情报工程 ISSN 2095-915X
第2卷 第6期 39-49 2016年12月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.6 39-49 Dec 2016

doi:10.3772/j.issn.2095-915x.2016.06.005

中文健康问句分类与语料构建

郭海红，李姣，代涛

（中国医学科学院医学信息研究所 北京 100020）

摘要：本文旨在构建一个中文健康问句分类方法，并通过对高血压相关的健康问句进行人工分类标注，分析公众的高血压相关健康信息需求，同时为研发高血压相关的智能中文问答系统提供语料基础。本研究基于临床问句分类及公众健康信息查询场景层次模型，构建一个四级中文健康问句主题分类方法，并由5位标注员独立地对从某中文健康网站上收集的将近10万条高血压相关提问数据中随机抽取的2000条样本数据进行人工分类标注，以优化和测试该问句分类方法的可靠性，构建标注语料库，并分析公众的高血压相关健康信息需求。5位标注员使用该分类方法进行独立标注的四级类目评判者间信度kappa值为0.63，意味着分类结果可靠，一级大类获得高度一致性（kappa=0.82），略优于国际上的同类研究。分布在治疗、诊断、健康生活方式、临床发现/病情管理、流行病学、择医六个一级类别中的问句分别占样本总量的48.1%、23.8%、11.9%、5.2%、9.0%和1.9%。所构建的健康问句分类方法可用于组织大型健康问题集，以提高检索效率；分类标注的样本问句可作为高血压相关健康问句自动分类研究的语料；得出的高血压相关健康问句主题分布有助于指导健康网站的知识资源建设。此外，所设计和采用的问句分类方法构建方式、语料标注流程、评判者间信度测量方法等，也可为开放领域及其他受限领域开展用户问句分类与语料构建提供借鉴。

关键词：健康问句，问句分类，语料构建，公众健康，信息需求

中图分类号：G202

Question Classification and Corpus Construction of Chinese Health

GUO HaiHong, LI Jiao, DAI Tao

(Institute of Medical Information, Chinese Academy of Medical Science, Beijing 100020, China)

Abstract: This study aimed to build up a Chinese health question classification schema and manually annotate hypertension related health question, so as to understand and specify hypertension related informational needs of

基金项目：本文受中国医学科学院中央级公益性科研院所基本科研业务费课题：中文公众健康问句分类与健康信息需求挖掘研究（2016ZX330011），国家社会科学基金资助项目：面向知识服务的健康知识组织体系构建研究（14BTQ032）的资助。

作者简介：郭海红（1987–），硕士，研究实习员，研究方向：自然语言处理、问答系统、公众健康信息学；李姣（1981–），博士，副研究员，研究方向：文本挖掘、语义网络、生物医学信息学，公众健康信息学；代涛，通讯作者，博士，研究员，研究方向：医学信息学、公众健康信息学、卫生政策与管理，E–mail：dai.tao@imicams.ac.cn。

the users, and further to lay a corpus foundation for hypertension related smart Chinese question and answering (QA) system. This paper built up a four-level classification schema of health questions based on taxonomies of generic clinical questions and a layered model of context for consumer health information searching. Five annotators independently and manually classified 2000 questions which were randomly selected from nearly 100 thousand hypertension-related messages posted on a Chinese health website to modify and test the reliability of the schema, as well as to build an annotated corpus for Chinese health QA system and to analyze the hypertension related information needs of health consumers. The results showed the kappa statistic for five annotators who independently annotated with the schema on the fourth level was 0.63, indicating "substantial" reliability, and reached “almost perfect” reliability (kappa=0.82) on the first level, which was slightly better than the similar studies oversea. Questions in the categories of treatment, diagnosis, healthy lifestyle, management, epidemiology, and health provider choosing were 48.1%, 23.8%, 11.9%, 5.2%, 9.0%, and 1.9% respectively. This study will do help to organize large collections of health question so as to improve retrieval efficiency, to train machine to automatically classify topics of hypertension related questions posted by health consumers, to guide the building of knowledge base of health websites. Besides, the methods for building the question classification schema, the procedure of corpus annotation, and the methods for evaluating the inter-rater reliability that we designed in this research can provide reference for studies about user question classification and corpus building in open domain and other restricted domain.

Keywords: Health questions, question classification, corpus building, consumer health, information needs

1 引言

随着社会经济的发展和生活水平的提高，人们对健康问题越来越重视，对健康信息的需求越来越大。随着网络通信技术的快速发展和普遍应用，互联网日益成为人们获取健康信息的重要渠道。截至 2015 年 12 月，我国互联网健康医疗用户规模为 1.52 亿，占网民的 22.1%[1]。高血压是最常见的慢性病，已成为威胁中国公众的主要因素之一，它是超过一半的心血管疾病（如脑卒中、冠心病等）的危险因素。据估计，2012 年中国有 2.7 亿高血压患者，即平均每 10 个成年人中至少有 2 人患有高血压，且其年发病率约为 3%[2]。相应的，公众对高血压相关的健康信息需求也日益增加，在互联网上提出的高血压相关问题也日益增多，且涉及该疾病的多个方面。以“高血压”为主题来分析用户的健康信息行为，发掘其健康信息需求，进而改善在线健康信息服务，已成为信息资源建设与服务领域的重要研究内容 [3]。

与传统的搜索引擎相比，自动问答系统能更好地满足用户从互联网上快速、准确获取信息的需求，它允许用户以自然语言形式提出问题，并采用自然语言处理技术自动地将简洁、正确的回答返回给用户，是目前自然语言处理和信息检索领域的一个研究热点 [4]。无论是 iPhone 手机上的热门 APP Siri，还是在美国电视竞答节目 Jeopardy 中打败人类冠军的 IBM Watson 系统，都与自动问答系统直接相关。在健康医疗领域，比较著名的自动问答系统有基于循证医学原则提供分级推荐意见的临床决策支持系统 UpToData[5]，基于 MEDLINE 生物医学文献数据库和网络数据为用户返回段落级答案的临床问答系统 AskHERMES[6] 和生物医学问答系统 GeneView[7] 等。针对公众的高血压相关健康信息需求，研发高血压相关的智

能中文问答系统将具有重要的社会应用价值。

问答系统一般包括问句分析、信息检索和答案抽取 3 个主要部分。问句分类作为问答系统所要处理的第一步，起着至关重要的作用，它有助于减少候选答案空间，提高系统返回答案的准确率，以及根据问句类型调整答案选择策略等 [8-10]。

2 问句分类相关研究

分类法是对问句进行分类的依据。当前的问句分类法按其分类依据主要分为 3 种 [9]：基于答案类型的问句分类，基于问句语义信息的问句分类，以及基于混合信息的问句分类。在开放领域，针对英文问句比较权威的分类是基于答案类型的 UIUC 层次分类 [11]，包括缩写、实体、人物、地点、描述和数字 6 大类，及其下属的 50 个不重复的小类。哈尔滨工业大学的文勖等 [12] 根据中文自身特点，构建了中文开放领域问句分类，包括人物、地点、数字、时间、实体、描述和未知 7 个大类，共 60 个小类。

具体到健康医疗领域，因其内容边界明确（Circumscribed）而又复杂多样（Complex）[13]，开放领域的问句分类难以满足其分类需求 [8]，还需要面向领域知识的主题分类。目前已有一些针对专业医学问句的分类，如进行医学主题分类的国际初级保健分类法（International Classification of Primary Care）[14] 和进行临床问句主题与形式分类的一般临床问句分类（Taxonomies of Generic Clinical Questions）[15] 等。这些分类已被证明适用于医务工作者的临床问句分类与信息需求分析 [16,17]，却不能直接用于公众的健康问句分类，因为公众的健康问句与医务工作者的临床问句有明显的区别 [18-20]。Yan Z[21] 构建的公众健康信息查询场景层次模型（Layered Model of Context for Consumer Health Information Searching）在认知层分类描述了公众认知到的感兴趣的话题，Roberts K 等 [22] 根据美国国立医学图书馆遗传和罕见病信息中心网站上的 1467 条公众提问信息，制定了一个分类规则，将健康问句分为解剖、病因、临床表现、诊断、管理、并发症、疾病的其他影响、预后、人群易感性、疾病的一般性信息、疾病的其他信息、人员与机构和非疾病信息 13 个类别。但这些分类尚未形成一个完整而系统的健康问句主题分类。另外，由于中文问句的语义语法更加灵活、表达更具有随意性、问句的形式更加复杂化等原因 [23,24]，使得这些针对英文问句的研究成果也不能直接应用于中文场景。因此，本研究旨在构建一个简洁可靠的中文健康问句分类，并通过对高血压相关健康问句的人工分类标注，构建标注语料库，进而为智能化的互联网用户高血压相关健康信息需求分析和研发高血压相关智能中文健康问答系统奠定基础。

3 资料与方法

3.1 数据收集

本研究所分析的高血压相关健康问句全部来自寻医问药网（www.xywy.com）。该网站成立于 2004 年，是目前中国领先的一站式互联网医疗服务平台，也是中国较早探索和实践互联网医疗服务的平台之一。截至 2015 年 6 月底，其注册用户超过 1 个亿，日独立访客超过 2000 万，月独立访客超过 3 亿，位居医疗健康服务行业前列 [25]。本研究收集了 2014 年 1 月 1 日至 8 月 10 日发布到寻医问药网有问必答栏目下标签为“高血压”或“血压”的问答信息，并导入数据库。根据提问内容去重后，共获得 98,032 条问答信息。为了对这些信息所包含的问句进行深入分析，本研究从库中随机抽取了 2000 条样本问句信息。

本研究中的“健康问句”是指公众（非医务工作者）发布在健康网站上的有关某个特定领域

的请求，以引出医务工作者的知识（回答）。应基于含义来判断，而不是问句的形式。另外，本研究聚焦于有关高血压的问句，而“高血压”有时会被表达成“血压高”，或简称为“高压”。因此，通过人工审核，剔除样本信息中与“健康问句”的定义不符，或者具有“高压”等形近词，但实际与高血压无关的问句，如含“高压氧”、“高压舱”、“高压锅”、“高压电”等词的信息。同时，以提问内容排序，对样本问句中残留的重复问句进行人工剔除。每剔除一条不符合要求的信息，同时从数据库中随机抽取一条问句信息，以确保样本数据的规模为 2000 条。

3.2 中文健康问句分类方法构建

基于前述一般临床问句分类 [15]、公众健康信息查询场景层次模型 [21] 和公众健康问句分类标注规则 [22] 预先构建一个基于主题的健康问句分类。具体来说，该分类中与临床相关的类别（包括诊断、治疗、病情管理、流行病学及其下位类）主要选自一般临床问句分类和公众健康问句分类标注规则，而非临床的类别（如健康生活方式、择医及其下位类）则主要选自公众健康信息查询场景层次模型。在此基础上，将一些类别进行细分，以便标注更为具体的健康信息需求。例如，参照药物治疗的下位类，将健康生活方式大类下的二级类目“饮食”细分为 5 个三级类目，分别为“饮食方式”、“食物选择”、“食物间相互作用”、“作用机制”、“一般性的，不具体的问题”。

具有医学和医学信息学专业背景的一位研究人员利用该分类对 2000 条样本数据进行预标注。在标注过程中新增部分类别，以适应那些未能归到既有类别中的问句。本研究制定了一系列的标注规则，并为每个最细的类目列举一些一般问句形式，以提高该分类的可用性和评判者间的信度。例如，针对 1.1.X.X（诊断→病因 / 临床发现的解释）类的问句，标注规则如下：若提问者提及一系列临床发现，想知道它们是由什么情况引起的（即知道临床发现，但不知道是什么情况），则归为该类。其细分类别的标注规则为：①若只提到一系列症状（如头晕、恶心、视物模糊等病人主观感受），询问是怎么回事，则归为 1.1.1.1（诊断→病因 / 临床发现的解释→症状）；②若只提到一系列体征（如心脏杂音，肺部罗音，血压升高（或血压值），反射异常等可通过体格检查发现的现象），询问是怎么回事，则归为 1.1.2.1（诊断→病因 / 临床发现的解释→体征）；③若只提到一系列实验室检查结果（如血糖值、甘油三酯值、心电图等），询问怎么回事，则归为 1.1.3.1（诊断→病因 / 临床发现的解释→检验检查发现）；④同时提到以上两类或两类以上，则归为 1.1.4.1（诊断→病因 / 临床发现的解释→不具体的发现或多种发现），该类问句的一般形式包括但不限于：给出临床发现 X1，X2，……，Xn，询问：

- 这是病态吗？ / 这是不是病？
- 是不是有情况 Y？ / 是不是由情况 Y 引起的？
- 这种情况严重吗？ / 要紧吗？
- 有问题么？ / 有什么问题？ / 有没有大问题？
- 这是怎么了？ / 这是怎么回事？ / 想知道怎么啦？ / 是怎么回事？ / 是什么情况？ / 是如何情况？
- 这些危害大吗？ / 是属于正常范围中吗？
- 是什么病症？ / 是什么病呢？ / 是啥病？
- 为什么会（这样）？ / 怎么会（这样）？

对于没有提问语句或提问词的问句，制定了以下分类规则：①仅提供临床表现或病情，诊断未知，归为 3.1.1.1（对病情 / 发现的管理）；②提供一个或多个疾病名称则视为诊断已知，未提及治疗措施，或提到的治疗措施不只包括药物治疗，归为 2.2.1.1（疾病治疗，包括但不限于药物

治疗）；③诊断已知，且仅提及药物治疗，归为2.1.2.1（疾病药物治疗）；④诊断已知，提及用药一段时间后恢复正常，归为2.1.1.3（用药时间选择），意指是否可停药。

通过这种方式，构建了初步的中文健康问句分类方法，包括101个主题类别和32条标注规则，由于篇幅所限，不在此逐一列举。

3.3 分类方法完善与问句语料标注

本研究通过以下步骤完善该中文健康问句分类方法，并进行健康问句语料分类标注：

首先，由另外4名标注人员（其中2人具有医学教育背景，另2人具有医学信息学背景）分别利用该分类方法，根据其类别划分、各类别标注规则与一般问句形式，独立地对从样本问句信息中随机抽取的200条问句信息进行分类标注。在标注过程中，每位标注员都提出了一些细化规则、增加部分类目以适应问句等的改进建议。作者比较了5份标注结果（包括3.2中预标注的一份），并将这200个已标注的问句分为3类：一是所有标注结果一致的问句（n=73），二是只有一个标注结果不一致的问句（n=63），三是有2个及以上标注结果不一致的问句（n=64）。再聚焦第三类，重点分析那些标注很不一致的问句，并通过细化标注规则和进一步准确描述作为示例的一般问句形式等方式来减少模糊因素。

然后，由前述5名标注人员分别利用改进后的分类方法独立地标注从余下的1800条样本问句信息中随机抽取的300条问句信息。本轮标注的目的主要是测试该分类方法的评判者间信度（interrater reliability），同时对其进行进一步的完善。

最后，由3名标注人员（其中2人具有医学教育背景，1人具有医学信息学背景）完成余下的1500条问句信息的标注，即每人独立标注500条。至此，加上前述为完善分类方法所标注的200条问句和为测试评判者间信度所标注的300条问句，全部2000条样本问句信息中的每一条都至少已被2人标注。作者对标注结果进行比较分析，对少数不一致的标注通过讨论达成了一致。然后，统计了每个类别下的问句数量，并将没有问句归为该类的类别剔除（如药物的物理特性、药效动力学、药物作用机制、医患沟通等）形式。

3.4 统计分析

采用描述性分析统计样本问句在各类别中的频率。采用可纠正随机一致性的kappa值来确定健康问句分类方法的评判者间信度。Kappa = (Po−Pe)/(1−Pe)，其中Po为观察到的一致度，Pe为随机一致性的期望值[26]。当类别数量较多，如本研究，则Pe趋近于0，kappa值趋近于Po。因此，直接将Po作为kappa值。Kappa值越大，一致性越好。

本研究假设，当用户提出多个问题时，回答其中的任何一个问题都是可以接受的。因此，采用了较为宽松的可靠性标准：只要某标注员所标的主要编码或次要编码中的任何一个与另一标注员所标的主要编码或次要编码中任何一个相同，则视为标注一致。

4 结果

4.1 中文健康问句分类

最终形成的中文健康问句分类是一个四级分类表，四级细目由最初的101个降至48个（表1），并设有35条标注规则。一级类目包括7个大类，分别为诊断、治疗、病情管理、流行病学、健康生活方式、择医及其他。病情管理类主要指那些询问该怎么办却又未特指诊断或治疗措施的问句[15]，

即这类问句既询问诊断又询问治疗。要回答这类问题，需先给出病情诊断，再提出治疗建议。二、三、四级类目所描述的主题越来越具体、细致。每个四级类目后都列出了一个或多个一般问句形式，如表 2 列举了类别“治疗→药物治疗→药物选择 / 适应症 / 效力→治疗”的部分常见问句。

表 1 中文公众健康问句分类

代码	一级类目	二级类目	三级类目	四级类目	频率（%）（主编码）	频率（%）（全编码）
1.1.1.1	诊断	病因 / 临床发现的解释	症状		39(2.0)	44(1.7)
1.1.2.1			体征		146(7.3)	152(5.8)
1.1.3.1			检验检查发现		7(0.4)	8(0.3)
1.1.4.1			不具体的发现或多种发现		286(14.3)	302(11.6)
1.2.1.1		标准 / 临床表现			35(1.8)	37(1.4)
1.3.1.1		检验检查	适应症 / 效力		36(1.8)	55(2.1)
1.3.2.1			精确度		4(0.2)	6(0.2)
1.3.3.1			时间选择 / 监测		6(0.3)	6(0.2)
1.3.4.1			方法		4(0.2)	5(0.2)
1.4.1.1		介绍	情况		4(0.2)	5(0.2)
1.5.1.1		费用			0(0.0)	2(0.1)
2.1.1.1	治疗	药物治疗	用药方式	一般性的	4(0.2)	7(0.3)
2.1.1.2				剂量	8(0.4)	11(0.4)
2.1.1.3				时间	38(1.9)	52(2.0)
2.1.2.1			药物选择 / 适应症 / 效力	治疗	324(16.2)	387(14.8)
2.1.2.2				预防	3(0.2)	7(0.3)
2.1.3.1			副作用	药物引起的	29(1.5)	46(1.8)
2.1.3.2			副作用	对副作用的管理	2(0.1)	5(0.2)
2.1.3.3			副作用	安全、禁忌症	23(1.2)	27(1.0)
2.1.4.1			互作用		20(1.0)	25(1.0)
2.1.5.1			名称查找		1(0.1)	2(0.1)
2.1.6.1			费用		1(0.1)	1(0.0)
2.1.7.1			可及性		1(0.1)	1(0.0)
2.1.8.1			品牌 / 生产厂家		2(0.1)	3(0.1)

代码	一级类目	二级类目	三级类目	四级类目	频率（%）（主编码）	频率（%）（全编码）
2.2.1.1		不限于但可能包含药物治疗	效力、适应症	治疗	473(23.7)	621(23.8)
2.2.1.2				预防	7(0.4)	16(0.6)
2.2.2.1			时间		4(0.2)	8(0.3)
2.2.3.1			治疗方式		1(0.1)	1(0.0)
2.2.4.1			安全 / 禁忌症 / 后遗症		12(0.6)	26(1.0)
2.2.5.1			费用		2(0.1)	8(0.3)
3.1.1.1	管理	病情 / 发现			114(5.7)	136(5.2)
4.1.1.1	流行病学	患病率 / 发病率			0(0.0)	1(0.0)
4.2.1.1		病因学 / 病原学	因果关系 / 相关性	危险因素 / 病原	111(5.6)	149(5.7)
4.2.1.2				遗传学	3(0.2)	4(0.2)
4.3.1.1		进程 / 预后			51(2.6)	82(3.1)
5.1.1.1	健康生活方式	饮食	饮食方式		4(0.2)	4(0.2)
5.1.2.1			食物选择	效力	69(3.5)	97(3.7)
5.1.2.2				禁忌症	29(1.5)	40(1.5)
5.1.3.1			相互作用		2(0.1)	4(0.2)
5.1.4.1			未特指的		19(1.0)	32(1.2)
5.2.1.1		运动			7(0.4)	15(0.6)
5.3.1.1		减肥			3(0.2)	3(0.1)
5.4.1.1		压力 / 情绪管理			2(0.1)	3(0.1)
5.5.1.1		未特指的			35(1.8)	107(4.1)
6.1.1.1	择医	医疗机构选择			10(0.5)	18(0.7)
6.2.1.1		医疗科室选择			11(0.6)	25(1.0)
6.3.1.1		医生选择			3(0.2)	6(0.2)
7.1.1.1	其他				5(0.3)	5(0.2)
合计					2000(100)	2607(100)

表 2 有关药物选择 / 适应症 / 效力的一般问句形式

病情 Y/ 情况 Y/ 临床发现 Y:	
该吃啥药?	有何 X 类药进行调理?
服用什么药?	用 X 类药物来调吗?
内服什么药合理?	X 类药物有很多种，怎样选?
能吃什么药（？）	吃药物 X，不吃别的药行吗?
吃什么药（好？）	不服药物 X 怎么控制病情 Y？
用什么药调比较好?	药物 X 是用来效果 E 的药物吗?
能不能吃药物 X？	吃药物 X 真能效果 E（如降糖降压）吗?
能否吃药物 X？	药物 X 可引起效果 E（如血压降低）吗?
可以吃药物 X 吗?	吃药物 X 有用吗?
可否用药物 X1，X2，…，Xn 调理一下?	药物 X1，X2，…，Xn 有 E 的功效吗?
还能继续服用（药物 X）吗?	吃什么效果 E（如降压）（的）药比较好?
药物 X（效果）怎么样?	吃什么效果 E 最快?
药物 X 管用不管用?	如何控制病情 Y/ 情况 Y/ 临床发现 Y？（提到了药物）
药物 X 是否能够根治病情 Y 呢?	请问用什么方法来调理？（提到药物类别）
吃什么 X 类药比较好?	

4.2 中文健康问句标注语料与一般主题分布

2000 条样本问句信息被标注为 2000 个主要主题和 607 个次要主题（表 1）。48% 的问句是关于治疗的，其中将近一半（45.9%）是特指药物治疗的，包括用药方式（5.6%）、选药方法（即药物效力 / 适应症，31.5%）、药物副作用（6.2%）、以及某种药可否与其他药一起用（即药物相互作用，2.0%）等。

23.8% 的问句是关于诊断的，其中绝大部分（19.4%）是用户希望得到有关其（或其关心的人）实际临床发现的解释，包括所感觉的症状（1.7%），从体检结果中得知的体征（5.8%），医院检查结果（0.3%），或他们所知的多种发现（11.6%）。有 5.2% 的问句被标注为 3.1.1.1（对病情或临床发现的管理），因为它们未特指是有关诊断的，还是有关治疗的。其中，超过一半的问句（54.4%）仅仅列举了一系列的临床发现，而没有任何提问句或提问词。

11.9% 的问句询问在日常生活中应该做什么或怎么做，以保持健康或帮助从特定的病情中恢复健康，其中超过一半（58.4%）是关于饮食或营养品的。有 9.0% 的问句是有关疾病流行病学的，关于危险因素的问句占总量的 5.7%，既包括其所患疾病的危险因素，也包括他们的病情是否会影响其他特定情况，如怀孕、分娩、母乳喂养、性生活、拔牙等。有 3.1% 的问句是关于预后的，但这些问句更像是在表达焦虑，而不是寻求真正的答案，提问者希望得到肯定的回答，以减轻其担忧 [21]。在占 1.9% 的择医问句中，有一半是关于在特定情况下应该选择什么科室就诊，提示开展就医导诊可能是健康网站的一个有前景的业务。

4.3 中文健康问句分类的评判者间信度

5 名标注员利用该分类进行高血压相关中文健康问句分类标注，在四级类别上评判者间信度 kappa=0.63，意味着“实质的（Substantial）”可靠性，超过了一些同类研究，如为一般临床问句进行主题标注（kappa=0.53）[15]。当仅考虑第一、

二级类目时，Kappa 值提高到 0.75。当仅考虑一级大类时，达到“几乎完美（Almost Perfect）”的一致性（kappa=0.82），略优于 Roberts K 等对英文公众健康问句的分类标注（kappa=0.80）[22]。

5 讨论

5.1 主要发现

本研究发现，尽管公众所询问的健康相关问题体量非常大，这些问句的一般主题却是有限的，并且每个主题类别都有一些特定的问句形式。这一发现表明可以用有限的主题类别和关键词来表示量大而多样的健康问句 [27]。另外，公众比较倾向于一次性提问多个问题，且往往涉及某个疾病的多个方面，即该问句信息需标注多个类别 [18,28]（示例如表 3）。我们通过对 2000 条样本问句信息的分类标注发现，有 26.35% 的问句信息被标注了 2 个及以上类别（表 4）。这一特点提示，进行健康问句分解，即提取出提问信息中的每一个健康问句，是使机器理解并进而回答健康问句的一个基本步骤 [29]。

表 3 一次提问多个问题的健康问句示例

提问信息	主要分类	次要分类
您好，我妈妈今年 48 岁，有高血压 1 年了，最近两天嘴皮发麻，有时候说不出话，一阵子一阵子的，还伴有头晕的状况，去检查确诊轻微的脑梗塞，请问要怎样治疗，平时需要注意些什么呢，心情是不是也非常重要啊？之前姥爷也脑梗，是遗传么？	2.2.1.1	5.5.1.1 5.4.1.1 4.2.1.2

表 4 健康问句分类标注类别数量分布

类别数	提问信息数	占比（%）
1	1473	73.65
2	460	23.00
3	54	2.7
4	13	0.65
合计	2000	100

关于提问主题，通过与 Ely 等 [15] 标注的 1396 个临床问句进行比较，发现公众与医务工作者都询问与诊断、治疗、病情和临床发现管理、流行病学有关的问题。除此之外，公众还询问有关健康生活方式的问题，因为人们已经意识到饮食、锻炼、减肥、情绪控制等也会影响其健康状况 [30,31]。医务工作者在给病人进行诊疗的过程中却很少询问此类问题，因为他们主要关注医疗服务而不是生活方式建议 [32]。同样，公众也从不询问有关与其他医务工作者协调、管理制度、伦理与法律等方面的问题，因为这些问题通常被认为是医务工作者的职责。

对于医学问题，公众与医务工作者都希望获得对于某种或多种临床发现的解释，但公众所提的问题要模糊得多，其包含多种发现的问题的频率超过医务工作者的相应问题的 2 倍。可能是因为公众难以判断哪种发现是最重要的，所以他们倾向于把所有知道的发现都列举出来，以帮助医务工作者诊断。尽管公众与医务工作者所提问题中，有关治疗的问题的频率基本相当（48.2% V.S. 43.7%），但医务工作者所提的问题更聚焦于药物治疗（37.2% V.S. 22.1%），并且他们有时候会询问一些特别专业的相关医学问题，如药物的物理特征、组成、药效动力学、作用机制、血清水平等 [15]，公众则很少问此类问题。这些发现再次证明，公众的健康信息需求与医务工作者的临床信息需求存在明显的差异，同时提示面向公众的智能健康问答系统研发过程中的健康问句处理需要独特的分类方法，不能直接套用既有的临床问句分类方法 [18]。

5.2 本研究的不足

本研究的主要不足在于对高血压相关样本问句的选取具有来源单一、样本量较小的局限性，即样本问句仅选自一个健康网站，且由于时间精力所限而仅随机抽取了 2000 条样本问句进行分类标注。因此，所构建的中文健康问句分类、相应的标注规则及标注语料仍有待进一步通过更大规模的、多样化的样本问句进行测试、完善和丰富。但是，千里之行始于跬步，我们希望本研究能起到一个抛砖引玉的作用，促进该领域的相关研究蓬勃发展。

6 结论

本研究借助用户提问数据构建了一个中文健康问句分类，并以寻医问药网上的高血压相关问句为例，采用人工标注的方式进行研究，以构建标注语料库，并归纳公众的健康信息需求及各类健康问句的一般形式。本研究的潜在用途如下：①所构建的中文健康问句分类有助于组织大型健康问题集，以提高检索效率；②所标注的高血压相关样本问句可作为后续研究的分析语料，如训练机器进行高血压相关中文健康问句的自动主题分类，进而实现健康热点 / 舆情监测、基于问句主题和关键词的自动答案生成等；③高血压相关健康问句的一般主题分布有助于健康网站的知识资源建设，例如优先建设有关公众询问较多问题的知识库。

此外，本研究所设计和采用的问句分类构建方式、语料标注流程、评判者间信度测量方法等，也可为开放领域及其他受限领域开展用户问句分类与语料构建提供一定的借鉴。

参考文献

[1] 中国互联网络信息中心 . 第 37 次中国互联网络发展状况统计报告 [EB/OL]. [2016-02-18]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/201601/P020160122469130059846.pdf.

[2] 国家心血管病中心 . 中国心血管病报告 2013[R]. 北京 : 中国大百科全书出版社 , 2014.

[3] 邓胜利 , 刘瑾 . 基于文本挖掘的问答社区健康信息行为研究——以“百度知道”为例 [J]. 信息资源管理学报 , 2016, 6(3):25-33.

[4] 毛先领 , 李晓明 . 问答系统研究综述 [J]. 计算机科学与探索 , 2012, 06(3):193-207.

[5] Isaac T, Zheng J, Jha A. Use of UpToDate and outcomes in US hospitals.[J]. Journal of Hospital Medicine An Official Publication of the Society of Hospital Medicine, 2012, 7(2):85-90.

[6] 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.

[7] Thomas P, Starlinger J, Vowinkel A, et al. GeneView: A Comprehensive Semantic Search Engine for PubMed[J]. Nucleic Acids Research, 2012, 40(W1): W585-W591.

[8] 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.

[9] 镇丽华 , 王小林 , 杨思春 . 自动问答系统中问句分类研究综述 [J]. 安徽工业大学学报 (自科版), 2015, 32(1):48-54.

[10] 陈玉 . 问答系统中问句分类算法研究 [J]. 软件工程师 , 2015(11):30-31.

[11] Rahman T A. Question Classification Using Statistical Approach: A Complete Review[J]. Journal of Theoretical and Applied Information Technology, 2015, 71(3): 386-395.

[12] 文勖 , 张宇 , 刘挺 , 马金山 . 基于句法结构分析的中文问题分类 [J]. 中文信息学报 ,2006,02:33-39.

[13] Mollá D, Vicedo J L. Question Answering in Restricted Domains: An Overview [J]. Computational Linguistics, 2007, 33(1): 41-61.

[14] Gebel R S, Okkes I M. International Classification of Primary Care (ICPC-2-NL)[M].

Utrecht: Nederlands Huisartsen Genootschap, 2000.

[15] Ely J W, Osheroff J A, Gorman P N, et al. A Taxonomy of Generic Clinical Questions: Classification Study[J]. BMJ, 2000, 321(7258): 429-32.

[16] Fiol G D, Workman T E, Gorman P N. Clinical Questions Raised by Clinicians at the Point of Care: A Systematic Review[J]. Jama Internal Medicine, 2014, 174(5):710-718.

[17] Schnall R, Cimino J J, Currie L M, et al. Information Needs of Case Managers Caring for Persons Living with HIV.[J]. Journal of the American Medical Informatics Association, 2011, 18(3):305-8.

[18] Roberts K, Demnerfushman D. Interactive Use of Online Health Resources: A Comparison of Consumer and Professional Questions [J]. Journal of the American Medical Informatics Association, 2016, 23(4):1-10.

[19] C é cile RLB, Frans JM. Classifying Health Questions Asked by the Public Using the ICPC-2 Classification and a Taxonomy of Generic Clinical Questions: An Empirical Exploration of the Feasibility[J]. Health Communication, 2010, 25(2):175-181.

[20] Guo H, Li J, Dai T. Consumer Health Information Needs and Question Classification: Analysis of Hypertension Related Questions Asked by Consumers on a Chinese Health Website [C]\\ MedInfo 2015: eHealth-enabled Health - Proceedings of the 15th World Congress on Health and Biomedical Informatics, So Paulo, Brazil, 19-23 August 2015. Studies in Health Technology and Informatics, IOS Press, 2015, 216: 810-814.

[21] Yan Z. Toward a Layered Model of Context for Health Information Searching: An Analysis of Consumer-Generated Questions[J]. Journal of the American Society for Information Science and Technology, 2013, 64(6): 1158-1172.

[22] Roberts K, Masterton K, Fiszman M, et al. Annotating Question Types for Consumer Health Questions[C]\\ Proceedings of the Fourth LREC Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing. 2014, [EB/OL].[2016-02-23], https://lhncbc.nlm.nih.gov/publication/pub7077.

[23] 高艳影 . 中文问答系统中的问题分类研究 [D]. 合肥 : 合肥工业大学 , 2011: 5.

[24] 李沛晏 , 朱露 , 吴多胜 . 问答系统综述 [J]. 数字技术与应用 , 2015, (4):69-71.

[25] 寻医问药网站简介 [EB/OL]. [2016-09-16]. http://www.xywy.com/about/index.html.

[26] Elliott A C, Woodward W A. Statistical Analysis Quick Reference Guidebook: with SPSS Examples[M]. SAGE Publications, Inc, 2006.

[27] Cao Y G, Cimino J J, Ely J, et al. Automatically Extracting Information Needs from Complex Clinical Questions[J]. Journal of Biomedical Informatics, 2010, 43(6): 962-971.

[28] Jadhav A S, Wu S, Sheth A P, et al. Online Information Seeking for Cardiovascular Diseases: A Case Study from Mayo Clinic.[C]. Medical Informatics Europe, 2014:702-706.

[29] Roberts K, Kilicoglu H, Fiszman M,et al. Decomposing Consumer Health Questions[C]\\ BioNLP Workshop. 2014: 29–37.

[30] 国家卫生和计划生育委员会宣传司 , 中国健康教育中心 . 2012 年中国居民健康素养监测报告 [R]. 2013:2-12.

[31] 聂雪琼 , 李英华 , 李莉 . 2012 年中国居民健康素养监测数据统计分析方法 [J]. 中国健康教育 , 2014, 30(2): 178-181.

[32] Reeder B, Le T, Thompson H J, et al. Comparing Information Needs of Health Care Providers and Older Adults: Findings from a Wellness Study[C]\\ MedInfo 2013: Proceedings of the 14th World Congress on Medical and Health Informatics, Copenhagen, Denmark, 20-24 August 2013. Studies in Health Technology and Informatics, IOS Press, 2013, 192: 18-22.