doi:10.16112/j.cnki.53-1223/n.2026.01.202502160001
引用格式:杨家全,苏适,冯勇,等. ERASum:基于实体关系感知的摘要生成方法[J]. 昆明理工大学学报(自然科学版),2026,51(1):102-111.
Citation: YANG Jiaquan, SU Shi, FENG Yong, et al. ERASum: Entity Relationship-Aware Summarization Method[J]. Journal of Kunming University of Science and Technology (Natural Science), 2026, 51(1): 102-111.

ERASum:基于实体关系感知的摘要生成方法

杨家全[1],苏　适[1],冯　勇[1],和学豪[1],马九顺[2*]
(1. 云南电网有限责任公司 电力科学研究院,云南 昆明 650217; 2. 昆明理工大学 信息工程与自动化学院,云南 昆明 650500)

摘要:为了提升大语言模型在文本摘要任务中的整体性能,特别是其对实体关系的理解与建模能力,以更有效地的方法生成覆盖关键信息的高质量摘要.为此,提出了一种基于实体关系感知的摘要生成方法 ERASum,结合显式与隐式的实体关系信息,引导模型更深入地理解实体之间的语义联系.在元素感知的 CNN/DailyMail 与 BBCXSum 数据集上进行的实验显示,ERASum 在 ROUGE-L 指标上分别比当前最优方法提升了 1.52 和 1.84,显著优于现有模型.结果表明,该方法能够有效增强摘要中实体关系的表达能力,为复杂语义建模提供了新的思路.

关键词:文本摘要;实体关系感知摘要生成方法(ERASum);大语言模型;思维链提示;实体关系
中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1007-855X(2026)01-0102-10

ERASum: Entity Relationship-Aware Summarization Method

YANG Jiaquan[1], SU Shi[1], FENG Yong[1], HE Xuehao[1], MA Jiushun[2*]
(1. Electric Power Research Institute, Yunnan Power Grid Co. Ltd, Kunming 650217, China;
2. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: This study aims to enhance the overall performance of large language models (LLMs) in text summarization tasks, with a particular focus on improving their ability to understand and model relationships between entities. To achieve this goal, we propose ERASum, an Entity Relationship-Aware Summarization approach that incorporates both explicit and implicit entity relationship information to guide LLMs in capturing inter-entity semantics more effectively. Experiments conducted on the element-aware CNN/DailyMail and BBCXSum datasets demonstrate that ERASum achieves ROUGE-L improvements of 1.52 and 1.84 over the current state-of-the-art methods, respectively. These results indicate that ERASum significantly enhances the representation of entity relationships in summaries and offers a promising direction for modeling complex semantic structures.

Keywords: text summarization; Entity-Relationship-Aware Summary method (ERASum); large language models; chain-of-thought prompting; entity relations

0 引言

文本摘要技术旨在生成自然语言摘要,以压缩长文本并提取关键信息.当前,基于神经网络的端到端

收稿日期:2025-02-16. **基金项目**:国家自然科学基金项目(U23A20388,U21B2027);云南省重点研发计划项目(202303AP140008,202302AD080003);云南省基础研究项目(202301AT070393);昆明理工大学"双一流"科技专项(202402AG050007).
作者简介:杨家全(1978-),男,硕士研究生.主要研究方向:新型电力系统规划与运行控制,智能配用电,文本摘要.
E-mail: yjquan99@163.com
通信作者:马九顺(2001-),男,博士研究生.主要研究方向:自然语言处理,文本摘要. **E-mail**: jsma@stu.kust.edu.cn

方法[1]通过对源文档进行编码并解码生成摘要,在该任务上取得了显著进展. 近年来,随着预训练语言模型的快速发展,摘要任务逐步从传统的序列到序列监督学习,转向基于预训练模型的微调与迁移学习,以提升摘要质量与泛化能力.

现有研究通常在大规模语料库上训练或微调语言模型,依赖于大规模的摘要数据集来进行模型训练. Narayan 等[2]训练了一个变换器模型,用于基于实体链和输入文本生成实体链与摘要. Liu 等[3]提出在生成模型中引入对比学习任务,以增强模型对不同质量摘要的排序能力. 此外,Liu 等[4]建议使用 GPT 生成训练数据,并通过对比学习引导下游模型以适应大模型的摘要能力. 然而,现有的合成摘要数据集包含噪声,如信息冗余和事实性幻觉[5-6],严重影响模型在摘要任务上的性能[7]. 除此之外,大量实验表明,这些标准数据集中的参考摘要在人类评估维度上的表现较差,特别是在连贯性、一致性和相关性方面[8].

近年来,大语言模型(LLMs)在自然语言处理任务中取得了突破性的进展. 通过在海量文本数据上的预训练,这些模型获得了丰富的语言和世界知识,从而具备了强大的泛化能力,无需针对特定任务进行微调[9-11]. Wang 等[8]提出的摘要思维链(SumCoT)方法通过从新闻文本中提取四个关键要素(实体、日期、事件和结果)来改进摘要生成. 具体而言,该方法首先识别这些核心要素,随后将其与原文共同输入语言模型,从而生成更具要素感知能力的摘要. 实验表明,SumCoT 不仅在微调预训练模型上表现优异,在零样本大型语言模型中也超越了现有最优方法. 为支持该方法评估,研究者还构建了新的要素感知测试集,该测试集通过客观衡量摘要对原文细粒度要素的覆盖程度,有效地解决了传统人工评估中存在的一致性问题. 然而,该方法未考虑核心元素间的依赖关系[12],且在涉及多实体的摘要任务中表现欠佳,主要受限于隐性关系的建模能力[13].

为此,本文提出一种基于实体关系感知的摘要生成方法(Entity Relationship - Aware Summarization, ERASum),以优化复杂实体场景下的摘要生成性能. 首先,ERASum 方法提取文档中涉及的所有实体. 其次,提取实体之间的显性关系(即文档中直接提到的关系). 再次,基于这些显性关系和文档中的隐性信息,推断实体之间的隐性关系. 模型根据这些隐性关系的可靠性进行评分,设定阈值以确定这些关系的可靠性,并排除低于阈值的关系. 最后,基于提取的实体和关系生成实体关系感知摘要. 为验证该方法的性能,实验在元素感知数据集 CNN/DailyMail 和 BBC XSum 上进行了广泛测试. 结果表明,所提出的方法相比现有最优方法,ROUGE - L 得分分别提升了 1.52 和 1.84.

1 摘要生成方法

1.1 基于预训练模型微调的摘要生成方法

抽象摘要方法通常使用深度学习模型,如循环神经网络(RNN)和变换器(Transformer),通过解读和重新表述源文本来生成摘要. 例如,Nallapati 等[14]利用 RNN 捕捉源文本中句子和词语之间的层次结构. Khandelwal 等[15]使用预训练的变换器解码器生成摘要. 此外,Huang 等[16]使用变换器构建文章元素的异构图,然后使用该异构图影响文本解码器生成简洁流畅的摘要. 尽管抽象摘要方法可以生成更接近人类语言风格的摘要,但它们可能在生成的摘要中出现信息遗漏和不一致的情况[17-19].

另外,基于预训练模型的自然语言处理技术取得了显著进展. 这些模型通常在大规模语料库上进行预训练,然后针对特定任务进行微调. 预训练语言模型,如 BERT[20-21]、GPT - 3[22]和 T5[23],在各种自然语言处理任务中表现出色,包括文本摘要生成. BERT 是一种双向深度学习模型,在庞大的文本语料库上进行预训练,采用掩蔽语言模型(MLM)和下一句预测(NSP)等任务. 当应用于文本摘要任务时,BERT 可以通过添加任务特定层进行微调,例如,将其用作编码器并与解码器结构结合生成摘要[24]. T5 模型通过将所有文本处理任务统一为文本到文本的格式,大大简化了任务设计和模型架构[23]. 在预训练阶段,T5 利用多种任务,包括完形填空、翻译和摘要生成. 对于文本摘要任务的微调,T5 将输入文本作为输入序列,摘要作为目标输出序列,从而使模型能够在统一框架内处理多种任务. Raffel 等[23]证明了这一统一框架在文本摘要任务中取得了最先进的成果. GPT - 3 是一个基于变换器的大规模生成模型,拥有 1 750 亿个参数. 与 BERT 不同,GPT - 3 主要依靠自回归文本生成. Liu 等[22]证明,GPT - 3 在少量学习的情况下,甚至在没有微调的情况下,也能在文本摘要任务中表现出色. 已有一些研究探索了将预训练模型与创新的微调技术结

合,以提高文本摘要的质量. 例如,Pegasus[25]是一种为文本生成优化的预训练模型,通过设计新的预训练任务,更好地捕捉摘要任务的特征. Pegasus 在多个文本摘要数据集上取得了优秀的结果.

1.2 基于思维链的大模型摘要生成

近年来,提升大语言模型(LLMs)推理能力和可解释性的研究,重点关注通过思维链技术将多步骤问题分解为中间阶段[26-27]. 例如,Wei 等[28]的研究表明,将复杂任务拆分为更小、更易于管理的步骤,可以显著提高 LLMs 的性能和透明度. 在自动摘要领域,Wang 等[8]提出了一种创新的摘要生成方法,利用抽象思维链引导大语言模型(LLMs)从源文档中提取更细粒度的元素,从而改善摘要过程,提高摘要的质量和精准度. 这一方法不仅提升了生成摘要的效果,还增强了摘要的连贯性,使其更准确、更细致地反映原文内容. 通过将信息分解成更小、更易消化的部分,这种方法促进了对材料的更深刻理解,从而能够合成出更具洞察力和更有意义的摘要.

在一种互补的方法中,Li 等[29]开发了一种生成局部摘要的策略,结合了滑动窗口技术和自一致性度量. 这一创新方法通过创建更小、更集中的摘要,捕捉文本中定义段落内的关键信息. 随后,这些局部摘要通过聚类过程和多数投票算法被聚合成一个全面的全局摘要,从而确保最终输出忠实于源材料. 通过将局部见解与强有力的聚合机制结合,Li 等的技术解决了摘要保真度和连贯性中的常见挑战,最终提供了更丰富、更可靠的摘要体验. 综上所述,这些研究展示了摘要方法的重大进展,强调了抽象思维和局部语境理解在生成高质量摘要中的重要性.

2 模型结构

如图 1 所示,标准摘要生成通常直接将源文档 S 与任务提示 Q 输入大语言模型,并以自回归方式生成摘要序列. 该过程主要依赖文本表层线索,难以显式建模文档中实体及其相互关系,从而可能导致关键信息遗漏或事实关联错误. 为此,本文提出 ERASum,其五步推理式摘要生成流程如图 2 所示:1)抽取实体及实体类型(Step 1);2)抽取文档中显式关系(Step 2);3)在显式关系基础上推断隐式关系(Step 3);4)通过评分代理筛选高置信隐式关系(Step 4);5)将实体与关系信息结构化注入提示以生成摘要(Step 5). 下文将依次对各步骤进行详细说明.

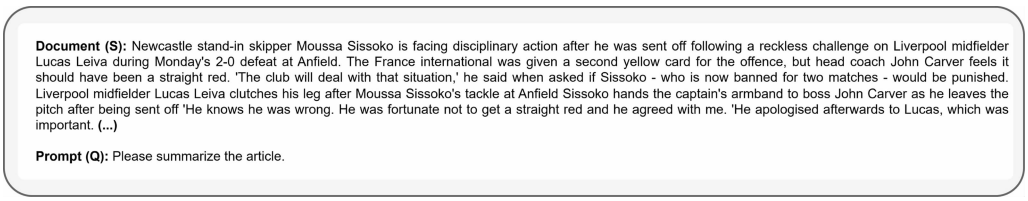

图 1 摘要的标准生成过程

Fig. 1 Standard summarization process

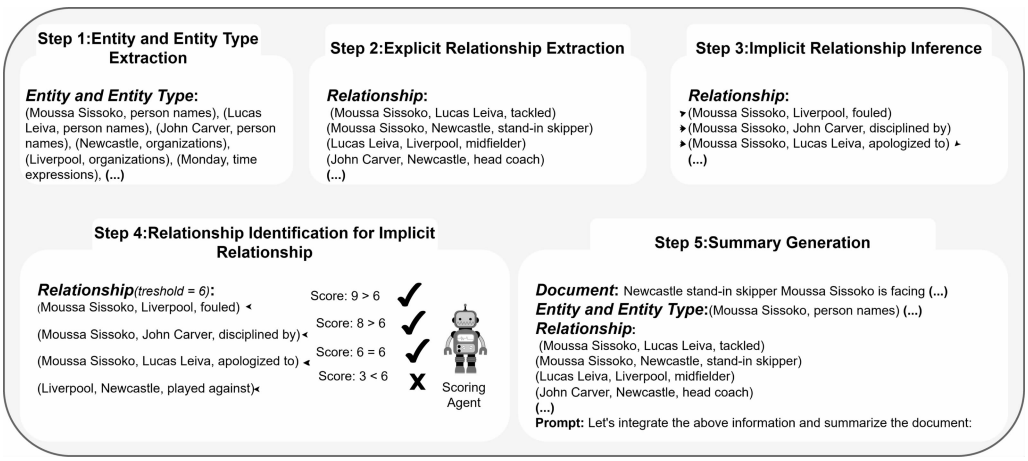

图 2 ERASum 的五步推理式摘要生成过程

Fig. 2 ERASum's five – step reasoning pipeline for summarization

2.1 问题定义

如图 1 所示,给定源文档 $S=[x_1,\cdots,x_N]$与任务提示 Q, LLMs 以自回归方式生成摘要序列 $Y=[y_1,\cdots,y_T]$,其条件生成概率可分解为:

$$P_\theta(Y|S,Q)=\prod_{t=1}^{T}P_\theta(y_t|S,Q,y_{<t}),y_t\in V \tag{1}$$

其中:θ 为模型参数,V 为词表,$y_{<t}=[y_1,\cdots,y_{t-1}]$表示已生成的前缀. ERASum 的核心思想是在生成前通过多步推理得到实体与实体关系信息,并将其结构化地注入到提示中,从而提高摘要的一致性与信息覆盖度.

2.2 实体及实体类型提取

利用 LLM 的信息抽取能力[30],本文首先从源文档 S 中识别实体及其类型,得到实体集合 $\varepsilon=\{e_i\}_{i=1}^{M}$,其中 $e_i=(m_i,t_i)$. m_i 为第 i 个实体在文档中的文本跨度,$t_i\in T$ 为其实体类型,T 为预定义类型集合. 给定提示 Q_1(例如“识别并列出文档中的所有实体及其类型……”),实体抽取步骤形式化为:

$$\varepsilon=LLM(S,Q_1) \tag{2}$$

为提高命名实体识别的可靠性,本文采用自一致性(Self-Consistency,SC)策略[31]:对同一文档在相同提示下独立采样 K 次抽取结果,并对候选实体进行多数投票;当某实体在 K 次结果中被识别为有效实体的次数不少于$[K/2]$时,保留该实体. 该策略用于降低偶发抽取错误带来的噪声.

2.3 显性关系提取

在摘要生成中,实体关系至关重要[32],因为它们有助于刻画文档事实结构并提升摘要的准确性与连贯性. 基于已抽取的实体集合 ε,本文进一步抽取文档中明确陈述(显式)的实体关系,关系以三元组形式表示为(e_i,r_{ij},e_j),其中 $e_i,e_j\in\varepsilon$,r_{ij}为两实体之间的关系类型(或关系描述字符串). 给定提示 Q_2(例如“提取文档中明确陈述的实体关系三元组……”),显式关系抽取形式化为:

$$R^{\exp}=LLM(S,Q_2,\varepsilon) \tag{3}$$

显式关系集合可写为:

$$R^{\exp}=\{(e_i,r_{ij},e_j)\,|\,e_i,e_j\in\varepsilon\} \tag{4}$$

上述步骤将文档中的显式事实连接转化为结构化三元组,为后续隐式关系推断提供可推理的图结构基础.

2.4 隐性关系推断

尽管显式关系抽取能够获得部分清晰可见的实体关系,但在多实体场景下仍存在大量需要背景推理的隐式关系[33]. 因此,本文在 $R^{\exp}$的基础上,对实体对进行关系推断以生成隐式关系候选. 直观地,隐式关系可由显式关系构成的实体链(路径)支持:若存在实体链 $e_{v_1}\to e_{v_2}\to\cdots\to e_{v_L}$,且链上边均来自 $R^{\exp}$,则可据此推断端点实体 e_{v_1}与 e_{v_L}之间可能存在未显式陈述的关系. 给定提示 Q_3(例如“根据已抽取的显式关系,推断实体之间所有可能的隐式关系……”),隐式关系推断得到候选集合:

$$\hat{R}^{\text{imp}}=LLM(S,Q_3,\varepsilon,R^{\exp}) \tag{5}$$

其中$\hat{R}^{\text{imp}}$可能包含同一实体对的多条候选关系,可写为:

$$\hat{R}^{\text{imp}}=\{(e_i,\hat{r}_{ij}^{(k)},e_j)\,|\,e_i,e_j\in\varepsilon,k=1,\cdots,K_{ij}\} \tag{6}$$

这里$\hat{r}_{ij}^{(k)}$表示第 k 条隐式关系候选,K_{ij}为对(e_i,e_j)生成的候选数量.

2.5 隐式关系的识别

隐式关系推断会为实体对生成多个候选关系,其中可能包含错误或无关关系. 为此,本文引入隐式关系识别步骤,将 LLM 作为评分代理对候选三元组进行置信度评估,并据此过滤噪声关系. 具体地,对每个候选三元组$(e_i,\hat{r}_{ij}^{(k)},e_j)$,评分代理输出一个置信度分数 $s_{ij}^{(k)}\in[0,10]$,分数越高表示该关系越可能正确. 给定提示 Q_4(例如“对每个关系进行评分(0-10)…”),评分过程形式化为:

$$s_{ij}^{(k)}=LLMScore(S,Q_4,(e_i,\hat{r}_{ij}^{(k)},e_j)) \tag{7}$$

设阈值为 τ_{th},则筛选后的隐式关系集合定义为:

$$R^{\text{imp}} = \{(e_i, \hat{r}_{ij}^{(k)}, e_j) \in \hat{R}^{\text{imp}} \mid s_{ij}^{(k)} \geqslant \tau_{\text{th}}\} \tag{8}$$

该步骤用于降低隐式推断引入的错误关系,从而提升后续摘要生成的一致性与准确性.

2.6 摘要生成

在摘要生成阶段,本文将实体集合与关系三元组进行结构化序列化(serialization),并与原始上下文共同组成最终提示输入. 令增强信息 $Z = [\varepsilon, R^{\text{exp}}, R^{\text{imp}}]$,给定提示 Q_5(例如"让我们整合上述信息并对文档进行总结:……"),摘要生成可形式化为条件序列生成:

$$P_\theta(Y \mid S, Q_5, Z) = \prod_{t=1}^{T} P_\theta(y_t \mid S, Q_5, Z, y_{<t}) \tag{9}$$

最终输出摘要 Y 由所采用的解码策略(如贪心或束搜索)从上述分布中生成. 该提示构建过程将实体、显式关系与高置信隐式关系共同注入模型输入,从而减少关键信息溃漏并提升摘要的逻辑连贯性.

3 实 验

本节在两个元素感知数据集上开展了系统性实验,全面评估 ERASum 方法的性能表现. 通过将 ERASum 与多种最先进方法进行基准对比,该方法在不同场景下的优势与改进潜力得以明确呈现. 进一步地,本文深入分析了 ERASum 中的每个模块,考察它们在提升整体性能方面的贡献和效果. 这一多维度的评估不仅验证了方法的有效性,而且客观地揭示了其技术边界与应用局限.

3.1 数据集

实验在新闻领域的元素感知测试数据集 CNN/DailyMail 和 BBCXSum[8] 上评估了所提方法的性能,这些数据集在长度和抽象性上具有代表性. 两个数据集都要求三位新闻专家独立为 200 个随机抽样的源文档编写专业摘要,基于四个核心元素(实体、日期、事件和结果),确保写作风格的全面性、客观性和一致性. 此外,由一位专家编写摘要以确保写作风格的一致性,而其他两位专家则根据流利性、一致性、连贯性和相关性四个维度对完成的摘要进行评判. 当出现评分分歧时,专家团队通过协商讨论达成一致意见,直至所有摘要均满足四个维度的评审标准.

3.2 基座模型

BART-BASE 是 BART(双向自回归变换器)[34]的一个较小变种,以其在摘要和翻译任务中的优异表现而著称. BART-LARGE 是 BART 的更大版本,在摘要任务中表现出色. T5-LARGE 是 T5 文本到文本模型家族的一部分[23],该家族将所有 NLP 任务转化为文本生成任务. PEGASUS-LARGE 是专门为抽象摘要任务设计的模型[25]. GPT-3(175B)[35-36]是一个大规模语言模型,包含 1 750 亿个参数. 它经过广泛的 NLP 任务微调,成为生成任务(如摘要生成)的强大基准. SumCoT[8]从源文档中提取细粒度元素,以引导 LLMs(GPT-3(175B))生成摘要.

3.3 评估指标

摘要生成质量的评估采用词汇重叠度度量方法,具体包括 ROUGE-1/2/L[37]和嵌入相似度 BERTSCORE[38]度量. ROUGE 指标包括 ROUGE-1、ROUGE-2 和 ROUGE-L,分别衡量生成摘要与参考摘要在 unigram、bigram 和最长公共子序列(LCS)上的重叠度. BERTSCORE 通过计算候选句子中每个词汇与参考句子中每个词汇之间的相似度分数来衡量生成文本的质量. *ROUGE* 的计算方法:

$$ROUGE\text{-}N = \frac{\sum_{S \in \{Refsum\}} \sum_{n\text{-}gram \in S} Count_{\text{match}}(n\text{-}gram)}{\sum_{S \in \{Refsum\}} \sum_{n\text{-}gram \in S} Count(n\text{-}gram)} \tag{10}$$

其中:*Refsum* 表示人工参考摘要集合,$S \in \{Refsum\}$ 表示遍历每一条参考摘要;$n\text{-}gram \in S$ 表示参考摘要 S 中包含的所有长度为 n 的连续词序列;$Count(n\text{-}gram)$ 表示该 $n\text{-}gram$ 在参考摘要中出现的次数;$Count_{\text{match}}(n\text{-}gram)$ 表示该 $n\text{-}gram$ 同时出现在系统生成摘要与参考摘要中的匹配次数(通常取两者出现次数的最小值).

3.4 实验细节

本文采用 GPT－3 的 text－davinci－002 版本[35－36]作为基础模型. 该模型具有 1 750 亿个参数,属于当前规模最大的预训练语言模型之一. 它在多种自然语言处理任务上的表现异常出色,包括文本生成、阅读理解、翻译和摘要生成,展现了强大的泛化能力和在不同应用场景中的多样性. 实验通过 OpenAI 官方 API 实现模型调用,该接口提供多配置选项支持,可灵活调整输入输出参数以满足实验需求. 为保障结果的可重复性,文本严格采用 API 默认参数设置,避免引入额外的超参数调整,从而确保实验结果的可靠性和模型能力的准确评估.

3.5 主要结果

实验结果表明,相较于其他微调模型的预训练语言模型,GPT－3 在元素感知测试集上的表现有了显著提升. 特别是在 BBC XSum 数据集上,GPT－3 的表现明显高于其他微调的预训练语言模型. 这表明 GPT－3强大的语言理解能力在生成更精细的摘要方面具有更强的潜力.

与表 1 中展示的最先进方法(SOTA)相比,基于GPT－3的元素关系感知方法(175B GPT－3(元素关系感知))在所有评估指标上都取得了显著的提升,其中在 CNN/DailyMail 测试集上,ROUGE－1、ROUGE－2 和 ROUGE－L 分数分别提高了 3.36、0.99 和 1.52. 在 BBC XSum 上,分数分别提高了 2.74、1.25 和 1.84. 这些结果表明,从源文档中提取精细化元素及其事实关系并进行排序,使得 LLMs 在生成摘要时具有更精细的能力,使生成的文本摘要更符合人类标准.

表 1　CNN/DailyMail 和 BBC XSum 数据集的 ROUGE(%)和 BERTScore(%)的 F1 分数

Tab. 1　F1 scores of ROUGE (%) and BERTScore (%) in CNN/DailyMail and BBC XSum datasets

Model	CNN/DailyMail				BBC XSum			
	ROUGE－1	ROUGE－2	ROUGE－L	BERTScore	ROUGE－1	ROUGE－2	ROUGE－L	BERTScore
BART－BASE	36.06	15.93	33.09	0.876 2	21.89	5.13	17.19	0.866 3
BART－LARGE	37.98	18.16	34.30	0.886 0	23.79	5.02	17.93	0.871 0
T5－LARGE	37.47	17.66	34.34	0.876 8	24.98	6.89	19.46	0.872 8
PEGASUS－LARGE	36.65	17.58	33.84	0.871 0	21.35	4.87	19.03	0.866 2
175B GPT－3	37.75	15.20	34.25	0.890 5	31.74	10.95	25.42	0.893 3
175B GPT－3(SumCoT)	43.03	19.51	38.67	0.902 3	35.70	15.31	30.19	0.901 8
175B GPT－3(ERASum)	**46.39**(↑ **3.36**)	**20.50**(↑ **0.99**)	**40.19**(↑ **1.52**)	**0.930 3**(↑ **0.028**)	**38.44**(↑ **2.74**)	**16.56**(↑ **1.25**)	**32.03**(↑ **1.84**)	**0.903 1**(↑ **0.013**)

3.6 人工评分

由于 ROUGE 无法准确且全面地表示摘要质量,实验引入了人工评估. 从 CNN/DailyMail(Element－aware)数据集中随机抽取 100 个英文摘要,由五名英语研究生独立评估. 每个摘要在流畅性(Fluency)、连贯性(Coherence)、一致性(Consistency)与相关性(Relevance)四个维度进行评分:流畅性(Fluency)衡量语法与可读性;连贯性(Coherence)衡量句间衔接与逻辑结构;一致性(Consistency)衡量摘要事实是否与原文一致;相关性(Relevance)衡量摘要是否覆盖原文关键信息并与主题一致. 与原始 175B GPT－3、175B GPT－3(SumCoT)和 175B GPT－3(ERASum)模型相比,该方法的实验结果如表 2 所示. ERASum 模型在所有评估指标上都表现出进一步的改进,尤其在流畅性(－0.09)和连贯性(－0.05)方面表现突出. 这表明该模型能够生成结构更连贯、语言更流畅的摘要. 此外,其一致性得分(－0.25)显著优于其他模型.

表2 CNN/DailyMail 与 BBC XSum(Element－aware)数据集的人工评价结果(Flu/Coh/Con/Rel)

Tab. 2 Human evaluation results (Flu/Coh/Con/Rel) on CNN/DailyMail and BBC XSum (Element－aware) dataset

Model	CNN/DailyMail (Element－aware)				BBC XSum (Element－aware)			
	Flu	Coh	Con	Rel	Flu	Coh	Con	Rel
175B GPT－3	－0.18	－0.30	－0.39	－0.72	－0.18	－0.47	－0.33	－0.56
175B GPT－3 (SumCoT)	－0.13	－0.08	－0.26	－0.28	－0.19	－0.23	－0.09	－0.25
175B GPT－3 (ERASum)	－0.09	－0.05	－0.21	－0.25	－0.13	－0.15	－0.07	－0.22

3.7 消融实验

如表3所示,为验证元素间关系的必要性,研究设计了消融实验.表中展示了在CNN/DailyMail(Element－aware)和BBC XSum(Element－aware)数据集上,使用不同技术生成的摘要的ROUGE和BERTScore F1分数.结果表明,175B GPT－3(ERASum)模型在这两个数据集上都取得了最高的ROUGE和BERTScore.这一发现证实,从源文档中提取细粒度元素并考虑它们的事实关系,能够有效地提升生成摘要的质量.更深入的分析揭示了该过程中不同步骤的作用:step 1是实体提取,step 2是显式关系提取,step 3是隐式关系推理,step 4是关系识别.175B GPT－3(ERASum)模型整合了这些步骤,取得了最佳的结果.

在CNN/DailyMail(Element－aware)数据集上,ERASum模型相比省略不同步骤的模型,ROUGE－1、ROUGE－2和ROUGE－L分别提高了3.36、0.99和1.52,同时BERTScore提高了0.028.同样,在BBC XSum(Element－aware)数据集上,ERASum模型在所有指标上都表现出显著的提升,特别是ROUGE－1和ROUGE－L分别提高了2.74和1.84.这些结果突显了完整ERASum过程的重要性,证明了考虑元素及其关系对于提高模型摘要生成质量至关重要.

表3 CNN/Daily Mail 和 BBC XSum 数据集的 ROUGE(%)和 BERTScore(%)的 F1 分数

Tab. 3 F1 scores of ROUGE(%) and BERTScore(%) in CNN/Daily Mail and BBC XSum dataset

Model	CNN/DailyMail (Element－aware)				BBC XSum (Element－aware)			
	ROUGE－1	ROUGE－2	ROUGE－L	BERTScore	ROUGE－1	ROUGE－2	ROUGE－L	BERTScore
175BGPT－3	37.75	15.20	34.25	0.890 5	31.74	10.95	25.42	0.893 3
175BGPT－3 (SumCoT)	43.03	19.51	38.67	0.902 3	35.70	15.31	30.19	0.901 8
175BGPT－3 (ERASum)	**46.39** (↑ **3.36**)	**20.50** (↑ **0.99**)	**40.19** (↑ **1.52**)	**0.930 3** (↑ **0.028**)	**38.44** (↑ **2.74**)	**16.56** (↑ **1.25**)	**32.03** (↑ **1.84**)	**0.903 1** (↑**0.013**)

3.8 案例分析

通过案例能直观地比较基于175B GPT－3的不同方法在CNN/DailyMail(Element－aware)测试集上生成摘要的质量.如图3所示,与其他方法相比175B GPT－3(ERASum)在生成摘要方面表现出了显著的优势.它不仅捕捉了所有细粒度的元素(如人物、日期、事件和结果),还准确地反映了这些元素之间的显式和隐式关系.显式关系指的是清晰、易于识别的事实连接,例如替补队长与中场球员在比赛中的合作,或者主教练指令对比赛结果的直接影响.隐式关系则涉及更为复杂的背景推理,例如先前的战术安排如何影响当前的决策,或者一名关键球员的情绪状态如何间接影响比赛的进程.

对比分析显示,175B GPT－3和175B GPT－3(SumCoT)都遗漏了关键细节(例如“Sissoko”获得了第二张黄牌),并且引入了不必要的内容,使摘要显得更加冗长且不准确.这一结果表明,ERASum这样考虑元素和关系的方法,能够生成更全面和精确的摘要,捕捉事实之间的显式和隐式联系.

Article: Newcastle ***stand-in skipper*** Moussa Sissoko is facing disciplinary action after he was sent off following a reckless challenge on Liverpool ***midfielder*** Lucas Leiva during Monday's 2-0 defeat at Anfield. The France international was given a second yellow card for the offence, but ***head coach*** John Carver feels it should have been a straight red. 'The club will deal with that situation,' he said when asked if Sissoko - who is now banned for two matches - would be punished. Liverpool midfielder Lucas Leiva clutches his leg after Moussa Sissoko's tackle at Anfield Sissoko hands the captain's armband to boss John Carver as he leaves the pitch after being sent off 'He knows he was wrong ...

Gold Summary: Newcastle ***stand-in skipper*** Moussa Sissoko was sent off for a reckless challenge on Liverpool ***midfielder*** Lucas Leiva on Monday's match. He was given a second yellow card but ***head coach*** John Carver thought it deserves a red card. Afterwards, Sissoko apologized to Lucas and he was punished by the club.

175B GPT-3: Newcastle's Moussa Sissoko faces disciplinary action for a reckless foul on Liverpool's Lucas Leiva. ***Manager*** John Carver said the club would deal with the situation and Sissoko apologized to Lucas. Carver also stressed the need for Newcastle to start winning games and not rely on other teams to avoid relegation.

175B GPT-3(SumCoT**):** Newcastle's ***interim captain*** Moussa Sissoko faces disciplinary action after being sent off for a foul against Liverpool at Anfield on Monday night. ***Head coach*** John Carver thought it should have been a straight red card. He said the club would deal with the situation. Carver also pointed out that the team cannot rely on the mistakes of others to maintain its Premier League status, and he hopes that the team can win the next two home games.

175B GPT-3(ERASum**):** Newcastle's ***acting captain*** Moussa Sissoko faces disciplinary action after receiving a second yellow card for a ***midfielder***'s foul on Lucas Leiva against Liverpool on Monday. ***Head coach*** John Carver thought it should have been a straight red card. Sissoko apologized to Lucas and Carver said the club would deal with the situation.

图 3　不同方法生成的摘要示例(加粗斜体表示实体关系)

Fig. 3　Examples of summaries generated by different methods (bold and italics indicate entity relationships)

4 结 论

自动摘要生成的关键在于从源文档中提取与事件相关的元素及其之间的关系,以生成简明扼要的摘要. 大语言模型(LLMs)具有强大的语言理解能力,能够通过提示利用更细粒度的事件信息和事件之间的关系,增强零样本摘要生成. 本文提出的基于实体关系感知的摘要生成方法首先利用大语言模型从文本中提取关键信息元素,如实体、日期、事件和结果. 然后,获取这些事件元素之间的显性和隐性关系,并利用这些信息和关系来引导大语言模型生成更详细的事件摘要.

实验结果表明,在 CNN/DailyMail(Element - aware)和 BBC XSum(Element - aware)数据集上的实验中,采用该方法的 ROUGE - L 评估得分分别比微调预训练语言模型和零样本大语言模型提高了 1.52 和 1.84. 总的来说,使用大语言模型和基于提示的方法可以有效地提升自动摘要,特别是在零样本场景中,元素关系使得生成的摘要能够包含源文档中的关键观点. 未来工作可进一步探索大语言模型在模拟人类写作中的潜力,以推动相关研究发展.

参考文献:

[1] WANG J A, MENG F D, ZHENG D, et al. Towards unifying multi - lingual and cross - lingual summarization[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, CanadaAssociation for Computational Linguistics, 2023: 15127 - 15143.

[2] NARAYAN S, ZHAO Y, MAYNEZ J, et al. Planning with learned entity prompts for abstractive summarization[C]//Transactions of the Association for Computational Linguistics. Cambridge, MAAssociation for Computational Linguistics, 2019, 9: 1475 - 1492.

[3] LIU Y X, LIU P F, RADEV D R, et al. BRIO: Bringing order to abstractive summarization[J]. Association for Computational Linguistics, 2022(1): 2890 - 2903.

[4] LIU Y X, SHI K J, HE K, et al. On learning to summarize with large language models as references[J]. Association for Computational Linguistics, 2024(1): 8647 - 8664.

[5] CHEN Y L, ZHANG H J, ZHOU Y J, et al. Revisiting cross - lingual summarization: A corpus - based study and a new benchmark with improved annotation[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada. Stroudsburg, PA, USA: ACL, 2023: 9332 - 9351.

[6] GAO M Q, WANG W Q, WAN X J, et al. Evaluating factuality in cross - lingual summarization[C]//Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada. Stroudsburg, PA, USA: ACL, 2023: 12415 - 12431.

[7] LIU S L, GAO Y, LI S S, et al. A hallucination detection and mitigation framework for faithful text summarization using LLMs[J]. Scientific Reports, 2026, 16: 1374.

[8] WANG Y M, ZHANG Z S, WANG R. Element - aware summarization with large language models: Expert - aligned evaluation and chain - of - thought method[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada. Stroudsburg, PA, USA: ACL, 2023: 8640 - 8665.

[9] LYU Q, HAVALDAR S, STEIN A, et al. Faithful chain - of - thought reasoning[C]//Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia - Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Nusa Dua, Bali. Stroudsburg, PA, USA: ACL, 2023: 305 - 329.

[10] SUZGUN M, SCALES N, SCHÄRLI N, et al. Challenging BIG - bench tasks and whether chain - of - thought can solve them[C]//Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada. Stroudsburg, PA, USA: ACL, 2023: 13003 - 13051.

[11] WANG B S, MIN S, DENG X, et al. Towards understanding chain - of - thought prompting: An empirical study of what matters[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada. Stroudsburg, PA, USA: ACL, 2023: 2717 - 2739.

[12] TAM D, MASCARENHAS A, ZHANG S Y, et al. Evaluating the factual consistency of large language models through news summarization[C]//Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada. Stroudsburg, PA, USA: ACL, 2023: 5220 - 5255.

[13] LIU Y M, PENG X Y, DU T Y, et al. ERA - CoT: Improving chain - of - thought through entity relationship analysis[J]. Association for Computational Linguistics, 2024(1): 8780 - 8794.

[14] NALLAPATI R, ZHOU B W, DOS SANTOS C, et al. Abstractive text summarization using sequence - to - sequence RNNs and beyond[C]//Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany. Stroudsburg, PA, USA: ACL, 2016: 280 - 290.

[15] KHANDELWAL U, CLARK K, JURAFSKY D, et al. Sample efficient text summarization using a single pre - trained transformer[EB/OL]. 2019: arXiv: 1905.08836. https://arxiv.org/abs/1905.08836.

[16] HUANG Y X, HOU S K, LI G, et al. Abstractive summary of public opinion news based on element graph attention[J]. Information, 2023, 14(2): 97.

[17] FABBRI A R, KRYŚCIŃSKI W, MCCANN B, et al. SummEval: Re - evaluating summarization evaluation[J]. Transactions of the Association for Computational Linguistics, 2021, 9(2): 391 - 409.

[18] MRIDHA M F, LIMA A A, NUR K, et al. A survey of automatic text summarization: Progress, process and challenges[J]. IEEE Access, 2021, 9: 156043 - 156070.

[19] MAYNEZ J, NARAYAN S, BOHNET B, et al. On faithfulness and factuality in abstractive summarization[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. OnlineAssociation for Computational Linguistics, 2020: 1906 - 1919.

[20] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre - training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, MinnesotaAssociation for Computational Linguistics, 2019: 4171 - 4186.

[21] MA T H, PAN Q, RONG H, et al. T - BERTSum: Topic - aware text summarization based on BERT[J]. IEEE Transactions on Computational Social Systems, 2022, 9(3): 879 - 890.

[22] LIU X, ZHENG Y N, DU Z X, et al. GPT understands, too[J]. AI Open, 2024, 5: 208 - 215.

[23] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text - to - text transformer[J]. Journal of Machine Learning Research, 2020, 21(140): 1 - 67.

[24] LIU Y, LAPATA M. Text summarization with pretrained encoders[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP - IJCNLP), Hong Kong, China. Stroudsburg, PA, USA: ACL, 2019: 3728 - 3738.

[25] ZHANG J Q,ZHAO Y,SALEH M,et al. PEGASUS: Pre - training with extracted gap - sentences for abstractive summarization[J]. Arxiv,2019(Proceedings of the 37 th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020):1 - 12. https://arxiv.org/abs/1912.08777.

[26] KHOT T,TRIVEDI H,FINLAYSON M,et al. Decomposed prompting:A modular approach for solving complex tasks[EB/OL]. 2022:arXiv:2210.02406. https://arxiv.org/abs/2210.02406.

[27] WANG L,XU W Y,LAN Y H,et al. Plan - and - solve prompting:Improving zero - shot chain - of - thought reasoning by large language models[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1:Long Papers),Toronto,Canada. Stroudsburg,PA,USA:ACL,2023:2609 - 2634.

[28] WEI J, WANG X, SCHUURMANS D,et al. Chain of thought prompting elicits reasoning in large language models[J]. Arxiv,2022(36th Conference on Neural Information Processing Systems):1 - 14.

[29] LI T J,LI Z,ZHANG Y. Improving faithfulness of large language models in summarization via sliding generation and self - consistency[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics,Language Resources and Evaluation (LREC - COLING 2024). Torino,ItaliaAssociation for Computational Linguistics,2024:8804 - 8817.

[30] ASHOK D,LIPTON Z C. PromptNER:Prompting for named entity recognition[EB/OL]. 2023:arXiv:2305.15444. https://arxiv.org/abs/2305.15444.

[31] WANG X Z,WEI J,SCHUURMANS D,et al. Self - consistency improves chain of thought reasoning in language models[EB/OL]. 2022:arXiv:2203.11171. https://arxiv.org/abs/2203.11171.

[32] 张娜,姜占宇,丁豪,等. 大型网站的实时流量分析技术研究——以 CNKI 网站为例[J]. 昆明理工大学学报(自然科学版),2025,50(2):88 - 97.
ZHANG N,JIANG Z Y,DING H,et al. Real - time traffic analysis techniques for large - scale websites: A case study of the CNKI website[J]. Journal of Kunming University of Science and Technology (Natural Science Edition),2025,50(2):88 - 97.

[33] 刘志坚,陶韵旭,刘航,等. 融合残差密集与生成对抗网络的红外巡检图像超分辨率重建[J]. 昆明理工大学学报(自然科学版),2023,48(5):120 - 129.
LIU Z J,TAO Y X,LIU H,et al. Super - resolution reconstruction of infrared inspection images by integrating residual dense and generative adversarial networks[J]. Journal of Kunming University of Science and Technology (Natural Science),2023,48(5):120 - 129.

[34] LEWIS M,LIU Y H,GOYAL N,et al. BART:Denoising sequence - to - sequence pre - training for natural language generation,translation,and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. OnlineAssociation for Computational Linguistics,2020:7871 - 7880.

[35] BROWN T,MANN B,RYDER N,et al. Language models are few - shot learners[J]. Advances in Neural Information Processing Systems,2020,33:1877 - 1901.

[36] OUYANG L,WU J,JIANG X,et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems,2022,35:27730 - 27744.

[37] LIN C Y. ROUGE:A package for automatic evaluation of summaries[C]//Text Summarization Branches Out. Barcelona,Spain Association for Computational Linguistics,2004:74 - 81.

[38] ZHANG T Y,KISHORE V,WU F,et al. BERTScore:Evaluating text generation with BERT[EB/OL]. 2019:arXiv:1904.09675. https://arxiv.org/abs/1904.09675.

(编　辑:朱银周)