软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(3):1240−1253 [doi: 10.13328/j.cnki.jos.007173] [CSTR: 32375.14.jos.007173]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

融合大规模医学事实的跨语言双层知识图谱*

王楚童[1,2], 李明达[1], 孙孟轩[1,2], 王 静[3], 杨雪冰[1], 牛景昊[1], 贺志阳[3], 张文生[1]

[1](多模态人工智能系统全国重点实验室 (中国科学院 自动化研究所), 北京 100190)

[2](中国科学院大学 人工智能学院, 北京 100049)

[3](讯飞医疗科技股份有限公司, 安徽 合肥 230088)

通信作者: 张文生, E-mail: zhangwenshengia@hotmail.com

摘 要: 得益于信息化技术的快速发展和医疗信息系统的普及, 医学数据库中积淀了海量的医学事实, 如患者临床诊疗事件以及医学专家共识等. 如何从医学事实中提炼出知识, 进而对其管理和合理利用, 是推进诊疗自动化和智能化的关键. 知识图谱作为一种新型的知识表示工具, 能够有效地挖掘和组织大规模医学事实中的信息, 受到医疗领域从业人员的广泛关注. 然而, 现有医疗知识图谱普遍存在规模小、限制多、可扩展性差等问题, 面向医学事实的知识表达能力有限. 为此, 提出一种双层医疗知识图谱架构, 通过对英文患者诊疗事件和中文专家共识进行信息抽取, 构建得到一个跨语言、多模态、动态更新、可拓展性强的 10 亿级医疗知识图谱, 可提供更加精准的智能医疗服务.

关键词: 医学事实; 医疗知识图谱; 双层知识表示; 信息抽取; 知识融合

中图法分类号: TP181

中文引用格式: 王楚童, 李明达, 孙孟轩, 王静, 杨雪冰, 牛景昊, 贺志阳, 张文生. 融合大规模医学事实的跨语言双层知识图谱. 软件学报, 2025, 36(3): 1240–1253. http://www.jos.org.cn/1000-9825/7173.htm

英文引用格式: Wang CT, Li MD, Sun MX, Wang J, Yang XB, Niu JH, He ZY, Zhang WS. Cross-language Bilayer Knowledge Graph with Large-scale Medical Facts. Ruan Jian Xue Bao/Journal of Software, 2025, 36(3): 1240–1253 (in Chinese). http://www.jos.org.cn/1000-9825/7173.htm

Cross-language Bilayer Knowledge Graph with Large-scale Medical Facts

WANG Chu-Tong[1,2], LI Ming-Da[1], SUN Meng-Xuan[1,2], WANG Jing[3], YANG Xue-Bing[1], NIU Jing-Hao[1], HE Zhi-Yang[3], ZHANG Wen-Sheng[1]

[1](State Key Laboratory of Multimodal Artificial Intelligence Systems (Institute of Automation, Chinese Academy of Sciences), Beijing 100190, China)

[2](School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China)

[3](Xunfei Healthcare Technology Co. Ltd., Hefei 230088, China)

Abstract: Benefiting from the rapid development of information technology and the widespread adoption of medical information systems, a vast amount of medical knowledge has been accumulated in medical databases, including patient clinical treatment events and medical expert consensus. It is crucial to extract knowledge from these medical facts and effectively manage and utilize them, which can advance the automation and intelligence of diagnosis and treatment. Knowledge graphs, as a novel knowledge representation tool, can effectively mine and organize information from abundant medical facts and have received extensive attention in the medical field. However, existing medical knowledge graphs often suffer from limitations such as small scale, numerous restrictions, poor scalability, and so on, leading to a limited ability to express knowledge from medical facts. To address these issues, this proposes a bilayer medical knowledge graph architecture and employs information extraction techniques on both English patient clinical treatment events and Chinese medical expert

* 基金项目: 科技创新 2030—“新一代人工智能”重大项目 (2018AAA0102100); 国家自然科学基金 (62203437)
收稿时间: 2023-09-21; 修改时间: 2023-11-25; 采用时间: 2024-02-28; jos 在线出版时间: 2024-06-14
CNKI 网络首发时间: 2024-06-19

consensus to construct a billion-scale medical knowledge graph that is cross-lingual, multimodal, dynamically updated, and highly scalable, aiming to provide more accurate, intelligent medical services.

Key words: medical fact; medical knowledge graph; bilayer knowledge representation; information extraction; knowledge fusion

医学事实是医学事件的真实情形. 它可以指在过去已发生的事件, 亦可指被验证成立且中立的陈述, 在医学中即共识、指南等蕴含医学知识与临床经验的概念. 伴随着现代医学设备、健康医疗场景、医疗信息系统的蓬勃发展, 医疗大数据在产生、传输和管理的全流程快速升级, 医学数据库中积淀了海量的医学事实数据[1]. 作为涉及国计民生的重要基础性战略资源, 医学事实中蕴含着大量的诊疗经验和临床规律, 如何从中挖掘不同医疗实体间的关联并提炼出知识, 进而对其管理和合理运用, 是实现智能化的医学知识检索、医疗辅助诊断以及医学档案管理的关键.

作为大数据时代的一种有效的知识表示与存储工具, 知识图谱能够有效地挖掘、组织和管理大规模数据中的知识, 提高知识信息服务质量, 从而也可以在医学领域为医生和病人提供更智能化的服务. 医疗领域中知识图谱被广泛用于对海量医学事实的存储组织[2], 其中, 知识图谱中医学事实表现为 (实体, 关系, 实体) 的三元组. 近 10 年来, 医疗知识图谱不断涌现, 国外如芝加哥大学的 LynxKB[3]、德克萨斯大学的 BMKG[4]、伍斯特理工学院的 BKG[5], 国内如中国中医科学院的中医药知识图谱[6]、哈尔滨工业大学的 EMKN[7]、北京大学联合鹏城实验室的 CMeKG[8]、北京交通大学的 R-MKG[9]、阿里的 DiaKG[10]. 整体而言, 知识来源从多源医疗知识库、电子病历到共识文献, 构建方式从静态手工标注到动态自动抽取, 规模从万级到亿级.

虽然医疗知识图谱的应用场景愈发广泛, 但真正落地、符合临床使用的知识图谱产品仍然短缺. 考虑到医疗领域信息维度多, 当前知识图谱受限于存储的知识类型和表达能力, 难以满足真实场景的临床诊疗需求. 具体而言, 当前阶段医生在临床诊疗过程中通常会同时考虑目标患者的相似病例和医学百科指南, 而现有医疗知识图谱的结构单一, 没有体现多种知识图之间的逻辑关系, 难以从多个角度辅助医疗决策; 现有图谱大多是由单语言描述的纯文本表示, 因此无法回答需要跨语言的医学知识查询; 由于缺乏医疗影像等不同于文本模态的数据, 导致难以应对多模态推理的需求, 限制了该类知识图谱在多模态医疗场景中的使用价值; 现有医疗领域的知识图谱主要用于存储静态医学知识, 缺乏面向患者动态诊疗事件的组织和表达能力, 难以有效地刻画医学事实的演化规律.

为了更加有效地表示复杂的医疗领域信息, 本文提出了一种新型面向大规模医学事实的双层知识图谱框架, 以医疗专家共识为上层宏观知识, 以患者诊疗事件为下层微观实例, 建立“个体诊疗指标-专家共识经验”的知识表示框架. 在此基础上, 对现有医疗数据进行自动化的信息抽取, 刻画不同语言、不同模态的医疗实体在不同时刻上的关联关系, 构建得到一个跨语言、多模态、动态更新的 10 亿级双层医疗知识图谱. 具体而言, 本文首先基于 UKB 数据库 (https://www.ukbiobank.ac.uk/) 的患者病例数据提取得到实体集和关系集, 并通过医学事实解码和多模态数据预处理构建得到包含约 50 万患者信息的英文病例图谱; 其次, 本文基于中华医学会 (https://www.cma.org.cn/) 的医学指南进行实体和关系抽取, 得到了包含疾病医学共识的中文百科图谱; 最后, 本文提出跨语言相似匹配方法, 将中英文图谱中的疾病实体进行对齐, 构建知识图谱的跨层连接, 进而得到一个跨语言、多模态、动态更新的双层医疗知识图谱, 实现了多层次、多模态、多尺度医疗知识的有机融合. 本文提出的双层医疗知识图谱将患者诊疗数据和专家共识经验以合理的逻辑关系结合起来, 有效地弥补了医疗知识图谱表达能力和医疗诊断需要之间的鸿沟, 可应用于药物推测、相似患者检索等多个方面, 为智慧医疗的实现提供了坚实的基础, 其主要贡献如下.

(1) 在理论层面, 提出了由“个体诊疗指标-专家共识经验”组成的双层知识图谱架构: 以疾病为中心的百科图谱, 用以关联专家知识和术语概念; 以患者为中心的病例图谱, 用以关联患者跨模态动态医学事实.

(2) 在技术层面, 基于双层知识表示架构, 提出了一种新型跨语言多模态的医学事实三元组自动化抽取技术, 可以有效地从结构化的英文 UKB 患者数据以及非结构化的中文医学指南进行信息抽取及跨语言关联.

(3) 所构建的双层知识图谱实现了英文病例图谱和中文百科图谱的跨语言连接, 建立了包含 17 261 282 个实体、805 065 729 个三元组的 10 亿级动态医疗知识图谱, 填补了国内外医疗信息结构化整合的空白.

本文第 1 节介绍医疗知识图谱构建的国内外研究现状, 然后对实体抽取、关系抽取和知识融合的技术进行介

绍. 第 2 节介绍知识图谱双层架构和跨语言知识图谱的具体构建流程和方法. 第 3 节展示实验结果并进行分析. 第 4 节对所构建的知识图谱进行可视化展示. 第 5 节是结论和未来的工作.

1 相关工作

1.1 医疗知识图谱

近年来, 医疗知识图谱蓬勃发展、方兴未艾. 得益于互联网时代的到来和信息化技术的快速发展, 国外在早期对医学领域的知识结构和组织形式投入了大量的研究, 并建立了一系列较为成熟的医疗知识库, 如: 描述基因和蛋白质供能的 Gene Ontology、描述医疗术语的 SNOMED-CT、描述疾病概念的 Disease Ontology 等[11]. 随着近年来医学信息系统的普及, 医学领域积淀了海量的医学知识、临床诊断和医疗影像数据, 国内外一些研究机构和知名公司逐步利用新兴的知识图谱技术来取代传统的知识库, 实现对海量医学数据的存储组织. 2014 年, 芝加哥大学的 Sulakhe 等人[3]将过往搜集的基因、蛋白质以及疾病相关的知识库进行整合, 构建 LynxKB 知识图谱, 为医疗用药和决策提供了辅助支持. 2018 年, 德克萨斯大学健康科学中心的 Cong 等人[4]将多源医学知识库 SemMedDB 和 Linked Open Data 的概念统一, 制定了标准 UMLS 术语, 构建得到融合多源医学知识的知识图谱 BMKG. 2022 年, 伍斯特理工学院的 Huo 等人[5]通过对近 10 年发表在 PubMed 上的文章进行信息抽取, 构建得到医疗百科知识图谱 BKG, 该图谱有效地展示了前沿医疗知识的发展和演变. 此外, 国外相关公司也在积极探索将医疗知识图谱集成至业务系统中, 进而实现医疗信息的有效存储, 提供高质量的医疗信息服务. 2015 年, Google 推出了医疗版“知识图谱”用于医疗搜索, 随后, IBM 在同一年以肿瘤和癌症领域的知识图谱作为数据源, 推出了 Watson Health 医疗决策系统.

国内在医疗知识图谱领域的研究起步较晚, 目前处于积极追赶阶段. 2015 年, 中国中医科学院中医药信息研究所基于已有的中医药学语言系统构建了中医药知识图谱, 其中包含约 100 万条中医药语义关系[6]. 2017 年, 哈尔滨工业大学的 Zhao 等人[7]基于海量电子病历构建得到医学知识图谱 EMKN, 并为基于症状的医疗诊断提供知识支持. 2019 年, 北京大学联合鹏城实验室的奥德玛等人[8]以人机结合的方式研发了中文医学知识图谱 CMeKG, 其中覆盖了疾病、药物和诊疗技术等各类医学知识. 2020 年, 北京交通大学的 Li 等人[9]利用结构化的电子病历数据, 从 1 600 万份诊疗数据中抽取出医学知识关系, 构建了包含超过 22 000 个医学概念节点的知识图谱 R-MKG. 在工业界, 北京妙医佳健康科技集团于 2021 年携手阿里巴巴集团、清华大学推出糖尿病知识图谱 DiaKG, 对公开发表的 41 篇糖尿病指南和共识进行实体和关系抽取, 涵盖了近年来最广泛的糖尿病领域研究热点和前沿进展[10]. 与上述知识图谱不同, 本文面向医学事实构建的大规模动态医疗知识图谱, 旨在为高质量、结构化的医疗知识整合提供基础性支撑.

1.2 知识图谱构建技术

知识图谱主要分为自底向上和自顶向下两种构建方式. 自底向上是从已有知识库中使用信息抽取和数据筛选手段, 采集置信度高的新事实加入知识库中[12]. 自顶向下是指先构建模式层, 确定知识图谱的本体结构, 根据模式层的规范将三元组存储构建数据层[13], 本文使用自底向上的构建方式, 主要涉及医疗实体抽取、医疗关系发现和医疗知识融合这 3 个方面.

医疗实体抽取也叫医疗命名实体识别, 常用技术可以分为基于字典、基于机器学习和基于深度学习的方法. 基于字典的方法是将相应数据库中的实体作为字典, 在此基础上对文本进行匹配, 如 Wu 等人[14]利用 CHV 和 SNOMED-CT 两个医学词典对医疗诊所笔记中的医学信息进行命名实体识别. 基于机器学习的方法通常将命名实体识别作为一个序列化标注任务, 需要标注一定规模的训练语料, 如 Wang 等人[15]采用 CRF 和 SVM 的级联模型, 精准识别出临床记录中的 10 种命名实体, Zhou 等人[16]考虑了医疗术语的特征, 利用多种词性特征训练 HMM 模型实现对医疗实体的识别. 基于深度学习的实体抽取算法近年来被广泛应用于命名实体识别, 不需要过多的特征工程且性能优越, 目前已经成为该领域的主流算法, 如 Chowdhury 等人[17]构建了 BiRNN 模型识别中文电子病历的医疗命名实体; Ji 等人[18]采用 Attention-BiLSTM-CRF 模型进行实体抽取, 并提出了一种自动校正的算法, 得到

了较高的 $F1$ 值; Tan 等人[19]提出了边界感知的神经网络模型, 并采用联合训练的方式, 解决了医学名词实体较长、识别边界困难的问题.

医学关系抽取目的是解决医疗实体间的语义关联问题, 与实体抽取类似, 早期关系抽取主要采用人工构造词典和基于机器学习的方法, 如郭宇捷等人[20]使用关系表达模板匹配和关系提示词、事件触发词规则约束相结合的方法抽取医疗事件因果关系, Abacha 等人[21]使用人工构建模板和 SVM 混合模型, 得到更为精确的分类效果. 近年来, 深度学习成为关系抽取的主流, 如 Quan 等人[22]提出了一个多通道卷积神经网络 (MCNN) 进行医学关系抽取; Zeng 等人[23]利用了使用序列到序列 (Seq2Seq) 的方法设计实体和关系提取的多任务学习复制模型, 解决了医学实体间普遍存在重叠关系的问题; Zhou 等人[24]提出一个基于 BERT 的关系抽取模型, 采用自适应阈值和局部上下文池化的方法来抽取文档级生物医学实体关系.

医学知识融合是以实体对齐为主要任务进行的不同数据源的医学知识的整合, 沈伟豪等人[25]采用基于词向量的语义相似的方法, 将不同数据源的实体文本使用 Word2Vec 转换成词向量, 利用词向量之间的余弦相似度计算判断是否对齐, E 等人[26]尝试将关系和属性三元组结合起来进行实体对齐, 采用参数共享联合方法和基于翻译的知识嵌入方法将它们联合嵌入. Dieng-Kuntz 等人[27]将不同医学数据库转换为医学本体, 在人工控制下对不同本体进行补全和融合, 通过知识的本体结构来实现两个医学知识库的融合. 翟霄等人[28]提出了一种多模态结构图模型, 利用图结构对多模态医学数据进行建模和表示. Lacoste-Julien 等人[29]设计了一种贪婪迭代算法, 利用实体匹配以及实体之间的关系图信息实现了实体对齐.

2 技术方法

2.1 双层知识图谱表示框架

本文提出了双层医疗知识图谱表示架构, 在逻辑上可分为患者层与疾病百科层两个层次, 其中底层是患者层 L_p, 该层以患者为中心, 主要由患者和患者关联的一系列诊疗事件组成, 诊疗事件包括疾病诊断和其他医学事实, 可形式化为 $L_p=\{h_p,r_p,t_p\}$, 其中 h_p 表示患者实体, t_p 表示诊疗结果, r_p 表示诊疗关系; 上层是疾病百科层 L_d, 该层以疾病为中心, 主要由疾病和百科知识组成, 这些百科知识包括症状、检查和治疗等疾病信息, 可形式化为 $L_d=\{h_d,r_d,t_d\}$, 其中 h_d 表示疾病实体, t_d 表示疾病百科知识.

双层知识图谱表示架构通过跨层连接的方式实现知识图谱的层间融合. 跨层连接的构建首先分别提取 L_p 和 L_d 中的实体类型集合 C_p 和 C_d, 查找两个集合中共现或存在相似语义的实体类型集合 $C=C_p\cap C_d$. 然后, 对于每一个实体类型 $c_i\in C$, 分别获取 L_p 和 L_d 中相应的实体集合 E_p^i 和 E_d^i, 通过制定相似度匹配规则识别出共指相同本体信息的实体对 $(e_p^{i,j},e_d^{i,j})$, 得到相应的三元组集合 $T_i=\{(e_p^{i,j},aligns.e_d^{i,j})\}$, 其中 $e_p^{i,j}\in E_p^i,e_d^{i,j}\in E_d^i$. 最终, 可以得到跨层连接三元组集合 $T=\bigcup T_i$, 实现知识图谱的跨层连接. 如后文图 1 所示, 具体而言, 本文构建的双层医疗知识图谱中, 在患者层和疾病百科层都存在疾病这一实体类型 (用蓝色节点表示), 因此, 本文对患者层的英文疾病实体和疾病百科层的中文疾病实体进行语义相似度匹配, 得到共指相同语义信息的疾病实体对集合, 实现跨层连接.

面向医学事实的双层知识图谱构建可主要划分为 3 个部分: UKB 动态多模态病例图谱构建、中文百科知识图谱构建以及中英文知识图谱融合, 整体构建流程如后文图 2 所示.

2.2 UKB 动态多模态病例图谱构建

2.2.1 UKB 病例数据获取

由于中文患者病例数据较少, 且因涉及患者隐私多数无法公开, 本节基于英国生物银行 (UK Biobank, 简称 UKB) 医学数据库构建病例知识图谱. UKB 是目前世界上最为知名和开放的生物银行, 自 2006 年建立以来已收集了英国超过 50 万名参与者的血液、尿液和唾液等样本, 获取了完善的人口学、社会经济、生活方式和患者诊断信息以及多种形式的医学影像数据, 具有样本量大、种类丰富、数据质量高的特点.

本文研究团队已获得 UKB 的授权, 从中提取了 502 409 个患者及相关的医学事实, 主要包括社会人口特征、生化测量、心理认知状态、在线随访信息、入院诊断数据和 MRI 成像等.

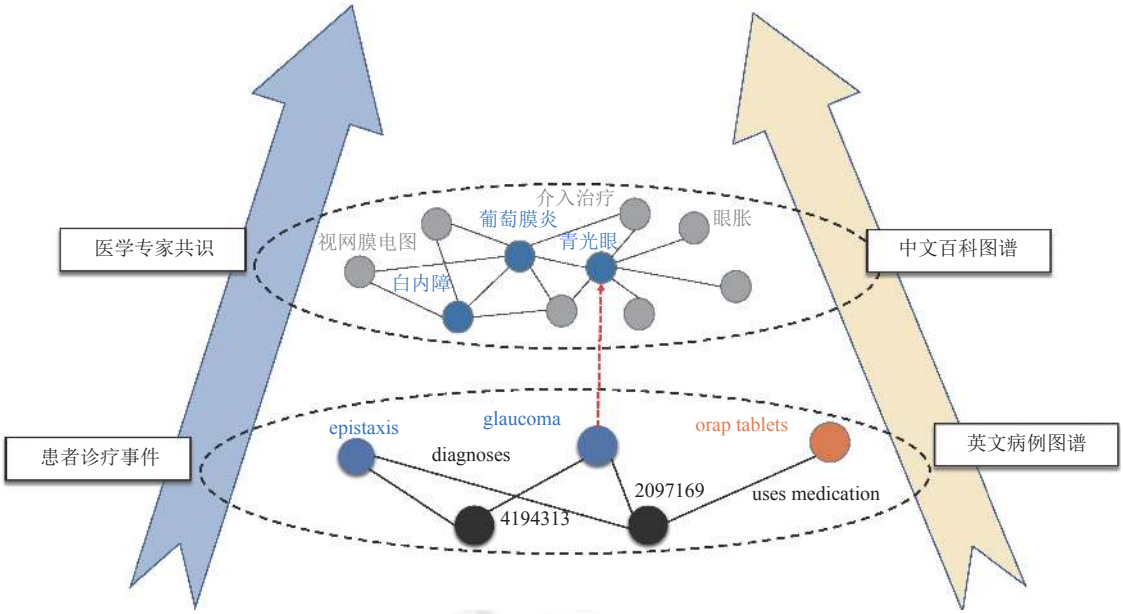


图 1 面向医学事实的双层知识图谱表示框架

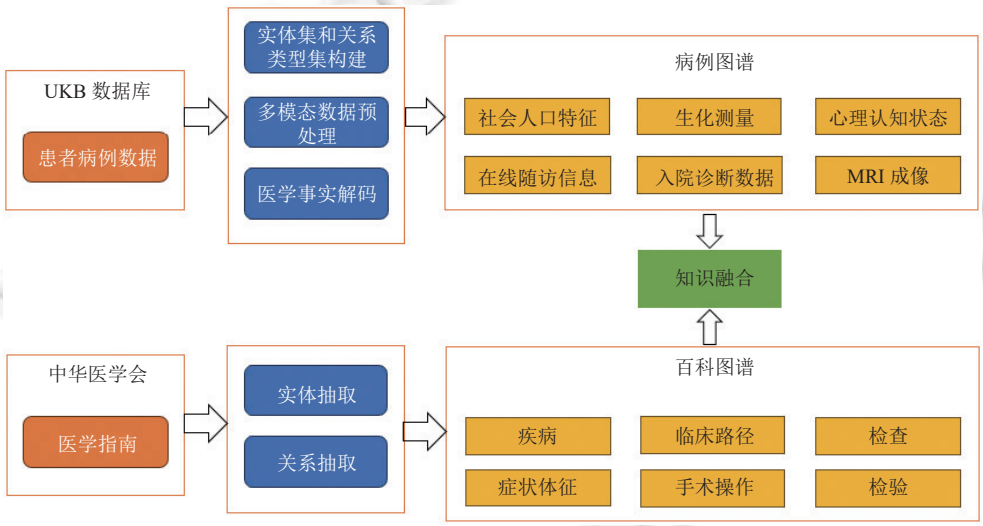


图 2 面向医学事实的知识图谱构建总体流程

2.2.2 病例医学事实三元组构建

在构建阶段之前, UKB 原始数据被处理为 csv 格式以便进行患者医学事实的提取. UKB 数据由 3 个 csv 格式的表格组成, 包括患者病例表、域名信息表和医学事实映射表, 分别记录了患者关联的医学事实、域名的相关信息和从域名编码、诊疗结果编码到诊疗结果的映射关系, 域名可以理解为与患者相关的某一个医学事实的属性. 在 UKB 中, 一些域名具有多个实例, 用来表示相应医学事实的采集时段.

在构建实体集阶段, 本文首先剔除了原始表中的空值, 然后分别从患者病例表和医学事实映射表提取患者实体 (以 eid 作为唯一标识) 和各类诊疗结果, 确定用于构建医疗知识图谱的全部实体. 具体而言, 对于数值型、离散型和文本型的诊疗结果, 直接将其名称作为实体名称构建实体, 对于 XML 格式的心电图、DCM 格式的影像等图像诊疗结果, 本文将其名称作为实体名称构建实体, UKB 提供了下载医疗图像的工具 ukbfetch 和批量下载的工具

ukbconv, 在批量下载图像实体后, 本文将各种格式的图像统一处理成 png 格式, 使用 Tomcat 为每个图像分配一个 URL 作为实体属性.

在构建关系集阶段, 本文将 UKB 中每一种域名作为一种关系类型. 对于多实例域名, 本文赋予相应关系时间属性, 将实例作为关系的时间属性值, 由此表达某一时段患者发生某一医学事实, 实现了 UKB 病例知识图谱的动态更新.

在构建事实三元组阶段, 本文对患者关联的每一个医学事实进行解码, 具体而言, 患者病例表的第 1 列是患者 eid, 其他列对应患者的不同医学事实, 列名由域名 ID、实例组成, 列值是诊疗结果编码, 对于每一列, 本文首先查询域名信息表, 得到域名 ID 对应的域名和域名编码, 从而得到对应的关系名称, 并将实例作为关系的时间属性值记录下来, 然后查询医学事实映射表, 得到域名编码和诊疗结果编码对应的诊疗结果. 通过上述方法, 本文得到了包含 50 多万的患者信息的医学事实三元组 (患者 eid, 关系, 诊疗结果), 其中部分关系类型被赋予时间属性, 用于表示该事实三元组发生的时刻.

2.3 中文百科知识图谱构建

2.3.1 百科医学知识数据获取

为了保证医疗知识图谱的科学性和权威性, 本文数据来源于国家一级出版单位和国家级医疗专业学会、协会, 主要有以下几个方面: (1) 权威行业协会、学会在国家卫生健康委员会领导和组织下制定的规范、指南、标准和路径等; (2) 国家卫生健康委“十二五”“十三五”医药卫生规划教材和权威医药卫生专著等; (3) 诊疗规范; (4) 相关医学文献; (5) 药物说明书; (6) 按照疾病组织专家编写适合基层医药诊疗需要的知识条目等.

2.3.2 百科医学事实三元组构建

在医学领域积累了海量的非结构化的医学指南, 其中蕴含丰富的医疗百科知识可认定为医学事实. 为了从中提取到高质量的中文百科知识图谱, 本文设计了一套完整的医疗知识挖掘和抽取技术, 旨在从大量的医疗指南中发现新的结构化的百科知识, 经清洗和人工审核之后, 添加到已有的知识库体系中, 从而保证知识库的完备性和实时性及迭代更新的能力.

医学指南实体抽取. 考虑到基于深度学习的实体抽取算法, 减少了对手工规则的依赖, 无需过多的特征工程, 在大规模医疗数据背景下能够取得较好的抽取性能, 因此本文采用基于 BERT (bidirectional encoder representation from Transformers) 加自注意力机制的多层序列化深度学习标注模型, 实现了嵌套实体和顶层实体的联合抽取.

如图 3 所示, 标注模型首先使用医疗领域预训练的 BERT 模型对输入句子进行编码, 在第 1 层, 首先识别底层实体元素, 该例中, 可以识别出“四肢”为身体部位, “乏力”为症状元素 (图中简写为“元素”); 识别出第 1 层的实体之后, 将识别出的实体中的每个字组合成一个整体表示, 用以识别更高层的实体. 在训练阶段, 设置一个固定的最大实体层数, 按上述方式依次计算每一层的损失, 将每一层的损失相加, 进行联合训练. 在测试阶段, 依次预测每一层的实体, 若在某一层没有标记出实体, 则预测结束. 通过上述方法, 本文对收集到的医学指南进行实体抽取, 得到包括疾病、症状、检查、药物、食物等类别的实体.

医学指南关系抽取. 医学领域相对专业封闭, 其实体类型可枚举 (如疾病、药品、特殊人群等), 因此实体之间的关系也是有限的、可枚举的, 可通过预定义实体与实体之间的关系, 将抽取任务转换成分类任务来解决. 如抽取病历中病原体与疾病之间的关系、疾病与手术之间的关系、症状与属性之间的关系、药物与病原体之间的关系等. 由于医学知识的快速更新, 仅依靠专家整理已经不能满足需求. 为了快速丰富知识库中的实体关系, 本文采用远程监督学习和人工核查的方案, 整体方案如图 4 所示.

首先, 通过医疗专家构建小型专家知识库, 从无结构化数据中 (如病历、医学百科、医学书籍等) 识别出实体, 并将其链接到已有知识库中; 接着, 从已整理好的知识库中抽取出关系三元组, 将其映射到无结构数据中, 从而自动构建含有噪声的训练数据; 利用上述自动构建的数据, 训练基于远程监督的关系抽取模型, 从而可以发现新的三元组; 新发现的三元组交由专家审核, 确认无误后再更新到知识图谱中. 如对于“C 型骨盆骨折的患者可以通过骨盆环前后联合固定的方式进行治疗”这一句子, 通过实体抽取模型可以识别出“C 型骨盆骨折”和“骨盆环前后联合固定”两个实体, 将其链接到已有的知识库中可以查出对应关系为“治疗方案”, 由此得到关系三元组 (C 型骨盆骨

折, 治疗方案, 骨盆环前后联合固定) 作为训练数据训练关系抽取模型. 该模型通过 BERT 获取训练数据的嵌入表示, 并通过一个多分类层输出句子中实体的语义关系. 本节构建的中文百科知识图谱主要包括并发、检查方法、治疗方案、用药、适宜食物、禁忌食物以及治愈率、治疗费用等关系.

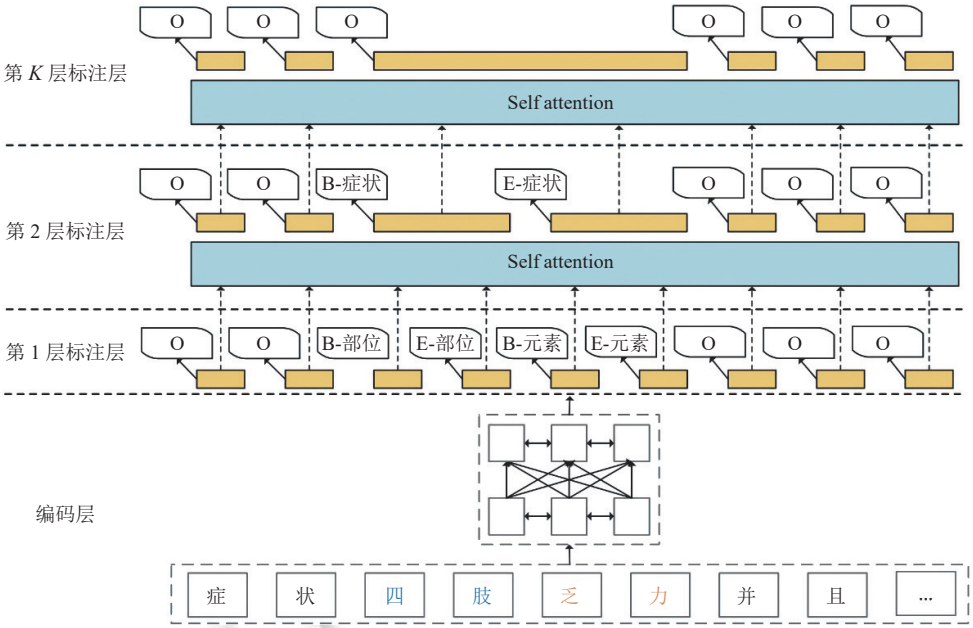


图 3 嵌套实体识别模型

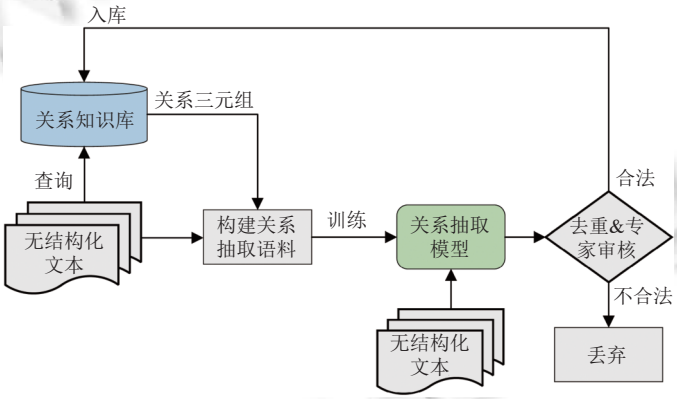


图 4 远程监督学习和人工核查

2.4 中英文知识图谱融合

如图 5 所示, 本文提出利用预训练翻译模型和字符串相似匹配实现英文 UKB 病例图谱与中文医疗百科知识图谱的融合.

如算法 1 所示, 本文首先从已构建的英文病例图谱和中文百科知识图谱中分别获取英文实体集和中文实体集, 其中, UKB 病例图谱中的疾病实体由 ICD10 和英文描述组成, 由空格符隔开, 本文通过正则表达式 "∧[A−Z]?[0−9]{3}\.[0−9]\s(.∗)"识别出英文描述作为实体构建英文疾病实体集 D_{en}; 中文实体类型可枚举且与实体关系具有明显的映射关系, 对于三元组 (h,r,t), 如果 r 是"存在症状", 就可以判断 h 是疾病实体, 本文通过映射规则识别出中文疾病实体并加入中文疾病实体集合 D_{zh}. 其次, 本文将英文疾病实体经预训练翻译模型 *BERTTranslation*

翻译为中文, 并利用中文疾病术语标准化函数 *diseaseStd* 将其进行术语标准化, 与原始英文疾病实体构成实体对 $(ent_{origin}, ent_{trans})$ 加入翻译实体对集合 P_{trans} 中. 然后, 对于该集合的每一个中文翻译 ent_{trans}, 本文采用字符串的 Jaccard 相似度匹配的方法遍历中文疾病实体集将阈值大于 0.85 的实体提取出来, 得到候选中文实体集, 并选择其中分数最高的中文实体, 与相应英文疾病实体 ent_{origin} 构成新增的实体对, 通过上述方式, 最终实现了英文病例图谱和中文百科图谱的融合.

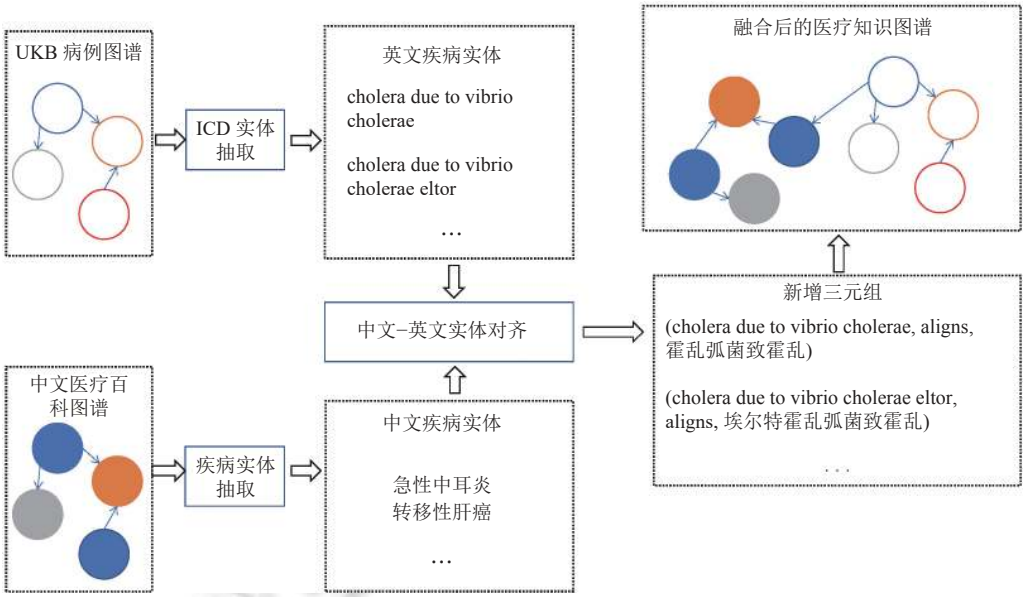


图 5 跨语言知识图谱融合

算法 1. 中英文知识图谱融合算法.

输入: 英文实体集 E_{en}, 中文实体集 E_{zh};
输出: 英文-中文实体对集合 P_{aligns}.

1. 初始化: $D_{en} \leftarrow \{\}, D_{zh} \leftarrow \{\}, P_{trans} \leftarrow \{\}, P_{aligns} \leftarrow \{\}$;
2. 根据正则表达式从 E_{en} 中选取实体加入 D_{en} 中;
3. 根据映射规则从 E_{zh} 中选取实体加入 D_{zh} 中;
4. **for** $ent_{origin} \in D_{en}$ **do**
5. $ent_{trans} \leftarrow BERTTranslation(ent_{origin})$;
6. $ent_{trans} \leftarrow diseaseStd(ent_{trans})$;
7. 将 $(ent_{origin}, ent_{trans})$ 实体对加入 P_{trans} 中;
8. **end for**
9. **for** $(ent_{origin}, ent_{trans})$ in P_{trans} **do**
10. $D_{candidate} \leftarrow \{\}$;
11. **for** ent_{zh} in D_{zh} **do**
12. **if** $Jaccard(ent_{zh}, ent_{trans}) > 0.85$ **then**
13. 将 ent_{zh} 和对应的 *Jaccard score* 加入 $D_{candidate}$ 中;
14. **end if**
15. **end for**

16. 从 $D_{candidate}$ 中选取 *Jaccard score* 最高的实体 ent_{zh}, 将 (ent_{origin}, ent_{zh}) 加入 P_{aligns} 中;
17. **end for**

通过应用上述信息抽取和知识融合技术对英文患者病例数据和中文疾病百科数据进行分析处理, 本文最终得到跨语言医疗知识图谱的统计信息, 如表 1 所示, 其类别信息如表 2 所示. 可以看出, 英文实体主要包括患者、疾病、药物和图像等, 其中, 患者是英文知识图谱的核心实体类型, 与包括疾病、药物在内的各种医学实体通过诊断、用药等关系相连. 中文实体主要包括疾病、症状、检查等类型的实体, 与英文知识图谱不同, 中文知识图谱以疾病为核心, 通过检查方法、用药等关系与各类实体相连.

表 1 跨语言知识图谱数据统计

类型	统计量
患者实体	502 409
实体总数	17 261 282
关系	8 219
事实三元组	805 065 729

表 2 跨语言知识图谱类别信息示例

类型	类别信息示例
英文实体	疾病 (constipation、glaucoma等)、药物 (ibuprofen、paracetamol等)、患者、图像等
英文关系	诊断、用药、脂肪测量、睡眠时间等
中文实体	疾病 (青光眼、白内障等)、症状 (眼胀、视力障碍等)、检查 (瞳孔检查、视野检查等)、药物 (盐酸左氧氟沙星片、颈复康颗粒等)、食物 (蜂蜜、羊肉等)、治疗 (中医治疗、介入治疗等)等
中文关系	别名、并发、检查方法、用药、禁忌食物、适宜食物、疾病症状、治疗方案等

3 实验与结果分析

3.1 用药推测

为了验证双层知识图谱架构的有效性, 本文设计了一个用药推测实验, 实验分为两组: 第 1 组仅使用英文病例图谱推测患者的用药情况, 由于疾病是患者用药的主要因素, 本文选取英文疾病信息作为患者特征; 第 2 组使用英文病例图谱和中文百科图谱推测患者的用药情况, 在英文疾病的基础上, 加入中文百科图谱中的疾病症状知识作为信息补充, 构建患者特征. 本实验的推断目标是患者是否服用 aspirin (阿司匹林) 药物, 是一个二分类问题. 本文选取了阳性样本和阴性样本各 500 个, 在将患者特征表示成 multi-hot 向量后, 训练多层感知机 (multi-layer perceptron, MLP) 输出患者的用药预测值, 并与真实标签进行比较, 得到的实验结果如表 3 所示.

表 3 用药推测结果对比

患者信息	*AUROC*	*AUPRC*	*F*1	*Accuracy*
英文疾病	0.943 3	0.959 9	0.891 6	0.880 0
英文疾病+中文症状	**0.948 8**	**0.967 6**	**0.904 1**	**0.906 7**

由结果可知, 与仅使用英文病例图谱相比, 双层图谱在 4 个指标上都获得了性能提升, 这是因为单一疾病名称无法全面反映患者的健康情况, 通过补充中文疾病百科图谱中某些与药物关联密切的疾病症状 (如头痛、乏力等), 可以更有效地推测患者的用药情况, 由此验证了双层架构的有效性.

3.2 跨层连接质量评估

为了验证跨语言相似匹配算法的有效性, 本文对双层知识图谱中存在的中英文疾病实体对的正确率进行了评估. 根据世界卫生组织 (World Health Organization, WHO) 提供的 ICD10 编码与英文疾病的对应关系 (https://icd.who.int/browse10/2019/en)、国家医疗保障局提供的 ICD10 编码与中文疾病的对应关系 (https://nhsa.gov.cn/

jbzd/public/dataWesterSearch.html), 本文生成了中英文疾病对齐的标签集 $Y=\{(y^i_{en},y^i_{zh})\}$. 然后, 对于本文知识图谱中的中英文疾病对 (d^i_{en},d^i_{zh}), 找到 d^i_{en} 在 Y 中对应的中文疾病, 如果其与 d^i_{zh} 完全匹配, 则判定 (d^i_{en},d^i_{zh}) 正确, 否则错误. 本文对眼、耳鼻喉、肿瘤、肾和心血管这 5 类疾病 446 个实体对进行了评估, 经统计, 正确率达到 78.3%.

经分析, 导致部分中英文疾病对存在错误的原因主要是中文疾病实体命名不规范, 由于本文通过对大量非结构化文本进行实体识别的方式抽取中文疾病实体, 难以保证原始语料的疾病实体符合 ICD10 的疾病命名规范, 如“高血压 1 级”等.

3.3 基于嵌入方法的患者信息质量评估

本节结合知识图谱嵌入方法对所构建图谱中患者信息的质量进行评估. 具体而言, 考虑到疾病是患者用药的主要因素, 本文选取英文知识图谱中与 1 000 个患者相关的 (患者, 诊断, 疾病) 三元组, 使用 TransE 模型学习患者的嵌入表示. 然后, 本文将患者表示线性映射到一维空间下, 推测患者对 aspirin 药物的服用情况. 如第 3.1 节所述, 此问题是一个二分类问题. 经与真实标签进行比较, 本文得到 *AUROC*=0.8984、 *AUPRC*=0.8988、*F*1=0.7304、*Accuracy*=0.7933 的实验结果, 证明了知识图谱中患者信息具有较高的质量.

4 可视化展示

本文利用 Neo4j 工具实现跨语言知识图谱的可视化展示. Neo4j 是一个高性能的 NoSQL 图形数据库, 它将结构化数据以节点、关系和属性的形式存储在图上, 其中节点代表实体, 连边代表实体关系, 两者均可以设置属性. 通过图数据的可视化界面, 用户可以直观地看出实体和实体信息之间的关联. Neo4j 提供了 Cypher 查询语言, 它是一种面向图分析、声明式的描述性图形查询语言[30], 可以更精准、高效地实现对图数据的查询.

4.1 跨语言知识图谱概览

本文以 eid 为 1000724 的患者为例, 对患者展开诊断和用药查询, 以说明知识图谱的概览和构成, 如图 6 所示. 可以看出, 该患者具有 8 个诊断信息和 5 个用药信息, 其诊断信息与中文疾病百科相关联. 具体而言, 患者关联的英文疾病节点 glaucoma 通过 aligns 关系连接到中文百科知识图谱的疾病“青光眼”上, “青光眼”节点包含了关于该疾病的症状、检查方法和治疗方案、适宜食物和禁忌食物等信息, 如通常会出现“眼胀”“视力障碍”等症状, 不宜吃“猪油”“鸭肝”等食物. 另外, 患者还诊断出“鼻出血”“鼻中隔偏曲”等疾病, 可以通过本知识图谱查询这些疾病的相关信息, 为该患者的临床诊疗提供参考意见和依据.

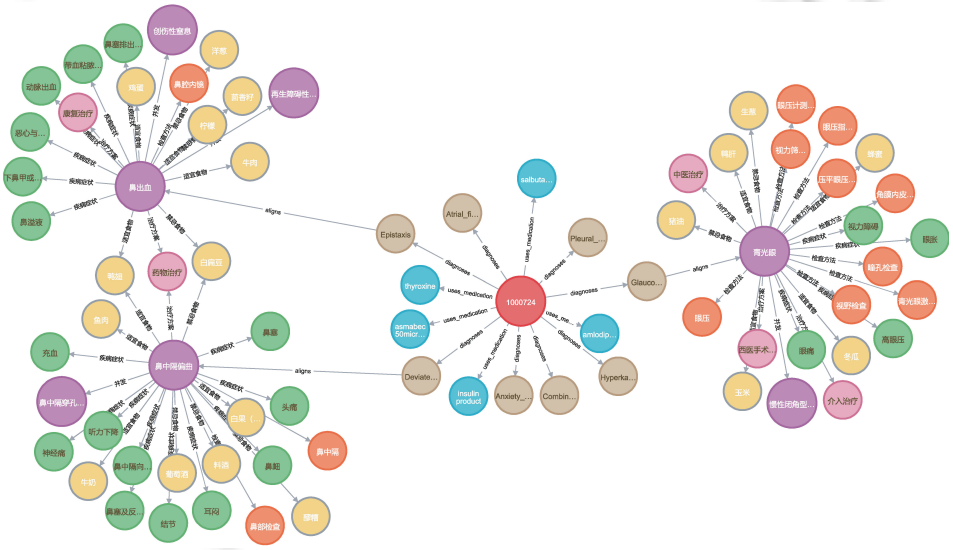


图 6 跨语言知识图谱概览

4.2 多模态信息展示

本文构建的知识图谱包含文本和多种格式的图像信息, 如 DCM 格式的患者影像、XML 格式的心电图等. 本节以英文病例图谱中的患者节点 1012980 的诊断、用药信息以及胰腺图像信息为例进行多模态展示, 如图 7 所示. 可以发现, 患者与多种诊断、用药和图像实体相连, 其中诊断和用药信息以文本形式存储, 图像信息通过 URL 进行访问. 具体而言, 患者通过 Abdominal MRI 关系与 Pancreas image 节点相连, 通过选中影像节点, 可以得到该节点的 URL 信息, 进而通过 URL 信息可以访问到医学影像. 通过上述示例表明, 本文构建的医疗知识图谱可以有效存储并展示患者的多模态诊断信息, 辅助医生进行更为精准、全面的患者医疗信息查询.

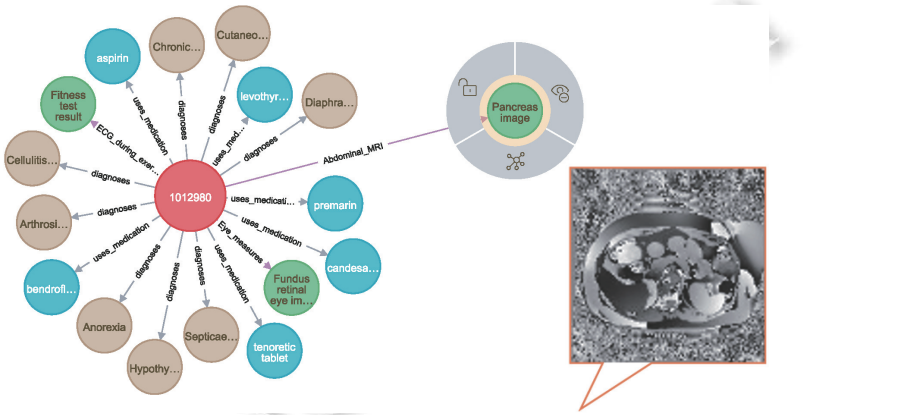


image **<id>**:2657350 **name**: Pancreas image **url**: http://172.18.116.37:8081/app/1012980_20260_0_0.png

图 7 多模态展示

4.3 知识图谱动态属性展示

本文构建的知识图谱中的患者-医学事实连边部分具有时间属性, 用以刻画患者在不同时刻的生理测量等医疗检测信息. 为了展示所构建图谱的动态属性, 本节以患者节点 1000074 为例, 通过指定连边的时间属性值, 分别查询患者在不同时间段的情况, 如 2006–2010 年时间段对应的查询语句为“MATCH (h:patient)-[r:Vitamin_D]->(t:fact) WHERE h.value=‘1000074’ and r.time=‘2006–2010’ RETURN t”. 如图 8 所示, 可以看出, 在 2006–2010 年时间段, 患者对维生素 D 的平均摄入量为 23.8 μg, 在后 3 个时间段的摄入量为 2.49 μg、2.08 μg、4.03 μg. 维生素 D 是一种固醇类衍生物, 具有增强肌体对钙、磷的吸收的作用, 患者在后 3 个时间段的维生素 D 摄入量低, 这说明有骨质疏松的风险. 上述实例表明, 时间演化信息可以辅助医生更直观地了解患者的病史和最新情况, 并为合理预测患者的未来情况提供了可能.

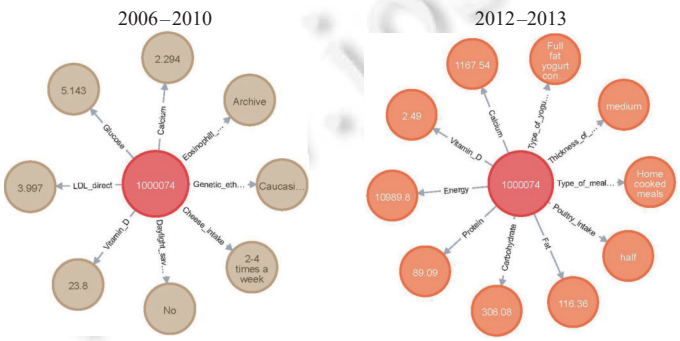


图 8 患者相关的医学事实不断更新

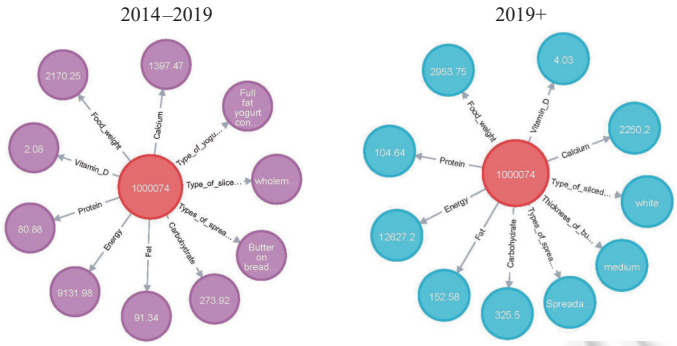


图 8 患者相关的医学事实不断更新 (续)

5 结论与未来的工作

本文提出了一种双层医疗知识图谱架构, 并通过对包括英文患者诊疗事件和中文专家共识的医学事实进行分析处理, 提取得到有价值的结构化数据, 构建一个跨语言、多模态、动态更新的医疗知识图谱. 具体而言, 本文首先从 UKB 患者病例数据中抽取到实体集和关系集, 对医学事实编码进行解码构建医学事实三元组, 得到英文病例知识图谱; 然后通过基于深度学习模型的实体抽取方法和基于远程监督的关系抽取方法得到中文百科知识图谱, 进而使用跨语言相似匹配的方法实现了实体对齐和知识融合. 本文通过用药推测实验、跨层连接质量评估和患者信息质量评估验证了所构建双层图谱的有效性, 并利用 Neo4j 图数据库对知识图谱进行了可视化展示, 验证了该图谱具有更为精准、全面的面向大规模医学事实的知识表达能力.

后续研究工作重点围绕 3 个方面展开. 一是随着新的患者病例不断引入, 医学指南不断更新, 需要探索针对现有知识图谱的动态更新机制, 自动化地高效引入并整合新的知识, 实现知识图谱的实时更新和规模扩展. 二是现代循证医学重视客观医学证据匹配, 在诊断过程中需要查询病例事实来提供临床诊断支撑, 未来将探索基于动态医学事实的医学证据涌现机制, 实现跨语言知识图谱对临床指南的辅助迭代与优化, 增强知识图谱面向真实医疗场景的实用性. 三是知识抽取技术创新. 本文目前基于 BERT 等模型实现中文百科知识抽取, 随着大模型技术的不断发展, 我们未来拟采用更大规模的语言模型来实现更高质量的知识抽取.

References:

[1] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. Nature Medicine, 2019, 25(1): 24–29. [doi: 10.1038/s41591-018-0316-z]

[2] Dong WB, Sun SL, Yin MZ. Research and development of medical knowledge graph reasoning. Journal of Frontiers of Computer Science and Technology, 2022, 16(6): 1193–1213 (in Chinese with English abstract). [doi: 10.3778/j.issn.1673-9418.2111031]

[3] Sulakhe D, Balasubramanian S, Xie BQ, Feng B, Taylor A, Wang S, Berrocal E, Dave U, Xu JB, Börnigen D, Gilliam TC, Maltsev N. Lynx: A database and knowledge extraction engine for integrative medicine. Nucleic Acids Research, 2014, 42(D1): D1007–D1012. [doi: 10.1093/nar/gkt1166]

[4] Cong Q, Feng ZY, Li F, Zhang L, Rao GZ, Tao C. Constructing biomedical knowledge graph based on SemMedDB and linked open data. In: Proc. of the 2018 IEEE Int'l Conf. on Bioinformatics and Biomedicine. Madrid: IEEE, 2018. 1628–1631. [doi: 10.1109/BIBM.2018.8621568]

[5] Huo CG, Ma ST, Liu XZ. Hotness prediction of scientific topics based on a bibliographic knowledge graph. Information Processing & Management, 2022, 59(4): 102980. [doi: 10.1016/J.IPM.2022.102980]

[6] Jia LR, Liu J, Yu T, Dong Y, Zhu L, Gao B, Liu LH. Construction of traditional Chinese medicine knowledge graph. Journal of Medical Informatics, 2015, 36(8): 51–53, 59 (in Chinese with English abstract). [doi: 10.3969/j.issn.1673-6036.2015.08.012]

[7] Zhao C, Jiang JC, Xu ZM, Guan Y. A study of EMR-based medical knowledge network and its applications. Computer Methods and Programs in Biomedicine, 2017, 143: 13–23. [doi: 10.1016/j.cmpb.2017.02.016]

[8] Byambasuren O, Yang YF, Sui ZF, Dai DM, Chang BB, Li SJ, Zan HY. Preliminary study on the construction of Chinese medical

knowledge graph. Journal of Chinese Information Processing, 2019, 33(10): 1–9 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2019.10.001]

[9] Li LF, Wang P, Yan J, Wang Y, Li SM, Jiang JP, Sun Z, Tang BZ, Chang TH, Wang SH, Liu YT. Real-world data medical knowledge graph: Construction and applications. Artificial Intelligence in Medicine, 2020, 103: 101817. [doi: 10.1016/j.artmed.2020.101817]

[10] Chang DJ, Chen MS, Liu CZ, Liu LP, Li DD, Li W, Kong F, Liu BC, Luo XB, Qi J, Jin Q, Xu B. DiaKG: An annotated diabetes dataset for medical knowledge graph construction. In: Proc. of the 6th China Conf. on Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction. Guangzhou: Springer, 2021. 308–314. [doi: 10.1007/978-981-16-6471-7_26]

[11] Kang L. Research and implementation of cardiovascular disease question answering system based on knowledge graph [MS. Thesis]. Guangzhou: South China University of Technology, 2020 (in Chinese with English abstract). [doi: 10.27151/d.cnki.ghnlu.2020.000819]

[12] van der Vet PE, Mars NJI. Bottom-up construction of ontologies. IEEE Trans. on Knowledge and Data Engineering, 1998, 10(4): 513–526. [doi: 10.1109/69.706054]

[13] Rodríguez H, Climent S, Vossen P, Bloksma L, Peters W, Alonge A, Bertagna F, Roventini A. The top-down strategy for building EuroWordNet: Vocabulary coverage, base concepts and top ontology. Computers and the Humanities, 1998, 32(2): 117–152. [doi: 10.1023/A:1001169525131]

[14] Wu ST, Liu HF, Li DC, Tao C, Musen MA, Chute CG, Shah NH. Unified medical language system term occurrences in clinical notes: A large-scale corpus analysis. Journal of the American Medical Informatics Association, 2012, 19(e1): e149–e156. [doi: 10.1136/amiajnl-2011-000744]

[15] Wang YF, Patrick J. Cascading classifiers for named entity recognition in clinical notes. In: Proc. of the Workshop on Biomedical Information Extraction. Borovets: ACM, 2009. 42–49.

[16] Zhou GD, Zhang J, Su J, Shen D, Tan C. Recognizing names in biomedical texts: A machine learning approach. Bioinformatics, 2004, 20(7): 1178–1190. [doi: 10.1093/bioinformatics/bth060]

[17] Chowdhury S, Dong XS, Qian LJ, Li XF, Guan Y, Yang JF, Yu QB. A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records. BMC Bioinformatics, 2018, 19(17): 499. [doi: 10.1186/s12859-018-2467-9]

[18] Ji B, Liu R, Li SS, Yu J, Wu QB, Tan YS, Wu JJ. A hybrid approach for named entity recognition in Chinese electronic medical record. BMC Medical Informatics and Decision Making, 2019, 19(S2): 64. [doi: 10.1186/s12911-019-0767-2]

[19] Tan CQ, Qiu W, Chen MS, Wang R, Huang F. Boundary enhanced neural span classification for nested named entity recognition. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 9016–9023. [doi: 10.1609/aaai.v34i05.6434]

[20] Guo YJ. Design and implementation of event evolution graph constructing system based on electronic medical records [MS. Thesis]. Shenyang: Shenyang Institute of Computing Technology, Chinese Academy of Sciences, 2022 (in Chinese with English abstract). [doi: 10.27587/d.cnki.gksjs.2022.000033]

[21] Abacha AB, Zweigenbaum P. A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. In: Proc. of the 12th Int'l Conf. on Computational Linguistics and Intelligent Text Processing. Tokyo: Springer, 2011. 139–150. [doi: 10.1007/978-3-642-19437-5_11]

[22] Quan CQ, Hua L, Sun X, Bai WJ. Multichannel convolutional neural network for biological relation extraction. BioMed Research Int'l, 2016, 2016: 1850404. [doi: 10.1155/2016/1850404]

[23] Zeng DJ, Zhang RR, Liu QY. CopyMTL: Copy mechanism for joint extraction of entities and relations with multi-task learning. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 9507–9514. [doi: 10.1609/aaai.v34i05.6495]

[24] Zhou WX, Huang K, Ma TY, Huang J. Document-level relation extraction with adaptive thresholding and localized context pooling. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2021. 14612–14620. [doi: 10.1609/aaai.v35i16.17717]

[25] Shen WH, Zhong YF, Wang JJ, Zheng Z, Ma AL. Construction and application of flood disaster knowledge graph based on multi-modal data. Geomatics and Information Science of Wuhan University, 2023, 48(12): 2009–2018 (in Chinese with English abstract). [doi: 10.13203/j.whugis20220509]

[26] E HH, Cheng R, Song MN, Zhu PC, Wang Z. A joint embedding method of relations and attributes for entity alignment. Int'l Journal of Machine Learning and Computing, 2020, 10(5): 605–611. [doi: 10.18178/ijmlc.2020.10.5.980]

[27] Dieng-Kuntz R, Minier D, Růžička M, Corby F, Corby O, Alamarguy L. Building and using a medical ontology for knowledge management and cooperative work in a health care network. Computers in Biology and Medicine, 2006, 36(7–8): 871–892. [doi: 10.1016/j.compbiomed.2005.04.015]

[28] Zhai X, Pan HW, Xie XQ, Zhang ZQ, Han QL. Parallel loading algorithm for multimode medical data fusion. Journal of Data Acquisition and Processing, 2018, 33(4): 758–768 (in Chinese with English abstract). [doi: 10.16337/j.1004-9037.2018.04.020]

[29] Lacoste-Julien S, Palla K, Davies A, Kasneci G, Graepel T, Ghahramani Z. SIGMa: Simple greedy matching for aligning large knowledge bases. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Chicago: ACM, 2013.

572–580. [doi: 10.1145/2487575.2487592]

[30] Zhang WC, Wang F, Huang Y. Knowledge modeling of big data policy in Guizhou Province based on graph database. Digital Library Forum, 2020(4): 30–38 (in Chinese with English abstract). [doi: 10.3772/j.issn.1673-2286.2020.04.005]

附中文参考文献:

[2] 董文波, 孙仕亮, 殷敏智. 医学知识推理研究现状与发展. 计算机科学与探索, 2022, 16(6): 1193–1213. [doi: 10.3778/j.issn.1673-9418.2111031]

[6] 贾李蓉, 刘静, 于彤, 董燕, 朱玲, 高博, 刘丽红. 中医药知识图谱构建. 医学信息学杂志, 2015, 36(8): 51–53, 59. [doi: 10.3969/j.issn.1673-6036.2015.08.012]

[8] 奥德玛, 杨云飞, 穗志方, 代达劢, 常宝宝, 李素建, 昝红英. 中文医学知识图谱 CMeKG 构建初探. 中文信息学报, 2019, 33(10): 1–9. [doi: 10.3969/j.issn.1003-0077.2019.10.001]

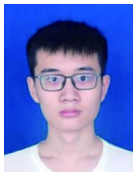
[11] 康莉. 基于知识图谱的心血管病问答系统的研究与实现 [硕士学位论文]. 广州: 华南理工大学, 2020. [doi: 10.27151/d.cnki.ghnlu.2020.000819]

[20] 郭宇捷. 基于电子病历的事理图谱构建系统的设计与实现 [硕士学位论文]. 沈阳: 中国科学院大学 (中国科学院沈阳计算技术研究所), 2022. [doi: 10.27587/d.cnki.gksjs.2022.000033]

[25] 沈伟豪, 钟燕飞, 王俊珏, 郑卓, 马爱龙. 多模态数据的洪涝灾害知识图谱构建与应用. 武汉大学学报 (信息科学版), 2023, 48(12): 2009–2018. [doi: 10.13203/j.whugis20220509]

[28] 翟霄, 潘海为, 谢晓芹, 张志强, 韩启龙. 支持多模态医学数据融合的并行加载算法. 数据采集与处理, 2018, 33(4): 758–768. [doi: 10.16337/j.1004-9037.2018.04.020]

[30] 张维冲, 王芳, 黄毅. 基于图数据库的贵州省大数据政策知识建模研究. 数字图书馆论坛, 2020(4): 30–38. [doi: 10.3772/j.issn.1673-2286.2020.04.005]



王楚童(1999－), 男, 博士生, 主要研究领域为深度学习, 数据挖掘, 表示学习.

杨雪冰(1991－), 男, 博士, 副研究员, 主要研究领域为机器学习, 大数据知识挖掘, 人工智能与辅助医疗及智慧气象的交叉应用.

李明达(1992－), 男, 博士, 助理研究员, 主要研究领域为机器学习, 数据挖掘, 知识表示与推理.

牛景昊(1993－), 男, 博士, 副研究员, 主要研究领域为知识表示学习, 可解释人工智能, 医疗大数据分析.

孙孟轩(1997－), 男, 博士生, 主要研究领域为机器学习, 医疗数据挖掘, 临床预测模型.

贺志阳(1981－), 男, 博士, CCF 专业会员, 主要研究领域为语音识别, 自然语言理解, 人工智能.

王静(1987－), 女, 硕士, 主要研究领域为医疗认知推理, 自然语言理解.

张文生(1965－), 男, 博士, 研究员, 博士生导师, CCF 专业会员, 主要研究领域为机器学习, 大数据分析, 知识挖掘.