Syntax-aware Multi-task Graph Convolutional Networks for Biomedical Relation Extraction

Diya Li[*], **Heng Ji**[†‡]
[*] Computer Science Department, Rensselaer Polytechnic Institute
lid18@rpi.edu
[†] Department of Computer Science [‡] Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
hengji@illinois.edu

Abstract

In this paper we tackle two unique challenges in biomedical relation extraction. The first challenge is that the contextual information between two entity mentions often involves sophisticated syntactic structures. We propose a novel graph convolutional networks model that incorporates dependency parsing and contextualized embedding to effectively capture comprehensive contextual information. The second challenge is that most of the benchmark data sets for this task are quite imbalanced because more than 80% mention pairs are negative instances (i.e., no relations). We propose a multi-task learning framework to jointly model relation identification and classification tasks to propagate supervision signals from each other and apply a focal loss to focus training on ambiguous mention pairs. By applying these two strategies, experiments show that our model achieves state-of-the-art F-score on the 2013 drug-drug interaction extraction task.

1 Introduction

Recently relation extraction in biomedical literature has attracted increasing interests from medical language processing research community as an important stage for downstream tasks such as question answering (Hristovski et al., 2015) and decision making (Agosti et al., 2019). Biomedical relation extraction aims to identify and classify relations between two entity mentions into pre-defined types based on contexts. In this paper we aim to extract drug-drug interactions (DDIs), which occur when taking two or more drugs within a certain period of time that alters the way one or more drugs act in human body and may result in unexpected side effects (Figure 1). Extracting DDI provides important clues for research in drug safety and human health care.

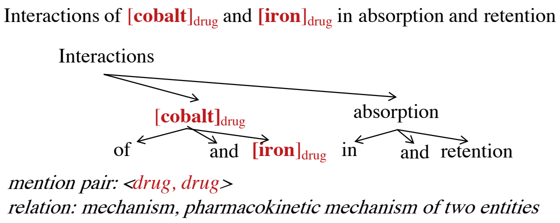
Interactions of [cobalt]drug and [iron]drug in absorption and retention
Interactions
[cobalt]drug
absorption
of
and
[iron]drug
in
and
retention
mention pair: <drug, drug>
relation: mechanism, pharmacokinetic mechanism of two entities

Figure 1: Example of drug-drug interaction on dependency tree.

Dependency parses are widely used in relation extraction task due to the advantage of shortening the distance of words which are syntactically related. As shown in Figure 1, the partial dependency path $\{iron \leftarrow cobalt \leftarrow interactions\}$ reveals that these two drugs are interactive, and the path $\{interactions \rightarrow absorption \rightarrow retention\}$ further indicates the *mechanism* relation between these two mentions. Therefore capturing the syntactic information involving the word *interaction* on the dependency path $\{iron \leftarrow cobalt \leftarrow interactions \rightarrow absorption \rightarrow retention\}$ can effectively help on the classification of the relation between these two mentions $\langle cobalt, iron\rangle$. In order to capture indicative information from wide contexts, we adopt the graph convolutional networks (GCN) (Kipf and Welling, 2016; Marcheggiani and Titov, 2017) to obtain the syntactic information by encoding the dependency structure over the input sentence with graph convolution operations. To compensate the loss of local context information in GCN, we incorporate the contextualized word representation pre-trained by the BERT model (Devlin et al., 2019) in large-scale biomedical corpora containing over 200K abstracts from PubMed and over 270K full texts from PMC (Lee et al., 2019) .

Moreover, we notice that data imbalance is another major challenge in biomedical text as the distribution of relations among biomedical mentions

Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pages 28–33
Hong Kong, November 3, 2019. ©2019 Association for Computational Linguistics
https://doi.org/10.18653/v1/D19-62

are usually very sparse. Over 80% candidate mention pairs have no relation in DDI 2013 (Herrero-Zazo et al., 2013) training set. To tackle this problem, we propose a binary relation identification task as an auxiliary task to facilitate the main multi-classification task. For instance, the detection of drug interaction on dependency path {*iron* $\leftarrow$ *cobalt* $\leftarrow$ ***interactions*** $\rightarrow$ *absorption* $\rightarrow$ *retention*} will assist the prediction of the relation type $mechanism$ by using the signals from binary classification as an inductive bias to avoid misclassifying it as no relation. We also exploit the focal loss (Lin et al., 2017) to potentially help the multi-class relation classification task by forcing the loss implicitly focus on ambiguous examples.

To recap, our contributions are twofold: First, we adopt the syntax-aware graph convolutional networks incorporating contextualized representation. Second, we further design an auxiliary task to solve the data imbalance problem, which achieves the state-of-the-art micro F-score on the DDIExtraction 2013 shared task.

2 Methods

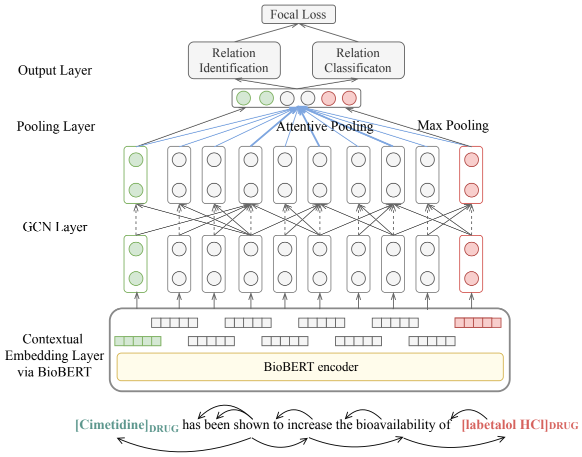

Figure 2: Framework of syntax-aware multi-task graph convolutional networks.

2.1 Contextual and Syntax-aware GCN

As a variant of the convolutional neural networks (LeCun et al., 1998), the graph convolutional networks (Kipf and Welling, 2016) is designed for graph data and it has been proven effective in modeling text data via syntactic dependency graphs (Marcheggiani and Titov, 2017).

We encode the tokens in a biomedical sentence of size n as $\boldsymbol{x} = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$, where $\boldsymbol{x_i}$ is a vector which concatenates the representation of the token i and the position embeddings corresponding to the relative positions from candidate mention pairs. We feed the token vectors into a L-layer GCN to obtain the hidden representations of each token which are directly influenced by its neighbors no more than L edges apart in the dependency tree. We apply the Stanford dependency parser (Chen and Manning, 2014) to generate the dependency structure:

$$h_i^{(l)} = \sigma(\sum_{j=1}^{n} \widetilde{A}_{ij} W^{(l)} h_j^{l-1}/d_i + b^{(l)})$$

where $\widetilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$ with $\boldsymbol{A}$ is the adjacent matrix of tokens in dependency tree, $\boldsymbol{I}$ is the identity matrix. $W^{(l)}$ is a linear transformation, $b^{(l)}$ is a bias term, and σ is a nonlinear function. Following Zhang et al. (2018), d_i is the degree of the token i in dependency tree with an additional self-loop.

We notice that some token representations are more informative by gathering information from syntactically related neighbors through GCN. For example, the representation of the token *interactions* from a 2-layer GCN operating on its two edges apart neighbors provides inductive information for predicting a $mechanism$ relation. Thus, we adopt attentive pooling (Zhou et al., 2016) to achieve the optimal pooling:

$$\alpha = softmax(w^T \tanh(h))$$

$$h_{attentive} = h\alpha^T$$

where w is a trained parameter to assign weights based on the importance of each token representation.

We obtain the final representation by concatenating the sentence from attentive pooling and the mention representations from max pooling. We finally obtain the prediction of relation type by feeding the final representations into a fully connected neural network followed by a softmax operation.

Graph neural networks (Zhou et al., 2018b) can learn effective representations but suffer from the loss of local context information. We believe the local context information is also crucial for biomedical relation extraction. For example, in the following sentence "*The response to [Factrel]*DRUG *may be blunted by [phenothiazines]*DRUG *and [dopamine antagonists]*DRUG", it's intuitive to tell *Factrel* and *phenothiazines* are interactive while *phenothiazines* and *dopamine antagonists* have no interaction according to the sentence order. However, GCNs treat the three drugs as interacting with each other as they are close in dependency structure with no order information.

BERT (Devlin et al., 2019) is a recently proposed model based on a multi-layer bidirectional Transformer (Vaswani et al., 2017). Using pre-trained BERT has been proven effective to create contextualized word embeddings for various NLP tasks (Han et al., 2019; Wang et al., 2019). The BioBERT (Lee et al., 2019) is a biomedical language representation model pre-trained on large-scale biomedical corpora. The output of each encoder layer of the input token can be used as a feature representation of that token. As shown in Figure 2, we encode the input tokens as contextualized embeddings by leveraging the last hidden layer of the corresponding token in BioBERT. As the BERT model uses WordPiece (Wu et al., 2016) to decompose infrequent words into frequent subwords for unsupervised tokenization of the input token, if the token has multiple BERT subword units, we use the first one. After getting the contextualized embedding of each token, we feed them into the GCN layer to make our model context-aware.

2.2 Auxiliary Task Learning with Focal Loss

In the DDIExtraction 2013 task, all possible interactions between drugs within one sentence are annotated, which means a single sentence with multiple drug mentions will lead to separate instances of candidate mention pairs (Herrero-Zazo et al., 2013). There are 21,012 mention pairs generated from 3,790 sentences in training set and over 80% have no relations. This data imbalance problem due to sparse relation distribution is a main reason for low recall in DDI task (Zhou et al., 2018a; Sun et al., 2019).

Here we address this relation type imbalance problem by adding an auxiliary task on top of the syntax-aware GCN model. To conduct the auxiliary task learning, we add a separate binary classifier for relation identification as shown in Figure 2. All classifiers share the same GCN representation and contextualized embeddings, and thus they can potentially help each other by propagating their supervision signals.

Additionally, instead of setting the objective function as the negative log-likelihood loss, here we optimize the parameters in training by minimizing a focal loss (Lin et al., 2017) which focuses on hard relation types. For instance, the *int* relation indicates drug interaction without providing any extra information (*e.g.*, *Some [anticonvulsants]*$_{\text{DRUG}}$ *may interact with [Mephenytoin]*$_{\text{DRUG}}$). This relation type only accounts for 0.82% in training set and is often misclassified into other relation types. We denote t_i and p_i as the ground truth and the conditional probability value of the type i in relation types C, the focal loss can be defined as:

$$\mathcal{L} = -\sum_{i}^{C} (\alpha_i (1 - p_i)^{\gamma} t_i \log(p_i)) + \lambda ||\theta||^2$$

where α is a weighting factor to balance the importance of samples from various types, γ is the focusing parameter to reduce the influence of well-classified samples in the loss. λ is the *L2* regularization parameter and θ is the parameter set.

The auxiliary task along with the focal loss enhances our model's ability to handle imbalance data by leveraging the inductive signal from the easier identification task and meanwhile down-weighting the influence of easy classified instances thus directing the model to focus on difficult relation types.

3 Experiments

3.1 Datasets and Task Settings

System	Prec	Rec	F1
CNN (Liu et al., 2016)	75.70	64.66	69.75
Multi Channel CNN (Quan et al., 2016)	75.99	65.25	70.21
GRU (Yi et al., 2017)	73.67	70.79	72.20
AB-LSTM (Sahu and Anand, 2018)	74.47	64.96	69.39
CNN-GCNN (Asada et al., 2018)	73.31	71.81	72.55
Position-aware LSTM (Zhou et al., 2018a)	75.80	70.38	72.99
RHCNN (Sun et al., 2019)	77.30	73.75	**75.48**
LSTM baseline	69.34	62.74	65.88
GCN baseline	71.96	67.14	69.47
–without attentive pooling	77.12	75.03	76.06
–without BioBERT	76.51	73.56	75.01
–without multi-task learning	76.01	71.92	73.91
Our Model	77.62	75.69	**76.64**

Table 1: Precision (Prec), recall (Rec) and micro F-score (F1) results on DDI 2013 corpus.

We evaluate our model on the DDIExtraction 2013 relation dataset (Herrero-Zazo et al., 2013). The corpus is annotated with drug mentions and their four types of interactions: *Mechanism* (pharmacokinetic mechanism of a DDI), *Effect* (effect of a DDI), *Advice* (a recommendation or advice regarding a DDI) and *Int* (a DDI simply occurs without extra information). We randomly choose 10% from the training dataset as the development set. Following previous work (Liu et al., 2016; Quan et al., 2016; Zhou et al., 2018a; Sun et al., 2019), we use a negative instance filtering strategy to filter out some negative drug pairs based on manually-formulated rules. Instances containing drug pair referring to the same thing and drug pair appearing in the same coordinate structure with more than two drugs (*e.g.*, drug1, drug2, and drug3) will be filtered. Entity mentions are masked with *DRUG* for better generalization and avoiding overfitting.

We train the model with GCN hidden state size of 200, the SGD optimizer with a learning rate of 0.001, a batch size of 30, and 50 epochs. Dropout is applied with a rate of 0.5 for regularization. The contextual embedding size from BioBERT is 768. The focusing parameter γ is set as 1. All hyper-parameters are tuned on the development set.

3.2 Results and Analysis

The experiment results are reported from a 2-layer GCN which achieves the best performance and shown in Table 1. Our model significantly outperforms all previous methods at the significance level of 0.05. To analyze the contributions and effects of the various components in our model, we also perform ablation tests. The ablated GCN model outperforms the LSTM baseline by 3.6% F1 score, which demonstrates the effectiveness of GCN on modeling mention relations through dependency structure. The utilization of contextualized embedding from BioBERT which encodes the contextual information involving sequence order and word disambiguation implicitly helps the model to learn contextual relation patterns, therefore the performance is further improved. We obtain a significant F-score improvement (2.7%) by applying multi-task learning. As over 80% mention pairs are negative samples, the multi-task learning effectively solves the problem by jointly modeling relation identification and classification tasks and applying focal loss to focusing on ambiguous mention pairs, and thus we also gain 3.8% absolute score on recall. Specifically, the F1 score of *int* type is increased from 54.38% to 59.79%.

For the remaining errors, we notice that our model often fails to predict relations when the sentence are parsed poorly due to the complex content which suggests us to seek for more powerful parser tools. Besides, we also observe some errors occurring in extremely short sentences. For example, in the following sentence "*[Calcium]$_{DRUG}$ Supplements/[Antacids]$_{DRUG}$*", our model cannot capture informative representations as the mentions are masked with *DRUG* and the sentence is too concise to offer indicative evidence.

4 Related Work

Traditional feature/kernel-based models for biomedical relation extraction rely on engineered features which suffer from low portability and generalizability (Kim et al., 2015; Zheng et al., 2016; Raihani and Laachfoubi, 2017). To tackle this problem, recent studies apply Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to automatically learn feature representations with input words encoded as pre-trained word embeddings (Zhao et al., 2016; Liu et al., 2016; Quan et al., 2016; Zhang et al., 2017; Zhou et al., 2018a; Sun et al., 2019). Learning representations of graphs are widely studied and several graph neural networks have been applied in the biomedical domain. Lim et al. (2018) proposed recursive neural network based model with a subtree containment feature. Asada et al. (2018) encoded drug pairs with CNNs and used external knowledge base to encode their molecular pairs with two graph neural networks. Here we directly apply syntax-aware GCNs on biomedical text to extract drug-drug interaction.

5 Conclusions and Future Work

We propose a syntax-aware multi-task learning model for biomedical relation extraction. Our model can effectively extract the drug-drug interactions by capturing the syntactic information through graph convolution operations and modeling context information via contextualized embeddings. An auxiliary task with focal loss is designed to mitigate the data imbalance by leveraging the inductive signal from binary classification and increasing the influence of decisive relation types. In the future, we plan to explore more informative parsers like the abstract meaning representation parser to create graph structure and consider leveraging external knowledge to further enhance the extraction quality.

Acknowledgments

This work was supported by the U.S. NSF No. 1741634 and Air Force No. FA8650-17-C-7715. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Maristella Agosti, Giorgio Maria Di Nunzio, Stefano Marchesin, and Gianmaria Silvello. 2019. A relation extraction approach for clinical decision support. *arXiv preprint arXiv:1905.01257*.

Masaki Asada, Makoto Miwa, and Yutaka Sasaki. 2018. Enhancing drug-drug interaction extraction from texts by molecular structure information. In *Proc. ACL2018*.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proc. EMNLP2014*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT2019*.

Rujun Han, Mengyue Liang, Bashar Alhafni, and Nanyun Peng. 2019. Contextualized word embeddings enhanced event temporal relation extraction for story understanding. *arXiv preprint arXiv:1904.11942*.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

Dimitar Hristovski, Dejan Dinevski, Andrej Kastrin, and Thomas C Rindflesch. 2015. Biomedical question answering using semantic relations. *BMC bioinformatics*, 16(1):6.

Sun Kim, Haibin Liu, Lana Yeganova, and W John Wilbur. 2015. Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. *Journal of biomedical informatics*, 55:23–30.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Sangrak Lim, Kyubum Lee, and Jaewoo Kang. 2018. Drug drug interaction extraction from the literature using a recursive neural network. *PloS one*, 13(1):e0190926.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. 2016. Drug-drug interaction extraction via convolutional neural networks. *Computational and mathematical methods in medicine*, 2016.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv preprint arXiv:1703.04826*.

Chanqin Quan, Lei Hua, Xiao Sun, and Wenjun Bai. 2016. Multichannel convolutional neural network for biological relation extraction. *BioMed research international*, 2016.

Anass Raihani and Nabil Laachfoubi. 2017. A rich feature-based kernel approach for drug-drug interaction extraction. *International journal of advanced computer science and applications*, 8(4):324–3360.

Sunil Kumar Sahu and Ashish Anand. 2018. Drug-drug interaction extraction from biomedical texts using long short-term memory network. *Journal of biomedical informatics*, 86:15–24.

Xia Sun, Ke Dong, Long Ma, Richard Sutcliffe, Feijuan He, Sushing Chen, and Jun Feng. 2019. Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. *Entropy*, 21(1):37.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Haoyu Wang, Ming Tan, Mo Yu, Shiyu Chang, Dakuo Wang, Kun Xu, Xiaoxiao Guo, and Saloni Potdar. 2019. Extracting multiple-relations in one-pass with pre-trained transformers. *arXiv preprint arXiv:1902.01030*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zibo Yi, Shasha Li, Jie Yu, Yusong Tan, Qingbo Wu, Hong Yuan, and Ting Wang. 2017. Drug-drug interaction extraction via recurrent neural network with multiple attention layers. In *International Conference on Advanced Data Mining and Applications*, pages 554–566.

Yijia Zhang, Wei Zheng, Hongfei Lin, Jian Wang, Zhihao Yang, and Michel Dumontier. 2017. Drug–drug interaction extraction via hierarchical rnns on sequence and shortest dependency paths. *Bioinformatics*, 34(5):828–835.

Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. *arXiv preprint arXiv:1809.10185*.

Zhehuan Zhao, Zhihao Yang, Ling Luo, Hongfei Lin, and Jian Wang. 2016. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics*, 32(22):3444–3453.

Wei Zheng, Hongfei Lin, Zhehuan Zhao, Bo Xu, Yijia Zhang, Zhihao Yang, and Jian Wang. 2016. A graph kernel based on context vectors for extracting drug–drug interactions. *Journal of biomedical informatics*, 61:34–43.

Deyu Zhou, Lei Miao, and Yulan He. 2018a. Position-aware deep multi-task learning for drug–drug interaction extraction. *Artificial intelligence in medicine*, 87:1–8.

Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018b. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proc. ACL2016*.