의생명과학 기반 기학습된 워드 임베딩을 이용한 의생명과학 논문 속의 돌연변이-약물 관계 추출 연구

김호준[1†] , 원성연[1†] , 강승우[3], 이규범[3], 김병건[3], 김선규[3], 강재우[3‡]
고려대학교 생명공학부[1], 생명과학부[2], 컴퓨터공학과[3]
e-mail: kangj@korea.ac.kr‡

Research on Identifying Mutation-Drug Relationship in Biomedical Literature Using Biomedical Context based pre-trained word embedding

Hojun Kim[1†] , Seongyeon Won[1†] , Seungwoo Gang[3], Kyubum Lee[3],
Byounggun Kim[3], Sunkyu Kim[3], Jaewoo Kang[3‡]
Dept. of Biotechnology[1], Korea University
Dept. of Life science[2], Korea University
Dept. of Computer science[3], Korea University

요 약

의생명과학분야가 계속 발전됨에 따라 매일 평균 3천여 편에 달하는 방대한 양의 의생명과학분야 문헌들이 나오고 있다. 많은 연구가 진행될수록, 새로이 규명된 관계를 습득하고 체계화하는 일이 연구자와 의료계 종사자들에게 더 중요해지고 있다. 하지만 현재로서는 의생명과학분야에 어느 정도의 지식이 있는 사람이 직접 논문을 읽고 해당 논문에서 밝히고 있는 정보를 정리해야만 하는 상황이며, 이로는 기하급수적으로 쌓이는 정보의 양을 대처하기 어렵다. 이를 해결하기 위해 본 논문에서는 기계학습을 통한 생명의료 객체관계 자동추출 연구를 이용하여 의생명과학분야의 정보를 체계화 하고자 한다.

본 논문에서는 돌연변이와 약물이 함께 등장하는 논문을 뽑아내어 글을 자연어 문장 단위로 나누었다. 추출한 돌연변이와 약물 간의 관계를 직접 사람에 의해 참거짓을 판명하였고, 해당 데이터셋을 기계학습에 이용하여 돌연변이와 약물 간의 관계를 학습시켰다. 최종적으로 GoogleNews의 기사들로 기학습된 워드임베딩, 의생명과학분야 문헌들을 이용하여 기학습된 워드임베딩을 이용하여 학습의 성능을 비교하였고, 의생명과학-문맥 특이적인 워드임베딩이 갖는 강점을 보고한다.

해당 연구를 통해 실제로 논문을 읽지 않고도 의생명과학분야 논문의 핵심적인 내용을 뽑아내는 자동화 시스템을 구축하는 데에 이바지하고, 의생명공학 연구자들의 연구에 핵심적인 도움이 되는 디딤돌이 되고자 한다.

1. 서론

21세기에 들어서며 인간 게놈 프로젝트의 성공, 많은 질병의 진단과 치료 기술의 개발 등에 힘입어 의생명과학분야의 연구는 가속되고 있다. PubMed 기준으로 2016년 한 해에만 총 1,210,896 편, 하루에 평균 약 3,300여 편에 달하는 생명의료분야 논문이 출판되었다.[1] 하지만 많은 양의 논문들이 쏟아져 나오고 있는 만큼, 논문이 담고 있는 정보를 인간이 직접 읽고 분석하는 것에는 시간적, 경제적 제약이 뒤따르고 있다. 이 같은 상황에서 연구를 통해 학계에 새로이 보고된 객체(돌연변이, 질병, 약물 등) 간의 관계를 사람이 일일이 읽지 않더라도 자동으로 추출하는 기술이 요구된다. 의생명과학분야에서 돌연변이와 질병의 관계, 질병과 약물의 관계의 보고는 맞춤의학, 정밀의료시대를 앞두고 보다 중요한 역할을 하고 있다. 본 논문에서는 문헌 마이닝을 이용해 의생명과학분야 논문 속 돌연변이(Mutation), 유전자(Gene), 약물(Drug)의 객체 정보를 추출해냈다. 추출된 객체 간의 관계가 실제로 존재하는지 여부를 이후 의생명과학분야 전공자들의 도움으로 확인하여 기계학습 모델에 사용 가능한 참거짓 데이터셋을 만들어냈다. 본 논문에서 보고된 워드 임베딩(word embedding)을 활용한 자동화 추출 방법은 의생명과학 정보가 효율적으로 활용될 수 있도록 함과 동시에 세계의 많은 의생명공학 연구자들을 도움으로써 데이터셋 분석과 의생명공학 분야 발전에 이바지 하고자 한다.

..... A small subset of BCR-ABL mutant clones have been associated with impaired responses to second-generation TKIs (tyrosine to histidine mutation at codon 253 [Y253H] .... for nilotinib .... neither nilotinib nor dasatinib is active against the threonine to isoleucine mutation at codon 315 (T315I). .....

PMID	돌연변이	약물	관계 유무
25774163	Y253H	nilotinib	1
25774163	T315I	dasatinib	-1
...			

그림 1 본 연구의 모식도

2. 배경 및 관련 연구

앞서 언급한 방대한 양의 의생명과학 문헌들을 정리하기 위해 이미 학계에선 문헌 마이닝을 통해 문헌들 내에 등장하는 특정 객체들의 관계를 찾아내는 연구가 진행되고 있다. 이러한 연구들은 주로 약물이나 질병, 돌연변이 등을 다루는 의생명과학 문헌에서 해당 객체들을 추출하여 추출된 객체들 간의 관계성을 입증하는 데에 초점을 두고 있다.

대표적인 예로는 Polysearch라는 프로그램을 들 수 있다.[2] Polysearch에 관심을 갖고 있는 한 가지 객체를 입력하면 그 객체와 관련된 -유전자나 단백질, 약물 등의- 모든 가능한 후보들이 나타난다. 예를 들면 유방암과 관련된 모든 '유전자'라는 키워드로 검색하도록 명령하면 실제로 문서 내에서 유방암과 관련된 유전자들이 검색 결과로 도출된다. 하지만 Polysearch의 원리가 문헌들 내에서 객체 간의 관계를 단지 높은 빈도로 판단하여 도출하는 데에 있다는 점에서 직접적인 관계가 없는 객체를 보고할 수도 있는 한계를 지니고 있다. 또한 Polysearch에 의해 보고되는 관계는 빈도에 의한 것이므로, 두 객체간의 어떤 관계인가에 대한 정보를 얻을 수 없다는 한계가 존재한다.

3. 연구 방법

이와 같은 문제를 해결하기 위해 우리는 객체간의 관계, 특히 돌연변이(Mutation)와 약물(drug) 간 특정 관계의 유무를 판단하는 기계학습 시스템을 만들어 보았다. 단순히 문헌에서의 출현빈도로 객체들의 관계를 예측하는 것이 아닌, (기계학습을 통해) 해당 문헌의 문맥을 반영하여 보다 높은 정확도로 객체간 관계를 예측하는 시스템을 만들었다.

1) Text Acquisition & Named Entity Recognition(NER)

먼저 논문의 의생명과학분야 논문 검색 사이트인 PubMed에서 돌연변이(Mutation)와 약물(Drug) 객체가 포함되어 있는 문서의 PMID를 Pubtator[3]를 이용하여 추출하였다. 이후 Stanford CoreNLP Sentence Parser[4]를 이용하여 글을 문장 단위로 나누었다.

2) Drug NER & 데이터셋 Generation

다음으로 BEST Entity Extractor[5]를 이용하여 앞서 추출된 문장 중 돌연변이와 약물이 동시에 등장하는 문장들만 추려내었다. 추출된 이들 문장 내에서 등장한 두 객체인 돌연변이와 약물이 실제로 관계가 있는지 없는지는 의생명과학 전공자들로 구성된 큐레이터(Curator)들이 직접 읽고, 돌연변이와 약물이 서로 관계가 있다면 +1, 관계가 없다면 -1, 알 수 없거나 상관이 없다면 0으로 표기하여 객체들의 관계를 분류하였다. 이후 본 단계에서 얻어진 +1 문장은 Positive Relation Sentence 데이터셋으로, -1 문장은 Negative Relation Sentence 데이터셋으로 기계학습에 사용되었다.

데이터셋의 양을 늘리기 위해 공개된 돌연변이-약물 간의 관계를 PharmGKB[6]에서 다운받았다. 공개된 돌연변이-약물 간의 관계를 규명하고 있는 문헌을 Stanford CoreNLP Sentence Parser를 이용하여 글을 문장 단위로 나누었다. 위 문장 중 돌연변이와 약물이 동시에 등장하는 문장을 추려내어 Positive Relation Sentence 데이터셋으로 추가하였다.

3) CoreNLP Dependent Parsing Tree

한 문장 내에는 두 객체 간의 관계를 추출하는데 도움이 되지 않는 많은 부분이 존재한다. 2)번 단계에서 추출된

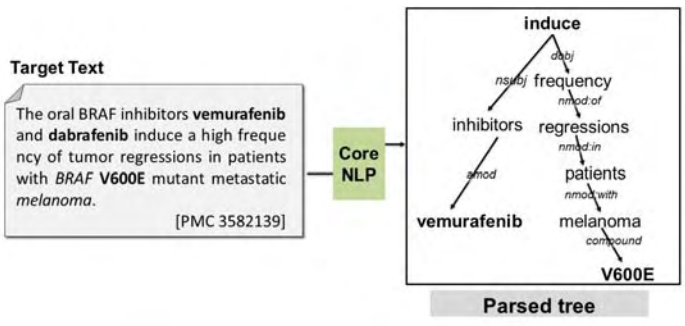

그림 2 문장에서 추출된 Parsed tree

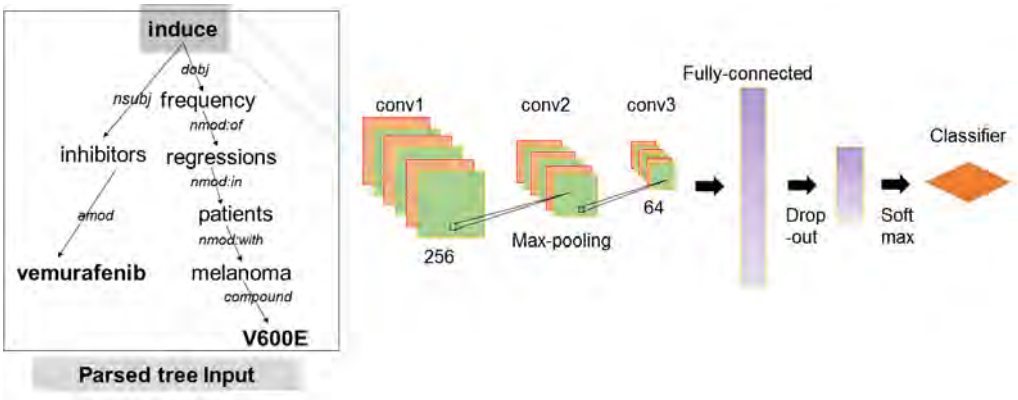

그림 3 CNN의 모식도

문장들은 기계학습의 데이터셋으로 사용되기 전에, 학습 효율을 높이기 위하여 그림 2처럼 Parsed tree 형태로 바꾸었다. 이로써 한 문장 내에서 돌연변이와 약물 간의 관계를 기술하는 주요 부분만을 추출하였다.

4) 워드 임베딩을 이용한 CNN방식의 기계학습

Convolutional Neural Network(CNN)는 주어진 데이터셋의 국소적인 특징을 이용해 일종의 깔때기와 같은 필터를 학습하는 기계학습 기법 중 하나이다.[7] 본 기법은 최초 컴퓨터의 이미지 처리를 위해서 만들어졌지만, 21세기에 접어들면서 자연어 처리의 다양한 영역에서 뛰어난 효과를 보이고 있는 추세이다. 우리는 Yoon Kim의 2014년 논문 코드를 기반으로 워드 임베딩이 가능하도록 부가적인 코드[8] 를 인용했고 이를 이용해 다양한 종류의 워드 임베딩을 이용하여 학습을 시킨 결과를 비교했다.

워드 임베딩은 특정 단어의 언어학적 문맥을 학습하기 위해서 사용되는 벡터이다. GoogleNews word2vec[9]은 Mikolov et al[10]이 2014년도에 보고한 것으로, 구글 뉴스에 등장하는 1,000억개 길이의 문서를 통해 학습시킨 워드 임베딩이다. PubMed는 의생명과학 주제에 대한 참조 및 요약을 담고 있는 MEDLINE 데이터베이스를 접근할 수 있게 해 주는 검색 엔진이다. PubMed 워드 임베딩은 PubMed에 등장하는 문헌으로 기학습된 워드 벡터로서, Google News 분야 문헌으로 학습시킨 GoogleNews word2vec과는 달리 의생명과학분야의 문맥을 지니고 있다. 본 연구에서는 워드 임베딩을 사용하지 않는 방식과 GoogleNews word2vec, PubMed 워드 임베딩 2 종류의 워드 임베딩을 사용하여 기계학습에 사용하였고 그 성능을 비교해보았다.

4. 성능 평가

1) 실험 과정

앞서 3-2)에서 만든 Manual Curation 데이터셋 중 300여개의 Positive Relation Sentence 데이터셋과 PharmGKB에서 얻어진 돌연변이-약물 관계를 토대로 만들어낸 Positive Relation Sentence 데이터셋을 모두 합하여 3,417개의 Positive Relation Sentence 데이터셋이 학습에 사용되었다. Negative Relation Sentence 데이터셋은 4,352개가 사용되었고 총 7,769개의 데이터셋이 학습에 사용되었다. CNN의 Filter의 크기(window)는 각각 3,4,5이고 각 필터마다 128개의 Feature map을 가지고 있으며 Dropout 비율은 0.5로 설정되었다.

2) 실험 결과

총 5회의 학습 후 얻어지는 정확도(accuracy) 의 평균을

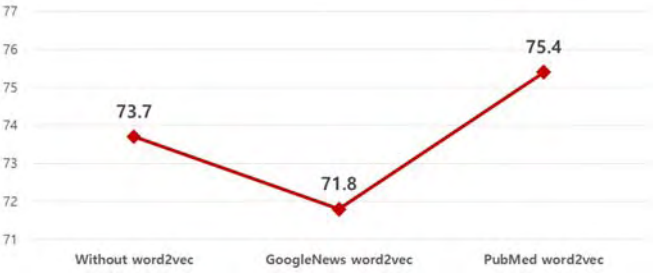

그림 4 서로다른 문서로 기학습된 워드 임베딩에 따른 정확도 비교

통해 각 기학습된 워드 벡터별로 성능을 비교하여 보았다. 그림 4와 같은 결과를 얻을 수 있었다.

5. 결론

특정 분야의 Context를 토대로 기학습된 워드 벡터를 사용하게 되면 자연어 처리의 기계학습 시에 정확도가 높아지는 것을 확인하였다. 본 연구에서는 이러한 예측대로 의생명과학 문맥을 지닌 PubMed 워드 임베딩의 정확도가

가장 높은 것을 확인하였다. 한편 전문적인 분야를 연구하는데 있어서 context를 반영하지 않은 워드 임베딩의 사용은 1,000억 개에 달하는 많은 학습량에도 불구하고 워드 임베딩을 사용하지 않았을 때보다도 정확도에 있어서 나쁜 결과를 낼 수 있다는 사실을 확인하였다. 이로써 워드 임베딩을 이용한 자연어 처리에서 워드 임베딩의 학습 데이터셋의 선정 또한 중요함을 알게 되었다.

Acknowledgement

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단 바이오.의료기술개발사업의 지원을 받아 수행된 연구임 (NRF-2016M3A9A7916996)

참고문헌

[1] https://www.ncbi.nlm.nih.gov/pubmed/

[2] http://polysearch.cs.ualberta.ca/index

[3] C.H. We et al, "PubTator: a web-based text mining tool for assisting biocuration." *Nucleic acids research* (2013)

[4] C.D. Manning et al, "The Stanford CoreNLP Natural Language Processing Toolkit" *ACL* (2014)

[5] Lee, Sunwon, et al. "BEST: next-generation biomedical entity search tool for knowledge discovery from biomedical literature." *PLoS One* 11.10 (2016)

[6] https://www.pharmgkb.org/downloads/

[7] Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).

[8] https://github.com/dennybritz/cnn-text-classification-tf

[9] https://code.google.com/p/word2vec/

[10] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." A*dvances in neural information processing systems.* 2013.

[11] LeCun, Yann A., et al. "Efficient backprop." *Neural networks: Tricks of the trade. Springer Berlin Heidelberg,* 2012. 9-48.

[12] Yih, Wen-tau, Xiaodong He, and Christopher Meek. "Semantic Parsing for Single-Relation Question Answering." *ACL* (2). 2014.