建立於Bigram的中文斷詞方法

B97901121 電機二 劉彥廷

June 26, 2010

動機

對於中文的斷詞問題，如果使用語法分析會相當複雜。在這個報告裡面，我試圖利用bigram語言模型來進行統計型的中文斷詞。

我首先觀察到中文的詞主要是雙字詞（首先、觀察、...），有一些三字詞（不一定...）。而四字詞除了成語之外，幾乎都可以再分解成更小的單位。因此可以透過強迫分割至最多三字詞的方法，對中文句子進行斷詞。

接下來一個問題是要在什麼樣的位置斷開。令一個中文句子爲$\{c_1, c_2, ..., c_n\}$，利用SRILM的函式庫，我們可以查出$P[c_i]$以及$P[c_i|c_{i-1}]$。由機率可以得到$P[c_{i-1}c_i] = P[c_i|c_{i-1}]P[c_{i-1}]$，我們可以找出整句當中機率最小的連接點，斷開後再利用遞迴的方法對於剩下的兩個部份進行斷詞，直到長度小於設定的斷詞長度。

演算法

```
Split(seq, start, end) {
    if(end - start < MAX_LENGTH) {
        print seq[start..end]
    }
    index = argmin {start<i<=end} (Prob[c[i]|c[i-1]] * Prob[c[i-1]])
    Split(seq, start, index)
    print " "
    Split(seq, index, end)
}
```

首先讀入一個中文字串seq，起點和終點分別是start和end。之後找出連接機率最小的點index，將字串分爲(start,index)和(index,end)。遞迴繼續分割，直到長度小於MAX_LENGTH爲止。

假設每一次分割都在最中間，運行時間的遞迴式是$T(n) = T(n/2) + O(n)$。根據Master Theorem，這個演算法的複雜度是$O(n \lg n)$。

程式實做

```
1  #include <cstring>
2  #include <cstdio>
3  #include <cfloat>
4  #include <string>
5  #include "Ngram.h"
6
7  using namespace std;
8
9  void split(string *wordlist, double *prob, int start, int end
       , int order) {
10   if(end - start <= order) {
11     for(int i = start; i < end; ++i) printf("%s", wordlist[i
           ].c_str());
12     return;
13   }
14   double min = DBL_MAX;
15   int index;
16   for(int i = start+1; i < end; ++i)
17     if(prob[i] < min) {
18       min = prob[i];
19       index = i;
20     }
21   split(wordlist, prob, start, index, order);
22   printf(" ");
23   split(wordlist, prob, index, end, order);
24 }
25
26 int main(int argc, char** argv) {
27   Vocab voc;
28   Ngram lm(voc, 2);
29   int order = atoi(argv[1]);
30   File lmfile(argv[2], "r");
31   FILE* to_split = fopen(argv[3], "r");
32   lm.read(lmfile);
33   lmfile.close();
34   char *line = NULL;
35   size_t len = 0;
36   ssize_t read;
37
38   while((read = getline(&line, &len, to_split)) != -1) {
39     line[strlen(line)-1] = '\0';    // get rid of \n
```

```
40      string wordlist[strlen(line)];  // storing the chinese
           characters
41      double prob[strlen(line)];      // storing prob[c_i, c_i
           -1]
42      VocabIndex prev, now;
43      char *tok = NULL, delim[] = "␣";
44
45      tok = strtok(line, delim);
46      if(tok == NULL) continue;
47      prev = voc.getIndex(tok);
48      wordlist[0] = string(tok);
49
50      // Pushing data into wordlist
51      int i;
52      for(i=1; ;++i) {
53        tok = strtok(NULL, delim);
54        if(tok == NULL) break;
55        wordlist[i] = string(tok);
56        now = voc.getIndex(tok);
57        VocabIndex bi[] = {prev ,Vocab_None};
58        VocabIndex uni[] = {Vocab_None};
59        prob[i] = lm.wordProb(now, bi) + lm.wordProb(prev, uni)
             ;
60        prev = now;
61      }
62      split(wordlist, prob, 0, i, order);
63      printf("\n");
64    }
65  }
```

編譯

在src/Makefile中更改SRIPATH以及MACHINE_TYPE之後，即可make完成編譯。

執行

```
zhsplit [n] [bigram] [file]
```

n代表斷詞後最長詞的長度，bigram爲SRILM格式的統計資料，file是已經將每個單字依空白分開的待斷詞文件。

程式運行

以作業三產生的bigram模型以及測試資料，可以來檢測斷詞的情形。以下分別限制斷詞後最大長度爲2、3、4：

	因應全球化市場國際競爭
4	因應｜全球化｜市場｜國際競爭
3	因應｜全球化｜市場｜國際｜競爭
2	因應｜全球｜化｜市場｜國際｜競爭

由這個例子以及我的測試發現，當限制斷詞後的最大長度爲3的時候，會有最好的斷詞效果。再看一個例子：

	刪除國片放映比例及外片徵收輔導金
4	刪除國片｜放映｜比例及外｜片｜徵收｜輔導金
3	刪除國片｜放映｜比例｜及外｜片｜徵收｜輔導金
2	刪除｜國片｜放映｜比例｜及外｜片｜徵收| 輔導｜金

可以發現在碰到功能字 ‘‘及’’的時候，發生了斷詞錯誤。由於功能字常出現，沒有意義的‘‘及外’’出現頻率比有意義的’’外片’’高，因此造成了這個問題。

結論

- 這個方法可以達到一定的斷詞效果，可以正確斷出大部分的二字詞，以最長斷詞長度爲3的結果最佳。此方法的好處爲訓練容易，只要對於一個語料庫作bigram分析即可。斷詞工作時演算法的時間複雜度爲$O(n \lg n)$，效率可以接受。

- 此方法的缺點爲斷詞結果受到最長斷詞長度的限制，若是3的時候將會強制切斷四字詞。另外碰到功能字的情況將會發生斷詞錯誤。

可能的發展

- 可以用特定領域的語料庫作成bigram來對特定領域的文件斷詞，如此可以提高斷詞的正確率。

- 建立功能字列表(都、及...)，事先將功能字斷開，應該可以避免功能字所帶來的斷詞錯誤。

- 建立四字詞及五字以上詞表，例如成語字典。因爲四字詞及五字以上詞的數量遠小於二字詞與三字詞，同時事先辨認四字詞錯誤的機率並不是很高，我們可以事先找出四字詞並斷開。如此一來就避免了當最長斷詞長度爲3時所造成的強迫斷詞問題。

附錄 - 斷詞結果

斷詞長度＝4

讓他十分 害怕
只 希望自己 明年度 別再這麼 苦命了
演藝 娛樂 產業加入 積極轉型 提升 競爭力
明天就是 年
台灣 將正式 加入世界 貿易 組織
因應 全球化 市場 國際競爭
演藝 娛樂圈 影視 產業 因爲是 文化 商品 並未 全面開放
多 認爲 影響不大
但業者都 積極 轉型 提升 競爭力
提供 閱 聽 大衆多元 優 質的服務 品質 選擇
海峽兩岸 將在明年 先後加入
對台灣在 包括電視 電影 唱片 界都 帶來多 面向 影響
其中 在電視 媒體產業 上
由於 境外 媒體 將大舉 進軍 市場
現有 電視台 都 釐清 自己的 市場 定位與 強化 競爭力
像台視 就 加強 與其他 電視台的 策略聯盟
並落實 改造 以製造 利潤
在電視 綜藝圈中
入會後將 面臨 外來節目 競爭
但 綜藝人多 認爲本土 綜藝 抓得住 觀衆口味
外來 客不見得 懂
抱持平常 心 看待
不過 倒 是有許多 綜藝大哥 包括 胡瓜 吳宗憲 黃安 高怡平 看上大陸 龐大市場 商機
到對岸 主持節目
是否會對 本地 綜藝界 帶來影響 仍待 後續觀察
台灣加入
刪除國片 放映 比例及外 片 徵收 輔導金
曾 讓 電影 人感嘆 前途多艱
加上 外國製片 商來台 拍 電影 機會增多
電影 工業化 也讓人 憂心將 衝擊 台灣本土 電影文化
但 行政院 新聞局 電影 處長江 傳清 認爲
入世 對台灣 電影 產業生態 不至於 產生 太大影響
電影人可 建立新 思維 以 跨國合作 瞄準 全球化 電影 市場
原本 盜版 猖獗 的唱片業
由於已 面對國際 化 競爭
多 認爲 在市場上 影響不大
但卻 肯定入世 會對 智慧 財產權 保護 更多
唱片盜版 預期會比 現在少
但民衆 如何 眞正 養成購買 正 版 商品 的消費 習慣

仍須 長期 宣導教育
綜 觀台灣 影視 產業
在進入上
都 有一定的 自 主 性與 限制
並未全面 市場開放
但加入 如經濟 聯合國後
因應 全球化 市場下
由於外來 演藝 娛樂 商品 在市場 開放 下 自由 競爭
閱 聽 大衆 將有 更多的 商品 選擇權
期待 能帶動 良性競爭
刺激本土 影視 產業 向上提升

斷詞長度 = 3

讓他 十分 害怕
只 希望 自己 明年度 別再 這麼 苦命了
演藝 娛樂 產業 加入 積極 轉型 提升 競爭力
明天 就是 年
台灣 將正式 加入 世界 貿易 組織
因應 全球化 市場 國際 競爭
演藝 娛樂圈 影視 產業 因爲是 文化 商品 並未 全面 開放
多 認爲 影響 不大
但 業者都 積極 轉型 提升 競爭力
提供 閱 聽 大衆 多元 優 質的 服務 品質 選擇
海峽 兩岸 將在 明年 先後 加入
對台灣 在 包括 電視 電影 唱片 界都 帶來多 面向 影響
其中 在電視 媒體 產業 上
由於 境外 媒體 將大舉 進軍 市場
現有 電視台 都 釐清 自己的 市場 定位與 強化 競爭力
像台視 就 加強 與其他 電視台 的 策略 聯盟
並落實 改造 以製造 利潤
在電視 綜藝圈 中
入 會後將 面臨 外來 節目 競爭
但 綜 藝人多 認爲 本土 綜藝 抓得住 觀衆 口味
外來 客 不見得 懂
抱 持平常 心 看待
不過 倒 是有 許多 綜藝 大哥 包括 胡瓜 吳宗憲 黃安 高怡平 看 上大陸 龐大 市場 商機
到對岸 主持 節目
是否 會對 本地 綜藝界 帶來 影響 仍待 後續 觀察
台灣 加入
刪除 國片 放映 比例 及外 片 徵收 輔導金
曾 讓 電影 人感嘆 前途 多艱
加上 外國 製片 商來台 拍 電影 機會增 多

電影 工業化 也讓人 憂心將 衝擊 台灣 本土 電影 文化
但 行政院 新聞局 電影 處長江 傳清 認爲
入世 對台灣 電影 產業 生態 不至於 產生 太大 影響
電影 人可 建立新 思維 以 跨國 合作 瞄準 全球化 電影 市場
原本 盜版 猖獗 的唱片 業
由於已 面對 國際 化 競爭
多 認爲 在 市場上 影響 不大
但卻 肯定 入世 會對 智慧 財產權 保護 更多
唱片 盜版 預期會 比 現在少
但民衆 如何 眞正 養成 購買 正 版 商品 的消費 習慣
仍須 長期 宣導 教育
綜 觀台灣 影視 產業
在進入 上
都 有一定 的 自 主 性與 限制
並未 全面 市場 開放
但加入 如經濟 聯合國 後
因應 全球化 市場下
由於外 來 演藝 娛樂 商品 在市場 開放 下 自由 競爭
閱 聽 大衆 將有 更多的 商品 選擇權
期待 能帶動 良性 競爭
刺激 本土 影視 產業 向上 提升

斷詞長度 = 2

讓他 十分 害怕
只 希望 自己 明 年度 別再 這麼 苦 命了
演藝 娛樂 產業 加入 積極 轉型 提升 競爭 力
明天 就是 年
台灣 將 正式 加入 世界 貿易 組織
因應 全球 化 市場 國際 競爭
演藝 娛樂 圈 影視 產業 因爲 是 文化 商品 並未 全面 開放
多 認爲 影響 不大
但 業者 都 積極 轉型 提升 競爭 力
提供 閱 聽 大衆 多元 優 質的 服務 品質 選擇
海峽 兩岸 將在 明年 先後 加入
對 台灣 在 包括 電視 電影 唱片 界都 帶來 多 面向 影響
其中 在 電視 媒體 產業 上
由於 境外 媒體 將大 舉 進軍 市場
現有 電視 台 都 釐清 自己 的 市場 定位 與 強化 競爭 力
像 台視 就 加強 與 其他 電視 台 的 策略 聯盟
並 落實 改造 以 製造 利潤
在 電視 綜 藝圈 中
入 會後 將 面臨 外來 節目 競爭

但 綜 藝人 多 認爲 本土 綜藝 抓 得住 觀衆 口味
外來 客 不見 得 懂
抱 持平 常 心 看待
不過 倒 是有 許多 綜藝 大哥 包括 胡瓜 吳宗 憲 黃安 高 怡平 看 上 大陸 龐大 市場 商機
到 對岸 主持 節目
是否 會對 本地 綜藝 界 帶來 影響 仍待 後續 觀察
台灣 加入
刪除 國片 放映 比例 及外 片 徵收 輔導 金
曾 讓 電影 人感 嘆 前途 多艱
加上 外國 製片 商 來台 拍 電影 機會 增 多
電影 工業 化 也 讓人 憂心 將 衝擊 台灣 本土 電影 文化
但 行政 院 新聞 局 電影 處長 江 傳清 認爲
入世 對 台灣 電影 產業 生態 不 至於 產生 太大 影響
電影 人可 建立 新 思維 以 跨國 合作 瞄準 全球 化 電影 市場
原本 盜版 猖獗 的 唱片 業
由於 已 面對 國際 化 競爭
多 認爲 在 市場 上 影響 不大
但卻 肯定 入世 會對 智慧 財產 權 保護 更多
唱片 盜版 預期 會 比 現在 少
但 民衆 如何 眞正 養成 購買 正 版 商品 的 消費 習慣
仍須 長期 宣導 教育
綜 觀 台灣 影視 產業
在 進入 上
都 有一 定 的 自 主 性與 限制
並未 全面 市場 開放
但 加入 如 經濟 聯合 國 後
因應 全球 化 市場 下
由於 外 來 演藝 娛樂 商品 在 市場 開放 下 自由 競爭
閱 聽 大衆 將有 更多 的 商品 選擇 權
期待 能 帶動 良性 競爭
刺激 本土 影視 產業 向上 提升