Classification and extraction of medical clinical trial screening standard texts based on Bi-LSTM and Attention mechanism

Lin Du¹, Dong Cao¹ and Jinghua Li², Hui Ye¹, *
¹School of medical information engineering, Guangzhou University of Chinese medicine, Guangzhou city, Guangdong province, China
²Institute of Information on Traditional Chinese Medicine (IITCM) China Academy of Chinese Medical Sciences, Beijing, China
*Corresponding author: wilsonyip2014@gzucm.edu.cn

Abstract. The medical recruitment of human subjects in clinical trials often requires manual comparison of the subject's medical records and clinical trial screening criteria, which is a time-consuming and labor-consuming method to determine candidates. In order to solve this problem, it has a great clinical value to study a method of classifying clinical trial data screening standards automatically. This paper attempts to propose a medical short text classification model of clinical medicine based on BiL-Att (Bi-LSTM + Attention) model. It uses Word2vec preprocessing to get short text vector as the model input. The results showed that the classification effect of BiL-Att model reached the highest Average F1 value of 80.26%.

1. Introduction
Clinical trial refers to the scientific research on the subjects, and the screening criteria are the main indicators for the clinical trial director to determine whether the subjects meet a certain clinical trial. Generally, the screening criteria are unstructured data in the form of irregular free text. In clinical trials, the recruitment of subjects needs to be completed by comparing the medical record form with the clinical trial screening standard manually. This method is time-consuming, laborious and inefficient. There are many problems such as long recruitment time, loss of subjects and difficulty in recruitment. Therefore, it is of great medical and clinical value to study a method of classifying clinical trial data according to screening criteria.

And this paper attempts to extract and classify the short text in the medical record table, and return the specific category of each screening standard through machine learning and natural language processing technology, with the given 44 screening standard categories and a series of description sentences of Chinese clinical trial screening standards. The example is as follows:
TABLE I. Example of filter standard classification.

| ID | Input (filter criteria)         | Output (category)      |
|----|---------------------------------|------------------------|
| S1 | Age > 80 years old              | Age                    |
| S2 | history of alcohol and drug abuse | Addictive Behavior    |
| S3 | Blood sugar <2.7mmol/L          | Laboratory Examinations|

Liao Xiaoqin used the BSP-CNN hybrid neural network model to study the classification of emotional tendencies in short text reviews of consumption, which effectively simplified the operation and achieved higher F1 value compared with the LSTM model and CNN model [1]. When Huang Wei studies the entity recognition model of terrorism related information, he uses Bi-LSTM network to predict and label initially, and then uses CRF layer to add constraints. His experiments show that the recognition accuracy and recall rate of Bi-LSTM-CRF model for terrorism related information are 90% [2]. Sun Chengai put forward ABGC model and added Attention mechanism to Bi-LSTM when he studied emotional tendentiousness of neural network, which avoided the gradient disappearance of LSTM, solved the problem of LSTM neglecting context implication, and effectively improved the accuracy of text classification [3]. Zhang et al. found that the accuracy of Bi-LSTM model was improved after using Attention mechanism when they studied emotion analysis model [4]. Therefore, the author attempts to use Bi-LSTM + Attention model (hereinafter referred to as BiL-Att) to classify medical short texts.

2. Research method

2.1. Training word vector with Word2vec

Word2vec is a tool for word vector computation and the first step in the study of natural language processing. There are two models available, CBOW and Skip-gram. Word2vec can transform natural language into dense vector that can be understood by computer as the input of the model. However, since the basic unit of natural language processing is word, word vector is generated by Word2vec. In this paper, it is necessary to obtain word vector for short text classification, so the author chooses to segment sentences in short text first, and calculate the average vector of word vector of all words in sentences to represent the sentence vector of the sentence. Similarly, according to the average vector of sentence vector, represents the text vector of the modified text.

![Fig 1. Word2vec getting short text vector.](image)
2.2. **Bi-LSTM mechanism of bidirectional long short memory network**

Bi-LSTM is a combination of forward LSTM and backward LSTM. It is trained from the front and back, and output from the same layer, so it has the ability to remember the past and future information. And it overcomes the difficulty that the traditional LSTM model cannot capture the context information due to the serialization problem, and can improve the accuracy of classification in theory.

![Fig 2. Structure of Bi-LSTM model.](image)

In Bi-LSTM, the coding process of a sentence "I love Apple" can be represented as the following figure:

![Fig 3. Bi-LSTM coding process.](image)
Forward LSTML input "I", "love" and "Apple" successively to get three vectors \{h_{L0}, h_{L1} and h_{L2}\}. Backward LSTMR input "I", "love" and "Apple" successively to get three vectors \{h_{R0}, h_{R1} and h_{R2}\}. Finally, the forward and backward implicit vectors are spliced to get \{[h_{L0}, h_{R2}], [h_{L1}, h_{R1}], [h_{L2}, h_{R0}]\}, namely \{h_0, h_1, h_2\}.

2.3. Attention mechanism

The Attention mechanism has shown great success in recent years in a variety of tasks from machine translation. Attention function can be described as the key-value pair set of

Mapping query and output, as shown in Figure 3 below. The main three steps to calculate the Attention value are as follows:
1) calculate the similarity of sequence and key value to get the weight;
2) normalize the weight calculated by using Softmax function;
3) Weighted sum of weights and corresponding keys to calculate the final Attention;

![Fig 4. Attention function structure diagram.](image)

The third step is that weighted sum of weights and corresponding keys to calculate the final Attention model. And the calculation formula used is as follows:

$$ a_t = \sum_{i=0}^{N} \alpha_i^t h_i \in R^h $$

(1)

Due to the gradient disappearance of LSTM and the neglect of context implication in Bi-LSTM, in order to solve this problem, the Attention mechanism was added in this paper to improve the classification accuracy by distinguishing the importance of different features, ignoring the unimportant features and paying Attention to the important features. The Attention model can be divided into the following three steps to solve the problems existing in Bi-LSTM:

a) Retain the intermediate output result of Bi-LSTM encoder to the input sequence;
b) Training a model of selective learning, taking the results in a) as input;
c) Associate the output sequence with b) model when Attention is output;
2.4. BiL-Att fusion model mechanism

The Bi-LSTM and Attention fusion model is to add Attention layer to the Bi-LSTM model, use the last time sequence output vector as the feature vector in Bi-LSTM, and select Softmax function for classification. Attention model first calculates the weight of each time sequence, then weights all the time sequence vectors and uses them as feature vectors, and then selects Softmax function for classification [6].

![Fig 5. BiL-Att fusion model structure diagram.](image)

3. Experimental steps

3.1. Word2vec to get the text vector

In this paper, Word2vec is used for data preprocessing to transform natural language words into vectors that can be recognized by the computer. The specific methods are described as follows: Each line of the document is extracted into a sentence, word segmentation is carried out for each sentence to obtain the word vector of each word, and the sentence vector of this sentence is calculated by weighting the word vector. Similarly, you can calculate the entire document text vector. So a normal natural language document is processed by Word2vec to generate a 100-dimensional vector that is the input to the model.

For example, a Disability class TXT document in the Training set is selected. The unprocessed documents are as follows:

Good sleep quality, no stay up in nearly 2 weeks;
A pattern of day and night reversal, or irregular sleep patterns.
Sleep latency, or more than 30 minutes before bed, is less than 85%

![Fig 6. Disability class documentation.](image)

The pre-processed vector of this document is:

```
[array([0.6541567,1.1093941,1.2360079,0.06075715,1.658712,-
0.22067167,0.30755034,1.0505137,-
0.17929547,1.464987],dtype=float32)......
```

![Fig 7. Disability class text vector.](image)
3.2. BiL-Att fusion model achieves medical text classification

BiL-Att fusion model means to add the Attention layer on the Bi-LSTM model, use the last time sequence output vector as the feature vector in Bi-LSTM [7], and select Softmax function for classification. The function of Attention layer is to calculate the weight of each time sequence first, then weight and sum all the time sequence vectors as feature vectors, and then select Softmax function for classification. BiL-Att model construction specific steps can be described as follows:

4. Result analysis

The data set adopted in this paper has a total of 44 categories, respectively expressed as C1..., Ci... C44. Each category has its own calculation accuracy, recall rate and F1 value. The evaluation indexes adopted in this paper are macro accuracy P value, macro recall rate R value and AverageF1 value. The calculation formula of accuracy and recall rate of each class is expressed as:

\[
\text{Accuracy } P_i = \frac{\text{the number of samples correctly predicted as category } C_i}{\text{the number of samples predicted as category } C_i}
\]

\[
\text{Recall rate } R_i = \frac{\text{number of samples correctly predicted as category } C_i}{\text{number of samples of real } C_i \text{ class}}
\]

The calculation expression of the macro accuracy P value is as follows:

\[
P = \left( \frac{1}{n} \right) \sum_{i=1}^{n} P_i
\]  \hspace{1cm} (2)

The calculation expression of R value of macro recall rate is as follows:
The calculation expression of AverageF1 value is:

\[ \text{Average} \cdot \text{F1} = \left( \frac{1}{n} \right) \sum_{i=1}^{n} \frac{2 \cdot P_i \cdot R_i}{P_i + R_i} \]  

TABLE II. Comparison of evaluation indexes of the three models.

| Model   | P     | R     | AverageF1 |
|---------|-------|-------|-----------|
| CNN     | 0.817160 | 0.599294 | 0.691472 |
| LSTM    | 0.823371 | 0.639107 | 0.719631 |
| BiL-Att | 0.879790 | 0.737877 | 0.802609 |

In the case of the same Word2vec data processing, there were significant differences in the training results under three different models. It was found in this paper that the Average F1 value of the CNN model was 69.14%, the Average F1 value of the LSTM model was 71.96%, and the Average F1 value of the BiL-Att model was 80.26% at most. The results indicated that BiL-Att fusion model was more effective in the classification of medical texts.

5. Conclusion
In this paper, by constructing sentence Word2vec word vector to calculate the text word vector, as the input of the model, medical text classification was carried out, trying to compare CNN and LSTM, two benchmark models, and found that BiL-Att model had the best classification effect. Attention mechanism was integrated into BiL-Att model. Attention was weighted and weighted by time sequence vectors as feature vectors, which overcame the problem of gradient disappearance and the neglect of context meaning in Bi-LSTM model. BiL-Att model had certain improvement in effect compared with traditional CNN and LSTM models. It is of great clinical value to be able to return the specific category of each screening criterion to provide research basis for automatic analysis of screening criteria in clinical trials. Future work can be considered to generate advances in data preprocessing and the imbalance of training set data.

Acknowledgment
This study was partially supported by 2017 National Key Research and Development Program. Cloud Computing and Big Data" Special Program (2017YFB1002300); experiment data is from China Health Information Processing Conference 2019. Corresponding Author: Hui Ye.

References
[1] Liao Xiaojin, Xu Yang. Classification study on emotional tendency of short text for BSP-CNN [J]. Computer engineering and application, 2019.
[2] Huang Wei. Study on entity identification model of terrorism-related information based on Bi-LSM-CRF [J]. School of economics and management, Hubei University of Technology, 2019.
[3] Sun Chengai. Analysis of emotional tendency of neural network based on GLU-CNN and Attention-Bi-LSTM [J]. Shandong University of Science and Technology, 2019.
[4] Zhang Liang, Yang Yuanfeng, Li Jinxiang, Jin Yi, Liu Yuanxia. Design and implementation of Bi-LSTM+Attention emotion analysis model [J/OL]. Electronic Technology and Software Engineering 177-179 [2019-10-15].
[5] Ullah A , Ahmad J , Muhammad K , et al. Action Recognition in Video Sequences using Deep Bi-directional LSTM with CNN Features[J]. IEEE Access, 2017, PP (99):1-1.
[6] Hu Tingting, Feng Yaqin, Shen Lingjie, Wang Wei. Selection of main features of LSTM speech emotion based on Attention mechanism [J]. Acoustics, 2019, 38 (04):414-421.
[7] Liu feng, gao si, yu bihui, guo fangda. Classification of entity relationship based on multi-head Attention and Bi-LSTM [J]. Computer system application, 2019, 28 (06):118-124.