數位語音概論 HW3　　b03705006 資管三 侯舜元

Environment

Ubuntu 16.04 / Kernel: 4.8.0-52-generic
gcc version 5.4.0 20160609 (Ubuntu 5.4.0-6ubuntu1~16.04.4)
Python 3.6.0 (For mapping.py to do the mapping)

How to compile

make　　make all, compile mydisambig.cpp
make map　　compile mapping.py and generate ZhuYin-Big5.map, executes the following script
`python mapping.py Big5-ZhuYin.map ZhuYin-Big5.map`

How to run

make run　　run and generate result2/1~10.txt

```
run:
    @#TODO How to run your code toward different txt?
    @for i in $(shell seq 1 10) ; do \
        echo "Running $$i.txt"; \
        ./mydisambig -text testdata/$$i.txt -map ZhuYin-Big5.map -lm bigram.lm -order 2 > result2/$$i.txt; \
    done;
```

fixed the input file ZhuYin-Big5.map and bigram.lm, and execute the script above to generate the result.

What I have Done

> mapping.py

mapping.py 負責將 Big5-ZhuYin.map map 到 ZhuYin-Big5.map 。先做注音 prefix 的 mapping，再做字到字的 mapping。

> mydisambig.cpp

預處理

先 load ZhuYin-Big5.map 建好對應關係的 dictionary *mMap*，以便之後做注音的 mapping。
將每一行字串切好 push 到 vector 裡，比較需要注意的是 big5 的中文字是 2byte，再切的時候要捨去不要的空白。另外一個句子前後也加上 <s>, </s>。

Viterbi

對每一個句子做 viterbi，*getBigramProb()* 回傳的是對數後的機率，故我們可以直接將其數值加總起來，取最大的 score 就是最大的機率。方法示意圖如下圖，並將每一個字跟字之間的 path 記錄下來。等全部算完每個字可能的 score 後，就從最後一個字</s>開始 backtrack 回去，即可得到我們要的 sequence。

(圖參考自：http://www.phontron.com/slides/nlp-programming-en-04-hmm.pdf)

結果觀察

結果其實比我想像中的還要好，基本上日常用語都能正確的顯示，某些地名例如『由 ㄒ ㄍ 新 城 電 台 主 辦』→『由 香 港 新 城 電 台 主 辦』也能夠正確對應到。而某些我自己也不是很確定的句子，如『連 歌 迷 ㄧ ㄕ 最 多 最 有 勁』，也能對應出合理的結果『連 歌 迷 也 是 最 多 最 有 勁』。

但是在專有名詞上就顯的比較弱一些，例如一些人名、公司名等（『華視』→『華社』）。或是一些日常上較少用的詞語『殊榮』→『逝榮』。