面向中文文本分类的字符级对抗样本生成方法

张顺香* 吴厚月 朱广丽 许 鑫 苏明星
(安徽理工大学计算机科学与工程学院 淮南 232001)
(合肥综合性国家科学中心人工智能研究院 合肥 230088)

摘 要：对抗样本生成是一种通过添加较小扰动信息，使得神经网络产生误判的技术，可用于检测文本分类模型的鲁棒性。目前，中文领域对抗样本生成方法主要有繁体字和同音字替换等，这些方法都存在对抗样本扰动幅度大，生成对抗样本质量不高的问题。针对这些问题，该文提出一种字符级对抗样本生成方法(PGAS)，通过对多音字进行替换可以在较小扰动下生成高质量的对抗样本。首先，构建多音字字典，对多音字进行标注；然后对输入文本进行多音字替换；最后在黑盒模式下进行对抗样本攻击实验。实验在多种情感分类数据集上，针对多种最新的分类模型验证了该方法的有效性。

关键词：对抗样本生成；文本分类；情感分类；多音字；字符级对抗样本
中图分类号：TN915.08; TP391.1 **文献标识码**：A **文章编号**：1009-5896(2023)06-2226-10
DOI: 10.11999/JEIT220563

Character-level Adversarial Samples Generation Approach for Chinese Text Classification

ZHANG Shunxiang WU Houyue ZHU Guangli Xu Xin SU Mingxing
(*School of Computer Science and Engineering, Anhui University of Science & Technology, Huainan* 232001, *China*)
(*Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei* 230088, *China*)

Abstract: Adversarial sample generation is a technique that makes the neural network produce misjudgments by adding small disturbance information. Which can be used to detect the robustness of text classification models. At present, the methods of sample generation in the Chinese domain include mainly traditional characters and homophones substitution, which have the problems of large disturbance amplitude of sample generation and low quality of sample generation. Polyphonic characters Generation Adversarial Sample (PGAS), a character-level countermeasure samples generation approach, is proposed in this paper. Which can generate high-quality adversarial samples with minor disturbance by replacing polyphonic characters. First, a polyphonic word dictionary to label polyphonic words is constructed. Then, the input text with polyphonic words is replaced. Finally, an adversarial sample attack experiment in the black-box model is conducted. Experiments on multiple sentiment classification datasets verify the effectiveness of the proposed method for a variety of the latest classification models.

Key words: Anti-sample generation; Text classification; Sentimental classification; Polyphonic characters; Character-level adversarial samples

1 引言

对抗样本起源于图像领域，通过对自动驾驶领域中的转弯图像进行修改，导致自动驾驶系统出现故障。在文本领域中，通过在文本中添加噪声的方式来生成对抗样本，会使分类器出现错误分类[1]，这启发了后续的文本对抗生成方法[2,3]和防御方法[4,5]。同时有学者[6]已经证实，当神经网络模型遭遇对抗样本攻击时，会出现准确率急剧降低的情况。在实际应用中，对抗样本常被用作检测模型鲁棒性的依据之一[7]。

目前，在中文文本对抗样本生成领域，生成对抗样本的方法主要有基于同音字替换[8]和繁体字替换[9]

收稿日期：2022-05-07；改回日期：2022-07-09；网络出版：2022-07-14
*通信作者： 张顺香 sxzhang@aust.edu.cn
基金项目：国家自然科学基金(62076006)，安徽高校协同创新项目(GXXT-2021-008)，安徽省研究生科研项目(YJS20210402)
Foundation Items: The National Natural Science Foundation of China (62076006), The University Synergy Innovation Program of Anhui Province (GXXT-2021-008), The Graduate Students Scientific Research Project of Anhui Province(YJS20210402)

等。但在进行繁体替换和同音字替换时，增加人工阅读障碍，扰动幅度大，容易被防御机制识别，生成的对抗样本质量不高。为最大程度保障语义，降低人工阅读障碍，本文提出一种字符级对抗样本生成方法(Polyphonic characters Generation Adversarial Sample, PGAS)，具体框架如图1所示。该方法采用改进的定向词删除评分机制进行关键词定位，找到影响分类的关键词；然后利用构建的多音字字典，用多音字替换的方法修改原始数据生成对抗样本，在多个最新的分类模型上进行试验。

2　相关工作

目前，文本领域对抗样本生成主要分为字符级、词级和句子级对抗样本生成方法。

2.1　字符级对抗样本生成方法

在字符级的对抗样本生成中，文献[10]提出一种字符级的对抗样本用作机器阅读理解模型的攻击验证。Niu等人[11]应用字符级对抗样本生成的方法，对生成样本采用最大边际法揭示了多种对话模型的弱点，提高了对抗模型的鲁棒性。Ebrahimi等人[12]通过研究字符级神经机器翻译的对抗样本，提出了以删除或改变翻译中的单词的两种攻击方法。GAO等人[13]提出通过修改核心词，使扰动编辑距离最小化的黑盒对抗样本生成方法(DeepWordBug)。文献[14]改进了Gao等人[13]的方法，提出快速生成对抗样本方法(FastWordBug)，对经常出错的单词进行更改，快速构造对抗样本。Ebrahimi等人[15]根据输入数据的重要性，提出强鲁棒性的字符级分类器。Song等人[16]提出一种基于梯度的搜索方法来输出欺骗目标分类器的自然文本。

2.2　词级对抗样本生成方法

在词级别的对抗样本生成中，Li等人[17]提出一种通过掩码填充，并利用上下文感知来修改语法输出的对抗样本生成模型。Tan等人[18]利用扰乱词形的变化，生成了看似合理和语义上相似的对抗样本。Li等人[19]提出一种利用预训练模型来生成对抗样本的高质量和有效的方法。Zang等人[20]将基于义元的词替换方法和基于粒子群优化的搜索算法结合，完善现有的词级攻击方法中的优化搜索算法。Cheng等人[21]考虑文本的离散性，提出了一种结合群套索和梯度正则化的投影梯度方法，来进行非重叠攻击和有针对性的关键字攻击。

2.3　句子级对抗样本生成方法

Jia等人[22]提出句末嵌入的句子级对抗样本生成方法，启发了后续句子级生成方法。Minervini等人[23]研究自然语言推理(Natural Language Inference, NLI)中违反给定1阶逻辑约束的对抗样本自动生成问题，最大限度地度量违反此类约束的程度。Wang等人[24]在Jia等人[22]的基础上，将生成的对抗样本嵌入到文本的不同位置，验证其模型的缺陷。Ribeiro等人[25]利用简单的扰动来检测单个句子中的语义改变问题。Iyyer等人[26]提出句法控制的释义网络，生成符合标准句法结构的对抗样本。Han等人[27]生成的对抗样本减弱了预测模型的结构化输出对输入中的小扰动过于敏感的现象。Wang等人[28]提出一种受控对抗文本生成模型，可以生成形式多样且流畅的对抗样本。

上述工作大都基于英文环境下进行对抗样本生成，在中文文本中效果不佳。而目前中文领域生成对抗样本的方法主要有同音字和繁体字替换等，扰动较大且生成质量不高，因此提出一种适用于中文领域生成高质量、小扰动的对抗样本的方法具有重要意义。

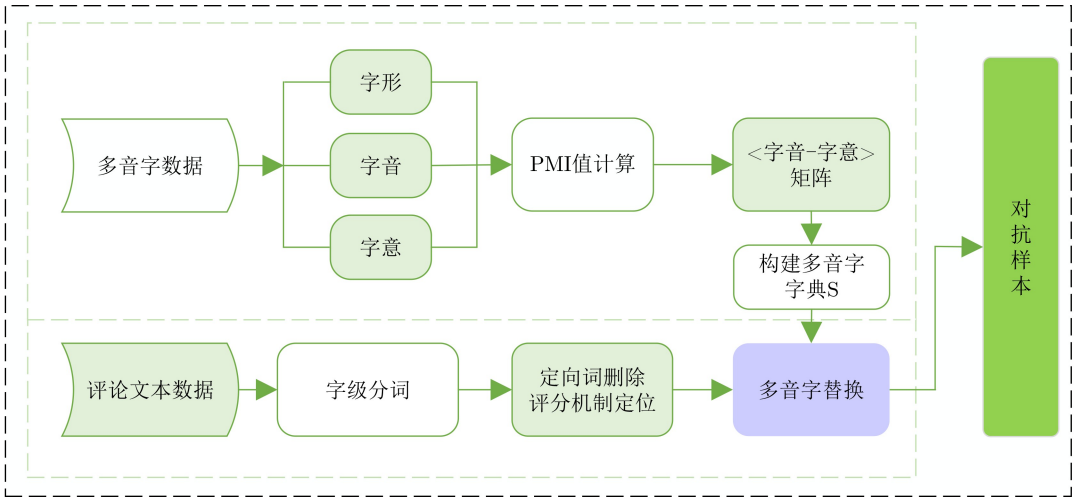


图 1 PGAS模型框架图

3 构建多音字字典

本文多音字字典的构建流程主要由两部分组成；是数据的获取与处理；进行多音字字典的构建。

3.1 数据获取与处理

传统中文中含有614个中文多音字[29]，其中共计1337个读音，其中3个读音以上的字共计91个，4个读音以上的字共17个，且不同发音代表的含义也不同。为确保数据的准确性，需对数据进行预处理，排除其中现代汉语不常用的多音字。其次，由于获取到的多音字为单个汉字，需要进行数据标注以区分具体读音对应的具体含义，本文考虑采用人工标注的方法。

3.2 构建多音字字典

针对多音字的结构特点，需要具体描述出不同读音下所表达的具体含义，因此进行下列定义。

定义 多音字字典。表示包含字符和读音之间关系的字典。它用于具体表述字符和读音之间的关系，可用四元组(w,x,y,i)描述。其中，w是多音字的中文表示，x是w的拼音表述，y是w的具体含义，i表示w的第i个读音，$i\in[1,7]$。

随着多音字读音的增多，其含义也逐渐变多。由定义知，在含多音字的句子中，需根据i值来确定w的具体含义y，而i值可通过x来确定。读音与字义之间的联系，采用点互信息PMI算法来完成，PMI可以较为准确地衡量读音与字义的相关性，其计算如式(1)所示

$$\begin{aligned}\mathrm{PMI}(x,y)&=\log_2\frac{p(x,y)}{p(x)p(y)}\\&=\log_2\frac{p(x\,|y)}{p(x)}\\&=\log_2\frac{p(y\,|x)}{p(y)}\end{aligned}\tag{1}$$

其中，若x与y无关，则$p(x,y)=p(x)p(y)$，表示该读音没有此含义；若x与y相关程度越高，则$p(x,y)$与$p(x)p(y)$比值越大。

设多音字w_i的含义集合为$W_i=\{w_1,w_2,\cdots,w_{l_i}\}$，则构建的<拼音-含义>集合$W$为

$$W=\begin{bmatrix}w_{11}&w_{12}&\cdots&w_{1n}\\w_{21}&w_{22}&\cdots&w_{2n}\\\vdots&\vdots&\ddots&\vdots\\w_{n1}&w_{n2}&\cdots&w_{nn}\end{bmatrix}\tag{2}$$

其中，w_{ij}表示单词W_i的第i个读音与对应的含义组成的<拼音-含义>集合，其中$i\in[1,7],n\in[1,7]$，并且每个多音字的不同读音可能有多个含义。

4 PGAS算法

4.1 算法假设

本文将评论数据作为输入得到输出结果得分s，将s与阈值进行比较，得到预测的分类结果。由于评论数据已经给定正例和负例，故仅需判断是否分类正确即可。训练集中的正负例分别标记为1和0，当$s>\lambda$时，判断该输入为正样本；当$s\le\lambda$时，则判断该输入为负样本。样本输入后得到得分s，若s在得分阈值α和β之间则为中性，情感倾向较弱或者不含情感倾向；$s>\alpha$则偏正面；$s<\beta$则偏负面。

评论文本中的多音字可能多个读音都不包含情感倾向，这不会对PGAS算法产生影响。因为无论是否包含情感倾向，在原句中进行多音字替换时，除了判断核心词的情感倾向外，核心词自身以及其他词含有多音字也会对最终的结果产生影响。

4.2 扰动定位

PGAS算法中，需要定位多音字的位置，根据WordHandling算法[8]的字删除评分方法的启发，提出了改进的定向词删除评分机制(Targeted Deletion Score, TDS)进行多音字位置的位置重要性判断，根据重要性进行多音字替换操作。对输入样本W进行分词得到$W=[w_1,w_2,\cdots,w_n]$，其中n表示输入样本的字符长度，再将输入样本进行拼音化处理，通过与构建的多音字字典中进行比对，找到输入样本中全部的多音字位置，对序列W中的第i个多音字，计算整个样本和删除该字之后样本的输入分数差值

$$\begin{aligned}\mathrm{TDS}(w_i)=&f(w_1,\cdots,w_{i-1},w_i,w_{i+1},\cdots,w_n)\\&-f(w_1,\cdots,w_{i-1},w_{i+1},\cdots,w_n)\end{aligned}\tag{3}$$

4.3 算法描述

PGAS算法的核心思想是通过对输入文本中的多音字进行替换来达到改变模型预测结果的目的。具体包含以下两个步骤：首先构建多音字字典，然后根据多音字字典来替换原始样本中的多音字，生成对抗样本。

PGAS算法进行对抗样本生成主要是通过多音字替换实现，在实际应用中，多音字的不同读音具有不同的含义，将不同读音的汉字视为相对独立的两个汉字，因此其对应的向量表示也完全不同。为了清晰地描述PGAS算法原理，展示多音字的读音不同导致的向量表示变化，相关描述如图2所示。

图2中，句子由$[X_1,X_2,\cdots,X_T]$等T个汉字组成，对其进行汉克尔矩阵化(Hankelization)操作，变形为$[\widehat{\boldsymbol{X}}_1,\widehat{\boldsymbol{X}}_2,\cdots,\widehat{\boldsymbol{X}}_T]$，其中$\widehat{\boldsymbol{X}}_i$表示为汉字对应的矩阵形式。通过PGAS算法，对含有多音字的汉字

执行替换操作，即将图3中红色框处的0变为1，得到改变后的矩阵$\widehat{\boldsymbol{M}}_i$，即可得到更新后的$X_{\text{new}}$。图2详细描述了PGAS算法通过矩阵变换得到不同含义且不同读音的同形字形式化流程。针对代替换字读音在2个以上的汉字，通过计算其IMD值(具体计算方法见5.3节)，选取IMD值最大的读音进行替换。IMD值越大，表明两读音之间的偏移量越大，原始语义偏离越大，越容易起到攻击的效果。

5　实验及结果分析

本文选用的数据分为两部分，构建多音字字典时，采用的多音字数据来源于魏星等人[29]提出的中文科技术语多音字表中数据。生成对抗样本数据来源于谭松波公开的酒店评论数据、微博评论数据以及商品评论数据，在针对数据集中的数据进行分词后，采用人工标注的方法对其中的多音字进行标注。

5.1　实验设置

本文在不同网络模型上进行了对抗样本有效性验证，通过对多种类型的情感分类文本数据集进行统计分析，数据集的相关信息汇总见表1。

本文使用多种类型的数据构建出试验数据集，

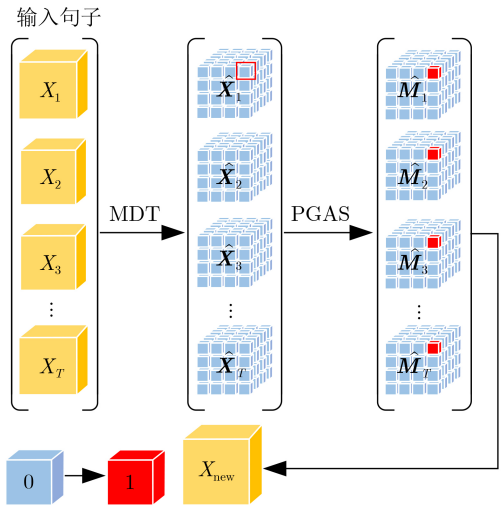


图 2 PGAS算法替换向量描述样例

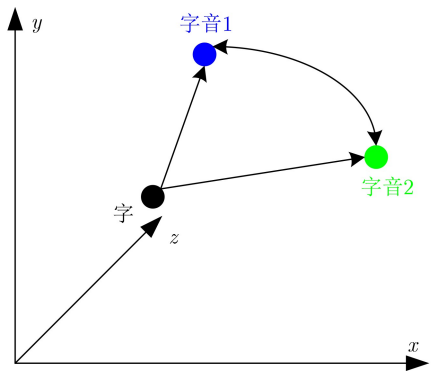


图 3 字音1和字音2在坐标系中的转移

每种评论文本平均有6000条。由于本实验仅需要验证含有情感倾向的多音字所属的评论语句，经过人工筛选后，评论中剩余5886条含有多音字的语句，正负样例比重相同。对评论语句的多音字进行人工标注，将标注好的数据分为训练集和测试集，训练集和测试的数据比例为3:7。在实验中，分类阈值λ设为0.5，α和β的值分别为0.6和0.4。使用PGAS生成对抗样本数据，并将对抗样本数据，传入长短期记忆网络(Long Short-Term Memory, LSTM)和卷积神经网络(Convolutional Neural Network, CNN)等传统模型和部分最新的情感分类模型测试生成样本的效果。

为了验证所提出来的PGAS的有效性，首先生成对抗样本，将这些对抗样本作为输入，对现有最新的情感分类模型实施黑盒攻击。对于PGAS攻击效果的衡量是根据神经网络模型对对抗样本检测的准确率下降程度体现，准确率下降越多，则攻击效果越好。

5.2　实验结果及分析

实验使用酒店评论数据、微博评论数据以及商品评论数据，各数据集的项目内容如表1所示。利用最新的情感分类模型对提出的PGAS算法生成的对抗样本进行验证。同时，为了与其他对抗样本生成方法作比较，特设置对照实验，以期验证本方法的优势。关于模型检测准确性如表2—表4所示，对比方法有词处理生成方法(WordHandling)[8]、词级黑盒对抗样本生成方法(CwordAttacker)[9]、黑盒对抗样本生成方法(DeepWordBug)[13]和快速生成对抗样本方法(FastWordBug) [14]。测试模型分别有：支持向量机(Support Vector Machines, SVM)[30]、长短期记忆网络(Long Short-Term Memory, LSTM)[31]、深度记忆网络(MemNet)[32]、方面交互网络(IAN)[33]、注意力集中注意网络(AOA)[34]、注意编码网络(AEN-GloVe)[35]、LSTM+SynATT[36]、目标依赖图注意网络(TD-GAT)[37]、特定方面图建卷积网络(ASGCN)[38]、卷积神经网络(Convolutional Neural Network, CNN)[39]和分层式卷积神经网络情感分类(pos-ACNN-CNN)[40]。

本文在相同的实验环境下，在多个公开数据集上与多种对抗样本生成方法生成的样本，用11种文

表 1　实验数据集

项目	酒店评论数据	微博评论数据	商品评论数据
任务类型	情感倾向性分类	情感倾向性分类	情感倾向性分类
分类数目	2	2	2
训练集(条)	4120	70000	42130
测试集(条)	1766	30000	18056
多音字数量(个)	2556952	7391456	6585441

表 2 在酒店评论数据集上的对比试验结果(%)

测试模型	无修改	对比方法								本文方法	
		WordHandling		CWordAttacker		DeepWordBug		FastWordBug		PGAS	
	准确率	准确率	降低幅度	准确率	降低幅度	准确率	降低幅度	准确率	降低幅度	准确率	降低幅度
SVM	76.35	72.19	4.16	71.03	5.32	69.18	7.17	70.15	6.20	52.36	23.99
LSTM	83.21	76.25	6.96	74.29	8.92	72.51	10.70	75.22	7.99	62.17	21.04
MemNet	77.12	70.31	6.81	72.59	4.53	70.15	6.97	69.19	7.93	58.63	18.49
IAN	86.31	81.25	5.06	83.26	3.05	78.32	7.99	78.29	8.02	64.92	21.39
AOA	79.91	71.26	8.65	73.29	6.62	68.25	11.66	70.53	9.38	60.15	19.76
AEN-GloVe	86.32	79.81	6.51	81.07	5.25	77.16	9.16	80.09	6.23	68.37	17.95
LSTM+SynATT	88.61	83.59	5.02	82.56	6.05	78.39	10.22	81.37	7.24	61.84	26.77
TD-GAT	78.36	72.20	6.16	73.21	5.15	72.19	6.17	71.24	7.12	60.23	18.13
ASGCN	82.97	77.18	5.79	77.41	5.56	71.05	11.92	73.08	9.89	61.08	21.89
CNN	82.36	74.21	8.15	76.38	5.98	69.91	12.45	69.51	12.85	59.39	22.97
pos-ACNN-CNN	76.28	70.15	6.13	72.53	3.75	68.25	8.03	66.19	10.09	58.18	18.10

表 3 在微博评论数据集上的对比试验结果(%)

测试模型	无修改	对比方法								本文方法	
		WordHandling		CWordAttacker		DeepWordBug		FastWordBug		PGAS	
	准确率	准确率	降低幅度	准确率	降低幅度	准确率	降低幅度	准确率	降低幅度	准确率	降低幅度
SVM	74.25	68.32	5.93	69.04	5.21	66.03	8.22	63.51	10.74	53.21	21.04
LSTM	79.66	72.58	7.08	71.22	8.44	68.25	11.41	69.79	9.87	59.74	19.92
MemNet	73.28	66.82	6.46	65.39	7.89	64.51	8.77	59.21	14.07	54.09	19.19
IAN	80.39	74.41	5.98	76.28	4.11	73.28	7.11	74.07	6.32	59.74	20.65
AOA	77.21	68.25	8.96	63.05	14.16	62.89	14.32	64.19	13.02	53.66	23.55
AEN-GloVe	85.31	74.29	11.02	73.08	12.23	74.85	10.46	76.28	9.03	66.23	19.08
LSTM+SynATT	89.07	72.14	16.93	75.44	13.63	77.60	11.47	81.18	7.89	67.04	22.03
TD-GAT	83.06	76.33	6.73	73.98	9.08	72.56	10.50	74.61	8.45	54.39	28.67
ASGCN	80.17	69.19	10.98	71.04	9.13	69.04	11.13	62.88	17.29	56.18	23.99
CNN	76.33	68.38	7.95	66.37	9.96	69.71	6.62	69.44	6.89	57.20	19.13
pos-ACNN-CNN	70.94	61.25	9.69	59.37	11.57	61.43	9.51	60.07	10.87	59.33	11.61

表 4 在商品评论数据集上的对比试验结果(%)

测试模型	无修改	对比方法								本文方法	
		WordHandling		CWordAttacker		DeepWordBug		FastWordBug		PGAS	
	准确率	准确率	降低幅度	准确率	降低幅度	准确率	降低幅度	准确率	降低幅度	准确率	降低幅度
SVM	73.28	64.21	9.07	66.07	7.21	63.19	10.09	65.03	8.25	54.64	18.64
LSTM	77.04	69.07	7.97	67.45	9.59	68.17	8.87	68.29	8.75	53.21	23.83
MemNet	82.36	73.85	8.51	71.04	11.32	73.22	9.14	72.94	9.42	63.02	19.34
IAN	74.07	62.25	11.82	66.38	7.69	65.31	8.76	65.83	8.24	56.41	17.66
AOA	78.25	69.44	8.81	68.51	9.74	67.14	11.11	68.07	10.18	54.20	24.05
AEN-GloVe	81.33	70.03	11.30	73.09	8.24	70.25	11.08	72.55	8.78	55.97	25.36
LSTM+SynATT	85.60	76.49	9.11	78.21	7.39	73.21	12.39	74.54	11.06	61.08	24.52
TD-GAT	84.92	72.17	12.75	75.60	9.32	75.09	9.83	74.60	10.32	62.04	22.88
ASGCN	83.64	76.05	7.59	79.03	4.61	74.32	9.32	76.59	7.05	64.29	19.35
CNN	75.91	63.21	12.70	64.24	11.67	65.39	10.52	65.02	10.89	59.31	16.60
pos-ACNN-CNN	86.49	72.71	13.78	77.30	9.19	79.02	7.47	72.74	13.75	68.03	18.46

本情感分析方法进行了对比试验。从表2—表4中数据可以清晰看到，本文提出的PGAS方法，相较于其他对抗样本生成方法，分类准确度下降幅度最大，证明PGAS方法生成的对抗样本可以大幅度改变原来的分类准确度。PGAS在多种最新的情感分类模型上都取得了很好的成绩。本文的模型与11种分类模型做了对比实验，在酒店评论数据集上使得分类结果下降了17.95%～26.77%；在微博评论数据集上使得分类结果下降了11.61%～28.67%；在商品评论数据集上下降了16.60%～25.36%。为了更直观地表明本文提出方法PGAS的下降效果，在表2—表4中将对比模型对分类效果影响程度大于10%的进行了加粗展示。实验证明。PGAS方法下降幅度远超其他模型。

同时，针对几种对比方法WordHandling[8], CwordAttacker[9], DeepWordBug[13]和FastWordBug[14]，进行时间复杂度与空间复杂度的比较分析，包括PGAS在内的方法都需要对原始数据中的所有字符进行一轮遍历，遍历的同时使用标记法记录替换词或删减词的位置，故所需时间复杂度为$O(n)$。在空间复杂度方面，本方法仅需记录替换词或删减词的位置信息，不涉及到额外的动态分配空间，因此空间复杂度为$O(1)$。

5.3 对抗样本质量度量

对于生成的对抗样本质量评估，在图像中通常采用L_p范数进行度量，但图像的连续性导致离散的文本不能利用范数进行度量。基于Kusner等人[41]提出的词移距离(Word Mover's Distance, WMD)来计算样本之间的相似度，WMD距离越大，表明相似性越低，反之越高则越相似，其语义偏离程度越低。WMD是基于词之间的关联度来进行衡量，基于此，本文提出了改进的词移距离(Improved Mover's Distance, IMD)来衡量对抗样本的质量，通过计算两样本之间的拼音信息来衡量其语义偏离程度。计算如式(4)所示。

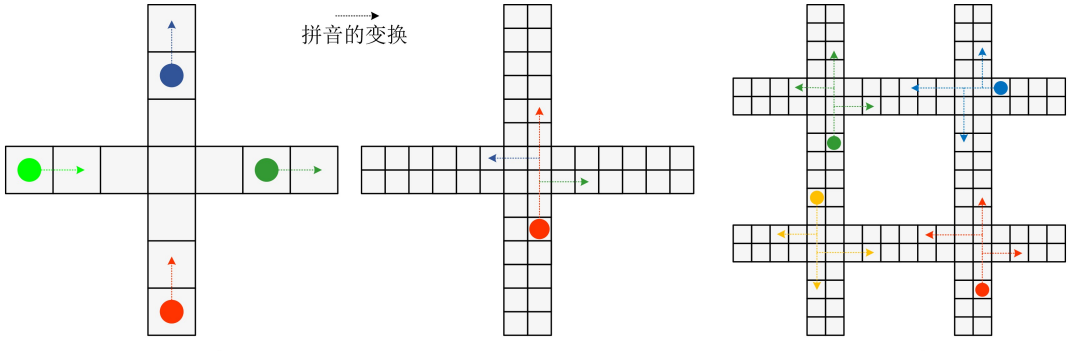
$$\left.\begin{aligned}&\min_{T\geq 0}\ \sum_{i,j}^{n} T_{ij}c(i,j)\\&\text{s.t.}\ \sum_{j=1}^{n} T_{ij}=d_i,\ \ \forall i\in\{1,2,\cdots,n\}\end{aligned}\right\}\tag{4}$$

其中，$c(i,j)$是多音字不同拼音词向量i和j之间的Euclidean距离，n为拼音的个数，d_i表示为多音字在原文中的TDS得分权重。由IMD的衡量方式可知，IMD主要考虑在拼音之间的移动距离还判断其语义偏离程度，因此计算不同拼音向量之间和的最小值来达到针对原始样本和对抗样本相比较的目的。在式(4)中，满足条件表示读音i和j之间的转换(如图3所示)。在坐标系中，同一个汉字的不同读音对应的坐标不同，在评价对抗样本质量时，应当考虑除了汉字之外，对抗样本和原文的读音之间的相似度情况。

图4分别描述了从字、词再到句子，拼音的变换情况。图4(a)为不同汉字的读音转移，汉字之间的读音不同，在进行不同字音转变时，具体表现为向量之间的转变(如图2)；图4(b)表示若词组中存在多个多音字，则多音字的变换方式会对生成的对抗样本产生影响；图4(c)则是将图4(b)的情况拓展到全文中。

本文利用WMD来计算对抗样本之间的偏离程度，利用IMD来进行拼音之间的相似度计算。若计算距离越大，则证明越不相似，反之则越相似。为了更好地验证本文提出方法的生成质量，分别从WMD和IMD两种衡量方法出发，与WordHandling[8], CwordAttacker[9], DeepWordBug[13]和FastWordBug[14]进行生成的样本质量做对比。在多种评论数据集生成的2 000条对抗样本中，各方法的WMD和IMD分布情况如表5所示。

从表5中可以明显看出，当生成2 000条对抗样本时，不同方法中，PGAS方法生成的样本在用WMD进行偏移程度测算时，全部在0～0.2范围内，而其



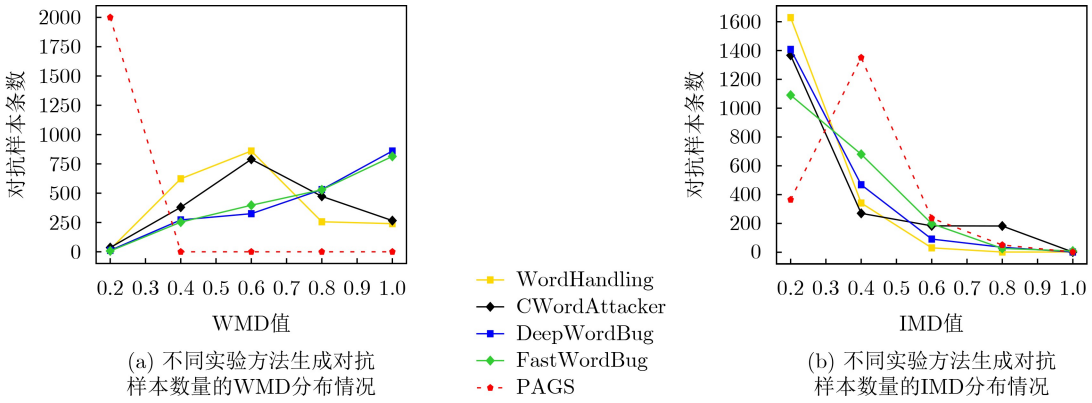
(a) 不同汉字　　(b) 不同词组　　(c) 不同句子

图 4 多音字3种拼音变换情况

他方法大多在0.4～0.6。同样地，在IMD进行拼音的偏移测算时，由于WordHandling方法[8]采用同音字替换方法，因此在拼音测算时比PGAS算法质量更优，但PGAS算法同样优于其他算法。可以得出结论，PGAS算法生成的对抗样本质量较高，且扰动幅度较小。数据折线图如图5所示。

表 5 不同实验方法生成对抗样本数量的WMD和IMD分布情况(条)

项目	值	对比方法				本文方法
		WordHandling	CWordAttacker	DeepWordBug	FastWordBug	PGAS
WMD	0-0.2	21	36	12	7	2000
	0.2-0.4	623	380	272	253	0
	0.4-0.6	860	789	325	397	0
	0.6-0.8	256	473	531	529	0
	0.8-1	240	266	860	814	0
IMD	0-0.2	1630	1368	1409	1091	364
	0.2-0.4	341	269	468	680	1352
	0.4-0.6	29	182	90	197	235
	0.6-0.8	0	181	33	25	49
	0.8-1	0	0	0	7	0



(a) 不同实验方法生成对抗样本数量的WMD分布情况

(b) 不同实验方法生成对抗样本数量的IMD分布情况

图 5 不同实验方法生成对抗样本数量的WMD和IMD分布情况

本文在实验操作时，只修改了字义和读音，并没有修改任何汉字。对于人类而言，阅读起来没有障碍，但是对机器来说，修改后的字义是符合机器模型运作的，对分类模型造成较强的干扰，达到了显著的实验效果。实验表明，通过PGAS算法，能够通过生成高质量的对抗样本来影响模型的分类结果。

6 结束语

本文针对目前中文领域生成的对抗样本扰动幅度大，质量不高的问题，提出一种面向中文文本分类的字符级对抗样本生成方法PGAS，并针对多种分类任务进行攻击实验。利用构建的多音字词典进行多音字替换，在多个评论文本数据集证明了该方法有效，且生成的对抗样本扰动幅度较小，语句含义表达完整。本文主要贡献在于：

(1) 通过PGAS方法生成了有效且高质量的对抗样本。利用PGAS方法生成的对抗样本，从字音和字形上同步进行质量检测，验证了生成的对抗样本保证了在最小扰动下，语义偏离最小，具有良好的可读性。

(2) 成功构建多音字字典。多音字字典的成功构建，为之后的从事多音字相关领域研究的学者提供了研究基础。本文构建的多音字字典，含有包括多音字的字音、字义以及不同读音和字义之间关系在内的多种信息，较为完整。

在今后的工作中，将进一步考虑多音字的多种含义在对抗样本生成中的影响，同时也会针对PGAS算法生成的对抗样本考虑防御措施，以期提高模型的鲁棒性。

参 考 文 献

[1] PAPERNOT N, MCDANIEL P, SWAMI A, *et al*. Crafting adversarial input sequences for recurrent neural

networks[C]. MILCOM 2016 - 2016 IEEE Military Communications Conference, Baltimore, USA, 2016: 49–54. doi: 10.1109/MILCOM.2016.7795300.

[2] WANG Boxin, PEI Hengzhi, PAN Boyuan, *et al.* T3: Tree-autoencoder constrained adversarial text generation for targeted attack[C/OL]. The 2020 Conference on Empirical Methods in Natural Language Processing, 2020: 6134–6150. doi: 10.18653/v1/2020.emnlp-main.495..

[3] LE T, WANG Suhang, and LEE D. MALCOM: Generating malicious comments to attack neural fake news detection models[C]. 2020 IEEE International Conference on Data Mining, Sorrento, Italy, 2020: 282–291. doi: 10.1109/ICDM50108.2020.00037.

[4] MOZES M, STENETORP P, KLEINBERG B, *et al.* Frequency-guided word substitutions for detecting textual adversarial examples[C/OL]. The 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021: 171–186. doi: 10.18653/v1/2021.eacl-main.13..

[5] TAN S, JOTY S, VARSHNEY L, *et al.* Mind your Inflections! Improving NLP for non-standard Englishes with Base-Inflection encoding[C/OL]. The 2020 Conference on Empirical Methods in Natural Language Processing, 2020: 5647–5663. doi: 10.18653/v1/2020.emnlp-main.455..

[6] 潘文雯, 王新宇, 宋明黎, 等. 对抗样本生成技术综述[J]. 软件学报, 2020, 31(1): 67–81. doi: 10.13328/j.cnki.jos.005884.
PAN Wenwen, WANG Xinyu, SONG Mingli, *et al.* Survey on generating adversarial examples[J]. *Journal of Software*, 2020, 31(1): 67–81. doi: 10.13328/j.cnki.jos.005884.

[7] MILLER D, NICHOLSON L, DAYOUB F, *et al.* Dropout sampling for robust object detection in open-set conditions[C]. 2018 IEEE International Conference on Robotics and Automation, Brisbane, Australia, 2018: 3243–3249. doi: 10.1109/ICRA.2018.8460700.

[8] 王文琦, 汪润, 王丽娜, 等. 面向中文文本倾向性分类的对抗样本生成方法[J]. 软件学报, 2019, 30(8): 2415–2427. doi: 10.13328/j.cnki.jos.005765.
WANG Wenqi, WANG Run, WANG Li'na, *et al.* Adversarial examples generation approach for tendency classification on Chinese texts[J]. *Journal of Software*, 2019, 30(8): 2415–2427. doi: 10.13328/j.cnki.jos.005765.

[9] 仝鑫, 王罗娜, 王润正, 等. 面向中文文本分类的词级对抗样本生成方法[J]. 信息网络安全, 2020, 20(9): 12–16. doi: 10.3969/j.issn.1671-1122.2020.09.003.
TONG Xin, WANG Luona, WANG Runzheng, *et al.* A generation method of word-level adversarial samples for Chinese text classiifcation[J]. *Netinfo Security*, 2020, 20(9): 12–16. doi: 10.3969/j.issn.1671-1122.2020.09.003.

[10] BLOHM M, JAGFELD G, SOOD E, *et al.* Comparing attention-based convolutional and recurrent neural networks: Success and limitations in machine reading comprehension[C]. The 22nd Conference on Computational Natural Language Learning, Brussels, Belgium, 2018: 108–118. doi: 10.18653/v1/K18-1011.

[11] NIU Tong and BANSAL M. Adversarial over-sensitivity and over-stability strategies for dialogue models[C]. The 22nd Conference on Computational Natural Language Learning, Brussels, Belgium, 2018: 486–496. doi: 10.18653/v1/K18-1047.

[12] EBRAHIMI J, LOWD D, and DOU Dejing. On adversarial examples for character-level neural machine translation[C]. The 27th International Conference on Computational Linguistics, Santa Fe, USA, 2018: 653–663.

[13] GAO Ji, LANCHANTIN J, SOFFA M L, *et al.* Black-box generation of adversarial text sequences to evade deep learning classifiers[C]. 2018 IEEE Security and Privacy Workshops, San Francisco, USA, 2018: 50–56. doi: 10.1109/SPW.2018.00016.

[14] GOODMAN D, LV Zhonghou, and WANG Minghua. FastWordBug: A fast method to generate adversarial text against NLP applications[J]. arXiv preprint arXiv: 2002.00760, 2020. doi: 10.48550/arXiv.2002.00760..

[15] EBRAHIMI J, RAO Anyi, LOWD D, *et al.* HotFlip: White-box adversarial examples for text classification[C]. The 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 2018: 31–36. doi: 10.18653/v1/P18-2006.

[16] SONG Liwei, YU Xinwei, PENG H T, *et al.* Universal adversarial attacks with natural triggers for text classification[C/OL]. The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021: 3724–3733. doi: 10.18653/v1/2021.naacl-main.291..

[17] LI Dianqi, ZHANG Yizhe, PENG Hao, *et al.* Contextualized perturbation for textual adversarial attack[C/OL]. The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021: 5053–5069. doi: 10.18653/v1/2021.naacl-main.400..

[18] TAN S, JOTY S, KAN M Y, *et al.* It's Morphin' time! Combating linguistic discrimination with inflectional perturbations[C/OL]. The 58th Annual Meeting of the

Association for Computational Linguistics, 2020: 2920–2935. doi: 10.18653/v1/2020.acl-main.263..

[19] LI Linyang, MA Ruotian, GUO Qipeng, *et al.* BERT-ATTACK: Adversarial attack against BERT using BERT[C/OL]. The 2020 Conference on Empirical Methods in Natural Language Processing, 2020: 6193–6202. doi: 10.18653/v1/2020.emnlp-main.500.

[20] ZANG Yuan, QI Fanchao, YANG Chenghao, *et al.* Word-level textual adversarial attacking as combinatorial optimization[C/OL]. The 58th Annual Meeting of the Association for Computational Linguistics, 2020: 6066–6080. doi: 10.18653/v1/2020.acl-main.540..

[21] CHENG Minhao, YI Jinfeng, CHEN Pinyu, *et al.* Seq2Sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples[C]. The 34th AAAI Conference on Artificial Intelligence, New York, USA, 2020: 3601–3608. doi: 10.1609/aaai.v34i04.5767.

[22] JIA R and LIANG P. Adversarial examples for evaluating reading comprehension systems[C]. The 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 2017: 2021–2031. doi: 10.18653/v1/D17-1215.

[23] MINERVINI P and RIEDEL S. Adversarially regularising neural NLI models to integrate logical background knowledge[C]. The 22nd Conference on Computational Natural Language Learning, Brussels, Belgium, 2018: 65–74. doi: 10.18653/v1/K18-1007.

[24] WANG Yicheng and BANSAL M. Robust machine comprehension models via adversarial training[C]. The 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, USA, 2018: 575–581. doi: 10.18653/v1/N18-2091.

[25] RIBEIRO M T, SINGH S, and GUESTRIN C. Semantically equivalent adversarial rules for debugging NLP models[C]. The 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 2018: 856–865. doi: 10.18653/v1/P18-1079.

[26] IYYER M, WIETING J, GIMPEL K, *et al.* Adversarial example generation with syntactically controlled paraphrase networks[C]. The 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, USA, 2018: 1875–1885. doi: 10.18653/v1/N18-1170.

[27] HAN Wenjuan, ZHANG Liwen, JIANG Yong, *et al.* Adversarial attack and defense of structured prediction models[C/OL]. The 2020 Conference on Empirical Methods in Natural Language Processing, 2020: 2327–2338.

[28] WANG Tianlu, WANG Xuezhi, QIN Yao, *et al.* CAT-Gen: Improving robustness in NLP models via controlled adversarial text generation[C/OL]. The 2020 Conference on Empirical Methods in Natural Language Processing, 2020: 5141–5146. doi: 10.18653/v1/2020.emnlp-main.417..

[29] 魏星, 王小辉, 魏亮, 等. 基于规范科技术语数据库的科技术语多音字研究与读音推荐[J]. 中国科技术语, 2020, 22(6): 25–29. doi: 10.3969/j.issn.1673-8578.2020.06.005.

WEI Xing, WANG Xiaohui, WEI Liang, *et al.* Pronunciation recommendations on polyphonic characters in terms based on the database of standardized terms[J]. *China Terminology*, 2020, 22(6): 25–29. doi: 10.3969/j.issn.1673-8578.2020.06.005.

[30] KIRITCHENKO S, ZHU Xiaodan, CHERRY C, *et al.* NRC-Canada-2014: Detecting aspects and sentiment in customer reviews[C]. The 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 2014: 437–442. doi: 10.3115/v1/S14-2076.

[31] TANG Duyu, QIN Bing, FENG Xiaocheng, *et al.* Effective LSTMs for target-dependent sentiment classification[C]. COLING 2016, the 26th International Conference on Computational Linguistics, Osaka, Japan, 2016: 3298–3307.

[32] TANG Duyu, QIN Bing, and LIU Ting. Aspect level sentiment classification with deep memory network[C]. The 2016 Conference on Empirical Methods in Natural Language Processing, Austin, USA, 2016: 214–224. doi: 10.18653/v1/D16-1021.

[33] MA Dehong, LI Sujian, ZHANG Xiaodong, *et al.* Interactive attention networks for aspect-level sentiment classification[C]. The 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 2017: 4068–4074.

[34] HUANG Binxuan, OU Yanglan, and CARLEY K M. Aspect level sentiment classification with attention-over-attention neural networks[C]. The 11th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, Washington, USA, 2018: 197–206. doi: 10.1007/978-3-319-93372-6_22.

[35] SONG Youwei, WANG Jiahai, JIANG Tao, *et al.* Targeted sentiment classification with attentional encoder network[C]. The 28th International Conference on Artificial Neural Networks, Munich, Germany, 2019: 93–103. doi: 10.1007/978-3-030-30490-4_9.

[36] HE Ruidan, LEE W S, NG H T, *et al*. Effective attention modeling for aspect-level sentiment classification[C]. The 27th International Conference on Computational Linguistics, Santa Fe, USA, 2018: 1121–1131.

[37] HUANG Binxuan and CARLEY K M. Syntax-aware aspect level sentiment classification with graph attention networks[C]. The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019: 5469–5477. doi: 10.18653/v1/D19-1549.

[38] ZHANG Chen, LI Qiuchi, and SONG Dawei. Aspect-based sentiment classification with aspect-specific graph convolutional networks[C]. The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019: 4568–4578. doi: 10.18653/v1/D19-1464.

[39] WANG Yuanchao, LI Mingtao, PAN Zhichen, *et al*. Pulsar candidate classification with deep convolutional neural networks[J]. *Research in Astronomy and Astrophysics*, 2019, 19(9): 133. doi: 10.1088/1674-4527/19/9/133.

[40] 唐恒亮, 尹棋正, 常亮亮, 等. 基于混合图神经网络的方面级情感分类[J]. 计算机工程与应用, 2023, 59(4): 175–182. doi: 10.3778/j.ssn.1002-8331.2109-0172.
TANG Hengliang, YIN Qizheng, CHANG Liangliang, *et al*. Aspect-level sentiment classification based on mixed graph neural network[J]. *Computer Engineering and Applications*, 2023, 59(4): 175–182. doi: 10.3778/j.ssn.1002-8331.2109-0172.

[41] KUSNER M J, SUN Yu, KOLKIN N I, *et al*. From word embeddings to document distances[C]. The 32nd International Conference on Machine Learning, Lille, France, 2015: 957–966.

张顺香：男，教授，研究方向为情感计算、人物关系挖掘.
吴厚月：男，硕士生，研究方向为对抗样本生成、关系抽取.
朱广丽：女，副教授，研究方向为情感计算、复杂网络分析.
许　鑫：男，硕士生，研究方向为自然语言处理、因果关系抽取.
苏明星：男，硕士生，研究方向为自然语言处理、情感计算.

责任编辑：余　蓉