计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(2): 314−335, 2025

基于信息检索的知识库问答综述

田 萱 吴志超
(北京林业大学信息学院 北京 100083)
(国家林业草原林业智能信息处理工程技术研究中心 北京 100083)
(tianxuan@bjfu.edu.cn)

Review of Knowledge Base Question Answering Based on Information Retrieval

Tian Xuan and Wu Zhichao
(*School of Information Science and Technology*, *Beijing Forestry University*, *Beijing* 100083)
(*Engineering Research Center for Forestry-oriented Intelligent Information Processing of National Forestry and Grassland Administration*, *Beijing*, 100083)

Abstract Knowledge base question answering is aimed to retrieval relevant information from the knowledge base for model inference, and return accurate answers. In recent years, with the development of deep learning and large language models, knowledge base question answering based on information retrieval has become the research focus, and many novel research methods have emerged. We summarize and analyze the methods of knowledge base question answering based on information retrieval from different aspects such as model methods and datasets. Firstly, we introduce the research significance and related definitions of knowledge base question answering. Then, according to the model processing stages, we explain the key problems and typical solutions faced in each stage from four stages: question parsing, information retrieval, model inference, and answer generation, and summarize the common network modules used in each stage. Then we analyze and sort out the inexplicability of knowledge base question answering based on information retrieval methods. In addition, relevant datasets with different characteristics and baseline models at different stages are classified and summarized. Finally, the summary and outlook are provided on each stage of knowledge base question answering based on information retrieval, as well as the overall development direction of the field.

Key words knowledge base question answering; information retrieval; deep learning; large language models; stage issues

摘 要 知识库问答旨在从知识库中检索相关信息用于模型推理,最终返回准确的答案.近年来随着深度学习和大语言模型的发展,基于信息检索的知识库问答研究成为焦点,涌现出许多新颖方法.从模型方法、数据集等不同方面对基于信息检索的知识库问答研究进行梳理总结.首先对知识库问答的研究意义和相关定义进行介绍.然后按照模型执行过程从问句解析、信息检索、模型推理、答案生成这4个阶段阐述每个阶段面临的关键问题以及典型解决方法,对每个阶段所使用到的共性网络模块进行总结.其次针对基于信息检索的知识库问答方法的不可解释性进行分析梳理.此外,对不同特点的相关数据集和不同阶段的基线模型进行了分类介绍与总结.最后对基于信息检索的知识库问答每个执行阶段以及该领域整体发展方向进行了总结和展望.

关键词 知识库问答;信息检索;深度学习;大语言模型;阶段性问题

中图法分类号 TP391

收稿日期: 2023-12-14; **修回日期**: 2024-06-19

DOI: 10.7544/issn1000-1239.202331013 **CSTR:** 32373.14.issn1000-1239.202331013

问答系统是一种基于自然语言处理和人工智能技术的软件系统，能够根据用户的提问返回准确而简洁的答案. 按照答案来源的不同，问答系统可分为3类：基于搜索引擎的问答系统、社区问答系统和知识库问答系统.

随着目前大规模知识库的发展，知识库问答系统因具有强大知识推理能力得到了研究者的关注和广泛的应用. 文献[1]将知识库问答研究划分为基于语义分析的知识库问答和基于信息检索的知识库问答2类. 基于语义分析的知识库问答方法首先将问句表示成符号逻辑表达式，然后在知识库上执行逻辑表达式获得最终答案，具有可解释性强、推理依据明确的特点，但过于依赖问句的逻辑解析模块，在逻辑表达式或查询图的制作上开销较大；而基于信息检索的知识库问答方法则是根据问句在知识库中检索出相关信息进行推理从而获得最终答案. 与基于语义分析知识库问答方法相比，基于信息检索知识库问答不需要人工制作逻辑表达式或查询图，该方法根据用户的问句经过模型推理即可得到答案，适用于端到端的训练方式，节约人力资源，因此其成为当前知识库问答领域的一个研究焦点.

随着深度学习和大语言模型技术的发展，越来越多的研究者将端到端的深度学习框架应用到基于信息检索的知识库问答领域，取得了显著成果. 本文梳理总结了基于信息检索的知识库问答领域相关研究，一方面从模型执行阶段性入手，另一方面从模型可解释性角度入手，分别介绍目前该领域各个阶段遇到的挑战和解决方法，如图1所示. 希望为后续研究者提供一定参考和帮助.

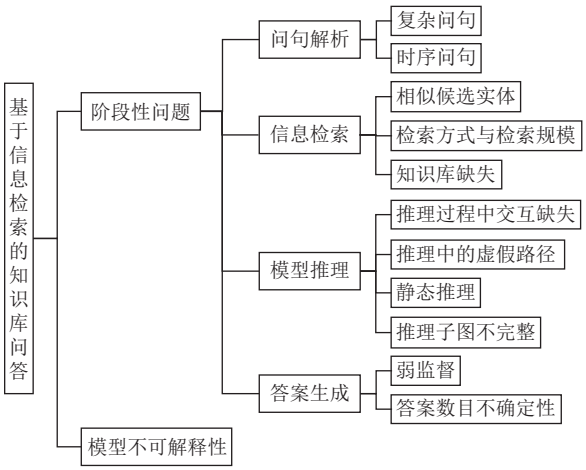


Fig. 1 The framework of survey

图1 综述框架

1 知识库问答相关定义

知识库问答旨在通过分析存储在知识库中大量丰富的实体以及实体与实体之间的关系来回答自然语言问句，如图2所示.

其中，知识库问答过程中常见的概念如下：

1）知识库（knowledge base, KB）. 知识库是用于知识管理的特殊数据库，通常由大量三元组构成，表示为$\mathcal{G}=\{\langle e_1,r,e_2\rangle|e_1,e_2\in E,r\in R\}$，其中$\langle e_1,r,e_2\rangle$表示头实体$e_1$和尾实体$e_2$之间存在关系$r$，集合$E$和集合$R$分别表示实体集和关系集.

2）知识库问答（knowledge base question answering,

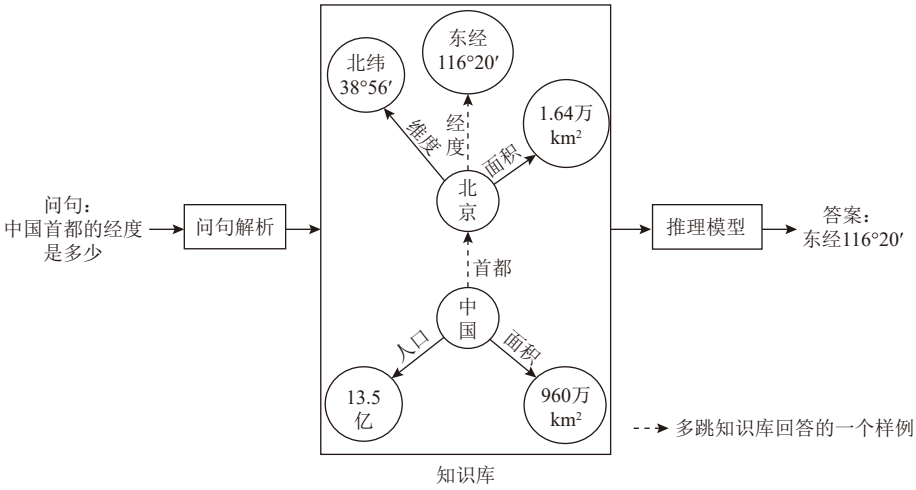


Fig. 2 The example of knowledge base question answering

图2 知识库问答示例

KBQA). 给定一个问句$q=\{q_1,q_2,\cdots,q_m\}$和一个知识库$\mathcal{G}$, KBQA的任务是根据问句q从知识库$\mathcal{G}$中获取答案实体.

3)单跳知识库问答(single-hop KBQA). 在知识库上进行1次推理, 就可以得到答案的过程.

4)多跳知识库问答(multi-hop KBQA). 相比于单跳知识库问答, 多跳知识库问答是指在知识库上需要进行$N(N>1)$跳推理才能得到答案. 多跳知识库问答需要在包含大量知识的知识库中找到多个有关联的三元组, 然后进行逐步推理, 直到找到正确答案. 例如图2中问句"中国首都的经度是多少", 这个问句涉及的关系有〈首都〉和〈经度〉, 需要三元组〈中国, 首都, 北京〉和三元组〈北京, 经度, 116°20′〉才能得到正确的答案.

5)时序问答(temporal question answering). 问句中包含时间因素或者隐藏时间因素的问答称为时序问答. 由于问句中存在时间限制作为答案约束, 时序问答的推理过程相对于单跳、多跳知识库问答更加困难.

基于信息检索的知识库问答主要是根据问句检索出相关信息进行推理获得最终答案, 其执行流程一般如图3所示: 首先识别出问句实体/特征, 然后将问句实体/特征链接到知识库获取问句的相关信息, 再利用模型根据检索信息进行推理, 最后通过排序模型或分类模型筛选出最终答案.

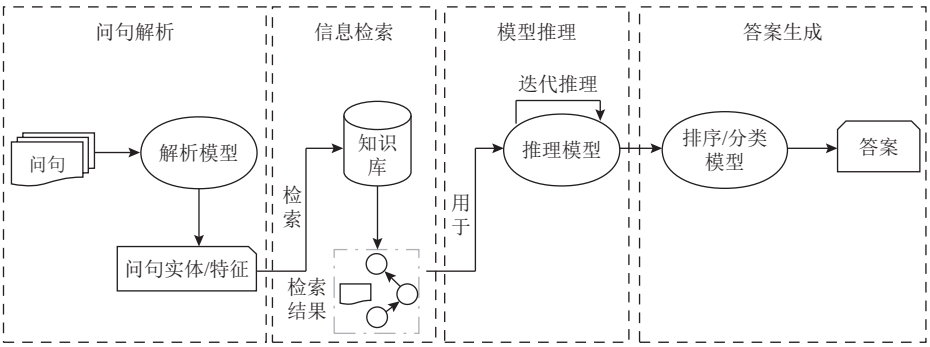


Fig. 3 The flowchart of knowledge base question answering based on information retrieval

图3 基于信息检索的知识库问答流程图

2 阶段性方法分析

参考Lan等人[1]的分析思路, 本文将基于信息检索的知识库问答过程划分为4个阶段: 问句解析、信息检索、模型推理和答案生成. 其中问句解析面临问句复杂难以解析、问句中存在时序约束等问题; 信息检索面临问句实体与知识库实体链接不对应、知识库中存在大量相似实体、如何控制检索子图的规模等问题; 模型推理面临推理信息缺少交互、推理过程中存在虚假路径、静态推理、推理子图不完整等问题; 答案生成面临弱监督、答案数目不确定等问题. 下面将对这些阶段性问题及解决方式进行分析、归纳和梳理.

2.1 问句解析

问句解析的目的是对用户提出的问句进行分析, 提取出问句主题实体或特征用于信息检索阶段. 面对用户多样性的问句形式, 问句解析阶段常存在着问句包含多个实体难以解析、问句中存在歧义实体、问句中包含噪声、问句存在时序性等问题和挑战.

2.1.1 复杂问句解析

当问句中包含多个实体、存在歧义实体或包含噪音时, 往往会影响问句主题实体的识别, 我们将这类问句归纳为复杂问句. 表1针对复杂问句解析的难点和相应方法进行总结归纳.

针对问句涉及多个实体干扰的情况, Xie等人[2]提出一个基于卷积的主题实体抽取模型(convolution based topic entity extraction model, CTEEM)来识别问

Table 1 Summary of Difficulties and Methods in Analyzing Complex Questions

表1 复杂问句解析难点与方法总结

难点	方法特点	代表模型
问句中存在多个干扰实体	捕捉问句中每个词更丰富的上下文特征, 更准确识别出问句主题实体	CTEEM
	使用大语言模型对问句进行凝练和压缩, 保留其中的关键信息	SBP
问句中存在歧义实体	利用歧义实体的相关信息来实现实体消歧	KaFSP
	引导大语言模型生成歧义问句的额外信息, 进而重写问句	query2doc
问句存在噪声	使用结果对问句中的噪声进行改进和纠错	VRN
问句中含有词语省略	捕获词的上下文信息从中提取语义相关特征, 从而弥补词语省略造成的信息缺失	AM-BI-LSTM
问句包含多个主题实体和关系	使用大语言模型对问句进行分解	self-ask, EIC

句主题实体以构建推理子图. CTEEM 首先将问句的单词分为属于主题实体和不属于主题实体 2 类标签，然后使用卷积操作捕捉问句中每个词的上下文特征，根据每个词的标签和上下文特征训练模型以识别该词是否为问句的主题实体. CTEEM 通过捕捉问句中每个词更丰富的上下文特征提升了复杂问句中的主题实体抽取的能力. CTEEM 不需要任何特征工程和额外的自然语言处理工具如词性标注等，就可在知识库问答上获得较高准确率. 然而，当问句中一个主题实体链接到知识库中多个实体时，该模型难以区分出最匹配问句实体，造成问答准确率严重下降. Zheng 等人[3] 提出了一个回步提示（step-back prompting，SBP）方法. 该方法首先会提示大语言模型对复杂问句进行凝练和压缩，保留其关键的实体和约束条件；然后在其基础之上对问句进行后续检索和推理，缓解了问句中干扰实体对问答准确率产生的影响.

在主题实体识别过程中，往往会有部分实体存在一词多义等歧义情况，导致问句解析出错. 针对该问题，Li 等人[4] 提出了 KaFSP（knowledge-aware fuzzy semantic parsing）模型，利用知识库中歧义实体的相关信息来实现消歧. KaFSP 模型图如图 4 所示，根据歧义实体的节点表示和相关信息以及歧义实体在问句中出现的上下文建模一个以 XLNet 为分类器的二元分类任务，实现实体消歧，提升主题实体识别准确率. 该方法在 CommonsenseQA 数据集的 8 种类型问句上的问答准确率优于基线模型，并取得了新的 SOTA（state-of-the-art）结果. 为了消除问句中的歧义信息，Wang 等人[5] 提出了一种查询扩展方法 Query2doc，该方法首先通过提示词使大语言模型生成歧义问句的相关信息，然后使用生成的信息对问句进行扩展说明，最后使用大语言模型根据扩展的问句对其进行重写，实现对歧义问句进行消歧. MSMARCO 和 TREC DL 数据集上的实验表明，通过提示大语言模型生成相关的信息有助于问句消歧以及提高后续检索的准确性.

在识别问句主题实体过程中，文本中的拼写错误等带来的噪声也会干扰问句解析. 为了降低问句解析中的噪声影响，Zhang 等人[6] 提出了一种基于知识图的变分推理网络（variational reasoning network，VRN）. 如图 5 所示，VRN 将主题实体识别与知识模型推理模块以端到端方式融合起来，把训练过程中的损失直接反馈到主题实体识别模块，有助于更准确地在噪声环境中识别出正确主题实体，提升问答准确率.

问句中词语省略带来的上下文不完整也会影响主题实体的识别. 针对该问题，Yu 等人[7] 提出一种注意力机制与多粒度分词相结合的模型 AM-BI-LSTM（attention mechanism and multi-granularity-based Bi-LSTM model）. 为减少不完整上下文导致的信息缺失，该模型在汉字级、词级和短语级进行多粒度分词，并利用基于注意力机制的 BiLSTM 捕获词的上下文从中提取语义相关特征来支撑主题实体识别. 该模型能够在问句表达不完整情况下更好地识别主题实体，提高了最终问答的准确率.

当问句中含有多个主题实体和多个关系时，需要多跳推理才能得到准确答案，多跳问句回答准确

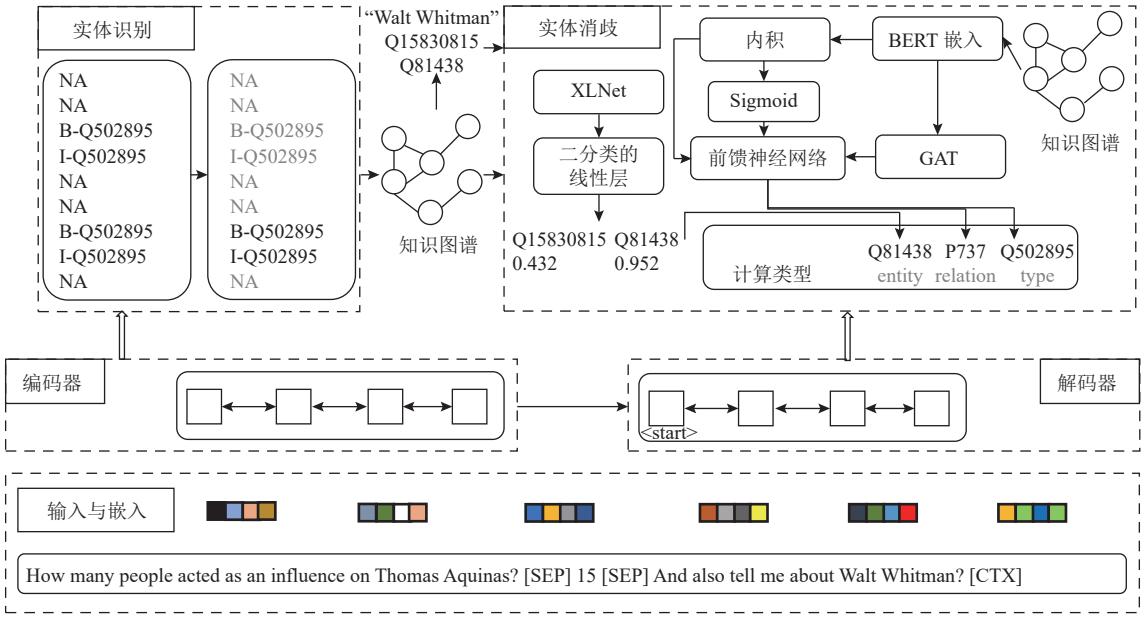


Fig. 4 KaFSP model diagram

图 4 KaFSP 模型图

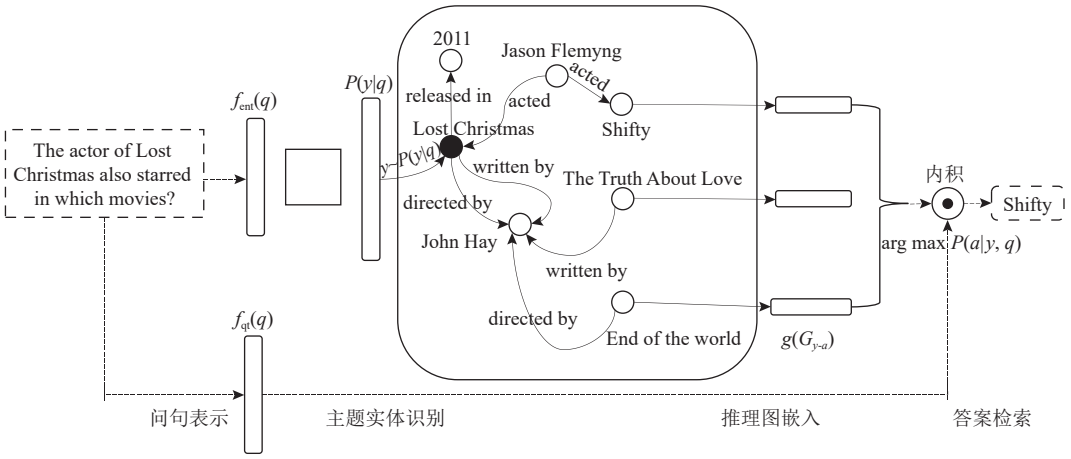


Fig. 5 VRN model diagram

图 5 VRN 模型图

率相比于单跳问句较低. 为了提高多跳问句回答的准确率, Press 等人[8] 采取分解问句的思想提出了 self-ask 的方法. 该方法采取思维链(chain of thought, CoT)提示大语言模型将复杂问句分解为回答准确率更高的子问句(单跳问句), 并在得出最终答案之前回答这些子问句, 以辅助最终问句的准确回答. WikiMultihopQA 数据集上的实验表明, 该方法在回答复杂问句方面的准确率明显提高, 并且随着模型规模增加, 准确率提高效果更显著. 与之类似, Haji 等人[9] 提出了探索性推理链的框架 EIC(exploratory inference chain), 其结合了大语言模型的生成能力和显示推理链. EIC 首先会使用大语言模型生成的问句所需信息作为关键词; 然后为每个关键词执行单跳推理, 若推理错误或者不充分, 则执行附加推理; 重复上述过程, 当获得足够信息后将它们进行聚合以提示大语言模型生成最终答案. WikiMultihopQA 数据集中 2 跳问句的实验结果表明, EIC 的准确率超过了现有基线模型, 说明基于大语言模型的复杂问句拆解有助于问答准确率的提升.

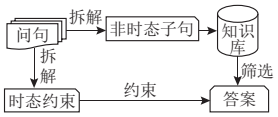
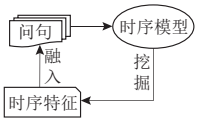
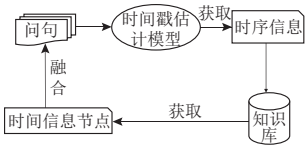
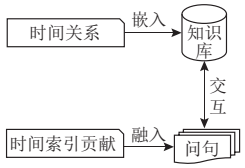
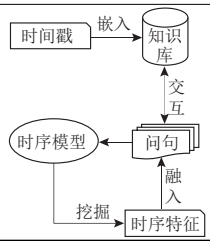
2.1.2 时序问句解析

时序问答中, 问句的答案通常与时间因素密切相关, 即随着时间推移问句的答案会不断发生变化, 本文将其称为时序问句. 时序问句解析的关键不在于找出问句中的主题实体, 而是在于如何解析出问句中的时序特征, 以提高答案准确性. 本文对时序问句解析的研究方法总结与对比如表 2 所示.

针对时序问句, Jia 等人[10] 提出了一种名为 TEQUILA 的时序问句分解模型. 模型首先检测问句是否具有时序性, 若有则将问句分解并重写为非时态子问句和时态约束, 之后从知识库中检索出子问句答案, 最后使用时态约束来选择整个问句的最终答案. TEQUILA 模型既具备处理时序问句的能力又具备处理复杂问句的能力, 在时序数据集 TempQuestions 和多跳问答

Table 2 Summary and Comparison of Methods for Analyzing Sequential Questions

表 2 时序问句解析方法总结与对比

方法特点	模型	优点	缺点	方法流程
分解问句	TEQUILA	可将时态问句重写为非时态问句和时态约束, 降低了时态问句回答的复杂度	重写问句会出现错误标记, 影响准确率	
增强问句时序特征	EXAQT	可增强问句中时间信号, 提高准确率	需要手工标注问句特征词, 费时费力	
结合知识库增强问句时序特征	TSQA	可推断出时序问句中未写的时间戳, 提高了准确率	缺乏扩展性	
知识库与问句交互增强问句时序特征	TERQA	通过与知识库交互, 在问句中添加了时间特征, 提高了时态问句回答的准确率	挖掘交互特征会带来大量的开销和计算量	
	CTRN	既使用模型挖掘问句中的时间特征, 又充分利用知识库中的时间特征	模型计算参数量较大	

数据集 ComplexWebQuestions 上具有较好的评测结果. 然而, 很多时序问句中并没有确切的时间信号, 采用上述拆分问句的形式难以保障答案准确率. 针对此类问题, Jia 等人[11] 设计了名为 EXAQT(explainable answering of complex questions with temporal)的模型, 采用手动标记的特征如时间感知实体嵌入、时间戳、时间信号和时间问句类别编码等与问句进行拼接融合, 增强了时态问句的时间特征. 该模型在时序数据集 TimeQuestions 上的测评结果优于 GraftNet, PullNet 等基线模型, 说明对时序问句时间特征的分析有助于提高问答结果准确性.

文献 [10−11] 的研究主要集中在如何分析时态问句上, 缺少对知识库中时间信息的应用. Shang 等人[12] 提出了一个时间敏感问答(time-sensitive question answering, TSQA)模型, 设计出一个时间戳估计模块用于推断问句中不确定的时间信息. 该模块首先把问句实体和关系映射到向量空间进行表示, 然后从知识库中搜索与这些实体和关系对应的时间节点融入到问句中. 数据集 Cronquestions 上的测试结果表明 TSQA 模型的 *Hit*@1 指标远超 EmbedKGQA 等基线模型, 说明该模型的时间戳估计模块可有效解决时序问句中不明确时间信息导致的答案准确率较低问题.

文献 [12] 从知识库中提取时间信息增强了问句的时序特征, 但忽略了问句表示中时态信息与知识库之间的隐藏特征, 使得模型不能很好捕捉到问句中的时间信息, 导致时态问答的回答准确率降低, 限制了时序问答研究的发展. 为此 Yao 等人[13] 提出了一种 TERQA 模型, 用于探索问句中时间信息与问句表示和知识库节点表示间的关系. 如图 6 所示, 首先定义了 13 类时间关系, 将它们嵌入到知识库的实体和关系表示中; 然后在问句表示中融入时间索引贡献(即问句特定位置处出现时间信息的概率); 如果问句特定位置与知识库中检索到的时间信息节点具有较高相似度, 则提升问句在该位置的时间索引贡献度, 以获取问句和知识库之间的时间交互特征. TERQA 模型在数据集 TempQuestions 上的问答正确率(*P*@1)比 EXAQT 基线模型提高了 11.3%, 表明通过挖掘问句与知识库的时间交互特征有助于提高最终问答的准确率. Jiao 等人[14] 提出了一个复杂的时间推理网络 CTRN, CTRN 首先采用时间图卷积网络捕获问句中隐含的时间特征, 然后将带时间特征的问句表示向量与问句涉及的知识库节点嵌入求和以增强问句表示, 最后使用门控机制来确定应该保留的时间特征比例. Cronquestions 数据集上的实验结果表明, 该方法在挖掘问句与知识库之间交互特征方面取得了不错的效果.

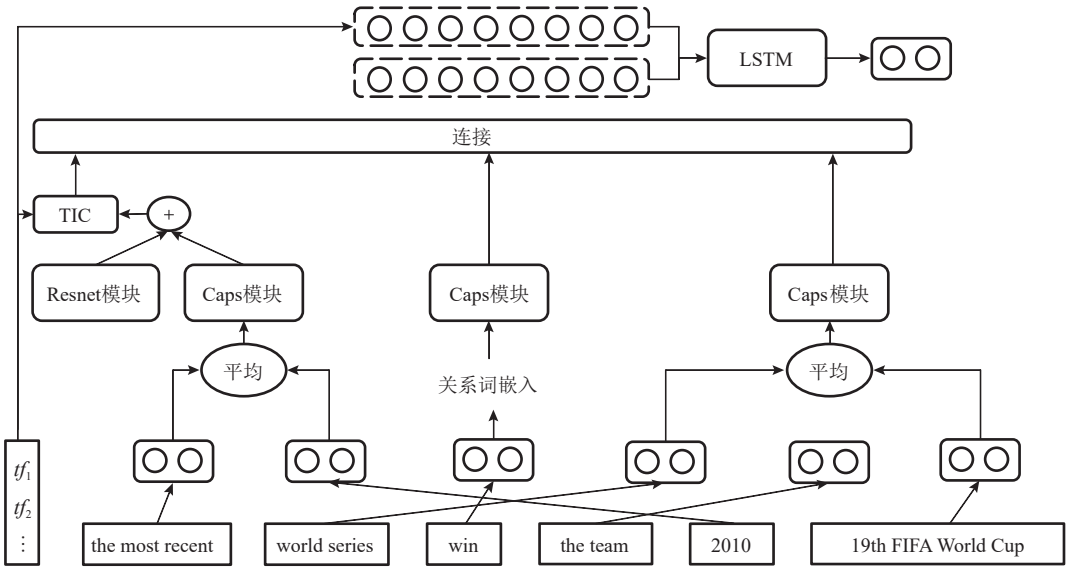


Fig. 6 TERQA model diagram

图 6 TERQA 模型图

2.2 信息检索

信息检索的目的是根据问句在知识库中检索相关信息, 将主题实体链接到知识库上进行检索, 检索出所有实体节点及其之间的关系作为推理知识, 以备后续推理出答案. 在实体链接过程中, 可能会出现主题实体在知识库中存在相似候选实体的问题; 在检索过程中, 可能会遇到检索规模过大或过小影响模型计算量或答案准确率的问题. 下文将对解决这

些问题的方法进行总结.

2.2.1 实体链接及相似候选实体问题

实体链接旨在将问句解析出的主题实体链接到知识库的实体上. 该过程面临的最大问题是问句中1个实体链接到知识库中时可能会存在多个相似候选实体, 这些实体表达相似而语义不同, 导致推理子图构建的合理性和有效性受到一定影响, 影响最终回答的准确性.

针对此问题, Wang 等人[15]提出了一种新的实体链接框架以提升问答准确率. 该框架首先比较相似候选实体间的差异特征和交互特征, 提取出关键性差异; 然后利用这些关键性差异对实体进行特征更新; 最后使用更新后的实体再次执行实体链接, 提高实体链接的准确性, 以便后续推理答案. 该框架在WebQuestion数据集上的召回率和*F*1指标均高于基线模型. 针对类似问题, Luo 等人[16]提出了一种单关系问答(single relation question answering, SRQA)方法, 该方法采用预训练的BERT和BiLSTM来降低链接过程中候选相似实体带来的噪声. SRQA首先使用基于BERT的序列标注模型识别问句中的实体词; 然后使用启发式算法将其链接到知识库实体节点, 并采用双向的LSTM对其中的候选相似实体节点提取上下文特征; 最后通过计算问句与这些候选相似实体节点的相似度计算筛选出与问句最相关的链接实体. 该方法在SimpleQuestion数据集上的问答测评准确率达到了80.9%, 说明降低链接过程中候选相似实体的噪声有助于提升问答准确性.

为了在链接实体时更好利用知识库中的关系特征区分相似候选实体, Cao 等人[17]提出了一种名为QAVGAE(question answering oriented variational graph auto-encoder)的模型, 来集成实体周围的关系特征提升对相似实体的区分. 该模型提出了一种基于图卷积网络的变分图自动编码器, 来学习实体周围的关系特征和计算关系权重, 并用于实体节点的聚合表示, 以提高候选相似实体间的区分度. WebQuestionSP和ComplexWebQuestion这2个数据集上的实验表明该模型的*Hits*@1和*F*1比基线模型均有所提升, 有助于增强知识库问答效果. 与之类似, Shamsabadi 等人[18]提出把相似候选实体周围的关系与问句中的关系相结合, 实现对相似候选实体的筛选, 也能够很大程度缓解知识库问答中候选实体相似带来的干扰.

2.2.2 检索方式及检索结果优化

检索的目的是依据问句中的主题实体及其他关键实体在知识库上检索出相关实体及其之间的关系组成推理子图或者推理知识以用于后续推理任务. 检索方式可以分为3种: 第1种是根据问句解析出主题实体在知识库中的*N*跳节点及其之间的关系, 再对其进行优化; 第2种是选择问句与知识库节点之间相关性程度高的实体节点, 再对其进行优化; 第3种是利用大语言模型辅助检索. 表3对以上3种检索方式进行了总结与对比.

Table 3 Summary and Comparison of Different Retrieval Methods

表3 不同检索方式的总结与对比

检索方式	模型	优点	缺点
*N*跳邻居检索	CLOCQ	使用知识库感知信号搜索空间, 修剪不相关部分	修剪时可能会遗漏答案实体, 影响准确率
	BAMnet	可以挖掘知识库实体更多相关信息, 检索更相关实体, 控制了检索规模	没有很好的利用答案信息, 影响准确率
	DAG	检索结果不需要进行裁剪工作	检索信息会存在缺失
问句相关性检索	DCRN	采取广度优先搜索算法进行剪枝, 减少了无关节点的数量	可能会排除与答案相关节点
	RASR	结合实体之间的关系信息, 控制了检索规模	模型复杂度高
大语言模型辅助检索	SKR	使大语言模型判断模型时具备回答能力以此引导检索过程	模型参数量大
	CON	使用大语言模型来生成问句相关信息, 用于检索更相关实体以及关系	模型需要内存较高

文献[19−21]提出使用问句主题实体在知识库中的*N*跳邻居节点及其之间的关系组成推理子图; 文献[22−23]提出使用问句与知识库节点之间的相关性进行检索. 其中这2种方式检索出的子图可能会面临着规模问题: 子图过小可能不包含答案, 子图过大则可能会引入许多与问题无关的实体节点, 增加模型复杂度, 导致效率降低. 因此, 现有文献大都首先尽可能检索更多的节点再进行子图优化以提升推理效率.

文献[19−21]选取问句主题实体在知识库中的*N*跳邻居节点及其之间的关系作为推理子图. Hao 等人[19]提出的有向无环图编码方式DAG(directed acyclic graph). 该模型首先识别给定问句中涉及的所有实体, 并分别链接知识库; 然后选取2跳范围内所有的实体节点的子图来构成候选DAG, 筛选出与问句具有top-*k*相似度的子图构成候选DAG集; 最后分别计算候选DAG集与问句的注意力权重, 选取与问句最相关的子图. 但是上述方式在面对多跳问句时会遗漏必要信息, 因此有必要检索更多的信息再对其进行优化. 针对检索出子图规模过大的优化问题, Christmann 等人[20]提出了一种CLOCQ(contracting answer spaces

with lists and top-*k* operators for complex QA)模型来对子图中不相关的部分进行裁剪. 该模型构建了一个结合问句相关性、候选实体一致性和在子图中连通性等信息的查询处理器, 通过该查询处理器从子图上筛选与问句相关的实体节点, 以获得规模更小的推理子图. CronQuestions 数据上的实验结果表明, CLOCQ 模型在答案存在性、检索规模和运行时间方面均优于 GraftNet 等基线模型, 验证了该方法在优化推理子图方面的有效性. 与之类似, Chen 等人[21] 利用知识库中实体节点的特征如实体类型、实体到问句主题实体的关系路径和实体上下文设计了一种名为 BAMnet(bidirectional attentive memory network)的新型双向注意力记忆网络来剪枝子图. BAMnet 直接对问句和子图之间的双向交互流进行建模: 根据子图关注问句的不同部分, 同时根据问句提炼最相关的子图实体. WebQuestions 数据集上的实验表明该模型比基线模型提高了 *F*1 指标, 验证了问句与知识库实体之间双向交互建模的有效性.

文献 [22−23] 采取了根据问句语义从知识库中筛选相关节点及其之间关系构成推理子图的方式. 针对该方式检索出的子图规模过大问题, Cai 等人[22] 设计了 DCRN(deep cognitive reasoning network)模型, 应用广度优先搜索 (BFS) 算法对检索出的子图进行剪枝. 其具体做法是从问句主题实体开始执行广度优先搜索, 在搜索过程中只保留已访问的实体节点, 并去除未访问的实体节点, 从而减少了无关实体的数量, 控制了检索的规模, 减少了模型复杂度. Luo 等人[23] 提出了一个关系感知信息检索 RASR(relation-aware subgraph retrieval)方法来构造一个规模小但答案存在率和覆盖率高的子图. 该方法首先计算问句实体在知识库中与其所有相邻关系的相似度; 然后依据相似度进行筛选, 由这些关系连接的实体构成一个 1 跳子图; 再继续对这个 1 跳子图进行上述相同操作; 最终每个关系的权值被聚合传递到实体节点, 权值较高的实体节点及它们之间的关系构成推理子图. 数据集 ComplexWebquestion 和 WebquestionSP 上的实验结果表明, 模型的 *Hits*@1 和 *F*1 均高于 GraftNet 和 PullNet 等基线模型, 表明该方法采取的子图优化在问答准确率方面取得了显著的效果.

随着大语言模型技术发展, 越来越多的推理过程中采用了大语言模型, 很多情况下模型可以越过检索过程就直接返回正确答案. 然而在缺乏必要知识时仍然需要检索才能返回正确答案, 检索过程中可以辅以大语言模型. Wang 等人[24] 提出了一种自我知识引导的检索方法 SKR(self-knowledge retrieval)如图 7 所示. 该方法首先会引导 InstructGPT 等大语言模型识别其是否具有回答该问句的能力. 若具有回答能力则由大语言模型直接返回答案; 若不具有回答能力则引导模型在知识库中进行检索, 以增强其推理能力. CommonsenseQA 数据集上的实验表明, 利用大语言模型指导检索过程有助于提升问答准确率. 与其类似, Yu 等人[25] 提出了检索验证模型 CON(chain-of-noting). 该模型同样会首先检查回答该问句是否需要检索外部知识, 但与之不同的是, 当需要检索外部知识时, CON 模型会利用 LLaMA2 模型为检索到的结果生成连续解释和说明, 从而深入评估检索结果与问句的相关程度, 选取相关程度高的检索信息以更好推理得出最终答案. WebQuestion, TriviaQA 数据集上的实验表明, CON 模型从外部知识库中检索信息的噪声明显减少, 其回答准确率也显著增加.

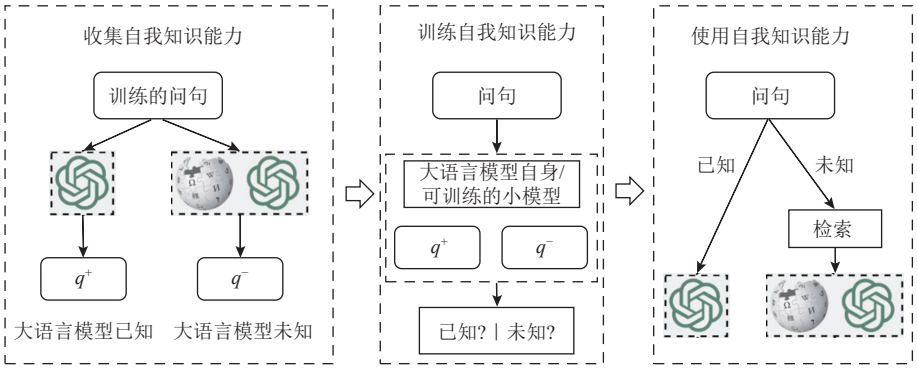


Fig. 7 SKR model diagram

图 7 SKR 模型图

2.2.3 知识库缺失问题

在检索过程中, 知识库的缺失可能会导致检索出的推理子图难以包含与最终答案足够相关的信息, 给后续有效推理带来了挑战. 因此如何利用缺失的

知识库检索出最终答案成为基于信息检索的知识库问答研究的一个焦点. 目前研究主要可分为 2 类方法: 一类是利用文本语料库和知识库共同检索相关信息; 另一类是通过模型学习补充知识库.

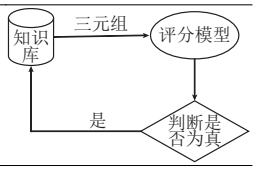
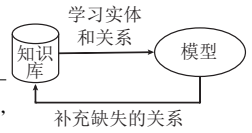
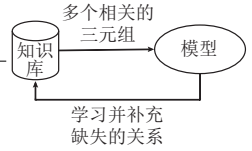
除结构化知识库之外, 大量文本语料库可用于信息检索以构建更为全面的推理子图. Sun 等人[26] 提出了一种名为 GRAFT-Net(graphs of relations among facts and text networks)的模型, 该模型分为 2 个模块: 一个用于检索知识库返回相关实体, 另一个检索语料库返回相关的文本; 然后通过实体链接将知识库中实体与文本进行关联, 即知识库中的实体在某一个文本中出现则该实体与该文本可以相互连接; 最终对检索的知识库实体与文本组成的异构图进行推理得到答案. 在 GRAFT-Net 融合知识库和文本基础上, Sun 等人[27] 提出了一种迭代方法 PullNet. 该方法使用图卷积神经网络识别和筛选语料库及知识库上的信息来迭代构建包含与问句相关的推理子图. 这种迭代方法可有效整合文本语料和知识库实现高效推理; 但会带来很多额外类似答案的实体, 对推理过程产生干扰.

上述工作大都结合知识库和语料库来检索子图, 而 Zhang 等人[28] 则使用特定知识如依赖关系和常识等来结合缺失的知识库, 提出了一个多维知识增强图网络(multidimensional knowledge enhanced graph network, MKGN)模型改进信息检索. MKGN 模型结合知识库进行检索时, 会首先分析知识库缺失了哪方面的特定知识, 如依赖关系和常识知识等, 然后使用这些特定的知识来结合知识库进行回答, 从而缓解了文献[26−27] 带来的噪声实体.

通过语料库或特定知识结合知识库进行检索固然可以解决知识库缺失问题, 但当文本语料库可用性不确定时, 也可能会引入干扰或不相关信息导致模型复杂度和计算量的增加. 当前, 很多研究人员尝试通过模型学习知识库自身特征来实现补充, 以完善缺失的知识库. 表 4 对通过模型学习补充知识库的 3 类方法进行了对比总结. 为补充缺失的知识库, Saxena 等人[29] 提出了一种基于嵌入的知识库问答方法 EmbedKGQA 来预测未知的有效连接. 该方法首先会学习知识库中实体节点和关系的高维嵌入, 使用一个评分函数来评估实体间隐藏的关系是否有效并用以补充缺失知识库, 在一定程度上缓解了知识库缺失问题. 数据集 MetaQA, WebQuestionsSP 的实验表明, 当随机删除知识库部分三元组后, EmbedKGQA 的 *Hit*@1 超过了 GraftNet, PullNet 等基线模型, 证明

Table 4 Comparison of Supplementing Knowledge Base Methods Through Model Learning

表 4 通过模型学习补充知识库方法对比

方法	模型	优点	缺点	方法流程
单个三元组学习	EmbedKGQA	学习知识库中实体和关系的高维嵌入, 补充结果更全面	容易过拟合, 泛化能力较差	
多个三元组学习	TREKBQA	减轻了节点分布偏差的影响	推理过程中监督信号弱可能会产生错误路径	
	CAPKBQA	结合多个三元组进行学习, 补充更加全面	计算量大, 运行时间长	
结合关系与实体进行学习	EIENAGE	同时捕获给定实体周围的实体特征和关系特征	缺乏对实体间高阶关系的分析和利用	
	DRGN	实体间关系聚合表示实体节点, 再动态补充关系	模型涉及参数量较多, 训练时间长	

了缺失知识库上的模型学习有助于提高回答准确率. 然而, EmbedKGQA 只是从单个三元组角度来学习实体和关系, 模型泛化能力较差.

针对知识库不完全性和节点分布的不均匀性, 文献 [30−31] 对文献 [29] 的不足进行了改进, 从路径角度出发进行知识库补充, 实现信息检索. TREKBQA (traversing knowledge graph embedding for multi-hop knowledge base question answering)[30] 是一种基于遍历知识库嵌入空间的知识库问答方法, 其通过对实体节点的邻居节点及其多条路径进行遍历, 对复杂关系进行建模来实现对缺失关系的预测补充, 以获得更完整的嵌入空间用于推理, 减轻了不均匀节点分布的偏差影响. APKGQA 模型[31] 则将所有头实体相同的三元组链接成一个 N 元链接, 然后依据 N 元链接之间的联系实现缺失关系的预测补充. 实验结果表明, 结合多个三元组实体补充知识库比利用单个三元组实体更有利于提高知识库补充效果.

上述研究方法在使用模型补充知识库关系时, 对实体和关系的学习大都独立进行, 无法体现和利用它们隐藏的内在联系. 为了结合邻居实体同时捕获更丰富的实体特征和关系特征, Zhao 等人[32] 提出了一种基于注意力图嵌入的实体重要性评估模型(entity importance estimation network of attention-based graph embedding, EIENAGE). 该模型由基于注意力的图嵌入

模块和实体重要性评估模块组成，基于注意力的图嵌入模块用于从一个实体的 N 跳内邻居及它们的关系上学习该实体的嵌入表示；然后实体重要性评估模块继续在其 N 跳内邻居中寻找重要性较高的实体；最后在两者间引入辅助边，实现关系预测以补充知识库. Zheng 等人[33]提出了一种新的图神经网络结构——动态相关图网络(dynamic relevance graph network, DRGN)，如图 8 所示，利用现有的实体间关系进行聚合来表示实体节点，再根据节点之间的相关性分数动态建立新的实体间关系，进而补充可能缺失的边，从而提高了后续推理回答的准确率. 上述方法补充缺失知识库子图时大都缺乏可解释性. 为了增强其中的可解释性，Cui 等人[34]提出了一种改进的强化学习方法 DCRN(dynamic completion reasoning network)来补充推理子图中缺失的关系. DCRN 根据环境在每个时间步提供的状态信息动态补充几个关系作为强化学习中智能体的额外动作，然后根据获得的奖励值进行调整. 数据集 WebQuestionSP 的测试结果表明，DCRN 在知识库缺失情况下的 *Hit*@1 高于其基线模型，证明了该方法在补充推理子图上的有效性.

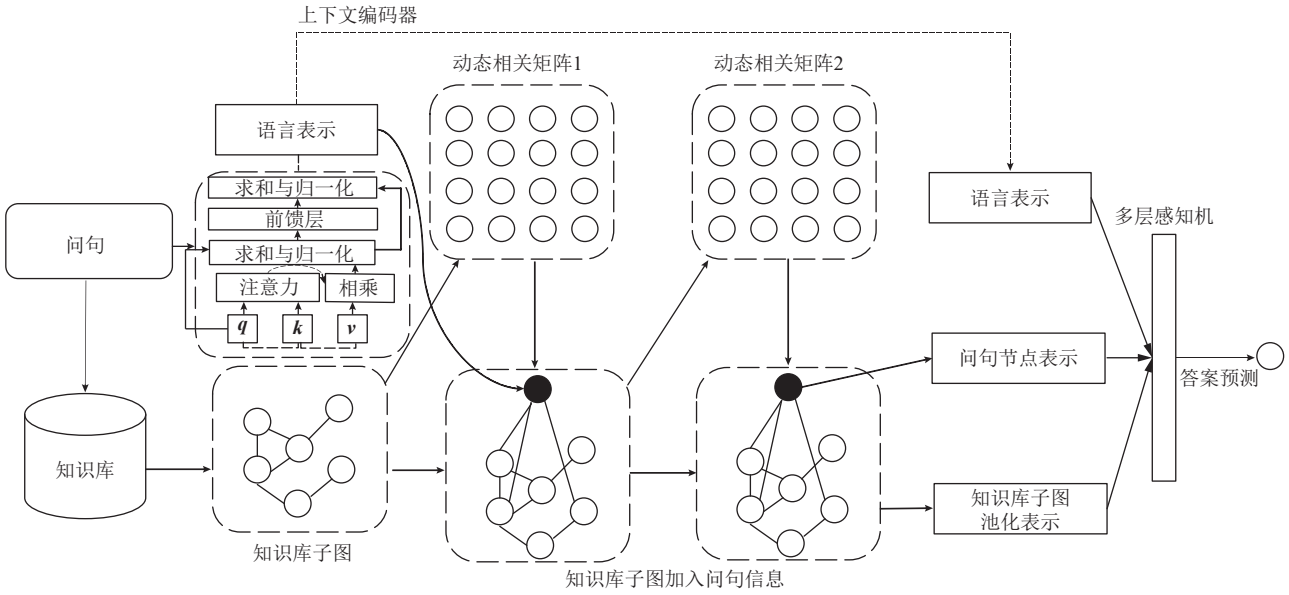


Fig. 8 DRGN model diagram

图 8 DRGN 模型图

2.3 模型推理

模型推理的目的是对信息检索得到的结果进行推理，推理结果能够直接决定问句最终答案生成的准确率. 模型推理中可能会存在推理过程交互缺失、推理中存在虚假路径、推理过程缺乏动态更新、推理子图不完整等挑战，下文主要对这些挑战的解决方法进行总结.

2.3.1 推理过程交互缺失问题

在推理过程中，很多研究大都将问句和推理子图中实体和关系进行单独更新，忽视了问句与推理子图、推理子图内部实体和关系间的交互性，使得模型推理能力下降.

为了结合问句与推理子图实现更好的模型推理，Yasunaga 等人[35]提出了一个 QA-GNN 模型. 如图 9 所示，该模型将问句与候选答案视为一个附加节点并将其连接到推理子图上，然后使用基于注意力的图神经网络更新问句和推理子图上实体节点的表示. 该模型通过问句与推理子图节点在图神经网络中的联合推理，加强了信息间的交互，增强了模型推理能力；但受限于问句表示，难以实现问句与子图节点更细粒度的交互. 在 QA-GNN 模型基础上，Sun 等人[36]提出了一个名为 JointLK(joint reasoning with language models and knowledge graphs)的联合推理模型，采用双向注意力把问句生成的 token 向量与推理子图节点进行融合，使得它们之间的交互性更强，推理效果更好. 同样为了在模型推理过程中获得问句与推理子图更细粒度的交互，Cao 等人[37]提出了一种名为 ReLMKG(reasoning with pre-trained language models and knowledge graphs)的联合推理模型. ReLMKG 将预训练的语言模型和图推理模型分别用于 token 向量和推理子图节点的更新表示，然后将 2 个模块输出的内容连接起来送入 2 层全连接网络获得推理结果. 数据集 WebQuestionSP 上的实验结果表明，ReLMKG 模型的 *Hits*@1 和 *F*1 均高于 KV-Mem, GraftNet 等基

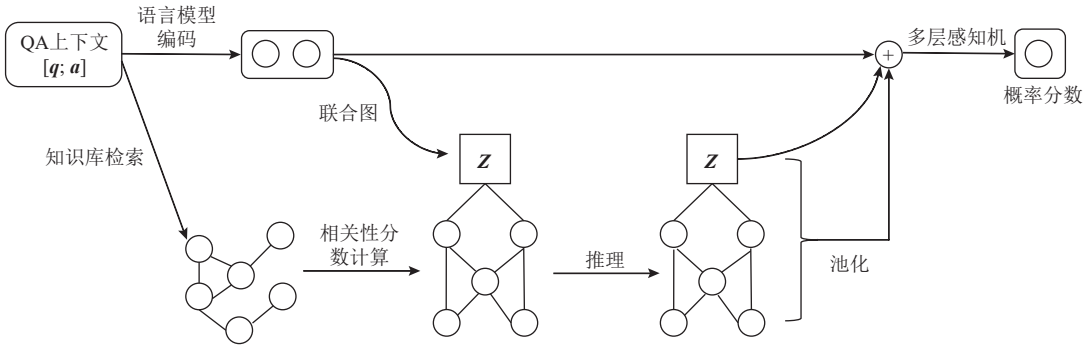


Fig. 9 QA-GNN model diagram

图 9 QA-GNN 模型图

线模型, 验证了问句与子图节点细粒度融合可以提高模型的推理能力, 有助于提升问答准确率.

为了更有效利用推理子图内部实体和关系间的结构交互性提升模型推理效果, Wang 等人[38] 提出了一个融合推理子图实体和关系的模型 RAID(the relation aware of subgraphs and implanting the direction information). 该模型在推理过程中将子图中关系边转化为关系节点融于推理子图中, 将关系节点与实体节点结合推理, 以更充分利用推理过程中的关系信息, 增强模型的推理能力. MetaQA 数据集上实验表明, RAID 模型的 *Hits*@1 相较于 KVMem, GraftNet 等基线模型有所提升, 验证了利用子图实体和关系的交互特征能够提升问答的准确率.

很多研究通过检索知识库相关信息来增强大语言模型的推理能力, 但缺失了检索信息与大语言模型自身知识之间的交互, 这不利于回答复杂的多跳问句. 针对此问题, Feng 等人[39] 提出了一种可以使检索信息与自身知识交互的 ITRG(iterative retrieval-generation)模型. 该模型首先利用 LLaMA 模型的生成能力生成问句相关文本来扩大检索范围以提高检索准确性, 然后利用检索结果结合问句再次生成问句相关文本, 多次迭代上述过程, 直到达到最大迭代次数. WikiMultiHopQA 数据集上的实验表明, ITRG 在多跳问句问答上的准确率有明显提升, 验证了将检索信息与大语言模型自身知识进行交互有利于增强其推理能力.

2.3.2 推理中的虚假路径问题

在模型推理过程中, 由于缺乏对推理路径的有效监督, 有时即使找到了正确答案, 但其路径信息在语义上与问句无关, 这类路径本文称之为虚假路径. 训练过程时, 虚假路径的存在可能会误导模型; 即便得到了正确答案, 该答案的推理过程也缺乏可解释性.

为了解决模型推理过程中这种虚假路径问题, Hu 等人[40] 提出了一种名为 CPR(core path reasoning)模型, 通过考虑问句与推理路径之间的语义关系来缓解虚假路径现象. CPR 模型首先使用深度优先搜索算法找到问句主题实体到答案实体的所有可能路径, 然后通过 2 个阶段筛选正确路径: 在第 1 阶段, 模型过滤掉语义不相关的路径, 选择与问句语义相似的路径; 在第 2 阶段, 模型过滤掉虚假路径, 利用期望最大化迭代选出与问句完全匹配的正确路径. WebQuestionSP 和 ComplexWebQuestions 的实验结果表明, CRP 模型的 *F*1 和 *Hit*@1 比 NSM, GraftNet 等基线模型均有所提高, 尤其是在解决长路径和伪路径问题上 CRP 模型性能表现更好. 与之类似, Niu 等人[41] 提出了一种名为 PKEEQA(path and knowledge embedding-enhanced multi-relational question answering)的模型, 该模型首先使用 LSTM 对路径信息进行编码, 然后选取余弦距离函数计算问句嵌入与路径嵌入之间的相似度作为问答任务的辅助监督信号, 在一定程度上缓解了虚假路径问题, 提高了答案准确性. 在 MetaQA 数据集上的测评结果显示 PKEEQA 模型在 1 跳和 3 跳推理中均比其基线模型有所提升, 表明结合路径信息进行监督问答任务更有助于找出正确答案. Bi 等人[42] 采用强化学习策略实现推理路径的有效监督, 提出了一种 RIPE(reward integration and policy evaluation)的强化学习推理模型. RIPE 通过评估每一步推理中模型所采取的行动, 设计一个即时奖励函数以加强监督, 将即时奖励与问句回答的终端奖励相结合, 引导模型沿着正确的推理路径寻找答案. 该方法可以缓解推理中的虚假路径现象, 有利于提高模型的鲁棒性, 使其在面对复杂问题时能够更好地理解问题以正确推理. 与上述建模正确路径和伪路径不同, Yoran 等人[43] 提出了一种促使大语言模型在多个思维链上进行推理的多链推理(multi chain reasoning, MCR)模型. MCR 模型能够根据不同思维链在不同路径上进行推理, 最终根据投票机制以及答案与问

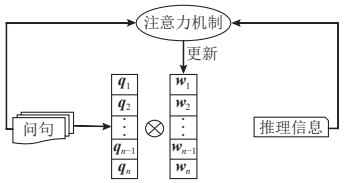
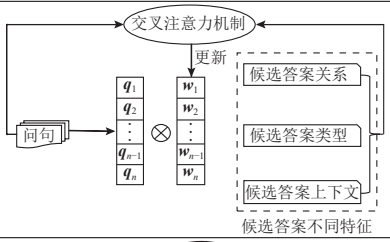
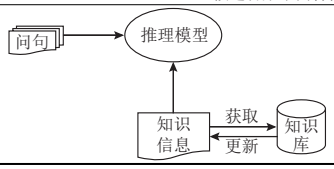
句的相关性选择出合适答案和推理路径. MCR 在解决虚假路径的问题上，能够使用户验证答案和路径，答案可解释性程度更高. 思维链技术可以指导模型多步推理，提高回答的准确率. Park 等人[44]提出了一个图形引导的 CoT 提示方法，引导大语言模型得到正确答案. 该方法首先利用 LLaMA2 构建一个问句推理图（其中包含每步推理所需要的信息），然后在每一步推理时结合问句和上一步骤的推理信息生成当前步骤的推理信息，通过问句推理图验证其是否合理，过滤掉不相关的内容，从而缓解了推理过程中的虚假路径问题. WikiMultihopQA 数据集上的实验结果表明，该方法在问答的 $F1$、准确率和召回率上均有所提升.

2.3.3 静态推理问题

大多推理模型在问句表示方面和子图连接结构上都是静态不变的，即问句向量表示和推理子图连接结构在推理过程中是始终不变的，其并不会随推理过程更新，这限制了模型的推理能力. 近年来，有很多文献将注意力放在了如何在推理过程中动态更新问句表示和推理子图连接结构上. 这种做法可以使模型在不同推理阶段获得与当前阶段更相关的信息，能够更加高效准确地利用信息，增强推理的准确性. 在模型推理过程中，大多数研究对于问句的表示从始至终保持不变，大都将问句作为一个静态向量处理，忽略了推理过程中问句需要更新的问题. 事实上，模型推理的每一次迭代对于问句各个部分的关注度并不同，候选答案的特征不同也会导致问句各部分表示不同，依据这些因素对问句表示进行适时动态更新有助于提升模型推理效果，促进答案更准确生成. 表 5 总结对比了推理中动态更新问句的 3 类方法.

Table 5 Summary and Comparison of Dynamic Update Question Methods in Reasoning

表 5 推理中动态更新问句方法总结与对比

方法	模型	优点	缺点	方法流程
推理迭代过程中的问句动态表示	IRN	根据上一步推理信息更新下一步对问句的关注度	不利于约束较多的问句，问句动态更新效率会较低	
	QA2MN	利用推理中涉及的三元组更新问句的表示，提高推理能力	算术运算的问题回答准确率较低	
	DRN	根据历史推理信息更新问句的表示，提高推理能力	随着推理跳数的增加，准确率逐渐下降	
	KBIGER	融合问句与子图信息对问句进行更新	模型计算量大	
候选答案影响下的问句动态表示	NCAM	使用交叉注意建模不同特征来动态更新问句表示，提高准确率	复杂问题的回答准确率低	
推理知识动态检索和变化	FLARE IR-CoT Self-RAG	根据上一步推理结果动态从知识库中获取相关知识	模型参数量大，相应时间长	

针对问句表示在模型推理每一次迭代过程中的更新问题，Zhou 等人[45]设计了 IRN（interpretable reasoning network）模型. IRN 能够动态决定每一步推理中问句应该被重视的部分，并选择子图上与其相对应的关系和实体，再利用其反向更新问句表示和推理过程状态，然后驱动下一步推理. 数据集 PathQuestion 上的评测结果表明 IRN 模型的 *Hit*@1 超过了 KV-Mem 等基线模型，展示出动态问句表示在推理中的有效性. 采取类似的思路，Li 等人[46]提出了问句感知记忆网络（question-aware memory network, QA2MN）模型. QA2MN 根据每步推理涉及的相关三元组动态更新问句的表示. 在数据集 PathQuestion 和 WorldCup2014 上的测试结果表明，QA2MN 模型的 *Hit*@1 都达到了 SOTA 结果，验证了不同推理步中更新问句表示能够对问答效果有所提升. Cui 等人[47]提出了一种动态推理网络 DRN（dynamic reasoning network），如图 10 所示，该模型采用 GRU 对上一步推理路径中的所有关系信息进行编码作为历史推理信息，依据历史推理信息对问句表示进行动态更新. 在 MetaQA 和 Pathquestion 这 2 个数据集上的实验结果表明，DRN 的 *Hits*@1 均

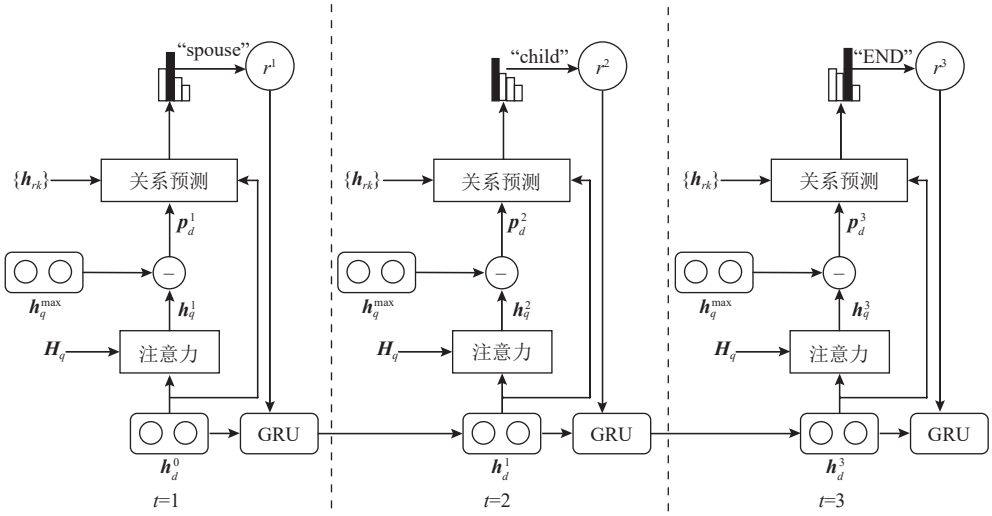


Fig. 10 DRN model diagram

图 10 DRN 模型图

高于基线模型 IRN 和 QA2MN，说明在推理阶段的动态更新问句有利于问答准确率的提升. 与文献 [47] 类似的做法，Du 等人[48] 提出了迭代更新问句表示的方法 KBIGER(knowledge base iterative instruction generating and reasoning)，该方法依据上一步推理子图的信息迭代生成下一步问句表示，用上一步推理子图的结果来更新问句表示，通过这种方式来更深入融合问句与推理子图之间的信息. WebQuestionSP 和 ComplexWebQuestions 数据集上的实验结果表明，KBIGER 模型的 *Hit*@1 和 *F*1 均高于 NSM 等基线模型，验证了利用历史推理信息对问句进行动态更新表示能够提升模型推理能力. 模型推理过程中，候选答案的不同特征也会影响对问句各部分的关注程度. 为了动态调整候选答案不同方面对问句的影响，Hao 等人[49] 提出了一个端到端的神经网络模型，根据候选答案的类型、关系、上下文等不同特征，使用交叉注意力机制迭代更新其相应权重以及问句表示. 该模型依据问句采用注意力机制选择候选答案的不同特征，从而调整问句对候选答案不同特征的关注度；再将这些候选答案不同特征采用注意力机制融入问句不同部分中，实现对问句表示的动态更新. WebQuestionsSP 数据集上的实验结果表明该模型的 *F*1 性能优于其基线模型，说明依据候选答案不同特征动态调整问句表示有助于提升模型推理能力并得到更准确的答案.

此外，文献 [35] 对问句节点连接知识库实体节点进行推理，但推理过程中子图与问句的连接结构是保持不变的，即推理的是静态子图，因此模型推理能力极其有限. 针对该问题，Wu 等人[50] 通过动态扩展子图的方法来消除静态模型推理的局限性. 其具体做法是：根据每一步推理中问句关注部分的不同，让问句连接推理子图中的不同节点，子图内部连接结构因此会根据问句关注的部分不同而动态变化，消除了静态模型推理的局限性.

事实上，仅仅通过 1 次检索所获取的信息是有限的. 面对复杂的多跳问句，在推理过程中所需要的推理知识需要动态检索和变化. 针对此类问题，Jiang 等人[51] 结合大语言模型提出了一种动态知识获取的方法 FLARE(forward-looking active retrieval). 该方法会根据当前推理结果预测下一步所需的内容，然后将预测结果作为查询检索相关知识，在推理过程中实现了推理知识的动态获取与更新. 数据集 WikiMultihopQA 上的实验结果表明，模型在准确率和召回率方面有了显著提高，说明了动态获取推理信息有助于回答准确率的提升. 沿用上述研究思路，Trivedi 等人[52] 提出了 IR-CoT，在文献 [51] 基础上引入了思维链，使用思维链来引导知识的动态获取，并利用获取的知识来改进思维链，从而保证了推理阶段的第 1 步中推理知识与问句的相关性. Asai 等人[53] 提出了一种自适应的的检索方法 Self-RAG，该方法优化了文献 [51] 动态获取知识的过程，并在此基础上加入了惩罚机制. Self-RAG 会在推理每一步骤中动态获取所需知识，并且根据每一步推理结果生成惩罚 token 来对结果进行评判，有助于提高检索信息的相关性.

2.3.4 推理子图不完整问题

在信息检索阶段，虽然研究者围绕知识库补充做了大量工作，但是经过补充的知识库仍然可能存在缺失导致推理子图不完整. 如何在不完整推理子图

上得到和问句匹配的正确答案是知识库问答研究关注的一个焦点，目前研究方法主要可分为 2 类：一类是利用不完整推理信息检索外部文本资源辅助答案推理；另一类是通过样例或相似问答对提升推理效果.

针对推理子图不完整问题，Xiong 等人[54] 借鉴了知识库缺失时结合文本检索处理的方式，设计出一个融合不完整推理子图和一组文本片段的知识库问答模型. 如图 11 所示，该模型包含子图阅读器和文本感知阅读器 2 个模块，子图阅读器负责从推理子图中积累实体知识，文本感知阅读器则利用积累的实体知识阅读外部文本片段以获得与子图相关的辅助资源，最后将推理子图和外部文本资源中的知识聚合生成答案. 和文献 [26−27] 采用文本语料库补充缺失知识库不同，这些方法基于不完整推理子图信息进行检索并得到相关文本片段进行回答，检索得到的文本更具相关性.

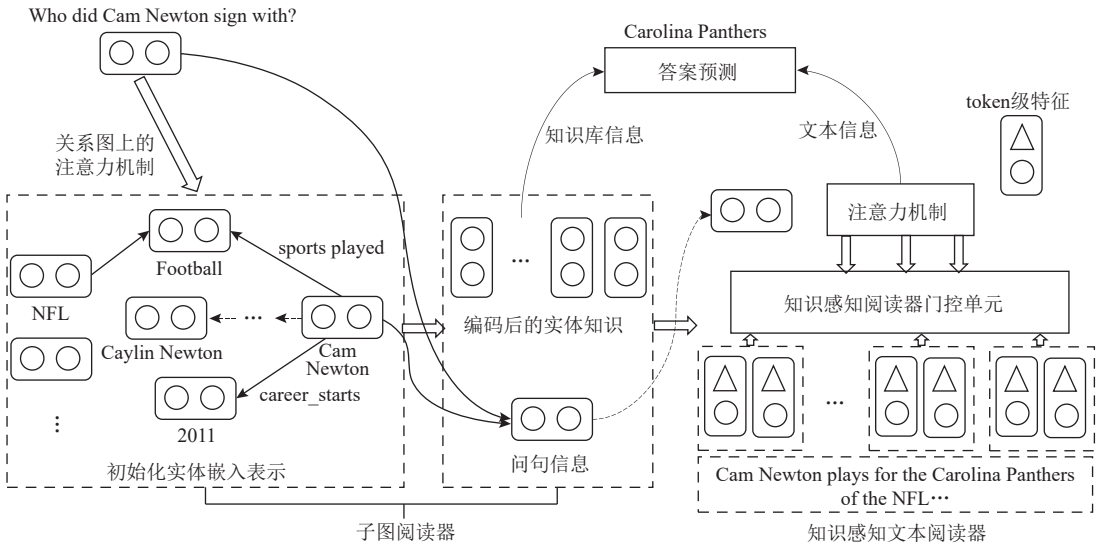


Fig. 11 Model diagram of subgraph reader and text aware reader

图 11 子图阅读器和文本感知阅读器模型图

利用大语言模型的生成能力，Yu 等人[55] 提出了一种名为 GENREAD(generate-then-read)的方法来补充缺失信息. GENREAD 首先利用 InstructGPT 等大语言模型的生成能力并根据问句生成相关的文本，然后在知识库中检索与问句相关的信息，将生成的文本与知识库中检索的信息结合并生成最终答案，目的是在知识库缺失时也能够返回问句的答案. WebQuestion 数据集上的实验表明，GENREAD 通过大语言模型生成文本能够有效提高在知识库缺失情况下问答的准确率.

Thai 等人[56] 提出了一种基于样例的推理方法 CBR-iKB(case-based reasoning over incomplete knowledge bases)，用于不完整推理子图情况下的知识库问答. 在 CBR-iKB 中，训练样例保存在样例库中，当 1 个新问句出现时，模型会在样例库中搜索相似问句及其推理路径，然后基于检索到的案例从推理子图中预测新的推理路径，并选择对应的答案. CBR-iKB 在 MetaQA 数据集上的 *Hit*@1 达到了 100%，提升了问答准确率. 类比人类在相似情景中的学习策略，Mei 等人[57] 提出了一种知识库问答模型 sMoCo，在不完整推理子图上使用相似问答对来推断正确答案. sMoCo 首先通过无监督方式从不完整推理子图中创建出与问句相似的问答对，再利用这些额外生成的相似问答对进行模型训练. 数据集 WebQuestionsSP 上的实验表明，该模型的 *Hit*@1 比基线模型 EmbeddKBQA 和 PullNet 分别提升了 2.2% 和 1.2%，说明额外的相似问答对生成和训练有助于提高问答准确率.

2.4 答案生成

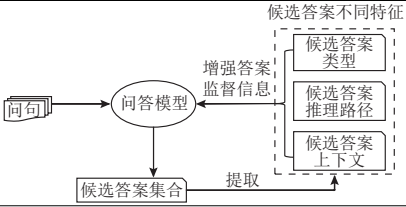
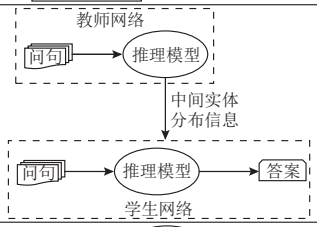
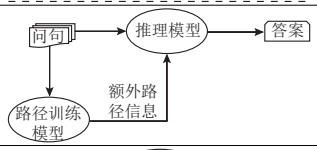
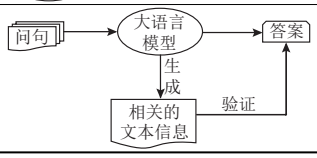
答案生成阶段的目的是通过排序模型或分类模型对候选答案进行筛选，从而获得最终答案. 答案生成时有时存在答案监督信息较弱、答案数目不确定等问题，致使答案生成存在一定的挑战.

2.4.1 弱监督问题

在答案生成阶段，目前大部分文献采用的方法是实体节点与问句表示之间的点积或激活函数计算来获得实体节点得分. 但是这种只依赖问句表示计算实体节点得分的方式缺乏多角度评价，准确率较低，存在弱监督问题. 现阶段针对弱监督问题，主要包含 4 类解决方法：从答案方面、从推理中间实体方面、从路径方面、从生成验证信息等方面增强监督信息. 表 6 对以上 4 类方法进行了总结与对比.

Table 6 Summary and Comparison of Methods for Weak Supervision Problem

表 6 弱监督问题方法总结与对比

增强方法	模型	优点	缺点	方法流程
从答案方面	MCCNNs	结合候选答案的不同方面对候选答案进行打分，增强了监督信息	处理复杂的问句时准确率下降	
从推理中间实体方面	NSM	利用正向和反向推理来增强中间实体分布的学习，增强监督信息	模型计算开销较大	
从路径方面	REKBQA	结合路径节点更丰富的周围节点信息，增强了监督信息	需要大量的人工标记成本	
	MRPQA	使用模型训练得到的最符合路径作为监督信息，减少了人工成本	通过训练得到路径的效果无法衡量	
从生成验证信息方面	LLatrieval	使用大语言模型根据检索信息生成额外文本，可以对文本进行验证	模型参数量较高，消耗资源高	

Dong 等人[58]提出多列卷积神经网络(multi-column convolutional neural networks, MCCNNs)模型从答案路径、答案类型、答案上下文 3 个方面对候选答案进行打分, 通过增加中间监督信号增强了答案的准确性. He 等人[59]则采用基于 NSM(neural state machine)模型的学生-教师网络框架来学习推理中的监督信号. 学生端旨在找到问句正确答案, 教师端旨在学习中间推理实体分布. 学生端可以根据从教师端学习到的中间实体分布来改进自身结果. 其中, 教师端结合了问句主题实体到答案(正向推理)和答案到问句主题实体(反向推理)双向推理机制, 学习的中间实体分布比单向学习到的中间实体分布更加有效. 同样为了在答案生成时引入额外的监督信号, Cao 等人[17]提出一个新的多任务模型 REKBQA, 以推理路径中实体节点周围信息和标记正确路径中真实信息损失为辅助任务, 利用预测的推理路径中实体节点周围信息来监督答案打分. 然而, 该方法在引入额外监督提升问答准确率的同时, 也增加了标记成本. 针对该问题, Wang 等人[60]提出了模型 MRPQA(mutliple reasoning paths KBQA), 将答案推理路径作为一个潜在的变量建模来找到最符合问句语义的路径, 并将此路径作为额外监督信号. WebQuestions, ComplexWebQuestion 数据集上的测评结果表明, 答案推理路径作为监督信号有助于提高问答准确率.

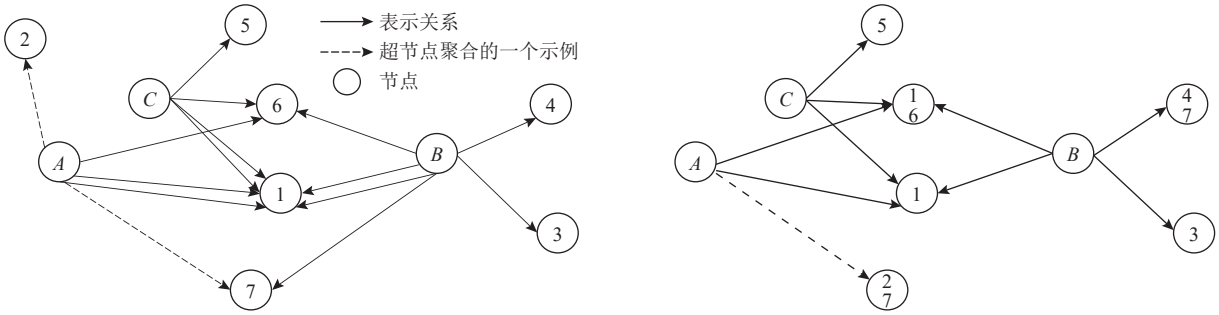
利用大语言模型强大的生成能力, Li 等人[61]提出了一种可验证的方法 LLatrieval, 该方法使用 ChatGPT3.5 根据检索结果生成可验证答案的相关文本, 从而使用户能够灵活验证答案, 即答案是否得到相应文本支持, 使得答案更加可信并且可解释性程度更高.

2.4.2 答案数目不确定性问题

一个问句会存在不确定数量答案的情况, 例如问句“哪些电影由成龙和洪金宝合作完成？”, 答案是一系列电影集合, 这类问句对答案集合的全面性有较高要求. 针对该问题, Li 等人[62]提出一种图摘要技术, 该技术将具有相同关系的节点聚合到 1 个超节点中, 如图 12 所示, 例如节点 A 通过相同关系连接了节点 2, 7, 则将节点 2, 7 聚合成 1 个超节点与节点 A 进行连接, 把原始子图转化为超节点子图, 并使用改进的图卷积神经网络将超节点子图中每个超节点进行嵌入表示, 最终选择 1 个最佳回答该问句的超节点以获取问句的全部答案. 当问句的答案数目不确定时, 该图摘要技术能够显著提高实际答案的召回率.

2.5 共性模块分析与总结

在上述各个阶段中, 大量工作都是采用深度学习网络的某些模块, 其中 BERT, LLM, GNN(GCN),



(a) 原始子图 (b) 超节点子图

Fig. 12 Original subgraph and hypernode subgraph in graph summary technology

图 12 图摘要技术中的原始子图与超节点子图

(self)attention 模块在不同阶段都有广泛的应用，本文对上述模块的作用以及应用阶段进行分析和总结如表 7 表示.

Table 7 Summary and Comparison of Common Network Modules

表 7 共性网络模块总结与对比

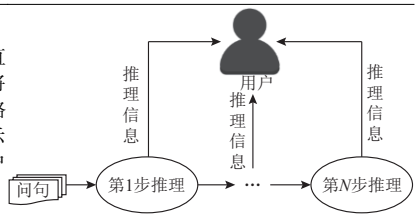
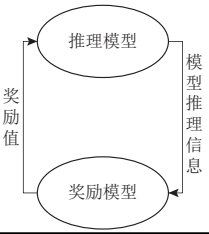
阶段	模块	模块作用	代表模型
问句解析	BERT	问句、知识库嵌入	VRN，CTRN，TSQA，KaFSP，DRGN，EmbedKGQA，ReLMKG，CPR，QA-GNN
	GNN (GCN)	捕获问句时间特征	CTRN，EXAQT
	(Self) Attention	为不同词或 token 分配权重	RASR，DCRN，BAMnet，DRN，NSM，MRPQA
	LLM	生成额外文本消除歧义	query2doc
		拆解问句	self-ask，EIC
信息检索	BERT	提取相似实体特征	SRQA
	GNN (GCN)	区分相似候选实体	QAVGAE
		扩建推理子图	PullNet
	LLM	辅助检索	SKR，CON
模型推理	GNN (GCN)	图节点进行更新计算	QA-GNN，KaFSP，QAVGAE，DRGN，ReLMKG，REAREV，JointLK，DGEN，RAID
	(Self) Attention	为图的邻居节点分配权重	KaFSP，ReLMKG，DGEN，QA-GNN
		为信息交互分配权重	ReLMKG，JointLK
	LLM	动态获取推理知识	FLARE，IR-CoT，Self-RAG
		推理得出问句答案	SKR，CON
答案生成	LLM	生成额外信息验证答案	LLatrieval

3 模型的不可解释性

目前基于信息检索的知识库问答研究大都采用端到端的深度学习模型，有效降低了人工成本，但这类端到端模型往往只输出最终结果，所有决策都隐含在模型本身，这种强烈的不可解释性导致用户对结果答案难以完全满意. 近年来，基于信息检索知识库问答模型的可解释性成为一个研究热点，目前研究方法主要包括 2 类：一类是通过给出中间推理路径表明推理过程来实现可解释性；另一类是采取强化学习方式在推理过程中显式奖励以实现可解释性. 表 8 对这 2 类方法进行总结与对比. 为了解决基于信息检索知识库问答的不可解释性，许多文献对推理过程中的推理路径进行研究，将推理路径可视化直观地展示给用户. Shi 等人[63] 提出了一个透明的问答模型 TransferNet，在每一步推理中关注问句不同部分，并计算推理子图中实体所连接关系边的激活分数，通过可视化每个推理步中被激活的关系和实体来跟踪模型动态. Zhou 等人[45] 提出了一种可解释推理网络模型，显式地对大规模多跳关系路径进行建模. 该模型通过这种方式提供了对复杂问句完整推理路径的可视化能力，有助于推理分析和错误诊断，从而可允许人工参与答案预测. Feng 等人[64] 提出一种新的知识感知方法 MHGRN(multi-hop graph relation network)，

Table 8 Summary and Comparison of Methods for Model Inexplicability

表 8 模型不可解释性的方法总结与对比

方法	代表模型	特点	方法流程
可视化推理路径	TransferNet MHGRN DRN	能够直观地将推理路径展示给用户	
强化学习显式奖励	SRN ARL RPGQA	在每一步推理中有明确的奖励值	

将基于路径的推理方法和图神经网络相结合. MHGRN 能够区分不同邻居或关系类型的重要性, 提供明确的关系路径行为解释, 显式地对关系路径进行建模表示, 在实体节点迭代更新过程中实现较好的可解释性. Cui 等人[47] 提出一种名为 DRN(dynamic reasoning network)的动态推理网络, 采用可解释的逐步推理过程来预测下一步关系, 把所有的中间关系形成可追溯的推理路径. 在逐步推理过程中, DRN 根据问句和决策历史, 为实体连接的每个关系计算语义分数并选择得分最高的关系来扩展推理路径, 增强了模型可解释性.

除了给出中间推理路径之外, 很多研究人员也借鉴强化学习根据奖励值探索推理路径的特征以解决问答系统的不可解释性. Qiu 等人[65] 把基于强化学习的神经模型 SRN(stepwise reasoning network)应用于问答, 将多跳问答任务转化为顺序决策任务. 模型从问句主题实体开始学习策略, 并在知识库上执行路径搜索以获得答案, 通过奖励函数给出的奖励值来纠正路径生成的偏差. 这种结合奖励推理过程中的路径生成增强了模型可解释性. Zhang 等人[66] 提出一种自适应强化学习(adaptive reinforcement learning, ARL)模型, 设计了扩展、桥接和联合 3 类原子操作, 自适应扩展推理路径. 在每一步推理中, “扩展”操作用于添加关系形成推理路径; “桥接”操作用于合并多个问句实体的路径; “联合”操作用于聚合问句实体的多个关系. ARL 通过每一步推理中获得的奖励来纠正推理路径的形成, 增强了模型的可解释性. Xiang 等人[67] 提出一个基于策略的强化学习模型 RPGQA(reasoning path generation question answering), 将知识库问答任务转化为知识库推理路径生成任务, 通过奖励策略引导来推理一条通往答案的路径. RPGQA 提出了一个包括最终问答奖励、语义奖励和推理路径奖励 3 部分组成的奖励策略, 以缓解奖励延迟和奖励稀疏问题. RPGQA 的奖励策略涵盖了问句答案、答案语义及有效路径 3 个方面, 特别有助于增强答案生成的可解释性.

上述基于强化学习的方法虽然有助于增强基于信息检索问答系统的可解释性, 但是强化学习的本质是在大空间中寻找最优解, 是介于穷举和贪婪之间的一种探索机制, 其面临的较大规模路径搜索的问题尚待解决.

4 数据集及基线模型

4.1 常见数据集

本节将不同问答任务中常见数据集概括为 3 类进行介绍: 简单问答数据集、复杂问答数据集和时序问答数据集. 根据答案推理过程所需的推理跳数是单跳还是多跳划分为简单问答数据集和复杂问答数据集; 若数据集问句中含有时间因素, 则为时序问答数据集. 常见的简单问答数据集有 SimpleQuestions; 复杂问答数据集有 WebQuestionsSP, MetaQA, ComplexWebQuestions; 时序问答数据集有 TempQuestions 和 CronQuestions.

1)SimpleQuestions[68] 是一个针对简单问题而构建的数据集, 是单跳的通用类数据集, 其问句采用人工标注方式根据知识库中的三元组事实生成; 答案则以 Freebase 为来源.

2)WebQuestions 的问句来自谷歌 Suggest API, 且以 Freebase 作为答案的来源. WebQuestionsSP[69] 是一个针对多跳问答而构建的数据集, 是单跳与多跳混合的通用类数据集, 它是 WebQuestions 数据集①的子集, 且以 Freebase 作为答案来源.

3)MetaQA[6] 是一个针对多跳问答而构建的数据集, 是多跳的影视类数据集, 其将 WikiMovies②的单跳问句拓展为多跳问句, 且以 Facebook 作为答案来源. ComplexWebQuestions[70] 是一个针对多跳问答而构建的数据集, 是多跳的通用类数据集, 通过向 WebQuestionsSP 数据集中问句添加约束条件来创建问句, 部分问句需要在知识库上进行多达 4 跳甚至更多的推理, 且以 Freebase 为答案来源.

4)TempQuestions[71] 是第一个针对时序问答构建的数据集, 是时间性的通用类数据集, 其在 Free917[68], WebQuestions, ComplexQuestions 上选择出时序问句构成 TempQuestions 的问句, 且以 Freebase 和 Wikidata 作为答案来源.

5)CronQuestions[72] 是一个规模更大的时序问答数据集, 是时间性的通用类数据集, 每个问句都涉及时间元素的推理, 且以 WikiData 作为答案来源.

以上数据集的分类总结如表 9 所示.

4.2 常见基线模型

依据对应阶段不同, 本节针对基于信息检索

① https://nlp. stanford.edu/software/sempre/

② https://paperswithcode. com/dataset/wikimovies

Table 9 Summary of Dataset Classification

表 9 数据集分类总结

类型	名称	提出年代	来源	数量	训练集	验证集	测试集	特点	侧重点
简单问答	SimpleQuestions①	2015	Freebase	108 442	75 910	10 845	21 687	单跳	通用类
复杂问答	WebQuestionsSP②	2016	Freebase	4 887	2 098		1 639	单/多跳	通用类
	MetaQA③	2018	Facebook	407 513	329 282	39 183	39 093	多跳	电影类
	ComplexWebQuestions④	2018	Freebase	34 689	27 734	3 480	3 475	多跳	通用类
时序问答	TempQuestions⑤	2018	Freebase	1 271				时间性	通用类
	CronQuestions⑥	2021	WikiData	410 000	350 000	30 000	30 000	时间性	通用类

知识库问答的常见基线模型进行分类总结如表 10 所示.

Table 10 Summary of Baselines Classification

表 10 基线分类总结

阶段	名称	说明	发表会议	作者单位
	KV-Mem	试图直接从原始文本中给出答案	ACL 2016	Facebook 公司
问句解析	EXAQT	增强问句中的时间信号	CIKM 2021	西南交通大学
	VRN	处理带有噪声的问句	AAAI 2018	格鲁吉亚理工学院
信息检索	EmbedKGQA	使用 1 个评分函数来评估实体间隐藏的关系是否有效	ACL 2020	班加罗尔印度科学研究所
	PullNet	迭代构建包含与问句相关的推理子图	EMNLP 2019	Google AI 公司
	GraftNet	将知识库中实体与文本进行关联	EMNLP 2018	卡内基梅隆大学
模型推理	IRN	根据上一步推理更新下一步对问句的关注度	COLING 2018	清华大学
	SGReader	融合不完整推理子图和一组文本片段的知识库问答模型	ACL 2019	加州大学
答案生成	NSM	设计教师-学生网络，增强监督信息	WSDM 2021	中国人民大学
可解释性	TransferNet	可视化每个推理步中被激活的关系和实体，增强可解释性	EMNLP 2021	清华大学
	SRN	使用强化学习将问答任务转化为顺序决策任务	WSDM 2020	中国科学院计算技术研究所

5 总结与展望

知识库问答作为自然语言处理的重要分支，近年来受到大量学者和研究人员的关注. 本文针对基于信息检索的知识库问答，从 4 个阶段和模型整体不可解释性等角度阐述了该领域当前的热点问题和解决方法，并对后续面临的挑战和未来展望总结如下：

1）问句解析阶段中融入大语言模型有助于加深对问句的深层次语义理解. 当前有 2 类问句回答准确率相对较低，分别为复杂问句和时序问句. 现有方法大都集中在基于知识库相关信息或者结合问句上下文进行细粒度建模提升复杂问句的理解能力上. 随着大语言模型的快速发展，其表现出语言解析的优异性能. 未来可以参考文献 [73] 的研究思路尝试把知识库和大语言模型结合来实现复杂问句分析. 大语言模型可以加深对问句的深层次语义理解，判断出问句中不同成分，更准确识别问句中的主题实体和关键特征，为生成准确的回答奠定基础. 此外，时序问句的解析和回答同样是问答领域的一大难题，因为模型需要复杂的时间推理才能得到正确答案. 鉴于目前大部分知识库都不包含时序关系，现阶段研究大都集中在如何对时序问题进行解析方面. 将大语言模型融入现有知识库以增强知识库时序性特征，并在此基础上设计相应的时序推理模型，包括对时间线的建模、时间间隔的计算、事件顺序的推断等，有助于提高时序问句回答的准确率.

2）信息检索阶段融入多模态和大语言模型生成的文本数据有助于提供更全面的知识信息. 受知识库存储容量的限制，知识库中有时可能不包含问答问句的必要信息. 目前针对知识库不完整主要有 3 种解决方法：基于大量文本语料库结合知识库进行回答[26-28]；通过模型学习方式补充知识库[29-34]；使用大语言模型生成相关信息来进行补充[55]. 但是，第 1 种方法经过

① https://worksheets. codalab. org/worksheets/0xb%20a659fe363cb46e7a505c5b6a774dc8a

② http://aka.ms/WebQSP

③ https://github. com/yuyuz/MetaQA

④ https://www. taunlp. org/compwebq

⑤ https://github. com/IBM/tempqa-wd

⑥ https://github. com/apoorvumang/CronKGQA

文本补充之后得到的知识库仍然有可能是缺失的;第2种方法则可能面临着知识库补充的准确性和可解释性无法考量问题;第3种方法对于大语言模型生成的文本相关性无法进行有效验证.类似于文献[74–75]使用知识库结合图像信息进行回答,未来可以参考文献[76]从多模态角度出发,融合知识库、文本语料、图像、视频及语音等多模态数据对异构知识库统一整合和嵌入,以及融入大语言模型并根据问句生成相关文本信息;并在此基础上进行跨模态推理和学习,以发现模态之间的关联和潜在的知识.通过不同模态数据的融合和联合推理学习,弥补知识库缺失,提供更准确、更全面的知识以提高问答准确性.

3)在模型推理阶段将强化学习和多任务学习融入大语言模型以提升推理泛化能力.多跳推理一直以来是模型推理过程中面临的挑战之一,需要在知识库推理过程的每一跳中找到适当的节点和路径.现阶段研究大都集中在如何提高模型对问句的理解和知识库的推理方面,并没有对于每一跳的推理准确性和合理性进行深入研究,未来可以从2个方面继续研究:一方面参考文献[65–67]的强化学习思路,将强化学习的思想引入大语言模型(类ChatGPT模型)中,使得模型能够在不断交互中,根据每一跳获得的奖励值来学习如何进行多跳推理,以获取更高的问答准确率;另一方面可以采取多任务学习方法,让模型能够正确回答问句的同时,将每一跳视为一个独立的任务,以学习到每一跳应该采取的路径,提高模型泛化能力和效果.

4)答案生成阶段引入思维链技术和用户反馈机制有助于提高模型回答的完整性和准确性.目前针对部分问句答案存在多个情况的解决方法较少.针对该问题,一方面可以参考文献[43]多链推理思想,提示大语言模型通过不同思维链沿不同路径进行推理,且根据各个路径上的推理结果对最终答案进行整合.另一方面可以增加用户反馈机制,鼓励用户对答案结果进行反馈,让模型根据用户反馈信息不断学习和改进,以提高回答的全面性和准确性.

总之,目前随着大语言模型技术发展,检索知识库相关信息以增强大语言模型在知识库问答领域取得了不错效果,其具有推理能力强、通用性强等优势,但是生成式模型存在的幻觉问题会造成答案不准确甚至会出现严重偏差,因此,对于精确化程度要求较高领域如医疗、金融等准确率仍然需要提升.未来可以从检索方法和策略出发,在模型推理前检索出更相关准确信息来增强大语言模型的推理能力,缓解生成式模型的幻觉问题.

作者贡献声明: 田萱提出指导意见并修改论文;吴志超负责论文撰写.

参 考 文 献

[1] Lan Yunshi, He Gaole, Jiang Jinhao, et al. A survey on complex knowledge base question answering: Methods, challenges and solutions[C]//Proc of the 30th Int Joint Conf on Artificial Intelligence. Freiburg: IJCAI, 2021: 4483−4491

[2] Xie Zhiwen, Zeng Zhao, Zhou Guangyou, et al. Topic enhanced deep structured semantic models for knowledge base question answering[J]. Science China Information Sciences, 2017, 60(11): 110103: 1−110103

[3] Zheng H S, Mishra S, Chen Xinyun, et al. Take a step back: Evoking reasoning via abstraction in large language models[J]. arXiv preprint, arXiv: 2310.06117, 2023

[4] Li Junzhuo, Xiong Deyi. KaFSP: Knowledge-aware fuzzy semantic parsing for conversational question answering over a large-scale knowledge base[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2022: 461−473

[5] Wang Liang, Yang Nan, Wei Furu. Query2doc: Query expansion with large language models[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 9414−9423

[6] Zhang Yuyu, Dai Hanjun, Kozareva Z, et al. Variational reasoning for question answering with knowledge graph[C]//Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 6069−6076

[7] Yu Xiaomei, Feng Wenzhi, Wang Hong, et al. An attention mechanism and multi-granularity-based Bi-LSTM model for Chinese Q&A system[J]. Soft Computing, 2020, 24(8): 5831−5845

[8] Press O, Zhang M, Min S, et al. Measuring and narrowing the compositionality gap in language models[J]. arXiv preprint, arXiv: 2210.03350, 2023

[9] Haji S, Suekane K, Sano H, et al. Exploratory inference chain: Exploratorily chaining multi-hop inferences with large language models for question-answering[C]//Proc of the 17th Int Conf on Semantic Computing (ICSC). Piscataway, NJ: IEEE, 2023: 175−182

[10] Jia Zhen, Abujabal A, Saha R R, et al. TEQUILA: Temporal question answering over knowledge bases[C]//Proc of the 27th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2018: 1807−1810

[11] Jia Zhen, Pramanik S, Saha R R, et al. Complex temporal question answering on knowledge graphs[C]//Proc of the 30th ACM Int Conf on Information & Knowledge Management. New York: ACM, 2021: 792−802

[12] Shang Chao, Wang Guangtao, Qi Peng, et al. Improving time sensitivity for question answering over temporal knowledge

graphs[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2022: 8017−8026

[13] Yao Junping, Wang Yijing, Li Xiaojun, et al. TERQA: Question answering over knowledge graph considering precise dependencies of temporal information on vectors[J]. Displays, 2022, 74: 102269

[14] Jiao Songlin, Zhu Zhenfang, Wu Wenqing, et al. An improving reasoning network for complex question answering over temporal knowledge graphs[J]. Applied Intelligence, 2023, 53(7): 8195−8208

[15] Wang Meiling, Li Min, Sun Kewei, et al. Entity difference modeling based entity linking for question answering over knowledge graphs[C]//Proc of the CCF Int Conf on Natural Language Processing and Chinese Computing. Berlin: Springer, 2022: 221−233

[16] Luo Da, Su Jindian, Yu Shanshan. A BERT-based approach with relation-aware attention for knowledge base question answering[C/OL]//Proc of the 2020 Int Joint Conf on Neural Networks (IJCNN). Piscataway, NJ: IEEE, 2020[2024-06-17]. https://ieeexplore-ieee-org-s.libyc.nudt.edu.cn/document/9207186/citations?tabFilter=papers#citations

[17] Cao Yong, Li Xianzhi, Liu Huiwen, et al. Pay more attention to relation exploration for knowledge base question answering[C]// Findings of the Association for Computational Linguistics (ACL 2023). Stroudsburg, PA: ACL, 2023: 2119−2136

[18] Shamsabadi A S, Ramezani R, Farsani H K, et al. Direct relation detection for knowledge-based question answering[J]. Expert Systems with Applications, 2023, 211: 118678

[19] Hao Zhifeng, Wu Biao, Wen Wen, et al. A subgraph-representation-based method for answering complex questions over knowledge bases[J]. Neural Networks, 2019, 119: 57−65

[20] Christmann P, Saha R R, Weikum G. Beyond NED: Fast and effective search space reduction for complex question answering over knowledge bases[C]//Proc of the 15th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2022: 172−180

[21] Chen Yu, Wu Lingfei, Zaki M J. Bidirectional attentive memory networks for question answering over knowledge bases[C]//Proc of the 2019 Conf of the North. Stroudsburg, PA: ACL, 2019: 2913−2923

[22] Cai Jianyu, Zhang Zhanqiu, Wu Feng, et al. Deep cognitive reasoning network for multi-hop question answering over knowledge graphs[C]//Findings of the Association for Computational Linguistics(ACL-IJCNLP 2021). Stroudsburg, PA: ACL, 2021: 219−229

[23] Luo Dan, Sheng Jiawei, Xu Hongbo, et al. Improving complex knowledge base question answering with relation-aware subgraph retrieval and reasoning network[C/OL]//Proc of the 30th Int Joint Conf on Neural Networks (IJCNN). Piscataway, NJ: IEEE, 2023[2024-06-17]. https://ieeexplore-ieee-org-s.libyc.nudt.edu.cn/search/searchresult.jsp?newsearch=true&queryText=Improving%20complex%20knowledge%20base%20question%20answering%20with%20relation-aware%20subgraph%20retrieval%20and%20reasoning%20network

[24] Wang Yile, Li Peng, Sun Maosong, et al. Self-knowledge guided retrieval augmentation for large language models[J]. arXiv preprint, arXiv: 2310.05002, 2023

[25] Yu Wenhao, Zhang Hongming, Pan Xiaoman, et al. Chain-of-note: Enhancing robustness in retrieval-augmented language models[J]. arXiv preprint, arXiv: 2311.09210, 2023

[26] Sun Haitian, Dhingra B, Zaheer M, et al. Open domain question answering using early fusion of knowledge bases and text[C]//Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 4231−4242

[27] Sun Haitian, Bedrax-Weiss T, Cohen W. PullNet: Open domain question answering with iterative retrieval on knowledge bases and text[C]//Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: ACL, 2019: 2380−2390

[28] Zhang Ying, Meng Fandong, Zhang Jinchao, et al. MKGN: A multi-dimensional knowledge enhanced graph network for multi-hop question and answering[J]. IEICE Transactions on Information and Systems, 2022, 105(4): 807−819

[29] Saxena A, Tripathi A, Talukdar P. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 4498−4507

[30] Shi Xiujin, Hu Jun, Sun Naiwen, et al. TrEKBQA: Traversing knowledge graph embedding for multi-hop knowledge base question answering[J]. Journal of Physics: Conf Series, 2023, 2424(1): 012−027

[31] Zan Daoguang, Wang Sirui, Zhang Hongzhi, et al. Complex question answering over incomplete knowledge graph as n-ary link prediction[C/OL]//Proc of the 29th Int Joint Conf on Neural Networks (IJCNN). Piscataway, NJ: IEEE, 2022[2024-06-17]. https://ieeexplore-ieee-org-s.libyc.nudt.edu.cn/search/searchresult.jsp?newsearch=true&queryText=Complex%20question%20answering%20over%20incomplete%20knowledge%20graph%20as%20n-ary%20link%20prediction

[32] Zhao Fen, Li Yinguo, Hou Jie, et al. Improving question answering over incomplete knowledge graphs with relation prediction[J]. Neural Computing and Applications, 2022, 34(8): 6331−6348

[33] Zheng Chen, Kordjamshidi P. Dynamic relevance graph network for knowledge-aware question answering[C]//Proc of the 29th Int Conf on Computational Linguistics. Southampton, UK: ICCL, 2022: 1357−1366

[34] Cui Hai, Peng Tao, Han Ridong, et al. Reinforcement learning with dynamic completion for answering multi-hop questions over incomplete knowledge graph[J]. Information Processing & Management, 2023, 60(3): 103283

[35] Yasunaga M, Ren Hongyu, Bosselut A, et al. QA-GNN: Reasoning with language models and knowledge graphs for question answering[C]//Proc of the 2021 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2021: 535−546

[36] Sun Yueqing, Shi Qi, Qi Le, et al. JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering[C]//Proc of the 2022 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies. Stroudsburg, PA: ACL, 2022: 5049−5060

[37] Cao Xing, Liu Yun. ReLMKG: Reasoning with pre-trained language models and knowledge graphs for complex question answering[J]. Applied Intelligence, 2023, 53(10): 12032−12046

[38] Wang Xu, Zhao Shuai, Cheng Bo, et al. Explore modeling relation information and direction information in KBQA[J]. Neurocomputing, 2022, 471: 139−148

[39] Feng Zhangyin, Feng Xiaocheng, Zhao Dezhi, et al. Retrieval-generation synergy augmented large language models[J]. arXiv preprint, arXiv: 2310.05149, 2023

[40] Hu Nan, Bi Sheng, Qi Guilin, et al. Improving core path reasoning for the weakly supervised knowledge base question answering[C]//Proc of the Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2022: 162−170

[41] Niu Guanglin, Li Yang, Tang Chengguang, et al. Path-enhanced multi-relational question answering with knowledge graph embeddings[J]. arXiv preprint, arXiv: 2110.15622, 2021

[42] Bi Xin, Nie Haojie, Zhang Guoliang, et al. Boosting question answering over knowledge graph with reward integration and policy evaluation under weak supervision[J]. Information Processing & Management, 2023, 60(2): 103242

[43] Yoran O, Wolfson T, Bogin B, et al. Answering questions by meta-reasoning over multiple chains of thought[J]. arXiv preprint, arXiv: 2304.13007, 2023

[44] Park J, Patel A, Khan O Z, et al. Graph-guided reasoning for multi-hop question answering in large language models[J]. arXiv preprint, arXiv: 2311.09762, 2023

[45] Zhou Mantong, Huang Minlie, Zhu Xiaoyan. An interpretable reasoning network for multi-relation question answering[C]//Proc of the 27th Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2018: 2010−2022

[46] Li Xinmeng, Alazab M, Li Qian, et al. Question-aware memory network for multi-hop question answering in human-robot interaction[J]. Complex & Intelligent Systems, 2022, 8(2): 851−861

[47] Cui Hai, Peng Tao, Bao Tie, et al. Stepwise relation prediction with dynamic reasoning network for multi-hop knowledge graph question answering[J]. Applied Intelligence, 2023, 53(10): 12340−12354

[48] Du Haowei, Huang Quzhe, Zhang Chen, et al. Knowledge-enhanced iterative instruction generation and reasoning for knowledge base question answering[C]//Proc of the CCF Int Conf on Natural Language Processing and Chinese Computin. Berlin: Springer, 2022: 431−444

[49] Hao Yanchao, Zhang Yuanzhe, Liu Kang, et al. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge[C]//Proc of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2017: 221−231

[50] Wu Wenqing, Zhu Zhenfang, Qi Jiangtao, et al. A dynamic graph expansion network for multi-hop knowledge base question answering[J]. Neurocomputing, 2023, 515: 37−47

[51] Jiang Zhengbao, Xu F, Gao Luyu, et al. Active retrieval augmented generation[J]. arXiv preprint, arXiv: 2305.06983, 2023

[52] Trivedi H, Balasubramanian N, Khot T, et al. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions[J]. arXiv preprint, arXiv: 2212.10509, 2023

[53] Asai A, Wu Zeqiu, Wang Yizhong, et al. Self-RAG: Learning to retrieve, generate, and critique through self-reflection[J]. arXiv preprint, arXiv: 2310.11511, 2023

[54] Xiong Wenhan, Yu Mo, Chang Shiyu, et al. Improving question answering over incomplete KBs with knowledge-aware reader[C]// Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 4258−4264

[55] Yu Wenhao, Iter Dan, Wang Shuohang, et al. Generate rather than retrieve: Large language models are strong context generators[J]. arXiv preprint, arXiv: 2209.10063, 2023

[56] Thai D, Ravishankar S, Abdelaziz I, et al. CBR-iKB: A case-based reasoning approach for question answering over incomplete knowledge bases[J]. arXiv preprint, arXiv: 2204.08554, 2022

[57] Mei Shijie, Hu Xinrong, Yao Xun, et al. Seeing the wood for the trees: A contrastive regularization method for the low-resource knowledge base question answering[C]//Findings of the Association for Computational Linguistics (NAACL 2022). Stroudsburg, PA: ACL, 2022: 1085−1094

[58] Dong Li, Wei Furu, Zhou Ming, et al. Question answering over freebase with multi-column convolutional neural networks[C]//Proc of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int Joint Conf on Natural Language Processing (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2015: 260−269

[59] He Gaole, Lan Yunshi, Jiang Jing, et al. Improving multi-hop knowledge base question answering by learning intermediate supervision signals[C]//Proc of the 14th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2021: 553−561

[60] Wang Yu, Jin Hongxia. A new concept of knowledge based question answering (KBQA) system for multi-hop reasoning[C]//Proc of the 2022 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2022: 4007−4017

[61] Li Xiaonan, Zhu Changtai, Li Linyang, et al. LLatrieval: LLM-verified retrieval for verifiable generation[J]. arXiv preprint, arXiv: 2311.07838, 2023

[62] Li Sirui, Wong KW, Fung C C, et al. Improving question answering over knowledge graphs using graph summarization[C]//Proc of the 28th Int Conf on Neural Information Proc (ICONIP 2021). Berlin: Springer, 2021: 489−500

[63] Shi Jiaxin, Cao Shulin, Hou Lei, et al. TransferNet: An effective and transparent framework for multi-hop question answering over relation graph[C]//Proc of the 2021 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 4149−4158

[64] Feng Yanlin, Chen Xinyue, Lin Bill Yuchen, et al. Scalable multi-hop relational reasoning for knowledge-aware question answering[C]// Proc of the 2020 Conf on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: ACL, 2020: 1295−1309

[65] Qiu Yunqi, Wang Yuanzhuo, Jin Xiaolong, et al. Stepwise reasoning for multi-relation question answering over knowledge graph with weak supervision[C]//Proc of the 13th Int Conf on Web Search and Data Mining. Stroudsburg, PA: ACL, 2020: 474−482

[66] Zhang Qixuan, Weng Xinyi, Zhou Guangyou, et al. ARL: An adaptive reinforcement learning framework for complex question answering over knowledge base[J]. Information Processing & Management, 2022, 59(3): 102933

[67] Xiang Yuxuan, Wu Jiajun, Wang Tiexin, et al. Reasoning path generation for answering multi-hop questions over knowledge graph[C]//Proc of the Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Int Conf on Web and Big Data. Berlin: Springer, 2022: 195−209

[68] Bordes A, Usunier N, Chopra S, et al. Large-scale simple question answering with memory networks[J]. arXiv preprint, arXiv: 1506.02075, 2015

[69] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs[C]//Proc of the 2013 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2013: 1533−1544

[70] Talmor A, Berant J. The Web as a knowledge-base for answering complex questions[C]//Proc of the 2018 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg, PA: ACL, 2018: 641−651

[71] Jia Zhen, Abujabal A, Saha R R, et al. TempQuestions: A benchmark for temporal question answering[C]//Proc of the Web Conf 2018 (WWW 18). New York: ACM, 2018: 1057−1062

[72] Saxena A, Chakrabarti S, Talukdar P. Question answering over temporal knowledge graphs[C]//Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2021: 6663−6676

[73] Hu Nan, Wu Yike, Qi Guilin, et al. An empirical study of pre-trained language models in simple knowledge graph question answering[J]. World Wide Web, 2023, 26(5): 2855−2886

[74] Jiang Lei, Meng Zuqiang. Knowledge-based visual question answering using multi-modal semantic graph[J]. Electronics, 2023, 12(6): 1390

[75] Yang Zhenyu, Wu Lei, Wen Peian, et al. Visual question answering reasoning with external knowledge based on bimodal graph neural network[J]. Electronic Research Archive, 2023, 31(4): 1948−1965

[76] Zhang Yingying, Qian Shengsheng, Fang Quan, et al. A question answering method based on multimodal knowledge perception attention mechanism[J]. Journal of Computer Research and Development, 2020, 57(5): 1037−1045 (in Chinese)
(张莹莹, 钱胜胜, 方全, 等. 基于多模态知识感知注意力机制的问答方法[J]. 计算机研究与发展, 2020, 57(5): 1037−1045)



Tian Xuan, born in 1976. PhD, associate professor. Senior member of CCF. Her main research interests include intelligent information processing and text mining.

田　萱, 1976 年生. 博士, 副教授. CCF 高级会员. 主要研究方向为智能信息处理、文本挖掘.



Wu Zhichao, born in 1999. Master candidate. Student member of CCF. His main research interests include knowledge base question answering and large language models.

吴志超, 1999 年生. 硕士研究生. CCF 学生会员. 主要研究方向为知识库问答、大语言模型.