

基于标点符号插入的中文对抗文本生成方法

张谦, 闫巧

(深圳大学计算机与软件学院, 广东 深圳 518060)

摘要: 自然语言处理模型容易受到对抗文本的影响。现有的针对中文对抗文本的生成方法主要采用形近字替换或同音字替换等方式, 在面对鲁棒性较强的预训练模型时, 对抗文本中的扰动数增加导致文本的流畅性和可读性下降, 从而生成的对抗文本质量差。而英文对抗文本中符号插入方法不能完全适用于中文。同时, 在黑盒场景下, 先验知识的缺乏导致难以生成高质量的对抗文本。面向中文文本分类任务, 提出了一种基于标点符号插入的中文对抗文本生成方法。在黑盒设置中, 该方法采用一种词性重要性计算方法, 并结合标点符号插入设计了一种适用于中文的字符级扰动方法, 以实现对抗文本的生成。实验证明, 针对文本分类任务, 所提方法在使用两个真实数据集训练的 LSTM (long short-term memory) 和 BERT (bidirectional encoder representations from transformers) 模型上攻击成功率显著提升, 同时, 成功避免了直接破坏原始句子, 保持了原始句意。在测试中, 该方法能够达到97%的语义相似度, 明显优于基线方法。

关键词: 中文文本分类; 对抗文本生成; 黑盒攻击

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2025026

Chinese adversarial text generation method based on punctuation insertion

ZHANG Qian, YAN Qiao

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

Abstract: The susceptibility of natural language processing models to adversarial texts has been a significant concern. Current methods for generating adversarial texts in Chinese were mainly based on replacing characters with visually similar or homophonic ones. However, when faced with robust pre-trained models, these methods led to increased perturbations in adversarial texts, resulting in reduced fluency and readability, and thus generating low-quality adversarial texts. Moreover, symbol insertion methods used in English adversarial texts were not entirely applicable to Chinese. Additionally, in a black-box scenario, the lack of prior knowledge made it difficult to generate

收稿日期: 2024-08-25; 修回日期: 2024-12-21

通信作者: 闫巧, yanq@szu.edu.cn

基金项目: 国家自然科学基金(61976142); 深圳科技计划项目(JCYJ20210324093609025)

Foundation Items: The National Natural Science Foundation of China (61976142), Shenzhen Science and Technology Planning Project (JCYJ20210324093609025)

引用格式: 张谦, 闫巧. 基于标点符号插入的中文对抗文本生成方法[J]. 网络与信息安全学报, 2025, 11(2): 161-174.

Citation Format: ZHANG Q, YAN Q. Chinese adversarial text generation method based on punctuation insertion[J]. Chinese Journal of Network and Information Security, 2025, 11(2): 161-174.

high-quality adversarial texts. A punctuation-based method for generating adversarial texts for Chinese text classification tasks was proposed. Under a black-box setting, a novel part-of-speech importance calculation was utilized and combined with punctuation insertion to design a character-level perturbation approach suitable for Chinese, achieving the generation of adversarial texts. Experiments were conducted, and the results demonstrated that for text classification tasks, the proposed method significantly improved the attack success rate on LSTM and BERT models trained with two real-world datasets. Furthermore, the method successfully avoided direct destruction of the original sentences and maintained the original meaning. In the tests, a semantic similarity of up to 97% was achieved, which was significantly better than the baseline methods.

Keywords: Chinese text classification, adversarial text generation, black-box attack

0 引言

深度学习技术在多个领域如计算机视觉、自然语言处理 (natural language processing, NLP) 和音频分析中发挥着重要作用。然而, 这些模型容易受到特定类型的攻击^[1-3], 即对抗样本攻击。这种攻击通过在标准样本中巧妙地引入微小变化, 导致深度学习系统产生错误的判断。研究如何生成对抗样本, 不仅可以评估这些模型的稳定性, 还能揭示它们的潜在缺陷, 从而采取预防措施。此外, 利用对抗样本进行训练可以增强模型的鲁棒性, 抵御潜在的攻击。

在自然语言处理领域, 对抗攻击同样是一个不可忽视的问题。2014年, Szegedy等^[1]首次揭示了通过在原始图像中添加微小的扰动可以生成导致深度学习模型误判的样本, 这类样本被称为对抗样本。早期的研究主要集中在计算机视觉上。然而, 近年来, NLP领域的对抗样本研究取得了显著进展。与图像数据的连续性不同, 文本数据是离散的, 这使得在文本中引入难以被人类察觉的微小变化更具挑战性。因此, 研究者提出了多种生成对抗文本的方法。文本分类作为NLP中的一项核心任务, 不仅为其他NLP任务提供了基础, 也是文本对抗研究的重点。目前, 大多数对抗文本生成技术是围绕文本分类任务^[4]设计的。当深度学习模型遭受对抗文本的攻击时, 可能会导致错误的分类结果, 进而使一些关键信息逃避模型的检测。

随着预训练语言模型的高速发展, 文本分类模型的稳健性得到了显著提升, 使得对抗性文本的创建变得更加困难。目前, 大多数针对文本分类任务的对抗性文本生成技术是为英文文本设计的, 而专门针对中文文本的相应方法相对较少。由于英文和中文之间存在显著的语言特性差异,

许多英文的对抗性文本生成策略并不完全适用于中文环境^[5]。而仅有的一些中文对抗方法大多采用同义字替换或同音字替换等方式, 在面对鲁棒性较强的模型时, 往往需要使用大量的扰动对原始语句中的字词进行替换, 使扰动比例增大, 从而导致对抗文本的流畅性和可读性下降, 甚至使原始句意发生改变。此外, 尽管词级攻击将原句中的词变为有扰动的同义词, 使这些扰动对人类几乎看不见, 但句子中多个同义词替换的累积效应会使样本更难理解。在攻击特定领域的文本时, 攻击必须找到扰动词同义词, 如果算法使用没有领域知识的一般词嵌入, 则可能具有挑战性。为了探究如何生成高质量的对抗文本, 相应的对抗文本生成方法也有待研究。

标点符号插入被认为是一种行之有效的攻击方法^[6], 其不受字符扰动或单词替换的限制, 因为语法检查器通常难以探测到标点符号的变化。同时, 这种方法不会彻底改变句子的意义, 且保持了良好的可读性。在一些针对英文的对抗研究中, 插入标点符号会导致深度学习模型的性能下降^[7], 因为标点符号包含了模型正确运行所需的关键信息^[8]。此外, 标点符号的插入也可以保留可能被恶意用户利用的对抗性下游信息^[9]。在英文的攻击方法中, 标点符号的插入通常发生在一个单词的字母之间。然而, 在英文文本阅读中, 单词被视作最小的单位。大量插入标点符号在一个单词中仍会对对抗文本的可读性造成影响。相比之下, 在中文中, 只有部分左右结构的汉字拆分后仍然保持可读性。但是, 在拆分的基础上再插入标点符号会极大地影响可读性。因此, 英文中常用的标点符号插入方法并不完全适用于中文文本。对于中文文本的对抗样本生成, 需要探索

更适合其语言特性的方法。在中文文本的对抗攻击中, 标点符号攻击仍然是一个尚未充分研究的领域。以前的研究工作只是随意地探索插入空格或句号, 而忽略了哪些具体的标点符号适合用于入侵攻击, 以及它们是否能够全面地反映攻击目标。这表明还有许多有待深入探索的机会, 以改进和优化中文对抗文本生成方法。

不同于英文, 中文缺乏天然的词间空格, 因此在文本处理中通常需要借助分词工具来确定不同词汇之间的边界。在对中文文本进行对抗攻击时, 采取选择特定符号, 并将其有意地插入对最终结果有重大影响的关键词中, 以干扰文本的分词结构, 从而扰乱深度学习模型的分类判断。此外, 研究发现, 模型对某些字符存在分类偏向, 因此有意地选择较高分类偏向的字符来进行插入扰动。例如, 在进行二分类任务时, 仅需插入两个字符即可实施对抗攻击, 这不仅减少了攻击操作的复杂性, 还降低了对模型的查询次数。本文针对两类常用的文本分类模型双向长短时记忆(bidirectional long short-term memory, BiLSTM)网络和经过预训练的基于Transformer的双向编码器(bidirectional encoder representations from transformers, BERT)进行研究。

本文的主要贡献包括3个方面。

1) 提出了基于标点符号插入的中文文本对抗样本生成方法, 标点符号的插入避免了直接破坏原句, 增强句子的可读性, 获得高达97%的语义相似度。

2) 提出了一种词语重要性排序方法, 结合标点符号插入, 构成一种黑盒攻击方法, 实验表明, 提出的黑盒攻击方法攻击效果强于随机攻击方法, 且仅次于灰盒攻击方法。

3) 在情感分析数据集ChnSentiCorp和新闻数据集THUNews上进行实验, 在BERT模型和BiLSTM模型上分别达到80%和95%以上的攻击成功率, 同时, 保证对抗文本具有90%以上的语义相似度。

1 相关工作

当前的自然语言处理攻击技术主要分为3种类型: 白盒攻击、灰盒攻击和黑盒攻击。白盒攻击依赖于对目标模型的参数、损失函数以及梯度信息的访问, 以产生对抗样本。这种攻击方式在

计算上成本较高, 并且通常需要对目标模型的内部机制有深入了解, 这在大多数实际应用场景中是不可行的。灰盒攻击则利用目标模型提供的置信度分数或类别概率来创建对抗性输入。先前的灰盒攻击研究通常分为两个关键步骤: 构建搜索空间和制定基于词语重要性进行搜索的策略。黑盒攻击则是依据目标模型的输出标签来生成对抗性文本, 这种攻击方式挑战性较大, 目前相关的研究还不多。

基于上述框架, 研究者提出了许多英文对抗样本生成方法。文献[10]研究了在灰盒设置下对文本分类和文本隐含模型的对抗文本生成方法TextFooler, 提出了一种词重要性排序方法, 并使用反拟合嵌入空间的同义词进行扰动。文献[11]提出了一种在白盒设置下新的词级攻击模型Sememe PSO, 其中包括基于语素的词替换方法和基于粒子群优化的搜索算法。文献[12]提出了一种灰盒设置下字符级英文文本对抗攻击方法DeepWordBug, 该方法首先使用Delete Score(DS)和Forward Score(FS)函数对单词进行打分并排序, 然后对单词中的部分字母进行替换、删除和增加等操作来生成对抗文本。文献[7]基于DeepWordBug提出了一种基于标点符号插入的英文对抗文本生成方法DeepWordBugPunc, 该方法在一个单词的相邻字母间插入字符进行扰动, 如将“basketball”变换为“bas.ket.ball”。但此方法不能直接应用于中文, 由于单个汉字不像单个单词那样容易分隔来进行插入操作。

学者提出了多种中文对抗性文本生成技术。WordHanding^[13]通过计算词的重要性, 并在灰盒环境中利用同音词和谐音字来生成对抗文本。该方法在词重要性评估中融入了TF-IDF评分机制, 以优化对文本中情感倾向关键词的处理, 从而提高情感分析任务的评分准确性。Argot^[14]专注于中文和英文的差异, 并提出了5种专门针对中文文本的扰动替换策略, 包括同音词和谐音词替换(Homophone)、汉字重新排序(Shuffle)、同义词替换(Synonyms)、字形相似字替换(Glyph)以及汉字拆分(splitting-character, SC)。Argot结合了DS词重要性评估方法, 为通用领域的文本分类任务设计了一套中文对抗文本生成方法, 实验

结果表明, Homophone、SC 和 Glyph 攻击效果最佳。但当扰动比例增加时, 会使生成的对抗文本的质量大大下降, 如 Argot 生成的一个对抗样本为“呢囚, 蓝球, 赛事较流力口”, 而原文本数据为“足球, 篮球, 赛事交流加”, 这样的对抗文本的可读性很差。WordChange^[15]通过应用 DS 排序算法对文本中的关键词汇进行重要性排序, 然后采用汉字替换、特殊符号添加和拆分文字等策略来创建对抗文本, 在符号插入时, 忽略了哪些具体的标点符号适合用于入侵攻击, 以及它们是否能够全面地反映攻击目标的鲁棒性。CWordAttacker^[16]同样利用 DS 算法对关键的词汇进行排序, 并结合繁体字转换、拼音表达重写、特殊符号添加和汉字替换等手段来生成对抗文本。CWordAttacker 同样只是对符号插入进行了简单的探索, 以改进和优化中文对抗文本的生成方法。Sentence-Pieces^[17]提出了一个预先训练的语言 BERT 模型作为替代生成器, 使用生成的具有相同意思的句子片段来生成汉语中的对抗文本。在生成的敌对例子中, 替换不是字符或单词, 而是“片段”。CBT (Chinese BERT tricker)^[18]则提出了一种基于汉字的词语重要性排序算法——ICCL (important Chinese character localization), 并在 BERT 模型的掩码任务基础上, 提出了一种包含两种替换方法的扰动策略, 专门用于中文环境中对关键词汇的替换。

由于中文与英文在语言特性上的差异, 英文文本中的标点符号插入使用规则并不适用于中文。因此, 本文针对中文文本的独特性, 对原本用于英文的标点符号插入技术进行了改进, 提出了一套适合中文的对抗文本生成策略。此外, 在黑盒环境下, 本文提出了一种基于词性和词重要性排序方法, 以增强对抗文本的生成效果。该方法有效地揭示了中文分类模型在面对文本对抗攻击时的潜在弱点, 为后续的防御措施提供了重要的参考依据。

2 方法论

2.1 对抗文本形式化表示

给定一个预先训练的文本分类模型 $F: X \rightarrow Y$, 将输入文本序列 X 映射到一组分类 Y , 对于一个正常的文本数据 $x \in X$, x 的真实标签 $y \in Y$, 目标是生成一个对抗样本 x_{adv} , 使 $F(x_{adv}) = t \ (t \neq y)$ 。

同时需要满足 $S(x, x_{adv}) \geq \varepsilon$, S 是一个相似度约束函数, ε 为约束值。

2.2 方法论

本文研究了在灰盒和黑盒设置中的标点符号攻击。具体来说, 在灰盒场景下, 攻击者只能获取受害者模型输出的各类别概率值, 本文使用 Textfooler^[10]结合标点符号插入进行灰盒攻击; 在黑盒场景下, 攻击者只能获取受害者模型输出的最高置信度类别的标签值, 本文使用提出的基于词性的词重要性排序方法结合标点符号插入进行黑盒攻击。

2.3 词重要性排序方法

词重要性排序是研究输入文本数据中哪些词对分类模型更敏感, 即哪些词具有更强的能力改变文本数据分类模型的预测标签。通过词重要性排序, 可以首先替换对分类模型最敏感的词, 以确保对抗样本与原始文本数据之间的差异最小, 下面介绍一种源自 Textfooler^[10]的词重要性排序方法, 用于灰盒攻击。

给定 n 个单词 $X = \{w_1, w_2, \dots, w_n\}$, 观察到只有一些关键词会成为模型 F 预测的影响信号。在对抗文本生成中, 许多攻击方法使用单词重要性排名搜索对抗文本。给定一个经过切词后含有 n 个词语的样本, $x = \{w_1, w_2, \dots, w_n\}$, 使用 I_{w_i} 来衡量单词 $w_i \in x$ 对分类结果 $F(x) = y$ 的影响。将删除词语 w_i (或使用未知 token 符号 <UNK> 代替) 后的句子表示为 $x_{/w_i} = x_{/\{w_i\}} = \{w_1, \dots, w_{i-1}, w_{i+1}, \dots, w_n\}$, 并使用 $F_y(\cdot)$ 来表示 y 标签的预测得分。重要性得分 I_{w_i} 计算为删除单词 w_i 前后的预测变化, 正式定义如下。

$$I_{w_i} = \begin{cases} F_Y(X) - F_Y(X_{/w_i}), & F(X) = F(X_{/w_i}) = Y \\ \left(F_Y(X) - F_Y(X_{/w_i}) \right) + \left(F_{\dot{Y}}(X_{/w_i}) - F_{\dot{Y}}(X) \right), & F(X) = Y, F(X_{/w_i}) = \dot{Y}, Y \neq \dot{Y} \end{cases} \quad (1)$$

根据计算的各个词语的分数从高到低依次排序来获取词重要性序列。

3 基于标点符号插入的文本对抗攻击

3.1 方法概述

为了更好地展现基于符号插入的文本对抗攻

击的效果，分别使用该方法进行灰盒和黑盒攻击。攻击过程分成3个阶段。①词重要性排序。采用分词工具对文本进行分词，以获得词语序列。在灰盒的设置下，运用词重要性排序方法对词语进行打分和排序；而在黑盒的设置下，利用第2.3部分中计算得到的词重要性排名，对文本中的词语序列进行排序。②构建字扰动空间。挑选经过筛选的符号，在单词的前部、中部和后部分别插入，从而形成该词的扰动替换集。③搜索生成最佳对抗文本。结合①中得到的词重要性排名，使用贪心算法进行搜索得到对抗文本。图1给出了符号攻击技术框架。

3.2 构建字扰动空间

构建字扰动空间主要包括特定符号的筛选和扰动替换集的构建。研究发现，模型对标点符号存在分类偏向，如在情感分类的BiLSTM模型上，单独使用“>”和“?”的分类概率为72%和95%，被识别为消极情感，而“<”和“#”的单独分类概率为92%和95%，则被识别为积极情

感。此外，当符号叠加时，分类概率值会提升。因此，只需使用两个标点符号即可构建扰动替换集。当攻击目标是积极情感的样本时，使用“>”，而当攻击目标是消极情感的样本时，则使用“#”。

一个词的候选替换集构建主要包括3个步骤，即使用筛选的符号进行左插入、中插入和右插入。以含有两个字的词“漂亮”（为例，通过插入“#”，可以获得3种情况（“#漂亮”“漂亮#”和“漂亮#”），这3种扰动组成了“漂亮”这个词的候选替换集： $\{“\#漂亮”，“漂亮\#”，“漂亮\#”\}$ 。对于给出的待扰动词 $t = “w_0w_1 \dots w_n”$ ， t 的扰动集 $\hat{T}$ 可具体表示为以下形式。

$$\hat{T} = \text{Tran}(P, t) = \{ “Pw_0w_1 \dots w_n” , “w_0Pw_1 \dots w_n” , “w_0w_1P \dots w_n” , “w_0w_1 \dots Pw_n” , “w_0w_1 \dots w_nP” \} \quad (2)$$

其中， P 表示插入的符号， $\text{Tran}(\cdot)$ 为扰动集生成函数。

3.3 词重要性排序

在黑盒设置中，只能够获取到文本的标签，

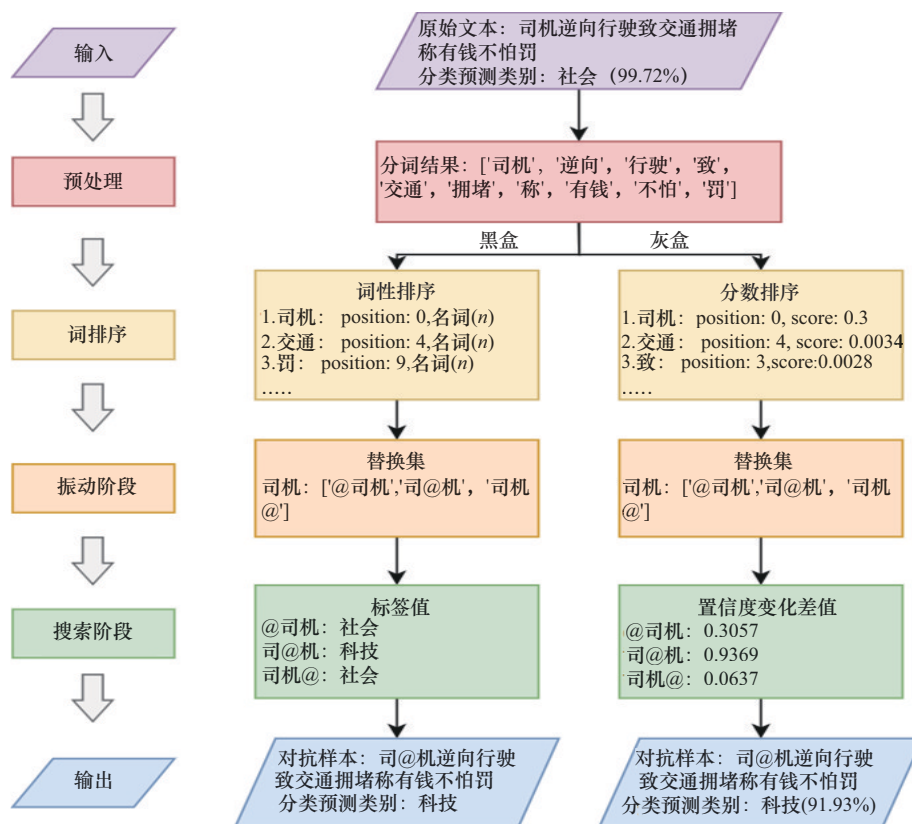


图1 符号攻击技术框架

Figure 1 Symbolic attack technology

无法使用 2.3 节中的方法计算词的重要性分数, 但可以将词性对模型预测的影响排序, 寻找在删除后使样本预测标签改变的词, 并统计这些词的词性出现的频率与它在数据集中出现^[19]频率的比值, 按照比值大小从高到低对其进行排序, 具体方法如下。

在中文分词模型中给出了 20 种词性构成集合 $L = \{l_1, l_2, \dots, l_{20}\}$, 选取 n 条样本作为统计数据集 $D = \{x_1, x_2, \dots, x_n\}$ 。首先统计每种词性在数据集中出现的数量 N_k , k 表示某种词性。

$$\text{framework} N_k = \sum_{i=1}^n \text{Count}(k, x_i) \quad (3)$$

$\text{Count}(k, x)$ 为计算 x 中词性 k 的数量, 给定一个经过切词后含有 n 个词语的样本, $x = \{w_1, w_2, \dots, w_n\}$, $x_{/w_j}$ 表示删除 w_j 后的样本, $x_{/w_j} = \{w_1, \dots, w_{j-1}, w_{j+1}, \dots, w_n\}$ 。 $f(x)$ 表示样本 x 在模型 f 上的预测标签, $N_{k'}$ 表示那些删除词性为 k 的词 w_i 后使样本标签改变的词的数量。

$$N_{k'} = \sum_{i=1}^n \text{Count}(\text{LAC}(w_j), x_i), f(x_i) \neq f(x_{i/w_j}) \quad (4)$$

$\text{LAC}(w_j)$ 为 w_j 的词性, 使用 I_k 来衡量 k 这个词性的重要性分数, 计算方法如下。

$$I_k = \frac{N_{k'}}{N_k} \quad (5)$$

通过对每种词性的重要性进行评分计算, 并按照得分从高到低的顺序进行排序, 最终得出了词性重要性的排名。词性分数计算算法见算法 1。

算法 1 词性分数计算算法

Input: 部分数据集 $D = \{x_1, x_2, \dots, x_n\}$, 目标函数 F , 词性集 $L = \{l_1, l_2, \dots, l_{20}\}$;

Output: 词性排序序列 Q

- 1) $N = \{l_1:0, l_2:0, \dots, l_{20}:0\}$;
- 2) $N' = \{l_1:0, l_2:0, \dots, l_{20}:0\}$;
- 3) for x_i in D do
- 4) $x_i = \text{lac}(x_i) = \{w_1, w_2, \dots, w_n\}$;
- 5) for w_j in x do
- 6) $k = \text{lac}(w_j)$;
- 7) $N[k]++$;
- 8) if $F(x_i) \neq F(x_{i/w_j})$ then
- 9) $N'[k]++$
- 10) end if

- 11) end for
- 12) end for
- 13) for each key li in L do
- 14) $I_k = N[li] / N'[li]$
- 15) end for
- 16) $Q = \text{Sort}(L)$ by I
- 17) return Q

3.4 灰盒攻击

作为灰盒标点符号攻击, 本文通过攻击框架实现了 Textfooler 的变体, 该框架只执行标点插入, 而不是进行词语替换。将这个标点符号变体表示为 TFP (text fooler punc)。TFP 有 3 个主要步骤。

步骤 1 词重要性排序 (2)~(6) 行)。计算词的重要性。用特殊符号 <UNK> 替代原词, 以分析单词替换时 logit 的变化。根据这些 logit 变化, 将其从高到低排序, 得到排序结果 T 。

步骤 2 构建扰动替换集 (7) 行)。为提升攻击效率, 在 BiLSTM 模型上, 使用经过筛选的字符进行攻击。在 BERT 模型上, 为了提升攻击成功率使用多字符组合攻击。

步骤 3 搜索对抗样本 (8)~(18) 行)。根据步骤 1 中 T 的词排序, 逐一查询步骤 2 生成的替换集并修改文本来搜索攻击空间。在 logit 和语义相似度评分方面保持最佳替换, 这个过程被称为使用字替换的贪心搜索。

总之, 对于示例 x , 该算法找到 T 中排名最高的词。它通过插入一个标点符号, 利用贪心算法调用受害者的模型来逐步修改, 直到文本标签改变。灰盒符号攻击算法见算法 2。

算法 2 灰盒符号攻击算法

Input: 原始样本 x , 相应的真实标签 Y , 标签 Y 下的置信度 P_Y , 目标函数 F

Output: 对抗样本 X_{adv}

- 1) $X_{\text{adv}} = X = \text{lac}(x) = \{w_1, w_2, \dots, w_n\}$
- 2) for w_i in x do
- 3) I_{w_i} via Eq.(i)
- 4) end for
- 5) Create set W via I_{w_i}
- 6) for w_j in W do
- 7) $C \leftarrow \text{Tran}(w_j)$

```

8)   for  $c_k$  in  $C$  do
9)      $X' \leftarrow c_k \rightarrow w_j$  in  $X_{adv}$ 
10)     $Y_k = F(X')$ 
11)     $P_k = F_{Y_k}(X')$  12: end for
13)    if  $Y_k \neq Y$  then
14)       $X_{adv} = X'$ 
15)      return  $X_{adv}$ 
16)    else if  $P_{Y_k}(X_{adv}) > P_k$  then
17)       $X_{adv} = X'$ 
18)    end if
19)  end for
20) end for
21) return None

```

3.5 黑盒攻击

本文提出了一种黑盒设置下的标点符号插入攻击方法——BBP (black box punc)，由于在黑盒设置中，只能获取到文本的标签，无法对词语进行打分，所以先对数据集中文本进行分析，找出哪一类词性会更影响结果。在黑盒设置下进行对抗文本生成时，优先对属于该类词性的词语进行符号插入。BBP 具体步骤如下。

步骤1 词重要性排序 (2)行)。使用词重要性计算方法，对选取的数据集中的部分数据进行计算，统计在删除后使样本预测标签改变的词的词性，并计算这些词性出现的频率与它在数据中出现频率的比值，以比值为标准进行排序。

步骤2 构建扰动替换集 (3)~4)行)。为提升攻击效率，在 BiLSTM 模型上使用经过筛选的字符进行攻击。在 BERT 模型上，为了提升攻击成功率，使用多字符组合攻击。

步骤3 搜索对抗样本 (5)-14)行)。根据步骤1中的词排序，依次通过查询经过步骤2的替换集修改样本来查找对抗文本。若在某个词替换后，替换集中的扰动都不能使样本标签改变，则随机选择一个替换词进行替换，继续对词顺序序列中的下一个词进行替换查询，直到样本标签改变。黑盒符号攻击算法见算法3。

算法3 黑盒符号攻击算法

Input 原始样本 x ，相应的真实标签 Y ，标签 Y 下的置信度 P_Y ，目标函数 F ，词性序列 Q

Output 对抗样本 X_{adv}

```

1)  $X_{adv} = X = \text{lac}(x) = \{w_1, w_2, \dots, w_n\}$ 
2)  $W = \text{sort}(X)$  via  $Q$ 
3) for  $w_j$  in  $W$  do
4)    $C \leftarrow \text{Tran}(w_j)$ 
5)   for  $c_k$  in  $C$  do
6)      $X' \leftarrow c_k \rightarrow w_j$  in  $X_{adv}$ 
7)      $Y_k = F(X')$ 
8)     if  $Y_k \neq P_k$  then
9)        $X_{adv} = X'$ 
10)      return  $X_{adv}$ 
11)    else
12)       $X_{adv} = \text{random}(X')$ 
13)    end if
14)  end for
15) else if  $P_{Y_k}(X_{adv}) > P_k$  then
16)    $X_{adv} = X'$ 
17) end if
18) end for
19) end for
20) end for
21) return None

```

4 实验

本节主要给出在两个真实的文本分类数据集上，攻击方法对 BiLSTM 和 BERT 模型的攻击效果，并对其多种攻击性能进行量化评估。

4.1 数据集

本文针对文本分类任务，分别在新闻数据集 THUCNews^[20] 及中文情感分析预料数据集 ChnSentiCorp^[21] 上进行相关实验。ChnSentiCorp 是百度收集整理中文情感分析数据集，包括对书籍、酒店和便携式笔记本的在线购物评价，平均长度为 110 个单词；共两个类别，其中 9 600 条数据作为训练集，1 200 条数据作为验证集。另外，本文提取了 THUCNews 数据集的文章标题，得到短文本数据，在现有的新浪新闻分类数据集的基础上，重新整合划分出 14 个候选分类类别（娱乐、财经、时政、彩票、体育、房产、游戏、股票、家居、教育、星座、社会、科技、时尚）。其中，752 471 条数据集作为训练集，80 000 条数据作为验证集。在此基础上，随机选择训练集以外数据中的 1 000 条

能够被模型正确分类的样本作为生成对抗文本的原始数据。

4.2 目标模型

将本文攻击方法应用于两个常见文本分类模型 BiLSTM^[22]和中文 BERT^[23]模型,使用上述两种数据集对 BiLSTM 模型进行训练和对 BERT 模型进行微调,来得到目标模型。

4.3 基线设置

1) Argot^[14]: 同音字、谐音字、字拆分、汉字互换、同义词等扰动方法。Textfooler 的词重要性计算方法依次与上述 5 种扰动方法组合,选取其中表现较好的 3 种基于字级攻击的基线。

① 拆字: 将能分解的汉字拆分成两个或多个部分,如“好”字可拆分为“女子”。

② 形近字替换: 使用汉字外形相似的字进行替换扰动,基于文献[14]得到的形近字进一步筛选了人类易于分别的。

③ 基于拼音的同谐音词替换: 使用同音、谐音和前后鼻音等发音相似的汉字进行替换,详见文献[14]。

2) CWordAttacker^[16]: 拼音符号替换。虽然 CWordAttack 展示了多种替换方法,为了使对抗文本具有较强的可读性,只选取拼音符号来替换原汉字这一种方法。拼音符号字替换即使用被替换字的拼音符号来代替这个字,如“好”字可扰动为“hao”。

3) Random: 为了与黑盒攻击采用的词重要性排序方法进行对比,本文使用 DeepWordBug-Punc^[7]中的随机排序方法对词重要性进行排序。

4) 本文方法。

① TFP: 使用灰盒攻击算法,通过攻击框架实现了 Textfooler 的变体,该框架只执行标点符号插入。

② BBP: 使用黑盒攻击算法,通过生成用于对抗样本的 1 000 条数据来作为词性重要性排序的数据集。

4.4 评估方法

本文专注于文本分类任务,通过在两个不同类型的真实数据集 THUCNews 和 ChnSentiCorp 上进行实验,从有效性和文本相似性两个角度与基线方法进行综合比较,以全面评估本文方法。

在这个背景下,主要关注以下两个方面:①关注对抗文本是否能够成功欺骗目标模型,这主要取决于有效性;②关注人类是否难以察觉对抗文本,这取决于文本相似性和流畅性。

4.5 灰盒攻击的有效性

“有效性”指生成的对抗样本是否能够欺骗目标模型。为了评估对抗样本生成方法的有效性,本文通过分析不同扰动率下模型的分分类准确率来衡量。在相同的扰动率下,准确率越低,说明该方法的有效性越高。为了进行有效性评估,本文分别在 THUCNews 和 ChnSentiCorp 数据集上使用 BiLSTM 和 BERT 模型进行实验。TFP 攻击在 THUCNews 数据集上的有效性评估如图 2 所示,在 ChnSentiCorp 数据集上的有效性评估如图 3 所示。这一系列实验旨在探讨对抗样本的有效性,以更深入地了解各种扰动率对模型性能的影响。

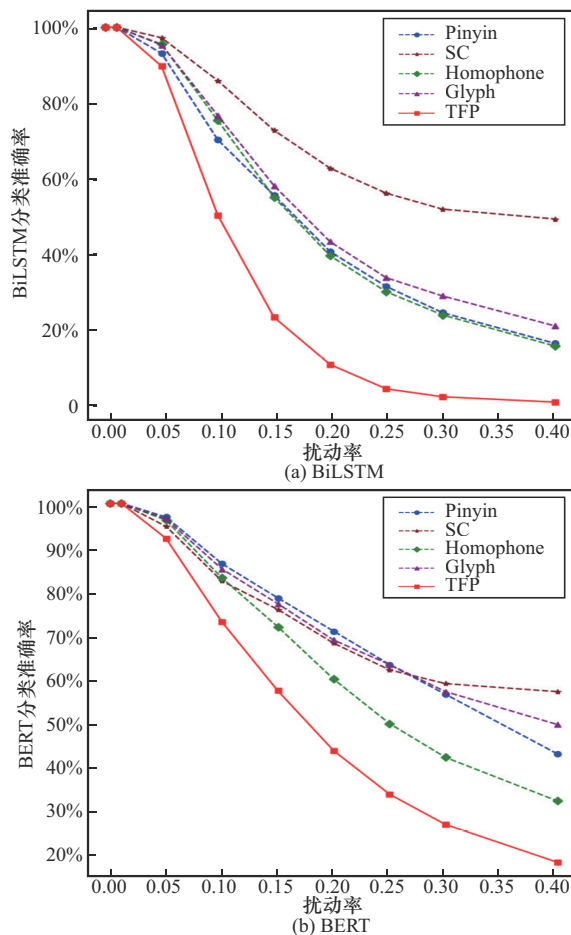


图2 TFP攻击在THUCNews数据集上的有效性评估
Figure 2 Evaluation of the effectiveness of TFP attack on the THUCNews dataset

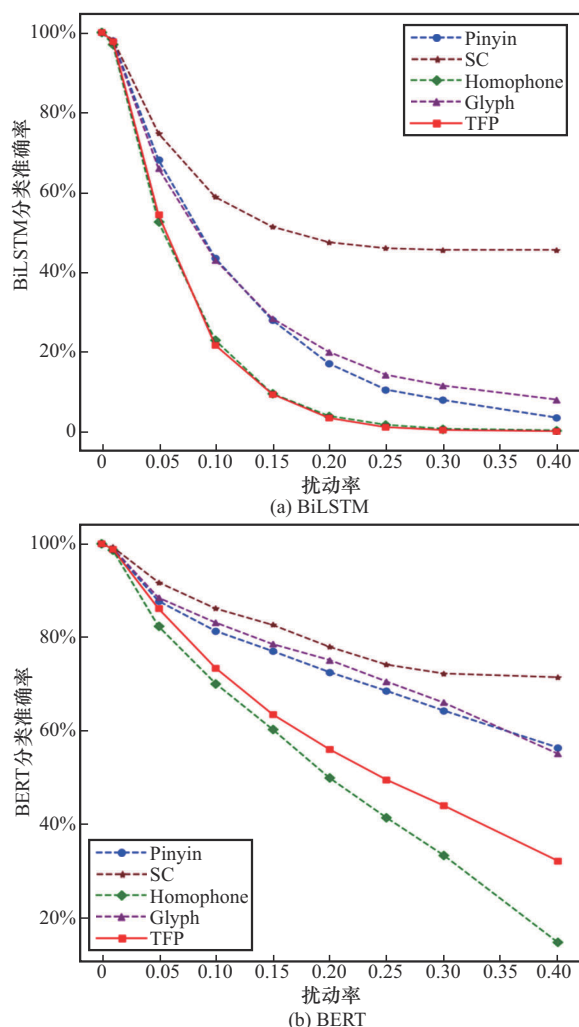


图3 TFP攻击在ChnSentiCorp数据集上的有效性评估
Figure 3 Evaluation of the effectiveness of TFP attack on the ChnSentiCorp dataset

由图2和图3可以得出,各个方法在两个不同的数据集上对模型的攻击效果表现出相似的趋势。并且,在与其他基线方法的对比中,本文提出的TFP方法显示出了卓越的攻击效果,能够显著地降低目标模型的分类型准确率。值得注意的是,Argot中基于拼音的形近字替换和拆字攻击仅适用于数据集中的一小部分字词,这一限制导致它们在攻击效果上表现一般。此外,Argot方法通过同音字替换在文本处理上取得了显著效果,但这种方法过度依赖谐音字、鼻音和舌音的拼音字替换,在一定程度上降低了文本的可读性。相比之下,本文方法更为细致,通过巧妙地插入标点符号来生成对抗样本,这种方法不直接破坏原文内容,有效维持了文本的可读性。

现有的用于生成英文对抗文本的基于标点符号的方法无法直接应用于中文。DeepWordBug-Punc基于DeepWordBug提出了一种针对英文的标点符号插入对抗文本生成方法,即将“basketball”变换为“bas.ket.ball”,然而这种方法不能直接用于中文,因为单个汉字与单个英文单词不同,难以像英文单词那样分隔进行插入操作。有部分汉字可以进行拆分,如“好”拆分为“女子”,本文选取了汉字拆分作为基线加入对比实验中,在实验中结果显示本文方法远优于汉字的拆字方法。

4.6 黑盒攻击的有效性

为了评估黑盒对抗样本生成方法的有效性,本文通过分析不同扰动率下模型的分类型准确率来衡量。在相同的扰动率下,准确率越低,说明该方法的有效性越高。为了进行有效性评估,分别在THUCNews和ChnSentiCorp数据集上使用BiLSTM和BERT模型进行实验。BBP在THUCNews数据集上的有效性评估如图4所示,而BBP在ChnSentiCorp数据集上的有效性评估如图5所示。这一系列实验旨在探讨黑盒对抗样本的有效性,以更深入地了解各种扰动率对模型性能的影响。由图4和图5可以得出,3种方法在两个不同的数据集上对模型的攻击效果表现出相似的趋势。本文提出的TFP方法显示出了卓越的攻击效果,能够显著降低目标模型的分类型准确率。同时,在相同扰动率下,本文提出的黑盒攻击方法优于随机进行排序的方法,且攻击效果接近于灰盒攻击的方法。由此,验证了本文的黑盒攻击方法的有效性。

在黑盒环境下,本文提出了一种全新的基于词性排序的算法,其效果优于采用随机的ZIP算法。ZIP是一种简单的黑盒攻击方式,以概率 p 随机扰动样本中的一个单词,然后以相同概率 p 在两个字母之间添加随机字符。在本文方法中,采用词性重要性排序方法对词进行预排序,按照重要性顺序依次对词进行扰动,同时优先选取对模型偏向性高的字符作为扰动插入,在黑盒攻击有效性验证中,证明本文方法(BBP)优于ZIP的随机方法(Random)。

4.7 相似度

文本相似度的衡量方法有多种,本文使用语义相似度生成的对抗文本与原文本的相似程度进行比

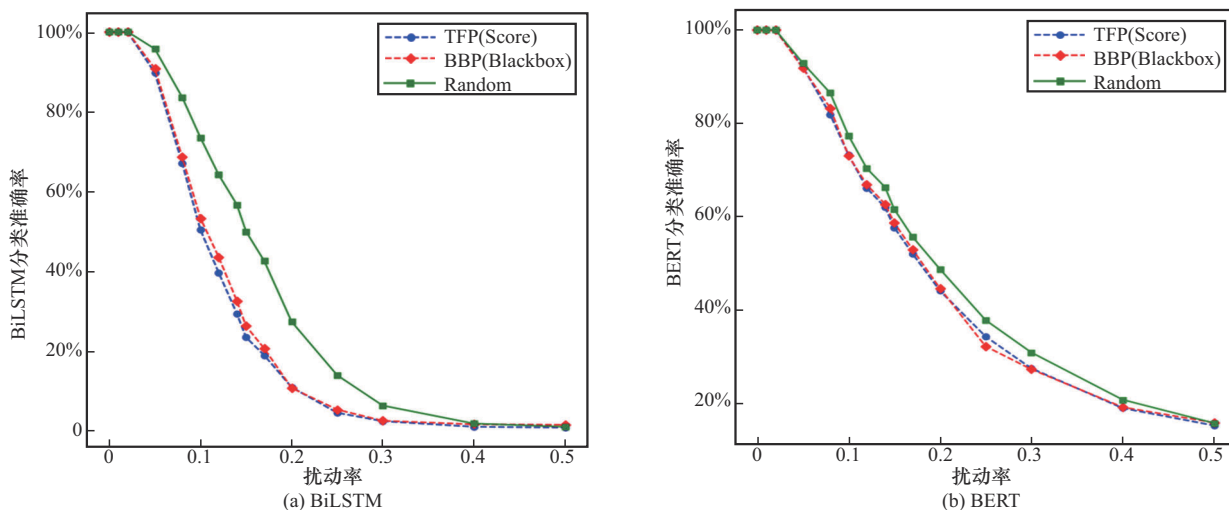


图4 BBP在THUCNews数据集上的有效性评估

Figure 4 Evaluation of the effectiveness of BBP on the THUCNews dataset

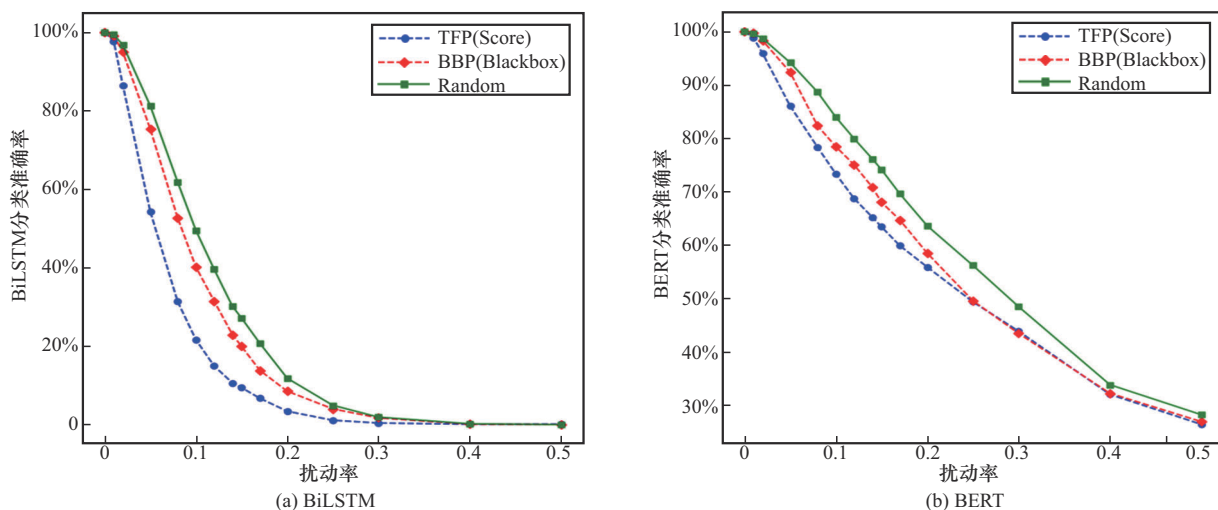


图5 BBP在ChnSentiCorp数据集上的有效性评估

Figure 5 Evaluation of the effectiveness of BBP on the ChnSentiCorp dataset

较, 语义相似度越高, 生成对抗文本的质量越高。在本文实验中, 对 THUCNews 和 ChnSentiCorp 数据集进行了评价, 利用相似度计算模型来衡量文本相似性。传统的相似度计算方法, 如词移距离和余弦相似度等, 通常仅仅将句子向量化后计算向量之间的距离, 而无法准确捕捉句子的语义信息。

为了解决这一问题, 本文进行了中文预训练的 BERT 模型微调, 并使用了 LCQMC^[24]数据集 (该数据集来源于百度知道, 包含了不同领域的中文问题匹配数据, 旨在填补中文领域大规模问题匹配数据的不足)。通过使用经过训练的相似度计算模型, 可以计算对抗样本与原始文本之间的语义相似度。

具体来说, 相似度值越接近 100%, 表示对抗

样本与原文的相似性越高。ChnSentiCorp 数据集和 THUCNews 数据集上对抗样本相似度对比分别如图 6 和图 7 所示。由图 6 和图 7 可以看出, 相比其他基线方法, 本文方法生成的对抗样本在文本相似度上远优于其他基线。上述结果展示了这种方法能够在显著降低中文模型的分准确度的同时, 很好地保持对抗文本与原始文本之间的相似度, 实现了有效性与文本相似性的双重优势。除了扰动率低, TFP 在攻击中表现得最好。这些结果展示了标点符号攻击的一个有趣特性, 即对抗样本的扰动率未必与语义相一致。因此, 可能存在一个高度扰动的样本 (就扰动词汇的百分比而言) 是可读的, 并且保留了原始信息。

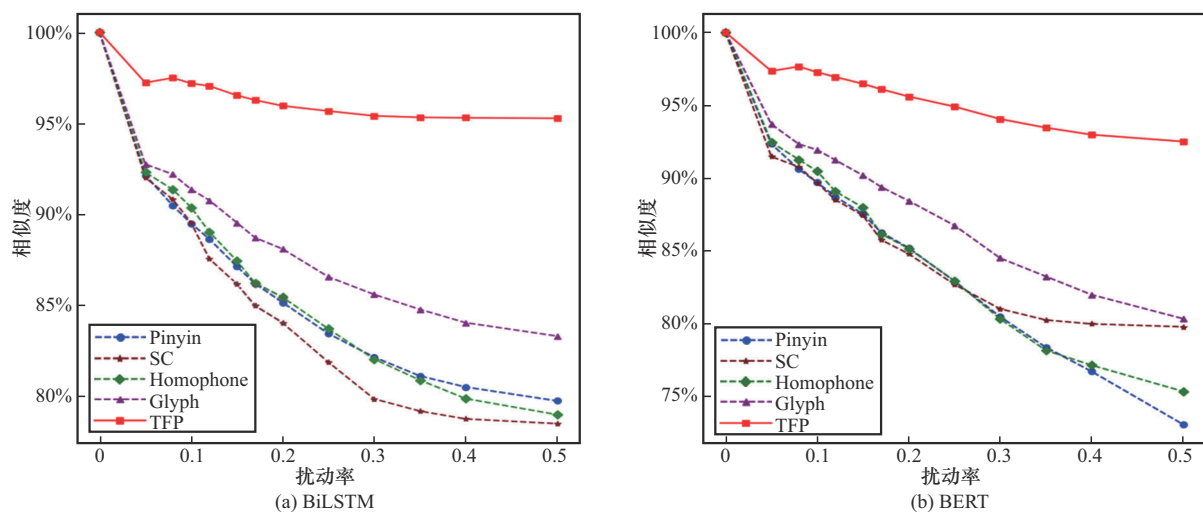


图6 ChnSentiCorp数据集上对抗样本相似度对比

Figure 6 Comparison of adversarial sample similarity on the ChnSentiCorp dataset

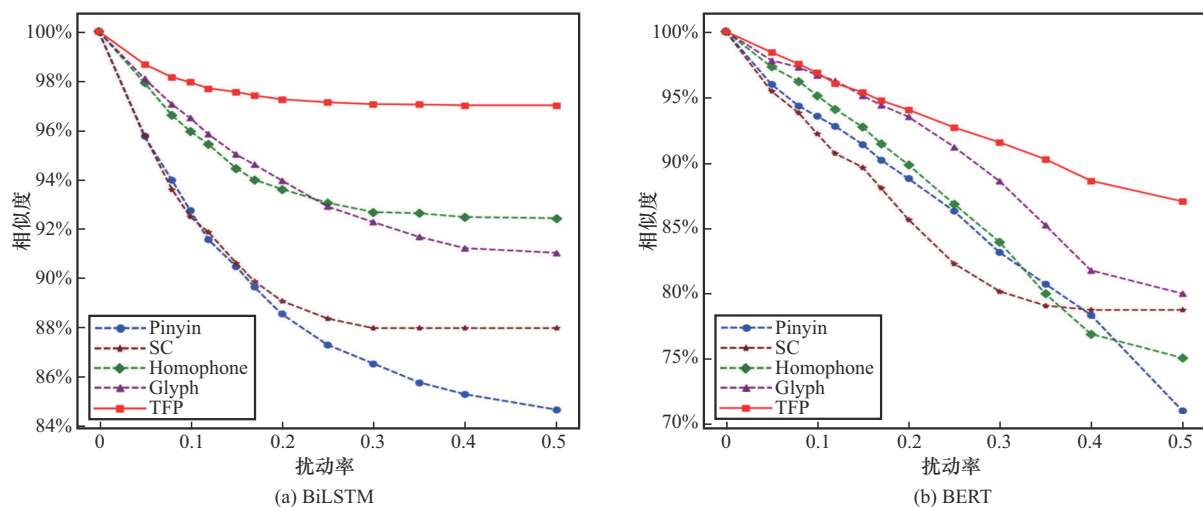


图7 THUCNews数据集上对抗样本相似度对比

Figure 7 Comparison of adversarial sample similarity on the THUCNews dataset

4.8 不同标点符号攻击的有效性

为验证攻击中选取的符号的有效性，本文选取不同的符号进行攻击。在去掉未在字典中出现的字符后剩余38个字符，表1展示了THUCNews数据集各个符号的分类概率。

在前文实验中，选取“@”符号来攻击原始标签不为6的数据，“（”符号来攻击原始标签为6的数据，其中“@”会参与到原数据标签不为6的其他13类数据的对抗样本生成中，作为主攻击符号；此外，“（”作为次攻击符号，只会参与到原数据标签为6的对抗样本生成中。为探究标点符号的置信度与攻击有效性的关系，根据表中标点符号置信度，选取了多组标点符号进行实验。表2

给出了实验设置的细节。不同符号组合的攻击效果如表3所示，其中，标点符号在其标签上的置信度的大小会影响到攻击效率，符号的置信度越高，攻击成功率就会越高，且平均扰动数和扰动率会越低。

4.9 普遍性

为了分析方法的普遍性（包括优势与局限性），本文补充了BBP方法在性能更优的中文Roberta^[25]模型上的有效性实验。BBP在ChnSentiCorp数据集上不同模型的有效性评估如图8所示。实验结果表明，本文方法在Roberta上的攻击成功率虽略低于在Bert上的表现，但仍能呈现出较好的攻击效果。

表 1 THUCNews数据集各个符号的分类概率

Table 1 The classification probability of each symbol in the THUCNews dataset

符号	标签	概率
.	6	0.795 6
,	6	0.528 9
;	4	0.378 3
=	6	0.413 5
!	10	0.288 0
(	4	0.653 7
)	4	0.300 1
[	4	0.257 8
]	0	0.252 9
‘	4	0.581 7
’	4	0.245 2
，	3	0.887 9
-	9	0.197 4
《	12	0.577 9
》	4	0.133 7
~	2	0.435 6
，	6	0.480 6
.	6	0.318 4
;	3	0.367 7
:	3	0.233 8
?	4	0.229 3
!	12	0.293 1
(	2	0.444 0
)	4	0.193 3
【	4	0.435 1
/	4	0.230 0
+	3	0.448 5
\$	13	0.345 6
#	2	0.349 5
@	6	0.985 7
&	4	0.273 3
「	6	0.572 5
」	6	0.410 8
"	4	0.133 7
...	3	0.313 4
、	4	0.315 3
"	4	0.133 7
】	4	0.403 3
:	4	0.230 0

表 2 实验设置细节

Table 2 Experimental setup details

符号选择	情况说明
@ (98%) 和 ((65%)	“@”攻击标签不为6的数据，“(”攻击标签为6的数据，所有“@”更多地参与到进攻中
@ (98%) 和# (35%)	验证次攻击符号的有效性，将更少参与进攻的“(”使用置信度更低的“#”替代
' (88%) 和# (35%)	验证主攻击符号的有效性，将更多参与进攻的“@”使用置信度更低的“”替代

表 3 不同符号组合的攻击效果

Table 3 Attack effects of different symbol combinations

符号组合	攻击成功率	平均扰动数	扰动率
@ (98%) 和 ((65%)	99.1%	2.23	11.8%
@ (98%) 和# (35%)	97.2%	2.38	12.7%
' (88%) 和# (35%)	96.6%	3.36	17.8%
. (79%) 和# (35%)	89.4%	3.81	20.3%
「 (57%) 和 ((65%)	87.3%	4.17	21.9%
》 (13%) 和# (35%)	71.0%	5.13	27.2%

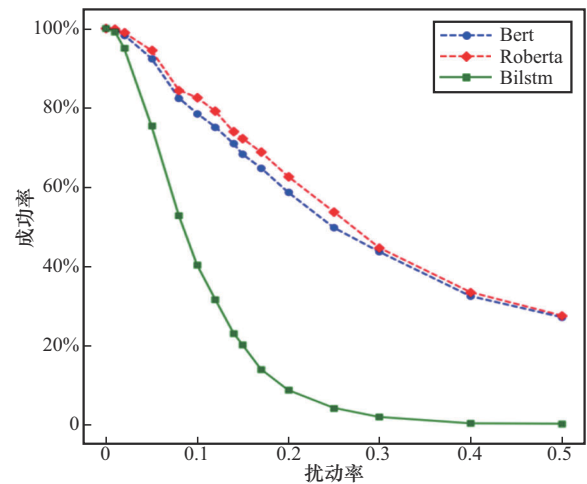


图 8 BBP 在 ChnSentiCorp 数据集上不同模型的有效性评估

Figure 8 The effectiveness evaluation of different models of BBP on the ChnSentiCorp dataset

在进行对抗样本生成时，期望选取表现最佳的模型作为目标模型，因为目标模型性能越优，越能彰显本文方法的优越性。训练数据的不平衡仅会影响目标模型的性能，而不会波及本文攻击方法的性能及表现。然而，当将生成的对抗样本加入训练集中进行对抗训练时，其是否会对模型的效果产生影响并不可知。用于防御的对抗训练

效果仍需进一步深入探索。

4.10 先进性

在对抗文本生成领域, 现有研究针对英文场景的标点符号插入方法已形成一定基础, 但中文场景下的应用仍存在显著差异与挑战。以 DeepWordBugPunc 为代表的英文标点扰动方法通过在单词字母间插入符号 (如将 “basketball” 转化为 “bas.ket.ball”) 生成对抗样本, 但该方法直接迁移至中文时面临根本性障碍: 汉字作为独立语义单元, 无法像英文单词般进行字母级拆分。尽管部分汉字可通过拆解结构 (如 “好” 拆为 “女子”) 实现类似扰动, 实验表明此类拆字方法的攻击有效性远低于本文提出的中文标点符号插入策略。进一步研究发现, 不同标点符号对模型的影响具有显著差异, 通过筛选对模型鲁棒性冲击最大的标点类型, 本文方法有效缩小了搜索空间并提升了攻击效率。

针对黑盒场景下的对抗生成, 现有方法多依赖随机搜索策略 (如 DeepWordBugPunc 的随机算法), 而本文引入词性排序作为先验知识。通过聚焦词重要性排序靠前的关键 token, 本文方法在攻击速度与成功率上均显著优于随机扰动策略, 实现了对抗样本生成过程的定向优化。在中文场景下, 现有研究 (如 CWordAttacker 和 WordChange) 虽尝试标点插入扰动, 但普遍缺乏系统性: 前者仅进行浅层探索, 后者则忽略标点类型对攻击效果的差异性影响。相较之下, 当前中文对抗文本生成领域仍以综合扰动方法为主, 如 Argot 提出的同音词替换、字形替换等 5 类策略, 以及 CWordAttacker 融合的繁体转换、拼音重写等方法。然而, 这些方法生成的对抗样本 (如 “呢囚, 蓝球, 寨事较流力口”) 往往严重破坏文本可读性, 而本文方法通过非破坏性标点插入, 在保持与原文本高相似度的前提下, 实现了相同扰动率下更高的攻击成功率, 展现出更优的实用性与隐蔽性。

总体而言, 本文通过深度解析中英文语言特性差异, 构建了适配中文语法结构的标点符号扰动框架, 同时结合词性排序机制优化黑盒攻击效率, 为中文对抗文本生成提供了兼具创新性与实用性的解决方案。

5 结束语

本文分析了汉语与英语相比的独特特点, 以及现有的基于标点符号插入的英文对抗性文本生成解决方案的局限性。基于中文的特点, 本文提出了一种基于标点符号插入的中文对抗样本生成方法, 针对两种攻击场景 (即在灰盒和黑盒设置下) 的非目标攻击。在灰盒场景下, 使用基于 Textfooler 的词语重要性排序方法配合标点符号插入来进行对抗样本的生成; 在黑盒场景下, 本文提出使用词性重要性排序方法来进行词语排序, 并与标点符号插入结合来生成对抗文本。本文评估了 TFP 和 BBP 生成文本的 3 个指标 (成功率、扰动率和相似度)。结果表明, 本文方法在相同对抗扰动率下能够产生高成功率, 同时, 生成的文本可以保持良好的可读性, 文本相似度远高于基线, 在不引起人注意的前提下, 显著降低中文模型的分类精确度, 实现了攻击效果与隐蔽性的优化平衡。在未来的研究中, 将从多个角度增强对抗文本生成技术的攻击能力, 并对更广泛的深度学习模型进行稳健性评估, 同时探索有效的防御策略。

参考文献:

- [1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.
- [2] CARLINI N, WAGNER D. Audio adversarial examples: Targeted attacks on speech-to-text[C]//Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018: 1-7.
- [3] EBRAHIMI J, RAO A, LOWD D, et al. HotFlip: white-box adversarial examples for text classification[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018. 31-36.
- [4] 郑洪浩, 郝一诺, 于洪涛, 等. 基于改进 Transformer 的社交媒体谣言检测[J]. 网络与信息安全学报, 2022, 8(4): 168-174.
ZHENG H H, HAO Y, YU H T, et al. Rumor detection in social media based on enhanced Transformer[J]. Chinese Journal of Network and Information Security, 2022, 8(4): 168-174.
- [5] LIU H Y, CAI C Y, QI Y J. Expanding scope: adapting english adversarial attacks to chinese[C]//Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing(TrustNLP). 2023: 276-286.
- [6] ZHANG Z H, LIU M X, ZHANG C, et al. Argot: generating adversarial readable chinese texts[C]//Proceedings of the 29th Int'l Joint Conference on Artificial Intelligence. 2021: 2533-2539.

- [7] FORMENTO B, FOO C.S, TUAN L.A, et al. Using punctuation as an adversarial attack on deep learning based NLP systems: an empirical study[C]//Proceedings of the Findings of the Association for Computational Linguistics: EACL. 2023: 1-34.
- [8] BEMARD E M. JONES. Exploring the role of punctuation in parsing natural text[C]//Proceedings of the 15th International Conference on Computational Linguistics. 1994: 421-425.
- [9] FORMENTO B, NG S. K, FOO C.S. Special symbol attacks on nlp systems[C]//Proceedings of International Joint Conference on Neural Networks (IJCNN). 2021: 1-8.
- [10] JIN D, JIN Z J, ZHOU J, et al. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. Menlo Park, 2020: 8018-8025.
- [11] ZANG Y, QI F C, YANG C H, et al. Word-level textual adversarial attacking as combinatorial optimization[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, Online, 2020: 6066-6080.
- [12] GAO J, LANCHANTIN J, SOFFA M L, et al. Black-box generation of adversarial text sequences to evade deep learning classifiers[C]//Proceedings of the IEEE Security and Privacy Workshops. 2018: 50-56.
- [13] 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.
WANG W Q, WANG R, WANG L N, et al. Adversarial examples generation approach for tendency classification on Chinese texts[J]. Journal of Software. 2019, 30(8): 2415-2427.
- [14] LIU M X, ZHANG Z H, ZHANG Y M, et al. Automatic generation of adversarial readable chinese texts[J]. IEEE Transactions on Dependable and Secure Computing. 2023, 20(2): 1756-1770
- [15] NOU C, CHANG G Q, GAO H C, et al. Wordchange: adversarial examples generation approach for chinese text classification[J]. IEEE Access. 2020, 20(8): 79561-79572
- [16] 全鑫, 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]. 信息安全学报, 2020, 20(9): 12-16.
TONG X, WANG L N, WANG R Z, et al. A generation method of word-level adversarial samples for chinese text classification[J]. Netinfo Security, 2020, 20(9): 12-16(in Chinese).
- [17] LI L Y, SHAO Y F, SONG D M, et al. Generating adversarial examples in Chinese texts using sentence-pieces[J]. arXiv preprint arXiv:2012.14769, 2020.
- [18] 张云婷, 叶麟, 唐浩林, 等. 基于掩码语言模型的中文 BERT 攻击方法[J]. 软件学报, 2023, 10(1): 1-18.
ZHANG Y T, YE L, TANG H L, et al. Chinese BERT attack method based on masked language model[J]. Journal of Software, 2023, 10(1): 1-18.
- [19] 袁里驰. 基于 BERT-BiLSTM-CRF 的中文分词和词性标注联合方法[J]. 小型微型计算机系统. 2023, 44(9): 1906-1911.
YUAN L C. Joint method for chinese word segmentation and part-of-speech tagging based on BERT-BiLSTM-CRE[J]. Journal of Chinese Computer Systems. 2023, 44(9): 1906-1911.
- [20] 孙茂松, 李景阳, 郭志芑, 等. THUCTC: 一个高效的中文文本分类工具包[M]. 北京: 清华大学. 2016.
SUN M S, LI J Y, GUO Z P, et al. THULAC: An efficient lexical analyzer for chinese[M]. Beijing: Tsinghua University. 2016.
- [21] JIA R, LIANG P. Adversarial examples for evaluating reading comprehension systems[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 2017: 2021-2031.
- [22] HPCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [23] DEVLIN J, CHANG M W, LEE K, TOUTANOVA K. 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, Minneapolis, Minnesota, 2019. 4171-4186.
- [24] LIU X, CHEN Q C, CHONG D, et al. LCQMC: a large-scale chinese question matching corpus[C]//Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico. 2018: 1952-1962.
- [25] LIU Y, MYLE O, NAMAN G, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv: 1907.11692, 2019.

[作者简介]



张谦 (2000-), 男, 湖北咸宁人, 深圳大学硕士生, 主要研究方向为深度学习、人工智能安全和自然语言处理。



闫巧 (1972-), 女, 广西桂林人, 深圳大学教授、博士生导师, 主要研究方向为网络安全、人工智能。