基于Web的问答系统综述

李舟军 李水华
(北京航空航天大学计算机学院 北京 100191)

摘 要 微软小冰引发了问答系统的新一轮研究热潮。作为一种新型的信息检索方式,问答系统能直接以自然语言与用户进行人性化的交互。而基于Web的问答系统能通过搜索引擎获取开放的互联网上的各种相关信息,并将以自然语言形式表述的准确答案返回给用户,因此此类系统同时具有搜索引擎和问答系统的优点。首先,对基于Web的问答系统的研究背景与发展历史进行了概述;然后,详细介绍了基于Web的问答系统的架构及其问题分析、信息检索、答案抽取这三大关键技术的研究进展;在此基础上,分析了基于Web的问答系统所面临的问题;最后,对基于Web的问答系统的未来发展趋势进行了展望。

关键词 问答系统,基于Web的问答系统,问题分析,信息检索,答案抽取
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.06.001

Survey on Web-based Question Answering

LI Zhou-jun LI Shui-hua
(School of Computer Science and Engineering, Beihang University, Beijing 100191, China)

Abstract Microsoft Xiaoice triggers a new round of boom on question answering research. As a new kind of information retrieval technology, question answering offers friendly interaction for users by using natural languages. Web-based question answering extracts answers in natural languages for users' questions from search results provided by search engines. Web-based question answering has both advantages of search engine and question answering. Firstly, background and history of web-based question answering were summarized. Then, the research progress of the three key technologies of web-based question answering (question analysis, information retrieval and answer extraction) were introduced in detail. Based on the above introduction, the problems to be solved of web-based question answering were analyzed. Finally, the future research trend of web-based question answering was discussed.

Keywords Question answering, Web-based question answering, Question analysis, Information retrieval, Answer extraction

1 引言

随着互联网的快速发展和个人计算机的广泛普及,越来越多的消息和数据通过超文本传输协议以电子文档的形式发布。于是,用户可以通过互联网这种更加便捷的途径获取他们所需要的各类信息。与此同时,海量的数据资源汇聚在一起形成了信息大爆炸。如何准确并及时地在浩如烟海的信息世界中获取用户所需的信息,已经成为互联网发展至今的一大难题。信息检索技术就是解决这一难题的良方。

搜索引擎技术作为一种成熟的信息检索技术,可以满足用户绝大部分的信息需求。但是,随着互联网数据的疯狂增长,搜索引擎的缺点逐渐显现。百度、谷歌、必应这类传统的搜索引擎,通常只能以关键词作为输入。而用户在检索信息时,需要将其查询需求凝练为若干简洁的关键词后提交给搜索引擎。对于普通用户而言,他们往往难以用少量的关键词来准确地表述其查询意图。此外,搜索引擎的返回结果不是一个简洁的准确答案,而是一个网页片段的列表。这些网页片段通常含有大量的噪声数据,用户仍然需要阅读这些网页片段乃至相应的原始网页,才能找到其所需的答案。在计算资源紧缺时,损失一些用户体验来提升信息检索的效率是合适的。其实,计算机能够很方便地分析与处理关键词和网页片段列表,搜索引擎本身以可实现这些相关的功能,以降低用户的使用难度。随着摩尔定律[2]的持续影响,计算机的性能获得了巨大的提升;特别是"互联网+"[3]的提出,预示着互联网将成为人们工作和生活的基础设施。此时,用户的查询体验就变得至关重要。

为了改善信息检索的用户体验,人们开始研究直接以自然语言作为输入与输出的问答系统(Question Answering,

到稿日期:2017-03-13 返修日期:2017-04-20 本文受国家自然科学基金项目(61672081,U1636211),国家863计划项目(2015AA016004),北京成像技术高精尖创新中心项目(BAICIT-2016001)资助。
李舟军(1963-),博士,教授,博士生导师,CCF高级会员,主要研究方向为数据挖掘与人工智能、网络与信息安全,E-mail:lizj@buaa.edu.cn;
李水华(1991-),男,硕士生,主要研究方向为自然语言处理。

QA)。对于问答系统,用户能够以文本或者语音的方式,使用自然语言直接地表达其查询需求。问答系统理解用户的查询意图后,通过一系列的检索、分析与处理,直接将以自然语言形式表述的准确答案返回给用户。因此,对用户来说,问答系统是一种更加方便、友好和精准的服务。

1999 年,信息检索评测组织 TREC(Text REtrieval Conference)设立了问答系统的评测任务,它极大地推动了问答系统的发展。

近年来,国内外各大公司在问答系统领域的激烈竞争也从侧面印证了问答系统蕴涵的庞大商机:苹果推出了 Siri,微软推出了小冰和小娜,百度推出了小度。

由于人们观察问答系统的角度各不相同,因此问答系统有多种分类方法。本文根据问答系统知识的不同来源,将其分为 3 类。

1)基于知识库的问答系统(Question Answering over Knowledge Bases,KBQA)[4-5]:主要以知识库作为问答系统的知识来源。

2)基于社区的问答系统(Community-based Question Answering,CQA)[6-7]:主要以问答社区(如知乎、百度知道、雅虎知道、搜搜问问等)作为问答系统的知识来源。

3)基于 Web 的问答系统(Web-based Question Answering,WQA)[8-9]:以开放的互联网上的 Web 文档作为问答系统的知识来源,从搜索引擎返回的相关网页片段从中抽取出用户所提问题的答案。

WQA 系统同时具有搜索引擎和问答系统的优点。

1)能通过现有成熟的搜索引擎来获取整个互联网上的各种相关信息,这些信息无所不包,不受领域的限制,且与时俱进,不断更新;

2)一般能够较好地处理和回答事实型问题;

3)能够利用自然语言进行人性化的交互。

2 发展历史

1950 年,英国著名数学家图灵提出了图灵测试[10],用于判断机器是否能够思考。在图灵测试中,让计算机用自然语言与用户对话的这种想法,其实就是问答系统的最初构想。

从 20 世纪 60 年代开始,相继出现了第一批问答系统,例如:Baseball[11]能够回答一些关于美国篮球联赛的相关问题;Lunar[12]能够对阿波罗月球探测任务取回的岩石样本,回答一些关于其分析结果的相关问题;ELIZA[13]能够通过与精神病人的对话,缓解其精神疾病。

1993 年,麻省理工学院人工智能实验室开发出了世界上第一个 WQA 系统——START[14]。START 能够回答数百万个有关地理、历史、文化、科技、娱乐等方面的英语问题。值得一提的是,START 是一个混合型问答系统,它综合利用了 KBQA 和 WQA 的相关技术。START 优先尝试通过知识库回答用户的问题;只有当知识库不包含用户所需的答案时,才会利用搜索引擎查找相关信息,并从中抽取答案返回给用户。

2002 年,密歇根大学开发了一个 WQA 系统 AnswerBus[15]。该系统的一个重要特性是支持多语言,用户可以自由地使用英语、法语、德语、西班牙语和意大利语等语言进行提问。同一时期,有影响力的问答系统还有 Webclopedia[16],Encarta[17]和 LAMP[18]等。

2009 年,Wolfram Research 公司推出了在线自动问答系统 Wolfram Alpha。该系统不仅能直接给出问题的答案,还能给出与答案相关的所有信息。其数据来源包括学术网站、出版物、商业网站以及公司与科学机构的数据等。Wolfram Alpha 是基于 Wolfram 的另一个旗舰产品 Mathematica 开发的,其底层运算和数据处理工作通过在后台运行的 Mathematica 实现。因为 Mathematica 囊括了计算机代数、符号和数值计算、可视化和统计功能的计算平台和工具包,所以 Wolfram Alpha 能够回答多种多样的数学问题,并将答案以清晰美观的图形化方式显示给用户。Wolfram Alpha 可以对用户上传的图片进行识别[19],并可完成数学、统计学、物理、化学、生命科学、计算机科学、经济学、社会学、语言学、文学、历史、文化、体育、音乐、天气等各个领域的查询、计算与分析。Wolfram Alpha 还为商务合作伙伴提供了一个应用程序接口[20]。

2011 年,IBM 的问答系统沃森(Watson)参加以问答为主的综艺节目《危险边缘》(Jeopardy!),并最终战胜了人类选手。沃森由 IBM 公司和美国德克萨斯大学联合打造,存储了海量的数据,而且拥有一套逻辑推理程序,可以推理出它认为最正确的答案。IBM 开发沃森旨在完成一项艰巨挑战:建造一个能与人类回答问题的能力相匹敌的计算系统。这就要求其具有足够的速度、精确度和置信度,并且能使用自然语言回答问题。沃森是一个混合型的问答系统,在实现上也借助了 WQA 的一些相关技术[21]。

Siri 最初以文字聊天服务为主,随后通过与全球最大的语音识别厂商 Nuance 合作,实现了语音识别功能。2010 年,苹果以 2 亿美金收购了 Siri[22],并在苹果的产品中内置了 Siri。2016 年 6 月 13 日,苹果开发者大会 WWDC 发布了 Siri 的新功能。

微软在 2014 年推出了聊天机器人小娜和小冰[23],并将小娜内置到了 Windows 系统中。小冰随后升级,新增了图像识别、语音对话等功能。小冰集合了中国网民多年来积累的公开数据,并凭借微软在大数据、自然语言理解、机器学习等方面的技术积累,精炼了几千万条真实而有趣的语料库,通过理解对话的语境与语义,实现了超越简单人机问答的自然交互。继登录微信、微博、京东等平台后,小冰作为全球英语培训机构 EF 英孚教育的品牌形象代言人,参与其全部广告市场活动。2015 年 12 月,小冰以见习主播身份登录东方卫视,负责每日天气播报板块。此外,微软还推出了日本版的小冰——Rinna(凛菜)和美国版的小冰——Zo。全新的第四代微软小冰不仅包括实时情感对话引擎、多种新感官、中日英 3 种语言,还有对应不同领域的功能插件平台。微软小冰引发了新一轮聊天机器人热潮,已经成为全球科技史上最大规模的一次图灵测试。

百度在 2014 年推出了小度机器人。依托百度强大的搜索能力,小度集成了自然语言理解、智能交互、语音与视觉等多种人工智能技术,能以自然的方式与用户进行信息、服务、情感的交流,并通过学习与进化不断提升各种技能。2017 年

1 月,在江苏卫视的《最强大脑》第四季的人机大战中[24],第一期由王峰对战小度,小度以 3:2 战胜人类"最强大脑"王峰;在第二期比赛中,小度与名人堂选手"听音神童"孙亦廷打成平手;在第三期比赛中,小度与"水哥"王昱珩进行人脸识别比赛,最终以 2:0 胜出。

这些是当下最热门的问答系统。虽然它们都没有公布实现细节,但是在使用的过程中不难发现它们都借助了一些网页数据,利用了 WQA 的相关技术。

3　基于 Web 的问答系统

经典的 WQA 系统[25]通常由以下 3 个模块构成。

1)问题分析模块:用户提出以自然语言表述的问题之后,问题分析模块负责分析用户的问题,理解用户的查询意图,并根据用户的查询意图生成相应的查询语句。根据具体情况,问题分析模块可能还需要对问题进行分类,提取问题的关键词或者生成一些其他描述用户查询意图的中间数据。这些中间数据将对 WQA 系统的其他模块提供非常重要的帮助。

2)信息检索模块:将问题分析模块得到的查询语句或关键词提交给搜索引擎,并对搜索引擎返回的搜索结果进行整理,从而得到一些可能包含正确答案的网页片段。由于信息检索模块需要调用搜索引擎来查找相关的网页,可能耗费大量的时间,因此该模块几乎是所有 WQA 系统的性能瓶颈。

3)答案抽取模块:综合利用信息抽取技术,从信息检索模块返回的相关网页片段中抽取出用户所提问题的最佳答案(疑似正确答案)。该模块可能还需要用到问题分析模块得到的问题类别、问题的关键词等刻画用户查询意图的数据。

基于 Web 的问答系统的架构如图 1 所示。

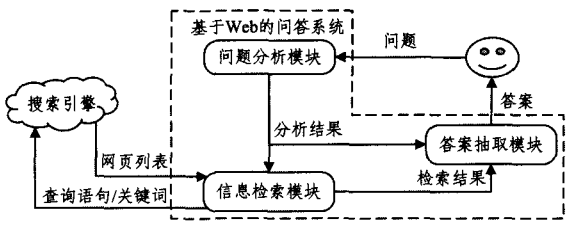

图 1　基于 Web 的问答系统的架构

目前,大部分 WQA 系统能够较好地处理和回答事实型问题[26],也有一些 WQA 系统可以解决列举型问题[27-29]。除非借用一些 KBQA 的相关技术[30]或者配合 CQA 系统一起使用[31],否则 WQA 系统不能很好地回答一些复杂的问题[32-34],例如定义型(Definition)、原因型(Why)、关系型(Relation)、比较型(Comparison)[35]、方法型(How-to)等问题。

下面将分别从问题分析模块、信息检索模块、答案抽取模块这 3 个方面介绍 WQA 的研究进展。

3.1　问题分析模块

首先,问题分析模块将对用户所提问题进行分词操作。众所周知,中文分词相对于英文分词要难得多,这给中文问题的分析增加了难度。然后,问题分析模块通常会对问题的类别进行分析[36],问题的类别是反映用户提问意图的重要信息。问题的关键词、同义词、相似问题之类的信息也常常被问题分析模块用于理解用户的查询意图[37]。此外,为了能够与搜索引擎更好地对接,问题分析模块有时候还会重写问题[38]。

3.1.1　问题分类

问题的类别其实也是答案的类别。问答的类别对其后的答案抽取具有重要的指导意义。国际上常用的问题分类体系如表 1[39]所列。针对中文的特点,哈尔滨工业大学信息检索研究室提出了中文问题的分类体系,如表 2[40]所列。

表 1　国际上常用的问题分类体系

大类(Coase)	小类(Fine)
ABBR	abbreviation, expansion
DESC	definition, description, manner, reason
ENTY	animal, body, color, creation, currency, disease, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
HUM	description, group, individual
LOC	city, country, mountain, other, state
NUM	code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight

表 2　中文问题的分类体系

大类(Coase)	小类(Fine)
人物(HUM)	特定人物,团体机构,人物描述,人物列举,人物其他
地点(LOC)	星球,城市,大陆,国家,省,河流,湖泊,山脉,大洋,岛屿,地点列举,地址,地点其他
数字(NUM)	号码,数量,价格,百分比,距离,重量,温度,年龄,面积,频率,速度,范围,顺序,数字列举,数字其他
时间(TIME)	年,月,日,时间,时间范围,时间列举,时间其他
实体(OBJ)	动物,植物,食物,颜色,货币,语言文字,物质,机械,交通工具,宗教,娱乐,实体列举,实体其它
描述(DES)	简写,意义,方法,原因,定义,描述其它
未知(Unknown)	未知

由于每一类问题都有一些非常明显的特征,因此很多 WQA 系统都采用了规则分类器[25,41]对问题进行分类。最简单的规则可以是:包含单词"谁(Who)"的问题属于人物类问题,包含单词"什么时候(When)"的问题属于时间类问题,包含单词"什么地方(Where)"的问题属于地点类问题。此外,存在一些比较复杂的规则,例如 LI 等人[42]提出的基于语义模式(SemanticPattern,SP)的问题分类方法,其语义模式是一种特殊的规则。对于问题"What book did Rachel Carson write in 1962?",可用语义模式"〈Target:Entity\Product〉〈Type:What〉[Physical_Object\Product] did [Physical_Object\Human] [Event\Action]?"进行匹配。除了问题的类型,这种方法还可以得到问题的目标、问题的约束等多种信息。

虽然规则分类器可以分类大多数问题,但它也有失效的时候,即没有一条规则能够与问题匹配成功。此时,WQA 系统会借助一些统计机器学习的方法[43]。支持向量机(Support Vector Machine,SVM)作为一种成熟的分类算法,被广泛用于 WQA 系统中的问题分类[41,44-49]。也有人使用 SNoW[50-51](Sparse Network of Winnow)分类器、语言模型[52-53](Language Modeling,LM)或者循环神经网络[54](Recurrent Neural Network,RNN)对问题进行分类,都取得了不错的效果。

3.1.2　关键词提取和扩展

关键词是问题的核心成分,既可作为搜索引擎的输入,也可辅助答案的抽取过程。WQA 系统通常在分词、去除停用

词之后进行关键词的提取，并将名词、动词、形容词等句子中的重要成分作为关键词。因此，可以用一些简单的规则提取关键词[55]："所有带形容词的名词都是关键词"，"所有在引号里的非停用词都是关键词"等。LIU 等人[41]利用自然语言处理工具分析问题的语法结构，然后抽取主语、宾语作为关键词。

关键词的扩展主要是为了解决关键词的同义词的匹配问题。很多时候，网页片段中并不包含问题中的关键词，但却包含关键词的同义词，例如：问题"第一次世界大战哪年爆发?"中存在关键词"爆发"，但是网页片段"第一次世界大战发生于1914 年。"中不包含"爆发"，却包含"爆发"的同义词"发生"。关键词的扩展通常需要一些同义词词库进行辅助[38]。

3.1.3 问题重写

除了问题的关键词，问题本身也可以作为搜索引擎的输入。但由于搜索引擎并不能很好地领会问题的语义，此时可能需要对问题进行重写[25,38]，以便 WQA 系统能从搜索引擎中获得更好的网页片段。

一些启发式的方法[56-57]依靠简单的字符串操作就可以实现问题重写，比如根据单词的词性可以将问题"When was the paper clip invented?"重写为"The paper clip was invented"。

有些 WQA 系统利用一些较为复杂的方法来优化问题的重写过程，比如 CHALI 等人[45]首先定义了一个针对问题的操作集合，然后利用概率模型选择最合适的操作来重写问题。

3.2 信息检索模块

信息检索模块是 WQA 系统和搜索引擎之间的桥梁，同时也是 WQA 系统和其他问答系统之间的最大区别。信息检索模块的主要功能是获取并解析搜索引擎返回的搜索结果，从而得到结构化的网页片段列表。信息检索模块可以调用搜索引擎提供的免费接口（通常有次数限制），也可以直接利用爬虫技术获取搜索结果[58]。

WQA 系统通常不会进一步获取网页片段所对应的原始网页[59]。虽然原始网页包含更丰富的文本数据，但是其中的噪音数据也更多。搜索引擎所具有的高质量摘要技术能过滤无关的文本数据，所以网页片段包含了原始网页中关键词密集出现的文本片段。这使得信息检索模块无需直接抓取并解析原始的网页，从而大大提高了 WQA 系统的效率与精度。另外，网页片段中常常含有不完整的句子，会影响一些依赖于语法结构的答案抽取算法[60-61]的效果。不完整的句子容易导致语法结构分析工具出错，也很容易丢失答案抽取模块所依赖的句子成分。

搜索引擎返回的网页片段是有顺序的，这种顺序反映了网页片段和查询语句的关联度。有的 WQA 系统[62]利用了这种搜索引擎提供的关联度，同时借助了一些简单的文本相似度算法对网页片段进行重排序。

3.3 答案抽取模块

答案抽取模块是 WQA 系统中的重点和难点。问题分析模块的分析结果和信息检索模块的检索结果都是答案抽取模块对答案进行抽取的重要依据。问题分析模块的分析结果包含问题类别、问题关键词等重要信息，这些信息能够很好地描述用户的提问意图。网页片段的列表作为信息检索模块的检索结果，是答案抽取模块抽取最佳答案的主要信息来源。答案抽取模块通过对上述信息的综合利用，使用信息抽取技术得到用户所需要的最佳答案。答案抽取模块通常包括两个主要步骤：候选答案抽取和候选答案排序。前者负责从网页片段中抽取出候选答案，后者负责对这些候选答案进行排序，从而得到最佳答案。

3.3.1 候选答案抽取

候选答案抽取是答案抽取模块的第一个步骤，其主要目的是缩小最佳答案的范围。抽取候选答案时，所得到的候选答案的数量不能太多，也不能太少。如果候选答案太多，则会给随后的候选答案排序带来很大的困难；如果候选答案太少，则很可能会遗漏所需的最佳答案。当候选答案集不包含最佳答案时，无论怎么优化候选答案排序算法都不可能召回最佳答案。

从网页片段中抽取候选答案的方法有很多，主要包括如下几种典型的方法。

1）手工编辑或自动生成名词词典，将词典中的所有名词都作为候选答案。这种极端的做法可能使得候选答案抽取失去了对最佳答案的筛选作用。著名的问答系统沃森[63]就将抽取维基百科的词条作为候选答案。这种方法产生的候选答案集非常大，从而导致候选答案排序的难度大大增加。由于候选答案集的维护、更新成本很大，这种方法很难应对新的领域和新的概念。

2）利用命名实体识别（Named Entity Recognition，NER）工具，从网页片段中抽取命名实体（Named Entity，NE）作为候选答案。XU 等人[64]实现的基于 NER 的答案抽取算法就是利用 NER 工具抽取出候选答案，然后利用一些启发式的规则筛选排序这些候选答案。这种方法的具体效果受问题分类算法和命名实体识别算法效果的影响。

3）根据手工编辑或自动生成的文本模式抽取候选答案[65-66]。这种方法有很高的准确率，但是由于文本模式过于精细，导致其在匹配过程中较为死板，无法适应新的数据。

3.3.2 候选答案排序

候选答案排序是 WQA 系统的最后一个步骤，其主要目的是通过对前一步骤得到的候选答案进行排序来找出最佳答案。虽然经过候选答案抽取的筛选已经缩小了最佳答案的范围，但是仍然存在大量的高质量候选答案需要排序。排序候选答案时，需要对候选答案进行综合考虑，并对候选答案的质量进行量化分析。只有将候选答案的质量进行量化，才能通过排序算法找出最佳答案。

目前存在很多候选答案排序及最佳答案选择的方法，主要包括如下几种典型的方法。

1）采用向量空间模型（Vector Space Model，VSM）计算候选答案与问题的相似度，并以此进行排序。余正涛等人[67]基于 VSM 方法，利用潜在语义分析（Latent Semantic Analysis，LSA）[68]实现了汉语问答系统的答案提取。

2）根据语法结构判断候选答案与问题的匹配度，并以此进行排序。YAO 等人[61]提出的基于语法树编辑距离的答案抽取方法以及 SUN 等人[69]提出的基于因子图的答案抽取方法，就是属于这一类的候选答案排序方法。

3)根据词汇特征、相似度特征、统计特征等多种特征进行综合排序[60-70]。KHODADI 等人[71]提出的基于遗传算法的问答系统就利用了遗传算法(Genetic Algorithm,GA)[72]擅长解决全局最优化问题的特点,从而实现了对候选答案的多特征综合排序。

4　WQA 面临的主要问题

通过分析 WQA 关键技术的研究进展可以看出,当前的 WQA 在问题分析、信息检索、答案抽取 3 个方面都取得了相当大的进步。特别是近几年来,国内外各大公司都投入了大量资源来研究问答系统,带动了 WQA 相关技术的发展。但从整体来看,WQA 仍有很大的提升空间,其主要面临以下几个问题。

1)问题分类有待改善:虽然现有的规则分类器、SVM 分类器、神经网络分类器等已经能够分类大部分的问题,但是这些分类器的效果仍然有待提升[54]。问题分类的本质是短文本分类,而短文本分类受制于特征稀缺的问题,一直没有特别大的突破[73]。

2)同义句子的理解需要解决:目前的 WQA 系统难以很好地理解文本的语义。自然语言非常灵活,同一句话可能有很多种不同的表达方式。无论是同义词的使用,还是语法结构的变化,都会使得 WQA 系统难以准确地抽取答案[74]。

3)高质量的问答对(Question-Answer Pairs)难以获取:WQA 系统的一个重要功能是解决事实型问题。目前,高质量的事实型问答对数据(特别是高质量的中文问答对数据)不足。

4)利用跨语言语料的能力较差:现在的 WQA 系统无法借助多种语言的语料数据。搜索引擎返回的网页片段是 WQA 系统进行答案抽取的重要依据。网页片段可能存在多种语言,但是目前的 WQA 系统还难以利用多种语言的文本数据来回答某一种特定语言的问题。

5)通用性不足:目前的 WQA 系统在回答特定领域的问题或者特定类型的问题时有较好的效果,但回答通用领域问题的能力尚有待进一步增强。

6)处理复杂问题的能力不足:目前的 WQA 系统一般能够较好地处理和回答事实型问题。但是对于定义型(Definition)、原因型(Why)、关系型(Relation)、比较型(Comparison)、方法型(How-to)等问题,WQA 系统尚不能给出满意的回答。

5　WQA 的发展趋势

通过分析 WQA 的研究现状,总结 WQA 面临的问题,本文认为 WQA 需要在以下几个方面继续进行深入的研究。

1)与其他问答系统的融合:每一种问答系统都有其独特的优势。WQA 系统与其他问答系统配合使用将会产生更好的效果。目前已出现了一些混合问答系统,但尚未完全成熟。本文认为,将 KBQA,CQA 和 WQA 3 种技术混合使用,将是问答系统的发展方向。一种简单的混用策略就是设定一个优先级,比如 KBQA>CQA>WQA,依次利用 KBQA 和 CQA 回答用户提出的问题,当 KBQA 和 CQA 失效时,则由 WQA 负责回答用户的问题。

2)通过答案摘要生成答案:现在的 WQA 系统大多数都是通过答案抽取的方式生成答案,但是 WQA 利用网页片段列表生成答案的过程的本质是解决一个多文档摘要的问题。当信息抽取技术不足以完全解决 WQA 的问题时,可以考虑利用多文档摘要技术。

3)自动生成高质量问答对数据:人工标注 WQA 系统所需的问答对不是一个好的解决方案,可以利用知识库自动生成一些问答对[75],或者从 CQA 的相应数据集中抽取出一些适合 WQA 系统使用的高质量问答对。大规模的问答对数据集能帮助答案抽取算法学习到更好的候选答案抽取知识与候选答案排序规律。

4)提升 WQA 系统处理复杂问题的能力:除事实型问题之外,WQA 系统还应能处理和回答一些更为复杂的问题,例如:原因型(Why)、定义型(Definition)、关系型(Relation)、比较型(Comparison)、方法型(How-to)等问题。对于定义型问题,可利用维基百科、百度百科、百科全书等资料抽取相关条目的定义。对于比较型问题,例如:"北京与上海,哪个城市更好?","苹果手机与华为手机,有什么不同?",则需要首先从问句中识别出两个比较的对象,然后通过搜索引擎收集到这两个比较对象的相关信息,最后列举出其在各个方面的异同点。对于方法型问题,则需要列出完成相关任务的一系列步骤,甚至给出说明一步一步完成该任务的视频。

5)跨语言能力、跨领域能力的进一步增强:跨语言能力和跨领域能力将会大大提高 WQA 的应用价值。随着人工智能和自然语言处理的不断发展,WQA 的这两大能力必然会继续增强。

6)与语音识别、语音生成等工具的进一步结合:随着语音识别与语音生成技术的成熟,个人和企业都可以非常方便地调用第三方封装好的语音接口。语音和文本是人们工作、生活中最重要的两种交流方式。WQA 系统也必将进一步与语音处理的相关工具结合使用。

7)辅助机器人:机器人是未来人工智能的一大热门方向。WQA 系统可以作为机器人内置的一种辅助系统,负责帮助机器人理解并回答人们提出的问题。

结束语　作为一种特殊的问答系统,WQA 从搜索引擎返回的搜索结果中抽取用户所需的答案。因此,WQA 同时拥有问答系统和搜索引擎的优点。WQA 可以单独使用,也可以与其他问答系统一起混合使用。对 WQA 进行深入研究能够促进问答系统的进一步发展。

本文阐述了 WQA 系统的发展历程,详细介绍了 WQA 系统关键技术的研究进展,并剖析了其在现阶段所存在的问题,最后对 WQA 的发展趋势进行了展望。

参 考 文 献

[1] MAO X L, LI X M. A survey on question and answering systems[J]. Journal of Frontiers of Computer Science & Technology, 2012, 6(3): 193-207. (in Chinese)
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.

[2] KEYES R W. The impact of Moore's Law[J]. IEEE Solid-State Circuits Newsletter, 2006, 20(3): 25-27.

[3] YANG D, YU K. "Internet+" Epoch Social Management Innovation: Challenge and Response-To the Case of Shanghai Taxi Operations Management[J]. International Journal of Social Science Studies, 2015, 3(6): 197-201.

[4] YANG M C, LEE D G, PARK S Y, et al. Knowledge-based question answering using the semantic embedding space[J]. Expert Systems with Applications, 2015, 42(23): 9086-9104.

[5] CUI W, XIAO Y, WANG W. KBQA: an Online Template Based Question Answering System over Freebase[C]// Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York: IJCAI/AAAI Press, 2016: 4240-4241.

[6] LIU Y, LI S, CAO Y, et al. Understanding and summarizing answers in community-based question answering services[C]// Proceedings of the 22nd International Conference on Computational Linguistics. Manchester: Coling 2008 Organizing Committee, 2008: 497-504.

[7] ZHANG K, WU W, WANG F, et al. Learning distributed representations of data in community question answering for question retrieval[C]// Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. San Francisco: ACM, 2016: 533-542.

[8] SUN H, WEI F, ZHOU M. Answer Extraction with Multiple Extraction Engines for Web-Based Question Answering[M]// Natural Language Processing and Chinese Computing. Shenzhen: Springer, 2014: 321-332.

[9] SU F, GAO D L, YE C. Study on Question Understanding of Web-based Question-answering System[J]. Journal of Test and Measurement Technology, 2012, 26(3): 207-212. (in Chinese)
苏斐, 高德利, 叶晨. Web 问答系统中问句理解的研究[J]. 测试技术学报, 2012, 26(3): 207-212.

[10] TURING A M. Computing machinery and intelligence[J]. Mind, 1950, 59(236): 433-460.

[11] GREEN B F, JR, WOLF A K, et al. Baseball: an automatic question-answerer[C]// Proceedings of the Western Joint Computer Conference. Los Angeles: ACM, 1961: 219-224.

[12] WOODS W A, KAPLAN R. Lunar rocks in natural English: Explorations in natural language question answering[J]. Linguistic Structures Processing, 1977, 5(1): 521-569.

[13] WEIZENBAUMJ. ELIZA--a computer program for the study of natural language communication between man and machine[J]. Communications of the ACM, 1966, 9(1): 36-45.

[14] KATZ B. From sentence processing to information access on the world wide web[C]// AAAI Spring Symposium on Natural Language Processing for the World Wide Web. Stanford: Stanford University, 1997: 22-25.

[15] ZHENG Z. AnswerBus question answering system[C]// Proceedings of the Second International Conference on Human Language Technology Research. San Diego: Morgan Kaufmann Publishers Inc., 2002: 399-404.

[16] HOVY E H, GERBER L, HERMJAKOB U, et al. Question Answering in Webclopedia[C]// Proceedings of The Ninth Text REtrieval Conference. Gaithersburg: National Institute of Standards and Technology, 2000: 53-56.

[17] DRENOYIANNI H, SELWOOD I, RIDING R. Searching Using 'Microsoft? Encarta?'[J]. Education and Information Technologies, 2002, 7(4): 333-342.

[18] ZHANG D, LEE W S. Web Based Pattern Mining and Matching Approach to Question Answering[C]// Proceedings of The Eleventh Text REtrieval Conference. Gaithersburg: National Institute of Standards and Technology, 2002: 129-144.

[19] STEPHEN. Wolfram Language Artificial Intelligence: The Image Identification Project[EB/OL]. http://blog.stephenwolfram.com/2015/05/wolfram-language-artificial-intelligence-the-image-identification-project.

[20] FIVEASH K. Wolfram Alpha given keys to the Bingdom[EB/OL]. http://www.theregister.co.uk/2009/11/12/bing_wolfram_alpha_deal.

[21] VOLKMER T, SMITH J R, NATSEV A P. A web-based system for collaborative annotation of large image and video collections: an evaluation and user study[C]// Proceedings of the 13th annual ACM international conference on Multimedia. Singapore: ACM, 2005: 892-901.

[22] SCOBLEIZER. BREAKING NEWS: Siri bought by Apple[EB/OL]. http://scobleizer.com/2010/04/28/breaking-news-siri-bought-by-apple.

[23] 麒麟会. 微软亚洲互联网工程院将分享"小冰"背后的故事[EB/OL]. http://tech.ifeng.com/a/20140928/40825530_0.shtml.

[24] 凤凰网. 就在今晚!百度机器人将大战"最强大脑"选手[EB/OL]. http://ent.ifeng.com/a/20170106/42804563_0.shtml?_zbs_baidu_bk.

[25] KWOK C, ETZIONI O, WELD D S. Scaling question answering to the web[J]. ACM Transactions on Information Systems, 2001, 19(3): 242-262.

[26] WANG M. A survey of answer extraction techniques in factoid question answering[J]. Computational Linguistics, 2006, 1(1): 1-14.

[27] YANG H, CHUA T S. Effectiveness of web page classification on finding list answers[C]// Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. Sheffield: ACM, 2004: 522-523.

[28] GONÇALVES P N, BRANCO A. Open-domain web-based list question answering with LX-listquestion[C]// Proceedings of the 4th International Confe-rence on Web Intelligence, Mining and Semantics. Thessaloniki: ACM, 2014: 43-49.

[29] GONÇALVES P N, BRANCO A H. A Comparative Evaluation of QA Systems over List Questions[C]// International Conference on Computational Processing of the Portuguese Language. Tomar: Springer, 2016: 115-121.

[30] REN H, JI D, TENG C, et al. A web knowledge based approach for complex question answering[C]// Asia Information Retrieval Symposium. Dubai: Springer, 2011: 470-478.

[31] SAVENKOV D. Ranking Answers and Web Passages for Non-factoid Question Answering: Emory University at TREC LiveQA[C]// Proceedings of The Twenty-Fourth Text REtrieval Conference. Gaithersburg: National Institute of Standards and Technology, 2015: 1-8.

[32] MOSCHITTI A, MàRQUEZ L, NAKOV P, et al. SIGIR 2016 Workshop WebQA II: Web Question Answering Beyond Factoids[C]// Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. Pisa: ACM, 2016: 1251-1252.

[33] AGICHTEIN E, CARMEL D, CLARKE C L A, et al. Web question answering: Beyond factoids: SIGIR 2015 workshop[C]// Proceedings of the 38th International ACM SIGIR Confe-rence on Research and Development in Information Retrieval. Santiago: ACM, 2015: 1143-1143.

[34] QUARTERONI S, MANANDHAR S. Designing an interactive open-domain question answering system[J]. Natural Language Engineering, 2009, 15(1): 73-95.

[35] LI S, LIN C Y, SONG Y I, et al. Comparable entity mining from comparative questions[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(7): 1498-1509.

[36] YIH W, MA H. Question Answering with Knowledge Base, Web and Beyond[C]// Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa: ACM, 2016: 1219-1221.

[37] WU G S, LAN M. Leverage Web-based Answer Retrieval and Hierarchical Answer Selection to Improve the Performance of Live Question Answering[C]// Proceedings of The Twenty-Fourth Text REtrieval Conference. Gaithersburg: National Institute of Standards and Technology, 2015: 1-8.

[38] CAO Z J, LI Z S, LIU C T. Study of Question Analysis in Question-Answering System[J]. Computer Science, 2005, 32(11): 158-160. (in Chinese)
曹志娟，李祖枢，刘朝涛. 自动问答系统中的问题理解研究[J]. 计算机科学，2005，32(11)：158-160.

[39] LI X, ROTH D. Learning question classifiers: the role of semantic information[J]. Natural Language Engineering, 2006, 12(3): 229-249.

[40] WEN X, ZHANG Y, LIU T, et al. Syntactic Structure Parsing Based Chinese Question Classification [J]. Journal of Chinese Information Processing, 2006, 20(2): 35-41. (in Chinese)
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，2006，20(2)：35-41.

[41] LIU Z J, WANG X L, CHEN Q C, et al. A Chinese question answering system based on Web search[C]// International Conference on Machine Learning and Cybernetics. Lanzhou: IEEE, 2014: 816-820.

[42] LI X, HU D, LI H, et al. Automatic question answering from Web documents[J]. Wuhan University Journal of Natural Sciences, 2007, 12(5): 875-880.

[43] ZHANG Z C, ZHANG Y, LIU T, et al. Advances in open-domain question answering[J]. Acta Electronica Sinica, 2009, 37(5): 1058-1069. (in Chinese)
张志昌，张宇，刘挺，等. 开放域问答技术研究进展[J]. 电子学报，2009，37(5)：1058-1069.

[44] ZHANG D, LEE W S. Question classification using support vector machines[C]// Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. Toronto: ACM, 2003: 26-32.

[45] CHALI Y, HASAN S A, MOJAHID M. A reinforcement learning formulation to the complex question answering problem [J]. Information Processing & Management, 2015, 51(3): 252-272.

[46] SUZUKI J, TAIRA H, SASAKI Y, et al. Question classification using HDAG kernel[C]// Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering. Sapporo: ACL, 2003: 61-68.

[47] MOSCHITTI A, QUARTERONI S, BASILI R, et al. Exploiting syntactic and shallow semantic kernels for question answer classification[C]// Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Prague: ACL, 2007: 776-783.

[48] POTA M, ESPOSITO M, DE P G. A Forward-Selection Algorithm for SVM-Based Question Classification in Cognitive Systems[M]. Switzerland: Springer, 2016: 587-598.

[49] MCROY S, JONES S, KURMALLY A. Toward automated classification of consumers' cancer-related questions with a new taxonomy of expected answer types [J]. Health Informatics Journal, 2016, 22(3): 523-535.

[50] YANG M H, AHUJA N. Learning to Detect Faces with Snow [M]// Face Detection and Gesture Recognition for Human-Computer Interaction. 2001: 123-150.

[51] LI X, ROTH D. Learning question classifiers: the role of semantic information[J]. Natural Language Engineering, 2006, 12(3): 229-249.

[52] MERKEL A, KLAKOW D. Language model based query classification [C]// Advances in Information Retrieval. Rome: Springe, 2007: 720-723.

[53] LIN S J, LU W H. Learning question focus and semantically related features from web search results for chinese question classification[C]// The Third Asia Information Retrieval Symposium. Singapore: Springer, 2006: 284-296.

[54] ANAND K M, SOMAN K P. Amrita_CEN@ MSIR-FIRE2016: Code-Mixed Question Classification using BoWs and RNN Embeddings[C]// Working notes of Forum for Information Retrieval Evaluation. Kolkata: CEUR-WS, 2016: 122-125.

[55] MOLDOVAN D, HARABAGIU S, PASCA M, et al. The structure and performance of an open-domain question answering system[C]// The 38th Annual Meeting of the Association for Computational Linguistics. Hong Kong: ACL, 2000: 563-570.

[56] BRILL E, DUMAIS S, BANKO M. An analysis of the AskMSR question-answering system [C]// Proceedings of the ACL-02 Conference on Empirical Methods in Natural Anguage. Stroudsburg: ACL, 2002: 257-264.

[57] BRILL E, LIN J J, BANKO M, et al. Data-Intensive Question Answering[C]// Proceedings of The Tenth Text REtrieval Conference. Gaithersburg: National Institute of Standards and Technology, 2001: 393-400.

[58] GONÇALVES P N, BRANCO A. Answering List Questions using Web as a corpus[C]// Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg: ACL, 2014: 81-84.

(下转第 42 页)

[12] CORBETT P, ENGLISH B, GOEL A, et al. Row-diagonal parity for double disk failure correction[C]// Proceedings of the 3rd USENIX Symposium on File and Storage Technologies (FAST' 04). 2004: 1-14.

[13] XU L, BRUCK J. X-Code: MDS Array Codes with Optimal Encoding[J]. IEEE Transactions on Information Theory, 1999, 45 (1): 272-276.

[14] CHAO J, HONG J, DAN F, et al. P-Code: A new RAID-6 code with optimal properties[C]// 23rd International Conference on Supercomputing. New York, June 2009.

[15] HUANG C, XU L. STAR: An Efficient Coding Scheme for Correcting Triple Storage Node Failures[J]. IEEE Transactions on Computers, 2007, 57(7): 889-901.

[16] LIN S, WANG G, STONES D S, et al. T-Code: 3-Erasure Longest Lowest-Density MDS Codes[J]. IEEE Journal on Selected Areas in Communications, 2010, 28(2): 289-296.

[17] AUTHORS U. WEAVER codes: highly fault tolerant erasure codes for storage systems[C]// Fast 05 Conference on File & Storage Technologies, 2005: 16.

[18] PLANK J S, SCHUMAN C D, ROBISON B D. Heuristics for optimizing matrix-based erasure codes for fault-tolerant storage systems[J]. IEEE/IFIP International Conference on Dependable Systems & Networks Annual, 2012, 122(12): 1-12.

[19] SIMD[EB/OL]. [2016-07-09]. https://en. wikipedia. org/w/index. php?title=SIMD&oldid=726575429.

[20] PLANK J S. A tutorial on Reed-Solomon coding for fault-tolerance in RAID-like systems[J]. Softw. , Pract. Exper. , 1997, 27 (9): 995-1012.

(上接第 7 页)

[59] SUN H, MA H, YIH W, et al. Open domain question answering via semantic enrichment[C]// Proceedings of the 24th International Conference on World Wide Web. Florence: ACM, 2015: 1045-1055.

[60] SEVERYN A, MOSCHITTI A. Automatic Feature Engineering for Answer Selection and Extraction[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: ACL, 2013: 458-467.

[61] YAO X, VAN Durme B, CALLISON B C, et al. Answer Extraction as Sequence Tagging with Tree Edit Distance[C]// Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics. Atlanta: ACL, 2013: 858-867.

[62] SREELAKSHMI V, JAMAL S. Web Based Question Answering System using Pattern Matching[C]// The International Conference on Information Science. Pattaya: IEEE, 2015: 1-4.

[63] CHU CARROLL J, FAN J. Leveraging Wikipedia Characteristics for Search and Candidate Generation in Question Answering [C]// Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2011: 872-877.

[64] XU J, LICUANAN A, MAY J, et al. Answer Selection and Confidence Estimation[C]// New Directions in Question Answering. Stanford: AAAI Press, 2003: 134-137.

[65] ZHANG D, LEE W S. Web Based Pattern Mining and Matching Approach to Question Answering[C]// Proceedings of The Eleventh Text REtrieval Conference. Gaithersburg: National Institute of Standards and Technology, 2002: 129-141.

[66] MEDITSKOS G, DASIOPOULOU S, VROCHIDIS S, et al. Question Answering over Pattern-Based User Models[C]// Proceedings of the 12th International Conference on Semantic Systems. Leipzig: ACM, 2016: 153-160.

[67] YU Z T, FAN X Z, GUO J Y, et al. Answer extracting for chinese question-answering system based on latent semantic analysis[J]. Chinese Journal of Computer, 2006, 29(10): 1889-1893. (in Chinese)
余正涛, 樊孝忠, 郭剑毅, 等. 基于潜在语义分析的汉语问答系统答案提取[J]. 计算机学报, 2006, 29(10): 1889-1893.

[68] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391.

[69] SUN H, DUAN N, DUAN Y, et al. Answer Extraction from Passage Graph for Question Answering[C]// Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing: IJCAI/AAAI, 2013: 2169-2175.

[70] FIGUEROA A G, NEUMANN G. Genetic algorithms for data-driven web question answering[J]. Evolutionary Computation, 2008, 16(1): 89-125.

[71] KHODADI I, ABADEH M S. Genetic programming-based feature learning for question answering[J]. Information Processing & Management, 2016, 52(2): 340-357.

[72] MA Y J, YUN W X. Research progress of genetic algorithm[J]. Application Research of Computers, 2012, 29(4): 1201-1206. (in Chinese)
马永杰, 云文霞. 遗传算法研究进展[J]. 计算机应用研究, 2012, 29(4): 1201-1206.

[73] MA C L, YAN Y H. Short Text Classification Based on Probabilistic Semantic Distribution[J]. Acta Automatica Sinica, 2016, 42(11): 1711-1717. (in Chinese)
马成龙, 颜永红. 基于概率语义分布的短文本分类[J]. 自动化学报, 2016, 42(11): 1711-1717.

[74] MA L. The Research and Implementation of Web-based Chinese Question Answering System[D]. Beijing: Beihang University, 2012. (in Chinese)
马琳. 基于 Web 的中文问答系统的研究与实现[D]. 北京: 北京航空航天大学, 2012.

[75] LEE J, KIM G, YOO J, et al. Training IBM Watson using Automatically Generated Question-Answer Pairs[C]// The 50th Hawaii International Conference on System Sciences. Hawaii: AIS Electronic Library, 2017: 1-9.