小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2024-0540
2026 年 2 月 第 2 期
Vol.47 No.2 2026

结合思维链评估和事实验证的小样本关系抽取方法

谭思莹[1,2],段建勇[1,2],范安宇[1,2],孙 婷[3],刘 杰[1,2]

[1](北方工业大学 信息学院,北京 100144)
[2](CNONIX 国家标准应用与推广实验室,北京 100144)
[3](军事科学院 战争研究院,北京 100850)
E-mail:18004292810@163.com

摘 要: 小样本关系抽取旨在从极少量标注数据中学习提取样本中实体之间关系的能力.一些研究表明,引入思维链生成推理过程辅助大模型进行推理的方法,可以有效完成小样本关系抽取任务.然而,现有的基于思维链的方法很少评估推理过程的准确性,其推理过程可能存在质量低的问题.因此,提出一种为基于思维链的小样本关系抽取任务设计的思维链评估方法,来检查生成的推理过程是否包含样本关键信息.同时,为了解决大模型推理存在的幻觉问题,提出了事实验证方法,旨在评估提取的三元组与原样本之间的事实一致性.实验结果表明,与之前的方法相比,该模型实现了性能上的提升,在 FewRel 1.0 和 FewRel 2.0数据集上性能最高提升了 1.3% 和 2%,这表明了这两个方法的有效性.

关 键 词: 关系抽取;小样本;思维链;上下文学习;大模型
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)02-0378-08

Chain-of-thought Evaluation and Fact-verification for Few-shot Relation Extraction

TAN Siying[1,2],DUAN Jianyong[1,2],FAN Anyu[1,2],SUN Ting[3],LIU Jie[1,2]
[1](School of Information Science and Technology,North China University of Technology,Beijing 100144,China)
[2](CNONIX National Standard Application and Promotion Lab,Beijing 100144,China)
[3](War Research Institute,Academy of Military Sciences,Beijing 100850,China)

Abstract: Few-shot relation extraction aims to learn the ability to extract relationships between entities in a sample from a small amount of labeled data. Some studies have shown that the method of using chain-of-thought generation reasoning process to assist the reasoning of large language models can effectively complete the Few-shot relation extraction task. However, the existing methods based on chain-of-thought rarely evaluate the accuracy of the reasoning process, and the generated reasoning process may suffer from low quality. Therefore, a chain-of-thought evaluation method specifically designed for the Few-shot relation extraction task based on the chain-of-thought is proposed. This method checks check whether the generated reasoning process contains the key information of the sample. At the same time, to address hallucinations that exist in Large Language Models, a fact verification method is proposed to evaluate the factual consistency between the extracted triplet and the original sample. The experimental results show that compared with the previous methods, the model achieves performance improvement, with the highest improvement of 1.3% and 2% on the FewRel 1.0 and FewRel 2.0 datasets, which indicates the effectiveness of these two methods.

Keywords: relation extraction; few-shot; chain-of-thought; in-context learning; large language models

0 引 言

少样本关系抽取(FSRE)是关系抽取(RE)的子任务,其目的是借助少量的样本数据,从非结构化文本中提取实体对之间的关系[1,2].传统的关系抽取方法往往依赖于大量的标注数据,这不仅耗费了大量的资源,而且在面对未知领域的新类型关系时,模型的泛化能力受到严重限制.因此,研究如何在少量标注数据的情况下实现高效的关系抽取具有重要的理论和实际意义.

用于解决 FSRE 的方法主要是基于传统预训练语言模型和基于大模型的方法.由于大多数传统的 FSRE 方法缺乏必要的先验知识,在执行关系抽取任务时表现出一定的局限性.故采取基于大模型的方法逐渐成为了执行小样本关系抽取任务的有效途径.近年来,大型语言模型(Large Language Models,LLM)凭借其强大的语言表示能力和泛化能力,在多个自然语言处理任务中取得了显著成果[3,4].GPT 系列等大型语言模型已经表现出了显著的上下文学习的能力,并在众多 NLP 任务中实现了显著的性能提升.此外,基于思维链的各

收稿日期:2024-12-06 收修改稿日期:2025-01-10 基金项目:新一代人工智能国家科技重大专项项目(2020AAA0109703)资助;国家自然科学基金项目(62476007,62076167)资助;北京市教育委员会科学研究计划项目(KM202210009002)资助. 作者简介:**谭思莹**,女,1999 年生,硕士研究生,研究方向为自然语言处理;**段建勇**,男,1978 年生,博士,教授,研究方向为自然语言处理;**范安宇**,男,1997 年生,硕士研究生,研究方向为自然语言处理;**孙 婷**,女,1990 年生,博士,助理研究员,研究方向为自然语言处理;**刘 杰**,男,1970 年生,博士,教授,研究方向为自然语言处理.

种提示方法使大模型在解决数学问题和常识推理问题时表现出了更强的推理能力.

为了更有效地处理小样本关系抽取任务,GPT-RE[5]研究基于上下文学习和思维链提出了使用黄金标签诱导大模型自动生成推理过程的方法,大幅提升了大模型在 FSRE 任务上的性能. CoT-ER[6]研究提出了一种在推理过程中引入实体和关系标签,使自动生成的推理过程可以为最终的推理提供更多信息. 尽管如此,以上几种方法显示出了性能提升,但与人工设计的思维链方法相比并未展现出明显的优势.

例如在处理某些相似的三元组时,模型会为示例样本生成类似的推理过程,进而在对查询集样本的推理中产生混淆. 譬如关系标签"winner"和"participant",虽然二者分别用于描述事件或进程中的胜出者和参与者,但二者头尾实体的相似度较高,都为涉及参与事件的个体、组织与事件本身之间的关系,不利于模型对二者的特征进行区分. 在实验中,关系标签"winner"极易被混淆为"participant". 这表明部分关系极度依赖于实体类型以及上下文信息,在这种情况下,大模型容易为这些关系类型的数据生成不符合常识的推理过程,进而导致在最终的推理步骤中产生错误.

基于以上问题,本文重新设计了诱导大模型生成推理过程的提示模板,通过在上下文中引入实体类型以及关系等信息,为最终的推理提供更多更完整的信息. 同时设计了一个思维链评估模块,通过对生成的多个推理过程的质量进行评估,挑选出最优的推理过程添加到支持集中. 在最后的推理阶段,设计了一个事实验证模块,通过检查最终抽取出的三元组是否与源文本所表达的事实一致来保证最终输出结果的正确.

本文的主要贡献如下:

设计了思维链评估模块,在思维链生成和评估阶段引入实体关系的信息,确保了思维链的全面性,解决了关键推理信息不准确的问题.

提出了事实验证模块,通过评估提取的三元组与源文本的事实一致性,让模型在多种事实验证场景中进行全面检查,降低了大模型在特定任务上的幻觉现象.

1　相关工作

1.1　思维链提示

Wei 等人[7]提出了思想链(Chain of Thought,简称 CoT)提示,模型通过学习如何生成问题解决方案的中间推理步骤,增强其处理复杂任务的能力. 随后,为了得到更精确、更全面的推理过程,一些研究工作采用迭代引导、自动生成提示等方式从多个角度优化模型性能,例如 Sun 等人[8]提出的 Iter-CoT,提出迭代引导,使大模型能够自主纠错,从而提高推理过程的准确性,Wang 等人[9]提出自一致提示方法,采样多个推理路径而非采用模型的贪婪路径,然后通过边缘化采样的推理路径来选择最一致的答案.

此外,先前的一些工作[10,11]表明大模型可以通过 CoT 提示解决具有挑战性的任务,Chung 等人[12]指出虽然大模型可以通过 CoT 提示解决新任务,但其有效性并不适用于较小的模型,CoT 提示需要依赖数千亿参数的大模型来获得最佳性能[13-15]. 故许多研究人员尝试使用微调 CoT 的方法,使较小的大模型具有推理能力,这样可以有效减少计算需求和推理成本.

以往的研究往往忽视了 CoT 质量的问题,因此通过在推理过程中引入一个思维链评估模块,来解决生成的关键推理信息不准确的问题.

1.2　上下文学习

In-Context Learning(ICL)充分利用模型的预训练知识,并通过在推理阶段提供相关的上下文信息来生成或调整模型输出.

自上下文学习出现以来,有大量工作聚焦于提升上下文学习模型的性能,Liu 等人[16]通过基于句子嵌入的 KNN 搜索来提升 ICL 检索到示例的相关性,Zhang 等人[17]通过前向计算聚合元梯度并将其应用于上下文学习来降低示例顺序对性能的影响. Min 等人[18]和 Zhao 等人[19]在 ICL 框架下,通过使用少量的演示示例,大模型在多种任务上的性能达到了传统的全监督方法的水平. Brown 等人[20]的研究证明 ICL 可用于替代大模型在特定数据集上针对特定任务进行微调的工作,进而在零样本或小样本情况下完成预期的任务.

此外,还有一些其他工作尝试将其应用到某些专业领域,如 Gutiérrez 等人[21]尝试在生物医学信息抽取任务上应用上下文学习.

1.3　小样本关系抽取

小样本关系抽取任务旨在解决在仅有少量标注样本的情况下预测实体对之间的语义关系的问题. 这为构建结构化知识,如知识图谱等,提供了基础.

最初,Han 等人[22]首次将小样本学习引入关系分类任务,并构建了大规模监督小样本关系分类数据集,同时引入 FERE 大规模基准. Soares 等人[23]通过结合 Harris 的分布假设扩展和 BERT 文本表示,从实体链接的文本中学习任务无关的关系表示,从而解决了通用关系提取器在泛化能力上的局限性. Zhang 等人[24]通过引入通用的特定领域的知识图谱作为外部知识整合到模型中,提高模型的领域适应能力.

随后,基于提示学习的方法表现了巨大的潜力,Liu 等人[25]以及 Gu 等人[26]提出基于提示学习的关系抽取方法,可以获得良好的任务效果. Zhang 等人[27]提出标签提示退出方法,通过在学习过程中随机删除标签描述来解决模型在面对未知的关系和文本标签时性能不佳的问题.

近期,Wang 等人[28]提出了一种新颖的句子增强方法,用于生成额外的训练数据,以提高小样本关系提取的性能.

2　方　法

2.1　任务定义

小样本关系抽取(FSRE)任务的目标是借助少量的训练数据提取出给定句子中实体对之间的关系. 通常由许多单独的 N-way-K-shot 关系抽取任务组成. 在每个单独的 N-way-K-shot 任务中,存在一个支持集 S 和一个查询集 Q. 支持集 S 包含 N 个关系类,每个类有 K 个标记实例. 查询集 Q 包含每个 $r \in R$ 的测试输入实例. 由于 N 和 K 通常非常小,因此在标记数据有限的查询实例中预测关系是一个重大挑战.

将小样本关系抽取任务视为一个基于大模型的文本生成

任务,对于给定的文本 s 以及头实体 h 和尾实体 t,需要从句子 s 中抽取这两个实体之间的关系 $r \in R$,这里 R 代表一组预定义的关系.

2.2 模型整体框架

CoTEV 模型整体结构如图 1 所示,其主要由 3 个模块组成,思维链评估模块旨在通过提示大模型使用人类标注的数据为支持集中每个实例生成推理过程,并对这些推理过程进行评估. 检索模块使用以文本匹配为基础的检索的方法,该模

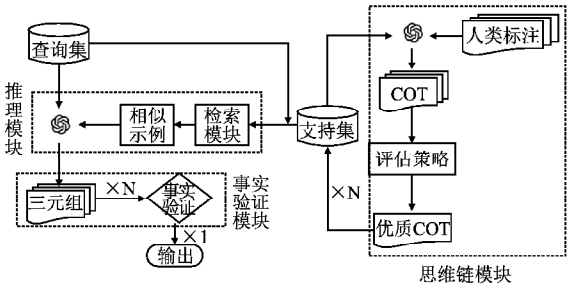

图 1 CoTEV 模型结构图

Fig. 1 Structure diagram of the CoTEV

块选择与查询样本具有高相关性的实例,作为最终提示模板中的示例. 此外,为了解决大模型可能出现的幻觉问题,提出了一个事实验证模块,用于判断预测的关系是否符合源文本中直接陈述的事实.

2.3 思维链评估模块

2.3.1 思维链生成

思维链提示是一种通过生成一系列的推理解释来增强大模型推理能力的技术,在该方法中,构建提示主要分为 3 步:

任务指令

Please extract a relationship triplet from the following sentence and make an inference process,structured as(head entity,relationship,tail entity),and follow the following restrictions:1.Place the final triplet result in curly braces (as shown below);2.Entity cannot be a personal pronoun,such as he/she/it;3.This relationship must be 'a place served by a transportation hub';4.Convert the outout to lowercase letters. These are examples.

小样本演示示例

Context:He also played in the l905 FA Cup Final at Crystal Palace, which ended in a 2-0 defeat to Aston Villa following two goals from Harry Hampton.
Given the context,what is the relation between l905 fa cup final and aston villa?
l.Subject entity "1905 FA Cup Final" refers to the entity of a sports event in the context.
2.Object entity "Aston Villa" is the name of the football club,which refers to the entity of the organization or club in the context.
3.According to the context,"ended in a 2-0 defeat to Aston Villa" indicates that "Aston Villa" was the opposing team in the "1905 FA Cup Final".
So,the relation between subject entity"1905 FA Cup Final" and object entity "Aston Villa" is opposing team in".

查询语句

Context: In l806,the Holy Roman Empire was dissolved by the last emperor Francis Ⅱ, after its defeat by Napoleon at the Battle of Austerlitz.
Given the context,what ls the relation between Battle of Austerlitz and Francis Ⅱ?

图 2 思维链生成的提示模板

Fig. 2 Prompt template generated by chain of thought

首先,设计一个与任务相关的思维链,如图 2 所示. 在这个思维链中,采用了一个特定的统一结构来丰富每个推理过程,例如:"头实体是[实体名称],尾实体是[实体名称],头尾实体之间的关系是[关系描述]".

其次,将少量人工构建的思维链提示作为示例,为大模型提供上下文学习. 这样可以使大模型学会生成特定格式的推理过程.

最后,将完整的数据输入到 LLM 中,补充数据中的推理过程.

2.3.2 思维链评估

为了控制思维链推理数据的质量,提出一种思维链评估方法,该方法通过优化选择过程来定位小样本关系抽取任务中最有价值的推理数据.

该方法通过比较输入实例与推理过程嵌入的一致性来评估思维链的质量. 采用 SimCSE 方法计算输入实例与大模型生成的推理过程之间的相关性,并选择与原实例相似性最高的推理数据,以此作为模型后续抽取所需的推理过程. 具体来说,在 CoT 生成阶段,为数据集中每个示例样本 $s \in S$,生成多个推理解释 $E = \{e_1, e_2, \cdots, e_m\}$. 使用语义相似性评估这些推理解释,选择评分最高的推理 e_i 作为最终的思维链示例.

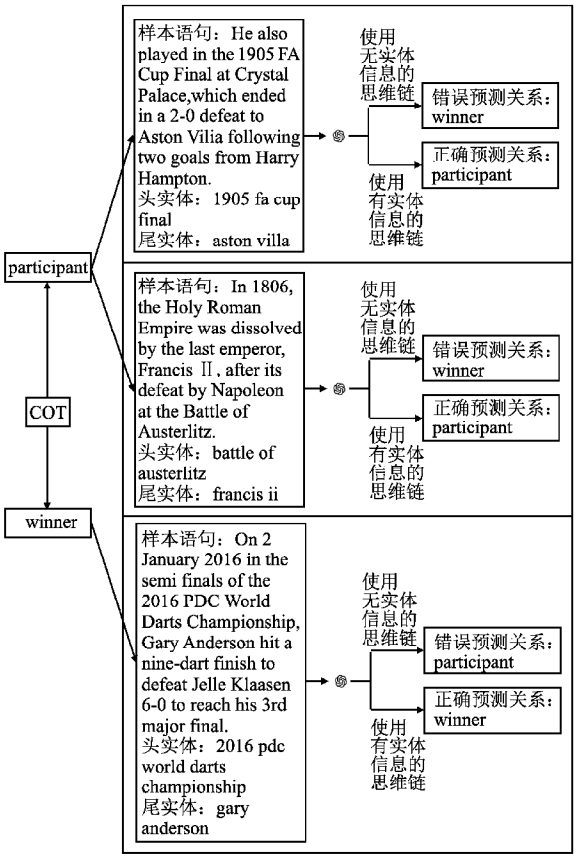

图 3 思维链评估案例

Fig. 3 Chain of thought evaluation case

除此之外,在模型生成推理过程时,如果句子中出现干扰信息或句子结构相似但实体类型等信息却有着很大的不同时,单一使用样本句子嵌入进行相似性对比有着严重的缺陷,尽管推理过程的信息是融合了三元组信息的推理数据,但仍可能存在偏离原实例语义的情况,如图 3 所示,这是一个描述句子结构相似,但抽取关系不同的情况的案例. 这表明缺少实

体描述信息的推理过程可能使得最终的预测不准确,因而有必要利用上下文信息来增强提示. 故通过引入实体以及关系信息来重构上下文. 该方法能够保证最终选择的推理过程和源文本输入实例之间具有信息一致性,如此的推理过程既能维持源文本的语义完整性,又保留了以三元组为核心的信息. 这样可以有效地解决关键推理数据不准确的问题,并在减少大模型生成推理数据上可能造成的偏差或错误的同时,确保了获得最优的思维链推理数据.

2.4　检索模块

在向模型输入演示示例时,并非支持集中的所有实例都对小样本关系抽取有利,同时考虑到大模型输入上下文长度存在限制,因此选择与查询实例相关的示例对模型进行准确的关系抽取至关重要. 大模型的预测结果往往会随着所选演示示例的不同而出现明显的波动. 在研究不同的上下文示例对预测结果的影响后,得出结论在向量空间中更接近测试样本的上下文示例会使模型生成更优的结果.

在 N-Way-K-Shot 任务的情况下,单次输入的示例样本可能无法包含支持集的所有实例,故引入额外的检索模块,采用文本匹配的方法来选择小样本的演示示例. 为了获得特定关系标签的向量表示,针对每个查询集样本 $q_i \in Q$,将单个支持集示例 $S=\{(s_j,h_j,t_j,r_j,e_j \mid j=1,\cdots,n\times k)\}$ 中的源文本 s_j、头实体 h_j、尾实体 t_j、关系标签 r_j、和推理过程 e_j 重新组合,获取它们的语义嵌入作为候选集合,随后根据每个候选实例与查询实例之间的欧氏距离从候选集合中选出 k 个距离较近的实例作为该查询实例 $q_i \in Q$ 的示例样本 $S_i^q=\{s_{i1},s_{i2},\cdots,s_{ik}\}$,最后重新构造提示模板,将由查询实例及其相似示例组成的提示输入到模型中,指导模型进行小样本关系抽取,并采用多轮并行的方式生成关系 $R_i=\{r_{i1},r_{i2},\cdots,r_{ij}\}$. 这种方法可以帮助模型从支持集中捕获更多有用的信息,减少信息冗余.

2.5　事实验证模块

通过先前的工作可以得知,大模型有表现出幻觉的倾向,

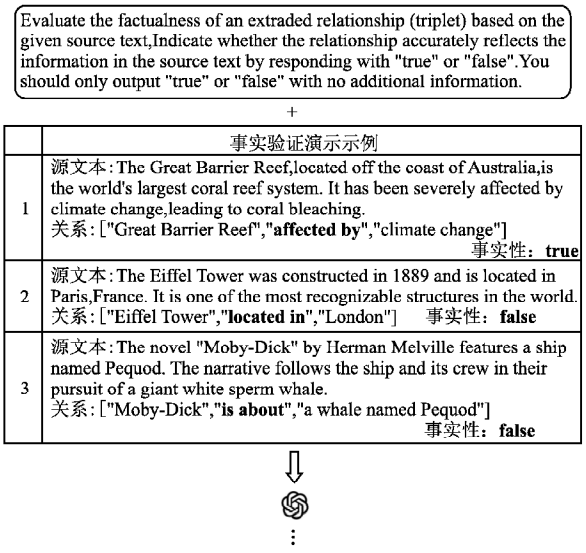

图 4　事实验证的案例

Fig. 4　Fact-verification case

其偶尔会生成看似合理,但可能会偏离用户输入或事实知识的结果. 因此本文通过评估提取的三元组与源文本信息的一致程度作为事实性验证来降低机大模型幻觉. 首先为不同种类的事实性错误设计一个特定的演示示例,随后,使用 GPT-3.5-Turbo 语言模型来评估信息的事实准确性. 该模板由 3 个不同的示例构成,如图 4 所示,这是一个描述模型是否能够准确识别源文本中直接陈述的事实的案例.

示例 1. 用于测试模型是否能够准确识别源文本中直接陈述的事实. 这个示例旨在验证模型在处理明确事实陈述时的能力.

示例 2. 用于评估模型对事实和常识的辨别能力,同时验证模型在检测错误信息方面的能力.

示例 3. 用于评估模型在更复杂的事实验证场景中的能力,即正确解释文学作品中的叙事背景和人物关系,这是一种更为微妙和复杂的事实理解.

这些示例分别用于校准模型识别直接事实陈述、辨别事实和常识、以及解释叙事背景和人物关系的能力,以确保模型在多种事实验证场景中得到全面检查,在事实检索模块中,首先,将该事实验证提示模板作为提示学习中的演示示例,随后将其与查询样本和已预测生成的关系标签输入到大模型中,诱导其输出二进制结果("真"或"假"),结果表示三元组是否来源于该文本的事实,即实现了在整个事实验证模块中,使用先验知识丰富的模型将其验证输出的结果作为事实一致性的评判标准. 针对查询集样本 $q_i \in Q$ 在推理阶段生成的多个结果,分别对这些结果进行事实验证,选取事实验证结果为真且出现次数最多的预测输出,将其中的预测关系作为最终该样本的关系预测结果 r_i.

3　实　验

3.1　数据集

本文分别使用了两个标准的小样本关系抽取数据集对模型进行评估:FewRel 1.0 数据集和 FewRel 2.0 数据集. FewRel 1.0 数据集是一个基于 Wikipedia 的开放数据集,常用于通用领域的小样本关系抽取任务. 它由 70000 个句子组成,标注了 100 个关系标签. 其中有 80 条关系的数据被公开,另外 20 条关系的数据则不公开. FewRel 2.0 数据集是一个基于生物医学领域的数据集,它共有 25 种关系,每个关系包含 100 个实例. 与 FewRel 1.0 数据集相比,其可以有效地检测模型是否能适应少量实例的新领域关系.

3.2　基线模型

为了更准确地评估方法的有效性,本文在小样本关系抽取任务上与两类基线模型进行了全面的比较. 小样本关系抽取基线可以分为两部分:基于大模型的方法以及基于预训练语言模型(训练数据为 100%)的方法. 对于基于大模型的方法,使用了两个模型进行对比. GPT-RE[5] 是一种通过在演示检索中加入任务感知表示的方法,同时使用黄金标签诱导推理逻辑丰富论证. CoT-ER[6] 是一种基于明确证据推理的思维链小样本关系抽取方法. 对于预训练语言模型的方法,使用了 GM_GEN[29]、RAPS[30]、FAEA[31]、HCPR[32] 模型进行对比.

3.3　实验细节

用于对比的基于大模型的基线模型大部分使用 text-da-

vinci-003 作为推理模型,为了公平比较,本文使用 text-davinci-003 作为小样本关系抽取推理模型,通过 OpenAI 的 api 完成调用,temperature 参数设置为 0. 在思维链评估模块的生成阶段中,为每个关系人工标注一些推理过程作为示例样本,使用 GPT-3.5-Turbo 为支持集生成推理数据,在思维链评估模块的评估阶段中,使用 SimCSE 对生成的推理过程进行评估. 对于检索模块,本文使用基于 KNN 的检索方法,对由样本语句和头尾实体重构的文本进行检索. 对于推理模块,使用 GPT-3.5-Turbo 作为事实验证模块的模型.

按照小样本关系抽取的标准配置,在 5-Way 1-shot、5-Way 5-shot、10-Way 1-shot 和 10-Way 5-shot 设置下进行了实验. 对所有基于大模型的方法,通过对验证集中每个 N-Way-K-Shot 任务采样 N×100 个样本来评估.

3.4 实验结果

CoTEV 在 FewRel 1.0 数据集和 FewRel 2.0 数据集上的实验结果如表 1 和表 2 所示,从实验结果中可以发现,在针对不同的 N-Way-K-Shot 任务时,CoTEV 在 FewRel 1.0 数据集和 FewRel 2.0 数据集上的表现均优于现有的基于大模型的基线,与传统的基于预训练语言模型的方法相比,CoTEV 在 5-Way 1-Shot 任务表现出更好的性能. 与之前基于大模型的方法结果相比,CoTEV 在 FewRel 1.0 数据集上的评分最高提升了 1.3%.

表 1 模型在 FewRel 1.0 数据集上的主要实验结果

Table 1 Main experimental results of the model on the FewRel 1.0 dataset

方法	5-Way 1-Shot	5-Way 5-Shot	10-Way 1-Shot	10-Way 5-Shot
预训练语言模型(使用 100% 训练数据)				
GM_GEN	96.97/97.03	98.32/98.34	93.97/94.99	96.58/96.91
FAEA	90.81/95.10	94.24/96.48	84.22/90.12	88.74/92.72
HCPR	94.10/96.42	96.05/97.96	89.13/93.97	93.10/96.46
RAPS	96.28/97.39	97.74/98.00	93.86/95.21	95.39/96.32
大模型(使用 0% 训练数据)				
GPT-RE	94.60/-	95.80/-	87.40/-	91.40/-
+ reasoning	95.40/-	96.40/-	87.60/-	92.40/-
CoT-ER	97.40/-	97.00/-	92.10/-	94.70/-
CoTEV	**97.80**/-	**97.20**/-	**93.40**/-	**95.00**/-

在 FewRel 2.0 数据集上 CoTEV 表现出了更高的性能,超越了大多数完全监督方法. 这表明,当提供高质量的关系信息和设计的推理过程时,GPT 系列的大模型有可能击败以前的完全监督方法. 同时与基于大模型的方法相比,CoTEV 的性能最高提升了 2%.

在使用大模型进行关系抽取的方法中,GPT-RE 是针对小样本关系抽取任务设计的基于大模型的方法,但是该方法用于辅助推理的思维链内容较为简单且质量较低,导致其性能较差. 由于大模型存在最大 token 长度限制,在提示中的演示示例数量较少,且检索模块没有引入实体对信息,GPT-RE 模型在处理较多类别的小样本任务时没有表现出显著的改进. 与之相比,CoTEV 在生成思维链阶段设计了详细的模板,向推理证据中添加了实体关系信息,并在检索示例样本时添加了实体对描述,为模型推理提供了尽可能多的信息. CoT-ER 设计了 3 个推理步骤,且在推理过程中引入了实体信息. 与基线相比,CoTEV 引入了思维链评估模块,使模型生成的推理过程更贴近样本语句和实体关系三元组所表达的信息,同时本文的事实验证模块也可以避免大模型生成样本语句中不存在的内容,在一定程度上解决了大模型机器幻觉的问题. 相比与传统的预训练模型,CoTEV 在一定程度上解决了关系抽取任务对大量训练数据的依赖问题. CoTEV 在 5-Way1-Shot 中表现优异,在其他设置下,CoTEV 可以也在不使用任何训练数据(即 0% 训练数据)的情况下,达到与传统大模型使用 100% 训练数据相近的效果. 这表明 CoTEV 在数据稀缺的场景下具有显著优势.

表 2 模型在 FewRel 2.0 数据集上的主要实验结果

Table 2 Main experimental results of the model on the FewRel 2.0 dataset

方法	5-Way 1-Shot	5-Way 5-Shot	10-Way 1-Shot	10-Way 5-Shot
预训练语言模型(使用 100% 训练数据)				
GM_GEN	76.67	91.28	64.19	84.84
FAEA	73.58	90.10	62.98	80.51
HCPR	76.34	83.03	63.77	72.94
RAPS	80.61	89.59	67.51	82.52
大模型(使用 0% 训练数据)				
GPT-RE	81.34	89.00	70.40	80.70
+ reasoning	78.34	89.80	66.00	73.20
CoT-ER	85.40	93.40	76.10	86.40
CoTEV	**86.00**	**93.60**	**77.50**	**88.40**

在 FewRel 2.0 数据集上,实验中不同模型取得的分数相较 FewRel 1.0 数据集有较大幅度的降低,这是由于 FewRel 2.0 数据集需要相关的医学知识. 但是在 FewRel 2.0 数据集上,CoTEV 有着明显的性能提升,且基于大模型的方法在大多数情况下取得了比传统方法更优的成绩,当提供高质量的关系信息和推理过程时,基于大模型的方法有可能优于以前的全监督方法. 这是因为大模型能够更好地利用上下文信息,包括医学术语之间的关联和语义层面的细微差别. 在关系抽取中,实体之间的关系往往依赖于文本的上下文. 大模型能够捕捉到这些细微的语义差异,使得模型能够在仅有少量标注样本的情况下也能有效学习. 与基于大模型的方法相比,CoTEV 通过思维链评估与事实验证模块为模型提供了更加精细和复杂的推理过程. 通过高质量的推理数据,模型能够进行更有效的逻辑推理,从而在少样本学习中实现更准确的预测.

4 消融实验

4.1 思维链评估模块消融实验

为了证明思维链评估模块的必要性,本节设置了 3 组实验来探讨思维链评估是否有利于关系抽取任务,在本实验中,CoTEV-UniCoT-WithTriplet 组在思维链生成和评估阶段引入三元组信息,生成单个推理过程并对原实例进行上下文重构映射对应推理解释,CoTEV-MultiCoT 组生成多条推理过程,CoTEV-MultiCoT-WithTriplet 组在思维链生成和评估阶段引入三元组信息,多轮生成推理过程并对原实例进行上下文重构映射对应推理解释. CoTEV-MultiCoT 组实验是在思维链生成和评估阶段均未引入三元组信息的 CoTEV-MultiCoT-

WithTriplet 组实验构成. 此外,3 组实验的其余部分保持一致. 实验结果如表 3 所示.

表 3　FewRel 1.0 数据集上思维链评估模块不同重构策略的比较

Table 3　Comparison of different refactoring strategies of the chain of thought evaluation module on the FewRel 1.0 dataset

方法	5-Way 1-Shot	5-Way 5-Shot	10-Way 1-Shot	10-Way 5-Shot
CoTEV-UniCoT-WithTriplet	96.30/-	96.90/-	92.00/-	94.40/-
CoTEV-MultiCoT	73.20/-	82.80/-	63.80/-	75.30/-
CoTEV-MultiCoT-WithTriplet	97.80/	97.20/-	93.40/-	95.00/-

观察发现,推理过程辅助模型推理生成关系的方法相比以前的解决方案有一定的增强,实验数据显示,CoTEV-MultiCoT-WithTriplet 组相比与 CoTEV-UniCoT-WithTriplet 组,它的正确率有着一定的提升,最高提升了 1.5%,这意味着控制推理数据质量是极其必要的. 对比 CoTEV-MultiCoT 组与 CoTEV-MultiCoT-WithTriplet 组的实验的数据,发现所有场景下,CoTEV-MultiCoT-WithTriplet 组的实验结果均优于 CoTEV-MultiCoT 组,性能上平均提升了 22%,这说明缺乏三元组信息的思维链的映射解释会导致 ICL 有效性较差,这意味着关键推理信息不准确的推理过程对模型的关系抽取并无优势. 同时,带有上下文重建的映射推理能确保知识的全面性且原子性,这更有利于大模型在小样本场景中执行关系抽取任务.

4.2　检索模块消融实验

为了探究检索模块在小样本关系抽取任务中的影响,本节设计了两组消融实验. 在 N-Way-K-Shot 任务设置下,CoTEV-NoEntity 组的单个支持集示例仅包含样本语句,而 CoTEV-Entity 组的单个支持集示例则结合了实体对的信息并对其进行了重新组合. 实验结果如表 4 所示.

表 4　FewRel 1.0 数据集上是否包含检索方法的比较

Table 4　Comparison of whether the search methods are included on the FewRel 1.0 dataset

方法	5-Way 1-Shot	5-Way 5-Shot	10-Way 1-Shot	10-Way 5-Shot
CoTEV-NoEntity	95.60/-	96.00/-	92.60/-	94.20/-
CoTEV-Entity	**97.80**/-	**97.20**/-	**93.40**/-	**95.00**/-

实验结果表明,在大多数情况下,使用结合实体对信息的示例能够对小样本关系抽取实现 0.8%-2.2% 的性能提升,这表明在现有的演示检索方法中,实体和关系的相关性不够强. 在 FewRel 1.0 数据集的实验结果中,CoTEV-Entity 组在 5-Way 1-Shot 和 5-Way 5-Shot 设置下分别比 CoTEV-NoEntity 组高出 2.2% 和 1.2%,在 10-Way 1-Shot 和 10-Way 5-Shot 设置下均高出 0.8%,由此发现 CoTEV-Entity 组实现了显著的性能提升. 这是由于在仅有少量示例的新领域中,该方法可以实现良好的领域自适应性,并确保检索的知识具有高相关性,这种改进与模型可以从支持集中捕获更多有用信息的能力密切相关.

4.3　事实验证策略消融实验

本节研究大模型幻觉是否可能导致提取的关系的偏离原文本的问题,消融实验的结果如表 5 所示,CoTEV-Verifier(text-davinci-003)组使用 text-davinci-003 模型自校验的方式. CoTEV-Verifier(GPT-3.5-Turbo)组使用 GPT-3.5-Turbo 作为校验器的核心. CoTEV-NoVerifier 组不包含任何事实验证部分. 此外,3 组实验的其余部分保持一致.

表 5　FewRel1.0 数据集上事实验证模块不同校验方法的比较

Table 5　Comparison of different verification methods for the fact verification module on the FewRel1.0 dataset

方法	5-Way 1-Shot	5-Way 5-Shot	10-Way 1-Shot	10-Way 5-Shot
CoTEV-Verifier(text-davinci-003)	97.60/-	97.10/-	93.00/-	94.80/-
CoTEV-Verifier(GPT-3.5-Turbo)	97.80/-	97.20/-	93.40/-	95.00/-
CoTEV-NoVerifier	96.80/-	96.40/-	92.30/-	94.70/-

实验结果显示,带有事实验证模块的 CoTEV-Verifier(GPT-3.5-Turbo)组和 CoTEV-Verifier(text-davinci-003)组均优于未进行校验的 CoTEV-NoVerifier 组,在不同设置下分别最高提升了 1.1% 和 0.8%. CoTEV-Verifier(text-davinci-003)组的提升显示大模型可以根据自己的回答做出是否正确的判断,这可能与模型并未理解最初的指令有关. 由 CoTEV-Verifier(text-davinci-003)组和 CoTEV-Verifier(GPT-3.5-Turbo)组的结果对比可以发现,使用 GPT-3.5-Turbo 作为验证模型相比使用 text-davinci-003 自验证有着轻微的性能提升,平均在 0.2% ~0.4% 之间,提升幅度不大,这可能时由于两个模型的先验知识和性能上都是相当的,考虑到 text-davinci-003 模型的成本较高,从成本角度,选择 GPT-3.5-Turbo 作为验证模型更为合适.

5　案例分析

为了证明 CoTEV 方法的效果,选择了一个典型的测试例子,如图 5 所示,这种情况需要模型正确识别“2016 PDC World Darts Championship”和“Gary Anderson”之间的关系标签

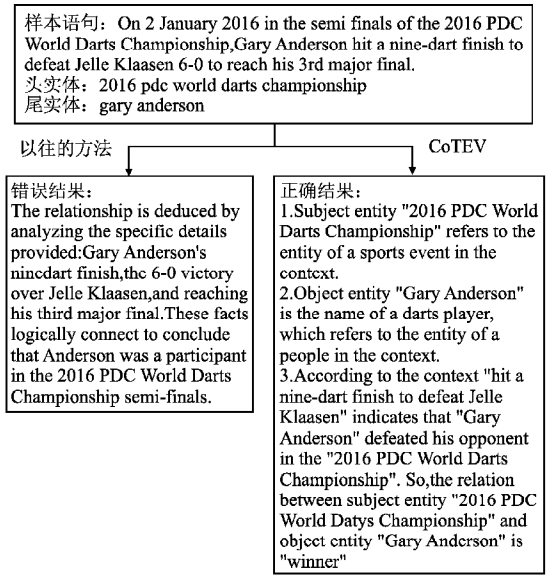

图 5　CoTEV 案例分析

Fig. 5　Case study of CoTEV

“winner”. 在 FewRel 1.0 数据集中,关系标签“winner”被描述为“winner of an event - do not use for awards, nor for wars or battles”. 然而,如果使用先前的以自动生成的思维链作为提示的方法会导致错误的预测. 模型会倾向于将关系标签预测为“participant”,该关系标签在数据集中被描述为“person, group of people or organization (object) that actively takes/took part in an event or process (subject).”. 从关系描述中可以看出二者的头实体和尾实体都可以表示为事件和人/组织. 而关系标签“winner”可以理解为“participants”的子集. 这种方法失败的主要原因是在生成思维链时缺乏更高层次的实体和关系信息,且未对其进行质量评估. 这是模型在理解类似的易混淆关系时所必需的信息. CoTEV 通过在生成思维链时引入实体和关系的描述来将这些关键信息纳入推理过程以解决这个问题,同时为了保证最后生成思维链的质量,将原样本语句和头实体、尾实体进行重构,在原样本中引入实体关系,将重构后的文本与生成的多个思维链进行语义相似性计算,选择结果最优的推理数据作为最终的候选提示. 通过以上方法,大模型在生成思维链时可以获取到足够的关键信息,同时经过评估过滤,可以生成质量更优的思维链,有助于进行后续的推理步骤.

6 结束语

本文提出了一种基于思维链评估与事实验证的生成式小样本关系抽取的方法,与之前的工作不同的是,本文的方法一方面提出了思维链评估的方法,目的是解决在推理过程中控制推理过程的质量,另一方面,提出了事实验证模块,目的是解决大模型幻觉可能导致提取的关系的偏离原文本的问题. 通过实验数据分析,在关系抽取任务的两个标准数据集上,CoTEV 在提升小样本关系抽取的性能方面取得了一定的提升.

尽管大模型在小样本的关系抽取任务上具有巨大的潜力,但是存在一定的计算需求和推理成本. 因此,未来的工作可以继续取探讨如何以较小的代价将 CoT 的推理能力迁移到小模型中,使之能够表现更出色的性能.

References:

[1] Gao T, Han X, Zhu H, et al. FewRel 2.0: towards more challenging few-shot relation classification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 6250-6255.

[2] Brody S, Wu S, Benton A. Towards realistic few-shot relation extraction[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 5338-5345.

[3] Wang Y, Si S, Li D, et al. Two-stage LLM fine-tuning with less specialization and more generalization[C]//International Conference on Learning Representations (ICLR), 2024: 1-19.

[4] Yang H, Zhang Y, Xu J, et al. Unveiling the generalization power of fine-tuned large language models[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024: 884-899.

[5] Wan Z, Cheng F, Mao Z, et al. Gpt-re: in-context learning for relation extraction using large language models[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023: 3534-3547.

[6] Ma X, Li J, Zhang M. Chain of thought with explicit evidence reasoning for few-shot relation extraction[C]//Findings of the Association for Computational Linguistics (EMNLP), 2023: 2334-2352.

[7] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Advances in Neural Information Processing Systems, 2022: 24824-24837.

[8] Sun J, Luo Y, Gong Y, et al. Enhancing chain-of-thoughts prompting with iterative bootstrapping in large language models[C]//Findings of the Association for Computational Linguistics (NAACL), 2024: 4074-4101.

[9] Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models[C]//International Conference on Learning Representations (ICLR), 2023: 1-24.

[10] Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners[C]//Advances in Neural Information Processing Systems, 2022: 22199-22213.

[11] Zhang Z, Zhang A, Li M, et al. Automatic chain of thought prompting in large language models[C]//International Conference on Learning Representations (ICLR), 2014: 1-25.

[12] Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models[J]. Journal of Machine Learning Research, 2024, 25(70): 1-53.

[13] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022: 30016-30030.

[14] Ho N, Schmid L, Yun S Y. Large language models are reasoning teachers[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 14852-14882.

[15] Chowdhery A, Narang S, Devlin J, et al. Palm: scaling language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.

[16] Liu J, Shen D, Zhang Y, et al. What makes good in-context examples for GPT-3? [C]//Proceedings of Deep Learning Inside Out (DeeLIO 2022): the 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, 2022: 100-114.

[17] Zhang K, Lü A, Chen Y, et al. Batch-icl: effective, efficient, and order-agnostic in-context learning[C]//Findings of the Association for Computational Linguistics (ACL), 2024: 10728-10739.

[18] Min S, Lyu X, Holtzman A, et al. Rethinking the role of demonstrations: What makes in-context learning work? [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023: 11048-11064.

[19] Zhao Z, Wallace E, Feng S, et al. Calibrate before use: Improving few-shot performance of language models[C]//International Con-

ference on Machine Learning,2021:12697-12706.

[20] Brown T B,Mann B,Ryder N,et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems,2020:1877-1901.

[21] Gutierrez B J,McNeal N,Washington C,et al. Thinking about gpt-3 in-context learning for biomedical ie? think again[C]//Findings of the Association for Computational Linguistics(EMNLP),2022:4497-4512.

[22] Han X,Zhu H,Yu P,et al. FewRel:a large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing,2018:4803-4809.

[23] Soares L B,FitzGerald N,Ling J,et al. Matching the blanks:distributional similarity for relation learning[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,2019:2895-2905.

[24] Zhang J,Zhu J,Yang Y,et al. Knowledge-enhanced domain adaptation in few-shot relation classification[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining,2021:2183-2191.

[25] Liu X,Ji K,Fu Y,et al. P-tuning v2:prompt tuning can be comparable to fine-tuning universally across scales and tasks[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics,2022:61-68.

[26] Gu Y,Han X,Liu Z,et al. Ppt:pre-trained prompt tuning for few-shot learning[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics,2022:8410-8423.

[27] Zhang P,Lu W. Better few-shot relation extraction with label prompt dropout[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing,2022:6996-7006.

[28] Wang T,Wang Z,Wang R,et al. Contextual information augmented few-shot relation extraction[C]//International Conference on Knowledge Science,Engineering and Management,2023:138-149.

[29] Li W,Qian T. Graph-based model generation for few-shot relation extraction[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing,2022:62-71.

[30] Zhang Y,Cen M,Wu T,et al. RAPS:a novel few-shot relation extraction pipeline with query-information guided attention and adaptive prototype fusion:RAPS for few-shot RE[C]//Proceedings of the International Conference on Modeling,Natural Language Processing and Machine Learning,2024:147-152.

[31] Han J,Cheng B,Lu W. Exploring task difficulty for few-shot relation extraction[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing,2021:2605-2616.

[32] Dou C,Wu S,Zhang X,et al. Function-words adaptively enhanced attention networks for few-shot inverse relation classification[C]//Proceedings of the 31st International Joint Conference on Artificial Intelligence,2022:2937-2943.

第十一届中国数据挖掘会议(CCDM 2026)征稿通知

中国数据挖掘会议(CCDM,China Conference on Data Mining)是由中国计算机学会和中国人工智能学会主办的数据挖掘盛会,每两年举办一次,已成功举办十届.第十一届中国数据挖掘会议(CCDM 2026)将于2026年7月31日~8月2日在山西太原举行,会议由中国计算机学会人工智能与模式识别专业委员会、中国人工智能学会机器学习专业委员会和山西大学承办,欢迎从事数据挖掘研究工作的产学研各界专家、学者以及学生踊跃投稿.

征文范围(包括但不限于):

1.数据挖掘理论与算法;2.特定数据类型的挖掘;3.人工智能与智能信息处理;4.人工智能与数据挖掘技术应用;5.机器学习理论及其应用

论文要求:

1.论文必须未公开发表过,仅接收中文论文,采用《计算机研究与发展》格式排版,字数8000~10000字.

2.论文应包括题目、作者姓名、作者单位、摘要、关键词、正文和参考文献.另附作者通讯地址、邮编、电话及E-mail地址.

3.学生(不包括博士后和在职博士生)第一作者的论文稿件请在首页脚注中注明.

4.会议网址:https://ccf.org.cn/CCDM2026

5.会议采用在线投稿方式,投稿地址:https://conf.ccf.org.cn/CCDM2026/paper

6.联系人:王老师　电话:13453145789　邮箱:ccdm2026@126.com

论文出版:会议录用的论文将推荐到《计算机研究与发展》、《模式识别与人工智能》、《计算机科学与探索》、《计算机工程与应用》、《计算机科学》、《智能系统学报》、《科技通报》、《小型微型计算机系统》、《计算机应用》、《数据采集与处理》、《数据分析与知识发现》、《南京大学学报(自然科学)》、《山东大学学报(工学版)》、《陕西师范大学学报(自然科学版)》、《南京师大学报(自然科学版)》、《南京师范大学学报(工程技术板)》、《郑州大学学报(理学版)》、《郑州大学学报(工学版)》、《华东交通大学学报》、《吉林大学学报(信息科学版)》、《常州大学学报》、《济南大学学报(自然科学版)》等期刊发表,具体发表期刊以最终推荐结果为准.

重要日期:论文投稿截止日期:2026年3月31日;录用通知日期:2026年5月15日;最终版论文提交日期:2026年6月10日

主办单位:中国计算机学会、中国人工智能学会

承办单位:中国计算机学会人工智能与模式识别专业委员会、中国人工智能学会机器学习专业委员会、山西大学

举办地:山西省太原市

特此通知,欢迎各界人士积极投稿参会!