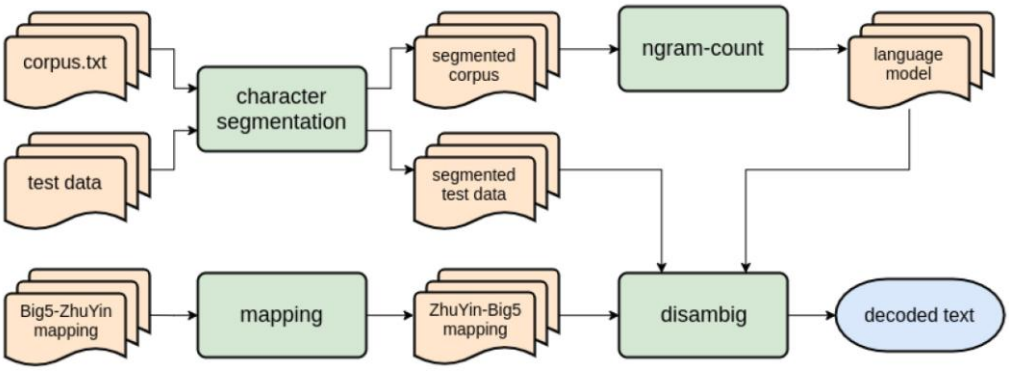
Digital Speech Processing Homework 3 Report

Name：鐘民憲　student ID：B06901017

1.Workflow



(1)Use the provided separator_big5.pl to generate segmented corpus and test data

(2)Use SRILM to generate count file from segmented corpus and then get language model from count file.

(3)Run mapping.py to transform Big5-ZhuYin mapping into ZhuYin-Big5 mapping.

(4)Run mydisambig to decode the segmented test data with the help of ZhuYin-Big5 mapping and language model.

(5)Validate the decode.txt by comparing it with the result produced by disambig provided by SRILM.

2.Mydisambig 實作技術與細節

(1)File I/O

由於所有檔案都是採 big5 編碼，也就是每個中文字跟注音都用兩個 char 來儲存，所以我花了些時間研究要如何讀檔跟寫檔。我的作法是用 char* buf=file.geline() 一次吃一行，再用 int(buf[i])來觀察每個字都是由哪兩個數字組成的，結果發現只有注音符號的第一個數字一定是-93，第二個數字則是ㄅ~ㄏ分別為 116~126、ㄐ~ㄩ為-95~70。根據這個規則，我可以將 ZhuYin-Big5 map 存到 37 個 vectors 中，給我任一注音符號，我便可以依據它的第二個數字去找到對應的 vector。而 language model 我則是直接使用 Ngram.h library 來儲存。至於要 decode 的檔案我也是一行一行讀取，先去掉多餘的空格(每個字之間只保留一個空格)，接下來進行 viterbi 運算，把注音替換成正確的文字，最後再一行行寫入 output file，前後分別加上<s>跟</s>。

(2)Viterbi Algorithm for Bi-gram

First step: $\delta_1(q_i) = P(W_1 = q_i)$

Induction step: $\delta_t(q_i) = \max_{W_{1:t-1}} P(q_i|W_{t-1})P(W_{1:t-1}) = \max_{q_j} P(q_i|q_j)\delta_{t-1}(q_j)$

Final step: $\delta_t(q_i) = \max_{W_{1:t-1}} P(W_1, \dots, W_{t-1}, W_t = q_i)$

簡言之，就是從一個長度為 T 的字串的左邊往右邊不斷累積 max_log_prob [t]。字串的第一個字為<s>，假設在中間第 i 個字遇到的是中文字，根據左邊是<s>、中文字或注音分為三種情況：(1)max_log_prob [i]=0 (2) max_log_prob[i]= max_log_prob[i-1] (3)先用字典找出所有注音可能對應到的中文字，再藉由 language model 找出兩個字相連的機率，然後加上 max_log_prob [i-1]，取其中的最大值便是 max_log_prob[i]。假設在中間第 i 個字遇到的是注音，那麼一樣可以分成三種情況，算法類似就不贅述。在計算 max_log_prob [i]的同時，也要把解答記錄下來。

3.觀察與遇到的困難

(1)Mydisambig 跑出來的結果在十一筆測資上皆與 SRILM 的 disambig 做出來的結果一模一樣，證明程式沒有問題。不過瀏覽 output file 可以發現有些字轉換出來仍然是錯誤的，如 10.txt2 第十二行「全國性勹放」會變成「全國性膀放」，原因是因為在 language model 裡，儘管「播放」的機率為-1.1，「膀放」的機率只有-1.7，然而「性膀」出現機率為-2.95，「性播」為 OOV，視為<unk>出現機率-5.51，加總起來「全國性播放」的機率為-5.62，「全國性膀放」的機率為-4.65，故 output 便是錯誤的。最好的解決辦法就是將這些 OOV 加入到 language model 內，便不會再答錯了。

(2)在將程式全部寫完拿去運行時，發現選擇的替換字都是錯誤的，仔細 debug 後發現在某些狀況下，從 language model 取出來的值是錯誤的，再追查之下發現原來是查詢時輸入的字串有誤，當時儲存一個字的字串時，只有用 char[2]，但是字串最後面還需要一個 char 去存"\0"，改成 char[3]之後便解決了這個問題。