

同行专家业内评价意见书编号：20250854379

附件1

浙江工程师学院（浙江大学工程师学院） 同行专家业内评价意见书

姓名： 宋紫君

学号： 22260005

申报工程师职称专业类别（领域）： 电子信息

浙江工程师学院（浙江大学工程师学院）制

2025年03月21日

填表说明

一、本报告中相关的技术或数据如涉及知识产权保护、军工项目保密等内容，请作脱密处理。

二、请用宋体小四字号撰写本报告，可另行附页或增加页数，A4纸双面打印。

三、表中所涉及的签名都必须用蓝、黑色墨水笔，亲笔签名或签字章，不可以打印代替。

四、同行专家业内评价意见书编号由工程师学院填写，编号规则为：年份4位+申报工程师职称专业类别(领域)4位+流水号3位，共11位。

一、个人申报

（一）基本情况【围绕《浙江工程师学院（浙江大学工程师学院）工程类专业学位研究生工程师职称评审参考指标》，结合该专业类别(领域)工程师职称评审相关标准，举例说明】

1. 对本专业基础理论知识和专业技术知识掌握情况(不少于200字)

我对电子信息专业的基础理论和专业技术知识有着扎实打的掌握。在理论方面，我系统学习了信号处理、模式识别、计算机视觉和深度学习等相关课程，具备较强的数学建模和算法分析能力。在专业技术方面，我熟练掌握Python编程，熟悉PyTorch、TensorFlow等深度学习框架，并能够应用卷积神经网络（CNN）、注意力机制等技术进行医学图像分析。在我的研究实践课题“基于深度学习的中医舌象辨识系统”中，我结合图像处理与深度学习方法，完成了舌象图像的特征提取与分类，优化了模型性能，并对舌色、舌苔等关键特征进行了有效识别。通过这一研究，我不仅加深了对本专业知识的理解，还提升了实践能力和创新能力。

2. 工程实践的经历(不少于200字)

在工程实践过程中，我结合图像处理技术与深度学习方法，构建了一套自动化中医舌诊系统，成功实现了系统的部署和落地，提升了舌诊的客观性与标准化水平。首先，我通过医院合作与公开数据收集，建立标准化的舌象数据集，并利用YOLOv5s进行舌体定位，同时采用等效圆检测法与灰度世界法校正图像偏色。其次，在舌体分割方面，我提出了基于边缘检测的T-EdgeNet模型，融合空洞卷积、注意力机制与边缘感知聚合模块，提高了舌体边缘检测与分割的精度。随后，在舌象分类与证型辨识方面，我采用多标签分类方法，结合嵌入空间注意力机制与融合共享机制进行舌象特征提取，并利用随机森林算法构建舌象特征与中医证型的关系，提高辨识准确性。最终，我完成了患者端在线诊断、报告下载及医生端结果核准等功能开发，实现了系统的应用落地。

3. 在实际工作中综合运用所学知识解决复杂工程问题的案例（不少于1000字）

在实际工作中综合运用所学知识解决复杂工程问题的案例

在电子信息工程领域，深度学习与图像处理技术的融合正推动着医学辅助诊断的智能化发展。针对传统中医舌诊依赖医生主观判断、难以实现标准化的痛点，我综合运用所学的深度学习、计算机视觉和模式识别等知识，构建了一套基于深度学习的自动化中医舌象辨识系统，实现了舌体特征的精准提取、分类以及证型辨识，为中医舌诊提供了一种客观化、标准化的智能辅助诊断方案。该系统的构建涉及多个复杂的工程问题，包括数据采集与预处理、舌体区域检测与分割、多标签分类及证型辨识算法设计，以及完整的系统架构与前后端开发。我在项目中综合运用了图像处理、深度学习模型设计、优化算法以及软件工程等多方面的技术，有效解决了多个关键难点。

1、数据采集与预处理方面：为了保证舌象识别系统的鲁棒性和泛化能力，我首先面临的挑战是如何构建一个高质量的标准化舌象数据集。数据来源主要包括与医院合作采集的真实患者舌象数据，以及从公开数据库获取的医学舌象图像。然而，这些数据在拍摄条件、色彩均衡、光照环境等方面存在较大差异，直接用于训练可能会导致模型偏差。因此，我运用了图像处理和颜色校正技术对数据进行标准化处理。在舌体区域检测方面，我使用了基于YOLOv5s的目标检测方法来精确定位舌部区域，并结合等效圆检测法校验图像的偏色情况。由于中医舌诊高度依赖于舌象颜色信息，我进一步采用了标准差加权的灰度世界法对偏色图像进行色彩校正，使得数据集中的舌象颜色尽可能接近真实情况。这一系列数据预处理操作提升了舌象数据的质量，为后续的舌体分割与特征提取奠定了基础。

2、舌体分割模型的设计与优化方面：由于舌体边缘往往存在模糊过渡、光照反射等问题，

传统的分割方法（如U-Net）在复杂环境下表现不佳。为此，我提出了一种基于边缘检测的舌体分割模型——T-EdgeNet，该模型融合了多种深度学习方法以提升舌体分割的准确性。T-EdgeNet由三个关键模块组成：（1）基于空洞卷积的残差语义提取模块，用于提取舌体的深层特征，同时保持高分辨率信息；（2）基于门控注意力机制的边缘检测模块，增强舌体边缘区域的检测能力，有效减少模糊边界带来的分割误差；（3）边缘感知聚合模块，将边缘信息与语义信息进行融合，提高分割结果的完整性。实验结果表明，该模型在多个舌象数据集上均取得了优于传统分割方法的性能，特别是在边缘区域的处理上，能够更好地保留舌体轮廓细节。

3、舌象分类与证型辨识方面：在舌象分析中，一个关键挑战是如何同时提取多种舌象特征（如舌色、舌苔、裂纹等），并将其与中医证型关联。针对这一问题，我采用了多标签分类模型，结合深度学习与传统机器学习的方法，提高了舌象分类与证型辨识的准确性。模型的训练分为两个阶段：第一阶段，针对舌象特征的多分类问题，我设计了基于嵌入空间注意力机制的深度学习模型，分别提取舌象的纹理特征和颜色特征。其中，纹理特征的提取采用了基于CNN的多分类网络，而颜色特征的提取则通过融合共享机制提高不同特征之间的信息传递。第二阶段，我结合中医先验知识，采用随机森林算法建立舌象特征与证型之间的映射关系，提高模型的可解释性。这种方法使得证型辨识不仅依赖深度学习的特征提取能力，还能够利用传统机器学习的透明决策过程，使医生更容易理解模型的判别依据。实验结果表明，该方法在多个舌象分类任务中取得了较高的准确率，同时提升了模型的可解释性，为智能化中医诊断提供了可靠的支持。

4、自动化舌诊系统的搭建与应用方面：在算法研究的基础上，我进一步完成了自动化舌诊系统的前后端开发，并将其应用于实际场景。该系统包括患者端和医生端两个模块，实现了舌象图像上传、在线诊断、报告生成与医生审核等功能。在患者端，用户可以通过手机或电脑上传舌象图像，系统会自动进行舌体检测、特征提取和证型辨识，并生成诊断报告。在医生端，医生可以查看患者的舌象分析结果，对系统给出的诊断进行核准，并结合自身经验调整诊断意见。整个系统采用前后端分离架构，前端基于Vue.js开发，后端采用Flask与PyTorch结合的方式进行模型部署，并通过数据库存储用户信息与诊断结果。此外，我优化了系统的推理速度，使其能够在普通计算机上流畅运行，满足实际应用需求。经过测试，该系统在多个舌象数据集上表现稳定，能够有效辅助中医医生进行舌诊，提高诊断的客观性和标准化程度。

在本项目中，我综合运用了电子信息专业的多项核心技术，包括深度学习、图像处理、模式识别、机器学习以及软件开发等，成功解决了舌象数据采集、舌体分割、特征提取、证型辨识等多个复杂的工程问题。通过提出T-

EdgeNet模型优化舌体分割精度，设计多标签分类方法提高舌象特征提取能力，并结合前后端技术实现自动化舌诊系统。本项目的研究不仅提升了我在深度学习与图像处理领域的工程实践能力，也加深了我对电子信息工程在医学影像分析方向的应用理解。通过解决实际工程问题，我积累了从数据采集、算法设计到系统部署的完整项目经验，为今后进一步研究和开发智能医学影像系统奠定了坚实基础。

(二) 取得的业绩(代表作)【限填3项, 须提交证明原件(包括发表的论文、出版的著作、专利证书、获奖证书、科技项目立项文件或合同、企业证明等)供核实, 并提供复印件一份】

1. 公开成果代表作【论文发表、专利成果、软件著作权、标准规范与行业工法制定、著作编写、科技成果获奖、学位论文等】

成果名称	成果类别 [含论文、授权专利(含发明专利申请)、软件著作权、标准、工法、著作、获奖、学位论文等]	发表时间/授权或申请时间等	刊物名称/专利授权或申请号等	本人排名/总人数	备注
一种基于边缘门控机制的舌象图像分割方法	发明专利申请	2024年11月01日	申请号: 202411548521.0	2/3	已进入实审阶段
A BERT-Based Named Entity Recognition Method of Warm Disease in Traditional Chinese Medicine	会议论文	2023年08月11日	ICIEA2023	1/5	EI会议收录


2. 其他代表作【主持或参与的课题研究项目、科技成果应用转化推广、企业技术难题解决方案、自主研发设计的产品或样机、技术报告、设计图纸、软课题研究报告、可行性研究报告、规划设计方案、施工或调试报告、工程实验、技术培训教材、推动行业发展中发挥的作用及取得的经济社会效益等】

(三) 在校期间课程、专业实践训练及学位论文相关情况	
课程成绩情况	按课程学分核算的平均成绩： 87 分
专业实践训练时间及考核情况(具有三年及以上工作经历的不作要求)	累计时间： 1 年(要求1年及以上) 考核成绩： 82 分
本人承诺	
<p>个人声明：本人上述所填资料均为真实有效，如有虚假，愿承担一切责任，特此声明！</p> <p style="text-align: right;">申报人签名：宋紫君</p>	



22260005

二、日常表现考核评价及申报材料审核公示结果

<p>日常表现 考核评价</p>	<p>非定向生由德育导师考核评价、定向生由所在工作单位考核评价： <input checked="" type="checkbox"/>优秀 <input type="checkbox"/>良好 <input type="checkbox"/>合格 <input type="checkbox"/>不合格 德育导师/定向生所在工作单位分管领导签字（公章）：  2015年3月21日</p>
<p>申报材料 审核公示</p>	<p>根据评审条件，工程师学院已对申报人员进行材料审核（学位课程成绩、专业实践训练时间及考核、学位论文、代表作等情况），并将符合要求的申报材料在学院网站公示不少于5个工作日，具体公示结果如下： <input type="checkbox"/>通过 <input type="checkbox"/>不通过（具体原因： 工程师学院教学管理办公室审核签字（公章）：) 年 月 日</p>

浙江大学研究生院
攻读硕士学位研究生成绩单

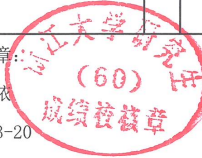
学号: 22260005	姓名: 宋紫君	性别: 女	学院: 工程师学院	专业: 电子信息	学制: 2.5年						
毕业时最低应获: 26.0学分		已获得: 30.0学分		入学年月: 2022-09	毕业年月:						
学位证书号:			毕业证书号:		授予学位:						
学习时间	课程名称	备注	学分	成绩	课程性质	学习时间	课程名称	备注	学分	成绩	课程性质
2022-2023学年秋季学期	新时代中国特色社会主义思想理论与实践		2.0	94	专业学位课	2022-2023学年秋冬学期	工程伦理		2.0	80	专业学位课
2022-2023学年秋季学期	研究生英语能力提升		1.0	免修	跨专业课	2022-2023学年秋冬学期	智能工业机器人及其应用		3.0	88	专业选修课
2022-2023学年秋季学期	研究生英语		2.0	免修	专业学位课	2022-2023学年春季学期	多相流检测技术		1.0	88	跨专业课
2022-2023学年秋季学期	研究生英语基础技能		1.0	免修	公共学位课	2022-2023学年春季学期	自然辩证法概论		1.0	83	专业学位课
2022-2023学年秋季学期	工程技术创新前沿		1.5	90	专业学位课	2022-2023学年春季学期	科技创新案例探讨与实战		2.0	85	专业选修课
2022-2023学年冬季学期	产业技术发展前沿		1.5	90	专业学位课	2022-2023学年春夏学期	高阶工程认知实践		3.0	88	专业学位课
2022-2023学年秋冬学期	研究生论文写作指导		1.0	94	专业选修课	2022-2023学年春夏学期	人工智能制造技术		3.0	91	专业学位课
2022-2023学年秋冬学期	数据分析的概率统计基础		3.0	90	专业选修课		硕士生读书报告		2.0	通过	

说明: 1. 研究生课程按三种方法计分: 百分制, 两级制(通过、不通过), 五级制(优、良、中、及格、不及格)。
2. 备注中“*”表示重修课程。

学院成绩校核章:

成绩校核人: 张梦依

打印日期: 2025-03-20





(12) 发明专利申请

(10) 申请公布号 CN 119359751 A

(43) 申请公布日 2025.01.24

(21) 申请号 202411548521.0

G06N 3/0985 (2023.01)

(22) 申请日 2024.11.01

G06T 5/60 (2024.01)

G06V 10/80 (2022.01)

(71) 申请人 浙江大学

地址 310058 浙江省杭州市西湖区余杭塘路866号

(72) 发明人 刘之涛 宋紫君 苏宏业

(74) 专利代理机构 杭州求是专利事务有限公司 33200

专利代理师 林超

(51) Int. Cl.

G06T 7/12 (2017.01)

G06T 7/13 (2017.01)

G06N 3/0442 (2023.01)

G06N 3/0464 (2023.01)

G06N 3/0455 (2023.01)

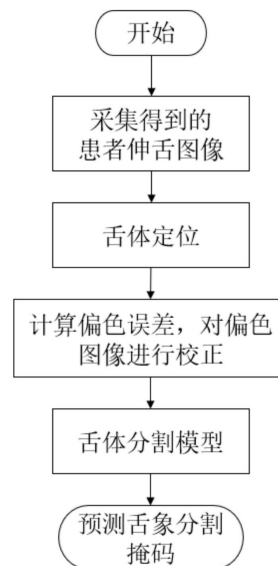
权利要求书2页 说明书8页 附图3页

(54) 发明名称

一种基于边缘门控机制的舌象图像分割方法

(57) 摘要

本发明公开了一种基于边缘门控机制的舌象图像分割方法。本发明针对移动设备采集的原始舌象数据,采用舌体定位与颜色校正的预处理操作。本发明针对舌体分割任务中舌体边缘提取粗糙的问题,设计了一种基于边缘门控机制的舌体分割模型,该模型是在Deeplabv3+模型基础上加入边缘分支作为改进,并设计了分割损失与边缘损失加权融合的损失函数对改进模型进行训练,从而提升了模型分割掩码的预测精度。本发明在解码器中使用了跳跃连接的方法,将通过编码器得到的特征向量与解码器的特征向量进行特征融合,并采用双层卷积与上采样机制将图像恢复至原尺寸,增强了特征捕捉能力,提高了解码器性能与分割精度。



A BERT-Based Named Entity Recognition Method of Warm Disease in Traditional Chinese Medicine

Zijun Song
Polytechnic Institute
Zhejiang University
Hangzhou, China
zjsong2000@zju.edu.cn

Wen Xu
Binzhou Medical University
Yantai, China
xuwenxw@yeah.net

Zhitao Liu
State Key Laboratory of Industrial
Control Technology
Institute of Cyber-Systems and Control,
Zhejiang University
Hangzhou, China
ztliau@zju.edu.cn

Liang Chen
State Key Laboratory of Industrial
Control Technology
Institute of Cyber-Systems and Control,
Zhejiang University
Hangzhou, China
aqcl@zju.edu.cn

Hongye Su
State Key Laboratory of Industrial
Control Technology
Institute of Cyber-Systems and Control,
Zhejiang University
Hangzhou, China
hysu@iipc.zju.edu.cn

Abstract—Traditional Chinese medicine (TCM) documents have been handed down through the ages, containing rich theoretical knowledge and clinical experience. These unstructured data are the foundation for building the digital knowledge system of TCM. However, written in ancient Chinese, the TCM books have complex grammatical rules and terms which are different from modern medicine, inducing difficulty in entity annotation and recognition. In order to solve the problem of lacking labeled data, we construct a dataset with *Wenbing Tiaobian*, a classic work of TCM on the warm disease, identify six entities and annotate the book with the BIOES method. The BERT-BILSTM-CRF model is used to conduct experiments on the dataset with an F1 value of 91.4%. The results verify the effectiveness of the model in NER tasks and advance the construction of knowledge graphs in TCM.

Keywords—Natural Language Processing, Named Entity Recognition, Deep Learning, Traditional Chinese Medicine

I. INTRODUCTION

Traditional Chinese Medicine (TCM) is a valuable cultural treasure of the Chinese people for over 5,000 years. The ancient books in Chinese medicine contain a wealth of Chinese medical theory, prescription, and clinical experience, which demonstrate how TCM practitioners understand and treat various illnesses. Warm Disease is a clinical discipline in the field of Chinese medicine. The discipline studies the occurrence and treatment methods of infectious and febrile diseases caused by external pathogens, such as wind, cold, and summer heat. Established in Qing Dynasty, the diagnosis and treatment system of the warm disease has had a profound impact on modern Chinese medicine. And the theory recorded in ancient books, such as *Wenyi Lun* and *Wenbin Tiaobian*, has also been widely used for the treatment of influenza, pneumonia, Covid-19, and so on. However, most TCM texts, including books on the warm disease, are written in ancient Chinese. These books are complex to read and understand, and the terms are different from the modern

medicine system, making it difficult for TCM researchers to obtain effective information. Therefore, it is necessary to promote the digitalization and intelligence of TCM knowledge, such as identifying named entities in ancient TCM texts and establishing a knowledge mapping network to build a digital knowledge system of TCM.

The Knowledge graph is a structured semantic network. Through the knowledge graph, the concepts and objects in the real world are abstracted into symbols and are presented in a network. The basic units of the knowledge graph are entities and relationships, which form "entity-relationship-entity" triples. The network represents complex semantic information in a way that is highly compatible with human cognitive patterns [1]. Named Entity Recognition (NER) is a crucial step in building the knowledge graph, and also an essential aspect of Natural Language Processing (NLP). It refers to the process of selecting meaningful terms, concepts, and other entities from unstructured digital data. The quality of entity recognition significantly impacts the accuracy of subsequent knowledge extraction, which has become a topic that scholars have been researching.

In the early days, the main methods for NER were rule-based and dictionary-based methods, as well as machine learning methods [2]. Based on the annotation standard proposed by the project "Informatics for Integrating Biology and the Bedside" (I2B2), Qu et al. [3] developed specifications for the labeling of Chinese electronic medical records. Traditional Chinese Medical Language System (TCMLS) was built by the Institute of Information on TCM, which defined TCM entities and established semantic relationships, forming a TCM knowledge graph [4]. To extract named entities from ancient texts that the usage of vocabulary and the writing convention are different from modern Chinese, it is necessary to define entities based on TCMLS and preprocess the texts, such as entity normalization and entity alignment [5]. Xu et al. [6] constructed the medicine ontology of the warm disease and built a semantic network that

represents the relationships between the disease and other types of entities. With the development of machine learning, Liu et al. [7] compared the effectiveness of conditional random field (CRF), hidden Markov model (HMM), and maximum entropy Markov model (MEMM) in NER experiments on TCM medical records.

In recent years, due to the superior performance of Neural Network models in feature extraction, deep learning has been widely used in NER tasks. The method does not require the manual definition of features and acquires features of texts through neural networks automatically. Recurrent Neural Network (RNN) is usually used to process sequential texts, while Long-Short Term Memory (LSTM) network improves the hidden layer structure of RNN and is able to process longer sequences. Bidirectional LSTM (BiLSTM) model fuses the textual information in both front and back directions, which makes up for the shortcomings of the LSTM network. With the conditional random field (CRF) for solving prediction tasks, the BiLSTM-CRF model proposed by Huang et al. [8] has been the mainstream model for NER for a period of time. Gao et al. [9] adopt the BiLSTM-CRF model to recognize the TCM entities in Huangdi Neijing with an F1 value of 85.32%. Deng et al. [10] annotate the TCM medical cases and construct word2vec-based word vectors to obtain a TCM-named entity recognition model with an F1 value of 88.34%.

Pre-trained models perform very well in all kinds of NLP tasks due to their large training datasets and portability under multiple tasks, especially the Bidirectional Encoder Representation from Transformers (BERT) proposed by Google in 2018 [11]. Transformers can capture long-range features more effectively than CNN models and can perform parallel computation with high efficiency, overcoming the shortcomings of RNN. Based on the dual-layer Transformer encoder with the self-attention mechanism, the BERT model can generate deep bidirectional language representation and adopt masked language modeling (MLM) and next-sentence prediction (NSP) strategies to train more powerful semantic representations. For Chinese NER tasks, scholars have also proposed various models based on BERT, for example, BERT-wwm [12], AMBERT [13], ChineseBERT [14], and so on. The pre-training model BERT has also been used for the TCM-named entity recognition tasks to construct the embedding representations. Zhang et al. [15] propose a semi-supervised embedded Semi-BERT-BiLSTM-CRF model that improves the entity recognition accuracy on TCM and reduces the manual labeling work. Liang et al. [16] utilize an unlabeled clinical corpus to fine-tune the BERT language model and achieve a state-of-the-art F1 score of 89.39%.

This paper constructs a dataset with Wenbing Tiaobian, one of the landmarks on the warm disease, and annotates the entities of disease, description, medicine, prescription, body, and treatment with the BIOES annotation method. A sequence labeling model based on BERT-BiLSTM-CRF is adopted to conduct NER experiments on the annotated dataset, resulting in successful entity recognition and providing technical support for the digitization of ancient Chinese medicine books.

II. METHOD

To improve the recognition accuracy of named entities for ancient books in traditional Chinese medicine, this paper proposes a BERT-BiLSTM-CRF model as shown in Fig. 1., which consists of 4 parts.

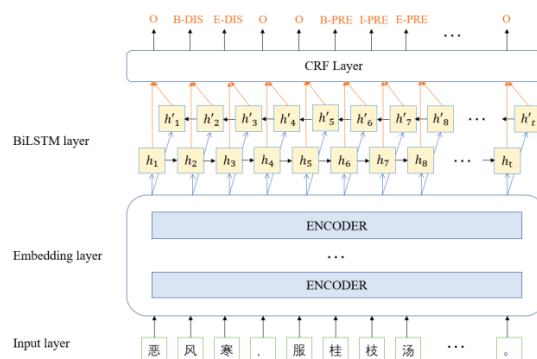


Fig. 1. The framework of the BERT-BiLSTM-CRF model.

- Input layer: Samples are input into the model in terms of sentences.
- Embedding layer: the Bert pre-training model is used to process the input sentences into the word vectors containing semantic embedding representations.
- BiLSTM layer: contextual information of the text is extracted through the BiLSTM model.
- CRF layer: the label sequence with the largest probability value is output to complete the labeling of the entity.

A. Embedding layer

BERT is a pre-trained model which generates deep bidirectional language representations with a great number of training parameters and corpus. The model structure is shown in Fig. 2.

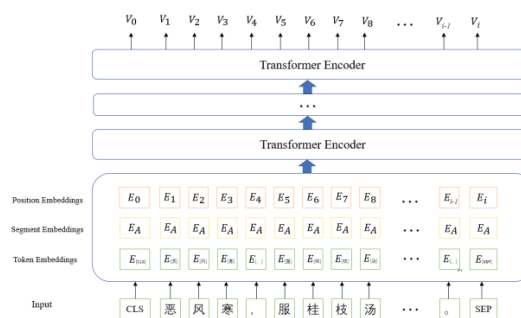


Fig. 2. The structure of the pre-trained model BERT.

The input layer adds the [CLS] and [SEP] token at the beginning and the end of the input sentence $X = \{x_1, x_1, x_2 \dots x_i\}$ which contains i tokens(characters), and assembles the token embeddings E_{token} , segment embeddings E_{seg} and position embeddings E_{pos} into the input representation $E = \{E_{x_0}, E_{x_1}, E_{x_2} \dots E_{x_i}\}$. Then the representation E is input into a

two-layer transformer structure for feature extraction. The transformer layer contains six encoder-decoders, using the self-attention mechanism to capture the contextual information. The final layer is the output in the form of word vector $V = \{v_0, v_1, v_2 \dots v_i\}$ that is passed into the BiLSTM layer.

B. BiLSTM layer

The LSTM model is a special kind of RNN. In a standard RNN network, the hidden layer h_t at time t is calculated from the hidden layer h_{t-1} at the previous time together with the input X_t at the current time, where the hidden layer function generally has only a simple structure, such as the tanh layer. In contrast, the hidden layer in LSTM consists of several gating units, including the traditional input and output gates, as well as the newly added forgetting gates. The computation flow of each gating unit is as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (2)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

where f_t represents the state of forget gate, h_{t-1} represents the state of the last hidden layer, \tilde{C}_t represents the effective information of the input state of x_t and the hidden layer h_{t-1} , i_t represents the state of input gate, which is used to determine the new information stored in the cell state, o_t represents the state of output gate, C_t represents the cell state updated, and h_t represents the hidden layer output state at time t .

The BiLSTM model is divided into two layers. In the forward layer, the model computes the output of the hidden layer h_t at each moment from moment t_0 to the current moment t in the forward direction, while in the backward layer, the model computes the output of the hidden layer h'_t at each moment along the opposite moment. Then the hidden layer states obtained during the forward and backward direction are calculated to obtain the final output o_t . The specific calculation process is as follows:

$$h_t = f(w_1 x_t + w_2 h_{t-1}) \quad (7)$$

$$h'_t = f(w_3 x_t + w_4 h'_{t+1}) \quad (8)$$

$$o_t = g(w_5 h_t + w_6 h'_t) \quad (9)$$

C. CRF layer

Conditional Random Field (CRF) is a conditional probability distribution model for solving the output label sequence, which can provide conditional constraints for the output of BiLSTM module. For example, the following constraints exist in this study: 1. entity labels must start with the

label B; 2. entity labels must end with the label O instead of the label I; 3. the label of each character in an entity must be same, such as “桂 B-PRE/枝 I-PRE/汤 E-PRE” instead of “桂 B-PRE/枝 I-DESC/汤 E-PRE”. For the labeled sequence $(x^i, y^i)_{i=1}^n$, the matching score $Score(x, y)$ is:

$$Score(x^i, y^i) = \sum_m \log \varphi_{emit}(y_m^i \rightarrow x_m^i) + \log \varphi_{trans}(y_{m-1}^i \rightarrow y_m^i) \quad (10)$$

where $x^i = \{x_1^i, x_2^i, \dots, x_m^i\}$ is the given input, $y^i = \{y_1^i, y_2^i, \dots, y_m^i\}$ is the given output, φ_{emit} represent the probability of the i^{th} word in x^i corresponding to the label y^i , φ_{trans} represent the transition matrix from y_{m-1}^i to y_m^i .

Then the probability of the labeled sequence y^i is:

$$P(x|y, \theta) = \frac{\exp(Score(x, y))}{\sum_{y'} \exp(Score(x, y'))} \quad (11)$$

The output label sequence is:

$$\theta^* = \arg \max \sum_i \log [P(x^i | y^i, \theta)] \quad (12)$$

The CRF layer can provide the BiLSTM model with conditional constraints by observing and learning the sequences, and prevent the model from producing wrong-labeled sequences effectively.

III. EXPERIMENT

A. Datasets

This paper constructs a dataset with Wenbing Tiaobian, an ancient book that has a profound influence on the development of the warm disease in the Qing Dynasty. The book is annotated with the BIOES annotation method. B represents the start position of the entity, I represents the middle position of the entity, E represents the end position of the entity, S represents the single-word entity, and O represents the rest of the non-entity parts. Based on TCMLS, we divides the entities in the text into six types of labels, including Disease, Description, Medicine, Prescription, Body, and Treatment. The specific annotation types are shown in Table I.

TABLE I. ANNOTATION TYPES OF THE DATASET

Entity Type	Single Word Entity	Phrase Entity	Examples
Disease	S-DIS	B-DIS, I-DIS, E-DIS	风温
Description	S-DESC	B-DESC, I-DESC, E-DESC	头痛
Medicine	S-MED	B-MED, I-MED, E-MED	当归
Prescription	S-PRE	B-PRE, I-PRE, E-PRE	桂枝汤
Body	S-BODY	B-BODY, I-BODY, E-BODY	肺
Treatment	S-TRE	B-TRE, I-TRE, E-TRE	苦辛通降
Non-entity	O	O	者

After crawling the full text from the web, we pre-process the data by clearing messy codes, unifying punctuation, converting all traditional characters to simplified ones, and converting phonetic loan characters to commonly-used ones. For the large number of different names that refer to the same entity in ancient texts, we conduct entity alignment before annotation, such as using standardized entities to rename them. While for various medicinal materials and prescriptions with abbreviated names, we restore them to their full names.

The annotated dataset consists of 95655 characters and some labeled sequences of the dataset are shown in Fig. 3.

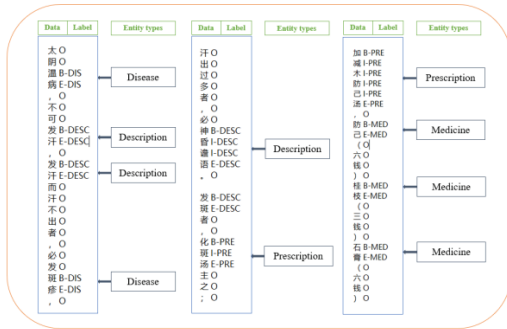


Fig. 3. Samples of labeled sequences in the dataset.

B. Experiment settings and Evaluation Indicators

For the completed annotated dataset, this experiment adopts a random sampling division method, dividing 80% of the dataset into a training set (Train), 10% into a validation set (Dev), and 10% into a test set (Test). The number of characters is shown in Table II.

TABLE II. DATASET SEGMENTATION STATISTICS

	Train	Dev	Test
Number of Characters	75973	9804	9876

The model structure used in this experiment is BERT-BiLSTM-CRF, which utilizes the basic architecture of the pre-trained model BERT and fine-tunes it to achieve semantic encoding instead of word embedding. The experimental hardware is Intel Core i7-11800H, the GPU version is GeForce RTX3060, the Python version is 3.7 and the PyTorch version is 1.11.0.

The hyperparameter settings are shown in Table III.

TABLE III. EXPERIMENTAL HYPERPARAMETER SETTINGS

Parameter	Value
Max sequence length	100
Learning rate	3e-5
Hidden dim	256
Epoch	300
Batch size	22

Precision P, recall R and F1 values are used to evaluate the effectiveness of the model. The precision rate P represents the percentage of truly positive samples in predicted positive samples. The recall rate R represents the percentage of truly positive samples in original positive samples. F1 represents an evaluation metric that takes both precision and recall into account and is calculated as follows:

$$P = \frac{TP}{TP+FP} \quad (13)$$

$$R = \frac{TP}{TP+FN} \quad (14)$$

$$F1 = \frac{2 \times P \times R}{P+R} \quad (15)$$

Where TP indicates the number of correctly-predicted positive samples, FP indicates the number of predicted positive samples from the negative classes, and FN indicates the number of predicted negative samples from the positive classes.

C. Experimental results and analysis

The results of the experiments are shown in Table IV, and the item Support in the table indicates the number of entities in the test set.

TABLE IV. EXPERIMENTAL RESULTS ON THE DATASET

Entity Type	Precision(%)	Recall(%)	F1(%)	Support
DIS	90.8	89.5	90.1	97
DESC	82	76.2	78.9	162
MED	97.6	95.1	96.3	273
PRE	96.2	98.1	97.1	104
BODY	96.3	97.5	96.9	77
TRE	88.1	66.7	75.3	11
Average	92.7	90.4	91.4	724

According to the table, the BERT-BiLSTM-CRF model has an average precision of 92.7%, an average recall of 90.4%, and an average F1 value of 91.4%. The column graph of the results is shown in Fig. 4. In terms of the recognition accuracy in each category, the model performs better in the four types of "DIS", "MED", "PRE" and "BODY", with F1 values above 0.9, while the F1 values are only 0.789 and 0.753 in the types of "DESC" and "TRE".

By analyzing the experimental results and annotated corpus, it can be concluded that the effectiveness of entity recognition is affected by the number of entities and the quality of annotation. The four types of entities, "disease (DIS)", "medicine (MED)", "prescription (PRE)", and "body (BODY)", have a high frequency of occurrence in the text. These types are evenly distributed within each chapter and have relatively fixed expressions, which will not cause significant annotation and recognition errors. However, the author tends to use different adjectives to describe the entities of description (DESC), and their frequency of occurrence is not high, which can easily

produce labeling noise. For the entities labeled with treatment (TRE), there are few samples distributed unevenly and are prone to nested words, which greatly affects the accuracy of entity recognition.

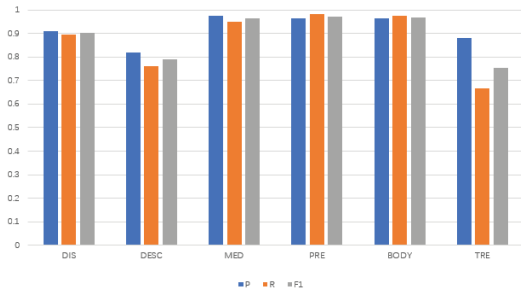


Fig. 4. Samples of labeled sequences in the dataset. The model performs better in the four types of “DIS”, “MED”, “PRE” and “BODY”, with F1 values above 0.9, while the F1 values are only 0.789 and 0.753 in the types of “DESC” and “TRE”.

To verify the efficiency of the model BERT-BiLSTM-CRF in NER, the following three experiments are set up for comparison: LSTM-CRF, BiLSTM-CRF, and BERT-CRF. The experimental comparative results are shown in Table V.

TABLE V. COMPARISON OF DIFFERENT MODELS

	<i>Precision(%)</i>	<i>Recall(%)</i>	<i>F1(%)</i>
LSTM-CRF	72.6	71.5	71.7
BiLSTM-CRF	90.7	82.9	85.9
BERT-CRF	91.8	89.6	90.7
BERT- BiLSTM-CRF	92.7	90.4	91.4

It can be seen that the LSTM-CRF model does not perform well on this dataset due to the inability of the one-way LSTM model to capture contextual information. While the BiLSTM-CRF model effectively solved this problem and improves recognition accuracy. The BERT-CRF model outperforms the BiLSTM-CRF model for the reason that the pre-trained model has been trained using a large dataset before fine-tuning, resulting in stronger representation capability. The overall performance of the BERT-CRF model is slightly lower than the BERT-BiLSTM-CRF model, which can also demonstrate that the BiLSTM module can provide more location and orientation information to the overall model, making it more capable of discovering dependencies in the input sequences. The experimental results show that the BERT-BiLSTM-CRF model has excellent performance for the entity recognition of Chinese medicine texts, and can be applied to promote the knowledge mining of TCM.

IV. CONCLUSION

This study focuses on the NER tasks of ancient books in traditional Chinese medicine. Due to the lack of publicly available datasets of TCM books, this paper constructs a dataset with “Wenbing Tiaobian”, a famous book on the warm disease, and defines six entity types according to the TCMLS standard.

In the named entity recognition experiments, the BERT-BiLSTM-CRF model obtained great recognition accuracy with an F1 value of 91.4 and performed best when compared with LSTM-CRF, BiLSTM-CRF, and BERT-CRF models. The experimental results demonstrate that the BERT-BiLSTM-CRF model can extract specific entities effectively from non-structured texts, which significantly promotes the digitalization of ancient Chinese medical books. In this way, knowledge graphs of the warm disease can be constructed to help relevant researchers inherit and develop traditional Chinese medicine.

ACKNOWLEDGMENT

This work was partially supported by National Key R&D Program of China (Grant NO. 2021YFB3301000); National Natural Science Foundation of China (NSFC:62173297); Natural Science Foundation of Shandong Province, China (ZR2020QH320); Zhejiang Key R&D Program (Grant NO. 2022C01035).

REFERENCES

- [1] L. Jia, J. Liu, T. Yu, Y. Dong, L. Zhu, B. Gao, and L. Liu, “Construction of traditional Chinese medicine knowledge graph,” *Journal of Medical Informatics*, pp.51-53, 2015.
- [2] X. Chu, B. Sun, Q. Huang, S. Peng, Y. Zhou, and Y. Zhang, “Quantitative knowledge presentation models of traditional Chinese medicine (TCM): A review,” *Artificial intelligence in medicine*, vol. 103, p.101810, 2020.
- [3] C. Qu, Y. Guan, J. Yang, and Y. Liu, “The construction of annotated corpora of named entities for Chinese electronic medical records,” *Chinese High Technol Lett*, vol. 25, pp.143-50, 2015.
- [4] T. Yu, M. Cui, H.Y. Li, and S. Yang, “Semantic network framework of traditional Chinese medicine language system: an upper-level ontology for Traditional Chinese medicine,” *China Digital Medicine*, vol.9, no. 1, pp.44-47, 2014.
- [5] M. Wang, “Research on the Construction, Knowledge Mining and Application of Infertility Knowledge Graph in Ancient Books of Traditional Chinese Medicine,” Beijing: China Academy of Chinese Medical Sciences, 2022.
- [6] W. Xu, “Study on Concept Semantic Network Building Based on the Knowledge of Warm Disease in Ancient TCM Books,” Beijing: China Academy of Chinese Medical Sciences, 2015.
- [7] L. Kai, “A Study of Named Entity Extraction of TCM Medical Records using Conditional Random Fields,” Beijing: Beijing Jiaotong University, 2013.
- [8] Z. Huang, W. Xu, and K. Yu, “Bidirectional LSTM-CRF models for sequence tagging,” *arXiv preprint arXiv:1508.01991*, 2015.
- [9] S. Gao, P. Jin, and D.Z. Zhang, “Research on named entity recognition of TCM classics based on deep learning,” *Technology Intelligence Engineering*, vol. 5, no.1, pp.113-123, 2019.
- [10] N. Deng, H. Fu, and X. Chen, “Named entity recognition of traditional Chinese medicine patents based on BiLSTM-CRF.” *Wireless Communications and Mobile Computing*, pp.1-12, 2021.
- [11] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [12] Y. Cui, W. Che, T.Liu, B. Qin, and Z. Yang, “Pre-training with whole word masking for chinese BERT,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp.3504-3514, 2021.
- [13] X. Zhang, P. Li, and H. Li, “AmbERT: A pre-trained language model with multi-grained tokenization,” *arXiv preprint arXiv:2008.11869*, 2020.
- [14] Z. Sun, X. Li, X. Sun, Y. Meng, X. Ao, Q. He, F. Wu, and J. Li, “ChineseBERT: Chinese pretraining enhanced by glyph and pinyin information,” *arXiv preprint arXiv:2106.16038*, 2021.
- [15] M. Zhang, Z. Yang, C. Liu and L. Fang, “Traditional Chinese Medicine knowledge Service based on Semi-Supervised BERT-BiLSTM-CRF

Model,” 2020 International Conference on Service Science (ICSS). IEEE, pp. 64-69, 2020.

[16] L. Yao, Z. Jin, C. Mao, Y. Zhang, and Y. Luo, “Traditional Chinese medicine clinical records classification with BERT and domain specific

corpora,” Journal of the American Medical Informatics Association, vol. 26, no. 12, pp.1632-1636, 2019.

经检索“Engineering Village”，下述论文被《Ei Compendex》收录。（检索时间：2024年12月13日）。

<RECORD 1>

Accession number:20234114875846

Title:A BERT-Based Named Entity Recognition Method of Warm Disease in Traditional Chinese Medicine

Authors:Song, Zijun (1); Xu, Wen (2); Liu, Zhitao (3); Chen, Liang (3); Su, Hongye (3)

Author affiliation:(1) Polytechnic Institute, Zhejiang University, Hangzhou, China; (2) Binzhou Medical University, Yantai, China; (3) Institute of Cyber-Systems and Control, Zhejiang University, State Key Laboratory of Industrial Control Technology, Hangzhou, China

Corresponding author:Song, Zijun(zjsong2000@zju.edu.cn)

Source title:Proceedings of the 18th IEEE Conference on Industrial Electronics and Applications, ICIEA 2023

Abbreviated source title:Proc. IEEE Conf. Ind. Electron. Appl., ICIEA

Part number:1 of 1

Issue title:Proceedings of the 18th IEEE Conference on Industrial Electronics and Applications, ICIEA 2023

Issue date:2023

Publication year:2023

Pages:1226-1231

Language:English

ISBN-13:9798350312201

Document type:Conference article (CA)

Conference name:18th IEEE Conference on Industrial Electronics and Applications, ICIEA 2023

Conference date:August 18, 2023 - August 22, 2023

Conference location:99 Min An Dong Lu, Yinzhou District, Zhejiang Province, Ningbo, China

Conference code:192573

Publisher:Institute of Electrical and Electronics Engineers Inc.

Number of references:16

Main heading:Deep learning

Controlled terms:Knowledge graph - Medicine - Natural language processing systems

Uncontrolled terms:Clinical experience - Deep learning - Knowledge system - Language processing - Named entity recognition - Natural language processing - Natural languages - Recognition methods - Traditional Chinese Medicine - Unstructured data

Classification code:461.4 Ergonomics and Human Factors Engineering - 461.6 Medicine and Pharmacology - 723.2 Data Processing and Image Processing - 723.4 Artificial Intelligence

Numerical data indexing:Percentage 9.14E+01%

DOI:10.1109/ICIEA58696.2023.10241595

Funding details: Number: 2022C01035, Acronym: -, Sponsor: -; Number: NSFC:62173297, Acronym: NSFC, Sponsor: National Natural Science Foundation of China; Number: ZR2020QH320, Acronym: -, Sponsor: Natural Science Foundation of Shandong Province; Number: 2021YFB3301000, Acronym: NKRDPC, Sponsor: National Key Research and Development Program of China;

Funding text:ACKNOWLEDGMENT This work was partially supported by National Key R&D Program of China (Grant NO. 2021YFB3301000); National Natural Science Foundation of China (NSFC:62173297); Natural Science Foundation of Shandong Province, China (ZR2020QH320); Zhejiang Key R&D Program (Grant NO. 2022C01035).

Database:Compendex

Compilation and indexing terms, Copyright 2024 Elsevier Inc.

注:

1. 以上检索结果来自 CALIS 查收查引系统。
2. 以上检索结果均得到委托人及被检索作者的确认。



教育部科技查新工作站 (Z09)

检索人 (签章): 李佳

2024年12月13日

