國立交通大學

電信工程研究所

碩士論文

使用階層式語言模型之大詞彙國語辨認系統

Large-Vocabulary Mandarin Speech Recognition using Hierarchical Language Model

研究生：楊雲舒

指導教授：王逸如　教授

中華民國九十九年八月

使用階層式語言模型之大詞彙國語辨認系統

Large-Vocabulary Mandarin Speech Recognition using Hierarchical Language Model

研 究 生：楊雲舒　　Student：Yun-Shu Yang

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

in

Communication Engineering

August 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年八月

使用階層式語言模型之大詞彙國語辨認系統

研 究 生：楊雲舒　　指導教授：王逸如 博士

國立交通大學電信工程研究所

中文摘要

本論文針對中文詞彙中的定量複合詞、人名、綴詞，利用此三類所具有的規則特性將之拆解，以較少數量的構詞單元來涵蓋全部的三類詞彙，可以降低此三種詞類的 OOV 問題。有別於傳統上以"字"為單元來評估辨認率(character error rate)為主，本研究希望以較長且具有意義的詞彙或者詞組(Word Chunk)來作為語音辨認效能的評估；透過詞彙的行為特性，藉由語法與語意資訊為此三種詞類建立可更精細的描述它們的語言模型，重新配置語言模型分數來找出最佳的辨識結果，以提升辨識效能。

由結果所分析，本研究之方法確實能運用此三類詞之語言模型，全面性的描述該詞類的特性，藉此辨識出包含更多語意之詞彙甚至是詞組；往後將再利用詞組本身所具有的結構、語意及語法來得到更多的資訊，建構更有系統且豐富之方法來輔助辨認。

Large-Vocabulary Mandarin Speech Recognition using Hierarchical Language Model

Student：Yun-Shu Yang　　Advisor：Dr. Yih-Ru Wang

Institute of Communication Engineering

National Chiao Tung University

Abstract

It’s difficult to list all words in recognizer’s vocabulary for large-vocabulary speech recognition, so we present an approach for modeling out of vocabulary (OOV) words. In this thesis, we choose three types of word in Mandarin such as determinative-measure compound word, person name and affixation to deal with this OOV problem. Words are converted to the sub-word units and searched for in the hypotheses to cover more new words through the use of flexible sub-word units.

The main focus of this study is to use the grammar and semantic information to construct a hierarchical language model for these three types of word. The language model will be added to promote the recognition performance and hope to recognize more meaningful long-term units such as word and word-chunk.

致謝

耶！我畢業了！！

首先非常感謝陳信宏老師及王逸如老師，這兩年來在課業及研究上的指導，兩位老師的觀點帶給自己不同的思考方向，也花了很多時間與我們咪聽討論研究上的問題，我能夠順利的完成碩士論文非常感激兩位老師的辛勞！

感謝博士班的學長們，感謝研究上花了很多心力幫助我且環保愛地球的希群；感謝關鍵時刻會指引明燈的性獸；感謝總讓我問你一堆哭笑不得問題的合哥；感謝總莫名對我比大拇指但常問我進度狀況的阿德；感謝好像很會喝酒的巴金，希望找工作有像你說的那樣簡單；感謝常講歷史經驗和故事給我們聽的輝哥，恭喜你博班畢業了！感謝上屆的學長，小帥哥、普烏、Q 哥、小宋，你們一畢業沒看到你們真是不習慣！也非常感謝我們同屆的夥伴們，感謝和我一樣是唯二的女生依玲，希望看到妳更多的作品；感謝常被我問問題又極度有耐心的一哥宥余，希望你也是工作部門裡的一哥；感謝很愛嘴砲的小卡，希望你不常看到早晨的星星；感謝竟然沒看到你跳 GeGeGe 的皓翔，你應該去報名食尚玩家主持人甄選成為下一個浩角翔起；感謝也常被我問蠢問題又不熟的 puma，我很難相信你說跟姓楊的不熟(除了我)；感謝很有毅力的承燁，好佩服你以 lab 為居這麼久，恭喜你也恭喜我們大家都畢業了！！感謝碩一的學弟們胖胖、大胖、啟全、銘節、豆腐、智障、小蝦，你們很愛互相調侃但也很有活力、熱血又認真，實驗室就是需要這樣的人，研究邁向世界第一吧！！

最後，還要感謝我的家人，給予我生活上無後顧之憂還有精神上的支持與鼓勵，讓我能順利的完成碩士學位。

目錄

圖目錄

表目錄

第一章 緒論

1.1 研究動機

至今科技發展日新月異，人類追求科技創新之外愈來愈注重實用性，電子產品出陳推新，不斷追求輕薄短小與可攜性，資訊的交流發達與快速使得人類的生活與各式各樣的數位產品緊密相連，連帶影響了現今人類的行為模式。聲音是人類最直覺且最便利的溝通方式，「語音」將逐漸取代文字扮演著重要角色，當作為和機器溝通的媒介必為一發展趨勢；而語音的應用領域具多元化，語音檢索、語言學習.....，其中發展語音辨識技術可為相關研究帶來極大效益。

中文語音辨認朝大詞彙語音辨認(LVSR)系統發展，中文和其他語言的最大相異處是詞的邊界模糊，造成詞彙定義的困難，並且詞彙種類繁多，受限於詞典詞條數限制，詞典無法收錄所有詞條，可能使得詞的涵蓋率不足，產生詞典外詞彙(out-of-vocabulary, OOV) 過多的問題【1】，都會影響訓練語言模型。因此針對此問題提出解決方法，考慮中文特性，對具有明顯之規則特性且數量多的定量複合詞、人名、綴詞這三種詞類來分析探討。

這三類詞皆能以有限的詞條組合出無限的詞彙，利用這點特性，將詞彙拆為較短的單元收錄至辨認詞典之中，進而增進詞典的涵蓋率，更能有效的解決 OOV 問題。另外，有別於傳統上以"字"為單元來評估辨認率(character error rate)【2】，本研究希望以較長且具有意義的詞彙或者詞組(Word Chunk)來作為語音辨認效能的評估，因此，針對定量複合詞更進一步的研究，透過定量複合詞再構詞得到更具有語意資訊的詞組，有架構、系統的分析產生其特有的語言模型，配合 two-stage 辨認架構【1,3-6】，加入這三類詞的語言模型分數，將這些詞正確地辨識出來以提高詞辨識率。

1.2 研究方向

本論文針對中文詞彙中的定量複合詞、人名、綴詞將之拆解，以較少數量的組合單元來涵蓋這三類詞彙，藉以降低這三種詞類的 OOV 問題，並更精細的描述它們的語言模型來幫助辨認。透過詞彙的行為特性，藉由語法資訊與語意資訊定義 OOV 的種類來減少中文 OOV 的問題，也重新定義中文語音系統的辨認單元，以含有更多語意資訊的詞彙或詞組作為語音辨認系統效能評估的依據。

以英文的電話詢問系統為例【1】，龐大的城市名無法全部收錄，大量低詞頻城市名使用相同的 backoff 機率也不易正確辨認出來，採用二階式辨認系統(two-stage ASR)將低頻的城市名拆解為 phone sequence 當作第一級的辨認單元，於第二級以查詢的方法將這些詞構回。

本研究同樣採用 two-stage ASR 系統，先以第一級之 LM 辨認產生混合 word 及其他較小的組合單元之 word lattice，再加入三種詞類的語言模型重新配置語言模型分數，藉此刪除不合理的路徑，並加強可靠的 lattice 路徑得到最終的辨認結果。

1.3 相關研究

語音辨識系統中，語言模型扮演重要的角色，透過語音辨識，將聲音轉換成文字，更能進一步分析文字的語意及語法，甚至良好的語言模型可以應用在各領域之中。

大部分的大詞彙語音辨認使用統計式語言模型(Statistical Language Model)，其中以 *N* 連語言模型(*N*-gram Language Model)最常被使用，然而在訓練 *N* 連語言模型時隨 *N* 值增大會有資料稀疏的問題，無法收集所有的 *N* 連組合詞彙，導致預估機率不準確，因此有平滑化(Smoothing)方法被提出，處理未出現在訓練語料中的詞彙組合，使其機率不為 0。

而為增進辨識系統效能，有許多學者提出方法與語言模型結合，以輔助 *N* 連語言模型，並且從不同角度探討語言模型，從詞彙衍生至類別資訊、語句結構、文章語意...等。*N* 連語言模型為訓練詞與詞之間的機率關係，但僅能得到相鄰距離的詞彙資訊，1993 年略詞模型(skipping model)被提出【7】，應用於 *N* 連語言模型時可略過數個詞的距離，利用歷史詞序列

中的詞來預估接下來的詞彙出現機率，以及觸發對模型(trigger pair)等以長距離資訊來輔助 *N* 連語言模型。

再者，1992 年 Brown 等人將詞彙進行分類，提出了 *N* 連類別模型(class-based N-gram model)【8】，加入了類別資訊來訓練語言模型，其優點為可將詞彙依特性分群並且資料的預估由詞彙的組合減少至類別的組合，進而改善資料稀疏的問題。1997 年結構化語言模型(Structured Language Model)被提出，利用語句結構的資訊，剖析歷史詞串並且以主導詞來預估詞彙發生的機率；也有以語句中潛藏的語意應用再語言模型上，例如潛藏語意分析(Latent Semantic Analysis, LSA)【9】分析詞彙與文章間的語意關係，將此資訊與 *N* 連語言模型結合。

1.4 章節概要

本論文的內容共分為五章：

第一章：緒論：介紹本論文之研究動機與研究方向。

第二章：傳統語言模型：此章節介紹語言模型的建立，包括：文字語料、辨認辭典的處理，以此作為和階層式語言模型辨識效能之比較基礎。

第三章：階層式架構語言模型：為建立更精細的的語言模型，說明如何建立三類詞的構詞模型及語言模型分數之重新配置。

第四章：實驗結果與分析：分析實驗結果以及辨識效能，並且與基本系統進行比較。

第五章：結論與未來展望。

第二章 傳統語言模型

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」。在本研究中，採用高詞頻六萬詞之辨認詞典所訓練的語言模型，即以傳統語言模型(General LM)稱之。

對於語言模型之建立，會在決定辨認詞典前，將文字資料庫經過文字前處理之流程以修正來符合語音辨認的形式。2.1 節中將對語言模型做基本的介紹，2.2 節介紹文字資料庫及文字的前處理細節，2.3 節說明詞典中詞彙的收錄方式和 OOV 的處理。

2.1 語言模型的基本介紹

在大詞彙連續語音辨認中，希望將輸入的語音辨認出合理的詞彙順序；為達到此目的，語言模型就必須考慮整段辨認語音中前後詞彙的關連性，不僅是單一字詞獨自出現的機率。

目前最廣泛使用的語言模型為 N-1 階馬可夫(Markov)假設，即 N 連(N-gram)語言模型，而目前我們所建立的語言模型為 N=2 的 Bigram 語言模型，即所預估的詞只和前一個詞有關。General LM 在此只採用 bi-gram 語言模型，而我們希望不單只採用 Bigram 語言模型，僅能使用相鄰距離的詞彙資訊，所得到的資訊太少。因此，本論文在進一步的研究方法中加入其他模型，如 N 連類別模型(Class-based N-gram Model)及觸發對模型(Trigger pair Model)，來改善 Bigram 語言模型並以下加以介紹。

2.1.1 N 連語言模型

式(2-1) $P(W)$ 是欲辨認詞串 W 的事前機率，其中 $W = w_1, w_2, \ldots, w_m$，$w_i \in V$ 代表 m 個詞所組成的詞串，V 則是詞典為所有詞的集合。使用貝式法則，$P(W)$ 可以分解成：

$$P(W = w_1, w_2, \ldots w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, w_2, \ldots w_{i-1}) \tag{2-1}$$

其中是 $P(w_i \mid w_1, w_2, \ldots w_{i-1})$ 詞 w_i 在給特定歷史詞串 $h_i = w_1, w_2, \ldots w_{i-1}$ 的情況下，會緊接著詞 w_i 出現的條件機率。

實際上在建立語言模型時，並不會將所有可能的參數 $P(w_i \mid w_1, w_2, \ldots w_{i-1})$ 都儲存起來；因為針對長度為 m，歷史長度為 $m-1$ 的詞串，所有可能的組合個數為 $|V|^m$；即使詞典為中等大小，只要詞串長度 m 稍長，參數量將會驚人的成長，因此必須做參數量的簡化。

簡化參數量的方法之一，就是裁減歷史詞串的長度。所謂 N 連語言模型，就是對參數 $P(w_i \mid w_1, w_2, \ldots w_{i-1})$ 做近似，此模型假設詞 w_i 出現的機率只和前面 N-1 $(\mathrm{N} < m)$ 個詞 $w_{i-\mathrm{N}+1}, w_{i-\mathrm{N}+2} \ldots w_{i-}$ 相關，而和 N 個之前的詞串 $w_1, w_2, \ldots w_{i-\mathrm{N}}$ 完全獨立，如此一來歷史詞串的長度便可輕易的裁減，模型的參數量也會因此大大的降低。根據此假設，則 $P(W)$ 可表示為：

$$P(W = w_1, w_2, \ldots w_m) \simeq \prod_{i=1}^{m} P(w_i \mid w_{i-\mathrm{N}+1}, w_{i-\mathrm{N}+2}, \ldots w_{i-1}) \tag{2-2}$$

實際估測 $P(w_i \mid w_{i-\mathrm{N}+1}, w_{i-\mathrm{N}+2}, \ldots w_{i-1})$ 的方式，是根據最大相似度估測法(maximum likelihood estimation, MLE)，得到下式：

$$P(w_i \mid w_{i-\mathrm{N}+1}, w_{i-\mathrm{N}+2}, \ldots w_{i-1}) = \frac{C(w_{i-\mathrm{N}+1}, w_{i-\mathrm{N}+2}, \ldots w_i)}{C(w_{i-\mathrm{N}+1}, w_{i-\mathrm{N}+2}, \ldots w_{i-1})} \tag{2-3}$$

其中 $C(\bullet)$ 表示詞串出現次數。

當 $\mathrm{N} = 2$ 時，成為雙連語言模型(bigram language model)，詞串 W 的機率可表示為：

$$P(W = w_1, w_2, \ldots w_m) \simeq \prod_{i=1}^{m} P(w_i \mid w_{i-1}) \tag{2-4}$$

及

$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})} \tag{2-5}$$

2.1.2 *N* 連類別模型

1992 年 Brown 等人提出的 *N* 連類別模型(Class-based *N*-gram Model)，將相近語意、詞性或是結構相似的詞彙分類，同一類別中的詞彙具有相似意義能共享參數，資料的預估可能由詞典的大小縮減至類別種類數量，甚至當訓練語料中不存在某些詞彙時還可預估其機率，改善資料稀疏的問題。

N 連類別模型的定義：

$$P(w_i \mid w_{i-n+1},...,w_{i-1}) \approx P(w_i \mid c_i)P(c_i \mid c_{i-n+1},...,c_{i-1}) \tag{2-6}$$

其中，$P(w_i \mid c_i) = \frac{count(w_i)}{count(c_i)}$ 為詞 wi 在其類別 ci 中出現的機率值， (2-7)

$P(c_i \mid c_{i-n+1},...,c_{i-1}) = \frac{count(c_{i-n+1},...,c_i)}{count(c_{i-n+1},...,c_{i-1})}$ 為出現該類別的機率。 (2-8)

2.1.3 觸發對模型

文章之中常有許多相關聯的詞彙，但不一定相鄰伴隨著出現，可能出現在同一句子當中，Bigram 語言模型較難得到如此資訊，以觸發對收集長距離的詞彙資訊，來補償 Bigram 語言模型長距離資訊不足。Trigger pair 可以 A → B 來表示，A 為觸發項，B 為被觸發項，當 A 出現於文章時 B 可能於後方伴隨出現，並且設定有長度限制的視窗(window)，移動視窗統計兩個詞彙共同出現的次數，其機率的估計我們利用以下定義：

$$P(w_2 \mid w_1) = \frac{N(w_1, w_2)}{\sum_j N(w_1, w_j)} \tag{2-9}$$

$N(w_1, w_2)$ 為 w_1, w_2 在視窗中共同出現的次數，再統計 w_1 觸發視窗內全部 j 個詞彙的數量來得到 $w_1 \rightarrow w_2$ 的統計機率。

2.1.4 語言模型平滑化

以雙連語言模型為例，在訓練資料中計算 $P(w_i \mid w_{i-1})$ 時，若 $C(w_{i-1}, w_i) = 0$ ，將會使得

$P(w_i \mid w_{i-1})$機率等於零，因為在訓練資料中並未出現，但是這並不是代表測試資料中不會出現，因此這種情況下機率的給定是不合理的。且當$C(w_{i-1}, w_i)$值很小時，所計算出的機率也是不準確的。所以必須對計算出的機率做平滑化的動作，使所有的機率均能被良好的估計。

退化平滑法(back off)及使用 Good-Turing discounting。若定義訓練語料中詞串出現的次數門檻值k，則可將詞串分為出現次數高於門檻值、出現次數低於門檻值及從未出現三種。則參數可表示為下式：

$$P(w_i \mid w_{i-\mathrm{N}+1}, w_{i-\mathrm{N}+2}, \ldots, w_{i-1}) = \begin{cases} \alpha(w_{i-\mathrm{N}+1}, \ldots, w_{i-1}) P(w_i \mid w_{i-\mathrm{N}+2}, \ldots, w_{i-1}) & , C(w_{i-\mathrm{N}+1}, \ldots, w_i) = 0 \\ d_c(w_{i-\mathrm{N}+1}, \ldots, w_{i-1}) \cdot \dfrac{C(w_{i-\mathrm{N}+1}, \ldots, w_i)}{C(w_{i-\mathrm{N}+1}, \ldots, w_{i-1})} & , 1 \le C(w_{i-\mathrm{N}+1}, \ldots, w_i) \le k \\ \dfrac{C(w_{i-\mathrm{N}+1}, \ldots, w_i)}{C(w_{i-\mathrm{N}+1}, \ldots, w_{i-1})} & , C(w_{i-\mathrm{N}+1}, \ldots, w_i) > k \end{cases} \quad (2\text{-}10)$$

2.1.5 語言模型評估—混淆度(Perplexity)

據前人研究顯示，混淆度已成為評估語言模型相當重要且通用的參考標準；混淆度是根據消息理論(information theory)而得：

$$H = -\frac{1}{m} \log P(W = w_1, w_2, \ldots, w_m) \quad (2\text{-}11)$$

上式為一個詞串$W = w_1, w_2, \ldots, w_m$，對於每個新詞提供的平均資訊量(entropy)，經過適當的化簡而得。而混淆度可直接使用式(4.8)進一步定義為：

$$PPL = \exp(H) \quad (2\text{-}12)$$

若$P(W = w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, w_2, \ldots, w_{i-1})$則可發現，混淆度就是$P(w_i \mid w_1, w_2, \ldots, w_{i-1})$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皆詞有較多的選擇，辨認時就越難找到確切的答案；反之，則較易找到正確答案。

2.2 文字資料庫

在訓練語言模型之前，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，之後再以訓練 word-based 之 General LM。文字前處理流程大致可分為：斷詞、文字正規化和處理 OOV，如圖 2.1 所示。將在以下小節分別說明文章所採用的斷詞方法及文字正規化的處理。

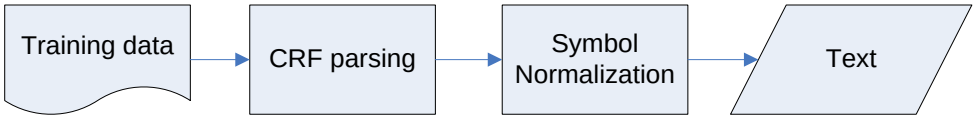


圖 2.1：文字資料處理流程

2.2.1 文字資料庫介紹

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：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。

針對訓練 General 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 2.1 中呈現，如下：

表 2.1：General 語言模型字數及詞數統計

語料庫	詞數(Word)	字數(Character)	詞條數	平均詞長
光華雜誌、NTCIR	116,173,318	219,893,736	1,073,479	1.80673

2.2.2 文章斷詞

語言模型是經由統計的方式建立，統計詞彙和詞彙之間的連接機率關係，所以把文章斷詞來統計詞彙和詞彙之間的機率，而語言模型的好壞也會和斷詞時所決定詞的邊界有關。

傳統的中文斷詞與詞性標記系統使用的是長詞優先及構詞規則，這些方法需要結合專門

領域的知識及大量的手工標記資源才能使中文斷詞研究做得不錯，最著名的是中央研究院的中文斷詞系統【10】。但自 2001 年起，由於條件隨機域(conditional random field ,CRF)方法【11】被提出，並有效的使用在自然語言處理斷詞器上。

以往斷詞所採用的為字典資訊，現今以 CRF 方法藉由學習句法結構與標記詞性來斷詞，和過去斷詞方法相比 CRF 斷詞可以產生較正確的斷詞結果，也能將很多辭典未收錄的詞彙正確斷出，大幅減少 OOV 所產生的連續短詞串問題。使用 CRF 斷詞也能得到更多辭典未收錄的詞，藉以擴充人名、詞綴詞清單，以強化後級構詞階段的正確性；由下圖 2.2 概略說明 CRF 斷詞器之處理流程。

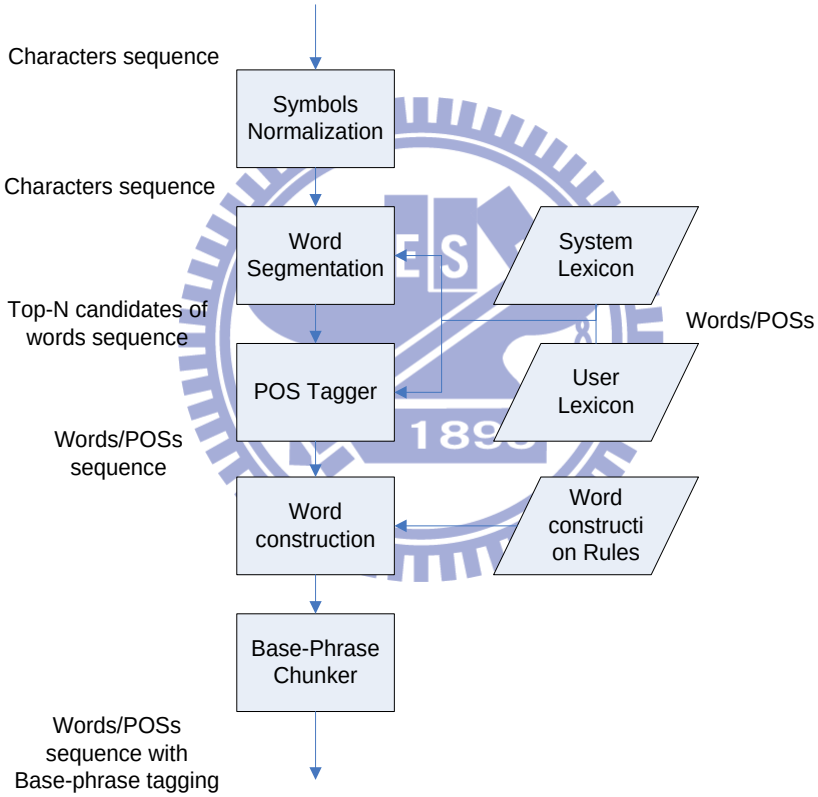


圖 2.2：CRF 斷詞流程

(1) 文字正規化單元(Symbol Normalization)

於此級處理了兩件事；其一，輸入的文句當中可能包含 ASCII code 或 BIG5 code，在進入斷詞單元之前，必須將所有的 ASCII code 轉成 BIG5 code 格式，使字串格式統一。其二，中文的用字歧異和標點符號的統一在此單元中解決，例如「裏」和「晒」轉為「裡」

和「曬」，而標點符號中如全形逗號「，」也統一轉成半形「,」。

(2) 斷詞單元(Word Segmemtaion)

斷詞單元是最前面的分詞單元，目的是將輸入文句做適當的斷詞，也是整個文句分析器當中最核心的部分。此單元找出所有可能的詞串組合，以供給下一級的「詞類標記單元」標記詞性。

(3) 詞類標記單元(Pos Tagger)

斷詞候選句產生後，依照學習到的句法結構，利用機率統計模型，給予候選句每一個詞彙一個最高機率數值，即最可能出現詞性標籤，來標記所有可能的詞性組合，並從候選句當中選出最合乎語法的句子。

(4) 構詞單元(Word Constuction)

於此級加入了構詞規則，再加上上一級所產生的 POS 資訊，將短詞組合為長詞彙；其中即加入了七條定量複合詞的組合規則來產生定量複合詞；甚至可將原本無法收錄至詞典當中的詞，其中有規律的利用構詞規則來合併詞彙提供了良好的斷詞結果。

(5) 輸出單元(Base-Phrase Chunker)

使用中央研究院中文句結構數資料庫【12】當作訓練語料，對斷詞後之 word sequence 標示出 ADVP、AP、NP、GP、PP、AP、ADVP、VP 各種詞組結構，其中分別是：副詞詞組、形容詞詞組、名詞詞組、方為詞詞組、介詞詞組、述詞詞組。

2.2.3 文字資料庫處理

對文字資料中所做的處理分別有：標點符號、英文串、文字正規化，以下之各小節詳細敘述之。

2.2.3.1 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，所以在文

章中標點符號的處理，利用點號中的四種符號(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。

2.2.3.2 英文串處理

由於我們的辨認目標為中文詞彙，聲學模型中並沒有訓練英文詞的聲音模型，所以文章中的英文詞以「LONGFW」符號來表示，我們將所有的英文詞當作一個類別看待它；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 OOV。

2.2.3.3 文字正規化

文字正規化可分為兩大部分：第一，文章的內容有些阿拉伯數字、詞彙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造成用字歧異，但在讀音上及語意上是相同的，需把這類的詞合併成同一個詞。這些處理過程以文字正規化稱之。

(1) 寫法轉讀法

將阿拉伯數字、詞彙或符號由文字書寫方式改為語音讀法格式，其中主要是數字部分的處理，其正規化的處理可由下表 2.2 範例示之。

表 2.2：文字正規化範例

正規化前	正規化後
120 號	一百二十號
90・23	九十點二三
35%	百分之三十五
二二，三三零人	二萬二千三百三十人

(2) 同音義異詞處理

某些詞在發音上甚至語意都是相同的，只在寫法上有所差異，而這類的詞若當作不同的詞彙對待會使得辨認上造成混淆，所以把這類的詞統一，合併為一個詞視之，如表 2.3。經過這個步驟可將文章的詞彙更集中，促使 OOV 量減少。

表 2.3：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
洩露國家機密	洩漏國家機密

2.3 建立辨認詞典

本研究的傳統式語言模型(General LM)之辨認辭典，其收錄方式為統計語料庫中全部的詞，並依照其詞頻排序，直接收錄了高詞頻的六萬筆詞條；以下表 2.4 列出詞典的字數分布狀況，並且以圖 2.3 表示詞典中詞彙的收錄數量與其對應的語料涵蓋率。

表 2.4：General 語言模型辨認詞典字數統計

General 語言模型辭典分析										
字數	一字	二字	三字	四字	五字	六字	七字	八字	九字	Total
數量	2908	35965	15327	4691	691	337	70	7	4	60000

由圖可看出，辨認詞典收至六萬詞時，語料庫的涵蓋率達 95.95%，OOV rate 約為 4.05%。

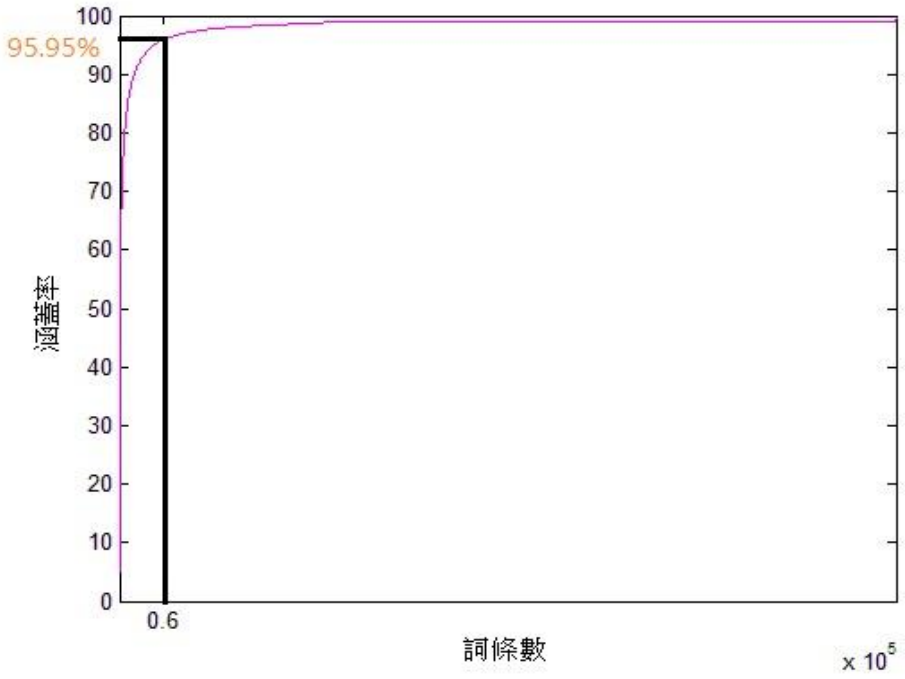


圖 2.3：詞典涵蓋率

2.3.1OOV 處理

過去 OOV 的處理方式為將所有 OOV 拆解為辨認詞典中的詞，因此所有的詞彙便可由詞典內的詞所組成，以此作法增加了 OOV 的涵蓋率；但現在這些低頻的 OOV 詞彙不再特別的處理，避免 OOV 拆解後的短詞干擾了詞典中的詞彙統計特性。

2.4 傳統語言模型之探討

受限於辨認詞典大小的限制，General LM 主要以收錄高詞頻的詞彙來提升對語料的涵蓋率，其對語料的涵蓋率雖可達一定之水準，然而希望採用其他方法來降低中文 OOV rate 同時提高辨識效率。因此，在第三章採用的階層式架構的語言模型當中，辨認詞典的收集方法將有所變化，藉著收錄 subwords 處理 OOV 問題及增加語料的涵蓋度，並且建立更精細的語言模型來輔助辨認。我們將比較此兩種系統之辨識效能，並探討語料的不同處理方式對辨識效能的影響。

第三章 階層式語言模型

中文詞彙數量繁多，詞與詞間也可再組合構成新詞彙，其中許多詞類為 open set 無法收集全數的詞彙組合，如：數詞[1](Neu)、專有名詞(Nb)、綴詞…等，而在辨認詞典有其詞彙量的限制下，無法將所有的中文詞彙收錄至詞典當中，會使得詞典的涵蓋率降低，語音辨識效能成長有限。因此，欲藉由修正詞彙的收集方式，以改善中文長久以來 OOV 過多的問題。

本研究則針對 open set 中數量較多且有明顯構詞規則的詞彙進行處理，如：定量複合詞(determinative-measure, DM)、中文人名(PN)及綴詞(MD)。因為人名、綴詞富有規則特性，可依據其規則拆解為較小的單元（以下將此單元以 subword 稱之）；而定量複合詞由多個詞彙所構成，雖然可由高階的 *N*-gram 預測此詞組的機率，但我們利用 DM 結構的完整性及與前後詞的強烈關係，依據 DM 的構詞規則為此建立語言模型。利用此三類所具有的規則特性，藉以收錄這三類詞彙的組合單元至詞典當中，以較少數量的 subword 來涵蓋全部的三類詞彙可以降低此三種詞類的 OOV 問題，並可更精細的描述它們的語言模型。

辨認系統採用二階式(two-stage)架構語音辨認系統，先初步的建立一個 LM 能夠產生高涵蓋率的 word lattice，再使用更精細的階層式 LM 並 rescoring 配置 LM 分數來找出最佳的辨識結果。在 3.1 介紹三類詞特性及辨認詞典的收集方式；3.2、3.3 節中，分別介紹為了有可更精細的描述三類的階層式詞組語言模型，最後於 lattice 上對語言模型機率重新調配；3.4 節介紹所採取的辨認系統之整體架構。

3.1 大詞彙語言模型之分析

本研究中，欲使用較小的單元可以涵蓋大量並有規則性的中文 open set，並利用語言學的知識來建立這些詞類的語言模型，於是，先建立一個 LM 以期能夠符合高涵蓋率的特性，之後再加入特定詞類的語言模型來輔助辨識。在以下小節介紹定量複合詞、人名、綴詞這三

[1]根據中研院詞庫小組，將中文詞彙依詞性(POS)可細分為 46 類標記【13】。

類詞並分析，隨後將這三類詞以其語言學規則特性來拆解成由多個 subword 所組合的詞彙，並且決定辨認詞典中這三類 subword 的收錄方式以達到建立此 LM 的目標。

3.1.1 三類詞之統計

建立此語言模型之語料中，其中的三類詞將被拆解為許多 subword，因此詞彙之長度變得較零散，重新統計此訓練語料之資訊，如下表 3.1。

表 3.1：大詞彙語言模型字數及詞數統計

語料庫	詞數	字數	詞條數	平均詞長
光華雜誌、NTCIR	124,579,263	219,893,736	663,981	1.76509

定量複合詞、人名、綴詞這三類都是一種擴張性強又極有規律的詞，詞頻集中分布在部分常用詞，大部分的詞條出現次數少卻同時種類繁多，無法被收錄的長詞會造成語言模型預估的困擾，由下表 3.2 統計出此三類詞約 9%。

表 3.2：三類詞之分佈

	DM	PN	MD	其他	總詞數
總量	5,600,480	1,537,228	3,515,465	105,520,145	116,173,318
百分比	4.82%	1.32%	3.03%	90.83%	100%

3.1.2 定量複合詞

定詞具有標示名詞組指涉或數量的功能，量詞是用以計量的單位詞，常和定詞構成定量式複合詞來修飾名詞【13】。DM 來源由使用 CRF 方法斷詞並標記 POS 之後，將符合 DM 之構詞語法規則的詞條挑選出來。實際上定量複合詞富變化性，但構成具規則性，以下舉例依其結構或語意將 DM 大致分為 4 大項，此外將非數詞定詞或量詞的詞類概括以修飾詞來稱之，數詞定詞則以數詞簡略之。

(1) 含(修飾詞、數詞、量詞)：頭[Nes]一[Neu]年[Nf]

(2) 含(修飾詞、量詞)：這[Nep]項[Nf]

(3) 含(數詞、量詞)：十二[Neu]年[Nf]

(4) 含(修飾詞、數詞)，可能省略了量詞：另[Nes]一[Neu]

在語句中可看出定量複合詞無法描述完整的語意，但實際上要修飾的名詞或者描述 DM 的修飾詞可能出現在距離定量複合詞幾個詞之外，這些詞類也會影響了量詞的選擇，如表 3.3 範例；根據這因素，分別以 POS 標記及文字語意這兩項訊息，統計 DM 與前、後詞彙及 POS 的關聯性，希望加入這些訊息以幫助 DM 構詞，加強 DM 詞組語意的完整性。

表 3.3：DM 詞組構詞範例

DM 中省略的名詞或者修飾詞出現於鄰近詞中
高(VH) 一六零[Neu]公分[Nf](DM)
高達(VJ) 三百[Neu]卡[Nf](DM) 的(DE) 熱量(Na)

除了針對 DM 之外，也將「時間」相關的詞一併處理。因為時間名詞(Nd)在語意上雖和 DM 不同，但觀察其文字特性與 DM 有極大的相似性。如十二月(Nd)與十二[Neu]月[Nf](DM)的差別在於，Nd 為時間名詞用來表示一時間狀態，而以 DM 為標記的十二月則是表示計量涵義；此類時間名詞結構與 DM 極為相同，因此一併視為如同 DM 來處理。

為構成更似完整語意的 DM 詞組，主要針對 DM 前後的修飾詞及名詞一起構成詞組，以下列之。

(1) DM 前方構詞

前方構詞參考 POS 資訊，利用 Da 為數量副詞的特性將之構入 DM。此外，條列出「長、寬、高、深、厚、重、粗」這些修飾詞，對 DM 中特定的量詞，參考如中研院詞庫小組整理之量詞表[附錄一]中的長度量詞、重量量詞、容量量詞來構詞。

(2) DM 後方構詞

DM 後方構詞，主要構入名詞(Na)及後置詞(Ng)這兩種詞類。其中，利用觸發對(trigger pair)方法，統計量詞與名詞間的相似度，再篩選適當的名詞來加入 DM；而後置詞 Ng 能更加完整描述 DM 詞組語意也一併將之構入 DM。

此外，將人們習慣會念成一個詞的例子構在一起，譬如；「一、二十人」、「甲、乙」或

連續國家縮寫「美、日」，將這些視為一個 prosodic word 看待，目前實驗室也積極的希望加入韻律資訊來協助語音辨認，藉由韻律模型來驗證構詞的合理性，使得詞的結構更加緊密不易產生混淆，因此對此類的詞彙進行構詞。

詞典的詞彙收錄，DM 詞組中除了數詞之外，其他皆以其原始構成方式拆解並直接收集。數詞的組合單元考慮了兼顧構詞能力與詞意的完整度，收集全部數字的變化型態，建構了數詞的組合單元集合如[附錄二]，然而，我們再將同結構、同性質的數詞組合單元歸至同一個集合，總共得到 25 種數詞組合單元集合。數詞的組合變化多，譬如：整數的數字串、只含個位的數字、含小數點與百分比的數字…等，但仍依據此集合來拆解數詞。

DM 詞組拆解為 subword 並全數收錄，使得 DM 詞組的涵蓋率達 100%。統計 DM 詞組成份，總共收錄了 6441 筆，詳細如下表 3.4。其中 554 筆為數詞 subword 單元集合，其餘的單元則為修飾詞、量詞、名詞…等。

表 3.4：DM 詞組成分分析

DM 詞組成分分析											
POS	Da	Nep	Nes	Neqa	Neqb	VH	Nf	Ng	Na	Nd	Num[2]	總量
數量	53	6	48	115	6	9	420	61	4797	372	554	6441

這些詞彙之中可能存在部分已是前 57000 的高詞頻一般詞，最後統計得知 DM 詞組的組合單元之中總共有 5485 筆為 57000 內的一般詞，實際上新加入的 DM 詞組構詞單元只再新增了 956 筆詞。以下表 3.5 列出，前 57000 高詞頻詞條所涵蓋的 DM 詞組比例，可看出已涵蓋了大部分 DM 詞組。

表 3.5：詞典對 DM 的涵蓋率統計

57K 辭典 DM 涵蓋率		57K 辭典之外 DM 涵蓋率	
3705885	66.17%	1894595	32.83%

[2] Num 在此代表數詞的組合單元

3.1.3 人名

本研究人名詞彙僅討論中文人名，人名可視為由為「姓氏」、「名字」此組合規則所構成，因此可依此原則將人名拆解。詞典數量有限，無法直接收錄所有的完整人名，故依照完整人名、姓氏與名字的詞頻高低分為三種形式收錄於詞典當中：第一種為詞頻較高的完整人名，直接收錄於詞典當中；第二種為詞頻較高的姓氏與名字；第三種為人名中詞頻較高的一字詞以提升人名的涵蓋率。

人名若出現於前 57000 詞條之中，直接以全名形式收錄了 3210 筆，而其餘低詞頻的詞彙依上述的說明方式來拆解。表 3.6 列出人名詞條的收錄數量，名字 subword 單元包含二字詞和一字詞共收錄 1991 筆，名字所相對應的姓氏則收了 262 筆，詞條數總共為 5463 筆。

表 3.6：詞典中人名的詞彙統計

	完整人名	姓氏	名字	Total
詞條數	3210	262	1991	5463
總量	999651	229726		1229377

然而，前 57000 詞條當中部分高詞頻詞條也有同樣是姓氏、名字的組合單元，譬如：自強、中興...等，是名字的組合單元亦為常見的一般詞，因此以圖 3.1 說明人名的收錄方式，以不同的詞彙形式收錄及其涵蓋率；橫坐標為與人名相關的詞彙數目，可能是全名、二字詞 subword 或是一字詞，縱座標為相對應的涵蓋率。

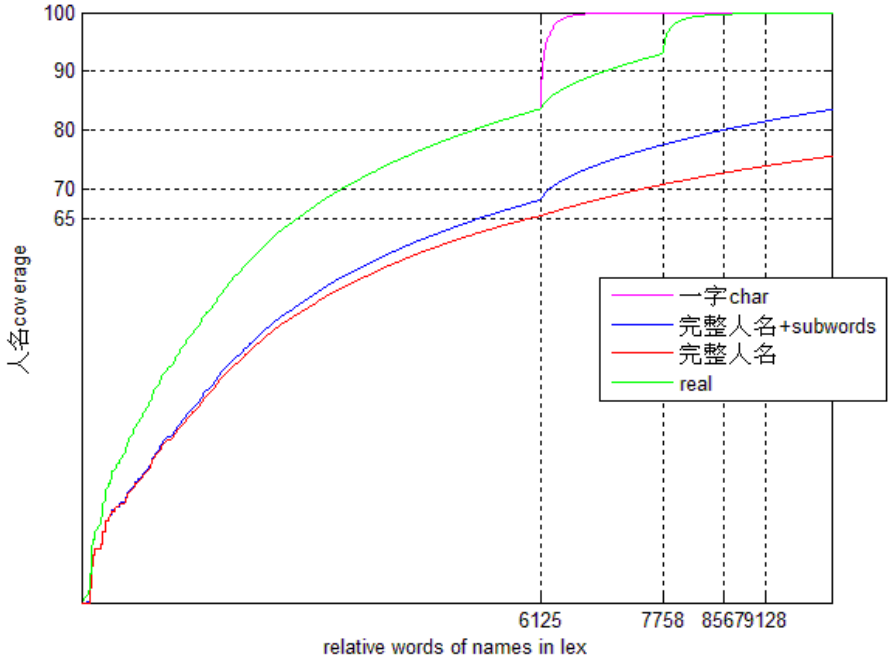


圖 3.1：詞典中人名的詞彙數量其對應之涵蓋率

橫坐標為 6125 時的意義，代表 57000 高頻詞條當中有 6125 筆詞條是與人名相關、並可涵蓋至人名的數量。紅色曲線表示以全名且未拆解的形式收錄，在 57000 詞條中全名收了 3210 筆涵蓋率佔 65.52%，總共必須收錄 97271 筆全名才能達到 100%覆蓋率；藍色曲線包含了全名涵蓋率之外，再多考慮收錄了 subword 的情形，也就是當以 subword 形式收錄時人名的涵蓋率，57000 詞條當中有 1493 筆可當作人名 subword 使涵蓋率提升至 68.21%，需再收 46306 筆 subwords 才能涵蓋所有人名；粉紅色曲線則代表當人名全部被拆為連續一字詞時，57000 內 557 筆一字詞即達涵蓋率 83.64%，再收 2441 筆一字詞其曲線成長速度極快能達到 100%涵蓋率。

最後，採取折衷的方式來收錄人名，即以綠色曲線來表示。於 57000 詞條當中人名的涵蓋率為 83.64%，再收錄 1633 筆 subwords 使得涵蓋率提升至 93.01%，最後再新增 100 個人名常用字，使涵蓋率上升至 97.36%。人名相關的詞條數如上表 3.6 所示共收錄 5463 筆，扣除大部分已重複出現於 57000 詞條中，新增於詞典的數量為 1733 筆。

3.1.4 綴詞

綴詞的組合結構為「詞幹(stem)」及「詞綴」，其中，詞綴又可分為前詞綴與後詞綴，並且皆為一字詞，如表 3.7。依中研院所統計綴詞數量龐大收不勝收，其中收錄了衍生性極強的詞幹與詞綴，而本論文將針對常出現的綴詞進行處理。綴詞同樣依照詞頻高低以兩種形式收於詞典當中：高詞頻的完整綴詞直接收進詞典，另一形式則拆解為「詞幹(stem)」及「詞綴」分別將之收錄。

表 3.7：綴詞範例

綴詞	詞綴
台灣 人	人
化妝 師	師
老 母雞	老

綴詞同上所敘述的方式拆解，收錄方式比人名簡易；出現在 57000 的高詞頻詞條中，代表此詞彙為常用詞而直接收錄長詞不再拆解，於 57000 之外拆解為詞幹與綴詞來收錄，表 3.8 列出綴詞的收錄總數量。收錄的 4438 筆詞幹當中，4020 筆為高詞頻一般詞，232 筆詞綴也為高詞頻一字詞，也就是說，綴詞 subwords 只新增了 418 筆，綴詞的涵蓋率已達到 96%。

表 3.8：詞典中綴詞的字彙統計

	完整綴詞	詞綴	詞幹	Total
詞條數	4948	232	4438	9618
總量	2876588	487104		3363692

3.1.5 大詞彙語言模型之辨認詞典分析

以往作法，60000 詞的辨認詞典直接將所有詞條依詞頻排序收錄至一定的數量，其餘空間保留收集構詞單元，甚至為了提高 OOV 的涵蓋率，收錄了許多低詞頻的一字詞，使得犧牲了其他的高詞頻詞彙；但過多的一字詞可能會造成辨認時一字詞混淆度增加，不易辨認出

正確的一字詞。因此，重新調整詞典的收錄方式，詞典中的詞彙以高詞頻的一般詞為主，再加入能以構詞規則組合回長詞的三類 subword 為輔。

首先，先行排除了所有 DM 詞組之後，依詞頻排序其餘詞條，直接收錄前 57000 的高詞頻詞條，其涵蓋率達 92.06%。而剩餘的 3000 詞典空間則收錄三類詞的 subwords，其辨認詞典的建立流程可見圖 3.2；其中紅色框所示的方塊圖表示三類詞的拆解方式，其處理過程由上述 3 小節分別詳細說明。

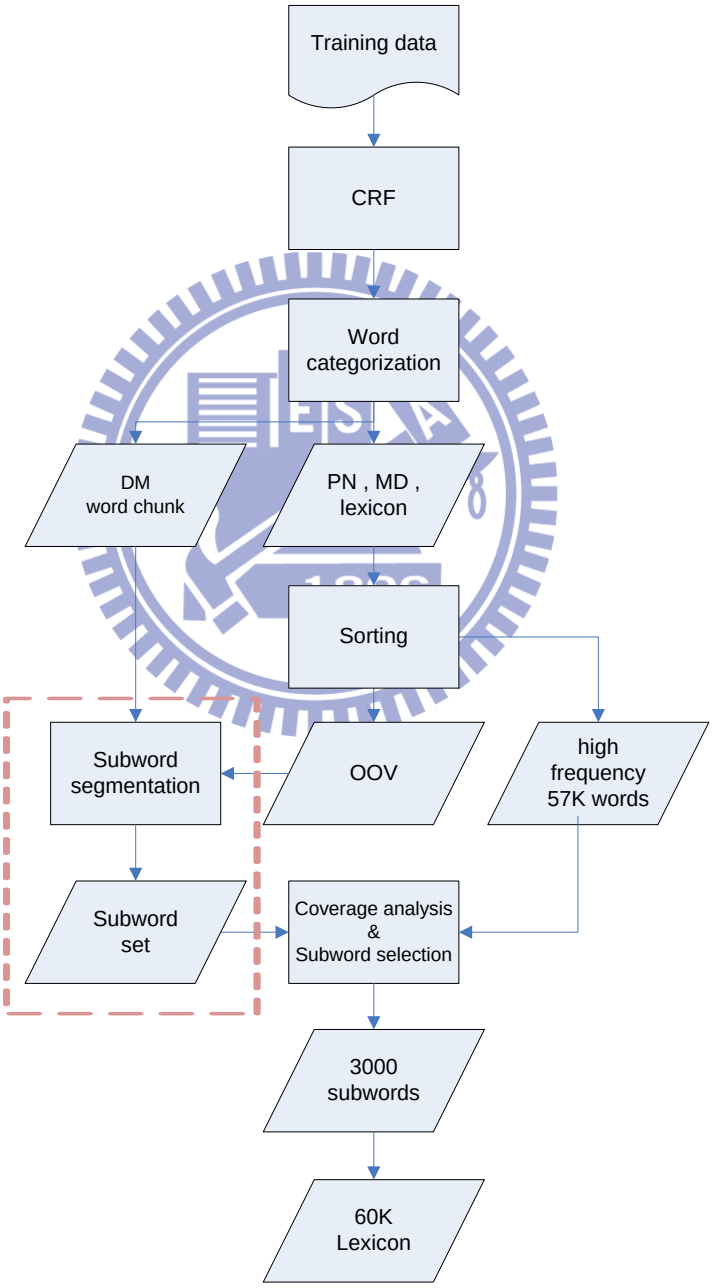


圖 3.2：辨認詞典建立前的處理流程

以 3000 的詞條空間來收錄 subword，能使 DM 詞組的涵蓋率達 100%，人名及綴詞也都涵蓋 96%以上，加上 57000 詞條的涵蓋率已涵蓋了 97.74%的語料，其餘 2.26%詞條則視為 OOV；因此，可假設此辨認詞典合乎所預期的，詞典中大部分所收錄的應是高詞頻的一般常用詞，而利用少許的空間收集了 3000 筆的 subword，於第二級再加入此三類的語言模型，給予新的語言模型分數將之辨認出來。

表 3.9：辨認詞典之涵蓋率

詞典詞數	涵蓋率
57000	92.06%
57000+3000	97.74%

以圖 3.3 列出辨認詞典的資料分析，如上述所分析，57000 高詞頻詞條之中其詞彙可兼顧多重角色，高頻詞可能為綴詞中的詞幹，或可和姓氏組合成人名，DM 詞組中的 subword 亦可能同時為高頻詞和綴詞的詞幹。對此詞典更加詳細分析其組合成份，詞典中與三類相關的詞彙約佔了 33%，排除未拆解的完整詞彙，更加確信許多構詞單元是高詞頻的常見詞，所收錄常用詞即同時兼顧三類詞的收集，達到在有限額度的空間中有系統的收集詞彙並且建立詞典。

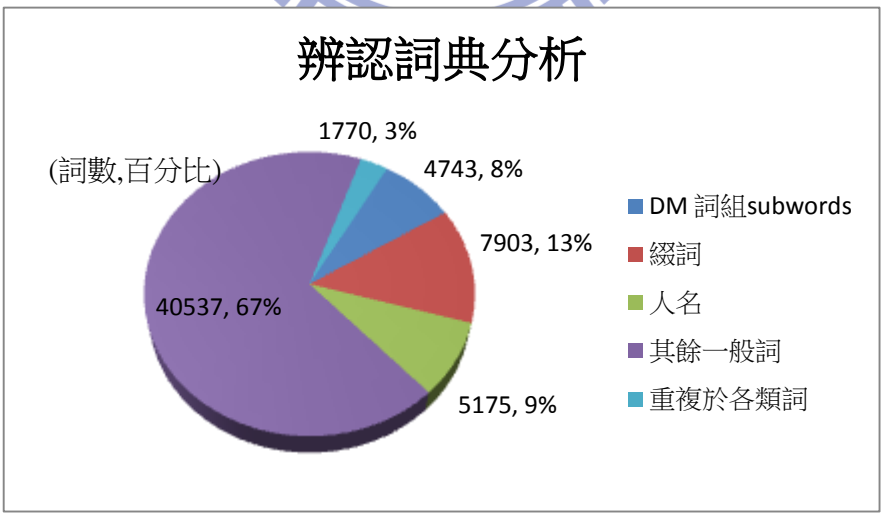


圖 3.3：辨認詞典分析

表 3.10 列出詞典的字數分布，詞典中以二字詞、三字詞為主，所佔比例約 87.7%；一字詞數量比以往減少許多，不再收錄過多的一字詞來增加混淆度；而多字數的詞彙量也減少許

多，起因於較長的定量複合詞詞組已經以構詞單元形式收集，大大減少了多字詞。

表 3.10：辨認辨認詞典字數分佈

字數分佈								
字數	一字	二字	三字	四字	五字	六字	七字	總量
數量	3145	37985	14750	4048	166	27	7	60128

3. 2 階層式詞組語言模型

於上一節三類詞的介紹，利用文字的規則特性將之拆解收錄短單元，同樣地，在此也依文字特性找出各類詞的構詞規則以建立更精細的詞組 LM。以各詞彙的組成結構來建立模型，因此詞彙的分群也是重要課題，本論文採用 class-based approach 構想【14】：在同一類別內的詞彙能使用相同的機率，解決部分資料稀疏的問題；對詞彙進行的分類，分析各類詞彙的組成方式及內部結構，不僅幫助詞典詞彙的收集，也利用其性質找出特有的結構關係並且有系統建立詞組 LM。

將三類分類的目的為透過有效率的方法來幫助辨識出長詞，在辨識時使得詞彙的競爭更可靠，競爭的數量由詞典的大小降為該類別下的數目，而類別之中的機率分布甚至會來得更加可信，構出有意義的長詞的可能性也增高。

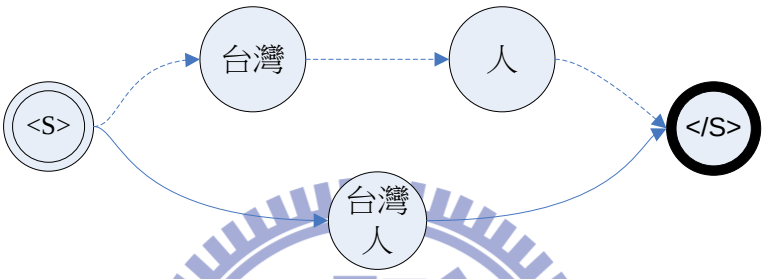
論文中針對 DM 做了更精細的研究，透過分析 DM 與其前後修飾詞、名詞等結構關係建立更大的 DM 詞組，藉此增進 DM 詞組語意的完整性，並依據其結構內詞彙間的關係建立強健的語言模型。已知 DM 詞組各有其特性，如詞組中的量詞「公尺」與「公斤」，因其語意的不同，會影響 DM 內部詞彙及其前後詞彙的預測，藉以這樣的特性能幫助語音辨識構出完整的長詞組單元。而兩類的人名、綴詞類則初步為其建立簡單的詞組語言模型，以列舉方式來建立詞組語言模型。

3.2.1 綴詞

綴詞的分群原則是以詞綴來作區分。中央研究院詞庫小組分析了常用詞首、詞尾字，定

義其詞義並舉例了其用法；我們仿其概念，認為綴詞中的詞綴涵義各有相異處，與詞綴相關聯的詞幹應有類似特性，如台北人、台灣人、北部人...等等，將此視為同一集合會比全部綴詞皆為一大類別來的適當，以其來訓練詞彙間的機率來的更可靠；目前共有 303 個綴詞類別。

本研究會將 lattice 上的「台灣」與「人」組合後至綴詞表中查詢，當「台灣人」在綴詞表之中表示可組合回「台灣人」，如圖 3.4，也產生新的辨識路徑，而後將重新計算該詞與前後詞彙間的機率，即統計新路徑實線的機率分佈。



組合出之綴詞

圖 3.4：新辨識路徑

3.2.2 中文人名

尚未觀察到人名的特性，難以將人名加以分類，目前人名歸為一類，不進行分群，和綴詞一樣將在 lattice 上相鄰的詞組合成長詞至表中搜索比對，有符合表中之長詞者則產生此長詞的新路徑。

其缺點為人名詞彙較特殊，難以分群，使得人名機率易隨著人名詞條數量的增加而變小，鑑別度將降低，機率預估相較於綴詞較不可靠，而一旦人名出現次數少時需藉由 smoothing 方式給予合理的機率來得到新的語言模型分數。

3.2.3 DM 詞組

DM 詞組數量龐大，其構成方式各有差異，將之分群至相似構詞型態的類別中，提升詞組內部的一致性。DM 詞組先以 POS 組合結構來分群，考慮 DM 詞組內部的複雜結構及所代表的不同語意，建立多個類別不只降低 DM 詞組內部混淆程度，也提高了詞組與前後詞彙

的關聯性；以此可不單只考慮 bigram 嘗試建立對特定詞彙間的機率關係，譬如量詞與名詞的關聯性，或者連續 DM 詞組相接的關係。

DM 詞組採用 FST 的架構建立其語言模型，分析詞組的內部構造並且掌握各組合單元，詞組單元必須有正確的分類，以便建立各個詞組單元之間的關聯性，每一詞組單元即為 FST 中的狀態，再透過 DM 詞組的詞性組合規則建立各狀態間的關係並轉換成一個 FST 機率模型，最後 DM 依詞性組合共細分為 112 群，詳細如[附錄三]說明。

(1)量詞：包含標準量詞、動量詞、暫時量詞、準量詞、跟述賓式合用的量詞、容器量詞、個體量詞、群體量詞等八類。

(2) 定詞：包含數詞定詞、特指定詞、指示定詞等三類。

a. 數詞定詞：幾、數字及其組合。

b. 特指定詞：前、後、另、其餘、這些...等。

c. 指示定詞：這、那、此...等。

(3) 修飾詞： 包含前修飾詞、後修飾詞、定詞量詞間修飾詞等三類。

a. 前修飾詞：約、近、不到...等。

b. 後修飾詞：左右、不等、以上、整...等。

c. 定詞量詞間修飾詞：多、餘。

DM 詞組所處理的範圍擴展到包含修飾詞、名詞等的完整語意單元，其中數詞由寫法轉讀法正規化之後型態更為多元，例如：純數字串、單一數字、十百千階層結構組合、含修飾詞...等，數詞型態細分了 17 項，每一群再建立其一層的 FST 架構。我們認為有些數詞的變化會侷限於量詞，例如月份其所搭配的數詞多為一至十二；甚至再有修飾詞時也會影響數詞的選擇，因此我們為 DM 詞組嚴以分類並建立個別模型，舉例如下列圖。

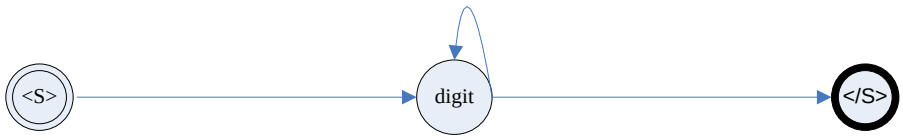


圖 3.5：數字串 FST

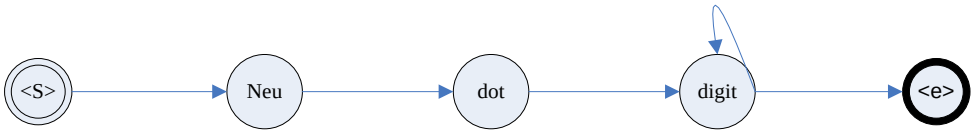


圖 3.6：含小數點的數詞 FST

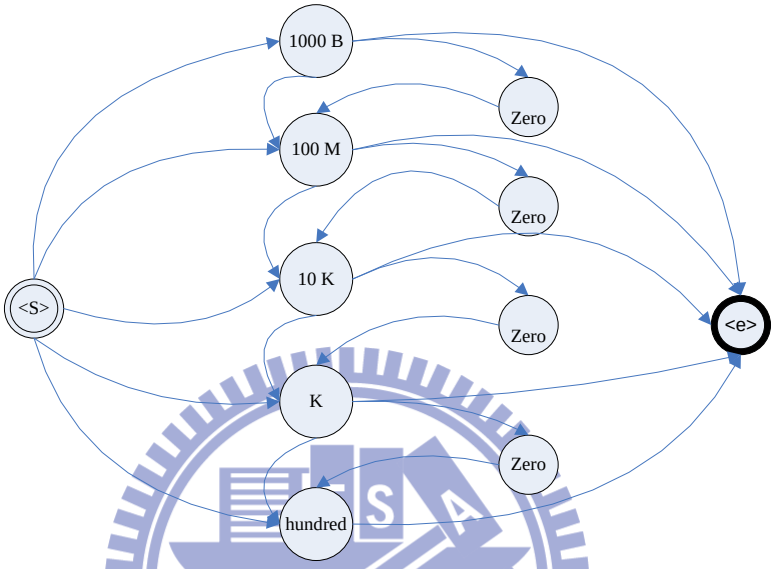


圖 3.7：複雜結構數詞 FST

訓練語料中的 DM 詞組以 data-driven 方式來建立語言模型，把詞組單元轉換成狀態序列，將這些序列依次數輸入至 FST 中產生 FST 狀態序列，以下舉〔這座、六十五年、三十五座、這份〕為範例建立 DM 詞組的 FST 架構：

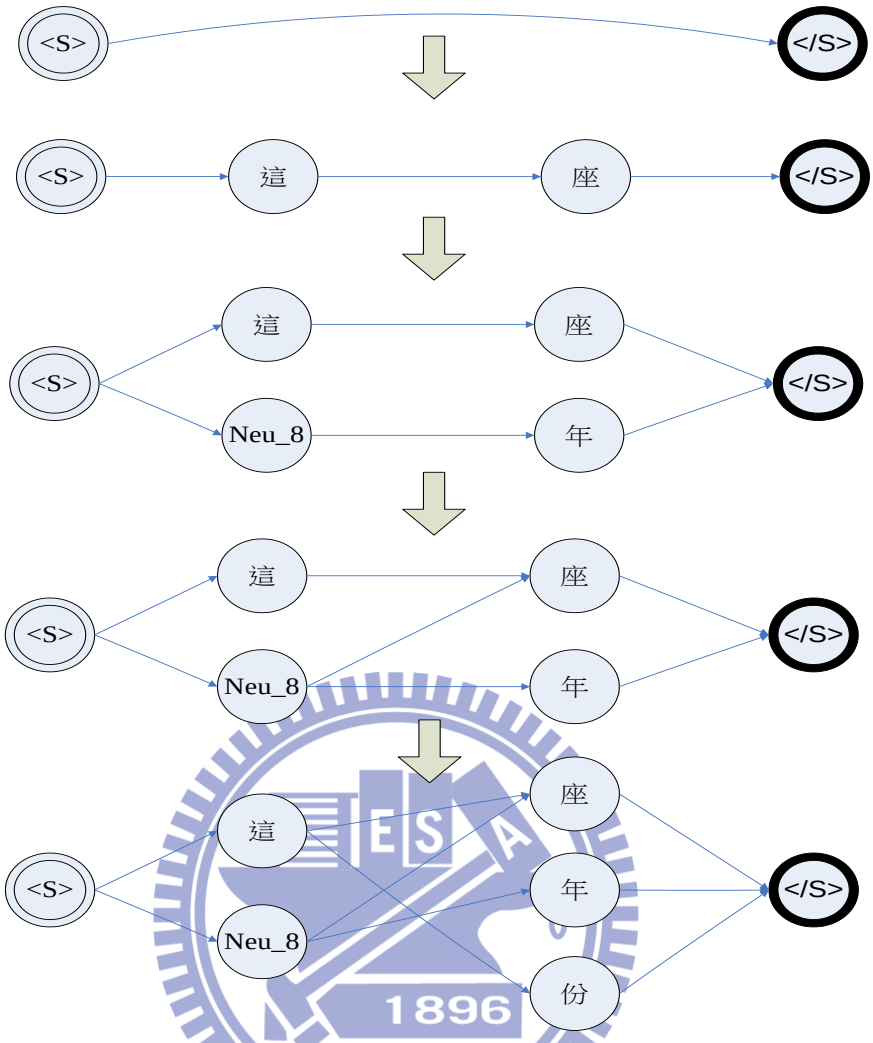


圖 3.8：DM 詞組 FST 建立流程

以上圖為例，「這」共用了狀態，共用狀態的優點為當一詞串出現次數少時能夠一起分享共同狀態的轉移機率，不會因詞串的出現次數多寡，或是測試語料中存在訓練語料沒有的詞串而難以構詞，只要是合理的詞串皆能夠產生新的路徑；而 Neu_8 表示「六十五」與「三十五」皆屬於 17 類之數詞型態中的第八個 class。另外，避免訓練過程中路徑產生迴圈造成錯誤，加入一個防止迴圈產生的機制，即訓練過程中每新增一個狀態，會先檢查是否產生迴圈，若會產生則新增一個新的狀態，以避免因迴圈造成該路徑有相當高的轉移機率，甚至無法走出 FST 結構。

最後定量複合詞詞組的模型架構為下圖 3.9 所示，建立成階層性架構之語言模型，包含第一層的 Inter Word Model 和第二層的 Intra DM chunk Model。為得到詞組語言模型的機率，

將定量複合詞詞彙視為類別，訓練此類別與其他詞彙間的 *N*-gram 機率，此為外部機率 (inter-word probability)；Intra DM chunk Model 中，皆以上方所述建立定量複合詞詞組的 FST 模型及內部數詞的 FST 模型，並且以訓練語料統計狀態間的轉移次數，得到內部機率 (intra-word probability)。

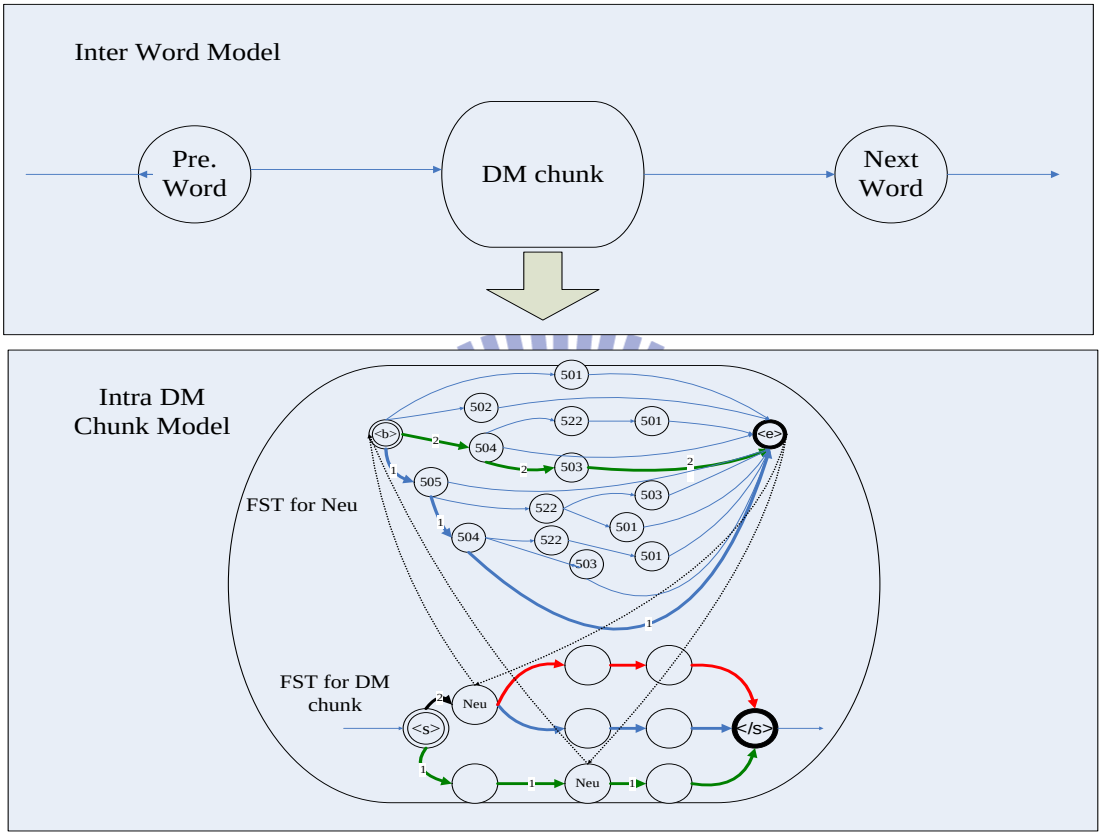


圖 3.9：DM 詞組 FST 模型架構

3.3 詞組語言模型機率

於 3.2 節說明建立 DM 詞組、人名、綴詞之語言模型，最後在辨認時必須使用新語言模型的機率來找出辨識結果。本研究將此三類詞分群而建立個別語言模型，其機率的計算概念以圖 3.9 所示分別考慮了外部機率和內部機率，由以下詳細敘述之：

(1) 外部機率—詞組外部機率的計算方式：

若詞彙為高頻詞則相信其可被信賴，將採用其 word bigram 機率；反之，對相對於

低頻的定量複合詞詞組、人名、綴詞，這三類的詞彙將之以類別標記，由此統計語料中詞彙間的關係，由此得到所有詞彙的機率值，包括如：詞和類別之間、類別和類別之間、詞和詞之間的 class bigram 機率。

(2) 內部機率—詞組內部機率的計算方式：

a. 綴詞(MD)：

將綴詞依詞綴細分為多個類別，非在單一類別中與所有綴詞一起競爭，之後統計某一綴詞於該類別內所出現的機率並搭配 good-tuning smoothing 作為內部機率。

b. 人名(PN)：

所有人名是為一集合，如同綴詞內部機率計算方式，統計各別人名於集合之中所出現的機率。

c. 定量複合詞(DM)：

依 3.2 節所述第二級構詞之 DM 詞組 FST 模型，定義了每個狀態所對應的 subword 集合，再透過構詞規則建立狀態之間的關連性(state transition)，最後訓練此架構的狀態轉移機率，其內部機率計算方式即為 FST 中各狀態轉移機率之乘積，以 lattice 上 DM 詞組 subword 詞串「約 一百 五十 人」為例，所對應的狀態序列為「<s> 約 Neu_9 人 </s>」，內部機率 $P(W_n|C_n)$=P(約|<s>)・P(Neu_9|約) ・P(一百五十|Neu_9)・P(人|Neu_9)・P(</s>|人)，以達到構詞目標為「約一百五十人」。

透過這些機制重新配置語言模型的分數，藉以辨識出正確的詞彙，並壓抑錯誤的辨識路徑藉以提高辨識之正確性。以下列出這兩種機率計算方式所得到新的語言模型機率值並舉例：

$$P(W_n \mid W_{n-1}) = \begin{cases} P(W_n \mid W_{n-1}) & , \text{for } count(W_{n-1}) \geq k \\ d \cdot \mathrm{P}(\mathrm{W_n} \mid \mathrm{W_{n-1}}) & , \text{for } \mathrm{count(W_n)} \geq \mathrm{k}, \mathrm{count}(\mathrm{W_{n-1}}, \mathrm{W_n}) = 1 \\ \alpha \cdot \mathrm{P(W_n)} & , \text{for } \mathrm{count(W_n)} \geq \mathrm{k}, \mathrm{count}(\mathrm{W_{n-1}}, \mathrm{W_n}) = 0 \\ P(C_n \mid C_{n-1}) \cdot P(W_n \mid C_n) & , \text{for } count(W_{n-1}) < k \ , C_n, C_{n-1} \in \{\mathrm{DM, MD, PN}\} \end{cases}$$

$$P(C_n \mid C_{n-1}) = \begin{cases} P(C_n \mid C_{n-1}) & , \text{for } C_n, C_{n-1} \in \{\mathrm{DM, MD, PN}\} \\ P(W_n \mid C_{n-1}) & , \text{for } C_n \in \text{other } and \ C_{n-1} \in \{\mathrm{DM, MD, PN}\} \\ P(C_n \mid W_{n-1}) & , \text{for } C_n \in \{\mathrm{DM, MD, PN}\} \ and \ C_{n-1} \in \text{other} \end{cases}$$

$$P(W_n \mid C_n) = \begin{cases} \prod_{j=1}^{L} P(S_j \mid S_{j-1}, C_n) & , \text{for } C_n \in \mathrm{DM}, \ S_j = state\ j, \ L = Length\ of\ DM \\ \dfrac{count(\mathrm{W_n})}{\sum_i count(\mathrm{W_i})} & , \text{for } C_n \in \mathrm{MD \text{ or } PN} \end{cases}$$

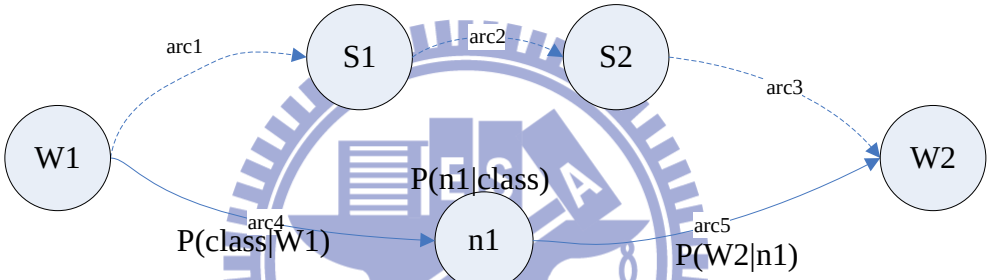


圖 3.10：新辨識路徑上之分數配置

n1 詞之機率估計為 P(n1|W1) = P(class|W1)·P(n1|class)，當 n1 詞彙不是常見高頻詞時，則以 class bigram 來計算新辨識路徑 arc4 的機率，P(class|W1) 是 inter probability 描述詞與類別之間的機率，P(n1|class) 是 intra probability 描述類別當中某一詞的機率；當該詞為定量複合詞詞組時，其內部機率為 subword 單元所對應至 FST 之中的狀態轉移機率之相乘，人名與綴詞則至所列舉的詞彙表中來得到該詞機率。

3.4 使用兩階段架構實踐階層式語言模型辨認器

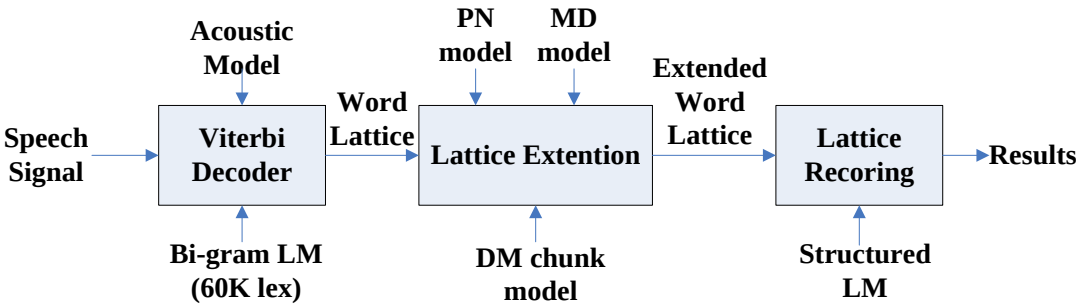


圖 3.11：階層式語言模型之語音辨認系統架構

因系統複雜度之考量辨認系統使用兩階段(two-stage)架構，先於第一級中之辨認詞典在有固定詞彙數量限制的下，希望能收錄較少量的詞彙來涵蓋更多的語料，使得第一級 LM 初步辨認時所產生的 word lattice 上之詞彙涵蓋率越高越好，再於第二級中為具有規則特性的特定詞類建立語言模型。

本研究之主軸為對定量複合詞、人名、綴詞建立更精細的語言模型，在此系統中第一級產生 lattice 後在第二級中加入這三種詞類的語言模型，透過這些語言模型得到更能詳細描述三類詞的資訊，重新訓練語言模型機率並於 Extended word lattice (EWL)上加入新的語言模型分數(rescoring)，最後重新辨識找出最佳的辨認結果。

第四章 實驗結果與分析

4.1 實驗目的

本研究對定量複合詞、人名、綴詞這三類詞彙拆解，以收錄拆解的小單元來取代原有的長詞，此方法詞典中能涵蓋至大多數的詞彙，藉此降低 OOV rate；再建立三類詞的語言模型並 rescoring，透過有效率的方法來輔助辨認出較有意義的長詞單元。

4.2 辨識語料

使用 TCC300 作為辨識語料，TCC300 為朗讀式麥克風語料庫，是由國立台灣大學、國立成功大學、國立交通大學各自擁有之語料庫集合而成，各校錄製之目的是為語音辨認研究。台灣大學語料庫主要包含詞以及短句，文字經過設計，考慮音節與其相連出現之機率，共 100 人；成功大學及交通大學為長文語料，其語句內容由中研院提供之 500 萬詞詞類標示語料庫中選取，每篇文章包含數百個字，再切割成 3 至 4 段，每段至多 231 字，分別各 100 人，且每人所朗讀之文章皆不相同。

TCC300 語音資料庫時間總長約為 26.4 小時；我們將所有可用語料的十分之九做為訓練語料，十分之一歸於測試語料，其中訓練語料的時間大約有 24 小時。聲學模型採用隱藏式馬可夫模型(Hidden Markov Model, HMM)，求取的 MFCC 語音參數為 38 維度的參數向量。而聲學模型以音節(syllable)為單位，共有 411 個模型，其中每一模型有 8 個狀態，而每一個狀態平均以 10 個 mixture 來描述。

語音參數求取時所使用之系統參數如下表 4.1 所示：

表 4.1：參數抽取設定檔

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

以 TCC300 中之 226 長句作為測試語料，本研究中第一級語言模型的好壞，進而影響辨識的效能，然而測試語料 TCC300 中三類詞的數量，也同樣會影響辨識效果，對此測試語料先行統計定量複合詞、人名、綴詞，表 4.2 列出這三類的比例數量：

表 4.2：TCC300 測試語料的詞類統計

TCC300 測試語料				
TCC300 測試語料	**詞條數**	**各類所佔比例**	**總量**	**各類所佔比例**
	4925		14746	
詞類細分				
定量複合詞詞組	339	6.88%	606	4.11%
人名	106	2.15%	218	1.48%
綴詞	202	4.10%	416	2.82%
一般詞	4278	86.87%	13506	91.59%

4.3 語言模型評估

在未以語音辨識系統測試辨識率之前，先行以複雜度來評估語言模型。語言模型帶有詞和詞相接的機率資訊，利用此訊息來預估下一個詞彙，當可相接的詞彙多，則表示預估至正確的詞彙次數將增多，較不易找到正確答案，複雜度也增大；因此，單單以評估語言模型而言，認為其複雜度應該越低越好。表 4.3 為 General 語言模型及第一級語言模型之複雜度：

表 4.3：語言模型之複雜度

Bi-gram語言模型	複雜度(PPL)
General LM	464.4016
第一級 LM	402.0596

4.4 三類詞於 word lattice 上之涵蓋率

本研究評估第一級 LM 所產生的 word lattice 上，定量複合詞、人名、綴詞的 subword

有多少能組合回原本的長詞。在 4.2 小節中統計了測試語料中的三類詞數量，由此可先行了解最多能夠正確辨識的數量及上限，進而對第一級 LM 產生的 word lattice 分析，統計理想上最佳路徑上有多少可以辨識回原始的三類長詞，並加以計算其涵蓋率。

- **General LM 之 lattice 上的涵蓋率**

表 4.4：General LM 之 lattice 上的的三類詞彙涵蓋率

詞彙分類	TCC300	最佳路徑	涵蓋率
定量複合詞	606	525	86.63%
人名	218	55	25.22%
綴詞	416	352	84.62%

由上表，General LM 的詞典所收錄的詞彙於 lattice 上對定量複合詞和綴詞都有達到一定的涵蓋率，人名的涵蓋率偏低是由於詞典是收錄高詞頻詞彙為主，人名用詞特殊較不具一般詞特性，僅收錄常見的高詞頻人名，因此 TCC300 測試語料中有詞典未收集的人名將較難辨識出來。

- **第一級 LM 之 lattice 上的涵蓋率**

表 4.5：第一級 LM 最佳路徑的上三類詞彙涵蓋率

詞彙分類	TCC300	最佳路徑	涵蓋率
定量複合詞	510	453	88.82%
定量複合詞詞組	96	90	93.75%
人名	218	117	53.67%
綴詞	416	402	96.63%

建立第一級 LM 的詞典當中原本就收錄了許多三類詞的詞彙，包含未拆解長詞以及三類詞 subword 單元；因此，lattice 上的詞彙除了原本就是完整的三類詞之外，還有可再經由串接成為三類詞的詞彙，其三類的涵蓋率都比 General LM 來的高。而人名詞彙特殊較不像一般詞、不容易在 lattice 上產生會降低人名的辨識率，而與 4.4 表比較人名涵蓋率提升幅度明顯許多，可初步評估以少許的詞典空間來收錄 subword 短詞可以提高三類詞的涵蓋數量，理想上可辨識的詞彙數量也將相對地增加，人名之正確辨識量的成長會較其他兩類更為顯著。

表 4.6：最佳路徑上之三類詞的 subword 數量分析

	一個 subword	二個 subword	三個 subword	三個 subword 以上
DM	0	386	61	6
DM 詞組	0	0	72	18
人名	55	25	37	0
綴詞	350	52	0	0

由表 4.6，可推論此 TCC300 測試語料中的三類詞可能為型態簡單且大部分是高頻的詞彙。Lattice 上可構回的人名數量不多，但仍有將近一半的人名為完整長詞，其餘由 subword 短詞串構成人名，綴詞此特性更為明顯許多；然而，定量複合詞絕大多數以兩個 subwords 所組成，TCC300 中定量複合詞詞組的結構相對簡單，以「定詞+量詞」為主，可能為「數字+量詞」或是「不含數字的定詞+量詞」為多數組合，反之構入的修飾詞或名詞並不多所致。

- **涵蓋數量之比較**

表 4.7：最佳路徑上之正確的三類數量比較

	General LM	第一級 LM	最佳路徑上可構回長詞的數量差
定量複合詞詞組	525	543	18
人名	55	117	62
綴詞	352	402	50

由上表 4.7，測試語料 TCC300 所含的綴詞和定量複合詞於兩個語言模型已達到不錯的涵蓋率，因為其辨認詞典皆收集了大多數的高詞頻詞彙，所以詞彙的涵蓋程度理當不會有誇張的落差；第一級 LM 之詞典中則收錄了三類詞 subword，藉由在限制詞典詞彙數量下，剔除部分長詞空間來收錄 subword 再提升詞彙的涵蓋率，以增進詞彙之可辨識上限。而這兩個不同的 LM 能在 lattiice 上找到的最佳辨識路徑，各自所能達到之最高辨識率以下表 4.8 所示，兩者不同 LM 的詞彙辨識率最高皆將近 90%，其中第一級 LM 之詞彙涵蓋率較高其理想之辨識率也比較高：

表 4.8：最佳路徑的辨識效能比較

	Deletion	Substitution	Insertion	Accuracy	total
General LM	85	1338	429	87.44%	14746
第一級 LM	81	1170	328	89.29%	14746

4.5 辨識結果分析

此小節將比較不同的語言模型之辨識效能，分別以字元、詞為辨識單元來評估辨識結果；另外，也統計辨識結果中三類詞之辨識能力並分析其構詞。

4.5.1 辨識效能結果

General LM 以收錄高詞頻之六萬詞為主之辨認詞典所訓練的語言模型;而另一方法為先建立一個包含三類詞之 subword 構詞單元的第一級 LM，再為三類建立更精細之語言模型，最後構成階層式 LM 來進行辨識。在此，比較兩個不同的語言模型其辨識率：

表 4.9：字元(character)辨識效能比較

	Deletion	Substitution	Insertion	Accuracy	Total
General LM	373	6446	223	73.41%	26486
階層式 LM	363	6411	224	73.58%	26486

表 4.10：詞(Word)辨識效能比較

	Deletion	Substitution	Insertion	Accuracy	Total
General LM	583	4017	626	64.56%	14746
階層式 LM	683	4056	474	64.65%	14746

由以上兩表格發現，兩個語言模型之辨識率的差異性極小，雖然階層式 LM 之辨識率皆稍微高於 General LM，但此方法之假設的整體效果並沒有比 General LM 來的顯著，為找出影響辨識率之因素，在下一小節中對辨識結果分析。

4.5.2 辨識結果之三類詞分析

表 4.11：辨識結果之三類詞分析

	TCC300 測試語料	General LM		階層式 LM	
	數量	數量	正確率	數量	正確率
定量複合詞	510	428	70.63%	349	76.87%
定量複合詞詞組	96			43	47.78%
人名	218	46	24.31%	76	34.86%
綴詞	416	321	77.16%	339	81.49%
總量	1240	795		807	

4.5.2.1 定量複合詞

辨識結果中，DM 詞組的正確辨識數量較 General LM 來得少，由結果分析得 General LM 之辨認詞典中已收了許多 DM，使得辨認結果中的 DM 大部分已辨認正確；相較之下，第一級 LM 之辨認詞典中則收錄 DM 的 subword 單元，未收錄任何一個完整的 DM，要正確地辨識出 DM 必須靠 DM 詞組語言模型來輔助辨識。

而本研究之目的為欲重新定義中文語音的辨認單元，以含有更多語意資訊的詞彙或詞組作為語音辨認系統效能評估的依據。基於此概念，第一級 LM 比 General LM 多了以下之優勢，且於下表 4.12 舉例。

以下 A、B 分別代表 General LM 和階層式 LM 之辨識結果：

(1) 可辨識出較長且語意更為完整的 DM 詞組：長度越長的詞組於語料中出現次數相對而言較少，加入詞組 LM 能幫助辨識出此類的 DM。

(2) DM 詞組語言模型之辨識可減少相近音詞之混淆度：當詞典中未收錄正確詞彙時容易辨識出其他相近音詞串，搭配良好的 LM 機率可辨識出正確的詞彙，減少相近音混淆的問題。

(3) 提升 DM 其前接、後接詞之正確率：當 DM 正確辨識出來時，間接影響前接詞與後

接詞之預估，凸顯鄰近詞彙之正確辨識。

表 4.12：DM 辨識結果分析 1

較長的 DM 詞組
新山 水庫 加高 **十五公尺之後** 我國 民法 **第三百七十三條規定** 買賣 標的物
減少相近音詞之混淆度
答案：八十一學年度 入學 的 國中生 A：**八十七學年度** 入學 的 國中生 B：**八十一學年度** 入學 的 國中生
答案：儲水 一千萬噸 A：儲水 **一千萬 盾** B：儲水 **一千萬噸**
答案：七十九年 第二期 農作物 A：七十九年 **第二季** 農作物 B：七十九年 **第二期** 農作物
提升 DM 其前、後詞之正確率
答案：南市 郭朝武 等 四名市議員 A：但是 **過 超過 二十名 市議員** B：南市 **郭朝武 的 四名市議員**
答案：民法 **第三百七十三條規定** 買賣 A：民法 **比賽 白旗 十三條 規定** 買賣 B：民法 **第三百七十三條規定** 買賣
答案：**一位老師** 全天候 **帶領** A：**一位 老師** 全天候 **待命** B：**一位老師** 全天候 **帶領**

針對 DM 詞組中分析，有 43%左右的 DM 詞組可以完整辨認正確，和 General LM 之辨識結果比較，階層式 LM 能辨識出較長的 DM 或 DM 詞組，是由於 General LM 雖然收集了許多 DM 詞，但長度越長的 DM 所出現的次數較少而未收錄至詞典當中，故較難將之辨識正確，而階層式 LM 能夠輔助長 DM 詞彙或詞組正確地辨識出來。

階層式 LM 雖然比 General LM 多辨識出 49 筆 General LM 未辨識正確的 DM/DM 詞組，並且多數為長度較長的 DM，然而歸納出階層式 LM 未能明顯改善 DM 辨識率之原因為：

(1) 二字詞或三字詞的短 DM 辨識效果較差：短 DM 容易與其他的相近音詞彙造成混

淆，或有錯誤的語言模型機率較為強勢並且音相近時往往會造成辨識錯誤。

(2) 詞組語言模型之分數較強勢；以下面之十一甲土地為範例，「甲」與「土地」有高度關係，而當鄰近詞彙音相似會容易辨識出錯誤的詞組。

以下列出辨識錯誤之範例：

表 4.13：DM 辨識結果分析 2

短 DM 之相近音辨識錯誤
答案：展開 **第一天** 會內賽 B：展開 **地點** 會內賽
答案：下午 **三時** B：下午 **三十**
答案：十**幾**個 B：十**七**個
答案：有 **幾種** 不同 B：有 **機種** 不同
答案：要求 **這種** 作法 B：要求 **折衷** 作法
詞組語言模型強勢
答案：徵收 的 **事宜 只要 土地** B：徵收 的 **十一甲土地**
答案：孩子 **是 二 足歲 到** 不滿 B：孩子 **十二座隧道** 不滿
答案：**時值 年終** 公務員 B：**四十年中** 公務員
答案：立刻 **施救 大冠鷲** B：立刻 **十九大碗酒**
答案：對象 **是 前 林姓 市議員** B：對象 **四千零七十一人**
答案：**而且 漁民** 若 違反 B：**二千餘名** 若 違反

General LM 之辨識產生了 122 筆不應該是 DM(over generate)的辨識結果，而詞組語言模型則為 155 筆，第一級 LM 之辨識詞典中未收錄完整 DM 只包含 DM subword 單元，然而其 over generate 的情況比 General LM 更多，推測部分詞組語言模型過於強勢導致

這樣的結果；然而，由上表 4.12 相信資訊完整的詞組能幫助提升前、後詞彙之辨識率，因此適當的 DM 詞組語言模型之分數配置對辨識之結果影響很大。改善以規則訓練之 DM 詞組語言模型，甚至將 subword 正確但非完整的 DM 詞組辨識成功，對詞組其前、後詞之辨識正確之可能性也相對提高，進而連帶提升整體之辨識率。

4.5.2.2 中文人名

由表 4.11 得人名之辨識數量相較於 General LM 有顯著的提升。General LM 辨識正確的人名僅 46 筆，佔全部人名數量 21%；而第一級 LM 之最佳路徑上的人名涵蓋率較高，相對的後續階層式 LM 能辨識出的人名數量也增多。

依據表 4.6 所分析共有 55 筆完整人名收錄於詞典當中，而階層式 LM 最後正確辨識數量共 76 筆，其中有 34 筆由 subword 組合辨識成完整之人名，舉例如表 4.14，由此可相信將人名拆解再建立其 LM 是個有效的方法，一方面提升詞彙涵蓋率並且增進辨識率。由人名之錯誤分析中可發現，有些人名詞彙被辨識成相近音的錯誤人名有 43 筆，甚至辨識為相似度極高但只有一字錯誤的人名；因此，若為人名建立更精準的 LM，透過鄰近詞彙間的關連性，如頭銜、職稱與人名的關係，相信能辨識出更多且正確的詞彙。

表 4.14：人名辨識範例

subword 組成之人名辨識正確
郭 振興
郭 朝武
楊 健 生
許 賜 雄
相近音之辨識錯誤
陳枝福(正確) vs. 陳 志 福
陳見生(正確) vs. 陳 建生
何振梁(正確) vs. 何 生 亮

4.5.2.3 綴詞

由表 4.11 知兩個不同語言模型之綴詞辨識數量差異不大，兩者的綴詞辨識率皆為 80% 左右，都達到不錯的辨識率。綴詞之分析中，階層式 LM 之辨識效果仍比 General LM 來得好，為綴詞建立其 LM 是有用的，舉例如表 4.15。依詞綴特性將綴詞分類，各類別保有不同詞綴間的特性，並如同人名一樣利用該詞彙與鄰近詞彙間的關連性，來訓練更精細的 LM，不僅幫助綴詞本身之辨識而其鄰近詞也更容易辨識成功。

表 4.15：綴詞辨識範例

subword 組成之綴詞辨識正確
志願 軍
集會 所
羽 球隊
相近音之辨識錯誤
市議員(正確) vs. 十億 元
音樂化(正確) vs. 一元化
私有地(正確) vs. 市有地
塑膠化(正確) vs. 塑膠 花
詞組 LM 輔助辨識
答案：合作金庫 **男子 羽球隊 不但** 在 A：**男士 與 球隊 不得** 在 B：合作金庫 **南市 羽球隊 不但** 在
答案：**美西區 理事長 陳進財 到** A：**沒 興趣 理事 長成 警察 來到** B：美西區 理事長 **陳進昌 來到**

第五章結論與未來展望

5.1 結論

由目前的辨識結果分析可得知，使用 General LM 及階層式 LM 這兩種不同 LM 的辨識效能差異不大。本論文之最終目的為一方面減少中文 OOV 問題，另一方面為如同本研究中之定量複合詞、人名、綴詞，這些具有語言學規則特性的詞類，歸納其規則所建構出更精細的 LM 是有效的，全面性的描述該詞類的特性，藉此辨識出包含更多語意、語法及結構之詞彙甚至是詞組來同時提高辨識率。

而探究目前之辨識效能不理想的原因：

(1) 訓練語料及測試語料之資料庫年代太相近：訓練 General LM 的六萬詞典中收錄了訓練語料中非常多的高詞頻 DM，因而 General LM 有很大的機率能將測試語料中，一般認為不大可能辨識正確的 DM 或特定時間詞(如；七十九年)正確地辨識出來。

(2) 測試語料中 DM 詞彙數量不多：DM 詞彙約佔 4%，並且 DM 詞組的數量很少，不容易凸顯出本研究之階層式 LM 的優勢。

5.2 未來展望

未來若能針對定量複合詞、人名、綴詞這三類來重新設計測試語料，並且以資料量更充足的語料來訓練語言模型，同時其中詞組語言模型之機率的預估也將更為準確，本研究之方法的辨識效能相信能夠大幅改善。

目前階層式 LM 仍有改善的空間，詞彙正確地辨識之外，也希望以辨識正確之完整的詞組，建立詞組與詞彙甚至是詞組與詞組間的緊密關係，掌握更大且更多資訊的辨識單元，例如連續 DM 或 Nd 的辨識：七十八年_十二月_二日，當辨識成一個大詞組將可得知同性質詞彙的範圍，以這個詞組再建立和其他詞彙的關連性；甚至是得到詞組內某個 subword 與詞組

範圍外特定詞彙的關係，例如：詞組內的量詞通常會影響鄰近詞彙的用詞。

當此方法成熟時，未來可將此概念運用到中文中同樣具有特定語言學規則特性的詞類，個別詞類有其語言模型更精準的描述各詞類的特性，以輔助中文之辨識效能；再且，未來將引進韻律模型的輔助，於辨認時加以確認詞彙單元是否真為一個有意義的詞彙單元，用以克服中文語音辨認長久以來的詞彙定義的問題。

參考文獻

【1】 O. Scharenborg, S. Seneff, L. Boves, “A two-pass approach for handling out-of-vocabulary words in a large vocabulary recognition task”, Computer Speech and Language 21 (2007) 206-218.

【2】 Y. C. Pan, L. S. Lee, “Lexicon Adaptation with Reduced Character Error (LARCE) — A New Direction in Chinese Language Modeling”, Interspeech, Antwerp, Belgium, August 2007, pp.610-613.

【3】 S Lee, K Hirose, and N Minematsu, ”Incorporation of prosodic modules for large vocabulary continuous speech recognition, ” in Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding, 2001.

【4】 F Gallwitz, A Batliner, J Buckow, R Huber, H Niemann, and E Noth, “Integrated recognition of words and phrase boundaries”, ICSLP1998.

【5】 D. Vergyri, A. Stolcke,VRR. Gadde, L. Ferrer, E. Shriberg, “Prosodic knowledge sources for automatic speech recognition, ” ICASSP 2003.

【6】 J. T. Huang, L. S. Lee, “Improved Large Vocabulary Mandarin Speech Recognition Using Prosodic Features,” Speech Prosody 2006, Dresden, Germany.

【7】 X. Huang, F. Alleva, H.W. Hon, M.Y. Hwang, K.F. Lee, and R. Rosenfeld, “The SPHINX-II speech recognition system: An overview,” *Computer, Speech, and Language*, vol. 2. pp. 137–148, 1993.

【8】 Peter F. Brown, Vincent J. DellaPietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. “Class-based *N*-gram models of natural language,” *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

【9】 J. R. Bellegarda, “A multispan language modeling framework for large vocabulary speech recognition,“ *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 6, no. 5,

pp.456-467, 1998.

【10】中央研究院的中文斷詞系統，http://ckipsvr.iis.sinica.edu.tw/。

【11】J. Lafferty, A. McCallum, and F. Pereira. “ Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, In Proc. of ICML, pp.282-289, 2001.

【12】中研院詞庫小組出版物，http://godel.iis.sinica.edu.tw/CKIP/publication.htm#t2。

【13】中央研究院詞庫小組，中央研究院平衡語料庫的內容與說明，詞庫小組技術報告 # 93-02，台北，1995。

【14】S. Onishi, H. Yamamoto, and Y. Sagisaka, “Structured language model for class identification of out-of-vocabulary words arising from multiple word-classes,” Eurospeech 2001.

附錄一：量詞表

量詞分類	量詞集合
個體量詞	本 把 瓣 部 柄 床 處 期 齣 場 朵 頂 堵 道 頓 錠 棟 檔 封 幅 發 分 服 個 根 行 件 家 戶 架 卷 具 闕 節 句 屆 捲(兒) 劑 隻 尊 盞 張 枝 樁 楨 只 株 折 灶 軸 口 棵 款 客 輛 粒 輪 枚 面 門 幕 匹 篇 片 所 艘 扇 首 乘 襲 頭 條 台 挺 堂 帖 顆 座 則 冊 任 尾 味 位 頁 葉 房 彎 班 員 介 丸 名 項 起 間 題 目 招 股 回
述賓式合用的量詞	通 口 頓 盤 局 番
群體量詞	對 雙 宗 翻 畦 餐 行 身 列 系列 排 副 套 蓬 筆 串 掛 幫 房 批 組 窩 綑 群 胎 桌 嘟嚕 部 種 類 樣(兒) 派 錄 壟 落 伙 束 簇 席 疊 紮 色 票 叢 攤 隊 式 項
部分量詞	些 分(兒) 部分 團(兒) 堆 泡 撮 把 股 攤 汪 陣 口 塊 滴 欄 捧 抱 層 重 帶 截(兒) 節(兒) 段(兒) 絲(兒) 點(兒) 片 縷 坨 匹 疋 階 坏 波 道
容器量詞	盒子 匣子 箱子 櫃子 櫥子 籃子 簍子 爐子 包兒 袋兒 池子 桶子 瓶子 罐子 盆子 鍋子 盤子 杓子 籠子 籮筐 罈子 擔子 杯子 勺子 碗 匙 筒子 茶匙 壺 盅 筐 瓢 鍬 缸
暫時量詞	手 地 池 身 腔 腳 嘴 頭 臉 肚子 屋子 家子 桌子 院子 鼻子
標準量詞—長度	丈 寸 尺 吋 米 呎 里 哩 浬 碼 釐 度 尋 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 營造尺
標準量詞—面積	甲 坪 畝 頃 公畝 公頃 市畝 英畝 營造畝 分 平方公里 平方公尺 平方公分 平方尺 平方英哩
標準量詞—重量	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 品脫
標準量詞—容量	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 營造升
標準量詞—時間	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
標準量詞—錢幣	分 角 毛 元 塊 先令 盧比 法郎 辨士 馬克 鎊 盧布 美元 美金 便士 里拉 日元 台幣 港幣 人民幣
標準量詞—其他	刀 文 令 卡 打 瓦 綸 赫 籮 千卡 千瓦 千赫 大籮 分貝 牛頓 仟卡 仟瓦 仟赫 瓦特 伏特 兆赫 安培 位元 周波 居里 馬力 毫巴 莫耳 焦耳 達因 爾格 赫茲 歐姆 燭光 燭光 卡路里 法拉第 毫安培 微居里 豪居里
準量詞	筆 劃(兒) 橫 豎 直 撇 捺 挑 剔 鉤 拐 點 格 國 省 州 縣 鄉 村 鎮

	鄰 里 郡 區 站 巷 弄 段 號 樓 衖 市 洲 地 街 部 司 課 院 科 系 級 股 室 廳 會 會兒 陣 世 輩 輩子 代 學期 學年 年代 下子 版 冊 編 回 章 面 小節 集 卷 面 方面 邊 頭 方 拍 板 眼 小節 程 作 倍 成 分 厘 毫 絲 圍 指 象限 度 開 聯 軍 師 旅 團 營 伍 班 排 連 球 波 端 回 合 折 摺 流 等 票 桿 棒 聲 次
動量詞	回 次 遍 趟 下 遭 番 聲 響 圈 步 把 仗 覺 頓 關 手 腳 巴掌 拳頭 拳 眼 口 刀 槌子 板子 鞭子 棒 棍子 陣 針 箭 槍 砲 場 度 輪 周 曲 跤 記 回合 票

附錄二：數詞單元集合表

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

附錄三：定量複合詞之類別

編號	詞性組合	編號	詞性組合
1	Nes Nf	57	Nes Neu Nf Na Ng
2	Nep Nf	58	Nep Neu Nf Ng
3	Nep Nf Na	59	Nep Neu Nf Na Ng
4	Nes Nf Na	60	Nep Neu Na Ng
5	Nes Nf Ng	61	Nep Neu Nf VH
6	Nes Nf Na Ng	62	Neqa Neu Nf Na Ng
7	Nep Nf Na Ng	63	Nes Neqa Nf Ng
8	Neqa Nf Neqb	64	Nes Neu Na Ng
9	Neqa Nf Neqb Na	65	Neqa Neu Nf Ng
10	Neqa Nf Neqb Na Ng	66	Nd Neu Nf
11	Neqa Nf Neqb Ng	67	Nd Neu Nf Ng
12	Neu Nf Neqb	68	Neu Neqa
13	Neu Nf Neu	69	Neu Neqa Nf
14	Neu Nf Neqb Ng	70	Nes Neu Neqa
15	Neu Nf Neqb Na Ng	71	Nep Neu Neqa
16	Neu Nf Neqb Na	72	Neu Neqa Nf Ng
17	Neu Nf Na Na	73	Da Neu Neqa Nf
18	Da Neu Nf Neqb Na	74	VH Neqa Nf
19	Nep Neu Nf Neqb	75	VH Neu Nf
20	Nep Neu Nf Neqb Ng	76	VH Da Neu Nf
21	Da Neu Nf Neqb	77	VH Neu Nf Ng
22	Neqa Nf	78	VH Neu Nf Na
23	Neqa Nf Na	79	VH Da Neu Nf Ng
24	Neqa Nf Na Ng	80	Neu Nf VH Na
25	Neu Nf Na	81	Da Neu Nf VH
26	Neu Nf	82	Neqa Nf VH
27	Neu Na	83	Neu Nf VH
28	Neu Na Ng	84	Nes Neu
29	Da Neu Na	85	Nep Neu
30	Da Neu Na Ng	86	Neu Nf Nf Ng
31	Da Neu Nf Na Ng	87	Neu Nf Nf Na
32	Da Neu Neqa Nf Ng	88	Nes Nf Nf
33	Da Neu Nf Ng	89	Neu Nf Nf
34	Da Neu Nf Na	90	Nep Nes Nf

35	Da Neu Nf	91	Nes Nes Neqa Nf
36	Neu Nf Ng	92	Nes Nes Neqa Nf Ng
37	Neu Nf Na Ng	93	Nep Nes Nf Na
38	Neqa Nf Ng	94	Neu Neqb
39	Neu Neqa Ng	95	Neu VH Nf Na
40	Nes Neqa Nf	96	Neu VH Nf
41	Nep Neqa Nf	97	Nes Neu Nf Na
42	Nep Neu Na	98	Neu Na Na
43	Nes Neu Na	99	Nep Neu Nf Na
44	Nes Neu Nf	100	Da Neu Nf Nf
45	Neqa Neu Na	101	Neu VH Nf Ng
46	Nep Neu Nf	102	Nd Nd Ng
47	Neqa Neu Nf	103	Neu Neu Nf
48	Neqa Neu Nf Na	104	Neu Neu Nf Na
49	Nes Neqa Nf Na	105	Neu Neu Nf Ng
50	Nep Neqa Nf Ng	106	Neu Neu Na
51	Nep Neqa Nf Neqb Ng	107	Neu Neu Nf Na Ng
52	Nep Neqa Nf Neqb	108	Neu Neqa Nf Na
53	Nep Neqa Nf Na Ng	109	Neu Neu Nf VH
54	Nep Neqa Nf Na	110	Nes Neu Nf Neqb Na
55	Nes Neqa Nf Na Ng	111	Nes Nes Neu Nf
56	Nes Neu Nf Ng	112	Neu Neqa Na