医疗多模态信息抽取技术评测数据集概述*

宗 辉 雷健波 李作峰

（同济大学 上海 200092）（北京大学医学信息学中心 北京 100091）（武田中国创新孵化器 上海 200126）

夏静波 陈漠沙

（华中农业大学信息学院 武汉 430070） （阿里巴巴 杭州 310000）

王晓玲 常德杰 康 波

（华东师范大学 上海 200062） （北京环球医疗救援 北京 100020） （医渡云（北京）技术有限公司 北京 100191）

李 姣 汤步洲

（中国医学科学院/北京协和医学院医学信息研究所 北京 100020） （哈尔滨工业大学（深圳）鹏城实验室 深圳 518055）

〔摘要〕 阐述医疗多模态信息抽取技术评测数据集的结构、构建方法、应用情况等，包括面向“基因－疾病”的关联语义挖掘数据集、中文医疗因果关系抽取数据集、医疗文本诊疗决策树抽取数据集、医疗材料要素提取数据集、临床诊断编码数据集，提出上述数据集有望为各种技术、算法以及系统的评估和实施提供有力支撑和参考。

〔关键词〕 中国健康信息处理会议；多模态信息抽取；医学数据集；人工智能；自然语言处理

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.12.001

Overview of Technology Evaluation Dataset for Medical Multimodal Information Extraction *ZONG Hui*, *Tongji University*, *Shanghai 200092*, *China*; *LEI Jianbo*, *Center for Medical Informatics*, *Peking University*, *Beijing 100091*, *China*; *LI Zuofeng*, *Takeda*

〔修回日期〕 2022－09－27

〔作者简介〕 宗辉，博士，发表论文7 篇；通信作者：汤步洲，博士，副教授，博士生导师。

〔基金项目〕 科技创新2030——“新一代人工智能”重大项目“儿科疾病复杂循证知识图谱构建研究”（项目编号：2021ZD0113402）；国家自然科学基金“基于电子病历深层语义分析的动态临床辅助决策方法研究”（项目编号：62276082）；国家自然科学基金“基于电子病历分析的慢病趋势预测方法研究”（项目编号：61876052）；广东省自然科学基金“基于深度联合学习的医疗实体及属性联合抽取”（项目编号：2019A1515011158）。

China, *Shanghai 200126*, *China*; *XIA Jingbo*, *College of Informatics*, *Huazhong Agricultural University*, *Wuhan 430070*, *China*; *CHEN Mosha*, *Alibaba Group*, *Hangzhou 310000*, *China*; *WANG Xiaoling*, *East China Normal University*, *Shanghai 200062*, *China*; *CHANG Dejie*, *Beijing Universal Medical Assistance*, *Beijing 100020*, *China*; *KANG Bo*, *Yidu Cloud (Beijing) Technology Co. Ltd.*, *Beijing 100191*, *China*; *LI Jiao*, *Institute of Medical Information*, *Chinese Academy of Medical Sciences & Peking Union Medical College*, *Beijing 100020*, *China*; *TANG Buzhou*, *Peng Cheng Laboratory*, *Harbin Institute of Technology*, *Shenzhen 518055*, *China*

〔**Abstract**〕 The paper expounds the structure, construction method and application of the technology evaluation dataset for medical multimodal information extraction, including "gene - disease" oriented association semantic mining dataset, Chinese Medical Causal Dataset (CMedCausal), Medical Text to Medical Decision Tree Dataset (Text2DT), Medical Material OCR Feature Extraction Dataset (MedOCR), clinical diagnosis coding dataset. The above datasets are expected to provide strong support and reference for the evaluation and implementation of various technologies, algorithms and systems.

〔**Keywords**〕 China Conference on Health Information Processing (CHIP); multimodal information extraction; medical datasets; Artificial Intelligence (AI); Nature Language Processing (NLP)

1 引言

随着医院信息化的普及，医疗健康领域得到飞速发展，积累了海量且类型多样的医学数据，例如发表文献、医疗指南、医学教材、检验数据、影像图片、医疗发票、在线医典百科、扫描报告图像等[1-2]。这些数据以文本、表格、图像等多模态形式存在，是进行临床决策支持、诊疗路径解释、智慧医院建设的重要资源[3-4]。

第八届中国健康信息处理会议（China Conference on Health Information Processing，CHIP 2022）是中国中文信息学会（Chinese Information Processing Society of China，CIPS）医疗健康与生物信息处理专业委员会开展的以“信息处理技术助力探索生命之奥秘、提高健康之质量、提升医疗之水平”为主旨的年度会议。CHIP 是中国健康信息处理领域的重要会议，是世界各地学术界、企业界和政府部门的研究人员和从业人员分享创意，进一步推广领域研究成果和经验的重要平台。中国健康信息处理会议自 2018 年以来每年都组织技术评测[5-8]。本次 CHIP 2022 技术评测围绕疾病主题，探索信息数字化技术、基因关联信息、症状体征检查知识、诊疗决策树构建和诊断自动编码等研究内容，公布了 5 项任务：“面向‘基因 - 疾病’的关联语义挖掘”“医疗因果实体关系抽取”“医疗文本诊疗决策树抽取”“光学字符识别（Optical Character Recognition，OCR）医疗清单发票”和“临床诊断编码”。

本文从医疗多模态信息抽取的角度梳理上述数据集，希望能为研究者提供一套测试技术、算法和系统的高质量数据集，为中国健康信息处理相关研究提供参考。

2 医疗多模态信息抽取技术评测数据集介绍

2.1 面向“基因 - 疾病”的关联语义挖掘数据集

2.1.1 数据集构建情况 在海量科学文献中，基因与疾病的关联机理通过突变和各类生物分子对象及其触发词进行描述，自然语言处理技术为自动挖掘这一隐性知识提供了可能，也为健康医学信息的自动化处理提供了解决方案。为了从文献中挖掘基因与疾病的关联语义知识，研究者基于 PubMed 摘要文本构建了活跃基因注释语料库（Active Gene Annotation Corpus，AGAC）[9]。该数据集注释了 8 类触发词实体，涵盖从分子水平到细胞水平的生物学现象和过程。实体类型包括 5 类生物概念实体（突变、相互作用、通路、分子生理活性、细胞生理活性）和 3 类调控概念实体（正调控、负调控、调控）。此外，AGAC 还通过主事和致事两个语义关系来描述主题和因果关系，从而呈现句子的语义信息。AGAC 数据集主要包含 3 个特点，分别为数据不平衡、选择性注释和潜在主题注释。基于该数据

集，科研人员可以提取阿尔茨海默症关键基因，研究抗癫痫药物重定位，挖掘冠状病毒病理知识。

2.1.2 子任务分析 在 CHIP 2022 评测中，任务 1 包括 3 个子任务：触发词实体识别、语义角色识别、“基因 - 调控类型 - 疾病”三元组关系抽取。每个子任务的训练集包含 250 篇文献，测试集包含 2 000 篇文献。子任务 1 是传统意义下的命名实体识别任务，用以识别 12 类与“基因 - 疾病”有关的分子对象及其触发词实体，包括疾病（disease）、基因（gene）、蛋白质（protein）、酶（enzyme）、突变（variation）、分子活性（molecular physiological activity）、互作（interaction）、通路（pathway）、细胞活性（cell physiological activity）、调控（regulation）、正调控（positive regulation）、负调控（negative regulation）。子任务 2 是一个语义角色标注任务，语义角色包括 ThemeOf 和 CauseOf。该子任务捕捉实体之间的语义依赖关系，用以构建“基因 - 疾病”关联。子任务 3 是一个三元组抽取任务，针对“基因 - 疾病”的关联机理调控类型进行相关语义的抽取，可利用子任务 1 和子任务 2 所获得的触发词和语义角色，挖掘其背后的深层语义。调控类型包含 4 种对突变基因的语义描述，即功能丧失、功能获得、功能调节和功能的复合变化。

2.2 中文医疗因果关系抽取数据集

2.2.1 数据集应用价值 现代医疗强调解释性，医生在诊断、治疗和评估上都要以患者为中心，突出医疗的因果关系。互联网搜索引擎和线上问诊平台中含有大量医学问答知识和诊疗信息，通过文本挖掘技术和深度学习技术，从中抽取医疗因果关系，构建因果关系解释网络和医疗因果知识图谱，可以提升诊疗结果的逻辑性和可解释性，也能有效改善患者就医体验。而目前国内外尚无医学因果解释和推理方向的公开数据集。

2.2.2 数据集构建情况 研究者构建了首个中文医疗因果关系抽取数据集（Chinese Medical Causal Dataset，CMedCausal）。数据来源于线上问诊和医典百科，均为网上公开问诊数据，未涉及患者隐私信息。该数据集标注了文本中出现的医学概念片段和医学概念片段之间的关系。其中，医学概念片段即为临床发现，内容限定在以疾病为中心的文本，也包括实验室检验结果和检查结果。数据集定义了 3 类关键的医学因果解释推理关系：因果关系、条件关系和上下位关系。数据集标注人员包括 1 名医学专家、1 名人工智能算法专家和 8 名医学专业学生，标注工作通过阿里巴巴夸克内部的标注平台完成。该数据集由 9 153 段医学文本组成，总计 79 244 对实体关系。

2.3 医疗文本诊疗决策树抽取数据集

2.3.1 数据集应用价值 临床决策支持系统旨在辅助临床医务人员更加高效地做出临床诊疗。临床诊疗可以看作是一个根据不同条件进行判断，然后做出不同决策的过程。这种临床诊疗过程可以被建模为诊疗决策树，诊疗决策树是由条件节点和决策节点组成的树型结构，条件节点表示需要做出的条件判断，决策节点表示需要做出的诊疗决策。诊疗决策规则是指将给定条件与医疗决策联系起来，帮助医生、患者和其他利益相关者对特定临床问题做出适当的管理、选择和决定。这些决策规则可以建模为诊疗决策树。目前，诊疗决策树的构建往往依赖于医学专家的人工标注，这种方式耗时费力，且新知识难以及时融入临床决策支持系统[10]。通过智能化的信息抽取技术从庞大且快速积累的医学文本中精确提取诊疗决策树是一个可行的解决方案，但目前缺乏可用于模型构建的公开可用数据集。

2.3.2 数据集构建情况 针对上述问题，研究者构建了医疗文本诊疗决策树数据集（Medical Text to Medical Decision Tree Dataset，Text2DT），用于从医疗文本中抽取诊疗决策树任务。Text2DT 数据集来源于权威医疗机构出版的临床实践指南和人民卫生出版社出版的临床医学教科书。数据集标注人员包括 2 名医学专家和 6 名相关领域研究人员。Text2DT 数据集包含 400 例文本 - 决策树对。三元组是诊疗决策树的主要组成部分，共有 6 种关系，即临床表现、治疗药物、治疗方案、用法用量、基本情况、禁用药物。诊疗决策树的深度从 2 层到 5 层。一般而言，在三元组抽取完成后，需要进一步生成树结

构，从而将信息串联形成一个完整的决策流程。Text2DT 的任务目标是从给定的医疗文本抽取出诊疗决策树。诊疗决策树表示简化的决策过程，即根据条件判断的不同结果做出下一个条件判断或决策。

2.4 医疗材料 OCR 要素提取数据集

2.4.1 数据集应用价值 在医疗和保险行业存在大量纸质文档形式的医疗数据，如就诊病历、缴费发票等。这些数据中含有丰富的信息，具有很高的商业和科研价值。目前这些数据通过业务人员手动录入的方式进行登记。光学字符识别和自然语言处理等人工智能技术的发展及其在生产生活中各种相关应用的普及，为医疗纸质材料的信息自动化抽取提供了新的思路。这种智能化的解决方案一般包括两个步骤，首先通过计算机视觉领域的目标检测和目标识别等算法将纸质扫描材料进行文本化；然后通过自然语言处理领域的信息抽取算法将这些文本信息结构化。此外，与传统方法不同，这些基于人工智能技术的新颖解决方案需要充足的标注数据进行模型训练，而缺乏高质量的标注数据是相关研究发展的最大障碍。

2.4.2 数据集情况介绍 医疗材料 OCR 要素提取数据集（Medical Material OCR Feature Extraction Dataset，MedOCR）是当前最新的数据集，共包括 1 700 张医疗材料图片。其中出院小结 340 张、购药发票 340 张、门诊发票 340 张、住院发票 680 张。数据集的原始数据来源于互联网，并经过了严格的人工审核，为每类数据都定义了特定提取属性。出院小结包含 8 个属性，购药发票包含 8 个属性，门诊发票包含 34 个属性，住院发票包含 37 个属性。MedOCR 数据集采用准确率作为评测指标，只有属性的预测值和标注值完全一致才判定为识别正确。这些来自于真实生活场景中的医疗材料图片质量不一、颜色清晰度各异，且含有各种干扰信息，对当前大多数模型都具有挑战性。该数据集样本量充足、类型多样，有望推动医疗信息处理领域针对图片文档进行信息抽取研究的发展。

2.5 临床诊断编码数据集

2.5.1 疾病分类与手术操作分类编码发展情况 疾病分类与手术操作分类编码是对患者疾病诊断和治疗信息的加工过程，是病案信息管理的重要环节。病案编码已成为医院科学化、信息化管理的重要依据之一，在评估医疗质量与医疗效率、设计临床路径方案、重点学科评价、医院评审、疾病诊断分级、传染病报告、医疗付款、合理用药监测等方面的应用越来越广泛、越来越深入。在诸多分类方案中，国际上最有影响力且最为普及的是国际疾病分类（International Classification of Diseases，ICD）。ICD 是世界卫生组织制定的国际统一的疾病分类方法，是目前国际上通用的疾病分类方法。中国也推出了《疾病分类与代码国家临床版 2.0》和《手术操作分类代码国家临床版 2.0》，并在部分医院中得到了应用。

2.5.2 数据集构建情况 在 CHIP 2022 评测中发布的临床诊断编码任务数据集，主要目标是针对中文电子病历进行诊断编码。给定一次就诊的相关诊断信息（包括入院诊断、术前诊断、术后诊断、出院诊断），以及手术名称、药品名称、医嘱名称，要求给出其对应的国家临床版 2.0 标准词。该数据集中所有就诊数据均来自于真实医疗数据，并以《疾病分类与代码国家临床版 2.0》词表为标准进行标注。其中训练数据 2 700 条，测试数据 337 条。数据集以准确率作为最终评估指标。

3 结语

医疗信息化的发展催生了海量且类型多样的多模态数据。本文介绍了中国健康信息处理会议评测任务发布的 5 项数据集，包括基于“基因－疾病”的关联语义挖掘数据集、中文医疗因果关系抽取数据集、医疗文本诊疗决策树抽取数据集、医疗材料 OCR 要素提取数据集、临床诊断编码数据集。这些数据集有望为各种技术、算法以及系统的评估和实施提供有力的支撑和参考。未来将继续补充类型更

（下转第 22 页）

策树中的关系并利用解码算法获得最终的决策树。实验表明，本文提出的方法与同类方法相比有明显改进，为未来诊疗决策树的自动抽取与大型临床决策支持系统自动化构建奠定了基础。

参考文献

1 李文念. 临床辅助诊断决策支持系统的原型研究 [D]. 北京：中国科学技术信息研究所，2013.

2 曹晓均. 基于规则库的临床辅助决策系统实践 [J]. 中国数字医学，2021，16（9）：16－20.

3 Saibene A，Assale M，Giltri M. Expert Systems：Definitions，Advantages and Issues in Medical Field Applications [J]. Expert Systems with Applications，2021（177）：114900.

4 杨宇辉，李素姣，喻洪流，等. 临床决策支持系统研究进展 [J]. 生物医学工程学进展，2021，42（4）：203－207.

5 Wei Z，Su J，Wang Y，et al. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction [C]. Seattle：Association for Computational Linguistics，2020.

6 Wang Y，Yu B，Zhang Y，et al. TPLinker：Single－stage Joint Extraction of Entities and Relations Through Token Pair Linking [C]. Barcelona：International Committee on Computational Linguistics，2020.

7 Wang Y，Sun C，Wu Y，et al. UniRE：a Unified Label Space for Entity Relation Extraction [C]. Online：Association for Computational Linguistics，2021.

8 Dozat T，Manning C D. Deep Biaffine Attention for Neural Dependency Parsing [C]. Toulon：OpenReview，2017.

9 Lei W，Yan W，Deng C，et al. Translating a Math Word Problem to an Expression Tree [C]. Brussels：Association for Computational Linguistics，2018.

10 Xie Z，Sun S. A Goal－Driven Tree－Structured Neural Model for Math Word Problems [C]. Macao：The 28th International Joint Conference on Artificial Intelligence 2019（IJCAI），2019.

11 鹏城实验室. PCL－MedBERT [EB/OL]. [2022－10－13]. https：//code. ihub. org. cn/projects/1775.

12 Cui Y，Che W，Liu T，et al. Pre－training with Whole Word Masking for Chinese Bert [J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2021（29）：3504－3514.

13 Liu Z，Lin W，Shi Y，et al. A Robustly Optimized BERT Pre－training Approach with Post－training [C]. Huhhot：Chinese Information Processing Society of China，2021.

（上接第5 页）

加丰富的数据，如影像数据、组学数据等，使医疗健康多模态大数据在真实世界研究中发挥应用价值。

参考文献

1 Esteva A，Robicquet A，Ramsundar B，et al. A Guide to Deep Learning in Healthcare [J]. Nature Medicine，2019，25（1）：24－29.

2 刘梦迪. 医疗大数据平台建设面临的困境及应对策略 [J]. 电脑知识与技术，2022，18（15）：19－21.

3 李赞梅，钱庆，李姣，等. 健康医疗科学数据共享标准体系框架构建 [J]. 医学信息学杂志，2018，39（11）：49－53.

4 张弘政，刘迷迷，李琳，等. 基于通用数据模型的健康医疗大数据平台数据治理研究 [J]. 医学信息学杂志，2022，43（6）：2－7，13.

5 黄源航，焦晓康，汤步洲，等. CHIP 2019 评测任务1 概述：临床术语标准化任务 [J]. 中文信息学报，2021，35（3）：94－99.

6 骆迅，倪渊，汤步洲，等. 基于竞赛视角探讨文本语义匹配技术在中文医学文本领域中的应用 [J]. 中国数字医学，2021，16（11）：99－103.

7 宗辉，张泽宇，杨金璇，等. 基于人工智能的中文临床试验筛选标准文本分类研究 [J]. 生物医学工程学杂志，2021，38（1）：105－110，121.

8 李雯昕，张坤丽，关同峰，等. CHIP 2020 评测任务1 概述：中文医学文本命名实体识别 [J]. 中文信息学报，2022，36（4）：66－72.

9 Wang Y，Zhou K，Kim J D，et al. An Active Gene Annotation Corpus and Its Application on Anti－epilepsy Drug Discovery [C]. San Diego：2019 IEEE International Conference on Bioinformatics and Biomedicine（BIBM），2019.

10 Saibene A，Assale M，Giltri M. Expert Systems：Definitions，Advantages and Issues in Medical Field Applications [J]. Expert Systems with Applications，2021（177）：114900.