Fine-tuning BERT for Joint Entity and Relation Extraction in Chinese Medical Text

Kui Xue[1], Yangming Zhou[1,*], Zhiyuan Ma[1], Tong Ruan[1], Huanhuan Zhang[1,*] and Ping He[2]
[1]School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China
[2]Shanghai Hospital Development Center, Shanghai 200041, China
*Corresponding authors
Emails: {ymzhou,hzhang}@ecust.edu.cn

***Abstract*—Entity and relation extraction is the necessary step in structuring medical text. However, the feature extraction ability of the bidirectional long short term memory network in the existing model does not achieve the best effect. At the same time, the language model has achieved excellent results in more and more natural language processing tasks. In this paper, we present a focused attention model for the joint entity and relation extraction task. Our model integrates well-known BERT language model into joint learning through dynamic range attention mechanism, thus improving the feature representation ability of shared parameter layer. Experimental results on coronary angiography texts collected from Shuguang Hospital show that the F_1-scores of named entity recognition and relation classification tasks reach 96.89% and 88.51%, which outperform state-of-the-art methods by 1.65% and 1.22%, respectively.**

***Index Terms*—Named entity recognition, Relation classification, Joint model, BERT language model, Electronic health records.**

I. Introduction

With the widespread of electronic health records (EHRs) in recent years, a large number of EHRs can be integrated and shared in different medical environments, which further support the clinical decision making and government health policy formulation [1]. However, most of the information in current medical records is stored in natural language texts, which makes data mining algorithms unable to process these data directly. To extract relational entity triples from the text, researchers generally use entity and relation extraction algorithm, and rely on the central word to convert the triples into key-value pairs, which can be processed by conventional data mining algorithms directly.

To solve the task of entity and relation extraction, researchers usually follow pipeline processing and split the task into two sub-tasks, namely named entity recognition (NER) and relation classification (RC), respectively. However, this pipeline method usually fails to capture joint features between entity and relationship types. For example, for a valid relation "存在情况(presence)" in Fig. 1, the types of its two relational entities must be "疾病(disease)", "症状(symptom)" or "存在词(presence word)". To capture these joint features, a large number of joint learning models have been proposed [2], [3], among which bidirectional long short term memory (Bi-LSTM) [4] are commonly used as the shared parameter layer.

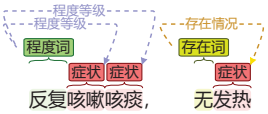

Fig. 1. An illustrative example of entity and relation extraction in the text of EHRs.

However, compared with the language models that benefit from abundant knowledge from pre-training and strong feature extraction capability, Bi-LSTM model has relatively lower generalization performance. To improve the performance, one of the solutions is to incorporate language model into joint learning as a shared parameter layer. However, existing models only introduce language models into the NER or RC task separately [5], [6], leading the joint features between entity and relationship types unable to be captured.

Given the aforementioned challenges and current researches, we propose a focused attention model based on widely known BERT language model [7] to jointly tackle NER and RC tasks. Specifically, through the dynamic range attention mechanism, we construct task-specific MASK matrix to control the attention range of the last K layers in BERT language model, leading to the model focusing on the words of the task. This process helps obtain the corresponding task-specific context-dependent representations. In this way, the modified BERT language model can be used as the shared parameter layer in joint learning of NER and RC task. We call the modified BERT language model as shared task representation encoder (STR-encoder) in the following paper. The main contributions of our work are summarized as follows:

- We propose a focused attention model to jointly learn NER and RC task. The model integrates BERT language model as a shared parameter layer to achieve better generalization performance.
- In the proposed model, we incorporate a novel structure, called STR-encoder, which changes the attention range of the last K layers in BERT language model to obtain task-specific context-dependent representations. It can make full use of the original structure of BERT to produce the vector of the task, and directly use the prior knowledge contained in the pre-trained language model.

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

- For RC task, we design two different MASK matrices to extract the required feature representation of RC task. The performances corresponding to the matrices are analyzed and compared in the experiment.

II. Related Work

Entity and relation extraction is to extract relational entity triplets. There are two kinds of approaches for the task, namely Pipeline and joint learning methods. The former decomposes the task into two subtasks, namely named entity recognition (NER) and relation classification (RC), while the latter attempts to solve the two tasks simultaneously.

A. Named Entity Recognition

In medical domain, we use NER to recognize disease, symptom, etc. In general, NER is formulated as a sequence tagging task using BIEOS (Begin, Inside, End, Outside, Single) [8] tagging strategy. Conventional methods in medical domain can be divided into two categories, i.e., statistical and neural network methods. The former are generally based on conditional random fields (CRF) [9] which relies on handcrafted features and external knowledges to improve the accuracy. Neural network methods typically use neural network to calculate the features without tedious feature engineering, e.g., bidirectional long short term memory neural network [10]. However, none of the above methods can make use of a large amount of unsupervised corpora, resulting in limited generalization performance.

B. Relation Classification

RC is closely related to NER task, which classifies the relationship between the entities identified in the text. The task is typically formulated into a classification problem that takes a piece of text and two entities in this text as inputs, and possible relation between the entities as output. The existing methods of RC can be roughly divided into two categories, i.e., traditional methods and neural network approaches. The former are based on feature-based [11] or kernel-based [12] approaches. These models usually spend a lot of time on feature engineering. Neural network methods can extract the relation features without complicated feature engineering. e.g., recurrent capsule network [13] and domain invariant convolutional neural network [14]. However, These methods cannot utilize joint features between entity and relation, resulting in lower generalization performance when compared with joint learning methods.

C. Joint Entity and Relation Extraction

Compared with pipeline methods, joint learning approaches are able to capture the joint features between entities and relations [15].

State-of-the-art joint learning methods can be divided into two categories, i.e., joint tagging methods and parameter sharing methods. Joint tagging methods transform NER and RC tasks into sequence tagging tasks through a specially designed tagging scheme, e.g., a novel tagging scheme proposed by Zheng et al. [2]. Parameter sharing methods share the feature extraction layer in the models of NER and RC. Compared to joint tagging methods, parameter sharing methods are able to effectively process multi-map problem. The most commonly shared parameter layer in medical domain is the Bi-LSTM network [16]. However, compared with language model, the feature extraction ability of Bi-LSTM is relatively weaker, and the model cannot obtain pre-training knowledge through a large amount of unsupervised corpora, which further reduces the robustness of extracted features.

III. Proposed Method

In this section, we first introduce classic BERT language model and the dynamic range attention mechanism. Then, we present a focused attention model for joint entity and relation extraction.

A. BERT Language Model

BERT [7] is a language model that utilizes bidirectional attention mechanism and large-scale unsupervised corpora to obtain effective context-sensitive representations of each word in a sentence. Owing to its effective structure and a rich supply of large-scale corpora, BERT has achieved state-of-the-art results on various natural language processing (NLP) tasks. The basic structure of BERT includes self attention encoder (SA-encoder) and downstream task layer. SA-encoder obtains the corresponding context-dependent representation using the sequence S and the MASK matrix:

$$H_N = SA\text{-}encoder(S, MASK) \tag{1}$$

The downstream task layer differs from task to task. In this work, we focus on NER and RC, which are further detailed in Section III-C2 and III-C4, respectively.

B. Dynamic Range Attention Mechanism

In BERT, MASK matrix is originally used to mask the padding portion of the text. However, we observe that, with the help of a specific MASK matrix, we can directly control the attention range of each word, thus obtaining specific context-sensitive representations.

Note that, when calculating the attention in BERT, the parameter matrix MASK$\in \{0,1\}^{T\times T}$, where T is the length of the sequence. If MASK$_{i,j}=0$, then we have (MASK$_{i,j}-1)\times\infty=-\infty$ and the Eq. (2), which indicates that the i-th word ignores the j-th word when calculating attention.

$$\begin{aligned} &Similar(i,j) \\ &= Softmax[\frac{QK^T}{\sqrt{d_k}}_{i,j} + (MASK_{i,j}-1)\times\infty] \\ &= Softmax(-\infty) \end{aligned} \tag{2}$$

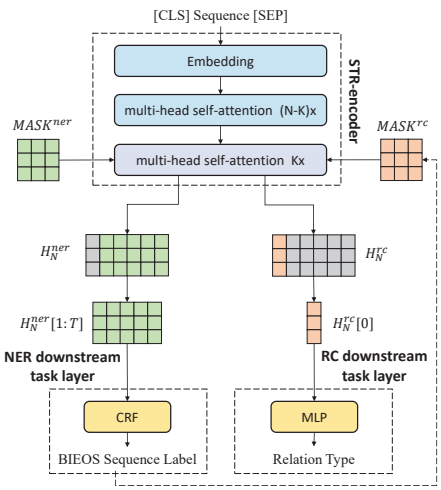

Fig. 2. The architecture of our proposed model.

When $\text{MASK}_{i,j} = 1$, we have $(MASK_{i,j} - 1) \times \infty = 0$ and the Eq. (3), which means the i-th word considers the j-th word when calculating attention.

$$\begin{aligned} &Similar(i,j) \\ &= Softmax[\frac{QK^T}{\sqrt{d_k}}_{i,j} + (MASK_{i,j} - 1) \times \infty] \\ &= Softmax(\frac{QK^T}{\sqrt{d_k}}_{i,j}) \end{aligned} \tag{3}$$

C. Focused Attention Model

The architecture of the proposed model is demonstrated in the Fig. 2. The focused attention model is essentially a joint learning model of NER and RC based on shared parameter approach. It contains layers of shared parameter, NER downstream task and RC downstream task.

The shared parameter layer, called shared task representation encoder (STR-encoder), is improved from BERT through dynamic range attention mechanism. It contains an embedded layer and N multi-head self-attention layers which are divided into two blocks. The first $N-K$ layers are only responsible for capturing the context information, and the context-dependent representations of words are expressed as H_{N-K}. According to characteristics of NER and RC, the remaining K layers use the MASK^{task} matrix setting by the dynamic range attention mechanism to focus the attention on the words. In this manner, we can obtain task-specific representations H_N^{task} and then pass them to corresponding downstream task layer. In addition, the segmentation point K is a hyperparameter, which is discussed in Section V-A.

Given a sequence, we add a $[CLS]$ token in front of the sequence and a $[SEP]$ token at the end of the sequence as BERT does. After the Embedding layer, the initial vector of each word in the sequence S is represented as H_0, which is same as BERT. Then we input H_0 to the first $N-K$ multi-head self-attention layers. In theses layers, attention of a single word is evenly distributed on all the words in the sentence to capture the context information. Given the output (H_{m-1}) from the $(m-1)$-th layer, the output of current layer is calculated as:

$$H'_m = LN[H_{m-1} + MHSA(H_{m-1}, MASK^{all})] \tag{4}$$

$$H_m = LN[H'_m + PosFF(H'_m)] \tag{5}$$

where $MHSA$, $PosFF$ and LN represent multi-head self attention, feed forward and layer normalization [17] and $\text{MASK}^{all} \in \{1\}^{T \times T}$ indicates each word calculates attention with all the other words of the sequence.

The remaining K layers focus on words of downstream task by task-specific matrix MASK^{task} based on dynamic range attention mechanism. Given the output (H_{m-1}^{task}) of previous $(m-1)$-th layer, the current output (H_m^{task}) is calculated as:

$$H_m^{'task} = LN[H_{m-1}^{task} + MHSA(H_{m-1}^{task}, MASK^{task})] \tag{6}$$

$$H_m^{task} = LN[H_m^{'task} + PosFF(H_m^{'task})] \tag{7}$$

where $H_{N-K}^{task} = H_{N-K}$ and $task \in \{ner, rc\}$.

As for STR-encoder, we only input different MASK^{task} matrices, which calculate various representations of words required by different downstream task (H_N^{task}) with the same parameters:

$$H_N^{task} = STR\text{-}encoder(S, MASK^{task}, MASK^{all}) \tag{8}$$

The structure has two advantages:

- It obtains the representation vector of the task through the strong feature extraction ability of BERT. Compared with the complex representation conversion layer, the structure is easier to optimize.
- It does not significantly adjust the structure of the BERT language model, so the structure can directly use the prior knowledge contained in the parameters of pre-trained language model.

Subsequently, we will introduce the construction of MASK^{task} and downstream task layer of NER and RC in blocks.

1) The Construction of MASK^{ner}: In NER, the model needs to output the corresponding BIEOS tag of each word in the sequence. In order to improve the accuracy, the appropriate attention weight should be learned through parameter optimization rather than limiting the attention range of each word. Therefore, according to the dynamic range attention mechanism, the value of the MASK^{ner} matrix should be set to $\text{MASK}_{ner} \in \{1\}^{T \times T}$, indicating that each word can calculate attention with any other words in the sequence.

2) The Construction of NER Downstream Task Layer: In NER, the downstream task layer needs to convert the representation vector of each word in the output of STR-encoder into the probability distribution of the corresponding BIEOS tag. Compared with the single-layer neural network, CRF model can capture the link relation between two tags [18]. As a result, we perform CRF layer to get the probability distribution of tags. Specifically, the representation vectors of all the words except $[CLS]$ token in the output of STR-encoder are sent to the CRF layer. Firstly, CRF layer calculates the

emission probabilities by linearly transforming these vectors. Afterwards, layer ranks the sequence of tags by means of emission and transition probabilities. Finally, the probability distribution of sequence of tags is obtained by softmax function:

$$H_p^{ner} = H_N^{ner}[1:T] \times W_{ner}^* + b_{ner} \tag{9}$$

$$Score(L|H_p^{ner}) = \sum_{t=1}^{T}(A_{L_{t-1},L_t} + {H_p^{ner}}_{t,L_t}) \tag{10}$$

$$p_{ner}(L|S, MASK^{ner}, MASK^{all}) = \frac{e^{Score(L|H_p^{ner})}}{\sum_J e^{Score(J|H_p^{ner})}} \tag{11}$$

The loss function of NER is shown as Eq. (12), and our training goal is to minimize L_{ner}, where L' indicates the real tag sequence.

$$L_{ner} = -Log[p_{ner}(L'|S, MASK^{ner}, MASK^{all})] \tag{12}$$

3) The Construction of MASK^{rc}: In RC, the relation between two entities are represented by a vector. In order to obtain the vector, we confine the attention range of $[CLS]$ token, which is originally used to summarize the overall representation of the sequence, to two entities. Thus, the vector of $[CLS]$ token can accurately summarize the relation between two entities. Based on the dynamic range attention mechanism, we propose two kinds of MASK^{rc} denoted as Eq. (13) and (14), respectively.

$$MASK_{i,j}^{rc} = \begin{cases} 1 & \text{if } i \in P_{CLS},\ j \in P_{CLS,EN1,EN2} \\ 1 & \text{if } i \notin P_{CLS} \\ 0 & \text{else} \end{cases} \tag{13}$$

$$MASK_{i,j}^{rc} = \begin{cases} 1 & \text{if } i,j \in P_{CLS,EN1,EN2} \\ 0 & \text{else} \end{cases} \tag{14}$$

where P_x represents the position of x in sequence S.

The difference between the two matrices is whether the attention range of entity 1 and 2 is confined. In Eq. (13), the attention range of entity 1 and 2 is not confined, which leads to the vector of RC shifting to the context information of entity. Relatively, in Eq. (14), only $[CLS]$, entity 1 and 2 are able to pay attention to each other, leading the vector of RC shifting to the information of entity itself. Corresponding to the RC task on medical text, the two MASK matrices will be further analyzed in Section V-A.

4) The Construction of RC Downstream Task Layer: For RC, the downstream task layer needs to convert the representation vector of $[CLS]$ token in the output of STR-encoder into the probability distribution of corresponding relation type. In this paper, we use multilayer perceptron (MLP) to perform this conversion. Specifically, the vector is converted to the probability distribution through two perceptrons with $Tanh$ and $Softmax$ as the activation function, respectively:

$$H_p^{rc} = Tanh(H_N^{rc}[0] \times W_{rc1} + b_{rc1}) \tag{15}$$

$$p_{rc}(R|S, MASK^{rc}, MASK^{all}) = Softmax(H_p^{rc} \times W_{rc2} + b_{rc2}) \tag{16}$$

The training is to minimize loss function L_{rc}, denoted as Eq. (17), where R' indicates the real relation type.

$$L_{rc} = -Log[p_{rc}(R'|S, MASK^{rc}, MASK^{all})] \tag{17}$$

D. Joint Learning

Note that, the parameters are shared in the model except the downstream task layers of NER and RC, which enables STR-encoder to learn the joint features of entities and relations. Moreover, compared with the existing parameter sharing model (e.g., Joint-Bi-LSTM [4]), the feature representation ability of STR-encoder is improved by the feature extraction ability of BERT and its knowledge obtained through pre-training. The loss function of the joint model (i.e., L_{all}) will be obtained as follows:

$$L_{all} = L_{ner} + L_{rc} \tag{18}$$

where L_{ner} and L_{rc} are defined in Eq. (12) and (17), respectively.

IV. Experimental Studies

A. Dataset and Evaluation Metrics

The dataset of entity and relation extraction is collected from coronary arteriography reports in Shanghai Shuguang Hospital. There are five types of entities, i.e., Negation, Body Part, Degree, Quantifier and Location. Meanwhile, five relations are included, i.e., Negative, Modifier, Position, Percentage and No Relation. 85% of "No Relation" in the dataset are discarded for balance purpose. The statistics of the entities and relations are demonstrated in Table I.

TABLE I
STATISTICS OF DIFFERENT TYPES OF ENTITIES AND RELATIONS

Entity Type	Number	Relation Type	Direction	Number
Negation	103	Negative	e2 to e1	406
Body Part	492	Modifier	e2 to e1	1,068
Degree	658	Position	e1 to e2	389
Quantifier	422	Percentage	bi-direction	100 / 256
Location	461	No Relation	none	1,975
Total	2,136	Total	none	4,194

* The bi-direction indicates there are two directions, i.e., e1 to e2 and e2 to e1.

In order to ensure the effectiveness of our experiment, we divide the dataset into training, development and test in the ratio of 8:1:1. In the following experiments, we use common performance measures such as Precision, Recall, and F_1-score [19] to evaluate NER, RC and joint models.

B. Experimental Setup

The training of focused attention model proposed in this paper can be divided into two stages. In the first stage, we need to pre-train the shared parameter layer. Due to the high cost of pre-training BERT, we directly adopted parameters pre-trained by Google in Chinese general corpus. In the second stage, we need to fine-tune NER and RC tasks jointly. Parameters of the two downstream task layers are randomly initialized. The two hyperparameters K and MASK^{rc} in the model will be further studied in Section V-A.

C. Experimental Result

To evaluate the performance of our focused attention model, we compare it with state-of-the-art methods on the task of NER, RC and joint entity and relation extraction, respectively.

Based on NER, we experimentally compare our focused attention model with other reference algorithms. These algorithms consist of two NER models in medical domain (i.e., Bi-LSTM [20] and RDCNN [21]) and one joint model in generic domain (i.e., Joint-Bi-LSTM [4]). In addition, we originally plan to use the joint model [16] in the medical domain, but the character-level representations cannot be implemented in Chinese. Therefore, we replace it with a generic domain model [4] in similar structure. As demonstrated in Table II, the proposed model achieves the best performance, and its precision, recall and F_1-score reach 96.69%, 97.09% and 96.89%, which outperforms the second method by 0.2%, 0.40% and 1.20%, respectively.

TABLE II
COMPARISONS WITH THE DIFFERENT METHODS ON THE TASK OF NER

Methods	NER		
	Precision	Recall	F_1-score
Bi-LSTM [18]	94.46	94.07	94.26
RDCNN [21]	96.49	94.90	95.69
Joint-Bi-LSTM [4]	93.84	96.69	95.24
Our model	**96.69**	**97.09**	**96.89**

To further investigate the effectiveness of the proposed model on RC, we use two RC models in medical domain (i.e., RCN [13] and CNN [22]) and one joint model in generic domain (i.e., Joint-Bi-LSTM [4]) as baseline methods. Since RCN and CNN methods are only applied to RC tasks and cannot extract entities from the text, so we directly use the correct entities in the text to evaluate the RC models. Table III illustrate that the focused attention model achieves the best performance, and its precision, recall and F_1-score reach 96.06%, 96.83% and 96.44%, which beats the second model by 1.57%, 1.59% and 1.58%, respectively.

TABLE III
COMPARISONS WITH THE DIFFERENT METHODS ON THE TASK OF RC

Methods	RC with Correct Entities		
	Precision	Recall	F_1-score
RCN [13]	90.77	93.65	92.19
CNN [23]	94.49	95.24	94.86
Joint-Bi-LSTM [4]	92.92	92.86	92.88
Our model	**96.06**	**96.83**	**96.44**

For the task of joint entity and relation extraction, we use Joint-Bi-LSTM [4] as baseline method. Since these two models are joint learning, we use the entities predicted in NER as the input for RC. From Table IV, we observe that our focused attention model achieves the best performance, and its F_1-score reaches 96.89% and 88.51%, which is 1.65% and 1.22% higher than the second method, respectively. These observations confirm that the feature representation of STR-encoder is indeed stronger than existing common models.

TABLE IV
COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE TASK OF JOINT ENTITY AND RELATION EXTRACTION

Methods	NER			RC with Predicted Entities		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
Joint-Bi-LSTM	93.84	96.69	95.24	93.64	81.75	87.29
Our model	**96.69**	**97.09**	**96.89**	**95.41**	**82.54**	**88.51**

V. EXPERIMENTAL ANALYSIS

In this section, we perform additional experiments to analyze the influence of different settings on segmentation points K, and different settings on MASK^{rc} and joint learning.

A. Hyperparameter Analysis

We further study the impacts of different settings on segmentation points K defined in Section III-C and different settings on MASK^{rc} defined in Section III-C3. As shown in Table V, when $K = 4$ and MASK^{rc} use Eq. (14), RC reaches the best F_1-score of 92.18%. When $K = 6$ and MASK^{rc} use Eq. (13), NER achieves the best F_1-score of 96.77%. One possible reason is that MASK^{rc} defined in Eq. (13) doesn't confine the attention range of entity 1 and 2, which enables the model to further learn context information in shared parameter layer, leading to a higher F_1-score for NER. For RC, the F_1-score with $K = 4$ is the lowest when MASK^{rc} uses Eq. (13), and reaches the highest when MASK^{rc} uses Eq. (14). One possible reason is that the two hyperparameters are closely related to each other. However, how they interact with each other in the focus attention model is still an open question.

TABLE V
COMPARISONS WITH DIFFERENT HYPERPARAMETERS ON THE TASK OF JOINT ENTITY AND RELATION EXTRACTION

K	MASK	NER			RC with Predicted Entities		
		Precision	Recall	F_1-score	Precision	Recall	F_1-score
2	Eq. (13)	95.07	97.93	96.48	97.08	**87.37**	91.97
4	Eq. (13)	94.98	98.20	96.56	97.06	86.84	91.67
6	Eq. (13)	**95.39**	**98.20**	**96.77**	98.20	86.32	91.88
2	Eq. (14)	94.67	97.93	96.27	95.95	**87.37**	91.46
4	Eq. (14)	94.77	97.87	96.29	**98.21**	86.84	**92.18**
6	Eq. (14)	94.47	97.93	96.17	96.51	**87.37**	91.71

B. Ablation Analysis

In order to evaluate the influence of joint learning, we train NER and RC models separately as an ablation experiment. In addition, we also use correct entities to evaluate RC, excluding the effect of NER results on the RC results, and independently compare the NRE and RC tasks.

As shown in Table VI, compared with training separately, the results are improved by 0.52% score in F_1-score for NER and 0.82% score in F_1-score for RC. It shows that

TABLE VI
COMPARISONS WITH TRAINING NER AND RC TASKS SEPARATELY

Methods	NER			RC with Correct Entities		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
Only NER	95.14	**97.64**	96.37	-	-	-
Only RC	-	-	-	96.00	95.24	95.62
Our model	**96.69**	97.09	**96.89**	**96.06**	**96.83**	**96.44**

joint learning can help to learn the joint features between NER and RC and improves the accuracy of two tasks at the same time. For NER, precision score is improved by 1.55%, but recall score is reduced by 0.55%. One possible reason is that, although the relationship type can guide the model to learn more accurate entity types, it also introduces some uncontrollable noise. In summary, joint learning is an effective method to obtain the best performance.

VI. CONCLUSION

In order to structure medical text, entity and relation extraction is an indispensable step. In this paper, we propose a focused attention model to jointly learn NER and RC task based on a shared task representation encoder which is transformed from BERT through dynamic range attention mechanism. Compared with existing models, our model can extract the entities and relations from the medical text more accurately. The experimental results on coronary angiography texts verify the effectiveness of our model.

VII. ACKNOWLEDGMENT

The authors would like to appreciate the efforts of the editors and valuable comments from the anonymous reviewers. This work is supported by the National Key R&D Program of China for "Precision Medical Research" (Grant No. 2018YFC0910500), the National Natural Science Foundation of China (Grant No. 61772201), the Special Fund Project for "Shanghai Informatization Development in Big Data" (Grant No. 201901043) and the Open Fund of Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University (Grant No. 2019MIP004).

REFERENCES

[1] T. D. Gunter and N. P. Terry, "The emergence of national electronic health record architectures in the united states and australia: models, costs, and questions," *Journal of Medical Internet Research*, vol. 7, no. 1, p. e3, 2005.

[2] S. Zheng, F. Wang, H. Bao, Y. Hao, P. Zhou, and B. Xu, "Joint extraction of entities and relations based on a novel tagging scheme," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1227–1236.

[3] C. Sun, Y. Gong, Y. Wu, M. Gong, D. Jiang, M. Lan, S. Sun, and N. Duan, "Joint type inference on entities and relations via graph convolutional networks," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 2019, pp. 1361–1370.

[4] S. Zheng, Y. Hao, D. Lu, H. Bao, J. Xu, H. Hao, and B. Xu, "Joint entity and relation extraction based on a hybrid neural network," *Neurocomputing*, vol. 257, pp. 59–66, 2017.

[5] C. Dogan, A. Dutra, A. Gara, A. Gemma, L. Shi, M. Sigamani, and E. Walters, "Fine-grained named entity recognition using elmo and wikidata," *arXiv preprint arXiv:1904.10503*, 2019.

[6] C. Alt, M. Hübner, and L. Hennig, "Improving relation extraction by pre-trained language representations," *arXiv preprint arXiv:1906.03088*, 2019.

[7] 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: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[8] V. Krishnan and V. Ganapathy. (2005) Named entity recognition. [Online]. Available: http://cs229.stanford.edu/proj2005/KrishnanGanapathy-NamedEntityRecognition.pdf

[9] M. Skeppstedt, M. Kvist, G. H. Nilsson, and H. Dalianis, "Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study," *Journal of Biomedical Informatics*, vol. 49, pp. 148–158, 2014.

[10] Q. Wang, Y. Zhou, T. Ruan, D. Gao, Y. Xia, and P. He, "Incorporating dictionaries into deep neural networks for the chinese clinical named entity recognition," *Journal of Biomedical Informatics*, vol. 92, p. 103133, 2019.

[11] B. Rink and S. Harabagiu, "Utd: Classifying semantic relations by combining lexical and semantic resources," in *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2010, pp. 256–259.

[12] D. Zelenko, C. Aone, and A. Richardella, "Kernel methods for relation extraction," *Journal of Machine Learning Research*, vol. 3, no. Feb, pp. 1083–1106, 2003.

[13] Q. Wang, J. Qiu, Y. Zhou, T. Ruan, D. Gao, and J. Gao, "Automatic severity classification of coronary artery disease via recurrent capsule network," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 1587–1594.

[14] S. Sahu, A. Anand, K. Oruganty, and M. Gattu, "Relation extraction from clinical texts using domain invariant convolutional neural network," in *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 206–215.

[15] Q. Li and H. Ji, "Incremental joint extraction of entity mentions and relations," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 402–412.

[16] F. Li, M. Zhang, G. Fu, and D. Ji, "A neural joint model for entity and relation extraction from biomedical text," *BMC Bioinformatics*, vol. 18, no. 1, p. 198, 2017.

[17] 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.

[18] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[19] Y. Liu, Y. Zhou, S. Wen, and C. Tang, "A strategy on selecting performance metrics for classifier evaluation," *International Journal of Mobile Computing and Multimedia Communications (IJMCMC)*, vol. 6, no. 4, pp. 20–35, 2014.

[20] M. Gridach, "Character-level neural network for biomedical named entity recognition," *Journal of Biomedical Informatics*, vol. 70, pp. 85–91, 2017.

[21] J. Qiu, Y. Zhou, Q. Wang, T. Ruan, and J. Gao, "Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field," *IEEE Transactions on NanoBioscience*, 2019.

[22] S. Sahu, A. Anand, K. Oruganty, and M. Gattu, "Relation extraction from clinical texts using domain invariant convolutional neural network," in *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 206–215.

[23] T. H. Nguyen and R. Grishman, "Relation extraction: Perspective from convolutional neural networks," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 2015, pp. 39–48.