Model程序说明:

1、程式部分

model 的程式部分由 seq2seq\_model.py、translate.py 和 data\_utils.py 三個指令檔組成。translate.py 是整個 model 的主程序。

seq2seq\_mode.py 和 data\_utils.py 配合 translate .py 使用，完成相关model的建立，训练模型验证，结果输出以及模型评估。

（1）、translate.py

translate.py是主程序，负责初始化参数，data的载入及预处理，RNN模型的建立，监督模型的训练，预测结果，计算BLEU以及输出结果。

read\_data():将数据从文件中读入，并切分处理后读入到bucket中。

get\_bleu():调用mult\_bleu.perl计算bleu score。

save\_result():将预测的结果存入到对应的file中，主要是在get\_bleu的时候用。

create\_model():创建Seq2SeqModel，并且判断是否已经有训练模型，若有则读入model数据，否则重新初始化参数预备模型的训练。

train()：模型训练的主函数，主要完成数据的预读入，预处理，模型的初始化，并执行训练过程，反复执行seq2seq\_model.py中的step训练操作， 对训练结果进行评估，计算BLEU的值。

decode():读入test.en的值，通过已有的训练模型数据，预测每一行的test的结果，并将结果输出成文件。

（2）、seq2seq\_mode.py

seq2seq\_mode.py是model程序。定义了多层的RNN网络模型，包括一个encoder和一个decoder模型。模型支持GRU cells和LSTMcells。

\_\_init\_():创建和初始化模型。根据在translate.py定义的各种初始化参数，初始化模型的输入输出，softmax loss。

step():根据feed的数据，进行训练。

get\_batch():从bucket中随机读取一个batch，该数据将会用在step()中。

（3）、data\_utils.py

data\_utils.py是数据读入和数据初始化预处理的程序。

maybe\_download():查看数据是否存在，不存在则下载。

gunzip\_file():解压数据

get\_wmt\_enfr\_train\_set():获取训练数据。

get\_wmt\_enfr\_dev\_set():获取valid数据。

basic\_tokenizer():对每行数据进行拆分。

create\_vocabulary():根据data file创建词表file。

initialize\_vocabulary():初始化词典file为array 和dict。

sentence\_to\_token\_ids():将string处理成tocken\_ids。

data\_to\_token\_ids():将data file根据给定的词表file，转成token\_ids。这个方面是从data path里面加载数据，然后调用sentence\_to\_token\_ids()，将结果存在target path里面。

prepare\_wmt\_data():调用以上方法，将data\_dir中，将数据读出，创建词表，并且tokenize数据。

可改进的地方：

（1） 机器翻译时，将输入句子中的单词顺序颠倒一下可以提高机器翻译的性能，这可能是因为减小了源句与翻译后的句子的对应词语之间的距离；

（2） 模型训练较慢，需要调整部分参数。