應用構詞法則與類神經網路於中文新詞萃取

梁婷，葉大榮

國立交通大學資訊科學系

Page 21 ~ 40

Proceedings of Research on Computational Linguistics Conference XIII (ROCLING XIII)

Taipei, Taiwan

2000-08-24/2000-08-25

應用構詞法則與類神經網路於中文新詞萃取

梁婷　　葉大榮

國立交通大學資訊科學系

新竹市 中華民國

email:tliang@cis.nctu.edu.tw

Fax: 886-3-5721490

摘要

中文自然語言的應用近年來越來越受到重視，例如中英翻譯、文件辨識等系統。在這些應用系統中，詞庫扮演著非常重要的角色。然而，新詞不斷的產生，會影響以詞庫為基礎的應用系統效能。因此在本論文裡，我們將建構一個二階段新詞萃取機制。在第一階段利用構詞學的原理建立三音詞萃取法則用以萃取三音詞，再以非詞彙篩檢法則來過濾掉非詞彙字組以減少第二階段的分辨量。第二階段的則以詞組間的特徵統計資訊，利用類神經網路作新詞的進一步的辨認。從實驗的結果可知我們所設計的篩檢與萃取法則將可迅速地萃取新詞。此外我們並探討特徵資訊的選取與多寡對作新詞的辨認成效影響。

1. 緒論

在許多以自然語言處理為導向的文件擷取系統，常是以詞庫來輔助系統，以提升效率。然而藉由人們的使用需求，新詞將不斷地產生、增加，因此新詞的萃取技術發展也就日顯重要。

以中文語料而言，新詞的萃取方法多與斷詞程序相連結。先將語料作切割再合併切割過短的字組形成長字組，經由篩選機制過濾掉可能非為詞彙的字組，最後再對有疑義的字組予以處理。目前所發展的技術可分為統計式與法則式。

統計式的方法多是利用語料中詞彙的組成或特徵資訊，並以統計法則計算作為萃

取原則。例如，相關度(association) [Sproat90]是以相互關連度(mutual information) [Church90]為基礎，並加入字元出現順序性的考量，用來衡量字組中字元與字組間的相關程度。骰子矩陣(dice matrix) [Smadja93, 96]亦是以相互關連度為基礎，改進字組的組成字元出現機率都很低的時候，相互關連度值會過大的問題。Smadja 等人利用骰子矩陣進行連字(collocation)的抽取與兩種語言間連字的翻譯。相對頻率(relative frequency) [Wu93]是將字組出現頻率正規化的統計式特徵，利用可能度比率模組結合相互關連度與相對頻率，作英文的複合詞萃取。熵(entropy) [Tung94]可用來考量字組與語料中相鄰字元間的相關程度，Tung 利用熵作新詞萃取，並應用到文字辨識系統中。這些統計式的方法多以門檻值來判讀詞或非詞彙。雖然，詞的特徵資訊統計值，如相關度、熵等，通常都較非詞彙的統計值高，但是統計值高的卻不一定是詞。因此 Wu[93]建立可能度比率模組，結合不同的特徵統計資訊，以考量字組是為詞彙或非詞彙的機率。Chang[97]除建立可能度比率模組，更進一步結合斷詞程式作遞迴式的中文新詞的萃取。

另一方面，法則式的萃取模組則是利用構詞學與構句學的理論，配合語意資訊或詞性進行萃取，例如詞性標籤(part of speech tag)等。Yeh[91]利用馬可夫模式斷詞，再使用語意與語法的分析選取最適當的斷詞。Lin[93]先作斷詞然後利用構詞學的法則修正斷詞的結果，並以可能度比率模組來萃取新詞，增加斷詞與辨詞的效能。Nie[95]使用 maximum-matching 與經驗法則來作斷詞。將斷詞的結果再利用構詞學的方法萃取三音詞，並且利用構詞學中不具語意功能的字元刪除候選字組。Chen[97]利用詞性標籤建立法則，並利用一部份已經切割好的語料作為訓練資料，用以挑選法則。

有別於上述的萃取技術，本篇論文將利用構詞法則和字組間的特徵統計值，直接從字組庫中而不考量斷詞程序來萃取出新詞。我們主要針對二、三音詞作萃取對象並希望能快速地將其挑選出。在本論文裡，我們將建構一個二階段新詞萃取機制。在第一階段利用構詞學的原理建立三音詞萃取法則用以萃取三音詞，再以非詞彙篩檢法則來過濾掉非詞彙字組以減少第二階段的分辨量。第二階段的則以詞組間的特徵統計資訊，利用類神經網路作新詞的辨認。從實驗的結果可知我們所設計的篩檢與萃取法則

將可迅速地萃取新詞；而不同的特徵資訊與多寡也將影響類神經網路作新詞的辨認成效。

本篇論文除了緒論外第二節將介紹所提的系統概觀。第三節將介紹法則式辨認模組及一些構詞學的基本原理，包括詞的主要構成方式，與詞的一些特性。訂定詞彙萃取法則，與非詞彙辨認法則，進行詞彙的萃取與非詞彙字組的辨認，並作實驗與分析。第四節將描述我們所應用的類神經網路辨認模組及分析統計式特徵，並以實驗結果作驗證。

2. 系統概觀

本系統主要包含兩個模組，一是法則式辨認模組，另一是類神經網路辨認模組，如圖2-1。首先我們為了提供類神經網路辨認模組統計式特徵，先計算字組的統計式特徵，建立特徵資料庫，然後利用系統詞庫將已知詞從字組庫中去除。在法則式的辨認模組中，我們利用構詞學的原理訂定三音詞萃取法則，然後來萃取三音詞，再利用非詞彙辨認法則來過濾屬於非詞彙的雙字組與三字組，經過法則式辨認模組之後，尚有一些無法決定是屬於詞彙或是非詞彙的字組，再交由類神經網路辨認模組結合統計式特徵資訊來作最後的判斷屬於詞彙或非詞彙。

3. 法則式辨認模組

3-1 詞的基本定義與構成

由於有些字組是屬於詞彙或是非詞彙，若我們不加以明確的定義，則難以評估系統的效能。因此我們參考語言學與中央研究院資訊科學研究所中文詞庫小組（以下簡稱詞庫小組）訂定的分詞標準，對詞加以定義。

在構詞學中詞素(morpheme) 是語言系統中具有語意或語法功能的最小的單位[Thompson 92]。有些詞素可以獨立而自由使用，如中文的『我』、『你』、『人』等等。這些稱為『自由詞素(free morpheme)』。 不加任何附著詞素的自由詞素稱為『詞根(root)』。而有些詞素則永遠不可以單獨使用，稱為『附著詞素(bound morpheme)』。附

著詞素也叫做詞綴(affix)，附在詞根前面的叫做『前綴』(或稱『詞頭』prefix)，例如『可微分』,『可』即屬於前綴;附在詞根後面的叫『後綴』(或稱『詞尾』suffix)，例如『正規化』,『化』即屬於後綴;加插在詞根中間的稱為『中綴』(或稱『詞嵌』infix)。

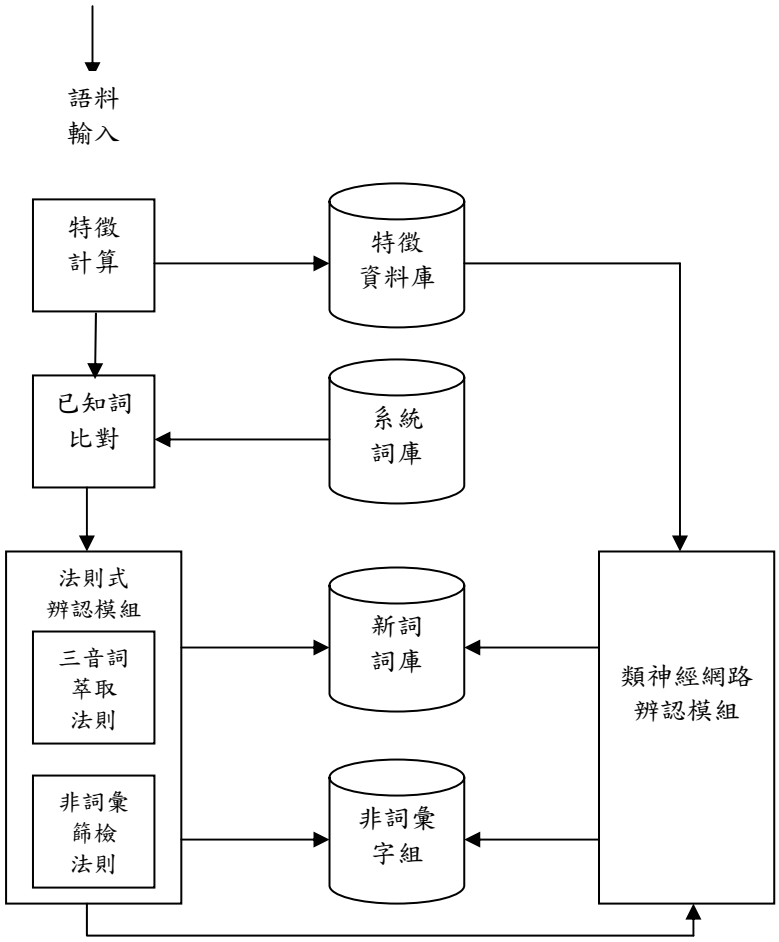

圖 2-1：系統流程圖

根據中文詞界研究與資訊用分詞標準中對詞的定義[詞庫小組]，詞為『一個具有獨立意義，且扮演特定語法功能的字串應視為一個詞』。根據詞性分類，則詞可以大略分為動詞、名詞、形容詞、副詞、定詞、量詞、介詞、方位詞、連接詞、語助詞等類。其中，動詞與名詞因為可具有詞組形式，所以有複合詞的認定問題。並且動詞、名詞是屬於『開放性詞集』[謝國平 86]，可能有新詞產生，因此在認定上困難度較高。『開放式詞集』是指可能會有新詞產生的詞集稱之，而『封閉性詞集』則是幾乎不可能有新詞產生的詞集。

詞的構成方式有許多種，比較重要的有『衍生(Derivation)』與『複合(Compounding)』兩種構詞方式[謝國平 86]。『衍生』是以衍生詞綴與詞根組合而成衍生詞的過程，例如『工業化』、『可微分』皆是衍生詞。『複合』則是指兩個詞併在一起構成另一個詞的過程，例如『遊戲』與『樹』是兩個詞，新詞『遊戲樹』可由此二詞合併得到。除了『衍生』與『複合』之外，詞的構成方式尚有許多，以下列舉幾種『略語(Acronym)』、『溶合(Blending)』、『反向構詞法(Back-formation)』、『借字(Borrowing)』、『簡縮(Abbreviation)』…等[謝國平 86]。

對於難以決定該歸於複合詞或是片語的詞組，我們依據中文詞界研究與資訊用分詞標準中所定的兩條基本原則與六條輔助原則加以分類[詞庫小組]。此外，我們對出自外來語翻譯的新詞認定，一方面考慮其在漢語的語意與語法，另一方面亦考慮原文的語意與語法，例如『能隙』(energy gap)一詞，每次在語料庫中出現都是其他詞的部分字組如『光能隙』(optical energy gap)，由於『能隙』明顯具有獨立的語意，因此認定為新詞。又例如『直方』每次在語料庫中出現都是『直方圖』(histogram)的部分字組，在原文中 'histo' 此字串在原文中並不是一個詞，而是一前接詞綴，有組織的意思，且『直方』在漢語中並無獨立的語意與語法，因此認定為非詞彙。對於化學式構成的新詞認定方面，我們將化學式視為不可切分的單位，因此化學式的部分字組一律視為非詞彙。

3-2 萃取法則

由於一般二音詞可以容易地從統計資訊萃取出來[Sproat90]，因此，本法則模組主要是針對新詞的三音詞部分提出萃取法則。萃取法則主要是依據衍生式構詞與複合式構詞，將我們所切分出來的字組中符合詞與複合詞頭、複合詞尾、衍生前綴、衍生後綴的字組視為詞。由於有一些組成詞素是屬於自由詞素或是附著詞素，各文件上的定義有所出入，但其與詞的組合都可歸為衍生詞或複合詞。

根據[詞庫小組]所附的語法詞綴、衍生詞綴、接頭/接尾詞一覽表，衍生前綴與接頭詞共有五十個，衍生後綴與接尾詞共有四百五十七個。由於我們使用詞彙萃取法則

的目的是想快速的將易於辨認出屬於詞彙的字組，所以我們觀察語料中經常出現的複合詞，挑選適合的衍生前綴與接頭詞有『主』、『副』、『非』、『多』、『超』、『子』、『單』、『雙』等共八個。挑選衍生後綴與接尾詞有『化』、『性』、『度』、『機』、『器』、『法』、『式』、『率』、『值』、『體』、『表』、『型』、『量』、『集』、『圖』、『碼』等共十六個。我們的詞彙萃取法則可以定義為以下兩條法則:

(三音詞萃取法則一):

若三字組($c_1c_2c_3$)，其中 c_1c_2 屬於雙音詞且 c_3 屬於衍生後綴或接尾詞者，並且三字組($c_1c_2c_3$)出現於語料中，其部分字組不得每次與相鄰字元構成雙音詞或三音詞。

(三音詞萃取法則二):

若三字組($c_1c_2c_3$)，其中 c_2c_3 屬於雙音詞且 c_1 屬於衍生前綴或接頭詞者，並且三字組($c_1c_2c_3$)出現於語料中，其部分字組不得每次與相鄰字元構成雙音詞或三音詞。

3-3 篩檢法則

詞常常因為語法功能相同而分為好些詞集，例如『開放性詞集』和『封閉性詞集』。『開放性詞集』包涵名詞、動詞、形容詞、及副詞，新詞往往出於此類 [謝國平 86]。反之『封閉性詞集』包涵介詞、連詞、冠詞等。這些詞類幾乎不會有新詞產生。以我們將字組分類來說，開放性詞集有可能與其他詞素結合成為新詞，封閉性詞集由於具有較固定的語法功能，與其他詞素構成新詞的機率較低。雖然，我們將詞分為『開放性』與『封閉性』，但並不保證封閉性的詞集不會有新詞產生。我們可利用封閉性詞集與其他詞素結合產生新詞機率較低的特性，加以訂定非詞彙字組的篩檢法則。

首先我們定義『封閉字集』以利於我們進行將字組篩檢。對於單字詞素能與其他詞素構成新詞的機率很低者我們稱為『封閉字』。『封閉字』的集合為『封閉字集』。我們所挑選的封閉字不一定來自封閉詞集，而是依據以下原則：

（封閉字挑選原則一）必須與其他詞素構成新詞的機率很低者。

（封閉字挑選原則二）挑選語意與語法功能簡單固定者。

（封閉字挑選原則三）盡量挑選出現頻率高者，因為這樣才能將較多的字組歸類。

原則一是我們挑選封閉字的最主要依據，因為若一中文字與其他詞素構成新詞的機率低，才符合封閉性的原則;原則二是因為語法與語意功能具有多種用法者，由其所產生新詞的機率不一定低;原則三是基於效率的考量，相對而言，通常出現頻率高者，有較多的字組可依照非詞彙篩檢法則將之歸為非詞彙。

根據上述三個原則，我們挑選的封閉字有『和』、『與』、『或』、『且』、『及』、『而』、『此』、『本』、『是』、『其』、『了』、『的』、『之』、『於』、『為』等十五個字。

首先『和』、『與』、『或』、『且』、『及』，這些字的詞性都是屬於連接詞，而且語法功能都相當明確且簡單。由連接詞的語法功能可知這些字會符合原則一與原則三。然而單字詞可能有兩種以上的語意或語法功能，例如『暖和』、『或然率』、『苟且』等，但這些情形大部分是詞庫中已經存在的詞，或是出現的機率很低。

『而』、『此』的語法及語意功能較為固定。『本』則有較多語意上的變化如『樣本數』、『超本文』(超本文會根據衍生詞構詞規則，歸為詞彙)，然而，『本』在我們的語料中單字的出現頻率是屬於出現頻率高的單字，且主要仍以『冠詞』的詞性出現，因此將『本』加入封閉字集中。『是』在我們語料中單字出現頻率是屬於高出現頻率單字，且語意語法固定，所以加入封閉字集。『其』的詞性歸於代名詞，之外用法『其他』、『其餘』。由於『其』的語意語法固定，與其他詞素結合為新詞的機率低，所以亦加入封閉字集。至於『了』我們視為構形詞綴如在『吃了』，並不改變語義。而『的』、『之』可視為修飾語與中心語之間的分隔標記且其使用頻率上高所以亦加入封閉字集。『於』、『為』則視為介詞。我們基於易於處理一律將『動詞＋於』、『動詞＋為』、『動詞＋成』視為非詞彙。

因此我們將字組歸為非詞彙的法則，即

(非詞彙字組篩檢法則一):

字組中包含{和、與、或、且、及、而、此、本、是、其、了、的、之、於、為}者則歸為非詞彙。

除了以封閉字將非詞彙字組篩檢出來以外，我們又使用另一簡單而有效率的非詞彙篩檢法則稱為部分詞彙篩檢法則：

***(非詞彙字組篩檢法則二)**:*

若一字組在語料庫中每次出現的情形，其部分字組皆與相鄰字元形成雙音詞或三音詞者，則歸為非詞彙。

3-4 實驗與分析

我們所使用的語料庫是由交通大學圖書館的中華碩博士論文查詢系統，所下載的資訊相關系所的 3,646 篇碩博士論文，來自資訊工程研究所、資訊及電子工程研究所、資訊科學研究所、資訊教育研究所、資訊管理技術研究所、資訊管理研究所、電子與資訊工程技術研究所、電機與資訊工程研究所等不同系所。在語料庫 3,646 篇碩博士論文中共有 1,163,928 字，包括 2680 個不同的字，本論文針對雙音詞與三音詞作新詞萃取，經字組抽取後共有 1,058,078 個雙字組與 956,046 個三字組，其中不同的雙字組共有 110,258 個，不同的三字組有 344,585 個。字組出現次數超過四次且不存在於系統詞庫中的雙字組與三字組各 22,172 個與 32,119 個。我們使用教育部所發展的詞庫作為系統詞庫來過濾已知詞，其中包含有 48,330 筆雙音詞，11,558 筆三音詞。

我們分別定義了新詞萃取正確率與召回率，非詞彙字組篩檢正確率與召回率，來衡量系統的效能：

$$\text{新詞萃取正確率} = \frac{\text{萃取正確之總數}}{\text{萃取為新詞之總數}} \tag{3.1}$$

$$\text{新詞萃取召回率} = \frac{\text{萃取正確之總數}}{\text{新詞之總數}} \tag{3.2}$$

$$\text{非詞彙字組篩檢正確率} = \frac{\text{篩檢正確之總數}}{\text{篩檢為非詞彙字組之總數}} \tag{3.3}$$

$$\text{非詞彙字組篩檢召回率} = \frac{\text{篩檢正確之總數}}{\text{非詞彙字組之總數}} \tag{3.4}$$

實驗中，三字組庫中所有的三音新詞共有 1246 個，經由詞彙字組萃取法則所取出來的

三字組共有 761 個三字組，其中正確的共有 707 個新詞，例如，壓縮率、門檻值、超媒體、子集合等; 錯誤的的共有 56 個錯誤，例如，利用圖、行程式等，因此新詞萃取正確率是 92.9%，召回率是 56.74%如圖 3-1。

造成詞彙字組萃取法則錯誤的原因，是因為雙音詞與接頭詞、接尾詞、詞綴的結合，在語料中並不一定會構成正確的句法結構。根據萃取錯誤的三字組與前後文的關係，可以將錯誤分為兩類，第一類是萃取錯誤的三字組中其部分字組是屬於另一雙音詞或三音詞。因為萃取錯誤的三字組中的部分字組是屬於一新詞，並非所有的新詞皆是由詞庫中的詞與詞綴、接頭詞或接尾詞的結合而成，例如『四元樹』、『波茲曼』等。

在尚未歸類的字組中有許多可以經由法則式非詞彙篩檢法則正確地篩檢出來，因此我們先利用法則式非詞彙篩檢法則將一些可正確歸類的字組篩檢出來。雙字組庫中原本有 22172 個雙字組，其中 21399 個非詞彙雙字組。根據兩條非詞彙篩檢法則，將 15884 個雙字組篩檢出來，其中 15882 個正確篩檢，正確率 99.99%，召回率 74.22% 見圖 3-2。經過三音詞萃取法則之後，剩下 31358 個三字組，其中包含 30858 個非詞彙三字組，依據兩條非詞彙篩檢法則篩檢出 21231 個三字組，其中有 21213 個正確的篩檢，正確率 99.92，召回率 68.74%見圖 3-3。

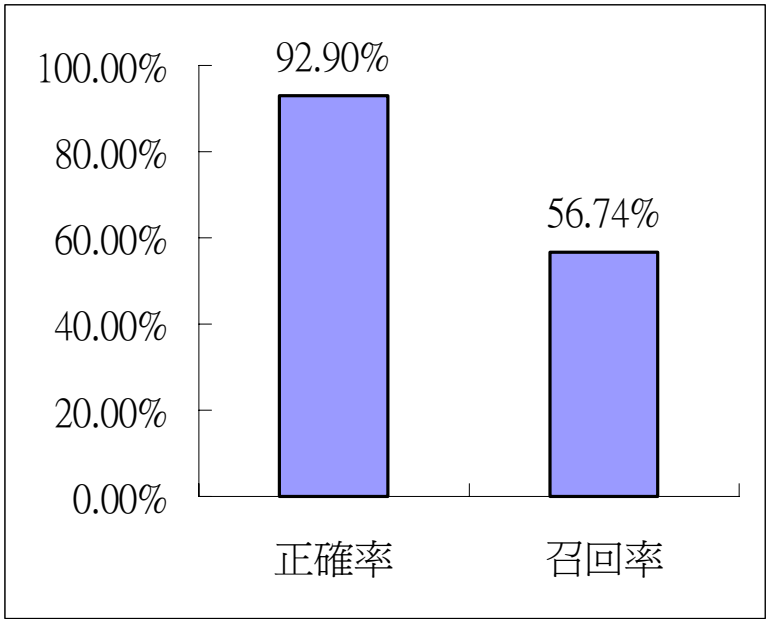

圖 3-1：法則式三音詞詞彙萃取正確率與召回率

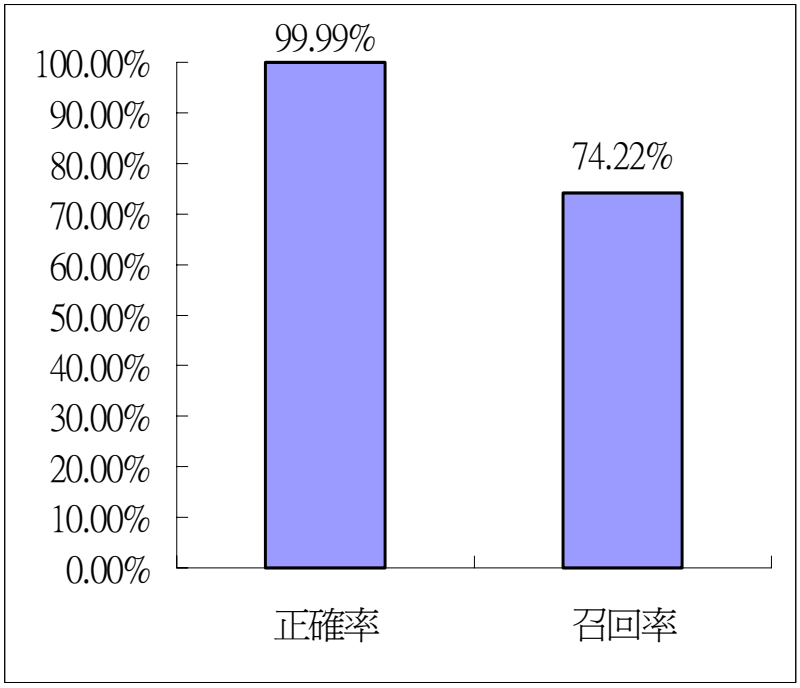

圖 3-2：雙字組非詞彙篩檢正確率與召回率

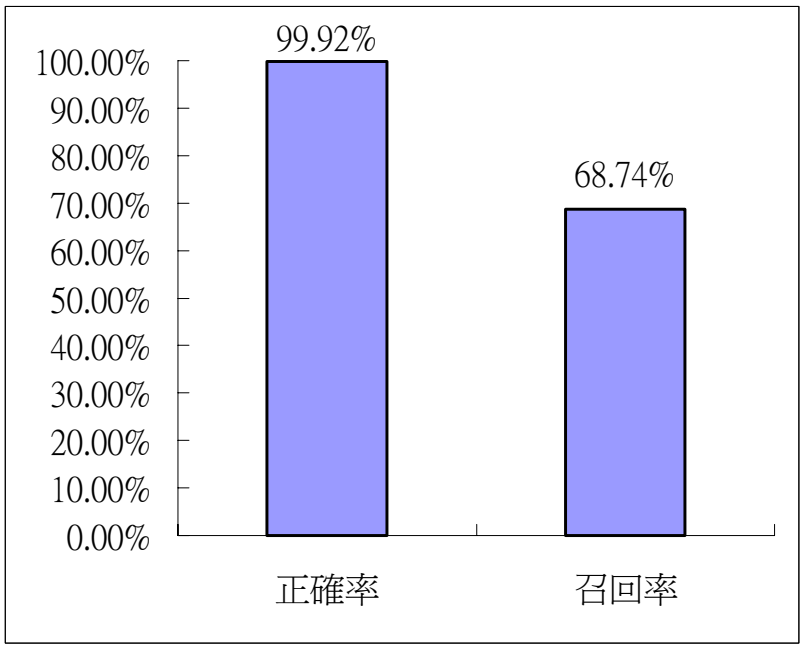

圖 3-3：三字組非詞彙篩檢正確率與召回率.

4. 類神經網路辨認模組

4-1 倒傳遞網路

由於詞的構成方式頗為複雜，且各中文字可能具有特殊的使用情形，因此若要以純法則式的辨認方法來判斷分類，必須要建立起很多的法則。另一缺點是建立的法則需考量不同的語料庫的特性。因此在第二階段的篩選時讀我們乃以字組間特徵值做為辨詞依據。

由於類神經網路中倒傳遞網路(multi-layer feed-forward with back propagation)具有學習正確率高、理論簡明[Zurada 92]，因此我們可將所挑選的特徵統計值做為此網路的輸入，建構成詞彙與非詞彙的分類器。我們使用具有一層隱藏層的倒傳遞網路，並且與輸入層是完全連接(full connect)且與輸出層亦是完全連接。在神經元的架構中，我們使用雙曲線正切函數(hyperbolic-tangent function)作為轉換函數。此函數具有微分容易的優點，可配合差距法則調整神經元間的權重，此函數當自變數趨於正負無限大時，函數值趨近於常數，其函數值域在[-1,1]之間。

4-2 特徵選取

統計式的特徵經常用到字組的出現機率，因此先定義字組的出現機率，再說明各種統計式的特徵。本論文將利用出現頻率來評估字組的出現機率如下

$$P(G_{ij}) = \frac{T(G_{ij})}{\sum_{j} T(G_{ij})} \tag{4.1}$$

其中 T（G_{ij}）表示長度為 i 的第 j 個字組 G_{ij} 的出現次數。字組間的特徵有

(1) 相對頻率(relative frequency count)[Wu 93]是將字組的出現次數除以所有字組的平均出現次數如公式(4.2)。

$$r_{ij} = \frac{f_{ij}}{K_i} \tag{4.2}$$

其中 r_{ij} 是指長度為 i 的字組庫中的第 j 個字組，f_{ij} 是 r_{ij} 的出現次數，K_i 是指長度為 i 的字組庫中所有字組的平均出現次數。一般的情況來說，相對頻率越高的字組，可能是屬於詞類的機率越高。

(2) 相關度(Association)[Sproat 90]定義如下：

$$A(ab) = \log_2 \frac{P(ab)}{P(a) \times P(b)} \tag{4.3}$$

其中 P(a)、P(b) 分別代表中文字 a 與 b 的出現機率。P(ab)代表雙字組 ab 的出現機率。此統計特徵有一缺點，當 P(a)、P(b)都很小的時候，A(ab)容易變得很大。三字組的相關度 A(abc)定義為：

$$A(abc) = \log_2 \frac{P(abc)}{P(a) \times P(b) \times P(c)} \tag{4.4}$$

其中 P(a)、P(b) 、P(c) 分別代表中文字 a、b 與 c 的出現機率，P(abc)則代表三字組 abc 的出現機率。

(3) 骰子矩陣 [Smadja93, 96]的定義如下：

$$D_2(x, y) = \frac{2P(x=1, y=1)}{P(x=1) + P(y=1)} \tag{4.5}$$

其中 P(x=1,y=1) 是中文字 y 緊跟著中文字 x 出現的機率，P(x=1)與 P(y=1)則分別是中文字 x、y 出現的機率。由上式可發現骰子矩陣與相關度很像，當 P(x=1)與 P(y=1)都很小的時候，則骰子矩陣是比相關度好的評量標準。三字組的骰子矩陣定義如下：

$$D_3(x, y, z) = \frac{3P(x=1, y=1, z=1)}{P(x=1) + P(y=1) + P(z=1)} \tag{4.6}$$

其中 P(x=1,y=1,z=1) 是中文字 z 緊跟著 xy 出現的機率，P(x=1)、P(y=1)與 P(z=1)則分別是中文字 x、y、z 出現的機率。

(4) 熵(Entropy) [Tung 94]是用來衡量字組與其相鄰字元的關係。若是有一字組的相鄰字元出現的分佈很亂，則可以想見在此字組很可能是一個詞。熵的定義如下：

$$H\_L(G_i) = -\sum_{C_j \in LN(G_i)} P\_L(C_j) \log_{T(G_i)} P\_L(C_j) \tag{4.7a}$$

$$H\_R(G_i) = -\sum_{C_j \in RN(G_i)} P\_R(C_j) \log_{T(G_i)} P\_R(C_j) \tag{4.7b}$$

$$P\_L(C_j):\frac{T(C_j-G_i)}{T(G_i)} \tag{4.7c}$$

$$P\_R(C_j):\frac{T(G_i-C_j)}{T(G_i)} \tag{4.7d}$$

其中 H_L(G_i)與 H_R(G_i)分別代表字組 G_i 的左熵與右熵，LN(G_i)與 RN(G_i)分別代表字組 G_i 的左相鄰字元集合與右相鄰字元集合，P_L(C_j)與 P_R(C_j)則分別代表字元 C_j 在 G_i 的左相鄰字元集合的出現機率，與右相鄰字元集合的出現機率。

從實驗中我們發現幾乎所有字組的相對頻率與骰子矩陣的特徵值都落在值域的最小百分之五，尤其骰子矩陣幾乎全都落在 0 到 0.05 之間，這樣的分佈幾乎顯不出字組的差異性。而二字組的相關度分佈情形相當接近高斯分佈（Normal Distribution），左熵與右熵除了特徵值為 0 的個數較多之外，其餘的分佈較為平均。因此我們首先以相關度、左熵與右熵作為系統的輸入特徵。若要考慮自動特徵選取的問題可以參考循序向前選取(Sequential Forward Selection)、Generalized "Plus l-Take Away r" Selection[Devijver 82]。若是原來特徵值分佈情形不佳的情形，可以透過一些轉換函數例如高斯函數(Gaussian Distribution)、雙彎曲函數(Sigmoid function)、雙彎曲正切函數(Hyperbolictangent function)來將值域與分佈情形加以轉換。利用自動特徵選取以獲得更好的系統效能是未來需要再加以研究的。

4-3 實驗與分析

在實驗中我們乃是以所提的倒傳遞網路辨識器和可能度(Likelihood)模組做分析比較[Duda 73]。可能度是評估在某一特定的情形下事件發生的機率。我們使用的可能度比率模組主要是修改 Wu [93]所提出用於抽取英文複合詞的模組。Wu 所選取的統計式特徵為相對頻率與相關度，因此利用雙變數的高斯函數的分佈作為可能度比率模組的機率分佈。以下是 Wu 所使用的詞類與非詞類雙變數機率函數：

$$f(A,R\,|\,Word)=$$
$$\frac{1}{2\pi\sigma_a\sigma_r\sqrt{1-r^2}}\exp\left\{-\frac{1}{2(1-r^2)}\left(\frac{(A-\mu_a)^2}{\sigma^2{}_m}-2r\frac{(A-\mu_a)(R-\mu_r)}{\sigma_a\sigma_r}+\frac{(R-\mu_r)^2}{\sigma^2{}_r}\right)\right\} \tag{4.8a}$$

$$f(A,R \mid Non-Word) =$$
$$\frac{1}{2\pi\sigma'_a\sigma'_r\sqrt{1-r'^2}}\exp\left\{-\frac{1}{2(1-r'^2)}\left(\frac{(A-\mu'_a)^2}{\sigma'^2_m}-2r'\frac{(A-\mu'_a)(R-\mu'_r)}{\sigma'_a\sigma'_r}+\frac{(R-\mu'_r)^2}{\sigma'^2_r}\right)\right\} \quad (4.8b)$$

其中 A 和 R 是代表相關度與相對頻率的變數。假設 A 和 R 都是屬於高斯分佈，而 μ_a 是詞類字組的相關度平均數、μ'_a 是詞類字組的相關度平均數，μ_a 是詞類字組的相對頻率平均數、μ'_a 是詞類字組的相對頻率平均數，σ_a 是詞類字組的相關度標準差，σ'_a 是非詞類字組的相關度標準差，σ_r 是詞類字組的相對頻率標準差，σ'_r 是非詞類字組的相對頻率標準差，r 是詞類字組相關度與相對頻率的相關係數，r' 是非詞類字組相關度與相對頻率的相關係數。定義了機率函數後，將機率函數套入對數可能度比率模組，若是 log λ 小於門檻值 T_{lrm} 則是屬於非詞類，若是 log λ 大於 T_{lrm} 則屬於詞類，在本系統 T_{lrm} 的預設值是 0。

我們所選取的特徵是相關度、左熵與右熵，因此我們使用多變數的高斯函數來作為可能機率函數[Duda 73]：

$$f(x \mid word) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right] \quad (4.9a)$$

$$\mu = E[x \mid word] \quad (4.9b)$$

$$\Sigma = E[(x-\mu)(x-\mu)^t] \quad (4.9c)$$

$$f(x \mid Non-word) = \frac{1}{(2\pi)^{d/2}\,|\Sigma'|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu')^t\Sigma'^{-1}(x-\mu')\right] \quad (4.9d)$$

$$\mu' = E[x \mid Non-word] \quad (4.9e)$$

$$\Sigma' = E[(x-\mu')(x-\mu')^t] \quad (4.9f)$$

其中 x 代表一個行向量[A LH RH]^t，A、LH 與 RH 分別代表相關度、左熵與右熵的變數，並假設此三變數是屬於高斯分佈，μ 是代表詞類字組的特徵平均值，μ=[μ_A μ_{LH} μ_{RH}]^t，μ_A μ_{LH} μ_{RH} 分別代表詞類字組的相關度、左熵與右熵的平均值，μ' 是代表非詞類字組的特徵平均值，μ'=[μ'_A μ'_{LH} μ'_{RH}]^t，μ'_A μ'_{LH} μ'_{RH} 分別代表非

詞類字組的相關度、左熵與右熵的平均值，Σ是代表詞類字組特徵的相關係數矩陣，Σ'代表非詞類字組的特徵相關係數矩陣。

我們利用在 3.4 節定義的正確率與召回率來評估系統的效能，另外以加權式正確召回率(weighted precision recall,WPR)做為衡量，

$$\text{加權式正確召回率} = W_1 \times \text{正確率} + W_2 \times \text{召回率}, \tag{4.10}$$

其中 W_1 與 W_2 皆設定為二分之一。

我們以亂數選取三分之二的字組作為訓練資料，分別使用可能度比率模組與類神經網路模組進行新詞萃取。我們使用相關度、左熵與右熵作為統計式特徵，並且利用多變數高斯函數作為可能度比率模組的機率分佈，計算訓練資料的平均值與相關係數矩陣，可得到高斯函數的參數，套入高斯函數後可得到可能度比率模組。

在類神經網路萃取模組方面，由於相關度的值域比左熵與右熵大許多，因此我們先將此三種特徵作一簡單的值域轉換，將特徵的值域轉換到[0.05，0.95]及[-0.95，-0.05]。因為 0 在倒傳遞網路中，是沒有作用的，因此避開 0。若是特徵 X 的值永遠大於零，則使用以下的轉換函數

$$f(x) = \frac{0.95 - 0.05}{\max - \min}(x - \min) + 0.05 \tag{4.11a}$$

否則使用此函數

$$f(x) = \frac{0.95 - 0.05}{\max}(x) + 0.05, \text{ if x } \geqq \ 0 \tag{4.11b}$$

$$f(x) = \frac{0.95 - 0.05}{\max}(x) - 0.05, \text{ if x} < 0 \tag{4.11c}$$

因為我們首先只使用三種特徵，所以輸入層的節點個數是三個，輸出值亦只有一個，我們使用的轉換函數是雙彎曲函數其值域為[-1，1]，若是輸出值大於 T_{mlff} 則視為詞彙，若是輸出值小於 T_{mlff} 則視為非詞彙類別，系統預設的 T_{mlff} 是 0。

圖 4-1 是調整不同門檻值 T_{mlff} 與 T_{lrm} 時，類神經網路模組與可能度比率模組雙音

詞萃取正確率與召回率的變化情形。在雙字組的新詞萃取方面，當高召回率的情形時，類神經網路模組的正確率優於可能度比率模組;而當低召回率的情形時，可能度比率模組的正確率則優於類神經網路模組。

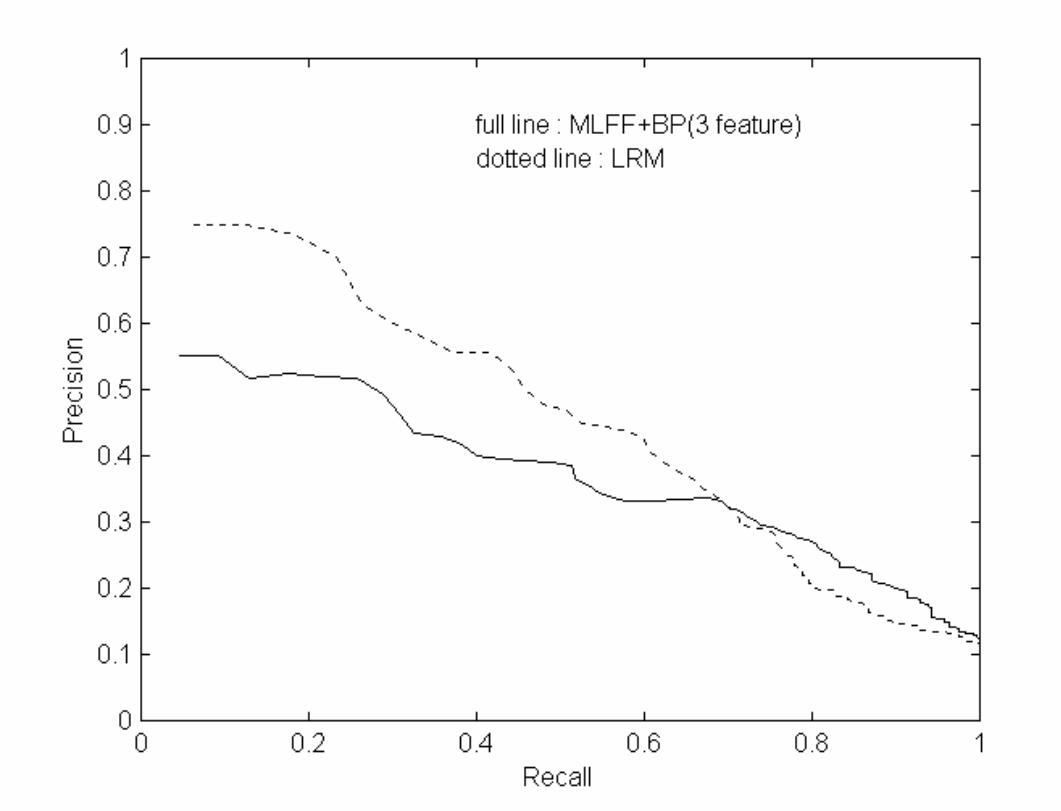

圖 4-1 雙音詞萃取效能比較圖

在三音詞的萃取方面，類神經網路模組在門檻值為預設值時，略優於可能度比率模組。觀察圖 4-2，發現類神經網路模組與可能度比率模組於三音詞的萃取能力並無明顯的優劣分別。

由於三字組中其二字組的資訊是有意義的因此在類神經網路模組三字組 $c_1c_2c_3$ 新詞萃取中除了原本使用的相關度、左熵與右熵的三個特徵外，我們另加入其部分字組 c_1c_2 與 c_2c_3 的相關度、左熵與右熵作為特徵，特徵個數增加為九個。在表 4.1 和圖 4-3 可知以特徵數的增加確實可提高分辨的正確率。

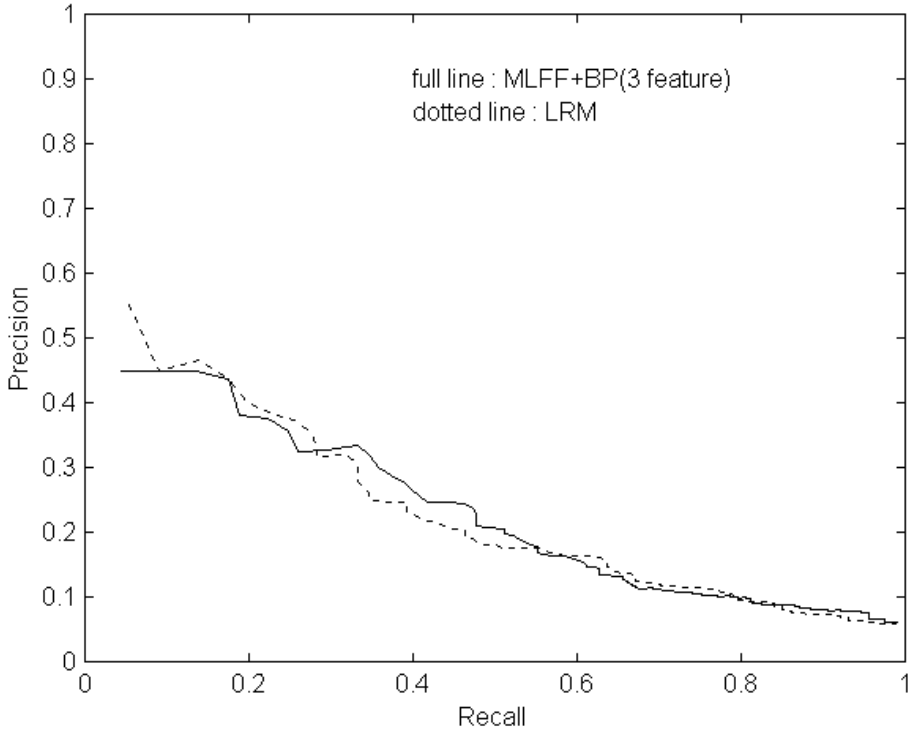

圖 4-2：三音詞萃取效能比較圖

	可能度比率模組	類神經網路模組（三種特徵）	類神經網路模組（九種特徵）
正確率	16.32%	13.68%	18.97%
召回率	59.3%	63.32%	77.89%
加權式正確召回率	37.81%	38.5%	48.83%

表 4-1 三音詞萃取效能比較表

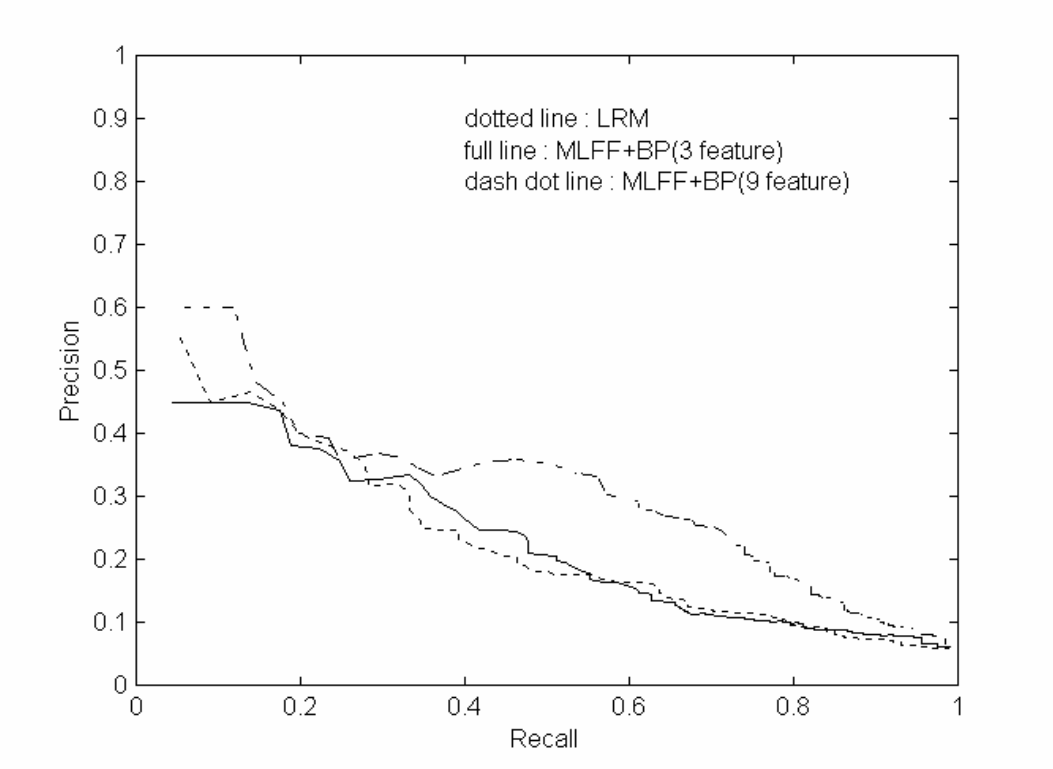

圖 4-3：三音詞萃取效能比較圖

5. 結論

本論文提出一個兩階段的中文新詞萃取技術，可應用於中文文件處理系統，將語料中有意義的新詞萃取出來。實驗數據的分析顯示利用構詞學的方法確實能有效的將三音新詞萃取出來，並且正確地將大部分非詞彙字組過濾掉。另一方面利用類神經網路結合各種統計式資訊來萃取新詞，可彌補構詞法則的侷限性。最後我們亦探討特徵的選取對於萃取的影響，並與可能度模組比較。從實驗的結果我們得知三音詞中二字組特徵的加入確實能提高三音詞新詞的正確率與召回率。

本論文的後續研究方向主要有特徵的自動選取。在使用類神經萃取模組時，選取的特徵的好壞會直接影響到系統的效能，在本論文使用分析其值域分佈情形來作特徵選取。但是當可使用特徵很多，導致難以逐個分析時，則需利用特徵自動選取來解決這個問題。與此問題相關的還有特徵的值域轉換問題，當各種特徵資訊的值域範圍相差太大時，就需要特徵的值域轉換，避免系統被少數幾個特徵所主宰。

6. 參考文獻

Li, Charles N. and Thompson, Sandra A., "Mandarin Chinese," University of California Press, New York, 1992.

Yeh, Ching-Long and Lee, His-Jian, "Rule-Based Word Identification for Mandarin Chinese Sentences – A Unification Approach," Computer Processing of Chinese & Oriental Languages, Vol. 5, No.2, March 1991.

Smadja, Frank, "Retrieving Collocations from Text: Xtract," Computational Linguistics, Vol. 19, No. 1, 1993, pp. 143-177.

Smadja, Frank, McKeown, K.R. and Hatzivasiloglou, V. "Translating Collocations for Bilingual Lexicons," A Statistical Approach," Computational Linguistics, Vol. 22, No. 1, 1996.

Zurada, Jacek M., "Introduction to Artificial Neural Systems", West Publishing Company, USA, 1992.

Nie, Jian Yun, Hannan, Marie-Louise and Hannan, Wanying, "Combining Dictionary, Rules and Statistical Information in Segmentation of Chinese," Computer Processing of Chinese and Oriental Languages, Vol. 9, No. 2, December 1995, pp. 125-143.

Chang, Jing Shin, "Automatic Lexicon Acquisition and Precision-Recall Maximization for Untagged Text Corpora", National Tsing-Hua University, P.h.D. thesis, 1997.

Chang, J. S., Chen, C. D. and Chen, S. D., "Chinese Word Segmentation through Constraint Satisfaction and Statistical Optimization," (in Chinese) Proceedings of ROCLING-IV, R.O.C. Computational Linguistics Conferences, Taiwan ROC, 1991, pp. 147-165.

Church, K. and Hanks, P., "Word Association Norms, Mutual Information and Lexicography," Computational Linguistics, Vol.16, March. 1990, pp. 22-29.

Chen, Keh Jiann, Bai, Ming Hong, "Unknown Word Detection for Chinese by a Corpus-based Learning Method," Proceedings of ROCLING X, Taipei, Taiwan, ROC, 1997, pp. 159-174.

Lin, M.Y., Chang, T. H. and Su, K. Y., "A preliminary study on unknown word problem in Chinese word segmentation," Proceedings of 1993 R.O.C. Computational Linguistics Conference, Taiwan, 1993, pp.119-137.

Wu, M. W. and Su, K. Y. "Corpus-based Automatic Compound Extraction with Mutual Information and Relative Frequency Count," Proceedings of ROCLING VI, Nantou, Taiwan, ROC, Sep. 1993pp. 207-216.

Sproat, Richard and Shin, Chilin "A Statistical Method For Finding Word Boundaries In Chinese Text," Computer Processing of Chinese & Oriental Language, Vol. 4, No. 4, March 1990.

Chen, S. C. and Su, K. Y. "The Processing of English Compound and Complex Words in an English-Chinese Machine Translation System," Proceedings of ROCLING I, Nantou, Taiwan, 1988, pp. 87-98.

劉興寰, “中文語料詞類自動標記,” 國立清華大學, 碩士論文, 1994。
謝國平, “語言學概論,” 三民書局, 1986。
詞庫小組, “新聞與語料詞頻統計表,” 1993。
詞庫小組, “搜文解字：中文詞界研究與資訊用分詞標準,” 1996。