Domain Specific Pre-training Methods for Traditional Chinese Medicine Prescription Recommendation

Wei Li[1], Zheng Yang[2], and Yanqiu Shao[1(✉)]

[1] School of Information Science, Beijing Language and Culture University, Beijing 100081, China
{liweitj47,shaoyanqiu}@blcu.edu.cn
[2] School of Traditional Chinese Medicine, Beijing University of Chinese Medicine, Beijing 100029, China
yangzheng@bucm.edu.cn

Abstract. Traditional Chinese Medicine (TCM) is an important constituent of medical treatment. During the development history of TCM, there have been a large number of medical records accumulated, which embody the experiential judgement of the TCM practitioners. There are usually the symptoms observed by the practitioner and the according treatment methods within the records. In the treatment procedure, TCM practitioners often refer to the classical records and the prescriptions within them, which makes recommending prescriptions from the records based on the observation of the symptoms valuable in practice. Based on these observations, we propose to model this problem as a matching based recommendation task. To precisely model the relation between symptoms and prescriptions, inspired by the success of pre-trained language models, we propose a TCM domain specific hybrid input construction method and multi-grained negative sampling methods and training objectives. To verify the effectiveness of the proposed method, we conduct extensive experiments on the symptom-prescription dataset. The experiment results show that our proposed method can accurately recommend suitable prescriptions with more abundant candidates for the reference of TCM practitioners, making it more valuable in practice.

Keywords: Prescription Recommendation · Traditional Chinese Medicine · Pre-trained Language Model

1 Introduction

In recent years, with the development of deep learning and natural language processing, artificial intelligence (AI) has been applied in numerous domains. Among them, the integration of AI and healthcare is considered one of the most promising directions. Current research on AI and healthcare primarily focuses on

W. Li and Z. Yang—Equal Contribution.

L. Fang et al. (Eds.): CICAI 2023, LNAI 14474, pp. 125–135, 2024.
https://doi.org/10.1007/978-981-99-9119-8_12

modern medical fields, while lacking attention to traditional Chinese medicine (TCM). Leveraging deep learning and natural language processing techniques to explore and utilize the rich knowledge inherited from the historical practices of TCM has significant theoretical and practical implications. Particularly, recommending suitable herbal formulas based on the diagnosis and symptom descriptions provided by TCM practitioners is an important application scenario with practical and theoretical significance.

Previous studies have explored the use of machine learning and deep learning methods for recommending herbal prescriptions based on diagnostic information. [4] initially proposed the use of a sequence-to-sequence model with an improved objective function to generate the herbal components of prescriptions based on textual symptom descriptions. [6,7], and [3] respectively suggested leveraging expert knowledge, attention models to learn the associations between symptoms and herbs, as well as associations between different herbs, and incorporating external herbal knowledge to assist in prescription generation. [8] applied transfer learning using a pre-trained bidirectional encoder, known as BERT (Bidirectional Encoder Representations from Transformers), to the task of generating traditional Chinese medicine prescriptions. These works primarily focus on recommendation through a generative approach. However, generative methods possess certain inherent limitations that are challenging to overcome, such as limited interpretability, difficulty in providing recommendation justifications, and relatively fixed patterns. In actual clinical practice, high reliability is crucial, and these limitations restrict the practical utility of generative methods in assisting traditional Chinese medicine practitioners during the diagnosis and treatment process.

Inspired by the application of next sentence prediction in prompt tuning [2,5] based on pre-trained language models [9], we propose using the next sentence classification objective to match diagnostic texts with herbal prescription components. To effectively leverage the information in herb names, we suggest incorporating both the textual representation and the ID identifier of the herbs as inputs to the model. This approach not only allows the model to capture the intrinsic characteristics of the herbs but also facilitates modeling of herbs that are difficult to automatically identify by considering their textual descriptions. Furthermore, for the constructed diagnostic-prescription inputs, we propose adapting the model through masked language modeling, enabling the establishment of associations at a finer-grained level, including the relationships between diagnosis-herb, herb-herb, and herb name-herb.

Considering the characteristics of traditional Chinese medicine record studies, in order to better train the matching between symptom descriptions and prescription compositions, we introduce two granularity levels of negative sample construction. This involves randomly replacing the original prescription at the prescription level and herb level, respectively, and requires the model to detect the substitutions at different granularity levels, enabling it to differentiate between differences in granularity among prescriptions.

We conducted extensive experiments on a dataset specifically transformed for recommendation scenarios. The results demonstrate that our proposed method achieves more accurate herbal prescription recommendations compared to generative methods. Additionally, the recommended prescriptions exhibit better diversity, thereby providing better assistance to traditional Chinese medicine practitioners in practical diagnosis and treatment processes. Furthermore, the experimental results indicate that directly utilizing the next sentence prediction objective from pre-trained language models for training and prediction does not yield satisfactory matching performance. However, by incorporating model designs that consider the characteristics of traditional Chinese medicine record studies, we achieve significant improvements in matching effectiveness.

The main contributions of this paper can be summarized as follows:

- We propose modeling the objective of recommending herbal prescriptions based on diagnosis as a retrieval-based recommendation task. We introduce the utilization of the next sentence prediction method, based on pre-trained language models, to match symptom descriptions with herbal prescriptions.
- Addressing the characteristics of traditional Chinese medicine prescription recommendation, we propose a hybrid model input construction pattern and a multi-granularity negative sampling method, as well as matching training objectives that align with tasks in the field of traditional Chinese medicine.
- We conducted extensive experiments and analysis on a diagnostic-prescription dataset to validate the effectiveness of the proposed approach.

2 Approach

In this section, we describe how we construct the inputs for the model based on the symptom-formula pairs, as well as how we create training examples and training objectives for matching training.

2.1 Input Construction for Pre-trained Language Model

Based on observations on the characteristics of traditional Chinese medicine (TCM) record data, we propose modeling the correspondence between symptom descriptions and herbal formulations as a next sentence prediction relationship. In other words, if there is a correspondence between the symptom description and the herbal formulation, they form a sentence pair relationship; otherwise, they do not form a sentence pair relationship.

Since the composition of a prescription consists of herbal medicine, we initially consider using the entire herb entity as the input unit. However, due to the nature of Chinese herbal medicine, which is derived from various natural sources, and taking into account the presence of non-standardized herb descriptions in ancient medical texts, directly using the entire herb entity as input would render these herbs out of vocabulary (OOV), thereby reducing the availability of effective context. Additionally, many herb names exhibit certain similarities with

their corresponding standardized herb names, with the only difference being the use of different names or the inclusion of preparation methods, places of origin, and other information. For example, "生地" actually refers to the same medicinal substance as "生地黄". In such cases, the textual descriptions of herbs themselves provide valuable information. Based on these observations, we propose combining the entire herb entity with the textual herb name as the input for the composition of a prescription.

Taking into account the considerations mentioned above and drawing inspiration from the input format of the next sentence prediction task in pre-trained language models, we propose constructing the symptom description-prescription pairs that require judgment in the following form:

$$[CLS]X[SEP]Y_sY_m[SEP]$$

Here, Y_s and Y_m respectively refer to the textual representation of the herb name and the ID identifier of the herb as a whole. The special symbol [CLS] is used to learn the representation at the sample level for the pair, while [SEP] is used to separate the symptom description and the prescription composition and marks the end of the input. Additionally, to differentiate the roles of the symptom description and the prescription, we follow the approach of BERT and incorporate token types in the model input. The symptom description is marked with 0, while the herb portion of the prescription is marked with 1.

2.2 Training Data Sampling

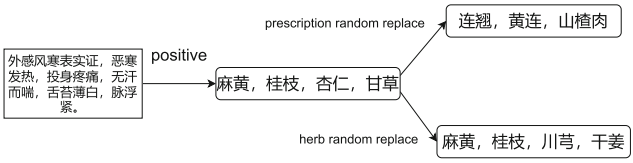

Fig. 1. Example of negative sample construction strategies.

The symptom-prescription pairs in the data naturally form positive examples for training the matching relationship. To train the model's ability to judge whether there is a match, we also need to construct negative examples. Drawing inspiration from the method of constructing negative examples for next sentence prediction and considering the characteristics of TCM medical records, we propose two granularities of random replacement schemes (as shown in Fig. 1). The first scheme is to randomly replace the entire prescription with another prescription from the training set, which is a prescription-level random replacement. The second scheme is to randomly replace certain herbs in the corresponding prescription with another herb, which is a herb-level random replacement.

The first scheme replaces the prescription with a valid prescription, but it may differ significantly from the original paired prescription, resulting in insufficient discriminative ability learned by the model, especially at a finer granularity level such as the herb level. The second scheme constructs prescriptions that are closer to the original paired prescription, allowing the model to learn the distinction and grasp the local herb information within the prescription. However, the constructed prescriptions may not be feasible in practice, meaning that the compatibility of the herbs in the prescription may be compromised. In the actual construction of negative examples, we choose to replace less than half of the herbs in the prescription to maintain the original framework of the prescription as much as possible, and the number of replacements itself is determined by random sampling.

For prescriptions with a small number of herbs (less than 3 ingredients), we adopt the first scheme of replacing the entire prescription because replacing individual herbs in this case wouldn't have much significance. For other prescriptions, we randomly choose one of the two schemes with an equal probability, meaning there's a 50% chance of using the first scheme (replacing the entire prescription) and a 50% chance of using the second scheme (replacing individual herbs).

2.3 Training Objective

To enable the model to learn both coarse-grained and fine-grained alignment information, we propose training the model using Masked Language Modeling (MLM) objective, the Symptom-Prescription Matching (SPM) objective, and the Herb Replacement Detection (HRD) objective to train the model from different perspectives.

Masked Language Modeling Objective. Due to the lack of ID representations for the complete herb names in the pre-trained word vector parameters of the pre-trained models, and considering that the training corpus of the pre-trained language models consists of general domain data, it is necessary to adapt the training set data using the input construction method described in Sect. 2.1. This adaptation involves employing a masked language model to learn herb word vector representations and the associations between herb textual descriptions and herb whole IDs proposed in this paper. When randomly replacing input tokens, we drew inspiration from BERT's approach, but with a modification. We only replace non-special characters with the "[MASK]" token with a probability of 15%, allowing the model to learn associations between the masked words or herbs and their contexts. To better capture the associations between symptoms and herbs, we slightly deviate from the original BERT model's masking strategy. Specifically, in some instances, we mask only the symptoms or the formulations separately, while in other instances, we perform completely random masking.

Symptom-Prescription Matching Objective. To directly model the overall matching relationship between symptoms and the composition of the formula, we

employ a training objective similar to that of the next sentence prediction task. We utilize the hidden vector representation of the special token [CLS], which is encoded by the BERT encoder, to predict whether the input pair is a match. If there is a match, a label of 1 is assigned; otherwise, a label of 0 is assigned. The loss function for the matching relationship is the cross-entropy between the predicted match and the ground truth label.

$$L_{match} = -\sum_{i=0}^{1} y_i log p_i \tag{1}$$

where p represents the predicted probability for the overall matching, y denotes the actual label indicating whether there is a match or not, and i takes the value of 0 or 1, indicating the match or non-match scenario, respectively.

Herb Replacement Detection Objective. In order to enable the model to differentiate more fine-grained matching information between the symptoms and individual herbs in the prescriptions (i.e., which parts of the herbs match the symptoms and which parts do not), we draw inspiration from the work of [1] and propose a method to train the model to detect specific mismatched herb information in non-matching prescriptions while simultaneously learning the overall matching relationship between symptoms and prescriptions. Specifically, for the negative examples constructed through the method of replacing local herbs mentioned in Sect. 2.2, we train the model to predict which herbs in the prescriptions are original (matching the symptoms) and which herbs are replaced (not matching the symptoms). We assign a label of 1 to the originally correctly matched herbs and a label of 0 to the replaced herbs. The logic behind label assignment is consistent with the coarse-grained labels, aiming to help the model learn the finer-grained reasons for mismatches. For the negative examples constructed through the method of randomly replacing prescriptions at the prescription level, we do not train the model to detect whether herbs are replaced, as the majority of herbs are replaced in this case. The loss function used in this context is similar to the cross-entropy used for coarse-grained labels but applied to each herb in the negative examples (for the method of replacing herbs):

$$L_{token} = -\frac{1}{L}\sum_{j}\sum_{i=0}^{1} y_i^j log p_i^j \tag{2}$$

The overall training loss is the sum of three components: the loss of the fine-grained masked language model, the loss of the coarse-grained symptom-prescription matching, and the loss of the replaced herb detection. For the fine-grained masked language model, the loss is calculated based on the predicted probability p of whether a herb is replaced, the actual label y indicating whether it is a replaced herb, the matching label i (0 or 1) indicating whether the symptom and prescription match, the position j in the sample, and the input sequence length L. The final training loss can be expressed as follows:

$$Loss = L_{mlm} + L_{match} + L_{token} \tag{3}$$

During the inference testing phase, we directly use the probability p_i of the coarse-grained matching judgment from Eq. 1 as the prediction probability. We then sort the probabilities in descending order and obtain the actual order of recommended prescriptions.

3 Experiment

In this chapter, we introduce the experimental setup, the data used in the experiments, the evaluation metrics employed, as well as the experimental results and analysis.

3.1 Setting

The BERT model used in this study is Guwen-bert (base)[1], which is pretrained on classical Chinese language corpus. The hidden layer size of the model is 768, with 12 layers and 12 heads in the multi-head attention mechanism. The masked language model was trained for 10 epochs on the symptom-prescription pairs data. The symptom-prescription matching objective was trained for 5 epochs. The model with the highest Macro-F1 score on the development set during training was selected as the test model. Regarding the size of the herb vocabulary, we selected the top 3000 herbs with the highest frequency of occurrence as the vocabulary for whole herbs when represented as characters. The remaining herbs (including noise that has not been cleaned) were represented in textual form. For each positive sample in the matching training, two negative samples were sampled. The batch size during training was set to 24 (limited by GPU memory). For the symptom description, the first 150 characters were extracted, and for the prescription, the first 50 herbs were extracted. The selection of hyper-parameters was based on the highest Macro F1 score obtained on the development set.

3.2 Data

Based on the Chinese medical record data used by [4], we transformed the data into a format suitable for the recommendation task. Using the Jaccard matching method, we first found the top 20 symptom-prescription pairs in the prescription database that were closest to the target symptom description (excluding the symptom-prescription pairs in the test set). These 20 identified prescriptions were mixed with the target prescription as negative examples, and the model was required to find the most suitable prescription for the target symptom from these 21 prescriptions. For the sake of comparison, we used the same test set as [4]. The test set was divided into two parts.

[1] GuwenBERT https://github.com/ethan-yt/guwenbert.

Table 1. Overall results on TextBook and Crawl test set. Precision@5, Recall@5 and F1@5 are provided after "/" for our proposed method. seq2seq and multi-label are the baselines applied in [4].

TextBook	MRR	MAP	MacroPrecision	MacroRecall	MacroF1
proposal	28.68	28.68	40.42/79.52	47.38/84.64	42.07/80.44
seq2seq	-	-	30.97/-	23.70/-	26.85/-
multi-label	-	-	13.51/-	40.49/-	20.26/-
Li and Yang [4]	-	-	38.22/-	30.18/-	33.73/-
Crawl					
proposal	21.17	21.17	24.07/54.53	24.73/55.60	23.21/52.97
seq2seq	-	-	26.03/-	13.52/-	17.80/-
multi-label	-	-	10.83/-	29.72/-	15.87/-
Li and Yang [4]	-	-	29.57/-	17.30/-	21.83/-

3.3 Evaluation Metrics

In this section, we introduce the evaluation metrics used in our experiments. To assess the performance of the model from different perspectives, we employ two types of evaluation metrics. The first type is commonly used in recommender systems, namely MRR (Mean Reciprocal Rank) and MAP (Mean Average Precision). These metrics focus on the relative ranking of the model's results, where higher scores are assigned when the correct answer is ranked higher by the model. Another type focuses on the degree of overlap between the herb composition of recommended prescriptions and the herb composition of standard answers, aiming for finer granularity. The higher the degree of overlap, the closer the recommended prescriptions are to the answers. This type of method includes Macro Precision, Macro Recall, and Macro F1.

3.4 Results

In Table 1, we present the experimental results of our approach in comparison to the results reported in previous work [4], which used the same dataset as ours. It can be observed that our proposed method achieved significant improvements in Macro F1 values compared to the results obtained by the previous generative models, particularly on the more accurate TextBook test set. The Macro F1@1 reached 42.07, a substantial improvement over the 33.73 achieved by Li and Yang's method [4].

Furthermore, our method achieved a Macro F1@5 of 80.44 on the TextBook test set and 52.97 on the Crawl dataset. A higher Macro F1@5 indicates that, in the context of prescription recommendation, providing the top 5 prescriptions that the model considers optimal as candidate recommendations can yield correct recommendations with a high probability, making our approach more practical compared to generative methods.

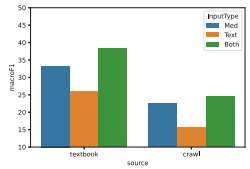

Fig. 2. Macro-F1 for different input construction methods. "Med" indicates only ID of herbs are used, "Text" indicates only textual names are used.

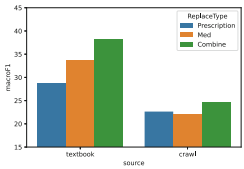

Fig. 3. Macro-F1 for different negative sampling methods. "Prescription" indicates prescription level negative sample, "Med" indicates herb level negative sampling.

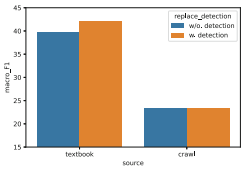

Fig. 4. Macro-F1 for whether herb replacement detection objective is applied.

Although our method did not achieve high scores for retrieval-related evaluation metrics such as MRR and MAP, this is due to the characteristic of Traditional Chinese Medicine records, where similar medical conditions may have different treatment approaches and, therefore, different prescription solutions. Additionally, in the context of prescription recommendation, the absence of an exact match with the prescriptions in textbooks or medical records does not necessarily mean the answer is incorrect. Some discrepancies may arise from non-standardized herb terminologies, while others may result from variations in diagnostic details while still providing prescriptions that are similar to the answers but with some additions or omissions of herbs. These aspects can be reflected in the Macro F1 value. We also provide specific examples in Appendix to further illustrate this.

3.5 Analysis

In this section, we will analyze the effectiveness of our proposed method from several different aspects.

Input Construction Effect. In Fig. 2, we present the results of Macro F1@1 obtained from different input construction methods mentioned in Sect. 2.1 on two test sets (other metrics show a similar trend to Macro F1@1). From the results, we can observe that the performance is weakest when solely using herb text as input (labeled as "Text" in Fig. 2). We believe this is because the herb-related text encountered by the pre-trained language model in the pre-training corpus is sparse, which makes it difficult for the model to accurately differentiate and recognize different herbs based solely on their textual representations. Although the training of the masked language model in our proposed method involves herb-symptom pairs, where herbs are more densely present, the overall quantity is still insufficient to support the model in learning precise herb recognition and differentiation abilities. On the other hand, using herb identifiers (IDs) as input

(labeled as "ID" in Fig. 2) yields better results compared to solely using herb text. We attribute this improvement to the fact that the model can more easily learn the associations between herb IDs and the symptom descriptions. IDs have a smaller semantic space compared to text, making it easier for the model to learn more accurate representations, especially for frequently used herbs. Compared to the two aforementioned individual herb input construction methods, the hybrid input construction method that combines herb text and herb IDs (labeled as "Both" in Fig. 2) provides richer information. It can capture the information of herb IDs for common herbs and the information of herb text representations for less common herbs. Additionally, it allows the model to learn the associations between herb text and herb IDs, leading to the best performance.

Negative Sampling Effect. In Fig. 3, we present the results of different negative sampling methods mentioned in Sect. 2.2 on Macro F1 (@1). It is important to note that the only difference here lies in the sampling methods, while the number or proportion of negative samples remains the same. From the graph, we can observe that both combined negative sampling methods proposed in this paper achieve the best performance on both test sets. On the TextBook test set, the effect of randomly replacing herbs at a finer-grained herb level is significantly better than randomly replacing herbs at a coarser-grained prescription level. On the Crawl test set, the two methods show similar performance. We believe this is because the TextBook test set has higher data quality, making it more sensitive to differences in the model's understanding of herb details. By combining negative samples at two different granularities, the model can better learn how to match symptom descriptions and prescriptions at different levels, resulting in the best matching performance.

Herb Replacement Detection Effect. In Fig. 4, we present the results of whether to use the replacement herb detection objective mentioned in Sect. 2.3 during training. From the results, we can observe that using this training objective brings some improvement on the TextBook test set, but the difference is not significant on the Crawl test set. We believe this phenomenon is due to the higher data quality of the TextBook test set, which better reflects the model's ability to grasp herb details. In fact, for the macro F1@5 metric (not shown in the graph), after adding the replacement herb detection objective, the macro F1@5 of TextBook improved from 76.89 to 80.44, and the macro F1@5 of Crawl improved from 52.18 to 52.97. This further confirms the effectiveness of this training objective from another perspective.

4 Conclusion

This article proposes a symptom-prescription matching method based on pre-trained language models for the task of recommending prescriptions based on symptom descriptions. In this method, we model the symptom-prescription

matching as the next sentence prediction task in pre-trained language models. Considering the characteristics of TCM medical records, we propose a hybrid medication input construction method, a multi-granularity negative sampling method, and training objectives that are adapted to the task, allowing the model to learn the associations and matching relationships at different levels between symptom descriptions, prescriptions, and herbs. Extensive experiments and analysis demonstrate that our proposed method can provide more accurate prescription recommendations compared to generative methods and offer more diverse candidate answers, thereby enhancing the practical diagnostic process for TCM practitioners.

Acknowledgements. This research project is supported by National Key R&D Program of China (2020YFC2003100, 2020YFC2003102), Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine. (No: ZYYCXTD-C-202001), Science Foundation of Beijing Language and Culture University (supported by "the Fundamental Research Funds for the Central Universities") (No. 21YBB19)

References

1. Chuang, Y.S., et al.: Diffcse: difference-based contrastive learning for sentence embeddings. arXiv preprint arXiv:2204.10298 (2022)
2. Han, X., Zhao, W., Ding, N., Liu, Z., Sun, M.: PTR: prompt tuning with rules for text classification. arXiv preprint arXiv:2105.11259 (2021)
3. Li, C., Liu, D., Yang, K., Huang, X., Lv, J.: Herb-know: knowledge enhanced prescription generation for traditional Chinese medicine. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1560–1567. IEEE (2020)
4. Li, W., Yang, Z.: Exploration on generating traditional Chinese medicine prescriptions from symptoms with an end-to-end approach. In: Tang, J., Kan, M.-Y., Zhao, D., Li, S., Zan, H. (eds.) NLPCC 2019. LNCS (LNAI), vol. 11838, pp. 486–498. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32233-5_38
5. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586 (2021)
6. Liu, Z., et al.: Attentiveherb: a novel method for traditional medicine prescription generation. IEEE Access **7**, 139069–139085 (2019)
7. Ruan, C., Luo, H., Wu, Y., Yang, Y.: TPGEN: prescription generation using knowledge-guided translator (2021)
8. Shi, Q.Y., Tan, L.Z., Seng, L.L., Wang, H.J., et al.: Intelligent prescription-generating models of traditional Chinese medicine based on deep learning. World J. Traditional Chin. Med. **7**(3), 361 (2021)
9. Sun, Y., Zheng, Y., Hao, C., Qiu, H.: NSP-BERT: a prompt-based zero-shot learner through an original pre-training task-next sentence prediction. arXiv abs/2109.03564 (2021)