行政院國家科學委員會專題研究計畫報告
總計劃：多語言資訊檢索與擷取之研究

計畫編號：NSC89-2213-E-002-059-
NSC89-2218-E-002-036-
NSC90-2213-E-002-046

執行期限：88 年 8 月 1 日至 91 年 7 月 31 日

主 持 人：陳信希　國立臺灣大學資訊工程學系 教授
共 同 主 持 人：簡立峰　中央研究院資訊科學所　副研究員
共 同 主 持 人：柯淑津　東吳大學資訊科學系　副教授
共 同 主 持 人：陳光華　國立臺灣大學圖書資訊學系 副教授

一、中文摘要

在網際網路的蓬勃發展下，資訊的傳播沒有國度、時間的限制，Internet 持續地累積多樣化的資訊，已經形成一個巨大、分散的多媒體。本整合型計畫在網際網路的環境中，建構訊息性的多語言資訊服務系統。

子計畫一第一年的重點為進行查詢句處理的研究，包括查詢句翻譯與擴展兩個部分。第二年的子計畫利用同義詞典資料進行檢索文件分類的研究，依處理文件的範疇，自動調適同義詞典所含的詞彙，以期得到一個具自調性的文件分類系統。第三年進行檢索文件翻譯的研究，利用機讀字典所蘊藏的詞彙資訊，以及由雙語語料庫中所得到的翻譯經驗，進行網路環境的檢索文件翻譯。

子計畫二主要是在詞彙頻率統計的基礎下，利用大量文獻資料，透過訓練的方式，達成控制詞彙索引自動指派之目的，控制詞彙可能由標題表或是索引典提供系統使用，基本上，標題表提供的是一般性詞彙；索引典則是特定領域的控制詞彙集合。該子計畫採用索引典提供的控制詞彙，分三年進行控制詞彙索引系統、控制詞彙分類系統、控制詞彙整合系統之研究。

子計畫三建立一自動機制，包括語料自動儲存分類，術語抽取，相似術語抽取，以及術語翻譯抽取等技術模組。這個動態辭典在設計時並考慮線上資訊服務系統的特性，包括累增，大量，即時等，所發展的技術都必須具備即時處理能力。

子計畫四研究文件自動摘要的生成，自動摘要就是想由文件中自動摘錄重要的資訊，提供使用者參考。本計畫分三年循序漸進的探討文件摘要自動化的技術，第一年專注於單一中文文件摘要技術的研究上，計畫第二年跳脫單一文件的限制，延伸到多文件摘要的問題，計畫第三年則擴大到多語言文件摘要的問題。

二、英文摘要

With the rapid development of Internet, the dissimulation of information is not bounded by space and is not limited by time. Nowadays, Internet becomes a huge channel of information and accumulates a great deal of information. This integrated project investigated an informative and multilingual system for information services.

Project 1 focused on the research of query processing in the first year, including query translation and query expansion. It carried out the document classification using thesaurus in the second year. In the third year, project 1 studied the document translation using machine-readable dictionary and examples extracted from bilingual corpora.

Project 2 studied on automatically controlled-vocabulary indexing. The first year devoted to the controlled-vocabulary indexing for information retrieval, the second year focused on controlled-

vocabulary classification for information retrieval, and the last year constructed a controlled-vocabulary integrated system for information retrieval.

Project 3 concerned the automatic construction of live dictionary for information retrieval and extraction. The first year dealt with automatic extraction of Chinese-English key terms. We extended the research results in the first year to automatically extract term translation. We focused on the automatic extraction of similar term in the third year.

Project 4 studied the technologies of automatic summarization systematically. The first year focused on summarization of single Chinese documents, the second year extended it to multiple documents, and the last year dealt with the general case: multilingual document summarization.

三、研究方法

計畫所執行的工作，依主題分為：檢索文件資料特質分析、中英詞彙語意資源建造，以及中英詞彙語意資源加值等三部分說明。在檢索文件資料特質分析，為了找出與標題內容具相關語意的詞彙，我們藉由詞彙網絡的豐富語意網絡，利用關聯性指標找出一個詞彙的同義詞與上位詞。在詞彙網絡內，有些一詞多義的詞彙會擁有多個詞義。因此在擴充詞彙之前，我們必須知道詞彙的正確詞義，也就是說我們需要先解決詞義歧異的問題。本計畫以詞彙網絡的定義與文件詞彙的重疊性來進行詞義歧異辨識。

在中英詞彙語意資源建造，本計畫透過屬類詞與上位詞的重覆用語，連結詞彙網絡的同義詞集與朗文英漢字典的詞目定義。並且，透過這種自動連結將詞彙網絡的同義詞集加上適當的中文翻譯詞，更重要的是讓詞彙網絡的豐富語意能雙語方式呈現。

在中英詞彙語意資源加值方面，自中文複合詞找出最能代表其意義的中心詞彙，及若干個特徵詞彙。其次，將這些詞彙進一步以語意概念形式表達出來。在中文詞彙轉為詞義概念的部分，辨識語意歧義的方法，我們除了用到詞彙的詞性之外，還透過詞彙網絡的上位關係來降低歧義度。

在控制詞彙的研究上，第一年已完成控制詞彙的自動指派機制；第二年完成的控制詞彙的自動分類機制。第三年建構一個控制詞彙輔助系統，探討其對於資訊檢索績效的影響。我們將以擴展使用者查詢詞彙的方式，如以控制詞彙擴展；以同義詞典擴展，以控制詞彙加權，探討各種不同關係的控制詞彙，其影響檢索績效的程度。為了分析控制詞彙對於檢索效益之影響，因此本研究設計以下三種不同的實驗：基礎檢索(Baseline)、以控制詞彙擴展、以同義詞典擴展，以控制詞彙修正。

在訊息抽取研究上，因為領域知識有關, 所以網路資源必需經過自動分類，才能有效提供不同領域的語料。以 PAT-tree 結構為基礎發展語料分類技術，將每一個以分好類的語料建成一 PAT-Tree， 而 PAT-tree 本身其實是一 N-gram Feature Vector。利用 PAT-tree 進行分類特徵比對具有容易更新, 不限特徵等特性。而統計斷詞的方法，先用統計算出字與字之間的關係，從結合度較低的字中間斷開，所得的詞再經詞典比對，過濾常用詞，最後可獲得較具代表性的關鍵詞。此外，進一步討論低頻術語抽取問題，想法是以上下文的語意(Context Semantic)預測可能的術語種類，如人名、地名等，在一類別用不同方法個別抽取。

在自動摘要的研究，單篇文件摘要著重於如何擷取合適的句子，構成一篇摘要。選取的線索包括句子或段落的位置、文本名稱、重要詞組、詞頻、詞連慣性、詞共現性、人名、地名、公司名、組織名等。在多文件摘要，包括文件分類、文件語義分析、句子相相似度分析、與摘要的呈現(重點式摘要和瀏覽式摘要)。在多語言摘要的研究，事件分群的架構需考慮語言的差異，翻譯與分群的先後，包括幾種模式：直接多語事件分群、單語事件分群再合併、翻譯的動作延遲到句子層次。不管哪種模式，翻譯是不可缺少的機制。