作业说明

本次机器翻译大作业的内容为实现使用Attention机制的seq2seq模型，完成英文到中文的机器翻译任务。

data文件夹下为中英文的训练语料，以及测试、验证语料。

notebook中的代码框架仅供参考，不一定要严格地进行代码填空，大家可以根据需要自行修改notebook中的内容，只要能够实现各部分的功能并跑通整个模型即可。

需要实现的主要部分为：

1. encoder的代码
2. attention模块以及使用了attention机制的decoder
3. 整个seq2seq模型
4. 模型训练部分的代码
5. 在验证/测试集上生成文本的代码

只用CPU训练，一个epoch耗时可能在5分钟左右，使用GPU加速一个epoch大约30秒。可以考虑使用Google Colab等在线计算资源，或者视训练情况而定适当减少epoch数。在训练过程中可以tqdm模块给for循环加上进度条，方便在每个epoch内部观察训练的进度。

单人版大作业要求

提交seq2seq模型的代码（jupyter notebook或者python文件），对所写代码进行适当的注释。

写一份简单的作业报告，包含关于模型实现的大致说明，训练情况和测试结果。

使用notebook提交的同学，可以在notebook中保留必要的输出结果，如果是运行python文件的同学，建议将输出结果单独保存成一个文件（比如在测试集上进行文本生成的结果，训练过程中的loss、bleu值等等）。

双人版大作业要求

完成单人版大作业，在此基础上完成以下内容：

1. 使用GIZA++进行文本对齐，生成双语词典。
2. 加入独立的multi-word alignment模块，发掘多词单位，对语料进行短语挖掘
3. 尝试以n-gram（短语）为单位进行翻译，对单人版的只使用attention机制的机器翻译模型进行优化

相关资料：

GIZA++：http://www.statmt.org/moses/giza/GIZA++.html

使用AutoPhrase的工具自动挖掘短语，github:https://github.com/shangjingbo1226/AutoPhrase

论文链接：

https://www.aclweb.org/anthology/J03-1002.pdf

https://arxiv.org/pdf/1702.04457.pdf

http://hanj.cs.illinois.edu/pdf/sigmod15_jliu.pdf

要求提交详细版的作业报告，除了基础版作业报告的内容，还需要在报告中介绍双语词典的建立、短语挖掘等部分，以及采用的优化方法，分析经过优化后的翻译模型与基础模型的效果对比。

提交时间：

非毕业年级：7月9日中午12点

毕业年级：6月21日晚上12点

补充说明：

预计在7月3日下午1点到3点进行一次线上交流，由助教介绍与NLP相关的几个小专题，感兴趣的同学可以先阅读一下这些文献：

- Seq2Seq中Unknown Words的处理
- Copy Mechanism：https://arxiv.org/pdf/1603.06393.pdf
- Coverage Mechanism：https://arxiv.org/abs/1601.04811，https://arxiv.org/pdf/1704.04368.pdf
- Back translation：https://arxiv.org/abs/1511.06709，https://www.aclweb.org/anthology/W18-2703.pdf