基于大模型增强的两阶段高效事件共指消解方法

吴耀宗[1]，齐帅[2]，王方圆[2*]，鲍琛龙[1]，唐晋韬[1]

[1]国防科技大学，计算机学院，湖南，410073
[2]中国科学院自动化研究所,复杂系统认知与决策重点实验室，北京，100190

wuyaozong@nudt.edu.cn，{shuai.qi,fangyuan.wang}@ia.ac.cn
{baochenlong,tangjintao}@nudt.edu.cn

摘要

本文针对两阶段事件共指消解方法存在的触发词词目启发机制缺乏同义词聚类能力和小模型理解触发词指代事件能力有限等问题，提出了一种基于大模型增强的两阶段高效的事件共指消解方法，一阶段引入大模型进行同义词聚类，二阶段大模型提供触发词解释文本增强小模型。此外，设计了引导小模型侧重触发词特征向量的损失函数。本文方法在保持近似线性时间复杂度的同时，在ECB+和GVC数据集上的CoNLL F1得分分别提升了2.9和8.0。

关键词： 事件共指消解；大语言模型；触发词词目；计算复杂度

An Efficient Event Coreference Resolution Method Based on Two-Stage Enhancement with Large Language Model

Wu Yaozong[1], Qi Shuai[2], Wang Fangyuan[2*], Bao Chenlong[1], Tang Juntao[1]

[1]National University of Defense Technology, School of Computer Science, Hunan, 410073
[2]Institute of Automation, Chinese Academy of Sciences, Key Laboratory of Complex Systems Cognition and Decision , Beijing, 100190

wuyaozong@nudt.edu.cn，{shuai.qi,fangyuan.wang}@ia.ac.cn
{baochenlong,tangjintao}@nudt.edu.cn

Abstract

To address the limitations of existing two-stage event coreference resolution methods, such as the lack of synonym clustering capability in the trigger word lemma heuristic mechanism and the restricted ability of small models to comprehend trigger word event references, this paper proposes a Large Language Model (LLM) enhanced efficient two-stage event coreference resolution approach. In the first stage, LLM is introduced for synonym clustering, while in the second stage, the LLM provides explanatory text for trigger words to enhance the small model. Additionally, a loss function is designed to guide the small model in focusing on trigger word feature vectors. Experimental results demonstrate that the proposed method maintains near-linear time complexity while achieving CoNLL F1 score improvements of 2.9 and 8.0 on the ECB+ and GVC datasets, respectively.

Keywords: Event Coreference Resolution; Large Language Models; Trigger Lemmas; Computational Complexity

1 引言

事件共指消解是指识别和链接指向同一事件的不同文本表述。例如，在新闻聚合时，需要聚合同一事件的不同新闻报道(Yang et al., 2022; Liu et al., 2024)；在构建事件图谱时，需要融

*通讯作者

合不同来源对同一事件的描述(Chen et al., 2023)。事件共指消解是支撑新闻聚合、知识图谱构建和问答系统等多种应用的重要任务。

当前，事件共指消解面临两个主要挑战(Ahmed et al., 2023)，一是如何准确理解事件指代，即从众多事件中精准链接具有指代关系的事件对；二是如何高效的完成共指消解，确保任务在可接受的时限内完成，即计算复杂度问题。如何高效、精准地实现事件的共指消解始终是众多学者研究的目标。为此，Held等人(Held et al., 2021)提出了一种两阶段方法，一阶段进行初步事件提及聚类，二阶段进行共指事件对判断分类，已成为现阶段事件共指消解的主流框架，但该方法在每个阶段均需要训练独立的模型，且在一阶段需要基于提取的每个事件的特征向量计算两两事件的相似度，计算复杂度是平方级。(Min et al., 2024)使用参数更大的二阶段判别器模型，并使用大模型生成每个事件的总结性文档作为额外特征，有效提升了事件共指消解的精准性，在多个数据集上取得了当前最好的消解效果，但其计算复杂度仍为平方级。(Nath et al., 2024)采用蒸馏的方法将大模型对事件共指对的判别能力迁移到高效语言模型上，取得了与(Min et al., 2024)接近的共指消解效果，但其大模型构造蒸馏语料的方法针对的是事件共指对，计算复杂度是事件个数的平方级。而(Ahmed et al., 2023)在一阶段引入触发词词目的先验统计方法替换原本训练模型以及基于特征向量相似度的计算方法，使得该阶段计算复杂度由平方级降为近似线性，并在二阶段选择参数量较小的高效语言模型作为判别器，大幅减少了时间消耗，但共指消解效果精准性退化较为严重。近年来，大语言模型在自然语言处理各任务上展现出强大的能力(Ma et al., 2023; Fu et al., 2024; Peng et al., 2024)，然而使用未经微调的通用大模型端到端解决事件共指消解问题，在消解效果上尚无法超过经过专业训练的高效语言模型(Nath et al., 2024)，并且大模型推理耗时也无优势。

简而言之，目前尚无有效的事件共指消解方法能够同时兼备线性计算复杂度和强大的共指消解能力。为此，本文提出了一种基于大语言模型增强的两阶段事件共指消解方法，该方法在设计上兼顾了性能和效率。为增强共指消解能力，本方法利用大模型强大的语义理解能力，在一阶段引入基于大模型的同义触发词词目聚类，结合基于触发词词目启发，生成更加完备的候选共指事件对集合；在二阶段结合触发词上下文信息利用大模型生成精准、简洁的触发词解释文本，增强专业判别器模型对触发词指代事件的理解能力；此外，为引导判别器模型更准确的抽取触发词指代事件特征，本文额外设计了一个简易分类损失函数，仅依赖候选共指对的两个触发词特征进行共指判断，融合经典的分类损失函数，可进一步增强共指消解性能。为减少共指消解计算开销，本文借鉴(Ahmed et al., 2023)方法的思路，在一阶段不训练单独模型，基于大模型的同义触发词词目聚类计算复杂度为O(t)，t是测试数据集的事件主题个数，而触发词词目启发的计算复杂度为O(n)，n是测试数据集的事件个数；在二阶段选择参数量较小的高效模型作为判别器，可有效减少训练、推理阶段的耗时。实验表明，本文在保持与(Ahmed et al., 2023)方法同样计算复杂度的同时，通过在两阶段各自引入大模型，在使用相同的Roberta-base(Ericka et al., 2023)高效判别器的情况下，在ECB+(Cybulska et al., 2014)和GVC数据集(Vossen et al., 2018)上的CoNLL F1得分分别为84.6和83.0，相比于(Ahmed et al., 2023)方法分别提高了2.9和8，并大幅缩小了与复杂模型(Min et al., 2024)的差距。

综上所述，本文探索提出了一种基于大模型增强的两阶段高效共指消解方法，在保持近似线性时间复杂度的前提下，显著增强了两阶段高效方法的事件共指消解效果，主要贡献包括：1）提出了一种基于大模型同义词聚类的触发词词目补全方法，能够提升共指事件对的召回效果；2）提出了一种基于大模型触发词解释增强的高效判别器事件共指消解方法，显著增强了判别器对触发词指代事件的理解能力；3）引入了一种辅助简易分类损失函数，进一步提升了共指事件对的判别效果。

2 相关工作

事件共指消解的研究经历了从规则系统和传统机器学习到深度学习的演变过程。早期方法(Baldwin , 1997; Lee et al., 2012; Stoyanov et al., 2012; Lee et al., 2013)主要依赖于手工设计的特征和规则。随着深度学习的发展，基于神经网络的方法逐渐成为主流，研究者开始使用卷积神经网络和循环神经网络对事件提及进行编码(Fang, J et al., 2019; Zeng et al., 2020)。预训练语言模型的引入标志着事件共指消解研究的新阶段，Caciularu等人(Caciularu et al., 2021)提出CDLM模型，通过全局注意力机制处理跨文档信息，但该模型参数量大，计算开销高。Cattan等人(Cattan et al., 2021)选择只使用句子级上下文，提高了效率但可能丢失重要的

文档级信息。Held等人(Held et al., 2021)提出由两个Bert模型构成的两阶段框架，首先在一阶段模型抽取每个事件提及的特征向量，之后通过两两计算相似性筛选出潜在的共指事件对，这些潜在的共指事件对进一步送入二阶段模型做最后的共指判断，两阶段的模型均需进行训练。虽然该方法在多个公开测试集上的性能表现出色，但其计算耗时比较大。相比之下，Ahmed等人(Ahmed et al., 2023)首次从问题分解角度提出使用触发词词目启发机制来解决训练数据正负样本不平衡和计算量大的问题，该方法将事件共指消解分为两个子任务：首先基于对已有的共指标注数据进行先验统计，得出潜在的触发词词目共指集合，然后再结合文本字符串的重叠度过滤掉约90%的非共指对，余下的事件对再送入二阶段判别器模型进行共指判断。此方法显著提高了计算效率，将计算复杂度从平方级降低到近似线性级别。

国内研究逐步从特征驱动模型过渡到结构增强的端到端框，主要集中于模型结构本身的优化。Huan等人(Huan et al., 2023)针对事件提及间的干扰问题，设计基于事件核心句的端到端模型，利用门控机制融合上下文补充信息，有效减少语义干扰并提升事件表示的准确性。Liu等人(Liu et al., 2024)进一步引入显式论元信息与事件链重构机制，通过缓解论元抽取误差传播和数据稀疏性，实现对事件共指链结构的一致性建模。此外，为提升大模型在共指消解任务上的效率，Liu等人(Liu et al., 2025)提出主题结构增强方法，将实体共指转化为问答任务，引入可学习提示模板注入主题信息，仅需少量参数微调即可获得接近全模型性能的结果。

大语言模型依托其强大的理解和生成能力，在自然语言处理的各类任务都能取得卓越的性能。例如，(Xu et al., 2024)基于大语言模型提出了一种信息抽取统一框架，将命名实体识别、关系抽取和事件抽取等任务转化为文本生成问题，通过结构化的输出格式直接生成实体、关系和事件信息；(Zhang et al., 2023)提出了问答重构策略，将信息抽取任务转换为一系列问答交互，利用大语言模型在问答场景中的优势来增强抽取效果；(Huang et al., 2022)研究了多语言生成语言模型在零样本跨语言事件要素抽取中的应用，展示了大模型在处理不同语言事件信息时的泛化能力。然而，直接使用大模型的零样本能力解决事件共指消解任务，面临计算成本高且性能较专业小模型并不占优势的问题(Nath et al., 2024)。目前，多数方法是在多阶段框架下，将大模型的能力蒸馏到小模型上(Nath et al., 2024)，或者使用大模型对小模型训练的语料特征进行增强(Min et al., 2024)。然而，(Nath et al., 2024)的方法在构建蒸馏语料方面时间复杂度为平方级，(Min et al., 2024)的方法在一阶段仍需要训练独立的小语言模型且推理时间复杂度也为平方级。区别于现有大模型增强的事件共指消解方法，本文研究基于(Ahmed et al., 2023)高效两阶段方法的大模型增强方法，在保持近似线性时间复杂度的同时，显著提升该方法的共指消解性能。

3 本文方法

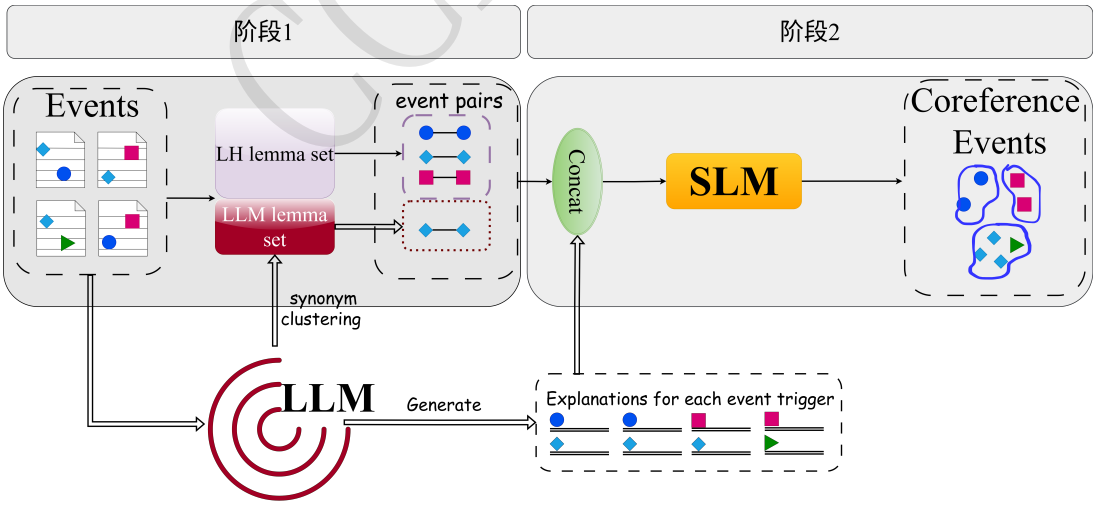

图1.本文事件共指消解方法的流程示意

本文提出的基于大模型增强的两阶段事件共指消解方法遵循两阶段框架，如图1所示。

1）在一阶段，本文方法融合触发词词目启发机制（Lemma Heuristic，LH）和大模型

（Large Language Model，LLM）同义词目聚类得到潜在共指词目集合，过滤掉大量非共指事件对；

2）在二阶段，本文方法利用一阶段输出的事件对训练一个小语言模型（Small Language Model，SLM）做为判别器进行事件共指判别，模型训练不仅采用触发词句子文本，还拼接了大模型对触发词指代信息的解释性文本，同时引入辅助简易分类损失函数，显著增强了高效判别器模型的事件共指理解能力。

3.1 融合触发词词目启发与大模型词目聚类的事件共指对过滤方法

本文方法一阶段采用的融合触发词词目启发（Lemma Heuristic，LH）与大模型词目聚类的事件共指对过滤方法的具体步骤如算法1所示。首先，对于测试数据集某个主题下事件提及集合，利用Spacy工具[0]提取所有事件的触发词词目；然后，利用大模型进行词目同义聚类得到候选共指触发词词目；再后，与利用触发词词目启发在训练数据集上统计得到的共指触发词词目集合进行合并；最后，利用Jaccard相似度，过滤掉阈值过低的候选共指事件，得到输入给二阶段的候选事件共指对集合。

输入: 测试数据集某个主题下事件提及集合Events，每个事件标有触发词triggers
lh_set为由训练数据集基于LH方法统计得到的共指词目集合
输出: 候选事件共指对集合C
C =set() # 初始化C为空集
$lemmas$ =Spacy(triggers)
llm_set = LLM($lemmas, lemma\_cluster\_prompt$) # $lemma_cluster_prompt$为聚类提示词
$our\_set = lh\_set \cup llm\_set$
for $(event_1, event_2)$ **in** $events\_pairs$:
$events_pairs$为Events集合中事件两两组合生成的候选事件对集合
if $[lemma(event_1), lemma(event_2)]$ ∈our_set **and** JC($event_1, event_2$) > $threshold$:
JC代表计算两个触发词各自所在句子间的重叠度
C.append(($event_1$, $event_2$))
return C

算法1.融合触发词词目启发与大模型词目聚类的事件共指对过滤方法

本阶段方法的第一个关键步骤是从训练数据集得到基于触发词词目启发的共指词目集合。Ahmed等人(Ahmed et al., 2023)通过分析发现，可以将所有的事件共指对分为四类，分别是简单共指事件对P^{+}_{easy}、困难非共指事件对P^{-}_{hard}、困难共指事件对P^{+}_{FN}和简单非共指事件对P^{-}_{TN}，其中约90%的事件对属于P^{-}_{TN}的范畴。LH机制作为一种精巧的预处理策略，通过先验统计得到的共指触发词词目集合能够快速剔除大量P^{-}_{TN}。LH机制的核心假设建立在语言学观察基础上：指向同一事件的触发词通常在词汇表达上具有相似性或关联性。这种简单而高效的策略将原本的平方级复杂度问题转化为更可控的近似线性复杂度。

本阶段方法的第二个关键步骤是基于大模型的同义词目聚类，主要解决由于测试集与训练集存在触发词词目分布差异情况下，同义的词目却无法指代共指的事件这一问题。图2展示了融合触发词词目启发与大模型词目聚类的事件共指对过滤的主要流程。如图所示，训练集中未出现过“crash”、“plummet”、“go down”等词目之间的关联，仅使用LH方法生成的词目集合，在测试时会导致“crash”与“plummet”等指代的潜在事件共指对遗漏，仅可得到(1，3)一个候选共指对，而通过引入基于大模型的同义词目聚类，并将得到的词目集合与LH词目集合合并，则可得到(1，3)、(1，2)和(2，3)三个候选共指对。

基于大模型的同义词目聚类通过提示词引导大模型完成，考虑到大模型可能存在的幻觉问题，为确保结果稳定可靠，本文在设计提示词时，将聚类过程分为两个任务：首先，剔除可能存在的含义较为广泛，在没有上下文语境的情况下无法确定其实际含义的触发词词目；然后，再对剩余的触发词进行同义词的聚类。具体提示词设计如表1所示，两个任务的描述是用下划线

[0]https://spacy.io/ model en_core_web_md v3.4

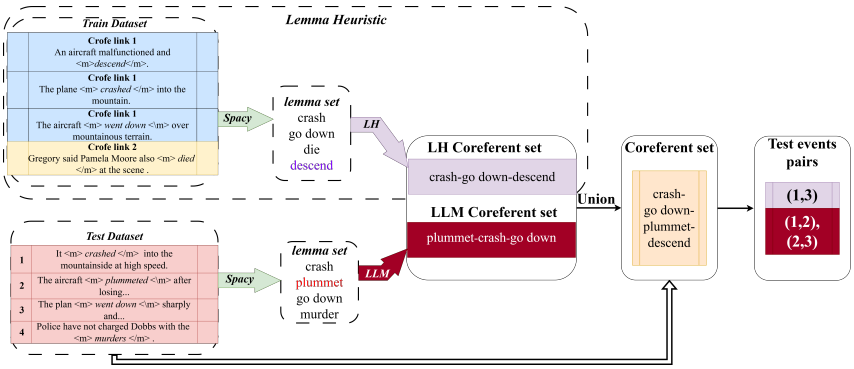

图2.融合触发词词目启发与大模型词目聚类的事件共指对过滤示例

标识。

大模型词目聚类提示词
Based on the given input list, select elements with broad or ambiguous meanings first. Secondly, group synonyms and elements with overlapping words together. Element that does not have a synonym constitutes a group with 1 element on its own. All elements in the input list must exist in output and each element must be only in one group. Words or phrases not in the input list should not exist in output. The format of output is: { 'Ambiguous_words': [...], 'Synonyms_words': [[...], [...],...] } Let's start ! Do not output step by step and output only the final JSON result. Here is the input list: **input_text**

表1.大模型触发词词目聚类使用的提示词

本文提出的基于大模型的同义词目聚类方法，在效果上缓解了LH方法在测试集与训练集的触发词词目分布存在差异导致的候选共指事件对遗漏问题；在计算消耗上，由于大模型每次进行聚类推理针对的是每个主题下所有的触发词词目集合,而同一主题下的触发词词目往往具有较高的重复性，某个主题下的所有触发词词目数量有限，当前大模型的输入长度限制可以满足此类文本输入长度需求，即一个主题可以通过一次推理完成，因此同义词词目聚类过程调用的大模型计算次数与主题个数相同，远远少于事件个数。

对比本文的基线方法(Ahmed et al., 2023)，本文方法在一阶段引入了基于大模型的同义词词目聚类，能够补充召回一些语义一致的触发词代表的事件对，在二阶段基线方法仅使用原始文本作为判别器模型的训练输入，而本文方法将大模型生成的增强信息文本和触发词所在句子结合作为二阶段模型的输入，缓解了只依赖原始文本导致的指代事件理解困难或错误的问题。

对比同样基于大模型增强且性能最好的方法(Min et al., 2024)，本文方法在一阶段不需要训练独立的模型，而是采用无须训练的词目启发机制与大模型同义词聚类方法，方法相对轻量；在二阶段(Min et al., 2024)的方法与本文方法均使用了大模型进行语义特征的增强，区别在于本文在生成触发词指代信息的解释性文本时更好的规避了引入可能会造成干扰的其他触发词信息，能够帮助模型更好理解触发词的指代，且本文方法使用了参数量更小的判别器模型。

3.2 基于大模型触发词指代文本增强的事件共指判别

一阶段过滤了大量的简单非共指事件对，保留下来的候选共指事件对正负样本较初始状态分布更加均衡，训练阶段这些样本用于训练事件共指判别器，可以缓解样本暴漏偏差的问题；测试阶段可以仅将保留下来的候选事件对输入模型进行判断。本阶段，本文提出的基于大模型触发词指代文本增强的事件共指判别方法如算法2所示。首先，对于候选共指对集合中的任意一对事件提及，利用大模型生成触发词指代的增强文本，并与触发词原始所在句子文本进行拼接得到增强后的特征文本；然后，将特征文本输入训练好的判别器模型，得到每对候选共指对的

共指得分；最终，依据阈值判断是否输出至判别的事件共指对集合中。

输入 候选共指对集合C
输出 最终共指对集合R
$R = \text{set}()$ # 初始化R为空集
$D = \text{DiscriminatorModel}()$ # 训练好的判别器模型
for (e_i, e_j) in C:
  $exp\_e_i = \text{LLM\_explanation}(e_i)$ # LLM生成的触发词指代文本增强信息
  $aug\_e_i = \text{Concat}(sen\_e_i, exp\_e_i)$ # 语义加强的句子
  $exp\_e_j = \text{LLM\_explanation}(e_j)$ # LLM生成的触发词指代文本增强信息
  $aug\_e_j = \text{Concat}(sen\_e_j, exp\_e_j)$ # 语义加强的句子
  $p_{ij} = D(aug\_e_i, aug\_e_j)$
  if $p_{ij} > \tau$:
    R.append $\{(e_i, e_j)\}$
return R

算法2.基于大模型触发词指代文本增强的事件共指判别方法

Original Sentence	As part of its efforts to <m> support </m> energy - efficient computing , HP Monday announced it has signed an agreement to acquire facilities consulting firm EYP Mission Critical Facilities for an undisclosed sum .
Explanation of Trigger (OUR)	HP's efforts to <m> support </m> energy-efficient computing.
Summary of Article (OUR)	HP has signed an agreement to acquire EYP Mission Critical Facilities, a facilities consulting firm, to support energy-efficient computing. The deal, expected to close by HP's first fiscal 2008 quarter, will integrate EYP MCF's 350 employees into a new division under HP Services, enhancing HP's ability to address customer needs for more efficient data centers.
Summary of Event (Min et al., 2024)	#support# refers to the actions taken by HP, a multinational information technology company, to promote energy-efficient computing. This support includes the acquisition of EYP, a green consultancy firm, and the development of energy-conserving technology. The acquisition was announced on November 12, 2007.

表2.LLM的信息增强文本对比

由算法2的描述可知，本文二阶段方法在框架上复用了(Ahmed et al., 2023)的思路，区别在于在特征送入判别器之前，对触发词指代信息利用大模型进行了解释增强。对于判别器训练或推理应该输入哪些信息，相关研究尚无明确定论。(Ahmed et al., 2023)曾对比了触发词所在句子和全篇章的效果，如果仅用触发词所在句子，可能会遗漏篇章级上下文的重要信息，而如果使用全篇章作为输入，会引入大量与触发词指代事件无关的信息，导致模型学习困难，且可能会超出小模型的输入词元长度限制。(Min et al., 2024)在利用大模型增强时，没有直接采用整个篇章的文本，而是利用大模型结合篇章对触发词事件进行了总结，压缩了文本的长度，取得了不错的效果。本文对比了篇章级总结和触发词级指代解释两种方式。篇章级总结方面，首先将文章中出现的所有触发词进行标记，然后将带标记的文章文本送入大模型，设计提示词要求大模型对整篇文章进行文本总结的同时,保证不丢失任何标记触发词的指代信息，这一要求是为了防止丢失关键背景事件信息.另外还在提示词中限制输出总结文本的长度，防止输入小模型时因过长而被截断。触发词指代解释方面，设计提示词让大模型对单个触发词的指代内容进行简短的文本解释，增强模型对单个触发词的信息捕捉能力，禁止大模型在解释时提及其他的触发词

和事件信息，保证解释文本中事件的独立性和唯一性。(Min et al., 2024)的解释文本和本文的两种解释文本示例如表2所示，其中，在(Min et al., 2024)生成的解释文本中用下划线标识了一些可能会干扰理解目标触发词指代信息的其他触发词。本文实验对比了几种解释方式，最终算法选择表现最优的触发词指代解释文本作为增强特征。该特征生成所用提示词设计如表3所示。

触发词指代解释文本
Based on the content of the entire text, identify the reference action associated with the trigger words located between <m> and </m>. The most crucial aspect is to distinguish between actions referred to by the same word when it appears in different positions within the text. The format of output is a dict:{"trigger":...,"explanation":...}.The value of "trigger" are the trigger words. The value of "explanation" must be a short sentence or phrase for explaining the action and do not mention events referred to by other unmarked trigger words. Here are some input and output examples for different types: ‴ ... ‴ Here is the input: ‴ **input_text** ‴ Output only one dict. Do not output step by step. Do not output your analysis.The words in explanation describing date, location, and individual must be clear. Words with unclear reference like "the man", "that day", "the region","he","it" are forbidden. Output:

表3.大模型生成触发词指代文本解释的提示词

本文提出的二阶段方法，在判别器模型选择方面与(Ahmed et al., 2023)一致，选择参数量较小的高效模型（如RoBERTa-base）以提升效率。

3.3 损失函数

通常，事件共指消解判别器模型，如(Ahmed et al., 2023)、(Min et al., 2024)等方法，选择使用类别交叉熵作为模型训练的损失函数，即潜在共指对的特征向量通过一个分类器之后计算共指得分p，并与标签y统计交叉熵，计算公式如下：

$$L_{cls} = -[y \log(p) + (1-y) \log(1-p)] \tag{1}$$

上述损失函数在计算特征向量的类别得分时，采用触发词、触发词句子或增强文本等拼接特征，但触发词特征可能缺失上下文信息，而额外拼接的上下文信息中可能存在一些干扰模型准确判别触发词指代的信息，不利于模型分类。为引导模型更准确的抽取触发词指代事件特征，本文额外设计了一个辅助简易分类器，仅依赖候选共指对的两个触发词进行类别判断，该部分的损失函数计算公式如下：

$$L_{aux} = \text{BCE}\left(\text{ReLU}\left(\frac{v_1 \cdot v_2}{||v_1|| * ||v_2||}\right), y\right) \tag{2}$$

其中，v_1和v_2分别代表两个对应触发词的特征向量，计算两者之间的余弦相似度，并利用ReLU用于将取值限制在[0, 1]之间，再利用二元交叉熵函数（Binary Cross Entropy, BCE）计算其与标签之间的损失。

本文判别器模型训练最终的损失函数由类别交叉熵损失函数L_{cls}和辅助交叉熵损失函数L_{aux}组成，即：

$$L_{total} = L_{cls} + L_{aux} \tag{3}$$

4 实验与分析

4.1 数据集与评价指标

本文实验使用事件核心标注数据集增强版（ECB+）(Cybulska et al., 2014)和枪支暴力语料库（GVC）(Vossen et al., 2018)两个常用数据集。其中，ECB+包含45个不同主题，主题1-35用于训练和验证，主题36-45用于测试，特点是主题间词汇区分度高。GVC围绕枪支暴力单一

主题展开，训练、验证和测试子主题分别为170个、37个和34个，特点是词汇相似性高，事件共指判断难度大。评价指标方面，本文采用标准指标：MUC、B^3、CEAFe、LEA以及CoNLL F1(Vilain et al., 1995; Bagga et al., 1998; Luo et al., 2005; Moosavi et al., 2019)，对每个指标，计算精确度(P)、召回率(R)和F1值，全面评估模型性能。

4.2 实验设置

本文所有实验均使用一张NVIDIA A800 (80GB显存)加速卡完成，无特别说明大模型选用Qwen2.5-72B-Instruct-AWQ[1]，大模型推理使用vllm[2]进行，推理温度系数设为0.5，触发词词目聚类使用的提示词如表1所示，触发词指代文本解释生成使用的提示词如表2所示。一阶段触发词词目启发、二阶段判别器模型超参配置均与(Ahmed et al., 2023)保持一致。

4.3 与基线方法的比较

本文主要是对(Ahmed et al., 2023)方法的改进，实验选取该方法作为基线方法，具体对比结果如表4所示。

			MUC			B^3			CEAFe			LEA			CoNLL
Method	Dataset	Encoder	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
Baseline	ECB+	RoBERTa-base	76.2	86.9	81.2	77.8	85.7	81.6	83.9	73.0	78.1	68.7	71.5	70.1	80.3
Ours	ECB+	RoBERTa-base	85.5	87.5	86.5	83.8	85.0	84.4	84.0	81.6	82.8	75.8	76.1	76.0	84.6
Baseline	ECB+	Longformer	80.0	87.3	83.5	79.6	85.4	82.4	83.1	75.5	79.1	70.5	73.3	71.9	81.7
Ours	ECB+	Longformer	84.7	85.3	85.0	83.8	81.4	82.6	80.9	80.3	80.6	74.9	71.9	73.4	82.7
Baseline	GVC	RoBERTa-base	87.0	89.6	88.3	82.3	67.9	74.4	62.0	55.2	58.4	77.6	57.8	66.2	73.7
Ours	GVC	RoBERTa-base	94.3	92.0	93.1	92.4	75.4	83.1	69.3	76.8	72.9	88.9	70.3	78.5	83.0
Baseline	GVC	Longformer	84.0	91.9	87.4	79.0	76.4	77.7	69.6	52.5	59.9	74.1	63.9	68.6	75.0
Ours	GVC	Longformer	91.7	91.2	91.4	88.0	78.4	82.9	70.6	72.1	71.3	83.4	71.5	77.0	81.9

表4.本文方法与基线方法在ECB+和GVC数据集上的结果对比

表4结果表明，本文方法相较基线方法在ECB+数据集上各项指标均得到提升，以二阶段选择RoBERTa-base编码器为例，MUC F1、B^3 F1、CEAFe F1、LEA F1和CoNLL F1分别绝对提升4.3、2.8、4.7、4.9和4.3；而在难度更大的GVC数据集上，本文方法的提升幅度更大，分别达到4.8、8.7、14.5、12.3和9.3，表明大模型增强对于困难样本的增益可能更大。二阶段使用Longformer作为编码器时的提升相对使用RoBERTa-base时较小，因为Longformer使用全文作为输入可能使得其他触发词信息干扰模型判别。

为验证本文方法的鲁棒性，在GVC数据集上使用相同的超参和模型配置进行了5次独立重复实验，五次实验CoNLL F1得分分别为82.6,82.7,82.9,83.0,83.2,平均误差为0.22。此外，为验证本文方法与基线方法的性能对比具有普适性，并非只在特定模型下有优势，大模型改为选取LLaMa3.3-70B [3]，在数据集ECB+和GVC上分别进行对比实验，CoNLL F1得分分别是84.1和82.6，较基线方法得分80.3和73.7，分别提升3.8和8.9，提升幅度与使用Qwen2.5-72B-Instruct-AWQ的表现类似。

Method	Complexity	LLM(s)	Stage-1(s)	Stage-2(s)	Total(s)
(Ahmed et al., 2023)	O(n)	-	2	69	71
(Min et al., 2024)	O(n)+O(n^2)	δ	71	720	791+δ
Ours	O(n)	161	2	84	247

表5.不同方法在ECB+测试集上的耗时统计

本文选择(Ahmed et al., 2023)作为基线方法的另一个原因在于，其是一种兼顾复杂度和性能的两阶段方法。本文在该方法的一阶段、二阶段分别引入大模型进行能力增强。

[1]https://huggingface.co/Qwen/Qwen2.5-72B-Instruct-AWQ
[2]https://github.com/vllm-project/vllm
[3]https://hf-mirror.com/meta-llama/Llama-3.3-70B-Instruct

具体而言，一阶段本文方法使用触发词词目启发高效过滤大量简单非共指对，该过程复杂度为O(n)，n为初始事件个数，后根据统计得到的潜在共指词目集合筛选候选共指事件对，送入二阶段判别器进行共指判断，该过程(Ahmed et al., 2023)通过实验指出时间复杂度近似为O(n)。本阶段的大模型同义触发词词目聚类的计算次数与事件的主题个数t一致（$t<<n$），因此本文方法一阶段的时间复杂度与基线方法一致。由于本文一阶段召回了更多的潜在共指事件对，本文方法二阶段耗时略高于基线方法，但由于本文方法是对事件触发词指代信息的解释增强，计算复杂度亦为O(n)，随着n 增大耗时线性扩张，本文方法可良好扩展至大规模语料。而(Min et al., 2024) 由于在语料蒸馏与推理阶段需对事件对做两两组合，其复杂度为$O(n)+O(n^2)$，当n 增大数量级时耗时显著增加。

本文方法与基线方法都使用RoBERTa-base作为判别器，在ECB+测试数据集上的耗时如表5所示。其中，LLM耗时包括一阶段同义词词目聚类和二阶段触发词文本解释的推理耗时，由于一阶段的聚类时间复杂度为O(t)，二阶段仅对过滤后的少量事件触发词进行增强，虽然单次大模型推理耗时较大，但整体次数有限，因此本文方法的LLM总耗时是可控的。

4.4 与其他方法的比较

表6给出了与近期事件共指消解代表性方法的性能比较。

			MUC			B^3			CEAFe			LEA			CoNLL
Method	Dataset	Encoder	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
LLaMA 2-7B	ECB+	–	84.2	76.3	80.1	82.7	73.2	77.7	67.5	77.2	72.0	–	–	–	76.6
GPT-3.5	ECB+	–	81.7	81	81.4	81.0	78.6	79.8	76.1	77.0	76.5	–	–	–	79.2
(Held et al., 2021)	ECB+	RoBERTa-base*2	87.0	88.1	87.5	85.6	87.7	86.6	80.3	85.8	82.9	74.9	73.2	74.0	85.7
(Nath et al., 2024)	ECB+	Longformer	84.1	92.0	87.9	82.4	91.7	86.8	88.9	80.5	84.5	–	–	–	86.4
(Min et al., 2024)	ECB+	RoBERTa-large*2	89.4	87.1	88.2	89.1	86.5	87.8	82.7	85.5	84.1	79.7	78.5	79.3	86.7
Ours	ECB+	RoBERTa-base	85.5	87.5	86.5	83.8	85.0	84.4	84.0	81.6	82.8	75.8	76.1	76.0	84.6
LLaMA 2-7B	GVC	–	93.9	84.3	88.8	89.5	38.1	53.4	28.9	54.9	37.9	–	–	–	60.0
GPT-3.5	GVC	–	88.6	81.9	85.1	82.6	35.4	49.6	27.1	41.1	32.7	–	–	–	55.8
(Nath et al., 2024)	GVC	Longformer	91.6	94.2	92.9	86.7	82.1	84.3	75.8	68.1	71.7	–	–	–	83.0
(Held et al., 2021)	GVC	RoBERTa-base*2	91.8	91.2	91.5	82.2	83.8	83.0	75.5	77.9	76.7	79.0	82.3	80.6	83.7
(Min et al., 2024)	GVC	RoBERTa-large*2	92.4	93.2	92.8	87.0	87.4	87.2	83.6	80.7	82.1	83.4	83.0	83.2	87.4
Ours	GVC	RoBERTa-base	94.3	92.0	93.1	92.4	75.4	83.1	69.3	76.8	72.9	88.9	70.3	78.5	83.0

表6.本文方法与其他方法在ECB+和GVC数据集上的结果对比

表6给出的结果表明:

1）如(Min et al., 2024)报道的LLaMA 2-7B(Touvron et al., 2023)和GPT-3.5[4]等大模型，在没有面向特定任务进行微调的情况下，其性能较专业小模型仍有一定差距，尤其在更具挑战性的GVC数据集上;

2）(Held et al., 2021)的方法由于在一阶段和二阶段分别构建了专业模型，并使用候选共指事件对两两比较的平方级复杂度（O(n^2)）方法过滤事件对，取得了较好的消解效果。本文方法则在保持近似O(n)时间复杂度的同时，取得了与该方法接近的消解效果（GVC上CoNLL F1仅相差0.7）;

3）(Min et al., 2024)的方法在(Held et al., 2021)的方法基础上，通过引入了更大的专业模型，并使用大模型进行语义增强，据本文所知是当前相关数据集上性能最好的方法。但其时间复杂度与(Held et al., 2021)一致为（O(n^2)），表5所示，该方法在ECB+上耗时显著大于本文方法和本文的基线方法，其中，δ是使用GPT4接口的耗时;

4）(Nath et al., 2024)的方法需要使用大模型构建蒸馏数据集，这过程时间复杂度是O(n^2)，后续需多阶段训练模型，训练成本较高。本文方法在GVC上性能和其相当，但是训练成本更低。

4.5 消融实验

4.5.1 本文方法消融

为验证本文方法的每个改进点对于最终事件共指消效果的贡献，在ECB+和GVC上进行消

[4]https://platform.openai.com/docs/model-index-for-researchers

融实验，结果如表7所示。

Method	Stage-1	Stage-2	Loss	CoNLL F1	
				GVC	ECB+
Baseline	LH	sentence	L_{cls}	73.7	80.3
Ours	LH	explanation	L_{cls}	76.9	82.4
Ours	LH	sentence + explanation	L_{cls}	80.3	82.8
Ours	LH	sentence + explanation	L_{cls}+L_{aux}	80.6	83.3
Ours	LH+LLM-SLC	sentence + explanation	L_{cls}+L_{aux}	83.0	84.6

表7.本文方法的消融实验

1）在一阶段，引入大模型触发词词目聚类（LLM-SLC）更为精准的召回候选共指对之后，模型在ECB+上提升1.3，而在GVC上则显著提升2.4，说明一阶段精准召回的重要性。

2）仅在二阶段将触发词词目句子替换为大模型提取的触发词指代解释（explanation），即可在ECB+和GVC上分别提升2.1和3.2；若将二阶段的触发词解释拼接原本的触发词词目句子，模型性能还会有轻微提升；对于二阶段模型训练，若引入面向触发词的辅助简易分类损失函数，模型在测试阶段的性能在两个数据集上继续绝对提升0.5和0.3。

3）整体而言，本文在一阶段在两个数据集上分别提升1.3和2.4,在二阶段分别提升3.0和6.9，一定程度上说明二阶段对于方法最终效果的影响更大。

4.5.2 大模型不同文本增强方式比较

本文方法最终选用基于大模型生成的触发词解释文本作为主要增强文本，除此方式之外，本文还对比了基于篇章级总结（Ours）和(Min et al., 2024)生成的事件总结性解释文本（截止本文撰写日期，只公开了ECB+部分）的效果，实验结果如表8所示。

Stage1	Stage2 Input	Loss	CoNLL F1	
			GVC	ECB+
LH	sentence	L_{cls}	73.7	80.3
LH	explanation (Ours)	L_{cls}	76.9	82.4
LH	sentence + explanation (Ours)	L_{cls}	80.3	82.8
LH	sentence + summary (Ours)	L_{cls}	80.4	82.2
LH	sentence + explanation(Min et al., 2024)	L_{cls}	–	82.3

表8.大模型不同文本增强对比实验

结果表明，篇章级总结文本增强和触发词指代解释增强均可提升专业模型对于事件共指的判别能力，而本文设计的触发词解释提示词生成的解释文本优于(Min et al., 2024)的方法生成解释文本的增强效果。

5 总结

本文提出了一种更高效的基于大模型增强的两阶段事件共指消解方法，一阶段引入大模型同义触发词词目聚类，有效缓解仅靠触发词词目启发先验统计导致的潜在共指对遗漏问题，二阶段引入大模型对高效判别模型进行触发词指代文本增强，并设计了一种简易辅助损失函数，显著增强了触发词指代事件的能力。实验表明，本文方法在保持与现有两阶段高效基线方法一致的时间复杂度的同时，在主流数据集上CoNLL F1等各项指标均取得了明显提升；较两阶段双模型方法，差距明显缩小，效率优势明显。

6 致谢

本文研究受项目(No.E4T8040301)的资助。

参考文献

Ahmed, S. R., Nath, A., Martin, J. H., and Krishnaswamy, N. 2023. $2 * n$ is better than n^2: Decomposing Event Coreference Resolution into Two Tractable Problems. *Findings of the Association for Computational Linguistics: ACL 2023*. 1569–1583

Amit Bagga, and Breck Baldwin. 1998. A modeltheoretic coreference scoring scheme. *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*. 563-566.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: Cross-document language modeling. *Findings of the Association for Computational Linguistics: EMNLP 2021*, :2648–2662.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Cross-document coreference resolution over predicted mentions. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 5100–5107.

Agata Cybulska, and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. *In Proceedings ofthe Ninth International Conference on Language Resources and Evaluation*. 4545—4552.

Breck Baldwin. 1997. CogNIAC: high precision coreference with limited knowledge and linguistic resources. *In Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

Chen, Liang-Ching, and Kuei-Hu Chang. 2023. An Extended AHP-Based Corpus Assessment Approach for Handling Keyword Ranking of NLP: An Example of COVID-19 Corpus Data. *Axioms*. 12(740).

Fang, J.; Li, P.; Zhu, Q. 2019. Employing Multi-attention Mechanism to Resolve Event Coreference. *Comput. Sci*, 46(x):277–281.

Fu, Yibin, Zhaoyun Ding, and Xiaojie Xu. 2024. LLM & Bagging for 1-shot Joint IE. *In 2024 IEEE 9th International Conference on Data Science in Cyberspace*. 204—208.

Huang K H, Hsu I, Natarajan P, Chang K W, and Peng N. 2022. Multilingual generative language models for zero-shot cross-lingual event argument extraction. *arXiv preprint arXiv:2203.08308*..

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. *In Proceedings ofthe 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 489—500

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational linguistics*. 39(4):885—916.

Zhigang Huan, Guoquan Jiang, Yujian Zhang, Liu Liu, and Kun Ding. 2023. 基于核心句的端到端事件共指消解. 计算机科学 50(11):185–191.

Liu, R, Mao, R, Luu, A.T, and Cambria, E. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*. 56(12):14439—14481.

Liu Liu, Guoquan Jiang, Zhigang Huan, Shanshan Liu, Ming Liu, and Kun Ding. 2024. 一种端到端的事件共指消解方法. 工程科学与技术 56(1):82–88.

Xiaoming Liu, Yanbo Wu, Yang Guan, Jie Liu, and Jiahao Wu. 2025. 主题结构增强的大模型实体共指消解方法. 计算机应用研究, 42(9). DOI: 10.19734/j.issn.1001-3695.2025.03.0044.

Letian Peng, Zilong Wang, Feng Yao, Zihan Wang, and Jingbo Shang. 2024. Metaie: Distilling a meta model from llm for all kinds of information extraction tasks. *arXiv preprint arXiv:2404.00457*.

Min, Q., Guo, Q., Hu, X., Huang, S., Zhang, Z., and Zhang, Y. 2024. Synergetic Event Understanding: A Collaborative Approach to Cross-Document Event Coreference Resolution with Large Language Models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 2985–3002

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A modeltheoretic coreference scoring scheme. *In Proceedings of the 6th Conference on Message Understanding, MUC6 '95*. 45-52.

Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube. 2019. Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4168–4178

Nath A, Manafi S, Chelle A, and Krishnaswamy, N. 2024. Okay, Let's Do This! Modeling Event Coreference with Generated Rationales and Knowledge Distillation. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3931—3946

Ovando-Becerril, Ericka, and Hiram Calvo. 2023. A Metaphorical Text Classifier to Compare the Use of RoBERTa-Large, RoBERTa-Base and BERT-Base Uncased. *In International Workshop on Artificial Intelligence and Pattern Recognition*. 248–259

Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers. 2018. Don't annotate, but validate: A data-to-text method for capturing event data. *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Touvron H, Martin L, Stone K, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *aarXiv preprint arXiv:2307.09288*.

Veselin Stoyanov, and Jason Eisner. 2012. Easy-first coreference resolution. *In Proceedings ofCOLING 2012*. 2519—2534.

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying discourse coherence theory to cross document coreference. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, :2648–2662.

Xu D, Chen W, Peng W, et al. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*. 18(6): 186357.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. *In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT'05*. 25-32.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! *In Findings of the Association for Computational Linguistics: EMNLP 2023*. 10572—10601.

Yang Y, Wu Z, Yang Y, Lian S, Guo F, and Wang Z. 2022. A Survey of Information Extraction Based on Deep Learning. *Applied Sciences*. 12(19):9691.

Zeng, Y., Jin, X., Guan, S., Guo, J., and Cheng, X. 2020. Event coreference resolution with their paraphrases and argument-aware embeddings. *Proceedings of the 28th International Conference on Computational Linguistics*, :3084–3094.

Zhang R, Li Y, Ma Y, Zhou M, and Zou L. 2023. LLMaAA: making large language models as active annotators. *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023*. 13088–13103.