Distantly Supervised Document-Level Biomedical Relation Extraction with Neighborhood Knowledge Graphs

Takuma Matsubara, Makoto Miwa, and **Yutaka Sasaki**
Computational Intelligence Laboratory
Toyota Technological Institute
2-12-1 Hisakata, Tempaku-ku, Nagoya, Aichi, 468-8511, Japan
{sd23439, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

Abstract

We propose a novel distantly supervised document-level biomedical relation extraction model that uses partial knowledge graphs that include the graph neighborhood of the entities appearing in each input document. Most conventional distantly supervised relation extraction methods use only the entity relations automatically annotated by using knowledge base entries. They do not fully utilize the rich information in the knowledge base, such as entities other than the target entities and the network of heterogeneous entities defined in the knowledge base. To address this issue, our model integrates the representations of the entities acquired from the neighborhood knowledge graphs with the representations of the input document. We conducted experiments on the ChemDisGene dataset using Comparative Toxicogenomics Database (CTD) for document-level relation extraction with respect to interactions between drugs, diseases, and genes. Experimental results confirmed the performance improvement by integrating entities and their neighborhood biochemical information from the knowledge base. [1]

1 Introduction

The number of documents reporting new interactions and functional/pathway relationships between biochemical entities continues to increase rapidly, and manual registration and maintenance of knowledge bases are becoming increasingly difficult to keep up with the pace (Davis et al., 2021; Wishart et al., 2017). Automatic extraction of biochemical relationships from documents could help in maintaining. Machine learning approaches have been the mainstream of relation extraction (RE) for more than a decade due to their high performance. However, machine learning requires a large amount of labeled data that require time-consuming and costly manual efforts to construct (Yao et al., 2019; Zhang et al., 2017; Miranda et al., 2021).

To alleviate the efforts, Mintz et al. (2009) proposed distantly supervised RE (DSRE), which constructs a distantly supervised corpus in which an unlabeled corpus is automatically labeled using existing knowledge bases. Unlike any manually labeled corpus, a distantly supervised corpus is directly connected to the knowledge base entries because it is labeled based on the target properties between entities described in the knowledge bases. However, existing DSRE uses only the distantly supervised corpus and does not take advantage of other rich information in the knowledge bases, such as the features of the entities, entities that do not appear in the corpus, and a wide range of relationships between entities.

To address the limitations and utilize the rich information in the knowledge base for RE, we propose a model for biomedical DSRE from documents that constructs neighborhood knowledge graphs and integrates them into RE using the distantly supervised corpus. A neighborhood knowledge graph consists of the neighbors of the knowledge base entities appearing in an input document. When constructing neighborhood knowledge graphs, we eliminate links between target entity pairs during distantly supervised learning to avoid label leakage. By using neighbors, we can avoid processing a huge amount of information in the knowledge base in processing each entity pair.

The contributions of this paper are summarized as follows:

- We propose to construct a neighborhood knowledge graph representing the information around all of the entities in an input document.
- We build a novel distantly supervised biomedical RE model that integrates the representations of entities acquired from the neighborhood knowledge graphs with representations

[1] The source code is available at https://github.com/tticoin/nkg-re.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 363–368

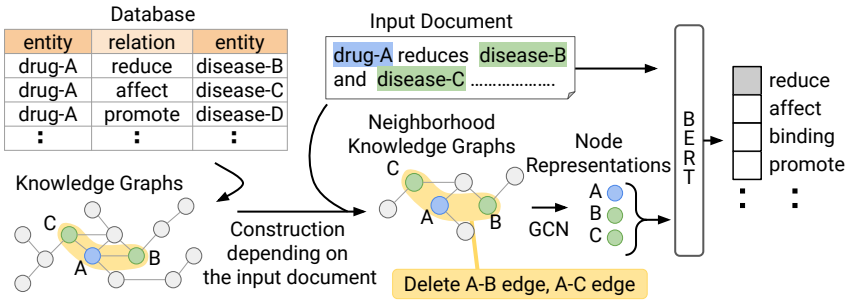

Figure 1: Overview of the proposed model

of the input document.

- Our model achieved the performance improvement by adding the representations of the entities from the Comparative Toxicogenomics Database (CTD) (Davis et al., 2021) over the ChemDisGene dataset (Zhang et al., 2022), a document-level DSRE dataset.

2 Related work

2.1 Distantly Supervised Relation Extraction

DSRE was proposed by Mintz et al. (2009) to train RE models without requiring manually labeled data. Distantly supervised data is built using knowledge bases and a large-scale unlabeled corpus. A relation is labeled between two entities when the two entities registered in the knowledge base co-occur in the unlabeled corpus.

Zhang et al. (2022) proposed ChemDisGene, a document-level distantly supervised relationship extraction dataset, and evaluated a relationship extraction model. ChemDisGene assigns interactions between drugs, diseases, and genes registered in CTD, an integrated database of drug, disease, gene, mutation, and metabolism interactions, to the titles and abstracts of 80,925 documents in the MEDLINE biomedical literature database (Coordinators, 2018). The RE model consists of a pre-trained model BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), a max pooling layer, and a fully connected layer. In addition, Wang et al. (2022) shows high performance on the ChemDisGene dataset using a positive-unlabeled (PU) learning under a prior shift of training data. They also proposed to use a squared ranking loss using a "NA" (none) class score as an adaptive threshold.

2.2 Knowledge Graph Representation Learning

Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017) are a type of Graph Neural Networks (GNNs) that represent the structure of a graph using neural networks and update the representation of each node by convolving the representation of the target node with those of its neighbor nodes.

In addition, several methods have been proposed to represent knowledge graphs using subgraphs extracted from the neighborhood of the target nodes. For example, Learning from Subgraphs, Embeddings and Attributes for Link Prediction (SEAL) proposed in (Zhang and Chen, 2018) shows high performance in the link prediction task using GNNs. They proved that heuristically extracting subgraphs can be approximated by extracting a set of nodes within a certain number of hops. However, there is no method that uses subgraphs in the neighborhood of targets in DSRE.

3 Proposed model

The overview of the proposed model is displayed in Figure 1. In our study, we propose to integrate the information in the neighborhood of the entities from the knowledge graphs that are present in the input document.

3.1 Construction of Neighborhood Knowledge Graphs

To avoid accessing the entire knowledge base in classifying each pair in a document and to utilize the information related to the target document, we construct the neighborhood knowledge graph based on triples that are close to all the entities in an input document in the knowledge base. To obtain the close triples, we extract a set of triples that are connected to the entities within a certain number of hops. We remove the triples between the pairs

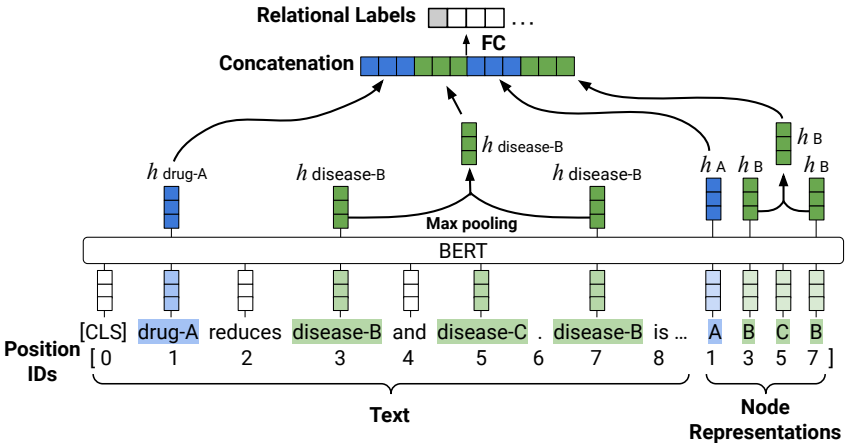

Figure 2: Proposed RE model

that provide the labels of the distantly supervised training data to avoid label leakage. To accelerate the training, we precalculate the neighborhood knowledge graph for each document once before training the model, while in prediction, we need to generate the neighborhood knowledge graph.

3.2 Integrated use of language information and neighborhood knowledge graphs

We perform RE by integrating entity information of the neighborhood knowledge graphs into the distantly supervised training data. The proposed RE model is illustrated in Figure 2.

For the base RE model, we employ the BERT-based RE model in Zhang et al. (2022). Specifically, we first extract the representations corresponding to the mentions of each entity of the target pair from the final layer of BERT. We next obtain the representation of each entity through max pooling. The representations of the entities are then concatenated and classified into relational labels through a fully connected layer (FC).

When integrating neighborhood knowledge graphs, the constructed neighborhood knowledge graph is processed by GCN, and the node representations corresponding to the entities are fed to BERT along with the input document. Specifically, the node representations corresponding to the entities in each input document are added to the document while matching the position identifiers of the entities in the document and the node representations of the graph (Zhong and Chen, 2021). We obtain the representations of each graph node of the target pair by pooling the corresponding representations from the final layer of BERT corresponding to the node, and we concatenate the pooled representations with the pooled representations of the entity mentions for classification.

GCN and BERT are trained simultaneously using Adam (Kingma and Ba, 2015) as the optimization method and the cross entropy loss.

4 Experimental Settings

We trained models using ChemDisGene (Zhang et al., 2022), distantly supervised data (CTD-derived) as the training and development data, and manually-labeled data (Curated) as the test data. The CTD-derived data contain 76,942 and 1,521 medical references as training and development data, respectively. The Curated data contain 523 medical references as test data. We used the micro-averaged F1 measure as the evaluation metric. The task is to classify entity pairs of chemical, disease, and gene into 14 relation types: two drug-disease relation types, ten drug-gene relation types, two gene-disease relation types, or no relation.

The entities of the neighborhood knowledge graphs are the drugs, diseases, and genes of CTD. The neighborhood knowledge graphs contain entities within two hops from any target entities corresponding to entities in each input document. All nodes with more than 100 edges are randomly restricted to 100 edges to reduce computation time. The statistics of the ChemDisGene and CTD are summarized in Appendix B.

The baseline model is the same as Zhang et al. (2022), excluding the neighborhood knowledge graphs from the proposed model. Following Zhang et al. (2022), we used PubMedBERT (Gu et al., 2021) for BERT. The input and output vectors are set to 768 dimensions, the maximum length of the text is 512, and the number of GCN layers is two.

	F1[%]
PubMedBERT	42.5 ± 0.2
PubMedBERT+BRAN	43.6 ± 0.4
PubMedBERT+Neighborhood KG	44.0 ± 0.3
PubMedBERT+BRAN+Neighborhood KG	43.7 ± 0.5
PubMedBERT (Zhang et al., 2022)	42.1
PubMedBERT+BRAN (Zhang et al., 2022)	43.8
PubMedBERT+SSR-PU+ATLOP (Wang et al., 2022)	48.6 ± 0.2

Table 1: The comparison with the existing studies (Zhang et al., 2022; Wang et al., 2022) using the curated test data

Relation	Title and Abstract
Gene-Dis: therapeutic Gene: mir-543 Disease: cervical cancer	**miR-543** inhibits **cervical cancer** growth and metastasis by targeting TRPM7. Dysregulation of **miR-543** has been implicated to play crucial roles in various human cancers. · · ·
Chem-Gene: transport-increases Chemical:glucose Gene: Fibroblast growth factor 21	**Fibroblast growth factor 21** secretion enhances **glucose** uptake in mono (2-ethylhexyl) phthalate-treated adipocytes. Previous cellular accumulation of mono (2-ethylhexyl) phthalate (MEHP) disturbed energy metabolism in adipocytes, where **glucose** uptake was significantly increased.· · ·

Table 2: Examples of correctly predicted relations not registered in the CTD

5 Results

The performance of extracting relationships between drugs, diseases, and genes from the documents on the Curated test data is shown in Table 3. The detailed performance for each relation type is summarized in Appendix A. The proposed model (+Neighborhood KG) improved the micro-averaged F-score by 1.5 percentage points compared to the baseline (PubMedBERT). This result suggests that the information of neighborhood knowledge graphs can improve prediction performance and that the relationship extraction can take the knowledge graph information into account.

Compared to the scores of Zhang et al. (2022), our baseline PubMedBERT result showed slightly higher performance than PubMedBERT, while our PubMedBERT+BRAN model showed comparable performance, as shown in Table 3. This shows we could correctly reimplement these models.

We also incorporated our neighborhood KG into the PubMedBERT+BRAN model, but it did not show performance improvement. Furthermore, compared to the scores of Wang et al. (2022), our proposed model showed a performance lower than PubMedBERT+SSR-PU+ATLOP. However, their model does not use the neighborhood knowledge graph, and our proposed method can be incorporated into their model. We will leave the adaptation of our approach to these models for future work.

In order to examine the influence of the neighborhood knowledge graph information on RE, we examined the cases that are not registered in CTD but annotated in the Curated test data. The baseline extracted 24.7 relations, whereas the proposed model extracted 40.2 relations on an average of 5 models. This result suggests that the neighborhood graph information is helpful in extracting new relations when the entities are in the knowledge base. The examples that are correctly extracted by the proposed model are shown in Table 2.

6 Conclusion

we proposed to integrate the information in the neighborhood of the entities from the knowledge graphs that are present in the input document. We trained and evaluated the proposed model on the ChemDisGene dataset and found that the introduction of the neighborhood knowledge graphs improved the micro-averaged F-score by 1.5 percentage points. We also confirmed that the proposed model is able to extract relationships that are not registered in the knowledge base, which could not be extracted by the model using only language information.

For future work, we will apply our model to state-of-the-art models. We will also investigate the way to utilize the extracted results in enhancing knowledge graphs to link the text and knowledge graph information more deeply and leverage the knowledge graph information more effectively.

Limitations

We have applied our neighborhood knowledge graphs to the PubMedBERT and PubMedBERT+BRAN models and show the effectiveness of the graphs on the PubMedBERT model. We have not deeply investigated how our approach cooperates with other enhancements, and the performance is lower than the state-of-the-art model (Wang et al., 2022).

Acknowledgements

This work was supported by JSPS Grant-in-Aid for Scientific Research JP20K11962.

References

NCBI Resource Coordinators. 2018. Database resources of the national center for biotechnology information. *Nucleic Acids Res.*

Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. 2021. Comparative toxicogenomics database (CTD): update 2021. *Nucleic acids research*, 49(D1):D1138–D1143.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. *ICLR*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Antonio Miranda, Farrokh Mehryary, Jouni Luoma, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. 2021. Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. In *Proceedings of the seventh BioCreative challenge evaluation workshop*.

Ye Wang, Xinxin Liu, Wenxin Hu, and Tao Zhang. 2022. A unified positive-unlabeled learning framework for document-level relation extraction with different levels of labeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4123–4135, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. 2017. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of ACL 2019*.

Dongxu Zhang, Sunil Mohan, Michaela Torkar, and Andrew McCallum. 2022. A distant supervision corpus for extracting biomedical relationships between chemicals, diseases and genes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1073–1082, Marseille, France. European Language Resources Association.

Muhan Zhang and Yixin Chen. 2018. Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pages 5171–5181. Curran Associates, Inc.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

A Detailed evaluation results

Table 3 shows the extraction performance of the baseline and our model for each relation type.

Relation	baseline PubMedBERT P [%]	R [%]	F1 [%]	Our model PubMedBERT+Neighborhood KG P [%]	R [%]	F1 [%]
Chem-Dis: marker/mechanism	79.65	30.61	44.18 ± 2.29	80.87	32.96	46.77 ± 1.05
Chem-Dis: therapeutic	79.35	23.11	35.60 ± 3.09	79.85	26.28	39.24 ± 5.68
Chem-Gene: activity-decreases	70.93	23.73	35.15 ± 4.19	71.73	24.66	36.35 ± 2.95
Chem-Gene: activity-increases	73.07	30.39	42.74 ± 3.48	72.92	27.58	39.77 ± 4.94
Chem-Gene: binding-affects	73.49	31.79	44.21 ± 4.04	61.04	48.76	54.00 ± 1.70
Chem-Gene: expression-affects	19.05	3.25	5.56 ± 8.33	36.09	2.20	3.63 ± 5.56
Chem-Gene: expression-decreases	77.81	38.30	51.16 ± 1.90	73.00	38.00	49.95 ± 1.52
Chem-Gene: expression-increases	63.21	47.17	53.95 ± 1.32	59.54	48.87	53.59 ± 1.36
Chem-Gene: localization-affects	51.40	43.09	46.75 ± 4.22	61.15	32.68	41.01 ± 8.75
Chem-Gene: metabolic processing-decreases	54.12	57.90	55.89 ± 2.83	53.67	51.58	51.06 ± 3.00
Chem-Gene: metabolic processing-increases	41.52	34.13	37.30 ± 3.39	44.75	36.32	39.74 ± 2.63
Chem-Gene : transport-increases	46.89	36.59	40.61 ± 8.13	55.93	36.59	42.19 ± 4.10
Gene-Dis: marker/mechanism	83.75	20.27	32.27 ± 8.25	81.88	23.39	36.15 ± 4.16
Gene-Dis: therapeutic	53.33	1.63	3.10 ± 3.48	61.67	1.95	3.74 ± 3.47
Micro-average	69.51	30.62	42.47 ± 0.16	67.73	32.57	44.01 ± 0.34

Table 3: Evaluation of manually labeled test data in Curated. The F-score shows the mean and standard deviation of the five evaluations.

data	paper	chem	dis	gene	relation
train	76,942	7,187	2,413	5,391	167,005
dev	1,521	759	283	852	3,290
test	523	670	318	887	3,833

Table 4: Statistics of ChemDisGene, a document-level DSRE dataset, where only tests are manually tagged. Each column shows the number of instances in each data.

domain	unique head	unique tail	triple
Chem-Gene	14,346	53,832	2,274,465
Chem-Dis	10,249	3,285	104,186
Gene-Dis	8,807	5,857	33,449

Table 5: Statistics of CTD used in this experiment. "domain" denotes the types of the entity pair of a triple, "unique head" and "unique tail" denote the numbers of unique nodes in the knowledge graph constructed from CTD, and "triple" denotes the number of instances.

B Dataset statistics

Tables 4 and 5 show the statistics of the data set.