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。

（一）以機讀雙語詞典找對應句

機讀英漢電子辭典可以用來找平行語料庫中英文詞的對應，進而猜測其句子的對應。作法又可以分為精確匹配與部分匹配兩種，前者只能找到很有限的翻譯對應，而後者雖可找到較多的翻譯對應，但其中與上下文不符合造成對應錯誤的情形相當多。我們將在後面提出實際的例子。

（二）統計方法

統計演算法的優點在於只需要大量語料庫不需要機讀辭典或語言知識即可找出句子的對應。統計的方法有兩類。一類是直接利用句子的長度關聯性的假設(Brown et al. (1991), Gale and Church (1993))，也就是如果原文某一句較長，那麼翻譯的句子應該也會較長，再利用動態規劃的技巧(dynamic programming)找出哪一句最有可能對應哪一句， Brown et al. (1991)及 Gale and Church (1993)利用加拿大國會英法雙語資料(Hansard)找出段落的標記後找句子對應，正確率在 93%以上。Gao (1998)在實驗後發現上述方法不適用中英雙語語料。如表（一）顯示英法雙語語料(Hansard)有 89%的句子是一對一對應。一對多或多對多的句子對應關係相當少。而利用光華中英雙語語料所做的實驗表（二）顯示無論以中文句點或逗點做為單位，與英文句子一對一的關係都不高，分別是 53%與 35%且多對多的對應關係相當的普遍。因此以句長的關連性來找中英文對應句相當困難。

表（一）Gale and Church (1993)句子的對應關係的機率

Category	Frequency	Prob(match)
1 - 1	1167	0.89
1 - 0 or 0 -1	13	0.0099
2 - 1 or 1 - 2	117	0.089
2 - 2	15	0.011

表（二）光華雜誌中英句子應關係的機率（以中文句點當作單位）

Bead Types	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)
frequency	0.53	0.32	0.06	0.06	0.03

表（三）光華雜誌中英句子應關係的機率（以中文逗點當作單位）

Bead Types	(1, 1)	(1,2)	(1,3)	(1,4)	(1,5)
frequency	0.35	0.38	0.17	0.06	0.04

另一種統計的方法是以詞的頻率與分佈情形來猜測詞的對應，進而找出句子的對應（例如： Kay and Roscheisen (1993), Fung and Church (1994)）。這種方法的缺點是受到頻率，語系,文類,風格等因素的影響很大。再者，根據詞在文章出現位置的分佈情形與出現頻率只能抽取一小部分頻率不高不低的詞彙（頻率太高可能是功能詞很難找到固定的翻譯，頻率太低則無法透過統計得到）。無論是利用統計或機讀電子辭典從中英平行語料庫自動擷取雙語詞彙對應句的困難在於翻譯並非一對一對應，而是隨著上下文語境而變化。如何有效地結合統計與語言知識（例如：辭典、詞類標記、與語法結構）成為研究的重點。

Gao (1998)測試並改良 Fung 與 Church (1994) 的演算法。Fung 與 Church (1994)提出 K-vec 演算法結合互見訊息(mutual information)與 t 值(t-score)等兩個統計方法來計算兩個詞在文件內部區段的共現關聯性。互見訊息(mutual information)是訊息理論(information theory)中的基本概念，計算的方式是兩個事件共同出現的機率除以個別事件出現的機率的積再取以二為底的對數。如果只考慮緊鄰的兩個詞，則可代入下列公式。其中 N 代表語料庫大小（即總詞數），f(x,y) 代表 x 與 y 一起出現的次數 f(x)，f(y)分別代表 x 出現的次數與 y 出現的次數。

互見訊息 $\mathrm{Log}_2 ((P(A \cap B) / (P(A) * P(B)) = \mathrm{Log}_2 (f(x,y)/f(x)*f(y))$

Ken Church (1991)與他的同事率先提出以互見訊息計算詞與詞之間的關聯性(word association)。互見訊息值越高表示詞的關聯性越高，當語料庫夠大時，而互見訊息值大於 1.65，表示這兩個詞常常一起出現。互見訊息可以視為一種相似度測量，T-值則可以視為相異度的測量。T-值(t-score)是計算語言學中常用的統計顯著性的檢定(statistical significance test)，也是 Ken Church (1991)與他的同事率先提出運用在計算語言學，通常與互見訊息搭配一起使用。T-值與標準差和信賴區間

(confidence interval)密切相關。當語料庫夠大時，而 T-值大於 1.65 時表示有 95%的信心證明差異是存在。計算 T 值的公式如下。

$$t = \frac{P(x|y) - P(x|z)}{\sqrt{\sigma^2 P(x|y) + \sigma^2 P(x|z)}}$$

其中 x|y 表示 y 出現時 x 出現的機率。σ 表示標準差。

T 值的計算可以採用下列簡化的公式。其中 N 代表語料庫大小（即總詞數）。f(x,y) 代表 x 與 y 一起出現的次數 f(x)，f(y)分別代表 x 出現的次數與 y 出現的次數。

$$t \approx \frac{f(x,y) - \frac{f(x)f(y)}{N}}{\sqrt{f(x,y)}}$$

Fung 與 Church (1994)的基本的假設是如果有兩篇互相對應的文章，某語言一個詞與另一個語言的一個詞在某些區段一起出現的機率大於個別出現的機率，則它們兩個詞有可能是翻譯。Fung 與 Church 將相對應的翻譯文章均分為 K 個區段（K 為文章長度的平方根），以 K 維向量來紀錄兩個語言中某個詞出現在哪幾個區段，例如在第一區段出現就將對應的向量值設為 1，否則設為 0。詞頻太低與太高的詞都不適合使用此演算法，因為若只出現一兩次的詞即使分佈區段完全相同也很有可能是巧合，而出現很頻繁的詞很可能是功能詞才會在很多區段一起出現，這些都必須先排除掉，否則會影響演算法的精確度。Fung 與 Church 建議使用詞頻在 5 次到 10 間的詞，以 K 維向量（K 為文章長度開平方）來表示其分佈情形之後再利用互見訊息與 t 值來計算頻率相近的中文與英文詞在相同區段一起出現的機率。Fung 與 Church (1994)使用下列聯方表。

表(四) 聯方表

a = k(A B)	b = k(~ A B)
c = k(A ~B)	d = k(~A ~B)

a 表示某個中文詞與英文詞一起出現的區段數， b 表示英文詞出現但中文詞沒有出現的區段數， c 表示中文詞出現但英文詞沒有出現的區段數， d 表示中文

詞與英文詞都沒有出現的區段數。再利用下列稍微修改過的互現訊息與 t 值，其中 P(Vc) 為某一中文詞出現在區段的機率，P(Ve) 為某一英文詞出現在區段的機率。

$$MI(V_c, V_e) = \log_2 \frac{P(V_c, V_e)}{P(V_c)P(V_e)}$$

$$P(V_c) = \frac{a+b}{a+b+c+d}$$

$$P(V_e) = \frac{a+c}{a+b+c+d}$$

$$t(V_c, V_e) = \frac{P(V_c, V_e) - P(V_c)P(V_e)}{\sqrt{\frac{P(V_c, V_e)}{K}}}$$

Gao (1998)使用中研院詞知識庫小組發展的分詞程式處理中文的分詞，並以中英對照之光華雜誌做實驗證明上述方法的精確度受到文類的影響很大，如下表所示，精確度有可能高至 70%也有可能低至 30%以下，此外利用此演算法實際能找到的對應詞相當有限。數千詞長度的對應文章，大多只能找到幾個對應詞。為了提高精確度，Gao (1998)改良 Fung 與 Church (1994)演算法。首先我們計算中文與英文的文章段落數目是否一樣，若一樣我們則將 K 設為段落數，若不一樣則依舊採用原先的定義。我們也合併賓州大學（University of Pennsylvania）與美國暑期語言學院（Summer Institute of Linguistics）所發展的構詞分析程式，將英文所有的名詞，動詞，形容詞換成原型，使計算共現機率時能更準確。此外我們不僅利用文章內部區段一起出現的機率，也收集數十篇對應文章，再以中文詞與英文詞出現在同一篇文章的機率（亦即文件的共現關聯性）來過濾 Fung 與 Church (1994)演算法所得到的結果，將精確度大幅提高至 90%以上。

除了改良 Fung and Church (1994)的 K-vec 演算法，我們也利用機讀英漢電子辭典得到部分翻譯對應，以便推論其它的翻譯對應。一般人認為利用機讀英漢電子辭典即可很容易得到平行語料中的詞彙對應關係，事實上撰寫程式呼叫電子辭

典自動查詢所得到的對應仍然非常有限。主要原因在於（1）一個詞可能有幾個翻譯，以機讀辭典判斷那一個詞對應哪一個詞，必須從上下文找訊息，相當困難，從實驗中我們發現功能詞的意義相當多，利用機讀辭典來得到翻譯對應非常不可靠。（2）利用完全字串匹配（exact string match）所能得到的翻譯對應相當有限。例如字典中 teacher 的翻譯是「教師」，實際上文章可能翻譯成「老師」，若採用部分字串匹配-（partial string match）則可以找到辭典中的翻譯與文章的翻譯有一個字「師」相同。採用部分字串匹配雖可以找到相當多可能的翻譯對應，但錯誤率也相對提高許多。為了解決這個問題，我們先排除最常出現的功能詞「的」。凡是部分字串匹配為「的」的翻譯對應一律排除。接著再找出相鄰兩個英文詞至少各有一個字與辭典翻譯吻合的連續詞。雖然利用上述緊鄰性(proximity)原則找到的詞組翻譯 (translation equivalents)相當有限，但精確度高達 90％以上。透過這一些正確率非常高的對應詞或詞組，我們即可得到某些翻譯句的對應。

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。

（三）採用 Champollion Tool Kit (CTK) 找出中英文對應句

CTK 利用雙語辭典，數字，及簡體中文中的英文詞的對應再加上句長對應的關聯性透過統計演算法找出中英文句對應。我們在實驗中發現 CTK 在簡體中文與英文句對應方面得到不錯的效果。我們透過 Perl 中文簡繁體字的對應程式，將繁體中文轉換成簡體中文再使用 CTK 這個工具程式找出文章裡面中英文句對應。圖一是我們利用 CTK 及其它工具程式所發展的中英平行語料句對應程式的介面，使用者貼入兩篇互為翻譯的中文及英文。圖二十一是這個程式輸出的中英文對應句。

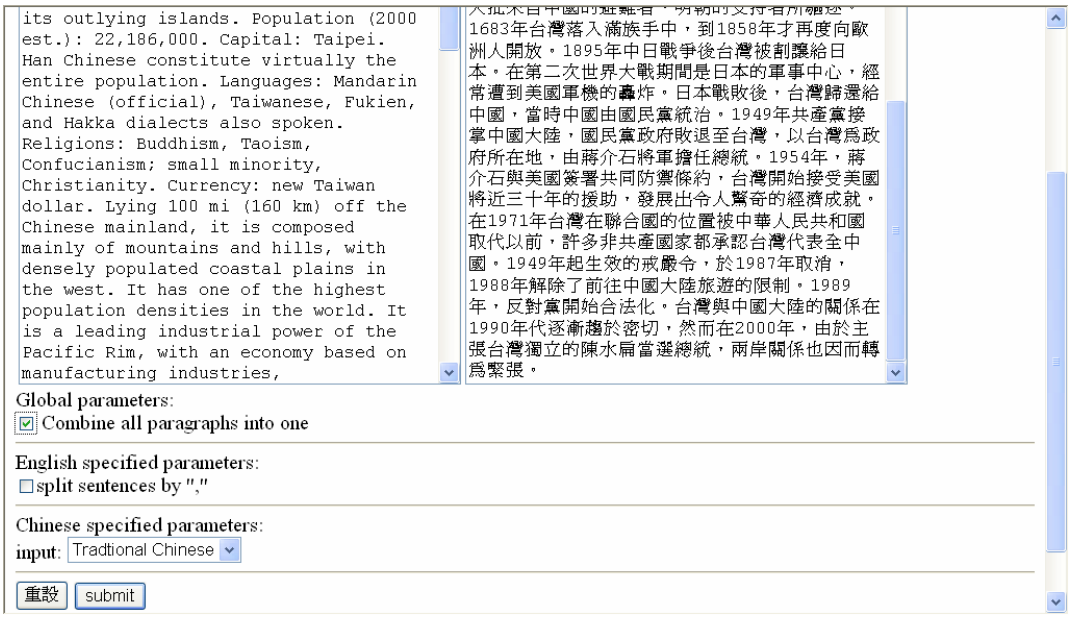
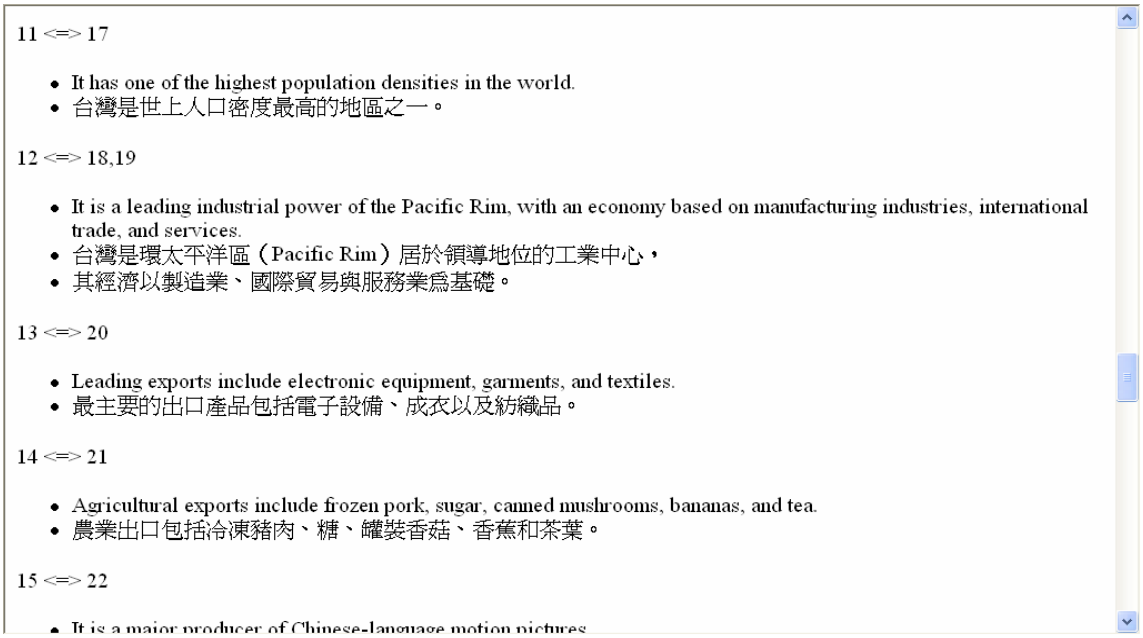


圖　二十 利用 CTK 所發展出來的中英平行語料對應程式介面



圖二十一　中英對應句程式輸出的結果

經過上述程式及分詞程式處理後，我們利用 Lucene 搜尋引擎將語料庫裡面所有中文文章的句子和英文翻譯自動做索引。我們所建構的這個中英雙語語料庫規模相當大，不但包括各種文類，來源語中屬於中文與英文比例接近，因此對-華語學習者或英語學習者而言都是一個很寶貴的資源。