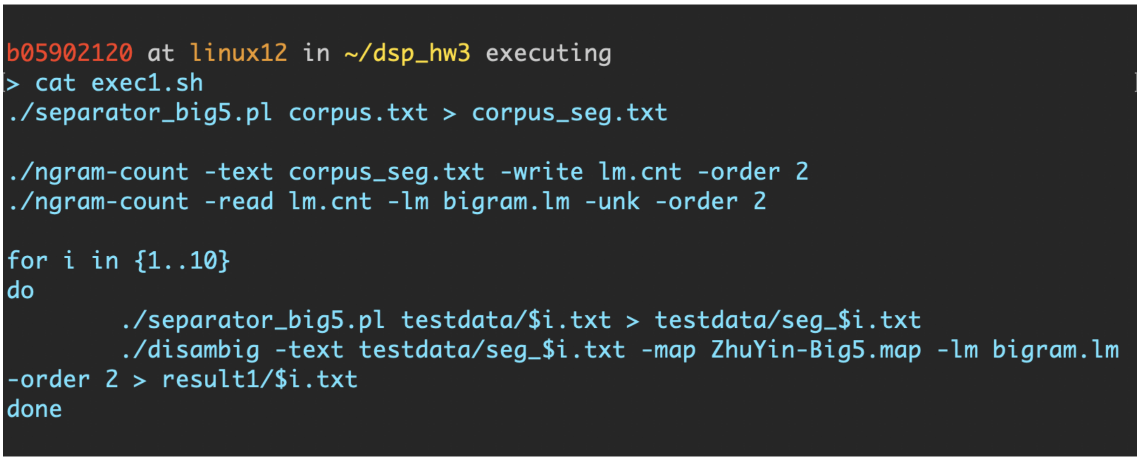
Homework #3

B05902120 / Yu-Ting, TSENG

Dec 12, 2018

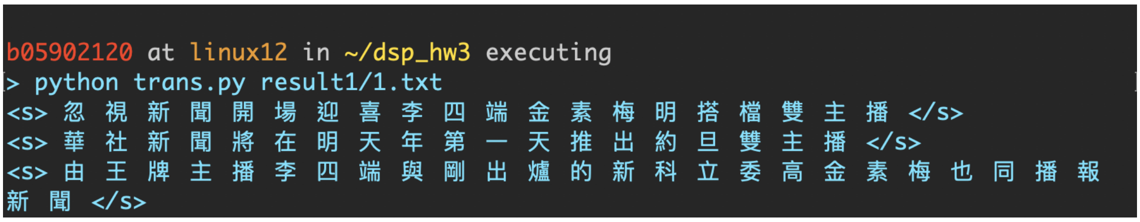
Basic Execution

1. First, setup `SRILM` on "CSIE workstation" so that we can use it.
2. Build `ZhuYin-Big5.map` by the command `python BuildTable.py`.
3. Separate the training data by space and train a language model (bigram).
4. Separate the testing data by space, input it to the language model and get the result.



```
b05902120 at linux12 in ~/dsp_hw3 executing
> cat exec1.sh
./separator_big5.pl corpus.txt > corpus_seg.txt

./ngram-count -text corpus_seg.txt -write lm.cnt -order 2
./ngram-count -read lm.cnt -lm bigram.lm -unk -order 2

for i in {1..10}
do
        ./separator_big5.pl testdata/$i.txt > testdata/seg_$i.txt
        ./disambig -text testdata/seg_$i.txt -map ZhuYin-Big5.map -lm bigram.lm
-order 2 > result1/$i.txt
done
```

5. Write the shell script to execute above mentioning things and get the result by command `python trans.py result1/$i.txt`. (`trans.py` is a script to decode file content with Big5)

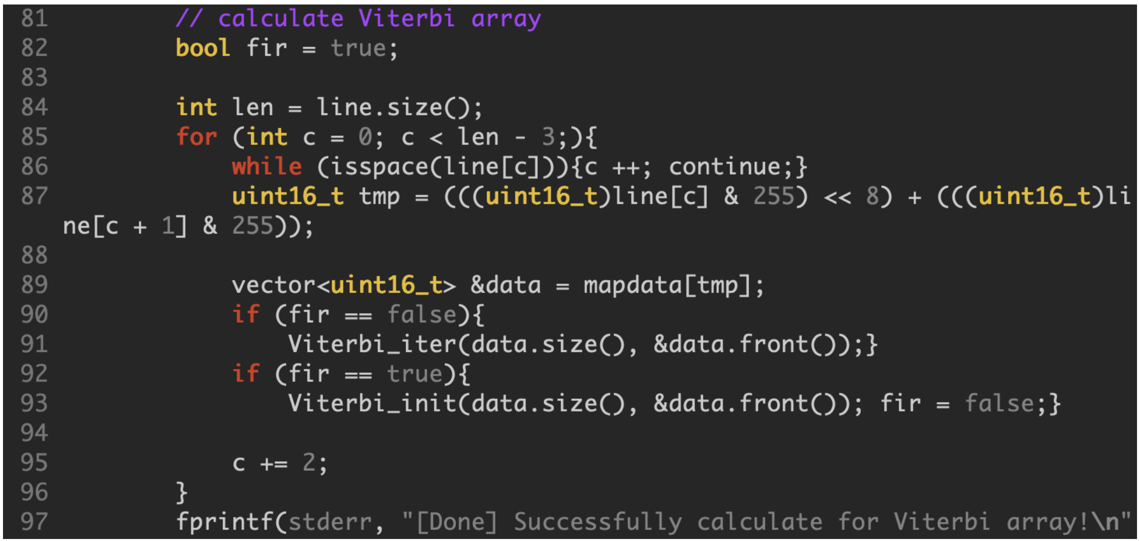


```
b05902120 at linux12 in ~/dsp_hw3 executing
> python trans.py result1/1.txt
<s> 忽 視 新 聞 開 場 迎 喜 李 四 端 金 素 梅 明 搭 檔 雙 主 播 </s>
<s> 華 社 新 聞 將 在 明 天 年 第 一 天 推 出 約 旦 雙 主 播 </s>
<s> 由 王 牌 主 播 李 四 端 與 剛 出 爐 的 新 科 立 委 高 金 素 梅 也 同 播 報
新 聞 </s>
```

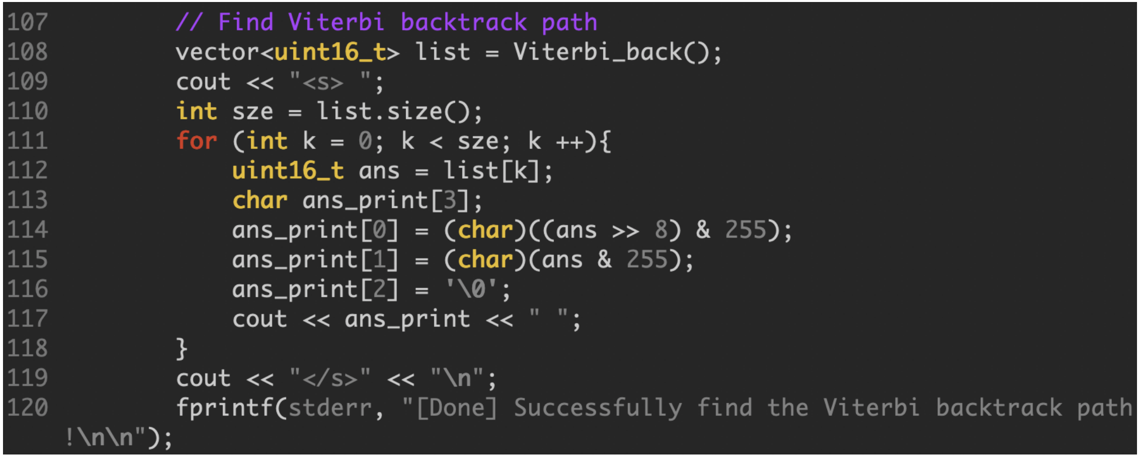
Advanced Execution

1. Write our own disambig program – `my_disambig.cpp`.

2. Read the input line by line, store characters by `uint16_t` and execute Viterbi.

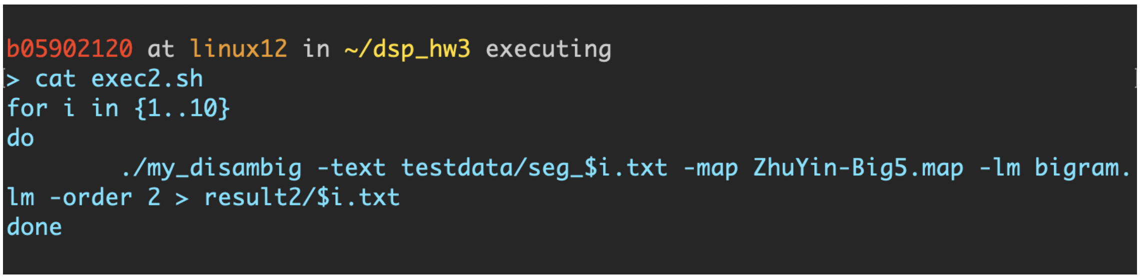


```
        // calculate Viterbi array
        bool fir = true;

        int len = line.size();
        for (int c = 0; c < len - 3;){
            while (isspace(line[c])){c ++; continue;}
            uint16_t tmp = (((uint16_t)line[c] & 255) << 8) + (((uint16_t)line[c + 1] & 255));

            vector<uint16_t> &data = mapdata[tmp];
            if (fir == false){
                Viterbi_iter(data.size(), &data.front());}
            if (fir == true){
                Viterbi_init(data.size(), &data.front()); fir = false;}

            c += 2;
        }
        fprintf(stderr, "[Done] Successfully calculate for Viterbi array!\n"
```



```
        // Find Viterbi backtrack path
        vector<uint16_t> list = Viterbi_back();
        cout << "<s> ";
        int sze = list.size();
        for (int k = 0; k < sze; k ++){
            uint16_t ans = list[k];
            char ans_print[3];
            ans_print[0] = (char)((ans >> 8) & 255);
            ans_print[1] = (char)(ans & 255);
            ans_print[2] = '\0';
            cout << ans_print << " ";
        }
        cout << "</s>" << "\n";
        fprintf(stderr, "[Done] Successfully find the Viterbi backtrack path!\n\n");
```

3. Compile the program by the command `make` and execute by `./my_disambig -text seg_$i.txt -map ZhuYin-Big5.map -lm bigram.lm -order 2`



```
b05902120 at linux12 in ~/dsp_hw3 executing
> cat exec2.sh
for i in {1..10}
do
        ./my_disambig -text testdata/seg_$i.txt -map ZhuYin-Big5.map -lm bigram.lm -order 2 > result2/$i.txt
done
```

Other Notes

- My program is able to do by trigram in order to enlarge the accuracy.
- The second word is just considered by bigram while other is considered by trigram.