

姓名: 楊宗桓

學號: b07901169

A. what I observed:

◆ disambig vs. MyDisambig:

我利用指令 `diff fileA fileB > out` 來確認 mydisambig 跟 disambig 對 testing data 預測的結果是否一致

結果:

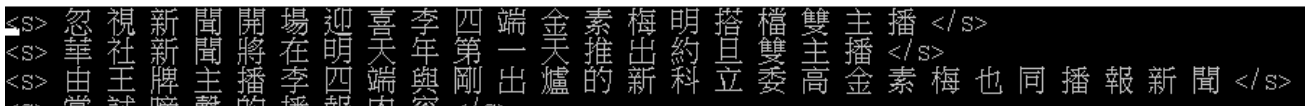
1. 對 example: 兩者完全相同
2. 對 1.txt(如下圖): 上者為 mydisambig 的預測，誤差產生的原因在於注音文出現在最後一個字，而我去查了 model 中這幾個字的機率: 時止 -3.001903 時指 -1.983562 止</s> -0.5525255 指</s> -2.70555，發現我的 model 因為不考慮字尾(也就是</s>)而只算到最後一個字的機率，但若將字後接</s>的機率也考慮進去的話，下面那個預測才是比較合理的(有較高的機率)。修正方法: 只要在跑 viterbi 時句尾記得加上</s>即可



< <s> 有別於以為一貫的摩的時指 </s>
> <s> 有別於以為一貫的摩的時止 </s>

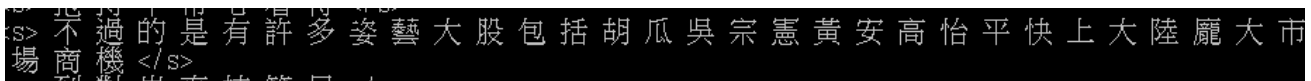
3. 對 2.txt: 一模一樣
4. 綜合以上，基本上只要考慮進去<s>、</s>，mydisambig 靠 viterbi algorithm 就可以達到跟 disambig 一樣的準確度

◆ 預測分析:



<s> 忽視新聞開場迎喜李四端金素梅明搭檔雙主 </s>
<s> 華視新聞開場迎喜李四端金素梅也同播報新聞 </s>
<s> 忽視新聞開場迎喜李四端金素梅也同播報新聞 </s>

1. 這是 mydisambig 對 1.txt 分析的前三行，可以很明顯看到第一行就出錯了，應該是華視新聞而不是忽視新聞，但這個分析結果也合理，因為忽視(-0.5303668)出現在 corpurs 的機率比華視(-2.496953)高太多，以致於 Model 將視判定成忽視而不是華視，若要修正這個缺點，可以把 n-gram 的階級再拉高，例如 4-gram 下華視新聞的出現機率相當然爾會比忽視新聞高(前提是有適當的 corpus)。



<s> 不過的是有許多姿藝大股包括胡瓜吳宗憲黃安高怡平快上大陸麗大市 </s>
<s> 場商機 </s>

2. 這是 mydisambig 對 example.txt 分析的第 23 行，前面幾個字應該要被判定為"不過倒是"但卻被判斷為"的"，可能是因為"的"這個字出現的機率太頻繁了，以致於就"的是"出現機率較"倒是"小，最後還是判定為"的是"，修正方法在於當 language model 建立時應該要對機率過高的字做處罰(例如稍微降低機率)。

B. What I have done

- ◆ Install SRILM on my machine(i686-m64)
- ◆ Generate segmental data
- ◆ Use SRILM to generate language model(order=2)

- ◆ my own Makefile
 - 1. required library
 - 2. Make map
- ◆ Mapping.py
- ◆ Mydisambig.cpp
 - Used library: (from SRILM)
 - a. File.h
 - b. Ngram.h
- ◆ Use Disambig and Mydisambig to generate the result of 1.txt – 10.txt
- ◆ Compare the result and improve my Mydisambig