跨語言資訊檢索：理論、技術與應用

Cross-Language Information Retrieval: Theories and Technologies

陳信希

國立台灣大學資訊工程學系

Hsin-Hsi Chen

Department of Computer Science and Information Engineering

National Taiwan University

E-mail: hh_chen@csie.ntu.edu.tw

關鍵詞 (Keywords)：**資訊檢索**（Information Retrieval）；**多語性**（Multilinguality）；**效能評估**（Performance Evaluation）；**詢問翻譯**（Query Translation）

【摘 要】

多語性是網路社會的重要特徵之一，如何將網路資源，介紹給不同語言的使用者，同時吸收其他語言所呈現的資訊，都是資訊國際化不能忽略的重要課題。跨語言資訊檢索，提供使用者以某種語言檢索另外一種語言表達的文件，為近年來很活躍的研究題目之一。本文嘗試將這個研究主題相關的理論和技術，介紹給有興趣的讀者。首先探討詢問翻譯、文件翻譯、和不翻譯等三類基本方法。接著考慮翻譯歧義性和目標多義性，以及專有名詞音譯等進階方法。評估是促進技術進步的必要工作，本文最後也介紹跨語言資訊檢索三大評比：TREC、CLEF、與 NTCIR。

【Abstract】

Multilinguality is one of the major characteristics in network era. The trend toward information globalization has brought new challenges for information management. On the one hand, it is often necessary to share the valuable resources on the web with users of different languages. On the other hand, it is also necessary for a user to utilize knowledge presented in a foreign language. This paper introduces related theories and technologies of cross language information retrieval, which is kernel in multilingual information management. The basic concepts are presented in sequence on the basis of the classification of query translation, document translation, and no translation. Besides, some advanced concepts like translation ambiguity and target polysemy, as well as proper name transliteration are discussed. Performance evaluation is indispensable for improvement. This paper also shows three world-wide IR evaluation, including TREC, CLEF and NTCIR.

1. 導論

新一代資訊傳播的特色是：網際網路突破空間距離，打造一個不分國界的資訊地球村。尤其透過全球資訊網，各地的資訊皆唾手可得，不但豐富且即時。在網際網路上流通的資訊除了數量非常龐大之外，所使用的語言種類也非常多。依據 2000 年ETHNOLOGUE目錄上的

統計，全世界語言數目高達 6,809 種[1]。真實世界語言使用人口數，前幾名依次爲中文、英文、印度文、西班牙文、葡萄牙文、孟加拉文、俄文、阿拉伯文、日文。但根據 2001 年 3 月份的統計估算，在網路世界語言使用人口數，前幾名依次爲英文(47.5%)、中文(9.0%)、日文(8.6%)、德文(6.1%)、西班牙文(4.5%)、韓文(4.4%)、法文(3.7%)、義大利文(3.1%)、葡萄牙文(2.5%)、俄文(2.1%)。而網際網路內容所使用的語言比例，前幾名依次爲英文(68.39%)、日文(5.85%)、德文(5.77%)、中文(3.87%)、法文(2.96%)、西班牙文(2.42%)、俄文(1.88%)、義大利文(1.56%)、葡萄牙文(1.37%)、韓文(1.29%)[2]。另外，Hershman在 1998 年，曾引用較舊的資料，指出在全球資訊網上，大約 80%的網站是英文網站，而將近 40%的網際網路使用者不會英文。因此，如何將網路資源，介紹給不同語言的使用者，同時吸收其他語言所呈現的資訊，都是資訊國際化不能忽略的重要課題。

多語言處理的應用相當廣，以目前極爲受到重視的數位圖書館(digital library)計畫爲例，數位圖書館所擁有的大量數位化資源，扮演網際網路內容提供者的重要角色，發揮文化傳播、教育、陶冶性情等多重功能。數位圖書館是內容與技術的整合，基本上有下列三個 M 的特色(Borgman, 1997)，其中多語言處理在許多數位圖書館計畫，都被列爲關鍵技術之一。

(a) 多媒體(multi-media)：透過不同媒體所呈現的內涵，引導不同層面的使用者，吸收數位圖書館的精華。

(b) 多語言(multi-linguality)：網際網路所帶來的無國界特質下，如何降低語言的障礙，呈現數位圖書館的內涵是重要課題。

(c) 多文化(multi-culture)：由於網際網路的特殊資訊傳播功能，各個數位圖書館典藏的交流，會越來越密切。重要內涵彼此的觀摩，更帶動多重文化的比較，促進文化的融合。

設計一套多語言資訊系統，必須考慮四項要素(Bian and Chen, 2000)：

(1)資料輸入：資料輸入的方法，例如中文有注音輸入法、倉頡輸入法等。

(2) 資料表現和傳輸方式：這牽涉到字元集合，編碼系統，和傳輸規約。

(3) 資料運算：例如資訊檢索。

(4) 資料輸出：資訊的呈現，例如字型對應及產生。

在這四項要素中，資料輸入、表現和輸出，已經很成熟，直接引用現有的系統。而在資訊傳輸與資源共用上，詮釋資料(metadata)(Andresen, 1997; Bearman, 1996; Caplan and Guenther, 1996)是不可缺少的重要機制，國內外有很多文獻探討這個課題，台灣大學-台灣師範大學圖書資訊系所團隊曾研發一套著名的系統：Metalogoy(Chen, Chen and Chen, 1999)。因此，本文專注於中英語言差異部份　跨語言資訊檢索理論與技術。所謂跨語言資訊檢索，是提供使用者以某種語言檢索另外一種語言表達的文件。過去這項研究，英文使用的名稱非常分歧，直至 1996 年在 ACM SIGIR Workshop for Multilingual Information Retrieval，經與會人員討論，將其定名爲 Cross-Language Information Retrieval。然而大約在同一時間，美國 Defense Advanced Research Project Agency (DARPA)，也將這項研究給予另一種稱呼：Translingual Information Retrieval。不管是那種稱呼，其研究目標一致，都是希望在多語的資訊時代，提供跨語的檢索服務。

在跨語言資訊檢索的研究上，近幾年有多項國際會議舉辦專題演講(Chen, 1997, 1998; Hovy and Idel, 1998; Grefenstette, 1998)、甚至舉辦特定主題國際會議 (Grefenstette, 1996; Oard, 1997a; Vossen, 1997)、數位元圖書館系列論文(Borgman, 1997; Oard, 1997b; Powell and Fox, 1998)。著名的計算語言學和資訊檢索領域國際會議，如 ACL Annual Meeting，ACM SIGIR99 (SIGIR00)等，都有特別的議程探討跨語言資訊檢索的發展。ACM SIGIR02 並有一研討會，由三個主要跨語言資訊檢索評比組織：TREC、CLEF、與 NTCIR 共同規畫，擬討論未來幾年的評比重點。本文擬介紹這個研究領域主要的問題、過去所發展的技術、以及評估方法，提供有興趣從事這領域研究的人員參考。

2. 主要的問題

在思考及提出可行的解決方案之前，我們先分析跨語言資訊檢索的特徵，以下列出其中幾個主要的問題：

(a) 詢問(query)與文件(document)分屬不同語言

這是跨語言資訊檢索主要的特徵，因此詢問與文件

[1] http://www.ethnologue.com/ethno_docs/distribution.asp

[2] http://www.emarketer.com

之間，必須有一個對應，翻譯是重要的運算之一。

(b) 詢問中的詞可能是多義

原始詢問(source query)的歧義性(ambiguity)，必須輔以歧義性分析機置，翻譯後目標詢問(target query)的多義性(polysemy)等問題必須克服。

(c) 詢問通常很簡短

由於使用者的習性，輸入的詢問非常簡短，這增加歧義性分析和翻譯的困難度，進而影響檢索的品質，適度的詢問擴張(query expansion)是可能考慮的方向。

(d) 詢問中詞的決定

一些語言，例如中文、日文、韓文等，詞與詞之間並沒有明顯的分隔符號，斷詞(segmentation)在此也是個問題。

(e) 文件的多語性

網際網路上的文件由不同的語言表達，語言識別(language identification)是檢索的基本工作。

(f) 輸出結果的呈現(visualization)

檢索所得的多語言文件，如何分辨彼此間的分數差異，以及合併不同語言文件檢索結果，呈現在使用者面前，也是跨語言資訊檢索必須面對的挑戰。

本文的重點，放在(a)、(b)、(c)等問題的介紹，斷詞部份請參考(Chen and Lee, 1996)，語言識別部份請參考(Su, et al., 1998)，合併不同語言文件檢索結果請參考（Lin and Chen, 2002）。

在分析問題之後，我們回過頭來看看，傳統的資訊檢索系統什麼地方必須加強，以解上述問題。圖一是傳統資訊檢索的模型，以及可能的切入點：

資訊檢索的基本目標是由龐大的數位化資料庫中，依使用者本身的需求，提供相關資料的特徵，以檢索出所要的資料。因此，資訊檢索基本架構主要的工作有三：

(a) 將每一物件的意義或內容表達儲存下來。

(b) 將所要檢索物件的特徵表示出來。

(c) 比較上述兩項表示式，以找出滿足條件的物件。

其中必須考慮的問題如下：

(a) 如何把物件表示出來？可供檢索的單位是什麼？他們之間的關係如何安排？如何把原來資料轉換成所要的結構？

(b) 如何將使用者的資訊需求表達出來？

(c) 如何比較敘述的相似度，以檢索出最近似的物件？如何將結果呈現出來？如何評估檢索過程的效能？

以跨語言資訊檢索而言，可以切入的點有四：

(a) 文件翻譯：原始文件(source document)翻譯成目標文件(target document)，再進入表示式的階段。

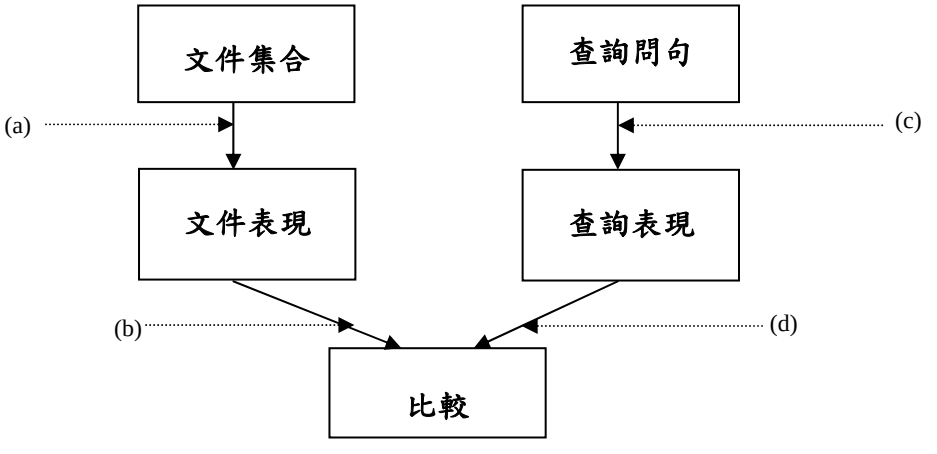
(b) 向量翻譯：原始文件轉換成特定的表示式後，代表原始文件的向量再經過翻譯。

(c) 詢問翻譯：原始詢問經過翻譯。

(d) 項向量翻譯：代表原始詢問的向量經過翻譯。在詢問非常簡短的情況下，(c)與(d)差異並不大。

3. 基本方法

過去跨語言資訊檢索可能的策略，基本上可以區分出詢問翻譯(query translation)、文件翻譯(document translation)、和不翻譯(no translation)等三類，以下小節分別描述各項技術的特點。



圖一　傳統檢索系統架構

3.1 詢問翻譯

根據翻譯所使用的資源，將詢問翻譯的模式進一步區分成知識爲本(ontology-based)、語料庫爲本(corpus-based)、和混和式(hybrid)的方法。

3.1.1 辭典為本的方法

辭典是很基本的知識來源，知識爲本的方法之一就是採用機讀辭典來做翻譯。這裡主要的問題是詞彙的歧義性，一個詞彙可能有多重意義，因此產生類似一般機器翻譯系統選詞(lexical selection)的問題。另一個問題是辭典本身的覆蓋度，詢問中的檢索詞彙在辭典中可能找不到。以選詞而言，有幾個考慮的因素：

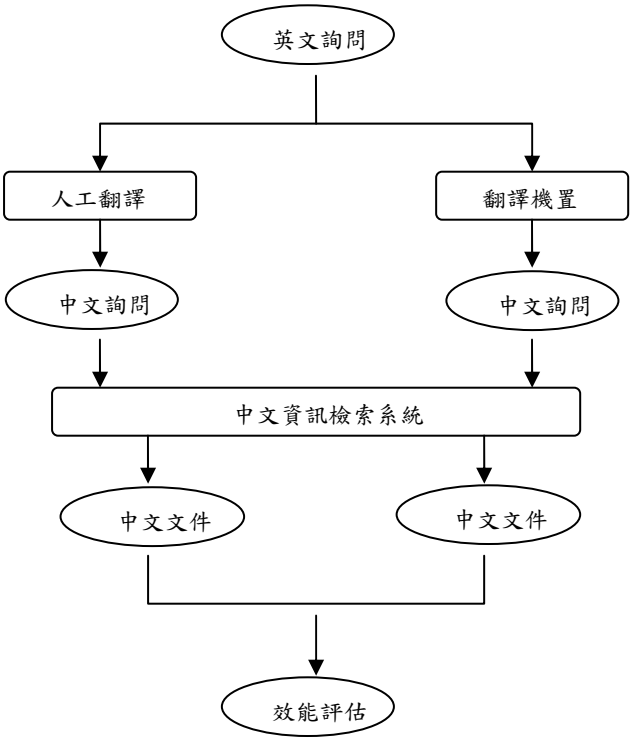
(a) 選擇的方法：全部選、任選 N 個、選「最好」的 N 個。

(b) 選擇的層次：詞彙、還是詞組。

圖二是標準的實驗架構，以英文詢問檢索中文文件集爲例。英文詢問以兩種翻譯：機器和人工，產生兩組中文詢問，模擬有無翻譯錯誤的情況。這兩組中文詢問分別輸入單語的中文資訊檢索系統，產生兩組中文文件檢索結果，最後比較效能差異。

Hull & Grefenstette (1996)曾做一個實驗，將辭典所查到的詞彙，不作任何歧義分析。在單語的資訊檢索系統，有 0.393 的平均準確度。在跨語檢索，則有 0.235 的平均準確度。跨語檢索只有單語檢索 59.8%的準確度，下降了近 40%。這項實驗反應，即使是最簡單的模型，也有不錯的效能。當然這跟測試的語料(詢問、檢索語料庫)、與環境(檢索系統)有極緊密的關係，數據只能供參考。Davis (1997)在 TREC 5 的實驗裡，採用同樣最簡單的方法，在不同的條件下，單語檢索平均準確度是 0.2895，跨語檢索有 0.1422，近 49.12%的效能，這同樣反應上述的論點。

當把選擇的方法換成任選 N 個，Ballesteros & Croft (1996)將詢問中的每個詞彙去查辭典，每個詞彙都只選擇第一種詞義，並以定義中的詞彙來代表原來的詞彙。在這樣的策略下，平均準確度比單語檢索掉了 50-60%。Kwok (1997)在英文-中文跨語檢索，做了類似的實驗，但由詞彙對應的三種詞義中各挑一個(第一個)。實驗顯示比所謂最好的翻譯(人翻譯)結果，差 30-50%。



圖二　標準實驗架構

Hull & Grefenstette (1996)分析最直觀的翻譯模式，發現主要錯誤的來源有二：多詞詞組(phrase)的翻譯和歧義性。Ballesteros & Croft (1997)針對多詞詞組翻譯做了實驗，發現”好”的詞組翻譯的確有很好的影響，但”不好”詞組翻譯卻產生反效果，比詞對詞直接翻譯差了39.3%。Davis (1997)對歧義性的問題進一步探討，嘗試以類別(part of speech)挑翻譯詞彙。整個相對效能爲：單語(0.2895)-全選(0.1422)-類別策略(0.1949)，平均準確度提昇到單語檢索的 67.3%。效果很不錯，當然類別標記程式會引進一些錯誤。Hayashi, Kikui and Susaki (1997)在他們的TITAN 系統(全球資訊網跨語查詢引擎)，以頻率爲標準，選頻率最高的詞彙爲翻譯，但這項作法沒有實驗數據可供參考。

3.1.2 語料庫為本的方法

語料庫(corpus)根據對應的程度，可分成詞彙對列(word alignment)、句子對列(sentence alignment)、文件對列(document alignment)、及不對列(no alignment)四種。詞彙對列是其中最細緻的雙語語料庫(bilingual corpus)，語料庫中不同語言詞彙間的關係，已經經過人工或機器建立連結。Oard (1996)和 Brown (1996) 都曾由這種語料庫粹取出翻譯表(某種型態的雙語辭典)，供查詢翻譯使用。在這種作法裡，語料庫對列的準確度對查詢翻譯影響相當大，而這也是對列(alignment)研究的主要問題(Chen and Chen, 1994)。

Davis & Dunning (1996)在TREC4 利用句子對列語料庫，做了一系列的實驗。第一種作法先用原始詢問(source query)由雙語語料庫找出 100 篇文件，統計這些文件詞彙出現的頻率，去除前 500 個頻率高的詞彙，由剩下的詞彙中找前 100 個頻率較高的詞彙當檢索詞彙用。這種作法規避了直接分析歧義性的問題。第二種作法，差異只在於用統計上的公式去篩選出檢索詞彙。其實驗包括單語檢索、辭典方法、Evoluationary Programming、χ^2篩選、Singular Value Decomposition、和頻率篩選之效能。實驗結果顯示辭典方法比語料庫的χ^2篩選和頻率篩選好，經分析後發現錯誤與語料庫的主題有關。

其次是運用文件對列的語料庫來解問題，這種語料庫可分成兩類：平行語料庫(parallel corpus)和比擬語料庫(comparable corpus)。前者是指同一文件，兩(多)種語言對譯；後者爲同一主題(或事件)，兩(多)種不同語言的描述。後者的定義較前者寬鬆，因此理論上較容易取得大量的文件。Sheridan & Ballerini (1996)就曾經把德文和義大利文的新聞，根據主題及時間對列，製作出虛擬的平行文件。接著，抽取出翻譯辭典，用來產生目標詢問(target query)。平行語料庫也可用來產生雙語辭典，其他用法後面會再說明。

最後來看未對列的語料庫，兩(多)個單語語料庫就構成未對列的語料庫，這個定義又比比擬語料庫更寬鬆，又更容易取得。但通常必須配合其他方法，例如辭典、區域性的回饋(local feedback)等，才能發揮功能。

3.1.3 混合式的方法

前面兩種作法的缺點，歸納如下：

(a) 辭典爲本(dictionary-based)的方法
不在辭典的詞就無法翻譯，通常直接送入檢索系統，這個詞的檢索功能相對有限。另外，歧義性加入不少錯誤的檢索詞。

(b) 語料庫爲本(corpus-based)的方法
平行語料庫取得不易，即使有量也不夠大、包含的主題不夠多，而且檢索效能跟對列的品質有極密確的關係。

雖然辭典爲本的作法有缺點，但卻已經有單語檢索 50%的效能。其實辭典和語料庫是互補的，辭典提供較廣(一般)、較淺的覆蓋度；反之，語料庫提供較窄(領域相關)、較深(即時反應現在用語)的覆蓋度。因此，如何將其整合是研究的重點。傳統單語檢索的詢問擴張技術，是整合兩種方法的橋樑。採用這種技術，必須考慮幾個要素：

(a) 語境：區域性回饋(local feedback)，和區域性語境分析(local context analysis)

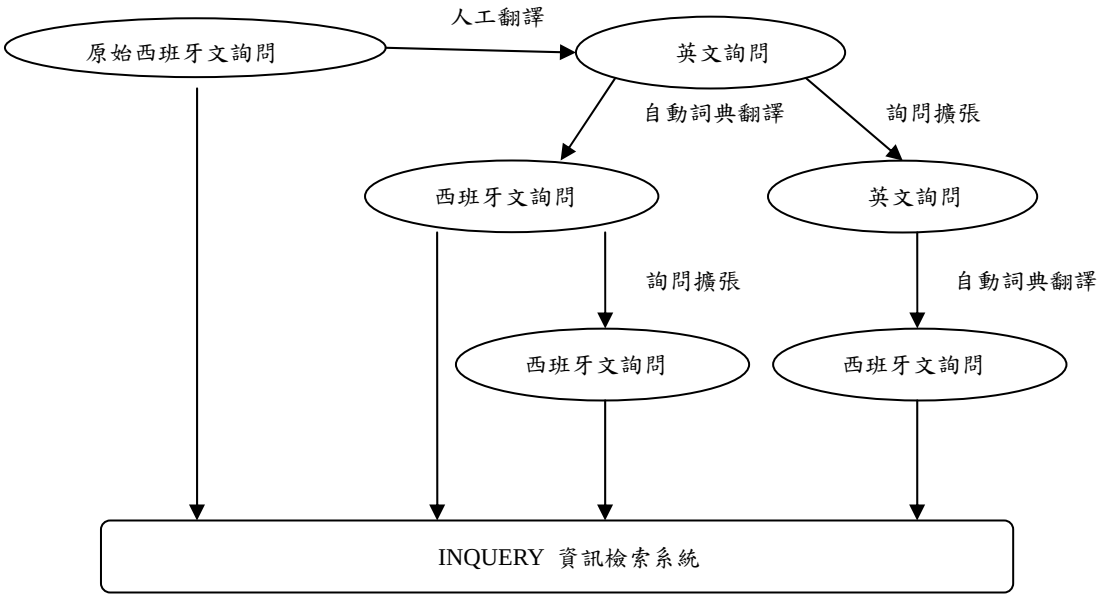
(b) 順序：翻譯前，或/且翻譯後。

Ballesteros & Croft (1997) 曾就不同的組合，做了一系列的實驗，圖三是其實驗規畫。實驗結果摘要如下：

(a) 翻譯前詢問擴張
區域性回饋相較於純辭典策略提昇 33.5%，區域性語境分析則增加 38.5%。

(b) 翻譯後詢問擴張
區域性回饋相較於純辭典策略提昇 11.3%，區域性語境分析則增加 24.1%。

(c) 翻譯前後都做詢問擴張
區域性回饋相較於純辭典策略提昇 51.0%，區域性語境分析則增加 65.0%。



圖三　Ballesteros and Croft (1997)的實驗規畫

整體而言，最好的方法仍比單語檢索差 32%。目前比較困擾的一個問題是：到底那種方法好，不同研究人員所做的實驗，都是在不同的條件下完成，即使使用 TREC 的訓練和測試語料，還會跟所使用的單語檢索系統(例如 Smart, Wais, Inquery 等)息息相關。因此，都只有相對效能供參考。我們再看另一組實驗，Davis (1997) 在 TREC5 設計的實驗。這項實驗比較詞性類別(part of speech)和語料庫爲本的歧義性分析之差異，結果爲全選、語料庫解歧義、類別、語料庫和類別兩者都用等四個策略，分別有單語檢索的 49.12%，39.83% ，67.32%和 73.47%。效能已經提昇到將近 75%，這項實驗少考慮詞組的因素。

3.2 文件翻譯

文件翻譯是把資料庫的文件翻譯成與詢問同一種語言，再進行檢索。如果以機器翻譯系統來做，馬上會有執行效率的問題。根據一項分析，如果嘗試翻譯 40 億篇網頁，以一部高速的個人電腦，需要 300 年。而以 3600 部個人電腦一起進行翻譯，則需要 1 個月。Oard(1998)曾以 4 部 SPARC 20 和 1 部 Ultra SPARC 1 發了 10 個機器-月的時間，將 251,572 篇原始文件翻譯成目標文件，進行跨語檢索實驗。這個層次的翻譯相對於詢問層次的優點是：語境比較寬，歧義性分析所能用的線索較多。Oard 以機器翻譯系統翻譯長的詢問，實驗結果比一般詢問翻譯的策略好。在較長的語境下，機器翻譯系統翻譯文件的結果比翻譯詢問顯著。關於這一點，Davis(1997)使用語料庫來解歧義，其實已經隱含這個目的。由於執行效率的考量，文件翻譯是有必要才做，例如已經確定要瀏覽某一網頁(Bian and Chen, 2000)。目前沒有實驗系統採用這樣的策略，但有人使用 Systran (Gachot, Lange and Yang, 1996)來產生文件對應雙語語料庫，供做區域性回饋。另一種想法是：對每個文件所對應的向量進行翻譯。

3.3 不翻譯

Bellcore (Deerwester, et al, 1990)曾在單語資訊檢索，提出一種稱爲 Latent Semantic Indexing (LSI)的方法。Dumais 等人 (1997)進一步把這種觀念引進到跨語資訊檢索，在其訓練過程，英法雙語文件、英語詞彙、法語詞彙都被對應到向量空間。英(法)文件向量可以英(法)文件向量、或英(法)詞彙向量表示。沿用 LSI 的基本想法，不管是不是同一種語言，這些文件可以在向量空間呈現出來。因此，這種作法不需翻譯。過去有多人在不同的語言配對上做過實驗，Berry & Young (1995)以希臘文-英文、Oard (1996) 以西班牙文-英文等。卡內基美隆大學語言技術研究所 (Carbonell, Yang, Frederking, et al., 1997)，曾對語料

庫導向的翻譯方法(TMT)、虛擬相關回溯(Pseudo Relevance Feedback, 簡稱 PRF)、一般化的空間向量模型(Generalize Vector Space Model, 簡稱 GVSM)、和 LSI 等四種方法，在相同的條件下，做了一系列的實驗，結果顯示 GVSM 比 LSI 稍微好一點，這兩種方法又比 TMT 和 PRF 好。

4. 進階方法

在基本方法的研究中，主要的目標是將詢問和文件均一化成單一語言，再進行資訊檢索。在常被採用的詢問翻譯策略，原始詢問的歧義性分析(translation disambiguation)是研究重點，但都沒有考慮目標詢問的多義性(target polysemy)(Chen, Bian, and Lin, 1999)。辭典覆蓋度的問題，是影響檢索效能的因素之一，如何進階式的自動建立雙語辭典，以及解決專有名詞未被收錄的問題，一直是研究人員努力的方向之一。在以下進階方法的討論，就針對原始詢問的歧義性與目標詢問的多義性、和專有名詞音譯(proper name transliteration)的問題作介紹。

4.1 翻譯歧義性與目標多義性

翻譯歧義來自原始詢問(source query)，而目標多義(target polysemy)則來自翻譯後的目標詢問(target query)。以中英資訊檢索爲例，中文檢索詞「銀行」本身沒有歧義性，但其對應的英文翻譯「bank」，則有 9 種意義(Longman, 1978)。當採用基本的詢問翻譯模式，「銀行」很直接得被翻譯成「bank」後，送入英文資訊檢索系統。因爲「bank」的多義性，可能會有不相關的文件被提出來。反過來當「bank」作爲英中資訊檢索系統的檢索詞，首先必須解其歧義性。如果知道正確中文翻譯是「銀行」，中文資訊檢索端直接找出含有「銀行」的文件。翻譯歧義性和目標多義性乘在一起挑戰性更大，例如「運動」有如下的意義(Lai and Lin, 1987)：(1) sport，(2) exercise，(3) movement，(4) motion，(5) campaign，和(6) lobby。而每一個對應的英文詞可能有一個以上的意義，例如「exercise」有“a question or set of questions to be answered by a pupil for practice”；“the use of power or right”等意義。

Chen 等人(1999)曾用原始詢問的語境，當作目標詢問的限制條件，來解決目標多義性的問題。例如中文詞「銀行」在中研院平衡語料庫(Huang, et al., 1995)的相關詞有“貼現”，“領出”，“押匯”，“匯兌”等詞彙，在送入英文資訊檢索系統時，造出虛擬的語境(pseudo context)，以限制「bank」的語意範圍。Chen 等人(1999)以共現模型(co-occurrence model)分析翻譯歧義，和限制語境模型分析目標多義，在 TREC-6 的資料集、和主題 301-350 評估條件下實驗，有單語檢索 62.92%的效能。與僅處理翻譯歧義性比較，提昇 10.11%的效能。

4.2 專有名詞音譯

根據 1995 年網路使用者，對 Wall Street Journal、Los Angeles Times 和 Washington Post 等新聞語料檢索的統計(Thompson and Dozier, 1997)，分別有 67.8%、83.4%、和 38.8%的檢索詞含專有名詞。我們知道辭典的覆蓋度，未收錄詞一直是詢問翻譯的重要問題，在專有名詞的翻譯更是挑戰。Chen 等人(1998)、Knight 和 Graehl(1998)、Lin 和 Chen(2000)、Wan 和 Verspoor(1998)都相繼提出機器音譯(machine transliteration)的方法，來處理這個問題。

音譯可以根據處理的方向，區分成正向音譯(forward transliteration)與反向音譯(backward transliteration)(Lin and Chen, 2000)。當一個語言的專有名詞，因爲沒有適當或是不容易以意譯來表示時，會採用正向音譯，將其音呈現出來。例如義大利的觀光勝地 Firenze，中文就音譯成「翡冷翠」，此爲正向音譯。反過來說，當看到一個中文的音譯人名「阿諾史瓦辛格」，如果想要找出原文是 Arnold Schwarzenegger，就是反向音譯。一般來說，使用羅馬字母的拼音文字語言，會保持原詞語字母的拼法，以原語言的發音規則，或是自己語言的發音規則來發音。但在象形文字與拼音文字語言之間作音譯時，則需要將聲音由原語言盡量用另外一種語言相近的音素來表示，而且要符合目的語言(target language)的語音組合規則。很顯然地，拼音文字與象形文字之間的音譯處理相對來說較爲困難，反向音譯比正向音譯更難。正向音譯允許某種程度的失真，所能夠接受的錯誤範圍較大，但反向音譯則不是。反向音譯較不允許錯誤，也就是在找出原文的過程中，必須要相當準確，否則反向音譯的結果應用性就較低。

Chen 等人(1998)提出一個將英文音譯成中文(目的語言)的音譯字，反向音譯回英文(原始語言)的模組，並應用於中英跨語言資訊檢索系統。這個系統是將可能的音譯字辨識出來，再進行反向音譯。首先利用漢字羅馬拼音系統(例如 Wade Giles (威翟)，或是漢語拼音(Pinyin))，把可能的音譯字(中文)轉成羅馬字母。接著將這個詞彙與一串

可能的專有名詞進行比對，藉此找出可能的原文(英文)。以下是一個範例，如果輸入的檢索詞是「埃斯其勒斯」，轉換成威翟羅馬拼音表示式是「ai.ssu.chi.le.ssu」，其對應的英文原詞彙是「Aeschylus」。這個方法嘗試計算兩個英文字串的相似度，可以考慮多個策略：

(a) 字元相似的個數

aeschylus
ais suchilessu

羅馬拼音共有九個字元，一共對應三個字元，分數爲 3/9。

(b) 音節內字元相似個數

aes chy lus
aissu chi lessu

共可切出三個音節，在音節對應的條件下，一共對應了六個字元，分數提高爲 6/9。

(c) 權重法

aes chy lus
AiSsu Chi LeSsu

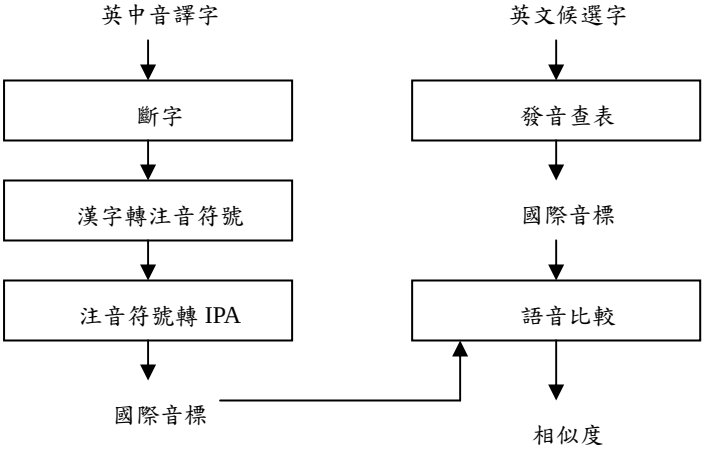
整合不同羅馬拼音系統，並給予第一個匹配的字元較大的權重，這個範例的分數提升到 0.83。

Chen 等人(1998)的研究，可以視爲在形素上比較相似度的反向音譯系統。由於羅馬拼音系統，主要並不是考慮語音上的相近來設計。例如漢語拼音就用到了 Zh、Q 與 X 等羅馬字母，來表示與字母發音完全無關的漢語語音。因此，英文音譯成中文的音譯字，在利用羅馬拼音系統轉換成羅馬拼音字母後，這些羅馬拼音字母，跟原來詞彙的拼音字母，在發音上並不十分相近。Lin 和 Chen(2000) 提出一個以音素相似度爲基礎的方法，以中文和英文的音譯爲例，進行反向異文字的音譯。圖四是實驗流程。

實驗結果顯示在音素上的比較，比在形素上的比較來得有效。在一個 1,261 個人名的候選名單中，執行配偶配對(mate matching)實驗，平均排名是 7.80，其中 57.65% 的排名爲第一名。這些專有名詞翻譯的想法，也被引入故宮的數位博物館跨語言資訊檢索(Chen, 2001)。

伍、跨語言資訊檢索評比

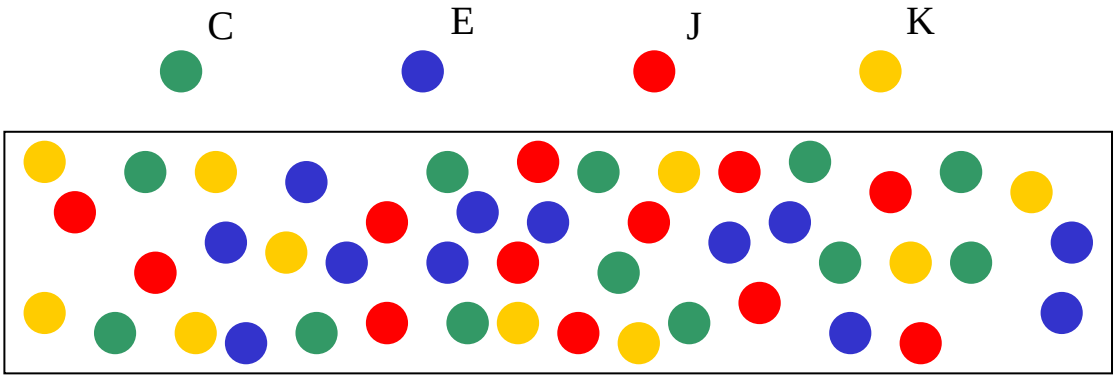
除了理論和技術外，評估也是系統發展過程重要的一環。資訊檢索有三個主要的評比單位：TREC(美國貿易部 NIST主辦)，CLEF(歐盟所支援的數位圖書館計畫)，和 NTCIR(日本文部科學省下的國立情報學研究所NII主辦)。TREC[3]過去曾舉辦跨語言資訊檢索的評比，例如 TREC6 的西班牙文-英文跨語檢索、TREC7 和TREC8 的英文/法文/德文/義大利文檢索英文/法文/德文/義大利文文件集、TREC9(1999)的英中跨語檢索等。在CLEF於 2000 年開始舉辦歐洲語言的跨語言檢索後，TREC就以英文爲主，並搭配一些戰略語言，如 2001 年的英文-阿拉伯文跨語檢索。CLEF[4] 則以歐洲語言爲主，但因應歐洲語言的多樣化，主題和文件集所涵蓋的語言數相對多起來，挑戰性也強很多。2000 年的主題包括荷蘭文、英文、法文、德文、義大利文、西班牙文、瑞典文、芬蘭文等。文件集包括英文、德文、法文、義大利文。2001 年主題增加俄文、日文、中文、和泰文，文件集增加西班牙文。



圖四　音素比對

[3] http://trec.nist.gov

[4] http://clef.iei.pi.cnr.it:2002/



圖五　NTCIR 跨語言資訊檢索工作示意圖

NTCIR[5]以亞洲語言爲主，初期是英文和日文的跨語言檢索評比， 2000-2001 年台灣大學圖資系陳光華教授和資訊系陳信希教授(Chen and Chen, 2001)，與日本文部科學省國立情報學研究所合作，共同舉辦中文資訊檢索評比和英中資訊檢索評比。2001-2002 年規模擴大至中、日、韓、英四國語言的跨語言資訊檢索評比，圖五是NTCIR跨語言資訊檢索示意圖。不同的顏色區分出不同的語言詢問和文件，目標是由五顏六色的袋子中，挑出相關的文件，並列出相關順序。

跨語言資訊檢索評比語料包括測試資料庫、檢索主題、和參考答案三部份，以下是檢索主題範例。基本的格式，延續 TREC 主題的定義，每一檢索主題包括標號<NUM>、標題<TITLE>、描述<DESC>、相關敘述<NARR>、和概念<CONC>。由於是跨語言資訊檢索，因此也提供不同語言的檢索主題，在每個主題加上語言的歸屬以示區別。<SLANG>表示主題定義者的母語，<TLANG>是主題呈現的語言。範例一是由母語是中文的主題制訂者提供，也以中文呈現，<SLANG>和<TLANG>都標記成中文(CH)。範例二是其英文翻譯，因此<TLANG>標記成英文(EN)。

範例一：

<TOPIC>
<NUM>010</NUM>
<SLANG>CH</SLANG>
<TLANG>CH</TLANG>
<TITLE>反聖嬰現象</TITLE>
<DESC>
查詢何謂反聖嬰現象及其與聖嬰現象的比較與影響
</DESC>
<NARR>聖嬰現象結束後接著而來的反聖嬰現象對全球氣候會有何影響？跟聖嬰現象的不同在何處？反聖嬰現象形成的原因、特徵、循環性等基本介紹視爲相關。個別國家因聖嬰現像造成的影響視爲不相關。</NARR>
<CONC>聖嬰現象，反聖嬰現象，氣候</CONC>
</TOPIC>

範例二：

<TOPIC>
<NUM>010</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>La Nina</TITLE>
<DESC>
To retrieve what the La Nina is and the comparison with El Nino</DESC>
<NARR>What are effects of La Nina following after El Nino on global climate? What is the comparison with El Nino? Its basic introduction, the way it is formed, its features and circulations are relevant. The influence on certain country made by El Nino will be regarded as irrelevant. </NARR>
<CONC>El Nino, La Nina, climate</CONC>
</TOPIC>

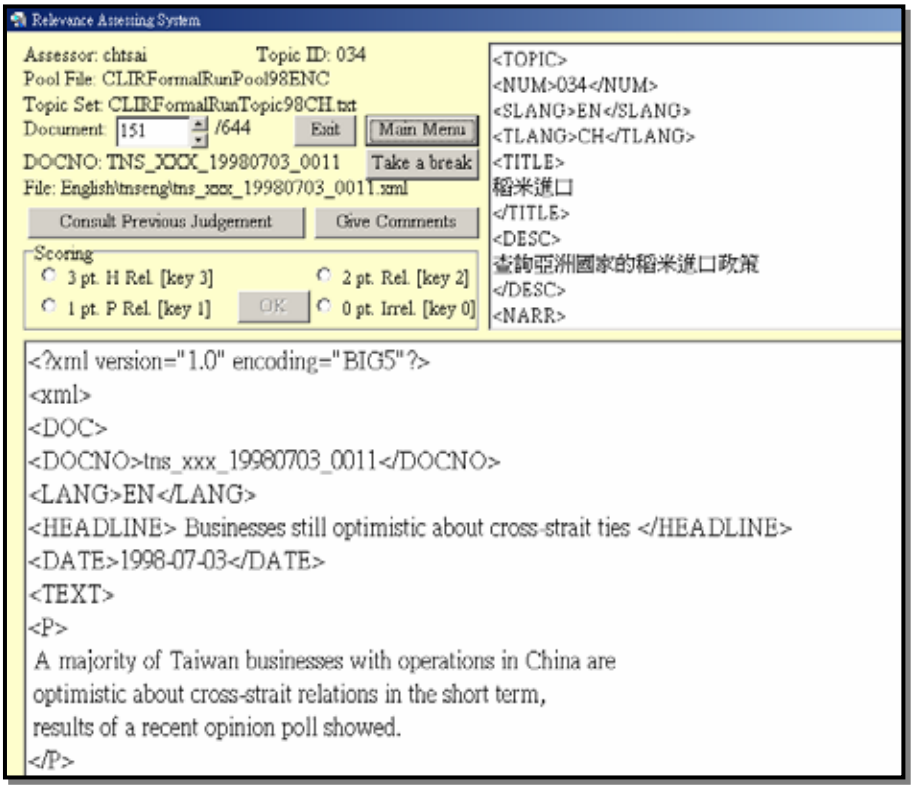
[5] http://research.nii.ac.jp/ntcir/workshop/work-en.html

表一摘錄 2001-2002 年第三屆 NTCIR 評比的文件集統計資料，中文(英文)、和日文(英文)分別收集於同一時期(1998-1999)台灣和日本報社的新聞，韓文爲 1994 年的經濟新聞。整體而言，中文文件量最多、日文文件次之，接著是韓文，英文新聞量相對較少。

表一・第三屆 NTCIR 評比文件集

語言	來源	報社	數量
中文	台灣	中國時報	38,163
		工商時報	25,812
		中時晚報	5,747
		中央日報	27,770
		中華日報	34,728
		聯合報	249,508
日文	日本	每日新聞	236,664
韓文	南韓	每日經濟	66,146
英文	台灣	台灣新聞	7,489
		中時電子報	2,715
	日本	每日新聞	12,723

參考答案一般採用聚合方法(pooling method)，由參與評比單位所提供的結果，經過合併之後，由主題訂定者做人工檢驗。爲了避免處理上有偏移的現象，會限制參與者送回答案組(runs)的個數。目前已經建立評估平台，協助評估人員(assessors)標記答案，並統計相關資料。圖六顯示此系統的畫面，視窗左上角顯示評估人員名字(Assessor)、主題編號(Topic ID)、pool 檔案名稱(Pool File)、文件序號(Document)、和文件編號(DOCNO)等。視窗右上角是主題的內容(Topic)，視窗下半部列出文件原文。assessor 對每一篇文件，給予相關程度判定(Scoring)，等級由 0 到 3，分別代表不相關、部份相關、相關、和極相關，assessor 並須加註做這項判定的線索(Give Comments)。assessor 也可以參考過去的記錄(Consult Previous Judgment)，作爲判定參考，或修正過去的判定。透過這套平台，assessor 所做的判定過程，被記錄到 log 檔案中，供進一步分析，作爲改善評估流程參考。



圖六　檢索評估平台

6. 結論與討論

本文對跨語言資訊檢索相關理論和技術，做深入的探討。在詢問翻譯上，實驗結果顯示辭典相當有用，即使是最簡單的方法，也有單語檢索大約百分之五十的效益。加上詞類、區域性回饋、區域語境分析、詞組、語料庫等，可以將平均準確度提昇到單語檢索的百分之七十五左右。語料庫對應的準確性、語料庫的選擇、辭典的覆蓋度和內容等，都是影響效能的重要因素。本文也考慮翻譯歧義性和目標多義性的雙重影響，以及專有名詞音義等問題。

從技術層面來看，可以發現詞彙語義資料庫的重要性，在查詢擴張、和詞義歧義分析等都會使用到。語義資料庫與單語資訊檢索的關係，可參考(Mandala, Takenobu, Hozumi, 1998)。在這方面的資源，中文相較於其他語言就顯得較弱。美國早在 1985 年，普林斯頓大學就提出建立英文詞彙語義資料庫的構想(Miller et al., 1990)，並在認知科學研究所 Miller 教授的帶領下進行研發，於 90 年代初免費提供各界使用，現已到 1.6 版(Soergel, 1998)，並廣泛的使用於多個應用領域，如資訊檢索系統、數位圖書館、自然語言處理系統、…。根據這個資料庫所發展出來的成果相當的多，http://www.cis.upenn.edu/~josephr/wn-biblio.html 列出部份參考資料。歐盟(European Commission)有鑑於歐洲主要語言也有類似的需求，在 1996 年提出 MLIS(Multilingual Information Society)的規畫(EC, 1996)，於 1997 年三月開始進行爲期三年的計畫(EC, 1997)，目的是嘗試建立多語知識庫(英語、荷蘭語、西班牙語、義大利語、法語、德語、捷克語、愛沙尼亞語)。我們也採用電腦輔助的方式，嘗試建立中英文對應語義詞彙庫 CEWordNet(Chen and Lin, 2000)，供跨語言資訊檢索使用(Chen, Lin and Lin, 2000; Chen, Lin and Lin, 2002)。由於此詞彙庫沒有經人工檢查，內含許多雜訊，在跨語檢索上的效益較不明顯，有待後續更進一步的研究與驗證。

參考文獻

Andresen, L. (1997) "Metadata: New Key Concept in Internet Circles," *Bibliotekspressen*, 6, March 1997, 152-153.

Ballesteros, L. and Croft, W.B. (1996) “Dictionary-based Methods for Cross-Lingual Information Retrieval,” *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*, 791-801.

Ballesteros, L. and Croft, W.B. (1997) “Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval,” *Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval*, 1-8.

Bearman, D. (1996) "Developments in Metadata Management Frameworks," *Archives and Museum Informatics*, 10(2), 1996, 185-188.

Berry, M.W. and Young, P.G. (1995) “Using Latent Semantic Indexing for Multilingual Information Retrieval,” *Computers and Humanities*, 29(6), 413-429.

Bian, Guo-Wei and Chen, Hsin-Hsi (2000) "Cross Language Information Access to Multilingual Collections on the Internet," *Journal of American Society for Information Science*, 51(3), 2000, 281-296.

Borgman, C.L. (1997) "Multi-Media, Multi-Cultural, and Multi-Lingual Digital Libraries: How Do We Exchange Data in 400 Languages," *D-Lib Magazine*, June 1997, http://www.dlib.org/dlib/june97/06borgman.html.

Brown, R.D. (1996) “Example-Based Machine Translation in the Pangloss System,” *Proceedings of the 16th International Conference on Computational Linguistics*.

Caplan, P. and Guenther, R. (1996) “Metadata for Internet Resources: the Dublin Core Metadata Elements Set and Its Mapping to USMARC,” Cataloging and Classification Quarterly, 22(3/4), 1996, 43-58.

Carbonell, J.G., Yang, Y., Frederking, R.E., Brown, R., Geng, Y., and Lee, D. (1997) “Translingual Information Retrieval: a comparative evaluation,” *Proceedings of the International Joint Conference on Artificial Intelligence*, 1997.

Chen, Hsin-Hsi (1997) “Cross-Language Information Retrieval,” *Proceedings of ROCLING Workshop on ED/MT/IR*, Academic Sinica, Taipei, 1997, 4-1~4-27.

Chen, Hsin-Hsi (1998) "Cross-Language Information Access on the Internet," Presented in *Symposium on Knowledge Discovery and Retrieval in the Network Era*, the IX Pacific Science Inter-Congress, November 18, 1998.

Chen, Hsin-Hsi (2001) "Cross-Language Information Retrieval for Digital Museums," *Global Digital Library Development in the New Millennium*, Ching-chih Chen (Editor), Tsinghua University Press, 33-40.

Chen, Kuang-Hua and Chen, Hsin-Hsi (2001) "Cross-Language Chinese Text Retrieval in NTCIR Workshop – Towards Cross-Language Multilingual Text Retrieval," *ACM SIGIR Forum*, 35(2), Fall 2001.

Chen, Kuang-Hua and Chen, Hsin-Hsi (1994) "A Part-of-Speech-Based Alignment Algorithm," *Proceedings of 15th International Conference on Computational Linguistics*, Kyoto, August 5-August 9 1994, 166-171.

Chen, Hsin-Hsi and Lee, Jen-Chang (1996) "Identification and Classification of Proper Nouns in Chinese Texts," *Proceedings of 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, August 5-9, 1996, 222-229.

Chen, Hsin-Hsi and Lin, Chi-Ching (2000) "Sense-Tagging Chinese Corpus," *Proceedings of 2nd Chinese Language Processing Workshop*, October 8, Hong Kong, 7-14.

Chen, Hsin-Hsi; Bian, Guo-Wei and Lin, Wen-Cheng (1999) "Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval," *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, University of Maryland, June 20-26 1999, 215-222.

Chen, Hsueh-Hua; Chen, Chao-Chen and Chen, Kuang-Hua (1999) "Metadata Interchange for Chinese Information," *IT and Global Digital Library Development*, Ching-chih Chen (Editor), West Newton: MicroUse Information, 65-74.

Chen, Hsin-Hsi; Lin, Chi-Ching and Lin, Wen-Cheng (2000) "Construction of a Chinese-English WordNet and Its Application to CLIR," *Proceedings of 5th International Workshop on Information Retrieval with Asian Languages*, September 30-October 2, Hong Kong, 189-196.

Chen, Hsin-Hsi; Lin, Chi-Ching and Lin, Wen-Cheng (2002) "Building a Chinese-English WordNet for Translingual Applications," *ACM Transactions on Asian Language Information Processing*.

Chen, Hsin-Hsi; Huang, Sheng-Jie; Ding, Yung-Wei and Tsai, Shih-Chung Tsai (1998) "Proper Name Translation in Cross-Language Information Retrieval," *Proceedings of 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, Montreal, Quebec, Canada, August 10-14 1998, 232-236.

Davis, M.W. (1997) "New Experiments in Cross-Language Text Retrieval at NMSU's Computing Research Lab," *Proceedings of TREC 5*.

Davis, M.W. and Dunning, T. (1996) "A TREC Evaluation of Query Translation Methods for Multi-lingual Text Retrieval," *Proceedings of TREC-4*.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R.A. (1990) "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, 41(6), 391-407.

Dumais, S.T., Letsche, T.A., Littman, M.L. and Landauer, T.K. (1997) "Automatic Cross-Language Retrieval Using Latent Semantic Indexing," *Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval*, 18-24.

European Commission (1996). *Language and Technology: from the Tower of Babel to the Global Village*, http://www2.echo.lu/mlis/en/l&t.pdf.

European Commission (1997). *Language Engineering: Progress and Prospects*, 1997.

Gachot, D.A., Lange, E. and Yang, J. (1996) "The SYSTRAN NLP Browser: An Application of Machine Translation Technologies in Multilingual Information Retrieval," *Proceedings of SIGIR96 Workshop on Cross-lingual Information Retrieval*.

Grefenstette, G. (Editor) (1996) *Proceedings of SIGIR'96 Workshop on Cross-Linguistic Multilingual Information Retrieval Workshop*, August 22, 1996, Zurich, Switzerland.

Grefenstette, G. (1998) "Multilingual Text Retrieval," *Third Biennial Conference of the Association for Machine Translation in the Americas*, Langhorne, Pennsylvania, October 28, 1998.

Hayashi, Y., Kikui, G. and Susaki, S. (1997) "TITAN: A Cross-linguistic Search Engine for the WWW," *Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval*, 58-65.

Hershman, T. (1998) "Real-Time Web Language Translators," *Byte*, June 1998, 5-10.

Hovy, E. and Ide, Nancy (Editors) (1998) *Multilingual Information Management: Current Levels and Future Abilities*, Montreal, Canada, August 16, 1998.

Huang, C.R., *et al*. (1995) "Introduction to Academia Sinica Balanced Corpus," *Proceedings of ROCLING VIII*, Taiwan, 81-99.

Hull, D.A. and Grefenstette, G. (1996) "Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval," *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, 49-57.

Knight, Kevin and Graehl, Jonathan (1998) "Machine Transliteration," *Computational Linguistics*, 24(4), 1998, 599-612.

Kowk, K.L. (1997) "Evaluation of an English-Chinese Cross-Lingual Retrieval Experiment," *Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval*, 110-114.

Lai, M. and Lin, T.Y. (1987) *The New Lin Yutang Chinese-English Dictionary*, Panorama Press Ltd, Hong Kong.

Lin, Wei-Hao and Chen, Hsin-Hsi (2000) "Similarity Measure in Backward Transliteration between Different Character Sets and Its Application to CLIR," *Proceedings of ROCLING*, Taipei, Taiwan, 2000, 97-113.

Lin, Wen-Cheng and chen, Hsin-His(2002) "*Merging Mechanisms in Mult; lingual Information Retrieval*," Procedings of 3rd Workshop of the Cross-Language Evaluation Forum, 2002, 97-102.

Longman (1978) *Longman Dictionary of Contemporary English*, Longman Group Limited.

Mandala, R.; Takenobu, T. and Hozumi, T. (1998) "The Use of WordNet in Information Retrieval," *Proceedings of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, 1998.

Miller, G.A. (1990). "WordNet: An ON-line Lexical Database," *International Journal of Lexicography*, 1990, 235-312.

Oard, D.W. (1996) *Adaptive Vector Space Text Filtering for Monolingual and Cross-language Applications*. Ph.D. Dissertation, University of Maryland, College Park.

Oard, D.W. (Editor) (1997a) *Proceedings of AAAI-97 Spring Symposium: Cross-Language Text and Speech Retrieval*, Stanford, California, March 24-26, 1997.

Oard, D.W. (1997b) "Serving Users in Many Languages: Cross-Language Information Retrieval for Digital Libraries," *D-Lib Magazine*, December 1997, http://www.dlib.org/dlib/december97/oard/12oard.html.

Oard, D.W. (1998) "A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval," *Proceedings of Third Conference of the Association for Machine Translation in the Americas*, Philadelphia, PA, October, 1998.

Powell, J. and Fox, E.A. (1998) "Multilingual Federated Searching Across Heterogeneous Collections," *D-Lib Magazine*, September 1998, http://www.dlib.org/dlib/september98/powell/09powell.html.

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

Soergel, D. (1998) "WordNet," *D-Lib Magazine*, October, 1998, http://www.dlib.org/dlib/october98/10bookreview.html/.

Su, Je-Jun; Ku, Lun-Wei; Lin, Chi-Ching; Chen, Hsin-Wei and Chen, Hsin-Hsi (1998). "A Multi-Language Identification System on WWW," *Bulletin of the College of Engineering*, National Taiwan University, 73, June 1998, 155-165 (in Chinese).

Thompson, P. and Dozier, C. (1997) “Name Searching and Information Retrieval,” *Proceedings of Second Conference on Empirical Methods in Natural Language Processing*, Providence, Rhode Island, 1997.

Vossen, P. (1997) *Proceedings of DELOS Workshop on Cross-Language Information Retrieval*, Zurich, March 5-7, 1997.

Wan, Stephen and Verspoor, Cornelia Maria (1998) “Automatic English-Chinese Name Transliteration for Development of Multilingual Resources,” *Proceedings of 17th COLING and 36th ACL*, 1998, 1352-1356.