數位語音處理概論 HW3

R07922004 資工所 吳星耀

A. 環境設置： OS: Ubuntu 16.04 LTS

B. 執行方式：

Part 1:

./separator_big5.pl corpus.txt >corpus_seg.txt

./separator_big5.pl testdata/xx.txt >testdata/seg_xx.txt

SRIBINPATH = $(SRIPATH)/bin/$(MACHINE_TYPE)

./$(SRIBINPATH)/ngram-count –text corpus_seg.txt –write lm.cnt –order 2

./$(SRIBINPATH)/ngram-count –read lm.cnt –lm bigram.lm –unk –order 2

./$(SRIBINPATH)/disambig –text seg_xx.txt –map ZhuYin-Big5.map –lm bigram.lm –order 2 > $(output)

Part 2:

1. Use makefile in R07922004 to compile:

make clean (clean previous files)

make MACHINE_TYPE=i686-m64 SRIPATH=/home/ta/srilm-1.5.10 all

2. Use the command below to build ZhuYin-Big5-map:

make mapping

or:

python3 mapping.py Big5-ZhuYin.map ZhuYin-Big5.map

3. Use the command below to execute mydisambig:

make MACHINE_TYPE=i686-m64 SRIPATH=/home/ta/srilm-1.5.10 LM=bigram.lm run

or***:***

./mydisambig –text $(testfile) –map $(ZhuYin-Big5.map) –lm $(LM) –order $(order) > $(output)

,where $(testfile) is the input file which you want to decode,

$(ZhuYin-Big5.map) is the map created by mapping.py,

$(LM) is the language model which can be a **bigram** or **trigram** model,

$(order) is the number determining if the language model is 2 or 3-gram.

C. 結果分析：

由 result1 中 disambig decode 出的結果可以發現，bigram 與 trigram 的結果都無法完美的還原實際答案，其中 trigram 做出的結果比 bigram 為佳：以 testdata/example.txt 為例：從句子為單位來看的話，bigram 的 decode 結果有 24 句與標準答案不同，錯誤率為 48%，而 trigram decode 的結果 50 句中有 20 句與標準答案不同，錯誤率為 40%。

如果從注音為單位來看：example.txt 中一共有 658 字，其中有 253 個注音，bigram decode 的結果與正確答案相比，不同的字數共 40 個，也就是錯誤率為 40/253 = 15.81%；trigram decode 的結果與正確答案相比，不同的字數共 28 個，也就是錯誤率為 40/253 = 11.07%，有明顯提升。因為 Mydisambig 產生的結果與 disambig 相同，不另外做分析。