Enhancing Dialogue Symptom Diagnosis with Global Attention and Symptom Graph

Qin Chen[1§], **Xinzhu Lin**[1§], **Xiahui He**[1], **Huaixiao Tou**[2], **Ting Chen**[3], **Zhongyu Wei**[1*]

[1]School of Data Science, Fudan University
[2]Alibaba Group, China
[3]Ping An Technology, China
{qin_chen,18210980055,18210980050,17210980029,zywei}@fudan.edu.cn
huaixiao.thx@alibaba-inc.com

Abstract

Symptom diagnosis is a challenging yet profound problem in natural language processing. Most previous research focus on investigating the standard electronic medical records for symptom diagnosis, while the dialogues between doctors and patients that contain more rich information are not well studied. In this paper, we first construct a dialogue symptom diagnosis dataset based on an online medical forum with a large amount of dialogues between patients and doctors. Then, we provide some benchmark models on this dataset to boost the research of dialogue symptom diagnosis. In order to further enhance the performance of symptom diagnosis over dialogues, we propose a global attention mechanism to capture more symptom related information, and build a symptom graph to model the associations between symptoms rather than treating each symptom independently. Experimental results show that both the global attention and symptom graph are effective to boost dialogue symptom diagnosis. In particular, our proposed model achieves the state-of-the-art performance on the constructed dataset.

1 Introduction

With the widespread use of electronic health records (EHRs) in medical treatment, symptom diagnosis based on EHRs have received a lot of attention in the natural language processing (NLP) research community (Linder et al., 2007; Shivade et al., 2013). Previous work on EHRs achieved great success in determining the diagnosis of clinical depression (Trinh et al., 2011), identifying community-acquired pneumonia (DeLisle et al., 2013), improving medication reconciliation (Persell et al., 2018) and infection detection (Tou et al., 2018).

Conversation
……
Patient: 孩子50天，嗓子有痰，而且咳嗽，是感冒吗?
My kid has been born for 50 days. He has a cough, with phlegm. Does he have a cold?
Doctor: 咳嗽，有痰，流黄鼻涕，说明有炎症了。
Coughing, sputum and a runny nose, indicate inflammation.
Doctor: 咳嗽频繁吗? 发烧吗?
Does he cough frequently? Had a fever?
Patient: 一阵一阵的，咳起来厉害，不发烧。
Occasionally, but his cough is very serious. No fever.
……
Symptom Diagnosis
True: ……有痰 (Phlegm) 咳嗽 (Cough) 流黄鼻涕 (Runny nose) 炎症 (Inflammation) ……
False: ……发烧 (Fever)……
Uncertain: ……感冒 (Cold)……

Table 1: An example of a doctor-patient dialogue and symptom diagnosis. Underlined phrases are symptom descriptions. *True*, *False*, and *Uncertain* are the inference results for whether the symptom exists in the patient.

However, EHRs usually contained historical information, such as the medical records or health records, which can not well reflect the current symptoms of the patients. In contrast, the dialogues between doctors and patients during the medical consultation process provide many valuable clues for the current symptom diagnosis. Only a few researchers focus on the dialogue between doctors and patients. (Wei et al., 2018) proposed a reinforcement learning based framework for medical dialogue system for automatic diagnosis. As shown in Table 1, a kid has a cough, the doctor asks the patient whether he has a fever. The patient describes his kid's real situation. Some symptoms like coughing appear, but some symptoms like fever don't appear. What's more, some symptoms are uncertain such as cold because doctor can not make a clear judgment at that time. Though the dialogues show great potential in med-

[§] Contributed equally
[*] Corresponding author

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing
and the 9th International Joint Conference on Natural Language Processing, pages 5033–5042,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

ical treatment, symptom diagnosis based on dialogues, namely dialogue symptom diagnosis, have rarely been studied. Moreover, there are no public datasets on dialogue symptom diagnosis as far as we know.

In this paper, we focus on the studies of dialogue symptom diagnosis and define it by two sub-tasks: symptom recognition and symptom inference. The symptom recognition aims to identify the symptom related entities from the dialogues, which is the basic step in finding symptoms or diseases. Symptom recognition is similar with the disease named entity recognition (NER) (Doğan et al., 2014) task that is generally considered as a sequence labeling problem (Chinchor and Robinson, 1997; Sang and De Meulder, 2003). Whereas, symptom recognition in dialogues is more challenging due to the short texts and nonstandard oral description. Regarding to the symptom inference, it focuses on making further decisions whether the symptom is *True*, *False*, or *Uncertain* with the patients, which helps the doctors diagnose the disease better in the next step.

To promote the research of dialogue symptom diagnosis, we collect a large amount of dialogues between patients and doctors from a Chinese online medical forum, and construct a dataset for the above two sub-tasks in dialogue symptom diagnosis. In addition, we provide several classical and advanced baselines on this dataset for further research. Furthermore, we propose an approach, which embeds a global attention and symptom graph to improve the performance of dialogue symptom diagnosis. Specifically, the global attention aims at incorporating more related information from the whole dialogue and corpus for better symptom entity representation, which will be used for symptom recognition and inference. Regarding to the symptom graph, it is built by treating each symptom as a node, and the edges are connected according to the true co-occurrence in the dialogues. We build the symptom graph to model the associations between symptoms rather than treating each symptom independently to improve the inference precision.

The contributions of this work can be summarized as follows:

- We provide a public dataset to promote the research of dialogue symptom diagnosis, which contains the annotation results in dialogues with respect to symptom recognition and symptom inference.
- We present a global attention mechanism, which captures more symptom related information from both dialogues and corpus to boost the performance of dialogue symptom diagnosis.
- We build a symptom graph to model the associations between symptoms, which further helps improve the precision of symptom inference.
- We perform extensive experiments, and the experimental results demonstrate the effectiveness of our proposed approach in the two sub-tasks of dialogue symptom diagnosis.

2 Related Work

Early attempts on biomedical NER task were based on rule-based dictionary matching method and machine learning method. (Lin et al., 2004) used maximum entropy as the underlying machine learning method incorporated with dictionary-based and rule-based methods for post-processing to identify biomedical entities. (Jimeno et al., 2008) used MetaMap which is provided from the National Library of Medicine and a dictionary matching method to identify diseases.

In recent years, researchers had proposed many neural network-based models on this problem. Most models use encoder-decoder architecture. (Collobert et al., 2011) used the convolutional neural network (CNN) as an encoder, and the conditional random field (CRF) (Lafferty et al., 2001) as a decoder. More recent works used LSTM as encoder which performed better in sequential problems, (Huang et al., 2015) used bidirectional LSTM as encoder, and the BiLSTM-CRF model achieved state-of-the-art on many datasets. Therefore, many researchers chose BiLSTM-CRF model as a baseline model when solving sequential problems. Some researchers made attempts to get better word representation. (Ma and Hovy, 2016) used an additional CNN to represent character-level features on the basis of BiLSTM-CRF. With character encoder, it can extract features inside words and get better representations.

In task of symptom NER, some symptom names entities are complex. There were many efforts to exploit features beyond individual sequences. (Yaghoobzadeh and Schütze, 2016) used knowledge base and aggregated corpus-level contextual

Patient	
Content	My kid has been born for 50 days. He has a **cough**, with **phlegm**. Does he have a **cold**? 孩 子 5 0 天 ， 嗓 子 **有 痰** ， 而 且 **咳 嗽** ， 是 **感 冒** 吗 ？
Recognition	O O O O O O O O B I O O O B I O O B I O O
Normalization	痰 咳嗽 普通感冒
Inference	True True Uncertain
Doctor	
Content	**Coughing**, **sputum** and **a runny nose**, indicate **inflammation**. **咳 嗽** ， **有 痰** ， **流 黄 鼻 涕** ， 说 明 有 **炎 症** 了 。
Recognition	B I O B I O B I I I O O O O B I O O
Normalization	咳嗽 痰 鼻流涕 发炎
Inference	True True True True

Figure 1: An example utterance with annotations of symptoms in BIO format (Symptom entities are in bold).

information to learn an entity's classes. To address the challenges of identifying rare and complex disease names, (Xu et al., 2019) proposed a method that incorporates both disease dictionary matching and a document-level attention mechanism into BiLSTM-CRF for disease NER. (Xu et al., 2018) used the document-level attention mechanism to capture long-range contextual dependencies and address clinical NER tasks. Symptom recognition is a very important step, but these researchers focus on symptom recognition only and do not further infer the recognized symptoms.

3 Dataset

In this section, we make a description of our dataset. We have constructed a Chinese dataset from the pediatric department of a Chinese online health community[1]. Patients can submit their health problems and then doctors start a conversation to know more about the patient and provide professional suggestions.

Annotation Symptoms reflect the abnormal state of the patient or the presence of the disease. The annotation consists of three parts, namely symptom recognition, symptom normalization and symptom inference. Figure 1 gives an example. We apply BIO (begin-in-out) schema at character level and each symptom is tagged with an extra label (*True*, *False* and *Uncertain*) which indicates whether the patient really has the symptom. Each symptom also links to the most relevant one on SNOMED CT[2] for normalization. In order to ensure the quality of the dataset, we hired three annotators with medical background. Each character is marked by two annotators and the inconsistent part is further judged by the third annotator. The Cohen's kappa coefficient (Fleiss and Cohen, 1973) among the annotators are between 91.80% and 92.71%.

Symptoms	Count	Ratio(%)
upper respiratory infection	480	23.22
functional dyspepsia	485	23.46
infantile diarrhea	546	26.42
bronchitis	556	26.90
Total	2,067	100.00

Table 2: Symptom distributions.

Description of the whole dataset	Avg	Std
# of sentence in each conversation	42.09	13.51
# of SNE in each conversation	15.14	9.53
# of character in each conversation	544.45	276.86
# of character in each sentence	6.47	14.18

Table 3: Statistics of the dataset.

Data Details Our dataset has a total of 2,067 conversations and we focus on four diseases, namely, "upper respiratory infection", "functional dyspepsia", "infantile diarrhea" and "bronchitis". The distribution of the diseases is shown in Table 2. Table 3 presents some statistics of the dataset. SNE stands for symptom named entity. Besides, the proportion of symptom status as *True*, *False* and *Uncertain* is around 63%, 12% and 25%. In order to get a reasonable comparison, we split the dataset by a 3:1:1 ratio to obtain the training set, validation set and test set[3].

[1] http://muzhi.baidu.com

[2] https://www.snomed.org/snomed-ct

[3] The dataset is available at: www.sdspeople.fudan.edu.cn/zywei/data/emnlp2019-cmdd.zip

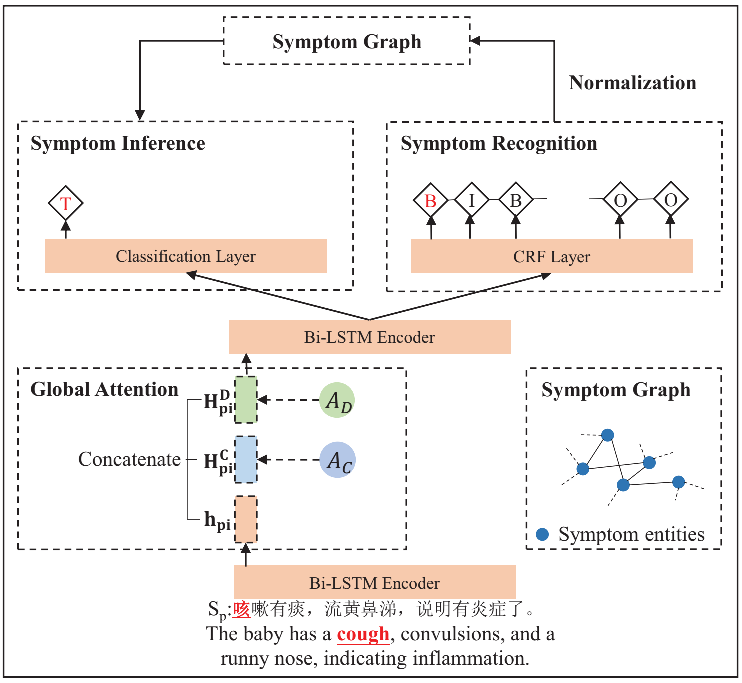
Symptom Graph
Normalization
Symptom Inference
Symptom Recognition
T
B
I
B
O
O
Classification Layer
CRF Layer
Bi-LSTM Encoder
Global Attention
$\mathbf{H}^{\mathbf{D}}_{\mathbf{pi}}$
A_D
Concatenate
$\mathbf{H}^{\mathbf{C}}_{\mathbf{pi}}$
A_C
$\mathbf{h}_{\mathbf{pi}}$
Symptom Graph
● Symptom entities
Bi-LSTM Encoder
S_p:咳嗽有痰，流黄鼻涕，说明有炎症了。
The baby has a cough, convulsions, and a runny nose, indicating inflammation.

Figure 2: Architecture of our dialogue symptom diagnosis model with global attention and symptom graph. A_D and A_C denote document-level and corpus-level attention respectively.

4 Proposed Model

The framework of our proposed model is presented in Figure 2. Our model consists of three parts, the first part is symptom recognition, the second part is symptom graph, and the third part is symptom inference. We first encode the word sequence by Bi-LSTM. Then we present a global attention mechanism to get the contextual information from document level and corpus level. Next, we re-encode the hidden states obtained above and decode by CRF to recognize the symptoms. To model the associations between disease entities in the dialogue, we construct a symptom graph, which is then incorporated into the classification layer for symptom inference. The detailed description of each step is shown in the following sections.

4.1 Bi-LSTM Encoder

In this work, we use the bidirectional long short-term memory network (Bi-LSTM) (Hochreiter and Schmidhuber, 1997) to encode the input sequences. Bi-LSTM has been widely-used to extract contextual text features. Bi-LSTM encodes the input from left to right and the same sequence in reverse (Huang et al., 2015). Given input sequence $X = (x_1, x_2, ..., x_n)$, we can get the hidden states $H = (h_1, h_2, ..., h_n)$ where h_t =Bi-LSTM(x_t). Formally, the basic units including hidden state h_t and the memory c_t are updated with following equations:

$$\begin{aligned}
\mathbf{f}_t &= \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \\
\mathbf{i}_t &= \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \\
\widetilde{\mathbf{c}}_t &= tanh(\mathbf{W}_c[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \widetilde{\mathbf{c}}_t \\
\mathbf{o}_t &= \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o) \\
\mathbf{h}_t &= \mathbf{o}_t \odot tanh(\mathbf{c}_t)
\end{aligned} \tag{1}$$

where σ is the sigmoid function and $\odot$ is the element-wise product. $\mathbf{x}_t$ is the input vector at time t. $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{o}_t$ denote the input, forget and output gate respectively.

4.2 Global Attention

Our global attention mechanism is shown in Figure 3, which consists of two parts, namely document-level attention and corpus-level attention. We will describe the details in the following.

Document-level Attention In a dialogue, the information provided by a single sentence is very limited and the same word may indicate different meanings due to the ambiguity. Therefore, we apply the document-level attention mechanism to make full use of the information in the whole dialogue to alleviate the ambiguity problem.

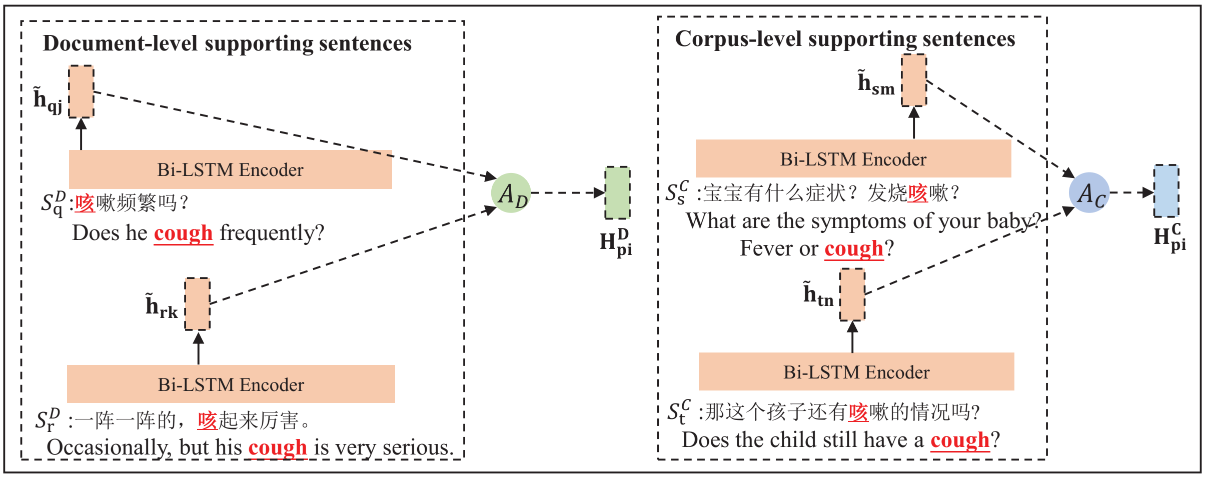
Document-level supporting sentences
$\tilde{\mathbf{h}}_{\mathbf{qj}}$
Bi-LSTM Encoder
S_q^D:咳嗽频繁吗？
Does he cough frequently?
$\tilde{\mathbf{h}}_{\mathbf{rk}}$
Bi-LSTM Encoder
S_r^D :一阵一阵的，咳起来厉害。
Occasionally, but his cough is very serious.
A_D
$\mathbf{H}_{\mathbf{pi}}^{\mathbf{D}}$
Corpus-level supporting sentences
$\tilde{\mathbf{h}}_{\mathbf{sm}}$
Bi-LSTM Encoder
S_s^C :宝宝有什么症状？发烧咳嗽？
What are the symptoms of your baby? Fever or cough?
$\tilde{\mathbf{h}}_{\mathbf{tn}}$
Bi-LSTM Encoder
S_t^C :那这个孩子还有咳嗽的情况吗?
Does the child still have a cough?
A_C
$\mathbf{H}_{\mathbf{pi}}^{\mathbf{C}}$

Figure 3: Global attention which consists of the document-level and corpus-level attentions. A_D and A_C denote document-level and corpus-level attention respectively.

We define a document (or dialogue) $D = (S_1, S_2, ...)$ and a sentence $S_p = (w_{p1}, w_{p2}, ...)$. S_p represents the pth sentence of the document and w_{pi} represents the ith word of S_p. $\mathbf{h}_{pi}$ is the hidden state of w_{pi}. We search for the sentence with the same word w_{pi} from the current document, and feed the found sentence into the same Bi-LSTM model. For example, as shown in Figure 3, w_{pi} represents the word "cough". The sentences as S_q^D and S_r^D in the current dialogue also contain "cough" . We add the hidden states of the word in the two sentences into a set $\tilde{\mathbf{h}}_{pi}$=$\{\tilde{\mathbf{h}}_{pi}^1, \tilde{\mathbf{h}}_{pi}^2, ...\}$. In Figure 3, $\tilde{\mathbf{h}}_{qj}$ and $\tilde{\mathbf{h}}_{rk}$ are $\tilde{\mathbf{h}}_{pi}^1$ and $\tilde{\mathbf{h}}_{pi}^2$ respectively. We weight the hidden states by document-level attention and the attentive representation is formulated as follows:

$$\begin{aligned} \mathbf{e}_{pi}^{D,j} &= \mathbf{v}^T \tanh(\mathbf{W}_h \mathbf{h}_{pi} + \mathbf{W}_{\tilde{h}} \tilde{\mathbf{h}}_{pi}^j + \mathbf{b_e}) \\ \boldsymbol{\alpha}_{pi}^D &= \mathbf{Softmax}(\mathbf{e}_{pi}^D) \\ \mathbf{H}_{pi}^D &= \sum_{j=1}^{k} \alpha_{pi}^{D,j} \tilde{\mathbf{h}}_{pi}^j \end{aligned} \tag{2}$$

$\mathbf{v}$, $\mathbf{W}_h$, $\mathbf{W}_{\tilde{h}}$ and $\mathbf{b}_e$ are the parameters to be learned. $\mathbf{H}_{\mathbf{pi}}^{\mathbf{D}}$ denotes the contextual information of the word w_{pi} in the dialogue.

Corpus-level Attention Noting that the same word in different dialogues may indicate additional associations, we devise a corpus-level attention mechanism to capture the extra information.

We define the corpus $C = \{D_1, D_2, D_3, ...\}$. Similar to the document-level attention, we find the supporting sentences in the corpus that contain the current word. In Figure 3, the sentences as S_s^C and S_t^C contain the word "cough", and $\tilde{\mathbf{h}}_{sm}$ and $\tilde{\mathbf{h}}_{tn}$ are the corresponding hidden states. We apply corpus-level attention to obtain the attentive representation of the hidden states in the corpus:

$$\mathbf{H}_{pi}^C = \sum_{j=1}^{k} \alpha_{pi}^{C,j} \tilde{\mathbf{h}}_{pi}^j \tag{3}$$

$\mathbf{H}_{pi}^C$ denotes the related information of the word w_{pi} in the corpus, $\alpha_{pi}^{C,j}$ is the attention weight for the corresponding hidden state in the corpus.

Both Document and Corpus-level Attention In order to integrate the information obtained based on document-level attention and corpus-level attention, we concatenate $\mathbf{h}_{pi}$, $\mathbf{H}_{pi}^C$ and $\mathbf{H}_{pi}^D$, and feed it into another Bi-LSTM model. Thus, the final hidden state of each word contains the complementary information from both the dialogue and corpus.

4.3 Symptom Recognition

In this work, we apply the Conditional Random Field (CRF) (Lafferty et al., 2001) as decoder for symptom recognition. CRF can compute the global optimal sequence and efficiently capture the dependencies among tags (e.g. label 'I' can not follow 'O') via jointly decoding the chain of labels. The Viterbi algorithm (Viterbi, 1967) is chosen for inference using dynamic programming. Given the representation of a sequence, we first map it to the tag space by a linear layer. Then, the score of the input along with a prediction y is given by:

$$s(X, y) = \sum_{i=1}^{n} (T_{y_{i-1}, y_i} + P_{i, y_i}) \tag{4}$$

where T is a transition matrix, $T_{i,j}$ represents the score of transition from tag i to tag j; P is a matrix of the output from the last layer, $P_{i,j}$ is the score of the j^{th} tag for the i^{th} word in the sentence. The goal is to predict the best tag path that given by:

$$y = argmax_{\widetilde{y}} s(X, \widetilde{y}) \tag{5}$$

4.4 Symptom Graph

Symptom entities have a certain probability of co-occurrence in a dialogue. For example, "fever" may appear in a conversation with "cold" at the same dialogue, and "cough" may appear in a conversation with "sputum" at the same dialogue. In order to capture the associations between the disease entities, we build a graph $G = (V, E)$, where $V = v_1, v_2, ..., v_m$ is the node set and $E \subset V \times V$ is the edge set. The edges $e_{i,j} = (v_i, v_j)$ in the graph is undirected. The nodes are the normalized symptom entities with status *True* from the training corpus. Edge $e_{i,j} = (v_i, v_j)$ indicates symptom entities v_i and v_j co-occur in a document. The co-occurrence number of two entities is normalized by min-max normalization to obtain the edge weight. Thus, the weight $w_{i,j} \in (0, 1)$.

4.5 Symptom Inference

Intuitively, the associations between symptom entities can help enhance symptom inference. Therefore, we first define a smoothness loss that quantitatively measures the entity associations in the constructed symptom graph:

$$S = \frac{1}{2} \sum_{i,j} w_{i,j} (y_i - y_j)^2 = \mathbf{y}^T L \mathbf{y} \tag{6}$$

where y_i is 0 or 1 that depends whether the entity has been recognized by the symptom recognition module. $L = D - A$ denotes the Laplacian matrix of an undirected subgraph G' with k nodes and m edges corresponding to the current document. $A \in R^{k \times k}$ denotes the weighted adjacency matrix of subgraph G'. D is a degree matrix and $D_{ii} = \sum_j A_{ij}$. $\mathbf{y}^T$ is a k-dimensional vector. Theoretically, if the model fully recognizes the symptoms, the value of S is 0 indicating the symptom graph is smooth. When some symptoms are not recognized, the value of S depends on the weights between the nodes of the incorrect symptoms and the neighbor nodes.

With the smoothness loss defined above, we then incorporate it into the loss function for symptom inference. Symptoms are classified into three categories: *True*, *False*, or *Uncertain*, and we adopt a softmax function in the classification layer to predict the probability of the symptom belonging to each category. The classification layer and the CRF layer share the same hidden states of the upper Bi-LSTM encoder. The objective is defined as minimizing the joint cross-entropy loss in classification and the smoothness loss in the graph:

$$loss = -\sum_{\mathcal{D}} (\sum_{i} \sum_{j} \log p^c_{i,j} - \lambda S) \tag{7}$$

where $\mathcal{D}$ is the document set, i is the index of the sentence, j is the index of the symptom, $p^c_{i,j}$ is the predicted probability of the gold-standard polarity class c for the jth symptom in the ith sentence of the document, and λ is a weight parameter to control the importance of the smoothness loss.

5 Experiments and Results

5.1 Experimental Setup

We use 200-dimensional Chinese embeddings trained on Wikipedia and fine tune them during model training by back-propagating the gradients. The parameters of the weight matrix are initialized by the Xavier method (Glorot and Bengio, 2010). The Stochastic gradient descent (SGD) method with a momentum of 0.9 is used for optimization. The initial learning rate $\eta_0 = 0.015$ and the learning rate gradually decreases with the increasing epoch. The specific update formula is $\eta_t = \eta_0/(1 + \rho t)$ where $\rho = 0.05$ and t is the number of trained epoch. Gradient clipping is set to 5 in order to avoid "gradient exploding". Other experimental settings such as the dropout rate is 0.5 and the Bi-LSTM hidden dimension is 200. We build the symptom graph from our training set. There are 1,646 edges and 162 nodes. We initialize the labels as 1 for all nodes.

Look-up Table and Stop Words For the attention mechanism, we select at most three document-level supporting sentences and three corpus-level supporting sentences. We build a look-up table that can quickly get the index of a word in each sentence and the index of a sentence in each document. Therefore, the time complexity of finding supporting sentences and words is O(1). Meanwhile, we use a stop-word list that contains 178 words. In this way, we can further reduce the time cost.

Symptom Normalization We consider the symptom normalization as a text classification problem. In our dataset, there are 162 normalized symptoms. We apply the convolution neural network (CNN) to classify all symptoms into 162 categories. The accuracy of symptom normalization on the test set is 97.04%. Thus, the symptom normalization doesn't bring too much noise to our model.

5.2 Performance of Symptom Recognition

Symptom recognition is the basis of symptom inference. We report the results of the recent advanced baselines as well as the variants of our proposed method. Specifically, we compare the performance of the following models:

- **Bi-RNNs** (**Dyer et al., 2015**): The models use LSTM or GRU for the sentence encoder, and treat symptom entity recognition as a classification problem with the softmax function. From now on, we use RNNs to denote LSTM or GRU for ease of description.
- **Bi-RNNs-CRF** (**Huang et al., 2015**): The models use RNNs for the sentence encoder and a CRF layer for decoder, which yields the tagging prediction for each token.
- **CNNs-Bi-RNNs-CRF** (**Ma and Hovy, 2016**): Compared with Bi-RNNs-CRF, the CNNs-Bi-RNNs-CRF models additionally incorporate the character level information with CNN for encoder.
- **Corpus-level Attention**: It is a Bi-LSTM-CRF model that incorporates corpus-level features via our corpus-level attention.
- **Document-level Attention**: It is a Bi-LSTM-CRF model that incorporates document-level features via our document-level Attention.
- **Both Corpus and Document-level Attention**: It is a Bi-LSTM-CRF model that incorporates both document-level and corpus-level features via our global attention.

The overall results of symptom recognition are shown in Tabel 4. We observe that the Bi-RNNs models including Bi-GRU and Bi-LSTM have the similar performance, which get about 81% in the F1 score on our dataset. The Bi-RNNs-CRF models perform much better than Bi-RNNs, which indicates the effectiveness of the CRF model for sequence tagging. In addition, the performance can be slightly improved by incorporating the character level information with CNN. Furthermore, by integrating either our corpus-level attention or document-level attention into the existing models, the performance can be significantly boosted. In particular, our model with global attention achieves the best performance in terms of all metrics.

Model	Prec.	Recall	F1
Bi-GRU	76.02%	88.09%	81.61%
Bi-LSTM	76.64%	87.60%	81.62%
Bi-GRU-CRF	86.44%	89.13%	87.77%
Bi-LSTM-CRF	89.93%	89.56%	89.74%
CNNs-Bi-GRU-CRF	87.08%	90.82%	88.91%
CNNs-Bi-LSTM-CRF	90.45%	90.48%	90.47%
Corpus-level attention	90.40%	91.02%	90.71%
Document-level attention	90.53%	91.67%	91.10%
Both Corpus and Document-level attention	91.09%	92.17%	91.62%

Table 4: Performance of various models for symptom recognition.

Method	inference of symptom	F1
Bi-LSTM CRF-inference	True	82.68%
	False	59.22%
	Uncertain	65.02%
Bi-LSTM CRF-inference with graph	True	83.79%
	False	60.04%
	Uncertain	65.80%
Our joint model	True	85.08%
	False	66.09%
	Uncertain	74.13%
Our joint model with graph	True	86.46%
	False	66.88%
	Uncertain	74.25%

Table 5: Performance of various models for symptom inference.

5.3 Performance of Symptom Inference

Table 5 presents the symptom inference results of the classical Bi-LSTM CRF-inference model and our proposed joint model (Figure 2). The results show that our proposed model with global attention significantly outperforms the Bi-LSTM CRF-inference model for symptom inference across all the categories. In particular, we achieve substantial improvements for inferring the *False* and *Uncertain* categories of symptoms, by utilizing the global information in the current dialogue and the whole corpus

To investigate the effect of the symptom graph for symptom inference, we compare the models with and without symptom graphs. The results in Table 5 show that when incorporating the symptom graphs for inference, the performance of each model can be further boosted. These observations have verified the effectiveness of modeling the associations between symptoms via graphs for symptom inference.

5.4 Qualitative Analysis

Table 6 presents a case of symptom recognition based on the baseline and our model. It is observed that the baseline Bi-LSTM CRF model only identifies the word “allergic” as a symptom. In contrast, our model can recognize the phrase “allergic rhinitis” by utilizing the related information (i.e., “allergies rhinitis” and “allergic caused rhinitis”) in the document-level and corpus-level supporting sentences, which is more accurate for the symptom description in this case.

Table 7 shows the results of symptom inference for a case by using the baseline and our joint model. From the patient’s answer, we know that the kid has no “allergy”. Whereas, the symptom “allergies” in the doctor’s question sentence is inferred as *uncertain* by the baseline. By incorporating the global attention mechanism, our joint model can correctly infer the symptom as *false*.

Model	Sentence
Bi-LSTM CRF	医生：相对来说，这个年龄的孩子出现过敏性鼻炎比较少见。 Doctor: Relatively speaking, allergic rhinitis is rare in children of this age.
Our Model	医生：相对来说，这个年龄的孩子出现过敏性鼻炎比较少见。 Doctor: Relatively speaking, allergic rhinitis is rare in children of this age.
Document-level supporting sentence	患者：如果是过敏性的鼻炎，会持续多久？ Patient: If it is allergic caused rhinitis , how long will it last?
Corpus-level supporting sentence	医生：那需要考虑过敏性鼻炎的可能。 Doctor: It may be allergic rhinitis.

Table 6: Symptom recognition results of the baseline and our methods. Underlined phrases are symptoms.

Model	Sentence
Bi-LSTM CRF- Inference	医生：孩子小时候湿疹重不重？平时易过敏吗？ Doctor: Is the child eczema serious? Is he susceptible to allergies in daily life? Inference:Uncertain Uncertain 患者：不过敏啊。 Patient: No allergy. Inference:False
Our joint model with graph	医生：孩子小时候湿疹重不重？平时易过敏吗？ Doctor: Is the child eczema serious? Is he susceptible to allergies in daily life? Inference:Uncertain False 患者：不过敏啊。 Patient: No allergy. Inference:False

Table 7: Symptom inference results of the Bi-LSTM CRF-inference model and our joint model with symptom graph. Underlined phrases are symptoms.

To have an insight of why the symptom graph can help boost symptom inference, we select several frequent symptoms in dialogues, namely “Cough”, “Sputum”, “Fever”, “Diarrhea”, “Snot”, “Cold” and “Indigestion”, and visualize the associations between the symptoms in Figure 4. The darker color indicates a larger association weight between the symptoms. We observe that the “cough” and “sputum” are highly associated, which corresponds to our intuition that the patient will probably have a cough and sputum simultaneously. To make it more clear, we show the inference results for each symptom with and without graph in Figure 5. The results show that our model with graph achieves larger improvements than that without graph for inferring the highly associated symptoms such as “cough” and “sputum”, which indicates the necessity to incorporate the symptom graph to enhance symptom diagnosis.

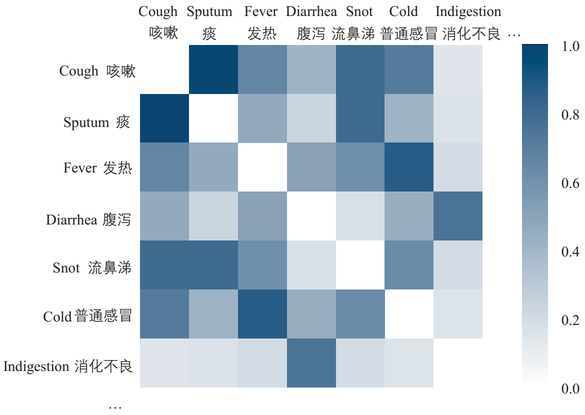

Figure 4: Symptom associations in the symptom graph.

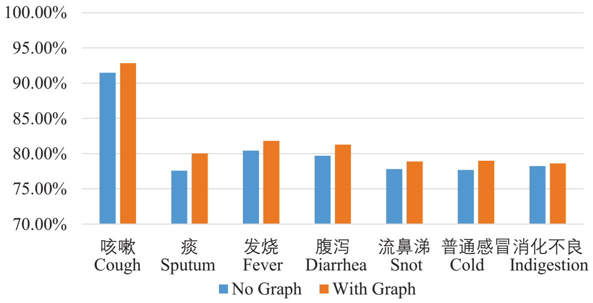

Figure 5: The impact of the symptom graph on F1 scores for symptom inference.

6 Conclusions and Future Works

In this paper, we construct a dataset for dialogue symptom diagnosis, and present a model with global attention and symptom graph for diagnosing symptoms in dialogues. Our global attention mechanism consists of the document-level and corpus-level attentions, which select supporting sentences from the current dialogue and corpus to overcome the information limitations. Experiments on our dataset show that our global attention can effectively boost the performance of dialogue symptom diagnosis. Furthermore, we build a symptom graph to model the associations between symptoms, which helps improve the performance of symptom inference.

In the future, we will build a larger symptom graph and use external medical information to further improve the performance of symptom diagnosis on dialogues.

Acknowledgments

This work is partially funded by National Natural Science Foundation of China (No. 61751201), National Natural Science Foundation of China (No. 61702106) and Shanghai Science and Technology Commission (No. 17JC1420200, No. 17YF1427600 and No. 16JC1420401).

References

Nancy Chinchor and Patricia Robinson. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, volume 29, pages 1–21.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.

Sylvain DeLisle, Bernard Kim, Janaki Deepak, Tariq Siddiqui, Adi Gundlapalli, Matthew Samore, and Leonard D'Avolio. 2013. Using the electronic medical record to identify community-acquired pneumonia: toward a replicable automated strategy. *PLoS One*, 8(8):e70944.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. 2015. Transition-based dependency parsing with stack long short-term memory. *arXiv preprint arXiv:1505.08075*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Antonio Jimeno, Ernesto Jimenez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga, and Dietrich Rebholz-Schuhmann. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. In *BMC bioinformatics*, volume 9, page S3. BioMed Central.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Yi-Feng Lin, Tzong-Han Tsai, Wen-Chi Chou, Kuen-Pin Wu, Ting-Yi Sung, and Wen-Lian Hsu. 2004. A maximum entropy approach to biomedical named entity recognition. In *Proceedings of the 4th International Conference on Data Mining in Bioinformatics*, pages 56–61. Springer-Verlag.

Jeffrey A Linder, Jun Ma, David W Bates, Blackford Middleton, and Randall S Stafford. 2007. Electronic health record use and the quality of ambulatory care in the united states. *Archives of internal medicine*, 167(13):1400–1405.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Stephen D Persell, Kunal N Karmali, Danielle Lazar, Elisha M Friesema, Ji Young Lee, Alfred Rademaker, Darren Kaiser, Milton Eder, Dustin D French, Tiffany Brown, et al. 2018. 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.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J Embi, Noemie Elhadad, Stephen B Johnson, and Albert M Lai. 2013. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230.

Huaixiao Tou, Lu Yao, Zhongyu Wei, Xiahai Zhuang, and Bo Zhang. 2018. Automatic infection detection based on electronic medical records. *BMC bioinformatics*, 19(5):117.

Nhi-Ha T Trinh, Soo Jeong Youn, Jessica Sousa, Susan Regan, C Andres Bedoya, Trina E Chang, Maurizio Fava, and Albert Yeung. 2011. Using electronic medical records to determine the diagnosis of clinical depression. *International journal of medical informatics*, 80(7):533–540.

Andrew Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-Fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 201–207.

Guohai Xu, Chengyu Wang, and Xiaofeng He. 2018. Improving clinical named entity recognition with global neural attention. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 264–279. Springer.

Kai Xu, Zhenguo Yang, Peipei Kang, Qi Wang, and Wenyin Liu. 2019. Document-level attention-based bilstm-crf incorporating disease dictionary for disease named entity recognition. *Computers in biology and medicine*, 108:122–132.

Yadollah Yaghoobzadeh and Hinrich Schütze. 2016. Corpus-level fine-grained entity typing using contextual information. *arXiv preprint arXiv:1606.07901*.