该项目中用到的数据集

bobsue.lm.train.txt
bobsue.lm.dev.txt
bobsue.lm.test.txt
bobsue.seq2seq.train.txt
bobsue.seq2seq.dev.txt
bobsue.seq2seq.test.txt
bobsue.vocab.txt

背景

该项目的数据来源于story generation，即给定前一句话，预测下一句话。该问题可以使用简单的语言模型或者seq2seq模型来实现。该项目我们会让大家来尝试研究这个问题。

1. 语言模型（10分）

利用bobsue.lm.train.txt 训练一个语言模型，请使用LSTM模型，以cross entropy loss作为训练用的loss function。使用bobsue.lm.dev.txt进行early stopping和各种parameter tuning，直到你得到了你认为的最好的模型。然后在bobsue.lm.test.txt上使用这个模型。汇报在dev set上得到的最好accuracy以及该模型在test set上的accuracy。

你的模型需要预测除了<s>之外的每个字符。在dev set上你需要预测7597个单词（包括</s>)，在test set上你需要预测8059个单词。当你预测下一个单词的时候你可以假设自己能够看到该句话前面已经出现的所有单词。

2. Sequence to sequence模型（15分）

利用bobsue.seq2seq.train.tsv训练一个sequence 2 sequence模型。该文件每行包括两句话，第二句话与bobsue.lm.train.txt完全相同。
你的任务与前一模块相同，也是预测第二句话中除<s>之外的每一个单词。请汇报在dev和test上的结果。
请使用encoder decoder模型做单词预测任务。请尝试下述几种encoder：

1. 使用一个LSTM当做encoder，使用最后一个hidden和cell vectors当做decoder的初始hidden和cell vectors。同样使用cross entropy loss进行训练。
2. 训练一个bi-directional LSTM，将forward和backward LSTM的最后一个hidden and cell vectors concatenate，然后用作decoder的初始hidden/cell vectors进行训练。
3. 使用一个average word embedding(拿第一句话每个单词的word embedding取平均)作为encoder，得到的context vector可以直接用作decoder的初始hidden/cell state，你也可以多加一层linear layer转换成hidden/cell vectors，再做decoding，请将你的做法写入报告中。

计算seq2seq模型的accuracy结果，对比语言模型和seq2seq模型。

3. 请提出一些方法来优化(2)的模型，例如加入attention，使用多层LSTM，使用GRU，使用CNN等等。使用的方法越有趣越好，请在报告中详细介绍自己使用的模型。如果有任何优化小方法小技巧，无论是自己想到的还是从别处看来的，欢迎一并写入报告中。请汇报accuracy结果。（15分）

4. 请使用(2)或/和(3)训练得到的seq2seq模型做next sentence generation。请随机从bobsue.lm.dev.txt选取10个句子，利用encoder encode句子，然后用decoder generate一些句子。挑选一些你认为有趣的结果写在报告中。（10分）

请提交：

1. 一份2页纸以内的项目报告
2. 所有项目中设计的代码，代码中请给出详细的注释。代码可以使用python文件或者jupyter notebook文件。代码建议使用tensorflow库，但是也支持同学们使用自己喜欢的库。