TCM-SD: A Benchmark for Probing Syndrome Differentiation via Natural Language Processing

Mucheng Ren[1], **Heyan Huang**[1], **Yuxiang Zhou**[1], **Qianwen Cao**[1], **Yuan Bu**[2], **and Yang Gao**[1]
[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[2]Xuzhou City Hospital of Traditional Chinese Medicine, Xuzhou, China
{renm,hhy63,yxzhou,qwcao,gyang}@bit.edu.cn
buyuantcm@gmail.com

Abstract

Traditional Chinese Medicine (TCM) is a natural, safe, and effective therapy that has spread and been applied worldwide. The unique TCM diagnosis and treatment system requires a comprehensive analysis of a patient's symptoms hidden in the clinical record written in free text. Prior studies have shown that this system can be informationized and intelligentized with the aid of artificial intelligence (AI) technology, such as natural language processing (NLP). However, existing datasets are not of sufficient quality nor quantity to support the further development of data-driven AI technology in TCM. Therefore, in this paper, we focus on the core task of the TCM diagnosis and treatment system—syndrome differentiation (SD)—and we introduce the first public large-scale benchmark for SD, called TCM-SD. Our benchmark contains 54,152 real-world clinical records covering 148 syndromes. Furthermore, we collect a large-scale unlabelled textual corpus in the field of TCM and propose a domain-specific pre-trained language model, called ZY-BERT. We conducted experiments using deep neural networks to establish a strong performance baseline, reveal various challenges in SD, and prove the potential of domain-specific pre-trained language model. Our study and analysis reveal opportunities for incorporating computer science and linguistics knowledge to explore the empirical validity of TCM theories.

1 Introduction

As an essential application domain of natural language processing (NLP), medicine has received remarkable attention in recent years. Many studies have explored the integration of a variety of NLP tasks with medicine, including question answering (Pampari et al., 2018; Tian et al., 2019), machine reading comprehension (Li et al., 2020; Yue et al., 2020), dialogue (Zeng et al., 2020), named entity recognition (Jochim and Deleris, 2017; He et al., 2020), and information retrieval (Liu et al., 2018). Meanwhile, numerous datasets in the medical domain with different task formats have also been proposed (Pampari et al., 2018; Li et al., 2020; Tian et al., 2019). These have greatly promoted the development of the field. Finally, breakthroughs in such tasks have led to advances in various medical-related applications, such as decision support (Feng et al., 2020; Panigutti et al., 2021) and International Classification of Disease (ICD) coding (Cao et al., 2020; Yuan et al., 2022).

However, most existing datasets and previous studies are related to modern medicine, while traditional medicine has rarely been explored. Compared to modern medicine, traditional medicine is often faced with a lack of standards and scientific explanations, making it more challenging. Therefore, it is more urgent to adopt methods of modern science, especially NLP, to explore the principles of traditional medicine, since unstructured texts are ubiquitous in this field.

TCM, as the representative of traditional medicine, is a medical system with a unique and complete theoretical basis formed by long-term medical practice under the influence and guidance of classical Chinese materialism and dialectics. Unlike modern medicine, in which medical professionals assign treatments according to disease type, TCM practitioners conduct in-depth analyses based on evidence collected from four diagnostics methods—inspection, auscultation and olfaction, interrogation, and palpation—to determine which type of **syndrome (zheng,** 证) the patient experiencing. Different treatment methods are then adopted according to the type of syndrome. Therefore, patients with the same

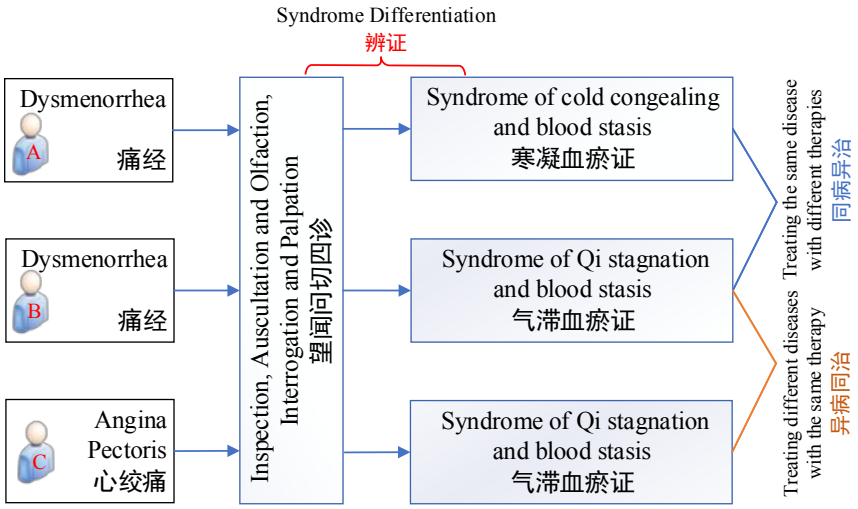


Figure 1: Concept of Traditional Chinese Medicine (TCM) syndrome differentiation.

disease may have different syndromes and thus receive different treatments, while patients with different diseases may have the same syndrome and thus undergo the same treatment. These concepts are called "treating the same disease with different therapies (同病异治)" and "treating different diseases with the same therapy (异病同治)," respectively, which are the core methods upheld by TCM.

For the example shown in Figure 1, patients A and B have the same disease—dysmenorrhea—but one is influenced by cold while the other is driven by Qi stagnation (which is a specific concept in TCM). Thus, different therapies would be assigned. However, patient C suffered from angina pectoris but shared the same syndrome as patient B. Therefore, they would be treated with similar therapies. Thus, the **syndrome**, instead of the disease, can be regarded as the primary operating unit in the TCM medical system, which not only effectively summarizes the patients' symptoms but also determines the subsequent treatment. In this process, known as **syndrome differentiation**, *the inferencing task of deciding which syndrome is associated with a patient based on clinical information*, is a vital pivot of the TCM medical system.

In recent years, with the discovery of artemisinin (Tu, 2016) and the beneficial clinical manifestations of TCM to treat COVID-19 (Yang et al., 2020; Zhang et al., 2020b), TCM has increasingly attracted attention. There have been some studies in which NLP techniques were used to explore SD tasks (Zhang et al., 2019; Zhang et al., 2020a; Wang et al., 2018; Liu et al., 2020; Pang et al., 2020), but the development has been significantly hindered by the lack of large-scale, carefully designed, public datasets.

Therefore, this paper aims to further integrate traditional medicine and artificial intelligence (AI). In particular, we focus on the core task of TCM—syndrome differentiation (SD)—to propose a high-quality, public SD benchmark that includes 54,152 samples from real-world clinical records. To our best knowledge, this is the first time that a textual benchmark has been constructed in the TCM domain. Furthermore, we crawled data from the websites to construct a TCM domain text corpus and used this to pre-train a domain-specific language model called as ZY-BERT (where ZY came from the Chinese initials of TCM). The experiments and analysis of this dataset not only explored the characteristics of SD but also verified the effectiveness of domain-specific language model.

Our contributions are summarized as follows:

1. We have systematically constructed the first public large-scale SD benchmark in a format that conforms to NLP, and established the strong baselines. This can encourage researchers use NLP techniques to explore the principles of TCM that are not sufficiently explained in other fields.

2. We proposed two novel methods, pruning and merging, which could normalize the syndrome type, improve the quality of the dataset, and also provide a reference for the construction of similar TCM datasets in the future.

3. We proposed a domain-specific language model named as ZY-BERT pre-trained with a large-scale unlabeled TCM domain corpus, which produces the best performances so far.

2 Preliminaries

To facilitate the comprehension of this paper and its motivation and significance, we will briefly define several basic concepts in TCM and analyze the differences between TCM and modern medicine.

2.1 Characteristics of Traditional Chinese Medicine (TCM) Diagnosis

The most apparent characteristic of TCM is that it has a unique and complete diagnostic system that differs from modern medicine. In modern medicine, with the assistance of medical instruments, the type of disease can be diagnosed according to the explicit digital indicators, such as blood pressure levels. However, TCM adopts abstract indicators, such as Yin and Yang, Exterior and Interior, Hot and Cold, and Excess and Deficiency.

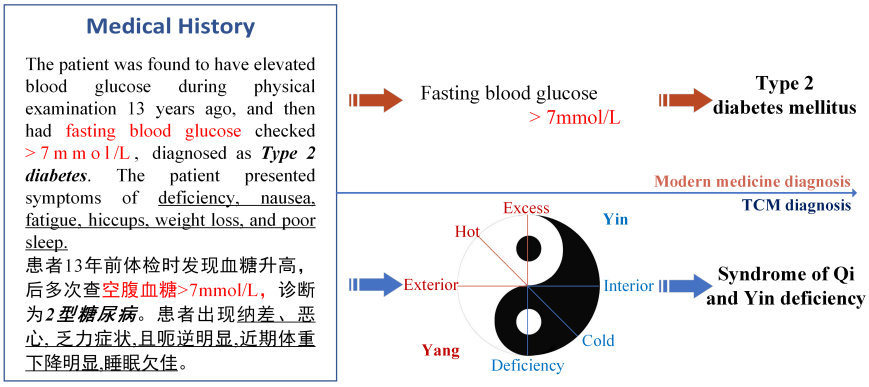


Figure 2: Different diagnostic processes of TCM and modern medicine for the same medical history.

As shown in Figure 2, given a medical history, modern medicine diagnoses the disease based on the level of fasting blood glucose, while TCM would map the various symptoms into a specific space with a unique coordinate system, analyze the latent causes, and combine them to determine a certain syndrome. Compared with the apparent numerical indicators of modern medicine, the concept of TCM is far more abstract and challenging to explain with modern medical theories.

However, TCM's difficult-to-describe nature does not mean that it has no value or rationality. In contrast, TCM has various complete and self-contained SD theories. Therefore, to explore TCM, we should not confine ourselves to the biomedical field. We may adopt NLP to explore TCM, which mainly consists of unstructured text. The linguistic characteristics may offer a scientific way to explain TCM theories. Therefore, in this paper, we present an SD dataset for further development.

2.2 Differences between ICD coding and Syndrome Differentiation

Automatic ICD coding is defined as assigning disease codes to Electronic Medical Records (EMR) , which is similar to TCM syndrome differentiation. Yet the two tasks are worlds apart in difficulty. Generally, the name of a patient's disease is directly recorded in EMR, and the task of the ICD coding is simply to normalize the names of these diseases in the manner of the ICD standard, without requiring a deep understanding of the context. For the example shown in Figure2, **Type 2 diabetes** has already described in the medical history so that ICD coding can be easily completed. While the syndrome differentiation not only requires collecting scattering evidence from the context through deep understanding, but also need to execute reliable and feasible inference, which brings a huge challenge to the model.

3 Related Works

There are three main streams of work related to this manuscript: medical dataset, natural language processing in syndrome differentiation and domain specific pre-trained language model.

	Medical system	Domain	# of syndromes	# of samples	Task Type	Is avaliable?	Language
This Work	Traditional Medicine	General	148	54,152	Class.,MRC	Yes	Chinese
Wang (2009)	Traditional Medicine	Liver Cirrhosis	3	406	Class.	No	Chinese
Zhang (2019)	Traditional Medicine	Stroke	3	654	Class.	No	Chinese
Wang (2018)	Traditional Medicine	Diabetes	12	1,180	Class.	No	Chinese
Pang (2020)	Traditional Medicine	AIDS	7	12,000	Class.	No	Chinese
Johnson (2016)	Modern Medicine	Critical Care	-	53,423	-	Yes	English
Stubbs (2015)	Modern Medicine	General	-	1,304	De-ID.	Yes	English
Dougan (2014)	Modern Medicine	General	-	6,892	DNR	Yes	English
Abacha (2019)	Modern Medicine	General	-	405;203;383	NLI;RQE;QA	Yes	English
Tian (2019)	Modern Medicine	General	-	46,731	MRC	Yes	Chinese

Table 1: Comparison of medical datasets in traditional and modern medicine. This table only includes textual data. The abbreviations in the table are defined as follows: classification (Class.), machine reading comprehension (MRC), de-identification (De-ID.), disease name recognition (DNR), natural language inference (NLI), recognizing question entailment (RQE), and question answering (QA).

3.1 Medical Datasets

In recent years, health record systems in hospitals have been moving towards digitalization and electronization, and a large amount of clinical data has been accumulated. To make more effective use of these data and provide better medical services, some studies led by MIMIC-III (Johnson et al., 2016) have shared these valuable data with medical researchers around the world (Stubbs et al., 2015; Doğan et al., 2014). Subsequently, with the development of AI, the domain characteristics of various studies have been combined to design various task-oriented datasets (Pampari et al., 2018; Li et al., 2020; Tian et al., 2019). These datasets have greatly promoted the development of AI in the medical field and have had a profound impact on society in terms of health and well-being.

However, as shown in Table 1, most of these publicly available datasets focus on modern medicine, there are far fewer datasets on traditional medicine. This is because, compared with traditional medicine, modern medicine has a rigorous, scientific, and standardized medical system, which can efficiently collect high-quality data. Furthermore, the standardization of traditional medicine is still in the development stage, which makes the collection and construction of relevant datasets extremely challenging. Thus the scarce TCM SD datasets has hindered the development of AI in this field. To alleviate this issue, we constructed the first large-scale, publicly available dataset for TCM SD.

3.2 Natural Language Processing (NLP) in Syndrome Differentiation

At present, most existing studies have treated SD as a multi-class classification task (i.e., taking the medical records as the input and output the predicted one from numerous candidate syndrome labels). Zhang (2019) used support vector machines to classify three types of syndromes for stroke patients. Zhang (2020a) also introduced an ensemble model consisting of four methods, a back-propagation neural network, the random forest algorithm, a support vector classifier, and the extreme gradient boosting method, to classify common diseases and syndromes simultaneously. Wang (2018) proposed a multi-instance, multi-task convolutional neural network (CNN) framework to classify 12 types of syndromes in 1,915 samples. Pang (2020) proposed a multilayer perceptron (MLP) model with an attention mechanism to predict the syndrome types of acquired immunodeficiency syndrome (AIDS). Similarly, Liu (2020) proposed a text-hierarchical attention network for 1,296 clinical records with 12 kinds of syndromes. However, these approaches only worked well for small-scale datasets. Our work established a series of strong baseline models and conducted comparisons on a larger-scale datasets.

3.3 Domain Specific Pre-trained Language Model

Large-scale neural language models pre-trained on unlabelled text has proved to be a successful approach for various downstream NLP tasks. A representative example is Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), which has become a foundation block for building task-specific NLP models. However, most works typically focus on pre-training in the general domain, while domain-specific pre-training has not received much attention. Table 2 summarizes common language

Model	Corpus	Domain	Language	Corpus Size
BERT (Devlin et al., 2018)	Wiki+Books	General	EN	3.3B tokens
RoBERTa-wwm (Cui et al., 2021)	Web Crawl	General	CN	5.4B tokens
MacBERT (Cui et al., 2020)	Web Crawl	General	CN	5.4B tokens
SciBERT (Beltagy et al., 2019)	Web Crawl	Science	EN	3.2B tokens
BioBERT (Lee et al., 2020)	PubMed	Medical	EN	4.5B tokens
ClinicalBERT (Alsentzer et al., 2019)	MIMIC	Medical	EN	0.5B tokens
BlueBERT (Peng et al., 2019)	PubMed+MIMIC	Medical	EN	4.5B tokens
PubMedBERT (Gu et al., 2021)	PubMed	Medical	EN	3.1B tokens
TCM-BERT* (Yao et al., 2019)	Web Crawl	Medical (TCM)	CN	0.02B tokens
ZY-BERT (Ours)	Web Crawl	Medical (TCM)	CN	0.4B tokens

Table 2: Summary of pre-training details for the various BERT models.

models pre-trained in either general domain or specific domain. In general, biomedical and science are mainstream fields of pre-training language model, but in the filed of TCM, there is no much work that has been conducted as for as we know.

The reasons may be two-fold. On the one hand, TCM lacks large-scale public text corpus, like Wikipedia and PubMed. We deal with this issue by presenting a corpus in TCM domain via crawling and collecting related documents from the websites and books. On the other hand, there is also a lack of downstream tasks that can verify the performance of the pre-training language model, thus we propose the syndrome differentiation task to measure its effectiveness.

To be noticed, an existing work already proposed a language model in the filed of TCM, named as TCM-BERT (Yao et al., 2019), but it did not undergo pre-training of large-scale corpus, but was only finetuned on small-scale nonpublic corpus (0.02B tokens). While, our work provide a more completed TCM-domain corpus (over 20 times larger) and verify its effectiveness during pre-training stage.

4 Benchmark and Methods

The TCM-SD benchmark that we collected contains over 65,000 real-world Chinese clinical notes. Table 3 presents an example. Specifically, each clinical note contains the following five components: **Medical history** is the critical information for completing SD. It mainly describes a patient's condition at admission; **Chief complaint** is a concise statement describing the main symptoms that appeared in the medical history; **Four diagnostic methods record (FDMR)** is a template statement consisting of four main TCM diagnostic methods: inspection, auscultation and olfaction, interrogation, and palpation; **ICD-10 index number and name** represents the name and corresponding unique ID of the patient's disease; **Syndrome name** is the syndrome of the current patient. However, the raw data could not be used directly for the SD task due to the lack of quality control. Therefore, a careful normalization was further conducted to preprocess the data.

4.1 Syndrome Normalization

Like ICD, TCM already has national standards for the classification of TCM diseases, named *Classification and Codes of Diseases and Zheng of Traditional Chinese Medicine* (GB/T15657-1995), which stipulates the coding methods of diseases and the zheng of TCM. However, TCM standardization is still in its early phase of development and faces inadequate publicizing and implementation (Wang et al., 2016). Some TCM practitioners still have low awareness and different attitudes toward TCM standardization, resulting in inconsistent naming methods for the same syndrome.

Therefore, based on the above issues, we accomplish syndrome normalization in two stages: merging and pruning.

Merging operation is mainly used in two cases. The first is cases in which the current syndrome has multiple names, and all appear in the dataset. For example, *syndrome of wind and heat* (风热证) and

Medical History
The patient began to suffer from repeated dizziness more than eight years ago, and the blood pressure measured in a resting-state was higher than normal many times. The highest blood pressure was 180/100 mmHg, and the patient was clearly diagnosed with hypertension. The patient usually took Nifedipine Sustained Release Tablets (20 mg), and the blood pressure was generally controlled, and dizziness occasionally occurred. Four days before the admission, the patient's dizziness worsened after catching a cold, accompanied by asthma, which worsened with activity. Furthermore, the patient coughed yellow and thick sputum. The symptoms were not significantly relieved after taking antihypertensive drugs and antibiotics, and the blood pressure fluctuated wildly. On admission, the patient still experienced dizziness, coughing with yellow mucous phlegm, chills, no fever, no conscious activity disorder, no palpitations, no chest tightness, no chest pain, no sweating, a weak waist and knees, less sleep and more dreams, forgetfulness, dry eyes, vision loss, red hectic cheeks, and dry pharynx, five upset hot, no nausea and vomiting, general eating and sleeping, and normal defecation.
患者8年余前开始反复出现头晕，多次于静息状态下测血压高于正常，最高血压180/100 mmHg，明确诊断为高血压，平素服用硝苯地平缓释片20 mg，血压控制一般，头晕时有发作。此次入院前4天受凉后头晕再发加重，伴憋喘，动则加剧，咳嗽、咳黄浓痰，自服降压药、抗生素症状缓解不明显，血压波动大。入院时：仍有头晕，咳嗽、咳黄粘痰，畏寒，无发热，无意识活动障碍，无心慌、胸闷，无胸痛、汗出，腰酸膝软，少寐多梦，健忘，两目干涩，视力减退，颧红咽干，五心烦热，无恶心呕吐，饮食睡眠一般，二便正常。
Chief Complaint
Repeated dizziness for more than eight years, aggravated with asthma for four days.
反复头晕8年余，加重伴喘憋4天。
Four Diagnostic Methods Record
Mind: clear; spirit: weak; body shape: moderate; speech: clear,..., tongue: red with little coating; pulse: small and wiry.
神志清晰，精神欠佳，形体适中，语言清晰, ... , 舌红少苔，脉弦细。
ICD-10 Name and ID: Vertigo (眩晕病) BNG070
Syndrome Name: Syndrome of Yin deficiency and Yang hyperactivity 阴虚阳亢证

External Knowledge Corpus:
A syndrome with Yin deficiency and Yang hyperactivity is a type of TCM syndrome. It refers to Yin liquid deficiency and Yang loss restriction and hyperactivity. Common symptoms include dizziness, hot flashes, night sweats, tinnitus, irritability, insomnia, red tongue, less saliva, and wiry pulse. It is mainly caused by old age, exposure to exogenous heat for a long period, the presence of a serious disease for a long period, emotional disorders, and unrestrained sexual behavior. Common diseases include insomnia, vertigo, headache, stroke, deafness, tinnitus, premature ejaculation, and other diseases.
阴虚阳亢证，中医病证名。是指阴液亏虚，阳失制约而偏亢，以头晕目眩，潮热盗汗，头晕耳鸣，烦躁失眠，舌红少津，脉细数为常见证的证候，多因年老体衰，外感热邪日久，或大病久病迁延日久，情志失调，房事不节等所致。常见于不寐、眩晕、头痛、中风、耳聋耳鸣、早泄等疾病中。

Table 3: A sample clinical record from the TCM-SD dataset with related external knowledge. An explicit match between the medical history and external knowledge is marked in blue, while the text in orange is an example of an implicit match that required temporal reasoning.

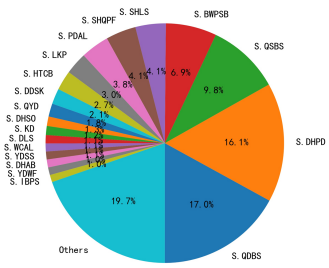
syndrome of wind and heat attacking the external (风热外袭证) belong to the same syndrome, and we would merge them into one unified name. In this case, we used the national standards for screening. Another is that the current syndrome name does not exist in a standardized form. Therefore, we recruited experts to conduct syndrome differentiation according to the specific case clinical records and finally merge the invalid syndromes into standard syndromes. For example, *syndrome of spleen and kidney yang failure* (脾肾阳衰证) would be merged into *syndrome of spleen and kidney yang deficiency* (脾肾阳虚证).

Pruning operation is mainly applied to syndromes with non-standard names that experts fail to differentiate due to vague features. In addition, since syndrome names are hierarchically graded, we pruned out syndromes with higher grades to ensure that the syndromes that appear in the current dataset are the most basic grade, that is the most specific ones that determine the subsequent treatment. For example, *syndrome of wind and cold* (风寒证) is a high-grade syndrome, and its clinical manifestations can be a *syndrome of exterior tightened by wind-cold* (风寒束表证) or *syndrome of wind-cold attacking lung* (风寒袭肺证); each has different symptoms and treatment methods.

4.2 Dataset Statistics

After normalization, the number of syndromes in the dataset was reduced from the original 548 categories to 244. Considering that some syndromes are infrequent, we further filtered out syndrome categories containing fewer than 10 samples when partitioning the dataset. Then, the processed dataset with 148 syndrome categories and 54,152 samples was divided into a training set, a development (Dev) set, and a

	Train	Dev	Test	Total
# Samples	43,180	5,486	5,486	54,152
# Vovabulary size	2,964	2,393	2,377	3,039
# Chief complaint tokens	13.4	13.5	13.5	13.4
# Medical history tokens	249.2	272.3	258.7	252.5
# FDMR tokens	131.9	132.3	132.2	132
# Syndrome tokens		4.9		
# Syndrome categories		148		
# Knowledge		1,027		
# Knowledge tokens		99.1		



Abbreviation	Full Name	证名
S.QDBS	syndrome of qi deficiency with blood stasis	气虚血瘀证
S.DHPD	syndrome of dampness and heat pouring downwards	湿热下注证
S.QSBS	syndrome of qi stagnation and blood stasis	气滞血瘀证
S.BWPSB	syndrome of blockade of wind-phlegm-static blood	风痰瘀阻证
S.SHLS	syndrome of stagnant heat in the liver and stomach	肝胃郁热证
S.SHQPF	syndrome of stagnated healthy qi and pathogenic factors	正虚瘀结证
S.PDAL	syndrome of phlegm and dampness accumulating in the lung	痰湿蕴肺证
S.LKP	syndrome of liver and kidney depletion	肝肾亏虚证
S.HTCB	syndrome of heat toxin congestion and binding	热毒蕴结证
S.DDSK	syndrome of dual deficiency of spleen and kidney	脾肾两虚证
S.QYD	syndrome of qi and yin deficiency	气阴两虚证
S.DHSO	syndrome of dampness and heat stagnation and obstruction	湿热瘀阻证
S.KD	syndrome of kidney deficiency	肾虚证
S.DLS	syndrome of disharmony of liver and stomach	肝胃不和证
S.WCAL	syndrome of wind and cold attacking the lung	风寒袭肺证
S.YDSS	syndrome of yang deficiency in spleen and stomach	脾胃虚寒证
S.DHAB	syndrome of dampness and heat accumulation and binding	湿热蕴结证
S.YDWF	syndrome of yang deficiency with water flooding	阳虚水泛证

Figure 3: The characteristics and syndrome distribution in the dataset.

test set with a ratio of 8:1:1. The dataset characteristics and syndrome distribution shown in Figure 3.

Since the data were collected from real-world scenarios, the distribution of syndromes was inevitably unbalanced, leading to a significant gap between the number of rare syndromes and the number of the common ones. The subsequent experiments demonstrate the challenges brought by long-tail distribution issues, and we show that this issue can be mitigated by introducing external knowledge and domain-specific pre-training.

4.3 External Knowledge

Current clinical records do not contain any relevant knowledge about the target syndromes, which causes models to have to rely on remembering patterns to complete the task. Therefore, we constructed an external unstructured knowledge corpus encompassing 1,027 types of TCM syndromes by web crawling for information on all the TCM syndromes on the online[1]. Specifically, the knowledge of each syndrome consisted of three parts: the cause of the syndrome, the main manifestations, and common related diseases. Table 3 shows an example. We demonstrate the effectiveness of this knowledge in the experimental section.

4.4 ZY-BERT

In general, ZY-BERT differs with TCM-BERT in two main parts: data and pre-training task.

First, the scale and quality of unlabelled text corpus directly affect the performance of pre-trained language models. Previous work TCM-BERT (Yao et al., 2019) directly used clinical records as pre-training corpus, resulting in monotonic data type and limited corpus size, which could not meet the needs of large-scale pre-training language model. To deal with this issue, we collected unlabelled data varies in different types from the TCM related websites, including books, articles from websites and academic papers from China National Knowledge Infrastructure (CNKI), counting over 400 million tokens.

Furthermore, the previous work TCM-BERT adopts char masking (CM) and next sentence prediction (NSP) as the pre-training tasks. However, Chinese words usually consist of multiple characters and masking single character might destroy the meaning of the whole word. For example, the word phrase *Yang Deficiency(*阳虚*)* consists of two characters. Thus, we borrowed the idea of Whole Word Masking from Cui (2021) and replace NSP with it , which could add challenges to the model training process and allow the model to learn more complex linguistic features.

Finally, the pre-trained language model consists of 24 Transformer layers, with input dimensionality of 1024. Each transformer contains 16 attention heads. Then we trained the model 300K steps with a maximum learning rate 5e-5 and a batch size of 256. Other hyperparameters and pre-training details are kept same as the ones used in Liu (2019).

5 Experiments

We selected the multi-class classification task as the primary form of SD to directly compare the performances of the existing models against the TCM-SD dataset, and used the accuracy and Macro-F1 as evaluation metrics. Specifically, the chief complaint and medical history were concatenated as the inputs,

[1] www.dayi.org.cn

Method	Dev				Test			
	Acc.	Macro-F1	Macro-R	Macro-P	Acc.	Macro-F1	Macro-R	Macro-P
DT	59.42%	20.68%	21.33%	21.52%	59.10%	21.67%	22.38%	22.20%
SVM	77.63%	32.13%	29.56%	43.10%	78.53%	36.37%	32.98%	49.35%
BiLSTM	69.30%	17.53%	15.08%	14.76%	69.65%	15.15%	15.65%	17.08%
BiGRU	73.57%	19.53%	20.12%	21.81%	74.43%	20.93%	21.90%	23.76%
CNN	77.56%	31.79%	30.39%	37.99%	78.58%	32.83%	31.29%	39.19%
BERT	79.44%	34.18%	34.12%	38.00%	80.17%	35.45%	34.99%	42.00%
distilBERT	79.09%	36.07%	36.62%	38.13%	80.46%	40.24%	39.99%	45.84%
ALBERT	79.62%	37.88%	37.65%	41.94%	80.51%	40.50%	39.57%	46.54%
RoBERTa	80.81%	43.18%	42.55%	47.68%	**82.26%**	47.55%	45.72%	54.15%
TCM-BERT	79.48%	37.84%	37.60%	42.00%	80.55%	41.58%	40.91%	48.47%
ZY-BERT(ours)	**81.43%**†	**49.47%**†	**48.89%**†	**54.08%**†	82.19%†	**51.01%**†	**49.42%**†	**57.70%**†

Table 4: Performance for the classification task. The marker † refers to p-value <0.01.

i.e. *[CLS] Chief Complaint [SEP] Medical History [SEP]*, where [CLS] and [SEP] are special tokens used for classification and separation. Then the model predicts the target syndromes from 148 candidate labels based on the representation of [CLS] token.

5.1 Baseline

The baseline methods we used consisted of four types: statistical methods, classical neural-network-based (NN-based) methods, language-model-based (LM-based) methods and domain-specific LM-based methods.

Statistical methods. These methods were the decision tree (DT) and support vector machine (SVM) methods. These two statistical methods have been widely used in previous studies on SD.

Classical NN-based methods. These methods included a Bi-LSTM (Schuster and Paliwal, 1997), a Bi-GRU (Qing et al., 2019), and a two-layer CNN (Kim, 2014). Word embeddings were retrieved from the Chinese version of BERT (Cui et al., 2021).

LM-based methods. These methods included several popular LMs, such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), distillBERT (Sanh et al., 2019), and ALBERT (Lan et al., 2019). These models concatenate multiple pieces of text with special tokens as inputs, make classifications based on the hidden states of the first token, or determine the start and end of the answer by training two classifiers.

Domain-specific LM-based methods. These methods are similar with LM-based ones but usually pre-trained on domain-specific corpus rather than general domain corpus. TCM-BERT (Yao et al., 2019) and our proposed ZY-BERT are the two LM used in this manuscripts.

5.2 Main Results

Table 4 presents the performances of all the methods for the classification task. Generally, all the methods had good accuracy, which demonstrated that the models were effective at fitting when enough examples were supplied. However, each syndrome in the TCM-SD dataset should have the same importance. Thus, the Macro-F1 is a more accurate metric to evaluate the performances of the models. The Macro-F1 scores achieved by the models were much lower than the accuracy, which demonstrated the challenges of the imbalanced TMC-SD datasets.

Moreover, the statistical methods achieved better scores than the classical NN-based methods. This is because the structures designed for focusing on contextualized representations, such as the Bi-LSTM and Bi-GRU networks, were not good at capturing features, and the performances were worse. In contrast, the SVM and CNN methods were good at extracting local features and obtained better scores. Nonetheless, the language models still achieved the highest scores, demonstrating the effectiveness of the large-scale corpus pre-training.

Method	Dev				Test			
	EM	Macro-F1	Macro-R	Macro-P	EM	Macro-F1	Macro-R	Macro-P
Medical History + All Syndromes								
BERT	77.27%	40.71%	41.10%	43.26%	78.20%	45.60%	45.32%	50.15%
RoBERTa	78.71%	45.09%	44.30%	49.38%	80.42%	47.57%	46.42%	51.89%
Medical History + Five Syndromes								
BERT	95.59%	77.12%	76.32%	81.04%	95.83%	82.33%	81.35%	86.34%
RoBERTa	95.75%	79.16%	78.74%	82.79%	95.86%	84.42%	84.92%	86.74%
Medical History + Five Syndromes + Knowledge								
BERT	95.24%	81.21%	81.33%	84.61%	96.06%	85.15%	84.48%	87.92%
RoBERTa	**95.33%**	**81.53%**	**81.76%**	**84.49%**	**96.26%**	**85.88%**	**85.59%**	**89.09%**

Table 5: Performance with the machine reading comprehension (MRC) task.

6 Discussion

6.1 Effect of Domain-specific Pre-training

The last two rows in Table 4 indicates the effects of domain-specific pre-training. To be noticed, our proposed ZY-BERT achieved the astonishing performance improvement and mitigated long-tail distribution issue greatly. On the one hand, Macro-F1 score achieved by ZY-BERT is over 4% larger than that achieved by RoBERTa, demonstrating the effectiveness of large-scale domain-specific corpus for domain-specific tasks. On the other hand, ZY-BERT also achieves over 10% Macro-F1 scores higher than the previous domain-specific model TCM-BERT, which proves the quality and reliability of the TCM domain corpus constructed by us.

6.2 Effect of Knowledge

To testify the effectiveness of the external knowledge corpus, we leveraged knowledge into the model by concatenating the relevant syndrome knowledge with the medical history. However, due to the length limits of the language models, feeding knowledge of all syndromes into the model is infeasible under classification setting. Thus we converted the task from classification to extractive MRC, and designed the following three settings shown in Table 5 to evaluate the significance of the knowledge.

Firstly, we concatenated the original inputs with all syndrome names, and asked the model to extract the target syndrome spans from the context. The competitive results shown between MRC and classification tasks demonstrated that the model had a consistent ability among different task formats without external knowledge. Then we further conducted two groups of experiments. In the first group, instead of concatenating all syndrome names, we only included five syndromes, where one was the target syndrome and the other four were randomly selected. In the second group, we appended the corresponding knowledge for each syndrome selected in the first group. The superior results achieved by the latter group demonstrate the importance of knowledge.

However, the outstanding performance, either with knowledge or without knowledge, was mainly due to the fact that we manually narrowed down the search range to five syndromes. We used the term frequency–inverse document frequency (TFIDF) to search for relevant knowledge from the knowledge corpus based on medical history, and P@5 was only 3.94%. Thus, knowledge is essential, but finding it is difficult.

6.3 Ablation Study

Table 6 shows the results of the ablation study on the TCD-SD dataset. Removing either the medical history or the chief complaint resulted in lower performances, especially if only the chief complaint was taken into account. This was because the chief complaint was typically too short to include sufficient features for classification. However, the chief complaint and medical history complemented each other in a coarse-to-fine fashion.

Method	Dev				Test			
	Acc.	Macro-F1	Macro-R	Macro-P	Acc.	Macro-F1	Macro-R	Macro-P
Only Chief Complaint								
BERT	70.56%	23.15%	26.34%	26.34%	71.58%	24.08%	25.38%	24.08%
RoBERTa	71.36%	28.55%	28.85%	33.13%	72.91%	30.78%	34.54%	34.54%
Only Medical History								
BERT	79.40%	33.50%	33.46%	37.90%	79.62%	35.57%	35.13%	42.18%
RoBERTa	79.80%	41.40%	40.12%	45.38%	81.83%	45.19%	43.03%	53.78%
Chief Complaint + Medical History								
BERT	79.44%	34.18%	34.12%	38.00%	80.17%	35.45%	34.99%	42.00%
RoBERTa	**80.81%**	**43.18%**	**42.55%**	**47.68%**	**82.26%**	**47.55%**	**45.72%**	**54.15%**

Table 6: Ablation study on the TCM-SD dataset.

6.4 Error Analysis

By analyzing the error cases, we found that the vast majority of errors occurred in the category with few samples, and fitting only according to the data distribution was still the most significant issue. Except for algorithmic problems, we concluded that there were three main error types:

Complex Reasoning. As shown in Table 3, besides the explicit match marked in blue, there was an implicit match marked in orange that required temporal reasoning. Additionally, the task also included complex reasoning, such as numerical reasoning, spatial reasoning and negative reasoning.

Incomplete Knowledge. The current models do not take into account the concepts that arise from the SD task, such as Yin and Yang. Therefore, the models do not know how to map the symptoms into the special coordinate system of the TCM diagnostics system.

Out-Of-Vocabulary. In the clinical records, there exists not only academic medical-related terms but also various rare traditional characters in TCM, which impeded the understanding of the context.

7 Conclusions

This paper introduced a meaningful task, SD, in TCM and its connection with NLP and presented the first public large-scale benchmark of SD: TCM-SD. Furthermore, a knowledge corpus supporting the model understanding and the large-scale TCM domain corpus for pre-training were constructed. Moreover, one domain-specific pre-training language model named as ZY-BERT was proposed. The experiments on this dataset demonstrated the challenges, the inadequacy of existing models, the importance of knowledge and the effectiveness of domain-specific pre-training. This work can greatly promote the internationalization and modernization of TCM, the proposed benchmark and associated baseline models provide a basis for subsequent research.

Acknowledgements

This work is supported by funds from the National Natural Science Foundation of China (No.U21B2009). The data used in this paper were only routine diagnosis and treatment data of patients, excluding any personal information of the patients (such as name, age, and telephone number). This study did not interfere with normal medical procedures or create an additional burden to medical staff, and no experiments were conducted on patients. **All the data have been desensitized**. Therefore, this paper does not involve ethical issues and waives the requirement of individual patient consent. We public TCM-SD dataset, TCM-domain corpus and ZY-BERT model at https://github.com/Borororo/ZY-BERT. We thank the reviewers for their helpful and constructive comments. And we thank M.D. Yonglan Zhou for her insightful and professional suggestions.

References

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379.

Emily Alsentzer, John Murphy, William Boag, et al. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Pengfei Cao, Chenwei Yan, Xiangling Fu, et al. 2020. Clinical-coder: Assigning interpretable ICD-10 codes to Chinese clinical notes. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 294–301, Online, July. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, et al. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online, November. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Jinyue Feng, Chantal Shaib, and Frank Rudzicz. 2020. Explainable clinical decision support from text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1478–1489, Online, November. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, et al. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Yun He, Ziwei Zhu, Yin Zhang, et al. 2020. Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614, Online, November. Association for Computational Linguistics.

Charles Jochim and Léa Deleris. 2017. Named entity recognition in the medical domain with constrained CRF models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 839–849, Valencia, Spain, April. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, et al. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, et al. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, et al. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Dongfang Li, Baotian Hu, Qingcai Chen, et al. 2020. Towards medical machine reading comprehension with structural knowledge and plain text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1427–1438, Online, November. Association for Computational Linguistics.

Ziqing Liu, Enwei Peng, Shixing Yan, et al. 2018. T-know: A knowledge graph-based question answering and infor-mation retrieval system for traditional Chinese medicine. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 15–19, Santa Fe, New Mexico, August. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ziqing Liu, Haiyang He, Shixing Yan, et al. 2020. End-to-end models to imitate traditional Chinese medicine syndrome differentiation in lung cancer diagnosis: Model development and validation. *JMIR Medical Informatics*, 8(6):e17821.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Others. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium, October-November. Association for Computational Linguistics.

Huaxin Pang, Shikui Wei, Yufeng Zhao, et al. 2020. Effective attention-based network for syndrome differentiation of AIDS. *BMC Medical Informatics and Decision Making*, 20(1):1–10.

Cecilia Panigutti, Alan Perotti, André Panisson, et al. 2021. Fairlens: Auditing black-box clinical decision support systems. *Information Processing & Management*, 58(5):102657.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy, August. Association for Computational Linguistics.

Li Qing, Weng Linhong, and Ding Xuehai. 2019. A novel neural network-based method for medical text classification. *Future Internet*, 11(12):255.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19.

Yuanhe Tian, Weicheng Ma, Fei Xia, et al. 2019. ChiMed: A Chinese medical corpus for question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 250–260, Florence, Italy, August. Association for Computational Linguistics.

Youyou Tu. 2016. Artemisinin—A gift from traditional Chinese medicine to the world (nobel lecture). *Angewandte Chemie International Edition*, 55(35):10210–10226.

Yan Wang, Lizhuang Ma, and Ping Liu. 2009. Feature selection and syndrome prediction for liver cirrhosis in traditional Chinese medicine. *Computer Methods and Programs in Biomedicine*, 95(3):249–257.

Juan Wang, Yi Guo, and Gui Lan Li. 2016. Current status of standardization of traditional Chinese medicine in china. *Evidence-Based Complementary and Alternative Medicine*, 2016.

Zeyuan Wang, Shiding Sun, Josiah Poon, et al. 2018. CNN based multi-instance multi-task learning for syndrome differentiation of diabetic patients. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1905–1911. IEEE.

Yang Yang, Md Sahidul Islam, Jin Wang, et al. 2020. Traditional Chinese medicine in the treatment of patients infected with 2019-new coronavirus (sars-cov-2): A review and perspective. *International journal of biological sciences*, 16(10):1708.

Liang Yao, Zhe Jin, Chengsheng Mao, et al. 2019. Traditional chinese medicine clinical records classification with bert and domain specific corpora. *Journal of the American Medical Informatics Association*, 26(12):1632–1636.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: Multiple synonyms matching network for automatic icd coding. *arXiv preprint arXiv:2203.01515*.

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020. Clinical reading comprehension: A thorough analysis of the emrQA dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4474–4486, Online, July. Association for Computational Linguistics.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, et al. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online, November. Association for Computational Linguistics.

Dongxue Zhang, Zhichao Gan, and Zhihui Huang. 2019. Study on classification model of traditional Chinese medicine syndrome types of stroke patients in convalescent stage based on support vector machine. In *2019 10th International Conference on Information Technology in Medicine and Education (ITME)*, pages 205–209. IEEE.

Hong Zhang, Wandong Ni, Jing Li, et al. 2020a. Artificial intelligence–based traditional Chinese medicine assistive diagnostic system: Validation study. *JMIR medical informatics*, 8(6):e17608.

Leyin Zhang, Jieru Yu, Yiwen Zhou, et al. 2020b. Becoming a faithful defender: Traditional Chinese medicine against coronavirus disease 2019 (covid-19). *The American journal of Chinese medicine*, 48(04):763–777.