개체명 인식을 이용한 소셜 미디어에서의 약물 부작용 표현 추출 및 분류

정현정, 김현희
동덕여자대학교 정보통계학과
hyeonjeong216@gmail.com, heekim@dongduk.ac.kr

Detecting and classification ADRs using Named Entity Recognition on social media

Hyeon-jeong Jeong, Hyon Hee Kim
Department of Statistics and Information Science,
Dongduk Women's University

요 약

의약품에 대한 안전성 정보 수집과 관리는 온라인, 오프라인을 통해 약물 이상 사례를 보고받는 형태로 진행되고 있다. 하지만 소비자들의 자발적인 참여로 이루어지므로 실제 발생하는 약물 부작용보다 데이터가 현저히 적다는 단점이 존재한다. 본 논문에서는 약물 이상 데이터 희소성 문제를 해결 할 수 있도록 소셜 미디어에서 약물 부작용 표현을 찾을 수 있도록 하였다. 소셜 미디어의 경우에는 표준 약물 부작용 용어를 사용하기보다는 일반인들이 자연어로 표현한 경우가 많으므로 개체명 인식 기법을 이용해 부작용을 추출할 수 있는 모델을 개발하였다. 또한 추출된 부작용 표현을 표준 용어로 분류할 수 있는 모델을 제시하였다. 실험 결과 제안한 두 가지 모델은 0.9 이상의 정확도를 얻을 수 있었으며, 일반 사용자들이 자연어로 표현한 약물 부작용 표현을 효과적으로 찾아내고 표준 부작용 용어로 매핑할 수 있음을 보여준다.

1. 서론

의약품은 판매 전에 임상시험을 거치지만 관찰 기간이 제한되고 대상이 한정적이기 때문에 모든 약물 이상 반응을 파악하는 것은 불가능하다.[1] 한국의약품안전관리원에서 약물 이상 사례를 온라인, 오프라인을 통해 보고 받고 있지만 자발적인 참여로 이루어져야 하고 수가 매우 적다. 본 연구에서는 약물 후기에서 부작용 표현을 찾아내고 분류함으로써 약물 이상 데이터 희소성 문제를 해결하고자 한다.

BERT[2], Bio-BERT[3], ClnicalBERT[4] 등 다양한 BERT 모델을 사용하여 사용자의 트윗에서 의약품 부작용 표현을 탐지하는 연구[5]와 Bio-BERT, EnDRBERT[6]를 사용해 트윗에서 의약품 부작용을 탐지하고 분류기에 EnDRBERT을 훈련시켜 약물 부작용 표현을 MedDRA[7] 용어로 정규화 시킨 연구[8]가 진행되었다. 하지만 트위터 게시물은 140자로 제한되고 약물 이름이 언급된 게시물 중에 실제 약물 부작용 표현이 포함된 것은 1% 미만[9]에 해당하기 때문에 본 연구에서는 의료 포럼 웹사이트의 의약품 후기를 이용하였다.

본 논문에서는 의료 포럼 웹사이트의 의약품 후기를 활용해 개체명 인식 기반의 약물 부작용 표현 추출 모델을 제시하고 이를 바탕으로 표준 약물 부작용 용어 사전인 MedDRA 용어로 자동 매핑하는 모델을 개발하였다. 개체명 인식(Named Entity Recognition) [10]이란 텍스트 내에서 단어나 구를 사람 이름, 지명, 시간 등의 특정한 개체명으로 찾아내는 것을 말한다.

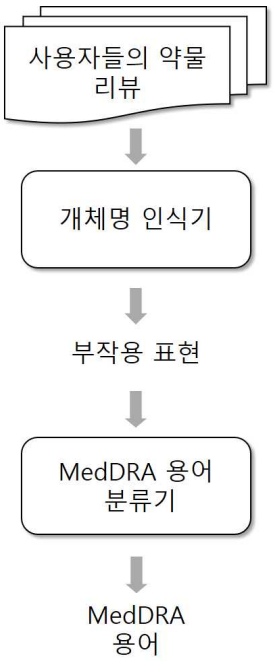
먼저 CADEC[11] 데이터를 개체명 인식 모델에 학습시킬 수 있도록 전처리와 태깅 작업을 진행하였다. 다음으로 BERT 임베딩을 적용한 BiLSTM-CNN-CRF [12] 구조의 개체명 인식 모델에 학습시켜 텍스트로부터 부작용에 해당하는 표현을 찾을 수 있도록 하였다. 그 다음은 찾아낸 부작용 표현들을 LSTM 구조를 이용해 알맞은 MedDRA 용어로 분류하는 모델을 만들었다.

제안하는 개체명 인식을 이용한 소셜 미디어 기반의 부작용 탐지 모델을 사용한다면 보고되는 사례 이외에도 더 많은 부작용을 관찰할 수 있을 것이다. 뿐만 아니라 딥러닝 기법으로 부작용을 추출, 분류

한다는 점에서 디지털 약물 감시에 활용될 수 있을 것으로 기대된다. 또한, 약물 분야에서 더 나아가 헬스케어 전반에서 부작용을 탐지하는 데 사용할 수 있을 것으로 기대된다.

본 논문은 다음과 같이 구성된다. 제2장에서는 본 연구에서 제시한 모델의 구조에 대해 설명한다. 제3장에서는 개체명 인식을 이용한 부작용 추출 모델에 대해 자세히 설명한다. 제4장에서는 추출한 표현을 MedDRA 용어로 분류하는 모델을 보이고, 제5장에서는 마지막으로 결론 및 향후 연구를 제시한다.

2. 약물 부작용 탐지 모델 구조



<그림 1> 약물 부작용 탐지 모델

<그림 1>은 제안하는 개체명 인식을 통한 약물 부작용 탐지 모델의 구조를 보여준다. 구조는 크게 개체명 인식기와 MedDRA 용어 분류기 두 부분으로 나눌 수 있다. 개체명 인식기는 약물 리뷰 전체로부터 부작용에 해당하는 표현을 찾는 역할을 한다. MedDRA 용어 분류기는 찾은 부작용 표현들을 알맞은 MedDRA 용어로 분류하고 매핑하는 모델이다.

소셜 미디어의 약물 후기에 특화된 개체명 인식기를 만들기 위해 약물 후기를 학습시켜 개체명 인식기를 만들었다. 학습 데이터로는 의약품에 대한 후기를 나누는 의료 포럼 홈페이지인 'AskaPatient'에서 수집한 1,250개의 사용자 리뷰로 구성된 CADEC 데이터를 사용하였다.

BERT 임베딩을 추가한 BiLSTM-CNN-CRF 구조를 이용해 부작용 표현을 찾아내는 개체명 인식기를 만들었다. 개체명 인식기를 사용하게 되면 약물 리뷰에서 부작용에 해당하는 표현들을 뽑아낼 수 있다.

다음으로는 추출한 표현들이 어떤 부작용에 속하는지 분류하기 위해 MedDRA 용어 분류기를 만들었다. MedDRA란 의약품에 대한 규제 정보를 공유할 수 있도록 국제의약품규제조화위원회(ICH)에서 개발한 표준화된 의약 용어집이다. MedDRA에 정의된 다양한 약물 부작용 중 CADEC 데이터에 가장 자주 등장한 부작용 20개를 선정하여 추출된 표현들을 20개의 부작용 표현으로 자동 매핑하도록 하였다.

3. 부작용 개체명 인식기

본 연구에 사용된 데이터인 CADEC은 CSIRO에서 만든 약물 리뷰 말뭉치로 약물에 대한 정보와 후기를 공유하는 웹사이트인 'AskaPatient'에서 수집한 1,250개의 약물 리뷰로 구성되어있다. 데이터에 약물 이름, 약물 유해 반응, 증상, 질병 등의 주석이 달려있는데, 본 연구에서는 약물 유해 반응 탐지를 목표로 하므로 약물 유해 반응의 주석만 사용하였다.

개체명을 인식하는 방법으로는 BIO 태깅을 이용하였다. 리뷰에서 부작용 개체명이 시작하는 부분은 Begin의 약자로 B-ADR 태그를, 부작용 개체명의 안쪽 부분에는 Inside의 약자로 I-ADR 태그를, 부작용에 해당하지 않는 단어는 O 태그로 설정하였다. 'Went to 20 mg and heart palpitations.' 라는 문장에서 heart palpitations는 약물 이상 반응에 해당하므로 각각 B-ADR, I-ADR의 태그를 달았고 나머지 단어들은 이상 반응에 해당하지 않기 때문에 O 태그를 달았다.

전체 1,250개의 리뷰 중 80%에 해당하는 1,000는 학습 데이터로 사용하였고, 나머지 250개는 테스트 데이터로 사용하였다. 모델 학습을 진행하기 전에 리뷰들을 소문자로 바꾸고 특수문자 제거 등의 전처리를 진행하였다.

부작용 개체명 인식기는 BiLSTM-CNN-CRF 구조를 사용하였고 크게 임베딩 층과 Bi-LSTM층 그리고 CRF층, 세 부분으로 나눌 수 있다.

임베딩 층에서는 단어를 벡터 값으로 변환하여 다음 층으로 전달할 수 있게 한다. 이 모델에서는 임베딩 층에 CNN (Convolutional Neural Network) 기반의 글자 표현과 BERT 워드 임베딩을 사용하였다. 양방향 LSTM (Bidirectional LSTM)은 기존의 LSTM에 역방향으로 데이터를 처리하는 LSTM을 추가한 것으로 단어의 앞, 뒤를 관계를 파악할 수 있다. 마지막 CRF 층에서는 예측한 태그들의 관계를 학습한 다음 가장 높은 확률을 가지는 태그를 출력하는 역할을 한다.

배치 사이즈는 8, 초기 학습률은 0.001, 학습률 감쇠 계수는 0.005로 설정하였다. early stopping을 적용한 결과 에포크가 9 일때 최적의 결과를 얻을 수 있었다. 실험 결과, 0.93의 정확도와 0.81의 F1-score를 얻었다.

추출된 부작용 표현
pain in my left leg
muscle pain from my neck & shoulders to my lower back
feel fatigued
severe intense left arm and shoulder pain
could not even walk
blurred vision
muscles were very stiff
severe stomach cramping

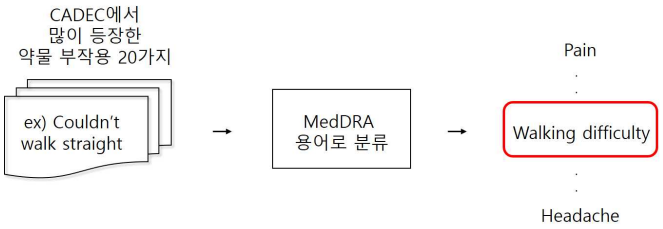
<표 1> 개체명 인식 모델로 찾은 부작용 표현

<표 1> 은 테스트 데이터에서 찾아낸 1,243개의 부작용 표현의 일부이다. ‘stomach cramping’ 같은 부작용 명이 포함된 표현뿐만 아니라 ‘muscles were very stiff’, ‘could not even walk’ 과 같은 정확한 부작용 명이 포함되지 않은 구 형태의 표현도 탐지가 가능했다.

4. MedDRA 용어 분류기

CADEC 데이터에서 가장 많이 등장한 약물 부작용 20가지를 사용하여 부작용 표현을 MedDRA 용어로 분류하는 분류기를 만들었다.

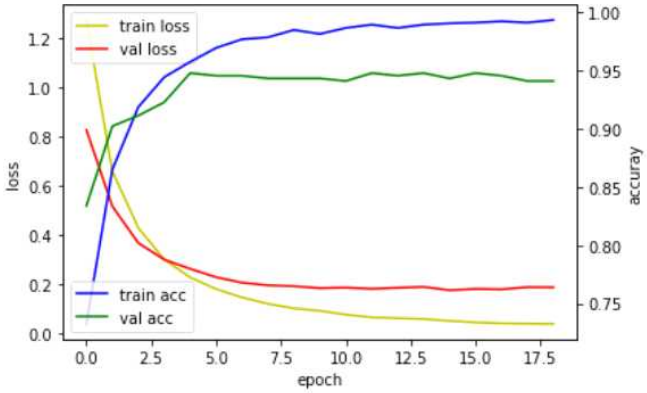
가장 많이 등장한 20가지 약물 부작용은 통증, 근육통, 피로, 우울증, 근육 경련, 속이 부글거림, 관절통, 두통 등이다.



<그림 3> MedDRA 용어 분류기 구조

CADEC 데이터에서 약물 이상 반응을 의미하는 표현과 그에 해당하는 MedDRA 용어를 훈련시켜 부작용 표현이 주어지면 그에 해당하는 MedDRA 용어로 분류하는 모델을 만들었다.

LSTM 구조를 기반으로 부작용 표현을 20가지 표준 용어로 분류하도록 하였다. 우선 임베딩 층에서 250개의 단어들을 100차원의 임베딩 벡터를 생성하도록 하였다. 그다음 100개의 뉴런을 LSTM 층으로 생성하고 dropout 비율은 20%로 설정하였다. 마지막 출력층에서는 20개의 부작용을 분류해야 하므로 20개의 뉴런을 사용하였다. 배치 사이즈는 64, optimizer는 adam, 활성화 함수는 softmax를 사용하였다. 총 2,192개의 표현 중 80%에 해당하는 1,753개의 표현은 훈련 데이터로 사용하였고 439개의 표현은 테스트 데이터로 사용하였다. Validation의 손실 값을 기준으로 Early Stopping을 적용하였고 에포크가 19일 때 최적 값을 얻을 수 있었다. Accuracy 값은 0.941로 수렴하였고 Loss 값은 0.187로 수렴하였다.



<그림 4> MedDRA 용어 분류기 성능 평가

<표 2>는 추출된 부작용 표현을 MedDRA 용어로 분류한 결과의 일부이다. 왼쪽의 부작용 표현은 일반 사용자들이 자연어로 많이 사용한 부작용 표현이며, 이러한 표현들은 표준 용어로 매핑되었다.

'pain in my lower leg'는 하체 통증에 매핑되었고, 구 형태의 'get tired lot' 과 couldn't walk도 각각 피로와 보행 장애를 뜻하는 MedDRA 용어로 정상적으로 매핑된 것을 확인할 수 있다.

부작용 표현	MedDRA 분류
pain in my lower leg	Pain of lower extremities
nagging muscle pain shoulder blades	Myalgia
gets tired lot	Tiredness
couldn't walk	Walking difficulty
arthritic type pain joints	Arthralgia
muscle weakness mouth area	Muscle weakness

<표 2> 부작용 표현을 MedDRA 용어 분류기로 분류한 결과

5. 결론 및 향후 연구

현재 약물 이상 반응 모니터링은 사용자들의 자발적인 보고로 이루어져야 한다는 단점이 있다. 본 연구에서는 보고가 이루어지지 않는 경우에도 소셜 미디어의 데이터로부터 약물에 대한 부작용을 탐지할 수 있는 모델을 제안하였다.

개체명 인식 기반의 부작용 탐지 모델은 BiLSTM-CNN-CRF 구조에 BERT 임베딩을 더하여 부작용 개체명 인식이 가능하게 하였다. 약물 부작용 시그널을 탐지한 이후에는 LSTM 기반의 분류 모델을 만들어 각각의 부작용 표현을 MedDRA 표준 용어로 매핑을 진행하였다.

본 연구는 딥러닝을 기반으로 약물 후기로부터 자동으로 부작용을 탐지한다는 점에서 약물 감시 데이터 희소성을 해결할 수 있을 것으로 보인다. 또한 약물 부작용 표현을 해당 의학 용어로의 분류를 통해 매핑을 자동화했다는 의의가 있다. 본 연구에서는 20 종류의 MedDRA 용어로 분류하였으며, 향후 부작용 용어를 확장할 계획이다. 또한 제안하는 모델을 코로나 치료제나 백신 등 다양한 약물에 적용하여 모델을 일반화하고 보다 빠르게 약물 부작용 시그널을 탐지하는데 기여하고자 한다.

참고문헌

[1] 윤나경, 강민구. "한국 의약품 부작용 보고제도에 관한 고찰" 대한약국학회지(Korean Journal of Community Pharmacy. Vol. 5. No. 1. 56-65. 2019

[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805, 2018.

[3] Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36.4 (2020): 1234-1240.

[4] Alsentzer, Emily, et al. "Publicly available clinical BERT embeddings." arXiv preprint arXiv:1904.03323 (2019).

[5] Biseda, Brent, and Katie Mo. "Enhancing Pharmacovigilance with Drug Reviews and Social Media." arXiv preprint arXiv:2004.08731 (2020).

[6] Tutubalina, Elena, et al. "The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews." Bioinformatics (2020).

[7] Mozzicato, Patricia. "MedDRA." Pharmaceutical Medicine 23.2 : 65-75. 2009

[8] KFU NLP Team at SMM4H 2020 Tasks: Cross-lingual Transfer Learning
with Pretrained Language Models for Drug Reactions

[9] O'Connor, Karen, et al. "Pharmacovigilance on twitter? Mining tweets for adverse drug reactions." AMIA annual symposium proceedings. Vol. 2014. American Medical Informatics Association, 2014.

[10] Jurafsky, Daniel, and James H. Martin. "Information extraction." Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition : 725-743. 2009

[11] Karimi, Sarvnaz, et al. "Cadec: A corpus of adverse drug event annotations." Journal of biomedical informatics 55 : 73-81. 2015

[12] Ma, Xuezhe, and Eduard Hovy. "End-to-end sequence labeling via bi-directional lstm-cnns-crf." arXiv preprint arXiv:1603.01354, 2016.