IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 30, 2016, accepted July 19, 2016, date of publication July 27, 2016, date of current version August 26, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2594839

A Convolutional Neural Network Model for Online Medical Guidance

CUILI YAO[1,2], YUE QU[3], BO JIN[2], (Senior Member, IEEE), LI GUO[4], CHAO LI[2], WENJUAN CUI[5], AND LIN FENG[2]

[1]School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China
[2]School of Innovation and Entrepreneurship, Dalian University of Technology, Dalian 116024, China
[3]Department of Biomedical Engineering, Dalian University of Technology, Dalian 116024, China
[4]School of Computer Science, Dalian Polytechnic University, Dalian 116034, China
[5]Computer Network Information Center, Chinese Academy of Sciences, Beijing 100049, China

Corresponding author: B. Jin (jinbo@dlut.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 91546123, in part by the National Key R&D Plan through the Ministry of Science and Technology of China under Grant 2016YFC0901900, and in part by the Fundamental Research Funds for the Central Universities of China under Grant DUT15QY44.

ABSTRACT The aging population of China is becoming increasingly more prominent, thus increasing the burden on medical resources. Therefore, the use of data mining technology to improve the efficiency of disease diagnosis has the following important significance. For hospitals, such technology can reduce the cost of providing one-on-one guidance to patients and the probability of registration errors. For patients, it can save time and energy spent on hospital visits; in addition, through remote access, patients can follow the automated guidance at home to complete registration, thereby enhancing admission efficiency. For internet users, such technology enables self-checking of these users' health conditions on a regular basis; based on certain main symptoms, possible diseases can be pre-diagnosed, thus providing a risk warning. Online medical guidance has become a very important step. To this end, we focus on employing the data mining technology to enhance the performance of online medical guidance. In this paper, we propose a medical diagnosis method called the named entity recognition method and a convolutional neural network model. We apply our proposed method and model as an innovative framework for hospitalization guidance to provide human-like, comprehensive and informative automated medical consultations. We perform experiments on real-world datasets. The experimental results show that our methods achieve state-of-the-art performance compared with baselines.

INDEX TERMS Medical guidance, convolutional neural network, name entity recognition.

I. INTRODUCTION

The internet is the key bridge for connecting patients with medical services. When people do not feel well, nearly 90% of them first go onto the internet to search for related medical information. The internet has already changed the eco-system of medical services for all major steps, including medical consultation, clinic visits, treatment and recovery as well as buying medication online [1]. According to a recent report, internet healthcare will have a total market value of one trillion.

Current medical network systems already have certain self-diagnosis functions, with registration systems available in certain medical network systems. However, the majority of medical network systems are based on expert systems, namely, experts use their experience to pre-diagnose diseases based on patients' symptoms and then symptomatically look for relevant experts and specialists within the network [2]. Such systems require multiple experts, and the summarization of different rules consumes substantial time, effort, and manpower. Furthermore, these expert systems suffer from high individual costs, thus making them unsuitable for wide application and unsatisfactory in terms of personalization [3].

In this paper, we will focus on data mining and machine learning technologies to research medical guidance with the objective of providing human-like, comprehensive and informative automated medical consultation. Most people, when they become sick, as a result of their lack of medical knowledge and experience, will describe their symptoms inaccurately in medical terms. The medical guidance model mentioned in this paper uses deep learning technologies, such as CNN and NLP, to perform feature construction and transformation on raw, noisy data. The current model covers 500 different types of diseases and a has rank-1 accuracy of nearly 70%.

When modeling the proposed method, we solve many problems. For the problem in which the word segmentation tool does not perform well on the medical field words, we use a mutual-information-based new word discovery method; for the problem of extracting symptoms from user queries, we use a named entity recognition algorithm.

The remainder of this paper is organized as follows. We begin with a brief survey. Then, we formulate the problem of online medical guidance and introduce the data employed in this paper. We present the medical named entity recognition method and convolutional neural network model in the next two sections. This is followed by the experimental results and analysis. Finally, we offer a conclusion to this paper.

II. RELATED WORKS

The basic of medical guidance depends on machine diagnosis, which is a historic topic. As early as 1966, Ledley and Lusted [4] proposed the idea of machine diagnosis. In 1972, Willcox *et al.* [5] also attempted to use a computer and applied Bayesian theory to identify a bacterial disease. In 2001, Saeys *et al.* [6] proposed applying a feature selection technique to biological information, which was published in Bioinformatics (Oxford University Press). He proposed that, although the feature selection technique has wide application in the field of bioinformatics, in the biological field, this application of the technique has just been initiated. Because the medical samples exhibit the features of large dimension and short length (as in medical text information), it is necessary to modify and optimize the feature selection technique according to such characteristics of medical data. In 2010, in an article also published in Bioinformatics, Abeel *et al.* [7] studied a feature selection algorithm to identify biochemical features in the diagnosis of cancer and used a support vector machine classification algorithm to apply the integrated feature selection technique to disease diagnosis. In recent years, relevant studies on the text classification technique in the medical field have become progressively more mature. Concerning the relevant diagnosis of heart diseases and neural diseases, Ahmed integrated an artificial bee colony algorithm and a modified full Bayesian network classifier and used this combined technique for the mixed estimation, therein achieving a nearly 100% accuracy for heart disease prediction [8]. Patil [9] applied the Jelinek-Mercer smoothing method and the Bayesian model to predict and diagnose heart diseases.

Traditional human-aided medical guide services are human resource intensive. This requires the collaboration of general practitioners and occupational physicians to improve the quality of their social medical guidance and the satisfaction of their patients [10]. Online medical guide services are almost human-aided and suffer from an uncertain wait time for patients. Existing medical guidance systems remain insufficient. Lin *et al.* [11] designed an intelligent medical guide system based on the TF-IDF algorithm, which consists of three modules: the User Interface, Nature Language Symptoms, and Medical Guide Calculations. Recently, the application of a network for disease diagnosis following automated guidance has also drawn increasing attention. In 2015, the ‘‘voice guidance’’ function was released online on the Baidu Doctor APP, which is a novel method based on the promising technology of voice-intelligent identification for solving a practical issue encountered by Baidu when combining the medical guidance module. This is also the first natural-language-based intelligent guidance technology in China to be applied under a mobile medical scenario.

III. PROBLEM FORMULATION

Based on the valuable data generated by the activities of physicians on the internet, we establish a guidance model for disease diagnosis using data mining to help users obtain artificial-intelligence-guided diagnoses prior to admission. In this paper, we propose a framework for artificial-intelligence-guided disease diagnosis. We first annotate the extracted disease described by users with a natural-language disease inquiry through named entity recognition (NER); then, we use word embedding to convert the disease inquiry data from users for the matrix expression. By training a convolutional neural network (CNN), we derive a model for artificial-intelligence-guided disease diagnosis and eventually achieve intelligent guidance for disease inquiry for newly acquired users through the artificial-intelligence-guided disease diagnosis model. In the artificial-intelligence-guided disease diagnosis model, we cause the disease diagnosis abstraction to become a classification question and use a relatively hot topic in deep learning research, the CNN model, to solve this classification problem; furthermore, we perform in-depth mining of the vast medical knowledge on the internet, therein obtaining an artificial-intelligence-guided disease diagnosis model with relatively good results. The research framework for artificial-intelligence-guided disease diagnosis proposed in this paper is illustrated in Figure 1.

IV. MEDICAL NAMED ENTITY RECOGNITION

In language usage, a named entity is defined as a text string with an independent meaning that is often used as an entirety in a sentence [12]. NER is the recognition of an entity with a specific meaning in the text, such as the named entity in the open field, which mainly includes names of people, places, organizations and institutions. There are also various classifications in different sub-divided fields [13]. For example, the entity recognized in the biomedical field has been expanded to technical terms such as names of genes and proteins [14].

Generally, the named entities in the medical field fall into four categories: disease, symptom, examination, and treatment. The artificial-intelligence-guided disease diagnosis studied in this paper is mainly for recognizing named entities in the category of symptom. NER is a pattern recognition task, namely, identifying boundary information and type information of the entity from a given sentence. A typical

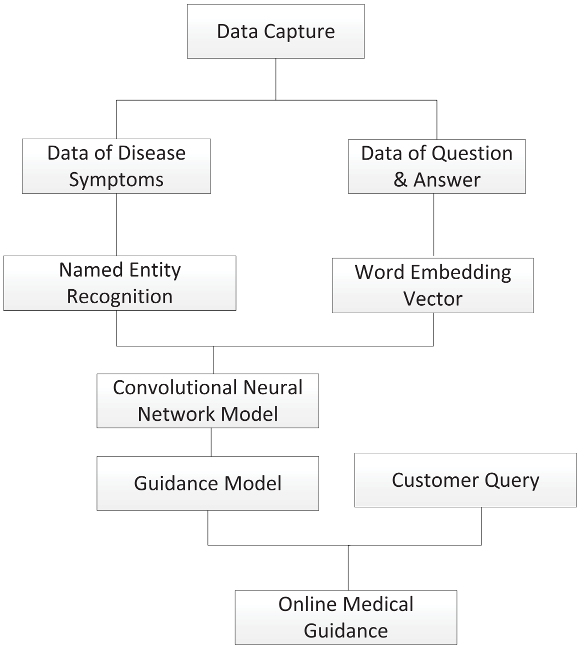


FIGURE 1. Framework of Medical Guidance.

method is to combine the boundary information and type information as a series of labels; then, the task of NER is converted into forecasting a label for each word in the sentence. A typical labeling method generates labels in the form of B_C and I_C, where B and I are position labels, C is a category label, B is the beginning of an entity, and I is the continuation of an entity. Content that does not belong to any entity is generally labeled with an O. For example, for an input sentence "Hi, I feel headache, nausea and vomiting. Do I have trigeminal neuralgia?", the result after labeling should be "Hi /O,/O I /O feel/O headache/B_S, /O nausea/B_S and/I_S vomiting /I_S ./O Did/O I /O have /O trigeminal/O neuralgia /O? /O", where B_S annotates the beginning of the symptom and I_S annotates the continuation of the symptom.

The NER includes two classes of methods. One class is a method based on classification, and the other class is a method based on the serialization of annotations. Because methods based on the serialization of annotations are superior to the methods based on classification in many respects, this paper adopts a method based on the serialization of annotations, i.e., the Conditional Random Field (CRF) model, to conduct the recognition of the symptom class among named entities in the medical field.

For two random variables X and Y, where X is the observation sequence to be labeled and Y is the label sequence, the conditional probability for a given observation sequence X being labeled by Y is $P(Y|X)$. Presume that graph $G = (V, E)$ is a undirected graph, where V is the set of apexes and E is the set of sides. If a random variable Y constitutes a Markov Random Field (MRF) expressed by G, namely,

$$P(Y_v|X, Y_w, w \neq v) = P(Y_v|X, Y_w, w \sim v) \quad (1)$$

which holds for any random node v, then the conditional probability distribution $P(Y|X)$ is CRF. Here, $w \sim v$ means all nodes in Figure G with a side connection to node v, w, and $w \neq v$ means all the nodes except for node v.

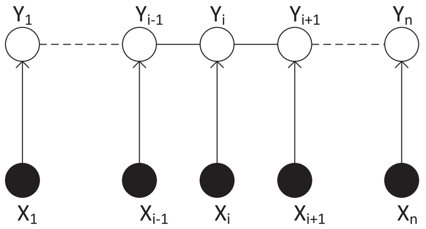


FIGURE 2. Linear Conditional Random Field.

As shown in Figure 2, in the annotation task of the sequence, the input variable X and the output variable Y are typically linear, thus constituting a linear CRF, where the input variables are $X = (X_1, X_2, \ldots X_n)$ and $Y = (Y_1, Y_2, \ldots Y_n)$. Then, under the condition of a given input observation sequence X, the distribution of the conditional probability for the output annotation sequence Y, $P(Y|X)$, satisfies the following property:

$$P(Y_i|X, Y_1, \ldots Y_{i-1}, Y_{i+1}, \ldots Y_n) = P(Y_i|X, Y_{i-1}, Y_{i+1}) \quad (2)$$

where $i = 1, 2, \ldots n$. Let X be x and Y be y; then, the linear CRF can be expressed in the following form:

$$P(y|x) = \frac{1}{Z(x)} exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,t} \mu_t s_t(y_i, x, i)) \quad (3)$$

$$Z(x) = \sum_{y} exp(\sum_{t,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)) \quad (4)$$

where λ_k and μ_l are the corresponding weights, t_k and s_l are the eigenfunctions, and $Z(x)$ is the normalization coefficient. Upon learning, we use a statistical method, the likelihood estimation of the training dataset, to derive the conditional probability model and input the data upon prediction to derive the output sequence when the conditional probability is maximal.

Because the CRF model is a supervised model, the results obtained when applying it to the extraction of named entities heavily relies on the quality of the annotation dataset. Therefore, the first step in the named entity extraction should be the construction of a standard annotation term set. However, because the cost is high and because the artificial annotation process for a named entity with respect to medical question data and answer data is complicated, it is extremely unrealistic to obtain the training set through artificial annotation. To reduce the workload of artificial labeling and to produce high-quality training data, we use a bootstrapping method in this paper to facilitate the annotation of an entity. The basic idea of the bootstrapping method is to use a relatively small annotation sample set as the seed set. We first use the seed set to train a basic model to annotate the named entity and then use this model to annotate the data. When the confidence level of an annotation result exceeds a certain threshold, it is added to the annotation set; we then use the new annotation data to re-train the model, etc., until the annotation results converge. The detailed processes are shown in Figure 3.

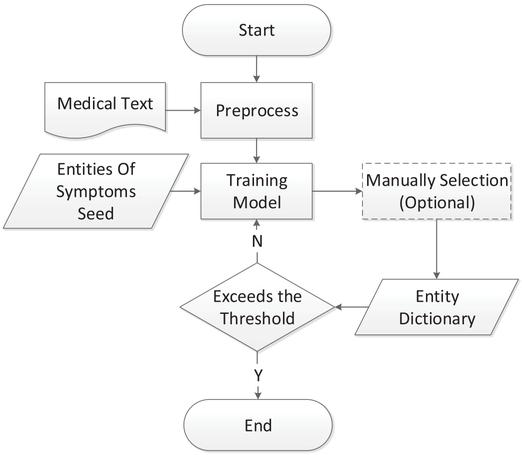


FIGURE 3. **Flow Chart for Bootstrapping Method.**

V. CONVOLUTIONAL NEURAL NETWORK MODEL

The objective of artificial-intelligence-guided disease diagnosis proposed in this paper is to extract the symptom description from the user inquiry and combine the symptoms to infer the most likely disease being suffered by patients under the combination of symptoms. In this paper, we perform the mathematical abstraction of this question, specifically, giving a series of symptoms to derive the disease classification for the patients exhibiting these symptoms. In this paper, we introduce the CNN model to solve this disease classification problem based on symptoms.

A. NETWORK STRUCTURE

In the field of natural language processing, CNNs typically exhibit excellent performance in sentiment analysis, spam detection, and topic classification, and studies on CNNs are continually emerging [15]. Notably, the network structure adopted by Kim [16] is simple and effective, and in this paper, we will use the same network structure for disease classification and prediction.

Figure 4 shows the CNN structure for the text classification process used in our model for artificial-intelligence-guided disease diagnosis.

Let $X_i \in \mathbb{R}^k$ be a k-dimensional vector of the i-th word in the corresponding sentence. A sentence of length n can be expressed as

$$X_{1:n} = X_1 \oplus X_2 \oplus \cdots \oplus X_n \tag{5}$$

where $\oplus$ represents the concatenation operator. Under a typical scenario, we use $X_{i:i+j}$ to represent the concatenation of $X_i, X_{i+1}, \ldots, X_{i+j}$.

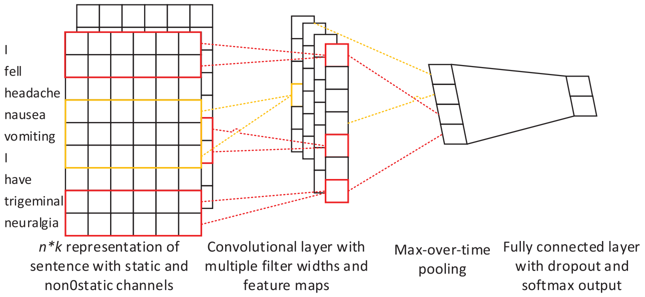

I
fell
headache
nausea
vomiting
I
have
trigeminal
neuralgia
n*k representation of sentence with static and non0static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

FIGURE 4. **Architecture of CNN for Text Classification.**

The convolutional operation contains a filter $w \in \mathbb{R}^{hk}$, and this filter is used to operate on the sliding window of length h to generate a new feature. For instance, a sliding window $X_{i:i+h-1}$ generates the feature c_i according to the following equation:

$$C_i = f(w \cdot x_{i:i+h-1} + b) \tag{6}$$

where $b \in \mathbb{R}$ is an offset term and f is a non-linear function such as the hyperbolic tangent function. This filter is applied to all possible windows in the sentence, $\{X_{1:h}, X_{2:h+1}, \ldots, X_{n-h+1:n}\}$, to generate the characteristic mapping

$$c = [c_1, c_2, \ldots, c_{n-h+1}] \tag{7}$$

Here, we have $c \in \mathbb{R}^{n-h+1}$. Next, each characteristic mapping is applied with a maximum pooling operation, where the specific operation is to take $\hat{c} = max\{c\}$ as the eigenvalue corresponding to the characteristic mapping. The goal of this operation is to extract the most important feature for each characteristic mapping, namely, the feature of maximum value.

As mentioned above, each filter will generate a feature. This model applies the different filters to windows of varying size to generate multiple eigenvalues. These eigenvalues constitute the penultimate layer in Figure 2 and pass a fully connected softmax layer. The output of this layer is the probability distribution on each classification label.

B. REGULARIZATION

To avoid the occurrence of over-fitting, we should adopt an appropriate method to perform over-fitting. Hinton *et al.* [17] applied the dropout method to the CNN for the first time. The dropout method improves the neural network performance by preventing the joint action of feature detectors [18]. Specifically, the dropout method sets a certain proportion of elements in the hidden layer to zero in the forward propagation.

In the neural network mentioned in this paper, if there are m filters, the original formula when we implement the forward operation on the penultimate layer, $z = [\hat{c}_1, \ldots, \hat{c}_m]$, is

$$y = w \cdot z + b \tag{8}$$

However, in the dropout method, in order to avoid over fitting problem, we use the formula

$$y = w \cdot (z \circ r) + b \tag{9}$$

to calculate the hidden element y. Here, "$\circ$" means the sequential multiplication operator of elements in the index group, and $r \in \mathbb{R}^m$ is a $0-1$ vector that obeys the Bernoulli random distribution.

C. FEATURE SELECTION

With respect to the classification task of disease diagnosis, age and gender are important features. The same symptom is often diagnosed as resulting from different diseases based on the consideration of age and gender, and therefore, in this paper, the features of age and gender are added to the neural network.

The network structure used in this paper, if there are m filters, will generate an m-dimensional vector on the penultimate layer; additionally, in this paper, we add two dimensionalities to the generated m-dimensional vector. One dimensionality takes the value of 0 or 1, representing gender (male indicated by 0 and female indicated by 1), and the other dimensionality takes the value of $0 \sim 8$ according to the age span, with each value representing a span of 10 years.

D. ALGORITHM

Through the softmax layer, the output of a network is the probability distribution for each label. In this paper, we select the diseases with the five highest probabilities and take the ratio between their probability and the probability of these five diseases as the predicted disease for output. The algorithm of the module used to achieve the artificial-intelligence-guided disease diagnosis proposed in this paper is shown in Algorithm 1.

Algorithm 1 Artificial-Intelligence-Guided Disease Diagnosis

```
Input: user inquiries, age, gender and word embedded
  vector, which are processed by word segmentation
Initialize the list matrix; the cursor i = 0
while cursor $i$ < length of sentence do
  Set the current word for the $i$-th word of the sentence
  if the cursor $i$ in the named entity results
    then double each dimension of embedded vector
    for current word, add them to the matrix
    else add the embedded vector of current word to
    the matrix
  end if
  i++
end while
set the generated matrix and gender age as the param-
  eters, and use the convolution neural network to
  predict
return forecasting results
```

VI. EXPERIMENTS

In this section, we empirically evaluate the effectiveness of the proposed approach for online medical guidance with intensive experiments on massive real-world datasets. All the data used in this paper are public data from the internet and have been acquired using a web crawler. In particular, the data involving questions/answers to/from doctors and the disease and symptom data are obtained from several major medical information websites in China. A detailed description of the data is provided in Table 1.

TABLE 1. Description of Medical Data.

Dataset	Introduction	Data Amount
1	Disease Information 1	7835
2	Disease Information 2	489
3	Symptom Information	6609
4	Question Answer 1	3263714
5	Question Answer 2	81965

For the disease information dataset and the symptom information dataset, we extract the disease names and symptom names and take the correlation between diseases and symptoms as the knowledge base, which will play a significant role in the diagnosis model. For the question-answer dataset, we extract the content, including the titles of user inquiry, user properties, content of user inquiry, and answers from doctors.

Because the object to be processed is Chinese text, word segmentation is an essential task. However, because the relevant data in the medical field are professional property, some questions from patients obviously include network language, and the word segmentation results for some words are not satisfactory. Therefore, in this paper, we use the method of new word extraction based on mutual information and extract portions of new words that cannot be recognized by the tools for word segmentation. During word segmentation, these unrecognizable words are loaded as a dictionary and therefore can more accurately implement the subsequent word segmentation and data process.

A. NAMED ENTITY RECOGNITION

Before recognizing the named entity of the medical information, we should extract the features of the named entity. The design of the feature extraction template depends on the format of the input file. In particular, there are two types of features: Unigram features and Bigram features. For the Unigram features each row $\%x[\#,\#]$ means to generate a function of a point in the CRF, $f(s, o)$, where s is the label at time t and o is the context at time t. For the Bigram feature, each row $\%x[\#,\#]$ means to generate a function of the side in the CRF, $f(s', s, o)$, where s' is the label at time $(t-1)$. The Bigram feature is often simplified to a B in the template, and then, the default generation $f(s', s)$ (namely, the previous output label and the current output label) is combined as the Bigram feature. Table 2 shows an example of the input file format used in the symptom recognition process, and Table 3

TABLE 2. Sample of training data for CRF.

Word	Part of speech	Label
Hi	ITJ	O
,	PUN	O
I	PNP	O
feel	VVB	O
headache	NN1	B_S
,	PUN	O
nausea	NN1	B_S
and	CJC	I_S
vomiting	VVG	I_S
.	SENT	O
Do	VDB	O
I	PNP	O
have	VHI	O
trigeminal	AJ0	O
neuralgia	SENT	O
?	PUN	O

TABLE 3. Feature Template.

Template Content	Template Interpretation	Sample
U00:%x[-2,0]	Row -2, Column 0	headache
U01:%x[-1,0]	Row -1, Column 0	,
U02:%x[0,0]	Row 0, Column 0	nausea
U03:%x[1,0]	Row 1, Column 0	and
U04:%x[2,0]	Row 2, Column 0	vomiting
U05:%x[-2,1]	Row -2, Column 1	NN1
U06:%x[-1,1]	Row -1, Column 1	PUN
U07:%x[0,1]	Row 0, Column 1	NN1
U08:%x[1,1]	Row 1, Column 1	CJC
U09:%x[2,1]	Row 2, Column 1	VVG
U10:%x[-1,0]/%x[0,0]	Combination of Row-1, Column0 and Row0, Column0	,/nausea
U11:%x[0,0]/%x[1,0]	Combination of Row0, Column0 and Row1, Column0	nausea/and
U12:%x[-2,1]/%x[-1,1]	Combination of Row-2, Column1 and Row-1, Column1	NN1/PUN
U13:%x[-1,1]/%x[0,1]	Combination of Row-1, Column1 and Row0, Column1	PUN/ NN1
U14:%x[0,1]/%x[1,1]	Combination of Row0, Column1 and Row1, Column1	NN1/CJC
U15:%x[1,1]/%x[2,1]	Combination of Row1, Column1 and Row2, Column1	CJC/VVG

presents the featured Unigram template used in this paper, where the third column is an example for the calculation on the word ''nausea''.

In this paper, we use dataset 5 in Table 1 for the first round of training and dataset 4 for the multi-round validation and training. We divide dataset 4 into five equal parts, where each part corresponds to one round of validation and training. We screen the symptom entity newly identified from the first round of results as the training data of the second round, etc. First, we punctuate the inquiry data in dataset 5; in this paper, we use the punctuation ''? ! . ?!'' to syncopate the sentence. The second step is to match the processed sentences with all the symptom names in dataset 3 to find the sentences containing the symptom named entity. We further use the matched entity for labeling.

The experimental results indicate that the first round of seed data contains 2,374 independent symptom named entities from 88,978 sentences. After the first round of training, we obtain 3,990 independent symptom named entities from 1,901,590 sentences, including 2,585 new named entities; after the artificial validation and confirmation, there are 2,396 effective named entities in total. The erroneously recognized entities are divided into the following categories:

1. Disease entity, such as ''multiple Takayasu arteritis'' and ''squamous cell carcinoma of the esophagus and medulla''; 107 errors are in this category, accounting for 56.6% of the problematic data;

2. Handling entity, such as ''radical resection of right lower lung cancer'' and ''intestinal perforation operation''; there are 42 errors in this category, accounting for 22.2% of the problematic data;

3. The entity segmentation is redundant or insufficient, such as ''it is said to be adenoidal hypertrophy'', ''with varus'', and ''father has tympanic membrane perforation''; there are 37 errors in this category, accounting for 19.6% of the problematic data;

4. Other data: an extracted test string that is excessively long can be eliminated according to the text string length.

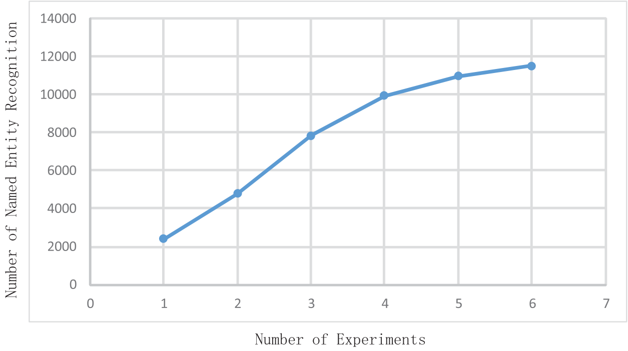


FIGURE 5. Trend for the number of named entities with the number of rounds.

Figure 5 shows the quantitative variation trend presented by the newly identified named entities after symptom entity recognition through five rounds of the bootstrapping method. Here, the horizontal axis is the number of experiments, with 1 as the initial setting, and the vertical axis is the total number of recognized entities.

Recognition of the disease named entity will provide the important symptom feature for the artificial-intelligence-guided disease diagnosis model. In addition, the nearly 12,000 disease named entities generated by the training will also help the word segmentation tool to obtain better word-segmentation results and thereby further improve the accuracy of the entire system.

B. WORD EMBEDDING TRAINING

The meaning of ''word embedding'' is to embed a word into a vector space, namely, using a vector to express the word [19]. In contrast to the Vector Space Model (VSM), word embedding uses many vector expressions for the training words without term supervision to make this vector expression rich in semantic information. Therefore, the word is expressed as a vector with a relatively low dimension; meanwhile, this vector also has certain abstract semantics. A vector's abstract semantics are expressed such that, for two words with very similar meaning, even the similarity of their characters is very small; for example, the vector expression of "diarrhea" and "enterorrhea" is very close, and the similarity is very high in the vector space. In addition, word embedding can also express the relationship between words; one classic example is the relationship between the four roles of queen, king, man, and woman: $V_{queen} - V_{women} + V_{man} \approx V_{king}$.

In this paper, we apply the word embedding training tool from the Skip-Gram model [20], Word2Vector, to train the word embedding on the medical question/answer text. For the obtained word embedding, we can use the cosine similarity to compute the degree of similarity. Table 4 lists the Top 20 similar words to the Chinese word "Fuxie", which means "diarrhea".

TABLE 4. Sample of similarity using word embedding.

Phrase(in Chinese)	Phrase(in English)	Similarity
Laduzi	Diarrhea	0.796444
Laxi	Diarrhoea	0.747393
Fuxie	Diarrhoea	0.709832
Ladu	Diarrhoea	0.708966
Futong	Bellyache	0.654137
Changyan	Enteritis	0.644247
Lashui	Diarrhoea	0.638210
Bianmi	Astriction	0.635700
Fuzhang	Ventosity	0.630602
Xiaohuabuliao	Indigestion	0.607682
Bianxie	Hematochezia	0.607306
Xibian	Diarrhea	0.605745
Xuebian	Hematochezia	0.604145
Lvbian	Greenish Stool	0.600339
Changming	Borborygmi	0.598407
Outu	Anabole	0.588085
Dutong	Collywobbles	0.581591
Shangtuxiaxie	Vomiting and Diarrhea	0.573924
Weichangyan	Gastroenteritis	0.573915

TABLE 5. Angology for word embedding.

Words	Similarity
Smecta	0.572514
Berberine	0.538834
Bifico	0.538151
Montmorillonite	0.519070
Siliankang	0.514807

Table 5 presents the remarkable results of word embedding in the analysis of an inter-word relationship. Table 5 also presents the results of the inquiry concerning what word pairs with "diarrhea", and the relationship between the formed pair is close to the relationship between "Qingkailing" (traditional Chinese medicine used for the common cold) and "common cold". In total, five words are listed in Table 5 as the closest inquiry results.

From the results in Table 4, we can see that the words close to "diarrhea" are all related to gastrointestinal diseases, which include both symptom and disease names. From the results in Table 5, we find that the results are all medicines used to cure diarrhea, and the relationship between the resultant words and "diarrhea" is close to the relationship between "Qingkailing" and "common cold". The two examples provided above prove that the word embedding is rich in semantics.

In this paper, the trained word embedding will be applied to the disease diagnosis model proposed in this paper.

C. CONVOLUTIONAL NEUTRAL NETWORK GUIDANCE MODEL

In this section, we introduce the experimental procedures and result analysis of the model for artificial-intelligence-guided disease diagnosis. Before the experiment, we first analyze the data and determine the range of artificial-intelligence-guided disease diagnoses. Then, we label the data (thereby generating the training model of the training datasets) and use both the validation method to verify the model's effectiveness and the method of artificial assessment to evaluate the model.

1) STATISTICAL ANALYSIS OF THE DATA

For the 7,835 pieces of disease data in dataset 1, we use the disease names to match the data after combining dataset 4 and dataset 5. If there is a matched disease, we correlate this inquiry with the disease. The top 10 matched diseases and their number of matches are listed in Table 6.

TABLE 6. Top 10 for disease match.

Disease	Amount	Precision(%)
Common cold	99483	2.973477133
Menoxenia	99170	2.964121782
Missed miscarriage	92325	2.75952953
Miscarriage	74799	2.235689676
Vaginitis	67358	2.013283402
Vitiligo	51147	1.528747976
Endocrine dyscrasia	45090	1.347708492
Hemorrhoid	43472	1.299347606
Hypertension	40027	1.196378971
Diabetes mellitus	39678	1.185947606

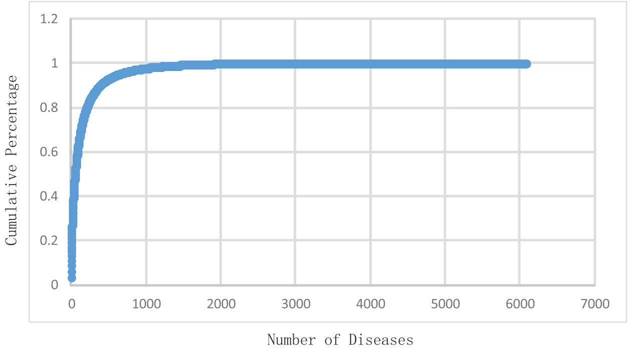


FIGURE 6. Disease Amount and Statistic Percentage.

The data statistics reveal that the diseases that rank at the top of the statistics are all common diseases. Figure 6 shows the statistics for the number of times diseases are matched and the corresponding percentage, where the horizontal axis is the number of matches and the vertical axis is the cumulative percentage. According to the statistics, the cumulative percentage of disease data whose number of matches ranks in the first 500 is 92.5%. Because the introduction of an excessive number of classifications will cause excessive training bias and inaccuracies, we divide all the classifications into 500 categories based on the statistical data results.

2) TRAINING OF THE GUIDANCE MODEL AND EVALUATION OF THE RESULTS

In this section, we will describe the screening and application process for the inquiry data using the above-mentioned method as the training data used to train the CNN to classify the text. We first introduce the screening process for the training data and then subsequently introduce the training method for the model and the evaluation of the results.

a: SCREENING OF THE TRAINING DATA

We obtained the medical question/answer data with disease labels through the process described in the previous sections. The medical question/answer data generally include two parts. One part is the symptom description, which assists doctors in diagnosing the disease, and the other part consists of understanding the relevant knowledge regarding a certain disease. The data required in this section are the former, namely, the inquiry of symptoms and the disease label.

Because we must screen the inquiry sentences with symptoms, the usage of the CRF model described above is shown here. Regarding the screening method for the sentences expressing the symptom inquiry, in this paper, we apply the CRF model to perform NER of symptoms on the inquiries to be screened. If for one sentence we can recognize two or more symptoms, we consider this sentence to be a symptom inquiry. In this paper, we use the aforementioned rules to screen the inquiry dataset, and a total of 507,411 pieces of inquiry data are screened. The disease distribution is roughly the same as the overall data distribution.

b: FEATURE EXTRACTION AND TRAINING

For the screened training data, we extract the features and adapt them to the CNN input. After inputting a sentence of the symptom inquiry, we first perform the word segmentation to eliminate the stop word and find the corresponding embedding vector of this word in the pre-trained set of embedding vectors as one row of the input matrix. In particular, if an input word is a symptom named entity or part of a symptom named entity, the value of each dimensionality in the vector is multiplied by two; this is done to strengthen the characterization role of the symptom in the text. After obtaining the input matrix, we take it as the input of the CNN, and its disease label is used as the classification category to train the CNN.

c: EVALUATION OF THE RESULTS

As a multi-classification problem, the evaluation method of this paper uses 10-fold cross-validation for the classification question as the method of model evaluation. In addition, according to the specificity of the medical questions, we also apply the sampling method to invite medical experts with professional knowledge in the medical field to evaluate the results.

The 10-fold cross validation divides the training data into 10 folds on average. At each iteration, we take the 9 folds as the training data for model training, and the other fold is used as the testing data to measure the accuracy of the computation model. In particular, the equation for calculating the precision is

$$precision = \frac{|correct|}{|predict|} \tag{10}$$

where $|correct|$ is the number of correct elements, namely the intersection of the output result and the standard result, and $|predict|$ is the number of elements in the output results.

Cross validation is repeated 10 times, and each piece of data is validated once. Eventually, we obtain the model's validation accuracy. Figure 7 shows the statistics concerning the accuracy for a validation result, where the horizontal axis is the number of experiments and the vertical axis is the accuracy of the experiments. The validation reveals that the accuracy of the present model for the task of disease diagnosis is 71.7%.

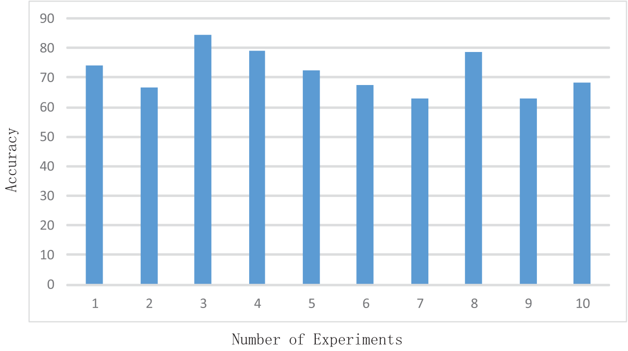


FIGURE 7. Result for Cross Validation.

TABLE 7. Precision with different answer counts.

Answer Number	Precision
1	71.72%
2	76.48%
3	79.35%
4	83.75%
5	85.41%

We can see from the results that, if only one disease name can be returned as the final answer, the results are not ideal. Table 7 shows the variation of the system accuracy with varying number of returned answers; if the standard answer is

TABLE 8. Case Study for evaluation.

Query	Answer	
	Disease	Precision
I got a bad headache and diarrhea. Is it the stomach flu?	Catch a cold (0)	0.362297429
	Diarrhea (17)	0.171501546
	Enteritis (31)	0.163804521
	Gastroenteritis (155)	0.153568569
	Gastroenteropathy (147)	0.148827935
Sometimes I have the mind on negative things, what's wrong with me?	Depression (57)	0.505206837
	Neurosis (244)	0.130453424
	Fetal malformation (307)	0.123073337
	Gastroenteropathy (147)	0.121412088
	Neurological Clisorders (274)	0.119854314
I get hiccups long after lunch. How to solve it?	Gastritis (12)	0.221776793
	Gastroenteropathy (147)	0.197842543
	Get inflamed (15)	0.195111134
	Gastroenteritis (155)	0.194377371
	Neurological Clisorders (274)	0.190892159
How to deal with insomnia?	Neurasthenia (46)	0.434840049
	Somnipathy (275)	0.174691083
	Neurological Clisorders (274)	0.139844428
	Chills (357)	0.125334927
	Damp-heat in the spleen and the stomach (312)	0.125289513
There are a little keloid on my palm. I feel itches when it exposed to heat.	Wart (60)	0.203118277
	Tinea manuum (437)	0.201948516
	Eczema (13)	0.199119034
	Urticaria (37)	0.198089090
	Hemorrhoid (7)	0.197725084

TABLE 9. Precision according to Manual Evaluation.

Answer Number	Precision
1	71%
2	75%
3	81%
4	84%
5	90%

among the answers returned by the system, the system answer is considered to be correct.

To evaluate the diagnosis capability of the diagnosis model in this paper, we randomly extract 100 inquiry data of symptoms from other data resources and use the diagnosis model to generate the results, which are labeled and evaluated by doctors as a case study. Table 8 lists a portion of the evaluation data, where age and gender are set as default values, and Table 9 shows the accuracy of this model, as evaluated by the doctors.

VII. CONCLUSION AND PROSPECT

In this paper, we introduced a novel framework of medical guidance. Specifically, we first annotated the extracted disease described by the users with the natural-language disease inquiry through named entity recognition (NER). Then, we employed word embedding to convert the disease inquiry data of users for the matrix expression. Finally, we proposed a model for artificial-intelligence-guided disease diagnosis and eventually provided intelligent guidance for disease inquiry for newly acquired users. We evaluated our methods with extensive experiments on real-world datasets. The experimental results clearly validate the effectiveness of our methods.

Potentially, this study has many future research directions. First, it would be interesting to investigate new guide diagnosis models to improve the accuracy of online medical guidance. Second, we plan to develop a medical online service platform basing on the models proposed in this paper to provide convenient online service.

REFERENCES

[1] S. Mertens, F. Gailly, and G. Poels, "Supporting and assisting the execution of flexible healthcare processes," in *Proc. Int. Conf. Pervas. Comput. Technol. Healthcare*, 2015, pp. 375–388.

[2] Y. Zhang, L. Sun, H. Song, and X. Cao, "Ubiquitous WSN for healthcare: Recent advances and future prospects," *IEEE Internet Things J.*, vol. 1, no. 4, pp. 311–318, Aug. 2014.

[3] M. R. Friesen and R. D. McLeod, "A survey of agent-based modeling of hospital environments," *IEEE Access*, vol. 2, pp. 227–233, 2014.

[4] R. S. Ledley and L. B. Lusted, "Reasoning foundations of medical diagnosis," *MD Comput., Comput. Med. Pract.*, vol. 8, no. 5, p. 300, 1991.

[5] W. R. Willcox, S. P. Lapage, S. Bascomb, and M. A. Curtis, "Identification of bacteria by computer: Theory and programming," *Microbiology*, vol. 77, no. 2, pp. 317–330, 1973.

[6] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[7] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, 2010.

[8] A. T. Sadiq and N. T. Mahmood, "A hybrid estimation system for medical diagnosis using modified full Bayesian classifier and artificial bee colony," *Iraqi J. Sci.*, vol. 55, no. 3A, pp. 1095–1107, 2014.

[9] R. R. Patil, "Heart disease prediction system using naive Bayes and Jelinek-mercer smoothing," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 3, no. 5, pp. 6787–6789, 2014.

[10] R. J. van Amstel, J. R. Anema, K. Jettinghoff, J. H. Verbeek, A. P. Nauta, and D. J. van Putten, "limited change in the quality of the social medical guidance and in the satisfaction of sick-listed patients, after collaborative projects between general practitioners and occupational physicians," *Nederlands Tijdschrift Geneeskunde*, vol. 149, no. 43, pp. 2407–2412, Oct. 2005. [Online]. Available: http://europepmc.org/abstract/MED/16277131

[11] Y. S. Lin, L. Huang, and Z. M. Wang, "An intelligent medical guidance system based on multi-words TF-IDF algorithm," in *Proc. 2nd Int. Conf. Energy Sci. Appl. Technol. (ESAT)*, 2015, p. 385.

[12] B. Mohit, "Named entity recognition," in *Natural Language Processing of Semitic Languages*. Springer, 2014, pp. 221–245.

[13] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Invest.*, vol. 30, no. 1, pp. 3–26, 2007.

[14] B. Settles, "Biomedical named entity recognition using conditional random fields and rich feature sets," in *Proc. Int. Joint Workshop Natural Lang. Process. Biomed. Appl.*, 2004, pp. 104–107.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[16] Y. Kim. (2014). "Convolutional neural networks for sentence classification." [Online]. Available: http://arxiv.org/abs/1408.5882

[17] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. (2012). "Improving neural networks by preventing co-adaptation of feature detectors." [Online]. Available: http://arxiv.org/abs/1207.0580

[18] S. Park and N. Kwak, "Cultural event recognition by subregion classification with convolutional neural network," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 45–50.

[19] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2177–2185.

[20] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

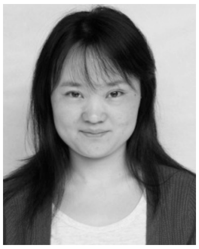
CUILI YAO received the B.S. degree in information and computing science and the master's degree in applied mathematics from Northeastern University, Shenyang, China, in 2007 and 2009, respectively. She is currently pursuing the Ph.D. degree with the School of Computer Science, Dalian University of Technology, Dalian, China. Her research interests include data mining, information system, and intelligent computing.

YUE QU received the bachelor's degree in biomedical engineering from the Dalian University of Technology, Dalian, China, in 2012. His research interests include data mining, information system, and intelligent computing.

BO JIN received the B.S. degree in mechanical design automation and the Ph.D. degree in computer application technology from the Dalian University of Technology, Dalian, China, in 2001 and 2009, respectively. He is currently an Associate Professor with the School of Innovation and Entrepreneurship, Dalian University of Technology. His research interests include information retrieval, data mining, and intelligent computing.

LI GUO received the B.S. degree in mechanical design automation and the master's degree in computer application technology from the Dalian University of Technology, Dalian, China, in 2002 and 2005, respectively, where he is currently pursuing the Ph.D. degree with the School of Computer Science. His research interests include data mining, information system, and intelligent computing.

CHAO LI received the B.S. degree in computer science and the master's degree in computer application technology from the Dalian University of Technology, Dalian, China, in 2013 and 2016, respectively. His research interests include data mining, information system, and intelligent computing.



WENJUAN CUI received the B.Sc. degree from Shandong University, China, in 2009, and the Ph.D. degree from the City University of Hong Kong in 2013. She is currently an Assistant Professor with the Computer Network Information Center, Chinese Academy of Sciences. Her research interests include data mining, recommender systems, semantic analysis, big data processing and bioinformatics.

LIN FENG received the B.S. degree in electronic technology, the M.S. degree in power engineering, and the Ph.D. degree in mechanical design and theory from the Dalian University of Technology, China, in 1992, 1995, and 2004, respectively. He is currently a Professor and Doctoral Supervisor with the School of Innovation and Entrepreneurship, Dalian University of Technology, China. His research interests include intelligent image processing, robotics, data mining, and embedded systems.