计算机研究与发展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(2): 378−392, 2026

基于生成模型的开放域关系推理方法

刘井平[1] 岑哲栋[1] 张维彦[1] 阮 彤[1] 王 超[2] 王昊奋[3]
[1](华东理工大学信息科学与工程学院 上海 200237)
[2](上海大学未来技术学院 上海 200444)
[3](同济大学设计创意学院 上海 200092)
(jingpingliu@ecust.edu.cn)

Open Domain Relation Reasoning Method Based on Generative Model

Liu Jingping[1], Cen Zhedong[1], Zhang Weiyan[1], Ruan Tong[1], Wang Chao[2], and Wang Haofen[3]
[1](*School of Information Science and Engineering*, *East China University of Science and Technology*, *Shanghai* 200237)
[2](*School of Future Technology*, *Shanghai University*, *Shanghai* 200444)
[3](*College of Design and Innovation*, *Tongji University*, *Shanghai* 200092)

Abstract Relation reasoning, an important task in natural language processing, aims to infer possible semantic relations between two or more entities. The reasoning process typically involves deriving new relations from known relations between entities, and the results can be widely applied in various downstream tasks such as knowledge graph completion, relation extraction, and commonsense knowledge question answering. Previous studies often face two main limitations: First, they are primarily based on the closed-world assumption, meaning the relation types are predefined and difficult to expand. Second, even if some methods focus on open domains, they typically only handle 1-hop reasoning, which is insufficient for complex multi-hop reasoning scenarios. To address these issues, we propose and define an open domain 2-hop relation reasoning task and construct a dataset for evaluating this task. Furthermore, we introduce an open domain 2-hop relation reasoning framework, named ORANGE (open domain relation reasoning method on generative model), which includes 3 key modules: entity generation, relation generation, and result aggregation. Firstly, the entity generation module generates unknown entities. Secondly, the relation generation module proposes potential new relations. Finally, the result aggregation module integrates the outputs of the preceding modules to determine the final result. Experimental results demonstrate that, when compared with the best existing methods, our approach achieves a 10.36% improvement in the average score. Moreover, when employing ORANGE's 3-module relation reasoning framework with large language models, it surpasses the conventional in-context learning prompt strategy, showcasing a 9.58% enhancement in the average score.

Key words relation reasoning; generative model; open domain; multi-hop reasoning; large language model

摘 要 关系推理是自然语言处理中的一项重要任务,旨在预测2个或多个实体之间可能存在的语义关系,推理过程通常为从已知的实体间关系中推导出新的关系,得到的结果可以在多种下游任务如知识图谱补全、关系抽取、常识问答中得到广泛应用。以往的研究主要存在2个局限性:首先,以往的方法主要集中于封闭域,关系类型都是已经事先定义好的,难以扩展;其次,即便存在少量针对开放域的关系推理方法,

收稿日期:2024-10-10;**修回日期**:2025-04-17
基金项目:国家自然科学基金项目(62306112);上海市青年科技英才扬帆计划项目(23YF1409400);上海市基础研究特区计划项目(22TQ1400100-20)
This work was supported by the National Natural Science Foundation of China (62306112), the Shanghai Sailing Program (23YF1409400), and the Shanghai Pilot Program for Basic Research (22TQ1400100-20).

也仅聚焦于单跳推理，难以满足更复杂的场景需求。因此，定义了开放域的2跳关系推理任务，并构建了一个用于评估该任务的数据集。面向该任务，提出了一种基于生成模型的开放域关系推理方法ORANGE，包括实体生成、关系生成、结果聚合3个模块。实验结果表明，ORANGE相比现有主流关系推理方法在平均得分上提高了10.36%。此外，当ORANGE的关系推理框架与大语言模型结合使用时，相较于传统的上下文学习提示策略，平均得分提高了9.58%。

关键词 关系推理；生成模型；开放域；多跳推理；大语言模型

中图法分类号 TP391.1

DOI: 10.7544/issn1000-1239.202440795 **CSTR:** 32373.14.issn1000-1239.202440795

关系推理旨在预测2个或多个实体之间的可能关系，推理过程通常为从已知的实体间关系中推导出新的关系[1]。这一过程可以被视为一种逻辑归纳问题，形如

$$\exists y:(x,\text{works at},y)\wedge(y,\text{is located in},z)\rightarrow(x,\text{works in},z)。\tag{1}$$

关系推理得到的结果可以在多种下游任务如知识图谱补全[2-4]、关系抽取[5]、知识问答[6]中得到广泛应用。

目前，针对关系推理任务的研究大多聚焦于封闭域，这意味着所有可能的关系类型都是预先定义好的。封闭域内的关系推理方法主要可以归纳为三大类：1）基于规则的推理方法，这类方法的核心在于探索知识图谱内部所蕴含的规则，并强调通过自动化手段从知识图谱中学习到具有高可信度的规则，代表性的技术有AMIE[7]，AMIE+[8]，AnyBURL[9]，LoTus[10]，RulePreM[11]等；2）基于图结构的推理方法，该类方法将事实表述为带有标签的有向图，其中实体被视为图中的节点，而关系则被看作是连接这些节点的边，之后即可运用图论中的算法来执行知识图谱上的推理任务，代表性的技术包括PRA[12]，DeepPath[13]，MA-TPath[14]，SRGCN[15]，RKLE[16]等；3）基于表示学习的推理方法，其目标是寻找一种映射机制，将符号化的实体和关系转换成向量空间中的数值表示，以此来揭示实体间及关系间的潜在联系，此类方法的代表有TransE[17]，RotatE[18]，Trans4E[19]，CPConvKE[20]，T-Temp[21]等。

尽管这些方法在关系推理方面取得了较好的进展，但其存在显著的局限性，主要体现在以下2方面：1）这些方法通常基于预定义的关系类型，这导致它们无法识别或处理那些没有预定义的关系类型。这一特性限制了模型的灵活性和适应新情况的能力。2）封闭域内的关系推理高度依赖于外部资源，例如知识图谱，这种依赖不仅影响了方法的独立性，而且当知识图谱更新或发生变化时，模型往往需要重新训练。这种对特定外部数据源的强依赖性，进一步削弱了模型的泛化能力和可扩展性。

因此，本研究重点讨论开放域关系推理问题。在本文中，开放域关系推理被定义为从给定关系推导出新关系的过程，其中给定关系和生成的新关系不受现有知识图谱中预定义关系类型的限制，可以为任何符合自然语言描述的关系。例如，当输入为$(x,\text{works at},y)$和$(y,\text{is located in},z)$时，能够推导出新关系$(x,\text{lives in},z)$。同样地，当输入改为$(x,\text{is working at},y)$和$(y,\text{is a place in},z)$，也能推出新关系$(x,\text{is living in},y)$。为了保证模型生成任意符合自然语言描述的关系，本研究将开放域的关系推理问题建模为生成问题，即给定2个关系，生成能概括它们的新关系。此外，为了摆脱对外部知识（如知识图谱）的依赖，本研究重点关注如何通过生成式语言模型来解决开放域关系推理问题，因为语言模型可以被认为是知识库，存储了丰富的基础知识[22]。类似的研究成果包括Orion[23]。Orion方法接收前提关系作为输入，并能输出从这些前提关系中推断出的假设关系。例如，给定$(x\text{ is founder of},y)$作为输入，Orion可以推断出$(x,\text{ is CEO of},y)$也可能是正确的。这类方法的出现和发展，标志着开放域关系推理技术的重要进步。

然而，Orion方法存在一些局限性，具体表现在以下2个方面：1）多跳推理能力不足。Orion主要集中在单跳关系的推理上，即假定2个实体之间已经存在直接的关联关系，并基于这种直接关系来推导出2个实体间其他可能的关系。但在实际应用场景中，多跳推理更加普遍，实体之间的联系往往不是直接的，而是通过一系列中间实体间接建立的。而Orion在处理此类复杂的多跳关系时显得有所局限。2）预训练语料库构建方法较为简单。Orion在构建预训练语料库时采用的方法相对简单，主要是通过对文本中的实体进行保留或删除操作来实现。这样的处理方式可能导致预训练语料库的质量不高，进

而影响模型的最终性能和准确性。

针对上述问题，本文提出了一种新的方法——基于生成模型的开放域关系推理方法(open domain relation reasoning method based on generative model, ORANGE)。与 Orion 类似，ORANGE 利用了预训练生成语言模型，这类模型在预训练阶段已经学习了复杂的语言结构和广泛的语义知识，可以看作是一个通用领域的知识库[22]，其具备强大的文本理解和文本生成能力，并且能够以自然语言的形式生成任何类型的关系描述，避免了以往的关系推理方法无法推理出预定义类型以外关系的局限性，使得关系推理任务从封闭域转向开放域成为了可能。此外由于预训练生成语言模型的容错性较高，即使输入的关系存在部分拼写错误，ORANGE 依然可以正常运行。同时与 Orion 相比，ORANGE 更加专注于解决多跳关系推理的问题。

引入生成语言模型也会带来一些挑战，由于生成语言模型生成的内容存在不确定性，因此很可能其生成的结果并非关系，而是代词、介词等其他内容。为了解决这个挑战，ORANGE 将开放域关系推理分为 3 个步骤并设计了对应的模块，同时构建了预训练语料库对模型进行微调，以约束模型的生成内容。实验结果证明，ORANGE 可以较好地解决相关挑战。

ORANGE 主要包含 3 个核心模块，分别为实体生成、关系生成、结果聚合。在第 1 个模块，本文设计了基于槽填充(slot filling)的实体生成策略。该策略接收 2 个前提关系作为输入，并输出满足这 2 个关系的 3 个实体及其概率，同时从这 3 个实体中选择第 1 个实体和最后 1 个实体形成一个实体对。在第 2 个模块中，利用生成模型填充 2 个实体间的槽标记及其概率，该标记表示每一对实体之间可能存在的潜在语义关系。最后，第 3 个模块负责汇总前 2 个模块生成的实体概率和关系概率，并通过一系列计算选择总概率最高的关系作为最终结果。此外，为了进一步提升模型的推理精度，本文构造了一套高效的预训练语料库。

本文的主要贡献如下：1)提出了一种专用于开放域的 2 跳关系推理方法 ORANGE，该方法包括实体生成、关系生成、结果聚合 3 个模块；2)构建了一套用于开放域 2 跳关系推理任务的预训练语料库，同时还构建了一个针对该任务的评估数据集；3)实验结果表明，ORANGE 在开放域 2 跳关系推理任务上表现出色，与现有的主流关系推理方法相比，ORANGE 的平均得分提高了 10.36%。特别地，通过将 ORANGE 的关系推理框架集成到大语言模型中，与传统的上下文学习提示策略相比，平均得分进一步提高了 9.58%。这些结果充分证明了 ORANGE 在开放域关系推理领域的有效性和优势。

1 相关工作

本文的相关工作可以分为以下 4 个类别：基于规则的关系推理方法、基于图结构的关系推理方法、基于表示学习的关系推理方法、基于语言模型的关系推理方法。

1.1 基于规则的关系推理方法

基于规则的推理是指通过自动化手段从知识图谱中挖掘出具有高可信度的规则。代表性工作包括 AMIE[7]，AMIE+[8]，AnyBURL[9]，LoTus[10]，RulePreM[11]。

AMIE[7] 是一个用于从不完全的知识图谱中进行关系推理的方法，包括规则生成、规则评估、规则修剪、规则扩展 4 个步骤，并通过置信度和支持度等指标确保规则质量。相比 AMIE，AMIE+[8] 在此基础上引入了类型信息和联合推理，提高了推理准确性，并通过优化机制提升了推理效率，使其适用于更大规模的知识图谱。AnyBURL[9] 采用自底向上的方法，从具体路径实例开始学习规则。与自顶向下方法不同，它不需要预先生成所有规则，因此对计算资源需求较低，适合资源有限的环境。LoTus[10] 结合了逻辑规则与文本表示以实现可解释的推理，并通过剪枝策略减少候选规则，降低推理复杂度和时间消耗。RulePreM[11] 通过构建关系 IO 图，并对关系进行编码，利用这些编码作为提示来编码实体。同时结合基于规则置信度的注意力机制，对候选实体进行打分预测。

基于规则的关系推理方法只能根据封闭域内的有限数量的规则进行推理，无法实现开放域中针对任意关系的推理。ORANGE 使用生成模型进行推理，可以利用生成模型中蕴含的丰富知识，不受封闭域内的规则限制，从而实现开放域的关系推理。

1.2 基于图结构的关系推理方法

基于图结构的推理方法是指将实体作为节点，关系作为节点之间的边，并利用图论算法实现图上的关系推理任务。代表性工作包括 PRA[12]，DeepPath[13]，MA-TPath[14]，SRGCN[15]，RKLE[16] 等。

PRA[12] 提出了一种基于路径约束的随机游走算法，通过加权组合多个路径来进行关系推理，并允许模型在不同上下文中灵活地使用不同的路径特征。DeepPath[13] 将路径学习过程建模为强化学习问题，并

提出了可控的关系推理方法。该方法通过策略梯度训练来优化推理路径的选择，提升路径的多样性和准确性。MA-TPath[14]使用2个智能体迭代地执行关系选择和实体选择。智能体使用长短期记忆网络从环境中捕获当前结果，并通过激活函数输出动作向量。同时，考虑到推理路径的多样性，设计了新的奖励函数。SRGCN[15]将实体间的多条路径视为子图序列，通过图卷积网络来更新实体嵌入，并整合不同路径的结构信息，以提升推理效果。RKLE[16]通过构建具有额外节点的逻辑结构，并采用逻辑奖励调整来帮助智能体选择更合理的路径，从而提升推理效率和可解释性。

基于图结构的推理方法需要利用额外的图结构进行推理，且推理的关系受限于图结构中已经定义好的关系类型。而ORANGE使用生成模型进行推理，无需依赖外部知识图谱，同时推理的内容也不受图结构限制，能够更好地应用于开放域关系推理问题。

1.3 基于表示学习的关系推理方法

基于表示学习的推理方法是指将符号化的实体和关系转换成向量空间中的数值表示，以揭示实体间及关系间的潜在联系。代表性工作包括TransE[17]，RotatE[18]，Trans4E[19]，CPConvKE[20]，T-Temp[21]等。

TransE[17]的核心思想是如果一个三元组(h,l,t)成立，那么头实体h的嵌入向量加上关系l的嵌入向量应该接近于尾实体t的嵌入向量。该方法简洁且适用于大规模知识图谱。与TransE不同，RotatE[18]将关系视为在复数空间中的旋转，头实体h和尾实体t的嵌入表示为复数向量，关系l则表示为一个旋转操作，使得h通过l旋转后接近t，从而更自然地捕捉关系的对称性和自反性。Trans4E[19]构建了一种新颖的嵌入模型，通过在四元数空间中进行操作，为每个头实体生成一个独特的嵌入向量，大大增加了向量空间的容量。CPConvKE[20]通过基于门控机制的路径嵌入方法和多尺度动态卷积滤波器来减少噪声对于嵌入学习的影响，从而增强知识图谱结构的表示能力。T-Temp[21]基于经典的张量分解技术，将时态知识图谱视为一个3阶张量，能够学习实体、关系、时间戳的嵌入表示，有效提高推理的准确性和效率。

基于表示学习的关系推理方法依赖于特定知识图谱进行训练，且训练结果仅能应用于该图谱，无法扩展到其他知识图谱，且可推理的关系类型也受限于原图谱定义的类型。与此不同，ORANGE通过预训练生成模型进行关系推理，可应用于任何领域的知识图谱，并且能够推理任何符合自然语言描述的关系，不受现有知识图谱的限制。

1.4 基于语言模型的关系推理方法

早期基于语言模型的推理是指利用预训练语言模型作为基础模型，并结合现有知识库（如Wikidata[24]、DBpedia[25]、Freebase[26]和维基百科）进行关系推理。代表性工作包括SPN[27]，COMET[28]，UIE[29]，Orion[23]等。SPN[27]将关系推理视为关系集合预测问题，并使用非自回归并行解码网络进行预测。COMET[28]基于已有知识图谱的种子知识，训练了一个基于Transformer的生成模型，并基于此模型来预测新知识。UIE[29]统一编码了包括关系推理在内的多个任务，并能够自适应地生成针对性的结构，以便从不同的知识源进行协同学习。Orion[23]提出了开放域关系推理的概念，允许推理出的关系类型不受预定义类型的限制。

随着具备强大文本生成能力的大语言模型（如GPT，LLaMa，Qwen，GLM）的出现，研究人员开始探索如何利用这些模型进行关系推理任务。具有代表性的工作包括KICGPT[30]、情景化蒸馏（contextualization distillation）[31]、CP-KGC[32]和KoPA[33]等。KICGPT[30]将大语言模型作为额外的知识库，利用其丰富的内部知识来弥补长尾实体信息的不足，并将关系推理任务形式化为对给定序列的重排任务。这种方法避免了大语言模型输出不受约束的问题，且无需对大语言模型进行额外的训练或微调。情景化蒸馏[31]是指从大语言模型中提取每个实体和三元组的描述性上下文，并通过上下文重构和上下文情景化这2个辅助任务来帮助关系推理模型学习这些丰富的上下文信息。CP-KGC[32]通过设计适应不同数据集的提示词，增强语义丰富性，并采用上下文约束策略，有效识别知识图谱数据集中的多义词，减少语义歧义带来的影响，从而提高关系推理模型的性能。KoPA[33]将子图结构、关系模式等结构信息整合到大语言模型中，使模型能够进行结构感知的推理。同时，通过结构嵌入预训练和知识前缀适配器实现了文本嵌入和结构嵌入之间的跨模态交互。

早期的基于预训练语言模型的关系推理仍聚焦于封闭域的关系推理任务，未能充分挖掘和利用语言模型的生成能力，从而限制了其在复杂场景中的应用潜力。ORANGE主要面向开放域的关系推理，显著提升了模型在开放环境下的适用性和适用范围。此外，目前仅有的关注开放域关系推理的研究方法Orion只针对单跳关系推理进行了探索，而ORANGE则实现了开放域内的多跳关系推理，为复杂场景的关系推理提供了新的研究思路。

2 任务说明

本节首先定义开放域2跳关系推理任务，然后介绍为解决该任务而设计的方法框架。

2.1 任务定义

开放域2跳关系推理任务的定义如下：输入2个关系r_1和r_2，假设存在实体e_a，e_b，e_c使得三元组(e_a,r_1,e_b)和(e_b,r_2,e_c)同时成立。输出一个新的关系r_3，使得三元组(e_a,r_3,e_c)成立。使用公式表示为

$$(e_a,r_1,e_b),(e_b,r_2,e_c)\rightarrow(e_a,r_3,e_c)。\tag{2}$$

式(2)可以简化为

$$r_1,r_2\rightarrow r_3。\tag{3}$$

即从2个给定的关系r_1和r_2推导出1个新关系r_3，使得r_3等价于r_1和r_2同时成立的组合。

2.2 方法框架

本文将$r_1,r_2\rightarrow r_3$的复杂关系推理问题转化为2阶段的实体生成和关系生成任务，并提出基于生成模型的预测策略来完成这2个任务，进而实现开放域的关系推理。而大部分现有的关系推理方法最终实现输出时通常使用一个分类模型，使得结果被限制在预定义好的关系类型中，无法做到真正的开放域关系推理。此外，尽管如BERT这样的掩码语言模型同样可以预测出句子中的缺失信息，但其通常只能预测单个字符，而生成模型则可以自己决定生成内容的长度，现实世界中的实体和关系往往长度是不确定的，因此生成模型是更加合适的选择。

如图1所示，ORANGE方法的框架主要由3个模块组成：实体生成、关系生成、结果聚合。在实体生成模块，本文提出了基于槽填充的实体生成策略。首先，给定2个前提关系，实体生成模块将2个关系转换为符合语法语义的自然语言句子。该句子中除了关系外，还包含3个掩码槽标记表示对应位置可能的实体名称。随后，实体生成模块利用生成模型填充这些句子中的掩码槽标记，从而推导出可能的3个实体及其概率。接着，从中提取第1个和最后1个实体形成一个新的句子，其中包含的1个掩码槽标记代表2个实体之间可能存在的关系。这个新的句子被输入到关系生成模块中，该模块同样利用生成模型填充掩码槽标记，从句子中推导出实体间潜在的新关系及其概率。然后，将前2个模块推导出的所有实体、关系及其概率送入结果聚合模块。结果聚合模块会计算出每个关系的总概率，并应用聚类操作过滤掉语义上高度相似的关系。最后，模块选择总概率最高的前k个关系作为最终输出。

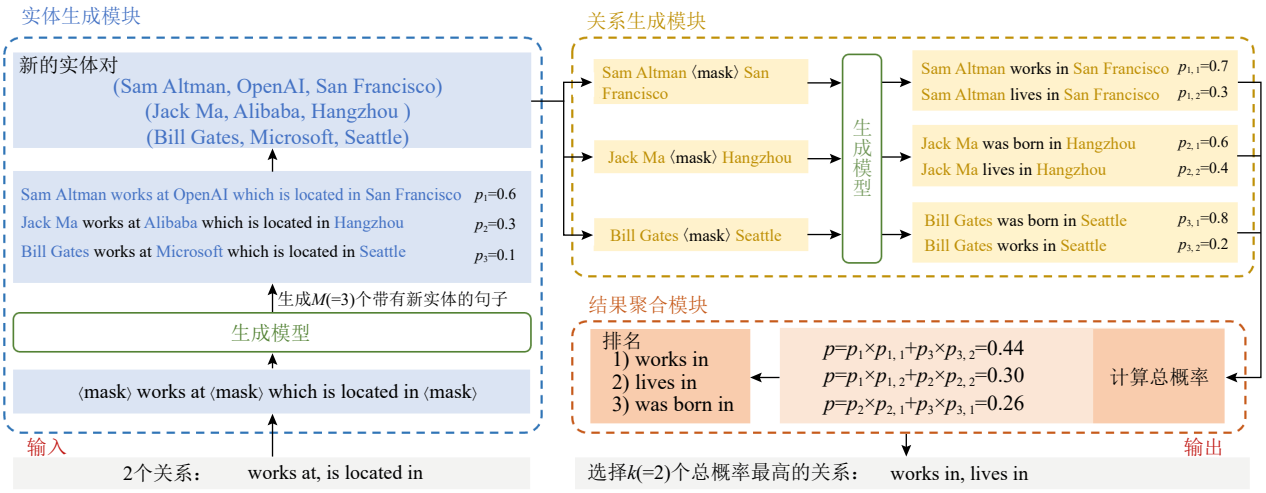


Fig. 1 Open domain relation reasoning method based on generative model

图1 基于生成模型的开放域关系推理方法

3 方 法

本节详细介绍了ORANGE框架中的3个关键模块：实体生成、关系生成、结果聚合的实现过程。此外，还探讨了构建预训练语料库时涉及的构建方法和相关考虑因素。

3.1 实体生成

3.1.1 输入输出

在实体生成模块中，目的是基于2个输入的关系生成2个实体对，使这些实体和关系能够组成2个合理的三元组，并且第1个三元组的尾实体与第2个三

元组的头实体相同，即确保生成的实体对能够合理地连接起来，以形成 1 个多跳关系推理路径。例如，给定输入的 2 个关系(works at, is located in)，输出可能是(Sam Altman, OpenAI)和 (OpenAI, San Francisco)这 2 个实体对，它们满足“Sam Altman works at OpenAI”和“OpenAI is located in San Francisco”的条件。

3.1.2 详细过程

具体来说，实体生成包含 2 个子步骤：句子构建和掩码槽填充。

1)句子构建。模块首先利用预定义好的句子模板将输入的关系转换为自然语言句子，其中包含掩码槽标记“〈mask〉”用于表示句子中的实体。例如，3.1.1 节中的案例将被转换为“〈mask〉works at〈mask〉which is located in〈mask〉”。需要注意的是，句子模板会为第 2 个关系引入一个关系代词(如 which, where, who 等)，因为省略它会导致语法错误，从而影响语言模型的理解效果。

2)掩码槽填充。使用预训练生成语言模型生成句子中掩码槽标记对应的内容。也就是说，在给定的关系 r_1 和 r_2 下，模型识别出最有可能满足这些关系的 3 个实体(记作 $T_e=(e_a,e_b,e_c)$)，从而最大化概率 $P((e_a,e_b,e_c)|r_1,r_2)$，这一过程可形式化地表示为

$$T_e=\arg\max_{(e_a,e_b,e_c)}P((e_a,e_b,e_c)|r_1,r_2)。\tag{4}$$

3.1.3 预训练语料库的构建

尽管预训练生成语言模型具有强大的生成能力，但直接使用模型来预测句子中的掩码槽标记可能无法在关系的两侧生成具体的实体名称，而可能生成代词或其他缺乏明确指代性的词汇，如“He works at a company which is located in a city”。为了确保语言模型生成具有明确指代的具体实体名称，本文构建了预训练语料库，以增强模型生成具体实体名称的能力。

在预训练语料库的构建过程中，本文设计了 3 种策略。1)从 Wikidata[24], FB15k-237[34], DBpedia[25], Yago3, WordNet[35] 等知识图谱中提取满足如下条件的 2 个关系三元组形成 1 对：1)第 1 个三元组的尾实体应为第 2 个三元组的头实体。然后将这些三元组对通过 3.1.2 节描述的预定义句子模板转换为自然语言描述的句子。2)利用描述多跳实体关系的自然语言句子的数据集，如 WebNLG[36]。具体而言，从 WebNLG 中选择描述 2 跳关系的句子，这些句子的形式非常类似于前文所述的句子模板。3)直接从维基百科中搜索描述 3 个实体间的 2 跳关系的句子。具体而言，首先使用命名实体识别工具对维基百科中的句子进行实体识别，过滤出恰好包含 3 个实体的句子，可以近似地认为这类句子描述了实体之间的 2 跳关系。然而，通过这种方法识别出的许多句子很可能只是在句子中提到了实体而并没有表示它们之间的关系，数据质量较低。为了提高数据质量，还会在以上数据的基础上过滤出包含关系代词(如 which, who, whose)的句子，并且这些关系代词必须位于第 1 个和第 3 个实体之间，避免得到以这些词语开头的疑问句。通过上述方法过滤出的句子有很大概率是一个定语从句，能贴切地描述 3 个实体之间存在的 2 跳关系。

之后，上述句子中的实体名称将被掩码槽标记替换，得到类似于“〈mask〉works at〈mask〉which is located in〈mask〉”这样的句子。模型的训练目标是准确预测出这些句子中掩码槽标记对应的具体实体名称。

3.2 关系生成

3.2.1 输入输出

关系生成模块的目标是预测出实体生成模块中输出的实体之间潜在的新关系。该模块的输入为实体生成步骤得到的 3 个实体，记作 $T_e=(e_a,e_b,e_c)$，输出则是描述其中的头实体 e_a 和尾实体 e_c 之间潜在关系的句子。例如，给定输入 T_{e_1}=(Sam Altman, OpenAI, San Francisco)，输出可能是“Sam Altman lives in San Francisco”。

3.2.2 详细过程

具体来说，和实体生成模块类似，关系生成模块同样包含 2 个子步骤：句子构建和掩码槽填充。

1)句子构建。该模块首先从输入的 3 个实体中提取出第 1 个实体和最后 1 个实体，并利用句子模板构建描述这 2 个实体之间关系的句子，该模板中包括 1 个掩码槽标记用于表示 2 个实体之间的潜在关系。例如在上述示例中，输入的实体组 T_{e_1} 将被转换为句子“Sam Altman〈mask〉San Francisco”。

2)掩码槽填充。使用生成模型生成句子中的掩码槽标记对应的内容。也就是说，给定实体 e_a 和 e_c，模型旨在预测出最可能的描述 2 个实体之间的关系 r_3，即最大化概率 $P(r_3|e_a,e_c)$。这一过程可形式化地表示为

$$r_3=\arg\max_{r}P(r|e_a,e_c)。\tag{5}$$

3.2.3 预训练语料库的构建

与实体生成模块类似，在关系生成模块直接使用生成模型预测句子中的掩码槽标记得到的输出很可能并不表示实体间的具体关系，而只是一些泛

泛的动词或短语, 例如“Sam Altman arrives in San Francisco”。为了确保生成模型能够预测出实体之间的具体关系, ORANGE 构建了预训练语料库, 以增强模型生成具体关系的能力。

预训练语料库的构建过程类似于实体生成模块, 主要包含 3 种策略。1)从 Wikidata, FB15k-237, DBpedia, YAGO3, WordNet 等知识图谱中随机提取一些关系三元组。这些三元组随后通过预定义的句子模板转换为自然语言描述的句子。考虑到最终预测的 2 个实体之间可能不一定存在关系, 本文还构造了描述实体之间不存在关系的句子, 这些句子的形式为“e_a has no direct relation with e_c”。e_a和e_c分别代表头实体和尾实体的具体名称。2)从 WebNLG 中选出描述 2 个实体之间单跳关系的句子。3)利用命名实体识别工具, 从维基百科中筛选出句首和句尾均为实体且仅包含这 2 个实体的句子。这些句子可以视为 2 个实体之间关系的近似描述, 有助于提升模型生成复杂关系的能力。

在获得上述句子后, 句子中的非实体部分被替换为掩码槽标记, 形成类似于“Sam Altman 〈mask〉 San Francisco”的句子结构。模型的训练目标是准确预测出句子中的掩码槽标记所代表的关系名称。

3.3 结果聚合

结果聚合模块的目标是根据实体生成模块的输出内容和关系生成模块的输出内容计算出具有最大总概率的关系作为最终输出。该模块的输入包括实体生成模块输出的所有实体及其概率, 以及关系生成模块输出的所有关系及其概率。该模块的输出为根据总概率进行排名的前k个关系。

具体而言, 根据联合概率公式, 可以计算出某个关系的总概率。例如, 给定生成的实体组T_{e_1}= (Sam Altman, Open AI, San Francisco), 其概率为p_1= 0.6; 实体组T_{e_2}= (Bill Gates, Microsoft, Seattle) 的概率为p_2= 0.1。给定生成的关系组$S_{1,1}$= (Sam Altman, works in, San Francisco), 其概率为$p_{1,1}$= 0.7; 关系组$S_{2,2}$= (Bill Gates, works in, Seattle) 的概率为$p_{2,2}$= 0.2。那么, “works in”关系的总概率为$p_1 \times p_{1,1} + p_2 \times p_{2,2} = 0.6 \times 0.7 + 0.1 \times 0.2 = 0.44$。

结果聚合步骤的计算过程可以形式化地表示为:

$$P(r_3|r_1,r_2) = \sum_{T_e} P(r_3|T_e)P(T_e|r_1,r_2)\text{。} \tag{6}$$

在式(6)中, r_1和r_2代表输入到实体生成模块的 2 个初始关系, T_e表示实体生成模块输出的 3 个实体, r_3表示关系生成模块输出的新关系。$P(r_3|r_1,r_2)$表示在给定初始关系r_1和r_2的情况下, 可推导得到新关系r_3的概率。由于$P(r_3|T_e)$实际上独立于T_e中的第 2 个实体e_b, 因此式(6)可以简化为

$$P(r_3|r_1,r_2) = \sum_{T_e} P(r_3|e_a,e_c)P(T_e|r_1,r_2)\text{,} \tag{7}$$

其中$P(T_e|r_1,r_2)$可由实体生成模块的模型预测得到, $P(r_3|e_a,e_c)$可由关系生成模块的模型预测得到。本文的最终目标是通过r_1,r_2得到概率最高的r_3, 即

$$r_3 = \arg\max_r P(r|r_1,r_2)\text{。} \tag{8}$$

在实际应用中, 期望实体生成模块生成所有可能的实体是不现实的。考虑到低概率实体对最终结果的影响微乎其微, 本文采用了一种近似计算方法, 该方法仅考虑概率最高的前M对实体, 并用这些实体计算最终关系的总概率, 具体为

$$\mathcal{T}_{\mathcal{E}} = \{T_e\} = \{T_e|\arg\max_{T_e} P(T_e|r_1,r_2)\}_{\text{top-}M}\text{,} \tag{9}$$

概率计算公式则变为

$$P(r_3|r_1,r_2) \approx \sum_{T_\eta \in \mathcal{T}_{\mathcal{E}}} P(r_3|e_a,e_c)P(T_\eta|r_1,r_2)\text{,} \tag{10}$$

M是一个可以根据具体需求调整的超参数。在实际实验中, M=10。在根据式(10)获得所有潜在新关系r_3的总概率后, 选择总概率最高的前K个关系作为最终答案。在实际实验中, K=10。实际上, 本文在实验中观察到许多候选关系之间高度相似。为了解决这一问题, 本文提出一种简单而有效的聚类策略。具体而言, 首先利用 TF-IDF 对概率最高的K个关系进行向量化表示; 然后采用K-means 算法对这些特征向量进行聚类, 并期望得到$k(k \leqslant K)$个聚簇以及每个簇中心对应的向量。在每个聚簇中, 选择最接近簇中心即与簇中心的欧氏距离最短的关系作为该聚簇的代表, 同时过滤掉聚簇中的其他关系。这一过程会产生k个不同的关系, 这些关系共同构成最终的答案集。比如, 给定 6 个候选关系“works at”“work at”“working at”“lives in”“live in”“living in”, 假设期望筛选出 2 个关系作为最终答案, 那么通过上述聚类算法最终得到的答案为“lives in”“works at”。聚类操作可过滤掉高度相似的关系, 避免了冗余结果的输出, 减小了生成模型由于生成内容难以控制导致的负面影响。

4 实 验

本节首先介绍了本研究所做的对比实验, 以验证 ORANGE 的有效性。接下来, 通过消融实验、细

节分析实验、案例分析来深入探讨ORANGE方法的内在机制。最后介绍将ORANGE应用于一个扩展应用，即自由文本关系推理后的效果，以进一步评估方法的有效性。

4.1 实验设置

4.1.1 预训练语料库

为了进一步提升模型在开放域关系推理上的性能，本文构建了2个预训练语料库，构建过程详见3.1.3节和3.2.3节。其中，用于实体生成的预训练语料库包含100万个样本，而用于关系生成的数据集包含5万个样本。具体统计信息如表1所示。

需要注意的是，本文的预训练语料库并非是必需的，任何具备生成能力的模型都可以直接应用本文的方法框架进行关系推理。但常规预训练语言模型（如BART[37]）的原始训练语料不足，生成实体和关系的能力较弱，因此需要使用本文的预训练语料库补充。针对常规的预训练语言模型，本文使用BART作为基模型，分别通过实体生成模块的预训练语料库和关系生成模块的预训练语料库对模型进行训练得到了2个新模型，并将它们作为实体生成模块和关系生成模块中实际使用的模型。而大语言模型在预训练阶段使用的预训练数据基本已经包含了本文的预训练语料库[38]，同时大语言模型通常经过基于人类反馈的强化学习（reinforcement learning from human feedback, RLHF）[39]训练来对齐人类指令，可以很好地遵循提示词中的要求，具备较高的学习能力。以往的研究（如GPT-3[40]、思维链[41]）已经证明，设计良好的提示词指令能够在不重新训练模型的情况下，使模型灵活地适应不同的任务和要求。因此可以不必使用预训练语料库对大语言模型进行训练，而可以直接在提示词指令中约束其生成的内容限于实体和关系，从而避免其生成无关信息。

Table 1 Statistical Information of Pre-trained Corpora

表1 预训练语料库统计信息

数据类型	不同数据来源对应条数			不同句子长度（包含的单词个数）对应条数					平均长度	总条数
	知识图谱	WebNLG	维基百科	1~5	6~10	11~15	16~20	>20		
实体生成	496 401	3 599	500 000	8 165	83 998	472 934	239 895	192 008	16.81	1 000 000
关系生成	20 104	4 896	25 000	4 441	22 713	8 188	5 451	9 207	13.48	50 000

4.1.2 评估数据集

由于现有关系推理数据集只针对封闭域上的多跳推理和开放域上的1跳推理，缺少针对开放域的2跳及以上的数据集。本文构建了自己的评估数据集。评估数据集的数据来源为包含关系三元组的知识图谱，包括Freebase[26]，Wikidata[24]，YAGO3。

具体而言，评估数据集的构建包含3个步骤：1）在数据源中搜索包含共同实体e_b的2个三元组实例$(e_a,r_1,e_b),(e_b,r_2,e_c)$，并且实体$e_a$和$e_c$还需满足新的关系三元组$(e_a,r_3,e_c)$；2）进行人工审核，以确保$r_3$可以在不依赖$e_a$，$e_b$，$e_c$的先验知识的情况下，直接从$r_1$和$r_2$推导出来，该步骤保证$r_3$不是一个仅存在于特殊情况中的关系；3）为了测试模型识别出e_a和e_c之间不存在关系的能力，本文还在数据源中搜索存在$(e_a,r_1,e_b),(e_b,r_2,e_c)$但$e_a$和$e_c$不满足其他任何关系三元组的实例。通过上述方法本文得到了600条数据，其中包括300条e_a和e_c之间存在关系的数据，以及300条e_a和e_c之间不存在关系的数据。获取到这些数据后，再使用3.1.2节中所述的句子模板将其转换为自然语言描述的句子。由于可能存在多个不同的关系r_3为正确答案，且即便相同的关系也可能存在不同的描述方式，因此本文提供尽可能多的正确答案选项和描述方式。对于e_a和e_c之间不存在关系的情况，本文使用“e_a has no direct relation with e_c”作为正确答案。本文将该评估数据集命名为OpenRelation600。

为了进一步说明ORANGE方法的鲁棒性，本文使用了FB15k-237[34]作为封闭域关系推理的评估数据集，并将ORANGE方法与其他封闭域方法进行对比。

4.1.3 对比基线

在OpenRelation600数据集上，本文采用了2类基座模型：预训练语言模型（BART）和大语言模型（Llama2, GLM3, Qwen）。对于第1类基座模型，对比的方法包括SPN[27]，COMET[28]，UIE[29]，Orion[23]。对于第2类基座模型，对比的方法是通过上下文学习（in-context learning, ICL）的方式来让基座大语言模型完成开放域的2跳推理任务。需要注意的是，由于大语言模型具备强大的文本生成能力，因此本文不使用构建的预训练语料库来训练大语言模型，而是通过在提示词指令中要求模型生成的内容局限于实体或关系来避免模型生成无关内容。

在FB15k-237数据集上，本文对比的方法包

括 AMIE+[8], LoTus[10], EGNN[42], RKLE[16], TransE[17], RotatE[18], CP-KGC[32]。其中 AMIE+, LoTus 是基于规则的推理方法, EGNN, RKLE 是基于图结构的推理方法, TransE, RotatE 是基于表示学习的推理方法, CP-KGC 是基于大语言模型的推理方法。

4.1.4 评价指标

在 OpenRelation600 上, 本文采用 BLEU-1/2/3/4, ROUGE-1/2/L, METEOR 以及这些指标的平均值来作为评价指标。上述指标均用于衡量模型输出与标准答案之间的相似性, 数值越高表示对齐程度越高, 内容准确性越高。

在 FB15k-237 上, 本文使用了 Hits@1, Hits@3, Hits@10, MRR 这些评价指标, 其中 Hits@1 表示正确答案为排名第 1 的结果, Hits@3 表示正确答案为排名前 3 的结果, Hits@10 表示正确答案位于排名前 10 的结果之中, MRR 表示正确答案在结果中排名的倒数平均值。

4.1.5 参数设置

在 OpenRelation600 数据集上, 针对 BART 作为基座模型的实验, 本文使用 BART-base 作为基础模型进行预训练, 设置训练数据的单批次大小为 100, 训练轮次为 10, 学习率为 2E−5, 优化器为 AdamW, 输出序列的最大 token 数设置为 100。模型的文本生成过程采用了束搜索策略, 束宽设置为 10, 每个生成结果代表 1 个关系, 同时采用 3.3 节中介绍的聚类策略筛选出最终答案, 类簇的数量设置为 4; 针对大语言模型作为基座模型的实验, 设置输出序列的最大字符数为 400, 文本生成过程采用束搜索策略, 束宽设置为 10, 所有生成的关系均包含在排名第 1 的生成内容中由分号分隔, 其余生成内容舍弃, 若生成内容中得到的关系数量小于等于 4 就直接输出所有关系, 大于 4 则使用聚类策略并设置类簇的数量为 4。

在 FB15k-237 数据集上, 本文使用和 OpenRelation600 数据集相同的经过预训练的 BART 模型作为生成模型, 文本生成过程同样采用了束搜索策略和聚类策略, 其中束宽设置为 30, 类簇的数量设置为 10。

为确保模型结果的可信度, 针对文中所有实验, 本文均进行了 3 次使用不同随机种子的实验, 并记录了平均结果。实验部署在运行 Ubuntu 20.04.6 LTS 操作系统的服务器上, 配备了 Intel Xeon Platinum 8336C CPU, 1.0 TiB RAM, NVIDIA RTX4090 24 G GPU。

4.2 开放域数据集对比

4.2.1 与预训练语言模型方法的对比

首先将本文提出的 ORANGE 与其他基于预训练语言模型的推理方法进行对比, 包括 Orion, COMET, UIE, SPN。此外, 本实验还评估了直接使用 BART 模型以“If-then”格式处理 2 个关系输入的方法, 例如输入“If *A* works at *B* and *B* is located in *C*, then”, 要求模型输出后续内容(简称为 BART(If-then))。本文还测试了未经过训练的 BART 模型用于实体生成模块和关系生成模块后的输出(简称为 ORANGE(w/o pt))。实验结果如表 2 所示。

Table 2 Compared with Methods Based on Pre-trained Language Models

表 2 与基于预训练语言模型方法的对比 %

方法	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	平均值
Orion	32.01	5.35	0.86	0.53	24.45	40.25	4.65	40.19	18.54
COMET	25.89	5.38	2.91	2.78	18.61	42.95	5.75	40.88	18.14
UIE	17.32	11.87	8.59	7.32	20.28	43.46	8.84	20.28	20.03
SPN	2.48	0.08	0.01	0.00	2.55	8.71	0.36	8.38	2.82
BART(If-then)	35.62	9.65	1.51	0.25	28.42	45.45	9.13	44.80	21.85
ORANGE(w/o pt)	37.58	9.90	1.97	1.50	27.37	47.19	9.36	46.45	22.75
ORNAGE(本文)	41.88	20.44	13.45	11.75	37.55	53.94	22.19	53.88	33.11

从表 2 可以得出以下结论: 1)ORANGE 在所有评价指标上均优于其他对比模型, 显示了 ORANGE 的有效性。特别是, ORANGE 在评价指标平均值上至少超过了对比方法 10%。2)与直接使用“If-then”提示进行推理相比, ORANGE 在输出的正确性和多样性上均有所提升, 这表明 ORANGE 的关系推理框架相较于简单的“If-then”提示具有明显优势, 即使在未进行预训练的情况下也是如此。3)相比使用未经预训练的 BART 模型, 预训练后 ORANGE 的性能显著提升, 突显了预训练过程在增强生成模型推理能力中的重要性。

4.2.2 与大语言模型方法的对比

本文通过比较传统的上下文学习提示策略和 ORANGE 的 3 步关系推理框架, 研究了将 ORANGE

应用于大语言模型上的表现。由于直接从大语言模型中获取生成内容的概率较为困难，因此在结果聚合步骤中，本文选择关系生成步骤的全部输出中出现频率最高的关系作为最终结果。实验中使用的大语言模型包括 Llama2, GLM3, Qwen。本文为大语言模型设计的上下文提示词如图 2 所示。

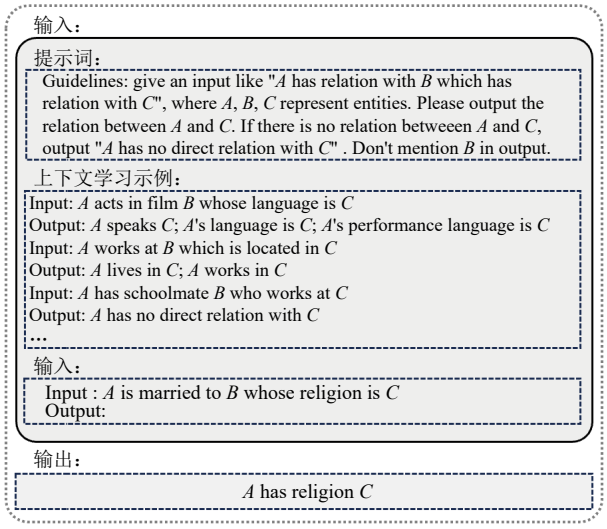


Fig. 2 Context prompt information of large language model

图 2 大语言模型的上下文提示信息

实验结果列于表 3 中。可以观察到，与使用传统的上下文学习提示策略相比，使用 ORANGE 的关系推理框架的大语言模型输出结果的准确性更高。具体而言，Llama2 使用 ORANGE 的关系推理框架后在 ROUGE-1 指标上的提升最为显著，从 41.73% 提高到 55.23%。GLM3 使用 ORANGE 的关系推理框架后在 METEOR 指标上的提升最大，从 31.92% 提升到了 40.57%。Qwen 使用 ORANGE 的关系推理框架后在 BLEU-1 指标上的提升最高，从 33.07% 提升至 48.75%。这些结果表明，与上下文学习提示策略相比，ORANGE 的关系推理框架能够更好地增强大语言模型进行开放关系推理的能力。

4.3 封闭域数据集对比

为了测试 ORANGE 在现有封闭域数据集上的效果，本文采用了 FB15k-237 作为测试数据集。这是一个典型的知识图谱补全数据集。在该数据集的基础上，本文构建了适用于关系推理的数据集。具体而言，该测试集根据输入的关系三元组 $(A,r_1,B),(B,r_2,C)$ 推出实体 A 与 C 之间的关系三元组 (A,r_3,C)，其中三元组 $(A,r_1,B),(B,r_2,C)$ 存在于 FB15k-237 的训练集中，(A,r_3,C) 存在于 FB15k-237 的测试集中。在具体实验过程中，将关系 r_1,r_2 作为 ORANGE 的输入，得到输出关系 r_3 作为测试结果，并与标准答案 r_3 进行对比。

实验结果如表 4 所示，可以观察到 ORANGE 方法在大部分指标上都优于其他方法，说明了其有效性。需要注意的是，ORANGE 并未使用 FB15k-237 的训练数据集进行专门训练，而仅仅使用了本文构建的预训练语料库进行训练，因此该实验结果更加体现了 ORANGE 方法具备较强的鲁棒性。

Table 3 Compared with Methods Based on Large Language Models

表 3 与基于大语言模型方法的对比 %

访求	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	平均值
Llama2（使用 ICL）	35.17	20.76	14.13	11.19	32.98	41.73	20.51	39.87	27.04
Llama2（使用本框架）	43.60	20.98	13.80	12.53	45.28	55.23	27.02	53.34	34.10
GLM3（使用 ICL）	31.92	20.89	14.15	11.07	30.86	48.92	24.11	46.89	28.60
GLM3（使用本框架）	40.57	21.96	18.59	17.67	41.63	49.10	24.43	49.03	32.87
Qwen（使用 ICL）	33.07	20.59	14.50	11.05	31.13	47.57	22.93	46.80	28.46
Qwen（使用本框架）	48.75	28.53	21.93	20.15	44.14	56.62	27.71	56.50	38.04

4.4 消融实验

为了评估 ORANGE 在不同条件下的有效性，本文进行了消融实验，具体包含以下实验内容：1）分别使用未经预训练的 BART 模型直接用于实体生成模块，以及未经预训练的 BART 模型直接用于关系生成模块，实验结果在表 5 中标注为“w/o EG”和“w/o RG”，同时还对上述 2 种情况进行了组合，在表中标注为“w/o EG & RG”。2）在结果聚合模块中，本文采用了联合概率计算公式来得出每个关系的总概率，并选择最终结果。为了评估这种方法的有效性，本文还进行了另一种方法的实验，即不使用总概率来选择关系，而是选择在关系生成模块产生的所有输出中出现频率最高的关系作为最终结果，表中记为“w/o RA”。同时该条件也和“w/o EG”以及“w/o RG”的 2 种条件组合进行实验，表中记为“w/o EG&RA”和“w/o RG&RA”。3）在 3.3 节中，本文提出在结果聚

Table 4 Experimental Results on FB15k-237

表 4 FB15k-237 上的实验结果 %

方法	Hits@1	Hits@3	Hits@10	MRR
AMIE+	17.4		40.9	
LoTus	24.9		50.9	32.1
EGNN	25.8	38.3	53.2	35.0
RKLE	26.8	35.1	45.3	31.0
TransE	19.8	37.6	44.1	27.9
RotatE	24.1	37.5	53.3	33.3
CP-KGC	25.1	36.5	51.6	33.8
ORANGE（本文）	26.8	39.0	55.1	35.1

合模块中将候选答案的文本编码进行聚类，并在每个类簇中选择最接近簇中心的关系形成最终的答案集。为验证该策略的有效性，本文还进行了不使用聚类策略的消融实验，实验结果在表中标注为“w/o cluster”。4)为了验证本文设计的包含关系代词的句子模板能够更好地优化语言模型的生成能力，本文探索了一种替代的预训练语料格式，这种格式直接将关系和掩码槽标记进行连接，而不添加关系代词，例如“〈mask〉was born in〈mask〉is a city of〈mask〉”。在此基础上，本文对使用这种格式构造的预训练语料库进行预训练的模型进行了测试，表中记为“w/o template”。

消融实验结果如表 5 所示，可以得出以下结论：1)在实体生成模块和关系生成模块中，本文构建的预训练语料库提升了原始的语言模型生成实体和关系的能力，说明了使用该预训练语料库对语言模型的训练能有效提升模型的推理能力。2)在结果聚合模块中，添加聚类的方法显著优于未使用聚类的方法，表明聚类策略是一种简单而有效的技术。3)在结果聚合步骤中，选取总概率最高的关系相比选取出现频率最高的关系表现更好，表明总概率计算是一种更为有效的结果聚合策略。4)使用本文设计的包含关系代词的句子模板构建预训练语料库而不是直接连接关系文本，能够取得更好的效果。这表明本文设计的句子模板能够更好地提升模型的生成能力。

4.5 实验分析

4.5.1 预训练语料库规模的影响

本节探讨了不同预训练语料库规模对实体生成模型性能的影响，并在 OpenRelation600 数据集上进行了测试。本文发现，当预训练语料库中的样本量少于 80 000 条时，模型性能明显下降。因此，本文将 80 000 条作为最低规模，并进一步在对比实验中考察了更大规模的语料库，具体为 200 000 条和 1 000 000 条。实验结果详见表 6，从结果中可以看出，增加预训练语料库规模对模型性能的提升有限，且增益与语料库规模的增长不成正比。具体而言，尽管数据集规模从 80 000 条增加至 1 000 000 条，增长了 12.5 倍，但平均分仅提高了 0.52%，BLEU-1 提高了 0.06%，

Table 5 Ablation Study

表 5 消融实验 %

方法	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	平均值
w/o EG	39.15	16.84	3.81	3.14	33.08	49.52	15.01	49.37	26.24
w/o RG	38.18	10.52	1.79	1.28	27.91	47.48	9.49	47.42	23.01
w/o RA	35.76	13.21	8.43	5.35	25.18	44.62	18.56	44.09	24.40
w/o EG&RG	37.58	9.90	1.97	1.50	27.37	47.19	9.36	46.45	22.75
w/o EG&RA	33.43	8.13	2.63	1.86	23.54	42.84	8.46	42.76	20.47
w/o RG&RA	31.87	6.22	1.42	0.95	21.39	41.32	6.17	40.08	18.68
w/o cluster	40.63	15.49	10.86	7.00	33.24	52.71	20.27	52.52	29.09
w/o template	40.90	18.39	11.65	10.82	33.87	52.90	20.71	52.76	30.25
ORANGE（本文）	44.26	22.80	15.49	14.77	37.55	53.94	22.19	53.88	33.11

Table 6 Impact of Pre-training Corpus Size

表 6 预训练语料库规模的影响 %

语料库规模	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	平均值
80 000	44.20	21.94	14.75	13.76	37.27	52.78	21.29	52.74	32.59
200 000	44.23	22.76	15.26	14.31	37.43	53.06	22.13	53.01	32.77
1 000 000	44.26	22.80	15.49	14.77	37.55	53.94	22.19	53.88	33.11

METEOR 提高了 0.28%, 而 ROUGE-1 则提升了 1.16%。这说明了语料的进一步增加并不能带来模型性能的明显提升。

4.5.2 与封闭域方法对比

传统的基于知识图谱的关系推理操作在封闭域假设下进行, 其中推断的关系类型是知识图谱中已经存在的。相比之下, ORANGE 可以推断出知识图谱中可能不存在的更广泛的关系。本文通过将 ORANGE 的结果与基于知识图谱的规则归纳系统 AMIE+进行比较来验证这一点。AMIE+[8] 是一种基于知识图谱的知名关系归纳方法。在 YAGO2[43] 上, AMIE+归纳出了总共 84 条规则, 这个规则的形式类似于式(2)。为了验证这些规则的有效性, 本文邀请了 3 位志愿者对其正确性进行评估并计算平均值, 确认了其中 26 条规则是正确的。根据本文的观察, 在所有规则中, 共有 57 种不同的 r_1 和 r_2 组合。这些组合随后被用作 ORANGE 的输入来生成新的关系 r_3。实验设置 $K=10, k=4$, 总共推断出 228 条新关系。经过 3 位志愿者的确认, 其中 64 条生成的知识是正确的。

实验结果如表 7 所示。从结果可以看出: 1)ORANGE 识别出的关系数量超过 AMIE+, 将新发现的关系数量从 84 条扩展至 228 条。2)尽管关系总数几乎是 AMIE+的 3 倍, 但正确关系的数量也从 26 条增加到 64 条, 最终关系的准确率保持几乎不变。

Table 7 Compared with AMIE+

表 7 与 AMIE+的对比

指标	AMIE+	ORANGE
新关系数量	84	228
正确的新关系数量	26	64
准确率/%	30.95	28.67

4.5.3 人工评估

除了自动评估, 本文还采用人工评估来验证 ORANGE 的有效性。具体而言, 本文选择了 BART-base, BART-large, gpt-3.5-turbo 作为基线模型, 并采用 Hits@1 和 Hits@3 作为评估指标。首先, 从 Freebase, Wikidata, YAGO3 中随机选择 300 对形如 (e_a, r_1, e_b), (e_b, r_2, e_c)的关系三元组作为保留数据集。然后, 本文使用 3.1.2 节中定义的句子模板, 将三元组构建成句子后, 推导 e_a 和 e_c 之间可能的新关系。对于 ORANGE, BART-base, BART-large, 本文使用实体生成、关系生成、结果聚合 3 步框架进行潜在新关系的推理。对于 gpt-3.5-turbo, 本文使用图 2 所示的上下文提示作为输入。最后, 邀请 3 位志愿者对每个结果进行评估。只有当所有志愿者一致认为生成的新关系确实可以由输入的关系得到, 才认为推理是正确的。

实验结果列于表 8 中。从结果中可以看到, ORANGE 在 Hits@1 和 Hits@3 指标上显著优于其他 3 种方法, 进一步说明了 ORANGE 的有效性。具体而言, ORANGE 在这 2 个指标上分别比 gpt-3.5-turbo 方法高出 6 个百分点和 18 个百分点。

Table 8 Results of Human Evaluation

表 8 人工评估结果 %

方法	Hits@1	Hits@3
BART-base	3.33	16.00
BART-large	11.23	22.56
gpt-3.5-turbo	43.33	63.33
ORANGE（本文）	49.33	81.33

4.6 案例分析

为了更深入地了解基于大语言模型的关系推理方法, 本文进行了案例分析。

根据本文的观察, 发现没有使用 ORANGE 关系推理框架的基于大语言模型的关系推理方法通常会出现 2 种错误。首先, 即使在输入中明确指示结果中不应包含实体 e_b, 输出结果中仍有可能包含实体 e_b。例如, 对于输入“*A* is the author of *B* whose language is *C*”, GLM3 输出“*A* wrote *B* in language *C*”。其次, 未使用 ORANGE 的大语言模型往往直接将 e_a 与 e_b 之间的关系作为 e_a 与 e_c 之间的关系。例如, 对于输入“*A*'s mother is *B* whose religion is *C*”, GLM3 输出“*A*'s mother is *C*”。然而, 这些问题在 ORANGE 的输出中均没有出现。具体而言, 对于第 1 个问题, ORANGE 在中间步骤中移除了实体 *B*, 因此实体 *B* 不可能出现在最终结果中。针对第 2 个问题, ORANGE 采用了包含实体生成和关系生成的 2 步生成过程, 在实体生成步骤中就已经获得了具体的实体 e_a 和 e_c, 避免了实体 e_a 与 e_b 之间关系被误认为是 e_a 和 e_c 之间的关系。表 9 提供了 ORANGE 生成的一些合理的新关系示例。

4.7 自由文本关系推理

在实际任务中, 通常涉及从一段自由文本(free text)中提取实体之间的关系。该问题定义为输入一段包含 3 个或更多实体的文本, 模型需要输出第 1 个和最后 1 个实体之间的关系。值得注意的是, 第 2 个实体与其他 2 个实体之间可能没有语义关系(这一设置是本文任务的扩展)。例如, 给定输入“Thomas Ebdon (1 738—1 811) was a British composer and organist born in Durham”, 其中包含 3 个实体“Thomas Ebdon”“com-

Table 9 Case Study of Generated New Relations

表 9 生成新关系的案例分析

输入	输出
spouse, country e.g., (Donald Trump, spouse, Melania Trump), (Melania Trump, country, United States)	1）is citizen of 2）nationality is 3）lives in 4）works in
lives in, language e.g., (Donald Trump, lives in, New York City), (New York City, language, English)	1)'s language is 2）speaks 3）is fluent in 4）has spoken language

poser"" Durham", 模型需要输出"Thomas Ebdon"和"Durham"之间的关系, 例如"was born in"。

本文应用了 ORANGE 来解决上述问题, 具体使用 Google-RE 数据集[22]作为实验数据。由于该数据集的输入和输出格式与本文的任务定义略有不同, 本文对原始数据集进行了修改。具体而言, 对于每个输入文本, 首先使用实体提取工具来选择包含 3 个或更多实体的文本。然后, 使用关系提取工具来提取第 1 个和最后 1 个实体之间的关系。最后, 使用掩码槽标记替换文本中的实体后作为本文方法的输入以输出实体之间的关系, 并与真实值进行比较。实验结果见表 10。从表 10 中可以观察到, 尽管该任务的输入数据格式与 ORANGE 的训练数据格式有所不同, ORANGE 仍然优于未经预训练的 BART-base 模型。具体而言, ORANGE 的平均得分比基线模型高出 2.03%, 显示了 ORANGE 在关系推理任务上拥有一定程度的泛化能力。

Table 10 Free-Text Relation Reasoning

表 10 自由文本关系推理 %

方法	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	平均值
BART（未经预训练）	37.31	14.72	5.52	2.27	29.12	54.13	54.25	14.56	26.49
ORANGE（本文）	41.08	17.45	5.84	4.99	30.49	54.80	54.91	18.60	28.52

5 总结和展望

本文专注于开放域中的 2 跳关系推理问题。为了解决这一任务, 本文提出了一个包括 3 个步骤: 实体生成、关系生成、结果聚合的关系推理方法。本文还构建了针对性的预训练语料库, 以更好地挖掘模型的推理能力。此外, 本文建立了一个新的评估数据集, 以评估模型在开放域 2 跳关系推理任务中的能力。实验结果证明, ORANGE 优于其他现有的推理模型。另外, 将 ORANGE 的关系推理框架应用于大语言模型时, 其性能超越了使用上下文学习提示策略的相同模型。

尽管 ORANGE 将开放域推理引入了 2 跳关系, 但仍面临若干限制: 1)在涉及 3 跳或更多跳的推理中, 虽然可以通过 2 跳推理不断扩展跳数, 但模型的预测结果在多次扩展后往往会显著偏离, 导致结果在实际应用中并不可用。2)评估数据集是半自动化构建的, 需要大量的人工进行筛选操作, 得到的数据库规模相对较小, 缺乏用于构建大规模的评估数据集的自动化手段。因此, 未来的工作包括将开放域关系推理任务扩展到更多跳数, 并构建更大规模、更全面的评估数据集。

作者贡献声明: 刘井平和岑哲栋负责算法思路设计、实验方案制定, 完成实验、论文撰写等工作; 张维彦负责论文撰写指导、绘图指导、技术支持; 阮彤完成了论文框架设计、整体内容规划; 王超和王昊奋负责论文撰写指导、实验方案完善、整体思路指导。

参 考 文 献

[1] Xie Shaorong, Liu Ruishen, Wang Xinzhi, et al. Hierarchical knowledge-enhancement framework for multi-hop knowledge graph reasoning[J]. Neurocomputing, 2024, 588: 127673

[2] Yang Yuji, Xu Bin, Hu Jiawei, et al. Accurate and efficient method for constructing domain knowledge graph[J]. Journal of Software, 2018, 29(10): 2931−2947 (in Chinese)
(杨玉基, 许斌, 胡家威, 等. 一种准确而高效的领域知识图谱构建方法[J]. 软件学报, 2018, 29(10): 2931−2947)

[3] Wang Fei, Liu Jingping, Liu Bin, et al. Survey on construction of code knowledge graph and intelligent software development[J]. Journal of Software, 2020, 31(1): 47−66 (in Chinese)
(王飞, 刘井平, 刘斌, 等. 代码知识图谱构建及智能化软件开发方法研究[J]. 软件学报, 2020, 31(1): 47−66)

[4] Wang Meng, Wang Haofen, Li Bohan, et al. Survey on key technologies of new generation knowledge graph[J]. Journal of Computer Research and Development, 2022, 59(9): 1947−1965 (in Chinese)

(王萌, 王昊奋, 李博涵, 等. 新一代知识图谱关键技术综述[J]. 计算机研究与发展, 2022, 59(9): 1947−1965)

[5] Li Dongmei, Zhang Yang, Li Dongyuan, et al. Review of entity relation extraction methods[J]. Journal of Computer Research and Development, 2020, 57(7): 1424−1448 (in Chinese)
(李冬梅, 张扬, 李东远, 等. 实体关系抽取方法研究综述[J]. 计算机研究与发展, 2020, 57(7): 1424−1448)

[6] Mutabazi E, Ni Jianjun, Tang Guangyi, et al. A review on medical textual question answering systems based on deep learning approaches[J]. Applied Sciences, 2021, 11(12): 5456

[7] Galárraga L A, Teflioudi C, Hose K, et al. AMIE: Association rule mining under incomplete evidence in ontological knowledge bases[C]//Proc of the 22nd Int Conf on World Wide Web. New York: ACM, 2013: 413−422

[8] Galárraga L, Teflioudi C, Hose K, et al. Fast rule mining in ontological knowledge bases with AMIE+[J]. The VLDB Journal, 2015, 24(6): 707−730

[9] Meilicke C, Chekol M W, Ruffinelli D, et al. Anytime bottom-up rule learning for knowledge graph completion[C]//Proc of the 28th Int Joint Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 3137−3143

[10] Pan Yudai, Liu Jun, Zhang Lingling, et al. Incorporating logic rules with textual representations for interpretable knowledge graph reasoning[J]. Knowledge-Based Systems, 2023, 277: 110787

[11] Cui Yuanning, Sun Zequn, Hu Wei. A pre-trained universal knowledge graph reasoning model based on rule prompts[J]. Journal of Computer Research and Development, 2024, 61(8): 2030−2044 (in Chinese)
(崔员宁, 孙泽群, 胡伟. 基于规则提示的知识图谱通用推理预训练模型[J]. 计算机研究与发展, 2024, 61(8): 2030−2044)

[12] Lao Ni, Cohen W. Relational retrieval using a combination of path-constrained random walks[J]. Machine Learning, 2010, 81: 53−67

[13] Xiong Wenhan, Hoang T, Wang W Y. DeepPath: A reinforcement learning method for knowledge graph reasoning[C]//Proc of the 2017 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2017: 564−573

[14] Bai Luyi, Chen Mingzhuo, Xiao Qianwen. Multi-hop temporal knowledge graph reasoning with multi-agent reinforcement learning[J]. Applied Soft Computing, 2024, 160: 111727

[15] Wang Zikang, Li Lingjing, Zeng Daniel. SRGCN: Graph-based multi-hop reasoning on knowledge graphs[J]. Neurocomputing, 2021, 454: 280−290

[16] Liu Ruinan, Yin Guisheng, Liu Zechao. Learning to walk with logical embedding for knowledge reasoning[J]. Information Sciences, 2024, 667: 120471

[17] Bordes A, Usunier N, Garcia-duran A, et al. Translating embeddings for modeling multi-relational data[J/OL]. Advances in Neural Information Processing Systems, 2013, 26[2025-03-20]. https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html

[18] Sun Zhiqing, Deng Zhihong, Nie Jianyun, et al. RotatE: Knowledge graph embedding by relational rotation in complex space[C/OL]// Proc of the 7th Int Conf on Learning Representations. Amherst, MA: Openreview. net, 2019[2025-03-20]. https://openreview.net/forum?id=HkgEQnRqYQ

[19] Nayyeri M, Cil G M, Vahdati S, et al. Trans4E: Link prediction on scholarly knowledge graphs[J]. Neurocomputing, 2021, 461: 530−542

[20] Yang Xiaohan, Wang Ning. A confidence-aware and path-enhanced convolutional neural network embedding framework on noisy knowledge graph[J]. Neurocomputing, 2023, 545: 126261

[21] He Peng, Zhou Gang, Chen Jing, et al. Type-enhanced temporal knowledge graph representation learning model[J]. Journal of Computer Research and Development, 2023, 60(4): 916−929 (in Chinese)
(何鹏, 周刚, 陈静, 等. 类型增强的时态知识图谱表示学习模型[J]. 计算机研究与发展, 2023, 60(4): 916−929)

[22] Petroni F, Rocktäschel T, Riedel S, et al. Language models as knowledge bases?[C]//Proc of the 2019 Conf on Empirical Methods in Natural Language Proc and the 9th Int Joint Conf on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: ACL 2019: 2463−2473

[23] Cui Wanyun, Chen Xingran. Open rule induction[J]. Advances in Neural Information Processing Systems, 2021, 34: 28536−28547

[24] Vrandečić D, Krötzsch M. Wikidata: A free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10): 78−85

[25] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A nucleus for a web of open data[C]//Proc of the 6th Int Semantic Web Conf. Berlin: Springer, 2007: 722−735

[26] Bollacker K, Evans C, Paritosh P, et al. Freebase: A collaboratively created graph database for structuring human knowledge[C]//Proc of the 22nd National Conf on Artificial Intelligence: Palo Alto, CA: AAAI, 2007: 1962−1963

[27] Sui Dianbo, Zeng Xiangrong, Chen Yubo, et al. Joint entity and relation extraction with set prediction networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(9): 12784−12795

[28] Hwang J D, Bhagavatula C, Le Bras R, et al. (COMET-) Atomic 2020: On symbolic and neural commonsense knowledge graphs[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 6384−6392

[29] Lu Yaojie, Liu Qing, Dai Dai, et al. Unified structure generation for universal information extraction[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics(ACL). Stroudsburg, PA: ACL, 2022: 5755−5772

[30] Wei Yanbin, Huang Qiushi, Zhang Yu, et al. KICGPT: Large language model with knowledge in context for knowledge graph completion[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 8667−8683

[31] Li Dawei, Tan Zhen, Chen Tianlong, et al. Contextualization

distillation from large language model for knowledge graph completion[C]//Proc of the 18th Conf of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2024: 458−477

[32] Yang Rui, Zhu Jiahao, Man Jianping, et al. Enhancing text-based knowledge graph completion with zero-shot large language models: A focus on semantic enhancement[J]. Knowledge-Based Systems, 2024, 300: 112155

[33] Zhang Yichi, Chen Zhuo, Guo Lingbing, et al. 2024. Making large language models perform better in knowledge graph completion[C]//Proc of the 32nd ACM Int Conf on Multimedia (MM'24). New York: ACM, 2024: 233–242

[34] Toutanova K, Chen Danqi. Observed versus latent features for knowledge base and text inference[C]//Proc of the 3rd Workshop on Continuous Vector Space Models and Their Compositionality. Stroudsburg, PA: ACL, 2015: 57−66

[35] Miller G A. WordNet: A lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39−41

[36] Gardent C, Shimorina A, Narayan S, et al. The WebNLG challenge: Generating text from RDF data[C]//Proc of the 10th Int Conf on Natural Language Generation. Stroudsburg, PA: ACL, 2017: 124−133

[37] Lewis M, Liu Yinhan, Goyal N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics(ACL). Stroudsburg, PA: ACL, 2020: 7871−7880

[38] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models[J]. arXiv preprint, arXiv: 2302.13971, 2023

[39] Stiennon N, Ouyang Long, Wu J, et al. Learning to summarize with human feedback[J]. Advances in Neural Information Processing Systems, 2020, 33: 3008−3021

[40] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877−1901

[41] Wei J, Wang Xuezhi, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24824−24837

[42] Zhang Yang, Chen Jiarui, Cheng Zhe, et al. Edge propagation for link prediction in requirement-cyber threat intelligence knowledge graph[J]. Information Sciences, 2024, 653: 119770

[43] Hoffart J, Suchanek F M, Berberich K, et al. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial Intelligence, 2013, 194: 28−61



Liu Jingping, born in 1991. PhD, associate professor. His main research interests include knowledge graph and natural language processing.

刘井平, 1991 年生。博士，副教授。主要研究方向为知识图谱、自然语言处理。



Cen Zhedong, born in 1999. Master candidate. His main research interest includes natural language processing.

岑哲栋, 1999 年生。硕士研究生。主要研究方向为自然语言处理。



Zhang Weiyan, born in 1994. PhD. Her main research interests include information extraction and large language model.

张维彦, 1994 年生。博士。主要研究方向为信息抽取、大语言模型。



Ruan Tong, born in 1973. PhD, professor, PhD supervisor. Member of CCF. Her main research interests include knowledge graph, data mining, and data quality assessment.

阮　彤, 1973 年生。博士，教授，博士生导师。CCF 会员。主要研究方向为知识图谱、数据挖掘、数据质量评估。



Wang Chao, born in 1988. PhD, lecturer. His main research interests include knowledge graph, commonsense reasoning, and machine learning.

王　超, 1988 年生。博士，讲师。主要研究方向为知识图谱、常识推理、机器学习。



Wang Haofen, born in 1982. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include knowledge graph, natural language processing, and data mining.

王昊奋, 1982 年生。博士，教授，博士生导师。CCF 高级会员。主要研究方向为知识图谱、自然语言处理、数据挖掘。