数据增强的方法

- Paraphrasing: 对句子中的词、短语、句子结构做一些更改，保留原始的语义
  - 利用MLM模型，例如BERT模型，Mask掉一部分词，作为生成新的句子
  - 嵌入层扰动，例如对抗训练
  - 利用相似嵌入向量替换：利用知识图谱中寻找找相似的向量
  - 同义词替换： 利用词典、知识图谱等外部数据，随机将非停用词替换成同义词或上位词，如果增加多样性的话还可以替换成相同词性的其他词
  - Back-translation（反向翻译）
  - 模型生成：利用seq2seq生成语义一致的句子
- Noising: 在保证label不变的同时，增加一些离散或连续的噪声，对语义的影响不大
  - 交互词位置
  - 删除部分词
  - 插入部分词
  - 替换部分词
  - Mixup方法，参考论文
- Sampling: 旨在根据目前的数据分布选取新的样本，会生成更多样的数据