자연어 처리 시스템을 구현해보겠습니다')
['한국어/Noun',
 '자연어/Noun',
 '처리/Noun',
 '시스템/Noun',
 '을/Josa',
 '구현/Noun',
 '해보다/Verb']
```

[그림 2] Tokenize 결과

[그림 2]와 같이 한국어로 작성된 문장을 입력하면 명사, 동사, 조사, 어미 등의 각 품사로 구분된다.

2.2 Word2Vec

본 단계는 앞서 처리한 Tokenize 데이터를 Word2Vec하는 과정이다. 보다 정확한 Embedding을 위해 단어마다 부여된 tag 정보까지 Word2Vec 학습에 사용한다. Word2Vec은 gensim에 구현된 Word2Vec 모델을 사용한다.[6]

Word2Vec의 학습은 CBOW와 Skip Gram 방식으로 구성된다.[7,8] CBOW는 주변 단어로 중심 단어를 예측하는 것이다. '배가 ___ 더 이상 먹을 수 없다'의 빈 칸이 예측 가능하듯 문장이 입력되면 문장의 각 단어들마다 단어를 예측하기 위한 네트워크를 구성한다. Skip Gram은 중심 단어로 주변 단어들을 예측하는 것이다. '아이스크림'이란 단어로 '차갑다, 먹는다, 녹는다' 등이 예측 가능하듯 문장이 입력되면 문장의 각 단어들의 주변 단어 등장을 학습한다.

Word2Vec 학습 조건을 설명하면 100차원 벡터 공간에 단어를 Embedding하고, 학습 단어의 앞, 뒤 8단어를 확인하고, 문서 내 출연빈도가 10번 미만인 것을 학습에 제외한다. 학습 방법으로 Skip Gram을 사용하며, 이 과정을 100번 반복 학습 한다.

3. 학문분야별 실험 및 결과

본 연구의 실험에 사용된 데이터는 DBpia(국내 학술저널, Conference Proceedings, 전문잡지, 전자책, 웹DB 등을 제공하는 온라인 서비스)에서 제공되는 생물정보학 분야 논문과 국문학 분야 논문 각 120건을 사용한다.

두 데이터 집합의 차이점은 생물정보학은 일반적으로 자주 사용되지 않는 생물정보 관련 전문용어가 다양하게 활용되는 분야이고, 국문학은 생물정보학에 비해 상대적으로 전문지식이 요구되는 전문용어의 활용이 낮은 분야이다. 이러한 기준으로 서로 다른 두 분야를 선택하여 각 분야에 대한 데이터의 자연어 처리 결과를 비교 분석한다.

3.1 Word2Vec 결과 분석

생물정보학 논문을 Tokenizing한 결과 416,200개의 단어로 분석되었고, 국문학 논문을 Tokenizing한 결과 1,224,035개의 단어로 분석되었다. 이어서 Tokenize된 논문 데이터를 Word2Vec 학습 시킨 결과 생물정보학 논문은 5,278개의 단어가 학습되었고, 국문학 논문은 7,714개의 단어가 학습되었다. 논문의 분량 차이로 분석된 단어 수의 차이는 있었지만, Word2Vec 학습된 모델은 2,500개 차이로 결과를 도출했다.

[그림 3]와 [그림 4]은 각 논문 분야에서 쓰이는 단어를 입력했을 때 유사도가 높은 단어 5개를 출력한 결과다. 그림을 간략히 설명하면 most_similar는 입력한 단어와 유사한 단어를 찾는 Word2Vec의 메소드다. 빨간색으로 표시된 단어를 입력했을 때 유사한 단어와 유사도를 나타내준다. 결과를 보면 알 수 있듯 입력 단어와 연관이 있는 단어들이 출력된 것을 볼 수 있다.

```
wv_model_bio.wv.most_similar(tokenize('유전'), topn=5)
[('형질/Noun', 0.5339716672897339),
 ('정보/Noun', 0.5305873155593872),
 ('자손/Noun', 0.5238721370697021),
 ('염색체/Noun', 0.5144349336624146),
 ('생명체/Noun', 0.5034881830215454)]

wv_model_bio.wv.most_similar(tokenize('신약'), topn=5)
[('투자/Noun', 0.5893573760986328),
 ('기업/Noun', 0.5856045484542847),
 ('발굴/Noun', 0.535403311252594),
 ('의약/Noun', 0.5342029333114624),
 ('개발/Noun', 0.5219202041625977)]

wv_model_bio.wv.most_similar(tokenize('말라리아'), topn=5)
[('열대/Noun', 0.72654128074646),
 ('열원충/Noun', 0.6778055429458618),
 ('키트/Noun', 0.6585235595703125),
 ('원충/Noun', 0.6442424058914185),
 ('치쿤구니/Noun', 0.6297529935836792)]
```

[그림 3] 생물정보학 Word2Vec 결과

[그림 4] 국문학 Word2Vec 결과

3.2 처리 결과에 따른 문제점 고찰

본 단계는 Word2Vec 학습 시 야기되는 문제를 제시한다. 앞서 말한 내용과 같이 Word2Vec은 Tokenize된 데이터를 이용해 학습을 진행하는데 Tokenize 과정은 전적으로 KoNLPy에 의존하게 된다. KoNLPy에 구성되어 있는 각 분석기는 Dictionay를 기반으로 단어를 분류하는데 만약 Dictionary에 포함되어있지 않은 단어는 음절단위로 구분되거나 의도한 방향과 다르게 분석된다. 따라서 전문용어가 빈번히 사용되는 생물정보학 같은 분야는 Word Embedding이 제대로 되지 않는 문제가 발생한다.

'생물학/Noun',	'클로르/Noun',	'메탄/Noun',
'에서/Josa',	'페/Noun',	'및/Noun',
'프로/Noun',	'네/Determiner',	'황화/Noun',
'테오/Noun',	'신/Noun',	'이/Josa',
'믹스/Noun',	'카르/Noun',	'메틸/Noun',
'의/Josa',	'바/Noun',	'제거/Noun',
'응용/Noun',	'메이트/Noun',	'특성/Noun',

[그림 5] 잘못된 Tokenizing의 예시

[그림 5]는 프로테오믹스, 클로르페네신카르바메이트, 황화이메틸 등 생물정보학에서 쓰이는 전문용어 중 잘못된 Tokenizing 예시다. 따라서 예시와 같이 황화이메틸은 '황화', '이', '메틸'로 구분되어 '황화이메틸'이란 단어는 제대로 Embedding되지 않는다. 이러한 문제로 인해 한국어 문서를 Word2Vec 시 전문용어 및 신기술 용어 학습이 어렵다는 문제점이 발생한다.

전문용어 및 신기술 용어를 포함하는, 현재의 Dictionary를 보완하는 방법을 추가 및 확장하여 한국어 문서를 Word2Vec하는 방법을 재고할 필요가 있다. 예를 들면, WordNet이나 온톨로지와 같은 방법을 통해, 각 분야별로 전문용어 말뭉치를 만들어 Tokenize 전에 형태소 분석기 Dictionary에 전문용어 말뭉치를 포함해 주어 처리하는 과정을 거친다면, 전문용어 역시 올바른 Tokenize 후 Word2Vec 학습이 가능해 질 것으로 사료된다.

4. 결론 및 향후 연구

본 연구의 목적은 한국어 문서를 Word2Vec 학습할 시 야기되는 문제점을 알아보고, 문제점에 대한 해결 방안을 재고하는 것이다. 과거 한국어 자연어 처리 연구는 기사, 소설, 댓글 등의 전문용어가 포함되지 않은 데이터를 다루었다. 본 연구에서는 전문용어가 포함된 문서를 한국어 자연어 처리 했을 때 문제점을 알아본다. 문제점을 알아보기 위하여 생물정보학 논문과 국문학 논문 각 120건을 KoNLPy의 Twitter를 이용해 Tokenizing 후 분석된 데이터를 각각 Word2Vec 학습했다. Word2Vec 학습 결과 일반적인 단어는 유의미한 결과를 출력했다. 하지만 생물정보학 논문에서 사용되는 전문용어는 올바르게 Embedding되지 않았다. 분석기 Dictionary에 전문용어가 포함되어있지 않기 때문이다.

앞선 문제점 해결하기 위해 본 연구에서 재고하는 방법은 각 분야별 전문용어 및 신기술 용어를 갖는 말뭉치를 만들어 Tokenize 전에 형태소 분석기 Dictionary에 말뭉치를 포함시키는 것이다. 위의 과정을 거치면 전문용어 및 신기술 용어 역시 올바르게 Tokenize되고, Word2Vec 학습 시 유의미한 결과를 도출할 수 있다.

향후 연구 방향으로는 분야별 전문용어 말뭉치를 만들어 Dictionary에 포함시킬 수 있도록 하고, 전문용어가 포함된 문서를 Word2Vec 학습시켜 유의미한 결과를 얻는 연구가 필요하다.

《참고문헌》

[1] 정영희, 기계학습 기반의 한국어 단문 감성분류 기법에 관한 연구, 101, 2017

[2] 손지영, word2vec을 이용한 거리 기반의 음악 가사 클러스터링 기법, 30, 2017

[3] Yoav Goldberg and Omer Levy, word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method, arXiv, 5, 2014

[4] Lucy Park, KoNLPy : Korean natural language processing in Python, Proceeding soft he 26th Annual Conferenceon Human & Cognitive Language Technology, 44, 2017

[5] https://www.unixuser.org/~euske/python/pdfminer/

[6] https://radimrehurek.com/gensim/models/word2vec.html

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, Proceedings of Workshop at ICLR, 12, 2013.

[8] Xin Rong, word2vec Parameter Learning Explained, 21 ,2016