구글 학술 검색 기반의 질병과 바이오마커 관계 분석

오병두○, 김유섭

한림대학교, 융합소프트웨어학과
iambd822@gmail.com, yskim01@hallym.ac.kr

Relation Analysis of Disease and Biomarker based on Google Scholar

Byoung-Doo Oh○, Yu-Seop Kim
Hallym University, Convergence Software

요 약

본 논문에서는 구글 학술 검색 기반의 데이터를 이용하여 질병과 폐질환과 관련된 바이오마커 단어의 유사도를 계산하는 방법을 제안한다. 질병과 바이오마커의 유사도를 계산할 때, 각 단어의 구글 학술 검색의 검색 결과를 이용하였다. 이를 통해 폐질환 관련 바이오마커와 다른 질병간의 관계를 파악하고자 하며, 의료 전문가에게 폐질환 관련 바이오마커와 다른 질병간의 새로운 관계를 제시하고자 한다. 이러한 데이터를 이용하여 계산한 결과, Wor2Vec의 결과를 이용한 코사인 유사도의 결과와 상관 계수가 약 0.64로 상당히 높은 상관 관계를 확인할 수 있었다. 따라서 이 방법을 통해 질병과 바이오마커의 관계를 파악하고자 하였다. 또한 Word2Vec을 이용한 질병과 바이오마커 단어의 벡터 값과 단어 유사도 계산 방법의 결과를 이용한 Deep Neural Networks (DNNs) 모델을 구축하고자 하며, 이를 통해 자동적으로 유사도를 분석하고자 하였다.

주제어: 질병, 바이오마커, 단어 유사도 분석, Deep Neural Networks (DNNs)

1. 서론

바이오마커는 일반적으로 단백질이나 DNA, RNA 등을 이용해 신체 내부의 변화를 알아낼 수 있는 지표를 의미한다. 이러한 바이오마커는 일반 생물처리 과정, 질병을 유발하는 과정, 그리고 질병을 치료하기 위한 약리학의 과정을 측정하거나 평가하는 부분에도 쓰이게 된다.[1] 바이오마커를 발굴할 때에는 대부분 임상적인 실험 또는 연구를 통해 발굴하게 된다. 이러한 임상적인 실험은 다양한 단계(또는 과정)를 통해 진행되어 많은 시간과 비용을 소모하게 되며, 원하는 결과를 얻지 못할 수도 있다.

기존부터 자연어처리 기술을 이용한 단어 관계 분석이 이루어졌다. 그 중 Information Content를 통한 단어 관계 분석은 꾸준히 연구되고 있는 분야이다. Information Content는 코퍼스에서 단어의 확률을 사용하는 방법이다. 이 방법은 연구마다 다양한 코퍼스의 단어 확률을 계산하는 방법들을 제시하고 있다. 또한 워드 임베딩은 문서에서 단어의 관계를 계산하여 각 단어들을 벡터로 표현해주는 방법이다. 비슷한 단어들은 유사한 벡터로 표현되었으며, 이를 통해 비슷한 의미를 가진 단어들을 알아낼 수 있었다. 또한 워드 임베딩의 결과인 벡터를 이용하여 코사인 유사도 계산을 통해 단어의 관계를 계산할 수 있다. 이러한 방법들을 이용하여 자연어처리 분야, BioNLP 분야 등에서 좋은 성능을 보였다.[1-3]

본 논문에서는 기존의 단어 간의 유사도 측정보다 단어에 대한 구글 학술 검색의 결과에 기반한 단어 간의 유사도를 측정하는 방법을 논한다. 질병은 암을 포함한 37가지의 질병을 선정하였고, 바이오마커는 폐질환과 관련되어 있다고 알려진 27가지의 바이오마커를 선정하였다. 이를 통해, 폐질환과 관련된 바이오마커 중 폐질환이 아닌 다른 질병과의 새로운 관계를 파악하고자 한다. 따라서 의학 분야의 전문가들에게 질병과 폐질환 관련 바이오마커의 기존에 알려지지 않은 관계에 대해 제시하고자 한다. 질병과 바이오마커의 유사도 방법을 계산할 때, 워드 임베딩 방법 중 하나인 Word2Vec[4]의 결과를 통해 계산한 코사인 유사도의 결과와의 상관관계를 비교하였다. 이 때, 상관계수는 약 0.64의 결과를 얻을 수 있었다. 또한 이러한 계산 방법을 통해 나온 결과를 DNNs를 이용한 학습 모델을 구축하여 다른 질병과 폐질환 관련 바이오마커의 관계를 분석하는데 특화된 학습 모델을 만들고자 한다.

2. 관련 연구

[5]는 워드 임베딩을 이용하여 질병과 미생물의 관계를 분석하였다. 이 연구에서는 워드 임베딩 방법 중 하나인 CCA (Canonical Correlation Analysis)를 이용해 문서의 단어들을 벡터화하고, 코사인 유사도를 계산하여, t-SNE를 통해 2차원으로 만들어 질병과 미생물들의 관계를 분석하였다. 이를 통해, 질병과 미생물들의 관계를 제시하였다.

[6]에서는 WordNet의 특성(경로의 길이, 단어의 깊이)을 이용해 두 단어의 유사도를 측정할 수 있는 방법을 활용하였다. 이러한 유사도 측정은 WordNet에서 단어의 깊이와 최단 경로를 균형 있게 조정할 수 있는 장점을 가지고 있다. 따라서 단어의 Information Content를 활용하여, 두 단어쌍의 동일한 경로와 깊이의 유사점을 찾

[1] https://ko.wikipedia.org/wiki/바이오마커

아낼 수 있는 similarity metric을 제안하였다.

3. 방법론

3.1 데이터

본 논문에서는 구글 학술 검색에서 단어를 검색한 결과의 개수를 이용하였다. 검색한 단어는 질병과 바이오마커를 선택하였다. 이를 통해, 해당 질병과 바이오마커의 유사도를 계산하고자 하였다.

또한 PubMed의 문서 데이터를 활용하여 Word2Vec의 결과를 얻었고, 그 결과를 이용해 코사인 유사도를 계산하고, DNNs 모델을 만들 때 입력 데이터로 사용하였다. 질병과 바이오마커의 종류는 다음의 표1과 같다.

표 1. 실험에 사용한 질병 및 바이오마커 종류

질병 (37 개)	바이오마커 (27 개)
만성폐쇄성폐질환, 천식, 부정맥, 심부전, 심근경색, 심근증, 뇌경색, 고지혈증, 동맥경화, 신부전, 신장결석, 당뇨병, 갑상샘과다증, 백혈병, 간경변, 고혈압, 유방암, 자궁경부암, 위암, 대장암, 폐암, 피부암, 간암, 자궁암, 전립선암, 췌장염, 백색증, 골연골이형성증, 혈우병, 신경섬유종증, 뇌종양, 혈소침착증, 식도암, 후두암, 쓸개암, 고환암	CC-16, rbp, cea, asph, Calprotectin, saa, sp-d, Igfbp-2, Endoglin, Endostatin, trail, Cyfra21-1, Ghrelin, Leptin, nse, pai-1, Angiostatin, ip-10, Adiponectin, il-10, Paraoxonase, C9, Eotaxin-1, dr5, ldl, Nf-kb, il-2

3.2 Word2Vec

Word2Vec[4]은 2013년, 구글의 Mikolov가 제안한 방법으로 인공신경망 모델을 기반으로한 방법이다. Word2Vec은 Skip-gram과 CBOW (Continuous Bag-of-words), 2 가지 모델이 있다. Skip-gram 모델은 하나의 단어를 통해 주변의 단어들을 예측하고, CBOW 모델은 주변의 단어들을 통해 해당 단어를 예측한다. 본 논문에서는 Skip-gram을 이용하여 PubMed의 문서 데이터에서 질병과 바이오마커에 대한 단어 벡터 값을 얻었다. Skip-gram 모델의 구조는 그림 1과 같다.

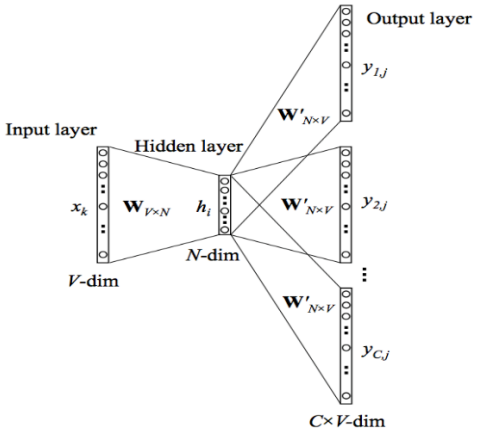

그림 1. Word2Vec의 Skip-gram 모델 구조

3.3 코사인 유사도

코사인 유사도는 내적 공간의 두 벡터간 각도의 코사인 값을 이용하여 측정된 벡터간의 유사한 정도를 측정하는 방법으로, 이 방법은 다차원의 공간에서의 유사도 측정에 자주 이용된다. 코사인 유사도는 두 벡터 X, Y의 내적, 벡터의 크기 등을 이용해 표현하게 된다. 본 논문에서는 코사인 유사도의 결과를 이용해 두 단어의 유사도 계산과의 상관관계를 비교하여 신뢰도를 측정하였다. 코사인 유사도의 계산은 다음의 (1)과 같다.

$$\text{similarity} = \cos\theta = \frac{X \cdot Y}{\|X\|\|Y\|} = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n}(X_i)^2} \times \sqrt{\sum_{i=1}^{n}(Y_i)^2}} \quad (1)$$

3.4 단어 유사도 계산

본 논문에서는 구글 학술 검색 기반의 데이터를 이용하여 계산을 하기 위한 계산 방법을 제안한다.

먼저 질병의 검색 개수(W_d)와 바이오마커의 검색 개수(W_m)를 더한 후, 질병과 바이오마커를 동시에 검색했을 때의 검색 개수(W_{d+m})를 나누어 준다. 이 식은 다음의 (2)과 같다.

$$p(W_d, W_m, W_{d+m}) = \frac{W_{d+m}}{W_d + W_m} \quad (2)$$

그 후, $p(W_d, W_m, W_{d+m})$의 결과에 상용 로그 함수를 취한다. 이 식은 다음의 (3)와 같다.

$$L(W_d, W_m, W_{d+m}) = log_{10}{}^{p(W_d, W_m, W_{d+m})} \quad (3)$$

마지막으로, $C(W_d, W_m, W_{d+m})$의 결과를 시컨트(SEC) 함수를 이용하여 계산한다. 이 식은 다음의 (4)과 같다.

$$f(W_d, W_m, W_{d+m}) = sec\, L(W_d, W_m, W_{d+m}) \quad (4)$$

이러한 3 단계의 계산을 통해 질병과 바이오마커의 유

사도를 분석하였다.

3.5. Deep Learning

Deep Neural Networks (DNNs)는 인공 신경망 기술로, 데이터 기반의 예측 방법 중 하나이다. Deep Learning은 현재 자연어처리, 컴퓨터비전 등 다양한 분야에서 좋은 성능을 보이고 있다. 이 방법은 데이터의 양이 많을수록 좋은 성능을 보이며, 또한 매개변수들을 어떻게 설정하는지에 따라 다양한 성능을 보이게 된다. Deep Learning에는 다양한 알고리즘이 존재한다. 본 논문에서는 그 중 DNNs를 이용하였다. 본 논문에서는 DNNs의 구조는 그림 2와 같다.

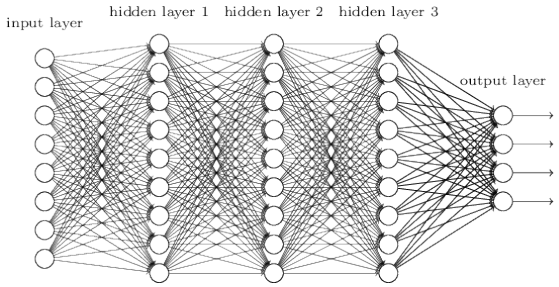

그림 2. DNN의 구조 예시

4. 실험 및 결과

4.1 실험

본 논문에서는 위와 같은 단어의 유사도를 계산하기 위해 질병과 바이오마커의 단어를 Word2Vec을 이용하여 단어에 대한 벡터 값을 얻었고, 이 결과를 이용해 코사인 유사도를 먼저 계산하였다. 그 후, 우리는 코사인 유사도의 결과와 구글 학술 검색 기반의 계산 결과에 대한 상관관계를 계산하여 위와 같은 계산 방법을 선택하였다. 이 때, COPD(만성폐쇄성폐질환)과 27 개의 바이오마커를 선택하여 코사인 유사도를 계산하였고, 이 결과와의 상관관계를 계산하였다. 상관관계를 계산할 때, 코사인 유사도와 3.4의 유사도의 결과에 대한 Rank를 각각 구하였고, 이러한 Rank에 대한 상관관계를 계산하였다. 그 예는 표 2과 같다.

표 2. Rank 상관관계 계산의 예

단어 쌍 (COPD)	코사인 유사도	Rank	계산 결과	Rank
CC-16	0.912140484	1	1.0294	9
SP-D	0.85229974	7	1.2003	16
Leptin	0.76273251	14	1.4872	23
C9	0.618226102	22	4.6947	26

이 때, 코사인 유사도와 단어 유사도 계산과의 상관계수 결과는 약 0.64로 상당히 높은 결과를 얻었다. 따라서 우리는 이러한 계산 방법을 이용하여 유사도를 계산해보고, 이를 기반으로 하여 DNNs 도구 중 하나인 TensorFlow를 이용해 이러한 결과를 학습한 모델을 구축하고자 하였다.

DNNs 모델을 구축할 때, 입력 데이터는 질병과 바이오마커의 단어의 Word2Vec을 이용한 벡터값으로 지정하였다. 그리고 출력 데이터는 계산 결과로 지정하였다. Training data는 질병 29 개, 바이오마커 27로 총 783 개의 데이터를 사용하였다. 그리고 Test data는 질병 8 개, 바이오마커 27개로 총 216 개의 데이터를 사용하였다.

4.2 실험 결과

본 논문에서는 구글 학술 검색 기반에 질병과 바이오마커 단어의 유사도를 계산할 때, 37 가지의 질병 단어와 27 가지의 폐질환과 관련된 바이오마커 단어의 유사도를 계산하였다. 이를 통해 기존에 알려지지 않은 질병과 폐질환 관련 바이오마커와의 관계를 파악하고자 하였다.

계산 방법은 3.4의 계산 방법을 통해 계산하였으며, 폐질환과 관련되지 않은 3 가지의 질병에 대한 계산 결과를 예로 제시한다. 그 예는 다음의 표 3와 같다.

표 3. 3 가지 질병의 유사도 결과 상위 4 가지

질병	바이오마커	유사도 결과
부정맥	ASPH	3.636684
	C9	2.546032
	TRAIL	-1.000802
	SP-D	-1.011931
동맥경화	SAA	34.982258
	Paraoxonase	22.201981
	I1-10	14.631821
	Adiponectin	3.945321
당뇨병	DR5	-1.004244
	SP-D	-1.005012
	Eotaxin-1	-1.011437
	CC-16	-1.017126

또한 질병과 바이오마커 단어의 벡터 값을 입력 데이터로 선정하고, 질병과 바이오마커의 단어에 대한 3.4의 계산 결과를 출력 데이터로 하여 TensorFlow를 이용해 DNNs 모델을 만들고자 하였다. 이를 통해, 3.4의 계산방법을 자동적으로 할 수 있는 모델을 만들고자 하였다. 이 때, 입력 데이터는 질병과 바이오마커의 단어로 각각 2 차원의 벡터, 5 차원의 벡터, 10 차원의 벡터로 총 3 가지의 데이터셋을 구축하였다. 이러한 3 가지의 데이터셋을 가지고 TensorFlow를 통해 어떤 차원의 수를 가진 데이터셋이 학습이 더 잘되는지 실험하였다. Training data로 학습을 한 후, Test data를 통해 예측한 값과 Test data의 정답과의 상관관계를 계산하여 성능을 평가

하였다. 이 때, 각 실험 중 가장 좋은 성능은 다음의 표 4와 같다.

표 4. DNN 모델의 Hyper parameter 및 성능

	2 차원	5 차원	10 차원
Num of hidden layer	5	3	5
Batch size	30	전체	전체
Num of node	80	10	130
Num of epoch	501	501	501
상관 계수	0.1933	0.1951	0.21

실험 결과, 코사인 유사도와 단어 유사도 계산의 상관관계만큼의 성능을 얻을 수 없었다. 3.4의 계산 결과와 예측된 값의 상관 계수는 약 0.2로 상당히 낮은 성능을 보였다.

5. 결론 및 향후 연구 방향

본 논문에서는 구글 학술 검색 기반의 질병과 바이오마커의 유사도를 계산하고자 하였다. 이 때, 구글 학술 검색 기반의 단어 유사도 계산방법은 단어의 벡터 값을 이용한 코사인 유사도의 결과와의 상관 계수가 약 0.64로 상당히 좋은 결과를 얻을 수 있었다. 이러한 문서에서의 질병 단어와 바이오마커의 유사도 계산을 통해 의료 전문가들이 어떠한 질병에 대한 새로운 영향을 가진 바이오마커를 제시하여 의학 분야에 도움이 될 것이라 판단된다. 그러나, DNN의 학습을 통한 예측 성능에서는 약 0.2로 낮은 성능 결과를 얻었다.

현재는 해당 질병과 바이오마커에 대한 단어 벡터값과 유사도 계산 결과를 학습하여 모델을 구축하려하였다. 그러나 향 후, 단어 벡터값과 유사도 계산을 학습하는 것이 아닌 질병과 바이오마커의 단어를 통해 자동적으로 유사도 계산이 가능한 질병과 바이오마커의 관계 분석에 특화된 모델을 만들고자 한다.

참고문헌

[1] T. Pedersen, Information content measures of semantic similarity perform better without sense-tagged text." Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010.

[2] Muneeb, T.H., Sahu, S.K. and Anand, A., Evaluating distributed word representations for capturing semantics of biomedical concepts, Proceedings of ACL-IJCNLP, 158, 2015.

[3] T. Slimani, Descirption and evaluation of semantic similarity measures approaches, arXive preprint arXiv:1310.8059, 2013.

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

[5] 윤영신·김유섭, “워드 임베딩을 이용한 미생물 관계 분석”, 한국정보과학회 학술발표 논문집, pp.461-463, 2016.

[6] Atoum, I., Bong, C.H., Joint Distance and Information Content Word Similarity Measure, Soft Computing Applications and Intelligent Systems SE-22, 378, 257-267, 2013.