Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

单词级文本对抗攻击

硕士研究生 程瑶

2023年05月28日

内容提要

- 背景简介
- 基础概念
  - 对抗样本
  - 文本对抗攻击
  - 开源工具TextAttack、 OpenAttack
- 算法原理
  - Word-Level Attack (WLTAACO)
  - CLARE
- 应用总结
- 参考文献

背景简介

- 预期收获
  - 1. 了解对抗攻击的背景和基本原理
  - 2. 理解文本对抗攻击的方法和难点
  - 3. 理解文本对抗攻击的应用
  - 4. 了解文本对抗攻击的前沿发展

背景简介　文本对抗、问题导入

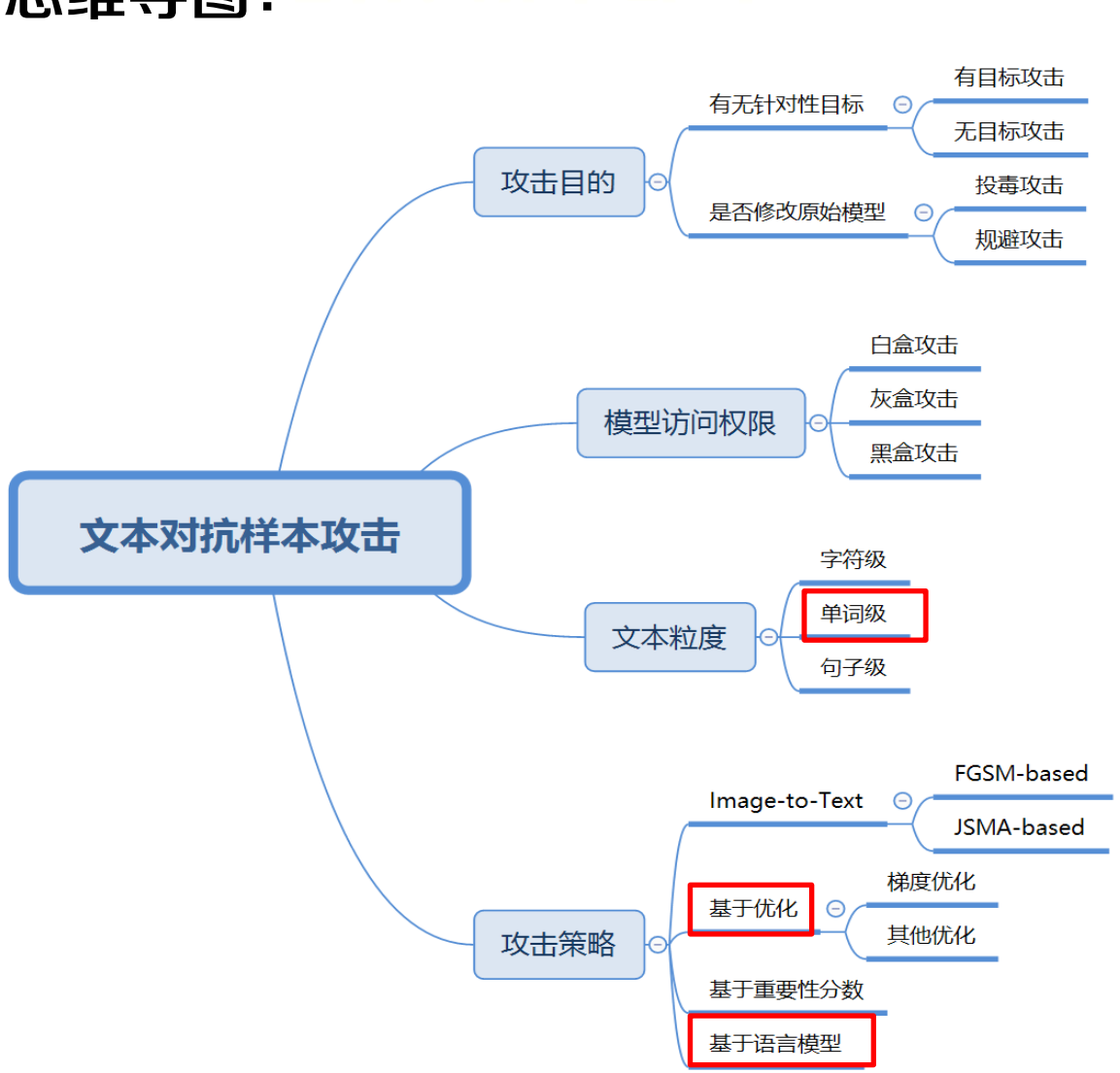
- 智能系统中的便利、漏洞和缺陷
  - 案例1：敏感词屏蔽
    - 社交网络中存在辱骂等敏感词汇，需要使用语言模型定位并“和谐”敏感词，维护网络秩序
    - 使用文本对抗技术可以欺骗语言模型，使其错误决策，以此绕过敏感词检测，不影响辱骂性质
  - 案例2：电子邮件中的漏洞
    - 垃圾邮件制造者在一封邮件里隐藏了很多复合附件，可Gmail只会显示最后一个附件，以此获得可靠域，逃过检测

阿里云-天池-安全AI挑战者计划
第三期 - 文本分类对抗攻击

- 原始文本："配[redacted]音乐，难听死了"
  - 模型预测：辱骂；真实标签：辱骂
- 对抗文本："配你ma的音乐，难听4了"
  - 模型预测：正常；真实标签：辱骂
- 真实标签的判断逻辑：辱骂文本需要具备攻击性，且能够通过字面快速辨识。
  - 反例："长亭外，古道边，芳草天"
    - 隐喻讥讽，不能够通过字面快速辨识。
    - 真实标签：正常
  - 反例："人是人[redacted]生的，妖是[redacted]生的"
    - 带有辱骂中常见词汇，但实际语义不具备攻击性。
    - 真实标签：正常

背景简介 对抗样本攻击

- 思维导图：



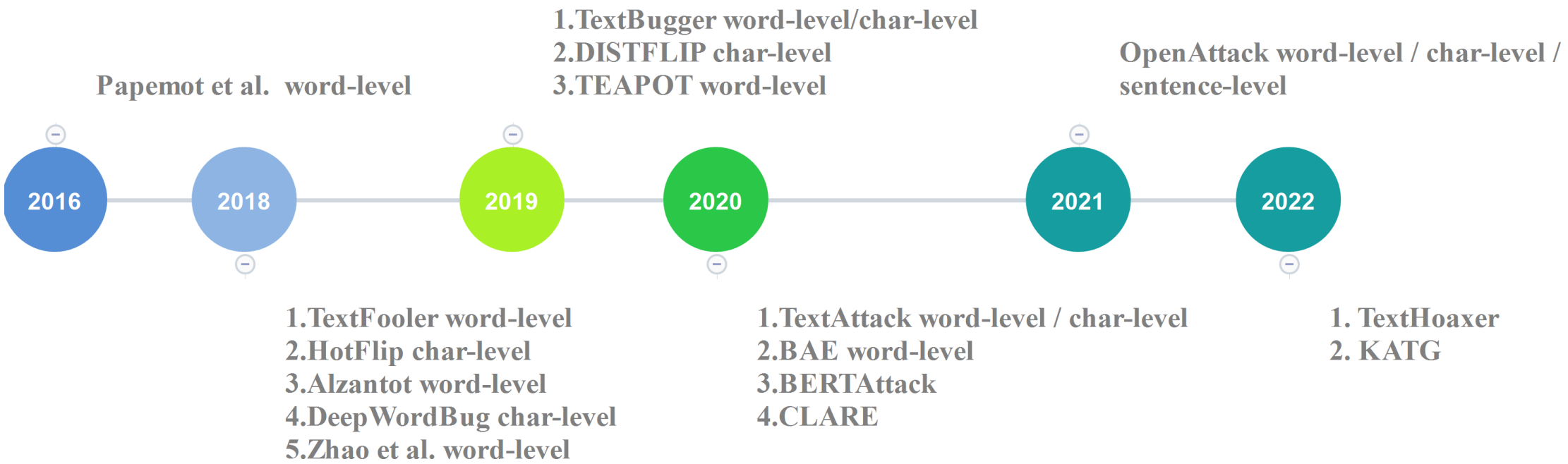
背景简介　对抗样本攻击

- 对抗样本攻击：文本 or 图像
  - 差异
    - 由于图像是近似连续的数据（图像像素值是0-255的整数值），但文本是离散数据
    - 图像扰动：对像素值添加微小改变就可以造成图像的扰动，并且很难被人眼察觉
    - 文本扰动：小扰动易被察觉，人类能猜出来原本表达的意义
  - 实例
    - 假设有一个one-hot编码为（00001）表示的是“道”字，那么改变一位的编码（00011）表示的字和“道”字不存在连续性

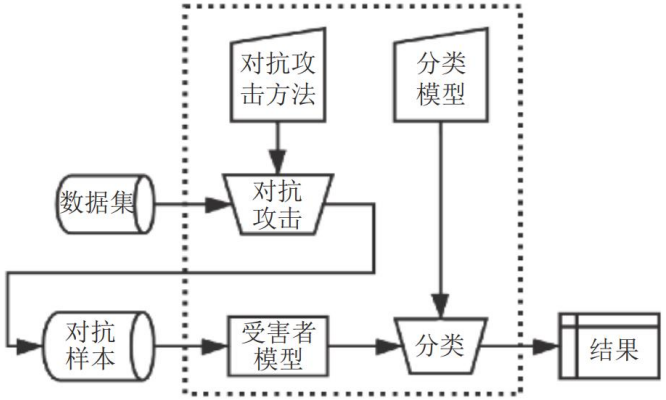
离散VS连续
易感知VS不易感知
富有语义VS无语义

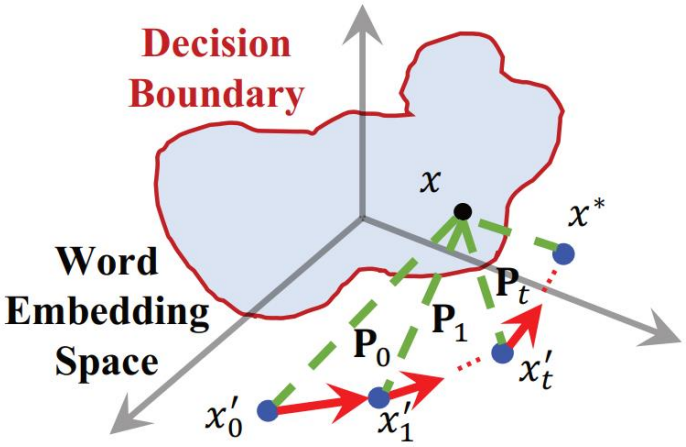
  - 参考学术报告（图像对抗攻击）
    - 组合对抗攻击的自动化搜索方法-关迎丹-2021.05.06
    - 特定安全攻防场景中的对抗样本生成方法-张荣倩-2021.07.26
    - 深度神经网络后门攻击-韩飞-2021.08.15

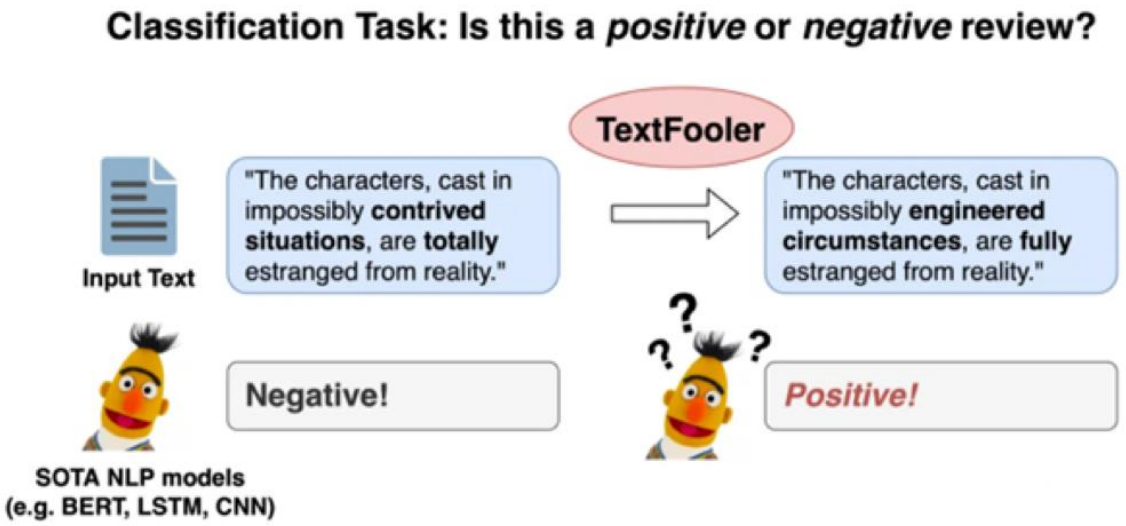
- 文本对抗攻击


Papemot et al. word-level
1.TextBugger word-level/char-level
2.DISTFLIP char-level
3.TEAPOT word-level
OpenAttack word-level / char-level / sentence-level
2016
2018
2019
2020
2021
2022
1.TextFooler word-level
2.HotFlip char-level
3.Alzantot word-level
4.DeepWordBug char-level
5.Zhao et al. word-level
1.TextAttack word-level / char-level
2.BAE word-level
3.BERTAttack
4.CLARE
1. TextHoaxer
2. KATG

- 文本对抗攻击（Text Adversarial Attack, TAA）
  - 是指对文本进行修改，使得文本的语义、可信度、真实性等方面受到影响
  - 攻击目标：对原始数据进行微小扰动使得预测错误结果最大化
  - 条件：
    - 保留语义，使其不影响人类理解
    - 在误导目标模型的同时不易让人察觉







基础概念

- 对抗样本
  - 以受害目标模型f为中心，假设f是一个文本分类器，仅允许访问f的输出（黑盒）
  - 给定一个输入序列$x = x_1 x_2 \ldots x_n$和标签y

$$f(x) = y$$

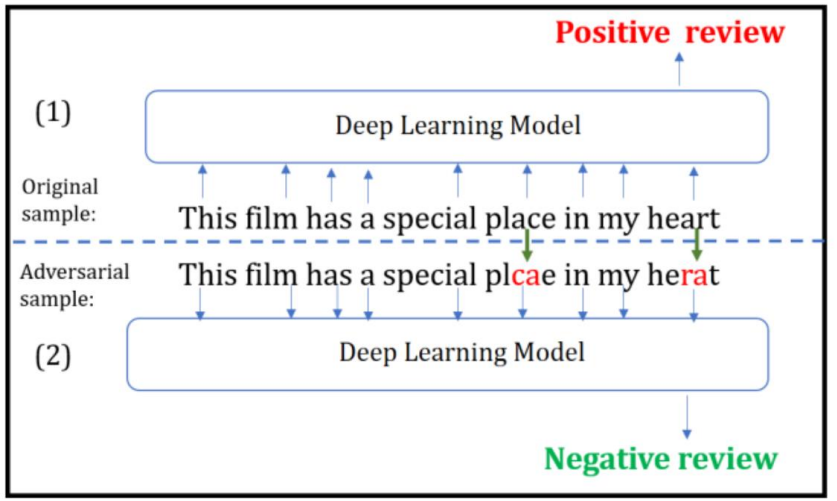
    - 修正原始文本序列x，添加微小扰动ε使得

$$f(x + \varepsilon) \neq y$$

    - 扰动条件：文本扰动最小，保留自然性和流畅性，人类对$x + \varepsilon$的预测结果保持不变
    - 数学表示：

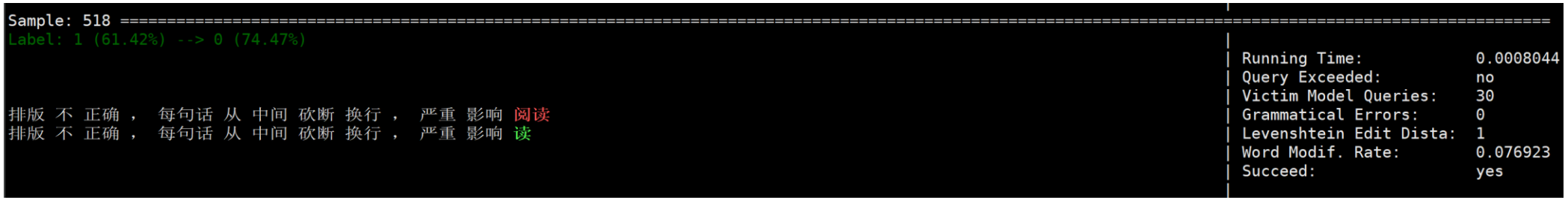
$$sim(x, x + \varepsilon) > \ell$$

    - 上式使用神经网络对句子进行编码，并计算句子对$(x, x + \varepsilon)$的余弦相似性

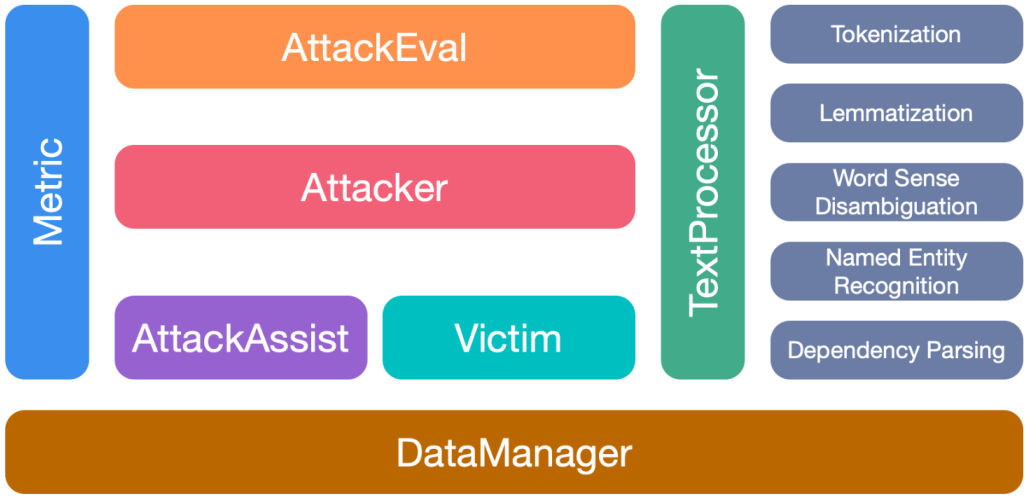


基础概念

- 攻击级别（文本粒度）：
  - 字符级攻击（char-level）：修改文本中的几个字符
    - 虽然ASR高，但拼写错误易被检测（拼写检查器）
  - 词级攻击（word-level）：对整个单词进行操作，而不是单词中的几个字符
    - 插入、删除和改变位置
    - 根据选择被操作词的方式：基于梯度的攻击、基于重要性的攻击、其他攻击
  - 句子级攻击（sentence-level）：更灵活
    - 修改后的句子可以插入文本的开头、中间或结尾（基于语义和语法正确的情况下）
    - 在文本分类中，这种攻击比其他两类攻击少得多



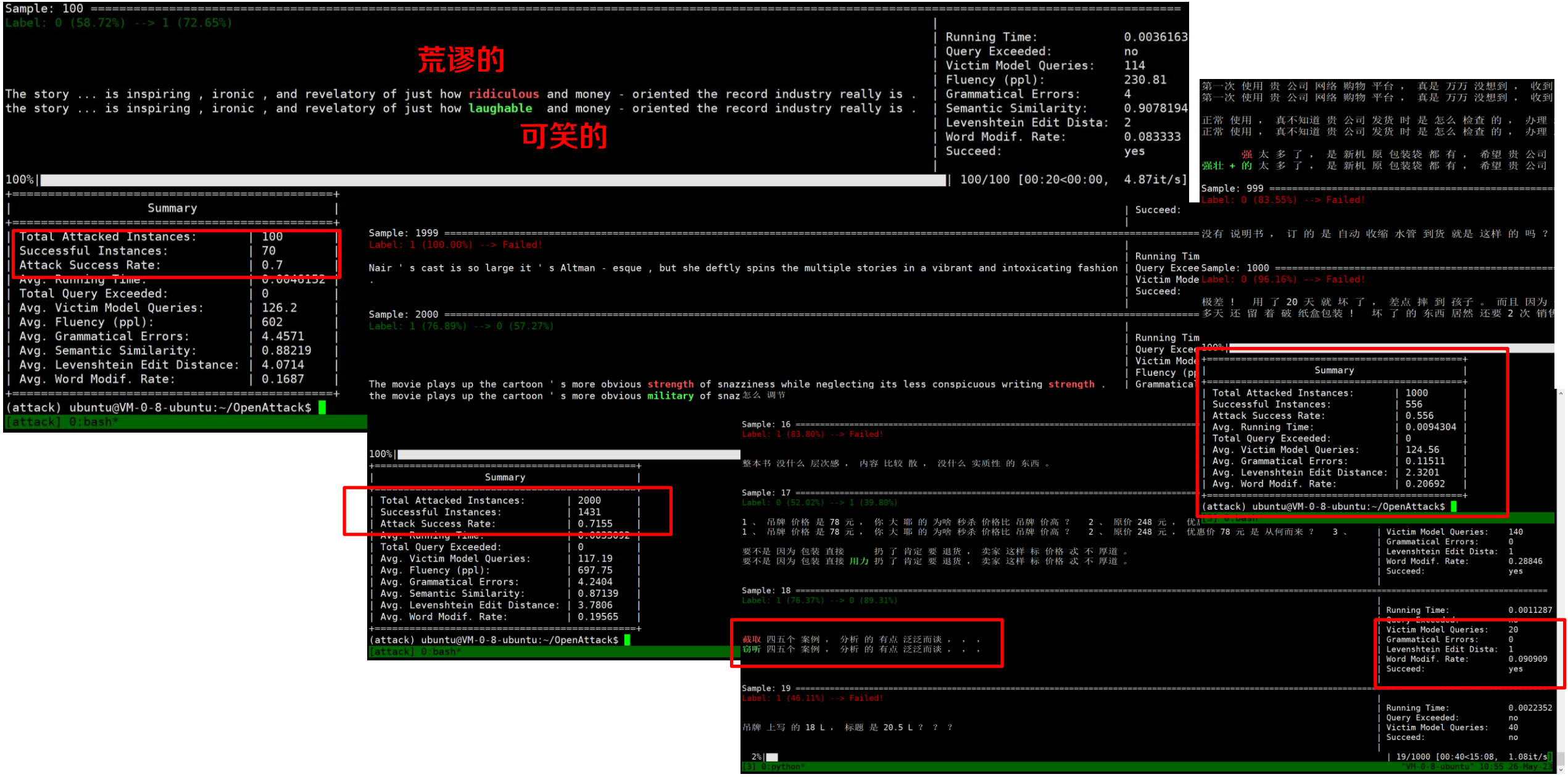
- OpenAttack
  - 清华大学THUNLP实验室开发的文本对抗攻击开源工具
  - 集成NLP主流的模型攻击方法，支持攻击模型多达13种
  - 支持中英双语



Model	Accessibility	Perturbation	Main Idea
SEA	Decision	Sentence	Rule-based paraphrasing
SCPN	Blind	Sentence	Paraphrasing
GAN	Decision	Sentence	Text generation by encoder-decoder
TextFooler	Score	Word	Greedy word substitution
PWWS	Score	Word	Greedy word substitution
Genetic	Score	Word	Genetic algorithm-based word substitution
SememePSO	Score	Word	Particle Swarm Optimization-based word substitution
BERT-ATTACK	Score	Word	Greedy contextualized word substitution
BAE	Score	Word	Greedy contextualized word substitution and insertion
FD	Gradient	Word	Gradient-based word substitution
TextBugger	Gradient, Score	Word+Char	Greedy word substitution and character manipulation
UAT	Gradient	Word, Char	Gradient-based word or character manipulation
HotFlip	Gradient	Word, Char	Gradient-based word or character substitution
VIPER	Blind	Char	Visually similar character substitution
DeepWordBug	Score	Char	Greedy character manipulation

https://github.com/thunlp/OpenAttack
https://github.com/QData/TextAttack

基础概念 OpenAttack



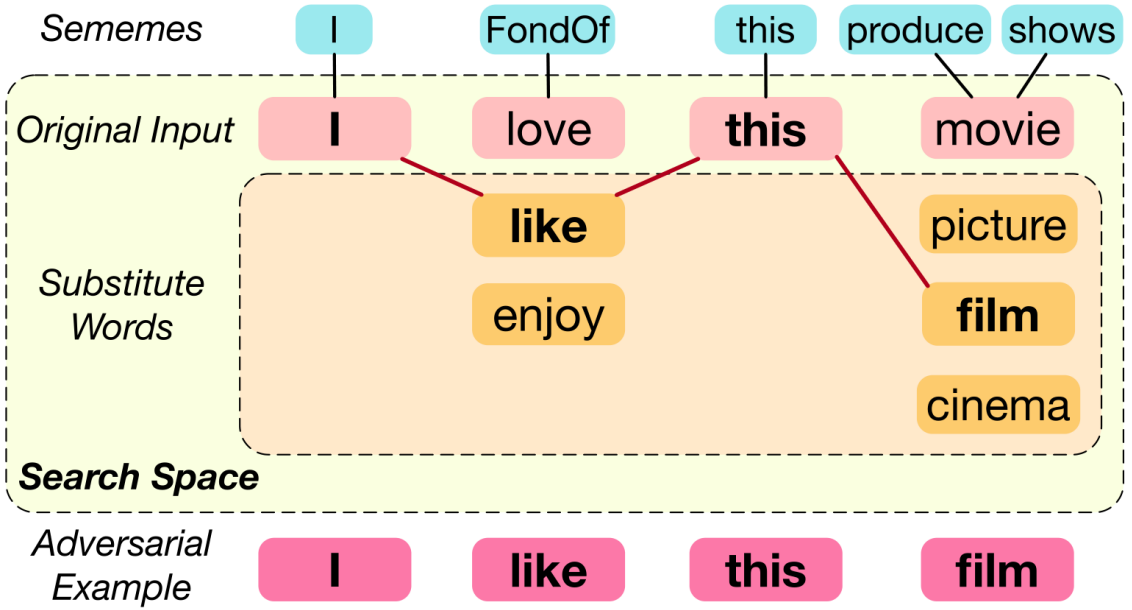
Word-Level Attack

WLTAACO TIPO

T	生成文本对抗样本
I	2个情感分类数据集、1个自然语言推理数据集 IMDB、SST-2、SNLI
P	1.使用义原替换方法缩减原始输入搜索空间； 2.使用粒子群优化算法以找到最佳的替换词； 3.生成对抗样本。
O	对抗样本

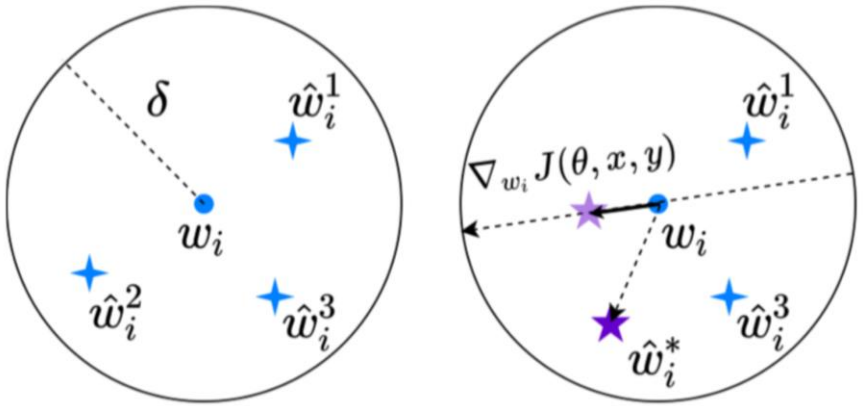
P	现有词级攻击模型中的搜索空间缩减方式不当、优化算法效率低下
C	基于语义相似度替换义原
D	如何选取多样化的攻击策略
L	ACL 2020 (CCF A)

- 算法原理
  - 缩小候选样本的空间
    - 排除无效或低质量的潜在对抗示例影响
    - 保留良好的语法性和流畅性的有效示例
  - 寻找最优搜索算法
    - 快速搜索与定位候选词
    - 找到可以成功欺骗目标模型的对抗样本

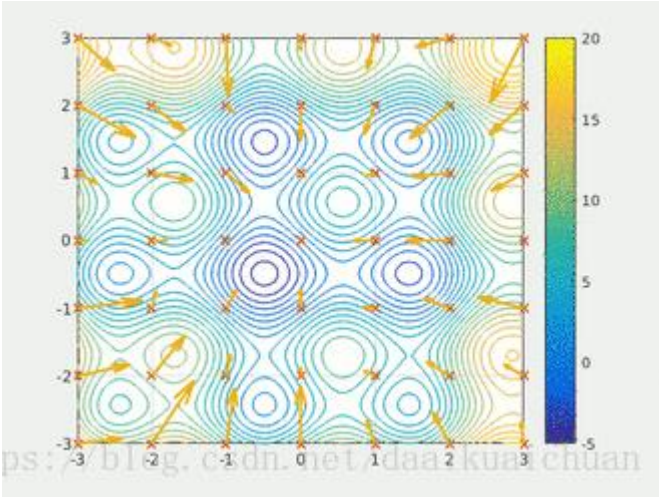


- 义原替换
  - 义原：
    - 即原子语义，是语言学意义上的最小的、不可再分的语义单位
    - 通常视为词语的语义标签，能够准确还原词语本意
  - HowNet：基于义原的大型中英文信息库
  - OpenHowNet：基于HowNet的开源信息库
- “基于义原”、“基于同义词”：
  - 义原能找到很多在同义词维度下难以发现的替换规则
  - 基于同义词的搜索空间稍微小一些

样本	We		love	**science**
义原	huaman	1stPerson	FondOf	knowledge
候选词	men	I	**like**	technology
	people			logic
对抗样本	People		like	science



- 超参数优化算法
  - 网格搜索（Grid Search）
    - 将可能的超参数空间划分为规则的网格，依次组合训练模型
  - 随机搜索（Random Search）
    - 区间范围内随机选择一组超参数
  - 贝叶斯优化（Bayesian Optimization）
    - 跟踪过去的评估结果和经验规律来迭代地更新概率模型
  - 贪婪算法（Greedy Algorithm）
    - 做出当前步最好的选择
  - 遗传算法（Genetic Algorithm）
    - 仿真生物遗传学和自然选择机理，通过人为方式构造的一类搜索算法
  - 粒子群算法（Particle Swarm Optimization, PSO）
    - 通过群体中个体之间的协作和信息共享来寻找最优解



- 基于PSO的对抗样本搜索算法
  - 效果：在缩减后的空间中使用一种基于粒子群的算法搜寻可以成功攻击目标模型的对抗样本，搜索复杂度降低
  - 改进的PSO算法中一个位置代表了一个对抗样本（句子），一个位置的每一个维度对应了一个单词，如下式：

$$\mathrm{x}^n = \omega_1^n \dots \omega_d^n \dots \omega_D^n, \ \omega_d^n \in \mathbb{V}(\omega_d^o)$$

- 位置信息x^n：包含了原始输入单词以及替换单词
- D：原始输入的单词数量
- ω_d^o：原始输入中的第d个单词
- $\mathbb{V}(\omega_d^o)$：由ω_d^o以及其替代词组成

- 基于PSO的对抗样本搜索算法
  - 初始化：（是否可以考虑随机初始化呢？）
    - 随机的替换原始输入中的一个单词来确定粒子的位置
    - 对于N个粒子，重复N次独立的类似操作
    - 粒子移动速度v^n的每一个维度 v_d^n都随机初始化到$[-V_{\max}, V_{min}]$
  - 终止条件：对于目前输入的任何一个对抗样本，目标模型都能预测出目标标签
  - 更新：由于搜索空间的离散性，通过下式更新速度信息：

随机初始化?

$$v_d^n = \omega v_d^n + (1-\omega) \times \left[\mathrm{I}(p_d^n, x_d^n) + \mathrm{I}\left(p_d^g, x_d^n\right)\right]$$

$$\mathrm{I}(a,b) = \begin{cases} 1, a = b \\ -1, a \neq b \end{cases}$$

$$\omega = (\omega_{max} - \omega_{min}) \times \frac{T-t}{T} + \omega_{min}$$

ω — 惯性系数
1. 在开始阶段粒子能够更快的搜寻更多位置
2. 最终更快的聚焦在最佳位置

- 基于PSO的对抗样本搜索算法
  - 基于文本的离散性
    - 不直接采用更新空间的方法，而是采用“是否移动”的概率思想
    - 位置更新方法：
      - 基于一种概率性的方法将粒子更新到最佳位置
      - 第一步，引入局部移动可能性P_i
      - 第二步，每个粒子根据另一个移动可能性P_g
  - 数学表示：

$$P_i = P_{max} - \frac{t}{T} \times (P_{max} - P_{min})$$

$$P_g = P_{min} + \frac{t}{T} \times (P_{max} - P_{min})$$

P_i——粒子是否单个移动到其最佳位置上

P_g——粒子是否整体移动到其最佳位置上

- 实验设计
  - 数据集
    - 3个基准数据集
  - 目标模型
    - BiLSTM
    - BERT
  - 评价指标
    - ASR：攻击成功率
    - Validity：人为检查的有效攻击率
    - Modification Rate：单词修改率
    - Grammaticality：语法检查率
    - Fluency：句子流畅率（困惑度Perplexity）
    - Naturality：句子自然性分数

Metrics	Evaluation Method	Better?
Success Rate	Auto	Higher
Validity	Human (Valid Attack Rate)	Higher
Modification Rate	Auto	Lower
Grammaticality	Auto (Error Increase Rate)	Lower
Fluency	Auto (Perplexity)	Lower
Naturality	Human (Naturality Score)	Higher

- 实验结果
  - 攻击任务
    - 本方法能达到更高的攻击成功率、产生更高质量的对抗样本
  - 文本相似性
    - 较低的PPL，表示生成的对抗样本更符合语法规则

Word Substitution Method	Search Algorithm	BiLSTM			BERT		
		IMDB	SST-2	SNLI	IMDB	SST-2	SNLI
Embedding/LM	Genetic	86.90	67.70	44.40	87.50	66.20	44.30
	Greedy	80.90	69.00	47.70	62.50	56.20	42.40
	PSO	96.90	78.50	50.90	93.60	74.40	53.10
Synonym	Genetic	95.50	73.00	51.40	92.90	78.40	56.00
	Greedy	87.20	73.30	57.70	73.00	64.60	52.70
	PSO	98.70	79.20	61.80	96.20	80.90	62.60
Sememe	Genetic	96.90	78.50	50.90	93.60	74.40	53.10
	Greedy	95.20	87.70	70.40	80.50	74.80	66.30
	PSO	**100.00**	**93.80**	**73.40**	**98.70**	**91.20**	**78.90**

Victim Model	Attack Model	IMDB			SST-2			SNLI		
		%M	%I	PPL	%M	%I	PPL	%M	%I	PPL
BiLSTM	Embedding/LM+Genetic	9.76	5.49	124.20	12.03	7.08	319.98	13.31	14.12	235.20
	Synonym+Greedy	6.47	4.49	115.31	10.25	4.65	317.27	12.32	21.37	311.04
	Sememe+PSO	**3.71**	**1.44**	**88.98**	**9.06**	**3.17**	**276.53**	**11.72**	**11.08**	**222.40**
BERT	Embedding/LM+Genetic	7.41	4.22	106.12	10.41	5.09	314.22	13.04	15.09	225.92
	Synonym+Greedy	4.49	4.48	98.60	8.51	4.11	316.30	**11.60**	11.65	285.00
	Sememe+PSO	**3.69**	**1.57**	**90.74**	**8.24**	**2.03**	**289.94**	11.72	**10.14**	**223.22**

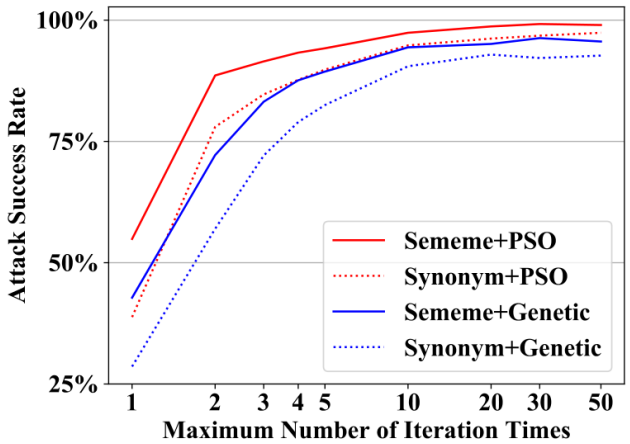
- 消融实验
  - Our sememe-based word substitution method
  - PSO-based search algorithm

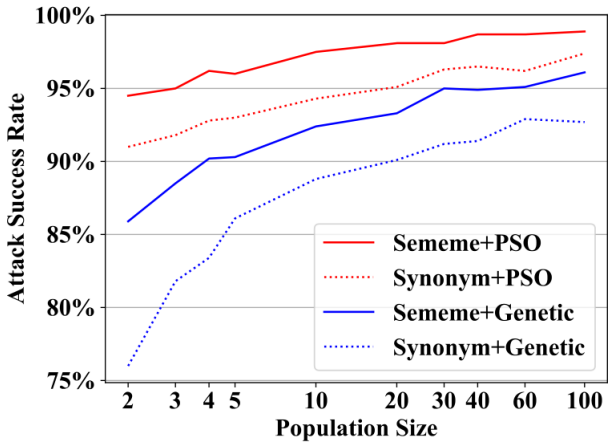
WordNet

She breaks the pie dish and screams out that she is not handicapped.		
Embedding/LM	Synonym	Sememe
tart, pizza, apple, shoemaker, cake cheesecake	None	cheese, popcorn, ham, cream, break, cake, pizza, chocolate, and 55 more

Word Substitution Method	IMDB	SST-2	SNLI
Embedding/LM	3.44	3.27	3.42
Synonym	3.55	3.08	3.14
Sememe	**13.92**	**10.97**	**12.87**

Victim	Attack Model	%Valid	NatScore
N/A	Original Input	90.0	2.30
BiLSTM	Embedding/LM+Genetic	65.5	2.205
	Synonym+Greedy	**72.0**	2.190
	Sememe+PSO	70.5	**2.210**
BERT	Embedding/LM+Genetic	**74.5**	2.165
	Synonym+Greedy	66.5	2.165
	Sememe+PSO	72.0	**2.180**





优劣分析

- 优势
  - 提出了一种优化搜索效率的单词级（word-level）攻击模型
  - 在攻击成功率、对抗样本的质量、迁移能力上都具有优越性和鲁棒性

- 劣势
  - 对抗样本在攻击对抗训练的模型时具有一定难度
  - 单一的词汇级替换方式，缺乏多样性

CLARE

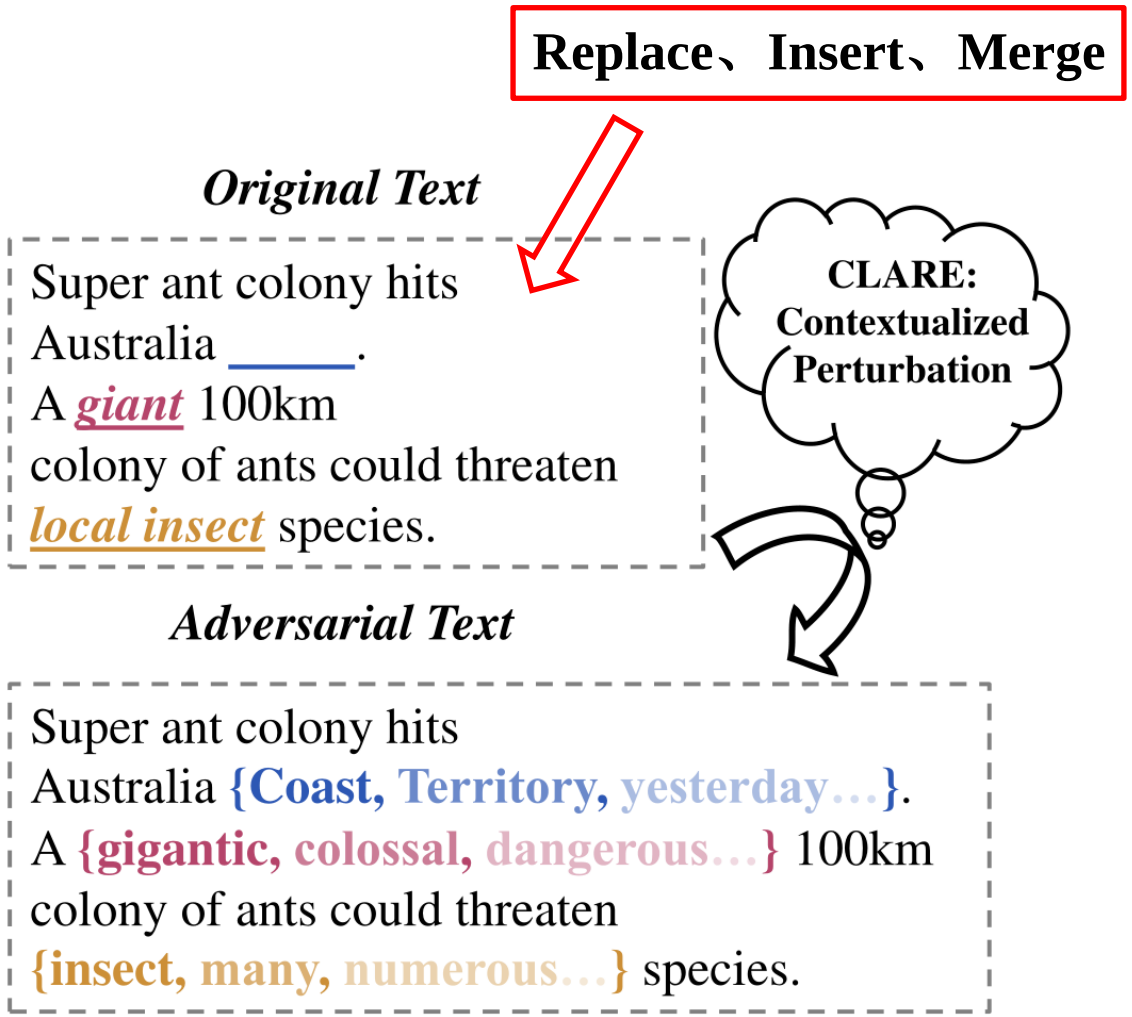
CLARE TIPO

T	生成自然且符合语法规则的对抗样本
I	文本分类数据集 Yelp Reviews, AG News, MNLI, QNLI
P	1.预训练：使用语言模型（如BERT）对原始文本预训练； 2.掩码生成、上下文扰动：使用掩码确定扰动的位置【随机、启发式】、Replace、Insert、Merge； 3.优先级填充； 4.对抗样本生成
O	对抗样本

P	对抗样本的自然性、流畅性不足，语法易出现错误
C	目标模型是预训练模型（如BERT）
D	如何保证语法正确性的同时，提升攻击成功率
L	NAACL 2021 (CCF A)

CLARE 算法原理

- 攻击算法提出的动机
  - 基于掩码语言模型生成近似词，实现基于上下文的扰动方式
  - 具体攻击算法原理如下：
    - 第一步，将掩码masks部署到特定文本中
    - 第二步，使用掩码语言模型生成替代词
  - 优势：
    - 生成不同长度的输出



CLARE 算法原理

- 掩码（masking）

The movie is fantastic.

The movie is amazing.

  - 在给定位置处应用不同策略的掩码操作
    - Replace：用掩码$[MASK]$替换x_i，然后从候选集$\mathcal{Z}$中选择一个token z 进行填充

$$\tilde{\mathbf{x}} = x_1 \dots x_{i-1}[MASK]x_{i+1} \dots x_n,$$

$$replace(\mathbf{x}, i) = x_1 \dots x_{i-1} z x_{i+1} \dots x_n$$

z是指从一个掩码语言模型中预测出的概率

$$p_{MLM}(z|\tilde{\mathbf{x}}) > k$$

$$sim(\mathbf{x}, \tilde{\mathbf{x}}_z) > \ell$$

ℓ是相似性概率值

  - 使用高的k和阈值ℓ会生成更像原始文本的自然句子
  - 但这样会降低攻击成功率，添加如下公式约束：

最小

$$p_f(y|\tilde{\mathbf{x}}_z)$$

选取k，ℓ平衡ASR和fluent of sentence

- 掩码（masking）
  - 语义自然性和流畅性可以通过构建候选集确定

$$\mathcal{Z} = \{z' \in \mathcal{V} | p_{MLM}(z'|\tilde{\mathbf{x}}) > k, sim(\mathbf{x}, \tilde{\mathbf{x}}_{\mathbf{z}'}) > \ell\}$$

k——5×10^{-3}
ℓ——0.7

    - $\mathcal{V}$是语言模型的词表数量
    - 减小了计算成本

$$|\mathcal{Z}| = 42(on\ average)$$

    - 词表大小$|\mathcal{V}| = 50{,}265$，候选集数量远小于词表大小
  - 攻击成功率通过下式实现
    - 计算标签预测为目标标签的概率最小值z：

$$z = \arg min_{z' \in \mathcal{Z}} p_f(y|\tilde{\mathbf{x}}_{\mathbf{z}'})$$

CLARE 算法原理

- 掩码（masking）

  I recommend.

  I highly recommend.

  – Insert插入
    - 添加额外的单词，文本长度加1
  – Merge合并
    - 合并$x_i x_{i+1}$，用一个位置的掩码填充
    - 可以看作是删除一个单词，文本长度减1

  New York is a beautiful city.

  York is a beautiful city.

$$\tilde{\mathbf{x}} = x_1 \dots x_{i-1}[MASK]x_{i+2} \dots x_n,$$

$$merge(\mathbf{x}, i) = x_1 \dots x_{i-1} z x_{i+2} \dots x_n$$

  – CLARE的优势：
    - 并行地为所有位置构建局部动作，即位置i的动作不受其他位置词变化的影响
    - 基于三种策略，可以选择执行策略的顺序

CLARE 算法原理

- 上下文顺序性扰动
  - 为位置i上的动作做3次变化，在长度为n的句子中，做$3n$次变化
  - 每个动作关联一个分数，计算该动作混淆f做决策的可能性

$$s_{(\mathbf{x},y)}(a) = -p_f(y|a(\mathbf{x}))$$

  - 选择扰动方式a——replace, insert, merge
  - 上式选择最值是为了避免对同一个位置进行多次修改操作

Algorithm 1 Adversarial Attack by CLARE

Input: Text-label pair $(\mathbf{x}, y)$; Victim model f
Output: An adversarial example
Initialization: $\mathbf{x}^{(0)} = \mathbf{x}$
$\mathcal{A} \leftarrow \varnothing$
for $1 \leq i \leq |\mathbf{x}|$ **do**
    $a \leftarrow$ highest-scoring action from $\{$ replace$(\mathbf{x}, i)$, insert$(\mathbf{x}, i)$, merge$(\mathbf{x}, i)$
    $\mathcal{A} \leftarrow \mathcal{A} \bigcup \{a\}$
end for
for $1 \leq t \leq T$ **do**
    $a \leftarrow$ highest-scoring action from $\mathcal{A}$
    $\mathcal{A} \leftarrow \mathcal{A} \setminus \{a\}$
    $\mathbf{x}^{(t)} \leftarrow$ Apply a on $\mathbf{x}^{(t-1)}$
    if $f(\mathbf{x}^{(t)}) \neq y$ **then return** $\mathbf{x}^{(t)}$
    end if
end for
return NONE

CLARE 实验设计

- 实验数据
  - Yelp Reviews
    - 基于餐厅评价的二元情感分类数据集
  - AG News
    - 包含四个类别的新闻文章集：World、Sports、Business、Science & Technology
  - MNLI
    - 每个实例都由一个前提假设对组成
    - 涵盖了来自各个领域的文本，从隐含、中性和矛盾的标签集来确定它们之间的关系
  - QNLI
    - Stanford Q&A 数据集转换而来的二元分类数据集
    - 任务是确定上下文是否包含问题的答案，主要基于wiki的英文文章

Dataset	Avg. Length	# Classes	Train	Test	Acc
Yelp	130	2	560K	38K	95.9%
AG News	46	4	120K	7.6K	95.0%
MNLI[6]	23/11	3	392K	9.8K	84.3%
QNLI	11/31	2	105K	5.4K	91.4%

受害目标模型在没有对抗性攻击的情况下在原始测试集上的准确性

CLARE 实验设计

- 实验设置
  - RoBERTa(distill)作为掩码语言模型，填充上下文
  - 消融实验中，比较了其和RoBERTa(base)以及BERT（base）的性能
  - 相似性函数是USE（通用语句编码器）中的通用函数
  - Victim model：微调了一个基于BERT的文本分类器
- 实验指标
  - ASR（A-rate）、Modification rate（Mod）、Perplexity（PPL）、Grammar error（GErr）、Textual similarity（Sim）
- 基线方法
  - TextFooler
  - TextFooler+LM
  - BERTAttack

CLARE 实验结果

- 实验结果
  - 最好的性能归因于
    - CLARE通过在任何位置组合3种不同的扰动而获得的更多样的攻击策略

	Yelp (PPL = 51.5)					AG News (PPL = 62.8)				
Model	**A-rate↑**	**Mod↓**	**PPL↓**	**GErr↓**	**Sim↑**	**A-rate↑**	**Mod↓**	**PPL↓**	**GErr↓**	**Sim↑**
TextFooler	77.0	16.6	163.3	1.23	0.70	56.1	23.3	331.3	1.43	0.69
+ LM	34.0	17.4	90.0	1.21	0.73	23.1	21.9	144.6	1.07	0.74
BERTAttack	71.8	10.7	90.8	0.27	0.72	63.4	7.9	90.6	0.25	0.71
CLARE	**79.7**	**10.3**	**83.5**	**0.25**	**0.78**	**79.1**	**6.1**	**86.0**	**0.17**	**0.76**
	MNLI (PPL = 60.9)					**QNLI** (PPL = 46.0)				
Model	**A-rate↑**	**Mod↓**	**PPL↓**	**GErr↓**	**Sim↑**	**A-rate↑**	**Mod↓**	**PPL↓**	**GErr↓**	**Sim↑**
TextFooler	59.8	13.8	161.5	0.63	0.73	57.8	16.9	164.4	0.62	0.72
+ LM	32.3	12.4	91.9	0.50	0.77	29.2	17.3	85.0	0.42	0.75
BERTAttack	82.7	8.4	86.7	0.04	0.77	76.7	13.3	86.5	0.03	0.73
CLARE	**88.1**	**7.5**	**82.7**	**0.02**	**0.82**	**83.8**	**11.8**	**76.7**	**0.01**	**0.78**

CLARE有最高的攻击成功率、最少的平均修改次数

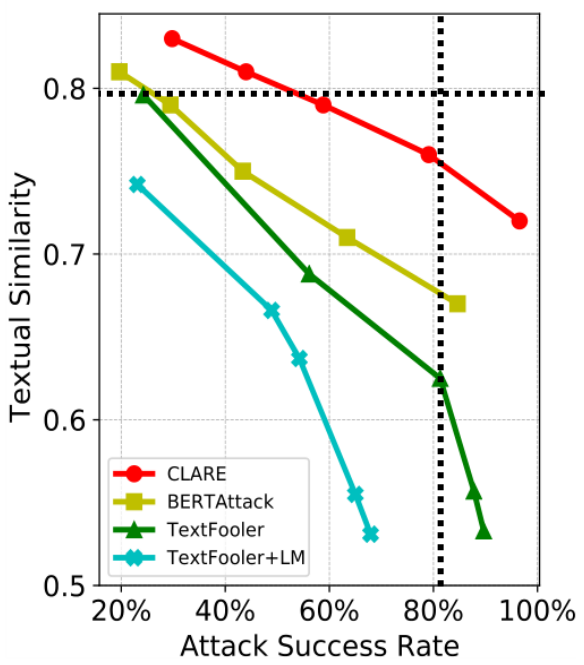
CLARE 实验结果

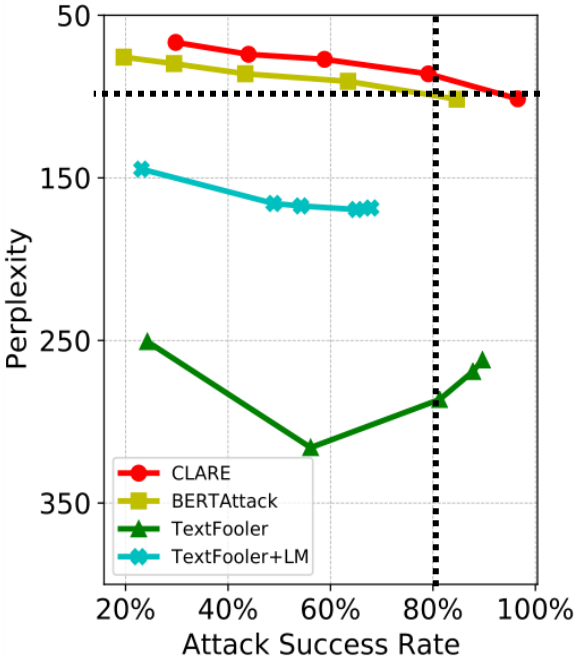
- 实验结果
  - 人为对CLARE与TextFooler生成的对抗样本打分指标和自动打分趋势一致
  - 在标签一致性指标上：
    - CLARE略微比TextFooler低，原因是AG News数据集某些类别存在重叠
    - 如Science，Business两类

Metric	CLARE	Neutral	TextFooler
Similarity	$56.1_{\pm 2.5}$	28.1	$15.8_{\pm 2.1}$
Fluency&Grammaticality	$42.5_{\pm 2.5}$	48.6	$8.9_{\pm 1.5}$
Label Consistency	$68.0_{\pm 2.4}$	-	$70.1_{\pm 2.5}$

ASR高
文本相似性高

ASR高
PPL低





CLARE 实验结果

- 实验结果

BERTAttack未考虑攻击位置

Module	A-rate↑	Mod↓	PPL↓	GErr↓	Sim↑
CLARE	79.1	**6.1**	**86.0**	0.17	0.76
MERGEONLY[8]	47.2	6.2	95.3	**0.08**	**0.79**
INSERTONLY	68.1	7.2	93.1	0.23	0.74
REPLACEONLY	66.7	7.7	85.6	0.10	0.72
BERTAttack	63.4	7.9	90.6	0.25	0.71
w/o sim $> \ell$	82.4	6.9	86.8	0.13	0.70
w/o $p_{\text{MLM}} > k$	**95.7**	6.8	162.8	0.71	0.61

一秒能处理的样本数量

MLM	A-rate↑	Mod↓	PPL↓	Sim↑	Speed↑
$\text{RoBERTa}_{\text{distill}}$	79.1	**6.1**	**86.0**	**0.76**	**0.14**
$\text{RoBERTa}_{\text{base}}$	**79.3**	6.3	88.9	0.75	0.07
$\text{BERT}_{\text{base}}$	78.4	8.3	95.2	0.71	0.06

AG (Sci&Tech)	Sprint Corp. is in talks with Qualcomm Inc. about using a network the chipmaker is building to deliver live television to Sprint mobile phone customers.
TextFooler (Business)	Sprint *Corps*. is in talks with Qualcomm Inc. about *operated* a network the chipmaker is *consolidation* to *doing viva* television to Sprint mobile phone customers.
CLARE (Business)	Sprint Corp. is in talks with Qualcomm Inc. about using a network Qualcomm is building to deliver *cable* television to Sprint mobile phone customers.
MNLI (Neutral)	*Premise*: Let me try it. She began snapping her fingers and saying the word eagerly, but nothing happened. *Hypothesis*: She became frustrated when the spell didn't work.
TextFooler (Contradiction)	*Premise*: *Authorisation* me *attempting* it. She *triggered flapping* her *pinkies* and *said* the word eagerly, but nothing *arisen*. *Hypothesis*: She became frustrated when the spell didn't work.
CLARE (Contradiction)	*Premise*: Let me try it. She began snapping her fingers and saying the word eagerly, but nothing **unexpected** happened. *Hypothesis*: She became frustrated when the spell didn't work.

CLARE 实验结果

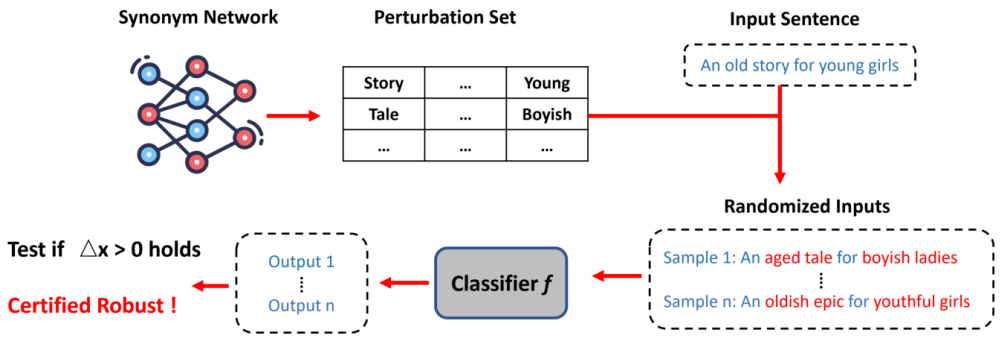
- 实验结果

AG (Business)	TECH BUZZ : Yahoo, Adobe team up for new Web services. Stepping up the battle of online search and services, Yahoo Inc. and Adobe Systems Inc. have joined forces to tap each other's customers and put Web search features into Adobe's popular Acrobat Reader software.
TextFooler (Sci&Tech)	TECH BUZZ : Yahoo, Adobe team up for *roman Cyberspace utilities*. Stepping up the battle of online *locating* and services, Yahoo Inc. and Adobe Systems Inc. have joined forces to tap each other's customers and put Web search features into Adobe's popular Acrobat Reader software.
CLARE (Sci&Tech)	TECH BUZZ : Yahoo, Adobe team up for new Web *Explorer*. Stepping up the battle of online search and services, Yahoo Inc. and Adobe Systems Inc. have joined forces to tap each other's customers and put Web search features into Adobe's popular Acrobat Reader software.
AG (Sport)	Padres Blank Dodgers 3 - 0. LOS ANGELES - Adam Eaton allowed five hits over seven innings for his career - high 10th victory, Brian Giles homered for the second straight game, and the San Diego Padres beat the Los Angeles Dodgers 3 - 0 Thursday night. The NL West - leading Dodgers' lead was cut to 2 1 / 2 games over San Francisco - their smallest since July 31 ...
TextFooler (World)	Dodger Blank *Yanks* 3 - 0. *Loos* ANGELES - *Adams Parades enabling* five hits over seven *slugging* for his career - high 10th *victoria*, Brian Giles homered for the second straight *matching*, and the *Tome José Dodger* beat the Los Angeles *Dodger* 3 - 0 Thursday *blackness*. The NL *Westerner* - *eminent Dodger*' lead was cut to 2 1 / 2 games over San *San* - their *tiny as janvier* 31 ...
CLARE (World)	Padres Blank Dodgers 3 - 0. Milwaukee **NEXT** - Adam Eaton allowed five hits over seven innings for his career - high 10th victory, Brian Giles homered for the second straight game, and the San Diego Padres beat the Los Angeles Dodgers 3 - 0 Thursday night. The NL West - leading Dodgers' lead was cut to 2 1 / 2 games over San Francisco - their smallest since July 31 ...

更少的修改，更自然的语法

应用总结

- 优势
  - Word-Level Attack
    - 提出了一种新的基于义原和PSO的单词级（word-level）攻击模型
  - CLARE
    - 使用三种不同的扰动方式，对输入的每个位置提供有效的攻击方案
    - 基于上下文产生不同长度的对抗样本，语言模型的优良性能
- 劣势
  - 产生的对抗样本语义相似性和自然性仍有待提高
- 未来工作
  - 基于同义词替换的快速梯度投影方法（FGPM）
  - 上下文语义偏差的文本对抗攻击
  - 对抗样本的可用性评估



参考文献

[1] Goyal S, Doddapaneni S, Khapra M M, et al. A Survey of Adversarial Defences and Robustness in NLP[J]. ACM Computing Surveys, 2022.

[2] Zang Y, Qi F, Yang C, et al. Word-level textual adversarial attacking as combinatorial optimization[J]. arXiv preprint arXiv:1910.12196, 2019.

[3] Li D, Zhang Y, Peng H, et al. Contextualized perturbation for textual adversarial attack[J]. arXiv preprint arXiv:2009.07502, 2020.

[4] Zeng G, Qi F, Zhou Q, et al. Openattack: An open-source textual adversarial attack toolkit[J]. arXiv preprint arXiv:2009.09191, 2020.

[5] Alzantot M, Sharma Y, Elgohary A, et al. Generating natural language adversarial examples[J]. arXiv preprint arXiv:1804.07998, 2018.

[6] Garg S, Ramakrishnan G. Bae: Bert-based adversarial examples for text classification[J]. arXiv preprint arXiv:2004.01970, 2020.

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。