

代码大语言模型赋能的知识图谱关键技术综述

李紫宣^{1,2}, 白龙^{1,2}, 任韦澄^{1,2}, 苏淼^{1,2}, 刘文轩^{1,2}, 陈磊³, 靳小龙^{1,2}

1. 中国科学院网络数据科学与技术重点实验室, 中国科学院计算技术研究所, 北京 100080;

2. 中国科学院大学计算机科学与技术学院, 北京 100080;

3. 中国工程院战略咨询中心, 北京 100088

摘要

传统知识图谱技术在将用自然语言表达的人类知识转化为用形式化语言表达的知识图谱再加以利用的过程中仍面临诸多挑战。近年来, 代码大语言模型具备了同时理解自然语言与形式化语言并将两者进行转化的能力, 有望为新一代知识图谱技术的发展带来重要突破。因此, 综述了代码大语言模型在知识图谱中的运用。首先, 从知识图谱构建、推理以及问答3个方面, 对代码大语言模型赋能的知识图谱关键技术进行了体系化梳理; 其次, 围绕上述3个方面, 对现有相关技术进行了较为详细的介绍; 最后, 对代码大语言模型赋能的知识图谱关键技术进行了总结与展望。未来基于编程语言的知识表示有望在知识图谱上实现更加高效、自动且复杂的操作, 实现知识编程。

关键词

知识图谱; 代码大语言模型; 大语言模型

中图分类号: TP391.1

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2025022

Review of key technologies in knowledge graphs powered by code large language models

LI Zixuan^{1,2}, BAI Long^{1,2}, REN Weicheng^{1,2}, SU Miao^{1,2}, LIU Wenxuan^{1,2}, CHEN Lei³, JIN Xiaolong^{1,2}

1. Key Laboratory of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China

2. School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100080, China

3. Center for Strategic Studies, Chinese Academy of Engineering, Beijing 100088, China

Abstract

Traditional knowledge graph technologies still face significant challenges in converting human knowledge, expressed in natural language, into a formal language-based knowledge graph and utilizing it effectively. In recent years, code large language models (LLMs) have demonstrated remarkable capabilities in understanding both natural and formal languages, as well as in translating between them. These advancements are expected to drive significant breakthroughs in developing next-generation knowledge graph technologies. This paper reviews the application of code LLMs in KGs. Firstly, this paper systematically analyzes the role of code LLMs in enhancing key knowledge graph technologies across

three critical areas: construction, reasoning, and question-answering. Secondly, a relatively detailed introduction to the existing methodologies in these areas is provided. Finally, this paper summarizes the current state in the field and offers insights into the future of knowledge graph technologies empowered by code LLMs. In the future, knowledge representation based on programming languages is expected to enable more efficient, automated, and complex operations on knowledge graphs, realizing knowledge programming.

Key words
knowledge graph, code large language model, large language model

0 引言

知识图谱以图结构的形式描述客观世界中的事物以及事物之间的复杂关系。知识图谱技术是对知识图谱中的结构化知识进行表示、获取、存储、检索、推理与问答的技术体系，包括知识的表示、抽取、推理、问答等。在该体系中，形式化语言作为精确表达知识与逻辑推理的关键工具，发挥着至关重要的作用。例如：人们利用形式化语言（如网络本体语言）可以定义实体、属性及其之间的关系，为知识图谱提供清晰的语义框架；通过形式化语言（如一阶谓词逻辑）实现基于规则的推理，使知识图谱能够从现有知识中推导出新知识，从而增强知识的可用性和可靠性；通过标准化的查询语言（如 SPARQL、Cypher 等），机器可以理解用户问题并检索相关知识，从而实现智能问答。同时，自然语言作为人类表达与传递知识的主要方式，既是知识图谱的主要知识来源，也是人与知识图谱交互的核心媒介。如何从自然语言中提取人类知识并以知识图谱的形式存储，以及如何利用知识图谱解决自然语言问题，是当前知识图谱技术面临的关键挑战。

随着深度学习的快速发展，国内外在

知识图谱技术上取得了显著进展。例如：基于深度学习的文本表示模型被用于实现更精准的知识抽取；深度学习模型与逻辑规则相结合，提升了知识推理的准确性；深度学习模型（如 Seq2Seq 模型）可用于理解自然语言问题，并将其解析为形式化语言查询（如 SPARQL），从而在知识图谱中执行并获取答案。受限于传统深度学习模型在自然语言理解与形式化语言理解方面的能力，知识图谱技术仍面临诸多挑战。近年来，随着大语言模型的发展，算法在自然语言理解能力上取得了突破性进展。在此基础上，代码大语言模型通过自然语言与形式化语言（如代码）的联合训练，能够同时理解自然语言与形式化语言，并具备将两者进行转化的能力，为知识图谱技术带来了新的发展机遇。如何利用代码大语言模型强大的自然语言与形式化语言的理解、生成及相互转化的能力，进一步提升知识图谱相关技术（如知识抽取、推理以及问答技术）的效果，成为当前研究的重要方向。针对上述问题，本文首先介绍代码大语言模型赋能的知识图谱关键技术体系，然后综述代码大语言模型在知识图谱构建、推理以及问答技术上的具体运用，最后对代码大语言模型与知识图谱技术深度融合的未来趋势进行展望。

1 代码大语言模型赋能的知识图谱关键技术体系

本文主要从知识图谱的构建、推理以及问答3个方面介绍代码大语言模型如何赋能知识图谱关键技术。如图1所示，在知识图谱构建方面，代码化的知识表示通过代码类的形式描述知识图谱中的各类概念及其相互关系，进而在本体层实现对知识的精准刻画。在该表示下，实例知识被进一步表示为相应类的实例化代码，通过代码大语言模型基于给定文本编写相应类的实例化代码，以实现代码化的知识抽取。在知识图谱推理方面，尽管目前尚无直接使用代码大语言模型对其赋能的研究，但已有相关工作通过代码大语言模型自动生成代码来实现图上的算子调用，从而实现推理。在知识图谱问答方面，代码大语言模型的编程能力为问答技术提供了新的范式。通过少量样例，代码大语言模

型能够将自然语言问题转化为形式化编程语言，并在知识图谱上执行得到答案。这种基于代码大语言模型的问答范式，显著提升了问答系统的灵活性与准确性。

2 代码大语言模型赋能的知识图谱构建技术

本节分别介绍代码大语言模型对知识图谱构建过程中的知识表示、抽取与融合方面的赋能。

在知识表示方面，可以通过大模型训练获得更好的知识图谱的表示，以提升知识图谱下游任务的性能。Code4Struct^[1]采用Python类定义事件类型，并通过类成员变量定义参与事件的论元类型。KnowCoder^[2]在Code4Struct的基础上进一步定义了实体类型以及实体类型之间的关系类型，并通过类继承描述概念之间的上下位关系，通过初始化函数中的变量类型

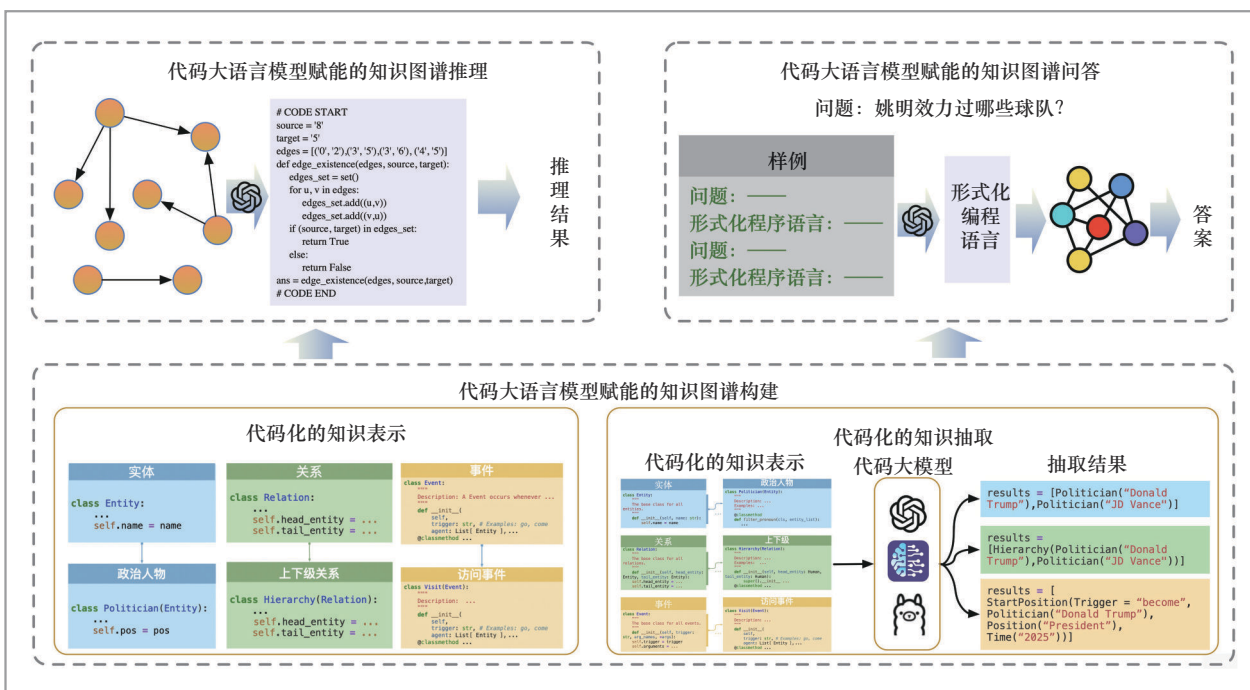


图1 代码大语言模型赋能的知识图谱关键技术体系

约束概念之间的关系，还为每一类实体定义了独特的类方法。CodeTaxo^[3]通过将实体类型定义为 Python 类的形式，并且通过增加类方法 add_parent() 完成实体类型归纳任务。

在知识抽取方面，可以通过形式化语言（如代码表示实体、关系等）将实体识别、关系抽取等任务转化为代码生成任务。该技术主要包括基于语境学习的方法与基于有监督微调的方法。早期的方法通过语境学习，基于人工构建的抽取样例，使模型学习抽取相关类型的知识。这类方法无须训练但依赖于大模型的上下文学习能力，目前主要基于闭源的通用大模型（如 ChatGPT^[4]等）。对于实体抽取任务，CodeIE^[5]通过定义抽取函数实现实体抽取，该函数的输入参数为待抽取的文本，其利用注释进行抽取任务描述，并将抽取结果定义成字典。对于事件抽取任务，Code4Struct^[1]提出以 Python 类的形式表示事件类型，进而将事件抽取任务转化为代码生成任务。Code4Struct 首先定义了基类“Class Event()”以及“Class Entity()”，并将事件类别以及实体类别定义为相应 Python 继承类，如 Class Transport(Event) 以及 Class Vehicle(Entity)，同时将参数角色（如 Transport 的 agent 论元）作为类初始化函数的输入参数。结果表明，在 50-shot 的情况下，Code4Struct 的性能甚至超过了基于全部数据微调的模型 Bart-Gen^[6]等。对于统一信息抽取任务，Code4UIE^[7]为不同的抽取任务设计了统一的代码格式。同时，它还利用检索增强技术选取与当前抽取任务语义相关度更高的样例，让模型更好地理解当前任务。Code4UIE 是当前基于语境学习的方法中的效果最好的方法。基于有监督微调的方法在知识抽取训练数据方面对大模型进行

微调，以取得更好的抽取效果。比如，GoLLIE^[8]为每一个待抽取概念设计了一个 Python 类，并在类注释中提供相应概念的标注指南，通过训练提升大模型遵循标注指南理解概念并抽取相应实体的能力。KnowCoder^[2]设计了一种更完备的、基于 Python 类的知识表示方法，通过类继承、类注释、类方法等代码特性描述概念，进一步构建了一个包含 30 000 余种不同类型的大型概念库。此外，为了更好地让模型理解概念并遵循指令，其提出了理解-遵循两阶段训练框架，通过代码预训练进行概念知识的理解，并通过指令微调提升大模型遵循给定的概念抽取相应实例的能力。在大规模训练之后，该模型在零样本、监督设置以及低资源设置的情况下均取得了目前最好的效果。

在知识融合方面，主要思路是将知识图谱转换为代码，以便大型语言模型能够更有效地理解待融合的实体信息，从而提升知识融合任务的性能。例如，ChatEA^[9]通过引入一个知识图谱代码翻译模块，将知识图谱转换为大语言模型容易理解的代码块。这一转换有效克服了以往的实体对齐方法过度依赖实体嵌入的问题，充分利用了大语言模型丰富、广泛的背景知识生成实体描述注释，显著提高了实体对齐任务的准确性。

3 代码大语言模型赋能的知识图谱推理技术

当前，利用代码大语言模型赋能知识图谱推理技术的研究处于早期阶段，主要分为两类较为相关的研究：大模型赋能的知识图谱推理与代码大语言模型赋能的图推理。

在知识图谱推理研究方面，传统的神经网络方法将知识图谱推理视为判断三元

组真假的分类过程，与此相对的生成式方法将知识图谱推理视为序列到序列的生成过程。例如，GenKGC^[10]将头实体、关系作为输出，基于编码器-解码器的预训练语言模型生成尾实体。在大模型被提出之后，也有相关研究尝试使用这类生成式方法进行知识图谱推理。

当前，使用大模型赋能知识图谱推理的技术主要分为两类，一类基于闭源大模型进行语境学习，另一类采用开源大模型进行指令微调。基于语境学习的方法将与推理任务相关的事实作为上下文输入大模型中，直接利用大模型的语境学习能力让模型输出答案。例如，KICKGC^[11]将知识图谱的推理过程分为候选实体检索和重排序两个阶段，在第一阶段使用检索器给所有候选答案实体打分，取打分最高的 N 个候选实体，使用大模型对其进行重排序。在知识图谱时序推理方面，Lee 等^[12]将与查询相关的历史候选事实依序排列，并将查询添加在候选事实之后，利用大模型生成缺失的答案实体。基于指令微调的方法通过构造包含查询及推理答案的训练样本微调大模型。例如，KG-LLM^[13]探索了 3 类知识图谱推理任务形式，分别是三元组分类、关系预测和实体预测。其中，三元组分类被转化为是非问答题，关系预测被转化为多项选择题，实体预测被转化为问答或者续写的形式。KoPA^[14]利用传统知识图谱方法得到实体和关系的低维向量表示，从而编码知识图谱的结构信息，并将其融入大模型中。此类方法目前主要使用自然语言描述知识图谱进行推理，如何利用代码的形式化表示与程序执行的能力还有待探索。此外，如何利用代码大语言模型增强多元知识图谱推理，也有待进一步的研究。

在图推理研究方面，近年来已经有诸多工作将大模型用于图推理^[15-16]，但使用代码

大语言模型进行图推理的工作相对较少。Code4Struct^[1]是较早利用代码大语言模型实现图结构预测的方法，然而该方法主要关注图结构的理解，并未过多涉及图推理。CodeGraph^[17]首次提出将代码大语言模型用于图推理，实现图上的节点数统计、边数统计、边存在性查询、节点度数计算、回路检测、邻居节点查询等。然而，目前此类方法主要关注同构图，该方法对节点与边类型极为丰富的异构图（如知识图谱）的效果有待优化。此外，与一般的图推理与计算不同的是，知识图谱推理还包括语义推理，即两个实体之间的语义关系可以通过多个关系复合而成的路径表示，如何使用代码大语言模型实现此类推理还有待探索。

综上所述，知识图谱推理相关研究虽然已经开始探索大模型相关技术，但是如何利用代码大语言模型赋能知识图谱推理，仍是一个亟待研究的问题。

4 代码大语言模型赋能的知识图谱问答技术

知识图谱问答技术旨在根据知识图谱回答自然语言问题，其中，基于语义解析的方法通过将自然语言转化为形式化语言表达的知识图谱查询，并在知识图谱上执行并获取答案。代码大语言模型同时具备自然语言与形式化语言理解与生成能力，有望赋能该方法。按照使用的形式化语言类型，知识图谱问答技术可以分为基于编程语言生成的知识图谱问答技术与基于逻辑形式生成的知识图谱问答技术。

4.1 基于编程语言生成的知识图谱问答技术

考虑到代码大语言模型主要基于编程

语言进行预训练和微调，一些方法设计了基于编程语言的函数或接口作为知识图谱查询语言，引导大模型生成更精准的查询代码，从而提升知识图谱问答的整体效果，本文称之为基于编程语言生成的知识图谱问答技术。比如，Nie等^[18]提出了一种代码风格的上下文学习方法，使用代码大语言模型将问题转换为Python函数调用序列。生成的函数调用序列通过Python解释器转换为逻辑形式，用来从知识图谱中得到答案。TrustUQA^[19]设计了一种类代码的查询函数，通过代码大语言模型编写相应的程序，实现了更准确的查询函数调用，这种方法可以在小样本提示的情况下获得较高的准确性。TrustUQA还使用动态样例检索方法进一步提高提示的质量，从而提高准确性。

4.2 基于逻辑形式生成的知识图谱问答技术

相比于生成编程语言，直接生成SPARQL等逻辑形式语言是传统知识图谱问答任务中更常见的做法。为了增强大语言模型对特定逻辑形式语言的理解与生成能力，研究者们提出诸多策略。例如，KB-BINDER^[20]设计了一个两阶段问答框架，首先利用代码大语言模型的语境学习能力，为给定样例生成一个“初步”的逻辑形式（即草稿），然后将知识图谱中的实体和关系填充到这个草稿中，进而构建一个完整且可执行的逻辑形式。Liu等^[21]提出了一种基于语境学习的问答模式理解方法，通过设计一种有效的混合样例检索策略，检索与问题模式相关的问题标注对，并利用大模型语境学习能力生成SPARQL查询。为了进一步降低SPARQL查询生成的错误率，D'Abramo等^[22]提出了一种动态小样本学习方法，利用语义搜索从训练

集中检索类似问题并丰富提示。Prog-TQA^[23]在KoPL算子的基础上增加了时序算子，从而实现对时序问题的解析。Prog-TQA还提出了一种迭代标注训练策略，首先利用大模型的语境学习能力基于样例编写程序，再执行程序得到结果，与正确答案进行比对，从而筛选出正确的程序作为训练数据训练模型，迭代上述过程实现大模型自进化。

除了选择不同的逻辑形式语言，研究人员还采用了模型训练和自我矫正等方式改进生成效果。为了应对知识图谱中实体数量多和逻辑形式语言多样的问题，Li等^[24]提出了FlexKBQA，首先通过与知识图谱的多轮交互生成多样化的程序，并基于程序生成自然语言问题；其次，在该部分数据上训练逻辑语言生成模型，并基于该模型标注用户问题；最后，利用标注的问题对模型进行继续训练。这种方法不仅提高了模型的泛化能力，还有效解决了数据稀缺的问题。KBLLaMA^[25]引入了生成逻辑形式语言的思维链，通过微调将生成逻辑形式语言的过程知识内化到模型中，从而提高代码大语言模型在知识图谱问答任务上的泛化能力。FuSIC-KBQA^[26]提出了一种逻辑形式语言生成的自校正机制，能够根据执行逻辑形式的结果为大语言模型提供反馈，提示其进行自我纠正。该机制不仅提高了模型的准确性，还增强了系统的鲁棒性。

5 结束语

本文综述了代码大语言模型在知识图谱技术中的应用及其发展趋势。首先，简要介绍了代码大语言模型赋能的知识图谱关键技术体系；其次，详细阐述了现有代

码大语言模型在提升知识表示、抽取、融合、推理和问答等方面的应用。通过对相关工作进行综述，可以看出编程语言一方面能够作为一种精准的知识表示语言，另一方面能够作为一种修改与操作知识的指令，有望在知识图谱上实现更加高效、自动且复杂的操作，笔者称之为面向知识的编程，即知识编程。然而，目前大多数工作聚焦于使用编程语言对结构化知识进行表示与抽取，以及使用编程语言对知识进行简单计算（如知识查询、统计等）。如何基于知识编程思想完成更复杂的知识获取与计算有待探索。例如：当知识图谱中触发了代表两人结婚的事件代码时，系统将自动更新两人的婚姻状态属性；触发美联储降息的事件代码时，将引发一系列相关事件的连锁反应，并在知识图谱中逐一执行这些事件，以反映最新的经济动态。虽然代码大语言模型在知识图谱问答方面已有一系列工作，但其主要聚焦于静态知识图谱问答，如何将代码大语言模型应用在更复杂的问答场景（如数值问答、时序问答）也是值得探索的方向。

参考文献：

- [1] WANG X Y, LI S, JI H. Code4Struct: code generation for few-shot event structure prediction[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 3640–3663.
- [2] LI Z X, ZENG Y T, ZUO Y X, et al. KnowCoder: coding structured knowledge into LLMs for universal information extraction[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2024: 8758–8779.
- [3] ZENG Q K, BAI Y Y, TAN Z X, et al. CodeTaxo: enhancing taxonomy expansion with limited examples via code language prompts[EB]. arXiv preprint, 2024, arXiv: 2408.09070.
- [4] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2022: 27730–27744.
- [5] LI P, SUN T X, TANG Q, et al. CodeIE: large code generation models are better few-shot information extractors[EB]. arXiv preprint, 2023, arXiv: 2305.05711.
- [6] LI S, JI H, HAN J W. Document-level event argument extraction by conditional generation[EB]. arXiv preprint, 2021, arXiv: 2104.05919.
- [7] GUO Y C, LI Z X, JIN X L, et al. Retrieval-augmented code generation for universal information extraction[EB]. arXiv preprint, 2023, arXiv: 2311.02962.
- [8] SAINZ O, GARCÍA-FERRERO I, AGERRI R, et al. GoLLIE: annotation guidelines improve zero-shot information-extraction[EB]. arXiv preprint, 2023, arXiv: 2310.03668.
- [9] JIANG X H, SHEN Y H, SHI Z C, et al. Unlocking the power of large language models for entity alignment[EB]. arXiv preprint, 2024, arXiv: 2402.15048.
- [10] XIE X, ZHANG N Y, LI Z B, et al. From discrimination to generation: knowledge graph completion with generative transformer[C]//Proceedings of the Web Conference 2022. New York: ACM, 2022: 162–165.
- [11] WEI Y B, HUANG Q S, ZHANG Y, et al. KICGPT: large language model with knowledge in context for knowledge graph completion[C]//Proceedings of the Findings of the Association for Compu-

- tational Linguistics: EMNLP 2023. Stroudsburg: ACL, 2023: 8667–8683.
- [12] LEE D H, AHRABIAN K, JIN W, et al. Temporal knowledge graph forecasting without knowledge using in-context learning[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2023: 544–557.
- [13] YAO L, PENG J Z, MAO C S, et al. Exploring large language models for knowledge graph completion[EB]. arXiv preprint, 2023, arXiv: 2308.13916.
- [14] ZHANG Y C, CHEN Z, GUO L B, et al. Making large language models perform better in knowledge graph completion[EB]. arXiv preprint, 2023, arXiv: 2310.06671.
- [15] TANG J B, YANG Y H, WEI W, et al. GraphGPT: graph instruction tuning for large language models[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2024: 491–500.
- [16] PEROZZI B, FATEMI B, ZELLE D, et al. Let your graph do the talking: encoding structured data for LLMs[EB]. arXiv preprint, 2024, arXiv: 2402.05862.
- [17] CAI Q L, WANG Z W, DIAO S Z, et al. CodeGraph: enhancing graph reasoning of LLMs with code[EB]. arXiv preprint, 2023, arXiv: 2408.13863.
- [18] NIE Z J, ZHANG R C, WANG Z Y, 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.
- [19] ZHANG W, JIN L, ZHU Y S, et al. TrustUQA: a trustful framework for unified structured data question answering[EB]. arXiv preprint, 2024, arXiv: 2406.18916.
- [20] LI T L, MA X G, ZHUANG A, et al. Few-shot in-context learning for knowledge base question answering[EB]. arXiv preprint, 2023, arXiv: 2305.01750.
- [21] LIU Y T, LI Z X, JIN X L, et al. An in-context schema understanding method for knowledge base question answering [M]//Knowledge science, engineering and management. Singapore: Springer, 2024: 419–434.
- [22] D’ABRAMO J, ZUGARINI A, TORRONI P. Dynamic few-shot learning for knowledge graph question answering[EB]. arXiv preprint, 2024, arXiv: 2407.01409.
- [23] CHEN Z, ZHANG Z, LI Z X, et al. Self-improvement programming for temporal knowledge graph question answering[EB]. arXiv preprint, 2024, arXiv 2404.01720.
- [24] LI Z Y, FAN S Q, GU Y, et al. Flex-KBQA: a flexible LLM-powered framework for few-shot knowledge base question answering[EB]. arXiv preprint, 2023, arXiv: 2308.12060.
- [25] ZHANG L X, ZHANG J, WANG Y L, et al. A learn-then-reason model towards generalization in knowledge base question answering[EB]. arXiv preprint, 2024, arXiv: 2406.14763.
- [26] PATIDAR M, SAWHNEY R, SINGH A, et al. Few-shot transfer learning for knowledge base question answering: fusing supervised models with in-context learning[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2024: 9147–9165.

作者简介



李紫宣（1995–），男，博士，中国科学院计算技术研究所特别研究助理，主要研究方向为时序推理、知识图谱、大语言模型等。



白龙（1993–），男，博士，中国科学院计算技术研究所特别研究助理，主要研究方向为知识图谱与自然语言处理技术等。



任韦澄（1998–），男，中国科学院计算技术研究所博士生，主要研究方向为事件抽取、大模型、知识图谱等。



苏森（2000–），女，中国科学院计算技术研究所博士生，主要研究方向为知识图谱、时序知识图谱问答、大语言模型等。



刘文轩（2002–），男，中国科学院计算技术研究所博士生，主要研究方向为事件抽取、大语言模型等。



陈磊（1983–），男，中国工程院战略咨询中心副研究员，国防科技大学计算机学院博士生，主要研究领域为网络安全等。



靳小龙（1976–），男，中国科学院计算技术研究所研究员、博士生导师，中国科学院大学岗位教授，主要研究方向为知识图谱、知识计算、大数据知识工程等。

收稿日期: 2024-10-16
通信作者: 陈磊, cl@cae.cn; 靳小龙, jinxiaolong@ict.ac.cn
基金项目: 国家自然科学基金项目 (No. 62306299, No. 62406308)
Foundation Items: The National Natural Science Foundation of China (No. 62306299, No. 62406308)