IMPROVING BIOMEDICAL NAMED ENTITY RECOGNITION WITH A UNIFIED MULTI-TASK MRC FRAMEWORK

Yiqi Tong[1,2]*, Fuzhen Zhuang*[1,2*]*, Deqing Wang*[2]*, Haochao Ying*[3]*, and Binling Wang*[4]

[1]Institute of Artificial Intelligence, Beihang University, Beijing 100191, China
[2]SKLSDE, School of Computer Science, Beihang University, Beijing 100191, China
[3]School of Public Health, Zhejiang University, Hangzhou 310058, China
[4]School of Informatics, Xiamen University, Xiamen 361005, China

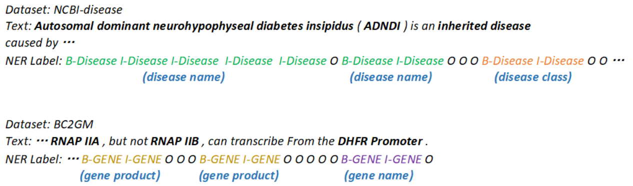
ABSTRACT

The prior knowledge, such as expert rules and knowledge base, has been proven effective in the traditional Biomedical Named Entity Recognition (BioNER). Most current neural BioNER systems use this external knowledge for pre-processing or post-editing instead of incorporate it into the training process, which cannot be learned by the model. To encode prior knowledge into the model, we present a unified multi-task Machine Reading Comprehension (MRC) framework for BioNER. Specifically, in the MRC task, the question sequences are derived from the standard BioNER dataset. We introduce three kinds of prior knowledge at query sequences, including Wikipedia, annotation scheme, entity dictionary. Then, our model adopts a multi-task learning strategy to joint training the main task BioNER and the auxiliary task MRC. Finally, experimental results on three benchmark datasets validate the superiority of our BioNER model compared with various state-of-the-art baselines.

Index Terms— Prior knowledge, Biomedical named entity recognition, Machine reading comprehension, Multi-task learning

1. INTRODUCTION

Biomedical Named Entity Recognition (BioNER) refers to the task of detecting the span and the semantic category of entity mentions (e.g., chemicals, diseases, and genes) from a chunk of text. Existing methods usually formalize the BioNER task into a sequence labeling problem. Compared with other NER tasks in the public domain like news, BioNER is more challenging due to its high degree of professionalism and some ethical concerns [1]. For example, given the mixed-case sentences shown in Figure 1, it is difficult to understand that "ADNDI" is a disease entity and "RNAP IIA" is a gene entity. On the other hand, even the same type of biomedical entities have differences in granularity. Intuitively, we usually treat specific disease names like "autosomal dominant neurohypophyseal diabetes insipidus " as biomedical entities. In fact, disease classes, such as "inherited disease" are also belong to the disease entity, which is frequently ignored. These cases illustrate the necessity of using prior knowledge in BioNER.



Dataset: NCBI-disease
Text: ***Autosomal dominant neurohypophyseal diabetes insipidus*** *(* ***ADNDI*** *) is an* ***inherited disease*** *caused by ···*
NER Label: B-Disease I-Disease I-Disease I-Disease I-Disease O B-Disease I-Disease O O O B-Disease I-Disease O O ···

Dataset: BC2GM
Text: ··· ***RNAP IIA*** *, but not* ***RNAP IIB*** *, can transcribe From the* ***DHFR Promoter*** *.*
NER Label: ··· B-GENE I-GENE O O O B-GENE I-GENE O O O O O B-GENE I-GENE O

Fig. 1. Example illustrating challenges in BioNER.

Traditional rule- or dictionary-based BioNER methods have proved the effectiveness of introducing external knowledge resources by well-designed feature engineering. However, these hand-crafted features are both model- and entity-specific, resulting in the weak generalization ability of the BioNER model [2]. In recent years, the neural network has been used in BioNER due to its outstanding performance [3]. Most neural BioNER models are implemented using Bi-directional Long Short-Term Memory (Bi-LSTM) networks or pre-training models, such as BioBERT [4], to automatically exploit semantic features, which no longer require costly manual work. To further improve the model performance, some prior works [5, 6] try to integrate prior knowledge like expert rules or knowledge base into neural networks. In particular, they employ this external knowledge for pre-processing or post-editing instead of incorporating it into the training process of the model. In this case, the model will not learn this valuable external information. Furthermore, these methods will reduce the generalization of the model, thus weakening the advantages of the deep learning approach.

This paper proposes a unified multi-task Machine Reading Comprehension (MRC) framework for BioNER to encode

[*] Corresponding author (E-mail: zhuangfuzhen@buaa.edu.cn)

This research work is supported by the National Key Research and Development Program of China under Grant No. 2021ZD0113602, the National Natural Science Foundation of China under Grant No. 62176014, U1836206.

prior knowledge into the model. In previous works, golden BioNER categories are merely class indexes and lack for semantic information, so we convert the main task BioNER into a SQuAD-style [7] MRC task. In the MRC task, each entity type is characterized by a natural language query, and entities are extracted by answering these queries with given contexts. When designing queries, we attempt to introduce different prior knowledge, such as Wikipedia, annotation scheme, and entity dictionary. Finally, we apply a multi-task strategy to joint training the main task BioNER and the auxiliary task MRC by sharing parameters, so as to make full use of label data.

We carry out experiments on three benchmark BioNER datasets of BC2GM [8], BC5CDR-chem [9], and NCBI-disease [10], respectively. The experimental results show that our model can achieve 0.46%, 0.30%, and 0.43% F1-score improvement over the state-of-the-art baseline model BioBERT, verifying the applicability and generality of our framework.

2. METHODOLOGY

2.1. Task Definition

Main Task We take BioNER as our main task. For a sequence of sentences $\{s_1, s_2, ..., s_N\}$ in a document, N is the number of training samples and $s_i (1 \leq i \leq N)$ of length n_i is represented as a sequence of tokens $s_i = \{x_1, x_2, ..., x_{n_i}\}$. The goal of BioNER is to use intra-sentence information of the sentence s_i to predict the tag sequence $t_i = \{y_1, y_2, ..., y_{n_i}\}$, where $y_i (1 \leq i \leq n_i) \in Y$. In this work, Y is predefined and each y_i is annotated manually by the standard BIO tagging scheme to represent the boundary of the entity and the entity type. For example, the **B**egining of the disease entity is labeled as "B-Disease", and the **I**nside of the disease entity is labeled as "I-Disease". All **O**ther tokens not describing entities of interest are labeled as "O".

Auxiliary Task We take MRC as our auxiliary task. In a collection of textual training samples $\{(p_i, q_i, a_i)\}_{i=1}^{N}$, where p is a text sequence, q is a question regarding p, and a is an answer about q. The goal of the MRC task is to learn a predictor f which takes p and a corresponding question q as inputs t and gives the answer a as output. In this work, p_i is the same as the raw input s_i in BioNER. The query q_i is designed by ourselves, and the answer a_i is the target entity span.

2.2. Model Architecture

The architecture of our proposed multi-task MRC BioNER model is shown in Figure 2, which includes two parts: the shared encoder and task-specific layers. For the BioNER task, following with [4], we add two special tokens at the start ([CLS]) and the end ([SEP]) to make the input $s_i = \{[CLS], x_1, ..., x_{n_i}, [SEP]\}$. Similarity, we make the input $m_i = \{[CLS], q_i, [SEP], p_i, [SEP]\}$ for the MRC task. According to the study of [2], changing the position of the query q_i and the text sequence p_i has almost no effect on model performance. Table 1 lists some examples of the queries we construct to introduce prior knowledge, where the sources include Wikipedia, annotation scheme, and entity dictionary. Wikipedia contains standard definitions of various entities, which can help the model understand what it represents. For example, according to the definition of the gene, we can know that RNA or gene products such as "RNAP IIA" are all gene entities. Annotation scheme notes are the golden guidelines provided to the annotators of the dataset by the dataset builder. Based on this knowledge, we can understand why "inherited disease" also belongs to the disease entity. Furthermore, the entity variables such as $GENE_i$, $Chemical_i$, and $Disease_i$ represent the annotated ones selected randomly from the training/development set of the target dataset. We note that all query constructions do not require any additional annotation data.

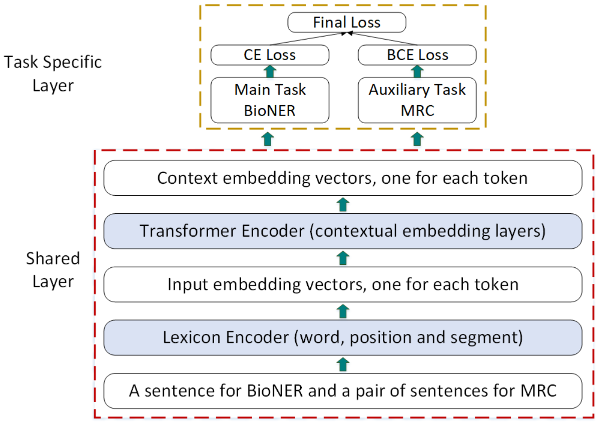

Task Specific Layer
Final Loss
CE Loss
BCE Loss
Main Task BioNER
Auxiliary Task MRC
Shared Layer
Context embedding vectors, one for each token
Transformer Encoder (contextual embedding layers)
Input embedding vectors, one for each token
Lexicon Encoder (word, position and segment)
A sentence for BioNER and a pair of sentences for MRC

Fig. 2. The overall framework of our model.

In the bottom encoder, we adopt pre-trained BioBERT-base v1.1[1] to embedding s_i and m_i as $H_i = \{h_1, ..., h_{n_i}\}$ respectively, where $h_i (1 \leq i \leq n_i) \in \mathbb{R}^d$ and d is the dimension of the hidden state. All hidden states are hard shared between the task-shared layers, including BioNER and MRC. In the task-specific layers, for the MRC task, the model first predicts the probability of each token being a start and end index, respectively, which is calculated as follows:

$$P_{start}^{i} = \text{softmax}(W_{start} H_i + b_{start}), \tag{1}$$

$$P_{end}^{i} = \text{softmax}(W_{end} H_i + b_{end}), \tag{2}$$

where the weight matrix W_{start}, W_{end} and the bias vector b_{start}, b_{end} are independent trainable parameters. Because a MRC input m_i may contain multiple entities, we use the following formula to get the start and end positions of multiple entities, respectively.

$$I_{start}^{i} = \{i | \operatorname{argmax}(P_{start}^{i}) = 1, i = 1, 2, ..., n_i\}, \tag{3}$$

[1] https://github.com/naver/biobert-pretrained

Dataset	Source	Query
BC2GM	Wikipedia	A gene is a sequence of nucleotides in DNA or RNA that encodes the synthesis of a gene product, either RNA or protein.
	Annotation scheme	A gene entity identified mention of a specific gene name or a gene product name.
	Entity dictionary	Can you detect entities like $GENE_i$?
BC5CDR-chem	Wikipedia	Chemical refer to a distinct compound or substance, especially one which has been artificially prepared or purified.
	Annotation scheme	A chemical entity is a physical entity of interest in chemistry including molecular entities, parts thereof, and chemical substances.
	Entity dictionary	Can you detect entities like $Chemical_i$?
NCBI-disease	Wikipedia	A disease is a particular abnormal condition that negatively affects the structure or function of all or part of an organism, and that is not due to any immediate external injury.
	Annotation scheme	A disease entity refers to a specific disease name or a disease class which could be described as a family of many specific diseases.
	Entity dictionary	Can you detect entities like $Disease_i$?

Table 1. Examples of constructed queries.

$$I^i_{end} = \{i|\,\mathrm{argmax}(P^i_{end}) = 1, i = 1, 2, ..., n_i\}. \quad (4)$$

In the end, the loss functions of the MRC task are defined as follows:

$$L_{start} = \mathrm{BCE}(P_{start}, Y_{start}), \quad (5)$$

$$L_{end} = \mathrm{BCE}(P_{end}, Y_{end}), \quad (6)$$

$$L_{MRC} = L_{start} + L_{end}, \quad (7)$$

where BCE denotes the binary cross-entropy loss function. For the BioNER task, we feed H_i to a linear layer to get the probability of the tag sequence t_i. Moreover, we still use the softmax classifier for a fair comparison with previous works [4, 11, 12] instead of sequence labeling algorithms such as Conditional Random Field (CRF). Finally, the total loss of our model is calculated as follows:

$$P^i_{BioNER} = WH_i + b, \quad (8)$$

$$L_{BioNER} = -\frac{1}{N}\sum_{i=1}^{N} logp(t_i|P^i_{BioNER}), \quad (9)$$

$$L_{total} = \alpha L_{BioNER} + \beta L_{MRC}, \quad (10)$$

where $\alpha, \beta \in [0, 1]$ are hyper-parameters to control the contributions towards the overall training objective.

2.3. Joint Training

At the training time, we first initialize the shared encoder and randomly initialize the task-specific layers, which are the same with [1]. In each epoch, we randomly select a mini-batch from the training set. Then we update the model according to the task-specific objectives of the main task and the auxiliary task. We use multi-task learning to jointly train the whole model in an end-to-end fashion, which is more effective and easy to train.

Dataset	Sentences	Entity Type and Counts
BC2GM	20,131	Gene/Protein (24,583)
BC5CDR-chem	13,938	Chemical(15,935)
NCBI-disease	7,287	Disease(6,881)

Table 2. Dataset description, we use BC2GM, BC5CDR-chem and NCBI-disease to conduct experiments.

3. EXPERIMENTS

3.1. Datasets and Settings

We evaluate the performance of our proposed approach on three BioNER datasets: BC2GM, BC5CDR-chem, and NCBI-disease. Table 2 summarizes these datasets based on the number of sentences and entities. We used these publicly available datasets[2] to make the experiments reproducible. To measure the performance of trained models, we adopt standard entity-level Precision (P), Recall (R), and F1-score (F1) as the evaluation metrics[3].

Following previous works [4], the original training and development sets are merged into a new training set in the experiments, and the test set is only used to evaluate the model. All datasets are trained with the batch size of 32, the maximum sequence length of 256 and a dropout with the probability of 0.1 after the shared encoder. We use the Adam optimizer with the learning rate from $1e^{-}5$ to $5e^{-}5$. The training procedure contains 50 epochs for BC2GM, BC5CDR-chem and 100 epochs for NCBI-disease. Finally, all models are trained on NVIDIA TITAN RTX GPUs.

[2] https://github.com/cambridgeltl/MTL-Bioinformatics-2016

[3] https://github.com/chakki-works/seqeval

Model	Dataset								
	BC2GM			BC5CDR-chem			NCBI-disease		
	P	R	F1	P	R	F1	P	R	F1
LSTM-CRF [3]	81.57	79.48	80.51	87.60	86.25	86.92	86.11	85.49	85.80
BiLM-NER [13]	81.81	81.57	81.69	88.10	90.49	89.28	86.41	88.31	87.34
Att-ChemdNER [14]	-	-	-	93.49	91.68	92.57	-	-	-
BERT [11]	82.16	81.85	82.01	93.13	89.88	91.48	82.98	88.85	85.81
MTM-CW [15]	82.10	79.42	80.74	93.09	89.56	91.29	85.86	86.42	86.14
CollaboNET [16]	80.49	78.99	79.73	94.26	92.38	93.31	85.48	87.27	86.36
MT-BioNER [12]	82.10	84.04	83.01	88.46	90.52	89.48	86.73	89.70	88.19
SciBERT [17]	83.04	83.65	83.34	93.54	93.61	93.58	87.04	90.63	88.24
Dic-Att-BiLSTM-CRF [5]	-	-	-	93.49	91.68	92.57	88.3	89.0	88.6
PubmedBERT [18]	84.17	84.35	84.26	93.30	94.42	93.86	88.57	88.75	88.66
BioBERT [4]	83.57	85.22	84.38	92.98	94.24	93.60	87.83	90.21	89.00
Ours	84.65	85.03	**84.84***	93.70	94.10	**93.90****	87.69	91.25	**89.43****

Table 3. Model performance comparison on the three benchmark datasets, where the best model and F1-scores are bolded. * and ** indicate our model have a significant difference compared with BioBERT with $p \leq 0.05$ and $p \leq 0.01$, respectively.

Rule	Dataset		
	BC2GM	BC5CDR-chem	NCBI-disease
None	84.44	93.69	88.75
Annotation	**84.84**	93.52	**89.43**
Wikipedia	84.37	**93.90**	89.08
Dictionary	84.31	93.51	89.23
Random	84.23	93.35	88.18

Table 4. The performance of our model with different queries, where the best F1-scores are bolded.

Model	Size		
	50%	25%	10%
CS-MTM	84.74	81.00	76.78
BioBERT	88.62	85.79	80.10
Ours	**88.89**	**86.50**	**82.08**

Table 5. The impact of data size on model performance, we conduct experiments on the NCBI-disease dataset.

3.2. Results and Comparisons

We compare our model with previous state-of-the-art models such as LSTM-CRF [3], BiLM-NER [13], Dic-Att-BiLSTM-CRF [5], domain-specific BERTs [4, 11, 17, 18], and multi-task models such as MTM-CW [15], CollaboNET [16], MT-BioNER [12]. The best baseline model is BioBERT. Since the BioBERT does not publish the hyper-parameter details of downstream tasks, we apply the same hyper-parameter and training strategy mentioned in the Settings.

Table 3 shows the overall performance of our model with the best configuration compared with existing approaches on three benchmark datasets. Experiment results show that our multi-task MRC-based model can achieve the best results, with improvements of 0.46%, 0.30%, and 0.43% F1-score over the BioBERT among three datasets, respectively.

Moreover, we study the influence of different prior knowledge on model performance. As shown in Table 4, where None represents we simply setting the $m_i = \{[CLS], p_i, [SEP]\}$ and Random represents randomly select from three prior knowledge during the query construction process. The results show that the model employing the Annotation scheme achieves the highest F1-score except the BC5CDR-chem dataset. Overall, our best model can achieve an average improvement of 0.43% F1-score over the None method, indicating that our method of introducing prior knowledge is effective.

Finally, as shown in Table 5, where CS-MTM is a multi-task model with the cross-sharing structure proposed by [19], and the best F1-score for each resource size is bolded. When the training set of NCBI-disease is reduced and the test set is kept, the missing information in removed sentences makes the results of all models worse. However, for 50%-size, 25%-size, and 10%-size datasets, our model can get 0.25%, 0.71%, and 1.98% F1-score improvements over the BioBERT. The improvement is more obvious with fewer training samples, which demonstrates our multi-task learning model is more robust in the low-resource scenario.

4. CONCLUSION

Current deep learning-based BioNER methods simply use prior knowledge for pre-processing or post-editing, which is sub-optimal. In this paper, we propose a multi-task MRC framework for BioNER, which introduces prior knowledge by the auxiliary task MRC through careful design of query sentences. The experimental results on three datasets demonstrate that our proposed method is effective.

5. REFERENCES

[1] Yiqi Tong, Yidong Chen, and Xiaodong Shi, "A multi-task approach for improving biomedical named entity recognition by incorporating multi-granularity information," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online, Aug. 2021, pp. 4804–4813, Association for Computational Linguistics.

[2] Cong Sun, Zhihao Yang, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang, "Biomedical named entity recognition using BERT in the machine reading comprehension framework," *J. Biomed. Informatics*, vol. 118, pp. 103799, 2021.

[3] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, 2017.

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[5] Kai Xu, Zhenguo Yang, Peipei Kang, Qi Wang, and Wenyin Liu, "Document-level attention-based bilstm-crf incorporating disease dictionary for disease named entity recognition," *Computers in biology and medicine*, vol. 108, pp. 122–132, 2019.

[6] Yiqi Tong, Jiangbin Zheng, Hongkang Zhu, Yidong Chen, and Xiaodong Shi, "A document-level neural machine translation model with dynamic caching guided by theme-rheme information," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 4385–4395.

[7] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li, "A unified mrc framework for named entity recognition," *arXiv preprint arXiv:1910.11476*, 2019.

[8] Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al., "Overview of biocreative ii gene mention recognition," *Genome biology*, vol. 9, no. S2, pp. S2, 2008.

[9] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu, "Biocreative v cdr task corpus: a resource for chemical disease relation extraction," *Database*, vol. 2016, 2016.

[10] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu, "Ncbi disease corpus: a resource for disease name recognition and concept normalization," *Journal of biomedical informatics*, vol. 47, pp. 1–10, 2014.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186, Association for Computational Linguistics.

[12] Muhammad Raza Khan, Morteza Ziyadi, and Mohamed AbdelHady, "Mt-bioner: Multi-task learning for biomedical named entity recognition using deep bidirectional transformers," *arXiv preprint arXiv:2001.08904*, 2020.

[13] Devendra Singh Sachan, Pengtao Xie, Mrinmaya Sachan, and Eric P Xing, "Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition," in *Machine learning for healthcare conference*. PMLR, 2018, pp. 383–402.

[14] Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang, "An attention-based bilstm-crf approach to document-level chemical named entity recognition," *Bioinformatics*, vol. 34, no. 8, pp. 1381–1388, 2018.

[15] Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han, "Cross-type biomedical named entity recognition with deep multi-task learning," *Bioinformatics*, vol. 35, no. 10, pp. 1745–1752, 2019.

[16] Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang, "Collabonet: collaboration of deep neural networks for biomedical named entity recognition," *BMC bioinformatics*, vol. 20, no. 10, pp. 55–65, 2019.

[17] Iz Beltagy, Kyle Lo, and Arman Cohan, "Scibert: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[18] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon, "Domain-specific language model pretraining for biomedical natural language processing," *arXiv preprint arXiv:2007.15779*, 2020.

[19] Xi Wang, Jiagao Lyu, Li Dong, and Ke Xu, "Multitask learning for biomedical named entity recognition with cross-sharing structure," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–13, 2019.