TcmYiAnBERT：基于无监督学习的中医医案预训练模型*

胡 为 刘 伟 盛 威 卢彦杰 石玉敬

（湖南中医药大学信息科学与工程学院 长沙 410013）

〔摘要〕 **目的/意义** 充分挖掘中医医案中的文本信息，提高中医药信息化程度和中医医案症状术语抽取、关系抽取等下游任务的准确率。**方法/过程** 通过光学字符识别和爬虫技术获取大量中医医案数据并进行预处理，构建面向中医医案领域预训练数据集，使用 BERT 模型预训练方法，经过多轮训练得到首个面向中医领域专有预训练模型 TcmYiAnBERT，并将该模型开源。**结果/结论** 中医领域专有预训练模型 TcmYiAnBERT 在中医命名实体识别任务中比未使用该模型的预训练模型 *F*1 值提高 2.8 个百分点。

〔关键词〕 中医医案；预训练模型；TcmYiAnBERT；命名实体识别；人工智能

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2023.07.010

TcmYiAnBERT: A Traditional Chinese Medicine Case Pre－training Model Based on Unsupervised Learning

HU Wei, LIU Wei, SHENG Wei, LU Yanjie, SHI Yujing

School of Informatics, Hunan University of Chinese Medicine, Changsha 410013, China

〔**Abstract**〕 **Purpose/Significance** To fully mine the text information in traditional Chinese medicine (TCM) medical records, to improve the degree of TCM informatization, and to improve the accuracy of downstream tasks such as symptom term extraction and relationship extraction in TCM medical records. **Method/Process** A large number of TCM medical case data are obtained through optical character recognition (OCR) technology and crawler technology, and data preprocessing is carried out. A pre－training data set for TCM medical case field is constructed. The first proprietary pre－training model, namely TcmYiAnBERT, for TCM field is obtained through multiple rounds of training by using the BERT model pre－training method, and the model is open source. **Result/Conclusion** The experiment shows that the recognition accuracy of TCM domain specific pre－training model TcmYiAnBERT in the task of TCM named entity recognition (NER) is 2.8 percentage points higher than that of other pre－training models.

〔**Keywords**〕 TCM medical records; pre－training model; TcmYiAnBERT; named entity recognition; artificial intelligence (AI)

〔修回日期〕 2022－12－24

〔作者简介〕 胡为，助教，发表论文 3 篇；通信作者：刘伟，博士，副教授。

〔基金项目〕 湖南省自然科学基金项目（项目编号：2022JJ30438）；湖南中医药大学校级自然科学基金项目（项目编号：2022XJZKC016）；湖南省教育厅科学研究项目（项目编号：20C1435）。

1 引言

中医医案是中医医师临床实施辨证论治的文字记录和实施临床诊断的宝贵经验[1]。但目前中医医案信息化程度低，各医案相对零散且缺乏联系，非结构化数据多，对医案结构化数据的定义难以形成统一标准，严重影响中医医案数据挖掘工作的推

进，因此，研究如何从海量中医医案数据中挖掘有用信息具有重要意义。近年来，人工智能在自然语言处理、计算机视觉等领域均取得重要突破，计算机技术已应用于中医药各领域，如中医医案命名实体识别、实体关系抽取、知识图谱构建等。2018 年双向编码器表征（bidirectional encoder representations from transformers，BERT）预训练模型[2]问世，并被应用于各类通用自然语言处理任务，准确率大幅度提升。预训练模型先在一个原始任务上预先训练一个初始模型，然后在目标任务上使用该模型针对目标任务特性，对初始模型精调，达到精准完成目标任务的目的。目前 BERT 预训练模型是基于通用语料训练的，受限于领域知识，难以在各中医医案自然语言处理下游任务中取得良好效果。因此，有些学者研究基于特定领域的预训练模型，如将面向古文、基于 BERT 的继续训练迁移至古汉语模型得到 SikuBERT 模型[3]，训练面向生物医学领域的预训练模型 BioBERT[4]、面向临床医学领域的预训练模型 ClinicalBERT[5]、面向科学领域的预训练模型 SciBERT[6]、面向专利领域的预训练模型 PatentBERT[7]，但目前还没有面向中医医案领域的预训练模型。本文基于 BERT 预训练模型技术，利用光学字符识别（optical character recognition，OCR）技术和爬虫技术获取大量中医医案数据语料，构建首个面向中医医案领域专有预训练模型（traditional Chinese medical YiAn bidirectional encoder representation from transformers，TcmYiAnBERT），并在多个中医医案自然语言处理下游任务实验中证明其优越性。

2 中医医案数据集与预训练模型构建

2.1 数据集构建

本文数据集来源于两方面。一是《中国现代名中医医案精粹》[8]《古今医案按》[9]《明清十八家名医医案》[10]《丁甘仁医案》[11]等中医医案经典书籍，通过 Python 语言编写 OCR 程序，将 PDF 文本转为 TXT 文本，得到 10 万条中医医案，共512万字。二是“中医中药网”“中医资源网”“经方派”“道医网”等中医药网站，通过 Python 语言提供的 Request 和 BeautifulSoup 等爬虫库及正则表达式等技术得到 120 万条中医医案，共 4 100 万字。剔除一些禁用字、识别错误的字、非中文标点符号的字等，最终得到的数据集共有汉字 44 121 245 个，对数据集中的每个句子按照 15% 进行掩码操作，最后按照 8∶1∶1 构建训练集、测试集和验证集。

2.2 预训练模型

在自然语言处理的各类任务中都需要考虑如何将文本信息转化为计算机能识别的数据形式，早期主流做法是以 Word2Vec[12] 和 Glove[13] 为代表的基于预训练静态词向量技术。该技术通常将词汇用多维向量表示，但构建的词向量没有考虑语境，因此无法解决一词多义问题。且该方法本身属于一种浅层结构的静态词向量，在句子较长时无法学习到长距离依赖的上下文语义信息。

基于深度学习算法的另一个特点是各类任务都需要考虑构建大规模带标注的数据集，以便使程序学习到相关的语法和语义信息，但标注数据需要消耗大量低技术的人工成本。自 2018 年以来，以 BERT[2] 为代表的超大规模预训练语言模型可有效弥补自然语言处理一词多义及处理数据标注耗时耗力的问题。BERT 是一个超大规模的语义表征预训练模型，它从维基百科大规模语料库通过无监督学习到通用的语义表示信息，且该模型可通过微调适应各种下游任务。该模型采用 Transformer 架构的 Encoder 模块，基于 Transformer 的双向编码模型具有强大的特征提取能力，最终生成的字向量融合字词上下文语义信息，能更充分地表征字词的多义性。

2.3 TcmYiAnBERT 预训练模型构建

本文预训练模型构建的总体流程，见图 1。

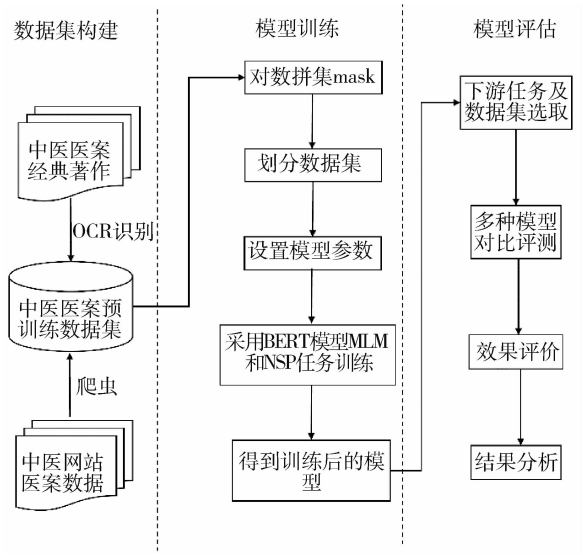

图 1 TcmYiAnBERT 预训练模型构建总体流程

2.3.1 模型的输入 对数据集中每一句文本进行编码得到 3 个向量，见图 2。其中字向量 Token Embeddings 通过查询字向量表将每个字转换为该字的字编码，句子嵌入编码 Segment Embeddings 表示文本的全局语义信息，位置嵌入编码 Position Embedding 表示文本中每个字所在的位置语义信息。

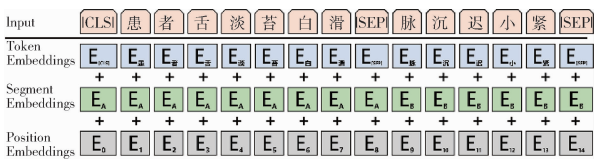

图 2 TcmYiAnBERT 模型输入

2.3.2 模型训练任务 以 BERT 模型的掩码语言模型（maskedlanguage model，MLM）和下一句预测（next sentenceprection，NSP）预训练任务为基础进行训练。MLM 任务采用在句子中随机掩盖掉若干字的方式学习该字在上下文的语义信息。如将“患者舌淡苔白滑脉沉迟小紧”这个句子掩盖若干字后变成了“患者［mask］淡苔［mask］滑脉沉［mask］小紧”，然后通过模型训练预测［mask］处的字信息，见图 3。NSP 任务主要目标是预测两个句子是否连在一起，令模型学习两个连续句子的关系，使模型具有更好的长距离上下文语义学习能力，如“患者舌淡苔白滑”和“脉沉迟小紧”同时输入模型，最后模型输出这两个句子是否连在一起。

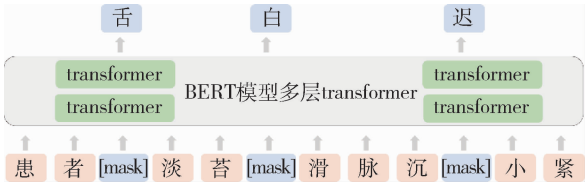

图 3 TcmYiAnBERT 模型 MLM 任务示例

2.4 TcmYiAnBERT 预训练模型发布

采用 Pytorch 1.15 框架、Python 3.7 版本环境，模型参数与 BERT 模型一致，设置 12 层 Transformer、12 个 Attention - head、768 个隐藏层单元，整个模型有 110 兆参数。实验过程中，将 BERT 模型参数最大句子长度设置成 256，批处理大小设置成 64，训练轮数设置为 100。最终在单台 V 100 显卡机器上经过 680 小时的训练得到本文的 TcmYiAnBERT 预训练模型，并且将该模型发布到 Huggingface 网站（https://huggingface.co/lucashu/TcmYiAnBERT）。

3 实验结果与分析

3.1 验证实验数据集和任务

采用中医医案命名实体识别任务对 TcmYiAnBERT 模型性能进行验证，验证模型包括 TcmYiAnBERT 预训练层、BiLSTM 层和条件随机场（conditional random field，CRF）层，见图 4。

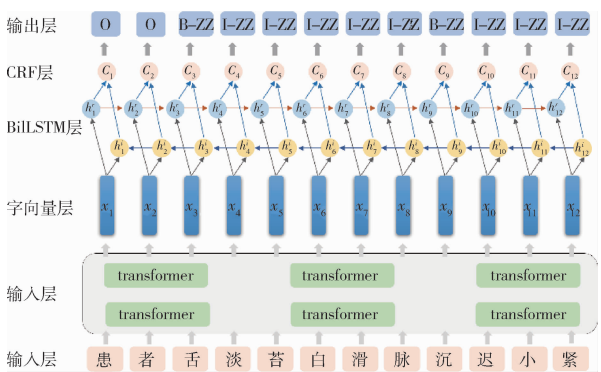

图 4 TcmYiAnBERT - BiLSTM - CRF 模型

验证实验用的数据集是从《中国现代名中医医案精粹》中选取的 1 000 条高质量医案，并由多位

经验丰富的中医学者对数据集标注。数据集采用命名实体识别任务通用的 BIO 标注法，共设计功效实体、辨证实体、治则实体、症状实体、方药实体、人群实体 6 类实体类别。

3.2 验证实验任务评价指标

采用命名实体识别任务常用的精确率（P）、召回率（R）和 $F1$ 测度值评估模型性能。假设 T_p 表示模型识别正确的实体个数，F_p 表示模型识别错误的实体个数，F_N 为模型没有识别出的实体个数。计算方式如下：

$$P = \frac{T_p}{T_p + F_p} \times 100\% \quad (1)$$

$$R = \frac{T_p}{T_p + F_N} \times 100\% \quad (2)$$

$$F1 = \frac{2P \times R}{P + R} \times 100\% \quad (3)$$

3.3 验证实验任务结果

选取命名实体识别 3 种经典模型，对比验证 TcmYiAnBERT 模型在下游任务的效果，见表 1。

表 1 对比模型实验结果

模型	P	R	$F1$
BiLSTM 模型	0.853	0.863	0.845
BiLSTM – CRF 模型	0.867	0.878	0.866
BERT – BiLSTM – CRF 模型	0.902	0.897	0.899
TcmYiAnBERT – BiLSTM – CRF 模型	0.930	0.926	0.927

TcmYiAnBERT – BiLSTM – CRF 模型得到的各类症状实体评价指标，见表 2。

表 2 TcmYiAnBERT – BiLSTM – CRF 模型各类实体识别结果

实体标签	P	R	$F1$	标签数量（个）
治则标签	0.894	0.884	0.888	338
人群标签	0.975	0.981	0.977	98
方药标签	0.953	0.976	0.964	1 656
辨证标签	0.847	0.838	0.842	298
症状标签	0.884	0.867	0.875	103
功效标签	0.865	0.871	0.867	486
平均	0.930	0.926	0.927	2 991

从结果可以看出，加入中医医案领域专有预训练模型 TcmYiAnBERT 后，在命名实体识别任务中不同实体标签结果存在一定差异。

3.4 实验结果分析

从实验结果可以看出，在中医医案自然语言处理命名实体识别任务中，相比于传统的 BiLSTM 模型，加入 CRF 后准确度有所提升，加入预训练模型 BERT 后准确率提升效果较明显，而加入中医医案领域专有预训练模型 TcmYiAnBERT 后，准确率有更大提升，说明在中医医案自然语言处理任务中，专有领域的预训练模型 TcmYiAnBERT 比通用领域的预训练模型 BERT 更具优势。其原因是预训练模型 TcmYiAnBERT 通过 MLM 任务和 NSP 任务已经提前学到了大量中医医案语义信息，在中医医案下游任务中，只需要在该模型上微调即可达到较高的准确率。

从 6 类实体识别结果可以看出，方药实体标签和人群实体标签的准确率较高，治则、辨证、功效实体标签的准确率低于前面两类标签。其原因既与命名实体识别任务数据集划分有一定关系，又与预训练模型的数据来源有一定关系，预训练模型的数据来源于中医医案经典著作和中医医案网站，在预训练模型训练过程中学到的方药实体标签和人群实体标签语义信息较多。后续将增大数据集，进一步提高中医医案各类自然语言处理任务的准确率。

4 结语

本文构建首个面向中医领域专有预训练模型 TcmYiAnBERT，并在中医医案自然语言处理下游任务命名实体识别任务上实验证明其优越性。该预训练模型对中医医案的分词任务和实体关系抽取也将有一定帮助，进而为中医药信息化提供技术支撑。

参考文献

1 韦昌法，罗丽琴，晏峻峰. 中医数字辨证配套医案智能

采集与分析系统构建研究［J］. 湖南中医药大学学报，2020，40（1）：70－74.

2 DEVLIN J，CHANG M W，LEE K，et al. BERT：pre－training of deep bidirectional transformers for language understanding［C］. Minneapolis：The 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，2019.

3 王东波，刘畅，朱子赫，等. SikuBERT 与 SikuRoBERTa：面向数字人文的《四库全书》预训练模型构建及应用研究［J］. 图书馆论坛，2022，42（6）：31－43.

4 LEE J，YOON W，KIM S，et al. BioBERT：a pretrained-biomedical language representation model forbiomedical text mining［J］. Bioinformatics，2020，36（4）：1234－1240.

5 ALSENTZER E，MURPHY J，BOAG W. Publicly available clinical BERT embeddings［C］. Minneapolis：The 2nd Clinical Natural Language Processing Workshop，2019.

6 IZ B，KYLE L，ARMAN C. SciBERT：a pretrained language model for scientific text［C］. Hong Kong：The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing（EMNLP－IJCNLP），2019.

7 LEE J S，HSIANG J. PatentBERT：patent classification with fine－tuning a pre－trained bert model［J］. World patent information，2020，61（6）：409－418.

8 王永炎. 中国现代名中医医案精粹［M］. 北京：人民卫生出版社，2010.

9 苏礼. 古今医案按［M］. 北京：人民卫生出版社，2007.

10 伊广谦，李占永. 明清十八家名医医案［M］. 北京：中国中医药出版社，2017.

11 邹鹤瑜. 丁甘仁医案［M］. 上海：上海科学技术出版社，2001.

12 MIKOLOV T，SUTSKEVER L，CHEN K，et al. Distributed representations of words and phrases and their compositionality［C］. New York：The 26th International Conference on Neural Information Processing Systems，2013.

13 PENNINGTON J，SOCHER R，MANNING C D. Glove：global vectors for word representation［C］. Doha：The 2014 Conference on Empirical Methods in Natural Language Processing，2014.

（上接第 44 页）

5 单治易，安新颖，关陟昊. 2015—2019 年国际医学信息学计量研究［J］. 中国数字医学，2021，16（1）：88－95.

6 MOED H F. Measuring contextual citation impact of scientific journals［J］. Journal of informetrics，2010，4（3）：265－277.

7 陈斯斯，刘春丽. 基于学科规范化引文影响力和相对引用率的论文学术影响力评价［J］. 中华医学图书情报杂志，2021，30（12）：25－30.

8 HIRSCH J E. An index to quantify an individual's scientific research output［J］. Proceedings of the national academy of sciences of the United States of America，2005，102（46）：16569－16572.

9 季淑娟，董月玲，王晓丽. 基于文献计量方法的学科评价研究［J］. 情报理论与实践，2011，34（11）：21－25.

10 钟伟金，李佳，杨兴菊. 共词分析法研究（三）——共词聚类分析法的原理与特点［J］. 情报杂志，2008（7）：118－120.

11 虞秋雨，徐跃权. 共词分析中高频词阈值确定方法的实证研究——以新冠肺炎文献高频词选取为例［J］. 情报科学，2020，38（9）：90－95.

12 丁敬达，陈一帆，刘超，等. 基于共词和 Word2Vec 加权向量的文献——主题语义匹配分析方法［J］. 图书情报工作，2022，66（12）：108－116.

13 周晓分，黄国彬，白雅楠. 科学计量可视化软件的对比与数据预处理研究［J］. 图书情报工作，2013，57（23）：64－72.

14 VAN ECK N J，WALTMAN L. Text mining and visualization using VOSviewer［J］. ISSI newsletter，2011，7（3）：50－54.

15 李艳，张悦，曾可，等. 文献信息分析工具的比较［J］. 中华医学图书情报杂志，2015，24（11）：41－47.

16 高凯. 文献计量分析软件 VOSviewer 的应用研究［J］. 科技情报开发与经济，2015，25（12）：95－98.

17 周晓英，裴俊良. 健康信息学的学科范畴与中国健康信息学的发展——兼述健康信息学学科建设与发展学术研讨会［J］. 中国图书馆学报，2022，48（2）：76－93.

18 钱庆. 医学信息学发展现状与展望［J］. 中华医学信息导报，2020，35（18）：10.

19 崔少国，陈俊桦，李晓虹. 融合语义及边界信息的中文电子病历命名实体识别［J］. 电子科技大学学报，2022，51（4）：565－571.