中国科学技术大学

《人工智能安全》课程论文

声明：我们证明本报告中所有非我们自己工作的材料都已注明出处

姓名：莫环欣

学号：PB22151796

学院：计算机科学与技术

日期：2025.05.20

签名：

文本对抗攻击样本生成文献调研*

以近年中文文本为例

莫环欣†
计算机科学与技术学院
中国科学技术大学
合肥 中国
wgdecade43@mail.ustc.edu.cn

摘要

本文简要回顾了对抗攻击领域从计算机视觉迁移到中文文本处理上的历程，并借助相关文献综述总结了当前领域近些年的工作。

文章首先罗列并简析了英文文本对抗攻击上的诸多成就，并以中英文文本处理区别为引，指出中文文本处理相较于英文的困难与中文文本对抗攻击的研究相对较少，随后重点关注近五年对抗攻击技术在中文文本处理方向的研究与技术发展，期间以及最后结合个人日常体验对部分对抗攻击方法进行简要讨论与评述。

CCS CONCEPTS

• 计算方法 • 人工智能 • 自然语言处理

关键词

中文样本，文本对抗攻击， 中英文区别

1 引言

2013 年，Szegedy 等人发表了第一篇基于图像数据集系统性研究对抗攻击的论文[1]，发现可以通过将一个网络的预测误差最大化来找到可以使其将图片分到错误类别中的扰动，并且这种扰动可以是极为微小且具有相当程度泛用性的（同样的扰动也可以干扰到其它网络的判断），为后续研究者与业界对模型鲁棒性的优化指明了一个可行方向，这之后图像领域的对抗攻防技术数量与质量蓬勃增长[11]。

然而由于文本数据与图片数据在连续性、敏感性、可感知性上的不同[20]，想要在不改变语义的前提下产生能稳定混淆模型判断的微小扰动会更困难一些，因此一直到 2016 在文本处理领域才有了相关跟进[18]，研究者们同样是在白盒（模型结构、参数等完全可见）场景下将对抗攻击问题转化为与梯度信息有关的优化问题，并针对文本数据在序列处理等方式上作了特殊处理后成功作用于递归神经网络上，得到了非常好的攻击效果。

之后在 2018 年 BERT（Bidirectional Encoder Representations from Transformers）预训练模型横空出世，其双向表示的编码器结构使得其对于上下文有了更好的理解，并且在多项性能测试中都远超同期其它同类预训练模型[19]，但其对于微小扰动仍旧不太鲁棒。

我在《深度学习原理与实践》课程中就曾基于预训练的 bert-base-uncased 模型预训练权重在 SST2 数据集（短句子数据集，用于微调分类任务）与 IMDB 数据集（长文本影评数据集，用于领域特征预训练）上微调过句子情感二分类任务，尽管模型主要性能指标都已达到较优值（94%+），但在未进行专门对抗训练[8]的情况下依然能够被一些简单且微小的扰动所误导，并且其自身也并未发觉（置信度与未作干扰时几乎可以认为没有区别）。

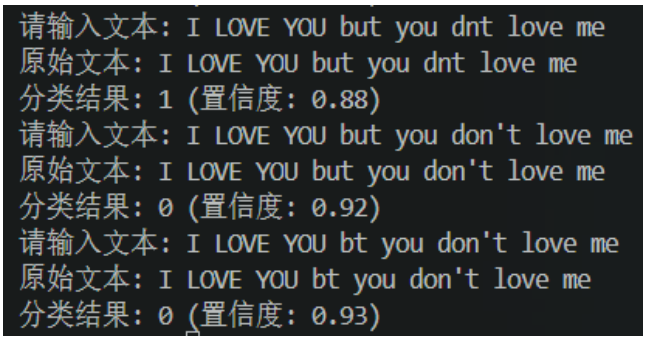

Figure 1: BERT 微调模型句子级情感二分类微小扰动测试

*A simple survey for the couese report.
†A Student of USTC.

USTC, May, 2025, He fei China

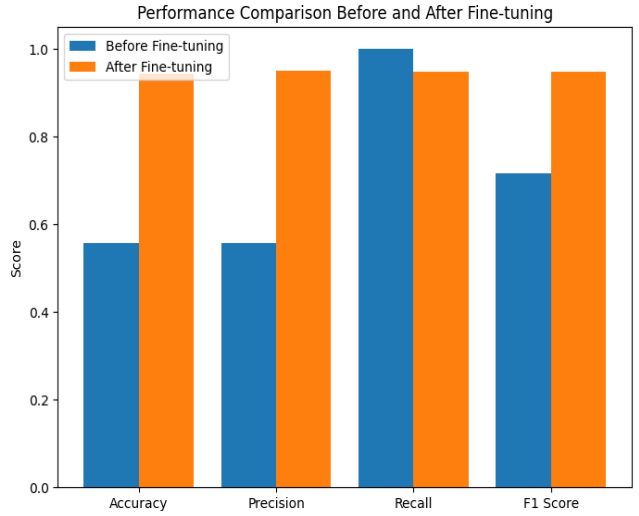

Figure 2: BERT 微调模型句子级情感二分类性能指标对比

近年来，众多研究者已经在文本相关对抗攻击上有了许多发现[5]，包括但不限于对句子/文本中的单词/字符进行增、删、改之类的直接操作，也可以对某些词语进行近义、同义词替换，但这些研究更多是在英文场景下的，由于中英文本身就在分词、词性、句法句式、指代关系等方面有着显著区别[21]，致使中文很难与英文一并处理，这部分困难同样也延续到了中文样本的对抗攻防中。

国内众多学者近年来在中文对抗样本生成上也有不少相关研究，提出了诸如使用拼音、拆分词项等方式以期生成优质的对抗样本。但现有调查多是针对于文本处理领域整体[13]或某些特殊方向[5]的，亦或是针对英文处理的，鲜少有针对中文处理方法的相关综述，对于硕士学位论文所提出方法的总结归纳更是稀少。

由于本人水平以及时间限制，同时该方向的研究相对比较少，本调查对于中文对抗样本生成部分将主要关注近几年国内公开硕士学位论文中的研究情况，从中节选出部分内容并与其它同方向研究相结合，以期为读者们提供国内当前研究情况概览。

对于本调查中所引参考文献的选取，学习借鉴了 Monserrat Vázquez-Hernández 等人在调研[5]过程中的文献选取方法，首先在 arXiv 上搜索与对抗攻击有关的综述，并根据综述中所引用的文献找到适合本主题的文章（例如有些内容是综述中没有提到但文献中有直接提到的，或直接引用高效的），同时在万方数据库中也检索类似的对抗攻防文献，以了解国内的研究情况，之后如果在撰文过程中觉得有需要引用的地方再另行搜索补充（例如前文的 BERT 与中英文区别），进而保持上下文的连贯性。

2 研究问题描述

“对抗攻击”由于出现时间相对比较短，且大众认知度不高（用户难以在业界应用上直接接触到相关概念，多作为幕后手段），目前还算是一个比较学术的概念，在百度百科与维基百科这种面向大众且相对比较权威的平台上其实并没有很明确的定义，出现比较多的地方还是在一些学术文章或专门介绍相关内容的科普文章中。其中相对比较通俗易懂一些的定义[26]可以是：通过对输入数据添加人类难以察觉的微小扰动，导致机器学习、深度学习模型输出错误结果的一类攻击方式。

与其它诸如网络安全（DDos 攻击）、系统安全（SQL 注入攻击）之类的传统计算机安全方面攻击方式的意义相似，人工智能领域的对抗攻击可以很好地检验模型的泛化能力与鲁棒性，同时出于深度学习的“训练”特性，这一领域由各种攻击方法生成出来的对抗样本还可以进一步作为对抗训练的数据强化模型应对复杂场景的能力与实际服务性能。自 2013 年计算机视觉领域的对抗攻击策略被 Szegedy 等人系统性证明有效以来，文本处理领域的研究者们也在不断跟进，并在多种文本处理任务都有了进展。

2019 年，Wei Emma Zhang 等人在他们的调研报告中基于当时已有的众多文本对抗攻击方法提出了五种[20]分类标准，之后的其他研究者又在各自具体的文本处理任务上基于此项标准提出了更为细化的分类标准，例如 Monserrat Vázquez-Hernández 等人在 2024 年的类似调研中[20]将文本情感分类任务的对抗攻击分类依据细化为了六种（引入了分析层级划分）。

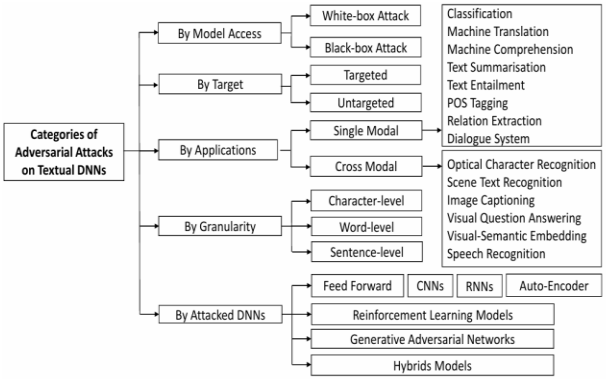

Figure 3: 文本对抗攻击分类参考[20]

基于上述分类标准，本调研报告将主要聚焦于模型参数可见性分支中的黑盒层面（模型参数与结构完全透明，仅能通过模型输入输出获知相关信息），着眼关注近几年国内硕士生所发表的与中文对抗样本生成策略相关的学位论文，下文将从中节选出几篇文章进行简述，并借助国内外相关领域类似研究作为补充以确保调研报告的连贯性、可读性与内容丰富性。

3 文献综述

3.1 英文文本对抗方法概述

计算机视觉上的图像对抗攻击成果令人瞩目，而自 2016 年文本处理对抗攻击也被证明有效后，针对语义、语序、词性、词形以及其它多种方面的英文文本对抗攻击方法如雨后春笋般涌现而出。

首先自然是 Papernot 等人里程碑式的研究[18]，他们成功将图像对抗攻击领域的 JSMA（Jacobian Saliency Map Attack）与 FGSM（Fast Gradient Sign Method）算法思想迁移到了文本对抗攻击领域，通过将循环神经网络的部分组件展开变为无环图的方法与逐步迭代从字典中将文本序列单词替换为更有误导性的单词的方法，成功解决了循环神经网络的梯度计算问题与对抗序列的生成问题，为后续研究者提供了可行思路。

紧接着 Robin Jia 等人在 2017 年的研究中基于文本对抗攻击扰动思想给出了两种新的模型理解能力评估方法[3]，通过在文本段落中添加无效内容（语法正确但上下文无关、无意义乱码）成功干扰了当时众多已有模型对于文本的理解与判断，同时通过对照组确保了这些修改后的文本不会影响到人类的正常阅读，进一步证明了相关文本对抗攻击方法的实用性并提供了评估指导。

同年四月 Bin Liang 等人则针对文本分类上的对抗攻击任务做了一定研究[2]，借助 HTPs（Hot Training Phrases）与 HSPs（Hot Sample Phrases）辅助识别文本中对模型分类结果影响最为关键的部分，在字符级别与单词级别的攻击层次上测试了上文所述的增、删、改三种攻击策略以及三者的结合策略，这些攻击措施最终在黑盒（借鉴了 Sutton 等人提出的模糊测试技术来找出 HTPs 与 HSPs，并引申出了一种能够减少生成样本数量的生成策略）与白盒（直接计算梯度以获取信息）两类不同场景下对多个深度模型（其中文本级共两种，字符级与单词级各有一种）在不同数据集下的攻击都取得了极好的效果，并且同样不会影响到人类的正常阅读理解与判断。

同年十二月 Javid Ebrahimi 等人提出的 HotFlip 方法[22]则是在白盒情况下重点关注于那些最能影响模型最终分类效果的字符，并基于梯度信息将其进行“翻转”，其因修改的幅度很少而更不容易被发觉，同时作者还证明了该方法也可用于单词上，不过也有研究指出其不适合被用于需要快速生成大量样本的场景。

之后几年的对抗攻击研究有很多，相关综述总结也有不少，例如 Wei Emma Zhang 等人在 2019 年的调研[20]中按攻击策略将二十八种攻击方法划分为七种白盒方式与六种黑盒方式，而郑海斌等人在 2021 年的调研[13]中则是按攻击细粒度将三十六种攻击方式划分为了三类。

我个人觉得 2020 年有几个比较有趣的研究，首先是四月份 Linyang Li 等人在前人找关键词、弱点词的基础上，利用 BERT 预训练模型本身对于上下文理解的优越性提出可以使用 BERT 自己来攻击自己[4]，在攻击速度变快的同时效果也变好了（相较同期策略成功率更高、扰动更小，同时还能使得生成样本具有较好的可读性）；到了九月份，Yuan Zang 等人又提出了一种优秀的攻击模型，其结合了一种替换单词的方法与新的搜索算法[7]，在文本对抗领域里成功引入了强化学习的概念，有效减少了搜索空间，提高了攻击的成功率与攻击效率。

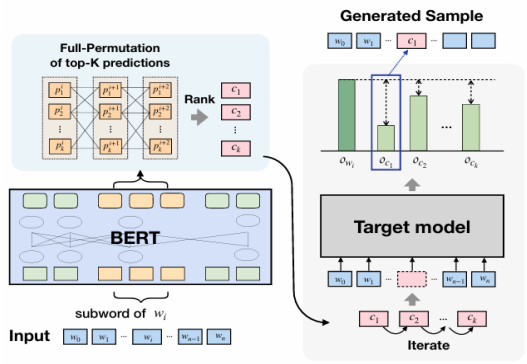

Figure 4: BERT-ATTACK 单轮攻击流程示意[4]

还有一个黑盒攻击也比较有趣且很有讨论意义，后文中的几种研究都与此相似，即 Di jin 等人于 2019 年预印、2020 年初发表的研究[6]，而他们的研究与 Bin Liang 等人在黑盒情况下的研究[2]也有点相似：算法通过获取模型对某个词删除前后的预测变化情况反馈（也就是置信度变化）来评估一个词的重要性，以此为依据对句子中的各非停用词（the、none 等）进行排序操作，并为每一个待攻击词都各自维护一个存储了 N 个与待攻击词余弦相似度最接近的词的近（同）义候选词表，随后从中选词进行替换攻击。这种攻击方法在 BERT 模型上的实验证明了其有效性。

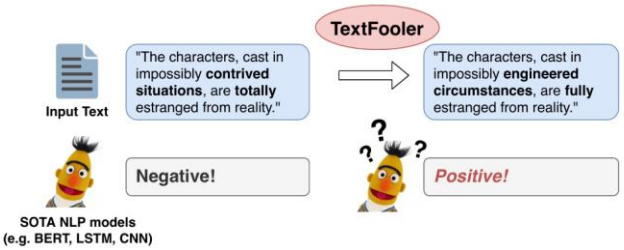

Figure 5: TextFooler 替换攻击效果展示图[6]

Di jin 等人在文章中建议将上述候选词表的 N 值设为 50 并要求余弦相似度至少大于 0.7，以在多样性与相似性之间取

得平衡，并且在实际攻击之前还需要对使用上述方法构建出的候选词表进行进一步筛查，确保其中各单词的词性全部与待攻击词的相同，且语义相似程度需要高于预设阈值，这样就得到了最终候选词表。

$$I_{w_i} = \begin{cases} F_Y(X) - F_Y(X_{\setminus w_i}), & \text{if } F(X) = F(X_{\setminus w_i}) = Y \\ (F_Y(X) - F_Y(X_{\setminus w_i})) + (F_{\bar{Y}}(X_{\setminus w_i}) - F_{\bar{Y}}(X)), & \\ \quad \text{if } F(X) = Y, F(X_{\setminus w_i}) = \bar{Y}, \text{ and } Y \neq \bar{Y}. & \end{cases}$$

Figure 6: 重要程度估计原理公式图[6]

之后只要取最相似的词进行替换即可达成攻击目的，如果最终候选词表为空则选择置信度降幅最大的词进行替换（目的是降低模型对整体的置信度，以便与后续攻击的贡献进行综合），然后按“关键词”顺序对下一个待攻击词进行处理（这意味着模型对于这个词错分类的概率最高，但是这样生成的样本质量就稍差一些），直到足够使模型的预测结果出现错误为止。

这种攻击方式在五种主流模型与两种文本任务（分类、蕴含，共七个数据集）上都取得了极高的成功率并且作出的改动幅度都很小，不过研究者在文末章节中承认其提出的方法容易受到三类因素的影响，如果词义有歧义、语法有错误或者内容对任务敏感，算法就容易出错；另外，文章中提出的方法虽然简单高效，但无法覆盖到所有情况，在少数极端情况下最终候选词表可能为空，此时就需要牺牲对抗样本的质量来换取攻击成功率，若连续出现这种情况就会大幅改变句意，从而容易被目标发觉，但其仍然是一个非常优秀的黑盒攻击算法。

3.2 中文对抗样本生成的困难

对抗攻击的思路是共通的，将英文文本上的对抗攻击方式加上中文编码并作一些适应性适配后“即可”应用到中文文本处理上，但这个“即可”中又蕴含着中文对抗样本生成的困难之处。

不同于英文分词在词性标注与词根分离上具有的便利性，中文语境下的很多字词大多无法简单通过分离偏旁部首的方式来确定含义，并且当前使用最广泛的简体字中已经去除了很多的信息，其前身也早已在演化过程中丢失了最初的象形特征。

中文在分词精细程度上的把控相对英文而言也是更为困难的，甚至于一个专业名词在经过有效拆分后依然是上下文连贯的专业名词，同时相较拆分前更利于自动化工具的检索与处理，并且这种拆分还可能同时通过多种形式进行，陈运文在其文章中[21]提到的对“中华人民共和国”分词的例子就可以很形象地说明这一点。

此外，还有学者认为困难点在于中英文语系的差别，语系区别带来的又是成分分析的差别，中文对句法的成分划分方式数目可以是英文的七倍[17]，同时中文常用词又为英文的十数倍，例如下图是对语系区别的举例：

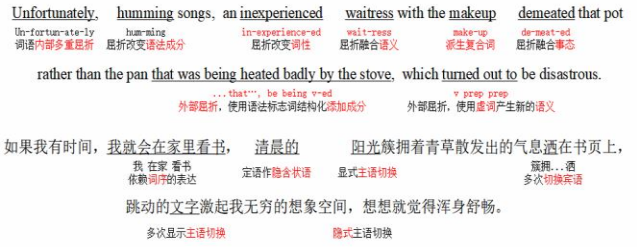

Figure 7: 屈折语（英）与孤立语（中）的区别示意图[17]

《语法六讲》[25]中曾提到过中文动、名词的区分较为困难，非常依赖于具体的语境。一个共识是，中文理解对于上下文的要求很高，一词多义、一词多用在这里十分常见，同一个词在形式相近的上下文中很有可能互为反例，甚至于同一句话也会因语气（可以体现在标点符号上）的不同而表达出多种不同的情感与态度。这一点英文中虽然也有涉及，但中文需要考虑的情况更为复杂。

考虑到诸如简体、繁体、文言、白话与方言，中文语境下包含了太多的文本形式。同时，中文互联网上每年都会出现很多的“生造词”与“旧意新解”词，例如“娄”字在某些场景下应该按上下结构拆开理解为“此女”；“蚌”、“绷”之类拼音相同但字义完全不同的字在类似“蚌埠住了”、“绷不住了”语境中实际表达的是一个意思；“资瓷”实为“支持”、“火钳刘明”实为“火前留名”；而如“yyds”（一般表夸奖）之类的“黑话”也数不胜数。这些词大多对于文本的情感倾向和语义内涵十分关键，每个词都考虑显然是不现实的、但若胡乱修改又无法保证生成样本的可读性。

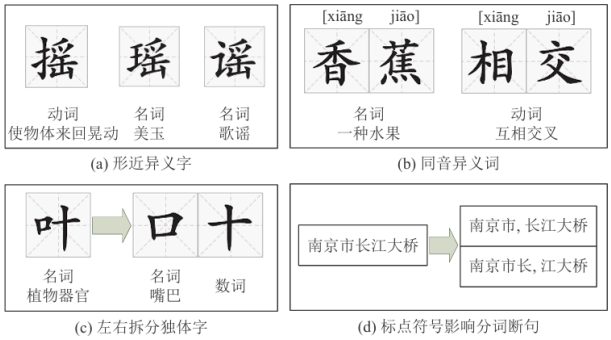

Figure 8: 中文字形、字音、拆字、分词示例[28]

而一些特殊场景下的文本也给对抗攻击算法的重要/关键/脆弱位置识别逻辑带来了挑战，例如一长串带有强烈正向情感倾向的文本最后可能带有对“三国杀”（一款卡牌游戏）的批评、“疯狂星期四 v 我 50”的文案、或者一段由“你说得对”起头的莫名其妙的学校、游戏介绍，而这串文本中最重要的就是后面这一部分，如果处理算法缺乏对于上下文的综合考虑能力，很可能会过于关注上文中的一些强烈情感表示，从而忽略上下文的转折处，并对那些无关紧要的强烈情感词识别为重要位置进行攻击，最终可能因替换位置不当而引起阅读者的察觉或完全起不到攻击效果。

3.3 近年中文样本对抗方法概览

中文处理领域的文本对抗攻击研究大多借鉴了研究众多且已有较为完善路径的英文文本对抗思想，其中有两种从英文对抗攻击上迁移而来的、较为早期的黑盒中文攻击算法在近些年的研究中时常被引用作为基准攻击方法与攻击性能参考，分别是 2019 年的 WordHanding[9] 与 2020 年的 CwordAttacker[10]。

前者在可读性上做了些研究，使用预训练模型处理中文文本输入，随后重点对汉语词语重要性的选择作出了修改，筛选出包含有效情感倾向的词并从中排除中性词与名词，最后按序对汉字使用相同拼音字进行替换攻击。不过作者在文章中并未说明同音字候选集的制作方法，并且这个策略考虑的情况相对来说还是有些简单了，针对这一点下文中有研究者做出了一些改进。

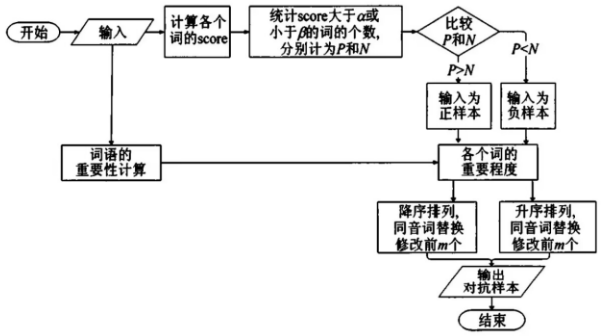

Figure 9: WordHanding 算法流程[9]

后者则更关注攻击的成功率，并为替换策略引入了繁体与拼音替换手段、给词语之间插入特殊符号以及交换词语内相邻字的顺序。这种方法的扰动幅度相对较大，不过在当时的中文样本生成领域中也算是比较具有创新性的成果，这一部分在下文中也会有较为详细的讨论。

除此之外，李相葛等人于 2023 年发表在软件学报上的研究[28]则是比较具有综合性的，他们在前人对中英文对抗攻击研究的基础上同时考虑从字音、字形、标点等多个维度入手生成对抗攻击样本，并借助 USE 模型编码（Universal Sentence Encoder）对生成样本施加语义约束，提出结合了攻击成功率、模型困惑度、对抗样本修改幅度等多个通用指标的自动化性能评估方案，最终命名为 CwordCheater，其整体结构如下图所示。不过 USE 主要是在英文语料下训练的，并且不太适合大规模与具有复杂语义的文本，接下来提到的研究就在中文编码上下了一定功夫。另外韩子屹等人同年发表的研究也与此类似[30]。

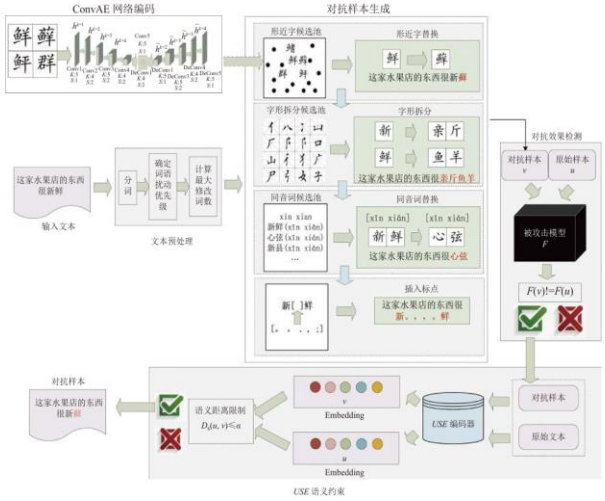

Figure10 : CwordCheater 算法整体结构图[28]

2023 年六月，蒙古科技大学的研究者弓燕在其硕士学位论文中提出了三种黑盒攻击算法[15]，其中两种依赖模型在输出中提供预测结果的置信度信息，都是在单个汉字级别执行相似字形替换策略的算法，前者主要考虑了字形的相似程度以维护可读性，后者则是在前者基础上牺牲了一部分相似性以换取更高程度的攻击成功率；最后一种则更具泛用性——仅依赖于模型输出。

对于第一种方法，作者首先将各汉字进行编码，其中又分为“音码”与“形码”：音码按拼音的声/韵母及声调三部分进行编码并排序，这里的编码均为二进制形式，只需根据声、韵母及声调的数量设置各部分二进制的位数即可，另外中间还需针对声韵母之间元音情况考虑添加补码。这里具体各部分音节的编码内容由作者直接给出，声、韵母采用格雷码，补码长度与前面两者相同且默认为全 0（若确实元音则按韵母处理并作为补码进行实际的填充），而四个声调则从 0 到 3 顺序排列，共需 17 位二进制表示；形码也分为三部分，描述了汉字的结构、形态特征与笔画数量，共需 36 位二进制表示。

完成编码后作者借助汉明距离（即两个等长字串之间在相同位置上的不同字符的个数，作者通过将其除以长度的手段抹去了等长限制）平滑了两部分编码的贡献，随后就可以计算各汉字之间的相似度得到最关键的一部分汉字加入候选列表中。

之后就是对文本中的各个字按重要程度逐个排序，从下图的重要性评估公式中就可以看到作者的这个算法思想与前文 Di Jin 等人提出的 TEXTFOOLER 算法确实有着异曲同工之妙，之后的算法思想也与 TEXTFOOLER 的类似，从最重要的字开始逐个替换直到模型被成功误导，并且攻击词的选取也是基于余弦相似度评估的，基本可以认为其是基于字形修改的 TEXTFOOLER 策略中文应用。

区别主要在于编码逻辑上（本身中英文在这一方面就有明显区别），以及作者在计算出重要分数后还对分数以及置信度变化程度进行了一个结合与归一化操作，并且当前算法中主要还是基于字形的替换攻击（而非 TEXTFOOLER 的同/近义词替换），在可读性上会稍差一些。另外也正如前提条件所述，本算法还有一个关键缺点在于需要模型提供置信度信息才能够正常工作。

$$T(x, w_i) = \begin{cases} P(y_{true} \mid x) - P(y_{true} \mid x \setminus w_i), & if\ P(x) = P(x \setminus w_i) = y_{true} \\ (P(y_{true} \mid x) - P(y_{true} \mid x \setminus w_i)) + (P(y' \mid x \setminus w_i) - P(y' \mid x)), & \\ if\ P(x) = y_{true}, P(x \setminus w_i) = y', y' \neq y_{true} & \end{cases} \quad (式\ 3.12)$$

其中 $T(x, w_i)$ 作为字词 $w_i \in x$ 对于分类结果 $P(x) = y_{true}$ 的字词显著性分数，$x \setminus w_i = w_1 \ldots w_{i-1} w_{i+1} \ldots w_n$ 表示删除字词 w_i 的句子，$P(y \mid x)$ 表示文本 x 预测为标签 y 的置信度。

Figure 11: 显著性分数估计原理公式及解释图[15]

随后的算法依然是基于置信度信息的，可以看作是前一种算法的强化，整体思想与流程相较前一种变化不大。作者在第一种算法对汉字所作编码的基础上额外引入了繁体等形式的形近字，用以扩充待攻击词的相似候选字替换列表，随后将对待攻击文本关键位置的定位逻辑从单个汉字扩展到了整个句子上，首先需要对文本中的所有句子按照重要性排一次降序，接着再按句子顺序逐个攻击，并且还给攻击方式从单一的汉字相似字形替换扩充了相邻汉字顺序交换与特殊字符插入，之后的攻击流程就与前一种相同——持续迭代直到成功误导为止。

其中后者需要使用作者特别建立的包含标点与特殊字符的候选集，而前者由于中文语义的特殊性其实也很好理解，例如下图所示是我于 2025 年四月份在学校“科乐喜剧社”社团群聊中与群友的聊天记录，可以发现中文文本在乱序情况下其实并不影响人类的正常理解，甚至于可能都注意不到这一变化，并且只要顺序变化不大，人类在阅读时更多会考虑的是句意，例如下面第二句就很容易产生误导。

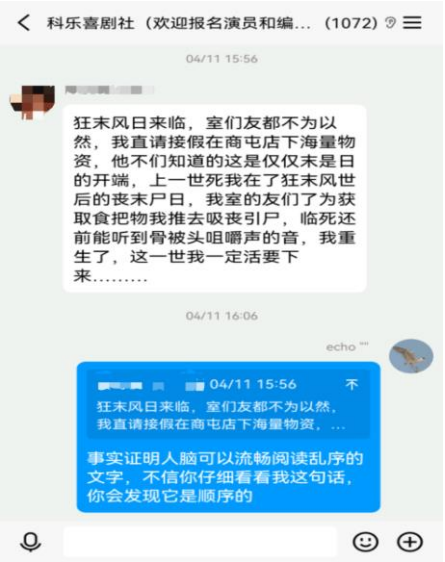

Figure 12: 乱序中文文本直观展示图

而由于这种方法相对来说比较简单，并不怎么影响语义；同时出现的时间已经比较久了，在其被证明有效后业界普遍已经将其应用于模型的训练当中，当前的众多大语言模型已经可以很好地理解其中句意，并且可以准确获知其正确语序。例如下图是豆包 AI 对这段话的归纳与部分解析，可以看到其能够完全理解文本含义与情感蕴含，甚至分析出了文本可能出现的场景与上下文。

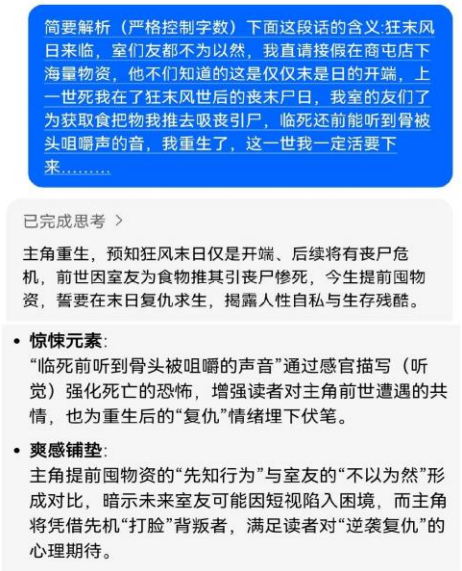

Figure 13: 豆包 AI 的解析展示图

第三种策略则在之前基础上稍微作了些修改，除了将编码依据改为汉字的字符位图外，还在扰动上除了考虑替换为同字形外还引入了同字音，并且按照词性进一步优化了汉字重要性评估方法。最后前两种方法在多个数据集上的攻击成功率平均能超越同期同类方法二十个百分点，第三种方法平均也能超越一到两个百分点。

孙嘉琪也在其学位论文中做了一些类似的研究[14]，主要工作同样是将候选词表设置为了字形与字音的结合，直接使用现有工具查找拼音、字形相近的字，针对上一小节中的“绷不住了”之类语境做了应对，并且其同年还有另一篇较为类似的文章[27]，也是在黑盒场景下基于字音、字形的对抗攻击，且依赖于模型返回置信度数据。

不同之处在于其没有特别为汉字做出编码，相对来说这份研究没有前一位研究者那么细致，主要精力放在了策略实现与后续的对抗防御上，不过替换效果同样比较优秀，在扰动率相当的情况下这种替换策略同样可以达到同期基准算法的两倍以上。

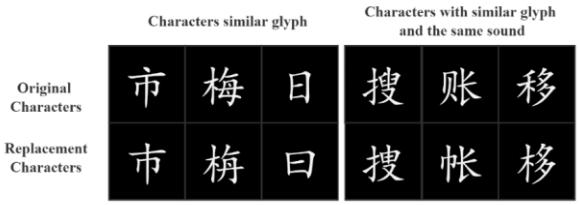

Figure 14:字形（左）与字音（右）扰动示意[14]

但是上面提到的算法与其它基于字形的攻击方式有着一些共通且较为显著的缺点，就是在一定程度上会降低文本的可读性，尤其是修改的位置比较多的时候；此外，这些研究攻击的基本都是 CNN（卷积神经网络）或者 LSTM（长短时记忆网络），专门针对中文模型的攻击相对来说比较少（就比如上面研究者攻击的也是 CNN 与双向的 LSTM）。

针对上面的两点不足，北京交通大学的研究者阮陈辉在他 2024 年的硕士学位论文中[12]，在黑盒攻击的场景与前提下，首先参考英文文本在对抗攻击领域的相关研究提出了一种使用近/同义词替换手段的攻击方法，以求在尽可能保持文本语义的前提下保持甚至达到更高的攻击成功率；随后在此基础上提出一种新方法，通过增加替换词并综合考虑各替换词之间的组合情况对可读性产生的影响，进一步提高了攻击成功率，同时还维护了算法所生成样本的高可读性，这些手段主要是针对于待攻击词的候选词表进行的。

对于作者提出的第一种方法，其整体的核心流程与 TEXTFOOLER 类似，同样是先找出关键位置，接着生成候选词集并对带攻击词进行替换操作。

不同之处在于，作者给词语的重要程度起了一个“脆弱值”的名字（从公式上看区别不大），在原有 DS（Deletion Score）与 TDS（Targeted Deletion Score）策略的基础上作出了一些改进并对应提出了两种改进策略，具体可以简述为额外考虑了删除一个词前后在非目标场景下的影响，并将其与原有影响相结合进行综合考虑。需要注意的是，这两个概念在逻辑上实际上是完全等价的，最重要的地方同样也是模型最脆弱、最容易被成功攻击的地方，只不过作者在这上面的计算方式略微进行了一些改变。

在同/近义词的生成逻辑上，作者并没有做出太多原理上的创新，而是直接使用了一个现有大型知识语言库与一个支持基于语义生成关系网络的大型词典，随后的计算公式也与上文提到的比较类似。

$$score^{(y_i)}_{decision}(s_i) = F^{(y_i)}(X_i) - F^{(y_i)}(X_0) + F^{(y_0)}(X_0) - F^{(y_0)}(X_i)$$
$$score^{(y_0)}_{decision}(s_i) = F^{(y_0)}(X_0) - F^{(y_0)}(X_i)$$

Figure 15:替换优先级依据公式（文中称为“决策值”）[12]

有了上面的信息之后，后续迭代替换时只需搜索候选的词并逐个尝试替换直到攻击成功即可，由于这个算法核心是基于语义的词替换，整体逻辑比上面一位作者的算法更贴近 TEXTFOOLER 的流程，只是对一些关键位置的指标作了简单修改，同时成功将其语义对抗攻击的流程迁移到了中文样本领域上，最终相较于中文文本攻击基准方法在多数情况下提升了 4 到 7 倍的性能。

另外，作者借助现有成熟工具以及知识库来生成同/近义词的方法很值得其他研究者进行参考，这一策略成功规避了中文样本生成方面的一部分难题，将文本处理的部分交给了更专精于此领域的专业工具，使得作者可以更专注于算法流程与策略本身，并且其基于中文预训练模型的研究也具有一定的参考意义。

在上面算法的基础上，作者进一步提出了改进措施以优化对抗样本与原文的相似程度。由于上面生成候选词表的过程是自动进行且未经筛查的，有些内容虽然近义，但存在着拉低文本可读性的隐患。

因此，作者在原有候选词的基础上引入了一个掩码语言模型，并利用其可以对句子中某个位置的内容进行预测的特点将模型的预测值也综合进行考虑，这样就为候选词表在近义的基础上补充了一些综合考虑上下文且比较连贯的文本，并且混

合后还在语义上做了进一步的筛查。值得一提的是，新增的这一部分的核心思想与 Linyang Li 等人的研究[4]较为相似，都是借助模型来生成攻击内容。

在这一算法上，作者作出的核心改进是考虑了各部分替换词之间进行组合的可能性，从而在更大程度、更大范围上维护语义之间的相似性以及上下文之间的可读性。

这一操作是通过将各待攻击位置综合考虑实现的，在搜索与应用替换词的同时就考虑组合的连贯性，并且作者指出在交错情况下（两词组合与单词组合交替）可以兼顾搜索效率（查询次数增长一倍）与攻击性能（语义相似情况与攻击的成功率），而在范围定长为两个词的情况下效果最差。由于这里引入了组合策略，最终的算法策略需要修改更多的位置并向模型查询更多次，在前一种算法的基础上又相较基准算法提高了两倍攻击成功率。

作者还发现，对候选词表的筛选如果太严格，对于语义性能的提升很小，但对攻击成功率的负面影响会很大，同时更容易过滤掉使用模型产生的候选词（因为这一部分是通过掩码-预测逻辑生成的，更多考虑的是上下文）。

纵观整个算法，个人感觉最后一步在黑盒场景下为了给攻击成功率“锦上添花”而引入的查询次数代价稍微有些大，不过前面结合知识库工具与掩码语言模型的策略还是比较不错的。

其余的研究大多也都与上文提及的各算法类似，例如研究者靳佳冀在其 2024 年的硕士学位论文[17]中借助中文的依存关系对文本按照具体成分进行划分，进而构建出语法树并对中文的分词方法进行了一些优化，同时还使用定长滑动窗口算法综合考虑了临近汉字之间的相互影响（这又与上一位研究者的组合策略思想类似），最后使用模型生成替换词并循环进行攻击。

最终作者通过实验证明了其引入特殊中文预处理策略后的算法可以出色地完成任务，并且相比其用作参考的 BERT-ATTACK[4]与 TEXTFOOLER[6]两种英文对抗算法更适合中文语境，在生成文本可读性与相似性大幅提升的同时准确率也有小幅度提升。

4 讨论和未来方向

对抗攻击的意义还是很重大的，研究者们提出的攻击算法不仅能简单地作为模型健壮性 、鲁棒性的评估手段，还能够生成对抗样本以作为模型对抗训练的基础，从而达到增强模型泛化能力与实用性的目的。

就拿中文来说，像是上面提到的一些基于字形、字音的修改，其实与早十几年的“火星文”（即读音、字形上相似但意义不同）非常类似，同时偶尔的错字在日常生活中可以说是十分常见且正常的。如果一些简单的字形扰动措施就能严重干扰一个模型判断，那么这个模型就无法面向大众平台，几乎没有实用性可言，更别提识别一些“火星文”交流了，届时只要在社交媒体上使用“火星文”交流有害信息那模型基本就无法正确识别。

另外从上文中对于中英文各自对抗攻击方法的概述中不难发现，各种对抗攻击策略实际上可以说是殊途同归，都是找到重要字/词/句，随后对其进行增/删/改/替的操作。现有英文文本对抗攻击的手段已经有了很不错的发展，而中文对抗攻击的研究者们在将对抗攻击措施进行迁移时面对的主要难点在于编码问题以及编码后的数据与算法的适配性上，由此可能还需要改进生成、搜索策略之类

原始样本	标签	对抗样本	标签
房间巨小, 电视成了摆设, 开不了, 服务员态度冷漠, 不睬我, 以后不会再进这家酒店。	消极	房间剧 xiao, 电视成了摆设, 开丕了, 服务员态度冷莫, bu 睬我, 以后布绘再进这家酒店。	积极
东西非常实惠, 快递真给力, 昨天下的单今天就到了, 新包装颜色鲜艳, 好评。	积极	东西非常湿蕙, 快递直铪厉, 昨天下的单今天就到了, 新包装颜色鲜燕,女子評。	消极
用完特别容易痒, 跟以前屈臣氏的根本不一样, 绝对的假货! 贪小便宜吃大亏!	消极	用完特别容易养, 跟以前屈臣氏的很奔步一样, 绝对的嘏货!谈小便宜吃大兮!	积极
酒店脚摩很有特色, 住客还给打折, 房间装修尚可, 位置稍偏。	积极	酒店脚摩恨油牛 寺涩, 住客还给 Da 浙, 房间装修伤渴, 位置稍偏。	消极

Figure 16:字音、字形、拼音修改结合策略示例 [31]

总结上面提到的多种算法，我们很容易发现其中增、删、改的扰动方式对于文本含义的影响很大，在一定程度上会降低（或者说显著降低）生成文本的可读性，不过其中有部分基于“改”的方法相对来说还是比较不错的，比如我们可以将左右结构的汉字拆开为两个字，这种做法在可读性上并不会造成太大影响，尤其是字号较小时，网络上比较常见的类似操作有很多，比如可以在回复他人时使用“彳亍”替代“行”字。

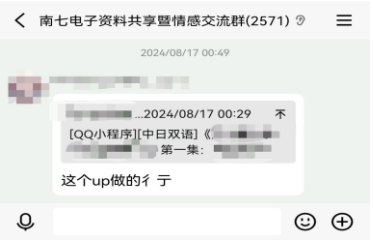

Figure 17:“行”与“彳亍”

至于基于“替”的方法，虽然在一定程度上维护了修改位置新内容与旧内容的相似性，但也由于这些策略大多只考虑新

旧内容相似性而忽视上下文，也存在着降低上下文连贯性的可能；一些学者在此基础上引入了字词组合，同时考虑多个词的组合替换，但上下文连贯性提升的代价却是倍增的查询开销。

另外，一些方法中还会借助于大型知识库或大语言模型来生成候选词表，极大地减轻了攻击负担，并且一些研究者[29]还会使用 GAN（Generative Adversarial Network）来在半白盒与黑盒场景下生成攻击样本，虽然开销稍大，且目前关于 GAN 的研究热度已经降下，但一样是一种值得学习的可行思路。

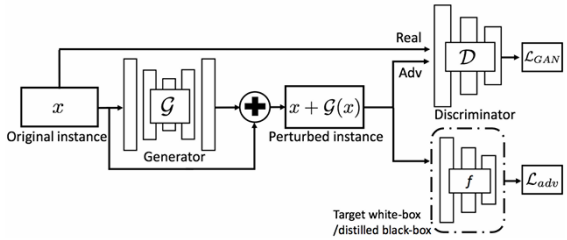

Figure 18:使用 GAN 生成对抗样本大致流程[29]

对于实际应用，上面提到的诸多文本对抗攻击手段其实在 prompt 工程中也是极为有效的，比如对于使用 GPT-SoViTs（一种简单高效的开源模型）之类文本生成语音模型的平台，直接输入敏感文本很容易被识别并拦截，而如果让文本语义变得混乱但保持音节相同则很有可能成功，这一点我曾在“番茄免费小说”平台里见到过（部分小说作者将敏感文本使用上面提到的一些方式呈现，成功规避了 AI 审核，并且 AI 听书功能的语音阅读向读者传达了作者真正想表达的含义），在这方面一个较为直观的例子是上文中提到的“火钳刘明”。

另外，上文中提到一些研究者在黑盒场景下逐步替换文本的重要信息，以降低模型对文本分类任务的置信度从而混淆分类结果，这种思想同样也可以用于其它类似领域，例如可以在 prompt 中引导模型逐步修改返回内容中的部分信息，最终得到诸如血腥暴力、地域歧视之类带有错误倾向或不良信息的敏感内容。

例如下面是利用最新版豆包 AI 生成的图片（仅作为示例，点到即止，不含任何情感倾向），利用上文的文本语义替换方法对模型进行了攻击（省略了迭代降低置信度的过程），接下来如果再对图中人物肤色之类的微小内容进行修改（实测是可以进行的）显然就能得到带有歧视含义的敏感内容——这些内容理应被模型拒绝生成，但是对 prompt 文本中部分重要位置的修改在不改变具体含义的前提下使得对抗样本成功欺骗并绕过了平台的检测模型，生成了带有敏感信息的返回结果。

Figure 19:对豆包 AI 进行提示词文本对抗攻击示例

5 结论

英文文本对抗攻击相对中文领域来说已经发展得较为成熟，常见方法一般是调整句子或文本关键位置的字词或字符，这其中又可以是添加无效内容、翻转字符、替换近/同义词等，并且部分如生成候选词表之类的工作还可以交给语言模型完成。

中文对抗样本生成的困难主要在于分词精细程度、语法成分分析、动名词区分、文本形式与特殊语境多样等，且很多词的具体含义十分依赖于文本中的上下文关系。

在中文文本对抗攻击场景下，现有白盒层面的对抗攻击算法相对较少，一部分原因其是在对抗攻击层面上不太具有现实意义，日常中基本不可能获取到模型的梯度等信息，更多时候连置信度信息都很难获取到，因此其一般更多地被用于对抗训练当中。

而黑盒攻击算法按是否需要模型在提供预测结果的同时返回置信度信息可以分为两种，普遍做法是按重要性对待攻击位置排序，这种排序可以是在句子内按字排序、按词排序，随后就可以使用迭代方法对待攻击位置进行诸如部分替换、插入、删除、部分增加、拆字之类的攻击，或者也可以维护近/同义候选词表随后从中选词攻击，直到模型预测结果符合预期为止。

在这个流程上，又可以先对文本中的句子按重要程度排序，再对句子内的单字/词排序与攻击；同时考虑多个词的组合情况，可以是定长、不定长或交错的；候选词表可以由查询知识库工具得来，也可以使用掩码预训练模型生成，还可以由

攻击者专门维护词表（不太实用），当然也可以综合使用各方法，取长补短。

不过针对这个流程，当前算法在候选词表上的搜索效率普遍不高，尤其是同时考虑多个重要位置的组合情况时向模型的查询开销倍增，后续研究者可以尝试从候选词表的生成策略、组合逻辑或搜索策略上入手进行优化，同时还可以优化重要位置的评估逻辑以从源头上减少候选组合的数目，最终实现语义相似、上下文连贯的对抗文本生成，以提高对抗攻击及其它相关领域算法的性能。

ACKNOWLEDGMENTS

感谢阮老师与三位助教老师一学期的教导，实验课与日常理论课的教学十分详细（并且答疑过程非常愉快），以通俗易懂的讲述帮助我从无到有建立起了人工智能安全相关知识思维链，尤其经过考核一后对于一些经典且常见的对抗攻击、对抗防御与后门攻击方式方法有了更深的理解，老师们实事求是的精神与传授给我的知识将深刻影响我今后的学习与工作。

参考文献

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Robert Fergus. 2013. Intriguing properties of neural networks. arXiv:1312.6199. Retrieved from https://arxiv.org/abs/1312.6199

[2] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, Wenchang Shi. 2017. Deep Text Classification Can be Fooled. arXiv:1704.08006. Retrieved from https://arxiv.org/abs/1704.08006

[3] Robin Jia, Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. arXiv:1707.07328. Retrieved from https://arxiv.org/abs/1707.07328

[4] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, Xipeng Qiu. 2020. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. arXiv:2004.09984. Retrieved from https://arxiv.org/abs/2004.09984

[5] Monserrat Vázquez-Hernández, Luis Alberto Morales-Rosales, Ignacio Algredo-Badillo, Sofía Isabel Fernández-Gregorio, Héctor Rodríguez-Rangel, María-Luisa Córdoba-Tlaxcalteco. 2024. A Survey of Adversarial Attacks: An Open Issue for Deep Learning Sentiment Analysis Models. *Applied Sciences* 14(11), 4614. https://doi.org/10.3390/app14114614

[6] Di Jin, Zhijing Jin, Joey Tianyi Zhou, Peter Szolovits. 2019. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. arXiv:1907.11932. Retrieved from https://arxiv.org/abs/1907.11932

[7] Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, Maosong Sun. 2020. Word-level Textual Adversarial Attacking as Combinatorial Optimization. arXiv:2009.09192. Retrieved from https://arxiv.org/abs/2009.09192

[8] Akbar Karimi, Leonardo Rossi, Andrea Prati. 2021. Adversarial Training for Aspect-Based Sentiment Analysis with BERT. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)* (Milan, Italy, January 10–15, 2021). https://doi.org/10.1109/ICPR48806.2021.9412167

[9] 王文琦,汪润,王丽娜,等. 2019. 面向中文文本倾向性分类的对抗样本生成方法. 软件学报. 30(8):2415-2427. DOI:10.13328/j.cnki.jos.005765.

[10] 仝鑫,王罗娜,王润正,等.2020. 面向中文文本分类的词级对抗样本生成方法. 信息网络安全. 12-16. DOI:10.3969/j.issn.1671-1122.2020.09.003.

[11] 王文萱,汪成磊,齐慧慧,等. 2025. 面向深度模型的对抗攻击与对抗防御技术综述. 信号处理. 41(2):198-223. DOI:10.12466/xhcl.2025.02.002.

[12] 阮陈辉. 2024. 高语义相似的中文对抗样本生成方法研究. 硕士学位论文. 北京交通大学(BJTU).

[13] 郑海斌,陈晋音,章燕,等. 2021. 面向自然语言处理的对抗攻防与鲁棒性分析综述. 计算机研究与发展. 58(8):1727-1750. DOI:10.7544/issn1000-1239.2021.20210304.

[14] 孙嘉琪.2023. 针对中文文本的对抗样本生成与防御研究.硕士学位论文. 天津理工大学(TUT).

[15] 弓燕. 2023. 针对中文文本分类的对抗攻击技术研究.硕士学位论文. 蒙古科技大学(IMUST)

[16] 吴超飞. 2023. 中文文本检测与对抗技术研究.硕士学位论文. 浙江工业大学(ZJUT)

[17] 靳佳冀. 2024. 中文文本情感分析任务的对抗攻击技术研究与应用.硕士学位论文. 电子科技大学(UESTC)

[18] Nicolas Papernot, Patrick McDaniel, Ananthram Swami, Richard Harang. 2016. Crafting Adversarial Input Sequences for Recurrent Neural Networks. arXiv:1604.08275. Retrieved from https://arxiv.org/abs/1604.08275

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805. Retrieved from https://arxiv.org/abs/1810.04805

[20] Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, Chenliang Li. 2019. Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey. arXiv: 1901.06796. Retrieved from https://arxiv.org/abs/1901.06796

[21] 陈运文.2019.微信公众号:一文详解中英文在 NLP 上的 10 大差异点. 来自 : https://mp.weixin.qq.com/s?__biz=Mzg4NDQwNTI0OQ==&mid=2247523303&idx=3&sn=9ed8aa10218bd57ad0a81b416cd87e75&source=41#wechat_redirect

[22] Javid Ebrahimi, Anyi Rao, Daniel Lowd, Dejing Dou. 2017. HotFlip: White-Box Adversarial Examples for Text Classification. arXiv:1712.06751. Retrieved from: https://arxiv.org/abs/1712.06751

[23] 李进锋. 2020. 面向自然语言处理系统的对抗攻击与防御研究. 硕士学位论文. 浙江大学(ZJU)

[24] 裴歌. 2020. 针对中文文本分类的对抗样本生成方法. 硕士学位论文. 西安电子科技大学(XDU)

[25] 沈家煊. 2011. 语法六讲, 第三讲: 为什么说汉语的动词也是名词. 商务印书馆.

[26] 怪友. 2020. CV|| 对抗攻击领域综述（adversarial attack）. 来自：https://zhuanlan.zhihu.com/p/104532285

[27] 王春东,孙嘉琪,杨文军. 2023. 基于矫正理解的中文文本对抗样本生成方法. 计算机工程. 49(2):37-45. DOI:10.19678/j.issn.1000-3428.0065762.

[28] 李相葛,罗红,孙岩. 2023. 基于汉语特征的中文对抗样本生成方法. 软件学报. 34(11):5143-5161. DOI:10.13328/j.cnki.jos.006744.

[29] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, Dawn. 2018. Generating adversarial examples with adversarial networks. IJCAI'18. AAAI Press, 3905-3911. Retrieved from: https://dl.acm.org/doi/10.5555/3304222.3304312

[30] 韩子屹,王巍,玄世昌. 2023. 多约束引导的中文对抗样本生成. 中文信息学报. 37(2):41-52. DOI:10.3969/j.issn.1003-0077.2023.02.004.

[31] 徐恩惠. 2022. 针对中文文本分类的对抗样本生成及防御技术研究. 硕士学位论文. 内蒙古科技大学（IMUST）