Structured information extraction for bone scan image report texts

Mengge Huang1,2, Jingyi Zhang1,2, Tao Deng1,2,3,*, Zhengxing Man1,2,3 and Qiang Lin1,2,3

1 Key Laboratory of China’s Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, 730030, P. R. China
2 Key Laboratory of Streaming Data Computing Technologies and Application, Northwest Minzu University, Lanzhou, 730030, P. R. China
3 School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou, 730030, P. R. China
*Corresponding author’s e-mail: dttom@lzu.edu.cn

Abstract. Extracting structured information from the bone scan image report text plays a crucial role in supporting clinical analysis and research. This study summarized the structure and characteristics of 3608 bone scan image report text using dictionary-based information extraction method, including data cleaning, entity recognition, building dictionary and extraction rules. This method was used to obtain the structured data of bone scan image report text required for clinical research, and the effect evaluation was carried out on 1000 randomly selected report texts, with the precision rate and recall rate higher than 90%. The method proposed in this study is practical and could have good effect on structured results for bone scan imaging report text.

1. Introduction
With the rapid development of medical informatization construction in China, massive medical data resources are generated by medical institutions every day. However, a large number of medical imaging report texts are still unstructured text data at present. Due to the huge gap between natural language and computer language, using computer processing or manual processing methods directly for unstructured text data is not only time-consuming, inefficient, but also difficult to ensure the quality of results. Therefore, how to extract valuable information from unstructured text data by using existing techniques and tools is becoming one of the hot spots in current research.

In recent years, there have been many researches on the structural processing of medical text data at home and abroad. Abroad, Hamon et al.[1] designed a system for analysing descriptive clinical documents to extract medication status and information related to medication. Jiang et al.[2] developed a new hybrid clinical entity extraction system, which integrates the module based on heuristic rules with the named entity recognition module based on ML. Savova et al.[3] constructed clinical text analysis and knowledge extraction systems, consisting of components executed in sequence to process clinical narratives.

Domestically, Yang et al.[4] built a VSM model to cluster medical image reports and extract key indicators to generate a medical vocabulary. The semantic dependency in the short sentence was analyzed and constructed. The corresponding dependency syntax tree is used to extract the key indicators and corresponding indicator values in the medical image report. Chen et al.[5] proposed a...
three-layer structure of short sentences. By constructing a calculation method of sample name similarity and index name similarity, the text clustering algorithm based on dictionary was proposed. Tian et al.[6] used the word2vec model to train word vectors, then used the cosine similarity to find synonyms to remove the polysemous phenomenon of a word. A tailoring strategy is proposed for the complex dependency relationship tree to extract the pathological index information in the report and the corresponding parameter values according to the characters of part of speech.

Aiming at the specific characteristics of the text data of medical imaging reports, this study proposes a simple and effective method of structured text processing for medical imaging reports. The specific process is as follows: 1) Perform data preprocessing operations; 2) Design the correspondence between text and label, and then manually label the data for model training based on the data features required to be extracted; 3) Obtain an entity dictionary by the training model, and the medical image report text is structured based on the dictionary.

2. System framework

2.1. Structured flow chart

In this study, a structured processing method for bone scan image report text is proposed, and its processing flow is shown in Figure 1.

![Figure 1. Structured processing flow.](image-url)
The original abbreviated medical terms are converted into complete medical terms through data cleaning. After the stage of entity recognition, the entity vocabulary can be obtained. In the end, the information extraction corresponds bone scan image report text to structured extraction rules, generating the structured output.

2.2. Problem description

$^{99m}$Tc-MDP Whole body bone scan is a method to detect bone metastases through the selective deposition of radioisotope tracer on the lesion[7]. The bone scan image report texts record the lesion location, the attribute characteristics of the lesion site and the objective description of the disease.

Generally, bone scan image report text, apart from the basic information of the patient and the report, is mainly divided into "examination description" and "diagnosis suggestions" fields, among which the description of lesion features is mainly focused on the "examination description" field.

3. Structured approach

3.1. Data preprocessing

Since some medical terms are abbreviated to facilitate the generation of the inspection report, the task of data cleaning is to standardize the data and improve the data utilization rate, mainly including completing the abbreviated medical terms and correcting the redundant spelling. Parts of the data cleaning scheme are shown in Table 1.

| category       | Match target                              | Replace content                                                                 |
|----------------|-------------------------------------------|---------------------------------------------------------------------------------|
| medical term   | L4 and 5 vertebral bodies                 | L4 vertebral body, L5 vertebral body                                           |
| Medical term   | Left side| right side| bilateral second and third front rib| rear rib| side rib | Left side| right side| bilateral second front rib| rear rib| side rib, left side| right side| bilateral third front rib| rear rib| side rib |
| Spelling       | Bilateral wrist joints joints             | Bilateral wrist joints                                                          |

3.2. Entity recognition

3.2.1. Sequence annotation. Since the bone scan image report text contains various parts of the human body and diseases, common Chinese word segmentation tools (such as jieba[8], IK Analyzer[9], FNLP[10] of Fudan University, NLPIR[11] of Beijing Institute of Technology, etc.) could not attain ideal word segmentation, so it is necessary to manually label the data for entity recognition[12-14].

The simplest way to solve the problem of joint annotation is to transform it into the original annotation problem. This study utilizes the BIO sequence annotation strategy: Labelling each element as "B-X", "I-X" or "O". "B-X" denotes that the fragment in which this element resides is of type X and that this element is at the beginning of the fragment, "I-X" means that the fragment in which this element resides is of type X and that this element is in the middle of the fragment, and "O" means that it is not of type.

According to the characteristics of the text data, the entity words of the bone scan imaging report text are divided into the following categories: Location entity words (Loc), Shape entity words (Shape), Degree entity words (Deg), Status entity words (Sta), and Disease entity words (Dis). Combining with the BIO sequence labelling strategy, in the specific labelling process, the labelling method shown in Table 2 is selected to distinguish.
Table 2. BIO tag categories.

| Entity tag | Start tag | Middle and end tags |
|------------|-----------|---------------------|
| Location   | B-Loc    | I-Loc               |
| Shape      | B-Shape  | I-Shape             |
| Degree     | B-Deg    | I-Deg               |
| Status     | B-Sta    | I-Sta               |
| Disease    | B-Dis    | I-Dis               |
| Nonentity tag | O     | O                   |

3.2.2. BiLSTM-CRF. Bilstm-crf is a popular sequence labelling algorithm at present. It combines BiLSTM and CRF[15], so that the model can not only consider the correlation between the sequence before and after like CRF, but also possess the feature extraction and fitting ability of LSTM[16]. The BiLSTM-CRF network structure in this study is shown in Figure 2.

Take the sentence in the bone scan image report text as the unit, and record a sentence (word sequence) containing n words as:

\[
(x_1, x_2, ..., x_n)
\]

Where \(x_i\) represents the id of the \(i\)-th word of the sentence in the dictionary, and then the one-hot vector of each word can be obtained, and the dimension is the dictionary size.

The first layer of the model is the look-up layer, which uses a pre-trained or randomly initialized embedding matrix to map each word \(x_i\) in the sentence from a one-hot vector to a low-dimensional dense word vector \(x_i \in \mathbb{R}^d\), \(d\) is the dimension of embedding. Before entering the next layer, set dropout to alleviate overfitting. The second layer of the model is the bidirectional LSTM layer, which automatically extracts sentence features. The character embedding sequence \((x_1, x_2, ..., x_n)\) of each word of a sentence is taken as the input of each time step of the bidirectional LSTM, and embedding the hidden state sequence \((h^+, h^-, ..., h^+_n)\) of forward LSTM output and reverse LSTM \((h^-, h^+, ..., h^-_n)\) output at various positions \(h_i = [h^+_i; h^-_i] \in \mathbb{R}^m\) to obtain a complete hidden state sequence \((h^+, h^-, ..., h^-_n) \in \mathbb{R}^{m\times m}\). After setting dropout, access a linear layer to map the hidden state vector from \(m\)-dimension to \(k\)-dimension, where \(k\) is the number of labels in the annotation set, so as to obtain the sentence features extracted automatically, which are denoted as matrix \(P=(p_1, p_2, ..., p_k) \in \mathbb{R}^{m\times k}\). Each dimension \(p_i\) of \(P \in \mathbb{R}^k\) is treated as a score that classifies the word \(x_i\) into the \(j\)-th label. The third layer of the model is the CRF layer, which carries out sentence level sequence annotation. The parameter of the CRF layer is a \((k+2)\times(k+2)\) matrix \(A\), \(A_j\) refers to the transfer score from the \(i\)-th tag to the \(j\)-th tag, and then the label that has been marked before can be used to mark a position. The reason for adding 2 is to add a start state for the first sentence and a stop state for the end of the sentence. If a sentence length is equal to the length of the label sequence \(y\), then the model scores the tag of sentence \(x\) equal to \(y\), the score of the entire sequence is equal to the sum of the scores of each position, and the score of each position is obtained from two parts, one part is determined by the \(p_i\) output by the LSTM, and the other part is determined by the transition matrix \(A\) of the CRF, and then the normalized probability

\[
p(y \mid x) = \frac{\exp(score(x, y))}{\sum_y \exp(score(x, y'))}
\]

can be obtained by using Softmax.
3.3. Information extraction based on dictionary

Through entity recognition, five types of attribute thesaurus of location, shape, degree, state and disease can be obtained, and keywords of these five types of attributes can be formed into a keyword list. According to certain rules, information extraction can be completed and structured data is finally generated. The specific algorithm is shown in Table 3.

The given text is denoted as Sentence, the keyword list is denoted as \( K = [K_1, K_2, ..., K_n] \), the start marker is denoted as BeginMark, the end identifier is denoted as EndMark, and the keyword type list Type=[] . Define the lists and dictionaries that need to be used: \( a = [], b = [], Tem = [], Results = \{\} \) and Finalresults=\{\}.

Table 3. Structured mapping rules.

| Algorithm 1: Structured mapping pseudo code |
|--------------------------------------------|
| 1: for i in K                              |
| 2:   if i in Sentence                      |
| 3:     for j in getIndex(str,i)            |
| 4:       item=(j,i,getType(i))             |
| 5:     Add item to a                       |
| 6:     for i in a                          |
| 7:      Add i[0]and i[2] to Tem            |
| 8:     while(True):                        |
| 9:       start=Tem.index(BeginMark)        |
| 10:     end=Tem[start:].index(EndMark)+start+1 |
| 11:     if start==--1 or end=start         |
| 12:     break                              |
| 13:     Add a[start:end] to b              |
| 14:     del Tem[:end]                      |
| 15:     a=a[end:]                          |
| 16:     return b                           |
17: for i in b
18: for tags in Type
19: if b[i][2]==tags
20: Add b[i][1] to tags
21: Add tags to Results
22: if DResults.keys() in LResults.keys()
23: return DResults.values() and LResults.values()

4. Experiment

4.1. Experimental data

The sample data for this experiment is the bone scan image report text, which comes from the Department of Nuclear Medicine, Gansu Provincial People's Hospital. Extract and clean the data from the data set. There are 3608 records in total, including six basic information fields: "ID", "age", "source", "clinical diagnosis", "injected drugs", "dose", and two main text fields: "Examination description" and "Diagnosis suggestions". In addition to the basic information of the patient and the basic information fields of the report, the data includes three fields: "ID", "Examination description" and "Diagnosis suggestions".

4.2. Results display

Take the data in Table 1 as an example. The " the fourth anterior rib on the right", " left ischium " and "bilateral knee joints" included in the report belong to the Location attribute; "Point-like" and "Flake" are Shape attribute; "Slight" belongs to the Degree attribute; "Concentration" belongs to the status attribute; "Bone metastasis" and "arthritis" belong to the Disease attribute. Therefore, according to entity recognition and algorithm 1, the structured results of the bone scan image report text in Table I can be obtained, as shown in Table 4.

| Examination description | Diagnosis suggestions | Structured results |
|--------------------------|-----------------------|-------------------|
| After intravenous injection of $^{99m}$TC-MDP for 3 hours, the anterior and posterior positions of whole-body bone imaging were performed: All the bones of the whole body were developed, the fourth anterior rib on the right and the left ischium showed slight concentration of spot-like radioactivity, and the bilateral knee joints showed sheet-like radioactivity concentration; the rest of the bone tissue showed no obvious abnormal concentration and defect areas. The both kidneys were developed, and the morphology was not abnormal. | 1. bone metastases of the fourth anterior rib on the right and left ischial bone; 2. Bilateral knee arthritis; | bone metastases:<the fourth anterior rib on the right, spot-like, slight, concentration > <left ischial bone, spot-like, slight, concentration > Arthritis: <bilateral knee joints, sheet-like, concentration > |

4.3. Analysis of results

In the entity recognition part, divide the data set 7:3, randomly select 2525 of them as the training data set, and the remaining 1083 as the test data set. The word vector dimension used by the BiLSTM-CRF model is set to 100, epoch is set to 100, Batch_size is set to 16, dropout is set to 0.5, and learning rate is set to 0.001. In order to evaluate the model, precision rate($P$), recall rate($R$)[17-18] and F1 value($F1$) are used as evaluation indexes in this study. The calculation formula of the evaluation index is as follows:
The BILSTM-CRF model was replaced with HMM model, CRF model and BiLSTM model, respectively, and implemented and applied to the experimental data set under the same experimental environment. The results are shown in Figure 3.

It can be seen from Figure 3 that the precision rate, recall rate and F1 value obtained by using the BiLSTM-CRF model for the bone scan image report text are better than the HMM model, CRF model and BiLSTM model. Compared with the BiLSTM model, the BiLSTM-CRF model adds a CRF layer to it, so that the model can consider the correlation between the entity category tags in the bone scan image report text.

In order to facilitate the evaluation of the effect of the structured method, 3608 bone scan image report texts were randomly selected 3 times, each time 1000 pieces of data were extracted. Algorithm 1 is applied to the extracted data, which uses precision rate \( \text{pre} \) and recall rate \( \text{rec} \) as evaluation indicators for the structured information extraction results. The calculation formula for the evaluation indicators is as follows:

\[
\text{pre} = \frac{l}{w} \quad (4)
\]
\[
\text{rec} = \frac{l}{t} \quad (5)
\]

Where, \( l \) represents the number of correct information bars extracted, \( w \) represents the number of information bars extracted in the sample, and \( t \) represents the number of information bars extracted.

According to the evaluation index: 1) The precision rate and recall rate of the structured results obtained through 3 random extractions are higher than 90%; 2) When the number of attribute words in the bone scan image report text is small, the precision rate and recall rate of extraction are relatively higher.

5. Conclusion
This study proposes a structured processing method for bone scan image report text, which uses the BiLSTM-CRF model to manually mark the text data, designs and trains the named body recognition model, and then structures the unstructured text data based on the dictionary. Experiments prove that the method proposed in this study has a good structured processing effect, and the precision rate of the processing result is 93.26%, which provides a certain reference for converting other image report text.
data into structured data. However, this method is restricted in an extraction rule formulated for a specific corpus. A better extraction rule will be developed in future, and applied to more different corpora.

Acknowledgments
This research was funded by the National Natural Science Foundation of China [No. 32060234], the Gankeji ([2020] 11) Natural Science Foundation (20JR5RA511).

References
[1] Hamon, T., Grabar, N. (2010) Linguistic approach for identification of medication names and related information in clinical narratives. J. Am. Med. Inform. Assoc., 17: 549–554.
[2] Jiang, M., Chen, Y., Liu, M., et al. (2011) A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J. Am. Med. Inform. Assoc., 18: 601–606.
[3] Savova, G. K., Masanz, J. J., Ogren, P. V., et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inform. Assoc., 17: 507–606.
[4] Yang, B., Nie, T., Shen, D., et al. (2019) Approach of Structured Information Extraction for Medical Text Data. Journal of Chinese Mini-Micro Computer Systems, 40: 1479–1485.
[5] Chen, D., Feng, J., Le, J., et al. (2016) Research on Structured Method for Chinese Pathological Text. Computer Science, 43: 272–276.
[6] Tian, C., Chen, D., Wang, M., et al. (2016) Structured Processing for Pathological Reports Based on Dependency Parsing. Journal of Computer Research and Development, 53: 2669–2680.
[7] Liu, Y., Zhang, J. (2011) The Diagnostic Significance of the Whole Body Bone Scanning in Bone Metastases in Lung Cancer. Journal of Basic and Clinical Oncology, 24: 325–326.
[8] Li, Y., Guo, F., Zhai, X., et al. (2020) Study on Doctors’ Portraits of Online Medical Website Based on Jieba Chinese Word Segmentation. Journal of Medical Informatics, 41: 14–18.
[9] Zheng, C., Sun, W., Deng, C., et al. (2016) Research of NPC proposals similarity based on IKAnalyzer and VSM. Information & Communications, 48–50.
[10] Wang, Y., Chen, T. T. (2019) An FNLP approach for planning energy-efficient manufacturing: wafer fabrication as an example. Procedia Manufacturing, 38: 439–446.
[11] Zhou, L., Zhang, D. (2003) NLPIR: a theoretical framework for applying natural language processing to information retrieval. Journal of the American Society for Information Science & Technology, 54: 115–123.
[12] Gong, C., Li, Z., Xia, Q., et al. (2020) Hierarchical LSTM with char-subword-word tree-structure representation for Chinese named entity recognition. Sci. China Inform. Sci., 63: 74–88.
[13] Zhang, D., Chen, W. (2020) Chinese Named Entity Recognition Based on Contextualized Char Embeddings. Computer Science.
[14] Liu, Z., Luo, M., Yang, H., et al. (2020) Named entity recognition for the horticultural domain. J. Phys.: Conf. Ser., 1631.
[15] Zong, C. (2013) Statistical natural language processing. Tsinghua University Press, Beijing.
[16] Wu, Y., Min, J., Lei, J., et al. (2015) Named entity recognition in chinese clinical text using deep neural network. Studies in Health Technology and Informatics, 216: 624–628.
[17] Xiang, C., Xiong, S. (2014) An extraction algorithm of web fragment information based on feature tree. Journal Of Lanzhou University Of Technology, 40: 104–107.
[18] Wu, L., Chen, X., Tang, J. (2011) Content Extraction Algorithm of HTML Pages Based on Optimized Weight. Journal of South China University of Technology (Natural Science Edition), 39: 32–37.