國 立 交 通 大 學

電信工程學系碩士論文

中文大詞彙語音辨認之

語言模型改進

Improvement on Language Modeling for

Large-vocabulary Mandarin Speech Recognition

研 究 生：周建邦

指導教授：陳信宏　博士

中華民國九十八年十二月

中文大詞彙語音辨認之語言模型改進

Improvement on Language Modeling for Large-vocabulary Mandarin Speech Recognition

研 究 生：周建邦　　Student： Chien-Pang Chou

指導教授：陳信宏 博士　　Advisor： Dr. Sin-Horng Chen

A Thesis
Department of Communication Engineering
College of Electrical Engineering and Computer Science
National Chiao Tung University
In Partial Fulfillment of Requirements
For the Degree of
Master of Science
In Electrical Engineering

December, 2009
Hsinchu, Taiwan, Republic of China

中華民國九十八年十二月

中文摘要

本研究之目的為探討中文大詞彙語音辨認之語言模型改進。傳統大詞彙語音辨認大多使用統計式語言模型，藉此計算數萬詞條之雙連文或三連文機率模型，然而，此方法仍有其缺失，因其無法對於不包含在辭典中之詞彙進行辨識，其中包含數量複合詞、專有名詞、不常出現之詞綴構詞等等，基此，本研究針對混合詞及半詞(subword)之統計式語言模型進行探討，期望藉此增進辭典之涵蓋率，降低無法進行辨識之詞條數目。

本研究分為三大主軸，首先，對於文字資料庫進行前處理，針對不適當內容(英文、文章標題等)進行刪減、對於錯誤文字予以更正、斷詞、文字正規化等；其次，建構混合詞及半詞統計式語言模型，探討字典收錄詞條之策略、將辭典未收納之詞彙拆解為半詞之方法、以及混合模型之建立，最後，採用兩階段(two-stage)辨認架構，針對辨認方法及實驗結果進行說明，並進一步分析與比較架構式模型和傳統方法模型之語音辨認結果之優劣，針對本研究考量之三種構詞(人名、詞綴及數量複合詞)的辨識效益進行深入分析。

為了驗證提出方法之效能，本研究採用 TCC300 麥克風語料為語音實驗語料，語言模型則由台灣光華雜誌(Taiwan Paramora)及中文檢索標竿(NTCIR3.0)文字語料庫求得，實驗結果顯示，相較於傳統採用之統計式語言模型，本研究所提出的混合模型對於大詞彙語音辨認系統效能有所改善，整體詞辨認率(word accuracy)由 **60.86%**提升至 **62.85%**，經過深入分析發現，使用所提出之兩階段辨認方法對於人名、詞綴及數量複合詞確實有所幫助，此三類辨認正確之數量增加驗證了提出方法的有效性。

Improvement on Language Modeling for Large-vocabulary Mandarin Speech Recognition

Student: Chien-Pang Chou Advisor：Dr. Xin-Hong Chen

Department of Communication Engineering

National Chiao Tung University

Abstract

The purpose of this research is the improvement of language modeling for large-vocabulary Mandarin speech recongnition. Traditionally, large-vocabulary speech recognition is almost to employ statistical language model. By calculating million of bigram(or trigram) probability model is also having the drawback. Because we can not recognition the OOV(out-of-vocabulary) words(including determiner-measure compoundi word, name entity, and affix word). Because these reasons, we probe into the statistic language model which mixs word and subword. By this way, we not only hope that increasing the coverage of lexicon, but also decreaing the number of words which we can't recognition correctly.

This thesis divides three parts. First, we explore the applicability of the corpus to be used to build the language model, and to observe the contents of corpus whether fit to build the language model or not. We delete the misfit contents and correct the wrong words. We hope to promote the whole recognition rate. Second, we want to train the statistic language model which mixs word and subword, and probe into the tactics that collect the entiry of recognition lexicon to building language model which have the word and the subword. Finally, we use two-stage framework to recognition, and further analyize the result of two-stage experiment.

In order to prove the efficiency of the method, we observed the numer of the class recognized correctly is obviously increasing, and recognition rate is 60.86% upto 62.85%. The phenomenon have identified that this framework is efficient.

誌謝辭

首先誠摯的感謝指導教授陳信宏教授、王逸如教授，兩位老師悉心的教導使我體會到語音領域的深奧，兩位老師教導的方式不同，一個對於研究的態度是大處著眼，另一個的態度則是小處著手，讓我學習到研究的重心在於理論的實現，在研究的過程中，好好的享受過程、了解探討過程、分析結果，觀看結果是否合理，並且驗證想法。老師對做學問的嚴謹更是我學習的典範。

這兩年多的研究所生活，幸好有大家的陪伴才得以順利度過，平常研究累了就會聽到大家言不及義的打屁交談，加入戰局更得精神百倍，不會忘記大家的加油聲，一起在實驗室相處的時光，因為有大家的加入，使得我在交大生活可以多采多姿。

感謝希群學長對我的研究給予建言及想法，當我在進退兩難時，給予協助；感謝阿德學長在我心情低落時，給予幾句關心及開導；感謝智合學長總是主動前來詢問我的進度狀況，以給予幫助；感謝 barking 學長讓我見識何謂程式上的強者；感謝 QQ、小宋、普烏同學，有你們的陪伴，我才能順利的度過研究所生活。

另外要特別感謝陪伴我六年多的黃小謬，也是我最、最..最可愛的女朋友，當我因研究受挫折時都一直給我安慰、鼓勵，要我不要氣餒，使我能再接再厲而不被挫折打敗，謝謝您一直陪伴著我，不管多遠都會到新竹看我，真的很感動，幸好生命終有你的陪伴，可以陪著我度過研究所時期，以後我們還要繼續加油，再一起手牽手走下去。

最後感謝我的雙親，給予學業上的信任與自由，以及支持我繼續努力，而得到這多彩多姿的交大研究生活，有你們的支持，使我能不顧一切的往前衝，我也才能努力到現在，謝謝您們！

目錄

圖目錄

表目錄

第一章 緒論

1.1 研究動機

近年來，語音辨識之相關研究大多自語音訊號層面著手，透過研究聲音之特性，藉此提昇辨識率。然而，語音和語言密不可分，若語音辨識能回歸於基本之語言，針對語言進行研究，對於辨識效果之提升將會有所助益。

在傳統大詞彙語音辨認中，語言模型所用之辨認辭典，大多是將語料中的詞依據詞頻排序，進而取其排序前六萬者納入辭典，然而，語料中並非每個詞均會被辨認辭典所收錄，進行語音辨認時，倘若出現辨認辭典未收錄之詞時，該詞將無法被辨認，造成辨認上的錯誤，而此些不在辨認辭典中的詞稱為「Out-of-Vocabulary」，簡稱 OOV words，所佔比例稱「OOV rate」。

OOV rate 大小將會影響辨識效能，數值越大代表著越多詞將無法被辨識，而此現象大多出現於拼音複雜或詞變化多的語言，如：德語、土耳其語、阿拉伯語、芬蘭語…等皆有此現象，其中，阿拉伯語之語音辨識研究為突破 ASR 系統無法完全收錄詞彙對於辨識效能之限制，運用 morphological analysis 建構語音辨認系統(R. Sarikaya et al, 2007)，試圖藉此解決拼音複雜的阿拉伯語辨認辭典詞彙收錄不足的問題。

自此反觀中文語音，中文語詞變化多元，特別是數量複合詞、專有名詞(本研究針對人名)、不常出現之詞綴構詞…等，此些類別之詞可任意組合且變化無限，故中文語音辨識亦存在著辨認辭典無法完整收錄詞彙而阻滯辨識效能成長的困境。基此，為突破辭典詞彙收錄有限之限制，本研究試圖提出一個階層式 word-subword based ASR 系統，針對混合詞及半詞(subword)之統計式語言模型進行探討，整體系統架構包含三個模組，包括：第一級 word/subword basd 辨認模組、第二級 lattice extension 構詞模組，以及第三級 lattice rescoring 模組，期望藉此增進辭典之涵蓋率，降低無法進行辨識之詞條數目，進而提升辨識成效。

1.2 研究方向

傳統基本語音辨識系統主要包含五大層面，包括：語音特徵參數的求取、聲學模型(Acoustic Model, AM)的訓練、語言模型(Language Model, LM) 的訓練、辭典選取以及辨識比對，如下圖 1-1 所示：

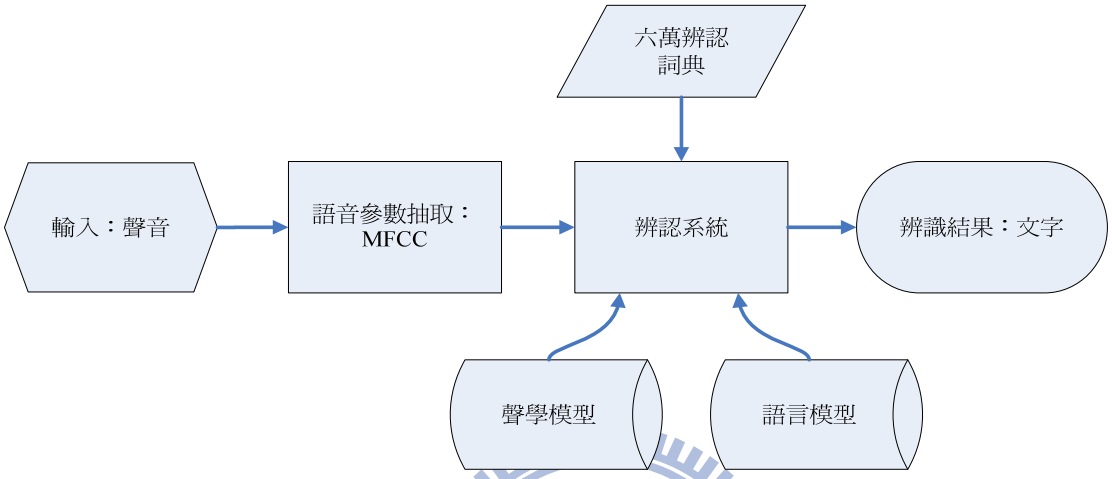

圖 1 -1：傳統式辨識流程

如先前所述，為改善傳統辨認辭典無法完整收錄詞彙的問題，本研究提出階層式 word-subword[1] based ASR 系統架構，以此作為研究之思考脈絡與架構，如下圖 1-2 所示：

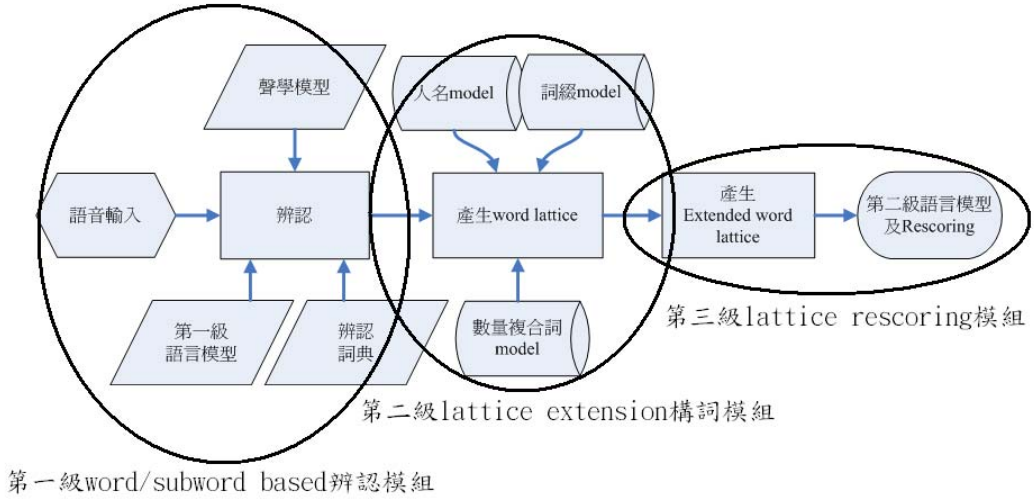

圖 1-2：階層式 word-subword based ASR 系統架構

[1] subword(半詞)：意指長詞拆解成數個半詞單位。

由上圖可知，階層式 ASR 架構由左至右可分為三大模組。首先，第一級模組為 word/subword based 辨認模組，本研究將針對如何建立 word/subword 混合式語言模型(第一級語言模型)進行探討，而如何於此模組中選取所需的 word 和 subword 作為辨認辭典，乃為本研究重要課題之一。

其次，第二級模組為 lattice extension 構詞模組，研究者將第一級模組之辨認辭典未收納的詞彙進一步拆解為數個 subword 半詞串，而第二級模組於 word lattice 上將運用此些 subword 短詞串進行構詞。然而，本研究所立基之實驗室目前除數量複合詞外，其餘尚未建立構詞模型，無法立即精確地偵測出需要構詞的 subword 半詞串，基於此限制，本研究運用查表法進行構詞，對於如何建立完整構詞模型，本研究尚無深入探討。

最後，第三級模組為 lattice rescoring 模組，該模組針對第二級模組構詞產生之 extended word lattice 的路徑重新給予語言模型分數，藉此產生第二級語言模型，基此，本研究將探討分數分配之相關議題，針對辨認方法及實驗結果進行說明，並進一步分析與比較階層式和傳統式語音辨認結果之優劣。

1.3 章節概要

基於本研究所欲探討之相關議題，本文之章節架構與內容於下概述：

第一章 **緒論**：介紹研究動機、研究方向及章節概要。

第二章 **語音辨認基本系統的建立**：回顧傳統聲學模型和語言模型的建立方式，包括：文字前處理、辨認辭典的建立、OOV 處理、smoothing 方法及 perplexity 計算，並運用 HTK toolkit 以訓練傳統語言模型，以此作為和階層式語言模型辨識效能之比較基礎。

第三章 **混合式語言模型**：將語料之人名、詞綴和數量複合詞(DM)以 subword 短詞方式進行拆解，並探討辭典收錄之範圍，並運用 HTK 工具以訓練 word/subword 混合式語言模型。

第四章 **第二級語言模型**：將第一級辨認產生的 word lattice 透過構詞形成第二級 extended word lattice，再針對其路徑給予分數配置，進而產生第二級語言模型。

第五章 **實驗解果與分析**：檢視有意義長詞之辨識效能、實驗相關分析，並與傳統語音辨認效能進行比較。

第六章 **結論與未來展望**：立基於本研究結果，對於未來研究提出建議與展望。

第二章 語音辨認基本系統的建立

近年來語音辨認之相關研究最常採用的聲學模型為隱藏式馬可夫模型(Hidden Markov Model, HMM)，透過該機率模型，描述發音過程之狀態(State)轉移現象與輸出結果，該方法之辨識效能佳，故本系統亦採用此模型，並加入語言模型，期望藉由語言模型之改進來幫助提升語音辨識率。

2.1 Acoustic model 的建立

對於聲學模型（Acoustic Model, AM）之建立透過以下兩小節來介紹，如下：

2.1.1 語音資料庫

目前語料採用的是PTSND第一年40小時。我們將所有可用語料的十分之九歸於訓練語料，十分之一歸於測試語料。其中訓練語料的時間大約有8.5個小時。

2.2.2 參數設定

在進行訓練或辨識之前，我們均會先將輸入的語音均進行前處理，即求取其語音參數。我們求取的語音參數是梅爾倒頻譜參數（Mel-Frequency Cepstral Coefficients；MFCC），利用語音在頻譜上具有短時間穩定的特性，並且MFCC有考慮到補償人耳的聽覺效應。

而語音參數求取時所使用之系統參數如下所示：

表 2-1：參數抽取設定檔

取樣頻率	16 kHz
音框長度	30ms
音框平移	10ms
Filter bank 個數	24 個梅爾刻度三角濾波器

我們求取的MFCC參數為13維度，再加上1維與2維的變化量。但是因為第0階的能量並不重要，因此第0階的能量會被省略，而第1、2階代表的是能量變化，因此會保留下來。所以最後求出的是一個38維度的語音參數向量。

為了能求得切割位置，我們先以Read speech（TCC300）所訓練之HMM模型來切割我們欲建立初始模型的訓練語料，初始訓練語料目前只選取單純的資料使用，也就是語料內容裡沒有背景聲（有背景聲下的語音辨識目前不是我們研究的方向），同時只能有國語411音節（Syllable）與Particle這兩種資料，資料的數量為665個sub-turn、字數35,388字，時間約2.05個小時。

因為Read speech不會去訓練Particle這類在Spontaneous speech常出現的聲音模型，因此在進行切割（Forced Alignment），Particle則使用相近411音節替代。

在求得411與Particle的切割資訊（Boundary Information）之後，我們會進行已知位置的初始模型訓練。訓練出的聲母和Particle採用3個狀態，韻母用5個狀態之HMM模式，Mixture個數均為16。

本研究採用 left-to-right HMM，儘管口腔聲道會隨時間而變，但語音訊號具備短時間的穩定特性，故假設同一音框（Frame）之口腔狀態是相同的。另外，本研究採用混和高斯模型（Gaussian Mixture Model），代表音框語音參數與各狀態相似程度之狀態觀測機率（State Observation Probability）。

本研究之訓練模型、估計參數時採用的方法則利用Baum-Welch參數估計法，從已知狀態序列，根據轉移規則，推測出每個音框所屬的最佳口腔狀態，並重複估測聲學模型至穩定為止；至於辨識工作的進行則是使用Viterbi search，讓每個音框均對所有模型進行估計，並找出最佳結果。另外，採用的訓練軟體為英國劍橋大學開發的HMM Tool Kit (HTK)【3】。

2.2 Language model 的建立

對於語言模型（Language Model, LM）之建立，傳統方法上會在決定辨認辭典前，將文字資料庫經過文字前處理之流程，之後可由統計方式得到辨認辭典及不收錄在辭典內的 OOVwords，在將此些 OOV words 經過拆解來解決 OOV 問題，經過這些步驟後之語料會對語言模型之效能有所助益，最後，利用處理過後的語料進行語言模型的訓練，訓練流程如下圖(圖 2-1)所示並且用以下數小節分開敘述之：

圖 2-1 文字語料庫處理流程

2.2.1 文字資料庫

辨識系統之語言模型，通常需要透過大量的文字資料來進行訓練，利用大量的文字資料訓練出一個涵蓋範圍廣泛、適用於各個領域的語言模型，基於此種模型的普遍性，稱為「General LM」。因此，一個好的語言模型，其所需要的條件，必須擁有大量的文字資料庫，而本研究使用下述兩個文字資料庫來建立語言模型：

(1)光華雜誌：內容為光華雜誌的文章，蒐集範圍為 1976 年到 2000 年之間。

(2) NTCIR：內容由各個不同學科領域之文章所構成，為建立資訊檢索系統的標竿測試集。

研究者針對訓練語言模型所使用之光華雜誌與 NTCIR 兩語料庫進行詞、字元數量統計，結果於表 2-2 中呈現，如下：

表 2-2：文字資料庫 word 及 character 統計表

語料庫	詞數（Word）	字數(Character)	詞條數
光華雜誌、NTCIR	112,966,482	210,480,091	518,539

文章的平均詞長= **1.8632**

由下圖（圖 2-2）可知，收納於語料庫的詞依據詞頻排序後，累加至次序到達辨認辭典限制值(六萬)時，語料庫詞條數涵蓋率到達 96.67%，其 OOV rate 約為 3.33%，換言之，將有 3.33%的詞數無法收錄於辭典中，而本研究將針對無法收錄者進行拆詞，以 subword 短詞形式呈現之，以解決 OOV 對於語音辨識造成錯誤結果之問題。

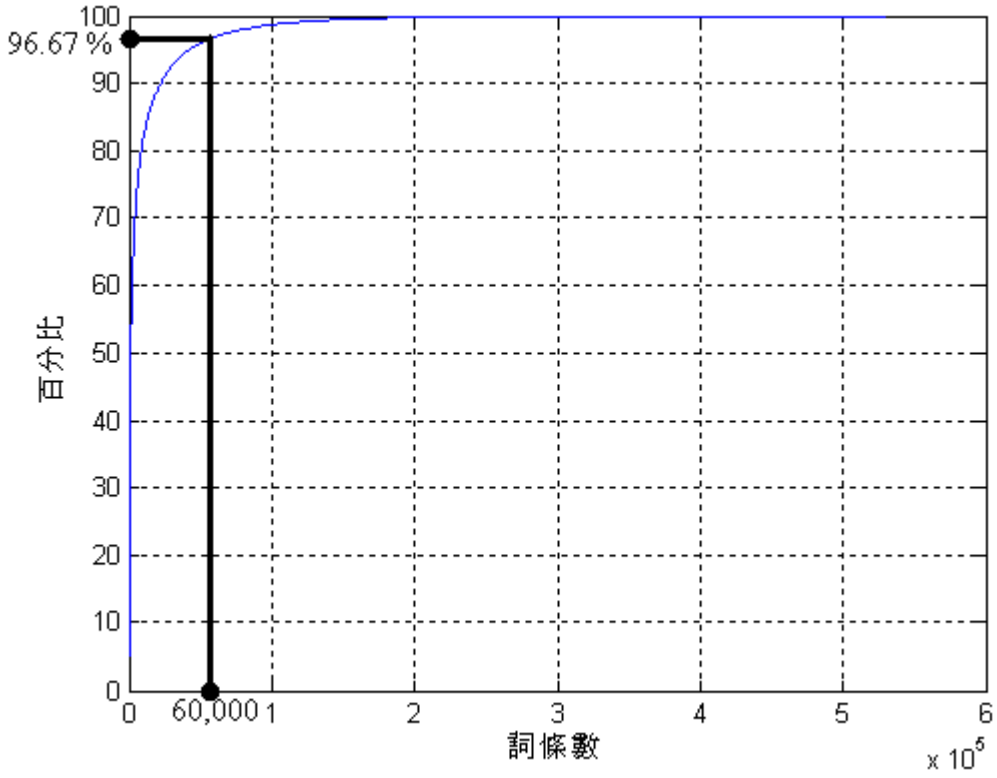

圖 2-2 文字資料庫 coverage 曲線圖

研究者針對光華雜誌與 NTCIR 兩大資料庫之可拆解為 subword 短詞者進行統計，統計結果如下(表 2-3)：

表 2-3 文字資料庫分析

文章內容分析如下				
	總詞條數	各類所佔比例(%)	總詞數	各類所佔比例(%)
NTCIR、光華雜誌	518,539		112,966,482	
文章細分如下				
人名	102,408	19.75%	1,588,822	1.40%
詞綴	18,364	3.54%	3,500,370	3.10%
數量複合詞	265,869	51.27%	4,894,156	4.33%
一般詞[2]	131,898	25.44%	102,983,134	91.17%

[2] 一般詞：意指不需進行 subword 短詞拆解者

2.2.2 文字前處理

在進行訓練語言模型之前，須先將語料庫的文章進行前處理，將文章中會影響辨認效能的內容移除或修改，再經由斷詞、正規化和消除 OOV 後，將此些經過處理後的文章，用以訓練 word-based 之語言模型。文字前處理流程如圖 3-1 所示，以下方塊內容將在 **2.2.2.1** 至 **2.2.2.4** 各小節分開敘述之。

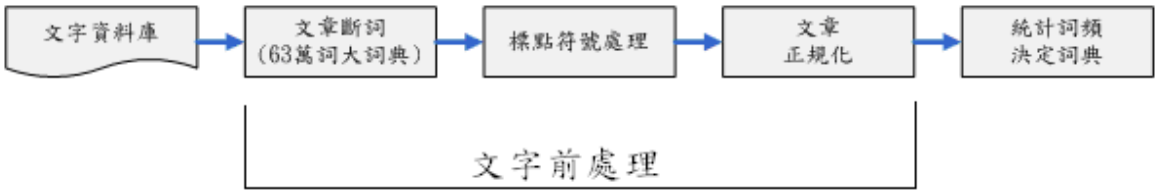

圖 2-3 文字前處理流程圖

2.2.2.1 文章斷詞

在此建立語言模型是經由統計的方式建立，語言模型是在統計詞和詞之間的連接機率關係，所以得把語料庫的文章斷詞成為以詞為單位的資料，來統計詞和詞之間的機率，而語言模型的好壞也會和斷詞時所決定詞的邊界有關，這個部分可參考相關論文【5】。斷詞之斷詞辭典所收錄的詞愈多，文章斷詞之後的結果，較不會有搶詞的現象出現，也就是斷詞辭典的大小，會影響到斷詞結果的正確率。斷詞辭典無法收錄所有的詞，這些未收錄的詞中，有些詞是有規則的，可以利用構詞規則產生，把輸入文句中符合構詞規則的詞成為斷詞的候選詞，再經由斷詞規則進行斷詞，其中這些有規則詞且和數量有關為「定量複合詞」、「數詞定詞」、「數量定詞」、「時間詞」、「地方詞」、「位置詞」等，這些我們稱為「數量複合詞」。在進行斷詞時，再加上構詞單元，能讓斷詞的結果更加完善【5】。我們使用的斷詞辭典為交大語音實驗室大辭典，收錄詞條數約為六十三萬條，其詞條分類如下：

(1)人名：主要來源有

(1.1)聯考榜單。

(1.2)文字資料庫(光華雜誌、NTCIR)經由現有的六萬詞辭典斷詞後，把連續的一字詞串，用半自動的方式，以姓氏為首的挑出連續兩個一字詞及連續三個一字詞。再以人工判別是否為人名。

(2)詞綴：分成前詞綴、後詞綴，主要來源包括

(2.1)中研院所提供的詞綴範例。

(2.2)文字資料庫(光華雜誌、NTCIR)經由現有的六萬詞辭典斷詞後，以半自動的方式取得，方法如下。

(a)以詞綴為首，分別挑出包含詞綴且詞頻高的二字詞及三字詞，以人工的方式去辨別是否為前(後)詞綴。

(b)詞頻少的二字詞及三字詞，不代表大部分都是前(後)詞綴，前(後)詞綴是三字詞可能性比二字詞高，所以現階段只針對三字詞處理。為了再縮小範圍，把挑出來的三字詞再給予詞類(POS)限制，由於詞綴的詞類大部分是名詞，所以挑出詞類(P 為名詞(Na、Nb、Nc、Nd)的三字詞，在以人工的方式辨別三字詞是否為前(後)詞綴。

(3)股票名：經由股票網站收集股票公司名。

(4)定量複合詞(DM)：文字資料庫(光華雜誌、NTCIR)經由現有的六萬詞辭典斷詞後，挑出詞類(POS)為 DM 的詞，由於構詞規則不夠嚴緊，定量複合詞(DM)有少數的錯誤，但大部分的定量複合詞(DM)都是正確的，所以這部分錯誤暫不討論。

(5)縣市鄉鎮：經由地圖資訊取得。

(6)學校及科系名：經由聯考榜單取得。

(7)核心詞：核心詞是由中研院八萬八千詞去除其他種類的詞後所剩的詞。

2.2.2.2 標點符號處理

中文所使用的標點符號(PM)共有十六種，可區分為標號與點號兩大類，其中標號

常用的有書名號、破折號、省略號、括號、引號等九種，而點號則有逗號、頓號、句號、冒號、分號、問號、驚嘆號共七種，這兩大類中又以點號跟說話時的停頓有較大的關聯性【6】，所以在文章中標點符號上的處理，利用點號中的四種點號(句號、分號、驚嘆號、問號)把文章分段，由於在聲學模型中並未有考慮到標點符號的模型，所以把文章中所有的標點符號先予以移除。

2.2.2.3 斷詞後的資料庫—英文串取代

由於我們的辨認目標為中文詞，聲學模型中並沒有去訓練這類英文詞的聲音模型，所以文章中的英文詞我們以一個類別看待它，把它歸類成「LONGFW」這個類別，在進行辨認的過程中，並不把這個類別收錄至辭典內，而將這個類別視為 unknown words（即 OOV）。

2.2.2.4 文字正規化

文字正規化可分為兩大部分，首先，由於文章的內容，有些是阿拉伯數字、詞和符號都必須由寫法轉為語音讀法；另一方面，文章內有些詞只是寫法不同，但在讀音上及語義上是相同的，需把這類的詞合併成同一個詞。這些步驟稱為文字正規化。

第一部份：寫法轉讀法

在文章之中，有些阿拉伯數字、詞或符號必須由寫法轉為語音讀法，這個過程稱為文字正規化，舉例來說「９０％」應該讀為「百分之九十」。經由蒐集整理的結果，我們發現到大部分由構詞規則構出的詞，也就是定量複合詞（DM）、數量定詞(Neqa)、數詞定詞(Neu)、時間詞(Nd)、地方詞(Nc)、位置詞(Ncd)等六類的數量複合詞，如果含有阿拉數字及特殊符號，都需要被正規化為讀法，文字正規化轉換範例如下表(表 2-4)：

表 2-4 文字正規化範例

未正規化之詞	已正規化之詞
90.2	九十點二
90%	百分之九十
90.20%	百分之九十點二零
一零零號	一百號
ＴＥＬ	電話號碼

第二部份：合併同音義異詞

某些詞和詞之間在發音上和語義上是相同，只是寫法上有所不同，而這類的詞會在辨認上造成混淆，所以把這類的詞統一合併成一個詞(如表 2-5)。經過這個步驟可將文章詞更集中，促使 OOV 量減少。

表 2-5 同音義異詞範例

同音義異詞	
佰、仟	百、千
部份	部分
佈告欄	布告欄
憤憤不平	忿忿不平
洩露國家機密	洩漏國家機密

2.2.3 辨認辭典建立

本研究所指的傳統語言模型辭典，其收錄方式為將語料庫中的詞統計按詞頻排序，進而加上所有中文的一字詞，以此累積至六萬詞，分佈如表 2-6：

表 2-6 傳統方式辭典統計表

Language Model Lexicon		
總詞數	**60,000**	
詞長	**詞數**	**百分比(%)**
1	13060	21.83
2	30125	50.14
3	11513	19.19
4	4208	7.01
5	700	1.17
6	344	0.57
7	45	0.075
8	5	0.009

LM 辭典的平均詞長 = 2.17

2.2.4 OOV 處理

針對未收錄於六萬詞辭典之 OOV word【7】進行拆解，將辭典未收錄之長詞，轉變成辭典可辨識之短詞，換言之，將 OOV word 拆解為可由辭典內的詞所構成。

2.2.5 N-gram training

由下圖(圖 2-4)可知，經過先前所述之文字前處理、OOV 處理動作後，語料庫內的文章將可利用 HTK tools 進行 bi-gram 語言模型的訓練、辨認，過程中將設定 cut-off 值，因而產生大量 bi-gram，進而轉成 word network 形式，藉由 bi-gram 將詞與詞相互串聯，形成 word network。

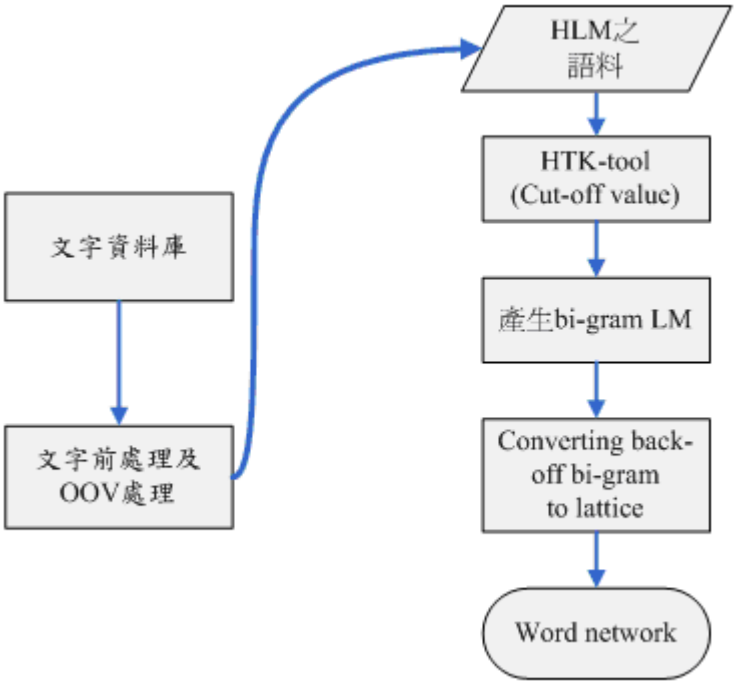

圖 2-4：傳統語言模型建立流程圖

2.2.5.1 機率的 Smoothing

在訓練 bi-gram 機率時，倘若分子的 *Count*(・)值為 0 時，即 bi-gram 機率等於零，因於 training data 中未出現者，並非代表 testing data 不會出現，故於此情況下，機率的給定是不合理的，而當詞接詞之 count 值很小時，所計算出的 n-gram 機率也是不準確。因此，須對計算出的機率予以 smoothing 動作【4】，使所有的 n-gram 機率均能被良好的估計。

Good-Turing discounting for n-gram 是常見的 smoothing 方法，它可表示如下：

$$P(w_i \mid w_{i-n+1},\mathrm{K}\ ,w_{i-1})$$

$$=\begin{cases} a(w_{i-n+1},\mathrm{K}\ ,w_{i-1})P(w_i \mid w_{i-n+2},\mathrm{K}\ ,w_{i-1}) & Count(w_{i-n+1},\mathrm{K}\ ,w_i)=0 \\ d_a \dfrac{Count(w_{i-n+1},\mathrm{K}\ ,w_i)}{Count(w_{i-n+1},\mathrm{K}\ ,w_{i-1})} & \min \le Count(w_{i-n+1},\mathrm{K}\ ,w_i) \le \max \\ \dfrac{Count(w_{i-n+1},\mathrm{K}\ ,w_i)}{Count(w_{i-n+1},\mathrm{K}\ ,w_{i-1})} & Count(w_{i-n+1},\mathrm{K}\ ,w_i) > \max \end{cases} \quad (2.2)$$

其中，$a(w_{i-n+1},\ldots,w_{i-1})$ 為 back-off 係數，當計算出次數為 0 時，則利用(n-1)-gram，乘上 back-off 係數，來表示出現次數為 0 的機率，並分配給它一個適當的機率值。$a(w_{i-n+1},\ldots,w_{i-1})$ 的選定，還會經過 normalization，令其滿足

$$\sum_{w\in V} P(w_i = w \mid w_{i-n+1},\ldots,w_{i-1}) = 1 \tag{2.3}$$

若是 Count 的值很小，造成機率預測不準確時，解決方式則為當詞串次數小於 max 次時，會乘上一個根據 Good-Turning discounting 所計算出來的值 d_a (Discount Coefficient Factor)，減低其機率，並將扣除的機率分給未出現的 n-gram 機率使用。

2.2.5.2 Perplexity 計算

利用建立好的 LM，將可以算出語言模型的 Perplexity (PPL)，而語言模型的好壞，可透過 perplexity 進行測量，perplexity 定義如下:

$$\begin{aligned} PP &= 2^{\hat{H}} \\ \hat{H} &= -\frac{1}{m}\log_2 P(w_1, w_2 \ldots, w_m) \end{aligned} \tag{2.4}$$

上式意指一個句子之內容乃由 m 個詞所組成，對於每個新詞提供的平均資訊量，entropy(熵值,以 H 表示)，經過 ergodic 的假設與適當地化簡，最後以上式來對 H 做近似。然而，計算 log probability 是以 10 為基底，因此數學式修改如下:

$$\begin{aligned} PP &= 10^{\hat{H}} \\ \hat{H} &= -\frac{1}{m}\log_{10} P(w_1, w_2 \ldots, w_m) \end{aligned} \tag{2.5}$$

其中

$$P(w_1, w_2, \ldots, w_m);\ \prod_{i=1}^{m} P(w_i \mid w_{i-1})$$

$$\begin{aligned} \log_{10} P(w_1, w_2, \ldots, w_m) &= \log_{10}\left(\prod_{i=1}^{m} P(w_i \mid w_{i-1})\right) \\ &= \log_{10} P(w_1) + \log_{10} P(w_2 \mid w_1) + \ldots + \log_{10} P(w_m \mid w_{m-1}) \\ &= \sum_{i=1}^{m} \log_{10} P(w_i \mid w_{i-1}) \end{aligned}$$

2.2.6 傳統 language model 之探討

傳統式語言模型通常會受限於辨認辭典大小之限制，對於沒有收錄之六萬辨認辭典的詞將會被拆解成辭典內的短詞，而其拆解方式並無一定地規則，故此，拆解後的詞在辨認結果中無法以有意義之長詞呈現，進而影響整體詞的辨識效能，因此傳統式辭典對於語料庫之涵蓋率（cover rate）高低會影響到語言模型之辨識效能，尤其是中文詞種變化繁多的語言更為明顯。

第三章 混合式語言模型

如先前第二章所述，中文語詞變化複雜，辭典無法完全收錄之，因而使得語音辨識效能成長有限，為突破此困境，本研究採取 word（詞）和 subword（半詞）並存之混合式語言模型，詞的部份是針對高詞頻的詞為保留長詞的形式針，而半詞的部份對人名、詞綴和數量複合詞三類進行規則性轉化，將之為拆解為 subword 半詞串，並在混合式辭典收錄高詞頻的詞及半詞，進而訓練混合式語言模型，再由辨認結果得到之 subword 半詞透過第二級構詞模組，將第一級拆解之半詞組合回原先有意義之長詞，讓詞辨識效能較傳統方式有所提升。

在進行混合式語言模型訓練之前，文字資料庫的處理過程如圖 3-1 所示，其中，文字前處理及 OOV 處理與訓練傳統式語言模型相同，而相異之處在於文字前處理之後增加了「三類詞拆解成半詞」及辨認辭典為「混合式」辨認辭典，以此建立第一級之混合型語言模型。

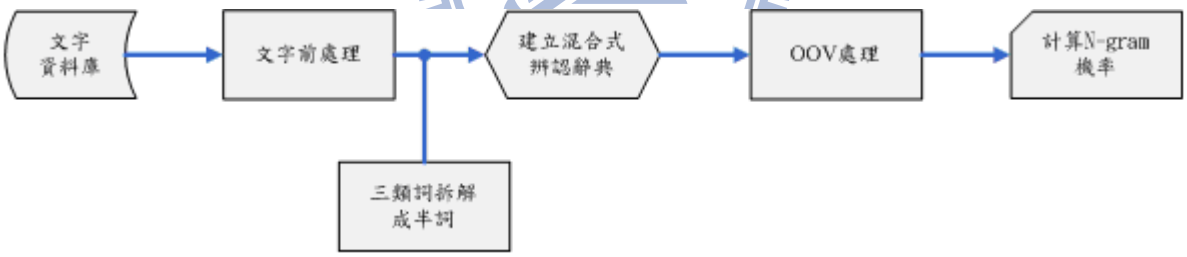

圖 3-1 文字資料庫處理流程

進行語音辨認時，一個語言模型需相對搭配一個適當的辭典。如下圖(圖 3-2)所示，本研究混合式辭典有別於傳統式辭典，除一般常見詞外，亦於辨認辭典中加入 subword 半詞，故 subword 集合的大小將相對地影響著辨認辭典內收錄一般常見詞之多寡，進而影響第二級構詞的效果。因此，如何於訓練第一級混合式語言模型之前，透過策略性方法來選取適當的 subword 數量，將是重要議題。本研究以 coverage 高低來決定 subword 詞條收錄範圍，反覆在 subword 集合中收集不同數量人名、詞綴和數量複合詞之 subword 半詞，直至 subword 集合使得辭典內的詞在文章中所佔 coverage

可高至某一程度時，即決定第一級辨認所需之辨認辭典。

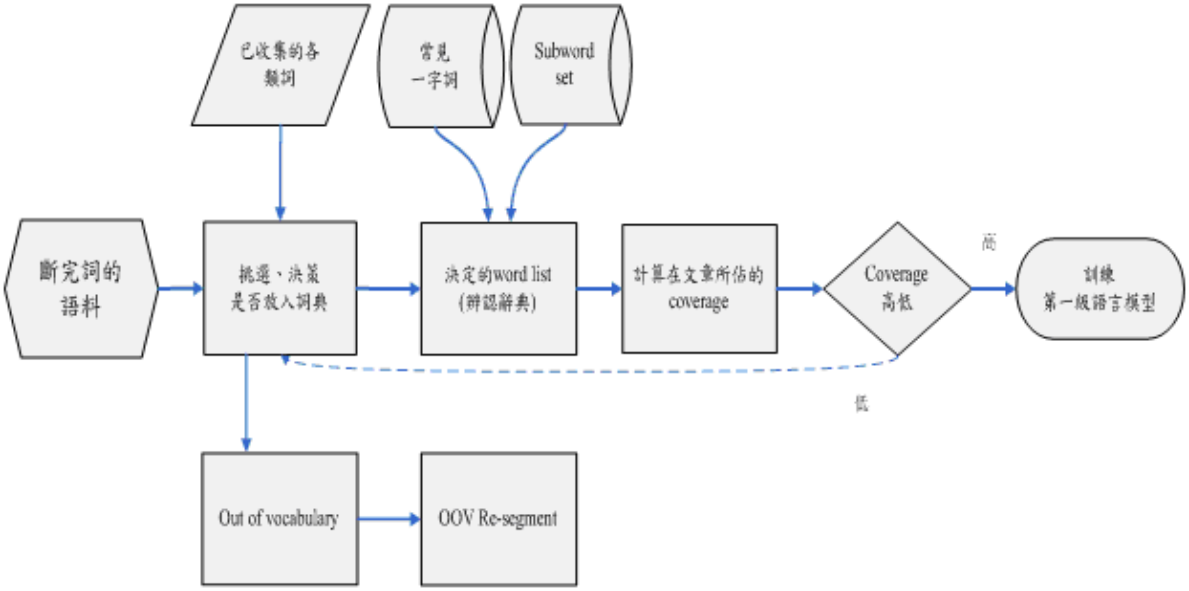

圖 3-2 混合式辭典挑選方法

然而，由於人名、詞綴和數量複合詞此三類在語料庫中變化多元且複雜，儘管辭典中無法完全收錄這三類詞，但此三類詞之組成結構上有各有其規則性，可依據其結構，將此些類別之詞個別拆解為 subword 半詞串，再將此些小單位 subword 半詞集合依詞頻高低，有比例性地收錄至辨認辭典中，促使第一級辨認後，word lattice 上相鄰節點可產生出 subword 半詞串，進而供給第二級構詞使用。

3.1 三類詞拆解方法

本研究針對人名、詞綴和數量複合詞三類進行拆解，拆解過程中，為避免辨認辭典收錄過多的 subword 短詞，反而犧牲原先未經拆解即可收錄之高詞頻一般詞空間，故將詞頻高之詞保留其長詞形式，不將之拆解為 subword 半詞串。研究者於詞頻排序後，依據範圍內之詞的 coverage rate，如表 3-1 所示，發現次序 45,000 之後，coverage rate 增加的速度趨緩，故此，將長詞的 threshold 設定為 45,000，次序於該值之內者則不進行拆解，以完整的長詞出現(如表 3-2 所示)，而次序在此值之外者，則進一步拆為 subword 半詞串。

表 3-1 各區間內的 coverage rate

次序	次序之內的 coverage rate
20,000	89.16 %
40,000	94.50 %
45,000	95.20 %
45,000+15,000[3]	95.48 %
60,000	96.67 %

表 3-2 各區間各類長詞分佈

詞彙分類	前四萬五千詞	四萬五千詞 至六萬詞	六萬詞之後
人名_詞條總數	1,440	1,102	99,866
人名_總數量	856,418	120,929	611,475
詞綴_詞條總數	3,511	1,720	13,133
詞綴_總數量	2,946,539	189,465	364,366
數量複合詞_詞條總數	3,706	1,759	260,404
數量複合詞_總數量	3,690,546	196,423	1,007,187
一般詞_詞條總數	36,343	10,419	85,136
一般詞_總數量	100,021,253	1,158,397	1,803,484

3.1.1 人名拆解方法

本研究之實驗室所收集的中文人名共有 414,280 條，而研究者針對文章中出現的中文人名詞條之詞進行拆解方法有兩種，於下詳述：

(1) 第一種：姓氏 + 名字

倘若將名字全收錄於辭典中，詞條數將會過多，約有 50,438 種，故研究者將常出現的名字以二字名字之 subword 形式收集於辭典中，人名收集於字典中，而此對於辨認用途比較小，通常只出現在人名部份，受收錄詞條數量所限制。

(2) 第二種：姓氏 + 名字的第一字 + 名字的第二字

對於未以二字名字收錄之 subword，研究者將其更進一步拆解為兩個連續一字詞相接，以更小單元收集之，而辭典內有收錄完整的一字詞，故此部份不多加考量收錄

[3] 此 15,000 指的是辭典內收錄 subword 之集合中詞數量

之多寡。

另一方面，在人名中，假使以名字為單位，可分成兩種情況(詳見下表 3-3)，分別為將保留名字全名，以及名字拆成連續一字詞此兩方式。

表 3-3 人名的拆解

人名	拆解結果
孔慶華	孔 慶華
方士豪	方 士豪
王文政	王 文政
王文珊	王 文 珊
林松源	林 松 源
蔣友華	蔣 友 華

3.1.2 詞綴拆解方法

本研究所收集之中文前後詞綴共 20,246 條，研究者針對文章中出現之中文詞綴詞條之詞進行拆解，而拆解方式有兩種(如表 3-4 所示)：

(1) 前詞綴：前綴詞 + 後接詞

(2) 後詞綴：前接詞 + 後綴詞

將詞綴拆解成以上兩種情形，而前接詞、後接詞若收錄在辨認辭典內，因為前後接詞具有一般詞的特性，出現頻率也比較高，則對於辨認上較有幫助，所以這裡將次序不在 45,000 threshold 值內的詞綴，均拆成前綴詞加上後接詞或前接詞加上後綴詞兩種，再收錄這些拆解過後產生的前接詞和後接詞 subword 短詞，而前後綴詞因我們會收錄全部一字詞，故這部份不考慮收錄的多寡。

表 3-4 詞綴的拆解

詞綴	拆解結果
下眼皮	下 眼皮
大工程	大 工程
大西瓜	大 西瓜
人事局	人事 局
上海人	上海 人
工作室	工作 室

3.1.3 數量複合詞拆解方法

本研究之數量複合辭之拆解方法，乃先建立一專門將數量複合詞斷成小單元 subword 短詞之斷詞辭典(如圖 3-3 所示)，辭典包含所有數字的變化、所有量詞(見附錄一)，進而運用所產生的數量複合詞 subword 斷詞辭典，將數量複合詞切割為 subword 小單元。

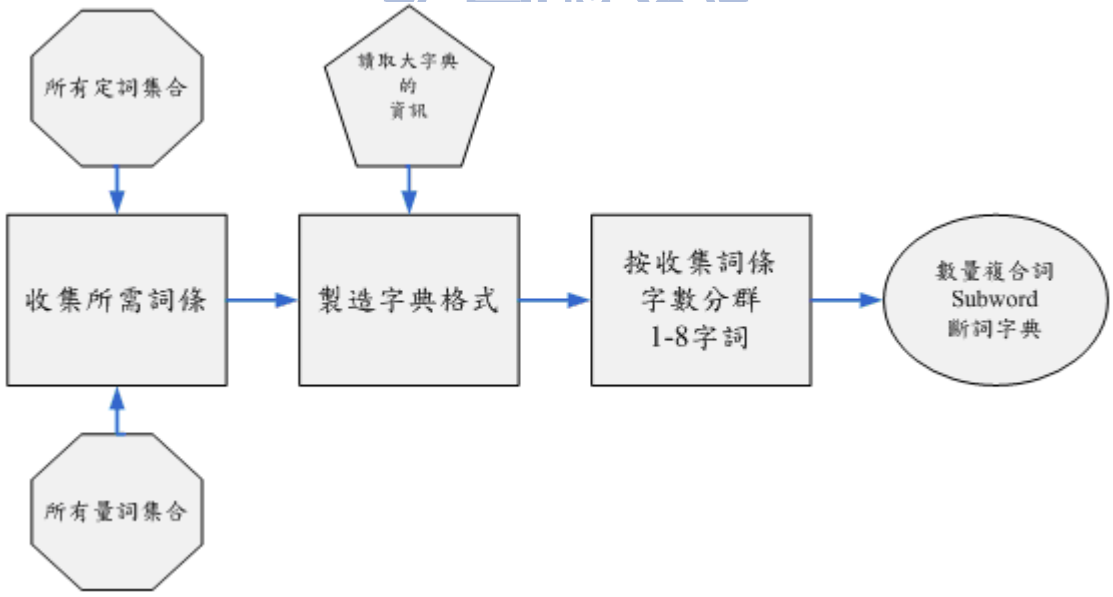

圖 3-3 數量複合詞 Subword 斷詞辭典

針對數量複合詞之拆解，本研究將次序 45,000 之內者保留長詞，次序 45,000 以上者均拆成小單元的 subword 半詞串，如下表(表 3-5)所示：

表 3-5 數量複合詞的拆解

構詞規則所構出的詞			拆解結果
定量複合詞	DM	兩萬平方公尺	兩萬 平方 公尺
數詞定詞	Neu	一千四百億	一千 四百億
數量定詞	Neqa	三十左右	三十 左右
時間詞	Nd	下午四點多	下午 四點 多
地方詞	Nc	六年十一班	六 年 十一 班
位置詞	Ncd	一百二十二度	一百 二十二 度

3.2 挑選辨認辭典的 subword 半詞

本研究將人名、詞綴和數量複合詞此三類有句法規則的詞拆解為 subword 短詞，因而產生此三類之 subword 集合。於此些 subword 集合中，有部份 subword 和詞頻排序 45,000 之內的詞重複，本研究之處理方法為將之視為已經收錄於辭典者，並針對未收錄至辭典的 subword 來斟酌收錄的 subword 詞條數量。

其中，詞綴、數量複合詞兩類之 subword 數量不多且重複性高，於語料庫出現之頻率較高，故本研究將此兩類之 subword 短詞全數收入辭典；然而，人名拆解後的名字由於通常只出現在人名，較不具有一般詞性質，故對名字類之 subword 收錄數量予以限制。換言之，收錄優先次序以詞綴、數量複合詞兩類為主，人名類則為最後考量。

如先前所述，去除詞頻排序 45,000 內本已存在的 subword 短詞後，其剩餘之詞乃為本研究考慮收錄之 subword 短詞，由表 3-6 可發現，姓氏、前後接詞、數量複合詞和全部一字詞此些集合之詞條數不多，故將之全數收錄至辭典中，在刪除各類彼此重複之 subword 後，共已收錄 11,932 詞；另一方面，辭典剩餘空間由人名之名字 subword 填補之，直至到達辭典 60,000 詞之容量上限為止，填補入辭典的名字 subword 數量為 3,068 詞，而其餘沒有以二字名字 subword 收錄的詞則拆成連續一字詞，而一字詞部分已全數收錄至辭典中。

表 3-6 各類詞剩餘 subword 詞條數量

分類	全部 subword	考慮收錄的 subword 短詞數
人名_姓氏	363	19
人名_名字	50,438	48,917
綴詞_前接詞	7,125	1,320
綴詞_後接詞	1,789	205
綴詞_前綴詞	136	0
綴詞_後綴詞	161	0
數量複合詞	1,729	873
全部一字詞	13,110	9,789

3.3 建立含 word 及 subword 的混合式辨認辭典

如先前所述，本研究跳脫傳統式辭典收錄方法，不再純粹單以詞頻大小來收錄 word 長詞，而是多運用某些辭典空間來收錄 subword 半詞，進而組成 word 和 subword 並存的混合式語言模型相對應之辭典。在下面圖 3-4 中，研究者試圖呈現本研究所採用之混合式辭典的建立流程，而該圖由左至右可分為三大步驟，每步驟之上半部方塊圖代表語料庫內容與狀態，下半部長條圖代表語料庫相對應的 wordlist，以細虛線代表次序在 45,000 詞的分界線，以粗虛線代表次序在 60,000 詞之分界線，如下：

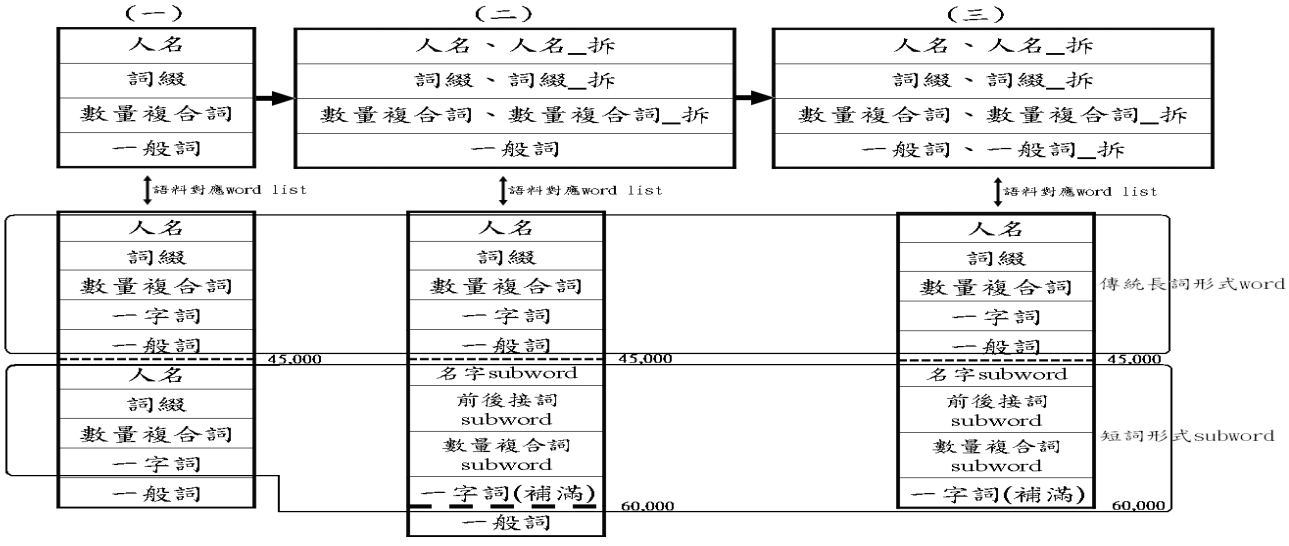

圖 3-4 Word/subword 混合式辭典建立流程圖

綜觀上圖(圖 3-4)，混合式辭典建立之三大步驟依序於下說明之。首先，步驟(一)之語料庫狀態為未進行人名、詞綴和數量複合詞之拆解，此三類保留其長詞型態存在於語料中，此時的語料庫經過錯誤文章的刪除、斷詞、標點符號處理和文字正規化之文字前處理動作；而其所對應之 word list 分為五種，包含：人名、詞綴、數量複合詞、一字詞和一般詞，經由詞頻排序後，對於次序在 45,000 內之詞不進行拆解，以完整的詞保存之，而次序於 45,000 以上之人名、詞綴和數量複合詞三類長詞則進一步拆解為 subword 短詞，接續邁入步驟(二)。

接下來，進一步檢視步驟(二)之 word list，研究者除將拆解後的 subword 短詞納入辭典外，並於辭典中補上未曾在次序 45,000 內出現的一字詞，進而運用辭典內次序 45,000(細虛線)到次序 60,000(粗虛線)之間的容量空間來收錄各類 subword 集合。如同 4.2 章所述，subword 集合收錄方法之優先次序以詞綴、數量複合詞兩類為主，人名類為輔；其中，儘管人名類拆解為姓氏 subword 與名字 subword，詞綴類拆解為前後綴詞 subword 及前後接詞 subword，但由於姓氏 subword 與前後綴詞 subword 均為本辭典已完整收納之一字詞，故僅考慮收錄名字 subword 以及前後接詞 subword。在收納完 subword 集合後，則大抵決定本研究之辨認辭典，然而，超出辭典容量限制(60,000 詞)之一般詞，即為步驟三所欲處理之 OOV。基此，就步驟(二)之語料庫狀態來說，其同時存在保留長詞型態與拆解後之 subword 短詞形式的人名、詞綴和數量複合詞，而一般詞則原始之長詞狀態呈現之。

最後，於步驟(三)，為處理 OOV 問題，將對於未收納於辭典內的一般詞進行拆解，而本研究解決 OOV 之策略為運用辭典內的六萬詞形成斷詞字典，將之切割為辭典內所收納之詞，故所有中文詞將均轉變為混合式辨認辭典內所涵蓋之詞。然而，本研究乃針對中文語音辨認進行研究，文字前處理之英文串(類別 LONGFW)未進行處理，因而於語音辨識時，類別 LONGFW 仍被視為 OOV。

混合式辭典建立過程中，三個階段語料庫所相對應之詞條數與詞總數變化如下表(表 3-7)所示：

表 3-7 混合式辭典建立各步驟之相對應詞條數與總數量

分類/單位	步驟(一)	步驟(二)	步驟(三)
詞條數/條	518,539	165,071	55,471
詞總數/個	112,966,482	119,652,666	128,092,663

由表 3-7 可發現，第三步驟之詞條數為 55,471，與研究原先設定之辭典容量的 60,000 條有所差距，其乃因本研究所設計的混合式辭典全數收錄一字詞，然而，原始語料資料庫中並無涵蓋所有中文一字詞，因而形成詞條總數之落差，由此可知，混合式辭典中仍有 4,529 條的集合空間尚未完整運用。

相較於傳統式語言模型之辭典，本研究所採用的辭典僅收錄詞頻次序於 45,000 以內之常見長詞，進而運用剩餘的詞條空間來收錄三類長詞經拆解後產生的 subword 短詞，如下表(3-8)所示：

表 3-8 word/subword 混合式語言模型辭典

Language Model Lexicon		
總詞數	60866	
詞長	**詞數**	**百分比(%)**
1	13060	21.83
2	32313	53.80
3	9950	16.58
4	3798	6.33
5	551	0.92
6	298	0.49
7	30	0.05

LM 辭典的平均詞長 = 2.124

3.4 OOV 處理

解決 OOV words 方式和訓練傳統語言模型時相同，針對未收錄於六萬詞辭典之 OOV word【7】進行拆解，將 OOV word 拆解為可由辭典內的詞所構成。

3.5 N-gram training

訓練 word/subword 語言模型過程與建立傳統式語言模型相似，兩者之差異，如先前所述，於長詞拆解為 subword 短詞部份，不再僅針對 OOV words，而是將人名、詞綴和數量複合詞三類均拆解成 subword 短詞，Word/subword 混合式語言模型建立過程如下圖所示(圖 3-5)：

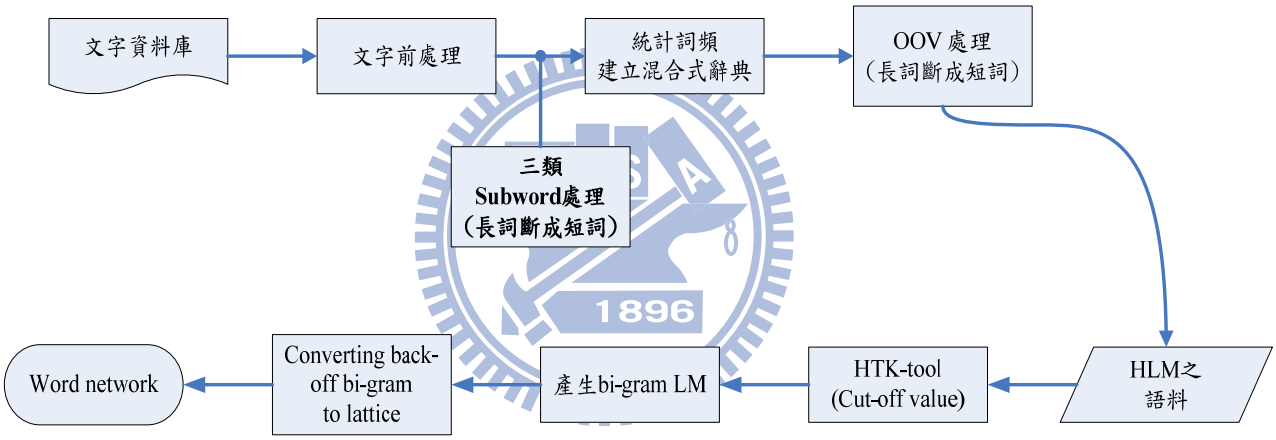

圖 3-5 Word/subword 混合式語言模型建立

3.6 Smoothing (cut off value)

本研究之 smoothing 採用較常見之 katz smoothing，數學式如下

$$P(w_i \mid w_{i-n+1},\mathrm{K}\ ,w_{i-1})$$

$$=\begin{cases} a(w_{i-n+1},\mathrm{K}\ ,w_{i-1})P(w_i \mid w_{i-n+2},\mathrm{K}\ ,w_{i-1}) & Count(w_{i-n+1},\mathrm{K}\ ,w_i)=0 \\ d_a \dfrac{Count(w_{i-n+1},\mathrm{K}\ ,w_i)}{Count(w_{i-n+1},\mathrm{K}\ ,w_{i-1})} & \min \le Count(w_{i-n+1},\mathrm{K}\ ,w_i) \le \max \\ \dfrac{Count(w_{i-n+1},\mathrm{K}\ ,w_i)}{Count(w_{i-n+1},\mathrm{K}\ ,w_{i-1})} & Count(w_{i-n+1},\mathrm{K}\ ,w_i) > \max \end{cases}$$

Smoothing 之 cut off 值的選擇，將影響到語言模型之參數量，當 cut off 值愈小時，

訓練語言模型則需較多的資料量，相對下使用 back off 之資料比率則較少，其語言模型之模糊度較低，辨識率較高，故就語音辨認而言，選擇適當的 cut off 值顯得相當重要。

3.7 混合式 language model 之探討

本研究中混合式語言模型是為階層式架構的第一級語言模型，藉由人名、詞綴及數量複合詞有規則地拆解成較短的半詞，使得此些有規則但未被收錄至辭典中的長詞，可在第一級辨認結果中能以半詞（subword）的型態被辨識出來，然而再透過第二級之構詞模組，將這些連續地半詞串進行構詞，使其恢復為原來的長詞型態，並經過 rescoring 模組產生第二級語言模型，透過此兩階段架構使語言模型增進長詞之辨識效能。

第四章 第二級語言模型

從第一級 word/subword 混合式語言模型延伸至第二級語言模型，主要是藉由第一級 word lattice 進行構詞擴充至第二級 extended word lattice，將 subword 短詞串構詞轉變為意義長詞，在構詞的過程中，extended word lattice 上會產生新的節點和弧，並給予新的路徑分數，而研究者將於本小節針對新產生之節點數和路徑選擇加以討論。

本研究採用 two-stage 方法來提升長詞之辨識率，而第二級主要可分為兩部分：

（1）word lattice 構詞：在 word lattice 上，將可構回長詞之 subword 串均構成長詞，目前人名和詞綴先以查表方式查詢 word lattice 上哪些 subword 可構回人名或詞綴此兩類長詞，而數量複合詞則運用 Finite State Machine (FSM)架構將此些 subword 構回數量複合詞。

（2）語言模型分數分配：在 extended word lattice 上將會產生新的節點(Node)、連接節點與節點間的弧(Arc)，而在 Arc 上必須給予此些新的路徑聲學模型分數和語言模型分數。

4.1 Word lattice 構詞

在第一級 word lattice 上，透過構詞模組來將 word lattice 擴展成第二級 extended word lattice，換言之，將第一級 word lattice 上符合長詞的 subword 短詞組合構回有意義之長詞，本研究中 word lattice 有進行構詞的詞類包括：數量複合詞、人名、詞綴以及 OOV，分別以下敘述之：

4.1.1 人名、詞綴及 OOV 構詞構詞

目前此部份尚未建立長詞構詞模型，暫時以查表法(search table)來替代之，將實驗室收集之人名、詞綴和第一級文字資料庫處理時產生的 OOV words 視為表(table)，進而可在 word lattice 把相鄰節點(均為詞)組合成長詞來表上搜尋(search)比對，如有符合表中之長詞者即將此長詞構出(即 word lattice 上產生新節點)。

4.1.2 數量複合詞構詞

數量複合詞的構詞本論文採用 FSM (Finite State Machine, 有限狀態機)的架構來產生，透過每個狀態(state)轉移時串接成詞來觀察是否符合構詞規則，凡是符合規則的詞則將路徑所經過的狀態依序串接產生構詞後的數量複合詞。在此之前，必須建立數量複合詞的 FSM 架構，其中，欲建立此架構則考慮三個部份：首先，決定數量複合詞的半詞集合，以及該集合與狀態之間的關連；其次，建立一個以 state transition 為主體的 FSM 架構；最後，訓練 FSM 架構下的 state transition probability。

經過以上三個部份來組成如下圖(圖 5-1)之架構：

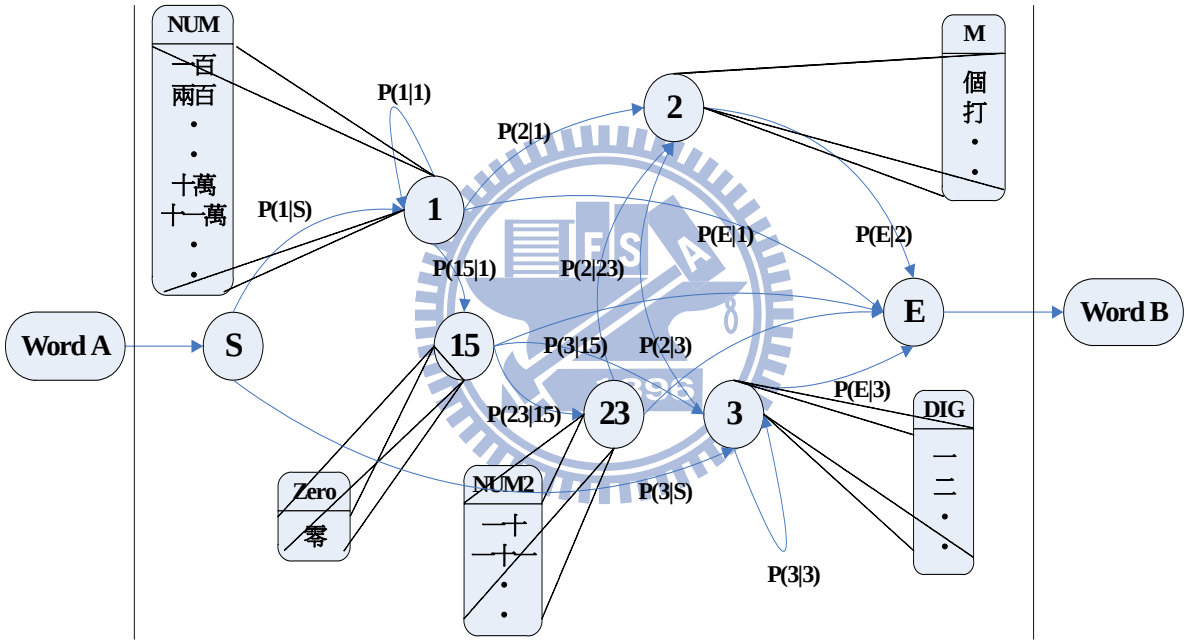

圖 4-1 數量複合詞構詞模型

在圖 4-1 中，該構詞模型乃以 Finite State Matchine(FSM)架構組成之，此架構中依據斷詞器構詞時所需之構詞集合來設定狀態(State)，再透過構詞規則使得狀態之間有規則地彼此相連，其中，S 代表構詞時長詞之開始狀態，而 E 代表構詞時長詞結束之狀態，經構詞後產生的數量複合詞意指將 S 和 E 之間經過之狀態彼此串接而形成的詞。此處之狀態共有 114 個，其中，數字狀態佔有 28 類(共 28 個狀態)，包含所有數字之變化型態，而量詞狀態則佔有 22 大類(共 86 個狀態)，是經由實驗室斷詞器中使用之

構詞規則中擷取出來的。

在 FSM 架構中須存在狀態轉移機率，透過這些機率來描述在數量複合詞該類別中可組合成各數量複合詞之情況，且這些半詞串接的總轉移機率將在第二級 extended word lattice 時被使用當做數量複合詞之內部機率(intra probability)來進行語言模型的分數配置。

4.1.2.1 定義定量複合詞半詞集合及該集合與狀態的對應關係

如第三章所述，數量複合詞半詞集合乃收集全部數字變化型態及構詞規則中之所有量詞集合，在此將該集合分成兩大類：一類是純粹與數字有關，另一類則是數字以外的定詞、量詞、修飾詞等。

- **數字集合**

針對數字串組合方式，以結構分析方式將數字組合分類，分類原則如下：

a. 兼顧構詞能力、平均詞長、詞義完整度

譬如『一百萬』可以拆成(1)『一』『百』『萬』 (2)『一百』『萬』 (3) 『一』『百萬』及(4)『一百萬』此四種。

其中，以構詞能力而言，優先順序為(1) > (2) (3)>(4)，而以平均詞長而言，順序為 (4)>(2) (3)>(1)，另外，以詞義完整度排序則為(4)>(2)>(3)>1。

最後在整體考量下，以收錄(4)形式之半詞為佳。

b. 同質性高者歸為一個集合

譬如『一百萬』與『兩百萬』同結構、同性質，且有同樣的語法角色，則將『一百萬』～『九百萬』歸為一個集合。

另外如『一十一』與『一十九』有同樣用法，則歸到同一個集合。而『一十一』與『十一』有不同用法，則將兩種組合分開到不同集合中。

以此原則將數字集合做分類，總共得到 28 類集合，分別給予集合與 state 對應如下表(表 4-1)所示：

表 4-1 數詞半詞集合與各集合對應 state ID

集合	State ID	集合	State ID
一～九	501	一千億～九千億	515
十～十九	502	一兆～九兆	516
一十一～九十九	503	十兆～十九兆	517
一百～九百	504	一十一兆～九十九兆	518
一千～九千	505	一百兆～九百兆	519
一萬～九萬	506	一千兆～九千兆	520
十～十九萬	507	一二～八九 一二十～八九十	521
一十一萬～九十九萬	508	零	522
一百萬～九百萬	509	百	523
一千萬～九千萬	510	千	524
一億～九億	511	萬	525
十一億～十九億	512	億	526
一十一億～九十九億	513	兆	527
一百億～九百億	514	兩	528

其中在上表(表 4-1)中，『一十一』～『一十九』沒有獨立成一個集合，原因是其角色在長結構的數詞串中與『二十』～『九十九』有相同角色(譬如『一百一十五』與『一百七十五』的例子)。

- **定詞、量詞、修飾詞集合**

以目前斷詞系統所採用的定詞、量詞集合和分類做集合類別與狀態做對應。共 22 大類，86 個集合(86 個狀態)，分布如下表(表 4-2)所示：

表 4-2 定詞量詞分類與其對應 state ID

定詞量詞類別	State ID	定詞量詞類別	State ID
標準量詞	92-97	動量詞	90
暫時量詞	91	特殊數量定詞	3-5
數詞定詞	0-2	特指定詞	12-13, 38-39, 41-43
疑問數量定詞	7	數字相關	17, 40, 44
跟述賓式合用的量詞	90	容器量詞	90

群體量詞	90	個體量詞	90
準量詞	90	指示定詞	11
程度數量定詞	9	前程度副詞	8
部分數量定詞	10	全體數量定詞	6
修飾詞	14-16, 18-26, 31-37,	地址班級	27-28, 30
時間	51-82	其他	29

4.1.2.2 以構詞規則建構 FSM 架構

依據上一小節 4.1.2.1 所決定出的狀態集合及其所對應的半詞集合，將目前實驗室斷詞系統所使用的構詞規則，轉換成以狀態表示法所形成的規則，以此來建構 FSM 系統所需之狀態轉移路徑。

在建構定量複合詞 FSM 時，為了減少 search space 大小，進而採用 subnet 概念實現，將數字部分與數字以外的部分各自建立一個 FSM 結構，再將兩個 FSM 結構結合起來以建構數量複合詞整體的 FSM 架構，此類架構以下圖(圖 4-2 和圖 4-3)來表示。

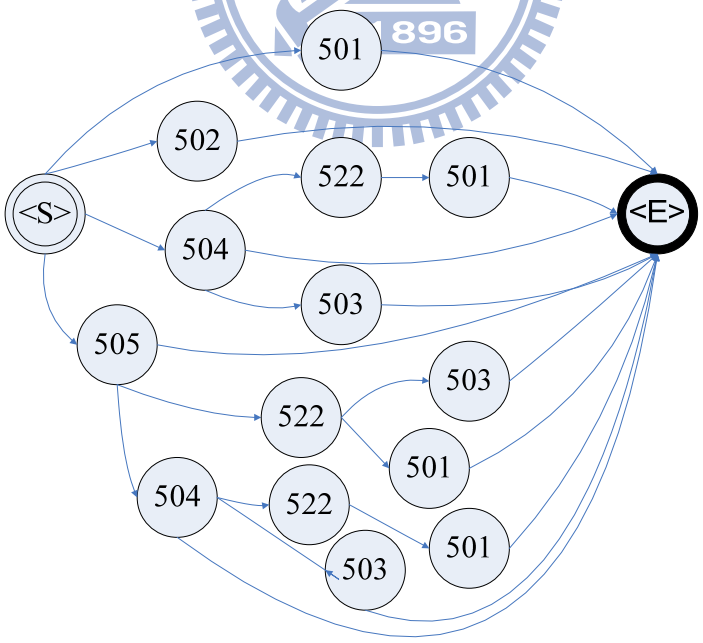

圖 4-2 數字串 FSM

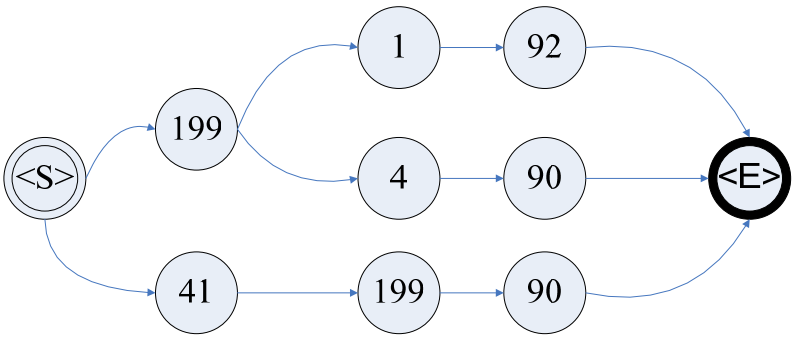

圖 4-3 定量複合詞 FSM

圖 4-2 中，數字串範圍為一到九九九九，各狀態代表意義如下：狀態 522 代表「零」、狀態 501 代表「一至九」、狀態 502 代表「十至十九」、狀態 503 代表「一十一至九十九」、狀態 504 代表「一百至九百」、狀態 505 代表「一千至九千」，此架構為數字串之 FSM 結構，在整體數量複合詞架構下視為「狀態 199」。

圖 4-3 中，為數字部分(狀態 199)和非數字部份搭配在一起的結構，其各自狀態連結起來代表意義如下：

I. 數字(199)+幾(1)+長度單位(92)

II. 數字(199)+多(4)+量詞(90)

III. 不到(41)+數字(199)+量詞(90)

由上述可觀察到此作法雖然降低 search space 的大小，但此舉會造成許多路徑共用同一個數字 state(狀態 199)的情形，並且必須多考慮 inter state transition probability 和 intra number state transition probability。然而此法的好處是，往後可以針對數字部分獨立處理，包含分出獨立類別、單獨建立各種型態的數字 FSM 結構。另外，將來也可以考慮將數字部分的 intra state transition probability 做正規化動作，讓某一結構的數字組合享有共同的機率。

4.1.2.3 訓練 FSM 架構下的狀態轉移機率(state transition probability)

建構出兩層式的數量複合詞 FSM 之後，必須利用訓練語料中出現的數量複合詞來訓練 FSM 中各狀態的轉移機率。訓練方式為將擷取出來的數量複合詞轉換成狀態序列(state sequence)，再將此些序列依次數均輸入至 FSM 架構中，而來統計每一次狀

態轉移時所產生的狀態轉移次數，以此些次數來計算狀態轉移機率。

在此以 4.1.2.2 所述之 FSM 架構為例，針對以下輸入語句做訓練過程的介紹。

步驟一. 將輸入詞串轉換成狀態序列

將訓練語料庫中數量複合詞串轉換成狀態序列，並作為 FSM 的輸入資訊。以下為幾個輸入詞串轉換為 state sequence 的例子。

語句輸入 => 轉換後的狀態序列對應

1. 一千兩百多個 => 505 504 4 90
2. 不到五百三十個 => 41 504 503 90
3. 一百五十幾公尺 => 504 503 1 92'

步驟二. 計算狀態轉移次數

基於前述兩層式 FSM 架構，計算不同輸入序列所造成的狀態轉移次數。當輸入詞串遇到數字狀態時，則進入階層式數字 FSM 中，同樣計算所走路徑次數。

下圖(圖 4-4)為將上述輸入詞串作為 FSM 訓練過程中計算狀態轉移次數的範例。

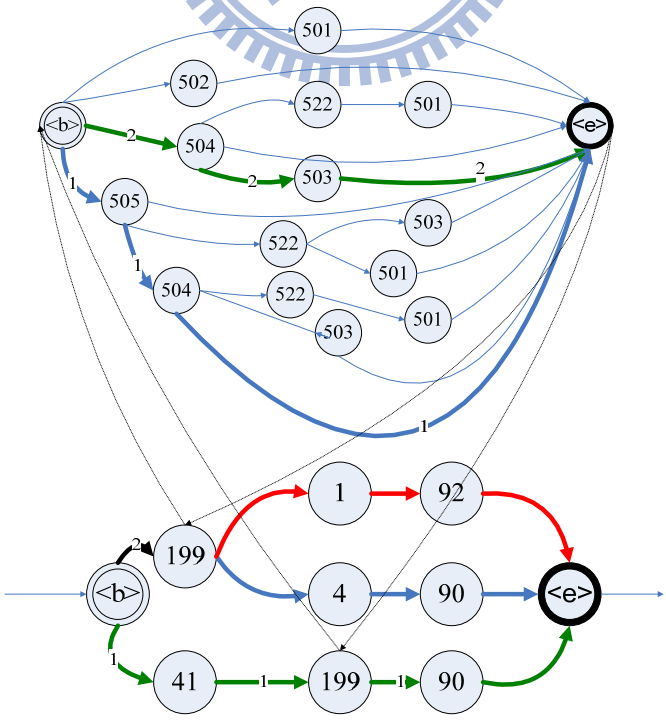

圖 4-4、計算輸入語句於 FSM 裡所造成的狀態轉移次數。

步驟三. 計算狀態轉移機率

依據不同路徑下的各狀態所記錄的下一狀態出現個數和次數，依轉移機率總和等於 1 的方式計算各狀態轉移機率。

由於不同路徑有不同的狀態轉移方式，即使是相同狀態，在不同路徑下也會有不同的狀態轉移機率，只有在相同路徑下的分支路徑才會共用狀態轉移機率，而這些轉移機率將會使用在第二級 extended word lattice 中數量複合詞之內部機率的分數配置。

步驟四. 建立定量複合詞的 baseline FST Model

基本的定量複合詞 FST Model 的建立，可以先將整個訓練語料中所含的 FT word 視為一個類別(FT class)，並在這個類別下定義兩層 FST，包含第一層 FST 的 Inter Word Model 和第二層 FST 的 Intra FT Word Model，架構如下圖所示。

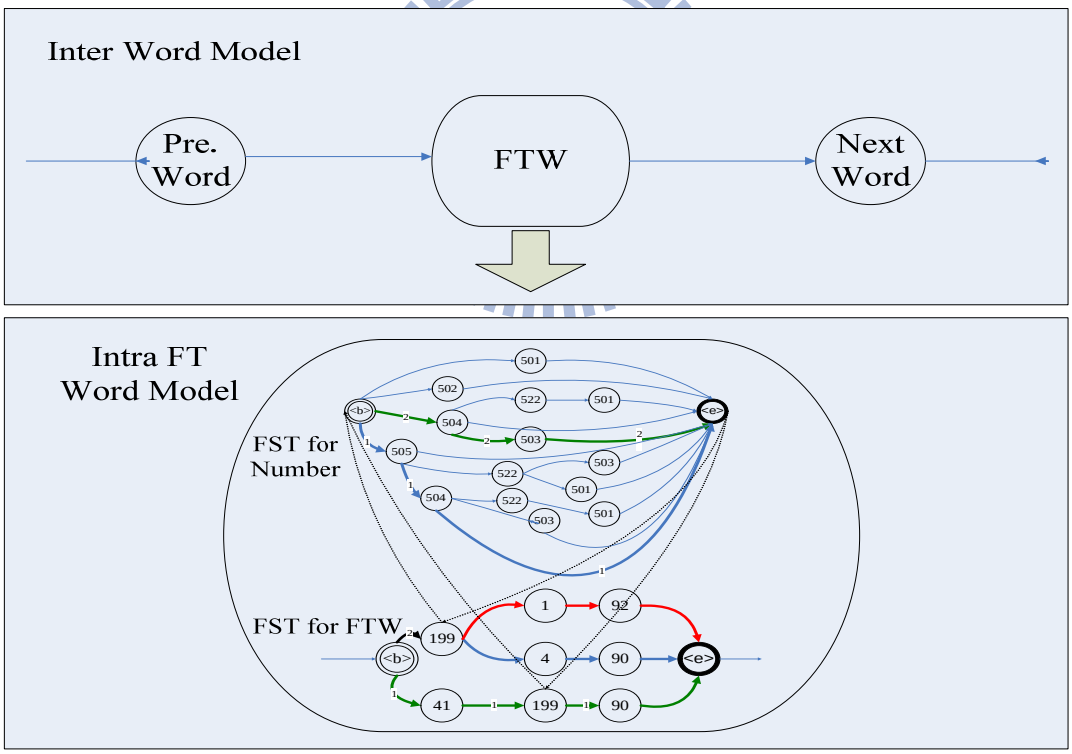

圖 4-5、適用於定量複合詞的兩層式 FST Model 架構

在圖 4-5 的 inter word model 中，將 FT word (FTW)視為一個類別，直接訓練數量複合詞此類別與其他詞彙間的 N-gram 機率模型，此為外部機率(inter probability)。而

在 intra FT word model 中，建立 FST for FT word(定量複合詞 FSM 結構)及 FST for number(數字串 FSM 結構)兩層 FST，針對第一層 FT word 的狀態轉移，以訓練語料統計出各狀態間的狀態轉移次數，並針對每一個狀態利用轉移機率和等於 1 來計算其狀態轉移機率。對於 FT word 的數字串狀態部分(狀態 199)，則再將數字狀態獨立建出一個純粹由數字半詞所建構的 FST，以同樣方式來計算所有數字組合所造成的狀態轉移機率。

4.2 語言模型分數配置

本研究對於 extended word lattice 新產生路經給予分數的方法有二種：

第一種方法為 **word penality 影響**：從未構詞前 subword 短詞串的分數直接相加給構詞後的弧，此時 word penality(數值不更變)之存在會使得原半詞串路徑將被壓抑，而選擇構詞後路徑，進而影響 extended word lattice 上最佳路徑，產生第二級辨認結果。

第二種方法為**改變語言模型分數**：將構詞的人名、詞綴和數量複合詞均以類別取代，進而透過 inter probability(外部機率)和 intra probability(內部機率)來分別計算一般詞和類別間的機率以及某類別內出現長詞之機率，藉此兩機率來給予新的語言模型分數。

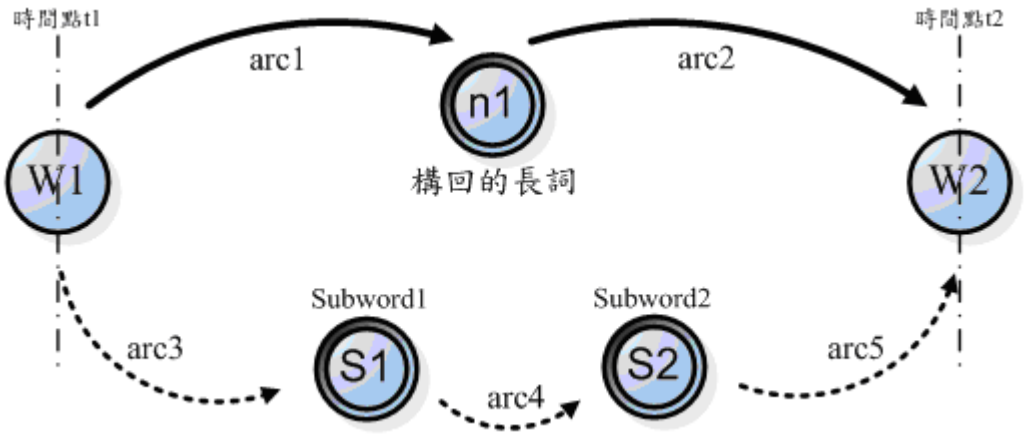

圖 4-6 新產生的節點和弧

在圖 4-6 中，arc1-arc2 和 n1 為 extended word lattice 新產生之弧(arc)和節點(node)，arc3 至 arc5、S1 和 S2 則為尚未構詞前第一級語言模型 word lattice 上的弧和節點，S1

和 S2 均為 subword 短詞，可進行構詞得到長詞 n1。

4.2.1 Word penality 影響

本研究不改變語言模型分數，而是將第一級 word/subword 語言模型中短詞串的分數直接相加，以此給予構出長詞弧上分數，進而藉由 word penality 參數 **p** 和 **s** 來影響最佳路徑，換言之，在 word penality 數值不變之下，會使得 subword 半詞串的原路徑會被壓抑，轉而比較相信構詞後之路徑。

- **Word penality 係數 p 和 s**

傳統上訓練語言模型時，產生 word lattice 時將會從統計的 bi-gram 機率來給予這些路徑分數，而當進行辨認時，在路徑上利用 word penality **p** 和 **s**(即第二級和第一級設定值相同)此兩個參數來影響弧上原始之聲學模型和語言模型的分數，重新尋找構詞後辨認語料之最符合、最接近的路徑，以求得到最高辨識效能。

微調前(弧上)：	**受 p 和 s 影響後(弧上)：**
聲學模型分數：A	聲學模型分數：A + p
語言模型分數：L	語言模型分數：L • s

- **受 word penality 影響之原理**

每經過一個弧，弧上的分數將會被 **p** 和 **s** 所影響，在某路徑上節點數越多時，其通過之弧數量也越多，分數變動將會越大，因此，在加入 **p** 和 **s** 後，倘若 **p** 為負值時，兩條原始分數相同的路徑，會因為該路徑上節點數越多導致分數會越來越低。基此，於最佳路徑時，將會選擇節點數較少路徑者。然而，我們在 word penality 沒有變動下，使得最佳路經相對地信任構詞後的新路徑進而選擇該構詞之路徑，則將第一級時拆解成 subword 短詞串透過第二級構詞後恢復為原有之有意義長詞。

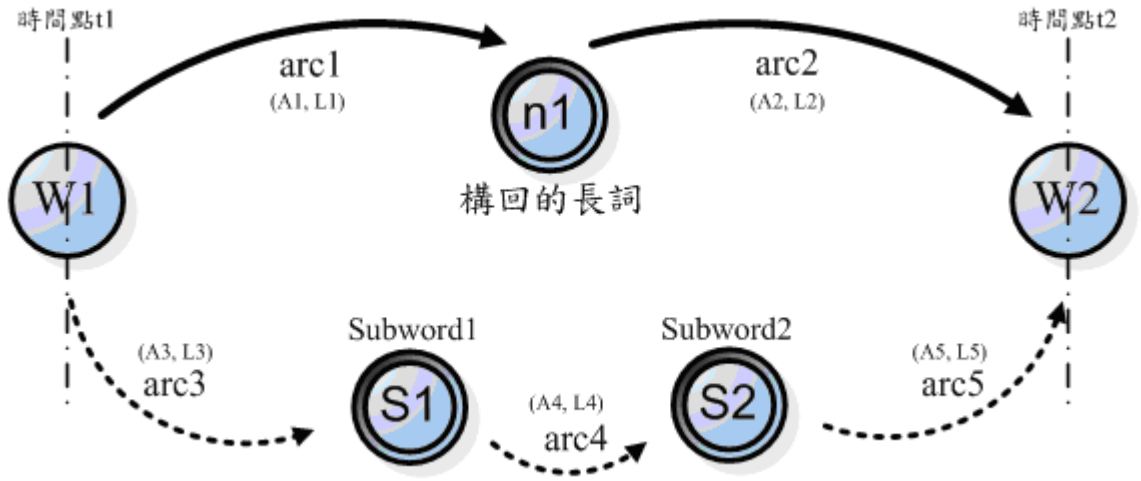

圖 4-7 弧上的 word penality

於上圖 4-7 中，A1 到 A5、L1 到 L5 分別為 word lattice 上 arc1 至 arc5 的原始聲學模型和原始語言模型分數，而長詞 n1 則是短詞 S1 和短詞 S2 構詞出來。

基此，新路經的分配如下：

arc1:	arc2:
A1 = A3 + A4	A2 = A5
L1 = L3 + L4	L2 = L5

因而產生出兩條分數相同的路徑，研究者進而透過 word penality 係數的影響，藉此選擇最佳路徑，而如先前所述，路徑經過越多節點(或弧)則會增加愈多 **p** 值，當 **p** 是負值時，則多節點之路徑分數將越低。

研究者於下列舉兩個例子進行說明：

例子一

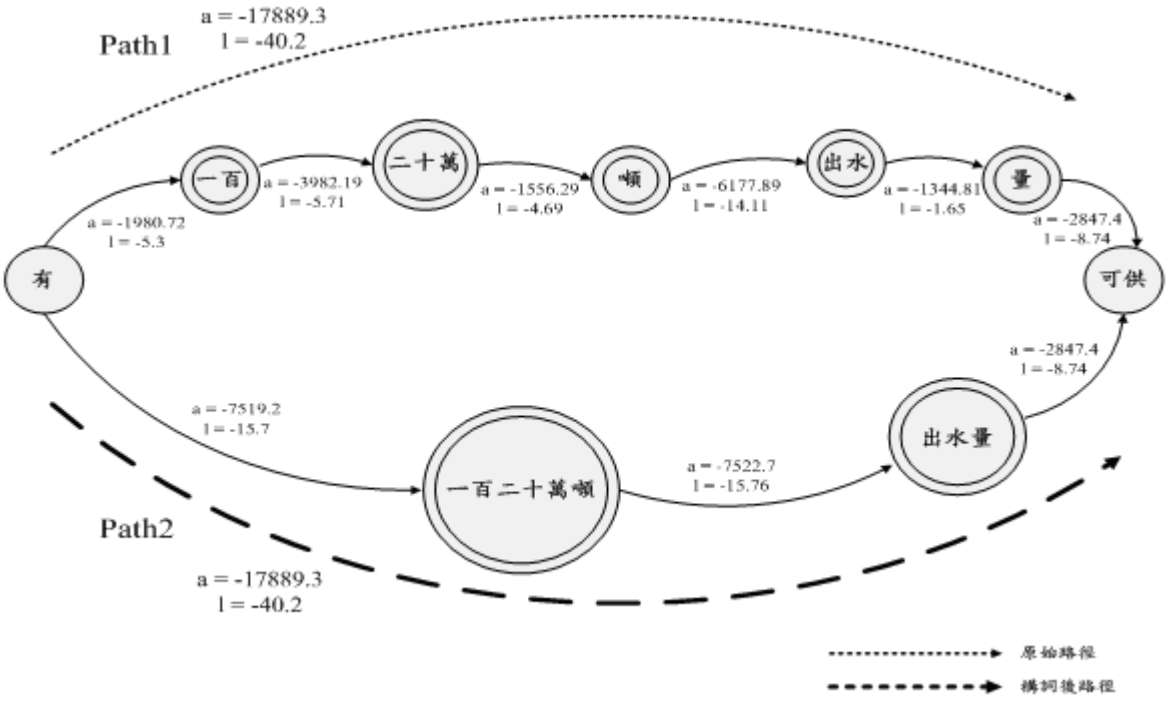

圖 4-8 正確構詞

在上圖 4-8 中，Path1 為第一級 word/subword 語言模型之最佳路徑，Path2 第二級語言模型構詞後產生的新路徑，由於 Path2 的分數是由 Path1 相關節點間相加得來，當 Path1 為第一級時最高分數的路徑時，此時將會有兩條最高分數的路徑，即 Path1 和 Path2，當兩條路徑分數相同時，因為 Path1 會經過 5 個節點和 6 個弧、Path2 經過 2 個節點和 3 個弧，故當加入 word penality 後，Path1 將會比 Path2 多增加三個 **p** 值，且當 **p** 是負值時，Path2 的分數則會比 Path1 來得大，基此，辨認結果可透過 word penality 影響(但不變更其數值，保留與第一級相同之值)，最後將會決定選擇 Path2 做為最佳路徑。

例子二

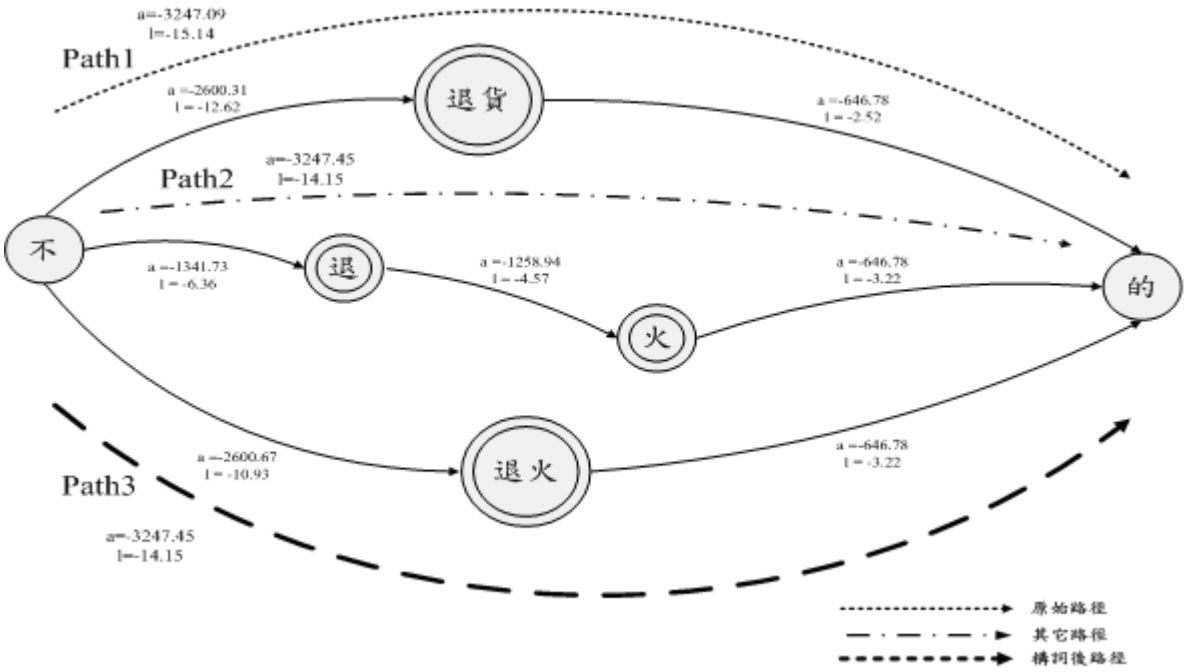

圖 4-9 錯誤構詞

於此例子中，顯示出在第二級構詞時，將會有「過分構詞」的現象，當第一級 word/subword 語言模型辨認正確路徑時，經過構詞後，節點數變少了，使得原先因 word penality 分數落後的路徑 Path3 再度取得分數領先，形成過分構詞的詞出現所導致的錯誤。

路徑分數變化如下：

第一級(未受word penality影響) ： Path2 > Path1
第一級(加入word penality之後) ： Path1 > Path2
第二級(未受word penality影響) ： Path2 = Path3
又 Path2 > Path1
⇒Path3 > Path1
在第二級(受word penality影響) ：Path3 > Path1
此時word penality亦和第一級時相同
所以透過一樣的word penality影響，會使得在第二級較信任構詞後的路徑
而改變原先之路徑

由圖 4-9 可知，Path1 為第一級的最佳路徑，Path2 則為第一級因 word penality (p = -15 , s = 15)影響而落後給 Path1 的路徑(Path2 原始分數領先 Path1 原始分數)，Path3 為第二級構詞後的最佳路徑。在第一級 word/subword 混合式語言模型中，那時並不存在 Path3，因為 Path3 是 Path2 的節點經過第二級構詞後所產生的，而新路徑 Path3 上的分數是由 Path2 的分數相加，因此，Path2 和 Path3 兩條路經在還未受 word penality 影響之前分數將會相同。

第一級時，未受 word penality 影響之前，Path2 分數大於 Path1 分數，而第一級辨認時加入 word penality 後，因而使得 Path1 分數大於 Path2，故 Path1 為第一級辨認結果。

第二級時，透過構詞產生了新路徑 Path3，Path3 分數和 Path2 分數相同(如先前所述，未受 word penality 影響之前，此兩路徑分數相同)，所以在第二級未加入 word penality 前，Path3 的分數會大於 Path1，但第二級辨認在使用 word peality(p 和 s 保持和第一級相同數值狀況下)後，由於 Path1 和 Path3 均為 1 個節點和 2 個弧，Path3 分數將會大於 Path1 分數，第二級辨認結果則變成 Path3。

4.2.2 改變語言模型分數

研究者於上一小節探討保持第一級和第二級 word lattice 在條件相同情況下，透過 word penality 之影響，使得原路徑被壓抑，進而選擇構詞後的新路徑。於本小節中，研究者將改變第二級語言模型分數，即第二級 extended word lattice 上路徑分數不再直接由第一級未構詞前之半詞串路徑相加而來，而是對構詞的人名、詞綴和數量複合詞視為三種類別，再由語料裡去統計類別和詞之間的機率(外部機率, inter-probability)，以及類別內出現某長詞的機率(內部機率, intra probability)，藉此兩機率對第二級之 extended word lattice 的路徑重新配置分數，再使用此語言模型進行辨認。

本論文則利用以下數學式來重新計算第二級語言模型分數：

$$P(W_n|W_{n-1}) = P(C_n|C_{n-1}) \cdot P(W_n|C_n) \quad , \quad \text{for } C_n, C_{n-1} \in \{\text{FT, MD, PN, other}\}$$

$$\text{其中 } P(C_n|C_{n-1}) = \begin{cases} P(W_n|W_{n-1}) & , \text{ for } C_n \in \text{other and } C_{n-1} \in \text{other} \\ P(C_n|C_{n-1}) & , \text{ for } C_n \in \text{FT, MD or PN and } C_{n-1} \in \text{FT, MD or PN} \\ P(W_n|C_{n-1}) & , \text{ for } C_n \in \text{other and } C_{n-1} \in \text{FT, MD or PN} \\ P(C_n|W_{n-1}) & , \text{ for } C_n \in \text{FT, MD or PN and } C_{n-1} \in \text{other} \end{cases}$$

$$P(W_n|C_n) = \begin{cases} \prod_{j=1}^{L} P(SW_j|SW_{j-1}, C_n) & , \text{ for } C_n \in \text{FT} \\ P(W_n|C_n) & , \text{ for} C_n \in \text{MD or PN} \\ 1 & , \text{ for otherwise} \end{cases}$$

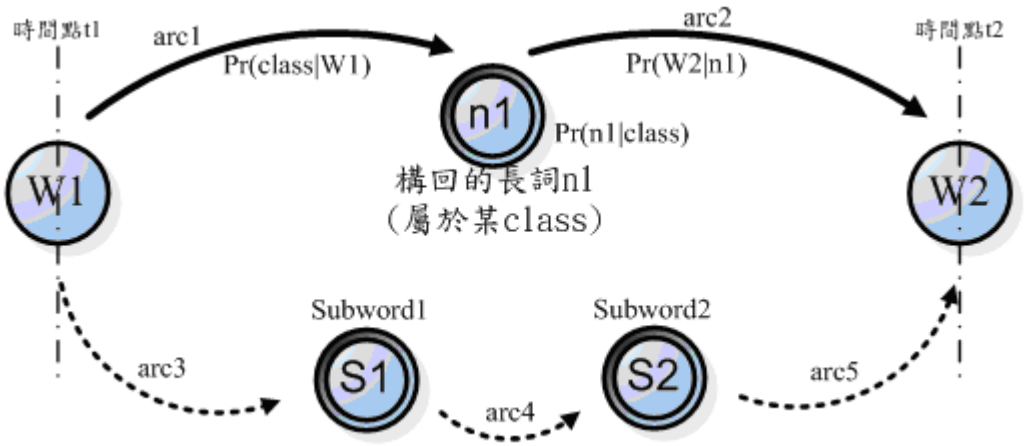

圖 4-10 弧上語言模型新分數

n1：新產生的節點(Node)，屬於某個 class，arc1：新產生的弧(Arc)

先將構詞出的長詞依分類視為某個類別(class)，有人名(PN)、詞綴(MD)及數量複合詞(FT)等三種 class，其中再將詞綴分類細緻化，將前後詞綴依綴詞分成前詞綴 142 個類別和後詞綴 159 個類別，最後統計計算某類別中出現某長詞 n1 的機率。

Pr (arc1) = Pr(class|W1) · Pr (n1|class) 給予新路徑弧上的分數
Pr(class|W1) 是 inter-word probability 描述詞和類別之間的機率
Pr (n1|class) 是 intra-word probability 描述類別中出現某長詞的機率

此處以下兩點來描述內外部(inter and intra)機率之求法

A. 外部機率(Inter-word probability) Pr(class|W1) **的求法：**

將第二級構詞所產生的詞中，把人名、詞綴和數量複合詞視為三大類別{PN, MD, FT}，再將三個大類別各自細分成多個小類別，計算一般詞和這些類別之間連接的機率。目前由於類別細分需要較長時間，目前實驗室在分類上尚未完整，究此，暫以簡單之類別進行分類 (以詞綴為例)，最後可觀察到類別細緻化後之實驗結果對本研究所提出路徑分數給予之方式有正面幫助，使辨認效果更佳。

在詞綴類別細分上，以「綴詞」來區分類別，將前後詞綴中具相同綴詞的詞放置同一類別，如此下來有 301 個詞綴類別及人名和數量複合詞兩個大類別共有 303 個類別，進而使用這些類別來計算分數配置時所需的外部機率，計算過程如下：

➢ **機率計算步驟如下：**

步驟 1：將語料中不在辨認辭典中人名、詞綴和數量複合詞分別以上述類別存在。

✧ 說明：目前由於人名和數量複合詞分類尚未建立，這裡僅針對詞綴加以分類，分類方法如前面所述，再將訓練語料中且不在辭典內的三類詞以這些類別替代。

步驟 2：對以類別替代後之訓練語料進行統計來求取詞和類別之間的 bi-gram 機率。

B. 內部機率(Intra-word probability) Pr (n1|class) **的求法：**

於此部分，機率採人名、詞綴和數量複合詞各自的類別分開獨立計算機率模型

a. 人名(PN)：

將語料內不在辨認辭典中的人名，依據其於文章中的出現次數列出，將這些人名「以 word 的形式」來計算某人名出現在此類別之機率，倘若辨認語料(TCC300 測試語料)中人名出現此些人名之外，則藉由 smoothing 方式來給予一個較低之分數，使其在辨認時亦可能辨認正確，在此 smoothing 方式採 Good-Turing smoothing。

b. 詞綴(MD)：

將語料內不在辨認辭典中的各類詞綴，依據其於文章中的出現次數列出，將這些詞綴「以 word 的形式」來計算某人名出現在該類別之機率，倘若辨認語料(TCC300

測試語料)中詞綴出現此些詞綴之外，則藉由 smoothing 方式來給予一個較低之分數，使其在辨認時亦可能辨認正確，在此 smoothing 方式採 Good-Turing smoothing。

c. 數量複合詞(FT)：

(1) 半詞二連文機率模型：將語料內不在辨認辭典中的數量複合詞，依據其於文章中的出現次數列出，將此些數量複合詞「以半詞(subword)的形式」來統計這些半詞互相連接的次數，進而計算數量複合詞半詞之間串接機率。

(2) Finite State Machine(FSM)機率模型：將語料內不在辨認辭典中的數量複合詞，依據其於文章中的出現次數列出，將此些數量複合詞「以狀態(state)的形式」呈現，統計訓練語料中這些狀態的轉移次數，再利用此些次數來計算在 FSM 架構內狀態間轉移的機率，此部分詳細計算方式為章節 4.1.2.3 所述。

此時各類內部機率所使用機率形態如下表 4-3 所示：

表 4-3 各類內部機率使用機率型態

	人名	**詞綴**	**數量複合詞**
機率型態	word uni-gram	word uni-gram	1. subword bi-gram 2. FSM transition probability

經由 **A** 和 **B** 兩方面分別可求得「詞和類別」之外部連接機率和類別內某詞出現之內部機率，藉由此內外部機率的關係來描述 extended word lattice 上新產生之節點與前一個節點之間連接機率，以此來決定新產生弧上之分數。

第五章 實驗結果與分析

5.1 辨識語料分析

本研究使用 TCC300 語料進行辨認，在辭典的收錄部分，將 15,000 詞條空間用以收錄人名、詞綴和數量複合詞等三類詞之 subword 半詞集合，因而犧牲原先可收錄常見一般長詞之辭典空間，故測試語料 TCC300 中三類長詞的多寡，將會影響使用 word/subword 混合式語言模型效果的好壞，進而影響其辨識結果。本研究針對 TCC300 語料中人名、詞綴、數量複合詞和一般詞加以統計，藉此了解這幾類長詞在辨認語料中所佔的比例，如下表(表 5-1)所示。

表 5-1 TCC300 辨認語料各類分佈

TCC300 測試語料分析				
TCC300 辨認語料	**總詞條數**	**各類所佔比例**	**總數量**	**各類所佔比例**
	5226		14098	
文章細分				
人名	107	2.05 %	219	1.55 %
詞綴	226	4.32 %	435	3.09 %
數量複合詞	240	5.89 %	409	2.90 %
一般詞	4653	87.74 %	13035	92.46 %

5.2 Perplexity 複雜度比較

Perplexity（PPL）用以測試語言模型之複雜程度，PPL 值高代表語言模型複雜度高，語言模型存在大量詞和詞連接的 n-gram 機率，透過這些機率能以當前的詞來預估下一個詞（以 bi-gram 為例），換言之，PPL 越高意味著透過目前的詞來預估下一個詞的命中機會相對較低，需要較多預估次數才會命中（預估命中次數最多不超過 PPL 本身的值），另一方面，PPL 值低，則其語言模型顯得較為單純，不需要過多的預測次數即可命中，故於 PPL 值越低時進行語音辨認，語言模型可能會呈現較好的辨識效能。

傳統式語言模型與 word/subword 混合式語言模型 Perplexity 值如下表 5-2 所示：

表 5-2 Perplexity 比較

分類	Perplexity
傳統方式語言模型	787.8169
Word/subword 混合式語言模型	744.0824

訓練語言模型之文章前處理時，將文章內容以「驚嘆號、問號、分號及句號」四種較具口語停頓的符號來進行分段，然而「逗號」的地方，在口語上比較連貫，前一句的最後一個詞通常和目前句子的第一個詞聽覺上沒有明顯的停頓，故該符號並不拿來分段，訓練語言模型時採取跨過逗號，使進行前文章處理時，內容可以被分段成一句句較長且合乎一般口語停頓的句子，也因如此 bi-gram 條目增多，會使得 PPL 值較為偏高。

5.3 Word lattice 上可涵蓋的三類長詞之比較

本研究於 word/subword 語言模型之建立上，辨認辭典新增 subword 半詞之主要目的為期望於第一級 word/subword 語言模型產生之 word lattice 上，能正確產生人名、詞綴和數量複合詞之 subword 半詞串，進而透過第二級構詞，將此些 subword 半詞串構回原本有意義的長詞，形成第二級新的 word lattice（這裡稱 extended word lattice），重新統計此些新產生出來長詞之機率，在 extended word lattice 上給予語言模型分數和聲學模型分數，最後產生第二級語言模型，基此，本研究將於下針對第一級語言模型產生的 word lattice 進行分析，觀察 suword 半詞串併在一起可以構回原來三類長詞之數量及其涵蓋率（cover rate）。

（1）傳統語言模型之涵蓋率

表 5-3 傳統語言模型 word lattice 上涵蓋率

詞彙分類	Test data	Coverword number	Cover rate
人名	219	59	26.94 %
詞綴	435	343	78.85 %
數量複合詞	409	331	80.92 %

在傳統語言模型中，被「cover（涵蓋）」於一個長詞中，意味 lattice 上的詞要整體符合長詞，由於傳統語言模型沒有採用規則將詞拆解成 subword，故無法以 two-stage 方法於第二級將短詞構回有意義長詞。另外，在上表(表 5-3)中可發現人名涵蓋率很低，因人名數量眾多、變化很大，而在辨認辭典中，本研究僅收錄語料出現頻率較高的人名，所以在 TCC300 測試語料中，出現辨認辭典中未收錄的人名，將會發生辨認錯誤，無法涵蓋到原本的人名。然而，詞綴和數量複合詞在語料中常出現的詞，即為收錄在辨認辭典中的詞，與測試語料中的此兩類詞有較高相同度，不像人名變化型態之多，故涵蓋率也較人名提升不少。

（2）Word/subword 混合式語言模型之涵蓋率

表 5-4　Word/subword 混合式語言模型 word lattice 三類詞涵蓋率

詞彙分類	Test data	Coverword number	Cover rate
人名	219	123	56.16 %
詞綴	435	401	92.18 %
數量複合詞	409	376	91.93 %

在 word/subword 混合式語言模型中，從上表(表 5-4)可發現倘若將人名、詞綴和數量複合詞有規則地切割為較小單位的 subword 半詞串，能使 word lattice 上此三類的涵蓋率較傳統式語言模型均有所提高，尤以人名提升的幅度較為明顯，故針對詞變化情況較大的人名，假使以較短的 subword 來代替，將使得迫於辨認辭典大小限制因而收錄不齊全的人名導致涵蓋率降低的影響性變小，涵蓋率均比傳統式來得高，亦可發

現 word/subword 混合式語言模型更適用於進行第二級構詞。

表 5-5 各類第二級 subword 構詞組成數量

	一個 subword	二個 subword	三個 subword	三個 subword 以上
人名	50	24	49	0
詞綴	338	63	0	0
數量複合詞	299	67	9	1

在人名部分，可拆解成「姓氏+名字」和「姓氏+名字_第一字+名字_第二字」兩種，由表(表 5-5)中可發現，在 word lattice 可以構成人名的部分，以沒拆解或全拆解的人名最多，意指 TCC300 辨認語料人名中的名字部份，會出現在收錄的 3068 個名字 subword 數量並不多，但供選擇的名字 subword 太多，目前僅能先將高詞頻的名字半詞優先選入辨認辭典，往後如果可有技巧地針對此方面半詞來收錄，使得所收錄的三類半詞集合可在第一級 word lattice 上涵蓋到更多詞，再經過構詞模組後，可產生更多對辨認結果有幫助的三類長詞，使得第二級語言模型辨認效果更佳。

- **兩個模型構詞涵蓋數量之比較**

表 5-6 三類在兩個語言模型構詞數量

分類	傳統式語言模型	Word/subword 混合式語言模型	兩種 lattice 上構回長詞數量差額
人名	59	123	**64**
詞綴	343	401	**58**
數量複合詞	331	376	**45**

由表 5-6 可知，人名有規則地切割成半詞後，於 word lattice 上可正確地涵蓋到 TCC300 測試語料原有的該類長詞數量增多，其增加之數量較詞綴、數量複合詞兩類

更為明顯，由於人名拆解後之名字 subword 較不具一般常見詞特性，大多僅出現於人名之中，因此，倘若能以規則性方式將人名拆解為 subword，將常見之名字 subword 收錄至辨認辭典內來進行辨認，進而由第二級構詞構回完整人名，該方法之辨識效能將會優於僅在辭典內收錄常見人名之方法。

另一方面，詞綴、數量複合詞兩類本身較具有一般詞之特性，故於語料之出現頻率較高，尤以詞綴最為明顯，其大多數的詞出現於次序 45,000 之內，以長詞形式存在，故許多該類詞均不須拆解成 subword 半詞串，也因如此，由 subword 半詞串構詞回有意義的長詞之數量較少。

另外，針對語料庫未出現的人名利用名字半詞進行涵蓋，如下表(表 5-7)所示：

表 5-7 名字 subword 對語料中未出現人名的涵蓋情況

		詞條數量
實驗室收集的人名		414,241
名字半詞	次序 45,000 內	14
	次序 45,000 外	3068
語料庫未出現的人名	總數量	308,696
	可涵蓋數量	22,266

在人名方面，人名將會被拆解為姓氏和名字兩種 subword 半詞，由於名字 subword 數量、詞條眾多，無法完全收錄於辭典內，故目前僅將名字 subword 中出現頻率較高者優先收錄之，共約 3068 個。而由上表(表 5-7)可知，本研究之研究實驗室所收集中文人名詞條數量共有 414,241 條，然而，並非所有收集的中文人名詞條均會出現於語料文章中，故進一步來觀察將這些未出現者(詞條數為 308,696 條)透過辨認辭典所收集的名字半詞可涵蓋詞條數，可得涵蓋詞條數量為 22,266 條，約佔語料庫未出現人名之總數量 7.21 %，使用較長的名字半詞所涵蓋數量有限。

5.4 Word lattice 構詞結果分析

第二級 extended word lattice 上所產生的新節點，是經過構詞所產生出來的長詞，在構詞時，針對符合 subword 半詞串可構出長詞之標準者，皆予以構詞，藉此滿足 word

lattice 上所有可能路徑需要的節點，如下表(表 5-8)所示。

表 5-8 構詞結果分析

	人名	詞綴	數量複合詞	總計
Word lattice 總構詞數	**18246**	**5241**	**357482**	**380969**
TCC300 測試語料	**219**	**435**	**409**	**1063**

於上表(表5-8)可得知，第二級構詞將會產生許多數量複合詞此類長詞，然而，在辨認上之最佳路徑，最佳路徑大多不會通過此類節點，在TCC300測試語中實際數量複合詞比例僅約有0.001144，相較於其他類別來說，其使用比例相當低，構詞後所付出代價較高，因過分構詞而導致辨認錯誤的風險也相對提高。

因此在 word lattice 上若把可以構詞的半詞串全數構出長詞，將會發生「過份構詞」的現象，原先第一級混合式語言模型辨認正確的詞，再經過第二級構詞模組後，產生的新路徑分數會因為 word penality 的加入或計算語言模型分數的處理不足，使得原本正確的詞經過第二級分數配置後，構詞新產生的路徑分數超越正確的路徑，過分構詞的詞反而在第二級語言模型進行辨認時被凸顯出來，反而造成辨認錯誤。

基此，往後或許於第二級構詞時，可透過建立有句法結構限制的構詞模型方式，藉此更精確地來偵測需要構詞之人名、詞綴、數量複合詞和 OOV，再來進行構詞，即使用更多資訊以將需要構詞之半詞串偵測出來，減少太多不必要之構詞，避免產生過多構詞，希望降低過分構詞導致錯誤辨認結果之影響性。

5.5 理想上最佳辨認效能

辨識結果指的是 word lattice 上分數最高的一條路徑，然而，辨識出的詞並非全為正確，其中亦可能出現因辭典中並未收錄導致辨認錯誤之現象，或者發生辭典有收錄，word lattice 上亦有出現，卻因在 word lattice 上須和其他相近音詞相互競爭而造成分數落後而導致辨認錯誤，由此可知，語料庫文章前處理之方式，將會對於連接詞的路徑

分數和辨認效能造成相當大的影響。基此，研究者於本小節針對最大辨識效能進行討論，換言之，倘若 word lattice 上之正確長詞均能同時在最佳路徑上出現，並能相互串接，此路徑將為 word lattice 上之最理想路徑，理想上最佳的詞辨識效能，藉此呈現當前語音辨認結果之改善空間，往後進而探討如何提升 word lattice 上正確詞之路徑分數。

A ： 傳統語言模型

B ： Word/subword 混合式語言模型，將人名、詞綴和數量複合詞構回長詞

C ： Word/subword 混合式語言模型，將人名、詞綴、數量複合詞及 OOV 構回長詞

表 5-9　Word lattice 上的最佳路徑比較

分類	Deletion	Substitution	Insertion	Accuracy	Total count
A	70	1796	1378	**76.99 %**	14098
B	54	1612	1143	**80.08 %**	14098
C	56	1213	607	**86.69 %**	14098

由上表(表 5-9)可知，在 A 和 B 兩者相互比較後可發現，僅針對人名、詞綴和數量複合詞三類在第二級進行構詞後，將三類 subword 半詞串構回有意義長詞，此與辭典不收錄 subword 半詞來進行構詞的傳統語言模型相比較，最佳辨識率上最多可提升 3.07 %的詞辨識效能；更進一步觀察，如果能把為了減少 OOV 而將此些不在辭典內即拆成短詞串的詞也一併加入第二級構詞後，如同 A 和 C 的比較結果，詞之辨識效能將可擴大範圍至 9.7 %，強化了構詞模組之效果，可使詞辨識效能更佳。

5.6 辨識效能比較

以下針對傳統式語言模型、第一級混合式語言模型和階層式第二級語言模型之字元及詞辨識效能進行比較，如下表(表 5-10、表 5-11)所示：

其中，

三類構詞(一)：數量複合詞內部機率(intra probability)使用 subword bigram 機率模型

三類構詞(二)：數量複合詞內部機率(intra probability)使用 FSM 的狀態轉移機率模型

表 5-10 字元(character)辨認效能比較

		Deletion	Substitution	Insertion	Accuracy	Total count
傳統語言模型		365	6121	101	**75.16 %**	26472
混合式語言模型	**未構詞**	353	6106	94	**75.25 %**	26472
	構詞	350	6092	95	**75.31 %**	26472
第二級語言模型	**未構詞**	342	6365	124	**74.20 %**	26472
	三類構詞(一)	317	6310	115	**74.53 %**	26472
	三類構詞(二)	326	6335	118	**74.39 %**	26472

表 5-11 詞(word)辨認效能比較

		Deletion	Substitution	Insertion	Accuracy	Total count
傳統語言模型		465	4105	948	**60.86 %**	14098
混合式語言模型	**未構詞**	436	4156	1050	**59.98 %**	14098
	構詞	500	3898	598	**64.21 %**	14098
第二級語言模型	**未構詞**	381	4272	1012	**59.82 %**	14098
	三類構詞(一)	704	4097	436	**62.85 %**	14098
	三類構詞(二)	692	4211	439	**62.10 %**	14098

由表 5-10 與表 5-11 可知，字元與詞的辨識效能在第二級語言模型在僅受 word penality 時均有能所提升，而在改變語言模型分數後，字元辨識上有些微的下降，但在詞辨認上有亦著明顯的改善，下一小節將對於辨認結果加以細部分析，觀察各類局部現象並加以說明。

5.6.1 辨識結果之細部剖析

在辨認結果中針對人名、詞綴及數量複合詞進行分析，並以 TCC300 測試語料作為比對的依據，觀察在傳統式做法和本研究所提出之方法對於此三類詞在辨認上的影響並進行說明。

首先，針對 TCC300 測試語料、傳統式語言模型辨認結果及第二級語言模型(階層式語言模型)辨認結果中的人名、詞綴和數量複合詞予以統計，來得知此三類詞在不同語言模型辨認出現的情況，分布情況如下表 5-12 所示。

表 5-12 辨識結果中各類所佔總數量

	人名總數量	**詞綴總數量**	**數量複合詞總數量**
TCC300 測試語料	219	435	409
傳統式語言模型	57	378	419
第二級語言模型	274	503	563
各類增加數量[4]	**217**	**125**	**144**

在上表（表 5-12）中，可發現傳統式語言模型辨識出人名和詞綴的數量較低，特別是在人名此部分上，遠低於 TCC300 測試語料中存在之人名數量，這是由於傳統式語言模型的辨認辭典有大小限制(約可收錄六萬詞條)，辭典中僅能收錄到高詞頻的人名，如果測試語料出現該辭典未收錄之人名，則無法正確辨識出此完整人名，因而凸顯出傳統式語言模型因辭典大小受限而影響辨識效能的缺點，有鑑於此，故在本研究提出階層式語言模型來克服此問題。

藉由階層式架構所得之第二級語言模型中可發現，第二級語言模型辨認出三類詞的數量明顯比傳統式語言模型增加許多，特別在人名部分增加幅度更為顯著，不過由

[4] 增加數量意指第二級語言模型數量減去傳統式語言模型數量而言

第二級語言模型辨認出三類詞的數量來觀察，亦可發現數量均超過 TCC300 測試語料原本存在的各類數量，這說明了第二級語言模型辨認出過多此三類詞，可能會造成原本傳統式可辨認正確的一般詞被誤認為此三類詞，對於此現象來加以分析，將三類詞細部情況透過 **5.6.1.1** 至 **5.6.1.3** 三小節來呈述，過程中使用以下三種代號{A,B,C}來表示不同情況。

A 代表 「應為{人名,詞綴,數量複合詞}但結果辨認錯誤之部分」

B 代表 「不為{人名,詞綴,數量複合詞}卻辨認成該類之部分」

C 代表 「該類辨認正確之部分」

其中 ，

「各類總數量[5] －B 類數量 －C 類數量」代表 A 類中有辨識出為該類別{人名, 詞綴, 數量複合詞}之詞，而此詞並非為正確辨識結果的數量。

5.6.1.1 人名

表 5-13 人名辨認情況

	A	B	C
傳統式語言模型	175	8	44
第二級語言模型	129	134	90

由上表（表 5-13）可知，採用階層式第二級語言模型可使得辨識正確的人名比傳統式語言模型成長許多，由原先傳統方式僅能使 44 個人名辨識正確改善至達到 90 個人名辨識正確，增加幅度為傳統式數量兩倍之多，但尚有許多改善空間，因為 TCC300 測試語料中有 219 個人名尚有超過一半辨識錯誤(此處辨識錯誤為 129 個)，此外，在建立階層式語言模型時，會在 word lattice 使用 subword 半詞進行構詞，讓 word lattice 上產生許多可為人名的節點，再透過對路徑分數的重新配置，讓這些人名可被正確地辨識出來，但往往為了將該類別的詞凸顯出來而分數給予過高，使得不該辨識為人名的地方反而出現人名，而相對地造成辨識上的錯誤，造成此類錯誤數量為 134 個。

[5] 總數量；指表 6-12 之各類總數量

為了增加人名的辨認效能(C 類)或減少辨認上的錯誤(A 類或 B 類)，或許可以試著在分數配置時加入一些限制，如：稱謂、尊稱等相關詞來更強化人名附近的分數配置，使得該出現人名地方能被加強，使得人名辨識程度能再往上提升。

5.6.1.2 詞綴

表 5-14 詞綴辨認情況

	A	B	C
傳統式語言模型	130	68	305
第二級語言模型	88	145	347

由上表（表 5-14）可知，詞綴在第二級語言模型中辨識效果亦有明顯的改善，可比傳統方式增加 42 個詞綴被正確地辨識出來，另有 145 個 B 類此種錯誤存在。此外，可發現詞綴在辨識上錯誤機會相對較小，TCC300 測試語料中有 435 個詞綴，而利用傳統式語言模型即可得到 305 個詞綴辨識正確，相較於人名或數量複合詞在傳統式辨識時正確比例來得高，此現象顯示出在傳統式辭典內已經收錄許多高詞頻的詞綴，使得辨認時詞綴這部份表現不錯，然而再透過本研究提出之階層式做法，發現可使此部分辨識效果更為提高，能正確辨識出之詞綴數量達到 347 個，該數量在 TCC300 測試語料詞綴中佔該類別的 80%(傳統方式為 70%)，因而辨識詞綴上獲得近一步改善。

5.6.1.3 數量複合詞

表 5-15 數量複合詞辨認情況

	A	B	C
傳統式語言模型	154	155	255
第二級語言模型	138	248	271

由上表（表 5-15）可知，階層式做法在數量複合詞正確辨認(C 類)比傳統方式有些微增加，但卻存在數量不少(數量為 248 個)的 B 類錯誤，比其他類別多出許多，表示辨識過程中有許多原本非數量複合詞的地方均被誤認為數量複合詞，而造成此結果的原因可能在於「分數配置」上，由 5.4 小節表 5-8 可知數量複合詞經過構詞後總構詞數量龐大(數量為 357482 個)，而本研究在構詞後的 extended word lattice 分數配置是採內外部機率(inter probability and intra probability)來給予，其中內部(intra)機率為使用該類別之 subword bigram 來計算，而從實驗結果來看，長度較長的數量複合詞由於拆解後 subword 半詞數量較多，故在分數配置上可以有效抑制，使其在 extended word lattice 上分數不至於過高，辨認上可得到不錯的效果，但相對於長度較短的數量複合詞較差，乃由於數量複合詞數量多且均處於同一類別(FT)，故一般詞彙與該類別(FT)串接的外部機率比較高，但此類數量複合詞串接的半詞數量少，因而較不易在內部機率中透過半詞機率串接的影響而讓分數被壓抑下來，使得該類數量複合詞會在分數配置上會比較偏高，導致在辨認時許多地方比較容易辨認成短數量複合詞，此為使用 subword bigram 做為數量複合詞內部機率(intra-probability)的主要缺點，其中，錯誤型態如下表(表 5-16)所示。

表 5-16 數量複合詞中各類錯誤型態

錯誤型態	**辨認結果**	**參考答案**
連續數字串	二一一	惡意 隱瞞
	五一一	無 疑義
	八二一八	拔 了 一把
{數詞}+{量詞}	四秒	寺廟
	十七架	時期 家家
{綴詞}+{數詞}+{量詞}	下七發	小西瓜
	近九十	困境 就是

然而數量複合詞之內部機率採 FSM 轉移機率時，則發現短長度的數量複合詞變成更容易被辨認出來，原因在於，分數配置上除了受過大的外部機率影響外，長度較短之數量複合詞受半詞數量少而不受壓抑的程度亦比使用 subword bigram 作為內部機率時更為嚴重，因為在 FSM 架構下，每個狀態為存在許多同質性的半詞集合，在訓練轉移機率時，會將該同質性的的半詞均轉為同一狀態，此動作使得狀態間的轉移機率會比 subword bigram 機率來得更大，使得長度較短的數量複合詞更容易受到過分構詞的影響，使得更多原本「非數量複合詞」的地方，被錯誤辨認成短數量複合詞，因而影響辨識效能。

由上述可知，本研究運用階層式 two-stage【12】【13】【14】之架構，採用 subword 短詞收錄至辭典之辨識方法，相較於僅收錄長詞到辨認辭典之傳統方式，本研究之方法辨認效能突破辨認辭典大小之限制，增加了人名、詞綴及數量複合詞可正確被辨認之數量，並使得整體辨識效能提升，辨識率由傳統式語言模型的 **60.86%**升至階層式語言模型的 **62.85 %**，而將數量複合詞使用 FSM 架構亦可達 **62.10 %**，均可使階層式語言模型得到進一步改善。

階層式架構所訓練之語言模型，可對於經過構詞模組後產生的 extended word lattice 進行分數配置，往後在此架構下可引入各種模型(如：韻律模型)、語法資訊等來加強 extended word lattice 上分數配置的可靠性，使得第二級語言模型具有更多資訊而得到強化，使辨識效能可加以進步。

.

第六章 結論與未來展望

6.1 結論

本研究使用 TCC300 語料庫進行語音辨識之相關研究，自基本系統之建立、文章內容之刪除、內容錯誤之更正處理此些層面著手，並針對人名、詞綴和數量複合詞此三類較常見、數量較多而無法收入至辭典中的長詞進行拆解，予以 subword 短詞處理。

綜而言之，本研究結果可分為四大要點，分述如下：

（1） 文章內容、文章斷詞結果的好壞，在訓練語言模型時，將會影響到估算詞和詞之間機率的正確性、可靠度，因而間接地影響到語言模型，對於辨認結果有所影響。

（2） 就辨認結果之目標而言，辨認標的為較有意義的長詞，然而，辭典又無法收錄全部的詞，故將有意義的長詞，如：人名、詞綴及數量複合詞，以 subword 半詞串取代之。

（3） 就 Subword 集合收錄至辭典的選擇來說，拆解後的 subword 集合其詞條數量眾多，無法均收錄至辭典內，需設立合理之收錄範圍，故對於短詞詞頻高者優先考量收錄之。

（4） 在 word lattice 上產生可構回長詞的 subword，經由人名、詞綴和數量複合詞的等三類和 OOV 進行構詞、word penality 微調及重新計算分數後，再進行辨認，可有意義長詞之辨識效能均獲得提升。

6.2 未來展望

語音辨認結果之最終目的乃為辨識出較長且具有意義的詞，本研究針對辨認結果和較長且有意義的詞做分析後，發現若將有意義的長詞，以較長的 subword 半詞串取代後，subword 集合詞條數量過於龐大，以致無法全收錄於辭典內，是故，subword 短詞之收錄方法就顯得極為重要。基此，建議未來研究可繼續運用 two-stage 概念，針對 word lattice 上的 subword 半詞串，將其構詞回到原先較有意義的長詞，並分別透過

建立有意義的各類長詞之構詞模型進行構詞，例如：人名的模型、詞綴模型和 OOV 等的構詞模型，不再以目前查表方式來當作構詞的依據，進而推動第二級之辨認，使辨認語料之更多 subword 半詞串得以被偵測出來並構回有意義的長詞；另一方面，在人名和數量複合詞內部機率之分數配置，可採取階層式的架構，人名可針對不同特性的人名，建立與前後詞的關連性，而數量複合詞則可在 FSM 架構下對數詞部份或量詞等地方加以細緻分類，依據相似結構，可針對分開建立的數詞 FSM 共享相同的機率。

另外，混合式辭典中仍有 4,529 條的集合空間尚未完整運用(原本收錄未出現之一字詞)，往後可運用此空間來收錄更多對於語音辨識效能有所助益的 subword 短詞。

參考文獻

【1】B.H.Juang and S.Furui,“Automatic recognition and understanding of spoken language—A first step towards natural human-machine communication,”in Proc IEEE,88,8,pp.1142-1165,2000

【2】L.R.Rabiner and B.H.Juang,“Fundamental of speech Recognition,”New Jersey,Prentice-Hall,Inc.,1993

【3】S.Young, G..Evermann, T.Hain, D.Kershaw, G,Moore, J,Odell,D.Ollan, D.Povey, V.Valtchev, P.Wooland,“The HTK Book(for HTK version 3.4)”

【4】Slava M. Katz,“Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer,”IEEE Transactions on Acoustic,Speech and Signal Processing,Vol.ASSP-35,NO.3,MARCH 1987

【5】江振宇(2004)。中文斷詞器之改進。國立交通大學電信工程學系碩士論文。

【6】張隆勳(2005)。國語廣播新聞語音基本辨認系統之建立。國立交通大學電信工程學系碩士論文。

【7】P.Geutner,“Using Morphology Towards Better Large-Vocabulary Speech Recognition Systems” in: Proc .Int. Conf. on Acoustics, Speech, and Signal Processing, Detroit, pp. 445-448 ,1995

【8】Mathias Creutz,Teemu Hirsimaki,Mikko Kurimo,Antti Puurula,Janne Pylkkonen,Vesa Siivola,Matti Varjokallio,Ebru Arisoy,Murat Saraclar,and Andreas Stolcke, ”Analysis of Morph-Based Speech Recognition and the Modeling of Out-of-Vocabulary Words Across Languages”,Helsinki University of Technology,2007

【9】Issam Bazzi and James R. Glass, “Modeling Out-of-Vocabulary words for robust

speech recognition", Laboratory for Computer Science Massachusetts Institute of Technology Cambridge, 2000

【10】Issam Bazzi and James R. Glass, "A Multi-class Approach for Modeling Out-of-Vocabulary words", MIT Laboratory for Computer Science Cambridge, 2002

【11】Ali Yazgan and Murat Saraclar, "Hybrid Language Models for Out-of-Vocabulary word Detection in Large Vocabulary Conversational Speech Recognition", Center for Language and Speech Processing, 2008

【12】Koichi Tanigaki, Hirofumi Yamamoto, and Yoshinori Sagisaka, "A Hierarchical language model incorporating class-dependent word models for OOV words recognition", ICLSP, 2008

【13】Shigehiko Onishi, Hirofumi Yamamoto, "Structured language model for class identification of out-of-vocabulary words arising from multiple word-classes", European Conference on Speech Communication and Technology, 2001

【14】Koichi Tanigaki, Hirofumi Yamamoto, Yoshihiki Ogawa, and Yoshinori Sagisaka, "Out-of-vocabulary word recognition with a hierarchical doubly markov language model", IEEE, 2005

附錄一

0 1 2 3 4 5 6 7 8 9 一 七 九 二 八 十 三 千 五 六 卅 廿 仟 四 兆 百 佰 兩 幾 萬 零 億
几 仟 伍 兆 玖 佰 拾 柒 捌 參 陸 壹 幾 貳 萬 肆 零 億
乙 丁 丙 戊 甲
大 小 整
半 正 多 足 許 餘 整 之多 出頭 好幾 開外
半 正 多 許 整
全 成 滿 整 一切 所有
多少 若干 幾多
太 再 多 好 更 怪 很 挺 真 夠 頂 最 極 滿 忒 十分 不大 尤其 比較 有點 那麼 非常 略為 異常 這麼 稍微 過分 過份
多 一些 不少 少許 少數 多數 好些 有些 有的 個把 泰半 許多 部分 部份 幾許 許許多多
半 有的 若干
那 哪 這 這些 那些 哪些
下 上 另 末 同 次 前 後 某 首 頭
本 何 別 旁 敝 貴 諸 啥 什麼
上下 不到 不等 以下 以上 左右
少 多
多 來 餘
兆 萬 億
點
又
分之
弱 強
半
雙
滿滿 整整
好幾
數
年
班
他
市 地 州 弄 村 里 巷 段 洲 省 站 郡 區 國 鄉 號 樓 鄰 縣 鎮 街
該

第
每
各
逐
近 另外 將近
此
其他 其它 其餘
任何
成
不到
一
那 這
平方 立方
個
分
秒
時 點 點鐘
點
小時
刻
元
年
月
日 號
號
月份
下 上 元 本 正 每
元 正
度
段
巷
弄
之
號
樓
華氏 攝氏

華氏
零下
午 晚 晨 下午 上午 子時 巳時 丑時 中午 午夜 午時 午間 半夜 卯時 未時 申時 亥時 戌時 早上 早晨 辰時 酉時 凌晨 寅時 晚上 晚間 晨間 清晨 深夜 傍晚 曷午
元始 元狩 元朔 元嘉 公元 正大 正朔 正統 正隆 正德 民國 永嘉 永樂 永曆 西元 明治 咸豐 宣統 建安 昭和 洪武 貞觀 開元 開皇 開運 道光 雍正 嘉靖 嘉慶 德光 德裕 寶元 中華民國
冬 春 秋 夏 大月 冬天 冬季 仲冬 仲春 仲秋 仲夏 早春 孟冬 孟春 孟秋 孟夏 炎夏 炎暑 初冬 初春 初秋 初夏 春天 春季 春秋 秋天 秋日 秋季 夏天 夏令 夏季 烈暑 深秋 盛夏 盛暑 寒冬 陽春 隆冬 隆暑 新春 暮春 暮秋 窮冬 嚴冬 徂暑
週一 週二 週三 週五 週六 週日 週四
三伏 中葉 今世 今生 公餘 凶年 凶歲 半晌 末代 年下 年假 年終 年關 早期 老年 邪世 例假 來世 來生 來年 旺季 花甲 花季 花期 初期 雨季 前夕 前期 後期 後葉 春假 風季 展期 朔望 衰世 乾季 假日 婚期 晚世 淡季 球季 寒假 寒暑 暑假 暑期 開春 亂世 新年 會期 歲末 歲首 歲暮 當世 當年 經期 農時 農閒 農隙 漁汛 漁期 暮歲 熱季 課餘 餘暇 檔期 濕季 糧季 髫年 齠年 一年半載 太平盛世 梅雨季節 黃金時代 過渡時期
今 七七 七夕 下元 下旬 上元 上旬 大雪 大寒 大暑 小雪 小寒 小暑 小滿 中元 中旬 中秋 今天 今日 元日 元旦 元宵 六甲 冬至 冬節 白露 立冬 立春 立秋 立夏 立雪 年節 旬日 次日 至日 佛誕 尾牙 良日 芒種 炎天 初一 初旬 長日 雨天 後日 春分 春日 春節 昨天 昨日 秋分 耶誕 重九 重陽 夏日 夏至 朔日 校慶 除夕 鬼節 國慶 望日 清明 陰天 陰壽 單日 寒食 晴天 週一 週二 週三 週五 週六 週日 週四 開年 暇日 當天 當日 聖誕 端午 端陽 暮節 熱天 穀雨 燈節 臘八 臘日 驚蟄 九九重陽 大年初一 良辰吉日 炎炎夏日 國定假日 黃道吉日 雙十國慶 九三軍人節 五一勞動節 行憲紀念日 開國紀念日 黑色星期五 臺灣光復節
刀 丸 勺 口 介 分 匹 升 天 戶 手 支 方 日 片 世 付 仗 代 仞 冊 包 司 只 台 句 市 本 疋 目 伙 伍

件 任 份 列 名 回 地 夸 州 年 式 曲 朵 次 色 行 串 位 作 匣 局 尾 床 弄 把 批 折 束 村 杓 步 汪 系 身 具 卷 味 宗 屆 帖 房 所 抹 招 拐 拍 抱 服 枝 杯 板 枚 波 泡 版 直 股 門 則 室 客 封 巷 度 指 挑 架 柄 段 洲 流 派 盆 盅 省 科 缸 胎 軍 重 面 頁 首 乘 倍 個 剔 員 套 家 峰 席 師 座 扇 拳 捆 挺 旅 根 桌 格 株 班 砲 站 級 記 起 郡 針 院 陣 隻 副 匙 區 圈 國 堆 堂 堵 宿 帶 張 捲 捧 掛 排 捺 桿 桶 條 毫 球 瓶 畦 盒 眼 票 粒 紮 組 處 袋 通 連 部 章 頂 圍 場 壺 尊 幅 幀 掌 期 棟 棵 棒 棍 款 番 發 程 等 筆 筐 筒 絲 街 軸 鄉 開 間 隊 階 集 項 塊 會 歲 盞 碗 節 綑 群 腳 落 葉 號 路 跤 道 遍 鈎 頓 劃 團 夥 對 幕 截 撇 摺 槍 槌 滴 種 窩 端 管 箇 臺 層 幢 撮 樣 樁 樓 盤 箭 箱 篇 編 蓬 課 豎 趟 輛 輩 遭 鄰 駝 劑 擔 橫 瓢 縣 艘 錠 頭 餐 幫 檔 營 簇 簍 縷 聲 聯 鍋 鍬 闋 顆 點 叢 罈 鎮 雙 鞭 題 龍 櫥 瓣 邊 關 類 籃 覺 齣 欄 響 彎 攤 灘 疊 籠 聽 襲 罐 廳 抔 坨 厘 灶 衖 綹 下子 下兒 勺子 口兒 小節 分兒 巴掌 手兒 方面 包兒 回合 池子 匣子 杓子 系列 杯子 板子 長列 長串 長段 長排 長條 長節 長截 段兒 盆子 面兒 拳頭 根兒 格兒 茶匙 陣子 圈兒 桶子 瓶子 盒子 袋兒 棍子 湯匙 筒子 絲兒 象限 開金 會兒 節兒 鈎兒 劃兒 嘟嚕 截兒 槍矛 槌子 樣兒 盤子 箱子 輩子 學期 擔子 頭兒 檔子 簍子 聲兒 鍋子 點兒 櫃子 罈子 鞭子 櫥子 邊兒 爐子 籃子 籠子 罐子 籮筐
手 地 池 身 腔 腳 嘴 頭 臉 肚子 屋子 家子 桌子 院子 鼻子
丈 寸 尺 吋 米 呎 里 哩 浬 碼 釐 度 噚 公丈 公寸 公分 公尺 公引 公里 公釐 公厘 台尺 市尺 光年 米尺 米突 英寸 英尺 英吋 英呎 英里 英哩 海里 海哩 海浬 毫米 微米 厘米
甲 坪 畝 頃 公畝 公頃 市畝 英畝
斤 克 兩 磅 噸 錢 公斤 公克 公兩 公噸 公擔 公衡 公錢 日斤 台斤 台兩 市斤 克拉 英兩 英磅 盎司 盎斯 毫分 毫克
升 斗 石 夸 斛 公勺 公升 公斗 公石 公合 公秉 公毫 公撮 日升 加侖 仟克 台升 市升 夸特 夸爾 西西 品脫 毫升
分 天 日 年 旬 更 周 夜 季 秒 紀 宿 週 歲 載 輪 鐘 小時 分鐘 年份 刻鐘 周年 周歲 星期 秒鐘 週年 微秒 禮拜 釐秒

刀 元 文 毛 令 卡 打 瓦 角 圓 塊 綸 赫 鎊 籮 千卡 千瓦 千赫 大籮 分貝 文錢 日元 日圓 牛頓 仟卡 仟瓦 仟赫 台幣 瓦特 伏特 兆赫 先令 安培 位元 里拉 周波 居里 法郎 法朗 便士 美元 美金 馬力 馬克 毫巴 莫耳 港幣 焦耳 塊錢 達因 爾格 赫茲 歐姆 盧比 盧布 辦士 燭光
・ • .
%
,
／
： ： ：
ＦＡＸ： ｆａｘ：
ＴＥＬ： ｔｅｌ：
ＡＭ ａｍ
ＰＭ ｐｍ
$
中 初 底
年級
學年 年度 學年度
年代
世紀
仁 平 孝 良 和 忠 信 勇 恭 智 愛 溫 義 簡 讓

附錄二

一 二 三 四 五 六 七 八 九 十 十一 十二 十三 十四 十五 十六 十七 十八 十九 二十 二十一 二十二 二十三 二十四 二十五 二十六 二十七 二十八 二十九 三十 三十一 三十二 三十三 三十四 三十五 三十六 三十七 三十八 三十九 四十 四十一 四十二 四十三 四十四 四十五 四十六 四十七 四十八 四十九 五十 五十一 五十二 五十三 五十四 五十五 五十六 五十七 五十八 五十九 六十 六十一 六十二 六十三 六十四 六十五 六十六 六十七 六十八 六十九 七十 七十一 七十二 七十三 七十四 七十五 七十六 七十七 七十八 七十九 八十 八十一 八十二 八十三 八十四 八十五 八十六 八十七 八十八 八十九 九十 九十一 九十二 九十三 九十四 九十五 九十六 九十七 九十八 九十九
一百 二百 兩百 三百 四百 五百 六百 七百 八百 九百
一千 二千 兩千 三千 四千 五千 六千 七千 八千 九千
一萬 二萬 兩萬 三萬 四萬 五萬 六萬 七萬 八萬 九萬 十萬 十一萬 十二萬 十三萬 十四萬 十五萬 十六萬 十七萬 十八萬 十九萬 二十萬 二十一萬 二十二萬 二十三萬 二十四萬 二十五萬 二十六萬 二十七萬 二十八萬 二十九萬 三十萬 三十一萬 三十二萬 三十三萬 三十四萬 三十五萬 三十六萬 三十七萬 三十八萬 三十九萬 四十萬 四十一萬 四十二萬 四十三萬 四十四萬 四十五萬 四十六萬 四十七萬 四十八萬 四十九萬 五十萬 五十一萬 五十二萬 五十三萬 五十四萬 五十五萬 五十六萬 五十七萬 五十八萬 五十九萬 六十萬 六十一萬 六十二萬 六十三萬 六十四萬 六十五萬 六十六萬 六十七萬 六十八萬 六十九萬 七十萬 七十一萬 七十二萬 七十三萬 七十四萬 七十五萬 七十六萬 七十七萬 七十八萬 七十九萬 八十萬 八十一萬 八十二萬 八十三萬 八十四萬 八十五萬 八十六萬 八十七萬 八十八萬 八十九萬 九十萬 九十一萬 九十二萬 九十三萬 九十四萬 九十五萬 九十

六萬
九十七萬 九十八萬 九十九萬

一百萬 二百萬 三百萬 四百萬 五百萬 六百萬 七百萬 八百萬 九百萬

一千萬 二千萬 三千萬 四千萬 五千萬 六千萬 七千萬 八千萬 九千萬

一億 二億 兩億 三億 四億 五億 六億 七億 八億 九億 十億 十一億 十二億 十三億 十四億
十五億 十六億 十七億 十八億 十九億 二十億 二十一億 二十二億 二十三億 二十四億
二十五億 二十六億 二十七億 二十八億 二十九億 三十億 三十一億 三十二億 三十三億
三十四億 三十五億 三十六億 三十七億 三十八億 三十九億 四十億 四十一億 四十二億
四十三億 四十四億 四十五億 四十六億 四十七億 四十八億 四十九億 五十億 五十一億
五十二億 五十三億 五十四億 五十五億 五十六億 五十七億 五十八億 五十九億 六十億
六十一億 六十二億 六十三億 六十四億 六十五億 六十六億 六十七億 六十八億 六十九億
七十億 七十一億 七十二億 七十三億 七十四億 七十五億 七十六億 七十七億 七十八億
七十九億 八十億 八十一億 八十二億 八十三億 八十四億 八十五億 八十六億 八十七億
八十八億 八十九億 九十億 九十一億 九十二億 九十三億 九十四億 九十五億 九十六億
九十七億 九十八億 九十九億

一百億 二百億 三百億 四百億 五百億 六百億 七百億 八百億 九百億

一兆 二兆 兩兆 三兆 四兆 五兆 六兆 七兆 八兆 九兆

一日 二日 三日 四日 五日 六日 七日 八日 九日 十日 十一日 十二日 十三日 十四日
十五日 十六日 十七日 十八日 十九日 二十日 二十一日 二十二日 二十三日 二十四日
二十五日 二十六日 二十七日 二十八日 二十九日 三十日 三十一日

一月 二月 三月 四月 五月 六月 七月 八月 九月 十月 十一月 十二月

一時 二時 三時 四時 五時 六時 七時 八時 九時 十時 十一時 十二時 十三時 十四時
十五時 十六時 十七時 十八時 十九時 二十時 二十一時 二十二時 二十三時 二十四時

一分 二分 三分 四分 五分 六分 七分 八分 九分 十分 十一分 十二分 十三分 十

四分 十五分 十六分 十七分 十八分 十九分 二十分 二十一分 二十二分 二十三分 二十四分 二十五分 二十六分 二十七分 二十八分 二十九分 三十分 三十一分 三十二分 三十三分 三十四分 三十五分 三十六分 三十七分 三十八分 三十九分 四十分 四十一分 四十二分 四十三分 四十四分 四十五分 四十六分 四十七分 四十八分 四十九分 五十分 五十一分 五十二分 五十三分 五十四分 五十五分 五十六分 五十七分 五十八分 五十九分
一秒 二秒 三秒 四秒 五秒 六秒 七秒 八秒 九秒 十秒 十一秒 十二秒 十三秒 十四秒 十五秒 十六秒 十七秒 十八秒 十九秒 二十秒 二十一秒 二十二秒 二十三秒 二十四秒 二十五秒 二十六秒 二十七秒 二十八秒 二十九秒 三十秒 三十一秒 三十二秒 三十三秒 三十四秒 三十五秒 三十六秒 三十七秒 三十八秒 三十九秒 四十秒 四十一秒 四十二秒 四十三秒 四十四秒 四十五秒 四十六秒 四十七秒 四十八秒 四十九秒 五十秒 五十一秒 五十二秒 五十三秒 五十四秒 五十五秒 五十六秒 五十七秒 五十八秒 五十九秒