國立交通大學

資訊工程學系

碩士論文

中文語音辨識中語言模型的強化之研究

The Study of Language Model Enhancement for Mandarin Speech Recognition

研究生　：呂宜玲

指導教授：傅心家 教授

中華民國九十四年七月

中文語音辨識中語言模型的強化之研究

The Study of Language Model Enhancement for Mandarin Speech Recognition

研 究 生：呂宜玲　　Student: Yi-Ling Lu

指導教授：傅心家　教授　　Advisor: Prof. Hsin-Chia Fu

國 立 交 通 大 學

資 訊 工 程 學 系

碩 士 論 文

A thesis Submitted to Institute of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Chiao Tung University

In Partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science and Information Engineering

July 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年七月

中文語音辨識中語言模型的強化之研究

研究生：呂宜玲　　指導教授：傅心家　教授

國立交通大學資訊工程學系碩士班

摘要

『資料稀疏』為語音辨識中語言模型極需克服之問題，目前常用的『凱氏平滑法』(Katz Smoothing)及『聶氏平滑法(Kneser-Ney Smoothing)』(包含『聶氏後退法(Kneser-Ney Backoff Smoothing)』及『聶氏內插法(Knerser-Ney Interpolation Smoothing)』)，其應用於雙連馬可夫模型時，對於欲估計的未知雙連詞串詞尾沒有出現於訓練語料的情況，並無適當的機率評估方法。針對此點，我們提出『強化凱氏平滑法』與『強化聶氏平滑法』加以改進，我們由詞尾曾出現於訓練語料但整個雙連詞串並無出現於訓練語料的未知雙連詞串做平滑化後的機率，再進一步扣除小部分機率值，將此被折扣的機率量分配給詞尾未出現於訓練語料的未知雙連詞串，並以混淆度(perplexity)為效能的評量標準。我們由華視網站收集一年的新聞語料，每月取 180 則新聞為測試語料，其餘新聞為訓練語料；由實驗結果可知『強化凱氏平滑法』比原來『凱氏平滑法』平均低了 6.65 個混淆度單位；『強化聶氏平滑法』比原來『聶氏平滑法』中效能較佳的『聶氏後退法』平均低了 4.50 個混淆度單位。此外，也將建構的雙連馬可夫模型及實驗結果中效能最佳的『強化聶氏平滑法』，應用於中文語音辨識系統的語言模型部分，經實際測試，系統正確率可達 88.62%，精確率可達 85.52%。

The Study of Language Model Enhancement for Mandarin Speech Recognition

Student : Yi-Ling Lu Advisor : Prof. Hsin-Chia Fu

Institute of Computer Science and Information Engineering
National Chiao Tung University

Abstract

In this thesis, we propose smoothing methods to solve the "data sparseness" problems of the language model to improve the efficiency of speech recognition.

"Katz Smoothing" and "Kneser-Ney Smoothing" which includes "Kneser-Ney Backoff Smoothing" and "Kneser-Ney Interpolation Smoothing" are the most popular smoothing methods. However, these methods for bigram models don't consider about the unseen bigrams with the last phrase which does not occur in training data. So we proposed the improving methods to discount small amount of probability from smoothed bigrams which does not occur in training data but the last phrase of them occurs in training data. And then, we distribute the discounted mass of probability to bigrams which does not occur in training data with the last phrase does not occur in training data, either. We use perplexity to measure the efficiency of our language model.

We collect the experiment corpus from daily news on Chinese TV System (CTS) website. We take 180 news from 12 months to be testing data, others to be training data. From "Enhanced Katz Smoothing", we obtain a perplexity which is 6.65 lower than for the "Katz Smoothing". And from "Enhanced Kneser-Ney Smoothing", we also obtain a perplexity which is 4.50 lower than for the "Kneser-Ney Backoff Smoothing".

Besides, we implement the bigram Markov language models and "Enhanced Kneser-Ney Smoothing" which performs best in our experiment to a Mandarin speech recognition system. The correct rate of system is 88.62%, and the accuracy of the system is 85.52%.

誌謝

謝謝傅老師在我念研究所的這兩年來所給予的照顧和指導，並幫助我在語音的領域找到研究的方向，也學到做研究時的方法與態度。同時，感謝永煜、柏伸、政龍、岳宏及士賢學長，不論在日常生活或學業上，給予我建議及指教，我由你們的身上學到許多。特別感謝士賢學長，不厭其煩地幫我解決困難及修改論文。感謝揚智及宗儒，這兩年來一起同甘共苦，互相加油及打氣。謝謝三位學弟，建榮、富評及政邦，實驗室因為你們而充滿了歡笑。感謝大學同學們，nash, yhyu, luph, siiy, killer, sjyang, chonsi, koncon…在生活上給我鼓勵及照顧。最後，感謝爸爸、媽媽、妹妹，給我無憂的生活環境，讓我可以專注於學業上的研究。

目錄

表目錄

圖目錄

第 1 章

前言

1.1 研究動機

隨著科技日漸發達，電腦已漸漸融入每個人的生活當中，因此，如何設計更方便的使用者介面，也逐漸變成一個重要的課題。當非專業使用者在透過電腦輸入資料時，尤其在中文環境下，傳統上以注音方式、拆解字形(如倉頡、嘸蝦米輸入法…等)或筆劃順序(如行列輸入法)等鍵盤輸入資料的方式，都較英文文字輸入困難許多，對於他們而言，必須長時間的學習，輸入方式不但緩慢且具有多重選擇的問題，這些無疑是使用上的障礙。基於上述原因，我們希望能找到一種簡單、迅速的輸入方式，以取代鍵盤輸入。

事實上，因為中文的特性是一個音至少對應到一個中文字，而音的總數有限，故使用聲音輸入，會是最理想的選擇。而今日對於語音的研究，已經可以讓使用者藉著聲音與個人電腦溝通，使得語音擔任起『人們與電腦間的橋樑』這個角色，也就是說，電腦可以透過語音訊號的分析，將聲音解碼成正確的中文字，來分辨人所說的話，因此，使用者只需透過聲音，就可以用最輕鬆、自然、友善的方式，讓電腦知道我們想要輸入的訊息。

因此，我們希望電腦能夠具備語音辨識的功能，也就是說，希望電腦能夠藉由使用者輸入的聲音訊號，透過聲音模型的參數比對，得到對應的『候選音節』，再由這些候選音節中找出最可能對應的『候選詞串』。而由『候選音節』找出『候選詞串』這個步驟，就必需藉由『語言模型』來加以完成。

目前最普遍的語言模型，也是本論文所選擇使用的雙連馬可夫語言模型，最主要的做法就是收集語料中的詞成對出現的機率，以評估候選詞串的機率。此模型的優點是模型雖然簡單，卻可擁有非常好的評估效果；但其最大的困難在於當實際語料中的詞沒有在訓練語料中出現時，將會造成資料稀疏的問題，這會造成整句話的機率為零的錯誤情況。

為了克服此問題，已經有許多平滑化方法被提出，最常見的是『凱氏平滑法』(Katz Smoothing)及『聶氏平滑法』(Kneser-Ney Smoothing)，它們應用於雙連馬可夫模型上的做法，都是將已知的雙連詞串機率加以折扣，再依照欲估計的未知的雙連詞串詞尾於訓練語料中出現的比例來加以分配，但若遇到未知的雙連詞串詞尾並沒有出現在訓練語料中的情況時，還是會產生機率為零的問題。因此，本論文提出『強化凱氏平滑法』及『強化聶氏平滑法』，目的是由詞尾曾出現於訓練語料但整個雙連詞串並無出現於訓練語料的未知雙連詞串做平滑化後的機率，再進一步扣除小部分機率值，將此被折扣的機率量分配給詞尾未出現於訓練語料的未知雙連詞串，以避免機率為零的情況。我們採用混淆度為語言模型的效能評估標準，希望上述方法能使得語言模型的混淆度降低，也就是使得機率的評估更為精確。

1.2 章節介紹

在以下的章節中，第二章對語言模型在語音辨識中所扮演的角色及語言模型的相關技術做了簡單的介紹，並介紹幾種語言模型中常見的平滑化方法；第三章則是本論文針對原有平滑化方法的問題所提出的解決策略及方法；第四章是各種平滑化方法的實驗結果比較；第五章是實作語音辨識系統中的語言模型；第六章則是結論及對未來的展望。

第 2 章

語音辨識之語言模型的相關研究

2.1 語音辨識

語音辨識的流程，大致上可用圖 2-1 表示

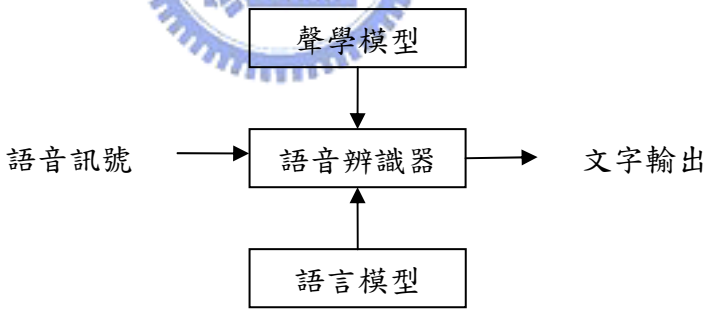


圖 2-1：語音辨識的流程

給定聲學的觀察值 $O = o_1 o_2 \ldots o_n$，語音辨識器的任務是找出擁有(2.1)式中所述之最大事後機率(maximum a posterior probability)的對應候選詞串 $\hat{W} = w_1 w_2 \ldots w_n$.

$$\hat{W} = \arg\max_W P(W \mid O) = \arg\max_W \frac{P(W)P(O \mid W)}{P(O)} \tag{2.1}$$

因為(2.1)式中的觀察值O是固定的，因此，要找出該式的最大值，也就等於使以下的式子最大

$$\hat{W} = \arg\max_{W} P(W)P(O \mid W) \tag{2.2}$$

由(2.2)式可知，語音辨識中的兩個重點就在於建立準確的聲學模型$P(O \mid W)$及語言模型$P(W)$，以正確地辨識出使用者所說的話。

目前的語音辨識技術，語言模型的設計大致可分為文法取向及統計取向:

- 文法取向的作法是根據文法及語意規則定義一些限制，再由剖析器(Parser)加以剖析。優點為易於擷取詞彙的意義，以用於語言處理。缺點為當用於語音辨識時，因為不能處理不合法的句子，因此也不易容忍前端辨識之錯誤。
- 統計取向的語言模型是由機率的觀點來建立語言模型，也就是找出$W = w_1, w_2, \ldots, w_n$的$P(W)$。而作法是先訓練大量的語料，得出機率的統計數據，辨識時再依據這些數據，找出機率最高的句子。優點為必然會辨識出一最有可能的句子，易於容忍前端辨識系統之錯誤。缺點為必需收集大量的語料。

2.2 N 連馬可夫模型

在日常對話中，我們常可藉著對方已經說出的話，去『猜想』接下來對方想要說的話，而 N 連馬可夫模型就是依照這種先前字詞對於下一字詞的預測機率，所發展出來的語言模型，是目前語音辨識上使用最成功的統計取向語言模型。雖然 N 連馬可夫模型並沒有使用高階的語言知識，如語意、文法結構…等等，但其容易建立模型，結構簡單及有效、快速的特性，使其成為語音辨識中語言模型的核心，

同時也是語言模型研究領域的主流，多應用於光學文字辨識、手寫辨識及語音辨識上。

N連馬可夫模型的原理，就是根據已知前N-1 個詞的情況下，預測下一個詞出現的機率【6】。假設W是由$w_1,w_2,\ldots,w_i$所組成，共i個詞的串列。根據貝氏定理，詞串W出現的機率為

$$\begin{aligned} P(W) &= P(w_1, w_2, \cdots, w_i) \\ &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)\cdots P(w_i \mid w_1, w_2, \cdots, w_{i-1}) \\ &= \prod_{k=1}^{i} P(w_k \mid w_1, w_2, \cdots, w_{k-1}) \\ &= \prod_{k=1}^{i} P(w_k \mid w_1^{k-1}) \end{aligned} \tag{2.3}$$

w_a^b 代表詞串$w_a,\ldots,w_b$，$a \leqq b$

而根據最大概似法(maximum likelihood，以下簡稱 ML)

$$P(w_i \mid w_1, w_2, \cdots, w_{i-1}) = \frac{C(w_1, w_2, \cdots, w_{i-1}, w_i)}{\sum_{w_k} C(w_1, w_2, \cdots, w_{i-1}, w_k)} \tag{2.4}$$

$$\text{即 } P(w_i \mid w_1^{i-1}) = \frac{C(w_1^i)}{\sum_{w_k} C(w_1^{i-1}, w_k)} \tag{2.5}$$

其中$C(w_i)$代表詞w_i於訓練語料中出現的次數，而(2.4)式也就代表給定過去可能詞串$w_1,w_2,\ldots,w_{i-1}$的情況下，詞w_i出現的機率

當i值變大時，要計算 $P(w_i \mid w_1, w_2, \cdots, w_{i-1})$將變得困難且不切實際，因此，需要一個簡化的模型，N連馬可夫模型就是假設詞w_i出現的機率只與w_i之前的N-1 個詞相關，而和其他的詞完全獨立。

當 N=1 時(即單連語言模型，或稱零階馬可夫模型)，即對於每一個詞的預測，完全只考慮它於語料中出現的機率，與它前面的詞無關

即 $P(w_i)=\frac{C(w_i)}{S}$，S 為出現在訓練語料中的詞的總數 (2.6)

當 N=2 時(即雙連語言模型，或稱一階馬可夫模型)，對於下一個詞的預測，取決於它和前一個詞同時出現的機率

即 $P(w_i \mid w_{i-1})=\frac{C(w_{i-1},w_i)}{\sum_{w_k} C(w_{i-1},w_k)}=\frac{C(w_{i-1},w_i)}{C(w_{i-1})}$ (2.7)

當 N=3 時(即三連馬可夫模型，或稱二階馬可夫模型)，對於下一個詞的預測，取決於它和前二個詞同時出現的機率

即 $P(w_i \mid w_{i-2},w_{i-1})=\frac{C(w_{i-2},w_{i-1},w_i)}{\sum_{w_k} C(w_{i-2},w_{i-1},w_k)}=\frac{C(w_{i-2},w_{i-1},w_i)}{C(w_{i-2},w_{i-1})}$ (2.8)

至於 N>3，因數量龐大，幾乎鮮少人使用。

2.3 中文語言模型的單位

在英文中，每一個 word(字)都是由許多 character(字母)組成，而最基本且有意義的單位就是一個字，因此，英文的語言模型，就是以"字"(word)為單位。然而，對於多數人而言，中文的『字』(character)及『詞』(word)兩字，在日常生活的使用上，並無明顯差別。事實上，在中文語言學的研究中，這兩者所代表的意義並不相同，每一個『字』都是一個單獨的方塊，但並非每個『字』都會有中文意思。舉例來說，我們都知道『葡萄』一詞，是由『葡』與『萄』兩個字所組成，但兩字分開時，並無任何意思，因此，在中文的語言模型中，『詞』就是最基本且有意義的單位。我們統稱由一個字組成，且有意義的『詞』為『一字詞』，以兩個字組成的『詞』為『雙字詞』，剩下的以此類推。

由語言模型的角度來看，以詞為基礎的模型是利用人類的知識，將某些字之間的關係轉化為辭典內的詞；也就是以詞與前後詞的關係，來代替組成此詞的字與其前後字的關係，這樣可以有效降低N連字串的數目。以『今天 天氣 很 好』為例，對於以詞為基礎的雙連語言模型而言，其計算量只有三個單位的雙連詞串，即(今天,天氣)、(天氣,很)、(很,好)，但是就以字為基礎的雙連語言模型而言，計算量則需要五個單位的雙連字串，即(今,天)、(天,天)、(天,氣)、(氣,很)、(很,好)。

當進一步分析可發現，中文的『詞』包含了明確的意義，所有的句子都可視為由長短不一的『詞』所構成，同時，由【16】的實驗結果可看出，以詞為辨識單位的語言模型，因為加入了人類的知識，會使得估測的結果較以字為單位更具強健性，且計算量也大大降低。

2.4 資料稀疏的解決方式 - 平滑化技術

在實際的情況中，許多詞與詞所形成的詞串，在訓練語料庫中是不存在的，此即資料稀疏(data sparseness)的問題，但並不表示這樣的組合不會在實際的測試樣本中出現。舉例來說，對於雙連語言模型而言，假設辭典中有八萬詞，則可能的雙連詞對數目將會有六十四億個(80,000*80,000=6,400,000,000)，但是一般的訓練語料庫中，通常只會有數千萬到數億個數目的雙連詞串，因此，未知事件(unseen event)將會佔很大的比例，使得許多雙連詞對的機率值等於零，如此會造成整句話的機率 $P(W)$ 為零的情況。

為了解決上述問題，必須對於未知事件的機率值加以平滑化(smoothing)處理，『平滑化』一詞的意思就是將原本的機率值做少許的調整，以產生較準確的機率值，而做法就是扣除小部分的已知事件(seen event)的機率值，分配給未知事件，使得未知事件的機率值不為零。同時，也會使得語言模型在應用了平滑化方法後，

擁有較高的正確率。目前，已有許多平滑化方法被提出，以解決資料稀疏的問題，底下列舉幾種最常見的技術：

2.4.1 加成平滑法(Additive smoothing)

此為最簡單、也最基本的平滑化技術【6】，其應用於雙連馬可夫模型上的做法就是假設每個雙連詞串出現的次數都比原來多了 δ 次，$0<\delta\leqq 1$，因此，根據原來雙連馬可夫模型的公式

$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1}, w_i)}{\sum_{w_k} C(w_{i-1}, w_k)} \tag{2.9}$$

分子多了 δ，分母則多了 V 個 δ，V 代表辭典中所有詞彙的個數(因為對於所有以 w_{i-1} 為詞首的雙連詞串而言，辭典中的每個詞都有可是雙連詞串的詞尾)，故雙連馬可夫模型的加成平滑法公式如下：

$$\begin{aligned} P(w_i|w_{i-1}) &= \frac{\delta + C(w_{i-1}, w_i)}{\sum_{w_k} (\delta + C(w_{i-1}, w_k))} \\ &= \frac{\delta + C(w_{i-1}, w_i)}{\delta V + \sum_{w_k} C(w_{i-1}, w_k)} \\ &= \frac{\delta + C(w_{i-1}, w_i)}{\delta V + C(w_{i-1})} \end{aligned} \tag{2.10}$$

V 代表辭典中所有詞彙的個數。

2.4.2 凱氏平滑法(Katz Smoothing)

凱氏平滑法【8】的主要概念是將所有的已知事件乘上一個折扣值 d_r $(0<d_r\leqq 1)$，再將被折扣的機率量，分配給未知事件。以雙連馬可夫模型而言，若將雙連詞串分為兩部分，一部分為詞首 w_{i-1}，另一部分為詞尾 w_i，凱氏平滑法就是

將整個雙連詞串(w_{i-1},w_i)都曾於訓練語料中出現的雙連詞串機率乘上折扣值，再將被折扣的機率量依照次低階的單連詞串機率，分配給(w_{i-1},w_i')沒有出現在訓練語料中，但詞首w_{i-1}與詞尾w_i'均曾出現於訓練語料中的詞雙連詞串。

在分配被折扣的機率量時，必需先統計所有在訓練語料中，沒有接在w_{i-1}之後的詞的次數和，再根據詞w_i' 的『次數』在這個次數和中所佔比例加以分配。公式如下(參照附錄)：

$$P_{katz}(w_i \mid w_{i-1}) = \begin{cases} \dfrac{C(w_{i-1}, w_i)}{C(w_{i-1})} & if\ C(w_{i-1}, w_i) > k \\ d_r \times \dfrac{C(w_{i-1}, w_i)}{C(w_{i-1})} & if\ 0 < C(w_{i-1}, w_i) \le k \\ \alpha(w_{i-1}) \times \dfrac{C(w_i)}{\sum_{w_j : C(w_{i-1}, w_j)=0} C(w_j)} & if\ C(w_{i-1}, w_i) = 0 \end{cases} \quad (2.11)$$

其中 $\alpha(w_{i-1}) = 1 - \sum_{w_j : C(w_{i-1}, w_j)>0} P_{katz}(w_j \mid w_{i-1})$，為所有被折扣的機率量

$0 < d_r \leqq 1, r = C(w_{i-1}, w_i)$

當$C(w_{i-1}, w_i)$>k時，表示(w_{i-1}, w_i)的次數夠大，使得其機率值非常可靠，因此將d_r設為 1，也就是不做折扣，通常取k=5；而對於k<5 的d_r值，則是給予其一個小於 1 的值，此值決定於古德圖靈估計(Good-Turing estimation)【6】【9】【13】。

在古德圖靈估計中，假設共有 s 個不同的 N 連詞串 $\alpha_1,\ldots,\alpha_s$，出現的機率分別為 $p_1,\ldots,p_s$，現在，想要估計 N 連詞串 α_i 在某個語料中出現 r 次的機率，就是希望得到 $E(p_i \mid C(\alpha_i) = r)$，其中 E 代表期望值，則

$$E(p_i \mid C(\alpha_i) = r) = \sum_{j=1}^{s} P(i = j \mid C(\alpha_i) = r) p_j \quad (2.12)$$

其中 $P(i = j \mid C(\alpha_i) = r)$ 表示一個未知的 N 連詞串 α_i 出現 r 次，且確實為第 j 個 N 連詞串 α_j 的機率

$$P(i=j|C(\alpha_i)=r)=\frac{P(C(\alpha_j)=r)}{\sum_{j=1}^{s}P(C(\alpha_j)=r)}$$
$$=\frac{\binom{M}{r}p_j^r(1-p_j)^{M-r}}{\sum_{j=1}^{s}\binom{M}{r}p_j^r(1-p_j)^{M-r}} \quad (2.13)$$
$$=\frac{p_j^r(1-p_j)^{M-r}}{\sum_{j=1}^{s}p_j^r(1-p_j)^{M-r}}$$

其中 $M=\sum_{j=1}^{s}C(\alpha_j)$，為所有次數的總和

將(2.13)式代入(2.12)式 ，可得

$$E(p_i \mid C(\alpha_i)=r)=\frac{\sum_{j=1}^{s}p_j^{r+1}(1-p_j)^{M-r}}{\sum_{j=1}^{s}p_j^r(1-p_j)^{M-r}} \quad (2.14)$$

現在，我們考慮 $E_M(n_r)$，n_r 為出現 r 次的 N 連詞串的個數，$E_M(n_r)$ 為在 M 個字中，N 連詞串出現 r 次的期望值

$$E_M(n_r)=\sum_{j=1}^{s}P(C(\alpha_j)=r)=\sum_{j=1}^{s}\binom{M}{r}p_j^r(1-p_j)^{M-r} \quad (2.15)$$

將(2.15)式代入(2.14)式

$$\text{得到 } E(p_i \mid C(\alpha_i)=r)=\frac{\frac{E_{M+1}(n_{r+1})}{\binom{M+1}{r+1}}}{\frac{E_M(n_r)}{\binom{M}{r}}}=\frac{r+1}{M+1}\frac{E_{M+1}(n_{r+1})}{E_M(n_r)} \quad (2.16)$$

這也就是一開始想求的N連詞串 α_i 出現r次的期望機率，則在M個詞的語料中，N連詞串 α_i 的期望次數r^*為

$$r^*=M\times P(\alpha_i)=M\times\frac{r+1}{M+1}\times\frac{E(n_{r+1})}{E(n_r)}\approx(r+1)\frac{n_{r+1}}{n_r} \quad (2.17)$$

此處的M為一個極大值，且以n_r的經驗值來估計n_r與n_{r+1}的期望值

而凱氏平滑法決定d_r值的做法就是將出現r次的N連詞串假設為出現r^*次，$r^*<r$

$$r^* = (r+1)\frac{n_{r+1}}{n_r} \tag{2.18}$$

此外，d_r必需符合以下兩點限制：

1. d_r需與古德圖靈估計的r^*值與原r值的比例成正比

   即對於$r \in \{1,\ldots,k\}$，$d_r = \mu\frac{r^*}{r}$，μ 為一常數值 (2.19)

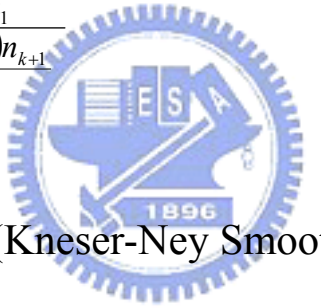
2. 古德圖靈估計分配給未知事件的次數，需與被折扣的次數和相等

   因為所有被分配給未知事件的次數為$n_0 0^* = n_0\frac{n_1}{n_0} = n_1$，所以

$$\sum_{r=1}^{k} n_r(1-d_r)r = n_1 \tag{2.20}$$

根據(2.19)式及(2.20)式 ，所解出的d_r值為

$$d_r = \frac{\frac{r^*}{r} - \frac{(k+1)n_{k+1}}{n_1}}{1-\frac{(k+1)n_{k+1}}{n_1}} \tag{2.21}$$



2.4.3 聶氏平滑法(Kneser-Ney Smoothing)

聶氏平滑法是根據絕對折扣法(Absolute Discounting Smoothing)所發展出來的，絕對折扣法的重點在於，每次由已知事件所扣除的折扣為一個固定的常數 D，D≦1，同時，在【13】中也建議

$$D = \frac{n_1}{n_1 + 2n_2} \tag{2.22}$$

其中n_1與n_2分別代表出現 1 次與 2 次的N連詞串數目。

當我們考慮以下情形，『菠』在日常生活中而言，算是常用字，但是『菠』的後面，只可以接『菜』這個字。若我們單純以字的出現次數來分配機率，可能『菠』後面接任意字詞的機率會很大，但這是錯誤的，因為在訓練語料中，『菠』的後面只接了一種字-『菜』，所以『菠』後面接其他字詞的機率值應該非常小，而聶氏平

滑法就是以此為原則發展出來的方法。

聶氏平滑法的分配原則與凱氏平滑法有些類似，都是將已知事件的機率值做折扣，再分配給所有詞串，或分配給未知事件。

對於雙連馬可夫模型的聶氏平滑法，我們定義以下幾個名詞：

1. 詞對(w_{i-1},w_i)中的w_i為『詞尾』
2. 詞對(w_{i-1},w_i)中的w_{i-1}為『詞首』
3. 在訓練語料中，詞尾之前所接的不同詞數為『前接詞數』
4. 在訓練語料中，詞首之後所接的不同詞數為『後接詞數』

聶氏平滑法又分為兩種：

1. 聶氏內插法(Kneser-Ney Interpolation Smoothing)，是聶氏於 1991 年提出的方法【11】，以雙連馬可夫模型為例，公式如下(參照附錄)：

$$P_{KN}(w_i \mid w_{i-1}) = \frac{\max\{C(w_{i-1},w_i)-D,0\}}{C(w_{i-1})} + \alpha(w_{i-1}) \times \frac{1}{V} \qquad (2.23)$$

其中 V 代表辭典中所有詞彙的個數

$\alpha(w_{i-1}) = 1 - \sum_{w_j : C(w_{i-1},w_j)>0} P_{KN}(w_j \mid w_{i-1}) = \frac{D \times N_{1+}(w_{i-1},\bullet)}{C(w_{i-1})}$ ，為所有被折扣的機率量

$$N_{1+}(w_{i-1},\bullet) = \left|\{w_k : C(w_{i-1},w_k) > 0\}\right|$$

(2.23)式主要的意義則是將被折扣的機率量$\alpha(w_{i-1})$與所有以w_{i-1}為詞首的詞串做內插，也就是將被折扣的機率量$\alpha(w_{i-1})$分配給所有以w_{i-1}為詞首的詞串，而非僅分配給以w_{i-1}為詞首但(w_{i-1}, w_i)未出現於訓練語料者；分配的方法是平均分配給所有可能的w_i，又因為所有於辭典中的詞都可能是w_i，所以是平均分配給辭典中的V個詞。

2. 聶氏後退法(Kneser-Ney Backoff Smoothing)【12】，是聶氏於 1995 年提出的方法，以雙連馬可夫模型為例，公式如下(參照附錄)：

$$P_{KN}(w_i \mid w_{i-1}) = \begin{cases} \dfrac{\max\{C(w_{i-1}, w_i) - D, 0\}}{C(w_{i-1})} & if\ C(w_{i-1}, w_i) > 0 \\ \alpha(w_{i-1}) \times \dfrac{N_{1+}(\bullet, w_i)}{\sum_{w_k : C(w_{i-1}, w_k) = 0} N_{1+}(\bullet, w_i)} & if\ C(w_{i-1}, w_i) = 0 \end{cases} \quad (2.24)$$

其中 $N_{1+}(\bullet, w_i) = \left|\{w_k : C(w_k, w_i) > 0\}\right|$ ，也就是詞尾w_i的前接詞數

$\alpha(w_{i-1}) = 1 - \sum_{w_j : C(w_{i-1}, w_j) > 0} P_{KN}(w_j | w_{i-1}) = \dfrac{D \times N_{1+}(w_{i-1}, \bullet)}{C(w_{i-1})}$ ，為所有被折扣的機率量

$N_{1+}(w_{i-1}, \bullet) = \left|\{w_j : C(w_{i-1}, w_j) > 0\}\right|$ ，也就是詞首w_{i-1}的後接詞數

(2.24)式主要的意義是將所有曾出現於訓練語料的雙連詞串(w_{i-1}, w_i)的機率值做折扣，再將被折扣後的機率量$\alpha(w_{i-1})$分配給(w_{i-1}, w_i')未成對出現於訓練語料，但w_{i-1}與w_i'均曾出現於訓練語料的雙連詞串。根據(2.24)式，被折扣的機率量$\alpha(w_{i-1})$會與詞首w_{i-1}的後接詞數$N_{1+}(w_{i-1}, \bullet)$成正比，而未知事件的分配原則是先計算所有沒接在w_{i-1}之後的詞之『前接詞數和』，再根據詞尾w_i'的前接詞數在此前接詞數總和中所佔的比例加以分配。

由此可知，聶氏後退法與凱氏平滑法最大的不同就是聶氏後退法是以『前接詞數』來分配機率，而凱氏平滑法是以該詞出現的『次數』來分配機率。

第 3 章

本論文提出的方法

3.1 語言模型的原則

在建構語言模型之前，有三個大原則需考慮：

1. 統計取向或文法取向？

統計取向的模型較能容忍前端語音辨識系統上的錯誤，因此本論文採用統計取向的語言模型，

2. 以『字』為單位或以『詞』為單位？

由中文語言的特性來看，詞包含了人類的知識，且能使得估測的結果較具強健性，還可有效地減少參數量，故採用以『詞』為單位。

3. 採用何種機率估計方法？

因為 N 連馬可夫模型具有容易建立模型、簡單、有效…等特性，是統計取向語言模型的主流。又在中文語言中，詞成對出現的機率較大，故採用雙連的馬可夫模型，如此，亦較符合記憶體資源的使用量。

3.2 強化凱氏平滑法與強化聶氏平滑法

當遇到恰好沒有在訓練語料中相鄰出現的詞對w_{i-1}與w_i時，將使得$P(w_i|w_{i-1})=0$，這會造成整句話機率為零的錯誤。若我們採取先估計單連字串機率，再往上推測未知雙連字串的機率值時，雖然可以推測出未雙連字串的機率值，但對於未出現於訓練語料中的詞而言，我們根本無法由訓練語料中得到該詞的相關資訊，也無法給它一個適當的機率估計方法。因此，在本論文中，我們的做法是在估計雙連詞串時，再對於未出現於訓練語料的詞給予其適當的機率估計方法，是分別針對凱氏平滑法及聶氏平滑法加以強化的方法，稱為『強化凱氏平滑法』及『強化聶氏平滑法』，並於第四章的實驗中，與原有的方法比較，以證明此兩種方法確實可提高模型的效能。

3.2.1 強化凱氏平滑法

『凱氏平滑法』的主要做法是將以w_{i-1}為詞首的雙連詞對中，在訓練語料裡出現次數小於k者(通常k=5)乘上一個折扣值d_r；至於在訓練語料中出現的次數大於k者，因其次數夠大，具有夠強的代表性，故不予以折扣。而在分配被折扣的機率量時，需先找出所有未與w_{i-1}相鄰出現在訓練語料中的w_i'，並計算這些w_i'出現在訓練語料中的次數總和，再根據w_i於此總和值中所佔的次數比例，來決定可以由被折扣的機率量中得到多少比例。但若詞w_i在訓練語料中出現的次數為零，則其由被折扣的機率量中，分配到的機率也會是零，將會造成錯誤，這是『凱氏平滑法』的一項缺點。

基於上述理由，我們參考原有的『凱氏平滑法』，提出『強化凱氏平滑法』，其應用於雙連馬可夫模型的公式如下(參照附錄)：

$$P_{katz}(w_i \mid w_{i-1}) = \begin{cases} \dfrac{C(w_{i-1}, w_i)}{C(w_{i-1})} & if\ C(w_{i-1}, w_i) > k \\ d_r \times \dfrac{C(w_{i-1}, w_i)}{C(w_{i-1})} & if\ 0 < C(w_{i-1}, w_i) \le k \\ \alpha(w_{i-1}) \times \dfrac{C(w_i)}{\sum_{w_j : C(w_{i-1}, w_j)=0} C(w_j)} & if\ C(w_{i-1}, w_i) = 0\ and\ C(w_i) > k \\ d'_r \times \alpha(w_{i-1}) \times \dfrac{C(w_i)}{\sum_{w_j : C(w_{i-1}, w_j)=0} C(w_j)} & if\ C(w_{i-1}, w_i) = 0\ and\ 0 < C(w_i) \le k \\ \beta(w_{i-1}) \times \dfrac{1}{T} & if\ C(w_{i-1}, w_i) = 0\ and\ C(w_i) = 0 \end{cases}$$

(3.1)

其中 $\alpha(w_{i-1}) = 1 - \sum_{w_j : C(w_{i-1}, w_j) > 0} P_{katz}(w_j \mid w_{i-1})$，為由出現於訓練語料的次數小於 k 的雙連詞對所折扣下來的機率量

$\beta(w_{i-1}) = \alpha(w_{i-1}) - \sum_{w_j : C(w_{i-1}, w_j)=0\ and\ C(w_j)>0} P_{katz}(w_j \mid w_{i-1})$，為當未與$w_{i-1}$相鄰出現在訓練語料中的$w_i$，其出現於訓練語料的次數小於$k$時，所折扣下來的機率量

T 為辭典中未出現於訓練語料的詞的個數

$$d_r = \frac{\dfrac{r^*}{r} - \dfrac{(k+1)n_{k+1}}{n_1}}{1 - \dfrac{(k+1)n_{k+1}}{n_1}},\quad r^* = (r+1)\frac{n_{r+1}}{n_r}$$

n_r為出現r次的雙連詞串數目

$$d'_r = \frac{\dfrac{r^*}{r} - \dfrac{(k+1)n'_{k+1}}{n'_1}}{1 - \dfrac{(k+1)n'_{k+1}}{n'_1}},\quad r^* = (r+1)\frac{n'_{r+1}}{n'_r}$$

n'_r 為出現 r 次的單連詞串數目

k=5

在我們的方法中，主要的做法有以下五點：

1. 當雙連詞串(w_{i-1},w_i)在訓練語料中有出現時，還是依原有的最大概似法計算其機率
2. 當雙連詞串(w_{i-1},w_i)沒有在訓練語料出現過時，先判斷詞尾w_i在訓練語料中是否有出現
3. 若有出現，且次數大於 k 次，則依照原凱氏平滑法做分配
4. 若有出現，且次數小於等於 k 次，則除了依原凱氏平滑法做分配外，還必需乘上一個 d_r' 值
5. 若詞尾w_i在訓練語料中根本沒有出現，則可將於第 4 點中被折扣下來的機率量 $\beta(w_{i-1})$，平均分配給所有未出現於訓練語料中的詞

如此，也考慮了雙連詞串(w_{i-1},w_i)的詞尾w_i於訓練語料中沒有出現的狀況，將使得辨識的估計值更為準確。

3.2.2 強化聶氏平滑法

『聶氏後退法』在分配機率時，先找出所有未與w_{i-1}相鄰出現於訓練語料的w_i'，再計算所有w_i'於訓練語料中前接詞數總和，以詞尾 w_i 的前接詞數於此前接詞數總和中所佔的比例來分配被折扣下來的機率。如果將這項原則應用到『聶氏內插法』時，可以使得具較多前接詞數 $N_{1+}(\bullet,w_i)$ 的詞尾 w_i ，分配到較大的機率值，而非平均分配給每個詞尾 w_i ，如此便可做更正確的預測，因此，我們將原有的『聶氏後退法』及『聶氏內插法』公式合併。但，這也會遇到與『凱氏平滑法』相同的問題，就是如果詞尾 w_i 並沒有出現在訓練語料中時，一樣會造成機率分配為零的錯誤，所以我們再針對這個缺點加以修正，稱為『強化聶氏平滑法』，公式如下(參照附錄)：

$$\begin{cases} P_{KN}(w_i \mid w_{i-1}) = \dfrac{\max\{C(w_{i-1}, w_i) - D, 0\}}{C(w_{i-1})} + \alpha(w_{i-1}) \times P'_{KN}(w_i) \\ P'_{KN}(w_i) = \dfrac{\max\{N_{1+}(\bullet, w_i) - D, 0\}}{\sum_{w_j} N_{1+}(\bullet, w_j)} + \beta \times \dfrac{1}{V} \end{cases} \tag{3.2}$$

其中 $\alpha(w_{i-1}) = 1 - \sum_{w_j : C(w_{i-1}, w_j) > 0} P_{KN}(w_j \mid w_{i-1}) = \dfrac{D \times N_{1+}(w_{i-1}, \bullet)}{C(w_{i-1})}$，是所有曾出現於訓練語料中，以 w_{i-1} 為詞首的雙連詞串被折扣的機率量總和

$P'_{KN}(w_i)$ 是分配 $\alpha(w_{i-1})$ 時的原則，當詞尾w_i的前接詞數越多，表示可接在w_i之前的詞種類很多，$P'_{KN}(w_i)$ 越大，因此由 $\alpha(w_{i-1})$ 分到的機率就越大；反之，當w_i的前接詞數越少，$P'_{KN}(w_i)$ 越小，因此由 $\alpha(w_{i-1})$ 分到的機率就越小

$\beta = \dfrac{D \times Q}{\sum_{w_j} N_{1+}(\bullet, w_j)}$，為計算 $P'_{KN}(w_i)$ 時，先將詞尾w_i的前接詞數做折扣後，再除上所有可能詞尾的前接詞數總和，進一步折扣下來的少量機率量

Q 為辭典中曾經出現於訓練語料的詞的個數

$N_{1+}(w_{i-1}, \bullet) = \left|\{w_j : C(w_{i-1}, w_j) > 0\}\right|$

$N_{1+}(\bullet, w_i) = \left|\{w_j : C(w_j, w_i) > 0\}\right|$

V 為辭典中所有詞彙的個數

$0 < D \leqq 1$

在上述的方法中，主要做法為：

1. 將所有以w_{i-1}為詞首的已知雙連詞串機率減去一個折扣值，所得的被折扣機率量總和為 $\alpha(w_{i-1})$
2. 將被折扣的機率量 $\alpha(w_{i-1})$ 分配給以w_{i-1}為詞首的所有已知與未知雙連詞串
3. 以w_{i-1}為開頭的雙連詞串可分成詞尾w_i於訓練語料有出現，與沒有出現這兩

種情況

4. 當詞尾的w_i於訓練語料有出現時，將第 1 點所述的被折扣的機率值$\alpha(w_{i-1})$，依照w_i的前接詞數於所有可能詞尾的前接詞數總和中所佔的比例$P'_{KN}(w_i)$加以分配機率，此處的w_i的前接詞數於計算時還會減去一個折扣值，得到被折扣下來的少量機率量為β。$P'_{KN}(w_i)$比例取決於w_i的前接詞數，當詞尾w_i的前接詞數越多，表示可接在w_i之前的詞種類很多，由$\alpha(w_{i-1})$分配到的機率就越大；反之，當w_i的前接詞數越少，由$\alpha(w_{i-1})$分配到的機率就越小

5. 將第 4 點中，被折扣的少量機率量β平均分配給所有可能接在w_{i-1}之後的詞，即辭典中的每一個詞。

第 4 章

實驗結果

4.1 實驗環境及資料來源

首先，本論文的實驗資料是取自中華電視公司全球新聞網站的新聞頻道【1】，資料的收集時間為 2004 年 3 月起，至 2005 年 2 月止每日的新聞內容，共 23,615 則新聞，5,206,223 個字，3,157,166 個詞，3,154,791 個四字以下的詞，詳細的統計如下表：

表 4-1 ：2004 年 3 月到 2004 年 8 月的語料統計

年/月	04/03	04/04	04/05	04/06	04/07	04/08
新聞則數	2,115	2,062	2,156	2,190	2,209	2,058
字數	468,103	449,496	477,708	469,219	488,523	445,577
總詞數	286,476	274,950	292,889	287,518	298,863	272,426
四字以下詞數	286,255	274,779	292,709	287,314	298,651	272,286

表 4-2 ：2004 年 9 月到 2005 年 2 月的語料統計

年/月	04/09	04/10	04/11	04/12	05/01	05/02
新聞則數	1,913	2,050	1,889	1,824	1,657	1,492
總字數	426,847	437,113	424,110	402,334	366,258	350,935
詞數	258,566	261,104	251,657	241,299	219,895	211,523
四字以下詞數	258,349	260,893	251,389	241,074	219,752	211,340

4.1.1 測試語料

在每個月的語料中，各隨機取出 180 則新聞來做為測試語料，統計如下：

表 4-3 ：2004 年 3 月到 2004 年 8 月的測試語料統計

年/月	04/03	04/04	04/05	04/06	04/07	04/08
字數	40,700	37,205	38,125	34,801	41,627	35,611
總詞數	25,235	22,836	23,410	21,309	25,718	21,969
四字以下詞數	25,213	22,819	23,392	21,289	25,697	21,955

表 4-4 ：2004 年 9 月到 2005 年 2 月的測試語料統計

年/月	04/09	04/10	04/11	04/12	05/01	05/02
字數	40,218	39,111	43,354	39,343	41,417	38,613
總詞數	24,783	23,874	26,199	23,828	25,231	23,265
四字以下詞數	24,761	23,853	26,172	23,805	25,217	23,247

4.1.2 訓練語料

即收集到的每月新聞中，不包含測試語料的新聞資料，共 21,455 則新聞，4,736,098 個字，2,869,509 個詞，2,867,371 個四字以下的詞，詳細的統計資料如下：

表 4-5 ：2004 年 3 月到 2004 年 8 月的訓練語料統計

年/月	04/03	04/04	04/05	04/06	04/07	04/08
字數	427,403	412,291	439,583	434,418	446,896	409,966
總詞數	261,241	252,114	269,479	266,209	273,145	250,457
四字以下詞數	261,042	251,960	269,317	266,025	272,954	250,331

表 4-6 ：2004 年 9 月到 2005 年 2 月的訓練語料統計

年/月	04/09	04/10	04/11	04/12	05/01	05/02
字數	386,629	398,002	380,756	362,991	324,841	312,322
總詞數	233,783	237,230	225,458	217,471	194,664	188,258
四字以下詞數	233,588	237,040	225,217	217,269	194,535	188,093

4.1.3 辭典

由中研院 CKIP(Chinese Knowledge and Information Processing)詞庫小組提供【2】，包含一般用詞、常用專有名詞、成語、慣用語…等等，並且只考慮單字詞到四字詞，辭典的大小為 93755 個詞，單字詞到四字詞的數目分別如下：

表 4-7 ：辭典的單字詞到四字詞數目統計

	單字詞	雙字詞	三字詞	四字詞
數目	7,672	56,562	17,419	12,102

4.1.4 斷詞程式

因為我們的語言模型是以『詞』為辨認單位，故實驗的語料均必須先以斷詞程式處理，採用的是中央研究院詞庫小組所提供的 Autotag 中文自動斷詞系統 1.0 版【3】。

4.2 實驗的評估方法

在建立語言模型後，最直接的做法是使用語音資料的辨識錯誤率來評量語言模型的優劣，但使用錯誤率做為評估語言模型的標準，會和語音辨識系統的好壞有極大的相關性，因此錯誤率不見得可以真正反應出語言模型的優劣。基於這個理由，混淆度(perplexity)已成為評估語言模型的一項最重要且最通用的量測標準【6】，此標準是依據資訊理論(information theory)中熵(entropy)的概念所發展出來的，它可以在不牽涉到語音辨識系統的情況下，測量語言模型的好壞。

根據資訊理論，對於一個語言模型及測試語料W而言，假設測試語料的詞共有N_W個，其語言模型預測詞出現的平均困難度，也就是詞的估計熵(estimated entropy) H_W定義為

$$H_W = -\frac{1}{N_W}\log_2 P(W) \tag{4.1}$$

由壓縮演算法的觀點而言，就代表說測試語料W可以用$-\log_2 P(W)$個位元來加以編碼。當H_W越小時，代表可以用越少的位元對測試語料加以編碼，因此，H_W值越

小越好。

若將測試語料接成一長串，即$W=(w_1,w_2,w_3,\ldots,w_{Nw-1},w_{Nw})$，$w_i$為一個詞，則對於雙連馬可夫模型而言，$H_W$可改寫成

$$\begin{aligned} H_W &= -\frac{1}{N_W}\log_2 P(w_1,w_2,\cdots,w_{Nw}) \\ &= -\frac{1}{N_W}\log_2 P(w_1,w_2)P(w_2,w_3)\cdots P(w_{Nw-1},w_{Nw}) \\ &= -\frac{1}{N_W}\sum_{i=1}^{Nw-1}\log_2 P(w_i,w_{i+1}) \end{aligned} \tag{4.2}$$

進一步定義詞的混淆度PP_W為 $PP_W = 2^{Hw}$ (4.3)

上述的介紹中，主要是針對西方語言的概念，但對於中文的『詞』而言，並沒有固定的長度單位，詞與詞之間也沒有明顯的邊界，因此有必要對於上述的定義加以修正如下。

假設測試語料的字共有N_C個，則字的估計熵H_C定義為

$$H_C = -\frac{1}{N_C}\log_2 P(W) \tag{4.4}$$

字的混淆度PP_C為 $PP_C = 2^{Hc}$ (4.5)

在本論文的實驗中，均以字的混淆度PP_C來做為語言模型的評估單位。

4.3 實驗的流程

本論文的實驗流程分兩部分：建立語言模型的部分及計算測試語料混淆度的部分

1. 建立語言模型部分：

因為本實驗所取得的語料為一篇篇的新聞文稿，所以必需先將文稿加以斷詞之後，依據辭典將每個斷出來的詞配上注音(此處以半自動方式加上注音，因為有些破音字詞的決定還是需要人工介入)，再以詞為單位，計算每個成對的詞與詞相鄰出現的機率(同字不同音的詞視為不同詞)，以建立雙連的馬可夫語言模型。

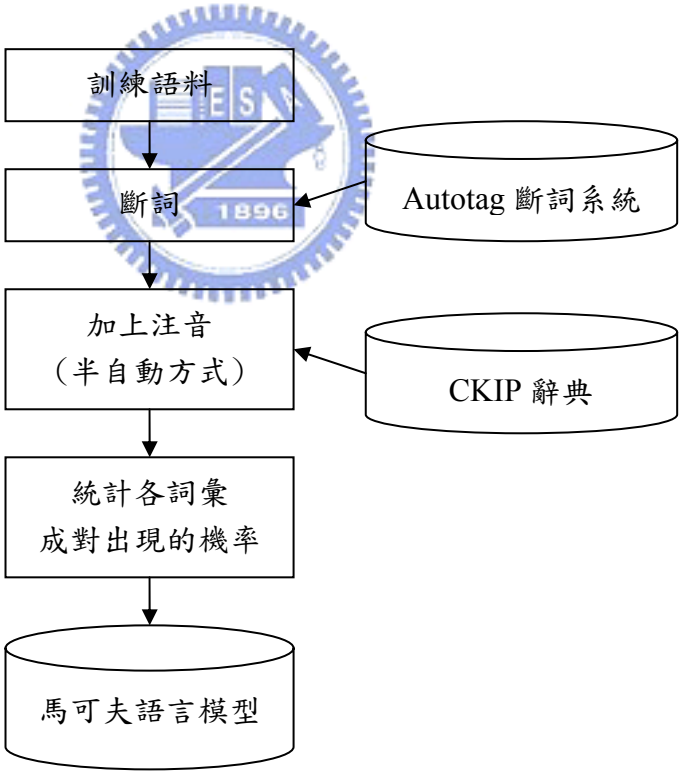


圖 4-1：建立語言模型的流程

2. 計算測試語料混淆度部分：

同樣地，我們必需先將要當做測試語料的新聞文稿加以斷詞，之後利用辭典以半自動的方式為文稿加上注音，再依據我們建立的馬可夫模型，計算測試語料的混淆度。

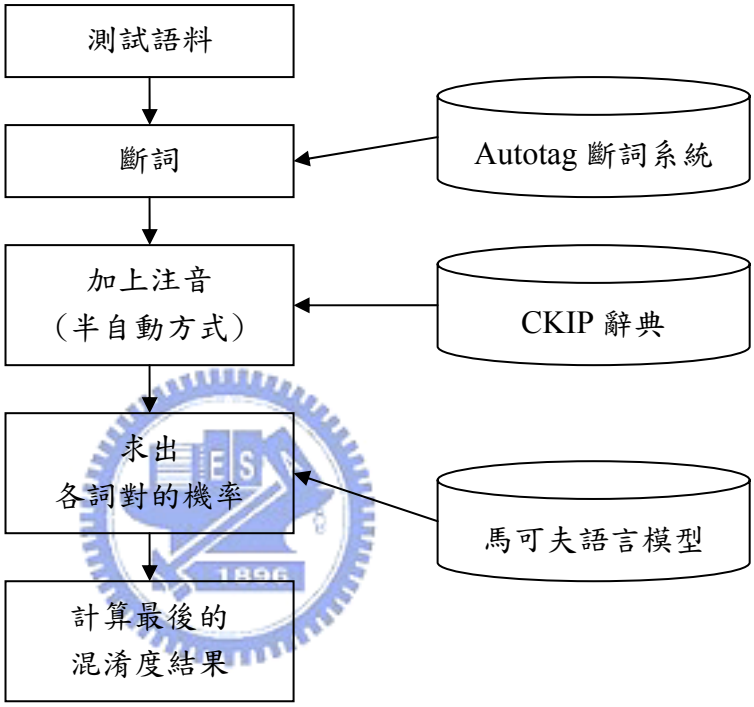


圖 4-2：計算測試語料混淆度的流程

4.4 實驗數據與結果

在本節中將列出我們的語言模型應用『凱氏平滑法』及『聶氏平滑法』時所需的一些數據，並實作各種平滑化方法，以驗證我們提出的強化式平滑法的確能增加模型的效能。

4.4.1『凱氏平滑法』及『強化凱氏平滑法』中的d_r值

在凱氏平滑法及強化凱氏平滑法中，當出現次數小於k時，必需乘上一個折扣值d_r，而折扣值d_r取決於出現r次的N連詞串。因為取k=5，所以必需統計出現 1 到 6 次的雙連詞串數目，以計算d_1到d_5及 1*到 5*的值。表 4-8 是於訓練語料中，出現 1 到 5 次的雙連詞串資料。

表 4-8 ：出現 6 次以下的雙連詞串數目，及其r*與d_r值。

出現次數 (r 值)	1	2	3	4	5	6
雙連詞串數目 (n_r值)	11,780	5,654	3,701	2,653	2,032	1,676
$r^* = (r+1)\frac{n_{r+1}}{n_r}$	0.96	1.96	2.87	3.83	4.95	
$d_r = \frac{\frac{r^*}{r} - \frac{(k+1)n_{k+1}}{n_1}}{1 - \frac{(k+1)n_{k+1}}{n_1}}$	0.73	0.87	0.71	0.72	0.93	

4.4.2 聶氏平滑法及強化聶氏平滑法中的前後接詞數

在聶氏平滑法與強化聶氏平滑法中，絕對折扣值的大小會與詞首的後接詞數有關，而未知事件在分配機率值時，則是與詞尾的前接詞數有關，因此，我們必需先統計所有詞的前接與後接詞數，才能將其代入聶氏平滑法中，求得平滑後的機率值。在訓練語料中，以『的』的前後接詞數最多，有 19527 種前接詞數、18606 種後接詞數。

表 4-9 ：『詞尾』的前接詞數分佈表

前接詞數	1-4	5-9	10-99	100-999	1000-5000	5000 以上
詞數	27,416	8,219	12,414	1,278	46	1

表 4-10：『詞首』的後接詞數分佈表

後接詞數	1-4	5-9	10-99	100-999	1000-5000	5000 以上
詞數	26,601	6,884	10,150	1,246	68	3

4.4.3 各種平滑化方法的比較

本論文實做了語言模型中使用『加成平滑法』、『凱氏平滑法』、『聶氏後退法』及『聶氏內插法』各種平滑化方法的情況，同時也實做了我們所提出的『強化凱氏平滑法』及『強化聶氏平滑法』，以下表格為各平滑化方法的混淆度。

表 4-11：各平滑化方法的混淆度之比較

方法 年/月	加成 平滑法	凱氏 平滑法	強化凱氏 平滑法	聶氏 內插法	聶氏 後退法	強化聶氏 平滑法
04/03	623.61	67.84	60.38	44.65	38.01	34.48
04/04	541.01	56.67	52.03	39.75	34.18	31.17
04/05	595.20	76.42	69.44	48.60	41.88	37.60
04/06	549.16	71.47	65.16	46.07	39.95	35.68
04/07	582.82	69.52	61.88	45.21	39.25	35.17
04/08	529.22	70.67	65.60	45.70	38.68	35.08
04/09	570.86	79.77	72.17	49.09	43.19	38.20
04/10	621.48	89.35	84.59	53.82	44.82	40.65
04/11	613.24	85.60	79.54	52.67	44.70	40.25
04/12	546.60	71.09	64.66	46.28	41.67	36.50
05/01	555.69	77.84	73.60	49.97	46.83	40.13
05/02	546.92	74.95	66.33	46.13	42.69	37.00

4.5 結果討論與分析

由表 4-11，我們將除了『加成平滑法』以外的五種方法繪製成比較圖 4-3 (因『加成平滑法』為本實驗的對照組，其實驗數據明顯與其他平滑化方法相差許多)，並觀察到以下結論：

1. 不論是使用『凱氏平滑法』或『聶氏平滑法』，所得的效果均較最基本的『加成平滑法』好上許多；而使用『聶氏平滑法』的混淆度會比使用『凱氏平滑法』來得小，這是因為『聶氏平滑法』所使用的鄰接詞資訊較為精確的緣故。

2. 單就『聶氏平滑法』而言，使用『聶氏後退法』的混淆度會比使用『聶氏內插法』來得小，這是因為『聶氏後退法』在分配時是依照詞尾的前接詞數來分配，比『聶氏內插法』平均給每個詞還要來得精確。

3. 我們所提的『強化凱氏平滑法』，較原『凱氏平滑法』平均小了 6.65 個混淆度單位；而『強化聶氏平滑法』，較原聶氏平滑法中效能較佳的『聶氏後退法』還平均小了 4.50 個混淆度單位，顯示我們所提的方法的確可增加語言模型的效能。

4. 『強化凱氏平滑法』與原『凱氏平滑法』比較時所增加的效能(6.65)，比『強化聶氏平滑法』與『聶氏後退法』比較時所增加的效能(4.50)還要多，這是因為『強化凱氏平滑法』是只將折扣後所剩的值分給機率為零的未知事件，而『強化聶氏平滑法』是分配給所有在辭典中的詞彙的緣故。

5. 整體看來，在我們所實做的六種平滑化方法中，『強化聶氏平滑法』擁有最佳的效能。

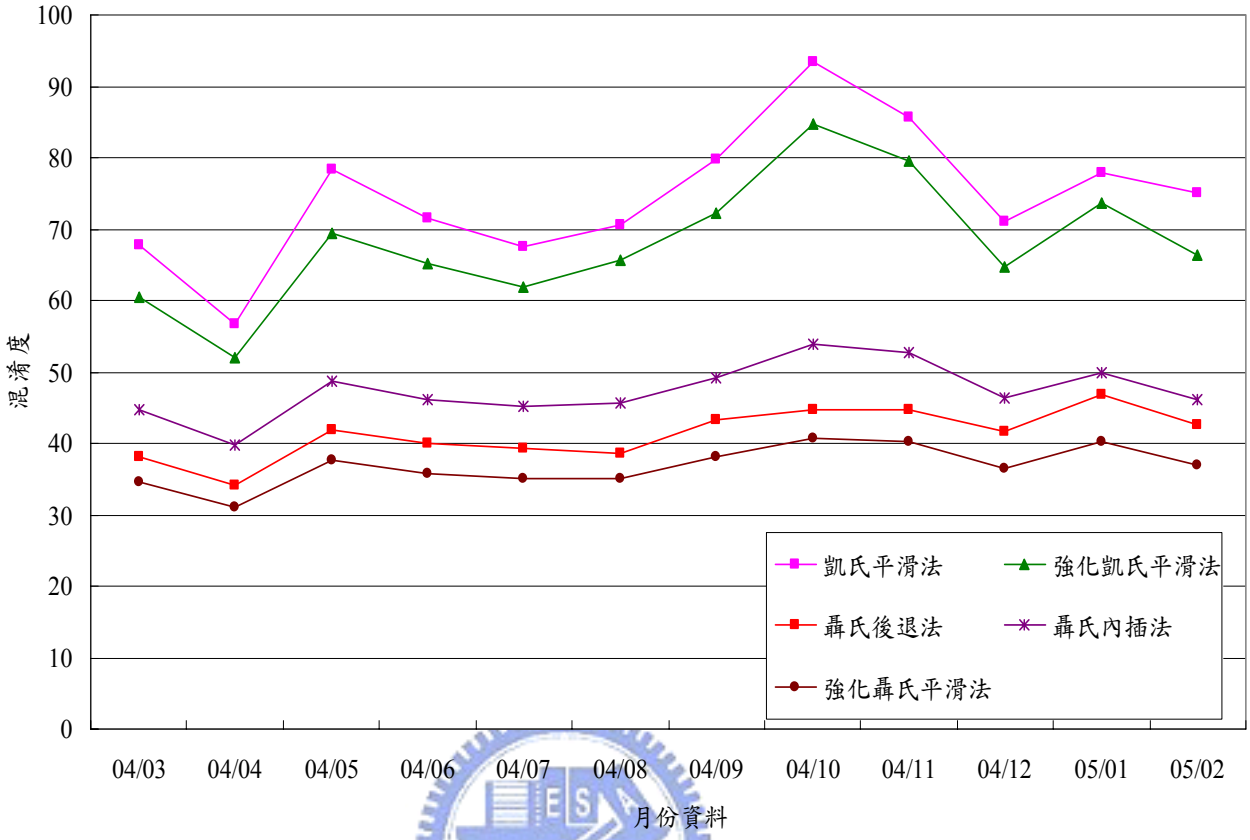


圖 4-3：除『加成平滑法』外的五種平滑化方法混淆度比較圖

第 5 章

系統應用：手持式語音辨識系統之實做

5.1 手持式語音辨識系統之建構

本論文將我們所提的平滑化方法及建構的雙連馬可夫語言模型，應用於我們所建構的手持式語音辨識系統中，我們的手持式語音辨識系統採用主從式的 Client-Server 架構，Client 端為手持式的個人數位助理設備(PDA, Personal Digital Assistant)，硬體設備為 HP iPAQ 5550，搭配 400M Hz 的中央處理器，與微軟視窗作業系統 Pocket PC 2003 Premium；Server 端為個人電腦，硬體設備為 Pentium 4 3.00G Hz 的中央處理器，1GB 的記憶體，搭配微軟視窗作業系統 XP 專業版。我們所建構的手持式語音辨識系統，能讓使用者由 PDA 輸入語音，並藉著無線傳輸，將資料傳送給 Server 端，透過個人電腦的輔助運算，再回傳對應的候選詞串給使用者。

圖 5-1 為我們實做的手持式語音辨識系統架構圖。Client 端（即行動裝置端）的使用者透過行動裝置端介面(User Interface)錄進語音訊號(WAV Recorder)之後，會做去除雜訊(Sound Enhance)與抽取聲音訊號特徵值(MFCC Converter)的動作【17】，之後將聲音訊號特徵值藉由無線傳輸至 Server 端的語音辨識器(HTK

Recognizer【5】)，再由語音辨識器的聲學模型做語音辨識【7】，而辨識出來的音節會傳送給語言模型(Language Model)，以判斷可能的詞串，語言模型根據辭典(Lexicon)及訓練語料庫(Corpus Database)判斷出最可能的前 N 個詞串，並回傳給 Client 端的使用者(User Interface)，如此便完成了語音辨識的工作。

再者，使用者在選擇正確詞串回送給 Server 端之後， Server 端還可依據使用者選擇的正確詞串音節，針對該使用者做語音模型的調適(Speaker Adaptation)，也就是將語音辨識器(HTK Recognizer)的模型參數調整為適合該使用者的參數，讓辨識的語音模型更適合該使用者【18】；若正確詞串均不包含在最可能的前 N 個詞串中，使用者也可藉著自行輸入正確的音，來做模型的調適。也正因為本系統具有針對不同的使用者分別調適其模型的功能，所以當使用者使用本系統的次數越多，辨識的效果也會隨之越來越好。

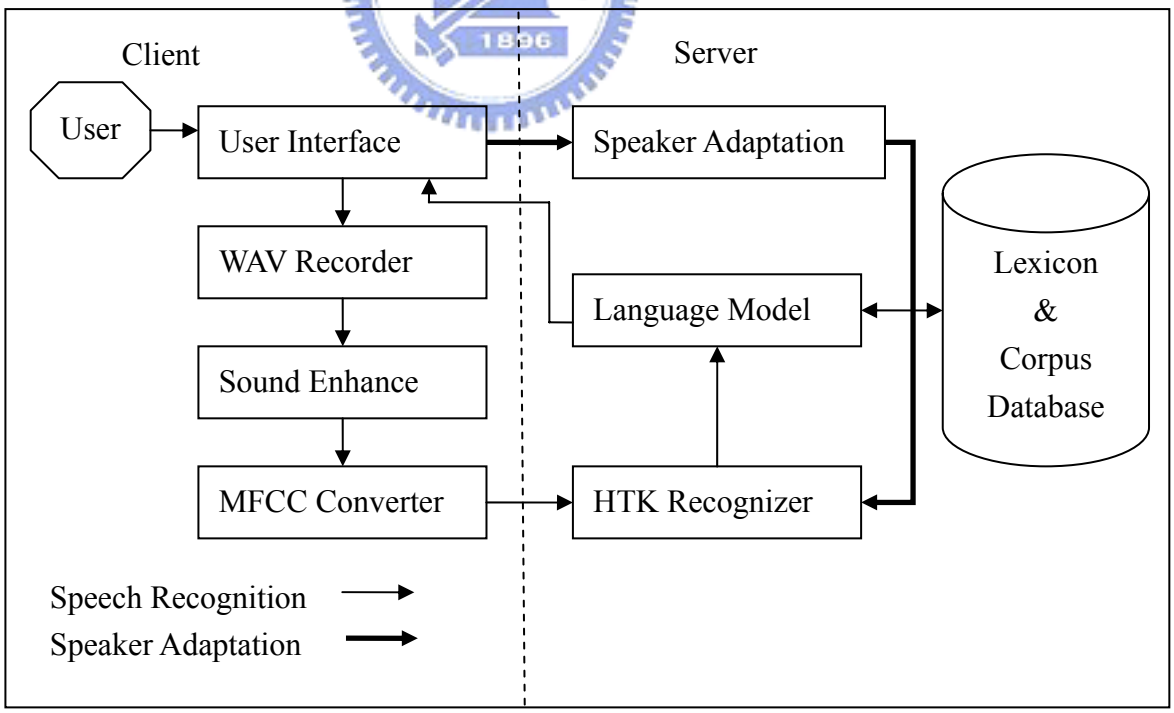


圖 5-1：語音辨識系統架構圖

5.2 語言模型於本手持式語音辨識系統之實做方法

我們的手持式語音辨識系統所採用的語音辨識方法分兩階段

第一階段：語音辨識器透過聲學模型參數，由使用者輸入的聲音訊號中辨識出可能的對應候選音節

第二階段：由候選音節中，透過語言模型，找出最可能的候選詞串做為輸出

而在將我們的語言模型實際應用到手持式語音辨識系統中時，需對候選音節先做構詞及搜尋的處理，才可將候選音節套用到我們的語言模型中。以下分別介紹我們在實做系統時所使用的構詞與搜尋的方法，及我們的手持式語音辨識系統中，語言模型的處理流程。

5.2.1 構詞

在本手持式語音辨識系統中，我們的做法是由聲學模型所辨認出的每個候選音節中，找出該音節所對應的『同音字』，與該音節及其前後音節相連時所形成的『同音詞』，以形成候選詞彙的集合，此步驟就稱為『構詞』。當這些候選詞彙相連後，形成如圖 5-2 的『格狀詞組』(word graph)【6】【14】【15】，而構詞後所形成的候選詞彙就是格狀詞組中的節點。我們想做的就是定義一個好的機率估計方式及一個有效的搜尋演算法，使得能由複雜的格狀詞組中，找出最佳的路徑當做輸出，以辨認出最可能的句子。

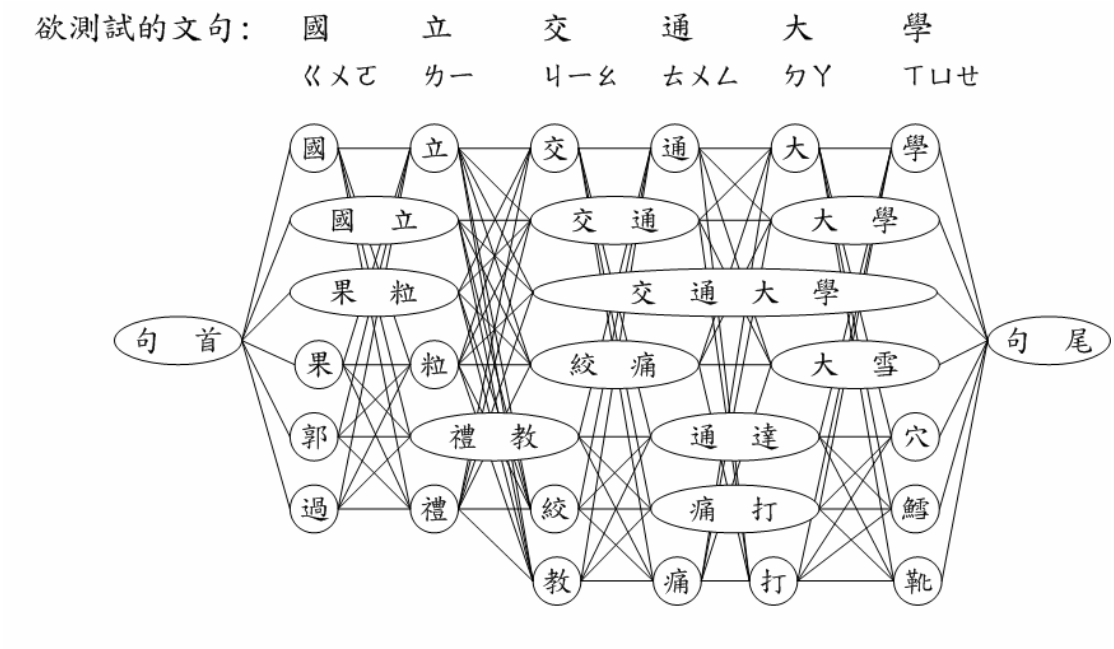


圖 5-2：格狀詞組示意圖

5.2.2 格狀詞組的搜尋

在本手持式語音辨識系統的語言模型中，因為使用雙連馬可夫模型，即下一個詞的預測只與前一個詞有關，所以在格狀詞組的搜尋上，我們使用動態規劃(dynamic programming)的維特比搜尋法(Viterbi Search)【4】。

維特比搜尋法使用遞迴方式來減少計算的複雜度，將由左而右的每個候選詞都看成一個節點，對於每個節點而言，都有一條到達此節點的最佳路徑，如圖 5-3。

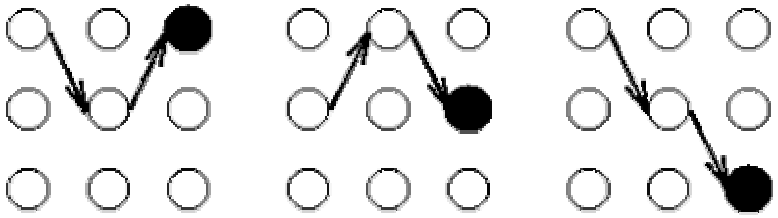


圖 5-3：時間 *t* 時到達各節點的最佳路徑示意圖

若將在時間 t 時結束於節點 i 的路徑中最大的機率稱為 $\delta(i,t)$，則可得

$$\delta(i,t)=\max_{j}^{N_{t-1}}\left\{\delta(j,t-1)\times\left(P\left(w_{i,t}\right)|\,P\left(w_{j,t-1}\right)\right)\right\}$$

(5.1)

$N_{t\text{-}1}$代表t-1 時間的候選節點個數

(5.1)式的意義就是，當要找到時間 t 於節點 i 的最大機率時，就是找出所有於時間 t-1 的 j 節點機率，與 j 節點至 i 節點的機率乘積之最大值即可。

最後再由N_t個 $\delta(i,t)$中，找出最大值 $\delta(t)$

$$\delta(t)=\max_{i=1}^{N_t}\delta(i,t)$$

並加以回溯，找出造成 $\delta(t)$的路徑，當 t 為句尾時，此詞串就是辨識出來的句子。

在本手持式語音辨識系統中，我們的做法是找出機率值最高的前 10 個詞串，做為使用者端的輸出，以供使用者點選。

5.2.3 本系統中語言模型的處理流程及系統效能評估

本系統的語言模型處理流程是先由輸入的候選音節中，每一到四個音節串成一個詞，並比對這個詞的注音是否出現在辭典中，若有，則將其視為格狀詞組中的一個節點，當全部的節點建好後，即完成構詞的步驟。再依照維特比搜尋法，找出到每個節點的最大機率，最後將機率排序，並列出機率值最高的前 10 串詞串做為使用者端的輸出；就完成了我們的語言模型在本手持式語音辨識系統中的工作。圖 5-4 為本手持式語音辨識系統中，語言模型處理時的流程圖。

在此系統的語言模型中，我們採用的平滑化方法為強化聶氏平滑法，因其於實驗中的效能表現最好。而系統前端也結合了沈揚智同學的去除聲音雜訊功能

【17】及謝宗儒同學的語者調適功能【18】。

我們請實驗室的 10 位同學做測試，測試語料為每人相同的 20 個短句，每句 4 到 8 個字，系統的正確率為 88.62%，精確率為 85.52%。

其中，正確率 $(Correct\ Rate)=\frac{H}{K}\times 100\%$，精確率 $(Accuracy)=\frac{H-I}{K}\times 100\%$

K 為測試文稿中所有字的數量，H 為辨識結果中正確的字的數量，I 為插入型錯誤(insertion error)的數量，即多辨識出不存在於文稿中的字的數量。

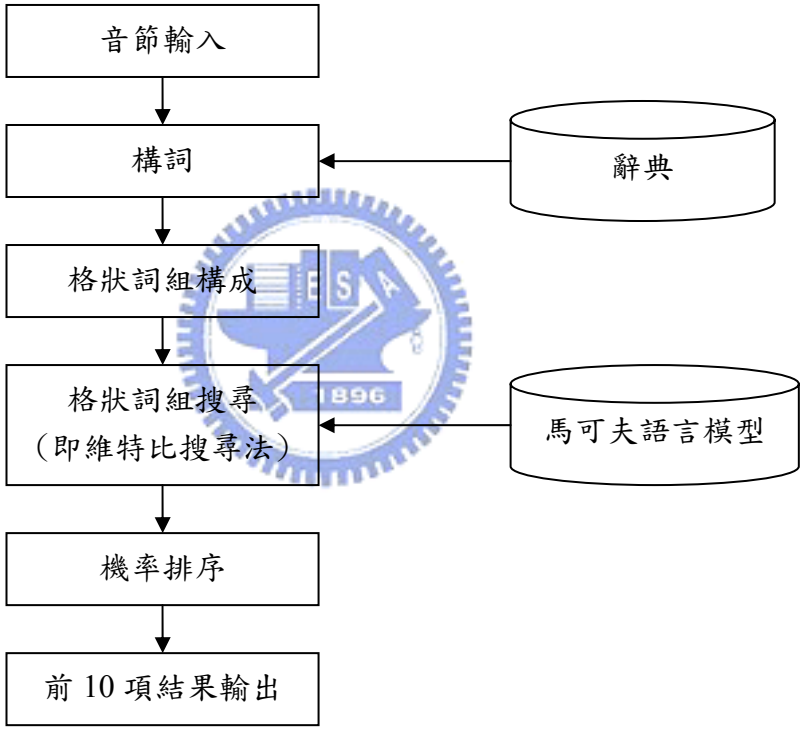


圖 5-4：語音辨識系統之語言模型的處理流程圖

第 6 章

結論及未來展望

6.1 結論

本論文針對語言模型最常遭遇到的『資料稀疏』問題，提出解決的平滑化方法，以增進語音辨識的效能。目前常用的『凱氏平滑法』及『聶氏平滑法』(包含『聶氏後退法』及『聶氏內插法』，其應用於雙連馬可夫模型時，對於欲估計的雙連詞串詞尾未在訓練語料中出現的情況，並無適當的機率評估方法。針對此點，我們由詞尾曾出現於訓練語料但整個雙連詞串並無出現於訓練語料的詞串做平滑化之後的機率，再進一步扣除小部分機率值，將此分配給詞尾未出現於訓練語料的雙連詞串，並以混淆度做為效能的評量標準。

我們由華視網站收集一年的語料，於每月的語料中，取出 180 則新聞做為測試語料，剩下的為訓練語料；由實驗結果可發現，『強化凱氏平滑法』比原來的『凱氏平滑法』低約 4 到 8 個混淆度單位，平均低了 6.65 個混淆度單位；『強化聶氏平滑法』也比原來『聶氏平滑法』中效能較佳的『聶氏後退法』低約 3 到 5 個混淆度單位，平均低了 4.50 個混淆度單位。由此可知，我們所提出的方法，的確可以降低語言模型的混淆度。

此外，也將我們建構的雙連馬可夫模型及實驗結果中效能最佳的『強化聶氏平滑法』，應用於中文語音辨識系統的語言模型部分，經實際測試，系統正確率可達 88.62%，精確率可達 85.52%。

6.2 未來展望

當我們實做的語言模型應用於語言辨識系統中時，還有一些可以改進的地方：

1. 整合高層次的語言能力

目前的系統完全依照機率統計時的高低，做為候選詞串的取捨，若我們能夠將高層次的語言能力，如文法、語意…等，結合在語言模型中，使得判斷出來的詞串更符合語法的規則。

2. 可調適的語言模型

因為本論文所建構的語言模型中，所使用的訓練語料與馬可夫模型都是固定的，並無法根據使用者所給予的候選詞串回饋來針對語言模型做調適的動作，因此，建立可調適的語言模型，將會使得語言模型更適合於該使用者。

3. 解決語音辨識時可能出現的插入型、替代型及刪除型錯誤

前端所使用的語音辨識系統目的是辨識大字彙的連續語音，所以難免會出現插入型錯誤(insertion error)、替代型錯誤(substation error)及刪除型錯誤(deletion error)，而我們的語言模型尚無法有效解決上述三種問題，因此，若加入高階的語言知識，或許可改善這三種問題。

4. 廣泛收集訓練語料

目前我們的訓練語料只取自於報紙，這會造成對於其他體裁的詞句辨識率效果

偏低，因此，應由各種不同的領域收集文章，以做為訓練語料，使得模型的應用層面能夠更廣泛。

參考文獻

【1】 中華電視公司全球資訊網-新聞頻道

"http://www.2cts.tv/default.aspx?ch=news"

【2】 中央研究院中文詞知識庫小組-中文詞知識庫

"http://ckip.iis.sinica.edu.tw/new/publication.htm#t5"

【3】 中央研究院中文詞知識庫小組-中文斷詞系統

"http://ckipsvr.iis.sinica.edu.tw/"

【4】 Introduction of Hidden Markov Models

"http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html"

【5】S.Young, et al., "The HTK Book 3.2.1," Cambridge University Engineering Department, 2001

【6】 X. Huang, A. Acero, and H. W. Hon, "Spoken Language Processing – A Guide to Theory, Algorithm, and System Development," Carnegie Mellon University, 2001

【7】 王小川, "語音訊號處理," 全華科技圖書股份有限公司, 台北, 民國 93 年 3 月

【8】 S. M. Katz, “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer,” IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 35, No. 3, pp. 400-401, Mar. 1987

【9】 W. A. Gale, “Good-Turing Smoothing Without Tears,” Journal of Quantitative Linguistics 2, 1995

【10】H. M. Meng, Z. Chen, Y. Shi and Y. C. Li, “A System for Spoken Query Information Retrieval on Mobile Devices,” IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 8, pp. 531-541, Nov, 2002

【11】H. Ney, and U. Essen, “On Smoothing Techniques for Bigram-Based Natural Language Modelling,” IEEE Int. Conf. Acoustic, Speech and Signal Processing, pp. 825-828, Canada, 1991

【12】R. Kneser, and H. Ney, “Improved Backing-Off for M-gram Language Modeling,“ IEEE Int. Conf. Acoustic, Speech and Signal Processing, Vol. 1, pp. 181-184, May. 1995

【13】S. F. Chen and J. Goodman, “An Empirical Study of Smoothing Techniques for Language Modeling,” 34-th Annual Meeting of the Association for Computational Linguistics, pp. 310-318, Santa Cruz, California, 1996

【14】楊燕珠, “Intelligent Language Modeling and Processing in Mandarin Speech Recognition,” 國立台灣大學, 資訊工程學研究所 碩士論文, 民國 82 年

【15】楊榮荃, “Language Modeling Techniques in Mandarin Speech Recognition,” 國立台灣大學, 資訊工程學研究所 碩士論文, 民國 83 年

【16】楊凱程, “Further Studies for Practical Chinese Language Modeling,” 國立台灣大學, 電機工程學研究所 碩士論文, 民國 87 年

【17】沈揚智, “The Study of Speech Enhancement in Additive Noise Environment for Speech Recognition,” 國立交通大學, 資訊工程學系研究所 碩士論文, 民國 94 年

【18】謝宗儒, “The Study of Speaker Adaptation for Speech Recognition,” 國立交通大學, 資訊工程學系研究所 碩士論文, 民國 94 年

附錄

【1】凱氏平滑法與強化凱氏平滑法的實例

w_{i-1}	w_i	$C(w_{i-1},w_i)$	$C(w_i)$	$P_{ori}(w_i\|w_{i-1})$	$P_{Katz}(w_i\|w_{i-1})$	$P_{EKatz}(w_i\|w_{i-1})$
大	片	100		100/200=0.5	0.5	0.5
大	大	65		65/200=0.325	0.325	0.325
大	又	30		30/200=0.15	0.15	0.15
大	手	3		3/200=0.015	d_r =0.9, 0.9*0.015=0.0135	0.0135
大	提	2		2/200=0.01	d_r =0.8, 0.8*0.01=0.008	0.008
大	工	0	150	0/200=0	0.0035*150/250=0.0021	0.0021
大	冬	0	95	0/200=0	0.0035*95/250=0.00133	0.00133
大	櫻	0	3	0/200=0	0.0035*3/250=$4.2*10^{-5}$	d'_r=0.9, $0.9*4.2*10^{-5}=3.78*10^{-5}$
大	篩	0	2	0/200=0	0.0035*2/250=$2.8*10^{-5}$	d'_r=0.8, $0.8*2.8*10^{-5}=2.24*10^{-5}$
大	籟	0	0	0	0	$(9.8*10^{-6})*(1/2)=4.9*10^{-6}$
大	糾	0	0	0	0	$(9.8*10^{-6})*(1/2)=4.9*10^{-6}$
總和		200	250	1	1	1

$$\alpha(w_{i-1}) = 1 - 0.5 - 0.325 - 0.15 - 0.0135 - 0.008 = 0.0035$$

$$\beta(w_{i-1}) = 0.0035 - 0.0021 - 0.00133 - 3.78\times10^{-5} - 2.24\times10^{-5} = 9.8\times10^{-6}$$

凱氏平滑法公式：

$$P_{katz}(w_i \mid w_{i-1}) = \begin{cases} \dfrac{C(w_{i-1},w_i)}{C(w_{i-1})} & if\ C(w_{i-1},w_i) > k \\ \dfrac{d_r \times C(w_{i-1},w_i)}{C(w_{i-1})} & if\ 0 < C(w_{i-1},w_i) \le k \\ \dfrac{\alpha(w_{i-1}) \times C(w_i)}{\sum_{w_j:C(w_{i-1},w_j)=0} C(w_j)} & if\ C(w_{i-1},w_i) = 0 \end{cases}$$

強化凱氏平滑法公式：

$$P_{Ekatz}(w_i \mid w_{i-1}) = \begin{cases} \dfrac{C(w_{i-1},w_i)}{C(w_{i-1})} & if\ C(w_{i-1},w_i) > k \\ \dfrac{d_r \times C(w_{i-1},w_i)}{C(w_{i-1})} & if\ 0 < C(w_{i-1},w_i) \le k \\ \dfrac{\alpha(w_{i-1}) \times C(w_i)}{\sum_{w_j:C(w_{i-1},w_j)=0} C(w_j)} & if\ C(w_{i-1},w_i) = 0, C(w_i) > k \\ \dfrac{d'_r \times \alpha(w_{i-1}) \times C(w_i)}{\sum_{w_j:C(w_{i-1},w_j)=0} C(w_j)} & if\ C(w_{i-1},w_i) = 0, 0 < C(w_i) \le k \\ \beta(w_{i-1}) \times \dfrac{1}{T} & if\ C(w_{i-1},w_i) = 0, C(w_i) = 0 \end{cases}$$

【2】聶氏平滑法(含聶氏內插法與聶氏後退法)與強化聶氏平滑法的實例

w_{i-1}	w_i	$C(w_{i-1},w_i)$	$N_{1+}(\cdot,w_i)$	$P_{ori}(w_i\|w_{i-1})$	$P_{BKN}(w_i\|w_{i-1})$	$P_{IKN}(w_i\|w_{i-1})$	$P_{EKN}(w_i\|w_{i-1})$
大	片	100	18	100/200=0.5	(100-0.5)/200=0.4975	0.4975+0.0125/10 =0.49875	0.4975+0.0125*(17.5/50+0.08/10)=0.501975
大	大	65	13	65/200=0.325	(65-0.5)/200=0.3225	0.3225+0.0125/10=0.32375	0.3225+0.0125*(12.5/50+0.08/10)=0.325725
大	又	30	9	30/200=0.15	(30-0.5)/200=0.1475	0.1475+0.0125/10=0.14875	0.1475+0.0125*(8.5/50+0.08/10)=0.149725
大	手	3	3	3/200=0.015	(3-0.5)/200=0.0125	0.0125+0.0125/10=0.01375	0.0125+0.0125*(2.5/50+0.08/10)=0.013225
大	提	2	2	2/200=0.01	(2-0.5)/200=0.0075	0.0075+0.0125/10=0.00875	0.0075+0.0125*(1.5/50+0.08/10)=0.007975
大	工	0	3	0/200=0	0.0125*3/5=0.0075	0.0125/10=0.00125	0.0125*(2.5/50+0.08/10)=0.000725
大	冬	0	1	0/200=0	0.0125*1/5=0.0025	0.0125/10=0.00125	0.0125*(0.5/50+0.08/10)=0.000225
大	櫻	0	1	0/200=0	0.0125*1/5=0.0025	0.0125/10=0.00125	0.0125*(0.5/50+0.08/10)=0.000225
大	籟	0	0	0	0	0.0125/10=0.00125	0.0125*(0.08/10)=0.0001
大	糾	0	0	0	0	0.0125/10=0.00125	0.0125*(0.08/10)=0.0001
總和		200	50	1	1	1	1

聶氏後退法：

$$P_{BKN}(w_i \mid w_{i-1}) = \begin{cases} \dfrac{\max\{C(w_{i-1},w_i)-D,0\}}{C(w_{i-1})} & \text{if } C(w_{i-1},w_i)>0 \\ \dfrac{\alpha(w_{i-1})\times N_{1+}(\bullet,w_i)}{\sum_{w_k:C(w_{i-1},w_k)=0} N_{1+}(\bullet,w_i)} & \text{if } C(w_{i-1},w_i)=0 \end{cases}$$

強化聶氏平滑法：

$$\begin{cases} P_{EKN}(w_i \mid w_{i-1}) = \dfrac{\max\{C(w_{i-1},w_i)-D,0\}}{C(w_{i-1})} + \alpha(w_{i-1})\times P'_{EKN}(w_i) \\ P'_{EKN}(w_i) = \dfrac{\max\{N_{1+}(\bullet,w_i)-D,0\}}{\sum_{w_j} N_{1+}(\bullet,w_j)} + \beta\times\dfrac{1}{V} \end{cases}$$

聶氏內插法：

$$P_{IKN}(w_i \mid w_{i-1}) = \frac{\max\{C(w_{i-1},w_i)-D,0\}}{C(w_{i-1})} + \alpha(w_{i-1})\times\frac{1}{V}$$