HW3 Report

B06602037 徐子程

What I observed ?

不論是 srilm 的 disambig 或 mydisambig，有些詞會沒有辦法正確還原，例如：「華視」會被解讀成 「忽視」；「大ㄐ」被解讀成「大金」。可能是 corpus 不足所造成。

Testing Environment

OS : Ubuntu 18.04.3 LTS
CPU : Intel i7 i7-8550U
RAM : DDR4 2133 12G

經測試給定之 test data 1~10，所花費時間如下

disambig

real 0m35.375s
user 0m35.219s
sys 0m0.117s

mydisambig

real 1m33.144s
user 1m32.951s
sys 0m0.136s

顯然，效能有些差距。

What I have done ?

我實做了 ZhuYin to Big5 mapping 和 bigram decoding (mydisambig)。與 srilm 套件的 disambig 比較，對於同一份輸入檔案，可獲得完全一致的結果。

其中實做上比較困難的部份是要將 HW1 的 Viterbi 轉移到這個作業上使用，以及衍生出決定要用何種資料結構的問題，不過也藉此機會重新瞭解 Viterbi 其中的原理，再回來寫就比較容易。

map 實做：

1. 逐行讀取中文與其對應到的注音，並將 Big5 字元轉為 unsigned 16bit integer。注音 ㄅ~ㄦ作為 key，對應到的中文字作為 value，存入 hash table 中。
   (`unordered_map<unit16_t>, vector<unit16_t> *`))
2. Traverse 整個 unordered map，逐一寫入注音與其對應到的中文字到輸出檔案。

mydisambig 實做：

1. 將 ZhuYin to Big5 mapping 檔案存入 hash table。
   (`unordered_map<string, vector<string> *`))
2. 利用 srilm 套件讀取 language model。
3. 逐行讀取 segemented file，進行 Viterbi，得到 decoded 結果，寫入輸至出檔案，直到將所有的行數讀完。

為了減少記憶體的使用，增加效率，每個注音的 mapping 只存一次，之後使用指標存取（指向 unordered map 內的 `vector<string> *`），不重複儲存。

另外是使用 srilm 套件要注意的細節，例如：要在每一個句子頭尾插入<s>, </s>，才能使用。以及其內部定義的機率是以 logrithm 表示，因此計算上要以相加代替相乘。

作業完成也要感謝與我討論的同學：
b06901180 鄭謹譯, b06602047 蔡宜倫