國語語音辨認中詞群語言模型之分群方法與應用

Methodology Implementation and Application of Word-Class Based Language Model in Mandrain Speech Recognition

張元貞[1]，林頌堅[1]
簡立峰[2]，陳克健[2]，李琳山[1,2]

[1]國立臺灣大學資訊工程研究所
[2]中央研究院資訊科學研究所
email:lsc@speech.ee.ntu.edu.tw

摘 要

統計式馬可夫語言模型由於實作容易，且在語音辨認上的正確率能夠維持相當的水準，因此近年來得到相當廣泛的使用。然而這類語言模型也存在一些困難，如訓練語料不足時連帶導致參數值可信度較低，及語言模型參數過多在使用上造成龐大的記憶體需求。

本文即是針對上述這些問題，提出改善方法。我們提出一個以詞群爲基礎的語言模型，且配合一種把詞自動分群處理的技術，將統計資訊相似的詞歸爲一群，利用同群之詞彙分享彼此統計資訊的特性來解決前述問題。藉由此方法所得的分群結果，經由觀察發現與文法上的詞類有相當程度的吻合。且將此結果應用於語音辨認上，由實驗的辨認率來看，以詞群爲基礎的語言模型接近於詞雙連語言模型，但所需的記憶體則遠少於詞雙連語言模型。

一、緒論

統計式馬可夫語言模型由於實作容易，且在語音辨認上的正確率能夠維持相當的水準，因此近年來得到相當廣泛的使用。然而這類語言模型也存在一些困難，如訓練語料不足時連帶導致參數值可信度較低，及語言模型參數過多在使用上造成龐大的記憶體需求。

以字或詞雙連文法在中文的應用爲例，一般認爲要訓練出可信的以字爲基礎的中文語言模型須要約 $(13000)^3$ 個字的語料庫，而要訓練出可信的以詞爲基礎的中文語言模型，則理論上將需要多達約 $(100000)^3$ 個詞的語料。且在使用時，13000 × 13000 的字雙連文法 (character bigram) 參數表及 100000 × 100000 的詞雙連文法 (word bigram) 參數表即使以稀疏矩陣 (sparse matrix) 的方式儲存，對於記憶體的需求仍是相當大的負擔。

本文主要是提出一個以詞群爲基礎的語言模型 (word-class-based language model) 以改善前述問題，這個語言模型是根據統計分群，所以不佔太多記憶體空間，且具有資訊分享的優點及快速的辨認時間及高正確率。近年來以統計方式分群的研究已有若干 [Brown 92,Ney and Essen 91,Schutze 93,Chang 93]，但仍少見成功應用在語音辨認的成果發表。由於先前我們已發展根據詞頭尾字分群之詞群雙連文法 [林 '93] 並成功應用在國語語音辨認，因此本文即是進一步利用統計特性分群來發展中文詞群語言模型並應用在國語語音辨認。

我們所採用之分群方法主要是根據詞的向量量化。要建立以詞群爲基礎的語言模型，必須先構成詞群。我們的作法是先統計出詞在語料庫中的相連情況，再以這些詞的相連情況做爲分群的依據，採用向量量化 (Vector Quantization)[Gray 84] 的方法來構成詞群。以實用的觀點而言，將詞分群相當費時；所以，我們採用奇異值分解 (Singular Value Decomposition) 的若干理論，在不損失太多資訊的情況下，加速分群的速度 [Schutze 93,Deerwester 90]。另外我們亦採用了一個兩階段式的分群流程來加速分群。其中第一階段處理是針對常用的高頻詞來進行分群，主要是希望利用常用高頻詞構成之群所蘊含之豐富的資訊來當做第二階段分群時所使用的特徵，以利第二階段分群的進行。而第二階段則是對所有

的詞進行分群。

以下第二節是有關分群方法之介紹，第三節是分群流程與結果之觀察及討論。第四節是將前述分群結果建立成中文語言模型及在國語音辨認之應用。第五節是實驗結果分析。最後第六節是結論

二. 分群方法

詞群的構成

在自然語言處理的研究中，關於詞的分群已經有相當多的研究成果發表 [Schtuze 93,Ney 91,Brown 92]，其主要用途是字義歧義的解決 (word sense disambiguity)、詞類預測 (category prediction) 及做爲輔助訊息以降低語言模型的文字複雜度 (perplexity) 等。不同的用途產生不同的分群概念。然而在語音辨認的應用上究竟如何分群較爲適當？由於我們希望藉由詞的分群來解決語言模型之參數值可信度不足及記憶體需求太大的問題，所以爲了配合此一需求，我們相信在語言模型的使用上具有相近統計性質的詞若能歸爲一群應屬恰當。由於我們主要要改進的對象是詞雙連 (word bigram) 語言模型，所以這裏所指的統計性質主要是指任一詞在語料庫中所有前接詞或後接詞的種類及頻率的統計，如圖 1 所示。簡言之，我們的觀念就是以資訊分享 (information sharing) 來達到解決問題的目的。

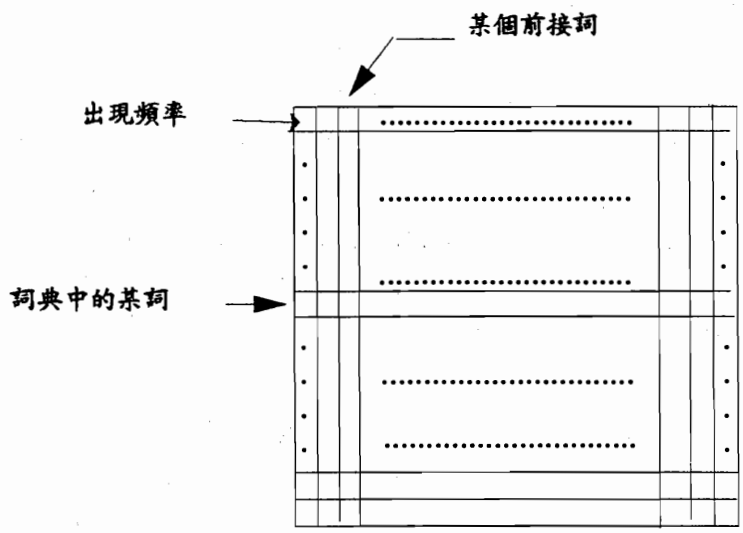

圖1 詞雙連參數表

雖然語言學上早有以文法觀點制定的詞類觀念存在[趙 80]，我們仍以實用觀點嘗試以統計的角度來進行詞分群的處理。舉例來說，「一月」和「五月」這兩個時間詞，前面所接的詞常常同是「在」之類的介系詞，即使有不同的情況發生，也不致於有太大的差異（相對其它詞而言）。因爲這些詞的統計訊息相似，如能順利地將它們構成一個詞群，則類似的統計訊息只需儲存一份，不必像原來一樣必須重覆儲存，如此，則達到了記憶體縮減(memory reduction)的目的。又假設「三月」這個詞，由於訓練語料不足產生了參數值不可信的情況，如「三月」能和「一月」、「五月」構成同一個詞群（因爲它們在使用上有極相似的特性），透過資訊的分享(information sharing)，就可以解決參數值可信度低的問題。

分群的方法－詞的向量量化

本文所使用爲分群訊息主要是根據以〝詞空間〞(word space)的概念[Schtuze 93]，也就是將語料庫中詞與其前接詞或後接詞的統計分佈情況當作這個詞的特徵(feature)，而以向量的型式來表示。於是分群的目的也就演變成是要將近似的向量聚集成群。我們發現在語音辨認處理中有一項相當重要的技術－向量量化(Vector Quantization)[Gray 84, 吳 91]，其觀念及目的與我們的需求非常的吻合，並且向量量化在實用上也已獲得證明，所以我們決定採用向量量化做爲我們的主要分群方法。

我們採用〝詞空間〞的概念，把詞視爲空間中的一點，每個詞以一個向量來代表(即圖1中矩陣之列)，向量中的每個維度則代表這個詞與某一個前接詞在語料庫中相鄰的頻率。原本在向量量化中失眞度的量測我們則改以相似度的觀點視之。而相似度的衡量是以向量間夾角的餘弦值做爲量度的標準，餘弦值愈大，表示兩向量的夾角愈小，意即其間的相似程度愈大，其成群的可能性愈大；反之，則表示其間的相似程度愈小，其成群的可能性愈小。

我們所設計詞的向量量化程序，主要分成二個階段，分別是初始化(initalization)及最佳化(optimization)。初始化的主要目的產生一個粗略的碼本，也就是將詞典中的所有詞做一個大概的分群。其作法是根據所需要的群數，也就是碼本大小，依照要達到向量量化最佳化的兩個必要條件，利用分離(split)的程序去切割詞

典的詞所構成的詞空間，並同時更新各群的中心點(centroid)。第一階段結束後即得一個初始的碼本。第二階段是最佳化的程序。在完成初始化程序後，所得到的分群結果只是一個雛形，存在有相當地的誤差，距離最佳化仍有一段距離。因此在最佳化的程序裏，根據最小失眞度準則，重新找出新屬之群，如此會造成某些單元脫離原屬之群，而分配至他群，因此會造成各群中心點的偏移，所以必須將各群的中心點重新調整。在這個階段我們希望整體的失眞度能夠透過疊代的過程使之不斷下降。在我們的應用中失眞度改以相似度視之，因此在衡量上我們是以所有詞與其所屬群之中心點的餘弦值之總和。只要連續兩次疊代過程失眞度之差小於某個臨界值或設定疊代次數即可終止最佳化程序。

其詳細流程如下：

階段一：初始化

重覆以下步驟

- 找出欲分群之群G中心點向量Co。
- 找出G中與中心點向量Co夾角最大的向量，做爲一個中心點向量C1。
- 找出G中與中心點向量C1夾角最大的向量，做爲另一個中心點向量C2。
- 將G中所有的向量分到以C1、C2爲中心點向量的兩群。
- 在現有的群中，找出平均相似度最小的群，做爲下次疊代所欲處理之群G。

直到群數等於應用所須之群數。

階段二：最佳化

重覆以下步驟

- 將所有向量依據最小失眞度準則重新分群。
- 修正重新分群後的各中心點向量，並且記錄總體相似度。
- 計算出此次總體相似度與上次總體相似度之差D。

直到D大於設定的臨界值或達到所設定的疊代次數。

分群的加速策略－奇異值分解

在瞭解前述詞的向量量化之詳細流程之後，我們進一步討論其執行效率。從其計算所花費之代價來看，如果在向量的維度上能夠縮減，則對執行效率能夠有相當大的助益，在參閱舒氏的相關論文[Schtuze 93]，了解本研究與其極相似的情況之下，採用奇異值分解(Singular Value Decomposition)的技巧在不產生太多失眞下，大量地縮減了向量的維度。因此，我們也嘗試採用SVD來做爲我們向量量化的加速策略，它主要是處理我們分群所依據的詞雙連參數矩陣，將其分解成U，V兩個正交化矩陣及∑這個對角化矩陣，選擇適當數目的奇異值k個，則可降低向量維度，加快分群處理的速度。有關奇異值分解的細節，可參照舒氏的論文。

三．分群的流程及結果之觀察與探討

由於目前我們所使用的中文語料庫字數計4,024,370字，斷詞後所得之詞數爲2,644,183詞，以量而言，極爲有限，與在西文的相關研究所使用動輒千萬字以上的訓練語料，仍爲不足。爲了解決這個問題，我們根據中研院中文詞知識庫小組(CKIP)的研究報告得知經由統計，高頻的5,000詞已經佔日常使用詞彙的90%以上。在瞭解這個訊息之後，我們參考舒氏[Schtuze 93]的作法，研擬出一個兩階段式的因應對策，在第一階段先對常用的5000詞進行分群，將其分成500群，當作特徵(feature)供第二階段使用。第二階段才對所有的詞分群。在詞的向量量化架構之下，由於訓練語料的不足，若以所有詞×所有詞的詞雙連矩陣做爲分群依據，會造成資訊過於分散，增加分群的難度。現在我們改以所有詞×500群的雙連矩陣替代，因爲這500群是由高頻詞所構成，任何其它詞的使用多少都會與高頻的共同出現，故在資訊上有其可靠性，而且可以將原本過於分散的資訊進行集中化，幫助分群程序的進行。以下則是詳細的流程：

第一階段：(對常用高頻詞分群，分群結果當做特徵，以供第二階段分群使用)

步驟1：從語料庫中選5000個高頻詞。

步驟2：統計出高頻詞×前接高頻詞的雙連矩陣。

步驟3：將高頻詞×前接高頻詞之雙連矩陣進行SVD處理，並且在可容許的誤差之下，選出k個奇異值，及對應的正交矩陣U及對角化矩陣∑，而k即代表

每個詞向量處理後的維度。

步驟 4：利用 U∑ 矩陣來進行詞的向量量化。目前是將 5000 高頻詞分爲 500 群。

第二階段：（以第一階段分群結果爲特徵，對所有的詞進行分群，分群結果主要是用來建立詞群語言模型）

步驟 1：統計出所有詞 × 500 群的雙連矩陣。

步驟 2：將所有詞 × 500 群的雙連矩陣進行 SVD 處理。

步驟 3：利用 SVD 處理所得的 U∑ 矩陣進行詞的向量量化，所欲群數隨應用而定。

在上述程序中高頻詞的數目及其所欲分成之群數，可視情況而加以調整：接下來，介紹進行分群處理時的各項實驗環境，語料庫部分是由中央研究院中文詞知識庫小組所提供 2,644,183 詞的語料庫，詞的長度只考慮到 5 字詞，欲進行分群的高頻詞 5033 目，其中單字詞 1610 目、雙字詞其 3162 目、三字詞 231 目、四字詞 30 目；而詞典中所有的詞共 85116 目，其中單字詞 14065 目、雙字詞 48454 目、三字詞字 11578 目、四字詞 10436 目，五字詞 583 目。SVD 的處理則採用 SVDPACKC 的套裝軟體 [Berry 92] 來進行。高頻詞分群處理時取 30 個奇異值，使原來詞向量的維度 5034 縮減爲 30。至於所有詞在進行分群時由於有 37596 目完全沒有統計訊息，所以先將它們歸成一群，SVD 處理後取 40 個奇異值，使詞向量的維度由 500 縮減至 40。在完成各階段分群之後，我們將兩階段的分群結果一一觀察，特別是高頻詞的分群結果，根據前接詞分群之部份高頻詞的分群結果可參見附錄一，其觀察結果請參見附錄二。

經由我們觀察的結果發現，在相當多的情況，分群的結果與文法上的詞類極爲近似，這說明了這套分群的架構對於近距離的一些語言現象有了相當不錯的掌握。前面的分群結果是依據前接詞的情況而產生的，主要是爲配合語言解碼的搜尋策略。由於單就將詞分群的觀點來看，我們也可依據後接詞的分佈加以分群，因此在相同的實驗條件下，我們也做了些實驗。其部分分群結果可參見附錄三。最後我們以文法詞類的觀點來比較依據前接詞及後接詞分群之優劣，結果請參見表 1。

詞類 / 依據	名詞	地方詞	時間詞	定詞	量詞	方位詞	代名詞	動詞	副詞	連接詞	感嘆詞
前接詞	△	△	△	○	○	○	▼	△	▼	▼	○
後接詞	△	△	△	▼	▼	▼	○	△	○	○	▼

○較好 △差不多 ▼較差

表1 分群優劣比較表

由上表我們可以發現依據前接詞分群對於定詞、量詞、方位詞、及感嘆詞的聚群能力較佳，而根據後接詞來進行分群則對代名詞、副詞及連接詞處理的較好，至於其它各種不同的文法詞類，在不同的情況下，各有其不錯的結果產生，例如依據前接詞會得到不錯的地名群，而根據後接詞，機關職稱群處理的較好，因此我們覺得文法上不同的詞類對於前文及後文會有不同的相依程度。

四. 詞群爲基礎的語言模型及國語語音辨認應用

國語語音辨認

一個大字彙的國語語音辨認系統一般可視爲由兩個核心子系統所構成(圖 2)[Lee 93a,b]，分別爲音節辨認子系統及字形確認子系統，其中音節辨認子系統主要是負責國語單音節的辨認。

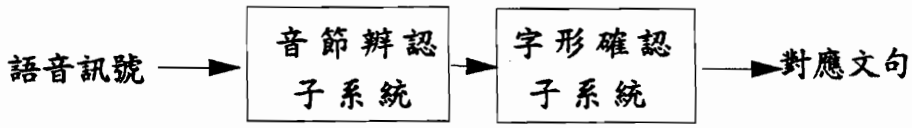

圖 2 大字集國語語音辨認系統

因爲國語單音節約有 1345 個，在實際的作法上通常分爲不計聲調的 408 個基本音節 (base syllables) 的辨認及 5 個聲調（四聲、輕聲）的辨認 [Lee 93a,b]。而這些基本音節及聲調的辨認結果，則構成若干候選音節交由字形確認子系統加以處理。所謂字形確認子系統主要的作用是要從前述候選音節中選擇出適當的對應文字。語言模型的作用即是利用前後文關係協助判斷最適合的對應文字。

詞群爲基礎之語言模型

一般所謂的語音辨認是指將所發出之語音轉換成對應文字的處理過程，通常以下列的數學模型來表示。

$$\underset{W}{\operatorname{argmax}}\, P(W|S) = \underset{W}{\operatorname{argmax}}\, P(S|W)P(W) \qquad (1)$$

$$P(S|W) = \prod_{i=1}^{n} P(s_i|w_i) \qquad (2)$$

$$P(W) \cong P(w_1|\text{boundary}) * \prod_{i=2}^{n} P(w_i|w_{i-1}) * P(\text{boundary}|w_n) \qquad (3)$$

其中 W 代表可能輸出詞串，S 代表候選音節串列，bourdary 表句首或句尾，w_i 代表第 i 個字元，s_i 代表其對應音節，式 (2) 是音節辨認子系統所得到的結果。而式 (3) 即爲詞雙連 (word bigram) 語言模型，如前所述，詞雙連語言模型可能遭遇參數量太大及參數值可信度不足的問題，而我們利用詞的分群來解決這些問

題。在進行分群後每個詞會對應到一個群，同群的詞表示其前接詞的情況極爲類似。因此我們將式(3)修正成式(4)及(5)，這就是以詞群爲基礎的語言模型。

$$P(W) \cong P(C(w_1)|boundary)P(w_1|C(w_1))*\prod_{i=2}^{n}P(C((w_i)|C(w_{i-1}))P(w_i|C(w_i)) \\ *P(boundary|C(w_n)) \quad (4)$$

$$P(w_i|w_{i-1}) \cong P(C(w_i)|C(w_{i-1}))P(w_i|C(w_i)) \quad (5)$$

以式(5)來加以解釋就是說在 w_{i-1} 之後出 w_i 的條件機率改以詞群 $C(w_{i-1})$ 之後出現詞群 $C(w_i)$ 的條件機率乘上 w_i 在其所屬之詞群 $C(w_i)$ 的出現機率。而 $P(C(w_i)|C(w_{i-1}))$ 及 $P(w_i|C(w_i))$ 的計算，則必須根據經過斷詞處理的訓練語料來加以統計。其 $P(C(w_i)|C(w_{i-1}))$ 要統計出事件 $C(w_{i-1})$ 及事件 $C(w_{i-1})C(w_i)$ 在語料庫中出現的機率如式(6)。而 $P(w_i|C(w_i))$ 則必須去統計事件 $C(w_i)$ 與 w_i 出現的機率如式(7)。

$$P(C(w_i)|C(w_{i-1}))=\frac{f(C(w_{i-1})C(w_i))}{f(C(w_{i-1}))} \quad (6)$$

$$P(w_i|C(w_i))=\frac{f(w_i)}{f(C(w_i))} \quad (7)$$

由式(6)與式(7)我們可以瞭解到只要分群後所得的詞群數遠小於詞數時，對於參數量會有相當程度的縮減。事實上，由於訓練語料的有限，在實際進行辨認工作時，仍極有可能會遭遇到 $P(C(w_i)|C(w_{i-1}))=0$ 的情況，所以仍然必須採用平滑化的技巧來加以處理。我們所使用的平滑化技巧是強化非線性內插法如[林93]; 另外格狀詞組搜尋法也一併參見[林93]。

五、初步實驗結果及分析

為了驗證以詞群(依據前接詞分群)為基礎的語言模型之效能，我們選擇詞雙連語言語模型做為基底模型(baseline model)，藉由辨認率及辨認結果去進行比較分析。首先，我們來對實驗的資源及環境進行介紹，其中包括詞典、訓練語料庫、測試文章及音節辨認結果等。

・詞典

單字詞　14065　目，雙字詞　48454　目

三字詞　11578　目，四字詞　10436　目

五字詞　583　目

・訓練語料庫

中國時報　民國80年7月份，自由時報　民國80年2月份

聯合報　民國79年12月份

三個月份報紙總計有351,026句、4,024,370字、2,664,183詞。

・測試文章

測試文章共八篇，其共包括四篇從報紙中選取的新聞(文章一至四，共2316字)，三篇選自天下雜誌的短文(文章五至七，共7764字)，及一篇短篇小說「明還」(文章八，共970字)，這些測試文章均沒有出現在訓練語料庫，也就是對語言模型進行外部測試(outside test)，以模擬實際的辨認情況。

・音節辨認結果

音節辨認子系統的辨認處理在不含聲調的音節方面，每個音節取5個候選音節，而在聲調辨認方面，每個音節選取3個候選聲調，經由組合則每個音節有15個候選音節，八篇文章的音節辨認結果以Top 1而言平均正確率是89.21%。在辨認時，將每個音節的分數併入機率評估公式中考慮。

接下來，表2我們來看實際的辨認結果。

% / 群數	測試文章								
	一	二	三	四	五	六	七	八	平均
750	89.97	93.75	82.97	86.58	90.65	91.67	82.41	80.62	87.33
1500	90.72	94.68	80.43	87.12	87.60	89.90	81.60	79.18	86.40
3000	88.47	91.57	80.16	86.85	87.60	90.51	80.76	82.16	86.01
6000	88.22	94.01	81.16	86.30	89.01	91.54	81.36	80.72	86.55
詞雙連	88.97	95.34	82.88	88.77	88.18	90.90	83.50	85.15	87.96

表 2. 不同群數之辨認率

其中測試文章一至四爲報紙報導，測試文章五至七爲天下雜誌的短文，而測試文章八爲短篇小說「明還」。表中的最後一列就是詞雙連語言模型所得到的辨認率。

我們從辨認率及實際的辨認結果去進行分析，觀察到一些現象如下。以詞群爲基礎的語言模型的辨認率略差於詞雙連語言模型的辨認率，但基本上差距不大，而從辨認的結果來看，有些時候詞群基礎的語言模型會得到較好的辨認結果，但是整體而言，詞雙連語言模型掌握得較爲精確。而不同的分群數，並未如我們所預期的分群越細而得到越高的辨認率，而是呈現波動的形式，其原因經由我們分析主要由於分群所造成的平滑化效果與分群恰當與否交互作用而造成的結果。此外訓練語料的不足亦是原因之一，所以在未來進一步的研究中將朝這幾個方向進行改進。然而以實際記憶體需求來看，詞群語言模型遠小於詞雙連語言模型，以辨認率與記憶體需求互爲取捨的原則來看，我們的詞群基礎語言模型有其實用的價值。

六 . 結論

本文即是針對統計式馬可夫語言模型之訓練語料不足及語言模型參數過多的問題，提出改善方法。我們提出一個以詞群爲基礎的語言模型來解決上述問題，它主要是配合一種把詞自動分群處理的技術，將統計資訊相似的詞歸爲一群，利用同群之詞彙分享彼此統計資訊的特性來解決前述問題。藉由此方法所得的分群結果，經由觀察發現與文法上的詞類有相當程度的吻合。且將此結果應用於語音辨認上，由實驗的辨認率來看，以詞群爲基礎的語言模型接近於詞雙連語言模型，但所需的記憶體則遠少於詞雙連語言模型。

參考文獻

[Berry 92]M.Berry,"Large-scale sparse singular value computations,",The International Journal of Supercomputer Application,",vol 6,no.1,pp.13-19,1992.

[Brown 92]P.F.Brown ,V.J.Della Pietra,D.V.deSouza,J.C.Lai,andR.L.Mercer,"Class-Based n-gram Models of Natural Language,"Computational Linguistics,vol.18,no.4,pp.467-479,1992

[Chang 93]C.H. Chang and C.D.Chen,"Automatic Clustering of Chinese Characters and Words", Proc of ROCOLING VI,pp.57-78,1993.

[趙 80]趙元任，"中國話的文法"，香港中文大學出版社 1980.

[Deerwester 90]S. Deerwester ,S.T.Dumais,G.W.Furnas,T.K.Landauer,R.Harshman,"Indexing by Latent Semantic Analysis", Journal of the American Society for Information Science ,vol 46,no.6, pp391-407.1990

[Gray 84]R.M.Gray,"Vector Quantization",IEEEASSP Magazine,pp.4-28,Apr.1984.

[Lee 93a]L.S.Lee,et.al.,"Golden Mandrain (I)-A Real -time Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary,",~~to appear in~~ IEEE Trans.on Speech and Audio Processing,vol 1,no.2,1993.

[Lee 93b]L.S.Lee,et.al.,"Golden Mandrain(II)-An Improved Single-chip Real-time Mandrain Dictation Machine For Chinese Language with Very Large Vocabulary,"ICASSP'93,pp.503-506, 1993.

[林 93]林頌堅，簡立峰，陳克健，李琳山，"國語語音辨認中詞群雙連語言模型的解碼方法"，Proc of ROCLING VI,pp.143-160.1993. [

[Ney 91]H.Ney,U.Essen,"On Smoothing Techniques for Bigram-based Natural Language Models,",ICASSP'91,pp.825-828,1991.

[Schutze 93]Hinrich Schutze, "Part-of-speech Induction from Scratch" in Proc. of ACL-93',pp. 251-258,1993.

[吳 92]吳永川，基於簡化機率模型之國語單音節辨認，國立台灣大學電機工程研究所碩士論文，中華民國八十年六月。

附錄一:依照前接詞頻率訊息分群之部份高頻詞分群結果

class 12 密集、低迷、老舊、激烈、辛苦、信任、重視、熱烈。
class 16 度、屆、季。
class 20 遍、步、票、片、頓、趟、劑、陣、系列。
class 26 噸、户、公噸。
class 28 根、顆、隻。
class 33 份、棟、套、句、座。
class 34 式、部會、階層、共和國。
class 39 局、事務所。
class 48 杯、番、顆、群、樁、聲、詞、絲。
class 52 不佳、消失、萎縮。
class 55 之上、而言。
class 58 部門、單位。
class 65 漸、趟、止、截止。
class 81 段、個、件、卷、次、位。
class 89 防範、顧及。
class 90 伯、培、佩、茂、敏、啓、秀、聰、森、耀、儀、玉。
class 99 總廠、財團、有限公司。
class 107 庭、街、城。
class 111 了。
class 114 謀、避免、防止、節省、減輕、確保。
class 122 版、點、桿、時、歲、德。
class 127 凌晨、清晨、深夜、上午。
class 129 的。
class 131 六月、十月、三月、四月、五月、元月、十二月。
class 135 撥、搬、鋪、罵、喊、動手、長大。
class 139 號、巷、日。
class 140 函、續、派員、下令、宣告。
class 160 低、烈、弱、頻繁、可觀、深刻。
class 161 難、差、高興、喜歡、容易。
class 164 篇、幅、場。
class 166 完、穩、排除、平衡、公平、可行、合理、景氣、小心。
class 180 課、顧問、辦事處。
class 181 條例、聯盟、路段、集團、中心、分公司。
class 194 偏低、很好、圓滿。
class 203 地區、條文、機構。
class 209 背景、基礎、前提、重點、主題。
class 218 滿意、清楚、愉快。
class 224 明白、知道。
class 231 吧、嗎。
class 247 不錯、多達、日益、越來越、愈來愈。
class 254 拜會、表決、停放、進出、限期、偵查、終止、審理、院會。
class 260 罷、啦、呀、年紀。
class 262 產量、營收、成長率、營業額。
class 265 報、島、區、鄉、船、電廠、會報、草案、基金會。
class 271 兵、屯、州、營、學院、中學、派出所。
class 297 把、幫、摸、提、伸、咬、聞、聽到、 想到。
class 322 不少、將近、這麼。
class 332 國、科、區公所。
class 348 表示、透露、指出、重申。
class 352 是。
class 370 廖、劉、李、陳、邵、……、將軍、……、 行政院、立法委員、紅十字會。
class 375 桃、竹、苗、洛、戈、熊、……板橋、屏東、高雄、恆春、鹿港、公所、海關、……、台北縣、台南縣、國民黨、加拿大、伊拉克、以色列、義大利、中華民國。

class 376 哩、公尺。
class 380 院、公司、協會。
class 381 覺得、承認。
class 394 八、叁、六、七、三、四、五、會計。
class 447 秒、呎、分鐘、公分、公里、周年。
class 428 很、非常、過於、較爲、極爲、有點、不可能
class 458 過來、進去、起來、出去、一下。

附錄二:依據前接詞分群之結果觀察

〔觀察 1〕class 111 了，class 129 的，class 231 吧、嗎，class352是，根據中央研究院中文詞知識庫小組的研究報告中這些詞類由於使用範圍相當的廣泛，爲了便於自然語言處理，它們須要獨立成類，我們的分群結果正符合上述說法。我們又參考趙元任先生所著之「中國話的文法」[趙 80]一書，書中指出「是」這個動詞它的造句性質相當特別，所以把它另立一類，類中只有這一個詞；而class 111、class 129、class 231是屬於語助詞。

〔觀察 2〕class 394若以一般語言學說法是數詞類，class 16、class20、class 26、class 28、class 33、class 48、class81、class 139、class 164、class 376、class 447是各式量詞，爲何這些群其結果看起來特別突出，我們認爲在文法上，定詞（包括數詞）及量詞大都可以列舉完全，因此各種定量式複合詞的組合形式是有限，且有相當的規則外，在詞序上較固定，因此這些詞類在統計上所表現出來的性質，跟其他的詞類可以較明顯地區隔，所以它的聚群性就會特別的強烈，通常不致有太多的雜訊出現。

〔觀察 3〕class 127、class 131是屬於時間詞，在一般的情況下，時間詞通常會伴隨下列的架構出現：「在…」、「到…」、「等到…」、「從…起」、「到…爲止」等。因此出現的型式也相當地固定，在這裏要特別提出說明的是class127是指一天內的各個時段，而class 131是月份，雖然同屬時間詞，但在我們所進行的分群處理，卻區隔的相當明顯，與語法上的詞類極爲吻合。

〔觀察 4〕class 107、class 265、class 271、class 332、class 375是地方詞。地方詞主要可以填至下列幾種主要的架構中「在…」、「到…」、「到…去」、「上…去」、「從…來」等等，這些地方詞包括有地名，縣市名、及國名必須與其它詞構成地方詞組的詞，如街、區、鄉、中學、區公所等。一般而言，不論是地方詞及地方詞組在所佔的造句位置並不會有太大的差異[趙 80]。

〔觀察 5〕class 34、class 39、class 180、class 181、class 380其視爲各種機構名詞的詞尾，透過適當的組合即可構成完整的機構名稱，例如「律師事務所」、「司法院」。

〔觀察 6〕class 370有相當多的姓在其中，而class 90爲名字上常用的單字詞，這些群的形成原因，我們認爲人名這種專有名詞在進行斷詞時，很難加以處理，因此在碰到人名時通常都斷爲單字詞，而造成了這樣的分群結果。

〔觀察 7〕class 58、class 89、class 209、class 224這些群的特性是同一群中的詞在使用時可以互相替代或者是其語意概念極爲近似，如果將每一詞的分群都能以此方向來進行的話，應可以解決自然語言處理中牽涉到同義詞處理之應用所遭遇到的問題。

〔觀察 8〕class 99、class 262、class 348很明顯地是由一些財經類的名詞所構成的，在這個地方給我們一個啓示，就是我可利用這樣一個特性來做爲語言模型進行領域調適(domain adaptation)的基礎，根據我們日常的經驗，可以瞭解到在一個特定的應用領域（如法律、醫藥），其詞彙的使用會比較不同於平常，會有特別常在這個領域的詞彙，而且它們通常會伴隨著一起出現，如果將它們構成一個群，我們就可以利用這種特性，在調整語言模型參數值時，整個群同時進行調整，所以調適的過程會更加有效率。

〔觀察 9〕其它與文法上所定義之詞類相當吻合的有class 297、class 458的動詞、class 161的狀態動詞、class 428的副詞等。

附錄三:依照後接詞頻率訊息分群之部份高頻詞分群結果

class 9 即日、昨天、昨日。
class 14 甚、越、愈、來得、越來越、愈來愈。
class 15 半、每、幾。
class 20 能否、率先、隨即、以便、務必。
class 27 八月、九月、七月、三月、四月、二月、一月、元月、十二月、十一月。
class 30 確是、像是、眞是。
class 34 或是、而是。
class 37 乃、必定、不論、大都、大多、絶不、往往、無論。
class 49 大型、個別、小型、外國。
class 53 博、邦、鵬、平、茂、富、德、登、田、南、郎、龍、陸、康、輝、鴻、吉、慶、欣、宅、忠、治、壽、山、聖、順、榮、財、林、松、安、雅、耀、陽、院、雲。
class 62 頗爲、極爲。
class 66 並、並未、立即。
class 73 顧問、捷運、中油、營造、瓦斯、自來水。
class 107 本來、不管、大概、多半、或許、或者、簡直、顯然、正好、實在、也許、尤其、一向、完全、一方面。
class 140 足以、不得不。
class 142 而、除了。
class 144 卻、彷彿、恐怕、好像、幾乎、向來、始終、似乎、原本。
class 166 不免、眞的、而且、有點,爲什麼。
class 176 陪、替、勸、交給。
class 178 合計、將近、計有、爲期。
class 187 僅是、只見。
class 188 幫、叫。
class 189 八、七、六。
class 191 逾、低於、高達、超過。
class 195 像、是、便是、不是、算是。
class 213 多達、長達。
class 219 成爲、做爲。
class 223 截、日、明天、當天、週六、週三、週二。
class 224 方、地院、農會、警局、及時、校方、縣府、郝柏村、環保局、環保署、檢查官、新聞局、鄉公所、鎮公所、市公所、衛生局。
class 241 大致、絶對。
class 245 晨、上午、夜間、晚上。
class 256 至於、以及。
class 260 盼、要。
class 265 比、及、以、是以。
class 269 發動、發射、提供、擬定、獲得、見到、選定、作成、引起。
class 288 它、她、牠、他們、你們、我們。
class 299 德州、宜蘭、基隆市。
class 301 明年、今年、去年。
class 317 僅、約、總計。
class 323 可、不能、可以。
class 324 千、億、萬、餘、千萬。
class 326 尚、未、不致、從未。
class 340 搞、咬、碰到、得到、聽到、看到、呈現。
class 354 美方、黨團、我們、紅十字會。
class 357 辛、官員、發言人、電信局、聯招會、國貿局、教育局、經建會、證管會。
class 369 您、你、我、當年、當時、它們、她們、自然、原來。
class 374 掀起、創下、採取、遇到。
class 378 的。
class 386 數、四、二、五。
class 399 劉、陳。
class 401 連續、共有、彩同。
class 405 還是、總是。
class 408 皆、常、仍、雖、亦、竟然、只要、除非、早已、曾經、已經。
class 416 拍、滿、打、帶、奪、聽、脫、唸、取、吸、繞、搖、想起。
class 424 李、邱、吳。
class 432 噴、倒、貼、停、掛、肯、夾、擠、住、聚集。
class 436 偏、太、較、極、朱、最、日益、最爲。
class 460 即、才、也、確實、依然、不一定。
class 462 都、到處、固然、既然、仍然、應該。
class 463 難道、如果、雖然。
class 494 能、須、曾、必須、不僅、陸續、逐漸、再次、無法。