Report

R10922132
吳峻銘

1. environment

software:使用提供的 docker 跑，並且用 vs code 來連接 docker 做編輯
hardware : macbook pro, 16GB ram

2. implement

mapping.py
使用 python 基本的讀檔操作就好，但 encoding 的部分要記得改成 big5 hkscs
mydisambig.cpp
使用 srilm 內建的讀檔操作會省掉很多步驟接著就實作 viterbi 就好

```
int main(int argc, char *argv[])
{
    VocabMap map(voc_key, voc_value);
    Ngram lm(voc, 2);

    File mapfile(argv[2], "r" );
    map.read(mapfile);
    mapfile.close();

    File lmfile(argv[3], "r" );
    lm.read(lmfile);
    lmfile.close();

    File inputfile(argv[1], "r");
    FILE *fp = fopen(argv[4], "w");

    char *inputline = NULL;
    while(inputline = inputfile.getline())
    {
        VocabString word_of_line[MAXLINE] = {};
        char decode[MAXLINE][3] = {};
        int len = Vocab::parseWords(inputline, word_of_line, MAXLINE);

        viterbi(word_of_line, decode, lm, map, len);

        char output[MAXLINE_LEN] = {};
        for(int i=0; i<len; ++i)
        {
            strcat(output, decode[i]);
            strcat(output, " ");
        }
        fprintf(fp, "<s> %s</s>\n", output);
    }
    inputfile.close();
    fclose(fp);
}
```

Viterbi 的部分只有實作 bigram 參考課程網頁 FAQ 上提供的公式

Bigram Derivation of Viterbi

To derive bigram part of Viterbi algorithm, define:

$$\delta_t(q_i) = \max_{W_{1:t-1}} P(W_1, \ldots, W_{t-1}, W_t = q_i)$$

where:

$$\delta_t(q_i) = \max_{W_{1:t-1}} P(q_i|W_{t-1})P(W_{1:t-1}) = \max_{q_j} P(q_i|q_j)\delta_{t-1}(q_j)$$

For $t = 1$, initialize first timestep like:

$$\delta_1(q_i) = P(W_1 = q_i)$$

3. observe

```
<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>
<s> 嘗 試 啼 聲 的 播 報 內 容 </s>
<s> 就 是 高 金 素 梅 自 己 明 天 上 午 應 總 統 府 之 邀 </s>
<s> 參 加 元 旦 升 旗 典 禮 唱 歌 歌 的 新 聞 </s>
<s> 華 社 今 天 召 開 董 事 會 </s>
<s> 總 經 理 周 蓉 參 與 經 營 團 隊 展 現 上 任 一 年 多 達 成 績 單 </s>
<s> 在 節 目 經 營 組 織 重 鎮 市 都 有 新 氣 象 </s>
<s> 在 會 長 董 事 郭 力 昕 提 案 討 論 與 台 商 策 略 聯 盟 的 可 行 性 </s>
<s> 董 事 會 決 定 以 客 源 節 流 </s>
<s> 對 外 積 極 合 作 </s>
<s> 尋 求 電 視 台 中 型 </s>
<s> 忽 視 在 明 年 首 先 提 出 誠 意 十 字 的 新 聞 </s>
<s> 讓 全 國 觀 眾 觀 眾 耳 目 以 新 </s>
<s> 李 四 端 特 別 在 新 的 一 年 開 始 邀 請 高 金 素 梅 搭 檔 雙 主 播 </s>
<s> 其 心 準 備 當 華 社 主 播 初 體 驗 的 她 還 將 美 國 九 十 年 度 最 後 一 天 </s>
<s> 獻 給 了 李 四 端 </s>
<s> 兩 人 在 試 播 搭 檔 後 十 分 有 信 心 </s>
<s> 相 信 一 定 會 在 明 天 發 揮 完 美 表 現 </s>
```

1.txt

基本上大致的 decode 都表現不錯，但有些詞組成句字會怪怪的，像是第一句應該要是「華視新聞」而不是「忽視新聞」，推測是「忽視」的 bigram 分數比「華視」高太多導致最後 decode 出來的結果是「忽視」，這種問題可以透過增 corpus 或是改變 bigram 背後的模型來解決。

4. challenge

用 vs code 當 IDE 時會因為編碼用 UTF-8 導致輸入中文字去 vocab 在編譯時 decode 不出來，後來改變 vs code 的編碼就解決了。

一開始實作 viterbi 演算法時陣列開太大導致 segment fault，所以可以了解實作 viterbi 不能用在太長的語音序列，不然 DP 策略會沒辦法執行。