바이오 마커와 질병 용어의 단어 표현 분석

윤영신○, 남경민, 김유섭
한림대학교 융합소프트웨어학과
pour657@gmail.com, jkre4030@naver.com, yskim01@hallym.ac.kr

Word Representation Analysis of Bio-marker and Disease Word

Young-Shin Youn○, Kyung-Min Nam, Yu-Seop Kim
Dept. of Convergency Software, Hallym University

요 약

기계학습 기반의 자연어처리 모듈에서 중요한 단계 중 하나는 모듈의 입력으로 단어를 표현하는 것이다. 벡터의 사이즈가 크고, 단어 간의 유사성의 개념이 존재하지 않는 One-hot 형태와 대조적으로 유사성을 표현하기 위해서 단어를 벡터로 표현하는 단어 표현 (word representation/embedding) 생성 작업은 자연어 처리 작업의 기계학습 모델의 성능을 개선하고, 몇몇 자연어 처리 분야의 모델에서 성능 향상을 보여주어 많은 관심을 받고 있다. 본 논문에서는 Word2Vec, CCA, 그리고 GloVe를 사용하여 106,552개의 PubMed의 바이오메디컬 논문의 요약으로 구축된 말뭉치 카테고리의 각 단어 표현 모델의 카테고리 분류 능력을 확인한다. 세부적으로 나눈 카테고리에는 질병의 이름, 질병 증상, 그리고 난소암 마커가 있다. 분류 능력을 확인하기 위해 t-SNE를 이용하여 2차원으로 단어 표현 결과를 맵핑하여 가시화 한다.

주제어: 바이오 마커, 질병 용어, 단어 표현

1. 서론

기계학습 기반의 자연어처리 모듈에서 중요한 단계중 하나는 모듈의 입력으로 단어를 표현하는 것이다. 대부분 이전 연구에서는 단어를 One-hot 형태로 표현한다. One-hot 형태는 vocabulary 사이즈와 같은 벡터가 있고 해당 단어만 1로 표현하고 나머지는 0으로 표현하는 방식이다 (즉,[0001000])[1]. 이런 단어 표현 방식은 크게 두 가지 문제점을 갖고 있다. 첫 번째로, 벡터의 사이즈가 매우 크고 두 번째로, 단어 간의 유사성의 개념이 존재하지 않는다.

단어 간의 유사성을 표현하기 위해서 최근 단어를 벡터로 표현하는 단어 표현(word representation) 작업은 자연어 처리 작업의 기계학습 모델의 성능을 개선하여 많은 관심을 받고 있다.

단어 표현은 One-hot 형태와 다르게 단어 별로 k 차원으로 축소된 단어 표현을 학습하게 된다. 이런 단어 표현 방식은 최근에 몇몇 자연어처리 분야의 모델에서 성능향상을 보여준다.

예를 들어 [2]는 단어 표현을 CRF의 자질 (feature)로 사용하여 개체명 인식 (Named entity recognition)의 성능을 향상 시켰다. 또한 최근 연구[3]에서는 바이오메디컬 도메인에서 word2vec, GloVe를 이용해서 단어쌍의 유사도를 확인하고 단어 표현의 효율을 입증했다.

본 논문에서는 [3]의 연구에서 더 나아가 단어 표현을 이용한 바이오메디컬 도메인의 특정 카테고리의 분류 능력을 확인 하고자 한다.

우리는 [3]에서 사용한 word2vec[4,5], GloVe[6]와 추가적으로 CCA[7] 모델을 사용한다.

실험에 사용된 단어 표현 모델들은 바이오메디컬 도메인에서 구문뿐만 아니라 단어의 잠재적인 의미를 찾을 수 있다. 우리는 PubMed의 바이오메디컬 논문의 요약을 사용하여 말뭉치를 구축하고 질병, 증상, 바이오 마커로 카테고리를 나눠 각 단어 표현 모델의 카테고리 분류의 능력을 확인한다. 바이오 마커로는 난소암 마커를 사용하였다.

분류 능력을 확인하기 위해 t-Distributed Stochastic Neighbor Embedding (t-SNE)[1] 을 이용하여 k 차원의 단어 표현 결과를 2차원으로 맵핑하여 가시화 한다.

2장에서는 본 논문에서 사용한 단어 표현에 대해서 살펴보고, 3장에서는 단어 표현에 사용한 데이터에 관해 설명하며, 데이터를 가지고 한 실험에 대한 내용은 4장에서 설명한다. 마지막으로 5장에서는 결론과 향후 연구에 대하여 논의한다.

2. 단어 표현

단어 표현 혹은 분산된 표현 (distributed representation) 은 주어진 말뭉치에 있는 모든 단어에 대한 벡터 표현을 학습하는 기술이다. 이 단어 표현은 단어 간의 유사도를 볼 수 없는 one-hot 형태와는 대조

1) http://lvdmaaten.github.io/tsne/

적으로 k차원으로 단어 표현을 학습한다. 본 논문에서는 많은 단어 표현 모델 중에서 Word2Vec, CCA, 그리고 GloVe 이 세 가지의 모델을 사용한다.

2.1. Word2Vec

Word2Vec[2] 은 일반적으로 서로 다른 언어 모델인 Continuous Bag of Word (CBOW) 와 Skip-gram을 갖는다 [4,5]. CBOW 모델은 주변의 단어를 가지고 단어를 예측한다. 따라서 CBOW모델의 입력은 예측하려는 단어의 주변단어가 된다. skip-gram 의 경우, 하나의 단어가 주어졌을 때 단어의 주변 단어 또는 문맥을 예측한다. 즉, skip-gram의 경우에는 문장이나 문서의 주변 단어를 예측하는데 유용한 단어표현이다. 이와 같이 두 모델은 신경망 기반의 언어 모델로 대량의 말뭉치가 입력으로 들어가고 말뭉치의 각 단어의 단어표현을 학습한다. 본 논문에서는 Word2Vec의 skip-gram모델을 사용한다.

2.2. CCA

canonical correlation analysis(CCA)[7]는 랜덤 변수 ($X, Y \in \mathbb{R}$)사이의 통계량 계수를 최대화 하고자 하는 모델이다. 즉, 랜덤 변수에서 단어 표현을 x, 그 단어와 관련된 문맥 표현을 y라고 했을 때, 두 변수의 상관관계를 최대화 하는 k차원의 투영 벡터를 찾는다.

2.3. GloVe

GloVe[6]는 Global Vector를 의미하며, 전역 문맥(global context) 을 고려할 뿐만 아니라 단어의 지역 문맥 (local context) 또한 고려한 하이브리드 방식의 단어 표현으로 볼 수 있다. GloVe의 w_x와 w_y 단어 학습을 위한 dot product는 co-occurrence count에 비례한다. 우리는 자유롭게 이용 가능한 GloVe 오픈 소스 툴[3]을 사용한다.

3. 데이터

본 연구에서는 PubMed의 바이오메디컬 논문 106,552개의 제목과 요약부분으로 말뭉치를 구축하였다. 구축한 말뭉치로 질병의 이름, 질병의 증상 그리고 바이오 마커로 카테고리를 만든다. 바이오 마커에는 난소암 마커를 사용하였다. 질병의 증상 같은 경우, [3]에서 사용한 벤치마크 데이터[4]에서 랜덤으로 추출한 데이터를 사용한다. 이렇게 구축되어진 말뭉치는 단어 표현 분석을 위한 input data로 사용한다.

[표 1]은 사용된 카테고리에 있는 질병 이름, 질병 증상, 바이오 마커 리스트이다.

질병이름	질병 증상	난소암 마커
폐렴	어지럼증	CA125
백내장	협심증	CA19-9
녹내장	빈혈증	EGFR
요도염	무취증	Myolobin
위염	패혈증	Tenascin-C
뇌수막염	저단백혈증	Apoa-i
결막염	골다공증	Apoc-iii
방광염	이상지질혈증	CRP
구내염	동맥경화증	FSH
기흉	진균증	Cortisol
천식	갑상선종	TTR
백혈병	전립샘증	CA15-3
선암	근이영양증	MIF
암	칸디다증	Leption
종양	심근증	IL-6
치매	균혈증	CEA
간염	혈전증	IL-8
고혈압	편모충증	Prolactin
당뇨병	백혈구감소증	OPN
결핵	브루셀라증	HE4
수두	혈소판감소증	MMp-7

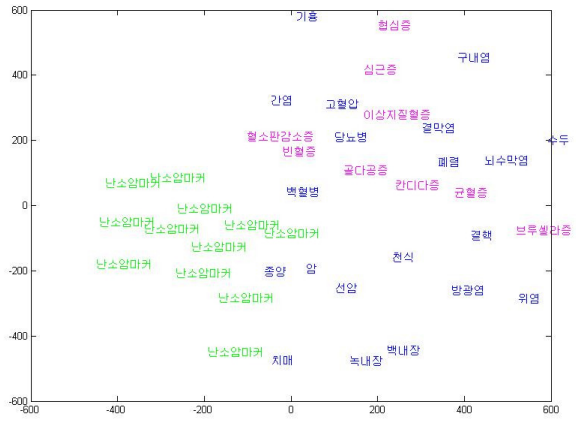
[표 1] 질병, 증상, 바이오 마커 카테고리 리스트

4. 실험

3장에서 구축한 바이오메디컬 도메인 카테고리를 Word2Vec, CCA, 그리고 GloVe 모델을 사용하여. 분류 능력을 확인하기 위해 t-Distributed Stochastic Neighbor Embedding (t-SNE) 을 이용하여 k 차원의 단어 표현 결과를 2차원으로 맵핑하여 가시화 한다.

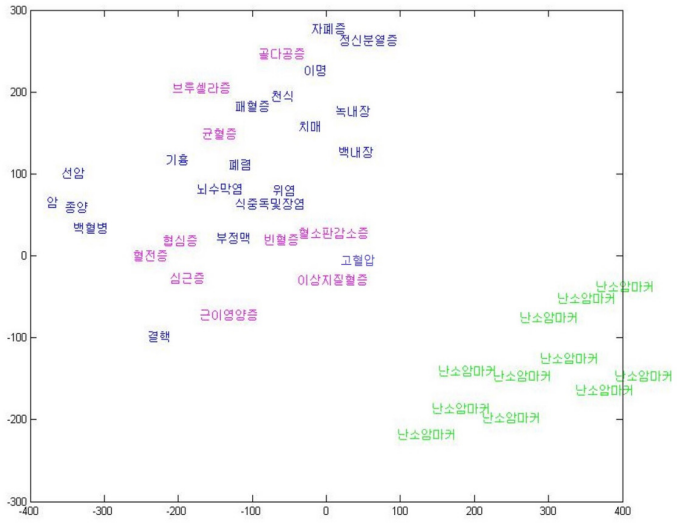
[그림 1-3]는 각각 Word2Vec의 skip-gram 모델, CCA, GloVe를 사용 하였을 때, 나온 단어 표현 결과이다. 각각의 그림의 파란색 글씨는 질병 이름, 자주색 글씨는 질병의 증상을 나타내며, 초록색은 난소암 마커이다.

2) https://code.google.com/p/word2vec/
3) http://nlp.stanford.edu/projects/glove/
4) http://rxinformatics.umn.edu.



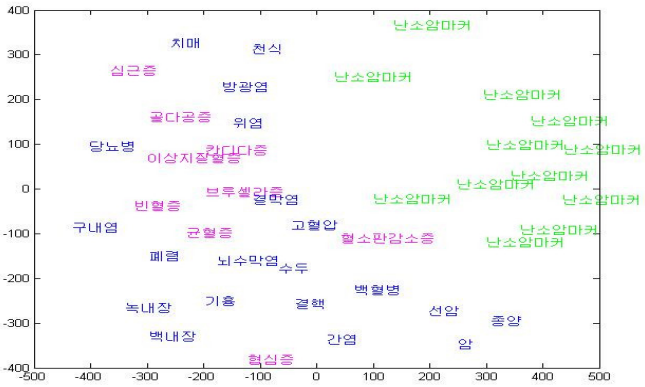
[그림 2] Word2Vec(skip-gram)모델 단어 표현 결과

Word2Vec모델의 단어 표현 결과를 보았을 때, 질병과 난소암 마커가 잘 구별이 되어 있는 것을 알 수 있다. 질병에서는 혈소판 감소증의 관련 질병에는 백혈병과 빈혈증이 있는데, 실제로 [그림 1]을 보았을 때, 혈소판 감소증과 백혈병, 그리고 빈혈증 이 3개의 질병과 증상의 거리가 서로 가까운 것을 확인 할 수 있다. 그리고 혈중에 총 콜레스테롤, LDL콜레스테롤, 중성지방이 증가된 상태거나 HDL콜레스테롤이 감소된 상태를 말하는 이상지질혈증의 관련 질병에는 고혈압이 있다. 암과 관련된 세 단어 암, 종양, 위, 장, 기관지 등의 점막을 비롯해서 이자의 선 조직이나 배설관에서 발생하는 선암도 잘 모여 있기 때문에 전반적으로 잘 구별 된다는 것을 알 수 있다. 이와 같이 관련이 있는 질병 단어들이 서로 가깝게 위치되어 있다는 것을 확인 할 수 있다.



[그림 3] CCA모델 단어 표현 결과

CCA모델 또한 앞서 보았던 Word2Vec 모델처럼 난소암 마커와 질병들이 서로 잘 구별 되어 있는 것을 확인 할 수 있다. [그림 2]를 보면, 앞서 [그림 1]에서 설명한 이상지질혈증과 관련된 질병인 고혈압이 서로 가깝게 위치 되어있는 것을 확인 할 수 있고, 혈소판 감소증과 빈혈도 서로 가까운 위치에 있는 것을 확인할 수 있다. 그리고 암, 종양 그리고 선암이 [그림 1]보다 아주 가깝게 위치되어 있는데, 이는 이 세 단어가 모두 암과 관련된 용어이기 때문이다. 그리고 Word2Vec모델, GloVe모델과는 다르게 추가적으로 염증 질환에 관련된 뇌수막염, 위염, 그리고 식중독 및 장염과 심장 질환에 관련된 증상인 협심증, 혈전증, 심근증 또한 서로 가까운 거리에 위치 해 있다는 것을 알 수 있다.



[그림 4] GloVe모델 단어 표현 결과

[그림 3]은 GloVe모델의 단어 표현 결과이다. [그림 2]과 같이 암에 관련된 단어들이 서로 멀지않은 곳에 위치되어 있는 것과, 눈과 관련된 질병인 녹내장, 백내장이 두 단어의 거리가 서로 가까운 것을 확인 할 수 있다.

이와 같이 세 단어표현 모델로부터 질병과 난소암 마커가 잘 분류 되는 것을 확인 하고, 서로 관련 있는 단어들이 가깝게 위치되어 있다는 것을 확인 하였다. 그리고 CCA모델에서는 Word2Vec, GloVe모델과 다르게 염증 질환에 관련된 질병들과, 심장 질환에 관련된 증상들끼리도 잘 구별을 하는 것을 확인 할 수 있었다.

5. 결론

본 논문에서는 PubMed의 바이오메디컬 논문의 요약을 사용하여 말뭉치를 구축하고 질병, 증상, 난소암 마커로 카테고리를 나눠 각 단어 표현 모델의 카테고리 분류의 능력을 확인하였다. 그 결과를 t-SNE를 이용하여 k차원의 단어 표현 결과를 2차원으로 맵핑하여 가시화 하였다.

4장에서 본 실험 결과처럼 세 모델에서 난소암 마커와 질병을 잘 분류 하고, 서로 관련 있는 질병을 단어표현으로 알 수 있었다. 향후에는 더 나아가 본 논문에서 쓰인 데이터보다 더 많은 데이터를 가지고 특정 질병과 관련된 새로운 바이오 마커를 찾는 연구를 하고자 한다.

참고문헌

[1] Ronan Collobert, et al. Natural language, processing (almost) from scratch. The Journal of Machine Learning Research, 12, 2011.

[2] Turian, Joseph, Lev Ratinov, and Yoshua Bengio. "Word representations: a simple and general method for semi-supervised learning." Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 2010.

[3] Muneeb, T. H., Sunil Kumar Sahu, and Ashish Anand. "Evaluating distributed word representations for capturing semantics of biomedical concepts." ACL-IJCNLP 2015, pp.158, 2015

[4] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[5] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

[6] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014) 12 (2014): 1532-1543.

[7] Karl Stratos, Michael Collins, and Daniel Hsu. "Model-based word embeddings from decompositions of count matrices." Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2015), 1282 - 1291, 2015.