软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(11):5158−5177 [doi: 10.13328/j.cnki.jos.007367] [CSTR: 32375.14.jos.007367]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

外部知识增强的事件共指消解方法*

徐 昇, 李培峰, 朱巧明

(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

通信作者: 李培峰, E-mail: pfli@suda.edu.cn

摘 要: 由于语言表述的多样性和复杂性, 事件共指关系有时是通过事件描述之间潜在的关联性体现的. 已有方法大多仅基于触发词、论元等事件内部信息采用语义相似度计算方法进行事件共指消解, 难以处理上述问题. 为此, 提出了一种外部知识增强的事件共指消解方法, 通过运用大语言模型生成共指相关外部知识 (主要包括篇章连贯性、逻辑关系知识和常识背景知识) 来发现共指事件之间的潜在联系. 首先, 运用超大语言模型 ChatGPT 构造包含外部知识的训练数据. 然后, 在数据上指令微调 FlanT5 等基础大语言模型, 使其学习到生成共指相关外部知识的能力. 最后, 运用微调后的大语言模型生成文档级事件摘要和思维链风格的共指推理路径, 结合事件内部信息和外部知识预测共指. 在 KBP 数据集上的实验结果显示, 该方法的性能优于当前先进的基线模型.

关键词: 事件共指消解; 外部知识; 指令微调; 思维链; 大语言模型

中图法分类号: TP18

中文引用格式: 徐昇, 李培峰, 朱巧明. 外部知识增强的事件共指消解方法. 软件学报, 2025, 36(11): 5158–5177. http://www.jos.org.cn/1000-9825/7367.htm

英文引用格式: Xu S, Li PF, Zhu QM. Event Coreference Resolution Method Enhanced by External Knowledge. Ruan Jian Xue Bao/Journal of Software, 2025, 36(11): 5158–5177 (in Chinese). http://www.jos.org.cn/1000-9825/7367.htm

Event Coreference Resolution Method Enhanced by External Knowledge

XU Sheng, LI Pei-Feng, ZHU Qiao-Ming

(School of Computer Science & Technology, Soochow University, Suzhou 215006, China)

Abstract: The diversity and complexity of linguistic expressions often lead to event coreference relations being reflected as latent correlations between event mentions. Existing methods predominantly rely on semantic similarity computations based on internal event features, such as triggers and arguments, which limits their ability to address such latent correlations effectively. To overcome this limitation, an external knowledge-enhanced event coreference resolution method is proposed. This approach leverages large language models (LLMs) to generate external knowledge related to coreference, encompassing discourse coherence, logical relationships, and common sense background knowledge. First, the ultra-large language model ChatGPT is utilized to construct training data enriched with external knowledge. Next, foundational LLMs like FlanT5 are fine-tuned on this data to acquire the ability to generate coreference-related external knowledge. Finally, the fine-tuned LLM generates document-level event summaries and chain-of-thought (CoT) style coreference reasoning paths. By integrating internal event features with external knowledge, the proposed method effectively identifies event coreference. Experimental results on the KBP dataset demonstrate that the proposed method outperforms previous state-of-the-art baselines.

Key words: event coreference resolution; external knowledge; instruction tuning; chain-of-thought (CoT); large language model (LLM)

信息抽取领域将事件定义为由某些参与者在某个时间某个地点发生的行为[1,2]. 文档中对事件的具体表述称为事件描述, 每个事件描述由最能代表事件发生的触发词以及参与者、时间、地点等一系列在事件中扮演特定角色的论元组成. 事件共指消解任务旨在判断单个或多个文档内的事件描述是否指向现实世界中的同一个事件, 从

* 基金项目: 国家自然科学基金 (62276177, 62376181)
收稿时间: 2024-07-24; 修改时间: 2024-09-08, 2024-10-29; 采用时间: 2024-11-13; jos 在线出版时间: 2025-08-20
CNKI 网络首发时间: 2025-08-20

而将共指的事件描述聚类成簇 (串联成链). 例 1 展示了一个包含 4 个事件描述 (*ev*1–*ev*4) 以及 9 个实体 (*en*1–*en*9) 的例子, 其中一些实体在事件描述中扮演攻击者、目标、时间等论元角色, 详细信息如图 1 所示.

*ev*1	
类型	攻击
触发词	砍击
攻击者	两名男子 (*en*1)
目标	一名英国士兵 (*en*2)
时间	上个月 (*en*3)

*ev*2	
类型	死亡
触发词	死
受害者	一名英国士兵 (*en*2)
时间	上个月 (*en*3)

*ev*3	
类型	攻击
触发词	谋杀
攻击者	这两名男子 (*en*5)
攻击者	阿德博拉 (*en*6)
攻击者	阿德博尔 (*en*7)

*ev*4	
类型	攻击
触发词	谋杀
攻击者	阿德博拉 (*en*8)
目标	两名警察 (*en*9)

图 1 事件描述具体论元角色的信息

例 1: 被控告在{上个月}$_{en3}${砍击}$_{ev1}${一名英国士兵}$_{en2}$ 致{死}$_{ev2}$ 的{两名男子}$_{en1}$ 出现在{分开的法庭}$_{en4}$ 接受听证. {这两名男子}$_{en5}$ 是 28 岁的{麦克·阿德博拉}$_{en6}$ 和 22 岁的{麦克·阿德博尔}$_{en7}$, 面临{谋杀}$_{ev3}$ 指控. {阿德博拉}$_{en8}$ 还被控犯有其他罪行, 包括企图{谋杀}$_{ev4}${两名警察}$_{en9}$.

在例 1 所示的例子中, 事件描述“砍击” (*ev*1) 和“谋杀” (*ev*3) 共指, 都指向“攻击英国士兵”事件, 在消解共指之后这两个事件描述会被放入到同一个簇中. 此时通过汇总它们的信息就能够描述出该事件的全貌, 即“28 岁的阿德博拉和 22 岁的阿德博尔在上个月谋杀了一名英国士兵”. 因此, 事件共指消解能够为依赖内容理解以及信息汇总的下游任务提供帮助, 包括文本摘要[3]、篇章分析[4]、事件抽取[5]等.

除了小部分研究[6,7]运用聚类方法直接生成事件共指簇, 大部分先前的工作将事件共指消解看作是一个事件对分类任务, 即预测文档中的任意两个事件描述是否共指. 因此, 如何分析事件描述之间的关联性是识别共指的关键. 早期的机器学习方法通过手工构建的触发词语义相似度、论元重合度等特征来匹配事件对[8,9], 后来的浅层神经网络方法则在运用卷积网络 (CNN)、循环网络 (RNN) 等编码器获得触发词嵌入的同时, 融入词汇、句法、篇章等语言学特征辅助预测[10,11]. 近年来, 随着 BERT 等大规模预训练语言模型的出现, 研究者一方面通过引入 SpanBERT、Longformer 等深度神经网络模型改进基于触发词的事件表示[12,13], 另一方面通过将论元信息融入事件编码[14]、设计专门的模型结构[15]等方式捕获论元、属性等其他事件元素层面的交互.

但是, 由于语言表述的多样性和复杂性, 事件描述存在表述多样性和信息缺失性. 例如事件通常只有在阐述不同方面或者补充额外信息时才会被重复提及, 因此即使共指的事件描述也可能因位于差异很大的上下文中而具有多样化的表述. 此外, 由于事件被再次提及时往往会省略部分甚至所有论元, 并且多个事件描述对于同一论元的表述也可能差异很大, 因此共指事件描述也可能具有不相似或者重合度很小的论元信息. 在图 1 展示的例子中, 事件描述“砍击” (*ev*1) 和“谋杀” (*ev*3) 共指, 但是“砍击”描述包含有较为完整的论元信息 (攻击者、目标、时间), 而“谋杀”描述侧重于阐述攻击者信息因而仅包含攻击者论元, 导致这两个共指事件描述的论元信息不匹配. 因此, 触发词、论元等事件内部信息在分析事件关联性方面存在不完备的问题, 一些语义不相似事件描述之间的共指是通过事件之间的潜在联系体现的. 然而, 已有工作大多仅基于触发词、论元等事件内部信息采用语义相似度计算方法分析事件关联性, 难以捕获到事件之间的潜在联系. 表 1 展示了几个仅通过事件内部信息很难识别共指的例子. 这些例子中的事件描述虽然共指, 却具有不相似的词汇和上下文, 因此仅通过分析触发词、论元等维度的关联性就可能会错误识别共指, 需要引入能够捕获事件潜在联系的其他信息. 考虑到研究[16–20]表明事件共指与篇章连贯性、逻辑关系知识、常识背景知识等外部知识存在联系, 本文尝试引入外部知识来发现事件之间潜在的关联性.

例如, 对于表 1 中的例子, 可以通过在事件内部信息之外引入如表 2 所示的外部知识来辅助识别共指. 例如对于第 2 个例子, 由于“使”描述 (*ev*1) 和“损伤”描述 (*ev*2) 分别侧重于阐述舞女受伤过程的细节和受伤之后的治疗过程, 因此它们在词汇和上下文层面的语义相似度都很低, 而通过引入逻辑关系知识, 模型就可以借助“袭击-受伤-治疗”这 3 个事件之间潜在的因果联系来发现这两个句子中存在共指事件的可能性.

与事件因果关系识别、时序关系识别等其他事件任务[21,22]相比, 事件共指消解任务的外部资源相对匮乏[23], 目前只有少数研究探索了引入外部知识的共指消解方法. 例如 Choubey 等人[16]设计约束项建模共指链之间的相关性、文档分布模式与共指之间的相关性等来改进性能. 但是该方法依赖于整数线性规划, 难以直接应用于当前

的深度学习方法. Ravi 等人[20]运用 GPT-3 模型为每一个事件描述推理可能发生在其之前和之后的其他事件, 然后将这些推理出的常识知识作为额外的上下文送入到共指打分器中. 但是该方法一方面依赖于标注数据, 另一方面仅将常识知识作为特征输入分类器, 无法激发出大语言模型在事件共指消解任务上的潜能. Nath 等人[24]首先运用大语言模型生成事件对共指的原因 (rationale), 然后将生成原因作为额外的标签训练小规模的编码模型以蒸馏出其中的上下文线索. 但是该方法仅将生成的推理线索用于提供远程监督信号, 同样没有充分利用大语言模型自身的推理能力识别事件共指.

表 1 仅通过事件内部信息很难识别共指的例子

缺少信息	内容	翻译
篇章连贯性	The president's {speech}$_{ev1}$ shocked the audience. … The {proposed}$_{ev2}$ policies will not surprise those who followed his campaign.	总统的{讲话}$_{ev1}$震惊了全场听众. … {提出}$_{ev2}$的政策不会让那些关注他竞选活动的人感到惊讶.
逻辑关系知识	Former Pakistani dancing girl commits suicide 12 years after horrific acid attack which {left}$_{ev1}$ her looking "not human". … She had undergone 39 separate surgeries to repair {damage}$_{ev2}$.	前巴基斯坦舞女在遭受可怕的{使}$_{ev1}$她变得看起来"不像人"的硫酸袭击12年后自杀. … 她接受了39次单独的手术来修复{损伤}$_{ev2}$.
常识背景知识	A barrage of US missiles {strikes}$_{ev1}$ Pakistan's North Waziristan tribal district on Tuesday. … They said further {strikes}$_{ev2}$ carried out by unmanned aircraft targeted the Nawab village on the border with Afghanistan.	美国周二对巴基斯坦北瓦济里斯坦部落地区发动一连串导弹{袭击}$_{ev1}$. … 他们表示, 无人机对阿富汗边境的纳瓦布村进行了进一步的{空袭}$_{ev2}$.

表 2 引入外部知识辅助识别事件共指的例子

事件内部信息	外部知识
*ev*1: 讲话、总统、听众 *ev*2: 提出、政策、竞选活动	[篇章连贯性] 这两个事件都围绕着总统在竞选活动上提出了一系列政策展开, 具有清晰的叙事连续性.
*ev*1: 使、舞女、硫酸袭击、自杀 *ev*2: 损伤、手术、修复	[逻辑关系知识] 袭击导致了受伤, 受伤又引发了手术治疗. 这种因果联系增加了这两个事件的关联性.
*ev*1: 袭击、美国、周二、巴基斯坦... *ev*2: 空袭、无人机、阿富汗...	[常识背景知识] 无人机是可以发射导弹的攻击装置, 攻击发生在巴基斯坦和阿富汗接壤的边境地区.

为此, 本文探索了事件共指相关外部知识的获取方法以及适合于共指消解任务的外部知识融入方法, 提出了一种外部知识增强的事件共指消解方法 CorefCoT, 通过在常规的触发词、论元等事件内部信息的基础上, 运用大语言模型生成共指相关外部知识来发现事件之间的潜在联系. 本文的主要贡献如下.

(1) 本文通过指令微调大语言模型使其学习到为事件对生成共指相关外部知识的能力, 从而使得模型可以通过引入外部知识来发现事件之间的潜在联系.

(2) 本文将共指识别分解为文档级事件摘要和共指推理路径生成两个步骤, 使得模型可以基于文档级信息以及思维链风格的推理路径识别共指, 充分激发出大语言模型的共指消解潜能.

1 相关工作

1.1 事件共指消解

先前的研究大多将事件共指消解视为一个有监督的事件对分类或者事件簇聚类问题. 早期的机器学习方法首先构建语言学特征, 然后运用机器学习模型识别共指. 例如 Chen 等人[8]在触发词和论元特征以外, 还构建了极性、模态、泛型等属性特征, 然后运用多个最大熵分类器识别共指. 但是这些方法逐项匹配各个事件元素, 一方面容易受到句法分析等上游环节中噪声的干扰, 另一方面也难以处理空事件元素槽问题. 随着深度学习的兴起, 后来的研究提出了基于神经网络的方法, 首先运用编码器围绕触发词获取事件表示, 然后送入到分类器中完成预测, 通常还

会结合手工构建的语言学特征来辅助判断. 例如 Krause 等人[10]首先运用 CNN 模型结合最大池化操作获得触发词表示, 然后将其与手工构建的类型匹配性、论元重合度等特征一起送入到全连接层中识别共指. Lai 等人[25]提出了一种上下文感知的门控机制, 引导模型在触发词之外有选择地融入类型、极性、模态等属性特征. 近年来, 随着 SpanBERT、Longformer 等深度神经网络模型被引入以获得更好的事件表示, 研究者大多专注于设计匹配结构或增强事件表示来改进性能, 例如将论元信息融入事件编码[14]、学习任务相关表示[6,13,26]、融入簇信息[7,27]等. 例如 Huang 等人[15]交替使用论元匹配度打分任务和事件共指打分任务来训练模型, 使得模型学习到论元匹配度与共指之间的联系. 本课题组 Xu 等人[13]使用全局编码器、局部编码器和事件主题模型分别获得文档级、句子级和主题级事件表示, 然后结合多层次表示识别共指. 此外, Kriman 等人[6]通过构建吸引和排斥损失使得共指事件的表示更相似、非共指事件的表示更不同. Phung 等人[7]提出了迭代聚类算法, 在每一轮迭代开始时使用当前的簇信息来改进事件表示. 与机器学习方法相比, 神经网络方法不仅获得了更好的触发词表示, 并且还在特征空间中结合多种事件信息分析了整体关联性, 从而可以缓解特征工程中的噪声以及空元素槽问题. 随着大语言模型的流行, 最近工作还尝试引入大语言模型技术来增强数据集或直接改进事件共指消解. 例如本课题组 Xu 等人[28]提出了专用于事件共指消解的提示方法, 通过分析事件类型和论元层面的整体关联性识别共指. Ding 等人[29]提出了一种基于大语言模型的数据增强方法, 通过修改触发词和上下文为事件对生成反事实 (counterfactual) 数据, 使得模型可以学习到事件对特征以及上下文语义和共指之间的因果联系. Min 等人[30]首先运用大语言模型为每一个事件描述生成对应的摘要, 然后将其作为额外的特征训练小模型识别共指. 得益于大语言模型优异的上下文理解以及文本生成能力, 这些工作能够用很小的成本就获取到高质量的共指相关标注以增强数据集. 本文的方法同样采用了该思路, 借助大语言模型来为事件对提供高质量的共指相关外部知识. 但是与这些方法依然使用小模型识别共指不同, 本文尝试将超大语言模型 ChatGPT 的共指推理能力蒸馏到 FlanT5 等基础大语言模型上, 并且借助生成思维链风格的推理路径以充分运用大语言模型的推理能力识别共指.

为了避免触发词识别等上游组件带来级联错误, 还有部分工作提出了联合模型, 例如在多个独立任务模型的输出上进行联合推理[31,32]或同步完成多个相关任务进行联合学习[33,34]. 例如 Lu 等人[12]引入实体共指消解领域的端到端联合模型[35], 同步完成实体抽取、实体共指、触发词抽取和事件共指消解, 并且通过构建软约束捕获跨任务依赖. 后来 Lu 等人[36,37]引入更多的相关任务来改进模型, 例如属性抽取、照应识别、事实性预测等. 得益于同时学习多个相关任务, 联合模型可以获取到更多的知识完成事件共指消解. 但是联合模型在运行过程中需要维持庞大的候选触发词集合, 具有较高的空间复杂度, 而且由于没有预先抽取出事件, 联合模型每次识别的是两个候选触发词片段之间的共指, 无法从多个层面分析事件关联性.

尽管这些工作改进了事件共指消解的性能, 但是它们都仅仅基于触发词、论元等事件内部信息分析事件关联性, 无法捕获到事件之间的潜在联系, 因此难以处理表 1 中语义相似度与共指不一致的情况. 考虑到研究[16−20]表明事件共指与常识背景知识等外部知识存在联系, 也有少数工作探索了引入外部知识的共指消解方法. 例如 Choubey 等人[16]通过设计约束项引入篇章连贯性知识来建模共指链之间的相关性以及建模文档分布模式与共指之间的相关性. 但是该方法依赖于手工设计的整数线性规划约束项, 难以直接应用于当前的深度学习方法. Ravi 等人[20]首先标注包含事件时序常识的小规模知识库, 然后在该数据上微调 GPT-3 模型使其学习到推理发生在事件前后其他事件的能力, 最后将生成的前后事件作为外部知识送入打分器辅助识别共指. 但是该方法一方面依赖于标注数据, 仅能够提供知识类型固定的时序常识; 另一方面仅将常识知识作为额外输入的特征, 依然使用传统事件对打分器识别共指, 无法激发出大语言模型在事件共指消解上的潜能. Nath 等人[24]首先运用大语言模型生成事件对共指的原因 (rationale) 作为额外标签以增强已有的共指语料库, 然后将生成的原因作为远程监督信号来训练小规模的编码模型以蒸馏出其中的上下文线索. 但是该方法仅将生成的推理线索用于提供远程监督信号, 同样没有充分利用大语言模型自身的推理能力识别共指. 本文提出的方法通过将共指识别分解为文档级事件摘要和共指推理路径生成两个步骤, 一方面使得模型可以使用文档级事件信息来识别共指, 另一方面使得模型可以通过自然语言的形式更好地融入外部知识以及激发出模型的共指推理潜能.

1.2 基于大语言模型技术的信息抽取

近年来, 随着越来越多的研究者尝试运用大语言模型来辅助完成任务, 部分信息抽取工作也对引入大语言模型进行了探索. 目前常用的大语言模型技术主要包括提示学习和指令微调.

提示学习技术通过精心设计的提示 (prompt) 将任务转换为与模型预训练任务类似的形式进行处理, 按照模板形式以及执行流程大致可以分为两种: (1) 基于填充提示 (cloze prompt) 的分类模型, 其首先构造包含特殊遮盖标记的文本模板将任务转换为一个完形填空问题, 然后运用预训练模型的遮盖语言建模头 (masked language modeling head) 来预测填充这些槽的标记; (2) 基于前缀提示 (prefix prompt) 的生成模型, 其首先构建包含所有引导信息的前缀提示, 然后将其送入到生成模型中继续生成后续的文本, 将任务转换为一个文本补全问题. 例如大语言模型中常用的上下文学习 (in-context learning) 技术即通过构建包含任务描述和示例的提示来引导模型生成合理的输出. 目前, 部分信息抽取研究也尝试运用提示技术来处理包括命名实体识别[38]、实体关系抽取[39−41]、事件抽取[42−44]、事件因果关系识别[45]等在内的各种任务. 例如 Lu 等人[46]提出了统一文本到序列生成框架, 首先通过结构抽取语言将实体、关系、事件等各种信息抽取任务转换为统一的文本表示, 然后构造基于模式的提示机制来引导模型学习跨任务的知识. Dai 等人[44]提出了一种迭代式的提示方法, 每轮迭代时首先拼接已预测出的事件论元来捕获论元之间以及论元与上下文之间的交互, 然后通过预测遮盖词来判断当前实体的角色. Ma 等人[43]通过人工设计以及自动构建的提示将事件论元抽取转换为槽填充问题, 然后基于预测出的角色相关查询来引导模型生成每一个槽对应的论元文本. 但是, 这些工作大多基于人工预设的模式 (schema) 运用模型直接输出结果, 只能处理抽取目标相对直接的简单任务, 难以充分激发出模型的推理能力.

因此, 为了提高模型解决复杂任务的性能, 研究者提出了思维链 (chain-of-thought, CoT) 提示学习技术[47], 通过将任务分解为多个子任务依次求解来充分激发出大语言模型的推理能力. 与上下文学习专注于构造前缀提示不同, 思维链会在提示中展现出一系列中间推理过程, 因此可以被视为是一种特殊的上下文学习. 近年来一些信息抽取领域的研究也探索了基于思维链提示技术的方法. 例如, Nguyen 等人[48]发现自然语言理解任务具有不同的粒度并且粗粒度的任务通常难度更低, 提出了从粗到细的思维链 (coarse-to-fine CoT, CoF-CoT) 提示方法来完成逻辑形式抽取等自然语言理解任务, 将前序粗粒度任务的输出作为提示的一部分来辅助完成后续的细粒度任务. Li 等人[49]提出了将实体关系抽取分解为 3 个步骤来完成的 SUMASK 方法, 首先生成包含头实体和尾实体语义关系的摘要, 然后生成每种关系对应的问题将关系标签转换为语言描述, 最后要求模型基于第 1 个步骤生成的摘要来回答第 2 个步骤中的问题以捕获实体和关系之间的语义联系. 考虑到事件共指消解也是一个依赖推理的复杂任务, 本文引导大语言模型生成思维链风格的共指推理路径来结合事件内部信息和外部知识作出预测.

尽管大语言模型展现出了优秀的上下文理解和推理能力, 但是这些模型的训练目标是解决通用领域的各种问题, 因此将它们直接应用于特定任务可能表现不佳. 先前的研究[50]表明, 直接运用大语言模型完成信息抽取等非生成式任务的性能甚至可能低于微调小模型. 因此, 研究者提出了指令微调 (instruction tuning) 技术[51], 将大语言模型的能力调整至特定目标, 其核心思想是基于领域相关的指令数据使用 Seq2Seq 损失按照监督学习的方式来微调大语言模型, 使得模型与领域更适配. Wang 等人[52]提出的统一信息抽取框架 InstructUIE 是将指令微调技术应用于信息抽取领域的代表性工作, 其通过设计描述性指令将包括命名实体识别、实体关系抽取、事件抽取在内的多种信息抽取任务都转换为自然语言生成问题, 然后微调 11B 规模的 FlanT5 模型完成任务. 近年来, 使用自指令 (self-instruct) 算法进行指令微调[53]已经成为一种经济且有效的方法来引导大语言模型拟合人类偏好, 这些方法首先运用超大语言模型 (例如 ChatGPT) 生成指令跟随数据, 然后使用这些数据指令微调基础大语言模型, 从而在不依赖于标注数据的情况下将大语言模型调整为符合人类偏好. 受这种做法的启发, 本文也首先运用超大语言模型 ChatGPT 生成包含共指相关外部知识的指令数据, 然后在该数据上指令微调 FlanT5 等基础大语言模型.

2 外部知识增强的事件共指消解方法

针对事件内部信息分析事件关联性不完备的问题, 本文提出了外部知识增强的事件共指消解方法 CorefCoT, 通过运用大语言模型生成共指相关外部知识来发现事件之间的潜在联系, 其工作流程如图 2 所示.

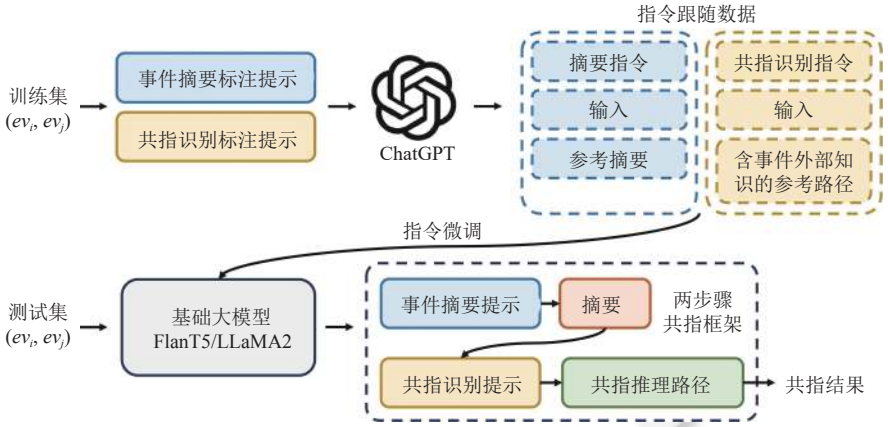

图 2 外部知识增强的事件共指消解方法 CorefCoT

考虑到人类在处理复杂的共指问题时也会将其分解为多个中间步骤 (inner monologue) 来解决, 从而在这些步骤中细致地推理两个事件描述之间在上下文、参与者、动作、地点等多个层面的联系[54,55], 而最近的研究[56]显示生成式大语言模型可以通过思维链提示技术模仿人类的这种逐步推理过程, 并且已经在部分共指任务上[20,57]证明了大语言模型在事件关系方面的溯因推理能力, 因此本文尝试借助大语言模型来辅助预测共指. 本文旨在通过指令微调技术将超大语言模型 ChatGPT 的共指推理能力蒸馏到 FlanT5 等基础大语言模型上, 以充分激发出模型的共指推理潜能. 具体来说, 本文首先设计用于事件摘要和共指识别任务的标注提示, 运用 ChatGPT 模型构建包含事件摘要以及共指相关外部知识的指令跟随数据. 然后在该数据上指令微调 FlanT5 等基础大语言模型, 使其学习到为事件生成文档级摘要以及为事件对生成共指相关外部知识的能力. 最后运用微调好的大语言模型执行本文提出的两步骤共指框架识别共指. 形式化地, 在训练阶段: 对于训练集中的每一对触发词 (ev_i, ev_j), 首先设计标注提示引导 ChatGPT 模型为每一个触发词生成摘要以及解释共指 (非共指) 的原因, 然后分别筛选出包含论元以及包含外部知识的样本构建指令数据集, 最后运用该数据集指令微调 FlanT5 等基础大语言模型. 在预测阶段: 将测试集中的触发词对 (ev_i, ev_j) 送入微调后的大语言模型, 依次执行事件摘要和共指推理路径生成任务, 通过生成思维链风格的推理路径来结合事件内部信息和外部知识预测共指, 最后从生成的推理路径中解析出共指预测结果.

由于文本生成过程具有较高的时间复杂度. 受过滤重排序框架[58]使用大语言模型对微调小模型置信分数较低的预测结果进行调整的启发, 本文也仅对当前先进的基线模型 CorefPrompt[28]的预测结果中那些低置信度的样本进行重新预测. 对 CorefPrompt 模型预测结果的分析表明, 模型对大部分预测结果都具有较高的置信度, 其中, 95% 结果的预测概率高于 0.98, 而低置信度结果仅占 5%. 此外, 高置信度结果的共指识别性能也优于低置信度结果, 因此影响微调小模型性能的主要就是低置信度样本. 具体的事件对共指识别性能如表 3 所示.

表 3 CorefPrompt 模型不同置信度预测结果的事件对共指识别性能 (%)

样本	占比	共指	非共指	Macro-$F1$
高置信度	95	48.3/81.7/60.7	99.4/97.1/98.2	79.4
低置信度	5	29.3/48.9/36.6	83.7/69.0/75.7	56.1

为了进一步分析低置信度样本判断错误的原因, 本文引入本课题组 Xu 等人[28]提出的事件语义相似度打分器根据相似度与共指是否一致对低置信度样本进行划分. 结果显示不一致样本的占比达 37.4%, 并且其事件对共指识别性能 (准确率: 13.8%/召回率: 18.4%/$F1$: 15.8%) 明显低于一致样本 (准确率: 36.4%/召回率: 69.2%/$F1$: 47.7%). 这验证了仅通过触发词、论元等事件内部信息难以处理语义相似度与共指不一致的情况.

2.1 两步骤共指框架

如图 2 所示, 两步骤共指框架是本文 CorefCoT 方法的核心, 模型的训练和预测阶段都围绕着该框架展开. 两

步骤共指框架将事件共指识别分解为文档级事件摘要和共指推理路径生成两个步骤, 一方面使得模型可以基于文档级事件信息识别共指, 另一方面使得模型可以借助思维链风格的推理路径自然地融入外部知识, 激发出大语言模型在事件共指消解任务上的潜能. 两步骤共指框架的具体结构如图 3 所示.

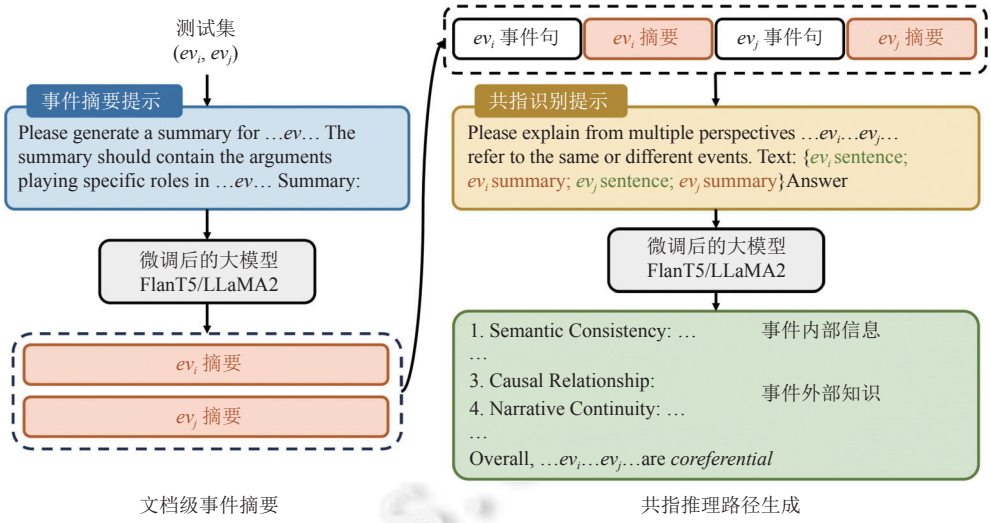

图 3 两步骤事件共指识别框架

具体来说, 对于文档 D 中的事件触发词对 (ev_i, ev_j), 两步骤共指框架首先构建事件摘要提示引导模型从文档 D 中选取事件相关信息为触发词 ev_i 和 ev_j 分别生成对应的文档级事件摘要. 然后, 将事件摘要与触发词所在的事件句相拼接作为包含文档级信息的增强事件句 es_i 和 es_j. 最后, 将增强事件句两两组合成对 (es_i, es_j) 再次送入模型, 通过构建共指识别提示引导模型生成思维链风格的推理路径, 结合事件内部信息和外部知识预测共指. 由于在训练阶段已经使用 ChatGPT 模型生成的事件摘要和共指识别数据对基础大语言模型进行了微调, 因此在预测阶段只需为每一对事件构建指令提示就可以运用微调后的大语言模型执行两步骤共指框架识别共指.

2.2 事件摘要和外部知识生成

考虑到近年来提出的大语言模型一方面通过在海量数据上预训练学习到了丰富的知识, 可以作为一种有效的事件外部知识库; 另一方面凭借强大的上下文理解以及语言生成能力, 可以遵循人类指令来完成各种复杂任务. 本文通过在 ChatGPT 模型构造的训练数据上指令微调基础大语言模型来为事件对生成共指相关外部知识.

事件摘要步骤旨在汇总文档中的事件相关信息. 考虑到 KBP 数据集标注有事件论元和实体共指信息, 本文提出了一种启发式方法来构建事件摘要标注提示 P'_{sum}, 引导 ChatGPT 模型选取与目标触发词存在关联的所有触发词和实体信息生成参考事件摘要. 假设截取的原文片段为 seg, 首先选出片段 seg 中与目标触发词 ev 共指的所有触发词组成集合 $C_e = \{t_1, \ldots, t_{|C_e|}\}$, $ev \in C_e$, 其中 $|C_e|$ 表示 C_e 集合的大小. 然后选出 C_e 中触发词对应的位于片段 seg 中的论元 (还包括与论元共指的实体) 组成集合 $A_e = \{a_1, \ldots, a_{|A_e|}\}$, 其中 $|A_e|$ 表示 A_e 集合的大小. 这样就得到了片段 seg 中与目标触发词 ev 相关的触发词集合 C_e 和实体集合 A_e. 特别地, 如果选出的多个实体共指, 只保留文本长度最长的实体以缩短提示长度. 为了进一步约束 ChatGPT 生成参考摘要的过程, 本文先从片段 seg 中挑选出包含触发词集合 C_e 和实体集合 A_e 中对象的句子组成新片段 seg', 然后使用该片段作为新的上下文来构建标注提示 P'_{sum}.

P'_{sum} : In the following text, the events [EVENT] t_1 [EVENT], …, [EVENT] $t_{|C_e|}$ [EVENT] represent the same event, please generate a summary of these events. The summary should contain the argument information of the events [EVENT] t_1 [EVENT], …, [EVENT] $t_{|C_e|}$ [EVENT], such as $\{a_1, \ldots, a_{|A_e|}\}$. Text: {$seg'$}

其中, [EVENT] 是特殊的事件标记, 它们也会被插入到新片段 seg'中以标记触发词的位置. 由于提示 P'_{sum} 显式地包含了与目标触发词 ev 相关的事件信息并且拼接了缩小范围的上下文 seg', ChatGPT 模型可以生成更遵循样本内

容的参考摘要. 最后, 过滤掉没有包含目标触发词 ev 及其论元的摘要以筛选出合格的训练样本.

共指推理路径步骤旨在生成包含外部知识的思维链风格路径识别共指. 为了使模型的训练和预测过程更接近, 本文使用前一个步骤中由 ChatGPT 模型生成的参考摘要来构建共指识别标注提示 P'_{Coref}. 具体来说, 对于目标触发词对 (ev_i, ev_j), 首先将上一个步骤中生成的参考摘要与原始事件句相拼接作为增强事件句 es_i, es_j, 然后将增强事件句作为上下文来构建共指识别标注提示 P'_{Coref}.

P'_{Coref} : In the following text, the two event trigger words [E1S] ev_i [E1E] and [E2S] ev_j [E2E] are {coreferential/non-coreferential}. Please concisely explain from multiple perspectives why they refer to {the same event/different events}. Text: {es_i; es_j}

由于提示 P'_{Coref} 显式地要求从多个角度解释事件共指 (非共指) 的原因, 这会鼓励 ChatGPT 模型生成包含更多事件信息维度的关联性描述. 最后, 过滤掉没有包含外部知识的描述以筛选出合格的训练样本. 此外, 由于本文在提示中使用标注的共指标签来引导 ChatGPT 模型生成共指相关外部知识, 这可以在事件描述对和潜在的共指原因之间建立一对一的映射, 从而更好地符合共指注释依赖于结构化知识库 (structured knowledge base) 的假设[59,60], 同时还降低了生成相关额外假设的成本 (包括计算成本和其他成本).

在运用 ChatGPT 模型获取到参考事件摘要以及参考事件关联性描述之后, 本文遵循指令微调技术的标准做法, 将生成的数据与对应的指令、输入和输出组合成 < instruction, input, output > 形式的 3 元组以构建指令跟随数据集. 得益于 ChatGPT 模型强大的文本理解和生成能力, 本文可以在几乎不依赖于人类标注的情况下获得大量高质量的训练数据.

接着, 本文在指令数据集上微调 FlanT5 等基础大语言模型, 使其学习到为事件对生成事件摘要以及共指相关外部知识的能力. 假设预训练语言模型的词表为 V. 对于编码器-解码器 (encoder-decoder) 结构的模型, 本文将包含指令和输入的提示 $x \in V^{T'}$ 作为模型的源序列, 将对应的输出 $y \in V^{T}$ 作为目标序列, 然后运用标准的序列到序列 (sequence-to-sequence, Seq2Seq) 损失来微调模型, 即最大化公式 (1) 所示的条件概率:

$$p(y|x) = \prod_{t=1}^{T} p(y_t|x, y_{<t}) \tag{1}$$

对于纯解码器 (decoder) 结构的模型, 本文直接拼接指令、输入和输出作为完整的目标序列 $y \in V^{T'+T}$, 然后运用标准单向语言模型损失来微调模型, 即最大化公式 (2) 所示的条件概率:

$$p(y) = \prod_{t=1}^{T'+T} p(y_t|y_{<t}) \tag{2}$$

考虑到纯解码器模型仅需要学习到生成输出的能力, 对于每一个样本, 本文还构建了长度为提示 (指令和输入) 标记序列长度的遮盖 (mask) 作为输入, 以便排除计算对提示标记预测结果的交叉熵损失.

通过使用共指相关推理数据 (包括与参与者、时间、地点等论元相关的事件摘要以及共指推理路径) 对 FlanT5 等开放权重大语言模型进行指令微调, 本文可以将模型的推理逻辑 (输出) 建立在事件描述对及其相关的文档级上下文信息上, 从而确保了互斥性 (mutual exclusivity) 并将大语言模型的推理能力适配到事件共指领域.

2.3 外部知识增强的事件共指识别

在完成指令微调后, 就可以运用微调后的大语言模型执行本文提出的两步骤共指框架, 通过依次生成事件摘要和共指推理路径识别共指. 具体来说, 对于事件对 (ev_i, ev_j), 首先为每一个触发词 ev 构建事件摘要提示 P_{sum}, 引导大语言模型生成包含论元信息的文档级摘要:

P_{sum}: Please generate a summary for the [ES] ev [EE] event. The summary should contain the arguments playing specific roles in this [ES] ev [EE] event, such as participants, time, locations, etc. Text: {*segment*} Summary:

其中, {*segment*}表示从原文中截取的片段, [ES] 和 [EE] 是标记触发词位置的特殊标记, 它们也会被插入到原文片段中以标记触发词位置. 然后, 将生成的摘要与触发词所在的原始事件句相拼接作为增强事件句 es. 最后, 将增强事件句组合成对 (es_i, es_j) 作为上下文来构建共指识别提示 P_{Coref}, 引导大语言模型生成思维链风格的共指推理

路径:

P_{Coref}: Please explain from multiple perspectives why the two event trigger words [E1S] ev_i [E1E] and [E2S] ev_j [E2E] in the following text refer to the same or different events. Text: $\{es_i; es_j\}$ Answer:

其中, [E1S]、[E1E] 等是特殊的事件标记, 它们也会被插入到增强事件句中以标记触发词的位置. 由于共指识别提示 P_{Coref} 采用了与标注提示 P'_{Coref} 相似的形式, 因此微调后的大语言模型会生成包含外部知识的推理路径.

与直接将共指识别转换为生成问题或者遮盖标记预测问题相比, 本文方法一方面通过将共指识别分解为事件摘要和共指推理路径生成两个步骤降低了完成任务的难度, 另一方面通过生成思维链风格的推理路径结合事件内部信息和外部知识预测共指, 能够更好地激发出大语言模型在事件共指消解方面的推理能力.

3 实验结果及分析

3.1 数据集及实验设置

为了与已有研究[13,15,28,36]进行比较, 本文同样在 KBP2015 和 KBP2016[2,61]数据集上训练模型, 然后在 KBP2017 数据集[62]上评估模型. KBP2015 和 KBP2016 数据集共包含 817 篇文档, 标注有分布在 14 794 个共指簇中的 22 894 个事件描述. KBP2017 则包含 167 篇文档, 标注有分布在 2 963 个共指簇中的 4 375 个事件描述. 与先前的工作[37]保持一致, 本文使用 KBP2015 和 KBP2016 数据集中的 735 篇文档进行训练, 使用剩余的 82 篇文档调整参数. 为了降低训练开销, 本文采用参数 k 设置为 1 的 CorefNM 策略[27]对训练集进行欠采样, 再通过随机采样使得正负样本比例平衡, 最终只使用了大约 1% 的训练数据. 此外, 为了测试模型的跨领域适用性, 本文还将训练好的模型迁移用于识别 ACE2005[1]数据集中的事件共指. ACE2005 数据集共包含 599 篇文档, 采用比 KBP 数据集更严格的共指定义. 跟随先前的工作[25,63], 本文同样在由 40 篇新闻文档构成的测试集上评估模型性能.

在实验细节方面, 本文使用 XL 版本的 FlanT5 模型和 7B 版本的 LLaMA2 模型作为骨架. 对于所有样本, 本文都围绕着两个事件触发词截取上下文. 对于 FlanT5 模型使得提示长度不超过 512, 对于 LLaMA2 模型使得提示和输出的长度之和不超过 1 024. 为了降低训练模型的计算资源消耗, 本文使用低秩调整技术 (low-rank adaptation, LoRA)[64]减少训练参数量, 并且运用模型量化技术 (quantization technique)[65]降低加载模型的显存占用. 对于 FlanT5 模型, LoRA 注意力维度设为 16, LoRA 缩放 Alpha 参数设为 32, 并且采用 8 位整数量化. 对于 LLaMA2 模型, 遵循 QLoRA[66]中的设置将 LoRA 注意力维度和 LoRA 缩放 Alpha 参数分别设为 64 和 16, 并且采用 4 位整数量化. 最终本文所有的实验都在单块消费级显卡 RTX3090 上进行. 批大小对 FlanT5 模型设为 16, 对 LLaMA2 模型设为 8 (带有梯度累积, 步数为 2), 训练轮数都设置为 5. 使用 Adam 优化器来更新模型参数, FlanT5 模型对应的学习率设为 1E–3, LLaMA2 模型的学习率设为 2E–4. 特别地, 对于 LLaMA2 模型, 本文使用线性学习率常数调度器, 预热 (warmup) 比例设置为 0.03. 所有组件中使用的随机种子都设置为 42. 由于大语言模型具有海量参数, 即使使用了参数高效微调方法 LoRA 和模型量化技术依然具有较高的训练开销, FlanT5 模型训练一轮 (epoch) 的时间开销大约在 40 min 左右, LLaMA2 模型大约在 5 h 左右.

与已有工作保持一致, 本文采用 MUC[67]、B^3[68]、$CEAF_e$[69]和 BLANC[70]这 4 个指标及其 $F1$ 分数的平均值 (AVG) 来评估事件共指消解性能, 采用准确率 (precision, P)、召回率 (recall, R) 和 $F1$ 分数来评估事件对共指识别性能. 其中, MUC 指标通过计算共指链的准确率、召回率和 $F1$ 分数来评估性能, 如公式 (3)–公式 (5) 所示.

$$P_{\text{MUC}} = \frac{\text{正确消解的共指链数量}}{\text{模型消解的共指链数量}} \times 100\% \tag{3}$$

$$R_{\text{MUC}} = \frac{\text{正确消解的共指链数量}}{\text{实际消解的共指链数量}} \times 100\% \tag{4}$$

$$F1_{\text{MUC}} = \frac{2 \times P_{\text{MUC}} \times R_{\text{MUC}}}{P_{\text{MUC}} + R_{\text{MUC}}} \times 100\% \tag{5}$$

而 B^3 指标则基于描述来计算得分, 如公式 (6)–公式 (8) 所示. 其中, S_{m_i} 和 G_{m_i} 分别表示模型输出结果中和实际标注中包含描述 m_i 的共指链; 权重 w_i 一般取 $1/N$, N 表示文档中描述的数量.

$$P_{m_i}=\frac{\left|S_{m_i}\cap G_{m_i}\right|}{S_{m_i}},\ R_{m_i}=\frac{\left|S_{m_i}\cap G_{m_i}\right|}{G_{m_i}} \tag{6}$$

$$P_{\mathrm{B}^3}=\sum_{i=1}^{N}w_iP_{m_i},\ R_{\mathrm{B}^3}=\sum_{i=1}^{N}w_iR_{m_i} \tag{7}$$

$$F1_{\mathrm{B}^3}=\frac{2\times P_{\mathrm{B}^3}\times R_{\mathrm{B}^3}}{P_{\mathrm{B}^3}+R_{\mathrm{B}^3}} \tag{8}$$

CEAF 指标会在模型输出的共指链与标注共指链之间建立映射, 通过计算共指链之间的相似度选择最匹配的映射来计算准确率、召回率和 $F1$ 分数. 常见的两种共指链相似度计算方法如公式 (9) 和公式 (10) 所示, 其中, G_i 表示标注的实际共指链, S_j 表示模型输出的预测共指链.

$$\phi_1(G_i,S_j)=\left|G_i\cap S_j\right| \tag{9}$$

$$\phi_2(G_i,S_j)=\frac{2\times\left|G_i\cap S_j\right|}{|G_i|+\left|S_j\right|} \tag{10}$$

CEAF 的准确率和召回率基于共指链对排序最优的相似度 $\Phi(g^*)$ 进行计算, 计算方式如公式 (11) 和公式 (12) 所示, 其中, 公式 (9) 和公式 (10) 两种相似度对应的指标分别称为基于描述的 $\mathrm{CEAF_m}$ 和基于共指链的 $\mathrm{CEAF_e}$.

$$P_{\mathrm{CEAF}}=\frac{\Phi(g^*)}{\sum_i\phi(S_i,S_j)},\ R_{\mathrm{CEAF}}=\frac{\Phi(g^*)}{\sum_i\phi(G_i,G_j)} \tag{11}$$

$$F1_{\mathrm{CEAF}}=\frac{2\times P_{\mathrm{CEAF}}\times R_{\mathrm{CEAF}}}{P_{\mathrm{CEAF}}+R_{\mathrm{CEAF}}} \tag{12}$$

但是前面这些指标仅关注描述之间的共指关系, 会影响对共指消解模型精度的评估, 因此 BLANC 指标在计算时同时考虑共指关系和非共指关系的识别性能, 计算方法如公式 (13)–公式 (15) 所示.

$$P_{\mathrm{c}}=\frac{\left|C_{\mathrm{g}}\cap C_{\mathrm{s}}\right|}{|C_{\mathrm{s}}|},\ R_{\mathrm{c}}=\frac{\left|C_{\mathrm{g}}\cap C_{\mathrm{s}}\right|}{\left|C_{\mathrm{g}}\right|},\ F1_{\mathrm{c}}=\frac{2\times P_{\mathrm{c}}\times R_{\mathrm{c}}}{P_{\mathrm{c}}+R_{\mathrm{c}}} \tag{13}$$

$$P_{\mathrm{n}}=\frac{\left|N_{\mathrm{g}}\cap N_{\mathrm{s}}\right|}{|N_{\mathrm{s}}|},\ R_{\mathrm{n}}=\frac{\left|N_{\mathrm{g}}\cap N_{\mathrm{s}}\right|}{\left|N_{\mathrm{g}}\right|},\ F1_{\mathrm{n}}=\frac{2\times P_{\mathrm{n}}\times R_{\mathrm{n}}}{P_{\mathrm{n}}+R_{\mathrm{n}}} \tag{14}$$

$$P_{\mathrm{BLANC}}=\frac{P_{\mathrm{c}}+P_{\mathrm{n}}}{2},\ R_{\mathrm{BLANC}}=\frac{R_{\mathrm{c}}+R_{\mathrm{n}}}{2},\ F1_{\mathrm{BLANC}}=\frac{F1_{\mathrm{c}}+F1_{\mathrm{n}}}{2} \tag{15}$$

其中, C_{g} 和 C_{s} 分别表示实际标注的和模型预测出的共指关系集合, N_{g} 和 N_{s} 分别表示实际标注的和模型预测出的非共指关系集合.

3.2 基线模型

为了展示本文事件共指消解方法 CorefCoT 的有效性, 本文选择以下基线模型进行比较.

(1) 基于论元匹配度的事件对模型 Huang19[15]. 先在自动标注的大规模论元数据集上训练一个论元匹配度打分器, 然后将打分器迁移用于完成事件共指识别任务.

(2) 非神经网络的联合模型 Lu&Ng20[71]. 先训练一个有监督的主题模型来为触发词识别任务提供主题信息, 然后运用结构化条件随机场同时完成触发词检测、主题建模和共指识别.

(3) 先进的联合模型 Lu&Ng21[36]. 联合建模实体抽取、实体共指消解、触发词抽取、事件共指消解、事件事实性预测等 7 个相关的事件和实体任务.

(4) 先进的微调事件对模型 MultiRep[13]. 通过在事件句之外引入文档级信息和事件主题信息来同时捕获两个事件描述的句子级、文档级和主题级交互识别共指.

(5) 先进的基于提示技术的方法 CorefPrompt[28]. 通过设计提示将事件共指识别转换为遮盖语言建模任务, 并且使用虚拟标签词分析事件类型和论元层面的整体关联性.

3.3 实验结果

基线模型与本文 CorefCoT 方法在 KBP2017 数据集上的事件共指消解性能如表 4 所示, 其中, CorefCoT(F) 和 CorefCoT(L) 分别表示使用 FlanT5 和 LLaMA2 模型作为骨架的方法性能. 结果显示 CorefCoT 方法取得了比基线模型更好的性能. 与 CorefPrompt 模型相比, CorefCoT 方法通过引入外部知识改进了对事件语义相似度与共指不一致的低置信度预测结果的判断, 因此进一步提高了性能, 关注共指链的 MUC 指标提高了 1.3 个百分点.

表 4 基线模型与 CorefCoT 方法在 KBP2017 数据集上的性能表现 (%)

模型	Pairwise	MUC	B^3	$CEAF_e$	BLANC	AVG
Huang19	—	35.7	43.2	40.0	32.4	36.8
Lu&Ng20	—	37.1	44.5	40.0	29.9	37.9
Lu&Ng21	—	45.2	54.7	53.8	38.2	48.0
MultiRep	41.8/38.5/40.1	46.2	57.4	59.0	42.0	51.2
CorefPrompt	42.1/39.0/40.5	45.3	57.5	59.9	42.3	51.3
CorefCoT(F)	40.9/42.0/41.5	46.3	57.3	59.8	42.4	51.5
CorefCoT(L)	41.8/40.6/41.2	46.6	57.6	60.1	42.5	51.7

实验结果显示, 得益于事件表示从基于句子级上下文变为基于片段级上下文, 基于文本块建模事件的方法 Lu&Ng20 和 Lu&Ng21 的性能都优于基于句子建模事件的模型 Huang19. 这证明了融入更大范围的上下文信息有助于识别事件共指. 之后的 MultiRep 模型通过引入更大范围的文档级上下文信息以及事件主题信息进一步改进了事件共指消解的性能. 而基于提示技术的 CorefPrompt 方法借助提示模板中引入的共指知识, 仅基于片段级上下文就取得了与 MultiRep 模型相当的性能. 这些结果都证明了文档级上下文中的事件信息以及提示技术在事件共指消解任务上的有效性. 本文提出的 CorefCoT 方法通过生成文档级事件摘要以及思维链风格的推理路径同时结合了文档信息以及提示技术, 并且还通过指令微调使得 FlanT5 等基础大语言模型能够生成共指相关外部知识来辅助识别, 因此取得了比基线模型更好的性能.

为了排除上游触发词检测环节级联错误带来的影响, 本文还在表 5 中给出了 MultiRep 模型、CorefPrompt 模型和本文方法在识别正确触发词上的事件对共指识别性能以及在标注触发词上的共指消解性能, MultiRep 模型和 CorefPrompt 模型均使用原论文公布的权重. 表 5 显示, 通过将共指识别分解为两个步骤并且引入外部知识辅助识别, 本文方法在对 CorefPrompt 模型的低置信度结果进行重新预测后, 显著提高了事件对共指识别的召回率, 在识别正确触发词和标注触发词上召回率分别提高了 3.2 个百分点和 3.2 个百分点.

表 5 识别正确触发词和标注触发词上的事件共指消解性能 (%)

模型	识别正确触发词	标注触发词					
	Pairwise	MUC	B^3	$CEAF_e$	BLANC	AVG	Pairwise
MultiRep	72.9/73.8/73.4	72.3	85.1	82.2	77.3	79.2	66.7/64.4/65.6
CorefPrompt	74.9/74.5/74.7	73.8	84.9	82.2	75.4	79.1	70.9/67.8/69.3
CorefCoT(L)	74.3/77.7/75.9	74.9	85.3	82.8	76.5	79.9	70.2/71.0/70.6

此外, 为了测试模型的跨领域适用性, 本文还将训练好的 CorefCoT 模型迁移用于识别 ACE2005 数据集中的事件共指 (基于标注触发词). 与 KBP 数据集上的设置保持一致, 本文同样仅仅对当前先进模型预测结果那些低置信度的样本进行重新预测. 考虑到 CorefPrompt 模型需要依赖识别出的论元信息, 为了与先前的工作 Yao 2023[63] 和 Lai 2021[25]进行公平比较, 本文选择对 MultiRep 模型的预测结果进行重新预测, 实验结果如表 6 所示.

实验结果显示, 本文训练好的 CorefCoT 模型在 ACE2005 数据集上同样表现出了很强的事件共指识别能力, 整体性能优于当前先进的基线模型 Yao 2023[63]. 在对 MultiRep 模型的低置信度结果进行重新预测后, 无论是以 FlanT5 还是 LLaMA2 模型作为骨架的 CorefCoT 模型都可以改进共指识别, 在所有共指指标上都取得了性能提升. 这充分证明了本文方法的有效性, 即生成的外部知识对于识别其他领域的事件共指同样有效.

表 6 基线模型与 CorefCoT 方法在 ACE2005 数据集上的性能表现 (%)

模型	MUC	B^3	$CEAF_e$	BLANC	AVG
Lai 2021[25]	—	—	—	—	84.0
Yao 2023[63]	79.2	92.4	88.8	86.2	86.6
MultiRep	80.6	92.1	88.5	84.3	86.4
CorefCoT(L)	80.9	92.4	88.9	85.4	86.9
CorefCoT(F)	81.9	92.4	88.6	85.5	87.1

3.4 实验分析

3.4.1 两步骤共指框架的有效性

考虑到本文在训练阶段仅使用了欠采样出的少量代表性样本, 在预测阶段仅对 CorefPrompt 模型低置信度结果进行重新预测 (重打分). 为了分析欠采样和重打分操作对事件共指消解的影响, 本文构建了两个基准事件对模型 ScoreBERT 和 ScoreRoBE, 它们使用与本文方法相同的训练集并且执行相同的重打分操作. 与 CorefCoT 方法的提示形式类似, ScoreBERT 模型和 ScoreRoBE 模型首先在片段中插入特殊的事件标记符以标记出两个触发词的位置, 然后运用 BERT/RoBERTa 编码器结合注意力机制生成事件表示, 最后将事件表示拼接后送入分类器完成预测. 此外, 为了评估大语言模型在零样本 (zero-shot) 实验设置下的性能, 本文还运用第 2.3 节中设计的共指识别提示 P_{Coref} 引导 ChatGPT 模型同样生成思维链风格的路径识别事件共指.

表 7 显示, 欠采样和重打分操作具有显著影响. 在使用欠采样出的代表性样本进行训练后, 基准事件对模型 ScoreBERT 和 ScoreRoBE 在共指样本识别上 (Coref) 就取得了比 CorefPrompt 模型更好的性能, $F1$ 值分别大幅提升了 18.3 个百分点和 21.4 个百分点. 但是由于基准事件对模型无法捕获事件之间的细粒度交互, 这种做法也会伤害对非共指样本的判断, 因此总体事件对性能 (Macro-$F1$) 和总体共指消解性能 (AVG) 都低于 CorefPrompt 模型. ChatGPT 模型凭借强大的上下文理解和推理能力, 仅使用本文设计的共指识别提示就可以在零样本设置下改进对共指样本的识别. 但是由于 ChatGPT 模型容易发现事件之间的潜在联系, 因此会倾向于将样本识别为共指, 导致对非共指样本的识别性能不佳, 总体性能也低于小规模的 CorefPrompt 模型. 本文提出的 CorefCoT 方法不仅使用思维链提示激发出大语言模型的推理能力, 而且通过指令微调使得大语言模型与事件共指消解任务更适配, 因此相比未经微调的 ChatGPT 模型取得了更好的性能. 这些结果说明, 尽管欠采样和重打分操作可以引导模型更关注共指样本, 但是也会干扰对非共指样本的识别, 因此本文提出的两步骤共指框架对于改进共指消解性能很重要.

表 7 执行欠采样和重打分操作的基线模型的事件共指消解性能 (%)

模型	所有样本	低置信度重新预测样本		
	AVG	Coref	Non-Coref	Macro-$F1$
CorefPrompt	51.3	13.8/18.4/15.8	74.7/67.4/70.7	43.3
ScoreBERT	51.2	22.3/72.4/34.1	78.2/28.2/41.5	37.8
ScoreRoBE	51.4	24.2/81.1/37.2	83.6/27.5/41.4	39.3
ChatGPT	51.2	22.5/62.2/33.1	78.4/39.0/52.0	42.6
CorefCoT(L)	51.7	21.7/46.9/29.7	77.4/51.7/62.0	45.8

为了更深入地分析事件摘要和共指推理路径生成步骤对事件共指消解的影响, 本文设计了以下消融实验: (1) –Summary: 移除事件摘要步骤, 直接使用共指识别提示 P_{Coref} 完成任务, 此时提示中的输入就变成了围绕两个触发词截取的原文, 而不再是事件句和事件摘要拼接出的增强事件句; (2) –CoT: 移除共指推理路径生成步骤, 直接要求模型基于增强事件句输出共指结果. 实验结果如表 8 所示.

表 8 显示, 事件摘要和共指推理路径生成步骤对共指消解都具有显著影响. 移除事件摘要步骤后, 以 FlanT5 和 LLaMA2 作为骨架的 CorefCoT 方法在共指样本识别上 (Coref) 都出现了性能下降, 尤其是召回率分别大幅下降了 9.1 个百分点和 5.6 个百分点. 背后的原因可能是因为没有预先从文档中筛选出事件相关信息, 模型在生成推理路径时, 一方面由于输入长度限制只能从触发词周围的局部上下文中挖掘线索, 另一方面难以聚焦于分析事件

关联性, 导致模型在一定程度上退化为基于语义相似度的方法, 因此会遗漏识别相似度与共指不一致的样本. 而移除推理路径生成步骤后, 大语言模型由于无法生成中间的分析步骤激发推理能力而退化为一个分类模型, 因此与 ScoreBERT 和 ScoreRoBE 模型类似, 受到训练数据的影响而倾向于将样本识别为共指, 在提高共指样本识别召回率的同时也干扰到非共指样本的识别, *F*1 值分别大幅下降了 8.4 个百分点和 16.5 个百分点, 造成总体性能的下降.

表 8 移除事件摘要和推理路径生成步骤后的事件共指消解性能 (%)

模型	所有样本	低置信度重新预测样本		
	AVG	Coref	Non-Coref	Macro-*F*1
CorefCoT(F)	51.5	22.9/72.4/34.8	79.5/30.5/44.1	39.5
–Summary	51.4	22.5/63.3/33.2	78.4/37.9/51.1	42.2
–CoT	51.4	23.8/84.7/37.1	83.9/22.7/35.7	36.4
CorefCoT(L)	51.7	21.7/46.9/29.7	77.4/51.7/62.0	45.8
–Summary	51.4	20.9/41.3/27.7	76.8/55.4/64.4	46.0
–CoT	51.4	22.8/70.4/34.4	79.1/32.0/45.5	40.0

3.4.2 外部知识的影响

为了分析外部知识对事件共指消解的影响, 本文首先按照外部知识的类型和数量对生成的推理路径进行划分, 然后比较了 CorefCoT 方法与 CorefPrompt 模型在这些样本上的性能差异. 特别地, 本文将生成数据中的领域相关知识 (例如司法类事件对应的法律知识) 统称为领域知识. 各类外部知识在 CorefCoT 方法生成的推理路径中的占比如表 9 所示. 为了简化表示, 本文将触发词、论元等事件内部信息合并统计.

表 9 各类事件信息在 CorefCoT(L) 方法生成的推理路径中的占比 (%)

事件信息	包含信息的路径比例
事件内部信息	97.6
篇章连贯性	35.7
逻辑关系知识	54.4
常识背景知识	41.5
领域知识	25.0

表 9 显示, 对于大语言模型, 已有研究中广泛使用的事件内部信息同样是识别共指的有效方式, 高达 97.6% 的推理路径中包含有事件内部信息. 而已有研究中较少考虑的逻辑关系知识、常识背景知识等外部知识在推理路径中的占比也很高, 这充分验证了本文 CorefCoT 方法的有效性, 即基础大语言模型通过指令微调学习到了为事件对生成共指相关外部知识的能力.

考虑到同一推理路径中可能包含多个相同类型的外部知识, 为了分析外部知识类型和数量对共指消解的影响, 本文按照以下两种方式对样本进行划分. 按知识类型划分: 包含一种外部知识 (one)、包含两种 (two)、包含 3 种及以上 (more). 按知识数量划分: 包含 3 个事件外部知识 (small)、包含 4 个 (medium)、包含 5 个及以上 (large). 表 10 展示了在这些情况下 CorefCoT(L) 方法与 CorefPrompt 模型在事件对共指识别性能上的对比.

表 10 显示, 大部分推理路径都包含一种或两种外部知识, 并且本文 CorefCoT 方法在这两种情况下都改进了对共指事件对的识别, *F*1 值分别显著提升了 20.4 个百分点和 14.8 个百分点. 但是在进一步增加外部知识多样性后性能却出现了下降, 甚至还不如仅使用事件内部信息的 CorefPrompt 模型. 这可能是因为模型借助新增加的外部知识捕获到了许多事件之间潜在的关联性, 在识别出遗漏共指事件对的同时, 也增加了将语义模糊的事件对错误识别为共指的风险, 因此在提高召回率的同时也造成了准确率的下降. 在外部知识数量方面, CorefCoT 方法在各种外部知识数量的情况下都改进了共指识别的性能, 尤其是在 Small 和 Large 场景下事件对共指识别 *F*1 值分别显著提高了 22.3 个百分点和 20.2 个百分点. 增加外部知识的数量不仅使得模型可以从更多角度分析事件关联性, 而且还可以汇总同一知识类型的多方线索, 因此有更多的机会发现事件之间共指的可能性. Large 场景在引入

更多事件外部知识后, 共指识别召回率大幅提高了 55.7 个百分点, 但是这也会对非共指事件对的识别造成干扰, 使得总体事件对识别性能 (Macro-*F*1) 出现下降. 综上所述, 增加外部知识的类型和数量在总体上都对事件共指消解具有积极影响, 但是过多的知识类型和数量也会使得模型过分关注事件之间的潜在联系而干扰到判断.

表 10 不同知识类型和数量下 CorefPrompt 和 CorefCoT 方法的事件对共指识别性能 (%)

情况	占比	CorefPrompt		CorefCoT(L)	
		Pairwise	Macro-*F*1	Pairwise	Macro-*F*1
One	42.3	7.6/10.5/8.8	43.1	19.3/59.6/29.2	46.3
Two	49.9	18.3/22.6/20.2	41.4	30.2/41.7/35.0	50.8
More	7.8	15.4/36.4/21.6	44.5	10.9/45.5/17.5	24.7
Small	17.1	6.3/7.4/6.8	42.1	19.3/59.3/29.1	44.2
Medium	51.6	14.7/21.2/17.4	43.8	18.5/24.2/21.0	47.2
Large	31.3	15.3/18.6/16.8	42.2	24.6/74.3/37.0	36.1

为了深入分析外部知识对 CorefCoT 方法的贡献, 本文还通过移除训练数据中的外部知识来构建模型变体, 使得微调后大语言模型生成的推理路径中仅包含事件内部信息. 实验结果如表 11 所示.

表 11 移除外部知识后的事件共指消解性能 (%)

模型	所有样本	低置信度重新预测样本		
	AVG	Coref	Non-Coref	Macro-*F*1
CorefCoT(F)	51.5	22.9/72.4/34.8	79.5/30.5/44.1	39.5
–Knowledge	51.3	22.6/75.0/34.8	79.1/26.9/40.1	37.4
CorefCoT(L)	51.7	21.7/46.9/29.7	77.4/51.7/62.0	45.8
–Knowledge	51.6	22.5/58.7/32.5	78.2/42.3/54.9	43.7

表 11 显示, 在移除外部知识后, 以 FlanT5 和 LLaMA2 为骨架的 CorefCoT 方法在共指消解总体 *F*1 分数 (AVG) 和事件对总体 *F*1 分数 (Macro-*F*1) 上都出现了性能下降. 由于训练数据不再包含外部知识, 指令微调后的基础大语言模型仅学习到了生成事件内部信息的能力, 在一定程度上退化为传统基于语义相似度的方法, 因此同样难以处理语义相似度与共指不一致的情况. 由于欠采样出的数据更关注共指样本, 微调后的大语言模型会倾向于将语义模糊样本判别为共指, 导致非共指样本的召回率下降, 从而影响到整体性能.

为了进一步验证外部知识的有效性, 本文还首先从 CorefCoT(L) 方法生成的推理路径中抽取出外部知识, 然后将其添加到第 3.4.1 节构建的基线模型 ScoreBERT 和 ScoreRoBE 中, 评估外部知识对基准事件对模型的影响. 具体来说, 对于 ScoreBERT 和 ScoreRoBE 模型, 本文在原始输入之外同步输入生成的外部知识, 然后同样使用 BERT/RoBERTa 模型编码后取出起始位置标记对应的嵌入作为外部知识表示, 最后运用一个门控单元融合原始事件表示和外部知识表示作为新的特征送入到分类器. 实验结果如表 12 所示.

表 12 融入外部知识后基准事件对模型的事件共指消解性能 (%)

模型	所有样本	低置信度重新预测样本		
	AVG	Coref	Non-Coref	Macro-*F*1
ScoreBERT	51.2	22.3/72.4/34.1	78.2/28.2/41.5	37.8
+CoTPath	51.5	24.3/78.1/37.1	83.1/30.8/45.0	41.0
ScoreRoBE	51.4	24.2/81.1/37.2	83.6/27.5/41.4	39.3
+CoTPath	51.6	25.0/81.6/38.2	85.2/30.1/44.4	41.4

表 12 中的结果显示, 基准事件对模型 ScoreBERT 和 ScoreRoBE 在融入 CorefCoT 方法生成的外部知识后都改进了事件共指消解的性能, 事件对总体 *F*1 分数 (Macro-*F*1) 分别提高了 3.2 个百分点和 2.1 个百分点. 尤其是 ScoreRoBE 模型在融入生成的外部知识后, 在共指消解总体性能 (AVG) 和事件对总体性能 (Macro-*F*1) 上都取得了与 FlanT5 模型作为骨架的 CorefCoT(F) 方法类似的性能. 这些结果充分验证了本文方法生成的外部知识在改

进事件共指识别方面的有效性. 为了直观地展现外部知识对识别事件共指的帮助, 本文在表 13 中展示了一些 CorefCoT(L) 方法在引入外部知识后成功修正 CorefPrompt 模型预测错误结果的例子.

表 13 CorefCoT 方法引入外部知识修正 CorefPrompt 模型预测错误结果的例子

内容	生成的外部知识
Mickey Mouse {arrested}$_{ev1}$ in Disneyland protest. … I'm going to venture to guess that when this dude joined the force he never imagined his photo of him {arresting}$_{ev2}$ a mouse would end up in the news! 米老鼠在迪士尼乐园的抗议中{被逮捕}$_{ev1}$. … 我大胆猜测这个家伙加入部队时, 从未想过他{逮捕}$_{ev2}$老鼠的照片最终会出现在新闻中!	4. Narrative Continuity: The narrative surrounding both phrases revolves around the unexpected arrest of a Disney character during a protest, creating a cohesive storyline that links the two phrases as part of the same event. … 4. 叙事连续性: 这两个短语的叙述围绕着迪士尼角色在抗议期间意外被捕展开, 创建了一根连贯的故事线将这两个短语作为同一事件的一部分串联起来. …
Apple announced Tuesday that it {hired}$_{ev1}$ Angela Ahrendts, the chief executive of Burberry, the British luxury fashion company, as a member of its executive team. … She will {start}$_{ev2}$ working for Apple in the spring. 苹果公司周二宣布, {聘请}$_{ev1}$英国奢侈时装公司博柏利首席执行官安吉拉阿伦茨作为其执行团队成员. … 她将于春季{开始}$_{ev2}$为苹果工作.	3. Causal Relationship: The hiring of Ahrendts is the cause for her subsequent employment. The act of hiring her directly leads to her starting work, creating a causal link between the two events. … 3. 因果关系: 雇用阿伦茨是她随后就业的原因. 雇用她的行为直接导致她开始工作, 从而在两个事件之间建立了因果关系. …
Initial reports said the {assailant}$_{ev1}$ was a man of North African appearance, about 30 years old, with a small beard. … Interior Minister Manuel Valls ordered the {terrorism}$_{ev2}$ investigation. 初步报道称, {袭击者}$_{ev1}$是一名北非男子, 年龄约 30 岁, 留着小胡子. … 内政部长曼努埃尔瓦尔斯下令进行{恐怖主义}$_{ev2}$调查. …	4. Geopolitical Context: The "assailant" event is localized within France, specifically in the transport station of La Défense, while the "terrorism" investigation extends to the broader international context, including France's involvement in military interventions in Africa and the potential threat posed by al-Qaida in the Islamic Maghreb. 4. 地缘政治背景: "袭击者"事件仅限于法国境内, 特别是在拉德芳斯的交通站, 而"恐怖主义"调查则扩展到更广泛的国际背景, 包括法国参与非洲军事干预以及伊斯兰马格里布基地组织构成的潜在威胁.

表 13 中的这些例子由于事件语义相似度与共指不一致而被 CorefPrompt 模型误判, 而本文 CorefCoT 方法通过生成篇章连贯性、逻辑关系知识、常识背景知识等外部知识使得模型可以结合事件之间潜在的关联性作出预测, 因此改进了对这些样本的共指识别. 例如对于表 12 中的第 2 个例子, 如果在识别事件共指时能够考虑到"聘请"和"开始工作"之间潜在的因果联系, 就更容易识别出这两个事件描述之间的共指, 而 CorefPrompt 等侧重于分析事件语义匹配性的模型就很难捕获到这种相似度之外的关联性.

3.4.3 序列到序列共指消解的性能

考虑到最近的实体指代研究[72]表明直接以文字序列形式生成共指簇也是一种可行的共指消解方案, 本文尝试将这种共指簇生成方法应用于事件共指消解任务, 评估在标准指令微调框架下直接进行序列到序列事件共指消解的性能. 受统一信息抽取框架 InstructUIE[52]的启发, 本文首先将包括事件共指消解在内的多个实体和事件任务转换为指令跟随的序列生成任务, 然后运用构建好的指令数据来微调大语言模型. 在预测时, 再按照每个任务对应的输出格式从模型生成的文字序列中解析出结果. 表 14 展示了几个代表性任务的提示形式.

本文选择与 CorefCoT 方法相同规模的 FlanT5 模型和 LLaMA2 模型进行实验, 并且基于相同的事件触发词集合来预测事件共指簇. 具体来说, 本文按照以下 3 种数据集设置来指令微调基础大语言模型: (1) 只使用事件共指消解任务的指令数据 (ECR); (2) 训练数据集中包含多种事件任务 (Evt), 包括事件共指消解、触发词检测、论元抽取等在内的共计 7 个事件任务; (3) 在所有的事件任务上进一步添加实体任务 (Evt+Ent), 包括实体识别、实体类别预测、实体共指消解等在内的共计 5 个实体任务. 实验结果如表 15 所示.

表 14 代表性实体和事件任务的提示形式

任务	提示
实体识别	Please find all entity spans belonging to given categories in the text and output the entity spans and corresponding types. Categories are facility (FAC), geopolitical entity (GPE), location (LOC), organization (ORG), person (PER). Output format is “type1: span1; type2: span2; type3: span3”. Text: {Segment}. Answer:
实体共指消解	Please cluster the listed entity spans according to whether they refer to the same entity in the following text. All spans in the same cluster refer to the same entity. Listed entity spans are “{[ID1]Span1; [ID2]Span2;...}”. Output format is “cluster1: {span1; span2}, cluster2: {span3; span4}, cluster3: {span5; span6}”. Text: {Segment}. Answer:
事件抽取	Please find trigger words expressing the occurrences of specific categories of events and arguments playing specific roles in the events, and output the trigger words and corresponding arguments. Event categories are... Argument roles are... Output format is “{type1: trigger1, [role1: argument1; role2: argument2]}, {type2: trigger2, [role3: argument3; role4: argument4]}”. Text: {Segment}. Answer:
事件共指消解	Please cluster the listed event trigger words according to their argument compatibilities and whether they express the occurrences of the same events in the text. All trigger words in the same cluster refer to the same event. Listed event triggers are “{[ID1]Trigger1; [ID2]Trigger2;...}”. Output format is “cluster1: {trigger1; trigger2}, cluster2: {trigger3; trigger4}”. Text: {... [ID1]Trigger1... [ID2]Trigger2...}. Answer:

表 15 使用不同数据集指令微调大语言模型后的事件共指消解性能 (%)

模型	数据	MUC	B^3	$CEAF_e$	BLANC	AVG
CorefPrompt	—	45.3	57.5	59.9	42.3	51.3
FlanT5	ECR	38.4	55.4	54.0	39.7	46.8
FlanT5	Evt	39.4	55.5	53.0	39.0	46.7
FlanT5	Evt+Ent	39.9	54.8	52.7	40.2	46.9
LLaMA2	ECR	31.1	54.9	52.2	37.1	43.8
LLaMA2	Evt	32.1	54.0	51.1	36.5	43.4
LLaMA2	Evt+Ent	35.2	53.5	50.6	37.0	44.1

表 15 中的结果显示, 与序列到序列模型在实体共指消解任务上取得成功不同, 使用不同数据设置 (ECR、Evt、Evt+Ent) 以及不同模型结构 (编码器-解码器、纯解码器) 指令微调出的大语言模型在事件共指消解任务上表现出了相似的性能, 甚至低于小规模的 CorefPrompt 模型. 背后可能有两方面原因: 一方面, 由于事件表述多样性和事件信息缺失性, 事件共指消解比实体共指消解任务更具挑战性; 另一方面, 实体任务在预训练数据中更为常见, 而大语言模型对事件概念的认识则较为模糊. 此外, 与已有研究[73]大多认为指令多样性对指令微调模型具有显著影响不同, 表 15 显示, 在增加事件和实体任务的指令数据后, 模型的事件共指消解性能并没有明显改善, 多事件任务设置下 (Evt) 模型的总体性能 (AVG) 甚至低于只使用共指消解任务数据的模型. 最后, 与本文 CorefCoT 方法的结果不同, 以 LLaMA2 模型作为骨架的序列到序列模型的性能都低于以 FlanT5 模型作为骨架的版本. 这可能是因为输入具有自然语言的形式而输出具有结构化表示, 由于纯解码器模型将输入和输出拼接起来作为整体进行训练, 因此受到的影响更大. 这些结果都表明, 仅采用标准指令微调框架训练大语言模型并不足以完成事件共指消解任务, 这也验证了本文 CorefCoT 方法在激发大语言模型共指推理能力方面的有效性.

4 总 结

针对触发词、论元等事件内部信息分析事件关联性不完备的问题, 本文提出了外部知识增强的事件共指消解方法 CorefCoT, 通过运用大语言模型生成篇章连贯性、逻辑关系知识、常识背景知识等共指相关外部知识来发现事件之间的潜在联系, 从而降低对事件内部信息的依赖. 该方法首先运用超大语言模型 ChatGPT 构造包含外部知识的训练数据, 然后在训练数据上指令微调 FlanT5 等基础大语言模型, 最后运用微调后的大语言模型生成文档级事件摘要以及思维链风格的共指推理路径, 结合事件内部信息和外部知识预测共指. 实验结果显示, CorefCoT 方法通过引入外部知识取得了比当前先进的基线模型更好的性能.

References:

[1] Doddington G, Mitchell A, Przybocki M, Ramshaw L, Strassel S, Weischedel R. The automatic content extraction (ACE) program-tasks, data, and evaluation. In: Proc. of the 4th Int'l Conf. on Language Resources and Evaluation. Lisbon: European Language Resources Association, 2004. 837–840.

[2] Mitamura T, Liu ZZ, Hovy EH. Overview of TAC KBP 2015 event nugget track. In: Proc. of the 2015 Text Analysis Conf. 2015. 1–11.

[3] Li ML, Ma TF, Yu M, Wu LF, Gao T, Ji H, McKeown K. Timeline summarization based on event graph compression via time-aware optimal transport. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 6443–6456. [doi: 10.18653/v1/2021.emnlp-main.519]

[4] Lee IT, Pacheco ML, Goldwasser D. Weakly-supervised modeling of contextualized event embedding for discourse relations. In: Findings of the Association for Computational Linguistics: EMNLP 2020. ACL, 2020. 4962–4972. [doi: 10.18653/v1/2020.findings-emnlp.446]

[5] Huang KH, Peng NY. Document-level event extraction with efficient end-to-end learning of cross-event dependencies. In: Proc. of the 3rd Workshop on Narrative Understanding. ACL, 2021. 36–47. [doi: 10.18653/v1/2021.nuse-1.4]

[6] Kriman S, Ji H. Joint detection and coreference resolution of entities and events with document-level context aggregation. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing: Student Research Workshop. ACL, 2021. 174–179. [doi: 10.18653/v1/2021.acl-srw.18]

[7] Phung D, Nguyen TN, Nguyen TH. Hierarchical graph convolutional networks for jointly resolving cross-document coreference of entity and event mentions. In: Proc. of the 15th Workshop on Graph-based Methods for Natural Language Processing. Mexico City: ACL, 2021. 32–41. [doi: 10.18653/v1/2021.textgraphs-1.4]

[8] Chen Z, Ji H, Haralick R. A pairwise event coreference model, feature impact and evaluation for event coreference resolution. In: Proc. of the 2009 Workshop on Events in Emerging Text Types. Borovets: ACL, 2009. 17–22.

[9] Cybulska A, Vossen P. "Bag of Events" approach to event coreference resolution. Supervised classification of event templates. International Journal of Computational Linguistics and Applications, 2015, 6(2): 11–27.

[10] Krause S, Xu FY, Uszkoreit H, Weissenborn D. Event linking with sentential features from convolutional neural networks. In: Proc. of the 20th SIGNLL Conf. on Computational Natural Language Learning. Berlin: ACL, 2016. 239–249. [doi: 10.18653/v1/K16-1024]

[11] Nguyen TH, Meyers A, Grishman R. New York University 2016 system for KBP event nugget: A deep learning approach. In: Proc. of the 2016 Text Analysis Conf. 2016. https://www.researchgate.net/publication/316471878_New_York_University_2016_System_for_KBP_Event_Nugget_A_Deep_Learning_Approach

[12] Lu J, Ng V. Span-based event coreference resolution. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021. 13489–13497. [doi: 10.1609/aaai.v35i15.17591]

[13] Xu S, Li PF, Zhu QM. Improving event coreference resolution using document-level and topic-level information. In: Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing. Abu Dhabi: ACL, 2022. 6765–6775. [doi: 10.18653/v1/2022.emnlp-main.454]

[14] Yu XD, Yin WP, Roth D. Pairwise representation learning for event coreference. In: Proc. of the 11th Joint Conf. on Lexical and Computational Semantics. Seattle: ACL, 2022. 69–78. [doi: 10.18653/v1/2022.starsem-1.6]

[15] Huang YJ, Lu J, Kurohashi S, Ng V. Improving event coreference resolution by learning argument compatibility from unlabeled data. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: ACL, 2019. 785–795. [doi: 10.18653/v1/N19-1085]

[16] Choubey PK, Huang RH. Improving event coreference resolution by modeling correlations between event coreference chains and document topic structures. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 1 (Long Papers). Melbourne: ACL, 2018. 485–495. [doi: 10.18653/v1/P18-1045]

[17] Danlos L, Gaiffe B. Event coreference and discourse relations. In: Korta L, Larrazabal JM, eds. Truth, Rationality, Cognition, and Music. Dordrecht: Springer, 2002. 191–209. [doi: 10.1007/978-94-017-0548-6_10]

[18] Qin HW, Cui R. Event eoreference and discourse coherence. Contemporary Linguistics, 2009, 11(1): 10–20 (in Chinese with English abstract).

[19] Choubey PK, Lee A, Huang RH, Wang L. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 5374–5386. [doi: 10.18653/v1/2020.acl-main.478]

[20] Ravi S, Tanner C, Ng R, Shwartz V. What happens before and after: Multi-event commonsense in event coreference resolution. In: Proc. of the 17th Conf. of the European Chapter of the Association for Computational Linguistics. Dubrovnik: ACL, 2023. 1708–1724. [doi: 10.

18653/v1/2023.eacl-main.125]

[21] Mostafazadeh N, Kalyanpur A, Moon L, Buchanan D, Berkowitz L, Biran O, Chu-Carroll J. GLUCOSE: Generalized and contextualized story explanations. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. ACL, 2020. 4569–4586. [doi: 10.18653/v1/2020.emnlp-main.370]

[22] Ning Q, Wu H, Peng HR, Roth D. Improving temporal relation extraction with a globally acquired statistical resource. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers). New Orleans: ACL, 2018. 841–851. [doi: 10.18653/v1/N18-1077]

[23] Liu K, Chen YB, Liu J, Zuo XY, Zhao J. Extracting events and their relations from texts: A survey on recent research progress and challenges. AI Open, 2020, 1: 22–39. [doi: 10.1016/j.aiopen.2021.02.004]

[24] Nath A, Avari SM, Chelle A, Krishnaswamy N. Okay, let's do this! Modeling event coreference with generated rationales and knowledge distillation. In: Proc. of the 2024 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers). Mexico City: ACL, 2024. 3931–3946. [doi: 10.18653/v1/2024.naacl-long.218]

[25] Lai T, Ji H, Bui T, Tran QH, Dernoncourt F, Chang W. A context-dependent gated module for incorporating symbolic semantics into event coreference resolution. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2021. 3491–3499. [doi: 10.18653/v1/2021.naacl-main.274]

[26] Kenyon-Dean K, Cheung JCK, Precup D. Resolving event coreference with supervised representation learning and clustering-oriented regularization. In: Proc. of the 7th Joint Conf. on Lexical and Computational Semantics. New Orleans: ACL, 2018. 1–10. [doi: 10.18653/v1/S18-2001]

[27] Barhom S, Shwartz V, Eirew A, Bugert M, Reimers N, Dagan I. Revisiting joint modeling of cross-document entity and event coreference resolution. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 4179–4189. [doi: 10.18653/v1/P19-1409]

[28] Xu S, Li PF, Zhu QM. CorefPrompt: Prompt-based event coreference resolution by measuring event type and argument compatibilities. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 15440–15452. [doi: 10.18653/v1/2023.emnlp-main.954]

[29] Ding BW, Min QK, Ma SK, Li YJ, Yang LY, Zhang Y. A rationale-centric counterfactual data augmentation method for cross-document event coreference resolution. In: Proc. of the 2024 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers). Mexico City: ACL, 2024. 1112–1140. [doi: 10.18653/v1/2024.naacl-long.63]

[30] Min QK, Guo QP, Hu XK, Huang SF, Zhang Z, Zhang Y. Synergetic event understanding: A collaborative approach to cross-document event coreference resolution with large language models. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics, Vol. 1 (Long Papers). Bangkok: ACL, 2024. 2985–3002. [doi: 10.18653/v1/2024.acl-long.164]

[31] Chen C, Ng V. Joint inference over a lightly supervised information extraction pipeline: Towards event coreference resolution for resource-scarce languages. In: Proc. of the 30th AAAI Conf. on Artificial Intelligence. Phoenix: AAAI, 2016. 2913–2920. [doi: 10.1609/aaai.v30i1.10392]

[32] Lu J, Venugopal D, Gogate V, Ng V. Joint inference for event coreference resolution. In: Proc. of the 26th Int'l Conf. on Computational Linguistics: Technical Papers. Osaka: The COLING 2016 Organizing Committee, 2016. 3264–3275.

[33] Araki J, Mitamura T. Joint event trigger identification and event coreference resolution with structured perceptron. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015. 2074–2080. [doi: 10.18653/v1/D15-1247]

[34] Lu J, Ng V. Joint learning for event coreference resolution. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1 (Long Papers). Vancouver: ACL, 2017. 90–101. [doi: 10.18653/v1/P17-1009]

[35] Lee K, He LH, Lewis M, Zettlemoyer L. End-to-end neural coreference resolution. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017. 188–197. [doi: 10.18653/v1/D17-1018]

[36] Lu J, Ng V. Constrained multi-task learning for event coreference resolution. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2021. 4504–4514. [doi: 10.18653/v1/2021.naacl-main.356]

[37] Lu J, Ng V. Conundrums in event coreference resolution: Making sense of the state of the art. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 1368–1380. [doi: 10.18653/v1/2021.emnlp-main.103]

[38] Cui LY, Wu Y, Liu J, Yang S, Zhang Y. Template-based named entity recognition using BART. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. ACL, 2021. 1835–1845. [doi: 10.18653/v1/2021.findings-acl.161]

[39] Chia YK, Bing LD, Poria S, Si L. RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In: Findings of the Association for Computational Linguistics: ACL 2022. Dublin: ACL, 2022. 45–57. [doi: 10.18653/v1/2022.findings-

acl.5]

[40] Chen X, Zhang NY, Xie X, Deng SM, Yao YZ, Tan CQ, Huang F, Si L, Chen HJ. KnowPrompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In: Proc. of the 2022 ACM Web Conf. Lyon: ACM, 2022. 2778–2788. [doi: 10.1145/3485447.3511998]

[41] Yang SC, Song DD. FPC: Fine-tuning with prompt curriculum for relation extraction. In: Proc. of the 2nd Conf. of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th Int'l Joint Conf. on Natural Language Processing, Vol. 1 (Long Papers). ACL, 2022. 1065–1077. [doi: 10.18653/v1/2022.aacl-main.78]

[42] Li HC, Mo T, Fan HC, Wang JK, Wang JX, Zhang FH, Li WP. KiPT: Knowledge-injected prompt tuning for event detection. In: Proc. of the 29th Int'l Conf. on Computational Linguistics. Gyeongju: International Committee on Computational Linguistics, 2022. 1943–1952.

[43] Ma YB, Wang ZH, Cao YX, Li MK, Chen MQ, Wang K, Shao J. Prompt for extraction? PAIE: Prompting argument interaction for event argument extraction. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics, Vol. 1 (Long Papers). Dublin: ACL, 2022. 6759–6774. [doi: 10.18653/v1/2022.acl-long.466]

[44] Dai L, Wang B, Xiang W, Mo YJ. Bi-directional iterative prompt-tuning for event argument extraction. In: Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing. Abu Dhabi: ACL, 2022. 6251–6263. [doi: 10.18653/v1/2022.emnlp-main.419]

[45] Shen SR, Zhou H, Wu TT, Qi GL. Event causality identification via derivative prompt joint learning. In: Proc. of the 29th Int'l Conf. on Computational Linguistics. Gyeongju: International Committee on Computational Linguistics, 2022. 2288–2299.

[46] Lu YJ, Liu Q, Dai D, Xiao XY, Lin HY, Han XP, Sun L, Wu H. Unified structure generation for universal information extraction. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics, Vol. 1 (Long Papers). Dublin: ACL, 2022. 5755–5772. [doi: 10.18653/v1/2022.acl-long.395]

[47] Wei J, Wang XZ, Schuurmans D, Bosma M, Ichter B, Xia F, Chi EH, Le QV, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 24824–24837.

[48] Nguyen H, Liu Y, Zhang CW, Zhang T, Yu P. CoF-CoT: Enhancing large language models with coarse-to-fine chain-of-thought prompting for multi-domain NLU tasks. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 12109–12119. [doi: 10.18653/v1/2023.emnlp-main.743]

[49] Li GZ, Wang P, Ke WJ. Revisiting large language models as zero-shot relation extractors. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: ACL, 2023. 6877–6892.

[50] Li B, Fang GX, Yang Y, Wang QS, Ye W, Zhao W, Zhang SK. Evaluating ChatGPT's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. arXiv:2304.11633, 2023.

[51] Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV. Finetuned language models are zero-shot learners. arXiv:2109.01652, 2022.

[52] Wang X, Zhou WK, Zu C, Xia H, Chen TZ, Zhang YS, Zheng R, Ye JJ, Zhang Q, Gui T, Kang JH, Yang JS, Li SY, Du CS. InstructUIE: Multi-task instruction tuning for unified information extraction. arXiv:2304.08085, 2023.

[53] Wang YZ, Kordi Y, Mishra S, Liu A, Smith NA, Khashabi D, Hajishirzi H. Self-instruct: Aligning language models with self-generated instructions. In: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics, Vol. 1 (Long Papers). Toronto: ACL, 2023. 13484–13508. [doi: 10.18653/v1/2023.acl-long.754]

[54] Bershon BL. Cooperative problem solving: A link to inner speech. In: Interaction in Cooperative Groups. The Theoretical Anatomy of Group Learning. Cambridge: Cambridge University Press, 1992. 36–48.

[55] Alderson-Day B, Fernyhough C. Inner speech: Development, cognitive functions, phenomenology, and neurobiology. Psychological Bulletin, 2015, 141(5): 931–965. [doi: 10.1037/bul0000021]

[56] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Lowe R. Training language models to follow instructions with human feedback. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 27730–27744.

[57] Zhao WT, Chiu J, Cardie C, Rush A. Abductive commonsense reasoning exploiting mutually exclusive explanations. In: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics, Vol. 1 (Long Papers). Toronto: ACL, 2023. 14883–14896. [doi: 10.18653/v1/2023.acl-long.831]

[58] Ma YB, Cao YX, Hong Y, Sun AX. 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. Singapore: ACL, 2023. 10572–10601. [doi: 10.18653/v1/2023.findings-emnlp.710]

[59] Vossen P, Ilievski F, Postma M, Segers R. Don't annotate, but validate: A data-to-text method for capturing event data. In: Proc. of the 11th Int'l Conf. on Language Resources and Evaluation. Miyazaki: European Language Resources Association (ELRA), 2018. 3034–3042.

[60] Tracey J, Bies A, Getman J, Griffitt K, Strassel S. A study in contradiction: Data and annotation for AIDA focusing on informational conflict in Russia-Ukraine relations. In: Proc. of the 13th Language Resources and Evaluation Conf. Marseille: European Language Resources Association, 2022. 1831–1838.

[61] Mitamura T, Liu ZZ, Hovy EH. Overview of TAC-KBP 2016 event nugget track. In: Proc. of the 2016 Text Analysis Conf. 2016. https://hunterhector.github.io/files/papers/Mitamura,_Liu,_Hovy_-_2017_-_TAC_2016.pdf

[62] Mitamura T, Liu ZZ, Hovy E. Events detection, coreference and sequencing: What's next? Overview of the TAC KBP 2017 event track. In: Proc. of the 2017 Text Analysis Conf. 2017. https://hunterhector.github.io/files/papers/Mitamura,_Liu,_Hovy_-_2018_-_TAC_2017.pdf

[63] Yao Y, Li ZC, Zhao H. Learning event-aware measures for event coreference resolution. In: Findings of the Association for Computational Linguistics: ACL 2023. Toronto: ACL, 2023. 13542–13556. [doi: 10.18653/v1/2023.findings-acl.855]

[64] Hu EJ, Shen YL, Wallis P, Allen-Zhu Z, Li YZ, Wang SA, Wang L, Chen WZ. LoRA: Low-rank adaptation of large language models. arXiv:2106.09685, 2021.

[65] Gholami A, Kim S, Dong Z, Yao ZW, Mahoney MW, Keutzer K. A survey of quantization methods for efficient neural network inference. arXiv:2103.13630, 2021.

[66] Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient finetuning of quantized LLMs. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 10088–10115.

[67] Vilain M, Burger J, Aberdeen J, Connolly D, Hirschman L. A model-theoretic coreference scoring scheme. In: Proc. of the 6th Conf. on Message Understanding Conf. 1995. 45–52.

[68] Bagga A, Baldwin B. Algorithms for scoring coreference chains. In: Proc. of the 1st Int'l Conf. on Language Resources and Evaluation Workshop on Linguistics Coreference. 1998. 563–566.

[69] Luo XQ. On coreference resolution performance metrics. In: Proc. of the Conf. on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver: ACM, 2005. 25–32. [doi: 10.3115/1220575.1220579]

[70] Recasens M, Hovy E. BLANC: Implementing the rand index for coreference evaluation. Natural Language Engineering, 2011, 17(4): 485–510. [doi: 10.1017/S135132491000029X]

[71] Lu J, Ng V. Event coreference resolution with non-local information. In: Proc. of the 1st Conf. of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th Int'l Joint Conf. on Natural Language Processing. Suzhou: ACL, 2020. 653–663. [doi: 10.18653/v1/2020.aacl-main.66]

[72] Zhang WZ, Wiseman S, Stratos K. Seq2Seq is all you need for coreference resolution. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 11493–11504. [doi: 10.18653/v1/2023.emnlp-main.704]

[73] Zhou CT, Liu PF, Xu PX, Iyer S, Sun J, Mao YN, Ma XZ, Efrat A, Yu P, Yu LL, Zhang SS, Ghosh G, Lewis M, Zettlemoyer L, Levy O. LIMA: Less is more for alignment. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 55006–55021.

附中文参考文献:

[18] 秦洪武, 崔蓉. 事件共指与话语连贯. 当代语言学, 2009, 11(1): 10–20.

徐昇(1994－), 男, 博士生, 主要研究领域为自然语言处理, 信息抽取.

朱巧明(1963－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为中文信息处理, Web 信息处理.

李培峰(1971－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为自然语言处理, 机器学习.