IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 20, 2020, accepted June 14, 2020, date of publication June 29, 2020, date of current version July 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3005684

Information Extraction for Intestinal Cancer Electronic Medical Records

SUFEN WANG[1], MINMIN PANG[2], CHANGQING PAN[3], JUNYI YUAN[3], BO XU[2], MING DU[2], AND HONG ZHANG[2]

[1]Glorious Sun School of Business and Management, Donghua University, Shanghai 201620, China
[2]School of Computer Science and Technology, Donghua University, Shanghai 201620, China
[3]Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding author: Changqing Pan (13386259762@189.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61906035, in part by the Shanghai Sailing Program under Grant 19YF1402300, and in part by the Artificial Intelligence Innovation and Development Project of Shanghai Economic and Information Commission under Grant XX-GRZN-01-19-6584.

ABSTRACT The data generated by the structured electronic medical records is helpful for mining and extracting medical data, and it is an effective way to make effective use of valuable data resources. However, the hospitals have accumulated a large number of unstructured data in electronic medical records, which cannot be effectively searched, resulting in serious waste of resources. In this paper, we study the problem of extracting attribute values from the unstructured text in electronic medical records. By observing intestinal cancer diagnostic texts, our attributes have two categories - discriminative attributes and extractive attributes, which use the text classification and the sequence labeling to tackle attribute values extraction problems. For discriminative attributes, we firstly divide the text into sentences/segments as instances. Secondly, we fine-tune the pre-trained word embedding to capture domain-specific semantics/knowledge. Thirdly, we also use an attention mechanism to select the most important instance for different attribute extractors. Finally, multi-tasking learning is used to share useful information to get better experimental results. For extractive attributes, we propose a novel model to get attribute values, including the BiLSTM layer, the CNN layer and the CRF layer. In particular, we use BiLSTM and CNN to learn text features and CRF as the last layer of the model. Experiments have shown that our method is superior to several competitive baseline methods.

INDEX TERMS Electronic medical records, text classification, sequence labeling, multi-task learning, attention mechanism.

I. INTRODUCTION

With the continuous development of science and technology, the research results on data have been gradually applied to various domains. In the medical domain, the data of the Electronic Medical Records (EMR) system has attracted the attention of researchers and has become the main issue of research. The EMR data contains a large number of patients' basic information, condition diagnosis reports and medical knowledge, which are valuable wealth in the medical domain. Effectively mining can support users' disease monitoring. However, only structured data can serve medical research. Therefore, the main work of this paper are to transform the unstructured intestinal cancer diagnostic text into structured data which provides clinical decision support and models the patients's health status for medical professionals. For example, in Figure. I, it shows three Chinese pathology reports of intestinal cancer specimens. The Chinese pathology report consists of multiple sentences and regulate the experimental data. At the same time, we assume that each sentence has one mention of cancer and one mention cannot span across more than one sentence. And these reports are unstructured text and contain a large number of descriptive information about the specimen, including specimen name, adenocarcinoma shape, tumor size, infiltration depth, cancer metastatic ratio, and the status of some cancer indicators. Information of those unstructured texts from electronic medical records is hard to understand for machines. In order to get more complete and useful information from electronic medical records, we need to complete the transformation from unstructured text into

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Callico.

ID	Sentences
1	直肠癌根治标本：腺癌II级（浸润型），肿瘤大小 1*0.5cm，浸润至肌层外纤维脂肪组织，侵犯神经；上切端、基底切缘均未见癌组织累及，另送“下切端”见癌组织累及；肠旁淋巴结13枚均未见癌转移。“左侧受侵犯部分卵巢”卵巢及输卵管组织伴出血、坏死及炎症反应，未见癌组织；“右输尿管旁组织”腺癌浸润/转移；另见淋巴结1枚未见癌转移。
2	“直肠”切除标本：腺癌II级（溃疡型），大小 1.5 × 1 × 0.5cm，浸润至肠壁外纤维脂肪组织，脉管内见癌栓；上切端、基底切缘均未见癌累及；找到肿块旁淋巴结4/4枚见癌转移。
3	“右半结肠根治标本”腺癌 II 级（隆起型），直径 0.5cm，浸润至外膜外纤维脂肪组织，未见侵犯神经，脉管内未见癌栓；上切端见癌组织累及，基底切缘、“下切端”均未见癌组织累及；肠旁淋巴结 4/7见癌转移 。“直肠息肉”管状腺瘤。

Classification

ID	Cancer involved in Upper Margin (CUPM,上切端癌累及情况)	Cancer involved in Bottom Margin (CBOM,下切端癌累及情况)	Cancer involved in Base Margin (CBAM, 基底切端癌累及况)	Nerve Invasion (NI,侵犯神经情况)	Vascular Invasion (VI,侵犯脉管情况)
1	NO(否)	YES(是)	NO(否)	YES(是)	UNKNOWN(未知)
2	NO(否)	UNKNOWN(未知)	NO(否)	UNKNOWN(未知)	YES(是)
3	YES(是)	NO(否)	NO(否)	NO(否)	NO(否)

Extraction

ID	Specimen Name (SN,标本名称)	Adenocarcinoma Shape (AS,腺癌形状)	Tumor Size (TS,肿瘤大小)	Infiltration Depth (ID,浸润深度)	Cancer Metastatic Ratio (CMR,癌转移比例)
1	Rectal Carcinoma Radical Specimen (直肠癌根治标本)	Infiltrated Type (浸润型)	1×0.5cm	Extra-muscularis fibrous and adipose tissue (肌层外纤维脂肪组织)	UNKNOWN(未知)
2	Rectal resection specimen (直肠切除标本)	Ulcerative Type (溃疡型)	1.5 × 1 ×0.5cm	Extra-intestinal fibrous and adipose tissue (肠壁外纤维脂肪组织)	4/4
3	Radical specimen of right hemicolectomy (右半结肠根治标本)	Eminence Type (隆起型)	0.5cm	Extra-adventitia fibrous and adipose tissue (外膜外纤维脂肪组织)	4/7

FIGURE 1. An example of intestinal cancer specimens and some extracted attribute values.

structured data. By observing intestinal cancer diagnostic texts, our attributes have two categories - discriminative attributes and extractive attributes.

For discriminative attributes, this paper uses text classification to solve the task of extracting attribute values. Extracting the status of some cancer indicators from cancer specimens includes whether the cancer is involved in the upper/bottom/base margins, whether it has invaded the nerves or vascular, etc. The status of cancer indicators has three categories: {YES, NO, and UNKNOWN}. We treat it as a typical multi-class problem. Firstly, we take each sentence/segment of the text as an instance. For each instance, the pre-trained word embedding is used to initialize parameters of the embedding layer in our neural network model. Secondly, the training data is used to fine-tune word embedding to capture domain-specific semantics/knowledge. Thirdly, considering the different importance of different instances for different attribute extractors, this paper uses the attention mechanism to select the most important instances for different attribute extractors and reduce the noise caused by other instances. Finally, the multi-tasking method is used to share useful information and reduce the risk of over-fitting to get better experimental results in the output layer.

For extractive attributes, this paper proposes sequence labeling to extract attribute values from text, including specimen name, adenocarcinoma shape, infiltration depth, tumor size, and cancer metastatic ratio, etc. In the first step, the pre-trained character embedding is used for each text to better initialize the parameters of the embedding layer in the neural network model. In the second step, the domain data is used to fine-tune the character embedding so that it can better represent the text. At the third step, BiLSTM and CNN are used together to learn text features, because CNN is good at learning and capturing spatial features, RNN is good at capturing timing features. To combine both the advantages, we concatenate the embedding from both models to get a better sentence representation. Finally, a CRF layer is used to learn the constraints of sentences, because CRF can consider the relationship of adjacent labels to give a globally optimal labeling result.

Our contributions can be summarized as follows:

- Firstly, for complex and diverse attribute values in intestinal cancer diagnostic text, the proposed solution is to use different methods and different models to solve the problem of attribute values extraction. For discriminative attributes and extractive attributes, this paper

uses the methods of text classification and sequence labeling methods to complete the task of attribute values extraction.
- Secondly, to adapt the domain data, we fine-tune the pre-trained word embedding with our electronic medical records to capture domain-specific semantics/knowledge.
- Thirdly, to tackle the data insufficient challenge, we propose a multi-task method for discriminative attribute extraction.
- Fourthly, in electronic medical record data, each attribute extractor focuses on a certain sentence in the input text. Therefore, we take each sentence in the input text as an instance and model it as a multi-instance learning model. And we use the attention mechanism to improve the effectiveness of the model.
- Finally, experiments show that our model can make full use of all informative sentences and effectively reduce the influence of useless instances.

This paper is structured as follows. In Section II, we review the related work. In Section III, we definite the problem and give an overview of our proposed model. In Section IV, we describe the details of extracting discriminative attributes from electronic medical records. In Section V, we describe the details of extracting extractive attributes from electronic medical records. In Section VI, we describe the experimental setup and discuss the results. Finally, we conclude this work in Section VII.

II. RELATED WORK

In this section, we review and summarize the works related to our research. Due to our task is to extract attribute values from unstructured medical texts, it includes information extraction, named entity recognition, text classification, and sequence labeling.

Information Extraction (IE) is to extract the content that people are interested in from unstructured or semi-structured text, and save it in a structured form, such as a relational database form or XML form. Information extraction technology can be applied to many fields, such as commodity search [37], text mining [10], biology [28], medical treatment [1], [13] and so on. Information extraction mainly includes named entity recognition [2], [24], relationship extraction [36], attribute extraction [23] and other tasks. In this paper, we study the problem of extracting attribute values from the unstructured text in electronic medical records.

Named Entity Recognition (NER) is a very important Natural Language Processing (NLP) task to the identification of entities with specific meanings in the text, including person names, place names, institution names, proper nouns, and so on. Various methods focused on a large set of fine-grained types [18]. They mainly relied on manual features [8], [30]. Neural network approaches, especially for the LSTM-CRF model [15], [16], [31], [34], can significantly improve the performance of the named entity recognition task in the medical field. But these methods are applied to an entire sentence.

Text Classification, as one of the core parts of Natural Language Processing (NLP), has attracted the attention of scholars. In the past few years, traditional machine learning has been widely used in document classification tasks [5], [33]. For example, Support Vector Machine classifier [22], [25] performed relatively well with high classification accuracy. In recent years, neural network model plays a dominant role in text classification for medical field [14], [17], [19]. And, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) based on Long Short-Term Memory (LSTM) are widely used in text classification tasks. However, those neural network models consider the entire context of a text to obtain a classification result. Our model extracts different attribute values from a text.

Sequence Labeling, such as Part-of-Speech (POS) tagging and Named Entity Recognition (NER), is one of the first stages in deep language understanding. The input of the model is a sequence of sentences expressed in natural language, and then the output of the model is a signed sequence of equal length. In recent years, Collobert *et al.* [4] proposed to use neural language model networks to pre-train word embeddings to better represent text information and complete sequence labeling problems. But they have heavy feature engineering. Zheng *et al.* [37] proposed the Opentag model, using BiLSTM to capture context information and CRF to complete sequence labeling, which is a new attention mechanism to provide interpretation for model decision-making. Similarly, the sequence labeling task is also widely used in the medical field [6], [11], [29], [32].

III. OVERVIEW

In this section, we first define our problem, and show an overview of our proposed model.

A. PROBLEM FORMULATION

The unstructured text of electronic medical records is not conducive to data mining. So our task is to extract attribute values from unstructured intestinal cancer diagnostic text, transforming unstructured text into structured data. Previous problems [14], [17], [19] either focus on extract discriminative nor extractive attributes from the unstructured text of EMR. In this paper, we focus on extract both the discriminative and extractive attributes, which is more practice in a real-world applications.

For the extraction of discriminative attribute values, we treat it as a *multi-class* problem, because the number of discriminative attribute values is known and small in the medical text. Let E is the set of cancer specimens, $S_e(maxsent\ 100)$ is the sentences of cancer specimen $e \in E$ in the Chinese pathological reporting. For each cancer specimen $e \in E$, our goal is to use attribute extractor a_i to infer the status v_i of cancer indicators i from S_e. And attributes include whether the cancer is involved in the upper, bottom, and base margins,

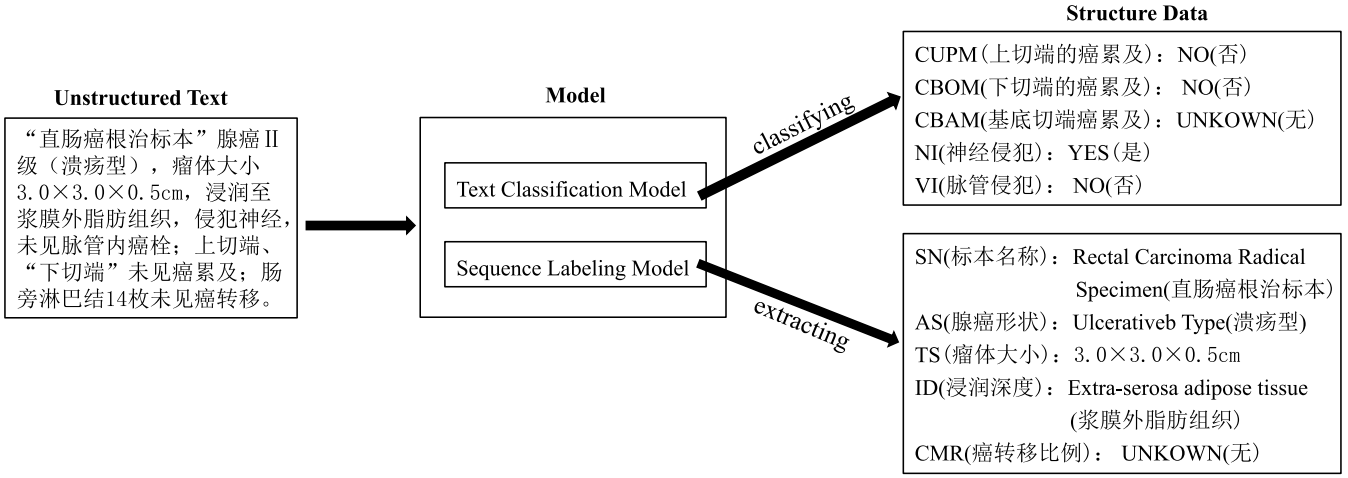


FIGURE 2. Framework for the way of extracting the attribute values from the unstructured text of electronic medical records.

whether it has invaded the nerves or vascular, etc. The status of cancer indicators has three classes: {YES, NO, UNKNOWN}.

For the extraction of extractive attribute values, we treat it as a *sequence labeling* problem, because the number of extractive attribute values is unknown and large in the medical text. Let E be the set of the cancer specimen. For each cancer specimen $e \in E$, the goal is to find the attribute value from e using the attribute tagger b_i to infer the values w_i of cancer indicators j.

B. SOLUTION FRAMEWORK

Different types of attributes use different methods to extract attribute values. The framework is shown in Figure. 2.

For discriminative attributes, we use fine-tuning, attention mechanism and multi-task learning to complete the extraction of attribute values. Firstly, we divide text into sentences and fine-tune the general word embedding to capture domain-specific semantics/knowledge. Secondly, we use the BiLSTM layer to capture the past and future features jointly to get a better instance representation. Then, considering that the importance of different instances for different attribute extractors is not equal, we use an attention layer to select the most important instances to compose text representation for different attribute extractors. Finally, the multi-tasking method is used to share useful information and reduce the risk of over-fitting to get better experimental results in output layer.

For extractive attributes, a novel neural network architecture, including the BiLSTM layer, the CNN layer, and the CRF layer. This architecture automatically learns from character representations. Firstly, we use domain data to fine-tune pre-trained character embedding to capture domain-specific semantics/knowledge, so that it can better represent text. The second step is to use BiLSTM and CNN to learn text features together, because CNN is good at learning and capturing spatial features. RNN is good at capturing timing features. To combine both the advantages, we concatenate the embedding from both models to get a better sentence representation. Finally, a CRF layer is used to learn the constraints of sentences. The CRF can solve the attribute value extraction by considering the relationship of adjacent labels, performing joint probability analysis on the labeled sequence, and giving a globally optimal label sequence.

IV. TEXT CLASSIFICATION

In this section, we mainly present the implementations of our text classification method to solve the problem of extracting the status of discriminative attributes from cancer specimens. Figure. 3 shows an example of using multi-task learning to extract the status of discriminative attributes, including CUPM, CBOM, CBAM, NI, and VI.

A. EMBEDDING LAYER

As shown in Figure. 3, for each cancer specimen $e \in E$, we first divide the text into sentences/segments S_e, and take each of them as an instance of the cancer specimen. Then for each instance, we divide the sentence into words. To avoid building complex features engineering and save a lot of computing resources, we use the pre-trained word embedding which are trained on large-scale general corpus, such as Tencent Embedding [27] and Word2vec [21]. However, since the pre-trained word embedding on general corpus cannot capture domain-specific semantics/knowledge [26], the performance of models using pre-trained word embedding during our implementation is limited. For example, the word ‘‘bottom margin’’ does not existed in any publicly pre-trained word embedding, which would affect the extraction effect of the attribute extractor of CBOM.

Hence, in this paper, we use the pre-trained word embedding to better initialize the parameters of our models, then we fine-tune them by using our domain corpus to capture domain-specific semantics/knowledge.

B. BiLSTM LAYER

Long Short Term Memory network is a cyclic neural network that adds memory function. The memory function is called

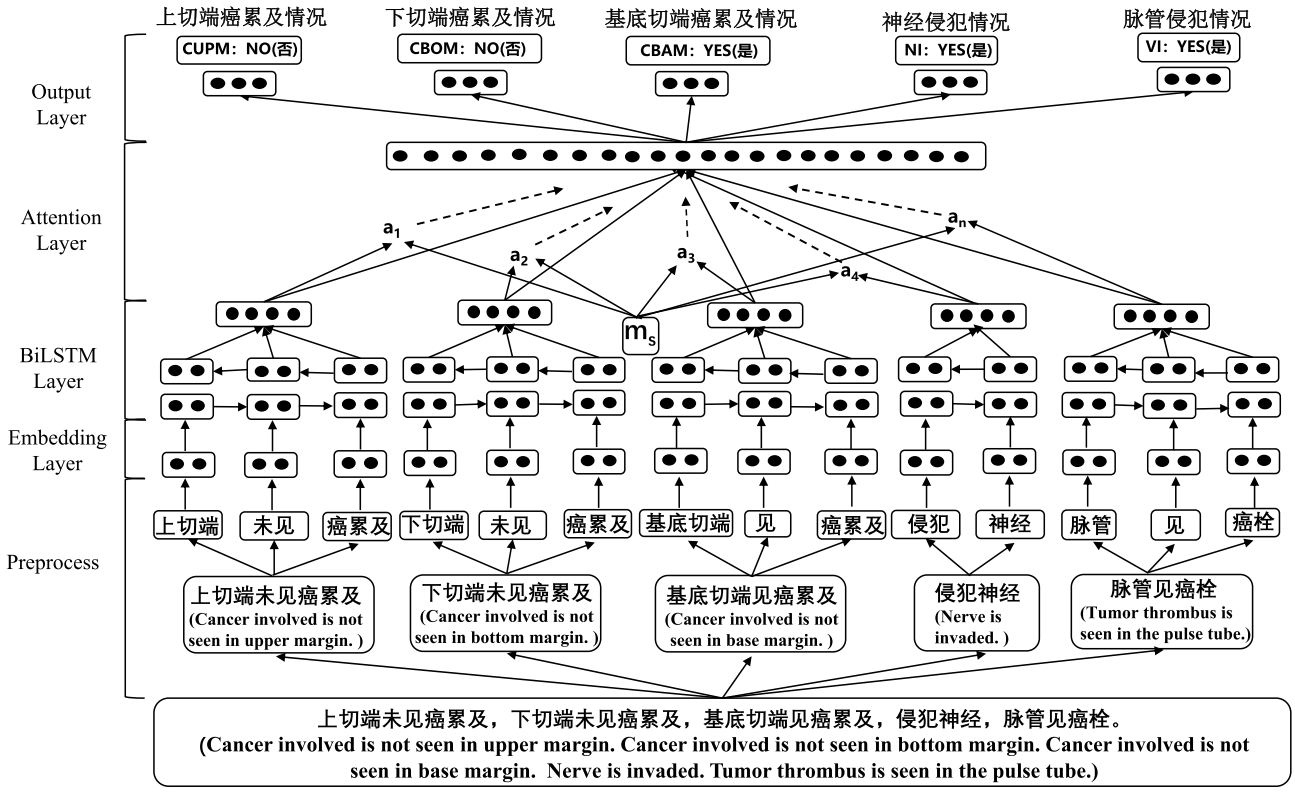


FIGURE 3. An example of using multi-task learning to extract the status of some cancer indicators from cancer specimens.

"cell", which consists of four elements: input gate I, forget gate f, output gate O, and self-circulate connected neurons. At each time step t, the LSTM formulas are as follows:

$$I_t = \sigma(W_i X_i + U_i h_{t-1} + b_i) \quad (1)$$
$$\tilde{C}_t = tanh(W_c X_t + U_c h_{t-1} + b_c) \quad (2)$$
$$f_t = \sigma(W_f X_t + U_f h_{t-1} + b_f) \quad (3)$$
$$C_t = I_t \odot \tilde{C}_t + f_t \odot C_{t-1} \quad (4)$$
$$O_t = \sigma(W_o X_t + U_o h_{t-1} + b_o) \quad (5)$$
$$h_t = O_t \odot tanh(C_t) \quad (6)$$

where σ is the sigmoid function and $\odot$ is the element-wise product. X_t is the input word embedding, W_i, W_f, W_o, W_c is the weight parameters and b_i, b_f, b_o, b_c is the bias term.

For many classification tasks, just looking at the words in front of them cannot meet all the requirements. Some tasks may need to be determined by the previous inputs and the latter inputs together, especially by the latter inputs, which will be more accurate. However, the hidden state of LSTM h_t, draws information only from the past and knows nothing about the information behind it. An effective solution and proven work [7] is a bidirectional long short-term memory neural network. The basic idea is to represent each sequence forward and backward as two independent sequences. The hidden state can capture past and future information separately. Then these two hidden states are connected to form the final output.

Therefore, we use a BiLSTM layer to capture the past and future features jointly to obtain a better sentence/instance representation [37]. For each sentence/instance S_{ei}, we use a hidden vector h_i to represent it:

$$h_i = [h_{i,1}, h_{i,2}, \ldots, h_{i,j}, \ldots h_{i,w}] \quad (7)$$

where $h_{i,j}$ is the hidden vector representation of the j-th word of sentence/instance S_{ei}, which concatenates the forward and backward LSTM along with a non-linear transformation σ. One with the standard sequence, and the other one with the sequence reversed:

$$h_{i,j} = \sigma([\overrightarrow{h_{i,j}}, \overleftarrow{h_{i,j}}]) \quad (8)$$

C. ATTENTION LAYER

Not all the sentences play an important role in the training process of the attribute extractors in the task of extracting the attribute values from multiple instances of the text. For example in Figure. 3, there are five sentences in the cancer specimen. However, only the first sentence "cancer involved in upper margin" is useful for the attribute extractor of CUPM.

In summary, we introduce an attention mechanism to identify the most important instances for different attribute extractors, which can dynamically reduce the weight of those noisy instances/sentences. Specifically, we use the attention

mechanism proposed by Yang *et al.* [35] to represent all the instances/sentences h, which is defined as follows:

$$h = \sum_i \alpha_i h_i \tag{9}$$

where h_i is instance of i, and α_i is the weight of instance h_i and defined as follows:

$$\alpha_i = \frac{\exp(\mathbf{m_i}^{\mathrm{T}} m_s)}{\sum_i \exp(\mathbf{m_i}^{\mathrm{T}} m_s)} \tag{10}$$

$$m_i = \tanh(w_s h_i + b_s) \tag{11}$$

where m_s and m_i are the instance/sentence-level context vectors, which can be randomly initialized and jointly learned during the training process [35].

D. OUTPUT LAYER

Aiming at the problem of the insufficient amount of data in intestinal cancer diagnostic text, we use multi-task learning to complete the extraction of multiple attribute values. The representation of all instances h is ultimately fed into different output layers. And, we use the soft-max function to obtain multi-class results.

$$P_c = softmax(w_c h + b_c) \tag{12}$$

where P_c is prediction probabilities for task c, w_c is the weight parameters and b_c is the bias vector.

E. MODEL TRAINING

We use Keras, which is a high-level neural networks library written in Python, to implement the model. The purpose of the training process is to minimize the cross-entropy of the predicted results and real results for all tasks. For a single task c, we use the categorical cross-entropy as a loss function. At the same time, stochastic gradient descent(SGD) is an optimizer.

And as learning progresses, the degree of reduction in different loss functions is not consistent. Therefore, we assign different weights to the contribution of the loss function of each task to the final loss.

$$Loss = \sum_{c=1}^{C} \gamma_c L(t_c, P_c) \tag{13}$$

$$L(t, P) = -\sum_{i=1}^{N} \sum_{j=1}^{M} t_{i,j} \log P_{i,j} \tag{14}$$

where C is the sum of tasks, γ_c is the weights for each task c, $c \in C$, $P_{i,j}$ is the probability result, $t_{i,j}$ is the true result, N is the number of training samples and M is the class number.

V. SEQUENCE LABELING

In this section, we mainly present the implementation of our sequence labeling method to extract extractive attribute values from cancer specimens, including `SN, AS, TS, ID, and CMR`. It is shown in Figure. 4. As we knew, CNN is good at learning and capturing spatial features. RNN is good at capturing timing features. To combine both the advantages, we concatenate the embedding from both models to get a better sentence representation. Finally, a CRF layer is used to learn the constraints of sentences.

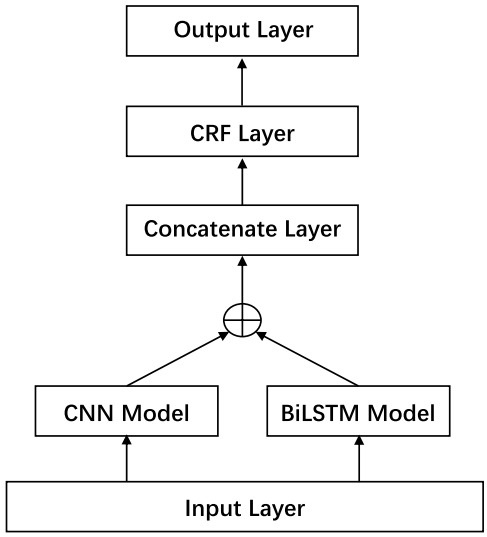


FIGURE 4. Framework of using BiLSTM-CNN-CRF to extract the attribute values.

A. EMBEDDING LAYER

The sentences in each cancer specimen $e \in E$ are divided into words, which is treated as separate individuals. In this paper, we first use the pre-trained character embedding to initialize the parameters of the embedding layer and then use our data to fine-tune the character embedding so that it can learn medical knowledge better and achieve a better effect.

B. CNN LAYER

CNN has the capability of representation and can perform convolution calculations [8]. It is known from existing literature [3], [9] that using CNN to extract character features can achieve very good results in the NER field.

After the embedding layer, the text is transformed into a character embedding matrix, then CNN is used to learn the character embedding information to form a character representation. The CNN model used in this paper is similar to Chiu and Nichols [3]. It consists of a character embedding matrix, a convolutional layer, and a fully connected layer, as shown in Figure. 5.

C. BiLSTM LAYER

Then we use a BiLSTM layer to capture the past and future features jointly to obtain a better text representation [37], as shown in Figure. 6. For each text e, we use a hidden vector h to represent it:

$$h = [h_1, h_2, \ldots, h_i, \ldots h_n] \tag{15}$$

where h_i is the hidden vector representation of the i-th word of text e, which concatenates the forward and backward LSTM along with a non-linear transformation σ. One with

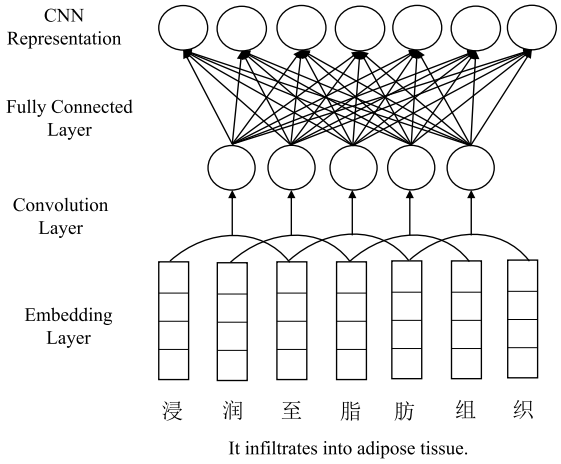


FIGURE 5. CNN is used to extract representations of "infiltration depth".

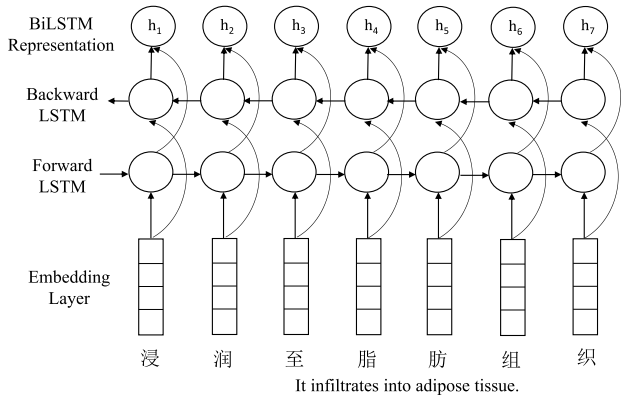


FIGURE 6. BiLSTM is used to extract representations of "infiltration depth".

the standard sequence and the other one with the sequence reversed:

$$h_i = \sigma([\overrightarrow{h_i}, \overleftarrow{h_i}]) \tag{16}$$

D. CRF LAYER

In the last layer of the model, this paper chooses to use Conditional Random Field (CRF). As BiLSTM can only extract the character characteristics of the context in the sentence, it cannot find the optimal sequence in the labeled sequence. However, CRF can perform joint probability analysis on the labeled sequence by considering the relationship between adjacent labels. It is shown in Figure. 7. We use the "BIOE" tagging strategy which is the most popular one. In the BIOE tagging strategy, "B" represents the beginning of an attribute, "I" represents the inside of an attribute, "O" represents the outside of an attribute, and "E" represents the end of an attribute.

Give a text $X = \{x_1, x_2, \ldots, x_n\}$ represents the input sequence, P is a state score matrix output of the BiLSTM neural network, the size is $n \times k$, where k is the number of different labels and $P_{i,j}$ is the score of the j-th label of the i-th word in the text. The score of a predicted label sequence

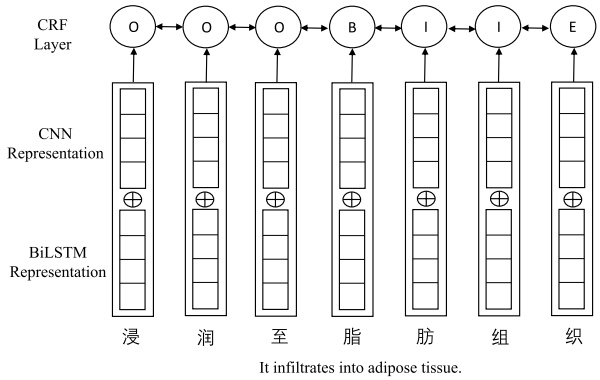


FIGURE 7. CRF is used to get values of "infiltration depth".

$y = \{y_1, y_2, \ldots, y_n\}$ is defined as:

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

where A is a fraction transfer matrix, $A_{i,j}$ represents the score of the label i transferred to the label j, and y_0 and y_n are the flags added respectively at the beginning and end of the text, which means A is a $k+2$ size Square matrix. The model uses the stochastic gradient descent(SGD) as the optimizer.

VI. EXPERIMENTS

In this section, we describe some details of the experimental setup. All models are trained on a server with Intel(R) Xeon(R) E5-2620 2.40GHz CPU and NVIDIA GeForce GTX 1080Ti, running Windows Server 2016 and 64GB memory.

A. DATA

In our experiment, our electronic medical record data set comes from a hospital. Our task is to extract discriminative attributes and extractive attributes from the unstructured text of EMR.

Discriminative attributes include:

- cancer involved in upper margin (CUPM)
- cancer involved in bottom margin (CBOM)
- cancer involved in base margin (CBAM)
- Nerve Invasion (NI)
- Vascular Invasion (VI)

And the status of discriminative attributes have three labels, including: {`YES`,`NO` and `UNKNOWN`}.

And extractive attributes include:

- Specimen Name (SN)
- Adenocarcinoma Shape (AS)
- Tumor Size (TS)
- Infiltration Depth (TD)
- Cancer Metastatic Ratio (CMR)

We use a crowdsourcing approach to label the data. A total of 10 experienced graduate students participated. For each intestinal cancer specimen report, it will be randomly assigned to three people, and each person needs to annotate all corresponding attribute values in it. If the results of the

three people are consistent, then we add this data to our annotation set. Finally, we get 8,818 labeled records. In the model experiment, 80% of them (7,054 labeled records) are randomly selected as training data and the remaining 20% (1,764 labeled records) as test data.

B. PARAMETER SETTING

Parameters of the neural layers are randomly initialized. We apply 10-fold cross-validation and different combinations of parameters are investigated, of which the best one is described in Table 1 and 2.

TABLE 1. Parameter setting for text classification.

Embedding size	200
Size of word-level BiLSTM layer	50
Size of sentence-level BiLSTM layer	50
Initial learning rate	0.01
Batch size	32

TABLE 2. Parameter setting for sequence labeling.

Embedding size	100
Size of character-level BiLSTM layer	50
Initial learning rate	0.01
Dropout rate	0.2
Batch size	32

C. BASELINES

We compare our method with variant methods, including different embedding methods and different learning models.

1) EMBEDDING METHODS OF TEXT CLASSIFICATION

- Tencent Embedding [27]. Tencent embedding is provided by Tencent AI Lab, which is pre-trained on large-scale high-quality data and contained over 8 million Chinese words and phrases, including Tencent News, daily express, internet pages, novel corpora, and also Wikipedia and Baidu Encyclopedia. It provides 200-dimension embedding.
- Word2vec Embedding [21]. In our experiment, we use word2vec to train word embedding from the training data.
- Fine-tuning Embedding1. In our experiment, we initialize our neural network models with the pre-trained Tencent embedding, then we fine-tune them by using the training data.

2) EMBEDDING METHODS OF SEQUENCE LABELING

- Wiki Embedding. It is pre-trained on large-scale and high-quality data based on 340,000 Wikipedias, which can well represent characters in general domain text. It provides 100-dimension embedding.
- Fine-tuning Embedding2. In our experiment, we initialize our neural network models with the pre-trained Wiki embedding, then we fine-tune them by using the training data.

3) LEARNING MODELS OF TEXT CLASSIFICATION

- TEXT-CNN [12] and MT-TEXT-CNN. TEXT-CNN takes the whole text as input and consists of the embedding layer and the CNN layer. MT-TEXT-CNN is the multi-task version.
- TEXT-LSTM [20] and MT-TEXT-LSTM. TEXT-CNN takes the whole text as input and consists of the embedding layer and the LSTM layer. MT-TEXT-LSTM is the multi-task version.
- TEXT-BiLSTM and MT-TEXT-BiLSTM. TEXT-BiLSTM takes the whole text as input and consists of the embedding layer and the BiLSTM layer. MT-TEXT-BiLSTM is the multi-task version.
- MI-BiLSTM and MT-MI-BiLSTM. MI-BiLSTM treats each sentence of the whole text as an instance, takes all the instances as input, consists of the embedding layer and the BiLSTM layer. MT-MI-BiLSTM is the multi-task version.
- MI-BiLSTM-ATT and MT-MI-BiLSTM-ATT. MI-BiLSTM-ATT treats each sentence of the whole text as an instance, takes all the instances as input, and consists of the embedding layer, the BiLSTM layer, and the attention layer. MT-MI-BiLSTM-ATT is the multi-task version.

4) LEARNING MODELS OF SEQUENCE LABELING

- BiLSTM-CNN-CRF, which takes the whole text as input and consists of the character embedding layer, the BiLSTM layer, the CNN layer, and the CRF layer.
- BiLSTM-CRF, which takes the whole text as input and consists of the character embedding layer, the BiLSTM layer, and the CRF layer.

D. PERFORMANCE

In this part, we introduce the empirical results of our model *MT-MI-BiLSTM-ATT* and *BiLSTM-CNN-CRF*.

1) METRIC

For discriminative attributes, we use multi-task learning to extract attribute values simultaneously and use the *accuracy* and *F1* metric to evaluate the effect of different models. For extractive attributes, we use sequence labeling to extract five attribute values separately and also use the *accuracy* and *F1* metric to evaluate the effect of different models.

2) RESULT OF TEXT CLASSIFICATION

We first compare our proposed model with all the basic models. The comparison results are shown in Table 3. No matter what kind of metric is used, our model(`MT-MI-BiLSTM-ATT`) is superior to several competitive baseline methods. Because our model uses multi-task learning, attention mechanism, and fine-tuning. The details are as follows:

Multi-task Learning. Even though different learning models use the same embedding method, they would achieve

TABLE 3. Performance of different methods for text classification.

Embedding	Learning Models	CUPM		CBOM		CBAM		NI		VI	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Tencent Embedding	TEXT-CNN	0.48	0.43	0.47	0.45	0.59	0.56	0.69	0.60	0.65	0.59
	TEXT-LSTM	0.50	0.49	0.49	0.50	0.63	0.60	0.70	0.62	0.66	0.62
	TEXT-BiLSTM	0.53	0.50	0.51	0.50	0.65	0.64	0.74	0.65	0.69	0.65
	MI-BiLSTM	0.55	0.54	0.56	0.55	0.69	0.68	0.77	0.67	0.73	0.69
	MI-BiLSTM-ATT	0.64	0.59	0.60	0.58	0.72	0.72	0.79	0.73	0.76	0.71
	MT-TEXT-CNN	0.50	0.45	0.58	0.53	0.62	0.62	0.71	0.63	0.73	0.72
	MT-TEXT-LSTM	0.57	0.51	0.60	0.54	0.64	0.65	0.74	0.66	0.75	0.73
	MT-TEXT-BiLSTM	0.58	0.56	0.61	0.57	0.67	0.66	0.75	0.69	0.77	0.76
	MT-MI-BiLSTM	0.60	0.59	0.61	0.59	0.70	0.69	0.80	0.73	0.79	0.76
	MT-MI-BiLSTM-ATT(our model)	0.65	0.63	0.67	0.63	0.74	0.73	0.83	0.76	0.80	0.78
Word2vec Embedding	TEXT-CNN	0.50	0.47	0.55	0.49	0.61	0.57	0.69	0.62	0.69	0.65
	TEXT-LSTM	0.56	0.51	0.58	0.51	0.65	0.60	0.73	0.69	0.72	0.69
	TEXT-BiLSTM	0.57	0.54	0.61	0.52	0.67	0.62	0.76	0.71	0.73	0.72
	MI-BiLSTM	0.59	0.56	0.60	0.54	0.72	0.64	0.80	0.74	0.77	0.73
	MI-BiLSTM-ATT	0.67	0.63	0.62	0.59	0.77	0.73	0.81	0.75	0.79	0.76
	MT-TEXT-CNN	0.56	0.51	0.61	0.57	0.79	0.69	0.80	0.74	0.79	0.71
	MT-TEXT-LSTM	0.61	0.56	0.62	0.57	0.80	0.69	0.82	0.79	0.80	0.74
	MT-TEXT-BiLSTM	0.84	0.77	0.82	0.74	0.80	0.71	0.83	0.78	0.80	0.76
	MT-MI-BiLSTM	0.87	0.80	0.84	0.79	0.81	0.73	0.82	0.78	0.81	0.79
	MT-MI-BiLSTM-ATT(our model)	0.92	0.85	0.88	0.83	0.83	0.76	0.85	0.79	0.84	0.83
Fine-tuning Embedding1	TEXT-CNN	0.59	0.54	0.59	0.55	0.63	0.60	0.70	0.69	0.73	0.69
	TEXT-LSTM	0.63	0.59	0.63	0.61	0.68	0.66	0.73	0.71	0.75	0.72
	TEXT-BiLSTM	0.65	0.61	0.68	0.65	0.72	0.70	0.79	0.77	0.76	0.74
	MI-BiLSTM	0.68	0.64	0.74	0.70	0.78	0.76	0.82	0.79	0.80	0.77
	MI-BiLSTM-ATT	0.73	0.70	0.79	0.76	0.81	0.79	0.84	0.80	0.82	0.79
	MT-TEXT-CNN	0.62	0.63	0.65	0.61	0.81	0.78	0.83	0.77	0.80	0.71
	MT-TEXT-LSTM	0.68	0.70	0.67	0.67	0.81	0.79	0.85	0.79	0.80	0.74
	MT-TEXT-BiLSTM	0.86	0.83	0.83	0.81	0.82	0.80	0.85	0.82	0.81	0.75
	MT-MI-BiLSTM	0.88	0.84	0.85	0.80	0.83	0.82	0.85	0.83	0.82	0.77
	MT-MI-BiLSTM-ATT(our model)	**0.93**	**0.87**	**0.89**	**0.85**	**0.83**	**0.84**	**0.86**	**0.80**	**0.86**	**0.81**

TABLE 4. Performance of different methods for sequence labeling.

Embedding	Learning Models	SN		AS		TS		ID		CMR	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Wiki Embedding	BiLSTM-CRF	0.65	0.62	0.68	0.63	0.73	0.70	0.67	0.63	0.75	0.70
	BiLSTM-CNN-CRF(our model)	0.72	0.69	0.73	0.69	0.80	0.76	0.72	0.69	0.81	0.75
Fine-tuning Embedding2	BiLSTM-CRF	0.68	0.65	0.71	0.68	0.76	0.72	0.70	0.68	0.78	0.72
	BiLSTM-CNN-CRF(our model)	**0.74**	**0.70**	**0.76**	**0.72**	**0.82**	**0.79**	**0.74**	**0.71**	**0.84**	**0.79**

different performances. For example, `MT-MI-BiLSTM` perform better than `MI-BiLSTM` and `TEXT-BiLSTM`, which demonstrates the effectiveness of our multi-task learning strategy.

Attention Mechanism. In the embedding method, models that use attention mechanisms work better. For example, `MT-MI-BiLSTM-ATT` performs better than `MT-MI-BiLSTM`, `MI-BiLSTM-ATT` performs better than `MI-BiLSTM` which demonstrates that the attention measure is very useful to select the most important instances.

Fine-tuning. In the learning method, different embedding methods would achieve different performances. The details are as follows:

- Firstly, the `Tencent Embedding` method achieves the worst performance, which demonstrates that the pre-trained word embedding in the general corpus has not captured capture medical domain semantics/knowledge.
- Secondly, the `Fine-tuning Embedding1` method achieves the best performance, demonstrating the effectiveness of our embedding strategy.

But this model is only applicable to the case where one attribute is described by one sentence. If the phenomenon that one attribute is described by two sentences occurs, the model does not support it. At the same time, non-standard text descriptions are also one of the reasons for poor results.

3) RESULT OF SEQUENCE LABELING

We first compare our proposed model(`BiLSTM-CNN-CRF`) with all the basic models. The comparison results are shown in Table 4. In Table 4, we can see that `BiLSTM-CNN-CRF` is better than `BiLSTM-CRF`, which proves that the spatial text features learned by CNN are helpful to improve the overall model effect.

And the `Fine-tuning Embedding2` method achieves the best performance, which proves the effectiveness of our strategy. Fine-tuning can make the character embedding capture the semantics/knowledge of the medical domain.

E. CASE STUDY

In order to intuitively understand the benefits of our model, we use two examples to illustrate. It is shown in Figure. 8.

ID	Sentences
1	直肠癌根治标本，腺癌，中分化，溃疡型，肿瘤大小6.0×4.0×1.5cm，浸润至肠壁外纤维脂肪组织，未见肯定的脉管瘤栓及神经侵犯；另见绒毛状管状腺瘤伴低级别上皮内瘤变；标本上切端未见癌累及，另送“下切缘”及基底切端均见癌累及；肠旁淋巴结0/13枚未见癌转移。
2	“右半结肠根治标本”，中分化腺癌，溃疡型，肿瘤直径0.65cm，侵透胃壁层至纤维脂肪组织，未见明确的神经侵犯，脉管内癌栓，上下切端均未见癌组织累及，基底切缘见癌累及，肠旁淋巴结3/20枚见癌转移。

discriminative attribute	ID-1					ID-2				
	CUPM	CBOM	CBAM	NI	VI	CUPM	CBOM	CBAM	NI	VI
MI-BiLSTM	NO (否)	NO (否)	YES (是)	YES (是)	NO (否)	NO (否)	NO (否)	NO (否)	YES (是)	NO (否)
MI-BiLSTM-ATT	NO (否)	YES (是)	YES (是)	YES (是)	NO (否)	NO (否)	NO (否)	YES (是)	YES (是)	NO (否)
MT-MI-BiLSTM	NO (否)	NO (否)	YES (是)	YES (是)	NO (否)	NO (否)	NO (否)	NO (否)	YES (是)	NO (否)
MT-MI-BiLSTM-ATT	NO (否)	YES (是)	YES (是)	NO (否)	NO (否)	NO (否)	NO (否)	YES (是)	YES (是)	NO (否)

extractive attribute	ID-1		ID-2	
	BiLSTM-CRF	BiLSTM-CNN-CRF	BiLSTM-CRF	BiLSTM-CNN-CRF
SN	Rectal Carcinoma Radical Specimen, (直肠癌根治标本，)	Rectal Carcinoma Radical Specimen (直肠癌根治标本)	“Radical specimen of right hemicolectomy”, (“右半结肠根治标本”，)	Radical specimen of right hemicolectomy (右半结肠根治标本)
AS	Ulcerative Type (溃疡型)	Ulcerative Type (溃疡型)	Ulcerative Type (溃疡型)	Ulcerative Type (溃疡型)
TS	6.0×4.0×1.5cm	6.0×4.0×1.5cm	UNKNOWN(无)	0.65cm
ID	Extra-intestinal fibrous and adipose tissue (肠壁外纤维脂肪组织)	Extra-intestinal fibrous and adipose tissue (肠壁外纤维脂肪组织)	layer of stomach wall to fibrous and adipose tissue (胃壁层至纤维脂肪组织)	fibrous and adipose tissue (纤维脂肪组织)
CMR	0/13	0/13	3/20	3/20

FIGURE 8. Our model and baseline model are used to extract discriminative attributes and extractive attributes values.

Our model (`MT-MI-BiLSTM-ATT` and `BiLSTM-CNN-CRF`) achieve the best performance than all baseline.

For discriminative attributes, in the first example of Figure. 8, for CBOM and CBAM attributes, an instance has two mentions of cancer, the baseline model cannot accurately extract attribute values. But our model accurately extracts attribute values, because considering that the importance of different instances for different attribute extractors is not equal, we use an attention layer to select the most important instances to compose text representation. And we use multi-task learning to share useful information to get better experimental results, so our model get the best result.

For extractive attributes, in the second example of Figure. 8, for ID attribute, the baseline model extracts redundant attribute values. But our model accurately extracts attribute values, because considering that CNN is good at learning and capturing spatial features, RNN is good at capturing timing features, so we concatenate the embedding from both models to get a better sentence representation. So `BiLSTM-CNN-CRF` is better than `BiLSTM-CRF`.

F. APPLICATION

For a description of the patients' condition, doctors are more inclined to use natural language to write electronic medical records. Since the diversity of natural language, the contents of electronic medical records written by different doctors are inconsistent. At the same time, unstructured electronic medical records are also not conducive to the analysis and mining of patient conditions. Therefore, tools are needed to normalize the records written by different doctors. So our method is very useful for the standardization of electronic medical records. We have completed the extraction of two types of attribute values. The one is the discriminative attribute. The text classification method is used to extract the attribute values of this part. The attributes include *CUPM, CBOM, CBAM, NI, and VI*. The other one is the extractive attribute. The sequence labeling method is used to extract the attribute values of this part. The attributes have five types, including *SN, AS, TS, ID, and CMR*. The application is shown in Figure. 9. If there is an error in the extraction of the attribute value, we can manually verify it and click the

FIGURE 9. **The application is used for attribute extraction tasks related to electronic medical records.**

"Submit Modification" button, which will be synchronized to the "Standard Format Text" for correction to ensure the accuracy of structured data.

VII. CONCLUSION

In this paper, we focus on extracting discriminative attributes and extractive attributes from the unstructured text of EMR. And we treat the attribute values extraction tasks of *text classification* and *sequence labeling*. For discriminative attributes, we first use the pre-trained word embedding to initialize the parameters, then we fine-tune them by using our domain corpus. Finally, we use an attention layer and multi-tasking method to get better experimental results. For extractive attributes, we propose a novel model that includes the BiLSTM layer, the CNN layer, and the CRF layer. Experiments have shown that our method is superior to several competitive baseline methods. In the future, we will use transfer learning methods to apply our model to more hospitals and more diseases.

REFERENCES

[1] M. Alawad, S. Gao, J. X. Qiu, H. Yoon, J. B. Christian, L. Penberthy, B. Mumphrey, X. Wu, L. Coyle, and G. D. Tourassi, "Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 1, pp. 89–98, 2020.

[2] C. Chantrapornchai and A. Tunsakul, "Information extraction based on named entity for tourism corpus," 2020, *arXiv:2001.01588*. [Online]. Available: https://arxiv.org/abs/2001.01588

[3] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 357–370, 2015.

[4] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, Feb. 2011.

[5] M. Cui, R. Bai, Z. Lu, X. Li, U. Aickelin, and P. Ge, "Regular expression based medical text classification using constructive heuristic approach," *IEEE Access*, vol. 7, pp. 147892–147904, 2019.

[6] P. Ding, X. Zhou, X. Zhang, J. Wang, and Z. Lei, "An attentive neural sequence labeling model for adverse drug reactions mentions extraction," *IEEE Access*, vol. 6, pp. 73305–73315, 2018.

[7] C. Dyer, M. Ballesteros, W. Ling, A. Matthews, and N. A. Smith, "Transition-based dependency parsing with stack long short-term memory," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, vol. 1. Beijing, China, 2015, pp. 334–343.

[8] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[9] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*. [Online]. Available: https://arxiv.org/abs/1508.01991

[10] M. Hussain, D.-J. Choi, and S. Lee, "Semantic based clinical notes mining for factual information extraction," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Barcelona, Spain, Jan. 2020, pp. 46–48.

[11] A. Jagannatha and H. Yu, "Structured prediction models for RNN based sequence labeling in clinical text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Austin, TX, USA, 2016, pp. 856–865.

[12] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1746–1751.

[13] Y. Kim and S. M. Meystre, "Ensemble method–based extraction of medication and related information from clinical texts," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 1, pp. 31–38, 2020.

[14] J. Krebs, M. Krug, G. Fette, G. Dietrich, M. Ertl, G. Güder, F. Puppe, and M. Kaspar, "Identifying heart failure patients by medical text classification," in *Proc. ICT Health Sci. Res. (EFMI)*, Hanover, Germany, 2019, pp. 251–252.

[15] L. Li and Y. Jiang, "Biomedical named entity recognition based on the two channels and sentence-level reading control conditioned LSTM-CRF," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 380–385.

[16] L. Li, L. Jin, Y. Jiang, and D. Huang, "Recognizing biomedical named entities based on the sentence vector/twin word embeddings conditioned bidirectional LSTM," in *Proc. CCL*, 2016, pp. 165–176.

[17] L. Qing, W. Linhong, and D. Xuehai, "A novel neural network-based method for medical text classification," *Future Internet*, vol. 11, no. 12, p. 255, Dec. 2019.

[18] X. Ling and D. S. Weld, "Fine-grained entity recognition," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 94–100.

[19] K. Liu and L. Chen, "Medical social media text classification integrating consumer health terminology," *IEEE Access*, vol. 7, pp. 78185–78193, 2019.

[20] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI)*, New York, NY, USA, Jul. 2016, pp. 2873–2879.

[21] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781

[22] Z. H. Moe, T. San, M. M. Khin, and H. M. Tin, "Comparison of naive bayes and support vector machine classifiers on document classification," in *Proc. IEEE 7th Global Conf. Consum. Electron. (GCCE)*, Nara, Japan, Oct. 2018, pp. 466–467.
[23] B. Peng, X. Zhang, Y. He, and Z. Li, "Attribute extraction by combing feature ranking and sequence labeling," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Shanghai, China, Jan. 2018, pp. 553–556.
[24] G. Popovski, B. Koroušić-Seljak, and T. Eftimov, "A survey of named-entity recognition methods for food information extraction," *IEEE Access*, vol. 8, pp. 31586–31594, 2020.
[25] V. Ranganayaki and S. N. Deepa, "Linear and non-linear proximal support vector machine classifiers for wind speed prediction," *Cluster Comput.*, vol. 22, no. S1, pp. 379–390, Jan. 2019.
[26] P. Kameswara Sarma, Y. Liang, and B. Sethares, "Domain adapted word embeddings for improved sentiment classification," in *Proc. Workshop Deep Learn. Approaches Low-Resource NLP*, Melbourne, VIC, Australia, 2018, pp. 37–42.
[27] Y. Song, S. Shi, J. Li, and H. Zhang, "Directional skip-gram: Explicitly distinguishing left and right context for word embeddings," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, vol. 2, 2018, pp. 175–180.
[28] M. Torii, C. N. Arighi, G. Li, Q. Wang, C. H. Wu, and K. Vijay-Shanker, "RLIMS-P 2.0: A generalizable rule-based information extraction system for literature mining of protein phosphorylation information," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 1, pp. 17–29, Jan./Feb. 2015.
[29] S. Wang, S. Sun, and J. Xu, "AUC-Maximized deep convolutional neural fields for protein sequence labeling," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Riva del Garda, Italy, Sep. 2016, pp. 1–16.
[30] R. Weegar, A. Pérez, A. Casillas, and M. Oronoz, "Recent advances in swedish and spanish medical entity recognition in clinical texts using deep neural approaches," *BMC Med. Informat. Decis. Making*, vol. 19, no. S7, Dec. 2019.
[31] H. Wei, M. Gao, A. Zhou, F. Chen, W. Qu, C. Wang, and M. Lu, "Named entity recognition from biomedical texts using a fusion attention-based BiLSTM-CRF," *IEEE Access*, vol. 7, pp. 73627–73636, 2019.
[32] J. Xu, Y. Wu, Y. Zhang, and H. Xu, "Detecting body location modifiers of disorders in clinical texts via sequence labeling," in *Proc. Amer. Med. Inform. Assoc. Annu. Symp.*, Washington, DC, USA, Nov. 2017, pp. 584–601.
[33] B. Yang, G. Dai, Y. Yang, D. Tang, Q. Li, D. Lin, J. Zheng, and Y. Cai, "Automatic text classification for label imputation of medical diagnosis notes based on random forest," in *Proc. Int. Conf. Health Inf. Sci.*, Cairns, QLD, Australia, Oct. 2018, pp. 87–97.
[34] X. Yang, Z. Gao, Y. Li, C. Pan, R. Yang, L. Gong, and G. Yang, "Bidirectional LSTM-CRF for biomedical named entity recognition," in *Proc. 14th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD)*, Jul. 2018, pp. 239–242.
[35] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, San Diego CA, USA, 2016, pp. 1480–1489.
[36] H. Yu, H. Li, D. Mao, and Q. Cai, "A relationship extraction method for domain knowledge graph construction," *World Wide Web*, vol. 23, no. 2, pp. 735–753, Mar. 2020.
[37] G. Zheng, S. Mukherjee, X. L. Dong, and F. Li, "Opentag: Open attribute value extraction from product profiles," in *24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 1049–1058.

SUFEN WANG received the B.S. degree in industrial automation from Jiangnan University, the M.S. degree in power transmission and automation from Shanghai Jiao Tong University, and the Ph.D. degree in control science from Donghua University, China. She is currently a Professor with the Department of Information System and Information Management, Donghua University. Her current research interests include validity analysis of information technology, natural language processing, and internet hospital.

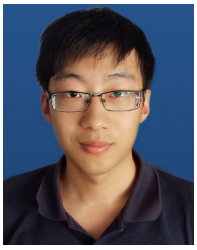
MINMIN PANG was born in Linyi, Shandong, China, in 1994. She received the B.S. degree in computer and information from Zaozhuang University, in 2017. She is currently pursuing the master's degree with the School of Computer Science and Technology, Donghua University, Shanghai. Her research interests include data analysis and mining and natural language processing.

CHANGQING PAN received the B.S. degree in medical from Shanghai Second Medical University, and the M.S. degree in public health from Fudan University, China. Since 2002, he has been a Chief Surgeon with the General Surgery Department, Shanghai General Hospital. Since 2016, he has been the President of the Shanghai Chest Hospital. His research interests include gastroenterology and hospital management.

JUNYI YUAN received the M.S. degree in software engineering from Donghua University. Since 2013, he has been the Director of the Information Center. He is currently a Senior Engineer with the Information Center, Shanghai Chest Hospital, China. His research interests include hospital information and medical data analysis.

BO XU received the Ph.D. degree from Fudan University, in 2018. He is currently a Lecturer with the School of Computer Science and Technology, Donghua University. He has published several papers at major conferences, including IJCAI, ICDE, CIKM, PKDD, DASFAA, and so on. His research interest includes knowledge base construction and applications. He received the Best Student Paper Award from NDBC, in 2014.

MING DU was born in Hulin, Heilongjiang, China, in 1975. He received the B.S. degree in electronic technology from Chinese Textile University, in 1998, and the M.S. degree in control theory and control engineering and the Ph.D. degree in management science engineering from Donghua University, Shanghai, China, in 2005 and 2013, respectively. He has been an Associate Professor with Donghua University, since 2010. His research interests include information retrieval techniques, data analysis and mining, and natural language processing.



HONG ZHANG is currently pursuing the master's degree with Donghua University, Shanghai, China. He is also a Visiting Graduate Student with the High Performance Computer Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include pathological image analysis and natural language processing.