

# Biomedical named entity recognition using BERT in the machine reading comprehension framework

Cong Sun<sup>1</sup>, Zhihao Yang<sup>1,\*</sup>, Lei Wang<sup>2,\*</sup>, Yin Zhang<sup>2</sup>, Hongfei Lin<sup>1</sup>, Jian Wang<sup>1</sup>

<sup>1</sup>*School of Computer Science and Technology, Dalian University of Technology, Dalian, China, 116024*

<sup>2</sup>*Beijing Institute of Health Administration and Medical Information, Beijing, China, 100850*

\*Corresponding author: yangzh@dlut.edu.cn, wangleibihami@gmail.com

**Abstract**—Recognition of biomedical entities from literature is a challenging research focus, which is the foundation for extracting a large amount of biomedical knowledge existing in unstructured texts into structured formats. Using the sequence labeling framework to implement biomedical named entity recognition (BioNER) is currently a conventional method. This method, however, often cannot take full advantage of the semantic information in the dataset, and the performance is not always satisfactory. In this work, instead of treating the BioNER task as a sequence labeling problem, we formulate it as a machine reading comprehension (MRC) problem. This formulation can introduce more prior knowledge utilizing well-designed queries, and no longer need decoding processes such as conditional random fields (CRF). We conduct experiments on six BioNER datasets, and the experimental results demonstrate the effectiveness of our method. Our method achieves state-of-the-art (SOTA) performance on the BC4CHEMD, BC5CDR-Chem, BC5CDR-Disease, NCBI Disease, BC2GM and JNLPBA datasets, with F1-scores of 92.38%, 94.19%, 87.36%, 90.04%, 84.98% and 78.93%, respectively.

**Index Terms**—text mining, named entity recognition, NER, machine reading comprehension, MRC

## I. INTRODUCTION

Biomedical named entity recognition (BioNER) aims to automatically recognize biomedical entities (e.g., chemicals, diseases and proteins) in given texts. Effectively recognizing biomedical entities is the prerequisite for extracting biomedical knowledge deposited in unstructured texts, transforming them into structured formats. Therefore, the BioNER task has important research value.

Traditionally, BioNER methods usually depend on well-designed feature engineering, i.e., the design of features using various natural language processing (NLP) tools and domain knowledge. Typical representatives of such models used in the biomedical domain include DNorm [1], tmChem [2], TaggerOne [3], Lou’s joint model [4], etc. Feature engineering, however, relies heavily on domain-specific knowledge and hand-crafted features. Furthermore, these features are both model- and entity-specific. In recent years, neural networks with automatic feature learning abilities have become prevalent in NER tasks [5], [6]. For the biomedical domain, several neural network methods [7]–[12] have been proposed to recognize biomedical entities. Within these methods, bidirectional long short-term memory (BiLSTM) [13] is usually employed to

learn vector representations of each word/token in a sentence, and then as the input to conditional random fields (CRF) [14]. Very recently, language models (e.g., ELMo [15] and BERT [16]) obtained state-of-the-art (SOTA) performance on many NLP tasks. In the biomedical domain, Lee et al. [17] used BioBERT (namely BERT pre-trained on biomedical corpora) and the softmax function to recognize biomedical entities, and their method achieved SOTA results on several biomedical datasets. Compared with feature engineering methods, neural network methods have the ability to automatically learn features and thus can achieve more competitive performance.

The existing methods usually formalize the BioNER task into a sequence labeling problem, i.e., training a sequence labeling model to assign a label to each token in a given sequence. However, the models mentioned above, i.e., BiLSTM-CRF and BioBERT-Softmax, both cannot effectively learn the semantic information in the sequence labeling framework. For BiLSTM, its representation ability has certain deficiencies, which limits the performance [17], [18]. For BioBERT, it is difficult to effectively use the semantic information learned by the final layer of BioBERT in the sequence labeling framework [19].

Inspired by the current trend of formalizing NLP tasks into machine reading comprehension (MRC) tasks [20]–[24], we use BioBERT in the MRC framework to perform BioNER (referred to BioBERT-MRC). In the MRC framework, each biomedical entity type can be encoded by a language query and identified by answering these queries. Take a sentence “[Meloxicam]<sub>chemical</sub> - induced liver toxicity .” from the BC5CDR-Chem dataset as an example. We can introduce more prior knowledge by designing queries in the MRC framework. Therefore, the original sentence can be formalized as a sentence pair “[Meloxicam]<sub>chemical</sub> - induced liver toxicity . Can you detect chemical entities like sodium or RA or cannabis?”, where ‘sodium’, ‘RA’ and ‘cannabis’ are chemical entities and can be obtained from the BC5CDR-Chem dataset. Compared with the sequence labeling framework, the MRC framework essentially has the advantage of introducing prior knowledge. More importantly, by converting the sequence labeling problem into the MRC problem, we can make full use of hidden representations of deep layers in BERT to learn semantic information. This can overcome the deficiency of existing methods and is more meaningful for BioNER tasks, such as using BioBERT-MRC for transfer learning to further

improve model performance.

We use BioBERT in the MRC framework to perform BioNER tasks and conduct experiments on six datasets, i.e., BC4CHEMD [25], BC5CDR-Chem [26], BC5CDR-Disease [26], NCBI Disease [27], BC2GM [28] and JNLPBA [29] datasets. Our method achieves SOTA performance on all these datasets. Moreover, we also analyze some typical layers of BERT. The experimental results show that BioBERT-MRC can effectively learn semantic information, thereby improving model performance. Finally, we further point out that using BioBERT-MRC for transfer learning is possible to obtain considerable performance improvement. This experimental finding may be more meaningful for the development of BioNER tasks. The code and datasets can be found at <https://github.com/CongSun-dlut/BioBERT-MRC>.

## II. RELATED WORK

### A. Language Model

Word embeddings can use a large amount of unlabeled data to learn the latent syntactic and semantic information of words/tokens, and map these words/tokens into dense low-dimensional vectors. In the past decade, several word-embedding methods have been proposed, among which the representative methods are Word2Vec [30], [31] and GloVe [32]. Word2Vec employs the Skip-Gram model [30] to predict surrounding words according to the current word/token or utilizes the Continuous Bag-Of-Words (CBOW) model [31] to model the current word/token based on the surrounding context. GloVe [32] uses a specific weighted least squares model that trains on global word-word co-occurrence counts, and thus makes efficient use of statistics and considers both the local and global features of the training corpus. However, the word/token trained by these methods is mapped to a certain vector. Therefore, word embeddings trained by these methods can only model context-independent representations.

Recently, language models such as ELMo [15] and BERT [16] have shown a considerable boost to NLP tasks. Unlike traditional word embeddings such as Word2Vec and GloVe, the embedding assigned to the word/token by the language model depends on the context. Therefore, the same word/token can have different representations in different contexts. ELMo [15] leverages the concatenation of independently trained left-to-right and right-to-left LSTM to model the contextual information of the input sequence. BERT [16] employs Transformer [33] to pre-train representations by jointly conditioning on both left and right context in all layers. Because the great success of BERT, it has gradually become a mainstream method that using a large corpus to pre-train BERT and fine-tuning it on the target dataset. Furthermore, the use of BERT for transfer learning has also become a promising research direction [34], [35].

### B. Machine Reading Comprehension

The MRC methods [20]–[24] can extract answer spans from the context through a given query. This task can be formalized as two classification tasks, namely to predict the start and

end positions of the answer spans. In recent years, there has been a trend of converting related NLP tasks into MRC tasks. For example, McCann et al. [21] used the question answering framework to implement ten different NLP tasks uniformly, and all achieved competitive performance. For the NER task, Li et al. [24] employed BERT to recognize entities from texts in the MRC framework (in the general domain). To the best of our knowledge, there is currently no specific research for BioNER in the MRC framework. Our work focus specifically on biomedical entities, which is significantly different from Li’s work [24]. More importantly, we are the first to explore the effect of different BERT layers on BioNER tasks in the MRC framework. In addition, we point out that BERT may achieve better BioNER performance by performing transfer learning in the MRC framework.

## III. METHODOLOGY

### A. Task Definition

Given an input sentence  $X = \{x_1, x_2, \dots, x_N\}$ , where  $x_i$  is the  $i$ -th word/token and  $N$  represents the length of the sentence. The goal of NER is to classify each word/token in  $X$  and assign it to a corresponding label  $y \in Y$ , where  $Y$  is a predefined list of all possible label types (e.g., CHEMICAL, DISEASE and PROTEIN). We formulate the NER task as an MRC task and thus need to convert the labeling-style NER dataset into a set of (Context, Query, Answer) triples. For each label type  $y$ , we first constructed a query  $Q_y = \{q_1, q_2, \dots, q_M\}$  for each sentence, where  $M$  represents the length of the query. Then, we obtained the annotated entities  $x_{start,end}$  according to the annotated labels  $Y$ , where  $x_{start,end}$  is a substring of  $X$  and  $start \leq end$ . For example, the sentence from the BC5CDR-Chem dataset, “Meloxicam - induced liver toxicity .”, its corresponding labels are “B O O O O O”. According to the labels, we can obtain the entities and their spans: ‘Meloxicam’<sub>0,0</sub>. Finally, we constructed the triple  $(X, Q_y, x_{start,end})$ , which is exactly the (Context, Query, Answer) triple that we need.

### B. Construct Queries

Different from the query generation in the general domain, e.g., Li et al. [24] utilized the annotation guideline notes as references to construct queries, we used biomedical entities from the training/development set of the target dataset to construct queries. Table I lists some examples of the queries

TABLE I  
EXAMPLE OF CONSTRUCTED QUERIES.

Entity type	Query
CHEMICAL	Can you detect chemical entities like <i>chemical</i> <sub>1</sub> or <i>chemical</i> <sub>2</sub> or <i>chemical</i> <sub>3</sub> ?
DISEASE	Can you detect disease entities like <i>disease</i> <sub>1</sub> or <i>disease</i> <sub>2</sub> or <i>disease</i> <sub>3</sub> ?
PROTEIN	Can you detect protein entities like <i>protein</i> <sub>1</sub> or <i>protein</i> <sub>2</sub> or <i>protein</i> <sub>3</sub> ?

Notes: *chemical* <sub>$x$</sub> , *disease* <sub>$x$</sub>  and *protein* <sub>$x$</sub>  are biomedical entities, which can be obtained from the training/development set of the target dataset.

we constructed. We conducted experiments on three types of biomedical entities (i.e., CHEMICAL/DRUG, DISEASE and PROTEIN/GENE). The entity variables (such as  $chemical_1$ ,  $chemical_2$  and  $chemical_3$ ) represent the annotated ones selected randomly from the training/development set of the target dataset. In this way, the queries can introduce more sufficient prior knowledge to the model, thereby enhancing the model's ability to recognize biomedical entities.

### C. Model Details

We exploited BERT [16] as our model backbone. In this study, we need to recognize entities in the biomedical domain, so we employed BioBERT [17] as the model weight. Figure 1 shows the BioNER task implemented in the MRC framework using BERT. Firstly, the query  $Q_y$  is concatenated with the context  $X$ , forming the combined sequence  $\{[CLS], x_1, x_2, \dots, x_N, [SEP], q_1, q_2, \dots, q_M, [SEP]\}$ , where '[CLS]' and '[SEP]' are special tokens. Then, the combined sequence is fed into BERT, which can be defined by the following formulas:

$$h_i^0 = W_e t_i + W_b \quad (1)$$

$$h_i^l = Trm(h_i^{l-1}) \quad (2)$$

$$H = [h_1^L; h_2^L; \dots; h_N^L] \in R^{N \times d} \quad (3)$$

where  $t_i$  is the embedding of the  $i$ -th token,  $W_e$ ,  $W_b$  are parameters,  $L$  means the total number of layers for BERT,  $l$  ( $1 \leq l \leq L$ ) is the number for the current layer, and  $Trm$  denotes the Transformer [33] block, including multi-head attention layers, fully connected layers and normalization layers. Furthermore,  $H$  is the output of BERT and  $N$  is the length of the context. We simply dropped query representations because they are not the target of the model prediction.

There are usually two strategies for choosing the span in the MRC framework. The first is to employ two  $n$ -class classifiers to respectively predict the start and end indexes, where  $n$  is the length of the context. Because the function is calculated on all tokens in the entire context, this strategy has the deficiency that one input sequence can only output one span. The other is to construct two binary classifiers. One is used to predict whether the token is a start index, and the other is employed to predict whether the token is an end index. This strategy allows multiple start and end indexes to be output for a given sequence, and therefore it has the potentials to identify all target entities based on  $Q_y$ .

In this study, we adopted the second strategy. Given the representation matrix  $H$  output by BERT, the model first predicts the probability of each token being a start index. The formula is as follows:

$$L_{start} = linear(HW_{start}) \in R^{N \times 2} \quad (4)$$

where  $linear$  is a fully connected layer and  $W_{start}$  is the weight to learn. Each row of  $L_{start}$  denotes a hidden representation of each index, used to determine the starting position of a target entity for a given query.

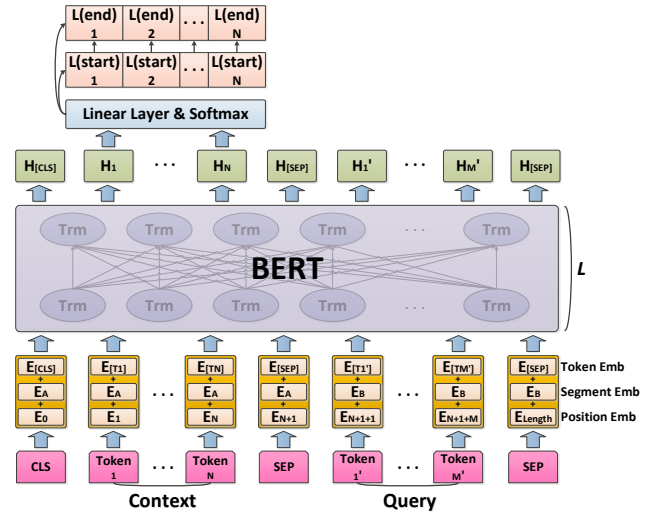


Fig. 1. Using BERT to perform BioNER in the MRC framework.

The end index prediction process is similar to the start index prediction process. To better predict the end index, we designed two strategies:

$$L_{end} = linear(HW_{end}) \in R^{N \times 2} \quad (5)$$

or

$$L_{end} = linear(HW_{end}; softmax(L_{start})) \in R^{N \times 2} \quad (6)$$

where  $linear$  is a fully connected layer,  $W_{end}$  is the weight to learn, and ';' means the concatenation operation.

Then, by applying the argmax function to each row of  $L_{start}$  and  $L_{end}$ , we can obtain the predicted indexes that might be the start or end positions, i.e.,  $I_{start}$  and  $I_{end}$ :

$$I_{start} = \{i \mid argmax(L_{start}^i) = 1, i = 1, 2, \dots, N\} \quad (7)$$

$$I_{end} = \{j \mid argmax(L_{end}^j) = 1, j = 1, 2, \dots, N\} \quad (8)$$

where the superscript  $i$  and  $j$  denote the  $i$ -th and  $j$ -th rows of a matrix, respectively.

In context  $X$ , there could be multiple entities of the same entity type. This means that multiple start/end indexes can be predicted from one input sequence at one time. Because the datasets we conducted experiments on are all flat NER datasets, we further used the nearest match principle to match the start and end indexes to obtain final answers. Specifically, for the case where one end/start index corresponds to multiple start/end indexes, only the nearest start/end index is matched.

### D. Training and Test

At the training time, context  $X$  is paired with two label sequences  $Y_{start}$  and  $Y_{end}$  of length  $N$  representing the ground-truth label of each token  $x_i$  being the start/end index of the biomedical entity. The loss function is defined as follows:

$$Loss_{start} = CE(L_{start}, Y_{start}) \quad (9)$$

$$Loss_{end} = CE(L_{end}, Y_{end}) \quad (10)$$

$$Loss = (1 - \lambda)Loss_{start} + \lambda Loss_{end} \quad (11)$$

where  $CE$  denotes the cross-entropy loss function and  $\lambda \in [0, 1]$  is a hyper-parameter used to control the overall training objective. In this study, we set  $\lambda = 0.5$  according to the performance of the model on the development set.

At the test time, the start and end indexes are first separately selected based on  $I_{start}$  and  $I_{end}$ . Then, the nearest match principle is used to match the start/end indexes to obtain the final answers.

#### IV. EXPERIMENTS

##### A. Datasets and Experimental Settings

Our method is evaluated on BC4CHEMD [25], BC5CDR [26], NCBI-Disease [27], BC2GM [28] and JNLPBA [29] datasets, all of which are pre-processed and provided by Lee et al. [17]. Among these datasets, BC5CDR has two sub-datasets, BC5CDR-Chem and BC5CDR-Disease, which are used to evaluate chemical and disease entities, respectively. Because most of the existing methods were evaluated on BC5CDR-Chem and BC5CDR-Disease respectively, we did the same. Table II lists the statistics of these datasets.

TABLE II  
STATISTICS OF BIONER DATASETS.

Dataset	Entity type	No. annotations	No. sentences
BC4CHEMD	Chemical/Drug	79,842	89,679
BC5CDR-Chem	Chemical/Drug	15,411	14,228
BC5CDR-Disease	Disease	12,694	14,228
NCBI-Disease	Disease	6,881	7,639
BC2GM	Protein/Gene	20,703	20,510
JNLPBA	Protein/Gene	35,460	22,562

Notes: Statistics are from Habibi et al. [7].

In the experiments, the original training and development set were merged into a new training set. Then 10% of the new training set was sampled as the validation set to tune hyper-parameters. The test set was only used to evaluate the model. Most existing works [3], [9]–[12], [17] pre-processed data in this way, and we also followed this way. We used BioBERT<sub>BASE</sub> v1.1 (+PubMed) [17] as the BERT weight. To facilitate comparison with the baseline method, most of our hyper-parameters are similar to those of Lee et al. [17], with only a few hyper-parameters (e.g., training epochs) different. The performance is measured with the F1-score ( $F1$ ), which attributes equal importance to precision ( $P$ ) and recall ( $R$ ). The formula is:  $F1 = 2PR/(P+R)$ . Moreover, all significance tests in this work are based on T-TEST.

##### B. Impact of Different MRC Strategies on Model Performance

In this work, we compared the impact of different MRC strategies on model performance. First, we explored the impact of different span strategies. As described in Section III.C, two

TABLE III  
PERFORMANCE COMPARISON FOR DIFFERENT END INDEX STRATEGIES.

Dataset	Model	$P(\%)$	$R(\%)$	$F1(\%)$
BC4CHEMD	BioBERT-MRC <sup>a</sup>	93.15	90.89	92.01
	BioBERT-MRC <sup>b</sup>	92.83	91.94	<b>92.38</b>
BC5CDR-Chem	BioBERT-MRC <sup>a</sup>	94.20	93.87	94.04
	BioBERT-MRC <sup>b</sup>	94.37	94.00	<b>94.19</b>
BC5CDR-Disease	BioBERT-MRC <sup>a</sup>	88.20	87.03	<b>87.61</b>
	BioBERT-MRC <sup>b</sup>	87.42	87.30	87.36
NCBI-Disease	BioBERT-MRC <sup>a</sup>	88.62	91.67	<b>90.12</b>
	BioBERT-MRC <sup>b</sup>	89.67	90.42	90.04
BC2GM	BioBERT-MRC <sup>a</sup>	86.10	83.65	84.86
	BioBERT-MRC <sup>b</sup>	86.05	83.94	<b>84.98</b>
JNLPBA	BioBERT-MRC <sup>a</sup>	75.10	82.40	78.58
	BioBERT-MRC <sup>b</sup>	75.97	82.13	<b>78.93</b>
Average	BioBERT-MRC <sup>a</sup>	87.56	88.25	87.87
	BioBERT-MRC <sup>b</sup>	87.72	88.29	<b>87.98</b>

Notes: BioBERT-MRC<sup>a</sup> denotes the end index prediction is designed on equation 5, and BioBERT-MRC<sup>b</sup> presents the end index prediction is designed on equation 6. The best scores are shown in bold.

TABLE IV  
PERFORMANCE COMPARISON FOR DIFFERENT TYPES OF QUERIES.

Dataset	Model	$P(\%)$	$R(\%)$	$F1(\%)$
BC4CHEMD	BioBERT-MRC <sup>†</sup>	93.36	91.15	92.24
	BioBERT-MRC <sup>‡</sup>	93.25	91.08	92.16
	BioBERT-MRC	92.83	91.94	<b>92.38</b>
BC5CDR-Chem	BioBERT-MRC <sup>†</sup>	93.75	93.80	93.77
	BioBERT-MRC <sup>‡</sup>	93.45	94.02	93.73
	BioBERT-MRC	94.37	94.00	<b>94.19</b>
BC5CDR-Disease	BioBERT-MRC <sup>†</sup>	87.21	86.64	86.93
	BioBERT-MRC <sup>‡</sup>	88.06	86.53	87.29
	BioBERT-MRC	87.42	87.30	<b>87.36</b>
NCBI-Disease	BioBERT-MRC <sup>†</sup>	90.35	88.75	89.54
	BioBERT-MRC <sup>‡</sup>	90.31	90.31	<b>90.31</b>
	BioBERT-MRC	89.67	90.42	90.04
BC2GM	BioBERT-MRC <sup>†</sup>	86.15	83.62	84.87
	BioBERT-MRC <sup>‡</sup>	86.26	83.75	<b>84.98</b>
	BioBERT-MRC	86.05	83.94	<b>84.98</b>
JNLPBA	BioBERT-MRC <sup>†</sup>	74.60	82.79	78.48
	BioBERT-MRC <sup>‡</sup>	74.33	83.29	78.56
	BioBERT-MRC	75.97	82.13	<b>78.93</b>

Notes: BioBERT-MRC<sup>†</sup> denotes the query is the simple query, and BioBERT-MRC<sup>‡</sup> represents the query using the none query. The best scores are shown in bold.

end index strategies are designed for BioBERT-MRC. The performance comparison is shown in Table III. BioBERT-MRC<sup>a</sup> denotes the end index prediction is designed on equation 5, and BioBERT-MRC<sup>b</sup> presents the end index prediction is designed on equation 6. It can be seen that the F1-scores of these two strategies are both competitive. This indicates the stability of BioBERT-MRC. On average, compared with BioBERT-MRC<sup>a</sup>, BioBERT-MRC<sup>b</sup> has a higher precision and thus obtains a higher F1-score. This may be due to the introduction of the start index information, leading to the model being able to learn more span information. Because the higher average F1-score of BioBERT-MRC<sup>b</sup>, in the experiments, we adopted BioBERT-MRC<sup>b</sup> as the BioBERT-MRC model.

Then, we explored the impact of constructed queries. Intuitively, the more information a query contains, the better its effect should be. Therefore, we constructed the queries as described in Section III.B. Moreover, we further designed two

TABLE V  
PERFORMANCE COMPARISON FOR DIFFERENT MODELS.

Dataset	Model	P(%)	R(%)	F1(%)
BC4CHEMD	BioBERT-Softmax [17]	92.80	91.92	92.36
	BioBERT-Softmax	92.66	91.69	92.17
	BioBERT-CRF	92.67	91.32	91.99
	BioBERT-BiLSTM-CRF	92.51	91.12	91.81
	BioBERT-MRC	92.83	91.94	<b>92.38</b>
BC5CDR-Chem	BioBERT-Softmax [17]	93.68	93.26	93.47
	BioBERT-Softmax	92.64	94.39	93.51
	BioBERT-CRF	93.30	93.82	93.56
	BioBERT-BiLSTM-CRF	93.57	93.48	93.53
	BioBERT-MRC	94.37	94.00	<b>94.19</b>
BC5CDR-Disease	BioBERT-Softmax [17]	86.47	87.84	87.15
	BioBERT-Softmax	85.04	88.02	86.50
	BioBERT-CRF	85.97	87.43	86.70
	BioBERT-BiLSTM-CRF	86.07	87.27	86.67
	BioBERT-MRC	87.42	87.30	<b>87.36</b>
NCBI-Disease	BioBERT-Softmax [17]	88.22	91.25	89.71
	BioBERT-Softmax	86.97	91.04	88.96
	BioBERT-CRF	86.66	90.00	88.30
	BioBERT-BiLSTM-CRF	88.68	88.96	88.82
	BioBERT-MRC	89.67	90.42	<b>90.04</b>
BC2GM	BioBERT-Softmax [17]	84.32	85.12	84.72
	BioBERT-Softmax	84.44	85.04	84.74
	BioBERT-CRF	84.80	84.60	84.70
	BioBERT-BiLSTM-CRF	83.61	85.57	84.58
	BioBERT-MRC	86.05	83.94	<b>84.98</b>
JNLPBA	BioBERT-Softmax [17]	72.24	83.56	77.49
	BioBERT-Softmax	72.80	84.30	78.13
	BioBERT-CRF	73.06	83.93	78.12
	BioBERT-BiLSTM-CRF	72.65	83.24	77.58
	BioBERT-MRC	75.97	82.13	<b>78.93</b>

Notes: The best scores are shown in bold.

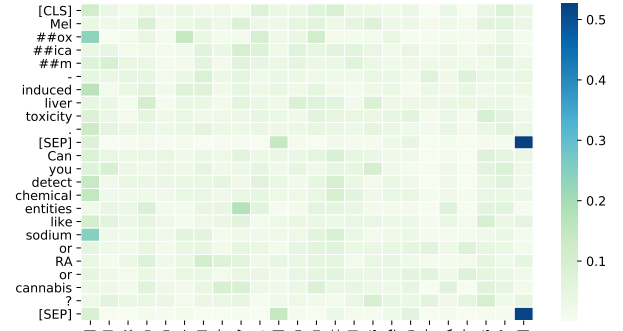
types of queries to explore the impact of different queries:

- **Simple query.** Simple language query, i.e., “Can you detect *X-type* entities?”. For example, for the CHEMICAL entity, the query is “Can you detect chemical entities?”.
- **None query.** The query is designed as “none”.

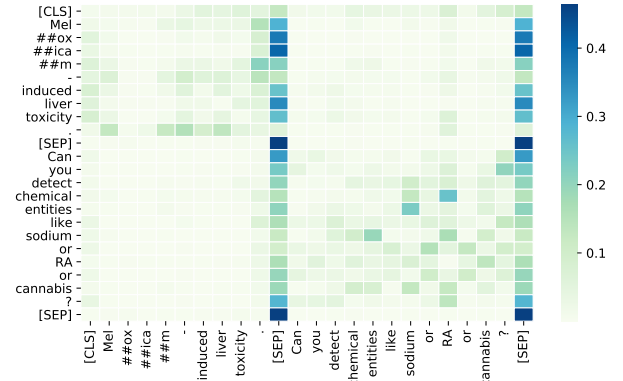
Table IV illustrates the performance comparison for different types of queries. BioBERT-MRC<sup>†</sup> and BioBERT-MRC<sup>‡</sup> denote the situation where the query we constructed is a simple query and none query, respectively. It can be seen that the performance of BioBERT-MRC is superior to these two models, which shows that using high-quality entities from the training/development set as queries can improve model performance.

### C. MRC vs Sequence Labeling

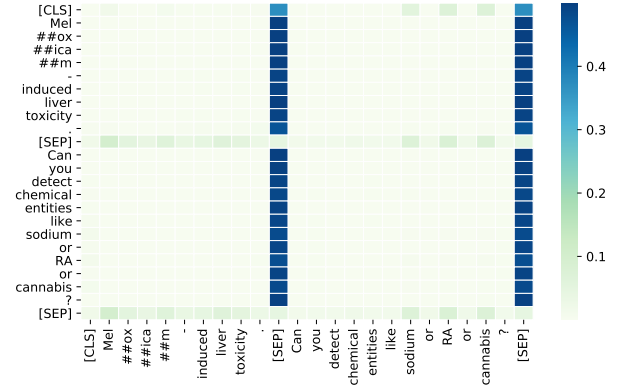
In this work, we explored the effect of BERT in the MRC framework and the sequence labeling framework in detail. First, we compared the performance of BERT on BioNER in the MRC framework and the sequence labeling framework. Table V shows the performance comparison for different models. BioBERT-Softmax, BioBERT-CRF and BioBERT-BiLSTM-CRF are commonly used methods in the sequence labeling framework. It can be seen that the performance of these three methods is similar. On the BC4CHEMD dataset, the performance of BioBERT-Softmax is superior to BioBERT-CRF and BioBERT-BiLSTM-CRF. The reason may be that the examples in the dataset are relatively long, while CRF



(a) Attention heat map of the first layer.



(b) Attention heat map of the sixth layer.



(c) Attention heat map of the final layer.

Fig. 2. Attention heat map of BERT on the BC5CDR-Chem dataset.

and LSTM seem not good at learning long-distance semantic information. However, the performance of these three methods on all the six datasets is inferior to BioBERT-MRC. BioBERT-MRC can greatly improve the precision while maintaining the recall, thereby improving model performance. The reason may be that BioBERT-MRC can learn high-level semantic information, thereby enhancing the model’s ability to identify target entities. These experimental results show that BERT’s ability to recognize biomedical entities in the MRC framework

TABLE VI  
PERFORMANCE COMPARISON WITH OTHER EXISTING METHODS.

Method	BC4CHEMD			BC5CDR-Chem			BC5CDR-Disease			NCBI-Disease			BC2GM			JNLPBA		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
tmChem [2]	89.09	85.75	87.39**	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
TaggerOne [3]	—	—	—	94.20	88.80	91.40**	85.20	80.20	82.60**	85.10	80.80	82.90**	—	—	—	—	—	—
Lou et al. [4]	—	—	—	—	—	—	89.61	83.09	86.23**	90.72	74.89	82.05**	—	—	—	—	—	—
D3NER [8]	—	—	—	93.73	92.56	93.14*	83.98	85.40	84.68**	85.03	83.80	84.41**	—	—	—	—	—	—
Luo et al. [9]	92.29	90.01	91.14**	93.49	91.68	92.57**	—	—	—	—	—	—	—	—	—	—	—	—
Sachan et al. [10]	—	—	—	—	—	—	—	—	—	86.41	88.31	87.34**	81.81	81.57	81.69**	71.39	79.06	75.03**
Wang et al. [11]	91.30	87.53	89.37**	—	—	—	—	—	—	85.86	86.42	86.14**	82.10	79.42	80.74**	70.91	76.34	73.52**
Collabonet [12]	90.78	87.01	88.85**	94.26	92.38	93.31*	85.61	82.61	84.08**	85.48	87.27	86.36**	80.49	78.99	79.73**	74.43	83.22	78.58*
BioBERT-Softmax [17]	92.80	91.92	<u>92.36*</u>	93.68	93.26	<u>93.47*</u>	86.47	87.84	<u>87.15*</u>	88.22	91.25	<u>89.71*</u>	84.32	85.12	<u>84.72</u>	72.24	83.56	<u>77.49*</u>
BioBERT-MRC	92.83	91.94	<b>92.38</b>	94.37	94.00	<b>94.19</b>	87.42	87.30	<b>87.36</b>	89.67	90.42	<b>90.04</b>	86.05	83.94	<b>84.98</b>	75.96	82.13	<b>78.93</b>

Notes: The first part (i.e., row 13) is feature engineering-based methods, and the second part (i.e., row 410) is neural network-based methods. Both BioBERT-Softmax and BioBERT-MRC are based on BioBERT v1.1 (+PubMed) [17]. \* and \*\* denote the method is significantly worse than BioBERT-MRC ( $p < 0.05$  and  $p < 0.01$ , respectively). The best scores are shown in bold, and the second best scores are underlined.

is superior to its ability to recognize biomedical entities in the sequence labeling framework.

Then, we explored what the model has learned from each layer of BERT, and why BERT can achieve better BioNER performance in the MRC framework. Figure 2 shows the attention heat map of BERT in the first layer, the middle layer (i.e., the sixth layer), and the final layer. The first layer mainly encodes the features of words/tokens, the middle layer tends to extract syntactic information (e.g., phrases and segments), and the final layer learns semantic information of the whole sequence. Unlike sequence labeling problems that focus on the target word/token and its surrounding words/tokens, MRC problems often require more attention to the semantic information of the entire sequence. Therefore, compared with using BERT in the sequence labeling framework, performing it in the MRC framework has more advantages in learning the syntactic and semantic information of the dataset. This may be the reason that BERT can obtain better BioNER performance in the MRC framework.

#### D. Performance Comparison with Other Methods

Table VI shows the experimental results on BioNER datasets. The first three methods all rely on feature engineering. TmChem [2] is a method that utilizes chemical knowledge to design features to recognize chemical entities. TaggerOne [3] and Lou’s model [4] both perform BioNER in a joint manner with named entity normalization (NEN), in which the feedback from NEN can be used to reduce NER errors. The difference is the former uses a semi-Markov linear classifier for BioNER, while the latter is based on a finite state machine. However, these methods are often model- and entity-specific and the performance is also unsatisfactory. In addition to the first three methods, the others are all neural network-based methods. Among these methods, the first five mainly use the LSTM-CRF model, while the latter two leverage the BioBERT model. D3NER [8] introduces various linguistic information to improve performance. Luo et al. [9] utilized document-level global information obtained by the attention mechanism to enforce tagging consistency across multiple instances of the same token in the document. Sachan et al. [10] trained a

bidirectional language model on unlabeled data and transfer the weights to pre-train an BioNER model. Wang et al. [11] proposed a multi-task learning framework for BioNER to collectively use the training data of different types of entities. CollaboNet [12] can also be regarded as a multi-task learning framework, which improves performance by combining multiple BioNER models. Generally, the performance of these LSTM-CRF-based models is superior to feature engineering-based models. Most recently, Lee et al. [17] utilized BioBERT to perform BioNER in the sequence labeling framework, which raises the performance of BioNER to a new level.

We formulated the BioNER task as an MRC problem and employed BERT to achieve BioNER in the MRC framework. It is encouraging to see that BioBERT-MRC obtains SOTA performance on all the six datasets. Compared with feature engineering-based and LSTM-CRF-based methods, our method usually obtains both the highest precision and recall performance. This shows that our method is far superior to those methods. Compared with BioBERT-Softmax, BioBERT-MRC can greatly improve the precision, thereby increasing the F1-score. This experimental result indicates that our method is also superior to BioBERT-Softmax. In conclusion, these experimental results show that BioBERT-MRC is superior in performance as compared with other existing methods.

#### E. Case Study

To further explore the difference between BioBERT-MRC and BioBERT-Softmax, we conducted the case study. The results of case study are shown in Table VII. For the first example, this is a case where BioBERT-Softmax recognized false negatives (FNs) while BioBERT-MRC corrected them as true positives (TPs). This example shows that BioBERT-MRC has certain advantages over BioBERT-Softmax in terms of learning syntactic information (e.g., phrases and segments). The second example is a consistency problem. It can be seen that BioBERT-Softmax only recognized one ‘cocaine’ in the whole sequence, while BioBERT-MRC corrected the error of BioBERT-Softmax. This example shows that BioBERT-MRC can alleviate the problem of label inconsistency by learning the semantic information of the entire sequence. The third and last



TABLE VII  
REPRESENTATIVE RESULTS OF CASE STUDY.

No.	Dataset	Gold standard	Model	Result
1	BC4CHEMD	lipoic acid (CHEMICAL)	BioBERT-Softmax	... and the natural antioxidant lipoic acid , without influencing the level of free ...
			BioBERT-MRC	... and the natural antioxidant <b>lipoic acid</b> , without influencing the level of free ...
2	BC5CDR-Chem	cocaine (CHEMICAL) DSE (OTHER)	BioBERT-Softmax	... cocaine use , we conducted a pilot study to assess the safety of <b>DSE</b> in emergency department patients with <b>cocaine</b> - associated chest pain .
			BioBERT-MRC	... <b>cocaine</b> use , we conducted a pilot study to assess the safety of DSE in emergency department patients with <b>cocaine</b> - associated chest pain .
3	BC5CDR-Disease	drug - induced disease (OTHER)	BioBERT-Softmax	Differential diagnosis for <b>drug - induced disease</b> is invaluable even for patients ...
			BioBERT-MRC	Differential diagnosis for drug - induced disease is invaluable even for patients ...
4	NCBI Disease	dominantly inherited neurodegeneration (DISEASE)	BioBERT-Softmax	This mutation may be valuable for developing models of dominantly inherited <b>neurodegeneration</b> ...
			BioBERT-MRC	This mutation may be valuable for developing models of <b>dominantly inherited neurodegeneration</b> ...
5	BC2GM	RXR ligand binding domain (PROTEIN)	BioBERT-Softmax	... influences the <b>RXR</b> ligand binding domain such that it is resistant to the binding of 9 - cis RA ...
			BioBERT-MRC	... influences the <b>RXR ligand binding domain</b> such that it is resistant to the binding of 9 - cis RA ...
6	JNLPA	peroxisome proliferator - activated receptor gamma (OTHER)	BioBERT-Softmax	Oxidized alkyl phospholipids are specific , high affinity <b>peroxisome proliferator - activated receptor gamma</b> ligands and agonists .
			BioBERT-MRC	Oxidized alkyl phospholipids are specific , high affinity peroxisome proliferator - activated receptor gamma ligands and agonists .

Notes: Green represents chemical entities, blue represents disease entities, and yellow represents protein entities.

examples are a case where BioBERT-Softmax recognized false positives (FPs) while BioBERT-MRC corrected them as true negatives (TNs). The fourth and fifth are segmentation problem examples. Compared with BioBERT-Softmax, BioBERT-MRC can better distinguish the boundary information of entities. These four examples all demonstrate the effectiveness of BioBERT-MRC in the syntactic learning.

Through the case study, we can infer that compared with BioBERT-Softmax, BioBERT-MRC has better performance in syntactic and semantic learning. Specifically, BioBERT-MRC can correct some FNs and FPs, accurately recognize entity boundaries, and alleviate the label inconsistency problem.

#### F. BioBERT-MRC for Transfer Learning

Transfer learning is an important method in NLP tasks. Because similar datasets can be used for model training, transfer learning may learn richer semantic knowledge. Among the six datasets, BC4CHEMD and BC5CDR-Chem are both chemical datasets, and BC4CHEMD has more annotated entities. Therefore, we tried to use transfer learning to further improve the performance of the model on the BC5CDR-Chem dataset. Compared with directly using BERT to perform BioNER on the BC5CDR-Chem dataset, using the weights obtained from BC4CHEMD to perform transfer learning may obtain better performance. As shown in Table V, BioBERT-Softmax and BioBERT-MRC obtain F1-scores of 92.17% and 92.38% on the BC4CHEMD dataset, respectively. We used these two weights as the weights of BioBERT-Softmax and BioBERT-MRC, and then performed transfer learning on the BC5CDR-Chem dataset. The performance comparison is illustrated in Table VIII. Directly using the BioBERT v1.1 (+PubMed) weight to perform NER on the BC5CDR-Chem

TABLE VIII  
PERFORMANCE COMPARISON FOR TRANSFER LEARNING

Method	Weight	P(%)	R(%)	F1(%)
BioBERT-Softmax	BioBERTv1.1(+PubMed)	92.64	94.39	93.51*
BioBERT-Softmax	BC4CHEMD <sup>a</sup>	93.85	94.15	94.00*
BioBERT-MRC	BioBERTv1.1(+PubMed)	94.37	94.00	94.19*
BioBERT-MRC	BC4CHEMD <sup>b</sup>	94.48	94.39	<b>94.44</b>

Notes: \* denotes the method is significantly worse than BioBERT-MRC ( $p < 0.05$ ). BioBERT v1.1 (+PubMed) can be obtained from Lee et al. [17]. BC4CHEMD<sup>a</sup> and BC4CHEMD<sup>b</sup> are the weights from BioBERT-Softmax and BioBERT-MRC on the BC4CHEMD dataset, respectively. The best scores are shown in bold.

dataset, BioBERT-Softmax and BioBERT-MRC obtain 93.51% and 94.19% respectively. On the contrary, using transfer learning, i.e., fine-tuning the weights obtained on the BC4CHEMD dataset to recognize chemical entities on the BC5CDR-Chem dataset, can achieve more excellent performance. By this way, BioBERT-Softmax and BioBERT-MRC obtain F1-scores of 94.00% (+0.49% higher than 93.51%) and 94.44% (+0.25% higher than 94.19%), respectively. These experimental results show that compared with fine-tuning directly on the target dataset, transfer learning can learn the knowledge of the source dataset, thereby improving the performance of the model on the target dataset. More importantly, the performance of BioBERT-MRC is superior to BioBERT-Softmax in terms of transfer learning. In the case of transfer learning, the F1-score of BioBERT-Softmax is 94.00%, while the F1-score of BioBERT-MRC is 94.44% (+0.44% higher than 94.00%). The reason may be that BioBERT-MRC learns more semantic information, which may be useful for transfer learning. This experiment result indicates that BioBERT-MRC is more suitable for transfer learning than BioBERT-Softmax.

## V. CONCLUSION

In this work, we use BERT to perform BioNER in the MRC framework. Compared with using BERT in the sequence labeling framework, performing it in the MRC framework can introduce more prior knowledge by well-designed queries, and no longer need decoding processes. The experimental results demonstrate the effectiveness of our method. We also explore the reasons for the excellent performance of BERT in the MRC framework. Through experimental analysis, we infer that BERT in the MRC framework has more advantages in learning the syntactic and semantic information of the dataset. Moreover, we further indicate the effectiveness of using BERT in the framework for transfer learning, which may be valuable for the development of BioNER tasks.

## REFERENCES

- [1] R. Leaman, R. Islamaj Doğan, and Z. Lu, “Dnorm: disease name normalization with pairwise learning to rank,” *Bioinformatics*, vol. 29, no. 22, pp. 2909–2917, 2013.
- [2] R. Leaman, C. Wei, and Z. Lu, “tmChem: a high performance approach for chemical named entity recognition and normalization,” *Journal of Cheminformatics*, vol. 7, no. 1, pp. 1–10, 2015.
- [3] R. Leaman and Z. Lu, “TaggerOne: joint named entity recognition and normalization with semi-Markov Models,” *Bioinformatics*, vol. 32, no. 18, pp. 2839–2846, 2016.
- [4] Y. Lou, Y. Zhang, T. Qian, F. Li, S. Xiong, and D. Ji, “A transition-based joint model for disease named entity recognition and normalization,” *Bioinformatics*, vol. 33, no. 15, pp. 2363–2371, 2017.
- [5] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural Architectures for Named Entity Recognition,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, 2016, pp. 260–270.
- [6] A. Jagannatha and H. Yu, “Structured prediction models for RNN based sequence labeling in clinical text,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 856–865.
- [7] M. Habibi, L. Weber, M. L. Neves, D. L. Wiegandt, and U. Leser, “Deep learning with word embeddings improves biomedical named entity recognition,” *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, 2017.
- [8] T. H. Dang, H.-Q. Le, T. M. Nguyen, and S. T. Vu, “D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information,” *Bioinformatics*, vol. 34, no. 20, pp. 3539–3546, 2018.
- [9] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, “An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition,” *Bioinformatics*, vol. 34, no. 8, pp. 1381–1388, 2018.
- [10] D. S. Sachan, P. Xie, M. Sachan, and E. P. Xing, “Effective Use of Bidirectional Language Modeling for Transfer Learning in Biomedical Named Entity Recognition,” in *Proceedings of Machine Learning Research*, 2018, pp. 383–402.
- [11] X. Wang, Y. Zhang, X. Ren, Y. Zhang, M. Zitnik, J. Shang, C. Langlotz, and J. Han, “Cross-type biomedical named entity recognition with deep multi-task learning,” *Bioinformatics*, vol. 35, no. 10, pp. 1745–1752, 2018.
- [12] W. Yoon, C. H. So, J. Lee, and J. Kang, “CollaboNet: collaboration of deep neural networks for biomedical named entity recognition,” *BMC Bioinformatics*, vol. 20, no. 10, pp. 55–65, 2019.
- [13] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [14] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, p. 282–289.
- [15] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep Contextualized Word Representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018, pp. 2227–2237.
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. 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*, 2019, pp. 4171–4186.
- [17] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2019.
- [18] P. Li, T. Fu, and W. Ma, “Why Attention? Analyze BiLSTM Deficiency and Its Remedies in the Case of NER,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
- [19] M. Kaneko and M. Komachi, “Multi-Head Multi-Layer Attention to Deep Language Representations for Grammatical Error Detection,” in *20th International Conference on Computational Linguistics and Intelligent Text Processing*, 2019.
- [20] O. Levy, M. Seo, E. Choi, and L. Zettlemoyer, “Zero-Shot Relation Extraction via Reading Comprehension,” in *Proceedings of the 21st Conference on Computational Natural Language Learning*, 2017, pp. 333–342.
- [21] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, “The Natural Language Decathlon: Multitask Learning as Question Answering,” *arXiv: Computation and Language*, 2018.
- [22] X. Li, F. Yin, Z. Sun, X. Li, A. Yuan, D. Chai, M. Zhou, and J. Li, “Entity-Relation Extraction as Multi-Turn Question Answering,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1340–1350.
- [23] Y. Shen, P.-S. Huang, J. Gao, and W. Chen, “ReasonNet: Learning to Stop Reading in Machine Comprehension,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1047–1055.
- [24] X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, and J. Li, “A Unified MRC Framework for Named Entity Recognition,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5849–5859.
- [25] M. Krallinger, O. Rabal, F. Leitner, M. Vazquez, D. Salgado, Z. Lu, P. Leaman, Y. Lu, D. Ji, D. M. Lowe et al., “The ChEMDNER corpus of chemicals and drugs and its annotation principles,” *Journal of Cheminformatics*, vol. 7, no. 1, pp. 1–17, 2015.
- [26] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegiers, and Z. Lu, “BioCreative V CDR task corpus: a resource for chemical disease relation extraction,” *Database*, vol. 2016, 2016.
- [27] R. I. Dogan, R. Leaman, and Z. Lu, “NCBI disease corpus: A resource for disease name recognition and concept normalization,” *Journal of Biomedical Informatics*, vol. 47, pp. 1–10, 2014.
- [28] L. H. Smith, L. Tanabe, R. J. N. Ando, C. Kuo, I. Chung, C. Hsu, Y. Lin, R. Klinger, C. M. Friedrich, K. Ganchev et al., “Overview of BioCreative II gene mention recognition,” *Genome Biology*, vol. 9, no. 2, pp. 1–19, 2008.
- [29] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, “Introduction to the Bio-Entity Recognition Task at JNLPBA,” in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, 2004, pp. 70–75.
- [30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
- [31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *International Conference on Learning Representations*, 2013.
- [32] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.
- [33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
- [34] C. Sun and Z. Yang, “Transfer Learning in Biomedical Named Entity Recognition: An Evaluation of BERT in the PharmaCoNER task,” in *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, 2019, pp. 100–104.
- [35] Y. Peng, S. Yan, and Z. Lu, “Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets,” in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019, pp. 58–65.