临床文本自动去识别方法比较*

都丽婷

（华中科技大学同济医学院医药卫生管理学院 武汉 430030）

罗 维

（成都中医药大学医学信息工程学院 成都 611137）

李 磊 林 斌

（四川九阵妙凰科技集团有限公司创新中心 成都 610041）

夏晨曦 马国庆 熊丹妮 马敬东

（华中科技大学同济医学院医药卫生管理学院 武汉 430030）

〔摘要〕 介绍临床文本自动去识别的常用方法，包括基于规则的方法、机器学习方法、综合方法等，阐述临床文本去识别研究在临床文本实用性、系统一般性和可扩展性等方面存在的挑战，分析今后的研究方向，为该领域研究者提供借鉴。

〔关键词〕 去识别；自动化；临床自由文本；受保护的健康信息

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2017.04.011

Comparison of Methods for Automatic De－identification of Clinical Texts *DU Li－ting, School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China; LUO Wei, Medical Information Engineering College, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China; LI Lei, LIN Bin, Arrcen Science&Technology Group Co., Ltd., Chengdu 610041, China; XIA Chen－xi, MA Guo－qing, XIONG Dan－ni, MA Jing－dong, School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China*

〔**Abstract**〕 The paper introduces the common methods for automatic de－identification of clinical texts, including the method based on rules, machine learning method, comprehensive method, etc., states the challenges for clinical texts practicability, system universality and scalability of clinical texts de－identification research, analyzes the further research direction, and provides reference for researchers of this field.

〔**Keywords**〕 De－identification; Automation; Clinical free text; Protected Health Information (PHI)

〔修回日期〕 2016－12－20

〔作者简介〕 都丽婷，硕士研究生；通讯作者：马敬东，博士，副教授。

〔基金项目〕 中央高校基本科研业务费资助项目“区域医疗机构知识网络形成机制研究”（项目编号：2015AE017）。

1 引言

信息技术的迅猛发展带动了医院的信息化建设，国家政策的支持为医学信息系统的建立打下了

坚实基础，由此带来了大量的医学数据，而其中临床文本受到了广泛关注。临床文本是医疗活动过程中产生的重要临床信息资源，包含大量有价值的信息，对于临床研究具有重要意义。然而，由于临床文本中大量受保护的健康信息（Protected Health Information, PHI）可能会暴露患者隐私，因此共享临床文本之前必须对隐私信息进行识别和替换。

去识别任务是识别和替换临床文本中 PHI 的过程。2006 年去识别任务作为 I2B2（Informatics for Integrating Biology and the Bedside）工程的子任务首次被提出。现有研究中，美国医疗保险可携带和责任法案（The Health Insurance Portability and Accountability Act, HIPPA）定义的 18 种 PHI 类型是目前去识别系统使用最广泛的标准。但也有越来越多的系统为了更全面地保护患者隐私，在 HIPPA 的基础上扩大了识别范围，如 I2B2 以 HIPPA 标准为基础重新定义了 25 种 PHI 类型[1]。为保证系统一般性，如今使用多种临床文本类型构建和测试系统模型也逐渐成为一种趋势[2-4]。

由于国内对于临床文本中的人员身份标识、隐私保护等问题没有出台相关法律法规，且缺乏中文医疗领域标注语料库，导致国内去识别研究相对较少，研究方法比较局限，仅徐益辉等[5]提出了一种基于中文分词技术识别并处理中文人名的算法。本文对国内外已有临床文本去识别方法进行了细致调研并对未来的研究方向予以展望。

2 资料与方法

以 Web of Science 以及 PubMed 数据库中 SCI - EXPANDED 作为数据来源，检索策略分别为：TS = ((deidentification OR de - identification OR anonymization OR "text scrubbing") AND (medical OR medicine OR biomedical OR clinical))、(deidentification[Title/Abstract] OR de - identification [Title/Abstract]) OR anonymization [Title/Abstract] OR"text scrubbing" [Title/Abstract]，时间限定为 2007 - 2016 年，文献类型限定为期刊文章。Web of Science 共检索出 164 篇，PubMed 共检索出 169 篇（检索时间为 2016 年 12 月 20 日），去重后得到 265 篇文献，除去仅描述人工去识别过程以及仅关注于结构化数据、影像学数据的文献，筛选出描述临床文本自动化去识别系统的代表性文献共 12 篇。本文从多方面对这些去识别系统进行归纳，见表 1，对去识别应用方法、现存的挑战以及今后的研究方向进行总结分析。

表 1 自动化临床文本去识别系统特点总结

第一作者	系统名称	时间	文档类型	PHI 类型	方法	语言	解决主要问题	Precision（%）	Recall（%）	F - measure（%）
Azad Dehghan	Systerm for the i2b2 de - identification challenge	2015	出院小结	i2b2，HIPAA	字典、规则、机器学习	英语	去识别	93.73	88.68	91.30
Bin He, Yi Guan	Systerm for the i2b2 de - identification challenge	2015	出院小结	i2b2，HIPAA	CRFs	英语	去识别	95.14	90.88	92.96
Soo - Yong Shin	bilingual de - identification system	2015	多种临床文本类型（训练集 20 种，测试集 33 种）	8 种类型（姓名、地址、电话号码、邮箱、身份证号、患者 ID、IP 地址、只包括年月的生日）	正则表达式规则	韩语和英语（双语）	去识别、实用性、一般性	100.00	92.97	99.12

续表 1

Jelena Jaimovi	automatic de – identification system for Serbian	2015	多种临床文本类型（出院小结、临床笔记、医疗鉴定）	HIPAA + 包括国家的地理位置和组织机构	规则	塞尔维亚语	去识别、一般性	94.00	94.00	94.00
Hui Yang	Systerm for the i2b2 de – identification challenge	2015	出院小结	i2b2，HIPAA	字典、规则、机器学习	英语	去识别	96.45	90.92	93.60
Zengjian Liu	Systerm for the i2b2 de – identification challenge	2015	出院小结	i2b2，HIPAA	规则、机器学习	英语	去识别	95.64	93.66	94.64
Emmanuel Chazard	FASDIM	2013	出院小结	HIPAA + 卫生保健人员及保健机构名称、卫生机构地理位置	模式匹配	法语	去识别、一般性	79.60	98.10	87.90
Louise Deleger	MIST1，MIST2，MCRF1，MCRF2	2013	多种临床文本类型（出院小结、临床笔记、医疗鉴定）	HIPAA + 包括小于 89 岁的年龄、医院或其他机构信息、包括年的日期	CRFs	英语	去识别、扩展性、实用性	95.73	92.91	94.30
Oscar Ferrández	aka BoB	2012	多种临床文本类型（来自于 VA，100 种）	HIPAA +国家、包括年的日期、医院或其他机构信息	规则、CRFs、SVM	英语	去识别、实用性、一般性	84.60	96.50	90.20
John Aberdeen	MIST	2010	多种临床文本类型（出院小结、实验室报告、letters、order summaries）	15 种类型	CRFs	英语	去识别、一般性	94.30	97.80	96.00
Ishna Neamatullah	MIT system	2008	多种临床文本类型（出院小结、护理记录）	HIPAA + 医生姓名和包括年的日期	模式匹配	英语	去识别	74.90	96.70	–
Szarvas	Systerm for the i2b2 de – identification challenge	2007	出院小结	i2b2，HIPAA	决策树	英语	去识别	98.89	96.42	99.75

3 结果

3.1 基于规则方法

去识别任务可以看作是命名实体识别任务，基于规则的方法一般总是将一些常用命名实体收入词典作为基础，对于词典中没有的命名实体，则通过规则方法来识别，规则的制定一般由正则表达式来完成，多采用语言学专家手工构造规则模板。基于规则的去识别方法，不需要带标注的训练语料，可

以通过添加规则和字典的方式快速简单地提高准确率；但人为编写规则对语言知识要求较高，且往往在语料和语言上会受到一定限制。系统常用的词典列表包括姓名、地理位置、机构、一般词汇（非 PHI 词汇）、医学词汇、UMLS 超级词汇表和拼写检查程序术语表。以模式和字符串相匹配为主要手段，基于规则的系统大多依赖于知识库和词典的建立。基于字典匹配的典型系统如 Ishna Neamatullah[6] 提出的模式匹配算法使用了 4 种类型的字典。已知的 PHI 查找表：患者和医务人员姓名列表（来自 MIMIC II 数据库）；潜在的 PHI 查找表：一般人的姓名、医院名称、地理位置名称（来自 Atkinson's Spell Checking Oriented Word Lists 或 UMLS）；PHI 指示词查找表：作为上下文线索经常出现在 PHI 之前或之后的关键词和短语；非 PHI 查找表：常用词字典（来自 Atkinson's Spell Checking Oriented Word Lists 或 UMLS）。基于正则表达式模式匹配的典型系统如 Shin 等[3] 通过构造 15 种正则表达式规则，设计适用于多种临床文本类型的双语（韩语和英语）临床文本去识别系统；Jacimovic 等[7] 基于已有的塞尔维亚语命名实体识别系统，设计基于规则的临床文本自动去识别系统；法语临床文本去识别系统 FASDIM[8]，基于模式匹配的方法，不需要任何预先定义的字典，可快速简单地进行 PHI 识别。

3.2 机器学习方法

由于与基于规则的方法相比，机器学习方法更具有一般性，移植性较好，因此越来越多的系统采用机器学习方法进行临床文本去识别。其中，使用较多的方法包括条件随机域（Conditional Random Field，CRF）、支持向量机（Support Vector Machine，SVM）和决策树（Decision Tree）等。应用这些方法常常需要大量的人工标注数据进行训练，这是机器学习方法的一个重要挑战。但是带标注的语料可以被分享，如 I2B2 在 2006 年和 2014 年组织的两次临床文本去识别竞赛中，与参与者分享了大量带标注的出院小结文本语料，为系统构建和评估提供基础。应用机器学习方法的去识别系统[4, 9-13]，通过提取不同的特征对模型进行训练，常见的特征包括词汇、拼写和语义特征。词汇特征包括文本中词的属性，当前词以及周围词的词元、词性，窗口长度一般为 3 或 5。拼写特征一般描述词的书写形式，包括词的长度、字母大写、数字、标点符号、首字母缩略词、前后缀等。另外，正则表达式也被看作是机器学习拼写特征的一部分，描述类型一般包括年龄、日期、电话号码、邮编等。语义特征主要指字典特征，常用的字典包括地理位置、日期、城市、职业、姓名、医院名称、疾病名称以及非 PHI 词汇，另外也包括章节标题[12-13]。其他特征还包括词频特征[13] 以及位置特征[10]，位置特征指词的绝对位置和边界特征。CRF 模型是目前去识别系统使用较多且评估效果较高的机器学习模型。He 等[9] 提出一种基于 CRF 模型的去识别系统，经过预处理以及 CRF 模型训练后，分别从实体水平和字符水平对结果进行评估，结果表明基于 HIPPA 标准的字符水平评估结果更优，且证明预处理模块可提高系统最终表现。去识别开源工具 MIST[14] 基于 CRF 模型，支持对不同临床文本类型进行快速自动化去识别。其他模型还包括：一种迭代机器学习的方法框架[13] 应用 Boosting 和 C4.5 算法进行模式识别，取得较好的评估结果。

3.3 综合方法

基于机器学习的去识别系统大多数会结合模式匹配的方式，一方面针对特定类型的 PHI 应用合适的方法，另一方面为了优化机器学习的识别结果。一般偏结构化的、表达方式较统一的 PHI 类型常使用基于规则的方法，如日期、联系方式，其他如医疗机构名等表达形式较复杂的 PHI 类型常结合机器学习方法。基于规则和机器学习的综合方法通过不同阶段优化识别结果，往往可以取得更优的结果。常见系统设计一般包括：基于字符水平和标签水平的 CRF 分类器与基于规则的分类器[12] 的双重识别；第 1 阶段利用规则、字典以及 CRF 分类器最大化召回率，第 2 阶段利用 SVM 分类器过滤假阳性率[4]；应用字典、正则表达式以及 CRF 模型对临床文本进行初始识别，利用初始识别形成的特定患者的字典对文本进行二次识别[10]；结合预处理过程、CRF 模

型构建及后处理规则匹配综合方法[11,15]，进一步提高系统精确度。

4 讨论

4.1 确保去识别后临床文本的实用性

临床文本去识别不仅需要最大化保护患者隐私，识别尽可能多的 PHI，而且还应考虑去识别过程带来的一些挑战[16]，包括临床文本实用性[17]、系统一般性[18]、可扩展性[11]以及逻辑推理[19]导致的隐私泄露。仅关注系统去除 PHI 的能力已不再符合当前的发展需求，对过度识别（错误地去除非 PHI 信息）进行详细测量并分析其对临床文本使用价值的影响是必要的。Meystre 等[17]通过量化临床信息丢失情况，评估系统对临床文本实用性的影响，结果发现 2010 年 I2B2 自然语言处理挑战语料中的疾病、诊断、治疗 3 类临床概念与系统识别标签的重合率小于 2%，证明该去识别系统对临床信息的影响不大但也不可忽略，可以通过改善临床缩略词以及以人名命名的临床概念间的歧义减少这些影响。Delege 等[11]评估了去识别文本在随后药物名称信息抽取任务中的表现，但是药物名称只代表了临床信息中很小的一部分，因此临床文本在随后信息抽取任务中的评估有待进一步深入。

4.2 提高去识别系统的一般性和可扩展性

基于机器学习方法的去识别系统，一般性往往是一个挑战，尤其对于训练和测试的数据集属于相同类型的情况。大部分系统虽选用了多种临床文本类型作为数据集，但均是通过随机分组的方式选取数据。Li 等[18]发现由不同临床文本类型组成的数据集，通过测量书写复杂度进行训练和测试集的分类，比随机分组方法取得了更好的结果。在语言扩展方面，已有英韩双语、塞尔维亚语、韩语等非英语语料去识别系统取得了较好结果。

面对临床文本数据量的不断增长，去识别系统的可扩展性问题同样需要解决，以大量且不同类型的临床文本为样本是未来的发展趋势。另外，对于逻辑推理导致的隐私泄露问题，Sánchez 等[19]提出一种临床文本自动化去识别方法，避免通过逻辑推理泄露患者敏感信息，且保证了处理后文本的实用性。

5 结语

本研究总结分析近 10 年来自动化临床文本去识别方法研究的发展状况、现存挑战及今后的研究方向。针对自动化临床文本去识别研究需要解决的主要问题，未来的研究方向主要包括：临床缩略语以及以人名命名的临床术语引起的歧义消除问题；测量临床文本在随后更全面的信息抽取任务中的互操作性[11]；解决逻辑推理对患者隐私带来的威胁[19]；利用大量多类型临床文本进行模型构建和测试，保证系统一般性，同时提升大数据环境下系统的计算性能[20]。

参考文献

1 Meystre SM, Friedlin FJ, South BR, et al. Automatic De-identification of Textual Documents in the Electronic Health Record: a review of recent research [J]. BMC Med Res Methodol, 2010, (10): 70.

2 Ferrández O, South BR, Shen S, et al. Evaluating Current Automatic De-identification Methods with Veteran's Health Administration Clinical Documents [J]. BMC Med Res Methodol, 2012, (12): 109.

3 Shin S, Park YR, Shin Y, et al. A De-identification Method for Bilingual Clinical Texts of Various Note Types [J]. Journal of Kroean Medcial Science, 2015, 30 (1): 7-15.

4 Ferrandez O, South BR, Shen S, et al. BoB, A Best-of-breed Automated Text De-identification System for VHA Clinical Documents [J]. Journal of the American Medical Informatics Association, 2013, 20 (1): 77-83.

5 徐益辉，姚琴，袁冬生，等. 中文医疗文本匿名化方法研究 [J]. 中国数字医学, 2014, 9 (7): 19-21.

6 Neamatullah I, Douglass MM, Lehman LH, et al. Automated De-identification of Free-Text Medical Records [J]. BMC Medical Informatics and Decision Making, 2008, (8): 32.

7 Jacimovic J, Krstev C, Jelovac D. A Rule-based System for Automatic De-identification of Medical Narrative Texts [J]. Informatica-Journal of Computing and Informatics,

2015, 39 (1): 43 - 51.

8 Chazard E, Mouret C, Ficheur G, et al. Proposal and Evaluation of FASDIM, a Fast and Simple De - identification Method for Unstructured Free - Text Clinical Records [J]. International Journal of Medical Informatics, 2014, 83 (4): 303 - 312.

9 He B, Guan Y, Cheng J, et al. CRFs Based De - identification of Medical Records [J]. Journal of Biomedical Informatics, 2015, (58): S39 - S46.

10 Dehghan A, Kovacevic A, Karystianis G, et al. Combining Knowledge - and Data - Driven Methods for De - identification of Clinical Narratives [J]. Journal of Biomedical Informatics, 2015, (58): S53 - S59.

11 Deleger L, Molnar K, Savova G, et al. Large - scale Evaluation of Automated Clinical Note De - identification and Its Impact on Information Extraction [J]. Journal of the American Medical Informatics Association, 2013, 20 (1): 84 - 94.

12 Liu Z, Chen Y, Tang B, et al. Automatic De - identification of Electronic Medical Records Using Token - level and Character - level Conditional Random Fields [J]. Journal of Biomedical Informatics, 2015, (58): S47 - S52.

13 Szarvas G, Farkas R, Busa - Fekete R. State - of - the - art Anonymization of Medical Records Using an Iterative Machine Learning Framework [J]. J Am Med Inform Assoc, 2007, 14 (5): 574 - 580.

14 Aberdeen J, Bayer S, Yeniterzi R, et al. The MITRE Identification Scrubber Toolkit: design, training, and assessment [J]. International Journal of Medical Informatics, 2010, 79 (12): 849 - 859.

15 Yang H, Garibaldi J M. Automatic Detection of Protected Health Information from Clinic Narratives [J]. Journal of Biomedical Informatics, 2015, (58): S30 - S38.

16 Velupillai S, Mowery D, South BR, et al. Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis [J]. IMIA Yearbook, 2015, 10 (1): 183 - 193.

17 Meystre SM, Ferrández ó, Friedlin FJ, et al. Text De - identification for Privacy Protection: a study of its impact on clinical text information content [J]. Journal of Biomedical Informatics, 2014, (50): 142 - 150.

18 Li M, Carrell D, Aberdeen J, et al. De - identification of Clinical Narratives Through Writing Complexity Measures [J]. International Journal of Medical Informatics, 2014, 83 (10): 750 - 767.

19 Sánchez D, Batet M, Viejo A. Utility - preserving Privacy Protection of Textual Healthcare Documents [J]. Journal of Biomedical Informatics, 2014, (52): 189 - 198.

20 Cyganek B, Grana M, Krawczyk B, et al. A Survey of Big Data Issues in Electronic Health Record Analysis [J]. Applied Artfiicial Intelligence, 2016, 30 (6SI): 497 - 520.

关于《医学信息学杂志》启用“科技期刊学术不端文献检测系统”的启事

为了提高编辑部对于学术不端文献的辨别能力，端正学风，维护作者权益，《医学信息学杂志》已正式启用“科技期刊学术不端文献检测系统”，对来稿进行逐篇检查。该系统以《中国学术文献网络出版总库》为全文比对数据库，可检测抄袭与剽窃、伪造、篡改、不当署名、一稿多投等学术不端文献。如查出作者所投稿件存在上述学术不端行为，本刊将立即做退稿处理并予以警告。希望广大作者在论文撰写中保持严谨、谨慎、端正的态度，自觉抵制任何有损学术声誉的行为。

《医学信息学杂志》编辑部