

ai rebuttal

1.1 Exploring Jiu-Jitsu Argumentation for Writing Peer Review Rebuttals Accepted at EMNLP Main Conference 2023


基础数据来源

- DISAPERE数据集（ICLR 2019-2020）：
 - 包含9,946条审稿句子（review sentences）和11,103条反驳句子（rebuttal sentences）。
 - 复用三层标注：
 - 审稿方面（Review Aspect）→ 作为态度根源（Attitude Root）（如Clarity、Substance等）。
 - 审稿-反驳链接（Review-Rebuttal Links）→ 用于关联审稿与对应反驳。
 - 反驳行为（Rebuttal Actions）（如Task Done、Reject Criticism等）。
- Peer-Review-Analyze数据集（ICLR 2018）：
 - 提供论文章节标签（Paper Sections）→ 作为态度主题（Attitude Theme）（如Methodology、Experiments等）。

数据集统计与结构

| 规模 | 组件 | 来源/方法 |
|-------------------------|-------|----------------------|
| 2,332条 | 审稿句子 | DISAPERE（负面审稿） |
| 8类（如Substance, Clarity） | 态度根源 | DISAPERE的审稿方面 |
| 143类（如Methodology） | 态度主题 | Peer-Review-Analysis |
| 302条 | 典型反驳句 | 人工筛选+PageRank |
| 16类（如Task Done） | 反驳行为 | DISAPERE原始标注 |

总结

 优点

1. 概念转化：


将心理学中的"态度根源"映射到学术审稿场景，用审稿方面和论文章节分别代理态度根源和主题。

2. 领域自适应：

通过MLM任务在审稿数据上继续预训练语言模型（如 `SciBERT_ds_neg`），提升主题分类效果。

3. 混合筛选策略：

结合自动分类器粗筛与人工偏好标注细筛，最终用图排序算法提取典型反驳。

 不足

• 数据仅来自ICLR，未覆盖自然科学或人文学科

• 稀疏映射：并非所有审稿句都有对应典型反驳，原始数据有缺失

• 论文"Limitations"部分提到，典型反驳作为模板需人工细化，需要避免直接替代人类写作

1.2 Re^2 : A Consistency-ensured Dataset for Full-stage Peer Review and Multi-turn Rebuttal Discussions

原文摘要：同行评审是AI等领域科学进步的关键环节，但提交量的激增使得评审系统不堪重负，这不可避免地导致了评审人员短缺和评审质量下降。除了研究热度持续攀升外，造成这种超负荷的另一个关键因素是不合格稿件的重复提交，这很大程度上源于作者在投稿前缺乏有效的工具进行自我评估。大语言模型 (LLMs) 在协助作者和评审人员方面展现出巨大潜力，但其性能从根本上受限于同行评审数据的质量。然而，现有的同行评审数据集面临三大局限：

1.数据多样性有限；2.由于使用的是修订后版本而非初始投稿，导致数据不一致且质量较低；3.对涉及反驳及评审人-作者互动的任务支持不足。

为应对这些挑战，我们推出了最大规模且确保一致性的同行评审与反驳数据集—— Re^2 。该数据集包含来自OpenReview平台上24个会议和21个研讨会的 19,926份初始投稿、70,668条评审意见以及53,818

份反驳。此外，我们将反驳和讨论阶段构建为多轮对话范式，旨在支持传统的静态评审任务以及动态交互式LLM助手，从而为作者优化稿件提供更具实践性的指导，并有助于缓解日益增长的评审负担。

[A Dataset of Peer Reviews \(PeerRead\): Collection, Insights and NLP Applications](#)这是peer read v1

他们根据这篇文章提出了peer read v2

- 从 PeerRead v1 的约 14,000 篇增长到 ~100,000 篇。
- 从 PeerRead v1 的约 4,000 条增长到 ~170,000 条。
- 覆盖范围45个学术活动，远超 v1 仅覆盖 3 个会议的范围。

数据来源

- 从OpenReview爬取 24个会议与21个研讨会，远超现有数据集比如如PeerRead仅14,700篇论文，无反驳数据
- 初始投稿获取方式
 - （1）通过会议截稿日期确定初始提交时间
 - （2）从论文修订历史中提取截稿前最新版本
 - 使用商用工具Doc2X保障公式识别精度，PDF转纯文本
- 关键技术挑战：
 - 作者/审稿人的多次连续回复合并为单轮对话
 - 针对多位审稿人的相似问题，将全局回复插入对话流
 - 混合人工与自动化方法剔除催促类回复，比如请审稿人回复
 - 结构化多轮对话，支持动态交互式LLM训练。

| 任务类型 | 评估指标 | 关键模 |
|------|-------------------------------|-------|
| 录用预测 | 准确率、F1 | SEA-E |
| 分数预测 | MAE, MSE | 微调后 |
| 评审生成 | BLEU/ROUGE/BERTScore/EmbedCos | 微调LL |
| 反驳对话 | LLM-as-Judge | 微调LL |

一些insight:

- 反驳对话任务中，模型在准确性和建设性表现最佳
- 零样本LLM倾向讨好作者
- 专用模型过于严格，评分均过低

局限

仅处理文本内容，未整合图表等多模态信息

注意点

- 需要处理不同会议使用的不同评审模板和格式，将其解析成结构化字段
- 对于双盲评审，评审人身份已匿名化。
- 遵循了 OpenReview 的使用条款，要求实验数据是公开可获取的
- 并非所有评审都是同等质量的。有些评审可能简短、无建设性，甚至带有偏见

1.3 APE: Argument Pair Extraction from Peer Review and Rebuttal via Multi-task Learning

原文摘要：同行评审及其间的反驳环节，伴随着丰富的互动和论证性讨论，天然是在挖掘论点的优质资源。然而，同时研究两者的工作却寥寥无几。本文中，我们针对同行评审和反驳环节，引入了一项新的论点对提取（Argument Pair Extraction, APE）任务，旨在研究其内容、结构以及两者之间的关联。为支持该任务研究，我们构建了一个具有挑战性的数据集，其中包含来自一个开放评审平台的 4,764 对经过完整标注的评审-反驳段落对。为从该语料库中自动检测论证性命题并提取论点对，我们将此任务建模为一个序列标注任务和一个文本关系分类任务的组合。基于此，我们提出了一个基于分层LSTM网络的多任务学习框架。广泛的实验与分析证明了我们多任务框架的有效性，同时也揭示了这项新任务的挑战性，并为未来研究提供了方向。

- 同行评审中的反驳阶段蕴含丰富的论辩结构，但现有研究多聚焦单文本论元挖掘，忽视审稿-反驳的互动性。
- 学术场景的论元对提取可应用于：
 - (1) 自动检查作者是否回应所有评审意见；
 - (2) 构建论辩知识库优化审稿流程；
 - (3) 为对话式审稿系统提供基础支撑。

数据收集与标注

- OpenReview 平台上的 ICLR 会议，覆盖 4,764组 审稿-反驳段落对。
- 为审稿论元分配索引，在反驳中标记对应回应的索引
- 5名专业标注员，标注准确率 98.4%
- 按结构清晰度划分难度：

- Easy (65.5%)：反驳直接引用审稿内容或有明确分隔符（如换行、编号）。
- Difficult (29%)：需深度理解语义才能对齐论元。
- 无有效论元对 (5.5%)：泛泛回应（如“感谢意见，我们会改进”）。

insight

- 反驳的论元更长、更密集
- 论元对词汇重叠率极低，说明任务依赖深层语义推理而非表面词匹配。

创新点

提出 MT-H-LSTM-CRF 框架，联合学习两个子任务：

1. 论元挖掘（序列标注）：

- 输入：审稿+反驳句子序列
- 输出：IOBES标签（标注论元边界）
- BERT词嵌入 → Token-BiLSTM → Sentence-BiLSTM → CRF

2. 句子对齐（二分类）：

- 输入：审稿句与反驳句的表示向量
- 输出：是否构成论元对
- 共享Sentence-BiLSTM输出 → 句对表示求和 → 三层线性层分类

基于LSTM而非Transformer，可能限制长距离依赖建模能力。

1.4 Incorporating peer reviews and rebuttal counter-arguments for meta-review generation.

原文摘要：同行评审是科学过程中的关键环节，研究论文需由多位评审人进行评估。在大多数顶级会议中设立的作者反驳阶段，为作者提供了针对评审意见为其工作进行辩护的机会。评审人指出的优缺点以及作者的回应，将由领域主席进行评估。最终决定通常会附带说明接受/拒绝理由的元评审。先前研究已探索使用基于Transformer的摘要模型生成元评审。然而，这些研究大多未充分考虑反驳内容本身以及评审意见与反驳论点之间的互动，而其中论据的说服力对最终决定具有重要影响。为生成能够有效整合评审人观点与作者回应、内容全面的元评审，我们提出了一种新颖的生成模型。该模型能够显式地建模复杂的论证结构，不仅涵盖评审人与作者之间的论点交锋，还包括审稿人内部的讨论。实验结果表明，在自动评估和人工评估两方面，我们的模型均优于基线方法，验证了所提方法的有效性。

- 采用端到端模型（MLMC）从评审-反驳对中提取ADU及其关系（表2 F1=54.15%）。
- 引入方面类型标注（Aspect Typology）：利用8类学术维度（如创新性、严谨性）增强ADU提取（F1提升至72.21%）

数据集构建 (PRRCA Dataset)

- OpenReview.net上ICLR 2017-2021的7,627篇投稿
- 包含25,316条评审-反驳对
- 标注论证结构和句子级方面类型
- 平均元评审长度152.5词，输入文本超长

Insight

- 边界分数论文（4-6分）：反驳更长，元评审更详细，生成难度最大
- 评分分歧：高分歧论文接收率更低

图构建：

- 节点：ADU（评审中的论证单元为绿色，反驳中的为黄色）
- 边：三层关系（文档内红色、讨论内蓝色、跨讨论绿色）
- 图增强：添加自环、反向边和超级节点以提升连通性。

1.5 Does my rebuttal matter? insights from a major nlp conference.

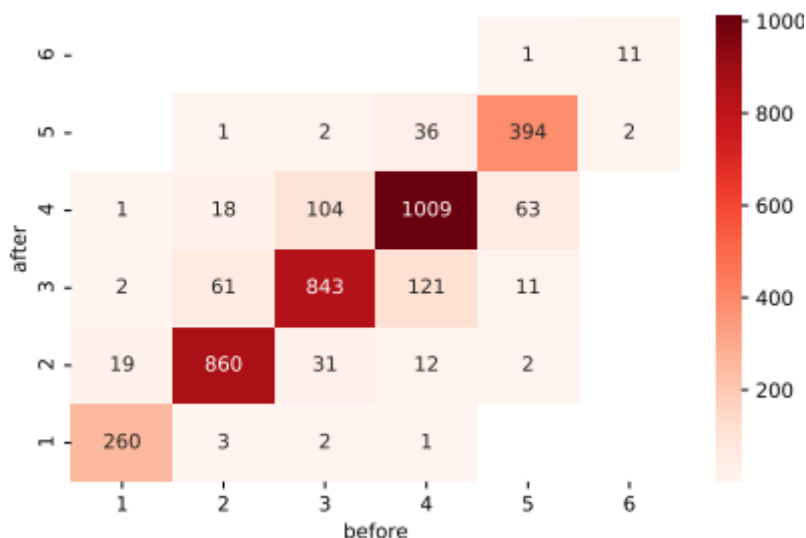
原文摘要：同行评审是科学过程的核心要素，在以会议为中心的领域（如机器学习ML和自然语言处理NLP）中尤其如此。然而，目前仅有少数研究对其特性进行了实证评估。为填补这一空白，我们构建了一个包含来自ACL-2018会议的4000余份评审意见和1200余份作者回应的语料库。我们对该语料库进行了定量和定性评估。这包括对评审人指出的论文弱点以及作者回应质量的初步分析。随后，我们聚焦于反驳阶段的作用，并提出了一项新颖的任务：基于初始评审意见和作者回应，预测反驳后（即最终）的评分。尽管作者回应确实对最终评分具有有限但显著的影响（尤其在边界论文上），但我们的结果表明，评审人的最终评分很大程度上由其初始评分及其与其他评审人初始评分的差异所决定。在此背景下，我们探讨了同行评审中固有的从众偏差，这一偏差在先前研究中长期被忽视。我们希望我们的分析将有助于更好地评估NLP会议中反驳阶段的有效性。

数据集构建与特性

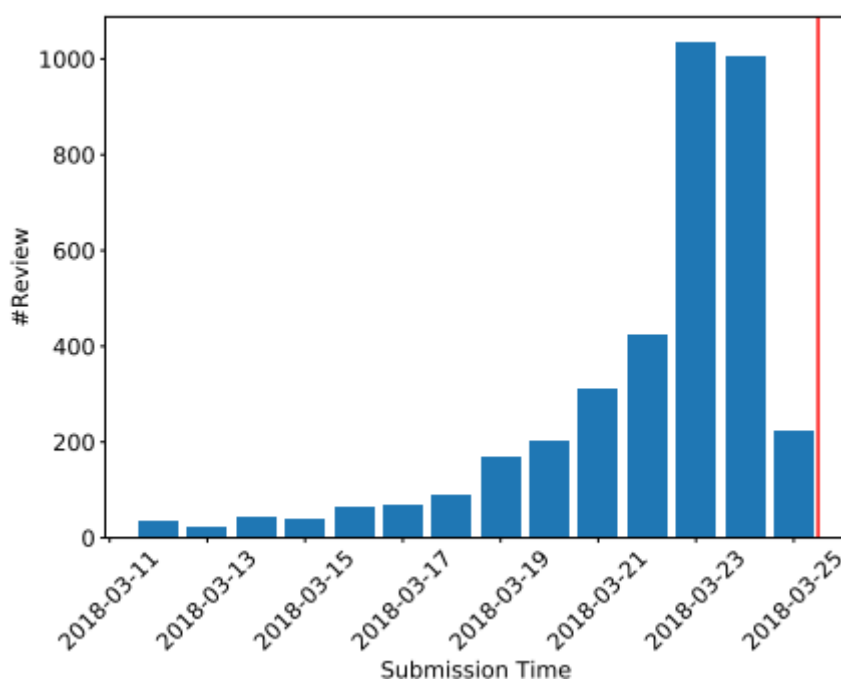
ACL-2018 Review Corpus包含4,054条评审（反驳前后完整记录）和1,227条作者反驳，覆盖1,542篇投稿（含接受/拒绝论文），是首个包含完整评审周期的NLP会议数据集

insight

- peer review评分均值差距是最大影响因素
- 文本特征仅在边缘论文中显著



- 高礼貌降低DEC概率，但对INC无促进
- 高质量反驳可阻止降分，但难以推动升分
- 低质量反驳显著增加降分风险
- 评审者强烈趋向同侪评分均值，有从众倾向
- 早期提交的评审倾向给低分，后期更可能升分



局限：仅31%作者同意共享反驳，高接受率论文可能过采样

1.6 What makes a successful rebuttal in computer science conferences?

原文摘要：随着计算机科学顶会投稿量呈指数级增长，越来越多的会议在同行评审流程中引入了反驳阶段。该阶段可建模为作者与评审间的社会互动，成功的反驳通常能提升评审分数。本文基于ICLR 2022（顶级计算机科学会议）的3,000余篇论文和13,000余条评论开展实证研究，旨在识别成功反驳的核心影响因素。研究发现：反驳阶段前后的评审分数存在显著差异，这对论文录用具有决定性影响；

通过符号化社交网络分析，首次揭示反驳后平衡网络结构比例显著提升；提出并验证了五种可量化反驳策略，证明其有效提升评审分数；构建的机器学习预测模型融合社交动态、策略指标等特征，验证多因子联合预测效能。本研究首创"作者-评审"双视角分析框架，为基于社交网络的评审优化奠定理论基础。

方法论创新：签名社交网络分析（SSNA）

将审稿人-作者互动建模为带符号网络（正/负链接=支持/反对），基于Heider平衡理论定义四类子图结构

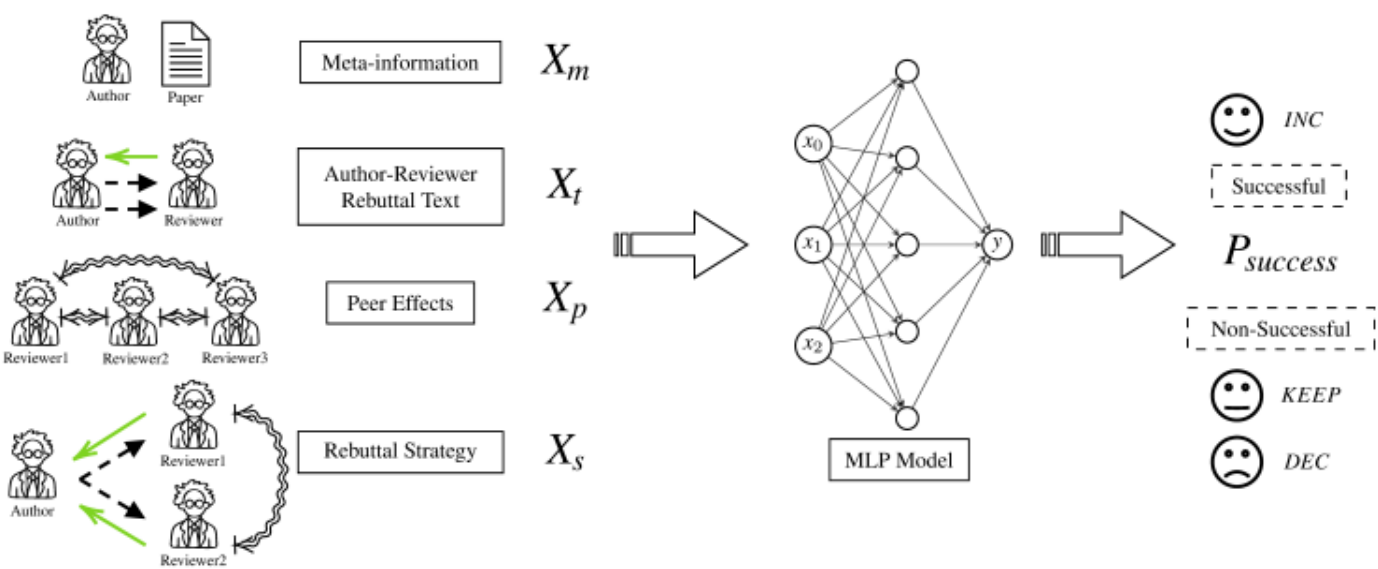
- 不平衡结构：审稿人 R_i 与 R_j 意见相左且互斥
- 平衡结构： R_i 受其他审稿人压力而修改分数

作者策略的量化验证表明，策略有效性排序为充分回复 > 保持礼貌 > 建立共识 > 引用文献

反驳成功预测模型

多因子机器学习模型

- 同伴效应特征 (X_p)：其他审稿人分数统计量（max/min/mean等）+ 平衡子图数量
- 作者策略特征 (X_s)：7类策略
- 文本特征 (X_t)：SPECTER编码的审稿意见与回复语义向量
- 最优模型：MLP整合全部特征显著优于单特征模型



insight

审稿人社交压力驱动分数变化

- 同伴效应是分数修改主因：当其他审稿人均支持时，反对者修改分数的概率提升32%。
- 边界论文（均分5-6）的审稿人更易受社交压力影响而修改分数
- 同伴效应特征贡献最大（AUC=0.7704），验证社交压力的核心作用；作者策略特征（AUC=0.7143），证明量化策略的有效性

局限

- 未分析领域主席对审稿人的影响（因数据缺失）
- 策略分析依赖均值检验，未进行因果推断

其他

[从 OpenReview 获取顶会接收论文集并保存至本地数据库](#)

[能不能用Chatgpt回复审稿人意见 （有趣的回答）](#)

[这个项目是一个用于从同行评审和反驳中提取论点对的自然语言处理工具](#)

关于bias

[Prior and Prejudice: The Novice Reviewers' Bias against Resubmissions in Conference Peer Review](#)

[Reviewer bias in single- versus double-blind peer review](#)

论元

[Argument Pair Extraction via Attention-guided Multi-Layer Multi-Cross Encoding](#)