面向文本摘要的反事实纠偏方法

陈 璐 张儒清 郭嘉丰 范意兴

(中国科学院计算技术研究所网络数据科学与技术重点实验室 北京 100190)

(中国科学院大学 北京 100190)

摘 要 文本摘要是自然语言处理领域中一项典型的文本到文本生成任务,旨在提取和概括一篇或多篇输入文档的关键信息，生成简洁、流畅又准确的摘要文本. 自动文本摘要技术涉及自然语言理解和自然语言生成技术，并能应用于多种实际场景，包括文档索引、标题生成和内容创建，因此受到学术界和工业界的长期关注. 近年来，基于神经网络的深度文本摘要模型得到广泛研究. 结合先进的预训练技术，现有的深度文本摘要模型已经具备流畅的语言表达能力，能够生成较为通顺的摘要. 然而，模型生成的摘要仍然存在表达不准确的问题，与原文存在信息偏差或包含原文以外的信息. 该问题被称为“幻觉”问题，仍是一个巨大的挑战. 针对这个问题，该文从因果的角度分析了基于预训练模型的深度文本摘要方法存在的偏差来源，并设计了去偏方法. 因果理论为理解和建模复杂系统提供了一个强大的框架. 在文本摘要系统中，因果推理可以帮助识别文档、摘要和语言先验之间的因果关系. 理解这些变量之间的因果关系，有助于设计出针对系统中潜在偏差来源的去偏方法. 具体来说，该文首先探究了文本摘要任务的因果结构，定义和分析了摘要任务的因果图. 分析表明，摘要会受到预训练过程中习得的语言先验的影响. 其中，语言先验包含的噪声会导致生成的摘要有偏. 由于先前的摘要模型没有考虑或规避语言先验中潜在噪声的影响，导致模型生成的摘要中容易出现原文没有的信息. 为此，该文根据因果理论提出了面向文本摘要的反事实纠偏方法. 受到人类行为的启发，该文根据是否和原文交互，显式地区分语言先验中的有用知识和噪声,然后建模噪声对摘要的影响并从总体影响中去除. 在 XSUM 和 CNN/DailyMail 数据集上的实验表明，该模型在 Rouge-1、Rouge-2 和 Rouge-L 指标上分别比基线方法 BART 提高 0.75%、0.54%和 0.46%以及 1.29%、2.08%和 1.20%，并且在人工评价中具备良好的流畅性和忠实性. 本文提出的方法是一个通用的框架，适用于不同的深度文本摘要模型. 通过利用因果理论，它对文本摘要领域以及其他文本生成任务有一定的启发，增加了该领域方法的可解释性.

关键词 深度文本摘要模型；因果效应；反事实推理；幻觉；去偏

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2023.02400

Counterfactual Debiasing for Text Summarization

CHEN Lu ZHANG Ru-Qing GUO Jia-Feng FAN Yi-Xing

(*CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

(*University of Chinese Academy of Sciences, Beijing* 100190)

Abstract Text summarization is a classic and critical text-to-text generation task in natural language processing (NLP), which aims to capture the key information of single or multiple documents and generate concise, fluent and accurate text as summary. Automatic text summarization technology involves natural language understanding and natural language generation techniques, and can be applied to a variety of practical scenarios, including document indexing, title generation, and content creation. Therefore, it has at-

收稿日期：2022-09-27；在线发布日期：2023-03-16. 本课题得到国家自然科学基金项目（Nos. 62006218，61902381）、中国科学院青年创新促进会（Nos. 20144310，2021100）、中国科学技术协会青年人才托举工程（No. YESS20200121）、联想-中科院联合实验室青年科学家项目资助. **陈 璐**，博士研究生，主要研究领域为自然语言处理、信息检索. E-mail: chenlu19z@ict.ac.cn. **张儒清**（通信作者），博士，副研究员，主要研究领域为自然语言处理. E-mail: zhangruqing@ict.ac.cn. **郭嘉丰**（通信作者），博士，研究员，主要研究领域为数据挖掘、信息检索. E-mail: guojiafeng@ict.ac.cn. **范意兴**，博士，副研究员，主要研究领域为数据挖掘、信息检索.

tracted long-term attention from both academia and industry. In recent years, deep text summarization methods based on neural networks have been widely studied. In combination with advanced pre-training technology, the existing deep text summarization methods have had powerful language expression capabilities and enhanced the fluency of generated summaries. However, despite these improvements, the generated summaries still suffer from the problem of inaccurate expression, containing information deviation from or not present in the original document. This problem, commonly referred to as hallucination, remains a significant challenge. To solve the problem of hallucination, this paper analyzes the source of bias, which exists in deep text summarization methods based on pre-trained models, from the perspective of causality, and designs a debiasing method accordingly. The theory of causality provides a useful framework for understanding and modeling complex systems. In the context of text summarization, causal inference helps identify the causal relationships between documents, summaries, and language priors. By understanding the causal relationships between these variables, it becomes possible to design debiasing methods that account for the potential sources of bias in the system. Specifically, this paper begins by exploring the causal structure of text summarization task, through defining a causal graph for this task and analyzing the causal relationships among documents, corresponding summaries, and language priors. The analysis reveals that language priors learned from the pre-training process may introduce noise that leads to biased summaries being generated. However, previous summarization models have not considered the influence of potential noise in the language priors or avoided it, resulting that the generated summaries often contain additional information without grounding in the input documents. To address this issue, this paper proposes a text summarization debiasing method based on counterfactual inference, according to the theory of causality. Inspired by human behaviors, this paper explicitly distinguishes between useful knowledge and noise in the language priors based on whether they interact with the source document. Then, this paper models the direct effect of noise on the summary and excludes it from the total effect. The experimental results show that the proposed method is 0.75%, 0.54% and 0.46% better than BART on XSUM and 1.29%, 2.08% and 1.20% better on CNN/DailyMail in terms of Rouge-1, Rouge-2 and Rouge-L metrics, respectively. Also, the proposed method demonstrates comparable fluency and better faithfulness in human evaluation. The method proposed in this paper is a general framework that can be applied to various deep text summarization models. By leveraging causal theory, it sheds some light on the field of text summarization or other text generation tasks, and increases the interpretability of work in this field.

Keywords　deep text summarization model; causal effect; counterfactual inference; hallucination; debias

1　引　言

自动文本摘要技术旨在从一篇或多篇输入文档中自动识别重要主题和关键信息，并根据它们生成准确、简洁又流畅的文本作为摘要. 该技术应用场景丰富，包括新闻标题生成[1]、科学文献摘要生成[2]、商品标题生成[3]等. 例如，各大商业搜索引擎在网页上提供检索文档的摘要，有助于用户确认文档的相关性.

早期的自动文本摘要技术主要采用人工设计的启发式规则和模板[4]. 随着深度学习的发展，基于神经网络的深度文本摘要技术得到广泛的关注和研究，其主流的方法可分为抽取式摘要和生成式摘要两大类. 其中，深度抽取式摘要方法得到的摘要文本流畅通顺，且语义与原文能够保持高度的一致性，因此在现实生活中得到广泛应用. 而深度生成式摘要方法虽然能够得到简洁精练的摘要文本，并且具备更高的灵活性和多样性，却在流畅性上表现欠佳. 近年来，预训练技术为自然语言理解和自然语言生成的研究提供了新的生机，同时也提升了文本摘要技术的性能表现[5]. 生成式摘要方法利用预训练模型强大的语言表达能力有效提升了生成文本的流畅

性，使实用性得到增强[6].

然而，基于预训练的生成式摘要方法依然存在表达不准确的问题，可能生成和原文信息不一致的摘要文本[7]. 如表 1 所示，对于一篇与油气开采的“水力压裂法（Hydraulic Fracturing）”技术相关的文章，全文并没有提到与“公众健康（Public Health）”相关的信息. 然而 BART 模型在生成摘要时，却生成了这样的字眼. 若是不参考原文，单独检查模型生成的摘要，可能不会看出其中的不准确. 类似的现象也发生在其他文本生成任务中，如对话[8]、翻译[9]等. 它们的共同点在于，生成文本看似正确，实际上存在偏差，包含与原文信息不一致的内容. 该现象也被称为“幻觉（Hallucinations）”.

表 1 BART 生成摘要示例

类型	内容
原始文档	“Fracking” involves pumping water and chemicals into shale rock at pressure. The joint report from the Royal Society and Royal Academy of Engineering say the technique is safe if firms follow best practice and rules are enforced. Exploratory fracking is being mooted in at least seven sites around the UK. The report was commissioned by the government's chief scientist, Sir John Beddington, following the decision last year to halt the UK's most advanced project, in Lancashire, after fracking caused small earth tremors. “Our main conclusions are that the environmental risks of hydraulic fracturing for shale can can be safely managed provided there is best practice observed and provided it’s enforced through strong regulation, ” said the report's chair… “水力压裂”是指在压力下将水和化学物质泵入页岩中. 英国皇家学会和英国皇家工程院的联合报告称，如果公司遵循最佳实践，并执行规则，这种技术是安全的. 在英国至少有 7 个地方正在讨论试探性水力压裂. 该报告是由政府首席科学家约翰• 贝丁顿爵士(Sir John Beddington)委托撰写的. 去年，在水力压裂法引发小规模地震后，英国决定停止位于兰开夏郡(Lancashire)的最先进项目. “我们的主要结论是，页岩水力压裂的环境风险是可以安全管理的，只要有最佳实践观察，并通过强有力的监管加以执行，”报告主席说……
参考摘要	A gas extraction method which triggered two earth tremors near Blackpool last year should not cause earthquakes or contaminate water but rules governing it will need tightening, experts say. 专家表示，去年在布莱克浦附近引发两次地震的一种气体抽取方法，应该不会引发地震或污染水源，但管理该方法的规定需要加强.
BART 摘要	Shale gas extraction can be carried out safely in the UK, but stronger regulations are needed to protect **public health** and the environment, a report says. 一份报告称，页岩气开采可以在英国安全进行，但需要更严格的法规来保护公众健康和环境.

为了让基于预训练的生成式摘要模型生成更准确的摘要文本，我们根据因果理论来分析该现象. 首先，我们探究了文本摘要任务中各个关键要素之间的因果关系. 除了原始文档和摘要之外，我们还考虑了预训练模型在预训练过程中从大规模语料库学到的语言先验对摘要的影响. 根据三者之间的因果关系，我们定义了一个因果图，并结合因果图分析了语言先验对摘要的影响方式. 具体来说，我们显式地区分语言先验中的有用知识和噪声，并将摘要中的偏差归因于语言先验中的噪声. 根据上述分析，我们基于反事实推理提出了摘要去偏方法，通过移除语言先验对摘要的直接影响来减少噪声带来的负面影响. 相应地，我们设计并实现了摘要去偏模型，并提出三个训练目标用于模型学习. 在摘要数据集 XSUM 和 CNN/DailyMail 上的实验表明，我们的方法能够有效地提高摘要的生成质量.

本文的主要贡献在于：

（1）定义摘要任务的因果图，用于描述原文、摘要和语言先验之间的因果关系，并结合因果图分析得出，摘要中的偏差来源于语言先验中的噪声；

（2）针对确定的偏差来源，提出基于反事实推理的摘要去偏方法，设计了具体的模型框架，并提出相应的训练目标；

（3）通过在 XSUM 和 CNN/DailyMail 数据集上的对比实验，验证了本文所提方法的有效性.

本文其余章节组织如下：第二章介绍相关工作，主要包括深度文本摘要方法、基于因果理论的去偏方法和预训练语言模型；第三章介绍因果理论的背景知识，包括结构因果模型和因果效应；第四章介绍我们所提出的方法，包括介绍文本摘要的结构因果模型、基于反事实推理的摘要去偏方法及其具体实现等；第五章介绍实验，包括实验设置、实验结果和相关分析；第六章总结了本文工作并提出未来计划.

2 相关工作

本章节将简单梳理与本文相关的研究工作，包括深度文本摘要方法、基于因果理论的去偏方法和预训练语言模型.

2.1 深度文本摘要方法

基于神经网络的深度文本摘要方法可分为深度抽取式摘要方法和深度生成式摘要方法. 深度抽取式摘要方法抽取原文中的句子子集作为摘要文本. 它通常将摘要任务转换为序列标注或排序形式的任务. 序列标注形式指的是对文档中每个侯选句子完成二分类标注，判断每个候选句子是否应作为摘要[10]. 排序形式指的是对文档中的每个候选句子按照重要程度排序，然后选择前几个句子构成摘要[11]. 在深

度抽取式摘要方法中，由于构成摘要的句子摘取自原始文本，因此得到的摘要文本天然具备表达流畅性和信息准确性. 然而该方式得到的摘要受限于原始文本的表达，可能具有内容重复、信息冗余等缺陷，更适用于原始文档长度较短的情况. 深度生成式摘要方法则将摘要任务视为序列到序列的生成任务，以词语或短语作为基本生成单元，从无到有、逐单元地生成摘要[12]. 在深度生成式摘要方法中，由于摘要文本不拘泥于原始文本的表达，生成的摘要具备更高的灵活性和多样性，并且在长度、详略程度等方面可控性高[13]，因此适用于各种长度文档的摘要，适用性强. 然而，该方法在流畅性上明显不如深度抽取式摘要方法.

近年来，蓬勃发展的预训练技术进一步为深度文本摘要方法带来了性能增益. 抽取式摘要方法能够利用预训练模型从大规模文本中学到的语义空间，更好地从原文中选择合适的句子作为摘要[14]. 生成式方法则从通用[6]或任务特定[15]的预训练模型中继承优秀的语言表达能力，有效地克服了自身流畅性欠佳的缺陷. 结合自身适用性强的优势，基于预训练模型的生成式摘要方法逐渐成为自动文本摘要技术中主流的研究课题.

然而，基于预训练模型的生成式摘要方法仍然面临着“幻觉”的巨大挑战，常常生成和原文信息不一致的摘要文本. 现有工作[16]提出了 COCO 指标，用于检测摘要偏差，衡量摘要的事实一致性. 为了克服“幻觉”的挑战，现有工作通过预处理[17]、后处理[18]等操作，或者结合对比学习[19]和多任务学习[20]等训练技巧来提升生成摘要文本与原始文本的一致性. 例如，CLIFF[19]通过对比学习的方式，使模型学会自行区分正确的摘要和存在“幻觉”的摘要. 这些工作通过数据驱动训练模型，具备一定的效果，却需要负例构造、多任务设计等额外操作，且缺乏可解释性.

本文引入因果理论解决基于预训练模型的生成式文本摘要方法中存在的“幻觉”问题. 首先探究偏差的来源，然后针对性地去除偏差. 该方法更具可解释性.

2.2　基于因果理论的去偏方法

基于因果理论的去偏方法在搜索[21]、推荐系统[22]、计算机视觉[23]和自然语言理解[24]等多个领域得到应用并具备亮眼的表现，主要解决数据集分布不均带来的偏差问题. 现有工作可分为平衡数据法和基于结构因果模型的因果推理法.

平衡数据法通过权重调整、插补法等方式来平衡数据分布[21]. 前者对观测数据调整权重，后者对未观测数据进行估计. 双稳健方法则同时运用了两种方式平衡数据分布[25].

基于结构因果模型的因果推理法则是在结构因果模型的基础上完成偏差来源的分析和去偏，包括因果干预法和反事实推理法. 因果干预法通过后门调整、前门调整等方式，消除由混淆因子引入的伪相关关系导致的偏差[26]. 反事实推理法则通过显式地建模偏差的影响，并将它从总体影响中消除来完成去偏[22]. 比如，CF-VQA[23]借助因果图描述和分析视觉问答任务中图像、问题和答案各要素的因果关系，然后定位答案的偏差源于问题的语言先验. 针对该偏差，CF-VQA 显式地建模了问题中的语言先验对答案的直接影响，并将其从总体影响中去除.

本文使用基于结构因果模型的反事实推理法对摘要去偏，原因在于：我们既要保留语言先验中有用知识对摘要的正面影响，又要去除语言先验中噪声对摘要的负面影响.

2.3　预训练语言模型

主流的预训练方法根据使用方式可以分为两大类，分别是基于特征表示的方法和基于微调的方法.

基于特征表示的预训练方法在早年间应用广泛，典型的方法包括静态词向量表示法 Word2Vec[27]、Glove[28]和动态词向量表示法 ELMo[29]. 该类预训练方法用于获取文本的向量表示，该向量表示的获得与下游任务的目标无关. 其中，获得的表示包括（1）词向量表示：常见的预训练目标包括经典的单向语言模型目标[30]，以及根据上下文内容预测中心词的目标 [27]. （2）句子向量表示[31]或段落表示[32]：常见的预训练目标包括对相邻句子的排序目标，根据已生成句子的表示生成下一个单词的目标[31]，或是基于去噪自编码器的目标[33]. 在许多自然语言处理任务中，直接使用预训练的向量表示能获得比从零开始训练的向量表示更好的性能. 以 ELMo 为例，通过拼接对给定文本提取的双向特征作为最终的词向量表示，ELMo 在多个任务上的性能表现获得了显著提升，包括问答任务[34]和情感分析任务[35].

基于微调的预训练方法在近年来成为研究热点，典型的方法包括 GPT[36]、BERT[37]、BART[6]等等. 该类预训练方法不仅为下游任务提供文本的向量表示，还能通过在下游任务上微调，即与任务相关的模型架构一起训练，获得下游任务的标签作为直接监督信号，从而进一步提升下游任务的性能表现. 该方法所获得的通常是句子、段落或文档的向量表示. 这些预训练模型大多由多层 Transformer 堆

叠而成，其主要组成模块 Transformer 以多头注意力机制为核心. 该类方法中，常见的预训练目标包括（1）自回归语言模型目标：根据前序生成词汇从左到右地预测下一个单词[36]，或借助注意力解码等方式改变预测顺序[38]；（2）掩码语言模型目标：根据上下文的内容来预测被掩盖的部分，包括预测离散的词汇[37]或连续的片段[39]；（3）相邻句子判别目标：将两个句子拼接，判断这两个句子是否是邻接的句子[37]；（4）基于去噪自编码器的目标：将经过噪声干扰的句子复原成原句子，噪声形式包括文档旋转、句子乱序、词汇增删改等[6]；（5）与下游任务适配的目标：如与检索任务适配的代表词预测目标[40]. 在各种预训练目标的指导下，预训练模型为更多的自然语言处理任务带来了卓越的性能提升.

然而，这些常见的预训练目标通常建立在极大似然估计的基础上. 这种形式可能带来的问题是：为了迎合训练目标，预训练模型可能通过捕获数据集中浅层的线索或表面的模式来学习文本中存在的相关关系，包括伪相关关系. 这种学习模式可能会导致模型继承语料库中不平衡的数据分布所导致的偏差. 这种偏差存储于预训练模型的参数中，会进一步致使预训练模型在下游任务上被有偏的先验所误导.

本文的目的就是为了避免预训练语言模型在下游任务上受语言先验中存在的偏差所误导.

3 因果理论的背景知识

本章节将介绍因果理论的背景知识，首先通过介绍“三级因果阶梯——关联、干预和反事实”引入反事实推理，然后介绍结构因果模型，最后介绍因果效应.

3.1 反事实推理

在“三级因果阶梯——关联、干预和反事实”中，第一层是“关联”，其通过观察被动获得信息，拟合可观测数据并建模相关关系，并未考虑因果关系；第二层是“干预”，其通过主动采取行动获取新的知识，探索因果关系；第三层是“反事实”，其通过想象创造新事物，以此学习不可观测的世界.

为了更好地理解这三者的含义，下面举一个实际的例子. 当我们想要探究“电商平台分发红包是否会促进消费”这个问题时，从“关联”、“干预”和“反事实”三个层级来分析问题的做法分别如下. （1）“关联”：从已有的历史观测数据中，观察和统计分发到红包的群体和没有分发到红包的群体的消费情况. 然而在该做法下，这两个群体的消费情况不可直接比较，这是因为这两个群体并非随机划分得到，自身可能具有不同的特性. 在没有约束除了“分发红包”之外的其他因素时，不能保证两个群体间的消费差异是由“分发红包”这一因素引起的；（2）“干预”：开展随机发放红包的实验，使得两个群体的其他特性不会造成消费情况的差异；（3）“反事实”：构造反事实样本，探究如果分发到红包的群体没有拿到红包，消费情况有何差异. 在该做法中，反事实样本是相对于可观测的样本而言的. 现实世界中，一个群体不可能在相同时间内、相同平台上既拿到红包、又没拿到红包. 为确保除了“分发红包”这一因素外的所有其他因素保持一致，需要设定与事实相反的条件来构造反事实样本. 此后通过比较反事实样本与真实样本，就能得到理想的因果关系.

像这样设定与事实相反的条件来构造假设的理想样本，并比较理想样本与真实的现存样本以确定变量之间的因果关系的过程就叫做反事实推理.

3.2 结构因果模型

结构因果模型由因果图和结构方程两部分组成. 因果图是一个有向无环图 $G=\{V,E\}$，其中 V 表示变量集合，E 表示变量之间的因果关系[22]. 两个变量之间如果存在有向路径，则意味着祖先节点是因，子孙节点是果. 其中，所有直接相连的父子节点中，父节点是其所有子节点的直接原因，会对子节点产生直接影响. 而存在有向路径却不直接相连的祖先节点会对子孙节点产生间接影响. 以图 1 所示的因果图为例，Z、Y 和 K 分别表示处理变量、目标变量和中介变量. 从 Z 到 Y 之间存在两条有向路径，表示 Z 是 Y 的因，改变 Z 时，Y 会相应的发生改变. 其中，$Z\to Y$ 表示 Z 对 Y 存在直接影响，$Z\to K\to Y$ 表示在中介变量 K 的传递作用下，Z 对 Y 存在间接影响. 综合两种影响，Y 的值可以表示成：

$$Y_{z,k}=Y\left(Z=z,K=k\right),$$

其中，大写字母表示变量，小写字母表示它的观测值. 同样，中介变量 K 的值可以表示成：

$$K_z=K\left(Z=z\right).$$

这两个式子可视为变量 Y 和 K 的结构方程. 结构方程用于确定变量在所有直接父节点的影响下的取值，描述了变量的生成机制.

3.3 因果效应

变量 Z 对 Y 的因果效应指的是目标变量 Y 在处

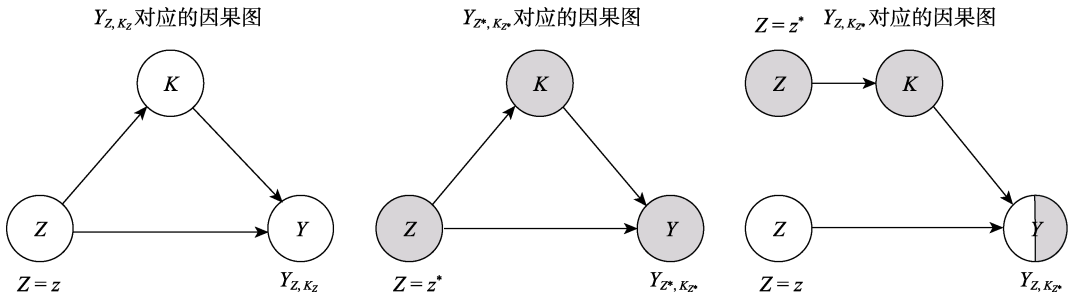

图 1　因果图示例

理变量 Z 发生变化时随之改变的量. 比如，在图 1 所示的因果图中，Z 对 Y 的总体效应（Total Effect，TE）定义为

$$TE = Y_{z,K_z} - Y_{z^*,K_z^*},$$

表示在 $Z=z$ 和 $Z=z^*$ 这两种情况下 Y 取值的差异. 其中，$Z=z^*$ 表示消除变量 Z 的影响，典型做法是将其置为空值，在图 1 因果图中用灰色节点表示. K_{z^*} 表示当 $Z=z^*$ 时 K 的取值. 根据图 1 因果图，变量 Z 对 Y 的总体效应可以分解为自然直接效应（Natural Direct Effect，NDE）和总体间接效应（Total Indirect Effect，TIE），分别对应了有向路径 $Z \to Y$ 和 $Z \to K \to Y$.

Z 对 Y 的自然直接效应指的是在 $Z \to Y$ 的直接影响下，当变量 Z 的值从 z 变成 z^* 时 Y 的取值变化量，表示为

$$NDE = Y_{z,K_{z^*}} - Y_{z^*,K_{z^*}},$$

其中，$Y_{z,K_{z^*}} = Y(Z=z, K=K_{z^*}=K(Z=z^*))$. 由于 $Z=z$ 和 $Z=z^*$ 在现实世界中不可能同时出现，属于反事实样本，因此该设置只能通过反事实推理实现.

相应的，Z 对 Y 的总体间接效应指的是在 $Z \to K \to Y$ 的间接影响下 Z 对 Y 的因果效应，表示为

$$TIE = TE - NDE = Y_{z,K_z} - Y_{z,K_{z^*}}.$$

4　模型方法

本章节将介绍基于反事实推理的文本摘要去偏方法. 首先结合文本摘要的结构因果模型介绍基于反事实推理的摘要去偏方法，然后介绍摘要去偏方法的具体实现，最后介绍模型训练目标和摘要预测方法.

4.1　文本摘要的结构因果模型

根据上一章介绍的表示方法，我们首先定义了文本摘要任务的结构因果模型，包括描述各个关键要素之间因果关系的因果图，以及相应的结构方程. 然后结合上一章的因果效应理论，我们提出了基于反事实推理的摘要去偏方法.

首先，我们探究了文本摘要任务中原文、摘要和语言先验之间的因果关系. 由于我们主要探究基于预训练模型的深度文本摘要方法，所以除了原始文档和摘要之外，我们着重考虑了预训练模型在预训练过程中从大规模语料库学到的语言先验对摘要的影响.

在生成摘要的过程中，不仅需要根据原文提取关键信息，还需要利用语言先验将关键信息组织成通顺、流畅的自然语言文本作为摘要. 其中，语言先验在不同的情况下可能扮演不同角色：既可能是有用的知识，也可能是噪声. 为了明确地区分语言先验的不同角色，更好地分析噪声的成因，我们参考并模拟了人类完成文本摘要的过程. 人类在成长过程中不断学习和积累语言先验，比如一些高频共现的词对（医院–病人、香港–迪士尼、成都–大熊猫、内蒙古–骑马、阿根廷–梅西）. 一方面，语言先验能够帮助人类更流畅地表达；另一方面，语言先验有时会让人自行产生丰富的联想. 当人类构思一篇文章的摘要时，他需要结合文章的内容，并根据自己的语言先验合理组织语言. 如果他仔细阅读了全文，那么他更有可能准确提取出文章的信息，并用流畅的语言将其组织为通顺的摘要文本. 如果他没有仔细地阅读全文，比如只看了文章的开头就根据语言先验联想揣测出了文章的内容，那么此时构思的摘要很有可能包含与原文不匹配的信息. 该情形也可以用学生在答题时审题不充分的情况作类比：在题干没看完的情况下，受到题海战术训练的学生很容易被自己的“固定思维”误导，给出错误的答案.

在人类行为的启发下，我们相应地在摘要任务中，根据语言先验使用方式的不同划分其不同角色：根据是否与原文交互，显式地区分有用的知识和可能造成误导的噪声. 其中，模型的预训练过程相当于人类成长时的学习过程，语言先验指的是模型在预训练过程中学到的信息，存储于预训练语言模型的参数中. 当先验和原文充分交互时，它对摘要的

生成起到正面的引导作用. 而当它没有和原文交互时，对应的是负面的影响，属于噪声. 此时的先验也称为“刻板印象”，可能会导致生成的摘要中包含一些原文没有的信息，与正确的摘要存在偏差[16].

根据上述分析，我们定义了一个因果图来描述三者之间的因果关系，如图 2 所示. 图中 X 、Y 和 P 分别表示原文、摘要和语言先验，M 表示的是原文 X 和语言先验 P 的交互结果.

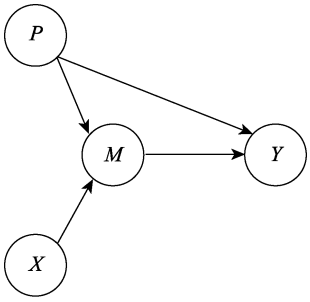

图 2 摘要的因果图

根据因果图我们可以看到，语言先验对摘要的影响通过两种方式作用：一是直接影响，对应了 $P \to Y$ 这条路径，二是间接影响，对应了 $P \to M \to Y$ 这条路径. $P \to Y$ 与 $P \to M \to Y$ 分别代表了语言先验对摘要的负面影响和正面影响. 综合两种影响方式，Y 的值可以表示成：

$$Y_{p,m} = Y(P = p, M = m).$$

其中，$m = M_{p,x} = M(P = p, X = x)$. 因此我们将 $Y_{p,m}$ 重新表示为 $Y_{p,M_{p,x}}$，x 代表给定的原始文档，p 代表语言先验，$M_{p,x}$ 代表先验 p 和原文 x 交互的结果.

进一步，我们结合具体的预训练模型 BART 分析模型所生成的摘要文本中偏差的来源. 在预训练的过程中，BART 模型学到了语言先验并存储在模型参数中. 然而，它没有显式地区分语言先验中的有用知识和噪声，在建模时将 $P \to Y$ 和 $P \to M \to Y$ 二者混在一起得到 $Y_{p,M_{p,x}}$. 所以在生成摘要时，既受到先验知识的正确指导，将原文的关键信息组合成流畅的自然语言，也受到先验在没有和原文充分交互时的直接影响，生成原文中没有的信息，导致了摘要中偏差的存在.

为了得到准确又流畅的摘要，我们只需要用到语言先验中的有用知识. 然而，我们难以从现有的预训练模型中单独提取有用的语言先验. 所以我们通过从语言先验的总体效应中减去自然直接效应的方式来得到总体间接效应. 这样就能尽可能的保留语言先验中有用知识的正面影响，同时剔除噪声的负面影响.

我们根据因果效应理论从语言先验对摘要的总体效应中剔除语言先验的自然直接效应，如图 3 所示. 首先，语言先验的总体效应可表示为

$$TE = Y_{p,M_{p,x}} - Y_{p^*,M_{p^*,x^*}}.$$

其中 * 表示对应的变量置为空值，不起作用. 自然直接效应可表示为

$$NDE = Y_{p,M_{p^*,x^*}} - Y_{p^*,M_{p^*,x^*}}.$$

此时，先验 p 只对摘要具有直接效应，间接效应通过为 p 和 x 赋空值被阻断. 由于在事实世界中，p 不可能同时被赋两种值，因此这种设定只会发生在反事实推理中.

然后，将总体效应减去先验对摘要的直接效应，便可得到先验对摘要的间接效应，可表示为

$$TIE = TE - NDE = Y_{p,M_{p,x}} - Y_{p,M_{p^*,x^*}}.$$

该等式即为摘要去偏方法的核心.

4.2 摘要去偏方法的具体实现

本文所设计的摘要去偏方法的具体实现如图 4

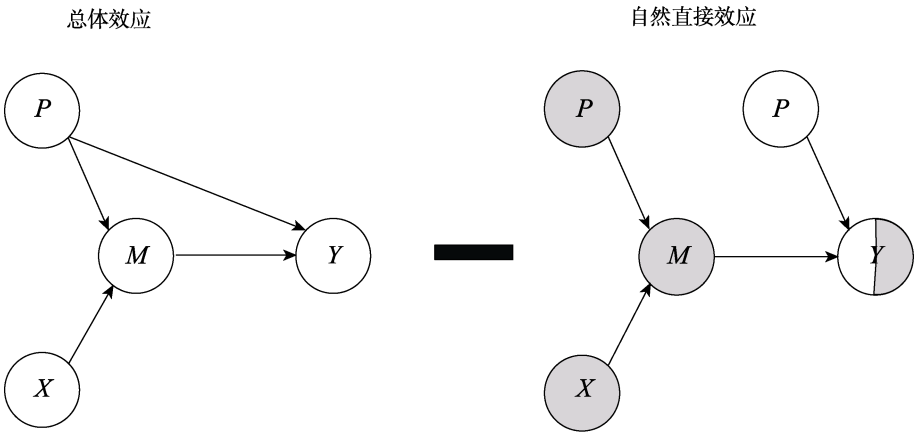

图 3 摘要去偏方法

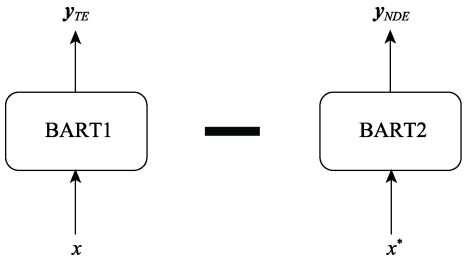

图 4　摘要去偏方法的具体实现

所示. 我们使用 BART 模型作为主体，并根据 TIE 等式完成去偏.

具体来说，我们的模型框架包含两个 BART 模型. 第一个模型用于建模语言先验对摘要的总体影响，对应图 2 因果图中 $P \to Y$ 和 $P \to M \to Y$ 两条边的作用，与传统的建模和使用方式保持一致. 第二个模型用于得到语言先验的直接影响，对应图 2 因果图中 $P \to Y$ 这条边的作用. 它们的输入分别为原始的原文和空字符串，即

$$\boldsymbol{y}_{TE} = \text{BART1}(x), \boldsymbol{y}_{NDE} = \text{BART2}(x^*),$$

其中，x^* 为空串，$\boldsymbol{y} \in R^{n \times v}$ 表示长度为 n 的摘要中每个位置在大小为 v 的词表上的分数分布. $\boldsymbol{y}_{TE}$ 是接收正常输入时所得到的分数分布，对应 TIE 等式中的 $Y_{p,M_{p,x}}$，$\boldsymbol{y}_{NDE}$ 是接收空串输入时所得到的分数分布，对应 TIE 等式中的 $Y_{p,M_{p^*,x^*}}$. 根据 TIE 等式，我们将两式相减，得到理想的分数分布：

$$\boldsymbol{y}_{TIE} = \boldsymbol{y}_{TE} - \boldsymbol{y}_{NDE}.$$

为了避免总体效应 $\boldsymbol{y}_{TE}$ 和自然直接效应 $\boldsymbol{y}_{NDE}$ 差距过大，导致 $\boldsymbol{y}_{TIE}$ 受其中一方的主导，我们引入去偏系数 $\boldsymbol{\lambda}$ 来平衡 $\boldsymbol{y}_{TE}$ 和 $\boldsymbol{y}_{NDE}$ 的大小：

$$\boldsymbol{y}_{TIE} = \boldsymbol{y}_{TE} - \boldsymbol{\lambda} \cdot \boldsymbol{y}_{NDE},$$

其中，$\boldsymbol{\lambda} \in R^v$，相当于对 $\boldsymbol{y}_{NDE}$ 按位调整大小，调节自然直接效应的影响效果.

在本工作中，我们将 $\boldsymbol{\lambda}$ 设置为可学习的向量. 具体来说，我们引入一个正则项，约束 $\boldsymbol{y}_{TE}$ 和 $\boldsymbol{\lambda} \cdot \boldsymbol{y}_{NDE}$ 两个分布之间的差距，避免其中一个分布的影响效果占主导地位. 当 $\boldsymbol{y}_{TE}$ 和 $\boldsymbol{\lambda} \cdot \boldsymbol{y}_{NDE}$ 之间的差距在适当范围内时，总体效应和调整过的自然直接效应的影响效果相当，此时得到的总体间接效应 $\boldsymbol{y}_{TIE}$ 更加合理. 正则项的设计见下一小节的 L_2.

4.3　摘要去偏模型的训练和预测

我们为摘要去偏模型设计了三个训练目标，分别用来拟合标准摘要、平衡语言先验的总体效应和自然直接效应的大小、以及从总体效应中区分自然直接效应和总体间接效应.

首先是用来拟合标准摘要的目标. 我们默认训练数据无偏，即参考摘要是无偏的，因此我们使用无偏的间接效应 $\boldsymbol{y}_{TIE}$ 去拟合无偏的参考摘要 $\boldsymbol{y}^G$，通过最小化交叉熵损失函数来实现：

$$L_1 = CE(\boldsymbol{y}^G, \boldsymbol{y}_{TIE}).$$

其次，为了得到合适的 $\boldsymbol{\lambda}$，用于平衡总体效应 $\boldsymbol{y}_{TE}$ 和自然直接效应 $\boldsymbol{y}_{NDE}$ 的影响效果，我们最小化 $\boldsymbol{y}_{TE}$ 和 $\boldsymbol{\lambda} \cdot \boldsymbol{y}_{NDE}$ 之间的 KL 散度：

$$L_2 = D_{KL}(\boldsymbol{y}_{TE} \| \boldsymbol{\lambda} \cdot \boldsymbol{y}_{NDE}) + D_{KL}(\boldsymbol{\lambda} \cdot \boldsymbol{y}_{NDE} \| \boldsymbol{y}_{TE}).$$

最小化 KL 散度能够使 $\boldsymbol{y}_{TE}$ 和 $\boldsymbol{\lambda} \cdot \boldsymbol{y}_{NDE}$ 两个分布相近，避免其中一个显著影响 $\boldsymbol{y}_{TIE}$. 这里我们同时使用正反向的 KL 散度来加强约束.

最后，为了更好地从总体效应中分辨直接效应和间接效应，我们最大化两个分数分布 $\boldsymbol{y}_{TIE}$ 和 $\boldsymbol{y}_{NDE}$ 之间的 KL 散度：

$$L_3 = -[D_{KL}(\boldsymbol{y}_{TIE} \| \boldsymbol{y}_{NDE}) + D_{KL}(\boldsymbol{y}_{NDE} \| \boldsymbol{y}_{TIE})].$$

综合三个目标，我们最终的目标函数为

$$L = L_1 + \beta L_2 + \gamma L_3.$$

其中，L_1 和 L_3 用于更新 BART 模型的参数，L_2 只用于更新去偏系数 $\boldsymbol{\lambda}$.

测试时，我们将去偏后的分数分布用于摘要预测，即根据 $\boldsymbol{y}_{TIE} = \boldsymbol{y}_{TE} - \boldsymbol{\lambda} \cdot \boldsymbol{y}_{NDE}$ 生成摘要文本.

5　实验和分析

本章节将介绍实验设置、实验结果以及对结果的相应分析.

5.1　实验设置

本节我们将介绍实验所采用的数据集、评价指标、对比的基线方法和模型的实现细节.

5.1.1　数据集

在本小节中，我们将介绍实验选用的两个数据集. 我们采用公开的摘要数据集 XSUM（eXtreme SUMmarization）[41]和 CNN/DailyMail[42]进行实验. 它们的统计信息如表 2 所示.

XSUM 是根据 BBC 的新闻自动构造的单句摘要数据集，将新闻的首句导语作为摘要，将新闻的其余部分作为原文. XSUM 属于生成式摘要数据集，与抽取式摘要数据集相比，该数据集的难度更大.

CNN/DailyMail 则收集了 CNN 和 Daily Mail 网站上的新闻作为原文，并将人工编写的多句新闻亮点作为摘要. CNN/DailyMail 属于抽取式摘要数据集，数据规模大于 XSUM.

表 2 数据集统计信息

数据集	训练集	验证集	测试集
XSUM	204,045	11,332	11,334
CNN/DailyMail	286,817	13,368	11,487

5.1.2 评价指标

在本小节中，我们将介绍实验采用的自动评价指标和人工评价指标.

我们采用 Rouge（Recall-Oriented Understudy for Gisting Evaluation）值作为自动评价指标. 作为经典的文本摘要任务评价指标，它通过计算参考摘要和生成摘要中共有 N 元词组（N-gram）的比例来衡量它们之间的匹配度. Rouge-N（RN）定义为

$$RN=\frac{\sum_{N-gram\in\{Reference\}}Count_{match}\left(N-gram\right)}{\sum_{N-gram\in\{Reference\}}Count\left(N-gram\right)}.$$

Rouge-L（RL）则对参考摘要和生成摘要的最长公共子序列（Longest Common Subsequence，LCS）计算匹配度，定义为

$$RL=\frac{LCS\left(Reference,Generated\right)}{m},$$

其中 m 是参考摘要的长度. 我们同时使用 Rouge-1（$R1$）、Rouge-2（$R2$）和 RL 三个指标来衡量生成摘要的质量.

为了更全面地评估模型生成的摘要文本质量，我们通过人工评价的方式评估摘要文本的流畅性和忠实性. 评价者共有两人，都是来自于计算机专业自然语言处理领域的研究生. 两人独立地对采样于两个数据集的各 30 个样本完成评估，每个样本提供<原文，参考摘要，我们的方法生成的摘要，基线方法生成的摘要>. 我们将摘要文本的流畅性和忠实性划分为四个等级，分别为“优秀（4 分）”、“良好（3 分）”、“一般（2 分）”、“差（1 分）”，并取分值的平均值作为最终评价结果. 下面说明流畅性和忠实性的评价标准：

（1）流畅性评价的是生成文本的通顺程度. 具体来说，“优秀”代表生成的文本非常通顺，完全符合人类语言习惯，比如语序正确、无重复表达；“良好”代表生成的文本基本通顺，只存在一处不符合人类语言习惯的表达，但总体可读性良好，不影响理解；“一般”代表生成的文本不够通顺，比如存在部分语序错误，包含一些重复的、无意义的表达，影响理解；“差”代表生成的文本语序混乱，可读性极差.

（2）忠实性评价的是生成文本忠于原文的程度. 具体来说，“优秀”代表生成的文本完全忠于原文，不存在任何原文之外的信息或是与原文不相符的信息，即不存在幻觉现象；“良好”代表生成的文本基本忠于原文，只存在一处无关紧要的幻觉现象，比如存在原文之外但相对合理或符合常识的信息；“一般”代表生成的文本存在个别不合理的幻觉现象，与原文信息明显不符；“差”代表生成的文本存在严重的幻觉现象，多处违背原文信息.

5.1.3 对比基线方法

在本小节中，我们将介绍实验采用的基线方法. 我们使用经典的摘要算法以及基于预训练的先进摘要算法作为基线方法，包括抽取式方法和生成式方法. 现分别介绍如下：

（1）LEAD/LEAD-3[43]将一篇文章的前几个句子作为摘要. 这是一种经典的抽取式摘要方法. 这里 LEAD 和 LEAD-3 分别指的是选择文章中的第一个句子和前三个句子作为摘要. 它们作为不同数据集的基线方法，对应了两个数据集各自的摘要特点：LEAD 用于 XSUM，对应了 XSUM 数据集中摘要只由一句话构成；LEAD-3 用于 CNN/DailyMail，对应了 CNN/DailyMail 数据集中摘要由多句话构成.

（2）PTGEN[44]将长短时记忆网络结合序列到序列模型，并通过指针机制选择原文词语复制到输出摘要，是一种生成机制和抽取机制相结合的方法. PTGEN+COV 则是该方法结合覆盖机制（Coverage）的变种，其中覆盖机制是为了防止重复文本的生成.

（3）CONVS2S[45]将卷积网络与序列到序列模型相结合.

（4）HBF[46]通过检测“事实幻觉”作为离线强化学习的奖励信号，以提升摘要文本的忠实性，其中“事实幻觉”指的是与原文不一致但符合事实的幻觉现象.

（5）BART[6]是将 Transformer 结构与序列到序列架构相结合的预训练模型，并通过多种基于去噪自编码器的预训练目标使得模型获得更通用的知识. 作为一种采用编码器–解码器架构的通用的预训练语言模型，它常用于各种自然语言生成的下游任务.

5.1.4 模型实现细节

在本小节中，我们将介绍模型的实现细节，包括主干模型、模型架构、优化器的选择、训练超参数的设置等.

我们使用 $\text{BART}_{\text{Large}}$ 作为主干模型，其隐变量的维度大小 1024，编码器和解码器都由 12 层

Transformer 堆叠而成，每一层使用的多头注意力机制共有 16 个头. 使用 BART 附带的原始词表，大小 v 为 50265. 我们基于公开的 Transformers 代码库 huggingface 实现我们的代码，其余未说明的参数全部遵循该代码库相应模块的默认参数设置.

训练时，我们使用 AdamW 优化器，并设置学习率为 5e-5，权重衰减率为 0.01；使用的训练批次大小为 8，并在累计计算 32 次后做一次梯度更新；warmup 的步数设置为 500. 对于 XSUM 数据集，我们设置超参数 β 为 1e-6，γ 为 1e-3；对于 CNN/DailyMail 数据集，我们设置超参数 β 为 1e-5，γ 为 5e-4. 我们在一张 NVIDIA Tesla 的 32GB V100 GPU 上训练模型，收敛步数在 15 k~25 k 之间. 所有超参数的设置都是基于公开数据集自带的验证集完成的.

5.2　自动评价结果分析

本小节中，我们将对比我们的方法和基线方法在两个数据集 XSUM 和 CNN/DailyMail 上的自动评价结果.

我们在生成式摘要数据集 XSUM 上得到的实验结果如表 3 所示. 由表 3 可知，抽取式方法 LEAD 和附带指针机制的 PTGEN 表现不佳，说明通过直接摘录原文首句的抽取式摘要方法和基于指针机制复制的摘要方法不适用于该生成式摘要数据集，可以看出基于抽取机制的摘要方法因为严格受限于原文的表达，存在一定的局限性. 基于卷积网络的 CONVS2S 利用自身的局部信息提取与整合能力表现略优于前两种方法，然而对于远距离信息的捕捉能力仍然不足，因此增益有限. BART 模型作为近两年通用的预训练模型，表现明显优于前几种基线方法，表明了 Transformer 结构的强大之处以及预训练的有效性. 我们的方法对 BART 模型去偏，又进一步超过了 BART 模型的表现，在 *R*1、*R*2 和 *RL* 三个指标上分别获得了 0.75%、0.54%和 0.46%的提升，侧面印证了 BART 模型中存在语言先验对摘要的负面影响，而去除该负面影响的模型能更恰当地拟合数据集. 此外，基于事实幻觉的检测信号来指导强化学习的 HBF 在 R1 和 RL 指标上的性能表现不如我们的方法，说明我们基于因果理论所设计的摘要去偏方法与其他提高忠实性的方法相比，更好地保留了拟合数据集的能力，表明基于反事实推理的去偏方法能够更有效地利用语言先验的有用知识去学习真实世界的可观测样本.

表 3　XSUM 实验结果对比

方法	*R*1	*R*2	*RL*
LEAD	16.30	1.60	11.95
PTGEN	29.70	9.21	23.24
CONVS2S	31.27	11.07	25.23
HBF	44.60	–	36.20
BART	45.14	22.27	37.25
OURS	**45.48**	**22.39**	**37.42**

我们在抽取式摘要数据集 CNN/DailyMail 上得到的实验结果如表 4 所示. 由表 4 可以看到经典抽取式摘要方法 LEAD-3 在该数据集上表现优秀，甚至超过了基于神经网络的 PTGEN 模型. 从该结果可以看出，该数据集的摘要存在明显的位置偏差，即，有相当一部分摘要直接摘自原文的开头部分. 生成机制和抽取机制相结合的基线方法 PTGEN 稍弱于 LEAD-3，而 PTGEN+COV 在 PTGEN 的基础上增加覆盖机制，有效避免了生成内容的重复问题，使模型性能获得了提升. BART 模型在该数据集上同样表现优秀，体现了其优秀的泛化能力和高度的适用性. 而我们的方法又能进一步超过 BART 的表现，在 R1、R2 和 RL 三个指标上分别获得了 1.29% 、2.08%和 1.20%的提升，说明我们的方法对于抽取式数据集也能有效去偏.

表 4　CNN/DailyMail 实验结果对比

方法	*R*1	*R*2	*RL*
LEAD-3	40.42	17.62	36.67
PTGEN	36.44	15.66	33.42
PTGEN+COV	39.53	17.28	36.38
BART	42.78	20.21	39.90
OURS	**43.33**	**20.63**	**40.38**

5.3　人工评价结果分析

本小节中，我们将对比我们的方法和基线方法在两个数据集 XSUM 和 CNN/DailyMail 上的人工评价结果.

由表 5 可知，在 XSUM 数据集上，我们的方法和 BART 的流畅性基本持平，而在忠实性的表现上，我们的方法会优于 BART. 评价者的一致性通过 Kappa 值来衡量，其平均值为 0.49. CNN/DailyMail 上的结论基本一致. 该实验结果说明了我们设计的去偏方法不仅能够有效利用先验知识，生成流畅的

表 5　人工评价结果对比

方法	XSUM		CNN/DailyMail	
	流畅性	忠实性	流畅性	忠实性
BART	**3.92**	3.18	3.77	3.68
OURS	3.87	**3.33**	**3.78**	**3.75**

摘要文本，还能够有效规避噪声的影响，减少 BART 模型生成的摘要文本中存在的偏差.

5.4 消融实验分析

为了更加细致地验证我们为摘要去偏模型所设计的不同训练目标的必要性，在本节中，我们在 XSUM 数据集上开展了针对性的消融实验，分别探究了学习目标 L_2 和 L_3 的影响.

首先，为了探究学习目标 L_2 的影响，以及探究平衡语言先验的总体效应和自然直接效应的大小的必要性，我们尝试将学习目标 L_2 的系数 β 设置为 0 来完成相应的消融实验. 此时，我们不使用学习目标 L_2 更新 λ，而是将其设置为固定的超参数. λ 的不同取值及相应的实验结果如表 6 所示. 由表 6 可以看到：（1）不使用 L_2 更新 λ，而是将 λ 固定为人工设定的超参数时，模型的性能明显下降. 以 R1 指标为例，当 λ 的取值固定为 1 的时候，模型的性能下降了 13.30%；（2）当 λ 的取值不同时，模型性能的差异较大. 以 R1 指标为例，λ 的取值为 5 时，模型性能比 λ 的取值为 1 时下降了 27.06%.

学习目标 L_2 的消融实验结果表明了通过学习目标 L_2 使 λ 自动调整至合适的数值的必要性，它有助于模型更好地生成摘要. 其中潜在的好处在于：（1）通过自学习的方式调整 λ 向量，可以做到为词表中不同的词赋予不同的去偏权重，更灵活且有针对性.（2）模型对该参数较为敏感，人工赋值时难以确定最优权重. 若要为不同的词赋予不同的去偏权重，则需要依赖精心设计的启发式方法，如根据每个词的 Tf-idf 值赋值，对设计者是一个更大的考验. 而自动学习参数的方式避免了设计去偏系数的费时费力.（3）调节语言先验的总体效应和自然直接效应的影响有效避免了生成摘要受其中一方的主导. 当受总体效应主导时，模型可能同基线方法一样，受到语言先验中噪声的干扰；当受自然直接效应主导时，相当于生成的摘要几乎由语言先验中的噪声（负值）所决定. 由于语言先验中的噪声无法感知原始输入文档，所以此时的模型不具备概括输入文档的能力，严重时甚至可能破坏模型原本具备的流畅表达能力.

表 6 学习目标 L_2 的消融实验结果

方法	*R*1	*R*2	*RL*
OURS	45.48	22.39	37.42
- L_2 (λ =1)	39.43	18.57	32.42
- L_2 (λ =5)	28.76	10.94	23.29

然后，为了探究学习目标 L_3 的影响，以及探究从总体效应中分辨直接效应和间接效应的必要性，我们尝试将学习目标 L_3 的系数 γ 设置为 0 来完成相应的消融实验. 相应的实验结果如表 7 所示. 由表 7 可以看到：当不使用 L_3 作为训练目标时，模型的性能下降. 以 R1 指标为例，不使用训练目标 L_3 时，模型的性能下降了 8.53%.

表 7 学习目标 L_3 的消融实验结果

方法	*R*1	*R*2	*RL*
OURS	45.48	22.39	37.42
- L_3	41.60	19.50	33.95

学习目标 L_3 的消融实验结果表明了通过学习目标 L_3 来区分总体效应中的间接效应和直接效应的必要性. 它提升模型性能的潜在原因是：（1）区分间接效应和直接效应对应的分布有助于更好地区分先验中的知识和噪声，帮助摘要生成所依赖的先验知识规避噪声的影响；（2）同时，确保想要消除的负面影响没有夹带有用的先验知识. 若是两者分布趋同，在扣除直接效应的时候可能导致有用信息的损耗. 下一节将对此做进一步分析.

5.5 关键参数分析

为了更好地获知关键参数对模型性能的影响，并进一步分析各学习目标对模型学习的效果，在本节中，我们在 XSUM 数据集上对学习目标 L_2 的权重 β 和学习目标 L_3 的权重 γ 开展分析实验.

首先，我们调整了学习目标 L_2 的权重 β，影响效果如表 8 所示. 实验结果表明，模型对 β 值的设置较为敏感，在相当小的数量级（1e-7）上调整 β 值仍然会对实验结果有明显影响. 在小范围内，随着 β 的增大，模型的性能表现提升. 以 R1 指标为例，当 β 值从 1e-7 增加为 8e-7 时，模型性能提升了 5.46%，而当 β 值增加为 1e-6 时，模型性能又进一步提升了 10.12%.

该分析实验说明了学习目标 L_2 对模型训练起到正向作用，进一步说明了使用学习目标 L_2 来平衡语言先验的总体效应和自然直接效应的必要性.

然后，我们调整了学习目标 L_3 的权重 γ，影响

表 8 调整 β 的影响效果

β 的取值	*R*1	*R*2	*RL*
β =1e-7	39.16	17.94	32.23
β =8e-7	41.30	19.61	34.03
β =1e-6	**45.48**	**22.39**	**37.42**

效果如表 9 所示. 实验结果表明，当 γ 值由 5e-4 增大到 2e-3 时，模型会在中间取得最优表现；而当 γ 值在 1e-3 的数量级上增加时，模型的性能表现会显著下降.

从该分析实验可以看出：（1）适当地区分直接效应和间接效应有助于模型更好地完成去偏，也说明了使用学习目标 L_3 来分辨直接效应和间接效应的必要性；（2）即使在不充分与原文交互的情况下，语言先验的直接效应也可能恰好生成正确的摘要文本，因此它得到的词表分数分布 y_{NDE} 未必会和 y_{TIE} 显著不同. 若是强硬地区分两个分数分布，结果可能适得其反.

对两个关键参数 β 和 γ 的分析可以看到超参数的调整会显著影响实验结果，从一定程度上说明了精准去偏的难度. 其主要难点在于，知识和噪声是语言先验所扮演的不同角色. 本文中的语言先验指的是模型在预训练过程中学到的信息，存储于预训练语言模型的参数中，所以“知识”和“噪声”原本是一体的，都源自于模型参数. 合理区分不同角色需要非常精细的方法设计.

表 9　调整 γ 的影响效果

γ 的取值	*R*1	*R*2	*RL*
γ =5e-4	42.86	20.47	35.17
γ =1e-3	**45.48**	**22.39**	**37.42**
γ =2e-3	23.35	9.76	20.15

5.6　生成摘要样例分析

为了更直观地感知我们模型的去偏能力，并观察生成摘要的效果，在本节中，我们在 XSUM 数据集上对比和分析了生成摘要的样例.

我们首先统计了不同模型生成的摘要中，与参考摘要完全一致的样例个数. BART 模型能够生成 18 个与参考摘要完全一致的摘要，而我们的摘要去偏模型能够生成 31 个，说明我们的摘要去偏模型具备更好地拟合数据集的能力.

然后我们通过分析一些生成的摘要样例，说明我们的模型确实能有效消除 BART 模型生成的摘要文本中存在的偏差. 表 10 展示了一些用我们的方法生成的摘要样例，并与 BART 模型生成的摘要对比. 其中，下划线标识的文本包含了原始文档中未出现的信息. 例如，在表 10 展示的样例一中，原始文档中并未出现与“美国（America）”或“洛杉矶（Los Angeles）”相关的信息，BART 模型却错误地生成了这样的文本（在表中用下划线粗体标识），而我们的方法能正确消除这样的偏差. 在表 10 展示的样例二中，原始文档中并未出现“第二周蝉联（has topped… for the second week）”的信息，BART 却出现了这样的错误（在表中用下划线粗体标识），我们的模型则能有效纠正这样的错误. 样例三和样例四也有类似的结论. 这些例子直观展现了我们模型的去偏能力. 与此同时，我们可以从样例一中看到，参考摘要中和我们模型所生成的摘要都包含了“the world's biggest convention for games”这个说法，然而原始文档中并没有提及该信息. 对于该现象，如果我们的评价准则是摘要信息对原文信息的忠实性，那么只要摘要中出现原文之外的信息或是与原文不相符的信息，都属于幻觉. 在未引入外部知识的情况下，我们通常采用该评价准则. 也就是说，即使是专用于摘要任务的数据集中，也可能包含属于幻觉的有偏数据. 对于该情况，我们将会在未来工作中从两个方面加以分析和改进：（1）适当引入外部知识，进一步判断摘要中原文以外的信息的准确性；（2）

表 10　生成摘要样例

样例	类型	内容
一	原始文档	Media playback is unsupported on your device 18 June 2015 Last updated at 15:10 BST All the big gaming companies including Microsoft, Nintendo and Sony are showing off what they've been working on. Virtual reality usually involves putting on some glasses that contain video screens. The specs give you the impression you're in a totally different world, that changes as you look around the room. For years, virtual reality has been tipped as the next big thing in technology - but now it looks like it might finally be ready for action. Radio 1 reporter Steffan Powell sent us this report from E3. 包括微软、任天堂和索尼在内的所有大型游戏公司都在展示他们正在研究的东西. 虚拟现实通常包括戴上一些包含视频屏幕的眼镜. 眼镜会给你一种置身于一个完全不同的世界的感觉，随着你环顾四周，它会发生变化. 多年来，虚拟现实一直被认为是技术领域的下一件大事，但现在它似乎终于可以付诸行动了. Radio 1 记者 Steffan Powell 从 E3 发回报道.
	参考摘要	Gaming experts are predicting a big future for virtual reality at E3 - the world's biggest convention for games. 在全球最大的游戏展 E3 上，游戏专家预测虚拟现实的未来前景广阔.
	BART 摘要	The E3 gaming show is **taking place in Los Angeles, America,** and it's all about virtual reality. E3 游戏展在美国洛杉矶举行，主题是虚拟现实.
	我们的摘要	Gaming experts are predicting a big future for virtual reality at E3 - the world's biggest convention for games. 在全球最大的游戏展 E3 上，游戏专家预测虚拟现实的未来前景广阔.

续表 10 生成摘要样例

样例	类型	内容
二	原始文档	The track broke streaming records, earning the highest first-week streams for a single in the UK with more than 2.03 million plays. Drag Me Down is One Direction's first single as a foursome since band mate Zayn Malik left and has had more than 17.7 million streams to date globally. The band shot the video for 'Drag Me Down' at NASA's headquarters in Houston yesterday. According to the Official Charts Company, Calvin Harris and Disciples' How Deep Is Your Love climbed from six to two and last week's number one, Black Magic by Little Mix, fell to three. 这首歌打破了英国的流媒体播放记录，首周播放量超过 203 万次，创下了单曲播放量的最高纪录.《Drag Me Down》是单向乐队(One Direction)自队友泽恩·马利克(Zayn Malik)离开后四人组合的第一首单曲，迄今为止在全球的播放量已超过 1770 万次. 这支乐队昨天在休斯顿的 NASA 总部拍摄了《Drag Me Down》的视频. 根据官方排行榜公司的数据，加尔文·哈里斯和门徒乐队的《How Deep Is Your Love》从第六名攀升至第二名，上周排名第一的 Little Mix 乐队的《Black Magic》则跌至第三名.
	参考摘要	One Direction have topped the UK singles chart with their new single Drag Me Down. 单向乐队凭借新单曲《Drag Me Down》登顶英国单曲排行榜.
	BART 摘要	One Direction's new single Drag Me Down **has topped** the UK singles chart **for the second week** in a row. 单向乐队的新单曲《Drag Me Down》已经连续第二周蝉联英国单曲排行榜冠军.
	我们的摘要	One Direction have topped the UK singles chart with their new single Drag Me Down. 单向乐队凭借新单曲《Drag Me Down》登顶英国单曲排行榜.
三	原始文档	Antares, built by Orbital Sciences Corp, combusted seconds after leaving the seaside launch pad at Wallops Flight Facility. The cause of the cargo ship malfunction has yet to be determined. The initial planned launch of the spacecraft on Monday was delayed due to a yacht in the surrounding danger zone. The flight was expected to be the third contracted mission with the National Aeronautics and Space Administration. The rocket was due to carry nearly 5,000lb (2,200kgs) of supplies to six astronauts aboard the International Space Station (ISS). It included equipment for astronauts to conduct tests on blood flow to the human brain and to analyse meteors. There was also equipment for experiments to examine the growth of pea shoots in orbit and how the body's immune system reacts to space travel. More than 1,300lb (600kg) of food was on board, including pre-packaged meals and freeze-dried crab cakes. 安塔瑞斯号由轨道科学公司建造，在离开沃勒普斯飞行设施的海边发射台几秒钟后起火. 货轮故障的原因还没有确定. 由于一艘游艇在附近的危险区，原定于周一发射的航天器被推迟. 这次飞行预计将是与美国国家航空航天局签订的第三次任务. 这枚火箭将为国际空间站(ISS)的 6 名宇航员携带近 5000 磅(2200 公斤)的物资. 它包括宇航员用于测试人类大脑血液流动和分析流星的设备. 还有用于实验的设备，以检查轨道上豌豆苗的生长，以及人体免疫系统对太空旅行的反应. 机上载有超过 1300 磅(600 公斤)的食品，包括预包装的餐食和冻干蟹肉饼.
	参考摘要	An unmanned supply rocket bound for the International Space Station has exploded shortly after its launch from the US state of Virginia. 一枚飞往国际空间站的无人补给火箭从美国弗吉尼亚州发射后不久发生爆炸.
	BART 摘要	A rocket carrying food and experiments to the International Space Station has crashed in the **Atlantic Ocean** off the coast of Virginia. 一枚向国际空间站运送食物和实验用品的火箭在弗吉尼亚州海岸附近的大西洋坠毁.
	我们的摘要	A rocket carrying food and experiments for the International Space Station has exploded shortly after blasting off from a Virginia launch pad. 一枚为国际空间站运送食物和实验用品的火箭从弗吉尼亚州发射台发射升空后不久发生爆炸.
四	原始文档	“Fracking” involves pumping water and chemicals into shale rock at pressure. The joint report from the Royal Society and Royal Academy of Engineering say the technique is safe if firms follow best practice and rules are enforced. Exploratory fracking is being mooted in at least seven sites around the UK. The report was commissioned by the government's chief scientist, Sir John Beddington, following the decision last year to halt the UK's most advanced project, in Lancashire, after fracking caused small earth tremors. “Our main conclusions are that the environmental risks of hydraulic fracturing for shale can can be safely managed provided there is best practice observed and provided it's enforced through strong regulation,” said the report's chair… “水力压裂”是指在压力下将水和化学物质泵入页岩中. 英国皇家学会和英国皇家工程院的联合报告称，如果公司遵循最佳实践，并执行规则，这种技术是安全的. 在英国至少有 7 个地方正在讨论试探性水力压裂. 该报告是由政府首席科学家约翰• 贝丁顿爵士(Sir John Beddington)委托撰写的. 去年，在水力压裂法引发小规模地震后，英国决定停止位于兰开夏郡(Lancashire)的最先进项目. “我们的主要结论是，页岩水力压裂的环境风险是可以安全管理的，只要有最佳实践观察，并通过强有力的监管加以执行，”报告主席说……
	参考摘要	A gas extraction method which triggered two earth tremors near Blackpool last year should not cause earthquakes or contaminate water but rules governing it will need tightening, experts say. 专家表示，去年在布莱克浦附近引发两次地震的一种气体抽取方法，应该不会引发地震或污染水源，但管理该方法的规定需要加强.
	BART 摘要	Shale gas extraction can be carried out safely in the UK, but stronger regulations are needed to protect **public health** and the environment, a report says. 一份报告称，页岩气开采可以在英国安全进行，但需要更严格的法规来保护公众健康和环境.
	我们的摘要	Shale gas extraction in the UK can be safely managed, according to a report commissioned by the government. 根据政府委托的一份报告，英国的页岩气开采是可以安全管理的.

对于确实违背事实的有偏数据集，探究如何准确地去除偏差，生成合理的摘要. 此外，从样例四中可以看到，虽然我们的方法能够有效地消除原文以外的信息，却也有一定概率会丢失原文中的重要信息“但管理该方法的规定需要加强（but rules governing it will need tightening）”. 其可能原因在于，我们的模型仍然在一定程度上混杂了语言先验中有用的知识和噪声，这表明我们的方法还需进一步确保噪声和有用知识的解耦.

6　总结和展望

本文作为我们结合因果理论设计文本摘要模型的初步尝试，首先通过因果视角分析了基于预训练模型的深度文本摘要方法中存在的偏差，并将偏差归因于语言先验中的噪声，然后提出基于反事实推理的摘要去偏方法. 在 XSUM 和 CNN/DailyMail 数据集上的实验验证了本文所提方法的有效性.

本文所提出的摘要去偏方法是一个通用的框架，适用于各种不同的摘要模型. 后续我们还将尝试 BART 之外的模型作为主体，进一步验证去偏方法的有效性和普适性.

在未来工作中，我们还将探索如何更好地从语言先验对摘要的总体效应中区分直接效应和间接效应. 此外，我们受人类行为的启发，在本工作中根据是否与原文交互显式地区分知识和噪声，具备一定的可解释性. 除了这一种可能的偏差来源，我们将从其他角度更细致地探索其他偏差的来源. 例如，训练数据集中存在的具体社会性偏见，包括性别偏见、职业偏见等，以及训练过程中由于优化目标设置不合理带来的偏差. 通过更全面地考虑不同偏差的来源，我们将设计更完善的因果图用于去偏工作；也将尝试从去偏之外的其他角度，如摘要任务背后的数据生成的角度，来探索因果图的设计，进一步学习知识和噪声更精细的解耦表达.

此外，如何提供可靠的理论证明，如可识别性证明，并更准确地计算出不同因素对摘要的影响效果仍然是一个开放的挑战. 我们希望我们的工作能给文本摘要领域，以及其他文本生成领域带来一些启示，激发更多基于因果理论和因果关系的工作，增加该领域工作的可解释性.

致　谢　该工作得到国家自然科学基金项目（No. 62006218, 61902381）、中国科学院青年创新促进会（No. 20144310, 2021100）、中国科学技术协会青年人才托举工程（No. YESS20200121）、联想–中科院联合实验室青年科学家项目资助.

参　考　文　献

[1] Zhang R, Guo J, Fan Y, et al. Structure learning for headline generation//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 9555-9562

[2] Xiao W, Carenini G. Extractive summarization of long documents by combining global and local context//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 3011-3021

[3] Sun F, Jiang P, Sun H, et al. Multi-source pointer network for product title summarization//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Torino, Italy, 2018: 7-16

[4] Mihalcea R, Tarau P. Textrank: Bringing order into text//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain, 2004: 404-411

[5] Liu Y, Lapata M. Text summarization with pretrained encoders//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 3730-3740

[6] Lewis M, Liu Y, Goyal N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 7871-7880

[7] Zhu C, Hinthorn W, Xu R, et al. Enhancing factual consistency of abstractive summarization//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Mexico City, Mexico, 2021: 718-733

[8] Shuster K, Poff S, Chen M, et al. Retrieval augmentation reduces hallucination in conversation//Findings of the Association for Computational Linguistics: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 2021: 3784-3803

[9] Lyu X, Li J, Gong Z, et al. Encouraging lexical translation consistency for document-level neural machine translation//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 2021: 3265-3277

[10] Nallapati R, Zhai F, Zhou B. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents//Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 3075-3081

[11] Zhou Q, Yang N, Wei F, et al. Neural document summarization by jointly learning to score and select sentences//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia,

2018: 654-663

[12] Rush A M, Chopra S, Weston J. A neural attention model for abstractive sentence summarization//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 379-389

[13] Zheng C, Cai Y, Zhang G, et al. Controllable abstractive sentence summarization with guiding entities//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain, 2020: 5668-5678

[14] Zhong M, Liu P, Chen Y, et al. Extractive summarization as text matching//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6197-6208

[15] Zhang J, Zhao Y, Saleh M, et al. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization//International Conference on Machine Learning. 2020: 11328-11339

[16] Xie Y, Sun F, Deng Y, et al. Factual consistency evaluation for text summarization via counterfactual estimation//Findings of the Association for Computational Linguistics: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 2021: 100-110

[17] Nan F, Nallapati R, Wang Z, et al. Entity-level factual consistency of abstractive text summarization//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 2727- 2733

[18] Cao M, Dong Y, Wu J, et al. Factual error correction for abstractive summarization models//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020: 6251-6258

[19] Cao S, Wang L. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 2021: 6633-6649

[20] Pasunuru R, Guo H, Bansal M. Towards improving abstractive summarization via entailment generation//Proceedings of the Workshop on New Frontiers in Summarization. Copenhagen, Denmark, 2017: 27-32

[21] Joachims T, Swaminathan A, Schnabel T. Unbiased learning-to-rank with biased feedback//Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. London, UK, 2017: 781-789

[22] Wei T, Feng F, Chen J, et al. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 1791-1800

[23] Niu Y, Tang K, Zhang H, et al. Counterfactual vqa: A cause-effect look at language bias//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12700-12710

[24] Zhang W, Lin H, Han X, et al. De-biasing distantly supervised named entity recognition via causal intervention//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 4803-4813

[25] Dudık M, Langford J, Li L. Doubly robust policy evaluation and learning//Proceedings of the 28th International Conference on International Conference on Machine Learning. Bellevue, USA, 2011: 1097-1104

[26] Zhang Y, Feng F, He X, et al. Causal intervention for leveraging popularity bias in recommendation//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 11-20

[27] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality//Proceedings of the 26th International Conference on Neural Information Processing Systems. Nevada, USA, 2013: 3111-3119

[28] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1532-1543

[29] Peters M, Neumann M, Iyyer M, et al. Deep contextualized word representations//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, USA, 2018: 2227-2237

[30] Mnih A, Hinton G E. A scalable hierarchical distributed language model//Proceedings of the 21st International Conference on Neural Information Processing Systems. Vancouver, Canada, 2008: 1081-1088

[31] Kiros R, Zhu Y, Salakhutdinov R R, et al. Skip-thought vectors //Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada, 2015: 3294-3302

[32] Le Q, Mikolov T. Distributed representations of sentences and documents//International Conference on Machine Learning. Beijing, China, 2014: 1188-1196

[33] Hill F, Cho K, Korhonen A. Learning distributed representations of sentences from unlabelled data//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA, 2016: 1367-1377

[34] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100,000+ questions for machine comprehension of text//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 2383-2392

[35] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 1631-1642

[36] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. 2018

[37] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies. Minneapolis, USA, 2019: 4171-4186
[38] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 5753-5763
[39] Joshi M, Chen D, Liu Y, et al. Spanbert: Improving pre-training by representing and predicting spans//Transactions of the Association for Computational Linguistics. Cambridge, USA, 2020: 64-77
[40] Ma X, Guo J, Zhang R, et al. Prop: pre-training with representative words prediction for ad-hoc retrieval//Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021: 283-291
[41] Narayan S, Cohen S B, Lapata M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 1797-1807
[42] Hermann K M, Kocisky T, Grefenstette E, et al. Teaching machines to read and comprehend//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada, 2015: 1693-1701
[43] Nenkova A. Automatic text summarization of newswire: Lessons learned from the document understanding conference//Proceedings of the 20th National Conference on Artificial Intelligence. Pittsburgh, USA, 2005: 1436-1441
[44] See A, Liu P J, Manning C D. Get to the point: Summarization with pointer-generator networks//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 1073-1083
[45] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning//International Conference on Machine Learning. Sydney, Australia, 2017: 1243-1252
[46] Cao M, Dong Y, Cheung J C K. Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland, 2022: 3340-3354

CHEN Lu, Ph. D. candidate. Her research interests include natural language processing and information retrieval.

ZHANG Ru-Qing, Ph. D., associate professor. Her research interests include natural language processing.

GUO Jia-Feng, Ph. D., professor. His research interests include data mining and information retrieval.

FAN Yi-Xing, Ph. D., associate professor. His research interests include data mining and information retrieval.

Background

Text summarization is an important problem in natural language processing, which aims to produce a fluent and condensed summary for a document, while preserving the core information. Recently, thanks to the advance of deep learning technology, deep neural networks have achieved promising results in text summarization, which automatically extract effective features from data and generate the summary in an end-to-end manner. Nevertheless, the mainstream of deep summarization models focuses on exploiting the correlations rather than the causal relationships. Among such correlations, there can be spurious ones which suffer from the language prior learned from the training corpus and therefore mislead the model into generating summaries with inaccurate expression and information deviation from the original document.

To address this problem, this paper analyzes the source of bias, which exists in deep summarization methods based on pre-trained models, from the perspective of causality. This paper firstly explores the causal structure of text summarization task, and then proposes a text summarization debiasing method based on counterfactual inference. The experiments on XSUM and CNN/DailyMail datasets show that the proposed model performs better than BART and other baselines.

This work is a preliminary attempt to design a text summarization model based on causal theory. We hope our work can shed some light into the community and motivate new causal ideas.

This work was funded by the National Natural Science Foundation of China (NSFC) under Grants Nos. 62006218 and 61902381, the Youth Innovation Promotion Association CAS under Grants Nos. 20144310, and 2021100, the Young Elite Scientist Sponsorship Program by CAST under Grants No. YESS20200121, and the Lenovo-CAS Joint Lab Youth Scientist Project.