·综述·

面向真实世界的知识挖掘与知识图谱补全研究（二）：非结构化电子病历信息抽取方法及进展

阎思宇[1]，李绪辉[1]，陈沐坤[2]，朱海锋[2]，谭杰骏[2]，高　旷[2]，王永博[1]，黄　桥[1]，任相颖[1]，靳英辉[1]，王行环[1]

1. 武汉大学中南医院循证与转化医学中心（武汉 430071）
2. 武汉大学计算机学院（武汉 430072）

【摘要】随着信息技术的普及和推广，健康医疗大数据呈指数级增长，基于健康医疗大数据的临床真实世界研究日益受到关注。医院电子病历记录了真实世界下患者的诊疗全过程，是最能为临床决策提供支持的数据源之一。但电子病历数据中大量非结构化文本数据的存在，增加了数据处理难度，制约了基于电子病历数据研究的开展。急需将信息技术、人工智能等先进的方法用于非结构化电子病历数据的处理，以加速数据价值转化。本文总结了当前非结构化医学数据处理的常用方法，包括基于词典和规则的方法、基于传统机器学习和深度学习的方法和以本体为代表的基于认知模型的方法，探讨了非结构化电子病历数据处理时的标准化问题及透明化报告问题，展望了相关发展。

【关键词】非结构化数据；电子病历；信息抽取；文本挖掘；自然语言处理；本体；真实世界数据

Research on real-world knowledge mining and knowledge graph completion (II): Methods and progress of information extraction from unstructured electronic medical records

Si-Yu YAN[1], Xu-Hui LI[1], Mu-Kun CHEN[2], Hai-Feng ZHU[2], Jie-Jun TAN[2], Kuang GAO[2], Yong-Bo WANG[1], Qiao HUANG[1], Xiang-Ying REN[1], Ying-Hui JIN[1], Xing-Huan WANG[1]

1. Center for Evidence-Based and Translational Medicine, Zhongnan Hospital of Wuhan University, Wuhan 430071, China
2. School of Computer Science, Wuhan University, Wuhan 430072, China
Corresponding authour: Ying-Hui JIN, Email: jinyinghui0301@163.com; Xing-Huan WANG, Email: wangxinghuan1965@163.com

【Abstract】With the popularization and promotion of information technology, healthcare big data is growing exponentially, and clinical real-world research based on healthcare big data is receiving increasing attention. The hospital electronic medical record (EMR) records the whole process of diagnosis and treatment of patients in the "real-world", and

DOI: 10.12173/j.issn.1004-5511.202301016
基金项目：国家自然科学基金面上项目（82174230）
通信作者：靳英辉，博士，副教授，硕士研究生导师，Email：jinyinghui0301@163.com
王行环，博士，教授，博士研究生导师，Email: wangxinghuan1965@163.com

is one of the most supportive data sources for clinical decision-making. However, the existence of a large number of unstructured text data in EMR data increases the difficulty of data processing and restricts the development of research based on EMR data. Advanced methods such as information technology and artificial intelligence need to be applied to the processing of unstructured EMR data to accelerate the transformation of data value. This paper summarizes the current common methods of unstructured medical data processing, including methods based on dictionaries and rules, methods based on traditional machine learning and deep learning, and methods based on cognitive models represented by ontology, and also discusses the problems of standardization and transparent reporting when processing unstructured EMR data and looks forward to the relevant development.

【Keywords】Unstructured data; Electronic medical record; Information extraction; Text mining; Natural language processing; Ontology; Real-world data

2015 年我国相继出台了《关于积极推进“互联网 +”行动的指导意见》和《促进大数据发展行动纲要》，2016 年国务院办公厅印发了《关于促进和规范健康医疗大数据应用发展的指导意见》，指出“健康医疗大数据是国家重要的基础性战略资源”[1]。健康医疗大数据已被提升至国家战略高度。在 2022 年中国医学发展大会上沈洪兵院士同样提到“要关注基于健康医疗大数据的临床真实世界研究，注重与信息技术、人工智能交叉融合”。随着健康医疗大数据的指数级增长，如何对健康医疗大数据进行充分挖掘和分析，提炼数据价值，已成为当今的研究趋势。真实世界研究并非方法学上新的研究类型，而是基于真实世界数据（real-world data, RWD）进行的研究，具有外部有效性高、数据来源广泛、易获取等优点，日益受到研究者的青睐[2-3]。

医院电子病历（electronic medical record，EMR）主要用于日常医疗实践管理，记录有真实世界下患者详细的就诊数据，是健康医疗大数据及 RWD 的重要来源之一。中国已有超过九成的医院在应用 EMR[4]。随着 EMR 的普及和诊疗数据的不断积累，虽然数据量一直在增长，但如何基于 EMR 数据生成高质量真实世界证据的困境一直存在。已有研究指出，医疗保健领域的最大问题是大约 80% 的医疗数据在创建后仍然是非结构化和未开发的（例如，文本、图像、信号等）[5-6]。为了便于医生灵活描述，EMR 中很大比例的信息是使用自由文本记录的非结构化数据，如病程记录、病理报告、影像学报告、手术记录、出院记录等。虽然 EMR 数据量大，但其中非结构化数据占比高，这让计算机难以理解，因此基于 EMR 数据的研究依然有限[7]。

机器学习、人工智能（artificial intelligence，AI）和其他现代统计方法正为利用先前尚未开发且极速增长的数据资源提供新的机会，以期让患者获益[8]。利用计算机算法从医疗健康数据中获取信息，以补充知识发现、促进循证医学、协助制定临床决策，已成为当前研究的热点[9]。

针对上述电子病历数据中非结构化数据普遍存在且处理困难的问题，本文将对现有的技术方法及新进展进行总结，以提供参考。

1 非结构化医学数据信息抽取的研究方法

从非结构化数据中提取结构化信息通常属于信息抽取（information extraction，IE）、文本挖掘（text mining，TM）或自然语言处理（natural language processing，NLP）领域的内容。一系列研究已经证明了从临床叙述性文本中提取结构化信息的可行性。一项纳入 263 篇有关 IE 在临床应用研究的综述显示，IE 可用于肿瘤、循环系统疾病等多个疾病研究领域，药物提取、药物不良反应等药物相关研究以及质量管理、不良事件等临床工作流程优化研究，所使用的非结构化数据主要包括出入院记录、手术记录等的临床记录和影像学报告、病理报告等的诊断报告[7]。

综合多项研究[10-12]，该领域的研究方法大体上可分为基于词典或规则的方法、基于统计的方法和基于认知模型的方法，这些方法也可混合使用以提升性能。每种方法的释义及在临床中的应

用举例总结如下，见图 1。

1.1 基于词典和规则的方法

（1）基于词典的方法：是指依靠术语词典，采取匹配算法进行简单文本匹配，该方法较为基本和直接，具有较强可解释性。适合于简单任务，如识别特定药物，但不适合复杂任务。早期医疗领域的多种代表性实体识别工具如 MedLEE、IBM 的 MedKAT 和 Mayo Clinic 的 cTAKES 都是采用基于词典的方法 [13]。朱彦等的研究通过建立中医药领域专业词典，来解决方剂数据自动结构化的问题 [14]。

（2）基于规则 / 模式匹配的方法：通常使用正则表达式技术，构建目标提取字段的模式（pattern），完成匹配和搜索。例如，使用基于模式匹配（pattern matching）的 NLP 算法解析非结构化电子健康记录数据，以识别研究人群中的老年综合征病例 [15]；使用正则表达式从前列腺癌病理学报告中提取 Gleason 评分 [16]；基于标注结果抽取模板，生成正则表达式，抽取中文电子病历中的糖尿病病史 [17]；基于规则的模式匹配方法对乳腺癌患者的病理报告进行信息抽取 [18]；使用正则表达式构建规则完成中医古籍中“崩漏”疾病相关的知识抽取 [19]。

基于词典或规则的方法依赖于手工建立的词典、抽取模式或规则，规则融合了领域知识和语言知识，领域相关性较高但可移植性较差。基于词典或规则的方法不涉及太复杂的计算机算法，对临床医生来说可解释性高，适用于较为简单、规范的非结构化文本的信息抽取任务，也适合于医学知识丰富、但无法掌握复杂计算机算法的临床医生。因此，临床 NLP 一直以基于规则的方法为主。一项有关临床信息抽取应用研究的综述显示，使用基于规则的方法进行信息抽取的文献在纳入的 263 篇文献中占比达 65%[7]；另一项有关临床概念提取的方法学综述显示，使用基于规则的方法进行信息抽取的文献在纳入的 228 个文献中占比达 48%[20]。但并非所有的自然语言都可以用确定性的规则来刻画，且捕获所有可能的变化需要大量的规则，规则的维护和更新也比较困难，因此学术 NLP 领域仍以基于统计的方法为主导。

1.2 基于统计的方法

基于统计的方法是通过构造模型进行信息抽取，可分为传统机器学习（machine learning，ML）方法和更先进的深度学习（deep learning，DL）算法。

（1）传统机器学习方法：按是否有标记的训练数据可分为无监督和有监督的 ML 方法。无监督机器学习方法指使用无任何标记数据的统计模型，最经典的方法为聚类，利用的是非结构化数据中上下文的相似性。如使用无监督机器学习方法从乳腺 X 线影像报告中自动提取信息 [21]；使用自动化手术术语聚类进行手术文本数据的预处理 [22]。有监督机器学习方法指使用标记的训练数据来训练模型，常用模型包括支持向量机、条件随机场模型、隐马尔可夫模型、决策树等。如使用支持向量机方法进行药物不良反应检测 [23]；使用条件随机场方法从急诊患者记录中提取儿科阑尾炎评分 [24]。

（2）基于深度学习的方法：常用模型包括卷积神经网络、循环神经网络、长短期记忆网络、Word2Vec 模型、基于变换器的双向编码器表示技术（bidirectional encoder representation from transformers，BERT）等。如使用多任务深度神经网络 [25]、卷积神经网络从癌症病理报告中自动提取信息 [26]；使用 BERT 等算法提取公开临床语料库中的临床概念 [27]。

基于机器学习和深度学习的方法是学术 NLP 领域的主流，但对于临床医生来说，较难掌握其复杂的算法。有学者指出信息抽取技术在临床 EMR 数据中未得到充分利用的原因之一就是 NLP 专家与临床医生缺乏密切合作 [7]，EMR 非结构化数据的提取工作需要多学科团队的参与。例如 2016 年美国启动的全球首个“癌症先进计算解决方案的联合设计”（Joint Design of Advanced Computing Solutions for Cancer，JDACS4C）项目，即为国家癌症研究所与美国能源部的跨机构合作，旨在借助计算、数据科学的深度学习技术加快抗癌研究，其中试点 3 项目就重点针对癌症患者病历数据的自动分析 [28]。

此外，ML、DL 算法在中文医学文本挖掘领域应用的另一制约因素是国内标注数据的稀缺性。医学领域目前没有像一般语料那样丰富的标记数据，尤其是电子病历数据，如何在共享中保护患者的隐私是需要考虑的问题。另一方面，医疗数据包含复杂、多样的医学知识，标注难度较大。虽然国内已有学者标注了部分临床文本，但尚无完整、公开共享的已标注的电子病历数据集 [29]。

因此，在使用 ML、DL 算法开展 EMR 非结构化数据信息抽取时，仍需要投入大量的时间、精力进行数据标注，这对时间宝贵的临床医生来说是个不小的挑战。

1.3 基于认知模型的方法

因语言理解具有明显的认知过程，所以除了上述方法外，基于认知科学的信息抽取研究也越来越多，常见的为基于本体的方法 [10]。学者们较为认可的本体（ontology）的定义是德国学者 Studer 等于 1998 年提出的“本体是共享概念模型的明确的形式化规范说明”[30]。本体可用来描述特定领域的知识，借助本体进行文本挖掘，相当于给挖掘过程“配备”了一名“领域专家”，指导整个挖掘过程 [31]，可增强对语义内容的理解、推理。由于本体具有能通过概念之间的关系来表达概念语义的能力，所以将本体应用于 NLP 领域，能够提高系统的召回率和准确率，优化提取结果。领域本体与信息抽取的结合，是当前的研究热点 [32]。基于本体的信息抽取一般是先建立领域本体，进而根据本体描述的概念、关系、层次结构和概念与关系间的约束等生成抽取规则，然后再根据规则对文档进行抽取 [10]。

为了适应特定的临床问题，通常将知识驱动的视角（如生物医学本体）与模型相结合，以定制模型 [20]。例如将 Word2vec 模型与心血管疾病本体相结合，提供定制解决方案，从生物医学文献中提取更相关的心血管疾病相关术语 [33]。Feichen 等的研究提供了一种基于不同知识存储库选择的人类表型本体生成自定义节点嵌入的方法，以便通过分析临床叙述中的患者表型表征来加速罕见病鉴别诊断 [34]。Popejoy 等的研究描述了一种护理协调本体，该本体旨在从护理笔记中识别和提取护理协调活动，并展示了如何量化这些活动 [35]。

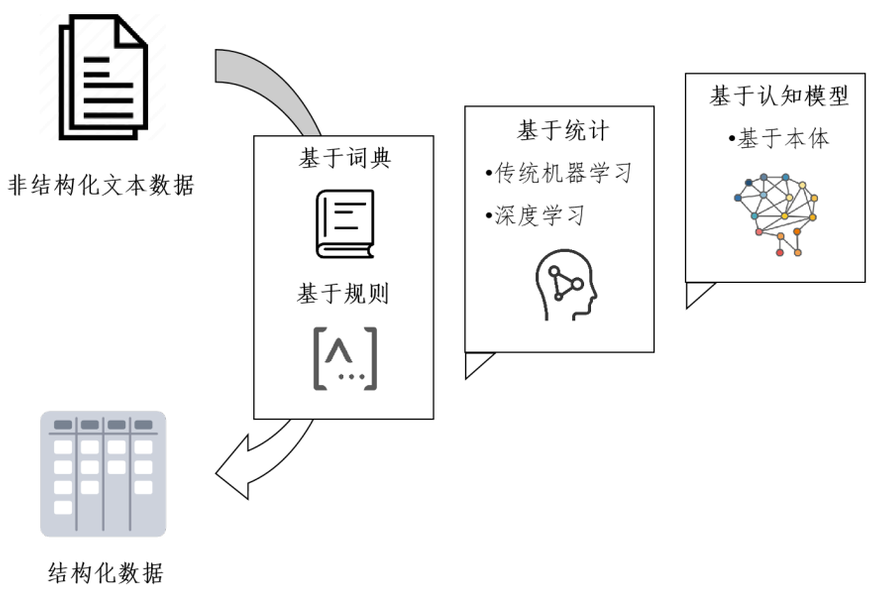

图1 从非结构化文本数据中提取结构化数据的方法学发展

Figure 1. Methodology development for extracting structured data from unstructured text data

2 非结构化电子病历数据处理时的标准化问题

使用 EMR 数据进行临床研究时，EMR 数据应满足临床研究数据的质量标准，如赖俊恺等的研究借鉴使用了临床数据交换标准协会（The Clinical Data Interchange Standards Consortium，CDISC）标准用于 EMR 数据到临床研究数据的标准转化，应用了 AI 领域的 NLP 技术，开发了临床研究中非结构化文本数据的电子来源（eSource）模式，根据 CDISC 标准填写病例报告表，以满足数据收集中的监管和可追溯性要求 [36]。

汇集不同来源的 EMR 数据时，需要构建一致性标准，以实现共享，具体包括对数据项类型、属性等的定义，或进行术语映射。如将提取到的标签与 CDISC ODM 特定术语表、国际疾病分类（ICD-10）进行映射，建立研究专用术语库 [36]。但 ICD-10 作为标准术语仍比较粗糙，无法完全匹配需抽取的临床术语。国外有比较成熟、广泛应用的医学术语系统、标准或本体，如医学系

统命名法—临床术语（systematized nomenclature of medicine-clinical terms，SNOMED CT）、统一医学语言系统（unified medical language system，UMLS），用于临床用语的规范化表达。这些术语集可以协调一致地在不同的学科、专业和机构之间实现对于临床数据的标引、存储、检索和聚合，便于计算机处理。它与 EMR 系统的结合可以实现在不同 EMR 系统之间协调一致地交换临床信息，方便数据挖掘与决策分析。如，英国制订的服务于电子病历管理的国民健康信息基础架构（National Health Information Infrastructure，NHII）就参考使用了 SNOMED CT 等一系列的术语标准[37]。而国内只在 2002 年由原卫生部授权对全国住院病人的诊断数据编码使用 ICD 标准，但尚未应用 SNOMED CT、UMLS 等普遍被认可的术语系统[4]。此外，这些术语系统基本基于英语语言开发，中英文的转换工作及中文医学术语的开发国内有学者团队正在进行。如，中国中医科学院中医药信息研究所开发了与 UMLS 对应的中医药语言系统（Traditional Chinese Medicine Language System，TCMLS），与 SNOMED CT 对应的中医临床术语系统（Traditional Chinese Medicine Clinical Terminology System，TCMCTS），与 MeSH 医学主题词表对应的中医药学主题词表（Chinese Medical Subject Headings，CMeSH）。但这些中文术语系统的推广使用目前还比较有限，尚未发现其与 EMR 系统结合的实践。非结构化中文医学文本挖掘的标准化工作需要国内更多可用的、细粒度的中文标准医学术语的开发和完善来推动。

3 非结构化电子病历数据处理的透明化报告问题

如何公开、透明地报告 RWD 数据治理过程，尤其是非结构化数据的处理，也是提高真实世界研究可信度的重要议题。2019 年哈佛医学院 Shirley V. WANG 团队发表了《使用非结构化电子健康数据开展真实世界研究比较效果和安全性研究的报告规范》[38]，列出了使用 NLP 和 ML 算法进行非结构化数据的数据挖掘相关研究中应公开报告的 9 项内容，如提供 NLP 和 ML 算法的完整描述，包括软件包的名称和版本、带有用于映射临床概念的本体引文或附录、算法中包含的输入和调整参数、输出的详细信息等，以规范相关算法研究的开展和报告。在进行非结构化电子病历数据处理时，各学者应进行过程的透明化报告，确保非结构化数据中提取变量的准确性和可复现性，以提升真实世界研究的质量。

4 结语

EMR 是开展真实世界研究的重要数据来源之一，但是由于其主要产生于日常医疗实践管理而非科研，其数据呈现多源异构的特点。大量非结构化数据的存在，增加了数据处理难度，显著制约了 RWD 向真实世界证据的转化效率。因此，有必要对现有非结构化电子病历数据标准化技术方法进行系统总结和分析。

处理非结构化电子病历数据可借助多种信息抽取或 NLP 技术，包括基于词典或规则的方法，基于传统机器学习或深度学习的方法，以及最近越来越热门的基于本体的方法，或者多种方法的融合使用。基于词典或规则的方法依赖于专家知识手工建立词典或规则，不涉及复杂的计算机算法，适合于较为简单、规范的非结构化数据处理任务，在临床 NLP 中应用广泛，但可移植性较差。基于机器学习或深度学习的方法是学术 NLP 的主流方法，大部分需要有已标注的训练数据、选择及训练模型，对计算机算法的掌握水平要求较高，因此应积极创建临床医生和计算机专家合作的环境，促进跨学科的交流，加速医疗数据合作挖掘，同时也应积极推动中文电子病历语料库的建设，在保护患者隐私的同时积极探索资源的共享模式。在信息的使用、重用、共享和互操作性方面，本体已经成功地应用于生成和提供领域知识。基于本体的方法，以本体知识为支撑，整合其他信息抽取技术，借助本体对领域共享概念的知识表达和推理能力，可优化提取结果，促进结果的标准化、共享、重用和互操作性，为进一步的数据融合打下基础。但当前中文医学本体、医学知识图谱的语义资源还非常稀缺，尤其缺乏细粒度的医学本体，因此需要加速中文医学本体的发展，以促进和带动基于本体的医学信息抽取的发展，最终助力健康医疗大数据的价值转化。

参考文献

1 国务院办公厅 . 国务院办公厅关于促进和规范健康医

疗大数据应用发展的指导意见（国办发〔2016〕47号）[EB/OL]. (2016-06-24) [2022-12-02]. http://www.gov.cn/zhengce/content/2016-06/24/content_5085091.htm.

2 施秀青, 阎思宇, 黄桥, 等. 真实世界研究：弥合临床实践指南与临床决策之间的距离 [J]. 协和医学杂志, 2023, 14(4), 859-867. [Shi XQ, Yan SY, Huang Q, et, al. Real world research: helping clinical practice guidelines span the distance between itself and clinical decision making[J]. Medical Journal of Peking Union Medical College Hospital, 2023, 14(4), 859-867.] DOI: 10.12290/xhyxzz.2022-0217.

3 CD Mack, L Parmenter, E Brinkley, et al. 利用补充真实世界数据研究获得更深层次的认识 [J]. 药物流行病学杂志, 2016, 25(1): 27-37. [CD Mack, L Parmenter, E Brinkley, et al. Using enriched real-world research for deeper insights[J]. Chinese Journal of Pharmacoepidemiology, 2016, 25(1): 27-37.] DOI: 10.19960/j.cnki.issn1005-0698.2016.01.009.

4 Zhang L, Wang H, Li Q, et al. Big data and medical research in China[J]. BMJ, 2018, 360: j5910. DOI: 10.1136/bmj.j5910.

5 Consultant AH. Why unstructured data holds the key to intelligent healthcare systems[EB/OL]. (2015-3-31) [2022-12-02]. https://hitconsultant.net/2015/03/31/tapping-unstructured-data-healthcares-biggest-hurdle-realized/.

6 Kong HJ. Managing unstructured big data in healthcare system[J]. Healthc Inform Res, 2019, 25(1):1-2. DOI: 10.4258/hir.2019.25.1.1.

7 Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review[J]. J Biomed Inform, 2018, 77: 34-49. DOI: 10.1016/j.jbi.2017.11.011.

8 Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness[J]. BMJ, 2020, 368: l6927. DOI: 10.1136/bmj.l6927.

9 He J, Baxter SL, Xu J, et al. The practical implementation of artificial intelligence technologies in medicine[J]. Nat Med, 2019, 25(1): 30-36. DOI: 10.1038/s41591-018-0307-0.

10 程显毅, 朱倩, 王进. 中文信息抽取原理及应用 [M]. 北京：科学出版社, 2010. [Cheng XY, Zhu Q, Wang J. Principle and application of chinese information extraction[M]. Beijing: China Science Publishing & Media, 2010.]

11 Malmasi S, Hosomura N, Chang LS, et al. Extracting healthcare quality information from unstructured data[J]. AMIA Annu Symp Proc, 2018, 2017: 1243-1252. https://pubmed.ncbi.nlm.nih.gov/29854193/.

12 吴宗友, 白昆龙, 杨林蕊, 等. 电子病历文本挖掘研究综述 [J]. 计算机研究与发展, 2021, 58(3): 513-527. [Wu ZY, Bai KL, Yang LR, et, al. Review on text mining of electronic medical record[J]. Journal of Computer Research and Development, 2021, 58(3): 513-527.] DOI: 10.7544/issn1000-1239.2021.20200402.

13 崔博文, 金涛, 王建民. 自由文本电子病历信息抽取综述 [J]. 计算机应用, 2021, 41(4): 1055-1063. [Cui BW, Jin T, Wang JM. Overview of information extraction of free-text electronic medical records[J]. Journal of Computer Applications, 2021, 41(4): 1055-1063.] DOI: 10.11772/j.issn.1001-9081.2020060796.

14 朱彦, 朱玲, 王俊慧, 等. 基于信息抽取的历代方剂药物知识发现方法及应用 [J]. 中华中医药杂志, 2015, 30(5): 1447-1451. [Zhu Y, Zhu L, Wang JH, et, al. An efficient approach of acquiring knowledge from ancient prescriptions and medicines based on information extraction[J]. China Journal of Traditional Chinese Medicine and Pharmacy, 2015, 30(5): 1447-1451.] DOI: CNKI:SUN:BXYY.0.2015-05-017.

15 Anzaldi LJ, Davison A, Boyd CM, et al. Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study[J]. BMC Geriatr, 2017, 17(1): 248. DOI: 10.1186/s12877-017-0645-7.

16 Miettinen J, Tanskanen T, Degerlund H, et al. Accurate pattern-based extraction of complex Gleason score expressions from pathology reports[J]. J Biomed Inform, 2021, 120: 103850. DOI: 10.1016/j.jbi.2021.103850.

17 包小源, 黄婉晶, 张凯, 等. 非结构化电子病历中信息抽取的定制化方法 [J]. 北京大学学报(医学版), 2018, 50(2): 256-263. [Bao XY, Huang WJ, Zhang K, et, al. A customized method for information extraction from unstructured text data in the electronic medical

records[J]. Journal of Peking University(Health Sciences), 2018, 50(2): 256–263.] DOI: 10.3969/j.issn.1671–167X.2018.02.010.

18 吴欢，应俊，王逸飞，等. 乳腺癌病理文本的结构化信息提取[J]. 解放军医学院学报，2020, 41(7): 746–751. [Wu H, Ying J, Wang YF, et, al. Structured information extraction from breast cancer pathological report texts[J]. Academic Journal of Chinese Pla Medical School, 2020, 41(7): 746–751.] DOI: 10.3969/j.issn.2095–5227.2020.07.022.

19 朱玲，朱彦，杨峰. 基于中医疾病相关语义关系的正则表达式及知识抽取研究[J]. 世界科学技术 – 中医药现代化，2016, 18(8): 1241–1250. [Zhu L, Zhu Y, Yang F. Knowledge extraction research for semantic expression of diseases in chinese medicine[J]. World Science and Technology–Modernization of Traditional Chinese Medicine, 2016, 18(8):1241–1250.] DOI: 10.11842/wst.2016.08.004.

20 Fu S, Chen D, He H, et al. Clinical concept extraction: A methodology review[J]. J Biomed Inform, 2020, 109: 103526. DOI: 10.1016/j.jbi.2020.103526.

21 Gupta A, Banerjee I, Rubin DL. Automatic information extraction from unstructured mammography reports using distributed semantics[J]. J Biomed Inform, 2018, 78: 78–86. DOI: 10.1016/j.jbi.2017.12.016.

22 Khaleghi T, Murat A, Arslanturk S, et al. Automated surgical term clustering: a text mining approach for unstructured textual surgery descriptions[J]. IEEE J Biomed Health Inform, 2020, 24(7): 2107–2118. DOI: 10.1109/JBHI.2019.2956973.

23 Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi–corpus training[J]. Biomed Inform, 2015, 53: 196–207. DOI: 10.1016/j.jbi.2014.11.002.

24 Deleger L, Brodzinski H, Zhai H, et al. Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department[J]. J Am Med Inform Assoc, 2013, 20(e2): e212–20. DOI: 10.1136/amiajnl–2013–001962.

25 Yoon HJ, Ramanathan A, Tourassi G. Multi–task deep neural networks for automated extraction of primary site and laterality information from cancer pathology reports[C]. Advances in Big Data: Proceedings of the 2nd INNS Conference on Big Data, 2016, Thessaloniki, Greece 2. Springer International Publishing, 2017: 195–204.

26 Alawad M, Yoon H J, Tourassi G D. Coarse–to–fine multi–task training of convolutional neural networks for automated information extraction from cancer pathology reports[C]. 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). IEEE, 2018: 218–221.

27 Yang X, Bian J, Hogan WR, et al. Clinical concept extraction using transformers[J]. J Am Med Inform Assoc, 2020, 27(12): 1935–1942. DOI: 10.1093/jamia/ocaa189.

28 Health NCIatNIo. Joint design of advanced computing solutions for cancer (JDACS4C). https://datascience.cancer.gov/collaborations/joint–design–advanced–computing.

29 张伟，张展鹏，张明淘，等. 医疗健康知识挖掘中的语义资源、数据集和工具[J]. 计算机技术与发展，2022, 32(4):21–27. [Zhang W, Zhang ZP, Zhang MT, et al. Semantic resource,dataset and tool for medical health knowledge mining[J]. Computer Technology and Development, 2022, 32(4):21–27.] DOI: 10.3969/j.issn.1673–629X.2022.04.004.

30 Studer R, Benjamins VR, Fensel D. Knowledge engineering: principles and methods[J]. Data & Knowledge Engineering, 1998, 25(1): 161–197. DOI: 10.1016/S0169–023X(97)00056–6.

31 姜丽华，张宏斌，杨晓蓉. 基于领域本体的文本挖掘研究[J]. 情报科学，2014, 32(12). [Jiang LH, Zhang HB, Yang XR. Research on text mining based on domain ontology[J]. Information Science,2014, 32(12). DOI: 10.13833/j.cnki.is.2014.12.024.

32 阳广元. 国内基于本体的信息抽取研究现状与热点分析[J]. 图书馆理论与实践，2017(5): 38–43. [Yang GY. Research Status and Hotspot Analysis of Ontology Based Information Extraction in China[J]. Library Theory and Practice, DOI: 10.14064/j.cnki.issn1005–8214.2017.05.008.

33 Arguello Casteleiro M, Demetriou G, Read W, et al. Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature[J]. J Biomed Semantics, 2018, 9(1):13. DOI: 10.1186/s13326–018–0181–1.

34 Shen F, Peng S, Fan Y, et al. HPO2Vec+: Leveraging heterogeneous knowledge resources to enrich node

embeddings for the Human Phenotype Ontology[J]. J Biomed Inform, 2019, 96: 103246. DOI: 10.1016/j.jbi.2019.103246.

35 Popejoy LL, Khalilia MA, Popescu M, et al. Quantifying care coordination using natural language processing and domain-specific ontology[J]. J Am Med Inform Assoc, 2015, 22(e1): e93-103. DOI: 10.1136/amiajnl-2014-002702.

36 赖俊恺, 王斌, 姚晨, 等. 从真实世界数据到临床研究数据的标准转化研究[J]. 中国食品药品监管, 2021, (11): 39-46. [Lai JK, Wang B, Yao C, et al. Research on standards transformation from real-world data to clinical research data[J]. China Food Drug Administration, 2021(11): 39-46.] DOI: 10.3969/j.issn.1673-5390.2021.11.005.

37 周晓音. SNOMED CT 在临床路径中应用探讨[J]. 医学信息学杂志, 2010, 31(9): 8-12. [Zhou XY. Discussion and research on the application of SNOMED CT in clinical pathway[J]. Journal of Medical Intelligence, 2010, 31(9): 8-12.] DOI: 10.3969/j.issn.1673-6036.2010.09.002.

38 Wang SV, Patterson OV, Gagne JJ, et al. Transparent reporting on research using unstructured electronic health record data to generate 'real world' evidence of comparative effectiveness and safety[J]. Drug Saf, 2019, 42(11): 1297-309. DOI: 10.1007/s40264-019-00851-0.

收稿日期：2023 年 01 月 05 日 修回日期：2023 年 01 月 28 日
本文编辑：桂裕亮 曹 越

引用本文：阎思宇, 李绪辉, 陈沐坤, 等. 面向真实世界的知识挖掘与知识图谱补全研究（二）：非结构化电子病历信息抽取方法及进展[J]. 医学新知, 2023, 33(5): 358-365. DOI: 10.12173/j.issn.1004-5511.202301016
Yan SY, Li XH, Chen MK, et al. Research on real-world knowledge mining and knowledge graph completion (II): Methods and progress of information extraction from unstructured electronic medical records[J]. Yixue Xinzhi Zazhi, 2023, 33(5): 358-365. DOI: 10.12173/j.issn.1004-5511.202301016