·综述·

智能问答系统在医学领域的应用研究

贺 佳,杜建强,聂 斌,熊旺平,罗计根

(江西中医药大学计算机学院,江西 南昌 330004)

摘 要:智能问答系统可以快速、准确地为用户提供信息服务,是自然语言处理领域的备受关注的研究方向。在医学知识服务领域,也具有很好的应用前景和发展空间。论文首先简述了医学领域智能问答系统的研究情况,其次就医学智能问答系统中的问题分析、信息检索、答案抽取三个组成部分及其关键技术进行了分别阐述;最后对其在中医方面的应用进行了阐述,并对医学智能问答系统的进一步发展提出了展望。

关键词:医学领域;智能问答系统;中医药

中图分类号:TP391 **文献标识码**:A **DOI**:10.3969/j.issn.1006-1959.2018.14.007

文章编号:1006-1959(2018)14-0016-04

Research on the Application of Intelligent Question-answering System in Medical Field

HE Jia,DU Jian-qiang,NIE Bin,XIONG Wang-ping,LUO Ji-gen

(School of Computer Science,Jiangxi University of Traditional Chinese Medicine,Nanchang 330004,Jiangxi,China)

Abstract:Intelligent question-answering system can provide information service to users quickly and accurately,which is the research direction of natural language processing.In the field of medical knowledge service,it also has a good application prospect and development space.Firstly,this paper briefly introduces the research situation of intelligent question-answering system in medical field, and then expounds the three components and key technologies of question analysis,information retrieval and answer extraction in medical intelligent question answering system.Finally,its application in traditional Chinese medicine is expounded,and the further development of medical intelligent question-answering system is prospected.

Key words:Medical field;Intelligent question-answering system;Traditional Chinese medicine

随着科学技术的发展,互联网应用开始普及于人类生活的方方面面,健康医疗与互联网相结合是医学信息化发展的必然趋势。医学信息化的发展影响着人们对健康知识的获取方式。目前,对医学相关知识的搜索主要通过传统搜索引擎,例如百度百科、360 搜索等,这种搜索方式一般只需用户输入关键字,便会返回大量的网页。然而这些方式难以满足用户的需求:一方面系统不能返回给用户最直接的答案,而是一些与问题相关的网页或者文档,用户需要再次从这些网页或者文档中寻找最终想要的答案。尤其对于非医学专业人员,他们对医学知识了解不深,寻找答案会更加困难。另一方面答案质量参差不齐,用户在查找、获取、理解方面会存在许多困难。智能问答系统不仅可以允许用户以自然语言方式提问,还能返回给用户准确、简洁的答案,不需要用户再次筛选合适的答案。将智能问答系统应用于医学领域,能够进一步提高人们获取健康知识的便捷性、准确性。

1 智能问答系统在医学领域研究概况

1.1 发展历程 智能问答系统的发展可追溯到图灵测试时期,其主要测试机器是否具备人类智能。20 世纪 60 年代,由于计算、数据资源有限,主要是限定领域智能问答系统发展,比如专家系统。这些系统中搜索答案的数据集来自于专家手工编写。90 年代以来,自然语言处理技术的兴起和语义信息的应用,以及随着网上的资源越来越丰富,智能问答系统得到了快速发展[1]。尤其在 1999 年国际文本检索会议(简称 TREC)引入了问答系统评测专项(简称 QA Track)后,QA Track 成为了最受欢迎的 TREC 评测项目之一,智能问答系统的发展速度越来越快。相比之下,医学领域智能问答系统研究起步较晚,当前尚处于初步发展阶段[2]。

1.2 研究现状

1.2.1 国外研究情况说明 国外在医学领域智能问答系统研究中已有了初步发展,国外的医学智能问答系统,见表 1。从面向的对象来看,MedQA、AskHERMES、MEANS、AskCuebee 主要针对医学专

基金项目:1.国家自然科学基金项目(编号:61363042);2.国家自然科学基金项目(编号:61562045);3.国家自然科学基金项目(编号:61762051);4.江西省自然科学基金重大项目(编号:20171ACE50021);5.江西省研究生创新专项资金(编号:YC2017-S349);6.江西省科技支撑计划(编号:20141BBE50031)

作者简介:贺佳(1992.6-),女,陕西渭南人,硕士研究生,研究方向:机器学习、医药自然语言处理

通讯作者:杜建强(1968.10-),男,江西南昌人,博士,副校长,教授,研究方向:医药信息与数据挖掘

业人员，如 AskCuebee 是一款用于寄生虫学家获取与寄生虫有关知识的系统。mnquireMe 则是针对大众群体，通过问题-答案对返回给用户想要的信息。从使用技术的不同来看，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。基于传统检索技术的问答系统有 MedQA、AskHERMES、mnquireMe，基于语义技术的问答系统有 MEANS、AskCuebee。基于传统检索技术的问答系统主要采用关键词匹配技术，这种方法的问答系统对于抽取的答案质量有一定局限性。基于语义技术的问答系统主要采用语义分析法对问题和抽取的答案进行分析，不再只是单纯的关键词匹配，而是从语义层面上对问题和答案进行了分析思考，答案质量有所提高。

1.2.2 国内研究情况说明　国内智能问答系统起步相对较晚。HestiaQA 是由 Zhang 等人[8]针对疾病咨询所做的中文问答系统。中科院计算研究所研究过一款医学检索系统[9]，这个系统采用深度问答方法对问题进行分析以及答案的抽取。由赵欣[10]发明的基于疾病圈的疾病自诊知识问答系统主要为了大众提供疾病知识的科学依据。其主要研究步骤是：建立疾病圈，根据疾病的不同建立不同的疾病子圈，系统从疾病圈抽取出一些问题用于该圈子的会员回答，另一方面，会员可以向题库中添加问题，由专家审核确认。运用此方法，疾病圈即知识库会越来越大。在社区类问答系统方面，国内出现了一些比较有名的医学信息服务类网站，如寻医问药网，快速问医生等[11,12]。这类网站虽然允许用户通过各种形式提问问题，但是返回给用户的答案较多，对于非专业人员，其获取准确答案较为困难。

表 1　国外医学智能问答系统

系统名称	数据来源	主要内容	优缺点
MedQA[3]	万维网、Medline 数据	采用浅句法对问句进行分块，获取名词短语作为关键词。	只能回答定义类问题；没有采用语义分析，其返回的答案有待提高。
AskHERMES[4]	维基百科、临床指南、PubMed 数据、Medline 数据	能够分析和理解复杂的临床问题	在冗长复杂问题的字数统计上具有良好的健壮性。
mnquireMe[5]	雅虎问答对	通过问答对构建知识库；使用关键词提取和加权相结合的衰变模型来匹配和评分问题答案对；	系统对于同义词的处理效果不明显。
MEANS[6]	Medline 数据	对问题进行了分类；答案提取采用语义查询技术	采用了查询松弛方法，用以处理一些情况下自然语言处理方法的错误或缺点；只能处理一般类型的问题，对于复杂问题的处理效果不佳。
AskCuebee[7]	TriTrypD、KEGG	回答关于寄生虫免疫学的相关知识。	否定关系、语句嵌套等类似的查询模式，不能被有效识别。

2 医学智能问答系统组成

一般来说，智能问答系统主要由三部分组成，分别是：问句分析、信息检索、答案抽取[13]。系统对用户提出的问题进行分析，将问题分析后所得的信息给信息检索环节，检索出相关文档或段落，利用答案抽取技术将最终答案返回给用户，见图 1。

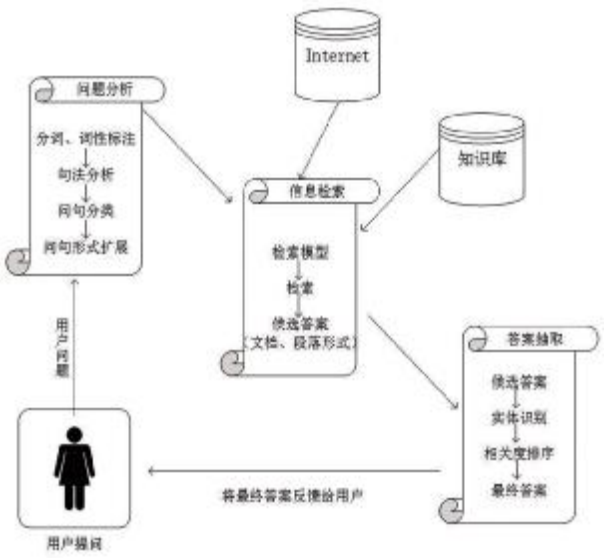

图 1　医学智能问答系统运行流程

2.1 问题分析　问题分析指将用户语言转化为计算机能够识别的语言。问题分析是智能问答系统首要环节，这一部分所用技术的成熟度影响着候选答案的精度。问题分析部分主要包括对问句进行中文分词、词性标注、句法分析，为了能够快速准确找到答案，还要对问句进行分类，最后进行关键词提取和拓展。其中，对于分词、词性标注等，可以采用哈工大社

会计算与信息检索研究中心开发的语言技术平台。对于问句分类,常采用支持向量机。关键词提取和拓展,一般用统计方法,其中含词频、共现频率等统计信息[14]。

李冬梅等[15]采用浅层句法分析和最大熵模型的语义分析算法对问题进行分析,利用构建的生物医学领域本体知识库进行 SPARQL 查询,进而实现结果的输出。刘凯等人[16]通过将条件随机场、隐马尔可夫模型、最大熵马尔可夫模型用于中医病历命名实体抽取实验,结果证实了条件随机场相比于其它两种方法具有较高的准确率和召回率。张芳芳等人[17]以糖尿病患者的饮食问题为例,采用支持向量机模型对问题进行分类,为深度自动问答系统提供了重要支撑。孟洪宇等[18]采用基于条件随机场方法,通过字本身、词性、词边界、术语类别标注的多特征融合模型对《伤寒论》中的术语进行了识别。

2.2 信息检索 信息检索旨在缩小答案存在的范围。该部分根据用户的问题从文档、网页或者知识库中提取可能相关的候选答案,候选答案可以是文档,也可以是段落,基于知识图谱的知识库最终得到的是拥有实体和实体关系链接的一个知识库子图。

在基于传统检索技术的医学问答系统中主要采用关键词匹配技术,。一般对于文档,可以使用检索模型如布尔模型、向量空间模型、语言模型等[19]检索候选答案。基于语义分析技术的医学问答系统中主要将问题的分析结果转化为 SPARQL 等查询语句,然后与本体知识库匹配。

Asiaee 等人[7]建立的知识库以 RDF 三元组作为存储形式,通过 SPARQL 语句进行信息检索。Wong 等人[5]利用关键词匹配技术从雅虎问答对中选取最贴近的问答对作为候选答案。

2.3 答案抽取 答案抽取即从候选答案中抽出最佳答案返回给用户。一般的答案抽取流程是:对候选文档或段落进行切分并形成候选答案集,根据问题类型对候选答案集进一步处理,排除冗余的句子,通过相似度计算对候选句子进行排序,对相似度高的句子再进行重新分析,选取出最佳答案。答案抽取的效果会直接影响返回给用户答案的好坏。

答案抽取环节,可以通过基于表层特征的答案提取方法、关系抽取答案的方法、模式匹配抽取、统计模型抽取答案方法[19]等完成。其中关系抽取答案的方法,可以改进表层特征的答案抽取。模式匹配抽取中通过机器学习方法得到的模式比手工模式要好很多。

潘昊杰等[20]列出与提取的生物医学相关概念所属的五个数据库链接,通过得分排名得出最终答案。刘宝艳[21]先通过语义相关性计算等方法找到候选答案的中心词,再利用相似度计算去掉重复段落,最后结合命名实体标注结果提取出最终答案。温思琦[22]通过构建中医冠心病本体来增强自然语言处理技术对中医术语的处理能力,同时采用关键词模糊匹配算法和神经网络词向量的相似度算法以提升问答系统的灵活性。

3 智能问答系统在中医方面的应用

中医学是中国传统文化中最宝贵财富之一,其中蕴含了丰富的医学哲理。面对飞速发展的科学技术,中医需要走出国门、接受全球的认可,中医现代化必不可少。自 1958 年至今,中医现代化研究已开展了 50 多年[23]。将智能问答系统运用于中医领域,推动了“互联网+中医药”的产业链发展模式[24],促进了中医现代化的发展。中医智能问答系统的发展为中医行业的创新和发展提供了技术支撑,以人为中心的健康管理模式越来越贴近现实。

实现中医智能问答系统,关键是对于中医知识的解析,即系统对用户问题和中医文本能够正确理解和分析。然而中医知识与中文词语有一定区别,主要包括:①中医文本多由古汉语表示,而古代汉语常常具有通假字现象且古文之间关系复杂。②中医古文中也可能含有医家写错的文字。③中医知识也具有中文词语特有的一词多义、同义词、歧义词等比较棘手的文法现象。这些都对中医问答系统造成了特别大的困扰。研究者们更多研究的是中医的实体识别,这也是实现中医智能问答系统最基本的环节。

中医智能问答系统已经有了一些发展。中国工程科技知识中心[25]在 2012 年启动了中草药专业知识服务系统子课题[26]的建设,其主要组成部分包含了对智能问答系统的建设。丁宏娟等[27]介绍的计算机中医问诊系统主要针对临床决策,根据该系统给出的问诊初步判断,临床医生可以有计划有目的的采集信息。计算机中医问诊系统的使用可以节省医生决策时间并提高辨证的准确率。陈程等[28]将中医药知识与知识图谱以及智能问答系统相结合,系统对用户的问题采用自然语言处理技术进行分析,在交互界面中借用知识图谱展示中医药知识。

4 总结

智能问答系统应用于医学领域,使得医学信息

资源的利用率有所提高，同时也为医学工作者提供了巨大的空间和选择余地[28]。另外，医学智能问答系统的发展也使得传统医学信息搜索中以疾病为中心的服务理念有所变化：以人为本的服务理念越来越实际。

医学智能问答系统的发展，可以从以下三个方面加以完善：①医学智能问答系统需要面向普通老百姓，而不单单只是专业医术工作者，这在一定程度上会为“就医难、看病难”贡献一份力量。②国内医学名词术语标准化还存有缺乏整体规划、权威术语标准数量不足、以及更新不及时等问题。尽力使医学专业词汇统一标准化，这不仅会降低智能问答系统中本体构建的难度，也会增强答案的准确性。③借助快速发展的自然语言处理技术和深度学习技术，寻找到适合解决医学领域智能问答系统的工具和方法，使医学智能问答系统更加趋向于从语义层面深度挖掘理解用户的问题。

参考文献：

[1]康海燕，李飞娟，苏文杰.基于问句表征的 web 智能问答系统[J].北京信息科技大学学报(自然科学版)，2011，26(1)：36-41.

[2]张芳芳，马敬东，王小贤，等.国外医学领域自动问答系统研究现状及启示[J].医学信息学杂志，2017，38(3)：2-6.

[3]Lee M，Cimino J，Zhu H R，et al.Beyond information retrieval medical question answering [J].Amia Annu Symp Proc，2006：469-473.

[4]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.

[5]Wong W，Thangarajah J，Lin P.Contextual question answering for the health domain[M].John Wiley&Sons，Inc.2012.

[6]Abacha AB，Zweigenbaum P.MEANS：A medical question-answering system combining NLP techniques and semantic Web technologies[J].Information Processing& Management，2015，51(5)：570-594.

[7]Asiaee A H，Minning T，Doshi P，et al.A framework for ontology-based question answering with application to parasite immunology[J].Journal of Biomedical Semantics，6，1(2015-07-17)，2015，6(1)：31.

[8]Zhang H，Zhu L，Xu S，et al.XML-Based Document Retrieval in Chinese Diseases Question Answering System[M]. Mobile，Ubiquitous，and Intelligent Computing.Springer Berlin Heidelberg，2014：211-217.

[9]吉宗诚，徐安莹，徐飞，等.医疗领域深度问答方法及医学检索系统，CN102663129A[P].2012.

[10] 赵欣. 基于疾病圈的疾病自诊知识问答方法及系统：，CN105678065A[P].2016.

[11]Ravichandran D，Hovy E.Lerning surface text patterns for a question answering system [C]//Meeting of the Association for Computational Linguistics，Proceedings of the Conference.2002：41-47.

[12]Echihabi A，Marcu D.A noisy-channel approach to question answering[C]//Meeting on Association for Computational Linguistics.Association for Computational Linguistics.2003：16-23.

[13]张宁，朱礼军.中文问答系统问句分析研究综述[J].情报工程，2016，2(1)：32-42.

[14]王煦祥.面向问答的问句关键词提取技术研究[D].哈尔滨工业大学，2016.

[15]李冬梅，张琪，王璇，等.基于浅层句法分析和最大熵的问句语义分析[J].计算机科学与探索，2017，11(8)：1288-1295.

[16]刘凯，周雪忠，于剑，等.基于条件随机场的中医临床病历命名实体抽取[J].计算机工程，2014(9)：312-316.

[17]张芳芳，马敬东，王小贤，等.面向深度自动问答的糖尿病饮食问题分类[J].医学信息学杂志，2017，38(3)：12-16.

[18]孟洪宇，谢晴宇，常虹，等.基于条件随机场的《伤寒论》中医术语自动识别[J].北京中医药大学学报，2015，38(9)：587-590.

[19]毛先领，李晓明.问答系统研究综述[J].计算机科学与探索，2012，6(3)：193-207.

[20]潘昊杰，周芳，张博文，等.生物医学文献检索方法与问答系统[J].情报工程，2016，2(5)：50-57.

[21]刘宝艳.面向生物医学领域的问答系统的研究与实现[D].大连理工大学，2007.

[22]温思琦.基于本体的中医冠心病自动问答系统的设计与实现[D].沈阳工业大学，2017.

[23]杨云松.关于中医现代化及传统中医未来发展的思考[J].中华中医药杂志，2017(3)：920-922.

[24]陈静锋，郭崇慧，魏伟.“互联网+中医药”：重构中医药全产业链发展模式[J].中国软科学，2016(6)：26-38.

[25]谢友柏.基于互联网的设计知识服务研究——分析中国工程科技知识中心(CKCEST)的功能[J].中国机械工程，2017，28(6)：631-641.

[26]中国工程科技知识中心中草药专业知识服务系统建设专家咨询会在浙江中医药大学召开 [J]. 浙江中医药大学学报，2014，38(06)：832.

[27]丁宏娟，何建成.计算机中医问诊系统的临床验证研究[J].辽宁中医杂志，2010(11)：2138-2139.

[28]陈程，翟洁，秦锦玉，等.基于中医药知识图谱的智能问答技术研究[J].中国新通信，2018，20(02)：204-207.

收稿日期：2018-4-11；修回日期：2018-4-25

编辑/成森