基于大语言模型少样本提示的金融关系抽取

马立开[1],袁梦霆[2]

(1. 武汉大学 国家网络安全学院;2. 武汉大学 计算机学院,湖北 武汉 430072)

摘 要: 针对金融领域语料实体种类多、关系类型复杂、专业术语密集、语义细微差异显著等特点,提出一种结合N-way K-shot少样本学习与多阶段大语言模型提示机制的关系分类方法。以金融语料为基础,构建支持集和查询集,利用预训练语言编码器选择语义相似的示例,并嵌入结构化提示词中引导大模型进行关系识别。在推理过程中,进一步引入自适应验证机制,对模型初步预测结果进行自我检查与校正,从而提升预测稳定性和准确性。实验结果表明,该方法在多个主流大语言模型上,较传统的少样本方法在准确率和召回率方面均有显著提升,且在零样本条件下依然保持出色的性能,验证了其在金融关系抽取中的实用价值与泛化能力。

关键词: 信息抽取;关系抽取;少样本学习;零样本学习;大语言模型

DOI: 10.11907/rjdk.251312 扫描二维码阅读全文:

中图分类号: TP18;TP391.1 **文献标识码:** A **文章编号:** 1672-7800(2026)007-0103-08

Financial Relation Extraction Based on Large Language Model with Few-Shot Demonstration

MA Likai[1], YUAN Mengting[2]

(1. *School of Cyber Science and Engineering*; 2. *School of Computer Science*, *Wuhan University*, *Wuhan* 430072, *China*)

Abstract: To address the challenges posed by financial domain corpora—characterized by diverse entity types, complex relationship structures, dense domain-specific terminology, and subtle semantic distinctions—this paper proposes a relation classification method that integrates N-way K-shot few-shot learning with a multi-stage prompting mechanism for large language models (LLMs). Based on financial text corpora, the method constructs a support set and a query set, leveraging a pretrained language encoder to select semantically similar examples. These examples are embedded into structured prompts to guide the LLM in relation recognition. During inference, an adaptive verification mechanism is introduced to perform self-assessment and correction of the model's initial predictions, thereby enhancing the stability and accuracy of the results. Experimental results demonstrate that the proposed method significantly outperforms traditional few-shot approaches in terms of precision and recall across several mainstream LLMs. Moreover, the method maintains robust performance under zero-shot settings, highlighting its practical value and strong generalization capability in financial relation extraction tasks.

Key Words: information extraction; relation extraction; few-shot learning; zero-shot learning; large language model

0 引言

关系抽取是自然语言处理任务中构建知识图谱的核心步骤之一[1-4]。给定非结构化的自然语言文本以及待判断的实体对,关系抽取任务旨在根据上下文信息对实体对之间的关系类型进行分类。近年来,基于深度学习的关系抽取方法已经成熟并取得了理想效果。然而,这些方法大多基于通用领域大量学习样本的有监督学习,在特定领域中,标注数据往往稀缺,传统有监督方法难以有效应用,限制了领域知识图谱的构建与发展[5-6]。因此,探索少样本条件下的关系抽取方法具有重要意义。

金融领域的关系抽取具有实体种类多、关系类型复杂、专业术语密集、语义细微差异显著等特点。金融文本通常涉及大量数值信息、日期和货币单位,实体之间的关系高度依赖上下文和领域知识,如公司间的并购、投资关系、债务结构等,且金融领域语料多为公告、年报、合同等正式文本,要求模型能捕捉细粒度语义和专业知识,以提

收稿日期: 2025-04-28

基金项目: 国家重点研发计划项目(2022YFB3304300)

作者简介: 马立开(1998-),男,武汉大学国家网络安全学院硕士研究生,研究方向为自然语言处理;袁梦霆(1976-),男,博士,武汉大学计算机学院教授、博士生导师,研究方向为类型系统、编译、高层次综合。本文通讯作者:袁梦霆。

升关系抽取精度和可靠性。目前,金融领域的关系抽取研究相对较少,特别是基于大语言模型少样本提示方法的研究尚属初步探索阶段。一些研究已尝试在金融领域利用领域特化的大语言模型(如FinBERT)提升关系抽取的性能[7]。近期也有研究探讨了如何通过大语言模型的少样本提示学习机制实现关系抽取任务,但这些研究多局限于通用场景,未充分考虑金融领域语料的特殊性与挑战[8-9]。因此,金融领域特化的少样本关系抽取仍有待进一步深入研究。

本文面向金融领域,针对少样本条件下难以支持深度学习模型进行有效监督学习的问题,研究一种高效且准确的关系分类方法,在给定实体对的前提下实现对领域语料中关系类型的准确判别。本文方法无需微调大语言模型,避免了使用大量训练资源,其少样本提示方法与基线少样本学习方法相比,在准确率和召回率上均有大幅度提升。此外,本文还在完全零样本(Zero-Shot)设定下对所提方法进行了系统评估,结果显示即便在无标签样本条件下,模型依然能够实现稳健的关系抽取性能,进一步证明了该方法的通用性与实用性。

1 相关工作

关系抽取(Relation Extraction,RE)是自然语言处理中的一项任务,旨在从文本中识别和提取实体之间的关系。它通常包括从句子或文档中识别出特定类型的实体(如人名、地点、组织等)以及这些实体之间的语义关系(如“位于”“属于”“合作”等)。关系抽取对于构建知识图谱、信息检索和问答系统等应用至关重要。传统的关系抽取方法依赖于规则和模板,而近年来,深度学习技术的引入使得基于神经网络的关系抽取模型取得了显著进展,能够从海量数据中自动学习复杂的关系模式。

1.1 关系抽取定义

关系抽取是结构化知识图谱构建任务中的关键子任务。给定文本T和一对实体(e_1, e_2),关系抽取模型需要从预定义的集合$R = \{r_1, r_2, ..., r_n\}$中,预测$(e_1, e_2)$之间的关系,如式(1)所示。

$$f:(T, e_1, e_2) \rightarrow r, r \in R \tag{1}$$

其中,f是关系抽取模型或方法,R是预设的关系集合。当(e_1, e_2)之间不存在任何已定义的关系时,f可返回无关系。该任务在构建领域知识图谱、信息检索等场景中具有重要应用。

1.2 相关研究

传统的关系抽取方法主要依赖于特征工程和统计模型。随着深度学习的兴起,基于神经网络的方法,如卷积神经网络(CNN)和循环神经网络(RNN),在关系抽取任务中取得了显著进展。近年来,类似于BERT的预训练语言模型(Pre-trained Language Model,PLM)已广泛应用于关系抽取,进一步提升了性能。Zhao等[2]对现有的深度学习关系抽取技术进行了全面综述,提出了新的分类方法,并讨论了其面临的挑战和未来方向。

少样本关系抽取(Few-Shot Relation Extraction,FSRE)聚焦于标注数据稀缺难题,旨在提升模型对新关系类型的识别能力[3,10-11]。元学习和提示学习是FSRE中的两种主要方法。Hang等[12]对结合元学习和提示学习的FSRE方法进行了全面调研,不仅对现有研究进行了细致分类和深入分析,还探讨了当前挑战和未来研究趋势。Ma等[13]提出显式证据推理的思维链方法,利用大型语言模型生成任务相关证据,并将其融入思维链提示中,实现了在零训练数据情况下与完全监督方法相媲美的性能。

在金融领域,Ettaleb等[14]探讨了大型语言模型在经济领域关系抽取任务中的表现。研究发现,通过微调,特别是对开源模型如Llama3的微调,可显著提升提取准确性,并在隐私保护上较传统小模型更具优势。

1.2.1 基于BERT有监督学习的关系抽取

现有基于深度学习和有监督学习的关系抽取方法通常基于预训练模型,如基于变换器的双向编码器表示技术(Bidirectional Encoder Representations from Transformers,BERT),对输入文本进行编码,以获取上下文敏感的词向量表示[15]。如式(2)所示。

$$H = \mathrm{BERT}(T) \tag{2}$$

其中,H是文本的上下文隐藏状态,T为输入的文本,通过大量语料的预训练对语义信息进行表征。

为了捕捉实体和实体对的关系特征,常见方法有:①[CLS]表示法,将BERT输出的“[CLS]”标记的隐藏向量H_{CLS}作为实体对在上下文中的全局表示;②实体标记表示法[16],在实体位置处添加特殊标记,如“[E1]”、“[E2]”等,并提取对应的隐藏状态向量,将其作为实体表示(如H_1、H_2等)。将特征向量通过全连接层和Softmax分类器进行关系预测,如式(3)所示。

$$r = \arg\max\left(\mathrm{Softmax}\left(W\left[H_{cls};\ H_1;\ H_2\right] + b\right)\right) \tag{3}$$

其中,W和b为可训练参数,r是预测的关系类别。模型通常通过交叉熵损失函数进行训练,但这种方法在通用数据领域充足时表现良好,在特定领域数据稀缺时表现会显著下降。

1.2.2 基于少样本学习的关系抽取

少样本关系抽取任务的核心要点在于,在给定极少标注数据的前提下,学习到有效的关系表示并泛化到新的关系类型。现有的少样本关系抽取方法通常依赖于预训练的语言模型(Pretrained Language Model,PLM),这些模型往往使用大量语料进行预训练,可以学习到丰富的语义信息。例如,基于度量和相似度的方法,该方法通过支持集样本,为每一个类别构建一个样本中心并映射到高维向量空间中,将查询集的样本映射到相同的样本空间中,计算其与每个样本中心的欧几里得距离,并将其分类为最接近

的样本中心所属的关系类型。此外，利用BERT预训练任务目标的特点，基于BERT模板的方法将关系分类任务转换为掩码语言模型（Masked Language Model，MLM）训练任务，将关系标签设计为模板填空，让BERT直接预测关系[18]。这种方法更贴近BERT的预训练目标，使其在少样本情况下泛化能力更强。然而，这些方法仍然需要少量的支持集样本进行训练。在新关系类型上，完全零样本（Zero-Shot）情况下效果仍然不理想，且对于未定义的关系，模型容易误分类，泛化能力下降。

1.2.3　基于大语言模型的关系抽取

随着大语言模型（Large Language Model，LLM）的流行，越来越多研究着眼于使用大语言模型进行关系抽取。大语言模型凭借其强大的生成能力，在少样本设定下表现出接近甚至超越全监督方法的性能。例如，Wadhwa等[19]使用GPT-3在多个数据集（如CoNLL04[20]和NYT[21]）上进行实验，通过N-way K-shot的设定在提示词中随机嵌入属于N个随机标签类型、每个标签K个样本，即使在不进行有监督微调的前提下，效果也超越了全监督学习方法。在金融领域，数据通常具有高度的专业性和敏感性，而构建大规模标注数据集的成本高昂且难以获取。因此，结合少样本学习与提示工程进行关系抽取具有重要意义。

2　基于大语言模型的少样本关系抽取

少样本方法可以在极少标注数据的情况下，将大模型泛化到新关系类型，而提示工程则能优化模型的输出格式，提高信息抽取准确性和可控性。金融关系（如企业并购、投资关系、股权结构等）往往涉及复杂的推理过程，因而少样本学习结合提示工程为金融领域的知识图谱构建、风险评估和监管合规等应用提供了一种高效、低成本的解决方案。

2.1　总体方案与流程

如图1所示，本文方案整体流程共分为5个步骤：①构建用于少样本关系分类的支持集与查询集，作为后续预测与验证的基础；②采用基于相似度的支持样本选择策略，从支持集中选取与查询样本语义接近的示例，提升模型泛化能力；③结合数据集的内容设计引导模型完成关系分类任务的结构化提示词；④基于所构建的提示模板，完成第一阶段的关系预测；⑤对第一阶段的预测结果进行验证与筛选，提升最终分类结果准确性与鲁棒性。本文创新主要体现在整体方案流程中的步骤③（提示词设计）和步骤⑤（关系自适应验证）。在步骤③中，针对金融领域关系抽取任务设计了结构化的提示词，以充分捕获细粒度的领域语义；在步骤⑤中，通过关系预测后的自适应验证机制有效降低了大语言模型的幻觉问题，提升了抽取结果的鲁棒性。

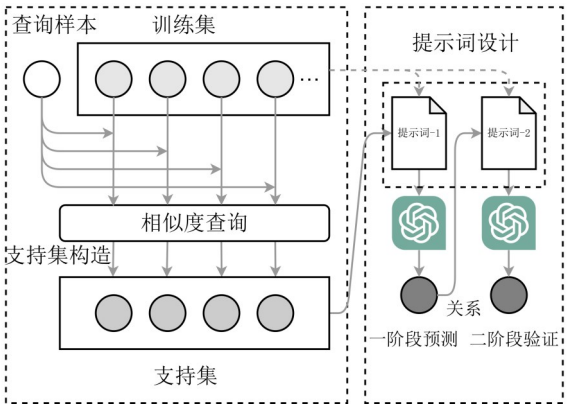

Fig. 1　The framework of the proposed method

图1　所提方法框架

2.2　支持集与查询集构造

少样本提示构建的核心目标是利用少量支持集实例，使大模型能够学习关系抽取任务，并通过适当的提示设计优化模型推理能力。在此设定下，本文方法构造一个支持集（Support Set）S和一个查询集（Query Set）Q，其中支持集S包含N个关系类型，每个关系类型选取K个样本，如式（4）所示。

$$S = \bigcup_{j=1}^{N} \left\{ \left(T_k, e_{1,k}, e_{2,k}, r_j \right) \right\}_{k=1}^{K} \tag{4}$$

其中，共有N个预设关系类型，T_k是第j个预设关系类型中第k个样本所对应样本中的文本，$e_{1,k}$和$e_{2,k}$分别是第k个样本中第1和第2个实体，r_j是第j个关系类型的标签。查询集Q包含需要预测关系的样本，目标是利用S进行推理，如式（5）所示。

$$Q = \left\{ \left(T_q, e_{1,q}, e_{2,q} \right) \right\}_{q=1}^{M} \tag{5}$$

其中，M是查询样本数量，T_q是第q个查询样本，$e_{1,q}$和$e_{2,q}$分别表示第q个样本中第1和第2个实体，并且$S \cap Q = \varnothing$，即查询集中的样本不会出现在支持集中。

2.3　基于语义增强的Few-Shot示例选择

使用BERT等工具获取语义相似的样本，可以确保少样本示例与查询样本的上下文匹配度更高，从而优化大模型的提示词构造。语义增强的示例选择减少了噪声，使模型更容易学习正确的关系模式，从而提升预测准确性[22]。本文使用预训练FinBERT模型计算查询样本$x_q \in Q$与关系r的候选样本$x_{r,i} \in D_r$之间的语义相似度，其中集合D_r为关系r的样本集合[7]。预训练的FinBERT是一种基于BERT架构的金融领域语言模型，它通过在大规模金融语料库上进行无监督学习，从而获得对金融文本的深层语义和上下文理解能力。FinBERT保留了原始BERT的双向Transformer结构，但在金融文本上重新训练，使其词向量和语言建模能力更加贴合金融领域。具体而言，FinBERT的目标是最小化如式（6）所示的联合损失函数。

$$\min_{\theta} \mathbb{E}_{X \sim \mathcal{D}_{pre}} \left[\mathcal{L}_{MLM}(\text{FinBERT}_\theta(X)) + \mathcal{L}_{NSP}(\text{FinBERT}_\theta(X)) \right] \tag{6}$$

其中，θ是模型的参数集合，X是金融预训练语料库

$\mathcal{D}_{pre}$的输入序列，$\mathcal{L}_{MLM}$是掩码语言建模的损失，掩码语言建模任务在输入句子中随机遮蔽部分词语，让模型根据上下文预测被遮蔽的词，学习词的语义和上下文依赖关系。$\mathcal{L}_{NSP}$是下一句预测的损失，目的在于让模型判断第二句话是否为第一句话的真实后续，用于学习句子级的逻辑与篇章结构关系。

设h_q和$h_{r,i}$分别为x_q和$x_{r,i}$的文本嵌入，如式(7)所示。

$$h_q = \text{FinBERT}\left(x_q\right),\ h_{r,i} = \text{FinBERT}\left(x_{r,i}\right) \tag{7}$$

相似度度量采用余弦相似度Sim(·)，如式(8)所示。

$$Sim\left(x_q, x_i\right) = \frac{h_q \cdot h_{r,i}}{\left\|h_q\right\| \cdot \left\|h_{r,i}\right\|} \tag{8}$$

其中，$\|\cdot\|$为取模运算，对于所有的候选样本，计算与查询样本x_q之间的余弦相似度，并选取K个最相似的样本，如式(9)所示。

$$S_{r,K} = \arg \max_{S' \subset D_r, |S'| = K} \sum_{x_i \in S'} \text{Sim}\left(x_q, x_i\right) \tag{9}$$

其中，$S_{r,K}$为最终选择的K个与查询样本最相关的少样本示例的集合。

2.4 提示词设计

金融领域特有的专业术语密集、关系复杂且语义细微差别显著，通常难以依靠传统深度学习方法获得大量标注数据以进行监督学习。而大语言模型经过大规模预训练，天然适合通过少样本提示方式进行知识迁移，在提示工程引导下能迅速适应金融文本的特殊语义环境。因此，本文选择基于大语言模型的少样本提示技术，将其用于金融关系抽取任务，能够有效弥补传统方法在数据稀缺情况下的不足，实现高效准确的关系识别。

提示词设计如图2所示。为了使模型能够充分理解任务需求和相关关系类型，本文结合支持样本语义内容，设计了结构化的提示词模板，引导模型完成关系分类任务。提示词设计主要包括如下部分：

(1)关系类型说明。为所有候选关系类型撰写自然语言说明，引入“e1”和“e2”作为实体占位符，用于表达该关系的一般语义结构。例如，“e1 was acquired by e2”表示“公司e2收购了公司e1”或者“公司e1被公司e2收购”。

(2)少样本示例构造。针对每种关系类型，从支持集中选择语义最相近的K个真实样本，将其中的实体名称替换为占位符“e1”“e2”，并以自然语言形式嵌入提示中，用作范例输入。

(3)实体标注方式。为了清晰标识预测目标中的实体对，真实文本中的实体使用特定符号(如“&&”“$$”等)进行包裹(例如“&& Badgeville && subsequently raised a $12M Series B round in July 2011, led by @@ Norwest Venture Partners @@ and El Dorado Ventures.”。其中，“Badgeville”表示头实体，“Norwest Venture Partners”表示尾实体)。此有助于模型准确聚焦于实体间的语义关系，并压缩输入长度、增强结构清晰度。

该提示模板不仅结合了任务语义信息和真实语料示例，还强化了模型对头尾实体角色的理解，确保大语言模型在推理过程中能够有效迁移已有知识，从而执行更稳定、精确的关系分类判断。

```
Give you a piece of text, head and tail entities from it, identity the relation between them. The head entity is embraced by two "&&"s and the tail entity is embraced by two "@@"s. You should distinguish head and tail clearly.

Candidate relations:
```json
{candidate_types}
```
{examples}
Notice : for an input, output one of the candidate relation types purely, do not output anything beyond that (including process of analysis, thinking, and reasoning).

Real Input:

{real_input}

Real Output:
```

Fig. 2 Prompt of stage 1

图2 阶段一使用到的提示词

2.5 第一阶段少样本关系预测

在完成支持集$S_{r,K}$和小样本提示词$P_{r,K}$的构造后，第一阶段的目标是利用这些提示信息，引导大语言模型对查询样本$x_q = (T_q, e_{1,q}, e_{2,q})$中的实体对关系进行初步预测。

具体而言，查询样本x_q通常为一段包含两个实体$e_{1,q}$和$e_{2,q}$的金融文本T_q，大语言模型通过接收基于支持集构建的提示词$P_{r,K}$，以少样本方式学习如何在相似语义上下文中识别关系类型。提示词中包含了若干由支持集中选出的语义相似示例，每个示例都明确标注了实体对及其对应关系，从而提供了学习范式。在该阶段，大语言模型不对参数进行微调，而是将任务转化为自然语言推理，通过提示驱动的生成方式输出预测结果。设模型输出的预测关系为r_q^*，则该预测过程可形式化为如式(10)所示。

$$r_q^* = f\left(x_q;\ P_{r,K}\right) \tag{10}$$

其中，函数$f(\cdot)$表示在提示词引导下，大语言模型对关系类型的分类函数，参数$P_{r,K}$表示构造的提示模板集，包含了每个关系类型对应的少样本示例。这一阶段的关键优势在于无需训练，即可利用预训练语言模型的知识实现关系识别，尤其适用于标注数据极为稀缺的金融场景。该过程也为后续第二阶段自适应验证提供了初步预测基础。

2.6 第二阶段少样本关系自适应验证

大模型的幻觉问题指模型在缺乏真实信息支持的情况下生成错误或虚假的内容，尤其在金融、法律等专业领域，幻觉可能导致误导性结论，影响关系抽取和决策可靠性[23]。对大模型的输出进行自我验证可以减少幻觉问题，通过让模型在初步预测后重新评估并确认其答案，结合上下文信息修正错误预测，从而提高关系抽取准确性和可靠性。

本文进一步提出了结构化的自适应验证机制，将第一阶段的预测结果r_q^*追加至验证提示词$P'_{r,K}$以构建出完整的二次提示结构。在少样本的设定下，该结构不仅包含与第一阶段相同的少样本支持集示例，还在末尾明确提示模型。

该机制的核心优势在于：不依赖额外人工示例或外部知识，仅通过增强语言提示结构，即可在少样本甚至零样本设定下提升推理稳定性。此外，相比传统思维链提示主要用于生成类任务，本文方法在关系分类任务中引入结构化自我验证，为高风险领域关系识别提供了一种轻量级、可扩展的方案。

将第一阶段的预测结果r_q^*添加到自适应验证提示词$P'_{r,K}$中，要求对第一阶段的结果进行自我检查，如式(11)所示。

$$r_q = f\left(r_q^*;\ P'_{r,K}\right) \tag{11}$$

其中，式(10)和式(11)使用相同的查询模型$f(\cdot)$。

3　实验与结果分析

3.1　实验基线

实验中的基线包括了基于原型网络[17]的Proto-BERT、BERT_{EM}[16]和BERT-Prompt[18]。Proto-BERT使用支持集示例创建关系类型原型，查询示例与关系原型通过计算相似性分数匹配。BERT_{EM}中，实体在上下文中的起始和结束位置被加入了特殊的实体标记符号，将起始实体标记位置的表示拼接起来，作为整个样本的表示，查询样本通过与支持样本的相似度进行分类。BERT-Prompt方法通过将关系分类任务转化为掩码语言建模问题，利用模板将输入改写为包含“[MASK]”的句子，模型根据上下文预测掩码处的词，并将每个关系映射为词表中的具体词项，最终根据预测词的概率选择关系类别，从而实现关系抽取。这种方法充分利用了预训练语言模型的知识，尤其适用于少样本场景。

3.2　实验设置

本文实验采用N+1-way K-Shot设定，即从训练集中每个关系标签r的集合D_r中挑选少样本构成支持集。其中，N为从预设关系集R中随机挑选的关系类型数量，且$N \leqslant |R|, K \in \{0,1,2,3,4,5\}, K \ll \min_{r \in R} |D_r|$，特别地，当$K=0$时，该任务为零样本学习任务。为了充分识别负样本，向支持集S中额外添加了一个负样本标签(None-of-the-Above，NOTA)，即不属于任何已知关系类型的样本。查询集Q来自于包含所有验证集和测试集的样本，用于评估模型的关系分析能力。生成模型分别选用ChatGPT-4o-mini、DeepSeek-R1和Claude 3-Haiku。此外，基线实验所使用的超参数如表1所示。

3.3　数据集

实验中使用FIRE数据集[24]和CORE数据集[25]以评估模型表现。FIRE数据集是一个面向金融领域的命名实体识别与关系抽取任务的数据集，共包含3 025个实例，涵盖13种实体类型和18种关系类型。该数据集主要来自SEC财报和金融新闻文章，所有样本均由单一注释者手工标注，以确保标签的一致性和质量。数据集按照70%训练集、15%验证集和15%测试集的比例划分，其中训练集包含2 117个样本，验证集和测试集各为454个样本。FIRE特别记录了每个样本的标注时间，可用于支持课程学习等机器学习策略。CORE数据集是一个用于评估模型少样本学习能力的金融领域数据集，包含了11个正类型标签和1个负类型标签。其中，训练样本有4 000个，测试样本有708个，将训练样本进一步划分为训练集和验证集，训练集有3 298个样本，验证集有702个样本。

Table 1　Hyperparameters of baseline experiments

表1　基线实验使用的超参数

序号	参数名称	参数解释	参数值
1	max_len	最长输入序列长度	256
2	batch_size	单次训练批大小	1
3	gradient_accumulation_steps	梯度累积步数	64
4	max_steps	总训练步数	500
5	learning_rate	学习率	2×10^{-5}

3.4　评估标准

实验中使用准确率(Precision)、召回率(Recall)和F_1分数以评价本文方法的表现。其中，准确率为模型预测为正类的样本中真正为正类的比例，描述了模型预测准确性；召回率为实际为正类的样本中，被正确预测的比例，描述了模型预测全面性；F_1为准确率和召回率的加权平均，综合衡量两者的平衡性。设真正例(True Positive，即预测正确的正例)为TP，假正例(False Positive，即预测错误的正例)为FP，假负例(False Negative)为FN，则准确率、召回率和F_1的计算方法分别如式(12)—式(14)所示。

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{14}$$

其中，当预测标签为正例且与真实标签完全一致时被视为真正例，否则被视为假正例；当预测标签为负例且真实标签为负例时被视为真负例，否则被视为假负例。

3.5　实验结果分析

表2和表3中，1、2和3是基线方法，分数均引用自Borchert等[25]，标注为“-”的为基线方法中不支持该项或者没有公开，4、5和6是本文研究的实验结果。由表2和表3可知，大语言模型在零样本条件下即具备较强的关系抽取能力。如ChatGPT-4o-mini和DeepSeek-R1在零样本的情况下表现已经接近甚至超越传统基于BERT和有监督方法的最好水平。这表明，现代大语言模型在没有额外示例的情况下仍能较为准确地理解和归纳关系类别。大语言模型在预训练时已经接触到了大量金融领域的知识，这也

预示着其在面临全新领域的信息抽取任务时仍能取得显著效果。

Table 2 Experimental results of few-shot relation extraction on the FIRE dataset

表 2 FIRE 数据集上少样本关系抽取实验结果 (%)

序号	模型名称	5-way 0-shot（零样本）			5-way 1-shot			5-way 5-shot		
		准确率	召回率	F_1	准确率	召回率	F_1	准确率	召回率	F_1
1	Proto-BERT[22]	–	–	–	–	–	27.04	–	–	30.04
2	BERT-Prompt[18]	–	–	–	–	–	75.43	–	–	75.39
3	$BERT_{EM}$[16]	–	–	–	–	–	49.42	–	–	75.56
4	GPT-4o-mini	50.16	82.14	62.28	53.47	93.43	68.01	61.46	77.29	68.47
	+自适应验证	50.85	70.03	59.96	53.27	89.95	66.92	61.62	76.09	68.09
5	DeepSeek-R1	79.25	90.31	84.42	88.77	88.77	88.77	80.20	89.25	84.48
	+自适应验证	80.90	88.26	84.42	90.58	89.61	90.09	83.45	92.86	87.90
6	Claude-3-haiku	32.22	94.70	48.08	53.68	82.16	64.94	54.84	91.67	68.62
	+自适应验证	28.53	84.68	42.68	24.25	85.58	37.79	37.01	87.92	52.09

Table 3 Experimental results of few-shot relation extraction on the CORE dataset

表 3 CORE 数据集上少样本关系抽取实验结果 (%)

序号	模型名称	5-way 0-shot（零样本）			5-way 1-shot			5-way 5-shot		
		准确率	召回率	F_1	准确率	召回率	F_1	准确率	召回率	F_1
1	Proto-BERT[22]	–	–	–	–	–	26.54	–	–	31.96
2	BERT-Prompt[18]	–	–	–	–	–	53.27	–	–	74.16
3	$BERT_{EM}$[16]	–	–	–	–	–	36.27	–	–	44.53
4	GPT-4o-mini	58.97	81.86	68.55	61.80	79.92	69.70	69.54	84.64	76.35
	+自适应验证	67.03	70.83	68.88	71.27	68.71	69.96	76.11	79.36	77.70
5	DeepSeek-R1	80.07	85.66	82.77	84.12	84.41	84.26	88.54	85.00	86.73
	+自适应验证	84.44	86.73	85.57	84.54	87.41	85.93	85.53	88.67	87.07
6	Claude-3-haiku	58.74	86.50	70.00	61.66	80.08	69.67	75.28	71.58	73.38
	+自适应验证	69.86	73.78	71.77	66.79	66.54	66.67	75.30	65.85	70.26

在 CORE 数据集上的实验结果说明了少样本提示应用于具体模型时效果上存在的差异。对于 ChatGPT-4o-mini，随着提示样本数量的增加，其准确率呈持续上升趋势，F_1 分数和召回率虽有波动，但整体表现有所改善，说明该模型对提示示例的语义归纳能力较强。而 DeepSeek-R1 在零样本时对任务的理解已较为充分，提示样本的增加对各项指标的影响相对较小，分数波动不大。这一现象表明，不同大语言模型对少样本提示的依赖程度不同，且其响应在准确率、召回率和 F_1 等不同指标上可能并不一致。对部分大模型而言，少样本提示的作用更多体现在优化边缘案例而非整体性能的显著提升。

从二阶段的自适应验证可以看出（见图 3），不同的大语言模型表现存在显著差异。在 ChatGPT-4o-mini 和 DeepSeek-R1 上，自适应验证在少样本设定下带来了 F_1 分数的小幅度提升。这表明，对于一些大语言模型，自适应验证有助于修正错误预测，提高推理稳定性。然而，在 Claude-3-haiku 模型上，自适应验证机制的引入使得其整体效果不佳。在零样本设定下，仅带来轻微的 F_1 分数提升，而在 1-shot 和 5-shot 条件下，F_1 分数不升反降，尤其在 5-shot 设定中，其 F_1 分数从 73.38% 下降至 70.26%。这表明，Claude 可能对二次推理过度敏感，导致错误修正，即在本不需要修改的情况下改变了正确答案。此外，在提示样本数量逐步增加时，二阶段的自适应验证会给预测结果带来不同的影响。在零样本和 1-shot 设定下，二阶段的自适应验证相较于不使用验证机制时表现出一定的性能提升。这可能是因为在提示样本较少的设定下，模型对关系的理解仍然依赖其内部知识，而非示例样本示例提供的明确模式。模型预测可能更受提示词结构、少样本示例质量的影响，导致部分预测偏差较大。而自适应验证可以让模型重新评估自身推理，提高结果稳健性。在多样本（如 5-shot）的设定下，自适应验证所带来的增益减弱，甚至可能导致下降。这一趋势表明，当少样本示例数量较多时，自适应验证的作用变得不稳定，甚至可能适得其反。

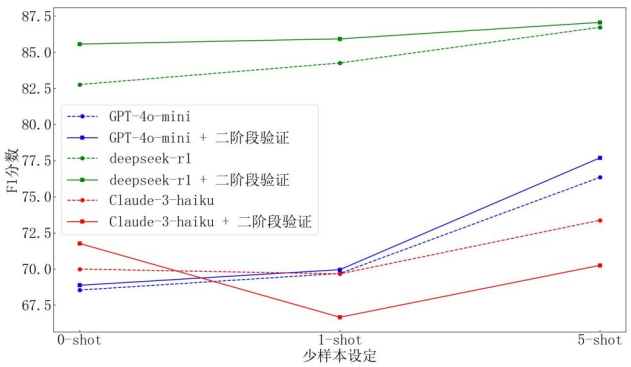

Fig. 3 Experiment results

图 3 实验结果

图 4—图 6 展示了在不同大语言模型（ChatGPT-4o-mini、DeepSeek-R1 和 Claude 3.5）下，针对随机挑选的 6 个关系类别（包含 5 个关系类型与 1 个无关系类别 undefined）在不同样本支持设置（5-way 0-shot、1-shot、5-shot）下的准确率表现，用于分析不同模型对少样本关系分类任务的敏感性与泛化能力。

从整体趋势看，不同关系类别在各模型中的表现具有显著差异，尤其在样本支持数量变化时的响应程度各不相

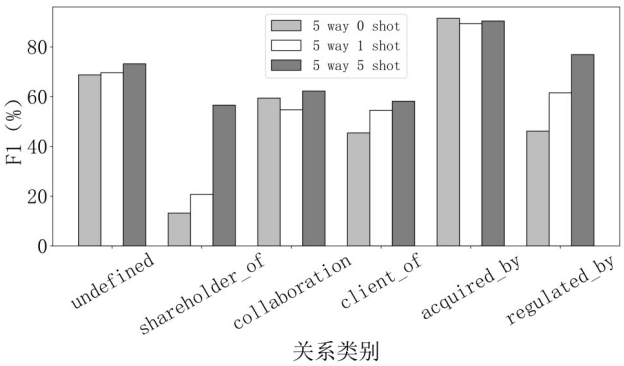

Fig. 4 Example analysis for specific relationship types（ChatGPT-4o-mini）

图 4 针对具体关系类型的实例分析（ChatGPT-4o-mini）

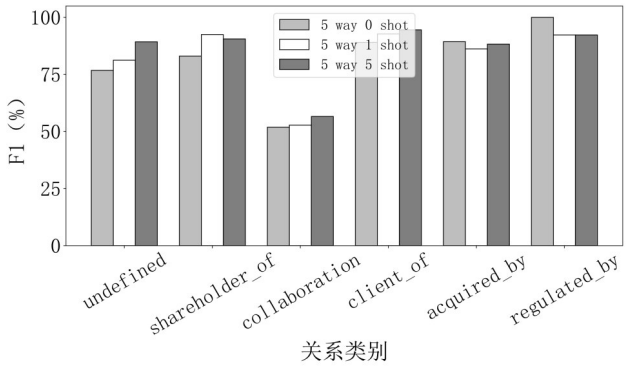

Fig. 5　Example analysis for specific relationship types（DeepSeek-R1）

图5　针对具体关系类型的实例分析（DeepSeek-R1）

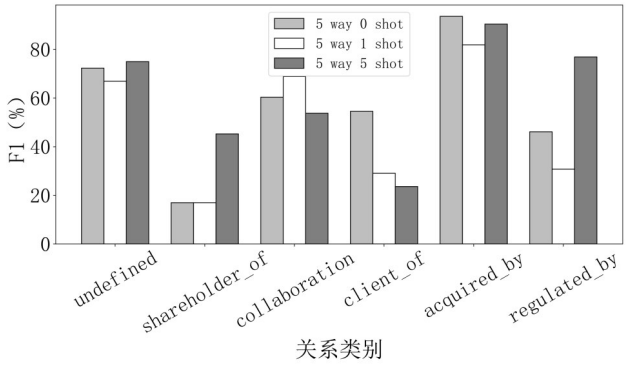

Fig. 6　Example analysis for specific relationship types（ Claude 3.5）

图6　针对具体关系类型的实例分析（Claude 3.5）

同。如图4所示，以ChatGPT-4o-mini为例，其在“acquired_by”与“collaboration”类别下的准确率在0-shot（零样本）情况下已处于较高水平，说明该模型对这类关系的理解能力较强，增加少量样本后准确率波动较小，提升空间有限；而对于如“shareholder_of”等初始准确率较低的类别，随着支持样本数量的增加，模型准确率显著提升，体现出良好的少样本适应能力。如图5所示，DeepSeek-R1表现相对更稳定，其在多数关系类别中，即使在0-shot场景下也能达到较高的准确率，且随样本数增加呈现出轻微但一致的上升趋势，说明该模型对不同关系具有更好的初始泛化能力，同时在few-shot设置下仍能进一步吸收支持样本带来的语义补充。相比之下，如图6所示，Claude 3.5在多数关系上的表现波动较大，缺乏统一趋势。例如，其在“client_of”与“regulated_by”类别下的准确率在添加支持样本后反而下降，可能表明该模型在少样本设置中提示响应不稳定，或存在对输入示例干扰敏感的问题。

因此，对于关系类型区分性强或语义明确的类别，部分大语言模型在无样本的情况下已具备较高判断能力，少样本支持对性能提升影响有限。对于关系语义模糊或模型先验较弱的类别，引入支持样本可以显著提高模型识别准确率。总体来看，引入少量高质量样本对于提升低基准性能类别的表现具有积极作用，而高基准关系类别则可通过结构设计优化效率而非依赖更多样本。

4　结语

本文结合基于类BERT模型相似度搜索的少样本示例选择方法和大语言模型，设计了一种在金融领域少样本甚至零样本条件下实现关系抽取的模型，可在样本稀缺的情况下有效抽取金融类关系。实验结果表明，当前基于大规模语料训练大语言模型在零样本设定下已经具有较强的关系识别能力，效果已大幅度超越了传统基于BERT的少样本学习模型的表现。这表明大语言模型在预训练阶段已经学习到了各领域具有一定深度的知识，使其在特定领域拥有了较强的泛化能力。

研究发现，少样本提示的作用因模型不同而存在差异，对于效果理想的大模型，少样本提示并未带来显著效果提升，甚至可能对模型表现存在负面影响。该现象可能源于大模型自身知识的充分性，使得外部知识价值降低。同时，主流大语言模型已具备较强的自主关系推理能力，对少样本选择策略的依赖性降低。

随着大语言模型的参数规模和预训练语料规模的增长，传统的少样本学习范式可能需要重新评估，并且在知识密集型任务中，大模型的零样本推理能力可能已经足够强大，以至于少样本示例的增量贡献有限。在应用层面，该研究表明，在金融NLP任务中，零样本关系抽取已经具有较高的实用价值，从而减少了对人工示例设计的依赖，从而为实际部署提供了更高的灵活性。

参考文献：

[1] DELAUNAY J, TRAN H T H, GONZÁLEZ-GALLARDO C E, et al. A comprehensive survey of document-level relation extraction (2016-2023) [DB/OL]. https://arxiv. org/abs/2309. 16396.

[2] ZHAO X, DENG Y, YANG M, et al. A comprehensive survey on relation extraction: recent advances and new frontiers [J]. ACM Computing Surveys, 2024, 56(11): 1-39.

[3] LIU B, XU Z M, TAO W, et al. A review on few-sample relation extraction research [J]. Computer Engineering and Applications, 2023, 59 (15): 27-37.
刘蓓，许卓明，陶皖，等. 少样本关系抽取研究综述[J]. 计算机工程与应用，2023，59(15)：27-37.

[4] BASSIGNANA E, PLANK B. What do you mean by relation extraction? a survey on datasets and study on scientific relation classification [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2022: 67-83.

[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 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 3534-3547.

[6] SUN C, JI W, ZHOU G, et al. FGSI: distant supervision for relation extraction method based on fine-grained semantic information [J]. Scientific Reports, 2023, 13(1): 14075.

[7] HUANG A H, WANG H, YANG Y. FinBERT: a large language model for

extracting information from financial text[J]. Contemporary Accounting Research,2023,40(2):806-841.

[8] XU X,ZHU Y,WANG X,et al. How to unleash the power of large language models for few-shot relation extraction?[C]//Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing(SustaiNLP),2023:190-200.

[9] KAUR S,SMILEY C,GUPTA A,et al. REFinD: relation extraction financial dataset[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval,2023:3054-3063.

[10] CHEN J,GENG Y,CHEN Z,et al. Zero-shot and few-shot learning with knowledge graphs: a comprehensive survey[J]. Proceedings of the IEEE,2023,111(6):653-685.

[11] ZHANG Y,GUO Y. A survey of research on few-shot relation classification[J]. Available at SSRN 4822235.

[12] HANG T,WU W,FENG J,et al. A survey of few-shot relation extraction combining meta-learning with prompt learning[DB/OL]. https://arxiv.org/abs/2007.02387.

[13] MA X,LI J,ZHANG M. Chain of thought with explicit evidence reasoning for few-shot relation extraction [DB/OL]. https://arxiv.org/abs/2311.05922.

[14] ETTALEB M,KAMEL M,AUSSENAC-GILLES N,et al. The contribution of LLMs to relation extraction in the economic field[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP),and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal),2025: 175-183.

[15] DEVLIN J,CHANG M W,LEE K,et al. Bert: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2019:4171-4186.

[16] SOARES L B,FITZGERALD N,LING J,et al. Matching the blanks: distributional similarity for relation learning[DB/OL]. https://arxiv.org/abs/1906.03158.

[17] SNELL J,SWERSKY K,ZEMEL R. Prototypical networks for few-shot learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems,2017:4080-4090.

[18] GAO T,FISCH A,CHEN D. Making pre-trained language models better few-shot learners[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing,2021:3816-3830.

[19] WADHWA S,AMIR S,WALLACE B C. Revisiting relation extraction in the era of large language models[C]//Proceedings of the Conference on Association for Computational Linguistics,2023:15566-15589.

[20] ROTH D,YIH W. A linear programming formulation for global inference in natural language tasks[C]//Proceedings of the Eighth Conference on Computational Natural Language Learning(CoNLL-2004),2004: 1-8.

[21] SANDHAUS E. The New York Times annotated corpus[M]. Philadelphia:Linguistic Data Consortium,2008.

[22] LEWIS P,PEREZ E,PIKTUS A,et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. Advances in Neural Information Processing Systems,2020,33:9459-9474.

[23] LI Y,DU Y,ZHOU K,et al. Evaluating object hallucination in large vision-language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,2023:292-305.

[24] HAMAD H,THAKUR A K,KOLLERI N,et al. FIRE: a dataset for financial relation extraction[C]//Findings of the Association for Computational Linguistics:NAACL 2024,2024:3628-3642.

[25] BORCHERT P,DE WEERDT J,COUSSEMENT K,et al. CORE: a few-shot company relation classification dataset for robust domain adaptation [C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,2023:11792-11806.

(责任编辑:孙 娟)