BioRED task DUTIR-901 submission: Enhancing Biomedical Document-Level Relation Extraction through Multi-Task Method

Jiru Li, Dinghao Pan, Zhihao Yang*, Yuanyuan Sun, Hongfei Lin, and Jian Wang

School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China

*Corresponding author: yangzh@dlut.edu.cn

Abstract
Biomedical relation extraction is central to biomedical natural language processing and is crucial for various downstream applications. The BioCreative VIII BioRED Track focuses on extracting entity relationships from biomedical literature titles and abstracts and classifying relations that are novel findings. This paper describes our method used to create submissions to identify all the relationships between human-annotated entities and implement an end-to-end system to identify all the asserted relationship subtasks. In our method, a multi-task training approach is employed for fine-tuning a pre-trained language model in the field of biology. Based on a broad spectrum of carefully designed tasks, our multi-task method not only extracts relations of better quality due to more effective supervision, but also achieves a more accurate classification of whether the entity pairs are novel findings. The official results on the test set show that our best submission achieves the F1-scores of 0.4441 on Subtask1 and 0.2334 on Subtask2.

Keywords: relation extraction; multi-task; biomedical natural language processing

Introduction
The development of biomedical natural language processing technology has made it possible to extract relationships between relevant concepts from text. Early research primarily focused on sentence-level relation extraction (RE). However, as more complex relationships are often expressed across multiple sentences, there is a growing trend in recent research towards document-level relation extraction tasks (1, 2).

Due to the increased complexity of entity interactions in the documents and the challenging nature of context modeling (3), we have employed a multi-task learning approach to enhance the performance of the language model in the field of biology. This approach aims to improve the model's ability to capture contextual information and entity type information by enhancing its capabilities in related tasks. Consider the motivating example in Figure 1:

(1) Coreference Resolution (CR): The descriptions "matrix metalloproteinases 2" and "MMP-2" in the document refer to the same entity in the gene database (4), representing different mentions of the same entity within the text. Modeling coreference relationships is crucial for the model to contextually capture entity semantics. Therefore, for a given document, we use

Coreference Resolution (CR) to parse various contextual representations conveyed by different mentions of the same entity.

(2) Entity Pair Typing (EPT): Entity pair types play a significant guiding role in relation type determination. In the BioRED (5) dataset, distinct entity pair types correspond to different sets of relation types. We construct entity pair types as labels used during training by extracting manually annotated entity type combinations from data for each sample.

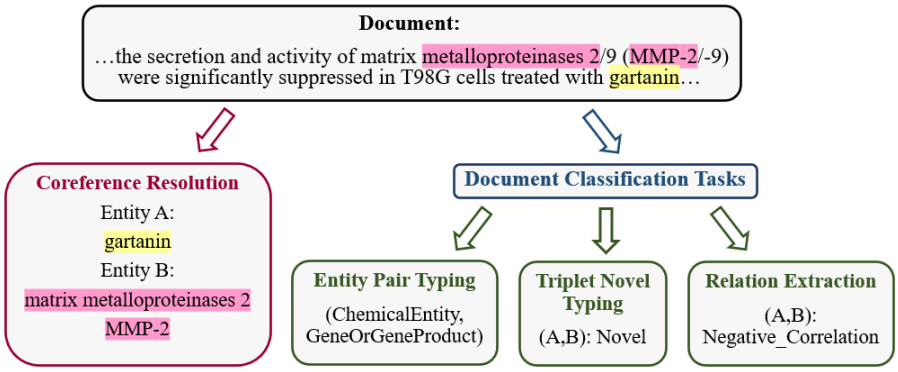

Figure. 1. The example is adapted from BioRED. There are four related tasks involved, from an input document with annotated entity mentions to RE and TNT output. These tasks are interconnected because CR, TNT, and RE capture text contexts, while ET preserves entity type information.

By incorporating these two additional designed tasks, we have constructed a multi-task model capable of handling four tasks: Coreference Resolution (CR), Entity Pair Typing (EPT), Relation Extraction (RE), and Triplet Novel Typing (TNT). Through multi-task tuning, the model shows significant performance improvement in both RE and TNT tasks.

Methods

Preprocess

We utilized a domain-specific pre-trained language model (PLM) called PubMedBERT (6) for context encoding in the field of biology. PubMedBERT is developed specifically for biomedical text and generates its vocabulary from scratch. It incorporates biomedical domain knowledge into the BERT-based pre-training, achieving state-of-the-art performance on various biomedical natural language understanding tasks.

In more detail, for each document d, we inserted "[CLS]" at the beginning and "[SEP]" at the end of the text. For every entity pair $\left(e_i^d, e_j^d\right)$ within the document d we constructed a sample s, regardless of whether they had a relationship. For each mention $m^d \in M_d$ in the sample, we inserted special symbols "@\E" and "E/@" to identify the positions of mentions.
During the training process, we filter out samples that involve entity pair types for which no candidate relationships exist. As shown in Eq. (1), we tokenize the document, composed of titles and abstracts, and input it at the token level into the PLM to obtain encoded representations of the text.

$$V = PLM(x_1, x_2, \ldots, x_m) \quad (1)$$

Where $V \in R^{\{m \times h\}}$ is the encoded representation of preprocessed input text. m is the max length of the input text. h is the embedding dimension of the PLM.

Multitask Method

A. Coreference Resolution(CR)

We define the relationships between mentions within the BioRED dataset based on whether they refer to the same standardized entity. We categorize the relation of mentions as coreference if the mentions refer to the same standardized entity. Our objective is to equip a pre-trained encoder with the capability to automatically identify whether mention pairs are coreference using this newly defined task. we use the representation of the special symbol "@\E" in V as the representation for mention m_d. consider a pair of mentions $\left(m_i^d, m_j^d\right)$, We use a bilinear layer to compute the probability of the coreference resolution task:

$$\mathrm{P}_{\mathrm{i,j}}^{\mathrm{CR}} = \delta\left(\mathrm{m}_{\mathrm{i}}^{\mathrm{d}} \mathrm{A}_{\mathrm{CR}} \mathrm{m}_{\mathrm{j}}^{\mathrm{d}} + \mathrm{b}\right) \quad (2)$$

Where δ is the softmax function, $\mathrm{A}_{\mathrm{CR}} \in R^{(\mathrm{h} \times 2 \times h)}$ is a trainable neural layer, which attends to the mentions simultaneously and $b \in R$ is the model prior bias. As shown in Eq. (3), we employ cross-entropy to calculate the loss for the coreference resolution task, which is used to update the parameters of the CR classifier and the encoding layer in the model.

$$L_s^{CR} = CrossEntropy\left(P_{i,j}^{CR}, \overline{y}_{i,j}^{CR}\right) \quad (3)$$

Where $\overline{y}_{i,j}^{CR}$ is the coreference label extracted from the annotation data.

B. Document Classification Tasks(DCT)

Within BioRED, there are inherent constraints and mapping relationships between entity types and relationship categories. Therefore, we believe that enhancing the entity type recognition capability of the model can significantly enhance the quality of relationship category recognition. To address this, we have established the Entity Pair Typing (EPT) task based on annotated data. In essence, it involves extracting manually labeled entity type combinations for each sample from existing data to construct entity pair types as labels for training.

In contrast to individually classifying entities, this approach provides a more cohesive task structure and allows for a broader focus on holistic information during classification. As we create a sample for each distinct entity pair within the same document, and all mentions of two different entities in a sample are enclosed by special symbols, we can define Entity Pair Typing (EPT), Triplet Novel Typing (TNT), and Relation Extraction (RE) tasks as a unified document classification format. Specifically, for a sample constructed for a pair of entities$\left(e_i^d, e_j^d\right)$, we create separate classifiers for different tasks $t_i \in \{ETP, RE, TNT\}$, and utilize the "[CLS]" representation $v_{i,j}^{[cls]}$ with three distinct classifiers to obtain predictions for these three tasks:

$$P_{i,j}^{t_i} = \delta\left(W_{t_i} v_{i,j}^{[cls]} + b_{t_i}\right) \quad (4)$$

Where $W_{t_i} \in R^{n \times h}$ is a trainable neural layer, $b \in R$ is the model prior bias and n is the number of classes in tasks. Similarly, we also use cross-entropy to obtain the loss for document classification tasks and update the parameters of the classifiers for each task and the shared encoding layer:

$$L_s^{t_i} = CrossEntropy\left(P_{i,j}^{t_i}, \overline{y}_{i,j}^{t_i}\right) \quad (5)$$

C. Multitask Train

In essence, our objective is to elevate the performance of both the RE task and the TNT task. We achieve this by integrating all tasks through the minimization of a multi-task learning objective. We allocate distinct loss weights to all tasks and train the models separately for the RE and TNT tasks. For the RE model, the training loss is expressed in Eq. (6):

$$L_s = L_s^{RE} + \sum \eta^{t_i} L_d^{t_i} \quad (6)$$

To enhance the performance and robustness of the single model, we use an adversarial training approach (7) after loss backpropagation. As shown in Eq. (7), based on the gradient g at the embedding layer we add an adversarial perturbation r_{adv} to the word vector to generate adversarial samples. The newly generated adversarial samples are fed into the model to generate the adversarial loss L_{adv}, and the losses are back-propagated to obtain the gradients generated by the adversarial sample losses at each layer of the model. Since the gradients generated by the original sample losses have not yet been used by the optimizer for parameter updating, the gradients of the original samples are accumulated with those of the adversarial samples. We use this accumulated gradient to update the parameters of the model, which effectively improves the performance and robustness of the model.

$$\begin{aligned} r_{adv} &= \epsilon g / \left|\left|g\right|\right|_2 \\ x_{adv} &= x + r_{adv} \\ L_{adv} &= Model(X_{adv}) \end{aligned} \quad (7)$$

Ensemble Model

For the final test, we use the k-fold cross-validation method to improve our performance. We integrate the original training set and validation set as a new training set, and use ten-fold cross-validation method to train ten different weights models, and then integrate the prediction results of each model through voting. Through this method, we use the diversity of the training set to obtain models with different targeted capabilities, thereby improving the RE and TNT performance and robustness of the entire system.

Results

Task1 Result

During this task, we submitted five runs as our final submissions. Our submitted five runs in the main task are based on the following configurations.

- Run1: We obtain the run1 results using the best five multi-task models trained on the training dataset of RE and TNT tasks.
- Run2: We integrate the multi-task models trained by k-fold cross-validation to obtain RE results, and use the same method as run1 to obtain TNT results.
- Run3: We integrate the multi-task models and one-task models trained by k-fold cross-validation to obtain the RE results, and use the same method as run1 to obtain TNT results.
- Run4: We integrate two sets of multi-task models with different task weight hyperparameters trained by k-fold cross-validation and one-task models trained by k-fold cross-validation to obtain the RE results. For TNT, we use the multi-task models trained by k-fold cross-validation to obtain TNT results.
- Run5: We use the same method as run1 to obtain RE results, and integrate the multi-task models and one-task models trained by k-fold cross-validation to obtain the TNT results.

Table 1: task1 results

Eval Schema	Runs						
	1	2	3	4	5	Median	Average
Entity Pair	75.35	75.49	75.00	**75.59**	75.59	73.56	67.03
+Relation Type	55.18	55.89	56.12	**56.67**	56.67	53.17	47.74
+Novelty	43.17	43.60	43.89	**44.41**	44.39	40.73	35.22

Our best ensemble model achieved a novelty F1 score of 44.41 on the test set, surpassing the average performance of all participating models by 9.19 points and outperforming the median by 3.68 points.

Task2 Result

During this track, we submitted four valid and distinct runs as our final submissions. Our submitted four runs in the main task are based on the following configurations. Due to the deadline, the method adopted by track2 is simpler than that of track1. For all runs, we used the PubTator API (8) to obtain entity recognition and entity normalization results.

- Run1: We use the best multitask model to obtain the results for both RE and TNT tasks.
- Run2: We integrate the results of the 5 best RE models and the results of the 5 best TNT models for the final results.
- Run3: We use train data and val data together to train the multitask model and stop training at epochs 50 for the RE task and epoch 15 for the TNT task.
- Run4: We integrate five models which are trained with train data and val data for the RE task and TNT task.

Our best ensemble model achieved a novelty F1 score of 22.34 on the test set, surpassing the average performance of all participating models by 6.09 points and outperforming the median by 5.55 points.

Table 2: task2 results

Eval Schema	Runs					
	1	2	3	4	Median	Average
NER	78.58	78.58	78.58	78.58	78.58	76.87
ID	76.35	76.35	76.35	**76.35**	66.81	63.36
Entity Pair	39.93	40.38	41.07	**41.27**	34.47	28.62
+Relation Type	30.40	**30.12**	30.66	31.03	25.40	21.39
+Novelty	23.05	23.22	23.26	**23.34**	17.79	16.25

Conclusion

In this paper, we propose explicitly training the model to capture context and entity type information related to the RE (Relation Extraction) and TNT (Triplet Novel Typing) tasks through the joint training of multiple relevant tasks. Leveraging a carefully designed set of diverse tasks, our approach extracts relationships of higher quality due to more effective supervision, resulting in improved accuracy for novel classification. Furthermore, we enhance the model's performance through model integration. The experimental results clearly demonstrate that our model outperforms the average performance of all participating models in both Track 1 and Track 2.

Funding

This work was supported by the grants from the Natural Science Foundation of China (No. 62276043) and the Fundamental Research Funds for the Central Universities (No. DUT22ZD205).

References

1. Yao, Y. , Ye, D. , Li, P., et al. (2019). Docred: a large-scale document-level relation extraction dataset. arXiv preprint arXiv:1906.06127.

2. Cheng, Q., Liu, J., Qu, X., et al. (2021). HacRED: A large-scale relation extraction dataset toward hard cases in practical applications. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 2819-2831).

3. Xu, B., Wang, Q., Lyu, Y., et al. (2021). Entity structure within and throughout: Modeling mentions dependencies for document-level relation extraction. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 16, pp. 14149-14157).

4. Lee, K., Lee, S., Park, S., et al. (2016). BRONCO: Biomedical entity Relation ONcology COrpus for extracting gene-variant-disease-drug relations. Database, 2016, baw043.

5. Luo, L., Lai, P. T., Wei, et al. (2022). BioRED: a rich biomedical relation extraction dataset. Briefings in Bioinformatics, 23(5), bbac282.

6. Gu, Y., Tinn, R., Cheng, H., et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), 1-23.

7. Miyato, T., Dai, A. M., and Goodfellow, I. (2016). Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605.07725.

8. Wei, C. H., Allot, et al. (2019). PubTator central: automated concept annotation for biomedical full text articles. Nucleic acids research, 47(W1), W587-W593.