计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(8):1−18 [doi: 10.15888/j.cnki.csa.009208]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

问答系统研究综述①

闫　悦[1], 郭晓然[2], 王铁君[2], 饶　强[1], 王铠杰[1]

[1](西北民族大学 中国民族信息技术研究院, 兰州 730030)
[2](西北民族大学 数学与计算机科学学院, 兰州 730124)
通信作者: 王铁君, E-mail: wtj@mail.lzjtu.cn

摘　要: 问答系统是人工智能和自然语言处理领域中具有广泛发展前景的研究方向之一. 早期的问答系统限定以自然语言形式进行提问和回答, 近年来, 随着多模态知识图谱、多模态预训练模型的发展, 支持文字、图片、音频、视频等多种模态间信息查询的广义问答系统逐渐成为新的研究热点, 其以多媒体方式展示结果, 更加直观、全面. 本文根据问答系统任务对象的变化, 将问答系统划分为 3 种类型: 专用问答系统、通用问答系统和多模态问答系统. 分析了这 3 种类型的问答系统发展过程中所面临的问题, 着重总结每个阶段所采用的关键技术与方法, 同时对问答系统在工业上的应用进行了举例说明, 并对未来研究方向进行了展望.

关键词: 问答系统; 问题分析; 信息检索; 答案生成; 智能问答

引用格式: 闫悦,郭晓然,王铁君,饶强,王铠杰.问答系统研究综述.计算机系统应用,2023,32(8):1–18. http://www.c-s-a.org.cn/1003-3254/9208.html

Survey on Question Answering System Research

YAN Yue[1], GUO Xiao-Ran[2], WANG Tie-Jun[2], RAO Qiang[1], WANG Kai-Jie[1]

[1](China National Information Technology Research Institute, Northwest Minzu University, Lanzhou 730030, China)
[2](School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou 730124, China)

Abstract: The question answering (Q&A) system is one of the promising research directions in the field of artificial intelligence and natural language processing. Early Q&A systems can only ask and answer in the form of natural language. In recent years, with the development of multimodal knowledge graphs and multimodal pre-training models, generalized Q&A systems supporting information queries of multiple modes such as text, image, audio, and video have gradually become a new research hotspot, and their display of results in a multimedia manner is more intuitive and comprehensive. This study classifies Q&A systems into three types according to their changing task objects: dedicated Q&A systems, general Q&A systems, and multimodal Q&A systems. The problems faced in the development of these three types of Q&A systems are analyzed, and the key technologies and methods used in each stage are highlighted and summarized. In addition, the industrial applications of Q&A systems are exemplified, and future research directions are prospected.

Key words: question answering (Q&A) system; question analysis; information retrieval; answer extraction; intelligent question answering

1　引言

问答系统是信息检索领域长期以来的研究重点, 通常使用自然语言形式的句子进行提问, 结合上下文语境为用户返回精准、可靠的答案, 改善用户的搜索

① 基金项目: 国家自然科学基金 (62166035); 甘肃省自然科学基金 (21JR7RA163); 中央高校基本科研业务费 (31920210090)
收稿时间: 2023-01-10; 修改时间: 2023-03-08; 采用时间: 2023-03-23; csa 在线出版时间: 2023-06-09
CNKI 网络首发时间: 2023-06-12

体验. 广义的问答系统不再仅局限于自然语言形式的问题和答案, 也支持用户输入图片、视频、音频等多媒体信息进行查询, 以多媒体形式展示答案, 例如: 微软小冰[1] 所提供的图片评论功能是对用户发出的图片进行回答; 阿里小蜜机器人[2] 在电商直播时担任虚拟主播和智能辅播, 可提供文字、图片、视频等多模态的商品展示和回复. 随着人工智能和深度学习的发展, 面对复杂问题时, 期望能够模仿人类运用不同模态的信息, 形成一个完整的思维链, 因此对多模态知识图谱、多模态预训练模型[3] 和多模态问答系统等诸多方面提出了新挑战.

1.1 问答系统的定义

早期的问答系统主要利用信息检索技术在对应知识库中找到合适的信息[4], 属于狭义的问答系统, 被定义为一个利用计算机等自动化机器回答自然语言问题的系统[5], 即问题和答案只有自然语言一种模态. 广义问答系统被重新定义[6,7], 允许用户输入任意模态信息, 通过提取用户输入内容中的关键信息, 给出问题所对应的准确答案. 例如, 用户可以提问“燕子的特征是什么?”, 也可以输入一张燕子图片, 提问“这个动物的特征是什么?”, 系统都应该返回一个精准的答案.

1.2 问答系统的处理框架

狭义的问答系统处理框架主要包括问题分析、信息检索、答案抽取 3 个部分. 添加其他模态信息后, 问答系统的框架发生变化, 问题分析阶段需对文本、图像、音视频信息进行特征提取, 之后将多种模态信息进行融合, 目的是将多个特征映射到同一个空间中, 最后根据融合的信息进行答案生成, 框架如图 1 所示.

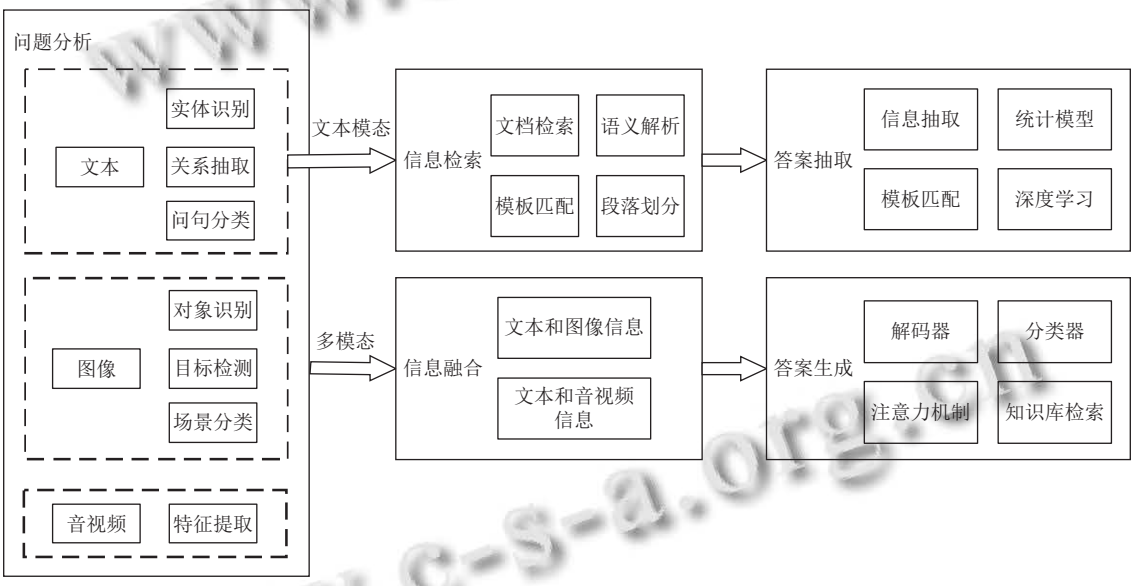

图 1 问答系统的处理框架

1.3 问答系统的发展脉络

问答系统概念的提出可追溯到 20 世纪 50 年代——图灵测试的提出, 发展到如今的多模态问答系统, 前后经历了 8 个阶段, 每个阶段所处理的数据格式和形式不同, 具体应用也有所不同, 如表 1 所示.

目前有许多研究人员针对问答系统的不同方面进行了综述, 陈子睿等人[8] 针对开放领域知识图谱问答方法进行了总结; 冯钧等人[9] 对问答系统中的复杂问题进行了总结; 姚元杰等人[10] 对问答系统中运用的深度学习方法进行了总结. 经过技术的不断发展, 本文根据任务对象将问答系统 8 个阶段划分为专用问答系统、通用问答系统和多模态问答系统 3 种类型. 接下来, 围绕不同类型问答系统所涉及的关键理论技术进行介绍.

2 专用问答系统

文中将任务对象为限定领域、结构化文本数据的问答系统统称为专用问答系统. 这一类问答系统是对问答系统这一概念的初提出与初实践, 由于当时的网络不是很发达, 受数据集内容与数量的限制, 留下的有代表性的问答系统多限定在某一领域.

表 1 问答系统发展脉络

类型	发展时期	任务对象	相关应用
专用问答系统	图灵测试 (20世纪50年代)	—	测定机器是否具有人工智能的方法
	AI时期 (20世纪60、70年代)	限定领域、结构化数据	Baseball系统 Lunar系统
	计算语言学阶段 (20世纪80年代)	限定领域、结构化数据	Unix Consultant (UC)
通用问答系统	基于大规模文档集的问答系统 (20世纪90年代)	开放领域、电子文档	聊天机器人ALICE 问答系统评测专项
	基于问题答案对的问答系统 (21世纪00年代)	开放领域、问题答案对	百度的AnyQ 腾讯知文—结构化FAQ问答引擎
	基于知识图谱的问答系统 (21世纪10年代)	开放领域、结构化数据	Google Knowledge Graph Amazon知识图谱
	基于大规模语言模型的问答系统 (21世纪20年代)	开放领域、非结构化数据	ChatGPT
多模态问答系统	多模态问答系统 (21世纪20年代)	文本、图像、视频、音频	阿里小蜜机器人 微软小冰 灵医小智

2.1 图灵测试

图灵测试最早来源于 1950 年图灵发表的一篇名为“Computing Machinery and Intelligence”[11] 的论文中. 图灵专注于研究计算机可否像人一样进行交谈, 提出了测定机器是否具有人工智能的一套方法——图灵测试, 它是人工智能最初的概念, 甚至早于“人工智能”这个词本身. 图灵测试采用“问”与“答”的模式, 即观察者通过控制打字机和测试对象通话, 观察者不断提出各种问题, 从而辨别回答者是人还是机器.围绕图灵测试展开的研究历程如图 2 所示. 图灵测试所使用的“问”与“答”的测试方法, 是问答系统最早的研究理论.

2.2 AI 时期

问答系统的雏形最早出现在 20 世纪 60 年代, 主要研究内容是如何使用自然语言检索结构化数据库. 这一时期有两个比较著名的系统 Baseball[12] 和 Lunar[13]. Baseball 系统是最早以“未来的人机交互将以自然语言进行交流的方式”为目标构建的系统, 用于回答用户通过自然语言提出的关于棒球联赛问题. Lunar 是在 NASA 载人航天器中心的支持下开发, 使月球地质学家无需学习编程语言和数据库的相关指令, 直接使用自然语言就能访问检索 NASA 数据库. Lunar 系统示意图如图 3 所示, 主要由 3 个部分组成: 第 1 部分用于将自然英语按照语法生成机器可理解的句法树; 第 2 部分的语义解释组件将句法表示进行转换, 对句子进行意图理解. 数据库检索和推理组件用于根据句子含义对数据库进行计算和检索; 第 3 部分确定查询的答案并根据结果对数据库进行更改.

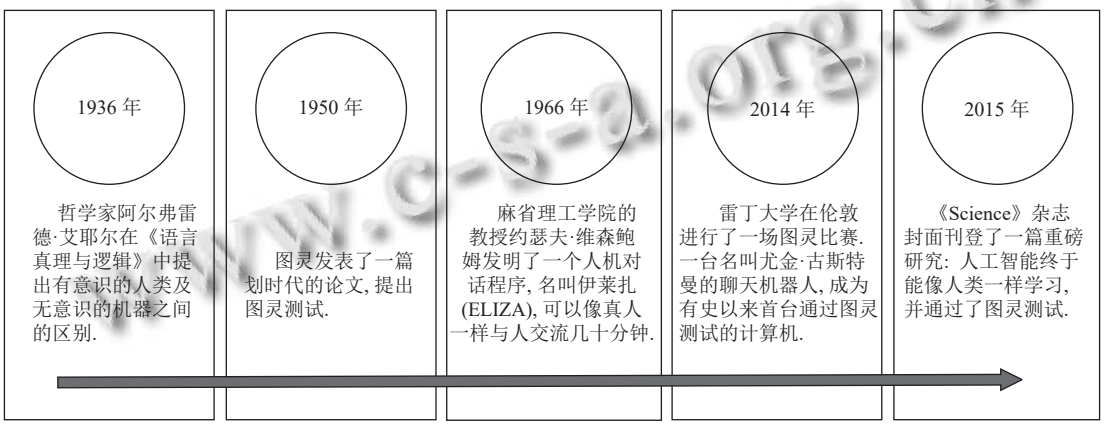

图 2 图灵测试研究历程

以上两个系统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. 这些早期系统大多针对特定领域, 处理的数据规模不大且为结构化数据. 这一时期还出现了一些可以进行对话的问答系统, 比如 ELIZA (心理咨询)[14]、SHRDLU (积木游戏)[15]、GUS (旅行信息咨询)[16]. 随着人们在探索问答系统中对句法和语义问题的理解不断加深, 问答系统的发展在 20 世纪 80 年代进入了计算语言学时期.

2.3 计算语言学时期

这一时期的问答系统更加关注于解决事实类型的

问题, 任务对象为限定领域和结构化文本数据, 代表系统为 UC 系统[17].

UC 提供一个智能、自然的语言界面, 让用户用英语与系统交流来了解 Unix 操作系统. 研究者们将其称为“智能帮助设施”, 具体指建立一个系统模拟人类顾问的实际功能. 它具备了分析用户的语言、确定用户操作的目标、给出解决用户需求的规划、决定需要与用户沟通的内容、以英语生成最终的对话内容以及根据用户对 Unix 系统的熟悉程度进行建模等功能. UC 的框架由大量组件组成, 组件之间大多以串行的方式调用, 功能如图 4 所示. 这一时期问答系统的处理流程已具备雏形, 首先分析用户的问题, 了解用户的意图, 然后获取有关用户问题主题的知识, 最后制定合理的回答并将答案返回给用户.

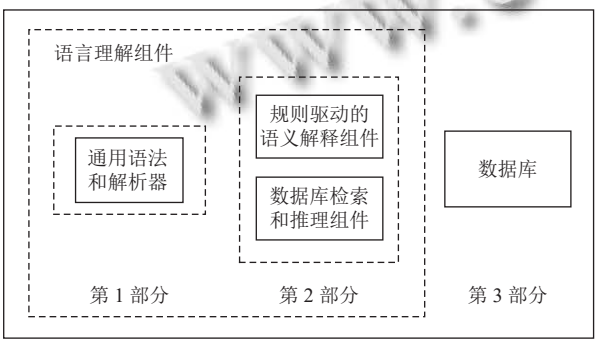

图 3 Lunar 系统组成部分

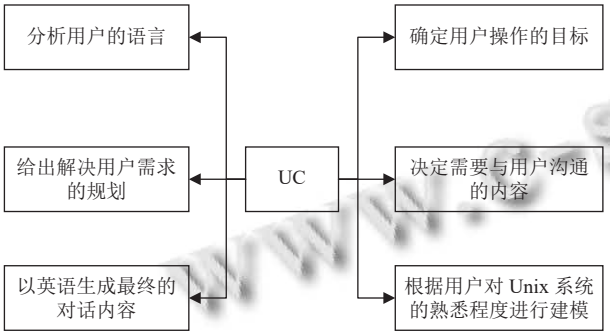

图 4 UC 的功能

3 通用问答系统

经历了上一阶段的专用问答系统后, 之后的研究对象大多为开放领域且在这一阶段的研究内容现在还在工业界或者学术界流通使用, 所以本文将基于大规模文档集的问答系统、基于问题答案对的问答系统、基于知识图谱的问答系统和基于大规模语言模型的问答系统归类为通用问答系统, 其特点为信息量大且不再限定信息领域.

3.1 基于大规模文档集的问答系统

20 世纪 90 年代, 问答系统的发展进入了基于大规模文档集的问答系统时期. 这一时期互联网快速发展, 产生了大量的电子文档, 数据的格式不再是固定的结构化数据, 问答系统的流程为从用户的自然语言问句中获取主题词, 利用主题词在网络文档中搜索相关的文档, 如图 5 所示.

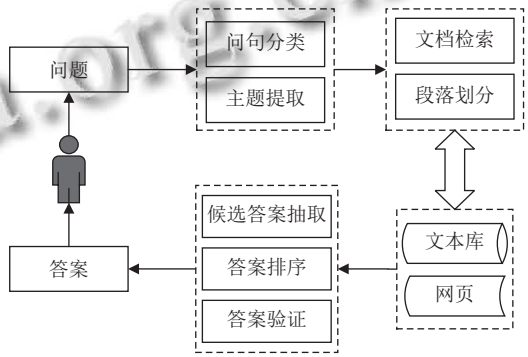

图 5 基于大规模文档集的问答系统处理流程

1995 年 Wallace 开发设计了第 1 个聊天机器人 ALICE[18]. ALICE 的知识库由 AIML (artificial intelligence markup language) 语言编写的文件组成, 每个 AIML 文件代表一个领域的可能话题. ALICE 的推理机制是将 AIML 知识库以树的结构形式加载到内存中形成内存知识树. 当用户输入要查询的句子时, ALICE 系统在这棵内存知识树中检索与用户输入语句最匹配的模式, 如果匹配到则对答复模板进行进一步处理后回复给用户.

这一时期问答系统研究的基本问题是问题分析、信息检索和答案抽取. 下面将针对这 3 个问题对基于大规模文档集的问答系统的相关研究进行综述.

3.1.1 问题分析

问题分析主要是对问句进行分析, 理解用户问题, 为信息检索做准备, 通常将问句分类和主题提取用于分析问题.

(1) 问句分类. 问句分类是在给定问句的情况下, 将用户问题分类到已规定的或自定义的分类结构中一个或几个类. 这一时期对问句分类的主要方法分为基于规则的分类方法和机器学习的方法, 最初 Lehnert 等人[19] 将问句分为了 13 个概念类, 但这样的分类方式在事实问题上不适用, 为解决这一问题, Li 等人[20] 利用

一种分类的机器学习方法, 将问题分为 6 个大类, 同时在这 6 个大类下又细分 50 个小类, 该方法较好满足了人们的需求. 2002 年, Magnini 等人[21] 利用规则的分类方法, 在 DIOGENE 系统中用一组手动定义的分类规则, 对问句进行分类, 可以正确对问句进行处理. 近年来问句分类主要是基于深度学习的方法, 利用 CNN 模型、LSTM 模型、BERT 模型进行, Xiao 等人[22] 利用 CNN 模型对中国法律问题进行分类, 该方法在粗粒度分类和细粒度分类上都取得了不错的效果. 杨建飞[23] 利用 BERT 模型对军事装备问题进行分类, 张永亮[24] 利用 TextCNN 模型对关于苹果病虫害的问题进行分类, 这些方法较好地完成了对应任务.

(2) 主题提取. 主题提取的结果将用于信息检索中获得与查询问题相关的文档, 主题词的选取会影响到信息检索的效果, 所以一般会选取问句的中心词和中心词的约束作为主题. 大多方法是基于 TF-IDF 模型及停用词过滤的方法, Mihalcea 等人[25] 参考 PageRank 算法[26] 提出了一种将文本作为图的无监督排序算法——TextRank 算法, 用于关键词抽取和对文档进行简明摘要. El-Beltagy 等人[27] 提出 KP-Miner 算法, 该算法首先利用标点符号和停用词将文档切分, 将切分出的词序列作为候选词, 然后通过候选词的频度以及设定的规则对候选词进行过滤, 计算总体权重, 最后根据权重值进行排名, 来提取关键词. 以上两种主题提取的算法应用较广.

3.1.2 信息检索

信息检索是根据所提取的主题剔除掉文档集中的无关文档, 减小信息搜索的范围, 以提高搜索的效率和精度. 信息检索的任务一般分为文档检索和段落划分, 检索出含有正确答案的文档或者段落越少, 提取出正确答案的可能性就越大.

(1) 文档检索. 文档检索是在所选数据库中查找出与用户问题相关的文档. 常用的检索模型有布尔模型、向量空间模型、概率模型、语言模型、机器学习排序算法等. 在问题分析的主题提取中提到过, 通常会选用问句的中心词和中心词的约束作为检索的内容, 但会遇到对长问句选取的关键词较多, 对短问句选取的关键词较少的问题, 这种情况会对检索产生困难. Moldovan 等人[28] 针对这个问题在查询过程中用迭代式调整技术对关键词进行调整, 若返回的文档过多, 说明关键词太少、查询限制宽松则需加上一些限制; 若返回文档较少, 说明若关键词太多、查询限制太严格就去掉一些限制.

(2) 段落划分. 段落划分是在文档检索的基础上再一次减少相关文档的数量. Tellex 等人[29] 在研究中发现布尔查询模式在问答任务中表现良好, 且基于密度的算法在段落划分中可以获得相对较好的效果. 但是这种算法只针对提取出的关键字, 不考虑上下文对答案的影响. Chu-Carroll 等人[30] 使用 XML 片段查询语言对文档进行搜索, 在关注关键词的同时, 也关注句子中词与词组的关系, 显著提高了划分范围的准确度. 针对查询语言与查询文档中句子不完全相关的问题, 李宇等人[31] 提出了一种文本片段化机制来进一步解决段落划分的问题, 综合考虑语义和词频的两个方面的影响, 计算查询语言与段落片段的相似度, 根据相似度进行排序, 实验结果表明该方法可以提高信息检索的性能.

3.1.3 答案抽取

答案抽取是对检索到的候选答案的段落集合进行提取, 从中获取正确的答案返回给用户, 这一时期常用的方法为模板匹配、信息检索及关系抽取.

模板匹配的方法通常是人工构建答案模板, 利用答案模板进行答案提取, 如文献 [32] 中的问答机制是搜索预定义的文本表达式, 这些表达式可以被解释为某些类型问题的答案. Yang 等人[33] 提出了一个经典的信息检索方法来进行答案提取, 该算法利用在信息检索中得到的候选答案的段落合集, 在该段落合集中进行命名实体识别, 将与问题类型与段落集合中的实体类型进行匹配, 选择与问题类型一致的作为候选答案, 然后将相似度最好的候选答案作为正确答案返回给用户. 上面两种方法第 1 种因为需要人工构建模板, 对可回答问题有类型限制, 只能回答有模板的问题, 第 2 种方法有性能的上限, 需要对多个段落进行实体识别. 关系抽取的方法被提出用来解决性能上限的问题, Lin 等人[34] 提出了一种无监督算法, 用于从文本中发现推理规则, 将其应用于解析语料库的依赖关系树中的路径. 在此基础上, Moldovan 等人[35] 利用一套可以通过对解析获得的语法树进行规则计算, 从而建立问题到答案的逻辑表示的工具, 这种方法与推理的方法类似, 通过找到各种关系来获得答案.

3.1.4 小结

1999 年, 文本检索会议 (text retrieval conference, TREC) 引入了问答系统评测专项 (question answering

track, QA track), 吸引了更多的研究者进行这方面的研究, 极大推动了问答系统的发展. 基于大规模文档集的问答系统的优点是, 可获取的知识来源更多, 不需要预先建立大规模的知识库. 这一方法虽然在一段时间内取得了较为不错的效果, 并使研究者对于问答系统中存在的问题更加聚焦, 但由于用户提问方式具有多样性以及自然语言的复杂性, 这一时期的研究仍不能较为准确的理解问题, 尤其是对于复杂问句. 并且由于数据的来源往往都是从网络文档中抽取的非结构化数据, 质量难以得到保证, 导致得到的效果差强人意, 问答的准确率及系统性能都比较低.

3.2 基于问题答案对的问答系统

进入 21 世纪后, 由于互联网技术的成熟与普及, 问答系统进入了基于问题答案对的问答系统时期. 这个时期提出了两种类型的问答系统, 一个是基于常问问题 (frequently asked questions, FAQ) 的问答系统, 另一个是基于社区问答 (community question answer, CQA) 的问答系统. 两种类型各有优劣: 与 CQA 相比, FAQ 有着质量高、组织好等优点, 使得系统回答问题的水平大大提高, 但是 FAQ 的数据获取成本高; CQA 为基于问答对的问答系统提供了可靠的问答数据来源, 但存在缺乏专业知识匹配、答案质量低等问题.

FAQ 问答系统以标准问题为桥梁, 将用户和答案连接起来, 通过预先整理好的一些常问问答对, 发布在网页上为用户提供服务. FAQ 问答系统的总体结构如图 6 所示, 整体框架符合狭义问答系统的定义.

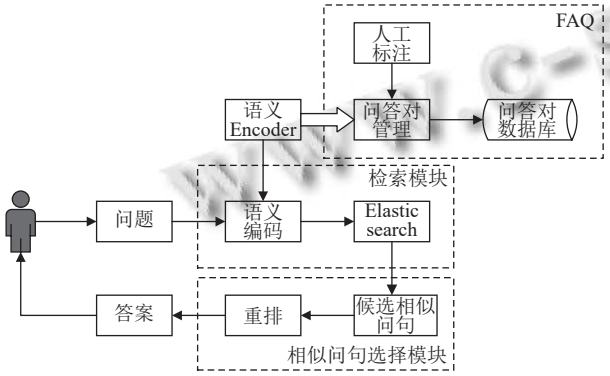

图 6 FAQ 问答系统的总体结构

CQA 提升了系统的自由度, 用户可以对他人的问题或者回答进行交互, 如点赞、评论等. 问答系统会检索社交媒体中的相似问题并将答案返回给用户. 有关 CQA 的研究主要包括专家推荐、答案质量评估、问句分类和相似问题检索等, CQA 问答系统的总体结构如图 7 所示.

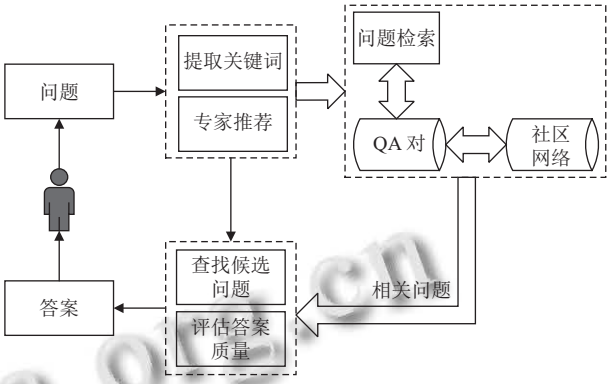

图 7 CQA 问答系统的总体结构

3.2.1 基本研究问题

FAQ 问答系统的研究内容与上一时期的基本一致, 主要处理客观、事实类型的问题. 但 CQA 问答系统中有大量的主观类问题, 主观类问题自由度更大, 答案不唯一, 所以 CQA 问答系统与 FAQ 问答系统的技术有所不同, 以下对几个不同之处进行综述.

(1) 专家推荐. 专家推荐是通过用户的行为特征对用户擅长的知识领域进行建模, 从而将问答系统中未解决的问题推荐给相关领域具有专业知识的用户去回答, 在用户权威评估上, Jurczyk 等人[36] 通过一条边将提问者和回答者之间连接起来, 建立用户网络, 用一种图结构算法来评估用户的权威性. Agichtein 等人[37] 将问题、用户答案与回答用户组成三元组, 判断相互之间的影响力, 并预测回答的质量. 使用这种方法得到的答案基本符合要求, 但是可能出现问题回复不及时的现象. Jeon 等人[38] 提出了一个使用非文本特征预测文档质量的框架, 综合考虑答案的接受率、回答者的积极程度、回答者的回答问题偏好、用户的推荐次数、点击次数等特征来向用户推荐问题.

(2) 答案质量评估. 答案质量评估有助于全面提升社区问答中的答案质量, 主要关注问题和答案之间的匹配程度. 如果已有的答案质量较高, 符合用户所想提问的问题, 用户就不用花时间在 CQA 问答系统中发布同样问题. 在特征工程中通常选用文本特征和链接特征来进行答案质量特征, 孔维泽等人[39] 综合考虑文本特征、链接特征以及时序、问题粒度和百度知道社区用户等特征, 从多方面评估答案质量, 实验表明在文本

特征与连接特征的基础上综合考虑其他特征能提升答案质量评估的效果. Bian 等人[40] 提出了 GB Rank 算法来进行答案质量评价; 沈旺等人[41] 提出一种新的回答质量评价指标, 融合用户评价标准和数据特征, 依据回答评论对文本进行向量化表示, 利用支持向量机对文本进行标签表示学习, 实验证明上述方法可以获得 85.32% 的分类准确率, 高于仅使用用户评价标准指标的 61.44% 和仅使用数据特征指标的 79.10%.

(3) 问句分类. 由于在 CQA 问答系统中, 用户可以进行提问也可以进行回答, 所以这种类型的问答系统会积累大量的信息. 为了保证系统的健壮性, 需要对这些信息进行分类存储. 熊大平[42] 针对用户不能确定问题类别而随便指定一个, 导致分类体系杂乱的问题, 提出一种结合问题分类和答案分类的组合模型, 等得到答案后, 再利用答案信息对问题进行分类. 延霞等人[43] 针对 CQA 问答提出了一个粗粒度的分类体系, 将问句分为 13 个类别, 以及提出一个使用分布策略的多标记多分类的问句分类算法 MLMC. Li 等人[44] 对于 CQA 问答提出了一个联合培训系统 CoCQA 框架, 该框架可以识别用户的主观或复杂的问题并对这些问题进行分类, 利用问题之间的关联, 自动判定答案.

(4) 相似问题检索. 相似问题检索是在已解问题的基础上利用问题特征相似程度来求解新问题的过程, 这一过程需对用户问题进行相似评估, 返回答案或者相似问题列表. Liu 等人[45] 提出了一种与语言无关的技术来解决数据提取问题, 此模型主要应用于网页中, 将页面转化为视觉区域树, 然后定位包含目标信息的区域, 最后从目标区域提取信息. Duan 等人[46] 首先将问题的主题和重点进行结合, 然后把这些信息加入到语言模型中, 提高了搜索的相关性. Cao 等人[47] 利用主题聚类和问题重点聚类的方法来呈现与查询相关的问题, 并生成查询问题与已存在问题的相关性分数; 熊大平[42] 提出了一种基于 LDA 的匹配框架来计算问句相似度, 分别利用基于 VSM 的统计模型、基于 WordNet 的语义模型、基于 LDA 的主题模型对问句的统计信息、语义信息和主题信息进行问句相似度计算, 在真实数据中的实验取得较好效果.

3.2.2 小结

基于问题答案对的问答系统在工业界有很多的应用, 例如知乎、腾讯知文-结构化 FAQ 问答引擎、一些购物 APP 中的客服机器人都是应用了 FAQ 问答系统的方法或者 CQA 问答系统的方法, 广泛的应用说明了这一时期问答系统的优越性, 但这其中仍有一些问题需要解决, 比如说如何自动获取有标记的相似文本训练数据、如何降低获取与维护高质量的问答对数据的成本、用户输入的问题类型较多如何解决以及在各个阶段使用模型的鲁棒性无法保证等问题.

3.3 基于知识图谱的问答系统

2012 年, 知识图谱 (knowledge graph, KB) 这一概念由谷歌提出, 同时还发表了 Google Knowledge Graph[48]. 2013 年之后工业界和学术界对这一概念的接受程度越来越高, 关于知识图谱的文献也越来越多. 本质上来说, KB 是一种语义网络的知识库, 其节点表示实体, 边表示实体之间的关系, 形成一个错综复杂的网状结构. 在进行问题处理的时候, 使用实体识别等技术将用户问句中的主题实体识别出来, 再在知识库中获取与主题实体相关的其他实体, 从而进行答案的获取.

基于知识图谱的问答系统, 亦可称之为知识库问答系统, 是根据对自然语言问题的理解并依赖知识库获取答案的一种问答方式, 用户除了可以得到答案还可以得到与其相关的内容. 目前基于知识图谱的问答系统采用的主要方法为模板匹配、语义解析和深度学习, 不同方法之间的比较如表 2 所示.

表 2 基于知识图谱的问答系统方法比较

方法	优点	缺点
基于知识图谱的模板匹配方法	准确、迅速	需要人工构建大量的模板且模板难以复用
基于知识图谱的语义解析的方法	可解释性强	对数据要求较高, 依赖高质量的解析算法
基于知识图谱的深度学习的方法	几乎不需要手工定义的特征	趋于黑盒, 缺少可解释性

3.3.1 基于知识图谱的模板匹配方法

基于模板的问答方法属于比较传统的方法, 需要预先构建好模板, 不用对问题进行分析而将其转化为三元组的形式, 再根据三元组寻找相匹配的查询模板, 最后根据匹配到的查询模板在知识库中进行数据匹配获得答案. 具体流程如图 8 所示.

初期, 基于模板的知识图谱问答通过构造一组模板参数, 形成查询表达式, 对问题文本进行匹配. 这样的流程可以较为容易的回答出用户提出的简单问题, 但对于复杂问题难以处理, 并且这种形式的模板需一一对应, 要耗费大量人力构建模板, 不够灵活, 所以一些研究者对模板匹配的方法进行了深入研究.

针对问答模板不能复用的问题, Tunstall-Pedoe 等人[49] 提出 True Knowledge 方法, 首先用问句与模板中的部分词进行对应, 其次根据问句的内容, 进一步映射填充到查询模板中, 这种方法不用为每一个问题都构建一个模板, 使得一个模板可以覆盖多个问题, 但是该方法需要大量人工处理才能形成模板且模板对数据库依赖性大, 只支持英语. Unger 等人[50] 提出一种模板改进的方法, 主要分为 4 步: 首先将自然语言问题通过句法分析映射为 SPARQL 模板, 其次进行模板实例化, 通过实体识别和关系识别将 SPARQL 模板进行槽位填充, 再次对多个模板进行排序, 最后用 SPARQL 模板从 RDF 数据查询获取结果. 这种方法不用规定用户必须使用固定格式的查询语言进行提问, 使用户可以用自然语言进行查询, 但是某些情况下, 生成的模板不一定可以和 RDF 数据库定义的结构对应上. Abujabal 等人[51] 提出了 QUINT, 可通过语料自动生成模板, 借助生成的模板将一个问题映射到一个知识库上进行查询, 该方法可以动态的学习新模板, 但不能解决聚合、排序等方面的问题. 为避免人工构建模板, Cocco 等人[52] 借助一个开放的 RDF 数据集, 通过机器学习方法在训练集上自动学习 SPARQL 模板. SPARQL 模板被提供给一个基于实例的分类器, 该分类器将用户问题与一对现有问题关联起来, 用于回答用户问题, 这是一种完全自动的方法, 但该方法准确率和召回率不高. 为了提高系统的性能, Liu 等人[53] 利用 BERT 预训练模型, 提出了一个基于 BERT 的知识库问答模型 BB-KBQA, 该模型可以捕获问题、实体和关系之间的深层语义信息, 取得很好的效果.

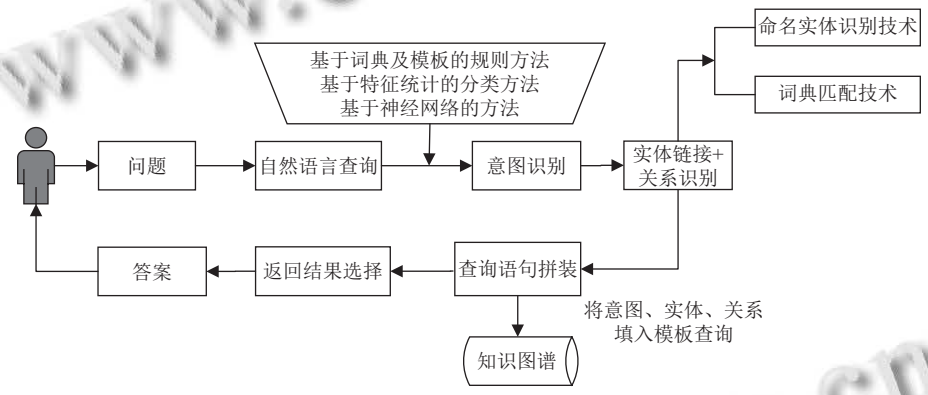

图 8 基于知识图谱的模板匹配方法总体结构

3.3.2 基于知识图谱的语义解析方法

基于知识图谱的语义解析方法与模板匹配方法的最大不同在于逻辑表达式, 模板匹配方法需要提前设定好表达的方式, 而语义解析方法是对用户提出的问句进行解析, 将其转化为一种形式化的意义表示, 如逻辑形式, 根据问句的逻辑形式查询知识库得到结果, 具体流程如图 9 所示.

基于语义解析的知识图谱问答方法的关键在于如何分析用户使用自然语言提出的问题, 并将分析结果转换为机器可理解的查询语言, 常用有 3 种方法: 基于词典－文法、基于语义图以及基于神经网络, 3 种方法之间的优缺点如表 3 所示.

(1) 基于词典-文法的语义解析方法. Berant 等人[54] 利用语义解析器进行知识图谱问答, 对问句连接、求交、聚合和桥接 4 种操作进行结点合并来构造语法树, 还介绍了一种经典的逻辑语言 λ-DCS, 用于减少逻辑表达的复杂性. Li 等人[55] 提出了上下文无关文法 SCFG, 该方法在英语、中文、泰语、德语和希腊语中都有不错的效果.

(2) 基于语义图的问句解析方法. Reddy 等人[56] 提出了一种利用图匹配将问句映射到知识库的新思路用于知识图谱问答. Yih 等人[57] 提出了一个适用于知识图谱问答的语义解析框架, 将语义解析简化为查询图生成, 缩小了搜索空间. 在提升问句解析的准确度上, Chen 等人[58] 提出 Sequence-to-Action 模型, 将语义分析建模为端到端的语义图生成过程, 并在解析过程中加入句法约束条件和语义约束条件, 提高了问句解析的准确度. 对于复杂的问题, Lan 等人[59] 改进了查询图的生成方法, 允许更长的关系路径存在, 并且将合并约束的过程融入进关系路径构造, 解决复杂问题解析路

径过多的问题. 针对大型知识库, Zhang 等人[60] 提出了一种因果增强的表填充器用于图结构生成, 可以克服序列建模中的问题并学习内部的因果关系.

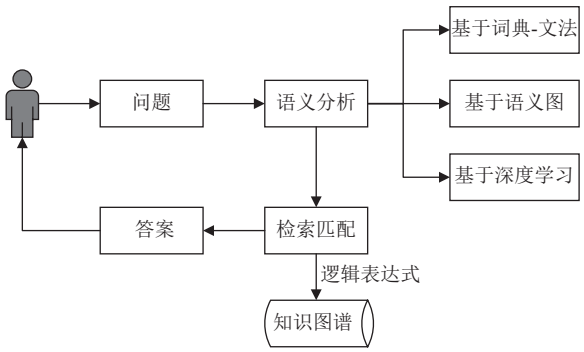

图 9 基于知识图谱的语义解析方法总体结构

表 3 基于知识图谱的语义解析方法比较

方法	优点	缺点
基于词典-文法的语义解析方法	解析过程清晰, 可解释性强	需学习语法知识, 且受限于词典的覆盖度
基于语义图的问句解析方法	与知识库联系紧密, 可用知识库辅助构建语义图	依赖特定手段构建语义图, 通用性不强
基于神经网络的方法	充分利用神经网络模型的表示能力和拟合能力	可解释性差, 模型训练时间长, 参数不易调节

(3) 基于神经网络的方法. Dong 等人[61] 提出了一种从粗到细的神经语义分析解码框架, 先生成忽略细节的表示, 再将细节填充到之前的表示中, 该框架可以较为容易的适应不同领域的表达. Chen 等人[62] 通过一个双向注意力记忆网络, 不仅从问句中提取特征进行语义分析, 且结合知识图谱中的一些信息, 使问答准确性提高. 不同于依赖实体识别、分段及分类等各种组件的问答系统, Lukovnikov 等人[63] 将研究聚焦于解决字符级问题, 训练出的网络可以让模型自己进行决策, 以端到端的方式回答简单问题, 实现了不错的效果.

3.3.3 基于知识图谱的深度学习方法

基于知识图谱的深度学习方法利用深度神经网络模型对问题内容与知识图谱三元组的高维抽象表达进行相似度计算, 利用预定义的评分机制获得最优答案, 又称向量建模的方法, 具体处理流程如图 10 所示[64].

在这一方法的研究过程中, 最为经典的是 2014 年 Bordes 等人[65] 提出的使用 embedding 模型, 该模型首先从句中的主题词对应的知识库实体出发, 找出几个与问句相关的实体关系, 作为候选答案, 根据实体关系与问句之间的相似度对答案进行排序, 选择相似度最大的返回给用户, 用户可获取到较为满意的答案, 但该模型没有考虑到问题的语言顺序. 2015 年, Dong 等人[66] 提出了一个基于 Freebase 数据集的 MCCNNs 的自动问答模型, 在传统的向量建模方法中融入了卷积神经网络, 通过 3 个通道来分别针对答案路径、答案背景信息及答案类型来学习理解问题, 使问答性能得到提升. 2016 年, Zhang 等人[67] 在前面研究的基础上, 使用双向 LSTM 并结合问题引入注意力机制提取问句特征, 效果优于 MCCNNs 方法. Hao 等人[68] 针对 MCCNNs 中的 3 个神经网络不灵活的问题, 提出了一个端到端的交叉注意神经网络模型, 根据主题实体联合知识库生成候选答案, 通过交叉注意力模型动态化的表示问题与候选答案的关联及计算相应的分数, 并将知识库本身作为训练数据, 捕捉整个知识库的全局结构, 解决词汇表外的问题. 2018 年, Qu 等人[69] 注重于句中的原始信息, 从而提出了一种基于相似矩阵的递归神经网络模型, 利用 RNN 和 CNN 互补的优势对句中信息进行提取, 使用注意机制同时关注实体和关系, 该模型可在未来扩展应用于解决复杂问题. 2020 年, Luo 等人[70] 关注问题和知识库事实之间的关系, 提出了一种基于 BERT 的单关系问答方法 SR-QA, 问答效果得以提升. 2021 年, Shi 等人[71] 针对复杂问题提出了 TransferNet, 该方法从问题的主题实体开始计算图中关系的分数, 然后将关系分数转化为一个邻接矩阵, 通过将实体得分向量与关系得分矩阵相乘, 以可微分的方式沿着关系进行多次查询得到答案, 在 MetaQA数据集的二跳和三跳问题上实现了 100% 的准确率.

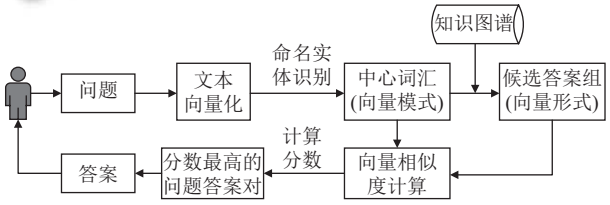

图 10 基于知识图谱的深度学习的方法总体结构

3.3.4 小结

基于知识库的问答系统无论是在学术界还是在工业界都很受关注, 以上总结了学术界在知识库问答上的许多研究, 工业界中很多公司也构建对应的领域知识图谱用于自身平台的不同应用场景, 例如 Facebook 社交知识图谱、Amazon 商品知识图谱、阿里巴巴商品知识图谱等.

3.4 基于大规模语言模型的问答系统

2022 年底, OpenAI 推出了一款专注于对话生成的语言模型 ChatGPT, 它可以很好地理解用户意图, 做到有效的多轮沟通, 回答内容完整、重点清晰、有概括、有逻辑、有条理. 除了聊天, 还可以根据用户提出的要求, 进行机器翻译、文案撰写、代码撰写等工作, 在众多行业领域有着广泛的应用潜力. 它的出现给自然语言处理研究范式带来了新的变革与新的思路, 推动了从大规模预训练语言模型 (large pre-trained language model, LPLM) 走向通用人工智能 (artificial general intelligence, AGI) 的转换, 学术界和企业界纷纷迅速跟进类 ChatGPT 模型研发. 以下将围绕 ChatGPT 的技术手段进一步解释它的出现对于问答来说改变了什么, 又是如何改变的.

3.4.1 ChatGPT 技术手段

ChatGPT 的出现离不开大量研究和技术积累, 基础生成模型的迭代创新、预训练模型的出现、参数量级的提升, 逐步构建了 ChatGPT 出现的关键技术要素支撑.

2018 年, Radford 等人[72] 提出了基于生成式预训练的语言模型 GPT, 该模型在 Transformer[73] 的基础上利用无监督预训练和有监督微调的手段训练模型, 在多个 NLP 任务上取得了当时的先进效果, 但同年 BERT 的出现导致 GPT 模型没有在学术界引起很大的重视. 2020 年, Brown 等人推出了 GPT3[74], GPT3 将 Transformer 层由 12 层扩展到 96 层, 使得该模型有了更好的零样本学习能力、小样本学习能力和自然语言生成能力, 这也是 ChatGPT 中应用的预训练模型.

理解用户意图是问答系统的关键, 但模型变大并不意味着它理解用户意图的能力更强. 为此, 研究人员在 ChatGPT 的训练流程中加入 instructGPT[75]. instructGPT 是研究人员人工收集了一组按照人说话的方式的数据集, 利用人类反馈的强化学习 (reinforcement learning from human feedback, RLHF)[76] 对模型进行微调, 结果证明利用该方法进行微调可以使语言模型与人类意图对齐. 在 ChatGPT 中, 提问数学问题时, 会给出整个问题的推导过程, 这是因为加入了思维链[77], 思维链可以用于提升大规模语言模型在算数、常识和符号上的推理能力, 这种能力仅是在训练集中增加中间步骤的说明, 无需重新训练或微调模型就可以获得. ChatGPT 还有即时学习的能力, 这种能力称作 In-Context Learning[78], 本质上是执行了一个隐式的微调. ChatGPT 在进行训练时使用了一种为下游任务设计出的一种被称为 Prompt[79] 的输入形式或模板, 它可以起到帮助预训练模型想起自己在预训练时接触到的知识的作用.

3.4.2 扩展应用

ChatGPT 背后涵盖的技术目前已在其他场景下有所应用, 例如 ResearchGPT, 这是一个适合于科研人员的问答系统, 只需要上传论文和提问问题, 就可以获取到答案, 但如果提问与论文中图表有关的问题就无法得到相应的答案, 微软也将 ChatGPT 引入 Bing 搜索引擎和 Edge 浏览器中, 处理更复杂的搜索提问. 2023 年 2 月, 北京大学深圳研究生院发布一款工具 ChatExcel, 用户可以直接用自然语言对表格中的数据信息进行查询、修改等操作. 同年 3 月, 微软发布的视觉聊天系统 Visual ChatGPT 将 ChatGPT 和多个 SOTA 视觉基础模型连接, 实现在对话系统中理解和生成图片. 可以接收和发送文本和图像提供复杂的视觉问答, 或者视觉编辑指令, 通过多步推理调用工具来解决复杂视觉任务, 该工作开启了 ChatGPT 借助视觉基础模型作为工具, 进行视觉任务处理的研究方向.

3.4.3 小结

ChatGPT 在问答系统的发展中是划时代的出现. 但它身上也有一些不足之处, 例如对于一个问答系统来说, 结合上下文语境为用户返回精准、可靠的答案是必要的, 而 ChatGPT 所回答的内容虽然流畅, 但有时存在事实性的错误, 而提问者无法判断是否准确; 另一方面, 换一个提问方式可能会得到不同的结果, 造成结果不稳定的现象; 同时训练成本高, 资源消耗大, 使得该模型难以获取新的知识. 有研究者提出对于 ChatGPT 在回答问题上出现的不足或许可以尝试与知识图谱相结合的方式进行改正. 目前, 对于高频且常规类的问答, 仍需要传统的技术来提供稳定且可靠的服务, 而 ChatGPT 提供的开放域问答能力, 可以用来被处理长尾且低频的问答. 所以尽管 ChatGPT 前景无限, 但并不会全面代替现有的问答技术, 对于垂直领域的问答来说现有的问答技术更加合适. 在未来的问答形式上, 现有问答技术可以和 ChatGPT 共存, 优势互补.

4 多模态问答系统

人们生活在一个多模态世界中, 语言的表达、文字的传递、视觉的感受, 综合处理不同模态的信息更

符合人类认知世界的方式. 不满足于只在文字一种模态进行问答, Tu 等人[80] 在 2014 年首次提出了视频问答 (video question answering), 用一个共同解析视频和文本的框架, 来了解事件并回答用户的问题. 2015 年, Agrawal 等人[7] 提出了视觉问答 (visual question answering), 将图片和文本联系起来, 生成符合图片的答案. 用户可以对一张有着人物的图片进行提问, 如“图中有几个女人?”, 系统给出相应答案, 这种形式的问答可以应用于对孩子的启蒙教育中[81]. 视频问答和视觉问答都是多模态问答系统的研究内容, 本文将这两种问答系统统称为多模态问答系统, 总体结构如图 11 所示.

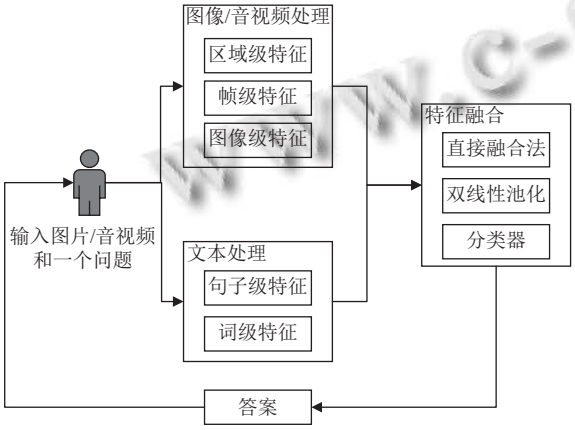

图 11 多模态问答系统总体结构

虽然视频问答要比视觉问答先提出, 但是视觉问答的研究进展得更快, 下文将对这两种问答分别进行总结.

4.1 视觉问答方法

视觉问答方法基本分成 4 类: 联合嵌入、注意力机制、神经网络架构和知识库增强, 这 4 种方法各有优缺点, 如表 4 所示.

表 4 视觉问答系统方法比较

方法类别	优点	缺点
联合嵌入	方法直接, 易理解	只能捕获训练集中存在的知识, 鲁棒性不强
注意力机制	通过增强模型对特征的辨识能力提高精度	对于需要更长的推理类问题效果不好
神经网络架构	有效利用监督学习方式提升信息综合能力	会忽略一些语法线索, 理解数学问题有难度
知识库增强	可以获取更多信息	模型的问题类型有限

4.1.1 基于联合嵌入法的视觉问答

联合嵌入法受到计算机视觉和自然语言处理中的深层神经网络的推动, 方法主要是使用 CNN 和 RNN 来学习图像和句子在公共特征空间中的嵌入, 再将它们一起输入分类器预测答案.

Zhou 等人[82] 在使用 CNN 和 RNN 来解决视觉问答的基础上提出了 iBOWING 模型, 该方法使用词袋模型来提取问题的文本特征, 用 GoogLeNet 模型提取图片的视觉特征, 将文本特征与图像特征连接起来得出答案. 同年, Malinowski 等人[83] 以 CNN 和 LSTM 为基础, 设计了一个预测结果长度可变的模型, 提取两种特征之后同时输入到 LSTM 编码器和解码器, 最后生成一个可变化长度的答案. Kafle 等人[84] 提出了在问题中预测答案类型并生成答案的框架, 使用贝叶斯算法对目标空间关系建模, 计算出每个答案概率. 以上视觉问答的方法主要是学习图像和问题的联合表示, Do 等人[85] 认为只关注图片和问题两个方面不够全面, 所以提出了一种三线性交互模型, 该模型在训练时学习图片、问题和答案之间的关联, 在测试时使用知识蒸馏将三线性交互模型转为双重交互模型, 这是首次将知识蒸馏引入视觉问答, 在思想上有较大的借鉴意义.

4.1.2 基于注意力机制的视觉问答

注意力机制模拟了人脑的认知模式, 允许模型对来自不同区域的特征赋予不同的重要性, 不同于联合嵌入法的使用全局特征会向预测阶段提供不相关或有噪声的信息. 2015 年, Xu 等人[86] 提出了将注意力应用于视觉任务的思路, 将图像描述生成的关注点集中在图像中的显著区域, 这一概念可以转移到视觉问答任务, 即关注与问题相关的图像区域, 这也说明在推理计算过程中需明确“看哪里”的问题.

2016 年, Shih 等人[87] 第 1 次将注意力机制用于问答, 提出了一个图像区域选择机制, 用于学习识别与问题相关的区域, 在回答“什么颜色”和“什么房间”等问题方面有了显著改进. Zhu 等人[88] 在 LSTM 中融入了关注图像特定区域的注意力机制, 每输入一个词就产生一个注意力图, 在特征融合的方法上选择了直接融合法. 以上两种方法的核心都是通过关注某一个图像的区域来回答问题, Yang 等人[89] 认为视觉问答中有一定的推理过程, 因此采用注意力机制实现分层关注的推理过程, 提出 SAN 模型来处理视觉问答任务, 证明了每次注意力机制都是一次推理的过程, 每一次使用注意力机制后都可以关注到更细的内容. 也有研究者认为现有注意力机制关注的区域与人类关注的区域不同,

Patro 等人[90] 提出通过差分注意力网络 DAN 和差分上下文网络 DCN 来获取一个微分注意力区域, 使用该方法计算出来的微分注意力比其他方法更接近人类注意力, 从而提高了回答问题的准确率.

4.1.3 基于神经网络架构的视觉问答

神经网络架构可以根据特定需要的功能设计不同的模块, 用模块的组成反映问题的结构, 利用这种形式可以在不同的任务中复用部分模块.

2016 年, Andreas 等人[91] 提出一种专为视觉问答设计的神经模块网络模型 NMN, 它由多个模块化的网络组合, 且模型的网络是根据问题的语言结构动态生成的, 这里面的所有模块都是独立并且可以自由组合的. Kumar 等人[92] 将动态存储网络 DMN 用于视觉问答, 模型主要由 4 个独立的模块组成, 输入模块、问题模块、情景记忆模块和答案模块, 每个模块都是使用 GRU 作为编码的基础模型且模型性能依赖于注意力机制的效果, 如果需要对某个模块进行改进的话, 可以在 GRU 和注意力机制上进行改进, 其中情景记忆模块是将问题、之前的记忆和客观事实作为当前迭代的输入, 然后在每次迭代中更新情景记忆获得当前的记忆. NMN 和 DMN 是神经网络架构的两个重要的模块化模型, 之后的很多研究都借鉴了这两个网络. Xiong 等人[93] 在 DMN 的基础上, 对输入模块、注意力机制和情景记忆模块都进行了更新, 在视觉问答上取得了较好的成果. 以上方法都是在数据量大的情况下进行的, 在数据量小的情况下, 对于新问题的判断准确度有限, Guo 等人[94] 提出了一个两阶段网络用于解决小样本的视觉问答任务, 并提出了一个小样本的视觉问答数据集. 考虑到 NMN 不是从数据中进行学习而是依靠解析器和相关网络架构来学习, 对复杂问题的适应能力不足, 进而阻碍模型的表示能力和泛化能力. 针对上述问题, Han 等人[95] 提出了一个用于视觉问答的动态模块化路由框架 SUPER, 该框架可以更好地捕捉特定于实例的视觉语义特征, 并对预测的判别表示进行优化, 这项工作对视觉问答的架构学习和表示校准提供了新的角度.

4.1.4 基于知识库增强的视觉问答

知识库增强方法是通过查询结构化知识库来解决外部数据的使用问题, 这样可以检索现有信息中不存在的信息, 解释推理的过程. 在知识库增强下还可以增加可回答问题的复杂性, 主要使用的外部知识库有结构化的 DBpedia 和 ConceptNet, 非结构化或半结构化的 Wikipedia 和 Visual Genome.

2015 年, Wang 等人[96] 提出了一种基于 DBpedia 的视觉问答网络 Ahab, 该网络首先用 CNN 提取视觉概念, 结合 DBpedia 中相似的概念通过推理获得答案, 还提出了一个需要视觉、常识、外部知识库共同进行回答的数据集 KB-VQA. 在这之后, 他们对 Ahab 模型进行了梳理和改进, 并提出了一个基于常识知识的视觉问答数据集 FVQA[97], 但是这一模型不能很好地处理出现同义词和同形词的用户问题. 为解决这一问题, Narasimhan 等人[98] 将一种端到端的方法用于具有知识库的视觉问答, 但是该方法在每个节点都引入了噪声, 导致模型不够灵活. 2019 年, Marino 等人[99] 提出了一个只含有需要外部知识回答问题的大规模数据集 OK-VQA, 同时使用 ArticleNet 预测查询内容在互联网上的非结构化数据中是否出现及出现位置, 辅助回答问题.

2020 年, Zhu 等人[100] 结合上述方法的优点提出了一个具有比较好的解释性的模型, 将图像表示成一个多模态的异构图, 来互相补充和增强视觉问答任务的信息, 该模型在 FVQA 上有很好的效果. 同年, Gardères 等人[101] 提出一个概念感知的端到端管道 ConceptBert, 聚合视觉、语言和外部知识嵌入来学习, 该方法不需要额外的知识标注或者是搜索查询, 降低了计算成本. 人类可以根据以往的经验知识对一个新的概念有所认识, 让机器仅从一个或一小撮样本中学习一个新的概念, 称作为零样本学习. 之前的方法大多忽略了可能会出现对候选答案外的结果不能进行预测的情况, 2021 年, Chen 等人[102] 针对这个问题提出了一种使用知识库的 ZS-VQA 算法, 通过掩码来调整答案预测分数, 同时提出了一个零次事实视觉问答数据集 ZS-F-VQA, 用于评估 ZS-VQA 看不见的答案. 2022 年, Ding等人[103] 提出了一个端到端的多模态知识抽取与积累模型 MuKEA, 通过域内和域外数据的训练, 模型积累了广泛的多模态知识并基于知识检索进行答案预测, 在两个经典的 KB-VQA 数据集上超越了以往的模型.

4.2 视频问答方法

视频问答与视觉问答的区别在于[104,105]: (1) 视频问答的任务对象不是单一的静态图像, 而是序列的图像信息; (2) 视频中存在大量时间线索, 问答上需要更多的时间推理; (3) 对于不同问题, 需要不同数量的帧

来得到答案. 由于有这些不同点, 所以不能将视觉问答的方法直接应用于视频问答, 而需针对任务对象的不同需要进行改变, 因此使用较多的方法为融入注意力机制和融入记忆网络.

4.2.1 基于注意力机制的视频问答

Zhao 等人[106] 从分层双层注意力网络学习的角度研究了视频问答, 用帧级的特征表示方法来获取视频中的对象外观和运动信息, 再利用分层双层注意力网络来学习词级和句子级的问题特征, 在实验中验证了该方法的有效性. 上述方法是针对单个视频的, Liang 等人[107] 提出了一种可以处理多个视频或者一系列照片的端到端的方法, 称为焦点视觉文本注意网络 FVTA, 利用分层过程来动态的确定在顺序数据中关注什么媒体和什么时间来回答问题. Li 等人[108] 提出了一种可学习聚合网络与多样性学习架构 LAD-Net, 用基于多路径金字塔共同关注机制来处理视频内容复杂多样的问题. Gao 等人[109] 构建了一个视频问答数据集 Env-QA, 其中的每个视频都由一系列关于在环境中探索和互动的事件组成, 并提出了一个视频问答模型, 引入事件级视频表示和相应的注意力机制的时间分割和事件注意力网络 TSEA, 该网络可以更好地提取环境信息并回答问题.

4.2.2 基于记忆网络的视频问答

Ge 等人[110] 探索如何在视频中关注到与问题相关的视频区域, 提出了一个忘记记忆网络 FMN, 该网络可以选择与问题相关的局部特征, 忽略无关特征. Gao 等人[104] 通过对比视觉问答和视频问答的区别提出了运动—外观共记忆网络, 利用一种共记忆注意力机制, 来记录视频中的运动和外观特征, 使用时间卷积—反卷积框架生成多层次的上下文事实, 并用一种动态事实融合的方法, 动态构造不同问题的时间表示. Kim 等人[111] 针对电影视频数据时常较长难以定位与问题相关位置和不同问题需要从视频或者字幕来推断答案的两个难点, 提出了针对电影故事问答的渐进式注意记忆网络 PAMN, 采取渐进式注意力机制来精确定位所需的视频部分, 并从记忆中过滤出与问题无关的信息, 利用融合记忆依次衡量每个答案的信度并校正每个候选答案的分数, 在 MovieQA 和 TVQA 这两个数据集上取得了很好的效果.

4.3 小结

多模态问答是人工智能 2.0 时代研究的热门方向, 对多种模态的任务对象: 文本、图像与音视频进行处理, 是有意义又充满挑战的问题, 其中还有很长的路要走. 对于视觉问答来说, 研究难点一方面是如何改进模型提升准确率, 另一方面还有来自数据集的问题, 现有的视觉问答数据集有一部分是用电脑自动生成的, 不太符合生活实际, 一个图片对应的描述通常只有一个, 但是现实生活中人们对于一张照片的理解是多种多样的. 视觉问答也有很广泛的应用场景, 如何将视觉问答实实在在的应用在现实中也是一个挑战. 对于视频问答来说, 发展速度没有视觉问答那么迅速, 虽然研究过程中的一些任务已经有初步解决办法, 但是效果还远不及其他的问答系统. 现有视频问答研究中普遍忽略了视频中含有的另一个模态信息即音频, 未来在视频问答中考虑利用文本、视频和音频 3 个模态进行交互, 可能会有不一样的效果.

5 总结与展望

从应用的角度来看, 问答系统及其衍生的设施, 在许多行业都有广泛的应用, 如医疗领域、教育和文化遗产保护领域、电商领域等, 智能问答系统逐渐融入人们生产生活的方方面面. 未来的问答系统将会推动互联网搜索引擎的进步, 从而使得从互联网获取知识的方式更加人性化与智能化.

虽然问答系统领域已经开展了大量的研究, 引入外部知识库或者其他模态的信息进行辅助, 努力解决问答中复杂问题的挑战. 但是大部分的方法由于不能真正地理解用户意图, 尤其是在多模态问答系统中这种情况更为突出, 所以问答的性能提升不明显, 使得用户的体验感欠佳. 因此, 基于多领域知识, 构建统一的、跨场景、多任务的多模态基础模型是智能问答以及 AI 的重点发展方向. 最新的多模态预训练模型研究有 Kosmos-1、GPT-4、BEiT v3、文心一言等, 通过对图片-文字进行联合表征学习, 并扩展到语音、视频等其他模态, 将在多模态问答取得明显优势. 大型语言模型的出现及越来越多的针对文本的扩散模型的研究也预示着问答系统会越发个性化与智能化. 使问答系统能够根据用户语言、行为、兴趣和偏好, 以更加自然、流畅的方式, 为用户提供准确、有用的答案. 此外, 人们希望问答系统能够拥有更高的自由度、更好的逻辑思维能力, 从而能够为用户解决决策类问题, 当用户面对类似于做与不做的问题时, 系统根据分析出的不

同利弊, 执行逻辑判断为用户做出选择, 这些都是未来研究的努力方向.

参考文献

1 Zhou L, Gao JF, Li D, *et al*. The design and implementation of XiaoIce, an empathetic social chatbot. Computational Linguistics, 2020, 46(1): 53–93. [doi: 10.1162/coli_a_00368]

2 Li FL, Qiu MH, Chen HQ, *et al*. AliMe Assist: An intelligent assistant for creating an innovative e-commerce experience. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore: ACM, 2017. 2495–2498.

3 Sun Y, Wang SH, Li YK, *et al*. ERNIE: Enhanced representation through knowledge integration. arXiv: 1904.09223, 2019.

4 MIT. START natural language question answering system. http://start.csail.mit.edu/index.php. (2022-02-08)[2023-02-22].

5 Mollá D, Vicedo JL. Question answering in restricted domains: An overview. Computational Linguistics, 2007, 33(1): 41–61. [doi: 10.1162/coli.2007.33.1.41]

6 Couto J. Introduction to visual question answering: Datasets, approaches and evaluation. https://tryolabs.com/blog. (2022-12-14)[2023-02-09].

7 Agrawal A, Lu JS, Antol S, *et al*. VQA: Visual question answering. International Journal of Computer Vision, 2017, 123(1): 4–31. [doi: 10.1007/s11263-016-0966-6]

8 陈子睿, 王鑫, 王林, 等. 开放领域知识图谱问答研究综述. 计算机科学与探索, 2021, 15(10): 1843–1869. [doi: 10.3778/j.issn.1673-9418.2106095]

9 冯钧, 李艳, 杭婷婷. 问答系统中复杂问题分解方法研究综述. 计算机工程与应用, 2022, 58(17): 23–33. [doi: 10.3778/j.issn.1002-8331.2201-0384]

10 姚元杰, 龚毅光, 刘佳, 等. 基于深度学习的智能问答系统综述. 计算机系统应用, 2023, 32(4): 1–15. [doi: 10.15888/j.cnki.csa.009038]

11 Turing AM. Computing machinery and intelligence. In: Parsing the Turing Test. Palo Alto: American Association for Artificial Intelligence, 1995.

12 Green BF, Wolf AK, Chomsky C, *et al*. Baseball: An automatic question-answerer. Proceedings of the 1961 IRE-AIEE-ACM (Western). Los Angeles: ACM, 1961. 219–224.

13 Woods WA. Semantics and quantification in natural language question answering. Advances in Computers, 1978, 17: 1–87.

14 Weizenbaum J. ELIZA—A computer program for the study of natural language communication between man and machine. Communications of the ACM, 1983, 26(1): 23–28. [doi: 10.1145/357980.357991]

15 Winograd T. Procedures as a representation for data in a computer program for understanding natural language. Technical Report, Boston: Massachusetts Institute of Technology, 1971.

16 Bobrow DG, Kaplan RM, Kay M, *et al*. GUS, a frame-driven dialog system. Artificial Intelligence, 1977, 8(2): 155–173. [doi: 10.1016/0004-3702(77)90018-2]

17 Wilensky R, Chin DN, Luria M, *et al*. The Berkeley UNIX consultant project. Artificial Intelligence Review, 2000, 14(1–2): 43–88.

18 Wallace R. ALICE. http://www.alicebot.org. (2022-12-28)[2023-02-24].

19 Lehnert WG. A conceptual theory of question answering. Proceedings of the 5th International Joint Conference on Artificial Intelligence. Cambridge: Morgan Kaufmann Publishers Inc., 1977. 158–164.

20 Li X, Roth D. Learning question classifiers. Proceedings of the 19th International Conference on Computational Linguistics. Taipei: ACL, 2002. 1–7.

21 Magnini B, Negri M, Prevete R, *et al*. Mining knowledge from repeated co-occurrences: DIOGENE at TREC 2002. Proceedings of the 11th Text Retrieval Conference. Gaithersburg: National Institute of Standards and Technology, 2002.

22 Xiao GY, Mo JQ, Chow E, *et al*. Multi-task CNN for classification of Chinese legal questions. Proceedings of the 14th IEEE International Conference on e-Business Engineering. Shanghai: IEEE, 2017. 84–90.

23 杨建飞. 基于知识图谱的军事装备问答系统设计与实现 [硕士学位论文]. 武汉: 华中科技大学, 2021.

24 张永亮. 基于知识图谱的苹果病虫害智能问答系统研究 [硕士学位论文]. 咸阳: 西北农林科技大学, 2022.

25 Mihalcea R, Tarau P. TextRank: Bringing order into text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona: ACL, 2004. 404–411.

26 Page L, Brin S, Motwani R, *et al*. The PageRank citation ranking: Bringing order to the Web. Stanford Digital Libraries Working Paper, 1998: SIDL-WP-1999-0120.

27 El-Beltagy SR. KP-Miner: A simple system for effective keyphrase extraction. Proceedings of the 2006 Innovations in Information Technology. Dubai: IEEE, 2006. 1–5.

28 Moldovan D, Paşca M, Harabagiu S, *et al*. Performance

issues and error analysis in an open-domain question answering system. ACM Transactions on Information Systems, 2003, 21(2): 133–154. [doi: 10.1145/763693.763694]

29 Tellex S, Katz B, Lin J, *et al*. Quantitative evaluation of passage retrieval algorithms for question answering. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto: ACM, 2003. 41–47.

30 Chu-Carroll J, Prager J, Czuba K, *et al*. Semantic search via XML fragments: A high-precision approach to IR. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle: ACM, 2006. 445–452.

31 李宇, 刘波. 文档检索中文本片段化机制的研究. 计算机科学与探索, 2020, 14(4): 578–589.

32 Soubbotin MM. Patterns of potential answer expressions as clues to the right answer. Proceedings of the 10th Text REtrieval Conference. Gaithersburg: TREC, 2002.

33 Yang H, Chua TS. The integration of lexical knowledge and external resources for question answering. Proceedings of the 11th Text Retrieval Conference. Gaithersburg: National Institute of Standards and Technology, 2002.

34 Lin DK, Pantel P. Discovery of inference rules for question-answering. Natural Language Engineering, 2001, 7(4): 343–360. [doi: 10.1017/S1351324901002765]

35 Moldovan DI, Harabagiu SM, Girju R, *et al*. LCC tools for question answering. Proceedings of the 11th Text Retrieval Conference. Gaithersburg: National Institute of Standards and Technology, 2007.

36 Jurczyk P, Agichtein E. Discovering authorities in question answer communities by using link analysis. Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management. Lisbon: ACM, 2007. 919–922.

37 Agichtein E, Castillo C, Donato D, *et al*. Finding high-quality content in social media. Proceedings of the 2008 International Conference on Web Search and Data Mining. Palo Alto: ACM, 2008. 183–194.

38 Jeon J, Croft WB, Lee JH, *et al*. A framework to predict the quality of answers with non-textual features. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle: ACM, 2006. 228–235.

39 孔维泽, 刘奕群, 张敏, 等. 问答社区中回答质量的评价方法研究. 中文信息学报, 2011, 25(1): 3–8.

40 Bian J, Liu YD, Zhou D, *et al*. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. Proceedings of the 18th International Conference on World Wide Web. Madrid: ACM, 2009. 51–60.

41 沈旺, 李世钰, 刘嘉宇, 等. 问答社区回答质量评价体系优化方法研究. 数据分析与知识发现, 2021, 5(2): 83–93.

42 熊大平. 社区问答中问句相似度计算和分类技术的研究 [硕士学位论文]. 大连: 大连理工大学, 2013.

43 延霞, 范士喜. 面向问答社区的粗粒度问句分类算法. 计算机应用与软件, 2013, 30(1): 219–222, 286.

44 Li BL, Liu YD, Agichtein E. CoCQA: Co-training over questions and answers with an application to predicting question subjectivity orientation. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu: ACL, 2008. 937–946.

45 Liu W, Meng XF, Meng WY. Vision-based web data records extraction. Proceedings of the 9th International Workshop on the Web and Databases. Chicago: WebDB, 2006.

46 Duan HZ, Cao YB, Lin CY, *et al*. Searching questions by identifying question topic and question focus. Proceedings of ACL-08: HLT. Columbus: ACL, 2008. 156–164.

47 Cao YB, Lin CY. Clustering question search results based on topic and focus: 20100030769. 2010-02-04.

48 Singhal A. Introducing the knowledge graph: Things, not strings. https://blog.google/products/search/introducing-knowledge-graph-things-not/. (2012-05-16)[2023-02-22].

49 Tunstall-Pedoe W. True knowledge: Open-domain question answering using structured knowledge and inference. AI Magazine, 2010, 31(3): 80–92. [doi: 10.1609/aimag.v31i3.2298]

50 Unger C, Bühmann L, Lehmann J, *et al*. Template-based question answering over RDF data. Proceedings of the 21st International Conference on World Wide Web. Lyon: ACM, 2012. 639–648.

51 Abujabal A, Yahya M, Riedewald M, *et al*. Automated template generation for question answering over knowledge graphs. Proceedings of the 26th International Conference on World Wide Web. Perth: International World Wide Web Conferences Steering Committee, 2017. 1191–1200.

52 Cocco R, Atzori M, Zaniolo C. Machine learning of SPARQL templates for question answering over linkedspending. Proceedings of the 28th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises. Napoli: IEEE, 2019. 156–161.

53 Liu AT, Huang ZQ, Lu HT, *et al*. BB-KBQA: BERT-based knowledge base question answering. Proceedings of the 18th China National Conference on Chinese computational linguistics. Kunming: Springer, 2019. 81–92.

54 Berant J, Chou A, Frostig R, *et al*. Semantic parsing on freebase from question-answer pairs. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: ACL, 2013. 1533–1544.

55 Li JH, Zhu MH, Lu W, *et al*. Improving semantic parsing with enriched synchronous context-free grammar. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015. 1455–1465.

56 Reddy S, Lapata M, Steedman M. Large-scale semantic parsing without question-answer pairs. Transactions of the Association for Computational Linguistics, 2014, 2: 377–392. [doi: 10.1162/tacl_a_00190]

57 Yih WT, Chang MW, He XD, *et al*. Semantic parsing via staged query graph generation: Question answering with knowledge base. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: ACL, 2015. 1321–1331.

58 Chen B, Sun L, Han XP. Sequence-to-Action: End-to-end semantic graph generation for semantic parsing. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 766–777.

59 Lan YS, Jiang J. Query graph generation for answering multi-hop complex questions from knowledge bases. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 969–974.

60 Zhang MH, Zhang RY, Li YZ, *et al*. Crake: Causal-enhanced table-filler for question answering over large scale knowledge base. Proceedings of the 2022 Findings of the Association for Computational Linguistics. Seattle: Association for Computational Linguistics, 2022. 1787–1798.

61 Dong L, Lapata M. Coarse-to-Fine decoding for neural semantic parsing. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 731–742.

62 Chen Y, Wu LF, Zaki MJ. Bidirectional attentive memory networks for question answering over knowledge bases. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 2913–2923.

63 Lukovnikov D, Fischer A, Lehmann J, *et al*. Neural network-based question answering over knowledge graphs on word and character level. Proceedings of the 26th International Conference on World Wide Web. Perth: International World Wide Web Conferences Steering Committee, 2017. 1211–1220.

64 袁博, 施运梅, 张乐. 基于知识图谱的问答系统研究与应用. 计算机技术与发展, 2021, 31(10): 134–140. [doi: 10.3969/j.issn.1673-629X.2021.10.023]

65 Bordes A, Chopra S, Weston J. Question answering with subgraph embeddings. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: ACL, 2014. 615–620.

66 Dong L, Wei FR, Zhou M, *et al*. Question answering over freebase with multi-column convolutional neural networks. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: ACL, 2015. 260–269.

67 Zhang YZ, Liu K, He SZ, *et al*. Question answering over knowledge base with neural attention combining global knowledge information. arXiv:1606.00979, 2016.

68 Hao YC, Zhang YZ, Liu K, *et al*. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: ACL, 2017. 221–231.

69 Qu YQ, Liu J, Kang LY, *et al*. Question answering over freebase via attentive RNN with similarity matrix based CNN. arXiv:1804.03317, 2018.

70 Luo D, Su JD, Yu SS. A BERT-based approach with relation-aware attention for knowledge base question answering. Proceedings of the 2020 International Joint Conference on Neural Networks. Glasgow: IEEE, 2020. 1–8.

71 Shi JX, Cao SL, Hou L, *et al*. TransferNet: An effective and transparent framework for multi-hop question answering over relation graph. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 4149–4158.

72 Radford A, Narasimhan K, Salimans T, *et al*. Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. (2018-12-23)[2023-02-27].

73 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

74 Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 159.

75 Ouyang L, Wu J, Jiang X, *et al*. Training language models to follow instructions with human feedback. arXiv:2203.02155, 2022.

76 Knox WB, Stone P. Augmenting reinforcement learning with human feedback. arXiv:1706.03741, 2017.

77 Wei J, Wang XZ, Schuurmans D, *et al*. Chain-of-thought prompting elicits reasoning in large language models. arXiv:2201.11903, 2023.

78 Dai DM, Sun YT, Dong L, *et al*. Why can GPT learn in-context? Language models secretly perform gradient descent as meta-optimizers. arXiv:2212.10559, 2022.

79 Liu PF, Yuan WZ, Fu JL, *et al*. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv:2107.13586, 2021.

80 Tu KW, Meng M, Lee MW, *et al*. Joint video and text parsing for understanding events and answering queries. IEEE MultiMedia, 2014, 21(2): 42–70. [doi: 10.1109/MMUL.2014.29]

81 Barra S, Bisogni C, De Marsico M, *et al*. Visual question answering: Which investigated applications? Pattern Recognition Letters, 2021, 151: 325–331. [doi: 10.1016/j.patrec.2021.09.008]

82 Zhou BL, Tian YD, Sukhbaatar S, *et al*. Simple baseline for visual question answering. arXiv:1512.02167, 2015.

83 Malinowski M, Rohrbach M, Fritz M. Ask your neurons: A neural-based approach to answering questions about images. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1–9.

84 Kafle K, Kanan C. Answer-type prediction for visual question answering. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4976–4984.

85 Do T, Tran H, Do TT, *et al*. Compact trilinear interaction for visual question answering. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 392–401.

86 Xu K, Ba JL, Kiros R, *et al*. Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR.org, 2015. 2048–2057.

87 Shih KJ, Singh S, Hoiem D. Where to look: Focus regions for visual question answering. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4613–4621.

88 Zhu YK, Groth O, Bernstein M, *et al*. Visual7W: Grounded question answering in images. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4995–5004.

89 Yang ZC, He XD, Gao JF, *et al*. Stacked attention networks for image question answering. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 21–29.

90 Patro B, Namboodiri VP. Differential attention for visual question answering. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7680–7688.

91 Andreas J, Rohrbach M, Darrell T, *et al*. Neural module networks. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 39–48.

92 Kumar A, Irsoy O, Ondruska P, *et al*. Ask me anything: Dynamic memory networks for natural language processing. Proceedings of the 33nd International Conference on Machine Learning. New York: JMLR.org, 2016. 1378–1387.

93 Xiong CM, Merity S, Socher R. Dynamic memory networks for visual and textual question answering. Proceedings of the 33rd International Conference on Machine Learning. New York: JMLR.org, 2016. 2397–2406.

94 Guo DL, Tao DC. Learning compositional representation for few-shot visual question answering. arXiv:2102.10575, 2021.

95 Han YD, Yin JH, Wu JL, *et al*. Semantic-aware modular capsule routing for visual question answering. arXiv:2207.10404, 2022.

96 Wang P, Wu Q, Shen CH, *et al*. Explicit knowledge-based reasoning for visual question answering. Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne: AAAI Press, 2017. 1290–1296.

97 Wang P, Wu Q, Shen CH, *et al*. FVQA: Fact-based visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(10): 2413–2427. [doi: 10.1109/TPAMI.2017.2754246]

98 Narasimhan M, Lazebnik S, Schwing AG. Out of the box:

Reasoning with graph convolution nets for factual visual question answering. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 2659–2670.

99 Marino K, Rastegari M, Farhadi A, *et al*. OK-VQA: A visual question answering benchmark requiring external knowledge. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020. 3190–3199.

100 Zhu ZH, Yu J, Wang YJ, *et al*. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama: Unknown publishers, 2021. 1097–1103.

101 Gardères F, Ziaeefard M, Abeloos B, *et al*. ConceptBert: Concept-aware representation for visual question answering. Proceedings of the 2020 Findings of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 489–498.

102 Chen Z, Chen JY, Geng YX, *et al*. Zero-shot visual question answering using knowledge graph. Proceedings of the 20th International Semantic Web Conference. Springer, 2021. 146–162.

103 Ding Y, Yu J, Liu B, *et al*. MuKEA: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 5079–5088.

104 Gao JY, Ge RZ, Chen K, *et al*. Motion-appearance co-memory networks for video question answering. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6576–6585.

105 Sun GL, Liang LL, Li TL, *et al*. Video question answering: A survey of models and datasets. Mobile Networks and Applications, 2021, 26(5): 1904–1937. [doi: 10.1007/s11036-020-01730-0]

106 Zhao Z, Lin JH, Jiang XH, *et al*. Video question answering via hierarchical dual-level attention network learning. Proceedings of the 25th ACM International Conference on Multimedia. Mountain: ACM, 2017. 1050–1058.

107 Liang JW, Jiang L, Cao LL, *et al*. Focal visual-text attention for visual question answering. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6135–6143.

108 Li XP, Gao LL, Wang XH, *et al*. Learnable aggregating net with diversity learning for video question answering. Proceedings of the 27th ACM International Conference on Multimedia. Nice: ACM, 2019. 1166–1174.

109 Gao DF, Wang RP, Bai ZY, *et al*. Env-QA: A video question answering benchmark for comprehensive understanding of dynamic environments. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 1655–1665.

110 Ge YY, Xu YJ, Han YH. Video question answering using a forget memory network. Proceedings of the 2nd CCF Chinese Conference on Computer Vision. Tianjin: Springer, 2017. 404–415.

111 Kim J, Ma M, Kim K, *et al*. Progressive attention memory network for movie story question answering. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 8329–8338.

(校对责编: 孙君艳)