A Traditional Chinese Medicine Text Relation Extraction Model Based on Bidirectional LSTM and GBDT: Postprint

Authors: Luo Jigen, Du Jianqiang, Nie Bin, Xiong Wangping, Liu Lei, He Jia

Date: 2018-10-11T00:00:00+00:00

Abstract

To address the problems of insufficient generalization capability of entity relationship recognition models when using Softmax as the classifier for Long Short-Term Memory networks, and their inadequate applicability to entity relationship extraction in Traditional Chinese Medicine, this paper proposes a relationship recognition algorithm that fuses Gradient Boosting Decision Trees with Bidirectional Long Short-Term Memory networks (BILSTM-GBDT). The approach first employs word2vec to perform vectorized representation of Traditional Chinese Medicine texts, then utilizes an attention mechanism-based Bidirectional Long Short-Term Memory network to extract high-order features, and finally adopts the ensemble classification model Gradient Boosting Decision Tree as the feature classifier to enhance relationship recognition effectiveness. Experimental results on multiple relationship corpora including Traditional Chinese Medicine demonstrate that the proposed model achieves higher precision, recall, and F-value compared with traditional SVM methods, GBDT methods, and deep learning methods.

Full Text

Preamble

Title: TCM Text Relationship Extraction Model Based on Bidirectional LSTM and GBDT

Authors: Luo Jigen, Du Jianqiang†, Nie Bin, Xiong Wangping, Liu Lei, He Jia
(School of Computer Science, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, China)

Abstract: To address the problem that using Softmax as a classifier for long short-term memory networks leads to insufficient generalization ability of entity relationship recognition models, making them poorly suited for Traditional Chinese Medicine (TCM) entity relationship extraction, this paper proposes a relationship identification algorithm (BILSTM-GBDT) that integrates a gradient boosting decision tree with a bidirectional long short-term memory network. First, word2vec is employed to vectorize TCM texts, then a bidirectional LSTM network based on an attention mechanism is used to extract high-order features, and finally, the ensemble classification model gradient boosting decision tree is adopted as the feature classifier to improve relationship recognition performance. Experimental results on multiple relational corpora including TCM demonstrate that the proposed model achieves higher precision, recall, and F-score compared with traditional SVM methods, GBDT methods, and deep learning methods.

Keywords: relationship extraction; LSTM; gradient boosting decision tree; attention mechanism; TCM text

0 Introduction

TCM diagnostic knowledge represents a treasure inherited from Chinese civilization over thousands of years and provides strong guidance for clinical TCM practice. With the continuous increase in TCM diagnostic data, the flexibility of sentence expression forms has grown, and entity relationships have become increasingly complex. TCM entity relationship recognition [1,2] constitutes part of information extraction in the TCM domain [3], referring to the identification of semantic relationships between two given entities within unstructured text. For example, the following sentence contains multiple entity types including prescriptions, herbs, symptoms, tongue images, pulse patterns, and syndrome types: “Bu Tian Da Zao Wan is composed of multiple Chinese herbs including Achyranthes bidentata, Angelica sinensis, and Foeniculum vulgare. Its main indications include cough, dyspnea, shortness of breath, expectoration of white frothy sputum, dull blood color, tidal fever, spontaneous sweating, night sweats, hoarseness or loss of voice, facial and limb edema, palpitations, purple lips, cold limbs, cold sensation, and mouth and tongue ulcers. The tongue coating is yellow and peeling, the tongue body is pale and glossy, and the pulse is fine, minute and rapid, indicating lung consumption with dual deficiency of yin and yang syndrome.”

In relationship extraction tasks, it is necessary to accurately identify semantic relationships such as between the entities “Bu Tian Da Zao Wan” and “Achyranthes bidentata”, “Angelica sinensis” and “Foeniculum vulgare”. The broad relationship category is “prescription-herb”, with the specific relationship being “composition”. The entity “Bu Tian Da Zao Wan” has a “treatment” relationship with “lung consumption with dual deficiency of yin and yang syndrome”, while “lung consumption with dual deficiency of yin and yang syndrome” has

a “pulse manifestation” relationship with “fine, minute and rapid pulse”, indicating that a fine and rapid pulse is the pulse pattern of lung consumption with dual deficiency of yin and yang. Additional relationships include syndrome-symptom, syndrome-tongue image, etc. The relationship representation for the entire sentence is shown in Figure 1 [Figure 1: see original paper].

1 Related Work

Relationship extraction holds significant research importance for information retrieval, discourse comprehension, and knowledge graph construction. Currently popular relationship extraction methods include feature engineering-based approaches, kernel function-based approaches, and deep learning-based approaches [4].

Feature engineering-based methods primarily utilize lexical features, syntactic features, and semantic features [5]. Although these methods have achieved decent results to some extent, the increasingly complex sentence expression forms make feature extraction progressively difficult, limiting the improvement of relationship extraction performance. Kernel function-based methods [6] differ from feature engineering by focusing on the structural information of sentences themselves without constructing high-dimensional data feature vectors. These methods use syntactic parse trees as input objects and perform relationship classification by comparing structural similarities between corpora through kernel functions. However, due to noise in hidden sentence features that cannot be recognized by humans, as well as different expressions with identical semantics and varying expressive capabilities of sentences of different lengths, kernel function-based relationship extraction also suffers from certain drawbacks.

With the continuous development of deep learning, its advantage in automatic feature extraction has led to wider application in relationship extraction tasks [7,8]. Vu et al. [9] proposed applying deep recurrent neural networks (DRNN) to relationship extraction by dividing sentences into two parts through parse trees and then feeding them into multi-layer recurrent neural networks. Zeng et al. [10] proposed a relationship extraction algorithm using convolutional neural networks (CNN) that incorporates position information. To effectively alleviate long-distance dependency issues, this algorithm considers N-gram features, but the limited size of filters in CNN prevents complete resolution of long-distance dependency problems. LSTM, an improved RNN model proposed by Hochreiter et al. [11], employs three gating mechanisms through memory and forget operations to address long-distance dependency issues present in RNN and CNN models. It has been increasingly applied to relationship extraction tasks, with Miwa et al. [12] utilizing LSTM with SPTree for relationship extraction. However, all these models adopt Softmax as the classifier, leading to insufficient generalization ability of entity relationship recognition models [13] and poor adaptability to TCM entity relationship classification problems.

To address these issues, this paper proposes a bidirectional long short-term memory (BILSTM) model integrated with the gradient boosting decision tree (GBDT) algorithm [14]. While using BILSTM for feature extraction, an attention mechanism is incorporated to capture keywords for sentence comprehension [7], solving the problem of interference from irrelevant words. Following feature extraction, GBDT is employed for relationship classification training and prediction. Due to the low variance and high bias advantages of GBDT's base models, the ensemble model achieves greater stability and can partially solve the generalization issues caused by using Softmax as an LSTM classifier.

---

2 BILSTM-GBDT Model for Relation Extraction

The BILSTM-GBDT model for relation extraction employs BILSTM to obtain deep implicit features from both forward and backward directions while effectively solving long-distance dependency problems inherent in traditional deep learning methods. Additionally, when utilizing BILSTM for feature extraction, an attention mechanism is incorporated to consider the impact of keywords on features, thereby obtaining more contextual information. The GBDT algorithm then classifies the extracted features to obtain final relationship categories. The model architecture is shown in Figure 2 [Figure 2: see original paper] and consists of two main components:

a) **Attention-based BILSTM feature extraction**: Word vectors from the training corpus are input into the BILSTM model, which uses an attention mechanism to calculate attention probabilities for analyzing the importance of key words in the BILSTM input, obtaining output features based on these attention probabilities.

b) **GBDT-based relation classification**: Features obtained from the BILSTM model are input into the GBDT algorithm, which iteratively constructs decision trees, improves the model using the negative gradient of the previous model, and builds new decision trees in the gradient direction of residual reduction.

2.1 Attention-Based BILSTM Feature Extraction

Since LSTM cannot directly process text data, Google's open-source tool Word2vec is first used to convert text into character vectors. Assuming an input sentence S with character set $w_1, w_2, ..., w_m$, where m is the sentence length, the character vector for the t-th character is $x_t \in \mathbb{R}^d$, where d is the dimension of the word vector. The input text is then represented as $S = [w_1, w_2, ..., w_m]^T \in \mathbb{R}^{d \times m}$.

LSTM is a special type of RNN that replaces the hidden layer neural units in RNN with LSTM cells. An LSTM unit consists of three gates: an input gate, an output gate, and a forget gate. Due to this special structure, LSTM

networks can partially solve long-distance dependency problems. At time step t, the updates for each component of the LSTM unit are as follows:

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
\tilde{C_t} &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C_t} \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
h_t &= o_t * \tanh(C_t)
\end{aligned}
$$

where σ represents the sigmoid activation function, $*$ denotes element-wise multiplication, x_t is the input vector at time t, h_t represents the hidden state, and W_f, W_i, W_C, W_o and b_f, b_i, b_C, b_o represent the weight matrices and bias terms for the forget gate, input gate, memory cell, and output gate, respectively.

To fully utilize contextual information and mine more implicit features for effective relation extraction, this paper designs a bidirectional LSTM neural network consisting of two LSTM networks in opposite directions. The model structure is shown in the BILSTM layer portion of Figure 1, where $\overrightarrow{h_t}$ is the forward LSTM output at time t and $\overleftarrow{h_t}$ is the backward LSTM output at time t. The output at time t is the concatenation of forward and backward outputs: $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$.

Since each word contributes differently to sentence classification, the attention mechanism [15] is employed for deeper feature extraction to improve relation classification accuracy. For example, in the sentence "Modern research: Xu reported using Tong Mai Si Ni Tang with modifications to treat 16 cases of Shaoyin Ge Yang syndrome, with all cases cured", a standard BILSTM network treats each word equally. By introducing the attention mechanism, the model focuses on keywords such as "treat" through attention weight allocation. The Attention layer structure connected after the BILSTM model is shown in Figure 2. The global output vector obtained through the attention layer is $H = \sum_{t=1}^{m} a_t h_t$, with calculations as follows:

$$
\begin{aligned}
u_t &= \tanh(W_w h_t + b_w) \\
a_t &= \text{softmax}(u_t^T u_w) \\
H &= \sum_t a_t h_t
\end{aligned}
$$

where h_t is the hidden unit, u_w is the sentence context vector, W_w is the attention vector, a_t is the attention weight, and b_w is the bias term, randomly initialized and learned during training.

2.2 GBDT-Based Relation Classification

Relation extraction can be viewed as a multi-classification problem. Dian Yujie et al. [16] proposed applying GBDT to microblog stance detection by manually extracting features from corpora for text classification. Duan Dagao et al. [17] proposed a false message detection method based on GBDT by extracting features from comment text content, user attributes, information propagation, and temporal information. GBDT is an ensemble learner that adopts Boosting to construct m weak learners that form a final strong learner through multiple iterations. It uses CART regression trees as weak learners, with each iteration aiming to reduce the residuals from the previous model by training new models in the gradient direction of residual reduction.

The features obtained from the attention-based BILSTM model and original category labels form the training dataset $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where x_i is the feature vector extracted from the i-th sentence in the corpus and y_i is the relationship category. Assuming the GBDT loss function is $L(y, f(x))$, its expression is:

$$L(y, f(x)) = \sum_{i=1}^{m} L(y_i, f(x_i))$$

The negative gradient of the loss function for the i-th sample in the t-th round is:

$$r_{ti} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)}$$

A CART regression tree can be fitted to obtain the t-th decision tree with corresponding leaf node regions R_{tj}, where J is the number of leaf nodes. For samples in each leaf node, the optimal output value c_{tj} that minimizes the loss function is:

$$c_{tj} = \arg\min_{c} \sum_{x_i \in R_{tj}} L(y_i, f_{t-1}(x_i) + c)$$

The decision tree fitting function for this round is:

$$h_t(x) = \sum_{j=1}^{J} c_{tj} I(x \in R_{tj})$$

The final strong learner model after this round is:

$$f_t(x) = f_{t-1}(x) + h_t(x)$$

2.3 Integrated BILSTM-GBDT Model

For entity relation extraction, the attention-based BILSTM extracts text feature vectors to obtain feature combinations H, which are then classified using gradient boosting decision trees for training and prediction to obtain the final relationship category for each sentence. The advantages of BILSTM-GBDT include solving the generalization problem that occurs when traditional deep learning methods handle relation extraction, while simultaneously improving extraction precision.

The specific algorithm flow of BILSTM-GBDT is as follows:

a) Perform embedding operations on training set samples using Word2vec, resulting in vector matrix $S = [w_1, w_2, ..., w_m]$ for each input sentence.

b) Input matrix S into the BILSTM model to compute forward output $\overrightarrow{h_t}$ and backward output $\overleftarrow{h_t}$ at time t, with BILSTM layer output features $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$.

c) Initialize attention weights for each node in the attention layer, obtain attention probabilities a_t through the attention formula, and compute final output features $H = \sum_t a_t h_t$.

d) Form training dataset $D = \{(H_1, y_1), (H_2, y_2), ..., (H_n, y_n)\}$ using attention layer outputs and category labels.

e) Initialize GBDT with a constant value: $f_0(x) = \arg\min_c \sum_{i=1}^{n} L(y_i, c)$.

f) For iteration rounds $t = 1$ to m: compute negative gradients r_{ti} for each sample, fit a CART regression tree to obtain $h_t(x)$, and update the model: $f_t(x) = f_{t-1}(x) + h_t(x)$.

g) After m iterations, the final ensemble learning model is $f_m(x)$.

h) Obtain final relationship categories through iterative model training.

---

3 Experiments

3.1 Experimental Setup

To verify the effectiveness of the proposed BILSTM-GBDT entity relation extraction algorithm, we validated the model using a curated TCM relationship corpus (TCM RECoupus). The corpus sources include ancient TCM texts, teaching materials, and TCM research papers. During construction, required entity pairs were first extracted from annotated documents, followed by sentence segmentation. The final corpus contains 26,855 sentences across 11 categories. To further verify the performance of the improved algorithm, we also conducted comparative experiments on the SemEval-2010 and ACL2007 relationship corpora. Detailed information for the three corpora is shown in Table 1 .

The corpus was split into training and test sets using a 7:3 ratio for each relationship type. BILSTM parameter settings are shown in Table 2 , where dropout, learning rate, and optimizer parameters were determined through multiple experimental comparisons.

To demonstrate the advantages of the proposed model, we adopted precision (P), recall (R), and F-score as evaluation metrics.

3.2 Experimental Results

Using the TCM entity relationship corpus training data for model training and the test data for BILSTM-GBDT model evaluation, we obtained precision, recall, and F-score for 11 relationship categories as shown in Table 3 .

As shown in Table 3, for the 11 custom TCM relationships, the precision, recall, and F-score for four categories—prescription-herb, prescription-syndrome, syndrome-symptom, and pulse-syndrome—exceed 90%. This is because TCM expressions in these four relationships are relatively simple, entity forms are fixed, and the corpus proportion is larger than other relationships, resulting in superior performance.

To verify the performance of the improved algorithm, we introduced four popular relationship extraction algorithms for comparison: Support Vector Machine (SVM), Gradient Boosting Decision Tree (GBDT), deep learning method BILSTM, and BILSTM with attention mechanism (BILSTM-ATT). Both deep learning models use Softmax for relationship classification after feature extraction. Comparative results are shown in Table 4 and Figure 3 [Figure 3: see original paper].

The experimental results in Table 4 and Figure 3 show that on the TCM corpus, the ensemble algorithm GBDT outperforms SVM in precision, recall, and F-score, demonstrating that ensemble learning enhances model anti-interference capability and generalization. Overall, BILSTM achieves 4.14% higher F-score than GBDT, indicating that automatically extracted deep sentence features benefit relationship classification. BILSTM-ATT adds attention mechanism to BILSTM, achieving 2.17% higher F-score than BILSTM due to weight allocation for each input character vector, reducing the impact of noise words. BILSTM-GBDT uses GBDT as the classifier while incorporating attention mechanism, achieving 2.11% higher F-score than BILSTM-ATT.

Results on the SemEval-2010 corpus (Table 4 and Figure 3) show that GBDT still outperforms SVM across all three metrics. BILSTM-GBDT achieves the highest scores among all algorithms, with F-score 2.25% higher than BILSTM-ATT.

On the ACL2007 corpus (Table 4 and Figure 3), GBDT maintains advantages over SVM. BILSTM-GBDT demonstrates significant advantages over Softmax-based models, achieving 86.06% F-score—1.66% higher than BILSTM-ATT.

To investigate the impact of the number of gradient boosting trees (m) on model performance, we conducted experiments on the three corpora, with results shown in Figure 4 [Figure 4: see original paper]. The number of decision trees increases in multiples of 10 across 10 iterations.

As shown in Figure 4, initially, as the number of decision trees increases, the F-score of BILSTM-GBDT on all three corpora shows an upward trend. When the number of gradient boosting trees reaches 60, the model achieves optimal performance, after which the F-score stabilizes or even decreases slightly.

In summary, BILSTM-GBDT utilizes an attention-based bidirectional LSTM to fully extract sentence features and employs ensemble learning GBDT as the classifier, which partially solves the generalization problem caused by using Softmax as a classifier and yields more stable results. SVM performs worst, while GBDT enhances model stability and generalization. However, the gap between GBDT and BILSTM-GBDT remains significant due to manual feature extraction in GBDT. BILSTM-GBDT first extracts high-order features using attention-based BILSTM, then leverages ensemble learning GBDT to iteratively form multiple decision trees, enhancing model generalization. The model achieves optimal performance when the number of gradient boosting trees reaches 60.

---

4 Conclusion

This paper addresses the problem of insufficient model generalization caused by using Softmax as a long short-term memory network classifier in relationship extraction tasks, which poorly adapts to TCM entity relationship classification. We propose a bidirectional long short-term memory model integrated with gradient boosting decision trees, leveraging the advantages of bidirectional LSTM for automatic feature extraction, incorporating attention mechanism to capture keywords for sentence comprehension and reduce interference from irrelevant words, and utilizing the low variance and high bias advantages of GBDT to enhance model robustness and generalization. Comparative experiments on the TCM relationship corpus and two other public domain corpora demonstrate that the proposed model significantly improves precision, recall, and F-score, making it suitable for relationship extraction in the specific TCM domain. However, the improved algorithm can only extract predefined relationships. Future work will focus on extracting new relationships and extending the approach to other domains.

References

[1] Rong Bohui, Fu Kun, Huang Yu, et al. Relation extraction based on multi-channel convolutional neural network [J]. Application Research of Computers, 2017, 34 (3): 689-692.

[2] Chen Yu, Zheng Dequan, Zhao Tiejun. Chinese relation extraction based on deep belief nets [J]. Journal of Software, 2012, 23 (10): 2572-2585.

[3] Guo Xiyue, He Tingting. Survey about research on information extraction [J]. Computer Science, 2015, 42 (2): 14-17.

[4] Duan Liguo, Xu Qing, Li Aiping, et al. Research on effect of entities semantic information on Chinese entity relation extraction [J]. Application Research of Computers, 2017, 34 (01): 141-146.

[5] Gan Lixin, Wan Changxuan, Liu Dexi, et al. Chinese named entity relation extraction based on syntactic and semantic features [J]. Journal of Computer Research & Development, 2016, 53 (2): 284-302.

[6] Chen Peng, Guo Jianyi, Yu Zhengtao, et al. Chinese field entity relation extraction based on convex combination kernel function [J]. Journal of Chinese Information Processing, 2013, 27 (5): 144-148.

[7] Wang Yequan, Huang Minlie, Zhu Xiaoyan, et al. Attention-based LSTM for Aspect-level Sentiment Classification [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2017: 606-615.

[8] Chen Qian, Zhu Xiaodan, Ling Zhenhua, et al. Enhanced LSTM for natural language inference [C]// Proc of Meeting of the Association for Computational Linguistics. 2017: 1657-1668.

[9] Vu N T, Adel H, Gupta P, et al. Combining recurrent and convolutional neural networks for relation classification [EB/OL]. (2016-05-24). https://arxiv.org/abs/1605.07333.

[10] Zeng Daojian, Liu Kang, Lai Siwei, et al. Relation classification via convolutional deep neural network [C]// Proc of the 25th International Conference on Computational Linguistics. 2014: 2335-2344.

[11] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9 (8): 1735-1780.

[12] Miwa Makoto, Bansal Mohit. End-to-end relation extraction using LSTMs on sequences and tree structures [EB/OL]. (2016-06-08). https://arxiv.org/abs/1601.00770.

[13] Hu Jie, Li Shaobo, Yu Liya, et al. A patent classification model based on convolutional neural networks and random forest [J]. Science Technology and Engineering, 2018, 18 (6): 268-272.

[14] Jerome H. Friedman. Greedy function approximation: a gradient boosting machine [J]. Annals of Statistics, 2001, 29 (5): 1189-1232.

[15] Baziotis C, Pelekis N, Doulkeridis C. DataStories at SemEval-2017 task 4: deep LSTM with attention for message-level and topic-based sentiment analysis [C]// Proc of International Workshop on Semantic Evaluation. 2017: 747-754.

[16] Dian Yujie, Jin Qin, Wu Huimin. Stance detection in Chinese microblogs via fusing multiple text features [J]. Computer Engineering and Applications, 2017, 53 (21): 77-84.

[17] Duan Dagao, Gai Xinxin, Han Zhongming, et al. Micro-blog misinformation detection based on gradient boost decision tree [J]. Journal of Computer Applications, 2018, 38 (2): 410-414.

Note: Figure translations are in progress. See original paper for figures.