CLINER: Clinical Interrogation Named Entity Recognition

Jing Ren[1,*], Tianyang Cao[1,*], Yifan Yang[2], Yunyan Zhang[2], Xi Chen[2], Tian Feng[1], Baobao Chang[1(✉)], Zhifang Sui[1(✉)], Ruihui Zhao[2], Yefeng Zheng[2], and Bang Liu[3(✉)]

[1] Key Laboratory of Computational Linguistics, Peking University, MOE, China
[2] Tencent Techonology Inc.
[3] University of Montreal
{rjj, ctymy}@pku.edu.cn, {tobyfyang, yunyanzhang, jasonxchen}@tencent.com, {fengtian0808, chbb, szf}@pku.edu.cn, {zacharyzhao, yefengzheng}@tencent.com, bang.liu@umontreal.ca

Abstract. The automatic generation of electronic medical record (EMR) data aims to create EMRs from raw medical text (e.g., doctor-patient interrogation dialog text) without human efforts. A critical problem is how to accurately locate the medical entities mentioned in the doctor-patient interrogation text, as well as identify the state of each clinical entity (e.g., whether a patient genuinely suffers from the mentioned disease). Such precisely extracted medical entities and their states can facilitate clinicians to trace the whole interrogation process for medical decision-making. In this work, we annotate and release an online clinical dialog NER dataset that contains 72 types of clinical items and 3 types of states. Existing conventional named entity recognition (NER) methods only take a candidate entity's surrounding context information into consideration. However, identifying the state of a clinical entity mentioned in a doctor-patient dialog turn requires the information across the whole dialog rather than only the current turn. To bridge the gap, we further propose CLINER, a **CL**inical **I**nterrogation NER model, which exploits both fine-grained and coarse-grained information for each dialog turn to facilitate the extraction of entities and their corresponding states. Extensive experiments on the medical dialog information extraction (MIE) task and clinical interrogation named entity recognition task show that our approach shows significant performance improvement (3.72 on NER F1 and 6.12 on MIE F1) over the state-of-art on both tasks.

Keywords: Clinical named entity recognition · Information extraction · Coarse-grained and fine-grained context · Historical pattern memory · BERT

1 Introduction

Electronic medical records (EMR) are widely used in modern health care information systems to store the information concerning individual health histories.

*Equal contribution.

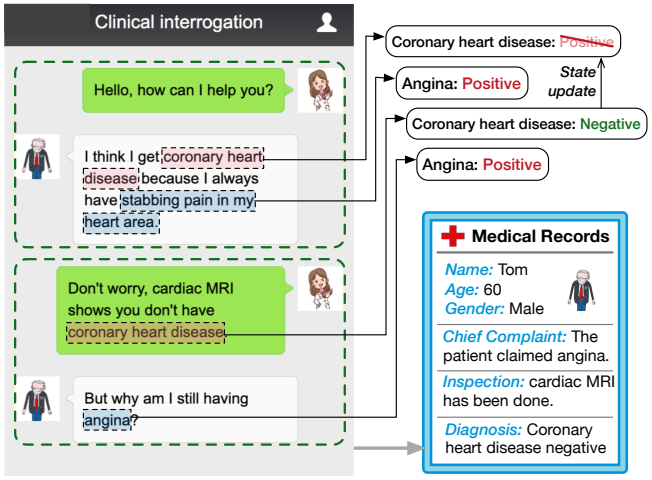

Fig. 1. Clinical interrogation named entity recognition helps medics automatically generate EMRs from the doctor-patient dialogs.

While EMRs play a key role in modern health care systems, it is an exhausting and time-consuming task for doctors to write an EMR for a patient. It was reported that the time that doctors spend on administrative work is almost twice as much as the time spent on consultation with patients, and the most time-consuming part is manually creating EMRs [21]. To relieve the heavy burden on doctors, the task of automatically converting doctor-patient dialogs into EMRs has become an emerging field in natural language processing domain in recent years [3, 19, 23]. To generate high-quality EMRs from doctor-patient dialogs, a core research problem is how to accurately extract the medical entities and their corresponding status from medical dialogs.

However, existing research focuses on extracting medical information in a turn-level context without considering the global consistency of entities' information. Taking Fig. 1 as an example for illustration, the patient initially thought he had coronary heart disease, but the doctor eventually overturned the hypothesis. In this case, the doctor will write "coronary heart disease: negative" on the EMR, where "coronary heart disease" is a medical named entity, and "negative" is the entity's state. However, existing approaches [6, 13, 28] only concentrate on extracting the medical items that are expressed in current turns and ignore their states, while the entity state (e.g., presence or absence of a symptom/disease) is critical for automatic generation of EMRs.

To bridge the gap between existing research and real-world scenarios of medical dialog, in this paper, we focus on accurately recognizing medical entities and their states that are expressed according to the entire dialog by considering both turn-level and dialog-level information. Specifically, we define the **state** of a medical item as its genuine condition expressed according to the dialog, including "positive", "negative" and "unknown", among which "positive" indicates that the patient actually suffers from one certain symptom or the doctor's diagnosis

confirms one has the disease, while "negative" does the opposite. Specially, "unknown" refers to uncertain condition, i.e., medical report hasn't been prepared or the doctor is explaining some general scientific knowledge of the disease. We then define the state update mechanism. The entity state should comprehensively consider the surrounding dialog turns other than only focus on the current turn. Take the Fig. 1 for illustration, the patient perceives he may get Coronary heart disease, however, the doctor's decision denies his hypothesis, thus the correct state should be "negative".

Based on the above principles, we first annotate and release a dataset for **CL**inical **I**nterrogation **N**amed **E**ntity **R**ecognition (CLINER) considering the state update mechanism as illustrated in Fig. 1. Our dataset distinguishes from the conventional NER datasets since the state of each entity is not simply determined by the context within the current turn, but may change as the dialog proceeds. Besides, one certain named entity can have a massive quantity of variants as the colloquial expressions in our dataset are much more diverse than the formal writing, where it also hinders the conventional NER tools to obtain promising results. For example, in Fig. 1, both *stabbing pain in my heart area* and *angina* refer to a same entity.

To address the two aforementioned challenges, we further propose a novel NER model to recognize named entities from interrogation dialogs. Specifically, we define a window as a dialog turn between the doctor and patient (e.g., we have two dialog windows in Fig. 1). Our model integrates multi-level context information from the dialog, to be specific, label-window interaction and inter-window interaction is combined to enhance the encoding representation of the current window. We then adopt a two-stage prediction manner. Firstly the model classifies whether the window contains a specific type of entity, then it determines the start and end positions of entity spans within the current window.

We evaluate the model on our dataset with two separate tasks: medical dialog information extraction (MIE) and clinical interrogation named entity recognition (CLINER). Experimental results show that the proposed model outperforms the state-of-the-art results by 3.72% and 6.12% F1 score in MIE and NER tasks, respectively.

2 Related Work

Named entity recognition (NER) [12, 27, 17, 18] is a well-studied field in the natural language processing community. There are also many efforts in medical NER from either EMRs or medical literature. The early attempts in the medical NER [8] tended to apply the conditional random field (CRF) model along with pre-defined features from a small labeled dataset. Recent works extract medical entities from the literature by introducing pre-trained models. For example, [26] addressed Chinese clinical NER task by fine-tuning BERT [4] on coronary arteriography reports. [2] leveraged data from the CCKS competitions and achieved state-of-the-art result with a BERT-BiLSTM-CRF model. [5] proposed ZEN, a BERT-based Chinese text encoder enhanced by n-gram representations, where

different combinations of characters were considered during training. However, previous attempts suffer from the limitations in text resource types [8] or predefined medical named entity categories [25]. Some other attempts tackled medical NER in different ways. [30] investigated a convolutional attention network for Chinese NER, capturing the information from adjacent characters and sentence contexts. [9] leveraged second-order lexicon knowledge of each character in the sentence, in order to provide more lexical word information including semantic and word boundary features for insufficient word information in Chinese NER.

The research on extracting information from medical dialog text just emerged in the past few years, thus we only investigate a few representative research works in this paper. Due to the limitation of available datasets, most of the works only aim to extract the symptom from the dialogue. [7] proposed a pipeline that mostly focus on the knowledge extraction module which combines rule-based methods with supervised machine learning methods. [6] aimed at extracting symptoms and their corresponding status with 186 symptoms and 3 pre-defined status. [13] annotated Chinese online medical dialogs with BIO (i.e., begin, inside, or other) schema but without considering the states, which is incompatible with the real scenario.

The most relevant work is MIE [28]. It proposed a more detailed annotation schema that contains 4 categories, 71 items and 5 status, as well as a pipeline model to classify the whole labels iteratively. Although our annotation label is consistent with MIE, there are two major differences: 1) MIE only provides the coarse-grained annotations on window-level without exact span annotation of an entity while we focus on annotating the named entities in each sentence. 2) MIE deploys an unreasonable state update mechanism that the later annotation states will blindly override the former ones. Based on our observation of the MIE dataset, it is clear that this mechanism works only for a small set of entities and is unreasonable for others.

3 Dataset and Annotation

In this section, we elaborate on our dataset and the annotation procedure. Our dataset originates from a Chinese online health community,[§] where patients can submit their health problems and then doctors kick off a conversation to communicate with the patients and provide professional suggestions. Compared to conventional NER datasets, our proposed dataset focuses on medical dialog text along with massive domain-specific items appearing in the text. It is a more challenging benchmark dataset since there are plenty of colloquial expressions amongst the dialog, which could be also annotated as entities. The dataset contains 5 categories, 72 types of entities as well as 3 types of states. Each entity is associated with a state and two labels with the same entity but different states are considered as different rather than progressive.

The most frequent types include *symptom:hyperglycemia, symptom:arrhythmia*, and so on. The detailed top-20 frequent labels can be found in Fig. 2.

[§]https://www.chunyuyisheng.com/. Our dataset will be released upon acceptance.

	Train	Dev	Test	Total
Interrogation dialog	3,633	756	685	5,074
Window	6,746	1,410	1,354	9,510
NER label	7,853	1,765	1,551	11,169
MIE label	6,887	1,513	1,371	9,771
Updated state	1,686	337	277	2,300

Table 1. The statistics of our proposed dataset.

As shown in Table 1, our dataset includes 5,074 clinical interrogation text dialog along with 11,169 NER labels and 9,771 MIE labels. The train/dev/test is splited as listed in Table 1. Since we focus on the transition of states for each label, we also count the number of states being updated, which are 2,300 in our dataset. Due to the inherent data imbalance problem, 21 of its 72 entity types only have fewer than 10 annotated instances.

An important characteristic of our dataset is that we take into consideration the fact that the state of entities may be changed along with the clinical interrogation. Based on our observation to the real data, the updating rules are complex and can't be listed with a finite set. We thus summarize our annotation schema as follows: 1) entity state in current window is implicitly related to both following and above window context. 2) entity state is not always updated with contextual information, e.g., above dialog turns make hypothetical description of the entity, which is redundant for state prediction. 3) state "unknown" aren't supposed to affect state "positive" and "negative", while the state transformation between "positive" and "negative" requires checking specific context.

Based on the criteria above, we invite six outsourcing staffs with medical background to participate in the dataset construction process. They cross-annotated each conversation and assigned the pre-defined labels to each entity by taking the whole dialog into account. If one entity was annotated differently by two annotators, the third annotator would be invited and give a final decision.

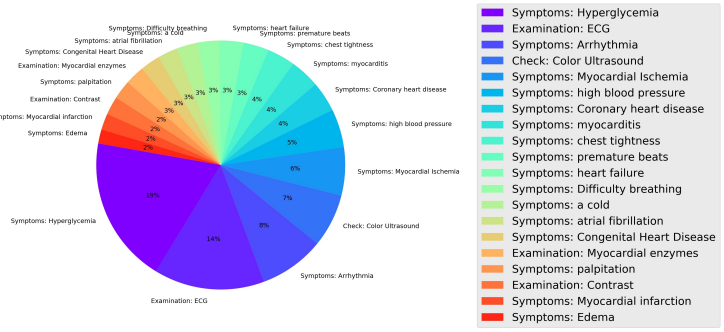

Fig. 2. Data distribution of the top 20 entity types in our dataset.

4 Proposed Method

In this section, we elaborate on the proposed CLINER, an end-to-end clinical interrogation named entity recognition model. The framework of CLINER is presented in Fig. 3, which contains five components: 1) a window encoder that converts the raw input to contextual representation; 2) a fine-grained dialog-aware aggregation module that preserves consistent entity mention semantics in the whole dialog to provide complementary information for each given token; 3) a label-aware fusion module that models the relevance of the label information with the window representation; 4) a coarse-grained global contextual aggregation module that takes the most informative following window into account; and 5) a predictor module that generates the tags for MIE task and NER task, respectively. In addition, we share the model parameters for both MIE and NER tasks, and jointly train them under a multi-task training framework.

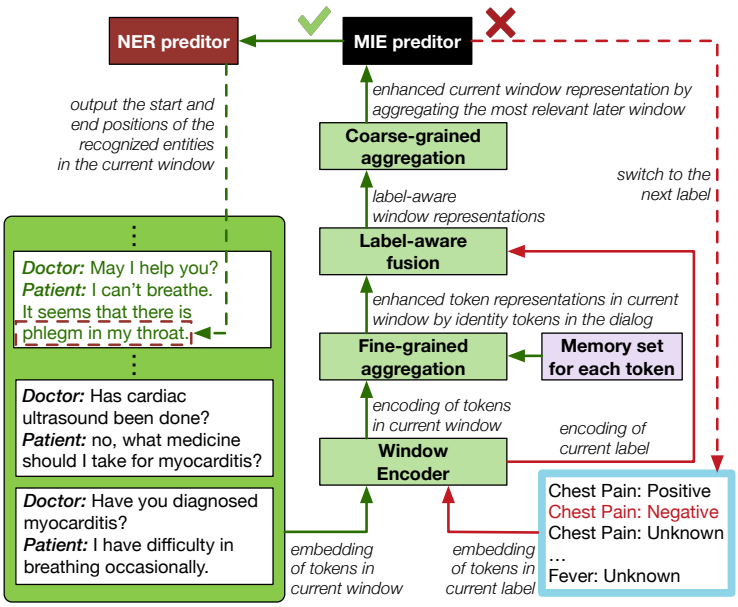

Fig. 3. Overview of the framework of CLINER.

4.1 Model Design

Window encoder. In our model, we define two consecutive utterances as a window, which is constituted by an utterance from the doctor and an utterance from the patient, respectively. Accordingly, the entire interrogation dialog $\mathcal{D}$ can be divided into multiple windows $\{X_1, X_2, ..., X_N\}$, where N denotes the number of the windows.

For each window X_i, we firstly concatenate the two utterances as a token sequence $X_i = \{x_1, x_2, ..., x_T\}$ and encode them into contextual token representations $H_i = \{h_1, h_2, ..., h_T\}$ as follows:

$$h_t = \text{Encoder}(x_t). \tag{1}$$

Different network architectures can be used for the Encoder, e.g., BiLSTM or BERT. For each candidate label l in the label set $\mathcal{L}$, we also adopt the above method to encode it into contextual semantic vector h^l.

Fine-grained dialog-aware aggregation. The state of an entity is related to all the identical entities in the entire dialog. Intuitively, we can infer that it is necessary for the model to leverage the holistic information from all other instances into the current entity in order to gain the contextual knowledge. We introduce the dialog-aware representation for each token by leveraging a key-value memory network [16]. Specifically, we define a memory set $\mathcal{M} = \{(k_1, v_1), (k_2, v_2), ..., (k_M, v_M)\}$ for each token to store all the identical tokens within the dialog, where the key k_m represents the positional information of the m-th instance, i.e., its window index amongst the dialog and its token index amongst the window. The value v_m represents the hidden state h_m of the given token. M is the number of instances. The hidden states of tokens are fine-tuned during training and used to update the value part of the memory.

For each token x_t, in order to aggregate the dialog-aware representation, its contextual representation h_t is adopted as attention key to calculate the attention scores amongst the other hidden states h_m in memory set $\mathcal{M}$ as follows:

$$s_{tm} = \frac{h_t \cdot h_m^\top}{\sqrt{d_e}}, \tag{2}$$

where d_e denotes the dimension of token embedding. Accordingly, the dialog-aware representation is computed as:

$$a_{tm} = \frac{\exp(s_{tm})}{\sum_{k=1}^{M} \exp(s_{tk})}, \tag{3}$$

$$h_t^d = \sum_{m=1}^{M} a_{tm} \cdot v_m. \tag{4}$$

Then, we integrate the original hidden state h_t with the dialog-aware representation h_t^d to form a new contextual representation g_t , which will be further fed to the following label-sentence attention module:

$$g_t = \lambda \cdot h_t + (1 - \lambda) \cdot h_t^d, \tag{5}$$

where λ is a hyper-parameter to balance the hidden state h_t and dialog-aware representation h_t^d.

Label-aware fusion. The label-sentence attention is devised to incorporate the label information h^l into the current window representation. Formally, we treat label representation h^l as a query in attention mechanism to compute the attention scores towards each contextual token representations g_t within the window. Then we can obtain the label-specific current window representation c^l as follows:

$$s_t^l = \frac{h^l \cdot g_t^\top}{\sqrt{d_e}}, \tag{6}$$

$$a_t^l = \frac{\exp(s_t^l)}{\sum_{k=1}^{T} \exp(s_k^l)}, \tag{7}$$

$$c^l = \sum_{t=1}^{T} a_t^l g_t, \tag{8}$$

where T is the sequence length of the current window. In this sense, the model is capable of determining the existence of entities of label l type in the current window, as well as predicting the entity spans.

Coarse-grained global contextual aggregation. Since the state in the current window is not only determined by the current context, but also the relevant information from the following windows, we need to take the interactions between windows into account. It is unnecessary to consider the windows prior to the current window, as the states only update based on the following interrogation text rather than the previous text.

We employ a dynamic attention mechanism to achieve this. Concretely, we take the current i-th window embedding $c_i^l = c^l$ as attention query $\mathbf{Q}$, the following window embeddings $\{c_{i+1}^l, ..., c_N^l\}$ as key matrix $\mathbf{K}$ and value matrix $\mathbf{V}$. It is so-called "dynamic" as the number of following window embeddings reduces when the dialog proceeds, and the last window does not have any following window embedding.

The following window embedding with the highest attention score c_g^l will be considered as the most informative embedding for the current window, and c_g^l will be adopted as our global contextual embedding and be concatenated to the current window embedding to facilitate predicting the state.

Formally, given the current window embedding c_i^l, we select the most informative following window by:

$$s_{ij}^l = \frac{c_i^l \cdot {c_j^l}^\top}{\sqrt{d_e}}, \tag{9}$$

$$a_{ij}^l = \frac{\exp(s_{ij}^l)}{\sum_{k=i+1}^{N} \exp(s_{ik}^l)}, \tag{10}$$

$$c_g^l = c^l_{\arg\max_j(a_{ij}^l)}, \tag{11}$$

where $i+1 \leq g \leq N$. Then the window embedding c_g^l with highest attention score is concatenated to our current window embedding c_i^l to form the global embedding c_G^l:

$$c_{i,G}^l = [c_i^l; c_g^l], \tag{12}$$

where ";" denotes concatenation operation, and c_g^l is set as zero vector if the current window is the last one in the interrogation text.

Predictor

MIE predictor. The output of the global contextual aggregation module $c_{i,G}^l$ is fed into this module. Specifically, we iterate over each MIE label $l \in \mathcal{L}$ and adopt a binary classifier to predict the MIE label:

$$\widetilde{y}_i^l = \text{Sigmoid}(\text{FFN}(c_{i,G}^l)), \tag{13}$$

where $\text{FFN}(\cdot)$ is a feed-forward neural network. A positive result indicates that at least one entity of type l exists in the current i-th window.

NER predictor. For each window, we first predict the MIE labels, then we align MIE labels to the corresponding spans within the current window as our predicted entities. Given an MIE label l, we adopt a PointerNet [24] to obtain the start and end positions of each entity:

$$\widetilde{y}_{i,t,start}^l = \text{Sigmoid}(\text{FFN}([g_{i,t}; h_l])), \tag{14}$$

$$\widetilde{y}_{i,t,end}^l = \text{Sigmoid}(\text{FFN}([g_{i,t}; h_l])), \tag{15}$$

where i denotes the i-th window.

Training In our model, we jointly train the MIE and NER tasks and optimize the cross entropy loss function as the following:

$$\mathcal{L}_{\text{MIE}} = \sum_{l \in \mathcal{L}} \sum_{i=1}^{N} y_i^l \log \widetilde{y}_i^l, \tag{16}$$

$$\mathcal{L}_{\text{NER}} = \sum_{p \in \{s,e\}} \sum_{l \in \mathcal{L}} \sum_{i=1}^{N} \sum_{t=1}^{T} y_{i,t,p}^l \log \widetilde{y}_{i,t,p}^l, \tag{17}$$

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{MIE}} + \mathcal{L}_{\text{NER}}, \tag{18}$$

where N is the number of windows in the dialog, and s, e represent the start and end positions, respectively.

5 Experiments

In this section, we carry out experiments on both NER and MIE tasks on the proposed dataset.

5.1 Baselines and Evaluation Metrics

For the NER task, on the one hand, we compare our proposed model with the traditional sequence labeling models, including **BiLSTM-CRF**, **BERT-SOFTMAX**, **BERT-CRF**, **BERT-SPAN**, and **BERT-LSTM-CRF**. On the other hand, we also compare our model with the state-of-the-art NER methods in both the general domain and Chinese medical domain [14, 20, 22, 11, 29]. Due to lack of source code and the difficulty of model replication, several Chinese dialog NER models are not selected as our baselines [7, 6, 13]. We report Precision/Recall/F1 score for evaluation.

For the MIE task, we only use the dialog-level metric, as window-level metric contains massive label redundancy, in which each label may be counted several times for evaluation. Specifically, we merge the results of the windows that belong to the same clinical interrogation, then we evaluate the results of each interrogation text.

5.2 Experimental Settings.

For a fair comparison, our setups are basically consistent with MIE [28]. We use 300-dimensional Skip-Gram [15] word embeddings pre-trained on medical dialog corpora from a Chinese online health community. Additionally, we also use Chinese-BERT-wwm [1] as our pre-trained model. The size of the hidden states of both feed-forward network and Bi-LSTM is 400. Adam is adopted for optimization [10], and we use dropout and $L2$ weight regularization to alleviate the overfitting problem and adopt early stopping using the F1 score on the development set. All experiments are run for three times and the averaged score is reported to achieve reliable results.

5.3 Results and Analysis

NER results. Table 2 shows the precision, recall and F1 score of our model and other baselines on the test set of the NER task. We observe that the BERT-based baselines achieve about 45% F1 score on our dataset. The current state-of-the-art models achieve similar results ranging from 45.42% to 48.13% in F1 score. Furthermore, by fully exploiting the information in the scope of the entire dialog, our proposed CLINER-LSTM outperforms the baseline models and obtained 48.61% F1 score even though we do not utilize pre-trained models. Finally, our CLINER-BERT gains the state-of-the-art result with 52.01% precision, 51.70% recall and 51.85% F1 score, respectively.

We also evaluate our model and baselines on the subset of the test set which only contains samples with updated states. As shown in Table 2, our model outperforms the baseline models and achieves the best performance of 49.06% F1 score. This is because our model involves more useful features from the whole dialog via the designed fine-grained and coarse-grained aggregation models.

MIE results. The experimental results are shown in Table 3. Both MIE-single and MIE-multi models obtain better results than the Plain-classifier model,

which indicates that MIE architecture is more effective than a basic LSTM representation method. Compared to the baseline model in MIE, our model can not only capture the interactions between utterances and labels but also integrate the information from the following windows. Therefore, our proposed CLINER-BERT achieves the state-of-the-art results with 88.95% and 63.97% F1 scores in category level evaluation and category-and-state level, respectively. Even if we utilize the BiLSTM as model encoder as they did in the baselines, our CLINER-LSTM still outperforms all these baselines by 2.78% and 0.58% F1 scores, respectively.

	Full test set			**Test set with updated states only**		
	Prec. (%)	Rec. (%)	F1 (%)	Prec. (%)	Rec. (%)	F1 (%)
LSTM-CRF	35.64	41.07	38.16	31.91	42.57	36.48
BERT-CRF	41.19	46.28	43.59	37.44	47.05	41.70
BERT-SOFTMAX	45.75	44.78	45.26	39.62	43.23	41.39
BERT-LSTM-CRF	42.15	49.49	45.53	37.67	49.67	42.85
BERT-SPAN	47.34	44.27	45.75	40.97	42.20	41.58
Qiu et al. [20]	44.90	45.95	45.42	39.49	43.76	41.52
FLAT [11]	44.44	48.24	46.26	41.07	45.61	43.22
Zhang et al. [29]	45.48	49.47	47.39	44.18	46.03	45.08
Sui et al. [22]	48.97	46.67	47.80	43.20	46.61	44.84
Ma et al. [14]	51.33	45.31	48.13	43.93	46.15	45.01
CLINER-LSTM	49.08	48.15	48.61	43.45	48.24	45.72
CLINER-BERT	**52.01**	**51.70**	**51.85**	**46.30**	**52.16**	**49.06**

Table 2. Experimental results of the NER task.

	Category only			**Category and state**		
	Prec.(%)	Rec.(%)	F1(%)	Prec.(%)	Rec.(%)	F1(%)
Plain-classifier	81.75	73.76	77.29	59.98	52.65	56.08
MIE-single	87.02	80.02	83.46	61.15	61.30	61.09
MIE-multi	85.32	80.48	82.83	60.30	60.78	60.54
CLINER-LSTM	88.48	84.10	86.24	59.85	63.60	61.67
CLINER-BERT	**91.02**	**86.97**	**88.95**	**62.31**	**65.72**	**63.97**

Table 3. Experimental results of the MIE task with dialog-level evaluation metric [28].

Ablation study. In this section, we estimate the effectiveness of the different model components in both NER and MIE tasks. CLINER-BERT represents the full model with all modules that achieves the best performance. The results shown in Table 4 suggest that getting rid of fine-grained dialog-aware representation deteriorates the F1 score with 1.68% and 2.30% drop. We can infer that the

	Category only			Category and state		
	Prec.(%)	Rec.(%)	F1(%)	Prec.(%)	Rec.(%)	F1(%)
CLINER-BERT	**62.31**	**65.72**	**63.97**	**52.01**	**51.70**	**51.85**
without memory	57.81	67.53	62.29	48.16	51.02	49.55
without global contextual aggregation	57.01	66.13	61.23	48.49	50.34	49.40
without label info in NER	56.34	64.40	60.10	46.92	48.54	47.72

Table 4. The ablation study with different settings.

	Positive NER			Negative NER			Unknown NER		
	P.(%)	R.(%)	F1.(%)	P.(%)	R.(%)	F1.(%)	P.(%)	R.(%)	F1.(%)
LSTM+CRF	39.04	56.20	46.07	15.68	7.76	10.39	10.01	13.71	11.56
BERT+CRF	**54.84**	50.19	52.41	41.66	27.78	33.33	31.94	34.67	33.25
CLINER-LSTM	52.32	55.69	53.95	50.01	26.38	34.55	39.46	36.87	38.12
CLINER-BERT	54.42	**57.01**	**55.68**	**52.85**	**30.14**	**38.39**	**41.26**	**38.75**	**39.97**

Table 5. State-specific evaluation on the NER task, where P., R., F. is abbreviate for Precsion, Recall and F1 score, respectively.

model performance is directly related to the information from all instances of the entity, which determines the change of the states in the entity-level. In addition, the model performance decreases by 1.06% and 0.15% without global contextual aggregation, as it incorporates the most informative window embedding in the following text to aid the current window in capturing the change of states. Finally, if we ignore the label information, the model suffers by 1.13% and 1.68% degradation in MIE and NER tasks, respectively. It indicates the label-sentence attention can capture the interaction between the current candidate label and the utterances as the most relevant tokens in the utterances will be highlighted, and NER task is strongly affected by the involvement of label information.

Case study. We analyze our model by visualizing utterance-level and window-level attention heat maps of our model on a prediction example. Fig. 4 presents the visualization of token-level and window-level attention heat maps on a prediction example of our model and baselines. The token-level attention visualization indicates that our model detects the tokens that are semantically related to the given category "myocardial infarction". We can easily find that the label "myocardial infarction" attends to the text "myocardial infarction" with the highest weight in the current window. To further determine the state, the model computes the attention score between the current window and the following windows. Window 2, which has the highest attention score, is selected as our global information to add to the state prediction. We notice in the heatmap that the token "no" is highlighted and further utilized as a crucial reference to correctly predict the state "negative" for the given label in the current window. On the contrary, conventional NER methods are impossible to predict this label properly without considering the information from window 2, which ultimately leads to the failure of the state prediction in this window.

Performance for state prediction. In this part, we analyze the effectiveness of the improvement for state prediction by our model. We carry out a quantitative

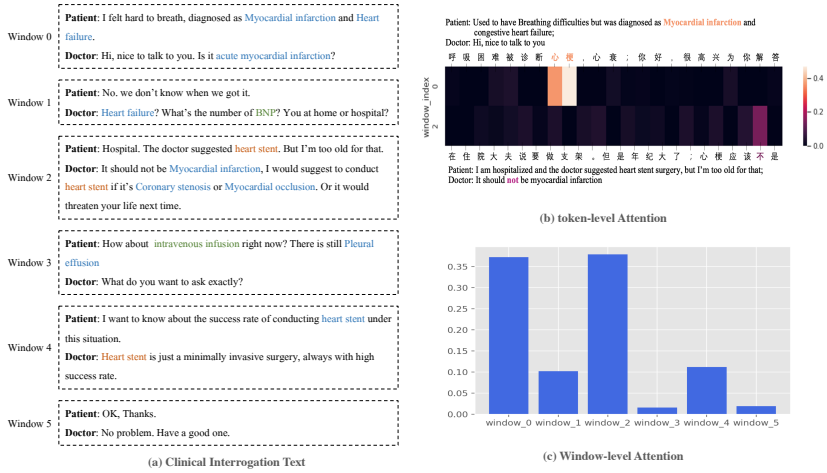

Fig. 4. The token-level and window-level attention visualization of our model on an example from the test set, which is assigned a ground-truth label "myocardial infarction: negative". (a) the clinical interrogation text; (b) token-level attention; and (c) window-level attention.

experiment to verify this. Specifically, we split the test set into three groups according to the states and evaluate them separately. The results in Table 5 show that our proposed model outperforms the baselines in NER tasks across all the state types. In particular, our model gains significant improvement over *unknown* by 6.72% and 4.87% in F1 score respectively. As "unknown" is the state that has been updated most frequently, our model can capture these variations and obtain a promising result.

6 Conclusion

In this paper, we built a clinical interrogation NER dataset, and introduced an effective model for the clinical interrogation NER task. Our proposed CLINER model better captures the update of entity states by fully exploiting the relevant context from the following windows of the current window. Experiments in both NER and MIE tasks showed that our model could effectively boost the performance and outperformed the baselines. Our research provides a promising solution for the automatic EMR generation based on clinical interrogation. For future work, we plan to further leverage the internal relations between labels and incorporate medical domain knowledge into our model.

Acknowledgement. This paper is supported by NSFC project U19A2065.

Bibliography

[1] Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., Hu, G.: Pre-training with whole word masking for chinese bert. arXiv preprint arXiv:1906.08101 (2019)

[2] Dai, Z., Wang, X., Ni, P., Li, Y., Li, G., Bai, X.: Named entity recognition using bert bilstm crf for chinese electronic health records. In: 2019 12th CISP-BMEI. pp. 1–5. IEEE (2019)

[3] DeLisle, S., Kim, B., Deepak, J., Siddiqui, T., Gundlapalli, A., Samore, M., D'Avolio, L.: Using the electronic medical record to identify community-acquired pneumonia: toward a replicable automated strategy. PLoS One **8**(8), e70944 (2013)

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

[5] Diao, S., Bai, J., Song, Y., Zhang, T., Wang, Y.: Zen: pre-training chinese text encoder enhanced by n-gram representations. arXiv preprint arXiv:1911.00720 (2019)

[6] Du, N., Chen, K., Kannan, A., Tran, L., Chen, Y., Shafran, I.: Extracting symptoms and their status from clinical conversations. arXiv preprint arXiv:1906.02239 (2019)

[7] Finley, G., Edwards, E., Robinson, A., Brenndoerfer, M., Sadoughi, N., Fone, J., Axtmann, N., Miller, M., Suendermann-Oeft, D.: An automated medical scribe for documenting clinical encounters. In: Proceedings of the 2018 NAACL. pp. 11–15 (2018)

[8] Gu, B., Popowich, F., Dahl, V.: Recognizing biomedical named entities in chinese research abstracts. In: Conference of the Canadian Society for Computational Studies of Intelligence. pp. 114–125. Springer (2008)

[9] Hu, D., Wei, L.: Slk-ner: Exploiting second-order lexicon knowledge for chinese ner. arXiv preprint arXiv:2007.08416 (2020)

[10] Kingma, D.P., Adam, B.J.: a method for stochastic optimization. 2014. arXiv preprint arXiv:1412.6980 **9** (2018)

[11] Li, X., Yan, H., Qiu, X., Huang, X.: Flat: Chinese ner using flat-lattice transformer. arXiv preprint arXiv:2004.11795 (2020)

[12] Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J.: Dice loss for data-imbalanced nlp tasks. arXiv preprint arXiv:1911.02855 (2019)

[13] Lin, X., He, X., Chen, Q., Tou, H., Wei, Z., Chen, T.: Enhancing dialogue symptom diagnosis with global attention and symptom graph. In: Proceedings of the 2019 EMNLP-IJCNLP. pp. 5036–5045 (2019)

[14] Ma, R., Peng, M., Zhang, Q., Huang, X.: Simplify the usage of lexicon in chinese ner. arXiv preprint arXiv:1908.05969 (2019)

[15] Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. arXiv preprint arXiv:1712.09405 (2017)

[16] Miller, A., Fisch, A., Dodge, J., Karimi, A.H., Bordes, A., Weston, J.: Key-value memory networks for directly reading documents. arXiv preprint arXiv:1606.03126 (2016)
[17] Nie, Y., Tian, Y., Song, Y., Ao, X., Wan, X.: Improving named entity recognition with attentive ensemble of syntactic information. arXiv preprint arXiv:2010.15466 (2020)
[18] Nie, Y., Tian, Y., Wan, X., Song, Y., Dai, B.: Named entity recognition for social media texts with semantic augmentation. arXiv preprint arXiv:2010.15458 (2020)
[19] Persell, S.D., Karmali, K.N., et al.: Effect of electronic health record–based medication support and nurse-led medication therapy management on hypertension and medication self-management: a randomized clinical trial. JAMA internal medicine **178**(8), 1069–1077 (2018)
[20] Qiu, J., Zhou, Y., Wang, Q., Ruan, T., Gao, J.: Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field. IEEE transactions on nanobioscience **18**(3), 306–315 (2019)
[21] Sinsky, C., Colligan, L., Li, L., Prgomet, M., Reynolds, S., Goeders, L., Westbrook, J., Tutty, M., Blike, G.: Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Annals of internal medicine **165**(11), 753–760 (2016)
[22] Sui, D., Chen, Y., Liu, K., Zhao, J., Liu, S.: Leverage lexical knowledge for chinese named entity recognition via collaborative graph network. In: Proceedings of the 2019 EMNLP-IJCNLP. pp. 3821–3831 (2019)
[23] Tou, H., Yao, L., Wei, Z., Zhuang, X., Zhang, B.: Automatic infection detection based on electronic medical records. BMC bioinformatics **19**(5), 117 (2018)
[24] Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. arXiv preprint arXiv:1506.03134 (2015)
[25] Wang, X., Li, J., Wu, Y., Li, J.: Bilstm-crf-based open concept relation extraction from chinese biomedical texts. Chinese Journal of Medical Library and Information Science **27**(11), 33–39 (2018)
[26] Xue, K., Zhou, Y., Ma, Z., Ruan, T., Zhang, H., He, P.: Fine-tuning bert for joint entity and relation extraction in chinese medical text. In: 2019 IEEE BIBM. pp. 892–897. IEEE (2019)
[27] Yan, H., Deng, B., Li, X., Qiu, X.: Tener: Adapting transformer encoder for named entity recognition. arXiv preprint arXiv:1911.04474 (2019)
[28] Zhang, Y., Jiang, Z., Zhang, T., et al.: Mie: A medical information extractor towards medical dialogues. In: Proceedings of the 58th ACL. pp. 6460–6469 (2020)
[29] Zhang, Y., Yang, J.: Chinese ner using lattice lstm. arXiv preprint arXiv:1805.02023 (2018)
[30] Zhu, Y., Wang, G., Karlsson, B.F.: Can-ner: Convolutional attention network for chinese named entity recognition. arXiv preprint arXiv:1904.02141 (2019)