·数据挖掘·

# 结合数据增强与实体映射CasRel模型的名家医案联合关系抽取

李钰欣[1]，向兴华[1]，杨航[1,2]，刘大胜[1]，王嘉恒[1]，赵志伟[1]，韩嘉旭[1,2]，吴孟洁[1]，车前子[1*]，杨伟[1*]

（1. 中国中医科学院 中医临床基础医学研究所，北京 100700；
2. 湖南科技大学 数学与统计学院，湖南 湘潭 411201）

［摘要］ **目的：**针对中医名家医案的非结构化文言表述、实体关系嵌套及标注数据稀缺问题，构建结合数据增强与实体映射的联合关系抽取框架，为中医诊疗知识图谱构建及临床规律挖掘提供技术支撑。**方法：**构建名家医案文本实体及其关系的标注结构，采用数据增强策略，整合多部古籍扩充医案关系抽取数据集，设计适配中医语义的基于级联二值标记的关系联合抽取（CasRel）模型，引入中医经典文本预训练双向编码器表征法（BERT）编码层，增强对古汉语的语义表征，采用头实体-关系-尾实体映射机制，同步解决实体嵌套与关系重叠问题。**结果：**相较于基于流水线的Bert-Radical-Lexicon（BRL）-双向长短期记忆网络-注意力机制（BiLSTM-Attention）模型，结合数据增强与实体映射的联合关系抽取CasRel模型展现出了更为显著的性能优势，在病症关系、舌证关系、因证关系、方证关系等共12类关系的综合精确率为65.73%、召回率为64.03%、$F_1$值为64.87%，比流水线的BRL-BiLSTM-Attention模型的综合精确率、召回率、$F_1$值分别提升14.26%、7.98%、11.21%。其中舌证关系（$F_1$值为69.32%，提升22.68%）提升显著，方证关系表现最优（$F_1$值为70.10%，提升9.93%）。**结论：**该研究通过数据增强与联合解码，显著改善中医文本的语义隐含与实体间复杂依赖性问题，为中医医案结构化挖掘提供可复用技术框架，所构建的知识图谱可支撑临床辨证选方与用药配伍优化，也为中医人工智能研究提供方法论参考。

［关键词］ 数据增强；名家医案；关系抽取；联合方法；基于级联二值标记的关系联合抽取（CasRel）模型；知识图谱
［中图分类号］ R242；R249；TQ018 ［文献标识码］ A ［文章编号］ 1005-9903（2026）02-0218-08
［doi］ 10.13422/j.cnki.syfjx.20251866
［网络出版地址］ https://link.cnki.net/urlid/11.3495.R.20250801.0938.002
［网络出版日期］ 2025-08-01 10:22:38 ［增强出版附件］ 内容详见http://www.syfjxzz.com或http://cnki.net

## Joint Relation Extraction of Famous Medical Cases with CasRel Model Combining Entity Mapping and Data Augmentation

LI Yuxin[1], XIANG Xinghua[1], YANG Hang[1,2], LIU Dasheng[1], WANG Jiaheng[1], ZHAO Zhiwei[1], HAN Jiaxu[1,2], WU Mengjie[1], CHE Qianzi[1*], YANG Wei[1*]

(1. *Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing* 100700, *China*; 2. *School of Mathematics and Statistics, Hunan University of Science and Technology, Xiangtan* 411201, *China*)

［**Abstract**］ **Objective:** To address the challenges of unstructured classical Chinese expressions, nested entity relationships, and limited annotated data in famous traditional Chinese medicine (TCM) case records, this study proposes a joint relation extraction framework that integrates data augmentation and entity mapping, aiming to support the construction of TCM diagnostic

［收稿日期］ 2025-04-30
［基金项目］ 国家自然科学基金项目（82405243）；中国中医科学院科技创新工程项目（CI2023C066YLL，CI2021B003）；中国中医科学院自主选题项目（Z0643，Z0723）
［第一作者］ 李钰欣，在读硕士，从事中医状态辨识与风险预测方法研究，E-mail：15696803840@163.com
［通信作者］ *车前子，博士，助理研究员，从事中医临床研究设计与数据统计研究，E-mail：cheqianzi123@126.com；
*杨伟，博士，副研究员，从事中医药观察性数据的统计学习及因果推断方法研究，E-mail：yangyxq@ruc.edu.cn

knowledge graphs and clinical pattern mining. **Methods**: We developed an annotation structure for entities and their relationships in TCM case texts and applied a data augmentation strategy by incorporating multiple ancient texts to expand the relation extraction dataset. A cascade binary tagging framework for relation triple extraction (CasRel) model for TCM semantics was designed, integrating a pre-trained bidirectional encoder representations from transformers (BERT) layer for classical TCM texts to enhance semantic representation, and using a head entity-relation-tail entity mapping mechanism to address entity nesting and relation overlapping issues. **Results**: Experimental results showed that the CasRel model, combining data augmentation and entity mapping, outperformed the pipeline-based Bert-Radical-Lexicon (BRL)-bidirectional long short-term memory (BiLSTM)-Attention model. The overall precision, recall, and $F_1$-score across 12 relation types reached 65.73%, 64.03%, and 64.87%, which represent improvements of 14.26%, 7.98%, and 11.21% compared to the BRL-BiLSTM-Attention model, respectively. Notably, the $F_1$-score for tongue syndrome relations increased by 22.68% (69.32%), and the prescription-syndrome relations performed the best with the $F_1$-score of 70.10%. **Conclusion**: The proposed framework significantly improves the semantic representation and complex dependencies in TCM texts, offering a reusable technical framework for structured mining of TCM case records. The constructed knowledge graph can support clinical syndrome differentiation, prescription optimization, and drug compatibility, providing a methodological reference for TCM artificial intelligence research.

[**Keywords**] data augmentation; famous medical cases; relationship extraction; joint learning approach; cascade binary tagging framework for relation triple extraction (CasRel) model; knowledge graph

中医名家医案作为历代医家临床经验的载体，系统记载了医者辨证论治的决策路径及其疗效验证过程[1]。然而，其非结构化文言表述、实体关系多元嵌套及标注数据稀缺等特性，制约了中医知识的现代化挖掘与应用。近年来，随着自然语言处理技术的发展，通过对医案知识进行挖掘研究以实现从经验描述向结构化知识的转化，这不仅为构建中医诊疗知识库提供了数据支撑，也为临床决策提供历史经验参照与数据化支持。目前，在中医领域面临着海量名家医案数据与临床实践需求之间的明显差距，如何突破中医文本的语言壁垒和结构复杂性，将非结构化的医案转化为可复用的诊疗知识体系，成为中医药现代化亟须解决的关键问题。

随着人工智能技术的发展，数字中医药已成为传承创新的重要突破口。国家层面正通过《“数据要素×”三年行动计划》在中医药等领域积极探索医疗健康数据流通新模式[2]。在此背景下，如何从海量非结构化中医文本中高效提取诊疗知识，成为支撑数据流通与临床决策的关键环节，而关系抽取技术则是实现这一目标的核心手段。关系抽取方法经历了从基于规则和词典的早期方法[3-4]，到传统机器学习[5-7]、深度学习[8-10]的发展过程，涌现了开放领域[11]、少次学习[12]和领域自适应[13]等更复杂和灵活的技术。但是古籍医案多以非结构化的文本方式进行记录和保存，这对信息的识别和提取带来了挑战：在语言表达层面，文本多以文言文记载，用词晦涩、语义含蓄；在知识结构层面，实体多嵌套、关系多重叠；在数据资源层面，缺乏适配古籍医案的统一标注规范，文本标注数据稀缺，标注成本高且领域适配性差。现有的医疗文本关系抽取研究多集中于疾病、症状和处方等特定要素[14]，其适用场景多局限于现代规范化病历，尚未构建符合中医语言特点的标注框架，也未有效解决实体嵌套与关系重叠问题，导致其在中医古籍知识抽取任务中的适用性受限。

针对中医古籍医案知识抽取面临的文言文语义解析困难、实体关系复杂嵌套及标注数据稀缺等问题，本研究提出一种方法，即构建适用于医案命名实体识别和实体关系抽取的9类实体及12类关系标注结构，提出一种结合数据增强与实体映射的基于级联二值标记的关系联合抽取（CasRel）模型，引入中医经典文本预训练双向编码器表征法（BERT）编码层，采用头实体-关系-尾实体映射机制，同步解决实体嵌套与关系重叠问题。与传统依赖简单同义词替换或句子重构的数据增强方法不同，本研究通过整合多部中医古籍文本扩展医案关系抽取的数据集，提升训练数据的多样性和覆盖面，以应对中医文本中的复杂表述和数据稀缺问题，增强模型的性能和应用范围。本文以高血压的中医古籍医案为例，以验证方法的有效性和实用性。

## 1 方法原理

随着深度神经网络和大型预训练语言模型的迅速发展，关系抽取的性能得到了大幅提升。这些方法通常分为两大类：流水线[15]关系抽取和联合[16]关系抽取，分别代表了不同的任务处理策略。

**1.1** 流水线关系抽取 流水线关系抽取方法分两步执行，第一步是通过命名实体识别（NER）模块实现文本内实体定位，第二步将检测出的实体按序两两组合为候选对，结合原句文本输入关系分类（RC）模块完成语义关联判断，见增强出版附加

材料。

以 Bert-Radical-Lexicon（BRL）神经网络模型为 NER 模块、双向长短期记忆网络-注意力机制（BiLSTM-Attention）模型为 RC 模块组成 BRL-BiLSTM-Attention 模型进行医案文本的流水线关系抽取。BRL 模型由嵌入层、编码层与解码层 3 部分组成，其中嵌入层将字符转化为嵌入向量，包含 3 个模块，分别为字符嵌入模块、关联词嵌入模块及部首嵌入模块[17]，显著提升了字符向量的语义丰富度，增强了字符表示的全面性，进一步提升了实体识别的准确性。BiLSTM-Attention 是一种用于关系分类的经典模型，结合了双向长短期记忆网络和注意力机制的优势，能够更有效地捕捉上下文之间的关系和聚集关键信息。

**1.2** 联合关系抽取 联合关系抽取是对实体识别与关系抽取任务进行联合建模，利用实体和关系间的交互信息同时提取实体和关系。CasRel 模型是一种基于联合解码的实体关系抽取模型，其核心思想是把关系建模为将头实体映射到尾实体的函数[18]，可同时提取医案文本的实体及关系，组成形如（实体 1，关系，实体 2）的关系三元组，不受重叠三元组问题影响。见图 1。

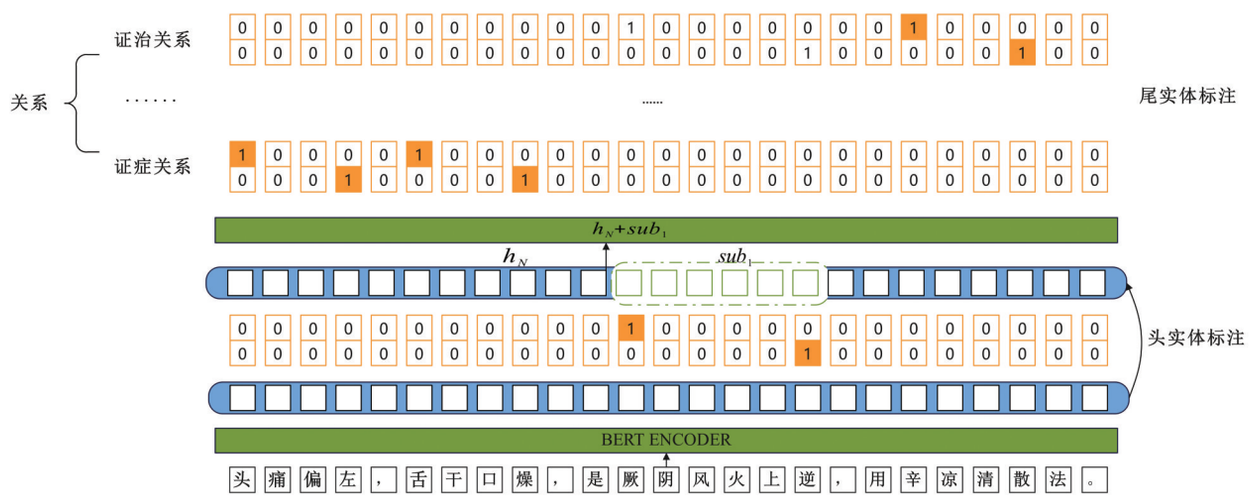


**图 1 CasRel 模型的结构**

**Fig. 1 Structure of CasRel model**

该模型包含编码端和解码端 2 个组成部分。编码端基于 BERT 的编码层用于获取上下文语义信息对字或者词进行表征；解码端又包括头实体识别层、关系与尾实体联合识别层：头实体识别层通过二分类的方式，使用线性层加上 sigmoid 激活函数来判断是否为头实体的开始或结束位置，然后通过最近匹配原则将识别到的开始和结束位置配对，以获得候选头实体集合；关系与尾实体联合识别层根据已识别的头实体，寻找可能的关系和尾实体。每一层尾实体识别层的结构与头实体识别层相似，但输入时会考虑头实体的特征。

**1.3** 数据增强流程 本研究所构建的中医古籍数据增强流程，采用四步递进式扩展关系抽取数据集。首先筛选多源古籍数据，系统整合中医典籍的 3 类核心古籍（本草类、方药类、医经类）。再挖掘关系特征，筛选潜在关系表述语句，通过关键词匹配技术从古籍中初筛出潜在关系的句子，进而构建初筛语料库。随后进行知识验证标注，由中医专家基于标注平台（“百部知识引擎”平台/百度 easydata 数据标注平台）对初筛语料库进行语义校验与关系修正。最后构建关系抽取数据集，将验证后的有效关系三元组进行整合，形成具有中医语言特征的增强型关系抽取数据集。

## 2 资料与方法

**2.1** 数据来源 《中华历代名医医案全库》[19]收录了 1750—1966 年间的 200 余部医案专著和 300 余位名医的诊疗记录，总计超过 15 000 则医案文献。本研究从中筛选出章节标题含“头痛”“眩晕”等高血压相关症状的医案共 408 篇。本文采取“数据标注、数据增强、实验设置、知识应用”的技术路线，见图 2。实验设计遵循了 NLP 领域的报告模板。

**2.2** 数据标注 本研究参考了国家标准《中医药学语言系统语义网络框架（GB/T 38324-2019）》[20]，并根据《中华历代名医医案全库（上中下）》文本内容的体例格式和医案文本中的医家诊疗思路提出适合医案关系抽取的标注结构，见增强出版附加材

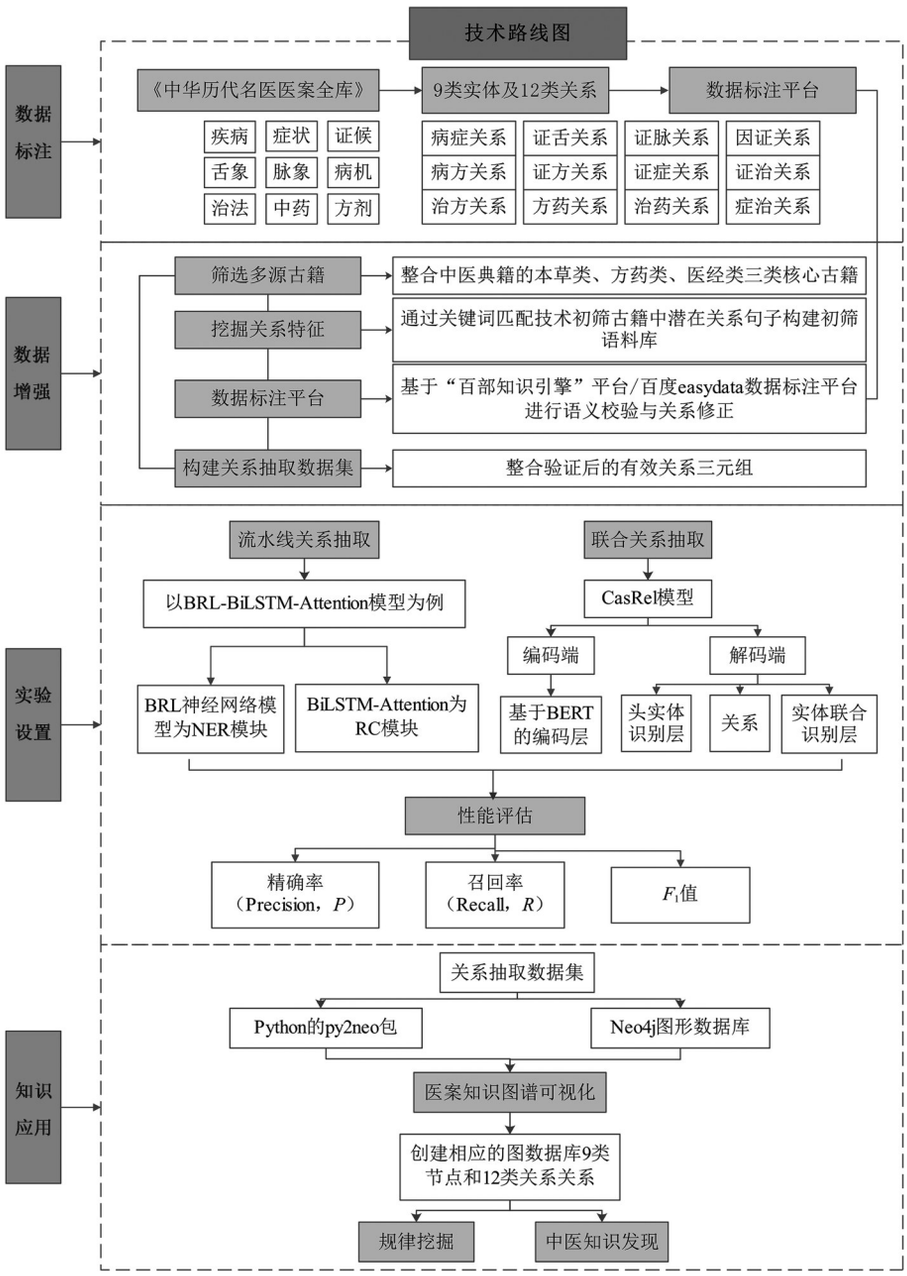


**图2　名家医案联合关系抽取的技术路线**

**Fig. 2　Technical route of joint relation extraction of famous medical records**

料。在这个标注结构框架中，节点代表医案中的实体，具体包括疾病、舌象、脉象、症状、病因病机、证候、治法、中药及方剂共9类实体。边则代表这些实体之间的关系，包括病症关系、舌证关系、脉证关系、因证关系、方病关系、方证关系、证症关系、证治关系、治方关系、方药关系、治药关系及症治关系共12类关系，见表1。

将408篇医案的PDF数据转换为txt格式，对转换后的数据进行人工校对，以修正字符错误和文本结构问题。将清洗完成后的txt文件上传至中国中医科学院中医药信息研究所开发的“百部知识引擎”标注平台，采用词典自动匹配与人工修正相结合的标注方式，对名家医案文本进行实体和关系的标注。

### 2.3　CasRel模型优化

**2.3.1**　数据增强　本研究筛选出包含两个实体且能明确表达表1中定义语义关系的关系三元组，这一过程旨在构建医案关系抽取数据集。然而，由于中医医案文本语言的表述特点与现代白话文存在显著差异，特别是其省略和简练的语言风格，导致实体关系的标注难度较大且数量有限。从408篇医案中初步只获得了239条关系数据。鉴于数据集规

**表1 实体关系定义**

**Table 1 Entity relationship definitions**

| 关系类型 | 关系描述 | 头实体 | 尾实体 |
|---|---|---|---|
| 病症关系 | 中医疾病临床所见症状 | 疾病 | 症状 |
| 证舌关系 | 证候表现出的舌象特征 | 证候 | 舌象 |
| 证脉关系 | 证候表现出的脉象特征 | 证候 | 脉象 |
| 因证关系 | 病因病机体现出的证候特征 | 病因病机 | 证候 |
| 病方关系 | 疾病适用方剂 | 疾病 | 方剂 |
| 证方关系 | 证候适用方剂 | 证候 | 方剂 |
| 证症关系 | 证候临床所见症状 | 证候 | 症状 |
| 证治关系 | 证候适合用某些治法 | 证候 | 治法 |
| 治方关系 | 治法适合用某些方剂 | 治法 | 方剂 |
| 方药关系 | 方剂由某些中药组成 | 方剂 | 中药 |
| 治药关系 | 治法适合用某些中药 | 治法 | 中药 |
| 症治关系 | 症状适合用某些治法 | 症状 | 治法 |

模较小，直接训练出的模型难以准确捕获实体间的复杂关系。为应对这一难题，本研究采取了数据增强策略[21-22]，通过从GitHub下载包括《神农本草经》等本草类著作、《肘后急备方》等方药类资料及《伤寒论》等医经类经典在内的704部中医古籍文本，补充关系标注数据的来源，从而提升模型的训练效果和泛化能力。从这些古籍文本中筛选出含有“以”“用”“乃”等关键字的潜在关系句子[23]共119 410条，并利用百度easydata数据标注平台进行标注工作。见增强出版附加材料。

**2.3.2** 联合抽取 (1)数据划分：通过人工标注，从古籍文本潜在关系句子中筛选出包含2个实体且能明确表达表1中定义语义关系的关系数据，本研究共获得4 161条关系数据。将这4 161条新标注的关系数据与先前从《中华历代名医医案全库(上中下)》获取的239条医案关系数据合并，最终形成了一个含有4 400条记录的医案关系抽取数据集。数据源自《中华历代名医医案全库》及古籍文本，经人工校验与格式清洗后无缺失值，故无需特殊处理。为了有效训练并评估关系抽取模型的性能，用python的random库对数据集进行随机划分，随机种子设置为200，该数据集被随机分成3个部分：训练集、验证集和测试集，划分比例为6∶2∶2。将训练集和验证集用于模型训练和筛选，以确定最优模型，最后使用测试集数据对两种最优模型进行评估。(2)实验参数：以CasRel模型进行医案文本的联合关系抽取实验，以BRL神经网络模型为NER模块、BiLSTM-Attention为RC模块进行医案文本的流水线关系抽取实验，实验在Windows系统环境下使用Python 3.6.15及包pytorch 1.10.2实现。其中最大句子长度设置为200，batchsize为8，epoch为50，dropout为0.5，学习率在$10^{-4}$~$10^{-3}$内经过微调得出在取值为$5\times10^{-4}$时表现出最优效果。(3)模型性能评估：对于关系抽取，当对应实体的左右边界和关系类型均正确时，则认为预测的关系是正确的。评估模型性能常用的指标为精确率($P$)、召回率($R$)和$F_1$值来评价。$TP$表示正确预测的三元组数量，$FP$表示错误预测的三元组数量，$FN$表示未预测的三元组数量。$P$表示正确预测的三元组与所有预测的三元组的比例，$R$表示数据集中正确预测的三元组与所有三元组的比例，$F_1$值是精确率和召回率的综合评价指标。$P$、$R$、$F_1$的计算公式如下。

$$P = \frac{TP}{TP + FP} \times 100\% \quad (1)$$

$$R = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\% \quad (3)$$

**2.3.3** 知识应用 采用CasRel关系抽取模型对中医古籍文本进行结构化处理，完成实体识别及关系抽取。对于抽取出的实体和关系数据，通过人工审核的方式剔除可能存在的错误数据和不准确的关联，确保数据的质量和准确性。人工审核过程包括对识别出的实体类别、实体名称及其之间关系的再次验证与修正，排除歧义和不合理的关联，最终形成高质量的实体节点文件和关系三元组文件。将抽取结果构建知识图谱，支持临床决策与规律挖掘。在知识图谱构建阶段，借助Python的py2neo包实现与Neo4j图形数据库的高效交互，基于前期提取的实体节点和关系三元组数据，创建相应的图数据库节点和关系。在Neo4j中，每个实体节点被表示为一个图中的节点，每一对实体间的关系被表示为节点之间的关系。构建完成的知识图谱不仅能够为中医古籍文本的深度分析提供强有力的支持，还为后续的知识发现、智能问答系统和中医药研究提供了丰富的资源和数据支持。

## 3 结果

**3.1** 性能对比 从综合性能来看，CasRel模型在精确率($P$=65.73%)、召回率($R$=64.03%)和$F_1$值($F_1$=64.87%)上均优于流水线方法BRL-BiLSTM-Attention模型($P$=51.47%、$R$=56.05%、$F_1$=53.66%)，其综合$F_1$值较BRL-BiLSTM-Attention模型提升11.21%，表明CasRel模型在中医文本的复杂语义理

解与关系建模中更具优势。在舌证关系三元组中，CasRel的$F_1$值达到69.32%，较流水线方法（46.64%）提升22.68%，其高召回率（70.55%）表明模型能有效捕捉舌象与证候间的隐含关联；在方证关系三元组中，CasRel的$F_1$值达70.10%，较流水线方法提升13.93%。在因证关系三元组中，CasRel的$F_1$值为54.75%，但也比流水线方法BRL-BiLSTM-Attention模型的$F_1$值提升11.25%。CasRel模型在多数任务中实现了$P$与$R$的均衡提升。在方药关系三元组中，其$P$（64.93%）与$R$（68.57%）均高于流水线（57.32%、57.98%），表明模型既能减少误判，又能覆盖更多真实关系。而在证症关系三元组任务中，其$R$（62.59%）显著高于流水线方法（61.22%），进一步验证了对多样化症状描述的泛化能力。模型在不同关系类型上的性能存在显著差异。CasRel模型在舌证关系中的$P$表现最为突出，在方证关系中的$R$值和$F_1$值均达到较高水平。相比之下，该模型在因证关系中的$P$、$R$及$F_1$均表现相对较低，表明其在识别因证关系时存在不足，需要进一步优化以提高适用性。BRL-BiLSTM-Attention模型$P$、$R$、$F_1$分别在证治关系、证症关系、方证关系表现较好，在舌证关系、因证关系表现较差。两类模型在因证关系上的$F_1$值均为最低，而在方证关系上表现最佳。见表2。

**表2 CasRel和BRL-BiLSTM-Attention模型关系抽取性能对比**

**Table 2 Comparative performance of CasRel and BRL-BiLSTM-Attention models in relation extraction**

| 关系三元组 | CasRel | | | BRL-BiLSTM-Attention | | |
|---|---|---|---|---|---|---|
| | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ |
| 疾病，病症关系，症状 | 65.82 | 61.65 | 63.64 | 48.64 | 52.53 | 50.51 |
| 证候，舌证关系，舌象 | 68.19 | 70.55 | 69.32 | 43.57 | 50.18 | 46.64 |
| 证候，脉证关系，脉象 | 68.07 | 60.44 | 64.25 | 59.14 | 51.51 | 55.32 |
| 病因病机，因证关系，证候 | 52.17 | 57.84 | 54.75 | 44.91 | 42.17 | 43.50 |
| 疾病，方病关系，方剂 | 58.35 | 61.65 | 60.00 | 48.54 | 52.11 | 50.32 |
| 证候，方证关系，方剂 | 67.81 | 72.55 | 70.10 | 57.89 | 62.63 | 60.17 |
| 证候，证症关系，症状 | 58.73 | 62.59 | 60.60 | 54.50 | 61.22 | 57.66 |
| 证候，证治关系，治法 | 63.58 | 60.79 | 62.15 | 60.52 | 57.73 | 59.09 |
| 治法，治方关系，方剂 | 62.55 | 66.53 | 64.48 | 47.53 | 51.61 | 49.49 |
| 方剂，方药关系，中药 | 64.93 | 68.57 | 66.71 | 57.32 | 57.98 | 57.65 |
| 治法，治药关系，中药 | 67.54 | 60.33 | 63.94 | 59.30 | 55.82 | 57.51 |
| 症状，症治关系，治法 | 60.22 | 58.97 | 59.60 | 58.13 | 54.29 | 56.14 |
| 综合 | 65.73 | 64.03 | 64.87 | 51.47 | 56.05 | 53.66 |

**3.2** 知识图谱展示 为了进一步提升模型的实用性，可以将医案关系三元组以知识图谱的形式可视化表示，并查询与分析相关实体间的联系，能挖掘出更多有价值的治疗方法和医学理论，为现代中医临床提供参考和启示。基于CasRel模型，通过整合名家医案与古籍预测数据构建三元组知识图谱，完成对9类实体及12类实体关系的结构化展示。该知识图谱包含68 929个实体实例及103 277条关系实例，实现了领域知识的体系化建模与关联表达。例如，头痛分为多种证型，例如厥阴头痛、血虚头痛、风热头痛等，其中风热头痛的症状有肢节烦疼、项背拘急，中药用石膏、荆芥穗、菊花、细辛、生绿豆，方剂用川芎散、龙脑芎辛丸；血虚头痛的症状包括遍身肢节痛、目痛脑疼、自觉头脑俱空、遍身痛，病因是产后伤风，脉象为脉近数，治法有润风燥经，药用当归、川芎、细辛等，方剂有养血胜风汤、芎归汤；厥阴头痛与血虚头痛均可使用细辛配伍治疗。见增强出版附加材料。

## 4 讨论

本研究针对中医古籍医案知识抽取中的文言文语义解析、实体关系复杂嵌套及标注数据稀缺问题，构建了融合数据增强与实体映射机制的CasRel联合抽取模型，该框架通过数据增强策略扩展标注数据集，结合头实体-关系-尾实体的映射机制实现双层解码，在名家医案关系抽取的整体$F_1$值和方证关系的抽取准确率方面展现了显著的性能优势，有效应对了中医文本的复杂语义结构与标注数据不足问题，为中医医案的结构化知识挖掘提供了可复用的技术框架，其核心方法在中医诊疗规律、人工智能应用发现等领域具有潜在的应用价值。在知识体系构建层面，形成包含9类中医实体与12类实体关系的标注体系，覆盖中医辨证论治的核心知识结构。在临床应用层面，识别和抽取中医医案里的实体及其相互关系，构建中医诊疗知识图谱，系统解析名医诊疗经验中的隐性治疗模式并将其转化为结构化知识网络，既为现代中医临床提供经典范式参考以优化治疗方案，也为中医诊疗规律的深度解析提供数据支撑。

相关研究通过联合学习与数据增强策略提升关系抽取性能，为本研究提供方法依据。中药专利文本实体关系联合抽取（TPSCRE）模型针对中药专利文本实体重叠问题，通过语义特征与多层交叉注意力机制实现联合抽取[24]；数据增强策略通过语义特征的全面学习，有效提升模型在复杂场景下的泛化能力。针对药物不良反应标注数据稀缺问题，提出数据增强与半监督学习融合方法，缓解标注不足

对检测模型性能的影响[25]；另有研究[26]整合多源医学文本，通过样本多样性扩充、语义索引构建实现知识图谱建模。本研究结合两种策略，探索其在名家医案关系抽取中的应用。性能提升主要源于中医预训练BERT强化古汉语语义表征，实体映射解决嵌套与重叠，多源数据增强覆盖12类关系分布。构建的9类实体及12类实体关系标注体系，不局限于单一要素或者几种特定要素的抽取[27]，实现了从碎片化要素提取到诊疗关系网络构建的范式升级。区别于通用领域的简单句子重构，数据增强策略通过整合中医3类核心古籍（本草类、方药类、医经类）以扩展标注语料，使模型习得文言文语境下的语义映射规律，通过多样化的临床表述提升模型泛化能力，缓解小样本场景下的过拟合问题，增强对中医复杂语义结构的适应性。然而，模型在因证关系上的表现较弱，在方证关系上的表现较好，可能与因证关系在数据集中样本量较少有关。这一结果也与中医临床实际情况相符，方证关系在中医诊疗中具有明确的临床意义和较高的出现频率，因而模型能够更好地学习和提取相关信息。

本研究也存在一定的局限，当前缺乏同类联合模型的对比分析，后续可在中医语料库不断完善的基础上，探索更多联合抽取模型在中医复杂关系场景中的适配性。实验验证基于高血压单病医案，尽管标注框架与技术方案具有跨疾病通用性，但其在消渴、中风等复杂疾病场景中的适配性仍需进一步验证。因证关系等低频复杂关系的抽取性能受限于样本稀疏性与语义隐含性，模型对长距离依赖及抽象病因的捕捉能力有待优化。未来可以引入图神经网络或Transformer长距离依赖建模机制，构建多层次语义关联网络，以增强对隐含病因、病机演变等深层语义关系的抽取能力，推动中医临床文本挖掘技术向精细化、智能化方向发展。

综上，本研究针对中医古籍名家医案的语义解析与知识结构化难题，构建融合数据增强与实体映射的CasRel联合抽取模型，通过多源语料扩展与双层解码机制，实现对中医名家医案实体关系的高效抽取，整体$F_1$值与方证关系准确率表现显著。这一方法不仅契合中医药临床案例强调“理法方药”逻辑一致性的特点，也为中医药临床案例“真实规范性、特色优势性与应用转化度”三维评价体系在知识挖掘层面提供了可操作的实现路径[28]，使中医个体化经验能够更加系统、规范地融入现代中医药知识体系，为中医药学术传承、循证研究的深化提供有力的技术支撑。

［利益冲突］ 本文不存在任何利益冲突。

［参考文献］

［1］ 何宇浩，李明，罗晓兰，等. 基于GPTs的中医知识图谱实体和关系抽取研究［J］. 上海中医药杂志，2024，58（8）：1-6.
HE Y H，LI M，LUO X L，et al. Research on entity and relation extraction for traditional Chinese medicine knowledge graphs based on GPTs［J］. Shanghai J Tradit Chin Med，2024，58（8）：1-6.

［2］ 国家数据局，中央网信办. 关于印发《“数据要素×”三年行动计划（2024—2026年）》的通知［EB/OL］.（2024-01-04）［2025-05-25］. https://www. nda. gov. cn/sjj/zhuanti/ztsjysx/qt/0902/20240830174038137859023_pc. html.
National Data Bureau，Cyberspace Administration of China. Notice on Issuing the "Three-Year Action Plan for Data Elements×"（2024-2026）［EB/OL］.（2024-01-04）［2025-05-25］. https://www. nda. gov. cn/sjj/zhuanti/ztsjysx/qt/0902/20240830174038137859023_pc. html.

［3］ GUO X Y，HE T T，YUAN J，et al. Relation dictionary construction and rule learning for PPI extraction from biomedical literatures［C］//Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine（BIBM）. Washington，USA：IEEE Computer Society，2015：1133-1140.

［4］ WANG X Z，LI J H，ZHENG Z，et al. Entity and relation extraction with rule-guided dictionary as domain knowledge［J］. Front Eng Manag，2022，9（4）：610-622.

［5］ LEI J B，TANG B Z，LU X Q，et al. A comprehensive study of named entity recognition in Chinese clinical text［J］. J Am Med Inform Assoc，2014，21（5）：808-814.

［6］ JIANG M，CHEN Y K，LIU M，et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries［J］. J Am Med Inform Assoc，2011，18（5）：601-606.

［7］ YADAV S，RAMESH S，SAHA S，et al. Relation extraction from biomedical and clinical text：Unified multitask learning framework［J］. IEEE/ACM Trans Comput Biol Bioinform，2022，19（2）：1105-1116.

［8］ SHERSTINSKY A. Fundamentals of recurrent neural network（RNN）and long short-term memory（LSTM）network［J］. Phys D Nonlinear Phenom，2020，404：132306.

［9］ GAO W C，ZHENG X H，ZHAO S S，et al. Named entity recognition method of Chinese EMR based on BERT-BiLSTM-CRF［J］. J Phys Conf Ser，2021，1848（1）：012083.

［10］ LI X Y，ZHANG H，ZHOU X H，et al. Chinese clinical named entity recognition with variant neural structures based on BERT methods［J］. J Biomed Inform，2020，107（5）：103422.

［11］ ZHAN J L，ZHAO H. Span model for open information extraction on accurate corpus［J］. Proc AAAI Conf Artif

Intell, 2020, 34(5): 9523-9530.

[12] WANG Y Q, YAO Q M, KWOK J T, et al. Generalizing from a few examples: A survey on few-shot learning[J]. ACM Comput Surv, 2021, 53(3): 1-34.

[13] RAKIN S, SHIBLY M A R, HOSSAIN Z M, et al. Leveraging the domain adaptation of retrieval augmented generation models for question answering and for hallucination reduction[C]//The 22nd International Conference on Information Technology-new Generations (ITNG 2025). Cham: Springer Nature Switzerland, 2025: 482-493.

[14] XU W X, WANG L, ZHANG M C, et al. A joint entity relation extraction method for document level traditional chinese medicine texts[J]. Artif Intell Med, 2024, 154: 102915.

[15] MIWA M, BANSAL M. End-to-end relation extraction using LSTMs on sequences and tree structures[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: Association for Computational Linguistics, 2016: 1105-1116.

[16] NAYAK T, NG H T. Effective modeling of encoder-decoder architecture for joint entity and relation extraction[J]. Proc AAAI Conf Artif Intell, 2020, 34(5): 8528-8535.

[17] 杨航, 彭叶辉, 杨伟, 等. 基于BRL神经网络模型的名家医案实体识别[J]. 中国实验方剂学杂志, 2024, 30(24): 167-173.
YANG H, PENG Y H, YANG W, et al. Entity recognition in famous medical records based on BRL neural network model[J]. Chin J Exp Tradit Med Form, 2024, 30(24): 167-173.

[18] WEI Z P, SU J L, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online. Association for Computational Linguistics, 2020: 1476-1488.

[19] 鲁兆麟. 中华历代名医医案全库[M]. 北京: 北京科学技术出版社, 2015: 1837-1936.
LU Z L. The complete compendium of medical cases by renowned physicians in Chinese history[M]. Beijing: Beijing Science and Technology Press, 2015: 1837-1936.

[20] 中国标准化研究院. 健康信息学 中医药学语言系统语义网络框架: GB/T 38324—2019[S]. 北京: 中国标准出版社, 2019.
China National Institute of Standardization. Health informatics-Semantic network framework of traditional Chinese medicine language system: GB/T 38324—2019[S]. Beijing: China Standards Press, 2019.

[21] 杨延云, 杜建强, 聂斌, 等. 融合数据增强和注意力机制的中医实体及关系联合抽取[J]. 智能计算机与应用, 2023, 13(8): 186-191, 196.
YANG Y Y, DU J Q, NIE B, et al. Entity and relationship joint extraction method for traditional medical textintegrating data augmentation and attention mechanism[J]. Intell Comput Appl, 2023, 13(8): 186-191, 196.

[22] LUO J G, YANG Y, DU J Q, et al. Joint extraction method of entity relationship in Chinese medicine based on data augmentation and deep learning[C]//Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering. New York, USA: Association for Computing Machinery, 2023: 1349-1358.

[23] 朱玲, 朱彦, 杨峰, 等. 基于中医疾病相关语义关系的正则表达式及知识抽取研究[J]. 世界科学技术—中医药现代化, 2016, 18(8): 1241-1250.
ZHU L, ZHU Y, YANG F, et al. Knowledge extraction research for semantic expression of diseases in Chinese medicine[J]. World Sci Technol Mod Tradit Chin Med, 2016, 18(8): 1241-1250.

[24] 邓娜, 喻卓群, 但文俊, 等. 一种融合语义特征和多层交叉注意力机制的中药专利文本实体关系联合抽取模型[J]. 数据分析与知识发现, 2025, 9(7): 141-153.
DENG N, YU Z Q, DAN W J, et al. TPSCRE: A joint model for entity and relation extraction in traditional Chinese medicine patent texts by integrating semantic features and multi-layer cross-attention mechanisms[J]. Data Anal Knowl Discov, 2025, 9(7): 141-153.

[25] 佘朝阳, 严馨, 徐广义, 等. 融合数据增强与半监督学习的药物不良反应检测[J]. 计算机工程, 2022, 48(6): 314-320.
SHE C Y, YAN X, XU G Y, et al. Adverse drug reaction detection combined with data augmentation and semi-supervised learning[J]. Comput Eng, 2022, 48(6): 314-320.

[26] 韩普, 马健, 张嘉明, 等. 基于多数据源融合的医疗知识图谱框架构建研究[J]. 现代情报, 2019, 39(6): 81-90.
HAN P, MA J, ZHANG J M, et al. The framework construction of medical knowledge graph based on multi-data source fusion[J]. J Mod Inf, 2019, 39(6): 81-90.

[27] 高佳奕, 杨涛, 董海艳, 等. 基于LSTM-CRF的中医医案症状命名实体抽取研究[J]. 中国中医药信息杂志, 2021, 28(5): 20-24.
GAO J Y, YANG T, DONG H Y, et al. Study on named entity extraction of TCM clinical medical records symptoms based on LSTM-CRF[J]. Chin J Inf Tradit Chin Med, 2021, 28(5): 20-24.

[28] 张凯歌, 张锋, 周波, 等. 建设中国中医药临床案例成果库, 探索中医药临床案例质量评价方法[J]. 中国实验方剂学杂志, 2026, 32(1): 271-276.
ZHANG K G, ZHANG F, ZHOU B, et al. Construction of the China Clinical Cases Library of Traditional Chinese Medicine and exploration of quality evaluation methods for TCM clinical case reports[J]. Chin J Exp Tradit Med Form, 2026, 32(1): 271-276.

[责任编辑 吕冬梅]