Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment

会议：2020AAAI

主要思想：

> 通过**词替换**产生对抗样本

具体做法：

- **选出重要词**

  同样用word saliency, 找出关键词排序后，将其中的停用词去掉

- **词替换原则**

  1. 尽可能与原句相似
  2. 能融入上下文
  3. 使得模型预测错误

- **具体流程**

  1. 初始化备选词列表
     - 先将词表里所有词与当前词做cos相似度，取出TopN

       > 这里使用2016年一篇paper《Counter-fitting word vectors to linguistic constraints》中提出的词向量表示，在SimLex-999数据集上达到了SOTA，专门用于判断词与词之间的语义相似度。

     - POS Checking

  > 过滤候选词里POS与当前词不同的词

  - Semantic Similarity Checking

    > 检查扰动完的句子与原句的相似度
    >
    > 这里用Universal Sentence Encoder(USE)对原句与扰动句计算相似度，相似度大于一定阈值才可以保留
    >
    > （经常看到有人用USE计算两句话相似度，可以尝试）