中英文專利文書之文句對列[1]

田侃文
國立政治大學
資訊科學系
96753027@nccu.edu.tw

曾元顯
國立臺灣師範大學
資訊中心
samtseng@ntnu.edu.tw

劉昭麟
國立政治大學
資訊科學系
chaolin@nccu.edu.tw

摘要

綜觀今日全球化的趨勢，世界各國皆進行跨語言的專利說明書翻譯工作。在中文專利說明書中譯英方面，為了追求更精確的翻譯品質，蒐集大量且正確的專利文書平行語料，能夠協助輔助式機器翻譯及資訊檢索的研究工作進行。因此本研究便希望利用中英技術名詞對應表，透過統計詞頻調整詞對應權重及計算中英文句子的向量相似度等機制，搭配動態規劃演算法，計算中英專利文書的句對相似度，最後產生對列結果。以精確率、召回率及輔助式機器翻譯系統，評比本對列系統的對列成效。實驗結果顯示本系統不僅在 1：1 對列模式[2]的精確率達到 0.995，且產出的大量中英文句對確實能夠提升輔助式機器翻譯系統的翻譯品質。

關鍵詞：專利說明書、電腦輔助機器翻譯、文句對列、動態規劃演算法

1. 緒論

在這資訊爆炸的時代，科技產業日新月異，因此在開發一項新的產品前，對於專利說明書的熟讀，更顯得格外重要，以避免在研發過程中，侵犯了他人的智慧財產權。近年來，中國大陸的經濟蓬勃發展，外商投資市場急速拓展，因此外語使用者對於中文專利說明書的查詢需求量便大幅增加。於 2007 年 10 月在拉脫維亞舉行的歐洲專利局 (European Patent Office，簡稱 EPO) 專利資訊會議，舉辦的二場主題皆與中文專利文書翻譯密切相關，顯示中文專利文書英譯的議題越來越被重視。

想要有高品質的專利文書翻譯效果，必須倚靠人力進行翻譯，不僅在時間及成本上的花費極高，在數量上也有所限制。為了在產品的生產前置作業上，減少跨語言專利查詢的成本，發展一套專利說明書英譯或是檢索系統，便成了刻不容緩的事情，不僅能方便使用者的查詢，更能間接促進科技產業的發展。

無論是進行專利文書翻譯或檢索研究，大量且正確的平行語料是不可或缺的。在跨語言資訊檢索研究中，針對查詢的問句進行翻譯有許多種策略，如以辭典為本(dictionary-based) 、索引典為本 (thesaurus-based) 等歸類於以知識為本 (knowledge-based) 的策略及語料庫為本 (corpus-based) 等策略[17]。其中以語料庫為本的策略，將大量對列好的雙語文句，透過計算詞彙對譯的強度建構詞彙對應表，利用此表，便能在使用者進行查詢時進行詞彙的翻譯動作[6]。而在輔助式機器翻譯的領域中，能夠利用這些對列好的雙語文句進行翻譯模型的訓練及其它相關的處理，供後續翻譯系統使用。

想要得到平行語料進行研究並非難事，為了保證語料本身的品質，除了利用人工對列的方式進行外，利用簡單且兼顧正確性的方法，如：僅利用雙語辭典進行詞彙比對，訂定保險的篩選門檻，也能夠產生平行的語料。但利用人工的方式進行對列需要大量的人力成本，而僅利用詞彙比對的方法為了兼顧正確性，往往得到數量較少的對列結果產出，仍然不敷成本效益。因此我們希望能夠改進現有的文句對列技術，透過加入專業領域的辭典及利用自然語言處理的方法，發展出一套適用於專利文書文本的文句對列系統。

為了統一詞彙的用法，以下所稱「英漢辭典」指的是本系統採用的英漢辭典；「原中文詞義」指的是英漢辭典未合併中文近義詞表的中文詞義，在中文近義詞表合併至英漢辭典後，我們統稱英漢辭典內的中文詞義及其中文近義詞為「中文詞義」；擷取出英文片語及技術名詞的步驟，簡稱為「詞彙擷取」；「英文詞彙」指的是英漢辭典及「中英技術名詞對應表」的英文詞，

[1] 本論文另有 25 頁的版本，請參照連結：http://www.cs.nccu.edu.tw/~chaolin/papers/rocling_tien.pdf。
[2] 例如：對列模式「1：1」代表一句來源語言句子對應至一句目標語言句子。

或經過詞彙擷取後，英文句子裡的片語、英文技術名詞及句子裡其它利用空白隔開的英文單詞，這些英文詞彙皆由一個以上連續的英文單詞組成，如：「of course」，我們稱為一個英文詞彙，由兩個連續的英文單詞組成，「Of course, I will help you.」，我們稱這英文句子內有五個英文詞彙：「Of course」、「I」、「will」、「help」及「you」；「詞形還原」(lemmatization) 指的是將英文詞彙，還原成其原形的處理步驟，例如：將「ate」還原成「eat」；「對應」一詞，指的是中英文互為翻譯的文章、句子或詞彙，例：「中英文對應句子」指的是中英文互為翻譯的句子；「句對」指的是完成對列後產生的中英文對應句子；「詞對」指的是在對列過程中，中英文句子在詞彙比對時，比對到的中英對應詞彙。而中英文句子在完成對列後，稱為「句對」，該句對包含的一組以上的詞對便形成「詞對集合」。

我們發展的系統（以下簡稱本系統）在前處理步驟中，利用簡單的規則對未經過段落對列處理 (paragraph alignment) 的中英文章進行斷句，透過 StanfordLexParser-1.6 進行英文詞彙的詞形還原，再利用詞彙量豐富的英漢辭典及中英技術名詞對應表，在中文的部份以長詞優先的技術進行斷詞，在英文的部份則進行詞彙擷取及片語保留的動作。為了增加中英文詞彙比對的成功率，透過 HowNet 及中研院現代漢語一詞泛讀系統尋找英漢辭典中的原中文詞義的近義詞彙，拓展英漢辭典的規模，並利用自「國立編譯館學術名詞資訊網」擷取、整理並建構的中英技術名詞對應表，讓本系統能適用於專利文書的文本。在進行各模式中英文句子的詞彙比對部份，我們沿用 Ma[15]於 2006 年提出的詞彙權重計算方法及其採用的動態規劃演算法。在對列的過程中，給予比對的來源句 (source sentence) 及目標句 (target sentence) 詞彙比對的分數，再利用衍生自餘弦相似度 (cosine similarity) [16]的計算原理，計算得到中英文句子之間的向量相似度分數，作為句子相似度的輔助評分權重，之後利用動態規劃演算法，計算得到整體相似度分數最高的對列組合，產生對列的結果。我們同時訂定篩選門檻，以句對平均比對詞數及向量相似度分數作為篩選條件，篩選出「1：1 信心句對」作為平行語料。

2. 文獻探討

跨語言文句對列的技術有很多種，發展至今，已有許多學者提出不同的方法，有針對來源語言 (source language) 及目標語言 (target language) 屬於同一語系(如:英文及法文同屬於印歐語系）的對列技術，也有跨語系的相關對列技術研究，如針對英日翻譯、英漢翻譯等翻譯文章。

近年來利用英漢辭典比對進行文句對列的方法，有[15]提出的「Champollion」這套對列工具。他認為任意兩個句子中的多個對應詞彙，並不應該一律給予相同的權重，於是透過修改 *tf-idf* [18]的權重計算公式，進行詞彙權重的調整，並搭配懲罰機制，依照不同的對列模式及句子長度差異進行扣分的計算。他同樣採用動態規劃演算法進行最佳對列模式的選擇，挑選出整體相似度分數最高的句對組合。

2007 年，Utiyama 與 Isahara[19]提出一套英日的專利平行語料庫，並參加第六屆 NTCIR 專利檢索 (patent retrieval) 的比賽。其語料庫包含了約 199 萬組經自動對列技術得到的英日句對。如此龐大的句對數量，必須倚賴自動對列技術才能蒐集完成。他們在專利說明書中發現「發明說明」段落所包含的兩個小段落：「先前技術」及「實施方法」翻譯較為整齊，因此選擇這兩個小段落進行對列處理。

在對列的方法上，他們利用英日及日英的辭典，搭配動態規劃演算法，對不同對列模式的句子進行相似度分數計算，接著再計算整篇文章的相似度總分及總句數的比例，最後選取在句子層級、文章層級及句數比例最為接近的對列結果，完成對列後約產生 700 萬組句對。接著他們對這些句對進行篩選，首先取出對列模式 1：1 的句對，再依照相似度分數進行排序，只取出以句號結尾的句對，同時過濾重複出現的句對，剩下約 390 萬組。為了保證這些句對的品質，他們進行區段抽樣的實驗，分析後決定取出相似度分數最高的前 200 萬組句對進行最後的篩選，去除句長太長以及長度差異太大的句對後，最後剩下約 199 萬組句對，他們便以這些句對建構專利文書的平行語料庫。除了提出文句對列及句對篩選的方法，他們在詞彙的層級將句對檢驗的評比分為三種等級：句對中的詞彙完全比對成功、50%以上比對成功及 50%以下的詞彙成功比對等三個等級；在語意的層級分成四個等級：句對的語意完全符合、80%以上的語意符合、80%以下的語意符合及語意完全不符合等四個等級。他們將這些專利文書語料依照國際專利分類號 (International Patent Classification，簡稱 IPC) 作分類，對詞彙及句長進行統計，並利用輔助式機器翻譯系統及翻譯指標進行翻譯效果的評比，證明該平行語料庫確實能夠勝任作為輔助式機器翻譯的訓練語料。

表一、語料來源、範圍與文章總數量統計

語料	文章數	範圍	來源
專利文書公開全文	7284	申請號 091132651 至 095145895	經濟部智慧財產局網站
專利文書公告全文	13675	申請號 084111615 至 096222166	經濟部智慧財產局網站
專利文書公告全文 摘要段落	417846	申請號 075102826 至 096213764	經濟部智慧財產局網站
科學人雜誌 中英對照電子書	2065	2002 年 1 月至 2009 年 1 月	國立政治大學圖書館
雙語網站 知識管理平台新聞	737	2005 年 8 月 30 日 至 2007 年 12 月 15 日	官方網站
自由時報 中英對照讀新聞	1553	2005 年 2 月 14 日 至 2009 年 5 月 27 日	官方網站
大考試題	131	2004 年至 2009 年	官方網站

表二、本研究採用訓練語料文章數及句數統計

語料	文章數	語言	總句數
專利文書公開全文	2271	中文	695326
		英文	518482
科學人雜誌中英對照電子書	319	中文	18966
		英文	19282

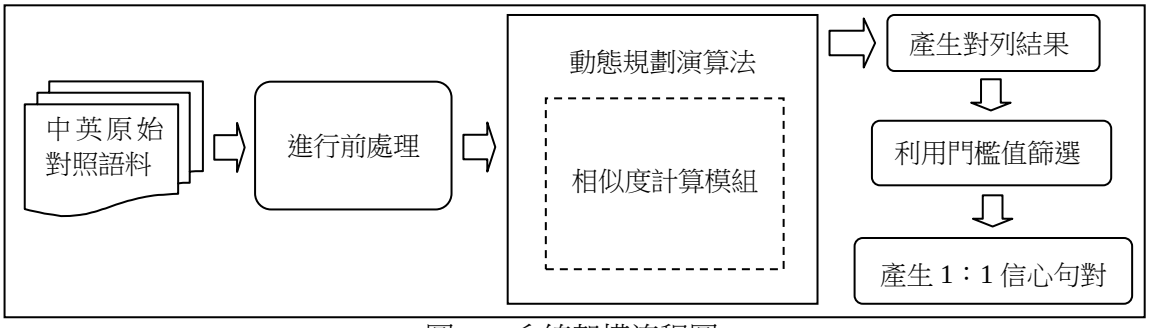
3. 語料來源

在專利說明書方面，本研究所需的語料來源，我們透過對從經濟部智慧財產局[9]擷取的資料整理及過濾，有中英文互為翻譯的「專利文書公開全文」及「專利文書公告全文」共約二萬篇及「專利文書公告全文摘要段落」約 42 萬篇（以下統稱為「專利文書的文本」）；在科普文學方面，我們有「科學人雜誌中英對照電子書」[5]從 2002 年 3 月創刊號至 2009 年 1 月約 2000 篇；在新聞文章方面，我們有「自由時報中英對照讀新聞」約 1500 篇[3]、「雙語網站知識管理平台新聞」[11]約 700 篇及包括了「四技二專統一入學測驗」、「學科能力測驗」及「大學指定科目考試」中「對話測驗」、「綜合測驗」及「閱讀測驗」等多個段落的「大考試題」[10]，共約 130 篇（以下統稱這四種語料為「其它主題的文本」），本研究詳細的語料來源統計數據如表一所示。由於語料數目相當龐大，在本系統開發時僅採用部份的語料作為訓練用途，如表二所示。

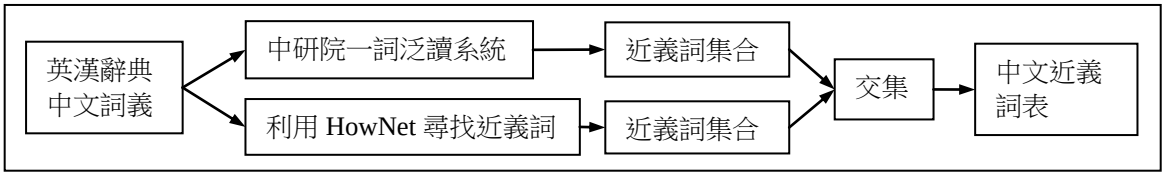
4. 研究方法

4.1 系統架構及對列流程

本系統的架構及處理流程如圖一所示，首先將任意中英互為翻譯、未經過段落對列處理的文章進行斷句、英文詞形還原、斷詞及詞彙擷取等前處理步驟，再透過相似度計算模組計算中英文句子的相似度，利用動態規劃演算法選取整體分數最佳之對列組合，產生對列結果，經過門檻值的篩選，再將高於門檻值條件的 1：1 對列模式句對取出，稱為「1：1 信心句對」，作為平行語料的用途。



圖一、系統架構流程圖



圖二、中文詞義近義詞表製作流程

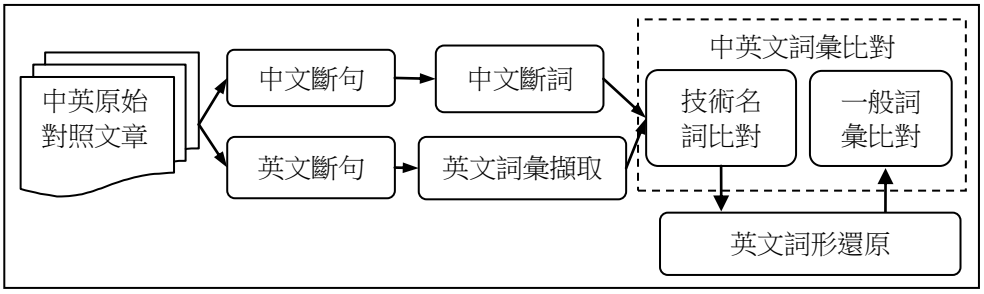
4.2 辭典之建置

在對列處理的過程中，利用英漢辭典進行中英句對的詞彙比對，為了提高詞彙比對的成功率，我們建置「中文近義詞表」，並將其合併至英漢辭典中，擴大英漢辭典的規模。除了建置中文近義詞表外，我們也建置「中英技術名詞對應表」，處理常見於專利說明書中的技術名詞。

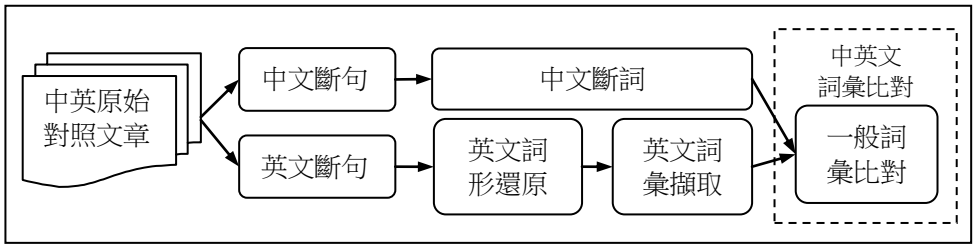
```
==============================
英文單字: chloride shift
氯轉移
鹽分移動
氯轉置
氯離子轉移
==============================
英文單字: chloride silver
氯化銀
==============================
英文單字: chloride stre
氯化物應力
==============================
英文單字: chloride stre corrosion crack
氯化物應力腐蝕裂縫
```

圖三、中英技術名詞對應表

- **英漢辭典**：本系統採用的電子英漢辭典，為國內知名企業英業達股份有限公司開發的 Dr.eye 譯典通線上辭典[12]，其含有 106269 筆英文詞彙及片語，在和牛津辭典比較詞目數量後，我們採用 Dr.eye 譯典通線上辭典作為本系統的英漢辭典。
- **中文近義詞表**：單純倚賴英漢辭典內之中文詞義進行中英詞彙比對雖然可行，但進行中英詞彙比對時的成功率並不高。為了增加中文詞彙比對的成功率，我們沿用呂明欣等學者[4]於 2007 年使用的中文近義詞尋找方法，來建構中文近義詞表，流程如圖二所示。在圖二中，我們將英漢辭典中的原中文詞義，透過 HowNet[13]及中研院一詞泛讀系統[2]尋找其意義相近的中文詞彙，在得到兩種方法產生的中文近義詞集合後，為了去除和原中文詞義偏差較大的中文近義詞，我們將兩個近義詞集合取交集，建構出中文近義詞表。我們將此中文近義詞表合併至英漢辭典中，作為中英文詞彙比對時的依據。
- **中英技術名詞對應表**：我們從「國立編譯館學術名詞資訊網」[6]擷取、整理並建構中英技術名詞對應表，中英文對應的詞目數量為 1660829 筆。在處理專利文書的文本時，除了一般性的英漢辭典外，若能再結合不同專業領域的中英技術名詞資訊，進行中英詞彙比對時，便不會將特殊、罕見的技術名詞視為未知詞，喪失了中英詞彙比對時應得的相似度分數。圖三為中英技術名詞對應表的部份內容，從圖中可以發現，單一英文技術名詞可能會對應至多個中文技術名詞，在中英詞彙比對的過程中，我們會將這些中文技術名詞皆納入比對的考量。
- **中文斷詞辭典**：中文斷詞辭典的詞彙涵蓋範圍，會影響以長詞優先方式進行斷詞的斷詞效果。許多進行斷詞相關研究的學者[14][20]認為較長的詞彙，能夠保留更完整的詞面資訊，因此，一套完整的中文斷詞辭典，在前處理的部份便佔了重要的角色。我們將合併中文近義詞表後的英漢辭典其全部的中文詞義，加上中英技術名詞對應表中全部的中文技術名詞，作為中文斷詞辭典，提供本系統在進行前處理時，進行中文斷詞的依據，中文斷詞辭典總詞目數量為 1835085 筆。



圖四、處理專利文書文本的前處理流程圖



圖五、處理其它主題文本的前處理流程圖

4.3 中英文原始文章前處理步驟

圖四為專利文書前處理流程圖。在圖四中，經過斷句處理後的中英文句子，接著進行中文斷詞及英文詞彙擷取，完成後即進入相似度計算模組中的中英文詞彙比對階段。在詞彙比對的過程中，首先利用中英技術名詞對應表進行技術名詞的比對，接著才進行英文詞形還原步驟。在處理專利文書的文本時，我們以「查表」的方式進行英文詞形還原，在完成英文詞形還原後，才利用英漢辭典進行一般性中英文詞彙的比對。

圖五為其它主題文本的前處理流程圖。可以發現圖五和圖四的最大不同點在於：完成斷句處理後的英文句子，在進行英文詞彙擷取前便進行了英文詞形的還原步驟，接著才進行詞彙擷取的處理及中英文詞彙的比對，這是因為在處理其它主題的文本時，並未進行技術名詞的比對。在處理其它主題的文本時，我們是以 StanfordLexParser-1.6 進行英文詞形的還原。

4.3.1 斷句

在中英文句子的斷句方面，我們僅利用「問號」、「驚嘆號」及「句號」三種符號進行斷句處理，而不將句子切分成太細的單位。我們利用簡單的規則避免英文句號與小數點混淆，並簡單地作人名及地名等名詞縮寫判斷，以避免斷句錯誤的情況發生。

4.3.2 英文詞形還原

StanfordLexParser-1.6 能夠在建立該英文句子的剖析樹 (parse tree) 後，根據該英文詞彙的詞性進行詞形還原處理。以圖六為例子，首先利用 StanfordLexParser-1.6 產生附帶著詞性的英文原句剖析樹，再利用 StanfordLexParser-1.6 的詞形還原套件對該剖析樹進行處理，使句中三個英文詞彙「plays」、「his」及「friends」進行詞形還原，產生這三個英文詞彙的原形「play」、「he」及「friend」。

StanfordLexParser-1.6 為 Java 語言所撰寫的套件，隨著電腦記憶體大小的不同，所能夠處理的英文句長也有限制。在專利說明書中，由於翻譯者的翻譯風格的不同，在文章中偶爾會出現極長的英文句子，這會造成剖析樹過深導致對列處理中斷，這樣的現象我們無法預期何時會發生。因此我們在對列其它主題的文本時，使用 StanfordLexParser-1.6 進行英文詞形還原的處理，而在對列專利文書的文本時，我們在完成斷句及斷詞後，在中英文詞彙比對的階段，完成技術名詞的比對之後才利用查表的方式進行英文詞形還原。我們沿用[15]開發的對列工具 Champollion 的英文詞形還原對應表，共有 136390 筆英文詞目及原形。有關中英詞彙比對的方式及步驟，會在 4.4.1 節作介紹及說明。

4.3.3 中文長詞優先斷詞、英文詞彙擷取

在完成中英文斷句處理後，在中文的部分首先對已經完成斷句的中文句子進行斷詞處理，在英文的部份，擷取出英文片語及技術名詞後再利用空白將其它的英文單詞分開，以利後續中英文翻譯詞彙的比對，我們以英漢辭典的英文詞彙及中英技術名詞對應表的英文技術名詞作為英文詞彙擷取的依據，詞目數量為 1151130 筆。

英文原句：Jim always plays baseball with his friends.
其剖析樹：(S(NP(NNP Jim))(ADVP(RB always))(VP(VBZ plays)(NP(NN baseball))(PP(IN with)(NP(PRP$ his)(NNS friends))))(. .))
詞形還原：Jim always play baseball with he friend.

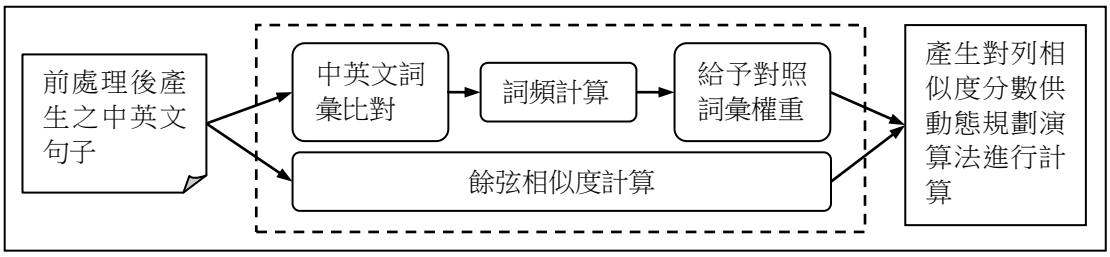
圖六、利用 StanfordLexParser-1.6 進行英文詞形還原

原中文句子：本發明提供一種形成半導體電子裝置的閘極堆疊的方法，其藉由晶圓接合含有高介電常數介電材料之至少一結構。

中研院斷詞：本 發明 提供 一 種 形成 半導體 **電子** **裝置** 的 **閘** **極** 堆疊 的 方法 ，其 藉由 **晶圓** **接合** 含有 高 **介電** **常數** **介電** **材料** 之 至少 一 結構 。

本系統斷詞：本 發明 提供 一種 形成 半導體 **電子裝置** 的 **閘極** 堆疊 的 方法 ，其 藉 由 **晶圓接合** 含有 高 **介電常數** **介電材料** 之至 少 一 結構 。

圖七、中研院中文斷詞及本系統斷詞結果比較



圖八、相似度計算模組（以虛線表示）

目前常見的中文斷詞工具為中研院的中文斷詞系統[1]，我們能透過網頁介面或撰寫程式進行查詢，提供中研院一段中文句子或文章，中文斷詞系統會進行詞性的標記及斷詞處理，並將查詢結果回傳。雖然中研院的中文斷詞系統為目前廣泛使用的斷詞工具，但為了處理專利說明書中常見的技術名詞，即時且大量地進行對列工作，我們不採用中研院的中文斷詞系統進行斷詞，而是利用中文斷詞辭典，以長詞優先的斷詞方式進行中文句子的斷詞處理。

圖七為利用中研院中文斷詞系統及中文長詞優先斷詞結果的比較，在技術名詞的斷詞差異處我們以粗體加上底線的字體表示。在圖七中，中研院的中文斷詞系統將「電子裝置」、「閘極」、「晶圓結合」、「介電常數」及「介電材料」等五個中文技術名詞錯誤斷詞，而利用長詞優先斷詞能夠正確地保留住這些中文詞彙。以圖中的中文詞「晶圓接合」為例，其對應的英文詞彙為「chip connection」，由於我們以「完全比對」的方式進行中英技術名詞的比對（也就是當中文詞必須完全相同才算比對成功），若「晶圓接合」被錯誤斷成「晶圓」及「接合」二詞，則在技術名詞比對的步驟便無法成功比對，而英漢辭典中沒有「chip connection」此英文片語，在中英文詞彙比對的步驟上，「chip connection」及「晶圓接合」這組中英文詞彙便失去了應得的詞彙比對分數。在圖七中的例子，共計有五組這樣的中英文對應詞彙被斷詞錯誤，累計起來失去的相似度分數便相當多，這也是我們採用長詞優先的方式進行中文句子的斷詞，並且將英文句子的片語及技術名詞擷取出來的主要原因。

4.4 相似度計算模組

相似度計算模組如圖八所示。相似度計算模組中，有兩個部份同時進行。模組中的第一個部份為中英文句子詞彙的比對及給分。第二個部份以計算向量之間相似度的原理，將中英文句子視為兩組向量，計算其之間的向量相似度。此計算方法不需要句長統計的數據，即可將中英文句子間的句長差異納入扣分考量。除此之外，對於句長相似但擁有不同權重詞彙之中英文句子（即不應該為翻譯對應之中英文句子），也能進行懲罰，使其得到較低的分數。將第一部份中英詞彙比對得到的分數，乘上第二部份之向量相似度作為扣分權重，最後得到該對列模式下的中英文句子相似度總分，這項分數接著提供給動態規劃演算法作為挑選最佳對列模式組合的依據。

4.4.1 翻譯詞彙的搜尋及比對方式

在中英文句詞彙比對的過程中，中英技術名詞對應表的參照順序在英漢辭典之前。進行中英文句子的詞彙比對時，以由左至右的順序將完成詞彙擷取處理的英文句子詞彙取出，並至中英技術名詞對應表進行搜尋，以完全比對的方式比對中文詞彙。

我們接著以同樣的方式，將英文句子中尚未比對到的詞彙取出，至英漢辭典搜尋。若能在英漢辭典中找到該英文詞彙，則將其中文詞義取出，至完成斷詞處理後的中文句子比對中文詞彙。為了兼顧比對的成功率及正確性，本系統以一字詞完全比對、二字詞以上部份比對的方式進行比對的動作，也就是英文詞彙及中文詞彙的最長共同子序列 (longest common subsequence)

為二字詞以上，也視為比對成功。當完成一組中英文句子的詞彙比對後，這些成功比對到的中英文詞對便形成詞對集合，我們予以記錄，供後續計算使用。

4.4.2 詞彙權重的計算

我們沿用[15]提出的方法，依照詞對集合內，各詞對的英文詞彙重要程度不同，給予比對到的詞對不同分數。

在資訊檢索的領域中，*tf-idf* (term frequency - inverse document frequency) [18]為一常見的計算公式，能夠計算文件中的詞彙權重值，*tf* 值為文件中某詞彙的出現次數，如公式(1)所示。而 *idf* 值代表該詞彙在所有語料的文件中的重要程度，*idf* 值的計算方式如公式(2)所示。在公式(2)中，*N* 為文件總數，*w* 為詞彙，*nd*(*w*)代表含有詞彙 *w* 的文件數量。詞彙的 *tf-idf* 權重值計算方式如公式(3)所示。

$tf(w) =$ 詞彙 w在文件中出現的次數 (1)

$$idf(w) = \log\left(\frac{N}{nd(w)}\right) \quad (2)$$

$$tfidf(w) = tf(w) \times idf(w) \quad (3)$$

[15]將計算文件中詞彙權重值的概念，應用到計算句子中的詞彙權重值。他認為中英文詞彙的比對，不應該將所有成功比對到的詞彙分數視為相同，太常出現的英文字，其強度代表性不如鮮少出現的英文詞彙。他將 *tf* 值的計算方式進行修改，並將名稱改寫為 *stf* (segment-wide term frequency) ，代表某中英文對應詞彙在該中英文句子中出現的次數如公式(4)所示，在公式(4)中，*e* 及 *c* 分別為某詞對集合中對應詞對的英文詞彙及中文詞彙，若進行中英文句子詞彙比對時，在英文句子中 *e* 出現三次，在中文句子中 *c* 出現三次，則在進行中英文詞彙比對的過程中，英文詞彙 *e* 及中文詞彙 *c* 會比對到三次，則其 *stf* 值為 3；[15]將 *idf* 值的計算方式進行修改，並將名稱改寫為 *idtf* (inverse document term frequency) ，代表句子中的詞彙在整篇文章中的重要程度，*idtf* 值的計算方式如公式(5)所示。在公式(5)中，*T* 代表文章的總詞頻（在此將總詞頻定義為完成「詞彙擷取」前處理步驟後的英文文章，統計其包含的英文詞彙出現的總次數，而非不同英文詞彙的數量，例如在某英文文章中只有三個英文詞彙：「really」、「really」及「good」，則該篇文章的英文總詞頻數為 3），*e* 代表詞對集合中詞對的英文詞彙，$O(e)$代表該英文詞彙 *e* 在英文文章中出現的次數。如公式(6)所示，將 *stf* 值及 *idtf* 值進行相乘，可以得到該中英文詞對的 *stf-idtf* 值。由於[15]在公式(6)中是利用英文詞彙 *e* 進行 *stf-idtf* 值的計算，我們在此予以沿用。

$stf(e,c) =$ 詞對集合中的中英文對應詞彙在該中英文句子出現的次數 (4)

$$idtf(e) = \frac{T}{O(e)} \quad (5)$$

$$stfidtf(e,c) = stf(e,c) \times idtf(e) \quad (6)$$

$$\text{中英文句子基礎相似度分數} = \sum_{i=1}^{k} \log(stfidtf(e,c)) \quad (7)$$

如公式(7)所示，假設某中英文句子間其比對到的詞對數共有 k 組（詞對集合中有 k 組詞對），在得到各詞對的 *stf-idtf* 值後，再取對數函數（以 10 為底數）將該 *k* 組詞對的分數進行加總，得到該組中英文句子的「基礎相似度分數」（在此稱為基礎相似度分數，是因為該組中英文句子還需經過向量相似度的計算，最後才會得到該組中英文句子的相似度分數）。

4.4.3 中英句子向量相似度的計算

我們在這篇論文中提出衍生自計算向量之間相似度的方法（餘弦相似度），作為中英文句子相似度的輔助計算，我們稱作「中英文句子間的向量相似度」，這也是本系統相似度計算模組中的第二個部份，常見的餘弦相似度的計算方式如公式(8)所示。

$$\cos\theta = \frac{V_1 \cdot V_2}{\|V_1\| \, \|V_2\|} \quad (8)$$

在公式(8)中，餘弦函數值介於 0 至 1 之間，當兩個向量的夾角越小，則其餘弦函數值越接近 1，代表這兩個向量越相似，反之則越接近 0。

$$ctf(w) = \frac{SO(w)}{L} \quad (9) \qquad idcf(w) = \log\left(\frac{SN}{n(w)}\right) \quad (10)$$

$$ctfidcf(w) = ctf(w) \times idcf(w) \quad (11)$$

我們利用[15]提出的 *stf-idtf* 概念，將 *tf-idf* 的公式，從文件的層級轉化為句子的層級，但不同於[15]，我們將 *tf* 值的計算方法作修改，並將名稱改寫成 *ctf*，如公式(9)所示。在公式(9)中，*w* 代表詞彙，*SO*(*w*)代表詞彙 *w* 出現在該句子中的頻率，*L* 則為該句子的詞數，即句長。我們將 *idf* 值的計算方式作修改，並將名稱改寫成 *idcf*，如公式(10)所示。在公式(10)中，*SN* 為一篇文章中的句子數量，*w* 代表詞彙，*n*(*w*)代表在文章中含有詞彙 *w* 的句子數量。我們將此方法應用於計算中英文句子之間的相似程度，視英文句及中文句各為一個向量，向量中屬性的值為每一個詞彙的 *ctf-idcf* 值，*ctf-idcf* 值的計算方式如公式(11)所示。

我們將 *tf-idf* 值的計算公式進行修改，因為 *tf-idf* 值代表的是某詞彙對其所在文件的權重，我們希望將詞彙對所屬文件的權重，轉化為詞彙對所屬句子的權重。原先存在文件之間的某詞彙，經過其 *tf-idf* 值的計算，我們可以知道其對其所屬文件的重要程度，轉化為句子的層級後，我們就可以知道某詞彙對其所屬句子的重要程度。和[15]提出的 *stf-idtf* 值計算原理不同，他所提出的 *idtf* 值計算方式雖然同樣計算句子中的詞彙重要程度，不過 *stf-idtf* 值乘上 *stf* 值的目的是為了要將 *idtf* 的值加倍，也就是當一組中英文句子間某一個中英文詞彙出現多次，則這一組中英文句子的相關性就更強，而我們提出的 *ctf-idcf* 值計算目的，則是為了要得到句子中的詞彙對於文件中該句子的權重，作為計算向量相似度時使用。

我們的構想為：當一篇中文文章和英文文章若互為翻譯，在正確地中文斷詞及英文詞彙擷取的情況下，英文句子中的每一個英文詞彙都應該對應至中文句子的一個中文翻譯詞彙。我們將中英文句子各視為一個向量，向量中屬性的值為每一個詞彙的 *ctf-idcf* 值，若在英文句子中去除 stop words 的前提下，和未去除 stop words 的英文句子相比較，有去除 stop words 的英文句子在進行中英文句子的向量相似度計算時，其值會較趨近於 1（本系統並未對英文文章進行去除 stop words 的前處理，因此僅將向量相似度計算得到的分數作為輔助分數）。但單純地利用向量相似度計算仍會碰到數個困難點，如：正常情況下，中英文句子長度通常不會相同，導致向量的維度不同而無法計算；中英互為翻譯的句子，因為語言本身的特性，常具有翻譯詞序不同的現象，這將導致在計算向量內積時，會對應到錯誤的的 *ctf-idcf* 值。

基於以上可能會碰到的問題，我們在進行相似度計算前，先對向量內的 *ctf-idcf* 值由小到大進行排序，以解決中英文詞序可能不同的問題。由於中英文句長可能不同，我們將維度較小的向量，進行補足維度的動作，也就是將維度較小的向量，補上 *ctf-idcf* 值為 0 的值，直到中英文句子的維度相同，得以進行相似度的計算。補足維度的方式，同時作為針對中英文句長差異的

E = { [Increasing][LED][directionality][makes][LEDs][more][attractive][for certain]
0.18 **0.18** **0.13** **0.15** **0.13** **0.15** **0.18** **0.18**
[applications][such as][projectors] }
0.18 **0.15** **0.18**

C = { [增加] [發光二極體] [的方向性] [可以] [使發光] [二極體] [對於] [例如] [投影機] [之]
0.10 **0.10** **0.13** **0.06** **0.13** **0.13** **0.11** **0.09** **0.13** **0.01**
[特定] [應用] [變得] [更] [吸引] }
0.11 **0.13** **0.10** **0.11** **0.13**

⇩ 將向量內的 *ctf-idcf* 值進行補齊、排序

V_E = { **0**, **0**, **0**, **0**, **0.13**, **0.13**, **0.15**, **0.15**, **0.15**, **0.18**, **0.18**, **0.18**, **0.18**, **0.18**, **0.18** }

V_C = { **0.01**, **0.06**, **0.09**, **0.10**, **0.10**, **0.10**, **0.11**, **0.11**, **0.11**, **0.13**, **0.13**, **0.13**, **0.13**, **0.13**, **0.13** }

句對餘弦相似度分數：**0.935816**

圖九、向量相似度計算範例

計分，即中英文句子之間的長度差異越大，則其向量相似度的分數，便會越小而越趨近於 0。利用此計算方法的優點在於，互為翻譯的中英文句子長度雖然有差異，但不全然會因為長度的差異而得到過低的向量相似度分數，當中英文句子的向量具有數量及數值差異較小的 *ctf-idcf* 值時，便能夠得到較高的向量相似度分數。由於可能發生權重值相同及句長相同的誤判情況，因此我們僅將向量相似度分數作為輔助扣分的依據。若向量相似度分數越趨近於 1，則在和第一部份的基礎相似度分數（中英文詞彙比對得到的相似度分數）相乘之後，仍保留住分數，反之，相乘之後則形同扣分的作用。

圖九為本系統進行向量相似度計算的範例。在圖九中，*E* 及 *C* 分別代表英文及中文的句子，V_E 及 V_C 分別為代表英文及中文句子的向量。我們將經過斷詞之後的中英文詞彙以括號予以區隔，而在中英文詞彙的下方，我們標記著該詞彙的 *ctf-idcf* 值。由於範例中的中英文句子句長不相同，在英文句子較短的情況下，我們將代表英文句子的向量補上四個值為 0 的 *ctf-idcf* 值，使代表中英文句子的向量維度相同，接著再對向量內的 *ctf-idcf* 值進行排序，最後進行向量相似度的計算，得到中英文句子的向量相似度分數。將此分數與相似度計算模組中的第一部份計算得到的相似度分數予以相乘便得到中英文句子的相似度分數。

4.4.4 對列模式與動態規劃演算法

考慮專利文書文本及其它主題文本的語料特性，我們僅採用 9 種對列模式，分別為：「1：0」、「0：1」、「1：1」、「1：2」、「2：1」、「1：3」、「3：1」、「1：4」及「4：1」。

$$S(i,j) = max\begin{cases} S(i-1,j) + sim(Seg_{i,i}, \emptyset) \\ S(i,j-1) + sim(\emptyset, Seg_{j,j}) \\ S(i-1,j-1) + sim(Seg_{i,i}, Seg_{j,j}) \\ S(i-1,j-2) + sim(Seg_{i,i}, Seg_{j-1,j}) \\ S(i-2,j-1) + sim(Seg_{i-1,i}, Seg_{j,j}) \\ S(i-1,j-3) + sim(Seg_{i,i}, Seg_{j-2,j}) \\ S(i-3,j-1) + sim(Seg_{i-2,i}, Seg_{j,j}) \\ S(i-1,j-4) + sim(Seg_{i,i}, Seg_{j-3,j}) \\ S(i-4,j-1) + sim(Seg_{i-3,i}, Seg_{j,j}) \end{cases} \quad (12)$$

中英文的句子依照對列模式的不同而進行組合，例如：對列模式「1：3」，則本系統會將三句中文句子進行組合成為一句，再將一句英文句子和這組合後的中文句子進行句子相似度的計算，經過相似度計算模組的計算，會得到一個相似度分數，我們沿用[15]採用的最佳對列模式組合的動態規劃演算法，如公式(12)所示，來挑選整篇中英文文章的最佳對列組合。

在公式(12)中，$S(i,j)$代表來源語言文章的 i 個句子及目標語言文章的 j 個句子之間的相似度分數，$Seg_{a,b}$代表中英文章內句子編號 a 至 b 的句子，$sim(Seg_{a,b}, Seg_{c,d})$ 則為兩組中英文句子 $Seg_{a,b}$ 及 $Seg_{c,d}$ 之間的相似度分數，從中英文文章的第一句計算至最後一句，在過程中每一次都記錄該回合最高分的對列模式，最後得到整篇中英文文章的對列總分 $S(i,j)$，沿著計算得到$S(i,j)$的路徑往回推算可以得到每一回合最高的對列分數及對列模式，最後將這些回溯得到的分數及模式列出，可以得到中文文章中第一句至第 i 句之間的句子和英文文章中第一句至第 j 句之間的句子對列結果。

4.4.5 「1：1 信心句對」篩選門檻

參考[19]的作法，我們將擷取平行語料的目標放在 1：1 模式的句對上。我們設計在本系統最後產生對列結果後，能夠依照向量相似度的分數及句對平均比對到的詞數這兩項門檻值進行篩選，讓使用者能夠得到正確率較高的 1：1 句對。句對平均比對詞數計算方式如公式(13)所示。在公式(13)中，將中英文句對比對到的中英文詞對數量除以英文句長，可以得到該句平均比對到的詞對數量，當平均比對到的詞對數量越多，表示本系統在該中英文句對之間比對到越多中英文詞彙，這也表示此句對的對應正確性就越高，越有可能是互為翻譯的中英文句子。

$$\text{句對平均比對詞數} = \frac{\text{句對比對到的詞對數}}{\text{英文句句長}} \quad (13)$$

表三、訂定篩選門檻值之測試數據

語料	語言	句數	總詞彙個數	平均句長	1:1 對列數	1:1 對列錯誤數	精確率
專利公開全文 50 篇	中文	13775	263051	19.1	7131	33	0.995
	英文	10847	146342	13.5			

另一項門檻值則為句對的向量相似度分數，因為我們觀察到具有較低向量相似度分數的中英文句對，通常不會是正確的對列結果，因此採用向量相似度分數作為另一項門檻值條件，利用兩項門檻值條件篩選出 1：1 的句對。為了能夠訂定較為客觀且正確的門檻值，在系統開發的過程中，我們對訓練語料以隨機產生檔案序號的方式抽選 50 篇「專利公開全文摘要」進行對列測試。在本系統完成對列程序後，共產生 7131 組 1：1 句對。我們針對這些句對進行檢查，得到 42 組錯誤的對列結果，在將對列過程中比對成功詞對數為 0 的 1：1 對列結果去除後（在有設定門檻值的情況下，本系統原本就不會對這些沒有比對到詞數的句對作篩選的處理），最後得到錯誤的組數為 33 組，詳細的測試數據，如表三所示。

我們針對這些錯誤的 1：1 結果進行分析，發現這些錯誤的對列結果其平均向量相似度分數為 0.835，句對平均比對到的詞數為 0.218。而這 33 組錯誤結果中，有 28 組的向量相似度分數在 0.94 以下，佔了錯誤組數量的 84.8%；有 32 組的句對比對詞數在 0.34 以下，佔了錯誤組數量的 97%，因此我們訂定「向量相似度分數 0.94」及「句對比對詞數 0.34」作為本系統進行對列結果測試的預設門檻值。

5. 系統效果評估

5.1 實驗語料來源

本實驗採用的對列實驗語料，主要分為專利文書的文本及其它主題的文本。而在輔助式機器翻譯系統翻譯品質評估方面，則依照 5.2 節的實驗設計，也將本系統產生的「1：1 信心句對」分成專利文書文本及其它主題文本兩種，並加以組合，分別進行測試。

本實驗的第一個部份為對列結果的評估，詳細的實驗語料統計數據如表四所示。我們以電腦產生隨機亂數序號代表進行測試的檔案，並以人工的方式作檔案抽取的動作。在專利文書方面隨機抽選的語料，共計有申請號範圍從 091132651 至 095121449「專利公開全文摘要」4998 篇、申請號範圍從 091132651 至 094101510 的「專利公開全文敘述」200 篇（包括了技術領域、先前技術、發明內容及實施方法等四個段落）。在其它主題文本方面，新聞文章有「雙語網站知識管理平台新聞」從 2005 年 8 月 30 日 至 2007 年 12 月 15 日共計 737 篇及「自由時報中英對照讀新聞」從 2005 年 2 月 14 日至 2006 年 12 月 31 日共計 686 篇；科普文章有「科學人雜誌中英對照電子書」從 2003 年 1 月至 2009 年 1 月共計 1745 篇文章及包括了「四技二專統一入學測驗」、「學科能力測驗」及「大學指定科目考試」中「對話測驗」、「綜合測驗」及「閱讀測驗」等多個段落的「大考試題」，共約 130 篇。

在本實驗的第二個部份我們為了進行「1：1 信心句對」的效果評估，將進行實驗的語料產生的「1：1 信心句對」分成專利文書及其它主題文本的句對各為 26401 句及 42010 句。

表四、對列實驗語料統計

語料	語言	文章數	總句數	總詞彙個數	文章平均句數	平均句長
專利公開全文摘要	中文	4998	19899	520475	3.98	26.2
	英文		18968	452150	3.80	23.8
專利公開全文敘述	中文	200	47985	1127704	239.93	23.5
	英文		42072	1016088	210.36	24.2
科學人雜誌中英對照電子書	中文	1745	112649	1871576	64.56	16.6
	英文		117785	2376440	67.50	20.2
雙語網站知識管理平台新聞	中文	737	9272	207580	12.58	22.4
	英文		9408	191051	12.77	20.3
自由時報中英對照讀新聞	中文	686	5523	123803	8.05	22.4
	英文		5594	104699	8.16	18.7
大考試題	中文	131	1534	27937	11.71	17.1
	英文		1604	24152	12.24	15.1

表五、TIMSS2003 實驗組別

八年級 2003 M 組	八年級 2003 S 組	四年級 2003 M 組	四年級 2003 S 組	八年級 2003 MS 組	四年級 2003 MS 組
國中數學 領域試題	國中科學 領域試題	國小數學 領域試題	國小科學 領域試題	國中數學及科學 領域試題	國小數學及科學 領域試題

5.2 實驗設計

表六、翻譯評比實驗組別

組別	建立範例樹語料
A	Chang 科學人
B	專利 1：1
C	一般 1：1
D	專利 1：1 + 一般 1：1
E	Chang 科學人 + 專利 1：1
F	Chang 科學人 + 一般 1：1
G	Chang 科學人 + 專利 1：1 + 一般 1：1

5.2.1 對列測試及對列結果隨機抽驗

在實驗的第一個部份，我們對實驗語料進行對列測試，再以精確率 (precision) 及召回率 (recall) 分別針對專利文書文本及其它主題的文本作評估。參考[19]的作法，我們以隨機抽樣的方式進行檢驗，利用電腦產生隨機檔案序號，並以人工選取這些檔案進行對列結果抽樣檢測及評比。

在對列結果檢測方面，我們的操作方法為：在事先沒有正確對列答案的情況下，我們首先將完成對列的檔案取出（此檔案為本系統對列結果），並且複製一份同樣的檔案進行句對的檢測，若該句對的對列結果是正確的，則註記本系統答對（即對列正確）；若該句對的對列結果是錯誤的，則予以修正至正確的對列結果。當完成此複製檔案的註記後，我們便得到了正確的對列結果及本系統答對的對列結果，加上原先產生的對列結果，便可以進行精確率及召回率的評比。

依照文本的不同，我們將對列結果的評估分為專利文書的對列結果、其它主題文本的對列結果及綜合對列結果三種進行評比。為了比較向量相似度計算機制的效果，特別將該機制移除後，同樣對這些隨機抽樣的檔案進行對列實驗。除了評比本系統的計算機制，我們也以同樣的方式，利用[15]提出的對列工具 Champollion 進行對列實驗，和其比較對列的效果。

5.2.2 利用輔助式機器翻譯系統進行翻譯

在實驗的第二個部份，透過張智傑及劉昭麟[8]於 2008 提出的輔助式機器翻譯系統，對 2003 年國際數學與科學教育成就趨勢調查 (Trends in International Mathematics and Science Study，簡稱 TIMSS) 的試題（以下簡稱 TIMSS2003 試題）進行翻譯效果的評估，以檢視本系統產生的大量平行語料，是否能夠提升現有輔助式機器翻譯系統的翻譯品質。TIMSS2003 試題的實驗組別如表五所示。

參考[8]採用的系統方法及流程，他們的作法分為建構範例樹及翻譯模組兩個部份。在第一個部份，他們將中英文句對作為語料，並利用 StanfordLexParser-1.6 產生英文句子的剖析樹，在中英文詞彙對應後，將中英文句子中對應詞彙順序有前後調換現象的句對，記錄其英文剖析樹及其子樹的樹葉詞彙順序編號及詞性，建構成為範例樹資料庫，在進行翻譯處理時，則將剖析後的目標英文句，依照其詞性進行範例樹資料庫的搜尋比對，若有比對到範例樹，則能夠依照其詞彙順序的編號進行翻譯時詞序的調動，由於僅記錄有詞序調換現象句子的剖析樹，因此大量的平行句對可能在建構範例樹資料庫的過程中便篩選剩下較少的句對。他們的系統第二個部份為翻譯模組，針對英文句子進行中譯的動作，第一個部份的正確詞序調動能夠幫助他們的系統在第二個部份作更正確的選詞。

我們將產生的「1：1 信心句對」視為平行語料，這些「1：1 信心句對」在透過建構範例樹之詞彙對列的篩選機制進行篩選後，專利文書文本及其它主題文本的句對分別剩下 556 句及 608 句。我們同時沿用[8]當時採用的實驗語料以進行比較：在建立範例樹的語料方面，共計有「科學人雜誌中英對照電子書」從 2002 年至 2006 年共 110 篇文章，經過詞彙對列篩選後共計有 30 組句對作為建構範例樹的語料[3]。而在訓練選詞機率模型方面，則有「科學人雜誌中英對照電子書」從 2002 年至 2006 年共 2685 個句對[4]及「自由時報中英對照讀新聞」從 2005 年至 2007 年共 4248 個句對。他們的語料並未和我們進行翻譯品質評估的語料重複。

[3] 該論文作者於之後碩士論文提出翻譯評比分數較高之範例樹語料組合，故於本實驗中予以沿用。

[4] 該論文作者於之後碩士論文提出更新後的數據。

表七、抽樣對列目標統計數據

		專利文書文本		其它主題文本				總和
		專利公開全文摘要	專利公開全文敘述	科學人雜誌中英對照電子書	雙語網站知識管理平台新聞	自由時報中英對照讀新聞	大考試題	
英文句		192	695	340	228	142	231	1828
中文句		177	845	326	228	151	228	1955
1：0 及 0：1	對列數	10	150	9	7	6	4	186
	比例	5.85%	20.6%	3%	3.3%	4.3%	1.8%	10.5%
1：1	對列數	138	425	210	159	115	179	1226
	比例	80.7%	58.4%	70%	76.1%	81.6%	82.5%	69.4%
1：2 及 2：1	對列數	18	81	70	41	16	31	257
	比例	10.5%	11.1%	23.3%	19.6%	11.3%	14.3%	14.6%
1：3 及 3：1	對列數	0	38	6	0	3	3	50
	比例	0%	5.2%	2%	0%	2.1%	1.4%	2.8%
1：4 及 4：1	對列數	2	5	2	0	0	0	9
	比例	1.2%	0.7%	0.7%	0%	0%	0%	0.5%
其它	對列數	3	29	3	2	1	0	38
	比例	1.8%	4.0%	1%	1.0%	0.7%	0%	2.2%
總和		171	728	728	300	209	141	217

1:1 的句對依照文本的不同分為專利文書文本及其它主題文本（以下分別簡稱「專利 1：1」及「一般 1：1」），並搭配[8]當時採用的實驗語料（以下簡稱「Chang 科學人」）共設計出 7 種組合，分別以代號 A 至 G 表示，如表六所示，並和 Google 的翻譯結果進行比較。

5.3 實驗結果與分析

5.3.1 對列效果分析

依照實驗的設計，我們對進行對列測試的語料檔案進行隨機抽樣評比，各文本的句數及對列模式統計數據如表七所示。在表七中可以觀察到，各文本皆以 1：1 的對列模式佔了最多的數量比例。而在「專利公開全文敘述」的抽樣語料中，「1：0 及 0：1」的對列模式佔了約 20%的比例，這顯示專利說明書的內文敘述有許多沒有完整的中英對應。觀察表七，這些語料的「其它」對列模式數量佔了 2.2%，這些多句對應的對列模式並不在我們對列處理的考量內。

表八、移除「向量相似度計算」機制後的抽樣對列結果數據

語料	專利文書文本		其它主題文本		全部文本	
模式	精確率	召回率	精確率	召回率	精確率	召回率
1：0 及 0：1	0.561	0.460	0.333	1.000	0.519	0.491
1：1	0.978	0.868	0.958	0.802	0.967	0.832
1：2 及 2：1	0.545	0.692	0.629	0.897	0.598	0.815
1：3 及 3：1	0.416	0.933	0.412	0.875	0.415	0.918
1：4 及 4：1	0.238	0.833	0.5	1.000	0.280	0.875

表九、加入「向量相似度計算」機制後的抽樣對列結果數據

語料	專利文書文本		其它主題文本		全部文本	
模式	精確率	召回率	精確率	召回率	精確率	召回率
1：0 及 0：1	0.579	0.619	0.393	0.923	0.530	0.661
1：1	0.989	0.913	1.000	0.908	0.995	0.910
1：2 及 2：1	0.652	0.929	0.812	0.981	0.744	0.961
1：3 及 3：1	0.413	0.868	0.588	0.833	0.443	0.860
1：4 及 4：1	0.333	0.429	1.000	0.500	0.400	0.444

表十、Champollion 抽樣對列結果數據

語料	專利文書文本		其它主題文本		全部文本	
模式	精確率	召回率	精確率	召回率	精確率	召回率
1：0 及 0：1	0.485	0.463	0.675	1.000	0.529	0.552
1：1	0.951	0.982	0.991	0.973	0.972	0.977
1：2 及 2：1	0.563	0.831	0.897	0.906	0.739	0.877
1：3 及 3：1	0.489	0.742	0.667	0.667	0.509	0.730
1：4 及 4：1	0.238	1.000	1.000	1.000	0.304	1.000

我們將向量相似度計算機制移除後，進行對列的結果數據如表八所示。從表八中可以發現，僅倚賴中英文詞彙比對的方式進行對列，在全部文本的對列表現，以 1：1 對列模式的精確率最高，但其餘對列模式的精確率皆不高，未達 0.6。而在召回率的部份，觀察數量最多的 1：1 對列模式在全部文本的表現，召回率未超過 0.9，這也表示僅倚賴中英文詞彙比對的方式進行對列，

確實有改進的空間，也因此在研究的過程中，我們加入了向量相似度計算的輔助機制，期望提升整體的對列效果。

本系統在加入向量相似度計算機制後的對列結果如表九所示。在表九中，在專利文書 1：1 對列模式的部份，精確率高達 0.989，這表示產生的 1:1 句對近乎完全正確，召回率也達到 0.913，表示能夠找出大部份正確的 1：1 句對正確。而在 1：2 及 2：1 對列模式的部份，雖然召回率很高，但三種組合在精確率的部份和 1：1 模式相較之下則下降許多，分析其原因，有許多較短的句子會因為詞彙數較少，在中英詞彙比對時並未成功比對到詞彙，但在動態規劃演算法整體計分下，最後會認為這樣的組合總分最高，而產生錯誤的對列結果。在其它的「一對多」對列模式下，也有相同的錯誤原因。從其它主題的文本對列結果可以觀察到，在精確率的部份，以 1：1 及「1：4 及 4：1」對列模式最高，在 1：2 及 2：1 對列模式方面也有 0.812 的精確率，表示本對列系統在 1：1 的對列模式下，不僅在專利文書的文本，在其它主題的文本同樣能有高精確率。在召回率的部份，表現則以 1：2 及 2：1 對列模式最佳，而「1：0 及 0：1」及 1：1 兩種模式的召回率有 0.9 以上，1：3 及 3：1 對列模式有 0.8 以上。和專利文書文本對列結果的精確率及召回率進行比較，我們可以發現在其它主題文本的表現都較佳，最大的因素為其它主題文本的翻譯品質較專利文書整齊，在對列的挑戰上，其它主題的文本較專利文書簡單許多。觀察全部文本的對列結果，在精確率的部份，以 1：1 對列模式表現最佳，高達 0.995，這也表示在近 1200 組 1：1 的模式下，本系統產生的對列答案近乎全對，其召回率也高達 0.910，這也是我們以 1：1 對列結果作為翻譯語料的原因之一。

我們用 Champollion 以同樣的方式進行對列，得到的對列結果如表十所示，與表九的本系統對列結果作比較，可以發現在其它主題文本的對列表現上，在精確率的部份，對列模式「1：0 及 0：1」、「1：2 及 2：1」及「1：3 及 3：1」的結果 Champollion 皆較本系統佳，而在專利文書文本的表現上，僅有對列模式 1：3 及 3：1 的表現優於本系統，這表示在此實驗中，本系統在專利文書文本的表現較優於 Champollion。從全部文本的對列結果進行綜合比較，Champollion 也僅有對列模式 1：3 及 3：1 的表現優於本系統，不過可以觀察到在 1：1 對列模式的精確率及召回率部份，Champollion 雖然精確率較本系統低，但其召回率卻較高，在 1：1 對列模式的部份，本系統與 Champollion 確實皆有很好的表現。

5.3.2 輔助式機器翻譯效果評估

在本實驗的第二個部份，我們利用 BLEU 及 NIST 指標對 TIMSS2003 進行翻譯評測，各實驗組別評比得到的分數如表十一所示。在表十一中，我們以斜粗體字表示各試題組別中除了 Google

表十一、各種組合之 BLEU 及 NIST 評比分數[5]

組別	八年級 2003 M 組		八年級 2003 S 組		四年級 2003 M 組	
指標	NIST	BLEU	NIST	BLEU	NIST	BLEU
A	4.7956	0.1506	4.4854	0.1454	4.1865	***0.1501***
B	4.7911	0.1510	4.4983	0.1467	4.1817	0.1461
C	4.7713	0.1490	4.4941	0.1470	4.1841	0.1464
D	4.7920	0.1518	4.5091	0.1485	4.1866	0.1464
E	4.7893	0.1512	4.4967	0.1466	4.1703	0.1459
F	4.7946	0.1508	4.5003	0.1484	***4.2759***	0.1473
G	***4.7986***	***0.1529***	***4.5111***	***0.1498***	4.1708	0.1459
Google	4.5930	0.1482	5.0538	0.1898	3.7682	0.1046
組別	四年級 2003 S 組		八年級 2003 MS 組		四年級 2003 MS 組	
指標	NIST	BLEU	NIST	BLEU	NIST	BLEU
A	4.2023	0.1072	4.9619	0.1487	4.5040	***0.1251***
B	***4.2262***	0.1075	4.9655	0.1494	4.5167	0.1235
C	4.2074	0.1075	4.9493	0.1483	4.5037	0.1237
D	4.2261	***0.1076***	4.9698	0.1506	4.5188	0.1238
E	4.2252	0.1074	4.9635	0.1495	4.5120	0.1235
F	4.2064	0.1075	4.9668	0.1500	***4.5352***	0.1239
G	4.2251	***0.1076***	***4.9747***	***0.1518***	4.5120	0.1236
Google	4.8162	0.1655	5.0646	0.1637	4.7315	0.1428

[5] 我們於 2008 年 6 月得到 Google 的翻譯結果，同樣的題目，Google 分數較 2007 年時來得高。

外，最高的 BLEU 及 NIST 分數。我們可以觀察得到最高分數的範例樹組合，除了 Google 之外，多數皆落在結合所有範例樹語料的 G 組上，證明了在產生之大量且正確的「1：1 信心句對」後，增加了範例樹的數量，能夠幫助輔助式機器翻譯系統在翻譯時能夠正確地進行詞序的調動，讓產生的 TIMSS2003 中文翻譯句更為通順。

6. 結論

在專利說明書文本中，中英文翻譯並不如其它主題的文本整齊，往往因為主題的不同，文句複雜程度也改變極大，對文句對列的任務而言為極大的挑戰。事實上在專利說明書方面，我們缺乏中英文翻譯對照的標準答案，在利用人工進行對列檢驗的前提下，極耗費時間及人力，而我們需要對更多的語料進行對列實驗的評比，以取得更客觀公正的數據。我們利用精確率及召回率進行對列實驗的檢驗，代表在 1：1 的對列模式下，確實有很好的對列效果；藉由輔助式機器翻譯系統，利用產生的平行語料進行翻譯實驗及比較，證明大量且正確的語料能夠增進 TIMSS2003 翻譯的品質。我們也期望在未來能夠針對[15]提出的詞彙權重計算方式，如：*stf-idtf* 值的計算，進行比較的實驗，並針對專利文書的文本，利用專利文書的平行語料進行中英翻譯的實驗，檢視本系統在專利文書翻譯方面的實質效果。

本中英文句對列系統之建置，透過自然語言處理等技術，以豐富的專業領域資源如：中英技術名詞對應表，改進現有文句對列的工具，藉由方法的改良，使系統能適用於不同文本主題的中英文句對列任務，能夠產生大量、正確且適合作為平行語料的中英文句對。

我們將現有對列工具進行改進，發展出適合於不同主題文本的文句對列技術，不僅止於專利說明書的文本，甚至在語文翻譯與學習的領域上，皆能夠透過本工具獲得豐富的中英文文句對列資源。

致謝

本研究承蒙國科會研究計畫 NSC-97-2221-E-004-007-MY2 的部份補助僅此致謝。我們感謝匿名評審對於本文初稿的各項指正與指導。雖然我們已經在從事相關的部份研究議題，不過限於篇幅因此不能在本文中全面交代相關細節。

參考文獻

[1] 中央研究院中文斷詞系統。
http://ckipsvr.iis.sinica.edu.tw/。最後造訪：2009 年 8 月 8 日。

[2] 中央研究院現代漢語一詞泛讀系統。
http://elearning.ling.sinica.edu.tw/CWordframe.html。最後造訪：2009 年 8 月 8 日。

[3] 自由時報中英對照讀新聞。
http://iservice.libertytimes.com.tw/Service/english/。最後造訪：2009 年 8 月 8 日。

[4] 呂明欣、劉昭麟、高照明及張俊彥。針對數學與科學教育領域之電腦輔助英中試題翻譯系統，*第十九屆自然語言與語音處理研討會論文集*，407–421，2007。

[5] 科學人雜誌中英對照電子書。
http://edu2.wordpedia.com/taipei_sa/。最後造訪：2009 年 8 月 8 日。

[6] 陳光華。超越資訊檢索的語言藩籬，*大學圖書館第二卷第一期*，87–99，1998。

[7] 國立編譯館學術名詞資訊網。
http://terms.nict.gov.tw/。最後造訪：2009 年 8 月 8 日。

[8] 張智傑及劉昭麟。以範例為基礎之英漢 TIMSS 試題輔助翻譯，*第二十屆自然語言與語音處理研討會論文集*，308–322，2008。

[9] 經濟部智慧財產局。
http://www.tipo.gov.tw/ch/。最後造訪：2009 年 8 月 8 日。

[10] 遠東高中．高職英文網站 - 歷年大考試題。
http://www.hsenglish.com.tw/2009/teach/resource/exam_paper.asp。最後造訪：2009 年 8 月 8 日。

[11] 雙語網站知識管理平台新聞。
http://design.taiwannews.com.tw/demosite/2005/rdec/ver10/htm/se-learning01.htm。最後造訪：2009 年 8 月 8 日。

[12] 譯典通線上辭典。
www.dreye.com/tw/dict/dict.phtml。最後造訪：2009 年 8 月 8 日。

[13] HowNet。
http://www.keenage.com/。Last visited on 8 August 2009。

[14] Y. Liu, Q. Tan and K. X. Shen, *Modern Chinese Word Segmentation Specification and Automatic Segmentation Methods for Information Processing (in Chinese)* , Beijing: Qinghua University and Nanning: Guangxi Science and Technology Press, 1994.

[15] X. Ma, Champollion: A Robust Parallel Text Sentence Aligner, *Proceedings of the Fifth International Conference of the Language Resources and Evaluation*, 489–492, 2006.

[16] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.

[17] D. W. Oard, Alternative Approaches for Cross-Language Text Retrieval, *Working Notes of the American Association for Artificial Intelligence Spring Symposiums on Cross-Language Text and Speech Retrieval*, 131–139, 1997.

[18] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1986.

[19] M. Utiyama and H. Isahara, A Japanese-English Patent Parallel Corpus, *Proceedings of the Eleventh Machine Translation Summit*, 475–482, 2007.

[20] P. K. Wong and C. Chan, Chinese Word Segmentation based on Maximum Matching and Word Binding Force, *Proceedings of the Sixteenth International Conference of the Computational Linguistics*, 200–203, 1996.