Expert Systems With Applications 231 (2023) 120709

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A dictionary-guided attention network for biomedical named entity recognition in Chinese electronic medical records

Zhichao Zhu [a], Jianqiang Li [a], Qing Zhao [a,*], Faheem Akhtar [b]

[a] Faculty of Information Technologyon, Beijing University Of Technology, Beijing, China
[b] Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan

ARTICLE INFO

Keywords:
Biomedical named entity recognition
Attention mechanism
Biomedical dictionary
Semisupervised method
Electronic medical record

ABSTRACT

Biomedical named entity recognition (BNER) is a critical task for biomedical information extraction. Most popular BNER approaches based on deep learning utilize words and characters as features to represent medical texts. However, many medical terminologies are composed of multiple words and characters, and splitting medical terminology into multiple words (or characters) and assigning weight values for each word (or character) by a standard attention mechanism may disperse the attention score and result in a lower weight value for the medical terminology. This paper proposes a Dictionary-guided Attention Network (DGAN) for BNER in Chinese electronic medical records (EMRs). First, the medical concepts are extracted as large-size words to supplement the comprehensive semantic information of the medical terminology by matching the EMR text to the biomedical dictionary. Then, based on the matched dictionary results, an optimized attention strategy is proposed to focus on the medical concept and adaptively assign higher weights to the characters contained in a concept. Furthermore, semisupervised learning is introduced to reduce the manual labeling of data and to handle the entities not defined in the medical dictionary. To validate our new model in recognizing biomedical named entities, we conduct comprehensive experiments on a real-world Chinese EMR dataset and the CCKS2017 dataset. Our promising results illustrate that our method not only achieves a state-of-the-art performance in BNER but also reduces manual data annotation.

1. Introduction

With the increased adoption of electronic medical record (EMR) systems, various clinical sources, including laboratory test results, medications and diagnostic information, are becoming available as a treasure trove for large-scale clinical data analysis (Li, Liu, Liu et al., 2015; Zhao, Kang, et al., 2018). Biomedical named entity recognition (BNER) is a fundamental step for extracting patient information from clinical text to support data-driven medical studies and health management systems (Zhao, Xu, et al., 2022; Zhao, Li, et al., 2022; Ma et al., 2022). The main goal of BNER is to automatically recognize the text spans that mention named entities from unstructured clinical records and classify them into predefined categories, i.e., diseases, symptoms, tests and treatments (Li, Sun, et al., 2018).

There have been two main ways to construct named entity recognition (NER) models, knowledge-driven and data-driven methods. The former relies on handcrafted rules and human-constructed domain knowledge graphs. Although these methods achieve a high precision, they might create high system engineering costs and a low recall rate. Therefore, existing NER studies adopt rule-based dictionary matching and feature-based supervised learning methods (Li, Sun, et al., 2018). The former relies on handcrafted rules based on syntactic-lexical patterns (Zhang & Elhadad, 2013). The latter excessively depends on sophisticated feature engineering, i.e., a set of valuable features are carefully designed by various natural language processing (NLP) tools from annotated data to represent training examples. Later, traditional machine learning algorithms, such as CRF-based models, were used to identify similar patterns from unseen data and are labor-intensive and skill-dependent tasks. Deep learning has received much attention in the machine learning community. It is composed of multiple layers of nonlinear units to automatically learn the data representations and reduce the workload of feature selection (LeCun et al., 2015). The bidirectional long short-term memory network, combining a conditional random field (BILSTM-CRF) among the deep neural networks, exhibits promising results in many NER tasks. Many current NER methods use the self-training BERT (Bidirectional Encoder Representation from

* Corresponding author.
E-mail addresses: Zhuzc@emails.bjut.edu.cn (Z. Zhu), lijianqiang@bjut.edu.cn (J. Li), zhaoqing@bjut.edu.cn (Q. Zhao), fahim.akhtar@iba-suk.edu.pk (F. Akhtar).

https://doi.org/10.1016/j.eswa.2023.120709
Received 15 September 2022; Received in revised form 9 April 2023; Accepted 3 June 2023
Available online 9 June 2023

Table 1
The segmentation method in different granularity.

Original sentence s in EMR text: 患者诊断为慢性阻塞性肺疾病 (The patient was diagnosed with chronic obstructive pulmonary disease).
s^c = ["患" (suffer), "者" (patient), "诊" (diagnosis), "断" (judge), "为" (as), *"慢" (slow), "性" (property), "阻" (block), "塞" (block), "性" (property), "肺" (lung), "疾" (illness), "病" (disease)*]**Granularity of features:** character-level
s^w = ["患者" (patient), "诊断" (diagnosed), "为" (as), *"慢性" (chronic), "阻塞性" (obstructive), "肺" (lung), "疾病" (disease)*]**Granularity of features:** word-level
s^e = ["患者" (patient), "诊断" (diagnosed), "为" (as), *"慢性阻塞性肺疾病" (chronic obstructive pulmonary disease)*]**Granularity of features:** concept-level

Transformers) model (Devlin et al., 2018) to encode the text. Unlike OpenAI GPT (Radford et al., 2018), every token can only fuse the previous tokens in the self-attention layer of the transformer, and BERT enables the combination of both right and left contexts. Furthermore, it uses "next sentence prediction (NSP)" to jointly pretrain representations of text pairs. BERT exhibits a state-of-the-art performance in many downstream tasks, such as NER and question answering. EMR text contains many short sentences, which lack fully synthetic and semantic indications of the linguistic composition. Considering that two sentences may not capture enough semantic information to learn a token representation, we adopt an improved model, RoBERTa (Robustly Optimized BERT Pretraining Approach) (Liu, Ott, et al., 2019), which removes NSP, and each input includes the full sentence that is sampled contiguously from one or two texts.

Most of the existing NER methods based on BILSTM-CRF use words (or characters) to represent text, and adopt standard attention to assign higher weights for important words (or characters) that help with entity inference (Luo et al., 2018; Li, Zhao, et al., 2018; Kim et al., 2017). Two challenges remain. First, words and characters cannot provide sufficient entity-related semantic information for BNER because medical terminology usually consists of multiple words or characters. Table 1 shows an example of the word-, character- and concept-level embedding methods. From Table 1, we can see that the medical entity "慢性阻塞性肺疾病" (chronic obstructive pulmonary disease) in sentence s can be divided into four words and eight characters. Existing NER methods first learn the semantic information of words (or characters) from their context; then, the entity can be inferred from the meaning of these words (or characters). However, this inference process might cause semantic ambiguities since EMR text contains many short sentences that lack full synthetic and semantic indications on the linguistic composition. Second, splitting medical terminology into multiple words (or characters) and assigning weight values for each word (or character) by standard attention may disperse the attention score and result in a lower weight value for medical terminology. Note that the word and concept are regarded as the word token in this paper.

To address the above problem, the dictionary-guided attention network (DGAN) is proposed for building the BNER model with Chinese EMRs, and is explained as follows. Initially, the concept is extracted as a large-size word to obtain the overall semantic information of the medical entities. To avoid the mistakes of entity boundary recognition caused by word segmentation, we only leverage the concept and character as inputs of our model and pretrain RoBERTa on a large number of sentences from Chinese EMRs to elaborate the concept and character embeddings. Later, based on dictionary matching results, an adaptive attention strategy is presented on the top of the BILSTM layer to focus on the medical concept and to assign a higher attention weight to the characters contained in a concept. Then, semisupervised learning is introduced to reduce the manually labeled data and to handle entities not defined in the medical dictionary. Finally, a BILSTM-CRF model is used to learn the context features and decode the predicted tags.

The main contributions of this paper can be summarized as follows: 1). The DGAN is proposed as a novel approach for building a BNER model with Chinese EMRs, where the encoded feature's discriminant capability is improved by extracting the concept in EMRs as the large-size word. 2). We present an adaptive attention strategy to focus on the medical entity and assign a higher attention weight to the characters contained in a concept. 3). A semisupervised learning process is introduced to reduce manual data labeling and to handle entities not defined in the medical dictionary. 4). In empirical experiments on two Chinese EMR datasets, the results show that DGAN outperforms all the baselines, which demonstrates the effectiveness and generalization of the proposed model.

The rest of this paper is organized as follows. Section 2 presents related work in named entity recognition, with a specific focus on constructing text representations of the multigranularity semantic features and capturing important semantic information by an attention mechanism. In Section 3, we discuss the problem definition. The training mode and the overall architecture of the proposed DGAN model are described in Section 4. Section 5 introduces our Chinese EMR dataset and the evaluation metrics of the model. Section 6 discusses comparisons with other baselines and the evaluation results. Finally, we conclude this study in Section 7 and Section 8.

2. Related work

The first NER task was held at the Sixth Message Understanding Conference (MUC-6) in 1996 (Grishman & Sundheim, 1996). Since then, many efforts have been devoted to this topic. Early NER approaches can be divided into two categories, i.e., the rule-based approaches and feature-based supervised learning approaches. The former does not need labeled data, as they rely on hand-crafted rules. These rules are designed based on domain-specific gazetteers or syntactic-lexical patterns. (Quimbaya et al., 2016) proposed a dictionary matching method to recognize named entities in electronic health records. The experimental results show that this method improves the recall rate and has a limited impact on precision. Due to the complexity of domain-specific rules and incomplete expert knowledge, high precision and low recall rates often exist in these approaches. Feature engineering is significant in supervised NER approaches. These features are carefully selected from a fully labeled dataset to represent training examples. Traditional machine learning algorithms, such as the hidden markov model (HMM), support vector machine (SVM) (Li, Liu, et al., 2020) and conditional random fields (CRF), are used to learn a model to identify similar patterns from unseen data. (Liu, Hu, et al., 2017) adopted multiple features containing word-clustering features, bag-of-characters (BOC), part-of-speech (POS) and dictionary, and the CRF model to recognize clinical entities from Chinese EMRs. (Liang et al., 2017) developed a cascade-type system incorporating a CRF-based clinical entity classifier and a SVM-based sentence classifier to identify drug names from 324 Chinese admission notes. Feature-based supervised learning approaches can perform well on NER tasks, however, they rely heavily on hand-engineered features and expert knowledge.

Deep learning is a field of machine learning; its key advantage is the representation learning and semantic composition capabilities, which enables vector representation and neural processing. It allows the machine to learn useful representations from raw data automatically. Therefore, deep learning-based methods can achieve state-of-the-art performances with very little feature engineering. Recurrent neural networks (RNNs) (Goodfellow et al., 2016) are based on simple RNNs, and long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997), bidirectional long short-term memory (BILSTM), gated recurrent units (GRUs) (Cho et al., 2014) and convolutional neural networks (CNNs) are the most leveraged networks in the NER field.

As an important extension of NER in the medical field, BNER frequently adopts the deep neural networks mentioned above to extract medical-related entities. (Chokwijitkul et al., 2018) compared the ability of CNN, RNN, LSTM, BILSTM and GRU models to recognize heart risk factors in EMRs and found that BILSTM achieved the best results. In addition, (Wu et al., 2017) evaluated different models such as a CRF, a CNN and a BILSTM for NER by leveraging the medical dataset of the

Table 2
The entity labeling symbols and examples of each category.

Medical entity category	Tag symbol
1. 症状 (symptom, S)	B-S/I-S
2. 检查 (Check, C)	B-C/I-C
3. 检查结果 (check result, CR)	B-CR/I-CR
4. 诊断或疾病(diagnosis or disease, DD)	B-DD/I-DD
5. 治疗或药物 (treatment or medicine, TM)	B-TM/I-TM
6. Nonentity	O

2010 i2b2 NLP challenge. The conclusion is similar to (Chokwijitkul et al., 2018), BILSTM outperformed all the other models. Many other methods extended a BILSTM model with a character embedding layer and a CRF layer. (Xu, Zhou, et al., 2018) leveraged the NCBI Disease Corpus to evaluate their proposed method, while (Unanue et al., 2017) leveraged three different medical datasets to evaluate their proposed model. The experimental results both showed that the CRF layer and the character layer could effectively improve the model's performance. This was already expected because the CRF layer can capture the relations among labels (e.g., label B-test cannot follow label I-disease in a sentence), and the character layer supplements more fine-grained semantic information for the model training process, further enhancing the semantic discrimination ability of the embeddings. Moreover, considering that the above methods do not distinguish the importance of the features, there are some ways to further enhance the NER task performance by introducing an attention mechanism. The attention mechanism allows neural networks to focus on the most informative elements of the inputs (Fu et al., 2021; Wang et al., 2022). The NER models (Li, Zhao, et al., 2019; Xu, Yang, et al., 2019) could capture the entity-related features in the input data by employing attention. For example, (Li, Zhao, et al., 2019) proposed a BILSTM-ATT-CRF model, which used the bidirectional maximum matching method to extract the entities in EMRs from the dictionary and applied attention for capturing important information from a character-level component in the BILSTM-CRF-based NER model. (Xu, Yang, et al., 2019) explored a Dic-Att-BILSTM-CRF model that matches entities in the text with the dictionary, and the attention assigns a certain weight for the matched entities to recognize the disease names. The experimental results show that the dictionary-integrated attention model is better than the model without a dictionary.

Although the above methods exhibit good performance on NER tasks, they mainly use words and characters as features to represent text. These fine-grained text representation methods frequently cause the loss of the overall semantic information of some longer medical entities. The focus of the attention mechanism on the overall entity is also distracted. Inspired by this, we propose to fuse medical concept embedding based on the character embeddings of the text; these coarse-grained language structures can add the overall semantic information of the entities. Moreover, we also propose a novel attention strategy, dictionary-guided attention, which can improve the overall attention of the attention mechanism to the entity and give the attention mechanism the capability to learn prior knowledge. Specifically, we first construct a biomedical dictionary and then match the text string with our biomedical dictionary to obtain the entities contained in the text. The concept embedding corresponding to these entities is integrated with the entity character embeddings to construct a multigranularity text representation. Meanwhile, the entities will obtain higher weight values by using an attention mechanism to assign adaptive weights for the matched entities by the calculated assignment loss.

3. Problem definition

As shown in Table 2, the entity type recognized in the Chinese EMR includes 6 categories. The BIO tagging schema is used to label the character tag in an entity, which indicates whether a character begins with an entity (B), is inside with an entity (I), or is outside of an entity (O). Each entity is annotated by one of six types, i.e., a symptom, check, check result, diagnosis or disease, treatment or medicine, and nonentity. An example of BIO tagging is shown in Fig. 1, where DD denotes the entity type of "diagnosis or disease".

For the biomedical named entity recognition task, a small subset of entities D_l is manually labeled from the Chinese EMR dataset D_{all} as the initial training data. D_m represents the biomedical dictionary, which contains a large number of medical entities. The goal of BNER is to iteratively train a classifier using D_l and D_m so that the classifier can be adapted to predict the entity tags (including the entity boundaries and entity categories) of all entities in D_{all}-$D_l = D_u$.

4. Methodology

This paper proposes a semisupervised model to extract the medical entities from Chinese EMRs. The model mainly consists of two stages, training the DGAN model and updating the data. An overview of our approach is shown in Fig. 2.

First, a small amount of labeled data is fed to train the DGAN model. After that, a large amount of unlabeled data is input into the trained model to predict the labels. Then, the relabeled-strategy sample selection method is utilized to select high-confidence samples and form the confidence set. The confidence set will be applied to expand the biomedical dictionary and train the next round of the model, and the model will be improved at the next iteration. The two stages are repeated iteratively until the stopping criterion is reached.

4.1. DGAN model

The core of this semisupervised model is to train the DGAN model. The architecture of the proposed DGAN model is shown in Fig. 3.

Specifically, we introduce concept embedding to supplement the overall semantic information of the entity and propose a dictionary-guided attention strategy to efficiently adapt the BILSTM-Att-CRF model to the BNER task. DGAN consists of three modules, dictionary matching, embedding and concept-guided classifier building. In the first module, we integrate biomedical resources to form a biomedical dictionary, and an efficient dictionary matching technique is introduced to match with this dictionary. Second, we utilize a large-scale real-world unlabeled EMR dataset to pretrain RoBERTa (Liu, Ott, et al., 2019) to elaborate semantic character embeddings. Meanwhile, the conceptual representation of the medical entity is integrated with the character embeddings to construct the feature representations with richer semantic information. Finally, the predicted sequence labels are calculated by the DGAN model constructed by leveraging BILSTM in

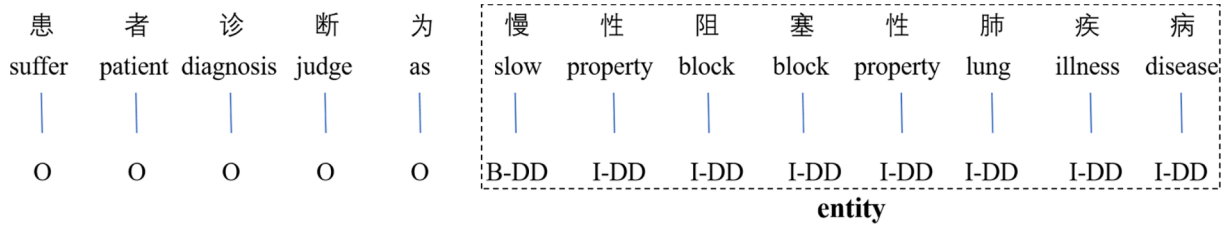


Fig. 1. BIO tagging schema.

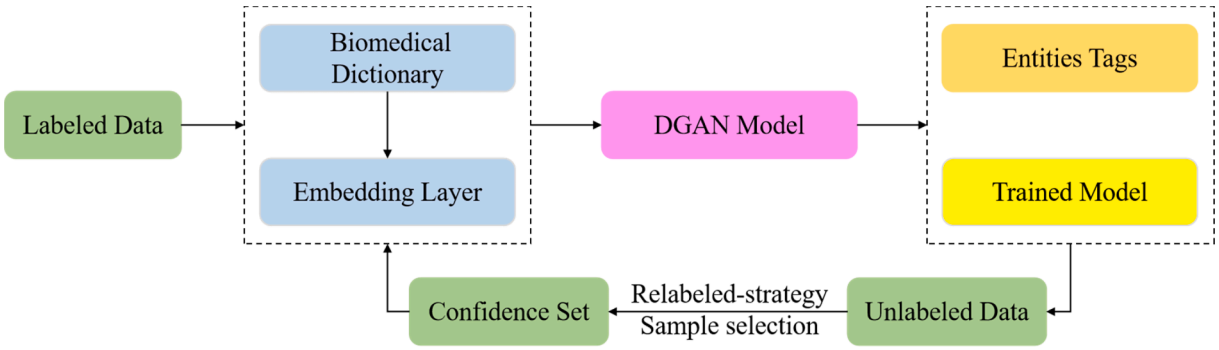


Fig. 2. Overview of the semisupervised model framework for BNER.

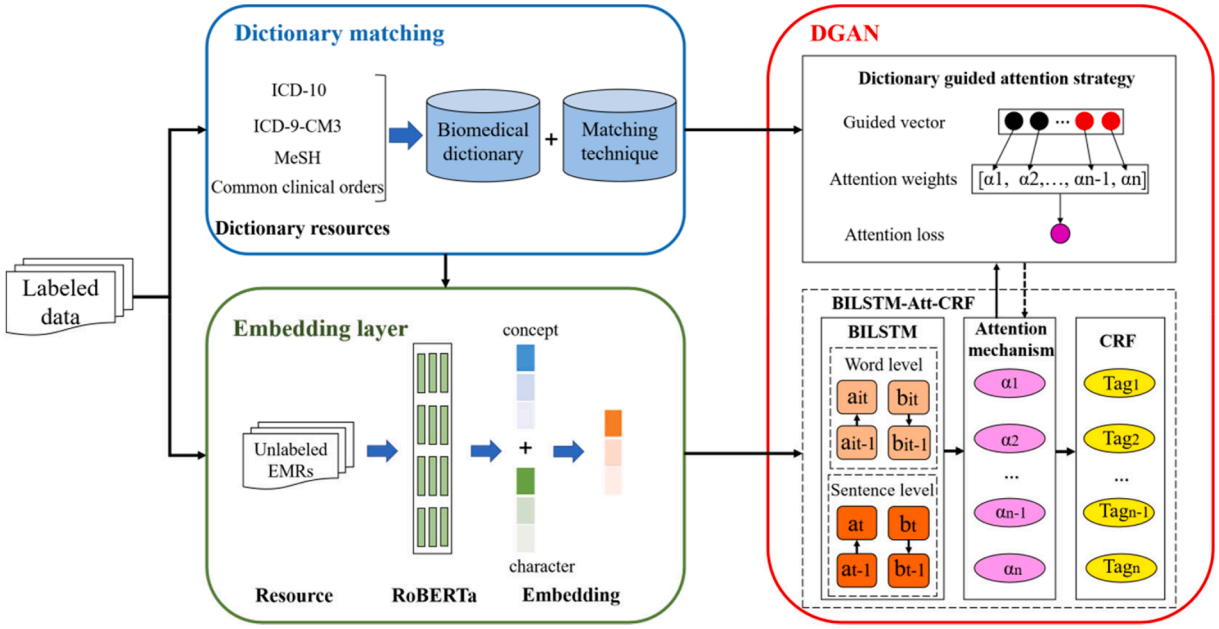


Fig. 3. The main architecture of our proposed DGAN model.

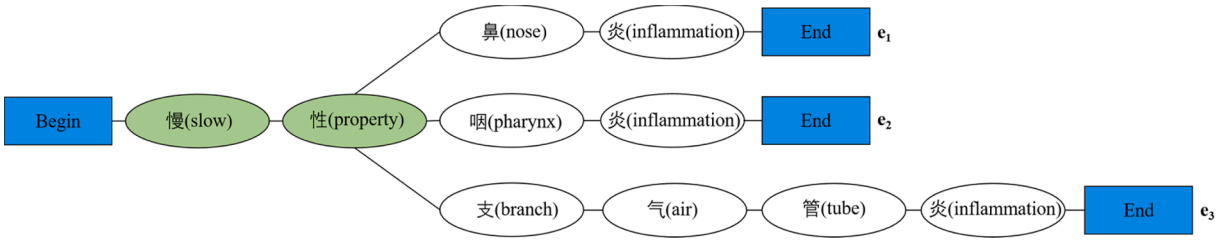


Fig. 4. An example of the data structure of a Trie tree.

conjunction with the attention mechanism, CRF and the dictionary-guided attention strategy.

4.1.1. Dictionary matching

The existing biomedical dictionaries contain considerable prior knowledge, which is effective in the BNER task. The construction of a medical dictionary can improve the model to more accurately focus on the medical named entities in the text. To maximize the coverage of the medical named entities, we pool a large number of existing medical resources into a medical dictionary. The constructed Chinese biomedical dictionary contains 30,743 medical entities based on the disease diagnosis code library ICD-10, medical subject headings (MeSH) (Lipscomb. 2000), ICD-10 medical insurance version 2.0, ICD-9-CM3 medical insurance version 2.0 and common clinical orders. Moreover, the dictionary will be gradually expanded with the iterative training of the model.

After the Chinese medical dictionary construction, we use the Trie tree (Fredkin, 1960) to store it. It is a kind of multifork tree structure that can store the dictionary efficiently. As shown in Fig. 4, three concepts are stored by a Trie tree, e_1[慢性鼻炎] (chronic rhinitis), e_2[慢性咽炎] (chronic pharyngitis) and e_3[慢性支气管炎] (chronic bronchitis). The blue parts represent the beginning and end of a word boundary, respectively, and the green parts are identical characters among the various concepts. The Aho-Corasick algorithm (Aho & Corasick, 1975) is a widely used method for string matching. To obtain an accurate

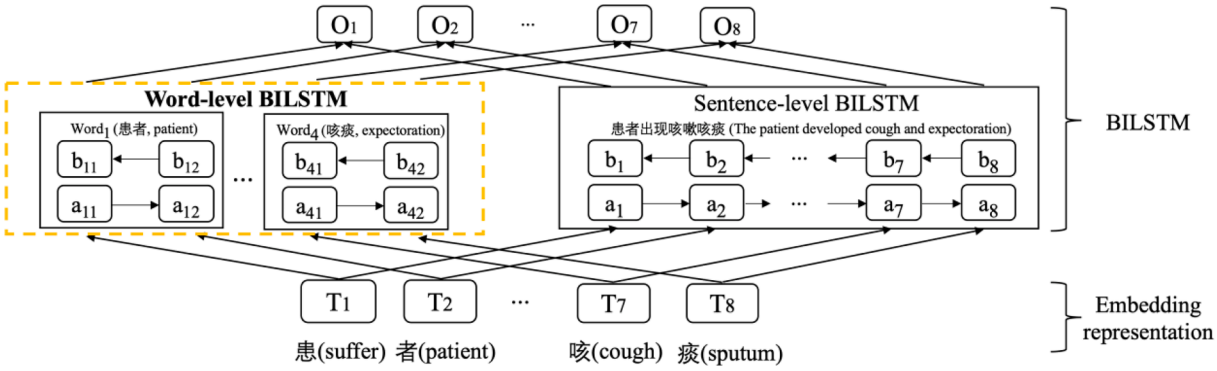


Fig. 5. BILSTM structure in our method.

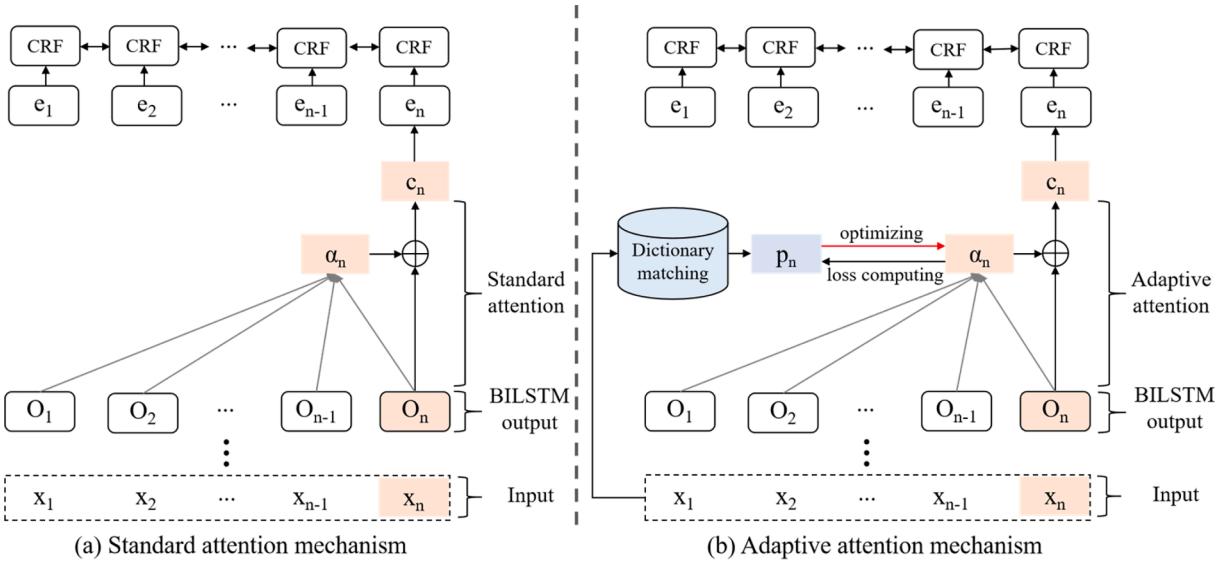


Fig. 6. Comparison between the standard attention mechanism (a) and adaptive attention mechanism (b). The feature representation layer and BILSTM are omitted in this figure.

guidance vector to optimize the weight assignment of standard attention, we only accept the matching results with complete concepts rather than a substring in a concept. Therefore, the time complexity of the Aho-Corasick algorithm is reduced to $O(n)$, compared to $O(mn)$, which accepts both the complete concept and substring.

4.1.2. Embedding layer

Our approach will leverage RoBERTa to implement character embedding generation, which can avoid the mistakes of entity boundary recognition caused by word segmentation. However, the commonly used RoBERTa has only been pretrained on Wikipedia, Book-Corpus, CC-NEWS, OPENWEBTEXT and STORIES. In contrast, medical texts consist of many technical terms that rarely appear in these corpora. Therefore, before generating the character embeddings, we obtained a large-scale real-world medical corpus to elaborate the embeddings. The corpus contains 12,518 unlabeled Chinese EMRs, including many different diseases, such as apoplexy, hyperlipidemia and adverse drug reactions. We adopt the recommended tips in (Devlin et al., 2018) because there is no obvious difference between RoBERTa and BERT in structure, which suggests running additional fine-tuning steps on a specific domain starting from the existing RoBERTa checkpoint. After the fine-tuning process is finished, we obtain a TensorFlow model that ends with.ckpt. Then, we leverage the tips mentioned in (Li, Zhang, et al., 2018) to transform the model into a PyTorch model ending with.bin. When the fine-tuned RoBERTa is obtained, the inputs of RoBERTa are the characters in a sentence of the EMRs, and the outputs are the feature vectors of the characters.

Moreover, considering that the single character-level embedding loses the overall semantic information of the entity, we add the coarse-grained concept embedding to the character-level embedding, i.e., for each character feature T_i, there are two versions. One version is only the character mention T_i itself, and the other is the summation between itself and the medical concept (appearing in the medical dictionary) it involves, where i is the i-th character in the text.

4.1.3. Concept-guided classifier building

(a) BILSTM module

In the context of NER, LSTM has been widely used to extract semantic features, which is a variant of RNN proposed to conquer the vanishing gradient problem. BILSTM is composed of two LSTM units in different directions. In addition to the sentence-level global semantic feature extraction, we consider word-level local features to further increase the available features. Therefore, BILSTM contains two parts, a word-level BILSTM and sentence-level BILSTM. The structure of BILSTM in our method is shown in Fig. 5.

In Fig. 5, the hidden nodes (a_{11}, b_{11}, a_1, b_1, etc.) represent LSTM units. Each LSTM unit can be expressed as Eqs. (1)-(5):

$$f_t = \sigma\left(W_{Tf}T_t \cdot W_{hf}h_{t-1} + W_{cf}C_{t-1} + b_f\right) \quad (1)$$

$$i_t = \sigma(W_{Ti}T_t \cdot W_{hi}h_{t-1} + W_{ci}C_{t-1} + b_i) \quad (2)$$

$$C_t = (1 - i_t)^* C_{t-1} + i_t \odot \tanh(W_{TC}T_t + W_{hC}h_{t-1} + b_C) \quad (3)$$

$$o_t = \sigma(W_{To}T_t \cdot W_{ho}h_{t-1} + W_{co}C_{t-1} + b_o) \quad (4)$$

$$h_t = o_t^* tanh(C_t) \quad (5)$$

where the subscript t represents the time step. f_t, i_t, C_t and o_t are the gates of forget, input and output in the LSTM unit. T_t, h_{t-1}, and C_{t-1} are inputs, and h_t and C_t are outputs. C_t is a temporary cell. W_f, W_i, W_C and W_o with subscripts T, h, and c, are the weight matrices of the input character representation T_t, hidden state h_t and temporary cell C_t, respectively. b_f, b_i, b_C and b_o represent the bias vectors. $\odot$ and σ represent the elementwise product and the sigmoid function, respectively.

For the word-level BILSTM, we can obtain the word-level text representation h_w by Eq. (1)-(5). The specific representation of h_w is shown in Eq. (6):

$$h_w = [h_{w1}, ..., h_{wm}] = [[a_{11}; b_{11}], ..., [a_{1k}; b_{1k}], ..., [a_{m1}; b_{m1}], ..., [a_{mk}; b_{mk}]] \quad (6)$$

where m is the total number of words in a sentence ($m = 4$, as the example shows in Fig. 5). k is the total number of characters in a word ($k = 2$, as *Word*1 shown in Fig. 5). [;] represents the vector concatenating operation.

Similarly, the sentence-level text representation h_s can be obtained by Eqs. (1)–(5). The specific representation of h_w can be expressed by Eq. (7):

$$h_s = [[a_1; b_1], ..., [a_{n-1}; b_{n-1}], [a_n; b_n]] \quad (7)$$

wheres n represents the total number of characters in the sentence ($n = 8$, as shown in Fig. 5).

Finally, by aligning and concatenating the text representation containing the word-level context information and sentence-level context information, the rich contextual text representation O is obtained, which is expressed by Eq. (8):

$$O = [O_1, O_2, ..., O_{n-1}, O_n] = [h_w; h_s] \quad (8)$$

(b) Attention module

Most medical terminologies are composed of multiple characters and words; character-level and word-level text segmentation methods split these entities into multiple fragments, and assigning weight values for each fragment by standard attention may disperse the attention score and result in a lower weight value for the medical terminology. We proposed dictionary-guided adaptive attention to optimize the standard attention by using a biomedical dictionary. Fig. 6 illustrates its overall framework, Fig. 6 (a) shows the standard attention and Fig. 6 (b) shows the dictionary-guided adaptive attention. It can be roughly divided into four stages: (1) guidance vector construction, (2) weight calculation by a standard attention mechanism, (3) assigning weights for characters, and (4) weight optimization by dictionary-guided adaptive attention.

(1) Guidance vector construction: The guidance vector is constructed by matching characters in the medical text into the constructed Chinese medical dictionary M. If a character sequence has a corresponding concept in M, this sequence is labeled 1 by a guidance vector; otherwise, it is labeled 0. We want the model to pay more attention to the matched characters. The matching function is shown in Eq. (9):

$$p_i = \begin{cases} 1, matched \\ 0, unmatched \end{cases} \quad (9)$$

For example, given a sentence X = [“患” (suffer), “者” (patient), “出” (out), “现” (appear), “咳” (cough), “嗽” (cough), “咳” (cough), “痰” (sputum)], after matching, there are four characters “咳” (cough), “嗽” (cough), “咳” (cough), “痰” (sputum) has corresponding concepts “咳嗽”(cough), “咳痰” (expectoration) in M, this sentence will label as P = [0, 0, 0, 0, 1, 1, 1, 1] by the guidance vector. It is worth noting that our matching strategy is a full string matching, rather than a substring. For example, a sentence [“心” (heart), “脏” (viscus), “结” (knot), “构” (structure), “未” (not), “见” (see), “明” (bright), “显” (apparent), “异” (different), “常” (normal)], the character “心” (heart) and “脏” (viscus) exists in the entity “心脏病” (heart disease), however, it is not a successful match because only two characters, i.e., “心” (heart) and “脏” (viscus), are matched. The guidance vector for the instance [“心” (heart), “脏” (viscus), “结” (knot), “构” (structure), “未” (not), “见” (see), “明” (bright), “显” (apparent), “异” (different), “常” (normal)] will be P = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0], rather than P = [1, 1, 0, 0, 0, 0, 0, 0, 0, 0]. This matching rule can effectively reduce the noise.

(2) Weight calculation by standard attention. The result of BILSTM will be leveraged as the input for the standard attention mechanism to obtain the weight coefficient. The calculation process is shown in Eqs. (10)–(11):

$$u_i = tanh\left(W_u\left[O_i; O_j\right]\right) \quad (10)$$

$$\alpha_i = \frac{exp(u_j)}{\sum_{k=1}^{n} \exp(u_{ik})} \quad (11)$$

where n is the number of characters in a sentence. i,j and k denote the i-th, j-th and k-th characters in the sentence (i,j,$k \in [1,n]$). u_i represents an alignment function. W_u needs to be learned in the process of training. O_i and O_j are the character representations output by BILSTM. α_i is the attention weight obtained by calculating the similarity between character representation O_i and the other character O_j in a sentence, and the attention weight will be larger with increasing similarity.

(3) Assigning weight for characters. The sentence representation $E = [e_1, e_2, \cdots, e_n]$ is obtained in the third step after adding the attention weight. The calculation process of the character representation e_i is shown in Eqs. (12)–(13):

$$c_i = \sum_{j=1}^{n} \alpha_j O_j \quad (12)$$

$$e_i = \tanh(W_e[c_i; O_i]) \quad (13)$$

where c_i is a global vector calculated as a weighted sum of the BILSTM output. e_i represents the weighted representation of each character, [;] represents the concatenated operation and W_e needs to be learned in the process of training.

(4) Weight optimization by dictionary-guided adaptive attention. Dictionary-guided adaptive attention is proposed to optimize the weight assignment of standard attention (in the second step) according to the guidance vector. The guidance vector is regarded as a soft threshold to adaptively assign more attention to the matched characters by calculating the loss between the guidance vector and standard attention. The benefit of this strategy is to adjust the weight coefficient in an adaptive manner and retain both the concept information in the dictionary and the indicative information of the context words. In our work, the mean-square error (MSE) is used as the loss function. As shown in Eq. (14):

$$MSE = -\frac{1}{n}\sum_{j=1}^{n}\left(\hat{y}_j - y_j\right)^2 \quad (14)$$

where n represents the total number of estimators, $\widehat{y}_j$ represents the estimator, and y_j represents the target value. In this paper, we regarded p_i as the target value labeled by the guidance vector, and α_i is the estimated value computed by standard attention. The loss function is calculated in Eq. (15):

$$L_{att} = -\frac{1}{q}\sum_{i=1}^{n}(\alpha_i - p_i)^2 p_i \tag{15}$$

where q represents the total number of target values of 1 in the guidance vector. The loss function is calculated according to the target value of the guidance vector. Note that if a character is labeled as 0, its attention loss will also be 0, which can help to reduce the error because the EMR text may contain concepts that do not appear in the dictionary. We expect the model to adaptively assign more weight to the characters involved in a concept while preserving the semantic information of the context to infer the entity's category and boundary.

(c) CRF module

In the model training process, the CRF model can automatically learn the constraints and effectively predict the dependency between tags, and it can model the labeled sequence to obtain the global optimal sequence. The matrix Q is a scoring matrix, $Q_{i,j}$ is a probability value that classifies the i-th character as the j-th tag, and $Trans_{i,j}$ is a state transition score from the i-th to the j-th tag. For an input sentence $X = [x_1, x_2 \cdots, x_n]$, its labeled sequence $Y = [y_1, y_2, \ldots, y_n]$ is scored as Eq. (16):

$$Score(X, Y) = \sum_{i=1}^{n} Q_{i,y_i} + \sum_{i=1}^{n} Trans_{y_i, y_{i+1}} \tag{16}$$

The maximum log-likelihood function is leveraged to train the CRF model. If sentence X is known, we can calculate the conditional probability of the labeled sequence Y by Eq. (17) and Eq. (18), where Y_X is all possible labeled sequences, and L is the loss function.

$$P(Y|X) = \frac{exp(Score(X, y))}{\sum_{\tilde{y} \in Y_X} exp(X, \tilde{y})} \tag{17}$$

$$L = log(P(Y|X)) \tag{18}$$

In the CRF model prediction process, the Viterbi algorithm is leveraged to solve the global optimal sequence, as shown in Eq. (19). Y^* is the sequence that obtains the highest score in the score function.

$$Y^* = \underset{\tilde{y} \in E_X}{argmax} Score(X, \tilde{y}) \tag{19}$$

(d) Training Objective

Assume that there are N medical sentences in the training dataset $[X_1, X_2, \cdots, X_N]$. We train the DGAN model, which contains both BILSTM-ATT-CRF and a dictionary-guided attention strategy. The objective function is shown in Eq. (20):

$$\min Loss = \sum_{i=1}^{N} L_i + \lambda L_{i_{att}},\ 0 \le \lambda \le 1 \tag{20}$$

where L_i and $L_{i_{att}}$ are the loss of the final entity recognition and the weight assignment loss of the attention mechanism, respectively. The calculation methods are shown in Eq. (15) and Eq. (18). λ represents the importance coefficient of attention weight loss.

4.2. Semi-supervised method

Considering that a biomedical dictionary and a large-scale labeled dataset are challenging and expensive to obtain, we use the self-training approach (Zhu, 2005), a classical semisupervised learning method, to expand the scale of the training dataset. Moreover, to avoid the error accumulation caused by too little initial training data, we optimize the self-training algorithm by introducing a relabeled strategy described in (Livieris, 2019) for unlabeled samples.

Self-learning predicts the unclassified data with the most reliable prediction by adopting the assumptions obtained from the unclassified samples (Yarowsky, 1995). The basic assumption of self-training is to use a model to predict the unlabeled samples, it iterates over samples with high confidence, and it correctly predicts (Didaci et al., 2006). For the BNER task, we have three datasets of D_m, D_l and D_u, where D_m is the biomedical dictionary, D_l is the labeled samples and D_u is the unlabeled samples. The optimized self-training algorithm leveraging the relabeled strategy is shown in Algorithm 1.

Algorithm 1: The self-training algorithm leveraging relabeled strategy

Input: The initial training dataset D_l and biomedical dictionary D_m, unlabeled dataset D_u. The Confidence level *ConL*.**Output:** A biomedical entity classifier.**Step 1:** Pretrain the Model M with the biomedical dictionary D_m and the labeled data D_L to obtain the pretrained Model M'. **Repeat:** **Step 2:** Adopt M' to predict the unlabeled dataset D_T. **Step 3:** Select the samples with the predicted probability more than *ConL* per iteration (D_T') leveraging relabeled strategy. **Step 4:** Expand D_m and D_L with D_T', i.e., $D_m + D_T' \rightarrow D_m$, $D_L + D_T' \rightarrow D_L$ and remove from D_T. **Step 5:** Train the Model M' with D_m and D_L. **Until:** Some stopped criteria are met or D_T is empty.

First, the incomplete biomedical dictionary D_m and the small-scale labeled data D_L are leveraged to pretrain the initial Model M. Then, the pretrained Model M' is adopted to predict the labels of the unlabeled sample set D_T and obtain the pseudolabels. The relabeled strategy is leveraged to select the samples with a prediction probability higher than the predetermined confidence *ConL* to generate the set D_T'. *ConL* is a specified threshold, which denotes the confidence level. The data and the labeled entities contained in D_T' are subsequently added to the initial training set D_L and the biomedical dictionary D_m to extend the scale of the training set and the dictionary iteratively (i.e., $D_L + D_T' \rightarrow D_L$, $D_m + D_T' \rightarrow D_m$) and increase their robustness. Meanwhile, remove D_T' from D_T. The newly obtained dictionary D_m and training set D_L are adopted to retrain Model M' until some stopping criteria are met or D_T is empty.

5. Experimental setup

5.1. Experimental dataset

Two datasets are leveraged to evaluate the superiority and generalization of the proposed DGAN model, namely:

COPD: A real-world Chinese dataset about chronic obstructive pulmonary disease (COPD) acquired from a cooperating hospital. A total of 138 EMRs are provided, which were manually labeled by the guidance of professional medical experts and teams. A total of 3,588 sentences can be leveraged after data processing. A total of 1,588 sentences are labeled as data and leveraged for the first training of the model, where 1300 sentences are leveraged as a training set and 288 sentences are leveraged as a test set. The other 2000 unlabeled data are used for iterative training, and 400 sentences are fed each iteration.

CCKS2017[1]: It is a public dataset provided by Beijing Jimuyun Health Technology Co., which includes five categories of medical entities, which are body, symptoms, check, disease and treatment. There are 800 labeled records after desensitization treatment (a single visit record of a single patient). We segment the records into sentences and follow the 7:1 probability to divide them into the training set and test set.

5.2. Evaluation metrics

The precision P, recall R, and $F1$ scores are chosen as evaluation

[1] https://biendata.com/competition/CCKS2017_1/.

Table 3
The hardware environment of experience.

Processor	Cache	GPU	Hard disk
Intel(R) Xeon(R) Silver 4216 CPU @ 2.10 GHz	128 GB	Nvidia RTX3090 24 GB	512 GB SSD + 4 T 5400 rpm

Table 4
Component comparison of different baselines.

Model	Domain-specific knowledge fine-tuning	Character embedding	Traditional machine learning	Deep learning	Attention mechanism
DM	√	×	×	×	×
HMM	×	×	√	×	×
BC	×	×	×	√	×
BBC	×	√	×	√	×
RSBGC	×	√	×	√	×
FBBCE	√	√	×	√	×
DABLC	√	×	×	√	√
DGAN	√	√	×	√	√

Table 5
Comparison with other advanced models.

Model	COPD			CCKS-2017		
	P (%)	R (%)	F1(%)	P (%)	R (%)	F1(%)
DM	45.54 ± 0.00	49.08 ± 0.00	47.24 ± 0.00	65.15 ± 0.00	67.65 ± 0.00	66.38 ± 0.00
HMM	62.45 ± 0.00	62.94 ± 0.00	62.69 ± 0.00	79.04 ± 0.00	80.76 ± 0.00	79.89 ± 0.00
BC	67.62 ± 1.53	66.21 ± 1.89	66.91 ± 1.51	83.55 ± 1.43	84.88 ± 1.07	84.21 ± 1.16
BBC	72.50 ± 1.38	72.51 ± 0.85	72.50 ± 0.60	87.66 ± 0.92	87.30 ± 0.95	87.48 ± 0.70
RSBGC	74.78 ± 0.26	74.37 ± 0.42	74.57 ± 0.27	89.09 ± 0.43	88.37 ± 0.31	88.73 ± 0.37
FBBCE	79.51 ± 0.15	78.44 ± 0.45	78.97 ± 0.31	91.96 ± 0.33	91.33 ± 0.18	91.64 ± 0.28
DABLC	80.14 ± 0.19	78.80 ± 0.49	79.46 ± 0.20	92.51 ± 0.16	91.93 ± 0.13	92.22 ± 0.11
DGAN-BERT (ours)	80.15 ± 0.46	79.37 ± 0.53	79.75 ± 0.44	94.37 ± 0.02	93.91 ± 0.06	94.14 ± 0.03
DGAN (ours)	80.30 ± 0.06	80.73 ± 0.02	**80.51 ± 0.03**	94.96 ± 0.06	95.39 ± 0.07	**95.17 ± 0.05**

performance metrics for the NER task. The *F*1 value is the weighted average of *P* and *R*. The specific calculation process is shown in Eqs. (21–(23):

$$P = \frac{TP}{TP + FP} \times 100\% \tag{21}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{22}$$

$$F1 = \frac{2^{*}P^{*}R}{P + R} \times 100\% \tag{23}$$

where *TP* represents the number of correctly recognized entities, *FP* represents the number of unrelated entities recognized, and *FN* represents the number of unrecognized entities. During prediction, the standard is to judge whether the prediction of a medical entity is completely correct, in that the boundary and category of the entity are predicted correctly at the same time.

5.3. Experimental environment and parameter settings

In this paper, the software environment of the BNER model is based on the Linux operating system, and Python 3.6 is the development language. We leveraged the PyTorch1.0 framework, Jieba, NumPy and other dependent libraries in the building process of the model. The hardware experiment environment is shown in Table 3.

In the model parameter settings part, the EMR text contains many short sentences, which lack fully synthetic and semantic indications on the linguistic composition. Considering that two sentences may not capture enough semantic information to learn a token representation, we adopt an improved model, RoBERTa, to replace BERT for pretraining. RoBERTa removes NSP, and each input includes a full sentence that is sampled contiguously from one or two texts. RoBERTa-wwm-ext is a version of RoBERTa that focuses on Chinese embedding generation. In our study, the dimension of the hidden layer of RoBERTa-wwm-ext is fixed at 768, the dimension of the multihead attention layer is set to 12 and the maximum sequence length is set to 512. To make RoBERTa-wwm-ext suitable for this work, it is further fine-tuned on our real-world Chinese EMRs. The dimension of the hidden layer of BILSTM is consistent with RoBERTa-wwm-ext. The learning rate is set to 0.0005, and dropout is used after the recurrent layer with a value of 0.1 to reduce overfitting. The batch size is fixed at 32. The importance coefficient of adaptive attention (λ value) is set to 0.05. The confidence level (*ConL* value) of the semisupervised strategy is set to 0.95. Adam is utilized for optimization.

6. Results

6.1. Comparison experiments of multiple advanced models

To ensure the validity of our proposed DGAN model, we reproduced some classical and advanced models as the baselines for comparison, including:

- DM (Quimbaya et al., 2016): It is a classical model based on string matching for NER.
- HMM (Teng et al., 2010): This method uses the Hidden Markov Model (HMM) for NER.
- BC (Xu, Zhou, et al. (2018)): This model uses word embeddings and the BILSTM-CRF model for NER.
- BBC (Dai et al., 2019): This model utilizes BERT to generate the character embeddings and uses the BILSTM-CRF model for NER.
- RSBGC (Yang et al., 2016): This model adopts RoBERTa to generate word embeddings, and leverages the Stacked BIGRU-CRF framework for NER.
- FBBCE (Li, Zhang, et al., 2020): This model utilizes the domain-specific medical knowledge and BERT to generate character embeddings, and then the BILSTM-CRF model is used to recognize entities.
- DABLC (Xu, Yang, et al., 2019): This model extracts concepts from the external dictionary to improve the standard attention mechanism for BNER.

We analyze the components of the above methods leveraging "domain-specific knowledge fine-tuning", "character embedding", "traditional machine learning", "deep learning" and "attention mechanism" in Table 4, where "√" denotes that the method contains the corresponding component, whereas "×" denotes that the corresponding component is not involved in the method.

To make the comparison study feasible, we implement two versions of the DGAN model. (1) DGAN-BERT uses BERT to encode the text. (2) DGAN: This version uses RoBERT to encode the text.

Table 5 shows the BNER performance of the proposed models (DGAN-BERT, DGAN) and 7 baseline methods. From Table 5, we can see that our models (DGAN-BERT, DGAN) outperform all baseline approaches, especially the DGAN, which improves the results significantly. Among the 7 baselines, five deep learning-based models, i.e., BC, BBC, RSBGC, FBBCE and DABLC obtain better performance than DM and

Table 6
The results of the *p*-value on two datasets after comparing different methods using each of the evaluation measures.

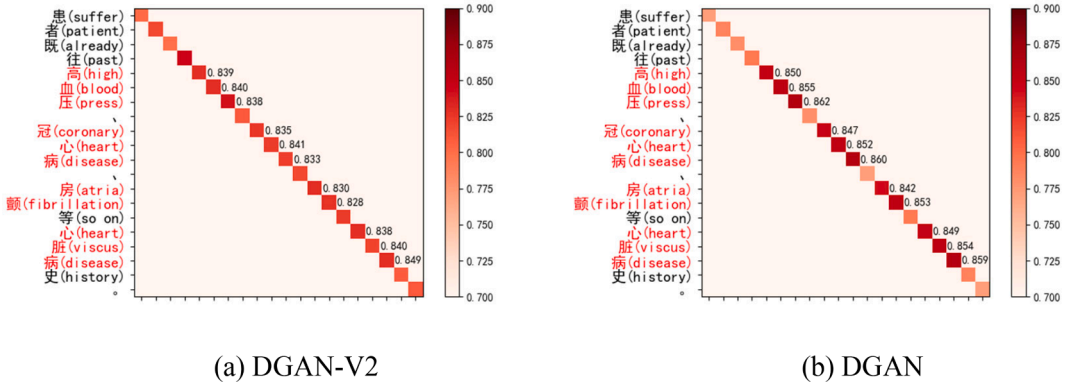
Model	*p*-value					
	COPD			CCKS-2017		
	P	R	F1	P	R	F1
DGAN/DM	4.83e−28	4.89e−32	1.10e−30	6.34e−28	1.39e−27	9.24e−28
DGAN/HMM	1.94e−25	8.72e−30	3.04e−28	1.70e−25	4.69e−25	2.77e−25
DGAN/BC	2.90e−12	2.08e−11	1.82e−12	6.92e−12	2.93e−11	5.74e−12
DGAN/BBC	4.81e−11	1.28e−12	4.66e−14	4.22e−12	6.90e−12	4.98e−13
DGAN/RSBGC	1.56e−15	1.34e−13	1.22e−15	3.84e−14	5.68e−15	3.99e−13
DGAN/FBBCE	1.26e−09	6.37e−07	9.98e−10	3.47e−19	6.92e−20	1.48e−19
DGAN/DABLC	1.58e−05	3.00e−11	1.42e−09	3.40e−15	5.37e−15	6.64e−15

Table 7
The effects of different components.

Model	COPD			CCKS2017		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
DGAN-V1	78.59 ± 0.46	78.82 ± 0.26	78.70 ± 0.27	94.44 ± 0.52	94.76 ± 0.81	94.60 ± 0.62
DGAN-V2	77.80 ± 1.53	78.70 ± 1.89	78.25 ± 1.51	91.71 ± 0.45	92.36 ± 0.15	92.03 ± 0.30
DGAN	80.30 ± 0.06	80.73 ± 0.02	**80.51 ± 0.03**	94.96 ± 0.06	95.39 ± 0.07	**95.17 ± 0.05**

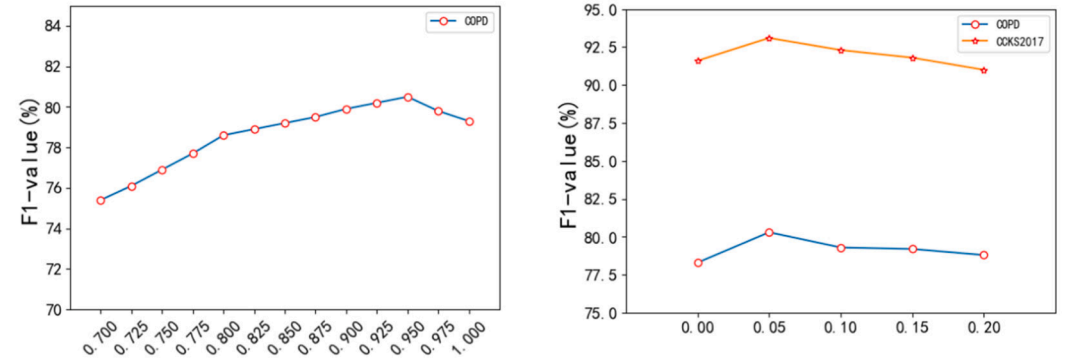
HMM. This demonstrates that the automatic learning of semantic features through deep learning enables the resulting model to better understand the meaning of the clinical texts. DGAN outperformed DGAN-BERT by 0.76% and 1.03% in terms of the F1 score in COPD and CCKS-2017, respectively, which validates the effectiveness of RoBERTa.

The DABLC is the closest one to our DGAN model, and it also utilizes medical concepts in the dictionary as knowledge to guide the attention assignment for BNER. The fact that the DABLC achieved a higher F1 score than other baselines on both datasets also illustrates the effectiveness of using medical concepts and the dictionary-guided attention mechanism as indicative information for BNER. DGAN-BERT and DGAN outperform DABLC on both datasets, which can be explained by two points. (1) Different from DABLC, which uses word2vec for feature representation, BERT (RoBERTa) is used in DGAN-BERT (DGAN) to pretrain feature representations from a large-scale medical corpus. (2) DABLC assigns attention weight by matching medical text into the domain dictionary, and the weight score is set to 1 for matched concepts. Our DGAN-BERT (DGAN) model utilizes an adaptive attention strategy that focuses not only on the matched concepts but also on context words with indicative information for entity inference. DABLC (using word2-vec for feature representation) and FBBCE (using BERT for feature representation) are better than the RoBERTa-based model RSBGC. This



(a) DGAN-V2 (b) DGAN

Fig. 7. Attention weight assignment (the red color represents the character involved in a medical concept).



(a) $ConL$ value (b) λ value

Fig. 8. The performance of DGAN with different key hyperparameters.

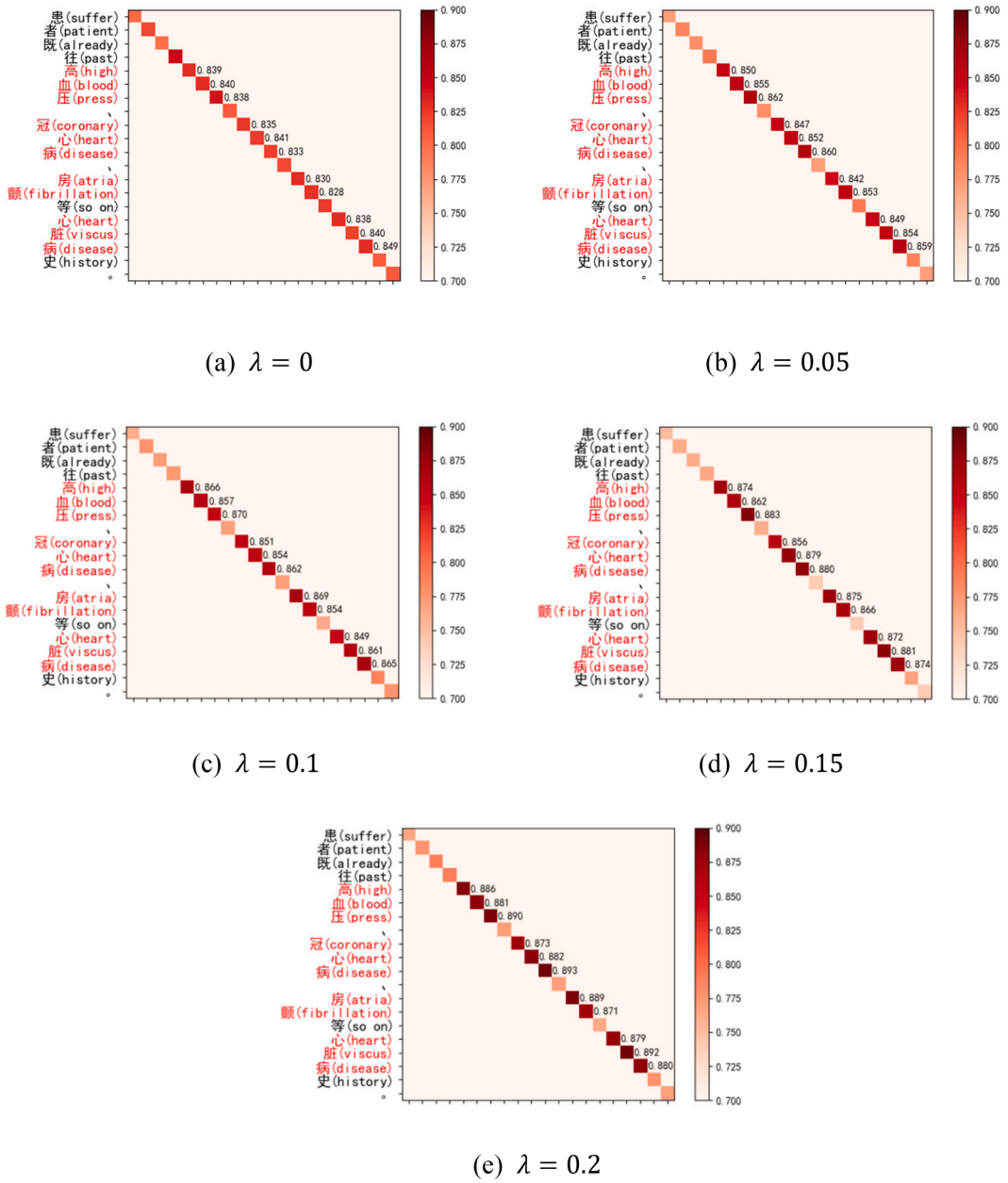


Fig. 9. The attention weight assignment of different coefficients (the red color represents the character involved in a medical concept).

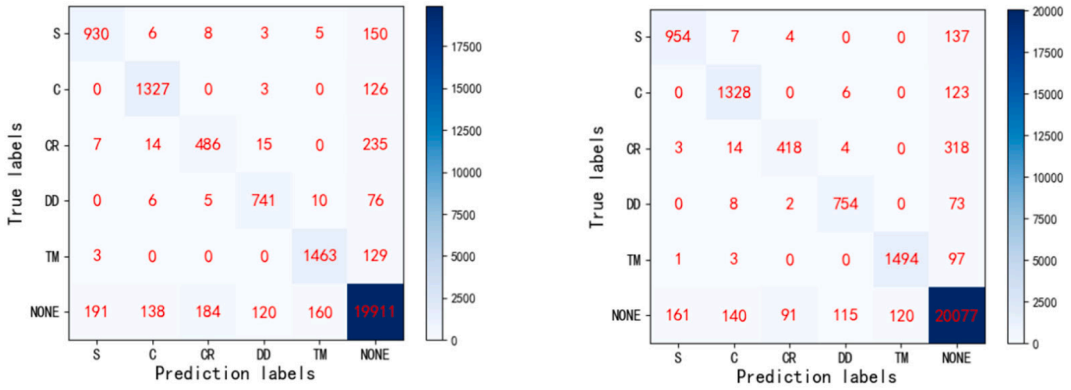
demonstrates that domain knowledge is more important than the neural language model. In addition, all baseline approaches employ the fully labeled dataset to train a BNER model, while our proposed models (DGAN-BERT and DGAN) use the partially labeled data as an initial training set. Then, the high confidence recognition results are added as training samples of classifier learning in the next iteration. It validates that our model provides great potential to reduce human efforts for training data labeling.

For the statistical significance of the results, we apply the *p*-value (Hollander & Sethuraman, 2001) to evaluate differences between the results produced by DGAN and each other approach over 10 runs. In this study, if the *p*-value is less than 0.05, the results of the two approaches are considered significantly different. From Table 6, we can see that our model is always better than the baselines, and the *p*-value is less than 0.05, which proves that DGAN outperforms other approaches with statistical significance.

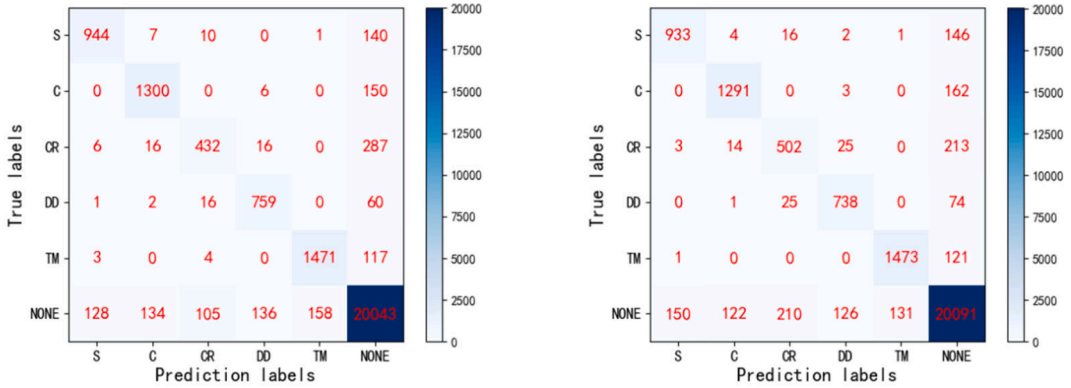
Meanwhile, it is necessary to note that the different promotion effects of our model's F1 score on the two datasets may be due to two reasons. First, the data scales of the two datasets are different. The COPD dataset has only 138 records, while the CCKS2017 consists of 800 records. Second, the data quality is different. The COPD dataset is provided by a cooperating hospital. It contains many abbreviations and writing errors of the medical entities, which can easily cause a dictionary matching failure. However, the CCKS2017 dataset is standard data with high quality and has little effect on dictionary matching. The performance of the model is proportional to the scale and quality of the data, which is generally true for all deep learning approaches. As a result, compared with the best baseline (i.e., DABLC), our proposed DGAN achieves a 2.95% F1 score improvement in the CCKS2017 dataset and only a 1.05% F1 score improvement in the COPD dataset.

6.2. Ablation study

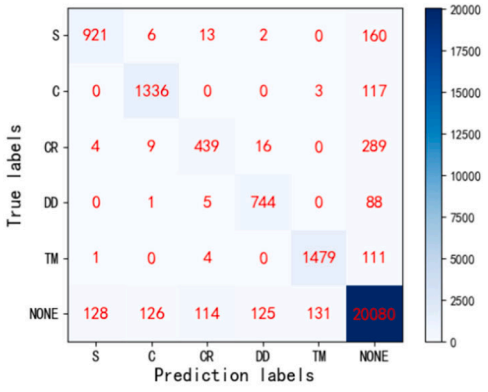
The aim of the ablation study (Meyes et al., 2019) is to evaluate the contributions of different components in the proposed DGAN model. In addition to the full version of the DGAN model, two simplified versions



(a) $\lambda = 0$ (b) $\lambda = 0.05$



(c) $\lambda = 0.1$ (d) $\lambda = 0.15$



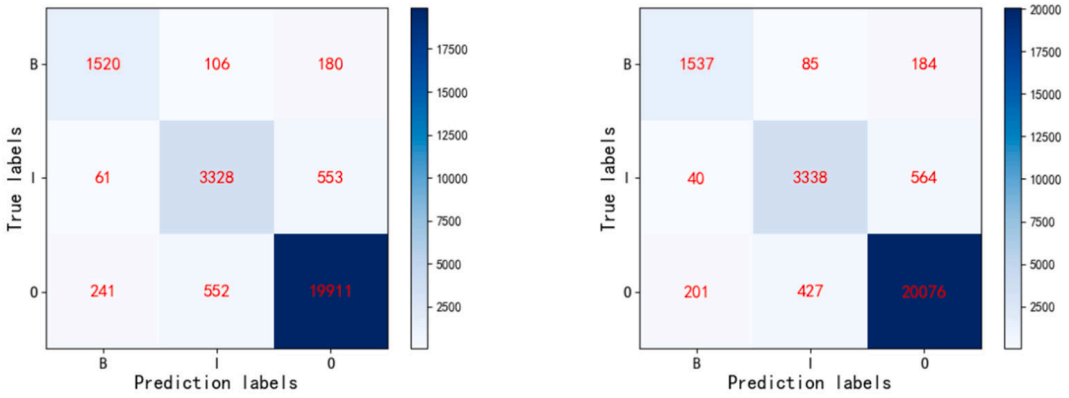
(e) $\lambda = 0.2$

Fig. 10. The entity category recognition results on the COPD dataset. "S" represents "symptom", "C" represents "check", "CR" represents "check result", "DD" represents "diagnosis or disease", "TM" represents "treatment or medicine" and "None" represents "nonentity".

of the DGAN model are implemented to evaluate the effects of different components in the DGAN model:

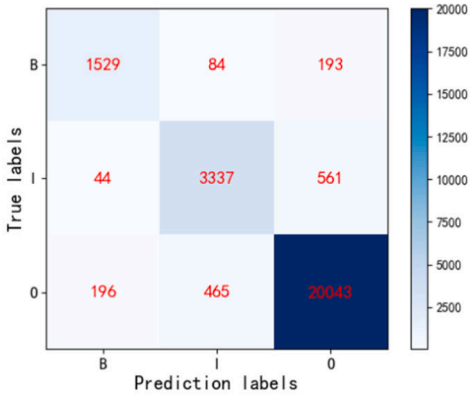
- DGAN-V1: The implementation of the DGAN model without concept embedding, however, the dictionary matching and adaptive attention strategy are considered in this model.
- DGAN-V2: The implementation of the DGAN model utilizes a standard attention strategy, which means that dictionary matching is not considered in this model.

The performance of two simplified versions of the DGAN model are illustrated in Table 7. From these figures, we find that both constituent components (i.e., concept embedding and dictionary-guided adaptive attention) are effective for BNER. The concept embedding improves the quality of BNER by 1.81% and 0.57% in terms of the F1 score on the COPD and CCKS2017 datasets, respectively. The dictionary matching component and adaptive attention strategy contributed more than concept embedding. Fig. 7 shows an example of standard attention (a) and adaptive attention (b). The example is randomly selected from the

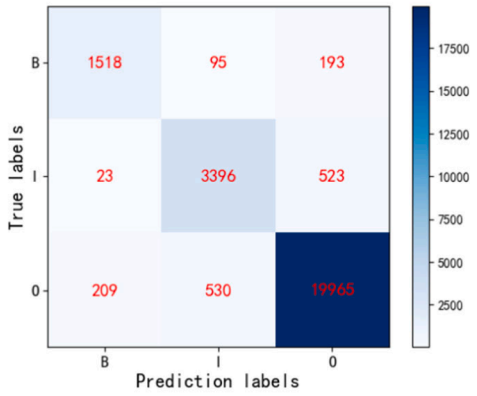


(a) $\lambda = 0$

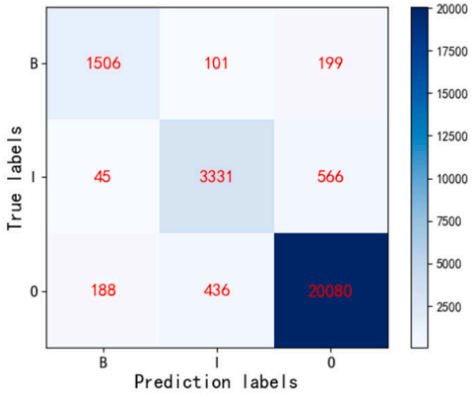
(b) $\lambda = 0.05$



(c) $\lambda = 0.1$



(d) $\lambda = 0.15$



(e) $\lambda = 0.2$

Fig. 11. The boundary category recognition results on the COPD dataset.

test data. The attention mechanism learns to impose higher weights upon important characters for BNER. The higher the weight is, the darker the intersection color is. From Fig. 7, we can see that compared to DGAN-V2, the characters involved in medical concepts (e.g., "高" (high), "血" (blood), "压" (press)) are assigned higher weights in our DGAN model. It can be further used to explain BNER because they are considered significant to make the classification by the model. Therefore, adaptive attention sheds some light on the interpretability of the BNER models.

6.3. Parameter analysis

For the key hyperparameters of the DGAN model, i.e., the confidence level (*ConL* value) of the relabeled strategy and the importance coefficient (λ value) of the attention weight loss, how to select their optimal values is important. Fig. 8 shows the results with different key hyperparameters.

For confidence level tuning, we evaluated the DGAN's performances with different confidence levels (*ConL* values) in the range of [0.7–1].

The results in Fig. 8 (a) show that when the *ConL* value is smaller (larger) than 0.95, the recognition performance is improved (decreased) with an increasing *ConL* value. Therefore, the *ConL* value is set to 0.95 in the learning of the DGAN model.

The control variable method is leveraged to select the importance coefficient of adaptive attention, i.e., λ value, we only change the weight of the importance coefficient in the range of [0, 0.05, 0.1, 0.15, 0.2], and the other parameters are fixed. The F1 score of the DGAN model with different importance coefficients on both the COPD and CCKS2017 datasets is reported in Fig. 8 (b). Meanwhile, to better observe the influence of the different importance coefficients on the adaptive attention mechanism, we randomly select a sample from the test set to reflect the change in the attention weights, as shown in Fig. 9.

From Fig. 9, we can see that as the importance coefficient increases from 0 to 0.2, the DGAN model pays more attention to the medical entities (i.e., "高血压" (hypertension), "冠心病" (coronary heart disease), "房颤" (atrial fibrillation), and "心脏病" (heart disease)). However, as shown in Fig. 8 (b), when λ is over 0.05, the F1 score of the proposed model decreases. We think that this result is because the guidance vector only provides the accurate location information of the medical entities and does not provide the category information of the entity and the entity boundary, whereas the correct entity and boundary category classification is key to measuring model performance. When we continually increase the importance coefficient, the model will gradually spend more attention on adjusting the weight assignment of the attention mechanism. The core task offset is the fundamental reason for the significant degradation in the recognition ability of the model, and it is revealed explicitly through the experiments, as shown in Fig. 10 and Fig. 11.

Fig. 10 and Fig. 11 show the recognition results of the entity categories and entity boundary categories under different importance coefficients on the COPD dataset. By comprehensive comparison of the experimental results, we can observe that as λ gradually increases, the recognition accuracy of the model for the entities declines obviously, which strongly confirms our previous conjecture. Meanwhile, we can conclude that the ideal value of the importance coefficient is $\lambda = 0.05$ in the current context.

7. Discussion

The proposed approach itself is highly generic but limited in Chinese EMRs. From the experimental results, we can observe that the proposed DGAN model can obtain excellent BNER results. It not only uses the complete concept representation to aggregate the fragmented character features but also introduces an external biomedical dictionary to guide the attention mechanism to pay high attention to the overall entity, which means that more indicative information can be provided to the model to identify the entities.

The main limitations of our study are obvious. The performance of the proposed DGAN model relies heavily on the quality of the biomedical dictionary, i.e., whether the entities mentioned in the EMR are defined in the dictionary. If the mentioned entities in EMR are not involved in the dictionary, the concept representation cannot be used to aggregate the character fragments, and the guidance vector cannot be constructed to guide the weight assignment of a target entity.

8. Conclusion

In this paper, we propose a dictionary-guided attention network named DGAN for biomedical named entity recognition in Chinese EMRs. To capture the complementary semantic information of entities, we construct a synthetical biomedical dictionary to learn the coarse-grained concept embedding, combining concept and character embeddings to encode the medical text. Meanwhile, a dictionary-guided attention strategy is proposed to adaptively assign higher weights to characters contained in a concept, which could prevent dispersion of the attention scores of the entities by the standard attention mechanism. In addition, semisupervised learning is introduced to reduce the manual labeling of data and to handle the entities not defined in the medical dictionary. The experiments on a real-world dataset and a public dataset demonstrate that our approach can optimize the weight assignment strategy of standard attention and achieve the best performance compared to the baselines in BNER. Since a Chinese EMR text contains many short sentences, in future work, we will explore the semantic completion method to extract more semantic indications of the entities.

CRediT authorship contribution statement

Zhichao Zhu: Conceptualization, Methodology, Investigation, Software, Data curation, Writing – original draft, Validation, Supervision, Writing – original draft. **Jianqiang Li:** Conceptualization. **Qing Zhao:** Conceptualization, Methodology, Writing – original draft. **Faheem Akhtar:** Data curation, Formal analysis.

Declaration of Competing 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.

Data availability

The authors do not have permission to share data.

Acknowledgment

This study is supported by the research project from the National Natural Science Foundation of China (No.62266041) and the R&D Program of Beijing Municipal Education Commission (KM202310005031).

References

Aho, A. V., & Corasick, M. J. (1975). Efficient string matching: An aid to bibliographic search. *Communications of the ACM, 18*(6), 333–340. https://doi.org/10.1145/360825.360855

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). *On the properties of neural machine translation: Encoder-decoder approaches*. arXiv preprint arXiv: 1409.1259.

Chokwijitkul, T., Nguyen, A., Hassanzadeh, H., & Perez, S. (2018, July). Identifying risk factors for heart disease in electronic medical records: A deep learning approach. In *Proceedings of the BioNLP 2018 workshop* (pp. 18–27).

Didaci, L., & Roli, F. (2006). Using co-training and self-training in semi-supervised multiple classifier systems. In *Structural, syntactic, and statistical pattern recognition: Joint IAPR International Workshops, SSPR 2006 and SPR 2006, Hong Kong, China, August 17-19, 2006. Proceedings* (pp. 522-530). Springer Berlin Heidelberg.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv: 1810.04805.

Dai, Z., Wang, X., Ni, P., Li, Y., Li, G., & Bai, X. (2019, October). Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. In *2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei)* (pp. 1-5). IEEE.

Fredkin, E. (1960). Trie memory. *Communications of the ACM, 3*(9), 490–499. https://doi.org/10.1145/367390.367400

Fu, G., Li, J., Wang, R., Ma, Y., & Chen, Y. (2021). Attention-based full slice brain CT image diagnosis with explanations. *Neurocomputing, 452*, 263–274.

Grishman, R., & Sundheim, B.M. (1996). Message Understanding Conference- 6: A Brief History. *International Conference on Computational Linguistics*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Sequence modeling: Recurrent and recursive nets. *Deep Learning*, 367–415.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780.

Hollander, M., & Sethuraman, J. (2001). *Nonparametric statistics: Rank-based methods*.

Kim, Y., Denton, C., Hoang, L., & Rush, A.M. (2017). *Structured Attention Networks*. ArXiv, abs/1702.00887.

Lipscomb, C. E. (2000). Medical subject headings (MeSH). *Bulletin of the Medical Library Association, 88*(3), 265. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC35238/.

Li, J., Liu, C., Liu, B., Mao, R., Wang, Y., Chen, S., Yang, J., Pan, H., & Wang, Q. (2015). Diversity-aware retrieval of medical records. *Computers in Industry, 69*, 81–91.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature, 521*, 436–444.

Liang, J., Xian, X., He, X., Xu, M., Dai, S., Xin, J., Xu, J., Yu, J.P., & Lei, J. (2017). A novel approach towards medical entity recognition in Chinese Clinical Text. *Journal of Healthcare Engineering*, 2017.

Liu, K., Hu, Q., Liu, J., & Xing, C. (2017). Named entity recognition in Chinese electronic medical records based on CRF. In *2017 14th Web Information Systems and Applications Conference (WISA)*, pp. 105–110.

Li, J., Zhao, S., Yang, J., Huang, Z., Liu, B., Chen, S., Pan, H., & Wang, Q. (2018). WCP-RNN: A novel RNN-based approach for Bio-NER in Chinese EMRs. *The Journal of Supercomputing, 76*, 1450–1467.

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., & Wang, J. (2018). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics, 34*, 1381–1388.

Li, J., Sun, A., Han, J., & Li, C. (2018). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering, 34*, 50–70.

Livieris, I. (2019). A new ensemble semi-supervised self-labeled algorithm. *Informatica* (03505596), 43(2).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach*. arXiv preprint arXiv:1907.11692.

Li, L., Zhao, J., Hou, L., Zhai, Y., Shi, J., & Cui, F. (2019). An attention-based deep learning model for clinical named entity recognition of Chinese electronic medical records. *BMC Medical Informatics and Decision Making, 19*, 1–11.

Li, J., Liu, L., Sun, J., Mo, H., Yang, J., Chen, S., Liu, H., Wang, Q., & Pan, H. (2020). Comparison of different machine learning approaches to predict small for gestational age infants. *IEEE Transactions on Big Data, 6*, 334–346.

Li, X., Zhang, H., & Zhou, X. H. (2020). Chinese clinical named entity recognition with variant neural structures based on BERT methods. *Journal of Biomedical Informatics, 107*, Article 103422.

Meyes, R., Lu, M., Puiseau, C. W., & Meisen, T. (2019). *Ablation studies in artificial neural networks*. ArXiv, abs/1901.08644.

Ma, Z., Zhao, L., Li, J., Xu, X., & Li, J. (2022). SiBERT: A Siamese-based BERT network for Chinese medical entities alignment. *Methods*.

Quimbaya, A. P., Sierra-Múnera, A., Rivera, R. A., Rodríguez, J. C., Velandia, O. M., Peña, A. A., & Labbé, C. (2016). Named entity recognition over electronic health records through a combined dictionary-based approach. *CENTERIS/ProjMAN/HCist.*

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

Teng, Q. Q., Ji, J. M., Zheng, R. T., & Li, N. (2010). Applicability analysis of Chinese named entity recognition method based on literatures. *Journal of Information, 29* (09), 157–161.

Unanue, I. J., Borzeshi, E. Z., & Piccardi, M. (2017). Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *Journal of Biomedical Informatics, 76*, 102–109. https://doi.org/10.1016/j.jbi.2017.11.007

Wu, Y., Jiang, M., Xu, J., Zhi, D., & Xu, H. (2017). Clinical named entity recognition using deep learning models. In *AMIA annual symposium proceedings* (Vol. 2017, p. 1812). American Medical Informatics Association.

Wang, R., Fu, G., Li, J., & Pei, Y. (2022). Diagnosis after zooming in: A multi-label classification model by imitating doctor reading habits to diagnose brain diseases. *Medical Physics*.

Xu, K., Zhou, Z., Hao, T., & Liu, W. (2018). A bidirectional LSTM and conditional random fields approach to medical named entity recognition. In *proceedings of the international conference on advanced intelligent systems and informatics 2017* (pp. 355-365). Springer International Publishing. https://doi.org/10.1007/978-3-319-64861-333.

Xu, K., Yang, Z., Kang, P., Wang, Q., & Liu, W. (2019). Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. *Computers in Biology and Medicine, 108*, 122–132.

Yarowsky, D. (1995, June). Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics* (pp. 189–196).

Yang, Z., Salakhutdinov, R., & Cohen, W. (2016). *Multi-task cross-lingual sequence tagging from scratch*. arXiv preprint arXiv:1603.06.

Zhu, X. J. (2005). *Semi-supervised learning literature survey*.

Zhang, S., & Elhadad, N. (2013). Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics, 46* (6), 1088–1098.

Zhao, Q., Kang, Y., Li, J., & Wang, D. (2018). Exploiting the semantic graph for the representation and retrieval of medical documents. *Computers in Biology and Medicine, 101*, 39–50.

Zhao, Q., Xu, D., Li, J., Zhao, L., & Rajput, F. A. (2022). Knowledge guided distance supervision for biomedical relation extraction in Chinese electronic medical records. *Expert Systems with Applications, 204*, Article 117606.

Q. Zhao J. Li L. Zhao Z. Zhu Zhao, Q., Li, J., Zhao, L., & Zhu, Z. (2022). Knowledge guided feature aggregation for the prediction of chronic obstructive pulmonary disease with Chinese EMRs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.