Model程式說明:

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，完成相關model的建立，訓練模型驗證，結果輸出以及模型評估。

（1）、translate.py

translate.py是主程序，負責初始化參數，data的載入及預處理，RNN模型的建立，監督模型的訓練，預測結果，計算BLEU以及輸出結果。

read\_data():將資料從檔中讀入，並切分處理後讀入到bucket中。

get\_bleu():調用mult\_bleu.perl計算bleu score。

save\_result():將預測的結果存入到對應的file中，主要是在get\_bleu的時候用。

create\_model():創建Seq2SeqModel，並且判斷是否已經有訓練模型，若有則讀入model資料，否則重新初始化參數預備模型的訓練。

train()：模型訓練的主函數，主要完成資料的預讀入，預處理，模型的初始化，並執行訓練過程，反復執行seq2seq\_model.py中的step訓練操作， 對訓練結果進行評估，計算BLEU的值。

decode():讀入test.en的值，通過已有的訓練模型資料，預測每一行的test的結果，並將結果輸出成檔。

（2）、seq2seq\_mode.py

seq2seq\_mode.py是model程式。定義了多層的RNN網路模型，包括一個encoder和一個decoder模型。模型支援GRU cells和LSTMcells。

\_\_init\_():創建和初始化模型。根據在translate.py定義的各種初始化參數，初始化模型的輸入輸出，softmax loss。

step():根據feed的資料，進行訓練。

get\_batch():從bucket中隨機讀取一個batch，該資料將會用在step()中。

（3）、data\_utils.py

data\_utils.py是資料讀入和資料初始化預處理的程式。

maybe\_download():查看資料是否存在，不存在則下載。

gunzip\_file():解壓數據

get\_wmt\_enfr\_train\_set():獲取訓練資料。

get\_wmt\_enfr\_dev\_set():獲取valid資料。

basic\_tokenizer():對每行資料進行拆分。

create\_vocabulary():根據data file創建詞表file。

initialize\_vocabulary():初始化詞典file為array 和dict。

sentence\_to\_token\_ids():將string處理成tocken\_ids。

data\_to\_token\_ids():將data file根據給定的詞表file，轉成token\_ids。這個方面是從data path裡面載入資料，然後調用sentence\_to\_token\_ids()，將結果存在target path裡面。

prepare\_wmt\_data():調用以上方法，將data\_dir中，將資料讀出，創建詞表，並且tokenize資料。

可改進的地方：

（1） 機器翻譯時，將輸入句子中的單詞順序顛倒一下可以提高機器翻譯的性能，這可能是因為減小了源句與翻譯後的句子的對應詞語之間的距離；

（2） 模型訓練較慢，需要調整部分參數。