

Intelligent Prescription-Generating Models of Traditional Chinese Medicine Based on Deep Learning

Qing-Yang Shi^a, Li-Zi Tan^a, Lim Lian Seng^b, Hui-Jun Wang^a

^aDepartment of Traditional Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin, China, ^bDepartment of Traditional Chinese Medicine, Malaysia Chinese Medical Association, Kuala Lumpur, Malaysia

Abstract

Objective: This study aimed to construct an intelligent prescription-generating (IPG) model based on deep-learning natural language processing (NLP) technology for multiple prescriptions in Chinese medicine. **Materials and Methods:** We selected the Treatise on Febrile Diseases and the Synopsis of Golden Chamber as basic datasets with EDA data augmentation, and the Yellow Emperor's Canon of Internal Medicine, the Classic of the Miraculous Pivot, and the Classic on Medical Problems as supplementary datasets for fine-tuning. We selected the word-embedding model based on the Imperial Collection of Four, the bidirectional encoder representations from transformers (BERT) model based on the Chinese Wikipedia, and the robustly optimized BERT approach (RoBERTa) model based on the Chinese Wikipedia and a general database. In addition, the BERT model was fine-tuned using the supplementary datasets to generate a Traditional Chinese Medicine-BERT model. Multiple IPG models were constructed based on the pretraining strategy and experiments were performed. Metrics of precision, recall, and F1-score were used to assess the model performance. Based on the trained models, we extracted and visualized the semantic features of some typical texts from treatise on febrile diseases and investigated the patterns. **Results:** Among all the trained models, the RoBERTa-large model performed the best, with a test set precision of 92.22%, recall of 86.71%, and F1-score of 89.38% and 10-fold cross-validation precision of 94.5% \pm 2.5%, recall of 90.47% \pm 4.1%, and F1-score of 92.38% \pm 2.8%. The semantic feature extraction results based on this model showed that the model was intelligently stratified based on different meanings such that the within-layer's patterns showed the associations of symptom-symptoms, disease-symptoms, and symptom-punctuations, while the between-layer's patterns showed a progressive or dynamic symptom and disease transformation. **Conclusions:** Deep-learning-based NLP technology significantly improves the performance of IPG model. In addition, NLP-based semantic feature extraction may be vital to further investigate the ancient Chinese medicine texts.

Keywords: Ancient books of Chinese medicine, bidirectional encoder representations from transformers, deep learning, intelligent prescription-generating models, pretrained models

INTRODUCTION

The rapid development of deep learning has popularized its applications in the medical field, including the use of computer vision in medical imaging and natural language processing (NLP) for electronic medical data records, reinforcement learning for robot-assisted surgery, and generalized deep learning for genetic analysis.^[1] Accordingly, deep learning can be applied extensively to Traditional Chinese Medicine (TCM) research. Existing studies include TCM tongue image recognition, TCM smell recognition, and TCM text classification.^[2-5] Although the intelligent identification of prescriptions in TCM has not been investigated widely, it has been performed based on machine learning,^[6] which is not effective owing to the lack of training data and the

simple methods involved. Deep learning has been employed as well,^[7] but its final classification is unsatisfactory for the following reasons: (1) in cases involving a small amount of data and direct deep-learning training results in insufficient model generalization capability, (2) data pertaining to TCM prescriptions used for training are unorganized,

Address for correspondence: Prof. Hui-Jun Wang,
Tianjin University of Traditional Chinese Medicine, Beihua South Road,
Jinghai District, 301617, Tianjin, China.
E-mail: whj-002@163.com

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

© 2021 World Journal of Traditional Chinese Medicine | Published by Wolters Kluwer - Medknow

Received: 26-06-2020, **Accepted:** 29-10-2020, **Published:** 09-08-2021

How to cite this article: Shi QY, Tan LZ, Seng LL, Wang HJ. Intelligent prescription-generating models of traditional chinese medicine based on deep learning. World J Tradit Chin Med 2021;7:361-9.

Access this article online

Quick Response Code:



Website:
www.wjtcn.net

DOI:
10.4103/wjtcn.wjtcn_54_21

thereby hindering the models from effectively learning their characteristics, and (3) the technical conditions for deep learning are limited previously. Therefore, based on previous studies, we performed our current study based on deep-learning NLP and transfer learning approaches.

Chinese medicine has a long history and is based on identifying and prescribing evidence, i.e. from the earliest written *Formulas for Fifty-two Diseases* to the *Treatise on Febrile and Miscellaneous Diseases* from the Eastern Han dynasty, *Formularies of the Bureau of People's Welfare Pharmacies* from the Tang and Song dynasties, and *Formularies for Universal Relief* from the Ming and Qing dynasties, all of which provide a wealth of prescription experiences. The most classic texts among these are *Treatise on Febrile Diseases* and *Synopsis of Golden Chamber*, which are recognized as the “members of the Four Classics” and the “ancestors of formularies.”^[8,9] Therefore, these two ancient texts were analyzed in this study to investigate the intelligent prescription-generating (IPG) model of TCM.

The ancient texts above, while authoritative and classic, contain an extremely low amount of valid data, whereas deep-learning models are generally complex and require a large amount of training data for generalization; otherwise, it can be easily overfitted. Therefore, transfer learning^[10] was adopted in this study to generalize the trained language model; subsequently, the generalized model was transferred to the TCM prescribing task, which was then fine-tuned and learned. The following three pretraining models are proposed herein: (1) a word-embedding vector model based on training *Imperial Collection of Four* (Word embedding, W2V),^[11,12] (2) the bidirectional encoder representations from transformers (BERT) model based on Chinese Wikipedia dataset training,^[13] (3) a fine-tuned BERT model (TCM-BERT) based on the TCM dataset,^[4] and (4) the robustly optimized BERT approach (RoBERTa) model based on the Chinese Wikipedia and a general database.^[14] In addition, data augmentation was performed on the original dataset to enhance the generalization capabilities of the models.^[15]

In summary, we compared the prediction performances of different deep-learning models on a testing set to identify the most suitable TCM intelligent prescribing model, extracted the semantic features of the models, and analyzed them visually.

MATERIALS AND METHODS

Data extraction and data augmentation

In this study, we used Zhao Kaimei's edition of the *Treatise on Febrile Diseases* and the *Synopsis of Golden Chamber* as the original data source and manually extracted all the formulas as well as their corresponding certification provisions. Finally, we extracted 385 prescription-syndrome records. The textual data augmentation method proposed by Wei and Zou in 2019^[15] was subsequently used; it comprised (1) synonym replacement, (2) random insertion, (3) randomly swap, and (4) randomly deleted, with the dataset expanded to 770 records.

In this study, Chinese word segmentation was not considered as preprocessing, but character-based basic sentences were trained. Character-level inputs were used because of the following reasons: (1) In 2019, Li *et al.* suggested segregating and not segregating Chinese words for deep learning. Although similar test results were obtained, the results without word segregation were slightly better.^[16] (2) This study is based on ancient Chinese texts, in which the ideographic ability of individual characters is stronger than that of modern Chinese texts. Therefore, character-level sentence inputs were used.

Model validation

Two methods were used to perform validation. The first is the hold-out approach, which randomly segregates the dataset into two portions at a ratio of 2:8. The “8” portion is used to train model, whereas the “2” portion is to test model. The second is a 10-fold cross-validation approach, which randomly segregates the dataset into 10 equal portions.^[17] Each time, 1 among 10 sets, is selected as the test dataset, whereas the others are used as the training dataset. Finally, the results of 10 trials are summarized in terms of the mean and standard deviation.

Performance metrics

Since the output of the model is an herbal prescription, i.e. a multilabel classification task, we selected classification task-related metrics, i.e. accuracy, error, precision, recall, and F1-score. However, the training data were extremely unbalanced. Hence, we did not focus on the zero values in the classification as they represented “no medication use.” The main purpose of the task was to detect whether prescription medications were administered appropriately. Therefore, we abandoned the metrics of accuracy and error rates because the accuracy is at least 85% when all predicted values are zero, representing a wrong model and the same error rate. Therefore, precision, recall, and F1-score were selected as the final metrics.

The formulas are presented as follows: $precision = \frac{TP}{TP + FP}$

; $Recall = \frac{TP}{TP + FN}$; and $F1 = \frac{2 * Precision * Recall}{Precision + Recall}$. Here,

true positive is the number of true positives, false positive the number of false positives, and FN the number of false negatives; furthermore, the F1-score is the harmonic mean of the precision and recall.

Model construction

Pretrained word-embedding models

The word-embedding vector is a distributed representation relative to one-hot encoding. One-hot encoding words are represented as sparse, high dimensional, and independent of each other. Whereas word embedding is dense, low dimensional, and derived from learning, word embedding considers more semantic relationships between words that offer greater advantages for textual tasks. The use of the pretrained word-embedding model proposed by Li *et al.* in 2018^[11] is proposed herein, where the pretrained data are from *Imperial*

Collection of Four, which is more consistent with the linguistic conventions of ancient Chinese texts as well as with the data set of this study; furthermore, the model is the best pretrained model among ancient Chinese word-embedding models.

Fully connected neural networks

Fully connected neural networks or deep neural networks are the most basic deep-learning models; however, they offer strong fitting capabilities. Therefore, the testing results of these models were used as a basis for other models in this study. The models were set as a pretrained word-embedding layer, followed by access to a two-layer 128-dimensional fully connected layer with the activation function set to a linear rectifier function (rectified linear unit (ReLU)). Each layer was followed by a dropout layer (0.5 ratio), and the output layer was a 110-dimensional fully connected layer, whose activation function was the sigmoid function, which enables multilabel classification. The loss function was set to binary cross-entropy, and the optimizer was set to adaptive moment estimation,^[18] where the learning rate and batch size were set to 0.001 and 8, respectively.^[19]

Recurrent neural networks and attention models

An RNN is based on serial data, where a class of neural networks in which all neurons are chained together as inputs are recursive in the direction of the sequence.^[20] This model was used because the textual data used in this study indicated the sequence's characteristics. Considering the vanishing gradient problem inherent in RNNs, i.e. the gradient is dominated by the proximity gradient, and the model cannot learn long-range dependencies,^[21] this study was modeled using a gated RNN, and the long short-term memory unit (LSTM) was used as its gating mechanism.^[22,23] The model was set as a pretrained word embedded in the model and then connected to a two-layer bidirectional 128-dimensional LSTM layer, with an input unit dropout ratio of 0.2, a loop unit dropout ratio 0.2 in each layer, and a single-layer 128-dimensional fully connected layer, where the activation function was set to ReLU.^[24] The loss function, optimizer, learning rate, and batch size settings were as detailed above.

In 2017, Vaswani *et al.* proposed an attention-based model that demonstrated the advantages of the self-attention layer in sequence modelling.^[25] Therefore, this study was modeled using a self-attention model, which was then compared with an RNN. The model was set to a pretrained word-embedding model after accessing the location information embedding layer, which was set to a 300-dimensional output vector and a word-embedding output. Subsequently, the summed vectors were incorporated into the two 128-dimensional self-attention layers, the activation function was set to sigmoid, and the connection layer was incorporated into the dropout layer with its rate set to 0.5. The output layer, loss function, optimizer, learning rate, and batch size setting were as described above.

Furthermore, the attention mechanism is essentially a matrix dot product model with no natural ability to model sequences. Hence, when modeling using the self-attention

model, the location encoding information is first introduced, and the location information embedding layer is constructed. Subsequently, the information embedding layer is added with the word-embedding layer as the sequence modeling information source, however, is not addressed well. Therefore, the models were hybridized from the first two sections, where the sequence modeling of RNNs replaced the location information embedding layer. The model was set as a pretrained word embedded in the model and then connected to two bidirectional 128-dimensional LSTM layers, in which the input unit dropout ratio and the recurrent unit dropout ratio were set to 0.2 in each layer, two layers of the 128-dimensional self-attention layer were accessed, and the activation function was set to sigmoid. Subsequently, the connection layer was flattened and connected to the dropout layer at a ratio of 0.5. The output layer, loss function, optimizer, learning rate, and batch size settings were as described above.

Bidirectional encoder representations from transformers series pretrained models

The BERT models^[13] are bidirectional encoder representations from transformers,^[25] where the transformer encoding structure is based on the multihead self-attention mechanism. Currently, the BERT structure (a series of transformer-based structures) is the best characterization extraction model in NLP; it integrates various structures such as word embedding, location information embedding, self-attention layer, and feedforward neural network layers in multiple language texts. It has demonstrated its best performance in several missions. Therefore, the official BERT pretraining model provided by Google (BERT-base),^[26] which is a Chinese Wikipedia, was used in this study. The model is structured as a 12-layer transformer block with 768 dimensions in each layer and 12 multiple attention counts. This enables it to access the 111-dimensional fully connected layer as an output layer at its (CLS) location after pretraining the BERT model. The fine-tuning layer was set to the multihead-attention layer in the encoder-12 layer, and the remaining layers were freezing weights. The loss function, optimizer, learning rate, and batch size settings remain unchanged.

Since the official BERT phenotype extraction is more inclined to modern Chinese habits and lacks information regarding Chinese medicine, further improvements can be considered. Therefore, the basic BERT model was fine-tuned with criterion-free data to introduce semantic information for TCM. *Yellow Emperor's Canon of Internal Medicine*, *Classic of the Miraculous Pivot*, and *Classic on Medical Problems* were selected as the fine-tuning datasets. The training method is based on advice provided by Google. First, a paired sentence structure is constructed for next-sentence prediction (NSP) training preparation; subsequently, the NSP is indicated as positive when two sentences are in a contextual relationship and negative otherwise. Second, the mask language model (MLM) is trained using the same basic BERT model. This model is known as the pretrained TCM-BERT model. Finally, the classification model is constructed, and the model is set to

pretrain the TCM-BERT model. Subsequently, it accesses the 111-dimensional full connection layer in the (CLS) position for classification, and the fine-tuning layer is set to multihead attention in the encoder-12 layer by freezing the remaining weights. The loss function, optimizer, learning rate, and batch size settings remain unchanged.

In addition, the RoBERTa model, also known as the robust optimal BERT approach,^[14] is an optimized and improved version of BERT; it spearheaded the NLP mission in 2019 and achieved state-of-the-art results in the current NLP field. Its main improvements were as follows: (1) the introduction of larger datasets, a longer training duration, and a larger batch size; (2) the use of the NSP method for training was abandoned; instead, whole-sentence training and filling to 512 words in length were performed directly; and (3) the static MLM method was replaced by a dynamic mask method, which generated a new mask strategy for each input sequence. In this study, a pretrained RoBERTa model was used additionally for modeling, the basic Chinese versions of the RoBERTa and RoBERTa-large models were selected based on large databases, and the pretrained corpus was the Chinese Wikipedia as well as the general Chinese corpus of various encyclopedias, news, quizzes, etc.^[27] Model parameters for 12- or 24-layer transformer blocks with 768 dimensions or 1024 in each layer and with 12 or 16 multiple attention layers were used, totaling 110M or 330M parameters, respectively. The classification layer settings were the same as those in the previous BERT model. The loss function, optimizer, learning rate, and batch size settings remained unchanged.

Semantic feature extraction and visualization

The most important information in a language model is the semantic features of a sentence, and the ability to learn semantic features is the key to assessing the model's goodness. For the trained model, the semantic information in its downstream layer is a good representation of the information features that are the most similar to the task requirements. Therefore, in this study, information regarding semantic features learned from the best model (RoBERTa-large) was extracted, and raw data articles were visualized to analyze the evidence and treatment patterns.

The transformer block in BERT serves as a feature extractor for the model, with each layer representing a different semantic feature, ranging from general linguistic and grammatical features in the upstream layer to those that are the most similar to the classification task in the downstream layer. The 768 dimensions in each layer represent different representations of features. (1) First, we extracted the semantic features of the transformer block in the downstream layer, which was an $n \times 768$ matrix (n for sentence length), representing the feature matrix. (2) Next, we performed a dot product of the matrix with its own transpose matrix, i.e. an $n \times n$ matrix, representing the autocorrelation matrix. (3) Subsequently, the dimensions of the $n \times 768$ matrix were reduced using isometric mapping in manifold learning,^[28] where manifolds that were locally identical to the embryonic

nature of Euclidean space were used – each point locally identifies the nearest neighbors through Euclidean distances to establish nearest-neighbor connections. The distance between nonneighboring points was infinite, and when a distance existed between any two points, it was scaled via multidimensional scaling for low-dimensional spatial coordinate calculations.^[29] After degradation, an $n \times 5$ matrix was obtained as a matrix of sentence feature weights for five-layer sentence visualization, and the sentence meaning was analyzed accordingly.

RESULTS

Traditional Chinese Medicine intelligent prescription-generating model fitting and prediction results

The experimental results of the fully connected neural networks were as follows: Training set loss value of 0.0164, accuracy of 93.39%, recall of 91.72%, and F1-score of 92.55%; test set loss value of 0.1560, accuracy of 62.76%, recall of 42.86%, and F1-score of 50.94%. Although the model froze the weights of the pretrained word embeddings and premature overfitting occurred after eight rounds of model training. The test set accuracy and recall rates no longer increased, and the F1-score remained stable at approximately 50.1%, which was not ideal and therefore only used as a basic effect in comparison.

The experimental results of the RNN were as follows: training set loss value of 0.0398, accuracy of 88.05%, recall of 74.43%, and F1-score 80.67%; test set loss value of 0.0623, accuracy of 78.74%, recall of 66.14%, and F1-score of 71.89%. Pretrained words are embedded and then incorporated into an RNN. Moreover, a sequence modeling approach was introduced, which yielded better results, indicating the importance of sequence or location information in language modeling.

The experimental results of the self-attention model were as follows: training set loss of 0.0247, accuracy of 93.4%, recall of 86.49%, and F1-score of 89.81%; test set loss of 0.0527, accuracy of 82.83%, recall of 66.86%, and F1-score of 73.99%. The final test F1-score was 2% points higher than that of the RNN, which was similar to the current mainstream test results and therefore not a significant advantage. The self-attention model performed slightly better than RNNs, with no significant advantage. However, owing to their parallel training ability, it was 10 times faster than the RNN.

The experimental results of the RNN self-attention model were as follows: training set loss of 0.0135, accuracy of 95.22%, recall of 93.16%, and F1-score of 94.18%; test set loss of 0.0540, accuracy of 88.34%, recall of 84.43%, and F1-score of 86.34%. The final test set F1-score increased by 12% points compared with that of the self-attention model, indicating improved performance. Introducing RNNs as a source of information for sequence modeling and hybrid building with self-attention models yielded good results, thereby demonstrating that the hybrid modeling approach is well suited for prescription model construction.

The experimental results of the BERT-based model were as follows: training set loss of 0.0075, accuracy of 98.16%, recall of 97.22, and F1-score of 97.68%; test set loss of 0.0504, accuracy of 88.29%, recall of 79.71%, and F1-score of 83.78%. The final model test set F1-score was 3% points lower than that of the RNN self-attention model, indicating relatively good performance. Although the BERT-based classification model was slightly less effective than the aforementioned hybrid model, it was significantly less expensive to train than the latter as it required only fine-tuning; moreover, it did not require any additional structures.

The experimental results of the TCM-BERT model were as follows: training set loss of 0.0064, accuracy of 98.58%, recall of 97.73%, and F1-score of 98.15%; test set loss of 0.0443, accuracy of 89.98%, recall of 83.43%, and F1-score of 86.58%. The final test set F1-score was 3% points higher than that of the BERT-based model and exhibited the same effect as the previous mixed model. After the introduction of TCM information, the model enhanced partially to the effect of the original hybrid model, indicating that the fine-tuning of knowledge was beneficial; however, the improvement was limited, primarily because the BERT-based model possessed sufficiently good results. However, the effect of this model satisfied our expectations, and the improved results were similar to those of mainstream testing in academia, i.e. knowledge fine-tuning can improve by 2%–3%.^[4]

The experimental results of the RoBERTa model were as follows: training set loss of 0.0091, accuracy of 97.66%, recall of 96.22%, and F1-score of 96.93%; test set loss of 0.0426, accuracy of 89.78%, recall of 84.14%, and F1-score of 86.86%. The final test set F1-score was the same as that of the TCM-BERT model. The RoBERTa model achieved the same results as the TCM-BERT model when it was not fine-tuned for specific knowledge, indicating a rare improvement in BERT. The success of omitting the fine-tuning of specific knowledge illustrates the generality of the language model not only in modern Chinese texts but also in ancient Chinese texts, indicating the necessity for a large pretraining corpus.

The results of the RoBERTa-large model were as follows: training set loss of 0.0029, accuracy of 99.66%, recall of 99.48%, and F1-score of 99.56%; test set loss of 0.0401, accuracy of 92.22%, recall of 86.71%, and F1-score of 89.38%. The final test set F1-score of the RoBERTa-large model was 3% points higher than that of the RoBERTa model, thereby rendering it the best model, i.e., comparable to the mainstream test improvement in academia. The degree of similarity was approximately 1% to 4% points. RoBERTa-large model has become the most effective model available because it uses a larger model structure, a larger dataset, and some details pertaining to optimization. Furthermore, this model is superior to the original TCM-BERT model as it does not require knowledge fine-tuning, thereby allowing the dataset to reach the optimal solution more easily.

Next, we provide the model fitting and prediction results of the main applications of this study and analyze the advantages and

disadvantages of the different models. We comprehensively compare the results of different model tests and introduce cross-validation results and experiments based on the supplementary datasets. Table 1 summarizes the assessment metrics for the hold-out and cross-validation methods. The values for the metrics calculated under the cross-validation method were higher than those calculated under the hold-out method; this may be due to the hold-out method ratio of 0.8–0.2, compared with the 10-fold cross-validation ratio of 0.9–0.1. Since more trained data were used, the basic dataset was enhanced by the EDA data treatment, which decreased the degree of overfitting. Therefore, theoretically, if more data are used for training, then the test results will be better; however, the results might be limited by a certain threshold, i.e. 0.9–0.1. The RoBERTa-large model demonstrated the best performance compared with the other models. In the cross-validation, because of random errors, the *t*-test revealed that the RoBERTa-large test F1-score was significantly higher than that of RoBERTa ($t = 2.259$; $df = 18$; $P = 0.0365$) but was not significantly different from that of TCM-BERT ($t = 1.838$; $df = 16.978$; $P = 0.0835$) or that of the mixed model (W2V + LSTM + SA) ($t = 1.9491$; $df = 17.991$; $P = 0.067$). However, because of the small sample size and insufficient statistical power, the results might be false negative results. In general, the classification based on the RoBERTa-large model is the best option for the IPG model because additional pretrained datasets and model structures are not required, i.e. only fine-tuning in the model downstream layer is required.

Semantic feature extraction and visualization analysis

The extracted feature matrix of article 1 in *Treatise on Febrile Diseases* is shown in Figure 1a, which shows a total of 768 layers of features, each of which differs from the other. It represents different meanings and can be regarded as each character embedded in a 768-dimensional space, and the 1×768 -dimensional vector represents its meaning. The dot product of the feature matrices yields an autocorrelation matrix, as shown in Figure 1b, which show that each word has the highest correlation with itself. In addition, “tai (太)” has a higher correlation with “yang (阳),” “taiyang (太阳)” has a higher correlation with “disease (病),” followed by “floating (浮),” “head (头),” and “cold (寒).” The correlation between “pulse (脉)” and “floating (浮)” is high, and the correlations between “floating pulse (脉浮)” and “head (头)” with “stiffness and hurt with the fear of severe cold (强痛而恶寒)” are high. The correlation between “head (头)” and “stiffness and hurt with the fear of severe cold (强痛而恶寒)” was high, and the in-between correlation of “stiffness and hurt with the fear of severe cold (强痛而恶寒)” is high. The correlations above show that this model searches for semantic associations between words and words in sentences to better understand the meaning of sentences. This model completely separated the meanings of words and punctuation marks, where the highest correlations were achieved between punctuation marks and the lowest correlations among words. Among them, “start” and “end” demonstrated the

Table 1: Test performance of different models						
Models	Precision	Recall	F1-score	CV- Precision	CV- Recall	CV-F1-score
W2V + DNN	62.76	42.86	50.94	62.71±1.7	48.6±4.1	54.67±2.8
W2V + LSTM	78.74	66.14	71.89	81.44±5	68.26±3.8	74.25±4.2
W2V + SA	82.83	66.86	73.99	83.35±3.3	71.67±5.2	76.96±3.6
W2V + LSTM + SA	88.34	84.43	86.34	92.7±2.6	87.17±3.5	89.82±2.7
BERT-base	88.29	79.71	83.78	88.71±4	85.62±5.1	87.02±3.6
TCM-BERT	89.98	83.43	86.58	91.81±3.9	87.64±5	89.57±3.6
RoBERTa	89.78	84.14	86.86	91.49±2.5	87.48±4.1	89.39±2.8
RoBERTa-large	92.22	86.71	89.38	94.5±2.5	90.47±4.1	92.38±2.8

W2V: Pretrained Word-to-Vector model; DNN: Deep Neural Network; LSTM: Long Short-Term Memory model; SA: Self-Attention; CV: 10-fold Cross-validation

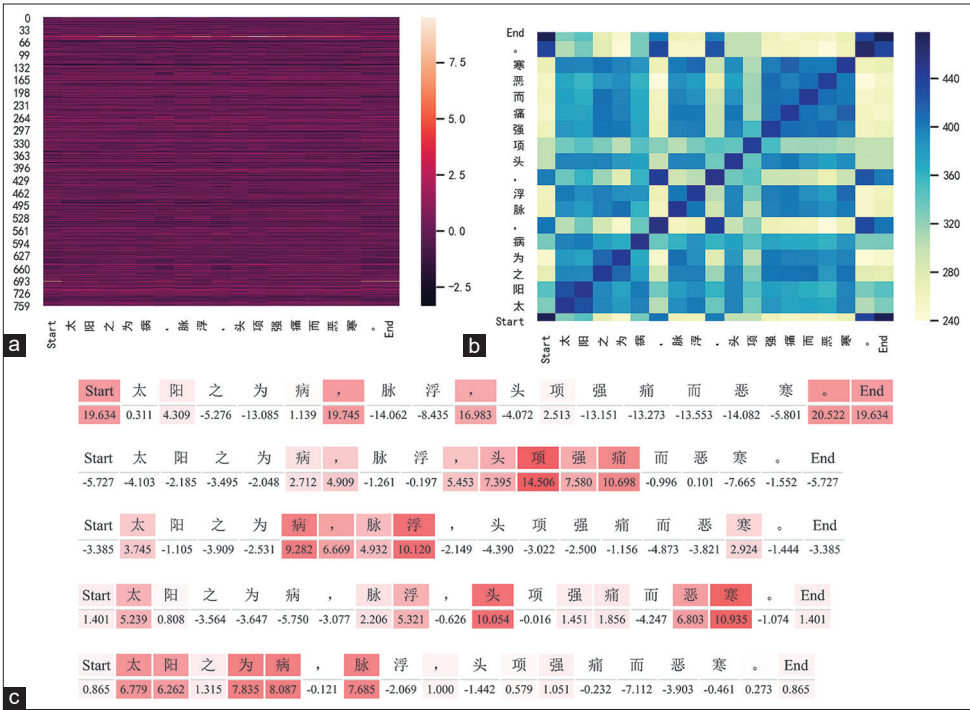


Figure 1: Visualization of semantic feature of article 1. (a) Extracted feature matrix. (b) Autocorrelation matrix. (c) Syntactic weights of layers

highest correlation with each other as start and end characters, respectively; additionally, they indicated higher correlations with punctuation marks and lower correlation with words. As shown in Figure 1c, the first level of sentence weighting represents the sentence interval, and the second level emphasizes the importance of “stiffness and hurt in head and on the back (头项强痛)” in the sentence. The third layer emphasizes “floating pulse (脉浮),” the fourth layer emphasizes “head (头)” and “fear the cold (恶寒),” and the fifth layer emphasizes “taiyang diseases (太阳病)” and “pulse (脉).” The different symptoms in the article are represented in different weight levels, which can be categorized into the following three symptoms: “floating pulse (脉浮),” “stiffness and hurt in head and on the back (头项强痛),” and “fear of cold (恶寒),” which are of equal importance and individually affect the final prescribing process.

The extracted feature matrix of article 2 in *Treatise on Febrile Diseases* is shown in Figure 2a, where the point product yields

the sentential autocorrelation matrix. As shown in Figure 2b, the correlation among “tai (太),” “yang (阳),” and “diseases (病)” is high; the correlation between “taiyang diseases (太阳病)” and “slow pulse (脉缓)” is high; the correlation between “fa (发)” and “heat (热)” is high; and the correlations between “getting fever (发热)” and “sweating (汗出)” as well as between “taiyang diseases (太阳病)” and “attacked by wind (中风)” are higher. These correlations imply that the symptoms are generally highly correlated, indicating the greater possibility of words that compose the symptoms occurring together with the words, as well as the correlation between some symptoms and other symptoms, such as “getting fever (发热)” and “sweating (汗出)” or “taiyang diseases (太阳病),” “slow pulse (脉缓),” and “attacked by wind (中风),” which are likely to be combined. As shown in Figure 2c, the syntactic weights of layers 1 and 2 are combined as sentence intervals, with layer 1 focusing on the first and last intervals and layer 2 focusing on intervals. The third and fourth levels

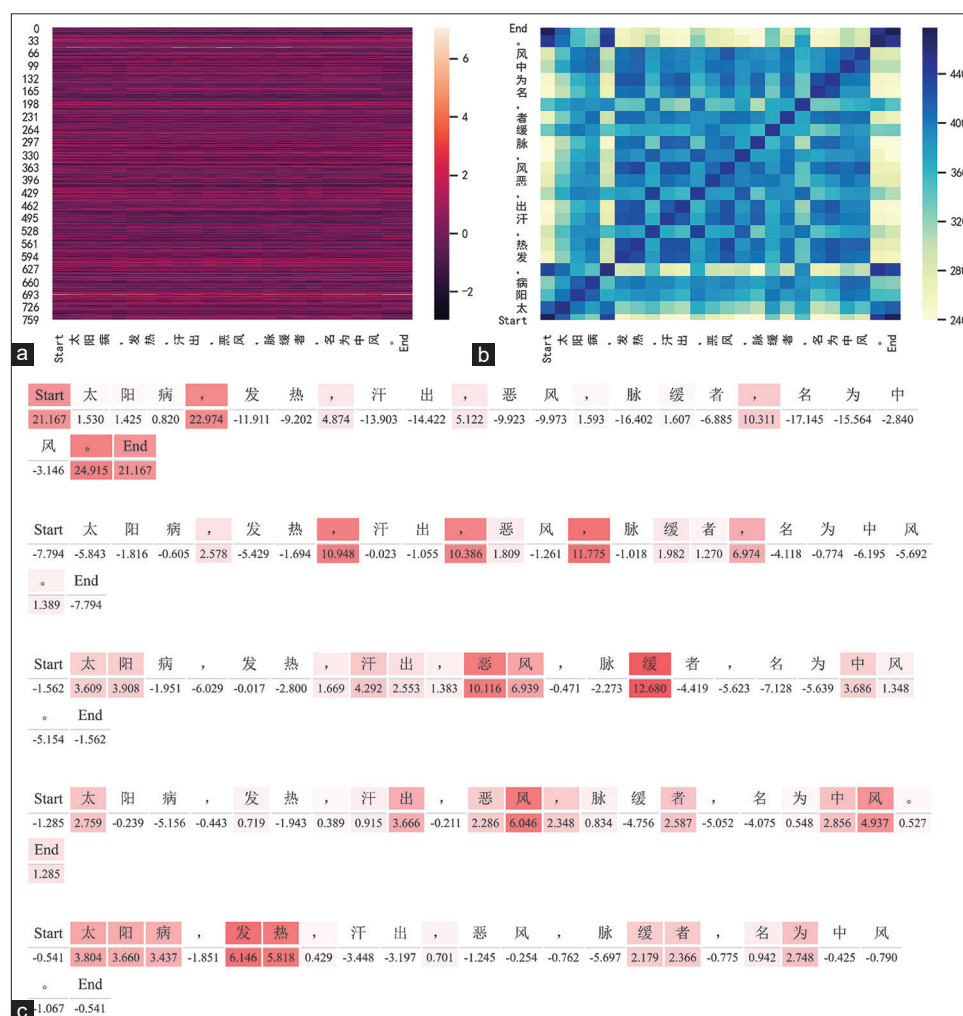


Figure 2: Visualization of semantic feature of article 2. (a) Extracted feature matrix. (b) Autocorrelation matrix. (c) Syntactic weights of layers

jointly emphasize “taiyang (太阳),” “sweating (汗出),” “fear of the wind (恶风),” “slow pulse (脉缓),” and “attacked by wind (中风).” The third level focuses on “sweat (汗),” “fear (恶),” and “slow (缓),” whereas the fourth level focuses on “out (出),” “pulse (脉),” and “wind (风).” The fifth level emphasizes “taiyang diseases (太阳病)” and “getting fever (发热).” Considering the sentence weights, the degree of association varies by the symptoms, e.g. “sweating (汗出)” is associated more closely with “fear of the wind (恶风),” whereas “taiyang diseases (太阳病)” is associated more closely with “getting fever (发热).” This implies that a hierarchical analysis between the different symptoms presented in *Treatise on Febrile Diseases* may be beneficial.

In this study, the features of original language statements from the model with the best test results were extracted and then presented visually. A few typical articles from *Treatise on Febrile Diseases* were selected for the analysis of patterns and treatment laws, although the specific manifestations of different articles were different. In general, the model effectively layered information from the text and identified autocorrelations embedded in the text by correlating the

information in each text to the entire book and then converting it to high-dimensional information. Consequently, the model achieved satisfactory results, which may provide a basis for future analysis approaches pertaining to ancient Chinese medical text studies.

DISCUSSION

In 2019, Esteva *et al.* discussed several important applications of deep learning in modern medicine and profoundly affected some areas of medical research.^[1] However, in the field of TCM, deep-learning research is limited to image recognition and simple text classification, and studies pertaining to intelligent Chinese medicine prescription are scarce. The Seq2Seq model was used to generate prescriptions^[7] in a study conducted in 2018; although the assumption was good, the results obtained were unsatisfactory. This is attributable to the following reasons: (1) failure to use the “pretrained,” “fine-tuned” model, which was directly trained with large datasets alone. However, in TCM, each prescription comprises a sentence, followed by the disease differentiation information; therefore, without using a pretrained model to extract the

linguistic characteristics, the model will undoubtedly fail. (2) In addition, owing to the lack of knowledge pertaining to TCM theories, most of the studies were performed based on computer science. Models were merely built by inputting and outputting data without an appreciation of the diverse features of different prescriptions. Hence, the simultaneous end-to-end training of all models will complicate the model and eventually yield unsatisfactory results.

This study differs from the studies above – the characteristics of existing studies are inherited and a research direction is established. (1) The prescription datasets should be small and compact, with clear and specific prescription ideas, as well as belong to the recognized classics in the field of Chinese medicine. (2) Using a transfer learning approach, a pretraining model of the language is first constructed to learn the semantic conventions of the language; subsequently, a transfer learning model of the downstream task is fine-tuned to learn the association between semantic features and prescribed medication.

Therefore, the following tasks were completed in the present study: Based on the basic theories of Chinese medicine, we selected *Treatise on Febrile Diseases* and *Synopsis of Golden Chamber* as the basic datasets, as well as *Yellow Emperor's Canon of Internal Medicine*, *Classic of the Miraculous Pivot*, and *Classic on Medical Problems* as supplementary datasets for fine-tuning; additionally, EDA data augmentation analysis was performed on the underlying dataset. We selected the word-embedding model based on the *Imperial Collection of Four* for the pretrained models; and the BERT model based on the Chinese Wikipedia and large databases, in which the base BERT model was fine-tuned with a fine-tuned dataset, as the TCM-BERT pretrained models. We constructed multiple deep-learning prescription models and conducted multiple experiments based on the underlying and supplemental datasets, where the test set and 10-fold cross-validation F1-scores were used as evaluation indicators to obtain the best results for the RoBERTa-large model. Based on the trained RoBERTa-large model, we extracted the semantic features and completed visualization analysis.

In the semantic feature extraction and visualization analysis, the following patterns were observed: For each of the articles in *Treatise on Febrile Diseases*, instead of merely memorizing the combinations of symptoms, the model first performs processing based on different meanings and patterns of symptoms within the different layers. Some of the symptoms were correlations between important symptoms, some were correlations between other symptoms, and some were punctuations in sentences representing the intervals. The patterns varied by layer, where some represented progressive relationships, whereas others can be interpreted as dynamic disease symptom development. In addition, the model searched for both intra-article autocorrelation and the overall correlation between the text and book. Since most of the current analyses of *Treatise on Febrile Diseases* involve the direct interpretation

of the sentence meaning and the memorization of symptom combinations, the model above performs a hierarchical analysis. However, the complexity of the hierarchical analysis increases exponentially with the number of layers; therefore, manual analyses can be laborious. In this regard, deep learning is highly advantageous because it projects low-dimensional information into a higher-dimensional space for analysis through a nonlinear approach, where the amount of observable information is increased significantly, and the processing of information in the high-dimensional space can yield unexpected effects, which is the key to successful model testing.

In this study, only the BERT model was used for fine-tuning when specific field knowledge was introduced to train the TCM-BERT model. This resulted in a 2%–3% improvement to the model, which, similar to the effect of the RoBERTa model, was not extremely significant. The introduction of TCM knowledge through an expert is expected to yield significant improvements because the introduction of TCM knowledge through fine-tuned methods is not the best option. This is because an expert can not only to understand and extrapolate from classical medical books but can also perform judgments based on his own TCM theoretical framework or a TCM knowledge network; clearly, these cannot be accomplished using general models. Knowledge mapping is an excellent option for introducing TCM knowledge networks. However, pure knowledge mapping techniques are used in smart queries and searches that cannot be used to build prescription models. Therefore, the idea of combining knowledge mapping with the BERT pretraining model is proposed, such that the extraction of linguistic semantic features can be retained while an *a priori* framework of TCM knowledge is introduced. However, this technology is not currently accessible, and the most recent study pertaining to it was that by Liu in 2019, where the K-BERT model was used. In that study, soft-position and visible matrices were proposed to solve the “knowledge noise” problem.^[30] Therefore, it provides a basis for future studies that entail combining the TCM knowledge map and the BERT series model.

In summary, the IPG model was successfully constructed through a series of strategies. The test set and 10-fold cross-validation F1-scores were 89.38% and 92.38% \pm 2.8%, respectively. These results are superior to those of existing studies. In addition, NLP-based semantic feature extraction and visualization analysis may be vital to the study of ancient Chinese medicine texts.

CONCLUSION

Deep-learning-based natural language processing technology improves the performance and visualization of the intelligent prescription-generating model.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, *et al.* A guide to deep learning in healthcare. *Nat Med* 2019;25:24-9.
2. Liu L, Yang F, Jing Y, Xin L. Data mining in Xu Runsan's Traditional Chinese Medicine practice: Treatment of chronic pelvic pain caused by pelvic inflammatory disease. *J Tradit Chin Med* 2019;39:440-50.
3. Liu M, Wang X, Zhou L, Tan L, Li J, Guan J, *et al.* Study on extraction and recognition of traditional Chinese medicine tongue manifestation: Based on deep learning and migration learning. *J Tradit Chin Med* 2019;60:835-40.
4. Yao L, Jin Z, Mao C, Zhang Y, Luo Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. *J Am Med Inform Assoc* 2019;26:1632-6.
5. Liang Z, Liu J, Ou A, Zhang H, Li Z, Huang J. Deep generative learning for automated EHR diagnosis of traditional Chinese medicine. *Comput Methods Programs Biomed* 2019;174:17-23.
6. Yang Y, Ruan C, Pei C, Yang M, Zhong Y, Zhang Y, *et al.* Exploration of introducing artificial intelligence to construct tcm prescription system for lung cancer. *Mode Tradit Chin Med Mater Med World Sci Technol* 2019;21:977-82.
7. Li W, Yang Z, Sun X. Exploration on generating traditional Chinese medicine prescription from symptoms with an end-to-end method. *arXiv:1801.09030*, 2018. Available from: <http://arxiv.org/abs/1801.09030>. [Last accessed on 2019 Nov 09].
8. Li P, Liu D. *Lecture Notes of Treatise on Febrile Diseases*. 1st ed. Shanghai, China: Shanghai Science and Technology Press; 2018.
9. Li K, Yang B. *Lecture Notes of Synopsis of Golden Chamber*. 1st ed. Shanghai, China: Shanghai Science and Technology Press; 2017.
10. Pratt L, Thrun S. Guest Editors' introduction. *Mach Learn* 1997;28:5-5.
11. Li S, Zhao Z, Hu R, Li W, Liu T, Du X. Analogical reasoning on chinese morphological and semantic relations. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics; 2018. p. 138-43.
12. Qiu Y, Li H, Li S, Jiang Y, Hu R, Yang L. Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings. In: *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Nanjing, China: Springer; 2018. p. 209-21.
13. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding; 2018. Available from: <https://arxiv.xilesou.top/abs/1810.04805v2>. [Last accessed on 2019 Nov 08].
14. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, *et al.* RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*, 2019. Available from: <http://arxiv.org/abs/1907.11692>. [Last accessed on 2019 Dec 22].
15. Wei J, Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv:1901.11196*, 2019. Available from: <http://arxiv.org/abs/1901.11196>. [Last accessed on 2019 Nov 08].
16. Li X, Meng Y, Sun X, Han Q, Yuan A, Li J. Is Word Segmentation Necessary for Deep Learning of Chinese Representations? *arXiv:1905.05526*, 2019. Available from: <https://arxiv.xilesou.top/abs/1905.05526v2>. [Last accessed on 2019 Nov 08].
17. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* 1995;14:1137-45.
18. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. Available from: <http://arxiv.org/abs/1412.6980>. [Last accessed on 2019 Nov 08].
19. Masters D, Luschi C. Revisiting small batch training for deep neural networks. *arXiv:1804.07612*, 2018. Available from: <http://arxiv.org/abs/1804.07612>. [Last accessed on 2019 Nov 08].
20. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Massachusetts: MIT Press; 2016.
21. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans Neural Netw* 1994;5:157-66.
22. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;9:1735-80.
23. Gers FA, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM. *Neural computation*. 2000;12:2451-71. [doi: 10.1049/cp:19991218].
24. Gal Y. *Uncertainty in Deep Learning*. Cambridge: PhD Thesis, University of Cambridge; 2016.
25. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, *et al.* Attention is all you need. In *Advances in neural information processing systems* 2017:5998-6008.
26. Google. Google AI Research; 2019. Available from: <https://github.com/google-research/bert>. [Last accessed on 2021 Jul 20].
27. Cui Y, Che W, Liu T, Qin B, Yang Z, Wang S, *et al.* Pre-training with whole word masking for Chinese BERT. *arXiv:1906.08101*, 2019. Available from: <http://arxiv.org/abs/1906.08101>. [Last accessed on 2019 Dec 22].
28. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. *Science* 2000;290:2319-23.
29. Cox TF, Cox MA. *Multidimensional Scaling*. New York: Chapman and Hall/CRC; 2000.
30. Liu W, Zhou P, Zhao Z, Wang Z, Ju Q, Deng H, *et al.* K-BERT: Enabling Language Representation with Knowledge Graph; 2019. Available from: <https://arxiv.xilesou.top/abs/1909.07606v1>. [Last accessed on 2020 Jan 03].