

国外医学领域自动问答系统研究现状及启示*

张芳芳 马敬东 王小贤 卢乃吉 夏晨曦

(华中科技大学同济医学院医药卫生管理学院 武汉 430030)

[摘要] 分析国外医学领域自动问答系统研究相关文献,对比自动问答与传统网络信息搜索的不同点,对自动问答系统的应用研究、常用工具及相关技术等方面进行阐述。结合国外相关研究探讨医学领域自动问答研究的关键问题及对我国的启示,以期为进一步研究提供参考。

[关键词] 问答系统;医学领域;大数据应用;自然语言处理

[中图分类号] R-056 **[文献标识码]** A **[DOI]** 10.3969/j.issn.1673-6036.2017.03.001

Research Status and Implications of Foreign Medical Automatic Question Answering System ZHANG Fang-fang, MA Jing-dong, WANG Xiao-xian, LU Nai-ji, XIA Chen-xi, School of Medical and Health Management, Tongji Medical College, HUST, Wuhan 430030, China

[Abstract] The paper analyzes the documents related to the foreign medical automatic Question Answering (QA) system, and compares the differences between automatic QA and traditional network information retrieval, describes the automatic QA system in the aspects of application researches, common tools and relevant technologies. Combining foreign related researches, it also discusses the current key problems of medical automatic QA system and the implications for China in order to offer reference for further research.

[Keywords] Question Answering (QA) system; Medical domain; Big data application; Natural Language Processing (NLP)

1 引言

随着大数据时代的到来及互联网技术的进一步发展,人们获取各类信息变得更加便捷、有效,现在人们已经习惯于通过网络来获取各种各样的信

息^[1]。在医学领域医学信息的获取主要依靠搜索引擎来实现^[2],包括医学数据库、专业的健康网站、综合搜索引擎中的健康板块等。搜索引擎可帮助人们快速地找到包含所需信息的网页,人们也只需输入一些关键字,搜索引擎便会马上搜索出许多相关的网页。但是这些传统搜索引擎存在很多不足之处,如返回相关网页太多,而且大量重复的文章没有过滤,面对海量的健康信息资源,即便是专业人员,仍需要耗费更多的时间来实现健康信息的搜索、浏览及获取,搜索的成功与否取决于用户的搜索技巧^[3]。而且由于医学专业的特殊性,面对质量参差不齐的健康信息^[4],非医学专业人员在查找、理解及获取方面存在诸多困难和障碍,难以满足用

[修回日期] 2016-10-13

[基金项目] 张芳芳,硕士研究生,发表论文1篇;通讯作者:夏晨曦,讲师。

[基金项目] 教育部网络时代的科技论文快速共享专项研究课题“基于互联数据的论文共享方法研究”(项目编号:0214516155)。

户健康需求^[5]。医疗大数据的应用对于临床医学研究、科学管理和医疗服务模式转型发展都具有重要意义^[6]，因此，如何管理和有效利用海量数据资源成为当前人们关注的热点问题^[7]。相对于传统信息检索，自动问答系统作为更智能的医学信息资源获取工具，为海量数据资源的有效利用提供可能。

从研究角度来看，自动问答作为自然语言处理领域的热门研究方向，在国外医学领域的应用已有了初步发展，但在国内尚未引起充分重视。为更有效地利用网络健康信息资源，推动中文领域医学自动问答研究和应用的发展，本文分析国外医学领域自动问答系统相关文献，重点介绍不同发展阶段的相关研究成果及应用，指出当前医学领域自动问答研究中存在的不足，以期为中文领域进一步开展相关研究提供借鉴。

2 国外医学领域自动问答概述

2.1 发展历程

国外自动问答的研究起步相对较早，1950 年的“图灵测试”是关于自动问答系统的最早构想^[8]。1999 年的文本检索会议设立问答系统评测专项^[9]，极大促进了问答系统的发展。相较于其他领域，问答系统在医学领域的应用起步较晚，尚处于初步发展阶段，按其发展阶段不同，可以分为基于传统检索技术的问答系统和基于语义技术的问答系统。目前医学领域自动问答的研究中，大部分是基于非语义技术的研究。

2.2 与传统医学信息获取模式对比分析

随着大数据时代的到来，面对网络上质量参差

不齐的医学信息，基于自动问答的医学信息搜寻模式对信息资源的利用具有重大意义。自动问答系统在医学领域的应用，可缓解医患之间信息不对称、提高医疗资源利用效率，同时也能更好地体现“以病人为中心”服务理念的转变。医学领域自动问答系统与传统医学信息检索模式在一些方面存在着差异，见表 1。

表 1 医学领域自动问答与传统医学信息搜寻模式对比

比较指标	传统医学信息搜寻	自动问答系统
服务理念	以疾病为中心	以病人为中心
信息搜寻方式	关键词	自然语言表达
检索时间	根据检索技巧及专业知识	系统自动整合、有效减少搜寻时间
答案返回形式	多个相关网页或文档	整合后精确、简短的片段

3 医学领域自动问答系统研究现状

3.1 应用研究

3.1.1 基于传统检索技术的问答系统相关研究
 基于传统检索技术的问答系统，在问题分析中将问题中的关键词与数据资源中的关键词进行匹配，进而获取可能相关的答案片段。其数据来源主要包括万维网、医学文献、临床指南、维基百科，以及抓取的在线健康网站的问题答案对，如雅虎问答中的问题答案对，通过对语料资源的整合，为自动问答的应用提供可及的基础语料资源。但此类应用研究仅限于关键词的匹配，对生成答案的质量有一定的局限。典型的医学领域自动问答应用具体，见表 2。

表 2 基于传统检索技术的自动问答系统相关研究

系统名称	针对人群	数据来源	主要内容	存在的不足
MedQA ^[10]	医学专业 人员	万维网和 Medline 临床指南、维基百科、	回答定义类的医学领域问题，针对医生提问	返回可能相关的段落，可能会影响系统性能 返回可能相关的片段，通常返回较多结果
AskHERMES ^[11]	医学专业 人员	医学文献	可回答复杂临床问题，针对医生的提问	—
MiPACQ ^[12]	医学专业 人员	医学文献	系统为多源集成平台，整合成可扩展的问答系统	针对其他语料库时，机器学习分类器促进重新排序程度有待进一步研究
enquireMe ^[13]	非医学专业 人员	雅虎问答对	从在线社区中提取问答对，形成知识库	未考虑同义词扩展等问题

3.1.2 基于语义技术的问答系统 基于语义技术的问答系统在问题处理方面,对自然语言问题进行语义处理,实现从语义层面理解用户提出的问题。在多元异构的医疗数据处理方面,通过采用相关语义技术,可将多元异构的医疗数据资源整合成以资源描述框架(Resource Description Framework, RDF)

3 元组形式存储的数据资源,从而实现整个医学领域资源的共享和利用。语义技术的不断进步为基于语义技术的自动问答在医学领域的应用提供可能。相关研究者对基于语义技术的问答系统进行研究,但相关的应用研究目前较少,见表 3。

表 3 基于语义技术问答系统相关研究

系统名称	针对人群	数据来源	特点	存在的不足
MEANS ^[14]	医学专业人员	Medline	答案提取基于语义查询松弛策略,可提高整体性能	把问题分为 10 类,只能解决其中 4 类问题
AskCuebee ^[15]	医学专业人员	TriTrypDB 和 KEGG 及研究小组内部试验数据	不把问题局限于特定集合或者模板,不需问题和模板进行匹配	语义关系发现中某些不能有效识别,如否定关系、使用嵌套等复杂查询模式

3.2 关键技术

3.2.1 问句处理模块工具及方法 在医学领域问答系统的研究中,每一模块都有比较成熟的算法或工具。问题预处理可利用现有工具,如 Stanford Parser、Stanford CoreNLP 和 OpenNLP 等。对于问题分类,支持向量机的应用最为广泛^[16]。重要实体识别中,条件随机域算法应用较多^[17]。此外,基于语义技术的问答系统研究中,语义关系的提取也有对应方法,如最近公共祖先^[15]。查询语言转换则将用户输入的自然语言转换为 RDF 3 元组形式,最后生成为 SPARQL 语句,以方便对信息进行检索。

3.2.2 信息处理模块关键技术 在信息处理模块,关键是问句与数据资源的匹配。对于传统的基于信息检索问答系统研究,主要包括问句与资源库文档匹配及问句与问答对集合中问题相似度。基于语义技术的自动问答研究中,则通过转换为 SPARQL 查询语句的问句与本体知识库匹配。相关研究如 Wong 等^[13]利用应用程序接口从雅虎问答中提取问答对存储在数据库中,匹配至少包含 1 个关键词的问答对作为候选答案;Asiaee 等^[15]以 RDF 3 元组形式存储作为知识库,使用查询语言 SPARQL 进行信息检索。相关研究总结,见表 4。

表 4 问答系统信息检索研究总结

研究内容	数据来源	分类
基于传统检索技术的自动问答	在线及本地文档资源库	问句与资源库中文档匹配
问题答案对	问句与问答对中间问句匹配	SPARQL 语句
基于语义技术的问答系统	RDF 存储及本体	与本体库的 3 元组匹配

3.2.3 答案生成模块关键技术研究 在答案生成模块,通过候选答案的排序、总结,最终以可视化的方法呈现给用户。基于传统检索技术问答系统中,答案生成关键在于将抽取的相关片段组合成连贯、简洁的语句呈现给用户。如 Yu 等^[18]采用 Google 定义和 TF * IDF 相结合方法,生成网络定义与 Medline 总结中前 5 位句子;Wong^[13]等根据得分函数及规则,对问答对排序,分数高的作为最终答案呈现给用户。在基于语义技术的问答系统中,答案生成模块则利用抽取 RDF 3 元组转换为自然语言,如 Asiaee 等^[15]利用已标注的 RDF 数据将 3 元组转换为用户可理解语言。

3.2.4 问答系统评价及数据来源 问答系统的研究中,需对开发的问答系统的性能进行评价,以检测系统的性能,进而对系统进行改进优化。对于问答系统的评价主要分为两大类,对准确率、召回率、F1 值作为评价标准进行评价^[15],及根据答案

质量、易用性、花费时间、操作数等评价指标对问答系统的性能进行认知评价^[18]。其中对于认知评价,因为局限性只能根据小样本进行,不能代表广泛人群,具有一定的主观性。

4 医学领域自动问答系统研究的关键问题

4.1 缺乏针对非医学专业人员知识需求研究

由于患者的医学专业知识相对比较缺乏,提出的问题针对性不强,因此,针对患者的医学领域问答系统研究较少。一些研究者为了弥补患者医学知识不足开发半自动算法简化医疗文本^[19],研究弥补患者与医学专业知识间差距^[20]。但随着医疗资源的供不应求及公众健康需求的不断增加,迫切需要针对健康需求者问题的问答系统研究。

4.2 中文领域问答系统研究不足

英文环境下有大量可用的知识及资源,为英文领域问答系统的发展提供了条件。大多医学领域问答系统的研究及开发是在英文环境下,而较少有研究涉及其他语言,尤其是中文^[21]。随着医学领域问答系统的发展,相关研究者对其他语言问答系统也进行研究,如 Brasch 等^[22]对丹麦语糖尿病领域问答系统进行研究, Yin 等针对健康需求者通过网络抽取数据开发问答系统 HealthQA^[21]。

4.3 基于语义技术的问答系统研究较少

由于医学相关信息的爆炸式增长,其中不仅包括结构化数据,还包含大量非结构化数据,传统的问答系统在医学信息资源的整合中存在困难。此外医学领域中 RDF 标注的资源越来越多^[23],对于用户直观访问和理解数据提出了挑战,传统的问答系统存在的这些问题也为基于语义技术的医疗问答系统的发展提供了机遇。但已有的本体,特别是中文本体并不能满足医学领域内日益增长的需求。在数据整合和推理方面,目前的方法和工具也并不成熟,还需要根据实际应用的情况进行改进。基于语义技术的问答系统还面临很多挑战,但具有巨大的发展潜力。

5 对中文领域医学自动问答系统研究的启示

5.1 关注非医学专业人员的信息需求

当前,我国该领域的研究及实践仍然是以满足医学相关人员的需求为主要目的^[24]。由于医学专业的特殊性,面对质量参差不齐的医学信息,非医学专业人员很难进行科学的搜寻、判断和筛选。因此,应从提升信息检索系统的功能入手,以弥补非专业人员因医学知识不足而导致的搜寻障碍,进而满足非医学专业人员的信息需求。

5.2 重视基础设施建设

自动问答系统的发展离不开大量可用知识及资源,但中文相关医学文献的标引发展还不完善,国内医学领域自动问答可用的语料资源相对较少,需要领域专家探索有效的方法对医学领域资源进行标注,从而提供可用资源。医学专业词汇标准化也是中文问答系统需要进一步研究的问题,如通过本体的构建将自然语言表达的词汇转换为标准化词汇,进而促进语义技术在医学领域自动问答中的研究。复旦团队在 BioASQ 竞赛中生物医学语义标引获得第 1 名^[25],但该研究是在英文环境中。因此在医学领域应构建中文问答系统评测平台,以支持问答系统性能的测验,促进中文领域医学领域问答系统发展。

5.3 确定适合中文领域的工具和方法

通过研究发现,目前对于医学领域问答系统的构建研究较少,适用于英文领域的工具和方法对于中文领域的适用性尚未有相关研究。因此,我国可以借鉴国外的研究成果,在完善基础语料库的基础上,增加对医学领域问答系统相关工具和方法的评价分析,尽可能准确地找出适合中文领域的工具和方法,促进问答系统达到更好的性能,促进医学领域自动问答更好地应用于中文环境。

5.4 重视基于语义技术自动问答的研究

随着语义技术的不断发展,逐渐应用到医学领

域的自动问答研究中也发挥了其巨大的优势,为大数据环境下医学资源的整合与共享提供了可能。通过文献分析可知,相较于传统医学领域自动问答,基于语义技术的自动问答在问题处理模块,对自然语言问句进行语义处理,实现从语义层面理解问句。在多元异构的医疗数据处理方面,通过相关语义技术,可将多元异构医疗数据资源整合成以 RDF 3 元组形式存储的数据资源,从而实现整个医学领域资源的共享和利用。因此,应重点关注基于语义技术的问答系统的研究,以弥补国内相关系统研究的不足。

6 结语

大数据时代的到来使医学领域的海量数据资源也面临着重大挑战。在多元异构的医疗数据处理方面,利用语义技术可将多元异构医疗数据资源整合成以 RDF3 元组形式存储的数据资源,通过问答系统将自然语言问题转换为 SPARQL 查询语句,从而实现整个医学领域资源的共享和利用。因此,基于语义技术的问答系统在医学领域虽然面临诸多挑战,但可以有效利用海量的医学信息,具有广阔的应用前景。

参考文献

- 施亦龙, 许鑫. 在线健康信息搜寻研究进展及其启示 [J]. 图书情报工作, 2013, 57 (24): 123 - 131.
- Stephens MB, Von Thun AM. Military Medical Informatics: accessing information in the deployed environment [J]. Military Medicine, 2009, 174 (3): 259 - 264.
- Wren JD. Question Answering Systems in Biology and Medicine - the Time is Now [J]. Bioinformatics, 2011, 27 (14): 2025 - 2026.
- 李迎娟. 网络医学信息资源高效利用的绿色评价研究 [J]. 医学信息学杂志, 2015, 36 (5): 75 - 78.
- 王若佳, 李月琳. 基于用户体验的健康类搜索引擎可用性评估 [J]. 图书情报工作, 2016, (7): 1 - 11.
- 汪鹏, 吴昊, 罗阳, 等. 医疗大数据应用需求分析与平台建设构想 [J]. 中国医院管理, 2015, 35, (6): 40 - 42.
- 张振, 周毅, 杜守洪, 等. 医疗大数据及其面临的机遇与挑战 [J]. 医学信息学杂志, 2014, 35, (6): 2 - 8.
- 李珊珊. 基于美食本体的问答系统研究与实现 [D]. 武汉: 武汉科技大学, 2015.
- Voorhees EM. The TREC - 8 Question Answering Track Report [J]. TREC, 1999, (99): 77 - 82.
- Lee M, Cimino J, Zhu HR, et al. Beyond Information Retrieval - Medical question answering [C]. Washington DC: Proceedings of AMIA Annual Symposium, 2006: 469 - 473.
- Cao Y, Liu F, Simpson P, et al. AskHERMES: an online question answering system for complex clinical questions [J]. Journal of Biomedical Informatics, 2011, 44 (2): 277 - 288.
- Cairns BL, Nielsen RD, Masanz JJ, et al. The MiPACQ Clinical Question Answering System [C]. Washington DC: Proceedings of AMIA Annual Symposium, 2011: 171 - 180.
- Wong W, Thangarajah J, Padgham L. Contextual Question Answering for the Health Domain [J]. Journal of the American Society for Information Science and Technology, 2012, 63 (11): 2313 - 2327.
- Abacha B, Zweigenbaum P. MEANS: a medical question - answering system combining NLP techniques and semantic Web technologies [J]. Information Processing and Management, 2015, 51 (5): 570 - 594.
- Asiaee AH, Minning T, Doshi P, et al. A Framework for Ntology - based Question Answering with Application to Parasite Immunology [J]. Journal of Biomedical Semantics, 2014, 6 (1): 1 - 25.
- Cruchet S, Gaudinat A, Boyer C. Supervised Approach to Recognize Question type in a QA System for Health [J]. Studies in Health Technology and Informatics, 2008, (136): 407 - 412.
- Cao Y, Cimino JJ, Ely J, et al. Automatically Extracting Information Needs from Complex Clinical Questions [J]. Journal of Biomedical Informatics, 2010, 43 (6): 962 - 971.
- Yu H, Lee M, Kaufman D, et al. Development, Implementation, and a Cognitive Evaluation of a Definitional Question Answering System for Physicians [J]. Journal of Biomedical Informatics, 2007, 40 (3): 236 - 251.
- Leroy G, Endicott JE, Mouradi O, et al. Improving Perceived and Actual Text Difficulty for Health Information Consumers Using Semi - automated Methods [C]. Chicago: Proceedings of AMIA Annual Symposium, 2012: 522 - 531.

(下转第 25 页)

中寻找匹配对象的传统的服务器, 这样选择的样本可更加有效地用于自身健康分析的参考样本。一种是获得了大量相似用户的数据, 有利于发现规律; 另一种是可以借助该共享网络为医生提供大量同种病例的对比信息, 非常有利于对医生的培养。

3 结语

本文以脉搏信号的采集、传输、分析为例, 设计远程移动医疗系统, 通过开发一系列基于智能手机应用的软件, 实现对脉搏数据的高速传输, 探究利用智能机构建自组织医疗共享网络, 构建一种基于智能手机的远程移动医疗系统模型。在这个网络中可以对比信息分析健康状况, 找到最佳健康恢复方案, 移动医疗的便携性真正在生活中体现出来, 更好地满足人们的需求, 扩大移动医疗的应用范围。由于条件有限, 本研究远程移动医疗的数据采集对象仅局限于脉搏信号。使用智能手机实现其他医疗信号采集、传输与分析, 需要进一步的探究。

参考文献

- 1 Hii P C, Chung W - Y. A Comprehensive Ubiquitous Healthcare Solution on an Android Mobile Device [J]. *Sensors*, 2011, 11 (7): 6799 - 6815.
- 2 樊宁. 基于 Android 平台的无线控制系统设计与实现 [D]. 上海: 华东理工大学, 2012.
- 3 Sparsh Agarwal, Chiew Tong Lau. Remote Health Monitoring Using Mobile Phones and Web Services [J]. *Telemedicine and e - Health*, 2010, 16 (5): 603 - 607.
- 4 H Zheng, C Nugent, P McCullagh, et al. Smart Self Management: assistive technology to support people with chronic disease [J]. *Journal of Telemedicine and Telecare*, 2010, 16 (4): 224 - 227.
- 5 Qing Pan, Pan Yang, Rui Zhang, et al. A Mobile Health System Design for Home and Community Use [C]. Hong Kong: IEEE - EMBS International Conference on Biomedical and Health Informatics, 2012: 116 - 119.
- 6 Kang K C, Heo S U, Bae C. S. Android/OSGi - based Mobile Healthcare Platform [C]. Seogwipo, Korea (South): International Conference on Advanced Information Management and Service, 2011: 125 - 126.
- 20 Nie L, Zhao Y, Akbari M, et al. Bridging the Vocabulary Gap between Health Seekers and Healthcare Knowledge [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2015, 27 (2): 396 - 409.
- 21 Yin YS, Zhang Y, Liu X, et al. HealthQA: a Chinese QA summary system for smart health [J]. Springer International Publishing, 2014, (8549): 51 - 62.
- 22 Braasch A, Huszka C, Henriksen L, et al. Creation and Use of Language Resources in a Question - Answering e - health System [C]. Istanbul: Proceedings of the Eighth International Conference on Language Resources and Evaluation, 2012: 23 - 25.
- 23 Ostankov A, Roehrbein F, Waltinger U. LinkedHealthAnswers: towards linked data - driven question answering for the health care domain [C]. Reykjavik: International Conference on Language Resources and Evaluation, 2014: 2613 - 2620.
- 24 彭琰, 严莉, 朱红. 大数据时代用户健康信息学的价值 [J]. *医学信息学杂志*, 2014, 35 (1): 1 - 6.
- 25 Tsatsaronis G, Balikas G, Malakasiotis P, et al. An Overview of the BIOASQ Large - scale Biomedical Semantic Indexing and Question Answering Competition [J]. *BMC Bioinformatics*, 2014, 16 (1): 1 - 28.

(上接第6页)