

國立台灣大學 資訊工程學研究所 碩士論文

指導教授：李琳山 博士

基於語料庫及辭典精緻化之

中文語言模型強化之研究

Chinese Language Modeling Enhancement

by Corpora and Lexicon Refinement

研究生：張碧娟 撰

中華民國九十三年六月

誌謝

在即將畢業的時候，回頭看在語音實驗室的兩年真的覺得收穫很多。要感謝的人也很多，首先是指導教授李琳山博士，由於老師在這兩年中細心的指導，才讓我有機會在碩士階段出國兩次參加國際性會議。跟老師學習到的，不只是學問、技術本身，更重要的是老師對作研究的方法和心得以及老師的學者風範。此外，在碩士班兩年中，有幸認識了簡立峰學長並參加中研院的seminar，在這個過程中，也受到中研院許多前輩的照顧及指導，非常感謝。

另外，要感謝在語音實驗室的學長姊和同學們，不只在知識的琢磨上給我很多幫助，也帶給我碩士班兩年許多歡樂的回憶。在研究方面，要特別感謝廖碩鵬學長在我碩一的時候跟我一起合作的研究，對我來說真的幫助很大。此外，潘奕誠學長的辨識程式，對這本碩士論文也是不可或缺的重要工具。

再來，我要感謝我的室友們，包括大學時代和碩士兩年的室友們，在台大的六年中，有幸擁有像你們一樣好的室友，感覺比一般朋友更加親近且珍貴，

最後，感謝我的家人對於我繼續做研究的支持，還要感謝家興在過去五年多的陪伴和支持，若是沒有這一切情感上的支援和知識經驗上的分享及充實，我的大學及碩士班六年中，絕不可能過得像現在一樣充實愉快。謝謝大家，也衷心希望我認識的每一個人都能身體健康、一切順心。

摘要

在語音辨識的研究中，使用一個對目標語料的語句有良好估測能力的語言模型，可以有效地提高語音辨識效能。因此各語音辨識系統相關的研究中，語言模型的部分向來是辨識系統中最核心的部分之一。在各式各樣的語言模型中，N 連統計式語言模型是使用在語音辨識系統中，最為有效且成功的一種模型。然而 N 連語言模型受到其訓練或調適所使用之文字語料的影響甚鉅，因此語料庫精緻化就成為語音辨識之語言模型研究中很重要的課題之一。此外，語音辨識的語言模型中，另一個重要的議題就是辭典的取得，這個問題在中文語言處理中更顯得重要，由於中文的詞與詞之間沒有明確的分隔，因此中文抽詞、斷詞問題向來也是研究的重點之一。故本論文中主要內容將分成語料庫精緻化、辭典精緻化、以及如何將這些技術整合使用在實際的問題上，以強化語言模型並得到更好的辨識效能。

語料庫精緻化，對於 N 連語言模型使用在辨識系統中的效能有極大的影響。本論文中首先討論兩種重大的目標語料與訓練語料的不匹配性問題，接著討論如何取得適當的訓練語料庫，本文中先提出兩種來源的語料庫——既有語料庫及衍生語料庫，其中既有語料庫指的是現存容易取得的、和目標語料有較高相關性的語料庫，而衍生語料庫指的則是以基礎轉寫結果為基礎，建構適當的查詢指令，從網際網路收集得到的語料庫。在處理廣播新聞語料的實驗中，由於高品質的既有語料庫（大量匹配性高的文字新聞語料）很容易取得，因此使用既有語料庫的辨識效能很明顯勝過使用衍生語料庫時的效能。然

而，針對其他不容易取得既有語料庫的應用領域，衍生語料庫是相當重要的。另外，針對衍生語料庫中使用的查詢指令之建構，本論文提出兩種建構法——分離式三連詞查詢指令建構法、相連式高信心量度查詢指令建構法。由實驗結果所得到的結論為：使用相連式高信箱速度查詢指令建構法時，可取得量較小但品質較好的衍生語料庫。

辭典精緻化，在中文語言模型中佔有很重要的地位。本論文中首先討論傳統詞的定義，並討論使用在語音辨識應用中的詞的統計式定義，包括一個好的詞必須要高的內聚力，並且其左右文相依性必須要低（亦即有良好的左右詞邊界）。接著本文中分別就兩種不同的抽詞法——派樹抽詞法、及迭代式組合式抽詞法，進行討論。再者，由於統計式抽詞法中，其演算法中使用的參數調整是一個重要的問題，因此在本論文中分別針對詞邊度量度的相異左/右相連詞個數門檻值、及最高特定左/右相連詞比例門檻值，還有內聚力量度的重疊子片段樣式之關聯基準量、及相鄰詞內聚力量度進行實驗及討論。由實驗的結果證明，在為了語音辨識的目的抽詞時，在詞邊度量度中的最高特定左/右相連詞比例門檻值事實上是應該被捨棄的。此外，實驗結果也顯示，迭代式組合式抽詞法的效能勝過於派樹抽詞法。推測其原因，是因為迭代式組合式抽詞法是從一初始辭典開始成長，因此不但具有統計式的長處，同時也善用了初始辭典所蘊含的詞的知識。

最後，本論文將各種語言模型強化的技術，整合使用在解決兩個實際的語音辨識問題——廣播新聞語料及訪談語料。由於這兩組目標語料的特性迥異，因此在語料庫的取得及精緻化、辭典的精緻化等問題上，會遭遇到不同的問題。本論文藉著將語言模型強化使用在此兩種目標語料上的實驗，來討論先前提過各種方法的效能。實驗的結果，說明了對於廣播新聞語料來說，由於其既有語料庫容易取得，因此採用既有語料庫，加上分群分類架構並搭配上辭典精緻化，便可得到相當好的辨識率的進步率。但在訪談語料的實驗中，可以發現其不匹配性高，且既有語料庫的取得困難，因此具有比較大的挑戰，本論

文中亦作了一些初步的嘗試，使得其辨識效能有些許的增進。

目錄

誌謝	v
摘要	vii
圖目錄	xv
表目錄	xvii
1 導論	1
1.1 研究動機	1
1.2 相關研究	2
1.3 研究方向及成果	5
1.4 章節安排	6
2 理論背景與實驗環境介紹	9
2.1 大字彙連續語音辨識問題	9
2.2 N連統計式語言模型	11
2.2.1 語言模型評估量度	15
2.2.2 統計式語言模型的調適	17
2.3 實驗環境	18
2.3.1 語音辨識系統	18
2.3.2 文字語料庫	18
2.3.3 語音語料	19
2.3.4 辭典	19
3 語料庫精緻化	21
3.1 目標與訓練語料庫的不匹配	21
3.1.1 主題不匹配	22

3.1.2	時間不匹配	23
3.2	精緻語料庫的取得	24
3.2.1	既有語料庫與衍生語料庫	25
3.2.2	分離式三連詞查詢指令建構法	26
3.2.3	相連式高信心量度查詢指令建構法	27
3.3	分群分類架構	28
3.3.1	文件分群	29
3.3.2	文件分類	29
3.3.3	分群分類法之應用	30
3.4	主題匹配性之實驗結果與比較	31
3.4.1	分群分類架構的影響	32
3.4.2	既有語料庫和衍生語料庫之比較	35
3.4.3	查詢指令建構法之比較	37
3.5	時間匹配性之實驗結果與比較	38
3.5.1	以長度一個月的滑動窗進行時間重疊性分析	39
3.5.2	如何細緻選取時間匹配語料庫	42
3.6	本章結論	44
4	辭典精緻化	47
4.1	詞的定義與問題	47
4.2	派樹抽詞法	49
4.2.1	片段樣式內聚力量度	50
4.2.2	片段樣式之左右文相依性	52
4.2.3	派樹抽詞法需調整之參數	54
4.3	迭代式組合式抽詞法	54
4.3.1	相鄰詞內聚力量度	55
4.3.2	左右文變異性統計 (Context Variation Statistics)	56
4.3.3	迭代式組合式抽詞法整體架構	57
4.4	實驗結果與比較	58
4.4.1	一字詞辭典及基礎辭典實驗	58
4.4.2	片段樣式左右文相依性之二門檻值— t_f 與 t_s	59
4.4.3	片段樣式之相異相連詞個數門檻值 t_f 與重疊子片段樣式之關 聯基準量 (內聚力) 門檻值 t_{MI}	64
4.4.4	迭代式組合式抽詞法之實驗	68
4.5	本章結論	70

5	語言模型強化之整合研究	73
5.1	廣播新聞語料	73
5.1.1	廣播新聞語料之特性	73
5.1.2	同時精緻化語料庫及辭典以強化語言模型之實驗	74
5.2	訪談語料	75
5.2.1	訪談語料之特性	75
5.2.2	本論文實驗使用之訪談語料介紹	75
5.3	針對訪談語料之語言模型強化	76
5.3.1	採用新聞語料訓練的語言模型及辭典進行辨識	76
5.3.2	採用平衡語言模型進行辨識	77
6	結論與展望	79
6.1	總結與討論	79
6.2	展望	81
	參考文獻	83

圖目錄

1.1	辭典 = $\{w_i, w_j\}$ 的二連語言模型之馬可夫模型表示法	3
2.1	統計語言模型調適的架構	17
3.1	根據高信心量度在詞圖上建構查詢指令	29
3.2	分群分類架構 (Clustering-Classification framework) 用以取得同質性高的語料	31
3.3	語言模型調適架構	32
3.4	字錯誤率及次音節錯誤率-文件群數	36
3.5	字/次音節錯誤率-內插參數圖	38
3.6	語言模型訓練/調適語料與目標語料的時間關係圖(2002年)	39
3.7	時間半重疊語料集與目標語料的時間關係圖(2002年)	40
3.8	目標語料字/次音節錯誤率相對於調適語料的時間匹配性關係圖	43
4.1	詞聚力與左右相依性示意圖	51
4.2	迭代式組合式抽詞法流程	57
4.3	(t_s, t_f) 對抽取出之詞數之三維關係圖	60
4.4	(t_s, t_f) 對抽取出之詞數等高線圖	61
4.5	(MI, t_f) 對抽取出之詞數之三維關係圖	65
4.6	(MI, t_f) 對抽取出之詞數等高線圖	66
4.7	相互機率與抽出詞數關係圖	68
4.8	相互機率與抽出詞數關係圖(部分放大)	69

表目錄

3.1	主題類別和混淆度的關係	22
3.2	主題類別和混淆度上升比例的關係	23
3.3	時間匹配性和混淆度的關係	24
3.4	時間匹配性和混淆度上升比例的關係	25
3.5	字錯誤率與次音節錯誤率及相對降低比率。	33
3.6	用兩種不同查詢指令建構法, 所產生的查詢指令總數, 以及從搜尋引擎 上檢索到的衍生語料庫總字數。	37
3.7	時間不重疊、半重疊及完全重疊之字錯誤率與次音節錯誤率及和基礎辨 識結果比較之相對降低率。	40
3.8	重疊段/非重疊段之字錯誤率與次音節錯誤率及和基礎辨識結果比較之 相對降低率。	42
4.1	一字詞辭典、辨識用基礎辭典的辨識效能及辭典統計特性	59
4.2	鄰近七萬字等高線的數組(t_s, t_f)之辨識效能及辭典統計特性	62
4.3	鄰近七萬字等高線的數組(t_s, t_f)之辨識效能及辭典統計特性	63
4.4	鄰近一萬字等高線的數組(t_s, t_f)之辨識效能及辭典統計特性	63
4.5	鄰近一萬字等高線的數組(t_s, t_f)之辨識效能及辭典統計特性	64
4.6	鄰近一萬字等高線的數組(MI, t_f)之辨識效能及辭典統計特性	67
4.7	鄰近一萬字等高線的數組(MI, t_f)之辨識效能及辭典統計特性	67
4.8	迭代式組合式抽詞法及派樹抽詞法得到之辭典和基礎辨識實驗的辨識 效能比較	69
5.1	字錯誤率與次音節錯誤率及相對降低比率。	75
5.2	字錯誤率與次音節錯誤率及相對降低比率。	77
5.3	採用平衡語料庫及中文詞庫進行辨識所得之字錯誤率與次音節錯誤率 及相對降低比率。	78

第一章

導論

1.1 研究動機

語音辨識的終極的目標，是希望不管在什麼樣的收音環境下，語者講話速度快慢變化不同，不論語者談論的是什麼主題，發音有什麼變異，都可以將語音辨識出來並加以理解。在實作的層面上，聲學訊號及口音的問題需靠強健的聲學模型或是發音變異模型加以解決；而在較高階的語言或認知層面上，語言模型就扮演了很重要的角色。要能夠準確的分析口語並設計有效率的語言模型以幫助語音辨識，其實是相當困難的。應用在語音辨識上時，目前效能最好且計算上可行(computationally feasible) 的語言模型是統計式語言模型 (Statistical Language Modeling)。

採用統計式語言模型時，用來估測語言模型的文字訓練語料與目標語音語料之間的差異性，對於語音辨識的結果有重大的影響。因此，語言模型的強化是一個相當重要的議題，其目的是希望能以我們對語音語料中特定語言知識的瞭解，來補償目標語料和文字訓練語料的不匹配。更明確的說，強化語言模型的目的是希望能夠調適語言模型，使此模型不論在詞彙、文法、內容及文體方面的估測能力能更加貼近目標語料的語言機率分佈。

事實上，解決統計式語言模型強化及調適的問題，有很多種不同角度的解法。例如，從模型的基本面著手，試圖設計將複雜的語言知識和統計方法整合在一起的語言模型。

或是從語言模型調適、平滑化的演算法著手。也可從文字語料庫的選取下手，例如選取寫作時間、文體、內容和目標語料相近的文字作為語言模型的訓練語料。此外語言模型和辭典的關係是密不可分的，特別是中文語音處理中，辭典的重要性更是不可忽視。因此從辭典精緻化來達到統計式語言模型的強化，也是一個重要的議題。

本論文的研究目的是針對中文的語言特性，探討如何強化統計式語言模型，以增加大字彙語音辨識率。為達到此目的，本論文的方法著重如何在如何針對不同的目標語音語料，選取較精緻的文字語料庫及辭典。

1.2 相關研究

語言模型在自動語音辨識系統中，扮演了非常重要的角色。語言模型不只可以縮小聲學分析的空間，幫助在眾多的文字假設 (text hypotheses) 中搜尋，也可以幫助選擇一條最佳的轉寫 (transcription) 路徑。倘若不使用語言模型，僅仰賴聲學模型做音素解碼 (phoneme decoding) 的話，由於並未將其他語言知識 (如詞彙、句型、文法、語意) 考慮在裡面，所以其辨識率會比使用語言模型時差許多。

目前在語音辨識上使用最成功的統計式語言模型，應屬 N 連語言模型 (N -gram language model)。N 連語言模型只用了詞頻 (word frequency)、限制長度詞序列 (constraint length word sequence) 頻率的統計，並以馬可夫模型的方式來模擬詞 (word) 與詞歷史 (word history) 之間的關係 (如圖 1.1)。

雖然 N 連語言模型並沒有用更為高階的語言知識，如語法結構 (syntactic structure)、語意 (semantics) 等等，但其容易建立模型以及有效、快速的特性，使其成為語音辨識上語言模型的核心。羅氏 (R. Rosenfeld) 曾就統計語言模型近二十年的發展與展望，做了摘要性的陳述 [1]。

N 連語言模型固然有其優勢，但也有其缺點。使用 N 連語言模型的語音辨識，會嚴

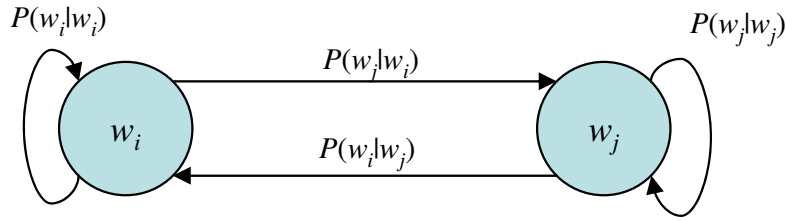


圖 1.1: 辭典 = $\{w_i, w_j\}$ 的二連語言模型之馬可夫模型表示法

重受到語音目標語料和文字訓練語料庫之間的不匹配所影響。「不匹配」有很多種型態，例如時間上 (temporal) 的不匹配、語料文體 (style) 的不匹配、語料主題 (topic) 的不匹配等等。舉例來說，如果想辨識一般日常電話中的對話，使用兩百萬字同樣為電話對話的人工轉寫語料作為文字訓練語料庫，會比使用一千四百萬字的電視或廣播新聞的人工轉寫語料來的好 [1]。

由於 N 連語言模型是語音辨識中最重要的一種語言模型，其強化與調適也一直是重要的研究議題。班氏 (J. Bellegarda) 的一篇文章中，整理了統計式語言模型調適相關的問題、並介紹以各種不同角度嘗試解決問題的方法 [2]。

語言模型調適 (adaptation) 的問題，一直以來有很多研究嘗試以不同方式解之。有些從模型的基本面改進，試圖設計將語意知識和統計方法整合在一起的語言模型調適法，例如劉氏等人 (R. Lau et al.) 所提出的詞觸發組 (word trigger pair) 調適法 [3]，是在指數模型 (exponential model) 的架構之下，以最大熵 (maximum entropy) 準則來學習語意上高度相關的詞組 (例如「醫生」和「醫院」)，以達到調適的作用。另有班氏 (J. Bellegarda) 提出以潛藏語意分析 (latent semantic analysis) 來處理詞觸發組的選取問題。潛藏語意分析是一個使用奇異值分解 (singular value decomposition) 以達到降低詞-文件矩陣 (word-document matrix) 維度的方法。除了語意知識以外，也有研究嘗試將句法知識 (syntactic knowledge) 整合到語言模型調適中。這類的調適法

使用背景的統計語言模型, 建構初始的句法模型 (syntactic modeling), 然後用調適語料來重新估測 (re-estimate) 各項參數, 以期達到在句法層面上, 從背景語言模型調適到符合欲辨識的目標語音語料。

有些則著重在以數學統計方法來增強調適法的強健性, 例如菲氏 (M. Federico) 的貝氏估測法 (Bayesian estimation) [4]、廣瀨氏等人 (K. Hirose et al.) 的最大後設機率估測法 (maximum a posteriori (MAP) estimation) [5, 6]、傑氏 (F. Jelinek) 的線性內插調適法 (linear interpolation) [7, 8], 另外還有將語言模型調適轉化成限制最佳化 (constrained optimization) 問題的解法, 例如菲氏的最小辨別資訊 [9] 調適法, 就是將調適語料的機率分佈視為新的限制, 將此限制加諸在背景語言模型上, 並由此限制將背景語言模型調適成符合最大熵 (maximum entropy) 準則的語言模型。由於這個最佳化問題在計算上相當昂貴, 因此菲氏提出一個特殊例子, 是僅使用調適語料的單詞機率分佈作為限制, 在這種情形下, 菲氏解出了一個閉鎖式解 (closed-form solution), 可以有效率的做單詞限制 (unigram constraints) 的最小辨別資訊調適。另外, 也有人從選取文字語料庫來處理調適的問題, 由於 N 連語言模型受到文字語料庫和目標語音語料不匹配的影響甚鉅, 因此根據目標語料的時間、文體、主題來選取適宜的調適語料是一個有效解決此問題的方法。調適語料的選擇, 可分為靜態的——人工選擇同主題、在對話系統中選用同一個對話狀態的語料作調適 [8]、選用同時間 [10] 的語料做為調適語料, 和動態的——例如混合模型 (mixture models)。其基本理念是將預先訓練好的調適語言模型集合中, 選出合適的模型, 與背景語言模型做線性內插 (linear interpolation), 而內插時使用的參數, 則以調適語言模型和目標語料的語言特性相近程度來計算 [11]。或以辨識結果直接做為調適語料 [12]; 或以辨識結果為查詢文字, 使用資訊檢索的方法從大量文字語料中, 擷取主題與與目標相近的文字, 做為調適語料 [13]。除此之外, 尚有以

資訊檢索的方式並搭配最小差異量 (minimum discrimination information) 調適法以選擇語料庫的研究 [14]。

此外, 好的辭典對於語言模型的優劣亦有重大的幫助, 特別在中文環境中更顯重要。中文辭典的建立, 傳統的以規則為基礎 (rule-based) 的抽詞法中, 大部分的人力花在從大量的語料庫中 取出詞或複合詞。近來, 統計方法大量的被使用, 楊氏等人提出在辭典中的詞, 其實並不一定要是傳統定義上的詞, 而可以定義為可降低訓練語料庫的整體混淆度 (perplexity) 的片段樣式 (segment patterns) [15]。同樣地, 類似的以混淆度為基礎的方法, 也被使用來作為增加、減少辭典中的詞或結合多個辭典中的詞成為複合詞的依據 [16], 除了英文的研究外, 在其他語言 (德語) 也有以結合複合詞以降低混淆度的研究 [17]。然而實務上, 要藉由混淆度的變化來找到一個最佳化的辭典, 是一件計算上很耗時的事情。因此陸續有許多的研究是採用較折衷的方法—使用其他較容易取得的統計特徵。例如, 中研院的簡立峰博士提出以派樹 (PAT tree) 為基礎的方法, 從線上文件中抽取特定領域 (domain specific) 的詞彙 [18], 利用派樹這個資料結構, 可以很容易取得兩項重要的統計特徵—正規化關聯係數 (associate norm) 和左右文相依性 (context dependency), 另有許多相關研究提出了不同的統計特徵 [19]。除了以混淆度、或以統計特徵抽詞的方法, 亦有研究是針對辭典、斷詞、和語言模型作整合性的最佳化, 高氏等人提出辭典的選取和斷詞在中文統計式語言模型中是隱藏程序 (hidden process) 的概念, 以期望值最大 (EM) 演算法來同時最佳化此隱藏程序和語言模型 [20]。

1.3 研究方向及成果

本論文的研究目的是強化中文語言模型, 使其應用在語音辨識系統上的效能達到更好。研究的方向, 是將中文語言模型強化的問題中兩個非常重要的部分 (語料庫及辭典) 分別拿出來討論。

在語料庫精緻化的研究方面，本論文中分析目標語音語料與訓練文字語料的不匹配，對於語言模型造成的影響，特別針對廣播新聞辨識的問題，分析兩種重要的匹配性問題（時間不匹配及主題不匹配）對辨識效能的影響。為了解決這些不匹配的問題，本論文提出精緻語料庫的取得之方法，包括分別取用了兩種不同特性的語料庫（既有語料庫與衍生語料庫）來進行語言模型強化的研究，特別在衍生語料庫方面的研究，本文提出從基礎轉寫結果，建構查詢指令，以檢索網際網路上的資源作為衍生語料庫。在查詢指令的建構法研究中，本篇論文提出了兩種方法——分離式三連詞查詢指令建構法，及相連式高信心量度查詢指令建構法，並以實驗分析兩種查詢指令建構法所取得的衍生語料庫之品質。

在辭典精緻化的研究中，本論文針對兩種不同風格的統計式抽詞法：派樹抽詞法，以及迭代式組合式抽詞法進行討論與分析。由於這兩個抽詞法都是統計式的，因此在本文中討論了其對「詞」定義使用的統計特徵，又由於統計式抽詞法中，對各個統計量所設定的門檻，對於整體的抽詞表現會有相當大的影響，因此本文中特別針對派樹抽詞法及迭代式組合式抽詞法中，所需設定的門檻值參數及如何調整這些參數作了一系列的實驗分析及討論。

1.4 章節安排

本論文的章節安排如下：

- 第二章 簡述統計式語言模型的背景知識及過去的研究，並著重在應用於語音辨識的語言模型的問題。接著簡介在大字彙連續語音辨識架構中的統計式語言模型。最後描述本論文實驗所使用的大字彙語音辨識程式以及實驗環境與語料。
- 第三章 討論語料庫精緻化的動機、遭遇到的不匹配問題，並針對兩種重要的匹配

性進行語料庫的精緻化。此外，本章中並分析兩種不同的外部語料庫所造成的辨識效能影響，並且針對衍生語料庫的取得方法，提出兩種查詢指令建構的方法。最後，提出一個整合了文件分群、分類方法的架構，用以得到更為精細的訓練/調適語料庫。

- 第四章 定義辭典精緻化對語言模型強化的重要性，並介紹此研究的定義及遭遇的問題。討論兩種不同的統計式抽詞法，並針對其對語音辨識系統的幫助，以及其演算法中參數的調校進行相關實驗及分析。
- 第五章 整合之前各章提過的語言模型強化之方法，實際處理兩種不同特性的目標語音辨識語料 — 廣播新聞語料及訪談語料。從最基礎的實驗開始，一步步討論如何採用之前討論過的各種技術，來強化語言模型，以達到更好的辨識結果。
- 第六章 總結與未來展望。

第二章

理論背景與實驗環境介紹

本章將簡述統計式語言模型的歷史及其相關研究課題，並著重在應用於語音辨識的語言模型上懸而未解的問題。在 2.1 節，對在大字彙連續語音辨識架構中的統計式語言模型做一個整體簡介，並介紹最常使用的 N 連語言模型及調適問題。接著在 3.4 節，描述本論文實驗所使用的大詞彙語音辨識程式以及實驗環境與語料。

2.1 大字彙連續語音辨識問題

自動語音辨識 (automatic speech recognition) 面對的問題，是必須同時在很多不同層次上進行解歧異的工作。特別是在大字彙連續語音辨識 (large vocabulary continuous speech recognition, LVCSR) 的問題上，由於我們無法不使用更高階的知識就達到正確地辨識出所有的音素或是字詞，因此在語音辨識的搜尋架構中，必須同時存在許多的音素/詞/句的可能性 (或稱「假設」, hypotheses)。然後在這樣巨大的搜尋空間中，再將其他更高階的知識—例如句法、語意、語用—加諸其上，以統計方法決定如何能達到最好的辨識正確率。

大字彙連續語音辨識問題，可以用以下的敘述表之：

$$\mathbf{W}^* = \arg \max_{w_1^n} P(w_1^n | x_1^T) \quad (2.1)$$

$$= \arg \max_{w_1^n} \frac{P(w_1^n)P(x_1^T | w_1^n)}{P(x_1^T)} \quad (2.2)$$

$$= \arg \max_{w_1^n} P(w_1^n)P(x_1^T | w_1^n) \quad (2.3)$$

$$= \arg \max_{w_1^n} P(\mathbf{W})P(\mathbf{X} | \mathbf{W}) \quad (2.4)$$

其中 $\mathbf{X} = x_1^T = x_1 \dots x_t \dots x_T$ 是時間 $t = 1 \dots T$ 觀測到的聲學特徵向量串；
 $\mathbf{W} = w_1^n = w_1 \dots w_n$ 代表長度為 n 的可能詞串；而 \mathbf{W}^* 是辨識結果，即為
 在所有可能的詞串中，給定觀測到的聲學特徵向量串條件之下，得到最高的
 條件機率的詞串。

$P(w_1^n | x_1^T)$ 代表語音辨識器辨識時希望能最大化的機率。也就是，語音辨識器將觀測到的聲音資訊轉成聲學特徵向量串 \mathbf{X} 之後，給定這樣的聲學特徵向量串的條件之下，語音辨識器希望能從所有可能的詞串 \mathbf{W} 中，找到最有可能的(條件機率最大)的詞串 \mathbf{W}^* 為語音辨識器的辨識輸出。從式(2.1)推導到至式(2.2)是一個貝氏轉換，由於分母部分是觀測到的聲學特徵向量串的機率，不會影響最大化的結果，因此最後將大字彙語音辨識問題推導至式(2.4)時，語音辨識架構中最重要的兩個模型就顯而易見了。 $P(\mathbf{W})$ 是本篇中所要討論的語言模型，其模型的目的是模擬出語者說出某一詞串 \mathbf{W} 的事前機率(prior probability)。建構語言模型的方法有很多種，可以從複雜精細的人類認知思考、語言產生的角度來建構；亦可忽略語言產生的過程，直接使用統計方法來描述產生的語言本身，用以估算各個詞串的出現機率，這就是統計式語言模型的概念。其中最直覺且容易建構的一種統計式語言模型就是 N 連語言模型，將在以下做進一步的介紹。

2.2 N 連統計式語言模型

在我們的生活經驗中，當我們聽到別人說了一些話但尚未說完時，常常可以大概猜到別人接下去會說什麼，雖然也許無法完全預測的準（事實上也不需要百分之百精準），但因為有這些預測，已經可以將對方接下去會說的話的可能性限制到一個比較合理的範圍。將類似的想法應用到以電腦解決大字彙連續語音辨識的問題上，在2.1中將問題解構為式(2.1)，可以看到在大字彙連續語音辨識程式中，這樣的限制和預測能力也扮演非常重要的角色。如前所述，語言模型可以描述成詞串 \mathbf{W} 的機率分佈 $P(\mathbf{W})$ ，此分佈可以反映出任一詞串 \mathbf{W} 在句子中出現的頻率。

有許多方法可以將機率分佈適當分配到各個可能詞串，以得到好的統計語言模型。最簡單的一種機率模型是假設語言中任何一個詞可能接在任何一個詞後面，在這種機率假設之下，無論是 w_j 出現在 w_i 之後，或是 w_k 之現在 w_j 之後，都分配以相等的機率。假設辭典中含有七萬詞，則任一詞出現在任一詞後面的機率均為 $\frac{1}{70000}$ 。

另外一個稍微複雜一點的詞串機率分配方法是使用相對頻率 (relative frequencies)，任何詞可以出現在任何詞的後面，但是這個機率是依詞在訓練語料庫中出現的相對頻率來估測。以雅虎奇摩中文新聞2003年整年的語料庫 (使用單字詞辭典) 為例，「的」出現3,553,330次，所有的詞數是148,941,101個單字詞，也就是說「的」這個頻率非常高的單字詞，在整個語料庫的相對頻率約為2.39%，另一個例子是「花」這個單字詞僅出現89,070次，相對頻率為0.06%，因此我們可以根據這些頻率來猜測下一個出現的詞的可能性。

更進一步，加上條件機率 (conditional probability) 來分配詞串的機率分佈會更為精準。舉例來說，如果是在一句話的最前面，並且不考慮和前一句話的關係的話，我們可以合理的假設用2.39%的機率來預估「的」，並以0.06%的機率預估「花」；但如果我們

知道前面出現的詞串為「蝴蝶採」的話，在這樣的情形之下，後面出現「花」看起來應該比出現「的」來的合理。這樣的觀察告訴我們，如果在給定前面的詞串歷史的條件之下，用條件機率來來預估一個詞出現出現的機率，會比單純使用各別詞的頻率來的精準。從以上這些語料庫、詞、詞串等關係的觀察，再回頭過來想詞串 \mathbf{W} 的機率分佈 $P(\mathbf{W})$ 應該如何估測。

如果將「每個詞出現在對的位置」這件事視為獨立事件，我們可以將 $P(\mathbf{W})$ 表示成爲 $P(w_1, w_2, \dots, w_{n-1}, w_n)$ ，套上連鎖律後，我們可以得到：

$$\begin{aligned} P(\mathbf{W}) &= P(w_1, w_2, \dots, w_{n-1}, w_n) \\ &= P(w_1)P(w_2|w_1)P(w_3|w_1^2)\dots P(w_n|w_1^{n-1}) \\ &= \prod_{i=1}^n p(w_i|w_1^{i-1}) \end{aligned} \quad (2.5)$$

其中 $P(w_k|w_1^{i-1})$ 是代表給定之前出現的詞歷史 (word history) 是 w_1, w_2, \dots, w_{i-1} 時， w_i 會接著出現的條件機率。

在式 (2.5) 中，假設使用的辭典 \mathcal{V} (大小爲 $|\mathcal{V}|$)，則有 $|\mathcal{V}|^{i-1}$ 種不同可能的詞歷史，爲了要能完整的表示 $P(w_i|w_1^{i-1})$ ，就必須要用上 $|\mathcal{V}|^i$ 個不同的值。然而事實上要能準確的預估 $P(w_i|w_1^{i-1})$ 並不是一件很容易的事，給定所有的詞串歷史，要估測下一個詞的出線機率，目前仍然沒有很好的方法，由於詞串歷史 w_1^{i-1} 的可能性非常多，用語料庫很難獲得一個穩定的統計量，因此爲了解決這個問題，前人做了若干種簡化，在詞串歷史上加諸一些合理的限制來得到原機率公式的近似值。一個最簡單的限制就是假設任一個詞出現的機率，僅和前一個出現的詞有關，這就是所謂的二連模型 (bigram model)，二連模型以 $P(w_n|w_{n-1})$ 來近似估測「給定所有詞串歷史來估測詞出現的條

件機率 $P(w_n|w_1^{n-1})$ 」。換句話說，本來應該估測 $P(\text{花}|\text{蝴蝶採})$ 的機率，我們假設「花」只和「採」有關，因此以 $P(\text{花}|\text{採})$ 來近似估測這個機率。

以上提到的「任何詞出現的機率，僅和前一個出現的詞有關」的假設，稱為馬可夫假設。馬可夫模型是一種特殊的機率模型，它假設在序列中，要預測未來會出現的東西時，不需要看太多過去出現的東西。第一章中提到的例子(圖 1.1)，是一個表示成馬可夫鍊 (Markov chain) 的示意圖，馬可夫鍊事實上是一種加權有限狀態自動機 (weighted finite-state automaton)，在這裡「馬可夫」的意思是源於：加權有限狀態自動機的下一個狀態，僅和有限的歷史有關 (因為有限狀態機中的狀態個數是有限的)。

我們可以將二連模型 (只往前看一個詞) 的概念延伸到三連模型 (往前看兩個詞)，甚至延伸到 N 連語言模型 (往前看 N-1 個詞)。從馬可夫模型的角度來看，二連模型等同於一階馬可夫模型 (first-order Markov model)，意即它指往過去看一個詞；三連語言模型則稱為二階馬可夫模型 (second-order Markov model)，表示下一個狀態和過去的前兩個狀態有關；而 N 連模型就是 N-1 階馬可夫模型。N 連語言模型的一般式，就是將預測下一個詞的條件機率變成一個近似的估測：

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1}) \quad (2.6)$$

式(2.6)是預測 w_n 出現的機率，在給定前面 $N - 1$ 個字的情況之下。將式(2.6)代入式(2.5)之後，可得到用二連語言模型計算整個詞串的機率估測：

$$P(w_1^n) \approx \prod_{k=1}^n P(w_k|w_{k-1}) \quad (2.7)$$

舉例來說，「蝴蝶採花蜜」的機率即為：

$$P(\text{蝴蝶採花蜜}) = P(\text{蝴} | < s >) P(\text{蝶} | \text{蝴}) P(\text{採} | \text{蝶}) P(\text{花} | \text{採}) P(\text{蜜} | \text{花})$$

$\langle s \rangle$ 代表句子的開始。

N連語言模型可藉由計數和正規化 (normalize) 來訓練, 正規化指的是在機率模型中, 機率值必須落在0和1之間。N連語言模型的機率估測, 可計算任一個 N 連詞串出現在語料庫中的機率, 然後除以第一個詞相同的 N 連詞串來加以正規化:

$$P(w_n | w_{n-1}^{n+N-1}) = \frac{C(w_{n-N+1}^{n-1} w_n)}{\sum_w C(w_{n-N+1}^{n-1} w)} = \frac{C(w_{n-N+1}^{n-1} w_n)}{C(w_{n-N+1}^{n-1})} \quad (2.8)$$

在式(2.8)中, 以某特定觀測到的詞串頻率除以詞串前綴 (prefix) 的頻率, 來估測 N 連語言模型機率, 這個比例就是我們之前提到的相對頻率 (relative frequency), 其目的其實就是達到最大相似度估測 (maximum likelihood estimation, MLE), 因為使用相對頻率作為機率模型 L 時, 可以使給定的訓練語料庫 T 的相似度 $P(T|L)$ 達到最大值。

最常用的 N連語言模型是三連語言模型 (trigram LM), 以 $N = 3$ 代入式(2.5)可以得到:

$$P(w_1^n) \approx \prod_{k=1}^n P(w_k | w_{k-2}, w_{k-1}) \quad (2.9)$$

但是, 即使將詞的關連性 (dependency) 簡化到三連語言模型, 還是有資料稀疏 (data sparseness) 的問題。未見事件 (unseen events) 和次數過低的詞串, 都會造成語言模型的估測失準。因此, 語言模型的平滑化是一個基本而重要的議題。統計模型平滑化的問題, 從早期研究生物統計的文獻就有許多基礎而重要的方法 [21], 針對語言統計模型的平滑化, 重要的研究包括凱氏 (Katz) 平滑化法 [22] 及尼氏 (Kneser-Ney) 平滑化法 [23]。

本論文實驗部分用的 N連語言模型均以古德—圖靈後撤平滑法 (Good-Turing back-off smoothing) [21] 處理。N連語言模型的概念雖然並不複雜, 但由於其容易訓練, 易於

使用在由左至右 (left-to-right) 的即時語音辨識 (real-time speech recognition) 應用中, 因此目前仍是語音辨識程式中的主流語言模型, 其中又以三連語言模型最為常用。

2.2.1 語言模型評估量度

研究語言模型時, 一個重要的問題是: 如何量化這個模型的優劣? 評估一個語言模型最重要而終極的方法, 就是看這個語言模型使用在對應的應用上, 對該應用造成的效能影響。由於本論文的應用是在語音辨識中, 我們主要將以辨識率作為評估語言模型優劣的基準; 另外, 由於本論文討論重點在中文語音辨識, 因此將以字錯誤率(character error rate, CER) 為主要的評估基準, 以排除斷詞歧異造成對詞錯誤率 (word error rate, WER) 的影響。字正確率的計算方法為: 以字為單位, 使用動態規劃校準法將人工轉寫答案與自動語音辨識程式轉寫結果對比。

$$\text{正確率 (Accuracy)} = \frac{N - D - S - I}{N} \times 100\% \quad (2.10)$$

N : 總字數;

D : 遺失型錯誤 (deletion errors);

S : 取代型錯誤 (substitution errors);

I : 插入型錯誤 (insertion errors)。

由於語音辨識的字錯誤率的原因較為複雜, 通常和語音辨識器中的各個元件均有關聯, 而且語音辨識的過程較耗時, 因此在過去的研究中, 若要調整語言模型的參數, 大部分不直接使用語音辨識的字錯誤率, 而是採用另一種替代的評估量度—混淆度 (perplexity)。混淆度常用於評估一語言模型 L 預測之前未曾看過的文字語料 T 的預測能力。這個預測能力, 可定義成 L 產生 T 的對數相似性 (log-likelihood), 亦即文字語料的機率

分佈 $P_T(x)$ 與模型的機率分佈 $P_L(x)$ 的交叉熵 (cross-entropy)。直覺上來說，交叉熵可想像成以語言模型 L 為一個人的背景知識去瞭解 T 的內容，這個過程中的熵 (或稱亂度, entropy)。更簡明的說，就是以 L 為背景知識時，閱讀 T 時的預期能力，亂度越低表示理解能力越好。以數學式表之：

$$H(P_T; P_L) = - \sum_x P_T(x) \times \log P_L(x) \quad (2.11)$$

混淆度的定義為：

$$PP_L(T) = 2^{H(P_T; P_L)} \quad (2.12)$$

混淆度可解讀為語言的字詞之間的分歧係數 (branchout factor) 的幾何平均數：混淆度為 X 的語言的複雜程度，可視為每個詞後面有 X 個不同可能的詞，如果假設每個詞出現的機率是一樣的。由式 (2.12) 可知，混淆度同時是語言模型和文字語料的函數，因此一旦替換不同的文字語料、不同的語言模型，都會影響混淆度的大小及其解讀法。因此，若要以混淆度為量度，來比較多個語言模型的效能的話，必須用同樣的文字語料和辭典才有意義。使用不同辭典時的混淆度，目前仍沒有較有意義的解讀法。辭典若不一樣，較小的辭典會使模型錯誤的偏向較小的混淆度。若使用同樣的辭典但不同文字語料，可能遭遇的問題是辭典和文字語料的不匹配，會造成不公平，舉例來說，以偏向體育領域的辭典，用在體育的文字語料和政治的文字語料上，在政治語料上的辭典外詞彙 (out-of-vocabulary words) 數量會大增。

在本論文中，在討論目標語料及辭典、訓練文字語料庫及語言模型的匹配程度時，會使用混淆度作為說明時參考的指標。由於本論文主要探討辭典和文字語料精緻化對於統計語言模型使用在語音辨識上的影響，因此用以衡量語言模型優劣的量度，主要仍為語言模型套用在語音辨識程式中的辨識結果的字錯誤率。

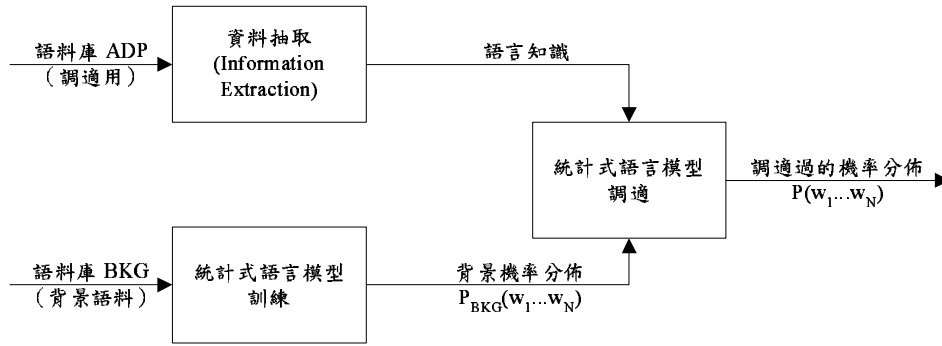


圖 2.1: 統計語言模型調適的架構

2.2.2 統計式語言模型的調適

統計語言模型調適的架構一般可以表示為圖 2.1。考慮兩個文字語料庫：一個是（較小的）調適語料庫 ADP，這個語料庫應選取和目前欲辨識的目標語料相關的；另外一個是（較大的）背景語料庫 BKG，這個語料庫可能包含比較過時，或是主題及文體和目標語料略有出入的語料。

統計語言模型調適的問題，目的是希望針對給定任一個包含 N 個詞的詞串 w_1, w_1, \dots, w_N ，都可以得到夠強健的語言模型機率估測：

$$P(w_1, \dots, w_N) = \prod_{k=1}^N P(w_k | h_k) \quad (2.13)$$

h_k 代表在時間點 k 時已存在的歷史詞串。

以 N 連語言模型的例子來說，由於其馬可夫模型（圖 1.1）的假設，可得到：

$$h_k = w_{k-n+1}, \dots, w_{k-1} \quad (2.14)$$

原本的式 (2.13) 是一個等式，但是由於在 N 連語言模型的假設下喪失了一些可能有用的資訊，因此若採用式 (2.14) 做為歷史詞串，式 (2.13) 便成為為一個近似值。在調適語言模型中，對任一可能詞串而言， $P(w_1, \dots, w_N)$ 的估測事實上來自兩個不同的知識來源：

1. 背景語料庫 BKG: 語料來源較廣, 但可能和目標語料不匹配。以 BKG 所估測的統計式語言模型機率分佈 $P_{BKG}(w_1, \dots, w_N)$ 。
2. 調適語料庫 ADP: 用來抽取出和目標語料相關的資訊, 其機率分佈不若 BKG 那樣平滑, 但包含比較精準的資訊。

針對 ADP 的取得和精緻化, 有許多不同的角度來做, 例如快取模型 (cache model)、取用主題 (topic) 相關語料等等, 其基本概念是希望以從 ADP 抽取出的資訊, 來動態修正背景統計語言模型的估測。

2.3 實驗環境

2.3.1 語音辨識系統

本論文實驗採用的語音辨識器為台大數位語音處理實驗室 2002 年所發展的: 大字彙中文連續一段式語音辨識系統 [24]。採用的聲學模型包含 151 個次音節模型 (Initial-Final sub-syllabic models), 其中有 112 個聲母 (initials), 38 個韻母 (finals) 和一個靜音模型。每個韻母由四個狀態組成, 聲母由三個狀態 (state) 組成, 但空聲母僅有兩個狀態。每一個狀態為十六高斯合成分佈 (16 Gaussian mixtures)。

2.3.2 文字語料庫

在本論文中, 文字語料庫的主要用途是做統計模型的訓練和調適, 以及使用文字語料庫來抽取新詞或刪去不適宜的詞, 以達到辭典精緻化。本論文實驗使用的文字語料有以下兩類:

- 基礎語料 (baseline corpus):

用來訓練基礎語言模型的語料, 大小為四千萬個中文單字, 來源是中央社 1997 至 1999 年的部分新聞 [25], 此基礎語言模型在本論文中作為線性內插調適時的背景語言模型。

- 調適及分析用語料(adaptation corpus):

用來做第三章及第四章的語料庫及辭典的精緻化相關實驗。來源是雅虎奇摩新聞網所整理的台灣各大新聞社的新聞[26]，雅虎奇摩合作的新聞媒體包括中時電子報、聯合新聞網、中央社、中廣新聞網、TVBS 新聞、民視、大成報、路透社、麗台運動報、中央日報、台灣日報，並將新聞分成十二類，以利讀者閱覽，其類別包含：政治、社會、國際、兩岸、財經、影視、體育、生活、休閒、科技、健康、新奇等十二類。雅虎奇摩新聞分類的原則並非十分嚴謹，且新聞來源較廣雜，因此在第三章中討論到用分類語料訓練模型時，主要仍採用自動分群的類別而不採用雅虎奇摩所分的十二類。本論文第三章前半分析部分是採用2003年的新聞，共148,941,101字,347,055則新聞；論文中討論主題及時間匹配性時，選用較小且和語音語料時間有重疊的語料庫，則是採用2002年8,9月的新聞，共58,663則,25,279,988字。

2.3.3 語音語料

語音語料均為測試語料，或稱目標語料，來源是台灣本地的新聞廣播電台 News98[27]，時間是2002年9月2-4, 9-13, 17-20, 23-27, 30號，共計收集到18天的廣播，共有506則，長度為3.7小時。

本論文中實驗的最終目的，是希望能藉由語料庫或辭典精緻化，使得統計語言模型更貼近欲解決的目標語料，而得到更好的辨識結果。將目標語料辨識到最好，是我們的目的，在這個過程中，可能會藉由多次迭代其辨識結果而達到精緻化語料庫的目標，因此語音語料在實驗中也扮演著資訊來源的功能。

2.3.4 辭典

本論文使用的辭典有二：

1. 辨識用基礎辭典 (baseline lexicon for speech recognition):

台大語音實驗室所開發的大字彙中文連續一段式語音辨識系統，使用的基礎辭典包含 61,522 個詞，詞長度最短為一字詞，最長為六字詞，平均詞長為 2.34 個字，詞長之標準差為 0.87。此辭典是由新聞語料庫中得來，因此其特性較貼近一般的新聞語料，因此在廣播新聞語音辨識上可達到不錯的精準度，又由於選用的詞彙時間相關性較低，因此被使用作為辨識用基礎辭典。

2. 一般語言處理(包括標音等) 使用之辭典:

除了上述辨識用基礎辭典以外，本論文中標音及部分實驗的抽、斷詞時，採用中華民國計算語言學學會的中文詞庫(Chinese Electronic Dictionary, CED) [28] 為基礎。中文詞庫由鄭良偉教授及畢氏 (Roger Bissonnett) 所整理發展，為一包含八萬詞的電子辭典。本詞庫收的詞包含一般用詞、常用專有名詞、成語、慣用語、常用派生詞 (derivative)、異體詞、合併詞 (或稱複合詞 (compound word)) 以及少數特殊領域用語和古漢語詞語。每個詞項包含的訊息有：注音、頻率、詞類、名詞語義分類等。本實驗所採用的詞，為詞頻大於一的詞，計有 54,475 詞，用到的辭典資訊，有詞、詞頻、注音 (所有詞均只有一種注音)。

為了在語言模型訓練時，能夠處理在基礎辭典中未出現的字詞以及未見字詞的標音，本實驗採用在 UNIX 系統下中文注音輸入法的表格，以此做為所有中文單字詞的注音定義 (含破音字)。中文多字新詞的標音方式也是一個有趣的課題，本論文中採用的是直接展開法，即直接將一個詞中所有可能不同的破音字展開。

第三章

語料庫精緻化

統計式語言模型的相關研究中，有許多不同的角度來解決相關的問題，包括從數學模型改進模型的效能、改進語言模型調適法等等。事實上，N 連統計式語言模型使用在語音辨識中的效能，和文字訓練調適語料與目標語音語料之間的關係，有很高度的相關性。因此，語言模型強化這個問題，可以從語料庫精緻化的角度切入。本章中將討論兩類重要的匹配性、分析兩種不同來源的調適語料庫對於辨識實驗的影響，並且提出一個整合兩種機器學習技術的語料庫精緻化架構。

3.1 目標與訓練語料庫的不匹配

由於三連語言模型僅紀錄了每個詞前面長度為二的詞歷史，可以說是視野非常窄小的一種模型。也因為這樣，三連語言模型受到訓練語料和目標語料的不匹配影響甚鉅。選取適宜的文字語料庫來估測或調適三連語言模型，可以得到較為匹配的機率分佈。在語言模型與目標語料匹配的狀況之下，語言模型預測的能力會較為精準。

而語言行為上的匹配與否，又可以從許多層面來說，例如寫作時間、文體、主題等等。本論文中實驗的語料均採用新聞文章，雖然不同媒體、不同記者所撰寫仍有不同，但由於均為新聞，因此文體方面沒有太大的差異度。

本節中將先針對主題不匹配和時間不匹配兩個問題作基本的分析，接下來討論兩種

訓練\測試	影視	社會	國際	政治	健康	財經	體育	All	1/7
影視	54.13	145.57	120.54	121.83	101.76	156.81	213.30	116.13	150.79
社會	175.84	46.27	104.73	59.23	69.07	119.39	240.67	97.06	185.98
國際	226.68	137.39	37.14	85.31	96.59	117.07	263.45	112.53	229.92
政治	231.42	52.51	84.61	36.13	82.64	105.28	302.88	103.49	233.97
健康	261.68	129.08	119.42	103.80	34.23	154.23	429.37	133.36	263.33
財經	300.93	161.46	106.47	93.43	115.33	33.17	322.29	125.76	308.00
體育	244.77	223.81	150.57	161.65	185.43	212.74	30.36	144.52	248.44
All	48.96	37.60	32.62	30.49	30.93	30.85	29.95	34.00	38.28
1/7	249.28	148.86	123.38	124.85	104.18	158.75	217.53	473.79	39.42

表 3.1: 主題類別和混淆度的關係

不同的調適語料庫來源, 並以實驗分析其異同處。

3.1.1 主題不匹配

究竟三連語言模型受到主題不匹配的影響為何, 可用以下一個小實驗來說明。這個實驗的文字語料部分, 選用雅虎奇摩新聞網 [26] 2003 年的七類新聞, 包括影視、社會、國際、政治、健康、財經及體育, 每個主題均選取約六十五萬字。除了此七組文字語料以外, 另外用一組將七個主題全部混合的語料 (在表中稱為「All」), 以及一組將七個主題混合, 但隨機選取使大小和原本一樣的語料 (在表中稱為「1/7」)。

在辭典的選用上, 爲了避免辭典對於領域的偏差影響, 採用單字詞辭典, 即所有辭典裡的詞彙均爲中文單字 (character)。以亂數選取訓練領域語料的 90% 做爲訓練語料, 訓練領域語料的 10% 做爲測試語料, 並重複二十次交叉測試求其平均的方式來觀察: 當選用某一主題的語料來訓練語言模型時, 將此語言模型使用在預測另一主題的語料時, 其混淆度的變化。

表 3.1 是使用七個主題的語料, 以亂數選取訓練領域語料的 90% 訓練、測試領域語料的 10% 測試, 並重複二十次交叉測試求其平均的方式, 計算交叉混淆度 (cross-perplexity)。爲了避免辭典造成的偏差, 一律使用單字辭典。表 3.2 則是將表 3.1 中每一行除以用同樣語料庫 (例如: 訓練=影視; 測試=影視) 所造成的混淆度。此表可以更

訓練\測試	影視	社會	國際	政治	健康	財經	體育	All	1/7
影視	1.00	2.69	2.23	2.25	1.88	2.90	3.94	2.15	2.79
社會	3.80	1.00	2.26	1.28	1.49	2.58	5.20	2.10	4.02
國際	6.10	3.70	1.00	2.30	2.60	3.15	7.09	3.03	2.69
政治	6.41	1.45	2.34	1.00	2.29	2.91	8.38	2.86	6.48
健康	7.65	3.77	3.49	3.03	1.00	4.51	12.55	3.90	7.69
財經	9.07	4.87	3.21	2.82	3.48	1.00	9.72	3.79	9.28
體育	8.06	7.37	4.96	5.32	6.11	7.01	1.00	4.76	8.18
All	1.44	1.11	0.96	0.90	0.91	0.91	0.88	1.00	1.13
1/7	6.32	3.78	3.13	3.17	2.64	4.03	5.52	12.02	1.00

表 3.2: 主題類別和混淆度上升比例的關係

容易看出訓練和預測的語料不匹配時，混淆度上升的情形。從表 3.1 和表 3.2 可以看出，當訓練和測試的語料是同一主題時，其混淆度比起用語言模型來預測不同主題的語料時低了許多。從這兩個表中，另外也可以觀察出一些現象。例如，體育類新聞和其他類新聞的文字語言行為差異較大，因此當試圖用體育訓練的三連語言模型來預測其他主題的語言行為時，其混淆度增加的幅度比其他交叉混淆度來的高上許多。此外，從這兩個表中可以看出，有些類別其實相當相近（例如社會和政治）。分析其原因，一方面是因為有些類別之間原本就有較大的相關度，因此會呈現這樣的差異，但也由於雅虎奇摩是一個收集多方來源的新聞語料庫（在 3.4 節中會再詳細描述之），不同媒體的分類方式或有不同，再加上其分類原則是讓新聞的讀者方便閱讀，並不必然符合機器學習時分類的原則。儘管如此，從表 3.1 和表 3.2 仍可窺見主題不匹配對三連語言模型訓練的影響。

3.1.2 時間不匹配

除了主題的不匹配以外，在新聞語料的處理中，時間的匹配性是不容忽視的。由於新聞語料中，很容易出現新的類專有名詞 (named entity)、新的用語，因此即使只是差距數天或數個月的時間不匹配，也可能造成三連語言模型預測語言行為的能力降低。

以下實驗選用雅虎奇摩新聞網 2003 年整年的新聞，以兩個月為一個單位，共有六個

訓練\測試	一、二月	三、四月	五、六月	七、八月	九、十月	十一、二月
一、二月	31.15	53.18	59.73	61.19	59.51	59.86
三、四月	54.65	30.23	50.50	59.75	58.88	59.73
五、六月	57.95	49.89	29.95	56.51	57.76	59.16
七、八月	55.79	51.93	50.28	32.38	52.64	55.38
九、十月	56.18	53.06	53.09	53.78	32.17	53.00
十一、二月	56.31	53.42	54.40	57.14	52.92	32.20

表 3.3: 時間匹配性和混淆度的關係

資料集，並爲了避免辭典對於不同時間語料的偏差影響，採用單字詞辭典，即所有辭典裡的詞彙均爲中文單字。以亂數選取訓練月份語料的90%做爲訓練語料，測試月份語料的10%做爲測試語料，並重複二十次交叉測試求其平均的方式來觀察：選用某一時段的語料訓練三連語言模型，並將此語言模型使用在預測另一時段的語料時，其混淆度的變化，特別是在時間差長短不同時，混淆度是否有明顯的差異。

表3.3是六組時間不重疊的資料集所產生的交叉混淆度，實驗的設定是將2003整年的語料分成六份，以亂數選取訓練月份語料的90%訓練、測試月份語料的10%測試，並重複二十次交叉測試求其平均的方式，計算交叉混淆度。且爲了避免辭典造成的偏差，一律使用單字辭典。而表3.4是將表3.3中每一行除以用同樣時段（例如：訓練=測試=一、二月）所造成的混淆度。此表可以更容易看出訓練和預測的語料在時間上不匹配時，混淆度上升的情形。在這兩個表中，以時間匹配的点爲中心往兩邊看，可以很明顯的看出混淆度最低點就是在時間匹配的情況下，而越往兩側去，其趨勢就是混淆度越來越大。這顯示出：當時間不匹配程度越大時，所建構的三連語言模型就越沒有能力描述或預測另一個時間點的語料。

3.2 精緻語料庫的取得

如前所述，N連統計式語言模型是目前最常用且最成功的語言模型，然而正如3.1節所

訓練\測試	一、二月	三、四月	五、六月	七、八月	九、十月	十一、二月
一、二月	1.00	1.71	1.92	1.96	1.91	1.92
三、四月	1.81	1.00	1.67	1.98	1.95	1.98
五、六月	1.93	1.67	1.00	1.89	1.93	1.97
七、八月	1.72	1.60	1.55	1.00	1.63	1.71
九、十月	1.75	1.65	1.65	1.67	1.00	1.65
十一、二月	1.75	1.66	1.69	1.77	1.64	1.00

表 3.4: 時間匹配性和混淆度上升比例的關係

述, 使用 N 連語言模型以預測詞出現的機率時, 必須使用匹配的訓練語料訓練 N 連語言模型。因此在語音辨識上使用 N 連語言模型辨識目標語料時, 最重要的是必須要取得和目標語料同質的文字語料。除了語料之外, 本論文的實驗將會統計語言模型調適技術來利用同質性高的調適語料, 來將語言模型調適至和目標語料更為接近。

本篇論文中將會討論使用兩種不同的外來語料來源作為調適語料, 這兩種不同來源語料庫, 本論文中分別稱為既有語料庫 (existing corpora) 及衍生語料庫 (derived corpora)。由於本論文語音辨識的目標語料主要為新聞語料, 因此對應的既有語料庫指的是從雅虎奇摩收集的新聞; 衍生語料庫指的是以基礎轉寫結果為參考, 產生適當的查詢指令 (query) 並透過搜尋引擎查詢網際網路, 以檢索方式試圖得到和目標語料有高同質性的調適語料。除了採用兩種不同來源的語料庫之外, 本章的實驗還採用了文件分群 (document clustering) 及文件分類 (document classification) 技術, 建構一個分群分類架構 (Clustering-Classification framework) 來幫助語料庫精緻化, 分群分類架構在 3.3 節有細部的討論。我們將先討論本論文使用的兩種語料庫及其精緻化。

3.2.1 既有語料庫與衍生語料庫

本論文中將討論兩種不同特性的語料庫 (既有及衍生語料庫), 其中「既有語料庫」指的是我們取用現成的、且和目標語料有類似特性的語料庫。

在本論文的實驗中, 由於欲辨識的目標語音語料是廣播新聞, 因此採用的既有語料庫

指的是網路上可取得的大量新聞資料庫。有了既有語料庫之後，我們可以針對主題、時間的同質性來作分析。例如，使用分群分類架構取得在詞-文件向量空間(word-document vector space)中具有相近特性的文件群(document clusters)，或者選用時間相近的語料來訓練語言模型。

衍生語料庫指的是，由目標語音語料中取得一些相關知識，來取得同質性高的精緻訓練語料。從目標語音語料取得資訊的方法之一，是利用基礎辨識器所得到的基礎轉寫結果，進行檢索以得到精緻化語料。

現在的問題是，當給定目標語音語料、及其有錯誤的基礎轉寫結果之後，要如何取得同質性更高的精緻語料庫？從資訊檢索及大量語料庫的需求來看，網際網路是一個相當具有潛力的資源；而最有效取得和目標語音文件同質性高的語料庫的方法之一是使用搜尋引擎，如 Google [29]、AltaVista [30]、Openfind [31] 等等。藉由搜尋引擎蒐集語料的一個缺點是，取得語料較費時，因此這裡討論的作法並不適用於需要即時處理的應用。不過仍有部分的語音辨識應用並不需要即時輸出辨識結果，例如本論文實驗所採用的廣播新聞語音辨識問題。

欲使用搜尋引擎檢索得到同質性高的精緻語料，其關鍵性問題在於要如何產生合適的查詢指令來查詢搜尋引擎。由於語音辨識轉寫結果往往包含許多錯誤，因此必須設計有容錯能力的查詢指令建構之方法。在 3.2.2 及 3.2.3 兩節中，將提出並實驗兩種不同的查詢指令建構(query-construction)方法：分離式三連詞查詢指令建構法(*disjunctive 3-word queries*)及相連式高信心量度查詢指令建構法(*conjunctive queries with higher confidence measure*)。

3.2.2 分離式三連詞查詢指令建構法

本論文中實驗都是做在中文語音辨識上，既然是處理中文，在建構查詢指令時就不能不

考慮到中文的特性。由於中文的詞與詞之間並沒有明顯的分界，且中文新詞可由任意的中文單字組合而成，因此中文出現的新詞（如類專有名詞）常常並不包含在基礎辨識器所使用的基礎辭典之中，且以單字詞型態出現。由於新詞以單字詞型態出現，從各個單字的角​​度來看，每個字也許重要性並不高，但從數個字組成的序列來看，的確存在一個新詞。在這樣的考量之下，資訊檢​​索中某些用以選取重要詞的方法——如詞頻倒文件頻 (TFIDF)——並不適合直接套用。

考慮正確性及檢​​索到的資料量的情況之下，在初步的實驗中嘗試了一個很直觀的查詢指令建構法：選取基礎轉寫結果中，任意三個連續的詞作為分離式三連詞查詢指令 (disjunctive 3-word queries)，以這些查詢指令來查詢 Google 並取得相關語料。

在這個方法中，我們使用分離式（即各個詞之間的邏輯關係是”OR”）查詢指令，而不採用相連式（即各詞之間的邏輯關係是”AND”）的原因是，我們選取了所有的任意三個連續的詞，而這些詞中很可能帶有辨識錯誤，在對辨識結果不夠確定的狀況之下，採用 AND 會使得限制過多，而且可能限制到錯誤的範圍，如果在這種情況下使用相連式查詢指令，會使得搜尋引擎找到的語料太少或甚至完全不相關。因此，在沒有其他指標能讓我們選用較正確的詞的狀況之下，選用任意三個連續的詞時應該建構成為分離式查詢指令。

3.2.3 相連式高信心量度查詢指令建構法

如同上節所述，基礎辨識器所產生的基礎轉寫結果總是含有一些錯誤，因此，使用一些過濾的機制來選取正確性較高的詞，是改進查詢指令建構的合理想法。從辨識器產生的基礎辨識結果中，選用正確性較高的詞來建構查詢指令，才能避免建構出的查詢指令有可能檢​​索不到相關的文件。本節中提出了一個加入信心量度 (confidence measure) 以建構查詢指令的方法。我們使用的信心量度 $C(w)$ ，是針對一個詞假​​設 (word hypoth-

esis) w 的事後詞機率 (*a posterior word probabilities*)。事後詞機率可由計算一個詞 w 所有的詞歷史和詞未來的機率總和而得到 [32]。

$$\begin{aligned}
 C(w) &= p(w|x_1^T) \\
 &= \sum_{h_2^{m-1}} \sum_{f_1^{m-2}} \frac{\Phi(h_2^{m-1}; w) \cdot \Psi(w; f_1^{m-2})}{p(x_1^T) \cdot (x_1^T|w)} \\
 &\quad \cdot \prod_{n=1}^{m-2} p(f_n|h_{n+1}^{m-1} w f_1^{n-1}) \tag{3.1}
 \end{aligned}$$

$\Phi(h_2^{m-1}; w)$ 代表長度為 m 的詞串假設中，最後一個詞假設是 w 且其詞歷史為 h_2^{m-1} 的前向機率 (forward probability); $\Psi(w; f_1^{m-2})$ 是長度為 m 的詞串假設中，第一個詞假設是 w 且其詞未來為 f_1^{m-2} 的後向機率 (backward probability)。

這個建構方法不再是只看最佳解 (single best)，而是需要用到辨識器產生的詞圖 (word graph) 資訊。首先我們將詞圖中的每個詞假設標上對應的信心量度，並設定一個信心量度的門檻值 (threshold)，然後選取詞圖中信心量度高於門檻值的詞假設，並將在詞圖上相鄰的詞作為相連式查詢指令，用以查詢搜尋引擎。舉例來說，圖 3.1 中，實線代表信心量度高於門檻值的詞假設，在這個例子中，詞串 $w_a w_b w_c$ 和 w_d 會被選為查詢指令。由於這個方法建構查詢指令時，已經將信心量度的資訊加入，因此用相連式高信心量度查詢指令檢索到的文件，會比使用分離式三連詞查詢指令檢索到的文件和目標語料的同質性更高。

3.3 分群分類架構

在 3.1.1 及 3.1.2 兩節的主題、時間不匹配性的實驗中，可以發現雖然大量的訓練語料對

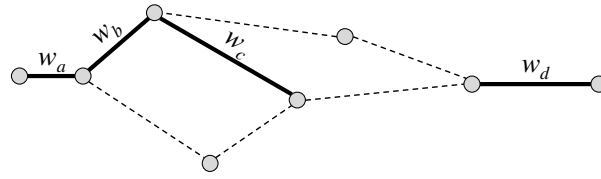


圖 3.1: 根據高信心量度在詞圖上建構查詢指令

於語言模型的效能有幫助，然而內容廣雜（異質性高）的大量訓練語料，有時候反而會傷害語言模型的精準度，因而影響到系統的效能。因此，與其使用大量但過於混雜的語料，我們寧可取用同質性高、但不必然要很大的精緻語料來做語言模型調適。

本論文中提出的分群分類架構，整合了文件分群、文件分類的機器學習(machine learning) 技術來幫助取得更精緻的語言模型調適語料庫，藉由精緻化的調適語料庫，我們可以將語言模型調適到更貼近目標語音語料的機率分佈。

3.3.1 文件分群

本論文中使用的文件分群技術是 K-means 演算法的一種變形：二分式 *K-means* 演算法 (*bisecting K-means*)。二分式 K-means 演算法的優點是比傳統的 K-means 演算法更有效率，且能產生大小較為一致的文件群 [33]。二分式 K-means 演算法的步驟如下：首先先選定一個文件群來切分 (split)，接著，使用傳統的 K-means 演算法，將這個選定的文件群分成兩個小的子文件群 (sub-clusters)，重複這個步驟直到達到原先設定的文件群數量。在本論文的語料庫精緻化架構中，文件分群這個部分是使用 CLUTO (A Clustering Toolkit) [34]，這個工具實作了二分式 K-means 演算法，適合用在分群分類架構上。

3.3.2 文件分類

在分群分類架構中使用文件分類的目的，是將每件目標語音文件分配到一個或多個預先定義的子語料庫中。本論文中使用的文件分類是 Naive Bayes 分類器，使用 Naive

Bayes 分類器的基礎假設是：選用的各個特徵必須互為獨立。當我們選用詞為特徵時，這個假設看起來不甚合理，然而在過去的文獻中曾有探討過，即使選用的特徵之間並未完全獨立，並不代表分類器的效能就會變差 [35]。此外，由於 Naive Bayes 分類器的計算效率很高，因此很適合大量資料的分類工作。

Naive Bayes 分類器是直接由訓練語料訓練得到，其目標是估測「給定一文件特徵向量，估測該文件為任一個類別的機率」。經由貝氏定理，我們可以預測一個文件（特徵向量 d ）屬於某一主題 T_j 的機率為：

$$P(T_j|d) = \frac{P(T_j)P(d|T_j)}{P(d)} \quad (3.2)$$

預測文件屬於哪一個主題時，只需要找到最高分的那個主題 T_j^* 即可。因此

$$T_j^* = \operatorname{argmax}_{T_j} P(T_j|d) = \operatorname{argmax}_{T_j} P(T_j)P(d|T_j) \quad (3.3)$$

在本論文實驗中，文件分類採用的程式為 Rainbow; Rainbow 是 Bow toolkit 的一部份，實作了 Naive Bayes 分類器。因此在本論文中使用 Rainbow 以達到文件分類的目的 [36]。此外，由於原本的 Rainbow 程式無法處理中文字，因此我們將 Rainbow 稍加修改，即成為使用於分群分類架構中的文件分類器。

3.3.3 分群分類法之應用

如圖 3.2 所示，給定一個大的文字語料庫集合 C ，首先使用文件分群將 C 分成許多小的子語料庫(sub-corpora) C_1, C_2, \dots, C_N ，由於文件分群採用的文件特徵由文件中所含的詞所得到的文件特徵向量 (document feature vectors)，因此每一個子語料庫都是在詞-文件向量空間中內聚力較高的文件群，也因此各個子語料庫可視為分別具有較高度相關的主題 T_1, T_2, \dots, T_N 。接下來，我們將這些同質性高的子語料庫作為文件分類的訓練語料，在訓練好文件分類器 (document classifier) 之後，將語音文件的基礎轉寫

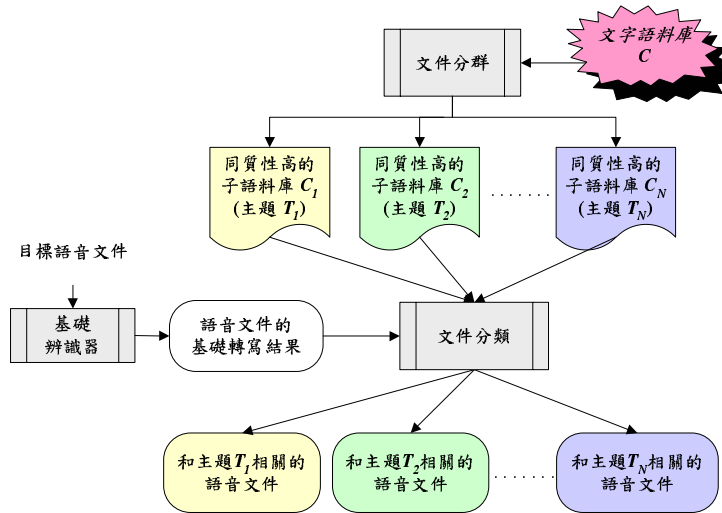


圖 3.2: 分群分類架構 (Clustering-Classification framework) 用以取得同質性高的語料

結果分類至 C_1, C_2, \dots, C_N 這些子語料庫下。到此為止分類分群法的第一部分完成，第一部份的目的是以分類分群架構，取得同質性高的精緻子語料庫。

第二部分如圖 3.3 所示，目的是分別以各個語音文件對應的子語料庫來做語言模型調適，並用對應的調適語言模型來重新辨識語音文件，最後取得精緻化的語音文件轉寫結果，並觀察其辨識率的進步量。

3.4 主題匹配性之實驗結果與比較

在本節的實驗中，以廣播新聞語料為目標語料，首先討論採用分群分類架構的影響，接著再討論使用兩種不同的語料庫：既有語料庫和衍生語料庫時不同的辨識結果。針對廣播新聞語料，實驗中選用的既有語料庫是雅虎奇摩 2002 年 8,9 月的新聞，共 63,253 則,29,076,437 字。衍生語料庫則是使用基礎語料器所辨識得到的基礎轉寫結果，分別以 3.2.2 和 3.2.3 兩節的查詢指令建構法產生查詢指令，並從 Google 上取得相關的文件，收集而成的語料庫。

此外，為了把觀察的重點著重在語料庫的主題匹配性，因此辭典固定使用辨識用基礎

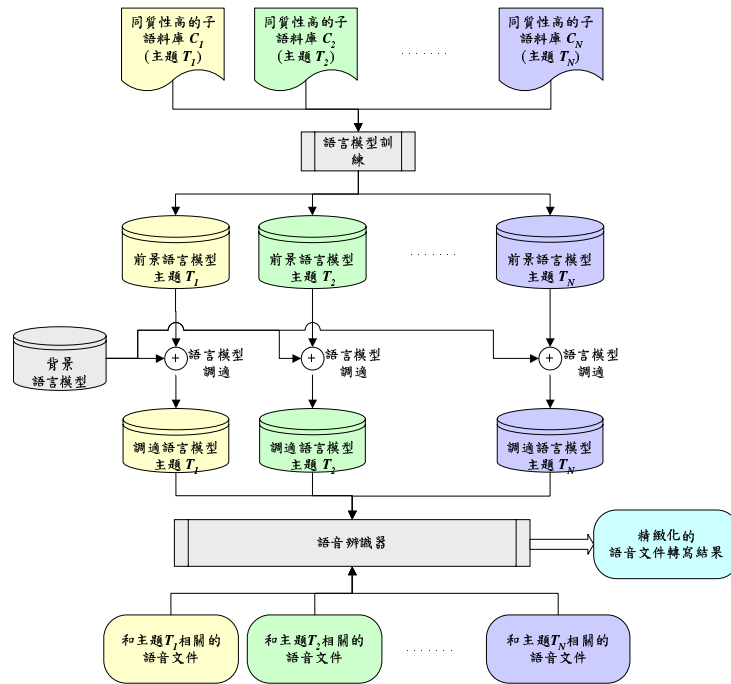


圖 3.3: 語言模型調適架構

辭典 (參考 2.3.4 節), 並不針對各個語料庫抽取新詞, 如此一來才能確定辨識率的變化是由語料庫的品質變動而來。

本章中均以辨識用基礎辭典, 以長詞優先斷詞法將語料庫進行斷詞, 再訓練語言模型。

3.4.1 分群分類架構的影響

為了瞭解採用分群分類架構, 在兩種不同語料庫上的影響, 我們設計了以下的實驗: 以 2002 年 9 月的 506 則廣播新聞語料作為目標語料, 分別使用 3.4 節所述的既有語料庫及衍生語料庫, 以辨識率的進步與否, 來觀察分群分類架構是否有達到取得精緻化語料庫的功能。

表 3.5 的結果中, 是以分群分類架構辨識廣播新聞語音語料, 得到的字、次音節錯誤率。括號中的百分比是指和不使用分群分類架構 (即「1 群」) 的比較的相對錯誤率降低

語言模型		字錯誤率	次音節錯誤率
基礎語言模型		23.17	13.17
以既有語料庫 調適語言模型	1 群	18.05	10.58
	10 群	17.32 (4.04%)	10.20 (3.59%)
	50 群	16.76 (7.15%)	9.94 (6.05%)
	100 群	16.75 (7.20%)	9.96 (5.86%)
	200 群	16.72 (7.37%)	9.96 (5.86%)
	500 群	16.86 (6.59%)	10.07 (4.82%)
以衍生語料庫 調適語言模型	1 群	20.18	11.65
	10 群	19.80 (1.88%)	11.43 (1.89%)
	50 群	19.83 (1.73%)	11.46 (1.63%)
	100 群	19.85 (1.64%)	11.40 (2.15%)
	200 群	20.29 (-0.55%)	11.71 (-0.52%)
	500 群	20.88 (-3.47%)	11.97 (-2.75%)

表 3.5: 字錯誤率與次音節錯誤率及相對降低比率。

比率。本實驗中採用相連式高信心量度查詢指令建構法。第一列是使用基礎語言模型得到的基礎轉寫結果之字、次音節錯誤率，這些基礎轉寫結果，也就是用來建構查詢指令的基礎語料，我們可以發現其字錯誤率為 23.17%，也就是說平均一百個字裡面，就有超過 23 個字是辨識錯誤的，這些錯誤即是查詢指令建構法必須具有容錯能力的最大原因。表 3.5 顯示使用分群分類架構後，無論是既有語料庫或衍生語料庫均可達到比未使用此架構更好的辨識率。其中最好的辨識率是當使用既有語料庫，分成 200 個文件群時，分群分類架構的功能讓我們得到 7.35% 相對字錯誤率降低，及 5.86% 相對次音節錯誤率降低，且分成 50 群到 500 群之間的增進率都相當好。對衍生語料庫來說，在這個表裡最好的文件群數是 50 到 100 之間。事實上，從表 3.5 中觀察得知，在尚未採用分群分類架構之前（亦即只使用 1 群），使用既有語料庫及衍生語料庫的字錯誤率，就已經從基礎實驗 23.17% 分別下降至 18.05% 及 20.18%，這個辨識率的上升顯示既有語料庫和衍生語料庫的確是較精緻的語料庫，既有、衍生語料庫的比較將在 3.4.2 節中詳述，在此不加贅述。採用分群分類架構之後，在既有語料庫的狀況之下，字錯誤率可以從 18.05% 降低至

16.72%; 在衍生語料庫的狀況下則可由 20.18% 降至 19.80%。

在這裡有兩點值得注意的：

- 調適語言模型所使用的語料量：由 3.1 節的實驗可以發現：有些狀況下，僅以同主題的語料訓練並不見得會贏過所有語料都拿下去訓練的效能。這是因為在文體沒有差異太大時，語料量的增大仍然可以提供比較好的語言模型。在表 3.5 的實驗中，使用分群分類架構會使調適語料量減小（因為文字語料會被分成數個文件群，不會同時使用全部來調適語言模型），在語料量減少的情況之下，我們可以看到使用分群分類架構的辨識結果仍有進步，這說明了分群分類架構並非隨機挑選語料，而是有達到語料庫精緻化的功效。
- 分群分類架構即濾除雜訊的想法：事實上，承上點所說分類分群架構可達到語料庫精緻化的功效。從另一個角度來想，就單獨一篇語音文件的辨識來說，在大量的文字語料庫中，並非每篇文字新聞對於預測該篇語音文件的語言行為都會有幫助。我們可以想像在新聞的概念空間 (concept space)¹ 中分佈著很多的點，其中有許多點群聚在一起，具有較相近的主題。分群即是將這些群聚的點找出來，而分類是把語音文件指派到和它相近的群聚點中，同時也排除了其他和它不相似的文章，這些不相似的文章如果拿來訓練語言模型，用以辨識該語音文件的話，會影響理想的機率分佈。從概念空間上來說，這些不相似的文章相當於是歧異點(outliers)，也就是雜訊 (noise)。

目前分群分類架構的一個缺點，其實和文件分群原生的問題有關，也就是如何選取群數。從概念空間及語言模型訓練的角度來看，群數的選取問題，就是應該選多少群，才

¹在本論文實驗中，文章的概念空間即為詞-文件向量空間。

能使每個群的內聚力夠強，而且每個群的數量又不能太少，這樣對語言模型調適會比較有利。

3.4.2 既有語料庫和衍生語料庫之比較

從表 3.5 及圖 3.4 中，可以看到既有語料庫和衍生語料庫之間明顯的差異：使用既有語料庫的辨識率一定比衍生語料庫的辨識率高。在本實驗設定之下，觀察到這個結果其實是非常合理的，這是由於本實驗的目標語料是廣播新聞，新聞的特性是文體較為一致，雖然常會有新詞、新類專有名詞出現，但由於其內容和當時發生的事件有絕對的相關性，因此如果能取得同時間的大量文字新聞語料，就相當於得到了同質性非常高的既有語料庫。又由於新聞是一個有大眾化市場需求的語料，因此網路上有許多現存而垂手可得的新聞網站，這也使得廣播新聞語音辨識這個問題，取得既有語料庫的困難度大為降低。相對來說，衍生語料庫是直接利用語音文件的基礎轉寫結果，透過搜尋引擎，從網際網路上搜尋相關的文件。從圖 3.4 中可以明顯看到：無論識字錯誤率或者是次音節錯誤率，使用既有語料庫的錯誤率（虛線）均比使用衍生語料庫的錯誤率（實線）來的低。這證明了衍生語料庫較既有語料庫更為雜亂，分析其原因主要有三點：

- 整個網際網路包含的語料的廣雜性：由於衍生語料庫的取得，是透過搜尋引擎存取網際網路，因此衍生語料庫的特性，和搜尋引擎的排行(ranking) 機制、搜尋引擎所貯藏的網際網路語料特性有關。本實驗中使用的搜尋引擎是 Google，其搜尋空間包含 4,285,199,774 個網頁，其內容包羅萬象，並不侷限於哪個領域或哪個主題，其文體也大異其趣。Google 使用的網頁排行機制是 PageRank，它利用回溯連結(Backlinks) 來決定一個網站的排行價值；也就是說，越多人連結的網頁，就可能擁有更高的 PageRank；另用 Google 內部的同步運算，將數以數十億計的網際網路網頁架構起來。排行機制可以說是能取得同質性高的衍生語料庫的關鍵

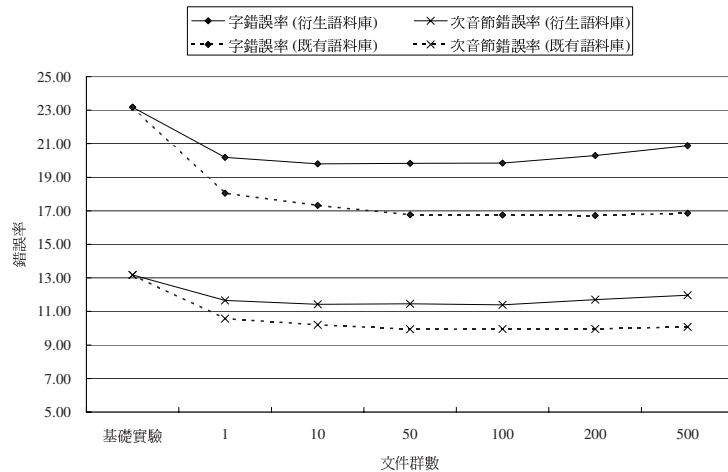


圖 3.4: 字錯誤率及次音節錯誤率—文件群數

之一。

- 基礎轉寫結果中帶有的錯誤：無論是採用哪一種查詢指令建構法，都難以避免受到基礎轉寫結果錯誤的影響。在使用分離式三連詞查詢指令建構法時，相當於沒有過濾任何錯誤的辨識結果，很可能產生了許多錯誤的查詢指令。在3.2.3提出的相連式高信心量度查詢指令建構法，有效地利用了詞圖和信心量度的資訊，降低了基礎轉寫結果中的錯誤造成的影響。
- 選取查詢指令的標準，光是選正確辨識的詞是不夠的：前面討論了辨識錯誤的情形，但即使加入了信心量度的指標來篩選詞，就算選出來的詞是正確的，也很可能並沒有辦法查詢到相關的文件。舉例來說，「今天 在」這樣的二連詞，即使其在詞圖中的信心量度夠高而被選取出來，將這樣的查詢指令送到Google去查詢，也未必能得到和語音文件相關的語料。因此仍應考慮加入其他條件（例如倒文件頻），來選取出更有意義的查詢指令。

查詢指令建構法	查詢指令總數	衍生語料庫字數
分離式三連詞建構法	31108	17092041
相連式高信心量度建構法	7301	3754509

表 3.6: 用兩種不同查詢指令建構法, 所產生的查詢指令總數, 以及從搜尋引擎上檢索到的衍生語料庫總字數。

3.4.3 查詢指令建構法之比較

在本節中, 我們以實驗比較在 3.2.2 與 3.2.3 兩節中提出的兩種查詢指令建構法。實驗的語音語料是 506 則廣播新聞, 而兩種查詢指令建構法都是從基礎轉寫結果開始產生查詢指令, 然後再去查詢搜尋引擎。表 3.6 中顯示: 使用了信心量度後, 檢索得到的衍生語料庫量, 比起沒有使用信心量度得到的衍生語料庫量, 降為四分之一。

爲了瞭解這兩種查詢指令建構法所取得的衍生語料庫, 能調適語言模型達到多少的辨識效能, 以下設計了初步的實驗來看可以達到多少的相對進步。以下的實驗, 針對調適語料的線性內插參數 λ_{adapt} 做 0.1 到 0.9 九組, 從圖 3.5 中, 上半是字錯誤率-線性內插參數 λ_{adapt} 作圖, 下半則是次音節錯誤率-線性內插參數 λ_{adapt} 作圖。首先我們可以發現, 對於分離式三連詞建構法來說, 最低的字錯誤率和次音節錯誤率分別是 20.18% 和 11.65%; 對相連式高信心量度建構法來說, 則分別是 20.16% 和 11.49%。由這裡的實驗數據, 在比照表 3.6 的語料庫大小, 我們發現相連式高信心量度建構法取得的衍生語料庫雖然只有四分之一的大小, 但其能達到最好的辨識率, 甚至比用分離式三連詞建構法取得的衍生語料庫更能達到的更好。這顯示出, 將基礎轉寫結果中的錯誤過濾掉, 對於衍生語料庫的精緻化是相當有幫助的,

在圖 3.5 中, 觀察 λ_{adapt} 的改變和辨識率的變化可以發現, 相連式高信心量度建構法對於 λ_{adapt} 的改變比較敏銳, 造成這個現象的原因, 可能是因爲這個方法取得的衍生語料庫量比較少, 雖然由於加入信心量度的關係會比較精準, 但因爲量比較少, 會比較偏

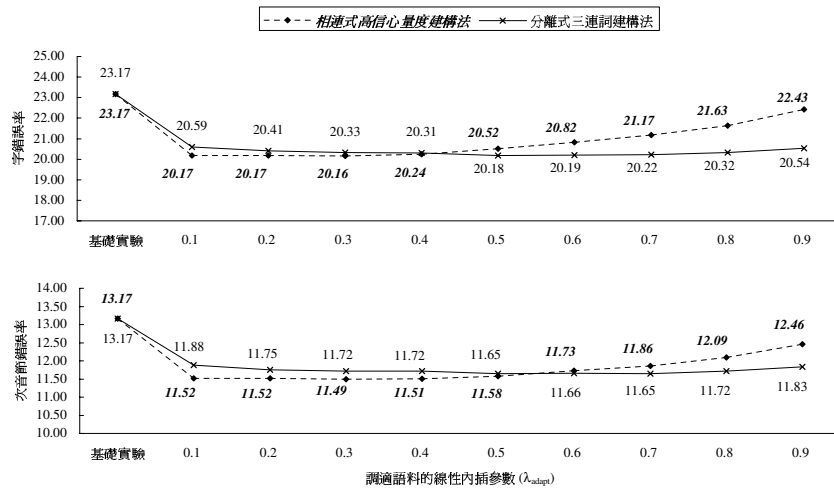


圖 3.5: 字/次音節錯誤率-內插參數圖

向且侷限於基礎轉寫結果的內容，因此，當相連式高信心量度建構法的 λ_{adapt} 調的太大的時候，辨識效能減低較為明顯。另一方面來說，分離式三連詞建構法收集的語料，雖然雜訊可能比較多，但是另一方面，語料量比較多會造成對 λ_{adapt} 的變動敏感度較低。

3.5 時間匹配性之實驗結果與比較

本節的實驗中，以2002年九月506則廣播新聞語料為目標語料，文字調適語料部分，則針對各段中不同的分析討論需求，從雅虎奇摩2002年8,9月的新聞中選取適當的子集。本節許取的文字語料包括:2002年8月的文字新聞，約15萬字；2002年9月的文字新聞，約19萬字；2002年8月15日至9月15日的文字新聞共41,782則，約17萬字。

此外，為了把分析的重點著重在目標語料和調適語料庫的時間匹配性之上，本節中辭典固定使用辨識用基礎辭典，並不針對不同時段的語料庫抽取新詞，這樣才能確定辨識率的變化是由語料庫的品質變動而來。在第五章中，將會整合時間和主題的匹配性來強化語言模型，並以實驗驗證分析其效果。

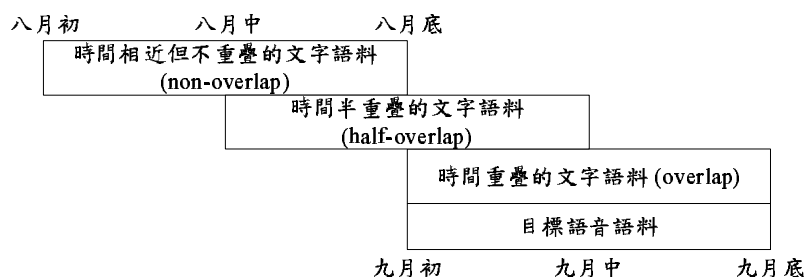


圖 3.6: 語言模型訓練/調適語料與目標語料的時間關係圖(2002年)

3.5.1 以長度一個月的滑動窗進行時間重疊性分析

首先我們先設計一個實驗，從較巨觀的角度來觀察目標語料和語言模型訓練及調適語料的時間重疊性，對於辨識結果的影響。本實驗以 2002 年 9 月的 506 則廣播新聞語料作為目標語料，分別以 2002 年 8 月 (時間不重疊)、2002 年 8 月 15 日至 2002 年 9 月 15 日 (時間半重疊)，及 2002 年 9 月 (時間完全重疊) 的雅虎奇摩新聞作為語言模型的訓練及調適語料，分析辨識率的變化，觀察時間性對辨識的影響。

圖 3.6 顯示：實驗採用的兩份不同語言模型語料及調適語料與目標語料的時間關係。實驗中，首先採用這兩份文字語料直接訓練而成的語言模型做語音辨識，接著在將其與基礎語言模型以線性內插方式調適，觀察辨識率的變化。辨識實驗的結果列於表 3.7。首先探討不考慮基礎語言模型，僅使用時間不重疊、半重疊、完全重疊各一個月的文字語料訓練語言模型，這樣得到的辨識率，我們可以發現使用完全重疊語料時，如我們所預期的，字、次音節錯誤率比起用過時的基礎語言模型得到的錯誤率，分別進步了 16.31% 及 12.38%；然而使用不重疊語料時，可以發現字、次音節錯誤率都有明顯幅度的退步；半重疊語料的部分的辨識結果則是介於兩者之間。這說明了：雅虎奇摩新聞一個月的文字語料量，已經足夠獨立訓練語言模型用以辨識。然而若單純選用一個月文字新聞，可以發現時間重疊時有顯著的進步，但在時間雖不重疊的時候，有明顯的退步，半重疊時

3 語料庫精緻化

語言模型			字錯誤率		次音節錯誤率	
基礎語言模型			23.17		13.17	
用以單獨訓練 語言模型	時間不重疊 (non-overlap)		25.39	(-9.58%)	14.11	(-7.14%)
	時間半重疊 (half-overlap)		21.88	(5.57%)	12.62	(4.18%)
	時間重疊 (overlap)		19.39	(16.31%)	11.54	(12.38%)
用以調適 基礎語言模型	時間不重疊 (non-overlap)		20.69	(10.70%)	12.10	(8.12%)
	時間半重疊 (half-overlap)		18.92	(18.34%)	11.36	(13.74%)
	時間重疊 (overlap)		17.51	(24.42%)	10.64	(19.21%)

表 3.7: 時間不重疊、半重疊及完全重疊之字錯誤率與次音節錯誤率及和基礎辨識結果比較之相對降低率。

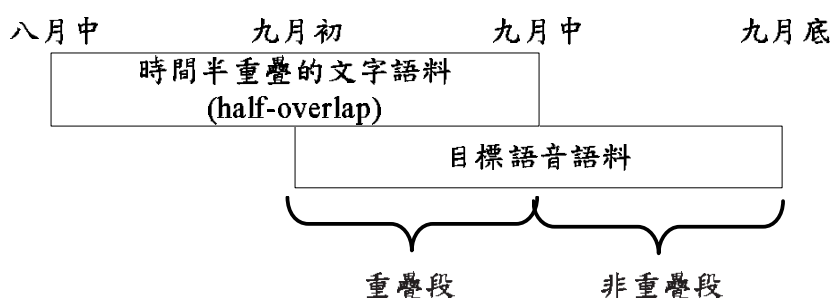


圖 3.7: 時間半重疊語料集與目標語料的時間關係圖(2002年)

則是介於中間的辨識率。另外，這兩組文字語料，在與基礎語言模型做線性內插的調適之後，個別的辨識率又有都進步，其原因固然可以推論為：混合兩語言模型後，相當於訓練語言模型的語料量增加，但由基礎語言模型向來穩定的表現，可得知除了取用時間匹配性高的文字語料以外，最好能存在一個仔細調整過的、效能穩定的基礎語言模型，以調適的方式產生新的語言模型。

在表 3.7 中，可以發現時間半重疊部分，其實可以從辨識語料的角度細分成前半個月（重疊段）及後半個月（非重疊段）來探討辨識結果。首先，由調適語料（時間半重疊）與目標語料的時間關係可看出重疊的部分，圖 3.7 為示意圖。目標語料與調適語料重疊的日期為 9 月 1–13 號，稱之為重疊段；而其餘沒有重疊的目標語料為 9 月 17–30 號，稱之為非重疊段。以下將更進一步地觀察並分析：時間半重疊語料集對於重疊段與非重疊段目標語料辨識率的影響。

在這個分析中，將目標語料，如上述分成重疊段與非重疊段。並比較三種語言模型：

1. 基礎語言模型。
2. 由時間半重疊語料(2002年8月中到9月中) 單獨訓練出來的模型。
3. 由基礎語言模型，根據時間半重疊語料做線性內插調適所得的強化語言模型。

辭典的部分，由於之前提過本章完全討論語料庫的影響，因此並未進行抽詞，辭典均採用基礎辭典。實驗結果表示於表3.8。

由表3.8的「用以單獨訓練語言模型」欄可看出在重疊段的字錯誤率，比使用基礎語言模型，進步了 14.93%；但在非重疊段，卻退步了 -5.65%。這現象可以視為使用時間匹配性高的語料，直接單獨訓練語言模型時，對於訓練語料（在此為調適語料集）有比較好的貼合度，但是也造成了過度貼合，使得其語言涵蓋率比基礎語言模型差，亦即其對未見語句的估測能力遜於基礎語言模型（由非重疊段比較得知）。因此必須要使用調適法，使得調適過的強化語言模型能兼具基礎語言模型的語言涵蓋率，和調適語料對即期新聞語料的估測能力。在「用以調適基礎語言模型」欄可看出：使用調適法不但可從單獨訓練語言模型中已有進步的重疊段字錯誤率進步至 17.60 (24.07%)，亦使得原先單獨訓練語言模型中退步的非重疊段，進步至 20.05 (10.77%)。重疊段的再進步，說明了調適語料仍不能完全涵蓋目標語料的語言特性，如果能同時考慮涵蓋率較廣的基礎語料，則可補其不足。非重疊段相對的大幅進步，表示基礎語料的確可以補救單獨訓練語言模型時，模型過度貼近調適語料的缺點，並能發揮調適語料對即期新聞主題的估測優勢，使得辨識率有所提升。整體而言，這表示對於一涵蓋率較廣的基礎語言模型，根據目標語料的語言特性做語言模型調適，對於語音辨識率有正面的幫助，且比單獨使用時間匹配性高之語料訓練的模型要來得好。

語言模型		字錯誤率	次音節錯誤率
基礎語言模型	重疊段	23.18	13.13
	非重疊段	22.47	13.21
用以單獨訓練語言模型	重疊段	19.72 (14.93%)	11.52 (12.26%)
	非重疊段	23.74 (-5.65%)	13.57 (-2.73%)
用以調適基礎語言模型	重疊段	17.60 (24.07%)	10.61 (19.19%)
	非重疊段	20.05 (10.77%)	12.00 (9.16%)

表 3.8: 重疊段/非重疊段之字錯誤率與次音節錯誤率及和基礎辨識結果比較之相對降低率。

進行到此, 回顧之前較巨觀的實驗 (表 3.7), 到進一步分析重疊段和非重疊段的實驗 (表 3.8), 我們得到了以下的結論:

- 時間重疊度與辨識率: 調適語料與目標語料時間重疊度越高時, 其辨識效能有越顯著的進步。
- 與基礎語言模型進行內插線性調適的必要性: 即使新進的調適語料集足夠單獨訓練語言模型, 其對未見事件的預測能力仍低於基礎語言模型, 因此最好的解決方案是進行調適, 以期同時擁有語言涵蓋率及即時新聞主題的估測能力。

3.5.2 如何細緻選取時間匹配語料庫

在之前的實驗中 (3.5.1), 選取時間匹配語料庫的滑動窗 (sliding window) 大小取為一個月, 除此之外, 圖 3.7 中更進一步分析重疊段及非重疊段的辨識率變化, 然而目前為止分析的精細度仍在以月份為單位的層次。接下來針對上述實驗, 設計更為細緻的實驗。本實驗設計的目的是探討調適語料及目標語料的時間匹配性, 對辨識率影響的趨勢。上一個實驗所測試的目標語料分成重疊段及非重疊段兩部分, 現在將這兩段語料依日期列成兩個序列, 分別為 (2, 3, 4, 9, 10, 11, 12, 13) 和 (17, 18, 19, 20, 23, 24, 25, 26, 27, 30), 各有 8 天和 10 天。接著為避免某一天因其他因素使得辨識率特別好或特別差, 以三天為一單位做一平滑窗 (smoothing window), 一次移動一天 (例如 (2, 3, 4)

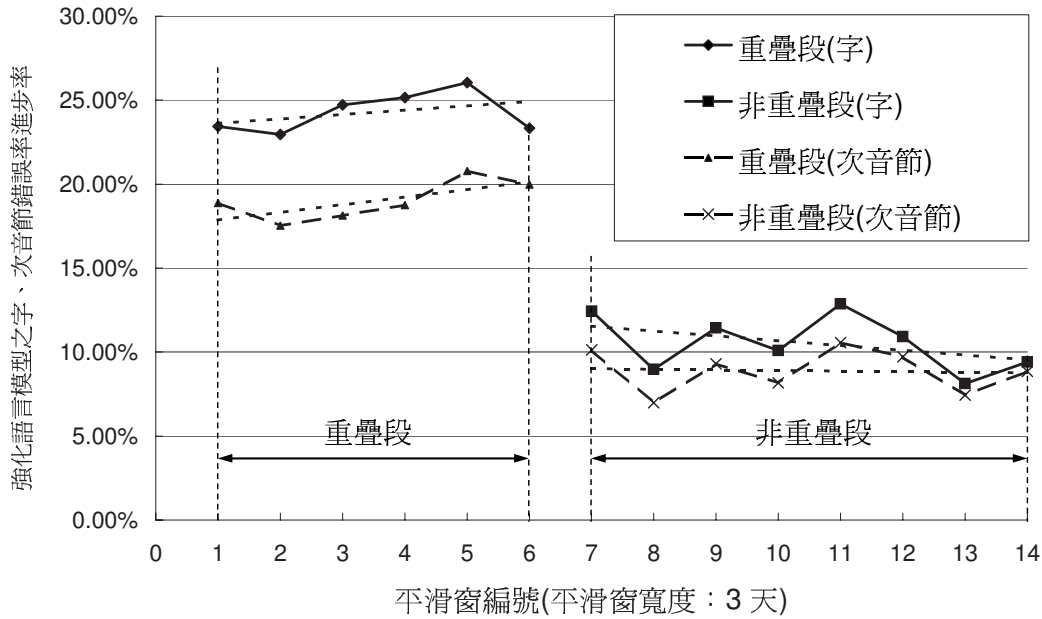


圖 3.8: 目標語料字/次音節錯誤率相對於調適語料的時間匹配性關係圖

→ (3,4,9)), 所以重疊段共得 6 個窗, 非重疊段共得 8 個窗, 依序編 1-14 號, 根據上個實驗中得到的結論, 本實驗中的語言模型使用時間匹配語料調適基礎語料而成的語言模型。求每一平滑窗語料的字錯誤率及次音節錯誤率之進步率, 結果見圖 3.8。

觀察圖 3.8, 從重疊段及非重疊段的線性迴歸線, 可發現在這兩段中的辨識率均為有較穩定的趨勢, 而在這兩段的中間有一個明顯的下降差距。重疊段部分的辨識率較為穩定, 而觀察非重疊段的線性迴歸線會發現, 隨著日期的差距越大, 辨識率有下降的趨勢, 且辨識率較不規則, 即變異比較大。非重疊段的下降現象可由時間半重疊的調適語料, 與辨識的目標語料之間的時間匹配性程度來解釋。假設一特定新聞發生在某日, 且此事件的相關報導持續一小段時間, 對新聞語料來說, 一個合理的推論是此新聞事件相關報導出現的機會, 會隨時間遞減。在重疊段的部分, 由於調適語料中很可能帶有與欲辨識的目標語料中相同事件的相關報導, 因此在重疊段的部分, 辨識率有非常穩定的進步率。在非重疊段的部分, 在九月中時候, 調適語料的新聞事件影響持續的可能較高, 而越靠近九

月底的時候，時間半重疊語料的影響力減小，這可以說明為何非重疊段的辨識率線性迴歸線的斜率為負值。

這解釋了圖3.8非重疊段中，相對於九月底的平滑窗之辨識率（如編號14）比起九月中的（如編號7）要低，且隨日期往後而下降。這是因為幫助非重疊段進步的，主要是九月初到九月中調適語料中新聞事件的延續，這使得九月底的目標語料，雖然處於非重疊段，但與九月初到九月中事件仍有程度上的時間一致性。然而其影響仍較弱，因此仍比不上重疊段辨識率的進步率。然而無論如何，在使用時間半重疊的語料調適基礎語言模型得到的強化模型，來進行辨識實驗時，無論在重疊段及非重疊段，都比未調適的基礎語言模型有更好的表現。此實驗更細緻地展現了時間匹配性對於語言模型強化的重要性。

3.6 本章結論

本章首先提出了兩種可能的目標及訓練語料庫不匹配及其造成的問題——主題不匹配及時間不匹配。並設計了兩組交叉混淆度的實驗，從實際資料的表現來說明不匹配造成的辨識率下降。接下來，本論文中分析兩種不同的語料庫取得來源——既有語料庫及衍生語料庫，此外，根據之前匹配性的推論，本論文提出了一整合文件分群及文件分類的語料庫精緻化架構——分群分類架構。

在實驗的部分分成兩部分，第一部份著重在主題匹配性的相關實驗，包括採用分群分類架構的效果、既有語料庫及衍生語料庫造成的影響及其特性、用以取得衍生語料庫的查詢指令建構法之比較等等。在分群分類架構的實驗中，觀察得知使用此架構取得主題匹配性高且雜訊少的語料時，在既有語料庫的狀況下，可使字錯誤率由基礎實驗的23.17%降至18.05%，使用衍生語料庫的狀況之下，字錯誤率則降至20.18%。在既有、衍生語料庫的影響及特性部分，實驗發現既有語料庫由於其先天與目標語料各方面的一

致性均很好，因此其辨識率有較穩定的進步。衍生語料庫則和其搭配使用的查詢指令建構法有相當密切的關係，使用分離式三連詞建構法時，最好的字錯誤率是20.18%，爲了能讓查詢指令建構法有更高的容錯性，使用加入詞圖及信心量度的相連式高信心量度建構法時，可以在僅僅取得四分之一的衍生語料庫的狀況之下，獲得稍微進步的字錯誤率20.16%。

實驗的第二部分則是時間匹配性的相關實驗，包括以長度一個月的滑動窗，分別採用時間不重疊、時間半重疊及時間完全重疊三個區段的調適語料來進行辨識。在此實驗中，發現時間重疊性越高時，辨識率有非常顯著的進步量。此外，發現使用調適語料結合基礎語言模型所得的強化語言模型，帶來比較大的效能增進。僅由調適語料建構的語言模型對於未見語句的平均辨識率甚至比基礎語言模型差，表示考慮基礎語言模型的強化語言模型比單使用調適語料的模型，兼顧著語言模型的涵蓋率與調適語料的貼近度。

由於使用長度一個月的滑動窗的實驗中發現了時間重疊性與辨識率的高度相關性，因此本論文設計了另一個更細緻的實驗，採用長度爲三天的平滑窗，並使用時間半重疊語料調適基礎模型得到的強化模型進行辨識，在這個實驗中，可以發現非重疊段的辨識率隨著時間距離的拉遠而降低，重疊段則持續有相當良好的辨識率。此實驗再次驗證了時間匹配性的重要性。

整體而言，調適語料庫的精緻化，對於語言模型的調適及訓練在語音辨識實驗上的表現，扮演相當重要的角色。爲了取得精緻化語料庫，必須重視目標語料及調適語料的主題、時間匹配性，此外，選取語料庫的來源也必須特別注意。在取得精緻化語料庫之後，還必須注意：比起單純只用調適語料建構的語言模型，調適後的強化語言模型有著較高的語言涵蓋率，亦即一方面語言模型能貼近調適語料，一方面又能兼顧對一般語言估測的能力。

第四章

辭典精緻化

使用在統計式語言處理的技術，幾乎都是以詞為基礎(word-based) 的。特別是以詞為基礎的統計式語言模型，已經被成功用在許多不同領域，包含語音辨識、資訊檢索等等。雖然以詞為基礎的語言模型在英文中運作地相當好，然而在中文處理時，以詞為基礎的想法遭遇了一些困難。世界上語言的書寫系統大致可以分成兩個重要的類別：表象文字(logographic languages) 及拼音文字(alphabetic languages)。在拼音文字如英文系的書寫系統中，詞與詞之間有明確的分隔符號(空白)，所以詞的定義很明確，也因此採用以詞為基礎的語言模型是相當直觀且有用的。然而，因為中文的書寫系統是屬於表象文字，幾乎每一個中文單字都有其獨立的意義，且字與字、詞與詞之間並無明確的分隔符號。亦即在中文處理中「詞」的並未有明確且統一的定義，也因此並未存在一個被大部分人所接受的辭典，辭典的問題這使得中文的語言模型問題更為複雜，因此在本章中，將討論如何將辭典精緻化，以期強化中文語言模型，使得語音辨識有更好的效能。

4.1 詞的定義與問題

中文的詞的定義並未有明確且統一的定義，這一點可以從存在許多分詞標準這個事實得知。中國大陸訂定的『信息處理用現代漢語分詞規範及自動分詞方法』[37] 及台灣訂定的『中央研究院平衡語料庫的內容與說明』[38] 是兩個中文最有名的分詞標準 (Word

Standards)。在這兩個分詞標準中，就有一定程度的歧異性。舉例來說，「全國」在中國大陸訂定的分詞標準中是一個二字詞，然而在台灣的分詞標準中並不是。除了這兩個分詞標準以外，尚有賓州大學中文句結構數的分詞指導 [39]，在這份分詞標準中亦有提到此三種分詞標準的比較。從以上這三種重要的分詞標準及其存在的歧異性，即可瞭解目前中文對「詞」的定義仍未有完全一致的見解。

除此之外，在中央研究院平衡語料庫的內容與說明中提到的詞的定義為：「一個具有獨立意義，且扮演特定語法功能的字串」[38]，然而，符合此定義的詞，並未必就是最適合語音辨識的語言模型使用的詞。由於中文是一個持續被使用的語言，其中詞彙的產生及消滅也是一直持續發生的，因此一個處理中文的系統，必須有自動加入新詞彙的功能，否則辭典外詞彙 (out of vocabulary, OOV) 會嚴重影響系統的效能。由於在語音辨識系統中，語言模型是一個非常重要的部分，而在中文統計式 N 連語言模型中基本的單位就是詞，辭典在中文語音辨識系統的重要性由此可見。因此，研究如何自動選取新詞，可說是中文自然語言處理中一個基礎而重要的議題。

傳統的抽詞法是以規則為基礎 (rule-based)，大量的人力花費在從語料庫中取出符合標準的詞，後來許多統計的方法被提出，例如在楊凱程等人的研究中發現使用單字詞的中文語言模型並無法達到很好的結果，同時提出使用在辨識系統的辭典中的詞，並不一定要是符合傳統定義的詞，而可定義為符合某些統計特徵的片段樣式 (segment patterns，也就是若干個字構成的一個小字串)[15]，常用的統計特徵包括：詞頻、詞的內聚力及詞邊界完整性 (左右相依性)。在本節中接下來將先討論兩種不同的統計式抽詞法——派樹抽詞法及迭代式組合式抽詞法，這兩種抽詞法就是利用了以上提到的幾種統計特徵，雖然這兩種抽詞法均考慮了詞內聚力及詞邊界完整性，但其演算法的基本精神不同，因此使用這兩個量度的方法也各異其趣。將會在 4.2 及 4.3 兩節中分別介紹。除此之

外，由於抽取出的辭典目的是希望能強化語音辨識系統中的語言模型，因此還有另一個詞統計特徵可以考慮，那就是詞長。在沙氏 (G. Saon) 及派氏 (M. Padmanabhan) 的英文複合詞 (compound words) 相關研究中，曾以實驗探討過語音辨識中音節長度較長的詞，較不容易發生語音辨識錯誤 [40]。因此在本章中亦會討論各抽詞法抽出新詞的平均詞長差異。在 4.4 節中將討論各抽詞法實際應用在中文語音辨識上的效能，及其參數調整的結果。

4.2 派樹抽詞法

使用派樹 (PAT tree) 這個資料結構從給定的語料庫中大量的不同長度 (various length) 的片段樣式中，抽取關鍵詞組 (keyphrase)，是中研院的簡立峰博士首先提出的 [18, 41]，當時將關鍵詞組抽取的重要特性定義出顯著性 (significance) 及完整性 (completeness) 兩種，其目的是希望能抽取出符合這兩種特性的關鍵詞組，以期幫助資訊檢索方面的應用。在語音辨識的應用中，我們得以重新分析這兩種特性之重要性：

- 顯著性：在資訊檢索的應用中，從一個給定的語料庫中抽取出所有的具有顯著性的詞彙以作為資訊檢索的索引詞彙是非常重要的。在簡立峰博士提出的方法中，為了排除一般性詞彙 (例如數字、時間、副詞片語等等) 並抽取出特殊主題中具有顯著性的詞彙，因此分別建立了一般性的派樹 (general PAT tree) 以及特定主題的派樹 (domain-specific PAT tree)，分析其統計分佈以達到取得具有顯著性詞彙並排除一般性詞彙的目的。

然而，本論文改進辭典的目的在於，在語音辨識的應用中增進其辨識率。為達到此目的，並不見得要排除掉一般性詞彙，甚至可以說一般性詞彙仍然有相當高的重要性，因此在本節中討論的派樹抽詞法，並未著重在顯著性的部分。雖然如此，抽

取具有顯著性的詞彙，在語音文件摘要、語音文件中類專有名詞抽取、語音文件標題產生、甚至語音文件分類分群中，都有一定程度的重要性，因此仍有其研究空間。

- 完整性：給定一個語料庫，判斷一個片段樣式(例如「關鍵詞」或「關鍵詞抽」)是否是一個具有完整性的詞彙。事實上，在判斷是否具有顯著性之前，就需要先知道候選詞彙中是否都具有完整的詞邊界，在語音辨識中，辭典中的詞並不一定要是最小的完整詞彙，但是由於詞彙的完整性對於中文的斷詞解歧異時很重要，也因此辭典中的詞彙完整性會影響到 N 連語言模型的品質，所以完整性這個特性，對於語音辨識的應用非常重要。

本論文中採用的派樹抽詞法，僅就完整性的部分討論，相當於是 [41] 中的抽取步驟僅進行至取得具有完整性的詞彙，並不討論其顯著性。此外，在 4.4 節的實驗部分，會討論派樹抽詞法的參數調整在語音辨識應用中的影響。

在本論文的派樹抽詞法中，主要是希望能使用派樹這個資料的結構，在檢索大量語料時可以很輕易的取得任一片段樣式及其出現頻率特性，因此可以有效率地得到在語料庫中，任一長度的片段樣式的相關統計特徵。在討論一個片段樣式的完整性時，派樹抽詞法提出一個好的詞具有夠強的內聚力及完整的邊界兩種統計特徵。圖 4.1 表達了這兩種判斷完整性的統計特徵與候選字串的關係。接下來的小節中，將針對派樹抽詞法中的片段樣式內聚力量度及邊界量度做討論。

4.2.1 片段樣式內聚力量度

片段樣式的內聚力指的是組成該片段樣式中的各字、詞之間的相關性，當一片段樣式有著高的內聚力時，表示在該語言的使用行為中，有相當高的可能會產生這樣一個片段樣

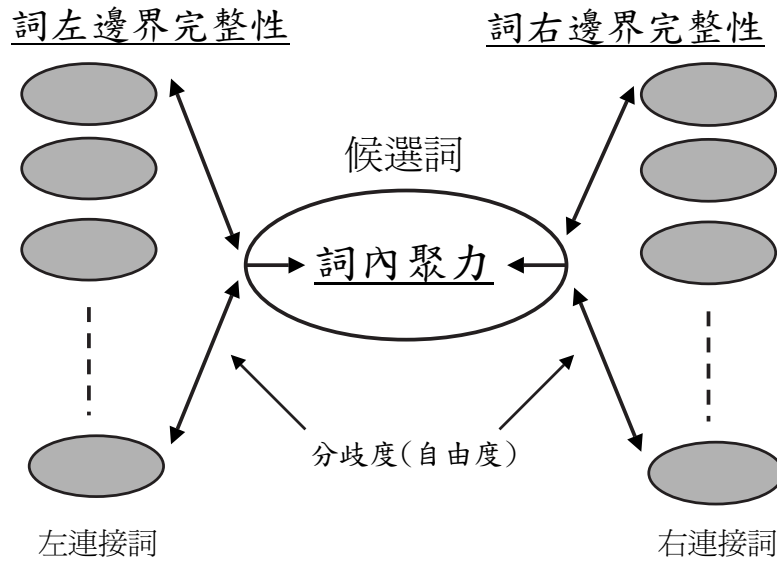


圖 4.1: 詞聚力與左右相依性示意圖

式，因此從統計式抽詞演算法的角度來看，具有高內聚力的片段樣式即代表這個樣式存在的可能性較高，也因此可以被抽取成爲一個詞。

內聚力的量度，一個最基礎的方法是採用資訊理論中的兩點式相互資訊 (pointwise mutual information):

$$I(x, y) \equiv \log_2 \frac{P(x, y)}{P(x)P(y)}, \quad (4.1)$$

其中 x 和 y 是兩個詞 (點)，各自出現的機率爲 $P(x)$ 及 $P(y)$; $P(x, y)$

是 x 和 y 的聯合機率 (joint probability)

從這個式子的實體意義來看，當事件 x 和 y 的出現是正相關時，則聯合機率 $P(x, y)$ 比起 $P(x)P(y)$ 來的大，因此 $I(x, y) \gg 0$ 。假若此二事件間沒有特殊的關係，表示其爲獨立事件，因此 $P(x, y) \approx P(x)P(y)$ ，即 $I(x, y) \approx 0$ 。反過來說，如果此二事件的出現爲負相關就會使得 $I(x, y) \ll 0$ 。

在派樹抽詞法中要考慮的是任一長度的片段樣式的內聚力，以便作為抽取新詞的考量，因此必須定義大於二字詞的詞內聚力量度。根據簡立峰在1997年的實驗，完整的詞的構成元素及重疊的子片段樣式中有相當高的相關性。因此任意長度片段樣式的內聚力量度，可定義為其重疊子片段樣式的關聯基準量 [18]：

$$\begin{aligned}
 AE_{PAT}(w) &\equiv \frac{P(w)}{P(w_h) + P(w_t) - P(w)} \\
 &= \frac{f(w)}{f(w_h) + f(w_t) - f(w)} \\
 &= \frac{\frac{C(w)}{N_C}}{\frac{C(w_h)}{N_C} + \frac{C(w_t)}{N_C} - \frac{C(w)}{N_C}} \\
 &= \frac{C(w)}{C(w_h) + C(w_t) - C(w)}, \quad (4.2)
 \end{aligned}$$

其中 $w = c_1, c_2, \dots, c_n$ 是目標片段樣式 (候選詞); w_h 是 w 的最長斷尾子字串 $w_h = c_1, c_2, \dots, c_{n-1}$; w_t 是 w 的最長斷頭子字串 $w_t = c_2, c_3, \dots, c_n$ 。 $f(\cdot)$ 和 $C(\cdot)$ 分別表示事件 \cdot 的相對頻率以及頻率。

本論文中的派樹抽詞法將使用 $AE_{PAT}(w)$ 作為抽詞法中量度各片段樣式的內聚力的公式。在4.3節中，迭代式組合式抽詞法採用的內聚力量度和派樹抽詞法最大的不同在於，迭代式組合式抽詞法的內聚力量度是探討兩個相鄰詞之間的內聚力，可以想像成是兩個原本相鄰但不連接的單位，如果其黏著力 (內聚力) 夠強便可考慮將其合起來成爲一新詞。迭代式組合式抽詞法使用的相鄰詞內聚力量度，將在4.3.1節中討論。

4.2.2 片段樣式之左右文相依性

片段樣式的左右邊界是否完整，可由其左右文相依性判斷。當相依性高時，代表其邊界之完整性低。

令 w 表示一段樣式 (候選詞), 定義 w 之左相連詞集合為 $L_w = \{w_l : C(w_l, w) > 0\}$, 右相連詞集合為 $R_w = \{w_r : C(w, w_r) > 0\}$ 。則「左文相依性 (left context dependency, LCD)」及「右文相依性 (right context dependency, RCD)」的量度法定義如下:

- w 具左文相依性(LCD), 若

$$|L_w| < t_f \quad \text{或} \quad \max_{w_l \in L_w} \frac{f(w_l, w)}{f(w)} > t_s, \quad (4.3)$$

其中 $|L_w|$ 為相異的左相連詞的個數, $w_l \in L_w$ 為左相連詞, t_f, t_s 為左右文相依性門檻值, $|L_w| = 0$ 視為 $|L_w| = \infty$ 。

- w 具右文相依性(RCD), 若

$$|R_w| < t_f \quad \text{或} \quad \max_{w_r \in R_w} \frac{f(w, w_r)}{f(w)} > t_s, \quad (4.4)$$

其中 $|R_w|$ 為相異的右相連詞的個數, $w_r \in R_w$ 為右相連詞, t_f, t_s 為左右文相依性門檻值, $|R_w| = 0$ 視為 $|R_w| = \infty$ 。

式(4.3)和式(4.4)兩個條件使用來確認 w 在語料庫中是否有完整的詞邊界。詞邊界完整性的判定, 是藉由考慮 w 這個片段樣式的左右文資訊, 來判斷其使用時的自由度。這個判定法的基本假設是: 如果 w 的左/右相連詞不多, 或者 w 有很高的比例和某些特定的左/右相連詞同時出現, 則 w 和左/右文有相依性, 因此在語意上是不完整的。從圖 4.1 中可看出, 片段樣式的左右變化性夠豐富時, 暗示著有完整的邊界。

在這裡值得注意的是, 處理過程中將 $|L_w| = 0$ 和 $|R_w| = 0$ 視為 $|L_w| = \infty$ 和 $|R_w| = \infty$ 。原因是本論文中的語料庫前處理僅考慮中文文字, 會將任何非中文字或標點符號視為片段樣式的結束, 並開始計算下一個新的片段樣式。因此左右相連詞數為

零的狀況，其實暗示著一個明顯詞邊界 (trivial boundary)。關於此一修正，亦可參考 [42, 43]。

4.2.3 派樹抽詞法需調整之參數

本論文中使用在語音辨識系統中的派樹抽詞法，和前人提出使用在資訊檢索上的抽詞法，不同點之一在於僅考慮目標片段樣式 (候選詞) 的完整性，而不考慮顯著性，其原因在 4.2 節一開始已提過。另一不同點在於完整性中的左右相依性，將明顯詞邊界考慮進去，避免了某些謬拒的情形。綜合 4.2.1 及 4.2.2 兩節，在本論文使用的派樹演算法中，需調整的參數包括有：

- 相異左/右相連詞個數門檻值 t_f
- 最高特定左/右相連詞比例門檻值 t_s
- 片段樣式內聚力(重疊子片段樣式之關聯基準量) t_{MI}

調整這三個參數會使得抽取出來的詞數、品質有所不同，關於參數的調整及詞數、辭典品質變化等議題，在本章的實驗部分將會有進一步的探討。

4.3 迭代式組合式抽詞法

在 4.2 節中討論派樹抽詞法時，討論到片段樣式的內聚力量度時採用的是可量度任意長度的片段樣式的量度法式 (4.2)，然而從資訊理論中最原始的兩點式相互資訊 (式 (4.1)) 中，可以發現該公式是適用於量度二詞 (點) 之間的黏合力。事實上，許多內聚力量度法均只考慮雙元複合詞組的內聚力。若希望能善加利用這些雙元的量度，但同時又希望能抽取出不限長度的多元複合詞，就必須提出不同於派樹抽詞法的架構。在 [42, 43] 中，

提出了迭代式組合式抽詞法，其迭代的特性使得此抽詞法可以抽出多元複合詞，而其每次組合雙元詞的特性，讓使用此抽詞法時，可以套入各種不同雙元內聚力量度。

4.3.1 相鄰詞內聚力量度

關於相鄰詞內聚力量度，因為可較直觀的套用資訊理論的相互資訊概念，因此其相關研究較豐富。首先是邱氏及漢氏 (K.W. Church and P. Hanks) 所提出的詞關連基準量 (word association norms) [44]，是以有序性相互資訊作為量度。此量度實為兩點式相互資訊 (式(4.1)) 的一個延伸：

$$AE_d(w_i, w_j) \equiv \log_2 \frac{P_d(w_i, w_j)}{P(w_i)P(w_j)}, \quad (4.5)$$

式(4.5)以 $P_d(w_i, w_j)$ 替換掉式(4.1)具對稱性 (symmetric) 的聯合機率 $P(x, y)$ ， $P_d(w_i, w_j)$ 是藉由計算 $f_d(w_i, w_j)$ 在整體語料庫中的比例得到的，其代表 w_i 出現後，取一長度為 d 的窗，出現 w_j 的機率。此窗的大小是一個可調整的參數，當時邱氏及漢氏提出：小的窗可以用以習得固定用語 (如俚語)，大的窗則可以捕捉到遠距離的語意或關係。在該研究中，提出以 $d = 5$ 為一個折衷的窗大小。在本論文中，此量度將被用來討論兩相鄰詞的關連性，因此我們取用的是 $d = 0$ 。值得注意的是 $P_{d=0}(w_i, w_j)$ 其中 w_i, w_j 仍是有序的參數，所以 $P_{d=0}(\cdot, \cdot)$ 為一非對稱的函數。因此，式(4.5)即為有序性相互資訊。之後的討論，將以 $AE(w_i, w_j) = AE_{d=0}(w_i, w_j)$ 。

解讀 $AE(w_i, w_j)$ 時，仍可以承續兩點式相互資訊的特性，當值越高，表示若詞 w_i 出現，之後立刻出現詞 w_j 的機率越高，表示詞組 w_i, w_j 有著越高的關聯性，亦即此二相鄰詞為一複合詞的機會越高。

除了式(4.5)以有序性相互資訊作為詞關連基準量以外，畢氏與賈氏 (C. Beaujard

and M. Jardino) 提出一種相互資訊法的變形，稱之為相互機率 [45]:

$$MP(w_i, w_j) \equiv \frac{P(w_i, w_j)}{\sqrt{P(w_i)P(w_j)}} \quad (4.6)$$

$$\begin{aligned} &= \sqrt{\frac{P(w_i, w_j)}{P(w_i)} \frac{P(w_i, w_j)}{P(w_j)}} \\ &= \sqrt{P(w_j|w_i)P(w_i|w_j)} \end{aligned} \quad (4.7)$$

式(4.6)是相互機率的定義，從中可以發現和兩點式相互資訊法最大的差別是分母項做了二次方根，這個計算的巧妙之處，在於式(4.6)可以轉換成式(4.7)，相當於是詞組 w_i, w_j 的前向機率 (forward probability) $P(w_j|w_i)$ 和後向機率 (backward probability) $P(w_i|w_j)$ 的幾何平均。更巧妙的是，由於相互機率是兩個機率值相乘後開二次方根，因此其值域即為一般機率值的值域。這一點比起傳統的相互資訊流派方法來的好，較容易調整其門檻值參數。

除了有序性的詞關連基準量、值域正規化的相互機率以外，另有一類似在4.2.1節中提到的內聚力量度法：張氏等人的「相互資訊」量度 [46]，基本上類似式(4.2)的架構，但處理單位由重疊字串改成相鄰的詞（這並不符合相互資訊的原始定義，但由於原文如此稱呼，因此在這裡沿用此稱呼）：

$$MI(w_i, w_j) \equiv \frac{C(w_i, w_j)}{C(w_i) + C(w_j) - C(w_i, w_j)}. \quad (4.8)$$

以上的三種計算相鄰詞組內聚力的量度法——邱氏的詞關聯基準量 (式(4.5))、張氏的相互資訊 (式(4.8)) 以及畢氏的相互機率 (式(4.6))——在廖碩鵬的碩士論文中，有抽詞及混淆度的分析比較 [42]。在該論文中，實驗的結果說明此三種量度的優劣結果是 $MI > MP \gg AE$ 。

4.3.2 左右文變異性統計 (Context Variation Statistics)

迭代式組合式抽詞法中，同樣也要使用到判斷邊界是否完整的方法。在迭代式組合式抽

詞法中採用的方法，和4.2.2節中，派樹抽詞法中的片段樣式左右文相依性只有一點些微的差異，那就是左右文變異性統計此種量度否定了最高特定左/右相連詞比例門檻值 t_s 的用途，換句話說，此種量度認為，在抽取語音辨識系統所用之辭典時，判斷候選詞左右邊界的完整性只需要用相異左/右相連詞個數門檻值 t_f 即可。根據式(4.3)及式(4.4)，拿掉 t_s 的限制，我們就得到左右文變異性統計 [43, 42]:

$$\text{若} \quad \begin{cases} |L_w| \geq t_{cvs} & \text{或} & |L_w| = 0 \\ |R_w| \geq t_{cvs} & \text{或} & |R_w| = 0, \end{cases} \quad (4.9)$$

則 w 有完整邊界，其中 t_{cvs} 為左右文變異性統計門檻值。

4.3.3 迭代式組合式抽詞法整體架構

雖然和派樹抽詞法一樣，都考慮了內聚力和邊界兩個統計特徵，但迭代式組合式抽詞法不同處在於，其採用的是相鄰詞的內聚力量度，因此每次只考慮一相鄰詞對是否可以黏合而成為一個新詞。在這樣的設定之下，爲了要能同時兼顧能抽出多元詞，因此迭代式組合式抽詞法採用了如圖 4.2 的抽詞流程。

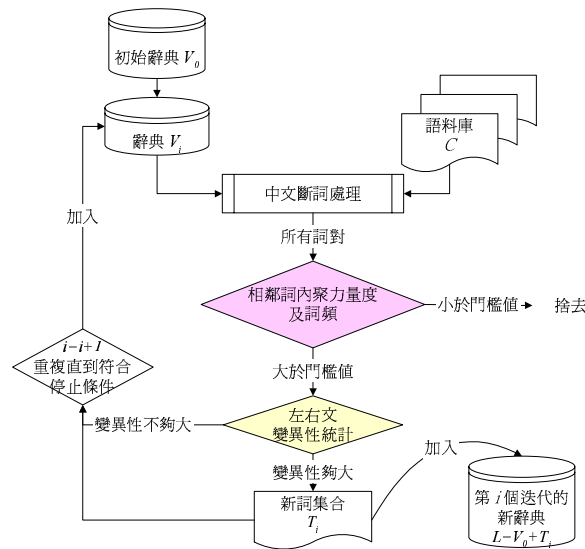


圖 4.2: 迭代式組合式抽詞法流程

剛開始的時候，存在一個語料庫 C 以及一個初始辭典(seed lexicon) V_0 。在任一第 i 次的迭代中，語料庫 C 都會使用當時的辭典 V_i ，以最大匹配斷詞演算法 [47] 進行斷詞。接下來，對斷詞語料庫中的任一詞對 (w_i, w_j) ，計算其相鄰詞內聚力量度及出現頻率，假若都超過門檻值，則此詞對會被加入一專門存放候選詞的優先權佇列 Q (依照其內聚力高低排序)。接著，從 Q 中選取可通過左右文變異性統計門檻值的前 n 個詞對，加入新詞集合 T_i 。因此，每一次迭代得到可供下一次迭代使用的新辭典是 $V_{i+1} \leftarrow V_i + T_i$ 。最後，當停止條件符合時，所有抽出的新詞都會包含在最後的辭典裡。

4.4 實驗結果與比較

本章的實驗討論的重點在於不同抽詞法、不同參數的調整，其最後得到的辭典的各種統計特性 (包括總詞數、平均詞長、最長詞字數等等)，以及這些辭典在辨識實驗上的效能好壞。因此，本章中的所有辨識實驗使用的語言模型，均採用同一套語料庫作為訓練，並且不與基礎語言模型作線性內插。根據第三章的實驗，我們知道對辨識率最有幫助的是採用時間完全重疊的語料，因此以下的實驗的文字語料部分，均採用2002年9月的雅虎奇摩新聞。

4.4.1 一字詞辭典及基礎辭典實驗

本章中的第一個實驗目的是為了建構往後抽詞實驗的對照標準。首先先以2.3.4節中提過的辨識用基礎辭典，以長詞優先斷詞法將2002年9月的雅虎奇摩新聞斷詞，並訓練其語言模型。

除此之外，為了分析多字詞對辨識結果的幫助，本節中同時實驗了一個僅含有一字詞的辭典，其辨識率的結果。這裡使用的一字詞辭典是將辨識用基礎辭典 (61,522 詞) 及一般語言處理使用之中文詞庫 (54,475 詞) 中所有相異的單字抽取出來，成為一含有

辭典	辨識效能		辭典統計特性			
	字 錯誤率	次音節 錯誤率	總詞數	平均詞長	最長詞長	詞長 標準差
一字詞辭典	22.44	13.15	6193	1.00	1	0.00
辨識用基礎辭典	19.39	11.54	61522	2.34	6	0.87

表 4.1: 一字詞辭典、辨識用基礎辭典的辨識效能及辭典統計特性

6,194 詞的一字詞辭典。

這兩個辭典分別以 2002 年 9 月的雅虎奇摩新聞作訓練語料，訓練得到的語言模型用以辨識 2002 年 9 月的 News98 新聞 506 則，得到的辨識效能（字錯誤率、次音節錯誤率）如表 4.1 所示。

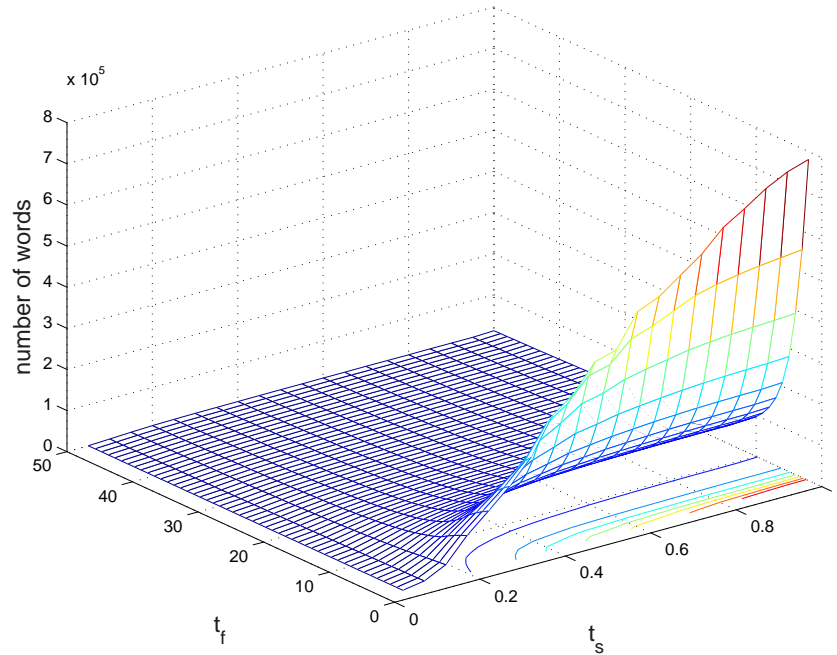
4.4.2 片段樣式左右文相依性之二門檻值— t_f 與 t_s

在 4.2 節，派樹抽詞法的左右文相依性的探討中有兩個重要的參數，其一是片段樣式左/右所連接相異的詞數，另一個是片段樣式左/右各個相連詞中，最多的相連詞與總出現次數的比例。使用派樹抽詞法時，必須針對這兩個參數設定門檻，也就是 4.2.2 節中所提到的前兩個參數：

1. 相異相連詞個數門檻值 t_f
2. 最高特定相連詞比例門檻值 t_s

首先，爲了瞭解這兩個門檻值的設定，和抽取出詞數的關連性，我們以 2002 年 9 月的雅虎奇摩新聞爲語料，將片段樣式內聚力（重疊子片段樣式之關聯基準量）先放鬆到最寬的標準，觀察此二門檻值與抽出詞數的關係，結果以三維曲面方式顯示在圖 4.3 中， x 軸是 t_s ，而 y 軸是 t_f ，其高度爲抽出的詞數。

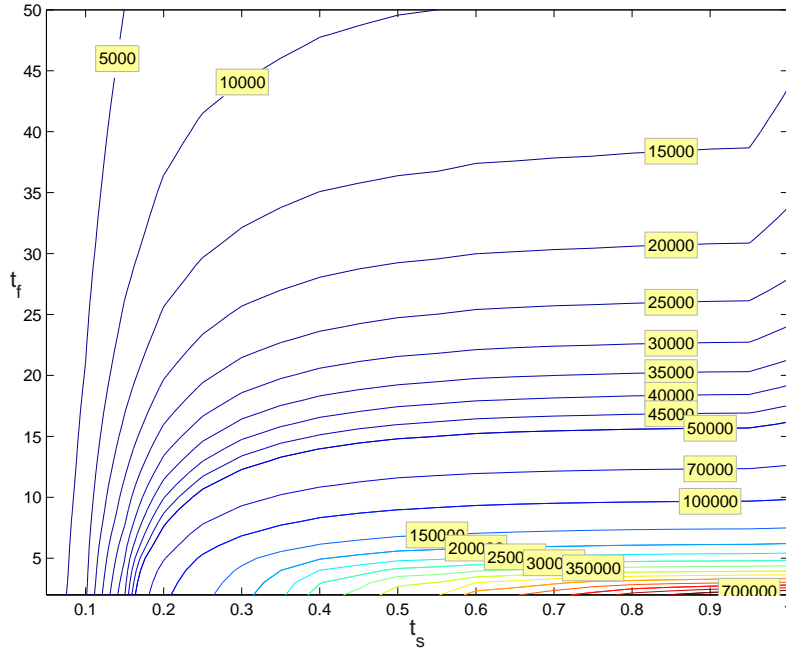
由此曲面圖可以發現， $(t_s, t_f) = (1, 0)$ 時相當於這兩個門檻都不設限，因此得到的詞數是最多的， t_s 越小、 t_f 越大，詞數就越少。若要詳細觀察詞數與此二門檻值的關係，可



x 軸為最高特定左/右相連詞比例門檻值 t_s , y 軸為相異相連詞個數門檻值 t_f , z 軸是在 (t_s, t_f) 的門檻值設定之下抽出的詞數。

圖 4.3: (t_s, t_f) 對抽出之詞數之三維關係圖

從其等高線圖 (圖 4.4) 觀察得知。在這個等高線圖中, 需注意的是等高線的間隔略有不同, 在圖的偏左上及中間的線條 (詞數一萬至十萬) 的間隔是一萬, 而圖的右下方 (詞數二十萬至七十萬) 的詞數間隔是取十萬。這是因為如果取一樣間隔, 將會發現右下方的線條非常的密集。從這個圖觀察 t_s 和 t_f 的交互關係, 從四萬到十萬這七條等高線可以發現, 當 $t_s > 0.5$ 時, 其變化幾乎對抽得的詞數沒有影響。從這個現象我們可以推得: 假若在派樹演算法中採用大於 0.5 的最高特定相連詞比例門檻, 就幾乎跟不對這個參數加以限制的結果是一樣的。當我們看圖 4.4 的左半邊 (即 $t_s \leq 0.5$) 時, 會發現 t_s 的門檻值對抽得的詞數有較明顯的限制作用。因此, 之後的實驗將僅就 $t_s \leq 0.5$ 做討論。此外, 在另一種邊界判定的方法—左右文變異性統計中, 認為最高特定相連詞比例門檻應被廢除 (即 $t_s = 1.0$), 因此在接下來的實驗除了討論 $t_s \leq 0.5$ 以外, 均再多加一組 $t_s = 1.0$ 的實驗, 以便觀察假設不考慮相異左/右相連詞個數門檻值 t_s 時, 辨識率有何變化。



橫軸為最高特定左/右相連詞比例門檻值 t_s ，縱軸為相異相連詞個數門檻值 t_f ，等高線上的標籤代表的是該等高線的詞數。本圖討論的是此二門檻值和抽出的詞數之關係。

圖 4.4: (t_s, t_f) 對抽取出之詞數等高線圖

爲了更深入觀察 (t_s, t_f) 對抽出的詞的品質好壞影響，以下設計一個實驗：用數組不同的 (t_s, t_f) 參數，各抽出大約相等的詞數（即選取某一條詞數等高線上的數個不同點），在這個實驗中，我們選取詞數最接近七萬詞的門檻值設定，包括以下五組： $(t_s, t_f) = (0.20, 5)$ 、 $(t_s, t_f) = (0.30, 10)$ 、 $(t_s, t_f) = (0.40, 11)$ 、 $(t_s, t_f) = (0.50, 12)$ 、 $(t_s, t_f) = (1.00, 13)$ ，由於本論文中的派樹抽詞法，抽出來的片段樣式最短是二字詞，因此除了派樹抽詞法抽出的詞以外，還會加入 4.4.1 節中提到的一字詞辭典。本實驗中跑了五組不同的 (t_s, t_f) 門檻值，所得到的辨識效能及辭典統計特性列於表 4.2。從表 4.2 中可發現幾個趨勢：

1. 當 t_s 門檻取的越嚴格而 t_f 門檻越放鬆時，相當於往圖 4.4 的左下角靠近，從表 4.2 可發現，其抽出來的平均詞長有略微增長的趨勢，詞長標準差也略有上升。這代表

(t_s, t_f)	辨識效能		辭典統計特性			
	字 錯誤率	次音節 錯誤率	總詞數	平均詞長	最長詞長	詞長 標準差
基礎辨識實驗	23.17	13.17	61522	2.34	6	0.87
(0.20, 5)	20.69	12.35	74513	2.99	19	1.25
(0.30, 10)	20.55	12.34	70231	2.73	19	1.15
(0.40, 11)	20.47	12.29	74718	2.68	19	1.13
(0.50, 12)	20.39	12.26	72809	2.63	19	1.12
(1.00, 13)	20.30	12.28	74513	2.99	20	1.25

表中所列之辭典，是用上列 (t_s, t_f) 為門檻值的派樹抽詞法抽詞，然後再加上一字詞辭典而成的。

表 4.2: 鄰近七萬字等高線的數組 (t_s, t_f) 之辨識效能及辭典統計特性

最高特定相連詞比例門檻值 t_s ，並不見得會阻止長詞的產生。相反地，在 t_s 受到限制時，平均詞長還可能有所提升。

- 承上，雖然說 t_s 限制較強時可以抽出平均詞長較長的辭典，但由表 4.2 可以發現，在這種情況下辨識效能卻略微低一點，比不上將 t_s 放鬆而限制 t_f 的狀況。其原因推論可能是因為這條條件會造成某些候選詞的謬拒 (false rejection)。如候選詞「陳水扁」的右連接詞在近期的大量語料統計，最常出現的是「總統」，雖然還有許多不同的右連詞，但「總統」此一最高比例的特定右連詞之比例高過門檻，而使其有著與右文相依的特性而不被視為一完整的詞。在語音辨識應用的情況之下，這種謬拒的狀況可能會使語音辨識的錯誤率升高，

除了加上一字詞以外，同時考慮另一種實驗設計法：將派樹抽詞法抽出的詞視為額外的新詞，加入辨識用基礎辭典中。在這樣的辭典之下，辨識的結果顯示在表 4.3。觀察表 4.2 和表 4.3 會發現，以辨識用基礎辭典加上派樹抽詞法抽出的新詞之實驗，其效能比起僅用一字詞辭典加上派樹抽詞法抽出的新詞略勝一籌。其原因是因為除了統計式方法抽得的新詞以外，再加上經由人選過的基礎辭典，可以補足統計方法未能抽出但對辨識

(t_s, t_f)	辨識效能		辭典統計特性			
	字 錯誤率	次音節 錯誤率	總詞數	平均詞長	最長詞長	詞長 標準差
基礎辨識實驗	23.17	13.17	61522	2.34	6	0.87
(0.20, 5)	19.82	11.98	117262	2.82	19	1.13
(0.30, 10)	19.87	11.98	109418	2.67	19	1.05
(0.40, 11)	20.13	12.12	112003	2.64	19	1.04
(0.50, 12)	20.01	12.07	109589	2.61	19	1.03
(1.00, 13)	19.98	12.07	109855	2.73	20	1.55

表中所列之辭典，是用上列 (t_s, t_f) 為門檻值的派樹抽詞法抽詞，然後再加上辨識用基礎辭典而成的。

表 4.3: 鄰近七萬字等高線的數組 (t_s, t_f) 之辨識效能及辭典統計特性

(t_s, t_f)	辨識效能		辭典統計特性			
	字 錯誤率	次音節 錯誤率	總詞數	平均詞長	最長詞長	詞長 標準差
基礎辨識實驗	23.17	13.17	61522	2.34	6	0.87
(0.20, 37)	20.70	12.19	15987	2.04	19	1.45
(0.30, 45)	20.36	11.93	15912	2.01	19	1.45
(0.40, 48)	20.69	12.14	16106	2.04	19	1.51
(0.50, 50)	20.44	11.95	16059	2.05	19	1.54
(1.00, 66)	20.30	11.81	16147	3.08	20	3.43

表中所列之辭典，是用上列 (t_s, t_f) 為門檻值的派樹抽詞法抽詞，然後再加上一字詞辭典而成的。

表 4.4: 鄰近一萬字等高線的數組 (t_s, t_f) 之辨識效能及辭典統計特性

有幫助的詞，

此外，除了抽出七萬詞的實驗以外，為了瞭解新詞詞量較少時的實驗結果，另外設計了一個抽出一萬詞的實驗。在這個實驗中，我們選取五組詞數最接近一萬詞的門檻值設定： $(t_s, t_f) = (0.20, 37)$ 、 $(t_s, t_f) = (0.30, 45)$ 、 $(t_s, t_f) = (0.40, 48)$ 、 $(t_s, t_f) = (0.50, 50)$ 、 $(t_s, t_f) = (1.00, 66)$ 。和七萬字的一樣，抽出一萬字的實驗也是分別和一字詞辭典以及辨識用基礎辭典混合後，分別顯示在表 4.4 及表 4.5。

綜合以上兩種詞數的設定（一萬詞及七萬詞），以及分別加入一字詞辭典，或是加入

(t_s, t_f)	辨識效能		辭典統計特性			
	字 錯誤率	次音節 錯誤率	總詞數	平均詞長	最長詞長	詞長 標準差
基礎辨識實驗	23.17	13.17	61522	2.34	6	0.87
(0.20, 37)	20.58	12.07	65367	2.40	19	1.02
(0.30, 45)	19.36	11.44	64828	2.40	19	1.03
(0.40, 48)	19.40	11.45	64857	2.40	19	1.05
(0.50, 50)	19.47	11.50	64840	2.41	19	1.06
(1.00, 66)	19.42	11.39	66527	2.65	20	1.85

表中所列之辭典，是用上列 (t_s, t_f) 為門檻值的派樹抽詞法抽詞，然後再加上辨識用基礎辭典而成的。

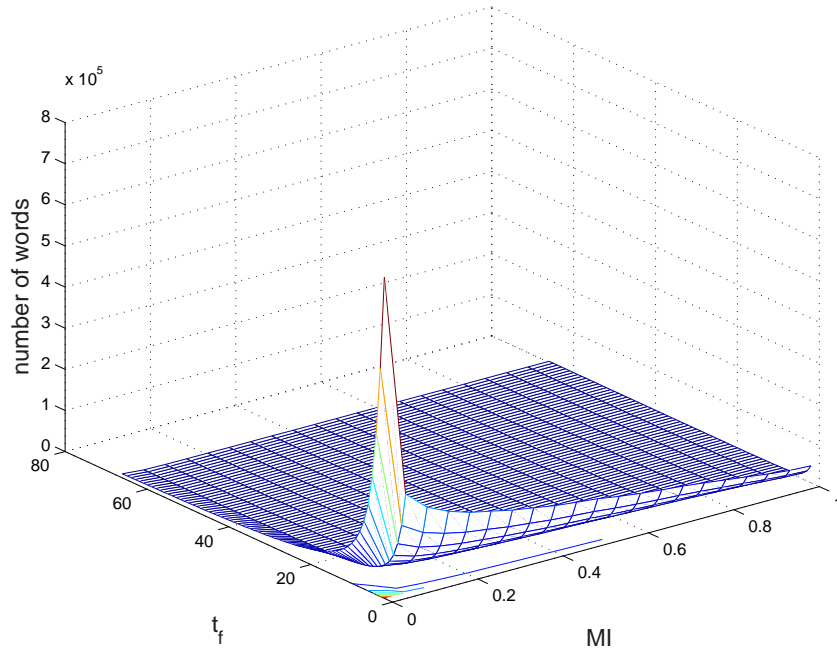
表 4.5: 鄰近一萬字等高線的數組 (t_s, t_f) 之辨識效能及辭典統計特性

辨識用基礎辭典兩種不同的實驗，可以得到以下幾個結論：

1. 無論是抽出約一萬詞或七萬詞，加上辨識用基礎辭典的辨識效能都比僅加上一字詞辭典的辨識效能來的好。如果考慮到詞數的影響，比較表 4.2 和表 4.5 也可以發現，抽出一萬新詞加入辨識用基礎辭典，得到約六萬五千個詞，但其辨識效能還是比抽出七萬新詞加入一字詞辭典得到的七萬詞左右來的高。這說明了僅使用統計式抽詞法的抽出來的詞，雖然對辨識率有幫助，但是完全用統計方法抽出來的詞，其效能還是不會比加入一原本就存在的基礎辭典來的好。
2. 抽出一萬詞和七萬詞的設定，加上一字詞辭典時辨識效能幾乎是完全一樣的，而再加上辨識用基礎辭典時，一萬詞的表現甚至略高一些。這說明了其實詞並不是抽越多越好，也顯示了抽到七萬詞時，其多抽的六萬詞對於辨識效能的幫助已經微乎其微了。

4.4.3 片段樣式之相連詞個數門檻值 t_f 與重疊子片段樣式之關聯基準量 (內聚力) 門檻值 t_{MI}

在 4.4.2 節中，我們討論過 t_s 和 t_f 的相互關連性，這兩個門檻值都是針對邊界完整性的



x 軸為重疊子片段樣式之關聯基準量 (內聚力) 門檻值 MI , y 軸為相異相連詞個數門檻值 t_f , z 軸是在 (MI, t_f) 的門檻值設定之下抽出的詞數。

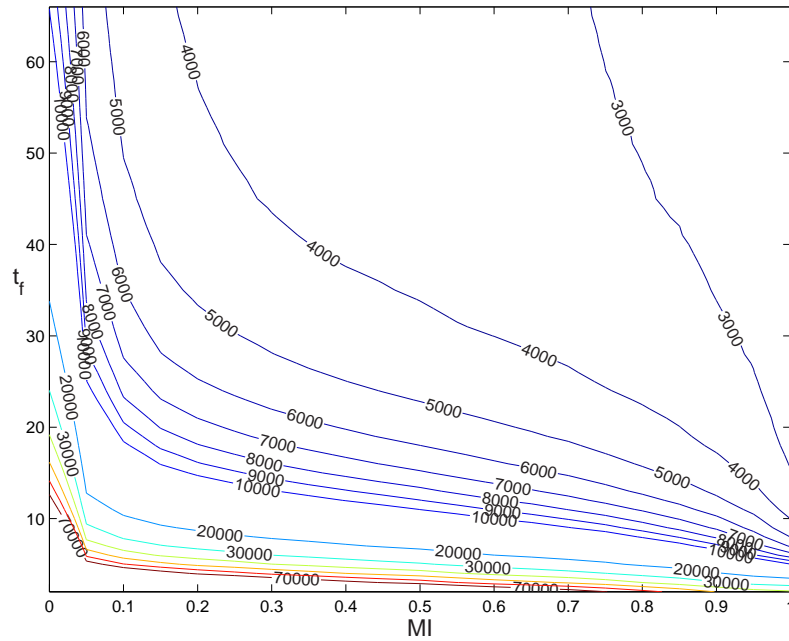
圖 4.5: (MI, t_f) 對抽取出之詞數之三維關係圖

判斷。在本節中, 將把重點轉移到同時更動邊界完整性判定門檻值以及內聚力門檻值時, 對於抽出詞數及辨識效能的影響。因此在本節中, 將針對片段樣式之相異相連詞個數門檻值 t_f 與重疊子片段樣式之關聯基準量 (內聚力) 門檻值 t_{MI} 這兩個參數的關係作比較分析。

首先, 為了瞭解這兩個門檻值的設定, 和抽取出詞數的關連性, 我們以 2002 年 9 月的雅虎奇摩新聞為語料, 觀察此二門檻值與抽出詞數的關係¹。結果以三維曲面方式顯示在圖 4.3 中, x 軸是 MI , 而 y 軸是 t_f , 其高度為抽出的詞數。

圖 4.5 中抽出詞數最多的點是 $(MI, t_f) = (0, 0)$ 時, 相當於這兩個門檻都不設限。 MI 及 t_f 越大, 詞數就越少。為了詳細觀察詞數與此二門檻值的關係, 另外繪製一等高線圖 (圖 4.6) 觀察得知。

¹最高特定相連詞比例門檻值 t_s 放鬆到最寬的標準 (設定為 1.0)



橫軸為重疊子片段樣式之關聯基準量(內聚力) 門檻值 MI ，縱軸為相異相連詞個數門檻值 t_f ，等高線上的標籤代表的是該等高線的詞數。本圖討論的是此二門檻值和抽出的詞數之關係。

圖 4.6: (MI, t_f) 對抽取出之詞數等高線圖

由於在前一節中，我們知道抽出一萬詞的效能比起抽出七萬詞來的好，因此本節中，我們用派樹抽詞法，用數組不同門檻值設定抽出約一萬詞，並分別加入一字詞辭典及辨識用基礎辭典做辨識實驗。選取四組抽出詞數最接近一萬詞的門檻值設定： $(MI, t_f) = (1.00, 5)$ 、 $(MI, t_f) = (0.60, 10)$ 、 $(MI, t_f) = (0.20, 14)$ 、 $(MI, t_f) = (0.00, 66)$ 。抽出約一萬字後，分別和一字詞辭典以及辨識用基礎辭典混合後，分別顯示在表 4.6 及表 4.7。

從表 4.6 和表 4.7 中的辭典統計特性，我們發現加入重疊子片段樣式之關聯基準量(內聚力) 的限制後，抽出的平均詞長變長。這是因為當詞長變長時，其出現次數及重疊子片段樣式的出現次數都比較少，也比較容易出現重疊子片段樣式之關聯基準量等於一的情形。(即斷頭或去尾的子片段樣式出現的次數，都恰好和原詞的出現次數一樣多。)此外，在這裡也觀察到，加入辨識用辭典的辨識效能略勝一籌，這個現象和 4.4.2 節得到的

(MI, t_f)	辨識效能		辭典統計特性			
	字 錯誤率	次音節 錯誤率	總詞數	平均詞長	最長詞長	詞長 標準差
基礎辨識實驗	23.17	13.17	61522	2.34	6	0.87
(1.00, 5)	22.69	13.28	16211	4.21	20	3.57
(0.60, 10)	22.00	12.91	16160	3.88	20	3.44
(0.20, 14)	21.95	12.97	16759	3.75	20	3.35
(0.00, 66)	20.30	11.81	16147	3.08	20	3.43

表中所列之辭典，是用上列 (MI, t_f) 為門檻值的派樹抽詞法抽詞，然後再加上一字詞辭典而成的。

表 4.6: 鄰近一萬字等高線的數組 (MI, t_f) 之辨識效能及辭典統計特性

(MI, t_f)	辨識效能		辭典統計特性			
	字 錯誤率	次音節 錯誤率	總詞數	平均詞長	最長詞長	詞長 標準差
基礎辨識實驗	23.17	13.17	61522	2.34	6	0.87
(1.00, 5)	19.61	11.56	71051	2.87	20	1.97
(0.60, 10)	19.35	11.46	70422	2.79	20	1.88
(0.20, 14)	19.44	11.49	70419	2.77	20	1.85
(0.00, 66)	19.42	11.39	66527	2.65	20	1.85

表中所列之辭典，是用上列 (MI, t_f) 為門檻值的派樹抽詞法抽詞，然後再加上辨識用基礎辭典而成的。

表 4.7: 鄰近一萬字等高線的數組 (MI, t_f) 之辨識效能及辭典統計特性

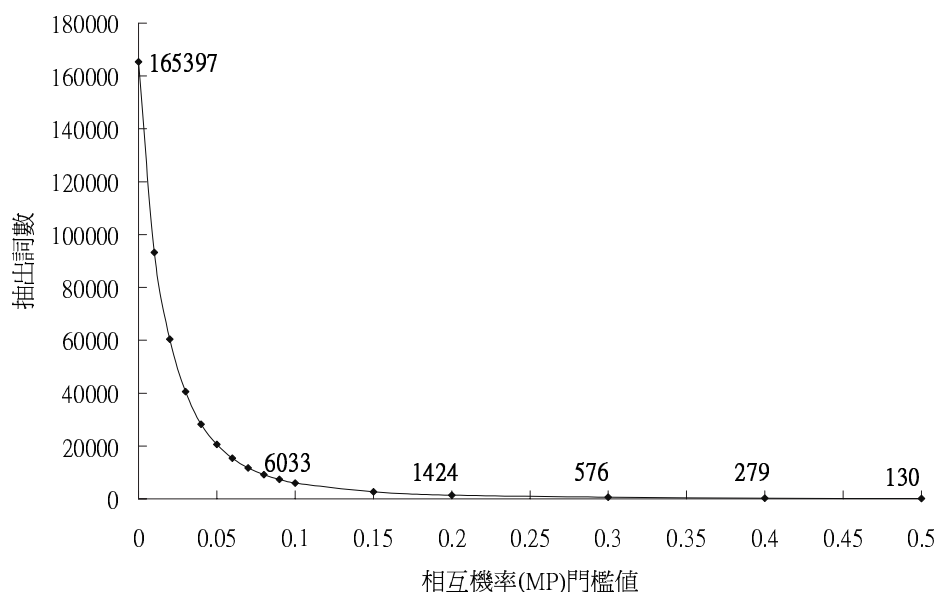


圖 4.7: 相互機率與抽出詞數關係圖

結論是一致的。

4.4.4 迭代式組合式抽詞法之實驗

在廖碩鵬的論文中，曾提過詞內聚力量度法的表現，為 $MI > MP \gg AE$ ，但 MI 與 MP 差距不大，因此建議以相互資訊內聚力量度 MI (式(4.1)) 或相互機率內聚力量度 MP (式(4.6)) 為詞內聚力量度法。在這部分的實驗中，將採用相互機率 (MP) 為詞內聚力量度法，原因是使用相互機率較容易設定其門檻值，因其已正規化至 $[0, 1]$ 的範圍。

迭代式組合式亦有許多參數需要調整，在本論文的迭代式組合式抽詞法的實驗中，針對雅虎奇摩2002年9月的中文新聞做抽詞，候選詞頻率門檻值設為10，左右文變異性統計門檻設為5，調整相互機率以得到不同的詞數，相互機率和詞數的變化如圖4.7所示。從圖4.7可以發現，如果想抽出一萬詞，則必須要調整相互機率門檻值在0及0.1之間，因此我們另外將這之間的相互機率語詞數變化關係，部分放大出來看，如圖4.8所示。

為了和先前的派樹抽詞法做比較，因此使用迭代式組合式抽詞法時，將相互機率門檻

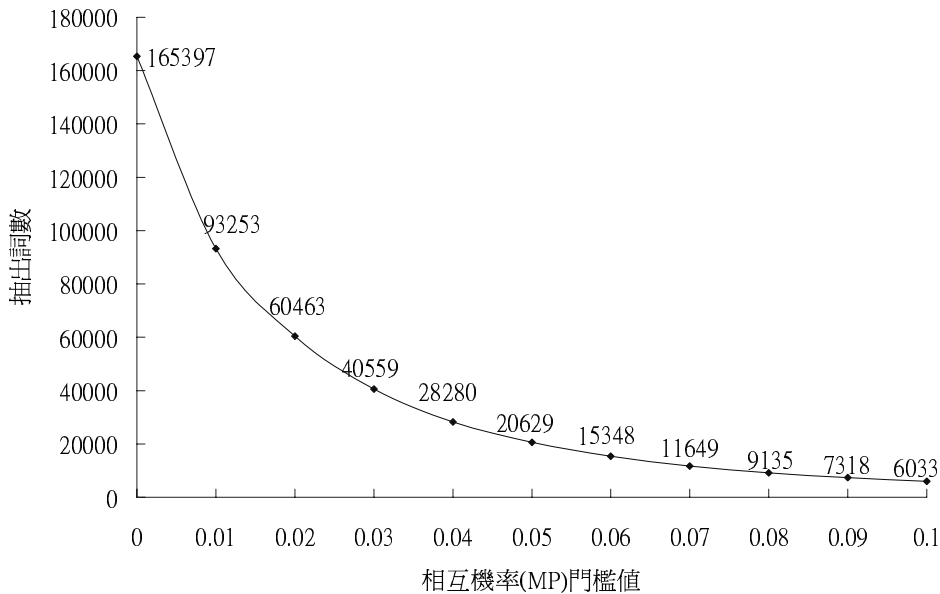


圖 4.8: 相互機率與抽出詞數關係圖(部分放大)

辭典	辨識效能		辭典統計特性			
	字 錯誤率	次音節 錯誤率	總詞數	平均詞長	最長詞長	詞長 標準差
基礎辨識實驗	23.17	13.17	61522	2.34	6	0.87
派樹抽詞法	19.35	11.46	70422	2.79	20	1.88
迭代式組合式抽詞法	19.03	11.37	70479	2.58	17	1.27

表 4.8: 迭代式組合式抽詞法及派樹抽詞法得到之辭典和基礎辨識實驗的辨識效能比較

值設定在 0.08, 以取得約一萬詞的新詞抽取量。在候選詞頻率門檻值設為 10, 左右文變異性統計門檻設為 5, 相互機率門檻設為 0.08, 並以辨識用基礎辭典作為初始辭典進行五次迭代抽詞的狀況之下, 所得到的新辭典及其辨識效能如表 4.8 所示。

從表 4.8 可以發現, 在使用派樹抽詞法和迭代式組合式抽詞法, 同樣抽出約一萬新詞並混合辨識用基礎辭典時, 在辨識效能上還是迭代式組合式抽詞法較為優越, 推測其原因可能是因為迭代式組合式抽詞法經由多次迭代, 可以組合出對於語音辨識應用中有意義的長詞, 雖然並不見得是符合傳統詞的定義的詞, 但有意義的長詞的確對正確的語音辨識有幫助。而且, 由於迭代式組合式抽詞法是以初始辭典為基礎, 因此抽出一萬左右

的新詞量，再加上本來的辭典就有約七萬詞，表 4.8 中列出的派樹抽詞法，是選用在前面 4.4.2 及 4.4.3 兩節中表現最好的一個，但其辨識效能仍落後迭代式組合式抽詞法的效能，推論其原因是派樹抽詞法並無根據一初始辭典，雖然最後有混入辨識用基礎辭典，但其增加出來的新詞，並未根據辨識用基礎辭典作為初始的知識，反觀迭代式組合式抽詞法，不但善用了原先存在的知識（初始辭典），而且其邊界判定法及內聚力門檻都是針對語音辨識的應用調整過的（例如採用左右文變異性統計以便能同時抽出「陳水扁」及「陳水扁總統」來）。因此能額外抽取出對於辨識有幫助的新詞來。

4.5 本章結論

1. 在語音辨識的應用中，無論是派樹抽詞法或是迭代式組合式抽詞法，最好的方法都是將抽出的新詞加入原本已存在的、有經由人工選取的辭典中。在迭代式組合式抽詞法中，由於其演算法的設計，因此初始辭典本來就會是最後產生的辭典的子集合；但派樹抽詞法中沒有採用初始辭典的概念，又派樹抽詞法可以抽出大量具有統計上意義的詞，但假若直接使用這些詞，而不加入一人工處理過的辭典的話，實驗結果會不盡理想（見表 4.2）。如果將派樹抽出的新詞能加入一辭典（例如本論文中的辨識用基礎辭典），再進行語音辨識的話，才能使其辨識效能比基礎實驗的效能進步（見表 4.3）。
2. 派樹抽詞法的邊界判定，其實可以僅使用片段樣式之相異相連詞個數門檻值 (t_f) 即可，最高特定相連詞比例門檻值 (t_s) 可以不用。因為當 t_s 設定太緊時反而使辭典品質下降，而設定太鬆 (> 0.5) 時又意義不大。甚至在 4.4.2 節中的實驗可以發現，限制 t_s 時，其辨識率不會比完全不考慮 t_s (即 $t_s=1.0$) 有明顯的進步。這一個觀察也證實了在迭代式組合式抽詞法中，用以判定邊界完整性的左右文變異性統

計僅使用相異相連詞個數門檻值的合理性。

3. 在派樹抽詞法中片段樣式之相異相連詞個數門檻值 (t_f) 與重疊子片段樣式之關聯基準量 (內聚力) 門檻值 (t_{MI}) 的關係中, 可以發現派樹抽詞法雖然同時也使用了邊界判定及內聚力量度, 對辨識率的增進仍不如迭代式組合式抽詞法。原因可能是因為斷頭去尾式的重疊子片段樣式之關聯基準量, 會造成很長的詞容易被選出來, 但完全偏好長詞而排除短詞的結果是, 當目標語音語料中出現了不同的短詞組合時, 就不容易辨識出來。例如先前提過的「陳水扁總統」一例, 最理想的狀況是能同時讓「陳水扁」及「陳水扁總統」和「總統」都存在於辭典之中。然而太偏好長詞的結果, 可能造成「陳水扁總統」一詞存在, 但排除了「陳水扁」。如此一來當目標語音語料中出現「陳水扁」但後面不接「總統」時, 便必須回到以一字詞或二字詞才有可能正確辨識出來。因此, 長詞固然對語音辨識有幫助, 但必須一個好的語音辨識辭典必須有適當的長短詞組合, 才能真正達到好的辨識效果。
4. 迭代式組合式抽詞法的優點是, 其抽出的詞是根基於一現存的辭典, 抽出的詞是根據原先就已存在的詞彙再去成長, 因此抽出來的新詞, 加入原辭典中使用的加成效果, 可能會比派樹抽詞法來的好。再加上先前在第1點中曾提到, 在語音辨識系統中的辭典, 最好是能將抽詞法抽出的新詞加入一現存的辭典, 可以有效的增進其效能。因此推論迭代式組合式抽詞法是比較適合語音辨識系統使用的抽詞法。

第五章

語言模型強化之整合研究

在第三章及第四章中，分別提出了以語料庫精緻化及辭典精緻化，來提升語音辨識系統的效能。之前的實驗都是固定一項變因，討論另一項的改變會造成整體效能的何種變化。但在實際將這些語言模型強化的方法使用在真實問題時，所有的方法都是可以加成並相互影響的，我們可以從時間/主題匹配的語料中抽取新詞，也可以同時用時間/主題匹配的語料作語言模型訓練及調適，也可以用現有的辨識結果，建構查詢指令並收集衍生語料庫等等。將這些語言模型強化的元素加成起來，如此一來辨識率的進步將更為顯著 [43, 48]。

除此之外，由於之前的實驗都是做在廣播新聞語料上，本章中亦嘗試將前面各章節提過的語言模型強化法使用在訪談語料上，因此本章將著重在針對兩個特性迥異的目標語音語料集——廣播新聞及訪談語料——進行語言模型的強化，來觀察其辨識效能會有怎樣的變化。

5.1 廣播新聞語料

5.1.1 廣播新聞語料之特性

很多從事語音辨識研究的學者，選用廣播新聞語料作為目標語料，其原因是廣播新聞較容易有系統地大量收集，並且主播講話通常相當清楚，且廣播新聞接近朗讀式語音 (read

speech), 再加上新聞的內容和時事有關, 而時事又會同時以文字新聞的型態存在。再者, 廣播新聞的辨識對於新聞資料庫的建立、檢索等等都有相當大的助益, 因此其重要性也不可忽視。以上種種的因素, 讓廣播新聞辨識成爲一個熱門且重要的研究題目。本章中使用的廣播新聞語料, 承續之前各章使用的同一個語音語料, 也是2.3.3節中提到的語音語料—2002年9月的506則新聞。

5.1.2 同時精緻化語料庫及辭典以強化語言模型之實驗

在第三章中, 由於討論的是語料庫精緻化, 各辨識實驗均爲固定使用辨識用基礎辭典; 在第四章中, 因爲討論的重點是辭典精緻化, 因此並沒有將分群分類架構、既有語料庫及衍生語料庫的實驗設定加入。然而, 在解決現實生活中的辨識問題時, 我們會盡量利用手邊所有的資源, 並同時進行辭典、語料庫的精緻化, 以期能強化語言模型並改進辨識效能。在本小節中, 將會同時使用第三章的既有、衍生語料庫, 及分群分類架構, 並同時進行抽詞, 然後觀察辨識效能的提升。

在第三章的分群分類價購之實驗結果 (表 3.5) 中, 可發現語料庫精緻化帶來的辨識效能增加, 但若能將辭典精緻化的技術加入, 便可得到表 5.1 的實驗結果。表 5.1 的實驗是以分群分類架構, 將既有語料庫精緻化後, 進行抽詞, 取得精緻化的辭典, 再用以辨識廣播新聞語音語料。括號中的百分比是指和不使用分群分類架構 (即「1群」) 的比較的相對錯誤率降低比率。

從表 5.1 可發現, 未使用分群分類架構前, 混合基礎語言模型時並加上使用迭代式組合式抽詞 (設定如第四章實驗) 所得到的新辭典後, 字錯誤率可以從 18.05% 降到 15.35%。在分群分類架構下, 分成 100 個文件群的條件之下, 則可以從 16.75% 降至 14.07%。

到這裡可以發現, 字錯誤率從基礎辨識實驗的 23.17%, 已經可以進步至 14.07%, 相

語言模型		字錯誤率	次音節錯誤率
基礎語言模型		23.17	13.17
未抽詞	1 群	18.05	10.58
	100 群	16.75 (7.20%)	9.96 (5.86%)
有抽詞	1 群	15.35	9.16
	100 群	14.07 (8.34%)	8.19 (10.59%)

表 5.1: 字錯誤率與次音節錯誤率及相對降低比率。

對進步率是 39.27%。使用的技術是包含了選用了精緻化的語料做語言模型的訓練並抽取精緻化的辭典，並用新語言模型和基礎語言模型做線性的內插，得到品質相當穩定且符合 2002 年 9 月廣播新聞的主題的語言模型，因此才能夠得到 39.27% 這麼高的字錯誤率進步率。

5.2 訪談語料

5.2.1 訪談語料之特性

訪談語料的特性和廣播新聞截然不同，第一點不同是，廣播新聞是接近朗讀式語音，而訪談語料則算是介於自發式語音 (spontaneous speech) 及計畫式語音 (planned speech) 之間。第二點不同處則是廣播新聞的型態屬於獨白式 (monologue)，而訪談則屬於會話式 (dialogue)，嚴格來說，如果要將訪談語料做細緻的處理，這個特性應不能被忽略。不過在本論文中的實驗均採用一般性的語音辨識系統，並無區分語者、紀錄對話狀態等等功能，因此將著重在如何使用語料庫精緻化及辭典精緻化來強化語言模型，進而增進辨識效能。

5.2.2 本論文實驗使用之訪談語料介紹

本論文中採用的訪談語料集，是一個相對較小的語料集，其總長度為 34 分鐘，來源是中央社所錄製的訪談節目，節目中由一個主持人訪問一個來賓。此中央社訪談語料集，和

新聞語料無論在主題或文體上完全沒有關係，因此此目標語音語料和辨識用基礎辭典及基礎語言模型的不匹配性甚嚴重，所以可以預期採用基礎辭典及語言模型進行辨識實驗的效能將會非常差。

5.3 針對訪談語料之語言模型強化

由於本論文之前談的都是使用新聞語料訓練得到的基礎辭典及基礎語言模型，因此雖然目標語料是與其不匹配的訪談語料，在後續的第一個實驗中，我們仍將先採用本論文中新聞之辨識用基礎辭典及語言模型做辨識，並以此得到之基礎轉寫結果來取得衍生語料庫，在分群分類架構之下來進行辨識。這些從不匹配的基礎設定出發之實驗結果，將在5.3.1節中詳述。

5.3.1 採用新聞語料訓練的語言模型及辭典進行辨識

第一個實驗中，將中央社訪談語料用基礎辨識器（採用由新聞語料訓練得到的辨識用基礎辭典、基礎語言模型及聲學模型）進行辨識，得到的辨識效能列在表5.2中。本實驗係以相連式高信心量度查詢指令建構法收集衍生語料庫，並以分群分類架構辨識訪談語音語料。括號中的百分比是指和基礎轉寫結果比較的相對錯誤率降低比率。從訪談語料的基礎實驗結果，和廣播新聞語料的實驗（表3.5）比較之下，可以發現其辨識效能很明顯的差了很多，原因是之前提過的訓練語料與目標語音語料之嚴重不匹配性。

產生了訪談語料的基礎轉寫結果後，我們可以利用它來取得外部的衍生語料庫。根據第三章得到的結論，相連式高信心量度查詢指令建構法的效能比較好，因此在這裡使用此建構法來取得衍生語料庫。另外在套用分群分類架構，分別實驗文件群數等於1, 5, 10的三種情形。又由於目標語料和基礎語言模型的嚴重不匹配，我們嘗試著迭代執行此步驟兩次，觀察從中可獲得多少辨識效能的進步。

語料來源		字錯誤率		次音節錯誤率	
基礎語言模型		72.52		56.11	
第一次迭代	1 個文件群	70.44	(2.87%)	55.19	(1.64%)
	5 個文件群	69.66	(3.94%)	54.74	(2.44%)
	10 個文件群	69.56	(4.08%)	54.90	(2.16%)
第二次迭代	1 個文件群	69.95	(3.54%)	54.82	(2.30%)
	5 個文件群	68.77	(5.17%)	53.98	(3.80%)
	10 個文件群	69.31	(4.43%)	54.69	(2.53%)

表 5.2: 字錯誤率與次音節錯誤率及相對降低比率。

從表 5.2 可觀察到，在第一次迭代中最大的字錯誤率進步量是 4.08%，而在第二次迭代中，字錯誤率則更進一步有 5.17% 的改善。從這個觀察可以推論，當我們能辨識出更多正確的字時，能取得的衍生語料庫才會更為精準。換句話說，在本實驗中迭代的進步量非常小，可歸因於一開始訪談語料的基礎轉寫結果錯誤就太多，若能用一個較為匹配語言模型及辭典做初始實驗，辨識效能或可改善。

5.3.2 採用平衡語言模型進行辨識

在 5.3.1 節的實驗中，採用的語言模型和辭典，完全是以為了辨識廣播新聞為目的而訓練的。因此其嚴重的不匹配其實是在意料之中。在本節中，將採用一專門針對語言分析所設計的中研院漢語平衡語料庫，搭配中研院中文詞庫 (CED) 所訓練而成的語言模型，在本文中稱之為平衡語言模型。採用這個平衡語言模型(辭典使用中研院中文詞庫、語言模型訓練語料使用中研院平衡語料庫) 對訪談語料進行辨識。並使用最初的轉寫結果，以相連式高信心量度查詢指令建構法收集衍生語料庫，並以分群分類架構辨識訪談語音語料，得到的辨識效能如表 5.3 所示，括號中的百分比是指和僅使用平衡語言模型之辨識結果比較的相對錯誤率降低比率。從表 5.3 和表 5.2 的比較可以發現，使用平衡語言模型的字錯誤率是 69.73%，比起使用新聞為主的基礎語言模型的字錯誤率 72.52% 已經有相對 4.00% 的進步率。此外，進行衍生語料庫的收集實驗後，可以發現平衡語言模型的字

語料來源	字錯誤率	次音節錯誤率
平衡語言模型	69.73	54.85
第一次迭代 1 個文件群	67.28 (3.51%)	53.65 (2.19%)
5 個文件群	65.54 (6.01%)	53.28 (2.86%)

表 5.3: 採用平衡語料庫及中文詞庫進行辨識所得之字錯誤率與次音節錯誤率及相對降低比率。

錯誤率進步率最高有 6.01%，比起基礎語言模型實驗第一次迭代中最高的 4.08% 高出一點。這點可以證實：當我們手邊有的正確辨識結果越多，能取得的相關衍生語料庫也越多，因此對於辨識效能的增進會比較有幫助。

此外，從以上不甚理想的辨識效能，可以推論對於訪談語料這種介於自發性及計畫式語音的語料，光是使用統計式 N 連語言模型作為語音辨識系統中的語言模型，想要達到很高的辨識效能是很困難的，即使收集了衍生語料庫，能夠增加的正確辨識詞彙，也僅止於網路上容易取得的相關詞彙，對於一些語者說話的不流利及一些語音中的填充語 (filler words)，並沒有辦法從網際網路這個資源取得充分的語料，因此其辨識效能雖然有部分的提升，仍無法有顯著的改善。

第六章

結論與展望

6.1 總結與討論

本論文針對中文大字彙語音辨識中的語言模型問題，提出強化語言模型的一些方法，使語言模型能對目標語料的語言行為有更好的估測能力，藉此增進語音辨識系統的效能。要增強語言模型對目標語料的估測能力，在 N 連語言模型中最直接相關的部分，就是如何選取訓練或調適語言模型的文字語料庫。因此本論文中專門針對文字語料庫的部分，提出語料庫精緻化的方法，並且分析兩種不同來源的語料庫使用在特定語音辨識問題時的效能。在廣播新聞的實驗結果，發現如果能取得與欲辨識的目標語料同質的既有語料庫，將可以有效的強化語言模型。但除了既有語料庫以外，本論文中還提出兩種根據基礎轉寫結果，建構查詢指令，用以從網際網路取得相關的衍生語料庫之方法。當欲辨識的目標語料，並不容易取得大量既有語料庫時，衍生語料庫的幫助就顯得更為重要。

另外，在語料庫精緻化的部分，一般認為語料量多會比語料量少來的好，然而本論文中說明其實若目標語料和訓練/調適語料的匹配性不高，則太多語料反而是變成雜訊，會造成混淆。與其拿大量不同質的語料，還不如想辦法取得適量而同質的語料，因此，本文提出使用文件分群及文件分類的方法建構的分群分類架構，在這樣的架構之下，無論是針對衍生語料庫或是既有語料庫，都可以達到去蕪存菁、取得相關性較高的精緻語料庫的目的。

此外，由於中文的文字特性，詞與詞之間沒有明確的分隔，因此在處理語言模型相關議題時，辭典的選取問題和語言模型的品質是密不可分的。不但如此，由於對語音辨識系統最有幫助的詞彙，並不見得是傳統認定的「詞」，因此在本論文中特別針對統計式中文抽詞法進行討論，首先先從資訊檢索研究中的派樹抽詞法討論起，並稍做修改，例如放寬詞的顯著性規定、加入明顯詞邊界的條件等等，即得到本論文中使用的派樹抽詞法。然而，在這種統計式抽詞法中，最麻煩的事情其實在各個參數的調校，本文中的派樹抽詞法共有三個參數需要調整，因此在實驗的部分設計了相關的實驗以分析要如何調整此三個參數。

除了派樹抽詞法以外，本論文中亦討論另一個抽詞法——迭代式組合式抽詞法。這個統計式抽詞法，對詞的定義和派樹抽詞法對詞的定義是很接近的（例如：內聚力、邊界完整性），然而，迭代式組合式抽詞法特別的地方在於它是一個以詞為本（word-based），並且是可以根基於一個已經存在的辭典，來增加新詞。因此迭代式組合式抽詞法可以說是同時採用背景知識（現存辭典）及統計方法來抽詞的一個想法。在迭代式組合式抽詞法中的邊界完整性，採用的是左右文變異性統計，此一量度和派樹抽詞法所用的量度，最大的差別在於派樹抽詞法的邊界完整性量度使用了兩個門檻值參數，而左右文變異性統計中只考慮了其中一種門檻值參數。在本論文中討論參數設定的實驗中，也證實了在判定邊界完整值時，同時考慮此兩個參數和只考慮其中一個參數，其實在辨識的效能上沒有明顯的差異性，因此也證明了左右文變異性統計的合理性。

在深入各別分析了語料庫精緻化，與辭典精緻化對語音辨識效能的影響之後，為了能夠了解實際處理語音辨識問題時，可能遇到的困境，以及強化語言模型到底有多少的幫助，在本論文中分別針對廣播新聞語料，及中央社訪談語料兩種風格迥異的目標語音語料進行一連串的语言模型強化，這部分等於是整合了先前所提到的語料庫精緻化及辭典

精緻化的部分，並實際觀察思考針對不同特性的辨識問題，應如何強化語言模型以期得到更好的辨識效能。在各個實驗當中，我們發現對於廣播新聞語料來說，使用既有語料庫和衍生語料庫，對於辨識率都有幫助，但是由於廣播新聞對應的既有語料庫容易取得且品質良好，因此使用既有語料庫的辨識效能進步量還是比較好。對於訪談新聞語料來說，由於無法取得適當的既有語料庫，因此退而求其次觀察衍生語料庫的幫助，從實驗結果得到幾項結論，第一點是，當最初的轉寫結果正確率高時，能收集到的衍生語料庫幫助也比較大；再來是從訪談語料不甚理想的辨識效能，研究其原因，是因為從網際網路收集衍生語料庫，雖然對於辨識出重要的詞彙有幫助，但是對於語音中的不流利以及填充字，並不容易取得有用的訓練語料，因此雖然在訪談語料中使用衍生語料庫有幫助，然而仍無法達到和廣播新聞一樣好的辨識效能。

6.2 展望

在語言模型的相關議題中，其實有很多研究主題是息息相關的，例如中文抽、斷詞，語言模型訓練、語料取得等等。特別是在語音辨識系統的環境之下，其目標是希望達到辨識效能的增加，因此語言模型強化的問題一旦分析起來，比起傳統使用混淆度作為量度情況來的複雜許多。在進行本論文中許多實驗時，可以深切感受到各個問題其實是環環相扣的，然而限於計算資源的限制，要對語言模型問題進行全域性最佳化，其實是非常困難的。過去就曾有研究指出，應該將語言模型的問題視為一整體的程序，進行整體性的最佳化來改進語言模型，這個方向應該還是有思考的空間。

此外，在本論文中採用了機器學習的技巧來收集同質性高的語料庫。事實上，雖然語言模型是一個基礎的研究議題，但很有趣的是，語言模型仍可以被其他較高階的技術輔助。舉例來說，本論文中實驗的訪談語料，因其沒有容易取得的同質既有語料庫，因此其辨識效能一直非常差。假若能從其基礎轉寫結果中，擷取一些信心量度較高的關鍵字，

並將這些關鍵字藉由其他來源的知識 (例如網際網路、或是相關的電子資料庫), 將同質性高的語料庫補足, 則不但可以增加辨識率, 也同時可以替該訪談的語音文件找到許多相關的資訊。再者, 對於訪談語料, 如果能針對統計式 N 連語言模型在語言知識上的不足, 加入一些可以幫助辨識語音不流利及填充字的方法, 應該也會有相當的助益。

因此, 我認為未來語言模型的研究, 若不是走向更細緻的數學模型研究, 就是與其他主題結合, 嘗試強化語言模型的同時, 也可以針對其他研究主題中的問題加以解決, 這是未來有潛力的語言模型相關研究方向。

参考文献

- [1] R. Rosenfeld, “Two decades of statistical language modeling: Where do we go from here?,” in *Proceedings of the IEEE*, November 7 2000.
- [2] J. R. Bellegarda, “Statistical language model adaptation: review and perspectives,” *Speech Communication*, vol. 42, pp. 93–108, December 2 2004.
- [3] R. Lau, R. Rosenfeld, and S. Roukos, “Trigger-based language models: A maximum entropy approach,” in *ICASSP*, December 4 1993.
- [4] M. Federico, “Bayesian estimation methods for n-gram language model adaptation,” in *Proc. ICSLP*, (Philadelphia PA), pp. 240–243, 1996.
- [5] K. Sasaki, H. Jiang, and K. Hirose, “Rapid adaptation of n-gram language models using inter-word correlation for speech recognition,” in *Proc. ICSLP*, (Beijing), pp. 508–511, October 2000.
- [6] T. Moriya, K. Hirose, N. Minematsu, and H. Jiang, “Enhanced MAP adaptation of n-gram language models using indirect correlation of distant words,” in *Proc. ASRU*, (Italy), December 2001.
- [7] F. Jelinek, “Up from trigrams! The struggle for improved language models,” in *Proc. EUROSPEECH*, pp. 1037–1040, 1991.

- [8] R. Solsona, E. Fosler-Lussier, H. Kuo, A. Potamianos, and I. Zitouni, "Adaptive language models for spoken dialogue systems," in *Proc. ICASSP*, 2002.
- [9] M. Federico, "Efficient language model adaptation through mdi estimation," in *Eurospeech*, December 4 1999. Unigram constraint, MDI, Eurospeech 1999.
- [10] M. Federico and N. Bertoldi, "Broadcast news LM adaptation using contemporary texts," in *Proc. EUROSPEECH*, 2001.
- [11] K. Seymore and R. Rosenfeld, "Using story topics for language model adaptation," in *Eurospeech*, December 4 1997.
- [12] T. Yokoyama, T. Shinozaki, K. Iwano, and S. Furui, "Unsupervised class-based language model adaptation for spontaneous speech recognition," in *Proc. ICASSP*, 2003.
- [13] L. Chen, J.-L. Gauvain, L. Lamel, G. Adda, and M. Adda, "Using information retrieval methods for language model adaptation," in *Proc. EUROSPEECH*, 2001.
- [14] L. Chen, J.-L. Gauvain, L. Lamel, and G. Adda, "Unsupervised language model adaptation for broadcast news," in *Proc. ICASSP*, 2003.
- [15] K.-C. Yang, T.-H. Ho, L.-F. Chien, and L.-S. Lee, "Statistics-based seg-

- ment pattern lexicon — a new direction for Chinese language modeling,” in *Proc. ICASSP*, (Seattle, WA), pp. 169–172, May 1998.
- [16] E. P. Giachin, “Phrase bigrams for continuous speech recognition,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1995.
- [17] A. Berton, P. Fetter, and P. Regel-Brietzmann, “Compound words in large-vocabulary german speech recognition systems,” in *Proceedings of the International Conference on Spoken Language Processing*, 1996.
- [18] L.-F. Chien, “PAT-tree-based keyword extraction for Chinese information retrieval,” in *SIGIR '97*, pp. 50–58, ACM, 1997.
- [19] P. Fung, “Extracting key terms from Chinese and Japanese texts,” 1998.
- [20] J. Gao, J. Goodman, M. Li, and K.-F. Lee, “Toward a unified approach to statistical language modeling for Chinese,” in *ACM Transactions on Asian Language Information Processing*, vol. 1, no.1, pp. 3–33, 2002.
- [21] I. Good, “The population frequencies of species and the estimation of population parameters,” *Biometrika*, vol. 40, no.3/4, pp. 237–264, 1953.
- [22] S. M. Katz, “Estimation of probabilities from sparse data for the language model component of a speech recognizer,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-35, no.3, pp. 400–401, March 1987.

- [23] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *International Conference on Acoustic, Speech and Signal Processing*, vol. 1, (Detroit, MI), pp. 181–184, May 1995.
- [24] Y.-C. Pan, "One-pass and word-graph-based search algorithms for large vocabulary continuous mandarin speech recognition," Master's thesis, National Taiwan University, 2001.
- [25] "Central news agency CNA news." <http://www.cna.com.tw>.
- [26] "Yahoo! Kimo News portal." <http://tw.news.yahoo.com>.
- [27] "News 98 FM-98.1." <http://www.news98.com.tw>.
- [28] L. W. Cheng and R. Bissonnett, "Chinese electronic dictionary, CED," corpus, Chinese Knowledge Information Processing Group, Sinica. <http://rocling.iis.sinica.edu.tw/ROCLING/corpus98/decform.html>.
- [29] "Google." <http://www.google.com/>.
- [30] "Altavista." <http://www.altavista.com/>.
- [31] "Openfind." <http://www.openfind.com/>.
- [32] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 288–298, March 2001.

- [33] M. Steinbach, G. Karypis, and V. Kumar, “A comparison of document clustering techniques,” in *KDD Workshop on Text Mining*, 2000.
- [34] G. Karypis, “Cluto: A clustering toolkit,” Tech. Rep. #02-017, University of Minnesota, Department of Computer Science, August 2002.
- [35] P. Domingos and M. Pazzani, “On the optimality of the simple bayesian classifier under zero-one loss,” *Machine Learning*, vol. 29, no. 2-3, pp. 103–130, 1997.
- [36] A. K. McCallum, “Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering.” <http://www.cs.cmu.edu/mccallum/bow>, 1996.
- [37] Y. Liu, Q. Tan, and X. Shen, *Segmentation Standard for Modern Chinese Information Processing and Automatic Segmentation Methodology*. Beijing:Tsinghua Press, 1993.
- [38] CKIP (Chinese Knowledge Information processing Group), “A study of Chinese word boundaries and segmentation standard for information processing (in Chinese),” tech. rep., Taiwan, Taipei, Academia Sinica, 1998.
- [39] F. Xia, “The segmentation guidelines for the Penn Chinese Treebank (3.0),” 2000.
- [40] G. Saon and M. Padmanabhan, “Data-driven approach to designing com-

- pound words for continuous speech recognition,” *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, May 2001.
- [41] L.-F. Chien, “PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval,” in *Information Processing and Management*, vol. 35, no.4, pp. 501–521, 1999.
- [42] S.-P. Liao, “Enhanced language modeling for Chinese speech recognition,” Master’s thesis, National Taiwan University, 2003.
- [43] P.-C. Chang, S.-P. Liao, and L.-S. Lee, “Improved Chinese broadcast news transcription by language modeling with temporally consistent training corpora and iterative phrase extraction,” in *Proc. EUROSPEECH*, 2003.
- [44] K. W. Church and P. Hanks, “Word association norms, mutual information, and lexicography,” in *Proc. Computational Linguistics*, vol. 16, no.1, pp. 22–29, 1990.
- [45] C. Beaujard and M. Jardino, “Language modeling based on automatic word concatenations,” in *Proc. EUROSPEECH*, 1999.
- [46] J. Zhang, J. Gao, and M. Zhou, “Extraction of Chinese compound words - an experimental study on a very large corpus,” in *The Second Chinese Language Processing Workshop attached to ACL2000*, 2000.
- [47] P.-K. Wong and C. Chan, “Chinese word segmentation based on maxi-

mum matching and word binding force,” in *Proc. of Computational Linguistics*, pp. 200–203, 1996.

- [48] P.-C. Chang and L.-S. Lee, “Improved language model adaptation using existing and derived external resources,” in *Proc. ASRU*, 2003.