行政院國家科學委員會專題研究計畫成果報告

智慧型知識擷取技術與應用研究II（總計畫）

Technologies and Applications for Knowledge Extraction (II)

計畫編號：NSC 87-2213-E-002-021
執行期限：86 年 8 月 1 日至 87 年 7 月 31 日

主持人：陳信希　　國立台灣大學資訊工程學系
共同主持人：陳光華　國立台灣大學圖書資訊學系
共同主持人：簡立峰　　中央研究院資訊科學所

一、中文摘要

雙語語料庫帶有許多語言的訊息，因而有許多可能的應用，本研究提出一些方法，以建立詞彙的對列。一般詞彙可根據是否有明顯的符號，粗分為公開的詞彙及空詞兩種。公開的詞彙可能是舊詞或新詞(未知詞)，舊詞會有新的用法，新詞則不斷的產生，因此判斷其規類一直是很多自然語言處理系統研究的重點。本子計畫針對這兩種類型的詞彙提出自動分群的方法。網路資源檢索系統可以協助使用者在資訊瀚海中有效掌握與利用真正有價值的資料，目前網路上中文資訊成長迅速，發展適合檢索中文資訊的高效率檢索技術與系統是相當重要。本子計畫發展出更高效率、具備資源自動發現、過濾、擷取技術、以語言知識處理為基礎、適合檢索網路中文資源的智慧型網路資源搜尋技術與相關系統。

關鍵詞：分群法，公開的詞彙詞彙，自然語言處理，空詞，對列，雙語語料庫，語料庫，語言模型。

Abstract

Bilingual corpus carries many kinds of linguistic knowledge such that they can be used in word-sense disambiguation, extracting translation templates, finding bilingual collocations, automatic translation in noun compounds, building bilingual dictionary, and so on. In this study, an approach for word alignment in English-Chinese corpus is presented.

There are two kinds of constituents - say, overt constituents and empty constituents. Overt constituents can be old words or new words. Old words may have new usage and new words are created dynamically. Thus to identify the use of a word is a research topic in natural language processing systems. This project investigates the clustering methods to recognize these two kinds of constituents.

Keywords: Alignments, Bilingual Corpus, Corpus, Empty Constituents, Language Models, Natural Language Processing, Overt Constituents, Segmentation.

二、緣由與目的

近年來，雙語語料庫已成為一種不可或缺的重要研究資源，這是因為雙語語料庫比單語語料庫帶有更多的語言訊息，因而可以用於多方面的應用。但在做各種應用之前，雙語語料必須先作對列。根據對應的單位來分，可以分為段落、句子、和詞彙三種對列。由於目前較少研究在探討詞彙的對列問題，且大部份都是針對同語系作實驗，少有跨越不同語系進行詞彙對列的研究。然而這項研究卻是其他後續研究的重要基礎，有了詞彙的對應關係，可以做進一步較複雜的應用所以，本研究便是在探討如何在英中雙語語料庫上做詞彙的對列。

一般詞彙可根據是否有明顯的符號，粗分為公開的詞彙及空詞兩種。公開的詞彙可能是舊詞或新詞(未知詞)，舊詞會有新

的用法，新詞則不斷的產生，因此判斷其規類一直是很多自然語言處理系統研究的重點。相對於公開的詞彙，空詞也是自然語言非常普遍的語言現象，中英文都有，但因其沒有明顯的記號，而且散佈在文本中，影響文句的處理非常多，因此如能判斷其類別，就可掌握其存在位置，以幫助後續的運作。本計畫針對這兩種類型的詞彙提出自動分群的方法。結果有助於檢索系統中關鍵詞更正確的統計。

目前網路上中文資訊成長迅速，但適合檢索中文資訊的高效率資源檢索系統則相當缺乏，由於資源檢索系統可以協助使用者在資訊瀚海中有效掌握與利用真正有價值的資料，發展適合檢索中文資訊的高效率檢索系統在目前看來是相當迫切與刻不容緩。以現有網路上的資源搜尋系統發展來看，整體系統由網路資源收集開始，包括資源自動發現，過濾收集，關鍵資訊擷取，索引建立，資訊檢索，宜人的人機介面等技術項目。這些技術功能上都必須達到快速、省空間、建立容易以及精確等系統要求，而且都必須愈來愈智慧化。子計畫三第一年發展出一些中文文件特徵抽取技術，第二年則集中在網路中文資源自動發現技術的研究，目前正進行第三年進一步發展網路中文資源自動過濾與抽取技術，使得網路中文資訊的檢索效率與資源的利用進一步提昇。

三、結果與討論

3.1 語料庫之設計與製作

在真正去找出詞彙之間的對應關係前，必須先對實驗的平行文件作些前處理，以便利爾後工作的進行。前處理包括了下列幾種工作：切分的工作，斷詞的工作，字根還原的工作，以及詞類標示的工作。由於文章的翻譯手法對於詞彙的對列工作影響很大，所以我們選擇二種不同風格的雙語語料作為實驗的材料。ROCLING 語料庫中的文章較偏逐字翻譯，其中的雙語語料度部份是電腦手冊。NTU 雙語語料則是由光華雜誌選出，翻譯較偏向意譯。

由於某個英文詞彙有可能對應到中文的任何一個詞彙，因此，為了避免盲目的配對，必須先估計到底是哪些配對的機率比較高，我們使用ϕ^2 統計值估算詞彙的相關程度。們進一步使用二個過濾程序，以消除不必要的ϕ^2 計算。第一是使用詞類訓練表過濾不可能互為翻譯的詞類配對；第二是利用頻率門檻過濾頻率過小的配對。通過上述過濾程序的配對，才真正進行ϕ^2 值的計算。

另外必須要考慮的問題是連語的現象。由於翻譯時經常有幾個詞彙被翻成一個詞彙，也就是把好幾個詞彙當成一個語意單位進行翻譯，這些詞彙構成了所謂的連語。我們使用 N-Gram 的方法分別對中文及英文抽取連語，將這些連語視為一個詞彙，即可採用前述相同的程序進行雙語詞彙對列的工作。

3.2 語言知識擷取技術之研究

以下是空詞辨識與分類系統的簡述：

步驟一：收集所有只含一個省略成分的 null fragment ，並褪去所有的 function tags 及刮號，如此所得的集合為 T-1。

步驟二：定義關鍵字。

步驟三：在T-1中，若某字為關鍵字，保留該字，否則保留詞類。這樣便形成了T-2。將T-2編碼，並去除雜訊，這就是T-3。

步驟四：將T-3分類。

步驟五：計算各類別中兩兩的LCS，其結果放在LCS-1中。

步驟六：為LCS-1再算一次LCS，得到LCS-2。

步驟七：濾掉LCS-2中出現頻率較低的以形成LCS-3。

步驟八：以LCS-3為規則，回授給訓練資料，去除低正確率者，正式的規則於是找到。

步驟九：為正式的規則找出使用限制便得到LCS-4。

步驟十：將LCS-4再回授一遍，同樣的，去除低正確率者。最後，便可得到LCS-5。

步驟十一：以LCS-5為規則來找出省略的成分。

3.3 中文資源自動發現技術

中文資源自動發現技術的研究內容包括進一步改進中文關鍵詞抽取的發展與應用， 網路資源自動發現技術的探討，持續高效率檢索系統開發。計畫執行迄今上述目標已經初步達成，除了發展出一些重要技術，也完成學術論文 14 篇 (其中 2 篇為期刊論文)，主要的成果如下所述：

在中文關鍵詞抽取的發展與應用方面，我們持續改進 PAT-TREE 技術，包括增加 Significance Analysis 技術，以過濾非重要詞彙；發展 CPAT-TREE 以提高存取效率；將關鍵詞檢索與語音辨認結合提高辨認正確率；以及應用在中文自動校稿，以及 OCR 辨認錯誤偵測。

在網路資源自動發現技術方面，包括利用 PAT-TREE 特性發展文件自動分類技術，可以擷取網路資源加以分類，這包括 News 以及 BBS 等，並且自動收錄分類關鍵詞彙。利用這項技術對抽取 Phrase Template，Collocations，雙語資訊等都有幫助。

在高效率檢索系統開發方面，包括進一步將尋易 (Csmart)資訊檢索系統發展成具備語音互動的輸入介面，可結合目標導向資料庫，可自動收集網路BBS, Web資源並施行中英文檢索的系統。

四、計畫成果自評

本研究提出一個可適用於不同語系的文件詞彙對列方法，並應不同的考慮因素，進一步區分三種語言模型。由於實驗文件各有其特性，而表現出不同的實驗結果。至於未來的研究方向，可以建立雙語辭典，或是將辭典中缺少的對譯詞彙列入。然而為了解決多詞對應的現象，應發展更好的演算法，以期更正確地抽取連語，並併入語言模型。以完成的詞彙對列雙語語料庫，可以作為其他研究的實驗材料。

另外，我們提出了一套為不定長度的句子作分類的方法，並提出如何決定類別個數與初始值的策略，其中衡量兩句的相似相異程度的公式也是一個新的構想。因為採用 LCS 來找出規則，原本令人頭痛的長距離相關的現象也獲得了相當程度的解決。我們的方法僅僅採用了語法的訊息，在樹狀語料庫日漸普及，而詞類標記的技術又很成熟的情況下，我們的作法是相當可行而且有高度的移植能力。此外，我們用來歸納出規則的方式也可推廣到其他的應用上。

整體而言，研究內容與原計畫所列的工作項目完全相符、並已經達成預期的目標、所提出的語言模型是多項應用的基礎、適合在學術期刊或會議上發表。

五、參考文獻

[1] Bai, Bo-Ren, Chun-Liang Chen, Lee-Feng Chien, Lin-shan Lee, Intelligent Retrieval of Dynamic Networked Information from Mobil terminals Using Spoken Natural Language Queries, *IEEE Transactions on Consumer Electronics*, N0. 1, Vol. 44, pp. 62-72, 1998.

[2] Bies, A. et al. (1995) Bracketing Guidelines for Treebank II style Penn Treebank Project, 1995.

[3] Brown, P. *et al.* (1991). Aligning Sentences in Parallel Corpora. In *Proceedings of 29th Annual Meeting of the ACL*, (1991): 169-176.

[4] Brown, P. *et al.* (1991). Word-Sense Disambiguation Using Statistical Methods. In *Proceedings of the 29th Annual Meeting of ACL*, (1991): 264-270.

[5] Brown, P.F., Pietra, V.J. *et al.* (1992), "Class-Based N-Gram Models of Natural Language". *Computational Linguistics*, 18(4), 467-479.

[6] Charniak, E. (1993) *Statistical Language Learning,* MIT press, 1993.

[7] Chun-Liang Chen ,Bo-Ren Bai, Lee-Feng Chien, Lin-Shan Lee,PAT-tree-based Language Modeling with Initial Application of Chinese Speech Recognition Output Verification, *International Symposium on Chinese Spoken Language Processing*, Singapore, 1998.

[8] Chun-Liang Chen and Lee-Feng Chien, OCR Text Verification and Retrieval Using a PAT-tree-based Ngram Indexing Technique, The First Asia Digital Library Workshop, Hong Kong, 168-175, Aug. 1998.

[9] Chun-Liang Chen, Bo-Ren Bai, Lee-Feng Chien and Lin-Shan Lee, CPAT-Tree-Based Language Models with an Application for Text Verification in Chinese, ROCLING'98.

[10] Chen, H.H. (1988) "A Logical Approach to Movement Transformation in Mandarin Chinese", *International Journal of Pattern Recognition and Artificial Intelligence* 2(1), 1988. pp. 71-86.

[11] Chen, H.H. (1992) "The Transfer of Anaphors in Translation", *Literary and Linguistic Computing,* 7(4), Oxford University press, 1992. pp. 231-238.

[12] Chen, H.H. and Yang, H.M. (1994). “Generating Logical Terms of Chinese Sentences”, *Journal of Information Science and Information Engineering,* 10(3), pp. 245-316.

[13] Chen, K.H. and Chen, H.-H. (1993), "A Probabilistic Chunker". In: *Proceedings of ROCLING VI*, 99-117.

[14] Chen, K.-H. and Chen, H.-H. (1994), "Extracting Noun Phrases from Large Scale Texts: A Hybrid Approach and Its Automatic Evaluation". In: *Proceedings of the 32nd Annual Meeting of ACL*, 234-241.

[15] Chen, K.H. and H.H. Chen (1994). A Part-of-Speech-Based Alignment Algorithm. In *Proceedings of the 15th COLING*, 166-171.

[16] Chen, K.H. and H.H. Chen (1995). A Corpus-Based Approach to Text Partition. In *Proceedings of the Workshop of Recent Advances in Natural Language Processing*, (1995): 152-160.

[17] Gale, W. and Church K. (1991). Identifying Word Correspondences in Parallel Texts. In *Proceedings of The Fourth DARPA Workshop on Speech and Natural Language*, (1991): 152-157.

[18] Lee-Feng Chien, PAT-Tree-Based Adaptive Keyphrase Extraction for Intelligent Chinese Information Retrieval, *Information Processing and Management ,* Elsevier Press, 1998.

[19] Lee-Feng Chien, Min-Chan Chen, Chun-Liang Chen, Bo-Ren Bai, Internet-based Chinese Text Corpus Classification and Domain-specific Keyterm Extraction, *First workshop on Computational Terminology,* Aug., 1998.

[20] Lee-Feng Chien, Min-Jer Lee, Hsiao-Tiech Pu and Lin-Shan Lee, Recent Results on Internet-based Chinese Text Corpus Classification Using Automatically Extracted Keywords, *First Oriental COCOSDA Workshop*, Japan, 1998.

[21] Chung, T.C.(1993). *Treatments of Ellipses in An English-Chinese Machine Translation System.* Master thesis, Department of Computer Science and Information Engineering, National Taiwan University, 1993.

[22] Crouch, R.(1995). “Ellipsis and Quantification: A Substitutional Approach”, *Proceeding of EACL-95.*pp. 229-236.

[23] Gazdar, G. and Mellish, C. (1989). *Natural Language Processing in PROLOG,* Addison-Wesley, 1989.

[24] Hartigan, J.A. (1975) *Clustering Algorithms,* Wiley, 1975.

[25] Hahn, U. et al. (1996). “A Conceptual Reasoning Approach to Textual Ellipsis”, *Proceeding of 12-th European Conference on Artificial Intelligence,* 1996, pp. 572-576.

[26] Hatzivassiloglou, V. and McKeown, K. (1993), "Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning". In: *Proceedings of the 31st Annual Meeting of ACL*, 172-182.

[27] Hindle, D. (1990), "Noun Classification from Predicate-Argument Structures". *Proceedings of 28th Annual Meeting of ACL*, 268-275.

[28] Jain, A.K. and Dubes, R.C. (1988) *Algorithms for Clustering Data,* Prentice Hall, 1988.

[29] Kehler, A. (1994) “Common Topics and Coherent Situations: Interpreting Ellipsis in the Context of Discourse Inference”, *Proceeding of ACL-94.* pp. 50-57.

[30] Chung-Hong Lee and Lee-Feng Chien, A Novel Integration of OODBMS and Information Retrieval Techniques for a Document Repository, *9th International Conference and Workshop on Database and Expert Systems Applications*, Feb. 1998.

[31] Lin, C.Y. (1995). Knowledge-Based Automatic Topic Identification. In *Proceedings of the 33rd Annual Meeting of ACL*, (1995): 308-310.

[32] Meyers, A., R. Yangarber, and R. Grishman (1996). Alignment fo Shared Forests for Bilingual Corpora. In *Proceedings of the 16th COLING*, (1996): 460-465.

[33] Min-Jer Lee and Lee-Feng Chien, Automatic Acquisition of Phrasal Knowledge for English-Chinese Bilingual Information Retrieval, *ACM SIGIR-98*, Melbourne, Australia, Aug, 1998.

[34] Marcus, M.P. et al.(1993) “Building a Large Annotated Corpus of English : The Penn Treebank”, *Computational Linguistics,* Volume 19, 1993, pp. 313-330.

[35] Salton, G. (1986). On the Use of Term Associations in Automatic Information Retrieval. In *Proceedings of the 11th COLING*, (1986): 380-386.

[36] Sells, P. (1985) *Lectures on Contemporary Syntactic Theories,* CSLI Publications, Stanford, 1985.

[37] Wu, D. (1995). An Algorithm for Simultaneous Brackets Parallel Texts by Aligning Words. In *Proceeding of the 33rd Annual Meeting of the ACL*, (1995): 244-251.

[38] Yamashina, M. and S. Obashi (1988). Collocational Analysis in Japanese Text Input. In *Proceedings of the 12th COLING*, (1988): 770-772.

[39] Zhai, C. and D. Evans (1996). Noun Phrase Analysis in Large Unrestricted Text for Information Retrieval. In *Proceedings 34th Annual Meeting of ACL*, (1996): 17-24.

[40] Zimmermann, H.-J.(1991) *Fuzzy Set Theory,* Kluwer Academic Publishers, 1991.