中医药自动问答系统中关于问题理解的研究*

贾李蓉 高 博 刘 静 刘丽红

（中国中医科学院中医药信息研究所 北京 100700）

〔摘要〕 介绍中医药自动问答系统功能及结构，详细阐述中医药自动问答系统中自然语言问句的理解方法和流程，包括问题预处理、问题分类、关键词提取与扩展、问题模型等。

〔关键词〕 中医药；自动问答系统；问题理解

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2021.07.006

Study on the Question Understanding in the Automatic Question and Answering System of Traditional Chinese Medicine

JIA Lirong, GAO Bo, LIU Jing, LIU Lihong, Institute of Information on Traditional Chinese Medicine, CACMS, Beijing 100700, China

〔**Abstract**〕 The paper introduces the function and structure of the automatic question and answering system of Traditional Chinese Medicine (TCM), expounds the understanding method and process of natural language questions in the automatic question and answering system of TCM in detail, including question preprocessing, question classification, keyword extraction and expansion, question model, etc.

〔**Keywords**〕 Traditional Chinese Medicine (TCM); automatic question and answering system; question understanding

1 引言

随着信息化技术的发展，自动问答系统相关研究日趋增多。与传统搜索引擎相比自动问答系统更快捷地为用户提供更准确的信息，较好地弥补了传统搜索引擎的不足，将成为下一代主流搜索方式[1]。中医药自动问答系统在用户用自然语言输入具体病症描述问题后，以中医药学语言系统为词库，利用中文分词、命名实体抽取等技术将自然语言转化为规范术语表示，基于中医药学语言系统语义网络查找相关知识体系，判断用户可能患有的病症并反馈相关信息，通过与用户进一步交互进行多次反馈从而形成较为可信的结论。同时利用信息搜索功能找出相似度较高的病案输出处方建议，为用户提供辅助诊疗建议[2]。

〔收稿日期〕 2021－01－05

〔作者简介〕 贾李蓉，硕士，副研究员，发表论文 30 篇；通讯作者：刘丽红，副研究员。

〔基金项目〕 国家重点研发计划中医药现代化重点专项“多源异构中医科学数据汇交系统研发”（项目编号：2017YFC703505）；国家重点研发计划“中医药优势病种全信息自动获取与理解的关键技术”（项目编号：2019YFC1709801）；中国中医科学院基本科研业务费自主选题项目“《中国中医药学主题词表》副主题词研究及平台升级”（项目编号：zz140301）；中央级公益性科研院所基本科研业务费专项“基于本体推理的中西药物相互作用自动发现示范研究”（项目编号：ZZ13－YQ－128）。

2 中医药自动问答系统

2.1 概述

2.1.1 定义 中医药自动问答系统是以中医药学语言系统为基础，结合中医药领域专家知识，利用自然语言处理、知识图谱、自动问答、信息检索等技术构建的一套智能知识问答系统。

2.1.2 功能 本系统以常见病“月经病”为示范，采集期刊文献、临床医案、相关教材数据构建数据库，将中医药学语言系统中的知识本体、专业概念术语、语义关系等规范知识表示与采集数据库有机结合，构建由病、证、症、方、药等知识实体构成的知识图谱，通过自动问答技术为用户提供更为精准的信息资源。

2.2 系统构成

自动问答系统主要由问题理解、信息检索和答案生成3部分组成，见图1。其中问题理解是系统第1步骤，也是最重要环节，对用户分析理解的精度和准确度将直接影响后续检索环节，继而影响最终给出的答案[3]。

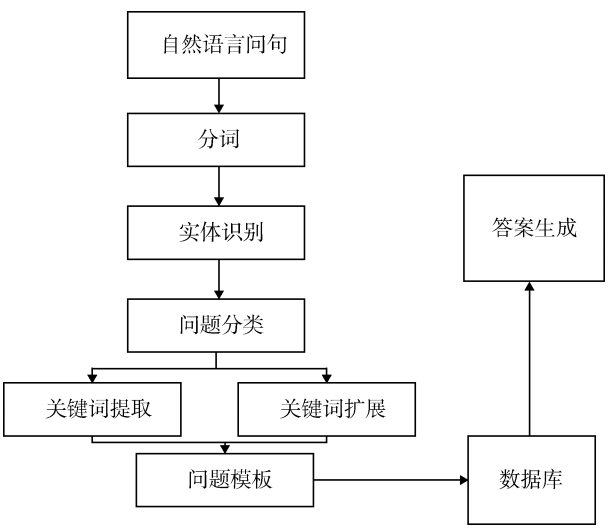

图1 自动问答流程

3 问题理解研究

3.1 定义

问题理解是指利用分词等技术将用户输入的自然语言转化为计算机可识别、理解的结构化语言的过程。该过程包括智能分词、实体识别、问题类型确定、问题关键词提取等，对后续信息检索和返回答案具有重要影响[4]。

3.2 问题预处理

3.2.1 概述 问题预处理阶段是指用户输入自然语言的问句后，问答系统处理用户查询语句的部分，包括智能分词与实体识别两个阶段。

3.2.2 智能分词 用户输入的自然语言问句是未经过分词的句子，将问句准确分词是问题理解的基本保障。系统使用开源汉语分词工具，与中医药学语言系统基础词库相结合，将新生成的特殊词和词族添加至分词词典中并保持更新[5]。分词过程可能存在切分歧义，如句子“大黄粉碎使用”正确切分为“大黄/粉碎/使用”，可能被错误切分为“大黄粉/碎/使用”。因此需要利用专业知识和专业领域词典解决分词歧义。在正确分词基础上利用马尔科夫链方法（基于规则和统计的方法）进行词性标注。基于马尔科夫链随机过程的n元语法统计分析方法，被证明在词性标注中能达到较高精度。

3.2.3 实体识别 命名实体识别模块主要负责识别语句中的实体名并加上特殊标记[6]。中医药自动问答系统基于Ansj分词自定义词典的方法进行命名实体识别，通过自定义词典识别出用户问句中的实体并进行词性标注，方便进行关键词分析和扩展。

3.3 问题分类

3.3.1 分类依据 根据统计和分析发现用户输入问题可以分为若干类型，利用不同方法处理不同类型问题。因此一般问答系统中都含有问题分类环节[7]。在中医药自动问答系统中，根据中医药行业主要涉及内容和用户最可能提问的问句将可能问句进行问题分类，见表1。

表1 主要问题举例

问题类型	疑问句举例	例子
询问疾病	什么病，怎么了，为什么…	经期小腹疼痛是什么病？
询问方药	什么药，什么方，哪些药…	月经不调用什么药治疗？吃什么药能缓解经期腹痛？
询问疗法	怎么治疗，如何治疗…	如何治疗气血虚弱性月经不调？
其他类型	–	–

3.3.2 分类方法 问题分类主要依靠句中特定词和疑问词，不同疑问词和特定词对问题分类的区别能力不同。例如出现“功效”“疗效”等特定词，可判断问题类型为询问药效；出现“治疗”等特定词，判断问题为询问疗法；出现“什么”“怎么”“为什么”“如何”等疑问词，则需要配合其他特定词来确定问题类型。

3.3.3 分类规则 针对不同问题类型需要制定相应规则，规则将在后续关键词提取和扩展、信息检索、答案抽取等阶段执行。例如对于询问方药问题判定，可以规定答案中必须含有“方剂”“中药”类的词。对于不容易划分问题可统一划归其他类型，此类问题无法制定详细规则，可以采用检索结果的前几句作为问题候选答案。

3.4 关键词提取

3.4.1 停用词 关键词提取直接影响后续检索结果，需在用户输入的自然语言中提取出对后续检索有价值的关键词。用户问句中的词并不都可以作为关键词，如疑问词、语气词、助词等一般不表示关键含义，通常在关键词提取过程中首先被过滤并汇集形成停用词表。

3.4.2 关键词 主要由名词、动词、形容词、限定性副词等组成。本研究将除疑问词以外的大部分词列为关键词。关键词分成两种：一般性关键词和“必须含有”的关键词。“必须含有”的关键词必须在答案句子中含有，而一般性关键词可以不被句子所包含。关键词被赋予不同权重，检索时以计算句子权重。

3.5 关键词扩展

3.5.1 关键词同义扩展 中医药知识中部分概念存在同义异名现象。此现象也存在于用户提问所用关键词中。用户问句所使用的关键词有时不是知识库答案中的词语，而是与其同义。例如用户提问“经期肚子疼是怎么回事”。其中“肚子疼”是常用口语表述，而在中医药知识库中所储存的同义表述“小腹痛”。需要将“肚子疼”同义对应到“小腹痛”概念上才能从知识库中获得答案“经期小腹痛是由于气血失调引起的”，否则将造成关键词查询失败。因此需要对关键词进行适当同义扩展以提高系统召回率[8]。中医药自动问答系统使用中医药学语言系统基础词库，包含大量中医药概念术语的同义对应，见表2。

表2 关键词同义扩展举例

疾病名称	同义词
月经先期	经期超前；经早
经期延长	经事延长
痛经	经行腹痛；原发性痛经
经断前后诸证	更年期综合征；绝经前后诸证
月经先后无定期	经乱；经水先后无定期；月经愆期
子宫内膜异位症	内异症；盆腔子宫内膜异位症
月经过少	经量过少；经水涩少
经行感冒	触经感冒
经行浮肿	经来遍身浮肿
经间期出血	排卵期出血
闭经	经闭；经水不通；女子不月；月事不来
经断复来	绝经后出血；年老经水复行
经行发热	经来发热
经行风疹块	经行瘖瘾；经行瘾疹
经行吐衄	倒经；逆经
经行泄泻	经来泄泻
月经过多	经水过多；月水过多
月经后期	经迟；经行后期；月经落后；月经延后

3.5.2 根据问题类型扩展关键词 某些类型问题所对应答案中经常会出现某种共同特征的词，例如询问病情的问题中经常出现“什么”“怎么”“为什么”等关键词，将此类词作为关键词进行扩展，见表3。

表 3 根据问题类型扩展的关键词举例

问题类型	扩展的关键词
询问病情	什么、怎么、为什么……
询问方药	什么、哪些……
询问疗法	怎么、如何……

3.6 问题模型

经过对问句理解以及多轮文本训练和规则修改最终形成“问疾病”“问证候”“问治疗方法”“问方剂”“问中药”“问病案”“问疾病简介”7 个主要问题模型，总结归纳 7 类问题模型最常见问句，见表 4。用户自然语言问句通过问句预处理、问题归类、关键词提取和扩展、归于问题模型等步骤进入知识库中检索并返回答案。

表 4 问句分类

问题类型	问句示例	说明
问疾病	最近 symptom 可能得什么疾病；什么疾病有 symptom 这些症状；symptom 是什么疾病的病症；symptom 是什么疾病；symptom 是什么病；symptom 什么病；symptom 什么疾病	symptom：表示用户输入的具体症状名词
问治疗方法	最近 symptom 怎么治疗；symptom 用什么方法治疗；symptom 怎么治疗；什么方法可以治疗 symptom；symptom 的治疗方法；symptom 治疗；symptom 怎么治疗；disease 这种病怎么治疗；什么方法可以治疗 disease；disease 这种病用什么方法治疗；disease 的治疗方法；disease 治疗；disease 怎么治疗	symptom：表示用户输入的具体症状名词；disease：表示用户输入的具体疾病名称
问病案	跟 disease 相关的病案有哪些；可作为 disease 参考的病案举例；disease 相关病案；disease 的参考病案；disease 参考病案；disease 的病案；disease 病案	disease：表示用户输入的具体疾病名称
问疾病简介	disease 相关的简介；什么是 disease；disease；disease 相关介绍；disease 介绍；disease 的简介；disease 简介；什么是 disease 的介绍	disease：表示用户输入的具体疾病名称
问中药	disease 的中药方法；disease 用什么中药；disease 的中药治法；disease 相关中药方法；什么中药能治 disease；能治 disease 的中药；symptom 怎么用中药治疗；能治 symptom 的中药；有 symptom 用什么中药；symptom 中药治疗；symptom 的中药方法；symptom 用什么中药；symptom 的中药治法；symptom 相关中药方法；什么中药能治 symptom	disease：表示用户输入的具体疾病名称；symptom：表示用户输入的具体症状名词
问方剂	disease 这种病用什么方剂治疗；disease 用什么方剂治疗；什么方剂可以治疗 disease；disease 相关的方剂	disease：表示用户输入的具体疾病名称
问证候	symptom 的证候是什么；symptom 证候表现；symptom 是什么证候；symptom 症状是什么证候的表现；什么证候有 symptom 症状	symptom：表示用户输入的具体症状名词

4 结语

自动问答系统是否好用的关键在于能否为用户提供简单且准确的答案，其主要前提是对问题有较好的理解。本文详细介绍中医药自动问答系统中自然语言问句的理解方法和流程以及独有问题模型，为实现系统对用户问题的精准理解提供参考。

参考文献

1 文勖. 中文问答系统中问题分类及答案候选句抽取的研究［D］. 哈尔滨：哈尔滨工业大学，2006.

2 贾李蓉，刘丽红，刘静，等. 基于中医药学语言系统的知识问答系统的设计与构建［J］. 中华医学图书情报杂志，2019，28（5）：11－14.

3 文勖，张宇，刘挺，等. 基于句法结构分析的中文问题分类［J］. 中文信息学报，2006（2）：33－39.

4 冶忠林，贾真，尹红风. 多领域自然语言问句理解研究［J］. 计算机科学，2017，44（6）：216－221，254.

5 胡金凤. 中文问答系统中的问题分析与答案抽取［D］. 石家庄：石家庄铁道大学，2016.

6 罗鹏，张宾，高亚斌，等. 中文知识图谱在商品咨询自动问答系统中的应用［J］. 自动化技术与应用，2020，39（6）：173－176.

7 宋文闯，刘亮亮，张再跃. 问答系统中问句相似度研究［J］. 软件导刊，2020，19（7）：148－152.

8 胡松江. 基于知识图谱的问答系统问句理解方法研究［D］. 成都：西华大学，2019.