

基于命名实体识别与 Neo4j 的中文电子病历知识图谱构建和应用

许思特 孙 木

(上海交通大学医学院附属瑞金医院 上海 200020)

〔摘要〕 基于真实中文电子病历与网络爬取数据, 构建病历实体识别模型, 确定实体关系, 进行知识图谱可视化展现, 搭建基于规则匹配的问答系统。探索适用于中文电子病历的知识图谱与知识体系构建方法, 提高医院统计部门相关审核工作效率, 为人工智能技术在医疗卫生行业应用奠定基础。

〔关键词〕 中文电子病历; RoBERTa; 命名实体识别; 知识图谱; Neo4j

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2022.12.010

Construction and Application of Knowledge Graph in Chinese Electronic Medical Records Based on Name Entity Recognition and Neo4j XU Site, SUN Mu, Ruijin Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai 200020, China

〔Abstract〕 Based on actual Chinese electronic medical records and web crawling data, the paper constructs an entity recognition model of medical records, determines entity relationships, presents the visualization of knowledge graph, builds a question and answer system based on rule matching. It explores the construction method of knowledge graph and knowledge system applicable to Chinese electronic medical records, improves the work efficiency of hospital statistics department related audit, and lays a foundation for the application of artificial intelligence technology in the medical and health industry.

〔Keywords〕 Chinese electronic medical records; RoBERTa; name entity recognition; knowledge graph; Neo4j

1 引言

电子病历是指医疗活动过程中, 实现存储、管理、传输和重现的数字化医疗记录^[1]。对其填写质量进行审核纠正, 形成体系化知识图谱, 是医院统计信息部门日常工作之一。随着大数据时代来临, 电子病历质量日益重要, 知识关联脉络日益庞杂。对审核工作的要求、构建知识图谱的难度, 都有了

显著的提高。同时, 专业人才的稀缺, 使得引入技术手段的需求更为迫切。因此, 为辅助统计部门进行审核工作、构建可视化知识图谱, 尝试引入人工智能技术, 对中文电子病历中的有效信息进行识别, 并将实体内容进行可视化, 从而达到提高审核工作效率、构建专业知识脉络的目的。大量自然语言处理 (Natural Language Processing, NLP) 学者聚焦命名实体识别 (Name Entity Recognition, NER) 和知识图谱 (Knowledge Graph, KG), 尝试高效识别包括中文病历在内的各领域关键信息, 并构建相关专业的知识图谱。Liu K 等^[2]对比模型实体识别效率, 探索最佳的电子病历特征模板。Ouyang E^[3]

〔修回日期〕 2022-10-07

〔作者简介〕 许思特, 硕士, 发表论文 5 篇。

结合分词、词性标注、医学词汇等语料特征，搭建实体识别系统。2019 年，Liu Y^[4]等建立 RoBERTa，修改 BERT 中的关键超参数，使其更好地推广到下游任务中。KG 的搭建也逐步渗透至各个领域，如星河军事 KG、沃森健康（Watson Health）医疗 KG、海致星图金融 KG、海信交管云脑 KG 等^[5]。为满足当下工作要求，将 RoBERTa - BiLSTM - CRF 的模型构建方式，引入中文病案 NER 领域中，达到有效识别中文病历关键信息，提高病案审核工作效率的目的。同时爬取网络数据，采取 Neo4j 属性图模型为相关实体建立成体系的 KG，具象化地展现医药领域的知识脉络。最终，搭建问答系统，针对不同病历提出参考建议。

2 资料与方法

2.1 数据来源

数据来源由 3 部分组成：全国知识图谱与语义计算大会（China Conference on Knowledge Graph and Semantic Computing, CCKS）2019 评测数据、上海交通大学医学院附属瑞金医院真实病历、网络爬取数据。其中，CCKS 2019 病历数据 1 379 条，瑞金医院病历相关数据 21 条，网络爬取数据 8 808 条。CCKS 是由中国中文信息学会语言与知识计算专委会定期举办的全国年度学术会议，CCKS 2019 包含 6 个评测任务。其中，实体识别任务由清华大学知识工程实验室组织，所涉及中文电子病历数据由专业人员进行标注，并严格脱敏^[6]。

2.2 NER 数据预处理

基于经审核的病历数据，采用联合标签方法实现序列标注。该方法通过将标签集对象的分段标签与命名实体的标签集进行联合生成新的标签集，常见的分段标签集有 BIO 和 BIOES 两种，其中，BIO 将实体边界分为两类，B 代表实体中的首个词，I 代表实体中的非首个词，O 代表目标实体类别外。BIOES 进一步细分实体边界，增加实体结尾词 E 和单个实体

词 S 两种边界类型。具体构造方法，见表 1。BIOES 较 BIO 提供更多分段信息，识别度更高。在 BiLSTM 模型下，BIOES 较 BIO 效果更优。因此，字粒度序列标注模型均采用 BIOES 联合标签编码模式。

表 1 标注对比

标注方式	标注内容（肝胆脾胃，左膝关节未见异常）
BIO 标签	B - B B - BB - BB - B O B - B I - B I - BI - BO OOOO
BIOES 标签	S - BS - BS - BS - B O B - B I - B I - B E - B O OOOO

将 CCKS 2019 病历数据分为 6 种主要命名实体标签。爬取的数据以 json 格式存储，对后续识别的实体进行 KG 构建。按 6:2:2 的比例，将 CCKS 2019 数据集分为训练集、验证集、测试集。训练集用以训练算法，验证集用以调整参数，测试集用以评估最优模型。

2.3 NER 建模方法

KG 的关键技术在于运用图模型描述知识和对关系进行建模^[7]。在构建 KG 的过程中，需要从大量的数据中抽取知识并建立联系，纯文本数据是知识的主要来源。实现从纯文本数据中获取实体信息依赖于实体抽取技术，KG 高适用性和准确度的保障是高精度的 NER 技术^[8]。为了有效地进行实体识别，采取 RoBERTa - BiLSTM - CRF 的 NER 建模方式。BERT 使用堆叠的 Transformer 作为模型的主要架构^[9]，通过在大规模语料上的预训练，获得强有力的句子语义提取能力。RoBERTa 基于 BERT，两者在模型层面基本一致，但 RoBERTa 更为精细。具体变化如下：取消 NSP 任务；使用更大的 mini - batch、训练数据、更长时间的训练；将 BERT 的静态掩码（mask）策略替换成动态掩码。

长短时记忆模型（Long Short - Term Memory, LSTM）是一种特殊的循环神经网络^[10]。该模型引入储存单元、输入门、遗忘门和输出门的控制机制，解决长序列的梯度消失和梯度爆炸问题。BiLSTM 即双向 LSTM，由于 LSTM 只能单向编码，所以一般使用 BiLSTM 对隐藏层拼接后得到的向量作为融合上下文信息的词向量。条件随机场（Condition-

al Random Field, CRF) 用以对序列标注进行建模。与隐马尔科夫模型相比, CRF 引入特征函数, 信息获取更加全面, 能够获得更优的序列标注效果^[11]。

采用 RoBERTa - BiLSTM - CRF 的模型构建方式, 将处理好的文本序列输入到 RoBERTa 中, 输出每个位置并与原标签序列对齐, 继而输入到 BiLSTM 中进行处理。随后, 将处理结果输入到 CRF 中, 得到预测序列, 见图 1。

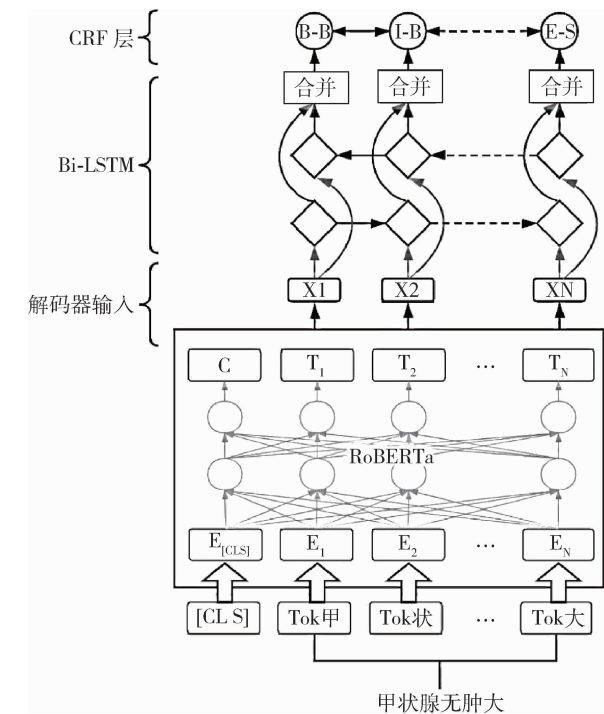


图 1 整体模型

2.4 NER 统计分析

NER 的相关深度学习算法基于 Tensorflow1.14 实现编写, 模型效果从准确率、召回率、F 值 3 方面进行展示。具体计算如下:

准确率 (P) =
$$\frac{\text{算法正确识别的词性总数}}{\text{人工识别的词性总数}} \times 100\%$$

召回率 (R) =
$$\frac{\text{算法正确识别的词性总数}}{\text{算法切分的词性总数}} \times 100\%$$

F 值 =
$$\frac{2 \times \text{识别准确率} \times \text{识别召回率}}{\text{识别准确率} + \text{识别召回率}}$$

2.5 知识图谱

KG 指使用语义检索, 从多种来源收集信息, 以提高搜索质量的知识库。本质上 KG 是将各种客

观实体、概念及其关系构成语义网络图, 以此形式化地描述真实世界中存在的关系。其主要内容为知识的数据结构, 包括实体、关系、属性等知识类的层次结构和层级关系定义, 约束数据层具体形式。随着信息时代数据量急剧增长, KG 的规模与日俱增, 对知识存储提出更高要求。现有研究中 KG 大多采用基于图的数据结构, 常见方式包括三元组图数据库、传统关系型数据库和图数据库。其中, 基于图数据库的存储是目前主流方式。其优点是以节点和边表示数据, 明确列出数据节点间依赖关系, 具有完善的图查询语言且支持各种图挖掘算法, 在深度关联查询速度上优于传统关系型数据库。常见图数据库包括 Neo4j、JanusGraph 和 HugeGraph 等^[5]。采用 Neo4j, 在 RoBERTa - BiLSTM - CRF 模型对中文电子病历进行关键实体抽取的基础上, 利用爬取的外延数据, 直观呈现实体的知识关系网络, 形成中文电子病历 KG。其中, 实体类型包括疾病诊断、操作检查、症状体征、药物、食物、科室类别等; 实体关系包括属于、常用药物、所需操作、常见症状、常见并发症、宜吃食物、忌吃食物等。

2.6 问答系统 (图 2)

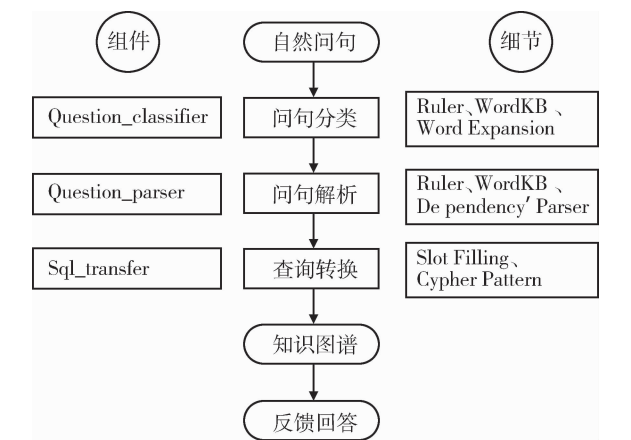


图 2 问答框架

由于中文电子病历涉及的医疗场景问答属于封闭域任务, 且病历数据相对规范, 对关系精准度有较高保证。因而关系抽取不采用远监督模型与联合抽取, 而是采取基于规则的关键词匹配, 对问题进

行穷举并分类。继而采用 Neo4j 的 Cypher 语法进行数据匹配查询，根据返回数据进行回答组装并输出。

3 实验结果

3.1 数据情况

实体识别模型基于 CCKS 2019 数据集进行搭

建。CCKS 2019 病历数据重点聚焦疾病诊断、解剖部位、影像检查、实验室检验、手术以及药物 6 种标注的命名实体，其中疾病诊断 2 798 个、解剖部位 1 933 个、影像检查 313 个、实验室检验 511 个、手术 905 个，药物 719 个。按 6: 2: 2 的比例，将数据集分为训练集、验证集、测试集。训练集用以训练算法，验证集用以调整参数，测试集用以评估最优模型，见表 2。

表 2 实体识别模型数据分布（个）

数据集	疾病诊断	解剖部位	影像检查	实验室检验	手术	药物
训练集	1 677	1 164	188	301	543	431
验证集	561	389	65	107	185	144
测试集	560	380	60	103	177	144
合计	2 798	1 933	313	511	905	719

KG 基于中文电子病历与网络爬取数据进行搭建，共获取 8 000 条 json 格式的扩展医疗数据。KG 的主要实体中，疾病诊断 8 000 个、操作检查 2 813 个、症状体征 5 123 个、药物 3 118 个、食物 4 057 个、科室类别 50 个，见表 3。

KG 的主要实体关系中，属于 8 015 个、常用药物 13 758 个、所需操作 36 852 个、常见症状 5 766 个、常见并发症 11 496 个、宜吃食物 38 156 个、忌吃食物 21 945 个，见表 4。KG 属性类型主要包括疾病名称、疾病简介、疾病病因、预防措施、治疗方

式、治疗周期、易感人群等。

表 3 实体情况

实体类型	实体数量（个）	样例
疾病诊断	8 000	窦性心动过缓
操作检查	2 813	冠状动脉造影
症状体征	5 123	胸闷
药物	3 118	盐酸曲美他嗪缓释片
食物	4 057	鸡蛋
科室类别	50	心内科

表 4 实体关系情况

关系类型	关系数量（个）	关联实体类型	样例
属于（belongs_ to）	8 015	疾病诊断 - 科室类别	窦性心动过缓，属于，心内科
常用药物（common_ drug）	13 758	疾病诊断 - 药物	窦性心动过缓，常用，盐酸曲美他嗪缓释片
所需操作（need_ check）	36 852	疾病诊断 - 操作检查	窦性心动过缓，所需操作，冠状动脉造影
常见症状（has_ symptom）	5 766	疾病诊断 - 症状体征	窦性心动过缓，常见症状，胸闷
常见并发症（acompany）	11 496	疾病诊断 - 并发症	窦性心动过缓，并发疾病，心肌炎
宜吃食物（do_ eat）	38 156	疾病诊断 - 食物	窦性心动过缓，宜吃，鸡蛋
忌吃食物（not_ eat）	21 945	疾病诊断 - 食物	窦性心动过缓，忌吃，浓茶

3.2 NER 模型效能

采用 CCKS 2019 数据集，进行 RoBERTa 实体识别模型实验。在 CCKS 2019 数据集中，RoBERTa - BiLSTM - CRF 模型效能如下：总体准确率

81.91%、召回率 83.03%、*F* 值 0.83；实验室检验中识别效能最佳，准确率 88.24%、召回率 86.35%、*F* 值 0.87；解剖部位中识别效能最差，准确率 78.22%、召回率 81.24%、*F* 值 0.79，见表 5。

表 5 模型效能

效能指标	总体	疾病诊断	解剖部位	影像检查	实验室检验	手术	药物
准确率 <i>P</i> (%)	81.91	80.65	78.22	86.27	88.24	84.64	86.34
召回率 <i>R</i> (%)	83.03	81.47	81.24	84.00	86.35	82.00	85.81
<i>F</i> 值	0.83	0.81	0.79	0.85	0.87	0.83	0.86

对比传统基于 Word2Vec（总体准确率 77.44%、召回率 84.90%、*F* 值 0.81）、BERT（总体准确率 79.18%、召回率 84.66%、*F* 值 0.82）的实体识别模型，RoBERTa - BiLSTM - CRF 拥有更优的效能。

3.3 Neo4j 知识图谱

基于 NER 模型识别的重点实体，采用 Neo4j 进行 KG 可视化呈现。以主疾病窦性心动过缓为例，从药品、食物、操作等角度搭建 KG 进行直观展现，见图 3。

KG 搭建完成后，可根据需要对重点关注的实体及其关系进行聚焦。以利巴韦林胶囊为例，对该药物可治疗的疾病、生产厂家等信息进行可视化呈现，见图 4。

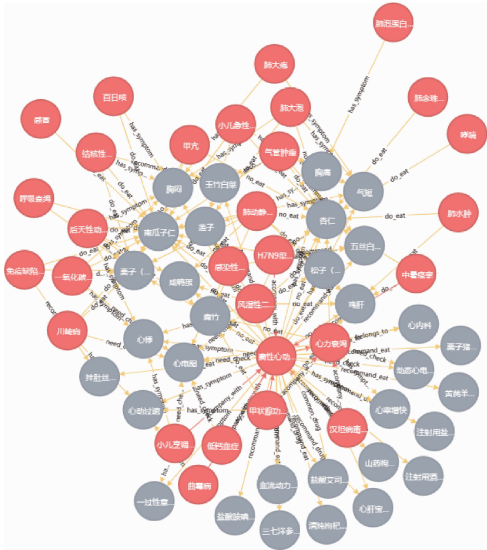


图 3 窦性心动过缓知识图谱

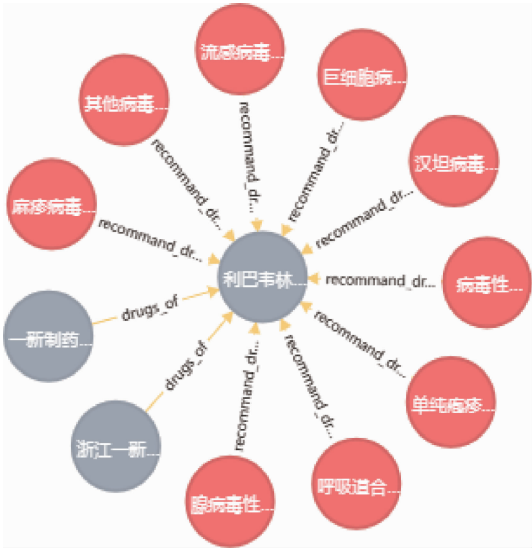


图 4 利巴韦林胶囊知识图谱

3.4 问答系统

基于 KG 数据库，进行问答系统搭建。以探究主疾病窦性心动过缓原因为例，进行展示。整个问答系统逻辑过程由以下几步完成：关键词匹配、关键词对问句分类、对问句解析、数据查询、反馈组装回答。以疾病原因为例，关键伪代码如下。

#关键词匹配

```
if 原因关键词 = [ '原因', '成因', '为什么', '怎么会',  
'怎样才', '咋样才', '怎样会', '如何会', '为啥', '为何',  
'如何才会', '怎么才会', '会导致', '会造成']
```

则关键词匹配→原因

#关键词对问句分类

```
if 原因关键词, 且问句包含疾病实体
```

则确认问句类型分类→疾病原因

#对问句解析

```
sql_ transfer ( 问句类型, 疾病实体)
```

#数据查询

```
"MATCH 疾病实体 return 疾病名称, 疾病原因" for i in entities
```

#反馈组装回答

```
answer = ' {0} 原因可能有: {1}'. format ( 疾病名称, ' ; '.  
join ( list ( set ( 疾病原因))))
```

以窦性心动过缓原因为例进行问答展示。
提问“为什么会窦性心动过缓”，得到回答
“窦性心动过缓原因可能有：迷走神经兴奋；窦房结功能受损；急性心肌梗死；自主神经张力改变”。

4 结论

4.1 不足与改进方向

基于 NER 与 Neo4j 的方式能有效构建中文电子病历 KG，但该方法仍有不足值得探究完善。针对实体识别，识别效能仍有进步空间，可考虑改进词典、引入 ALBERT 模型^[12]、融合字形^[13]等方式。由于 RoBERTa 等预训练模式最初为英文设计，而汉语作为符号语言，符号包含了一些额外的语义信息。因此原始预训练语言模型形式缺失了字形信息和拼音信息。而将字形和拼音信息融入到预训练语言模型中的方式，已在中文多个领域都达到了最优，后续将以此作为探索方向。针对关系抽取，主要基于规则模板的匹配方式，会导致信息缺乏覆盖率、规则冲突等问题，可考虑引入更为前沿的抽取方法，如远监督关系抽取、实体关系联合抽取^[14-16]等。针对 KG 仅涉及知识抽取阶段，后续知识融合、加工、推理仍有较大挖掘空间。另外对于

可视化的展现，后续研究将以知识超图作为切入点。

4.2 未来展望

现阶段国内病案相关信息化工作主要集中在软件开发、无纸化阶段。后续对于人工智能技术在病案审核、监管、分类中的应用应该保持开放态度，积极进行成果融合。NER、KG 作为 NLP 中的基础工作，在中文电子病历中应用效果良好。在未来，实体识别与实体链接联合任务、深度迁移学习、利用辅助资源进行基于深度学习的非正式文本分析等，都会是相关病案信息化发展中的有效助力。KG 能够提高医疗信息系统智能化水平，为医疗领域提供从海量医学文本和图像信息中抽取结构化知识的手段。基于医疗 KG，可以实现医疗知识问答、智能辅助诊疗、医疗质量控制及疾病风险评估等，具有广阔发展前景。总体而言，KG 将赋能认知智能，具有广泛且多样的应用需求，能够产生巨大的社会价值，对社会结构产生深远影响。

参考文献

- 1 原卫生部. 关于印发《电子病历基本规范（试行）》的通知 [EB/OL]. [2021-11-19]. http://www.gov.cn/zwqk/2010-03/04/content_1547432.htm.
- 2 Liu K, Hu Q, Liu J, et al. Named Entity Recognition in Chinese Electronic Medical Records Based on CRF [C]. Liuzhou: 2017 14th Web Information Systems and Applications Conference (WISA), 2017.
- 3 Ouyang E, Li Y, Jin L, et al. Exploring N-gram Character Presentation in Bidirectional RNN-CRF for Chinese Clinical Named Entity Recognition [C]. Chengdu: CCKS - Tasks 2017, 2017.
- 4 Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach [EB/OL]. [2021-11-19]. <https://arxiv.org/pdf/1907.11692.pdf>.
- 5 田玲, 张谨川, 张晋豪, 等. 知识图谱综述——表示、

- 构建、推理与知识超图理论 [J]. 计算机应用, 2021, 41 (8): 2161–2186.
- 6 中国中文信息学会. 全国知识图谱与语义计算大会 (CCKS 2019) 技术评测任务发布 [EB/OL]. [2021–11–19]. <http://www.cipsc.org.cn/chcontent.php?xuhao=2018031804>.
- 7 Xie X, Zou Y F, Zhang S W, et al. Research on the Construction of Knowledge Map of Chinese Film and Television [J]. Journal of the University of South China (Natural Sciences), 2019, 33 (1): 73–79.
- 8 Zhai S P, Guo L, Gao S, et al. Method for Knowledge Graph Completion Based on Bayesian Reasoning [J]. Journal of Chinese Computer Systems, 2018, 39 (5): 995–999.
- 9 Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [EB/OL]. [2021–11–19]. <https://arxiv.org/pdf/1810.04805.pdf>.
- 10 Hochreiter S, Schmidhuber J. Long Short-term Memory [J]. Neural Computation, 1997, 9 (8): 1735–1780.
- 11 Lafferty J, McCallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [EB/OL]. [2021–11–19]. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers.
- 12 Lan Z, Chen M, Goodman S, et al. ALBERT: a Lite BERT for Self-supervised Learning of Language Representations [EB/OL]. [2022–05–10]. <https://arxiv.org/pdf/1909.11942.pdf>.
- 13 Sun Z, Li X, Sun X, et al. ChineseBERT: Chinese Pre-training Enhanced by Glyph and Pinyin Information [EB/OL]. [2021–11–19]. <https://arxiv.org/pdf/2106.16038.pdf>.
- 14 刘千军, 桂前进, 张祺, 等. 实体关系抽取技术研究进展综述 [J]. 计算机应用研究, 2020, 37 (S2): 1–5.
- 15 杨穗珠, 刘艳霞, 张凯文, 等. 远程监督关系抽取综述 [J]. 计算机学报, 2021, 44 (8): 1636–1660.
- 16 郑德权, 胡佳. 远程监督关系抽取方法综述 [J]. 计算机应用, 2021, 41 (S1): 7–14.

(上接第 44 页)

- 6 李丹亚, 胡铁军, 李军莲, 等. 中文一体化医学语言系统的构建与应用 [J]. 情报杂志, 2011, 30 (2): 147–151.
- 7 储戟农, 彭莉, 崔蒙, 等. 论《中国中医药学主题词表》在中医药信息学中的作用 [J]. 国际中医中药杂志, 2014, 36 (12): 1060–1063.
- 8 阮学平. 关于医学分类表与主题词表再对应的思考 [J]. 医学情报工作, 2002 (3): 181–183.
- 9 张晶. 《中国中医药学主题词表》用户使用需求调查分析 [C]. 北京: 中国中医科学院中医药信息研究所 2011 年学术年会, 2012.
- 10 尹爱宁, 张汝恩. 建立中医药学一体化语言系统的方法学研究 [J]. 中国中医药信息杂志, 2003 (9): 90–92.
- 11 贾李蓉, 刘丽红. 基于中医药学语言系统的文献检索服务平台 [J]. 医学信息学杂志, 2012, 33 (1): 54–56.
- 12 杨阳, 李园白, 崔蒙. 建立中医临床术语集探索性研究 [J]. 中国中医药信息杂志, 2006 (12): 105–110.
- 13 于彤, 李敬华, 于琦. 中医药知识服务研究概述 [C]. 北京: 第一届中国中医药信息大会, 2014.
- 14 于彤, 杨硕, 李敬华. 中医药知识库系统的研究进展 [J]. 中国医学创新, 2014, 11 (18): 142–144.
- 15 龙凯, 李雄文. 早期肠内营养支持对重症脑卒中患者营养状况及并发症的影响 [J]. 实用心脑血管病杂志, 2014, 22 (9): 114–115.
- 16 甄建华, 黄光瑞. 溃疡性结肠炎中医病名、病因、病机的古今比较和回顾 [J]. 环球中医药, 2019, 12 (8): 1286–1289.
- 17 胡徽星, 梁发俊, 张静波, 等. 俞募配穴针刺联合益气化痰方治疗卒中后相关性肺炎 40 例 [J]. 环球中医药, 2021, 14 (9): 1705–1708.
- 18 王见义, 竺丽明, 陈建新, 等. 基于熵聚堆方法对肺炎中西医病名标准的对照研究 [J]. 中医杂志, 2010, 51 (S2): 165–166.