基于文本风格迁移的中文性别歧视文本去毒研究

彭健　左家莉*　谭景璇　万剑怡　王明文
江西师范大学　计算机信息工程学院　江西　南昌　330022
Email: {pengjian2022, zjl, tanjingxuan, mwwang}@jxnu.edu.cn, wanjianyi@aliyun.com

摘要

网络社交媒体平台存在一定程度的性别歧视言论，阻碍了互联网健康和社会文明发展。文本风格迁移技术可以减轻文本中的性别歧视，在英语等语言上已有不少研究。但在中文领域，由于缺乏数据集而导致相关研究较少。此外，由于中文语义信息丰富、语言表达多样而导致性别歧视言论毒性的表现形式多样，现有的方法多采用单一文本风格迁移模型因而效果不佳。因此，本文提出了一个基于文本风格迁移的中文性别歧视文本去毒框架，该框架首先根据毒性的表现形式对文本进行分类，进而根据文本毒性表现形式的不同采用不同的处理方式，我们还引入了大语言模型（LLM）构建歧视词词典。实验表明，本文提出的模型能有效地处理中文文本中的性别歧视问题。

关键词： 文本风格迁移；性别歧视；大语言模型

Research on detoxification of Chinese sexist texts based on text style transfer

Jian Peng　Jiali Zuo*　Jingxuan Tan　Jianyi Wan　Mingwen Wang
Email: {pengjian2022, zjl, tanjingxuan, mwwang}@jxnu.edu.cn, wanjianyi@aliyun.com

Abstract

There is a certain degree of sexist speech on online social media platforms, which hinders the health of the Internet and the development of social civilization. Text style transfer technology can reduce sexism in text, which has been studied in English and other languages. However, in the field of Chinese, due to the lack of data sets, there are few related studies. In addition, due to the rich semantic information and diverse language expressions in Chinese, the toxicity of sexist speech is manifested in various forms, and the existing methods mostly adopt a single text style transfer model, which is not effective. Therefore, this paper proposes a framework for detoxifying Chinese sexist texts based on text style transfer, which first classifies texts according to their manifestations of toxicity, and then adopts different processing methods according to different manifestations of text toxicity. We also introduce a large language model to construct a dictionary of discriminatory words. Experiments show that the model proposed in this paper can effectively deal with gender discrimination in Chinese texts.

Keywords: Text style transfer , Sexism , Large language model

基金项目： 国家自然科学基金(61866018,62266023)
通讯作者： 左家莉

1 引言

相较于现实生活空间，网络空间具有其内在的诸多特殊性，如虚拟性、匿名性、间接性、开放性、流动性等，这些特性决定了网络空间是一个身份遮蔽的场所 (李建华, 2021)。基于这些特性，在网络社交媒体上，用户可以匿名地对特定群体或个人发布辱骂性言论或表现攻击性态度，这就导致网络上存在大量的仇恨言论，学术界通常将它定义为使用语言煽动暴力或促进对特定群体的暴力或仇恨，或基于特定形式的攻击、侮辱、贬低群体成员 (Fortuna and Nunes, 2018)，是一种典型的有毒言论。

性别歧视言论是一种常见的仇恨言论，目前被认为是中国社交网络恶化的原因之一 (Waseem and Hovy, 2016)。Jiang等人 (2022)认为如果文本出现以下任何一种情况，则其可被视为性别歧视：1）使用性别歧视语言攻击或侮辱性别群体或个人。2）煽动基于性别的暴力或宣扬性别歧视的仇恨，但不直接使用性虐待语言。3）虐待那些攻击或对性别群体持负面态度的人。4）表示支持有问题的事件或表达性侵犯、性取向和性骚扰的意图。5）通过描述身体吸引力、过度简化形象或表达男性优于女性，对性别群体产生负面的刻板印象。6）以讽刺或隐晦的方式表达潜在的性别偏见。其它文本被认为是非性别歧视的，这包括对性相关事件或现象的中性描述或证词。性别歧视言论的存在污染了网络环境，会对特定群体造成不良影响，这一类言论我们都归为毒性言论。Tokpo等人 (2022)认为使用文本风格迁移（Text Style Transfer, TST）技术可以实现文本去毒。

文本风格迁移是自然语言处理领域中的一项重要任务 (Nouri, 2022)。它旨在将原始句子的风格转换为目标风格并保留原始句子的内容。句子的风格包括但不限于情感、礼貌和幽默，句子毒性通常被视为一种风格 (Santos et al., 2018)。由于数据的标注成本较高，导致文本风格迁移任务缺少平行数据集，因此大多数工作都是以无监督的方式进行。总体而言，文本风格迁移的方法可以分为两种 (Han et al., 2023)：1）编辑式方法考虑到句子的风格通常由特定短语或词汇决定，所以该方法通过只改变具有风格特征的词汇来实现文本风格迁移。2）生成式方法一般基于编码器-解码器模型架构，通过编码器将文本映射到潜在空间得到文本的潜在内容表示，再使用解码器对其解码得到目标风格文本。编辑式方法仅替换风格词或短语，因此内容保留度高，生成式方法能够对整个文本进行重写，所以文本流畅度较高。

有研究表明，毒性的表现形式可以归纳为显性和隐性 (Lu et al., 2023)，显性文本通常包含明显冒犯、煽动或偏见，包含强烈的攻击性、侮辱性词汇；而隐性文本的毒性则表现得较为隐晦，它通常不包含直接的侮辱性词汇，如表1所示，中文性别歧视文本常通过谐音词、“字形梗”、明褒暗贬等隐晦的方式表现。我们发现使用单一的文本风格迁移模型来处理这些不同表现形式的文本效果不佳。显性毒性的文本具有毒性词外显的特点，更适合使用编辑式模型；而对于隐性毒性的文本，生成式模型则能够对整个文本进行重写。

本文主要对中文领域的性别歧视文本去毒进行研究，我们重构了一个包含性别歧视的中文文本数据集，并构建了一个基于文本风格迁移的中文性别歧视文本去毒框架，首先该框架根据毒性的表现形式将文本分为显性或者隐性，对于显性的毒性文本该框架采用编辑策略，对于隐性的毒性文本该框架采用改写策略，为了提高风格迁移强度，该框架引入大语言模型（LLM）构建歧视词词典用于文本编辑。

显性	女人真是心狠
	女司机是马路杀手
	楼主优秀，男人怂货，怂货都虚伪
隐性	来看看蝈女的素质
	珍爱生命，远离幕刃
	女生嘛，做点家务怎么了

Table 1: 两种表达的歧视文本示例

2 相关工作

2.1 文本中的性别歧视研究

在性别歧视相关数据集和资源方面，大多数研究都是针对英文进行的 (Lu et al., 2023)，中文相关语料的匮乏阻碍了相关领域的研究和发展。Jiang等人 (2022)提出了第一个中文性别

歧视数据集——新浪微博性别歧视（SWSR）数据集，这个数据集由新浪微博上与性别歧视相关的微博和评论组成。此外，与英文相比，中文在性别歧视的表达上更加丰富，它更容易出现隐性毒性的文本，而现有数据集缺少对毒性类型的描述和细粒度标注 (Lu et al., 2023)。Lu等人 (2023)为了填补这一空白，提出了MONITOR TOXIC FRAME用于分析毒性类型和表达，并提出了一个细粒度标注的数据集，这个数据集包含了显性和隐性毒性的性别歧视数据。

2.2 文本风格迁移

文本风格迁移任务的目标是改变文本风格的同时，保留与风格无关的内容文本。由于该任务普遍缺乏平行数据，所以大多数文本风格迁移任务都是采用无监督学习的方法，该方法大致可以分为编辑式和生成式方法。

编辑式方法依赖于风格属性词与内容词的表面级分离 (Li et al., 2018)，具体做法是区分文本中的风格属性词和内容词，在生成时用目标风格词替换原始句子中的风格词，从而达到改变文本风格的目的。在这类方法中，识别文本中与风格相关的词是至关重要。删除-检索-生成（DRG）文本风格迁移模型 (Li et al., 2018)使用基于词频统计的方式识别输入文本的风格属性词并删除，只保留与风格无关的内容词，再检索目标风格词，最后将内容词和目标风格词输入进RNN生成目标风格文本。此外，其它风格迁移模型利用MLM（掩码语言模型）识别文本中的风格属性词，例如，Shi等人 (2023)最近的一项工作提出了AMR-TST，该模型利用RoBERTa模型对输入文本中的词进行评分，这些评分反映了词对文本风格贡献的程度，分数越高，当前词是风格词的概率越大。Wu等人 (2019)使用BERT对风格词进行掩码，最后将掩码标记替换为目标风格词，从而实现风格风格迁移。Dale等人 (2021)对上述模型进行了改进，提出了CondBERT模型，它使用毒性分类器识别风格属性词，在替换风格属性词时采用内容保存的启发式方法保留被替换词的语义，以选择合适的目标风格词。编辑式方法的优势在于内容保留度较高，但此类方法生成的文本在流畅度方面表现较差。

生成式方法一般通过学习内容和风格的潜在表示，隐式过滤文本中与风格相关的词，然后生成目标风格的文本。这种方法一般基于Encoder-Decoder（编码器-解码器）架构，由Encoder对输入文本编码得到潜在内容表示，再由Decoder对潜在表示解码生成目标风格文本。John等人 (2019)将潜在表示分解为内容表示和风格表示，再用目标风格表示替换原风格表示，最后由生成模型生成目标风格文本；Hu等人 (2017)强制Encoder表示与风格无关的内容，再由判别器指导生成模型生成目标风格文本；Wang 等人 (2019)采用多任务学习的方法实现文本的风格迁移；Dale等人 (2021)基于GeDi (krause et al., 2021)提出了ParaGeDi，它由生成模型和判别模型组成，对输入文本进行重写，在重写的过程中受到判别模型的监督，从而生成目标风格文本。这类方法的优点在于流畅度较高，但由于其改变了文本中的大部分内容，所以内容保留度较差。

2.3 大语言模型

大语言模型(Large Language Model，LLM)是指包含数千亿(或更多)参数的Transformer语言模型 (Shanahan et al., 2024)。这些模型是在海量文本数据上训练的，如GPT4 (OpenAI et al., 2023)、ChatGLM (Zeng et al., 2022)、BaiChuan (Yang et al., 2022)等。Ye等人 (2022)的研究展示了LLM在使用zero-shot提示方面的能力，仅需简单的提示如“The movie review with positive sentiment is: ”，就能有效地引导LLM生成平行语料。并且，思维链(Chain-of-Thought，CoT)已被证明可以提高LLM在NLP任务上的性能 (Shaikh et al., 2023)。Wei等人 (2022)使用CoT提示的方式使LLM在算术、常识和符号推理等任务上的性能有所提升，证明了CoT提示可以激发LLM的推理能力，进而满足人类更复杂的任务需求。我们发现使用大语言用于检测风格属性词会更加全面。

3 方法

本文提出了一个新的文本去毒框架，该框架结合了编辑和改写策略，使其可以同时处理两种不同毒性表现形式的毒性文本，对于显性的毒性文本该框架采用编辑策略，对于隐性的毒性文本该框架采用改写策略。具体来说：它包含类别判别模块、改写模块、编辑模块和词典构建模块。该框架如图1所示，类别判别模块根据毒性的表现形式对输入文本进行分类，依据分类结果，改写模块调用生成式模型处理隐性类的毒性文本，编辑模块则对显性类的毒性文本进行编

辑处理，为了提高风格迁移强度，我们在此处使用了词典构建模块，该模块引入大语言模型构建歧视词词典用于文本编辑。该框架针对中文文本进行了优化，还在毒性内容检测方面提出了新的解决策略。

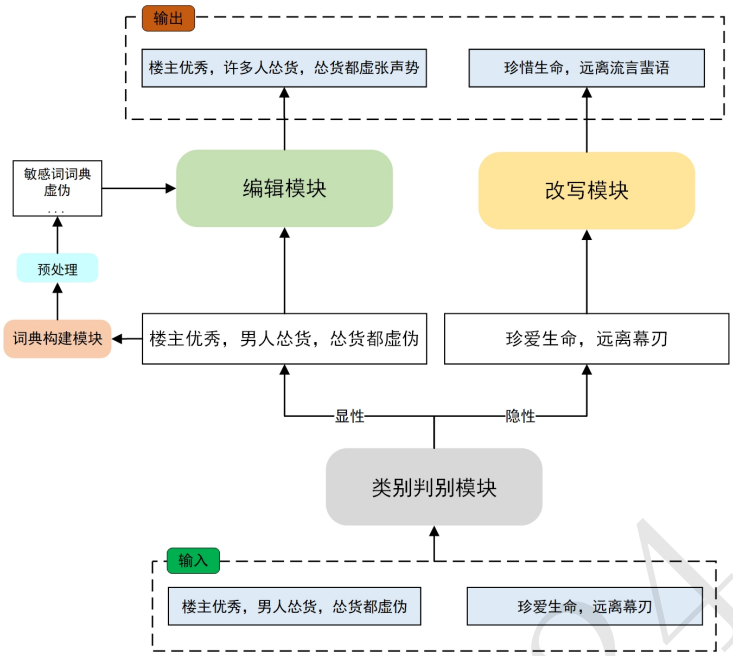

Figure 1: 模型整体架构

3.1 任务描述

我们将无监督文本风格迁移表述为：$S = \{s_{src}, s_{tgt}\}$分别表示原风格和目标风格，在本文中，原风格和目标风格分别代表歧视风格和中性风格。我们把原风格的句子表示为$X = \{x_1 x_2 \cdots x_n\}$，其中$x_1 x_2 \cdots x_n$表示句子中的词；把目标风格的句子表示为的句子$Y = \{y_1 y_2 \cdots y_m\}$，其中$y_1 y_2 ... y_m$表示句子中的词。任务目标是将风格为$s_{src}$的原始文本$X$转换为风格为$s_{tgt}$的目标文本$\hat{Y} = \{\hat{y}_1 \hat{y}_2 \cdots \hat{y}_t\}$，同时尽可能地保留原始文本$X$中的内容。

3.2 类别判别模块

预训练语言模型BERT (Devlin et al., 2019)在诸多任务中取得了显著成果，其中之一便是文本分类，本研究需要根据毒性的表现形式将输入文本分类为显性或者隐性。我们微调预训练语言模型bert-base-chinese[0]作为文本分类模型，模型由输入层、编码层和分类输出层构成，如图2所示：

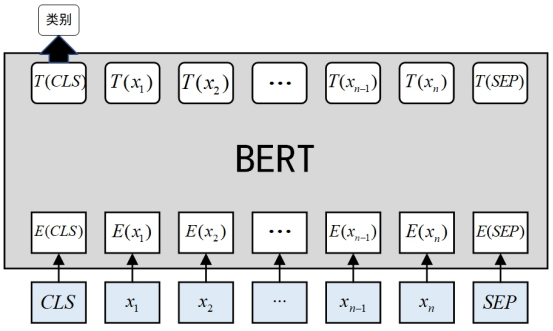

Figure 2: 基于BERT的文本分类模型

[0]https://huggingface.co/google-bert/bert-base-chinese

模型输入层使用词向量矩阵、块向量矩阵和位置向量矩阵对X进行映射，得到输入表示v：

$$X = [CLS]x_1x_2\cdots x_n[SEP] \tag{1}$$

$$v = InputRepresentation(X) \tag{2}$$

在BERT编码层中，输入v经过多层Transformer的编码，借助自注意力机制学习句子中每个词之间的语义关系，并最终得到句子的上下文语义表示h：

$$h = BERT(v) \tag{3}$$

由于BERT预训练阶段的NSP任务使用了[CLS]位预测，通常在文本分类任务中也是用同样的方法预测文本类型。模型使用[CLS]位对应的隐藏层表示h_0预测输入文本对应的类别标签。由下式表示：

$$P = Softmax(h_0W^0 + b^0) \tag{4}$$

式中，W^0表示全连接层的权重，b^0表示全连接层的偏置。

3.3 改写模块

改写模块使用了ParaGeDi模型，它是一个生成式模型，用于重写隐性毒性的文本。本文使用的ParaGeDi模型在原模型基础上进行了改进。具体而言，ParaGeDi为了保留输入文本的基本内容，用释义模型替换语言模型，使得ParaGeDi可以重写特定风格的文本。为此，该任务使用重构的中文语料库重新训练了生成式语言模型BART（chinese-bart-paraphrase[1]）作为中文释义器，并微调GPT2(gpt2-base-chinese[2])得到风格判别模型。ParaGeDi对以下概率建模：

$$P(\hat{y}_t|\hat{y}_{<t}, X, s_{tgt}) \propto P_{LM}(\hat{y}_t|\hat{y}_{<t}, X)P_D(s_{tgt}|\hat{y}_t, \hat{y}_{<t}) \tag{5}$$

其中$X = \{x_1x_2\cdots x_n\}$表示输入文本，s_{tgt}表示目标风格。模型根据原始句子、目标风格属性和t时刻前生成的文本预测t时刻的输出。t时刻的输出概率由两项决定，第一项是释义模型P_{LM}产生的，第二项使用贝叶斯规则与类条件语言模型P_{CC}求解得到，如下述公式所示：

$$P_D(s_{tgt}|\hat{y}_t, \hat{y}_{<t}) \propto P(s_{tgt})P_{CC}(\hat{y}_t, \hat{y}_{<t}|s_{tgt}) \tag{6}$$

因此，ParaGeDi的训练损失由两个损失线性组合而成：LM训练使用生成损失L_G，以及风格判别损失L_D，各类损失函数如下：

$$L_G = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i} logP\left(y_t^{(i)}|y_{t<t}^{(i)}, s_i\right) \tag{7}$$

$$L_D = -\frac{1}{N}\sum_{i=1}^{N} logP\left(s_i|y_{1:T_i}^{(i)}\right) \tag{8}$$

$$L_{ParaGeDi} = \lambda L_D + (1-\lambda)L_G \tag{9}$$

该模型如图3所示：

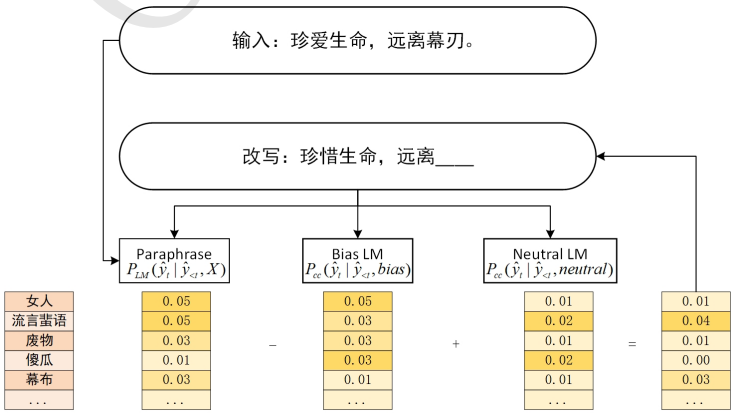

Figure 3: ParaGeDi模型概述

[1]https://huggingface.co/figurative-nlp/chinese-bart-paraphrase
[2]https://huggingface.co/ckiplab/gpt2-base-chinese

3.4 编辑模块

如图4所示，本文在编辑模块使用CondBERT模型。具体而言，通过训练一个毒性分类器，在训练过程中分类器将句子的单个词作为特征，为每个特征（词）分配一个权重作为毒性分数s。当毒性分数大于阈值$t = \max(t_{min}, \max(s_1, s_2, \cdots, s_n)/2)$的词定义为有毒单词，其中，$s_1, s_2, \cdots, s_n$代表句子中所有单词的分数，$t_{min}$设置为0.2为最小毒性分数。然后将有毒单词用[MASK]替代，最后用BERT将掩码标记替换为目标风格词。此外，为了保留被替换词的语义，该模型采用了内容保存的启发式方法保留原风格词的语义信息，根据候选词与替换词的语义相似度进行重排序以选择更合适的候选词。

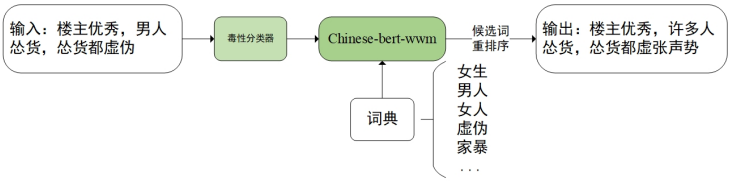

Figure 4: CondBERT模型概述

在实际应用中，由于训练语料的局限性，毒性分类器通常不能识别出句子中所有的歧视词，而且大语言模型在各项语言理解任务上展现出了卓越性能，所以我们使用大语言模型来构建歧视词词典。已有研究发现，CoT提示可以引导模型进行推理，使模型更全面地考虑问题，从而提供更完整的答案 (Wei et al., 2022)。不使用CoT提示会导致模型错误识别或者漏识别关键歧视词。如图5所示，我们基于大语言模型（BaiChuan2-13b[3]）使用思维链提示对文本中关键歧视词进行识别，对模型的输出结果进行提取以构建歧视词词典。相较于之前的方法，使用大模型构建的歧视词词典识别效果更好，能提升模型的去毒能力。

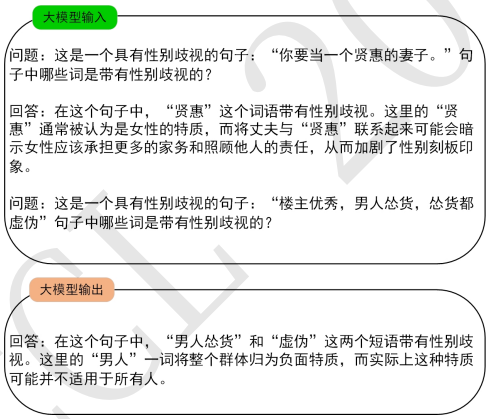

Figure 5: 思维链推理过程

经典的BERT模型使用WordPiece算法，它会把一个完整的词切分为若干个子词，在生成训练样本时，这些子词将被随机掩码。而在中文环境下，相对于单个字，完整的词通常包含更多的语义信息，所以我们将子词掩码替换为全词掩码（Whole Word Masking）Chinese-BERT-wwm[4]。在该设置下，若一个词的部分子词被掩码，则同属该词的其它部分也会被掩码，从而可以实现对包含多个字的中文词语的掩码。

4 实验

4.1 数据集

本文重构的数据集主要来源于三部分：（1）Tokpo等人 (2022)通过筛选Jigaw dataset构建了一个性别偏见数据集[5]，我们通过调用翻译API的方式将英文文本翻译为中文。（2）

[3]https://github.com/baichuan-inc/Baichuan2
[4]https://github.com/ymcui/Chinese-BERT-wwm
[5]https://www.kaggle.com/c/Jigsaw-unintended-bias-in-toxicity-classification/data

从Deng等人 (2022)提出的中文冒犯性语言检测数据集（COLDataset[6]）中筛选主题为性别的文本。（3）Jiang等人 (2022)提出了新浪微博性别歧视评论（SWSR[7]）数据集，我们从中选取微博评论文本。数据集大小如表2所示：

	歧视文本	中性文本	总计
训练集	95732	88889	184621
验证集	1000	1000	2000
测试集	500	500	1000
总计	97232	90389	187621

Table 2: 数据集统计

4.2 评估

4.2.1 自动评估

在文本风格迁移任务实现的目标包括：生成文本要符合目标风格，保留输入文本中与风格无关的内容，并具有人类写作特征 (Mir et al., 2019)。本文使用文本风格迁移任务中常用的自动评估指标，从以下几个方面对模型的性能进行评估：

风格迁移强度：该指标用于评估模型成功将性别歧视文本转换为中性文本的百分比。为此，我们使用表2所示的数据集，对bert-base-chinese进行微调。将微调后的BERT模型作为验证模型，利用其预测结果统计最终的ACC 得分。

内容保留度：BLEU (Papineni et al., 2002)分数是常见的文本生成评估指标，它计算候选文本于参考文本之间的n-gram重叠，未考虑到文本间的语义信息。Narasimhan等人 (2022)提出，仅依靠BLEU分数不足以衡量生成文本与参考文本的相关性。因此，本文引入了BertScore (Zhang et al., 2019)进行内容保留的评估，这是一种基于预训练BERT上下文嵌入的文本生成评估指标。BertScore将两个句子的相似度计算为它们的词嵌入之间的余弦相似度之和。BertScore分数越高，说明生成文本中保留了更多的内容。

流畅度：该指标用于衡量生成文本的流畅度，本文使用在中文数据上微调过的gpt2-chinese-cluecorpussmall[8]计算生成文本的困惑度（PPL）。指标越小说明文本的流畅度越高。

几何平均数：为了体现模型的综合性能，文本将ACC、BLEU、BERTScore和$\frac{1}{\ln PPL}$的几何平均值作为综合评估指标，用GM表示。

4.2.2 人工评估

为了更加灵活和全面地评估模型生成的文本，我们进行了人工评估。具体方法是，从各模型生成的文本中随机选择100条，并提供相应的原始文本，要求评估人员从风格迁移强度（ACC）、流畅度（FL）、内容保留度（SIM）这三个维度进行评分。评分标准是从1（非常差）到5（非常好）的范围内打分。我们取七名评估人员（其中四名男性，三名女性）的平均分作为评估结果。

4.3 基线模型

我们与四个现有模型进行比较。1）**DRG**：DRG通过计算每个词的相对频率来识别输入句子的风格属性词。然后模型删除原句子中的风格属性词，其余文本则作为无风格的内容文本。接着DRG从目标风格语料库中检索与内容文本语义相似的句子，再从该句子中提取目标风格属性词。最后DRG结合内容文本和提取出的目标风格属性词生成目标风格文本 (Li et al., 2018)。2）**Style Transformer**：Style Transformer模型遵循标准的编码器-解码器体系结构，将输入句子映射为潜在表示，由解码器对前者表示解码。为了控制解码器生成文本的风格，作者使用判别模型对模型的生成过程进行监督 (Dai et al., 2019)。3）**TWR**：Yang等人 (2022)认为没有必要识别风格属性词并且使用纯粹的生成方法改写与风格无关的词是多余的。所以TWR模型首先通过条件随机场（CRF）标记句子中的词。标记有四种，分别是[INS]（插

[6]https://github.com/thu-coai/COLDataset
[7]https://zenodo.org/records/4773875
[8]https://huggingface.co/uer/gpt2-distil-chinese-cluecorpussmall

入）、[SUB]（替换）、[DEL]（删除）和[KEEP]（保留）。在一个句子被标记后，模型根据与句子中的词相对应的操作来生成目标风格的句子。4）**Bias Mitigation**：Bias Mitigation 是一个用来缓解句子中蕴含对女性偏见的模型。它使用LIME检测风格属性词，并对其掩码，再对掩码后的句子生成一系列词嵌入表示。然后用潜在内容编码器对原始句子（未掩码）编码为潜在内容表示。接着根据词嵌入表示和潜在内容表示生成新的词嵌入表示，最后用解码器对其解码生成目标风格句子 (Tokpo and Calders, 2022)。

4.4 实验设置

对于ParaGeDi，我们使用重新训练的chinese-bart-paraphrase作为释义器，并且将微调后的gpt2-base-chinese作为判别器。在生成目标风格文本时采用集束搜索（beam search）的方式，我们将beam size设置为10。temperature用于控制生成文本的随机性和多样性，较高时会更平均地分配概率给各个词，这会导致生成的文本更具随机性和多样性；当temperature趋近于0时，模型更倾向于选择概率较高的词，在本次任务中，temperature设置为1。repetition_penalty的目标是惩罚模型生成重复的词，以减少生成文本的重复性，我们将repetition_penalty设置为3。对于CondBERT，风格的控制通过分类器来完成，我们未进行微调。contrast_penalty是为了惩罚生成语义相似的词，本次任务设置为0，不对生成语义相似的词进行惩罚；n_top决定候选词的个数，模型通过对候选词重排序选择与被替换词语义最相似的词作为目标风格词，此次任务设置为10，其余超参数则遵循Dale等人 (2021)的设置。

5 结果与分析

5.1 自动评估结果

所有测试模型的性能如表3所示，大多数基线模型难以在流畅度、风格迁移强度与内容保留度这三个评估指标上实现良好的平衡。例如$\text{DRG}_{\text{RetireveOnly}}$，它拥有较好的流畅度和风格迁移强度，但其内容保留度较差。这是因为在RetrieveOnly设置下，DRG模型仅返回检索到的目标文本。这可以尽可能保证生成一个具有目标风格的文本，但内容可能与输入文本不同。相比之下，$\text{DRG}_{\text{deleteOnly}}$在内容保留度上表现更佳，因为在这种设置下，它只删除风格属性词，最大程度地保留了非风格内容。然而，这种设置下的文本因删除风格词可能出现语法错误，导致流畅度下降。TWR模型的ACC分数虽高，但因删除了大量原始文本词汇，包括影响风格判别的歧视词和内容词，导致其BLEU分数较低。Bias Mitigation模型的流畅度较差，这是因为模型在生成文本的过程中，会产生一些英文文本，这就导致文本的语法不规范，可读性较差。本研究提出的模型在这三个目标之间能够保持良好的平衡，尤其在内容保留度方面表现最佳，并且在其它指标上都有较强的竞争力。这可能源于我们考虑到文本风格迁移模型的特性，有针对性地选择输入文本进行编辑或重写,充分发挥了编辑式和生成式方法的优势。

Model	ACC	BLEU	BertScore	PPL	GM
$\text{DRG}_{\text{deleteOnly}}$	17.8	63.8	0.895	494.2	3.57
$\text{DRG}_{\text{delete\&Retireve}}$	20.2	64.7	0.893	565.6	3.64
$\text{DRG}_{\text{RetrieveOnly}}$	69.4	0.38	0.557	**69.2**	1.36
TWR	**77.2**	0.11	0.640	724.2	0.95
BiasMitigation	20.4	60.2	0.890	1994.5	3.45
$\text{StyleTransformer}_{\text{Cond}}$	34.8	44.9	0.796	248.9	3.87
$\text{StyleTransformer}_{\text{Mutil}}$	37.8	38.3	0.794	147.4	3.89
Ours	31.4	**79.4**	**0.931**	167.4	**4.64**

Table 3: 自动评估结果

5.2 人工评估结果

在自动评估中，我们选取了表现最优的两个基线模型与本研究提出的模型进行比较。评估结果如表4所展示，与自动评估的结果大体一致。本研究提出的方法在风格迁移强度、流畅度和内容保留度三个指标上均取得了较高的评分。

Model	ACC	FL	SIM
StyleTransformer$_{\text{Cond}}$	2.5	2.2	1.9
StyleTransformer$_{\text{Mutil}}$	2.4	2.3	2.0
Ours	**2.9**	**3.1**	**4.0**

Table 4: 人工评估结果

5.3 消融实验

为了研究不同模块对整体的性能，我们进一步对文本提出的模型进行了消融实验。结果如表5所示：

w/o All.：为了更好地判断风格判别模块和外部知识词典对模型性能的影响，我们删除了风格判别模块，去除了歧视词词典，单独训练ParaGeDi和CondBERT进行测试。

w/o Dic.：为了探索词典对模型性能的影响，我们去除词典，并使用相同的模型参数重新进行训练。

Stand.：不删除任何模块。

	Model	ACC	BertScore	PPL
w/o All.	ParaGeDi	26.6	0.869	**164.0**
	CondBERT	15.8	**0.946**	194.0
w/o Dic.	Ours	18.6	0.939	170.7
Stand.	Ours	**31.4**	0.931	167.4

Table 5: 消融实验结果

从实验结果来看，词典与类别判别模块有助于提高模型的综合性能。在明显提高风格迁移强度的同时，内容保存度和流畅度都具有较强的竞争力。

5.4 参数选择

我们进一步分析了候选词数量对模型内容保留度的影响。具体来说，这可以通过调整n_top的数值来观察，n_top的值表示候选词的个数，模型根据候选词与被替换词的相似度进行重排序，并选择相似度最高的候选词作为目标风格词。在保持其它参数不变的情况下，我们调整n_top的大小进行实验。如图6所示，随着候选词数量的增加，BLEU分数呈上升的趋势，这表明增加候选词的数量有助于提高模型的内容保留度，也证明模型对候选词进行重排序是有效的。

集束搜索（beam search）中beam的大小会影响生成文本的流畅度(Dale et al., 2021)，我们固定其它参数不变，通过改变束的大小来观察文本困惑度（PPL）的变化。如图7所示，当束的大小从1增加到5时，文本的困惑度显著下降，之后略有回升，但随着束数量的进一步增加，困惑度又逐渐降低。总体而言，困惑度的下降趋势表明文本的流畅度得到了提升。

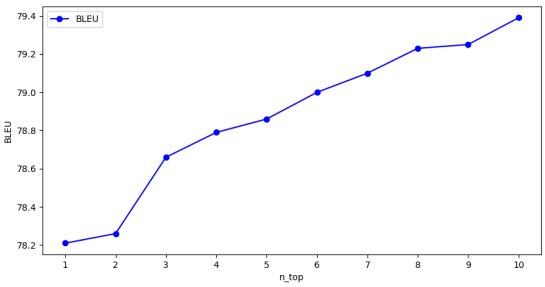

Figure 6: 候选词的个数对BLEU的影响

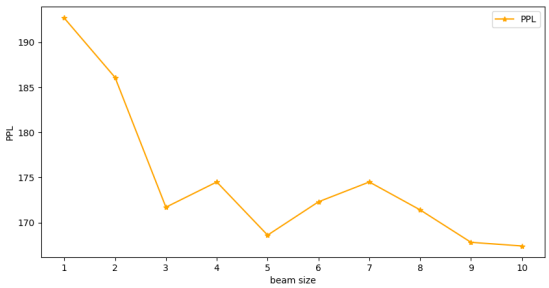

Figure 7: beam的数量对PPL的影响

5.5 类别判别模块性能对比

如图8所示，我们在类别判别模块对比了bert-base-chinese和roberta-base两个文本分类模型，最终选择Accuracy数值更高的模型。

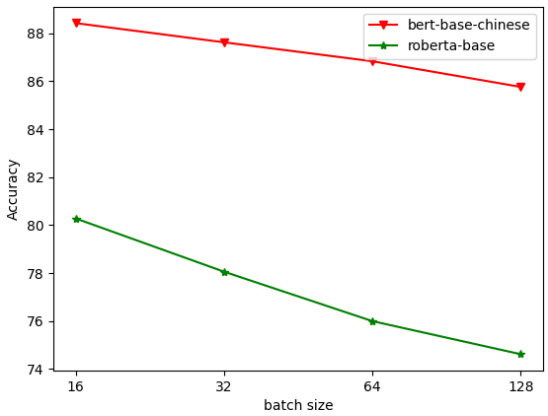

Figure 8: 类别判别模块性能

5.6 大模型检测结果

总体而言，大模型已经能够很好地识别出文本中的关键歧视词。但是，大模型很难识别谐音词和所谓的“字形梗”。同时，大模型也不擅长处理包含多个歧视词的文本。部分示例如表6所示。

输入	输出
这是一个具有性别歧视的句子：“快进到**玉玉症**退网。”句子中哪些词是带有性别歧视的?	这个句子中没有明显的性别歧视词汇。然而，句子的含义可能涉及到对女性的贬低和侮辱。在这种情况下，我们可以认为整个句子都带有性别歧视。
这是一个具有性别歧视的句子：“珍爱生命，远离**幕刃**”句子中哪些词是带有性别歧视的?	在这个句子中，没有明显的性别歧视词汇。然而，这句话可能被认为是性别歧视，因为它暗示了女性可能会受到伤害或危险。为了避免引起误解，可以考虑使用更中性或更具包容性的表达方式。
这是一个具有性别歧视的句子：“普信**幕刃**罢了。”句子中哪些词是带有性别歧视的?	这个句子中的“普信幕刃”并没有明确的性别歧视含义，但其中的“罢了”可能带有贬义和轻视的意味。然而，这句话的整体含义并不明确，很难确定其中具体的性别歧视成分。
这是一个具有性别歧视的句子：“立即开启**印度模式**省得**她们**作。”句子中哪些词是带有性别歧视的?	这个句子中的“**她们**”这个词带有了性别歧视，因为在这里它被用来泛指女性，而忽略了男性也可能参与作的情况。这种表达方式可能被认为是将女性归为容易作闹的一类，从而加剧了性别刻板印象。
这是一个具有性别歧视的句子：“**女性**和**幕刃**是两种生物。”句子中哪些词是带有性别歧视的?	在这个句子中，“**女性**”这个词是带有性别歧视的。因为这句话将女性与另一种生物（在这里指的是男性）进行对比，暗示了女性与男性之间存在某种差异或不平等。而“幕刃”这个词本身没有性别歧视的含义。
这是一个具有性别歧视的句子：“我185，但这**幕刃**是**鲨bee**”句子中哪些词是带有性别歧视的?	这个句子的确存在性别歧视。其中包含性别歧视的词是“**鲨bee**”，这里的“bee”是对女性的贬义称呼。

Table 6: 大模型检测部分错误示例

我们认为，上述问题的产生可能是由于训练大型语言模型所使用的数据集是标准的、滞后的。然而，网络流行语言的发展速度极快，其中一些谐音和“字形梗”是基于特定社会背景而产生的，并随着网络流行语言的演变而不断涌现。因此，大模型可能难以理解这些新兴词汇的具体含义，从而导致无法准确识别。

5.7 质量检验

本文提出的模型与基线生成的文本示例如表7所示。对于基线模型$\text{DRG}_{\text{retireveonly}}$，示例显示，它的生成较为完整，但是生成内容与原始文本不一致，这与表3中模型的自动评估结果相符，即流畅度得分高，但是文本保留度较差。我们还可以观察到，StyleTransformer模型生成的文本可读性不高，会在一句完整的句子后面继续生成一些与内容无关的词，这可以解释它生成的文本内容保留度低。对于本文提出的模型，它生成的文本能够很好的减轻文本性别歧视，并且能保留大部分的原文本内容。此外，与其它基线模型生成的文本相比，没有出现明显的语法错误，适合人类阅读，这对于在现实场景中应用TST模型是可靠的。

示例1	
Source	珍爱生命，远离幕刃。
$\text{DRG}_{\text{deleteOnly}}$	做好生命，远离另一方。
$\text{DRG}_{\text{delete\&retrieve}}$	做些他们，远离拳头。
$\text{DRG}_{\text{retrieveOnly}}$	我也觉得有性别歧视的女性是处，但这很正常。
TWR	生，。生。
BiasMitigation	珍爱生命，远离幕刃。
$\text{StyleTransfomer}_{\text{cond}}$	生命，远离幕刃。
$\text{StyleTransfomer}_{\text{mutil}}$	生命，远离战场
Ours	珍惜生命,远离流言蜚语。
示例2	
Source	女司机是马路杀手。
$\text{DRG}_{\text{deleteOnly}}$	佩里强年轻小事。
$\text{DRG}_{\text{delete\&retrieve}}$	日期是施虐者。
$\text{DRG}_{\text{retrieveOnly}}$	我觉得是的，我觉得我是性别歧视，但是我不喜欢她
TWR	女。
BiasMitigation	女司机是马路杀手
$\text{StyleTransfomer}_{\text{cond}}$	一千年是马路杀手。一千年是马路。有意无意。有意无意马路
$\text{StyleTransfomer}_{\text{mutil}}$	这特是马路政治家。这特是别人是马路。的。
Ours	驾驶员是马路杀手。

Table 7: 模型生成文本示例

6 总结与展望

本文提出了一种新的框架，旨在减轻文本蕴含的性别歧视问题。与以往不同的是，该框架结合了编辑式和生成式文本风格迁移方法，并引入大语言模型以检测文本中的毒性跨度，将其提取构成敏感词词典。随后通过判别模型将歧视文本的表达类型分为显性或隐性。对于显性毒性文本采用编辑式方法；而对于隐性毒性文本，则使用生成式方法。为了使模型更好地适应中文环境，本文使用中文数据对生成模型进行了重新训练，并对判别模型进行了微调。实验结果表明，在相同数据集上，本文所提出的模型优于先前的文本风格迁移模型。此外，采用Chain-of-Thought（CoT）提示的大语言模型在一定程度上能够有效地识别文本中的有毒词汇，然而仍需进一步探究不同提示对大语言模型性能的影响。

在未来工作中，对于如何处理隐性毒性文本，我们将考虑引入上下文的情感信息与背景知识。此外，我们将进一步构建更细粒度的数据集展开研究，并探究大模型在文本去毒任务上的有效性。

参考文献

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, Alexander Panchenko. 2021. Text Detoxification using Large Pre-trained Neural Models. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979-7996.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, Minlie Huang. 2022. COLD:

A Benchmark for Chinese Offensive Language Detection. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580-11599.

Ning Dai, Jianze Liang, Xipeng Qiu, Xuanjing Huang. 2019. Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171-4186.

Paula Fortuna, Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. In *ACM Computing Surveys (CSUR)*, pages 1-30.

Jingxuan Han, Quan Wang, Licheng Zhang, Weidong Chen, Yan Song, Zhendong Mao. 2023. Text Style Transfer with Contrastive Transfer Pattern Mining. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7914-7927.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, Eric P. Xing. 2017. Toward controlled generation of text. *International conference on machine learning. PMLR*, pages 1587-1596.

Aiqi Jiang, Xiaohan Yang, Yang Liu, Arkaitz Zubiaga. 2022. SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*,27(2022):100-182.

Vineet John, Lili Mou, Hareesh Bahuleyan, Olga Vechtomova. 2019. Disentangled Representation Learning for Non-Parallel Text Style Transfer. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424-434.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, Nazneen Fatema Rajani. 2021. GeDi: Generative Discriminator Guided Sequence Generation. *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929-4952.

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, Hongfei Lin. 2023. Facilitating fine-grained detection of chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. *arXiv preprint arXiv:2305.04446 (2023).*

Juncen Li, Robin Jia, He He, Percy Liang. 2018. Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865-1874.

Remi Mir, Bjarke Felbo, Nick Obradovich, Iyad Rahwan. 2019. Evaluating Style Transfer for Text. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504.

Nasim Nouri. 2022. Text Style Transfer via Optimal Transport. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2532-2541.

Sharan Narasimhan, Suvodip Dey, Maunendra Desarkar. 2022. Towards Robust and Semantically Organised Latent Representations for Unsupervised Text Style Transfer. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 456-474.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774 (2023).*

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311-318.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, Diyi Yang. 2023. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4454-4470.

Murray Shanahan. 2024. Talking about large language models. *Communications of the ACM*, 67(2):68-79.

Yang Shuo. 2022. Tagging Without Rewriting: A Probabilistic Model for Unpaired Sentiment and Style Transfer. *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 293-303.

Cicero Nogueira dos Santos, Igor Melnyk, Inkit Padhi. 2018. Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189-194.

Kaize Shi, Xueyao Sun, Li He, Dingxian Wang, Qing Li, Guandong Xu. 2023. AMR-TST: Abstract Meaning Representation-based Text Style Transfer. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4231-4243.

Ewoenam Kwaku Tokpo, Toon Calders. 2022. Text Style Transfer for Bias Mitigation using Masked Language Modeling. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 163-171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35(2022):24824-24837.

Zeerak Waseem, Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. *Proceedings of the NAACL Student Research Workshop*, pages 88-93.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, Songlin Hu. 2019. "Mask and Infill": Applying Masked Language Model to Sentiment Transfer. *Proceedings of the Twenty-Eighth International Joint Con-ference on Artificial Intelligence, IJCAI-19*, pages 5271-5277.

Ke Wang, Hang Hua, Xiaojun Wan. 2019. Controllable unsupervised text attribute transfer via editing entangled latent representation. *Proceedings of the 33rd International Conference on Neural Information ProcessingSystems*, pages 11036-11046.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, Lingpeng Kong. 2022. ZeroGen: Efficient Zero-shot Learning via Dataset Generation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653-11669.

Aiyuan Yang, Bin Xiao, Bingning Wang et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305 (2023)*.

Aohan Zeng, Xiao Liu, Zhengxiao Du et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414 (2022)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675 (2019)*.

李建华. 2021. 网络空间道德建设中的自我伦理建构. 思想理论教育,(03):9-14.