“十三五”国家重点图书出版规划

中国人工智能学会推荐

人工智能丛书

智能问答

Question Answering

段楠　周明

高等教育出版社

“十三五”国家重点图书出版规划

中国人工智能学会推荐

人工智能丛书

智能问答

Question Answering

段楠 周明

高等教育出版社·北京

内容简介

智能问答是人工智能领域中一个极为重要的课题，并且在现代搜索引擎和人机对话系统中发挥着极为关键的作用。全书共分10章，力求对智能问答研究中涉及的典型任务和主流方法进行全面系统地介绍。首先，对智能问答的历史沿革、任务分类和常见问答数据集进行简单回顾和总结。然后，对智能问答研究中涉及的常见机器学习理论进行概要说明，保证了本书在内容上的系统性和完备性。接下来，重点讲述智能问答研究中五个最具代表性的任务，包括知识图谱问答、表格问答、文本问答、社区问答和问题生成。针对每类任务，本书选择典型的并具有持续发展性的方法进行介绍。最后，总结全文，并对未来发展进行展望。此外，本书末尾还通过FAQ的形式回答读者最可能提出的常见问题。

本书既可以作为人工智能专业和自然语言处理专业的教学参考书，也可以供从事智能问答相关产品的开发人员参考和借鉴。

图书在版编目(CIP)数据

智能问答／段楠，周明主编. --北京：高等教育出版社，2018.9

ISBN 978-7-04-050244-2

Ⅰ. ①智… Ⅱ. ①段… ②周… Ⅲ. ①智能技术 Ⅳ. ①TP18

中国版本图书馆CIP数据核字(2018)第170279号

策划编辑 张江漫	责任编辑 张江漫	封面设计 赵 阳	版式设计 徐艳妮
插图绘制 于 博	责任校对 陈 杨	责任印制 韩 刚	

出版发行	高等教育出版社	网 址	http://www.hep.edu.cn
社 址	北京市西城区德外大街4号		http://www.hep.com.cn
邮政编码	100120	网上订购	http://www.hepmall.com.cn
印 刷	廊坊市文峰档案印务有限公司		http://www.hepmall.com
开 本	787mm×960mm 1/16		http://www.hepmall.cn
印 张	12.5		
字 数	220千字	版 次	2018年9月第1版
购书热线	010-58581118	印 次	2018年9月第1次印刷
咨询电话	400-810-0598	定 价	39.90元

物料号 50244-00

智能问答

段楠　周明

1　计算机访问http://abook.hep.com.cn/1253561，或手机扫描二维码、下载并安装Abook应用。

2　注册并登录，进入“我的课程”。

3　输入封底数字课程账号（20位密码，刮开涂层可见），或通过Abook应用扫描封底数字课程账号二维码，完成课程绑定。

4　单击“进入课程”按钮，开始本数字课程的学习。

课程绑定后一年为数字课程使用有效期。受硬件限制，部分内容无法在手机端显示，请按提示通过计算机访问学习。

如有使用问题，请发邮件至abook@hep.com.cn。

http://abook.hep.com.cn/1253561

人工智能丛书编委会

前　言

1950年，图灵在《Computing Machinery and Intelligence》一文中提出图灵测试(Turing Test)这一概念。从那时起，构建能够完成特定任务的人机交互系统，就成为人类为提高自身生产力而不断追求和奋斗的目标。早期的人机交互系统主要以基于规则的受限领域专家系统为主。由于这类系统严重依赖领域专家总结和撰写的规则模板，因此扩展性很弱。20世纪90年代，互联网的兴起极大地促进了开放领域搜索引擎的出现和发展。以Yahoo和Google为代表的搜索引擎允许人们使用自然语言输入查询，并从整个互联网中检索与用户查询最相关的网页集合作为返回结果。进入21世纪，随着移动终端设备的普及和语音识别技术的成熟，以Apple Siri、Microsoft Cortana、Amazon Alexa和Google Assistant为代表的智能语音助手比以往任何一个时期都更贴近人们的日常生活。该类系统允许人们使用自然语言文字或语音与之进行交互，并返回自然语言形式的响应结果。作为搜索引擎和智能语音助手的核心功能，智能问答(question answering)近年来受到学术界和工业界的一致关注和深入研究，各种问答数据集和方法层出不穷。

智能问答旨在为用户提出的自然语言问题自动提供答案。答案候选来自预先指定的问答知识库，例如知识图谱、表格集合、文档集合、<问题，答案>对集合等。按照所使用问答知识库类型的不同，本书将智能问答任务分为四类，并将逐一进行系统介绍。

(1) 知识图谱问答(knowledge-based question answering)，使用知识图谱作为问答知识库，问题的答案既可以是来自知识图谱的实体集合，也可以是基于知识图谱推理出来的内容。该类任务可以采用语义分析和答案排序两种方法完成。前者首先将输入问题转化为知识图谱能够理解的语义表示进行查询，然后据此从知识图谱中检索得到问题对应的答案。后者首先根据输入问题从知识图谱中筛选得到答案实体候选集合，然后通过对不同答案实体候选进行打分和排序，得到问题对应的答案。

(2) 表格问答(table-based question answering)，使用从互联网上抓取得到的

表格集合作为问答知识库，问题的答案既可以是某个表格，也可以是基于表格推理得到的内容。前者对应表格检索任务，其核心目标是从给定表格集合中找到与输入问题最相关的表格。后者对应答案生成任务，其核心目标是基于表格检索的返回结果，推理并生成问题对应的精准答案。和知识图谱问答类似，答案生成的过程同样可以采用基于语义分析的方法或基于答案排序的方法来完成。

(3) 文本问答(text-based question answering)，使用无结构文本作为问答知识库，问题的答案既可以是输入文本中的句子，也可以是输入文本中的单词或短语。前者对应答案句子选择任务，其核心目标是计算问题和答案句子候选之间的相关性。后者对应机器阅读理解任务，其核心目标是根据问题定位答案在输入文本中的起始和终止位置。

(4) 社区问答(community question answering)，使用从社区问答网站上抓取的<问题，答案>对作为问答知识库，问题的答案来自于和输入问题语义最为匹配的已有问题对应的答案。该类任务的核心目标之一是计算问题和已有问题之间的相似度。

除了上述四类问答任务外，本书还将介绍问题生成(question generation)任务。作为智能问答的对偶任务，问题生成任务根据给定内容生成该内容可能对应的问题。实验证明，该技术能够从数据和模型训练两个角度进一步提升智能问答系统的性能。

本书力求系统地介绍智能问答研究领域中的任务和方法，并尽量选择具有代表性和延续性的工作进行具体说明。随着人工智能热潮的来临，关于智能问答的研究成果层出不穷。对于尚未包含在内的新任务、新方法和新观点，计划在后续写作中予以补充。

本书初稿完成后，微软亚洲研究院自然语言计算组的同事和学生对本书内容提出许多宝贵意见，在此向他们表示衷心的感谢。作者还要感谢高等教育出版社对本书出版的全力支持。

由于作者水平有限，书中难免有错误和不当之处，欢迎专家和读者给予批评和指正。来函请发至 qa-book-nanduan@ outlook.com。

作者

2018 年

目　录

第1章 概 述

智能问答(question answering,QA)旨在为用户提出的自然语言问题自动提供答案。得益于互联网数据的海量增长、硬件计算能力的飞速提高以及自然语言处理和深度学习技术的长足进步,智能问答方法在近年来取得了突飞猛进的发展,其应用场景(例如搜索引擎和智能语音助手等)比以往任何一个历史时期都更贴近人们的日常生活。

本书的目的是系统介绍智能问答研究涉及的典型任务和方法,希望能够帮助读者快速了解该领域的相关进展,以便开展进一步学习和工作。作为开端,本章将简要回顾智能问答的历史沿革,对智能问答任务进行分类,并介绍典型的智能问答评测任务。

1.1 历史沿革

1.1.1 受限领域问答

20世纪60年代,智能问答研究主要是针对数据库自然语言接口(natural language interfaces to database)任务,即如何使用自然语言检索结构化数据库。Green等人提出BASEBALL系统[1],用于回答用户通过自然语言提出的关于美国棒球联赛的问题。Woods提出LUNAR系统[2],允许月球地质学家使用自然语言检索NASA数据库,以获取月球岩石和土壤相关的信息和数据。上述两个工作都需要依赖人工撰写的规则模板,完成从自然语言问题到结构化数据库查询语句的转换工作。

20世纪70年代,智能问答研究聚焦于对话系统(dialogue system),并在70年代末期拓展到阅读理解任务(reading comprehension)。Winograd提出

SHRDLU 系统[3]，支持用户使用自然语言控制模拟程序中的积木完成各种操作，并允许用户对积木的状态进行自然语言形式的提问。该系统是智能问答在对话系统中的早期应用之一。Lehnert 提出 QUALM 系统[4]，用于完成阅读理解任务。该系统将自然语言问题分为 13 个类别，针对不同类型的问题采用不同的策略从文本中寻找问题对应的答案。

20 世纪 80 年代，受限领域知识库的构建工作继续向前发展，进一步推动了基于知识库的专家系统研究。Shortlie 提出 MYCIN 系统[5]，用于识别引发感染的病毒并根据患者体重等信息推荐抗生素。该系统基于推理引擎和一个包含 600 条规则的知识库。Wilensky 等人提出 UNIX Consultant 系统[6]，允许用户使用自然语言完成 UNIX 操作系统中的特定任务。和 MYCIN 类似，该系统基于人工撰写的规则完成自然语言理解和回复生成任务。

早期智能问答系统大都针对特定领域构建，并且需要领域专家撰写大量领域相关的规则用于问题理解和答案生成，这极大地限制了该类智能问答系统的规模和通用性。此外，受限于自然语言处理水平和计算机的运算能力，这类系统的准确度也差强人意。

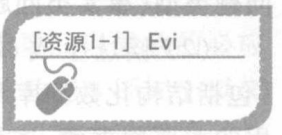
1.1.2 开放领域问答

1993 年，第一个基于互联网的智能问答内容系统 START[7] 由 MIT（麻省理工学院）开发上线。该系统使用结构化知识库和非结构化文档作为问答知识库。对于能够被结构化知识库回答的问题，系统直接返回问题对应的精准答案。否则，START 首先对输入问题进行句法分析并根据分析结果抽取关键词；然后，基于抽取出来的关键词从非结构化文档集合中找到与之相关的文档集合；最后，采用答案抽取技术从相关文档中抽取可能的答案候选进行打分，并选择得分最高的句子作为答案输出。

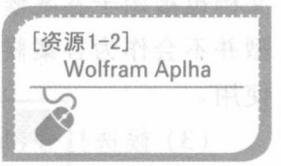
1999 年，TREC（text retrieval conference）举办第一届开放领域智能问答评测任务 TREC-8[8]，目标是从大规模文档集合中找到输入问题对应的相关文档。该任务从信息检索角度开创了智能问答研究的一个崭新方向，受其影响，越来越多的研究者投身到智能问答的研究中来。可以说，TREC 问答评测是世界范围上最受关注和最具影响力的问答评测任务之一。

受 TREC 影响，CLEF（cross-language evaluation forum）在 2000 年提出跨语言问答评测任务[9]。在该任务中问题和检索文档集合分别使用不同的语言，这就需要在问答中考虑并运用机器翻译技术。此外，CLEF 引入了包括推理类和动机类在内的复杂问题，用于检验问答系统处理这类问题的水平。

Evi[资源1-1]是2005年上线的问答型搜索引擎。和传统搜索引擎不同，Evi基于开放领域知识库对用户提出的自然语言问题进行问题理解，并根据问题理解的结果从知识库中查找出问题对应的精准答案，其核心技术是基于知识图谱的智能问答技术。该产品在2012年被Amazon公司收购。



Wolfram Aplha[资源1-2]是2009年上线的计算知识引擎(computational knowledge engine)。该系统源于Wolfram早期的旗舰产品Mathematia，因此支持计算机代数、符号和数值计算、可视化和统计等功能。此外，Wolfram Aplha还提供基于结构化知识库的问答功能。



2011年，由IBM构建的智能问答系统Watson[10]参加了美国电视问答比赛节目Jeopardy!，并在比赛中击败了人类冠军选手。Jeopardy! 问答比赛涵盖了包括历史、语言、文学、艺术、科技、流行文化、体育、地理和文字游戏等多方面问题。每个问题对应的线索(例如He was a bank clerk in the Yukon before he published "Songs of a Sourdough" in 1907)在逐条展示给选手的过程中，选手需要根据已有提示尽快猜出问题对应答案。图1-1给出IBM Watson的系统架构图，它主要由下述四个模块组成：

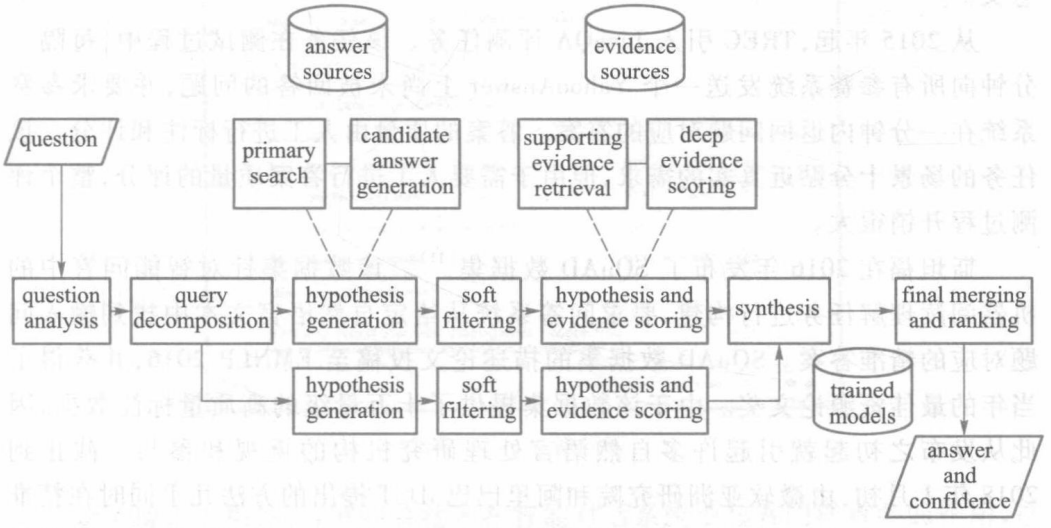


图1-1 IBM Watson的系统架构图[10]

(1) 问题分析(question analysis)。该模块负责对输入问题进行包括命名实体识别、关系分类、句法解析、指代消歧等在内的多种自然语言处理操作，并根据处理结果抽取问题焦点(question focus)、词汇化答案类型(lexical answer type)和

问题类型(事实类问题/非事实类问题)。

(2) 候选生成(hypothesis generation)。该模块负责从不同类型的数据源(包括结构化数据库和非结构化文档)中抽取答案候选。维基百科是 Watson 使用的主要数据源,这是由于 Jeopardy! 中 95%的问题对应的答案都能在维基百科的标题中找到。除了维基百科外,知识图谱、词典、新闻文档和其他类型百科文档也被用于答案候选抽取。需要注意的是,问题分析模块预测得到的答案类型并不会作为答案候选抽取的硬性过滤条件,而是在答案候选排序中作为特征使用。

(3) 候选打分(hypothesis scoring)。该模块使用多种特征,从不同角度对每个答案候选进行置信度打分。

(4) 答案合并和排序(answer merging and ranking)。该模块基于每个答案候选对应的多个置信度得分,使用逻辑斯蒂回归计算每个答案候选对应的最终得分。答案候选集合可能包含同一个答案的不同表达形式(例如 JFK、John Fitzgerald Kennedy 和 President Kennedy 都表示同一个人),系统将对不同表达形式的得分进行合并。

IBM Watson 是现代智能问答系统在特定领域成功应用的典范,其背后所采用的系统架构和所使用的问答知识库类型,对现代智能问答研究极具借鉴意义。

从 2015 年起,TREC 引入 LiveQA 评测任务。该任务在测试过程中,每隔一分钟向所有参赛系统发送一个 YahooAnswer 上尚未被回答的问题,并要求参赛系统在一分钟内返回问题对应的答案。答案的质量由人工进行标注和评分。该任务的场景十分贴近真实的需求,但由于需要人工进行答案质量的评分,整个评测过程开销很大。

斯坦福在 2016 年发布了 SQuAD 数据集[11]。该数据集针对智能问答中的机器阅读理解任务进行构建,要求问答系统从给定自然语言文本中找到输入问题对应的精准答案。SQuAD 数据集的描述论文投稿至 EMNLP 2016,并获得了当年的最佳资源论文奖。由于该数据集提供了十万量级的高质量标注数据,因此从发布之初起就引起许多自然语言处理研究机构的重视和参与。截止到 2018 年 1 月初,由微软亚洲研究院和阿里巴巴 iDST 提出的方法几乎同时在精准匹配(exact match, EM)这一指标上超过了 Amazon Mechanical Turk 标注者的平均水平,这可以说是深度学习模型在自然语言处理任务上的一次成功应用。在中文方面,百度发布了中文机器阅读理解数据集 DuReader,并在 2018 年 3 月联合中国中文信息学会和中国计算机学会共同举办了“2018 机器阅读理解技术竞赛”。哈尔滨工业大学讯飞联合实验室也从 2017 年起开始举办“讯飞杯”中文

机器阅读理解评测。

在现代搜索引擎中,智能问答技术所起的作用越来越重要。以微软的必应搜索引擎为例,下图给出其支持的三种不同类型的智能问答功能。不同于以文档作为检索和结果展示基本单位的传统搜索引擎,必应搜索引擎中的智能问答模块能够基于知识图谱问答(图 1-2)、表格问答(图 1-3)和文本问答(图 1-4)等问答知识库,直接生成问题对应的精准答案。

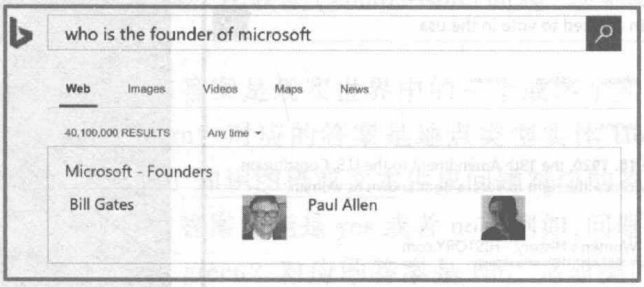


图 1-2 知识图谱问答

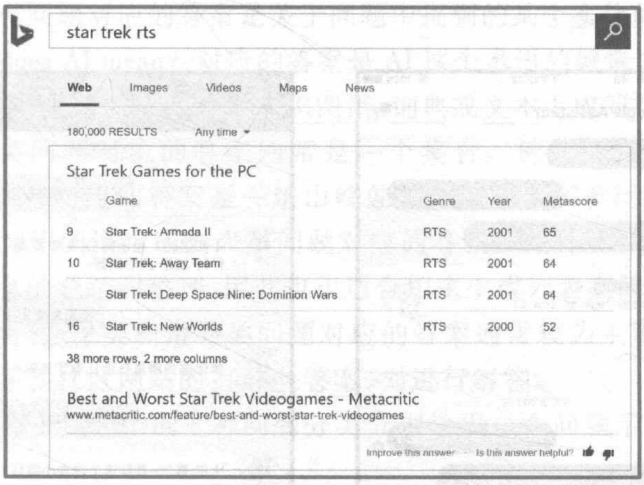


图 1-3 表格问答

除了搜索引擎外,智能问答技术在智能对话系统中起着同样重要的作用,这是由于人机对话中相当比例的场景都是问答场景。Apple、Google 和 Microsoft 等公司发布的智能对话产品中都集成了智能问答模块作为其重要组成部分,下面给出三个具体的例子,分别是 Apple Siri(图 1-5)、Google Assistant(图 1-6)和 Microsoft Xiaoice(图 1-7)。

图 1-4 文本问答

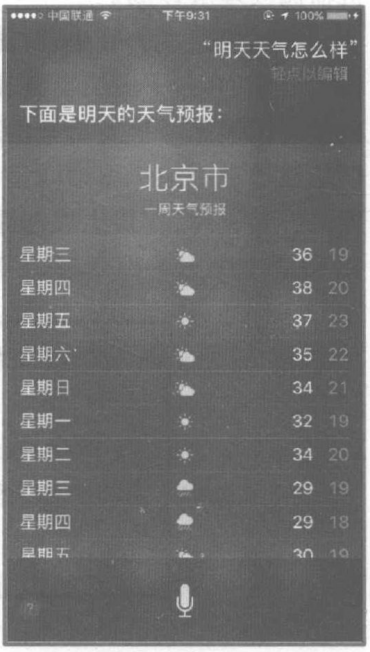


图 1-5 Apple Siri

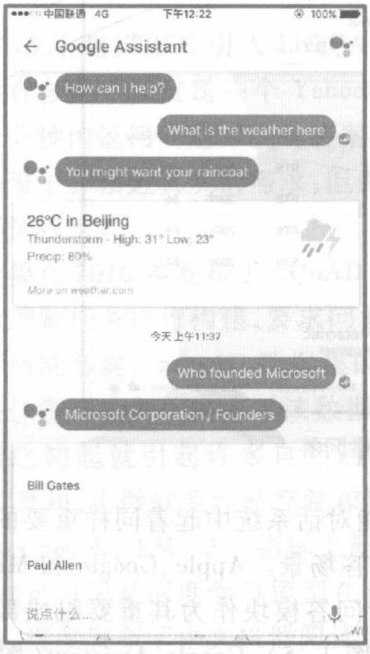


图 1-6 Google Assistant

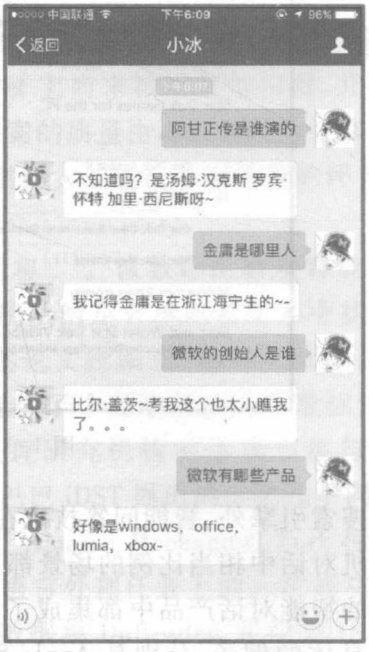


图 1-7 Microsoft Xiaoice