TextBugger

用途：产生对抗样本，使目标模型分类错误

特性：针对英文，分类任务

局限性：未在Transformers系列下测试

概念：

- 一个分类器M，接收一段文本a，输出logits = M(a)（各个标签的概率），标签y = argmax(logits)
- TextBugger寻找一个对抗样本b，满足dis(a, b) <= threshold，即b与a差别不会很大，不影响人的正常阅读，使得输出标签 z = argmax(M(b)) != y

TextBugger

场景：白盒与黑盒，实际环境中，在白盒场景下做对抗攻击较少，因此主要讨论黑盒环境

术语：bug->将原文本中的一个单词使用某攻击手段攻击后的产物

攻击手段：

- Insert
- Delete
- Swap
- Sub-C (形相似)
- Sub-W （意相似，基于glove）

Original	Insert	Delete	Swap	Sub-C	Sub-W
foolish	f oolish	folish	fooilsh	fo0lish	silly
awfully	awfull y	awfuly	awfluly	awfu1ly	terribly
cliches	clich es	clichs	clcihes	c1iches	cliche

TextBugger

黑盒攻击流程：

1. 寻找重要句子：对每一个句子做查询，按初始label的概率排序
2. 寻找重要单词：对一个句子中的每一个词，查询删除该词后本句的初始label概率，差距越大则越重要
3. Bug替换：将重要的词使用所有bug进行尝试，保留影响最大的bug，后替换下一个单词
4. 失败条件：当某次bug替换后，对抗文本与原始文本的语义差距过大，文中使用Universal Sentence Encoder 作为指标

TextBugger

复现：

- 推荐 hugging face transformers 库
- 流程：
  - 训练一个简单的文本分类器（按照教程即可）
  - 实现攻击手段（简单的字符串操作，其中sub-w有库可以使用）
  - 实现攻击框架（对pytorch、transformers有一定了解，能获取输出的概率即可）

TextShield

用途：转化对抗样本，使目标模型能正确识别

TextBugger的中文化：

- 分词工具：spacy、jieba、hanlp等，但也不能完美分词
- 更多的攻击手段：拼音、同音字等

Bug	Example	Bug	Example
Insert	傻逼 → 傻&逼	Sim2Trad/1	裸体 → 裸體
PyConvert/1	智障 → zhi zhang	Sim2Trad/2	裸体 → 裸&體
PyConvert/2	智障 → zhi zha.ng	GlyphSim/1	赌博 → 堵博
PyConvert/3	智障 → zhi zhan	GlyphSim/2	赌博 → 堵搏
PyConvert/4	智障 → zhi zhnag	GlyphSim/3	赌博 → 堵t搏
PyConvert/5	智障 → ℤhi zhang	PhoneticSim/1	色情 → 涩情
Split/1	炸弹 → 火乍弓单	PhoneticSim/2	色情 → 涩o情
Split/2	炸弹 → 炸弓/单		

TextShield

原理：通过Text2Text(NMT)进行对抗训练，将对抗样本还原

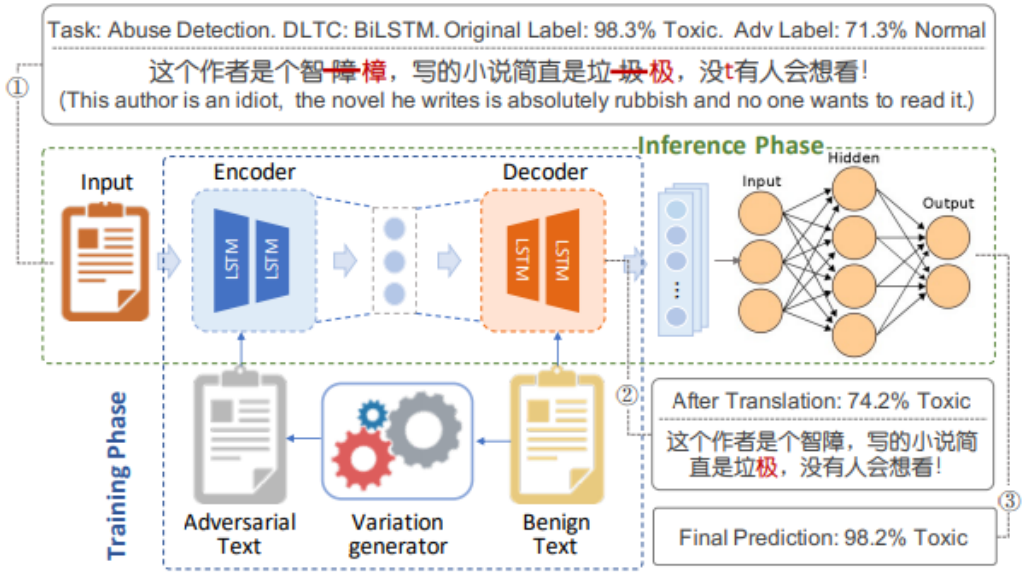

Figure 1: The framework of TEXTSHIELD.

TextShield

多重embedding：

- 语义：skip-gram
- 图形：CNN
- 读音：转拼音后skip-gram

组合：

- EMF：简单拼接

$$\boldsymbol{V} = [\boldsymbol{V}^{(S)} \oplus \boldsymbol{V}^{(G)} \oplus \boldsymbol{V}^{(P)}].$$

- IMF：三个embedding再通过一个简单的神经网络转化后拼接

$$\boldsymbol{V} = [F_s(\boldsymbol{V}^{(S)}) \oplus F_g(\boldsymbol{V}^{(G)}) \oplus F_p(\boldsymbol{V}^{(P)})],$$

TextShield

复现：

- 没有直接复现，使用mT5简单训练后，发现效果较差，所以没有进一步全面复现

局限性：

- 文中主体仍为LSTM、skip-gram等“过时”技术，
- 没有对bert等新模型尝试；
- 文中NMT训练了约100000 epoch，loss稳定约为
- 20000epoch，想达到文中效果是困难的；
- 使用全体NMT还原，很可能导致过拟合的情况；