Reevaluating Adversarial Examples in Natural Language

会议：2020EMNLP

内容：评价攻击方法并提出改进

思想：认为现有sota攻击模型生成的样本质量低，应该添加更严格的限制

受益之处：

- 提出应该缩小1.替换单词的余弦相似度的阈值（0.9） 2.攻击样本与原样本之间的Sentence Enconding的余弦相似度的阈值（0.98）
- 单词用Counter fitting word vectors论文里的词向量，句子应该是用USE
- 提出用Language Tool检查语法错误