智慧型知識擷取技術與應用研究-總計畫(III)

計畫編號：NSC 88-2213-E-002-033-
執行期限：87 年 8 月 1 日至 88 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

子計畫一建構適用於評估資訊檢索的標竿測試集，同時亦可用於訓練語言模型、建構計算機制的語料庫。子計畫二所擷取的述語參數結構，對於語言分析及產生有很大的幫助。在剖析句子時，它能減少歧異樹的個數；在以轉換為本的機器翻譯系統中，我們首先要知道不同語言間的述語參數對應規則；在概念為基礎的資訊檢索系統，述語參數結構提供文本概念基本的骨架。子計畫三發展中文資源自動過濾與抽取技術，目的是希望在研發出的中文 Robot 軟體，增加針對特定主題自動過濾與抽取功能，使使用者能隨時自動獲取有興趣的資源。研究內容包括中英文文件過濾技術差異分析，高效率中文文件過濾技術，特定主題中文文件特徵抽取技術，個人化資訊服務技術等。

關鍵詞：中文語言處理，計算語言學，語料庫，資訊檢索，知識擷取，自然語言處理。

Abstract

Project 1 constructs a Chinese benchmark for performance evaluation. Besides, it can be also used to train language models. Project 2 extracts predicate argument structures. That can reduce ambiguities in parsing, propose mapping rules for transfer-based machine translation systems, and provide concept skeletons for concept-based information retrieval systems. Project 3 develops Chinese information filtering and discovering technologies.

Keywords: Chinese Language Processing, Computational Linguistics, Corpus, Information Retrieval, Knowledge Extraction, Natural Language Processing.

二、緣由與目的

知識是一切智慧型系統的根本, 知識的來源非常廣，並以不同的方式存在，其中語言文字是人類最自然、最常用的媒体，在這種型態的媒体裡，知識就隱藏(implicitly embedded)在文本中。以現階段網際網路(Internet)被廣泛使用的情況下, 知識的供應源不是太少，反而是太多。因此如何有效的擷取知識，是項刻不容緩的研究課題。本整合型計畫乃在這項規畫下，就語料庫之設計與製作、語言知識和網路中文資源自動擷取等三個子計畫通盤考慮這個問題。

第一個子計畫的主要目的是建構適用於評估資訊檢索的標竿測試集，同時亦可用於訓練語言模型、建構計算機制的語料庫。第二個子計畫則是探究語料庫中述語參數結構擷取技術，並考慮其應用。第二個子計畫在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。三個子計畫間的關係非常密切，語料庫除了提供語言知識擷取資料源外，也提供網路資源自動擷取系統評估的語料；語言知識擷取系統不僅從語料庫學習知識，提供網路資源自動擷取系統，而且可回溯給語料庫設計部份，擴大其規模；由網路中所擷取分類好的資料，又可當語料庫的基本材料。

三、結果與討論

3.1 語料庫之設計與製作

本計畫主要的目標是建立一個可實際應用的資訊檢索系統標竿測試集，首先要確立測試集的主題。主題確立後則必須進

行初步的使用者需求分析，使用者檢索的方式會影響標記主題的格式，以及描述主題的用語。接下來重要的工作即是蒐集大量的文件資料。此外在蒐集文件的同時，必須使用先前制訂的標記集進行文件的整理、組織、與標示的工作。對於製作與各主題相關文件的相關判斷是本計畫比較困難的部份，這牽涉了主觀判斷的問題。

本研究已實際建構完成一包含文件集、查詢問題以及相關判斷的完整測試集，也初步驗證了此建構程序是可行的。與現行其他測試集相較，本測試集的規模已在中等以上，在文件集與查詢主題方面，均盡量使其能接近真實之檢索環境，提高其測試的效度，而相關判斷的部分，亦結合多位判斷者進行，減低了判斷結果可能出現偏差機率。在各界急於研發中文資訊檢索系統的今日，預期此測試集之建置與出現，應能稍微解除目前國內中文完全無從取得測試資料的現狀，使中文資訊檢索系統的發展能有更高的可行性，也期望它能成為後續相關研究的基礎。

3.2 語言知識擷取技術研究

研究中的困難處在於首先必須確定動詞後的語法成分的左右邊界，然後決定那些成分併接在 VP，並且對於併接在 VP 的成分，區分是參數還是修飾語。第二，必須決定句子的省略成分以及位置，以得到正確的參數結構。最後，必須考慮動詞後每個成分所扮演的語意的角色。

在達成的目標上，(1)提出一個述語參數結構自動擷取系統。包含一個名詞片語擷取器，以減少句子的變異性；用一個有限狀態機制來得到動詞後面的語法成分；再利用樹狀語料庫評估不同策略的績效。(2)提出不同於最長優先策略的的參數結構選擇策略。在我們的實驗裡，利用『切點位置決定法則』，以提出所有可能的參數結構。數據顯示，利用從樹狀語料庫中訓練出來的 PAS 字典的最高機率優先策略，績效比最長優先策略好。(3)利用計算 Lexical Association，來決定 PP 的併接位置藉以減少可能的參數結構個數，並利用樹狀語料庫所提供的資訊，區分修飾語與參數。(4) 我們也將 transformation-based error-driven learning 的技術，應用在我們的加強模型中；實驗結果顯示，應用這種 learning 的技巧，可以彌補系統的績效。

尚待努力的課題有：(1)由於動詞後面的語法成分所扮演的語意角色，關係其為修飾語或參數，所以，決定每個語法成分所扮演的語意角色，亦非常重要。(2)可以從動詞的語意，來決定其參數個數及態。

3.3 網路中文資源自動擷取技術的研究

研究內容包括改進中英關鍵詞自動抽取技術，持續加強中英雙語檢索技術，以及發展中英文資訊過濾技術。在中文關鍵詞抽取的發展與應用方面，由於網路上的資訊是日新月異、變動頻繁的，為解決這個問題，先前我們採用 PAT-tree 為基礎的統計自動學習方法，自動偵測新詞發生，判斷其重要性，以即時獲得較具代表性的關鍵詞，本年度進一步研發改進方法，使得包括低頻人名術語，中英術語翻譯都有機會為線上資料中自動擷取出。在中英雙語檢索技術技術方面，我們試著從網路上利用分類與關鍵詞抽取技術發展出相近主題雙語語料自動收集技術，藉此有助於雙語檢索時雙語辭典的建構與收集。此外本年度也持續研究中英文分類技術，著重更種分類方法比較分析，發展適合動態環境分類技術。在中英文資訊過濾技術方面，目前開始探討利用借書記錄，Bookmark, Hyper-links 等不同使用者資源，研究個人化資訊檢索的可行性。

4. 自評

本計畫的研究內容與原計畫完全相符，並已達成預期目標。研究成果具有學術和應用價值，論文成果請參考各子計畫。