捌、雙語語料庫建構技術

一、如何建構雙語平行語料庫

由於自動對齊雙語文章的句子是計算語言學界近年來積極研究的議題，且牽涉到相當複雜的計算，我們留到下一節敘述。我們先探討是否有中英雙語資料可以不經過複雜的自動句子對齊程序來建立一個電腦輔助翻譯工具。答案是肯定的。有一些雙語資料由於有特殊的段落或句子標記可以輕易的找出對應的句子或段落。Resnik, Olsen, and Diab (1999)就注意到聖經的每一章節段落與詩篇都有數字標記，透過這些標記即可找到對應的句子或段落。類似這樣有句子或段落標記的雙語語料還可以從開放程式碼(Open Source)軟體的說明文件找到一些。

對於沒有明顯段落標記的雙語資料，如果翻譯者在翻譯原文時相當忠實的保留了原文的段落，沒有增加或刪減，那麼我們可以紀錄每一個詞出現在哪幾篇文章的哪幾個段落並做成索引檔，使用者輸入一個詞後，程式查索引檔得到詞出現的檔案及段落位置，即可顯示出包含關鍵詞的段落及對應的翻譯。為了幫助使用者快速找到正確的翻譯，關鍵詞及包含關鍵詞的段落及可能的翻譯以較顯目的顏色標示出來，從使用者的角度來看，這樣的工具雖然在找對應段落的正確率不是特別高，但因為正確的段落對應通常落在程式判斷的段落附近，所以仍然有相當高的實用性。

如前所述，利用段落對應來找對應句並不是一個很可靠的方法，因為翻譯者在翻譯原文的時候多少會做一些增減。另一個困難是中文對於句子的定義相當模糊，有些時候用逗點，有些時候則用句點，不同的人對同一段文字通常就會有不同的標法。這些都是嘗試以中英平行語料庫自動找翻譯對應句時會遭遇的困難。下面是光華雜誌的例子。

(1) 近年來，校園民主的呼聲日切，大學生自主意識越來越高，中國文化中

特有的「尊師重道」、「一日為師，終身為父」倫理觀念，也在時代的衝擊下逐漸解體。

'In recent years calls for democratization of campuses have grown more insistent. Traditional Chinese concepts of the proper ethical relationship between students and teachers, in which students accorded teachers the same level of respect they accorded their own fathers, are dissolving.'

(2) 在大學校園裡，這樣的故事越來越不是特例；許多教師感覺到，經過了社會泛政治化和民主化的洗禮、新「大學法」的頒布實施，和女性主義在校園中蔚然成風的衝擊，大學校園裡，師生之間似乎隱隱形成了角力戰，關係也愈來愈微妙。

'Stories like these are less and less exceptional on university campuses. Manyprofessors have come to believe that a number of factors have laid hidden bones of contention in teacher-student relations in recent years' politicisation of all aspects of life, democratisation in society, the promulgation of the new

"University Law" three years ago, rising feminism. . . . Relations have become much more subtle and complex.'

從上面的例子我們可以發現由於標點符號使用不嚴謹，中文句子有時以逗點有時以句點表示。在找對應句時，如果以英文句子為單位來找中文的對應句將會相當困難。Gao (1998)提出中文的句點，驚嘆號，問號是比句子大的言談單位(discourse unit)的標記，以這些標點符號為單位來找英文對應單位比較容易。

二、如何從平行語料庫中自動找對應句

隨著語料庫計算語言學的興起，研究人員發現可以用機讀雙語詞典或統計方法從平行語料庫自動抽取翻譯對應句。以下簡述幾種常用的方法及所面臨的問題。