问答系统研究综述

李娉

**摘要：** 文中首先会介绍问答系统的定义和历史，并介绍现有的主流问答系统，以及分析其核心技术和评测机制， 最后将对问答系统的发展趋势和方向进行展望。

**关键词：** 问答系统；问答式检索系统；信息检索;信息抽取

**1 研究意义**

问答系统（也可称为人机对话系统），是指的一个机器系统，该系统的功能是对于用户通过自然语言输入的问句，它能够给出既简洁又准确的回答。由于用户的需求越来越精细化、多样化，基于关键词组合或者基于浅层语义分析的检索系统越不能满足用户的需求，智能问答系统可以通过数据经过深度加工处理形成具有某种固定结构的知识库，并通过自然语言处理技术解析用户需求，从而快速准确地为用户提供所需要的信息[8]。

问答系统及其相关领域有重要的研究价值和实际应用价值。其涉及的领域很广，其中主要技术有知识的抽取和表示，用户问句的语义理解和通过知识推理得到答案，该领域如果有重要突破，对于文本分类和推荐系统等领域都会有相当大的促进作用。对于实际应用价值，问答系统能快速准确的找到信息可以让人们的生活变得十分便利，比如：查询天气，股票价格，飞机航班等情况，甚至可以部分代替人工劳动，用机器来替代人工客服进行自动问答，可以大大减少人力成本。

**2 发展历程**

问答系统已有70年的发展历史。早期的上世纪去六十年代问答系统大多针对特定的领域，如Baseball[1]和Lunar[2],一般只接受特定形式的自然语言问句，训练数据少，未能广泛使用。进入九十年代后，互联网发展和TREC-QA评测的推出，极大推动促进了智能问答系统的发展，研究人员在该语料库上训练和测试各种问答模型，先后提出基于逻辑推理[3]、模板匹配[4]、集起学习[5]、数据冗余性的方法等很多领先方法[6]。在该阶段，人们主要利用信息检索或浅层语义理解技术去从大量候选集中寻找答案从而构建智能问答系统，取得了巨大发展，但其缺陷就是答案中至少要包含用户问句中的字词，用户问句是自然语言，其有复杂性，所以检索式问答不能很好的处理和解决用户的需求。

缺乏高质量的数据和强大的技术一直是阻碍发展的两个主要因素，但是随着一些常见的互联网应用兴起，比如百度百科、维基百科，使得缺乏高质量数据的问题被得到，像Freebase等知识库被建立；此外，自然语言处理技术的发展得益于机器学习和深度学习的兴起。

近年来，有很多智能问答系统诞生了，比如IBM研发的问答机器人Waston在《Jeopardy》节目中战胜人类选手，该机器人是融合了统计机器学习、信息抽取、知识库集成和知识推理等各种技术为一体的人工智能系统。又比如：微软的“小冰”、Apple的Siri和中科汇联公司的“爱客服”，其不仅提供情感聊天还提供智能客服的功能。问答系统正在快速发展中，未来将会有满足和解决更多用户需求的问答系统诞生。

**3 现有的主流问答系统**

主流的问答系统，大致可以分为：聊天机器人、基于知识库的问答系统、基于自由文本的问答系统等。

3.1聊天机器人

聊天机器人主要功能给出较为人性化且符合人类语言习惯的答案。比如：ALICE。ALICE是由Richard S.Wallac开发的，在2000年以及接下来的两年，都获得了“Loebner Prize”比赛冠军。其使用AIML表示其知识，采用了模式匹配的方法，并定义了丰富的标签，用Java作为引擎分析用户输入，并在知识库中搜索最合适的回答来返回给用户。

3.2基于知识库的问答系统

该系统首先采用自定义词典分词和CRF模型相结合的方法识别问句中的主体；其次，采用模糊匹配方法将问句中的主体与知识库中实体建立链接；然后，通过相似度计算以及规则匹配等多种方法识别问句中的谓词并与知识库实体的属性建立关联；最后，进行实体消歧和答案获取。

3.3 基于自由文本的问答系统

自由文本，又称原始文本、非结构化文本，是指 未经人工处理的文档、网页等。 基于自由文本的问答系统，是接受用户以自然语言提交问题； 利用信息检索等技术，从系统的自由文本库中检索出 相关的文档、网页； 利用答案抽取等技术，从这些检索出来的自由文本中 抽取出问题的答案并提交给用户。 与其它问答系统相比，基于自由文本的问答系统： 不需要建立大规模知识库，而是基于自由文本进行知 识问答，节省了大量的人力物力； 系统返还给用户的，是用户问题的具体答案而不只是和用户查询相关的文本或者网页。

**4 研究前沿**

智能问答系统涉及的 领域很广，其中主要关键技术有知识的抽取和表示，用户问句的语义理解和通过知识推理得到答案，这些领域都各自相对独立的存在并且采用了非常不同的方法，并且这些方法都有着各自的瓶颈。所以智能问答系统的现代研究基本围绕这三方面展开。[7]

4.1问句理解

给定用户问题，自动问答首先需要理解用户所提问题。用户问句的语义理解包含词法分析、句法分析、语义分析等多项关键技术，需要从文本的多个维度理解其中包含的语义内容。在词语层面，需要在开放域环境下，研究命名实体识别（Named Entity Recognition）、术语识别（Term Extraction）、词汇化答案类型词识别（Lexical Answer TypeRecognition）、实体消歧（Entity Disambiguation）、关键词权重计算（Keyword Weight Estimation）、答案集中词识别（Focused Word Detection）等关键问题。在句法层面，需要解析句子中词与词之间、短语与短语之间的句法关系，分析句子句法结构。在语义层面，需要根据词语层面、句法层面的分析结果，将自然语言问句解析成可计算、结构化的逻辑表达形式（如一阶谓词逻辑表达式）。

4.2文本信息抽取

给定问句语义分析结果，自动问答系统需要在已有语料库、知识库或问答库中匹配相关的信息，并抽取出相应的答案。传统答案抽取构建在浅层语义分析基础之上，采用关键词匹配策略，往往只能处理限定类型的答案，系统的准确率和效率都难以满足实际应用需求。为保证信息匹配以及答案抽取的准确度，需要分析语义单元之间的语义关系，抽取文本中的结构化知识。早期基于规则模板的知识抽取方法难以突破领域和问题类型的限制，远远不能满足开放领域自动问答的知识需求。为了适应互联网实际应用的需求，越来越多的研究者和开发者开始关注开放域知识抽取技术，其特点在于：1）文本领域开放：处理的文本是不限定

领域的网络文本；2）内容单元类型开放：不限定所抽取的内容单元类型，而是自动地从网络中挖掘内容单元的类型，例如实体类型、事件类型和关系类型等。

4.3知识推理

自动问答中，由于语料库、知识库和问答库本身的覆盖度有限，并不是所有问题都能直接找到答案。这就需要在已有的知识体系中，通过知识推理的手段获取这些隐含的答案。例如，知识库中可能包括了一个人的“出生地”信息，但是没包括这个人的“国籍”信息，因此无法直接回答诸如“某某人是哪国人?”这样的问题。但是一般情况下，一个人的“出生地”所属的国家就是他（她）的“国籍”。在自动问答中，就需要通过推理的方式学习到这样的模式。传统推理方法采用基于符号的知识表示形式，通过人工构建的推理规则得到答案。

但是面对大规模、开放域的问答场景，如何自动进行规则学习，如何解决规则冲突仍然是亟待解决的难点问题。目前，基于分布式表示的知识表示学习方法能够将实体、概念以及它们之间的语义关系表示为低维空间中的对象（向量、矩阵等），并通过低维空间中的数值计算完成知识推理任务。虽然这类推理的效果离实用还有距离，但是我们认为这是值得探寻的方法，特别是如何将已有的基于符号表示的逻辑推理与基于分布式表示的数值推理相结合，研究融合符号逻辑和表示学习的知识推理技术，是知识推理任务中的关键科学问题。

**5 总结与展望**

一方面，基于维基百科等高质量的知识资源库的建立和更新，另一方面，基于机器学习的自然语言处理技术的快速发展，将给问答系统的快速高效发展提供了资源基础和技术基础。但是，仍然存在问题，比如网络中有大量涉及多个领域的知识资源库，需要将这些多个领域的知识资源统一起来，来更好的满足用户的统一查询的需求；此外这些资源库中大多是事实性知识、缺乏常识性知识，这个问题不利于问答中知识推理过程，因为在人类的推理过程中，常识知识起着重要作用，所以如何将常识性的知识融入到问答系统中也值得深入研究；最后，如何更好的将机器学习结合深度模型来应用于问答系统，包括语义解析、信息检索与推理性问题等，这将是未来的研究趋势。

**参考文献**

[1] Green Jr, B. F., Wolf, A. K., Chomsky, C., and Laughery, K. Baseball: an automatic question-answer. In Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference (1961), ACM, pp. 219–224.

[2] Woods, W. A. Progress in natural language understanding: an application to lunar geology. In Proceedings of the June 4-8, 1973, national computer conference and exposition (1973), ACM, pp. 441–450.

[3] Moldovan， D. & Rus， V. Logic form transformation of WordNet and its applicability to question answering， in Proceedings of the 39th Annual Meeting on Association for Computational Linguistics， 2001.

[4] M. M. Soubbotin， S. M. Soubbotin. Patterns of Potential Answer Expressions as Clues to the Right Answers. Tenth Text REtrieval Conference (TREC-10). Gaithersburg， MD. November 13-16，2001.

[5] H. Yang， T.-S. Chua. The Integration of Lexical Knowledge and External Resources for Question Answering. Eleventh Text REtrieval Conference (TREC-2002). Gaithersburg， MD. November 2002.

[6] Kwok， Etzioni， Weld: Scaling Question Answering to the Web. Proc. WWW10， Hong Kong.

[7] 吴友政, 赵军, 段湘煜等.问答式检索技术及评测研究综述[J]. 中文信息学报, 2005, 19(3): 2-14.

[8] 詹晨迪. 基于知识库的自然语言问答方法研究[D]. 2017.

**Overview the Research of Question-Answering System**

liping

**Abstract:** The paper first introduces the definition and history of the question and answer system, introduces the existing mainstream question and answer system, and analyzes its core technology and evaluation mechanism. Finally, it will look forward to the development trend and direction of the question and answer system.

**Key words:** Question -answering system; question-answering retrieval system; information retrieval; information extraction