Artificial Intelligence In Medicine 150 (2024) 102813

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Research paper

Online biomedical named entities recognition by data and knowledge-driven model

Lulu Cao [a,1], Chaochen Wu [c,*,1], Guan Luo [b,**,1], Chao Guo [d], Anni Zheng [b]

[a] Department of Rheumatology and Immunology, Peking University People's Hospital, 100044, China
[b] State Key Laboratory of Multimodal Artificial Intelligence Systems Institute of Automation, Chinese Academy of Sciences, China
[c] Renmin University of China, Beijing, 100872, China
[d] Department of Cardiology, Fuwai Hospital CAMS and PUMC, Beijing, 100037, China

ARTICLE INFO

Keywords:
Biomedical named entity recognition
Neural network
Pre-training
Knowledge representation
Online text

ABSTRACT

Named entity recognition (NER) is an important task for the natural language processing of biomedical text. Currently, most NER studies standardized biomedical text, but NER for unstandardized biomedical text draws less attention from researchers. Named entities in online biomedical text exist with errors and polymorphisms, which negatively impact NER models' performance and impede support from knowledge representation methods. In this paper, we propose a neural network method that can effectively recognize entities in unstandardized online medical/health text. We introduce a new pre-training scheme that uses large-scale online question-answering pairs to enhance transformers' model capacity on online biomedical text. Moreover, we supply models with knowledge representations from a knowledge base called multi-channel knowledge labels, and this method overcomes the restriction from languages, like Chinese, that require word segmentation tools to represent knowledge. Our model outperforms other baseline methods significantly in experiments on a dataset for Chinese online medical entity recognition and achieves state-of-the-art results.

1. Introduction

Taking advantage of natural language understanding (NLU) methods, biomedical researchers and physicians can use computational tools to analyze a large amount of electrical medical records, online health texts, and academic publications. In addition, we can employ NLU models as a core component for building intelligent medical applications to reduce the health expenses of customers and educate people to prevent some diseases. Named entity recognition (NER) is an important task for NLU, which aims to extract and classify entities from text. NER for medical text has an additional challenge: some medical entities are complicated. They are constructed by several words with specific means; they usually appear in complex sentences with a low frequency; the cost of sentences annotation is much higher than other NER topics because the annotation of medial text requires a lot of supports from medical professionals, so dataset sizes of annotated medical corpora are small for biomedical NER.

Different from accurate and precise named entities from academic documents, recognizing named entities from the online biomedical/health text have extra obstacles (data examples in Fig. 1): first, many biomedical entities from the online text are typos; second, people without professional background usually place biomedical entities in the wrong place; third, people may use different expressions they prefer to refer standardized biomedical entities. Although the NER for online biomedical text is a difficult task, it has important research and industry applications. For example, by biomedical NER, we can mine people's thoughts and their need for specific diseases or drugs. What is more, with the artificial intelligence application that is powered by online biomedical NER, patients can get quality answers instantly after they post their questions to websites.

Several machine learning methods were applied to the named entity recognition of medical text. A common machine learning approach for NER is the conditional random fields (CRF), and it was used to recognize biological entities [1] and medical entities [2]. The structural support vector machine (SSVM) also had a promising performance for clinical named entity recognition [2]. In the past few years, some bio-medical named entity recognition studies have started using neural network methods. The BiLSTM-CRF overcomes drawbacks of Bidirectional long short-term memory (BiLSTM) in sequence labeling and generates impressive results for medical NER [3]. The self-attention mechanism can additionally improve the BiLSTM-CRF model

* Corresponding author.
** Correspondence to: State Key Laboratory of Multimodal Artificial Intelligence Systems Institute of Automation, Chinese Academy of Sciences, China.
E-mail addresses: wuchaochen2021@ruc.edu.cn (C. Wu), gluo@nlpr.ia.ac.cn (G. Luo).
[1] Contribute equally

https://doi.org/10.1016/j.artmed.2024.102813
Received 31 March 2023; Received in revised form 15 December 2023; Accepted 12 February 2024
Available online 21 February 2024

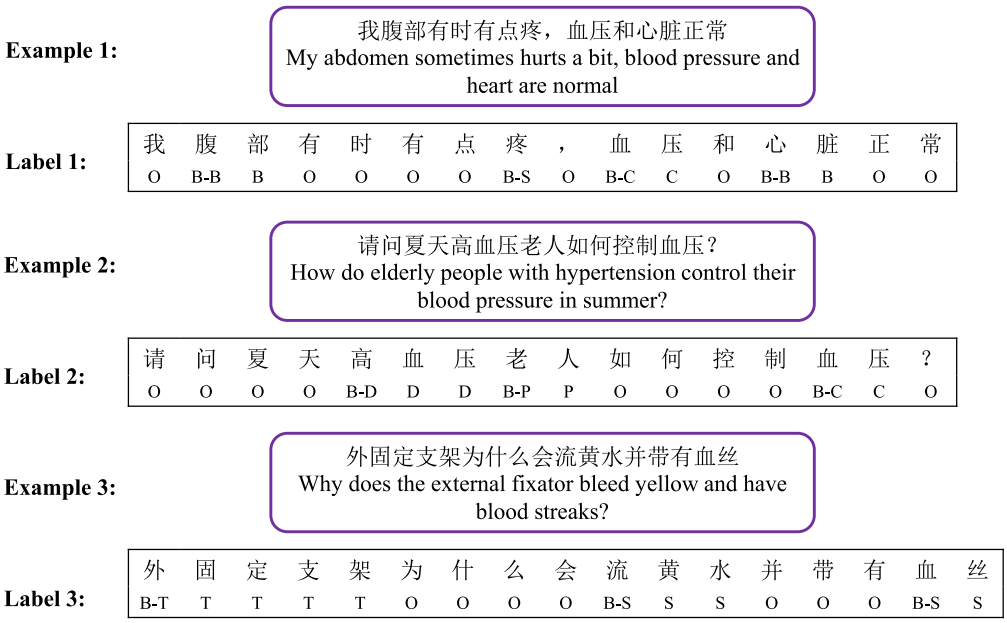


Fig. 1. Data examples for online biomedical NER. It shows problems in online biomedical text and their NER labels.

performance on chemical named entity recognition [4], and the CNN layer can boost models' name entity recognition ability on Chinese clinical text [5]. Recently, the transformer-based model achieved state-of-the-art results in many NLU tasks, which imply transformer-based models, like pre-training of deep bidirectional transformers (BERT) [6], could achieve better results for biomedical NER tasks. For example, pre-trained deep-learning model [7–10] have been used to effectively recognize biomedical entities from different datasets, and experiment results showed the pre-trained model generates promising results across different corpora. Besides pre-trained methods, biomedical NER models that are powered by knowledge representations achieved impressive accuracy gains [11–13].

Although the current NER models have achieved outstanding progress in the clinical text and academic text, the NER of online medical or health text is still a challenging task. Named entities from online medical texts show higher polymorphism. For example, some users prefer to use "stroke", but others prefer "cerebrovascular accident (CVA)".; typos appear much more frequently in online health text than in electronic health records (EHR) and academic documents; people without medical backgrounds may mistakenly use some medical words, somebody may say "how do I measure hypertension?", but the true phase is "how do I measure blood pressure?". In this paper, addressing challenges in biomedical NER for online medical/health text, we provide one data-driven strategy and one knowledge-driven strategy to improve the transformer model performance on biomedical NER tasks. We called it Data and Knowledge-driven Biomedical NER (DKD-BiomedNER). We employ enormous resources from online health communities, which use question-answering pairs to pre-train the transformer model. Moreover, we used a knowledge base to process medical text to generate a method for knowledge representations called multi-channel knowledge labels. This method supports online biomedical NER by significantly improving the model's capacity to use knowledge from knowledge bases. For language without delimiters between words, retrieval knowledge from knowledge bases exists challenge: tokens produced by word segmentation tools are disappointing for Chinese biomedical text [2], which restricts the effect of knowledge representations. Our method solves this problem without word segmentation tools and uses knowledge to improve model performance. We provide an open-accessible dataset[2] for online biomedical NER, and we hope our dataset could let more researchers pay attention to online biomedical NLU and facilitate people to contribute to this area.

2. Method

2.1. Online question-answering pairs pre-training

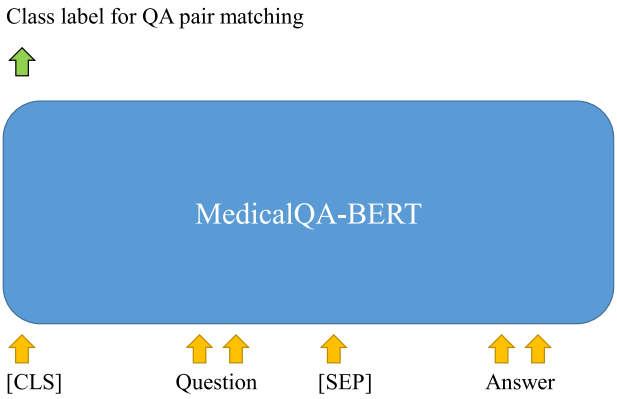
The BERT model was pre-trained by large text corpora like English Wikipedia. However, the medical text has specific words and sentence expressions, and the pre-training on corpora other than health social media text, biomedical literature, and clinical notes may not fully excavate the BERT model ability for NLU for medical/health text. Therefore, we could use medical/health text to fine-tune the BERT language model and improve its performance on the medical/health text. Existing studies used EHR or academic publications to pre-train transformer model [7–10], but these pre-trained strategies may not fit for online biomedical text. The online health/medical text written by users is usually much shorter than academic and clinical texts, which impedes pre-train tasks to learn inter-sentence relationships. In this paper, we focus on NER for online medical texts, which contain professional or unprofessional expressions that approach conversations between patients and doctors. The question-answering pairs data from online health communities are exactly fit for this goal, which is easy to collect and less likely to leak patients' sensitive information. On health community websites, a person may ask a question about his or her health issues anonymously, and doctors can share their answers. One question may have zero to multiple answers. In our pre-training process, we only select one answer for a question.

The original BERT pre-training has two tasks: the masked language model (MLM) and the next sentence prediction (NSP). For the question-answering pre-training process, we did not change the MLM task and used the question-answering prediction (QAP) task to replace the NSP task. We input BERT with one sequence that contains one question, one answer, and one *[SEP]* token between them. The positive one is question-answering pairs from the original dataset, and we produce the negative one by a question and randomly select an answer from other pairs. The QAP could be represented by:

$$Q = QAP(S_1, S_2) \tag{1}$$

As this formula, S_1 is a question, and S_2 is its answer, Q could indicate whether they are QA pairs. We treat the QAP task as a binary

[2] https://github.com/yuanxiaoheben/OnlineChineseBiomedNER



Question: Hi doctor! Could you tell me the syndrome of COVID-2019?
Answer: Generally speaking, young people and patients without chronic diseases will have flu-like symptoms, with fever or slight cough in the early stage, accompanied by discomfort such as soreness and headache. The frail elderly often progress more quickly and may develop into severe pneumonia can even cause death from respiratory failure.

Fig. 2. Using online health question-answering pairs to pre-train the BERT model by question-answering prediction task.

classification (is a QA pair or not). The loss function of QAP can be computed by the cross-entropy:

$$Loss_{QAP} = CrossEntropy(Q_{true}, Q_{target}) \tag{2}$$

We used the loss from the QAP task to replace the loss for the NSP task for BERT pre-training:

$$Loss_{MeidcalQA} = \alpha Loss_{QAP} + (1-\alpha) Loss_{MLM} \tag{3}$$

Where α is the weight parameter. Fig. 2 presents how we train the BERT model by a health question-answering pair. In our experiment, we used the BERT model with parameters that were already tuned by large corpora, and then we used online medical QA pairs to pre-train the model.

2.2. Multi-channel knowledge label

The understanding of some complicate medical entities is hard for human annotators, even some students or professionals with medical majors. Hence, supplying additional knowledge may improve NLU model performance on medical text. However, Finding and representing medical knowledge from online medical/health texts is challenging. To prepare knowledge for NER, we built a knowledge base that covers 10 knowledge label types: *position*, *disease*, *medical instrument*, *people*, *food*, *operation*, *medicine*, *sport*, *diagnosis*, *syndrome*.

The errors in online medical/health text restrict us from using knowledge to improve NLU model performance. What is more, for some languages without delimiters between words, we usually need to use word segmentation tools to generate a sequence of words from a sentence for word embedding. However, current word segmentation tools cannot be competent for online medical texts. Therefore, we proposed a method for labeling Chinese medical text without word segmentation, and we call it multi-channel knowledge labeling. For example, if a word with the disease knowledge is in a sentence, we will label 1 in the disease channel and label 0 in other channels. Noticeably, some words may have two or more types of knowledge, or parts of it have other types of knowledge. For instance, in Chinese, *Hypertension* is a disease word, but its last two characters mean *blood tension*, so we also label *diagnosis* on its last two characters. For a sentence with n characters, we used this method to get a sequence of vectors to represent knowledge labels K. The dimension of vectors k is equal to the number of classes in the knowledge base, which is 10 in our study.

$$K = (k_1, k_2, \ldots, k_n) \quad k \in R^{1\times 10} \tag{4}$$

A sample sentence with its knowledge label is shown in Fig. 3. In labeled text preparation, we used each word in the knowledge base to search for the corresponding word in sentences and used the word's knowledge to generate knowledge vectors from each character in the word's position. The efficient searching algorithm (detail in Supplementary Materials) for knowledge labeling is represented by:

$$K = \Omega(S) \tag{5}$$

K is the multi-channel knowledge label for the sentence S that is generated by the searching algorithm Ω. However, during training or testing processes of NER, using a neural network module that can dynamically generate multi-channel knowledge labels is a solution for faster and easier deployment.

Therefore, we used a neural network model to train with the text S and its knowledge labels K to generate a sequence of knowledge vectors for a sentence dynamically:

$$K^* = \Omega^{RT}(S) \tag{6}$$

K^* is approximate the multi-channel knowledge label for the sentence S that is generated by the neural network knowledge labeling module Ω^{RT}. In our experiment, we labeled the question corpus by the searching algorithm Ω and trained a BERT model that focuses on generating knowledge labels as Ω^{RT}, and we let its parameters update with the NER training process. We illustrate the multi-channel knowledge labeling by Fig. 4.

BERT assisted by knowledge labels

After we get knowledge label representations, we used it to improve BERT performance on the medical text NER task. We used the self-attention mechanism to generate a matrix to represent knowledge in a sentence.

$$AttentionK = softmax(V_a(tanh(W_a K^* T))) \tag{7}$$

Then we used the attention matrix and the BERT output to generate the sequence of NER labels. The $\oplus$ symbol is the concatenation operation.

$$NER = Linear(BERT \oplus AttentionK) \tag{8}$$

The architecture of this NER model is shown in Fig. 5.

爸	爸	得	了	高	血	压	,	胸	闷	,	吃	氢	氯	噻	嗪	可	以	吗	Label
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	People
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Medical instrument
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Operation
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	Diagnosis
0	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	Disease
0	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	Syndrome
0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	Medicine
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Sport
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	Position
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Food

Fig. 3. Labeling a sentence by knowledge labels. The English translation of this sentence is: *My father has hypertension, can he eat the hydrochlorothiazide?*.

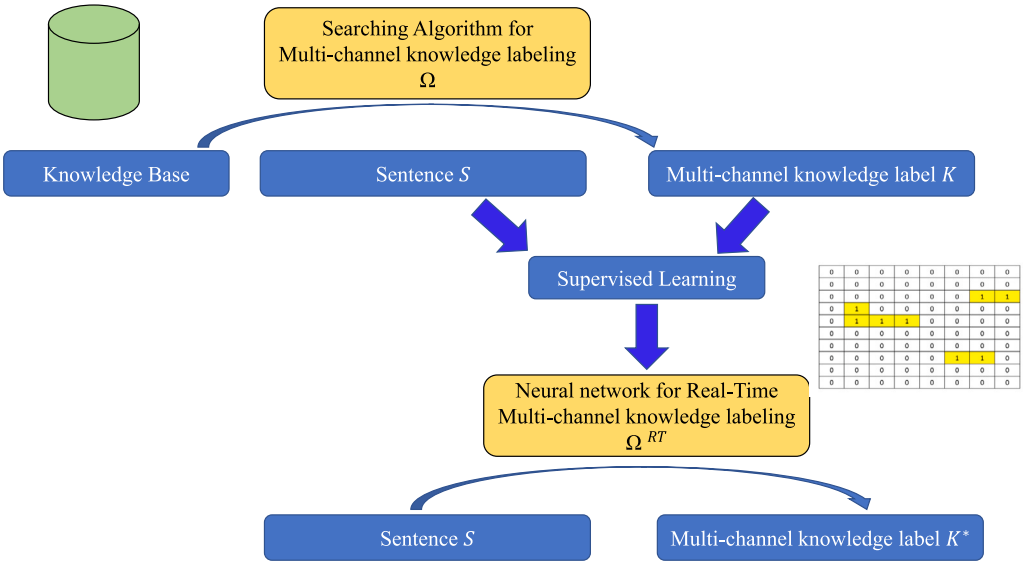


Fig. 4. Using the knowledge base to generate multi-channel knowledge label K for a sentence S by the searching algorithm Ω. The real-time knowledge labeling modelΩ^{RT} is learning by K and S, and it can generate K^* without the knowledge base.

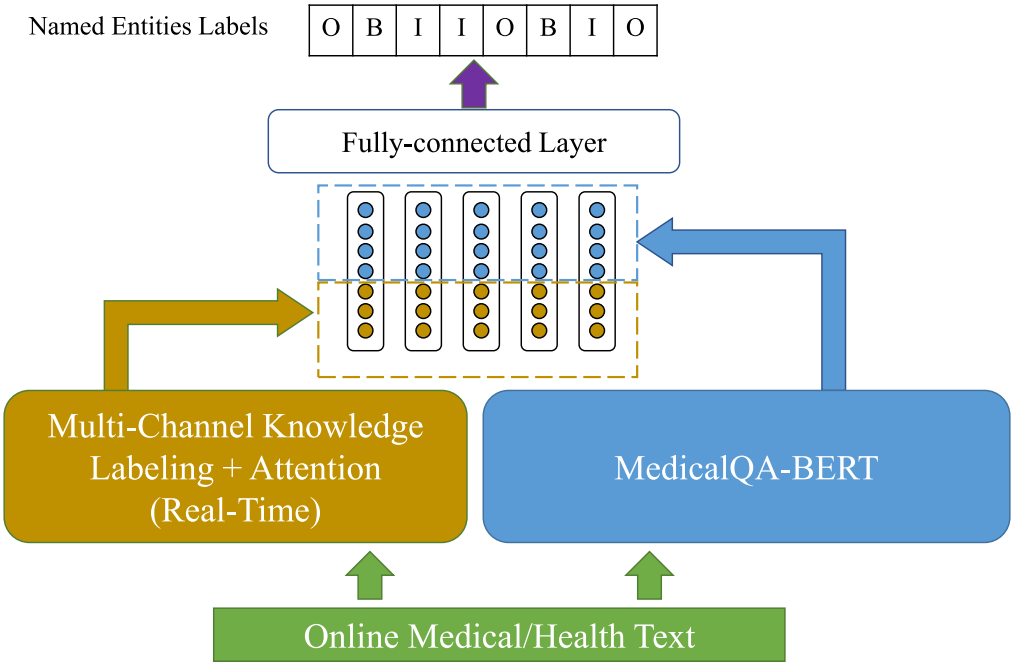


Fig. 5. The biomedical NER model architecture. It uses both MedicalQA BERT and multi-channel knowledge labeling to generate biomedical entities from online text.

Table 1
Results of NER models on MedQ-NER dataset.

Method	F1	Precision	Recall
SSVM	0.7415	0.7451	0.7379
CRF	0.7818	0.7975	0.7666
LSTM-CRF-wordvec	0.7726	0.7631	0.7824
LSTM-CRF-fasttext	0.7890	0.7850	0.7929
BERT	0.8142	0.7931	0.8364
K-BERT(CN-Dbpedia)	0.8130	0.7966	0.8301
K-BERT(MedicalKG)	0.8142	0.7976	0.8316
Lattice-LSTM	0.8272	0.8290	0.8254
BERT-CRF	0.8310	0.8250	0.8369
DKD-BiomedNER	**0.8419**	**0.8370**	**0.8470**

3. Data and experiments

3.1. Datasets

We collected 100,000 medical question-answering pairs from an online medical question-answering community.[3] Each question-answering page, it has one question from a person and one or more answers from physicians. After we collected question-answering pages, we processed them by extracting text data from pages with the python *beautifulsoup4* library; cleaning question-answering text, we removed unrelated characters and discarded too-short text; if one answer has two or more answers, we only keep one answer, and answers from experienced physicians and have suitable length had priority to be selected.

We used an annotated dataset to evaluate our model: MedQ-NER, which was annotated by medical researchers. This dataset has 6199 online medical user questions, and it has six types of medical entities: *disease*, *treatment*, *syndrome*, *checkup*, *people*, and *position*. The MedQ-NER used the BIO tagging scheme, where *B*, *I*, and *O* respectively represent the beginning, inside, and outside of the medical entity name. The MedQ-NER's annotation standard was designed by physicians, and the dataset was labeled by two experienced medical researchers. The inter-annotator agreements were evaluated by the kappa statistic, and it is over 0.83 on 500 questions during MedQ-NER annotation. Datasets' sizes for training, testing, and validation are 4599, 800, and 800 respectively.

We also made another dataset to evaluate our model performance called CMID-NER. The CMID-NER uses the corpus from a benchmark dataset (CMID) for Chinese medical question intent classification [14]. We selected 1200 medical questions for annotation. The annotator for CMID-NER is different from MedQ-NER, and the entities set and annotation standards are the same.

3.2. Experiments setting and evaluation

We used the BERT model that implemented by the PyTorch.[4] We first loaded the BERT model which had been trained on cased Chinese simplified and traditional text, and then we used medical question-answering pairs to trained it by 11 epochs. The knowledge labeling model was implemented by a pre-trained BERT model, and it was trained by automatically labeled question text by 1 epoch. The hyper-parameters used in BERT-NER models training are: the learning rate is 0.00005, and the weight decay is 0.0007.

We applied micro-averaged precision, recall, and F1 for model performance evaluations [15].

$$MicroPrecision = \frac{\sum_1^C TP_c}{\sum_1^C TP_c + \sum_1^C FP_c} \tag{9}$$

[3] https://www.familydoctor.com.cn/
[4] https://github.com/huggingface/transformers

Table 2
Results of NER models on CMID-NER testing dataset.

Method	F1	Precision	Recall
BERT	0.8406	0.8474	0.8339
K-BERT(CN-Dbpedia)	0.8158	0.8056	0.8264
K-BERT(MedicalKG)	0.8203	0.8100	0.8308
Lattice-LSTM	0.8091	0.7801	0.8404
BERT-CRF	0.8385	0.8493	0.8280
DKD-BiomedNER	**0.8676**	**0.8830**	**0.8529**

Table 3
Ablation studies on MedQ-NER dataset.

Method	F1	Precision	Recall
BERT	0.8142	0.7931	0.8364
MedicalQA-BERT	0.8243	0.8058	0.8436
KL-BERT	0.8290	**0.8374**	0.8209
DKD-BiomedNER(Jieba)	0.8276	0.8153	0.8403
DKD-BiomedNER	**0.8419**	0.8370	**0.8470**

$$MicroRecall = \frac{\sum_1^C TP_c}{\sum_1^C TP_c + \sum_1^C FN_c} \tag{10}$$

$$MicroF1 = 2 \times \frac{MicroPrecision \times MicroRecall}{MicroPrecision + MicroRecall} \tag{11}$$

For evaluating C types of entities, TP_c, FP_c, and FN_c are the number of cth type of true positive entities, cth type of false positive entities, and cth type of false negative entities, respectively. The entity matching was under a "strict" criterion, which means we checked whether model-predicted entities were exactly matched with golden labels by entity types, start positions, and end positions.

4. Result

4.1. Main results

We compared our model with several baseline methods, including CRF [1], BiLSTM with CRF [3], SSVM [2], BERT [6], and K-BERT [11], Lattice-LSTM [16], BERT-CRF [17], and these models achieved state-of-the-art performance in previous studies. For the BiLSTM with CRF model, we tried two pre-training methods: the word2vec [18,19] and the fastext [20]. For K-BERT, we use two different knowledge graphs as knowledge resources: the CN-Dbpedia [21] and a medical knowledge graph(MedicalKG) they constructed. We measure these models' micro-averaged precision, recall, and F1 scores over the MedQ-NER dataset, and experiment results are shown in Table 1.

In our experiments, pre-trained transformer-based models outperform other methods except Lattice-LSTM. The LSTM-CRF model that uses word vectors that were pre-trained by the fasttext had a higher F1 score than the LSTM-CRF model with word2vec. The BERT model got over 0.81 F1 scores, which is higher than LSTM-CRF's results. In our experiment, K-BERT, as one of the state-of-the-art NER methods, only generated similar results compared with BERT, which shows that unstandardized biomedical entities restrict knowledge graphs from improving models' performance. Lattice-LSTM and BERT-CRF also had outstanding performances in previous studies, but our method still outperforms them in the online biomedical NER task.

What is more, we take CMID-NER as the test dataset for evaluating models that are trained by MedQ-NER additionally. The result is shown in Table 2. We can see that our method, DKD-BiomedNER, performed better in the CMID-NER testing dataset, and it outperforms other methods significantly. Experiments on CMID-NER suggest that our method can produce promising online biomedical entities across different corpora.

Table 4
BERT models performance on each entity.

Model	BERT			MedicalQA-BERT			DKD-BiomedNER		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Disease	0.8617	0.8433	0.8808	0.8660	0.8574	0.8747	0.8800	0.8713	0.8889
Treatment	0.8899	0.8866	0.8933	0.8874	0.8816	0.8933	0.9098	0.9072	0.9124
Syndrome	0.7175	0.7004	0.7355	0.7376	0.7016	0.7776	0.7544	0.7290	0.7816
Checkup	0.7512	0.7368	0.7662	0.7726	0.7596	0.7861	0.8010	0.7913	0.8109
People	0.9022	0.8904	0.9144	0.8976	0.8692	0.9279	0.9079	0.8846	0.9324
Position	0.6690	0.6879	0.6510	0.7261	0.7143	0.7383	0.7291	0.7267	0.7315

Table 5
Performance comparison between pre-training corpora.

Method	F1	Precision	Recall
BERT	0.8142	0.7931	0.8364
BERT-Clinical	0.8150	0.8126	0.8173
BERT-MedBook	0.8163	0.8051	0.8278
TCM-BERT	0.8203	0.8198	0.8208
MCSCSet-BERT	0.8195	0.8215	0.8175
MedicalQA-BERT	0.8243	0.8058	0.8436

4.2. Ablation studies and analysis

To investigate how knowledge and pre-training affect the online biomedical NER task, we made a model with knowledge label only (KL-BERT) and QA pairs pre-training only (MedicalQA-BERT). Additionally, we use a Chinese word segmentation tool, Jieba,[5] to replace multi-channel knowledge labeling (DKD-BiomedNER (Jieba)) for generating knowledge labels. Results are presented in Table 3.

By comparing the BERT model with the multi-channel knowledge labels (KL-BERT) and the original BERT model, the multi-channel knowledge labels method can significantly improve BERT performance in a larger scope. Strikingly, with the support of additional knowledge labels, the BERT model that was pre-trained with medical QA pairs had an outstanding performance, which proved knowledge labels and medical QA pairs can significantly improve BERT model performance on the biomedical NER task. Interestingly, medical question-answering pairs pre-training improve BERT performance in recall score, and the ability of knowledge labels enhances the BERT reflected in its precision score.

We present three pre-trained models' performance on each entity in Table 4. MedicalQA-BERT has additional improvement for *Syndrome* and *Checkup* over BERT for entity recognition, which implies question-answer pairs pre-training help the model improve its ability on medical terms recognition. We observed that in the entity recognition of *Disease*, *Syndrome*, *Checkup*, and *Position*. The transformer model that is powered by data and knowledge-driven method (DKD-BiomedNER) shows obvious improvement compared with the BERT model and the BERT model pre-trained with Medical QA pairs (MedicalQA-BERT). The knowledge labels may help the BERT model to process complex biomedical entities and generate proper labels according to medical knowledge.

We further discussed the effects of different pre-training corpora on our dataset and show experiment results in Table 5. The medical QA pre-training improved BERT performance on the medical NER dataset obviously. BERT-Clinical is the BERT model that was pre-trained with clinical notes from the CCKS-2017 test.[6] BERT-MedBook is the model that was pre-trained with a medical book.[7] In addition, we also used pre-trained weights from a BERT model (TCM-BERT) that trained with traditional Chinese medicine corpora [9] and a model (MCSCSet-BERT) that trained with a large-scale specialist-annotated dataset [22] for comparison.

Compared with BERT models that only pre-trained with non-medical text, clinical notes or medical books only help the BERT get limited improvement. Our method that used question-answer pairs to pre-train the BERT model helps BERT improve over the NER task for online medical questions. The difference summarizing between different medical corpora was shown in Table 6. Because online QA pairs can be easily collected and obviously improve BERT model performance, they are extraordinarily suitable for BERT model pre-training for NLU tasks on online medical text. In the online biomedical NER task, our pre-trained model (MedicalQA-BERT) outperformed reported models (TCM-BERT and MCSCSet-BERT) that trained with different medical corpora, which proves the effectiveness of online QA pairs pre-training.

5. Discussion

Although it is controversial that neural network models can "learn" knowledge from training data, we enhance the pre-trained transformer capacity on online text's biomedical NER by incorporating knowledge representations and pre-training them with domain corpora. In our study, we show the complexity of online biomedical terms, which could be an obstacle for the building named entity recognition systems for online Chinese medical texts, and we provide a multi-channel knowledge labeling method that labels characters that are composed of Chinese medical terms. Because using medical text can improve transformer models' performance on online medical NLU, we proposed a method that using online medical question-answering pairs to pre-train the BERT model. Our pre-training strategy is effective because it provides a large source of patient and professional expression of medical phrases, so the BERT model showed significant improvement over biomedical named entity recognition. Our experiment results proved that knowledge labels and QA pairs pre-training achieve promising results for biomedical NER for online medical/health text. Moreover, the obvious improvement for biomedical NER implies domain corpora and knowledge representations could further extend transformer-based models' ability to medical natural language understanding. Our study further discusses the performance improvement for different pre-training corpora, and we found suitable pre-training corpus is critical for natural language understanding of online medical texts. Using question-answer pairs to pre-train the BERT model is more suitable than clinical notes and medical books for the NER of online questions text.

Despite our method showing its effectiveness in online biomedical NER, it still has some limitations. First, the knowledge label method that we proposed is limited by the scope we set. In our study, we use ten knowledge labels, but it is hard to apply when we encounter some knowledge that cannot be described by these labels. Second, we use concatenation to fuse features from the knowledge and the pre-trained transformer model, and using multi-head attention to fuse knowledge and encoded text features may produce a better result. Third, our method's performance on clinical NER datasets could be evaluated in future studies.

6. Conclusion

Extracting biomedical named entities from online medical text is a challenging task. In this study, we provide a knowledge-driven and

[5] https://github.com/fxsjy/jieba
[6] https://www.biendata.xyz/competition/CCKS2017_2/
[7] Internal Medicine (7th Edition).

Table 6
Characteristics of different types of medical text.

Corpora	Clinical notes	Online questions	Book/Publications
Resource	Relatively rich	Rich	Relatively poor
Accessibility	Ethics/privacy restriction	Less/no restriction	Copyright restriction
Sentences Size	Short	Short, middle	Relatively long
Medical terms	Relatively standard	Not standard	Strictly standard
Accuracy	Accurate	Unclear and indistinct	Precise and accurate
Specialization	Middle	Low	High

data-driven method to enhance the transformer model capacity for biomedical named entity recognition. The knowledge-drive method uses a knowledge base to label sentences and generates knowledge labels for each character; The data-driven method uses online-collected medical question-answering pairs to pre-trained the BERT model by the MLM task and the QAP task. By the collaboration of these methods, our model shows outstanding performance for online medical named entity recognition in F1, precision, and recall metrics. The superiority of our model for biomedical NER can help medical professionals and technology companies build automated medical services for patients and mine medical and health information from social media.

CRediT authorship contribution statement

Lulu Cao: Conceptualization, Formal analysis, Funding acquisition, Methodology, Writing – original draft, Writing – review & editing. **Chaochen Wu:** Conceptualization, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **Guan Luo:** Project administration, Supervision, Writing – review & editing, Writing – original draft. **Chao Guo:** Resources, Supervision. **Anni Zheng:** Resources.

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.

Acknowledgment

Funding

This work was supported by grants from the National Natural Science Foundation of China (82201988). This work was supported by grants from Peking University People's Hospital Research and Development Funds (RDJP2022-01).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.artmed.2024.102813.

References

[1] He Y, Kayaalp M. Biological entity recognition with conditional random fields. In: AMIA annual symposium proceedings, vol. 2008. American Medical Informatics Association; 2008, p. 293.
[2] Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. A comprehensive study of named entity recognition in Chinese clinical text. J Am Med Inform Assoc 2014;21(5):808–14.
[3] Ji B, Liu R, Li S, Tang J, Yu J, Li Q, et al. A BiLSTM-CRF method to Chinese electronic medical record named entity recognition. In: Proceedings of the 2018 international conference on algorithms, computing and artificial intelligence. 2018, p. 1–6.
[4] Luo L, Yang Z, Yang P, Zhang Y, Wang L, Lin H, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics 2018;34(8):1381–8.
[5] Tang B, Wang X, Yan J, Chen Q. Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF. BMC Med Inform Decis Making 2019;19(3):74.
[6] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018, arXiv preprint arXiv:1810.04805.
[7] Weber L, Münchmeyer J, Rocktäschel T, Habibi M, Leser U. HUNER: improving biomedical NER with pretraining. Bioinformatics 2020;36(1):295–302.
[8] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.
[9] Yao L, Jin Z, Mao C, Zhang Y, Luo Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. J Am Med Inform Assoc 2019;26(12):1632–6.
[10] Jin Q, Dhingra B, Cohen W, Lu X. Probing biomedical embeddings from language models. In: Proceedings of the 3rd workshop on evaluating vector space representations for NLP. 2019, p. 82–9.
[11] Liu W, Zhou P, Zhao Z, Wang Z, Ju Q, Deng H, et al. K-bert: Enabling language representation with knowledge graph. 2019, arXiv preprint arXiv:1909.07606.
[12] Yuan Z, Liu Y, Tan C, Huang S, Huang F. Improving biomedical pretrained language models with knowledge. In: Proceedings of the 20th workshop on biomedical language processing. Online: Association for Computational Linguistics; 2021, p. 180–90.
[13] Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Med Inform Decis Making 2019;19(3):71.
[14] Chen N, Su X, Liu T, Hao Q, Wei M. A benchmark dataset and case study for Chinese medical question intent classification. BMC Med Inform Decis Making 2020;20(3):1–7.
[15] Chowdhury S, Dong X, Qian L, Li X, Guan Y, Yang J, et al. A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records. BMC Bioinform 2018;19(17):499.
[16] Zhang Y, Yang J. Chinese NER using lattice LSTM. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers). 2018, p. 1554–64.
[17] 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.
[18] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013, arXiv preprint arXiv:1301.3781.
[19] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. 2013, p. 3111–9.
[20] Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. 2016, arXiv preprint arXiv:1607.01759.
[21] Xu B, Xu Y, Liang J, Xie C, Liang B, Cui W, et al. CN-DBpedia: A never-ending Chinese knowledge extraction system. In: International conference on industrial, engineering and other applications of applied intelligent systems. Springer; 2017, p. 428–38.
[22] Jiang W, Ye Z, Ou Z, Zhao R, Zheng J, Liu Y, et al. Mcscset: A specialist-annotated dataset for medical-domain Chinese spelling correction. In: Proceedings of the 31st ACM international conference on information & knowledge management. 2022, p. 4084–8.