

文章编号: 1003-0077(2026)08-0076-08

基于解耦与组合的大模型知识问答方法

施圣泽¹, 秦利², 胡军¹, 鲁晶晶³, 王丹丹¹

(1. 中国科学院 软件研究所 智能博弈重点实验室, 北京 100190;
2. 军事科学院 军事科学信息研究中心, 北京 100091;
3. 93010 部队, 辽宁 沈阳 110000)

摘要: 传统的知识库问答模型依赖大量标注数据和复杂训练架构, 难以适应新的知识库, 限制了其应用范围。最近, 大语言模型展示了卓越的代码生成与泛化能力, 但在直接生成知识库查询语句时, 面临推理结构复杂和知识语义多样等挑战。为了解决以上问题, 本文提出了集成大语言模型的知识库问答框架 DC-KBQA, 设计了一种解耦与组合策略, 将知识库查询语句的生成任务解耦为推理结构和语义组件的检索与组合任务。该框架在公开数据集上进行了广泛实验, 结果表明, 该框架提高了知识库查询语句的可执行性和模型的泛化能力。

关键词: 知识问答; 大语言模型; 代码生成

中图分类号: TP391

文献标识码: A

Leveraging Decoupled Composition in Large Language Models for Knowledge Base Question Answering

SHI Shengze¹, QIN Li², HU Jun¹, LU Jingjing³, WANG Dandan¹

(1. State Key Laboratory of Intelligent Game, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China;
2. Academy of Military Science, Military Science Information Research Center, Beijing 100091, China;
3. Unit 93010 of PLA, Shenyang, Liaoning 110000, China)

Abstract: Traditional knowledge base question answering (KBQA) models rely on large amounts of annotated data and complex training architectures, making it difficult for them to be adapted to new knowledge bases. Recently, large language models (LLMs) have demonstrated remarkable capabilities in code generation and generalization. However, they face challenges in directly generating knowledge base query statements due to the complexity of inference structures and the diversity of knowledge semantics. To address these issues, this paper proposes a knowledge base question answering framework, DC-KBQA, that integrates large language models. The framework employs a decoupling and combination strategy, breaking down the task of generating knowledge base query statements into the retrieval and combination of reasoning structures and semantic components. Extensive experiments on public datasets show that this framework improves the performance of knowledge base query statements and enhances the model's generalization ability.

Keywords: knowledge base question answering; large language model; code generation

0 引言

知识库问答系统(Knowledge Base Question Answering, KBQA)^[1]旨在通过自然语言为非专业用户提供访问大规模知识库的接口^[2], 是自然语言处理领域内的一个长期研究热点^[3]。鉴于知识库的庞大规模

及问题的多样性, 构建高性能的 KBQA 模型通常需要依赖大量的标注数据和复杂的训练架构, 这不仅提高了成本, 也限制了模型的应用范围^[4]。特别是, 已训练的模型往往只能在其训练的特定知识库上表现出色^[5], 而难以适应新的、未经训练的知识库^[6]。因此, 设计一种无须依赖特定训练数据且具备更广泛泛化能力的新型 KBQA 框架是至关重要的。

最近,大语言模型(Large Language Models, LLMs)在多个领域显示出卓越的泛化能力^[7],旨在采用通用框架来处理多种不同任务^[8],展现了对新颖场景下任务的出色处理能力^[9]。已有研究表明,LLMs 能实现从自然语言处理到结构化代码的转换^[10],这与 KBQA 的核心任务高度相关^[11]。基于此,本文提出将 LLMs 集成到 KBQA 框架中,以充分发挥其强大的代码生成与场景泛化能力。

然而,将 LLMs 应用于 KBQA 任务面临着若干挑战。首先,对于 LLMs 而言,直接生成完整的知识库查询语句尤为困难。知识库查询语句通常包含两个关键方面的信息:推理结构和语义组件。其中,推理结构涵盖了逻辑类型和推理步数等,而语义组件则涉及实体、关系和类型等元素。有研究指出,将复杂任务分解成更简单的子任务可以有效提升 LLMs 的处理效果^[12]。在 KBQA 任务中,从一个粗粒度的查询语句提取出细粒度的组件,也可以显著增强模型对不同问题的泛化处理能力^[13]。受此启发,本文提出一种新策略,将查询语句中的推理结构和语义组件进行解耦,把生成完整知识库查询语句的复杂任务拆分成逻辑层面和知识层面的检索与组合任务。

同时,LLMs 在预训练阶段很少涉及 KBQA 任务中知识库查询语句的知识,导致直接采用 LLMs 生成的推理结构存在以下两方面的问题:一方面,推理结构的格式错误率较高^[14],常常不符合查询语句的语法规则,导致其在知识库上无法执行;另一方面,LLMs 对复杂问题的逻辑解析能力较弱^[15],导致推理结构的逻辑类型和推理步数错误率较高,无法得到正确的查询结果。为了应对这些挑战,本文考虑将推理结构的生成任务转换为检索与解析任务,这不仅降低了

LLMs 的逻辑推理难度,也提高了查询语句的可执行性。

此外,语义组件作为查询语句中的核心内容,其数量和质量直接影响 LLMs 推理结果的准确性。语义组件的抽取面临以下两方面的问题:一方面,与问题相关的知识可能过多,导致 LLMs 对于语义组件的理解负担过大;另一方面,当与问题不相关的知识被注入 LLMs 中时,会误导 LLMs 得出错误的答案^[16]。为此,本文设计了合适的知识抽取与检查策略,确保检索所得的语义组件数量少且质量高。

综合以上考虑,本文提出了一种基于解耦与组合的 KBQA 推理框架 DC-KBQA (Leveraging Decoupled Composition in Large Language Models for Knowledge Base Question Answering)。首先,将知识库查询语句中的推理和语义信息进行解耦,把直接生成查询语句的复杂任务分解为推理和语义两方面组件的检索与组合任务。具体来说,由 LLMs 驱动检索器在代码库和知识库中检索出相关组件。在推理层面,从已有的查询语句代码库中检索出少量的推理结构,并由 LLMs 根据问题的逻辑类型进行二次筛选。在语义层面,从知识库中检索出少量的语义组件,并设计了基于本体和实例级别的模式检查,由 LLMs 来约束语义组件的数量和质量。接着,由 LLMs 驱动组合器将选取的推理结构和语义组件进行组合,构建出完整且高质量的查询语句列表,并根据推理和语义层面的细粒度信息,择优选取最终的查询语句答案。图 1 为原有方法与本文方法的框架对比,本文方法不仅提升了查询语句的可执行性,还充分发挥了 LLMs 的泛化能力,从而更加高效地应对了 KBQA 任务中的各种挑战。

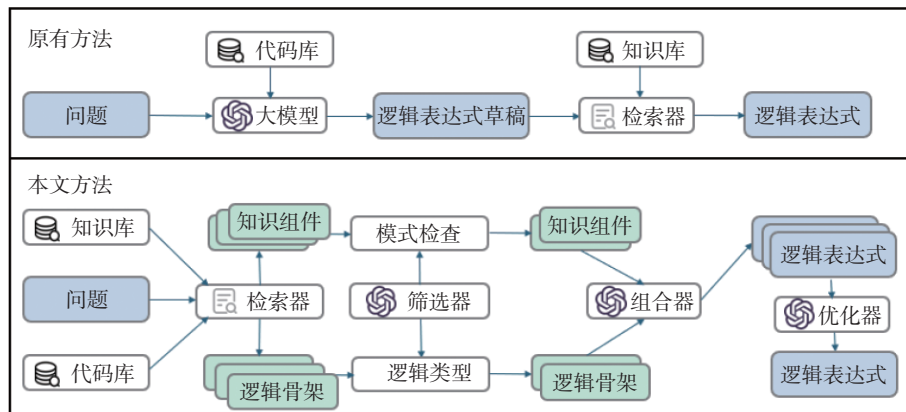


图 1 现有方法与本文方法的对比

1 相关工作

1.1 知识库问答

KBQA 旨在利用知识库作为其知识来源来回答自然语言问题。现有的 KBQA 方法主要分为基于信息检索的方法和基于语义解析的方法。基于信息检索的方法^[17]通过提取与问题相关的子图来生成答案。相较之下,基于语义解析的方法^[18]将问题解析为查询语句,并在知识库中执行以获得答案,这种方法更具可解释性。大多数先进的 KBQA 模型采用了基于语义解析的方法,并使用 S-expression 的逻辑形式,因为它具有紧凑性、组合性和可读性。具体而言,基于语义解析的方法可以分为基于排名的方法^[3]和基于序列到序列的方法^[19]。前者枚举大量的查询语句候选项,并根据与问题的相关性进行排名;后者则直接将问题转换为查询语句。

1.2 大语言模型在知识库问答中的应用

最近,大语言模型在各种任务中展现出强大的少样本上下文学习能力^[20],包括问答、代码生成等^[21],这与基于语义解析的 KBQA 任务相似。研究表明,大语言模型的逻辑推理能力有限,而把复杂任务分解为逐步解决的简单任务有利于提升性能^[22]。FC-KBQA^[13]指出,相较粗粒度的查询语句,用更细粒度的组件(如类别、关系等),可以提高模型的泛化能力。

在这方面,与本文工作最相关的是 KB-Binder。KB-Binder^[20]首次探索了大语言模型在 KBQA 上的应用,利用大语言模型的上下文学习能力来生成查询语句草稿,并替换查询语句中的组件,枚举得到大量候选表达式,选择可执行的表达式来得到答案。然而,KB-Binder 采用生成的方式来获得逻辑草稿,导致格式错误率较高,同时采用替换组件的枚举方式,面临着组合爆炸的问题。与 KB-Binder 不同,本文将粗粒度的查询语句生成任务解耦成推理层面和语义层面细粒度的子任务,并以检索的方式获取少量组件并完成组合,可以提高查询语句的可执行性,从而实现整体性能的提高。

2 方法

2.1 概述

本文提出了一个基于推理结构和语义组件的组合框架 DC-KBQA。对于给定的问题,本文把直接生成完整知识库查询语句的复杂任务从推理和语义两个层面进行解耦,并由 LLMs 驱动的检索器和组合器来完成推理和语义两个层面的组件检索与组合。该策略基于 FC-KBQA^[13]和 KB-Binder^[20]等前沿研究的思想,进一步改进了查询语句的分解方式,并优化了语义组件的模式检查。具体而言,如图 2 所示,该框架包括以下几个模块:查询语句分解、语义组件模式检查、LLMs 驱动的检索器和组合器。

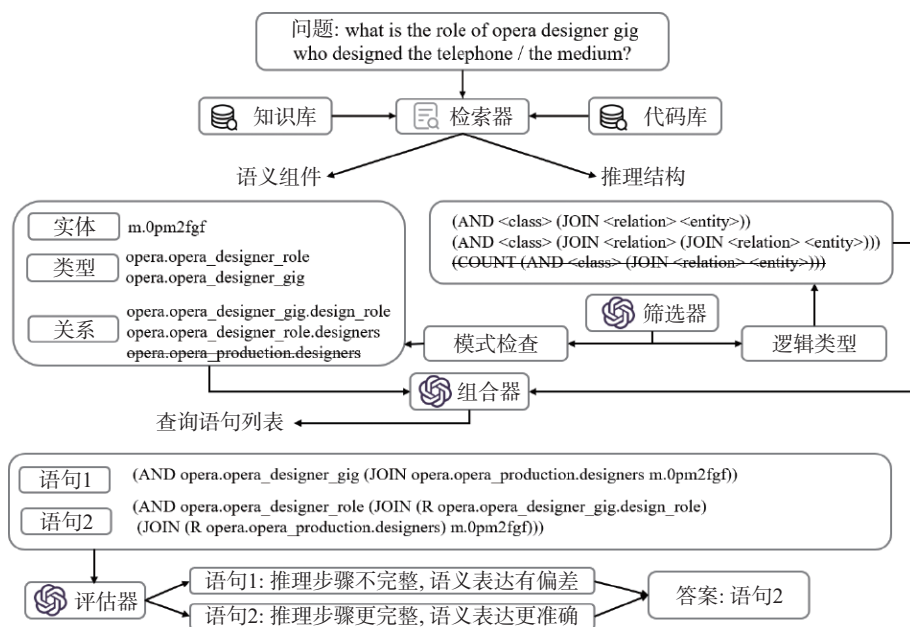


图 2 DC-KBQA 方法框架

2.2 查询语句分解

一个完整的查询语句可表示为 $Q=\langle S,T \rangle$, 其中, S 表示语义成分, T 表示推理结构。具体而言, 语义成分 $S=\{e,r,c\}$, 其中, $e \in E$ 表示知识库中的实体集合, $r \in R$ 表示关系集合, $c \in C$ 表示类别集合。推理维度的组件主要与问题的逻辑结构有关, 包括逻辑类型、推理步数等。本文基于现有研究^[23]总结的细粒度推理结构, 通过解耦方式进一步优化了这些推理组件的检索与组合。为了解耦出语句中的逻辑和知识的信息, 本文用 $\langle \text{class} \rangle, \langle \text{rel} \rangle, \langle \text{entity} \rangle, \langle \text{literal} \rangle$ 等占位符来分别替换类型、关系、实体、数字等语义维度的具体组件, 并保留“AND”“JOIN”“COUNT”等包含推理信息的组件, 如图 3 所示, 来作为该表达式的推理结构。通过对具体知识的遮蔽和推理结构的抽象, 推理结构可以实现对不同模式下逻辑信息的泛化表示。通过使用占位符替换具体知识, 可以在推理结构上灵活地填充不同的语义组件。

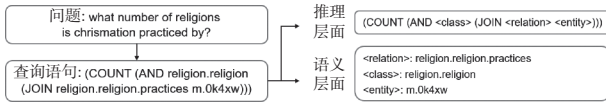


图 3 查询语句分解示意图

2.3 语义组件模式检查

为了确保查询语句的可执行性, 我们通过实体级别与本体级别的双重模式检查, 对语义组件进行严格筛选(图 4)。设定给定实体 e , 对于一跳关系检查, 我们定义一跳关系集合 $R_1(e)=\{r \mid (e,r,e') \in K\}$, 其中, K 为知识库, r 表示从实体 e 出发可以连接的所有关系。对于两跳及以上的多跳关系检查, 我们进一步通过定义本体级关系连接 $R_n(e)=\{r_n \mid r_{n-1} \in R_{n-1}(e), r_n \in K\}$ 来构建多跳候选关系。在通过以上两种检查之后, 本文将一跳关系和多跳关系候选项合并表示成关系组合的形式, 作为最终的关系组件候选项。经过以上检查, 本文结合了知识库的模式信息, 在缩减候选关系数量的同时, 也可以提高查询语句的可执行性。

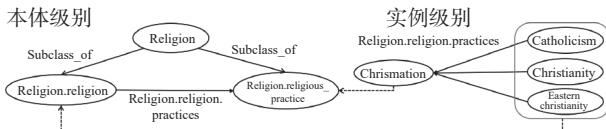


图 4 语义组件模式示意图

2.4 LLMs 驱动的检索器

在查询语句生成过程中, 直接从代码库和知识库

中提取的推理结构和语义组件, 数量众多且质量不一, 难以直接进行组合。为了有效解决这一问题, 本文设计了一种由大语言模型(LLMs)驱动的检索器, 主要包括以下两个步骤:

首先, 采用无监督检索方法, 根据输入的自然语言问题, 从代码库和知识库中获取初步的推理结构和语义组件。这一步骤利用基于语义相似性计算的方法, 从大量的候选项中筛选出少量且相关的组件, 以降低后续问题处理的规模。

在推理层面, 推理结构需要从代码库中提取。在代码库中, 每个样例问题都有对应的查询语句, 也可以提取出对应的推理结构。本文利用无监督检索器为目标问题检索出最相似的若干个样例问题, 并提取对应的推理结构。接着, 本文统计样例中出现的频次最高的若干个推理结构, 将其作为初步检索结果。

在语义层面, 语义组件需要从知识库中提取。一个完整的查询语句涉及的语义组件主要有实体、类型和关系。对于实体组件, 本文采用现成的实体链接器来获取与问题相关的实体候选项。对于关系组件, 本文使用无监督检索器从知识库中检索与问题语义最相似的若干个关系候选项。对于类型组件, 本文基于已获取的实体和关系候选项, 根据知识库的模式信息, 提取出符合模式定义的类型候选项。

通过以上步骤检索得到的组件, 虽然数量相对较少, 但仍然存在质量参差不齐的问题。因此, 本文利用 LLMs 对初步检索到的推理结构和语义组件进行评估与筛选, 来进一步提升组件的质量。为了确保评估的一致性和准确性, 本文从逻辑和语义层面分别设计了评估标准。

对于推理结构而言, 逻辑类型是评估的重要标准。推理结构往往涉及计数、最值、比较等逻辑类型, 而自然语言问题表述中往往有比较明显的关键词, 可以作为逻辑类型判断的参考依据。因此, 本文利用 LLMs 对目标问题进行基于关键词的逻辑类型判断, 将问题进一步划分为普通、计数、最值、比较四种类型中的一种。接着, 根据逻辑类型对初步检索得到的推理结构结果进行筛选, 仅保留逻辑类型相同的候选项, 进一步缩小推理结构的范围。

对于语义组件而言, 模式正确性和语义相关性是评估的重要标准。首先, 本文对初步检索得到的语义组件进行模式检查, 以确保每个组件都符合知识库中的模式定义。接着, 本文利用 LLMs 对语义组件进行语义分析, 通过比较语义组件与问题描述的语义匹配程度, 筛选出与问题最为相关的语义组件, 进一步提高语义组件的质量。

2.5 LLMs 驱动的组合器

在检索到高质量的推理结构和语义组件之后,需要将它们高效组合,生成准确且符合语义的查询语句。为此,本文设计了一种由 LLMs 驱动的组合器。该组合器主要包括两个步骤:

首先,以推理结构为基础,利用 LLMs 完成语义组件的填充。具体而言,LLMs 根据推理结构的占位符位置,从语义组件中选择合适的实体、关系和类别等进行填充,生成完整的查询语句。在这个过程中,LLMs 通过充分理解推理结构的占位符需求和语义组件的语义特点,确保填充组件之间的高度契合。

接下来,利用 LLMs 对生成的查询语句进行评估。具体而言,LLMs 从推理结构和知识语义角度对查询语句进行全面分析,评估其逻辑合理性、语义一致性。逻辑合理性是指查询语句在逻辑推理结构上的正确性和完整性。LLMs 通过解析表达式中的每个逻辑操作和推理步骤,检查其是否存在逻辑操作不准确或推理路径不完整的情况,确保生成的查询语句在理论上是严谨且自洽的。语义一致性涉及查询语句在知识层面上的一致性和准确性。LLMs 利用其语义理解能力,对表达式中的实体和关系进行校验,检查其是否存在知识语义上的偏差或问题表述内容的遗漏,确保其与目标问题的意图保持一致且完整。通过以上评估,可以提高最终生成的查询语句的质量,确保其在知识库上的可执行性。

3 实验

3.1 概述

3.1.1 数据集

本文在三个公共 KBQA 数据集上进行了方法评估,数据集规模特征如表 1 所示。

表 1 公共 KBQA 数据集规模情况

数据集	问题数量	逻辑形式	领域
GrailQA	64 331	4 969	86
WebQSP	4 737	794	56
GraphQA	5 166	500	70

GrailQA^[23]是一个基于 Freebase 的大规模数据集,用于评估三种层次的泛化能力:独立同分布、组合泛化和零样本泛化。它包含复杂的逻辑形式,涉及最多 4 跳关系和多种聚合函数,包括比较级、最高级和计数操作。

WebQSP^[24]是一个中等规模的数据集,包含来自 Google 查询日志的问题。它主要用于评估独立同分布类型问题的泛化能力。其问题主要涉及最多 2 跳关系,偶尔包括实体和聚合约束。

GraphQA^[6]是一个通过图查询的句子级复述构建的中等规模数据集,旨在评估组合泛化能力。由于引入了句法复杂性和表达多样性,因此具有较大的挑战性。

3.1.2 评估指标

本文使用官方脚本和各种评估指标来评估不同数据集上的表现。对于 GrailQA,本文使用精确匹配(Exact Match, EM)来衡量查询语句的准确性,并根据预测答案和标准答案计算 F_1 得分。对于 WebQSP 和 GraphQA,本文按照标准评估方法报告 F_1 得分。

3.1.3 基线

本文主要和以下两种类型的方法作对比:直接采用 LLMs 进行推理的基线方法,基于 LLMs 推理框架的方法。前者直接采用几种主流的 LLMs 进行推理,包括 ChatGPT、Davinci-003 和 GPT-4,为 KBQA 的方法提供可靠的性能基准。后者的代表性方法主要有 KB-Binder^[20]和 KB-Coder^[14],其中 KB-Binder 是一种基于上下文学习范式的知识库问答方法,通过使用多个(问题,逻辑形式)对作为演示示例,并引导大语言模型生成正确的逻辑形式。KB-Coder 则是一种代码风格的上下文学习方法,专注于知识基础性问题回答。它们在 KBQA 推理框架中结合了 LLMs 代码生成的能力,进一步提高了模型的推理性能。

3.1.4 实现

在无监督检索步骤中,本文采用 BM25 检索器,分别从代码库和知识库中对推理结构和语义组件进行检索。对于推理结构而言,本文初步检索了 50 个相似问题,并保留前 5 个频率最高的推理结构。对于语义组件而言,本文初步检索了 100 个关系组件。在 LLMs 驱动的评估筛选步骤中,本文进一步对语义组件进行模式检查,并由 LLMs 筛选出 10 个最相关的语义组件。

3.2 主要结果

在实验中,本文评估了 DC-KBQA 在多个数据集不同设置下的性能表现。总体而言,与基于大语言模型的无须训练的方法相比,DC-KBQA 在多种场景下都表现出优越的性能。

如表 2 所示,在 GrailQA 数据集上,DC-KBQA 在独立同分布、组合和零样本三种任务中均取得了性能的提升。具体而言,在零样本任务中,DC-KBQA 表现

尤为突出,表明其在处理未知问题时具备强大的泛化能力。这得益于本文框架中推理结构与语义组件的

解耦和有效组合,使得生成的查询语句既具备高可执行性,又能准确覆盖不同类型的问题。

表 2 GrailQA 数据集结果

(单位: %)

方法	独立同分布		组合		零样本		总体	
	EM	F_1	EM	F_1	EM	F_1	EM	F_1
ChatGPT	18.6	19.6	16.1	17.0	29.6	31.2	24.0	25.3
Davinci-003	22.3	23.5	20.8	22.0	34.5	36.4	28.5	30.1
GPT-4	23.7	25.0	19.5	20.6	37.1	39.2	30.1	31.7
KB-Binder	43.6	48.3	44.5	48.8	37.5	41.8	40.8	45.2
KB-Coder	43.6	49.3	44.0	49.6	37.7	43.2	40.9	46.3
DC-KBQA(ours)	47.9	49.6	46.9	49.1	49.0	50.8	48.0	50.4

如表 3 所示,在 WebSQL 数据集中,DC-KBQA 相比其他基于大语言模型的方法有一定提升,进一步验证了其在实际应用场景中的有效性。

表 3 WebSQL 数据集结果

方法	$F_1/\%$
ChatGPT	47.2
Davinci-003	49.6
GPT-4	51.2
KB-Binder	56.6
KB-Coder	60.5
DC-KBQA(ours)	61.3

如表 4 所示,在 GraphQA 数据集中,DC-KBQA 的表现再次验证了其优越性。与其他方法相比,DC-KBQA 的 F_1 得分最高,这表明 DC-KBQA 在处理图结构查询方面具有较强的能力。

表 4 GraphQA 数据集结果

方法	$F_1/\%$
ChatGPT	24.5
Davinci-003	25.7
GPT-4	30.8
KB-Binder	34.5
KB-Coder	35.8
DC-KBQA(ours)	37.5

总体而言,DC-KBQA 通过推理结构和语义组件的有效解耦和组合,有效提升了基于大语言模型的 KBQA 推理框架的性能。

3.3 消融实验

本文为模型设计了以下几种设置,以分析 DC-KBQA 各组件如何影响性能,实验结果如图 5 所示。

(1)采用 LLMs 直接生成的推理结构,而非从代码库中检索,命名为 LK-KBQA。在这一设置中,跳过了

从已有代码库中检索推理结构的步骤,直接让 LLMs 生成推理结构。实验结果显示,直接生成的推理结构在准确性上不如经过筛选的推理结构,导致整体性能有所下降。这表明代码库中的推理结构提供了更稳定和可靠的基础,有助于提高查询语句的质量。

(2)直接采用检索的推理结构,而不进行逻辑类型筛选,命名为 LT-KBQA。在这一设置中,直接使用从代码库中检索到的推理结构,而不再根据问题的逻辑类型进行二次筛选。结果表明,未经筛选的推理结构质量参差不齐,导致推理结果的准确性有所下降。由此可见,基于逻辑类型的筛选过程对于提升推理结构的质量和适用性至关重要。

(3)直接采用从知识库中提取的语义组件,不进行模式检查,命名为 SC-KBQA。该设置下,直接使用从知识库中提取的语义组件,而不进行本体和实例级别的模式检查。实验结果显示,未经模式检查的语义组件存在大量冗余和无效信息,导致查询语句的可执行性和准确性显著下降。这表明本体和实例级别的模式检查对于过滤出高质量的语义组件至关重要,有助于提高查询语句的可执行性。

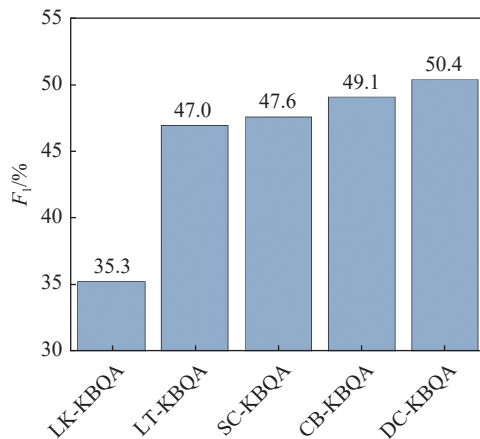


图 5 消融实验结果

(4)直接采用组合形成的查询语句,而非 LLMs 进行组合,命名为 CB-KBQA。在这一设置中,直接将推理结构和语义组件进行组合,而不再通过 LLMs 进行优化组合。结果表明,直接组合的查询语句在语义一致性和逻辑结构上存在较多问题,导致整体性能下降。由此可见,LLMs 在组合推理结构和语义组件过程中发挥了重要作用,有效提升了查询语句的整体质量。

3.4 参数分析

为了进一步优化 DC-KBQA 框架的性能,本文对以下参数进行了详细分析:推理结构数量、语义组件数量,如图 6 所示。

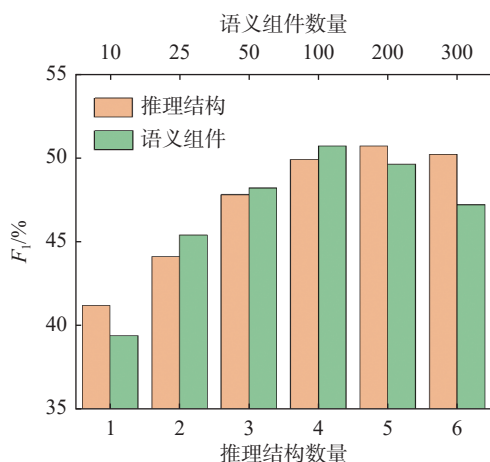


图 6 参数分析

(1)推理结构数量。本文研究了不同数量推理结构的设置下模型性能的变化趋势。实验结果表明,推理结构数量在一定范围内增加,可以显著提升模型的准确性和可执行性。然而,当数量超过一定阈值时,性能提升趋于平缓,甚至出现下降趋势。这是因为过多的推理结构增加了筛选和组合的复杂度,可能引入冗余信息。

(2)语义组件数量。本文对语义组件的数量如何影响模型的性能进行了实验研究,结果显示,适当增加语义组件数量可以提高查询语句的丰富性和准确性。然而,过多的语义组件会导致模式检查和组合过程中的计算开销增加,且容易引入噪声信息。以 GrailQA 数据集为例,当语义组件数量在 100 到 200 之间时,模型表现最佳。超过这个范围后, F_1 得分开始下降,表明过多的语义组件反而不利于查询语句的生成。

3.5 案例分析

以图 2 为例,现有方法如 KB-Binder,通常倾向于直接生成完整的逻辑表达式(语句 1),然而这种方式

往往忽略了推理步骤的细粒度分析,导致语义表达存在偏差,推理过程不够严谨。相比之下,本文提出的方法基于解耦与组合的思想,能够逐步细化推理路径,确保每一步推理都具备逻辑上的完整性,从而得到正确且完整的逻辑表达式(语句 2)。

4 结论

本文提出了基于解耦与组合的知识库问答推理框架 DC-KBQA,将知识库查询语句的生成任务分解为推理和语义两方面的检索与组合任务,从而降低了大语言模型的逻辑推理难度,提高了模型的泛化能力和查询准确性。尽管该框架在依赖大语言模型进行检索和组合任务时可能对计算资源需求较高,但未来的工作将致力于探索更轻量化的模型或高效的检索算法,以提升计算效率,减轻对资源的依赖。该框架为知识库问答领域提供了新的解决方案,为后续研究和应用奠定了基础。

参考文献

- [1] BERANT J, CHOU A, FROSTIG R, et al. Semantic parsing on freebase from question-answer pairs[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2013: 1533-1544.
- [2] YIH S W, CHANG M W, HE X, et al. Semantic parsing via staged query graph generation: Question answering with knowledge base[C]//Proceedings of the Joint Conference of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing of the AFNLP, 2015: 1321-1331.
- [3] LAN Y, JIANG J. Query graph generation for answering multi-hop complex questions from knowledge bases[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 969-974.
- [4] LAN Y, HE G, JIANG J, et al. A survey on complex knowledge base question answering: Methods, challenges and solutions[C]//Proceedings of the 30th International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, 2021: 4483-4491.
- [5] ZHANG Y, DAI H, KOZAREVA Z, et al. Variational reasoning for question answering with knowledge graph[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 6069-6076.
- [6] SU Y, SUN H, SADLER B, et al. On generating characteristic-rich question sets for QA evaluation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016: 562-572.
- [7] BROWN T, MANN B, RYDER N, et al. Language models

- are few-shot learners[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020, 33: 1877-1901.
- [8] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models[J/OL]. Transactions on machine learning research, 2022. <https://openreview.net/pdf?id=yzkSU5zdwD>.
- [9] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]// Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022, 35: 24824-24837.
- [10] GAO L, MADAAN A, ZHOU S, et al. PAL: Program-aided language models[C]//Proceedings of the 40th International Conference on Machine Learning. PMLR, 2023: 10764-10799.
- [11] CHEN W, MA X, WANG X, et al. Program of thoughts prompting: disentangling computation from reasoning for numerical reasoning tasks[J/OL]. Transactions on machine learning research, 2023. <https://openreview.net/forum?id=YfZ4ZPt8zd>.
- [12] POURREZA M, RAFIEI D. Din-SQL: Decomposed in-context learning of text-to-SQL with self-correction[C]// Proceedings of the 37th International Conference on Neural Information Processing System, 2024, 36: 36339-36348.
- [13] ZHANG L, ZHANG J, WANG Y, et al. FC-KBQA: A fine-to-coarse composition framework for knowledge base question answering[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 1002-1017.
- [14] NIE Z, ZHANG R, WANG Z, et al. Code-style in-context learning for knowledge-based question answering[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(17): 18833-18841.
- [15] CRESWELL A, SHANAHAN M, HIGGINS I. Selection-inference: exploiting large language models for interpretable logical reasoning[C]//Proceedings of the 11th International Conference on Learning Representations, 2023.
- [16] JI Z, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.
- [17] SHI J, CAO S, HOU L, et al. Transfernet: An effective and transparent framework for multi-hop question answering over relation graph[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 4149-4158.
- [18] CAO S, SHI J, YAO Z, et al. Program transfer for answering complex questions over knowledge bases[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 8128-8140.
- [19] CHEN S, LIU Q, YU Z, et al. ReTraCk: A flexible and efficient framework for knowledge base question answering[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, 2021: 325-336.
- [20] LI T, MA X, ZHUANG A, et al. Few-shot in-context learning on knowledge base question answering[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 6966-6980.
- [21] NASHID N, SINTAHA M, MESBAH A. Retrieval-based prompt selection for code-related few-shot learning[C]// Proceedings of the IEEE/ACM 45th International Conference on Software Engineering, IEEE, 2023: 2450-2462.
- [22] GU Y, DENG X, SU Y. Don't generate, discriminate: A proposal for grounding language models to real-world environments[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 4928-4949.
- [23] GU Y, KASE S, VANNI M, et al. Beyond iid: Three levels of generalization for question answering on knowledge bases[C]//Proceedings of the Web Conference, 2021: 3477-3488.
- [24] YIH W, RICHARDSON M, MEEK C, et al. The value of semantic parse labeling for knowledge base question answering[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 201-206.



施圣泽(2001—), 硕士, 主要研究领域为知识图谱与大模型。

E-mail: shishengze2023@iscas.ac.cn



胡军(1979—), 通信作者, 博士, 高级工程师, 主要研究领域为知识图谱。

E-mail: hujun@iscas.ac.cn



秦利(1975—), 通信作者, 硕士, 正高级工程师, 主要研究领域为智能信息系统。

E-mail: qinli_ams@163.com