基于自然语言处理的问答系统综述

蒲 伟,王 恒
(宁夏大学 信息工程学院,宁夏 银川 750000)

摘 要:目前,问答系统已经广泛应用到社会的各个方面,其作为整个计算机领域的重点研究方向,引起了人们的高度重视。文章对问答系统的基础概念、发展历史、关键技术及未来的发展趋势进行简要阐述,旨在为问答系统今后发展打下良好基础。文章主要分析自然语言相关的背景、运用方法以及应用的领域。

关键词:问答系统;信息检索;答案抽取;问题分析;深度学习
中图分类号:TP391.1 文献标志码:A 文章编号:2095-2945(2021)22-0077-03

Abstract: At present, the question answering system has been widely used in all aspects of society. As a key research direction in the entire computer field, it has attracted people´s attention. This article briefly describes the basic concepts, development history, key technologies and future development trends of the question answering system, aiming to lay a good foundation for the future development of the question answering system. The article mainly analyzes the background, application methods and application fields of natural language.

Keywords: question answering system; information retrieval; answer extraction; question analysis; deep learning

语言是人类生活中不可或缺的一种沟通方式,自然语言是一种表达直接且简单的工具,自然语言处理(Natural Language Processing,NLP)是一种机器语言,可以将人类的交流转换为机器语言,以便于让计算机理解人类的想法。伴随着网络的发展,自然语言处理在人工智能方面迅速发展,被越来越多的人所熟知和运用。伴随着网络的飞速增长,促使网络信息量不断增加,人们获得信息就要更加精确。利用传统的搜索引擎技术就很难实现这些高要求,而智能问答技术成为解决这个问题的有效手段。早在 20 世纪 60 年代人工智能研究刚开始的时候,人们就提出了要让计算机像人一样用自然语言来回答人们的问题,实现“人机对话”,这就是问答系统[1]。智能问答就是指将用户的需求输入到计算机中利用计算机自动生成答案并输出,问答系统不像传统的搜索引擎那样将问题分解成关键字。问答系统在收到用户的问题后,将问答系统和自然语言处理技术结合起来,对问题进行解析处理,利用算法和模型,将用户需要的答案直接输出,不像搜索引擎输出的是相关的网页。所以智能问答系统和传统搜索引擎相比可以更有效地为用户解决问题。在问答系统中,我们可以根据答案的来源分类,可以分为基于知识库的问答系统,基于文档的问答系统和答案选择,按照应用的领域不同,我们又可以将问答系统分为基于限定领域的问答系统和开放领域的问答系统。限定域问答系统只能解决限定在某些范围或者某些范围的问题,常见的酒店预订、网上订餐等问答系统都是属于限定域问答系统。开放域问答系统指的是回复的问题不限定在某些特定范围。

1 自然语言处理的发展

自然语言处理(Natural Language Processing)是人工智能(AI)的一个子领域。自然语言处理是研究人与人以及人机交互的语言问题的一门学科。其发展分为三个阶段:20 世纪 50 年代开始是萌芽期;20 世纪 60 年代是发展期;20 世纪 90 年代是繁荣期。

早期计算机刚刚问世的时候,英国工程师布斯和美国工程师威弗最先提出了利用计算机进行翻译,但是起初机器翻译系统的粗糙导致翻译出来的质量非常低,人们慢慢就对机器翻译失去了好感,有的人甚至认为机器翻译是永远不可能实现的,意味着第一次机器翻译实验就失败了。在 20 世纪 50 年代是计算机科学发展的基础时段,当时提出来的理论都是基于图灵机的模型。随着发展在基于图灵机模型的基础上提出正则表达式以及有限自动机。在 1956 年,Chomsky 提出了一种关于上下文无关语法的模式,同年在人工智能诞生之后,自然语言处理迅速融入该领域之中,在快速发展期,上下文无关语法的提出使得该领域的研究分为了基于规则的符号派和基于概率的随机派,促使了未来的很多年人们都在研究这两种方法到底哪种方法更有效。虽然机器翻译面临着各种困难,但是在法国、日本等国家仍然在坚持研究机器翻译。直

作者简介:蒲伟(1995-),男,硕士研究生在读,研究方向:自然语言处理,问答系统。

到 20 世纪 70 年代的时候，机器翻译的研究者逐渐找到了研究的思路，在机器翻译的过程中要使原句的语义和机器翻译出来的语义一致，好的机器翻译系统就是能够将原句的语义准确无误地翻译出来，从此机器翻译就出现了复苏发展的趋势。至此，机器翻译中的语义分析就受到了越来越多研究者的重视，繁荣期最突出的是机器翻译的研究走向实用化，市场上出现了非常多的机器翻译系统，逐步进入了商业化模式并且运用在多种行业。

2 问答系统研究方向

2.1 视觉问答

视觉问答 [1] 将图片中提及的问题用自然语言输出，想要准确地回答问题，首先需要知道照片所表示的内容以及问题的含义，其次还需要了解图片和文字之间存在的对应关系。

在视觉问答系统中常见的通过以下两种方法实现。Kushal[2]基于贝叶斯方法实现了视觉问答系统，该模型通过对问题和图片特征建模共现统计概率，使用贝叶斯模型对问题、图片和答案进行推断，然后计算每个答案的边缘概率，将概率最高的作为问题的答案。

在基于深度学习的视觉问答系统方法中，有学者[3-4]将注意力机制引入视觉问答系统的研究中。通过注意力机制关注到图片部分的重要区域，在图片上产生较大的权重，从而给出更准确的答案。

2.2 基于知识图谱的问答

基于知识图谱的问答系统已经成为一种访问大型知识图谱的流行方式。通过访问知识图谱的结构化数据，其可以使用自然语言来准确地回答事实性问题。知识图谱是一种大规模的语义网络系统，可以将一些不同类型的信息链接在一起，形成知识图谱的关系网络结构，可以帮助人们直接找到各个物质之间的关系。目前基于知识图谱的问答系统已经运用在多种领域，张楚婷[5]研究并实现了基于知识图谱的旅游问答系统，在旅游高峰期的时候可以帮助游客解决一些问题，不再通过人工咨询的方式获取信息。帮助游客在游玩的时候减少一些不必要的时间损耗。基于知识图谱的问答系统在教育、医疗、汽车、农业、金融、电影等领域都得到了充分的研究和应用，由于知识图谱的网络结构，充分体现了良好的推理能力，在公安情报分析以及推理、医疗系统问诊以及开药等系统中都得到了较好的效果。

在基于知识图谱的问答系统中，在旅游领域，张楚婷[5]运用了基于 BiLSTM-CRF 的细粒度问答模型用于候选主实体以及实体的选择，并且在关系抽取中用了注意力机制和 CNN 抽取之间的关系。在实体识别和关系抽取中的准确率和识别率得到提高。韩薇[6]在张楚婷[5]的基础上进行改进用 BiLSTM-CNN-CRF 模型进行实体识别，进一步提高了实体识别的准确率和效率。在属性链接上，在 CNN 和注意力机制用作关系识别的基础上，加入了 Droppout 方法，目的是防止模型训练过程中拟合数据集。在教育方面，李轩[7]将企业和高校学生之间联系在一起，企业在招聘的时候需要不同的人才，基于 BiLSTM+CNN-CRF 的实体识别模型，对职位信息数据、技术领域以及个人能力等实体进行抽取，通过序列标注问题，利用 Keras Embedding 模型进行词嵌入矩阵，在实体识别后进行实体间的关系抽取，并且搭建一个基于知识图谱的教育问答系统，可以通过问答系统进行人机交互，学生能在问答系统中提问，了解不同岗位以及不同的领域需求，可以提前规划自己的方向，在明确求职目标，岗位的工作范畴来提高学习效率。在法律领域，黄薇屹[8]提出基于法律领域的知识图谱问答系统，并且引入少量样本和迁移学习模型运用在基于知识图谱的法律问答系统中，在迁移学习模型和少量数据上进行实验发现更快的迭代，实验效果明显提升。

目前构建知识图谱的问答系统主要基于以下几种方法。基于规则的方法，Mekhaldi[9]使用该方法将问题映射成谓词然后进行结构化查询，这种方法的优点是准确率较高但是规则是由人设计的，所以泛化能力较差。构建模板的问答方法，该方法主要是使用已知模板成分匹配句子中的内容。Cui[10]提出了一种基于模板的问题表示方法，针对简单事实问答，在大规模模板自动化生成方面，提出优化方案。语义解析的问答方法，基于词典-文法的语义解析方法，基于神经网络的方法，基于知识图谱嵌入学习的问答方法及多跳推理的知识图谱问答等。

3 相关理论技术

基于不同类型的问答系统在数据处理以及技术实现方面存在着差异，但是主要技术有问题预处理、信息检索以及答案生成。

3.1 问题预处理

在问答系统工作时，进行下一步之前的首要任务就是对问题进行预处理。问题预处理包括进行分词、去除停用词、词性标注以及语法分析等任务。问答系统的类型不同，在问题类别的判别上就是一项重要的工作，答案的类别就是依据问题的类别判断出来的，所以这个过程对问答系统来说必不可缺。问答系统遇到的问题都是由若干个词或者词组组成，所以要获得关键词信息就要对问题进行分词，提取问题中的关键词；关键词的提取是问题分析的核心步骤，问答系统中信息检索就是通过关键字进行检索，所以关键字抽取与信息检索的准确度高低有密切的联系。

3.2 信息检索

用户在问答系统中通过提问的方式得到最准确的答案，主要是通过信息检索从文档或者知识库中检索出答案。问答系统中信息检索是必不可少的一步。首先需要了解用户问句所表示的含义与意图，利用信息检索技术在知识库中抽取相似的信息作为回复的答案。在不同类型的问答系统中检索方式也各不相同；如基于文本类的问答系统，主要是对文档或者相关文字缩小答案范围，最后抽取出最精准的答案；基于问答对的问答系统主要根据在问答语料库中匹配相似度较高的句子进行检索；基于知识图谱的问答系统，主要通过实体链接将问句中的实体和知识图谱相映射查找相对的实体信息。

3.3 答案生成

问答系统的最后一个步骤就是答案生成，在基于文本和问题答案对的问答系统中都是通过信息检索得到数据，依据文档中的信息，抽取出与问句相似的句子作为答案返回给用户 在基于结构化数据类型的问答系统中，主要是通过在答案库中抽取出来与之对应的实体给用户。答案生成模块主要是用候选答案抽取，在文档或者段落中抽取出可能的答案作为一个答案集，然后在答案集中通过实体的类别与问题中关键字或者实体进行比较，计算各个实体之间的权重，抽取出相似度最高的生成答案，然后将该答案返回给用户。

4 结束语

随着数据的增加，查找数据难度越来越大，用户越来越需要通过问答系统获取想要的答案，现阶段问答系统刚刚处于起步阶段，只能处理一些简单的问答和推理。大量学者也在不断地研究问答系统，由于数据变多，基于知识图谱的问答系统研究的相关技术不断突破，应用在不同的场景。

参考文献：

[1]Antol S，Agrawal A，Lu J S，et al. VQA：Visual Question Answering [C]//2015 IEEE International Conference on Computer Vision (ICCV)，December 7-13，2015. Santiago，Chile. Piscataway，NJ：IEEE，2015.

[2]Kushal Kafle，Christopher Kanan. Answertype prediction for visual question answering [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE，2016：123-132.

[3]Jin Hwa Kim，Kyoung Woon On，Jeonghee Kim，et al. Hadamard product for low-rank bilinear pooling [C]//In International Conference on Learning Representations，2017：236-245.

[4]Akira Fukui，Dong Huk Park，Daylen Yang，et al. Ual Question Answering and Visual Grounding [C]//In Conference on Empirical Methods in Natural Language Processing (EMNLP)，2016：235-242.

[5]张楚婷.基于知识图谱的旅游问答系统研究与实现[D].桂林：桂林电子科技大学，2019.

[6]韩馥.基于知识图谱的山西旅游饮食问答系统[D].太原：中北大学，2020.

[7]李轩.基于知识图谱的教育领域知识问答系统的研究与应用[D].长春：吉林大学，2019.

[8]黄薇屹.基于知识图谱的深度法律内容问答模型[D].深圳：中国科学院大学(中国科学院深圳先进技术研究院)，2020.

[9]S. Ou，C. Orasan，D. Mekhaldi，et al. Automatic Question Pattern Generation for Ontology-based Question Answering [C]//In FLAIRS，2008：183-188.

[10]Cui W Y，Xiao Y H，Wang H X，et al. KBQA：Learning Question Answering over QA Corpora and Knowledge Bases [J]. Proceedings of the VLDB Endowment，2017，10(5)：565-576.

(上接 76 页)

为实现各地市输电线路巡查需求而对共同使用的作业资源实行时间分配，是一个典型的组合优化问题。通过利用精确算法求解 NP 难问题，即使能得到最优解，也存在所需计算时间过长，难以直接应用。因此，可以采用近似算法，通过输入上述方案模型所需参数进行求解计算。基于智能优化的近似算法是基于一定的优化搜索机制，并具有全局优化性能的一类算法[4]。常见的算法有 模拟退火算法(SA)、遗传算法(GA)、蚁群算法(ACO)、路径重连算法(PR)、迭代局部收缩算法(ILS)、禁忌搜索算法(TS)、分散搜索算法(SS)、粒子群算法(PSO)等等。作为一种通用算法框架，智能化算法主要是针对具体问题对框架结构进行局部修改，具有较好的实践通用性，且能在较快地处理大规模数据的同时得到可接受的解，在工业实际问题组合优化求解方面具有一定优势。

参考文献：

[1]陈兰波.电力线路无人机巡检方案研究[J].科技与创新，2020(11)：36-38+41.

[2]输电线路运维策略及管控机制实施细则[S].广东：广东电网有限责任公司，2019.

[3]Hamidi Mustapha. Solution of P versus NP problem [J]. Algorithms Research，2015，4(1)：7.

[4]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+1210.