中医诊治高血压医疗实体提取问题研究

庞 震 顾继昱 吴宇飞 颜仕星

（中国中医科学院西苑医院 北京 100091） （北京中医药大学研究生院 北京 100105） （上海道生医疗科技有限公司 上海 201203）

李汪洋 孙 越

（上海中医药大学 上海 201203） （国家卫生健康委卫生发展研究中心 北京 100044）

〔摘要〕 提出一种基于三元组信息抽取策略的新型实体提取模型，以解决传统命名实体识别方法应用于高血压中医电子病历医疗实体识别时出现的实体离散问题，阐述实验数据集及相关处理、实验方法与结果，为中医医疗实体自动化抽取提供方法学参考。

〔关键词〕 高血压；中医病历；命名实体识别；BERT 联合抽取

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2021.09.009

Study on the Medical Entity Extraction Problems in Traditional Chinese Medicine Diagnosis and Treatment of Hypertension

PANG Zhen, *Xiyuan Hospital*, *China Academy of Chinese Medical Sciences*, *Beijing 100091*, *China*; *GU Jiyu*, *WU Yufei*, *Graduate School of Beijing University of Chinese Medicine*, *BeiJing 100105*, *China*; *YAN Shixing*, *Shanghai Daosh Medical Technology Co. Ltd.*, *Shanghai 201203*, *China*; *LI Wangyang*, *Shanghai University of Traditional Chinese Medicine*, *Shanghai 201203*, *China*; *SUN Yue*, *China National Health Development Research Center*, *Beijing 100044*, *China*

〔**Abstract**〕 The paper proposes a new entity extraction model based on the triplet information extraction strategy to solve the entity discrete problem when the traditional Named Entity Recognition (NER) method is applied to the medical entity recognition of hypertension Electronic Medical Record (EMR) of Traditional Chinese Medicine (TCM), expounds the experimental data set and related processing, experimental methods and results, and provides methodological references for automatic extraction of TCM medical entity.

〔**Keywords**〕 hypertension; Traditional Chinese Medicine (TCM) medical records; Named Entity Recognition (NER); BERT joint extraction

1 引言

推行电子病历、规范中医术语是传承临床经验的必要手段。引入医疗实体识别，进行中医医疗实体识别研究能够起到促进作用。医疗实体识别旨在

〔修回日期〕 2021－08－25

〔作者简介〕 庞震，硕士，工程师，发表论文 5 篇；通讯作者：孙越，博士。

从医疗文本数据中划分实体边界、检测实体类别，是自然语言处理任务中的基础研究之一。

2 命名实体识别研究现状

2.1 研究成果

当前命名实体识别（Named Entity Recognition，NER）研究已取得一定成果。NER 常用方法包括基于词典和规则、基于统计学习、基于深度学习的方法等。近年来深度学习发展迅速，从早期应用的循环神经网络（Recurrent Neural Network，RNN），卷积神经网络（Convolutional Neural Network，CNN），到如今主流的长短期记忆人工神经网络（Long - Short Term Memory，LSTM），门控循环单元（Gated Recurrent Unit，GRU），NER 方法得到更广泛的应用和发展，在医疗实体识别中占据重要地位。其中 LSTM 是一种特殊的 RNN 类型，能够解决 RNN 长依赖问题。GRU 则为 LSTM 变体，在克服长依赖问题的同时在数据量较大时比 LSTM 网络具有更快的训练速度。

2.2 存在问题

传统 NER 方法将命名实体识别归类为简单序列标注范畴，即给定 1 个序列（1 段文本），对序列中的每个元素（汉字）仅做 1 个标记，再使用算法识别某些特定标记模式。然而在中医医学病历抽取领域，医疗实体大多存在实体嵌套和实体非连续等问题，即实体识别中的实体序列分散导致边界样本混淆，无法正确识别医疗实体的问题[1]。如医疗文本“后背痛，有发紧抽动感”“发紧”“抽动感”等症状表现词语和“后背”症状部位词语是非连续的。传统基于序列标注的 NER 方法抽取的医疗实体仅为“后背痛”“发紧”“抽动感”，对于“发紧”“抽动感”无法识别具体部位（“后背”）。实际需抽取的医疗实体应表示为“后背痛”“后背发紧”“后背抽动感”，见表 1、图 1。

表 1 高血压医疗实体提取

示例	文本	传统 NER 抽取实体	实际医疗实体	关系抽取方案
示例 1	后背痛，有发紧抽动感	“后背痛” “有发紧”“抽动感”	“后背痛” “后背发紧” “后背抽动感”	（痛，位于，后背）、 （发紧，位于，后背）、 （抽动感，位于，后背）
示例 2	苔根部薄黄	“苔根部薄黄”	“苔根部薄” “苔根部黄”	（黄，位于，苔根部）、 （薄，位于，苔根部）
示例 3	脉沉无力	“脉沉无力”	“脉沉”“脉无力”	（沉，位于，脉）、 （无力，位于，脉）
示例 4	不伴恶心、呕吐、流泪、畏光	“不伴恶心”“呕吐” “流泪”“畏光”	“不恶心”“不呕吐”“不流泪” “不畏光”	（不，描述，恶心）、（不，描述，呕吐）、（不，描述，流泪）、 （不，描述，畏光）
示例 5	右腿外侧麻木，天冷时撕裂样疼痛	“右腿外侧麻木” “撕裂样疼痛”	“右腿外侧麻木” “右腿撕裂样疼痛”	（麻木，位于，右腿外侧）、 （撕裂样疼痛，位于，右腿外侧）
示例 6	后背凉不适	“后背凉不适”	“后背凉” “后背不适”	（凉，位于，后背）、 （不适，位于，后背）

2.3 解决方法

本文提出一种基于三元组抽取策略的高血压医疗实体提取方法，通过关系抽取技术解决中医实体识别出现的实体离散问题。例如针对上述举例中的医疗文本，通过关系抽取技术抽取三元组信息（痛、位于、后背）（发紧、位于、后背）（抽动感、位于、后背），重组后得到医疗实体“后背痛”“后背发紧”“后背抽动感”。在针对真实世界高血压中医病历的计算机模拟实验中选择 BiGRU、BERT - base - chinese、BERT_ TCM、词嵌入、CASREL、对抗训练等多种关系抽取算法和训练技术进行组合比较，以找到

最佳的高血压医疗实体三元组抽取策略。

3 实验数据集及相关处理

3.1 实验数据来源

实验数据来自中国中医科学院西苑医院，为非结构化高血压门诊病历数据集。设定病案纳入标准为：参照《中国高血压防治指南（2018 年修订版）》，临床诊断为高血压的患者；排除标准为：血压、血糖控制不达标者；心功能Ⅳ级者，急性冠脉综合征者，心源性休克、恶性心律失常、心脏瓣膜患者；肝肾功能不全、呼吸系统疾病者；脑卒中、肿瘤疾病者及合并感染者；精神异常不能配合者及药物过敏者。根据设定病案纳入与排除标准筛选病案数据，选取 2 000 例高质量病案进行研究。

3.2 数据标注内容

3.2.1 概述 定义 3 类细粒度的症状实体类型对症状词组进行拆分，定义 2 类关系类型对划分出的细粒度实体进行重组。

3.2.2 实体类型 根据症状实体类别在医疗文本中所占成分的不同，将症状细粒度分为“症状”“部位”和“程度”。（1）症状。症状语素：可独立表示某一症状具体含义的最小语言单位，用于描述患者异常感觉或体征，如“痛”“失眠”等；症状性质：反映症状语素特征，如“颈肩部刺痛”中的“刺”。（2）部位。人体部位：包括人体解剖部位、形体官窍和脏器等，如头、背、腹、经络、腧穴等；分泌物与排泄物：包括可直接观察到的人体正常分泌物、排泄物和病理产物，如汗、脓液等；生命特征：包括体温、呼吸、心跳、脉象、舌象、面色、食纳、睡眠、情志等；症状方位：表示症状发生的具体方位，如“小腿后侧感觉障碍”中的“后侧”。（3）程度。症状变化：表现症状改变及发展趋势，如“膝肿胀好转”中的“好转”；症状程度：症状发作的频率、严重程度等，如“频繁腹泻”中的“频繁”。（4）暂时不考虑的实体类型。症状发生时间：症状发生或依赖的时间；症状发生条件：影响疾病发生的体内外各种因素。根据上述定义对实验数据进行实体信息统计，见表 2。

表 2 实体信息统计

实体类型	个数	百分比（%）
部位	19 154	47.82
症状	20 897	52.18
程度	4 448	11.11
总计	40 051	100.00

3.2.3 关系类型 为保证高血压症状抽取中重组实体的便捷性和症状的完整性，首先根据已设定的实体类型将抽取目标具体为：<部位 + 症状（语素）：描述 >。由于全身性症状如“过敏”“打嗝”“晕厥”等在病历记述中未提到症状发生部位，因此“部位”实体为可选类型。据此可进一步将需要抽取的关系定义为两类，即症状实体“位于”部位实体、程度实体“描述”症状实体：一是 <Subject：症状，predicate：位于，Object：部位 >；二是 <Subject：程度，predicate：描述，Object：症状 >。对实验数据进行关系标注统计，见表 3。此外在试标注过程中发现一些关于分泌物和病理产物异常的表述组合为“部位”实体 + 若干“描述”实体，已定义实体关系无法覆盖此类情况，新增加关系类型可能导致标签分散、样本不均衡，因此约定标注时将所有分泌物和病理产物类异常的描述性实体统一标注为“症状”，如“痰/部位黄/症状语素黏/症状语素”，见图 1。

表 3 关系标注统计

关系类型	个数	百分比（%）
位于（located_ at）	19 149	78.41
描述（is_ a_ description_ of）	5 274	21.59
总计	24 423	100.00

3.3 数据标注方案

采取单人标注及监督校对的方式进行数据标注。标注者要求具备本科以上中医临床相关专业学位；选取少量病历文本供 5 位参与研究的标注者进行标注练习并组织一致性讨论，熟悉标注规范并形成标注共识；将全部 2 000 份病历文本随机、平均分成 5 份，分别由 5 位标注者进行标注；选择 1 名资深高血压术语专家作为监督者，对标注结果进行 1 轮审核和校正。

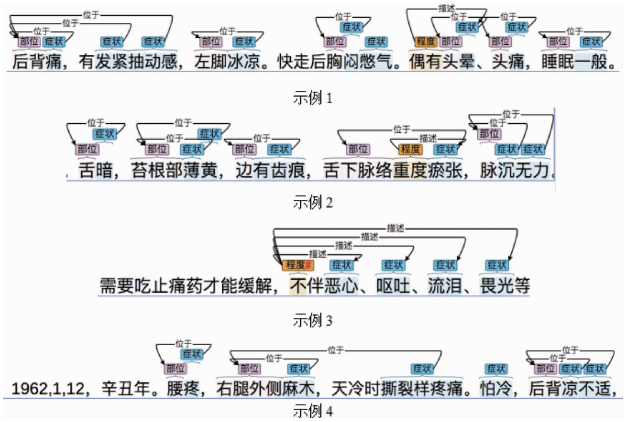

图 1 标注示例

4 实验方法

4.1 概述

组合多种深度学习自然语言处理算法和策略，训练出数个基于深度学习的关系抽取模型，在此基础上提取、重组三元组信息，获取并比较最终症状抽取结果。实验环境为深度学习服务器 1 台，配置 NVIDIA 3090TI 显卡 1 块，内存 24G，CPU8 核。

4.2 实验环境与算法介绍

4.2.1 BiGRU 模型 GRU 是 RNN 的一种变体，与 LSTM 一样可解决长期记忆和反向传播中的梯度问题，但相对 LSTM 网络而言结构更简单、效果更佳，是当前应用较多的一种网络。单向神经网络结构通常是从前往后输出状态。在关系抽取中，如果当前时刻输出能与前后某一时刻状态均产生联系，将利于文本深层次特征提取。因此同时使用两个单向且反向的 GRU 组成 BiGRU 神经网络模型。在每一时刻同时提供两个输入并由两个 GRU 共同决定输出。

4.2.2 联合抽取模型 该模型[2]共 5 个组件：编码层，将输入序列转换为字符级嵌入；双向 LSTM 层，进行命名实体识别；条件随机场（Conditional Random Fields，CRF）层，进行命名实体识别；标签嵌入，输入实体标签，输出学习到的标签嵌入表示，编码命名实体知识，应用于关系抽取；多头选择评分的预测模型层。给定序列 token $S_{char} = [char_1, char_2, \cdots, char_n]$，BERT 模型提供预训练，表征 token 进行词，将 token 作为 NER 送入 biLSTM 和 CRF 层，再输出预测实体标签 $h_i = [\overrightarrow{h_i}, \overleftarrow{h_I}]$。通过最小化交叉熵损失 L_{ner} 来训练 NER，获取标签嵌入层预测的实体标签 $char_i$，见图 2。实验中关系抽取为多头选择问题，Head 表示实体的最后一个 token；多头即假设每个实体都可能与任何其他实体有关系。给定 1 个关系标签集合 R 以及 token 向量 S_{char}: $char_i$ and $char_j$，多头选择模型旨在预测这两者在属于 R 的标签关系中的概率。给定关系 r_k，关系抽取的 score 计算如下：

$$s^{(r)}(z_j, z_i, r_k) = V^{(r)} f(U^{(r)} z_j + W^{(r)} z_i + b^{(r)}) \quad (1)$$

其中 z_j、z_i 是 token 和的 NER 阶段输出。通过降维和激活将得到的 score 转化为 token$char_j$。

被选为关系标签的 r_k 的头部的概率 $char_i$ 为：

$$Pr(head = char_j, label = r_k \mid char_i) = \sigma(s^r(z_j, z_i, r_k)) \quad (2)$$

其中 $\sigma(\cdot)$ 表示 sigmoid 函数。

关系抽取训练过程是最小化损失函数的交叉熵：

$$L_{re} = \sum_{i=0}^{n} \sum_{j=0}^{m} -logPr(head = y_{i,j}, label = r_{i,j} \mid char_i) \quad (3)$$

其中 y_i 属于输入序列，r_i 属于关系集 R；y_i 和 r_i 分别表示头部的标注及关系标签；m 为关系标记数。基于等式（2），训练后将设置 1 个阈值来决定是否保留 $\widehat{y_i}$ 和 $\widehat{r_i}$ 组合。对于实体和关系抽取的联合学习，最终目标损失为 $L_{ner} + L_{re}$。

4.2.3 CASREL 模型[3] 为基于主语感知的层叠式指针网络，该模型抽取过程为：先通过遍历标注样本所有主语建立训练集，再依次抽取主语、谓语和宾语。依次定义 1 个头实体标注器、关系特定标注器，找寻句子中所有可能的头实体，再对每个找到的头实体使用所有关系的关系特定标注器找到相关关系和对应的尾实体。同时实现对主语的感知表征并通过改进 BERT 分词器适应中英文和特殊标点同时出现的场景，见图 3。

4.2.4 对抗训练 引入对抗训练快速梯度法（Fast Gradient Method，FGM）与投影梯度下降[4]（Projected Gradient Descent，PGD）提高模型性能。通过对抗训练将噪声引入模型，利用大量未标注的样本对参数进行正则化，可以有效防止模型过拟

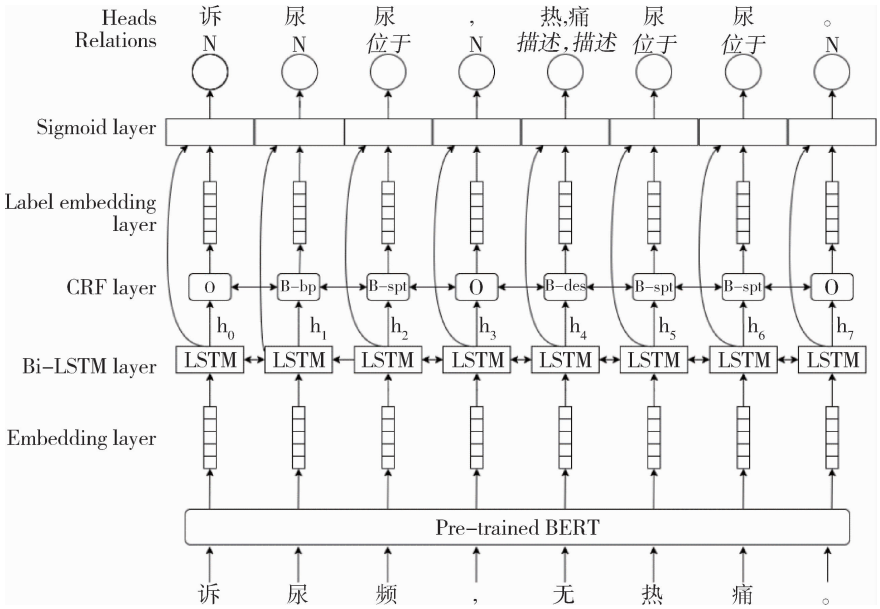

图 2 联合抽取框架

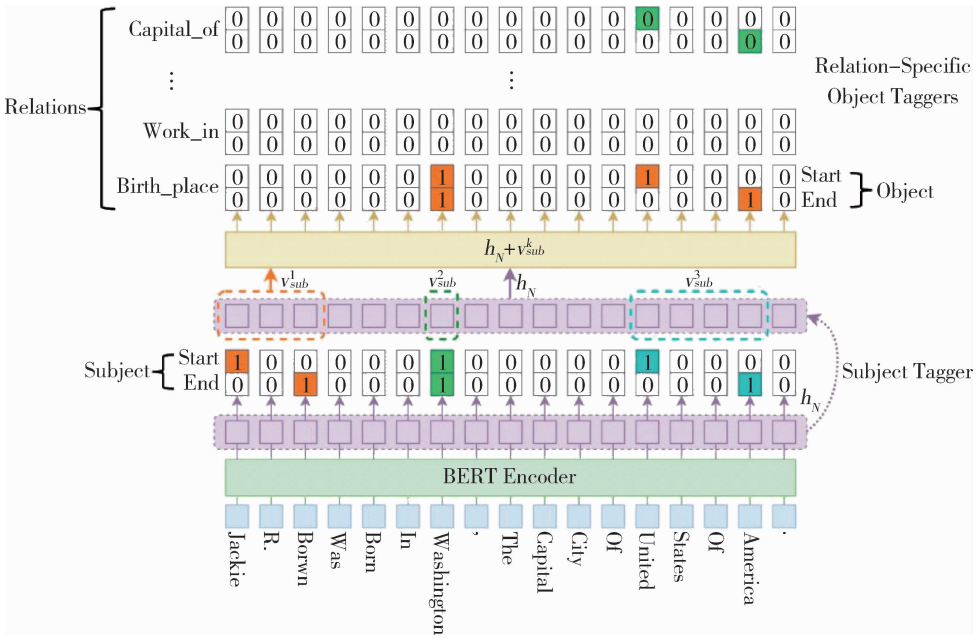

图 3 CASREL 模型框架

合，提升其泛化能力与鲁棒性。使用 BERT 预训练的语言模型进行参数初始化，在关系抽取及实体提取的嵌入层、BERT 的词嵌入层添加对抗扰动。

4.3 模型评价（评估指标）

为评估三元组抽取策略的可行性和所应用模型在高血压中医病历中的性能，使用精确率、召回率、*F*1 值作为评估指标。其中预测标签与证型标注标签中任意一个重合则为正确，将各个标签正确率按位加权，得出最终正确率指标。*F*1 值计算公式如下：

$$F1\ 值 = \frac{2precision * recall}{precision + recall} \quad (4)$$

4.4 实验步骤及参数设置

一是文本处理。使用 BERT 模型[5]提取文本特征信息。二是训练与调参。联合抽取模型参数设置：词嵌入维度为 512；Dropout 为 0.5；学习率为 1e－06。CASREL 模型参数设置：词嵌入维度为

512；Dropout 为 0.5；学习率为 5e－5。

5 实验结果

5.1 基于不同 BERT 的联合抽取模型性能比较

5.1.1 概述 在大型中医临床语料库基础上训练针对中医特定场景的 BERT_ TCM，以此预训练模型作为联合抽取模型中的 pre－trained BERT，对实验数据进行实体和关系联合抽取。与文献[6]中的 BiGRU 模型进行对比，以研究 BERT 在提升关系抽取任务性能方面的效果。

5.1.2 无词嵌入时的性能比较 基于 BERT_ TCM 的联合抽取模型的各项性能指标远高于 BiGRU 模型，特别是代表平均性能的 *F*1 值。这表明 BERT 预训练语言模型充分利用了无监督的预训练和人工标注的训练数据，可提升关系抽取性能，见表 4。

表 4 无词嵌入时基于 BERT_ TCM 的联合抽取模型与 BiGRU 模型的关系抽取性能比较

模型	是否词嵌入	*F*1 值	精确率	召回率
基于 BERT_ TCM 的联合抽取	否	0.782 9	0.820 2	0.783 0
BiGRU	否	0.689 5	0.825 3	0.592 1

5.1.3 加入词嵌入后的性能比较 进一步研究词嵌入技术对关系抽取模型性能的影响。结果表明在有词嵌入的条件下，基于 BERT_ TCM 的联合抽取模型 *F*1 值比 BIGRU 模型高出 14%；同时两种模型在词嵌入后效果均提升，如联合抽取模型的 *F*1 值在词嵌入后提升了 8%，见表 5。

表 5 加入词嵌入后基于 BERT_ TCM 的联合抽取模型与 BiGRU 模型的关系抽取性能比较

模型	是否词嵌入	*F*1 值	精确率	召回率
基于 BERT_ TCM 的联合抽取	是	0.863 7	0.910 2	0.821 7
BiGRU	是	0.725 9	0.835 5	0.641 7

5.1.4 联合抽取模型性能比较 通过比较基于 BERT_ TCM 的联合抽取模型与基于 BERT－base－chinese（通用中文预训练 BERT）的联合抽取模型可知，BERT_ TCM 相比 BERT－base－chinese 在高血压中医病历关系抽取任务上效果有所提升，见表 6。

表 6 基于 BERT_ TCM 与基于 BERT－base－chinese 联合抽取模型的关系抽取性能比较

模型	是否词嵌入	*F*1 值	精确率	召回率
基于 BERT－base－chinese 的联合抽取	否	0.702 8	0.827 1	0.6110
	是	0.859 2	0.915 5	0.809 5
基于 BERT_ TCM 的联合抽取	否	0.782 9	0.820 2	0.783 0
	是	0.863 7	0.910 2	0.821 7

5.2 CASREL 模型和联合抽取模型性能比较

CASREL 模型作为一种新型级联二进制标注框架，能够将关系建模为一个从头实体映射到尾实体的函数，将任务中心从标注分类转移至找寻三元组信息，能够适应中医医学病历中三元组实体重叠的场景，在同一时间提取医疗文本中的多个关系三元组。本文添加 PGD 和 FGM 两种对抗训练技术作为防御机制，对比不同条件下 CASREL 模型和联合抽取模型的性能差异，以期找到性能最优的高血压病历三元组抽取策略。基于 BERT_ TCM 的 CASREL 模型在未添加对抗时，*F*1 值 = 0.880 7，远超过联合抽取模型。即使使用通用的 BERT－base－chinese，CASREL 模型在未添加对抗时 *F*1 值仍高于联合抽取模型。此外添加 PGD 和 FGM 后基于 BERT_ TCM 的 CASREL 模型性能有所提升，见表 7。

表 7 CASREL 抽取实验结果

模型	对抗训练	*F*1 值	精确率	召回率
基于 BERT_ TCM 的联合抽取模型	无	0.863 7	0.910 2	0.821 7
	PGD	－	－	－
	FGM	－	－	－
基于 BERT－base－chinese 的 CASREL 模型	无	0.872 5	0.859 6	0.885 9
	PGD	0.884 1	0.881 0	0.887 2
	FGM	0.885 0	0.889 2	0.880 9
基于 BERT_ TCM 的 CASREL 模型	无	0.880 7	0.872 7	0.888 9
	PGD	0.888 7	0.884 1	0.893 2
	FGM	0.885 8	0.883 0	0.888 6

5.3 实验结果举例

对于文本“多汗，晨起手肿胀无力，盗汗，舌苔薄黄腻，舌质淡，口唇色红，大便 3～5 天一行，脉沉细弱，脉左稍弦”，获取三元组：[（‘肿’‘位于’‘手’），（‘胀’‘位于’‘手’），（‘无力’‘位于’‘手’），（‘薄’‘位于’‘舌苔’），（‘黄’

‘位于’‘舌苔’)，(‘腻’‘位于’‘舌苔’)，(‘淡’‘位于’‘舌质’)，(‘色红’‘位于’‘口唇’)，(‘3～5 天一行’‘位于’‘大便’)，(‘沉’‘位于’‘脉’)，(‘细’‘位于’，‘脉’)，(‘弱’‘位于’‘脉’)，(‘弦’‘位于’‘脉左’)，(‘稍’‘描述’‘弦’)，(‘多’‘位于’‘汗’)]。重组三元组从而获取实际的症状抽取结果为：“手肿”“手胀”“手无力”“盗汗”“舌苔薄”“舌苔腻”“舌苔黄”“舌质淡”“口唇色红”“大便 3～5 天一行”“脉沉”“脉细”“脉弱”“脉左稍弦”。

6 结语

本文提出一种基于三元组抽取策略的高血压医疗实体提取模型，有效解决传统 NER 无法解决的中医实体识别中出现的实体离散问题。实验发现基于大型中医临床语料库训练出的针对中医特定场景的 BERT_ TCM，与常规中文语料库训练出的 BERT - base - chinese 相比，在中医高血压病历关系抽取任务中具有更好的性能。与仅进行单一关系抽取的 BiGRU 模型相比，联合抽取模型显著提高了各项性能指标。可能的原因是联合抽取同时提取实体和关系，避免实体识别任务中 CRF 层语义信息丢失。CASREL 模型性能比联合抽取模型更加优越，在使用相同预模型 BERT_ TCM 且未添加对抗情况下，CASREL 模型的 *F*1 值远超联合抽取模型。此外引入对抗训练技术能够有效提升模型鲁棒性。

参考文献

1 Zheng S, Wang F, Bao H, et al. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme [C]. Vancouver: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017.

2 Giannis B, Johannes D, Thomas D, et al. Joint Entity Recognition and Relation Extraction as a Multi - head Selection Problem [J]. Expert Systems with Application, 2018 (114): 34 - 45.

3 Wei Z, Su J, Wang Y, et al. A Novel Hierarchical Binary Tagging Framework for Relational Triple Extraction [EB/OL]. [2020 - 06 - 22]. https://arxiv.org/abs/1909.03227v2.

4 Qin C, Martens J, Gowal S, et al. Adversarial Robustness through Local Linearization [C]. Vancouver: Proceedings of the 33rd Annual Conference on Neural Information Processing Systems, 2019.

5 Devlin J, Chang M W, Lee K, et al. BERT: Pre - training of Deep Bidirectional Transformers for Language Understanding [C]. Minneapolis: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

6 Yi R, Hu W. Pre - trained BERT - GRU Model for Relation Extraction [C]. Beijing: Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition, 2019.

(上接第 16 页)

14 Library of Congress. Recommended Formats Statement 2020 - 2021[EB/OL].[2020-10-19]. https://www.loc.gov/preservation/resources/rfs/RFS%202020 - 2021.pdf.

15 ANDS Guide. File Formats [EB/OL]. [2020 - 10 - 19]. https://www.ands.org.au/_ _data/assets/pdf_file/0003/731775/File - Formats.pdf.

16 Libraries Digital Conservancy. About the Data Repository [EB/OL].[2020 -10 -21]. https://conservancy.umn.edu/pages/drum/.

17 University of Leicester. Good Practice and Guidance - document Version Control Chart (Draft) [EB/OL]. [2020 - 10 - 19]. https://www2.le.ac.uk/services/research - data/old - 2019 - 12 - 11/documents/UoL_ VersionControlChart_ d0 - 1.pdf.

18 Deankin University Library. Where Should I Store My Digital Data?[EB/OL]. [2020 - 10 - 21]. https://www.deakin.edu.au/library/research/manage - data/store/where - should - i - store - my - digital - data.

19 中华人民共和国国家质量监督检验检疫总局，中国国家标准化管理委员会. GB/T32843 - 2016 科技资源标识 [EB/OL]. [2021 - 07 - 20]. https://wenku.baidu.com/view/77931fbd7dd5360cba1aa8114431b90d6c8589ae.html.

20 钱毅. 基于长期保存视角的电子档案格式管理研究 [J]. 档案学通讯，2016，4 (6)：52 - 57.

21 DuraSpace Organization. Fedora Commons Repository Developer Documentation [EB/OL]. [2021 - 07 - 20]. http://docs.fcrepo.org/.