IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 13, 2019, accepted June 11, 2019, date of publication June 14, 2019, date of current version July 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923221

Automatically Structuring on Chinese Ultrasound Report of Cerebrovascular Diseases via Natural Language Processing

PENGYU CHEN[1], QIAO LIU[2,3,4], LAN WEI[1], BEIER ZHAO[5], YIN JIA[5], HAIRONG LV[2,3,4], AND XIAOLU FEI[1]

[1]Information Center, Xuanwu Hospital, Capital Medical University, Beijing 100053, China
[2]Ministry of Education Key Laboratory of Bioinformatics, Tsinghua University, Beijing 100084, China
[3]Bioinformatics Division, Center for Synthetic and Systems Biology, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
[4]Department of Automation, Tsinghua University, Beijing 100084, China
[5]Beijing INFI-SAGACITY Technology Company, Ltd., Beijing 100086, China

Corresponding authors: Hairong Lv (lvhairong@tsinghua.edu.cn) and Xiaolu Fei (feixiaolu@xwh.ccmu.edu.cn)

This work was supported in part by the Capital Medical University Research and Development Fund under Grant PYZ2018125, in part by the National Natural Science Foundation of China under Grant NSFC U1736210, in part by the National Key Research and Development Program of China under Grant SQ2018YFC090002, and in part by the Tsinghua-Fuzhou Institute for Data Technology under Grant TFIDT2018004.

ABSTRACT The current ultrasound reports in Chinese hospitals are mostly written in free-text format. Important clinical information, such as stenosis rate and plaque location, is recorded in long sentences, especially for ultrasound reports of cerebrovascular diseases. They cannot be directly used for further automatic analysis due to the lack of structure and standardization. The goal of this paper is to assess the feasibility of applying natural language processing technology to automatically extract disease entities and relate information such as the stenosis rate and plaque location from free-text ultrasound reports of cerebrovascular diseases. A structured model using conditional random fields (CRFs) is first constructed. Then, the clause optimizing and segmentation process is performed on a word level to achieve data structuring. Seven categories of terms, including symptoms, plaque locations, diseases, and degree, in 1980 de-identified ultrasound reports were manually annotated as a training dataset. With this model, 7937 ultrasound reports were automatically processed to structure data within 40 min. The true positive rate of the model for each category of terms is 96%, 94%, 97%, 100%, 100%, 100%, and 97%, respectively. The CRF model can be used in Chinese natural language processing to provide support for unstructured data analysis. The standardized segmentation results can be obtained based on medical ontology libraries. However, real-time processing and scientific annotation remain a challenge if intelligent clinical decision making needs to be applied to a real-world clinical environment.

INDEX TERMS Natural language processing (NLP), conditional random fields (CRF), ultrasound report.

I. INTRODUCTION

Stroke (cerebrovascular disease) is the fifth leading cause of death in the United States and is a major cause of serious disability for adults [1], [2]. Approximately 795,000 people in the United States have a stroke each year [2]. In China, stroke has been the leading cause of death in recent years [3], constituting almost one-third of the total number of deaths from stroke worldwide [4]. By 2013, 27 of 33 provinces in China had stroke as the leading cause of death [5]. Many actions have been taken to reduce the prevalence of cerebrovascular disease in China. Head and neck vascular ultrasound is one of the most popular methods of screening people with cerebrovascular disease because of its high accuracy, mobility, safety, and low cost. Head and neck vascular ultrasound can also provide reliable objective imaging and measured data for clinicians to select the appropriate treatment methods accordingly.

Recently, artificial intelligence has been investigated to be applied to improve stroke diagnosis and treatment [6], [7].

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

The results of head and neck vascular ultrasound examinations are very useful data for developing methods to automatically recognize stroke-threatening findings to prevent cerebrovascular disease regarding its efficacy in detecting bleeding. Although the quantity of head and neck ultrasound reports has been very large in China [8], important clinical information such as stenosis rates and plaque locations still cannot be used in automatic stroke risk analysis directly because most of the reports are recorded in free text format and only be interpreted can by physicians manually, which is very time-consuming.

Natural language processing (NLP) can provide a way to analyze a large number of documents and structure unstructured text information for data processing for further analysis [9]–[11]. In this study, we investigated developing a Chinese NLP pipeline using conditional random fields (CRF) to extract the significant attributes from ultrasound reports of cerebrovascular diseases. The extracted attributes provide support for performing further statistical analysis and generate intelligent treatment suggestions concerning cerebrovascular diseases.

II. METHODS

In this study, 1,980 reports were manually annotated to train and validate the CRF model and then 7,937 reports were processed using this model. Python on the Linux platform was used. After extracting the key attributes through CRF, structured attributes can be obtained based on a self-developed word segmentation tool, especially for unknown words.

A. CRF MODEL IMPLEMENTATION

CRF is a class of statistical modeling methods used for structured prediction in natural language processing that is a discriminant probability undirected graph learning model proposed by Lafferty in 2001 based on the maximum entropy model and hidden Markov model [12], [13]. The CRF model performs particularly well for labeling and segmenting ordered data [14].

The CRF model can be used to express contextual relationships and describe the characteristics of overlapping and long-distance dependencies between nodes [15]. Feature sets can be arbitrarily selected according to the characteristics of text medical records, and global optimal values are obtained after globally normalizing all features, which avoids the problem of label offset.

Let $G = (V, E)$ be an undirected graph, and $Y = \{Y_v | v \in V\}$ is a set of random variables indexed by node v in G. Given the condition of X, if each random variable Y_v obeys the Markov attribute, it is represented by the following formula:

$$p\left(Y_v \mid X, Y_w \mu \neq v\right) = p\left(Y_v \mid X, Y_w \mu \sim v\right)$$

Then (X, Y) constitutes a conditional random field, where $\mu \sim v$ denotes that μ and v are adjacent edges. The simplest and most commonly used is a first-order chain structure, such as a linear chain structure. The definition is provided below.

Let $\mathrm{x} = (x_1, x_2, \ldots, x_n)$ denote the observation sequence and $\mathrm{y} = \{y_1, y_2, \ldots, y_n$ is a set of finite states. Based on the basic theory of random fields, we obtain the following formula:

$$p(\mathrm{y}|\mathrm{x}, \lambda) \propto \exp(\sum_j \lambda_j t_j\left(y_{i-1}, y_i, x, i\right) + \sum_k \mu_k s_k(y_i, x, i))$$

where $t_j\left(y_{i-1}, y_i, x, i\right)$ is the transfer eigenfunction between $i-1$ and i for the observed sequence, and $s_k(y_i, x, i)$ is the state eigenfunction at the position of the observed sequence i. The two eigenfunctions are unified to $f_j(y_{i-1}, y_i, x, i)$. So the above formula can be described as the following two formulas:

$$p\left(\mathrm{y} \mid \mathrm{x}, \lambda\right) = \frac{1}{Z(x)} \exp(\sum_{i-1}^{n} \sum_j \lambda_j f_j(y_{i-1}, y_i, x, i))$$

$$\mathrm{Z}\left(\mathrm{x}\right) = \sum_j \exp(\lambda_j f_j(y_{i-1}, y_i, x, i))$$

From the above analysis, we can see that the linear chain conditional random field is completely determined by the eigenfunction t_j, s_k and the corresponding weights λ_j and μ_k.

B. DATA PREPROCESSING

We collected data from head and neck ultrasound reports and related medical records in a tertiary hospital with neurology specialty in China from 2015-2017 to build the training and testing dataset. The collected data must first be pre-processed for verification, including the three steps of data format unification, data cleaning, and data storage.

(1) **Data format unification**: The existence of data in the database includes multiple formats such as word, txt, and html. We analyzed the files in different formats and stored them in a unified manner.

(2) **Data cleaning**: We cleaned the unified data with errors, spaces, special symbols, etc., to express the data in a unified format. The main cleaning data include the following three categories:1) **Incomplete data**: We tested whether the data were incomplete or not. Generally, incomplete data refer to the fields that have no data associated with the medical record, which cannot be used by the system. 2) **Error data**: Error data generally refers to data in different parts conflict. For example, a patient was admitted to the hospital on May 12, 2015, but there was a discharge record on December 5, 2014. 3) **Duplicate data**: Duplicate data generally refer to repeated occurrences of the same data, such as two identical admission records in a medical record

(3) **Data storage**: The data that have been cleaned are stored in the database based on the patient's ID number and examination time as a unique identifier.

C. MANUAL ANNOTATION

The manual annotation of data was to establish a training set for the CRF model. We randomly selected 1,980 copies in 9,917 reports, and the data of the training set accounted for

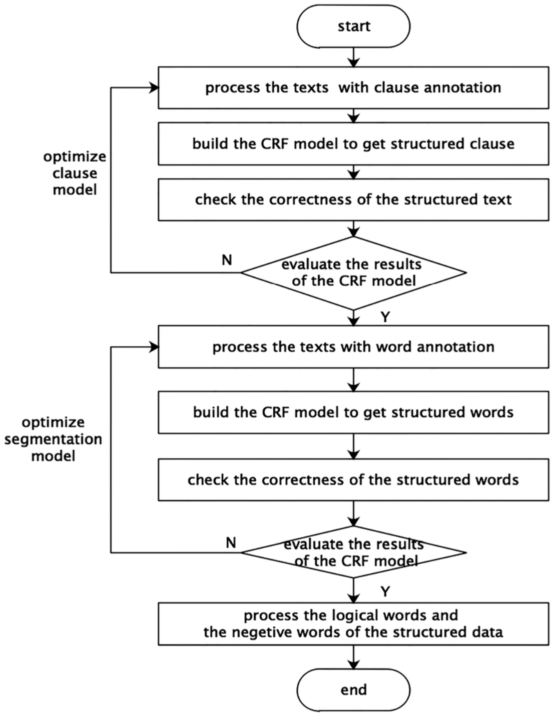

start
process the texts with clause annotation
build the CRF model to get structured clause
check the correctness of the structured text
evaluate the results of the CRF model
N
Y
optimize clause model
process the texts with word annotation
build the CRF model to get structured words
check the correctness of the structured words
evaluate the results of the CRF model
N
Y
optimize segmentation model
process the logical words and the negetive words of the structured data
end

FIGURE 1. **Flow chart of the data processing using the CRF model.**

approximately 20% of the total data. The corpus of reports was manually reviewed by physicians to label key attributes in each report into seven categories, including symptom, disease, three types of location, degree word, and stenosis rate. Sub-categories such as "plaque quantity" and "side" were annotated as second-level attributes under the symptom and location category. The labeled data were used as a gold standard for further word segmentation and classification. The model was built according to the seven categories of key attributes. Manual annotation of the neck vascular ultrasound report required a total of 165 hours.

D. THE PROPOSED FRAMEWORK

1) THE OVERALL PROCESS

First, through the system interface, we obtained substantial amounts of healthcare data and stored them in the database. Next, structured attributes were obtained based on the CRF model after analysis, cleaning, and desensitization. Finally, we assessed cerebrovascular disease specifically through statistical analysis. The process is illustrated in Figure 1.

Figure 1 shows that according to the data processing configuration file, the unstructured text data are first marked by a clause. Then the CRF model is created based on the result of the clause marking. After marking, the correctness of the structured data is checked, and then we evaluate the integrity of the segmentation. If it cannot be concluded, the clause model needs to be optimized until a predetermined threshold is reached. After marking the words, the word segmentation results are evaluated continuously until the predetermined threshold is reached. Finally, the structured processing data are processed logically and negatively.

TABLE 1. **Feature template.**

No	Feature template type	Feature Template
1	Unigram	U00:%x[-3,0]
2	Unigram	U01:%x[-2,0]
3	Unigram	U02:%x[-1,0]
4	Unigram	U03:%x[0,0]
5	Unigram	U04:%x[1,0]
6	Unigram	U05:%x[2,0]
7	Unigram	U06:%x[3,0]
8	Unigram	U07:%x[-3,0]/%x[-2,0]
9	Unigram	U08:%x[-2,0]/%x[-1,0]
10	Unigram	U09:%x[-1,0]/%x[0,0]
11	Unigram	U10:%x[0,0]/%x[1,0]
12	Unigram	U11:%x[1,0]/%x[2,0]
13	Unigram	U12:%x[2,0]/%x[3,0]
14	Bigram	B

When writing the feature templates, this study used Unigram, which could realize binary operation of one-dimensional features plus a binary feature of a Bigram template. Then we combined the adjacent features in Unigram to form features. The specific template is shown in Table 1.

2) PARAMETERS ADJUSTMENT

(1) **-a CRF-L2 or CRF-L1**: This is the normalized algorithm selection, of which the default is CRF-L2. This parameter is on the loss function, plus the square of the feature coefficient. The goal is to avoid overfitting and to make the parameters smaller and then more accurate. In general, the CRF-L2 algorithm is slightly better than the CRF-L1 algorithm, so we adopted the CRF-L2 algorithm.

(2) **-c float**: This sets the hyper-parameter of the CRF model. The larger the value of c, the higher the degree to which the CRF fits the training data. This parameter can adjust the balance between overfitting and underfitting. The larger the –c float, the higher the fit.

(3) **-f NUM**: This sets the cut-off threshold for the feature. In the training data CRF++ uses at least NUM features, of which the default is 1. This option will work when using CRF++ for large-scale data, as there may be millions of features that appear only once. Since our data volume does not reach the million level, we use the default value.

(4) **-p NUM**: NUM is the number of threads. If the computer has multiple CPUs, the training speed can be increased using multiple threads.

TABLE 2. The surgical description of the patients with stents before and after standardization.

Before standardisation	After standardisation
左侧颈动脉支架置入术后 Left carotid artery stent placement	颈总动脉远段支架置入术后 Distal common carotid artery stenting
右侧颈动脉支架置入术后 Right carotid artery stent placement	颈总动脉远段支架置入术后 Distal common carotid artery stenting
右侧颈总动脉远段及颈内动脉近段支架置入术后 Distal right common carotid artery and proximal carotid artery stent placement	颈总动脉远段支架置入术后 Distal common carotid artery stenting
右侧颈总动脉、颈内动脉支架术后 Right common carotid artery, internal carotid artery stent	颈总动脉远段支架置入术后 Distal common carotid artery stenting
右侧颈总动脉支架置入术后 Right common carotid artery stent placement	颈总动脉远段支架置入术后 Distal common carotid artery stenting
左侧颈总动脉支架植入术后 Left common carotid artery stent implantation	颈总动脉远段支架置入术后 Distal common carotid artery stenting
右侧颈总动脉分叉处及颈内动脉起始处支架置入术后 Right carotid bifurcation and internal carotid artery at the beginning of stent placement	颈总动脉远段支架置入术后 Distal common carotid artery stenting
左侧颈动脉支架术后 Left carotid artery stent	颈总动脉远段支架置入术后 Distal common carotid artery stenting
双侧颈动脉支架置入术后 Bilateral carotid artery stent placement	颈总动脉远段支架置入术后 Distal common carotid artery stenting
左侧颈总动脉、颈内动脉支架置入术后 Left common carotid artery, internal carotid artery stent placement	颈总动脉远段支架置入术后 Distal common carotid artery stenting
双侧颈总动脉、颈内动脉支架术后 Bilateral common carotid artery, internal carotid artery stent	颈总动脉远段支架置入术后 Distal common carotid artery stenting
左侧颈总动脉、颈内动脉支架置入术后 Left common carotid artery, internal carotid artery stent placement	颈内动脉支架置入术后 Internal carotid artery stenting
右侧颈总动脉、颈内动脉支架术后 Right common carotid artery, internal carotid artery stent	颈内动脉支架置入术后 Internal carotid artery stenting
双侧颈总动脉、颈内动脉支架术后 Bilateral common carotid artery, internal carotid artery stent	颈内动脉支架置入术后 Internal carotid artery stenting

3) DATA NORMALIZATION

After processing the data, some are still ambiguous and cannot be used further directly. Therefore, we standardized these data to avoid ambiguity and make them more authentic and easier to understand. For example, the standardization of the surgical description of patients with stents is demonstrated in Table 2.

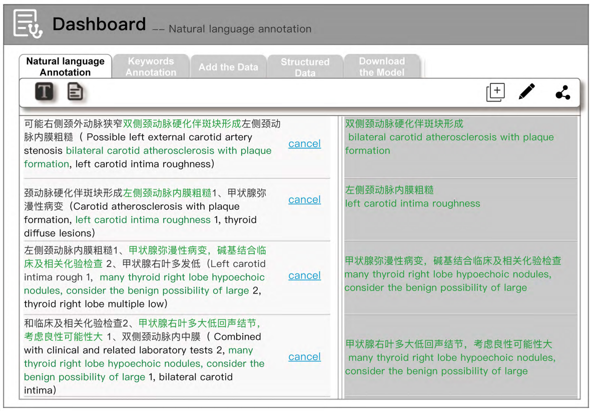
III. RESULTS

A total of 7,937 neck vascular ultrasound reports were processed with the CRF models and the total processing time was 40 minutes. The quantities of related words that are categorized into 7 types of attributes including the symptoms, diseases, Location 1, Location 2, Location 3, degree words, and stenosis rate are shown in Table 3.

The following is an example of the model implementation process. The first process level is the clause level. We need to segment a paragraph into clauses. Figures 2 and 3 illustrate the process with manual annotation (in green) of the training dataset and automatic segmentation results of the testing dataset (in red). After model training, the automatic segmentation results can reach the quality of manual annotation. We evaluated the segmentation results in a manner of classification considering different types of attributes (Table 4). In the type of disease, our CRF model achieves a precision of 0.958 and recall of 0.939, respectively. The F1-score, which considers both precision and call, can reach as high as 0.948. For the largest category, we collected 254 related words as ground truth, our CRF model also achieves a precision, recall, F1-score more than 0.94, which shows the consistent superior performance. It is also the most

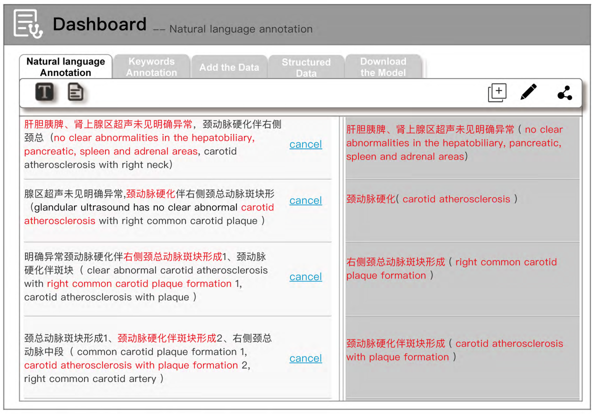
TABLE 3. The attributes and their quantity after processing.

Attribute category	Quantity
Disease	48
Symptoms	256
Plaque Location 1	150
Plaque Location 2	36
Plaque Location 3	10
Degree words	10
Stenosis degree	38


Dashboard -- Natural language annotation
Natural language Annotation
Keywords Annotation
Add the Data
Structured Data
Download the Model

可能右侧颈外动脉狭窄双侧颈动脉硬化伴斑块形成左侧颈动脉内膜粗糙（ Possible left external carotid artery stenosis bilateral carotid atherosclerosis with plaque formation, left carotid intima roughness)	cancel	双侧颈动脉硬化伴斑块形成 bilateral carotid atherosclerosis with plaque formation
颈动脉硬化伴斑块形成左侧颈动脉内膜粗糙1、甲状腺弥漫性病变（Carotid atherosclerosis with plaque formation, left carotid intima roughness 1, thyroid diffuse lesions)	cancel	左侧颈动脉内膜粗糙 left carotid intima roughness
左侧颈动脉内膜粗糙1、甲状腺弥漫性病变，碱基结合临床及相关化验检查 2、甲状腺右叶多发低（Left carotid intima rough 1, many thyroid right lobe hypoechoic nodules, consider the benign possibility of large 2, thyroid right lobe multiple low)	cancel	甲状腺弥漫性病变，碱基结合临床及相关化验检查 many thyroid right lobe hypoechoic nodules, consider the benign possibility of large
和临床及相关化验检查2、甲状腺右叶多大低回声结节，考虑良性可能性大 1、双侧颈动脉内中膜（ Combined with clinical and related laboratory tests 2, many thyroid right lobe hypoechoic nodules, consider the benign possibility of large 1, bilateral carotid intima)	cancel	甲状腺右叶多大低回声结节，考虑良性可能性大 many thyroid right lobe hypoechoic nodules, consider the benign possibility of large

FIGURE 2. The model interface of the manual annotation on the clause level.


Dashboard -- Natural language annotation
Natural language Annotation
Keywords Annotation
Add the Data
Structured Data
Download the Model

肝胆胰脾、肾上腺区超声未见明确异常，颈动脉硬化伴右侧颈总（no clear abnormalities in the hepatobiliary, pancreatic, spleen and adrenal areas, carotid atherosclerosis with right neck)	cancel	肝胆胰脾、肾上腺区超声未见明确异常（ no clear abnormalities in the hepatobiliary, pancreatic, spleen and adrenal areas)
腺区超声未见明确异常,颈动脉硬化伴右侧颈总动脉斑块形（glandular ultrasound has no clear abnormal carotid atherosclerosis with right common carotid plaque ）	cancel	颈动脉硬化(carotid atherosclerosis)
明确异常颈动脉硬化伴右侧颈总动脉斑块形成1、颈动脉硬化伴斑块（ clear abnormal carotid atherosclerosis with right common carotid plaque formation 1, carotid atherosclerosis with plaque ）	cancel	右侧颈总动脉斑块形成（ right common carotid plaque formation ）
颈总动脉斑块形成1、颈动脉硬化伴斑块形成2、右侧颈总动脉中段（ common carotid plaque formation 1, carotid atherosclerosis with plaque formation 2, right common carotid artery ）	cancel	颈动脉硬化伴斑块形成（ carotid atherosclerosis with plaque formation ）

FIGURE 3. The model interface of the automatic segmentation results on the clause level.

challenging type for segmentation. As for the rare types such as plaque location 3 and degree words, our CRF models successfully identifies all the words correctly. For the last type stenosis degree, only one false word was misclassified as true one by our CRF model. If we consider all types of words together, our CRF model can achieve a precision of 0.958, a recall of 0.949 and a F1-score of 0.954. The high measuring metrics show the superior segmentation ability of our CRF model.

The key attributes are the segmented on a word level. Figures 4 and 5 illustrate the process of "Plaque Location 1" category segmentation with the manual annotation (in green) of the training dataset and automatic segmentation of the testing dataset (in red).

Finally, the segmented key attributes are stored in structured tables that can be used for statistical and other analyses.

IV. DISCUSSION

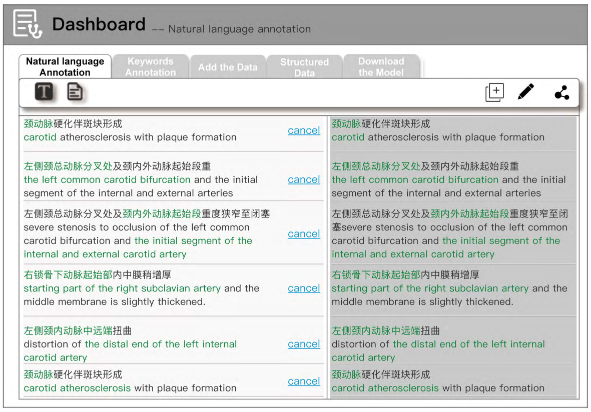
Currently, healthcare information systems such as electronic medical records (EMR) and radiology information systems (RIS) have become very popular in China and play important roles in hospital management. In a survey by the Chinese Hospital Information Management Association (CHIMA), EMR and RIS were continuously among the top 3 most important healthcare information systems from 2013 to 2016. Meanwhile, although many clinical data have already been digitized in China, large amounts of key information that can offer valuable insights still cannot be used in statistical analysis and decision making. The greatest obstacle is that the information is recorded in long sentences and has diversity of expression in the Chinese language. Therefore, to make further use of the vast amount of information in these clinical documents, developing an automatic natural language processing tool for the Chinese language to realize the automatic structuring and standardization of key clinical information with minimal human intervention is necessary and urgent.

CRF is a mature method for word segmentation and this study was conducted mainly to verify its efficacy and efficiency in processing Chinese healthcare words. The results showed that the CRF model can be used in Chinese natural language processing to provide support for unstructured data analysis. Its efficacy and efficiency both meet the needs of utilization of clinical documents in Chinese. Additionally, to verify whether the established model is restricted to certain specific examinations or can be widely used in healthcare textual data segmentation, this model can be also applied for processing medical records of physician examinations. Future research should expand the data sources to other electronic records such as admission notes and treatment records.

Standardization and normalization are very import processes to assist with data usage in statistical analysis. For example, descriptions of stenosis may vary between ultrasound doctors because of different educational backgrounds or work experience. The expression of stenosis in ultrasound report templates may change over time. Additionally, full-width and half-width characters are also substantial noise factors in the Chinese language. Under such conditions, it would be very difficult to accurately filter all patients with severe stenosis (75% and above) of the carotid artery who had not undergone surgery and were examined within one year if data standardization and normalization were not performed. One of the advantages of our study is the availability of the medical ontology library in Chinese. This library came

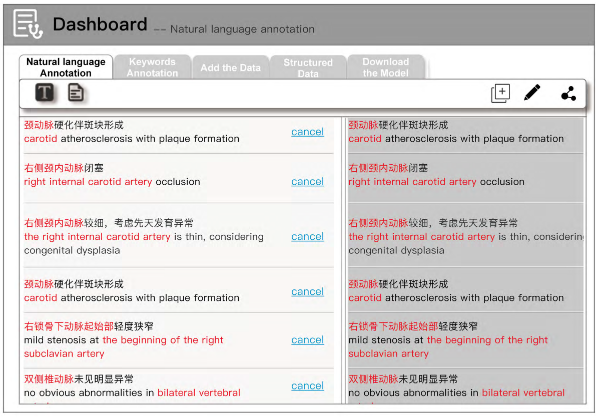
TABLE 4. The accuracy and specificity of the model for the seven categories.

Attribute category	True positives	False positives	False negatives	Precision	Recall	F1-score
Disease	46	2	3	0.958	0.939	0.948
Symptoms	241	15	13	0.941	0.949	0.945
Plaque Location 1	145	5	8	0.967	0.948	0.957
Plaque Location 2	36	0	2	1	0.947	0.973
Plaque Location 3	10	0	0	1	1	1
Degree words	10	0	0	1	1	1
Stenosis degree	37	1	2	0.974	0.949	0.961


Dashboard -- Natural language annotation
Natural language Annotation
Keywords Annotation
Add the Data
Structured Data
Download the Model
T

颈动脉硬化伴斑块形成 carotid atherosclerosis with plaque formation	cancel	颈动脉硬化伴斑块形成 carotid atherosclerosis with plaque formation
左侧颈总动脉分叉处及颈内外动脉起始段重 the left common carotid bifurcation and the initial segment of the internal and external arteries	cancel	左侧颈总动脉分叉处及颈内外动脉起始段重 the left common carotid bifurcation and the initial segment of the internal and external arteries
左侧颈总动脉分叉处及颈内外动脉起始段重度狭窄至闭塞 severe stenosis to occlusion of the left common carotid bifurcation and the initial segment of the internal and external carotid artery	cancel	左侧颈总动脉分叉处及颈内外动脉起始段重度狭窄至闭塞severe stenosis to occlusion of the left common carotid bifurcation and the initial segment of the internal and external carotid artery
右锁骨下动脉起始部内中膜稍增厚 starting part of the right subclavian artery and the middle membrane is slightly thickened.	cancel	右锁骨下动脉起始部内中膜稍增厚 starting part of the right subclavian artery and the middle membrane is slightly thickened.
左侧颈内动脉中远端扭曲 distortion of the distal end of the left internal carotid artery	cancel	左侧颈内动脉中远端扭曲 distortion of the distal end of the left internal carotid artery
颈动脉硬化伴斑块形成 carotid atherosclerosis with plaque formation	cancel	颈动脉硬化伴斑块形成 carotid atherosclerosis with plaque formation

FIGURE 4. The model interface of manual annotation of "Location 1" category on the word level.


Dashboard -- Natural language annotation
Natural language Annotation
Keywords Annotation
Add the Data
Structured Data
Download the Model
T

颈动脉硬化伴斑块形成 carotid atherosclerosis with plaque formation	cancel	颈动脉硬化伴斑块形成 carotid atherosclerosis with plaque formation
右侧颈内动脉闭塞 right internal carotid artery occlusion	cancel	右侧颈内动脉闭塞 right internal carotid artery occlusion
右侧颈内动脉较细，考虑先天发育异常 the right internal carotid artery is thin, considering congenital dysplasia	cancel	右侧颈内动脉较细，考虑先天发育异常 the right internal carotid artery is thin, considerin congenital dysplasia
颈动脉硬化伴斑块形成 carotid atherosclerosis with plaque formation	cancel	颈动脉硬化伴斑块形成 carotid atherosclerosis with plaque formation
右锁骨下动脉起始部轻度狭窄 mild stenosis at the beginning of the right subclavian artery	cancel	右锁骨下动脉起始部轻度狭窄 mild stenosis at the beginning of the right subclavian artery
双侧椎动脉未见明显异常 no obvious abnormalities in bilateral vertebral	cancel	双侧椎动脉未见明显异常 no obvious abnormalities in bilateral vertebral

FIGURE 5. The model interface of automatic segmentation of "Location 1" category on the word level.

from the combination of the ICD 10 (Beijing version), ICD 10 (military version), and Practical Internal Medicine 15th edition.

In this study, we adopted ICD Beijing version and some related knowledge base designed by ourselves.

ICD10 Beijing version keep the same with the category (One uppercase English character followed by two 0~9 digits) and subcategory (one 0~9 digit) of ICD-10 coding system. At the same time, it extended the coding structure in the following two aspects:

1. It extends detailed categories and extended nodes under detailed categories itself.
2. it adds a new table in which the elements are diagnoses actually used in clinical environment. Those elements can be exactly mapped to certain nodes under detailed categories

The segmented attributes were mapped with the terminologies in the library according to the normalization relationships built by the training dataset after they were labeled using the CRF model. The query with this terminology library more efficiently standardizes textual data.

The CRF models built in this study managed more than 7,900 reports within only 40 minutes, an average of approximately 0.3 second per report. With its fast calculation speed, it can manage many annoying texts and effectively combine prior knowledge. In clinical application, this model can provide clinicians with a relevant list of disease identifications rather than a large number of text reports, which will make it easier for clinicians to analyze those data. It can also be provided to Clinical Decision Support Systems (CDSS), which can give more precise suggestions about clinical treatment.

In future research, we will attempt to use different methods to improve the speed of natural language processing and achieve real-time processing. We expect that while the ultrasound doctor is writing the textual ultrasound report, the free-text report can quickly generate a structured, well-marked report and form an annotation on the ultrasound image. After the ultrasound doctor completes the written report, the generated annotations are checked, the labels that accurately reflect the characteristics of the images are retained, and the correlation between the text of the ultrasound reports and the images is established. While the medical process is realized, this association can complete the scientific annotation of the ultrasound image data and improve the follow-up analysis of the ultrasound image. This association will not only reduce the workload of labeling images, but also reduces the errors caused by the different labeling personnel and ultrasound doctors, which could improve the accuracy and reliability of image annotation.

V. CONCLUSION

In this study, we proposed a CRF model for automatically segmentation of Chinese ultrasound reports. The CRF model built in this study showed a superior performance of accuracy and efficiency in processing neck vascular ultrasound reports in the Chinese language, which demonstrates that CRF model can well capture the textual information of different categories. The demonstrated CRF model shows great power in automatically segmenting Chinese ultrasound reports into several defined categories. One major application of our CRF model is for indexing Chinese ultrasound reports for obtaining the specific terms, such as drug related terms. Our model can also be applied in the processing of Chinese ultrasound reports classification. The automatically structured ultrasound reports can be further utilized in further downstream applications.

Our CRF model can further be improved in the following aspects. First, the current model can only handle with pre-defined categories. It may give wrong classification when dealing with terms that are not included in pre-defined categories. Second, it can also be extended to other kind of electric medical record. Thus the generalization ability across different electric medical record can be measured. Third, further research into building models to establish more specific relationships between reports and objective images should also be performed to improve the data reliability.

To sum up, we proposed an automatically segmentation framework based on CRF model. Our model can be used in real-world Chinese healthcare information systems to help improve data usability. Hope our model can shed light on the processing of medical electric record data.

ACKNOWLEDGMENT

(Pengyu Chen and Qiao Liu contributed equally to this work.)

REFERENCES

[1] D. Mozaffarian, E. J. Benjamin, A. S. Go, D. K. Arnett, M. J. Blaha, M. Cushman, and V. J. Howard, "Executive summary: Heart disease and stroke statistics—2016 update: A report from the American Heart Association," *Circulation*, vol. 133, no. 4, pp. 447–454, 2016.

[2] D. Mozaffarian, E. J. Benjamin, A. S. Go, D. K. Arnett, M. J. Blaha, M. Cushman, and M. D. Huffman, "Executive summary: Heart disease and stroke statistics—2015 update: A report from the American Heart Association," *Circulation*, vol. 131, no. 4, pp. 434–441, 2015.

[3] L. Liu, D. Wang, K. L. Wong, and Y. Wang, "Stroke and stroke care in China: Huge burden, significant workload, and a national priority," *Stroke*, vol. 42, no. 12, pp. 3651–3654, 2011.

[4] V. L. Feigin, R.V. Krishnamurthi, P. Parmar, B. Norrving, G. A. Mensah, D. A. Bennett, S. Barker-Collo, A. E. Moran, R. L. Sacco, T. Truelsen, S. Davis, J. D. Pandian, M. Naghavi, M. H. Forouzanfar, G. Nguyen, C. O. Johnson, T. Vos, A. Meretoja, C. J. L. Murray, and G. A. Roth, "Update on the global burden of ischemic and hemorrhagic stroke in 1990-2013: The GBD 2013 study," *Neuroepidemiology*, vol. 45, no. 3, pp. 161–176, 2015.

[5] M. Zhou *et al.*, "Cause-specific mortality for 240 causes in China during 1990–2013: A systematic subnational analysis for the Global Burden of Disease Study 2013," *LANCET*, vol. 387, no. 10015, pp. 251–272, 2016.

[6] D. L. Labovitz, L. Shafner, M. R. Gil, D. Virmani, and A. Hanina, "Using artificial intelligence to reduce the risk of nonadherence in patients on anticoagulation therapy," *Stroke*, vol. 48, no. 5, pp. 1416–1419, 2017.

[7] L. Mirtskhulava, J. Wong, S. Al-Majeed, and G. Pearce, "Artificial neural network model in stroke diagnosis," in *Proc. 17th UKSim-AMSS Int. Conf. Modelling Simulation (UKSim)*, Mar. 2015, pp. 50–53.

[8] Y. Sun, H. Gregersen, and W. Yuan, "Chinese health care system and clinical epidemiology," *Clin. Epidemiol.*, vol. 9, p. 167, Mar. 2017.

[9] A. Kao and S. R. Poteet, *Natural Language Processing and Text Mining*. Springer, 2007.

[10] T.-T. Kuo, P. Rao, C. Maehara, S. Doan, J. D. Chaparro, M. E. Day, C. Farcas, L. Ohno-Machado, and C.-N. Hsu, "Ensembles of NLP tools for data element extraction from clinical notes," in *Proc. AMIA Annu. Symp.*, 2016, pp. 1880–1889.

[11] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?" *J. Biomed. Inform.*, vol. 42, no. 5, pp. 760–772, 2009.

[12] M. Hayashida, M. Kamada, J. Song, and T. Akutsu, "Prediction of protein-RNA residue-base contacts using two-dimensional conditional random field with the lasso," *BMC Syst. Biol.*, vol. 7, no. 2, p. S15, 2013.

[13] M. J. Shafiee, Z. Azimifar, and A. Wong, "A deep-structured conditional random field model for object silhouette tracking," *PLoS ONE*, vol. 10, no. 8, 2015, Art. no. e0133036.

[14] A. R. Kinjo, "Profile conditional random fields for modeling protein families with structural information," *Biophysics*, vol. 5, pp. 37–44, 2009.

[15] L. van der Maaten and E. Hendriks, "Action unit classification using active appearance models and conditional random fields," *Cognit. Process.*, vol. 13, no. 2, pp. 507–518, 2012.

PENGYU CHEN received the double bachelor's degree in engineering and management from Harbin Engineering University, in 2017. He is currently pursuing a degree in biomedical engineering with Capital Medical University. His research interests include natural language processing and knowledge graph in medical imaging reports.

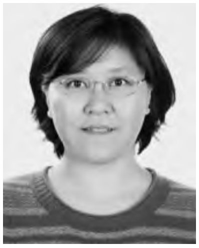
QIAO LIU received the B.E. degree from the ShenYuan Honors College, Beihang University, Beijing, China, in 2016. He is currently pursuing the Ph.D. degree in automation with Tsinghua University, Beijing. He has been with the MOE Key Laboratory of Bioinformatics, since 2016. His research interests include applied machine learning, and biomedical and bioinformatics.

LAN WEI is currently a Senior Engineer with Xuanwu Hospital, Capital Medical University. Her research interests include statistical reports and data analysis of the hospital, and the construction of hospital information platform and data platform.

BEIER ZHAO received the M.B.A. degree in computer science from the Beijing University of Posts and Telecommunications. He is currently the CEO of Beijing INFI-SAGACITY Technology Company, Ltd. His research interests include big data product design and development, and big data and artificial intelligence product technology architecture and distributed computing.

YIN JIA is the Medical Business Operator with Beijing INFI-SAGACITY Technology Company, Ltd. Her research interests include the collection of medical field experts' experience and opinions, combined with real medical needs, designing business goals, and integrating Internet big data to form medical solutions.

HAIRONG LV received the Ph.D. degree from Tsinghua University, in 2007, where he is currently an Associate Researcher with the Department of Automation. His research interests include artificial intelligence and blockchain.



XIAOLU FEI received the B.S. and M.S. degrees in biomedical engineering from Tsinghua University, and the Medical Doctor degree in medical imaging and nuclear medicine from Capital Medical University. She serves as the Director of the Information Technology Department, Xuanwu Hospital. Her research interests include clinical engineering, especially for introducing the strategies of Information Governance and Data Governance into healthcare environment, improving the quality and accessibility of healthcare data, and investigating to perform value-based assessment using data mining methods to evaluate and analyze the safety and outcome of technologies and productions adopted under the highly complex digital medical environment.