DOI:10.3979/j.issn.1673-825X.202409020228

面向生成式文本摘要模型的内在幻觉优化方法

李　能,于成成,刘　群
(重庆邮电大学 计算智能重庆市重点实验室,重庆 400065)

摘　要:生成式文本摘要模型能够为摘要生成新的表达,即使是最先进的摘要模型也可能产生与原始文本矛盾或无法验证准确性的内容,这种现象被称为“幻觉”。为解决这一问题,提出了一种内在幻觉优化方法,用于改进摘要生成过程。该方法分别从数据层面、模型训练层面和摘要生成策略层面提出了摘要模型幻觉优化方法。在 2 个数据集上的实验验证均取得最佳性能。实验结果表明,对比基线模型,在 CNNDM 数据集上 R-1 得分平均提升 8.58%;在 XSUM 数据集上 R-1 得分平均提升 7.26%。该方法不仅能够提升摘要生成效果,而且有效减少了生成摘要中的幻觉问题,为生成式文本摘要模型落地和应用提供了参考。

关键词:生成式文本摘要;内在幻觉;候选摘要;大语言模型
中图分类号:TP393　　**文献标志码**:A　　**文章编号**:1673-825X(2025)05-0688-08

An intrinsic hallucination optimization method for generative text summarization models

LI Neng, YU Chengcheng, LIU Qun
(Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, P. R. China)

Abstract: Generative text summarization models can produce novel expressions in summaries, but even the most advanced models may generate content that contradicts the source text or lacks factual verifiability—a phenomenon known as hallucination. To address this issue, this paper proposes an intrinsic hallucination optimization method to improve the summarization generation process. The proposed approach mitigates hallucinations from three perspectives: data-level optimization, model training-level optimization, and summary generation strategy-level optimization. Experiments conducted on two benchmark datasets demonstrate the superior performance of the proposed method. Compared with baseline models, the proposed approach achieves an average improvement of 8.58% in R-1 score on the CNNDM dataset and 7.26% on the XSUM dataset. The results indicate that the method not only enhances summary quality but also effectively reduces hallucinations, providing a valuable reference for the practical deployment of generative text summarization models.

Keywords: generative text summarization; intrinsic hallucination; candidate summaries; large language model

收稿日期:2024-09-02　**修订日期**:2025-03-05　**通讯作者**:刘　群　liuqun@cqupt.edu.cn
基金项目:国家自然科学基金重点项目(61936001);重庆市教委重点合作项目(HZ2021008)
Foundation Items: Key Project of National Natural Science Foundation of China (61936001); Key Cooperation Project of Chongqing Municipal Education Commission (HZ2021008)

0 引　言

文本摘要是自然语言处理(natural language processing,NLP)中的一项重要任务,广泛应用于信息检索[1]、新闻媒体摘要[2]、社交媒体评论摘要[3]和学术文章总结[4]等方面。常见的文本摘要方法是抽取式摘要生成方式,其大都依赖于特征工程和监督学习技术,如文本相似性、关键字频率和位置权重。典型的算法包括统计方法,如 TF-IDF[5]、TextRank[6]和 PageRank[7]。

近年来,随着人工智能的蓬勃发展,生成式文本摘要方法表现出了令人印象深刻的性能。这些模型主要基于 Transformer 架构[8],并结合自注意力机制。例如,自 BERT 模型[9]提出后,其在各种下游任务的应用中都取得了出色的结果,展示了对文本中潜在的含义和深度的表示学习、表达的理解能力。基于深度学习的模型不仅能从源文本中提取关键信息,而且还能用自然语言生成新的表达,从而得到更准确、连贯和富有表现力的摘要。因此,生成式文本摘要技术的应用越来越广泛。然而,生成式文本摘要仍然面临一系列挑战。其中,由于长文本中关键信息的分散性和内容的广泛性,以及常用的 Transformer 架构引入的注意力可能会随着文本长度增加而被稀释。因此,在摘要生成过程中经常出现幻觉问题,这对文本摘要的自然流畅性和事实一致性提出了复杂的挑战,使其难以在实践中实施。例如,摘要模型会错误地生成“7 月 1 日发布的提案”,但事实是“7 月 1 日实施”,生成的摘要中出现了幻觉问题。

为了解决这个问题,本文深入研究了生成式文本摘要生成的各个具体过程,并提出了一种模型训练优化方法,用于解决生成式文本摘要模型中的内在幻觉问题。首先,使用预训练的语言模型,通过生成具有质量约束的多个摘要作为候选摘要来增强训练数据。然后,设计了一种改进的摘要生成过程训练策略,该策略可以让模型自行评估生成摘要的质量并调整训练参数。最后,在摘要生成阶段引入标签平滑技术,缓解由曝光偏差引起的幻觉现象,并使模型能够根据更高质量的候选摘要生成最终的摘要,最大限度地减少幻觉内容的产生。

综上所述,本文的主要贡献如下。

1)为了提高训练数据的质量,本文使用预训练模型(bidirectional and auto-regressive transformers,BART)通过质量约束为训练创建高质量的多候选摘要训练数据集。

2)为了解决摘要生成过程中的曝光偏差问题,本文设计了一种新的训练策略,可以在模型训练时评估生成摘要的质量。并在摘要生成过程中引入标签平滑技术,为更准确地候选摘要分配了更高的概率,以鼓励模型学习更好地表示。

3)将本文的方法与多个生成式文本摘要模型相结合后,在 CNNDM 数据集上,其 ROUGE-1 分数平均提高了 8.58%,在 XSUM 数据集上 ROUGE-1 分数平均提高了 7.26%。此外,根据另外的一些定性实验对比结果也能有效证明,本文提出的方法有效地减少了生成摘要中的幻觉问题出现。

1 相关工作

大部分生成式模型被表述为序列到序列(Seq2Seq)模型[9]。当前,Seq2Seq 模型的著名例子包括 GPT-3[10]、ChatGLM[11]和 GPT-4[12]等。基本上所有主流的生成式模型都是基于序列到序列的模型。这一类模型主要是通过不断增加训练数据量和模型参数量的大小,来让它们具有令人印象深刻的能力。近年来,文本摘要的序列到序列建模一直是自然语言处理领域的一个热门研究课题。

其中,生成式摘要模型在学习过程中以自回归的方式生成摘要。如果之前的预测结果不好,将严重影响后续的预测,并导致最终生成结果不佳,进一步会导致幻觉问题的出现。在自然语言处理领域,幻觉问题通常是指模型生成的内容毫无意义或与提供的原内容不符。自然语言生成任务中的幻觉[13]可分为 2 种主要类型:内在幻觉和外在幻觉。内在幻觉属于与原内容冲突的大型语言模型(large language model,LLM)的输出。外在幻觉是指 LLM 的输出,无法从原内容中验证,原内容可能与外在知识不一致。

近年来,为了缓解生成式文本摘要模型中的幻觉问题,研究人员提出了各种策略,包括基于提示的方法和训练额外的判别模型的方法。

1)基于提示词的方法。这种方法旨在通过向模型中输入各种指令来进行训练,在模型生成摘要时根据指令的约束来生成与约束条件一致的文本,通过向模型提供具体的背景和预期的结果,来对模型的生成效果进行控制,以期望能够达到减少摘要中的幻觉问题出现的可能性。例如,文献[14]提出了一种元训练(Meta-Prompt)的提示生成模型(prompt generation model,PGM),具体方法是使语言模型能够从生成的提示词来创建对应的上下文中进行稳健的学习。PGM 通过在 COLIEE 文本蕴涵任

务上进行实验,探索了这种方法在法律推理任务中的具体效果。通过使用提示词的方法,他们的实验结果可以超过当时性能最好的模型。

尽管提示词方法可以有效地改善生成文本的幻觉问题,但基于提示词的方法在很大程度上依赖于提示词的质量。如果生成的提示词的质量较差,则模型对应的生成文本的质量将严重下降,甚至比原来的效果更差。此外,很难设计出针对所有数据都有效的提示词。

2)训练额外判别模型的方法。一些研究人员认为,评估事实一致性的最直观方法是计算生成的摘要和原文本间的事实重叠。例如,文献[15]首先尝试使用 OpenIE 工具来提取原始文本和摘要文本中的三元关系,以评估摘要中的事实是否与原始文本中的事实一致。文献[16]认为,事实一致的摘要在语义上是跟原文本相一致。文献[17]提出的基于问答的摘要生成方法(question answering and summary generation,QAGS)和文献[18]提出的基于问答自动度量的方法(finite element computations on quantum annealers,FEQA)都是通过训练问答模型来对源文本和生成的摘要来生成问题,然后让模型来根据源文本和摘要来生成问题的答案。即如果生成的摘要与原始文件事实一致,那么对摘要及源文本提出的问题将得到相一致或者相类似的答案。以上的所有这些方法都在一定程度上改善了生成摘要时的幻觉问题。然而,上述模型并没有从摘要模型的训练方式的角度去优化和改善模型中的幻觉问题,并且此类方法还需要训练额外的判别模型,不仅增加资源的消耗,还无法保证判别模型的性能。

为解决上述问题,本文提出了一种用于生成式文本摘要模型内在幻觉问题的优化框架。

2 生成式文本摘要任务

生成式文本摘要的目标是训练一个摘要生成模型 g ,在模型接收到原文本 D 之后,根据原文本生成恰当的摘要 S。可以表示为

$$S \leftarrow g(D) \tag{1}$$

生成式摘要模型的目的是能够训练一个表现效果好的摘要模型 g。模型的训练算法采用最大似然估计(maximum likelihood estimation,MLE)[19]。其目的是最大化候选摘要 S^* 的可能性,表达式为

$$\theta^* = \text{argmax} \sum_i \log p_{g\theta}(S^*_{(i)} \mid D_{(i)};\theta) \tag{2}$$

式(2)中:θ 表示 g 的参数;$P_{g\theta}$ 表示这些参数所包含的概率分布;$\{D_{(i)}, s^*_{(i)}\}$ 表示第 i 个训练样本,其中 $D_{(i)}$ 表示第 i 个参考摘要,$s^*_{(i)}$ 表示第 i 个候选摘要。

为方便计算,将其转换成求其最小化负对数似然之和,表达式为

$$L_{\text{Loss}} = -\sum_{i=1}^{K} y_i \log q_i \tag{3}$$

式(3)中:y_i 表示第 i 个位置的预测值;q_i 表示预测分布。q_i 的计算公式为

$$q_i = \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)} \tag{4}$$

(4)式中:z_i 表示当前位置的预测概率;z_j 表示整个句子的预测概率。

3 生成式文本摘要内在幻觉问题优化方法

本文所提出的用于文本摘要的内在幻觉优化方法模型如图 1 所示。

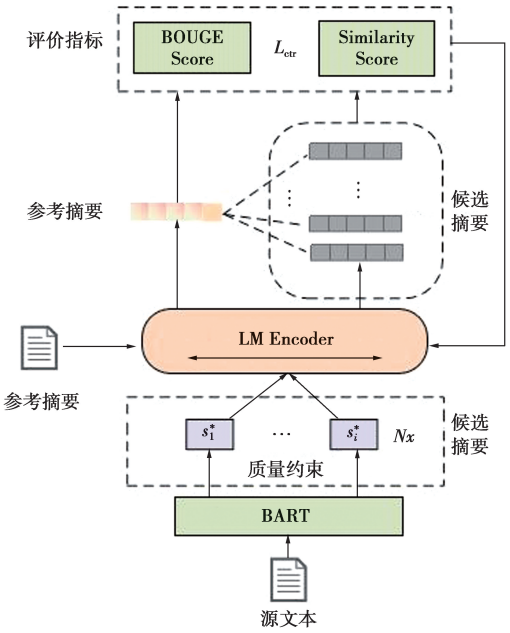

图 1 生成式文本摘要内在幻觉优化方法(模型训练阶段)

Fig.1 Optimization method for intrinsic hallucination in generative text summarization (summary training stage)

针对模型的各个训练过程,本文的方法分 2 个阶段进行优化:数据预处理阶段和模型训练优化过程。在数据预处理阶段,使用预训练的模型 BART[20] 为源文本生成多个摘要,同时以召回率作为质量约束将满足条件的摘要保留作为候选摘要。在训练阶段,设计了一种改进的训练策略,该策略能够让模型拥有评估候选摘要质量的能力,保证模型在生成摘要的时候为高质量的摘要分配更高的概率。本文模型生成摘要的阶段引入标签评估技术,

如图 2 所示,通过软化目标标签,并设计新的损失函数,让模型能够生成更准确的摘要文本。

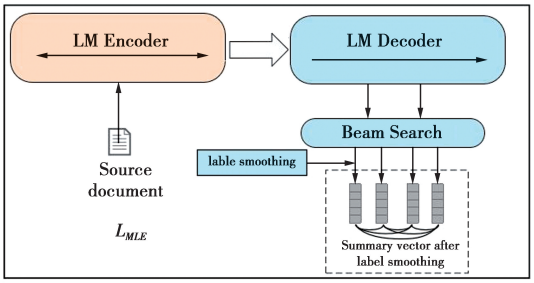

图 2　生成式文本摘要内在幻觉优化方法(摘要生成阶段)

Fig.2　Optimization method for intrinsic hallucination in generative text summarization (summary generation stage)

3.1　候选摘要生成

本文主要使用 ROUGE 度量来评估候选摘要的质量。ROUGE 评分指标由 LIN[21] 于 2003 年提出,主要是基于召回率(Recall)的,ROUGE 是一种常用的机器翻译和文章摘要评价指标。其主要计算方法是计算文本之间的重叠率,能在一定程度上反映生成摘要的质量。本文使用预训练的摘要模型 BART 为每个源文本生成多个摘要,为保证候选摘要的质量,本文选择了召回率作为质量约束条件,并在生成的摘要的召回率分数达到 40 分时将其作为候选摘要。训练数据的质量在数据层面得到加强,并提供给模型进行训练。

3.2　标签平滑技术

如果在训练模型时,只考虑预测正确的位置标签的损失,而不考虑错误预测的位置标记的损失,容易导致模型的过拟合,降低模型的鲁棒性。在正常的推理过程中,模型总是根据之前的预测结果来进行下一个位置的结果预测,即使之前的预测是不正确的,此即被称为曝光偏差问题。它会在生成的摘要中带来内在的幻觉问题。

为了解决这个问题,本文引入标签平滑技术[22],当模型在波束搜索生成摘要时,在损失计算过程中使用该技术有助于减少过拟合,提高模型的泛化能力。

标签平滑是分类问题中广泛使用的有效技术,通过软化硬标签,将分类正确和错误位置上的“1”“0”标签软化成 0—1 的小数,其具体转换公式为

$$p_{\text{true}}(s \mid D, s_j^*) = \begin{cases} 1-\partial, s = s_j^* \\ \dfrac{\partial}{N-1}, s \neq s_j^* \end{cases} \tag{5}$$

式(5)中:∂是固定值,一般取值为 0.1,N 为文本总长度;s_j^* 表示候选摘要。

具体来说,就是用转换后的标签向量来替换原来的标签向量,即

$$\hat{\boldsymbol{y}} = y_{\text{hot}}(1-\alpha) + \alpha/k \tag{6}$$

式(6)中:k 为多分类的类别总个数;a 是一个较小的参数(一般取 0.1),该公式可以表示为具体的每个标签位置转换如下

$$\boldsymbol{y}_i = \begin{cases} 1-\alpha, i = \text{目标位置} \\ \alpha/k, i \neq \text{目标位置} \end{cases} \tag{7}$$

通过标签平滑策略对目标标签进行平滑处理,将正确类别的概率分配给其他不正确类别,在一定程度上引入噪声,增强了模型的鲁棒性。

在本文中,标签平滑技术主要用于摘要生成阶段,在模型生成摘要向量的时候加入标签平滑,软化向量中的预测标签。具体来说,重新设计的标签平滑损失如下

$$L_{\text{MLE}} = -\sum_{j=1}^{l}\sum_{s} p_{\text{true}}(s \mid D, S_{<j}^*)\log p_{g\theta}(s \mid D, S_{<j}^*;\theta) \tag{8}$$

式(8)中,$p_{\text{true}}(s|D,S_{<j}^*)$表示模型在当前位置的预测经过标签平滑软化后的标签。在模型训练完成之后,生成最终的摘要。

3.3　多候选摘要模型幻觉优化方法

在模型训练阶段,给予摘要模型评估候选摘要质量的能力,模型能够通过计算候选摘要的 ROUGE 分数和 BARTScore 分数[23]来评估其质量,从而增强了训练数据集的多样性。为了降低幻觉的可能性,使用多个候选摘要和参考摘要作为输入,并鼓励更高的概率根据以下损失函数分配给质量更好的候选摘要,摘要质量评价阶段的损失函数为

$$L_{\text{ctr}} = \sum_{i}\sum_{j>i}\max(0, f(S_j) - f(S_i)) \tag{9}$$

式(9)中,S_i 和 S_j 是 2 个不同的候选摘要,并且 $\text{ROUGE}(S_i;S^*) > \text{ROUGE}(S_j;S^*)$,$f(S_i)$是长度归一化估计对数概率,计算公式为

$$f(S) = \frac{\sum_{t=1}^{l}\log p_{g\theta}(s_t \mid D, s_{<t};\theta)}{|S|^{\delta}} \tag{10}$$

(10)式中:s_t 表示当前生成位置;$s_{<t}$表示已经生成的序列。

图 2 中,使用预训练的 BART 摘要模型生成多个候选摘要之后,将参考摘要和多个候选摘要作为输入,输入到语言模型编码器中(LM Encoder),模

型作为评估者,根据 ROUGE 分数和 BARTScore 进行综合评估候选摘要的质量,同时模型进行参数学习,选择综合评分最高的候选摘要保留下来,然后将最好的候选摘要作为摘要生成阶段的输入。

因为 ROUGE 分数只是生成的摘要和参考摘要之间重叠的度量,所以它不能有效地反映模型生成的摘要的质量。因此,本文还使用 BARTScore 分数来评估生成摘要的质量。BARTScore 是一种用于评估文本生成任务的评估指标,它考虑了生成文本和参考文本之间的语义相似性,能够很好地反映了类似人类对生成质量的感知。本文中,在模型评估候选摘要质量的时候更倾向于选择更高 BARTScore 分数的候选摘要,模型根据更高质量候选摘要进行参数学习和更新。

首先,在模型评价候选摘要并学习更新参数之后,模型在最终生成摘要阶段的时候引入之前的标签平滑技术,通过波束搜索生成最终摘要的时候引入标签平滑软化生成的摘要向量,并设计新的损失函数 L_{MLE}。图 3 中的 LM Encoder 和 LM Decoder 使用的均是前一阶段训练使用的语言模型,2 个阶段共享学习参数。然后,模型根据设计的标签平滑损失函数进行参数学习和更新。为了验证本文方法的有效性,本文使用多种基线模型进行实验。最后,利用上述 2 个损失来训练模型,2 个训练阶段共享学习参数,可以在令牌级别实现摘要质量的优化。

联合 2 个训练阶段的损失函数式(8)和式(9)。在模型训练阶段,本文不仅鼓励模型将更高的概率分配给质量更好的候选摘要,而且还在生成摘要阶段根据表现最好的候选摘要进行更深入的学习和预测生成摘要。具体损失函数为

$$L_{\mathrm{mul}} = L_{\mathrm{ctr}} + \beta L_{\mathrm{MLE}} \tag{11}$$

式(11)中:L_{ctr}表示模型训练阶段评估候选摘要质量的损失函数;L_{MLE}表示摘要生成引入标签平滑技术后的损失函数。在 2 个训练阶段给予模型不同的角色,前一阶段模型作为评估者评估候选摘要质量,后一阶段模型作为摘要生成器,根据最好的候选摘要进行摘要生成和参数学习,β 是超参数,用来衡量 2 个指标之间的重要度,在实验中具体设置大小。

4 实 验

4.1 实验数据集

为验证提出方法的有效性,本文在 2 个公开英文数据集上进行实验,证明了模型在真实数据上的实用性,数据集详细分析如表 1 所示。

表 1 实验数据集

Tab.1 Experiment dataset

数据集	主要内容	数据划分		
		Train	Test	Val
CNNDM	News	287 113	13 368	11 490
XSum	News	204 045	11 332	11 334

1)CNNDM 数据集是大规模新闻数据集,是文本摘要任务中常用数据集。包含训练样本 287 113 条,测试样本 13 368 条。将新闻文本作为源文本,相关亮点作为参考摘要,源文本和参考摘要长度都较长。

2)XSum 数据集是英国广播公司(BBC)的高度抽象的文章数据集,训练集中包含训练样本204 045 条,测试样本 11 332 条,其源文本和参考摘要长度都较短。

4.2 实验参数说明

由于 2 个数据集的源文本和参考摘要的长度差距较大,所以需要单独设置实验参数。对于 CNNDM 数据集,输入源文本的最大长度设置为 1024,生成摘要的最大长度设置为 140;对于 XSum 数据集输入源文本最大长度设置为 512,生成摘要的最大长度设置为 60。另外,训练轮数设置为 5,波束生成搜索大小设置为 8,使用 Adam 优化器,学习率设置为 0.001。

4.3 对比模型

实验对比了多种近年来文本摘要领域的方法,选取 BART 作为基本框架模型。对比模型简介如下。

1)Pegasus。基于 Transformer 结构的一种新的摘要生成预训练模型。相比于其他通用预训练模型,Pegasus 模型的架构设计更贴近下游的摘要生成任务,其在抽取式摘要上的表现相比其他模型表现更好。

2)SimCLS。基于 BART 模型,再使用抽取式方法的额外训练来提高最终生成摘要的效果。

3)T5。其训练方式是将每个文本处理问题都看成“Text-to-Text”问题,即将文本作为输入,生成的新文本作为输出。其可以完成各种文本任务,在文本摘要任务中也表现优异。

另外,还有近年来在文本摘要领域中表现优异的其他一些模型。

4.4 实验结果与分析

本文提出的方法在 2 个英文数据集上的实验结

果分别如表 2 和表 3 所示。本文在 2 个数据集之上均取得了最佳的 ROUGE 分数。与多个基线模型相比,本文在 CNNDM 的 ROUG-1 评分上实现了平均 8.18%的改善,在 XSum 的 ROUGE-1 评分上平均提升了 6.68%。

表 2　CNNDM 数据集的实验结果

Tab.2　Results on CNNDM. R-1/2/L are the ROUGE-1/2/L F1 scores

模型	评估指标		
	R-1	R-2	R-L
BART(2020)	44.16	21.28	40.90
Pegasus(2020)	44.17	21.47	41.11
SimCLS(2021)	46.67	22.15	43.54
T5(2020)	43.52	21.55	40.69
GOLD(2021)	45.40	22.01	42.25
SEASON(2022)	46.27	22.64	43.08
Fourier Transformer (2023)	44.76	21.55	43.34
Ours(BART)	**47.95**	23.52	**44.20**
Ours(Pegasus)	47.45	24.21	42.37
Ours(T5)	47.24	**24.30**	42.01

注:加粗数据表示最佳性能。

表 3　XSum 数据集的实验结果

Tab.3　Results on XSum. R-1/2/L are the ROUGE-1/2/L F1 scores

模型	评估指标		
	R-1	R-2	R-L
BART(2020)	45.14	22.27	37.25
Pegasus(2020)	45.20	24.56	39.25
GOLD(2021)	45.75	22.26	37.30
SimCLS(2021)	47.61	24.57	39.44
HAT-BART(2021)	45.92	22.79	37.84
SummaReranker(2022)	48.12	24.95	40.00
Ours(BART)	**48.42**	**25.26**	**40.06**
Ours(Pegasus)	47.96	25.12	40.01
Ours(T5)	48.06	24.68	39.85

注:加粗数据表示最佳性能。

尽管本文提出的方法在每个数据集上都获得了最高的 ROUGE 分数。然而,ROUGE 分数仅计算生成摘要和参考摘要之间的重叠指标,不能有效反映生成摘要的质量。为进一步说明本文所提方法的性能,将对生成的摘要的质量采用额外的评估方法。

4.5　分析与讨论

以 CNNDM 数据集为例,本文从不同的角度和不同的指标进行分析,深入地了解本文方法的性能。

1)标签平滑损失。本文比较了使用标签平滑前后的 ROUGE 得分,得分结果如表 4 所示。实验结果表明,标签平滑的引入有效地提高了生成摘要的质量。其原因是生成的序列不一定完全与参考摘要相同,通过结合标签平滑之后,可以减轻模型的过度自信,进一步会减少摘要中的幻觉问题。

表 4　添加标签平滑前后 ROUGE 评分的比较

Tab.4　Comparison of ROUGE score before and after adding label smoothing

方法	R-1	R-2	R-L
no-label smoothing	46.27	23.31	43.80
label smoothing	**47.95**	**23.52**	**44.20**

注:加粗数据表示最佳性能。

2)增加波束宽度。理论上,增加波束搜索的宽度将允许生成更多的候选令牌,并提高模型的性能。但许多现有模型的性能往往会下降,因为模型难以将其与高质量的候选令牌区分开来。

表 5 给出了不同波束宽度下的 ROUGE 分数。可以观察到 BART 的性能随着波束宽度的增加而降低。相反,结合本文提出的方法模型在更大的波束宽度下获得更好的性能。这表明本文提出的方法软花了概率分布,模型从而获得了更高质量的摘要表示。

表 5　不同波束宽度的 CNNDM 数据集实验结果

Tab.5　Results on CNNDM dataset in beam width

波束宽度	BART		本模型	
	R-1	R-2	R-1	R-2
4	44.29	21.17	47.68	23.43
8	43.96	20.85	47.86	23.66
16	43.68	20.68	**47.95**	23.52
32	43.21	20.41	47.90	**23.80**

注:加粗数据表示最佳性能。

3)新颖性比较。本文比较了 BART 模型结合本文方法前后生成摘要中新词组合的比例,以进一步反映本文方法生成摘要的新颖性,文本选择使用“Novelty”指标[24]进行度量,实验结果如图 3 所示。

$$Novelty(D,S^*)=\frac{\sum_{g\in G_{S^*}}1(g\notin G_D)}{|G_{S^*}|} \tag{12}$$

式(12)中:D 和 S^* 分别表示源文档和参考摘要;G_D 和 G_{s*} 是 n 元词汇集合。结合本文提出方法之后,BART 模型更具“新颖性”。尽管与参考摘要相比,

BART 可以生成许多不同的单词,但结合本文方法之后模型生成了更通用和更符合人们日常阅读习惯的摘要文本。

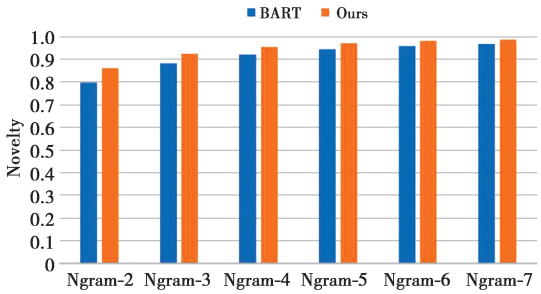

图 3 摘要中生成新词汇

Fig.3 Generate new vocabulary in the summary

4)Pearson 和 Spearman 相关性。为了体现本文方法能够有效地改进摘要模型的幻觉问题,使用本文方法训练的 BART 模型生成的摘要进行 Pearson 和 Spearman 相关性评分[25]。通过细粒度的计算“Consistency”,“Fluency”分数。同时,计算 BLEU 和“Relevance”分数用于评估生成的摘要。这些分数都是针对摘要结果中的某一个细粒度的层次去评价摘要的质量,比如“Consistency”即“一致性”分数,就是针对摘要中的一致性问题去对生成的摘要结果进行评分,分数越高,则表明生成的摘要更贴合源文本。实验结果列于表 6 中。本文的方法在所有指标上都比基线模型表现更好。

表 6 生成摘要的 Spearman 和 Kendall-Tau 相关性

Tab.6 Summary-level Spearman and Kendall-Tau correlations of different metrics

指标	Consistency	Fluency	Relevance	BLEU
BART	0.541	0.624	0.406	0.178
Ours	0.574	0.668	0.448	0.268

5)人工评价。为了更直观地展示本文模型生成摘要的质量,本文从 2 个数据集中各随机选择了 100 条数据,以人工评价方式评估模型生成摘要的质量。表 7 展示了人工评估的实验结果。从表 7 可以看出,分别在 2 个数据集之上对比基线模型,本文的方法在 2 个数据集上随机挑选的 100 数据的摘要生成结果中,在 CNNDM 数据集上,有 78 条数据的生成摘要结果比基线模型更好,15 条效果相似,只有 7 条生成摘要结果更差;在 XSum 数据集上,有 82 条数据的生成摘要结果比基线模型更好,10 条效果相似,只有 8 条生成摘要结果更差。

表 7 人工评估生成摘要与基线摘要评估结果

Tab.7 Human evaluation to generate summary results

数据集	人工评估		
	Better	Nearly	Worse
CNNDM	**78**	15	7
XSum	**82**	10	8

5 结束语

本文提出了一种用于生成式文本摘要模型的内在幻觉优化方法。为了优化模型中的幻觉问题,首先,通过在训练阶段设计一个模型自模型评分模式,给予模型评估候选摘要质量的能力。然后,引入了一种改进的标签平滑技术,推动模型生成与参考摘要一致的文本,并使模型生成策略更加灵活。这些有助于模型根据源文本生成忠实的摘要。实验结果表明,本文提出的方法可以有效缓解生成摘要中的内在幻觉问题。

在未来的研究中,将继续深入研究摘要生成过程,并研究如何更细粒度地优化摘要生成,从多方面综合提升摘要的质量,进一步提高模型性能。

参考文献:

[1] LIU J, KONG X, ZHOU X, et al. Data mining and information retrieval in the 21st century: A bibliographic review[J]. Computer Science Review. 2019(34): 100193.

[2] CHAKRABORTY R, BHAVSAR M, DANDAPAT S K, et al. Tweet summarization of news articles: An objective ordering-based perspective [J]. IEEE Transactions on Computational Social Systems. 2019, 6(4): 761-77.

[3] BLEKANOV I, TARASOV N, BODRUNOVA S S. Transformer-based abstractive summarization for Reddit and Twitter: single posts vs. comment pools in three languages [J]. IEEE Transactions on Computational Social Systems, 2022, 14(3): 69-93.

[4] DÖNMEZ I, IDIN S, GÜLEN S. Conducting academic research with the ai interface, ChatGPT: Challenges and opportunities[J]. Journal of STEAM Education, 2023, 6(2): 101-18.

[5] AIZAWA A. An information-theoretic perspective of tf—idf measures[J]. Information Processing & Management, 2003, 39(1): 45-65.

[6] ZHANG M, LI X, YUE S, et al. An empirical study of TextRank for keyword extraction [J]. IEEE Access, 2020(8): 178849-178858.

[7] BIANCHINI M, GORI M, SCARSELLI F. Inside page

rank[J]. ACM Transactions on Internet Technology, 2005, 5(1): 92-128.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//In Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California: NIPS, 2017: 1-11.

[9] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language. Technologies, Minneapolis, USA: ACL, 2019: 4171-4186.

[10] FLORIDI L, CHIRIATTI M. GPT-3: Its nature, scope, limits, and consequences[J]. Minds and Machines, 2020(30): 681-94.

[11] WAN W, LIAO B, CAO J, et al. Transforming Automated Customer Service Responses Using ChatGLM Model: A Case Study of Expressway Toll Station Information System[C]//In 2023 IEEE 7th Information Technology and Mechatronics Engineering Conference (ITOEC). Chongqing, China: IEEE, 2023: 1968-1973.

[12] KATZ D M, BOMMARITO M J, GAO S, et al. Gpt-4 passes the bar exam[J]. Philosophical Transactions of the Royal Society A, 2024, 382(2270): 20230254.

[13] CHEN Y, FU Q, YUAN Y, et al. Hallucination detection: Robustly discerning reliable answers in large language models[C]//In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. New York, USA: Computing Machinery, 2023: 245-255.

[14] HYEONMIN H A, JIHYE L, WOOKJE H, et al. Meta-Learning of Prompt Generation for Lightweight Prompt Engineering on Language Model as a Service[C]//In Findings of the Association for Computational Linguistics, Singapore. Association for Computational Linguistics. Singapore: ACL, 2023: 2433-2445.

[15] GOODRICH B, RAO V, LIU P J, et al. Assessing the factual accuracy of generated text[C]//In proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. Toronto, Canada: ACM, 2019: 166-175.

[16] FALKE T, RIBEIRO L F, UTAMA P A, et al. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference[C]//In Proceedings of the 57th annual meeting of the association for computational linguistics. Florence, Italy: ACL, 2019: 2214-2220.

[17] ESIN D, HE H, MONA D. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization[C]//In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2019: 5055-5070.

[18] ALEX W, KYUNGHYUN C, MIKE L. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries[C]//In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: ACL, 2020: 5008-5020.

[19] MYUNG I. Tutorial on maximum likelihood estimation[J]. Journal of Mathematical Psychology. 2003, 47(1): 90-100.

[20] LEWI S, MIK E, LI U, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension[C]//In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Washington, USA: ACL, 2020: 7871-7880.

[21] SCHLUTER N. The limits of automatic summarization according to rouge[C]//In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Melbourne, Australia: ACL, 2017: 41-45.

[22] MÜLLER R, KORNBLITH S, HINTON G. When does label smoothing help?[J]. Advances in Neural Information Processing Systems. 2019(32): 1-15.

[23] YUAN W, NEUBIG G, LIU P. Bartscore: Evaluating generated text as text generation[J]. Advances in Neural Information Processing Systems. 2021(34): 27263-77.

[24] JAIN S, KESHAVA V, SATHYENDRA S M, et al. Multi-Dimensional Evaluation of Text Summarization with In-Context Learning[C]//In Findings of the Association for Computational Linguistics. Toronto, Canada: ACL, 2023: 8487-8495.

[25] MINGKAI D, BOWEN T, ZHENGZHONG L, et al. Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation[C]//In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Bangkok, Thailand: ACL, 2021: 7580-7605.

作者简介:

李　能,硕士研究生,主要研究方向为自然语言处理,文本摘要。E-mail:s220201050@stu.cqupt.edu.cn。

于成成,硕士研究生,主要研究方向为自然语言处理。E-mail:s230201155@stu.cqupt.edu.cn。

刘　群,教授,博士生导师,工学博士,主要研究方向为数据挖掘、深度学习、自然语言处理、复杂网络等的理论与应用研究。E-mail:liuqun@cqupt.edu.cn。

(**编辑**:田海江)