中文糖尿病问题分类体系及标注语料库构建研究

钱晓波[1] ， 谢文秀[2] ， 龙绍沛[1] ， 兰牧融[1] ， 慕媛媛[3] ， 郝天永[1,*]
[1]华南师范大学，计算机学院，广东广州
[2]香港城市大学，电脑科学系，香港
[3]巢湖学院，外国语学院，安徽合肥
xiaoboqian1221@outlook.com，vasiliky@outlook.com，Shaopei-Lauv@m.scnu.edu.cn，
1460685366@qq.com，myy@chu.edu.cn，haoty@m.scnu.edu.cn

摘要

糖尿病作为一种典型慢性疾病已成为全球重大公共卫生挑战之一。随着互联网的快速发展，庞大的二型糖尿病患者和高危人群对糖尿病专业信息获取的需求日益突出，糖尿病自动问答服务对患者和高危人群的日常健康服务也发挥着越来越重要的作用，然而存在缺乏细粒度分类等突出问题。本文设计了一个表示用户意图的新型糖尿病问题分类体系，包括6个大类和23个细类。基于该体系，本文从两个专业医疗问答网站爬取并构建了一个包含122732个问答对的中文糖尿病问答语料库DaCorp，同时对其中的8000个糖尿病问题进行人工标注，形成一个细粒度的糖尿病标注数据集。此外，为评估该标注数据集的质量，本文实现了8个主流基线分类模型。实验结果表明，最佳分类模型的准确率达到88.7%，验证了糖尿病标注数据集及所提分类体系的有效性。Dacorp、糖尿病标注数据集和标注指南已在线发布，可以免费用于学术研究。

关键词： 糖尿病 ； 问题分类 ； 分类体系 ； 语料库建设 ； 标注

The Construction of Question Taxonomy and An Annotated Chinese Corpus for Diabetes Question Classification

Xiaobo Qian[1],Wenxiu Xie[2],Shaopei Long[1],Murong Lan[1],Yuanyuan Mu[3],Tianyong Hao[1,*]
[1]School of Computer Science, South China Normal University,Guangzhou, Guangdong
[2]Department of Computer Science,City University of Hong Kong, Hong Kong
[3]School of Foreign Languages,Chaohu University,Hefei, Anhui
xiaoboqian1221@outlook.com，vasiliky@outlook.com，Shaopei-Lauv@m.scnu.edu.cn，
1460685366@qq.com，myy@chu.edu.cn，haoty@m.scnu.edu.cn

Abstract

As a typical chronic disease, diabetes has become one of the major global public health challenges. With the rapid development of the Internet, the huge group of type 2 diabetes patients and high-risk people has shown an increasing demand for specialized information on diabetes. The automated diabetes Question Answering (QA) services also play a vital role in providing daily health services for patients and high-risk people. However, issues like fine-grained classification are still unsolved in many QA services. In this paper, we design a new diabetes question classification taxonomy which represents the user intent, including 6 coarse-grained categories and 23 fine-grained categories. We also construct a new Chinese diabetes QA corpus DaCorp that contains 122,732 questions-answer pairs, collected from two professional medical QA websites. Meanwhile, we annotate 8,000 diabetes questions in DaCorp as a fine-grained diabetes dataset. To evaluate the quality of the proposed taxonomy and the annotated dataset, we implement 8 mainstream baseline classifiers for diabetes question classification. Results show that the best-performing model gained an accuracy of 88.7%, demonstrating

第二十一届中国计算语言学大会论文集，第395页-第405页，南昌，中国，2022年10月14日至16日。

the validity of the annotated diabetes dataset and the efficacy of the proposed taxonomy. The Dacorp, annotated diabetes dataset, and annotation guidelines are published online and free for academic research.

Keywords: Diabetes , Question Classification , Classification Taxonomy , Corpus Construction , Annotation

1 引言

随着经济的快速发展和生活方式的迅速变化，糖尿病的患病率呈急剧上升趋势，已成为当今重要的公共卫生挑战之一。根据国际糖尿病联合会（International Diabetes Federation，IDF）全球糖尿病地图集的最新报告[1]，2021 年，全球有近5.37亿成年人（20-79岁）患有糖尿病，其中670万人在同年死亡，这意味着每5秒钟就有1人死于糖尿病。中国目前糖尿病患者已达1.4亿，是世界上糖尿病患者人数最多的国家[1]。糖尿病已成为中国21世纪最具挑战的公共健康问题(Jia, 2014)。然而，近50%的糖尿病患者未得到确诊且不了解自己的病情[1]，即存在自诊率低的问题。另外，据世界卫生组织（World Health Organization，WHO）报告[2]，在糖尿病的患病人群中，超过95% 的患者患有二型糖尿病。与无法预防的一型糖尿病不同，二型糖尿病可以通过改变生活方式和提高健康管理能力来预防(Powers et al., 2015)。因此，高质量的糖尿病管理知识和信息对糖尿病患者和糖尿病高危人群至关重要。

据中国互联网络信息中心（China Internet Network Information Center，CNNIC）报告[3]，截至2021年12月，我国网民规模达10.32亿，互联网普及率达73.0%。互联网已成为患者寻找健康信息、表达健康信息需求的重要工具。许多在线健康问答社区和论坛已成为患者提问和分享信息的热门平台。然而，互联网上的健康信息质量参差不齐(Kanthawala et al., 2016)，因而向患者提供可靠的健康信息至关重要。由于现有的问答服务存在缺乏细粒度分类等突出问题，导致患者不能快速地找到与自己病情最相关的糖尿病信息。因此，从患者需求出发并对糖尿病问题进行细类度分类是自动问答服务一个亟需解决的问题，同时也是向用户提供可靠信息的一种有效方式。

本文为辅助患者快速获得最相关的糖尿病信息，设计了一个表示用户意图的新型糖尿病问题分类体系，包括饮食、治疗、预防、并发症等细粒度类别。同时，构建了一个糖尿病中文问答语料库DaCorp，包含122732个问答对。根据提出的分类体系，本文对DaCorp中的8000个糖尿病问题进行人工标注，形成一个糖尿病问题标注数据集，为糖尿病自动问答服务的发展提供数据支撑。Dacorp、标注数据集和标注指南已在网站发布可免费下载及用于学术研究。此外，本文实现了8个主流分类模型，并对各个模型在该标注数据集上的分类性能进行了对比与分析，验证了标注数据集的质量及提出分类体系的有效性。

本文的主要贡献如下：1）设计了一个表示用户意图的糖尿病问题分类体系，以辅助患者快速获得最相关的糖尿病信息；2）构建了一个糖尿病中文问答语料库DaCorp和糖尿病问题标注数据集；3）通过8个主流基线分类模型在糖尿病标注数据集上的分类实验，验证了分类体系的合理性和数据集的有效性，本文的实验结果可作为糖尿病问题分类任务的基准。

2 相关工作

2.1 医学问题分类体系

现有的医学问题分类体系研究主要分为两类，一类是基于问题性质和主题的分类体系，另一类是面向用户意图的分类体系。Ely et al. (2000)提出了一种针对临床问题的类型和性质的分类体系，即通用临床问题分类（Taxonomy for Generic Clinical Questions，TGCQ），TGCQ是一个四层结构的分类体系，包含64个通用问题类型。该分类体系

[1]International Diabetes Federation（2021）IDF diabetes atlas.https://diabetesatlas.org/

[2]World Health Organization.https://www.who.int/news-room/fact-sheets/detail/diabetes

[3]China Internet Network Information Center：China Internet Network Development State Statistic Report.http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/202202/P020220407403488048001.pdf

由来自152位医生提出有关患者护理的临床问题构建而成。另一种针对问题主题的分类体系是基层医疗国际分类[3]（International Classification of Primary Care-Version 2，ICPC-2），它由世界家庭医生组织（WONCA）开发，并广泛用于多个国家。该体系可以用于对患者的就诊原因、管理、干预措施等相关问题进行分类。然而，Boot and Meijman (2010)调查了这两种专业分类体系对患者提出的健康问题进行分类的可行性，发现这两种分类体系并不能直接用于患者健康问题的分类。他们的研究指出患者和医生的信息需求之间存在着差异。在现实生活中，存在很多患者经常询问但医生很少提出的问题，而患者常问的问题无法与这两个分类体系中的类别对应。例如，没有关于标准医学知识（例如，“X 综合征是否存在？”）和患者常问的饮食建议（例如，“X疾病吃什么食物有好处？”）相对应的问题类别。因此，直接使用以医生信息需求为基础的问题分类体系来表达患者意图并对患者所提问题进行分类是不完整且不合适的。

近年来，研究者开始构建和研究面向用户意图的问题分类体系。McRoy et al. (2016)提出了一个癌症问题分类体系，该体系将用户提出的癌症相关问题分为10个类别。Wang et al. (2020)使用卷积神经网络将用户提出的糖尿病健康问题分为9类，构建了一个糖尿病健康问题的分类体系，同时对中国糖尿病患者的健康信息需求进行了分析。Luo et al. (2020)将1000个高血压患者问题分为7个类别，构造了一个高血压问题的分类体系。尽管这些分类体系在一定程度上可以表达用户的健康信息需求，但它们都停留在对问题的表层分类即粗粒度类别分类，缺乏对用户提出的健康问题进行细类度划分，因而不能有效地满足患者快速检索与自身健康最相关的信息需求。基于现有糖尿病问题分类体系存在的问题，本文提出了一个新的问题分类体系，将问题依据用户意图进行细粒度划分，从而能够有效地辅助用户快速检索到与自身健康状况密切相关的问答信息。

2.2 医学问题语料库

近年来，不少国内外研究者对医学问题语料库进行了构建和研究。Ely et al. (2000)使用所提出的64个通用类型对临床问题进行人工标注并构建了一个临床问题语料库。该语料库是由152位家庭医生提出的有关患者护理的1396个临床问题组成。Roberts et al. (2014)构建了一个包含2937个遗传罕见疾病的患者问题语料库，该语料库的问题被标注为13个类别。相较于英文的医学问题语料库，中文医学问题语料库起步较晚。Guo et al. (2018)创建了一个中国健康问题的标注语料库。该语料库由5000个人工标注的中文健康问题组成。这些健康问题由2000个高血压相关问题和3000个来自内科、外科、妇产科、儿科、传染病、中医6个不同疾病领域的问题构成。

尽管现有中文问题语料库的疾病种类范围广，但很少有人构建和研究与糖尿病相关的中文问题语料库。据调研，只有Guo et al. (2020)提出了一个面向自动答疑服务的糖尿病问题语料库。该语料库由6401个带有<实体类型，意图类型>标签的糖尿病问题构成，采用人工标注来保证语料库的质量。然而，该语料库中的糖尿病问题主要与糖尿病的主要特征相关，例如，BMI指数（Body Mass Index）、葡萄糖、糖化血红蛋白、高血压和肌酐，并不能有效地体现用户意图和信息需求。据我们调研，在常用的专业医疗问答网站中，很多糖尿病患者并没有医学背景，且常问的问题更多地集中在与糖尿病相关的一般性问题和日常健康管理相关的问题，例如“二型糖尿病可以吃X食物吗”、“如何预防糖尿病”等。基于此，本文从两个大型专业医疗问答网站爬取和糖尿病直接相关的问答数据，构建了一个中文糖尿病问答语料库和糖尿病问题标注数据集。与其他研究不同的是，本文提出的语料库在分类体系上进行创新，对糖尿病问题进行需求细粒度分析，同时提供了主流分类模型在标注数据集上的问题分类性能，进一步对标注数据集的质量和分类体系的有效性进行评估。

3 糖尿病问题分类体系

分类体系是表示用户意图并系统地分析和记录用户对健康信息需求的一种方法。在问答系统中，问题分类在缩小检索范围和提高返回答案的准确性等方面发挥着重要作用(Zhen et al., 2015)。由于现有的面向用户意图的分类体系缺乏对问题的细类度划分，因此本文基于TGCQ分类体系设计了一个新的可以表示用户意图的糖尿病问题双层分类体系，包括6个大类和23个细类。尽管TGCQ分类体系是基于医生提出的有关患者护理的问题构建而成，并不能直接用于对

[3]International Classification of Primary Care-Version 2（ICPC-2）：https://www.ehelse.no/kodeverk/icpc-2e–english-version

用户健康问题的分类(Boot and Meijman, 2010)。但由于医生和患者提出的问题在“治疗”、“诊断”类别存在共性，本文对TGCQ分类体系进行调整和扩展后可用于用户糖尿病健康问题的分类。

针对TGCQ分类体系在用户糖尿病问题分类上存在的问题，本文对分类体系进行了调整和改进。首先，由于患者和医生的信息需求差异，存在很多患者经常询问但医生很少提出的问题，因此患者的常问问题无法与该分类体系中的类别相对应。基于此，为了尽可能涵盖患者提出的糖尿病问题，对于第一层的大类（coarse categories），本文增加了“常识”、“健康生活方式”和“其他”三个类别。保留了TGCQ分类体系中的“诊断”、“治疗”和“流行病学”三个大类，去除TGCQ分类体系中的“管理”和“非临床”两个类别。

其次，在TGCQ分类体系中的细类很多并不适宜用于糖尿病的问题分类，例如，“物理特性”、“社区服务”等。因此，对于第二层的细类（fine-grained categories），本文只保留了“临床解释”、“症状/表现”、“检查”和“病因学”4个类别。同时，增加了更符合糖尿病患者意图相关的19个类别，例如，“并发症”、“饮食”、“锻炼”、“预防”等。本文通过多轮人工标注对分类体系进行不断修改，最终确定的糖尿病问题双层分类体系为：第一层包括6个大类：“诊断”、“治疗”、“常识”、“健康生活方式”、“流行病学”和“其他”，第二层细类主要包括“检查”、“药物选择”、“生育力”、“饮食”和“预防”等23个类别。该分类体系对TGCQ分类体系进行了调整和扩展，使其能有效地体现患者意图和信息需求，从而可以辅助患者快速获得最相关的糖尿病信息。具体的分类体系结构如图1所示。

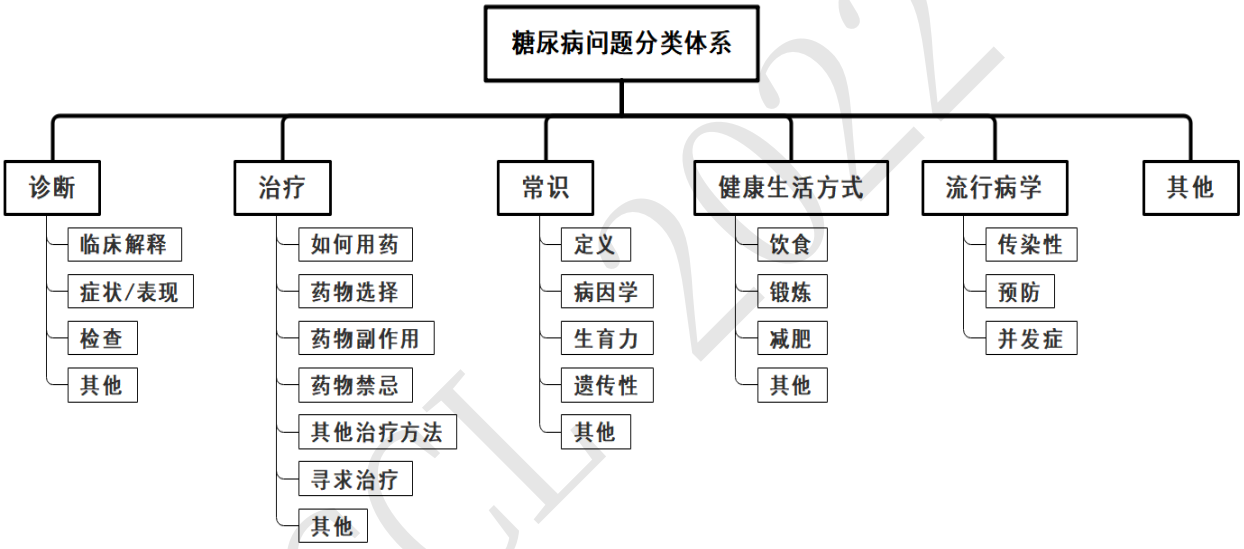

Figure 1: 糖尿病问题分类体系结构

4 语料库标注与构建

4.1 数据收集

糖尿病问答语料库DaCorp的数据收集于两个中国大型医疗问答网站:“39健康”[4]与“有问必答”[5]。在这两个网站上，用户可以提出与医疗护理相关的问题，并在提交问题时提供病情的详细描述。此外，用户提交的每个问题都会由国内医院的专业医生给出回复，同时用户咨询的问题具有高覆盖性和多样性等特点。语料库的构建主要包括医疗问答网站原始数据的收集和数据清洗两个步骤。首先，本文开发了一个基于Beautiful Soup 库[6]的python脚本，自动从网站上爬取以“糖尿病”为关键词查询到的问答数据。数据包括用户提出的问题和医生给出的相应回复，共收集196915条。其次，我们对收集的原始数据进行数据清洗和预处理，以便后续的人工标注。

[4]http://www.39.net/
[5]http://club.xywy.com/
[6]Python Library Beautiful soup. https://pypi.org/project/

第二十一届中国计算语言学大会论文集，第395页-第405页，南昌，中国，2022年10月14日至16日。

数据清洗主要是对重复和不相关的数据的进行过滤，去除如广告等非健康相关的内容。此外，由于网站上的问题是由患者提出的，他们大多没有医学背景，因而问题中存在大量的自然语言描述，非常口语化且存在很多错别字。例如，许多患者可能会将“二甲双胍”输入为“二甲双瓜”，将”妊娠糖尿病”输入为“妊辰糖尿病”等。因此，我们对问题中的错别字进行预处理，即人工纠正。在数据清洗和预处理后，最终的糖尿病问答语料库DaCorp包含122732 条问答对。

4.2 数据标注

在糖尿病问答语料库DaCorp构建完成后，本文从语料库中随机抽取了8000个问题进行人工标注，并形成一个糖尿病问题标注数据集。图2为数据的人工标注流程。数据标注分为标注准备和数据正式标注两个阶段。在标注准备阶段，通过对现有的分类体系和标注指南的研究和分析，本文设计了最初版本的糖尿病问题分类体系和对应的标注指南，其中标注指南包括每个类别的标注规则，并给出了相应的通用问题模式和示例问题，以提高分类体系的合理性、可用性和标注一致性。表1是类别“治疗”中每个细类的标注指南。分类体系的完整标注指南可从我们发布的网站中获取。同时，为了减少人工标注的繁重工作量并加快标注过程，本文基于Tkinter库[7]开发了一个简易的人工标注工具来辅助数据标注工作。在数据标注的正式标注阶段，由三位具有标注经验的硕士研究生参与标注工作。首先，数据集中的8000个糖尿病问题由一位具有医学信息学背景的标注者进行标注，另外两名标注者对数据集中随机抽取的2000个糖尿病问题进行标注，每人标注1000个。初始标注完成后，我们对三位标注者的标注结果进行一致性评估，对出现分歧的问题进行讨论，在协商一致后，修改分类体系和对应的标注指南。然后，三位标注者按照修改后的标注指南，独立标注剩余的6000个问题，每人标注2000个。最后，比较三位标注者的数据标注结果，对标注过程中出现的分歧进行讨论以达成一致，同时修改和完善标注指南的最终版本。

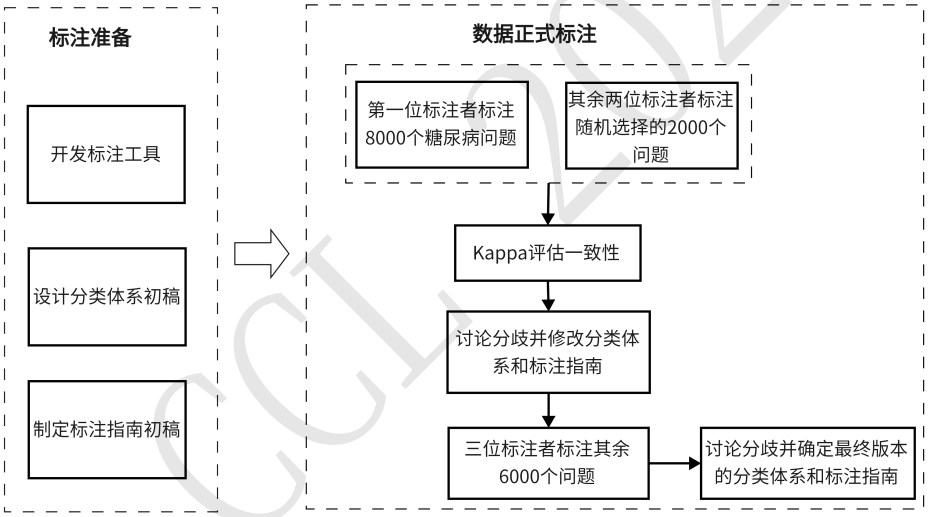

Figure 2: 数据标注流程

为了评估数据集标注的一致性和质量，发现标注者之间可能存在的分歧，本文使用Kappa统计量来检验标注者间一致性（Inter Annotator Agreement，IAA）。Kappa是IAA的衡量标准(McHugh, 2012)，它可以纠正偶然发生的一致性。Kappa的计算如公式1所示。

$$Kappa = \frac{(P_o - P_e)}{(1 - P_e)} \tag{1}$$

其中P_o是总体分类精度，即每一类正确分类的样本数量之和除以总样本数，P_e是偶然发生的一致性(Elliott and Woodward, 2007)。Kappa值越大，表明标注一致性越好。在本研究中，如果标注者标注的大类和细类都相同，则视为一致。在标注过程中，三位标注者的平均标注一致性为0.78，表明人工标注结果的差异较少，分类体系的制定合理且有效。

[7]Python interface to Tcl/Tk. https://docs.python.org/3/library/tkinter.html

第二十一届中国计算语言学大会论文集，第395页-第405页，南昌，中国，2022年10月14日至16日。

细类	标注规则	通用问题模式（部分）	示例问题（部分）
如何用药	患者知道使用什么药物，但是不知道药物使用的时间、剂量、注意事项。	X疾病服用Y药物是饭前还是饭后?	二型糖尿病吃二甲双胍是饭前吃还是饭后吃?
药物选择	患者已知自己的病情，需要了解合适的药物。	X疾病Y症状需要服用什么药物?	一型糖尿病人发烧能吃些什么药物?
药物副作用	患者不确定服用某种药物是否有副作用，以及副作用的详情。或者患者不确定某种症状是否是药物的副作用。	X疾病服用Y药物有副作用吗?	糖尿病吃盐酸二甲双胍缓释片有副作用吗?
药物禁忌	患者想了解服用某种药物的注意事项和禁忌。	X疾病服药期间可以打Y疫苗吗?	糖尿病患者服药期间可以打乙肝疫苗吗?
其他治疗方法	患者想了解能否通过非药物的方式治疗疾病，以及非药物治疗的类型、效果和风险。	X疾病能手术治疗吗?	二型糖尿病能手术治疗吗?
寻求治疗	患者描述自身的症状想寻求帮助或治疗。	X疾病患有Y症状如何治疗?	糖尿病人身上痒如何治?
其他	患者的问题与治疗相关，但不属于其他细类。	X疾病最好的治疗方法是什么?	一型糖尿病最好的治疗方法是什么?

Table 1: “治疗”类别中每个细类的标注指南

4.3 标注结果

糖尿病标注数据集的大类标注结果分布如图3所示。在8000个糖尿病问题中，从第一层大类的角度看，常识（C）和健康生活方式（D）类别在数据集中出现频率最高，分别为1655（20.7%）和2226（27.8%）个，其次是治疗（B）类别2026（25.3%）个。这些数据表明，在医患平台上查询糖尿病相关问题的用户中大多为糖尿病患者，且对糖尿病的日常健康管理的关注度较高。同时，还有717（9%）个问题分到诊断类别（A）中，此数据反映了可能存在相当数量的人群已有糖尿病症状却还未被诊断。其余，807（10.1%）个问题与流行病学（E）相关，其他类型（F）的问题有569个（7.1%）。特别地，标注者在标注过程中发现用户对健康生活方式（D）中饮食（D1）类别的关注度尤为突出，这反映了患者越来越重视饮食在糖尿病预防和病情控制过程中的作用。

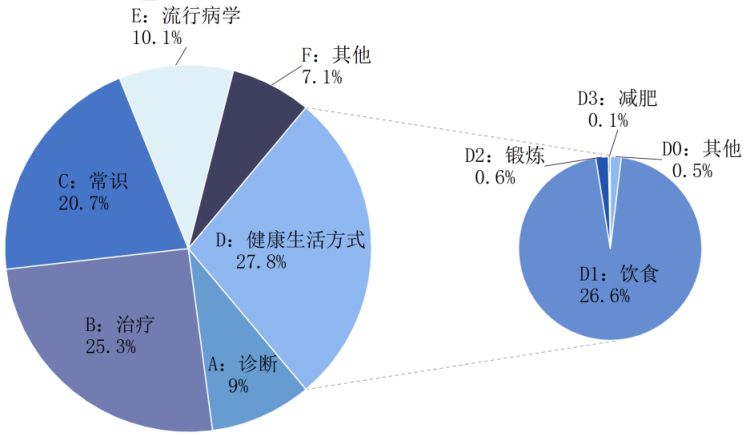

Figure 3: 语料库大类标注结果分布

糖尿病标注数据集的细类标注结果如表2所示。从第二层细类的角度看，与治疗（B）相关的问题，主要是寻求治疗（B6；788/2026）和药物选择（B2；514/2026）。用户寻求治疗的问题大都集中在患有糖尿病的情况下如何对正常疾病进行治疗，例如“糖尿病人喉咙痛怎么办？”、“糖尿病人感冒了能吃一般的感冒药吗”。尽管“喉咙痛”和“感冒”是很常见的症状和疾病，但是糖尿病患者仍然担忧常见疾病发生在糖尿病患者身上需要特别处理和治疗。同时，我们从用户询问药物选择相关的问题中发现，用户除了想了解糖尿病常见治疗药物之外，对常见疾病的治疗药物也很关注。此外，在有关诊断（A）的患者询问问题中，用户想通过临床解释和症状来诊断糖尿病的问题占主体的88.7%（636/717）。在糖尿病常识（C）类别中，患者询问的问题包括定义（C1）、病因学（C2）、糖尿病的生育力（C3）和糖尿病的遗传性（C4）,其中“其他”类别（C0）所占C类的比例为31.8%(527/1655)，通过仔细观察和分析语料库，发现患者提出的与糖尿病常识相关的问题，种类非常繁多且不集中，例如“2型糖尿病可以拔牙吗”、“糖尿病患者需要补充什么维生素”、“糖尿病患者能上大学吗”、“糖尿病患者在夏季需要注意什么”等问题，这表明患者对于与糖尿病相关的常识问题具有种类多、数量高的特征。

同时，患者也咨询了保持健康生活方式（D）的各种方法，包括饮食（D1）、锻炼（D2）、减肥（D3）等许多方面。一些患者意识到不良的生活方式可能会加重糖尿病的病情，而良好的生活方式可以缓解病情、改善健康。此外，患者还咨询了流行病学（E）的相关问题，他们主要想了解糖尿病的传染性（E1）、预防（E2）以及并发症（E3）相关的信息。最后，有569个如“1型糖尿病不治疗能活多久”、“甲亢合并糖尿病可怕吗”、“糖尿病患者可以带胰岛素上飞机吗”等不属于以上5大类别的问题被归于其他类别。

大类	细类	数量
A:Diagnosis(诊断)	A0: other(其他)	16
	A1:interpretation of clinical（临床解释）	344
	A2:symptom/manifestations（症状/表现）	292
	A3:test（检查）	65
B:Treatment（治疗）	B0:other（其他）	388
	B1:how to use drug（如何用药）	101
	B2:drug choice（药物选择）	514
	B3:adverse effects of drug（药物副作用）	69
	B4:contraindications of drug（药物禁忌）	4
	B5:Other Therapy（其他治疗方法）	162
	B6:Treatment Seeking（寻求治疗）	788
C:Common Knowledge（常识）	C0:other（其他）	527
	C1:Definition（定义)	152
	C2:Etiology（病因学）	860
	C3:Fertility（生育力）	67
	C4:Hereditary（遗传性）	49
D:Healthy lifestyle（健康生活方式）	D0:other（其他）	43
	D1:Diet（饮食）	2126
	D2:Exercise（锻炼）	50
	D3:weight-losing（减肥）	7
E:Epidemiololgy（流行病学）	E1:Infect（传染性）	31
	E2:Prevention（预防）	34
	E3:Complication（并发症）	742
F:Other（其他）	569	

Table 2: 语料库标注结果

语料库的标注结果表明，患者主要关心糖尿病诊断的方法和症状，如为什么患有糖尿病会出现某些症状，如何治疗，他们是否能够服用特定药物，使用是否有副作用或禁忌，糖尿病的遗传性和生育力，以及他们在日常生活中可以采取怎样的措施来改善或预防他们的病情。

第二十一届中国计算语言学大会论文集，第395页-第405页，南昌，中国，2022年10月14日至16日。

5 实验与结果

5.1 实验设置

为了评估分类体系的合理性和标注数据集的质量，本文比较了8个主流分类模型在糖尿病标注数据集上大类的分类性能，并将8000条糖尿病标注问题进行随机划分，其中6000条数据作为训练集，1000条作为验证集，其余1000条作为测试集。分类模型采用的是6个深度神经网络模型Text CNN(Chen, 2015)、Text RNN(Liu et al., 2016)、Text RCNN(Lai et al., 2015)、Text RNN Attention(Zhou et al., 2016)、fastText(Joulin et al., 2016)、DPCNN (Johnson and Zhang, 2017)，以及两个预训练的大规模语言模型BERT(Devlin et al., 2018)和ERNIE(Sun et al., 2019)进行实验。对于深度神经网络模型，本文使用常用的Jieba[8]分词工具对数据进行中文分词。除fastText可自行训练词向量外，其它模型使用预训练的搜狗新闻词向量(Li et al., 2018)作为特征。其中，Text RNN和Text RNN Attention隐藏层数为128，Text RCNN和fastText隐藏层数为256。同时我们使用Adam(Kingma and Ba, 2014)优化器以0.001的学习率最小化交叉熵，并使用提前停止机制避免过拟合的问题。对于预训练的语言模型，本文主要对BERT和ERNIE进行微调，将学习率设置为5e-5并采用提前停止，epoch数目设置为12。

5.2 实验结果

本文采用的评测指标是分类准确率（Accuracy），即对于给定的测试数据集，分类模型正确分类的样本数与模型总样本数之比。不同分类模型在标注数据集上的分类性能结果如表3所示。从实验结果可以看出，在深度神经网络模型中，Text CNN的表现最佳，准确率为86.1%；DPCNN性能最低，准确率仅为82.8%。尽管DPCNN拥有最深的网络结构和最多的参数，能够捕获文本的长距离特征，但用户提出的糖尿病问题通常比较简短，因而TextCNN在短文本分类中性能要优于DPCNN。对于预训练的语言模型，ERNIE模型的准确率（88.7%）要高于BERT 模型（87.8%）。与BERT模型主要学习字级别的信息相比，ERNIE能利用文本的词法结构和语法结构，直接对先验语义知识单元进行建模，增强了模型完整概念的语义表示能力，有助于对文本进行理解和分类。同时，预训练的大规模语言模型BERT和ERNIE要优于其他深度神经网络模型，但性能差别不大。这与分类任务的训练样本数量有关，预训练语言模型BERT和ERNIE在大数据集上的优势应该会更明显。因此，在后续的工作中，本文将会继续扩大标注语料，完善标注数据集。

模型	Text CNN	Text RNN	Text RNN Attention	Text RCNN	fastText	DPCNN	BERT	ERNIE
准确率(%)	86.1	83.8	83.6	84.2	84.7	82.8	87.8	88.7

Table 3: 糖尿病标注数据集在8个分类模型上的实验结果

此外，本文对最佳分类模型ERNIE在糖尿病标注数据集上每个大类的分类性能进行对比和分析，实验结果如表4所示。采用的评价指标是查准率（Precision）、召回率（Recall）和F1值（F1-score），其中F1值是模型查准率和召回率的一种调和平均。从模型的分类结果中可以看出，“健康生活方式”类别的分类性能最好（F1：94.44%）。因患者关于“健康生活方式”的糖尿病问题主要集中在“饮食”方面，而“饮食”相关问题的文本特征明显（例如，“糖尿病可以吃X食物吗？”），因此利于模型分类。其次，模型在“治疗”、“诊断”、“常识”和“流行病学”类别上也具有较高的分类性能，F1均高于85%。而模型在“其他”（F1：64.%）这个类别的分类性能较低。模型的分类性能与数据集的数量相关，类别数据量较大时，模型能够得到充分训练，故分类效果较好。由于训练集中“其他”这个类别的数量较少，因此模型在这个类别的训练上可能存在欠拟合的问题，因而分类性能较低。

主流文本分类模型在标注数据集上的分类结果显示，分类模型的准确率均高于82%，最佳模型ERNIE在6个大类的平均F1为86.06%，证明本文提出的糖尿病问题分类体系合理且有效，同时人工标注的糖尿病问题数据集具有较高的质量，可为糖尿病问题分类任务和相关的问答服务系统研究提供有效的数据支撑。

[8]https://github.com/fxsjy/jieba

	Precision(%)	Recall(%)	F1-score (%)
诊断	94.05	90.80	92.40
治疗	84.98	93.96	89.25
常识	87.50	83.87	85.65
健康生活方式	95.51	93.41	94.44
流行病学	90.11	91.11	90.61
其他	70.18	58.82	64.0

Table 4: ERNIE在糖尿病标注数据集上每个大类的分类性能

5.3 语料库访问

为了方便用户和研究人员访问和使用语料库，本文开发了一个语料库网站，可以通过URL[9]访问数据。该网站主要分为4个主要模块，分别为：分类体系、标注指南、标注数据、数据下载。在“分类体系”页面，用户可以浏览糖尿病问题的分类体系。在“标注指南”页面，如图4所示，用户可以查看每个类别的标注规则，以及对应的问题模式和示例问题。在“标注数据”页面可以查看标注数据集中的糖尿病问题和答案，以及每个问题对应的大类和细类。最后，用户可以在“数据下载”页面选择XML格式下载标注的糖尿病问题和DaCorp中所有的糖尿病问答对。

Figure 4: 语料库网站的“标注规则”页面

图5显示了XML格式的问题示例，以便用户根据需求选择特定格式的语料库。从标注示例来看，第一行是XML声明，它定义了XML的版本。下一行描述了根元素<QAPairs>，它在XML文件中是唯一的。子元素<QAPair>包含了所有的糖尿病问题和答案。其中，子元素<question>、<answer>、<url>分别指的是患者提出的糖尿病问题，医生给出的答复和问答对的来源。<main_category>和<sub_category>表示示例问题被标注的大类和细类。XML格式文件将语料库可视化，以加强对问答语料库的直观理解。

[9]http://47.102.207.52:9090

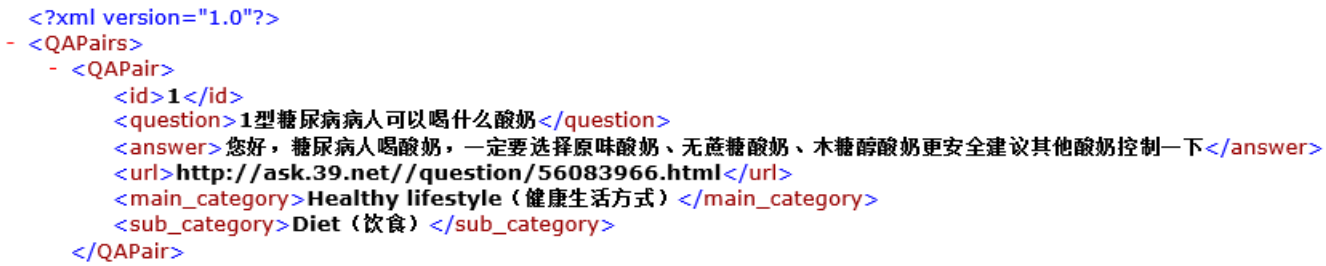

Figure 5: XML格式的标注问题示例

6 总结与未来工作

本文为辅助患者快速获得最相关的糖尿病信息，设计了一个表示用户意图的新型糖尿病问题分类体系。同时，构建了一个糖尿病问答语料库DaCorp，并使用新的分类体系对问题进行人工标注形成糖尿病标注数据集。最后，本文评估了8个主流分类模型在标注数据集上的分类性能，实验结果验证了数据集的有效性以及提出分类体系的合理性。据调研，本文提出的标注语料库是目前最大的糖尿病问题标注语料库。该标注语料库可通过网站公开访问，用于训练机器理解糖尿病患者的中文健康问题。本研究将为糖尿病问题分类、问答匹配等NLP相关的任务以及自动问答系统的开发提供数据支撑。此外，我们将不断完善糖尿病问题分类体系，并尝试使用主动学习来进行半自动化标注，扩大标注语料，提高标注效率，进一步挖掘和分析糖尿病问题的特征和糖尿病患者的信息需求。同时，利用现有语料库和分类体系开发糖尿病问题的高性能分类模型也是未来的重要工作。

参考文献

Cécile RL Boot and Frans J Meijman. 2010. 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. *Health communication*, 25(2):175–181.

Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alan C Elliott and Wayne A Woodward. 2007. *Statistical analysis quick reference guidebook: With SPSS examples*. Sage.

John W Ely, Jerome A Osheroff, Paul N Gorman, Mark H Ebell, M Lee Chambliss, Eric A Pifer, and P Zoe Stavri. 2000. A taxonomy of generic clinical questions: classification study. *Bmj*, 321(7258):429–432.

Haihong Guo, Xu Na, and Jiao Li. 2018. Qcorp: an annotated classification corpus of chinese health questions. *BMC medical informatics and decision making*, 18(1):39–47.

Xusheng Guo, Likeng Liang, Yuanxia Liu, Heng Weng, and Tianyong Hao. 2020. The construction of a diabetes-oriented frequently asked question corpus for automated question-answering services. In *Proceedings of the 2020 Conference on Artificial Intelligence and Healthcare*, pages 60–66.

Weiping Jia. 2014. Diabetes: a challenge for china in the 21st century. *The Lancet Diabetes & Endocrinology*, 2(4):e6–e7.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Shaheen Kanthawala, Amber Vermeesch, Barbara Given, Jina Huh, et al. 2016. Answers to health questions: internet search results versus online health community responses. *Journal of medical Internet research*, 18(4):e5369.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. *arXiv preprint arXiv:1805.06504*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

Aijing Luo, Zirui Xin, Yifeng Yuan, Tingxiao Wen, Wenzhao Xie, Zhuqing Zhong, Xiaoqing Peng, Wei Ouyang, Chao Hu, Fei Liu, et al. 2020. Multidimensional feature classification of the health information needs of patients with hypertension in an online health community through analysis of 1000 patient question records: observational study. *Journal of Medical Internet Research*, 22(5):e17349.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Susan McRoy, Sean Jones, and Adam Kurmally. 2016. Toward automated classification of consumers' cancer-related questions with a new taxonomy of expected answer types. *Health informatics journal*, 22(3):523–535.

Margaret A Powers, Joan Bardsley, Marjorie Cypress, Paulina Duker, Martha M Funnell, Amy Hess Fischl, Melinda D Maryniuk, Linda Siminerio, and Eva Vivian. 2015. Diabetes self-management education and support in type 2 diabetes: a joint position statement of the american diabetes association, the american association of diabetes educators, and the academy of nutrition and dietetics. *Diabetes care*, 38(7):1372–1382.

Kirk Roberts, Kate Masterton, Marcelo Fiszman, Halil Kilicoglu, and Dina Demner-Fushman. 2014. Annotating question types for consumer health questions. In *Proceedings of the Fourth LREC Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Tian-Hao Wang, Xiao-Feng Zhou, Yuan Ni, and Zhi-Gang Pan. 2020. Health information needs regarding diabetes mellitus in china: an internet-based analysis. *BMC Public Health*, 20(1):1–9.

Lihua Zhen, Xiaolin Wang, Sichun Yang, et al. 2015. Overview on question classification in question-answering system. *Journal of Anhui University of Technology (Natural Science)*, 32(1):48–54.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.