VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

A Comparison of Single Concatenated Input and Indenpendent Multiple-input for Convolutional Neural Networks to Predict Chemical-induced Disease Relations

Pham Thi Quynh Trang, Bui Manh Thang, Dang Thanh Hai*

Bingo Biomedical Informatics Lab, Faculty of Information Technology,
VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam

Received 21 October 2019
Revised 17 March 2020; Accepted 23 March 2020

Abstract: Chemical compounds (drugs) and diseases are among top searched keywords on the PubMed database of biomedical literature by biomedical researchers all over the world (according to a study in 2009). Working with PubMed is essential for researchers to get insights into drugs' side effects (chemical-induced disease relations (CDR), which is essential for drug safety and toxicity. It is, however, a catastrophic burden for them as PubMed is a huge database of unstructured texts, growing very fast (~28 million scientific articles currently, approximately two new deposits per minute). As a result, biomedical text mining has been empirically demonstrating its great implications in biomedical research communities. Biomedical text has its own distinct challenging properties, attracting much attention from natural language processing communities. A large-scale study in 2018 showed that incorporating information into indenpendent multiple-input layers outperforms concatenating them into a single input layer (for biLSTM), producing better performance when compared to the state-of-the-art CDR classifying models. This paper demonstrates that the opposite is right for a convolutional neural network (CNN), in which concatenation is better for CDR classification. To this end, a CNN based model is developed with multiple input concatenated for CDR classification. The study experimental results on the benchmark dataset demonstrate their outperformance over the other recent state-of-the-art CDR classification models.

Keywords: Chemical-induced disease relation prediction, convolutional neural network, biomedical text mining.

1. Introduction

Drug manufacturing is an extremely expensive and time-consuming process [1]. It requires approximately 14 years, with a total cost of about $1 billion, for a specific drug to be available in the pharmaceutical market [2]. Nevertheless, even when being in clinical uses for a while, side effects of many drugs are still unknown to scientists and/or clinical doctors [3]. Understanding drugs' side effects is

* Corresponding author.
E-mail address: hai.dang@vnu.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.237

essential for drug safety and toxicity. All these facts explain why chemical compounds (drugs) and diseases are among top searched keywords on PubMed by biomedical researchers all over the world (according to [4]). PubMed is a huge database of biomedical literature, currently with ~28 million scientific articles, and is growing very fast (approximately two articles added per minute).

Working with such a huge number of unstructured textual documents in PubMed is a catastrophic burden for biomedical researchers. It can be, however, accelerated with the application of biomedical text mining, hereby for drug (chemical) - disease relation prediction, in particular. Biomedical text mining has been empirically demonstrating its great implications in biomedical research communities [5-7].

Biomedical text has its own distinct challenging properties, attracting much attention from natural language processing communities [8, 9]. In 2004, an annual challenge, called BioCreative (Critical Assessment of Information Extraction systems in Biology) was launched for biomedical text mining researchers. In 2016, researchers from NCBI organized the chemical disease relationship extraction task for the challenge [10].

To date, almost all the proposed models have been only for prediction of relationships between chemicals and diseases that appear within a sentence (intra-sentence relationships) [11]. It is important to note that those models that produce the state-of-the-art performance are mainly based on deep neural architechtures [12-14], such as recurrent neural networks (RNN) like bi-directional long short-term memory (biLSTM) in [15] and convolutional neural networks (CNN) in [16-18].

Recently, Le et al. [19] developed a biLSTM based intra-sentence biomedical relation prediction model that incorporates various informative linguistic properties in an independent multiple-layer manner. Their experimental results demonstrate that incorporating information into independent multiple-input layers outperforms concatenating them into a single input layer (for biLSTM), producing better performance when compared to the relevant state-of-the-art models. To the best of our knowledge, there is currently no study confirming whether it is still held true for a CNN-based intra-sentence chemical disease relationship prediction model by far. To this end, this paper proposes a model for prediction of intra-sentence chemical disease relations in biomedical text using CNN with concatenation of multiple layers for encoding different linguistic properties as input.

The rest of this paper is organized as follows: Section 2 describes the proposed method; Experimental results are discussed in Section 3; and, Section 4 concludes the paper.

2. Method

Given a preprocessed and tokenized sentence containing two entity types of interest (i.e. chemical and disease), the proposed model first extracts the shortest dependency path (SDP) (on the dependency tree) between the two entities. The SDP contains tokens (together with dependency relations between them) that are important for understanding the semantic connection between the two entities (see Figure 1 for an example of the SDP).

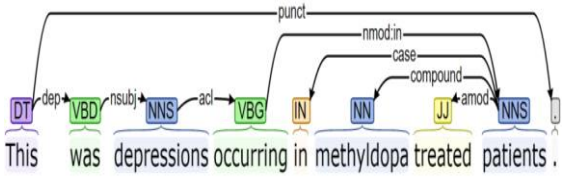

Figure 1. Dependency tree for an example sentence.

The shortest dependency path between the two entities (i.e. *depressions* and *methyldopa*) goes through the tokens "*occurring*" and "*patients*".

Each token *t* on a SDP is encoded with the embedding e^t by concatenating three embeddings of equal dimension *d* (i.e. $e^w \oplus e^{pt} \oplus e^{ps}$), which

represents important linguistic information, including its token itself (e^w), part of speech (POS) (e^{pt}) and its position (e^{ps}). The two former partial embeddings are fine-tuned during the model training. Position embeddings are indexed by distance pairs [d^l%5, d^r%5], where d^l and d^r are distances from a token to the left and the right entity, respectively.

For each dependency relation (r) on the SDP, its embedding has the dimension of *3*d*, and is randomly initialized and fine-tuned as the model's parameters during training.

To this end, each SDP is embedded into the R^{NxD} space (see Figure 2), where N is the number of all tokens and dependency relations on the SDP and $D=3*d$. The embedded SDP will be fed as input into a conventional convolutional neural network (CNN [20]) for being classified if there is or there is not a predefined relation (i.e. chemical-induced disease relation) between two entities.

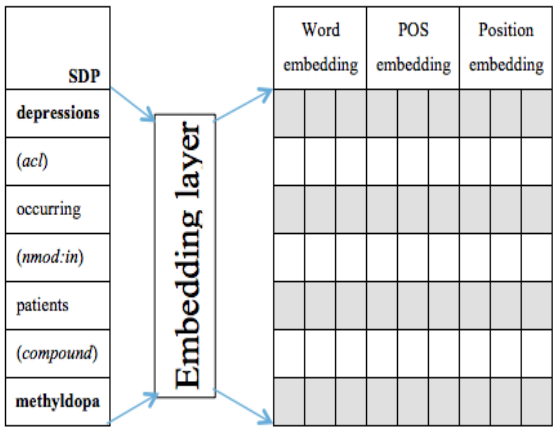

Figure 2. Embedding by concatenation mechanism of the Shortest Dependency Path (SDP) from the example in Figure 1.

2.1. Multiple-Channel Embedding

For multi-channel embedding, instead of concatenating three partial embeddings of each token on a SDP, three independent embedding channels are maintained. Channels for relations on the SDP are identical embeddings. As a result, SDPs are embedded into R^{nxdxc}, where n is the number of all tokens and dependency relations between them, d is the dimension number of embeddings, and $c=3$ is the number of embedding channels.

Feature maps for CNN are calculated using the scheme in the work of Kim 2014 [21]. Each CNN's filter f_i is slided along each embedding channel (c) independently, creating a corresponding feature map $\boldsymbol{\mathcal{F}}_{ic}$. The max pooling operator is then applied on those created feature maps on all channels (three in our case) to create a feature value for filter f_i (Figure 3).

2.2. Hyper-Parameters

The model's hyper-parameters are empirically set as follows:

- Filter size: n x d, where d is the embedding dimension (300 in our experiments), n is a number of consecutive elements (tokens/POS tags, relations) on SDPs (Figure 3).
- Number of filters: 32 filters of the size 2 x 300, 128 of 3 x 300, 32 of 4 x 300, 96 of 5 x 300.
- Number of hidden layers: 2.
- Number of units at each layer: 128.
+ The number of training epochs: 100.
+ Patients for early stopping: 10.
+ Optimizer: Adam

3. Experimental Results

3.1. Dataset

The experiments in this study are conducted on the Bio Creative V data [10]. It's an annotated text corpus that consists of human annotations for chemicals, diseases and their chemical-induced-disease (CID) relations at the abstract level. The dataset contains 1,500 PubMed articles divided into three subsets for training, development and testing. Most of the 1,500 articles were selected from the CTD data set (accounting for 1,400/1,500 articles). The remaining 100 articles in the test set were carefully selected from different sources. All these data are manually curated. The detail information is shown in Table 1.

3.2. Model Evaluation

The training and development subsets of the BioCreative V CDR are merged into a single training dataset, which is then divided into the new training and validation/development data with a ratio of 85%:15%. To stop the training process at the right time, the early stop technique on F1-score on the new validation data is used.

The entire text will be passed through a sentence splitter. Then, based on the name of the disease and the name of the chemical marked from the previous step, all the sentences containing at least one pair of chemical-disease entities are filtered out. With all the sentences found, the relation for each pair of chemical-disease entities can be classified. Model training and evaluating are performed 15 times on the new training and development set, and the averaged F1 on the test set is chosen as the final evaluation result across the entire dataset to make sure that the model can work well with strange samples.

Finally, the models that achieve the best results based on the sentence level will be applied to the problem on the abstract level to compare with the other most recent state-of-the-art methods.

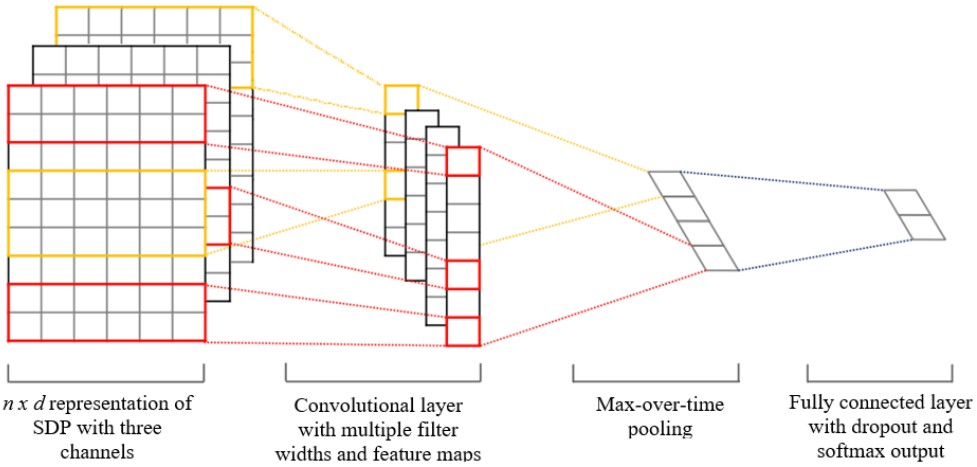

Figure 3. Model architecture with three-channel embedding as an input for an SDP.

Table 1. Statistics on BioCreative V CDR dataset [10]

Dataset	Articles	Chemical		Disease		CID
		Mention	ID	Mention	ID	
Training	500	5203	1467	4182	1965	1038
Development	500	5347	1507	4244	1865	1012
Test	500	5385	1435	4424	1988	1066

3.3. Results and Comparison

The experiment results show that the model achieves the averaged F1 of 57% (Precision of 55.6% and Recall of 58.6%) at the abstract level. Compared with its variant that does not use dependency relations, an outperformance of about 2.6% at F1 is observed, which is very significant (see Table 2). This indicates that dependency relations contain much information for relation extraction. In the meanwhile, POS tag and position information are also very useful when contributing 0.9% of the F1 improvement to the final performance of the model.

Compared with the recent state-of-the-art models such as MASS [19], ASM [22] and the tree kernel based model [23], the proposed model performs better (Table 3). The proposed model and MASS only exploit intra-sentence

information (namely, SDPs, POS and positions), ignoring prediction for cross-sentence relations, while the other two incorporate cross-sentence information.

Table 2. Performance of the proposed model with different linguistic information used as input

Information used	Precision	Recall	F1
Tokens only	53.7	55.4	54.5
Token, Dependency (depRE)	55.7	56.8	56.2
Tokens, DepRE and POS tags	55.7	57.5	56.6
Tokens, depRE, POS and Position	55.6	58.6	57.0

It is noted that cross-sentence relations account for 30% of all relations in the CDR dataset. This probably explains why ASM could achieve better recall (67.4%) than the proposed model (58.6%).

Table 3. Performance of the proposed model in comparison with the other state-of-the-art models

Model	Relations	Precision	Recall	*F1*
Zhou et al., 2016	Intra- and inter-sentence	64.9	49.2	56.0
Panyam et al., 2018	Intra- and inter-sentence	49.0	67.4	56.8
Le et al., 2018	Intra-sentence	58.9	54.9	56.9
Proposed model	Intra-sentence	55.6	58.6	57.0

4. Conclusion

This paper experimentally demonstrates that CNNs perform better prediction of abstract-level chemical-induced disease relations in biomedical literature when using concatenated input embedding channels rather than independent multiple channels. It is vice versa for BiLSTM when multiple independent channels give better performance, as shown in a recent related large-scale study [Le et al., 2018]. To this end, this paper presents a model for prediction of chemical-induced disease relations in biomedical text based on a CNN with concatenated input embeddings. The experimental results on the benchmark dataset show that the proposed model outperforms the three recent state-of-the-art related models.

Acknowledgements

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under Grant 102.05-2016.14.

References

[1] Paul SM, D.S. Mytelka, C.T. Dunwiddie, C.C. Persinger, B.H. Munos, S.R. Lindborg, A.L. Schacht, How to improve R&D productivity: The pharmaceutical industry's grand challenge, Nat Rev Drug Discov. 9(3) (2010) 203-14. https://doi.org/10.1038/nrd3078.

[2] J.A. DiMasi, New drug development in the United States from 1963 to 1999, Clinical pharmacology and therapeutics 69 (2001) 286-296. https://doi.org/10.1067/mcp.2001.115132.

[3] C.P. Adams, V. Van Brantner, Estimating the cost of new drug development: Is it really $802 million? Health Affairs 25 (2006) 420-428. https://doi.org/10.1377/hlthaff.25.2.420.

[4] R.I. Doğan, G.C. Murray, A. Névéol et al., "Understanding PubMed user search behavior through log analysis", Oxford Database, 2009.

[5] G.K. Savova, J.J. Masanz, P.V. Ogren et al., "Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications", Journal of the American Medical Informatics Association, 2010.

[6] T.C. Wiegers, A.P. Davis, C.J. Mattingly, Collaborative biocuration-text mining development task for document prioritization for curation, Database 22 (2012) pp. bas037.

[7] N. Kang, B. Singh, C. Bui et al., "Knowledge-based extraction of adverse drug events from biomedical text", BMC Bioinformatics 15, 2014.

[8] A. Névéol, R.L. Doğan, Z. Lu, "Semi-automatic semantic annotation of PubMed queries: A study on quality, Efficiency, Satisfaction", Journal of Biomedical Informatics 44, 2011.

[9] L. Hirschman, G.A. Burns, M. Krallinger, C. Arighi, K.B. Cohen et al., Text mining for the biocuration workflow, Database Apr 18, 2012, pp. bas020.

[10] Wei et al., "Overview of the BioCreative V Chemical Disease Relation (CDR) Task", Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, 2015.

[11] P. Verga, E. Strubell, A. McCallum, Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction, In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1 (2018) 872-884.

[12] Y. Shen, X. Huang, Attention-based convolutional neural network for semantic relation extraction, In: Proceedings of COLING 2016, the Twenty-sixth International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 2526-2536.

[13] Y. Peng, Z. Lu, Deep learning for extracting protein-protein interactions from biomedical literature, In: Proceedings of the BioNLP 2017 Workshop, Association for Computational Linguistics, Vancouver, Canada, 2016, pp. 29-38.

[14] S. Liu, F. Shen, R. Komandur Elayavilli, Y. Wang, M. Rastegar-Mojarad, V. Chaudhary, H. Liu, Extracting chemical-protein relations using attention-based neural networks, Database, 2018.

[15] H. Zhou, H. Deng, L. Chen, Y. Yang, C. Jia, D. Huang, Exploiting syntactic and semantics information for chemical-disease relation extraction, Database, 2016, pp. baw048.

[16] S. Liu, B. Tang, Q. Chen et al., Drug-drug interaction extraction via convolutional neural networks, Comput, Math, Methods Med, Vol (2016) 1-8. https://doi.org/10.1155/2016/6918381.

[17] L. Wang, Z. Cao, G. De Melo et al., Relation classification via multi-level attention CNNs, In: Proceedings of the Fifty-fourth Annual Meeting of the Association for Computational Linguistics 1 (2016) 1298-1307. https://doi.org/10.18653/v1/P16-1123.

[18] J. Gu, F. Sun, L. Qian et al., Chemical-induced disease relation extraction via convolutional neural network, Database, 2017, pp. 1-12. https://doi.org/10.1093/database/bax024.

[19] H.Q. Le, D.C. Can, S.T. Vu, T.H. Dang, M.T. Pilehvar, N. Collier, Large-scale Exploration of Neural Relation Classification Architectures, In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2266-2277.

[20] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, In Proceedings of the IEEE. 86(11) (1998) 2278-2324.

[21] Y. Kim, Convolutional neural networks for sentence classification, ArXiv preprint arXiv: 1408.5882.

[22] C. Nagesh, Panyam, Karin Verspoor, Trevor Cohn and Kotagiri Ramamohanarao, Exploiting graph kernels for high performance biomedical relation extraction, Journal of biomedical semantics 9(1) (2018) 7.

[23] H. Zhou, H. Deng, L. Chen, Y. Yang, C. Jia, D. Huang, Exploiting syntactic and semantics information for chemical-disease relation extraction, Database, 2016.