第一章　緒論

1.1　研究動機與背景

任何語言處理系統都必須先能分辨本文中的詞才能進行進一步的處理，例如機器翻譯、語言分析、語音辨識、資訊抽取、還有近來熱門的網路搜尋引擎。為了能讓這些語言處理能有更好的效能，因此中文斷詞成了語言處理不可或缺的技術。大多數歐美語系國家所使用的語言，由於其字與字之間都有明顯的空白做為分隔，所以很容易的就能判斷出單字。但中文不同於其他歐美的語言，中文是一長串文字的組合，字詞與字詞間並沒有一明顯符號來做區隔，也因此中文斷詞將是困難且值得研究的課題。

過去有許多的斷詞技術不斷的被發表出來[1][2][3]4]，而最常見的中文斷詞技術主要可以分為三大類—詞庫式斷詞（word identificstion）、統計式斷詞法（statistical word identification）與綜合前面兩種斷詞方法的混合式斷詞法（hybrid word identification）。詞庫式斷詞（word identificstion）即是利用已事先建立的詞庫作為斷詞的基礎，透過比對出現在中文資料中並與詞庫裡的字詞相符的中文字詞來達成斷詞的目的，是目前最普遍被採用的斷詞方法。但詞庫式斷詞法的缺點在於遇到新詞就會大幅降低斷詞正確性，且要建立一個完整的詞庫並不容易，所以當文件中出現新詞時就會產生斷詞上的錯誤，因此斷詞的效果與詞庫的大小和多樣性有相當的關係，也因此必須時常對詞庫的內容加以維護，除了增加新詞的收藏外，由於詞庫過大將影響系統的效能，所以也必須淘汰詞庫中不常出現的字詞避免詞庫的無限擴大。而解決這些缺點就是本篇論文最主要的目的。

詞庫式斷詞（word identificstion）技術所遇到的最大困難就在於新詞的擷取。新詞代表的就是辭典中所沒有的辭彙，也就是所謂的未知詞。傳統上詞庫是參考現有中文辭典所建立的，但沒有任何辭典能包含所有中文字詞，尤其在現今資訊發達的世界裡，每天都有大量的中文新詞被創造出來，新詞在各個領域中不斷的出現，並且文章中新詞所出現的比例愈來愈高，甚至新詞往往是文章中最關鍵最重要的辭彙，當斷詞系統[6]所擁有的詞彙不足時，斷詞的結果很容易是錯誤的，並嚴重影響中文處理的正確性。因此如果仍舊依賴舊有的詞庫，斷詞系統必須對於未知詞做進一步的處理[7][8][9]，其效能將會受到很大的限制。但詞庫的擴充並不是一項簡單的任務，必須耗費大量的人力和金錢，搜集資料並從中擷取新詞出來，這是很沒有效率的做法但卻是過去唯一的選擇。

在數位化的現代社會裡，網路上流通著大量的資訊，資料的蒐集變的非常簡單，透過分析過濾這些資料，新詞的擷取將比已往更加容易，已經有許多新詞擷取技術利用網路資源來實現[10][11][12]。拜眾多新聞媒體競爭激烈所賜，所以每天都有無數的新聞事件呈現在眼前，而在這些新聞事件裡隱藏著許多新詞。如能使用一套有效的方法，新詞的自動擷取將變成可能。本篇論文將利用統計式斷詞法（statistical word identification）的概念來擷取隱藏在網路新聞資料裡的新詞。統計式斷詞法（statistical word identification）乃參考一大型語料庫(corpus)上的統計資訊，單純以鄰近字元同時出現頻率高低作為斷詞的依據。由於語料庫屬於領域相關(domain dependent)，不同語料庫間的統計資訊不適合互用。再者，統計式斷詞常受限於一階馬可夫模式(first-order Markov models)，進一步擴充此模式會提高演算法的時間複雜度，所以大多只針對二字詞進行處理，三字詞如：「大賣場」、四字詞如：「小額投信」等就無法有效擷取。而一些出現頻率較低的字詞例如姓名、地名、也容易被忽略。本篇論文以統計式斷詞法（statistical word identification）的概念為基礎，使用更有效率的方法搭配 Google 提供的中文新聞[5]服務成功地解決這些問題。

由於各家媒體所報導的新聞事件重複性很高，而新聞標題則含有新聞事件最關鍵的辭彙。透過蒐集各家媒體對於同一新聞事件所訂的標題，利用簡單的統計方法擷取這些新聞標題中重複的詞彙，就能得到單一新聞事件最具代表性的字詞。不需依賴人工的方法去整理同一新聞事件的所有標題，Google 所提供的中文新聞服務[5]，已經整理好每日各家網路新聞媒體所報導的新聞事件，不同新聞媒體所報導的同一新聞事件，在 Google 提供的新聞服務裡已被歸類在一起。本篇論文提出一個方法，藉由自動提取 Google 裡同一新聞事件的所有標題，經過統計，分析和過濾等方法，擷取每日新聞中所蘊含的詞彙，這裡頭含有舊有詞庫所包含的詞彙，但重要的是，不只是舊有的詞彙，新聞資訊裡包含著大量新詞，而這些新詞往往代表某一新聞事件最關鍵的辭彙，重複出現在同一新聞事件的所有標題，利用本文提出的方法，將能擷取到各個領域的中文字詞。由於詞庫容量過大將影響斷詞系統的效率，必須淘汰一部份不合時宜的字詞，例如一些曾經當紅的人物，其人名雖然在某一段時間經常出現，但經過歲月的流逝這些人名將不再被提起，如果這些人名依舊存放在詞庫裡，將成為一種累贅並影響斷詞系統的效率。無論是哪一種新詞擷取方法，都無法達到百分之百的正確率，而這些被擷取出來的錯誤字詞也將佔據詞庫一大空間，並影響斷詞系統的正確率。因此本篇論文將建立一個機制來避免發生此一問題，並把詞庫的正確率提升到接近百分百。除了避免詞庫過大，由於一些出現詞頻過低的字詞無法透過統計的方法有效的被擷取出來，因此本篇論文將透過分析舊有詞庫來填補統計方法的缺陷。一些新聞類相關但詞頻較低無法被擷取出來的字詞，將透過分析舊有詞庫來建立。隨著時間的流動，本篇論文所提出的方法將不斷的新增與淘汰新聞類專業詞庫裡的字詞，成為一實用且有效率的新聞類專業詞庫。

為證明本文所提出的方法，實驗的部份使用實際的新聞事件來套入本文提出的新詞自動擷取方法。並與 Yih-Jeng Lin and Ming-Shing Yu [14]所提出的方法比較其字詞擷取的數量與正確率。正確率將交由中文語言專家來判斷。實驗證明本

文提出的新詞自動擷取系統其擷取新詞的數量與正確率都明顯有較佳的表現。

1.2 研究目的

本研究欲達成之目的如下:

1. 自動提取 Google 所提供的新聞資訊並典藏成一新聞資料庫，以供未來之研究。
2. 建構一新詞自動擷取方法，實際應用 Google 提供的中文新聞服務[5]來擷取新聞類專業字詞。
3. 透過分析新聞資料來獲得新詞自動擷取方法的內部參數,期待本系統能達到最高效率。
4. 建立一符合現今需求的新聞類專業詞庫,使用在詞庫式斷詞系統,增加其效率與正確性。
5. 分類新聞類專業詞庫裡的字詞，以供未來不同之研究者使用。

1.3 研究方法

為達成預定的研究目的，本研究將採取以下的研究方法:

1. 將以排列組合的計算方式取代一般詞頻計算方式,並透過實驗選取最佳的系統參數。
2. 使用 entropy 理論來增加詞頻較低的字詞之正確性，確保新聞類專業詞庫之容量與正確性。
3. 計算新聞類專業詞庫所包含之字詞其生命週期,用以判定詞庫的內容是否符合需求。
4. 使用 PHP 軟體來撰寫本系統，並搭配 MySQL server 作為本研究之資料

庫。透過 PHP 具備之跨平台特性，使本研究所提出的方法能更加容易與其他系統結合，並方便與其他研究者之交流。

5. 以中央研究院自然語言研究小組開發的六萬詞庫為基礎,逐步建立一新聞類專業詞庫。

1.4 研究步驟

1. 建立明確的研究目標。
2. 擬定研究計畫:探討文獻，確定研究方向，擬定研究目的、方法與研究步驟。
3. 理論分析與文獻探討:蒐集有關中文斷詞系統、中文新詞擷取之相關文獻，作為本研究之理論基礎。並探討詞庫所能應用的範圍與領域，作為日後研究之參考。
4. 初步實驗:透過分析初步實驗結果，求得最適當的系統參數用以建置最終的實體系統。
5. 系統建置:以 PHP 軟體為基礎，並與 MySQL 資料庫做結合，完成一完整的新聞類專業詞庫。
6. 歸納結論與建議:將本研究心得歸納並作一總結，檢討實驗結果並提出未來可能之研究方向以供參考。
7. 撰寫研究報告:將本研究所設計之系統與實際建立的新聞類專業詞詞庫，加以彙整並完成研究報告。