基于因果推断在消除自然语言处理（NLP）模型偏差中的研究进展

赵剑波[1]，何琪[2]，闫懿涛[1]，徐龙[1]
1. 南宁师范大学 广西人机交互与智能决策重点实验室，广西 南宁 530100
2. 南宁市大学东路小学，广西 南宁 530000
DOI: 10.61369/SDME.2025090038

摘　　要： 近年来，随着自然语言处理（NLP）技术的迅猛发展，大规模预训练模型在诸多实际场景中取得了显著成就，但与此同时，其内在偏差问题也日益凸显，严重影响了模型在公平性、鲁棒性以及社会伦理等方面的应用。传统的偏差消除方法主要依靠数据增广、微调及人工标注等手段，虽然在一定程度上缓解了偏差问题，但往往难以从根本上消除模型所固有的系统性偏见，且在处理复杂语义和跨领域任务时存在适用性不足。本文聚焦于因果推断在 NLP 模型偏差消除中的应用，系统梳理了基于因果干预和反事实推理的最新研究成果。

关 键 词： 因果推断；反事实推理；因果干预；NLP 偏差；模型公平性

Research Progress on Eliminating Bias in Natural Language Processing (NLP) Models Based on Causal Inference

Zhao Jianbo[1], He Qi[2], Yan Yitao[1], Xu Long[1]
1.Guangxi Key Laboratory of Human-Computer Interaction and Intelligent Decision Making, Nanning Normal University, Nanning, Guangxi 530100
2.University East Road Primary School of Nanning, Nanning, Guangxi 530000

Abstract: In recent years, with the rapid development of natural language processing (NLP) technology, large-scale pre-trained models have achieved remarkable results in many practical scenarios. However, their inherent bias issues have become increasingly prominent, seriously affecting the application of these models in terms of fairness, robustness, and social ethics. Traditional bias elimination methods mainly rely on data augmentation, fine-tuning, and manual annotation. Although they alleviate bias to a certain extent, they often fail to fundamentally eliminate the inherent systemic bias of the models and have insufficient applicability when dealing with complex semantics and cross-domain tasks. This paper focuses on the application of causal inference in eliminating bias in NLP models and systematically reviews the latest research results based on causal intervention and counterfactual reasoning.

Keywords: causal inference; counterfactual reasoning; causal intervention; NLP bias; model fairness

引言

（1）研究背景与意义

近年来，自然语言处理（NLP）技术取得了显著突破，尤其是预训练语言模型的出现，极大地推动了文本理解、生成和交互任务的发展。然而，在这些模型迅速普及和实际部署的过程中，数据中固有的社会、文化和语言偏见逐渐暴露出来，导致模型在实际应用中出现了严重的偏差问题。这些偏差不仅体现在模型预测的准确性上，还可能影响决策的公平性和系统的鲁棒性，更进一步可能引发伦理和社会公正方面的争议。例如，在文本分类[1]、情感分析[2]以及对话系统等应用中，模型偏差可能固化刻板印象，加剧社会不平等，对用户体验和信任度产生负面影响。因此，如何有效识别和消除 NLP 模型中的偏见，成为当前学术界和工业界亟待解决的重要问题。

（2）现有方法与局限性

为应对模型偏差问题，研究者们提出了多种传统消偏方法，包括数据增强、微调和人工标注等策略。这些方法通常通过调整训练数

作者简介：赵剑波（1997—），男，硕士，主要从事自然语言处理（NLP）、因果推理。E-mail：1019657026@qq.com
通信作者：何琪（1998—），女，本科，主要从事教育学。E-mail：836144906@qq.com

Copyright © This Work is Licensed under A Commons Attribution-Non Commercial 4.0 International License.

据分布或在后处理阶段修正模型输出来缓解偏差。然而，受限于数据本身的局限性以及人为干预的局部性，这些技术往往只能针对表面现象进行修补，而难以根除深层次的系统性偏见 [2]。

（3）综述范围与贡献

本文旨在从因果推断的视角系统回顾 NLP 模型偏差消除的最新研究进展。具体而言，我们首先介绍因果推断的基本理论框架和关键概念，包括结构化因果模型、因果干预及反事实推理等 [6]；随后，重点讨论基于因果推断的各类应用场景，如文本分类 [1]、情感分析 [2]、多轮对话 [3]、多模态任务 [4] 中的偏差修正方法，并对比分析传统方法与因果消偏技术在实际应用中的优劣。最后，我们探讨当前研究面临的主要挑战，包括数据质量、因果模型构建中的假设验证、计算复杂度和泛化能力等问题，并展望未来将因果推断与深度学习等前沿技术相结合的潜在发展方向。通过本综述，我们期望为后续构建更公平、鲁棒且透明的 NLP 系统提供理论依据和实践参考。

一、因果理论基础知识

（一）因果推断基础

因果推断旨在区分变量间的简单统计关联与真实因果关系，其基本思想可通过“因果阶梯”来理解 [6]。该阶梯分为三个层次：第一层是关联，即通过观察数据揭示变量之间的相关性；第二层是干预，通过人为改变某一变量的状态（采用 do-operator），探讨该干预对其他变量产生的直接影响；第三层是反事实推理，借助假设性情景回答“如果未发生干预，结果将如何变化”的问题。在此过程中，混杂因素（confounders）常常同时影响因果链上的多个变量，从而引入虚假的关联。为解决这一问题，后门准则（backdoor criterion）和前门准则（Frontdoor criterion）被提出，通过选取合适的变量集合来阻断非因果路径，确保评估出的因果效应更为纯粹和准确。

（二）因果图

因果图也称为有向无环图（DAG），是一种用来直观展示变量间因果关系的图形工具。图中节点代表变量集合 V，而有向边则表示变量之间的因果影响关系。以图1为例，变量 T（Treatment）对变量 Y（Outcome）存在直接的因果效应（T → Y）；同时，T 还通过中介变量 M（Mediator）对 Y 产生间接影响，即 T → M → Y，其中 M 在这一过程中发挥了中介作用，形成了分层的因果结构。

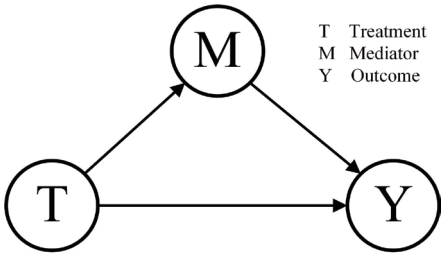


图 1：因果图

二、基于因果推断的 NLP 偏差消除方法

（一）文本分类中的因果推断方法

在文本分类任务中，因果推断方法主要通过因果干预和反事实推理来消除因混杂变量引入的偏差。例如，有研究者提出了利用反事实思维来修正文本分类中的偏见，其方法通过构造虚拟情境来对比实际输入与假定“无偏”输入之间的差异，从而实现对模型决策过程的解释与调整。陈乾等人则采用因果图构建与后门调整策略 [1]，提出 CORSAIR 模型，明确识别并控制混杂因素，确保模型在捕捉文本语义时能剔除外部干扰因素，这类方法通常需要首先构建一个合理的因果图，然后选取合适的混杂变量，并利用 do-operator 进行干预，最终通过后门准则实现偏差的有效消除。

（二）情感分析与方面级情感分类

在情感分析领域，模型往往受到数据中隐含情感偏差的干扰，尤其在方面级情感分类任务中，不同方面的信息可能相互交叉影响。周杰等人以及吴嘉龙等人则分别利用多变量因果推断方法 [2]，构建了针对特定方面的因果模型，从而在保持语义完整性的同时，有效抑制了因果路径中虚假相关性的传递。这些方法通过精细化地建模各因素之间的因果关系，为情感分析任务提供了更为公正和准确的决策依据 [7]。

（三）对话系统与文本生成

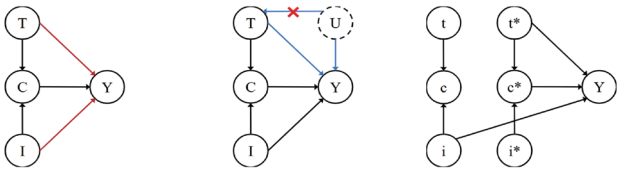
在多轮对话系统中，模型的偏差可能在连续对话中被不断放大，从而影响最终对话质量。王旭等人提出利用因果推断的方法 [3]，通过对多轮对话中各轮之间的因果关系进行建模，实现对话生成过程中的偏差校正。与此同时，Goyal 等人提出了 CaM-Gen，他们通过构建因果图来指导文本生成过程，从而确保生成文本在语义连贯的同时尽可能减少由数据偏见引起的不良影响。这些方法为对话系统和文本生成任务提供了一种系统性的偏差消除思路。

（四）多模态与事实验证应用

在假新闻检测和多模态任务中，文本和图像等不同模态之间可能存在复杂的交互关系，而这种交互关系也容易导致跨模态偏差。陈志云等人利用因果干预方法构建了跨模态因果图，对文本和图像特征之间的因果关系进行分离和校正，从而提高了假新闻检测的准确性。与此同时，陈子伟等人采用反事实推理来探索各模态之间的潜在偏见 [4]，图2所示文章采用后门调整的方法在因果推理和计算因果效应在训练阶段使用微积分 P（Y|(T)），从根本上不同于传统的概率 P（Y|T），后利用反事实推理，通过想象一个反事实世界的图2(c) 来应用因果干预，前面2.1节提到的第三层 [8]。

通过引入虚拟参照值，有效地抑制了跨模态信息传递中的虚假关联。这些方法显示了因果推断在融合不同模态信息时的独特优势，有助于实现更为精准和公平的事实验证。

 Copyright © This Work is Licensed under A Commons Attribution-Non Commercial 4.0 International License.



(a) 真实世界的因果图 (b) 混杂因素的因果图 (c) 反事实世界因果图

图2：给出了假新闻检测的因果图。在这个因果图中，T 表示文本特征，I 表示图像特征，C 表示多模态特征，可以理解为图像和文本特征的融合。Y 表示新闻标签，U 表示混杂因素。文章特别指出，一个星号（*）表示一个参考值。

（五）国内外相关研究综述

在国际上，基于因果推断的偏差消除方法已在多个 NLP 任务中展现出显著效果，如文本分类[1]、情感分析[2]和对话系统[3]等。与此同时，国内学者也在这一领域做出了积极探索。例如，孙圣杰等人提出了融合干预与反事实的知识感知型去偏推理模型[5]。总体来看，国内外研究在方法构建、数据集选取以及实验设计上各有侧重，但都强调因果推断在根除模型系统性偏见方面的潜力[9]。

三、挑战与未来展望

（一）当前方法的局限性

1. 数据质量与样本偏差问题

因果推断方法依赖于高质量的数据集，但在 NLP 任务中，数据往往存在标注不准确、数据稀缺或样本不均衡等问题，这些问题可能影响因果模型的训练效果和偏差消除的有效性。特别是在处理多样化的自然语言文本时，数据质量问题对模型的性能影响较大，可能导致模型未能完全捕捉到因果关系，从而无法有效消除偏见[10]。

2. 计算复杂度与模型泛化性问题

因果推断模型通常需要对数据进行复杂的推理计算，这导致了较高的计算复杂度。对于大规模 NLP 任务（如大规模文本分类、生成任务等），因果推断方法的计算开销可能难以接受。与此同时，模型的泛化性也是一个亟待解决的问题，尤其是在面对不同领域或不同任务的数据时，因果推断方法的适用性和效果可能受到限制。

（二）未来研究方向

1. 多模态数据中更全面的因果建模方法

随着多模态数据（如文本、图像、视频等）的广泛应用，因果推断方法需要考虑不同模态之间的相互影响。未来的研究应致力于构建能够处理多模态信息的因果推断模型，深入挖掘不同模态之间的因果关系，进而更好地消除跨模态偏见。例如，在多模态假新闻检测中，因果推断可以用来分析文本和图像特征间的交互作用，从而提高检测效果。

2. 推广因果推断在更多 NLP 任务中的应用

尽管目前因果推断方法在文本分类、情感分析和对话系统中取得了成功，但其在更多 NLP 任务中的应用仍然相对较少。未来的研究应进一步拓展因果推断在问答系统、机器翻译、文本摘要等其他 NLP 任务中的应用，探索因果推断如何帮助改善这些任务中的偏见问题，并提高系统的公平性和透明度。

四、总结

因果推断方法为自然语言处理（NLP）领域的偏差消除提供了全新的理论框架和技术手段，通过构建因果模型、因果干预和反事实推理，能够有效识别并消除数据偏见、混杂因素和虚假相关性带来的偏差。这些方法在文本分类、情感分析和对话系统等领域展现了显著潜力，促进了更公平、透明和鲁棒的 NLP 系统构建。然而，因果推断仍面临数据质量、假设检验和计算复杂度等挑战，未来研究需深入探索因果推断与深度学习结合、多模态因果建模及领域知识应用等方向，进一步扩展其在更多 NLP 任务中的应用，推动人工智能系统的公平性和透明度。

参考文献

[1] 陈乾，冯福利，温立杰，马春平，谢鹏军 . 基于反事实推理的文本分类去偏 . 载于《第 59 届计算语言学协会年会暨第 11 届国际自然语言处理联合会议论文集（卷 1：长篇论文）》第 5434–5445 页 .2021.

[2] 吴嘉龙，张林海，周德宇，许国强 .DINER: 基于多变量因果推理的方面级情感分析去偏 . 载于《计算语言学协会年会论文集 :ACL2024》3504–3518 页 .2024.

[3] 王旭，张海南，赵帅，陈洪申，丁卓烨，万志国，程博，蓝雁燕 . 基于因果推理的多轮对话推理中的反事实语境去偏 .IEEE/ACM 音频、语音与语言处理期刊 .2023.

[4] 陈子伟，胡林梅，李威新，邵颖霞，聂立强 . 基于因果干预和反事实推理的多模态假新闻检测 . 载于《第 61 届计算语言学协会年会论文集 (卷 1: 长篇论文)》第 627–638 页 .2023.

[5] 孙圣杰，马廷淮，黄凯 . 融合干预与反事实的知识感知型去偏推理模型 [J].《计算机科学与技术前沿》18(12).2024 年 .

[6] 朱迪亚・珀尔 .《因果关系：模型、推理与推断（第二版）》. 剑桥大学出版社，美国 .2009.

[7] 虞燕波，支德源 .PCSK9 抑制剂的应用与消化道肿瘤之间的因果关系：一项药物靶向孟德尔随机化研究 [J]. 临床和实验医学杂志 ,2024,23(22):2353–2356.

[8] 宋春雨 . 客观归责理论视角下重大责任事故罪因果关系的判断 [J]. 中国检察官 ,2024,(22):26–30.

[9] 孙丽华，王文静，朱宏涛，等 . 基于故障相关变量因果关系分析的工业过程故障根因诊断 [J]. 化工自动化及仪表 ,2024,51(06):1035–1044.

[10] 伍杨，李甜，王顺娜，等 . 两样本孟德尔随机化分析幽门螺杆菌感染与慢性乙型肝炎的因果关系 [J]. 胃肠病学和肝病学杂志 ,2024,33(11):1475–1480.

Copyright © This Work is Licensed under A Commons Attribution-Non Commercial 4.0 International License.