基于医案文本的名老中医诊疗知识图谱构建方法及应用*

高晓苑[1]，高文佳[1]，王欣宇[1]，朱　垚[2]，丁海雁[3]，杨　涛[1**]，胡孔法[1]

（1. 南京中医药大学人工智能与信息技术学院　南京　210023；2. 南京中医药大学第一临床医学院　南京　210023；
3. 江苏法迈生医学科技有限公司　南京　210000）

摘　要：目的　以名老中医临床病案为研究对象，设计中医诊疗知识图谱构建方法并进行应用。方法　首先，设计深度学习和正则表达式相结合的医案文本实体抽取方法，对非结构化医案文本的疾病、症状、病机和中药实体进行自动抽取；其次，定义实体关系，利用HAN方法计算实体之间的相关性，形成“实体-关系-实体”三元组；最后，利用图数据库Neo4j进行知识存储、Gephi进行可视化展示并在名老中医诊治肺癌医案上进行应用验证。结果　知识抽取模型对医案实体抽取的精确率、召回率和F_1分别为88.49%、90.02%、89.25%，各指标优于对比方法；通过实体相关性计算共提取了1077条三元组，并成功构建了知识图谱，能够反映名老中医诊治肺癌“病-症-机-药”之间的关系。结论　本文方法可以有效解决名老中医临床医案文本知识的抽取、组织和表达问题，实现了从医案文本到知识图谱的自动化构建过程，相关研究思路和方法可以为名老中医诊疗知识图谱构建提供参考。

关键词：中医医案　名老中医　临床诊疗　知识图谱

doi: 10.11842/wst.20220810008　　中图分类号: R197.323.1　　文献标识码: A

名老中医是当代中医药学术发展的优秀典范，代表着当前中医学术和临床发展的最高水平，开展名老中医经验挖掘和传承具有重要的意义。医案是名老中医学术思想和临床经验的重要载体，深入分析挖掘其中蕴含的专家经验意义重大。然而，医案大多以非结构化或半结构的文本形式存在，文字表述具有专家的个人特点，分析处理较为复杂；此外，如何高效地组织和呈现解析出的医案，以便更好地整理专家经验，也成为困扰中医传承和发展的难题。知识图谱[1]是以“语义网络”为骨架的知识系统，用可视化方式描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系[2]，可以为名老中医临床诊疗经验研究提供技术和方法支撑。

近年来，知识图谱作为一种知识管理和应用的新思路和新方法，受到各行各业学者们的广泛关注和研究。在中医药领域，知识图谱也越来越受到重视，诸多学者开展了知识图谱相关研究。例如，朱丹[3]对姚乃礼教授治疗脂肪性肝病的数据进行人工整理，利用Neo4j数据库构建了名老中医治疗脂肪性肝病的知识图谱，以可视化方式展示脂肪性肝病的病因、病证、治法、方药之间的关系。吕子畔等[4]从名医源流整理、学术思想挖掘和临床经验整理等多个角度探索知识图谱构建和挖掘研究，为名医学术经验传承的系统化、规范化建设提供了参考。刘凡[5]以姚乃礼主任医师脾胃病诊疗临床经验相关文献资料为基础，采用实体抽取、知识融合等技术构建名老中医诊疗经验知识图

收稿日期：2022-08-10

修回日期：2023-02-19

*　国家科学技术部国家重点研发计划课题（2022YFC3500201）：中医肿瘤癌毒病机理论创新研究，负责人：沈卫星；国家自然科学基金委员会面上项目（82174276）：知识和数据协同驱动的中医藏象智能辨证方法研究，负责人：杨涛；国家自然科学基金委员会面上项目（8207458）：基于知识图谱的现代名老中医诊治肺癌用药规律及其机制研究，负责人：胡孔法。

**　通讯作者：杨涛，副教授，主要研究方向：中医药信息学。

谱，并对其治疗慢性胃炎辨治规律进行总结，形成个体化诊疗方案，可以为临床诊疗提供参考。

在中医知识图谱的研究领域，以名老中医的临床医案为基础，构建名老中医诊疗知识图谱较为多见[6-9]。然而，传统依靠人工方式进行医案整理和知识提取，耗时费力，迫切需要建立科学高效的研究模式和方法。鉴于此，本文提出名老中医诊疗知识图谱自动化构建模式，实现从医案文本到知识图谱的自动化构建过程，以期为中医临床医案的自动化分析，以及名老中医经验分析和展示提供参考。

1 中医诊疗知识图谱自动化构建流程及方法

中医诊疗知识图谱构建过程包括知识抽取、实体相关性计算、知识图谱构建3个核心步骤（见图1）。首先，对原始医案数据进行预处理（清洗和标注），形成标注数据集；其次，设计知识抽取模型，在标注数据集上进行训练和测试，当模型达到预设目标后，利用该模型对医案进行知识抽取，形成结构化数据；之后，利用实体相关性方法，计算疾病、症状、病机、药物等体之间的相关性；最后，定义模式层图谱结构，将抽取的知识填充到模式中，形成知识图谱。

1.1 知识抽取方法

医案文本中包含病史、诊断、方药等，其中病史的解析最为复杂，而诊断和中药相对规范。针对医案文本特点和解析需求，本文设计深度学习和正则解析相结合的方法解决复杂的医案文本解析问题，利用深度学习方法解析病史，利用正则表达式进行解析诊断和中药，由于正则解析相对简单，因此不作为本文讨论的重点，而就深度学习方法进行重点讨论。本文使用的深度学习方法包含四层，分别为输入层、ERNIE（Enhanced language Representation with Informative Entities）层、BILSTM（Bi-directional Long Short-Term Memory）层和CRF（Conditional Random Fields）层，模型结构如图2所示。

输入层：采用标准的BIO（Beginning-Inside——Outside）标注方法，将医疗文本中的每个字符标注为“B-X”、“I-X”或者“O”。其中，“B-X”表示此元素所在的片段属于X类型并且此元素在此片段的开头，“I-X”表示此元素所在的片段属于X类型并且此元素在此片段的中间位置，“O”表示不属于任何类型。

ERNIE层：ERNIE[10-11]模型是在BERT[12]模型的基础上，采用基于短语和实体的掩码策略，把一个短语或一个实体作为一个基本单元。在训练期间，同一单元中的所有单词都被掩蔽。通过这种方式，可以在训练过程中依靠大规模的无标注语料，学习到丰富的词级信息和语义信息，有效提升了中文实体识别和关系匹配的效果。Transformer[13]编码器是ERNIE的核心，通过内部的多头注意力机制，学习文本内在的长程语义关联。其中注意力机制相应数学公式为：

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

其中，Q、K、V表示输入向量矩阵，d_k为输入向量的维度，通过计算输入序列中每个词与这个序列中其他词的相互关系，反映序列中不同词之间的关联性及重要程度，然后调整每个词的权重获得每个词的新的表征。

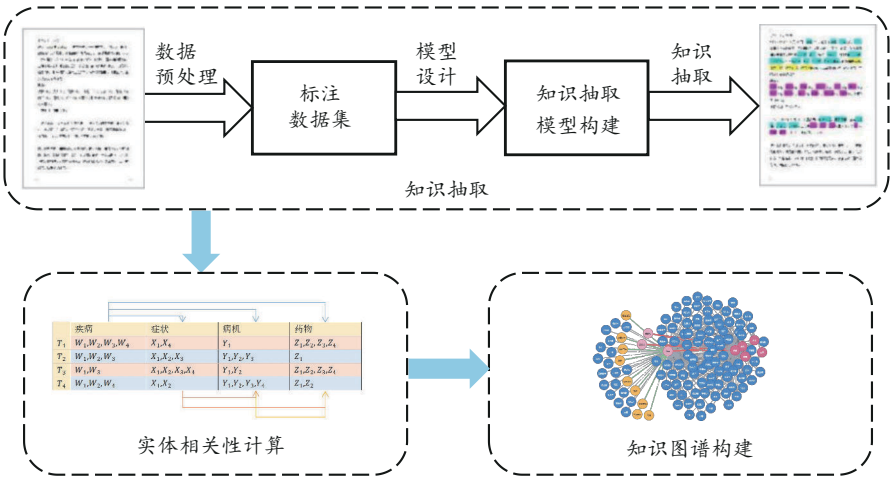


图1 技术流程

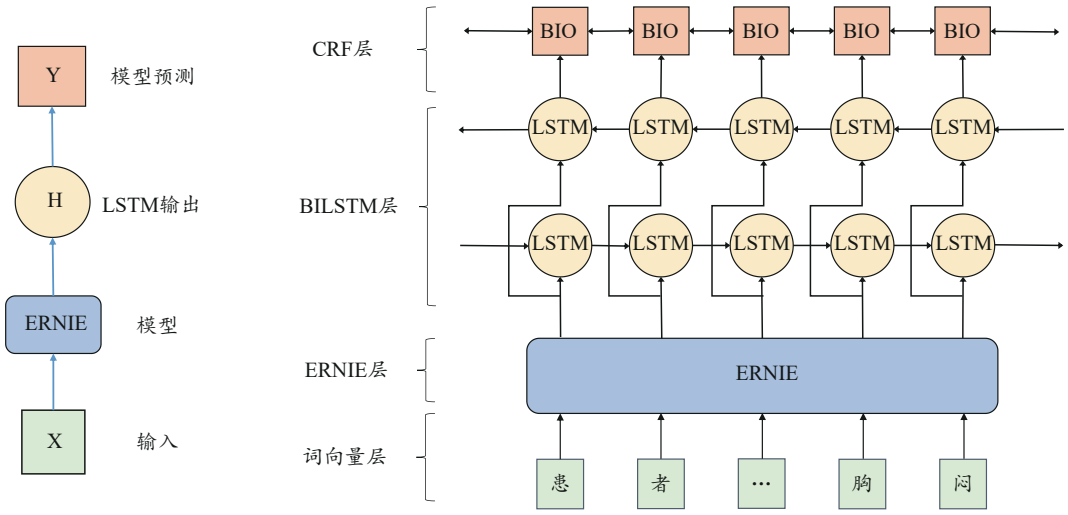


图2　模型结构

BILSTM层：BILSTM[14]模型由前向LSTM（Long Short-Term Memory）和后向LSTM组成。LSTM是一种特殊的循环神经网络，通过增加遗忘门、记忆门和输出门，使得自循环的权重变化，从而解决梯度消失和爆炸问题。将前、后向LSTM结合可以充分利用文本上下文信息特征，公式如下所示。

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(X_t)$$
$$\overleftarrow{h_t} = \overleftarrow{LSTM}(X_t)$$
$$h_t =< \overrightarrow{h_t}, \overleftarrow{h_t} > \quad (2)$$

其中，$\overrightarrow{h_t}$和$\overleftarrow{h_t}$分别代表BILSTM模型的前后向隐藏层的状态，h_t为包含了文本前后向信息的特征。

CRF层：CRF[15]是给定一组输入序列的条件下，另一组输出序列的条件概率分布模型。在命名实体识别任务中，BILSTM解决长距离的文本信息后，标签之间的依赖性问题也亟待解决。例如，预测的标签一般不会出现*B-BW*、*I-CD*并列的情况，因此需要条件随机场（CRF层）来标记全局的最优序列。对给定数据集中输入序列（BILSTM层的输出序列）$x = [x_1, x_2, ..., x_n]$进行处理，输出序列设为$y = [y_1, y_2, ..., y_n]$，为计算最优预测序列的条件分布，再对序列的评价值进行*softmax*[16]归一化，得到相应标注预测的概率值，并以此来进行预测全局的最优序列。针对特定x_i的某种隐含状态y_i的得分计算公式如下：

$$Score(x,y) = \sum_{i=0}^{n} H_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \quad (3)$$

其中，H表示转移矩阵，P_{i,y_i}表示每一个字对应预测标签的分数。

1.2　实体相关性计算方法

中医医案中包含了疾病、症状、病机、中药等不同实体类型，通过矩阵运算计算两两实体之间的相关性，可以为后续知识图谱中不同实体之间的联系紧密程度表征提供依据[17]。本文在团队前期提出的异质关联网络（HeterogeneousAssociated Network，HAN）[18]的基础上进行模型优化，用以解决“病-症-机-药”两两元素之间的组合规律。其计算过程如下：

第一步，构建0-1矩阵。通过命名实体识别模型提取出原始医案中关键信息，得到包含“疾病”、“症状”、“病机”“药物”的l条医案实体集。以“病-症”关系为例，令W表示疾病集的输入集合，X表示疾病集的输入集合；$W(w_1, w_2, \cdots, w_m)$为m种不同的疾病，$X(x_1, x_2, \cdots, x_n)$为n种不同的症状，可以构建相应的0-1矩阵，其中每一行代表一条医案，每一列代表不同的疾病和病机，并分别设为A（疾病）、B（症状）的0-1矩阵。

第二步，筛选二元规则。通过遍历0-1矩阵A、B的每个元素并将矩阵中每两列元素对应相乘后求和计算出两元素的共现频率$f(w_\alpha, x_\beta)$，设定频数阈值$\min f$，设定$f(w_\alpha, x_\beta) \geqslant \min f$的规律$w_\alpha \rightarrow x_\beta$为有效的二元规则。选取*Kulc*系数和不平衡因子*Ir*作为二元规则提取依据，其中*Kulc*系数就是对两个置信度做一个平均处理，而不平衡因子考虑两组数据间的平衡性，值越小，平衡性越高。相关定义如下：

$$Kulc(w_\alpha, x_\beta) = \frac{conf(w_\alpha \rightarrow x_\beta) + conf(w_\beta \rightarrow x_\alpha)}{2} \quad (4)$$

其中：$conf(w_\alpha \rightarrow x_\beta)$为置信度，代表在先决条件$w_\alpha$发

生的情况下，由关联规则 $w_\alpha \rightarrow x_\beta$ 推出 x_β 的比例，即在含有 w_α 的集合中，含有 x_β 的可能性。

$$Ir(w_\alpha,x_\beta)=\frac{|supp(w_\alpha)-supp(x_\beta)|}{(supp(w_\alpha)+supp(x_\beta)-supp(w_\alpha,x_\beta))} \quad (5)$$

其中：$supp(w_\alpha,x_\beta)$ 为支持度，代表 w_α 集合与 x_β 集合共同出现的比例。

1.3 知识图谱构建方法

通过定义疾病、症状、病机、中药实体之间的约束关系（见表1），结合二元规则，得到（实体-关系-实体）三元组，利用 Neo4j 对三元组进行存储，形成知识图谱。

2 中医诊疗知识图谱自动化构建方法应用

2.1 数据准备

本文数据主要来源于国医大师周仲瑛工作室，均为名老中医治疗肺癌的临床病案，经过数据审核、清洗之后，最终得到986诊次医案。其中包括了病史、疾病诊断、病机诊断、处方等内容。

纳入标准：①病案中明确记载诊断包含肺癌的患者；②就诊时的主诉辩治是以肺癌为主者；③数据完整，至少包含标准化西医诊断、标准化临床诊断和标准化病机、标准化方药等内容。

排除标准：①其他系统病症转移到肺癌患者，为非原发性肺癌；②病案记录存在明确错误或缺失。

表1 约束关系汇总表

头实体	关系	尾实体
疾病	有	症状
症状	对应	病机
中药	治疗	疾病
中药	治疗	症状
中药	治疗	病机
病机	引起	疾病

表2 模型的最优结果

模型	最佳 Epoch	Precision	Recall	F_1
BILSTM-CRF	91	0.8077	0.7890	0.7983
Bert-BILSTM-CRF	54	0.8377	0.8500	0.8438
ERNIE-BILSTM-CRF	64	0.8849	0.9002	0.8925

2.2 知识抽取

2.2.1 模型训练及参数设置

采用BIO方法对数据集进行标注，选择80%数据作为训练样本，20%作为测试样本，利用 ERNIE-BILSTM-CRF 模型对数据进行训练和测试。为了验证知识抽取模型的效果，本文还将 ERNIE-BILSTM-CRF 模型与 BILSTM-CRF[19]、BERT-BILSTM-CRF[20]模型进行比较。

模型参数设置为：dropout 为 0.5，选用 SpanFPreRecMetric，以 span 的方式计算准确率（Precision）、召回率（Recall）、F_1，全局学习率（learning_rate）为2e-5，最大训练轮数（max_epoch）为100。

2.2.2 抽取结果

由表2可知，使用 ERNIE-BILSTM-CRF 模型在准确率、召回率和 F_1 上均效果明显优于对比模型，说明本文模型相较于经典模型，对肺癌医案知识抽取表现更佳，抽取更为准确。利用训练好的 ERNIE-BILSTM-CRF 模型对名老中医治疗肺癌医案进行自动化处理，成功抽取出医案中所蕴含的临床症状信息（见表3），为下一步进行实体相关性计算分析提供支撑。

表3 “病-症-机-药”抽取结果展示（TOP 10）

疾病	频次	症状	频次	病机	频次	药物	频次
肺癌	982	咳嗽	404	气阴两虚	346	北沙参	811
糖尿病	81	舌质暗	234	痰瘀阻肺	168	南沙参	767
高血压	79	口干	227	热毒痰瘀阻肺	140	山慈菇	757
前列腺增生	50	苔黄薄腻	218	癌毒走注	90	泽漆	740
脑梗塞	33	舌质暗红	208	肺虚络损	65	太子参	631
肝囊肿	28	大便	154	脾胃虚弱	38	仙鹤草	631
鼻咽部肿瘤	26	脉细	139	肝肾亏虚	25	猫爪草	625
肾囊肿	23	细滑	133	肺虚饮停	24	白花蛇舌草	605
胆结石	17	舌有裂纹	123	肺虚	21	炙僵蚕	535
腰椎间盘突出症	14	脉细滑	121	肺肾两虚	16	肿节风	534

注：疾病和症状为 ERNIE-BILSTM-CRF 模型抽取；病机和药物为正则表达式抽取。

从表3可以看出：从疾病列来看，肺癌病人的病情复杂，除了肺癌还有合并糖尿病、高血压等疾病。从症状和病机来看，患者以咳嗽、舌质暗、口干等症状多见，病机以气阴两虚、痰瘀阻肺最为多见；治疗上以北沙参、南沙参、太子参等药滋阴补气、以山慈菇、泽漆化痰祛瘀。

表4　高频症状相关的“症状-药物”组合规则（Top 10）

症状	药物	*Kulc* 系数	不平衡因子 *Ir*
咳嗽	北沙参	0.7780	0.1488
咳嗽	南沙参	0.7686	0.1043
咳嗽	麦冬	0.7584	0.1044
咳嗽	泽漆	0.7562	0.0747
咳嗽	山慈菇	0.7510	0.0798
舌质暗	北沙参	0.7468	0.2068
舌质暗	麦冬	0.7384	0.165
舌质暗	南沙参	0.7344	0.1635
舌质暗	泽漆	0.7272	0.1352
咳嗽	猫爪草	0.7254	0.0652

注：按 *Kulc* 系数降序排列。

2.3　相关性分析

为了进一步分析“病-症-机-药”之间的相关性。本文以 *Kulc* 大于0.5作为阈值，并从中选取不平衡因子 *Ir* 小于0.5的规则，得到“疾病-症状”74条有效规则、“症状-病机”147条有效规则、“药物-疾病”75条有效规则、“药物-症状”689条有效规则、“药物-病机”88条有效规则、“病机-疾病”4条有效规则。以高频症状相关的“症状-药物”组合规则为例，其结果如表4所示。通过 *Kulc*、*Ir* 指标，从数值角度表征不同变量之间的关系。就“咳嗽”而言，与其相关性较强的中药依次为北沙参、南沙参、麦冬、泽漆、山慈菇等，这与肺癌中晚期，气阴两虚、痰瘀阻肺的治疗思路一致。

2.4　知识图谱构建

按照表1的约束关系，将表4结果转换成三元组（头实体，关系{*Kulc*:*a*，*Ir*:*b*}，尾实体）的形式，其中，关系{*Kulc*:*a*，*Ir*:*b*}表示头实体与尾实体之间的相关性。将所有的三元组导入Neo4j中，总共得到809个节点（实体），1077条边（关系）。由于Neo4j自带的可视化

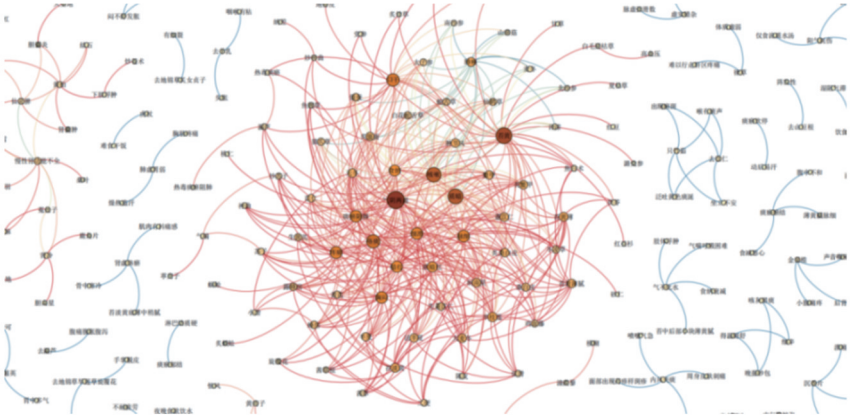
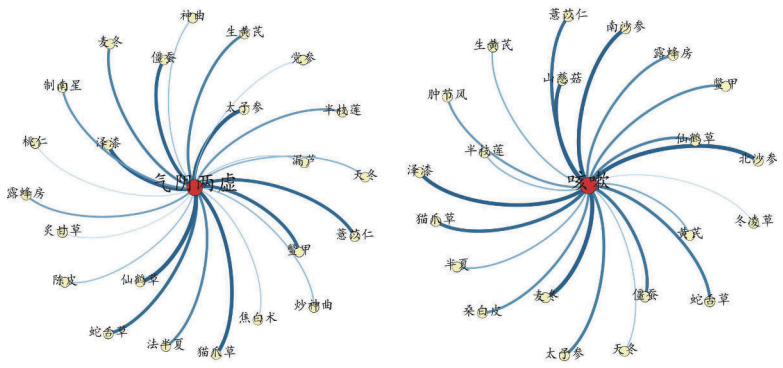


图3　名老中医诊治肺癌知识图谱



(A)　(B)

图4　子图查找

工具无法展示边的强弱关系，因此，本文后端采用Neo4j进行知识存储和管理，前端采用Gephi绘制知识图谱（见图3），提取肺癌主要病机：气阴两虚、痰瘀阻肺等；临床症状：咳嗽、舌质暗、苔黄和咯痰等，其中节点的大小表示该节点的度，边的颜色粗细表示边的权重，即两节点之间的关系亲疏。

Cypher是Neo4j图数据库特有的查询语言，程序化查询匹配某种模式的数据，从而实现语义搜索。例如可以通过查询语句“MATCH(n)-[t:‘治疗’]-(d:‘病机’{name:‘气阴两虚’}) RETURN n;”查找肺癌病例中治疗“气阴两虚”相关的药物，结果见图4A，可观察出针对气阴两虚主要治疗药物有太子参、麦冬、生黄芪等；通过“MATCH p=()-[r:‘治疗’]->(d:‘症状’ {name:‘咳嗽’}) RETURN p;”查询治疗咳嗽相关的药物，结果见图4B，可观察出针对咳嗽这一症状主要治疗药物有北沙参、南沙参、泽漆、山慈菇等。

3 分析及讨论

中医诊疗知识图谱是名老中医知识的有效载体，通过可视化方式，可以直观地提取和展示名老中医的诊疗经验。由于中医临床表述复杂多变，语义关系复杂，采用传统的人工整理方式耗时费力，迫切需要从技术方法层面进行变革。随着自然语言技术的飞速发展，为自动化中医知识抽取和知识图谱构建提供了可能。

本文将自然语言处理技术、关联分析、图谱可视化等诸多技术相结合，提出了一种面向医案文本的中医知识图谱自动化构建方法，并在名老中医医案数据上进行了应用。具体而言，设计深度学习和正则解析相结合的中医医案知识抽取方法，其中重点利用ERNIE-BILSTM-CRF深度模型自动抽取疾病、症状、病机等实体。ERNIE-BILSTM-CRF模型通过对词、实体等语义单元的掩码学习得到完整概念的语义表示，可以有效表征字的多义性，增强了模型语义理解能力，在一定程度上提高了知识抽取效果，有助于更好地自动识别医案文本中的实体并进行分类。在成功抽取实体的基础上，将异质元素（实体）之间的潜在关系转为两两元素之间的有向组合规则，以矩阵运算为基础，利用*Kulc*系数和不平衡因子*Ir*选出有效组合规则；之后将提取出的实体作为节点集合，规则作为边集，利用图数据库Neo4j和可视化展示工具Gephi，构建诊治肺癌的“病-症-机-药”知识图谱。

从实验情况来看，本方法能够初步完成名老中医肺癌诊疗知识图谱构建和展示，“病-症-机-药”间的关系基本符合名老中医临床诊疗规律。研究过程中，仍然遇到以下问题：

①训练数据有限，模型效果有待提升。相较于自然语言处理领域数据集，名老中医临床诊疗的数据规模十分有限，如何在有限的数据集上训练出优秀的中医信息抽取模型显得尤为关键。本文在经典的ERNIE模型的基础上，通过小样本微调技术，虽然能够达到85%以上的精度，但仍存在模型参数量较大、泛化性不强等问题。

②临床信息表述多样，抽取结果需要审核确认。模型虽然能够识别出文本的中医实体，但是由于临床信息表述多样，导致识别出的实体存在多词一义的问题，需要进一步进行实体对齐，并结合人工审核，这无疑给后续图谱构建带来很大的挑战。

③二元关系复杂，筛选依据有待规范。由于患者往往合并多种疾病，诊疗过程综合辨证用药，致使“病-症-机-药”之间的关系极其复杂，利用关联分析从概率统计角度可以发现不同变量之间的二元关系，但是如何选择合理的阈值，如何进行规则筛选，这是本研究面临的一个难题。在研究中，根据文献报告，结合领域知识进行了参数设置，后续还应进一步优化。

在中医临床的诊疗过程中，复杂多变的中医临床症状和表达形式独特的中医专业术语给中医临床的信息抽取带来了极大挑战。中医药领域知识图谱构建目前还面临标准规范不统一、关系抽取技术不成熟、质量控制环节不完善等种种问题，迫切需要从领域标准、技术方法、实施方案、应用场景等角度进行探索。在今后的研究中，要进一步扩大样本量，尝试运用数据增强技术扩增样本，让模型能够充分地学习；其次，尝试引入外部医学知识参与训练，特别是名老中医个性化的临床术语库的加入，有望进一步提升模型对特殊临床表述的识别和抽取能力；最后，要进一步研究二元关系的筛选指标和依据，尝试运用多指标、多方法综合评价，同时充分考虑数据分布不均衡带来的结果偏差，进而为高质量名老中医知识图谱构建提供参考。

参考文献

1 陈悦, 刘则渊. 悄然兴起的科学知识图谱. 科学学研究, 2005, 23(2): 149-154.

2 路畅. 基于科学知识图谱的大数据研究可视化分析. 杭州: 浙江工业大学硕士学位论文, 2019.

3 朱丹. 名老中医治疗脂肪性肝病的证治规律研究及知识图谱构建探索. 北京: 中国中医科学院博士学位论文, 2019.

4 吕子畔, 黄仲羽, 刘凤斌, 等. 基于知识图谱的名医经验传承模式探究. 世界科学技术-中医药现代化, 2020, 22(12):4200-4204.

5 刘凡. 基于知识图谱技术的名老中医慢性胃炎辨证论治方案研究. 北京: 中国中医科学院博士学位论文, 2020.

6 司宜蓓, 郭静, 王永博, 等. 临床实践指南实施性促进研究之四: 中医/中西医结合指南知识图谱知识抽取、存储与实例展示. 医学新知, 2022, 32(2):99-107.

7 左冉, 李敬华, 王映辉, 等. 名医传承知识图谱的构建方法与应用研究. 中国数字医学, 2021, 16(3):33-36.

8 王东军, 孙璇, 田春颖, 等. 基于CiteSpace的亚健康中医体质知识图谱可视化与文献计量分析. 西部中医药, 2022, 35(4):101-105.

9 刘凡, 王明强, 李凌香, 等. 名老中医临床经验知识图谱构建方法探索. 中华中医药杂志, 2021, 36(4):2281-2285.

10 Sun Y, Wang S, Li Y, *et al*. ERNIE: Enhanced representation through knowledge integration. https://arxiv.org/abs/1904.09223.

11 毕云杉, 钱亚冠, 张超华, 等. 基于ERNIE模型的中文文本分类研究. 浙江科技学院学报, 2021, 33(6):461-468.

12 Devlin J, Chang M, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. *Comput Language*, 2018, 10:4171-4186.

13 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all You need. Proceedings of the 31st International Conference on Neural Information Processing Systems. December 4-9, 2017, Long Beach, California, USA. New York: ACM, 2017:6000-6010.

14 Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw*, 2005, 18(5-6):602-610.

15 Lafferty J , Mccallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proc. 18th International Conf. on Machine Learning. 2001.

16 Brown P, Dellapietra V, Souza P, *et al*. Class-based n-gram models of natural language. *Comput Linguist*, 1992, 18(4):467-479.

17 李文林, 赵国平, 陆建峰, 等. 关联规则在名医临证经验分析挖掘中的应用. 南京中医药大学学报, 2008, 24(1):21-24.

18 于婧, 张宁, 杨涛, 等. 基于异质关联网络的辨证规律挖掘方法设计及应用. 世界科学技术-中医药现代化, 2020, 22(6):1955-1961.

19 Lample G, Ballesteros M, Subramanian S, *et al*. Neural Architectures for Named Entity Recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016:260-270.

20 Zhang W T, Jiang S H, Zhao S, *et al*. A BERT-BiLSTM-CRF model for Chinese electronic medical records named entity recognition. 2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA). October 26-27, 2019, Xiangtan, China. IEEE, 2020:166-169.

Research on the Construction Method and Application of Diagnosis and Treatment Knowledge Graph of Famous TCM Physicians Based on Medical Records

Gao Xiaoyuan[1], Gao Wenjia[1], Wang Xinyu[1], Zhu Yao[2], Ding Haiyan[3], Yang Tao[1], Hu Kongfa[1]

(1. School of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing 210023, China; 2. The First Clinical Medical College of Nanjing University of Chinese Medicine, Nanjing 210023, China; 3. Jiangsu Famous Medical Technology Co., LTD, Nanjing 210000, China)

Abstract: Objective A knowledge graph construction pipeline of traditional Chinese medicine (TCM) diagnosis and treatment was designed and applied, aiming at the automatic construction of the "disease-symptom-pathogenesis-and medicine" knowledge graph based on the medical records of famous TCM physicians, to analyze, organize and present medical records efficiently. Methods Firstly, The entity extraction method of medical records combining Deep Learning and Regular Expression was designed to extract disease, symptom, pathogenesis, and TCM entities from unstructured medical records automatically; secondly, entity relationships were defined and the correlations between entities were

calculated using HAN method, and then the "entity-relation-entity" triplets were built; the graph database Neo4j and Gephi were used for knowledge storage and visual display separately; Finally, the application was verified in the Medical records of lung cancer treated by the old famous TCM physicians. Results The precision, Recall and F_1 of the knowledge extraction model for medical records entities extraction are 88.49%, 90.02% and 89.25%, respectively, and each index is better than the comparison methods. A total of 1077 triples are extracted through entity correlation calculation, and the knowledge graph is successfully constructed. It can reflect the relationship between ′disease-symptom-pathogenesis-medicine′ in the treatment of lung cancer by the famous specialists of TCM. Conclusion The method in this paper can effectively solve the extraction, organization and expression of clinical medical records of famous TCM physicians, and realize the automatic construction process from the text of medical records to the knowledge graph. Relevant research ideas and methods proposed in this paper could provide a reference for the construction of the diagnosis and treatment knowledge graph of famous TCM physicians based on medical records.

Keywords: Chinese medical records, Famous TCM physician, Clinical diagnosis and treatment, Knowledge graph

（责任编辑：李青）