國立交通大學

電信工程研究所

碩士論文

基於加權有限狀態轉換器國語語音辨認系統之設計

System design of a WFST-based Mandarin speech recognizer

研究生：蘇仲銘

指導教授：王逸如博士

中華民國一百零二年十一月

基於加權有限狀態轉換器國語語音辨認系統之設計

System design of a WFST-based Mandarin speech recognizer

研 究 生： 蘇仲銘　　Student：Chung-Ming Su

指導教授： 王逸如 博士　　Advisor：Dr. Yih-Ru Wang

國 立 交 通 大 學

電信工程研究所

碩 士 論 文

A Thesis

Submitted to Institute of Communication Engineering

College of Electrical and Computer Engineering

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in Communication Engineering

November 2013

Hsinchu, Republic of China

中 華 民 國　一百零二 年　十一 月

基於加權有限狀態轉換器國語語音辨認系統之設計

研 究 生：蘇仲銘　　指導教授：王逸如 博士

國立交通大學電信工程研究所碩士班

中文摘要

本論文主要針對語音辨識系統中之語言模型做改善，對訓練語料做正規化，包含合併同義詞、異體詞、又讀詞等等，選詞依照詞性分類，提高開放式詞類選入詞典之門檻，降低封閉式詞類之門檻，並且考量詞彙在訓練語料中分布之均勻性，最後再藉由音節解碼來評估語言模型。

實驗結果得知，在相同辨識效能下，加權有限狀態機的辨識速度比傳統辨識系統快將近 20 倍，因此本論文主要探討如何使用加權有限狀態轉換器來建構中文大詞彙連續語音辨識系統，首先介紹加權有限狀態轉換器的相關演算法，以及不同層級之語音模型如何以有限狀態機圖形來表示，並且以最佳化來縮小有限狀態機之路徑。接著調整建構語言模型之參數，改變加權有限狀態轉換器之大小，且改變辨識時所需之參數，探討這些因素和辨識率與辨識速度之間的關係。

System design of a WFST-based Mandarin speech recognizer

Student：Chung-Ming Su Advisor：Dr. Yih-Ru Wang

Institute of Communication Engineering
National Chiao Tung University

Abstract

This thesis is mainly focus on improving language model in Automatic Speech Recognition(ASR). The studies normalize the training data including combining synonym, variant word, multi-pronunciation. The words are categorized by word class to choose dictionary. Raise the opening word class threshold and reduce the closing word class threshold when choosing dictionary. We also consider the word distribution in training data when choosing word in dictionary. Using syllables to decode to estimate language model whether good or not after training language model.

We can find that the recognition rate of WFST is 20 times faster than traditional recognition system at the same recognition rate, hence this thesis is mainly studying how to use Weighted Finite-State Transducer (WFST) to build Large Vocabulary Continuous Mandarin Speech Recognition. We first introduce the algorithm of WFST and represent different ASR layer with WFST also use optimization to minimize WFST. Modify the features when building language model. Finally, we change the size of WFST and the features when recognizing, so we can find the relationship between recognition rate and recognition time.

致謝

無論如何要先謝謝爸爸、媽媽、妹妹，讓我可以寫這篇致謝拿到碩士學位，現在才知道舞台上那些得獎者總是第一個感謝家人，這行為不是例行公事是由衷發自內心想說的話；秀琇，謝謝你比我聰明，比我漂亮，比我亮眼，比我多人追，比我苗條，比我高個那麼一點，比我活耀，比我活潑，比我懂待人處事，比我有包容心，比我有遠見，比我有愛心，比我懂說話之道，比我說話流暢，比我會唱歌，比我懂得享受人生，讓我不斷地有往前邁進的動力，讓我更成熟，也謝謝你的陪伴。

謝謝王老師的嚴厲和堅持，教導如何學習和做研究，很幸運地當老師的學生，給我了磨練。感謝讓我找出研究盲點的性獸老師，謝謝你的耐心和親和力。

感謝相處一年的學長姐們，教了我許多實驗室生存之道的輝哥，實驗室第一個交談的智誠學長，願意讓我問問題的小蝦學長，願意讓我當跟班的子軒學長，口口聲聲賤 X 子的喬華學長，聰明偶爾給冷箭的企鵝學長，心情不好總是揍我的睿詮學長，跟我在夜店被拒絕一整夜的俊翰學長，耐心十足和教導我 WFST 的昂星學長，影片提供源的宅宅昌祐學長，籃球打贏我的維陽學長，趙同學口中 707 小 S 的雅婷學姊；感謝相處兩年的同儕們，帶領我融入交大文化和帶我去看他如何追學妹的子睿，滿口正妹經和充滿獄望之不定時炸彈的良基，充滿政治色彩和爬文忠實會員的奕勳，愛自各 murmur 和需要感謝我和良基關懷之實驗室邊緣人物的婉君，讓實驗室的人知道許多八卦和滿口瘦身經的靖觀；感謝相處一年半的學弟妹們，陪我奮鬥最後半年的小鋒，總是陪我閒晃的仲毛，一起出餿主意的王柏、阿璋和阿駿，被害者實驗室創造傳奇者的 ML，讓實驗室生活變精彩，愛看偶像劇和遊記的佩樺。

目錄

中文摘要.....3
Abstract.....4
致謝.....5
目錄.....6
表目錄.....8
圖目錄.....9
第一章　緒論.....10
1.1 研究動機.....10
1.2 文獻回顧.....10
1.3 研究方向.....13
1.4 章節概要說明.....13
第二章　語音辨識系統介紹.....14
2.1 聲學模型.....14
2.1.1 語音語料簡介.....14
2.1.2 聲學模型建立.....15
2.2 語言模型.....16
2.2.1 文字語料庫簡介.....16
2.2.2 文本前處理.....16
2.2.3 形音義分合詞處理.....18
2.2.4 選詞.....24
2.2.5 n-gram 語言模型.....26
2.2.6 Perplexity.....27
2.2.7 建立語言模型.....28
第三章　加權有限狀態機之語音辨識分析.....30
3.1 有限狀態機.....30
3.1.1 有限狀態機的簡介.....30
3.1.2 組合演算法.....33
3.1.3 確定性與非確定性.....35

3.1.4 最小化 36
3.1.5 取代演算法 37
3.2 語音系統中的有限狀態機 37
3.2.1 聲學模型 38
3.2.2 發音詞典 39
3.2.3 語言模型 41
3.2.4 合併各層有限狀態機 42
第四章 加權有限狀態機之實驗分析 44
4.1 文本前處理的分析 44
4.1.1 語料庫加入維基百科 44
4.1.2 依詞性分類選詞法 46
4.1.3 語言模型 PPL 的比較 50
4.2 音節的解碼來評估語言模型 52
4.3 HTK 辨識分析 55
4.4 WFST 辨識分析 57
4.5 辨識率與辨識時間的分析 58
4.5.1 有限狀態機大小的調整 59
4.5.2 加權有限狀態機大小與辨識率的關係 60
4.5.3 加權有限狀態機辨識率與速度的關係 61
4.5.4 測試語料的切短 62
第五章 結論與未來展望 64
參考文獻 65
附錄一:實驗所用 Variant Word Pair 表 67

表目錄

表 2. 1: TCC-300 語料庫統計表 15
表 2. 2: MFCC 參數抽取設定檔 15
表 2.3: 數量詞切短範例 18
表 2.4: 英文量詞範例 18
表 2.5: 漢字形音義異同表 20
表 2.6: 相差一字的四字詞 22
表 2.7: 相差一字的三字詞 22
表 2.8: 相差一字的二字詞 23
表 2.9: 四字詞兩兩對調 23
表 2.10: 合併前 ppl 28
表 2.11: 合併後 ppl 28

表 4.1: 混淆度的比較(對 inside test) 45
表 4.2: 由 IDF 移除的詞 47
表 4.3: 新增選詞法移除的詞彙 48
表 4.4: 六萬詞內各詞性詞彙更動的個數 49
表 4.5: 選詞前的混淆度(對 inside test) 50
表 4.6: 不同 discount 對 inside test 算 PPL 51
表 4.7: 不同語言模型對 inside test 算 PPL 的比較 55
表 4.8: 不同語言模型對 outside test 算 PPL 的比較 56
表 4.9: 不同語言模型的詞辨識率與字辨識率 56
表 4.10: 兩個語言模型的比較 58
表 4.11: 最終之有限狀態機之狀態與轉移數 58

圖目錄

圖 2.1: 文本前處理的步驟.....17
圖 2.2: 語言模型的建立流程.....29

圖 3 1: 有限狀態轉換機.....31
圖 3 2: 組合演算法的例子.....34
圖 3 3: 確定性的有限狀狀態機.....35
圖 3 4: 最小化的有限狀態機.....36
圖 3 5: 取代演算法.....37
圖 3 6: 聲學模型的有限狀態機.....39
圖 3 7: 聲學模型的 WFST.....39
圖 3 8: 線性詞典有限狀態機.....40
圖 3 9: 樹狀詞典有限狀態機.....40
圖 3 10:發音詞典的範例.....41
圖 3 11: 語言模型的有限狀態機.....42
圖 3 12: 語音辨識系統架構圖.....43

圖 4.1: 各個詞數的詞頻與涵蓋率的比較.....45
圖 4.2: 詞典中各詞性的涵蓋率.....50
圖 4.3: 不同 cutoff 對 inside test，PPL 與 arc 數的比較.....51
圖 4.4: 【研究生命起源】的正確音節序列作範例.....53
圖 4.5: 【研究生命起源】的正確音節為輸入，和 L。G 做合併的範例圖.....53
圖 4.6: 【研究生命起源】最佳路徑的範例圖.....54
圖 4.7: 各個辨識系統的語言模型評估值.....54
圖 4.8: 各個加權有限狀態基的轉移數.....59
圖 4.9: PPL、hypotheses 與評估語言模型之辨識率的關係圖.....60
圖 4.10: hypotheses 5000 與 7000 之辨識率與速度的關係圖.....61
圖 4.11: tg172 的辨識率與速度之關係圖.....62
圖 4.12: 句子切短前後之辨識率與速度的關係圖.....63

第一章　緒論

1.1 研究動機

科技不斷的進步，人類以最直接的方式聲音與機器之間的溝通不再是夢想，語音系統日趨成熟，廣泛的應用於生活中各個科技產品，大大提升了生活上的便利性。

隨著網路崛起，資訊的傳播不再受到區域的限制，使得資訊隨手可得且資訊量大增，為了因應廣泛的資訊，語音系統追求著含蓋龐大的資訊量，也由於電腦硬體的進步，實現了技術上的可行性，近年來大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)成為語音系統的趨勢，中文詞(word)的變化很大，例如:名詞、專有名詞、地方詞等等，此類的詞彙會隨著時間而增加，因此在詞典的數目受到限制情況下，以有限的詞典大小選出有效率的詞典成為重要的課題。

語音系統的普及化，人們開始追求語音系統能即時反應且精準的辨識，使語音系統能達到實用性，本研究欲以有限狀態機(Finite State Machine, FSM)以最佳化提升辨識速度，並且確認影響速度與辨識的要素，此為本論文研究的重要議題。

1.2 文獻回顧

有限狀態機(Finite State Machine)是一種簡單而有效率的數學演算法，近年來廣泛的運用在語音辨認系統中，如:用來表示前後文相關(Context dependency)的發音模型、表示大詞彙的字典…等等。最早由 A&T 實驗室的莫氏(M. Mohri)等人[1]-[3]提出使用加權有限狀態機(Weight Finite State Machine)，在狀態間的轉移上賦予一個加權值，將語音辨識中重要的機率分數以加權有限狀態機中之加

權值表示。加權有限狀態機之特性為利用組合運算整合了傳統語音辨認中各自獨立的聲學模型、發音詞典以及語言模型，將之整合為單一一個加權有限狀態機，並且運用了莫氏等人提出的確定性(determinization)以及最小化(minimization)演算法[1]、[4]來移除多餘之路徑，能有效的節省時間與空間。確定性可以在搜尋時減少有限狀態機存活的狀態數，最小化可求得狀態最少的等價有限狀態機；加權推移(Weight pushing)演算法[5]使語言模型的分數可以提早利用，辨識時移除掉不必要的路徑。臺灣大學的余氏[6]與交通大學的姜氏[7]曾以有限狀態自動機實做過中文大詞彙連續語音辨識，他們的論文核心在敘述有限狀態機的基本定義、建構流程，如何以有限狀態機建構一套大詞彙連續中文語音辨識系統，經由實驗結果可以證實，加權有限狀態機和傳統辨識相比，加權有限狀態機在相同辨識率下減少所需要的辨識時間。

在大詞彙語音辨識系統中，N 連語言模型(n-gram language model) [8]最常被使用，此模型以統計的方式來描述詞與詞之間相連接的機率，但我們無法收集所有 N 連詞彙的組合。而往後有學者提出方法來加強語言模型，1992 年 Brown 等人提出類別式 N 連語言模型(class-based n-gram language model)[9]，加入了類別資訊來訓練語言模型，將詞彙依照特性分群，則資料的預估由詞彙組合數降低為類別的組合數，能夠有效改善資料稀疏的問題。

以往中文大詞彙辨識採用的詞典，多數依照各詞彙在語料中的詞頻來排序，取其順位高者優先納入詞典中，由於中文構詞的多元與彈性，無法收入所有的詞彙，但不在詞典中的詞彙在辨識時即無法辨識出，這些不在詞典中的詞稱為 OOV word(Out-of-Vocabulary，OOV)，周氏[10]在其論文中提出階層式的辨識系統，針對中文構詞最為彈性的人名、定量複合詞與詞綴三個類別，以構詞學的角度出發，依照各種詞類的特性將之拆解，以較少數量的構詞單元收錄以提升詞的涵蓋率。

語言模型用來計算一個句子的機率，計算方式為預測之下一個詞藉由前一個

或前 n 個詞來得知該詞出現的機率，前 n 個詞即為 n-gram 語言模型之 n， n 越大，預測之詞所帶的資訊量越多，辨識時之錯誤率越低，反之，亦然。

大詞彙語音辨識系統中，OOV 出現是不可避免的，當文章中出現 OOV，Bart Decadt[11]等人提出大詞彙語音辨識系統提升輸出文章的可看性，概念為利用 phoneme-to-grapheme（P2G）後處理 phoneme 辨識出的不確切詞彙；以往處理 OOV 的方式為將之忽略或以已知的詞彙取代但處理方式不佳，Bart Decadt 等人將每個詞彙加上資訊以分數，表示此辨識後詞彙的可信度，若分數低於門檻，則此詞彙以 phoneme 經 phoneme-to-grapheme 轉換後輸出的詞彙取代之，本研究以 phoneme-to-grapheme 的概念來評估語言模型。

語言模型為語音辨識系統中最困難突破的瓶頸，在發達的科技時代中，語言模型在辨識時需要快速，大小需要精小不能過於龐大，但往往語言模型內會存至千萬個 n-gram，因此對語言模型的儲存空間為一大挑戰；Zhijian OU[12]提出語音辨識系統以 WFST 表示時，轉移(arc)含有五個參數，包含初始狀態、終止狀態、輸入字元、輸出字元和權重，每個轉移記憶體需要 20bytes 儲存，狀態(node)帶有指向轉移的資訊，每個狀態記憶體需要 4bytes 儲存，因此一個加權有限狀態機記憶體需 4× |WFST_nodes|+20|WFST_arcs|，但並非每個轉移皆需儲存初始狀態，因此，記憶體儲存一個加權有限狀態機改為 4× |WFST_nodes|+16|WFST_arcs|。

語音辨識為根據前面的參數找出與正確解答最相近的詞串，在搜尋時，相似的詞串相當多，通常在辨識時，只會搜尋部分，仍可能在辨識一個句子會歷經語言模型中千百條路徑，因此辨識所需之時間費時，但搜尋的路徑越多，辨識結果會越為準確；反之，搜尋的路徑越少，辨識結果的準確度會下降，但辨識速度會變快，因此，時間、空間與辨識率之間的取捨成為學者們探討的問題。

1.3 研究方向

本論文中之語音辨識系統主要針對語言模型做改善，包含訓練語料的正規化、選詞方式皆有做更新，提升語音辨識系統辨識的準確性，並且針對語言模型提出直接的方法來評估語言模型；語音辨識系統以加權有限狀態機取代傳統辨識系統，在辨識速度上有明顯的提升，再討論影響辨識率與辨識速度之因素，探討 n-gram 之 n 與語言模型之後撤平滑化(back-off smoothing)改變時，語音辨識系統大小與系統涵蓋資訊的改變，導致語音辨識系統的辨識效能與辨識速度之變化，找出最適宜之加權有限狀態機。

1.4 章節概要說明

本論文一共分為五章，其各章節內容分配如下：

第一章：緒論

第二章：語音辨識系統介紹

第三章：加權有限狀態機之語音辨識分析

第四章：加權有限狀態機之實驗分析

第五章：結論與未來展望

第二章　語音辨識系統介紹

本章介紹辨識系統中各個層的建立，包含聲學模型、發音詞典與語言模型，聲學模型使用 TCC-300 語料庫建立，以隱藏式馬可夫模型表示；語言模型的建立則由文字語料開始介紹，經由文字前處理，選詞方式改為先依詞性分類，再選詞，最後以 n-gram 建立語言模型。

2.1 聲學模型

本節先介紹訓練聲學模型的語音語料庫，再介紹聲學模型的相關參數設定。

2.1.1 語音語料簡介

在本研究中使用 TCC-300 麥克風語音資料庫，此資料庫由國立台灣大學、國立成功大學及國立交通大學的 300 位同學共同錄製，中華民國計算語言學學會所發行，屬於麥克風朗讀語音,主要目的是為了提供語音辨識研究，檔案統計資料如表 2.1 所示。語句取樣頻率皆為 16000 赫茲（Hertz），取樣位元數為 16 位元，音檔檔頭為4096位元。將此語料庫再區分為訓練語料及測試語料，訓練語料的部分約占 90%，共 274 位語者，長度共約 23 小時；測試語料的部分約 10%，共 29 位語者，長度約 2.43 小時。在進行辨識時，所使用的測試語料為交通大學與成功大學的長句音檔，共 29 位語者 226 句長句音檔，長度約 2 小時，詞總數量為 15497，每個句子平均含有 117.2 個音節。從中挑選十分之一的檔案做測試音檔，包含 29 個語者，15 位男性，14 位女性，音節總數為 26472 個，十分之九則做為訓練音檔，音節總數為 300836 個。

表 2. 1: TCC-300 語料庫統計表

學校名稱	文章屬性	語者總數		總音節數		音檔總數	
台灣大學	短文	男	50	男	27541	男	3425
		女	50	女	24677	女	3084
		總數	100	總數	52218	總數	6509
交通大學	長文	男	50	男	75059	男	622
		女	50	女	73555	女	616
		總數	100	總數	148614	總數	1238
成功大學	長文	男	50	男	63127	男	588
		女	50	女	68749	女	582
		總數	100	總數	131876	總數	1170

2.1.2 聲學模型建立

在語音辨識時，首先需對輸入的語音抽取出語音參數，由於考量到短時間穩定特性與人耳聽覺效應的補償作用，因此使用 MFCC 參數(Mel-Frequency Ceptral Coefficients，梅爾倒頻譜參數)，它的成份含 12 維 MFCC 加上 1 維能量共 13 維，並取其 Delta 和 Delta-Delta term 用以描述參數變化訊息，最後共可得 39 維參數。本次實驗訓練的模型為中文單音節(mono-syllable)模型一共 411 個音節，每個音節使用 8 個狀態(state)的隱藏式馬可夫模型(HMM)表示之，並使用 HTK 中的 MMI 鑑別性訓練得到。訓練相關設定如下表:

表 2. 2: MFCC 參數抽取設定檔

Frame size	32ms
Frame shift	10ms
Filter bank number	24
Sampling frequency	16kHz
Pre-emphasis Filter	First order with coefficient 0.97

2.2 語言模型

本節介紹訓練語言模型所使用的文字語料庫，接著介紹文字語料庫的處理，並且選詞產生發音詞典，最後介紹使用 n-gram 計算詞與詞之間的機率建立語言模型。

2.2.1 文字語料庫簡介

用於訓練語言模型的文字資料庫共有以下來源：

1.) 光華雜誌(Sinorama)：內容為一般雜誌的文章，蒐集的資料年代範圍介於 1976 年到 2000 年之間。

2.) NTCIR：為一個建立資訊檢索系統的標竿測試集，其內容由數種不同學科領域文章構成。

3.) 中研院平衡語料庫(Sinica)：它是一套由中研院收集，內容包含多種主題，以語言分析研究為目的的資料庫。

4.) Chinese Gigaword：由 Linguistic Data Consortium (LDC)整合發行，內容包含台灣中央社、北京新華社等國際新聞。

5.) 維基百科語料(Wiki) ：維基百科為領域廣泛且資訊較為新，可以使語言模型更加多元，資料庫增加。

2.2.2 文本前處理

由於語料中有各種的文字組合，所以在訓練語言模型前先把文字、語句盡可能的整合，文字都應以從口中說出為出發點，例如特定符號、數字轉為文字，本實驗針對中文做研究，異國詞彙則不採用，移除文章中不合理的語句，結合以上種種的文字處理，文本前處理大致上分為 CRF 斷詞、文字正規化...等等，以下為文本前處理流程:

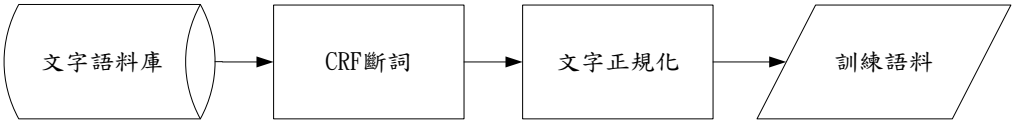


圖 2.1: 文本前處理的步驟

1.) CRF 斷詞

斷詞是決定語言模型好壞的一大重要元素，把語料斷為以詞為單位，詞的邊界成為斷詞的關鍵，以條件隨機域(conditional random field, CRF) [13]方法進行斷詞，此法主要藉由標記詞性與學習句法結構來進行斷詞，相較於傳統以詞典為基礎之長詞優先的斷詞規則，使用 CRF 斷詞可產生較正確的辨認結果，也能將詞典未收錄的詞正確斷出，一方面減少 OOV 所帶來的連續短詞串問題，另一方面也能擴充人名詞、詞綴詞等清單。本研究斷詞為王逸如老師所撰寫，目前的 F-measure 為 97.49%。

2.) 文字正規化:

2.1) 文字正規化處理:

斷詞後的詞必須符合口語的發音文字,語料中出現符號、數字需文字化;語料中部份內容為強調、註解、翻譯，在語音讀法時，這類的內容通常不會被讀出，需移除；寫法不同，但讀音和語義相同的詞，需合併等等的現象，都需要文字正規化。

2.2) 部份數量詞切短:

經由 CRF 斷詞後，屬於 Neqa、Neu 詞性的詞皆為長詞，我們將若干長詞切短，以短詞的形式收錄於詞典中，由於詞典的容量有限且數量定詞和量詞會有各種可能的組合，無法將其全部收錄於詞典中，如此也能增加詞典的涵蓋率。

表 2.3: 數量詞切短範例

正規化前	正規化後
百分之五十_Neqa	百分之_Neqa2 五十_Neu2
一點五二_Neu	一_Neu 點五二_Neu
八億五千萬_Neu	八億_Neu6 五千_Neu4 萬_Neu5

2.3) 符號、POS、異國詞彙處理

中文共有十六種標點符號(PM)，其中又可分為標號與點號兩類，而點號與說話的停頓有較大的關聯性。我們藉由點號中的句號、驚嘆號、問號、分號將文章分段，並將除此之外所有的標點符號移除，並一併將 POS 標記也移除；部份語料的內容在語音辨識上，是不會出現的，包括註解或強調語句，在一般口語上不會出現，這些語句會造成詞和詞之間的統計數據不準確，對語音辨識是沒有幫助的，所以需要將之移除，例如:今(十三)日修改為今日；將部份符號取代為文字，例如；連接詞符號～取代為文字"至"。

本實驗的辨認目標為中文詞彙，故將文章中的英文詞以「FW」標記為同一個類別，FW 類別並沒有收錄進訓練詞典中，而是視為一個 OOV 對待，且量詞符號由文字取代。

表 2.4: 英文量詞範例

正規化前	正規化後
℃	度西
km	公里
kg	公斤

2.2.3 形音義分合詞處理

中文辨識因應形音義的異別，有些詞類其實是可以合併訓練，例如異體字

在發音上與語意上皆相同，僅有字形不同，若視為不同詞彙看待會在辨認時造成混淆，字形、字音、字義三者的關係與分合情況容易影響到訓練與辨識，以下將簡介漢字的特質，並就各式特質提出我們在處理語料時所應對的方法。

漢字具有的三大要素為：「形、音、義」，其中字義為我們語文的核心，字形、字音都是為字義而存在。在文化的演進中，有些字形變得不一致、或因沒有創字而借用，各種複雜的因素使得漢字形成了「多形、歧音、異義」的狀況，所以目前所使用的「漢字」呈現出字形不一、字音分歧、字義寬廣的特質。

1.) 字形不一

歷史上漢字有甲骨、金文、篆、隸、楷、行、草等不同形體，如今使用者也有簡體／繁體的差別。也存在有結構上同字但異形的差別，例如兇宅的「兇」一字也有人寫作「凶」、曝曬的「曬」寫作「晒」、藥局的「藥」寫作「葯」。在字形不一的情況下，影響到的是斷詞器的訓練、斷詞詞頻統計、詞典收錄、語言模型統計...等。

2.) 字音分歧

一字多音一向是漢語的特色，當中音變而意思不同者俗稱破音字。例如「了（ㄌㄧㄠˇ）解」、「作了（ㄌㄜ˙）手腳」。字音的分歧所影響的是詞典收錄，就破音字意義不同的層面來看，也影響了語言模型的訓練（尤其指單字詞的情況）。

3.) 字義寬廣

在漢字中有一字多義的情況存在，相同的字形可同時代表不同意義，如「叫(大聲喧嘩)囂」、「叫(開價)價」、「叫(喝采)好」、「叫(叫喊冤屈)屈」。

「多義字」、「同音字」的異別可以經由 *n*-gram 語言模型學習到，而「又讀字」可用 multi-pronunciation 形式收錄在發音辭典中，額外處理的三個項目為：

(A) 異體字:

異體字僅字形上的差異,字義為相同,若異體字皆收入進詞典中，會使機率分散，也會使 perplexity 增高，為了避免，將異體字合併為同一詞納入詞典。

(B) 破音字；

同形異音異義的「破音字」和同形異音同義的「又讀字」,這些同形歧音(multi-pronunciation)傳統上都會將歧音字同時收入進詞典中，但由於此斷詞後的詞不帶音節，所以無法辨別歧音，如此等同在訓練語言模型時將所有歧音字合併訓練。本實驗採用將歧音詞納入發音詞典。

(C) 同義詞:

針對單字詞的情況，同義字是不該被合併訓練的，儘管「足」跟「腳」係屬同義字，但前後文通常存有差異，故不對單字詞同義字進行處理。延伸的情況為同義詞(variant word)，所指為語義相同但字形不盡相同的詞，我們希望同義詞能在語言模型中共享相同的分數。

表 2.5: 漢字形音義異同表

形	音	義	現象	例子
同	同	異	多義字	作(改變臉色、戲弄、當作...)
同	異	同	又讀字	多(ㄉㄨㄛ/ㄉㄨㄛˊ)麼不容易
異	同	同	異體字	凶/兇、姐/姊
異	異	同	同義字	足/腳、首/頭
異	同	異	同音字	ㄗㄨㄛˋ(坐、座、作...)
同	異	異	破音字	樂(ㄩㄝˋ/ㄌㄜˋ)

4.) 針對同義詞、又讀詞、異體詞的處理

在眾多詞的詞典中，仍然有許多詞寫法不同但為同義，這些具有相同意思的詞同時存在詞典中，會使機率分散；將這些詞合併，不但能降低辨識時混淆，增加詞典的涵蓋率，還可以減少 OOV 數目，是影響語言模型的重要一環；找出辨識時容易混淆的詞，例如:詞頻相近的詞等等，由於語料庫相當的龐大，無法用人工的方式尋找，Data Drift(飄移資料)為快速且有效率的尋找方式，以下為兩種漂移資料:

(A) 相差一字的漂移資料:

語料中有許多詞彙相差一個字，由實驗結果觀察得知，語料中詞彙和詞彙間相差一個字，可以有效的找出詞頻相近的詞，也可以使落在發音詞典外的詞彙，因為詞的合併，而有機會被選為發音詞典，同時能找出語料中詞彙有錯字的詞；這裡針對四字詞、三字詞和二字詞，這些混淆詞彙可以至教育部網站查詢正確的寫法，若寫法都正確，將採用詞頻較高、較多人使用的詞彙，以下為詞頻相近的詞，四字詞和三字詞中可以看出有許多詞為外國人名或外國地名等等，因為翻譯的關係，外語翻成中文時，往往會被翻成不同的字。

表 2.6: 相差一字的四字詞

詞彙	詞頻	詞彙	詞頻
克什米爾	1133	喀什米爾	915
根深蒂固	512	根深柢固	488
藉題發揮	150	借題發揮	248
精疲力竭	215	筋疲力竭	195
洛克哈特	307	羅克哈特	599
辛辛那提	286	辛辛那堤	114
斯馬庫斯	205	司馬庫斯	154
劃地自限	152	畫地自限	123
乘虛而入	131	趁虛而入	157

表 2.7: 相差一字的三字詞

詞彙	詞頻	詞彙	詞頻
科索伏	3603	柯索伏	4919
渡假村	1012	度假村	1369
威廉絲	1632	威廉斯	1812
保特瓶	426	寶特瓶	1267
路易士	731	路易斯	1008
大姆指	200	大拇指	613
禁得起	524	經得起	834
富比世	700	富比士	368

表 2.8: 相差一字的二字詞

詞彙	詞頻	詞彙	詞頻
復建	4963	復健	5186
布設	196	佈設	374
紀事	435	記事	296
檢查	74893	檢察	7894
勘查	8159	勘察	4956
巡查	1661	巡察	2343

(A) 四字詞中兩字的轉移

語料中，可以發現有許多四字詞，前兩字和後兩字對調後仍為同義，實驗結果可以看出，這些四字詞多半為成語。

表 2.9: 四字詞兩兩對調

詞彙	詞頻	詞彙	詞頻
物美價廉	610	價廉物美	507
不發一語	500	一語不發	131
光明正大	372	正大光明	96
傾盆大雨	538	大雨傾盆	100

5.) 處理外來語註解

維基百科中有許多人名、地名、專有名詞等等的註解，且註解內多數為外來語，這些外來語的註解不但對漢語研究沒有幫助，還會影響詞與詞相連接的機率，因此，將語料中註解內詞性皆為外來語(FW)的情形移除，例如:地質學(Geology)是對地球的起源，由範例可知，地質學應與是對地球的起源相連接，但由於外來

語註解的關係，使之拆開，如此影響了語言模型中 bi-gram 和 tri-gram 的分數，因此將註解中皆為外來語移除。

2.2.4 選詞

漢語詞類可以分為兩類，實詞與虛詞；實詞有詞彙意義和語法意義，大多屬於開放類，開放類的成員會隨著時間而日益增加，數目為無限的，詞性屬於開放類的有名詞、動詞、形容詞、非謂形容詞、副詞、數詞、量詞、代詞、感嘆詞、擬聲詞。虛詞只有語法意義，多數屬於封閉類，封閉類的成員不會隨著時間而增加，數目為有限的，詞性屬於封閉類的有介詞、連詞、助詞、語氣詞。

選詞是建立語言模型的重要一環，受限於記憶體的大小，無法將語料中所有的詞彙皆列入詞典中，因此應將有限數目的封閉類詞彙盡可能的收錄進詞典中，而慎選數目無限的開放類詞彙，可以使選出的詞典在有限的數目達到更高的效率，另外，詞彙還會受到時間、地點等等影響出現的頻率，例如前陣子受到鳥類傳播流感病毒的關係，H7N9 這個詞彙頻頻的出現；若身處於印度，想必伊斯蘭教這名詞將耳熟能詳，以上即為詞彙受時間地點的影響，某些詞彙只會出現在特定文章中，在特定領域的文章中，該詞彙即頻頻出現，在其他領域的文章中卻微乎其微，本實驗研究並非針對特定領域做中文辨識，因此僅參考詞頻而直接收錄高詞頻的詞彙做為詞典並不合理，我們欲收錄普遍出現的詞彙，也就是廣泛出現在各個文章中的詞彙。

本研究先將詞彙依照詞性分類，在選詞時，可以避免移除一些不必要移除的封閉類詞彙，並且提高開放類詞彙的篩選門檻，降低封閉類詞彙的門檻，目的為使詞典中的封閉類詞彙增加，開放類詞彙下降，詞彙依照詞性分類後，篩選詞彙的方式，傳統以詞頻高低來選取詞典，這種選詞法可以降低 OOV 的數目，但沒有考量詞彙是否只會出現在特定領域，我們想要找廣泛出現在各個文章中的詞彙，因此學長採用 IDF(inverse document frequency)，IDF 表示一個詞彙普遍的重要性，

代表其詞彙類別區分能力隨著在語料庫各文章中出現的頻率反比下降，是一種用於資料檢索(IR-Information Retrieval)的常用加權技術，用於評估一個詞對於一個文件或一個語料庫中的其中一份文件的重要程度。算式如下:

$$idf_i=\log\frac{|D|}{|\{d:d\ni t_i\}|} \quad (2.1)，$$

其中 D 表示所有文件的集合，分子|D|表示語料庫中的文件總數，d 表示文件，t_i表正在處理的詞條，分母則表示包含該詞條t_i的文件數目。由於"出現的文章數"在分母項，因此我們將挑選 IDF 分數低的詞收錄進詞典中。

最終目標是希望能依照詞頻與出現文章篇數，均勻的移除詞彙，而不單只是移除出現文章篇數少的詞彙，IDF 做法僅依據詞的出現篇數，把出現篇數極端低的詞剔除，出現篇數偏高但仍分布不均勻的詞彙無法被移除，因此新增一選詞法來篩選詞典，運算式如下:

$$[1-(1-P)^n]\times N \quad (2.2)，$$

其中 P 如公式(2.3)，N 表示文章的總篇數，n 表示平均一篇文章中的詞彙數，算法為如公式(2.4)。

$$P=\frac{(正在處理的詞頻)}{(總詞頻)} \quad (2.3)，$$

$$n=\frac{(總詞頻)}{(文章總篇數)} \quad (2.4)，$$

上式依據各個詞彙的詞頻，計算出各別應出現的文章篇數，我們再參考各個詞彙實際出現的文章篇數，將實際出現的文章篇數除以應出現的文章篇數，得到一個比值，依照這個比值大小排序，即可得知各詞彙的分布狀況，比值越高表示該詞出現的文章篇數越接近應當出現的文章篇數；反之，比值越低表示該詞彙實

際出現的文章篇數和應當出現的篇數越有差距，我們即可判定該詞只出現在某些特定文章中，分布不均勻，這些詞就不適合選入詞典。

2.2.5 n-gram 語言模型

各種語言都有它的文法規則，透過大量的文字資料來進行訓練，統計出其語言的規則，利用這種文法規則建立出的機率模型稱之為語言模型，在大詞彙連續語音辨識時，會利用語言模型，考慮前後詞彙的關聯性，使輸入的語音能辨識出一串有意義的詞串。這裡使用廣為人知的 n-gram 語言模型，此模型假設任一個詞在詞串中只受到前 $n-1$ 個詞的影響，令 $W= w_1 w_2 \cdots w_N$ 為一個 N 個詞的詞串，其中 w_k 代表句中第 k 個詞，則第 k 個詞所出現的機率表示為 $P(w_k|w_{k-n+1} w_{k-n+2} \dots w_{k-1})$，這個 N 詞長的 W 詞串之出現機率可展開為:

$$P(W)=P(w_1)\cdot P(w_2|w_1)\cdots P(w_i|w_{i-n+1} \dots w_{i-1}) \dots P(w_N|w_{N-n+1} \dots w_{N-1}) \quad (2.5)，$$

$$P(w_i|w_{i-n+1} \cdots w_{i-1})= \frac{\text{Count}(w_{i-n+1},\dots,w_i)}{\text{Count}(w_{i-n+1},\dots,w_{i-1})} \quad (2.6)，$$

由於 n-gram 語言模型是統計式的模型，如果訓練語料中沒出現該詞語組合，就無法預估其機率，且隨著 n 值上升，所需的訓練語料也呈指數成長。為了解決這些問題，以後撤平滑化(back-off smoothing)來調整模型的機率分布。當詞串 $w_{i-n+1} \dots w_{i-1}$ 不存在時，捨棄距離最遠詞的資訊，以低一階的 $w_{i-n+2} \dots w_{i-1}$ 機率乘上後撤加權值 α，變為 $\alpha(w_{i-n+1} w_{i-n+2} \dots w_{i-1}) \cdot P(w_i|w_{i-n+1} w_{i-n+1} \dots w_{i-1})$。若也沒有 $P(w_i|w_{i-n+2} w_{i-n+3} \dots w_{i-1})$ 的資訊，繼續後撤並逐一乘上後撤加權值。改寫機率預估式如下:

$$P(w_i|w_{i-n+2}, \dots, w_{i-1})$$

$$= \begin{cases} \alpha(w_{i-n+1} \cdots w_{i-1}) P(w_i|w_{i-n+2} \cdots w_{i-1}) & ,\text{Count}(w_{i-n+1}, \dots, w_i) = 0 \\ d_\alpha \cdot \frac{\text{Count}(w_{i-n+1},\dots,w_i)}{\text{Count}(w_{i-n+1},\dots,w_{i-1})} & ,1 \le \text{Count}(w_{i-n+1}, \dots, w_i) \le k \\ \frac{\text{Count}(w_{i-n+1},\dots,w_i)}{\text{Count}(w_{i-n+1},\dots,w_{i-1})} & ,\text{Count}(w_{i-n+1}, \dots, w_i) > k \end{cases} \quad (2.7)，$$

後撤加權值$\alpha(w_{i-n+1}\cdots w_{i-1})$需經過正規化(normalization)處理，並滿足條件式:

$$\sum_{w\epsilon V} P(w_i = w|w_{i-n+1} \ldots w_{i-1}) = 1 \quad (2.8)，$$

另外，當 Count(·)的數值很小時，可能造成預估的不準確性，因此以d_α(Discount Coefficient Factor)來進行平滑化。當一詞串組合的出現次數小於某設定的次數時，我們將原始預估的 n-gram 機率乘上d_α值，d_α依據 Good-Turing discounting 計算，並將 discounting 扣除的機率值再平分給詞串沒有出現的 n-gram 機率使用。

另外，音節序列越長，表示擁有相同音節序列的詞彙序列越少，辨識出錯誤詞彙的機率越低，因此平均詞彙的長度越長越好。

2.2.6 Perplexity

Perplexity 為評估語言模型的重要依據，他的中文叫混淆度，簡寫為 PPL，混淆度是根據消息理論(information theory)而得，式子如下:

$$H = -\frac{1}{m}\log P(W = w_1，w_2，\cdots，w_n) \quad (2.9)，$$

上式為一個詞串$W = w_1，w_2，\cdots，w_n$，對於每個新詞提供的平均資訊量(entropy)，經過 ergodic 的假設和適當化簡而得。而混淆度可以直接進一步定義為:

$$PP=\exp(H) \quad (2.10)，$$

若 $P(W = w_1，w_2\cdots w_n) = \prod_{i=1}^{n} P(w_i \mid w_1，w_2\cdots w_{i-1})$則可發現，混淆度就是$P(w_i \mid w_1，w_2，\cdots，w_{i-1})$的幾何平均數的倒數。因此混淆度可以解讀為語言模型估測一個歷史詞串後面，平均可能的可接詞數；混淆度越高，表示一個歷史詞串後接詞有較多的選擇，辨認相對的就越難找到確切的詞彙；反之，混淆度越低，則越容易找到正確的詞彙。

$$ppl= 10^{(-\text{logprob}/\text{words}-\text{OOVs}+\text{sentences})} \quad (2.11)，$$

$$ppl1= 10^{(-\text{logprob}/\text{words}-\text{OOVs})} \quad (2.12)，$$

其中 words 表示詞的總數，OOVs 表示詞典外的詞(out-of-vocabulary)數量，sentences 表示為測試的句子數量；logprob 包含</s>句子結束符號，所以平均每個 word 的 perplexity 是以 ppl 為計算方式，若排除</s>句子的結束符號，則 perplexity 以 ppl 為計算方式。

前面介紹的同義詞處理,這裡以 perplexity 將同義詞合併前與合併後做比較，在訓練語言模型碰到同義詞時，會將相同意思的同義詞視為一組，並將這組同義詞合併做訓練，又讀詞和異體詞亦是如此，當語言模型層展開至詞典層時，會將先前合併的 variant word 補回語言模型中。

異體字與同義字的判別有難度，因此，這些詞的搜尋絕大多數採用人工搜尋，決定是否合併，截至目前為止，異體字與同義字被修正的詞條數共為 2046，總詞數為 3772518，下表為同義字與異體字在語料庫中合併前後的混淆度比較，計算混淆度的對象為 inside test。

表 2.10: 合併前 ppl

Order	ppl	ppl1
3	195.255	240.109

表 2.11: 合併後 ppl

Order	ppl	ppl1
3	189.736	233.092

2.2.7 建立語言模型

此研究分別建立 unigram、bigram 與 trigram，並針對不同的 smoothing 做了比較，本實驗訓練語言模型所使用的軟體為 SRILM[14]，建立語言模型的流程如

下:

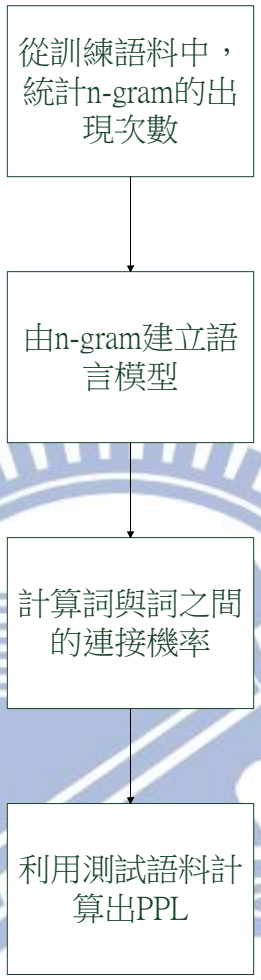


圖 2.2: 語言模型的建立流程

第三章　加權有限狀態機之語音辨識分析

傳統語音辨識系統受限於詞典大小與辨識速度，大詞彙語音辨識系統近年來廣泛的使用加權有限狀態機(Weighted Finite-State Transducer,WFST)，讓語音辨識系統更往前邁進一大步。

本章先簡介加權有限狀態機與有限狀態機的演算法，再介紹將語音辨識系統中每一層轉換至加權有限狀態機表示，並且合併與最佳化。

3.1 有限狀態機

本節介紹加權有限狀態機，與語音辨識系統使用的演算法，例如:組合演算法、確定性與最小化等等。

3.1.1 有限狀態機的簡介

有限狀態機可以分為兩類，有限狀態自動機(finite state automata)和有限狀態轉換機(finite state machine)，有限狀態自動機的圖中，可看到點(node)在此稱為狀態(state)、邊(arc)在此稱為轉移(transition)及邊上的字元為此轉移的輸入字元(input symbol)，初始狀態(initial state)由粗線圈表示，終止狀態(final state)由雙線圈表示，若狀態同時為初始與終止狀態，則以粗雙線圈表示，有限狀態自動機由五個元素(Q, I, F, Σ, δ)組成，Q 為所有狀態的集合，I 為初始狀態，F 為終止狀態的集合，Σ 為輸入字元集合，δ 為轉移函式，若給一串輸入字串(input string)，經由狀態轉移後，得到輸出為接受(accept)或拒絕(reject)；相較於有限狀態自動機，有限狀態轉換機則在邊上多了輸出字元(output symbol)，因此，輸入一串輸入字串，有限狀態轉移機輸出不僅得到接受或拒絕，還會得到一連串經由狀態轉移後

的輸出字串，有限狀態轉換機可以看成是有限狀態自動機的延伸，實際應用通常邊上會有權重，加了權重後的有限狀態自動機與有限狀態轉換機分別為加權有限狀態自動機與加權有限狀態轉換機。

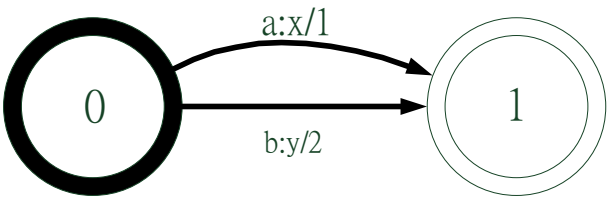


圖 3 1: 有限狀態轉換機

有限狀態機由六個元素(Q, I, F, Σ, △, δ)所組成:

1) Q:所有狀態的集合，Q={0,1}
2) I:初始狀態，有限狀態機只會有一個初始狀態，I={0}
3) F:終止狀態，有限狀態機的結束狀態，有限狀態機至少有一個以上終止狀態，F={1}
4) Σ:輸入字元集，Σ={a, b}
5) △:輸出字元集，△={x, y}
6) δ:轉移函式，表示輸入字元經過狀態後，轉移至另一狀態且輸出字元。

下面為上述的名詞解釋:

1) 狀態

   有限狀態機必須為有限個狀態，其中包含一個初始狀態與一個以上的終止狀態，開始接由初始狀態進入，經過一連串的狀態轉移，當最後一個狀態轉移完成後，若此時的狀態停止在終止狀態，則表示此條路徑為接受；反之，非停止於終止狀態，則此路徑為拒絕。

2) 游標

有限狀態機的輸入字串有一游標，表示有限狀態機目前正在執行的字元，起始時，游標位於輸入字串的第一個字元，每當一轉移發生，游標即往後移動一個字元，直到字串結束為止。

3) 轉移

轉移表示狀態與狀態間的轉移關係，由轉移函式 δ 表示，轉移 t 包含來源狀態(source state，符號為 s[t])、目的狀態(destination state，符號為 d[t])、輸入字元(input symbol，符號為 i[t])、輸出字元(output symbol，符號為 o[t])和權重(weight，符號為 w[t])，轉移寫作為 I:O/W，I 為輸入字元，O 為輸出字元，W 為權重；轉移方式為狀態由來源狀態轉移至目的狀態，當游標指一輸入字元時，若有限狀態機無對應的來源狀態和輸入字元，則輸出為拒絕。

4) 輸入輸出

在轉移上以文字表示，輸入和輸出會以冒號隔開，若輸入為 x，輸出為 y，則表示為 x:y。

5) 空轉移

空轉移的表示法為ε(eplison)，當轉移上的輸入字元為ε時，表示此轉移不需要輸入也可以進入下一狀態；而當轉移上的輸出字元為ε時，表示此轉移並不會輸出字元

6) 路徑

路徑由一連串相連的轉移所組成，假設 $P=p_1p_2 \dots p_n$ 為一條路徑，其中 p_i 表示路徑上第 i 個轉移，$s[p_{i-1}]=d[p_i]$，當結束狀態為 $d[p_n]$，則此為一條被接受的路徑。

7) 加權值

有限狀態機可以藉由加權值賦予不同的權重，以 w[t]表示轉移 t 的加權

值，在描述語音辨識所用的有限狀態機時，利用加權值來表示各種模型的分數，除了在轉移上會帶有權重之外，每個結束狀態也可以再賦予加權值，一般採用 log semi-ring 的數學模型，對機率的轉移取 negative nature log，尋找最佳路徑時則為搜尋累積加權最小的路徑。

8) 等價性

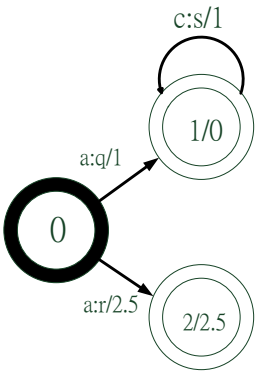
若兩個有限狀態機所接受的語言相同，稱它們為等價(equivalent)，但它們的狀態與轉移不一定相同。語音辨識中，提升搜尋效率與有效的利用空間相當重要，如何使處理後的有限狀態機在等價的情況下，將有限狀態機的狀態和轉移數減少是一大學問。

3.1.2 組合演算法

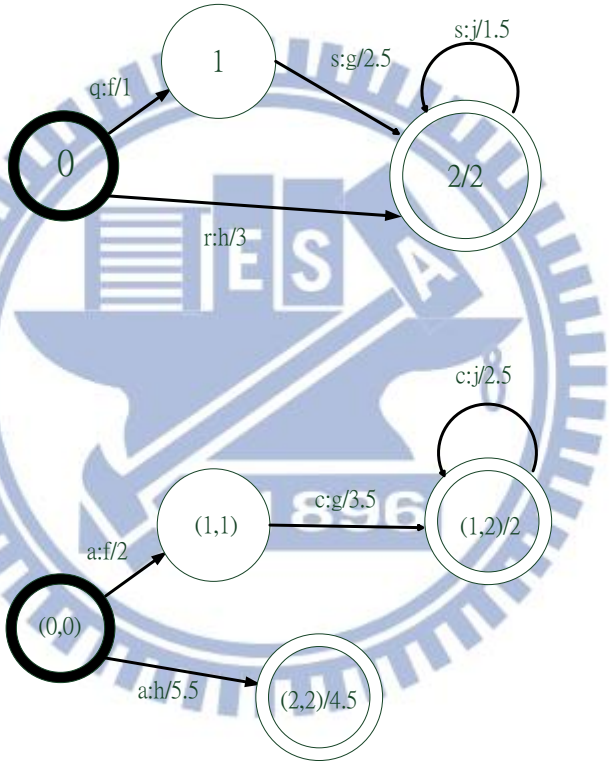
有限狀態機的一大特性為，可將各個不同的層級進行組合演算法整合在一起，得到一個整合的有限狀態機，運用在語音辨識上，將聲學模型的有限狀態機和發音詞典的有限狀態機作組合演算法合併，可使得語音訊號轉換為單獨的詞彙，將發音詞典的有限狀態機和語言模型的有限狀態機作組合演算法合併，可使得音節序列轉換為詞彙序列，最終將聲學模型、發音詞典和語言模型的有限狀態機合併為一個有限狀態機，其中各個層級的有限狀態機皆為獨立建構，因此各個層級做更改也相當容易。

令兩個有限狀態機 A 和 B，若 A 轉移上的輸出字元和 B 轉移上的輸入字元相同，則路徑連接，反之，則 A 有限狀態機的路徑無法和 B 有限狀態機的路徑連接，最後成功連接的路徑整合為有限狀態機，為 A 和 B 整合出的新有限狀態機 C，寫作 C=A。B，C 的每個狀態和轉移皆由 A 和 B 所組成。

A:



B:



C=A。B:

圖 3 2: 組合演算法的例子

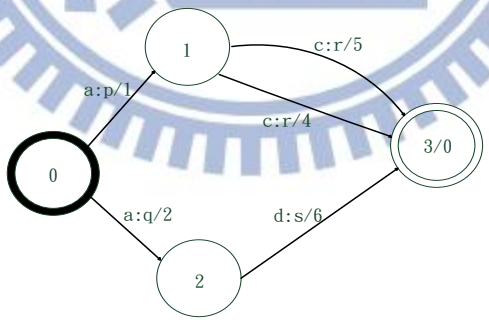
但一般組合演算法後的有限狀態機皆相當的龐大，因此需要對其壓縮減少它的大小，有兩種使有限狀態機縮小的方式，分別為確定性(determinize)與最小化(minimize)，以下為兩種方式的介紹。

3.1.3 確定性與非確定性

當有限狀態機中，來自所有相同來源狀態的轉移，其輸入字元皆為重覆，則稱此有限狀態機具有確定性(deterministic)，確定性的有限狀態機在任何時刻皆只會有一個狀態的可能；反之，非確定性的有限狀態機則會有多於一種的可能，令 t 與t′為一非確定性的有限狀態機中的兩個轉移，其中 s[t]=s[t′]且 i[t]=i[t′]且 d[t]≠d[t′]，當狀態為 s[t]時，則 d[t]與 d[t′]皆為可能轉移至的狀態，非確定性的有限狀態機亦可解釋為，每個非確定性的有限狀態機皆能找到與其等價的確定性的有限狀態機。

本研究的有限狀態機中包含輸入字元ε，當空轉移出現時，即使沒有輸入字元也可以轉移至下一個狀態，即為非確定性的有限狀態機；當輸入一字串時，由於為非確定性的有限狀態機，只要有一個以上的狀態為終止狀態，則此結果為接受；反之，當未出現任何一個終止狀態，則結果為拒絕。

A:



Determinize of A:

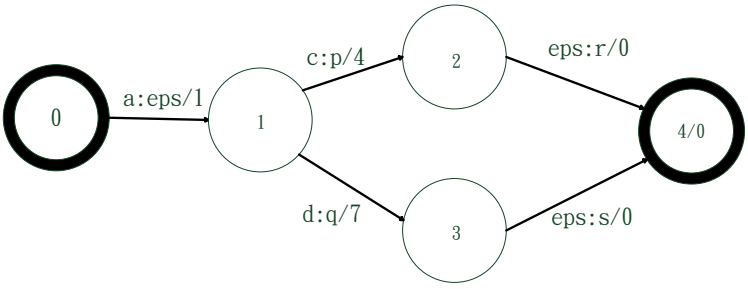
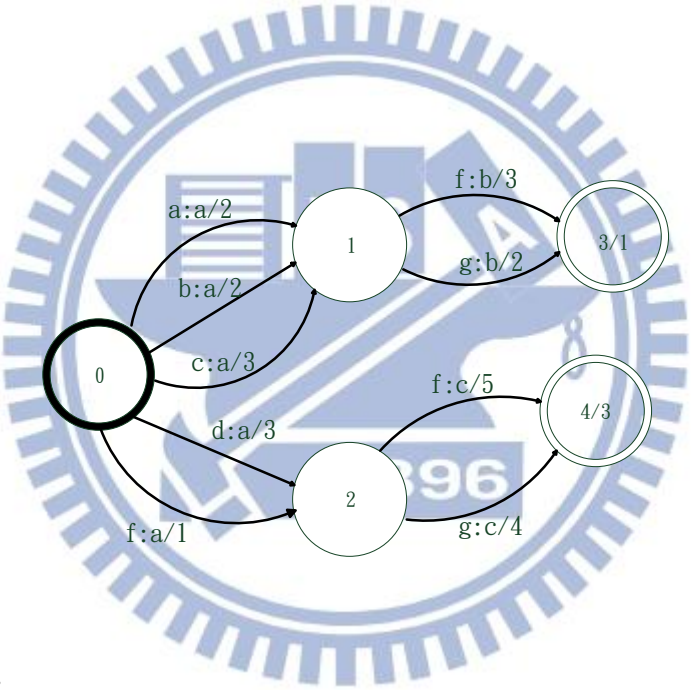


圖 3 3: 確定性的有限狀狀態機

3.1.4 最小化

當兩個有限狀態機為等效的情況，減少其中一個有限狀態機的狀態和轉換至無法再減少，此即為最小化的有限狀態機；最小化的操作只有當有限狀態機執行完確定性後才能執行，執行最小化有兩個步驟，首先，對有限狀態機的轉換進行 push weight 操作，將有限狀態機中成功路徑的權重移至前端，以減少狀態與轉移，接著，採用最小化演算法進行最小化的操作，將輸入、輸出與權重似為一個符號。

A:



Minimize of A:

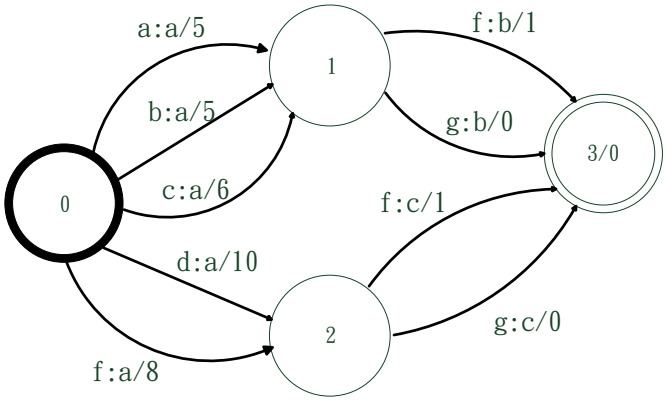
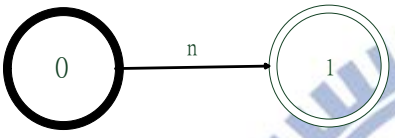


圖 3 4: 最小化的有限狀態機

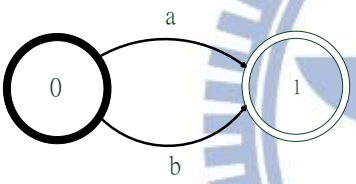
3.1.5 取代演算法

將一個有限狀態機中的轉移取代為另一個有限狀態機；令一個從狀態 s 到狀態 d 的轉移上的輸出字元為 n，欲將有限狀態機 F 取代該轉移，取代的方式為，先將此輸出字元 n 換為ε，接上這個有限狀態機 F，再把 F 的終止狀態接到原先的狀態 d。

A:



B:



B 取代 A 轉移上的 n:

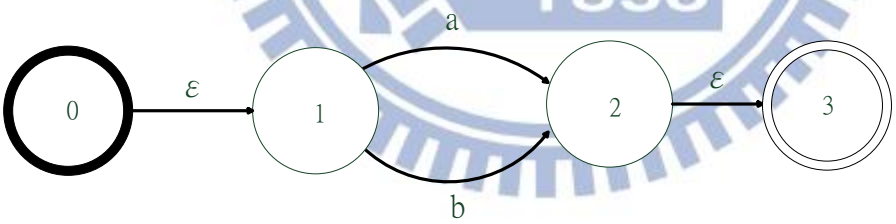


圖 3 5: 取代演算法

3.2 語音系統中的有限狀態機

這裡介紹的辨識系統與傳統的辨識系統在辨識的演算法上相同，差異在於此以有限狀態機來描述每個模型，語音模型同樣由三個模型所構成，分別為隱藏式馬可夫模型(Hidden Markov Models,HMMs)、發音詞典(Lexicon)與語言模型

(Language Model)；傳統大詞彙連續語音辨識中，分別對三個不同層級的語音模型各別處理，首先，在聲學模型中，利用隱藏式馬可夫模型來描述每個不同單元音的發音過程，再以每個單元音藉由發音詞典串接成不同詞彙的隱藏式馬可夫模型序列，並且使詞彙的首尾相接，形成一連續的隱藏式馬可夫模型的搜尋空間，最後再藉由語言模型引入文法觀念將詞與詞之間相接的情形以機率表示。

搜尋的演算法使用維特比光束搜尋(Viterbi beam search)來搜尋最佳路徑，由於大詞彙語音辨認的搜尋空間十分的龐大，提升了硬體實作的困難度，因此藉由維特比光束搜尋，以動態的方式來搜尋較為可能的部分，如此一來，使辨識系統更有效率，也增加了實作的可行性。

本實驗將聲學模型和語言模型的有限狀態機合併，藉此，只要在一個搜尋空間中找出一條最佳的路徑即可，不需要再分別對聲學模型與語言模型各別搜尋。

3.2.1 聲學模型

聲學模型為語音辨識系統中的第一層，其中，隱藏式馬可夫模型為語音辨識中最常被應用的模型，用來統計分析語音訊號，細節可以參考雷氏(L.Rabiner)著作，藉由目前和過去時間點的每個狀態來描述聲音的特性，以一連串的狀態轉換表示，狀態與狀態間有轉移機率，當建立有限狀態機時，將隱藏式馬可夫模型中的狀態機率以有限狀態機中的狀態轉移表示，如此一來，有限狀態機即以隱藏式馬可夫模型的狀態機率分布對應到聲學模型表示；圖 3.6 為一個描述聲學模型的有限狀態機，輸入的表示法為$a_2{}^+a_3{}^+a_4{}^+$，輸入為所有可接受的字串，輸出為 xin，輸出符號可在任何狀態之間，但必須在不同狀態的轉換之間，如此才是一個成功的路徑。

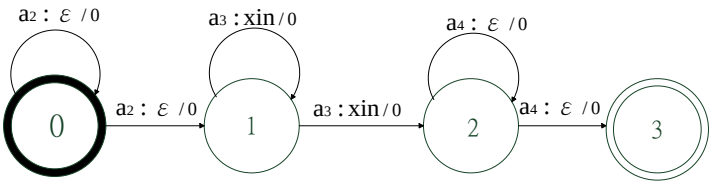


圖 3 6: 聲學模型的有限狀態機

各別的隱藏式馬可夫模型皆建立完成後，以聯集演算法將所有獨立的有限狀態機聯合成一個單一的有限狀態機，並在每個終止狀態新增加一條路徑，回到初始狀態，如此即完成聲學模型的有限狀態機。

Juicer 辨識器的聲學模型分數與語言模型分數為分開計算，並且 HMM 分數獨立計算之，因此 Juicer 聲學模型的有限狀態機以前後相關(context-dependent)的 WFST 表示，而非 HMM 模型。本研究採用與前後文無關的單音節模型，僅以單一狀態表示。

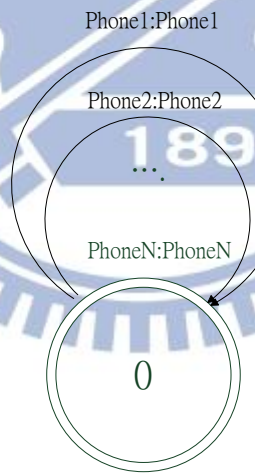


圖 3 7: 聲學模型的 WFST

3.2.2 發音詞典

發音詞典為聲學模型與語言模型之間的橋樑，將聲學模型輸出的聲音序列組合成有意義的詞彙，而每個詞彙對應至 HMM 序列。

我們藉由線性方式建立出線性詞典，線性詞典可以建立出每個詞彙的有限狀態機，詞彙的輸入為一串聲音序列，而此詞彙可以放置於此有限狀態機的任何轉移上的輸出，其餘的輸出為空轉移。

樹狀詞典，以樹狀架構來建立詞彙的有限狀態機，詞典中若有詞彙彼此字首相同，則可以共用同一字首結構，因此使詞典的空間縮小。

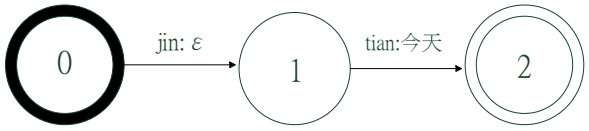


圖 38: 線性詞典有限狀態機

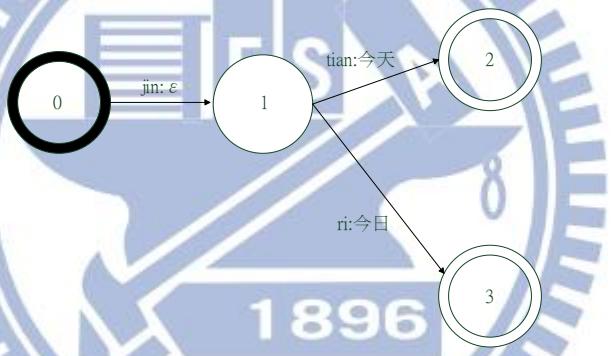


圖 39: 樹狀詞典有限狀態機

建立發音詞典的有限狀態機，方法與聲學模型的有限狀態機雷同，首先，建立各個詞彙的有限狀態機，初始狀態皆由空轉移聯集而成，聯集成一個有限狀態機，並且各個終止狀態連一條空轉移回至初始狀態。本研究的聲學模型並未加入聲調資訊，因此在面對音節相同的詞彙時，發音詞典會在序列的末端加上輔助符號 (auxiliary symbol) 來區別詞彙的差異，用以進行確定性演算法 (determinimization)。

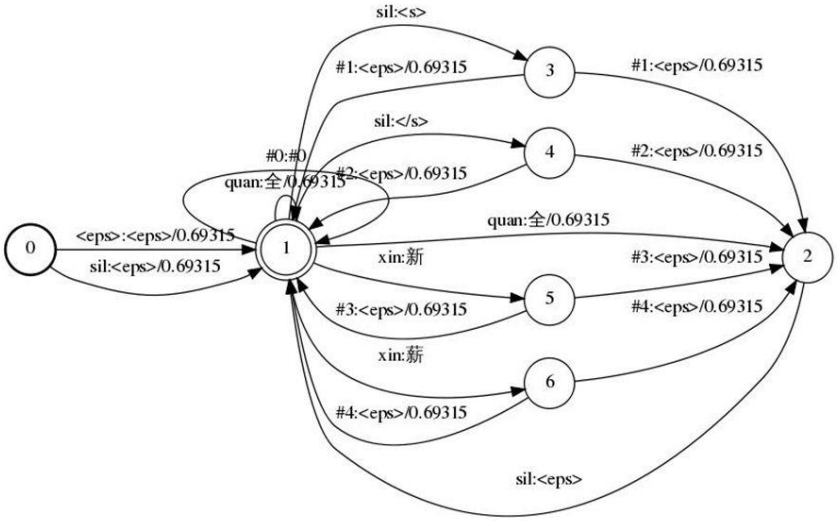


圖 3 10:發音詞典的範例

3.2.3 語言模型

本研究將 n-gram 語言模型的 ARPA 格式轉換為有限狀態機，藉由有限狀態機來表達詞與詞之間相連的機率，其中後撤平滑化可以以空轉移來描述。

WFST 補回同義詞的方式，本研究使用取代演算法，將欲取代的轉移由一個小 WFST 取代之，其中小 WFST 為欲取代與被取代的轉移，轉移上的字元彼此為同義詞，且加權值相同，即為這組同義詞享有相同的語言模型分數；細說取代的方式為，被取代的轉移由狀態 s 開始至狀態 d，欲用 WFST A 取代，狀態 s 會以空轉移連接至 A 的初始狀態，並且 F 的終止狀態會同樣以空轉移連接至狀態 d。

圖為一 bi-gram 的有限狀態機，其中狀態二的作用為，當沒有有效的輸入時，會藉由空轉移走至狀態二，並且附上後撤分數，而狀態二走至其他狀態所附上的分數，即為經由後撤後的 uni-gram 的 n-gram 分數；其餘的狀態，接有兩種轉移，一種是輸入為有效詞的轉移，另一種是沒有符合輸入詞的轉移，即為空轉移，因此藉由空轉移來做後撤，後撤會使詞串對應至非唯一的路徑，如此一來，不同的

路徑可能產生相同的輸出結果，其中未經過後撤路徑的機率會高於經過後撤路徑的機率。

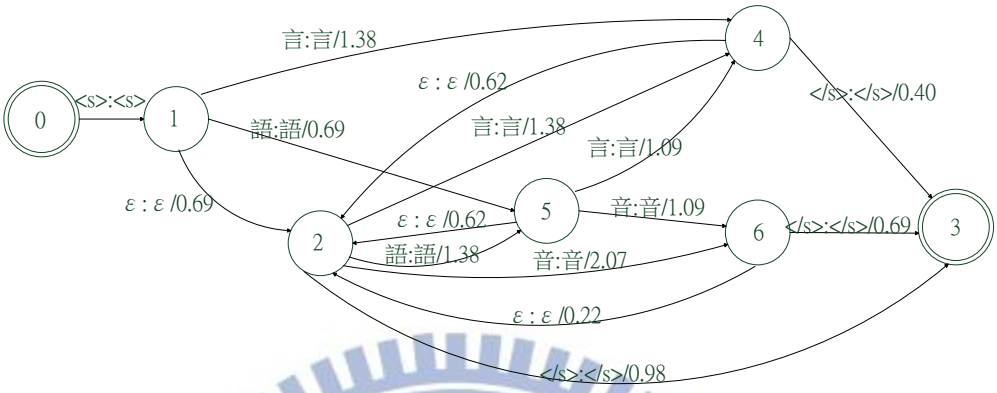


圖 3 11: 語言模型的有限狀態機

3.2.4 合併各層有限狀態機

介紹完各層有限狀態機後，我們將這三層的有限狀態機合併為一有限狀態機，並且將之整合為一有效率的有限狀態機；我們將聲學模型、發音詞典、語言模型轉換為有限狀態機並整合為一龐大的語音辨識系統模型，再經由確定化與最小化等演算法將辨識系統進行最佳化。

其中本研究使用的辨識器為 Idiap Research Institute 所開發的 Weighted Finite State Transducer Decoder – Juicer[15]。處理 WFST 相關演算法則使用 Google Research and NYU's Courant Institute 發展 的 OpenFst library[16]進行。

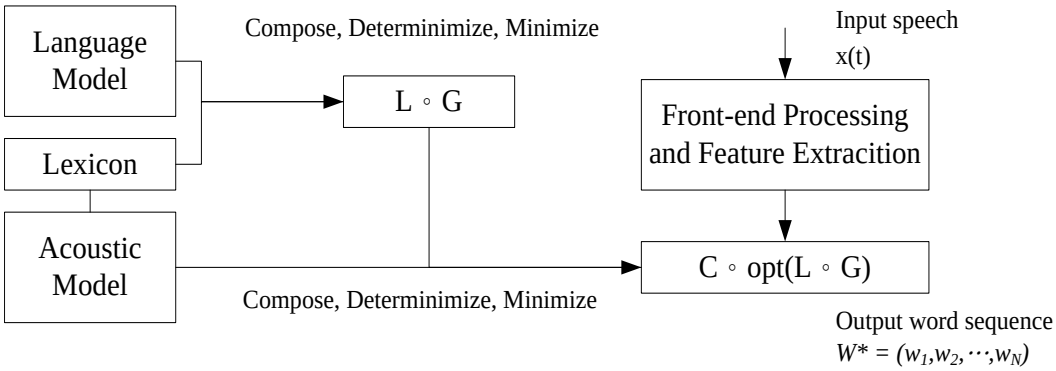


圖 3 12：語音辨識系統架構圖

第四章　加權有限狀態機之實驗分析

本章節分析加入維基百科後的語料，改變選詞方式後對語言模型的改善，並且以正確音節序列利用 WFST 轉換為詞序列，計算其辨識率以評判此語言模型的優劣，最後再比較 HTK 辨識系統與 WFST 辨識系統的差異，最後分析辨識率和辨識速度之影響要素，並且探討辨識率和辨識速度之關係。

4.1 文本前處理的分析

本節對語料加入維基百科分析，並且討論改變選詞方式後對語言模型的改善。

4.1.1 語料庫加入維基百科

隨著科技的興盛，電腦功能快速成長，採集廣泛的語料，訓練大規模的語言模型，成了語音辨識系統追求的目標，語料量的增加，可以使訓練出的語言模型更加穩定且可信度更高。

維基百科語料普遍被眾人所信任，語料的內容領域廣泛，且持續的在更新，為一個高品質的語料。

維基百科語料經處理後，總詞數約為六千萬，使得目前語料庫總詞數由 3.8 億增至 4.4 億，詞條數亦由一百七十萬增至兩百九十萬，文章的篇數為 8771 篇，語料庫的增大，使得經由大量文字語料訓練而成的語言模型涵蓋範圍更為廣泛。

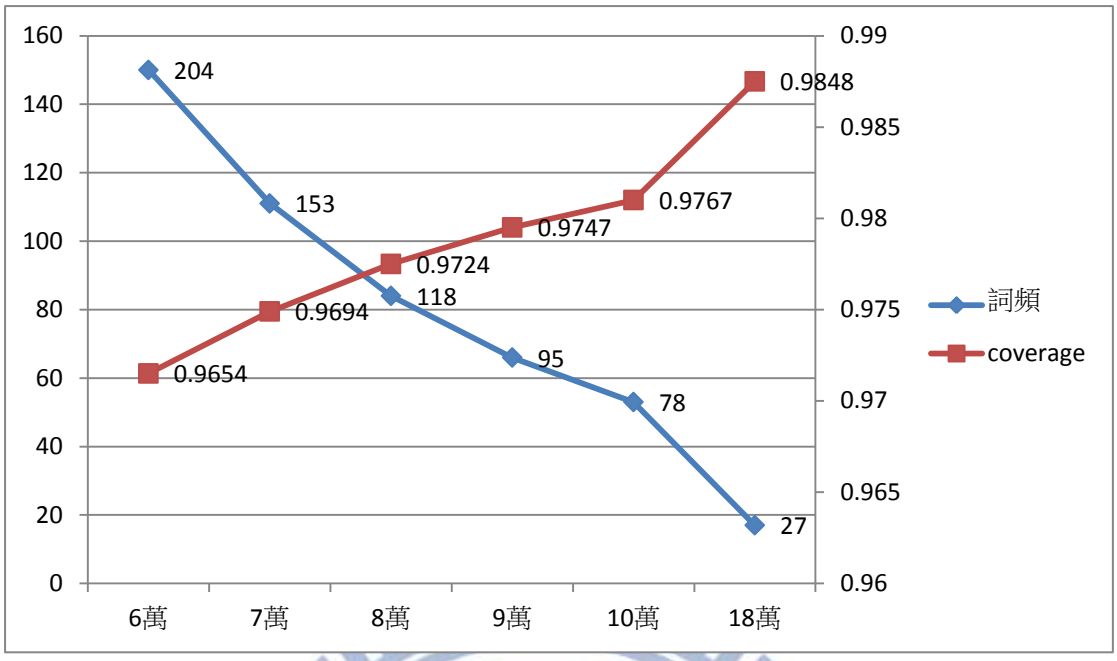


圖 4.1：各個詞數的詞頻與涵蓋率的比較

目前詞典選出第六萬詞的詞頻為 149，涵蓋率為 96.53%，若語料僅維基百科的涵蓋率為 91.99%，由此可知，維基百科語料和光華雜誌、NTCIR、中研院平衡語料庫、Chinese Gigaword 內容領域方向有些許不同，且證實了維基百科語料的領域多元；另外，混淆度(Perplexity，ppl)的分析，選詞後，tri-gram 的混淆度為 189.736，相對於語料庫未加入維基百科的混淆度為 176.127，混淆度上升的原因為，計算混淆度時，不在詞典中的詞(OOV)會被忽略，因此，隨著維基百科加入語料庫，OOV 增加，導致混淆度也提升，下表為加入維基百科語料的混淆度(A)與語料僅有維基百科的混淆度(B)比較。

表 4.1：混淆度的比較(對 inside test)

語料	order	ppl	ppl1
A	3	189.736	233.092
B	3	195.432	267.461

4.1.2 依詞性分類選詞法

以往選詞方式為 IDF，依據詞彙在語料庫中出現的文章篇數來判定詞彙的重要性，此方式雖然比起傳統直接收錄高詞頻的詞彙作為詞典來的優異，但並未考慮至詞彙的詞頻，因此選詞方式仍然不夠嚴謹。

中文詞可以分為實詞與虛詞，實詞數量為無限的，虛詞數量為有限的，在詞典數量有限的情況，我們應慎選收錄進詞典中的實詞，使詞典更有效率，因此先將詞彙依實詞與虛詞做分類，為了對這兩類以不同的標準選詞，選詞方式除了過去使用的 IDF 選詞法，這裡新增了一個更為嚴謹的選詞方式，算出各個詞彙在語料庫中應出現的文章篇數，和實際出現的文章篇數做比例，我們可以按照此比例移除在語料庫中分布不均衡的詞彙。

本研究將處理後的詞彙依照詞性地方詞(Nc)、非謂形容詞(A)、連接詞(Cab)、連接詞(Cbb)、副詞(D)、位置詞(Ncd)、數詞定詞(Neu)、狀態不及物動詞(VH)、動作使動動詞(VAC)、動作及物動詞(VC)、動作前程度副詞(Dfa)、時態標記(Di)、普通名詞(Na)、專有名詞(Nb)、指代定詞(Nep)、後置數量定詞(Neqb)、感嘆詞(I)、語助詞(T)、雙賓動詞(VD)、動作謂賓動詞(VF)狀態類及物動詞(VI)、狀態句賓動詞(VK)、SHI、Nv、後置詞(Ng)、V_2 做分類，再依詞性屬性分群，分為開放類和封閉類兩群，開放類詞性的詞彙即提高選詞標準，降低收錄進詞典中的數量，封閉類詞性的詞彙則降低標準，提高收錄進詞典中的數量。

接著選詞，先以 IDF 刪除出現文章篇數低的詞彙，這階段僅以各詞彙出現文章的篇數做考量，下表為 IDF 法刪除的詞彙。

表 4.2: 由 IDF 移除的詞

詞彙	詞頻	篇數
艾賽克斯號	417	23
利奇馬	394	35
海德格爾	388	35
帕希佐	382	112
庫德民主黨	377	104
紅火蟻	372	49
潮下帶	360	13
土衛	336	33
劍魚座	323	22

由實驗結果可以觀察出，IDF 選詞法雖然可以移除僅出現在極端少數文章中的詞彙，但由於只考慮到出現文章數，詞頻並未列入考量，因此在選詞時會受到限制，例如，塔利班，該詞的詞頻為 5638，出現文章數為 649，從詞頻與出現文章數的落差可以視該詞為只出現在特定領域的詞彙，若以 IDF 移除該詞，會使得例如駛進(詞頻為 727，出現文章數為 601)、前行(詞頻為 801，出現文章數為 640)等等較為泛用的詞彙連帶被移除，因此增加一選詞方式，針對出現文章篇數非極端少數但分布不平均的詞彙，詞頻與出現文章數皆為考量，算出各個詞彙的應出現文章數，再由應出現文章數與實際出現文章數做比較，此階段的選詞法更為嚴謹，實驗結果可以看出，這階段的選詞法確實可以移除分布較不平均的詞彙。

表 4.3: 新增選詞法移除的詞彙

詞彙	詞頻	實際文章數	應出現文章數	Ratio
馬其頓	14041	1428	7001	0.2040
車臣	11076	1273	6290	0.2024
親民黨	11610	1318	6436	0.2048
引種	10711	135	6184	0.0218
賽季	9683	704	5862	0.1201
東帝汶	8707	1031	5520	0.1868
球會	5993	733	4341	0.1688
塔利班	5638	649	4159	0.1560
程泉	4368	254	3440	0.0738

由於辨識結果只會輸出有收錄在詞典中的詞彙，因此希望將所有存在的詞彙皆收錄進詞典中，但受限於記憶體的大小，限制開放類詞性的詞彙，其中以普通名詞(Na)、專有名詞(Nb)、地方詞(Nc)移除的數目最多，這些詞性的詞彙容易隨著時間、地點等而變化；選詞後，六萬詞內更動的詞共有 3980 個，第六萬詞的詞頻為 149。

表 4.4: 六萬詞內各詞性詞彙更動的個數

詞性	IDF	新增選詞	詞性	IDF	新增選詞
Nc	869	87	A	11	.0
Cab	0	0	Cbb	2	0
D	15	1	Ncd	0	4
Neu	1	0	VH	27	6
VAC	0	0	VC	68	0
Dfa	1	0	Di	0	0
Na	1149	6	Nb	3730	0
Nep	0	0	Neqb	0	0
I	0	0	T	0	0
VD	0	0	VF	0	0
VI	0	0	VK	0	0
SHI	0	0	Nv	0	0
Ng	0	0	V_2	0	0

下圖為選詞後，各詞類詞典的涵蓋率，其中 Nb 的涵蓋率最低，詞性 Nb 中許多詞彙為人名，詞典內收錄人名對辨識系統幫助不大，因此我們僅讓 Nb 的涵蓋率為 57.35%。

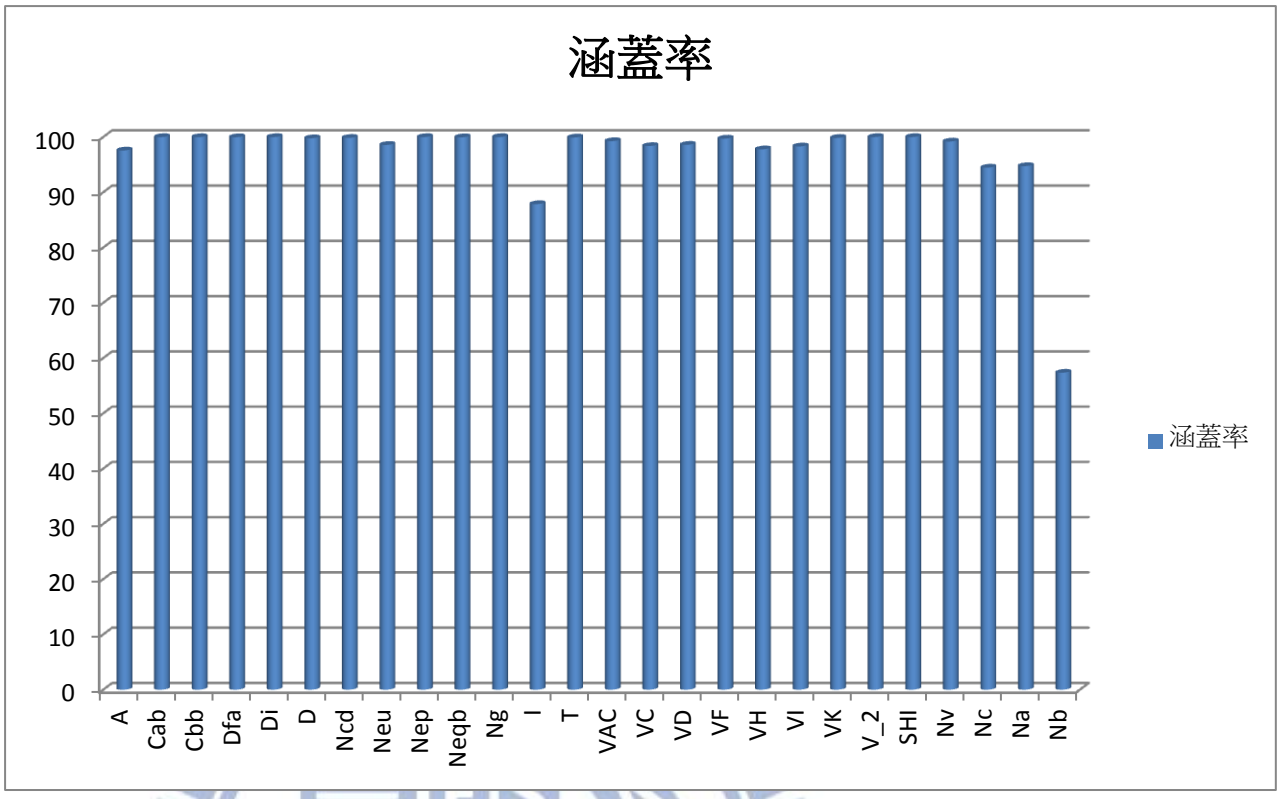


圖 4.2: 詞典中各詞性的涵蓋率

經由上述方法選詞後，選出六萬詞的詞典在訓練語料的涵蓋率為 96.36%，在測試語料的涵蓋率為 97.23%，平均詞長為 2.36 個字，下表為選詞前後對訓練語料計算混淆度。

表 4.5: 選詞前的混淆度(對 inside test)

	order	PPL	PPL1
選詞前	3	189.373	232.54
選詞後	3	189.736	233.092

4.1.3 語言模型 PPL 的比較

不同的 cutoff 與 discounting 會影響語言模型的效能，cutoff 表示詞頻低於該

cutoff 次數的詞彙序列改由出現頻率較高的短詞彙所取代，cutoff 設定的越高，語言模型越小，辨識效能較差，但辨識所需要的時間較短，下圖為 tri-gram cutoff 比較。

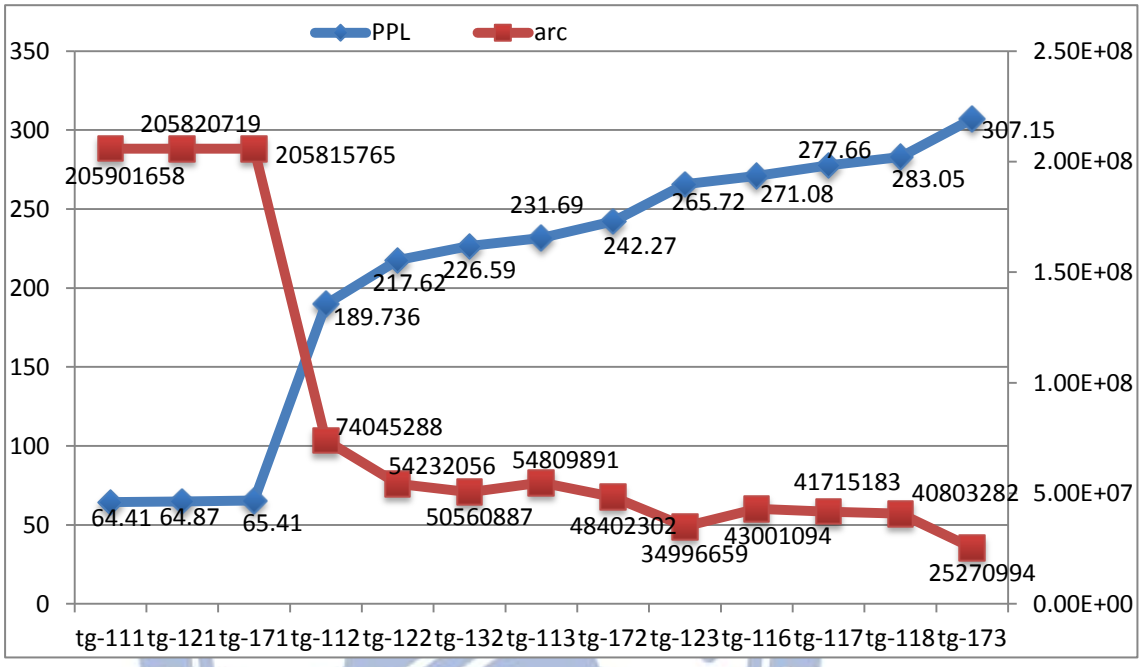


圖 4.3: 不同 cutoff 對 inside test，PPL 與 arc 數的比較

下表為語言模型在同樣的 cutoff 時，uni-gram bi-gram tri-gram 的 cutoff 分別 112，不同 discount 的比較。

表 4.6: 不同 discount 對 inside test 算 PPL

discount	PPL	PPL1
Good-turing	189.736	233.092
Ristad's Natural	188.522	231.543
Witten-bell	199.528	245.607
Abolute	190.811	234.466
Kneser-ney	212.109	261.721

4.2 音節的解碼來評估語言模型

以往評估語言模型，通常是以計算 PPL(perplexity)來判斷，但實際上 PPL 為語言模型估測一個歷史詞串後面平均可能的可接詞數，語言模型的好壞與 PPL 無直接性的關係，由此可知使用 PPL 評估語言模型，並非一個最好方法，因此本研究藉由 OpenFst 系統發展了一套直接性評估語言模型的方法。

正規的 OpenFst 語音辨識流程為，輸入語音訊號的 MFCC 參數，藉由聲學模型的加權有限狀態機做第一層級的語音辨識，接著將辨識結果的音節序列經由發音詞典的加權有限狀態機對應至有意義的詞彙序列，最後再藉由語言模型的加權有限狀態機加入了語言文法做最後一層級的語言辨識，得到了最後的辨識結果，但如果僅考慮語言模型的話，正規的語音辨識系統多了聲學模型辨識因素，因此，在此將聲學模型的加權有限狀態機移除，剩下發音詞典與語言模型兩個層級，由於辨識流程未通過聲學模型的辨識，輸入訊號更改為以正確音節序列為輸入，經由發音詞典的加權有限狀態機對應至有意義的詞彙序列，再藉由要評判的語言模型加權有限狀態機做辨認，得到最後的詞彙序列，接著以最終辨識的詞彙序列與正確解答的詞彙序列做比較，算出詞彙的錯誤率，因此，可以藉由不同的語言模型，算出各別的詞彙錯誤率，當詞彙錯誤率較低者，表示訓練出為較佳的語言模型，另外，還可以藉由本研究的辨識結果，分別看出聲學模型與語言模型的改善空間；由上可知，此研究的輸入為正確音節序列，得到的輸出為辨識後的詞彙序列，研究概念與注音輸入法雷同，注音輸入法為一種以注音符號來輸入漢字的中文輸入法。

本研究處理加權有限狀態機的相關演算法採用 OpenFst 進行，首先，正確音節序列依照 L。G 的輸入表編輯成加權有限狀態機的格式，並且檔案轉換為二進位的檔案格式。

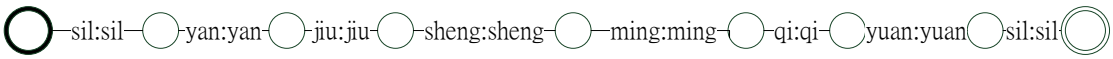


圖 4.4: 【研究生命起源】的正確音節序列作範例

同時準備先前訓練的發音詞典與語言模型結合的加權有限狀態機(以 L。G 表示)，使用 OpenFst 指令中的 fstcompose 將正確音節序列與 L。G 合併，合併後涵蓋了辨識後所有可能性的詞彙序列。

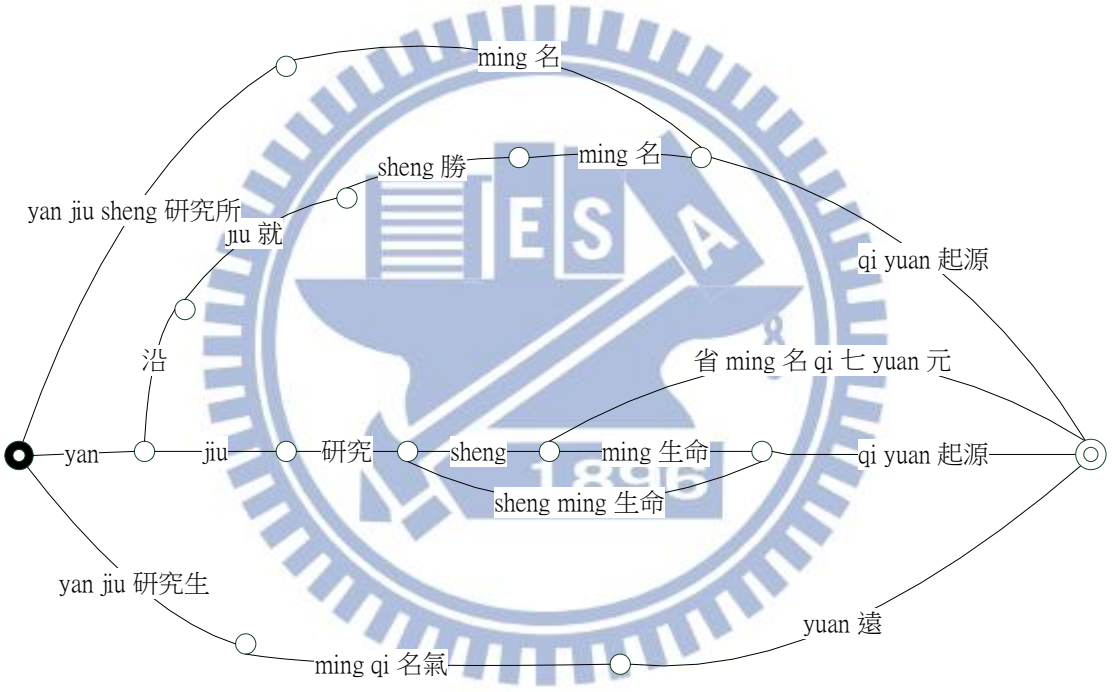


圖 4.5: 【研究生命起源】的正確音節為輸入，和 L。G 做合併的範例圖

我們以 OpenFst 指令中的 fstshortestpath 找出最佳的辨識結果，n-shortestpath 以路徑分數最低的前 n 條路徑做為辨識最佳的前 n 個結果，這裡以 n=1 選出一條最佳的路徑，由於，fstshortestpath 指令限制轉換上權重的型態為 non-log，本研究轉換上權重的型態皆以 log 存在，因此，需先將檔案的檔頭資訊轉換為 non-log 型態再執行 fstshortestpath 指令，完成 fstshortestpath 指令得到的輸出即為最佳路徑。

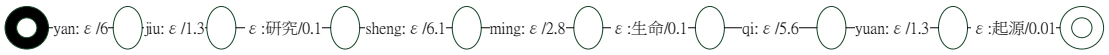


圖 4.6: 【研究生命起源】最佳路徑的範例圖

目前辨識系統 tri-gram 的辨認率為 87.34%，本研究實驗結果的辨識率為 91.04%，這說明了，聲學模型約有 3.7%的改善空間，而語言模型約有 8.96%的改善幅度，由此觀察出，語言模型可以改進的幅度大於聲學模型，並且有明確數據可以準確的得知各個模型改善的幅度。

下圖為各個語言模型所建構出的語音辨識系統，藉由本節的方式算出的評估值，可以從圖中證實，tg172、tg182、tg1112 的值相近，但明顯高於 bg12、bg17、tg173，證實了評估值可以評斷語言模型的好壞。

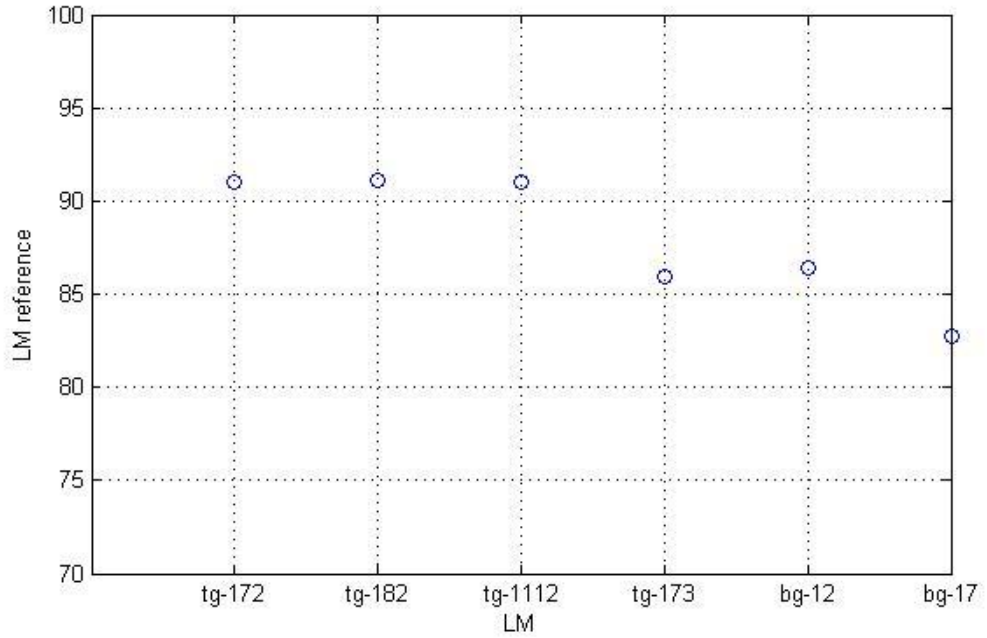


圖 4.7: 各個辨識系統的語言模型評估值

4.3 HTK 辨識分析

HTK[17]辨識為，首先，產生 HTK Word Network 的語言模型，由於一般語言模型皆由 SRILM 訓練而成，因此將 SRILM 訓練出 ARPA 格式的 bi-gram 轉換為 HTK 的 Word Network，原本這裡要將在訓練語言模型時即合併的同義詞使用 sub-lattice 展開，但受限於記憶體的關係，只能對又讀詞展開；有了 bi-gram Word Network 後，接著做 Forced Alignment，藉由詞典、音節表、語言模型和聲學模型對聲學參數用 Viterbi search 做辨識，產生 bi-gram 的 lattice；最後再做 rescore，由前面產生的 bi-gram lattice 藉由 tri-gram 語言模型重新調整 lattice 上的語言模型權重產生 tri-gram lattice，其中 rescore 的速度比一般辨識來的快；若要做 4-gram 辨識，則使用工具 SRILM，藉由 4-gram 語言模型對 tri-gram lattice 做 rescore 產生 4-gram lattice，HTK 只能辨識至 tri-gram。

下面分別為 bi-gram、tri-gram、4-gram 的 PPL、字辨識率與詞辨識率。

表 4.7: 不同語言模型對 inside test 算 PPL 的比較

LM	PPL	PPL1
Bi-gram	417.647	529.213
Tri-gram	242.248	300.471
4-gram	13.889	15.399

表 4.8: 不同語言模型對 outside test 算 PPL 的比較

LM	PPL	PPL1
Bi-gram	538.941	592.849
Tri-gram	64.209	68.390
4-gram	6.884	7.088

表 4.9: 不同語言模型的詞辨識率與字辨識率

LM	詞辨識率	字辨識率	weight
Bi-gram	72.62%	80.00%	lm:17 penalty:-12
Tri-gram	87.34%	90.31%	lm:19 penalty:-11
4-gram	90.79%	92.50%	lm:20 penalty:-15

實驗結果可以看出，隨著 n-gram 的 n 提升，辨識率也提升，其中 bi-gram 至 tri-gram 的提升幅度最大，4-gram 辨識率也有明顯的成長，辨識時還需要調整 lm 與 penalty 的權重，lm 的權重越大，表示此語言模型越可以被信任，語言模型越好，由上表可以證實，以 4-gram 的 lm 權重最大，另外，PPL 也會因為 n 的上升而下降，尤其以 4-gram 對 outside test 的 PPL 僅剩 6.884。

目前 TCC300 測試語料共有 15488 個詞，OOV 詞數有 446 個，辨識過程中，OOV 詞會影響附近的詞彙，這裡算出辨識中一個 OOV 平均造成約 2.25 個詞辨識錯誤。

4.4 WFST 辨識分析

HTK 辨識系統中，詞典的大小會受到限制，無法加入同義詞，也無法處理大詞彙語音辨識，且辨識過程需要花很長的時間，相較於 HTK，WFST 詞典沒有限制，辨識時間也快許多。

本研究使用的辨識器是由 Idiap research institute 發展的辨識器 Juicer，Juicer 使用有限狀態機(Weighted Finite State Transducer)來表示搜尋空間針對大詞彙語音做辨識，將語言模型層、發音詞典層與聲學模型層合為一層，再做最佳化，即完成搜尋空間，辨識時，以 Viterbi Search 找出一條最佳路徑。

WFST 的好處為可以經由最佳化使搜尋空間縮小，減少辨識所需的時間，另外，WFST 皆以相同表達方式表示各個層級，各個模型皆以狀態和轉移表示，方便系統以標準化的方式處理。

WFST 的缺點為，訓練語言模型時，語料中未出現的詞彙會以後撤平滑化(back-off smoothing)來調整語言模型，而語言模型以有限狀態機表示時，將後撤平滑化以空轉移(ε)表示，語言模型的有限狀態機與發音詞典的有限狀態機做組合演算法時，空轉移會使有限狀態機複雜化且增大有限狀態機的大小，因此在有限狀態機做組合演算法與最佳化時，時常因為記憶體不足的原因而無法順利進行，在做辨識時，由於需要運算龐大的有限狀態機，因此也需要相當可觀的記憶體；另外，Juicer 在處裡 n-gram 語言模型的有限狀態機時，當 n 超過 3 時，容易發生問題。

當本研究語言模型選擇 tri-gram 且 cutoff 中 gt1min、gt2min、gt3min 分別為 1、1、2 時，對 L。G 執行確定化(determinimize)會出現記憶體不足的問題，因此改用 gt1min、gt2min、gt3min 分別為 1、7、2 的語言模型，雖然辨識結果會差一些，但辨識所需的時間會較少。

表 4.10: 兩個語言模型的比較

LM	PPL	arc 數
tg_112	189.736	74051225
tg_172	242.248	48409754

將語言模型、發音詞典與聲學模型合併和最佳化後，最終的有限狀態機的資訊如下:

表 4.11: 最終之有限狀態機之狀態與轉移數

狀態數	48,440,760
轉移數	97,579,638

WFST 和 HTK 辨識出相同效能的情形下，WFST 的辨識時間可以比 HTK 辨識速度快至 20 倍左右。

4.5 辨識率與辨識時間的分析

研究如何提升辨識率後，接下來我們希望能讓語音辨識系統能即時且準確，使語音辨識系統達到實用性。

本研究實驗語音辨識系統的影響因素包含:語言模型的 cutoff、n-gram 語言模型的 n、perplexity 語言模型的評估值、hypotheses，其中 hypotheses 為辨認器中每個音框傳遞時所保留的 beam width；由前面的探討，可以先得知，n-gram 語言模型的 n 越大，語言模型越複雜，語言模型的 cutoff 越高，語言模型的 n-gram 詞彙越少。

4.5.1 有限狀態機大小的調整

加權有限狀態機的語音辨識系統由轉換、狀態與加權值所組成，由 Zhijian OU[12]可以得知，轉換與狀態數會決定加權有限狀態機所需要記憶體的多寡，這裡利用調整 gt1min、gt2min 與 gt3min 的方式來改變加權有限狀態機的轉移數和狀態數。下圖為各個加權有限狀態基的轉移數目，圖中可以觀察出 bg17 加權有限狀態機的轉移數最少，最多的為 tg172，轉移數越多，表示語音辨識系統的 search space 越大。

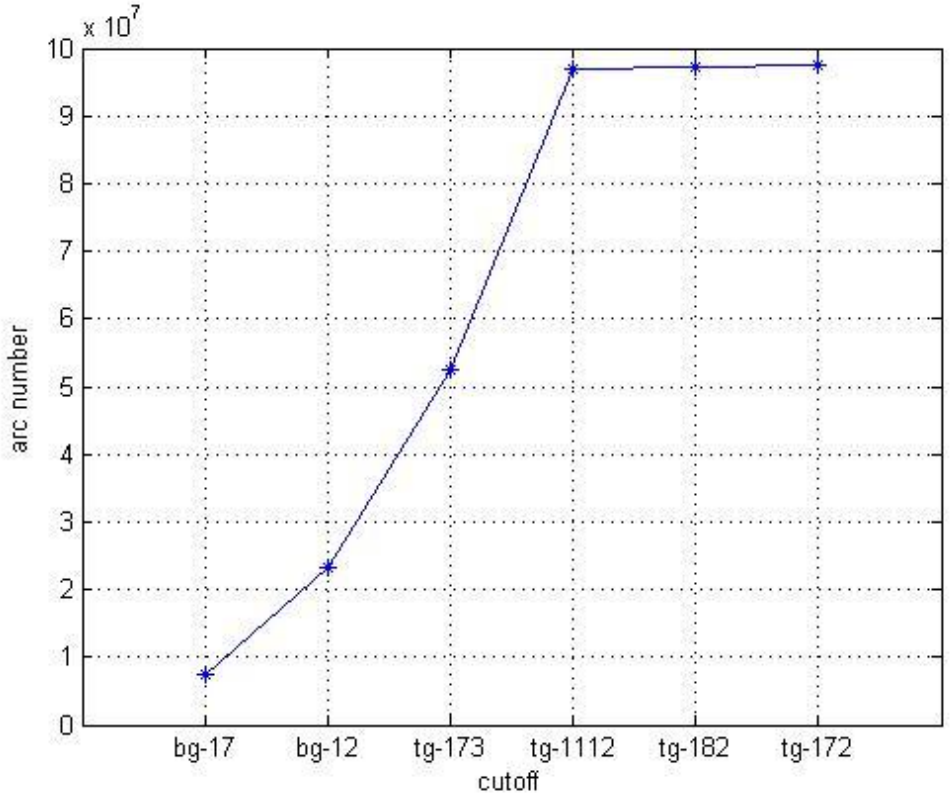


圖 4.8: 各個加權有限狀態基的轉移數

4.5.2 加權有限狀態機大小與辨識率的關係

接著以辨識系統的 cutoff 和 maximum hypotheses 兩個參數，觀察辨識率、語言模型估計值和 perplexity 的關係，加權有限狀態機做辨識時，會給每個音框中的 hypotheses 一個上限，稱為 maximum hypotheses，hypotheses 保留的數目會影響辨識時 beam search 的快慢與辨識結果。

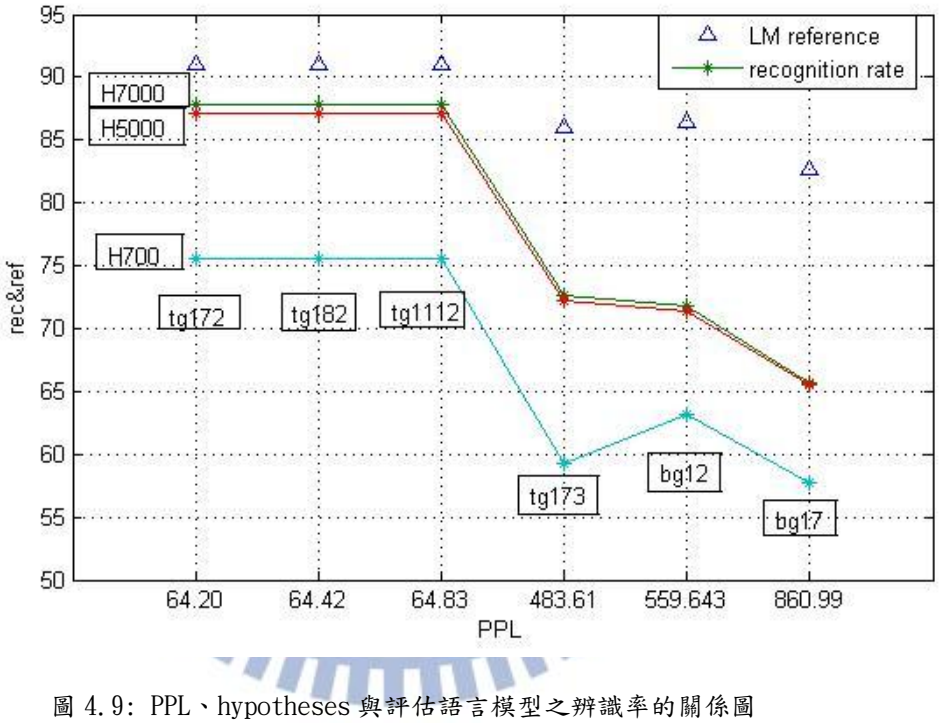


圖 4.9: PPL、hypotheses 與評估語言模型之辨識率的關係圖

由上圖可以觀察出，當 PPL 越低時，辨識效能越好；評估語言模型的辨識率也是以 tg172、tg182、tg1112 高於 tg173、bg12 與 bg17，前者的轉移數與語言模型相近但優於後者許多，證實了計算評估語言模型的辨識率確實可以做為評估一個語音辨識系統中的語言模型好壞，且在同樣的 maximum hypotheses 下，辨識率與評估語言模型值之間的間距，前者的值小於後者，可以評斷原因為前者的語言模型較好，使得辨識率更接近上限；由圖中可以發現，當 maximum hypotheses

為 5000 與 7000 時，辨識率高出 maximum hypotheses 為 700 的許多，且 maximum hypotheses 5000 與 7000 的辨識率十分的相近，接下來探討這兩個值得辨識速度來決定何者辨識系統較佳。

4.5.3 加權有限狀態機辨識率與速度的關係

首先說明本研究之工作環境，作業系統為 Linux 作業系統，使用之 CPU 型號為 Intel(R) Xeon(R) CPU x5650 @ 2.67GHz。

影響辨識速度的因素有 maximum hypotheses 與加權有限狀態機的大小，這裡以不同 cutoff 的加權有限狀態機分別對 maximum hypotheses 5000 與 7000 做辨識率與速度的關係圖。

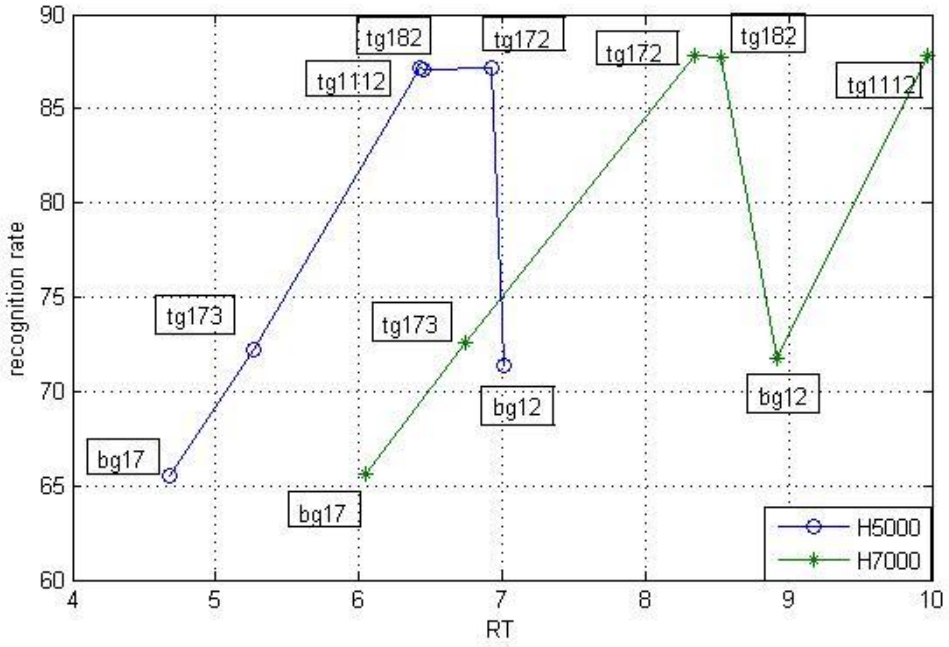


圖 4.10: hypotheses 5000 與 7000 之辨識率與速度的關係圖

由圖中可知，maximum hypotheses 5000 與 7000 之辨識率差不多，但 hypotheses 5000 速度快了許多，因此本研究的語音辨識系統 maximum

hypotheses 設定為 5000 較佳。

另外，看針對單一語音模型辨識系統，調整其 maximum hypotheses，可以更清楚的看出辨識率與速度之間的關係。

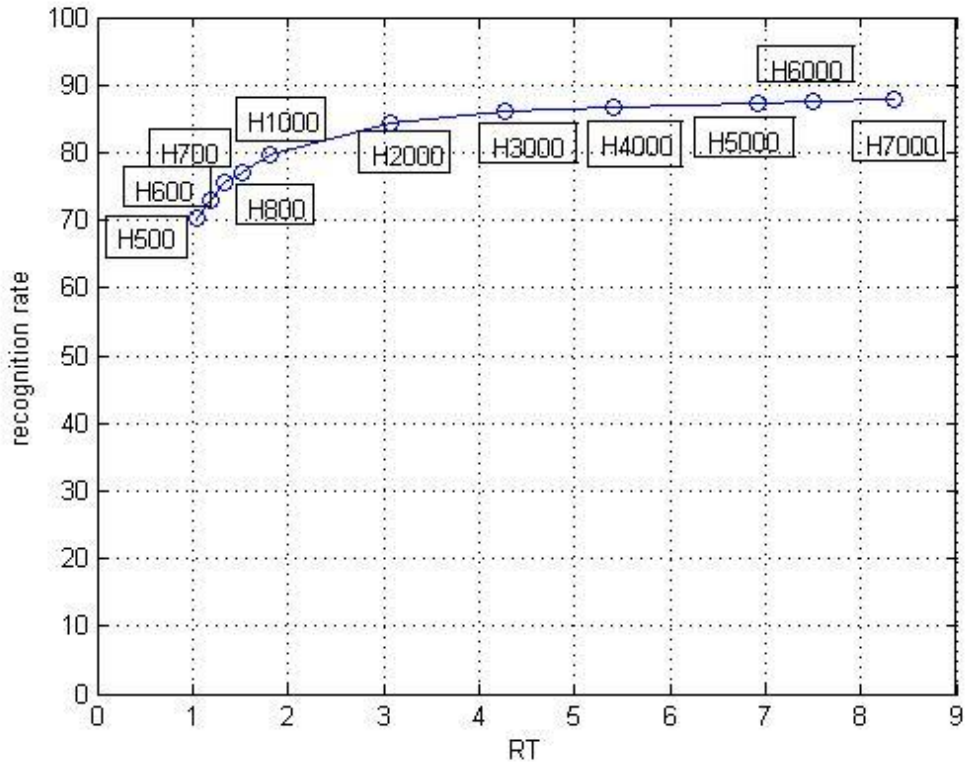


圖 4.11: tg172 的辨識率與速度之關係圖

從圖中可知，每個音框中的 hypotheses 保留的越多，辨識時，beam search 的時間越長，導致辨識速度慢，但辨識率高；反之，亦然。

由實驗可以得知，辨識率和辨識速度為 trade off 的關係，兩者不可兼得。

4.5.4 測試語料的切短

目前的測試語料有 15488 個詞彙，分為 226 個句子，平均一句約有 68 個詞彙，但實際在語音辨識時，測試語料中並不會出現如此長的句子，因此本節研究

將測試語料中之句子調整貼近實際測試語料之長度，並且觀察句子切短後辨識率與辨識速度之關係。

本研究採用測試語料中之十個句子，總詞彙為 710，將之切短成 29 個句子，平均一句約有 24 個詞彙，並且尋找和上一節 maximum hypotheses 為 5000 時相近之辨識率，藉以比較句子切短前後之速度與保留 hypotheses 的變化。

下圖分別為句子切短前後測試語料的辨識結果，在辨識率為相近的情形下，句子切短後的 maximum hypotheses 僅需 710 即可辨識率相近的效果，此說明了測試語料句子越長，辨識時所需保留的 hypotheses 越多；由於句子短的測試語料之 hypotheses 較低，因此辨識時所需花費的時間較少。

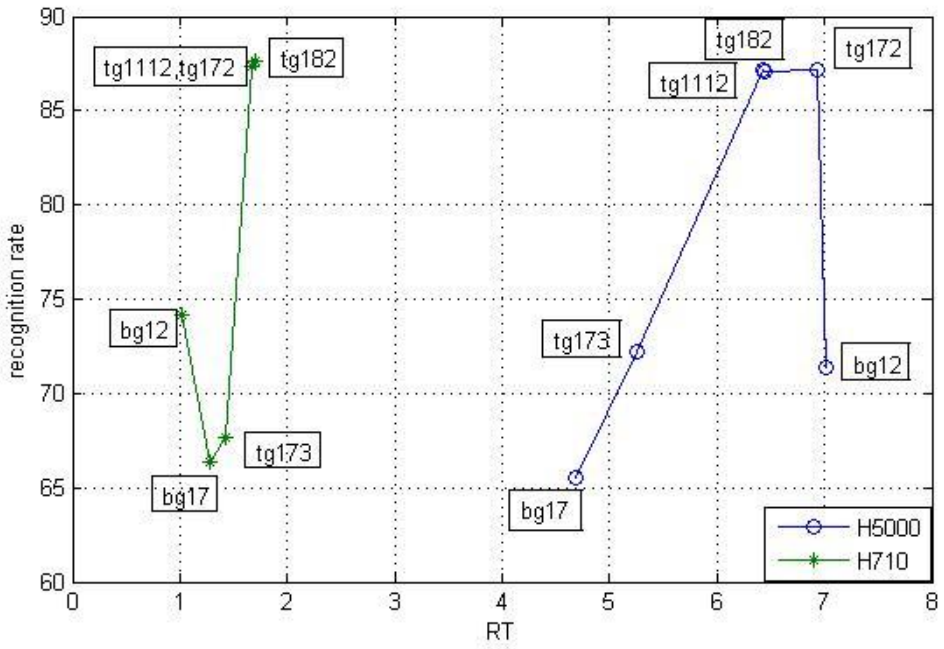


圖 4. 12：句子切短前後之辨識率與速度的關係圖

第五章　結論與未來展望

本研究主要針對語音辨識系統中之語言模型做改善，語料的整理並改變選詞方式，有效的提升辨識率，再藉由音節解碼來評估語言模型，使評估語言模型的方式更可信；本研究採用加權有限狀態機建立語音辨識系統，由實驗結果，了解各個影響辨識率與辨識速度之因素與彼此之間的關係，並加以調整，找出最適宜之中文連續語音辨識系統。

訓練語料中，尚有許多未被整理的異體字與同義字，由本研究可以看出，這些詞彙經文字正規化後，確實能有效的改善語音辨識系統的效能，增加辨識的準確性。未來可以在加權有限狀態機之 word lattice 加入韻律資訊，包含韻律邊界停頓與音節韻律狀態，使其帶有語速訊息，並藉由加權有限狀態機之最佳化縮小 word lattice 的大小，加快辨識速度。另外，未來可以先藉由 bigram 並且增加辭典的大小產生加權有限狀態機之 word lattice，再對 trigram 做 rescoring，不僅能提升辨識速度且提升辨識率，使語音辨識系統辨識時間能更接近 real time。

參考文獻

【1】 Mehryar Mohri, “Finite-State Transducers in Language and Speech Processing,” AT&T Labs – Research, 1997

【2】 M. Mohri, F. Pereira, M. Riley, “Weighted finite-state transducers in speech recognition,” Proc. of ASR2000, pp. 97–106, 2000.

【3】 Mohri, M., Pereira, F., Riley, M.l.: Weighted finite-state transducers in speech recognition. Computer Speech and Language 16(1), 69–88 (2002)

【4】 Mohri, Mehryar, and Michael Riley. "Weighted determinization and minimization for large vocabulary speech recognition." *Eurospeech*. 1997.

【5】 Mehryar Mohri, Michael Riley, “A Weight Pushing Algorithm for Large Vocabulary Speech Recognition,” AT&T Labs – Research

【6】 Chia-Hsing Yu, “Large Vocabulary Continuous Mandarin Speech Recognition Using Finite-State Machine,“ NTU Digital Speech Signal Processing Lab, 2004

【7】 Shang-Yao Chang, “Large Vocabulary Continuous Mandarin Speech Recognition Using Finite-State Machine,” NCTU Speech Processing Lab, 2008

【8】 Daniel Jurafsky and James H. Martin, ”SPEECH and LANGUAGE PROCESSING,”2008

【9】 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*

【10】 Chien-Pang Chou, “Improvement on Language Modeling for Large-Vocabulary Mandarin Speech Recognition,” NCTU Speech Processing Lab, 2009

【11】 Decadt, Bart, and Walter Daelemans. "Phoneme-to-grapheme conversion for out-of-vocabulary words in speech recognition." (2001).

【12】 Ou, Zhijian, and Ji Xiao. "A study of large vocabulary speech recognition decoding using finite-state graphs." *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*. IEEE, 2010.

【13】 J. Lafferty, A. McCallum, F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," In proc. ICML01, 2001

【14】 Stolcke, Andreas. "SRILM-an extensible language modeling toolkit."*INTERSPEECH*. 2002

【15】 D. Moore, J. Dines, M. Magimai Doss, J. Vepa, O. Cheng, and T. Hain, "Juicer: A weighted finite state transducer speech decoder," in Proc. MLMI (to appear), Washington DC, May 2006.

【16】 C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri. OpenFst: A general and efficient weighted finite-state transducer library. In Proceedings of the 12th International Conference on Implementation and Application of Automata (CIAA 2007), Prague, Czech Republic, July 2007, volume 4783 of Lecture Notes in Computer Science, pages 11–23. Springer, Heidelberg, 2007

【17】 Young, Steve, et al. "The HTK book." *Cambridge University Engineering Department* 3 (2002): 175

附錄一:實驗所用 Variant Word Pair 表

Variant Word Pair		Variant Word Pair	
愈來愈	越來越	鄭重其事	慎重其事
巴金森氏症	帕金森氏症	鼎鼎有名	赫赫有名
無人問津	乏人問津	除此以外	除此之外
身臨其境	身歷其境	勢所難免	在所難免
非同尋常	非比尋常	努力以赴	全力以赴
意興風發	意氣風發	生氣勃勃	生氣蓬勃
同聚一堂	齊聚一堂	言之成理	言之有理
共聚一堂	齊聚一堂	自嘆不如	自嘆弗如
不亢不卑	不卑不亢	從頭至尾	從頭到尾
成竹在胸	胸有成竹	無怨無尤	無怨無悔
精疲力盡	精疲力竭	耶誕	聖誕
垂手可得	唾手可得	耶誕節	聖誕節
難分軒輊	不分軒輊	耶誕老人	聖誕老人
無分軒輊	不分軒輊	聖誕老公公	聖誕老人
百折不撓	不屈不撓	損毀	毀損
堅毅不撓	不屈不撓	星期一	週一
縛手縛腳	綁手綁腳	禮拜一	週一
指手劃腳	比手畫腳	星期二	週二
粥少僧多	僧多粥少	禮拜二	週二
餐風宿露	餐風露宿	星期三	週三
心驚膽顫	膽顫心驚	禮拜三	週三
聞名遐邇	名聞遐邇	星期四	週四
縮衣節食	節衣縮食	禮拜四	週四
豐功偉績	豐功偉業	星期五	週五
臨機應變	隨機應變	禮拜五	週五
如醉如癡	如癡如醉	星期六	週六
煞有介事	煞有其事	禮拜六	週六
洋洋得意	得意洋洋	星期天	週日
命在旦夕	危在旦夕	星期日	週日
峻工	完工	禮拜天	週日
一脈相承	一脈相傳	禮拜日	週日
猶疑不決	猶豫不決	醫師	醫生
和衷共濟	同舟共濟	耶誕夜	聖誕夜
一新耳目	耳目一新	耶誕樹	聖誕樹

一言不發	不發一語	耶誕卡	聖誕卡
鼎鼎大名	大名鼎鼎	教部	教育部
來歷不明	來路不明	教局	教育局
僕僕風塵	風塵僕僕	市銀行	市銀
聳人聽聞	駭人聽聞	昨天	昨日
譬如說	比如說	日昨	昨日
天淵之別	天壤之別	明天	明日
迫在眉梢	迫在眉睫	今天	今日
前所未聞	前所未見	多姿多采	多采多姿
萬不得已	逼不得已	紛紛議論	議論紛紛
迫不得已	逼不得已	藏龍臥虎	臥虎藏龍
一語不發	不發一語	瘡痍滿目	滿目瘡痍
繪影繪聲	繪聲繪影	背井離鄉	離鄉背井
正大光明	光明正大	價廉物美	物美價廉
半夜三更	三更半夜	大雨傾盆	傾盆大雨
敵愾同仇	同仇敵愾	無補於事	於事無補
百孔千瘡	千瘡百孔	泰然處之	處之泰然
避兇趨吉	趨吉避兇	人心大快	大快人心
舌劍唇槍	唇槍舌劍	高深莫測	莫測高深
難解難分	難分難解	麗質天生	天生麗質
久久長長	長長久久	萬貫家財	家財萬貫
大同世界	世界大同	無止無休	無休無止
流離顛沛	顛沛流離	萬水千山	千山萬水
萬里晴空	晴空萬里	利己利人	利人利己
地地道道	道道地地	載沉載浮	載浮載沉
獨往獨來	獨來獨往	春風滿面	滿面春風
鱗次櫛比	櫛比鱗次	送舊迎新	迎新送舊
海角天涯	天涯海角	連臺好戲	好戲連臺
八面威風	威風八面	人傑地靈	地靈人傑
陳倉暗渡	暗渡陳倉	山盟海誓	海誓山盟
玉殞香消	香消玉殞		
富貴榮華	榮華富貴		
鴻圖大展	大展鴻圖		
進退應對	應對進退		
老馬識途	識途老馬		
起居飲食	飲食起居		
晨鐘暮鼓	暮鼓晨鐘		
茹苦含辛	含辛茹苦		


E S A
1896