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

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 F1-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.

Introduction

UTF8gkai 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 formulationBIBREF0. 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. Fig. FIGREF1 shows an example of entity and relation extraction in the text of EHRs. The text contains three relational entity triples, i.e., $<$咳嗽, 程度等级, 反复$>$ ($<$cough, degree, repeated$>$), $<$咳痰, 程度等, 反复$>$ ($<$expectoration, degree, repeated$>$) and $<$发热, 存在情况, 无$>$ ($<$fever, presence, nonexistent$>$). By using the symptom as the central word, these triples can then be converted into three key-value pairs, i.e., $<$咳嗽的程度等级, 反复$>$ ($<$degree of cough, repeated$>$), $<$咳痰的程度等级, 反复$>$ ($<$degree of expectoration, repeated$>$) and $<$发热的存在情况, 无$>$ ($<$presence of fever, nonexistent$>$).

UTF8gkai To solve the task of entity and relation extraction, researchers usually follows pipeline processing and split the task into two sub-tasks, namely named entity recognition (NER)BIBREF1 and relation classification (RC)BIBREF2, 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. FIGREF1, 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 BIBREF3, BIBREF4, among which bidirectional long short term memory (Bi-LSTM) BIBREF5, BIBREF6 are commonly used as the shared parameter layer. 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, a simple solution is to incorporate language model into joint learning as a shared parameter layer. However, the existing models only introduce language models into the NER or RC task separately BIBREF7, BIBREF8. Therefore, the joint features between entity and relationship types still can not be captured. Meanwhile, BIBREF9 considered the joint features, but it also uses Bi-LSTM as the shared parameter layer, resulting the same problem as discussed previously.

Given the aforementioned challenges and current researches, we propose a focused attention model based on widely known BERT language model BIBREF10 to jointly for 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 NER and RC task. We call the modified BERT language model shared task representation encoder (STR-encoder) in the following paper.

To sum up, 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 can directly use the prior knowledge contained in the pre-trained language model.

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

The rest of the paper is organized as follows. We briefly review the related work on NER, RC and joint entity and relation extraction in Section SECREF2. In Section SECREF3, we present the proposed focused attention model. We report the experimental results in Section SECREF4. Section SECREF5 is dedicated to studying several key factors that affect the performance of our model. Finally, conclusion and future work are given in Section SECREF6.

Related Work

Entity and relation extraction is to extract relational entity triplets which are composed of two entities and their relationship. Pipeline and joint learning are two kinds of methods to handle this task. Pipeline methods try to solve it as two subsequent tasks, namely named entity recognition (NER) and relation classification (RC), while joint learning methods attempt to solve the two tasks simultaneously.

Related Work ::: Named Entity Recognition

NER is a primary task in information extraction. In generic domain, we recognize name, location and time from text, while in medical domain, we are interested in disease and symptom. Generally, NER is solved as a sequence tagging task by using BIEOS(Begin, Inside, End, Outside, Single) BIBREF11 tagging strategy. Conventional NER 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)BIBREF12 and hidden Markov models BIBREF13, BIBREF14, which relies on hand-crafted features and external knowledge resources 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 BIBREF15 and residual dilated convolutional neural network BIBREF16. However, none of the above methods can make use of a large amount of unsupervised corpora, resulting in limited generalization performance.

Related Work ::: Relation Classification

RC is closely related to NER task, which tries to classify the relationship between the entities identified in the text, e.g, “70-80% of the left main coronary artery opening has stenosis" in the medical text, there is “modifier" relation between the entity “left main coronary artery" and the entity “stenosis". The task is typically formulated into a classification problem that takes a piece of text and two entities in the text as inputs, and the 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-basedBIBREF17, BIBREF18, BIBREF19 or kernel-basedBIBREF20 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., convolutional neural network BIBREF21, BIBREF22, BIBREF23, recurrent neural network BIBREF24 and long short term memory BIBREF25, BIBREF26. In medical domain, there are recurrent capsule network BIBREF27 and domain invariant convolutional neural network BIBREF28. However, These methods cannot utilize the joint features between entity and relation, resulting in lower generalization performance when compared with joint learning methods.

Related Work ::: Joint Entity and Relation Extraction

Joint entity and relation extraction tasks solve NER and RC simultaneously. Compared with pipeline methods, joint learning methods are able to capture the joint features between entities and relations BIBREF29.

State-of-the-art joint learning methods can be divided into two categories, i.e., joint tagging and parameter sharing methods. Joint tagging transforms NER and RC tasks into sequence tagging tasks through a specially designed tagging scheme, e.g., novel tagging scheme proposed by Zheng et al. BIBREF3. Parameter sharing mechanism shares 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 BIBREF9. 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.

Proposed Method

In this section, we introduce classic BERT language model and how to dynamically adjust the range of attention. On this basis, we propose a focused attention model for joint entity and relation extraction.

Proposed Method ::: BERT Language Model

BERT 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, e.g. ELMO BIBREF30 and GPT BIBREF31. Since its effective structure and a rich supply of large-scale corporas, BERT has achieved state-of-the-art results on various natural language processing (NLP) tasks, such as question answering and language inference. The basic structure of BERT includes self attention encoder (SA-encoder) and downstream task layer. To handle a variety of downstream tasks, a special classification token called ${[CLS]}$ is added before each input sequence to summarize the overall representation of the sequence. The final hidden state corresponding to the token is the output for classification tasks. Furthermore, SA-encoder includes one embedded layer and N multi-head self-attention layers.

The embedding layer is used to obtain the vector representations of all the words in the sequence, and it consists of three components: word embedding (${e}_{word}$), position embedding (${e}_{pos}$), and type embedding (${e}_{type}$). Specifically, word embeddings are obtained through the corresponding embedding matrices. Positional embedding is used to capture the order information of the sequence which is ignored during the self-attention process. Type embedding is used to distinguish two different sequences of the input. Given an input sequence (${S}$), the initial vector representations of all the words in the sequence (${H}_0$) are as follows:

where ${LN}$ stands for layer normalization BIBREF32.

N multi-head self-attention layers are applied to calculate the context-dependent representations of words (${H}_N$) based on the initial representations (${H}_0$). To solve the problems of gradient vanishing and exploding, ResNet architectureBIBREF33 is applied in the layer. In N multi-head self-attention layers, every layer can produce the output (${H}_{m}$) given the previous output of $(m-1)$-th layer (${H}_{m-1}$):

where ${H}_{m}^{\prime }$ indicates intermediate result in the calculation process of m-th layer, ${MHSA}_{h}$ and ${PosFF}$ represent multi-head self-attention and feed-forward that are defined as follows:

where h represents the number of self-attention mechanisms in multi-head self-attention layer and ${Att}$ is a single attention mechanism defined as follows:

where Q, K and V represent “query”, “key” and “value” in the attention calculation process, respectively. Additionally, MASK matrix is used to control the range of attention, which will be analyzed in detail in Section SECREF14.

In summary, SA-encoder obtains the corresponding context-dependent representation by inputting the sequence S and the MASK matrix:

Finally, the output of SA-encoder is passed to the corresponding downstream task layer to get the final results. In BERT, SA-encoder can connect several downstream task layers. In terms of the content in the paper, the tasks are NER and RC, which will be further detailed in Section SECREF25 and SECREF32.

Proposed Method ::: Dynamic Range Attention Mechanism

In BERT, MASK matrix is originally used to mask the padding portion of the text. However, we found that by designing a specific MASK matrix, we can directly control the attention range of each word, thus obtaining specific context-sensitive representations. Specially, when calculating the attention (i.e., Equation (DISPLAY_FORM12)), the parameter matrix $MASK\in {\lbrace 0,1\rbrace }^{T\times T}$, 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 Equation (DISPLAY_FORM15), which indicates that the i-th word ignores the j-th word when calculating attention.

While $MASK_{i,j} = 1$, we have $(MASK_{i,j}-1)\times \infty = 0$ and the Equation (DISPLAY_FORM16), which means the i-th word considers the j-th word when calculating attention.

Proposed Method ::: Focused Attention Model

The architecture of the proposed model is demonstrated in the Fig. FIGREF18. 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 former $N-K$ 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 SECREF47.

Given a sequence, we add a $[CLS]$ token in front of the sequence as BERT does, and a $[SEP]$ token at the end of the sequence as the end symbol. After the Embedding layer, the initial vector of each word in the sequence S is represented as H_0, and is calculated by Equation (DISPLAY_FORM9). Then we input H_0 to the former $N-K$ 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:

where $MASK^{all}\in {\lbrace 1\rbrace }^{T\times T}$ 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 model calculate the current output (H_m^{task}) as:

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

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:

This 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.

Proposed Method ::: Focused Attention Model ::: The Construction of @!START@$MASK^{ner}$@!END@

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 $MASK_{ner}\in {\lbrace 1\rbrace }^{T\times T}$, indicating that each word can calculate attention with any other words in the sequence.

Proposed Method ::: Focused Attention Model ::: 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 BIBREF34. 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 after self attention layer. Firstly, CRF layer calculates the emission probabilities by linearly transforming these vectors. Afterwards, layer ranks the sequence of tags by means of transition probabilities of the CRF layer. Finally, the probability distribution of tags is obtained by softmax function:

H_N^{ner} is the output of STR-encoder when given $MASK^{ner}$, $H_N^{ner}[1:T]$ denotes the representation of all words except $[CLS]$ token. H_p^{ner} is the emission probability matrix of CRF layer, $Score(L|H_p^{ner})$ represents the score of the tag sequence L, A_{L_{t-1},L_t} means the probability of the $(t-1)$-th tag transfering to the t-th tag, and ${H_p^{ner}}_{t,L_t}$ represents the probability that the t-th word is predicted as an L_t tag. $p_{ner}(L|S,MASK^{ner},MASK^{all})$ indicates the probabilities of the tag sequence L when given S, $MASK^{ner}$ and $MASK^{all}$, and J is the possible tag sequence.

The loss function of NER is shown as Equation (DISPLAY_FORM29), and the training goal is to minimize L_{ner}, where $L^{\prime }$ indicates the real tag sequence.

Proposed Method ::: Focused Attention Model ::: The Construction of @!START@$MASK^{rc}$@!END@

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 Equation (DISPLAY_FORM31) and ().

where P_{CLS}, P_{EN1} and P_{EN2} represent the positions of $[CLS]$, entity 1 and 2 in sequence S, respectively.

The difference between the two matrices is whether the attention range of entity 1 and 2 is confined. In Equation (DISPLAY_FORM31), 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 Equation (), 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 SECREF47.

Proposed Method ::: Focused Attention Model ::: 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 carry out this conversion. Specifically, the vector is converted to the probability distribution through two perceptrons with $Tanh$ and $Softmax$ as the activation function, respectively:

H_N^{rc} is the output of STR-encoder when given $MASK^{rc}$, $H_N^{rc}[0]$ denotes the representation of $[CLS]$ in the output of STR-encoder, H_p^{rc} is the output of the first perceptron. $p_{rc}(R|S,MASK^{rc},MASK^{all})$ is the output of the second perceptron and represents the probabilities of the relation type R when given the sequence S, $MASK^{rc}$ and $MASK^{all}$.

The training is to minimize loss function L_{rc}, denoted as Equation (DISPLAY_FORM34), where $R^{\prime }$ indicates the real relation type.

Proposed Method ::: 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-LSTMBIBREF6), the feature representation ability of STR-encoder is improved by the feature extraction ability of BERT and its knowledge obtained through pre-training.

Proposed Method ::: Additional Instructions for MASK

Due to the limitation of deep learning framework, we have to pad sequences to the same length. Therefore, all MASK matrices need to be expanded. The formula for expansion is as follows:

where $maxlen$ is the uniform length of the sequence after the padding operation.

Experimental Studies

In this section, we compare the proposed model with NER, RC and joint models. Dataset description and evaluation metrics are first introduced in the following contents, followed by the experimental settings and results.

Experimental Studies ::: 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. 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 TABREF39 and TABREF40, respectively.

In order to ensure the effectiveness of the 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 to evaluate NER, RC and joint models.

Experimental Studies ::: 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 parameters are optimized by Adam optimization algorithmBIBREF35 and its learning rate is set to 10^{-5} in order to retain the knowledge learned from BERT. Batch size is set to 64 due to graphics memory limitations. The loss function of the model (i.e., L_{all}) will be obtained as follows:

where L_{ner} is defined in Equation (DISPLAY_FORM29), and L_{rc} is defined in Equation (DISPLAY_FORM34).

The two hyperparameters K and $MASK^{rc}$ in the model will be further studied in Section SECREF47. Within a fixed number of epochs, we select the model corresponding to the best relation performance on development dataset.

Experimental Studies ::: Experimental Result

In order to fully verify the performance of focused attention model, we will compare the different methods on the task of NER, RC and joint entity and relation extraction.

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-LSTMBIBREF36 and RDCNNBIBREF16) and one joint model in generic domain (i.e., Joint-Bi-LSTM BIBREF6). In addition, we originally plan to use the joint modelBIBREF9 in the medical domain, but the character-level representations cannot be implemented in Chinese. Therefore, we replace it with a generic domain model BIBREF6 with similar structure. As demonstrated in Table TABREF44, 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.

To further investigate the effectiveness of the proposed model on RC, we use two RC models in medical domain (i.e., RCN BIBREF27 and CNN BIBREF37) and one joint model in generic domain (i.e., Joint-Bi-LSTMBIBREF6) 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 TABREF45 illustrate that 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.

In the task of joint entity and relation extraction, we use Joint-Bi-LSTMBIBREF6 as baseline method. Since both of the models are joint learning, we can use the entities predicted in NER as the input for RC. From Table TABREF46, we can observe that focused attention model achieves the best performance, and its F$_1$-scores reaches 96.89% and 88.51%, which is 1.65% and 1.22% higher than the second method.

In conclusion, the experimental results indicate that the feature representation of STR-encoder is indeed stronger than existing common models.

Experimental Analysis

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

Experimental Analysis ::: Hyperparameter Analysis

In the development dataset, we further study the impacts of different settings on segmentation points K defined in Section SECREF17 and different settings on $MASK^{rc}$ defined in Section SECREF30.

As shown in Table TABREF48, when $K=4$ and $MASK^{rc}$ use Equation (), RC reached the best F$_1$-score of 92.18%. When $K=6$ and $MASK^{rc}$ use Equation (DISPLAY_FORM31), NER has the best F$_1$-score of 96.77%. One possible reason is that $MASK^{rc}$ defined in Equation (DISPLAY_FORM31) 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. In contrast, $MASK^{rc}$ defined in Equation () only allows $[CLS]$, entity 1 and 2 to pay attention to each other, which makes the learned features shift to the entities themselves, leading to a higher F$_1$-score of RC.

For RC, the F$_1$-score with $K=4$ is the lowest when $MASK^{rc}$ uses Equation (DISPLAY_FORM31), and reaches the highest when $MASK^{rc}$ uses Equation (). One possible reason is that the two hyperparameters are closely related to each other. However, how they interact with each other in focus attention model is still an open question.

Experimental Analysis ::: 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 use correct entities to evaluate RC, exclude the effect of NER results on the RC results, and independently compare the NRE and RC tasks.

As shown in Table TABREF49, compared with training separately, the results are improved by 0.52% score in F$_1$score for NER and 2.37% score in F$_1$score for RC. It shows that 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.

Conclusion and Future Work

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, the model can extract the entities and relations from the medical text more accurately. The experimental results on the dataset of coronary angiography texts verify the effectiveness of our model.

For future work, the pre-training parameters of BRET used in this paper are pre-trained in the corpus of the generic field so that it cannot fully adapt to the tasks in the medical field. We believe that retrain BRET in the medical field can improve the performance of the model in the specific domain.

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" under grant 2018YFC0910500.