# CORGI-PM：中文性别偏见探测与缓解语料库

Ge Zhang[1 3 4] *、Yizhi Li[2] *、Yaoyao Wu[5], Linyuan Zhang [6]、Chenghua Lin [2] †、Jiayi Geng[7], Shi Wang [3] †、Jie Fu [1]
[1] 中国北京人工智能研究院 [2] 英国谢菲尔德大学计算机科学系 [3] 中国科学院计算技术研究所 [4] 美国密歇根大学安娜堡分校 [5] 美国科罗拉多大学博尔德分校 [6] 中国四川大学 [7] 加拿大麦吉尔大学 {yizhi.li, c.lin}@sheffield.ac.uk[2], gezhang@umich.edu[1], wangshi@ict.ac.cn[3]

arXiv:2301.00395v1 [cs.CL] 1 Jan 2023

## 摘要

随着性别偏见自然语言处理（NLP）成为一个重要的跨学科主题，当前普遍采用的数据驱动技术（如大规模语言模型）却受到数据不足和有偏语料库的制约，尤其是在中文等资源匮乏的语言中。为此，我们提出了中文性别偏见探测与缓解语料库（**CORGI-PM**[1]），该语料库包含 32.9k 个句子，标签质量高，并依据专为中文语境中的性别偏见制定的标注方案生成。此外，我们针对自动缓解文本性别偏见的三项挑战开展研究，这类任务要求模型能够检测、分类并缓解文本性别偏见。我们还使用最先进的语言模型开展实验，以提供基线。据我们所知，CORGI-PM 是首个用于性别偏见探测与缓解的句子级中文语料库。

## 1 引言

越来越多的人达成共识，认为识别并防止有害的性别态度和刻板印象对社会至关重要（Blodgett 等人，2020）。由于性别偏见信息可能以文本形式呈现并广泛传播，因此开发用于检测和缓解文本性别偏见的自动化方法至关重要。

自然语言处理（自然语言处理）已广泛应用于与文本相关的应用中，并对性别偏见议题产生了重大影响（Costa-jussà，2019）。一方面，大规模语言模型（LM）作为现代自然语言处理的关键技术，已被证实会从训练语料库中习得主观性别偏见，甚至会放大这种偏见（Zhao 等人，2017）。另一方面，越来越有希望应用前沿的自然语言处理技术用于探测和缓解性别偏见。

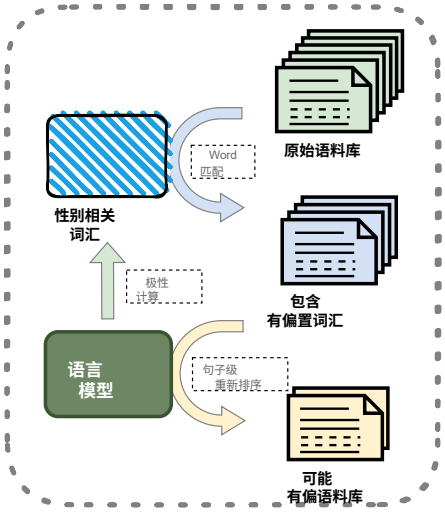



图1：从原始语料库中检索并筛选可能存在偏见的句子，以供人工标注的流程。

构建高质量文本语料库一直是改进自然语言处理应用、消除文本中性别刻板印象的关键方向之一（Sun 等，2019）。一些研究人员引入自动标注技术，例如基于性别互换的方法，以创建用于性别偏见缓解的语料库（Lu等人，2020；Zhao 等，2018；Rudinger等人，2018）。尽管无需投入大量人力即可构建大型语料库颇具吸引力，但基于性别互换的自动方法高度依赖基础语言模型的质量，而且容易生成无意义句子（Sun 等，2019）。为解决这一问题，一些研究致力于开发用于性别偏见缓解的人工标注语料库。然而，这些语料库要么主要关注词汇层面或语法级偏见（Webster等人，2018；Zhu 和 Liu，2020；Sahai和Sharma，2021；Zhou等人，2019），要么只关注性别歧视相关主题（Jiang等人，2022；Chiril等人，2021，2020；Parikh等人，2019）。

此外，现有关于性别偏见的研究仅

* 两位作者对本研究贡献相同。
† 通讯作者。[1]我们的代码可在 GitHub

关注英语（Costa-jussà，2019），而其他具有影响力的语言（如中文）可用的数据集寥寥无几。（注：有关生成的含有不合逻辑中文句子的性别偏见语料库的详情，见附录D）。我们旨在提供一个高质量的中文人工标注语料库，用于上下文层面的性别偏见探测与缓解，从而解决上述问题。

为此，我们提出中文性别偏见探测与缓解语料库（**CORGI-PM**）数据集，其中包含32.9k 条人工标注的句子，既有性别偏见样本，也有无偏样本。为收集初始数据，我们提出一种自动方法，从现有的大规模中文语料库中构建一个可能存在性别偏见的句子集合。受Bolukbasi 等人 (2016);Jiao 和 Luo(2021提出的利用语言模型计算性别偏见分数的指标启发，我们召回包含性别偏见分数较高单词的样本，然后根据其句子级别的性别偏见概率重新排序并筛选，如图 1所示。为确保语料库的质量，我们精心设计了标注方案，并选用教育背景合格的标注人员，对检索出的句子进行进一步标注和改写。

此外，我们基于 CORGI-PM 应对三项挑战，即即性别偏见检测、分类和缓解，并为性别偏见探测与缓解相关的自然语言处理任务提供明确的定义和评估协议。为针对所提出的挑战提供参考基线和基准，我们以均衡标签进行随机数据划分，并在零样本、上下文学习和微调范式下，使用前沿语言模型开展实验。我们在 §3中讨论实验设置并分析结果。实现细节请参阅附录C。

总之，我们提供了一个经过完善标注、用于性别偏见探测与缓解的中文语料库，并明确界定了相应的挑战。借助精心设计的标注方案，CORGI-PM 提供了高质量语料库，可帮助模型检测文本中的性别偏见。更重要的是，除了22.5k 个由人工标注的无偏样本外，我们语料库中的全部 5.2k 个带偏见的句子也都由标注人员进一步标注了性别偏见子类和公司，并提供了相应的无偏见版本

| 样本类别 | | 数量 | | |
|---|---|---|---|---|
| | | 训练集 | 验证集 | Test |
| Biased | AC | 1.90k | 235 | 237 |
| | DI | 2.70k | 334 | 337 |
| | ANB | 2.47k | 306 | 309 |
| 无偏 | | 21.4k | 516 | 526 |
| 总体 | | 30.1k | 1391 | 1409 |

表 1：CORGI-PM 数据集的总体统计。符号 **AC**、**DI** 和 **ANB** 表示 §2.2 中所述的特定偏见标签。

人员。我们的代码和数据集将会发布，以造福社区。

## 2 数据收集

### 2.1 样本筛选

我们提出了一种自动处理方法，利用从词级到句子级的两阶段筛选，从原始语料库中召回、重新排序并筛选标注候选项，如图 1 所示。中文句子样本主要从 SlguSet（Zhao 等人，2021）和 CCL 语料库（Weidong 等人，2019）中筛选得到。

为了召回性别偏见词或检索具有性别偏见分数的候选句子，我们将目标词/句子的表示与种子方向进行比较，该方向可通过she和he的词嵌入相减计算得到（Bolukbasi 等人，2016；Jiao 和 Luo，2021）。我们利用不同的中文语言模型，包括 ERNIE（Zhang 等人，2019）、CBert（Cui 等人，2020）以及中文词向量（Qiu 等人，2018），以获取词级和句子级表示。对于词级筛选，我们使用上述指标构建高偏见分数的词汇表，并通过精确匹配从原始语料库中召回包含这些词的句子。我们计算爬取句子的性别偏见分数，并根据上一阶段获得的性别偏见关键词对其进行分组，以进行句子级筛选。随后，根据特定的全局阈值性别偏见分数和组内阈值排名，选出最终用于标注的句子。词级筛选过程以词云形式展示，详见附录 B.1。

### 2.2 标注方案

该标注方案旨在进行性别偏见探测和缓解。在性别偏见探测任务中，标注人员需要针对给定句子提供以下信息：是否存在性别偏见；如果存在，偏见是如何形成的。在性别偏见缓解任务中，还需要提供带偏见句子修正后的无偏见版本。我们

| 语言学 | 无偏 | | | 有偏 | | | 校正后的有偏 | | |
|---|---|---|---|---|---|---|---|---|---|
| 信息 | 训练集 | 验证集 | Test | 训练集 | 验证集 | Test | 训练集 | 验证集 | Test |
| Word | 724k | 18.9k | 17.7k | 228k | 24.8k | 28.3k | 265k | 27.1k | 30.0k |
| 词典 | 574k | 14.4k | 14.1k | 167k | 18.4k | 20.4k | 191k | 19.9k | 21.5k |
| 字符 | 1,156k | 30.1k | 28.1k | 358k | 39.2k | 44.4k | 417k | 42.8k | 46.9k |
| 句长 | 53.952 | 58.397 | 53.473 | 85.837 | 76.087 | 85.214 | 99.839 | 82.853 | 89.939 |

表2：语料库的语言特征。单词、词典和字符分别表示特定类别中的中文词语总数、中文词语去重后的总数以及字符总数。句长定义为所含字符的数量。

将在下文进一步介绍标注方案的详细内容。

**存在性与分类**。

标注人员需要标注句子在语境层面或词级上是否存在性别偏见（**B**）或无偏（**N**），并进一步说明偏见是如何形成的。鉴于我们的原始数据是通过性别相关关键词或从性别相关语料库中收集的，标注为不存在性别偏见的样本可作为由人工标注的有效负样本，用于检测性别偏见。为进一步提供有关性别偏见分类的信息，我们将性别偏见类型分为三个子类：（1）性别刻板化的活动与职业选择**(AC)**；（2）性别刻板化的描述与推断**(DI)**；以及（3）表达出的性别刻板化的态度、规范与信念**(ANB)**。分类标准参考了（King 等人，2021），并进一步归纳为上述子类。

**偏见缓解**。 标注人员还需要在保留原有语义信息的同时，缓解所选句子中的性别偏见。我们也要求标注人员在适用时采用多样化的表达方式。主要的修订模式可概括如下：**(1).** 将 性别特定代词替换为中性代词。**(2).** 将性别特定形容词替换为语义定义相近的中性描述。**(3).** 添加 额外评论，以中和无法直接缓解偏见的句子。

### 2.3 语料库分析

本节报告 CORGI-PM 的语言学统计数据，如表1所示。我们在构建验证集和测试集时，综合考虑全局分布中的负-正样本比例和偏见子类比例，设计了一个平衡的数据划分。如表 2[2],所示，将有偏样本与原文样本进行比较后，我们观察到两个主要差异：表达更长且更加多样（注意，表2中的句子长度和词汇量也体现了这一点）。我们推测，这是因为人工标注员在缓解性别偏见的同时，有意保持语义信息不变并确保句子连贯。与某些带有内在性别偏见的表达相比，他们可能会使用更多连词和更长的描述。有关质量管理与控制的更多细节，请参阅附录 B.1 和 B.2。

## 3 性别偏见缓解面临的挑战

为了明确自动化文本性别偏见探测与缓解任务的定义，我们提出了相应的挑战，并制定了标准化的评估协议。我们针对性别偏见探测提出了两项任务：检测和分类，并将性别偏见缓解挑战形式化为一项文本缓解任务。

### 3.1 检测与分类面临的挑战

我们将性别偏见检测和分类挑战都视为监督分类任务，并使用一致性指标对其进行评估。**定义**。性别偏见检测挑战可视为一项二元分类任务，要求模型预测给定句子包含性别偏见的概率。如 § 2.2所述，有偏样本还会被进一步归入一种或多种类别。因此，我们可以将性别分类挑战作为一项多标签分类任务来处理。在这两项挑战中，我们选取精确率、召回率和 F1 分数作为主要指标。我们要求通过验证集和测试集提供分类别指标以及宏平均汇总评估，以展示语言模型的性能。

**基线**。 我们对来自三种具有代表性的不同预训练范式的中文语言模型进行微调，即中文 BERT、Electra 和 XL-Net Cui 等人 (2020)，通过添加一个额外的密集预测层来完成检测和分类任务。 [3] 我们还提供了 GPT-3（Brown 等人，2020）Curie 在检测和分类任务上的少样本表现

[2]我们使用结巴 进行解析。

[3]预训练模型可在HFL Anthology中找到。

| 模型 | 指标 | 分类（验证集） | | | | 分类（测试集） | | | | 检测（验证集） | | | 检测（测试集） | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AC | DI | ANB | Avg. | AC | DI | ANB | Avg. | N | B | Avg. | N | B | Avg. |
| BERT | 精确率 | .609 | .729 | .533 | .624 | .556 | .615 | .521 | .564 | .699 | .950 | .824 | .742 | .980 | .861 |
| | 召回率 | .594 | .665 | .543 | .601 | .493 | .652 | .585 | .577 | .971 | .591 | .781 | .985 | .662 | .823 |
| | F1 分数 | .602 | .695 | .538 | .612 | .522 | .633 | .551 | .567 | .813 | .729 | .771 | .846 | .790 | .818 |
| Electra | 精确率 | .587 | .727 | .544 | .619 | .556 | .630 | .516 | .568 | .691 | .936 | .814 | .745 | .974 | .860 |
| | 召回率 | .758 | .687 | .386 | .610 | .680 | .685 | .373 | .579 | .961 | .570 | .766 | .983 | .656 | .820 |
| | F1 分数 | .661 | .706 | .451 | .606 | .612 | .656 | .433 | .567 | .804 | .708 | .756 | .848 | .784 | .816 |
| XLNet | 精确率 | .587 | .696 | .523 | .602 | .544 | .589 | .527 | .553 | .713 | .928 | .820 | .772 | .959 | .865 |
| | 召回率 | .622 | .643 | .495 | .587 | .545 | .614 | .514 | .558 | .953 | .620 | .787 | .968 | .722 | .845 |
| | F1 分数 | .604 | .669 | .509 | .594 | .544 | .601 | .520 | .555 | .816 | .743 | .780 | .859 | .824 | .841 |
| Curie | 精确率 | .695 | .907 | .010 | .537 | .622 | .887 | .009 | .506 | .763 | .665 | .714 | .635 | .825 | .730 |
| | 召回率 | .395 | .802 | .375 | .524 | .395 | .804 | .010 | .403 | .576 | .825 | .700 | .975 | .584 | .780 |
| | F1 分数 | .504 | .851 | .019 | .458 | .508 | .852 | .019 | .460 | .656 | .736 | .696 | .769 | .684 | .727 |

表3：性别偏见检测与分类任务的基线结果。总体指标指 Marco 平均值。模型名称及缩写参见 § 3.1。类别定义参见 § 2.2。

| 指标 / 模型 | BLEU | METEOR | ROUGE-L | | | 人工评估 | |
|---|---|---|---|---|---|---|---|
| | | | 召回率 | 精确率 | F1 | 连贯性 | 性别偏见 |
| *Davinci | .776 | .879 | .203 | .211 | .205 | 5.25 | 0.96 |
| Ada | .288 | .429 | .407 | .180 | .250 | 5.98 | 1.13 |
| Babbage | .359 | .504 | .716 | .310 | .432 | 6.32 | 0.69 |
| Curie | .364 | .506 | .692 | .316 | .434 | 6.21 | 1.20 |

表4：基线结果。性别偏见纠正任务的指标详情见附录C。*表示使用 the 零样本模式，其他的则指微调。

。检测和分类的基线结果表明，分类任务颇具挑战性；正如表3所示，CORGI-PM 中的性别偏见检测性能仍有提升空间。

### 3.2 缓解挑战

**定义**。性别偏见缓解挑战可以被视为一项自然语言生成任务：要求模型生成带偏见句子的修正版本，并以人工标注的版本作为参考。

**基线**。我们在微调实验设置下测试了 GPT-3（Brown 等人，2020）在 CORGI-PM 上的表现，采用三种不同参数规模的模型，分别为 Ada（350M）、Babbage（1.3B）和 Curie（6.7B）；并在零样本实验设置下测试 Davinci（175B）。我们只提供 Davinci 的零样本结果，因为它是唯一已发布的 GPT-3 编辑模型。更多实现和评估细节见附录C。

**讨论**。 我们同时提供人工评估和自动化指标进行评估。表 4 表明，语言模型能够学习缓解性别偏见的标注模式，零样本编辑模型也展现出有竞争力的表现。人工评估和 ROUGE-L 的结果显示，微调后的 Babbage 优于规模大得多的零样本 Davinci，这表明 CORGI-PM 有望作为性别偏见缓解任务的强监督信号。我们注意到，与微调后的模型相比，Davinci 往往采取更保守的编辑方式。因此，Davinci 编辑的句子保留了大部分原句，并且始终只修改原句中的代词和形容词，这有利于 BLEU 等侧重精确率的自动化指标（Papineni 等人，2002）和 METEOR（Agarwal 和 Lavie，2007）。人工评估与自动化指标之间的表现差异，揭示了人类与语言模型在写作风格上的差异。

## 4 结论

我们提出了 CORGI-PM，这是首个同时用于性别偏见探测与缓解的中文人工标注语料库。我们还基于 CORGI-PM，为三个挑战明确了定义和评估指标，并测试了最先进的语言模型的性能。我们提出的挑战可作为基准，用于衡量语言模型检测、分类和缓解文本性别偏见的能力。实验表明，带有细粒度子类标签的句子能够帮助语言模型进行性别偏见探测，而我们人工撰写的平行去偏数据则可为生成式语言模型提供强有力的监督。总之，我们认为，未来利用 CORGI-PM 的工作将有助于推动自然语言处理领域在性别偏见探测与缓解方面的发展。

## 局限性

本研究工作存在若干主要局限性。由于对标注人员的要求较高，

在标注性别偏见句子并纠正这类句子时，我们只选择了受教育程度较高的标注人员，这可能导致潜在的认知偏差。此外，我们仅对常用中文语言模型在我们提出的新挑战中的表现进行了有限的实现和测试实验。未来可以在这些挑战中进一步探索更多语言模型和技术。

## 伦理声明

我们在收集过程中仔细考虑了伦理影响。CORGI-PM 语料库中句子的收集仅依赖于公开可用、用于研究目的的语料库。我们已向标注人员说明数据集的潜在用途及相关隐私问题，并在开始标注前获得了确认。

## 参考文献


Abhaya Agarwal 和 Alon Lavie。2007。Meteor：一种用于机器翻译评估的自动指标，与人工判断具有高度相关性。*WMT-08* 论文集。

Su Lin Blodgett、Solon Barocas、Hal Daumé III 和 Hanna Wallach。2020。语言（技术）即权力：对自然语言处理领域“偏见”的批判性综述。发表于*ACL*。

Tolga Bolukbasi、Kai-Wei Chang、James Y. Zou、Venkatesh Saligrama 和 Adam Tauman Kalai。2016。男人之于计算机程序员，犹如女人之于家庭主妇？消除词嵌入中的偏见。发表于*NIPS*。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901。

Patricia Chiril、Farah Benamara 和 Véronique Moriceau。2021。“对妻子好一点！餐馆都关门了”：性别刻板印象检测能否改善性别歧视分类？载于 计算语言学协会 *EMNLP 2021* 研究成果，第 2833–2844 页。

Patricia Chiril、Véronique Moriceau、Farah Benamara、Alda Mari、Gloria Origgi 和 Marlène Coulomb-Gully。2020。用于检测法语推文中性别歧视的标注语料库。载于 第 *12* 届语言资源与评估会议论文集，第 1397–1403 页。

Marta R Costa-jussà. 2019. 自然语言处理中的性别偏见研究分析。《自然机器智能》，1(11):495–496。

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shi- jin Wang, and Guoping Hu. 2020. 重新审视用于中文自然语言处理的预训练模型。*arXiv* 预印本 *arXiv:2004.13922*。Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. SWSR：用于在线性别歧视检测的中文数据集和词典。《在线社交网络与媒体》，27:100182。Meichun Jiao and Ziyang Luo. 2021. 中文词嵌入背后隐藏的性别偏见：以中文形容词为例。载于《第三届自然语言处理性别偏见研讨会论文集》，第 8–15 页。

Tania L King、Anna J Scovelle、Anneke Meehl、Allison J Milner 和 Naomi Priest。2021。幼儿期的性别刻板印象与偏见：系统性综述。澳大拉西亚幼儿期教育期刊，46(2)：112–125。

Chin-Yew Lin。2004。ROUGE：一种用于自动评估摘要的工具包。载于文本摘要分支拓展，第 74–81 页。

Kaiji Lu、Piotr Mardziel、Fangjing Wu、Preetam Amancharla 和 Anupam Datta。2020。神经自然语言处理中的性别偏见。载于逻辑、语言与安全，第 189–202 页。施普林格。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。2002。BLEU：一种用于机器翻译自动评估的方法。载于计算语言学协会第 *40* 届年会论文集，第 311–318 页。

Pulkit Parikh、Harika Abburi、Pinkesh Badjatiya、Rad- hika Krishnan、Niyati Chhaya、Manish Gupta 和 Vasudeva Varma。2019。使用神经框架对性别歧视相关账号进行多标签分类。*arXiv*预印本 *arXiv:1910.04602*。

Yuanyuan Qiu、Hongzheng Li、Shen Li、Yingdi Jiang、Renfen Hu 和 Lijiao Yang。2018。重新审视词嵌入内在评估与外在评估之间的相关性。载于基于自然标注大数据的中文计算语言学与自然语言处理，第 209–221 页。施普林格。

RachelRudinger、JasonNaradowsky、Brian Leonard 和 Benjamin Van Durme。2018。共指消解中的性别偏见。*arXiv* 预印本 *arXiv:1804.09301*。

Saumya Sahai 和 Dravyansh Sharma。2021。法语语法性别的预测与解释。载于第三届计算类型学与多语言自然语言处理研讨会论文集，第 90–96 页。

Aarohi Srivastava、Abhinav Rastogi、Abhishek Rao、Abu Awal Md Shoeb、Abubakar Abid、Adam Fisch、Adam R Brown、Adam Santoro、Aditya Gupta、Adrià Garriga-Alonso 等。2022。《超越模仿游戏：量化并外推

语言模型的能力》。*arXiv* 预印本 *arXiv:2206.04615*。

Tony Sun、Andrew Gaut、Shirlyn Tang、Yuxin Huang、Mai ElSherief、Jieyu Zhao、Diba Mirza、Eliza- beth Belding、Kai-Wei Chang 和 William Yang Wang。2019。减轻自然语言处理中的性别偏见：文献综述。*arXiv* 预印本 *arXiv:1906.08976*。

Zeerak Waseem。2016。你是种族主义者，还是我看错了？标注者对 Twitter 仇恨言论检测的影响。载于自然语言处理与计算社会科学第一届研讨会论文集，第 138–142 页，美国德克萨斯州奥斯汀。计算语言学协会。

Kellie Webster、Marta Recasens、Vera Axelrod 和 Ja- son Baldridge。2018。关注差距：一个平衡的带有性别含义的歧义代词语料库。计算语言学协会汇刊，6:605–617。

占卫东、郭睿、常宝宝、陈一戎和龙晨。2019。北京大学 CCL 语料库的建设。载于《中文语料库语言学》。

徐慧敏、张章、吴凌飞和王成军。2019。灰姑娘情结：词嵌入揭示电影和书籍中的性别刻板印象。*PLOS ONE*，14(11):e0225385。

张正言、韩旭、刘知远、姜鑫、孙茂松和刘群。2019。ERNIE：借助信息丰富的实体增强语言表征。*arXiv* 预印本 *arXiv:1905.07129*。

Jieyu Zhao、Tianlu Wang、Mark Yatskar、Vicente Ordonez 和 Kai-Wei Chang。2017。男性也喜欢购物：使用语料库级约束减少性别偏见放大。*arXiv* 预印本 *arXiv:1707.09457*。

Jieyu Zhao、YichaoZhou、李泽宇、王伟和 Kai- Wei Chang。2018。学习性别中立词嵌入。*arXiv* 预印本 *arXiv:1809.01496*。

JishunZhao、杜冰洁、Shucheng Zhu 和 Pengyuan Liu。2021。中文句子级性别无偏数据集的构建及预训练语言中的性别偏见评估。载于第*20*届中国计算语言学大会论文集，第564–575页。

周沛、施维佳、Jieyu Zhao、Kuan-Hao Huang、陈慕豪、Ryan Cotterell 和 Kai- Wei Chang。2019。考察具有语法性别的语言中的性别偏见。*arXiv* 预印本 *arXiv:1909.02224*。

Shucheng Zhu 和 Pengyuan Liu。2020。《伟大的男性与固执的女性：中文形容词中基于语料库的性别偏斜历时研究》。载于第十九届全国计算语言学学术会议论文集，第 31–42 页。

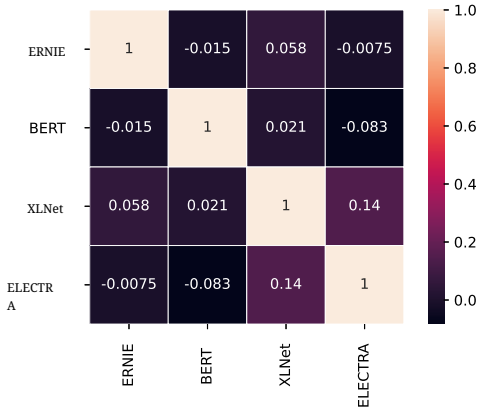



图 2：职业词汇的词级性别偏见比较。

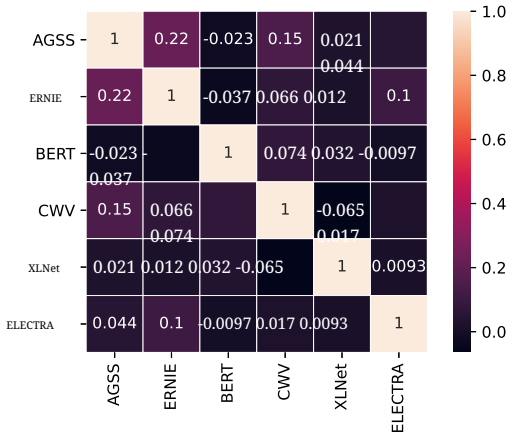



图 3：形容词的词级性别偏见比较。**CWV** 表示使用 Qiu 等人提出的混合大型语料库训练的中文词向量Qiu 等人。

## 中文语言模型的性别偏见分析

### A.1 评估方法与数据集

我们通过实验探究广泛用于研究和工业领域的中文语言模型中所包含的性别偏见。我们采用 Bolukbasi 等人 (2016)；Jiao 和 Luo (2021)提出的方法来评估性别偏见。单词的性别偏见分数$\vec{}$ $\vec{}$根据其词向量计算，计算方式为 $\vec{w} \cdot (she - he)$。正值表示该单词与女性的相关性更高，负值则表示该单词与男性的相关性更高。性别偏见分数的绝对值越大，该单词表现出的偏见越强。

Srivastava 等人 提出了一个大型基准，其中包含一个列出已有中文职业词汇的数据集。Zhu 和 Liu 提出了 AGSS，这是一个人工创建的、包含性别偏见的中文词级形容词列表。为了衡量语言模型中包含的性别偏见，我们首先根据 Srivastava 等人，2022；Zhu 和 Liu，2020）提供的词表，使用投影法计算词语的性别偏见分数（Bolukbasi 等人（2016）；Jiao 和 Luo （2021）。我们比较职业词汇和形容词的性别偏见分数向量，以了解语言模型对词级习得性别偏见的影响。为了使观察结果更加清晰，我们进一步对职业词汇和形容词的性别偏见分数向量应用符号函数。热力图使用的相似度函数为皮尔逊相似度。

我们将 AGSS 作为黄金标准（Zhu 和 Liu，2020）、ERNIE（Zhang 等人，2019）以及基于混合语料训练的中文词向量（Qiu 等人，2018）之间的形容词进行上述比较，

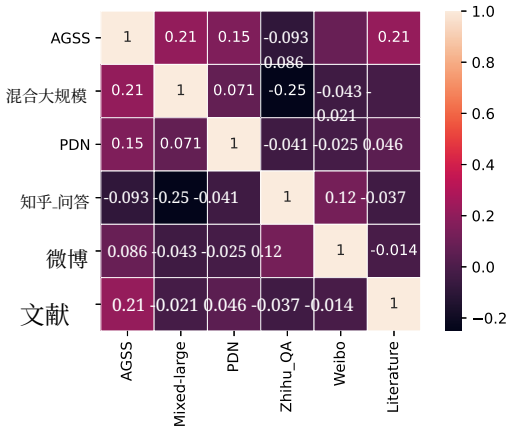



图4：使用不同语料库预训练的语言模型的形容词词级性别偏见比较。**PDN** 表示人民日报语料库。

并对 tecui-etal-2020-revisiting 提出的 Chinese-XLNet、Chinese-Bert 和 Chinese-Electra 进行比较，以生成图3。我们对 ERNIE（Zhang 等人，2019）与 tecui-etal-2020-revisiting 提出的 Chinese-XLNet、Chinese-Bert 和 Chinese-Electra 之间的职业词汇进行上述比较，以生成图2。由于基于混合语料训练的中文词向量的词典中缺少相当数量的职业词汇，因此没有使用这些词向量进行上述职业词汇实验。

我们没有提供黄金标准向量（Sri- vastava 等人，2022），因为他们没有对职业词汇进行人工性别偏见分析。我们还对不同语料库预训练的中文词向量中的形容词进行了上述对比，包括 Mixed-large 语料、人民日报、知乎问答数据集、微博和中文文学数据集，结果见图4，并分析由使用不同数据集进行预训练所导致的性别偏见差异




图5：ERNIE 和 Chinese-XLNet 的词云分析示例。**Ch** 表示中文。**En** 表示单词的英语机器翻译。**Man** 和 **Woman** 分别表示嵌入更接近男性和女性的单词。**Adj** 表示形容词。**Career**表示职业词汇。

来预训练语言模型。

### A.2 讨论

根据图3，开源中文语言模型中存在可观察到的性别偏见，尤其是在 Ernie 和中文词向量中。我们推测，这一现象与所使用的语料库密切相关。Cui 等人声称，他们使用的语料库由 Chine- seWiki 和其他一些通用中文数据集组成，包括百科全书、新闻和问答数据集。与之形成鲜明对比的是，Ernie 和中文词向量使用的语料库包含来自文献、论坛和其他社交媒体的句子，这可能导致模型产生性别偏见。

根据图4，人民日报和中文文学语料库都存在可观察到的性别偏见。这一观察表明，研究人员在使用文学数据训练语言模型时应更加谨慎。我们还推测，这是由文学语料库和人民日报造成的，因为其中包含更多描述性表达。

## B 语料库

### B.1 词云分析

我们在形容词和职业词汇部分提供了 Ernie 和 Chinese-Electra 的词云分析。更多词云分析结果可在我们的公共代码库中获取。单词根据其性别偏见分数的绝对值进行排序，该分数按照Bolukbasi 等人；Jiao 和 Luo所用的方法计算。词云显示出明显的词级性别刻板印象。例如，男性强壮，女性慈爱；男性适合当健身教练，女性适合当编舞师。我们还对使用不同语料库预训练的语言模型进行了词云分析。

### B.2 质量监控与控制

我们采用标准化操作方法，并对标注人员进行培训，以实现高质量标注，具体如下：

(1). **标注人员** 我们共有 6 名标注人员，均为中文母语者。标注人员

只有在阅读了若干关于社会学（King 等人，2021；Xu 等人，2019）和计算机科学研究（Sun 等人，2019；Zhao 等人，2018）的著作，并在进行标注流程前了解性别偏见后，才有资格参与标注。所有标注人员均拥有学士学位。Waseem 指出，专家标注员会更加谨慎，并能大幅提升语料库质量，这证明了开展训练流程的必要性。我们还确保男性标注员和女性标注员人数相等。

(2). **原始语料库的性别平等**在原始数据收集过程中，我们确保与男性相关的关键词数量和与女性相关的关键词数量相等，并使依据不同关键词召回的样本数量保持平衡。因此，原始数据和最终数据都应保持性别平等。

(3). **标注流程**我们的标注流程分为两个阶段。第一阶段，我们鼓励标注人员不要录入自己无法确定的样本。第二阶段，我们安排标注人员交叉核查标注结果。我们没有录入任何相互矛盾的样本。

(4). **标注者间一致性**鉴于该数据集的领域和用途，我们希望尽可能构建高质量的数据集。在完成首轮由 6 名标注人员参与的标注后，我们还在表5中报告了标注者间一致性，以验证标注可靠性：三名标注人员在偏见分类、检测和缓解方面的标注者间一致性分别为 0.802、0.935 和 0.987。

| | 分类 | 检测 | 缓解 |
|---|---|---|---|
| IAA | 0.802 | 0.935 | 0.987 |

表 5：标注者间一致性（标注者间一致性）

## C 实现细节

对于**性别偏见分类挑战**，我们使用了经过微调的 Chinese-BERT-wwm 、 Chinese-ELECTRA-180g-base 和 Chinese-XLNet-base（Cui 等人，2020），以及在上下文学习范式下使用的 GPT-3（Curie）。我们首先使用训练集，将多个已标注样例保存到一个具有特定文件 ID 的文档中。然后，我们使用测试集对已保存的文件执行分类查询。性别偏见分类的处理时间约为 1 小时。我们计算了精确率、召回率和 F1 分数，以分析模型性能。

对于**性别偏见检测挑战**，我们使用与分类挑战相同的基线模型集合。我们测试了对“是”和“否”进行检测的性能。检测任务也使用 GPT3（Curie）的分类端点；由于训练和测试均使用了更大的数据集，因此相比分类需要更多时间。

对于**性别偏见缓解挑战**，由于成本高昂，且与 Curie 和 Babbage 相比未观察到性能提升，我们没有提供在 CORGI-PM 上对规模最大的 Davinci（175B）GPT-3 进行微调的实验结果。微调实验的设置遵循 GPT-3 官方 API 中补全模型的教程，并将人工标注人员提供的真实编辑结果视为原始句子的补全内容。对于零样本实验的设置，我们采用 GPT-3 编辑模型，并将指令设为“消除句子中包含的性别偏见。”

对于所使用的指标，一方面，我们从性别偏见和连贯性两个方面，对 CORGI-PM 进行了广泛的人工评估。针对性别偏见和连贯性，我们打乱了人工标注员和不同模型的更正结果，并请标注人员在不了解来源信息的情况下，按照 1—完全没有至 7—极度有性别偏见／极其流畅的评分范围为结果打分。另一方面，我们还提供了模型在 CORGI-PM 上的自动化指标结果，包括 BLEU（Papineni 等人，2002）、ROUGE-L（Lin，2004）和 METEOR（Agarwal 和 Lavie，2007）。BLEU 是最早且应用最广泛的机器翻译和自然语言生成任务指标。METEOR 引入了 WordNet 和其他外部资源，以提升基于 BLEU 的指标的稳健性。与 BLEU 相比，ROUGE 更注重召回率。

在性别偏见纠正任务中，我们对 GPT3 的 Ada 、 Babbage 和 Curie 模型进行微调，并使用上述指标测试其性能。在训练集上进行微调，Ada 需要 31 分钟，Curie 需要 35 分钟，Babbage 需要 43 分钟。在生成过程中，每个微调后的模型约需 30 分钟才能完成。对于零样本范式，我们使用“消除句子中的性别偏见”作为指令，并使用 Davinci。与使用微调模型进行句子纠正相比，零样本范式需要更长时间（约

| 原文 | 参考译文 | 性别互换 |
| --- | --- | --- |
| 但是再坚强的你也是个女人吧。<br>（但你仍然是个~~女人，~~即使你很强大。） | 可是，再坚强的你，也会有脆弱的时候吧。<br>（但即使是你们中最坚强的人，也会有脆弱的时候，对吧？） | 但是再坚强的你也是个男人吧。<br>（但你仍然是个~~男人，~~即使你很强大。） |
| 可怕可恨的是，有些女性自己也是这样给自己定位的——没有反对，没有抗争，有的只是心甘情愿、死心塌地遵照执行。 | 女人不应该这样给自己定位——没有反对，没有抗争，有的只是心甘情愿、死心塌地遵照执行。 | 可怕可恨的是，有些女性自己也是如此给自己定位——没有反对，也没有抗争，只有心甘情愿、死心塌地地遵从执行。 |
| 可怕的是，有些女性<br>她们对自己的定义就是如此<br>她们自己——没有反对，也没有抵抗；只有一种意愿<br>严格遵循它。） | （一位~~女性~~不应该这样定义自己<br>不该以这种方式——没有反对，<br>没有抵抗，只有一种<br>愿意严格遵循它的态度。） | （可怕的是，有些男人<br>对自己的定义如此明确<br>——毫无异议，也不抵抗，<br>只有甘愿<br>严格照办。） |
| 那个时候，能驾驭gold high heels的女人得多摩登啊！<br>(那个时候，女人能够<br>驾驭金色高跟鞋的人一定很现代！ | 那个时候，能驾驭金色高跟鞋的人一定很摩登啊！<br>(那时候，得多时髦的人才<br>能驾驭那双金色<br>高跟鞋！） | 那个时候,能驾驭gold high heels的男人得多摩登啊！<br>(那时候，男人能驾驭<br>那双金色高跟鞋一定非常现代！） |
| “鸿雁队”是清一色的“理工男”。<br>（“鸿雁队”团队全是男性，都从事科学、技术、工程和数学领域。） | “鸿雁队”是清一色的“理工科学生”。<br>（全体“鸿雁队”队员中的科学、技术、工程和数学领域学生。） | “鸿雁队”是清一色的“理工女”。<br>（全体“鸿雁队”队员中女性，从事科学、技术、工程和数学领域。） |

图 6：性别互换方法生成的无意义句子案例研究。

| | 原句 | 修改后的句子 |
| --- | --- | --- |
| 更改代词 | 清洁女工一边扫地一边赞扬。 | 清洁工一边扫地一边赞扬。 |
| | （女清洁工一边<br>扫地一边赞扬。） | （清洁工一边赞扬，同时扫着地。） |
| | 我，有时文静，有时却调皮得像个男孩。 | 我，有时文静，有时调皮。 |
| | （我有时文静，但有时<br>像个男孩一样调皮。） | （我有时文静，有时调皮。） |
| 更改性别特定形容词 | 在小王眼里，李某高大帅气、温柔体贴，而且风趣幽默，是一位十分优质的年轻男性。 | 在小王眼里，李某身材高大、外表好看、温柔体贴，而且风趣幽默，是一位十分优质的青年。 |
| | （在小王眼里，李某身材高大，<br>英俊、温柔体贴，而且<br>风趣幽默，是一位十分优质的年轻<br>男性。） | （在王看来，李身材高挑，长相出众、体贴温柔——<br>而且风趣，是一位<br>非常优秀的年轻人。） |
| | 沙丘起伏，金光灿灿，宛如一座金山，像绸缎一样柔软，像少女一样娴静。 | sand dunes起伏，金光灿灿，宛如一座golden mountain，像绸缎一样柔软、宁静。 |
| | （沙峰起伏不平，<br>金光灿灿，宛如一座金山，<br>像丝绸般柔软，像一位少女般宁静。<br>少女。） | （沙丘起伏绵延，<br>金灿灿的，宛如一座金山，柔软得<br>如丝绸般宁静。） |
| Add 评论 | 我想要世界，而世界当时属于男人们。 | 我想要世界，而世界当时属于男人们。评：世界应当属于人们，与男女无关。 |
| | （我想要世界，而世界那时<br>属于男人。） | （我曾想要世界，而世界那时<br>属于男人。评论：世界<br>应该属于所有人，而不是<br>男人和女人。） |
| | 哎哟，果然每个追逐梦想的男人背后，都有一个不世俗的后方！ | 哎哟，果然每个追逐梦想的男人背后，都有一个不世俗的后方！评：这种感慨是错误的，将男女的家庭分工固定化，剥夺女性就业的权利，应予以鄙弃。 |
| | （哎呀，的确，每个追逐梦想的男人背后——<br>追逐男人，有一个<br>粗俗的女人！） | （哎呀，的确，每个追逐梦想的男人背后——<br>都有一个粗俗的<br>女人！评论：这种感觉是错误的<br>这固定了家庭劳动分工<br>在男女之间的划分，并剥夺<br>女性的就业权利，<br>这理应受到鄙视。） |

图 7：缓解标注模式案例研究。

1 小时）。

## D 案例研究

如图所示 6，性别互换方法难以缓解由性别相关描述和归纳所表达的性别偏见，以及所表达的性别刻板化的态度、规范与信念。因此，在某些情况下，性别互换方法可能会生成无意义句子。

我们还采用了基本的缓解标注模式（图 7）。这三种主要的缓解标注模式并非仅单独用于标注，也可选择组合使用。除上述三种模式外，我们还运用其他几种语言处理技巧，包括删除性别特定代词，以及替换性别相关隐喻中的交通工具，以缓解性别偏见，同时保持语义信息不变。