小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0080
2026 年 4 月 第 4 期
Vol.47 No.4 2026

利用微调大语言模型的检索增强文档级多事件抽取

施蒂妮,曾剑平

(复旦大学 计算机科学技术学院,上海 200433)
(教育部网络信息安全审计与监控工程研究中心,上海 200433)
E-mail:zjp@fudan.edu.cn

摘 要: 过去对于文档级事件抽取的研究为了提升对长文本的整体理解能力,需要通过实体识别尽可能地获取全部的可能论元.这对于包含了很多数值词的金融文档级数据而言是一个挑战,实体识别效果往往不理想,错误会传播到后续的事件解码任务.本文提出了一种针对文档级、多事件抽取任务的新方法(RADME),该方法结合了贝叶斯方法实现高效的检索增强技术,并利用了微调大语言模型处理长文本的优势,该方法缩小事件类型检索范围,并从知识库中检索和输入文本最相似的文档的事件类型作为大语言模型输入参考,有效地捕捉全局事件信息.实验结果表明,RADME 在两个公共数据集上的 F1 分数超越了之前最先进的基线方法,分别提升了 3.2% 和 16.5%,证实了该方法的有效性和优越性.

关 键 词: 文档级事件抽取;大语言模型;监督微调;检索增强生成
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)04-0784-09

Retrieval-augmented Document-level Multi-event Extraction with Fine-tuned Large Language Models

SHI Dini, ZENG Jianping
(School of Computer Science Department, Fudan University, Shanghai 200433, China)
(Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, Shanghai 200433, China)

Abstract: Previous research on document-level event extraction has traditionally focused on enhancing comprehension of lengthy texts by identifying all potential arguments through entity recognition. This idea poses challenges for financial documents, which often contain numerous numerical terms, leading to suboptimal entity recognition and error propagation in subsequent event decoding tasks. In this paper, we propose the novel approach of Document-level Multi-event Extraction (RADME), which integrates Bayes to implement efficient retrieval-augmented generation techniques and leverages the advanced long-text processing capabilities of fine-tuned LLMs. RADME narrows the scope of event type retrieval and utilizes the most similar documents as reference inputs for LLMs, effectively capturing global event information. Experimental results demonstrate that RADME outperforms state-of-the-art methods in two public datasets, improving F1 scores by 3.2% and 16.5%, respectively, validating the effectiveness and superiority of the method.

Keywords: document-level event extraction; large language models; supervised fine-tuning; retrieval augmented generation

0 引 言

事件抽取(Event Extraction,EE)是信息抽取研究领域中非常重要的组成部分,旨在将事件信息从非结构化的纯文本中提取为结构化的形式.之前的大量研究主要在 ACE2005 数据集[1]上进行,集中于句子层面的事件抽取(Sentence-level Event Extraction,SEE).然而,在金融领域,由于存在大量文档级别的长文本,如财务公告等,文档级事件抽取(Document-Level Event Extraction,DEE)变得尤为重要.尽管已有一些研究尝试探索文档级事件抽取,但其中的两大挑战仍未得到有效解决:

1)论元分散.与 SEE 任务中事件论元仅限于单一句子提取不同,DEE 任务中的事件论元可能分布于多个不相邻的句子中.如图 1 所示,事件“股份质押”的论元分别出现在不同的句子(S_3、S_4 和 S_8)中,跨度较大.这要求模型不仅能够对整个文本有全局性的把握,还需要具备跨句上下文的长期记忆与理解能力,进而将跨句的论元与事件类型有效结合.

2)多事件.如图 1 所示,文本中同时包含两个事件“股份质押”和“股份减持”的交错叙述.随着文本内容的增加,一个文档中可能涉及的事件数量增多,模型必须能够准确识别事件类型的数量并提取相应的论元.此外,多个事件可能共享某些实体图 1 中“股份质押”事件中的“总持股数量”与“股份减持”事件中的“后续持股数量”存在关联.这种情况对模型的整体理解能力提出了更高的要求,特别是在处理事件间的依赖关系和实体共享时,模型需要具备有效的跨事件信息整合能力.

为了解决文档级事件抽取任务面临的挑战,之前的研究通常采用以下技术手段:1)首先在句子层面进行实体识别,

收稿日期:2025-02-25 收修改稿日期:2025-04-18 基金项目:长三角科技创新共同体联合攻关项目(2023CSJZN0500)资助. 作者简介:施蒂妮,女,2001 年生,硕士研究生,研究方向为自然语言处理、事件抽取;曾剑平(通信作者),男,1973 年生,博士,副教授,研究方向为大数据安全、人工智能安全和金融舆情安全.

以捕获所有潜在的候选论元;2)然后设计解码策略,用于识别事件类型并将识别出的实体填充到预定义的事件模式中.这将 DEE 任务转化为一个事件表格填充任务.

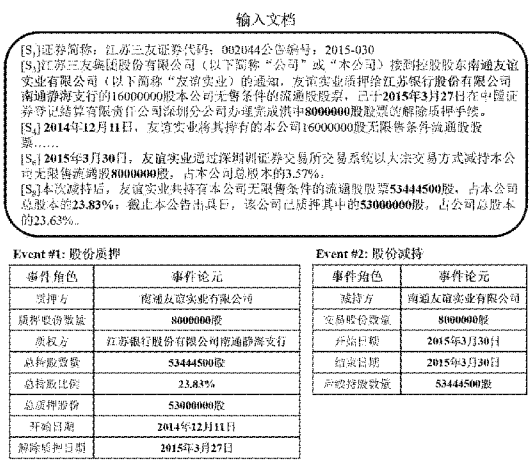

图 1　多事件抽取样例

Fig. 1　Example of multi-event extraction

对于前者,之前的研究通常采用通用的实体识别方法,如基于 BIO 标注的 BiLSTM-CRF 模型[2,3].然而这种方法会引入新的问题,比如图 1 中的“2015 年 3 月 30 日”既是开始日期,也是结束日期,而 BIO 标注难以表达同一实体的不同角色.此外,金融文档中包含大量的数值词(如股票数量和交易金额),早期研究中的通用模型常常错误识别这些实体,实体抽取通常是流水线任务的起点,产生的错误将会引起灾难性的错误传播[4].

对于后者,通常通过图的方法实现多事件类型的解码.Doc2EDAG 通过路径扩展策略从实体构建有向无环图,但这倾向于产生局部最优结果[5].ProCNet 通过使用事件代理节点捕获全局事件信息,但由于不直接对句子之间的关系进行建模,该方法捕获跨句子的长期依赖关系的能力有限,这限制了其在理解多个事件时的准确性[3].

随着大语言模型(Large Language Model,LLM)在各种自然语言处理任务中取得了显著的成果,基于大模型的事件抽取研究也取得了令人瞩目的进展.GOLLIE 通过将事件抽取任务转化为代码语言的格式对大模型进行微调,超越了现有的最先进技术[6].然而,这些基于 LLM 的事件抽取研究主要在 ACE2005 数据集上进行评估,重点集中在句子级的单一事件抽取[7,8],而文档级多事件抽取任务尚未得到充分解决.

为了解决上述问题,本文利用 LLM 和检索增强生成(Retrieval Augmented Generation,RAG)技术来应对金融领域的文档级事件抽取任务,从而摒弃了传统 DEE 中的实体识别任务,将任务分解为事件检测(Event Detection,ED)和事件论元抽取(Event Argument Extraction,EAE)两个子任务.RAG 用于从大规模语料库中检索相关文档,并将其作为上下文输入提供给微调后的 LLM.由于 LLM 在大规模数据集上的广泛预训练和针对特定任务的微调,LLM 具备强大的语义理解能力,能够捕捉复杂的上下文关系和跨句依赖性.通过利用 LLM 处理长文本的能力,本文方法提高了对跨句分布的多个事件和论元的检测能力.此外,ED 任务的输出可以自然地作为 EAE 任务的输入,从而形成一个递归的回答系统,这种集成使得 RAG 成为支持 LLM 提取更准确、上下文适配性更强的结果的理想工具.

本文的主要贡献如下:

1)提出了 RADME(Retrieval-Augmented Document-level Multi-event Extraction)方法,充分利用了大语言模型在长文本理解方面的能力以及检索增强生成在事件检索方面的高效性.该方法有效地解决了文档级事件抽取)任务,特别是在处理多个事件时.与端到端模型不同,RADME 能够成功应对事件检测和事件论元抽这两个任务,这在金融领域尤为重要.

2)在预检索阶段引入了触发词词典和贝叶斯方法,以缩小候选事件类型的范围.通过使用基于阈值的过滤机制,比传统的 top-方法实现了更轻量的解决方案.由于阈值对最终的 F1 分数不太敏感,因此更容易确定.

3)据作者所知,该方法是首个探索 LLM 在 DEE 任务中效果的研究.在两个公开数据集上展示了单事件检测、多事件检测以及相应论元抽取的显著改进.特别是在“被约谈”事件的检测上,提升了 35.9%,充分展示了该方法的有效性.

1　相关研究

1.1　句子级事件抽取

以往的事件抽取研究主要集中于句子级别的任务.Chen 等人将句子级事件抽取任务划分为两个子任务:事件触发词分类和事件论元分类[9].早期的研究通常依赖于手工设计的特征[10,11],或通过构建神经网络模型自动学习句子级特征[9,12,13].近年来,随着语言模型的发展,基于提示学习的方法在 SEE 任务中获得了广泛关注.Hsu 等[14]和 Ma 等[15]通过提示学习,利用预训练语言模型的知识来进行事件抽取,取得了显著的进展.

随着大语言模型(LLM)的出现,研究者们开始探索如何将 LLM 应用于事件抽取任务.GoLLIE[6]和 InstructUIE[7]通过指令微调引导大语言模型执行事件抽取任务,并取得了良好的效果.Code4UIE[16]则提出通过检索技术快速定位与当前任务相关的数据,以辅助 LLM 生成更精确的回答,从而提高事件抽取的准确性.

然而,上述事件抽取方法大多基于 ACE 2005 基准数据集进行实验,研究主要集中在句子内部事件的抽取上.对于跨句事件抽取,尤其是多事件的抽取,仍然面临着较大的挑战.

1.2　文档级事件抽取

在文档级事件抽任务中,研究者们面临的主要挑战是如何从整个文档中有效提取并整合事件信息.这一任务的复杂性源于事件信息通常分布在多个句子或段落中,因此需要捕捉跨句子之间的依赖关系和长距离的上下文信息.

Yang 等人首次提出了关键事件检测模型 DCFEE,通过识别文档中的关键事件句子来指导事件抽取,并在相邻句子中寻找其他相关论元[17].随着 DEE 研究的深入,研究者们开始探索基于图神经网络的方法来建模文档中事件之间的关系.Zheng 等人将事件表格转化为有向无环图,通过路径扩展捕捉跨句子的事件关系[5].Xu 等人在此基础上提出了 GIT 模型,构建了一个异构图神经网络,并通过跟踪机制增强了对

部分解码事件的处理能力[18]. Huang 和 Jia[19]提出了一种通过构建句子社区的方式来捕捉实体与事件之间关系的图模型. Liang 等人则引入了关系增强注意力 Transformer 来建模实体之间的关系[20]. PTPCG 通过构建一个完整的图表示文档中的所有可能事件及其论元关系,并利用伪触发词概念来引导和优化图的构建过程[2]. ProCNet 则通过引入事件代理节点代表文档中的事件,并利用这些节点捕捉事件之间的关系[3]. Wan 等人提出通过构建基于 Token 的双向事件完成图来捕捉文档中词与词之间的关系,进而识别事件触发词和论元[21].

这些方法在建模实体之间的相互关系时,往往依赖于实体识别的准确性,这可能导致错误传播. 图神经网络中的许多节点与实体、句子、事件类型及其论元角色相关,随着节点数量的增加,训练所需的时间显著增加,并且需要大量的计算资源. 此外,部分方法容易陷入局部最优解,或在捕捉跨句子的长期依赖关系时存在局限性.

2 本文方法

本文提出的 RADME 架构如图 2 所示. 该框架结合了大语言模型和检索增强生成技术,将文档级事件抽取任务分解为事件检测和事件论元抽取,将复杂问题简化为两个可顺序解决的子任务. 该迭代回答过程自然契合 RAG 中的任务分解策略,通过检索相似事件,大语言模型可以提供更为准确的回答.

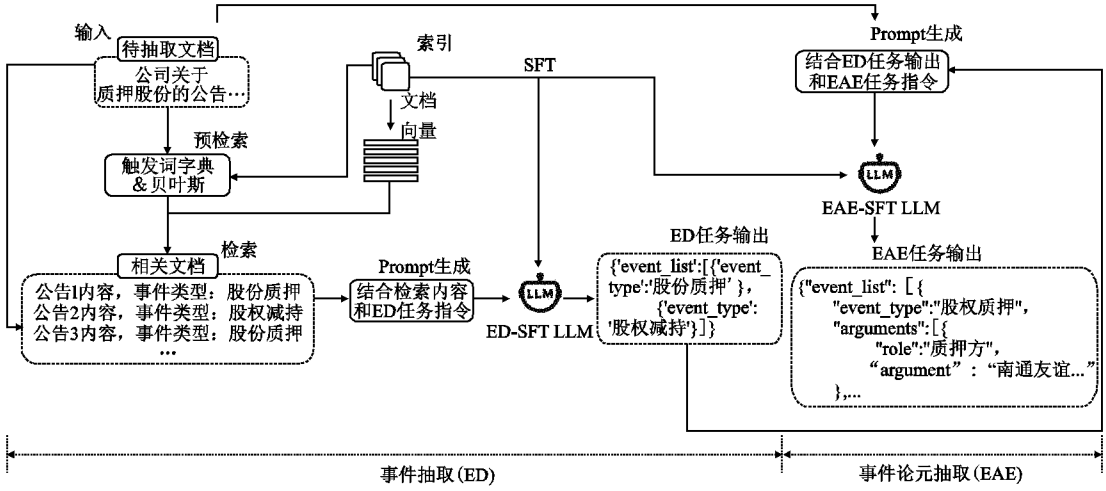

图 2 RADME 模型整体架构图

Fig. 2 Overview of RADME

在金融领域中事件检测是一个关键子任务. 事件类型的准确识别,诸如并购、破产或市场波动,对于风险评估、决策制定和预测分析至关重要[22,23]. 类似地,在法律领域,识别正确的法律事件类型,如合同违约或知识产权争议,对于整理案件信息和推动法律流程也具有重要意义[24,25]. 然而,大多数现有方法依赖于端到端模型,这些模型通常在没有明确进行事件检测的情况下直接进行论元抽取[2,5]. 这类模型往往难以同时满足事件检测和论元抽取的要求,而 RADME 则能够成功解决这两项任务.

2.1 事件检测

本文使用检索增强生方法来检索高度相似的文本,这些文本随后作为候选输入提供给大语言模型. 这一方法在实践中非常有效,因为金融文档通常遵循行业规范和术语,具有较高的一致性[26]. 例如,财务报表和市场分析报告中常使用标准化术语,如“质押股份”和“股权比例”. 因此,包含相似术语、表达方式、主题和内容的文档很可能属于同一类型,这为判断文档类型提供了可靠的依据.

具体来说分为以下 3 个主要部分:

2.1.1 索引

索引阶段是 RAG 的关键组成部分,文档将被处理并转换为嵌入向量,存储在向量数据库中. 索引构建的质量决定了在检索阶段是否能够获得正确的相似度高的上下文. 在预处理数据集的过程中,针对不同格式的数据集都仅保留最少的必要信息,如文本和事件类型. 如果数据集包含触发词(例如在 DuEE-Fin 数据集[27]中)也会将其保存. 同时为了保证后续检索的正确性,对于包含多种事件类型的文档按事件类型进行拆分,因此索引数据库中只包含具有单一类型单实例或单一类型多实例结构的文档.

为了提升文本语义表示的质量,本文使用了 BAAI 通用嵌入(BAAI General Embedding, BGE)模型,该模型在 C-MTEB 基准测试中的检索和语义文本相似度任务中表现优异,且支持最长 8192 个 token 的输入,具备一定的长文本处理能力[28,29]. 区别于其他采用分块策略的 RAG 技术[30],考虑到文档级文本中的事件提及可能跨句,分块处理会导致事件提及被划分到不同块,从而破坏语义完整性并降低检索准确性,因此本文选择保留文本的完整性,避免使用分块策略.

最后,本文使用先进的向量索引技术存储并索引生成的文档向量,利用 FAISS[31]进行高效的相似度搜索.

2.1.2 预检索

在正式进入检索阶段之前,本文进行预检索以优化索引并缩小搜索范围. 此步骤有助于防止检索到与金融事件无关的信息. 例如,图 1 中的文档涉及关于“南通友谊实业有限公司”的事件,如果仅依赖相似性进行检索,会检索到与该公司相关的事件. 通过引入贝叶斯方法来结合事件触发词,增强了

触发词在事件检索中的作用,从而避免了这种问题.

此外,由于本文后续事件检测与事件论元抽取任务依赖于微调后的大语言模型(LLM),其推理过程本身对计算资源具有较高要求,若检索增强生成(RAG)流程设计过于复杂,将进一步增加整体系统的计算负担. 鉴于预检索阶段的核心目标是快速获取与输入文本相关的候选信息集合,而非直接生成最终结果,因此本文在该阶段采用了轻量化的触发词过滤与贝叶斯概率筛选方法,以在保证初步筛选准确性的前提下,显著降低计算开销. 该策略有助于提升整个方法的可扩展性,便于部署于大规模金融文本场景中. 具体实现细节如下:

首先,对于每种类型的金融事件,统计词频构建了事件触发词字典,例如公司上市事件中的"IPO"和企业破产事件中的"破产",并通过同义词搜索扩展字典范围.

然后,基于训练数据中各个事件类型的频率计算每个事件类型 e 的先验概率 $P(e)$、条件概率 $P(t \mid e)$,即在事件类型为 e 的情况下,触发词 t 出现的概率. 使用拉普拉斯平滑来计算条件概率,公式如下:

$$P(t|e) = \frac{\text{trigger\_count}(e,t) + 1}{\sum_{t'} \text{trigger\_count}(e,t') + T} \tag{1}$$

其中,触发词计数 $\text{trigger\_count}(e,t)$ 表示事件类型 e 下触发词 t 的计数,T 是触发词表的大小.

最后,对于给定的输入文本和触发词列表 $\text{triggers} = \{t_1, t_2, \cdots, t_m\}$,计算其在每个事件类型在下的对数概率,这个得分用于评估每个事件类型相对于其他事件类型的可能性:

$$\text{score}(e) = \log P(e) + \sum_{j=1}^{m} \log P(t_j \mid e) \tag{2}$$

然后将对数概率转换为实际概率,并计算归一化后的事件类型概率,选择概率大于阈值 θ 的事件类型作为候选事件类型. 本文通过该阈值平衡事件检测任务的精确度和召回率,较低的阈值会扩大潜在事件类型的范围,从而增加捕捉所有相关事件的机会. 通过调整该阈值,可以在优化 ED 性能的同时,确保 EAE 任务中的 F1 分数保持稳定.

2.1.3　检索

在检索阶段,索引向量支持快速准确的文档检索. 输入的文本被转换成向量,在获得可能的候选事件范围后,对检索进行过滤,限制检索的范围,仅在候选事件范围中使用相似性度量与索引向量进行匹配,从而确保检索到最相关的 top-k 文档. 由于在预检索阶段已设定了阈值,因此此阶段的值不再是决定性因素.

2.1.4　生成

选择检索出的最相似的文档的事件类型和事件检测任务指令与待检测文档内容结合,输入到微调后的 LLM 中以获得事件检测结果. 构建的提示语如图 3 所示.

为了提升通用 LLM 的性能,本文采用了低秩自适应技术(Low-Rank Adaptation, LoRA)[32],在 LlamaFactory 框架[33]下对 ChatGLM3-6B 模型[34]进行了监督微调(Supervised Fine-Tuning, SFT). 尽管通用 LLM 具备广泛的知识和语言处理能力,但它们并未针对特定任务进行优化. 具体而言,事件检测任务需要识别特定类型的金融事件及其相关的上下文信息,而这些要求难以在通用模型的训练数据中得到充分体现. 因此,本文使用带有事件类型标签的监督数据对模型进行指令微调,从而使其在这一特定任务中能够更高效、更具适应性地执行.

Instruction: 你是一个高度智能和精确的金融领域事件检测模型。你将文本作为输入，并检测其中存在的事件类型，事件类型共有以下5种：[股份减持,股份增持,股份回购,股份质押,股份冻结]。一段文本中事件类型可能不止一种，输出文本中含有的所有事件类型，如果文本中没有事件类型，则输出为空。输出应采用以下JSON格式。

Retrieved Content: 输入文本中很有可能含有以下事件类型，但不限于此: [事件类型1, 事件类型2, 事件类型3,...].

Input: 待检测的文档

图 3　事件检测任务的提示词

Fig. 3　Prompt for the ED task

此外,微调对齐了事件检测任务的输入与输出,确保结果可以顺利地传递到后续的 EAE 任务中. 通过微调,模型适应了 ED 任务的数据格式,并确保按照指令生成特定格式的响应. 如果不进行微调,输出可能不符合标准化的 JSON 格式,

提供的文本包含以下事件类型:
- 股权冻结
- 企业布局加速
- 新零售破局新风口

根据您提供的文本，我为您生成的 JSON 输出如下:

```
{"event_list": [
    {"event_type": "股权冻结"},
    {"event_type": "企业布局加速"},
    {"event_type": "新零售破局新风口"}]}
```

图 4　未微调的 LLM 的事件检测任务输出

Fig. 4　ED Response form LLM without SFT

且需要手动调整,另外还可能包括超出指定范围的事件类型,例如图 4 中的"企业布局加速"和"新零售破风"事件类型.

2.2　事件论元抽取

在完成事件检测任务后,根据预定义的事件模式识别与每种事件类型对应的论元角色. 通过将这些论元角色与特定指令相结合,构建 EAE 提示,并输入到 LLM 中以提取事件论元,如图 5 所示. 同样,本文使用带有事件论元标签的监督数

Instruction: 您是一个高度智能和精确的金融领域事件论元抽取模型。您将文本作为输入，并根据提供的事件类型和论元模板，抽取出相应的论元信息。请按照JSON格式输出事件类型和事件论元，如果文本中没有事件类型则输出为空，如果文本中没有该论元则不输出，不需要输出其他内容。

Retrieved Content: 以下输入文本中最可能包含事件类型是: [事件类型1,...]，对应论元模板：[{'event_type': 事件类型1, 'arguments': [....]}]

Input: 待检测的文档

图 5　事件论元抽取任务的提示词

Fig. 5　Prompt for the EAE task

据对模型进行指令微调,以提升其在 EAE 任务中的表现. 该微调过程确保模型能够更准确地识别和提取相关的事件论元.

3　实验设置

3.1　数据集

本文在两个文档级多事件抽取数据集 ChFinAnn[5] 和

DuEE-Fin[27]上进行实验评估. 数据集分布如表 1 所示.

表 1 数据集分布

Table 1 Distributions of datasets

类目	ChFinAnn			DuEE-Fin		
	Train	Dev	Test	Train	Dev	Test
总数量	4806	1602	1602	4911	1637	1637
单事件	3421	1150	1133	3461	1167	1195
多事件	1385	452	469	1450	470	442

ChFinAnn 是迄今为止最大的中文文档级事件抽取数据集,其中 98% 的事件记录论点分散在不同的句子中. 数据集重点研究了 5 种事件类型:股权冻结(Equity Freeze,EF)、股权回购(Equity Repurchase,ER)、股权减持(EquityUnderweight,EU)、股权增持(EquityOverweight,EO)和股权质押(Equity Pledge,EP). 考虑到计算资源的限制本文抽样了 1/4 的样本,按照 6∶2∶2 的比例重新划分训练集、验证集和测试集.

DuEE-Fin 是另一个广泛应用于文档级事件抽取研究的数据集,涵盖了更多样化的事件论元角色,包含约 11,900 篇财务文档和 13 种事件类型,分别是:质押(Equity Pledge,EP)、股份回购(Equity Repurchase,ER)、解除质押(Pledge Release,PR)、被约谈(Regulatory Talk,RT)、企业收购(Business Acquisition,BA)、股东增持(Equity Overweight,EO)、高管变动(Executive Change,EC)、中标(Win Bidding,OB)、公司上市(Company Listing,CL)、企业融资(Enterprise Financing,EF)、亏损(Financial Loss,FL)、股东减持(Equity Underweight,EU)和企业破产(Enterprise Bankrupt,EB). 与 ChFinAnn 不同的是 DuEE-Fin 的每个事件提及均标注了触发词,但未提供序列标注信息. 由于该数据集未发布测试集标签,本文从公开的训练集和测试集中按 6∶2∶2 的比例重新划分了训练集、验证集和测试集.

3.2 评估指标

本文遵循先前研究中设定的评价指标[5]. 对于每个预测记录,通过匹配相同事件类型且共享最多论元的标准记录来进行选择. 通过对比论元,计算 P,R,F1 分数. 由于事件类型通常包含多个角色,采用计算论元角色级别的分数作为直接反映 DEE 能力的最终度量.

3.3 Baselines

为了全面评估本文提出的方法,本文选取了一系列针对金融领域设计的 DEE 任务基线进行比较. 鉴于本文方法仅在 ChFinAnn 数据集的 1/4 子集上进行训练,为确保实验的可比性,本文仅选择了公开代码可复现的方法进行对比实验,包括以下模型:

1)DCFEE[17]通过检测关键事件句子,从识别的中心句子中提取论元,再查询周围句子补充缺失的论元. 该模型有两个变体:DCFEE-O 只从一个文档中生成一条事件记录,而 DCFEE-M 提取多个事件.

2)Doc2EDAG[5]将文档级事件抽取转化为基于实体路径扩展的填充事件表任务,GreedyDec 是 Doc2EDAG 的变体,通过使用已识别的实体角色贪婪填充事件表条目.

3)GIT[18]在 Doc2EDAG 的基础上通过设计了一个异构的基于图的交互模型来捕捉全局交互,并使用跟踪器模块来跟踪路径扩展解码中的事件.

4)PTPCG[2]使用非自回归解码方法,通过修剪的完全图将事件和论元组合在一起.

5)ProCNet[3]通过事件代理节点建立连接,捕捉全局信息,进而利用全局上下文信息来提高模型对不同事件之间关系的理解,再通过最小化 Hausdorff 距离直接优化训练损失.

3.4 LLM 和 SFT 实现细节

由于受到计算资源的限制,本文选择 10B 以下在各个 NLP 任务中表现出色的 ChatGLM3-6B 开源大语言模型进行实验. 为了高效进行微调,使用 LLaMA_Factory 框架并结合 LoRA 技术. 为了减少输出的随机性,将 top_p 设置为 0.7,temperature 设置为 0.1. 实验在两张 NVIDIA GeForce RTX4090(24GB)GPU 上进行,LoRA 适配器的秩(rank)设置为 8,alpha 值为 16,单 GPU 的批次大小(batch size)为 2,梯度累积步数为 8,学习率为 5e-5.

4 实验分析

4.1 整体结果分析

表 2 展示了 RADME 在 ChFinAnn 和 DuEE-Fin 数据集上的实验结果,基准结果使用开源代码进行了重现,其中 F1(S)和 F1(M)分别表示在单一事件(S)和多事件(M)集合下的得分,ΔF1 = F1(S) − F1(M). 本文的方法在这两个数据集上均表现优于基准方法,分别达到了 83.1% 和 83.2% 的平均 F1 分数,在 ChFinAnn 数据集上提升幅度为 3.2% 到 38.1%,在 DuEE-Fin 数据集上提升幅度为 16.5% ~ 53.8%. 结果还表明,采用简单论元补全策略的 DCFEE-O 和 DCFEE-M 表现最差. GreedyDec 相较于 DCFEE 有一定提升,但仍不及

表 2 在 ChFinAnn 和 DuEE-Fin 数据集上的整体 P,R,F1 结果

Table 2 Overall precision (P),recall (R),and F1 scores (F1) on the ChFinAnn and DuEE-Fin datasets

模型	ChFinAnn						DuEE-Fin					
	P	R	F1	F1(S)	F1(M)	F1	P	R	F1	F1(S)	F1(M)	F1
DCFEE-O	54.2	46.2	49.6	55.2	40.9	14.3	51.2	29.7	36.5	42.6	24.1	18.4
DCFEE-M	46.0	44.4	45.0	48.1	40.3	**7.8**	29.9	30.6	29.4	33.2	20.8	12.4
GreedyDec	69.3	42.0	52.0	64.4	33.1	31.3	52.4	36.0	42.2	48.9	31.9	17.0
Doc2EDAG	75.1	60.0	66.3	74.7	55.7	19.0	64.7	40.4	48.6	53.8	39.3	14.6
GIT	75.5	70.6	72.8	81.4	61.0	20.4	63.5	44.4	51.4	54.7	43.0	11.7
PTPCG	80.8	69.3	74.2	83.7	62.6	21.1	71.1	49.3	57.7	66.1	43.3	22.8
ProCNet	83.8	76.4	79.9	89.1	67.6	21.5	72.9	62.3	66.7	70.6	59.4	11.3
RADME(ours)	**85.3**	**81.1**	**83.1**	**90.4**	**73.4**	17.1	**86.3**	**80.5**	**83.2**	**87.0**	**77.5**	**9.5**

Doc2EDAG,尤其是在多事件论元抽取方面. GIT、PTPCG 和 ProCNet 的表现优于早期方法.

还观察到所有模型在处理单事件时 F1 分数普遍高于多事件任务,这突显了从文档中抽取多事件的难度. RADME 模型在 F1(S)和 F1(M)上都取得了较高的分数,并且 ΔF1 保持在较低水平,这种平衡展示了其在处理不同复杂度任务时的鲁棒性和可靠性,使其非常适合实际的多事件抽取应用.

此外,尽管 ChFinAnn 和 DuEE-Fin 的训练集大小相似,但 DuEE-Fin 的任务更具挑战性,因为其包含了更复杂、数量更多的事件类型和论元角色. 之前基于图神经网络的研究在面临事件类型和论元角色数量增长时,其拓扑结构的复杂度会呈现指数级膨胀,导致参数优化困难与特征传播效率下降,所以在 DuEE-Fin 上的表现都有显著下降,这直接反映在 F1 值较 ChFinAnn 平均下降 15.3 个百分点的现象中. 相较之下,本文利用了预训练大模型而非显式构建图结构,有效规避了结构复杂度对模型表现的约束,在两个数据集上的表现相近,凸显了其在不同数据集复杂度下的有效性和鲁棒性.

随着金融领域事件抽取这一任务被越来越多的研究者关注到,近期也有很多研究有不错的表现,比如 SIAT[35] 和 TGIN[36],前者通过计算实体的相对位置编码表示空间交互特征,将其与多粒度语义交互相结合,增强了每对实体之间交互的建模;后者提出了一种新颖的两阶段图推理网络方法,构建了一个异构的文档级图,以捕捉不同粒度节点之间的复杂交互,从而获取文档感知特征,并带有注意力机制的关键信息聚合器,显式地聚合与实体对相关的关键句子. 由于二者并未公开代码,以下引用原文中的结果进行对比:SIAT 方法在 ChFinAnn 数据集上的总体 P,R,F1 分别是 85.9/78.8/82.2,在 DuEE-Fin 数据集上的总体 P,R,F1 分别是 68.1/61.5/64.6,TGIN 方法在 ChFinAnn 数据集上的总体 P,R,F1 分别是 88.8/76.5/82.2,TGIN 未在 DuEE-Fin 数据集上进行实验. 针对实验结果的对比分析,本研究在仅使用 25% 训练集规模的约束条件下,在 ChFinAnn 数据集上取得了与 SIAT(F1 = 82.2)和 TGIN(F1 = 82.2)相当的性能表现,在 DuEE-Fin 数据集上的表现更是远超 SIAT,展现了本文方法的显著优势.

4.2　各事件类型的结果

表 3 和表 4 展示了在 ChFinAnn 数据集上 5 种事件类型和在 DuEE-Fin 数据集上 13 种事件类型的评估结果. 在 ChFinAnn 数据集上,RADME 在 5 种事件类型中的 4 种表现优于其他模型,仅在“股份回购”事件类型上,ProCNet 表现较为突出. 在 DuEE-Fin 数据集上,RADME 在所有 13 种事件类型上均表现出显著的提升,始终超越了现有的最先进基准.

表 3　ChFinAnn 数据集上各事件类型的 F1 分数

Table 3　F1 scores on ChFinAnn dataset with 5 event types

模型	EF	ER	EU	EO	EP
DCFEE-O	41.8	66.5	48.0	41.3	50.7
DCFEE-M	36.5	62.4	45.1	35.6	45.7
GreedyDec	43.3	70.8	51.2	41.4	53.2
Doc2EDAG	52.6	79.0	69.4	64.3	66.4
GIT	66.2	80.8	72.4	72.4	72.1
PTPCG	66.1	82.9	75.5	72.5	74.2
ProCNet	72.0	**91.0**	83.5	72.8	80.0
RADME(ours)	**76.6**	89.2	**84.6**	**81.8**	**83.0**

值得注意的是,在“被约谈”事件类型上,本方法实现了 35.9% 的性能提升,而其他基于图的方法 F1 均低于 50(表 4RT 列). “被约谈”事件在训练集中占比极少,仅为 1%,这使得基于图的方法难以学习到如何有效抽取该事件的论元角色. 基于图神经网络的建模范式本质上依赖于充足的标注数据来构建有效的拓扑表示,从而捕捉实体间的潜在交互模式. 然而,在长尾分布的事件类型下,基于显式图结构的建模方法面临双重挑战:首先,稀疏的标注样本无法支撑图卷积操作所需的邻域统计量,从而导致实体关联矩阵信息缺失显著;其次,GNN 的过平滑现象在低频事件类型中被放大,导致高阶节点特征趋于同质化. 这些结构性缺陷最终使得传统方法在低资源事件类型上的 F1 值显著下降.

表 4　DuEE-Fin 数据集上各事件类型的 F1 分数

Table 4　F1 scores on the DuEE-Fin dataset with 13 event types

模型	EP	ER	PR	RT	BA	EO	EC	WB	CL	EF	FL	EU	EB
DCFEE-O	38.9	54.5	44.3	16.7	30.2	27.5	16.4	33.1	31.6	52.6	52.2	38.4	37.8
DCFEE-M	30.9	45.5	32.2	7.4	24.8	29.2	18.8	26.9	25.0	45.1	45.6	29.0	22.2
GreedyDec	46.5	66.7	59.2	20.7	36.6	36.7	30.6	40.6	29.2	47.4	63.3	43.1	28.0
Doc2EDAG	64.9	72.2	70.5	20.3	39.2	42.5	32.2	49.5	34.5	48.3	68.0	49.5	40.3
GIT	65.1	76.7	68.9	26.2	43.4	44.3	40.9	52.4	34.9	50.7	71.2	56.5	36.8
PTPCG	61.3	83.6	59.3	37.2	54.1	53.1	42.4	66.1	47.8	58.4	72.4	59.3	55.1
ProCNet	69.6	89.8	73.5	49.9	67.1	57.8	50.5	75.8	54.6	68.3	84.7	68.3	57.7
RADME(ours)	**82.7**	**94.5**	**83.3**	**85.8**	**80.3**	**71.0**	**80.9**	**89.1**	**78.1**	**85.9**	**92.2**	**82.6**	**74.9**

与此不同,本方法中,基于大型语言模型的预训练知识能够弥补这一不足,通过预先学习的知识有效补充数据稀缺带来的弊端,利用其强大的上下文理解能力在有限数据的情境下仍能取得较好的效果. 此外,“被约谈”事件的论元角色在 DuEE-Fin 数据集中最少,仅有 4 个:公司名称、被约谈时间、约谈机构和披露时间. 该事件类型论元角色的简洁性使其成为一个较为简单的抽取任务,使得 LLM 能更容易地通过上下文信息和语义线索高效地识别和捕捉事件的触发词及其关联论元,从而超越了基于图的方法.

4.3　单事件 & 多事件

表 5 展示了在 ChFinAnn 数据集上,不同事件类型的单事件(S)和多事件(M)情况下各模型的 F1 得分,其中 SD 列代表各事件 F1 得分的标准差. RADME 在大多数事件类型上始终优于所有基线模型,证明了其在处理单一事件和多事件方面的稳健性. 此外,该方法在不同事件类型上的提取性能变化

较小,标准差始终保持在较低水平.

表 6 和表 7 分别是在 DuEE-Fin 数据集上不同事件类型的单事件(S)和多事件(M)情况下各模型的 F1 得分. 在更具有挑战性的 DuEE-Fin 数据集上,之前的模型普遍表现不佳,特别是在“被约谈”和“高管变动”等事件类型上,SOTA 在多事件抽取中的得分未能超过 60%. 相比之下,该方法在所有事件类型上展现了强劲且稳定的表现,在单事件和多事件抽取中均取得了最低的标准差,表明其结果更加平衡且可靠.

表 5 ChFinAnn 数据集上单事件和多事件抽取的 F1 分数

Table 5 F1 scores of single-event and multi-events on the ChFinAnn dataset

模型	EF		ER		EU		EO		EP		SD	
	S	M	S	M	S	M	S	M	S	M	S	M
DCFEE-O	49.4	35.5	71.7	46.3	53.7	38.9	44.3	38.0	56.9	45.7	10.4	4.8
DCFEE-M	38.1	35.2	65.0	51.2	47.4	41.8	39.1	31.8	50.9	41.7	10.9	7.4
GreedyDec	60.4	26.7	76.7	43.6	60.7	34.2	54.5	23.8	69.7	37.0	8.8	8.0
Doc2EDAG	63.8	44.4	83.9	62.5	76.2	58.9	71.3	56.1	78.1	56.4	7.6	6.8
GIT	79.4	56.0	86.6	59.1	78.5	63.0	79.2	64.5	83.2	62.4	3.4	**3.4**
PTPCG	82.1	52.0	86.1	71.7	83.8	61.9	81.3	62.6	85.1	65.0	**2.0**	7.1
ProCNet	88.6	59.7	**95.8**	**70.1**	89.4	74.4	83.0	61.0	88.8	72.8	4.5	6.8
RADME(ours)	**89.0**	**67.2**	94.8	68.9	**89.6**	**78.4**	**89.2**	**74.6**	**89.5**	**77.8**	2.5	5.1

表 6 DuEE-Fin 数据集上单事件抽取的 F1 分数

Table 6 F1 scores of single-event on the DuEE-Fin dataset

模型	EP	ER	PR	RT	BA	EO	EC	WB	CL	EF	FL	EU	EB	SD
DCFEE-O	38.1	59.5	47.4	25.0	25.9	42.4	27.2	35.5	40.3	57.1	61.4	42.0	51.9	12.5
DCFEE-M	29.4	49.7	39.0	8.7	21.0	36.1	23.6	28.7	31.1	49.2	53.3	32.6	28.6	12.5
GreedyDec	50.4	75.5	63.2	30.8	43.0	41.7	34.0	44.6	32.1	57.7	76.9	54.2	31.2	16.0
Doc2EDAG	60.4	79.4	69.5	29.4	46.6	47.1	36.3	53.0	38.1	57.5	78.9	59.6	44.1	15.7
GIT	58.5	84.6	66.5	17.1	51.3	49.8	40.8	56.1	39.7	60.4	84.6	60.1	41.3	18.4
PTPCG	55.3	89.7	62.7	52.0	63.1	66.9	45.5	74.0	59.4	70.6	86.4	74.2	59.8	12.8
ProCNet	56.1	91.0	75.1	53.9	72.8	59.0	56.5	79.7	62.7	80.8	88.4	72.8	69.5	12.4
RADME(ours)	**74.8**	**97.1**	**78.1**	**82.1**	**88.6**	**75.3**	**77.7**	**93.5**	**92.5**	**92.0**	**97.8**	**89.0**	**92.6**	**8.3**

表 7 DuEE-Fin 数据集上多事件抽取的 F1 分数

Table 7 F1 scores of multi-events on the DuEE-Fin dataset

模型	EP	ER	PR	RT	BA	EO	EC	WB	CL	EF	FL	EU	EB	SD
DCFEE-O	39.0	46.1	43.4	0.0	34.9	17.8	14.7	23.9	17.5	0.0	44.2	32.4	0.0	17.3
DCFEE-M	31.1	38.6	30.7	0.0	28.7	25.0	18.2	20.3	13.8	0.0	39.4	24.0	0.0	13.9
GreedyDec	46.0	50.7	58.4	12.5	24.5	30.5	29.4	29.3	24.5	21.1	44.4	29.6	13.8	14.0
Doc2EDAG	65.4	59.3	70.7	13.3	26.4	36.9	31.0	40.4	28.1	25.1	53.9	37.8	22.2	17.8
GIT	65.8	62.6	69.4	32.7	31.1	38.2	40.9	43.0	26.2	27.6	53.5	52.6	15.4	16.7
PTPCG	62.0	73.1	58.8	25.4	37.6	34.0	41.5	43.5	27.5	27.6	52.8	41.2	37.7	14.6
ProCNet	71.6	86.2	73.1	48.8	57.8	56.9	48.7	59.9	41.7	45.6	77.5	64.5	39.3	14.6
RADME(ours)	**83.4**	**88.5**	**83.7**	**86.4**	**71.5**	**68.2**	**81.7**	**75.9**	**66.3**	**77.9**	**84.2**	**78.2**	**61.8**	**8.3**

4.4 对比基于大模型的方法

尽管 LLM 在多个领域的应用日益广泛,但在金融领域的事件抽取方面的相关研究仍较为稀缺. 本文通过对比实验,探讨了 LLM 在金融事件抽取中的性能表现. 具体而言,本文选用了 ChatGLM3-6B(本文方法的微调基座模型)、Qwen-1.8B[37] 和 Qwen2.5-7B[38](阿里云研发的大型语言模型)进行零样本(zero-shot)实验. 此外,为了进一步验证大模型在事件抽取中的潜力,本文还对比了采用启发式驱动类比链接提示(Heuristic-Driven Link-of-Analogy prompting, HD-LoA)的通用的基于提示工程的事件抽取方法[39].

结果如表 8 所示,文档级事件抽取任务本身具有较高的复杂性,输入文本通常较长,且涉及多种事件类型和复杂的论元角色. 此外,金融领域的事件通常包含大量专业术语和数值词汇,这进一步增加了任务的难度. 因此,直接采用 LLM 进行零样本学习或简单的提示工程往往难以取得理想的效果. 这是因为,零样本学习和提示工程方法通常依赖于模型对任务的普遍理解,然而在面对具有多样事件类型和层次化角色关系的复杂任务时,这些方法可能无法有效捕捉任务中的细粒度特征及其上下文依赖.

表 8 基于大模型的方法在 ChFinAnn 数据集上的结果

Table 8 Results of the methods based on LLM on the ChFinAnn dataset

模型	P	R	F1
ChatGLM3-6B	35.2	31.8	33.4
Qwen-1.8B	33.7	31.2	32.4
Qwen2.5-7B	36.1	33.7	34.9
HD-LoA	38.3	35.5	36.8

相较而言,如表 2 中所示本文的方法 RADME 在 ChFi-

nAnn 数据集上的总体 P,R,F1 分别达到了 85.3/81.1/83.1,这说明了针对特定任务进行大模型微调能够有效优化模型的参数,使其能够专注于特定任务的细节,从而更好地理解和抽取复杂事件类型及多样化的论元角色.因此,微调大语言模型成为提升文档级事件抽取性能的关键步骤.

4.5　消融实验

为了验证 RADME 框架中各主要组件的有效性,本文在更具挑战性的 DuEE-Fin 数据集上进行了消融实验.如表 9 所示,去除 RAG 会导致大多数事件类型的性能下降.这表明,RAG 确实对模型的整体表现起到了积极作用.然而,性能的相对较小下降也表明经过微调的 LLM 已经能够有效地识别各种金融事件,因此 RAG 带来的增益有限.表 9 还展示了在不同贝叶斯阈值下,使用未经微调的 LLM 进行事件检测的结果.实验结果显示,当 $\theta=0.1$ 时,模型的平均 F1 得分最高.随着阈值的增加,更多事件类型被过滤掉,从而导致提取性能逐渐下降.这一变化趋势与表 10 中在不同阈值下事件检索的 F1 得分波动一致.

表 9　DuEE-Fin 数据集上消融实验

Table 9　F1 scores of ablation on the DuEE-Fin dataset

Ablation	EP	ER	PR	RT	BA	EO	EC	WB	CL	EF	FL	EU	EB	Avg
RADME(ours)	**82.7**	**94.5**	**83.3**	85.8	**80.3**	**71.0**	80.9	89.1	**78.1**	**85.9**	**92.2**	**82.6**	**74.9**	**83.2**
w/o RAG	77.6	94.1	78.2	**86.1**	79.9	70.4	**81.3**	**89.3**	77.9	85.2	91.7	78.8	73.9	81.9
w/o SFT($\theta=0$)	18.7	45.1	16.4	5.4	36.0	60.6	41.1	35.7	34.9	44.1	57.3	62.1	24.5	37.1
w/o SFT($\theta=0.1$)	19.7	48.6	16.8	4.8	37.7	55.6	41.2	34.7	29.9	43.6	61.2	65.2	34.4	38.0
w/o SFT($\theta=0.2$)	16.6	45.6	11.8	5.6	39.5	54.9	40.9	34.8	30.7	45.0	60.6	62.3	30.7	36.8
w/o SFT($\theta=0.4$)	16.8	47.1	12.2	7.0	38.9	54.3	42.6	33.1	28.0	44.7	61.0	64.5	29.0	36.9
w/o SFT($\theta=0.6$)	18.6	46.4	14.0	5.5	38.1	51.4	41.8	35.5	26.9	43.4	60.4	62.7	28.0	36.4
w/o SFT($\theta=0.8$)	15.2	44.0	12.9	7.1	41.4	43.8	46.4	33.9	26.6	45.2	61.0	57.1	27.2	35.5

表 10　不同 θ 下事件检索的表现

Table 10　Performance for event retrieval at different θ

θ	0.01	0.1	0.2	0.4	0.6	0.8
P	72.7	75.52	75.6	74.74	71.5	65.25
R	94.13	90.21	83	82.14	82.87	85.57
F1	82.04	82.22	79.13	78.26	76.77	74.04

实验结果表明,模型的 F1 得分对贝叶斯阈值的变化相对不敏感,这使得用户能够在较小的范围内更容易找到一个平衡性能和效率的最优阈值.这种不敏感性在微调后的 LLM 中表现得尤为明显.相比之下,传统 RAG 中设置适当的值以检索前个相似事件可能更具挑战性,因为该值可能会根据具体向量数据库的大小而有较大差异.

5　结论与未来工作

为了应对文档级事件抽取(DEE)任务中论点分散和多事件处理的挑战,本文提出了 RADME 方法,该方法将经过微调的 LLM 的长文本理解能力与 RAG 的事件检索效率相结合.RADME 在处理事件检测(ED)和事件论元抽取(EAE)任务方面表现出了特别的优势,尤其在金融领域具有重要的应用价值.该方法通过触发词字典和贝叶斯方法实现高效的预检索,并采用基于阈值的过滤机制,相较于传统的 top-k 方法,简化了处理流程.该阈值易于优化,因为其对 EAE 任务最终 F1 得分的影响较小.据作者所知,RADME 是首次探索使用 LLM 处理 DEE 任务的方法.在 ChFinAnn 和 DuEE-Fin 数据集上的广泛实验结果验证了该方法在单事件和多事件检测中的有效性和鲁棒性.

本文提出的 RADME 方法,通过结合微调的大型语言模型(LLM)与检索增强生成(RAG)技术,在文档级多事件抽取任务中显著提升了性能.然而,如何进一步提高模型的泛化能力,以应对更广泛的领域与复杂场景,仍是未来研究的关键挑战.未来可结合最新的大型模型优化与强化学习技术,通过动态调整奖励函数优化模型的决策能力,以提升模型对多样化事件模式的适应能力.比如在文档级事件抽取任务中,可设计分层奖励机制:1)局部奖励,侧重事件触发词识别和论元抽取的精确度;2)全局奖励,促进模型识别跨句子事件依赖关系与实体共享模式;3)领域适应奖励,通过对比学习引导模型适应不同领域(如金融、法律、医疗)的文档结构与术语差异.利用 GRPO 等技术,减少模型对于标注数据的依赖,使得能够通过交互反馈自我优化,提升抽取的鲁棒性.

References:

[1] Walker C,Strassel S,Medero J,Maeda K. ACE 2005 multilingual training corpus[EB/OL]. https://catalog. ldc. upenn. edu/LDC2006T06, 2006-02-15.

[2] Zhu T,Qu X Y,Chen W L,et al. Efficient document-level event extraction via pseudo-trigger-aware pruned complete graph[C]// Proceedings of the 31st International Joint Conference on Artificial Intelligence,2022:4552-4558.

[3] Wang X,Gui L,He Y. Document-level multi-event extraction with event proxy nodes and hausdorff distance minimization[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1:Long Papers),2023:10118-10133.

[4] Wan Q,Wan C,Xiao K,et al. Joint document-level event extraction via token-token bidirectional event completed graph[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics(Volume 1:Long Papers),2023:10481-10492.

[5] Zheng S,Cao W,Xu W,et al. Doc2EDAG:an end-to-end document-level framework for Chinese financial event extraction[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP),2019:337-346.

[6] Sainz O,García Ferrero I,Agerri R,et al. Gollie:annotation guidelines improve zero-shot information-extraction[J]. arXiv preprint arXiv:2310.03668,2023.

[7] Wang X,Zhou W,Zu C,et al. Instructuie:multi-task instruction

tuning for unified information extraction[J]. arXiv preprint arXiv:2304.08085,2023.
[8] Xu D,Chen W,Peng W,et al. Large language models for generative information extraction:a survey[J]. Frontiers of Computer Science,2024,18(6):186357,doi:10.1007/s11704-024-40555-y.
[9] Chen Y,Xu L,Liu K,et al. Event extraction via dynamic multi-pooling convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing(Volume 1:Long Papers),2015:167-176.
[10] Li Q,Ji H,Huang L. Joint event extraction via structured prediction with global features[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics(Volume 1:Long Papers),2013:73-82.
[11] Li X,Nguyen T H,Cao K,et al. Improving event detection with abstract meaning representation[C]//Proceedings of the 1st Workshop on Computing News Storylines,2015:11-15.
[12] Chan Y S,Fasching J,Qiu H,et al. Rapid customization for event extraction[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:System Demonstrations,2019:31-36.
[13] Liu J,Chen Y,Liu K,et al. Event extraction as machine reading comprehension[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),2020:1641-1651.
[14] Hsu I H,Huang K H,Boschee E,et al. DEGREE:adata-efficient generation-based event extraction model[C]//Proceedings of the-Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,2022:1890-1908.
[15] Ma Y,Wang Z,Cao Y,et al. Prompt for extraction? PAIE:prompting argument interaction for event argument extraction[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics(Volume 1:Long Papers),2022:6759-6774.
[16] Guo Y,Li Z,Jin X,et al. Retrieval-augmented code generation for universal information extraction[C]//CCF International Conference on Natural Language Processing and Chinese Computing,Singapore:Springer Nature Singapore,2024:30-42.
[17] Yang H,Chen Y,Liu K,et al. Dcfee:a document-level chinese financial event extraction system based on automatically labeled training data[C]//Proceedings of ACL,System Demonstrations,2018:50-55.
[18] Xu R,Liu T,Li L,et al. Document-level event extraction via heterogeneous graph-based interaction model with a tracker[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1:Long Papers),2021:3533-3546.
[19] Huang Y,Jia W. Exploring sentence community for document-level event extraction[C]//Findings of the Association for Computational Linguistics,EMNLP,2021:340-351.
[20] Liang Y,Jiang Z,Yin D,et al. RAAT:relation-augmented attention transformer for relation modeling in document-level event extraction[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,2022:4985-4997.
[21] Wan Q,Wan C,Xiao K,et al. Joint document-level event extraction via token-token bidirectional event completed graph[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics(Volume 1:Long Papers),2023:10481-10492.
[22] Carta S,Consoli S,Piras L,et al. Event detection in finance using hierarchical clustering algorithms on news and tweets[J]. PeerJ Computer Science,2021,7:e438,doi:10.7717/peerj-cs.438.
[23] de Salles D S,Gea C,Mello C E,et al. Multi-scale event detection in financial time series[J]. Computational Economics,2024,65:211-239,doi:10.1007/s10614-024-10582-9.
[24] Yao F,Xiao C,Wang X,et al. LEVEN:alarge-scale Chinese legal event detection dataset[C]//Findings of the Association for Computational Linguistics(ACL),2022:183-201.
[25] Chen Z Z,Ma J,Zhang X,et al. A survey on large language models for critical societal domains:finance,healthcare,and law[J]. arXiv preprint arXiv:2405.01769,2024.
[26] Alali F,Cao L. International financial reporting standards—credible and reliable? An overview[J]. Advances in Accounting,2010,26(1):79-86.
[27] Han C,Zhang J,Li X,et al. DuEE-fin:a large-scale dataset for document-level event extraction[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Cham:Springer International Publishing,2022:172-183.
[28] Zhang P,Xiao S,Liu Z,et al. Retrieve anything to augment large language models[J]. arXiv preprint arXiv:2310.07554,2023.
[29] Xiao S,Liu Z,Zhang P,et al. C-pack:packaged resources to advance general chinese embedding[J]. arXiv preprint arXiv:2309.07597,2023.
[30] Gao Y,Xiong Y,Gao X,et al. Retrieval-augmented generation for large language models:a survey[J]. arXiv preprint arXiv:2312.10997,2023.
[31] Douze M,Guzhva A,Deng C,et al. The faiss library[J]. arXiv preprint arXiv:2401.08281,2024.
[32] Hu E J,Shen Y,Wallis P,et al. LoRA:low-rank adaptation of large language models[C]//International Conference on Learning Representations,2022,doi:10.48550/arXiv.2106.09685.
[33] Zheng Y W,Zhang R C,Zhang J H,et al. LlamaFactory:unified efficient fine-tuning of 100 + language models[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics(Volume 3:System Demonstrations),2024:400-410.
[34] Glm T,Zeng A,Xu B,et al. Chatglm:a family of large language models from glm-130b to glm-4 all tools[J]. arXiv preprint arXiv:2406.12793,2024.
[35] Tao Z,Wang C,Tian Z,et al. SIAT:document-level event extraction via spatiality-augmented interaction model with adaptive thresholding[J]. ACM Transactions on Asian and Low-Resource Language Information Processing,2024,23(10):1-21.
[36] Zhong Y,Shen B,Wang T. TGIN:document-level event extraction with two-phase graph inference network[J]. Neural Networks,2024,176:106343,doi:10.1016/j.neunet.2024.106343.
[37] Bai J,Bai S,Chu Y,et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609,2023.
[38] Yang A,Yang B,Zhang B,et al. Qwen2.5 technical report[J]. arXiv preprint arXiv:2412.15115,2024.
[39] Zhou H,Qian J,Feng Z,et al. LLMs learn task heuristics from demonstrations:aheuristic-driven prompting strategy for document-level event argument extraction[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics(Volume 1:Long Papers),2024:11972-11990.