IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 August 2023, accepted 13 September 2023, date of publication 18 September 2023,
date of current version 27 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3316219

RESEARCH ARTICLE

PreGenerator: TCM Prescription Recommendation Model Based on Retrieval and Generation Method

ZIJUAN ZHAO[1], XUETING REN[1], KAI SONG[2], YAN QIANG[1], JUANJUAN ZHAO[1,3], JUNLONG ZHANG[4], AND PENG HAN[5]

[1]College of Information and Computer, Taiyuan University of Technology, Taiyuan 030600, China
[2]College of Physics, Taiyuan University of Technology, Taiyuan 030600, China
[3]School of Information Engineering, Jinzhong College of Information, Jinzhong 030800, China
[4]School of Basic Medicine, Shanxi University of Traditional Chinese Medicine, Taiyuan, Shanxi 030600, China
[5]North Automatic Control Technology Institute, Taiyuan 030006, China

Corresponding author: Yan Qiang (qiangyan@tyut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61972274, in part by the Major Project of National Natural Science Foundation of China under Grant U21A20469, and in part by the Natural Science Foundation of Shanxi Province under Grant 202103021224066.

ABSTRACT The generation of Traditional Chinese Medicine (TCM) prescription is one of the most challenging tasks in the research of intelligent TCM. Current researches usually use transfer learning methods to apply the relevant technology of text generation to this task simply and roughly. Either they need to train a model with large number of standardized data set, or they ignore the domain knowledge and expertise of TCM. In order to solve these problems, we propose a hybrid neural network architecture for TCM prescription generation—*PreGenerator*. It includes a novel hierarchical retrieval mechanism, which can automatically extract prescription and herbal templates to facilitate accurate clinical prescription generation. Firstly, *PreGenerator* uses the Symptom-Prescription Retrieval (SHR) module to retrieve the most relevant prescriptions for a given patient's symptoms. In order to follow the rule of compatibility of herbs, the Herb-Herb Retrieval (HHR) module is introduced to retrieve the next most relevant herb according to the conditioned generated herbs. Finally, the prescription decoder (PreD) fuses the symptom features, the retrieved prescription and herbal template features to generate the most relevant and effective Chinese medicine prescription. The validity of the model is verified by automatic evaluation and manual evaluation on the real medical case dataset. In addition, our model can recommend herbs that do not appear on the prescription label but are useful for relieving symptoms, which shows that our model can learn some interactions between herbs and symptoms. This research also lays a foundation for the future research on intelligent query and prescription generation of traditional Chinese medicine.

INDEX TERMS Text generation, traditional Chinese medicine, prescription, intelligent Chinese medicine, herb retrieval.

I. INTRODUCTION

As an ancient medical practice system that differs from modern medicine in substance, methodology, and philosophy, Traditional Chinese Medicine (TCM) has played an indispensable role in the healthcare of the Chinese people for thousands of years and is now gaining recognition and utilization in Western countries. Through extensive research, TCM has made significant contributions to modern medicine, exemplified by the discovery of artemisinin extracted from Artemisia annua and the establishment of a wormwood drug database [1].

In TCM, a prescription is a group of herbs (mineral and animal medicines are also used. Here, we will use the word

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

FIGURE 1. An example of TCM prescription, including prescription name, herbs composition, usage and indication symptoms.

"herbs" to refer to the herbs in the prescription) that have been the main way of treating diseases for thousands of years. In the long history of China, people have invented many prescriptions to treat diseases, and more than 100,000 kinds of prescriptions have been recorded [2]. Fig. 1 shows an example of a Chinese medicine prescription in the Traditional Chinese Dictionary, including the prescription name, herbs, usage, and indication symptoms. As can be seen from Fig. 1, the production of TCM prescriptions is a complex process. In this process, doctors need to carefully select herbs and make reasonable combinations according to patients' specific symptoms and physique, so as to achieve the therapeutic effect. Accurate treatment based on syndrome differentiation often rely on rich clinical experience and theoretical knowledge of TCM. The quality of TCM prescriptions directly determines the therapeutic effect of TCM. However, due to the inherent ambiguity in TCM, each doctor has room to develop their own experiences. When facing the same set of symptoms, different doctors may induce different patterns and prescriptions due to the lack of standardized diagnostic criteria within the industry. Additionally, the vast variety of herbal medicines, along with their complex properties and compatibility rules, make it challenging to fully explore and comprehend this knowledge using traditional manual methods. As a result, the rational application of herbal medicines and the development of innovative herbal formulations are limited.

With the rise of artificial intelligence (AI) technology, the introduction of machine learning and deep learning has provided new avenues to address the aforementioned issues. Researchers are exploring the use of artificial intelligence to automatically generate herbal prescriptions from symptom descriptions, mine and analyze the correlations and compatibility patterns among different herbs, and uncover hidden patterns and rules within massive datasets. This approach aims to guide the rational combination of herbs and facilitate the design of innovative herbal formulations.

Existing Chinese herbal prescription recommendation models can generally be classified into two categories: topic models and non-topic models. Topic models [3], [4], [5] are commonly used methods for studying the patterns of Chinese herbal prescriptions. They discover underlying topics by analyzing the distribution of symptoms and herbs, which are often considered as the syndromes of TCM theory. Besides, prior knowledge of specific elements can be integrated into the theme model [6] to improve the efficiency of prescription pattern recognition. Because of these advantages, the topic model can not only recommend appropriate herbs, but also explore their indications and compatibility. However, there are more than 1000 commonly used herbs, and there are always higher-order herb or symptom correlations, so it is difficult to train the theme model. This makes them inadequately used in drug prediction tasks. Non-topic models [7] typically regard prescription generation as a multi-label classification problem, where each herb is treated as a class label, and the goal is to predict the probabilities of each herb being included in the prescription. These models focus on modeling binary relationships between sets of symptoms and sets of herbs, utilizing graph convolutional networks and other techniques to explore the dependencies among multiple symptoms and herbs [8], [9], [10]. However, due to the vast number of herbal species and the imbalanced distribution of herbal data, this presents a significant challenge for the classifier, impacting the model's robustness and generalization capabilities.

Inspired by the success of natural language generation tasks such as neural machine translation and text summarization, some researchers regarded the task of generating TCM prescriptions as a translation task from symptoms to herbs. The sequence to sequence (seq2seq) model was used to generate traditional Chinese medicine prescriptions [11], [12], [13]. The seq2seq model includes an encoder that maps symptoms to latent spaces and a decoder that generates herbs. In addition, doctors usually pay special attention to one or some primary symptoms when prescribing, rather than paying equal attention to all symptoms. Therefore, attention mechanism was introduced to focus on different symptoms in each step of prescription generation, which is also widely used in seq2seq tasks [14], [15], [16].

However, in the above researches, the experimental results showed that the number of clinical medical records should be at least more than 6000 to achieve good performance. As we all know, the medical records used for training models are clinical records of high-level medical experts in the field of traditional Chinese medicine. The number of structured and standardized medical records from these medical experts is

relatively small, and the number of samples subdivided into specific diseases is much smaller. The number of samples of complex deep learning models in experiments is far from enough. Therefore, the model trained by these samples is difficult to be applied in the clinical environment. Moreover, the generation of TCM text is limited by professional format, grammar, compatibility of traditional Chinese medicine, etc. It is hard for conventional generation methods to learn some specific rules, which brings great challenges to the automatic generation of traditional Chinese medicine. The above problems seriously restrict the development of intelligent traditional Chinese medicine, especially the generation of TCM prescriptions.

In order to solve the above problems, the method of Chinese text based on retrieval came to us [14], [15], generating standardized reports from the predefined retrieval database, which is somewhat similar to the process of prescribing in real life. TCM physicians memorize some traditional prescription templates, and based on this, they add, subtract, and combine prescriptions according to the patient's key information extracted in the four diagnosis, so as to generate new prescriptions that are consistent with the current patient.

Inspired by the above ideas, in this study, we propose a Seq2Seq learning model based on retrieval to generate Chinese medicine prescriptions, called *PreGenerator*. Specifically, we use three modules to simulate this process. Firstly, the *PreGenerator* uses the patient's symptom features as a query to generate a prescription level feature template from the Classical Prescription Retrieval pool, called Symptom-Prescription Retrieval (SPR) module. For the more accurate prescriptions generation, *PreGenerator* also extracts key herbs from the prescriptions we retrieved. At the same time, we propose a multi-query attention mechanism to learn prescription level template representation. Secondly, in order to make the generated prescriptions more consistent with the compatibility rules of TCM prescriptions, we propose Herb-Herb Retrieval (HHR) module, which aims to learn the herbal level template required for the next generation of herbal medicine by analyzing the correlation between the herbs in the retrieval prescriptions. Finally, the Prescription decoder (PreD) is used to generate reasonable prescriptions using symptom features, prescription level and herbal level template features. The coverage mechanism adopted in this paper can ensure the generation of non repetitive herbs, and the two-level memory retrieval mechanism designed is helpful to generate accurate and diverse TCM prescriptions. To sum up, the main contributions of this paper are as follows:

(1) As far as we know, this study is the first to apply memory retrieval mechanism to the task of generating TCM prescriptions. By simulating the generation of standardized Chinese herbal medicine in real life, our memory retrieval mechanism effectively uses the existing hierarchical templates in the text of TCM prescriptions. This design enables *PreGenerator* to generate more accurate and reasonable TCM prescriptions.

(2) Based on the retrieval module, a new multi-query attention mechanism is proposed to integrate the retrieved information into the prescription generation. The fused information can well fuse the existing symptoms, herbs information, thus improving the quality of prescription generation.

(3) Experiments on our TCM records dataset show that *PreGenerator* has achieved better performance compared with the latest baselines such as LSTM seq2seq and Herb-Know. The case study shows that the Chinese prescriptions generated by *PreGenerator* are more accurate and reasonable, and conform to the compatibility specifications of Chinese medicine through the evaluation of experts in the field.

II. RELATED WORK

A. MEDICAL TEXT GENERATION

As a sub task in the medical field, medical text generation mainly focuses on the generation from medical image to reports. At present, the automatic generation model of image report mainly draws on the Encoder-Detector framework in the field of machine translation, which uses the Convolutional Neural Network(CNN) to extract image features, and then uses the Recurrent Neural Network (RNN) to generate text descriptions of images. For example, CNN-RNN [17], LRCN [18] and AdaAtt [19] are applied to medical report generation tasks. In order to further improve the generation of long text with domain specific knowledge, the descendant based method introduced layered Long Shot Term Memory(LSTM) with common attention [20] or medical concept features [21] to guide report generation. On the other hand, the concept of reinforcement learning [22] was used to ensure that the generated radiology report correctly describes the clinical results. To avoid generating clinically non-informative reports, external domain knowledge like knowledge graphs and anchor words [23] are utilized to promote the medical values of diagnostic reports. Clara [24] also provided an interactive solution to integrate doctors' judgments into the generation process. This paper applies this idea to the automatic generation of TCM prescriptions, in which the transformer model is used to encode symptom features, and then LSTM decoding is used to generate TCM prescriptions.

B. RETRIEVAL-BASED TEXT GENERATION

The retrieval-based method is usually combined with the generation-based method to improve the rationality and readability of the generated text. The retrieval module is used to return equivalent documents to explicitly introduce external knowledge, while the generation module is used to generate the target sequence. For example, Lewis et al. [25] explored a general fine-tuning method for the Retrieval-Augmented Generation(RAG) model, which combines pre-trained parameters and nonparametric memory for language generation. Izacard et al. [26] studied a simple open domain question answering method. The retrieval module tried to

retrieve similar documents in different ways, and then the title returned by the retrieval module and the corresponding document were spliced together through special characters and input to the generation module to generate the representation of all retrieval documents, and then all document representations were spliced together and processed with the generation model. Zhang et al. [27] alleviated data constraints by jointly training label generators and document retrievers, retrieved the most effective documents in the generation process for rewards, and used expert hybrid integration to collectively combine them to generate subsequent texts. In the field of medical reporting, Li et al. [28] used abnormality graphs to retrieve most related sentence templates during the generation. HRGR Agent [29] combines the retrieved sentences into an enhanced learning framework for generating medical reports. However, they all require a template database as model input. Unlike these models, *PreGenerator* can automatically learn prescription level and herbal level templates from data, which greatly enhances the applicability of the model.

III. METHODS

As shown in Fig. 2, we propose a new framework called *PreGenerator*, which consists of three modules. The Symptom-Prescription Retrieval (SPR) module works on the prescription level and uses symptom features to find the most relevant template prescriptions based on a multi-view TCM diagnostic records. The Herb-Herb Retrieval (HHR) module works on the herb level and retrieves a series of candidate herbs that are most likely to be the next herb from the retrieval pool. Finally, the Prescription decoder (PreD) is used to generate accurate, diverse, and syndrome-specified TCM prescription using symptom features, prescription level and herbal level template features. The coverage mechanism adopted in this paper can ensure the generation of non repetitive herbs, and the two-level memory retrieval mechanism designed is helpful to generate accurate and diverse TCM prescriptions. To improve the effectiveness and efficiency of retrieval, we also first pretrain SPR and HHR modules to build up a retrieval pool for TCM prescription generation.

A. PROBLEM DEFINITION

The task of TCM prescriptions generation requires that the model mines the patient's TCM symptom types based on the patient's symptoms and other pathological information, output a group of Chinese herbs, and formulate a TCM prescription that can effectively alleviate the given symptoms. Considering the facts of TCM clinical records, we regard the patient's symptoms and medical history (pathological information) as a symptom sequence $s = \{s_1, s_2, \ldots, s_n\}$, each $\boldsymbol{s}_i$ represents a symptom. TCM prescription is composed of a series of medicinal herbs h. The model proposed in this paper mainly includes three core modules: Symptom-Prescription Retrieval (SPR) Module, Herb-Herb Retrieval (HHR) Module and Prescription Generation (PreD) Module. The SPR module mainly uses symptom features to match the most relevant prescription template. HHR module mainly searches a series of candidate sets of the next most likely herbs in the retrieval pool based on the previously generated herbs. According to the symptoms, herbs and compatibility features obtained by SPR and HHR modules, the PreD module can generate accurate TCM prescription related to symptoms. Let $s = \{s_1, s_2, \ldots, s_n\}$ represents the symptom sequence text in the TCM record, and $\boldsymbol{s}_i$ represents the symptom feature in the patient's text, such as ''fever, headache, aversion to cold'', which is expressed by a word embedding. $p = \{p_1, p_2, \ldots, p_k\}$ represents the candidate prescription set retrieved by the SPR module. Therefore, the task of this paper is to generate a collection of herbs $H = \{h_1, h_2, \ldots, h_T\}$ based on S and P, so as to form a prescription Y highly related to disease symptoms.

B. SYMPTOM-PRESCRIPTION RETRIEVAL MODULE PRETRAINING

The SPR module aims to retrieve the most relevant TCM prescriptions for a given symptom description from the training prescription pool. The retrieved prescriptions will be further used to learn an abstract template to generate new high-quality prescriptions. To this end, we introduced a self-supervised pretraining task to judge whether the symptom prescription pair came from the same medical record report, that is, the symptom-prescription matching. It is based on an intuitive assumption that the symptom prescription pairs from the same patient's medical record share some common semantics. This assumption is consistent with the diagnosis and treatment of TCM based on syndrome differentiation, that is, symptoms and prescription should correspond to the same type of syndrome. Therefore, in the pretraining task, we also considered the TCM syndrome.

1) SYMPTOM-PRESCRIPTION ENCODER

The input of SPR module is a series of symptom sets and corresponding prescription pairs issued by doctors *(s, p)*, where $s = \{s_i\}_{i=1}^{N}$ means that the symptom text contains N symptom representations, and $p = \{h_i\}_{i=1}^{T}$ represents there contains T herbs in a prescription. In this paper, we use Transformer [30] model as the encoder of symptom and prescription text pairs, which is composed of n layers of stacks. Each layer mainly consists of two parts: multi-head self attention layer and position full connection feedforward network layer. We first put the symptom text into the template layer and the position embedding layer. Then, self attention layer linearly transforms the input into query vector Q, key vector K and value vector V, and calculates the attention weight and context vector. These vectors reflect the correlation within the symptom sentence:

$$Attention(Q, K, V) = \boldsymbol{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

where d_k indicates the dimension of a matrix *Q, K, V*.Multi-heads refer to repeating self attention h times in different representation subspaces, and then concatenate the results of

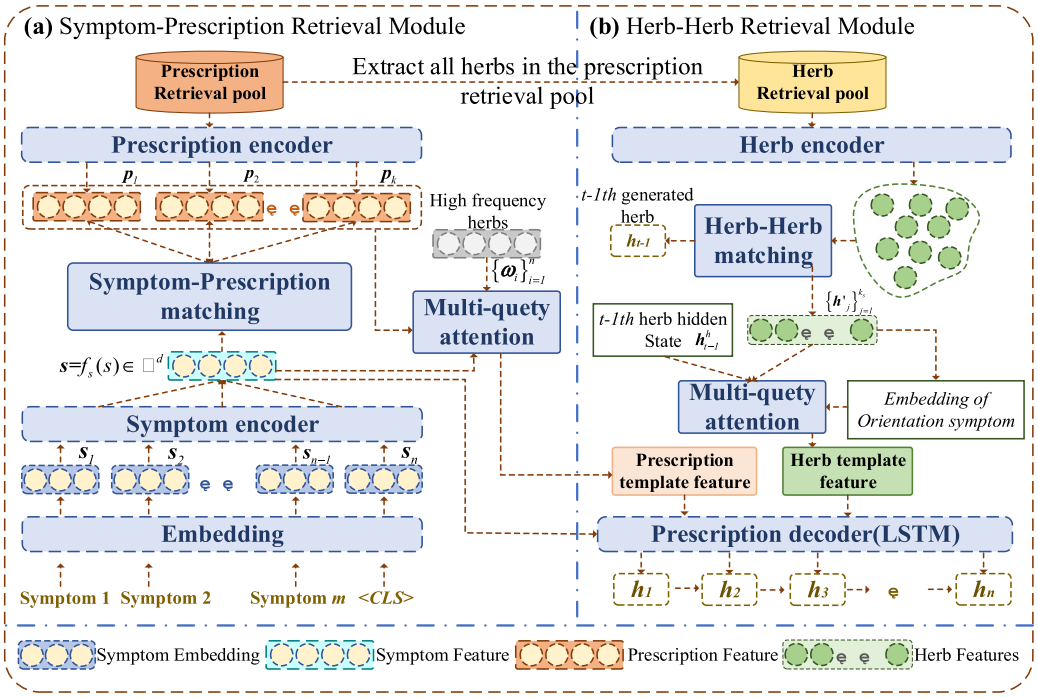

FIGURE 2. Structure diagram of PreGenerator model for TCM prescription generation. The left part learns the prescription template representation through the Symptom-Prescription Retrieval (SPR) module. The right part shows the details of the herb-Herb retrieval (HHR) module and the Prescription Decoder(PreD) used to generate the herbs one by one.

each head, described by the following formula:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \text{head}_h)W^O$$
$$\text{Where, head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

In the above formula, W_i^Q, W_i^K, W_i^V and W^O are learnable parameter matrixes, h refers to the number of Multi-head attention.

The output of multi-head attention is operated through a residual connection and layer normalization:

$$FFN(x') = \max(0, xW_1 + b_1)W_2 + b_2 \quad (3)$$

where, W_1, W_2, b_1 are b_2 learnable parameters. x' indicates the output value of the multi-head attention module.

After the transformer encoder, we obtain the context aware hidden state of each tag of the symptom text.

2) SYMPTOM-PRESCRIPTION MATCHING

The next training task of SPR is to judge whether the symptom prescription pair belongs to the same medical record report. In this subtask, the input symptom-prescription pair *(s, p)* is embedded as a feature vector pair ($\boldsymbol{s}$, $\boldsymbol{p}$), such as, $\boldsymbol{s} = f_s(s) \in \mathbb{R}^d, \boldsymbol{p} = f_s(p) \in R^d$. Then, the probability of input pair $\left(\{s_i\}_i^N, \{h_j\}_j^T\right)$ belonging to the same patient can be calculated as:

$$prob_{sp} = \text{sigmoid}(\boldsymbol{s}^T, \boldsymbol{p}) \quad (4)$$

Given the probability $prob_{sp}$ and ground-truth label of whether the input symptom-prescription pair belongs to the same of the TCM medical record, the cross entropy loss is used to optimize the learning objective.

C. HERB-HERB RETRIEVAL MODULE PRETRAINING

Traditional Chinese medicine prescriptions usually correspond to the symptoms of patients from primary to secondary, that is, the principle of "monarch, minister, assistant and envoy", which means that different herbs play different roles in a prescription. In addition, different Chinese herbal medicines have their own natures, tastes and meridians. There are different compatibility laws among herbs, such as single line, mutual necessity, mutual use, mutual fear and mutual killing. Automatic learning of these features and laws should help *PreGenerator* generate high-quality TCM prescriptions. For this reason, we suggest to pretrain the HHR module in advance to generate the most appropriate herb for the next step. In particular, we introduced a self supervised pretraining task for HHR to determine whether the two medicinal materials come from the same prescription, that is, herb-herb matching.

Similar to SPR module, we use the Transformer model $f_h(\cdot)$ as the herbal encoder, embedding the input herb pair $\{h_i, h_j\}$ into the herbal token vector $\{\boldsymbol{h}_i, \boldsymbol{h}_j\}$, and then to calculate the probability that the two herbs come from the same prescription:

$$prob_{hh} = sigmoid(\boldsymbol{h}_i^T \cdot \boldsymbol{h}_j) \quad (5)$$

Similarly, cross entropy loss is used again to optimize the given probability of learning objectives. The ground truth label depends on whether the input herbal pair $\{h_i, h_j\}$ comes from the same prescription. If $\{h_i, h_j\}$ comes from the same TCM prescription, the probability label is 1, otherwise it is 0.

D. RETRIEVAL-BASED PRESCRIPTION GENERATION

Using the pretrained SPR and HHR modules, *PreGenerator* uses the novel proposed hierarchical retrieval module and prescription decoder according to the given symptom sequence $\{s_i\}_{i=1}^{N}$ to generate the most relevant TCM prescriptions.

1) PRESCRIPTION-LEVEL RETRIEVAL IN SPR MODULE

a: PRESCRIPTION-LEVEL RETRIEVAL

Let $\mathcal{D}_p^{(tp)} = \{p_j\}_{j=1}^{N_{tp}}$ denote the set of all the training prescriptions, where N_{tp} is the number of prescriptions in the training dataset. For each prescription p_j, we first obtain its vector representation using $f_s(\cdot)$ in the SPR module, which is defined as $\boldsymbol{p}_j = f_s(p_j)$. Let $\mathcal{P}_h = \{\boldsymbol{p}_j\}_{j=1}^{N_{tp}}$ denote the set of training prescription representations. Given patient symptom representation $\{\boldsymbol{s}_j\}_{j=1}^{N}$. The SPR module aims to return top k TCM prescription samples $\{p_j^{'}\}_{j=1}^{k_r}$ and top l key Chinese herbs appeared in the top k prescriptions.

In particular, *PreGenerator* uses the encoder mentioned in the previous section to extract the symptom text hidden feature $\boldsymbol{S}$. And then according to formula (4), to calculate matching score of symptoms $\boldsymbol{S}$ and every prescription $\boldsymbol{p} \in \mathcal{P}_h$. In this way, the top k_p prescriptions $\{p_j'\}_{j=1}^{k_p}$ with the highest matching score are considered to be the most relevant treatment prescription templates for input symptoms. From these prescription templates, we can count the most important n key Chinese herbs $\{\omega_i\}_{i=1}^{n}$ according to their frequency of occurrence.

b: PRESCRIPTION TEMPLATE REPRESENTATION LEARNING

The retrieved prescriptions are highly correlated with the symptoms of a given patient, which will certainly help to generate the correct prescriptions. In order to make full use of them, we need to use symptom embedding vector $\boldsymbol{S}$, retrieved prescription feature representation $\left\{\boldsymbol{p}_j^{'}\right\}_{j=1}^{k_p}$, and the embedding representation $\{\boldsymbol{\omega}_i\}_{i=1}^{n}$ of key Chinese herbs $\{\omega_i\}_{i=1}^{n}$ to learn a prescription template representation.

In this paper, we propose a new multi-query attention mechanism to learn prescription template representation. Specifically, we utilize the symptom feature $\boldsymbol{S}$ as the key vector K, the retrieved prescription representation $\left\{\boldsymbol{p}_j^{'}\right\}_{j=1}^{k_p}$ as the value vector $\boldsymbol{V}$, and the words embedding of key herbal tokens as the query vector $\boldsymbol{Q}$. Here, we improve the original self attention [30] to multi-query attention. For each query vector Q_i, the corresponding attention features are obtained at first, and after concatenating every attention feature, we get the transformed prescription template vector $\boldsymbol{p}_s$.

$$\begin{aligned}\boldsymbol{p}_s &= \text{MultiQuery}(\{Q_i\}_{i=1}^{n}, K, V) \\ &= \text{concat}(\text{attn}_1, \cdots, \text{attn}_n)W^O \end{aligned} \tag{6}$$

where, $\text{attn}_i = \text{Attention}(Q_i, KW^K, VW^V)$, W^K, W^V are W^O transfer matrixes. In general, attention function can be caculated in the following way:

$$Attention(Q_g, K_g, V_g) = softmax(\frac{Q_g K_g^T}{\sqrt{d_g}})V_g \tag{7}$$

In the formula, *Q, K, V* indicate query vector, key vector and value vector respectively. d_g represents the dimension of query vector.

2) HERB-LEVEL RETRIEVAL IN HHR MODULE

As the retrieved prescriptions $\{p_j'\}_{j=1}^{k_r}$ are highly related to the symptoms of the imported patient, the herbs in these prescriptions must contain some beneficial pathology and syndrome differentiation rules, which are conducive to the generation of herbal medicine level. To do this, we first select herbal medicine from the retrieved prescriptions, and then learn herbal template representation.

a: HERB-LEVEL RETRIEVAL

First of all, all the herbs $\{h_j\}_{j=1}^{L}$ appeared in the retrieved classical prescriptions are collected in the herbal retrieval pool of the HHR module. Then, the pretrained encoder $f_h(\cdot)$ of HHR module is used to get the herbal level feature pool $H = \{f_h(h_j)\}_{j=1}^{L} = \{\boldsymbol{h}_j\}_{j=1}^{L}$. Suppose that the herb generated at time t is recorded as o^t, its embedding vector is $\boldsymbol{o_t} = f_h(o_t)$, which is used to select k_h herbs with the highest probability $\left\{h_j^{'}\right\}_{j=1}^{k_h}$ from the candidate herb pool. The query method is shown in Section 2.3.

b: HERB TEMPLATE REPRESENTATION LEARNING

Similar to template prescription representation learning, the multi-query attention mechanism is still used in the representation learning of herbal template medicine. From the retrieved top k_h herbs, the top k_s symptomatic symptoms $\left\{s_j^{'}\right\}_{j=1}^{k_s}$, which are encoded by symptom encoder $f_s(\cdot)$ to get main symptoms features $\left\{\boldsymbol{s}_j^{'}\right\}_{j=1}^{k_s} = \left\{f_s(s_j^{'})\right\}_{j=1}^{k_s}$. Here, we let input symptom embedding $\left\{\boldsymbol{s}_j^{'}\right\}_{j=1}^{k_s}$ as query vector, key herb embeddings $\left\{\boldsymbol{h}_j^{'}\right\}_{j=1}^{k_h}$ as value vector. And the hidden state of the last time $\boldsymbol{h_{t-1}^{dec}}$ are used as key vector, which will be illustrated in section 2.4.3. According to the following formula, we can get the herbal template representation at time t, using the symbol $\boldsymbol{u_t}$ represents.

$$\boldsymbol{u}_t = MultiQuery(\left\{\boldsymbol{s}_j^{'}\right\}_{j=1}^{k_s}, \boldsymbol{h}_{t-1}^{dec}, \left\{\boldsymbol{h}_j^{'}\right\}_{j=1}^{k_h}) \tag{8}$$

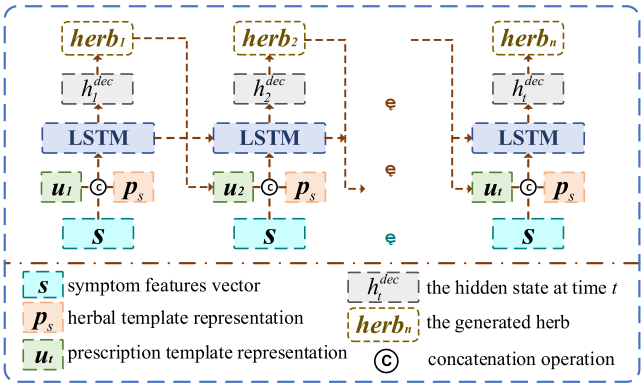

FIGURE 3. The architecture of prescription decoder in PreGenerator.

3) PRESCRIPTION DECODER

With features extracted from the above retrieval mechanism used, we apply the hierarchical prescription decoder to generate TCM prescriptions according to the matching rule of "monarch, minister, assistant and envoy" in Chinese medicine prescriptions. The decoder consists of two layers, the LSTM decoder for outputting the hidden state of herbal medicine and the LSTM decoder for decoding the hidden state of herbal medicine into natural herbs. In this way, prescriptions can be generated in a certain herbal order. (The decoding generation process is shown in the Fig. 3)

To generate the *t-th* herb, *PreGenerator* uses the herb-level hidden state $\boldsymbol{h}_{t-1}^{dec}$ at the previous moment, input symptom features $\boldsymbol{s}$, prescription template representation $\boldsymbol{p}_s$ and current herbal template representation $\boldsymbol{u}_t$ to learn the hidden state $\boldsymbol{h}_{t^{dec}}$ at the current time together. Specifically, *PreGenerator* uses formula(3) to learn the symptom features $\boldsymbol{s}$. Then, multi-query attention is utilized to learn herbal template representation $\boldsymbol{u}_t$ (Formula (8)). Finally, the symptom features $\boldsymbol{s}$, prescription templates representation $\boldsymbol{p}_s$ and herbal template representation $\boldsymbol{u}_t$ are concatenated to input LSTM to generate herb at moment *i+1*. The formula is as follows:

$$\boldsymbol{h}_t^{dec} = \text{LSTM}_h(\text{concat}(\boldsymbol{s}, \boldsymbol{u}_t^{dec}, \boldsymbol{p}_s), \boldsymbol{h}_{t-1}^{dec}) \tag{9}$$

$$herb_{t+1} = argmax(\text{softmax}(FFN(h_t^{dec}))) \tag{10}$$

where $FFN(\cdot)$ represents feedforward neural network. For the generation of the first herb, we set $\boldsymbol{u_0}$ to 0, $\boldsymbol{h}_0$ is a randomly initialized vector.

a: COVERAGE MECHANISM

Different from other natural language generation task, there are no duplicate herbs in TCM generation task. When the *PreGenerator* model is directly applied in this task, the decoder often generates some frequently observed herbs repeatedly. Although we can remove duplicate herbs through post-processing to trim them, the maximum length of the prescription is limited, which will still affect the recall performance. Therefore, this paper uses an<EOS>flag to indicate where the generation should stop.

In order to encourage the decoder to generate more diversified and reasonable herbal tokens, this paper introduces a coverage mechanism to enable the model to perceive the generated herbs. The coverage mechanism is mainly used to provide a fertility vector to indicate how much input information has been used in the attention calculation process, so as to help the decoder focus more on the parts that are not concerned.

In the *PreGenerator* model in this article, we do not use fertility vector to adjust the attention weight of the symptom encoder. The reason is that the symptoms are interrelated and together describe the whole disease. Even so, inspired by its motivation, we apply the coverage mechanism to the decoder, where the coverage vector is input to the LSTM unit together with the vector. Then replace formula (8) with the following:

$$\begin{aligned} a_t &= \tanh(WD_t + b) \\ \boldsymbol{h}_t^h &= \text{LSTM}_h(\text{concat}(\boldsymbol{x}^h, \boldsymbol{u}_t^h, \boldsymbol{p}_s), \boldsymbol{h}_{t-1}^h, a_t) \end{aligned} \tag{11}$$

where, a_t represents the coverage vector at time t during decoding. D_t represents the one-hot vector of the generated herbal medicine up to time t. $W \in \mathbb{R}^{V \times H}$ represents learnable parameter matrix, V represents the dimension of Chinese herbal medicine vocabulary, and H represents the dimension of hidden state. By providing the coverage vector (that is, the sketch of the generated herbal medicine) as part of the input to the LSTM decoder in this paper, our model can give more probability to the unpredictable herb gently. This will encourage the model to generate new herbs instead of repeatedly predicting frequently occurring herbs, thus increasing the recall rate.

b: SOFT LOSS FUNCTION

Generally, cross entropy loss is applied to optimize the network in Seq2Seq models, which will have strict restrictions on the order of generating herbs. However, in the process of generating prescriptions, we do not need to strictly follow the order of labels when generating herbal tokens one by one. Therefore, this paper uses soft cross entropy [8] to calculate the loss, and the formula is as follows:

$$L = -\sum_t \hat{q}_t \log(p_t) \tag{12}$$

where, p_t represents the probability distribution generated by *PreGenerator* at time t, $\hat{q}_t$ represents the smoothed target probability distribution, which is calculated based on the original one-hot label q_t at time t and multi-hot of target herbs label q_v.

$$\hat{q}_t = \frac{1}{2}(\frac{q_v}{M} + q_t) \tag{13}$$

where M represents the number of herbs on the prescription label.

症状	草药
小儿惊热，心烦不得睡卧。	龙脑，麝香，甘草，牛蒡子，栀子仁，牛黄，马牙消，郁金
肝脏壅热，两眼赤痛	龙脑，栀子仁，黄芩，麦门冬，地骨皮，川升麻，犀角屑，牛黄，川大黄，甘草
中寒中暑，感冒秽浊，昏迷气闭，四肢厥冷，呕吐恶心，腹痛作泄。	苍术，橘皮，五加皮，厚朴，阔羊花，茯苓，槟榔，冰片，百草霜，猪牙皂，藿香，灯心，雄黄粉，朱砂粉，细辛，麝香，牛黄

FIGURE 4. Example of specification of data.

In general, when the model generates herbs in the wrong order, the loss can be reduced by adding the information of the overall target label. Our model aims to maximize the conditional probability of herbal prescription by training the encoder and decoder.

IV. EXPERIMENTS

A. DATASET COLLECTION

The model proposed in this paper needs two datasets: (1) TCM clinical medical record dataset. (2) Pretraining symptom-prescription pair and herb-herb pair datasets.

The dataset used in this paper is from the Chinese herbal medicine professional knowledge service system (http://zcy.ckcest.cn/tcm/). A total of 2634 cases of traditional prescription data were extracted, and each case included the patient's gender, age, symptoms, syndrome types and corresponding prescriptions. In this study, due to the uniqueness of the prescription generation task, we focus more on the correspondence between symptoms and herbs. Based on the above conditions, we have cleaned and structured this datasets. For this dataset, we first use regularization to filter out the symptom and herbal composition, subsequently removing the quantity words from the herbs, retaining only the names. And we unify the different names of the same herb. As shown in the Fig. 4, one herb corresponds to only one name, e.g., Borneolum Syntheticum, borneol, etc., which we collectively refer to as borneol. Finally, we conducted a statistical analysis of the combinations of symptoms and herbs present in 2,643 sets of prescription data. We counted the total number of symptoms and herbs included in these combinations, as well as the number of symptoms and herbs present in each individual sample. After prosessing, there are 186 different symptoms and 204 different herbs in the dataset. Each patient has a different number of symptoms (minimum 1, maximum 20) and herbs (minimum 1, maximum 21).

For the pretraining stage, we randomly match the symptoms and prescriptions of different patients, and obtain the SOP label at the same time. If the symptoms and prescription belong to the same patient, the label is 1, otherwise it is 0. 5349 symptoms-prescription matching pairs are extracted for the pretraining of the symptom-prescription retrieval module. Then, all the herbs in the prescription are extracted to form a herbs retrieval pool, we also match herb-herb pairs randomly from herbs retrieval pool. If two kinds of herbs appear in the same prescription, the label is 1; otherwise, the label is 0. Finally, 8462 pairs of herb-herb pairs are randomly generated. The dataset of herb pairs is used for the pretraining of HHR module. Table 1 shows the details of the pre-training data pairs.

TABLE 1. Pre-training data statistics in retrieves modules.

	symptoms-prescription pairs	herb-herb pairs
Matched	2643	3924
Unmatched	2706	4538

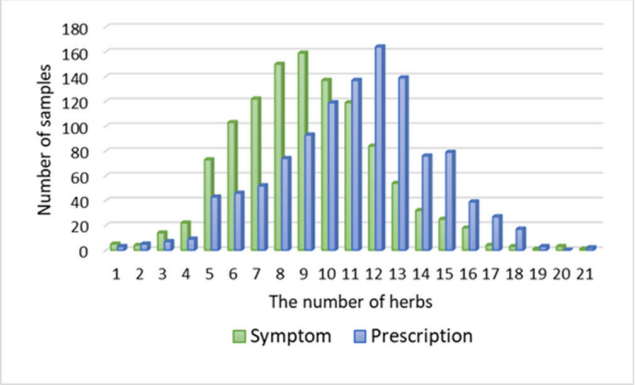

FIGURE 5. The length distribution of symptom sequence and herb sequence(TCM prescription).

B. EXPERIMENT SETTING

The input symptoms were first filled to the maximum length and embedded into the 512 dimension embedding vector. On this basis, we used 4 heads in the symptom encoder to calculate the multi-head attention, so that the model achieved the best performance. In the feedforward network, we set the dimensions to 1024 and 512 respectively. The hidden state size of the prescription LSTM decoder was also 512. Adam optimization method [31] was used to train the model, and the learning rate was $\alpha = 0.001$, momentum parameter $\beta_1 = 0.9, \beta_2 = 0.99$. In addition, we trained our model with 150 epoch, and the batch size was 25. In order to obtain fair performance comparison, 1134 medical records were randomly shuffled in each epoch, which were divided into training set (80%), verification set (10%) and test set (10%). we don't change the validation set during training. To effectively evaluate the model performance on a limited dataset, we use 10-folding cross-validation.

We have also calculated and displayed the length distribution of symptom sequence and herb sequence in each medical record. As shown in Fig. 5, the range of symptom sequence length is larger than that of herb sequence length. Most medical records contain less than 19 kinds of Chinese herbs or symptoms. This showed that it may be rare to see more than 19 kinds of symptoms or Chinese herbs in the clinical medical record. If these samples are included in the dataset,

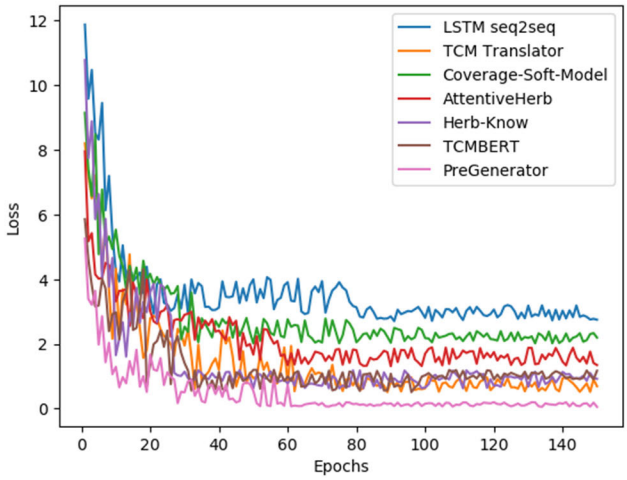

FIGURE 6. The training loss of all included models.

the model can easily be over fitted. Therefore, to avoid this problem, we removed them from the dataset.

C. BASELINES

We apply the following baseline models to validate our proposed approach. They come from the topic model and classic TCM prescription generation model:

1) LSTM seq2seq: LSTM sequence to sequence generation model. A basic framework for TCM prescription generation.

2) PTM: Prescription Topic Model, to explore the generation process of TCM prescriptions [3]. In this paper, the best parameter set is used for comparison. The number of topics is 20, and 5 herbs are recommended.

3) TCM Translator: A Seq2Seq Chinese herbal medicine translation model based on the combination of transformer encoder and LSTM decoder [12].

4) Coverage-Soft-Model: It can be regarded as LSTM seq2seq based on attention [11].

5) AttentiveHerb: LSTM seq2seq model based on two-stage attention mechanism [14].

6) Herb-Know: A seq2seq model based on transformer and enhanced by external knowledge of herbal medicine [15].

7) TCMBERT: A two-stage transfer learning model is used to generate TCM prescriptions from a small number of medical records and TCM literature resources [16].

D. EVALUATION INDICATORS

In this paper, the traditional Chinese medicine prescription generation task is regarded as a sequence generation task, rather than a classification generation task. Considering the principle of traditional Chinese medicine and the role of prescription generation, there is no appropriate evaluation standard for TCM prescriptions generated by artificial intelligence model. Because TCM prescriptions do not completely contain linguistic relations, but contain some compatibility and exclusion principles between different herbs, which is not reasonable to just count the number of herbs in the label prescriptions in that the same symptoms may be caused by different reasons. Therefore, it is necessary to find the real reason for using symptomatic Chinese herbs under the guidance of syndrome differentiation and treatment. The model can select different effective herbs according to the interaction between learned medicinal materials and symptoms (different TCM doctors may recommend different herbs for the same patient, but all prescription are provided under the guidance of syndrome differentiation). Calculating the number of herbs based on the labeled prescription alone does not reflect the diversity of useful herbs. For this reason, it is unreasonable to simply compare indicators commonly used in machine learning, such as BLEU [34], sensitivity, recall rate, accuracy, etc.

Only considering the similarity between the generated prescription and the labeled prescription, the model could not evaluate results reasonably roundly. However, similarity measurement can provide us with some analysis perspectives. For example, from the perspective of similarity, the higher the similarity between the generated prescription and the labeled prescription, the more effective the prescription generated by the model. Therefore, this paper used two strategies to evaluate all the prescriptions generated by models; Similarity evaluation and Human evaluation. Similarity evaluation provides objective evaluation results, while TCM practitioners provide more comprehensive and accurate subjective evaluation results. Then, with the help of these doctors, the results were statistically analyzed.

1) SIMILARITY EVALUATION

In terms of similarity evaluation indicators, this paper uses Hamming Loss (HL) to measure the misclassification instance and label prescription pair, and uses micro-F1(F1), Precision(P) and Recall(R) to evaluate the overall prediction performance under different views.

2) HUMAN EVALUATION

Since the automatic generation of TCM prescriptions is a very complex task, we invited two professors from Shanxi University of Traditional Chinese Medicine, which is one of the best TCM universities in Shanxi province. Both professors have more than five years of practical experience in Chinese medicine. All evaluators were asked to evaluate the generated prescriptions from the following two aspects: 1) Herbs Effectiveness (HE); 2) Herbs compatibility (HC). The value range of these two scores is [0.5]. The higher the score, the better the effectiveness and compatibility of the prescription, and vice versa. Doctors evaluated according to the theory, principle, prescription and their own experience of traditional Chinese medicine. In addition to the generated prescriptions, the invited doctors need to score the labeled prescriptions as the baseline for generating prescriptions. Their evaluation results are listed in Table 2. Pearson's correlation coefficient between the two evaluators was 0.72, and spearman's correlation coefficient was 0.79. P values were less than 0.01, indicating that they were highly consistent. Different from the

automatic evaluation method, the manual evaluator focuses on the potential efficacy of the candidate answers, not just the literal similarity. This evaluation method is more reasonable and closer to reality.

E. EXPERIMENTAL RESULTS AND ANALYSIS

1) OVERALL PERFORMANCE ANALYSIS

We compared the performance of *PreGenerator* and the baseline models on the test set. The objective comparison results are shown in Table 2. "−" indicates that the lower the numerical value, the better the model performance. "+" is the opposite. Compared with other baseline models, our method achieved the best performance and greatly exceeds the conventional Seq2Seq methods, such as LSTM seq2seq and TCM Translator. This shows that integrating information retrieval methods into the deep sequence generation framework can not only use the retrieved herbal information as a template to help generate prescriptions, but also overcome the monotony and repeatability of using only symptoms for generation. In addition, Herb-Know has slightly lower performance than the method in this article, introducing external knowledge into the calculation, which helps the model to understand label dependencies and background knowledge to some extent. Compared with the Seq2Seq method, the Prescription Topic Model has great flexibility and can include different data parts into the calculation, such as symptoms, herbs and prior knowledge, which makes their performance is higher growth and upside potential. Moreover, AttentiveHerb and TCMBERT also have excellent performance, which shows that attention mechanism and transfer learning strategy are much better than other methods in capturing symptoms and label relevance.

To sum up, the proposed *PreGenerator* has not only incorporated the correlation between symptoms and labels into the calculation, but also can properly match the automatic prescription generation task by using the information of prescriptions and herbal template. Beyond that, any valuable information, such as previous knowledge and experience, can be injected into the entire architecture through the method of multi attention embedding, which can be easily extended in the future.

On the other hand, we also compared the performance of *PreGenerator* and the baseline model from the perspective of human evaluation. Because the evaluation process is very time-consuming (each project takes more than 1 minute), we only require the evaluator to judge the results from the test set. As shown in Table 3, only 134 test samples were scored, and LSTM-seq2seq scored the lowest, 2.2 points. AttentiveHerb, Herb-Know and TCMBERT are 6.2, 6.4 and 6.8 respectively, which is relatively high, indicating that adding attention or background knowledge to the model can generate more practical prescriptions. Although TCM Translator and Coverage Soft Model performed fairly well in similarity evaluation, they performed poorly in human evaluation, indicating that compatibility and exclusion between herbs are ignored in the generation process. *PreGenerator* scored 7.6, the highest performance of all models, and closed to the label prescription score. This showed that the generation performance of *PreGenerator* is close to the level of experts in the field of traditional Chinese medicine and has great medical potential in the discovery of simple prescription generation.

We have noticed that the first four models are all supervised learning models, which are heavily dependent on TCM record training data. The more medical records used for training, the better performance of these models can be obtained. It shows that the usage of prior knowledge (such as TCM theoretical books and TCM experimental literature) or attention mechanism can improve the effectiveness of prescription generation when the data volume is small. In addition, Herb-know can obtain relatively higher score, indicating that external herbal knowledge can provide useful information for prescription generation. And TCMBERT is pretrained on TCM text resources, which enables the network to have prior knowledge related to TCM, and has a certain role in solving the problem of small samples. The difference between Herb-know and *PreGenerator* in HE and

TABLE 2. The similarity metrics results between generated prescription and label prescription.

Models	HL(-)	P(+)	R(+)	F1(+)
LSTM seq2seq	0.028	0.32	0.27	0.34
PTM	0.055	0.41	0.36	0.39
TCM Translator	0.116	0.39	0.35	0.38
Coverage-Soft-Model	0.012	0.45	0.41	0.30
AttentiveHerb	0.017	0.51	0.33	0.46
Herb-Know	0.010	0.57	0.53	0.50
TCMBERT	0.009	0.59	0.54	**0.60**
PreGenerator (our method)	**0.006**	**0.63**	**0.57**	0.59

"-" indicates that the lower the value is, the better the model' performance. "+" is the opposite. Bold indicates the best result of each experiment.

TABLE 3. Human evaluation results of *PreGenerator* and baseline models.

Models	HE(+)	HC(+)	Total
Lable	3.4	4.5	7.9
LSTM seq2seq	1.0	1.2	2.2
PTM	2.7	1.3	4.0
TCM Translator	1.9	1.6	3.5
Coverage-Soft-Model	2.5	0.9	3.4
AttentiveHerb	3.7	2.5	6.2
Herb-Know	3.8	2.8	6.4
TCMBERT	3.9	2.9	6.8
PreGenerator (our method)	**4.2**	**3.4**	**7.6**

HC indicates that the external knowledge used in Herb-know may significantly affect the HE of prescription generation. In addition to therapeutic efficacy, Chinese prescriptions also have their specific format specifications and compatibility rules. *PreGenerator* can learn the format specifications and herbal rules in prescription templates through retrieval, which is more in line with the idea of professional text generation. Therefore, *PreGenerator* is designed and trained with this idea. The experimental results on the test set also demonstrate the effectiveness and practicability of the method. In the case of small amount of data and strong specialization, the proposed model has achieved the optimal performance in the task of generating traditional Chinese medicine prescriptions.

In order to study the learning efficiency of these models, we analyzed the training process of these models. The training process of the model is shown in Fig. 4. It can be observed that in 150 training epochs, the loss of all models decreases gradually with the increase of training rounds. The loss of *PreGenerator* converges to 0.074, while the loss of TCM Translator and Herb-Know is 0.472 and 0.813 respectively. The losses of AttentiveHerb, Coverage-Soft-Model and LSTM-seq2seq are about 1.4, 2.2 and 2.9 respectively. The *PreGenerator*'s loss converges rapidly and reaches the minimum value in all models. This is attributed to the utilization of the retrieval modules, which incorporates retrieval of knowledge from the Chinese medicine domain's knowledge base or dataset. This assists the model in better understanding the data and contextual relationships, thereby enhancing its generalization and generation capabilities. We also observed that AttentiveHerb, Coverage soft model and LSTM-seq2seq still have high losses and slow convergence. These three models are all faced with the same situation, gradually decreasing at the beginning, and then fluctuating in the remaining training stages. The reason is that these three models are purely supervised learning models, requiring a huge sample of labels. However, in this study, the training set only contains 907 labeled samples. *PreGenerator* algorithm had the least loss and the loss decreases rapidly, which demonstrated the learning efficiency and effectiveness of the algorithm from the other aspect.

Furthermore, combined with the results in Table 2, we noticed that TCMBERT's training loss was lower than *PreGenerator's*, but its manual evaluation results were poorer. We believe that TCM Translator has over fitted after several iterations, while the prescription rules have just converged. Using template information through retrieval in *PreGenerator* can improve the performance of the model in professional learning scenarios. Compared with the baseline models, the learning efficiency of *PreGenerator* is significantly better than that of all comparison baselines, which proves the superiority of the proposed method.

2) ABLATION STUDY

We conducted ablation research in our own dataset to verify the effectiveness of each module in *PreGenerator*. In each of the following studies, we only change one module without changing other modules.

TABLE 4. The results of ablation study.

Modles	HL(-)	P(+)	R(+)	F1(+)
PreGenerator without SPR	0.127	0.56	0.41	0.49
PreGenerator without HHR	0.052	0.52	0.43	0.45
PreGenerator without PreD	0.023	0.60	0.55	0.54
PreGenerator	**0.006**	**0.63**	**0.57**	**0.59**

"-" indicates that the lower the value is, the better the model' performance. "+" is the opposite. Bold indicates the best result of each experiment.

a: REMOVING THE SPR MODULE

In this experiment, the prescription template feature $\boldsymbol{p}_s$ is neglected and the first herb is generated only based on symptom features. The HHR module remains unchanged. However, it does not search for the herb with the highest probability from the retrieved prescriptions, but retrieves the herb most likely to appear next from all prescriptions pool. As can be seen from Table 3, removing the SPR module (''without SPR'') resulted in an average performance reduction of 2%. This shows that the symptom-prescription retrieval module can outline the dialectical direction of the entire prescription. The rest of herbal generation is largely affected by prescription level semantic information.

b: REMOVING THE HHR MODULE

In this experiment, the generation of the $t+1$ herb is based on the global prescription feature $\boldsymbol{p}_s$ and symptom feature $\mathbf{x}$ without using the herbal information retrieved in Eq. (5). Table 3 shows that, compared with the complete model, the deletion of the HHR module (''without HHR'') resulted in a 4% decrease in the average evaluation score. This confirms that the HHR module plays an important role in generating coherent and effective clinical prescription reports.

c: REPLACING PRESCRIPTION DECODER

In this experiment, we used a single-layer LSTM, which treated the entire prescription as a long sentence and generated it word by word. Table 3 shows that replacing our prescription decoder with a single-layer LSTM (''without PreD'') will significantly reduce performance. This phenomenon shows that our prescription generation model can effectively and greatly improve the performance of TCM professional text generation task.

V. CASE DISCUSSION

To better demonstrate the feasibility of our proposed method in practical application, we provide an example generated

TABLE 5. The actual generation results of various models for a given case. We have selected several models with good generation results. The doctor's prescription means the label prescription given in the dataset.

【症状】	咳嗽痰多，疲惫不堪，精神萎靡，高热，脉细而无力，舌质淡红，形体消瘦，苔白
Symptoms	Cough phlegm, fatigue, lethargy, high fever, thin and weak pulse, pale red tongue, thin body, white moss
【医生处方】	黄芪，白术，防风，太子参，茯苓，陈皮，浮小麦，炙甘草，大枣
Clinical Prescription	Radix Astragali, Rhizoma Atractylodis Macrocephalae, Divaricate Saposhnikovia Root, Radix pseudostellariae, Poria cocos, Dried tangerine or orange peel, Blighted wheat, Honey-fried licorice root, Chinese date
LSTM seq2seq	白术，桂枝，茯苓，玉竹，柴胡，黄芩，黄芪，
	Rhizoma Atractylodis Macrocephalae, Cassia Twig, Poria cocos, Fragrant Solomonseal, Radix bupleuri, Scutellaria, Radix Astragali
TCM Translator	茯苓，黄芪，黄芩，附子，半夏，柴胡，大枣
	Poria cocos, Radix Astragali, Scutellaria, Prepared Common Monkshood Daughter Root, Pinellia Tuber, Radix bupleuri, Chinese date
Herb-Know	黄芪，白术，橘皮，炙甘草，附子，五味子，菟丝子
	Radix Astragali, Rhizoma Atractylodis Macrocephalae, Exocarpium Citri Leiocarpae, Honey-fried licorice root, Prepared Common Monkshood Daughter Root, Schisandra chinensis, Semen Cuscutae
PreGenerator	黄芪，白术，半夏，茯苓，干姜，炙甘草，人参，党参，陈皮，玉竹
	Radix Astragali, Rhizoma Atractylodis Macrocephalae, Pinellia Tuber, Poria cocos, Dried Ginger, Honey-fried licorice root, Ginseng, Root of Pilose Asiabell, dried tangerine or orange peel, Fragrant Solomonseal

by clinical experts, TCM Translator, PTM (topic model) and the proposed *PreGenerator*. The input symptoms and output herbs are shown in Fig. 5.

For the above case, these symptoms are caused by insufficient qi and blood and low external defense function, so the treatment principle of supplementing qi and strengthening the exterior should be adopted to strengthen the physique. We have noticed that the prescriptions generated by LSTM-seq2seq are basically based on the symptoms of patients to select the corresponding herbs, such as Rhizoma Atractylodis Macrocephalae and Radix Astragali for the treatment of qi deficiency, fatigue and other symptoms, Fragrant Solomonseal for the relief of cough and fever, and Radix bupleuri for the reduction of high fever. However, the cause of the symptoms was not considered, and the type of symptoms could not be identified. For example, although B Radix bupleuri has antipyretic effect, it is cold in nature, and is mainly used for the heart stomach gas stagnation caused by the exchange of cold and heat, which has little effect on the symptoms in this case. In this case, the functions of the viscera of the elderly gradually decline. And the physical performance such as imbalance of qi, blood, yin and yang, the insufficiency of blood essence and body fluid, and the low defense function of the body always appeared to them. So, the susceptibility to external pathogens is strong, and the main treatment principles are to replenish qi, strengthen the surface, and harmonize. We observed that TCM Translator, Herb-Know and *PreGenerator* all took this key factor into account in the formulation generated by the three models, and selected Rhizoma Atractylodis Macrocephalae, Poria cocos, Radix Astragali, and Honey-fried licorice root as the main herbs to benefit the spleen and stomach, and to replenish qi and restore pulse. However, TCM Translator has seriously neglected the compatibility principle of herbs. The combination of Prepared Common Monkshood Daughter Root and Pinellia Tuber will increase the burden on the stomach and even cause poisoning. Due to the introduction of external knowledge, Herb-Know model generated herbs that basically conform to the treatment principles and compatibility rules, but also ignored some symptoms caused by the exogenous cold syndrome of patients. The model proposed in this paper can not only generated five herbs in the labeled prescription in terms of similarity, but also took advantage of Rhizoma Atractylodis Macrocephalae, poria cocos and honey-fried licorice root to nourish the spleen and stomach, Radix Astragali to replenish qi and strengthen the surface. Pinellia Tuber and dried tangerine or orange peel were usesd to relieve cough and phlegm, Fragrant Solomonseal was generated to relieve cough and thirst, deficiency and fever. And at the same time, these herbs cooperated with Root of Pilose Asiabell, dried tangerine or orange peel and other herbs to regulate qi and strengthen the spleen, so as to achieve preventive effect, which is not considered in other models.

From the above analysis, it can be seen that compared with the basic LSTM-seq2seq model, the *PreGenerator* model proposed by us can not only recognize the relationship between symptoms and have a better overall understanding of diseases, but also fully understand the compatibility rules between herbs, so as to generate a more suitable combination of herbs. This advantage is consistent with the prescription principle of traditional Chinese medicine, that is, prescriptions should focus on ''differentiation of syndrome'' (the reason behind the symptoms), rather than superficial ''symptoms''.

VI. CONCLUSION

For traditional Chinese medicine, the most widely used form of preserving medical knowledge, experimental results, medical analysis results and clinical records is text. Using these literature resources is a challenging task. In this work, we aim to depict the hierarchical retrieval process in the prescription generation. Specifically, a retrieval enhanced based TCM prescription generation model is proposed. The purpose is to make full use of unstructured resources and knowledge of traditional Chinese medicine, as well as limited clinical records, to generate TCM prescriptions. Firstly, the SPR module is used to retrieve the most relevant prescriptions for a given patient's symptoms. In order to follow the rule of compatibility between herbs, the HHR module is introduced to retrieve the next most relevant herb according to the previously generated herbs. Finally, the prescription decoder combines the symptom features and the retrieved prescription and herbal medicine features to generate a reasonable and effective prescription. The experiments verify the rationality and effectiveness of our model and the results in Table 3 fully illustrate the advantages of each module. The experimental results also show that the model has a strong ability to learn knowledge representation from unstructured resources, and can effectively learn the principles of traditional Chinese medicine, the interaction between traditional Chinese medicine and symptoms, and the role of herbal medicine in prescriptions from clinical medical records.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive advice, and also would like to thank the hard works of the doctors from Affiliated Hospital, Shanxi University of Traditional Chinese Medicine, Taiyuan, China, who help them to estimate the generated prescriptions.

REFERENCES

[1] D. Andrzejewski, X. Zhu, M. Craven, and B. Recht, "A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1171–1177.

[2] J. Qiu, "Traditional medicine: A culture in the balance," *Nature*, vol. 448, no. 7150, pp. 126–129, 2007.

[3] L. Yao, Y. Zhang, B. Wei, W. Zhang, and Z. Jin, "A topic modeling approach for traditional Chinese medicine prescriptions," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1007–1021, Jun. 2018.

[4] F. Lin, J. Xiahou, and Z. Xu, "TCM clinic records data mining approaches based on weighted-LDA and multi-relationship LDA model," *Multimedia Tools Appl.*, vol. 75, no. 22, pp. 14203–14232, Nov. 2016.

[5] W. Ji, Y. Zhang, X. Wang, and Y. Zhou, "Latent semantic diagnosis in traditional Chinese medicine," *World Wide Web*, vol. 20, no. 5, pp. 1071–1087, Sep. 2017.

[6] J. Wood, P. Tan, W. Wang, and C. Arnold, "Source-LDA: Enhancing probabilistic topic models using prior knowledge sources," in *Proc. IEEE 33rd Int. Conf. Data Eng. (ICDE)*, Apr. 2017, pp. 411–422.

[7] Y. Jin, W. Zhang, X. He, X. Wang, and X. Wang, "Syndrome-aware herb recommendation with multi-graph convolution network," in *Proc. IEEE 36th Int. Conf. Data Eng. (ICDE)*, Apr. 2020, pp. 145–156.

[8] Y. Jin, W. Ji, W. Zhang, X. He, X. Wang, and X. Wang, "A KG-enhanced multi-graph neural network for attentive herb recommendation," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 5, pp. 2560–2571, Sep. 2022.

[9] Y. Yang, Y. Rao, M. Yu, and Y. Kang, "Multi-layer information fusion based on graph convolutional network for knowledge-driven herb recommendation," *Neural Netw.*, vol. 146, pp. 1–10, Feb. 2022.

[10] S. Li, W. Yue, and Y. Jin, "Patient-oriented herb recommendation system based on multi-graph convolutional network," *Symmetry*, vol. 14, no. 4, p. 638, Mar. 2022.

[11] W. Li and Z. Yang, "Exploration on generating traditional Chinese medicine prescriptions from symptoms with an end-to-end approach," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.* Cham, Switzerland: Springer, 2019, pp. 486–498.

[12] Z. Wang, J. Poon, and S. Poon, "TCM translator: A sequence generation approach for prescribing herbal medicines," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 2474–2480.

[13] C. Rong, X. Li, X. Sun, and H. Sun, "Chinese medicine prescription recommendation using generative adversarial network," *IEEE Access*, vol. 10, pp. 12219–12228, 2022.

[14] Z. Liu, Z. Zheng, X. Guo, L. Qi, J. Gui, D. Fu, Q. Yao, and L. Jin, "AttentiveHerb: A novel method for traditional medicine prescription generation," *IEEE Access*, vol. 7, pp. 139069–139085, 2019.

[15] C. Li, D. Liu, K. Yang, X. Huang, and J. Lv, "Herb-know: Knowledge enhanced prescription generation for traditional Chinese medicine," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 1560–1567.

[16] Z. Liu, C. Luo, D. Fu, J. Gui, Z. Zheng, L. Qi, and H. Guo, "A novel transfer learning model for traditional herbal medicine prescription generation from unstructured resources and knowledge," *Artif. Intell. Med.*, vol. 124, Feb. 2022, Art. no. 102232.

[17] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[18] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[19] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 375–383.

[20] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," 2017, *arXiv:1711.08195*.

[21] J. Yuan, H. Liao, R. Luo, and J. Luo, "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*. Shenzhen, China: Springer, Oct. 2019, pp. 721–729.

[22] G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, and M. Ghassemi, "Clinically accurate chest X-ray report generation," in *Proc. Mach. Learn. Healthcare Conf.*, 2019, pp. 249–269.

[23] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12910–12917.

[24] S. Biswal, C. Xiao, L. M. Glass, B. Westover, and J. Sun, "CLARA: Clinical report auto-completion," in *Proc. Web Conf.*, Apr. 2020, pp. 541–550.

[25] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9459–9474.

[26] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," 2020, *arXiv:2007.01282*.

[27] Y. Zhang, S. Sun, X. Gao, Y. Fang, C. Brockett, M. Galley, J. Gao, and B. Dolan, "RetGen: A joint framework for retrieval and grounded text generation modeling," 2021, *arXiv:2105.06597*.

[28] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 6666–6673.

[29] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

ZIJUAN ZHAO received the M.D. degree from the College of Information and Computer, Taiyuan University of Technology, China, in 2020, where she is currently pursuing the Ph.D. degree in computer applications technology. Her current research interests include medical image processing, TCM text analysis, and deep learning.

XUETING REN received the M.D. degree from the College of Information and Computer, Taiyuan University of Technology, China, in 2020, where she is currently pursuing the Ph.D. degree in computer applications technology. Her current research interests include medical image processing, data quality upgrade, and deep learning.

KAI SONG received the master's degree from the College of Information and Computer, Taiyuan University of Technology, China, in 2022. He is currently pursuing the Ph.D. degree in optical engineering. His current research interest includes computational imaging.

YAN QIANG received the Ph.D. degree from the Department of Computer Application Technology, Taiyuan University of Technology (TYUT), China, in November 2010. He is currently a Professor with the College of Information and Computer, TYUT. His recent research interests include data mining, medical image processing, and cloud computing.

JUANJUAN ZHAO received the Ph.D. degree from the Department of Computer Application Technology, Taiyuan University of Technology (TYUT), China, in November 2010. She is currently a Professor with the College of Software Engineering, TYUT. Her current research interests include medical image processing and deep learning.

JUNLONG ZHANG received the Ph.D. degree in the major of basic theory from Chinese Medicine, Shandong Traditional Chinese Medicine University, China, in 1997. He is currently a Professor with Shanxi Traditional Chinese Medicine University. His current research interests include anti-tumor metastasis and its mechanism.

PENG HAN received the M.D. degree from the College of Information and Computer, Taiyuan University of Technology, China, in 2020. He is currently a Technical Design Engineer with the North Automatic Control Technology Institute, Taiyuan, China.