한국컴퓨터정보학회 동계학술대회 논문집 제27권 제1호 (2019. 1)

Multi-Channel PCNN 모델을 활용한 약물-약물 상호작용 관계 추출

박찬희[O], 조민수[*], 박장원[*], 박상현[*]
[O*]연세대학교 컴퓨터과학과
e-mail: {channy_12, minsoo0104, adueujw, sanghyun}@yonsei.ac.kr[O*]

Relation Extraction of Drug-Drug Interaction using Multi-Channel PCNN Model

Chanhee Park[O], Minsoo Cho[*], Jangwon Park[*], Sanghyun Park[*]
[O*]Dept. of Computer Science, Yonsei University

● 요 약 ●

DDI 추출은 생물 의학 문헌으로부터 약물-약물 상호작용(Drug-Drug Interaction) 관계를 추출하는 작업으로, 기존에 알려지지 않은 인체 내 약물 간의 효과 또는 부작용 정보를 제공하는데 중요한 역할을 한다. 본 연구에서는 PCNN 모델을 활용하여 특징 추출 과정을 자동화하고 약물 개체 간의 구조 정보를 포착해 개체 간 관계를 효율적으로 추출하였으며, 생물 의학 문헌에서 쓰이는 생소한 용어를 보다 풍부하게 표현하기 위해 5가지 버전의 단어 임베딩을 PCNN의 채널로 사용하였다. 본 연구에서 제안하는 MC-PCNN 모델의 성능 평가를 위해 DDI'13 Corpus 데이터를 사용하여 비교 실험을 진행하였으며, 그 결과 기존 연구보다 F_1 점수 기준 최대 2.05%p 향상된 성능을 보이며 DDI 관계 추출에서 효과적인 방법론임을 확인하였다.

키워드: 약물-약물 상호작용(Drug-Drug Interaction), 관계 추출(Relation Extraction), 심층 신경망 (Deep Neural Network), 멀티 채널(Multi-Channel)

I. 서 론

약물-약물 상호작용(Drug-Drug Interaction)은 두 개 이상의 약물을 동시에 투여할 때, 인체에서 발생할 수 있는 숨겨진 약물 간의 상호 관계를 추출하는 것을 목적으로 한다. 둘 이상의 약물이 결합되었을 때, 그 효과가 증대되거나 약화될 수 있을 뿐 아니라 인체에 매우 유해한 결과를 초래할 수 있기 때문이다. 따라서 이러한 부작용을 사전에 방지하고자 DrugBank[1], Drugs.com[2]과 같은 다양한 DDI 데이터베이스가 존재하고 있다. 하지만 생명 의학 문헌의 빠른 발생 속도로 인해 데이터베이스가 최신의 DDI 정보를 수집하는 데에는 한계가 존재한다. 또한, 자연 언어로 쓰인 DDI 정보를 DDI 데이터베이스에 수동으로 정리하는 것은 비용과 시간 측면에서 비효율적이라 할 수 있다. 따라서, 이러한 과정을 자동화한 효율적인 DDI 추출 시스템이 필요하게 되었다.

기계 학습을 기반으로 한 기존의 DDI 관계 추출 방법은 크게 특징 기반(Feature-based)과 커널 기반(Kernel-based)의 두 가지 방법으로 나눌 수 있다. 특징 기반 방법은 관계 분류를 위한 단서들을 특징 벡터로 추출하여 활용하므로, 경험을 기반으로 적합한 특징을 설계해야 할 필요가 있다[3-4]. 또한, 커널 기반 방법은 서로 다른 커널을 사용하여 직접 특징을 추출하지 않고 구문 분석 트리(Parse tree), 의존성 그래프(Dependency graph)와 같은 데이터 객체의 구조적 표현을 통해 두 객체 간의 유사성을 계산한다[5]. 그러나 이러한 접근법은 모두 분류기(Classifier)를 학습시키기 위해 도메인 전문가와 자연 언어 처리 툴을 통한 특징 엔지니어링(Feature engineering)에 의존하므로, 많은 시간과 비용을 필요로 할 뿐 아니라 고품질의 특징을 설계하기 어렵다는 단점이 존재한다.

최근에는 심층 신경망을 사용하여 학습을 통해 특징 추출 과정을 자동화한 연구가 활발히 진행되고 있다. 본 연구에서도, CNN(Convolutional Neural Network)을 사용하여 복잡한 전처리 과정 없이 DDI 관계 추출을 위한 특징 추출 모형을 학습시켰다. 특히, Max Pooling 과정에서 은닉층(Hidden layer)의 크기를 급격히 축소시키고 2개의 약물 개체 사이의 구조 정보를 포착할 수 없는 한계를 극복하고자 [6]에서 제안한 PCNN(Piece-wise CNN)을 독립적인 구조로 사용하였다. PCNN은 관계 추출 데이터의 증대(Augmentation)를 위해 사용되는 Distant Supervision 방법론의 잘못된 레이블(Wrong label) 문제를 해결하고자 [6]에서 제안되었으나, 독립적인 DDI 관계 추출 구조로 적용된 사례는 확인된 바가 없다. 또한, 생명 의학 문헌에서 쓰이는 생소한 용어의 의미를 보다 정확하게 표현하고자 여러 버전의 단어 임베딩(Word embedding)을 CNN의 채널(Channel)로 사용하는 Multi-Channel 임베딩[7]을 사용하였다. 제안하는 MC-PCNN 모델의 학습 및 성능 평가를 위해 DDI'13 Corpus(말뭉치)를 사용하였으며, 추가적인 언어 자질을 사용

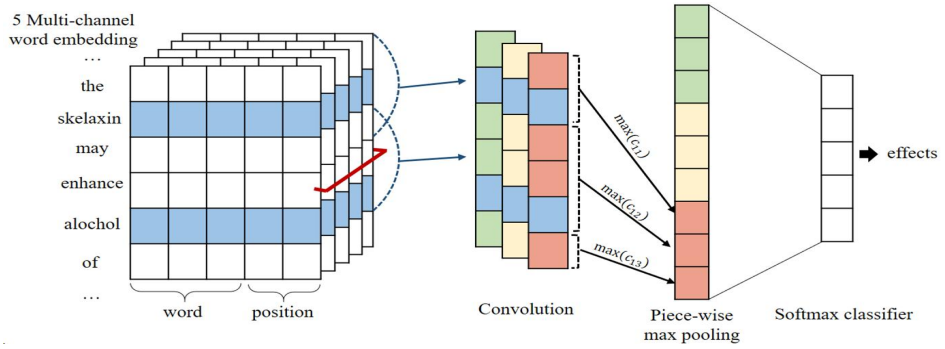

Fig. 2. Architecture of the Multi-Channel Piecewise CNN model

하지 않음에도 불구하고 선행 연구보다 F1 점수를 기준으로 최대 2.05%p 향상된 성능을 보임을 확인하였다.

II. 방법론

1. Multi-Channel 임베딩

본 연구에서는 네 가지 종류의 생물학 말뭉치를 Word2vec과 CBoW 방법론을 사용하여 학습시킨 여러 버전의 단어 임베딩을 PCNN의 입력으로 사용한다. 여러 버전의 단어 임베딩을 모두 사용함으로써, 보다 많은 단어를 벡터로 표현할 수 있으며, 이를 통해 단어 임베딩의 대표적인 문제인 미등록 단어 문제(Out-Of-Vocabulary)를 완화시킬 수 있다. 미등록 단어 문제란 특정 단어가 임베딩 사전에 존재하지 않을 경우, 이를 <UNK>토큰으로 처리하여 단어의 뜻을 제대로 반영하지 못하는 문제를 일컫는다. 특히 생물학 데이터의 경우, 생소한 생물학 용어의 사용으로 미등록 단어 문제가 빈번하게 발생한다. 이러한 문제를 보완하고자 본 연구에서는 PubMed, PMC, MedLine, Wikipedia와 같은 생물학 말뭉치와 일반 말뭉치를 모두 사용하였다. 표 1과 같이 해당 말뭉치를 이용한 5가지 버전의 단어 임베딩을 입력의 각 채널로 사용함으로써 미등록 단어 문제를 완화하고, 각 단어의 의미를 보다 풍부하게 표현하도록 하였다.

Table 1. Word embeddings of 5 version

	#. of words	Train corpus
1	2515686	PMC
2	2351706	PubMed
3	4087446	PMC and PubMed
4	5443656	Wikipedia and PubMed
5	650187	MedLine

2. Position 임베딩

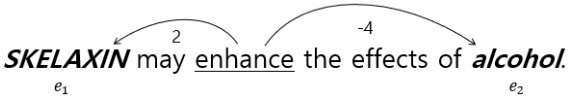

Fig. 1. Example of the position embedding

본 연구에서는 단어 임베딩 외에도, 선행 연구[8]에서 적용된 Position 임베딩을 추가하여 모델의 입력 값으로 사용하였다. Position 임베딩은 문장 내 두 개체와 나머지 단어 사이의 상대적 거리를 각 단어별로 나타낸 것으로, Position 임베딩 정보를 통해 문장 내 두 개체의 위치를 나타낼 수 있다. 예를 들어, 아래 문장에서 세 번째 단어인 enhance라는 단어와 두 약물 개체(SKELAXIN, alcohol)에 대한 상대적 거리는 각각 2와 -4이다.

이와 같이 모든 단어에 대해 두 개체와의 상대적 거리를 계산하고, 이를 n차원의 임베딩 벡터로 각각 변환시켜 그림 1과 같이 Multi-Channel 임베딩과 연결(Concatenate)한다. 각 문장에 대한 최종 임베딩은 $S \in R^{s \times d}$로, s는 각 문장의 길이를 나타내며, d는 $d = d_w + d_p * 2$ 로 단어 임베딩과 두 개의 Position 임베딩을 결합한 최종 입력 임베딩을 나타낸다.

3. PCNN 모델

본 연구에서는 CNN의 확장 모델인 PCNN 모델을 문장 내 두 약물 개체의 관계 추출 및 예측을 위해 적용한다. 최초의 CNN 모델은 주로 이미지 분야에서 사용되었지만, 최근에는 자연 언어 처리 분야에서 우수한 성능을 보여, 다양한 모델로 확장되고 있다. 자연어 처리에서 CNN 모델의 필터는 단어 간의 지역적인 정보 및 문맥 정보를 학습하고 자질 맵(Feature map)을 생성하는데 사용된다. 생성된 자질 맵은 Max Pooling 과정을 거쳐, 그 의미를 대표하여 반영할 수 있는 가장 중요한 자질 한 개로 추출된다.

일반적인 CNN 모델에서는 이와 같이 컨벌루션 결과에 한번의 Max Pooling 과정을 통해 자질 맵의 크기를 빠르게 축소시킨다. 하지만, PCNN 모델에서는 두 개체간의 구조적인 정보를 포착하기 위해 그림 2와 같이 두 개의 약물 개체를 기준으로 컨벌루션 결과를 3개의 Segment로 분리하고, 각각에 대해 Max Pooling을 수행한다. 이를 통해 개체간의 구조적인 정보를 더 세밀하게 포착하여, 두 개체의 관계를 효율적으로 추출하도록 하였다.

$$m_k = f(\sum_{i=1}^{c} V^i[k:k+h-1] \odot W^i) \quad (1)$$

PCNN의 컨벌루션 과정은 수식 1과 같다. 수식 1의 $V^i \in R^{N \times d}$와 $W^i \in R^{h \times d}$는 각각 채널i의 입력 단어 임베딩과 크기가 h인 필터를 나타내며, $\odot$와 f는 각각 Element-wise 곱과 활성화 함수 ReLU를 의미한다. 컨벌루션 결과로 생성된 자질 맵은 $c = [m_1, m_2, m_3, \cdots m_{N-h+1}]$로 나타낼 수 있으며, 서로 다른 크기의 필터 n개가 적용될 경우, $c_i = [c_1, c_2, c_3, \cdots c_n]$와 같이 나타낼 수 있다.

$$p_{ij} = \max(c_{ij}) \quad (1 \le i \le n, 1 \le j \le 3) \qquad (2)$$

컨벌루션 결과 생성된 c_i는 수식 2와 같이 세 개의 segment c_{i1}, c_{i2}, c_{i3} 로 나누어져, c_{ij} 각각에 대해 Max Pooling이 수행된다. 이후에 Max Pooling 결과인 p_{ij}를 모두 연결하여 활성화 함수 ReLU에 적용한다. 그 결과 값을 최종적으로 소프트맥스(Softmax) 함수에 적용하여 두 약물 개체의 관계를 예측한다.

III. 실험 및 결과

1. 데이터

본 연구에서는 모델의 학습 및 성능 평가를 위해 DDIExtraction 2013 Challenge에서 제공하는 DDI 태스크를 위한 DDI'13 Corpus를 사용하였다. 생명 의학 문헌인 784개의 DrugBank 텍스트와 233개의 MedLine 요약문으로 구성된 DDI'13 Corpus는 총 5,021개의 약물간 상호작용(DDI)을 포함하고 있으며, DDI 유형에는 Advice, Effect, Mechanism, Int, 4가지 유형과 어떠한 유형에도 속하지 않는 False 유형을 포함하여 총 5개의 유형이 존재한다.

DDI'13 Corpus에서 파싱되어 추출된 데이터는 모델의 성능 향상을 위해 전처리 과정을 거쳤다. 약물의 이름이 DDI 관계를 추출하는데 중요한 역할을 하지 않으므로, 각 문장에서 관계 추출에 해당하는 두 개체는 Drug A, Drug B로, 두 개체 외에 나머지 약물은 Drug N으로 대체하였으며, 문장에서 독립적으로 등장하는 숫자는 '#' 기호로 변경하였다.

또한, 본 연구에서는 DDI'13 Corpus에 Negative Instance Filtering을 적용하여, 유효한 유형보다 False 유형이 많은 데이터의 불균형 문제를 해결하고자 하였다. Negative Instance Filtering은 관계 추출 분야에서 데이터의 균등한 분포를 위해 False 유형을 필터링 하는 기법으로, 사전에 정의해 놓은 규칙에 해당하는 문장을 제거하는 방식으로 수행된다. 규칙의 예로, 'Drug A such as Drug B'와 같이 한 문장 내 동일한 약물을 지칭하는 관계를 가진 규칙 또는 'Drug A, Drug B, and Drug N'과 같이 'and' 또는 'or'와 같은 특정 관계를 지칭하는 단어를 포함하는 규칙 등이 존재한다. 본 연구에서는 [9]에서 제공하는 Negative Instance Filtering에 해당하는 문장 ID 정보를 활용하여 데이터의 False 유형을 필터링하였다. 그 결과, 학습 데이터와 평가데이터에서 각각 False 데이터를 14,785개, 2,733개 줄여, False 데이터와 나머지 유형의 데이터의 비율을 대략 1:2로 맞추었다. 전처리 과정을 거친 최종 DDI'13 Corpus 데이터는 총 15,861개의 문장으로 구성되었으며, 학습 데이터와 평가 데이터의 문장 개수 각각 12,841개, 3,020개 이다.

2. 실험 환경 및 모수 설정

Table 2. Hyperparameter of the model

Hyperparameter	Value
Epoch	20
Batch size	64
Word embedding size	200
Position embedding size	10
CNN kernel size	3,5,7,9
Number of filters	100
Dropout rate	0.45
Learning rate	3e-4

본 모델은 Python 기반의 Keras 프레임 워크를 활용하여 구현되었으며, 실험은 NVIDIA GeForce GTX 1080 GPU가 장착된 PC에서 진행되었다. 단어 임베딩 벡터는 [8]에서 제공하는 5개 버전의 단어 임베딩을 200 차원으로 사용하였다. MC-PCNN 모델의 학습을 위해서는 Adam Optimizer[10]를 사용하였으며, 사용한 모수는 표 2와 같다.

3. 실험 결과 분석

제안하는 모델의 객관적인 성능 평가를 위해 DDI 관계 추출의 선행 연구에서 사용하는 F_1 점수를 평가 척도로 사용하였다. F_1 점수는 정밀도(Precision)와 재현율(Recall)의 조화 평균으로 계산되며, 데이터의 유형이 불균형할 경우 주로 사용된다. DDI 관계 추출에서 사용하는 데이터셋 또한 각 유형별 데이터가 균등하지 않으므로, 불균등한 데이터의 분포를 고려하고자 F_1 점수를 모델의 성능 평가 지표로 사용한다.

표 3은 제안하는 모델(MC-PCNN)과 기본 모델(CNN, Bi-LSTM), 그리고 선행 연구 모델(PCNN[6], MCCNN[8])과의 실험 결과를 비교하였다. 제안하는 모델이 기본적인 CNN, Bi-LSTM과 각각 4.51%p, 2.75%p의 성능 향상을 보였다. 기본적인 CNN은 Max Pooling 과정에서 은닉층을 급격히 축소시키고 2개의 약물 개체 사이의 구조 정보를 포착할 수 없기 때문에 이러한 성능 차이가 존재하는 것으로 분석된다. 또한 Bi-LSTM의 경우에도 여러 은닉층을 거치면서 문장의 구조 정보가 유실되는 한계가 존재하는 것으로 판단된다.

기존 선행 연구인 PCNN, MCCNN과 제안하는 모델을 비교하였을 때에는 각각 2.05%p, 1.22%p의 성능 향상을 확인할 수 있었다. 이는 PCNN 모델을 통해 두 약물 개체간의 구조적인 정보를 보다 세밀하게 포착해 DDI 관계를 효율적으로 예측할 뿐 아니라, Multi-Channel 임베딩을 통해 생명 의학 문헌에서 등장하는 생소한 용어들을 보다 정확하고 풍부하게 표현하는 것이 가능하기 때문인 것으로 분석된다.

Table 3. Experiment results on proposed model and other models

	Precision	Recall	F1-score
CNN	66.50	67.31	66.90
Bi-LSTM	70.62	66.80	68.66
PCNN	70.58	68.18	69.36
MCCNN	69.94	70.44	70.19
MC-PCNN	**71.97**	**70.85**	**71.41**

IV. 결 론

본 연구에서는 약물 간의 상호 작용을 예측하고자, 생명 의학 문헌에서 DDI 관계 추출을 위한 MC-PCNN 모델을 제안하였다. PCNN을 통해 약물 개체 사이의 구조 정보를 포착하여 두 개체의 관계를 효율적으로 추출하였다. 또한, 생명 의학 문헌의 특성을 반영하기 위해 생물학 말뭉치와 일반 말뭉치를 모두 사용한 5가지 버전의 단어 임베딩을 적용하여 생소한 용어들을 보다 정확하고 풍부하게 표현하였다. 제안하는 모델의 검증을 위해 DDI'13 Corpus를 사용하였으며, 선행 연구와 비교한 경우에도 최대 2.05%p의 성능 향상을 확인하였다.

하지만 DDI 관계 추출이 약물의 부작용을 방지하는 매우 중요한 작업임에도 불구하고 다른 관계 추출 작업과 비교했을 때 상대적으로 낮은 성능을 보이고 있다. 따라서 추후에는 성능 향상을 위해 멀티 태스크 학습을 적용한 확장 연구를 진행할 계획이다.

ACKNOWLEDGEMENT

이 논문은 과학기술정보통신부와 한국연구재단의 방사선기술개발사업으로 연구 지원한 (2017M2A2A7A02020213)의 결과물입니다.

REFERENCES

[1] V. Law, C. Knox, Y. Djoumbou et al., "DrugBank 4.0: sheddingnew light on drug metabolism," *Nucleic Acids Research*, vol. 42, no. 1, pp. D1091–D1097, 2014.

[2] Drugs.com [Internet] Prescription drug information, interactions and side effects. 2000; Available from: *https://www.drugs.com/*.

[3] N. Kambhatla, "Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations," *Proc. of ACLdemo*, 2004.

[4] F. M. Suchanek et al., "Combining linguistic and statistical analysis to extract relations from web documents," *Proc. of KDD*, pp. 712–717, 2006.

[5] D. Tikk et al., "A detailed error analysis of 13 kernel methods for protein–protein interaction extraction," *BMC Bioinformatics*, 14(1):12, 2013.

[6] D. Zeng et al., "Distant supervision for relation extraction via piecewise convolutional neural networks," *Proc. of 2015 EMNLP*, pp. 1753–1762, 2015.

[7] D. Zeng et al., "Relation classification via convolutional deep neural network," *Proc. of COLING*, pp. 2335–2344, 2014.

[8] C. Quan et al., "Multichannel convolutional neural network for biological relation extraction," *BioMed Research International*, vol. 2016, 2016.

[9] S. Lim, et al., "Drug drug interaction extraction from the literature using a recursive neural network," *PloS one*, 13(1), e0190926, 2018.

[10] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," *Proc. of the 2015 ICLR*, arXiv preprint arXiv:1412.6980, 2014.