專題報告

Team\_10朱雅 R02944047

**一、構想**

**第一階段:**

任務：在給定評論文章中標記Aspect和Opinion Word並建立對應辭典

構想：建立一定的搜尋規則，自動從文章中搜尋並抓取出符合規則的Aspect和Opinion Word，並根據每個term出現的frequency作為重要程度排名，選擇排名靠前的若干建立辭典。

通過對corpus的觀察，認為可以嘗試的搜尋規則有：

1) POS規則：Aspect大多是名詞，Opinion Word大多是形容詞

2) Dependency規則：Aspect和Opinion Word 通常成對出現（global的opinion word除外）；而且在一個句子中，兩者距離較近，有一定的dependency。

譬如：環境<-不錯，良好->體驗

規則2的構想參考資料：

<http://www.cs.uic.edu/~liub/publications/computational-linguistics-double-propagation.pdf>

**第二階段:**

任務：根據已有的訓練語料，在新給定的評論文章中標記每個Aspect和整篇文章的極性（正或負）

構想：首先根據已標記極性的訓練語料判定opinion word辭典中每個opinion word的固有極性（正、負、中性），再用第一階段的方法找出測試語料中每篇文章的aspect以及修飾這個aspect的opinion word，然後由這個opinion word的極性和這個詞前面的副詞（如“不”）來決定對應aspect的極性（正或負）。

對於隱性的aspect的解決方案構想：

1) 根據對訓練語料和給定aspect辭典的分析，列出辭典中每個aspect的同義詞，也就是在文章中經常出現的與這個aspect意思相同的詞彙，以及經常在文章中出現但不在辭典內的這個aspect的下級詞彙，例如（“設備”的下級詞彙可能會有“冰箱”，“衛生間”的下級詞彙可能會有“馬桶”）。然後把這些詞彙作為拓展的aspect辭典去測試語料中搜尋，按照同理標記極性，但是用原始的aspect辭典來標記這個aspect。

2) 根據對訓練語料和兩個辭典的分析，列出opinion word經常修飾的aspect，當在測試語料中出現這個opinion word但沒有對應aspect的時候可以標識特定的aspect，譬如“乾淨”對應的aspect可能是“衛生”，那麼當句子“這裡很乾淨”中出現了“乾淨”，就可以標記隱性aspect“衛生”。

整體極性標註的解決方案的兩種構想：

假設aspect極性正=+1，負=-1，當一篇文章中多次出現相同aspect（包括隱性）時，將每次出現的極性相加，若結果為正就認為這個aspect的極性是正，反之為負。

方案1： 將這篇文章中所有提到的aspect的極性相加，結果為正則整體極性是正，反之為負。

方案2： 將辭典中每個aspect作為一個feature，其value為+1或者-1，那每篇文章可以表示成一個50維的vector，然後將整體極性標註看作classification的問題，正極性的label=1，反之=-1。利用訓練語料得到training model，然後用這個model對測試語料進行標記。

**二、方法**

**第一階段:**

1. Based on POS規則：將檔案經過預處理、段詞和詞性標記，再用名詞和形容詞對應的POS tag來搜尋aspect和opinion word
2. Based on Dependency 規則：對經過預處理的檔案進行dependency parsing，再用適合的dependency tag來搜尋。

**第二階段:**

Step1: 對給定的opinon word 辭典標記極性（正、負、中）。分別統計每個opinion word在正、負評論文章中出現的frequency，如果在正評論文章中出現次數較多則判定為正極性，反之為負，如果相同則標記為中性。

Step2: 針對隱性aspect的問題， 建立aspect替代詞辭典和opinion word指向aspect辭典。

利用POS和dependency parsing的結果，可以分別統計：1）每個aspect經常同時搭配的名詞；2）每個aspect經常搭配的形容詞；3）每個opinion word經常修飾的名詞；4）每個opionion word經常同時搭配的形容詞。

然後通過propagation的方法可以得到aspect的替代詞辭典（包括同義詞和它的下級詞彙），然後經過人工刪選可以選定一些有特定意義的opinion word（和它的同義詞），用它們作為標記aspect的備用辭典。

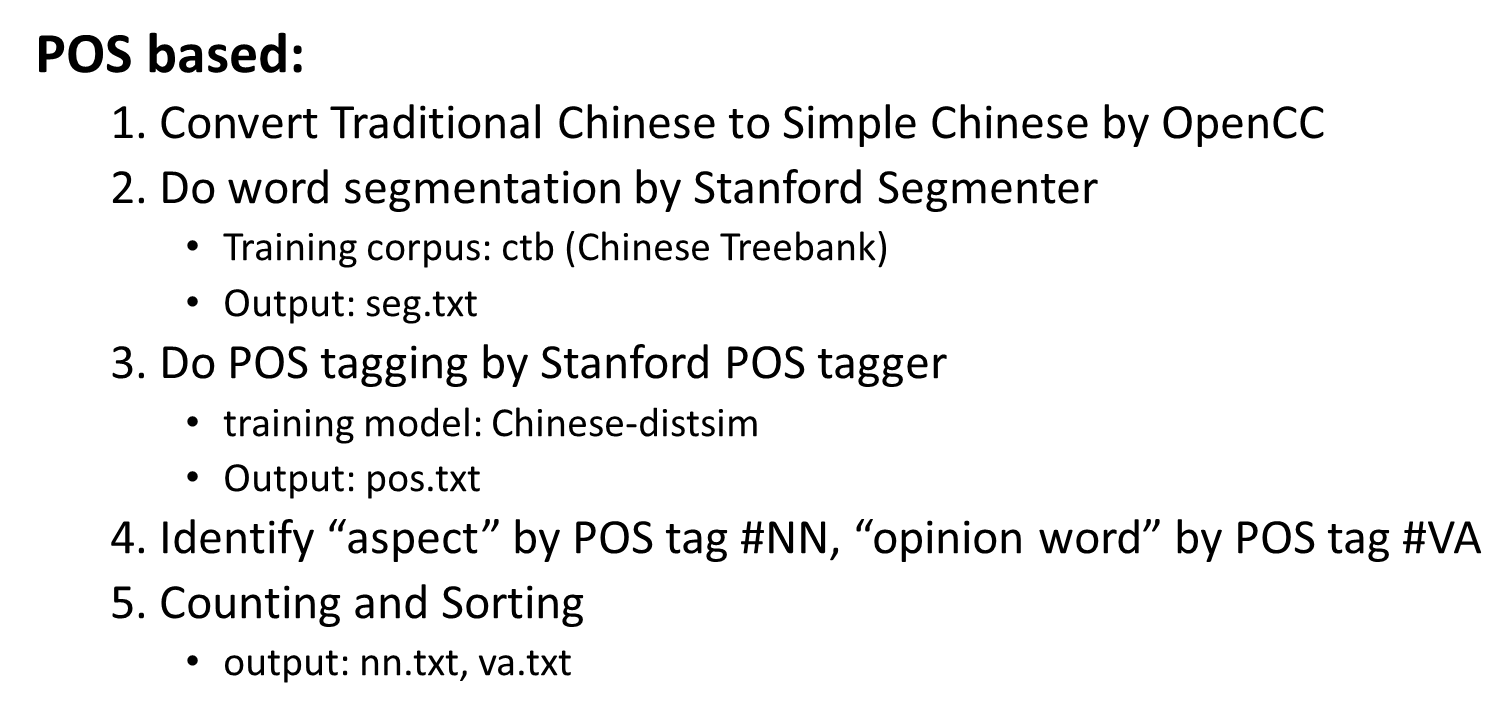
Step3: 用以上4個辭典（aspect辭典、word\_sentiment辭典、aspect\_extension辭典、opinion\_refer\_to\_aspect辭典）+第一階段的方法來刪選aspect以及標註aspect的極性。具體步驟：用aspect，aspect extension，opinion word辭典篩選出aspect和opinion word的candidates，用原始的aspect詞來替代extension的詞，然後用第一階段的方法，也就是POS和dependency還有距離規則來找到成對的opinion word和aspect。

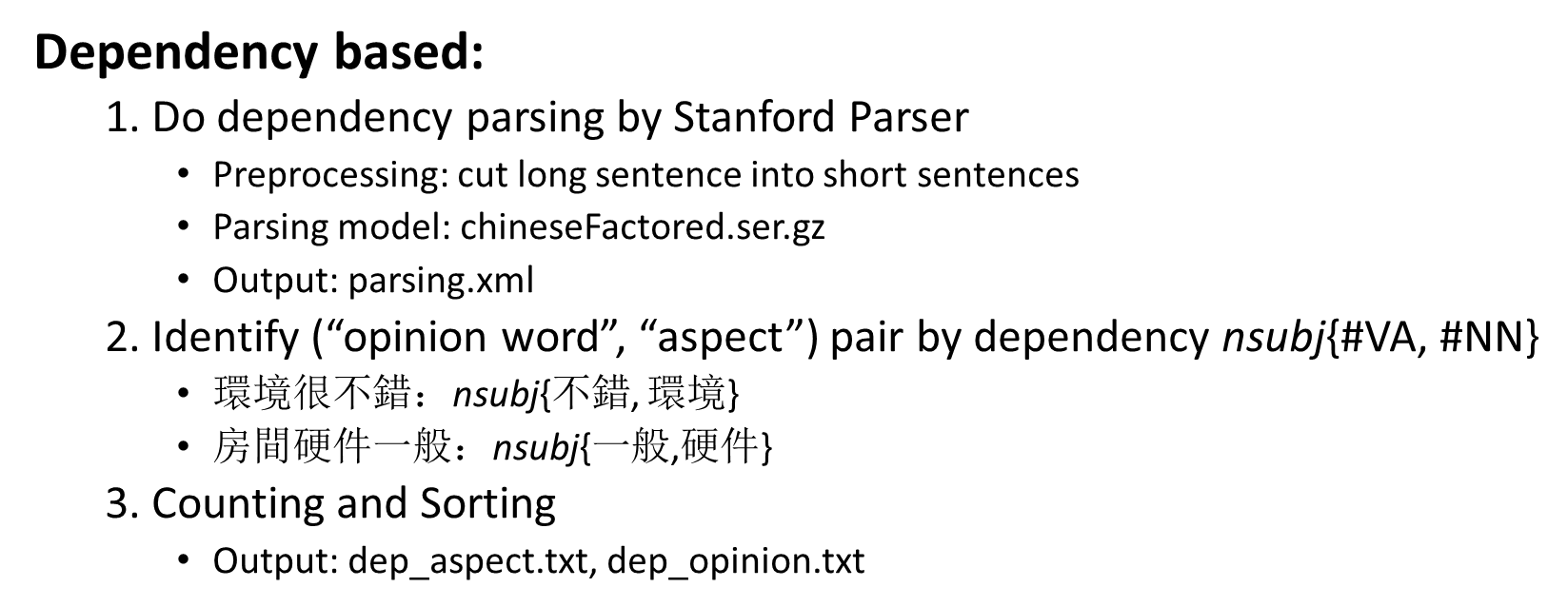
對於落單的opinion word，查詢opinion\_refer\_to\_aspect辭典，加上對應的aspect；對於落單的aspect，觀察看是否有遺漏的term或者rule，加入標記詞性的code中。

Step4: 按照構想中提到的方案標註統計每個aspect的極性和整體極性。

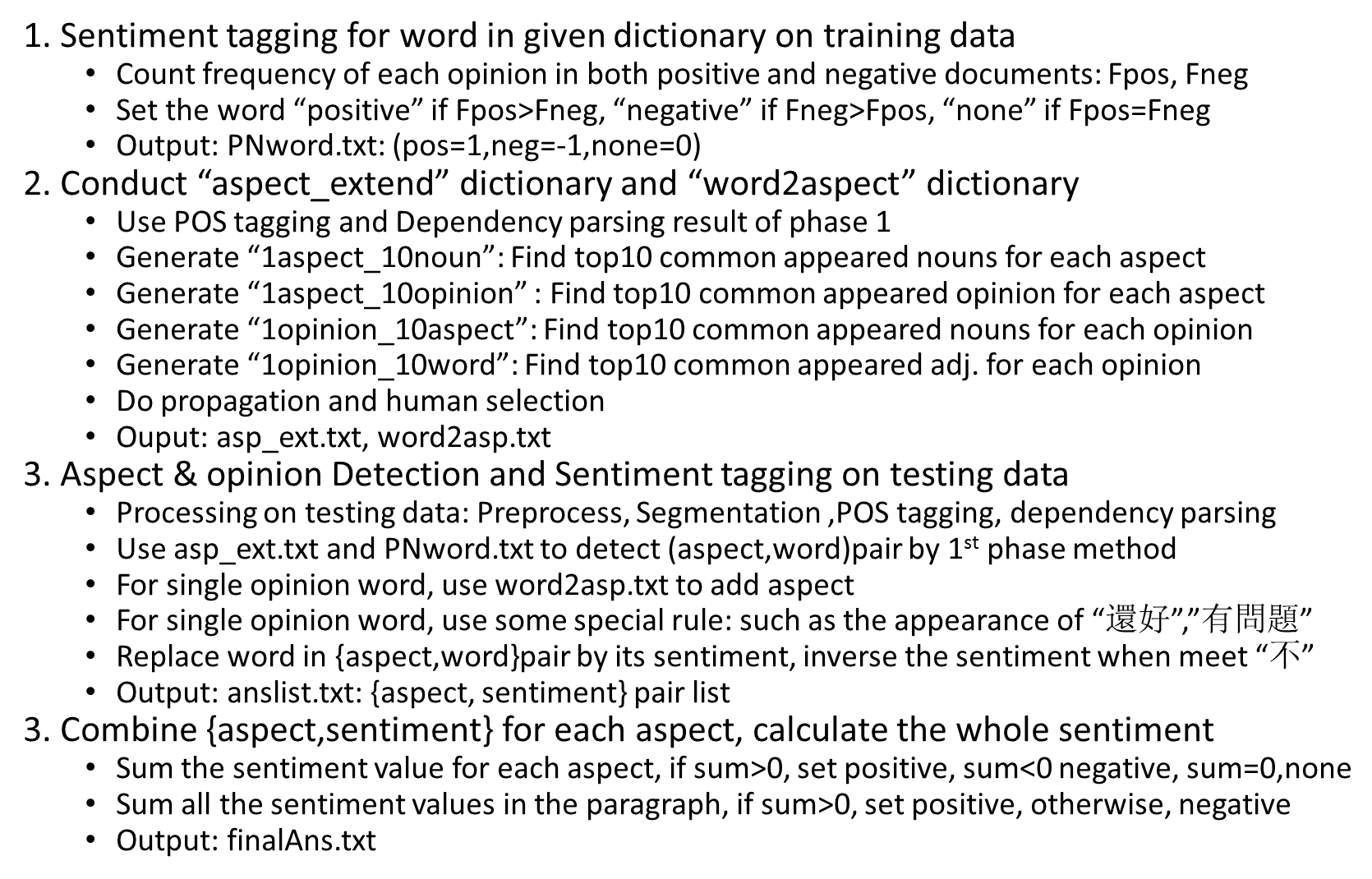
**三、實驗設計**

**第一階段:**

****

****

**第二階段:**

****

**四、效能評估**

**第一階段:**

用POS tag來搜尋的方法，其結果比預想中要好，因為aspect和opinion word的詞性都比較固定，雜訊不會太大，但是精準度還不是很高，因為有些經常出現的noun並不一定就是aspect，譬如“酒店房間很大”和“酒店員工很熱情”中的“酒店”與“房間”、“員工”相比其實並不是aspect。

下圖是兩種方法的對比，左邊是POS方法，右邊是dependency方法。可以看到Dependency的方法在aspect的detection上要更好，將經常被修飾的名詞如“環境”、“位置”等排名提升了，而將經常出現但並未被形容詞修飾的名詞如“人”、“星”等的排名降低了。

但是在opinion word的detection上，dependency的方法效果要比POS差一些，這是因為一些global的形容詞，如“滿意”、“乾淨”會被這個方法遺漏。



**第二階段:**

由於沒有標準答案，所以我人工檢查了前50組，整體極性的正確率是98%，aspect的precision大概在75%～85%左右，recall在80%～90%左右。

這裡分析一下錯誤產生的原因：

1. 段詞錯誤+詞性標註錯誤，因為在這個階段尤其是針對opinion word的分析非常依賴段詞和詞性標註的結果，但是從人工檢查的結果來看，這兩者都帶入了不少錯誤。
2. 極性標註錯誤，因為我的方法是只在training data上計算的，所以首先，用文章的整體極性來標註每個opinion word的極性本來就不太嚴僅，然後那些在正負文章中出現次數相差不大的opinion word的極性標註錯誤的可能性也很大，另外，有些opinion word針對不同aspect的時候的極性可能是不一樣的，無法用一種極性來標記它。所以這部分的造成的錯誤比較大。
3. Opinion word的作用有限。文章中有些對aspect的描述並不是用幾個opinion word，而是會用一些結構不規則的句子，而沒有明顯出現opinion word甚至都沒有出現形容詞。所以我的方法在遇到這些句子的時候是基本無能為力的。（只加入了“……有……問題”這一個常見負面意義的規則）
4. 之前提到的classification的方法由於時間關係沒有嘗試，而將各個aspect的極性相加作為整體極性的方法，是否能符合語料的真正規律則不得而知。

**備用辭典的效果：**

這兩個辭典對找到隱性aspect的貢獻很大，大大提高了recall。但是同時也會帶入一定的錯誤，尤其是word2asp辭典的建立要十分謹慎，比如“方便”這個詞通常用來形容“交通”，但有的時候會出現在“用起來很方便”這樣的句子中，就可能多餘添加了“交通”這個aspect。而像“製冷”雖然是形容“空調”比較多，但是因為通常“空調”很少是隱性的aspect，所以沒有必要為它做辭典。

**五、心得**  
 這次專題實作非常有趣，用的是網路抓取的text，一方面裡面會有很多雜訊（譬如，網頁廣告，商家回覆等需要刪除的內容），另一方面沒有現成的標準答案，需要全班共同標註，這樣肯定也會造成很多雜訊，但是也讓我體會到了處理現實生活中的data是非常困難的。

關於這次第一階段的結果其實有很多不妥的地方，譬如“酒店”和“賓館”，“人員”和“員工”，“位置”和“地方”這樣的同義詞，在標註的時候令人困惑，其實也是沒有必要的，當然更有歧義的是“味道”這個詞，到底是指“氣味”呢還是“美味”呢？還有“客人”、“問題”這些詞看起來並不像是重要的aspect，在實作中我甚至發現“問題”作為opinion word更加合適。而“隔音”這個aspect其實非常重要，卻沒有出現在最後的辭典中，這讓人在第二階段遇到它的時候處理起來也非常困惑。

當然第一階段的成果討論課讓我收穫很大，拓寬了思路，尤其是那個用google搜尋的方法，非常有啟發性也非常實用。借助現成的rich source做NLP也是一種趨勢。

還有一個心得就是發現在NLP的實作中（尤其是closed domain），首先觀察corpus的規律是非常有幫助的，可以避免複雜的操作還能取得更好的結果。但是人工總結規律總是無法涵蓋複雜多變的NL句子，這也是NLP領域研究的難點和樂趣所在吧。