中文詞彙與跨語詞彙抽取技術在數位佛典上的研發與應用
－階段成果研討報告－

釋法源 李家名 黃乾綱
財團法人中華佛學研究所
國立台灣大學工程科學及海洋工程學系
ktang92@mail.chibs.edu.tw, trueming@chibs.edu.tw, ckhuang@ntu.edu.tw

摘要

本計畫的目標及成果是要支援建立一個方便佛教學者建立知識架構的環境與研究平台。研究平台的構想，是建立一個友善並有效率的介面，佛教學者可以透過它對龐大的數位佛教資源庫，進行統計分析、資訊檢索及抽取、文件分類與分群、資料探勘及機器學習等各項工作，以提供研究者不同於傳統佛學的研究方法及更多樣的參考資源和結果。

中文抽辭及跨語詞彙抽取的研究，是達成上述研究平台目標的重要基礎工作。本計畫分四個主要步驟：1. 資料整理。2. 中文抽辭研究。3. 跨語辭彙抽辭研究。4. 網路服務。

本篇報告將說明，目前本計劃在抽辭及跨語研究兩個主要工作上，所使用的研究方法、執行過程、階段成果及後續工作討論等。

關鍵詞：抽辭，檢索，平行語料庫，跨語。

1. 前言

由於佛教的文獻被傳譯為多種語文，如梵文、巴利文、漢文、藏文等，佛教文獻的研究者，常需比對不同語文的經典。中文抽辭工作配合跨語字辭典的應用，是提供文章分類、詞彙分類以及多語言資料平行比對的基礎。精確的多語言版本資料比對與跨語詞彙抽取，可提供佛教學者大量跨語檢索及資訊研究統計的資訊。

雖然中文抽辭結果，是進一步進行跨語詞彙研究與建立平行語料庫的基礎，本計劃實際執行，仍採抽辭與跨與詞彙研究同時進行的方式。跨語工作所需的詞彙，先由人工判斷和擷取字辭典兩種方式取得，待抽辭工作完成，便套用自動抽辭結果的詞彙來完成跨語研究的目標。

接下來將分別討論"中文抽辭"與"跨語詞彙抽取"兩項工作階段性的進行狀況與成果。

2. 研究方法、執行過程與階段結果

2.1 中文抽辭

〔研究方法〕

統計抽辭的方法長久以來都是中文抽辭的主要方式之一[1-4]

目前抽辭工作主要參考簡立峰老師 1999 年所做大量新聞資源抽辭工作中的演算法[1]。如圖 1 所示，主要的辨別依據是如果一個字串（Lexical Pattern）的左右兩邊出現字的種類越多，該處應被視為斷開點的特性就越強，也就是越應該切斷。相反的，如果某個字的下一個字的可能性永遠都只有一種，那這兩個字就必定不可切斷。

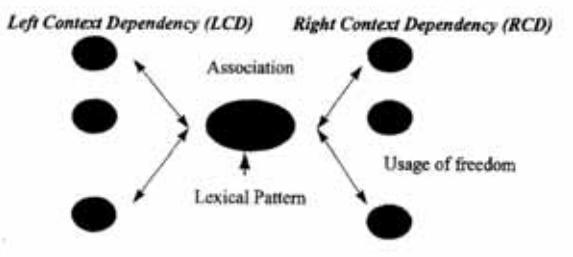

圖 1 中文抽辭方法概念圖

以上述的標準，進一步產生下列參數與演算法，來進行抽辭篩選的依據：

參數：

fx: 每個字串(例：abcde)出現的頻率
fy: 字串去掉最後一個字後(例：abcd)，出現的頻率
fz: 字串去掉第一個字後(例：bcde)，出現的頻率
|R|: 字串兩邊出現單字的種類
fxB: 字串兩邊出現最多次的單字的次數

公式：

設定篩選值：n1, n2, n3，並抽取符合以下條件者。
a. |R| > n1
b. fxB / fx < n3
c. AE > n2 *AE = fx / fz+fy-fx
其中公式 a 與公式 b 擇一個執行，再拿其結果執行公式 c。(a or b) and c

〔執行過程〕

1. 蒐集要抽辭的檔案

將所有檔案依順序合併為一個大的檔案。將合併後的檔案反向存成另一個檔案，如表 1 所示。由於是使用大量統計的方式計算結果，因此所收檔案量越多越好。

本計劃中準備三個檔案群：

◎《大正藏》全文
使用 Cbeta 全文檔，228mb 約一億單字
◎《卍續藏》全文
使用 Cbeta 全文檔，98mb 約 4 千萬字

◎佛學學報全文

華岡、中華佛學學報、中華佛學研究、佛學研究中心學報，共 57mb 約 1700 萬字

以文獻種類來分，《大正藏》與《卍續藏》屬於佛學原典，而學報則屬於當代文獻資源。因為語言的使用有時代性，因此原典與當代文獻分開抽辭預計會有不同的結果。

表 1 正反檔案內容範例

原檔案	中國佛教的內涵…太虛大師《大智度論》出現的意義不是…
反向檔案	…是不義意的現出》論度智大《師大虛太…函內的教佛國中

2. 建立正反兩個檔案的索引檔

索引方式：Suffix Array

索引值：連續中文字字串字首的 offset

排序依據：Unicode 字碼順序

表 2 索引檔（Suffix Array）建立流程表例

檔案內容	有關法顯大師的歷史資料，…
取出字串 * 紅色數字為每一字串起始位置。 * 非中文字元視為字串的結束點。	0 有關法顯大師的歷史資料 1 關法顯大師的歷史資料 2 法顯大師的歷史資料 3 顯大師的歷史資料 4 大師的歷史資料 5 師的歷史資料 6 的歷史資料 7 歷史資料 8 史資料 9 資料 10 料
排序字串	8 史資料 4 大師的歷史資料 5 師的歷史資料 10 料 0 有關法顯大師的歷史資料 7 歷史資料 2 法顯大師的歷史資料 6 的歷史資料 9 資料 1 關法顯大師的歷史資料 3 顯大師的歷史資料
取得索引	8, 4, 5, 10, 0, 7, 2, 6, 9, 1, 3

3. 套用演算法篩選字串完成抽辭步驟

取得每個字串的 fx, fy, fz, |R|, fxB 等值，套用並調校演算法取得最佳抽辭結果，步驟如表 3 所示。

表 3 抽辭步驟

使用表 1 的檔案，配合表 2 建好的索引檔，依序取出表 2 中排序好的字串，並累計出演算公式所需的值。 以右欄字串為例，”大師”這個 pattern 的各項值如下： fx = 4 fy = 回檔案中統計”大”的次數 fz = 回檔案中統計”師”的次數 |R| = 2（的、說） fxB = 2（的）	例： … 大人 大師 大師的僧俗大德 大師的歷史資料 大師說到 大德 …
使用演算法公式，依上述各字串的各項值決定是否抽取該字串。 公式： （|R| > n1 || fxB / fx < n3）& （fx / fz+fy-fx > n2）	符合篩選標準的字串

4. 與現有佛學字辭典比對結果

抽辭結果與現有佛學辭典比對，看出交聯集及分析各種特性。目前搜集佛光大辭典與丁福保佛學大辭典，以此兩部權威性與使用量都很高的資源做這個步驟中比較的對象。

〔目前成果〕

1. 當代文獻抽辭結果

當代文獻抽辭結果列舉
條件：|R| > 20 & AE > 0.001 （紅色詞彙為佛光大字典中亦出現者）： 一九三七年、一切外道、不一而足、世界佛學名著譯叢、互相矛盾、佛教義理、修多羅、俱舍論記、六祖壇經、南嶽懷讓、因陀羅網、掌握、政策、明儒學案、智者大師、梁啟超…

圖 2 顯示當代文獻字串以 |R| > 20 的條件篩選後，各字串與左右 |R| 值的相關狀況。圖中顯示，左右兩邊出現數量的狀況幾乎相同。出現兩種狀況者最多，然後依序遞減。

20 次到 200 次之間有一小高峰，是接下來進行 AE 值計算的字串群，2 到 20 次間的字串，則需要用 fxB / fx 最大數的方式來篩選。

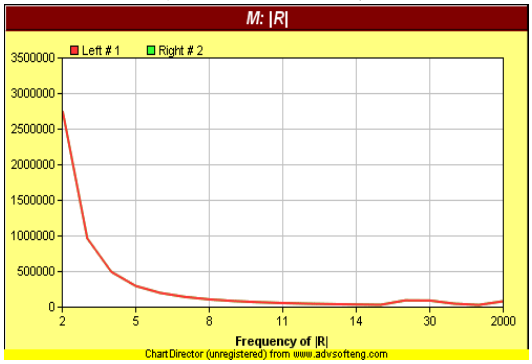

圖 2 當代文獻字串數與其左右|R|值的分布圖

用左右 |R| > 20 的結果進一步進行 AE > 0.001 的計算，抽出 57,962 個辭彙。這 57,962 個詞比對佛光大辭典 24,798 個辭結果如圖 3 所示。

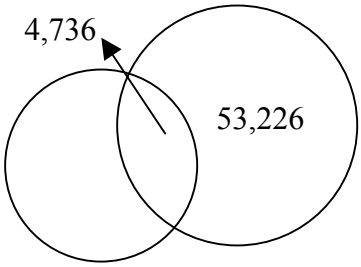

圖 3 當代文獻抽辭結果與佛光辭典比較圖

2. 卍續藏抽辭結果

卍續藏抽辭結果列舉
條件：|R| > 20 & AE > 0.003 （紅色詞彙為佛光大字典中亦出現者）： 一二三四五六七、一切有為法、一念三千、古德、我昔所造諸惡業、下化眾生、有漏無漏、無量壽經、歡喜無量、沙門婆羅門、洪武元年滄洲、涅槃…

卍續藏字串與左右 |R| 值的統計表，除字串數量上卍續藏本來就多於當代文獻的現象外，其餘趨勢幾乎一樣，見圖 2 及圖 4。

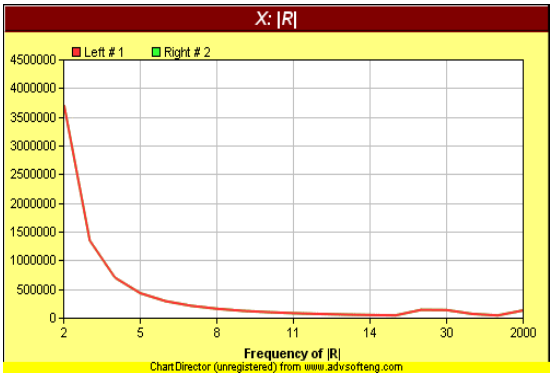

圖 4 卍續藏字串數與其左右|R|值的分布圖

用左右 |R| > 20 的結果進一步進行 AE > 0.003 的計算，抽出 54,546 個辭彙。這 54,546 個詞比對佛光大辭典 24,798 個辭結果如圖 3 所示。

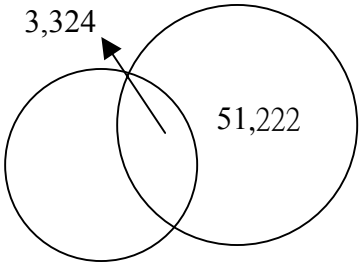

圖 5 卍續藏抽辭結果與佛光辭典比較圖

2.2 跨語詞彙抽取

〔研究方法〕

(一) 運用跨語辭典 (cross language dictionary)：

佛學辭典是佛教基本知識的來源；但是，詞彙有歧義性的問題，一個詞彙可能有多重的意義，因此會遇到選詞(lexical selection)的問題。另外，是有關辭典覆蓋度的問題，即某些詞彙可能在辭典中找不到。雖然有這些問題；不過，跨語辭典在跨語檢索仍然是很重要的參考資料。本研究擬採用以下二種佛教跨語辭典：1.中日韓佛教名詞辭典、2. MONIER WILLIAMS 梵英字典，來執行跨語檢索，並配合語料庫來提高準確度。

表 4 中日韓佛教名詞辭典

辭條	說明
菩提	[py] pútí [wg] p'u-t'i [ko] □□ pori [ja] ボダイ bodai ‖ A transliteration of the Sanskrit/Pali term bodhi, meaning wisdom, enlightenment or awakening. (1) The wisdom of the true awakening of the Buddha. Enlightenment. The function of correct wisdom. The situation of the disappearance of ignorance due to the functioning of awakened wisdom. … [Dictionary References] Naka1221d Iwa735 [Credit] cmuller(entry) cwittern(py)

本辭典又稱為 CJK(Chinese, Japanese and Korean)佛教名詞辭典，係由日本 Toyo Gakuen 大學的 Charles Muller 等學者於 1986 年開始編輯，約蒐集 30 萬個佛教名詞，內容亦包含英文說明的資料。

表 5 MW 梵英字典

辭條	說明
tulana	n. lifting Mr2icch. ix , 20 ; weighing , rating , iii , 20 ; N. of a high number Buddh. L. ; (%{A}) f. rating ib. ; equalness with (instr. or in comp.) Prasannar. ii , 16.

本辭典係由 MONIER WILLIAMS 所編輯，牛律大學出版社發行，主要依以英文來解釋梵文的字義，按照梵文字母順序排列；電子檔為 MWSDD V 1.5 Beta. 版，提供使用者查詢的介面。

本研究擬嘗試運用兩部辭典中的英文解釋，作為梵語詞條與漢語詞條的特徵向量，以文件比對的方式，找出對應的梵漢詞彙。詳如圖 6。

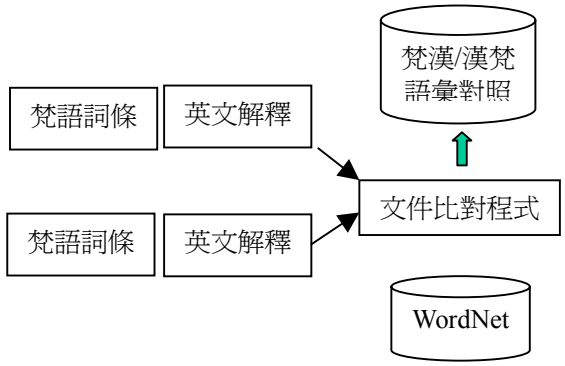

圖 6 跨語辭典的運用方式

(二)運用平行語料庫(parallel corpus)

以統計方式利用平行與料庫抽取對應詞彙,過去有相當多的研究, 但主要都是以現代語文為主.[5-9]

語料庫(corpus)根據對應的程度,通常可以分成詞彙對列(word alignment)、句子對列(sentence alignment)、文件對列(document alignment)、及不對列(no alignment)等四種。本研究的平行語料庫(parallel corpus)，則嘗試以段落對應單位，或是以句為對應單位的方式，取出對應的佛典語句；再統計模式找出最可能對應的詞彙，最後再利用 Bi-partite graph 中尋找 Max Matching pair 的方式，找出其它可能對應的詞彙，詳如圖 7。

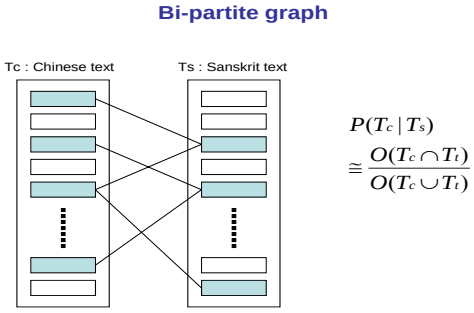

圖 7 平行語料庫的運用方式

由於本研究所採用的佛教文獻為《妙法蓮華經》，它的文本具有多種語言對譯，包括梵文、漢文、英文、藏文等，而且呈現句子或段落對列，故擬採用平行語料庫，來建立跨語檢索，或用以產生雙語辭典。如下表 6：

表 6 妙法蓮華經

語研	內容
《漢文》	如是我聞，一時佛住王舍城耆闍崛山中， 與大比丘眾萬二千人俱， 皆是阿羅漢，諸漏已盡，無復煩惱，逮得己利，盡諸有結，心得自在。
《梵文》	ekasmin samaye bhagavān rājagrhe viharati sma grdhakūṭe parvate/ mahatā bhikṣusaṃghena sārdhaṃ dvādaśabhir bhikṣuśataiḥ/ sarvair arhadbhiḥ kṣīṇāsravair niḥkleśair vaśobhūtaiḥ suvimuktacittaiḥ …
《英譯》	Thus have I heard. Once upon a time the Lord was staying at Râgagriha, on the Gridhrakûta (1)mountain, with a numerous assemblage of monks, twelve hundred monks, all of them Arhats, stainless, free from depravity, self-controlled(2), …

〔執行過程〕

有關佛教名詞的平行語料庫的詞彙抽取，其執行步驟約略可分為三個階段：第一階段是先將《妙法蓮華經》的梵文、漢文、英文、藏文等具有 TEI（Text Encoding Initiative）標記的 xml（Extensible Markup Language）數位檔案，依其科判的標記（markup tag）作為對應，以 DOM(Document Object Model）程式取出對照的平行段落（parallel paragraph）。第二階段則執行計算每個詞彙的字頻，與可能對應的機率值，來決定平行語料庫（parallel corpus）。第三個階段再利用 Bi-partite graph 中尋找 Max Matching pair 的方式，找出其它可能對應的詞彙，並進一步與跨語辭典做比對。詳如圖 8。

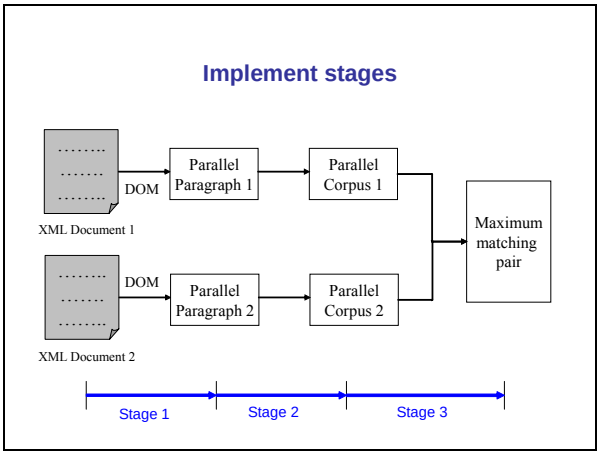

圖 8 跨語研究執行階段及步驟

〔目前成果〕

本研究目前的階段性成果，主要有「平行詞彙的對照」及「詞彙的機率統計」等項，分別說明如下：

(一)平行詞彙的對照

現階段是以《妙法蓮華經》的梵文及漢文的數位文獻檔案做測試，將二個文獻依其科判的標記（markup tag）作為段落，並以 DOM（Document Object Model）程式取出對照的平行段落（parallel paragraph）。取出對照的平行段落（parallel paragraph）後，再以程式計算每個詞彙的字頻、出現的段落，以及可能對應的機率值，用以決定平行語料庫（parallel corpus）。詳如圖 9。

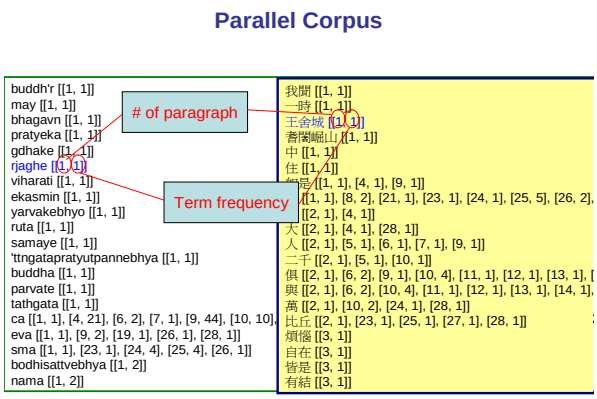

圖 9 平行詞彙的計算

(二)詞彙的機率統計

取出的平行詞彙，再以程式記錄詞彙出現的段落，詞彙若出現在該段落則記錄值為 1 ，若無出現記錄值為 0 ，而形成 signature file，詳如圖六。再將此 0 與 1 組成的 signature file 數列，套入機率公式計算，即可進一步決定跨語的詞彙對照關係，見圖 10。

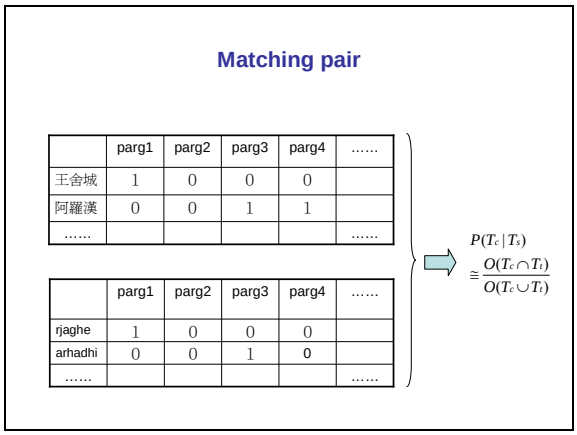

圖 10 平行詞彙機率統計

3. 階段結論與後續工作

3.1 中文抽辭

(一) fxB/fx 計算的使用：

目前尚未用到 fxB/fx（所接可能性最高次數）的計算，因此在 $1 < |R| < 21$ 之間的詞（被 $|R| > 20$ 條件刪掉的詞）並沒有做進一步的篩檢。由進行 $|R| > 20$ 篩選前後的結果發現，各自與佛光大字典比對的詞彙數量相差（$|R| > 20$ => 4,736 vs. $|R| > 1$ => 13,202）。在 $1 < |R| < 21$ 之間，有 $|R| > 20$ 中兩倍的詞在佛光大字典裡出現。

(二) |R| 與 AE 條件的調整與字辭典的選擇：

目前的結果只各用了一組的值來抽辭，未來需要測試各種條件已獲得最好結果。結果好壞的判定，可以比對字辭典中辭彙交集多少作為一依據。

另外佛光大字典外，如丁福保辭典及其他重要佛學辭典也是未來將最為比較的工具。

(三) 字辭典以外的辭彙：

尚有許多字辭典比對以外的辭彙需要整理，其中包括很多中文語言敘述特性上的用詞，這些需要請專人看過做進一步的分類與分析。

(四) 中研院斷詞系統的使用：

申請使用中研院所研發的中文斷詞系統，將抽出的詞彙作為斷詞的詞庫，並對所有資源再做一次斷詞工作，應可對中文語言上特性的用詞做出更準確的分析。[10]

(五) 演算法的改進：

除了前文所介紹的方式外，還計劃使用 Entropy 亂度計算模式來進行抽辭。

(六) 大正藏：

目前的古典文獻資源是進行卍續藏的部份，若以原點或當代文獻分類，大正藏應與卍續藏合併作業。目前卍續藏的檔案總大小為 97mb，而大正藏的總檔案大小為 228mb。未來需解決在建索引時避免檔案太大而造成電腦硬體資源不足或時間太慢的問題。

(七) 同字異碼的現象：

同字異碼或異形的現象不少，尤其在面對跨語的狀況時。例如大陸的簡體與日文的漢字，歲然是同一個字，或是相同字型不同意義等狀況，都有可能造成電腦判斷上的差別。在大量統計上或許可以暫時忽略，但在日後提供服務時，必須要將這些案例考慮進去。列舉如表 7。

表 7 同字異碼異形舉例表

狀況一：同一字本身可用兩種以上的字互代	狀況二：因為輸入法對應的字型不同造成的差異
歡 喜 踊 躍	欲 令 眾 生
歡 喜 踴 躍	欲 令 衆 生

3.2 跨語研究

未來跨語研究的工作，包括「佛典多語詞彙的抽取與檢索服務」、「佛學詞彙關聯性分析服務」、「佛學詞彙出處的比對服務」等。這些希望能整合為跨語佛典詞彙的服務平台，詳如圖 11。未來多語佛典文獻語料庫（parallel corpus）與跨語搜尋引擎（Cross-language Search Engine）的建立，將可做為佛教知識管理與多語知識本體的基礎。

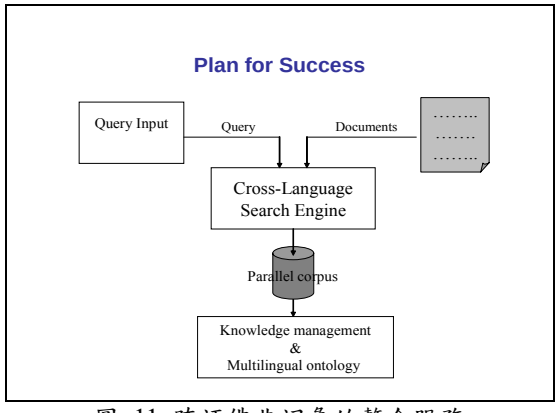

圖 11 跨語佛典詞彙的整合服務

參考文獻

[1] Chien, L.-F. "PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval" Information Processing and Management 35 (1999) 501-521。

[2] Wu, D. and Xia X., “Large-scale automatic extraction of an English-Chinese translation lexicon” Machine Translation 9:3-4 (1994) 285-313.

[3] Kwong, O.Y. and Tsou, B.K. 2001. “Automatic Corpus-Based Extraction of Chinese Legal Terms.” In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001), Tokyo, Japan.

[4] Tsou, B.K., Tsoi, W.F., Lai, T.B.Y., Hu, J. and Chan, S.W.K. 2000. LIVAC, “A Chinese Synchronous Corpus, and Some Applications.” In Proceedings of the ICCLC International Conference on Chinese Language Computing, pages 233-238, Chicago.

[5] Kwong, O.Y., Tsou, B.K., Lai, B.Y., Luk, W.P., Cheung, Y.L. and Chik, C.Y. “A Bilingual Corpus in the Legal Domain and its Applications” Workshop on Language Resources in Asia, Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (2001), 39-46

[6] Kwong, OY, Tsou, BK, Lai, TBY “Alignment and extraction of bilingual legal” Terminology, 2004 - ingentaconnect.com 10:1 (2004), 81–99.

[7] Cheung, L, Lai, T, Luk, R, Kwong, OY, Sin, KK, Tsou, BK, “Some considerations on guidelines for bilingual alignment and terminology extraction” International Conference On Computational Linguistics 18 (2002), 1-5.

[8] Fujii, A., Ishikawa, T., Lee, J.H., “Term Extraction from Korean Corpora via Japanese” CompuTerm 2004: 3rd International Workshop on Computational Terminology (COLING 2004), 71-74

[9] Yang, C., Luk, J., “Automatic generation of English/Chinese thesaurus based on a parallel corpus in laws” Journal of the American Society for Information Science and Technology 54:7 (2003), 671 – 682

[10] Chen Keh-Jiann, Wei-Yun Ma, 2002, "Unknown Word Extraction for Chinese Documents", Proceedings of Coling 2002, pp.169-175.