面向临床决策支持的肿瘤知识图谱构建与应用研究*

张 琴
（神州医疗科技股份有限公司 北京 100080）

林伟龙
（国家癌症中心/国家肿瘤临床医学研究中心/中国医学科学院 北京协和医学院肿瘤医院 北京 100021）

王军辉
（中国医学科学院/北京协和医学院医学信息研究所 北京 100020）

杨雅婷 马匡宇
（神州医疗科技股份有限公司 北京 100080）

郭 强
（国家癌症中心/国家肿瘤临床医学研究中心/中国医学科学院 北京协和医学院肿瘤医院 北京 100021）

洪 娜
（神州医疗科技股份有限公司 北京 100080）

〔**摘要**〕 基于肿瘤领域中文临床指南和相关文献构建肿瘤知识图谱，依据该肿瘤知识图谱实现智能问答应用，详细阐述该图谱构建方法及应用结果，指出该图谱可为肿瘤诊疗提供一定的临床决策支持，为其他领域知识图谱构建和应用提供参考。

〔**关键词**〕 肿瘤；知识图谱；问答；临床决策支持
〔**中图分类号**〕R－058 〔**文献标识码**〕A 〔**DOI**〕10.3969/j.issn.1673－6036.2021.12.008

Study on the Building and Application of Cancer Knowledge Graph for Clinical Decision Support

ZHANG Qin, Digital Health China Technologies Co. Ltd., Beijing 100080, China; LIN Weilong, National Cancer Center/ National Clinical Research Center for Cancer/ Cancer Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100021, China; WANG Junhui, Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China; YANG Yating, MA Kuangyu, Digital Health China Technologies Co. Ltd., Beijing 100080, China; GUO Qiang, National Cancer Center/ National Clinical Research Center for Cancer/ Cancer Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Bei-

〔**修回日期**〕 2021－09－24
〔**作者简介**〕 张琴，博士，高级工程师，发表论文 6 篇；通讯作者：洪娜，博士，医学数据研发总经理，发表论文 80 余篇。
〔**基金项目**〕 国家重点研发计划项目“基于人工智能的临床辅助决策支持新型服务模式解决方案”（项目编号：2018YFC0116900）。

jing 100021, China; HONG Na, Digital Health China Technologies Co. Ltd., Beijing 100080, China

〔**Abstract**〕 A cancer Knowledge Graph (KG) is built based on Chinese clinical guidelines and related literatures in the field of cancer, and the intelligent question answering application is realized according to the cancer KG. The paper elaborates the building method and application results of the KG in detail, and points out that the KG can provide certain clinical decision support for cancer diagnosis and treatment, and provide references for the building and application of KG in other fields.

〔**Keywords**〕 cancer; Knowledge Graph (KG); question answering; clinical decision support

1 引言

1.1 相关概念

1.1.1 知识图谱（Knowledge Graph, KG） 知识图谱是一种使用图数据结构模型集成数据的知识库，着重于构建实体之间的联系，在搜索引擎、问答系统和社交网络中有广泛应用。在医疗领域组成知识图谱的节点主要是“药品”或“疾病”等类型实体，知识图谱的边由“药品治疗疾病”等实体间的关系表示，实体和关系构成三元组形式的图结构数据，例如“钙（LV）200 mg/m2 静脉滴注 2h”中提取出 <静脉滴注 2h, LV, 用药方法 - 药品名称 > 信息。

1.1.2 实体识别 其方法主要分为 3 种：一是基于规则的方法。Quimbaya A P、Múnera A S 和 Rivera R A G 等[1]提出一种基于字典的电子健康记录命名实体识别方法，字典收录词汇越全，基于规则的系统性能越高，但是这些规则较难移植到其他领域。二是基于特征的监督学习模型方法。该方法依赖于算法和精心设计的特征，例如 Settles B[2]使用词汇特征和领域语义特征基于条件随机场（Conditional Random Fields, CRF）模型识别生物医学文本中的实体。三是基于深度学习模型方法。较常用的是 Huang Z、Xu W 和 Yu K[3]提出的 BiLSTM - CRF 模型，采用拼写特征、上下文特征、词嵌入和地名词典特征进行命名实体识别。基于深度学习模型方法，以端到端方式自动从输入中检测所需表示形式，本研究使用该方法抽取文本中的实体。

1.1.3 关系抽取 识别出文本中的实体后需要根据上下文对实体间语义关系进行抽取并分类。与实体识别方法相似，关系抽取方法主要分为基于规则的方法、基于特征的方法和基于深度学习模型的方法 3 类，其所采用的具体规则和学习模型不同。Califf M E 和 Mooney R J[4]提出抽取系统 RAPIER，使用单词、词性标注、填充词和文本中的语义类约束构成的模板诱导新规则。Nguyen D P T、Matsuo Y 和 Ishizuka M[5]提取两个实体子树特征，结合实体类型特征使用支持向量机（Supported Vector Machine, SVM）分类器识别关系类型。为避免语言分析模块引起的错误，Wu S 和 He Y [6]提出利用预训练语言模型 BERT 结合实体信息，从 BERT 模型输出得到实体嵌入，结合句子作为多层神经网络的输入进行分类。

1.2 相关研究

目前国内外已有医学知识图谱构建和应用相关研究。阮彤、孙程琳和王昊奋等[7]从国家标准、医院存储的临床知识库 3 个来源抽取数据，构建中医药知识图谱并进一步应用在基于模板的中医药知识问答和辅助开药两个场景中。奥德玛、杨云飞和穗志方等[8]利用规则引擎和命名实体识别（Named Entity Recognition, NER）技术从医学标准术语集、医学教材、临床指南和医学百科数据中提取三元组，构建由疾病、药物和诊疗技术类实体及其关系、属性组成的中文医学知识图谱 CMeKG。吴思竹、修晓蕾和王安然等[9]采集医学文献、医学健康网站和医学资源数据库等数据并从中抽取疾病、药物、检查检验、手术操作、医生、医院和企业 7 类信息，经过清洗、加工审核和数据对齐转换成 RDF 形式存储到 Openlink Virtuoso 数据库中。

1.3 本文研究目的及创新性

调研发现当前针对中文肿瘤知识领域构建知识

图谱并进行应用的研究尚不多见。构建肿瘤领域医学知识图谱，将大量知识进行整合，可支持多种肿瘤领域临床应用，具有重要价值和意义。构建智能问答系统为临床诊疗提供精准的决策支持是医疗信息化发展目标之一。目前各类机器学习、深度学习和大数据技术发展已进入实用阶段，配合知识图谱应用可在提升医疗信息系统智能化程度方面发挥重要作用。本研究主要针对肿瘤领域的知识图谱构建方法和应用进行探索和实践，基于肿瘤临床指南和相关文献，通过自然语言处理方法获取肿瘤相关医学实体和关系、构建肿瘤相关知识图谱并进一步依据该图谱完成智能问答模块构建，从而提供智能问答服务。本研究主要创新性是提出了一套利用肿瘤领域中文临床指南和相关文献构建肿瘤知识图谱的方法，具体包括数据预处理，利用命名实体识别和实体关系抽取（Relation Extraction，RE）两个自然语言处理技术完成实体和关系信息抽取，进行临床术语标准化和三元组存储，基于该肿瘤知识图谱实现智能问答应用，以期为其他领域知识图谱构建和应用提供参考。

2 数据与方法

2.1 总体流程

本研究基于肿瘤临床指南和权威文献，通过自然语言处理方法从文本数据中抽取实体，对包含两个实体的句子进行实体关系识别，得到（实体 1，实体 2，实体关系）类型三元组，再将三元组进行知识融合并写入图数据库，最终得到肿瘤相关知识图谱，见图 1。为控制模型输出，采用先进行命名实体识别、再抽取实体间关系这种管道式信息抽取方法，先用实体识别算法提取句子中的实体，再将实体组合分类到具体关系类别中进而得到实体间关系。

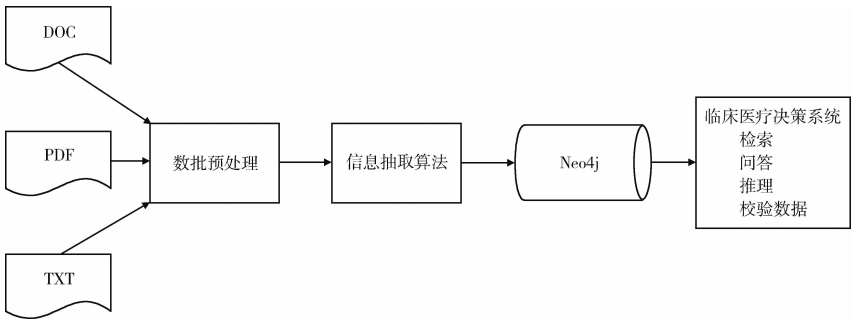

图 1 构建肿瘤知识图谱流程与方法

2.2 数据预处理

肿瘤临床指南数据是医生诊疗参考材料，由专业医学团队编写并定期更新，是专家经验梳理总结以及疾病治疗依据，可作为研究证据。此外相关学术文献是临床试验结果的重要医学证据来源。本研究使用肿瘤领域的中文临床指南和相关文献，由于其数据以 TXT、PDF 或 DOC 文档多种形式存储，利用文本预处理模块从中提取非结构化文本数据，然后经过人工校对，通过标准化接口输出。

2.3 实体识别方法

在医学自然语言处理中，命名实体识别是指抽取出文本中的疾病、药品名称、用药频率等专有名词。BiLSTM－CRF 模型结构可以有效地保留文本序列中的关键信息，采用该模型完成肿瘤文本数据中的命名实体识别工作。将标注结果转化成“BIO”形式（分别代表 begin、inside 和 other）标签，将文本通过 BiLSTM－CRF 模型得到识别标签。具体而言，使用 word2vec 语言模型[10]将文本中的字符映射成向量，经过 BiLSTM 网络的输入门和遗忘门抽取特征，每个序列元素都可以输出 1 个隐藏向量，将正反两个方向的隐藏向量拼接起

来得到包含完整信息的向量，作为序列元素向量输入到 CRF 模型中进行分类，最终得到元素标签类别。

2.4 关系识别方法

2.4.1 概述 实体识别模型的输出作为关系抽取模型的输入，将关系识别任务看作一个分类问题，采用分类算法对包含两个实体的句子进行关系识别，得到两个实体间的语义关系类型。本研究采用双向循环神经网络和多级别注意力机制实现关系分类任务，根据文本特征判断实体对的语义关系类型，主要包括特征构建、模型结构和训练 3 部分。

2.4.2 特征构建 使用预训练的字向量和文本位置特征对文本中字符进行编码并输入模型，其中预训练向量由 word2vec 模型在医学数据上训练得到，位置特征为该字符和实体的相对位置。

2.4.3 模型结构 网络结构由双向门控循环单元（Gate Recurrent Unit，GRU），字符级别注意力机制和句子级别注意力机制 3 部分组成。第 1 个 GRU 网络从前向后对句子进行建模，可以利用前半句信息，第 2 个 GRU 网络从后向前对句子进行建模，可以利用后半句信息，将生成的两个向量进行拼接得到完整向量表示。注意力机制通过赋予元素权重值，然后权重值乘以该元素的编码向量实现，采用双曲正切函数 tanh 作为激活函数。字符级别注意力机制可以保留更多关键信息，去除非关键信息，句子级别注意力机制可以避免错误标注造成模型产生偏差。

2.4.4 训练 使用交叉熵作为损失函数训练模型，得到每个句子的交叉熵损失后，将一组实例中所有句子的损失相加得到该组实例的损失值，通过梯度下降法求解最大似然函数，得到最终的实体间关系语义类型。

2.5 临床术语标准化、存储与查询

不同来源的文本抽取出的实体存在表达不一致情况。因此需要将抽取出的知识进行融合，涉及术语标准化工作，旨在将识别出来的实体链接到指定知识库，从而达到对齐目的。自然语言处理模块识别到实体和实体关系，经过实体和关系校准等人工校验后，进一步通过临床术语标准化和规范化，采用的标准化系统有：医学系统命名法 - 临床术语（Systematized Nomenclature of Medicine Clinical Terms，SNOMED CT），国际疾病分类中文版（International Classification of Diseases - 10th Edition - Chinese，ICD - 10 - CN），国际疾病分类肿瘤学专辑第 3 版（ICD - O - 3）和中国临床药物编码（Normalized Chinese Clinical Drug，NCCD）等国际及国内标准医学术语体系。通过知识融合合并重复实体节点和节点关系信息，以属性方式添加到对应实体和关系三元组中以构建语义一致的三元组，将这些三元组保存到 Neo4j 图数据库中形成知识图谱一致化语义表示。

2.6 基于肿瘤知识图谱开发智能问答应用

知识图谱作为结构化数据来源可以为问答系统提供更加高质量的知识。基于知识图谱的问答系统包含两部分内容：问题语义解析和答案检索模块。问题解析功能包括问题实体识别、意图识别和构造查询语句；答案获取模块包含结果检索、组织答案返回给用户，见图 2。

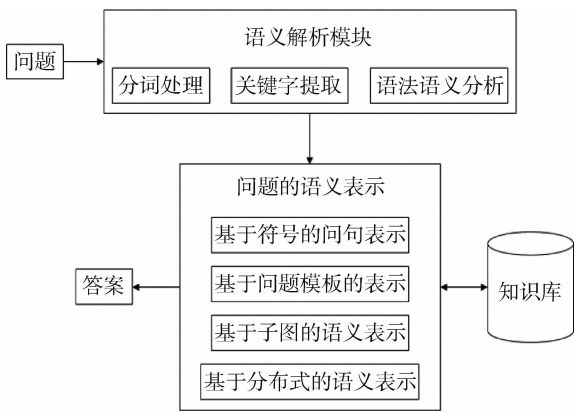

图 2 基于知识图谱的问答技术框架

3 结果

3.1 实体和关系识别性能

本研究共采集 209 篇肿瘤领域临床指南和相关文献，经过数据预处理提取出文本内容。临床医生

制定标注指南指导数据标注工作，完成 100 篇文本标注。将这些标注数据按照 6∶2∶2 的比例分为训练集、验证集和测试集，完成相关实体和实体间关系抽取，实体识别算法结果，见表 1；关系抽取算法结果，见表 2。命名实体识别方面，与使用专门词嵌入研究[11]结果类似，本研究结果比使用随机向量的模型实体识别结果好，比使用预训练词向量的模型和加入词汇特征的模型性能差，说明加入医学词汇信息较为重要。对比相关研究结果[12]，分析本研究采用的基于注意力机制的深度残差网络模型在中文医学数据上的性能。将本研究实体抽取结果与其他研究中支持向量机、卷积神经网络、长短期记忆人工神经网络（Long - Short Term Memory，LSTM）和加入注意力的 BiLSTM 网络等模型结果进行比较可知，本研究结果优于上述模型。

表 1 实体识别算法性能指标（%）

实体类别	精准率	召回率	*F*1 - score
疾病	93.98	81.53	87.31
身体部位	84.20	91.67	88.02
药品名称	78.90	83.18	80.98
手术	78.90	83.18	80.89
临床表现	80.00	83.87	81.89
检查方法	98.46	81.82	89.37
非药治疗	86.96	76.92	81.63
平均	85.91	81.17	84.30

表 2 关系抽取算法性能指标（%）

关系类别	精准率	召回率	*F*1 - score
身体部位→疾病	76.81	87.85	81.96
药品名称→疾病	84.93	72.17	78.03
手术→疾病	82.48	80.71	81.59
临床表现→疾病	83.77	86.22	84.98
检查方法→疾病	77.20	76.18	76.69
非药治疗→疾病	79.21	80.46	79.83
临床表现→药品名称	75.00	63.64	70.71
临床表现→手术	71.67	79.63	75.44
临床表现→身体部位	76.81	87.85	81.96
临床表现→非药治疗	79.55	73.26	74.12
平均	78.33	78.58	78.19

3.2 实体和关系抽取结果

利用实体识别模型和实体关系抽取模型对剩余文本进行初步识别并人工修正识别结果，识别出最终实体和语义关系。最终共抽取出 9 442 个实体，以及 6 589 个实体间语义关系，所有实体都作为知识图谱中的 1 个节点，关系作为节点间的边。具体实体类别和数量，见表 3；实体关系类型和数量，见表 4。针对临床术语标准化结果的测评发现，频次在 100 以上的高频术语在 SNOMED CT、ICD - O - 3、ICD - 10 - CN 和 NCCD 术语体系中的自动映射率达到 61.2%。术语标准化程度有待提高，术语表覆盖程度、映射的自动推荐方法均对结果有一定影响。本地化术语表扩展和人工校对将逐步丰富至实体的三元组属性中。将提取的关系三元组导入图数据库完成知识图谱建立，选择 Neo4j 图数据库存储和查询经术语标准化、实体校准、关系校准等人工校验后的三元组。基于知识图谱实现校验、检索和推理模块在临床系统应用中的相应功能。校验数据模块用于校验知识图谱中数据是否正确，经过校验的数据支持定期更新知识图谱，同时同步到 Neo4j 中完成数据更新。此外可支持通过校验数据模块删除和增加数据。

表 3 实体抽取结果

类别实体	数量（个）
身体部位	1 442
疾病	100
药品名称	275
手术	1 364
临床表现	2 779
检查方法	3 384
非药治疗	98

表 4 关系抽取结果

实体关系	数量（个）
身体部位→疾病	639
药品名称→疾病	292
手术→疾病	20
临床表现→疾病	1 450
检查方法→疾病	2 273
非药治疗→疾病	142
临床表现→药品名称	317
临床表现→手术	275
临床表现→身体部位	1 027
临床表现→非药治疗	154

3.3 基于肿瘤知识图谱的智能问答应用

本研究尝试将所构建的肿瘤知识图谱应用于智能问答场景中，致力于人工智能辅助回答医护端或患者端提问。根据用户输入的问题和问句意图类型，获知其意图后，根据识别出的实体和问句意图构造查询语句，在肿瘤相关知识图谱中查询对应实体或属性，将查询结果生成符合对话逻辑且语法通顺的答案返回给用户。问句意图类型包括：询问疾病发生的部位、询问用药方式、询问病因、询问可能的疾病、询问疾病的临床表现、询问药品的用药剂量、询问身体部位的临床表现、询问疾病的手术治疗、询问肿瘤的检查方法、询问疾病的非药治疗、询问肿瘤的治疗药品、询问手术后的临床表现、询问药物治疗的副作用、询问非药治疗后的临床表现等。

4 结语

本研究以肿瘤为应用领域，在大规模医疗知识和文献基础上构建中文知识图谱并基于该知识图谱开发智能问答应用，探索基于知识图谱辅助临床决策问答的技术实现可行性，验证其潜在应用价值。研究重点为知识获取，即从已有资源和数据中抽取有意义的实体关系三元组的构建实践。未来研究将从两个方向进行拓展：一方面，对知识获取后的标准化、融合、存储及推理过程进行深入研究，从而为智能问答提供更丰富、完善的知识支撑；另一方面，针对随医学知识不断更新而不断增加的实体和关系，后续将对相关医学知识进行人工迭代评审和质量保证，以支持中文肿瘤知识图谱内容更新，最终构建准确、完善和更加智能化的临床决策支持系统。

参考文献

1 Quimbaya A P, Múnera A S, Rivera R A G, et al. Named Entity Recognition Over Electronic Health Records Through a Combined Dictionary - based Approach [J]. Procedia Computer Science, 2016 (100): 55 - 61.

2 Settles B. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets [C]. Geneva: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (NLPBA/BioNLP), 2004: 107 - 110.

3 Huang Z, Xu W, Yu K. Bidirectional LSTM - CRF Models for Sequence Tagging [EB/OL]. [2021 - 05 - 10]. https://arxiv.org/abs/1508.01991.

4 Califf M E, Mooney R J. Relational Learning of Pattern - match Rules for Information Extraction [C]. Orlando Florida: Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence, 1999: 328 - 334.

5 Nguyen D P T, Matsuo Y, Ishizuka M. Relation Extraction from Wikipedia Using Subtree Mining [C]. Vancouver: Proceedings of the 22nd National Conference on Artificial Intelligence, 2007: 1414 - 1420.

6 Wu S, He Y. Enriching Pre - trained Language Model with Entity Information for Relation Classification [C]. Beijing: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019: 2361 - 2364.

7 阮彤，孙程琳，王昊奋，等. 中医药知识图谱构建与应用 [J]. 医学信息学杂志, 2016, 37 (4): 8 - 13.

8 奥德玛，杨云飞，穗志方，等. 中文医学知识图谱 CMeKG 构建初探 [J]. 中文信息学报, 2019, 33 (10): 1 - 7.

9 吴思竹，修晓蕾，王安然，等. 中文疾病知识图谱构建初探 [J]. 医学信息学杂志, 2019, 40 (7): 14.

10 Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [EB/OL]. [2021 - 05 - 10]. https://arxiv.org/abs/1301.3781.

11 Li Z, Zhang Q, Liu Y, et al. Recurrent Neural Networks with Specialized Word Embedding for Chinese Clinical Named Entity Recognition [C]. Chengdu: CEUR Workshop Proceedings, 2017: 55 - 60.

12 Zhang Z, Zhou T, Zhang Y, et al. Attention - based Deep Residual Learning Network for Entity Relation Extraction in Chinese EMRs [J]. BMC Medical Informatics and Decision Making, 2019, 19 (2): 171 - 177.