專利雙語語料之中、英對照詞自動擷取

Automatic Term Pair Extraction from Bilingual Patent Corpus

曾元顯　Yuen-Hsien Tseng
國立臺灣師範大學
資訊中心
Information Technology Center
National Taiwan Normal University
samtseng@ntnu.edu.tw

劉昭麟　Chao-Lin Liu
國立政治大學
資訊科學系
Department of Computer Science
National Chengchi University
chaolin@nccu.edu.tw

莊則敬　Ze-Jing Chuang
威知資訊股份有限公司
WebGenie Information Ltd.
terry@webgenie.com.tw

摘要

針對 50 多萬筆的專利中英雙語語料，本文提出兩種翻譯對照詞彙的自動擷取方案，一種是精確導向、另一種是召回導向。在精確導向的方案中，我們提出了一種詞彙擷取方法，並比較了六種詞彙對列作法，以實際資料驗證，得出可供參考的經驗。我們發現 EM（Expectation Maximization）方法效果最好，但其最花時間，也難以找出多對多的同義翻譯。而即便是最差的 MI（mutual information）法，其排序在前頭的正確詞對跟 EM 法不同，因此可以作爲輔助的詞對擷取方法，爲後續合併或混用多種對列方式的研究，開啓了可能性。在召回導向的方案中，我們提出了簡單的想法與有效的實做，可從雙語對列語料庫中召回大量的新詞對，供後續應用，讓既有的上百萬條雙語詞庫，再增加約 20%的新詞對。

Abstract

This paper proposes two approaches to extract translation term pairs from Chinese-English bilingual corpus with more than 500,000 patents. One approach is precision-oriented, in which we compare six term alignment methods. Based on our experiments, we find that the EM (Expectation Maximization) method is the best. However, it is time-consuming and hard to extract many-to-many translations for the same concept. While the MI (mutual information) method performs worst, the term pairs extracted may be totally different from those by EM. This may inspire subsequent researches to study the possibility of hybrid term alignment methods. The other approach is recall-oriented, in which a simple idea was proposed. With an efficient implementation, 20% more term pairs were extracted based on an existing lingual lexicon which already has more than one million term pairs merged from several sources.

關鍵詞：專利語料庫，機器翻譯，專利跨語分析，詞彙對列，新詞擷取

Keywords: Patent Corpus, Machine Translation, Cross-lingual Patent Analysis, Term Alignment, Term Extraction.

一、緒論

世界各主要國家的專利機構（IPO, Intellectual Property Office，或是 PO, Patent Office），如世界智慧財產權組織的 WIPO、歐洲的 EPO、美國的 UPSTO、日本的 JPO、韓國的 KPO、大陸的 SIPO 以及我國的 TIPO，都有專利網站提供免費（或付費）的專利單語檢索服務。然而，在專利誇語言的服務上（特別是內容翻譯），最近才開始陸續提供。

目前台灣已是美國專利核准案中第四大外國申請國 [1]，且台灣專利申請案中有 40%爲外國申請案 [2]，其中比重最大的日本亦曾向經濟部反映希望增設專利英語檢索功能[3]，顯見目前台灣相關市場上對於專利跨語言檢索與翻譯的迫切需求，但目前智慧財產局與民間尚無提供跨語言的專利自動翻譯服務，顯然與世界趨勢與國內需求嚴重脫節。由國科會登錄的研發計畫與人力來看，國內專利分析與翻譯方面的研究非常匱乏，更看不見技術與應用端的上下游整合，顯示出此一問題的嚴重性。

事實上，我國智慧局自民國 87 起即進行本國專利摘要英譯的業務，以便與世界專利機構交換資料，至今完成 33 萬餘件。雖然每年智慧局都編列預算進行專利摘要人工英譯的工作（平均每年約 3 萬篇），至今卻苦無相關的資訊系統提供英譯對照、專利近似文句的譯文範例，使得翻譯過程辛苦而緩慢、翻譯內容與品質難保一致、翻譯的知識無法累積，而不能爲廣大的使用者運用於跨語言的專利檢索與分析。

專利文件的特色，在於技術名詞繁多、艱深，使得翻譯者專業知識的差異，容易造成英譯名詞的不一致。另外，專利中新詞不斷衍生，且散落於不同領域文件，即便是專家，也難以知悉相同概念的所有相關譯名。因此，在能夠進行有效的專利自動翻譯之前，能否獲取大量的專利雙語對照詞彙（aligned term pairs），乃此項自動化服務的先決條件。

本文的目的，在試圖從上述既有的專利中英雙語語料中，擷取出中英對照詞彙，一方面可做爲人工翻譯的參考，以便提高後續翻譯的一致性，一方面也可做爲未來機器翻譯的基礎語料，讓翻譯專家的語言知識，得以變成機器翻譯可依賴的詞彙知識。

二、待解問題

在擷取中英對照詞彙的過程中，預計將遭遇兩種類型的問題，其各有不同的發生時機與解決目標：

問題一：給定一組特定領域的中、英雙語對照語料庫，自動找出該領域的中、英雙語對照詞彙。

問題二：給定一組特定領域的中、英雙語對照語料庫，以及一組專屬該領域的中、英雙語對照詞庫，自動找出未曾出現在詞庫中的詞對。

問題一發生於一開始還沒有該領域（專利）的專屬雙語詞庫時，需要以自力更生（bootstrap）、精確導向（precision-oriented）的方式，找出專屬領域的翻譯詞對（translation term pairs）。一旦專屬的雙語詞彙越來越多時，待解問題將變成問題二的類型，其目標在召回導向（recall-oriented）的來找出更多的翻譯詞對。

實務上，在解決上述問題時，所有可用的資源皆可使用，例如用到既有的非專屬領域雙

語詞庫、單語詞庫、甚至其他的語料庫。當用到這些資源時，問題一與問題二會變得有點模糊，因爲所謂專屬或非專屬領域的雙語詞彙，其實不易界定。在問題一的解法中使用到非專屬雙語詞庫，好像變成問題二。事實上，我們對問題一的解決方法，可以找出問題二解法所無法找出的翻譯詞對。這是因爲問題一的精確導向特性，與問題二的召回導向要求，使得其解法不同，而有相異的結果。

另外，專利文件其文句大多很長，非常不利於機器的自動翻譯。以統計式翻譯範例而言，其先透過翻譯模型（translation model）將每個字的可能翻譯列舉出來，組合成所有可能的英文候選句，再透過語言模型（language model）來評估每個英文候選句的英文符合程度。當文句很長，裡面的詞彙很多的時候，就容易碰到組合爆炸（combinational explosion）或是翻譯時間過長的情形。因此，在 Manning 與 Schutze 的書中（第 490 頁）[4] 甚至建議超過 30 個字的句子，就不給機器作翻譯訓練。但這對專利文件是行不通的。我們認爲，解決的辦法，在降低其組合數。亦即對付的句子雖然長，但只要其「已知如何翻譯的片段越長」，組合出來的候選句就會越少，而仍舊可以有效率的進行機器翻譯。在此所謂「已知如何翻譯的片段越長」，表示我們擷取出來的雙語詞對越長、越多，對長句的專利翻譯，會越有幫助。

基於上述特性，本文提出兩種解決方案，各別對付上述兩類問題。其中我們用到的語料，如表一所示。

表一、本文中使用到的語料庫

語料	數量	使用場合
專利中英摘要語料庫	506510 篇	擷取中英詞對
中文單語詞庫（一般領域用詞）	123280 條詞	中文斷詞
國立編譯館學術名詞	約 160 萬條中、英詞對（註）	驗證擷取的詞對

註：我們取得的國立編譯館的學術名詞有 77 個類別，共 170 個檔。每個檔案裡的格式不同，中、英詞對的標註習慣也不一致，例如：

熱乾 ⇔ Drying，heat	向熱性液晶⇔liquid crystal; thermotropic
黃[土]風 ⇔ yellow wind	液晶顯示 ⇔liquid crystal; display; LCD
向列的(液晶)，絲狀的⇔nematic	液晶顯示器⇔LCD (=liquid crystal display)
鼠【魚+銜】⇔ dragonet	液晶顯示器⇔liquid crystal display (LCD)
蒔蘿【火+青】⇔ cymenes	液晶顯示器⇔liquid-crystal display {= LCD}

在剖析成電腦可用的中英詞對時，常有剖析規則（在同一領域同一檔案中）互相衝突的情形，因而產生不少雜訊（即錯誤的詞對）。但由於數量太多，且專業領域知識有限，目前無法一一對其檢驗進行更正。因此，剖析規則的不同，最後的詞條數也不同。我們後來另有一個版本，約 110 萬條，其用到的剖析規則較多、較嚴、較精確，但可能遺漏較多。

三、解決方案一

針對問題一，我們提出的解決方案，如圖一所示，分下列步驟詳述。

<table>
<tr><td>步驟</td><td>芸香劑方中藥組成製造方法</td><td colspan="3">Method for producing rutin Chinese medicine composition</td><td>所需資源</td></tr>
<tr><td>斷詞</td><td>芸香劑方 中藥 組成 製造 方法</td><td colspan="3">method, producing, rutin, Chinese medicine, composition</td><td>片語擷取</td></tr>
<tr><td>過濾</td><td>芸香劑方 中藥 組成 製造 方法</td><td colspan="3">rutin, Chinese medicine, composition, producing, method</td><td>既有詞庫</td></tr>
<tr><td>對列</td><td colspan="2">Run 1:
Co(芸香劑方, rutin)
Co(中藥, Chinese medicine)
Co(組成, composition)

Co(芸香劑方, rutin)
Co(中藥, composition)
Co(組成, Chinese medicine)

Co(芸香劑方, Chinese medicine)
Co(中藥, rutin)
Co(組成, composition)</td><td>Co(芸香劑方, Chinese medicine)
Co(中藥, composition)
Co(組成, rutin)

Co(芸香劑方, composition)
Co(中藥, Chinese medicine)
Co(組成, rutin)

Co(芸香劑方, composition)
Co(中藥, rutin)
Co(組成, Chinese medicine)</td><td>Run 2:
Co(芸香劑方, rutin)
Co(中藥, Chinese medicine)

Co(芸香劑方, Chinese medicine)
Co(中藥, rutin)

Run 3:
Co(芸香劑方, rutin)</td><td>統計資料</td></tr>
</table>

圖一、中英文詞彙對列方法示意圖

（一）起始步驟：

首先，從專利中英摘要語料庫中，取出每一篇中英對列文件，根據標題內的中、英文用語，進行詞彙對列的計算工作。每一篇中英對列文件，除專利號外，只有標題與摘要。目前只用標題的文句，而不用摘要內的文句，來進行詞彙對列。這是因爲我國專利中英摘要的文句對照情況並不好。常常其中文摘要只有一句，但英文摘要卻有三句。因此，若要用到摘要的文句，來豐富可用的語料，需先進行文句對列的工作，而且不能只對到句子，還要對到正確的子句。然而這項工作並不容易。因此，在目前精確導向的考量下，我們決定只用標題的文句。

（二）斷詞處理：

斷詞處理的目標，是要將中英文對照的詞彙，擷取出來。像圖一中的範例，中文不能只斷出「芸香、劑方」兩個詞，而要整個「芸香劑方」都斷出來，才有可能匹配到英文的「rutin」一詞。反之，英文也是如此。否則就要考慮多對一、一對多、多對多的詞彙對列情況，而這樣會讓對列步驟過於複雜，成效也不見得較好。

因此，在斷詞處理中，要具備擷取片語、新生詞彙或未知詞彙的能力，否則不易找出既有雙語詞庫以外的詞對。爲此目的，我們採用 Tseng 的關鍵詞擷取演算法 [5-7]，其可擷取「最大的」（maximal）重複字串。在此所謂「最大的」，是指字串最長的或出現頻率最高的。由於此演算法沒有用到詞彙知識，僅憑字串的重複出現與其重複出現時左右字詞會有所不同的特性來擷取詞彙，因此，只要在文中出現兩次，即可被擷取出來，詞

彙擷取的門檻很低。另外，其也可以運用於 OCR 的錯誤詞彙擷取（具有正確左右邊界但內含辨識錯誤文字的詞彙） [8]，甚至是 MIDI 數位音樂的關鍵旋律擷取 [9]，可見其可擷取新詞的能力。

如同前述，我們只以標題進行詞彙對列，但在運用此演算法時，我們是將標題與摘要合併起來進行詞彙擷取，以便讓標題中的重要詞彙能夠重複出現，而被擷取出來。

然而，專利摘要的寫法，其第一句常包含整個專利標題。由於此演算法相當貪婪（greedy），造成整個標題常被擷取成單一個詞彙，而無法進行詞彙對列。我們的解決方法，是利用一般詞庫先對中文標題進行斷詞（英文則依空格斷詞），將斷詞後的單一字詞屬於連接詞、代名詞等停用詞者，代換成逗號，從而將整句標題斷開，而不會與摘要中的第一句完全相同。圖二的例子展示了此過程。其中，我們使用的「一般詞庫」，乃從網路上下載，並自行加入 3 千多個新聞詞彙而來，總共有 123280 個中文詞。

原始標題	紗線製法及其組成結構	method for making yarns and constitution structure thereof
斷詞及過濾後的標題	紗線, 製法, , , , 組成, 結構	, , making yarns , constitution structure ,
標題及摘要的重複字串	線捲:4,線紗:3,紗線製法:2,組成結構:2, 管狀:2	making yarns:2, winding:4, yarn:3, weaving:2, constitution structure:2
屬於標題中的重複字串	組成結構, 紗線製法	making yarns, constitution structure

圖二、標題中重要詞彙的擷取過程範例(冒後後面的數字爲該詞彙在該文件的出現次數）

（三）詞彙過濾：

此步驟目的，是要將不可能的詞對過濾掉，以節省後續對列計算的負擔。我們可以用單語詞庫，將一般性的詞彙刪除（因爲可以假設其對應的翻譯詞已知，或容易從他處取得），或是利用既有非專屬領域的雙語詞庫，將已知的詞對刪除，以減少後續不必要的詞彙對列。當然，若沒有任何詞庫資源，此步驟也可以不做，只是會浪費力氣去處理很多不可能的（implausible）詞對。

（四）對列計算：

一旦句對中的詞彙擷取出來，並運用既有資源將不可能的詞對排出後，接下來即可大膽的將中、英詞彙進行盲目的配對。如此累積一篇篇、一句句的詞對後，透過對列分析，正確的詞對，大多會和錯誤的詞對有統計上的區別。常用的對列分析列舉如下：

1. 相互資訊（MI，mutual information）

MI 計算公式爲 [4]：

$$MI(c,e) = \log_2 \frac{p(c,e)}{p(c)p(e)}$$

在我們的應用中，p(c,e)表示中文詞 c 與英文詞 e 一起出現在同一句對的機率，而 p(x)則表示單語詞彙 x 出現在某一句對的機率。這些機率可用最大可能性估計（maximum likelihood estimation）算出，亦即用其出現的句對數，除以全部句對數來求得。

MI 的值是對稱的（symmetric），亦即 MI(c,e)=MI(e,c)，而且 MI 的值都爲正數，但沒有上界。雖然 MI 很早就被統計式自然語言處理採用在這類應用上 [4]，但後續的研究，不斷的驗證出 MI 其實效果並不佳[10]。但由於其計算簡單，我們也納入此方法，以供比較驗證。

2. 相關分析（CC, correlation coefficient）

文獻中常用 Chi-square 來分析兩個事物的相關性 [4]：

$$\chi^2(c,e) = \frac{(f_{11} \times f_{22} - f_{12} \times f_{21})^2}{FcFc^* \, FeFe^*}$$

其中f_{11}、f_{21}、f_{12}、f_{22}分別代表中文詞c出現而英文詞e也出現的句對數、c出現但e沒出現的句對數、c沒出現但e出現的句對數、以及兩者都沒出現的句對數，如表二所示，其中 Fc、Fc*、Fe、Fe* 與N，分別代表各行與各列的合計。

表二、中文詞 c 與英文詞 e 在句對中出現次數的交叉分析

	c 出現	c 沒出現	合計
e 出現	f_{11}	f_{12}	Fe
e 沒出現	f_{21}	f_{22}	Fe*
合計	Fc	Fc*	N

Chi-square 可用來做相依性統計考驗（test for dependence），而不需假設事件是否常態分佈。此值介於 0 到 1 之間，其實是底下相關係數的平方：

$$CC(c,e) = \frac{(f_{11} \times f_{22} - f_{12} \times f_{21})}{\sqrt{FcFc^* \, FeFe^*}}$$

相關係數的值介於-1 到+1 之間，可顯示兩事物從負相關到正相關的程度。在本應用中，我們選用相關係數 CC 以分析對照詞彙的共現性（co-occurrence）。Chi-square 與 CC 都是對稱的指標。

3. 可能性比例（LR, likelihood ratios）

文獻上提到 Chi-square（或 CC）對事件發生次數不多的情形，效果較差，因此建議最好不要用於事件次數低於 20 次的場合 [4]。LR 則比較適合於資料稀少的情形。其計算公式，原爲負值 [4]，我們將其改爲正值，如下：

$$LR(c,e) = \log L(f_{11}, Fc, p(e \mid c)) + \log L(f_{12}, Fc^*, p(e \mid c^*)) \\ - \log L(f_{11}, Fc, p(e)) - \log L(f_{12}, Fc^*, p(e))$$

其中$L(k,n,x)=x^k(1-x)^{n-k}$, p(e)=Fe/N, $p(e|c)=f_{11}/Fc$, $p(e|c^*)=f_{12}/Fc^*$。檢視其定義，並經過驗算，LR的值是對稱的。

4. Dice 係數（DC, Dice coefficient）

在資訊檢索中 Dice 係數常用來衡量兩事物的相似度：

$$DC(c,e) = 2f_{11}/(Fc+Fe) = 2f_{11}/(2f_{11}+f_{12}+f_{21})$$

其值是對稱的，且計算雖簡單，文獻上卻顯示其成效較不受出現次數多寡的影響 [10]。

5. 分數累積（FC, fractional count）

上述的各項分析，所依賴的資訊，都屬於跨句對的（inter-sentence），而沒有用到句對內的（intra-sentence）資訊。圖一中顯示，盲目的配對，有六種可能的組合，因此每一種配對組合裡的每一種翻譯，應該只獲得 1/6 的分數。在同一句對中累積這些分數，可以知道每一個詞對翻譯機率，例如「芸香劑方」翻成「rutin」的機率是 2/6。累積詞對在所出現句對中的翻譯機率（而非次數），則爲我們所謂的 FC 值。其計算公式表達如下：

$$FC(c,e) = \sum_{for\_all\_i, s.t.(c,e)\in sp(i)} p_{sp(i)}(c,e)$$

其中sp(i)表示第i個句對（且同時包含中文詞c與英文詞e），而$p_{sp(i)}(c,e)$表示在第i個句對中詞對(c,e)的翻譯機率。這樣的FC，其值是對稱的。

6. EM 分析（Expectation-Maximization analysis）

上述五種分析方式，都沒有考慮到一個因素，即：較差詞對的懲罰（penalization of implausible alignment）。我們當然可以利用上述五種方法之任一種，依照其指標數值排序，將排序在前面的詞對視爲正確的翻譯，然後回頭過濾不可能的盲目配對，如此反覆循環，如圖一的 Run 2 與 Run 3 所示。這樣，也可以刪除掉（懲罰了）較差的詞對。

然而，在我們反覆進行詞對的過濾與分析以前，利用 EM 方法，就可以在進行詞對分析時，懲罰出現次數較少的詞對，同時凸顯出現較多的翻譯詞對。亦即若某個中文詞與某個英文詞，在所有的句對中，有較多的配對關係，則該中文詞就不太可能再跟同句對中的其他英文詞配對。底下的 E 步驟與 M 步驟的反覆計算方法，可以自動達到這種特性：

E-step：在所有中文詞 c 與英文詞 e 同時出現的句對中，累積條件機率 p(e|c)的值

$$s(c,e) = \sum_{for\_all\_i, s.t.(c,e)\in sp(i)} p(e|c) = p(e|c) * f_{11}$$

其中f_{11}定義如前，而p(e|c)的初始值可任意設定爲 1，或設定爲FC(c,e)。

M-step：根據上述累積的結果，重新評估條件機率

$$p(e|c) = \frac{s(c,e)}{\sum_v s(c,v)}$$

其中v爲任意使得s(c,v)不爲 0 的英文詞。表三顯示一個中文詞c=「驅動電路」的範例。在所有的句對中，它與其他四個英文詞共同出現的次數如f_{11}一欄所示。在初始值p(c|e)都爲 1 的情況下，經過四次EM步驟後，c跟e4 互爲翻譯的機率越來越高，而其他則越來越低，果真達到我們的期望。

表三、EM 分析範例

		Loop 1		Loop 2		Loop 3		Loop 4	
c=驅動電路	f_{11}	s(c,e)	p(e|c)	s(c,e)	p(e|c)	s(c,e)	p(e|c)	s(c,e)	p(e|c)
e1= display devices	2	2	2/8	0.5	0.1818	0.3636	0.1081	0.2162	0.0584
e2= electroluminescent lamp	1	1	1/8	0.125	0.0455	0.0455	0.0135	0.0135	0.0036
e3=lamp driving circuit	1	1	1/8	0.125	0.0455	0.0455	0.0135	0.0135	0.0036
e4= driving circuit	4	4	4/8	2	0.7273	2.9090	0.8649	3.4595	0.9343

四、成效分析

將前述的 506510 篇專利中英標題與摘要，共 463MB 純文字檔案，以起始步驟、斷詞處理、詞彙過濾三步驟在桌上型電腦上計算（Pentium 4, 3.20GHz, 2GB RAM），總共花費 5945 秒，統計結果如下：

表四、專利中英標題與摘要之處理結果統計

Number of bilingual texts processed (463MB)	506510
Number of empty Chinese Title	0
Number of empty Chinese Abstracts	281
Number of empty English Title	0
Number of empty English Abstracts	166532
Number of Doc. with both abstracts empty	2
Number of bilingual texts yielding empty term pair	322309
Number of term pairs generated	1180281
Number of unique term pairs after merging	455696

將上述 455696 個詞對，以前一節中的所有對列分析方式計算，總共花費 1340 秒（Intel CPU T2500 2.0 GHz, 2 GB RAM），其中讀取所有詞對花費 15 秒，以 EM 分析同時求 p(c|e) 與 p(e|c)用五個迴圈所花費的總時間爲 1296 秒，而其他五種分析方法總花費時間則爲 29 秒，平均每一種方法不到 6 秒鐘。將此對列分析結果，以 FC 由大到小排序，檢視其前 10 名的詞對，如表五所示：

表五、以 FC 排序的前 10 個詞對

c	e	FC	p(e|c)	p(c|e)	DC	MI	CC	LR	f_{11}	Fc	Fe
半導體裝置	semiconductor device	2078.30	1.00	1.00	0.79	5.58	0.79	-1	2455	2764	3429
顯示裝置	display device	867.83	1.00	1.00	0.70	6.22	0.70	-1	1225	1909	1589
液晶顯示裝置	**liquid crystal display device**	772.08	**0.98**	**1.00**	0.64	6.21	0.65	-1	1078	2072	1292
電連接器	electrical connector	750.25	1.00	1.00	0.80	7.15	0.79	-1	830	1045	1029
電子裝置	electronic device	326.45	1.00	1.00	0.57	6.90	0.59	-1	541	1180	706
背光模組	backlight module	323.25	1.00	1.00	0.74	7.70	0.74	-1	498	807	546
液晶顯示裝置	**liquid crystal display**	292.15	**0.01**	**1.00**	0.27	5.03	0.27	-1	451	2072	1227
半導體記憶裝置	semiconductor memory device	237.50	1.00	0.98	0.57	7.74	0.59	-1	302	394	661
*** 液晶顯示裝置**	**liquid crystal**	218.15	**0.01**	**1.00**	0.27	5.10	0.27	-1	431	2072	1119
薄膜電晶體	thin film transistor	197.08	1.00	1.00	0.71	8.43	0.72	-1	280	467	320

* 註：錯誤的翻譯詞對

在表五中，中文詞「液晶顯示裝置」對列到三個英文詞「liquid crystal display device」、「liquid crystal display」與「liquid crystal」，其中前兩是正確的，第三個是錯的，但只有第一個的翻譯機率 p(e|c)接近 1，其餘接近 0。從 EM 法的計算公式可知，由於競爭性懲罰的關係，對同一個中文（或英文）而言，只有一個對列會存活而獲得較高的翻譯機率，其他的對列，其機率都會退化成接近 0。亦即其無法擷取多對多的同義翻譯詞對。

至於表中的 LR，其計算過程有 log(0)的情形，但因該值沒有數學定義，我們遂將其值設爲-1。事實上，若以 LR 排序，其前 10 名的詞對如表六所示。

表六、以 LR 排序的前 10 個詞對

c	e	FC	p(e|c)	p(c|e)	DC	MI	CC	LR	f_{11}	Fc	Fe
環氧樹脂組成物	epoxy resin composition	44.62	0.97	1.00	0.87	10.47	0.87	757.59	99	114	113
照明系統	illumination system	67.64	1.00	1.00	0.82	10.32	0.83	731.76	98	133	106
記錄載體	record carrier	50.93	1.00	1.00	0.87	10.58	0.87	703.34	91	105	104
感光性樹脂組成物	photosensitive resin composition	65.58	1.00	1.00	0.72	9.86	0.72	697.28	102	137	148
資料處理系統	data processing system	52.29	1.00	1.00	0.87	10.65	0.87	675.94	87	100	100
熱交換器	heat exchanger	60.58	1.00	1.00	0.81	10.40	0.81	662.40	89	119	102
基地台	base station	38.96	1.00	1.00	0.80	10.42	0.80	638.43	86	111	104
資訊儲存媒體	information storage medium	48.62	1.00	1.00	0.83	10.59	0.83	633.72	83	103	96
半導體晶片	semiconductor chip	55.06	1.00	1.00	0.59	9.33	0.59	630.94	101	186	155
矽晶圓	silicon wafer	51.56	1.00	1.00	0.80	10.55	0.81	602.65	80	106	93

一般而言，我們可以依照各個對列方法的結果一一排序，然後檢視其前N名的詞對品質，以瞭解該分析方法的成效。然而實做結果顯示，有好幾個分析方法，其最大值的詞對很多，不是唯一，以致於前N名沒有區別性。如表七所示，若依照EM的結果以平均翻譯機率排序，其機率最大值 1 者，就有 1 萬 8 千多個。排除出現次數少於 5 次的詞對，還有 94 個並列第一。爲了再加以區別，可用其他的數值，例如用該詞對的出現次數（欄位f_{11}），做爲第二個排序條件。

表七、各項排序方式其最大值的詞對個數

最大數值範圍	詞對數
(p(c|e)+p(e|c))/2=1.0	18322
(p(c|e)+p(e|c))/2=1.0 and f_{11}>=5	94
DC=1.0	156353
DC=1.0 and f_{11}>=5	51
MI=17.49= maximum	152907
MI=17.49 and f_{11}>=5	33
CC=1.0	156353
CC=1.0 and f_{11}>=5	51
LR>186.13（註 1）	249
LR>186.13 and f_{11}>=5	249
FC>311.01（註 2）	6
FC>311.01 and f_{11}>=5	6

註 1：此數爲其平均數＋3＊標準差=26.07+3*53.35

註 2：此數爲其平均數＋3＊標準差=28.20+3*94.27

經由上述二階排序後，我們人工檢視了各個對列分析法前 n=50 個詞對，結果如表八所示。另外，我們也分析各個方法前 50 個正確詞對的重疊比率，結果如表九所示。這兩個表裡面的數據，印證了我們的觀察：{EM>LR>FC} > {DC=CC} >> MI，亦即 EM 與 LR、FC 的排序效果最好，惟三者中仍有些爲差距，其次是 DC 與 CC，兩者特性非常相似，最差的是 MI。

表八、各個方法前 n 個詞對人工判斷結果

排序方法	FC	EM	DC	MI	CC	LR
人工檢視的詞對數	50	50	50	50	50	50
錯誤的詞對數	3	0	6	39	6	1

表九、各個方法前 n=50 個正確詞對的重疊個數與重疊率

	FC	EM	DC	MI	CC	LR
FC						
EM	9 (18%)					
DC	0	0				
MI	0	0	0			
CC	0	0	45 (90%)	0		
LR	2 (4%)	10 (20%)	0	0	0	

表九透露出各個方法擷取的詞對重疊率不高（DC 與 CC 除外），顯示其能找出各有特色的正確詞對。爲了進行大規模的驗證，我們以國立編譯館學術名詞，共約 160 萬個中英詞對，來過濾前述的 455696 個詞對，得出其中的 7050 個，已出現在此雙語詞庫中，我們稱其爲「舊詞對」，其範例如表十所示；而另有約 25963 個詞對，雖不在既有詞庫中，但詞對裡中、英文詞彙的每個子字串都互有翻譯的關係，我們稱其爲「新詞對」，其範例如表十一所示。

表十、「舊詞對」範例

c	e	FC	p(e|c)	p(c|e)	DC	MI	CC	LR	f_{11}	Fc	Fe
半導體裝置	semiconductor device	2078.3	1	1	0.79	5.58	0.79	-1	2455	2764	3429
半導體裝置	semiconductor devices	28.583	0	1	0.03	4.93	0.09	130	48	2764	105
顯示裝置	display device	867.83	1	1	0.70	6.22	0.70	-1	1225	1909	1589
顯示裝置	display devices	8.0595	0	1	0.02	6.20	0.08	61.7	16	1909	21
電連接器	electrical connector	750.25	1	1	0.80	7.15	0.80	-1	830	1045	1029
電子連接器	electrical connector	4	1	0	0.01	7.00	0.06	-1	5	7	1029
電氣接頭	electrical connector	1	1	0	0.00	7.48	0.03	-1	1	1	1029
電子裝置	electronic device	326.5	1	1	0.57	6.90	0.59	-1	541	1180	706
電子裝置	electron device	2.667	0	1	0.01	6.80	0.05	21.1	5	1180	7
薄膜電晶體	thin film transistor	197	1	1	0.71	8.43	0.72	-1	280	467	320
薄膜電晶體	thin-film transistor	28	0	1	0.15	8.39	0.27	215	39	467	46
記錄媒體	recording medium	162.4	1	1	0.65	7.98	0.66	-1	315	556	414
記錄媒體	recording media	15.87	0	1	0.09	7.88	0.19	135	27	556	38
記錄媒體	record medium	7.667	0	1	0.04	8.13	0.13	58.4	11	556	13
記錄媒體	record media	0.143	0	0.25	0.00	8.37	0.04	5.80	1	556	1
記錄介質	recording medium	5.333	1	0	0.04	8.10	0.11	-1	8	13	414
電腦系統	computer system	159.9	1	1	0.85	8.74	0.85	-1	313	374	361

表十一、「新詞對」（註）範例

c	e	FC	p(e|c)	p(c|e)	DC	MI	CC	LR	f_{11}	Fc	Fe
半導體裝置	semiconductor assembly	1.5	0	0.56	0.0014	5.06	0.02	5.66	2	2764	4
半導體裝置	semiconductor equipment	0.5	0	0.0	0.0007	3.06	0.01	1.29	1	2764	8
半導體裝置	semiconductor arrangement	0.5	0	0.0	0.0007	3.06	0.01	1.29	1	2764	8
半導體器件	semiconductor device	3	0.99	0	0.0017	4.75	0.02	-1	3	6	3429
顯示器裝置	display device	88.2	1	0	0.1347	6.39	0.23	-1	118	163	1589
顯示器設備	display device	0.5	0.5	0	0.0013	6.86	0.03	-1	1	1	1589
顯示器器件	display device	0.5	1	0	0.0013	6.86	0.03	-1	1	1	1589
液晶顯示裝置	liquid crystal display device	772	0.98	1	0.6409	6.21	0.66	-1	1078	2072	1292
液晶顯示裝置	lcd device	21.9	0	1	0.0265	5.82	0.09	97.1	28	2072	44
液晶顯示裝置	lcd apparatus	9.3	0	1	0.0105	5.85	0.06	38.4	11	2072	17
液晶顯示裝置	lcd equipment	0.5	0	0.5	0.001	6.47	0.02	4.49	1	2072	1
電連接器	electric connector	31.1	0	1	0.0879	6.95	0.18	212	49	1045	70
電連接器	cable connector	2	0	0	0.0037	3.18	0.01	2.67	2	1045	39
電氣連接器	electrical connector	81.9	1	0	0.1803	7.11	0.28	-1	105	136	1029
電接頭	electrical connector	2	0.97	0	0.0077	7.16	0.06	-1	4	5	1029
電連接器插座	electrical connector	1	1	0	0.0019	7.48	0.03	-1	1	1	1029
電路連接器	electrical connector	0.3	0.5	0	0.0019	7.48	0.03	-1	1	1	1029

註：以第一列而言，此詞對不在既有的雙語詞庫中，但中文「半導體裝置」中的「半導體」在其英文詞對中有「semiconductor」對應（亦即此詞對出現在既有的雙語詞庫中），而其中的「裝置」也有「assembly」對應，而且其中沒有任何中文沒有對應的英文，反之亦然，我們因此稱此其爲新詞對。

五、解決方案二

經過前一階段之處理，我們已經獲得一組可信的專業領域中英詞對，至少有 25963 條。在此階段主要的工作，是基於此一可信度較高的結果，自動擴充更多的中英詞對。基本的擴充概念很簡單：將已有的中英詞對組合加長後，再回到原專利文件中檢查是否出現在對應的文件中。此概念的範例，如圖三所示，其中既有的詞對以不同的底線與頂線顯示，當鄰近的詞彙找到既有詞對時，就可以將其組合出新的詞對。

標題：背光模組、串接模組及其導電塊	Title: Backlight modules with connector modules and conductive blocks thereof
摘要：一種適用於大尺寸平面顯示器之背光模組，包括一背板、複數個第一燈管、複數個第二燈管、一第一串接模組以及一光學構件組。…	Abstract: Backlight modules for large size flat panel displays are provided. A backlight module comprises a plate, a plurality of first and second lamps, a first connector and an optical assembly. …

圖三、基於既有詞對的翻譯詞對擷取示意圖

上述新詞對的擷取構想很簡單，也符合一開始提到的專利機器翻譯資源建構目標，亦即「擷取出來的雙語詞對越長、越多，對長句的專利翻譯，會越有幫助」。本階段的問題，在於如何提高執行效能。圖四是此構想的直覺演算方法。

我們目前有 506510 篇雙語專利摘要，並且有前一階段擷取的專利新詞對 25963 個，與國立編譯館的學術名詞約 160 萬條，再加上從一般性電子字典取得的中英詞彙，總共約 170 萬條對照詞。若直接使用上述演算法，其時間複雜度爲$O(NM^2)$，意謂著必需執行 1445000 兆次「詞對是否出現在某專利中」的檢查（50 萬 × 170 萬 × 170 萬）。顯然這並非是一個合理時間內，可執行完成的演算法。有鑑於此，我們利用檢索系統的索引結構與查詢功能，來提高效率，其演算法分三步驟，如下：

1. Set i = 1 to M；M 爲所有的中英詞對數
2. Set j = 1 to M；
3. Set $c_{ij} = c_i c_j$；c_i表示第i筆中英詞對中的中文詞（組合成較長的中文詞）
4. Set $e_{ij} = e_i e_j$；e_i表示第i筆中英詞對中的英文詞（組合成較長的英文詞）
5. Set k = 1 to N；N 爲所有的專利篇數
6. 若c_{ij} 出現在CP_k且e_{ij} 出現在EP_k，則 (c_{ij}, e_{ij}) 爲新增詞對；其中CP_k及EP_k分別代表第k篇專利的中文及英文部分。

圖四、新詞對擷取的基本演算法

1. 建立所有專利文件的反向索引檔 I，裡面記錄每個詞出現在哪些文件的資訊。

2. 建立一個記錄詞對編號的陣列 A，其維度爲專利資料的筆數 N。
3. Set i = 1 to M；M 爲所有的中英詞對數
4. 以 (c_i, e_i) 查詢索引I以檢索出所有包含 (c_i, e_i) 的專利編號集合$S = \{s_1,...,s_P\}$
5. Set $j = s_1$ to s_P
6. push i to A[j]；A[j]將記錄所有出現在專利 j 的詞對 i

7. Set p = 1 to N；N 爲專利資料的筆數
8. Set x = 所有包含在 A[p] 中的詞對編號
9. Set y = 所有包含在 A[p] 中的詞對編號
10. Set $c_{xy} = c_x c_y$；c_x表示第x筆中英詞對中的中文詞
11. Set $e_{xy} = e_x e_y$；e_x表示第x筆中英詞對中的英文詞

若c_{xy} 出現在CP_p且e_{xy} 出現在EP_p，則 (c_{xy}, e_{xy}) 爲新增詞對；其中CP_p及EP_p分別代表第p篇專利的中文及英文部分。

圖五、新詞對擷取的改良演算法

上述算法只針對「確定出現在同一篇專利文件的中英詞對」，才進行兩兩組合。根據實際處理結果，同一篇文件會出現的詞對數，平均約爲 4 組。因此只需檢查大約 200 萬組（50 萬 × 4）詞對是否在文件中即可，在速度上的增加非常明顯。

表十二爲演算法各步驟的執行時間。結果顯示，此演算法確實可在合理時間內自動擷取新的詞對。經此算法自動擴充的中英詞對數量爲 406184 筆，對既有的上百萬條雙語詞對而言，約可增加 20%（=40/(170+40)）的數量。不用驗證，**它們都是正確的**（錯誤的詞彙可用組合規則過濾掉），而且是**不在既有詞庫裡的專業領域新詞對**。

上述演算法只組合兩個既有詞彙，事實上也可以考慮組合三個、四個、甚至 n 個詞彙。然而，一方面合法的長詞會越來越少，二方面若既有詞對已夠多，可以先只組合兩個詞彙。若怕遺漏，將找出來的新詞對，以此方法再演算一遍，即可找出原來需要三個甚至四個詞彙組合的新詞。這種方式，比直接組合多個詞彙還要快。

表十二、新詞擷取演算法各步驟的執行時間

步驟	時間(秒)
索引檔建立	2971
查詢所有中英詞對	25920
檢查新增中英詞對	1560

六、結論

從相關計畫的分析得知，翻譯詞庫的持續更新，是不斷提升機器翻譯品質的必要任務，也是輔助人工翻譯或進行跨語檢索的基礎工作。本文以精確導向及召回導向角度的探討此議題，提出了因應不同目標的解決方法。

在精確導向的任務中，我們比較了六種詞彙對列方式，說明它們的優缺點，並以實際資料作驗證，從而得出合理的解讀結果。亦即就詞對的排序效果而言，{EM>LR>FC} > {DC=CC} >> MI。其中 FC 方法僅作局部（單篇）的分數計算再全部累加，計算最簡單；而 LR 的計算也不算太困難，效果卻更好；EM 則最花時間，但效果最好，只是難以找出多對多的同義翻譯，而 LR 與 FC 都可以（因爲其具有對稱性）；即便是最差的 MI 法，其排序在前頭的正確詞對跟這三種最好的方式不同，因此可以作爲輔助的詞對擷取方法，爲後續合併或混用多種對列方式的研究，開啓了可能性。

而一旦精確導向的詞對擷取結果出爐，立刻可用於召回導向的任務。我們的成果顯示，簡單的想法加上有效的實做，即可從雙語對列語料庫中召回大量的詞對。對後續不斷累積擴大的專利雙語語料而言，本文詳述的方法，可供後續一再地應用，以自動的方式，擷取出更多的專利新生中英詞對。

參考文獻

[1] Patents By Country, State, and Year - Utility Patents (December 2008). 2009; Available: http://www.uspto.gov/go/taf/cst_utl.htm. [Accessed: July 10, 2009].

[2] 96 年 專 利 統 計 . 2008; Available: http://www.tipo.gov.tw/ch/MultiMedia_FileDownload.ashx?guid=e24d3489-c729-4159-ab2a-eab9e6788d49.pdf. [Accessed: July 10, 2009].

[3] 台北市日本工商會 and 日本知的財產協會. 致經濟部建議書. 2007; Available: http://kousyoukai.japan.org.tw/tokusennchu.pdf.

[4] Christopher D. Manning and Hinrich Schutze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 2001.

[5] Tseng, Y.-H. Multilingual Keyword Extraction for Term Suggestion. in 21st International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '98. 1998. Australia.

[6] Tseng, Y.-H., Automatic Thesaurus Generation for Chinese Documents. Journal of the American Society for Information Science and Technology, 2002. 53(13): p. 9.

[7] Tseng, Y.-H., C.-J. Lin, and Y.-I. Lin, Text Mining Techniques for Patent Analysis. Information Processing and Management, 2007. 43(5): p. 1216-1247.

[8] Tseng, Y.-H., Automatic Cataloguing and Searching for Retrospective Data by Use of OCR Text. Journal of the American Society for Information Science and Technology, 2001. 52(5): p. 12.

[9] Tseng, Y.-H., "Content-Based Retrieval for Music Collections," Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '99, Aug. 15-19, Berkeley, U.S.A., 1999, pp.176-182.

[10] W. Bruce Croft, Donald Emtzler, Trevor Strohman, *Search Engine: Information Retrieval in Practice*, Addison-Wesley, 2009.