mathematics

MDPI

Article

Leveraging Part-of-Speech Tagging Features and a Novel Regularization Strategy for Chinese Medical Named Entity Recognition

Miao Jiang †, Xin Zhang †, Chonghao Chen, Taihua Shao * and Honghui Chen

Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, No. 109 Deya Street, Changsha 410073, China; jiangmiao20@nudt.edu.cn (M.J.); zhangxin16@nudt.edu.cn (X.Z.); chenchonghao@nudt.edu.cn (C.C.); chenhonghui@nudt.edu.cn (H.C.)

* Correspondence: shaotaihua13@nudt.edu.cn

† These authors contributed equally to this work.

Abstract: Chinese Medical Named Entity Recognition (Chinese-MNER) aims to identify potential entities and their categories from the unstructured Chinese medical text. Existing methods for this task mainly incorporate the dictionary knowledge on the basis of traditional BiLSTM-CRF or BERT architecture. However, the construction of high-quality dictionaries is typically time consuming and labor-intensive, which may also damage the robustness of NER models. What is more, the limited amount of annotated Chinese-MNER data can easily lead to the over-fitting problem while training. With the aim of dealing with the above problems, we put forward a **B**ERT-**B**iLSTM-**C**RF model by integrating the part-of-speech (**P**OS) tagging features and a **R**egularization method (BBCPR) for Chinese-MNER. In BBCPR, we first leverage a POS fusion layer to incorporate external syntax knowledge. Next, we design a novel **RE**gularization mothod with **A**dversarial training and **D**ropout (READ) to improve the model robustness. Specifically, READ focuses on reducing the difference between the predictions of two sub-models through minimizing the bidirectional KL divergence between the adversarial output and original output distributions for the same sample. Comprehensive evaluations on two public data sets, namely, cMedQANER and cEHRNER from the Chinese Biomedical Language Understanding Evaluation benchmark (ChineseBLUE), demonstrate the superiority of our proposal in Chinese-MNER. In addition, ablation study shows that READ can effectively improve the model performance. Our proposal does well in exploring the technical terms and identifying the word boundary.

Keywords: Chinese-MNER; BERT-BiLSTM-CRF; part-of-speech; regularization

MSC: 68T50

Citation: Jiang, M.; Zhang, X.; Chen, C.; Shao, T.; Chen, H. Leveraging Part-of-Speech Tagging Features and a Novel Regularization Strategy for Chinese Medical Named Entity Recognition. *Mathematics* **2022**, *10*, 1386. https://doi.org/10.3390/math10091386

Received: 10 March 2022
Accepted: 13 April 2022
Published: 21 April 2022

1. Introduction

Named Entity Recognition (NER) is one of the core objectives in natural language processing (NLP) [1,2], whose purpose is to determine the underlying entities and their categories from the unstructured text [3]. As an essential component in many downstream NLP tasks, for instance, the correlation extraction [4], information retrieval [5], sarcasm detection [6], and so on, NER is always a hot research direction and attracts much attention in the NLP community. In general, most of the previous works are devoted to the English NER task and achieve promising performances by integrating word-level features [7]. Compared with English, the East Asian languages (e.g., Chinese) typically lack explicit word boundaries and have complex composition forms, which brings greater challenges for these languages for the development of a competitive NER model. For example, the property of present Chinese state-of-the-art (SOTA) models are much lower than the English SOTAs, with a gap of nearly 10% in terms of F1 metric [8]. What is more, recent studies pay

more attention to the domain-specific NER, e.g., medicine, which is much more complicated and requires external domain expertise [9,10].

In particular, in the current work, we pay attention to the research of Chinese Medical Named Entity Recognition (Chinese-MNER), which is considered as a character-level sequence labeling problem, while it is word level for English [11]. Recently, deep learning methods have been extensively employed in Chinese-MNER [10,12–15] due to their excellent ability in automatically extracting features from massive data. For instance, previous works attempt to leverage the Bi-directional Long Short-Term Memory (BiLSTM) network for acquiring sequence features and achieve comparable results [16]. In addition, on account of the excellent ability of the pre-trained language models in extracting the contextual features, transformer-based models (e.g., BERT [17]) are becoming a new paradigm for Chinese-MNER [15,18–21].

Specially, in the medical domain, the external expertise is beneficial in understanding the technical terms and identifying the word boundary, which motivates recent research to incorporate the dictionary knowledge on the basis of traditional BiLSTM-CRF or BERT architecture [9,10]. However, the construction of high-quality dictionaries is typically time consuming and labor-intensive, which may also damage the generalization and robustness of NER models [22]. Compared with the dictionary knowledge, the part-of-speech (POS) tagging features [23] are now readily available, which does not require additional manpower and material resources. The POS tagging features [24] can be regarded as supervised signals to guide the model to explicitly identify the word boundary for the reason that it contains potential word segmentation information [25]. Therefore, we argue that the POS tagging features are more suitable to be used for Chines-MNER than the dictionary knowledge. Last but not least, due to the restrictions of high specialization degree, ethics, and privacy, the annotated Chinese-MNER data are difficult to obtain and usually small in scale, which can result in the over-fitting problem easily when training the Chinese-MNER model [26].

For the sake of alleviating the above issues, we present a **B**ERT-**B**iLSTM-**C**RF with **P**OS and **R**egularization (BBCPR) model for Chines-MNER, which leverages a POS fusion layer to incorporate external syntax knowledge as well as introduces a novel **RE**gularization method with **A**dversarial training and **D**ropout (READ) to improve the model robustness. In general, our proposal is based on a combined MC-BERT [27] and BiLSTM-CRF modeling framework. We first utilize the MC-BERT to generate the context representation of each token in the Chinese medical text. Then, we design a POS fusion layer to integrate the part of speech tagging features and send them into a BiLSTM module as inputs. Finally, a standard conditional random fields (CRF) [25] layer is employed for decoding the sequence labels. Particularly, besides the traditional learning objective, we introduce an external Kullback–Leibler (KL) divergence loss based on READ. In detail, READ can generate the adversarial word embeddings through a Fast Gradient Method (FGM) as well as a dropout mechanism, which are subsequently put into a Softmax layer for forecasting the label distributions. After that, we can regularize the model predictions through minimizing the bidirectional KL divergence between the adversarial output and original output distributions for the same sample [28].

For proving the effectiveness of the proposal, we implement comprehensive experiments on two public data sets from the Chinese Biomedical Language Understanding Evaluation (ChineseBLUE) benchmark [27], i.e., cMedQANER and cEHRNER. The experimental results suggest that our presented BBCPR model is superior to the SOTA baseline, and the overall improvement for the F1 score on cEHRNER and cMedQANER datasets is 2.48% and 2.87%, respectively. Furthermore, the effectiveness of our designed modules is verified by the ablation studies.

In summary, the major contributions of this research can be concluded as below:

- We design a POS fusion layer that can explicitly learn the word boundary feature for the task of Chinese-MNER by incorporating the POS tagging features.

- We put forward a novel regularization approach READ to alleviate the over-fitting problem for Chinese-MNER and enhance the robustness of the model on small data.
- We conduct comprehensive experiments on two public datasets. The performance comparisons over several competitive baselines indicate the superiority of our proposal.

2. Related Work

In this section, we first summarize prior studies of Chinese-MNER and illustrate the differences between our proposal and prior works in Section 2.1. Then, we describe the related regularization and adversarial training methods in Section 2.2.

2.1. Deep Learning-Based Chinese-MNER

Deep learning approaches have been extensively applied in the task of Chinese-MNER [10,12,14,15] due to their excellent ability in automatically extracting features from massive data. Before the popularization of the pre-trained language model, most of the prior works leverage the convolutional neural networks [13,29,30], such as the recurrent neural networks [13,31], as well as their variants (i.e., bidirectional long short-term memory [11,32,33]) to represent the contextual features [15,16,34]. In addition, they usually adopt the conditional random fields to predict the label sequence. Among these models, the BiLSTM combined with CRF yields the best performance [16].

In the past several years, on account of the outstanding ability of pre-trained language models in representation learning, transformer-based models (e.g., BERT [17]) have become a new paradigm for Chinese-MNER [15,18–21]. BERT can apply prior semantic knowledge obtained from large unlabeled corpora to the downstream tasks through fine-tuning [17]. For instance, Xu et al. [20] leverage the contextual features learned by BERT to enrich the word semantics and incorporate them to the model of Bi-LSTM-CRF. Inspired by BERT, Lee et al. [18] introduce a pre-trained biomedical language representation model for biomedical text mining. Similarly, as a variant of BERT, RoBERTa [19] is also applied to learn the medical features, which uses the dynamic masking and eliminates the next sentence prediction task in pre-training.

Specially, in the medical domain, the external expertise can help the model understand the technical terms and identify the word boundary, which motivates recent research to incorporate the extra knowledge on the basis of traditional BiLSTM-CRF or BERT architecture [9,10]. For example, Li et al. [13] propose to incorporate the pre-trained medical dictionary as the model input. In addition, Dong et al. [35] adopt a radical-level LSTM to obtain pictographic root characteristics of Chinese.

However, the construction of high-quality dictionaries is typically time consuming and labor-intensive, which may also damage the generalization and robustness of NER models [22]. Compared with the dictionary knowledge, the POS tagging features [23,24] are now readily available, which does not require additional manpower and material resources. Therefore, in the present work, the POS fusion layer is designed to incorporate the POS tagging features, which can act as supervised signals to guide the model to explicitly identify the word boundary.

2.2. Regularization vs. Adversarial Training

When training the neural network on the small training set, the deep learning-based models usually perform poor generalization ability on the test data. To prevent the deep neural networks from suffering from the over-fitting problem, most of recent works introduce the regularization methods in their models, which includes weight penalties of L1 and L2 regularization, dropout, and batch normalization [36], etc.

Dropout is a typical regularization method and has been widely used to regularize the fully connected neural network due to its simplicity and efficiency [37]. It drops neurons from each layer of the neural network at random with probability p during the training process [38]. On this basis, Wan et al. [39] propose a novel type of dropout, called DropConnect. Different from randomly setting activation units within each layer as

zero, it randomly sets weights within the network as zero. However, the above methods typically work on the fully connected layer. However, for the convolution layer, the activation units are interrelated spatially. Thus, information can also flow in the network even if some neurons are dropped. To deal with this issue, Ghiasi et al. [40] design a structured dropout method named DropBlock to regularize the convolutional networks through dropping the units together in adjacent areas of the feature map. Instead of using dropout alone, some works combine it with other training frameworks. For instance, Gao et al. [41] take the standard dropout as noise and integrate it into a comparative learning framework, which advances the SOTA sentence embedding. Liu et al. [42] use the dropout to generate the positive sentence samples from the feature space and then train the encoder by a contrastive learning-based objective. When using the dropout method, inconsistency between the training samples and the inference samples may arise because of the randomness introduced by dropout. In response to this problem, Wu et al. [28] propose a R-Drop method, which regularizes the output distributions of two sub-models by minimizing the bidirectional KL-divergence for each data sample in the training.

In addition to the over-fitting, the robustness of the model is also a urgent problem to be solved, since traditional neural networks are easily cheated by slightly disturbed samples [22]. To address this issue, the adversarial training is recently introduced in the representation learning; among these methods, FGM [43] is a popular model used to generate adversarial examples, which makes neural network models robust against perturbations. The basic principle is to add disturbance to construct adversarial samples during model training, thus enhancing the model robustness when it meets adversarial samples. Goodfellow et al. [44] propose a rapid and simple approach for producing adversarial examples, which shows that adversarial training can give an extra regularization benefit in addition to the benefit of utilizing dropout alone. In addition, the experiments in this paper demonstrate that adversarial back-propagation, as a stand-alone regularizing method, performs well in improving the generalization and robustness of the network.

Inspired through the above approaches, in this study, we design a new regularization method combining R-Drop and FGM to deal with the over-fitting problem and enhance the model robustness. In accordance with generating adversarial samples, we conduct two dropouts and shorten the distance between the two sub-models by KL clustering.

3. Approach

In this work, we principally pay attention to the problem of Chinese-MNER. Here, we first formulaically define the Chinese-MNER problem and introduce the main notations employed in this study (see Section 3.1). Then, we show the technical specific information about our presented model of BBCPR (see Section 3.2). Finally, we describe how to integrate the designed regularization method READ into BBCPR (see Section 3.3).

3.1. Problem Definition and Notations

The Chinese-MNER task is intended to identify and predict entities (such as *diseases*, *symptoms*, *drugs*, etc.) from the unstructured Chinese medical text. In this paper, following Chen and Kong [8], Li et al. [15], Xu et al. [20], Zhou et al. [45], and Liang et al. [46], the Chinese-MNER task is treated as the sequence labeling problem. Given one piece of Chinese medical text $X = \{x_1, x_2, \ldots, x_n\}$ with n tokens as the input, the objective of a Chinese-MNER algorithm is to predict each token x_i in X with the BIO tag (*Begin*, *Inside*, *Outside*) and finally obtain a label sequence $Y = \{y_1, y_2, \ldots, y_n\}$ as the output. An instance in the real world of the labeled entities in the Chinese medical text is presented in Table 1. For the purpose of clarity, we conclude the major notations applied in this article in Table 2.

Table 1. An instance of the labeled entities in Chinese medical text. B-s denotes the beginning of entity symptom, I-s presents the interior of entity symptom, O stands for external entity.

Sentence	重	度	感	冒	患	者	容	易	出	现	高	热	呕	吐
		(Patients	with	severe	colds	are	prone	to	high	fever	and	vomiting.)		
BIO	O	O	B-d	I-d	B-p	I-p	O	O	O	O	B-s	I-s	B-s	I-s
POS tagging	a	a	n	n	n	n	a	a	v	v	a	a	v	v
Entity type			disease	disease	person	person					symptom	symptom	symptom	symptom

Table 2. Major notations applied in this paper.

Variable	Description
x_i	The expression of the i-th token in Chinese medical text
e_i	The BERT embedding of the i-th token
e'_i	The adversarial embedding of the i-th token
m_i	The output embedding of BERT for the i-th token
t_i	The POS tagging features of the i-th token
p_i	The POS embedding of the i-th token
v_i	The output fusion embedding of POS fusion layer
H	The final hidden representations produced by BiLSTM
P	The input matrix of CRF layer
y_i	Prediction label of the i-th token in Chinese medical text
$P(Y \mid X)$	The probability distribution of the original BERT embedding E
$P'(Y \mid X)$	The probability distribution of the adversarial embedding E'
$\mathcal{L}$,$\mathcal{L}_{NER}$,$\mathcal{L}_R$	The final, basic NER and regularization method loss
λ	The trade-off parameter for balancing $\mathcal{L}_{NER}$ and $\mathcal{L}_R$

3.2. Model Architecture

The overall architecture of our presented model BBCPR is presented in Figure 1, which principally contains four layers: namely, the MC-BERT layer (see Section 3.2.1), POS fusion layer (see Section 3.2.2), BiLSTM layer (see Section 3.2.3) and CRF layer (see Section 3.2.4), respectively. Since BERT-BiLSTM-CRF is the state-of-the-art workflow in NER [15,16,20], we follow Li et al. [15] and adopt the same workflow in this work. Differently, on the basis of this workflow, we replace the BERT with the MC-BERT [27], which is specially pre-trained and more appropriate in the field of medicine. Moreover, we design a POS fusion layer and propose a READ strategy to improve model performance in our workflow. Below, we will present the detailed process of each component in BBCPR.

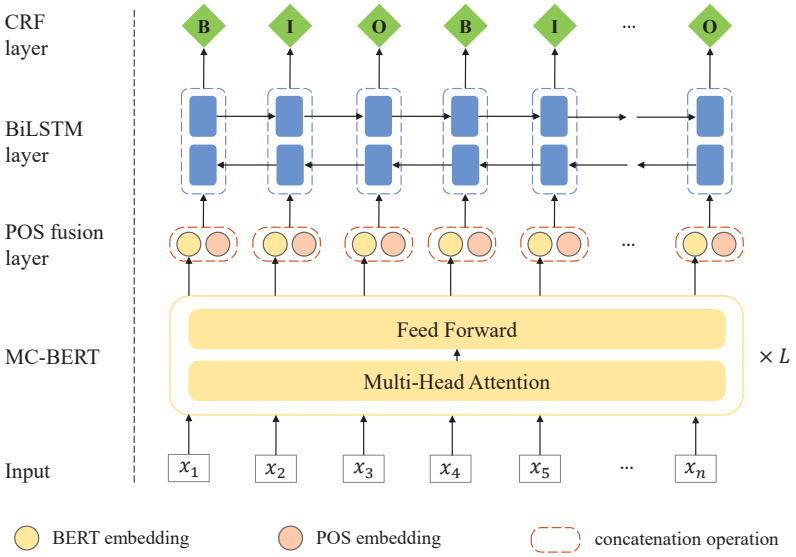


Figure 1. The architecture of BBCPR model.

3.2.1. MC-BERT

As is known to all, BERT is pre-trained on the Wikipedia corpus [17]. However, medical texts typically contain professional terms that seldom appear in general corpus. To bridge such a semantic gap, Zhang et al. [27] propose a variant of BERT, i.e., MC-BERT, which is further trained on the Chinese medical corpus and performs well in extracting the medical contextual features. Thus, MC-BERT is specially adopted in this work.

For the input $X = \{x_1, x_2, \ldots, x_n\}$, we first adopt MC-BERT to convert them into a sequence of BERT embeddings $E = \{e_1, e_2, \ldots, e_n\}$ through summing the position embedding, segment embedding, as well as token embedding.

Generally, MC-BERT retains the same structure as BERT, which is composed of a pile of L same layers. For convenience, the output of the l-th layer together with input of the first layer are represented as M_0 and M_l, respectively. The output representations M_{l-1} of the previous layer are placed into the Multi-head Self-Attention (MSA) sub-layer to acquire contextual-level representation $\tilde{M}_l$:

$$\tilde{M}_l = \text{LayerNorm}(\text{MSA}(M_{l-1}) + M_{l-1}). \tag{1}$$

Next, we gather the output representation of each encoder layer through feeding the contextual-level representation through a Feed-Forward Network (FFN) sub-layer. We formulate these operations as:

$$M_l = \text{LayerNorm}\big(\tilde{M}_l + \text{FFN}(\tilde{M}_l)\big), \tag{2}$$

where $M_l \in \mathbb{R}^{n \times d_{bert}}$, $l \in \{1, 2, 3, ..., L\}$ and d_{bert} denotes the hidden size of MC-BERT. Then, the final output embedding $M_L = \{m_1, m_2, \ldots, m_n\}$ is fed to the POS fusion layer.

3.2.2. POS Fusion Layer

MC-BERT processes the input Chinese medical text as a collection of token X and generates token-level embeddings M. However, a word is generally recognized as the smallest unit of semantic expression in Chinese, which results in the loss of semantic and increases the difficulty in extracting the entity boundary as well.

To tackle this issue, we design a POS fusion layer to incorporate the POS tagging features into the BBCPR model. Different from the existing dictionary strategy where labor costs are invariably high [10], the POS tagging features are simple, straightforward, and easily accessible. The POS is defined as the features of words that contains verbs, nouns, modal particles, adjectives, and so on, which can accurately label common words and thus distinguish medical entities from the edges of common words [25].

Formally, we employ the Baidu LAC toolkit [47] to generate the POS tagging features for each token in X as:

$$T = \{t_1, t_2, \ldots, t_n\}, \tag{3}$$

where T is the collection of POS tags; t_i is the corresponding POS tag of the i-th token x_i.

Next, with the aim of mapping the dispersed POS tag into the consecutive semantic space to conduct the model training, we generate the corresponding POS embedding p_i for each POS tag t_i as:

$$p_i = W_p t_i, \tag{4}$$

where $W_p \in \mathbb{R}^{d_p}$ is the learnable network parameter; and d_p stands for the POS-embedding dimension.

Subsequently, the concatenation operation is utilized to fuse each MC-BERT output embedding m_i and its corresponding POS embedding p_i. The formula is shown as follows:

$$v_i = [m_i; p_i], \tag{5}$$

where $v_i \in \mathbb{R}^{d_{bert}+d_p}$ is the concatenated fusion embedding v_i of the i-th token x_i.

3.2.3. BiLSTM Layer

Afterwards, for acquiring more comprehensive context features of entities, we further employ a BiLSTM layer to encode the fusion embeddings, which can make the most of both past and future input features. Following Huang et al. [16], in our proposed BBCPR model, the operation inner of an LSTM unit at step t can be expressed as below:

$$i_t = \sigma(W_i[h_{t-1}, v_t] + b_i), \tag{6}$$

$$f_t = \sigma\Big(W_f[h_{t-1}, v_t] + b_f\Big), \tag{7}$$

$$\widetilde{C}_t = \tanh(W_c[h_{t-1}, v_t] + b_c), \tag{8}$$

$$o_t = \sigma(W_o[h_{t-1}, v_t] + b_o), \tag{9}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t, \tag{10}$$

$$h_t = o_t \odot \tanh(C_t), \tag{11}$$

where o_t, f_t, and i_t represent the output, forget as well as input gate. v_t and h_t represent the input vector and the hidden state at step t. σ means the Sigmoid function and $\odot$ denotes dot product function. W_i, W_f, W_c, and W_o denote the trainable network weight parameters of the input $[h_{t-1}, v_t]$. In addition, b_i, b_f, b_c, and b_o represent the deviation parameters. C_t and $\widetilde{C}_t$ stand for the cell state and candidate cell state at t step, respectively. We formulate the computation of BiLSTM as follows:

$$\overrightarrow{h_t} = \mathrm{LSTM}\Big(v_t, \overrightarrow{h_{t-1}}\Big), \tag{12}$$

$$\overleftarrow{h_t} = \mathrm{LSTM}\Big(v_t, \overleftarrow{h_{t+1}}\Big), \tag{13}$$

$$h_t = \Big[\overrightarrow{h_t}; \overleftarrow{h_t}\Big], \tag{14}$$

where $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ denote the hidden state at step t of the forward LSTM and the backward LSTM, respectively. After that, the hidden representation of the input Chinese medical text X produced by BiLSTM module can be denoted as $H = \{h_t\}_{t=1}^{n} \in \mathbb{R}^{n \times 2d_{LSTM}}$, where d_{LSTM} is the hidden size of the LSTM.

3.2.4. CRF Layer

For a typical NER task, the relationship between adjacent labels is sequential and should also follow some constraint rules. For example, the label I-Symptom must appear after the label B-symptom. Due to the fact that BiLSTM only focuses on the long-term contextual features rather than the dependency between labels, a CRF layer is the preferred choice for decoding the ultimate sequence labels in the current research [4,15,16,20], as it can model the sequential relationships between labels by learning the adjacent constraint.

We first convert the output of the BiLSTM H to the input matrix P of the CRF with a linear function as follows:

$$P = W_p H + b_p, \tag{15}$$

where $W_p \in \mathbb{R}^{k \times 2d_{LSTM}}$, $b_p \in \mathbb{R}^k$ are learnable parameters, and k is the label types number. The CRF module is subsequently deployed to count the conditional probability $P(Y \mid X)$ of the random label sequence $Y = \{y_1, y_2, \ldots, y_n\}$ under the situation of a given Chinese medical text $X = \{x_1, x_2, \ldots, x_n\}$. In form, the probability $P(Y \mid X)$ of the ultimate optimal label sequence can be counted as:

$$p(Y \mid X) = \frac{e^{s(X,Y)}}{\sum_{\tilde{y} \in Y_X} e^{s(X,\tilde{y})}}, \tag{16}$$

$$s(X, Y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i}, \tag{17}$$

in which $\tilde{y}$ represents the basic-truth label sequence. Y_X denotes all the probable label sequences. A_{y_i,y_i+1} stands for the transition possibility from the label y_i to the label y_{i+1}, with the transition probability matrix A being a learnable model parameter. P_{i,y_j} indicates the non-normalized probability that the i-th token will be mapped to the named entity label y_i.

During the training process, the below loss of negative log-likelihood (NLL) is minimized for the optimization of the model as:

$$\mathcal{L}_{NLL} = -\log p(Y \mid X). \tag{18}$$

Moreover, to predict the labels of X, the Viterbi algorithm [48] is applied for decoding the overall optimal label sequence. The output label sequence Y^* containing the highest score will be produced as:

$$Y^* = \underset{\tilde{y} \in y_X}{\operatorname{argmax}}\, s(X, \tilde{y}). \tag{19}$$

3.3. Regularization Method

As a result of the restrictions of high specialization degree, ethics, and privacy, the annotated Chinese-MNER data are hard to gather and generally small in scale. The models are more prone to over-fitting problems. The dropout method has been used in most of the research works to alleviate the over-fitting problem. Dropout is a perturbation addition in essence [37]. In addition, Goodfellow et al. [44] propose an adversarial training strategy to increase the diversity of samples by adding noise perturbation and apply it into the field of computer vision. On the basis of the previous studies, Miyato et al. [49] extend the adversarial training to the text classification task. However, most of the existing research works only consider the addition of a single perturbation [42,43].

Focusing on the Chinese-MNER task, we expect to improve the performance of our proposal in identifying named entities by increasing the diversity of perturbations. As a result, we put forward a new regularization mechanism to regularize two distributions of the same sample, i.e., the original distribution and the distribution intervened by the adversarial perturbation and the dropout perturbation. We name such regularization mechanism as READ, that is, REgularization method with Adversarial training and Dropout. Specially, the READ mechanism regularizes the model predictions from two sub-models produced by the dropout and the adversarial perturbation. Unlike the previous works that consider only a single perturbation, READ amplifies the variability of the same sample by combining different perturbations, thus enhancing the robustness of the model.

The architecture of READ is shown in Figure 2. For the input $X = \{x_1, x_2, \ldots, x_n\}$ and the output sequence $Y = \{y_1, y_2, \ldots, y_n\}$, we originally put X into the pre-trained language model MC-BERT to acquire the embeddings E. In READ, we alternatively apply an adversarial perturbation on the original embeddings E to generate the adversarial embeddings E' as follows:

$$E' = E + \delta, \tag{20}$$

where δ is an adversarial perturbation produced by FGM [49]. FGM employs L_2-norm to scale the specific gradients to achieve better perturbation, which is calculated as:

$$\delta = \epsilon \frac{g}{\|g\|_2}, \tag{21}$$

$$g = \nabla_E \mathcal{L}_{NER}(X, \theta), \tag{22}$$

where ϵ is a constant that presents perturbation degree and g denotes the gradients of loss.

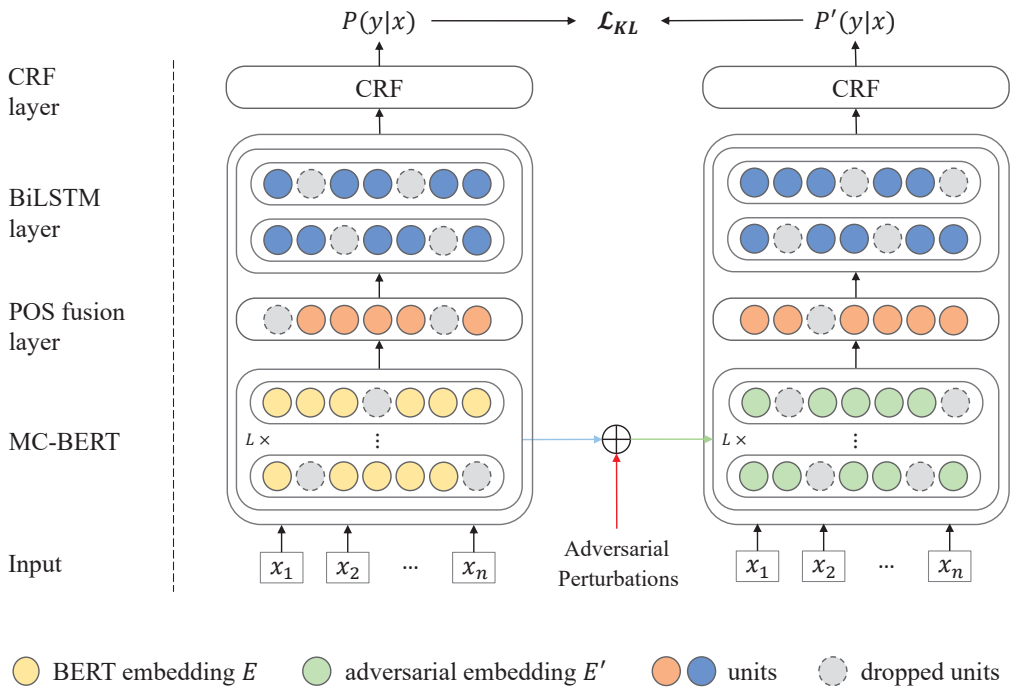


Figure 2. The architecture of the READ.

As shown in Figure 2, after randomly applying different dropout into the neural network, we can obtain two different sub-models for training. Then, we feed separately the original BERT embeddings E and adversarial embeddings E' into the above two sub-models to produce two different distributions for the output label sequence, that is, $P(Y \mid X)$ and $P'(Y \mid X)$, which generate two losses. So, we take the average of the two losses as the basic NER learning object $\mathcal{L}_{NER}$:

$$\mathcal{L}_{NER} = -\frac{1}{2}\left(\log \mathcal{P}(Y \mid X) + \log \mathcal{P}'(Y \mid X)\right). \tag{23}$$

The adversarial perturbation and dropout noise will shift the representation away from the one of original input. In the training step, READ focuses on reducing the difference between the predictions of the two sub-models by minimizing the bidirectional KL-divergence between these two output distributions of the same sample. Formally, we denote this process as follows:

$$\mathcal{L}_R = \frac{1}{2}\left(\mathcal{D}_{KL}\left(\mathcal{P}(Y \mid X) \| \mathcal{P}'(Y \mid X)\right) + \mathcal{D}_{KL}\left(\mathcal{P}'(Y \mid X) \| \mathcal{P}(Y \mid X)\right)\right), \tag{24}$$

where $\mathcal{L}_R$ denotes the loss function of the regularization method. $\mathcal{D}_{KL}(\mathcal{P} \| \mathcal{P}')$ denotes the KL divergence between two distributions $\mathcal{P}$ and $\mathcal{P}'$.

The ultimate training goal is to minimize the joint loss $\mathcal{L}$ for data (X, Y), which is calculated as:

$$\mathcal{L} = \mathcal{L}_{NER} + \lambda \mathcal{L}_R, \tag{25}$$

in which λ is the trade-off parameter for balancing the $\mathcal{L}_{NER}$ and $\mathcal{L}_R$. We further provide the pseudo-codes to detail the major steps of READ in Algorithm 1.

Algorithm 1: Pseudo Codes of READ.

Input: Dataset $\mathbb{D}$, model f, trade-off parameter λ
Output: model parameters θ
for *sampled mini-batch* $(X, Y) \in \mathbb{D}$ **do**
  Obtain the original embedding E of X;
  Generate adversarial embedding $E' = E + \delta$;
  Generate model $f^{'}$ by applying different dropout on model f;
  Feed embedding E to model f, obtain the distribution $\mathcal{P}(Y \mid X) = f_\theta(E)$;
  Feed embedding E' to model $f^{'}$, obtain the distribution $\mathcal{P}'(Y \mid X) = f_\theta^{'}(E')$;
  Calculate negative log-likelihood loss $\mathcal{L}_{NER}$ by Equation (23);
  Calculate regularization loss $\mathcal{L}_R$ by Equation (24);
  Update parameters θ to minimize total loss $\mathcal{L}$ of Equation (25).
end

4. Experiments

4.1. Datasets and Evaluation Metrics

For confirming the effectiveness of our proposal, we evaluate the model performances on two public datasets, that is, cMedQANER (https://github.com/alibaba-research/ChineseBLUE/tree/master/data/cMedQANER) together with cEHRNER (https://github.com/alibaba-research/ChineseBLUE/tree/master/data/cEHRNER) released by the ChineseBLUE benchmark [27]. The cMedQANER and cEHRNER datasets are annotated from the Chinese community question answering and the Chinese electronic health records, respectively. In detail, cMedQANER contains 2063 annotated instances altogether with eleven kinds of the medical named entities, such as *Crowd*, *Body*, etc. In addition, cEHRNER contains 999 annotated samples altogether with seven kinds of the medical named entities, for instance *Operation*, *Diagnosis*, *Disease*, and so on. The statistics of cMedQANER and cEHRNER are shown in Table 3. We also list the specific statistics of various kinds of the entities in cMedQANER and cEHRNER in Table 4 together with Table 5, respectively. The limited amount of annotated instances and the complex types of medical named entities make it challenging for Chinese-MNER to achieve a promising performance.

To measure the model performances for Chinese-MNER, we employ the scores of F1, precision (P), and recall (R) for evaluation, which are widely used metrics for sequence labeling tasks [34].

Table 3. The statistics for the cMedQANER together with the cEHRNER dataset.

Dataset	Training Set	Dev Set	Test Set
cMedQANER	1673	175	215
cEHRNER	914	44	41

Table 4. The statistics of various kinds of the medical named entities in cMedQANER.

Type	Training Set	Dev Set	Test Set
Body	2443	203	234
Crowd	735	48	78
Department	146	13	8
Disease	3890	332	431
Drug	541	44	61
Feature	311	27	28
Physiology	384	32	45
Symptom	2277	130	229
Test	485	70	49
Time	212	18	32
Treatment	1066	107	145
Total	**12,490**	**1024**	**1340**

Table 5. The statistics of different types of medical named entities in cEHRNER.

Type	Training Set	Dev Set	Test Set
Disease and diagnosis	3824	173	149
Operation	946	52	43
Anatomical part	5623	252	220
Drug	1646	84	72
Symptom	2095	78	88
Imaging examination	889	55	29
Laboratory test	1113	55	31
Total	**16,136**	**749**	**632**

4.2. Model Summary

We examine the effectiveness of BBCPR by comparing it against several competitive NER methods. Note that word2vec embeddings are employed for the BiLSTM and the BiLSTM-CRF model in this work. The baseline models are summarized as follows:

- **HMM** [50]: A traditional linear statistical model that is proposed to solve the sequence labeling problem.
- **BiLSTM** [16]: A BiLSTM model that employs a gated memory cell to capture long-range and bi-directional semantic dependencies within the sequence information.
- **BiLSTM-CRF** [16]: It extends the BiLSTM model by combining the BiLSTM network and a CRF module to decode the final sequence labels.
- **MC-BERT** [27]: A variant of BERT that is further trained on the Chinese medical corpus, which performs well in extracting the medical contextual features.
- **MC-BERT-CRF** [15]: A recent NER model that combines BERT with CRF. For a fair comparison, we replace BERT with the same MC-BERT used in our method.
- **MC-BERT-BiLSTM-CRF** [20]: It extends the MC-BERT-CRF model by applying a BiLSTM module to capture more comprehensive contextual features.

4.3. Research Questions

We comprehensively examine the effectiveness of our proposed BBCPR model by focusing on the following research questions:

(**RQ1**) Can BBCPR achieve better performance than the competitive baselines for the Chinese-MNER task? (See Section 5.1)

(**RQ2**) How about the contribution of the POS fusion layer and the READ module in BBCPR? (See Section 5.2)

(**RQ3**) What is the influence of different pre-trained models on the performance of BBCPR? (See Section 5.3)

(**RQ4**) How about the performance of BBCPR under different hyperparameters, i.e., POS embedding size, trade-off parameter λ? (See Section 5.4)

4.4. Experimental Details

In this paper, all the experiments are conducted in Python with the deep learning toolkit PyTorch (https://pytorch.org), where we run each experiment on both cMedQANER and cEHRNER datasets for five times under random seeds and then report the average results as well as the standard deviation. We employ the Baidu LAC toolkit [47] to obtain the POS tagging features for cMedQANER and cEHRNER datasets. In addition, the pre-trained language model MC-BERT [27] is specially selected as the contextual embedding layer for our model, with 12 layers, 768-dimensional of the hidden size, as well as 12 self-attention heads. Set the BiLSTM hidden size and feed-forward network dimension as 256 and 1024, respectively. The POS embedding is randomly initialized from the standard normal distribution, where the size of POS embedding is 512. The model is trained under a mini-batch strategy, and the maximum sequence length and the batch size are 256 and 32, respectively. Following Li et al. [15] and Xu et al. [20], we employ the Adam optimizer [51]

where β_1 is 0.9 and β_2 is 0.998. For the BiLSTM and MC-BERT, their learning rates are 7×10^{-5}. The learning rate of CRF is set to 5×10^{-3}. During training, we adopt a linear decay schedule to vary the learning rate with the weight decay being 0.01. The dropout rate and regularization loss weight λ are set to 0.2 and 2.0, respectively. Our model is trained for 50 epochs at most, and it stops on the optimal model for testing.

5. Results and Discussion

In this section, we first discuss the comparison of overall performance between BBCPR and the competitive baseline models (see Section 5.1). Next, we analyze the effectiveness of each component we propose in BBCPR (see Section 5.2). Furthermore, we explore the influence brought by different pre-trained language models (see Section 5.3) and different hyper-parameters (see Section 5.4). Finally, we present a case study to clearly demonstrate the superiority of BBCPR (See Section 5.5).

5.1. Overall Evaluation

To answer **RQ1** , we check the overall entity recognition performance of the baselines and our proposal in terms of all evaluation metrics on the cMedQANER and cEHRNER datasets. Table 6 shows the detailed outcomes between the models discussed.

First, we focus on the property of the baselines. In accordance with Table 6, compared to the traditional statistical learning model, i.e., HMM, deep learning based-models report obvious improvements in terms of all metrics. For example, BiLSTM-CRF beats HMM by 23.95% and 44.50% for the score of F1 on the cMedQANER and cEHRNER datasets, respectively, which demonstrates that the capability of learning context features is essential for the Chinese-MNER task. Specially, we can obverse that those models using MC-BERT as an encoder show nearly 7.08–8.87% and 4.59–5.31% improvements for the metric F1 against BiLSTM-CRF on the cMedQANER and cEHRNER dataset, respectively, indicating an impressive superiority of BERT in representation learning. In addition, equipping the model with additional CRF layer yields better performance than the original ones, which may be attributed to the CRF being able to well model the dependencies between tag sequences.

Next, we focus on the comparison between the baselines and our proposal. As revealed in Table 6, BBCPR achieves the best performance among all discussed models on both the cMedQANER and cEHRNER datasets. Specially, it can be found that our approach achieves the SOTA performance with the improvements of 2.44% and 2.87% in terms of F1 score against the best baseline MC-BERT-BiLSTM-CRF on two datasets, respectively. There is a similar phenomenon in terms of P and R. For the metric P, our proposal model beats the best baseline by 2.50% and 4.05% on the cMedQANER and cEHRNER dataset, respectively. For the metric R, our method shows 2.36% and 1.68% improvements over the best baseline on the cMedQANER and cEHRNER dataset, respectively. The improvements acquired from BBCPR can be explained by the fact that using POS tagging features, which imply potential word segmentation, can provide an extra supervision signal to distinguish the edges between ordinary words and the medical entities. In addition, the adversarial samples generated by FGM can enhance the model's robustness.

Table 6. Model performance comparison on the cMedQANER and cEHRNER datasets. The subscript indicates the standard deviation. The results generated by the best performer in each column are boldfaced.

Model	cMedQANER			cEHRNER		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
HMM	$61.35_{0.41}$	$63.13_{0.35}$	$61.91_{0.30}$	$51.92_{0.54}$	$56.96_{0.46}$	$53.73_{0.30}$
BiLSTM	$62.19_{0.71}$	$66.25_{0.51}$	$64.00_{0.29}$	$62.58_{0.46}$	$66.04_{0.81}$	$64.26_{0.28}$
BiLSTM-CRF	$79.91_{0.72}$	$73.94_{0.66}$	$76.74_{0.25}$	$79.27_{0.53}$	$76.10_{0.79}$	$77.64_{0.23}$
MC-BERT	$80.15_{0.55}$	$83.02_{0.37}$	$81.56_{0.31}$	$78.32_{0.77}$	$83.22_{0.60}$	$80.69_{0.24}$
MC-BERT-CRF	$81.16_{0.79}$	$83.23_{0.56}$	$82.17_{0.23}$	$78.98_{0.68}$	$83.55_{0.38}$	$81.20_{0.24}$
MC-BERT-BiLSTM-CRF	$83.21_{0.37}$	$83.91_{0.45}$	$83.55_{0.22}$	$80.04_{0.49}$	$83.57_{0.66}$	$81.76_{0.27}$
BBCPR	$\mathbf{85.29_{0.28}}$	$\mathbf{85.89_{0.55}}$	$\mathbf{85.59_{0.23}}$	$\mathbf{83.28_{0.52}}$	$\mathbf{84.97_{0.61}}$	$\mathbf{84.11_{0.24}}$

5.2. Ablation Study

To answer **RQ2** , we conduct comprehensive ablation studies on cMedQANER and cEHRNER datasets to verify the effectiveness of each key module of BBCPR. The detailed results of ablation studies are presented in Table 7. It can be observed that when any certain module is taken out, the model performances decrease obviously in terms of almost all metrics, which verifies the effectiveness of our proposed modules in BBCPR. Specifically, for the metric F1 and R, removing the POS fusion layer and the DP mechanism (see Row 6, Table 7) results in the biggest drop of model performance. Specifically, on both of the cMedQANER and cEHRNER datasets, model performances show a 1.54% and 2.11% decrease in terms of F1 and a 1.68% and 1.36% decrease in terms of R. This observation indicates that incorporating POS tagging features into the neural networks can evidently enhance the superiority of the Chinese-MNER model, for which POS tagging features add extra potential entities' boundary information. At the same time, the DP mechanism can help alleviate the over-fitting problem of the model and thus reduces the prediction error. As for the metric P, the removal of the READ (see Row 2, Table 7) declines the model performance most, with a 1.55% and 3.29% decrease on the cMedQANER and cEHRNER dataset, respectively. This can be due to the fact that the READ module amplifies the variability of the original samples, thus enhancing the learning ability of the model.

In addition, we can observe that removing AP (see Row 4, Table 7) or removing DP (see Row 3, Table 7) leads to a 0.70% and 0.82% decrease for the score of F1 on the cMedQANER dataset, respectively. Similar results can be found on the cEHRNER dataset, where the model reveals a 0.58% and 0.96% drop. The reason why DP performs better than AP may be that it is applied on the whole model, while AP only works on the BERT-embedding layer. In addition, it is worth mentioning that using either the AP or DP alone is not as effective as the combination of the two, i.e., the READ. For example, without the READ (see Row 2, Table 7), the decreases in F1 score are 1.29% and 1.82% on two datasets, respectively. The reason may be that each of the perturbations is relatively simple; employing only one mechanism can only introduce a small perturbation. Meanwhile, diverse perturbations can increase the dissimilarity of the representation from the same sample.

Moreover, in the condition of using DP by default (See Row 7, Table 7), adding AP (see Row 5, Table 7) results in a 0.32% and 0.70% increase in terms of F1 on the cMedQANER and cEHRNER datasets, respectively. Similarly, when adding POS (see Row 4, Table 7), the score of F1 increases by 0.41% and 0.90%, respectively. Compared with AP, POS has a greater impact on the neural network that has a dropout by default. We attribute this phenomenon to the fact that POS can directly increase the features of entities and thus bring useful information to the model. However, AP utilizes an indirect way to enhance the learning ability of the model by adding perturbations. Under the condition that uses

DP by default, the F1 score drops by 1.10% and 1.48% on two datasets when removing POS and AP, which indicates that the diversity added by both POS and AP has a positive impact on the model performance.

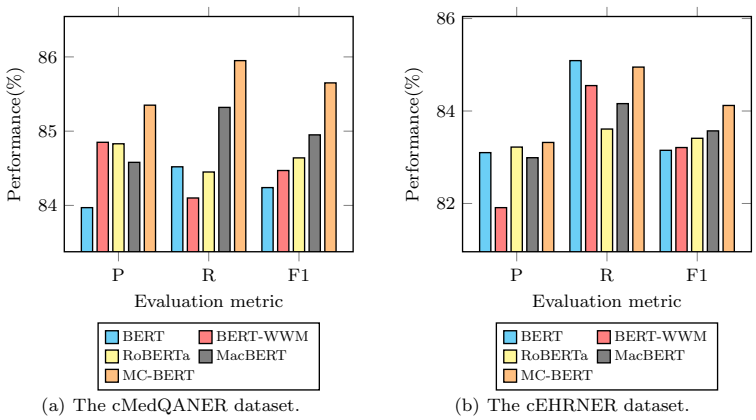
Table 7. Performances in terms of P, R, and F1 without different modules on the cMedQANER and cEHRNER datasets. AP and DP denote adversarial perturbation and dropout perturbation in READ, respectively. POS denotes the POS fusion layer. The subscript indicates the standard deviation.

POS	AP	DP	cMedQANER			cEHRNER		
			P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
✓	✓	✓	$85.29_{0.28}$	$85.89_{0.55}$	$85.59_{0.23}$	$83.28_{0.52}$	$84.97_{0.61}$	$84.11_{0.24}$
✓			$84.00_{0.34}$	$85.02_{0.40}$	$84.51_{0.23}$	$80.64_{0.79}$	$84.71_{0.91}$	$82.62_{0.24}$
✓	✓		$84.63_{0.80}$	$85.19_{0.54}$	$84.91_{0.22}$	$81.79_{0.48}$	$84.92_{0.18}$	$83.32_{0.19}$
✓		✓	$85.38_{0.59}$	$84.64_{0.27}$	$85.01_{0.18}$	$81.95_{0.51}$	$85.39_{0.66}$	$83.63_{0.17}$
	✓	✓	$85.07_{0.63}$	$84.79_{0.49}$	$84.93_{0.17}$	$82.35_{0.66}$	$84.61_{0.70}$	$83.46_{0.16}$
	✓		$84.12_{0.64}$	$84.49_{0.58}$	$84.30_{0.21}$	$81.01_{0.88}$	$83.83_{0.61}$	$82.39_{0.22}$
		✓	$84.60_{0.20}$	$84.72_{0.30}$	$84.66_{0.18}$	$81.24_{0.26}$	$84.59_{0.37}$	$82.88_{0.18}$

5.3. Influence of Pre-Trained Language Models

To answer **RQ3,** we further conduct comparative experiments to analyze the effect brought by different pre-trained language models, such as BERT, BERT-WWM, RoBERTa, MacBERT, and MC-BERT. First, **BERT** is able to obtain more informative contextual representation through employing the masked language model and next sentence to forecast training targets. Furthermore, **BERT-WWM** employs a whole word masking strategy for the Chinese corpus. For the **RoBERTa**, it robustly optimizes the BERT pre-training method by four simple and effective modifications. **MacBERT** masks the words with their similar words in the Chinese corpus, while **MC-BERT** is further trained on the Chinese biomedical corpus based on BERT.

As shown in Figure 3, MC-BERT significantly outperforms other pre-trained models in terms of all evaluation metrics. In particular, as reflected in Figure 3a, it can be observed that MC-BERT shows nearly 0.83–1.67%, 0.59–1.65% and 0.74–2.21% improvements in terms of F1, P, and R score than other pre-trained language models on the cMedQANER dataset. Figure 3b indicates that MC-BERT increases nearly 0.66–1.17%, 0.12–2.48%, and −0.16–1.60% of F1, P, and R scores compared to other pre-trained language models on the cEHRNER dataset. The possible reason may be due to the fact that MC-BERT adapts the whole entity masking strategy and the whole span masking strategy to inject medical domain knowledge for Chinese biomedical text, which can help generate better contextual representation for the Chinese-MNER task. Accordingly, we choose the MC-BERT model [27] as the contextual embedding component in the following experiments.



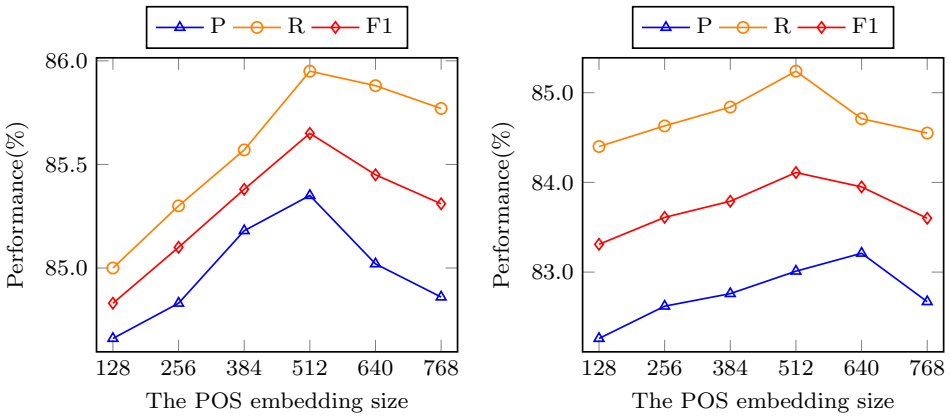
(a) The cMedQANER dataset. (b) The cEHRNER dataset.

Figure 3. The influence of the pre-trained language models: the cMedQANER together with the cEHRNER test set.

5.4. Analysis on Different Hyperparameters

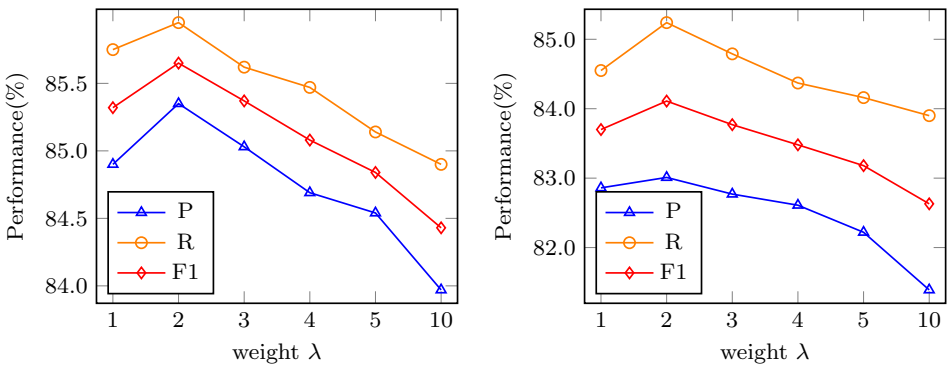
To answer **RQ4,** we conduct the experiments on cMedQANER and cEHRNER datasets to explore the BBCPR property under different hyperparameters, i.e., POS embedding size, and trade-off parameter λ. For the POS embedding, the size of it is set to 128, 256, 384, 512, 640, and 768 in our experiments, respectively. As displayed in Figure 4a, the property of our model initially grows when the size of the POS embedding increases and the difference between the best and worst performance can be as high as 0.97%, 0.82%, and 1.12% in the terms of F1, P, and R, respectively on the cMedQANER dataset. Figure 4b shows that there are similar outcomes on the cEHRNER dataset. This is probably explained through the fact that adding the neural network size can enhance the model complexity to obtain more powerful representation capability. However, when the size increases further, worse performance is achieved due to model over-fitting. As shown in Figure 4, when the POS embedding size is 512, all the metrics achieve their best performance. Therefore, we choose 512 as the POS embedding size in our experiments.

Further, we analyze the effect resulted from the trade-off parameter λ. We vary the λ in $\{1, 2, 3, 4, 5, 10\}$ and conduct extensive experiments. According to Figure 5a,b, either too large or too small λ will make our model perform poorly. When $\lambda = 2$, the model realizes the optimum property, and subsequently, our model performs worse and worse as λ increases. Specifically, the performances of the model on the cMedQANER dataset have drops by 0.33–1.45%, as shown in Figure 5a and 0.41–1.80% on the cEHRNER dataset, as shown in Figure 5b. In terms of P and R, there are floats of 0.15–1.64% and 0.24–1.24% on the cMedQANER dataset, while they are 0.19–1.99% and 0.53–1.60% on the cEHRNER dataset. When the λ is at 2, all the metrics achieve their best performance. Therefore, we select 2 as the regularization loss weight for our proposed BBCPR model.



(a) The cMedQANER dataset. (b) The cEHRNER dataset.

Figure 4. The influence of the POS embedding size against the cMedQANER and cEHRNER test sets.



(a) The cMedQANER dataset. (b) The cEHRNER dataset.

Figure 5. The effect of regularization loss weight λ on the cMedQANER and cEHRNER test sets.

5.5. Case Study

In the current section, we conduct a case study to demonstrate the superiority of our model against baselines. In particular, we compared the predictive results of our proposal and MC-BERT-BiLSTM-CRF on two cases from the cMedQANER and cEHRNER datasets, respectively. The detailed input medical text and the corresponding predictive results of the two cases are presented in Figure 6.

As shown in Figure 6, in Case #1, the baseline model fails to identify "药物治疗 (*drug treatment*)" as a whole unit. However, our model could completely and correctly recognize the entity "药物治疗 (*drug treatment*)" as the four tokens have the same part-of-speech (i.e., noun). Likewise, in Case #2, we can see that the baseline MC-BERT-BiLSTM-CRF identifies "升结肠恶性肿瘤及肝内外胆管结石 (*malignant tumor of ascending colon and intrahepatic and extrahepatic bile duct stones*)" as an independent entity. However, the token "及 (*and*)" is a conjunction, while "瘤 (*tumor*)" and "肝 (*liver*)" are nouns. On the contrary, our model could accurately recognize two entities "升结肠恶性肿瘤 (*malignant tumor of ascending colon*)" and "肝内外胆管结石 (*intrahepatic and extrahepatic bile duct stones*)" through identifying obvious entity boundaries before and after "及 (*and*)". Overall, the above cases demonstrate that BBCPR can explicitly learn word boundary information by introducing the POS tagging features, which is conducive to enhance the entity recognition accuracy.

Case # 1	Medical Text	**药物治疗**(treatment)是比较低廉的... **Drug treatment** (treatment) is relatively cheap ...
	Baseline	药物**治疗**(treatment)是比较低廉的...
	BBCPR	**药物治疗**(treatment)是比较低廉的...
Case # 2	Medical Text	患者2-月前因诊断为**升结肠恶性肿瘤**(disease and diagnosis)及**肝内外胆管结石**(disease and diagnosis)。 The patient was diagnosed with **malignant tumor of ascending colon** (disease and diagnosis) and **intrahepatic and extrahepatic bile duct stones** (disease and diagnosis) 2-months ago.
	Baseline	患者2-月前因诊断为**升结肠恶性肿瘤及肝内外胆管结石**(disease and diagnosis)。
	BBCPR	患者2-月前因诊断为**升结肠恶性肿瘤**(disease and diagnosis)及**肝内外胆管结石**(disease and diagnosis)。

Figure 6. Qualitative comparison of Baseline and BBCPR. Case #1 and Case #2 are from the cMedQANER and cEHRNER test set, respectively. The entities are marked in **bold**. The ground truth labels are in orange. The labels of baseline prediction are in red. The labels of BBCPR prediction are in blue.

6. Conclusions and Future Work

In our work, we propose a model named BBCPR for improving the performance of the Chinese-MNER task, which leverages a POS fusion layer to explicitly learn word boundary information by incorporating external syntax knowledge. What is more, we also design a novel regularization method READ to deal with the over-fitting problem and improve the model robustness. In detail, READ regularizes the predictions of the two sub-models through minimizing the bidirectional KL-divergence between the adversarial output and original output distributions for the same sample. Comprehensive experiments conducted on two benchmark datasets confirm the advantage of our proposal for the Chinese-MNER task. In addition, an ablation study proves that the POS fusion layer and READ can effectively improve the model performance.

For future research, we want to explore how to obtain more features for entities by introducing the contrastive learning [41], which can pull the same type of entities closer and push apart different types of entities [52]. Furthermore, we have interests in verifying

the effectiveness of our proposal in other domains, e.g., financial domain, legal domain. Finally, mining more potential supervisory signals from the unlabeled samples and then training the model in an unsupervised setting may also be a promising direction.

Author Contributions: Funding acquisition, T.S.; Project administration, H.C.; Validation, X.Z.; Writing—original draft, M.J.; Writing—review and editing, C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Postgraduate Scientific Research Innovation Project of Hunan Province under No. CX20200054.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Akkasi, A.; Varoğlu, E.; Dimililer, N. Balanced undersampling: A novel sentence-based undersampling method to improve recognition of named entities in chemical and biomedical text. *Appl. Intell.* **2018**, *48*, 1965–1978. [CrossRef]
2. Pan, J.; Zhang, C.; Wang, H.; Wu, Z. A comparative study of Chinese named entity recognition with different segment representations. *Appl. Intell.* **2022**. [CrossRef]
3. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 50–70. [CrossRef]
4. Magge, A.; Scotch, M.; Gonzalez-Hernandez, G. Clinical NER and relation extraction using bi-char-LSTMs and random forest classifiers. In Proceedings of the 1st International Workshop on Medication and Adverse Drug Event Detection, Brest, France, 3–5 May 2018; pp. 25–30.
5. Banerjee, P.S.; Chakraborty, B.; Tripathi, D.; Gupta, H.; Kumar, S.S. A information retrieval based on question and answering and NER for unstructured information without using SQL. *Wirel. Pers. Commun.* **2019**, *108*, 1909–1931. [CrossRef]
6. Savini, E.; Caragea, C. Intermediate-Task Transfer Learning with BERT for Sarcasm Detection. *Mathematics* **2022**, *10*, 844. [CrossRef]
7. Klein, D.; Smarr, J.; Nguyen, H.; Manning, C.D. Named entity recognition with character-level models. In Proceedings of the Seventh Conference on Natural Language Learning at NAACL, Edmonton, AB, Canada, 31 May–1 June 2003; pp. 180–183.
8. Chen, C.; Kong, F. Enhancing Entity Boundary Detection for Better Chinese Named Entity Recognition. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, ACL, Online, 1–6 August 2021; pp. 20–25.
9. Song, M.; Yu, H.; Han, W.S. Developing a hybrid dictionary-based bio-entity recognition technique. *BMC Med. Inform. Decis. Mak.* **2015**, *15*, 1–8. [CrossRef]
10. Wang, Q.; Zhou, Y.; Ruan, T.; Gao, D.; Xia, Y.; He, P. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. *J. Biomed. Inform.* **2019**, *92*, 103133. [CrossRef]
11. Zhang, Y.; Yang, J. Chinese NER using lattice LSTM. *arXiv* **2018**, arXiv:1805.02023.
12. Habibi, M.; Weber, L.; Neves, M.; Wiegandt, D.L.; Leser, U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **2017**, *33*, i37–i48. [CrossRef]
13. Li, Z.; Zhang, Q.; Liu, Y.; Feng, D.; Huang, Z. Recurrent neural networks with specialized word embedding for chinese clinical named entity recognition. *J. Biomed. Inform.* **2017**, *1976*, 55–60. [CrossRef]
14. Xu, G.; Wang, C.; He, X. Improving clinical named entity recognition with global neural attention. In Proceedings of the Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data, Guangzhou, China, 23–25 August 2018; pp. 264–279.
15. Li, X.; Zhang, H.; Zhou, X.H. Chinese clinical named entity recognition with variant neural structures based on BERT methods. *J. Biomed. Inform.* **2020**, *107*, 103422. [CrossRef] [PubMed]
16. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
17. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
18. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef] [PubMed]
19. Wu, Y.; Huang, J.; Xu, C.; Zheng, H.; Zhang, L.; Wan, J. Research on Named Entity Recognition of Electronic Medical Records Based on RoBERTa and Radical-Level Feature. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 2489754. [CrossRef]

20. Xu, L.; Li, S.; Wang, Y.; Xu, L. Named Entity Recognition of BERT-BiLSTM-CRF Combined with Self-attention. In Proceedings of the International Conference on Web Information Systems and Applications, Kaifeng, China, 24–26 September 2021; pp. 556–564.
21. Liu, W.; Zhou, P.; Zhao, Z.; Wang, Z.; Ju, Q.; Deng, H.; Wang, P. K-bert: Enabling language representation with knowledge graph. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 2901–2908.
22. Zhang, W.; Lin, H.; Han, X.; Sun, L. De-biasing Distantly Supervised Named Entity Recognition via Causal Intervention. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, ACL, Online, 1–6 August 2021; pp. 4803–4813.
23. Toutanova, K.; Klein, D.; Manning, C.D.; Singer, Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 North American Chapter of the Association for Computational Linguistics, NAACL, Edmonton, AB, Canada, 27 May–1 June 2003; pp. 252–259.
24. Gimpel, K.; Schneider, N.; O'Connor, B.; Das, D.; Mills, D.P.; Eisenstein, J.; Heilman, M.; Yogatama, D.; Flanigan, J.; Smith, N.A. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, OR, USA, 19–24 June 2011; pp. 42–47.
25. Cai, X.; Dong, S.; Hu, J. A deep learning model incorporating part of speech and self-matching attention for named entity recognition of Chinese electronic medical records. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 101–109. [CrossRef]
26. Tong, Y.; Chen, Y.; Shi, X. A multi-task approach for improving biomedical named entity recognition by incorporating multi-granularity information. In Proceedings of the Findings of the Association for Computational Linguistics, Online Event, 1–6 August 2021; pp. 4804–4813.
27. Zhang, N.; Jia, Q.; Yin, K.; Dong, L.; Gao, F.; Hua, N. Conceptualized representation learning for chinese biomedical text mining. *arXiv* **2020**, arXiv:2008.10813.
28. Wu, L.; Li, J.; Wang, Y.; Meng, Q.; Qin, T.; Chen, W.; Zhang, M.; Liu, T.-Y. R-drop: Regularized dropout for neural networks. *arXiv* **2021**, arXiv:2106.14448.
29. Chiu, J.P.; Nichols, E. Named entity recognition with bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 357–370. [CrossRef]
30. Ma, X.; Hovy, E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv* **2016**, arXiv:1603.01354
31. Papernot, N.; McDaniel, P.; Swami, A.; Harang, R. Crafting adversarial input sequences for recurrent neural networks. In Proceedings of the MILCOM IEEE Military Communications Conference, Baltimore, MD, USA, 1–3 November 2016; pp. 49–54.
32. Jie, Z.; Lu, W. Dependency-guided LSTM-CRF for named entity recognition. *arXiv* **2019**, arXiv:1909.10148.
33. Sachan, D.S.; Zaheer, M.; Salakhutdinov, R. Revisiting lstm networks for semi-supervised text classification via mixed objective function. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 6940–6948. [CrossRef]
34. Ma, X.; Hovy, E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 1064–1074.
35. Dong, C.; Zhang, J.; Zong, C.; Hattori, M.; Di, H. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. In *Natural Language Understanding and Intelligent Applications*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 239–250.
36. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
37. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
38. Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv* **2012**, arXiv:1207.0580.
39. Wan, L.; Zeiler, M.; Zhang, S.; Le Cun, Y.; Fergus, R. Regularization of neural networks using dropconnect. In Proceedings of the 30th International Conference on Machine Learning, PMLR, Atlanta, GA, USA, 16–21 June 2013; pp. 1058–1066.
40. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Dropblock: A regularization method for convolutional networks. *arXiv* **2018**, arXiv:1810.12890.
41. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *arXiv* **2021**, arXiv:2104.08821.
42. Liu, F.; Vulić, I.; Korhonen, A.; Collier, N. Fast, Effective and Self-Supervised: Transforming Masked LanguageModels into Universal Lexical and Sentence Encoders. *arXiv* **2021**, arXiv:2104.08027.
43. Zuo, C. Regularization effect of fast gradient sign method and its generalization. *arXiv* **2018**, arXiv:1810.11711.
44. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
45. Zhou, Y.; Zheng, X.; Huang, X. Chinese Named Entity Recognition Augmented with Lexicon Memory. *arXiv* **2019**, arXiv:1912.08282.
46. Liang, J.; Xian, X.; He, X.; Xu, M.; Dai, S.; Xin, J.Y.; Xu, J.; Yu, J.; Lei, J. A novel approach towards medical entity recognition in Chinese clinical text. *J. Healthc. Eng.* **2017**, *2017*, 4898963. [CrossRef] [PubMed]
47. Jiao, Z.; Sun, S.; Sun, K. Chinese Lexical Analysis with Deep Bi-GRU-CRF Network. *arXiv* **2018**, arXiv:1807.01882.
48. Huang, W.; Hu, D.; Deng, Z.; Nie, J. Named entity recognition for Chinese judgment documents based on BiLSTM and CRF. *EURASIP J. Image Video Process.* **2020**, *2020*, 52. [CrossRef]
49. Miyato, T.; Dai, A.M.; Goodfellow, I. Adversarial training methods for semi-supervised text classification. *arXiv* **2016**, arXiv:1605.07725.

50. Zhou, G.; Su, J. Named entity recognition using an HMM-based chunk tagger. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 473–480.

51. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
52. Jin, T.; Zhao, Z. Contrastive Disentangled Meta-Learning for Signer-Independent Sign Language Translation. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, China, 20–24 October 2021; pp. 5065–5073.