大學圖書館
2卷1期(民國87年1月) 頁87-99

超越資訊檢索的語言藩籬
Traverse Language Barriers of Information Retrieval

陳光華 **Kuang-hua Chen**
國立臺灣大學圖書館學系助理教授
Assistant Professor, Department of Library Science
National Taiwan University
E-mail: khchen@steelman.ls.ntu.edu.tw

【摘 要】

資訊檢索研究的目的在解決人類對於資訊的需求，發展至今不斷地消除一道道的資訊藩籬。隨著電腦網路的普及，網際網路快速地深入世界的每一個角落，普羅大眾對於「世界村」觀念感同身受的同時，語言的藩籬變得具體而殘酷，使用者很難檢索「不同文」的文獻資料。本文說明資訊檢索的語言藩籬，討論目前語言技術用於處理語言藩籬的可能方案，並且比較現有超越語言藩籬的資訊檢索系統與傳統資訊檢索系統之間的系統績效。

【Abstract】

The purpose of research for information retrieval (IR) is to fulfill information need. IR has eradicated many information barriers from the very beginning. With the fast development of computer networks, the Internet has turned the world into a global village. But when the Internet users try to locate documents in different languages, the once implicit language barrier becomes real and crucial. This paper attempts to explore the language barrier underlying the process of information retrieval, the possible solutions of this barrier, and a comparison of the performance of cross-language information retrieval and monolingual information retrieval systems.

關鍵詞：資訊檢索；自然語言；跨語資訊檢索
Keywords： Information retrieval; Natural language; Cross-language information retrieval

壹、緒論

圖書館一直扮演知識寶藏的角色，盡力地蒐羅圖書資料，而長久以來充實館藏也是圖書館重要的政策。（註 1）然而，若是讀者無法得知有哪些館藏是符合其需求，則再多的館藏對其而言，仍然是沒有使用的價值。所以，圖書館一方面依據政策與文獻計量的原則擴充館藏或淘汰無用館藏，另一方面必須發展圖書資料的組織與整理的方式，並使用某種有效的檢索方式協助讀者取用資訊。

前述的組織與整理正展示於圖書館學發展過程中幾個重要的里程碑，如杜威分類法（DDC）、美國國會分類法（LCC）、美國國會標題表（LCSH）、中文圖書標題表、機讀編目格式（Machine Readable Catalog，簡稱MARC）、英美編目規則（AACR2），與衍生的中國機讀編目格式（Chinese MARC）、中國編目規則等。不同的分類法、標題表以相異的切入角度描述圖書資訊的主題編目（Subject Catalog），而編目規則規範如何進行圖書資訊的記述編目（Descriptive Catalog）；而MARC則記錄著主題編目與記述編目的資料。

早期的讀者使用卡片目錄檢索圖書資料，透過著者卡、題名卡、標題卡提供不同的檢索點（Access Point），滿足讀者各式各樣的資訊需求。圖書館學的學者專家很快地發現前述的方式無法與圖書館其他的作業整合，造成許多重複的作業以及人力的浪費，同時卡片目錄越來越龐大，也很難整理並維護，成本也太高；對讀者而言，使用卡片檢索圖書同樣也非常的不方便，必須先決定要檢索著者、題名、抑或是標題，然後在相應的卡片櫃找尋描述所需圖書的卡片。爲了整合圖書館的作業流程並解決前述的問題，圖書館自動化系統便應運而生，使吾人超越了第一道資訊檢索的藩籬。

採用自動化系統的圖書館，讀者或是使用者只要坐在終端機之前，透過電腦檢索系統，依照各種不同的檢索需求，使用系統提供的檢索點，不必遊走於各個卡片櫃，就能夠瞭解圖書館典藏的圖書是否符合其資訊需求。館員也能夠透過自動化系統，使圖書館的流通作業、編目作業、採訪作業更有效率，提昇整體的服務品質。然而第二道藩籬卻依然豎立於前方等待吾人克服。

當讀者有某種資訊需求時，必須前往圖書館，然後透過檢索系統檢索館藏是否能滿足其需求。然而，當其到達圖書館使用檢索系統後，才發現該館並沒有收藏其所需要的圖書，造成讀者白跑一趟，浪費寶貴的時間與精力，此時橫亙在前的是實際距離的藩籬。當電腦網路出現後，基本上提供吾人消除距離造成的資訊藩籬的工具。讀者可以在家裡或是任何有網路連線的地方檢索圖書資料，並且可以線上預約。更重要的是，透過電腦網路讀者可以翻山越嶺、飄洋過海，檢

索全世界的圖書館或資料中心。從讀者或是使用者的觀點而言，網路世界的圖書館好似一個龐大的聯合圖書館，能夠檢索所有現存的線上圖書資料，有效擴展吾人的視野。然而，這時讀者赫然發現另一道藩籬又出現了，這是當吾人放眼全球時，必然會面對的藩籬－－語言的藩籬。

這道藩籬事實上早已存在，只是以往並不明顯，因爲圖書館外文圖書的使用者通常爲較高級的知識份子，他們在使用這一類圖書時並沒有困難。但是隨著電腦網路逐漸地全民化、商業化，解決這道藩籬的需求越來越急迫，這也是目前全球各地的圖書館學、資訊科學、電腦科學各相關領域的學者專家投入大量研究經費與人力，從事這項重要研究課題的原因。（註2）

本文主要討論的是如何消除第三道藩籬，亦即是語言的藩籬，筆者將說明語言藩籬的屬性，解決方案的初步架構，之後並詳細地討論各種策略的作法與其優缺點。最後則是簡要的結語。

貳、跨語資訊檢索

跨語資訊檢索（Cross-Language Information Retrieval, Translingual Information Retrieval，簡稱CLIR或TIR，註3）研究的目的即在消除因語言的差異而導致資訊取得的困難。至於何謂跨語資訊檢索，有學者以"select information in one language based on queries in another"作爲CLIR的定義，翻成中文則爲：「使用不同於書面語的查詢語言進行資訊的檢索」。希望設定明確的界線，便於學者專家的討論。（註4）

既然牽涉兩種以上的語言，並且限定是以不同的查詢語言檢索圖書資料，因此查詢與圖書資料兩者之一必須進行翻譯，如此查詢問句與圖書資料就屬於同一種語言，之後的處理方式和單語資訊檢索相同。依據前述的作法，吾人可以消除檢索時語言的藩籬，然而使用者閱讀檢索所得之圖書資料時的語言藩籬仍然存在，如果要完全消除語言的藩籬，顯然還是必須引入機器翻譯系統，將檢索所得的圖書資料翻譯爲使用者或讀者所能閱讀的語言。

機器翻譯是極具挑戰的研究領域，對於一般人而言，要真正理解一段文字事實上就不是簡單的工作，遑論使用機器進行翻譯。因爲這牽涉到字（Character）、詞（Word）、語　法（Syntax）、語義（Semantics）、語用（Pragmatics）等層次的知識。例如如何處理未知詞、介詞組的修飾對象爲何，多義詞彙的詞義如何決定，照應詞如何處理等等。基本上，幾乎所有自然語言的現象都得到一個妥善的解決方案時，才能夠建構一套優秀的機器翻譯系統。（註5）日本政府也曾經詳細評估，以目前日本的科技水準與發展的情形，約在2020年才可能有一套商業運轉且績效良好的日英機器

翻譯系統。然而如果跨語資訊檢索系統是使用於特定的領域，則使用機器翻譯系統會有比較好的成效，這是因爲特定領域的自然語言趨於一定的使用方式，比較容易處理。

另一個觀點是將機器翻譯當作輔助的工具，一旦檢索所得的圖書資料翻譯爲使用者熟悉的語言之後，即使翻譯品質不佳，使用者仍然可以判斷圖書資料的相關性，如果有必要再進一步仔細閱讀圖書資料，或是請人潤飾譯稿。這一類的應用以目前全球資訊網（World Wide Web，簡稱 WWW）的文件檢索最爲風行，台灣大學資訊工程學研究所自然語言處理實驗室目前也提供這項服務。（註 6）

一般而言，跨語資訊檢索面臨的問題如下所示：

* 必須翻譯使用者下達的查詢或是檢索的文獻。
* 查詢的問句通常都很短，以美國的研究爲例，通常少於 2 個詞，不會超過 4 個詞（註 7）。因此很難判定詞義。
* 查詢問句中的詞彙通常都有歧義性（Ambiguity）。
* 查詢問句可能必須先分詞（Segmentation），如中文、日文、泰文等等。
* 檢索的文獻可能使用不同的語言，必須先辨識語言（Language Identification）。

目前用於處理跨語資訊檢索的相關技術可以分門別類如圖一所示，主要可分爲翻譯查詢問句與翻譯文件兩類，至於不翻譯的情形不屬於本文討論的範圍。

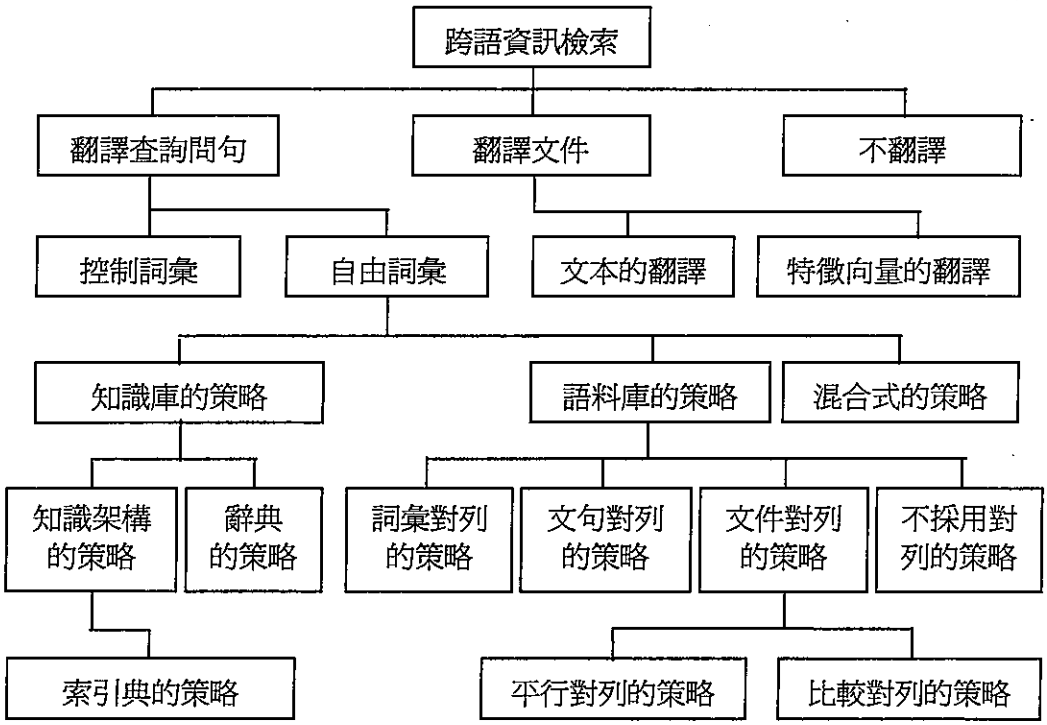
翻譯文件的作法所需的處理時間隨文件的不同而有極大的差異，而且計算量過於龐大，極少有系統採用這種作法。比較實際而且主流的作法是遵循「翻譯查詢問句」（Query Translation）的研究路線。在翻譯查詢問句的範疇之下，目前可能的策略有：知識庫的策略、語料庫的策略、以及混合知識庫與語料庫式的策略。

本文以下的數節將依據圖一說明目前跨語資訊檢索的研究狀況，主要描述辭典爲本（Dictionary-Based）、索引典爲本（Thesaurus-Based）、以及語料庫爲本（Corpus-Based）的策略，討論使用的技術及其優缺點，並且提供現有跨語資訊檢索系統績效的數據，供有興趣的讀者參考。（註 8）

參、辭典爲本的策略

初期的跨語資訊檢索研究多數採用辭典爲本的策略，因爲這是最直接了當的作法（Naive Approach），任何人都想得到，其作法是利用一部雙語機讀辭典（Bilingual Machine Readable Dictionary）將使用者下達的 S 語言查詢詞彙轉換爲 T 語言詞彙。然而，這種作法有幾個問題必須處理：第一是如何處理詞彙的歧義性（Ambiguity）；第二是如何處理未知詞（Unknown Words）。

一、詞彙的歧義性



圖一、跨語資訊檢索的相關技術（註 9）

詞彙通常都有歧義性，如英文的 Bank 有銀行的意思，也有河岸的意思，到底是什麼意思則必須由前後文決定。初期的跨語資訊檢索研究都沒有真正地處理這個問題，主要的因素是處理時間過長，影響使用者檢索的意願。因此替代的作法有如下幾種：

* 選擇排列第一的意義（Select First）
  採用這種作法的學者認爲辭典中詞彙的第一個意義通常是最常用的，因而直接取用第一個意義。
* 選擇所有的意義（Select All）
  因爲無法判斷到底意義爲何，所以所有的意義都視爲詞彙的意義。因而查詢問句中每一個 S 詞彙可能被數個 T 詞彙取代，轉換後的查詢問句相對地變得很大。
* 任選 N 個意義（Select N Randomly）
  因爲選擇所有可能的意義造成轉換後的查詢問句變得太大，修正的作法是任選 N 個意義以控制查詢問句的任意膨脹。
* 選擇最佳 N 個意義（Select Best N）
  採用任意選擇的方式並不具有說服力，因而部份學者利用語料庫計算詞彙不同意義出現的頻率，然後選擇頻率最高的 N 個。這種作法使用了語料庫，也可視爲混合式的作

法，然而最終的意義是由辭典取得，因此本文將其列爲辭典爲本的策略。

採用辭典爲本策略的跨語資訊檢索系統，其檢索成果的精確率（Precision）大約爲原來單語資訊檢索系統的 40-60%。如 Hull 與 Grefenstette 在 1996 年的實驗採用 Select All 的策略，結果顯示由 0.393 降爲 0.235（註 10）；Davis 在 1996 年的實驗也是採用 Select All 的策略，實驗的結果顯示精確率由 0.290 降爲 0.142（註 11）；Ballesteros 與 Croft 於 1996 年的實驗採用 Select N Randomly 的策略，平均精確率降低 50-60%（註 12）；Davis 另外也做了 Select N Best 的實驗，由 0.290 降低爲 0.195。（註 13）辭典爲本的作法，最大的問題在於無法有效處理詞組（Phrase Terms），導致檢索的精確率偏低，Ballesteros 與 Croft 在 1997 年針對詞組對於跨語資訊檢索的影響作了評估，其實驗結果顯示：如果詞組能夠正確地翻譯，可以有效提昇檢索的精確率達 150.3%；若無法正確地翻譯則會降低檢索的精確率達 39.3%。（註 14）因而，有效而正確地轉換詞組將是跨語資訊檢索重要的研究課題。

二、未知詞的處理

未知詞一直是自然語言處理的大問題，由於辭典爲本的策略是以系統的辭典做爲詞彙判斷的依據，因此所謂的未知詞可分爲幾種情形：

* 詞彙正確但辭典並未收錄

因爲辭典不可能完全收錄所有的詞彙，而且隨著時間的推演，新的詞彙不斷的出現，所以未知詞情形也不可能完全消除。

* 專有名詞（人名、地名、機構名）

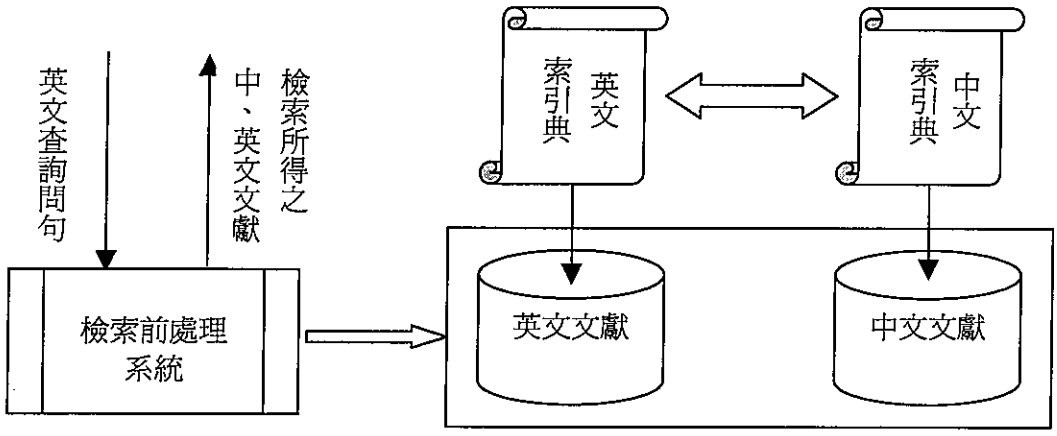
人名、地名等專有名詞也是一直困擾著資訊檢索與自然語言處理的學者專家，這些專有名詞不可能全然收錄，所以如何利用語境判斷詞彙的意義就成爲重要的研究課題。訊息理解會議（Message Understanding Conference，簡稱 MUC）也知道專有名詞的重要性，因而歷年的會議都將資訊檢索系統辨識專有名詞的正確率作爲評估系統的一項依據。（註 15）

* 詞彙錯誤

如果全然是拼字錯誤或用詞錯誤是無法檢索所需的圖書資料。但是，事實上使用者仍然希望能夠檢索出相關的文件，因此模糊處理（Fuzzy Processing）也就成爲這類系統必須採用的處理技術。在這種情形之下，檢索系統必須判斷是否是詞彙錯誤還是未知詞，如果是詞彙錯誤可以使用最近似的正確詞彙取代原來的詞彙，然後再進行檢索。

肆、索引典爲本的策略

以目前實際運作的跨語資訊系統而言，多數是使用索引典爲本的作法，（註 16）索引典提供控制詞彙用以索引文獻資料，不同語言的文獻資料使用各自的索引典，而各索引典之



圖二、索引典架構之跨語資訊檢索

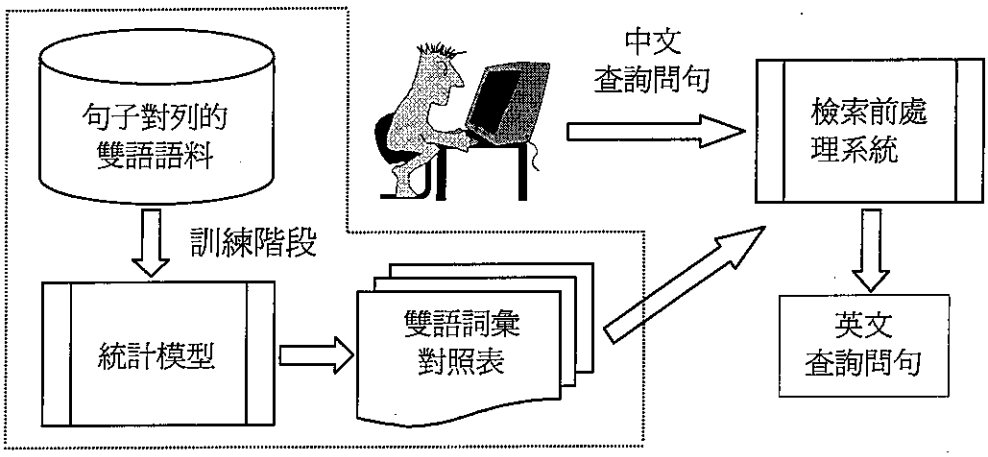
間有對映的關係，透過這種對映關係實現跨語資訊檢索，參見圖二。採用索引典的最大難題是，爲文獻資料設定索引詞彙需要大量的人力與時間，且建構一部高品質的索引典不是一件容易的事；此外，不同語言索引典之間的對映並非直接了當。索引典可視爲組織知識的階層架構，歐洲共同體資助的EuroWordNet（EWN）計畫（註17），目的就希望發展多語言（義大利語、荷蘭語、西班牙語、英語）的概念知識庫，吾人亦可將之視爲多語索引典。

從使用者的角度而言，採用索引典索引文獻資料有另一個難題。索引典內的詞彙通常稱之爲控制詞彙（Controlled Vocabulary），使用者一般並不知道何者是控制詞彙，造成無法有效檢索文獻資料。因此，還必須輔以檢索詞彙與索引詞彙的轉換介面，從而解決上述的問題。

Salton 曾做過跨語資訊檢索的實驗，結論是只要仔細地建構雙語索引典，跨語資訊檢索系統能夠達到單語資訊檢索系統的系統績效。（註18）雖然「典型在夙昔」，然而目前網際網路發展的情況畢竟不同於當時Salton做實驗時的環境，期望跨語資訊檢索系統能夠擁有單語資訊檢索系統的成效，必須投入更多的研究心力，還有很長的一段路要走。

伍、語料庫爲本的策略

基本上，語料庫爲本的技術是希望由大量的語言素材抽取語言知識，而跨語資訊檢索需要的語言知識正如其表面文字所描述的，必須透過兩種以上的語言素材，運用知識擷取的技



圖三、詞彙層次的跨語資訊檢索系統

術，取得跨語資訊檢索系統的跨語言的知識，如詞彙的對照、專有名詞的對譯、詞組的轉換，甚而詞彙的重排。由於系統建構者運用的技術不同，造成所需知識的層次（Granularity）亦有所不同，可以分爲詞彙（Word）、詞組（Phrase）、句子（Sentence）、文件（Document）等層次。以下分別說明各個層次必須處理的問題。

詞彙層次的處理亦即是希望建構跨語詞彙對照表，並且進一步統計詞彙對照的機率。複雜的系統甚至考慮語境（Context）的參數，能夠更加精確地轉換詞彙。基本的作法說明如下，蒐集大量的雙語語料（其中一種語言稱之爲S語言，另一語言稱之爲T語言），並且對列（Align）雙語語料的句子（亦即S語言的句子與T語言的句子一一對應），接著採用統計技術，計算S語言詞彙與T語言詞彙的對應關係。圖三說明詞彙對照的產生方式，以及如何運用於跨語資訊檢索系統。

這一類系統採用的統計模型有許多種不同的作法，以下筆者將說明一種最基本的作法。以S以及T分別代表兩種不同的語言，SS與TS則分別代表雙語語料不同語言的句子，SW與TW代表不同語言的詞彙。假設雙語語料庫已經在句子層次對列完成，因此 SS_1 對應 TS_1、SS_2 對應 TS_2、SS_i 對應 TS_i、...、SS_n 對應 TS_n。若 SS_i 句子包含詞彙 SWi_1、SWi_2、...、SWi_j、...、SWi_m，而且 TS_i 句子則包含 TWi_1、TWi_2、...、TWi_k、...、TWi_l。基於該兩個句子對列的事實，可以認爲這兩個句子中的詞彙應具有對譯的情形，然而目前卻不知道哪些詞彙是互相對

譯的。吾人可假設 SS_i 的詞彙與 TS_i 的詞彙都可能對譯，亦即 SWi_1 與 TWi_1、TWi_2、...、TWi_k、...、TWi_l 都具有對譯的情形，可以記錄爲（SWi_1, TWi_1），（SWi_1, TWi_2），...，（SWi_1, TWi_l），也就是共同出現過一次，對於 SWi_2、...、SWi_j、...、SWi_m 也做同樣的處理。如果雙語語料的數量夠大的話，吾人有理由相信，經過上述的處理，真正對譯的詞彙，其（SWi_j, TWi_k）出現的次數一定最大，可以採用共容資訊（Mutual Information，簡稱MI）計算詞彙對譯的強度。（註 19）MI 的數學式如下所示：

$$MI(SW,TW)=log\frac{P(SW,TW)}{P(SW)\times P(TW)}$$

共容資訊的意義是，當 SW 與 TW 經常一起在語料庫出現，聯合機率P(SW, TW)會甚大於 P(SW)(P(TW)，因此 MI (SW,TW)會甚大於 0；當 SW 與 TW 出現的方式是背道而馳時，MI(SW,TW)會甚小於 0；當彼此沒有什麼關係時（以機率論的術語而言，也就是互相獨立），因此 P (SW,TW)≈P (SW)×P(TW)，所以 MI(SW,TW)接近於 0。

經過上述的計算方式，當共容資訊甚大於 0，可得知該二詞彙對譯的可能性極大，吾人可據以建構一張詞彙對照表。當使用者下達 S 語言的查詢問句時，系統可以參考詞彙對照表，將 S 語言的詞彙轉換爲 T 語言的詞彙，然後再依照傳統資訊檢索系統的作法進行資訊檢索的工作。

詞組層次的處理和詞彙層次大致相同，但是在進行跨語對列處理之前，單一語言必須先行各自處理詞組的辨識（Phrase Identification）或是複合詞的辨識（Compound Identification），然後將詞組與複合詞視爲一般的詞彙，進行詞彙的對列。當然也可以直接進行詞組的對列，然而計算的複雜度相對變高，使得詞組的對列困難許多。在此筆者將簡要說明如何辨識詞組或是複合詞。

所謂的複合詞是指連續的詞彙組合而成，雖然詞組可能由分隔的詞彙組成，但是筆者在這裡不討論這種情形，因此可以將複合詞與詞組一併討論。屬於複合詞的詞彙一起出現於語料的情形一定比隨機出現的機率大很多，基於這個假設，吾人可以由大量的語料庫觀察或計算連續的詞彙組合一起出現的頻率，然後根據統計值判定這種一起出現的現象是否是出於隨機，如果不是，可以相信它們是複合詞或是詞組。這種作法以 Smadja 於 1990 年提出的 N-Gram 的研究方法最著名（註 20），國內的學者專家也曾採用類似的方法建構複合詞辭典。（註 21）

至於實際採用語料庫爲本的跨語資訊檢索系統有Dumais於 1997 年提出以跨語隱含語意索引（Cross-Language Latent Semantic Indexing，簡稱CL-LSI）的策略建構跨語資訊檢索系統（註 22），Oard 則是於 1996 年提出建構於詞彙對列的跨語資訊檢索系統。（註 23）前者提出的系統有一個非常特別

的優勢，亦即使用 LSI 的單語檢索系統與使用CL-LSI的跨語檢索系統，兩者的系統績效相差無幾。而後者的跨語系統仍然比單語系統差（大約降低50%），Oard 計畫未來使用涵蓋面比較廣的雙語辭典配合語料庫改進跨語檢索的精確率與回現率。

前述所提語料庫爲本的策略都需要大量的相互對譯的雙語語料庫，亦即所謂的「平行對列語料庫」（Parallel Aligned Corpus），然而事實上大量的平行對列語料庫並不多，爲了解決這個問題，有些學者專家提出以「比較對列語料庫」（Comparable Aligned Corpus）替代平行對列語料庫。所謂的比較對列語料庫，指的是雙語語料並非相互對譯，僅僅是類似的雙語語料，或許是討論相同的主題，或許是屬於同一類型的語料。事實上，吾人可將比較對列語料庫視爲是由兩個性質相似的單語語料庫構成，因而能夠比較容易蒐集大量的語料，從而取得所需的雙語知識。Sheridan與Ballerini於1996年使用比較對列語料庫建構索引典，然後利用索引典做爲詞彙轉換的依據，實驗結果顯示跨語資訊檢索的平均精確率比之單語資訊檢索降低54%。（註 24）

陸、結論

試圖跨越語言障礙的研究並非始於資訊檢索，發其軔者應爲機器翻譯。機器翻譯的研究幾乎與電腦的發明同時展開，其目的不外乎希望能夠聯繫兩種不同的語言，超越無形的語言、文化的藩籬。早期資訊檢索的研究一直侷限在同一種語言文字的範疇內，因此只需要考慮查詢問句與圖書資料內容如何匹配的問題。但是，網際網路的無遠弗屆卻將這個範疇擴大爲全球，跨越語言的資訊檢索很自然地成爲不可或缺的資訊服務。

檢視目前的跨語資訊檢索系統，多數使用索引典爲本、辭典爲本、以及語料庫爲本的技術，本文亦詳細討論這三種技術的發展現況，表一摘錄不同策略的跨語資訊檢索系統其系統績效的降低幅度。

表一、跨語資訊檢索系統之系統績效

技術分類	系統設計者	相較單語檢索系統績效的降低幅度
辭典文本	Hull and Grefenstette (1996): Select All	51.0%
	Davis (1996): Select All	49.1%
	Davis (1996): Select N Best	32.8%
索引典爲本	Salton (1970)	相近（註 25）
語料庫爲本	Oard (1996): Word-Aligned	50%

吾人可以發現其系統績效普遍比單語資訊檢索系統降低50-60%，這顯示跨語資訊檢索還有很多的發展空間。正因爲如此，有識之士認爲知識的來源不僅一種，完成一項工作的方法也不只一個，每一種知識來源、每一個方法都有其優勢與弱點，因而最新的發展趨勢是結合各種技術，研發混合式的跨語資訊檢索系統。

筆者認爲隨著時間的遞移，資訊交流的方式與資訊科技的結合只會更加地緊密，以全球爲範疇的資訊流動會更加地頻繁，超越語言藩籬的資訊需求會越來越緊迫，跨語資訊檢索的服務將會是解決這項需求的一個重要關鍵技術。

註釋

註1：圖書一詞的來源可以追溯到「河圖洛書」，傳統的圖書泛指所有提供資訊的載體，包括所謂的紙本資料、非書資料等等。在這個意義之下，資訊與資訊載體是合而為一的，因為載體僅能使用一次。然而，在廣泛使用電腦科技的環境中，資訊的載體可以重複使用，此時的資訊成為獨立的個體，不再依存於資訊的載體，因而，筆者在本文使用「圖書資料」或「圖書」一詞時，實際指的是資訊本身，而不論資訊載體的形式。

註2：Salton於1970年就曾經做過類似的研究。而這幾年的資訊檢索研究很明顯展示了這個趨勢，例如1997年美國人工智慧學會（The American Association for Artificial Intelligent，簡稱AAAI）舉辦了AAAI Spring Symposium on Cross-Language Text and Speech Retrieval；而由計算機學會舉辦的資訊檢索領域的重要會議SIGIR有關跨語資訊檢索的論文也越來越多；至於評鑑資訊檢索系統的會議（Text Retrieval Conference，簡稱TREC）也於1996年開辦跨語資訊檢索系統的評鑑工作。

註3：一個研究領域的形成通常經過時間的醞釀，跨語資訊檢索的研究歷經同樣的過程。由於研究者會為這樣的研究設定一個專有名詞作為討論的依據，而且不同的地區可能會訂定不同的專有名詞，對於跨語資訊檢索的研究而言，ACM SIGIR96 Workshop on Cross-Linguistic Information Retrieval訂定的術語為"Cross-Language Information"，而 Defense Advanced Research Project Agency（簡稱 DARPA）則為"Translingual Information Retrie-val"。本文採用的術語是" Cross -Language Information Retrieval "（CLIR），中文為「跨語資訊檢索」。

註4：為了便於討論的進行，本文在討論到多種語言時，使用的詞彙是「雙語」（Bilingual）而不是「多語」（Multilingual），然而並不代表僅限於兩種語言的情形，而且稱使用者使用的語言為S語言，另一語言則稱為T語言；若很明顯地討論跨越兩種以上語言的處理現象時，則使用「跨語」（Cross-language），例如，在說明

資訊檢索系統使用「跨語資訊檢索系統」，以反應使用者可以經由這一類的資訊檢索系統，查檢不同語言的文獻資料。因而，吾人可以說「雙語」指的是靜態地描述物件（Object）的現象；而「跨語」則是動態地說明如何處理或連結兩種以上不同語言的物件，可能是使用者下達的查詢問句（Query），也可能是資訊系統內收藏的文獻資料（Document-like Object）。

註5：陳光華，「語彙知識之擷取與混合式機器翻譯系統之研究」（博士論文，國立台灣大學資訊工程學研究所，民國85年）。

註6：參考文獻請見 Guo-Wei Bian and Hsin-Hsi Chen, "An MT Meta-Server for Information Retrieval on WWW," in Proceedings of AAAI-97 Spring Symposium Series on Natural Language Processing for the World Wide Web, (1997): 10-16.線上輔助翻譯服務請連線 <http://nlg3.csie.ntu.edu.tw/mtir.html>。

註7：Larry Fitzpatrick and Mei Dent, "Automatic Feedback Using Past Queries: Social Searching?" in Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (1996): 306-313.

註8：本文不探討因語言編碼不同造成的檢索藩籬。如中文資料就有 BIG5 碼、CCCII 碼、EUC 碼、GB 碼等等。

註9：本圖引用台灣大學資訊工程學系陳信希教授於民國86年6月2日發表的演講稿。陳信希。「跨語資訊檢索」。電子辭典、機器翻譯與資訊擷取研討會，（民國 86 年 6 月 2 日）。

註10：David A. Hull and Gregory Grefenstette, "Experiments in Multilingual Information Retrieval," in Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (1996).

註11：Mark Davis, "New Experiments in Cross-Language Text Retrieval at NMSU's Computing Research Lab.," The Fifth Text Retrieval Conference (TREC-5), (1996).

註12：Lisa Balleseros and W. Bruce Croft, "Dictionary Methods for Cross-Lingual Information Retrieval," In Proceedings of the 7th International DEXA Conference on Database and Expert Systems, (1996): 791-801.

註13：同註12。

註14：Lisa Ballesteros and W. Bruce Croft, "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval," in Proceedings of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, March 1997, <http://www.ee.umd.edu/medlab/filter/sss/papers> (1 June, 1997).

註 15：訊息理解會議每年針對不同的主題評估參與比賽的系統，以 1997 年為例，設定三種比賽項目，分別是專有名詞的辨識、照應詞的解析、以及腳本樣版資訊的擷取。

註 16：Carol Peters, " Across Language , Across Cultures," D-Lib Magazine, May 1997, http://www.dlib.org/dlib/may97/peters/05peters.htm> (29 May, 1997).

註 17：Julio Gilarranz, Julio Gonzalo and Felisa Verdejo, "An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database, " in Proceedings of AAAI-97 Spring Symposium Series on Cross-Language Text and Speech Retrieval, (1997): 51-57.

註 18：Gerard Salton, " Automatic Processing of Foreign Language Documents," Journal of the American Society for Information Science, 21 (1970): 187-194.

註 19：Kenneth Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," Computational Linguistics, 16:1 (1990): 22-29.

註 20：Frank Smadja and Kathleen McKeown, "Automatically Extracting and Representing Collocations for Language Generation," in Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics, (1990): 252-259.

註 21：中央研究院資訊科學研究所以及台灣大學資訊工程學研究所自然語言處理實驗室都有類似的研究。

註 22：S.T. Dumais et al., "Automatic Cross-Language Retrieval Using Latent Semantic Indexing," in Proceedings of AAAI-97 Spring Symposium Series on Cross-Language Text and Speech Retrieval, (1997): 18-24.

註 23：Douglas Oard, "Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Applications" (Ph.D. diss., University of Maryland, College Park, 1996).

註 24：P. Sheridan and J.P. Ballerini, " Experiments in Multilingual Information Retrieval Using the SPIDER System," in Proceedings of the 19th ACM SIGIR Conference, (1996): 58-65.

註 25：前提是必須仔細小心地製作索引典。