山西大学学报(自然科学版)47(3):528—538,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024001

中文文本去毒任务的研究

刘江盛,左家莉*,胡玉婷,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)

摘 要:文章旨在研究如何有效去除中文文本的毒性。针对此任务,文章重构了一个中文毒性语料集,以此作为任务研究的数据基础。基于此数据集文章探究了文本的毒性表现形式,同时对特定类别的毒性文本成因展开了分析。基于上述分析结果,文章使用基于编辑式、生成式两类文本风格迁移模型进行文本去毒,并进一步探究了大语言模型基于不同Prompt时去除文本毒性的表现。据实验结果表明,基于编辑式的模型能有效去除显式毒性文本的毒性,且具有较高的内容保存度,生成式模型生成的文本则有更高的流畅度。基于Prompt的大语言模型在一定程度上可以去除句子毒性,但相较于特定的风格迁移模型而言,小参数大语言模型的去毒能力还有待提高。

关键词:文本风格迁移;文本去毒;大语言模型
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2024)03-0528-11

Research on Detoxification Task of Chinese Texts

LIU Jiangsheng, ZUO Jiali*, HU Yuting, WAN Jianyi, WANG Mingwen
(*School of Computer Information Engineering, Jiangxi Normal University, Nanchang* 330022*, China*)

Abstract: The purpose of this paper was to study how to effectively remove the toxicity of Chinese texts. For this task, this paper reconstructed a Chinese texts toxicity corpus set, which was used as the data basis for task research. Based on this data set, this paper explored the toxic manifestations of texts, and analyzed the causes of specific types of toxic texts. Based on the analysis results above, this paper used two types of text style transfer models based on editing and generating to remove text toxicity, and further explored the performance of removing text toxicity based on different Prompts in large language models. According to the experimental results, the edited model can effectively remove the toxicity of explicit toxic text, and has a higher degree of content preservation, while the generated text has a higher degree of fluency. Prompt-based large language model can remove sentence toxicity to a certain extent, but compared with specific style transfer models, the detoxification ability of small parameter large language model needs to be improved.

Key words: text style transfer; text detoxification; large language model

0 引言

毒性文本主要是指句子中出现不尊重、身份攻击、侮辱、淫秽、粗鲁、威胁、讽刺等言论[1]。这类有毒文本的出现会损害和谐、健康的网络环境,对人类的心理健康造成不良影响。此外,大语言模型的训练依赖网络数据,若毒性文本成为大模型的预训练数据,则可能会导致模型产生令人反感和不道德的内容,这将极大阻碍大模型的广泛应用[2]。

收稿日期:2023-12-15;**接受日期**:2024-01-12
基金项目:国家自然科学基金(61866018)
作者简介:刘江盛(1998-),男,江西赣州人,硕士研究生,研究方向为自然语言处理。E-mail:sheng@jxnu.edu.cn
* **通信作者**:左家莉(ZUO Jiali),E-mail:zjl@jxnu.edu.cn
引文格式:刘江盛,左家莉,胡玉婷,等.中文文本去毒任务的研究[J].山西大学学报(自然科学版),2024,47(3):528-538. DOI:10.13451/j.sxu.ns.2024001

近年来,文本毒性的检测任务[3]受到了NLP(Natural Language Processing)领域较多学者的关注,但重写毒性文本,以保留句子中重要内容的同时去除句子毒性[4]这一任务的相关研究较少[5]。在现有的工作中,句子毒性通常被视为一种风格[6-8],使用文本风格迁移(Text Style Transfer,TST)能在保留句子原义的基础上,将句子的风格从“毒性”迁移到“中性”,从而达到去除句子毒性的目的。

由于数据的标注成本较高,这就导致TST任务(如情感迁移、文本去毒等)普遍缺乏监督训练所需的平行语料,因此大多数TST任务都采用无监督学习的方法。一般而言,无监督的TST方法主要基于两种框架:(1)编辑式,即句子的风格由特定词语构成,可通过检测、删除、替换等操作修改原句以实现句子的风格转换;(2)生成式,通过训练基于编码器-解码器的生成模型改写原句以实现风格转换,其中一种做法是将风格分类器用于指导语言模型,以生成目标风格的句子。

但值得注意的是,去除文本毒性存在一定难度。一方面,文本的毒性表现通常有两类[1]。一类是显式毒性,该类文本因毒性词外显等特点而使人更易理解;另一类是隐式毒性,这类文本不同于一般的负面情感文本,它的毒性较为隐晦,往往难以识别,具体示例如表1所示。另一方面,对毒性文本缺乏更细粒度的探究,如地域歧视、性别歧视等文本,是否因句子中包含某些片段而导致歧视发生。

表1 两种毒性类型的文本示例

Table 1 Text examples of two toxicity types

毒性类型	示例
显式毒性文本	没有人会想成为这款商品首次上线就购买的笨蛋!
隐式毒性文本	小仙女的事你少管!

本文主要研究了中文领域的文本去毒任务。为此,(1)我们重构了一个中文毒性语料集,以此作为研究的数据基础;(2)对毒性的表现形式进行了探究;(3)使用文本风格迁移模型和大语言模型进行文本去毒。

首先,我们采用跨多个领域收集数据的方式进行语料的收集,这种构建数据集的策略可以有效地降低人力和时间成本。具体而言,本文通过翻译、爬取和从现有中文冒犯性语料中筛选等三种方式重构语料。语料集包含约100 k条毒性句子和规模相当的无毒句子。

其次,依据自动评估和人工评估的方法判断文本的毒性表现形式。进一步,通过基本原理提取机制(Rationale Extraction)[9]结合毒性二元分类器的方式检测毒性文本中的毒性词,以此分析毒性文本的成因。

最后,我们根据分析结果使用编辑式和生成式文本风格迁移模型去除文本毒性,此外,在模型去除文本毒性这一方面的表现,将特定的风格迁移模型与基于输入提示词的大语言模型进行了对比。

1 相关工作

1.1 文本风格迁移

文本风格迁移任务的目标是重写具有相同内容文本的同时,改变构成文本的风格[10]。由于该任务普遍缺乏平行语料,大多数TST任务都采用无监督学习的方法,该方法主要分为两类。

一种是以编辑式框架为主的显式TST方法,思想是预先假定文本的风格特征只存在于特定的单词中,让模型学习并定位这些“风格词”,随后用目标风格词进行替换。如Li等[11]提出的删除-检索-生成模型(Delete-Retrieve-Generate,DRG),该模型基于词频统计的方式生成风格词列表,随后根据风格词表删除源风格词并检索目标风格词,最后生成目标风格的文本。

除了基于词频统计的方法识别风格词外,Xu等[12]提出,基于注意力机制的分类模型可以显式地分离文本的内容和风格词,即在训练过程中分类模型通常会赋予风格词更高的注意力权重。随后Zhang等[13]针对情感迁移任务提出了基于自注意力机制的情感风格词删除方法。此外,因基于Transformer的双向编码表征(Bidirectional Encoder Representations from Transformers,BERT)拥有强大的特征提取能力,Sudhakar等[14]将该模型应用到DRG框架,提出了盲式生成风格转换器(Blind Generative Style Transformer,B-GST)和引导式生成风格转换器

(Guided Generative Style Transformer, G-GST)两个模型实现情感风格迁移。随后Malmi等[15]提出了MASKER模型,该模型在源风格域和目标风格域训练两个掩码语言模型,找到两个模型在似然性方面不一致的词片段,从而识别原句中待删除的词。

另一种则是以生成式框架为主的隐式TST方法。如,让模型自动学习句子内容和风格的潜在表示并进行风格的分离与转换;或通过特定风格分类器指导语言模型生成目标风格文本。相关工作如Mueller等[16]采用基于潜在表示编辑的方法,以生成包含目标风格的文本;Wang等[17]采用多任务学习的方法实现文本的风格迁移;Liu等[18]提出渐进式地优化文本在隐空间中的表示以实现文本的风格迁移;Dale等[5]提出基于生成式的模型去除文本毒性。

总的来说,当文本中出现风格词时,检索、删除属性词并用目标风格词替换的显式方法可以较好地实现文本的风格迁移。倘若难以分离文本的风格与内容,通过风格分类器指导模型生成或多任务学习的隐式文本风格迁移方法则有不错的表现。

1.2 文本去毒任务

目前,文本去毒任务的研究主要集中于英文领域,且仅有少部分工作采用有监督的方法。如Logacheva等[4]通过管道的方式构建了一个毒性-非毒性的英文平行语料,然后训练seq2seq模型[19]进行文本去毒。

因缺乏平行语料,无监督学习同样是完成文本去毒任务的主流方法。如Santos等[6]通过训练一个具有额外风格分类和周期一致性损失的自编码器用于文本去毒;随后Tran等[7]使用搜索引擎找出与给定有毒句子语义相似的无毒句子,并通过掩码语言建模的方式填充句子中的空白部分,最后由seq2seq模型编辑生成更加流畅的无毒句子;Laugier等[8]则通过微调T5[20]模型实现文本的去毒。Dale等[5]提出了两种有较强解毒能力的基于Trransformer的条件双向编码表示模型(Condtional Bidirectional Encoder Representations from Transformers, CondBERT)和ParaGeDi,CondBERT模型使用掩码语言建模的方式识别文本中的毒性词并用语义相似的中性词进行替换,ParaGeDi模型则通过引入一个预训练好的风格判别器用于指导语言模型生成中性的文本,上述两种方法均有较强的文本去毒能力。

Floto等[21]随后提出了一种混合扩散模型进行文本去毒,作者指出,混合扩散模型相较于条件文本生成模型具有更高的输出多样性,同时能有效解决因数据集匮乏导致生成文本流畅度低的问题。此外,通过改写文本以消除冒犯性的含义同样可以减轻句子的毒性,但改写毒性程度较为微妙的句子存在难度。为此,Hallinan等[22]提出了一种结合可控文本生成和文本重写的去毒算法,通过有毒、无毒语言模型生成的单词概率分布找到要掩码的毒性词,并用中性词进行替换。

2 基线模型

2.1 基于编辑式框架的文本风格迁移模型

2.1.1 T&G模型

该模型由Madaan等[23]提出,包含Tagger和Generator两个模块。Tagger模块负责使用[TAG]标记替换原句中的风格属性词,若该模块未识别出句子中的属性词,则在句子中合适的位置添加[TAG]标记;Generator模块则负责将带有[TAG]标记的句子生成目标风格的文本。该模型的训练可分为两阶段。具体而言,第一阶段训练Tagger模块,使其生成与原句子$x_i^{(1)}$风格无关的内容$z(x_i)$;随后对Generator模块进行训练,使其能够将$z(x_i)$作为输入并生成带有目标风格Γ_2的句子$\hat{x}_i^{(2)}$。该模型的流程图如图1所示。

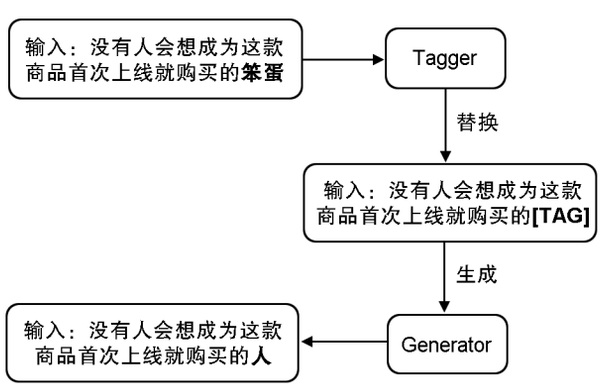

图1 T&G模型流程图

Fig. 1 Flow chart of T&G model

2.1.2 Bias Mitigation模型

该模型由Tokpo等[24]提出用于去除文本数据中的性别偏见。模型使用显示关键词替换与潜在内容相结合的方式改写文本。由于作者尚未给模型命名,因此本文将该模型命名为Bias Mitigation Model。

Bias Mitigation模型由四个模块构成,属性掩码器(Attribute Masker)、词编码器(Token Embedder)、潜在内容编码器(Latent-content Encoder)和解码器(Embedding Decoder)。属性掩码器模块用于掩盖文本中的偏见词;潜在内容编码器负责生成不包含性别偏见且与原句语义相似的内容表示,该模块的训练采用双目标训练的方式。词编码器负责编码带有掩码标记的文本,解码器则负责解码带有潜在内容表示的词嵌入,以生成中性文本。具体而言,潜在内容编码器与源内容编码器均将源文本x_b作为输入,输出$\hat{z}$和z。$\hat{z}$和z内容的相似性,通过余弦相似度指标进行度量。内容相似度训练损失函数如下公式(1)所示,定义的第二个损失函数,如公式(2)所示,使$\hat{z}$的内容与偏见无关:

$$L_{\text{sim}}=\frac{1}{N}\sum_{j=1}^{N}(\text{con\_sim}(\hat{z},z)-1)^2, \tag{1}$$

$$L_{\text{acc}_{\hat{z}_j}}=-\sum_{j=1}^{N}\log P(s_a|\hat{z}_j), \tag{2}$$

其中,$P(s_a|\hat{z}_j)$表示分类器预测$\hat{z}_j$为中性的概率。该模块最终的损失函数为如公式(3)所示,并通过引入平衡参数λ来控制每个目标的影响程度。

$$L_{\text{CE\_loss}}=(1-\lambda)L_{\text{sim}}+\lambda L_{\text{acc}_{\hat{z}_j}}。 \tag{3}$$

2.1.3 CondBERT模型

BERT模型[25]通过填充句子中的掩码标记作为模型的一项预训练任务。Wu等[26]则提出,在掩码的位置插入中性的词语,可以实现文本的去毒。Dale等对上述模型进行了改进,提出了CondBERT模型,该模型的流程图如图2所示。具体而言,通过训练一个单词级的毒性分类器,并在训练过程中将单词作为分类特征,把模型分配给单词的权重值作为对应的毒性分数。当毒性分数高于阈值$t=\max(t_{\min},\max(s_1,s_2,\cdots,s_n)/2)$时,用[MASK]替换毒性词,其中$s_1,s_2,\cdots,s_n$表示句子中所有单词的得分,$t_{\min}=0.2$为最小毒性得分。最后,BERT模型将[MASK]标记替换为中性的词。此外,该模型采用了内容保存的启发式方法保留被替换词的语义,以选择合适的候选词。

2.2 基于生成式框架的文本风格迁移模型

ParaGeDi模型

该模型由Dale等基于GeDi模型所提出。常规的GeDi模型[27]由一个生成式预训练转换器模型(Generative Pre-trained Transformer,GPT2)和一个生成式判别模型组成,判别模型是在包含句子级别风格标签的数据集上训练得到的类别条件语言模型(Class-conditional LMs)[27]。因此,GeDi模型可生成指定风格或主题的文本。ParaGeDi模型在GeDi模型的基础上进行改进,即为了保留原始句子的内容,用释义模型替换语言模型,使得ParaGeDi模型可对特定风格的文本进行改写,并生成目标风格的文本。该模型的流程图如图3所示。

给定毒性文本$x_i\in X_i$作为模型的输入,y表

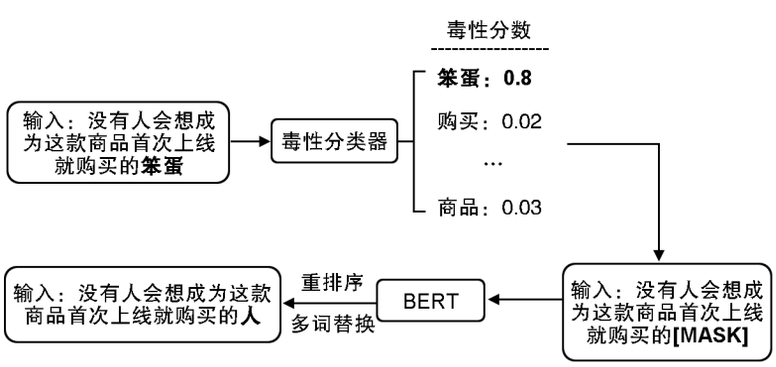

图2 CondBERT模型流程图

Fig. 2 Flow chart of CondBERT model

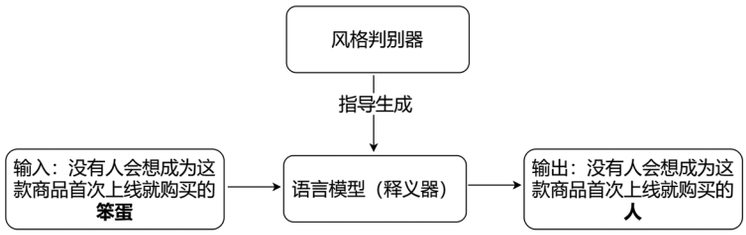

图3 ParaGeDi模型流程图

Fig. 3 Flow chart of ParaGeDi model

示长度为T的输出文本，c表示目标风格属性(毒性或者情感)，ParaGeDi模型的求解概率将遵循下述公式：

$$P(y_t|y_{<t},x,c)\propto P_{\mathrm{LM}}(y_t|y_{<t},x)P(c|y_t,y_{<t},x)\approx P_{\mathrm{LM}}(y_t|y_{<t},x)P_D(c|y_t,y_{<t}), \tag{4}$$

根据原始文本x、目标属性c和t时刻前生成的文本预测t时刻的输出。t时刻的输出概率包含两项，第一项由主语言模型P_{LM}产生，第二项则使用贝叶斯规则与类条件语言模型P_{CC}求解得到，如公式(5)所示：

$$P_{\mathrm{D}}(c|x_t,x_{<t})\propto P(c)P_{\mathrm{CC}}(x,x_{<t}|c)。 \tag{5}$$

3 中文毒性数据集

3.1 数据来源

据调研，因存在的中文数据集相对匮乏，当前文本去毒研究主要集中于英文领域。基于此，本文通过组合和人工标注的方式，重构了一个中文的毒性语料集。

语料集中的训练集源于三部分：(1) Jigsaw 2018 dataset[28]：我们使用翻译应用程序接口(Application Programming Interface，API)将英文毒性文本翻译成中文，随后使用自动筛选的方式选取出翻译质量较好、表达清晰流畅的文本，通过该方式得到的毒性文本平均长度为24.41；(2)中文冒犯性语言检测数据集(Chinese Offensive Language Detection Dataset，COLDataset)：该数据集由Deng等[2]构建，涵盖了种族、性别和地域三类主题的冒犯性和非冒犯性文本，我们选取了性别和地域主题中的部分冒犯性文本，得到的文本平均长度为48.77；(3) Weibo数据：根据毒性文本的定义，本文使用爬虫技术并结合关键字查询、从相关主题抓取两种策略收集毒性文本，最后从微博平台爬取了一万余条毒性文本，文本的平均长度为17.52。毒性文本的训练集和验证集统计结果如表2所示。并通过人工改写的方式得到600条毒性-中性的平行测试数据。

表2 数据集统计结果(毒性文本)

Table 2 Statistical results of data set (Toxic text)

参数	Jigsaw 2018	COLDataset	Weibo
训练/验证集	80 k	10 k	10 k
句子平均长度	24.41	47.89	17.52

3.2 数据分析

分析清楚文本的毒性表现形式有助于选择更有效的去毒模型。主要内容包括：(1)探究毒性的表现形式。即文本属于显式毒性或隐式毒性，该步采用自动评估与人工评估的方式；(2)文本呈现毒性的原因，这一步将采用先检测文本毒性词再分析的方式。

3.2.1 毒性表现形式分析

通常而言，根据句子中是否存在明显毒性词可将毒性文本分为显式毒性文本和隐式毒性文本。为此本文采用自动和人工评估的方式对毒性文本划分。针对自动评估，我们构建了一份毒性词汇表，随后通过检索的方式对毒性文本逐一检索，倘若一段文本中包含了毒性词汇表中的词，则认为该文本为显示毒性文本，否则为隐式毒性文本。我们从三个数据集中分别随机抽样一定量的文本进行分析，自动分析方法的分析结果如表3所示。

人工注释的方式则是将上述所选数据集分发给工作人员进行注释，并统计出注释为显式

表3 自动注释统计结果(句)

Table 3 Automatically annotate statistical results (sentences)

数据集	Label 0	Label 1	总计
Jigsaw 2018	2 615	2 385	5 k
COLDataset	1 669	831	2.5 k
Weibo	1 367	1 133	2.5 k

毒性文本(label:0)和隐式毒性文本(label:1)的比例,人工注释的结果如表4所示。

表4 人工注释统计结果(句)

Table 4 Manually annotated statistical results (sentences)

数据集	Label 0	Label 1	总计
Jigsaw 2018	3 036	1 964	5 k
COLDataset	1 746	754	2.5 k
Weibo	1 254	1 246	2.5 k

根据表3、表4和图4可知,语料集中显式毒性文本的占比高于隐式毒性文本,且同类型的毒性文本,如显式毒性文本,人工注释的数量同样高于自动注释。

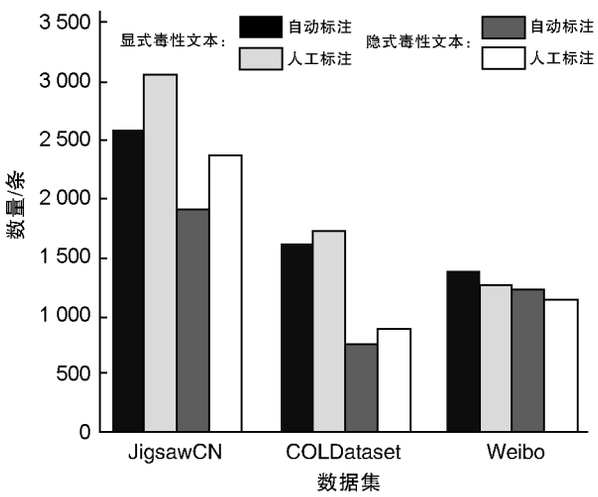

图4 数据集统计结果

Fig. 4 Statistical results of dataset

3.2.2 毒性词检测及毒性成因分析

Pavlopoulos等[9]认为,通过基于注意力的原理提取机制(Attention-based Rationale Extraction,ARE)结合二元毒性分类器的方式可以有效检测文本中的毒性片段。基于此理论,本文使用基于注意力机制的二元毒性分类器预测句子是否有毒,并在模型推理时调用其注意力分数来获得句子中的毒性词。

本文使用BERT模型[25]和强力优化的BERT预训练方法(Robustly Optimized BERT Pretraining Approach,RoBERTa)模型[29]作为二元毒性分类器,并根据两个模型最后一层头部的注意力平均分作为检测依据。本文基于Pavlopoulos等[9]提出的毒性片段检测评估方法,使用$F1$值作为模型检测毒性词的评估指标,该指标的求解公式如式(6)所示:

$$F_1^t(A_i,G)=\frac{2*P^t(A_i,G)*R^t(A_i,G)}{P^t(A_i,G)+R^t(A_i,G)},\quad(6)$$

其中,t表示测试集中的句子,A_i表示模型预测的毒性词,S_A^t表示所有预测结果的集合,G表示毒性词的真值,$P^t(A_i,G)$和$R^t(A_i,G)$的求解公式如下所示:

$$P^t(A_i,G)=\frac{|S_A^t\cap S_G^t|}{|S_A^t|},\quad(7)$$

$$R^t(A_i,G)=\frac{|S_A^t\cap S_G^t|}{|S_G^t|}。\quad(8)$$

根据表5的结果可知,基本原理提取机制结合二元毒性分类器的方式能有效检测出句子中的毒性词。为此,对涉及地域冒犯及性别冒犯的毒性文本进行成因分析。具体而言,通过调用分类器在推理时的注意力分数得到导致文本呈现毒性的词,注意力热图如图5—图6所示。不论是地域冒犯或性别冒犯,文本中带有辱骂、仇恨色彩的词是导致文本呈现毒性的主要原因,若仇恨或攻击的对象从个人延伸至地域、性别等目标群体,便会发生地域和性别冒犯。

表5 毒性词检测评估结果

Table 5 Detection and evaluation results of toxic words

模型	F_1/%	P/%	R/%
BERT+ARE	**72.8**	72.7	73.1
RoBERTa+ARE	**73.0**	72.9	73.1

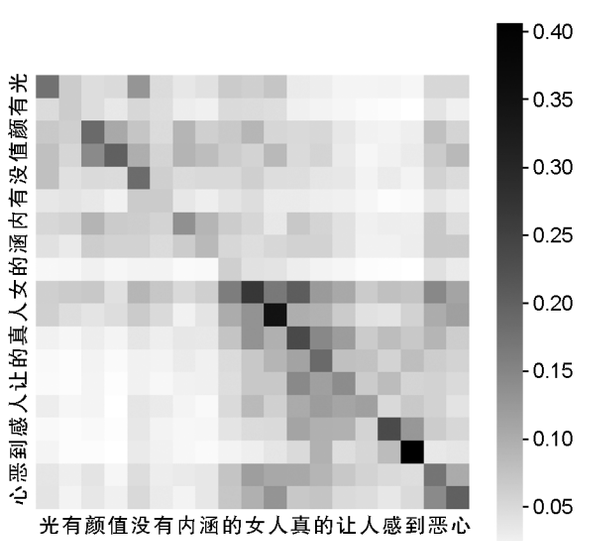

图5 分类模型的注意力分数热图例1

Fig. 5 Sample 1 of attention fraction heat map of classification model

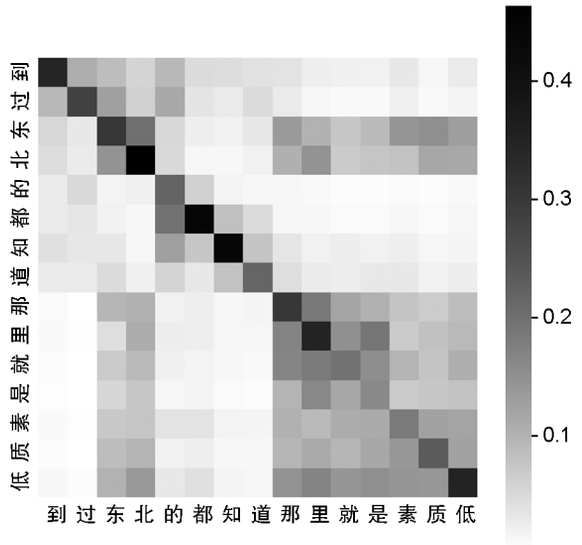

图6 分类模型的注意力分数热图例2

Fig. 6 Sample 2 of attention fraction heat map of classification model

4 实验与结果分析

4.1 评估指标

风格转换准确率ACC:该指标用于评估模型成功将有毒文本转换为无毒文本的百分比。为此,我们使用预训练好的毒性二元分类器作为验证模型,利用其预测结果统计最终的ACC得分。

内容保存度SIM:该指标用于衡量生成文本与原始文本内容上的一致性程度。我们通过计算生成文本与原始文本之间的相似度来衡量句子的内容保存,即计算原始文本嵌入与生成文本嵌入之间的余弦相似度。SIM得分越高,说明生成文本中保留了越多的内容。

流畅度FL:该指标用于衡量生成文本的流畅性。我们使用预训练好的语言模型,通过计算生成文本的困惑度值(Perplexity,PPL)来衡量句子的流畅度,该指标越小说明句子流畅程度越高。

4.2 模型设置

T&G模型:我们使用4层的Transformer训练Tagger和Generator,每个Transformer有4个注意力头,嵌入层和隐藏层维数均为512。推理过程中,使用波束搜索(beam search),其中beam size =5,来解码生Tagger和Generator生成的结果。

Bias Mitigation模型:最大序列长度(sequence size)设置为60;潜在内容编码器的平衡参数λ设置为0.5;去除属性词的阈值μ设置为0.1,其他超参数的设置将遵循BERT-base-Chinese[①]模型的默认值。

CondBERT模型:该模型且未进行任何微调,且风格的控制通过毒性分类器来完成,因此我们将BERT替换成适用于中文任务的BERT-base-Chinese模型。模型超参数我们遵循Dale等人的设置。

ParaGeDi模型:我们使用预训练中文释义生成模型Chinese-Bart-Paraphrase[②]作为主语言模型,并在重构的中文毒性语料集上微调GPT2-base-Chinese[③]得到风格判别模型。在生成候选词汇时采用波束搜索的方式,并将其值设置为beam size=10。

4.3 实验结果分析

4.3.1 结果分析

由表6可知,现有的文本风格迁移模型针对中文文本去毒任务是有效的。根据表7—表8的实验结果可知,编辑式模型能有效去除显示毒性文本的毒性,同时具有更高的内容保存度,这是由于该类模型能准确定位句子中的毒性词并删除,同时保留句子中的非毒性词。对于隐式毒性的文本,基于生成式的模型的去毒能力则更加优异,且生成的句子具有更高的流畅度。主要原因在于该模型的架构基于生成式的GPT语言模型,故而生成的句子不仅流畅度较高,且在外部风格判别器的作用下,模型尽可能避免产生毒性词。由于隐式毒性文本中无明显毒性关键词,所以编辑式模型无法识别以及替换毒性词,因此对这一类的毒性文本的去毒能力稍有欠缺。

此外,本文使用大语言模型并结合Prompt的方法对毒性文本进行了改写。针对ChatGLM系列模型,我们采用Suzgun等[30]提出的四种Prompt模板,如表9所示,以此对比在不同

①https://huggingface.co/bert-base-chinese

②https://huggingface.co/figurative-nlp/chinese-bart-paraphrase

③https://huggingface.co/ckiplab/gpt2-base-chinese

表6 基线模型在测试集上的评估结果

Table 6 Evaluation results of baseline model on test set

模型	ACC/%
Original*	0.46
Tag&Generator	47.96
Bias Mitigation	66.40
CondBERT	70.47
ParaGeDi	74.06

表7 自动评估结果(显式毒性文本)

Table 7 Automatic evaluation results (explicit toxicity text)

模型	SIM/%	FL	ACC/%
Tag&Generator	54.12	113.87	45.86
BiasMitigation	88.52	45.01	47.13
CondBERT	94.92	89.20	58.53
ParaGeDi	54.40	39.97	**64.60**

表8 自动评估结果(隐式毒性文本)

Table 8 Automatic evaluation results (implicit toxicity text)

模型	SIM/%	FL	ACC/%
Tag&Generator	52.03	130.08	48.04
BiasMitigation	86.29	101.46	59.77
CondBERT	91.86	100.03	61.32
ParaGeDi	52.55	**47.88**	**74.13**

单样本输入提示下的输出文本间的差异,结果见表10所示。同时使用少量的数据对CPM-bee-1B模型进行微调,使其适应文本解毒任务,并对比微调前后两个版本的ChatGLM-6B模型的去毒表现,实验结果如表11所示。

从结果可以发现,使用不同的Prompt作为模型的输入,对最终的输出并未有太大的影响,并且在给定一个改写样例时,模型改写毒性文本的能力并未有太大提升。此外,微调后的大语言模型在去毒任务上的表现有一定提升。这是由于微调前的大语言模型侧重生成更加通用的文本,而微调后的模型更加适用去毒任务。大语言模型相较基线模型而言,生成的文本具有较高的流畅度,但句子毒性没有明显降低,基线模型在内容保存度上与大语言模型相当,且改写后的文本毒性程度更低。

表9 大模型输入提示模板

Table 9 Input prompt template of large model

模板名称	模板内容
模板一 Vanilla:	这是一段文本:[d_1][$x^{(s_1)}$][d_2],这是重写后的文本,它的风格为[s_2]: [d_1] Here is a text: [d_1][$x^{(s_1)}$][d_2] Here is a rewrite of the text, which is [s_2]: [d_1]
模板二 Contrastive:	这是一段风格为[s_1]文本: [d_1][$x^{(s_1)}$][d_2],这是改写后风格为[s_2]的文本: [d_1] Here is a text which is [s_1]: [d_1][$x^{(s_1)}$][d_2] Here is a rewrite of the text, which is [s_2]: [d_1]
模板三 Negation-v1:	这是一段风格为[s_1]的文本: [d_1][$x^{(s_1)}$][d_2],这是改写后的文本,它不属于风格[s_1]: [d_1] Here is a text which is [s_1]: [d_1][$x^{(s_1)}$][d_2] Here is a rewrite of the text, which is not [s_1]: [d_1]
模板四 Negation-v2:	这是一段风格不是为[s_2]的文本: [d_1][$x^{(s_1)}$][d_2],这是改写后的文本,它属于风格[s_2]: [d_1] Here is a text which is not [s_2]: [d_1][$x^{(s_1)}$][d_2] Here is a rewrite of the text, which is [s_2]: [d_1]

4.3.2 错误示例分析

本文对基线模型未成功去除文本毒性的句子示例进行了定性分析。错误类别可分为两类:(1)句子中存在多个毒性词的情况下,模型

表10 四种Prompt模板在ChatGLM模型上的生成结果

Table 10 Generation results of four Prompt templates on ChatGLM model

Prompt	输入:	模型预测结果:	真实结果
模板一 Vanilla	这是一句文本:{你可以从什么都不懂的笨蛋变为天才},这是改写后的中性文本:{你可以从什么都不懂的人变为天才},这是一句文本:{没有人会想成为商品首次上线就购买的白痴},这是改写后的中性文本:{	虽然有些人在首次购买商品时会犹豫,但并不是所有人都会这样做	没有人会想成为商品首次上线就购买的人}
模板二 Contrastive	这是一句有毒文本:{你可以从什么都不懂的笨蛋变为天才},这是改写后的中性文本:{你可以从什么都不懂的人变为天才},这是一句有毒文本:{没有人会想成为商品首次上线就购买的白痴},这是改写后的中性文本:{	虽然有些人在首次购买商品时会犹豫,但并不是所有人都会这样做	没有人会想成为商品首次上线就购买的人}
模板三 Negation-v1	这是一句文本:{你可以从什么都不懂的笨蛋变为天才},这是改写后的非毒性文本:{你可以从什么都不懂的人变为天才},这是一句文本:{没有人会想成为商品首次上线就购买的白痴},这是改写后的非毒性文本:{	虽然有些人在首次购买商品时会犹豫,但并不是所有人都会这样做	没有人会想成为商品首次上线就购买的人}
模板四 Negation-v2	这是一句非中性文本:{你可以从什么都不懂的笨蛋变为天才},这是改写后的中性文本:{你可以从什么都不懂的人变为天才},这是一句文本:{没有人会想成为商品首次上线就购买的白痴},这是改写后的中性文本:{	虽然有些人在首次购买商品时会犹豫,但并不是所有人都会这样做	没有人会想成为商品首次上线就购买的人}

未能完全识别并进行改写;(2)原句子中的毒性词与候选词之间的词义差异太大,导致模型生成的句子语义被改变。具体示例如表12所示。

表11 大语言模型实验结果

Table 11 Experimental results of large language model

模型	SIM/%	FL	ACC/%
ChatGLM-6B	39.57	39.21	50.47
ChatGLM2-6B	41.29	38.79	50.48
ChatGLM-6B*(微调)	71.61	**38.54**	44.76
ChatGLM2-6B*(微调)	57.44	**30.06**	57.42
CPM-bee-1B*(微调)	83.55	**43.73**	40.77

表12 模型输出错误示例

Table 12 Model output error example

	句子示例及错误类型
源句子1	你的回答散发着一股众人皆醉我独醒的优越感,实在是恶心透了
中性句子1	你的回答散发着一股子众人皆醉我独醒的味,实在是棒透了
错误类型1	**语义改变**
源句子2	看照片中,一群可怜的失败者
中性句子2	看着照片,很可怜的猫
错误类型2	**语义改变**
源句子3	你总是这么无知,还是因为愚蠢?
中性句子3	你总是这么无知,还是因为别的?
错误类型3	**未完全识别毒性词**

导致第一种错误产生的原因是模型在学习过程中倾向于关注毒性程度高的词汇。我们发现无论是通过毒性分类模型的注意力权重,或是训练单词级别的词袋分类模型等方法来识别毒性词,模型都容易忽略毒性程度较低的词。

而导致句子语义变化的主要原因是,模型在生成候选词时更倾向于生成偏中性风格的词,未充分考虑候选词与原词间的语义关系。此外,当句子较短且被掩码的毒性词较多时,模型无法充分学习句子中的上下文信息,这种情况下即便替换词是中性,句子的语义也可能改变。

5 总结与展望

本文重构了一个针对中文领域的文本去毒任务数据集,并基于该数据集进行了文本去毒研究。我们发现,不同类型的文本风格迁移模型去毒能力存在差异,即基于编辑式的模型天然地更适用于句子中包含毒性词的情况,而生成式模型在句子毒性较为隐晦时更加适用。实验结果也表明,对于显式毒性的文本,编辑式模型能有效去除其毒性,同时最大程度保留原文本中的内容信息;生成式模型生成的文本则具有更高的流畅性。但尚未有模型能够同时有效去除上述两类毒性。此外,采用Prompt的大语言模型具有一定的去毒能力,且在少量特定模板数据上微调后的大语言模型去毒能力得到了提升,但相较于风格迁移模型,小参数大语言模型的去毒性能还有待提高。

中文的隐式毒性文本不仅形式多样,且表达更为隐蔽,这也给中文的文本去毒任务带来了新的挑战。比如“小仙女的事你少管!”中的“小仙女”一词本身是褒义词,但在该句中则是在含蓄地攻击女性。未来,对该类隐式毒性文本的去毒,需要考虑引入外部情感词汇知识和上下文的情感信息。此外,在本文工作的基础上,我们将进一步在Lu等[1]构建的细粒度毒性数据集上展开研究。

参考文献:

[1] LU J Y, XU B, ZHANG X K, *et al*. Facilitating Fine-grained Detection of Chinese Toxic Language: Hierarchical Taxonomy, Resources, and Benchmarks[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2023: 16235-16250. DOI: 10.18653/v1/2023.acl-long.898.

[2] DENG J W, ZHOU J Y, SUN H, *et al*. COLD: a Benchmark for Chinese Offensive Language Detection[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2022: 11580-11599. DOI: 10.18653/v1/2022.emnlp-main.796.

[3] ZAMPIERI M, MALMASI S, NAKOV P, *et al*. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (Offensval)[C]//Proceedings of the International Workshop on Semantic Evaluation (SemEval). USA: Association for Computational Linguistics, 2019: 75-86.DOI:10.18653/v1/S19-2010.

[4] LOGACHEVA V, DEMENTIEVA D, USTYANTSEV S, *et al*. ParaDetox: Detoxification with Parallel Data[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Pa-

pers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2022: 6804−6818. DOI: 10.18653/v1/2022.acl-long.469.

[5] DALE D, VORONOV A, DEMENTIEVA D, *et al*. Text Detoxification Using Large Pre-trained Neural Models[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 7979−7996. DOI: 10.18653/v1/2021.emnlp-main.629.

[6] NOGUEIRA DOS SANTOS C, MELNYK I, PADHI I. Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 189−194. DOI: 10.18653/v1/p18-2031.

[7] TRAN M, ZHANG Y P, SOLEYMANI M. Towards a Friendly Online Community: an Unsupervised Style Transfer Framework for Profanity Redaction[C]//Proceedings of the International Conference on Computational Linguistics(COLING). Spain (Online): International Committee on Computational Linguistics, 2020: 2107−2114. DOI: 10.18653/v1/2020.coling-main.190.

[8] LAUGIER L, PAVLOPOULOS J, SORENSEN J, *et al*. Civil Rephrases of Toxic Texts with Self-supervised Transformers[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 1442−1461. DOI: 10.18653/v1/2021.eacl-main.124.

[9] PAVLOPOULOS J, LAUGIER L, XENOS A, *et al*. From the Detection of Toxic Spans in Online Discussions to the Analysis of Toxic-to-civil Transfer[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2022: 3721−3734. DOI: 10.18653/v1/2022.acl-long.259.

[10] TOSHEVSKA M, GIEVSKA S. A Review of Text Style Transfer Using Deep Learning[J]. *IEEE Trans Artif Intell*, 2022, **3**(5): 669−684. DOI: 10.1109/TAI.2021.3115992.

[11] LI J C, JIA R, HE H, *et al*. Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 1865−1874. DOI: 10.18653/v1/n18-1169.

[12] XU J J, SUN X, ZENG Q, *et al*. Unpaired Sentiment-to-sentiment Translation: a Cycled Reinforcement Learning Approach[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 979−988. DOI: 10.18653/v1/p18-1090.

[13] ZHANG Y, XU J J, YANG P C, *et al*. Learning Sentiment Memories for Sentiment Modification without Parallel Data[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 1103−1108. DOI: 10.18653/v1/d18-1138.

[14] SUDHAKAR A, UPADHYAY B, MAHESWARAN A. "Transforming" Delete, Retrieve, Generate Approach for Controlled Text Style Transfer[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP−IJCNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 3269−3279. DOI: 10.18653/v1/d19-1322.

[15] MALMI E, SEVERYN A, ROTHE S. Unsupervised Text Style Transfer with Padded Masked Language Models[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 8671−8680. DOI: 10.18653/v1/2020.emnlp-main.699.

[16] MUELLER J, GIFFORD D, JAAKKOLA T. Sequence to Better Sequence: Continuous Revision of Combinatorial Structures[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. New York: ACM, 2017: 2536−2544. DOI: 10.5555/3305890.3305943.

[17] WANG K, HUA H, WAN X J. Controllable Unsupervised Text Attribute Transfer via Editing Entangled Latent Representation[EB/OL]. arXiv Preprint: 1905.12926. 2019. https://doi.org/10.48550/arXiv.1905.12926.

[18] LIU D, FU J, ZHANG Y D, *et al*. Revision in Continuous Space: Unsupervised Text Style Transfer without Adversarial Learning[J]. *Proc AAAI Conf Artif Intell*, 2020, **34**(5): 8376−8383. DOI: 10.1609/aaai.v34i05.6355.

[19] SUTSKEVER I, VINYALS O, LE Q V. Sequence to Sequence Learning with Neural Networks[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2. New York: ACM,

2014: 3104−3112. DOI: 10.5555/2969033.2969173.

[20] RAFFEL C, SHAZEER N, ROBERTS A, *et al*. Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer[EB/OL]. arXiv Preprint: 1910.10683. 2019. https://doi.org/10.48550/arXiv.1910.10683.

[21] FLOTO G, ABDOLLAH POUR M M, FARINNEYA P, *et al*. DiffuDetox: a Mixed Diffusion Model for Text Detoxification[C]//Findings of the Association for Computational Linguistics: ACL 2023. Stroudsburg, PA, USA: Association for Computational Linguistics, 2023: 7566−7574. DOI: 10.18653/v1/2023.findings-acl.478.

[22] HALLINAN S, LIU A, CHOI Y, *et al*. Detoxifying Text with MaRCo: Controllable Revision with Experts and Anti-experts[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2023: 228−242. DOI: 10.18653/v1/2023.acl-short.21.

[23] MADAAN A, SETLUR A, PAREKH T, *et al*. Politeness Transfer: a Tag and Generate Approach[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 1869−1881. DOI: 10.18653/v1/2020.acl-main.169.

[24] TOKPO E K, CALDERS T. Text Style Transfer for Bias Mitigation Using Masked Language Modeling[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop. Stroudsburg, PA, USA: Association for Computational Linguistics, 2022: 163−171. DOI: 10.18653/v1/2022.naacl-srw.21.

[25] DEVLIN J, CHANG M W, LEE K, *et al*. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//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). Minnesota: Association for Computational Linguistics, 2019: 4171−4186. DOI: 10.18653/v1/N19-1423.

[26] WU X, ZHANG T, ZANG L J, *et al*. "Mask and Infill": Applying Masked Language Model to Sentiment Transfer[EB/OL]. arXiv Preprint: 1908.08039, 2019. https://doi.org/10.48550/arXiv.1908.08039.

[27] KRAUSE B, GOTMARE A D, MCCANN B, *et al*. GeDi: Generative Discriminator Guided Sequence Generation[EB/OL]. arXiv Preprint: 2009.06367, 2020. https://doi.org/10.48550/arXiv.2009.06367.

[28] CJADAMS, JEFFREY S, JULIA E, *et al*. Toxic Comment Classification Challenge. Kaggle[EB/OL]. 2018, https://kaggle. com/competitions/jigsaw−toxic−comment−classification−challenge.

[29] LIU Z, LIN W, SHI Y, *et al*. A Robustly Optimized BERT Pre-training Approach with Post-training[C]//China National Conference on Chinese Computational Linguistics. Cham: Springer. China: Chinese Information Processing Society of China, 2021: 471−484.10.1007/978−3−030−84186−7_31

[30] SUZGUN M, MELAS-KYRIAZI L, JURAFSKY D. Prompt-and-Rerank: a Method for Zero-shot and Few-shot Arbitrary Textual Style Transfer with Small Language Models[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2022: 2195−2222. DOI: 10.18653/v1/2022.emnlp-main.141.