國 立 交 通 大 學

電信工程學系

碩 士 論 文

中文連續語音辨認後處理之進一步研究

A Further Study on Post-Processing of Continuous Mandarin Speech Recognition

研 究 生：張志豪

指導教授：陳信宏　博士

中華民國九十七年八月

國立交通大學
電信工程學系碩士班

論文口試委員會審定書

本校 電信工程學系 碩士班 張志豪 君

所提論文(中文) 中文連續語音辨認後處理之進一步研究

(英文) A Further Study on Post-Process of Continuous Mandarin Speech Recognition

合於碩士資格水準、業經本委員會評審認可。

口試委員：王小川　陳信宏

王逸如

李祈均

指導教授：陳信宏

系主任：陳信宏 教授

中 華 民 國　97 年　8 月　25 日

中文連續語音辨認後處理之進一步研究

研究生：張志豪　　指導教授：陳信宏 博士

國立交通大學電信工程學系碩士班

中文摘要

本論文分成兩個部份，第一部分探討建立語言模型時所使用的文字資料庫的適用性，觀察文字資料庫的內容是否適合建立語言模型，刪除不適合的內容及更正錯誤文字，希望能提升整體的辨識率。第二部分是針對辨認結果，以有意義的長詞為目標，而非只和辨認用詞典中的詞比對，為此我們多考慮了二種構詞，包括數量複合詞及人名，結果辨認率下降許多，顯示原先辨識結果將許多有意義的這二類長詞辨識成意義不完整或錯誤的短詞。由於辨認用詞典無法包含所有構詞，我們因此嘗試將常被用來構成這些詞的一字詞或 subword 加入詞典，希望這些構詞被辨認成正確的短詞串，以便在未來經後處理產生正確構詞。實驗結果顯示以 subword 作為構詞成分較一字詞為佳。

A Further Study on Post-Processing of Continuous Mandarin Speech Recognition

Student: Zhi-Hao Zhang Advisor：Dr. Xin-Hong Chen

Department of Communication Engineering

National Chiao Tung University

Abstract

The thesis divided into two parts, one is to 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. The second part is that aim at the recognizable result. We use the meaningful long term for goal, not the meaningless short term. For these, we consider two compound words that include determiner-measure compound and name entity . The result is that the recognition rate goes down a lot. That shows the recognizable result let many meaningful these two kinds of long term to recognize incomplete meaning or wrong short term. Because our recognition can not include all compound words, we try to put one length word or subword which are often used to compound these words into lexicon. We hope these compound wards can be recognized the correct strings of word, then it can produce the right compound words in the future. The experimental result is that the subword is better than the string of word to be the component of compound words.

誌謝

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。

兩年的研究所生活，大部分的時間都在研究室及宿社度過，和學長討論學術上的知識、言不及義的鬼扯、讓人又愛又恨的線上遊戲、因研究進度太慢深怕遇到老師...，感謝眾位學長、同學、學弟的勉勵指教，有你們的陪伴讓兩年的研究所生活變得多彩多姿。

感謝振宇學長在學術上及程式上的指教，當我在無助時，都有你的陪伴；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；感謝小廣、阿宅、翔耀、柯達同學，有你們的陪伴，我才能順利的度過研究所生活。感謝我可愛的女友，當我對研究不順心時，總是安慰我、鼓勵我、認真的對我說「Just do it」；感謝愛好攝影的朋友們，陪我拍盡無數美麗的照片，讓我在心情低落時，看到自己拍的照片也會大笑。

最後感謝我的雙親，給予學業上的信任與自由，以及支持我放棄北科大研究所，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

現代科技中的一項重要發展是用電腦來處理語言問題，最終的目標，就是利用語言辨識技術來建立人與機器之間溝通的橋樑【1】，由於信號處理、演算法和電腦硬體設備的進步，語音辨識技術在過去十到二十年間有長足的發展，例如聲學模型和語言模型建立方式，以及基於動態編輯程序（Dynamic Programming-based）之搜尋方法等【2】。

近年來，大家都從語音訊號的方面著手，研究聲音的特性，提昇辨識率。若回歸基本面，在語音辨識上，語音和語言密不可分，因此對語言這方面做研究，或許能夠提高辨識的效果。基於以上理由，本論文將探討中文語音辨認的語言模型，對訓練語言模型的文字語料庫的整理及詞典的建立做深入的探討，希望對整體辨認率的提升有所助益。

1.2 研究方向

基本的語音辨識系統方塊圖如下：

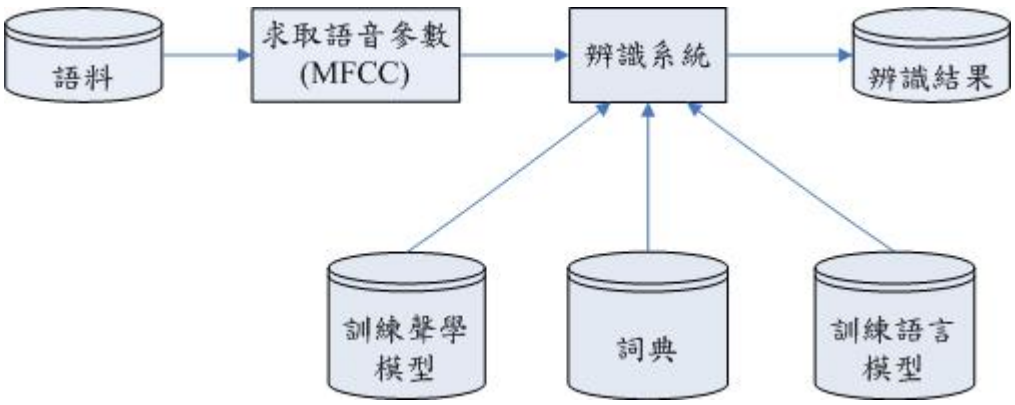

圖 1-1：辨識流程

如上圖所示，一個基本語音辨識系統包含：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、詞典以及辨識比對。

近年來，我們用的詞典大小都是根據統計得來的，但是我們從未探討文章的內容正確與否，是否間接影響詞典的統計結果，以及所用的詞典是否適合辨識系統。因此本論文的研究方向是在深入探討文字語料庫中文章內容，是否有些文章的內容對於訓練語言模型是沒幫助的，將其去除，以建立有效詞典及語言模型，改進整體的辨認效能。

1.3 章節概要

本篇論文章節內容說明如下：

第一章 **緒論**：介紹研究動機、研究方向及章節概要。

第二章 **語音辨認基本系統建立**：訓練聲學模型、語言模型，並檢視其辨識效能。

第三章 **訓練語料的分析、修正與討論**：刪除語料的書寫形式、修正語料內容。

第四章 **語音辨認後處理基本分析**：語料內的定量複合詞(DM)、人名使用小單位的詞取代，檢視其辨識效能。

第五章 **語音辨認後處理之改良**：語料內的定量複合詞、人名使用較長的短詞取代，檢視其辨識效能。

第六章 **結論與展望**

第二章 語音辨認基本系統建立

近年來語音辨認的研究最常採用的聲學模型是隱藏式馬可夫模型(Hidden Markov Model, HMM)，藉由這種機率模型，來描述發音過程的狀態(State)轉移現象和輸出結果，這種方法有不錯的辨識效能。所以本系統也採用這個模型，並加入語言模型，藉由語言模型的幫助提高辨識率。

2.1 Acoustic model 建立

進行語音辨認系統之訓練、測試，首先前處理的工作就是將語音參數從輸入的語音中抽取出來。因為語音訊號具有短時間穩定特性(Short Term Stationary)，加上考慮到人耳聽覺效應，語音辨認常使用的參數為梅爾倒頻譜參數(Mel-Frequency Cepstral Coefficients，MFCC)，它的成份包括 12 維 MFCC 加上能量共 13 維，取 Delta 和 Delta- Delta，加起來總共 39 維，而能量的大小對於語音辨認較不重要，所以一般省略了能量參數，最後得到的參數是 38 維的語音參數。

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

我們採用 left-to-right HMM，雖然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，因此假設在同一音框（Frame）中，口腔狀態是相同的。此外，代表音框語音參數與各狀態的相似程度的狀態觀測機率（State Observation Probability），使用混和高斯模型（Gaussian Mixture Model）來表示。

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，推出每個音框所屬的最佳口腔狀態，並重複估測直到

收斂為止。下圖為 HMM model 的訓練流程。採用的訓練軟體為英國劍橋大學開發的 HMM Tool Kit (HTK) 【3】。

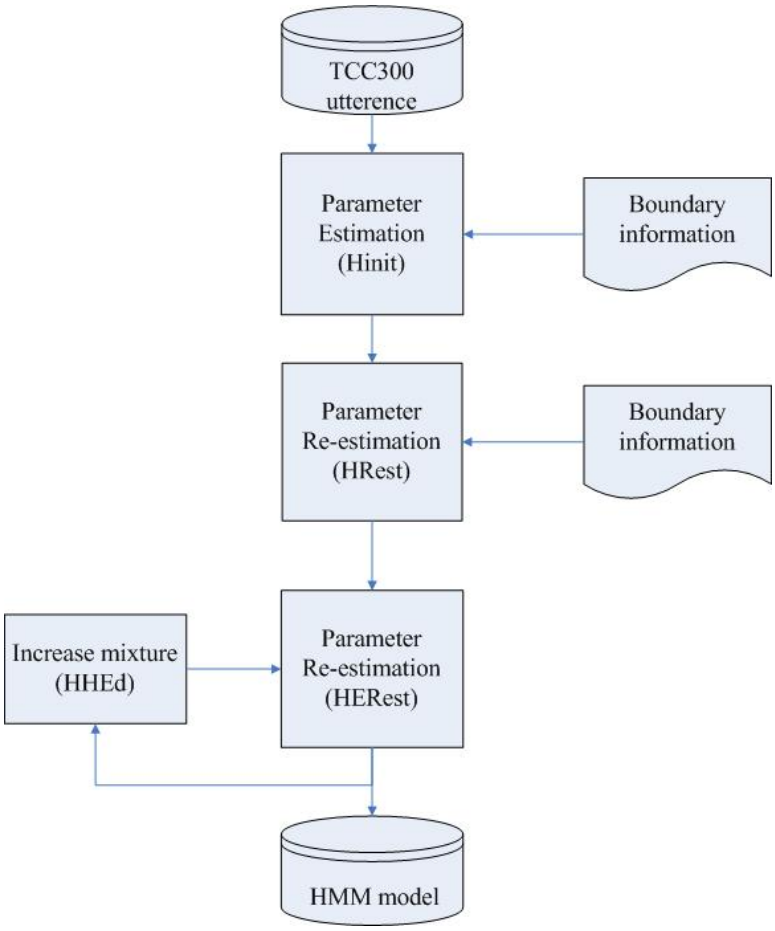

圖 2-1：已知切割模型之訓練圖

2.2 Word-based bi-gram language model 建立及基本觀念

語言模型區分為兩種，一種是根據語言的文法、詞性，訂定文章出現符合規則之語言模型（Rule-based LM）;另一種則是藉由處理大量的文字資料，利用統計的方式，計算詞和詞之間的聯結規則而建立的語言模型(Statistic-Based LM)。

2.2.1 n-grams 語言模型

假設有一個句子，句子構成單元為詞，若句子中有 m 個詞$(w_1, w_2, ..., w_m)$，其中「w_i」表示句子中的第 i 個詞。此句子的發生的機率可表示為

$$
\begin{aligned}
P(w_1, w_2, ..., w_m) &= \prod_{i=1}^{m} P(w_i \mid w_1, w_2, ..., w_{i-1}) \\
&\simeq \prod_{i=1}^{m} P(w_i \mid w_{i-n+1}, ..., w_{i-1}) \qquad \text{(n-gram)} \\
&\simeq P(w_1) \prod_{i=2}^{m} P(w_i \mid w_{i-1}) \qquad \text{(bi-gram)}
\end{aligned}
\tag{2.1}
$$

其中

$$P(w_i \mid w_{i-1}) = \frac{Count(w_{i-1}, w_i)}{Count(w_{i-1})}$$

$Count(\bullet)$ 表示詞串出現次數

2.2.2 機率的 Smoothing

在訓練 bi-gram 機率時，若分子的 $Count(\bullet)$值為 0 時，就是 bi-gram 機率會等於零。因為在 training data 中未出現，但不代表 testing data 不會出現，因此這種情況下機率的給定是不合理的。當詞接詞之 count 值很小時，所計算出的 n-gram 機率也是不準確。所以必須對計算出的機率做 smoothing 的動作【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：

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

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

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，解決方式是當詞串次數小於 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。

2.2.3 語言模型建立流程

辨識器的語言模型通常需要由大量的文字資料來訓練，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」。一個好的語言模型，它所需要的條件，必須擁有大量的文字資料庫，在此論文中我們使用下列兩個文字資料庫來建立語言模型：

(1)光華雜誌：內容是一般雜誌的文章，蒐集範圍為 1976 年到 2000 年之間。

(2) NTCIR：內容是各個不同學科領域的文章構成，是建立資訊檢索系統的標竿測試集。

藉由輸入大量的文字資料，統計出各種詞串在文章中累積出現的次數，接者利用（2.2）式和（2.3）式的 smoothing 方法，計算出建立語言模型所需的 n-gram 機率。在此我們建立出了 unigram 及 bigram 機率。但是要把語言模型及聲學模型在辨識系統中共同使用，我們需將語言模型轉換成 word-net 的形式，其描述著詞與詞之間的串接機率。訓練語言模型使用的軟體為 HTK Toolkit【3】。整體語言模型建立流程可參考下面流程圖：

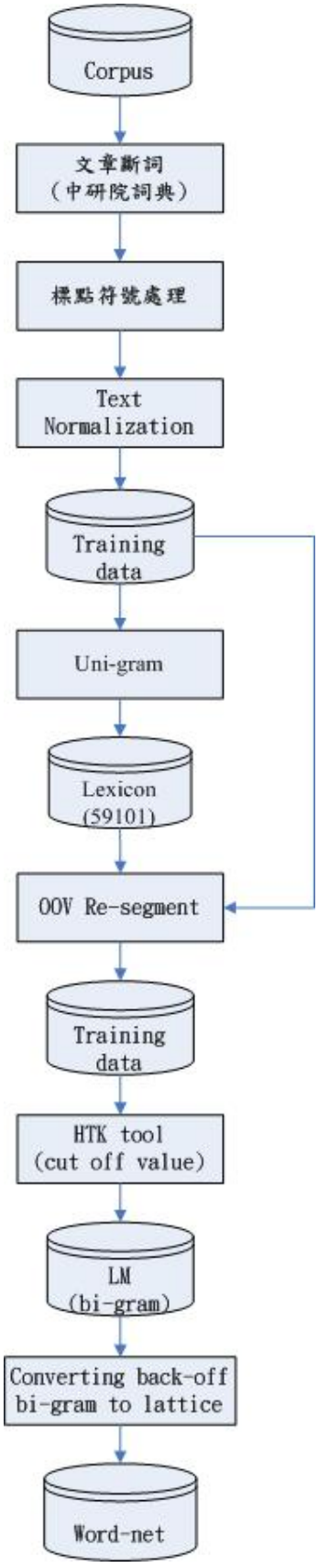

圖 2-2：語言模型建立流程圖

2.2.4 Perplexity 計算

利用建立好的 LM，可以算出 testing data 的 Perplexity (PP)，語言模型的好壞我們可以由 perplexity 來測量，perplexity 定義如下:

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

上式是一個句子之內容由 m 個詞所組成，並對於每個新詞提供的平均資訊量，entropy (H)，經過了 ergodic 的假設與適當的化簡，最後以上式來對 H 做近似。計算 log probability 是以 10 為底，因此數學式修改如下:

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

其中

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

$$\log_{10} P(w_1, w_2, ..., w_m) = \log_{10}(\prod_{i=1}^{m} P(w_i \mid w_{i-1}))$$
$$= \log_{10} P(w_1) + \log_{10} P(w_2 \mid w_1) + ... + \log_{10} P(w_m \mid w_{m-1})$$
$$= \sum_{i=1}^{m} \log_{10} P(w_i \mid w_{i-1})$$

2.2.5 語言模型測試

由於辨識語料是 TCC300，所以將 TCC300-train data 視為測試語料，下表為測試語料統計表及所算出的 Perplexity。

表 2-2 測試語料 Database 統計

測試語料	詞數（Word）	字數（Character）
TCC300-train data	181,762	300,867

表 2-3 測試語料 Perplexity

LM	Perplexity
Word-based	381.1629

2.2.6 實驗一：TCC300 辨識效能

實驗中使用到的語言模型是使用兩大語料庫建立的 bigram LM，配合聲學模型進行辨識，為了加快 Viterbi search 的速度，使用 beam search。因為使用語言模型，辨識結果的基本單元將不再是音節，辨識器輸出單元將以詞為主。因此我們可以計算詞(word)、字元（character）和音節(syllable)三種不同的辨識率。

加入 word-based language model，下表為辨識結果。

表 2-4：實驗一之辨識結果

	Deletion	Substitution	Insertion	Accuracy	Total count
word 辨識率	5.17%	24.12%	2.17%	68.53%	19215
character 辨識率	1.15%	20.41%	0.30%	78.11%	31412
syllable 辨識率	1.16%	12.11%	0.31%	86.41%	31412

實驗一中 word 的辨識率是以辨認詞典中所含的 60,000 詞為準，辨識出來的 word 中有許多是較沒意義的短詞。由於我們語音辨識的最終目標，是辨識結果都為有意義的長詞，所以我們把辨識答案跟含有意義長詞的標準答案做比對，來計算 word 的辨認率。由於這是一個複雜的問題，本論文僅考慮對斷過詞的文章做三類的構詞，包括數量複合詞、人名及含常用前後詞綴的複合詞，將其視為標準答案。實驗一的測試語料(TCC300_測試語料)經此處理後，總詞數降為 18,034，其統計分佈如表 2-5。

表 2-5：TCC300_測試語料之詞統計分佈

TCC300_測試語料分析如下				
	總詞數	詞數佔多少比例(%)	總字數	字數佔多少比例(%)
TCC300_test	18,034		31,421	
文章細分如下：				
TCC300_test_DM	625	3.47%	1,670	5.31%
TCC300_test_Name	245	1.36%	723	2.30%
TCC300_test_Prefix	643	3.57%	1,338	4.26%

TCC300_test_Suffix	629	3.49%	1,784	5.68%
TCC300_test_Other	15,892	88.12%	25,906	82.45%

將實驗一的辨識結果與新的標準答案比對，結果如下：

表 2-6 實驗一-辨識結果跟有意義的長詞比較

Outside test：Total 18034 words			
Deletion	Substitution	Insertion	Accuracy
3.92%	28.07%	7.26%	60.73%

從表 2-6 我們發現 word 的辨識率，降為 60.73%，這顯示原先辨識結果將有意義的三類長詞辨識成意義不完整或錯誤的短詞。

第三章 、文章內容的分析、修正與討論

基本語音辨識系統包含：求取語音特徵參數、訓練聲學模型、訓練語言模型、詞典、以及辨認比對。本論文加入辨識器的語言模型，屬於 word-based language model，而 word-based 的語言模型需要大量的訓練語料，才可以訓練出好的模型。在語音辨識系統中，有些文字資料庫的內容在語音辨識上，是不會出現的，這些內容大部分都是註解或是強調功能，在一般口語上不會出現的內容，所以這些類似註解的內容對於我們在訓練一個語言模型時，只會造成詞和詞之間的統計數據較不可靠，對語音辨識來說是沒幫助的，所以得把文字資料庫中，有這種類似註解的內容，做進一步的處理。

3.1 文字資料庫

辨識器的語言模型需要大量的文字資料，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」。一個好的語言模型，所需要的條件，必須擁有大量的文字資料庫，在此建立的語言模型共有兩個文字資料庫，分別如下：

(1)光華雜誌：內容是一般雜誌的文章，蒐集範圍為 1976 年到 2000 年之間。

(2) NTCIR：內容是各個不同學科領域的文章構成，是建立資訊檢索系統的標竿測試集。

3.2 文章內容分析與刪除

文章是由報章雜誌、不同學科領域的文章構成，文章內容的正確與否會影響到語言模型，因為文章的結構是書寫形式，語音辨識是語音讀法形式，所以文章的內容得再進一步處理，把書寫形式轉換成語音讀法。文章的內容是修飾、強調、註解或翻譯功能，在語音讀法時，這類型的內容通常不被讀出，所以文章針對這

些問題做進一步的修改、刪除和分析。

(1)文章內容_半型轉為全型

由於文章的書寫形式都不一樣，所以文章內的文字要統一成單一形式，把文章內的文字全轉換成全型的形式，半型變為全型，而全型仍為全型。做此處理是為了文字形式統一，因為有些字在半型和全型上，都是相同語義、相同發音和相同字形，但寫法上有些許差異，在訓練語言模型(LM)時造成詞跟詞之間的不確定性，也就是同一個字，有不同寫法。

例如：阿拉伯數字、標點符號、英文符號。

表 3-1：文章內容半型轉全型

	半型	全型
阿拉伯數字	0 1 2 3 4 5 6 7 8 9	０ １ ２ ３ ４ ５ ６ ７ ８ ９
標點符號	, ` . ; : ? !	， 、 。 ； ： ？ ！
英文符號	a b c d e f g h I j k l m n o p q r s t u v w x y z	ａ ｂ ｃ ｄ ｅ ｆ ｇ ｈ ｉ ｊ ｌ ｋ ｍ ｎ ｏ ｐ ｑ ｕ ｒ ｓ ｕ ｖ ｗ ｘ ｙ ｚ
	A B C D E F G H I J K L M N O P Q R S T U V W X Y Z	Ａ Ｂ Ｃ Ｄ Ｅ Ｆ Ｇ Ｈ Ｉ Ｊ Ｋ Ｌ Ｍ Ｎ Ｏ Ｐ Ｑ Ｒ Ｓ Ｔ Ｕ Ｖ Ｗ Ｘ Ｙ Ｚ

(2)文章內容_刪除空白符號

在中文語音辨認上，空白符號在中文字串中是沒意義的字元，所以在文章中文部分，刪除空白的符號。在英文的空白則是保留，因為在英文部分，空白算是英文單字的分界，但在此只對中文語音辨識做為研究重點，對英文部分則不加以

討論。

(3)文章內容_刪除標題 (Title)

標題是為突顯文章的主要內容所給予的一串文字，它通常是由較重要的詞串接而成，但常常不符合語法。這標題內的文字，在訓練語言模型時，估算詞和詞之間的關係較無直接的關連性，在訓練語言模型時，並未考慮納入標題，所以把標題部分從文章中刪除。標題的例子如下表：

表 3-2：文章內容_刪除標題

台灣南部墾丁公園
日月潭文武廟
春節舞龍
陽明山辛亥樓
張大千繪慈湖圖

移除字串總個數= 37。

(4)文章內容_刪除括弧【】及括弧【】內的中文字串

當文章的內容是各個不同學科領域的文章構成，括弧及括弧內的中文字串，幾乎都是標示這篇文章是由某單位或某人物記錄，強調此篇文章的來源，對於語音辨認上，這類的資訊在辨識上，是沒有幫助的，所以移除這方便的資訊。例子詳列於附錄一。移除字串總個數= 116806。

(5)文章內容_刪除括弧（）及括弧（）內的中文字串

當文章的內容是一般雜誌時，括弧（）及括弧（）內的中文字串，幾乎都是註解的字串，對於語音辨認上，這類的資訊是沒有幫助的，也就是人類不會在說話時加上括弧註解，所以可移除這類資訊。依照括弧（）前的字找出一致性的關係。經觀察發現括弧（）內的字串都為時間詞，大都為數字，表示某一天的日期，這些字串都是註解功能，這類的文章對訓練語言模型沒有幫助，所以刪除。括弧（）及括弧（）內的字串格式（＊表示某一個中文字）如下表：

表 3-3：括弧（）及括弧（）內的字串格式

＊＊＊今（＊＊＊）
＊＊＊昨（＊＊＊）
＊＊＊明（＊＊＊）
＊＊＊後（＊＊＊）
＊＊＊去（＊＊＊）
＊＊＊前（＊＊＊）
＊＊＊本（＊＊＊）
＊＊＊一（＊＊＊）
＊＊＊二（＊＊＊）
＊＊＊三（＊＊＊）
＊＊＊四（＊＊＊）
＊＊＊五（＊＊＊）
＊＊＊六（＊＊＊）
＊＊＊日（＊＊＊）

移除字串總個數= 26,989。例子詳列於附錄二。

(6)文章內容_刪除括弧（）及括弧（）內的英文字串

當文章的內容是一般雜誌，且括弧內為英文字串時，幾乎都是表示某組織單位的英文縮寫、某會議的英文縮寫、專有名詞的英文或英文人名，在語音辨認上，這類的資訊是沒有幫助的，所以移除這方面的資訊。括弧（）及括弧（）內為英文的例子如下表：

表 3-4 括弧（）及括弧（）內為英文

原始文字串	修改過後文字串
亞太經濟合作會議（ＡＰＥＣ）二樓	亞太經濟合作會議二樓
中共永久正常貿易關係（ＰＮＴＲ）	中共永久正常貿易關係
世界貿易組織（ＷＴＯ） 身懷「花花公子」（ＰｌａｙＢｏｙ）雜誌	世界貿易組織 身懷「花花公子」雜誌
國民生產總值（ＧＮＰ）上	國民生產總值上
那斯達克（Ｎａｓｄａｑ）指數	那斯達克指數
得獎人莫頓（ＲＯＢＥＲＴＭＥＲＴＯＮ）博士	得獎人莫頓博士

移除字串總個數=51,322 個。

(7)文章內容_修正錯誤的符號

訓練語言模型需要大量的文字資料，而文章內容的正確性影響到統計參數，所以文章內容的完整性和正確性，直接影響語言模型好壞。文章中有些應該是0(零)的字，卻是以符號○表示，此問題需更正，所以得修正文章中此符號，用“零”取代。錯誤的符號之例子如下表：

表 3-5 更正錯誤的符號

原始文字串	修改過後文字串
日本在二○○三年前	日本在二**零零**三年前
九百○五億二千五	九百**零**五億二千五
加坡幣偏低○・五四％	加坡幣偏低**零**・五四％
三十八點○二元	三十八點**零**二元
刑法第一二○三條	刑法第一二**零**三條

修改字串總個數=74331 個。

(8)文章內容_刪除錯誤的符號

文章內容愈正確，對語言模型有較好的參數量，也間接影響到語音辨識系統，文章內容對統計詞和詞之間的統計機率影響深遠。而文章中有一些符號對於語音辨認上來說是沒幫助的，所以移除文章中出現的此符號”（）”。錯誤的符號之例子如下表：

表 3-6 刪除錯誤的符號

原始文字串	修改過後文字串
最佳動畫短片：兔子（）Ｂｕｎｎｙ	最佳動畫短片：兔子Ｂｕｎｎｙ
日精技研專務（）總經理竹村進。	日精技研專務總經理竹村進。
另外可將生菜（）切細絲	另外可將生菜切細絲
每月付７５（）元	每月付７５元
人類學家孔恩博士（）所說的	人類學家孔恩博士所說的

移除字串總個數= 1,794 個。

(9)文章內容_刪除括弧及括弧內的中英文字串

當文章的內容是一般雜誌，括弧及括弧內中英文夾雜的字串，幾乎都是某公司的英文縮寫、某會議的英文名稱或是註解，在語音辨認上，這類的資訊是沒有幫助的，也就是人類不會在說話時，加上括弧註解，並且本論文研究的是中文辨識，所以移除這方面的資訊。

括弧（）及括弧（）內為中英文的例子如下表：

表 3-7 括弧（）及括弧（）內為中英文

原始文字串	修改過後文字串
基金會（簡稱ＩＦＰＩ）的統計	基金會的統計
環–單磷酸鳥甘（簡稱ＣＧＭＰ）	環–單磷酸鳥甘
領養的女嬰（即為Ａｉｍｅｅ）	領養的女嬰
庫德工人黨（土稱ＰＫＫ）武裝游擊隊	庫德工人黨武裝游擊隊
ＤＶＤ－Ｒ（可燒錄一次ＤＶＤ）及ＤＶＤ－ＲＡＭ	ＤＶＤ－Ｒ及ＤＶＤ－ＲＡＭ
ＨＤＤＶＤ（高畫質ＤＶＤ）時代	ＨＤＤＶＤ時代
八百億比次秒（８０Ｇｂｐｓ）數位電路容量	八百億比次秒數位電路容量
「艾卡彼卡索」(暫譯，ＡＫＡＰｉｃａｓｓｏ)	「艾卡彼卡索」

移除字串總個數= 1,396 個。

(10)文章內容_刪除括弧及括弧內的中文字串

文章的內容是一般雜誌，括弧及括弧內的中文字串，在文章中幾乎都是註解或修飾，在語音辨認上，這類的資訊是沒有幫助的，也就是人類不會在說話時，加上括弧註解，所以移除這方面的資訊。刪除的字串格式詳列於附錄三。刪除的字串例子詳列於附錄四。移除字串總個數= 60521 個。

3.3 文字資料庫_斷詞

在此建立語言模型是由統計式的方式建立，在統計詞和詞之間的聯結關係，所以得把文章斷成以詞為單位的資料，統計詞和詞之間的機率。斷詞之詞典所收錄的詞愈多，文章斷詞之後的結果，較不會有搶詞的現象出現，也就是詞典的大小，會影響到斷詞結果的正確率。詞典無法收錄所有的詞，這些未收錄的詞，有

些是有規則的，我們可以利用構詞規則，把輸入文句中符合規則的詞構出，即為「定量複合詞」、「數詞定詞」、「數量定詞」、「時間詞」、「地方詞」、「位置詞」。在斷詞時，加上構詞單元，能讓斷詞的結果更加完善【5】。我們使用的斷詞詞典為交大語音實驗室_大詞典，其詞條分類如下：

(1)人名：主要來源：

(1.1)聯考榜單。

(1.2)文字資料庫(光華雜誌、NTCIR)經由現有的六萬詞詞典斷詞後，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：

(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)經由現有的六萬詞詞典斷詞後，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)經由現有的六萬詞詞典斷詞後，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。

斷詞的詞典(Segment_Lexicon)詞條分類統號如下表：

表 3-8 斷詞的詞典詞條分類

分類	細分	詞數	備註
人名	平衡語料庫	9,291	去除重複後詞數 409,847
	光華雜誌 &NTCIR	75,108	
	榜單	329,593	
詞綴	前詞綴	14,997	來源：半自動，以人工挑出
	後詞綴	40,436	
股票名		5,027	來源：網路
學校及科系	校名	100	來源：榜單
	科系名	352	
定量複合詞(DM)		256,068	來源：語料庫自動斷詞
縣市鄉鎮	縣市	19	來源：地圖資訊
	鄉鎮	327	
中研院八萬八千詞辭典	原始辭典	88,142	
	DM 成分	536	
	科系成分	8	
	詞綴構成詞成分	13,384	
	人名	240	
	股票名	130	
	剩下的核心詞	73,852	

3.4 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，利用點號中的五種點號(逗號、句號、分號、驚嘆號、問號)把文章分段，由於在聲學模型中並未考慮到標點符號的模型，所以把文章中所有的標點符號移除。

在訓練語言模型時，因為句子由太少的詞構成，所以句子只包含一或二個詞，這類的句子對我們在訓練 Bi-gram LM 時，會造成估算詞和詞之間的機率被

高估，所以這類的句子在訓練語言模型時不採用。

下表則是兩大資料庫斷詞經標點符號處理後的 word 及 character 統計表：

表 3-9：標點符號處理後的 word 及 character 統計表

訓練語料	詞數（Word）	字數（Character）
光華雜誌	8,945,660	15,691,089
NTCIR	108,217,471	201,075,931
合計	117,163,131	216,767,020

文章的平均詞長= 1.85。

下表則是兩大資料庫斷詞經標點符號處理後詞的分析如下表：

表 3-10 資料庫斷詞經標點符號處理後詞的分析

文章內容分析如下				
	總詞數	詞數佔多少比例(%)	總字數	字數佔多少比例(%)
文章 CIRB、光華	117,182,354		216,941,145	
文章細分如下：				
文章 CIRB、光華_DM	5,088,836	4.34%	15,842,895	7.30%
文章 CIRB、光華_Name	1,744,102	1.49%	5,075,250	2.34%
文章 CIRB、光華_Prefix	5,849,677	4.99%	12,356,210	5.70%
文章 CIRB、光華_Suffix	5,355,880	4.57%	14,952,253	6.89%
文章 CIRB、光華_Other	98,840,959	84.35%	167,492,428	77.20%
文章 CIRB、光華_Eng	302,900	0.258%	1,222,109	0.563%

3.5 斷詞後的資料庫_文字正規化

文字正規化主要分兩部分：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。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，需把這類的詞統一成同一個詞。經由這兩個步驟的過程稱為文字正規化。

3.5.1 文字正規化：

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)，如果含有阿拉數字及特殊符號，都需要被正規化為讀法。例子如下表：

表 3-11 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
1980/3/7	一九八零年三月七日
ＴＥＬ	電話

3.5.2 同音同義異詞：

某些詞和詞之間在發音上和語義上是相同，只是寫法有所不同，而這類的詞會在辨認上造成更多的不確定性，所以得把這類的詞給統一成同一個詞。同音同義異詞且詞頻少的詞，經過此步驟會減少 OOV 的量(當詞典未收錄同音同義異詞時)。例子如下表：

表 3-12 同音同義異詞範例

同音同義異詞	
佰	百
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

3.6 OOV 的分析_長詞變短詞

為了比較語料庫經前處理過後是否有比較好的統計機率，所以利用現有的詞典(約六萬詞)，而不在這六萬詞內的詞即為 OOV Words，使用長詞變短詞的技巧，利用詞典的詞取代 OOV words【7】。下表則是經長詞變短詞後的 word 及

character 統計表：

表 3-13 長詞變短詞後的 word 及 character 統計表

訓練語料	詞數（Word）	字數（Character）
光華雜誌	9,741,583	15,699,877
NTCIR	121,653,079	201,389,272
合計	131,394,662	217,089,149

3.7 建立 LM 之流程圖

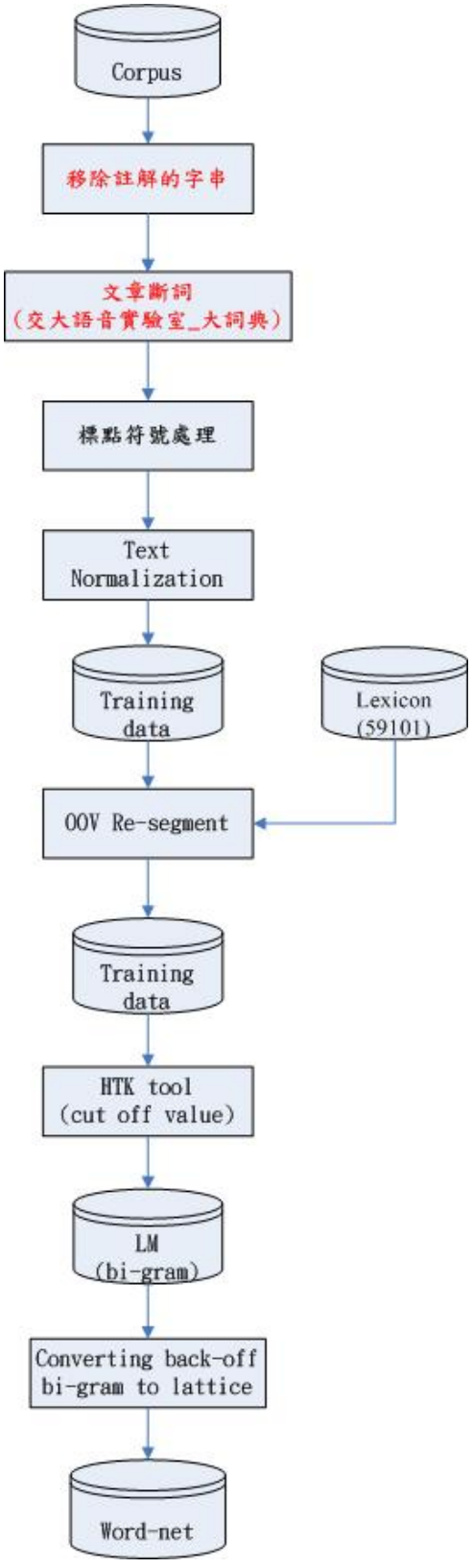

圖 3-1：實驗二-語言模型建立流程圖

藉由輸入大量的文字資料，統計出各種詞串在文章中累計的出現次數後，利用 smoothing 的方法建立 LM 所需的 n-gram 機率，使用 HTK tool 分別求出 unigram、bigram。

3.7.1 HTK Training data：

Database 經由 Segment_Lexicon (表 3-8)斷詞、標點符號處理、文字正規化、同音同義異詞處理、不在 Language Model_Lexicon 內的詞(OOV)，經由詞典的詞取代。

3.7.2 Language Model_Lexicon：

表 3-14 實驗二詞典統計表

Language Model_Lexicon		
總詞數	59101	
詞長	詞數	百分比(%)
1	9794	16.57
2	33632	56.91
3	9389	15.89
4	5874	9.94
5	231	0.39
6	126	0.21
7	33	0.06
8	22	0.04

LM 詞典的平均詞長= 2.22。

3.7.3 Smoothing (cut off value)：

選擇一個適當的 cut off 值，將影響到語言模型的參數量，當 cut off 值愈小，相對在訓練語言模型時，需要較多的資料量，使用到 back off 的資料就比較少，語言模型的模糊度相對的低，對於辨認時，由於語言模型的資料量較大，辨認時需花較多的時間，而當 cut off 值愈大，所需要的資料量愈少，使用到 back off 的資料就比較多，語言模型的模糊度相對的高，而語言模型資料量小，辨認時所花的時間較小，所以需在這之間找一個適當的 cut off 值。

3.7.4 語言模型測試：

由於辨識語料是 TCC300，所以將 TCC300-train data 視為測試語料，下表為 Perplexity。

表 3-15：測試語料 Perplexity

LM	Perplexity
Word-based	355.2593

3.8 實驗二：修改語料庫後_辨識效能

表 3-16：實驗二-辨識結果

	Deletion	Substitution	Insertion	Accuracy	Total count
word 辨識率	4.40%	23.57%	2.47%	69.47%	19215
character 辨識率	1.12%	20.16%	0.32%	78.39%	31412
syllable 辨識率	1.13%	12.04%	0.33%	86.48%	31412

實驗一：原語料庫。

實驗二：原語料庫經修改、刪除。

語料庫經修改後，所影響的是詞和詞之間的統計機率更加可靠，在詞的辨認上有所進步，因為實驗一的語料庫，有些文章不是口說時會出現的文章，這類的

內容，會造成詞跟詞之間的機率被高估，導致有些詞頻少的詞被低估。文章的改進影響到詞的辨認率，提升了 0.94%，對於 character、syllable 的辨認率來說，影響較小，但還是有小小的提升。

實驗二 word 的辨識率，辨識出來的 word 是較沒意義的短詞。由於我們語音辨識的最終目標，是辨識結果有愈多有意義的長詞，所以我們把辨識答案跟新的標準答案比對，結果如下表：

表 3-17：實驗二-辨識結果跟較有意義的長詞比較

Outside test：Total 18034 words			
Deletion	Substitution	Insertion	Accuracy
3.38%	27.46%	7.78%	61.36%

下表為實驗一跟實驗二的辨識答案跟較有意義的長詞做辨認。

表 3-18：實驗一跟實驗二的辨識答案跟較有意義的長詞

	Deletion	Substitution	Insertion	Accuracy
實驗一	3.92%	28.07%	7.26%	60.73%
實驗二	3.38%	27.46%	7.78%	61.36%

文章的內容修改間接的影響到 word 的辨認率，詞的辨認率也提升了 0.94%，針對我們的目標，想在辨認結果上，辨認出的是有意義的長詞，對詞辨認上，相對有提升 0.63%。

第四章 語音辨認後處理基本分析

4.1 分析人名及構詞規則構出的詞

中文字很容易就因結構上、語法上，能構出一個新的詞，由於詞典無法收錄所有的詞，並且詞典內收錄的詞有限，無法全部收錄，當有些詞不是詞典內所收錄的詞，我們稱此詞為 Out Of Vocabulary (OOV) words，OOV words 在辨認時又無法被辨認出來，會影響到辨認結果。為了減少 OOV 這問題，盡可能把 OOV words 減少。

OOV words 主要是人名及構詞規則所構出的詞(「定量複合詞」、「數詞定詞」、「數量定詞」、「時間詞」、「地方詞」、「位置詞」)，針對這兩大類的詞做進一步的處理，以減少當詞典未收錄所造成的 OOV words。

表 4-1 人名及構詞規則所構出的詞

構詞規則所構出的詞			人名
定量複合詞	DM	兩萬平方公尺	陳水扁
數詞定詞	Neu	一千四百億	馬英九
數量定詞	Neqa	三十左右	周杰倫
時間詞	Nd	下午四點多	郭台銘
地方詞	Nc	六年十一班	梁朝偉
位置詞	Ncd	一百廿度五	劉嘉玲

(1)構詞規則構出的詞以構詞最小單元代替

文章在斷詞時，使用構詞單元，為了使斷詞的結果更好。構詞規則構出的詞種類很多，當我們詞典沒包含到這些詞時，這些詞即為 OOV，且構詞規則構出的詞是造成 OOV 的元兇，所以用構詞的最小單元取代【5】。構詞規則的詞數量很多，可以用構詞最小單元取代，當詞典收錄構詞最小單元，可以減少 OOV 的數量。構詞的最小單元詳列於附錄五。例子如下表格：

表 4-2 構詞規則所構出的詞以構詞最小單元代替

構詞規則所構出的詞			以構詞最小單元代替
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	一百二十二度十	一 百 二 十 二 度 十

(2)人名用一字詞串代替

文章使用交大語音實驗室詞典(表 3-8)斷詞，詞典包含大量的人名，這對斷詞結果是有幫助的，當詞典沒收錄人名此詞時，會造成 OOV，而大量的 OOV 大多都是人名造成的，所以把人名以一字詞串取代，以減少 OOV 的數量。例子如下表格：

表 4-3 人名以一字詞串代替

人名	**一字詞串代替**
陳水扁	陳 水 扁
馬英九	馬 英 九
周杰倫	周 杰 倫
郭台銘	郭 台 銘
梁朝偉	梁 朝 偉

下表則是兩大資料庫經長詞變短詞(構詞規則構出的詞，以構詞最小單元代替、人名用一字詞串代替)之後的 word 及 character 統計表：

表 4-4 長詞變短詞後的 word 及 character 統計表

訓練語料	詞數（Word）	字數（Character）
光華雜誌	9,526,689	15,699,898
NTCIR	120,844,181	201,390,164
合計	130,370,870	217,090,062

文章的平均詞長= 1.67。

(3)OOV 的分析_長詞變短詞

原始文字資料庫經由以上的程序，處理後的文章可經由資料統計，統計出每一個詞在文章中出現的次數，詞頻愈大表示此詞在文章中出現的次數愈多，此詞會比其他詞頻小的重要。在語言模型的詞典或是聲學模型的詞典上，都是經由統計詞頻，把詞頻高的納入詞典，不在詞典的詞即為 unknown word，就是 OOV(out of vocabulary)。詞典所收錄的詞，在文字資料庫所佔的比例為 Cover rate，而 OOV 佔的比例為 OOV rate，OOV 愈多，表示在語音辨認中有很多詞無法被辨認，因為詞典沒有收錄那個詞，所以 OOV rate 會間接影響到語音辨認結果。

統計詞頻，詞頻高的前六萬詞作為詞典，不在這詞典的詞，就是 OOV Words(不包含人名、構詞規則構出的詞)，利用前六萬詞和一字詞取代 OOV Words，就是長詞變短詞，降低 OOV rate。兩大資料庫經長詞變短詞(OOV words 轉換六萬詞和一字詞取代)後的 word 及 character 統計表：

表 4-5 OOV 以短詞取代後 word 及 character 統計表

訓練語料	詞數（Word）	字數（Character）
光華雜誌	9,772,232	15,699,898
NTCIR	122,442,525	201,390,164
合計	132,214,757	217,090,062

文章的平均詞長= 1.64。

4.2 建立 LM 流程圖

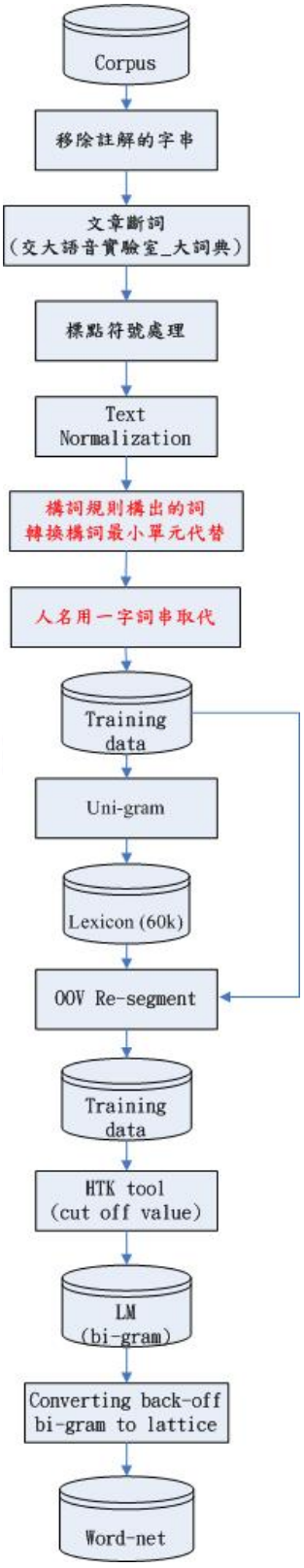

圖 4-1：實驗三-語言模型建立流程圖

4.3 語言模型建立

藉由輸入大量的文字資料，統計出各種詞串在文章中累計的出現次數後，利用 smoothing 的方法建立 LM 所需的 n-gram 機率，使用 HTK tool 分別求出 unigram、bigram、trigram。

4.3.1 HTK Training data：

Database 經由 Segment_Lexicon(表 3-8)斷詞、文字正規化、同音同義異詞、構詞規則構出的詞轉換構詞最小單元代替、人名用一字詞串取代的處理。不在 Language Model_Lexicon 內的詞(OOV)，經由詞典的詞和一字詞取代。

4.3.2 Language Model_Lexicon：

Database 經由 Segment_Lexicon(表 3-8)斷詞、經由文字正規化、同音同義異詞、構詞規則構出的詞轉換構詞最小單元代替、人名用一字詞串取代的處理後，統計文章中的詞頻，取詞頻高前六萬名。

下表為詞典詞長分佈：

表 4-6 實驗三詞典統計表

Language Model_Lexicon		
總詞數	60,000	
詞長	詞數	百分比(%)
1	4234	7.06
2	36941	61.57
3	13390	22.32
4	5054	8.42
5	313	0.52
6	68	0.11

7	0	0
8	0	0

LM 詞典的平均詞長= 2.34。

4.3.3 語言模型測試：

由於辨識語料是 TCC300，所以將 TCC300-train data 視為測試語料，下表為 Perplexity。

表 4-7 測試語料 Perplexity

LM	Perplexity
Word-based	355.6665

4.4 實驗三：修改語料庫&人名、DM 拆成短詞_辨識效能

表 4-8 實驗三-辨識結果

	Deletion	Substitution	Insertion	Accuracy	Total count
word 辨識率	5.14%	24.89%	2.41%	67.56%	19510
character 辨識率	1.15%	21.47%	0.32%	77.04%	31412
syllable 辨識率	1.17%	12.64%	0.33%	85.85%	31412

文章在斷詞後，把人名及構詞規則所構出的詞，使用短詞取代，詞典能收錄較多的短詞，目標是想在辨認時，辨認出較有意義的詞，而不是一字詞串，因為一字詞，在語意上較無意義，長詞會比短詞更有意義。從實驗二和實驗三的比較結果得知，辨識率都有下降的趨勢，從下表格得知：

表 4-9：實驗二和實驗三的辨識率比較

	word accuracy	character accuracy	syllable accuracy
實驗二：辨識率	69.47%	78.39%	86.48%
實驗三：辨識率	67.56%	77.04%	85.85%
辨識率分別下降	1.91%	1.35%	0.63%

TCC300 測試語料中的構詞規則所構的詞和人名。

以 word 統計如下表格：

表 4-10 測試語料中的構詞規則所構的詞和人名之 word 比例

	Number of Word	佔全文章 word 數之百分比(%)
TCC300 測試語料	19,510	
TCC300 測試語料_構詞規則所構的詞	592	3.03%
TCC300 測試語料_人名	245	1.26%
TCC300 測試語料_構詞規則所構的詞 &人名	**837**	**4.29%**

以 character 統計如下表格：

表 4-11 測試語料中的構詞規則所構的詞和人名之 character 比例

	Number of Character	佔全文章 character 數百分比(%)
TCC300 測試語料	31,412	
TCC300 測試語料_構詞規則所構的詞	1,652	5.26%
TCC300 測試語料_人名	723	2.30%
TCC300 測試語料_構詞規則所構的詞 &人名	**2,375**	**7.56%**

word 佔了全部測試語料的 4.29%，character 佔了全部測試語料的 7.56%。

實驗二和實驗三的主要差別，在構詞規則所構的詞和人名以短詞取代，探討是否長詞拆成太多的短詞，造成辨識率下降，也就是辨認是構詞規則所構的詞和人名的正確性，是否為造成辨識率下降的原因。下表格以構詞規則所構的詞來討論：

表 4-12 構詞規則所構的詞結果比較

	Number of Word	Number of Character
TCC300 測試語料_構詞規則所構的詞	592	1,652
實驗二辨識結果_構詞規則所構的詞	587	1,539
實驗三辨識結果_構詞規則所構的詞	548	1,434
實驗三跟實驗二比較，實驗三沒辨認出來	**39**	**105**

下表格以人名來討論：

表 4-13 人名結果比較

	Number of Word	Number of Character
TCC300 測試語料_人名	245	723
實驗二辨識結果_人名	93	268
實驗三辨識結果_人名	64	182
實驗三跟實驗二比較，實驗三沒辨認出來	**29**	**86**

實驗二和實驗三比較，實驗三少辨認出 39 個構詞規則所構的詞，105 個字元。少辨認出 29 個人名，86 個字元。沒辨認出來的詞，會辨認出其他錯誤的詞，辨認出來錯誤的詞會影響到前後的詞，造成前後詞也跟著辨認出錯誤的結果。下表是統計實驗二和實驗三比較，實驗三沒辨認出來的詞數。

表 4-14 實驗二和實驗三比較

	Number of Word	Number of Character
TCC300 測試語料	19,510	31,412
實驗三跟實驗二比較，實驗三沒辨認出來	**68**	**191**

實驗二和實驗三的 word 辨識率比較，實驗三 word 辨識率掉了 1.91%，估算是由沒辨認出來的詞造成前後詞錯誤的部分，長詞由短詞取代後，短詞辨認的結果有的對、有的錯，但在最壞的情形，短詞全辨認錯誤。以人名來說，最多被拆成三個連續一字詞。以構詞規則所構的詞，最多被拆成五至六個詞。估算錯誤的詞，數學式如：68*(4+2)/19510*100=2.09% (數學式裡的「4」表示長詞被拆成短詞的平均詞數，「2」表示前後詞)，這估計是在最壞的前提下，所以 2.09%是高估的值。

實驗二和實驗三的 character 辨識率比較，實驗三 character 辨識率掉了 1.35%。估算字元的錯誤率，數學式如右式：191*3/31,412*100=1.82% (數學式裡的「3」是假設前後字元及本身的字元都辨識錯誤)，估算是由沒辨認出來的字，所造成前後字錯誤的部分，佔了全文的 1.82%。但不代表前後詞一定錯，只是錯的機會比較大，所以 1.82%是高估的值。

從上述得知，沒辨認出來的詞，會影響到此詞及前後詞的正確性，所以我們把有意義的詞，拆成短詞，但拆成的短詞太短，會影響短詞的辨識結果。本實驗三主要把有意義的長詞以短詞取代，辨認的結果是較長且有意義的詞，但拆的太短，會造成辨識率下降。假如把所有的詞都拆成一字詞的組合時，詞典只要收錄全部的一字詞，辨認結果當然會更加的混亂，因為同音不同形的字很多。本實驗三，人名的部分是拆成連續一字詞，拆解的太短，若人名拆成姓和名。構詞規則所構的詞以構詞最小單元取代，構詞最小單元包含很多一字詞，若把構詞規則所構的詞，以較短的構詞規則所構的詞(subword)取代，對辨認率會有所提升。

實驗三 word 的辨識率，辨識出來的 word 是較沒意義的短詞。我們語音辨識的最終目標，是辨識結果有愈多有意義的長詞，所以我們把辨識答案跟較有意義的長詞做辨認。結果如下表：

表 4-15 實驗三-辨識結果跟較有意義的長詞比較

Outside test：Total 18034 words			
Deletion	Substitution	Insertion	Accuracy
3.55%	28.82%	8.78%	58.82%

實驗一：原語料庫。

實驗二：原語料庫經修改、刪除。

實驗三：原語料庫經修改、刪除，並且把有意義的長詞(人名、構詞規則所構出的詞)用短詞取代。

實驗三有意義的長詞被拆成較短的詞，word 的辨識率下降 1.91%。針對結果跟有意義的長詞做辨認，word 的辨識率也是下降。比較實驗二和實驗三，辨識結果跟有意義的長詞做辨認，實驗三比實驗二掉了 2.54%，所以說長詞用短詞取代時，短詞不能太短，否則辨識結果錯誤，錯誤的詞會影響到前後詞的正確率，而要辨認出有意義的長詞是不可能的。

第五章 語音辨認後處理之改良

5.1 分析人名及構詞規則構出的詞、以 Subword 取代

若把構詞規則構出的詞，以 subword 取代、人名用一字詞及二字詞取代(姓、名)。太多的一字詞，會造成辨認錯誤，同音異義異形的情形間接影響前後詞的辨認結果，有意義的長詞若以較長 subword 取代，在語言模型上，相對會減少不確定性。

5.1.1 構詞規則所構出的詞以長的短詞代替

構詞規則構出的詞，以構詞最小單元代替，由於構詞最小單元有太多一字詞，在辨認時，會造成詞和詞之間的不確定性，也就是同音同義異形的一字詞，會造成辨認時的錯誤，導致影響到前後詞辨認結果，使詞的辨識率下降。解決這一字詞造成的不確定性，所以把構詞規則所構的詞，以 subword 取代。構詞的最小單元及 subword，分別詳列於附錄五和附錄六。例子如下表格：

表 5-1 構詞規則構出的詞拆成構詞的最小單元及較長的詞

構詞規則所構出的詞			以構詞最小單元代替	以 Subword 代替
定量複合詞	DM	兩萬平方公尺	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	一千 四百億
數量定詞	Neqa	三十左右	三 十 左右	三十 左右
時間詞	Nd	下午四點多	下午 四 點 多	下午 四點 多
地方詞	Nc	六年十一班	六 年 十 一 班	六 年 十一 班
位置詞	Ncd	一百二十二度十	一 百 二 十 二 度 十	一百 二十二 度 十

由上可知，以構詞規則所構出的詞，以 subword 取代後，包含較少的一字詞，使不確定性下降。

5.1.2 人名用一字詞和二字詞代替

人名以一字詞串取代，一字詞辨認錯誤，造成前後詞受到影響，所以人名以姓、名拆解，使連續的一字詞出現頻率降低，以增加詞和詞之間的可信度，間接影響到辨識人名的正確率，相對的比連續一字詞串的辨認，少了同音異義異形的困擾。例子如下表格：

表 5-2 人名以姓、名拆解

人名	一字詞串代替	長詞代替
陳水扁	陳 水 扁	陳 水扁
馬英九	馬 英 九	馬 英九
周杰倫	周 杰 倫	周 杰倫
郭台銘	郭 台 銘	郭 台銘
梁朝偉	梁 朝 偉	梁 朝偉
劉嘉玲	劉 嘉 玲	劉 嘉玲

5.2 建立 LM 流程圖

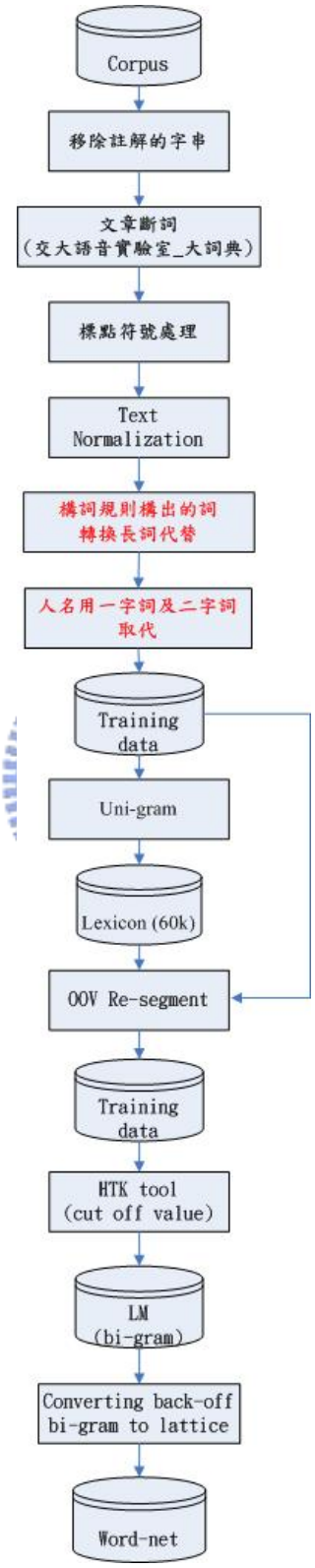

圖 5-1 實驗四-語言模型建立流程圖

5.3 語言模型(LM)建立

藉由輸入大量的文字資料，統計出各種詞串在文章中累計的出現次數後，利用 smoothing 的方法建立 LM 所需的 n-gram 機率，使用 HTK tool 分別求出 unigram、bigram、trigram。

5.3.1 HTK Training data：

Database 經由 Segment_Lexicon(表 3-8)斷詞、文字正規化、同音同義異詞、構詞規則構出的詞以構詞最小單元及 subword 取代、人名用一字詞(姓)和二字詞(名)取代。不在 Language Model_Lexicon 內的詞(OOV)，經由詞典的詞和一字詞取代。

5.3.2 Language Model_Lexicon：

Database 經由 Segment_Lexicon(表 3-8)斷詞、文字正規化、同音同義異詞、構詞規則構出的詞以構詞最小單元及 subword 取代、人名用一字詞(姓)和二字詞(名)取代後，統計文章中的詞頻，取詞頻高前六萬名。

下表為詞典詞長分佈：

表 5-3 實驗四詞典統計表

Language Model_Lexicon		
總詞數	60,000	
詞長	詞數	百分比(%)
1	3979	6.63
2	38458	64.10
3	12358	20.60
4	4864	8.11

5	283	0.47
6	58	0.10
7	0	0
8	0	0

LM 詞典的平均詞長= 2.32。

5.3.3 語言模型測試：

由於辨識語料是 TCC300，所以將 TCC300-train data 視為測試語料，下表為 Perplexity。

表 5-4 測試語料 Perplexity

LM	Perplexity
Word-based	379.7348

5.4 實驗四：修改語料庫&人名、構詞規則所構出的詞，以較長的短詞取代_辨識效能

表 5-5 實驗三-辨識結果

	Deletion	Substitution	Insertion	Accuracy	Total count
word 辨識率	5.22%	24.54%	2.10%	68.13%	19182
character 辨識率	1.15%	21.15%	0.32%	77.36%	31412
syllable 辨識率	1.16%	12.46%	0.33%	86.03%	31412

從上述可知，把人名及構詞規則所構出的詞，使用較長 subword 取代，使辨認率上升，詞和詞之間的不確定性降低，估算詞和詞之間的機率更加準確。長詞拆成較長的 subword，可降低連續一字詞的辨認錯誤。由實驗三和實驗四比較，實驗三是以較短的短詞取代，實驗四是以 subword 取代，可知長詞以 subword 取代的辨識率有提升。實驗三和實驗四的比較結果，從下表格得知，辨識率都有上升的趨勢。

表 5-6 實驗三實驗四的辨識率比較

	word accuracy	character accuracy	syllable accuracy
實驗三：辨識率	67.56%	77.04%	85.85%
實驗四：辨識率	68.13%	77.36%	86.03%
辨識率分別提升	**0.57%**	**0.32%**	**0.18%**

下表格以構詞規則所構出的詞來討論：

表 5-7 構詞規則所構的詞結果比較

	Number of Word	Number of Character
TCC300 測試語料_構詞規則所構出的詞	592	1,652
實驗二辨識結果_構詞規則所構出的詞	587	1,539
實驗三辨識結果_構詞規則所構出的詞	548	1,434
實驗四辨識結果_構詞規則所構出的詞	572	1,522
實驗三跟實驗四比較，實驗四多辨認出來	**24**	**88**

下表格以人名來討論：

表 5-8 人名結果比較

	Number of Word	Number of Character
TCC300 測試語料_人名	245	723
實驗二辨識結果_人名	93	268
實驗三辨識結果_人名	64	182
實驗四辨識結果_人名	106	308
實驗三跟實驗四比較，實驗四多辨認出來	**42**	**126**

由上兩表可知，把人名拆成姓和名，會增加詞和詞之間的可信度，不會因為一字詞的錯誤，造成前後詞的錯誤。構詞規則所構出的詞，也是同樣的道理，只是本階段的拆解是針對數字的部分，把數字拆解成 subword。

實驗三 word 的辨識率，辨識出來的 word 是較沒意義的短詞。我們語音辨識的最終目標，是辨識結果有愈多有意義的長詞，所以我們把辨識答案跟較有意

義的長詞做辨認。結果如下表：

表 5-9 實驗四-辨識結果跟較有意義的長詞比較

Outside test：Total 18034 words			
Deletion	Substitution	Insertion	Accuracy
3.83%	28.64%	6.88%	60.64%

實驗一：原語料庫。

實驗二：原語料庫經修改、刪除。

實驗三：原語料庫經修改、刪除，並且把有意義的長詞(人名、構詞規則所構出的詞)用短詞取代。

實驗四：原語料庫經修改、刪除，並且把有意義的長詞(人名、構詞規則所構出的詞)用較長的短詞取代(subword)。

由於實驗三有意義的長詞被拆成較短的詞，實驗三 word 的辨識率跟實驗二的辨識率比較，word 的辨識率下降 1.91%。針對結果是有意義的長詞做辨認，當然 word 的辨識率也是下降的，而實驗四則把有意義的長詞，以較長的短詞(subword)取代，實驗四 word 的辨識率跟實驗三的辨識率比較，提升了 0.57%，由此可知，有意義的詞拆成較長的 subword，有較好的辨識結果。

實驗四是針對數字的部分，把數字拆解成較長的 subword，而人名則是拆成姓和名。若未來把有意義的詞(構詞規則所構的詞)，拆解的更好、更有意義，相信對辨識率會有可觀的提升。

第六章 結論與展望

6.1. 結論

在本論文中，我們使用 TCC300 語料庫來進行語音辨識的相關研究，從基本系統的建立、文章內容的刪除、內容錯誤的更正處理，為了辨識的結果為有意義的長詞，而非是較無意義的短詞，所以針對構詞規則所構出的詞及人名，進行長詞變短詞處理。短詞的長度會影響到辨認結果與語言模型的不確定性，本論文將幾個重點分列如下：

（1） 文章內容的好壞，在統計語言模型時，估算詞和詞之間機率的正確性、可靠度，間接的影響到語言模型，內容的好壞對於辨認結果有所影響。

（2） 由於想要辨認結果是較有意義的長詞，詞典又無法收錄全部的詞，所以把較有意義的長詞，用短詞取代，減少 OOV 的數量。主要目的是要辨認結果是較有意義的短詞，而非是一字詞串。

（3） 有意義的長詞以較短的短詞取代，在辨認時短詞的錯誤會間接影響到前後詞的辨認結果，導致辨認率下降，若長詞以較長的 subword 取代，較少的一字詞串，對於辨認時減少一字詞串的不確定性。所以有意義的長詞，以較長的 subword 取代後的辨認率，會比以較短的短詞取代的辨認率來的好。

6.2. 展望

辨認結果的最終目的是辨認出來的詞為較長且有意義的，本實驗只針對結果和最後目的做分析，也就是辨認結果和較長且有意義的詞做分析，沒進一步處理。若把有意義的長詞，用較長的 subword 取代後，並且短詞全收錄在詞典內，而短詞的收錄極為重要，就是本論文提到，以較短的短詞取代長詞會使辨識率下降，所以短詞的收錄是一門學問。未來可利用 two-stage【8】【9】【10】的想

法，從辨認結果的 word lattice 重新估算分數，使辨認結果構回較有意義的長詞，就是在第一級辨認出 word lattice，再分別建立有意義的長詞模型，例如：人名的模型、構詞規則所構出的詞之模型，利用模型及 word lattice 重新計算分數，使辨認出來的詞是一句長且有意義的長詞，而非第一級所辨認的短詞串。

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，中文斷詞器之改進，國立交通大學電信工程學系碩士論文，民國九十三年七月

【6】張隆勳，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，民國九十四年七月

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

【9】Shigehiko Onishi, Hirofumi Yamamoto, “Structured language model for class identification of out-of-vocabulary words arising from multiple word-classes”

【10】Koichi Tanigaki, Hirofumi Yamamoto, Yoshihiki Ogawa, and Yoshinori Sagisaka, “Out-of-vocabulary word recognition with a hierarchical doubly markov language model”

附錄一

括弧【】及括弧【】內的例子：（＊表示某一個中文字）

原始文章的例子	格式	刪除個數
【前言】	【前言】	11
【記者葉建田中正機場報導】 【記者葉建田台北報導】 【記者吳秉鍇高雄報導】	【＊＊＊＊＊＊＊報導】	97,401
【本報東南亞特派員梁東屏二十日電】 【本報二十一日上海電】 【本報十五日香港電】	【＊＊＊＊＊＊＊＊＊電】	9,950
【記者張佩芬台北－香港電話採訪】 【記者謝文大陸巡迴採訪】 【記者邱裕榮調查採訪】	【＊＊＊＊＊＊＊採訪】	561
【記者于國欽新聞分析】 【記者魏冠中新聞分析】	【＊＊＊＊＊＊＊分析】	614
【高雄訊】 【鳳山訊】 【本報訊】	【＊＊＊＊＊＊＊＊訊】	5,710
【記者胡清揚／好萊塢專訪】 【記者徐秀美台北專訪】 【記者林宛諭/專訪】	【＊＊＊＊＊＊＊專訪】	802
【本報記者康文炳、江今葉、仇佩芬摘譯】 【陳虹妙摘譯】 【謝富旭摘譯】	【＊＊＊＊＊＊＊摘譯】	7
【中央社記者鍾行憲華盛頓特稿】 【費國禎／特稿】 【記者潘罡特稿】	【＊＊＊＊＊＊＊特稿】	673
【美聯社紐約十三日】 【東京記者王淑珍／廿二日】	【＊＊＊＊＊＊＊＊日】	9
【本報記者郭崇倫】 【本報記者江靜玲】	【本報＊＊＊＊＊＊＊】	310
【新加坡十日美聯社】 【昆明二日法新社】 【路透社】	【＊＊＊＊＊＊＊＊社】	10

【編者按】	【＊＊＊＊＊＊＊＊按】	2
【記者謝文台北報告】 【記者呂曼文台北報】 【記者錢劍民執筆】	【記者＊＊＊＊＊＊＊】	142
【專題報導／林憲洋】 【專題報導／李采洪】 【專題報導熊正容】	【專題＊＊＊＊＊＊】	10
【文／記者李碧勳】 【文／方素惠】	【文／＊＊＊＊＊＊＊】	6
【文接昨日第十二版】 【文轉九版】 【前瞻版】	【＊＊＊＊＊＊＊＊版】	7
【法新社威瑪市二十日雷】 【法新社波昂廿日雷】 【中央社香港雷】	【＊＊＊＊＊＊＊＊雷】	13
【本報特派記者宋秉忠、徐尚禮整理】 【記者張錫銘記錄整理】 【王榮章／整理】	【＊＊＊＊＊＊＊整理】	322
【新聞分析／記者謝孟儒】 【新聞幕後／夏珍】 【新聞縫裡】	【新聞＊＊＊＊＊＊＊】	10
【呂庭華／新聞切片】 【劉寶傑／新聞切片】	【＊＊＊＊＊新聞切片】	11
【戎撫天／新聞探索】	【＊＊＊＊＊新聞探索】	2
【李祖舜／新聞幕後】 【張麗伽／新聞幕後】	【＊＊＊＊＊新聞幕後】	17
【社評】	【社評】	87
【記錄整理／陳碧芬】	【記錄整理＊＊＊＊＊】	1
【慕尼黑特派員陳玉慧／三日奧斯陸長途電話訪問】	【＊＊＊＊＊＊＊訪問】	2
【劉黎兒／東京傳真】 【陳世宗／社會傳真】	【＊＊＊＊＊＊＊傳真】	34
【郭淑敏、劉寶傑／人物側寫】 【李義／人物側寫】	【＊＊＊＊＊＊＊側寫】	6
【看問題／楊肅民】 【看問題】	【看問題＊＊＊＊＊＊】	3
【作者簡介】	【作者＊＊＊＊＊＊＊】	1
【前瞻】	【前瞻】	4

【社論】	【社論】	4
【文：田媛】	【文：＊＊＊＊＊＊＊】	1
【轉載自本期新新聞周報】	【＊＊＊＊＊＊＊周報】	2
【簡介】	【簡介】	1
【精油內部情報】	【＊＊＊＊＊＊＊情報】	18
【李芬芳／取材自時代雜誌】	【＊＊＊＊＊＊＊雜誌】	2
【摘自福爾摩沙之憾】	【摘自＊＊＊＊＊＊＊】	1
【人物側寫／李宛蓉】	【人物側寫＊＊＊＊＊】	1
【廖嘯龍、孟祥傑／刑案追蹤】	【＊＊＊＊＊＊＊追蹤】	2
【富邦證券研究員陳永俊】	【富邦證券研究員＊＊】	1
【夏韻芬／編輯台】	【＊＊＊＊＊＊編輯台】	4
【黃文財／整體】	【＊＊＊＊＊＊＊整體】	1
【京華投顧詹俊南】	【京華投顧＊＊＊＊＊】	1
【本文作者林文山為旅居印尼台商】	【本文作者＊＊＊＊＊】	1
【一一二一專案】	【＊＊＊＊＊＊＊專案】	2
【政壇走筆張啟楷】	【政壇走筆＊＊＊＊＊】	1
【國際投信基金經理林長杰】	【國際投信＊＊＊＊＊】	4
【關於協同辦案對象自選】 【關於主任檢察官票選】	【關於＊＊＊＊＊＊＊】	3
【公式一】	【公式＊＊＊＊＊＊＊】	2
【託者袁世珮／綸合美國外電軒】	【＊＊＊＊＊＊外電軒】	1
【屏東訊導】	【屏東訊導】	1
【中新杜泰安九月九日重】	【＊＊＊＊＊＊＊＊重】	1
【路透北京六日需】	【＊＊＊＊北京六日需】	1
【視覺空間類】 【生活消費類】	【＊＊＊＊＊＊＊＊類】	6
【普】	【普】	5
【國】	【國】	1

附錄二

括弧（）及括弧（）內的例子：

原始文字串	修改過後文字串	修改的個數
今（六十五）年外銷 決定在**今**（三十一）日前往北京	今年外銷 決定在今日前往北京	3,352
陳方安生**昨**（二十三）日偕同 **昨**（三十一）日深夜十一時	陳方安生昨日偕同 昨日深夜十一時	14,857
將於**明**（六十六）年初陸續完成 預計**明**（十九）日松花江幹流	將於明年初陸續完成 預計明日松花江幹流	1,599
人權協會將於**後**（四）日舉辦 報稅最快**後**（八十九）年實施	人權協會將於後日舉辦 報稅最快後年實施	262
都已於**去**（８４）年十二月 迄**去**（八十七）年底為止	都已於去年十二月 迄去年底為止	242
上午與**前**（二十一）日下午接見 半年**前**（八十七年七月廿一日）臺指期貨	上午與前日下午接見 半年前臺指期貨	627
國家建設研究會**本**（八）月一日上午 財政部公布的**本**（四）月上旬	國家建設研究會本月一日 財政部公布的本月上旬	687
下周**一**（二十九日）抵達上海 下周**一**（廿六日）下午於警政署	下周一抵達上海 下周一下午於警政署	962
下周**二**（九日）依往例 本周**二**（十二日）晚間進行制海	下周二依往例 本周二晚間進行制海	489
已於本周**三**（一日）原則通過 本周**三**（二十日）屆滿兩周年	已於本周三原則通過 本周三屆滿兩周年	495
理事會上周**四**（二日）決定 上周**四**（二十一日）通過「國防法」	理事會上周四決定 上周四通過「國防法」	349
預定本周**五**（十七日）即將 黨團決定本周**五**（五日）召開	預定本周五即將 黨團決定本周五召開	554
央行在上周**六**（一日）宣布 本周**六**（六日）晚間返回臺北	央行在上周六宣布 本周六晚間返回臺北	1,043
一月五**日**（星期二）年度 下周**日**（七日）高雄市重新辦理	一月五日年度 下周日高雄市重新辦理	1,471

附錄三

括弧（）及括弧（）內的字串格式：（＊表示某一個中文字）。

分類	格式
文章來源註解	＊＊＊（＊＊＊訊）
	＊＊＊（＊＊＊導）
	＊＊＊（＊＊＊攝）
	＊＊＊（＊＊＊理）
	＊＊＊（＊＊＊報）
	＊＊＊（＊＊＊譯）
	＊＊＊（美聯社）
	＊＊＊（作者＊＊）
	＊＊＊（＊＊＊版）
	＊＊＊（中國日報＊）
人名註解	＊＊＊（＊＊＊歲）
	＊＊＊（＊＊＊圖）
	＊＊＊（＊＊＊縣）
	＊＊＊（＊＊＊市）
	＊＊＊（＊＊＊人）
	＊＊＊（＊＊＊籍）
	＊＊＊（＊＊＊音）
	＊＊＊（＊＊＊州）
	＊＊＊（＊＊＊男）
	＊＊＊（＊＊＊鄉）
	＊＊＊（＊＊＊譯音）
時間註解	＊＊＊（＊＊＊年）
	＊＊＊（＊＊＊月）
	＊＊＊（＊＊＊日）
	＊＊＊（＊＊＊時）
	＊＊＊（＊＊＊分）
	＊＊＊（周＊＊＊）
職稱、單位數註解	＊＊＊（＊＊＊家）
職稱註解	＊＊＊（＊＊＊員）
	＊＊＊（＊＊＊授）
	＊＊＊（＊＊＊傷）

	＊＊＊（＊＊立委）
	＊＊＊（＊＊記者）
組織註解	＊＊＊（＊＊＊會）
	＊＊＊（＊＊＊團）
	＊＊＊（＊＊＊派）
	＊＊＊（＊＊＊組）
	＊＊＊（＊＊＊店）
	＊＊＊（＊＊＊所）
	＊＊＊（＊＊＊織）
	＊＊＊（＊＊＊黨）
	＊＊＊（＊＊＊無）
文章來源、公司縮寫	＊＊＊（＊＊＊電）
	＊＊＊（＊＊＊司）
學歷註解	＊＊＊（＊＊＊校）
	＊＊＊（＊＊＊院）
單位、時間註解	＊＊＊（＊＊＊度）
圖片註解	＊＊＊（＊＊＊表）
	＊＊＊（＊＊＊左）
	＊＊＊（見圖＊）
單位註解	＊＊＊（＊＊＊票）
	＊＊＊（＊＊＊元）
	＊＊＊（＊＊＊％）
	＊＊＊（＊＊＊區）
路名註解	＊＊＊（＊＊＊段）
書名註解	＊＊＊（＊＊編著）
影片註解	＊＊＊（＊＊＊護）
無意義註解	＊＊＊（＊＊＊㊣）
國家名註解	（新加坡）（大陸）（台灣）（日本）（韓國）（香港）（臺灣）

附錄四

括弧（）及括弧（）內為中文的例子：

分類	原始文字串	修改過後文字串	刪除個數
文章來源註解	（何凱彰・臺北**訊**）那魯灣	那魯灣	7,823
	（綜合外電報**導**）美國亞特蘭大	美國亞特蘭大	1,140
	投手布朗（右圖，路透**攝**）掛帥主投	投手布朗掛帥主投	318
	（黃秀仁・整**理**）社福王	社福王	2,575
	（紐約時**報**）台灣	台灣	624
	（記者賈亦珍／輯**譯**）衝動性性濫交等	衝動性性濫交等	866
	通訊公司。（**美聯社**）	通訊公司。	289
	直接翻譯（**作者**一寫好，即拷貝交翻譯）	直接翻譯	715
	「瞬間收藏家」（全由格林文化出**版**）	「瞬間收藏家」	679
	（**中華日報**記者陳舜華）我想延續	我想延續	444
人名註解	李明勳（廿九**歲**），身高	李明勳，身高	11,906
	周妙玲（見**圖**）、世界	周妙玲、世界	731
	陳明壽（卅二歲，彰化**縣**）	陳明壽	473
	黃群寶（三十四歲、住台北**市**），黃瑞景（五十五歲，桃園**人**）駕駛的	黃群寶，黃瑞景駕駛的	2,355 3,111
	黃主文（國民黨**籍**）	黃主文	142
	姚儀（譯**音**）和入籍	姚儀和入籍	829
	傑福茲（佛蒙特**州**）	傑福茲	228
	黃聰仁（**男**）、李光	黃聰仁、李光	118
	陳信安是同學（均十六歲，住高雄縣大寮**鄉**）	陳信安是同學	208
	主教任澤強（**譯音**）在米蘭晚郵報	主教任澤強在米蘭晚郵報	829
時間註解	二零零一年（民國九十**年**）	二零零一年	2,166

	新進黨因而分裂(一九九六年九**月**)	新進黨因而分裂	540
	時間（十二月十四日至十九**日**）	時間	1,698
	上午９時（台北時間１０**時**）在總統	上午９時在總統	698
	五十五分(臺北凌晨二時五十五**分**)	五十五分	1,126
	三台前天（**周**日）晚上的	三台前天晚上的	302
職稱、單位數	薩雪莉（哥斯大黎加「經濟現況雜誌」社長暨專欄作**家**）	薩雪莉	705
職稱	特別獎（最有價值球**員**）為高苑	特別獎為高苑	1,045
	孫中興（臺大社會系教**授**）Ｄ持人	孫中興Ｄ持人	665
	邱創煥（台灣省主**席**）	邱創煥	205
	主任（國民黨青工會主**任**）、主委	主任、主委	475
	職歷（立委或非**立委**）	職歷	306
	曾萬（中華電視公司**記者**）、	曾萬、	947
組織註解	委員會（簡稱精省委員**會**）第二次	委員會第二次	798
	青年團（馬青**團**）總團長	青年團總團長	159
	張清源（屬農會**派**）	張清源	154
	弊端小組（東部機動**組**）	弊端小組	195
	企業化（酒**店**）經營	企業化經營	199
	台灣證券交易所（證交**所**）也將在明年	台灣證券交易所也將在明年	174
	沙國科技中心（類似我國國科會的組**織**）	沙國科技中心	183
	綠色本土清新黨（簡稱綠**黨**）	綠色本土清新黨	446
	林正杰（**無**）	林正杰	427
文章來源、公司縮寫	（特派記者阮玫芬廿六日專**電**）為介紹	為介紹	125
	佐川急便（快遞公**司**）的金錢醜聞	佐川急便的金錢醜聞	497
學歷註解	梅博凱（泰國中華國際學**校**）	梅博凱	149
	廖若芳（菲律賓中正學**院**）	廖若芳	335
單位、時間註解	明年（八十八學年**度**）首度	明年首度	440

圖片註解	招生單位（詳見附**表**）	招生單位	692
	基諾李維（見右圖**左**）在「駭	基諾李維在「駭	434
	好下場了（**見圖**表）。	好下場了。	377
單位註解	聖先師孔子票數最高（十**票**）	聖先師孔子票數最高	297
	就業獎助（每人每月一萬五千**元**）	就業獎助	4,373
	韓國（一．一%）還低	韓國還低	1,538
	第四選區（新興、前金、苓雅**區**）	第四選區	505
路名註解	松平路（松仁路至松智路**段**）兩旁早	松平路兩旁早	175
書名註解	戀戀北投溫泉（洪德仁**編著**）	戀戀北投溫泉	5
影片註解	樓學賢演的「三十而立」（**護**）	樓學賢演的「三十而立」	57
無意義註解	家專資投與（㊣㊣㊣）終數指業工	家專資投與終數指業工	112
國家名註解	（**新加坡**）（**大陸**）（**台灣**）（**日本**）（**韓國**）（**香港**）（**臺灣**）		394

附錄五

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 昌午
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 居 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 鉤 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 關 顆
點 叢 罈 鎮 雙 鞭 題 壟 櫥 瓣 邊 關 類 籃 覺 齣
欄 響 彎 攤 灘 疊 籠 聽 襲 罐 廳 抔 坨 厘 灶 術
絡 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 鉤兒 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 辨士 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

附錄六

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