단어 표현에 기반한 연관 바이오마커 발굴

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

Biomarker Detection of Specific Disease using Word Embedding

Young-Shin Youn○, 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. 2차원으로 맵핑된 결과 값을 코사인 유사도를 사용하여 질병과 바이오 마커간의 유사도를 구한다. 이 유사도 결과 값 상위 20쌍의 결과를 가지고 실제 연구가 되고 있는지 구글 스콜라를 통해 관련 논문을 검색하여 확인하고, 검색 결과를 점수화 한다. 실험 결과 상위 20쌍 중에서 85%의 쌍이 실제적으로 질병과 바이오 마커 간의 관계를 파악하는 방향으로 진행 되고 있으나, 나머지 15%의 쌍에 대해서는 실질적인 연구가 잘 되고 있지 않은 것으로 파악되었다.

주제어: 질병 용어, 난소암, 바이오 마커, 워드 임베딩, CCA, 코사인 유사도, t-SNE

1. 서론

바이오 마커란 질병을 발견하거나 치료, 모니터링 그리고 예측하는데에 사용되는 중요한 도구이다. 이는 물리적인 상태와 질병의 과정의 변화와 지표를 표시하는 생물학적인 분자이다. 이러한 바이오 마커의 발굴방법에는 다양한 방법이 있다. 바이오 마커를 발굴하는 방법의 가장 간단한 방법은 혈액을 이용하거나[1] 단백질 정보를 이용하는 것이다 [2]. 그리고 더 나아가 바이오 마커를 발굴 할 때 워드 임베딩을 사용 할 수도 있다. 워드 임베딩은 단어 간의 유사도를 표현하기 위하여 단어를 벡터로 표현하는 방법이다. 이러한 방법은 벡터의 사이즈가 크고 해당되는 단어만 1로 표현하고, 나머지는 0으로 표현하는 one-hot방식[3] 과는 다르게 단어 별로 k차원의 축소된 표현을 학습한다. [4]에서는 바이오메디컬 도메인에서 워드 임베딩 모델 중에서 Word2Vec, GloVe를 이용하여 bio-NLP 영역에서의 효율성도 입증하였다.

본 논문에서는 난소암과 관련된 바이오 마커가 나온 PubMed 바이오메디컬 도메인의 제목과 요약부분을 새로운 말뭉치로 구축한다. 구축된 말뭉치를 가지고 워드 임베딩 모델 중에서 CCA를 사용한다. CCA의 임베딩 결과 값을 t-SNE(t-distributed stochastic neighbor embedding)[5] 을 사용하여 2차원으로 맵핑하고 결과를 가시화 한다. 2차원으로 맵핑된 결과 값을 코사인 유사도를 사용하여 질병과 바이오 마커 간의 유사도를 구하고, 이 결과 값을 기준으로 유사도 상위 20쌍의 결과가 실제 연구 되고 있는지 관련 논문을 구글 스콜라를 통해 검색한다. 검색 결과 유사도 점수가 높은 경우 85%는 실제로도 많은 연구가 질병과 바이오 마커간의 관계를 파악하는 방향으로 진행되었으나, 나머지 15% 쌍에 대해서는 실질적인 연구가 잘 되지 않는 것으로 파악되었다.

2장에서는 본 논문의 주된 목적인 바이오마커 발굴에 대하여 간략히 설명 하고, 3장에서는 본 연구에 사용된 실험 방법들에 대하여 설명한다. 4장에서는 실험에 사용한 데이터에 관하여 설명한다. 실험에 대한 내용은 5장에서 설명하고, 마지막으로 6장에서는 결론과 향후 연구에 대하여 논한다.

2. 바이오 마커 발굴

바이오 마커는 일반적으로 단백질등을 이용하여 몸 안의 변화를 알아 낼 수 있는 지표를 의미한다. 이는 많은 과학적 분야에 이용되며, 평범한 생물처리 과정이나, 치료를 위한 약리학의 과정을 측정하거나 평가하는데에 사용 된다[1]. 의학적으로 바이오 마커는 장기의 기능을 검사하는데에 추적 가능한 물질을 의미한다. 이는 질병을 감지하거나 약물 개발 등 다양하게 사용된다. 바이오 마커 발견, 발굴이란 바이오 마커를 검출하는 과정을 뜻한다. 바이오 마커를 발굴하는 방법은 크게 4종류로 나뉜다. 첫 번째는 게놈을 이용한 접근방법이고, 두 번째는 단백질의 정보를 이용하는 것이다. 세 번째로는 Metabolomics(대사 체학)를 이용한 접근 방법이 있다. 마지막 방법은 Lipidomics를 이용한 접근 방법이다.

그 외에도 [6]에서는 암을 발견하기 위해서 데이터 마이닝 기법을 적용하였다. 이와 같이 최근에는 머신 러닝, 데이터 마이닝, 그리고 비지도 학습(Unsupervised Learning)과 워드 임베딩을 이용하여 바이오 마커를 발굴 할 수 있다 [7-9].

3. 실험 방법

본 논문에서는 크게 두 가지 방법을 사용하여 질병과 특정한 바이오 마커 간의 관계를 파악한다.

3.1 워드 임베딩

워드 임베딩은 주어진 코퍼스에 있는 모든 단어에 대한 벡터 표현을 학습하는 기술이다. 대부분의 이전 연구들에서는 단어를 one-hot 형태로 표현하였다. 이 표현 방식은 단어가 사전의 사이즈와 같은 벡터들을 가지고 있으므로 벡터의 사이즈가 크고, 단어 간의 유사성의 개념이 존재하지 않아서 단어가 본질적으로 다른 단어와 어떤 차이점을 가지는지에 대해 이해 할 수 없다는 단점이 있다. 워드 임베딩은 이러한 one-hot 방식과는 다르게 k차원으로 단어 표현을 학습한다. 본 논문에서는 워드 임베딩의 많은 모델 중 CCA모델을 사용하여 특정한 질병과 난소암 마커 간의 관계를 파악한다.

CCA[10]는 변수들의 상관관계를 살피는 기법이다. [11]에서는 CCA가 두 단어의 관계를 조사하기 위해서 유용한 도구가 될 수 있음을 보였다. CCA는 랜덤 변수 x는 단어 표현을, 그 단어와 관련된 context 표현을 y라고 했을 때, 이 둘의 상관관계를 최대화하는 k차원의 투영 벡터를 찾는 모델이다.

3.2 코사인 유사도

코사인 유사도는 내적 공간의 두 벡터간 각도를 코사인 값을 이용하여 측정된 벡터간의 유사도이다. 이는 벡터 공간 모델에서 가장 많이 사용되는 문서와 질의어 간의 유사도 계산법이다. 코사인 유사도는 벡터 A와 벡터 B가 주어졌을 때, 내적과 벡터의 크기 등을 이용하여 표현된다.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

계산된 유사도는 -1에서 1 사이의 값을 가진다. 예를 들어 벡터 A가 (1,2)값을 가지고, 벡터 B가 (3,4)의 값을 가질 때, 두 벡터의 코사인 유사도는 0.9839가 된다.
코사인 유사도를 사용하여 각각의 질병과 각각의 난소암 마커 간의 유사도 결과값을 토대로 상위 20개의 쌍을 뽑는다.

4. 데이터

본 논문에서는, 1,080,194개의 PubMed 바이오 메디컬 도메인의 제목과 요약 부분에서 난소암과 관련된 문서들을 사용하여 말뭉치를 만들었다. 말뭉치는 [표 1]처럼 각각 질병 이름과 난소암 바이오 마커로 분류하였다. 카테고리를 구출 할 때, 질병 이름에는 연구자를 제외한 다른 사람들에게 잘 알려진 질병들을 위주로 하여 구축하였고, 난소암 마커의 경우 [12]에서 사용한 난소암 마커를 사용하였다. 구축된 말뭉치로 분류한 카테고리는 리스트로 작성하여 워드 임베딩의 input 데이터로 사용한다.

질병 이름	선암, 천식, 암, 결막염, 방광염, 치매, 당뇨병, 위염, 녹내장, 간염, 고혈압, 백혈병, 뇌수막염, 기흉, 폐렴, 구내염, 결핵, 종양
난소암 바이오 마커	apoa-i, apoa-iii, CA125, CA15-3, CA19-9, CEA, Cortisol, CRP, CYFRA21-1, EGFR, FSH, HE4, IL-6, IL-8, MIF, MMP-7, Myoglobin, OPN, Prolactin, Tenascin-C, TTR

[표 1] 질병 이름, 난소암 마커 카테고리 리스트

5. 실험

본 논문에서는 4장에서 구축한 질병 이름과 난소암 바이오 마커 리스트를 워드 임베딩의 모델 중 CCA를 이용하여 서로의 관계를 파악한다. CCA로 나온 질병과 바이오 마커의 k차원 벡터들을 t-SNE를 사용하여 2차원으로 만든다. 2차원으로 만든 결과 값을 가지고 코사인 유사도를 사용하여 유사도 계산을 한다. 이처럼 난소암 바이오마커와 다른 질병들이 관련이 있는지 맵핑하는 이유는 난소암을 가진 환자가 가지고 있는 난소암 바이오마커를 토대로 걸릴 수 있는 질병들을 예측하거나 다른 질병들에 대한 예방을 할 수 있지 않을까 하기 때문이다. 이를 생물학적이 아닌 텍스트 문서에서도 도출이 가능한지에 대하여 살펴보기 위해 워드임베딩을 이용하여 난소암 마커 말뭉치와 특정한 질병과의 유사도 계산을 한 결과 값을 기준으로 상위 20쌍을 선출한다.

[표 2]는 유사도 상위 20쌍의 결과 중 5개만 선출한 결과표이다.

[표 2] 유사도 상위 5개의 질병-마커 쌍

1) https://en.wikipedia.org/wiki/Biomarker

질병 이름	바이오 마커	유사도
암	HE4	0.99934
선암	HE4	0.99927
치매	Cortisol	0.99915
구내염	CRP	0.99889
치매	CRP	0.99827

[표 2]에 있는 유사도 상위 20쌍의 질병과 마커가 실제로 관련이 있으며 연구가 되고 있는지에 대해 관련 연구를 구글 스콜라를 통해 검색한다. 검색 할 때, 질병 이름과 바이오 마커가 검색 키워드가 되며 구글 스콜라 검색 결과 상위 20개의 논문들의 제목과 요약부분에서 질병이름과 바이오 마커가 얼마나 나오는지 확인하고 이를 점수화한다. 구글 스콜라에서 상위 20개의 논문을 가지고 평가하는 이유는, 논문 검색 시에 인용이 많이 되어진 순서로 나오기 때문에 많은 연구자들이 보았던 논문이기 때문에 상위 20개의 논문을 기준으로 평가를 하였다. 만약 질병 이름과 바이오 마커가 논문의 요약에 같이 나오는 경우, 두 단어 사이에 얼마나 많은 단어가 존재하는지도 확인한다. 이와 같이 유사도 상위 20쌍의 검색 결과를 [표 3]과 같이 정리한다.

질병이름	바이오마커	유사도	words	Score_title	Score_abst
암	he4	0.99934	1	17.32	102.53
선암	he4	0.99927	1	12.73	83.64
치매	cortisol	0.99915	2	17.49	66.51
구내염	crp	0.99889	6	8.94	35.99
치매	crp	0.99827	1	10.91	72.66
치매	leptin	0.99751	2	10.95	87.46
폐렴	myoglobin	0.99562	5	10.82	34.64
고혈압	leptin	0.99493	1	15.87	77.63
고혈압	prolactin	0.99259	1	13.49	74.46
고혈압	il-6	0.99230	1	15.43	54.12
위염	crp	0.99203	2	9.49	79.37
구내염	cortisol	0.99183	8	10.95	27.66
종양	ca125	0.99157	1	16.43	112.98
종양	cea	0.98999	0	12.73	159.05
치매	prolactin	0.98292	1	11.22	37.34
고혈압	myoglobin	0.97864	8	10.25	36.06
뇌수막염	crp	0.97791	1	14.87	110.89
치매	il-6	0.97592	4	13.08	70.70
천식	egfr	0.97468	2	12.49	59.90
종양	egft	0.97095	0	12.33	150.40

[표 3] 특정 질병과 바이오 마커의 유사도 20쌍의 검색 결과

[표 3]의 Score_title과 Score_abst는 각각 구글 스콜라에 나온 검색 결과를 가지고 상위 20개의 논문의 제목과 요약부분에 질병-마커 쌍이 각각 등장하는 횟수와 질병-마커가 모두 등장한 것을 기준으로 점수화 한 것이다. Words는 요약부분에 질병이름과 마커가 모두 나온 경우, 한 문장 안에서 질병-마커 사이의 등장하는 단어의 개수를 나타낸다.

이와 같이 실험을 하였을 때, 20쌍 중 85%의 쌍이 실제적으로 연구가 이루어지고 있으며, 나머지 15%의 쌍의 경우는 실질적으로 연구가 잘 되고 있지 않다는 것을 확인 할 수 있다.

6. 결론

본 논문에서는 난소암 마커가 출현한 PubMed 바이오메디컬 도메인의 제목과 요약 부분을 가지고 말뭉치를 구축하여 특정한 질병과 난소암 마커 간의 관계를 워드 임베딩을 사용하여 알아본다. 그 결과 나온 벡터 값들을 t-SNE를 사용하여 2차원으로 줄이고, 그 결과 값을 코사인 유사도를 사용하여 각각의 유사도를 계산한다. 유사도 계산 결과 상위 20개의 질병과 바이오 마커가 실질적으로 연구가 이루어지고 있는지 확인한다. 구글 스콜라에 질병과 마커를 키워드로 넣어 상위 20개의 논문의 제목과 요약 부분에 등장한 횟수를 점수화 하여 이를 확인한다.

실험 결과 5장에서의 [표 3]처럼 유사도가 높은 20쌍 중에서 85%의 쌍은 실제로도 많은 연구가 질병과 바이오 마커간의 관계를 파악하는 방향으로 진행중이나, 나머지 15%의 쌍에 대해서는 실질적인 연구가 잘 이루어지고 있지 않은 것으로 파악되었다.

참고문헌

[1] Feng, Qinghua, Mujun Yu, and Nancy B. Kiviat. "Molecular biomarkers for cancer detection in blood and bodily fluids." Critical reviews in clinical laboratory sciences 43.5-6 (2006): 497-560.

[2] Srinivas, Pothur R., et al. “Proteomics for cancer biomarker discovery.”, Clinical chemistry 48.8 (2002): 1160-1169.

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

[4] 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

[5] Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE."Journal of Machine Learning Research 9.Nov (2008): 2579-2605.

[6] Zhao, Ying-Yong, et al. "Lipidomics applications for disease biomarker discovery in mammal models." Biomarkers 9.2 (2015): 153-168

[7] Li, Lihua, et al. "Data mining techniques for cancer detection using serum proteomic profiling." Artificial intelligence in medicine

32.2 (2004): 71-83
[8] Sajda, Paul. "Machine learning for detection and diagnosis of disease." Annu. Rev. Biomed. Eng. 8 (2006): 537-565.
[9] Nam, Kyeong-Min, et al. "Find Alternative Biomarker via Word Embedding."The 4th International Conference on Artificial Intelligence and Application. 2015.
[10] Mikolov, Tomas, Chen, K., Corrado, G., Dean, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301. 3781 (2013).
[11] Weenink, David. "Canonical correlation analysis." Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam. Vol. 25. 2003.
[12] M.K.Jang, Y.S.Kim, C.Y.Park, H.J.Song and J.D.Kim, “Integration of Menopausal Information into the Multiple Biomarker Diagnosis for Early Diagnosis of Ovarian Cancer”, International Journal of Bio-Science and Bio-Technology, vol. 5, no. 4, (2013), pp. 215-222.