**DSP HW3**

**B05902001 資工三 廖彥綸**

一、架設環境

資工系工作站(i686-m64)

二、檔案說明

**result1/** 使用SRILM disambig 產生的結果

**hw3.sh** 第一部分的shell code，需要有原始的資料檔corpus.txt、lm.cnt、separator\_big5.pl、Big5-ZhuYin.map、未分割的testdata/X.txt、執行檔ngram-count(來自SRILM)輸出seg\_corpus.txt、testdata/seg\_X.txt、bigram.lm、ZhuYin-Big5.map和result1下的結果。(這次不會使用)

**hw3\_my.sh** 第二部分的shell code，配置和上述相似 (這次不會使用)

**Makefile** 作業主要執行的檔案

**mapping.py** 執行Big5-ZhuYin.map 到ZhuYin-Big5.map

**mydisabig.cpp**

三、執行方式

1. bigram.lm, Big5-ZhuYin.map, testdata 放到目錄下

2. make MACHINE\_TYPE=$機器型號 SRIPATH=$srilm-1.5.10的路徑 all

3. make map(目錄下有Big5-ZhuYin.map時輸出ZhuYin-Big5.map)

4. make MACHINE\_TYPE=$機器型號 SRIPATH=$srilm-1.5.10的路徑 run (在result2/下產生結果)

四、其他

1. Makefile中已提供result2/下的預測結果，testdata目錄下必須要有檔案X.txt

(X = 1 to 10)。若要直接執行mydisabig.cpp以下列格式

**./ mydisabig -text $file -map $map -lm $LM -order $order > $output**

2. mydisabig 和disabig輸出的內容相似度約90%，未能做到完全模仿，正確度比較則各有對錯。

3. mapping.py以python撰寫可以直接得到ZhuYin-Big5.map

**python3 mapping.py**

五、mapping.py

1. 讀寫使用encoding = big5hkscs 如果只使用big5會有部分字遺失。

2. 創建字典，字或注音為key，整段可能結果為value。

六、mydisabig.cpp

1. 使用Ngram.h函式庫讀入bigram.lm

2. 建立map，讀入ZhuYin-Big5.map，字或注音為key，整段可能結果為value的vector，同時去除tab、空白鍵、換行符號等。

3 讀入testdata，繪出所有可能性(注音符號用一串vector表示)。

4. 應用Viterbi尋找路徑。

5. 應用Viterbi的backtrack，找出最大機率的路徑。

6. 回溯的路徑放入stack再依序印出。