國立台灣師範大學
資訊工程研究所碩士論文

指導教授：　陳柏琳 博士

多種鑑別式語言模型應用於語音辨識之研究

Exploiting Discriminative Language Models for Speech Recognition

研究生：　劉家妏　撰

中華民國　　九十九 年　八　月

摘要

N 連(*N*-gram)語言模型在語音辨識器中扮演著關鍵性的角色，因為它可幫助辨識器從其大量輸出的候選詞序列中，區辨出正確與非正確的候選詞序列。然而，因 *N* 連語言模型的訓練目標為最大化訓練語料的機率，而不是以最佳化語音辨識評估量為目標，導致在語音辨識效能表現上有所侷限。本論文首先探討多種鑑別式語言模型(Discriminative Language Model, DLMs)，雖然它們的訓練目標函數不同，但皆符合提昇語音辨識率的直覺精神；同時，也從理論與實用的觀點比較這些鑑別式語言模型的效用。再者，我們提出測試語句相關之鑑別式語言模型(Test Utterance-driven Discriminative Language Model, UDLM)，此模型可即時推算其模型參數而適應於語音辨識的應用。最後，我們將最大化事後機率法(Maximum a Posterior, MAP)結合測試語句相關之鑑別式語言模型以希望最大化事後機率法所產生的辨識結果，能幫助測試語句相關之鑑別式語言模型獲致更佳的語音辨識率提昇。本論文的實驗皆建立在臺灣所收集的中文廣播新聞語料上，實驗結果顯示本論文所提出之作法似乎有其可行性。

Abstract

N-gram language modeling is a crucial component in any speech recognizer since it is expected to help the recognizer distinguish the correct hypothesis from the other incorrect ones in an extremely large output space of the recognizer. However, the *N*-gram language models are inadequate since they usually set the goal of training at maximizing the likelihood of a large amount of training text, but not at optimizing the final performance measure of speech recognition. In this thesis, we first investigate a wide variety of discriminative language models (DLMs), which have their roots stemming from different training objectives but are consistent with the intuition of enhancing recognition performance. The utilities of these DLMs are compared both theoretically and empirically. Further, we also propose a test utterance-driven DLM (UDLM) that can efficiently infer its model parameters on-the-fly and accommodate itself well to speech recognition applications. As a final point, we pair UDLM with the maximum *a posteriori* probability (MAP) language model adaptation approach for better recognition performance. All experiments are conducted on a Mandarin broadcast news corpus compiled in Taiwan, and the associated results seem to demonstrate the feasibility of the proposed methods.

誌謝

感謝父母與家人的陪伴，因為有您們在背後的支持，我才能心無旁騖的完成我的學業；您們從小到大對我無私的栽培，是我一輩子都報答不了的。在求學的這段期間總是讓您們擔心，謝謝您們。

感謝指導教授陳柏琳博士在我專題生到研究所時期這三年的教導，不管在研究或是待人處事方面，老師總是不厭其煩的一遍遍提醒，老師的教誨我會謹記在心。也謝謝老師在最後那段時間，不管多忙仍然抽空跟我討論，給我研究上許多建議與鼓勵，老師對學生的奉獻與付出，只有由衷的感謝與感激。

感謝口試委員古鴻炎博士、洪志偉博士及李俊仁博士對論文的指正與建議，讓我的論文能更趨完整。

感謝實驗室的學長姐，斯涵學姐、鴻欣學長、韋豪學長及鳳萍學姐，謝謝你們在學業或人生方面給予的建議與幫助，都讓我受益良多；士翔學長、冠宇學長及永典學長，因為有跟你們在學業上的討論，讓我能夠從不同角度思考，謝謝你們耐心的指導與建議；謝謝鈺玫，認識妳讓我的研究生活多了許多樂趣，一起努力和互相鼓勵的這段時間，真的很慶幸我的同梯是妳，以後也要互相加油。也謝謝實驗室的學弟妹，珮寧、紋儀及敏軒，你們的加入讓實驗室多了許多歡笑，讓我覺得實驗室就像一個大家庭一樣；而在我們忙碌的時候，你們也盡全力幫忙我們，真的很謝謝你們，接下來也要加油。

研究生活兩年一轉眼就過了，學生生涯也劃下句點，在語音實驗室的這兩年時間，對我來說是一輩子珍貴的回憶，謝謝大家。

家妏 謹誌

圖目錄

表目錄

第 1 章 緒論

1.1 研究背景

隨著時代的進步、科技的發展，許多電子產品的發明改變了我們的日常生活，為了方便使用者的攜帶，體積輕巧幾乎已是所有可攜式裝置必備的條件；但隨著產品的迷你化，手動的操作已漸漸無法滿足人們的需求，甚至造成了使用上的不方便。為了因應這樣的改變，語音操作成為其中一個解決方法。

語音是人類最自然且直接的溝通方式，而如何讓電腦達到如人類般具備「聽、說、讀、寫」能力就是語音技術的最大課題。而應用在電子產品的操作上，首先要做到的是如何讓這些電子產品能夠「聽」懂使用者的語音輸入，將語音資訊轉換成文字後，再對文字進行語意(Semantics)分析，進一步達到使用者的要求；而其中將語音資訊轉換成文字的過程，即是自動語音辨識(Automatic Speech Recognition, ASR)的技術。可想而知，若辨識錯誤過多，可能會影響語意的理解，因此無法達到使用者的要求。

過去數十年來，在眾多專家學者努力下，自動語音辨識的技術已得到大幅的提升，現今只要提供足夠的訓練資料，經過一連串的訓練後，即可建立一套有一定辨識率的語音辨識系統。而透過語音人機互動這樣以往無法想像的功能，已透過現今科技的發展得以實現，成為生活的一部分。除了人機互動的應用，語音辨識還可應用在許多領域，如語音文件檢索與摘要(Spoken Document Retrieval and Summarization)，也需要自動語音辨識技術將語音資料轉換成文字後，再進行分析。因此，語音辨識的技術在未來人類的日常生活中將扮演舉足輕重的角色。

本章針對研究需要，將簡介語音辨識的研究內容，並於接下來的章節說明本論文的研究重點。

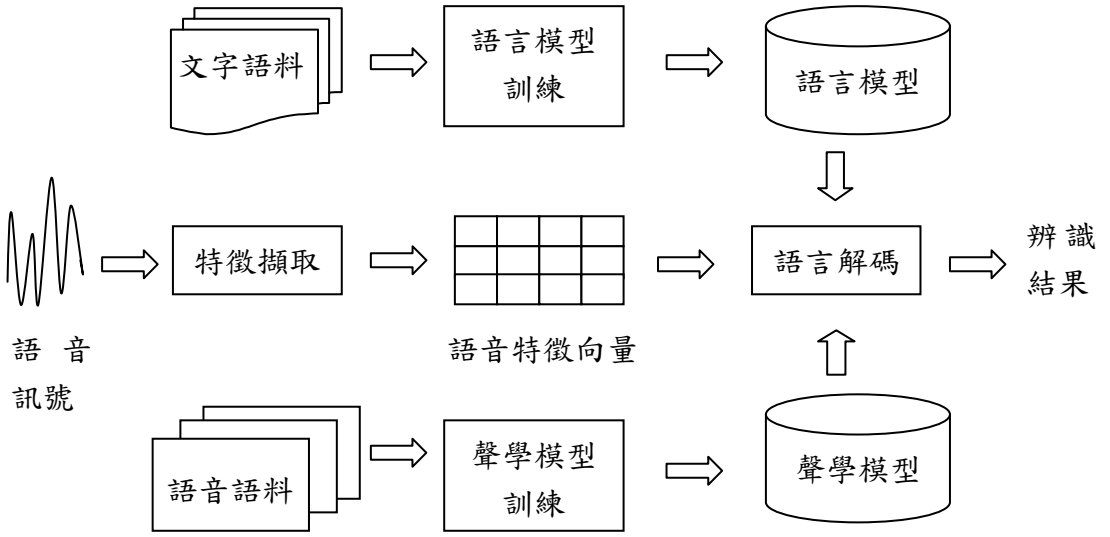

圖 1-1 自動語音辨識流程圖

1.2 語音辨識簡介

語言對人類來說是最自然且直接的溝通方式，而電腦將語音訊號轉換成文字的過程，則須透過自動語音辨識(Automatic Speech Recognition, ASR)來達成。在實現自動語音辨識系統時，大致可分為四個部份，特徵擷取(Feature Extraction)、聲學模型訓練(Acoustic Model Training)、語言模型訓練(Language Model Training)及語言解碼(Decoding)，其流程如圖1-1所示。我們首先須將聲音數位訊號經由特徵擷取(Feature Extraction)而產生出能代表語音的聲學特性(Acoustic Characteristics)且易於電腦處理的聲學特徵向量；接著，將語音語料轉換的聲學特徵向量透過機率模型建立起其對應的聲學模型(Acoustic Model)，串連起聲音與文字間的對應關係；最後，再由使用大量文字語料訓練而成的語言模型(Language Model)估測詞彙或語句出現的機率，以減輕因詞彙發音相近所造成的混淆，並找出語音訊號最可能對應的詞序列，此過程一般是以統計式方式與最大事後機率(Maximum a Posteriori, MAP)來實行，以數學式子表示即如式(1-1)所示[Jelinek 1999]：

$$
\begin{aligned}
W^{*} &= \arg\max_{W} P(W \mid X) \\
&= \arg\max_{W} \frac{P(W)p(X \mid W)}{p(X)} \\
&\approx \arg\max_{W} P(W)p(X \mid W)
\end{aligned}
\tag{1-1}
$$

在輸入一語音訊號X後，自動語音辨識主要目的就是找出一串最有可能的對應詞序列W^{*}，因$P(W \mid X)$無法直接估算，所以經由貝氏定理轉換；又因為對同一語音訊號X來說$p(X)$皆相等，不會影響排序，因此就可將式子化簡為$\text{argmax}_{W} P(W)p(X \mid W)$，即可找到相對應最可能的詞串。式(1-1)中$p(X \mid W)$為聲學模型，代表給定一詞串$W$，產生某語音訊號$X$的機率；而$P(W)$為語言模型，表示產生某一詞串$W$的機率。

接下來分別概述自動語音辨識過程中，特徵擷取、聲學模型、語言模型、語言解碼四個部份：

（一）特徵擷取：特徵擷取是從一段語音訊號中擷取出其較重要的參數，轉換成語音辨識系統易分析、使用的資料型態，如特徵向量(Feature Vectors)，期望語音特性可表現在擷取出的特徵向量上。特徵擷取常見的方法有梅爾倒頻譜係數(Mel-frequency Copstral Coefficients, MFCC)等。

（二）聲學模型：由於語音具有時序性，一般而言採用由左至右(Left-to-right)的隱藏式馬可夫模型(Hidden Markov Model, HMM)來建立聲學模型。隱藏式馬可夫模型包含多個狀態(State)，對每個狀態而言，都有其各自的觀測機率分布(Observation Probability Distribution)，亦有相對應的狀態轉移機率(State Transition Probability)，用以控制狀態停留或轉移。

在中文語音辨識中，中文音節可視為兩個次音節(Subsyllable)，分別為聲母(Consonant)與韻母(Vowel)，因此需分別建立對應的聲學模型，稱為INITIAL與

FINAL模型。常見的訓練方法有最大化相似度訓練法(Maximum Likelihood, ML)[Bahl *et al.* 1983]、最大話交互資訊(Maximum Mutual Information, MMI)[Bahl *et al.* 1986]、最小化分類錯誤(Minimum Classification Error, MCE)[Juang and Katagiri 1992]或是最小化音素錯誤(Minimum Phone Error, MPE)[Povey 2004]等。

（三）語言模型：由於聲學模型只能辨識某一段語音訊號發出的音節序列，無法確認其對應的詞與詞間的連接關係，因此便需要語言模型的存在，以解決聲學上的混淆。目前最廣泛使用的是統計式語言模型，利用 $N-1$ 階馬可夫假設($N-1$-order Markov Assumption)，即為 N 連(N-gram)語言模型。而近年來，一種新型的語言模型訓練法被提出，稱為鑑別式語言模型，其目的在於希望能直接減少語音辨識錯誤率來進行訓練，而不是如傳統 N 連語言模型般，希望找出機率最高的詞序列。

（四）語言解碼：進行語言解碼時，是利用特徵擷取時取得的特徵向量，與此特徵向量在聲學模型上的相似度及對應詞序列在語言模型上的機率，找出最可能的詞序列，如上述的式(1-1)所示。一般使用維特比動態規劃搜尋(Viterbi Dynamic Programming Search)。除此之外，也會進行路徑裁剪(Pruning)技術將可能性較低的詞先行排除。

1.3 論文貢獻

本論文主要貢獻有幾個：(I)介紹多種基於不同訓練準則的鑑別式語言模型，並比較它們在語音辨識重新排序的表現；(II)另外，我們提出針對個別測試語句，透過線性組合不同訓練語料所訓練的語言特徵之權重向量的方法(或稱之語句相關之鑑別式語言模型)，以改進傳統鑑別式語言模型在測試過程中，所有測試語句皆同樣地使用由所有訓練語料訓練出的權重向量的缺點，讓不同測試語句擁有各自的組合係數來線性結合不同訓練語料所訓練而得的語言模型特徵權重參數向量，以期新的權重向量能更加符合測試語句的特性；(III)再者，我們將語句相關之鑑別式語言模型所訓練的權重參數向量與傳統鑑別式語言模型所訓練的權重參數向量利用插補法(Interpolation)結合，以期兩者依不同訓練目標所訓練的權重參數向量能達到互補的效果，並獲得更顯著的效果。(IV)我們將最大化事後機率法(Maximum a Posterior, MAP)結合語句相關之鑑別式語言模型，期望藉由最大化事後機率法所產生的辨識結果，能幫助語句相關之鑑別式語言模型獲致更顯著的語音辨識率提昇。實驗結果顯示本論文所提出的語句相關之鑑別式語言模型，在部份的組合方法下，可相較於僅使用三連語言模型、或使用傳統鑑別式語言模型的基礎大詞彙連續語音辨識系統，有一定的語音辨識率提升；而將語句相關之鑑別式語言模型與傳統鑑別式語言模型的權重參數向量結合的方式，更能有相當程度的語音辨識率提升。

1.4 論文章節安排

本論文接下來的章節概要如下：

第 2 章 介紹傳統 *N* 連語言模型及其改進方法與鑑別式語言模型的相關文獻。

第 3 章 介紹鑑別式語言模型的基礎精神與多種不同訓練精神的鑑別式語言模型。

第 4 章 介紹本論文提出的語句相關之鑑別式語言模型的三種權重參數向量產生方法。

第 5 章 實驗結果討論與分析。

第 6 章 結論及未來展望。

第 2 章 文獻回顧

2.1 *N* 連語言模型與其改進方法

語言模型在自動語音辨識中的主要的作用，是代表語言使用的習慣與規律。當前最常使用的語言模型為統計式語言模型(Statistical Language Model, SLM)，它統計一個詞在訓練語料中的出現情形，給予該詞一個機率，以代表該詞於某種語言使用環境下的重要性。

而其中最基本的是 *N 連*語言模型，其設計理念與 Claude Elwood Shannon 在資訊理論(Information Theory)研究中所提出的問題有關：「給定一個字母序列，下一個字母最有可能會是什麼？」。可藉由從訓練資料中求得給定歷史序列下所有字母的機率分布解決這個問題，而統計式 *N 連*語言模型即為一個基於歷史資訊之模型(History-based Model)：每一個詞 w_i 的出現，都只與其前 $N-1$ 個詞有關，此 $N-1$ 個詞組成的詞序列 $w_{i-N+1}w_{i-N+2}...w_{i-1}$ 即為詞 w_i 之歷史詞序列 h_i 。

因此辨識的過程，即是估測每一個詞在其前緊鄰 $N-1$ 個歷史詞序列已知的情況下的條件機率：

$$P(W)=\prod_{i=1}^{n}P(w_i \mid h_i)\approx\prod_{i=1}^{n}P(w_i \mid w_{i-N+1}w_{i-N+2}...w_{i-1}) \qquad (2\text{-}1)$$

其中 $P(w_i \mid h_i)$ 的訓練是透過最大相似度估測(Maximum-Likelihood Estimation, MLE)來做估測，以三連詞為例，其估測值為：

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

$C(W)$代表詞序列W在訓練語料出現的次數。接著由語音辨識器所產生的詞圖(Lattice)或前 M 條最佳(在此指機率或分數最高)辨識候選詞序列(M -best Recognition Hypotheses)中，選取相對於此語音訊號機率最高的詞序列(Word Sequence)做為最後語音辨識結果

另外，除了基本的 N 連語言模型外，過去已有許多使用不同層次資訊的語言模型被提出，依語言資訊層次的不同，我們可將語言模型概略分成四類[邱炫盛等人 2007]：

(一)詞相關語言模型：為了改進N連語言模型只能捕捉短距離詞彙資訊的限制，此類語言模型嘗試捕捉更長距離的詞彙資訊。如混階層馬可敷模型(Mixed Order Markov Model)[Saul and Pereira 1997]、觸發對語言模型(Trigger-based Language Model)[Troncoso *et al.* 2004]等。

(二)詞類別相關語言模型：類似N連語言模型，詞與詞的關係可以透過固定或非固定的詞類別建立。建立詞之間的關係後，因為歷史詞序列亦是由詞組成，因此可以進一步找出序列中的詞與欲預測詞之間的關係。例如 N 連類別模型(Class-based N-gram Model) [Brown *et al.* 1992]、聚合式馬可夫模型(Aggregate Markov Model, AMM)[Troncoso *et al.* 2004]等等。

(三) 語句結構相關語言模型：此類模型通常會同時使用詞彙與詞類別的資訊，所以能夠根據歷史詞序列的句型決定辨識詞的可能性。例如結構化語言模型(Structured Language Model) [Chelba and Jelinek 2000]等等。

(四) 文件主題相關語言模型：透過隱藏或非隱藏的參數估測，針對一篇或一群文件的主題性建立模型。歷史詞序列可視為尚未完成的文件，假設完成的程度

已經能呈現某些主題，透過此類模型可找出其主題性。例如混合主題式語言模型(Mixture-based Language Model)[Clarkson and Robinson 1997]、潛藏語意分析(Latent Semantic Analysis, LSA)[Bellegarda 2005]、機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA) [Gildea and Hofmann 1999]、潛藏狄利克雷分配(Latent Dirichlet Allocation, LDA) [Tam and Schultz 2005]等皆屬於此類模型。

另外，由於在詞圖(Lattice)或前 M 條最佳辨識候選詞序列(M-best List)中，可能還存在其它錯誤率較低的候選詞序列可供語音辨識器做選擇，因此使用傳統 N 連語言模型所找出最高機率詞序列未必是最佳(錯誤率最低)的語音辨識結果；而重新排序(Rearanking)的研究便是希望透過使用更多語言特徵的語言模型與較好的模型訓練演算法來解決此問題。例如，近年來有許多以最小化辨識錯誤率為目標的鑑別式語言模型被提出，接下來的章節將詳細介紹鑑別式語言模型之相關研究。

2.2 鑑別式語言模型

有別於傳統統計式語言模型，鑑別式語言模型(Discriminative Language Model)是直接以最小化辨識錯誤率為訓練目標；也就是希望藉由調整語言特徵在語言模型中所對應的權重參數，以最佳方式結合各式語言特徵，使語音辨識器能再對僅使用基礎 N 連語言模型所產生的前 M 條最佳辨識候選詞序列進行重新排序，使得錯誤率較低的候選詞序列能有較高的排序來做為最終辨識結果，而降低辨識錯誤率。

鑑別式訓練一開始被應用於自然語言處理領域[Collins and Koo 2000]與語音辨識的聲學模型領域[Povey 2004]，已有顯著的成效；近幾年來拓展到語音辨識的語言模型領域[Kuo and Chen 2005；Roark *et al*. 2007；Oba *et al*. 2010]，並能獲致不錯的效果。鑑別式語言模型於語音辨識結果重新排序方法的流程圖如圖 2-1 所示，首先利用基礎辨識器對訓練語料產生的 M 條最佳辨識結果進行鑑別式訓練，訓練出來的鑑別式語言模型為相對應每一維語言特徵的語言特徵權重參數；接著將基礎辨識器對測試語料產生的 M 條最佳辨識結果，利用訓練語料所訓練的鑑別式語言模型進行重新排序，以期錯誤率較低的候選詞序料能有較高的排序，將會在第 3 章做詳細的解說。

在 1998 年，Rigazio 等人[Rigazio *et al*. 1998]以最小化分類錯誤為目標，對語言模型機率及語言權重作鑑別式的訓練及調適，其目標在於訓練分類器(Classifier)，使其能從 M 個最佳辨識結果中擇其預期錯誤率(Expected Error Rate)最小者。語言模型機率代表詞序列在特定語言使用環境下的重要性，而語言權重(Language Weight)則是代表在辨識系統中，語言模型與聲學模型二者相較之下的可信賴度(Relative Reliability)。

之後，Warnke 等人[Warnke *et al*. 1999]在 1999 年提出利用最大相互資訊估

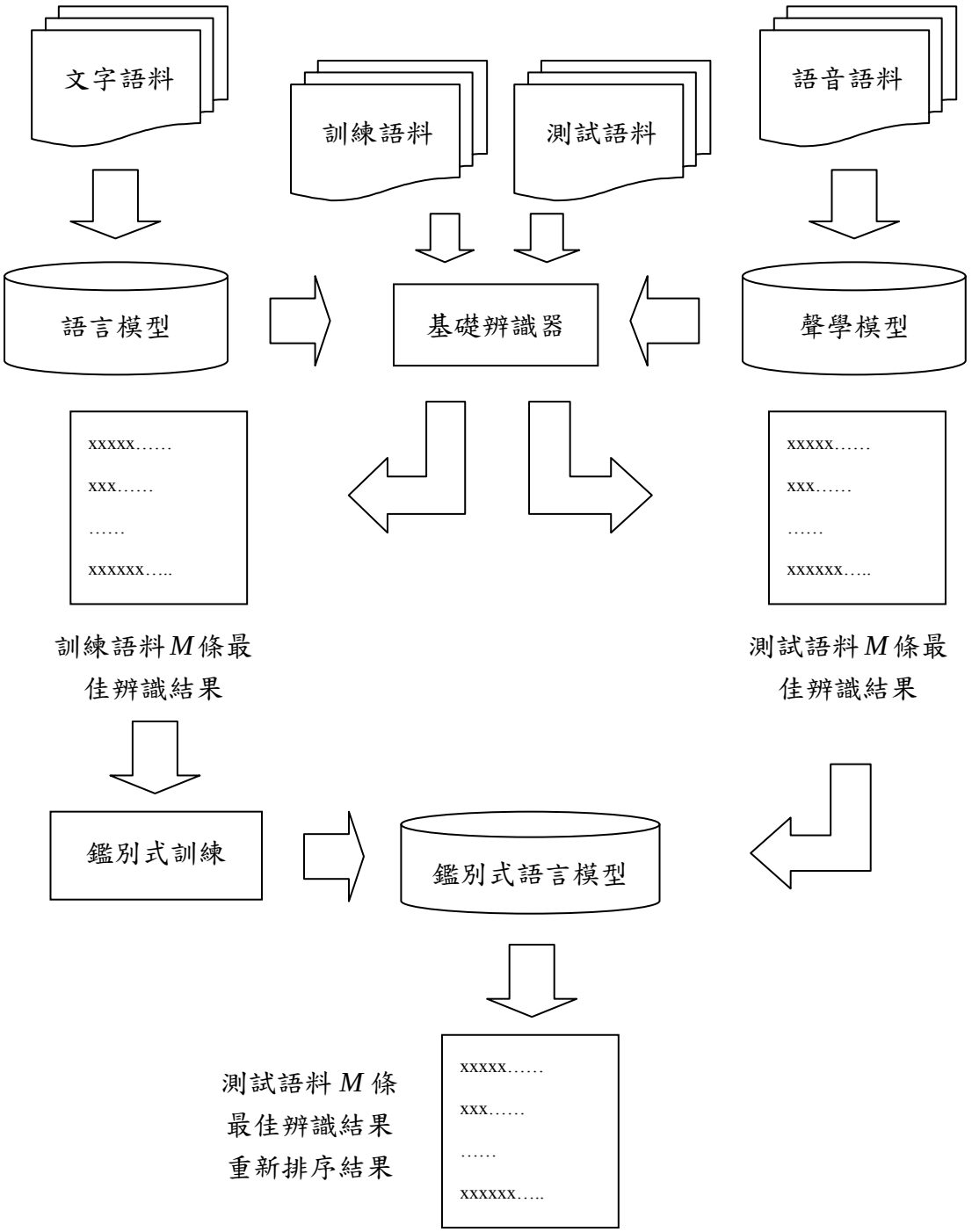

圖 2-1 鑑別式語言模型流程圖

測(Maximum Mutual Information Estimation, MMIE)與最小化分類錯誤(Minimum Classification Error, MCE)來訓練語言模型插補法(Language Model Interpolation)的權重，並達到一定程度的辨識錯誤率降低。

接著在 2000 年，Collins 與 Koo[Collins and Koo 2000]提出了 Reranking Boosting 演算法的鑑別式訓練方法，將其應用在自然語言處理領域。此方法主要是利用多種語言特徵的結合，對語法分析器的結果進行重新排序，期望較高正確率的語句經由重新排序後能有較高排序。

而在 2002 年 Kuo 等人[Kuo *et al*. 2002]則提出以最小分類錯誤為基礎的鑑別式語言模型訓練，目的在於區分最接近正確辨識結果的候選詞序列與其他候選詞序列。其方法為比較 *N* 連詞在正確轉寫語句與候選詞序列中的出現情形，以決定如何增減該候選詞序列之機率值。

同年，Collins[Collins 2002]將感知器演算法(Perceptron Algorithm)應用在自然語言處理領域中。他們將感知器演算法視為一種最大化熵值法(Maximum-Emtropy, ME)與條件式隨機域(Conditional Radom Fields, CRF)的變形。此演算法利用維特比解碼(Viterbi Decoding)並加上更新的概念所形成；他並對感知器演算法運用在分類問題時，會在有限訓練次數收斂至無訓練錯誤提出了理論上的證明。

同樣的，在機器翻譯(Machine Translation)領域中，也有傳統的統計式訓練方法的訓練準則跟最終結果的評估值的關聯性不高的問題；於是在 2003 年，Och[Och 2003]提出最小化錯誤率訓練(Minimum Error Rate Training, MERT)方法應用於機器翻譯領域；此方法期望能以直接最小化最終結果的錯誤率為目標進行訓練，並達到更佳的效果。

而在 2004 年時，Roark 等人[Roark *et al*. 2004a]提出以權重有限狀態機(Weighted Finite-state Automata)實作的感知器演算法，將鑑別式語言模型應用到語音辨識領域，以最小化平方錯誤(Minimum Square Error, MSE)為目標，針對基礎語音辨識器的結果進行重新排序，以期錯誤率最低的候選詞序列能有較高的排

序，作為最終辨識結果。此外，亦採用條件式隨機域方法進行鑑別式訓練[Roark *et al*. 2004b]。

在 2005 年，Kuo 與 Chen [Kuo and Chen 2005]則是提出最大化訓練語料中詞圖的期望正確率(也就是最小化詞錯誤(Minimum Word Error, MWE))來估測語言模型機率。其方法使用延伸波式(Extended Baum-Welch)演算法[Gopalakrishnan et al. 1991]推得語言模型參數估測之更新公式，透過遞迴更新語言模型機率。

同年，Gao 等人[Gao *et al*. 2005]介紹了 4 種語言模型調適方法：包括最大化事後機率法(Maximum a Posteriori, MAP)、與 3 種鑑別式語言模型調適法，Boosting 演算法、感知器演算法與最小化樣本風險法(Minimum Sample Risk Methods, MSR)，並運用在日文假名與漢字的轉換上，並比較了這些方法在字錯誤率降低的效能。

而同一時期，Collins 等人[Collins *et al*. 2005]將句法(Syntactic)的特徵加入鑑別式語言模型中，對前 1000 條最佳辨識結果，利用感知器演算法進行重新排序，並得到不錯的效果，代表句法的特徵對鑑別式語言模型是有效的特徵。

接著在 2006 年，Zhou 等人[Zhou *et al*. 2006]採用 Reranking Boosting 演算法的鑑別式訓練方法運用於語言模型調適上，並比較多種鑑別式訓練方法的訓練時間與效果，其中包括感知器演算法、最小化樣本風險法(MSR)，與排序式支援向量機(Ranking Support Vector Machine, Ranking SVM)。

而 Roark 等人[Roark *et al*. 2007]在 2007 年則是利用感知器演算法建立鑑別式全域線性模型，用全域條件式對數線性模型(Global Conditional Log-linear Model, GCLM)的方法進行參數調整。

同年，Singh-Miller 與 Collins[Singh-Miller and Collins 2007]將以觸發為基礎

(Trigger-based)的語言特徵運用到鑑別式語言模型中。所謂觸發為基礎的語言特徵是指在對話中同時出現的詞彙特徵。

在 2008 年，Kobayashi 等人[Kobayashi *et al.* 2008]將最小化錯誤率訓練方法應用於語音辨識領域，可將錯誤率視為一種樣本權重(Sample Weight)的資訊。所謂的樣本權重代表候選詞序列的重要性，因此最小化錯誤率訓練方法即是用錯誤率代表各候選詞序列的樣本權重，以期望加入樣本權重資訊可訓練出準確率更高的模型。

同年，Zhou 與 Meng[Zhou and Meng 2008]將鑑別式語言模型與傳統 *N* 連語言模型進行結合。因為鑑別式語言模型可視為一種線性的架構，因此可將兩者結合使鑑別式語言模型擁有 *N* 連語言模型的性質，而能運用在更廣泛的領域中，如重新計分(Rescoring)上。實驗結果顯示重新計分的結果的確能有效降低錯誤率。

接著，Roark[Roark 2009]在 2009 年出版的關於語音辨識的書中，對鑑別式語言模型運用在大詞彙連續語音辨識領域做了概括性的解說，他針對三個方面來探討鑑別式語言模型：訓練語料、學習演算法、與特徵方面。

Magdin 與 Jiang[Magdin and Jiang 2009] 在同年提出一種新型的鑑別式訓練演算法，運用在大詞彙連續語音辨識上。他們利用最大化相互資訊估算(Maximum Mutual Information Estimation, MMIE)設計目標函數。此種非線性的最大化相互資訊估算的目標函式利用一線性的最大期望(Expectation Maximization, EM)輔助函式來近似，將鑑別式 *N* 連語言模型簡化成線性問題，其實驗結果能獲致一定程度的效能提昇。

另外， Xu 等人[Xu *et al.* 2009] 同樣在 2009 年提出一種自身監督型(Self-supervised)鑑別式訓練方法，將其運用在自動語音辨識估測語言模型上。不

像傳統鑑別式語言模型需要語音訊號的正確轉寫文字，此模型只需要語音訊號跟大量的文字即可。此種語言模型也是以指數形式為基礎，但只利用那些易混淆(confusing)的詞彙資訊來訓練語言模型。

Rastrow 等人[Rastrow *et al*. 2009] 在同年提出一種條件式(Constrained)鑑別式語言模型，他們提出了三種技術來提高鑑別式語言模型的效能：(1)對沒見過的事件更新其後撤式機率(back-off probability)、(2)正規化 *N* 連更新以確定保持機率特性、和(3)對 *N* 連機率更新增加一相對熵值(relative-entropy)為基礎的全域條件。

接著，Kaufmann 等人[Kaufmann *et al*. 2009] 在同一年將一種精確文法當作鑑別式語言模型的特徵，運用在新聞廣播上，對前 *M* 條最佳辨識結果進行重新排序，實驗結果顯示此種特徵對新聞廣播語音辨識有一定的效果。

在 2010 年，Oba 等人[Oba *et al*. 2010]將全域條件式對數線性模型加入樣本權重的元素進行改良，提出權重式全域條件式對數線性模型(Weighted Global Conditional Log-linear Model, W-GCLM)並與最小化錯誤率訓練方法及加入樣本權重的 Reranking Boosting 演算法進行比較。

而 Magdin 與 Jiang[Magdin and Jiang 2010]也在 2010 年利用之前提出的最大化相互資訊估算法為基礎，融合大邊界估算(Large Margin Estimation)來設計目標函數。其觀念為希望能最大化正確轉寫語句與候選詞序列間的最小邊界。如同最大化相互資訊估算法，也利用線性的最大期望(EM)輔助函式來近似目標函式，使鑑別式 *N* 連語言模型簡化成線性問題。

緊接著，Arisoy 等人[Arisoy *et al*. 2010]也在同年利用分別定義幾種文法的資訊當作鑑別式語言模型的特徵，運用在土耳其語的語音辨識上，也得到不錯的效果。

同一時期 Huang 等人[Huang *et al*. 2010]也提出了一種鑑別式訓練方法，其不需要任何轉寫的聲學資料。他們利用最小化詞序列條件式熵值的方法來進行，並做了兩種設定來進行學習。

第 3 章 鑑別式語言模型

3.1 鑑別式語言模型訓練之基礎定義

本節將介紹鑑別式語言模型的基本概念與基礎定義。鑑別式語言模型以直接最小化辨識錯誤率為目標，希望對基礎辨識器所產生前 M 條最佳辨識候選詞序列(或詞圖中所有詞序列)重新排序，以期擁有較低辨識錯誤率的詞序列可有較高的排序。以前 M 條最佳辨識候選詞序列為例，每一條候選詞序列分別以語言特徵向量表示，其中每維特徵值皆有其對應的特徵權重參數。鑑別式語言模型的訓練目標即在於訓練出最佳的特徵權重向量，使前 M 條最佳辨識候選詞序列中最低錯誤率的詞序列，其語言特徵向量與特徵權重向量內積後的分數能為最高。以下將對鑑別式訓練定義其參數：

(1) 給定一語音訊號 x_i，假設基礎辨識器對應此語音訊號產生的前 M 條最佳辨識候選詞序列集合為 $\mathrm{GEN}(x_i)=\{W_{i,j}\}$，其中 j 介於 1 到 M 之間。

(2) 將訓練語料表示成 $\{x_i, W_i^R\}$ 的集合，i 介於 1 到 L 之間，L 為訓練語料句數；W_i^R 為 $\mathrm{GEN}(x_i)$ 中最低錯誤率詞序列；而測試語料則表示成 $\{y_k\}$ 的集合，k 介於 1 到 K 之間，K 為測試語料之句數。

(3) 對每條候選詞序列 $W_{i,j}$ 定義 $D+1$ 個特徵 $f_d(W_{i,j})$，d 介於 0 到 D 之間，每個語言特徵皆為一個將候選詞序列 $W_{i,j}$ 對應到實數值之函數。$f_0(W_{i,j})$ 為基礎語言特徵，本論文定義為三連詞語言模型與聲學模型乘積之對數值，而其餘的語言特徵則可定義為候選詞序列 $W_{i,j}$ 中，各 N 連詞出現的次數；本論文使用單連詞

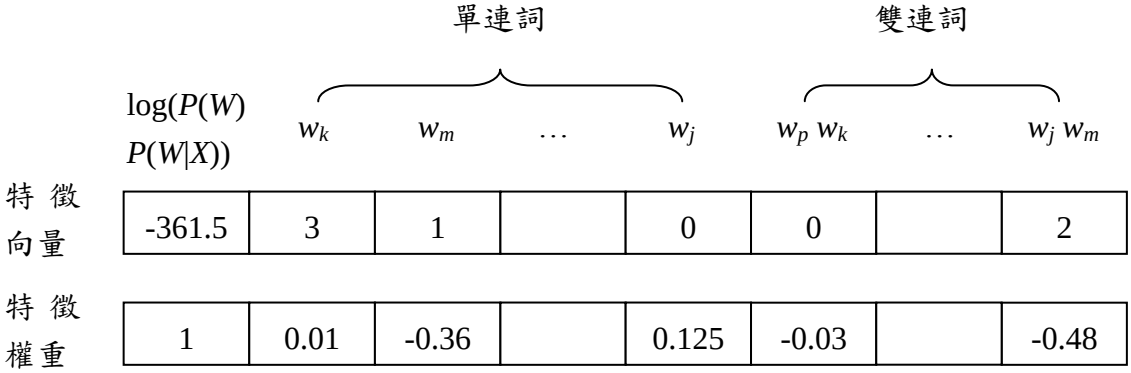

圖 3- 1 特徵向量與特徵權重向量

(Word Unigram)與雙連詞(Word Bigram)出現的次數做為其餘的語言特徵。

(4) 每一語言特徵定義都有其對應的權重參數值，為一 $D+1$維的參數向量 $\boldsymbol{\lambda}=[\lambda_0,...,\lambda_D]$，每一語言特徵及其對應的權重參數值如圖 3-1 所示。候選詞序列 $W_{i,j}$ 的排序分數則定義為特徵權重向量$\boldsymbol{\lambda}$與特徵向量$\mathbf{f}(W_{i,j})$之內積：

$$Score(W_{i,j},\boldsymbol{\lambda})=\boldsymbol{\lambda}\cdot\mathbf{f}(W_{i,j})=\sum_{d=0}^{D}\lambda_d f_d(W_{i,j}) \tag{3-1}$$

則排序分數最高的候選詞序列W_i^*即為重新排序結果：

$$W_i^*=\arg\max_{1\le j\le M} Score(W_{i,j},\boldsymbol{\lambda}) \tag{3-2}$$

鑑別式語言模型的訓練目標，即是求得最佳權重參數解，能使排序分數最高的候選詞序列W_i^*與候選詞序列集合$\mathrm{GEN}(x_i)$中的最低錯誤率詞序列W_i^R相等。而在測試階段，則是利用訓練階段時求得的最佳權重向量，使用式(3-1)的評分機制，從測試語料的候選詞序列集合$\mathrm{GEN}(y_k)$中找出得分最高之詞序列W_k^*，將其作為重新排序後的輸出結果。

接著介紹幾種常見的鑑別式語言模型。

Initialize all parameters in the model i.e. $\lambda_0 = 1$ and $\lambda_d = 0$ for $d = 1...D$
For $t = 1...T$. where T is the total number of iterations
  For each training sample $(x_i, W_i^R), i = 1...L$
    Use current modle $\boldsymbol{\lambda}$ to choose W_i^* from $\mathrm{GEN}(x_i)$
    For $d = 1...D$
      $\lambda_d^i = \lambda_d^i + \eta\left(f_d(W_i^R) - f_d(W_i^*)\right)$, where η is the size of the learning step.

圖 3- 2 感知器演算法[Zhou *et al*. 2006]

3.2 一般鑑別式語言模型

3.2.1 感知器演算法(Perceptron)

感知器[Rosenblatt 1958]最初是被應用在人工類神經網路(Artificial Neural Network)領域，它是一種二元分類器，把矩陣上的輸入**x**(實數值向量)映射到輸出值 $f(\mathbf{x})$ 上(二元數)，如式(3-3)所示：

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{w}\cdot\mathbf{x} + b > 0 \\ 0 & \text{else} \end{cases} \tag{3-3}$$

其中**w**實數權重向量，b 為一常數，表偏移量。此方法可視為一種最簡單形式的前饋式(Feed-Forward)人工神經網路。感知器演算法亦可以視為條件式隨機域(Conditional Random Fields, CRFs)的一種變形[Lafferty *et al*. 2001]，是一種用來最佳化排序減損函數(Loss Function)的增量訓練程序(Incremental Training Procedure)，1999 年時被提出投票與平俊形式的感知器演算法[Freund and Schapire 1999]，於 2002 年被應用在自然語言處理領域[Collins 2002]，並於 2005 年被應用在語言模型調適[Gao *et al*. 2005]上。

感知器演算法的排序減損函數的觀念為最小平方誤差(Least Squared Error, LSE)，如式(3-4)所示：

$$F_{Perc}(\boldsymbol{\lambda}) = \frac{1}{2}\sum_{i=1}^{L}\left(Score\left(W_i^R, \boldsymbol{\lambda}\right) - Score\left(W_i^*, \boldsymbol{\lambda}\right)\right)^2 \tag{3-4}$$

其中 W_i^* 為 x_i 所有候選詞序列中最接近 W_i^R 的一條詞序列。則最佳權重向量 $\boldsymbol{\lambda}^*$ 即為滿足 $\boldsymbol{\lambda}^* = \arg\min_{\lambda} F_{Perc}(\boldsymbol{\lambda})$ 的權重向量，即最小化所有訓練語句之最高排序分數候選詞序列 W_i^* 與最低錯誤率詞序列 W_i^R 排序分數的平方誤差總和。為了求得 $\boldsymbol{\lambda}^*$，我們可採用梯度下降法(Gradient Descent Method)將式(3-4)對每一維特徵權重參數 λ_d 作偏微分，以求得每一維特徵權重參數的更新量，如式(3-5)所示。

$$\begin{aligned}
\frac{\partial F_{Perc}(\lambda)}{\partial \lambda_d} &= \frac{\partial \frac{1}{2}\sum_{i=1}^{L}\left(Score\left(W_i^R, \boldsymbol{\lambda}\right) - Score\left(W_i^*, \boldsymbol{\lambda}\right)\right)^2}{\partial \lambda_d} \\
&= \frac{1}{2}\sum_{i=1}^{L}\left[\begin{array}{l}\left(Score\left(W_i^R, \boldsymbol{\lambda}\right) - Score\left(W_i^*, \boldsymbol{\lambda}\right)\right)\left(f_d\left(W_i^R\right) - f_d\left(W_i^*\right)\right) \\ + \left(f_d\left(W_i^R\right) - f_d\left(W_i^*\right)\right)\left(Score\left(W_i^R, \boldsymbol{\lambda}\right) - Score\left(W_i^*, \boldsymbol{\lambda}\right)\right)\end{array}\right] \\
&= \frac{1}{2}\sum_{i=1}^{L} 2\left[\left(Score\left(W_i^R, \boldsymbol{\lambda}\right) - Score\left(W_i^*, \boldsymbol{\lambda}\right)\right)\left(f_d\left(W_i^R\right) - f_d\left(W_i^*\right)\right)\right] \\
&= \sum_{i=1}^{L}\left[\left(Score\left(W_i^R, \boldsymbol{\lambda}\right) - Score\left(W_i^*, \boldsymbol{\lambda}\right)\right)\left(f_d\left(W_i^R\right) - f_d\left(W_i^*\right)\right)\right]
\end{aligned} \tag{3-5}$$

然而，由於 $F_{Perc}(\boldsymbol{\lambda})$ 可能存在許多局部最佳解(Local Minimum Solutions)，導致無法保證梯度下降法可求得全域最佳解(Global Minimum Solution)。因此，感知器演算法採取隨機近似法(Stochastic Approximation) [Gao *et al*. 2005]，相對於梯度下降法同時對所有訓練語句計算權重更新量，隨機近似法使用增量式(Incremental)的方法計算更新量，也就是將式(3-4)對每一訓練語句的每一維特徵權重參數 λ_d 作偏微分，以求得每一訓練語句 x_i 對每一維特徵的權重調整量：

$$\left(Score\left(W_i^R, \boldsymbol{\lambda}\right) - Score\left(W_i^*, \boldsymbol{\lambda}\right)\right)\left(f_d\left(W_i^R\right) - f_d\left(W_i^*\right)\right) \tag{3-6}$$

隨機近似法於是可視為最佳化個別訓練語句的排序減損函數：

$$F_{Perc}(i,\boldsymbol{\lambda}) = \frac{1}{2}\left(Score\left(W_i^R,\boldsymbol{\lambda}\right) - Score\left(W_i^*,\boldsymbol{\lambda}\right)\right)^2 \tag{3-7}$$

而感知器演算法對每一訓練語句的每一維特徵權重更新式可表示成：

$$\hat{\lambda}_d^i = \lambda_d^i - \eta \cdot \left(Score\left(W_i^R,\boldsymbol{\lambda}\right) - Score\left(W_i^*,\boldsymbol{\lambda}\right)\right)\left(f_d\left(W_i^R\right) - f_d\left(W_i^*\right)\right) \tag{3-8}$$

其中 η 為學習步調常數。關於權重調整量亦有學者提出省略 $\left(Score\left(W_i^R,\boldsymbol{\lambda}\right) - Score\left(W_i^*,\boldsymbol{\lambda}\right)\right)$ 項直接計算特徵的差值 $\left(f_d\left(W_i^R\right) - f_d\left(W_i^*\right)\right)$ 來更新權重[Roark *et al*. 2007]，由於 $\left(Score\left(W_i^R,\boldsymbol{\lambda}\right) - Score\left(W_i^*,\boldsymbol{\lambda}\right)\right)$ 項恆負，因此權重更新式則表示成 $\lambda_d^i = \lambda_d^i + \left(f_d\left(W_i^R\right) - f_d\left(W_i^*\right)\right)$，本論文使用此方式更新特徵權重。演算法如圖 3-2 所示。

另外有學者提出以每一訓練語句在每一次遞迴訓練後各自的特徵權重參數之平均值，當成最後的特徵權重參數[Freund and Schapire 1999]：

$$\left(\lambda_d\right)_{avg\_global} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{L}\left(\left(\lambda_d^{t,i}\right)_{Local}\right)}{T * L} \tag{3-9}$$

此方法稱之為平均全域特徵權重感知器演算法(Averaged Perceptron Algorithm)，本論文使用式(3-9)來表示最後的語言模型特徵權重參數。

接著為了證明感知器演算法會在有限訓練次數內收斂，我們需先定義：

定義 1：讓 $\mathrm{GEN}'(x_i) = \mathrm{GEN}(x_i) - \{W_i^R\}$，當存在某向量 $\mathbf{U}$ 滿足 $\|\mathbf{U}\| = 1$ 時，我們可以說一訓練語句 (x_i, W_i^R) 可用一邊界(margin) $\delta > 0$ 分離，也就是說：

$$\forall i, \forall W \in GEN'(x_i),\ \mathbf{U} \cdot f\left(W_i^R\right) - \mathbf{U} \cdot f(W) \geq \delta \tag{3-10}$$

Initialize all parameters in the model i.e. $\lambda_0 = 1$ and $\lambda_d = 0$ for $d = 1...D$
For $t = 1...T$. where T is the total number of iterations
  For each training sample $\left(x_i, W_i^R\right), i = 1...M$
    For $d = 1...D$
      $\lambda_d = \lambda_d + \eta \cdot \frac{\partial F_{GCLM}}{\partial \lambda_d}$, where η is the size of the learning step.

圖 3- 3 全域條件式對數線性模型

根據此定義，就可以得到以下定理：

定理 1：對任何可用邊界δ分離的訓練語句$\left(x_i, W_i^R\right)$，在任何訓練次數T下

$$N_e = \frac{R^2}{\delta^2} \tag{3-11}$$

其中R是一常數，滿足$\forall i, \forall W \in GEN'(x_i)\ \left\| f\left(W_i^R\right) - f(W) \right\| \le R$，而 N_e 代表錯誤次數。

此定理表示如果有一參數向量**U**可以使訓練集沒有錯誤，那最多$\frac{R^2}{\delta^2}$次訓練，演算法就可收斂到無訓練錯誤。此處的重點在於，錯誤的次數跟候選詞序列的個數間的關係是獨立的，因為在自動語音辨識上候選詞序列的個數是以語音訊號的長度為準，呈指數成長，因此這結果非常重要。

3.2.2 全域條件式對數線性模型(GCLM)

早先全域條件式對數線性模型被應用在自然語言處理領域[Ratnaparkhi *et al.* 1994] [Johnson *et al.* 1999]，於 2007 年第一次被應用在語音辨識領域，語言模型重新排序問題上[Roark *et al.* 2007]。全域條件式對數線性模型利用權重向量對語音訊號x_i的所有候選詞序列定義一個條件分佈：

$$p_{\lambda}(W \mid x_i) = \frac{1}{Z(x_i, \boldsymbol{\lambda})} \exp(Score(W, \boldsymbol{\lambda})) \tag{3-12}$$

其中 $Z(x_i, \boldsymbol{\lambda}) = \sum_{j=1}^{M} \exp(Score(W_{i,j}, \boldsymbol{\lambda}))$ 代表一正規化的常數。全域條件式對數線性模型希望給定一訓練語句，其最低錯誤率詞序列的對數條件機率能越大越好，因此根據上式的定義，全域條件式對數線性模型的目標函數可表示成：

$$F_{GCLM}(\boldsymbol{\lambda}) = \sum_{i=1}^{L} \log p_{\lambda}(W_i^R \mid x_i) = \sum_{i=1}^{L} \log \frac{\exp(Score(W_i^R, \boldsymbol{\lambda}))}{\sum_{j=1}^{M} \exp(Score(W_{i,j}, \boldsymbol{\lambda}))} \tag{3-13}$$

我們可將式(3-13)視為一種最低錯誤率詞序列與其他後選詞序列間的邏輯回歸(Logistic Regression)。另外，此目標函數跟其它鑑別式訓練研究非常相像，如它與最大化交互資訊估測(Maximum Mutual Information Estimation, MMIE) [Bahl *et al.* 1986]的目標函數有相同的表示式，而最小分類錯誤(Minimum Classification Error, MCE) [Juang and Katagiri 1992]則可看做此方法的延伸。

為了避免過度訓練(Overtraining)，我們可在式(3-13)加上一權重參數的零均值高斯事前機率(Zero-Mean Gaussian Prior) [Lafferty *et al.* 2001] [Johnson *et al.* 1999]：

$$F_{GCLM}(\boldsymbol{\lambda}) = \sum_{i=1}^{L} \log \frac{\exp(Score(W_i^R, \boldsymbol{\lambda}))}{\sum_{j=1}^{M} \exp(Score(W_{i,j}, \boldsymbol{\lambda}))} - \frac{\|\boldsymbol{\lambda}\|^2}{2\sigma^2} \tag{3-14}$$

σ 值控制對數機率項與高斯事前機率間的相互影響，可利用發展集語料(Development Set)估算。而最佳權重參數需符合 $\boldsymbol{\lambda}^* = \arg\max_{\lambda} F_{GCLM}(\boldsymbol{\lambda})$，因為 $F_{GCLM}(\boldsymbol{\lambda})$ 為一凸函數(Convex Function)，因此可求得其全域最佳解(Globally Optimal Solution)，為了求取最佳參數向量，利用梯度下降法將此目標函數對每

維權重參數 λ_d 偏微分後即可求得 $\boldsymbol{\lambda}^*$ 的權重調整量：

$$\begin{aligned}
\frac{\partial F_{GCLM}}{\partial \lambda_d} &= \frac{\partial \sum_{i=1}^{L} \log \frac{\exp\left(Score\left(W_i^R, \boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M} \exp\left(Score\left(W_{i.j}, \boldsymbol{\lambda}\right)\right)} - \frac{\|\boldsymbol{\lambda}\|^2}{2\sigma^2}}{\partial \lambda_d} \\
&= \frac{\partial \sum_{i=1}^{L} \left[Score\left(W_i^R, \boldsymbol{\lambda}\right) - \log \sum_{j=1}^{M} \exp\left(Score\left(W_{i.j}, \boldsymbol{\lambda}\right)\right) \right] - \frac{\|\boldsymbol{\lambda}\|^2}{2\sigma^2}}{\partial \lambda_d} \\
&= \sum_{i=1}^{L} \left[f\left(W_i^R\right) - \sum_{k=1}^{M} \frac{\exp\left(Score\left(W_{i,k}, \boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M} \exp\left(Score\left(W_{i.j}, \boldsymbol{\lambda}\right)\right)} \frac{\partial Score\left(W_{i,k}, \boldsymbol{\lambda}\right)}{\partial \lambda_d} \right] - \frac{2\lambda_d}{2\sigma^2} \\
&= \sum_{i=1}^{L} \left[f\left(W_i^R\right) - \sum_{k=1}^{M} \frac{\exp\left(Score\left(W_{i,k}, \boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M} \exp\left(Score\left(W_{i.j}, \boldsymbol{\lambda}\right)\right)} f_d\left(W_{i,k}\right) \right] - \frac{\lambda_d}{\sigma^2}
\end{aligned} \tag{3-15}$$

於是對每一維特徵分別更新其權重，以求得最佳權重參數：

$$\lambda_d = \lambda_d + \eta \cdot \sum_{i=1}^{L} \left[f\left(W_i^R\right) - \sum_{k=1}^{M} \frac{\exp\left(Score\left(W_{i,k}, \boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M} \exp\left(Score\left(W_{i,j}, \boldsymbol{\lambda}\right)\right)} f_d\left(W_{i,k}\right) \right] - \frac{\lambda_d}{\sigma^2} \tag{3-16}$$

全域條件式對數線性模型的演算法如圖 3-3 所示。

3.3 考慮樣本權重(Sample Weight)之鑑別式語言模型

3.3.1 權重式全域條件式對數線性模型(WGCLM)

權重式全域條件式對數線性模型[Oba *et al.* 2010]是將全域條件式對數線性模型加入樣本權重要因素作延伸。因此它的目標函數定義為：

$$F_{WGCLM}(\boldsymbol{\lambda}) = \sum_{i=1}^{L} \log \frac{\exp\left(Score\left(W_i^R, \boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M} \omega_{i,W_{i,j}} \exp\left(Score\left(W_{i,j}, \boldsymbol{\lambda}\right)\right)} \tag{3-17}$$

可發現與全域條件式對數線性模型不同點在於，每一候選詞序列$W_{i,j}$因樣本權重$\omega_{i,W_{i,j}}$而有不同的重要性，另外，與全域條件式對數線性模型相同，在參數調整過程中為了避免過度訓練，可加入零均值高斯事前機率：

$$F_{WGCLM}(\boldsymbol{\lambda}) = \sum_{i=1}^{L} \log \frac{\exp\left(Score\left(W_i^R, \boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M} \omega_{i,W_{i,j}} \exp\left(Score\left(W_{i,j}, \boldsymbol{\lambda}\right)\right)} - \frac{\|\boldsymbol{\lambda}\|^2}{2\sigma^2} \tag{3-18}$$

而最佳權重參數需符合$\boldsymbol{\lambda}^* = \arg\max_{\lambda} F_{WGCLM}(\boldsymbol{\lambda})$。從式(3-19)可看出權重式全域條件式對數線性模型的目標函數也是一凸函數，因此也可求得其全域最佳解，與全域條件式對數線性模型相同，利用梯度下降法將此目標函數對每維權重參數λ_d偏微分後即可求得$\boldsymbol{\lambda}^*$的權重調整量：

$$\frac{\partial F_{WGCLM}}{\partial \lambda_d} = \frac{\partial \sum_{i=1}^{L} \log \frac{\exp\left(Score\left(W_i^R, \boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M} \omega_{i,W_{i,j}} \exp\left(Score\left(W_{i,j}, \boldsymbol{\lambda}\right)\right)} - \frac{\|\boldsymbol{\lambda}\|^2}{2\sigma^2}}{\partial \lambda_d}$$

$$= \frac{\partial \sum_{i=1}^{L} \left[Score\left(W_i^R, \boldsymbol{\lambda}\right) - \log \sum_{j=1}^{M} \omega_{i,W_{i,j}} \exp\left(Score\left(W_{i,j}, \boldsymbol{\lambda}\right)\right) \right] - \frac{\|\boldsymbol{\lambda}\|^2}{2\sigma^2}}{\partial \lambda_d}$$

$$= \sum_{i=1}^{L} \left[f\left(W_i^R\right) - \sum_{k=1}^{M} \frac{\omega_{i,W_{i,k}} \exp\left(Score\left(W_{i,k}, \boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M} \omega_{i,W_{i,j}} \exp\left(Score\left(W_{i,j}, \boldsymbol{\lambda}\right)\right)} \frac{\partial Score\left(W_{i,k}, \boldsymbol{\lambda}\right)}{\partial \lambda_d} \right] - \frac{2\lambda_d}{2\sigma^2}$$

$$= \sum_{i=1}^{L} \left[f\left(W_i^R\right) - \sum_{k=1}^{M} \frac{\omega_{i,W_{i,k}} \exp\left(Score\left(W_{i,k}, \boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M} \omega_{i,W_{i,j}} \exp\left(Score\left(W_{i,j}, \boldsymbol{\lambda}\right)\right)} f_d\left(W_{i,k}\right) \right] - \frac{\lambda_d}{\sigma^2} \tag{3-19}$$

因此可對每一維特徵分別更新其權重，以求得最佳權重參數$\boldsymbol{\lambda}^*$：

$$\lambda_d = \lambda_d + \eta \cdot \sum_{i=1}^{L} \left[f\left(W_i^R\right) - \sum_{k=1}^{M} \frac{\omega_{i,W_{i,k}} \exp\left(Score\left(W_{i,k}, \boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M} \omega_{i,W_{i,j}} \exp\left(Score\left(W_{i,j}, \boldsymbol{\lambda}\right)\right)} f_d\left(W_{i,k}\right) \right] - \frac{\lambda_d}{\sigma^2} \tag{3-20}$$

3.3.2 最小化錯誤率訓練(MERT)

最小化錯誤率訓練一開始被提出運用在機器翻譯領域[Och 2003]，近年才被使用在語音辨識領域[Kobayashi *et al.* 2008]。其排序減損函數定義為：

$$F_{MERT}\left(\boldsymbol{\lambda}\right) = \sum_{i=1}^{L} \sum_{k=1}^{M} \frac{\omega_{i,W_{i,k}} \exp\left(Score\left(W_{i,k}, \boldsymbol{\lambda}\right) - Score\left(W_i^R, \boldsymbol{\lambda}\right)\right)^{\beta}}{\sum_{j=1}^{M} \exp\left(Score\left(W_{i,j}, \boldsymbol{\lambda}\right) - Score\left(W_i^R, \boldsymbol{\lambda}\right)\right)^{\beta}} \tag{3-21}$$

若式(3-21)中的變數$\omega_{i,W_{i,k}}$設為候選詞序列$W_{i,k}$的錯誤率時，則此排序減損函數可

視為前 M 條最佳辨識候選詞序列錯誤率的期望值，因此最小化錯誤率訓練的訓練目標即為最小化前 M 條最佳辨識候選詞序列錯誤率期望值。由於最小化錯誤率訓練存在許多局部最佳解，因此使用 β 項來平滑化(Smooth)此排序減損函數，以減少局部最佳解的個數。經運算後，我們可以很技巧地將式(3-21)中的 $\exp\left(Score\left(W_i^R,\boldsymbol{\lambda}\right)\right)^\beta$ 項消去，使排序減損函數化簡為：

$$F_{MERT}\left(\boldsymbol{\lambda}\right)=\sum_{i=1}^{L}\sum_{k=1}^{M}\omega_{i,W_{i,k}}\frac{\exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)^\beta}{\sum_{j=1}^{M}\exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)^\beta} \tag{3-22}$$

另外，從式(3-22)可看出，最小化錯誤率訓練在訓練語言模型時不僅考慮降低訓練語句的候選詞序列中錯誤率詞最低者的錯誤率，同時也考慮降低其它候選詞序列的錯誤率。實作時，最小化錯誤率訓練的最佳權重參數需符合 $\boldsymbol{\lambda}^*=\arg\min_{\lambda}F_{MERT}\left(\boldsymbol{\lambda}\right)$，因此我們可對式(3-23)的每維權重參數 λ_d 偏微分而可求得 $\boldsymbol{\lambda}^*$ 的權重調整量：

$$\frac{\partial F_{MERT}}{\partial\lambda_d}=\frac{\partial\sum_{i=1}^{L}\sum_{k=1}^{M}\omega_{i,W_{i,k}}\frac{\exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)^\beta}{\sum_{j=1}^{M}\exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)^\beta}}{\partial\lambda_d}$$

$$=\sum_{i=1}^{L}\sum_{k=1}^{M}\omega_{i,W_{i,k}}\frac{\left[\begin{array}{l}\frac{\partial\sum_{j=1}^{M}\exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)^\beta}{\partial\lambda_d}\exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)^\beta\\-\frac{\partial\exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)^\beta}{\partial\lambda_d}\sum_{j=1}^{M}\exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)^\beta\end{array}\right]}{\left(\sum_{j'=1}^{M}\exp\left(Score\left(W_{i,j'},\boldsymbol{\lambda}\right)\right)^\beta\right)^2}$$

$$= \sum_{i=1}^{L}\sum_{k=1}^{M} \omega_{i,W_{i,k}} \frac{\left[\begin{array}{l}\left[\sum_{j=1}^{M} \exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)^{\beta}\cdot\beta\cdot f_d\left(W_{i,j}\right)\right]\exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)^{\beta} \\ -\left[\exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)^{\beta}\cdot\beta\cdot f_d\left(W_{i,k}\right)\right]\sum_{j=1}^{M}\exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)^{\beta}\end{array}\right]}{\left(\sum_{j'=1}^{M}\exp\left(Score\left(W_{i,j'},\boldsymbol{\lambda}\right)\right)^{\beta}\right)^{2}}$$

$$= \sum_{i=1}^{L}\sum_{k=1}^{M} \beta\cdot\omega_{i,W_{i,k}}\cdot\exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)^{\beta}\cdot\frac{\left[\begin{array}{l}\sum_{j=1}^{M}\exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)^{\beta} f_d\left(W_{i,j}\right) \\ - f_d\left(W_{i,k}\right)\sum_{j=1}^{M}\exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)^{\beta}\end{array}\right]}{\left(\sum_{j'=1}^{M}\exp\left(Score\left(W_{i,j'},\boldsymbol{\lambda}\right)\right)^{\beta}\right)^{2}}$$

$$= \sum_{i=1}^{L}\sum_{k=1}^{M} \omega_{i,W_{i,k}}\cdot\beta\cdot\exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)^{\beta}\cdot\frac{\sum_{j=1}^{M}\exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)^{\beta}\left(f_d\left(W_{i,j}\right)-f_d\left(W_{i,k}\right)\right)}{\left(\sum_{j'=1}^{M}\exp\left(Score\left(W_{i,j'},\boldsymbol{\lambda}\right)\right)^{\beta}\right)^{2}}$$

(3-23)

因此與前述的鑑別式語言模型相同，我們可對每一維特徵分別更新其權重，以求得最佳權重參數：

$$\lambda_d = \lambda_d - \eta\cdot\sum_{i=1}^{L}\sum_{k=1}^{M}\omega_{i,W_{i,k}}\cdot\beta\cdot\exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)^{\beta}\cdot\frac{\sum_{j=1}^{M}\exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)^{\beta}\left(f_d\left(W_{i,j}\right)-f_d\left(W_{i,k}\right)\right)}{\left(\sum_{j'=1}^{M}\exp\left(Score\left(W_{i,j'},\boldsymbol{\lambda}\right)\right)^{\beta}\right)^{2}}$$

(3-24)

表 3- 1 鑑別式語言模型比較

	是否考慮樣本權重	是否考慮 W_i^R	一般化能力	是否有全域最佳解	訓練速度
Perceptron	否	是	差	否	快
GCLM	否	是	略佳	是	慢
WGCLM	是	是	略佳	是	慢
MERT	是	否	佳	否	極慢

3.4 鑑別式語言模型之比較

本節對上述所介紹的鑑別式語言模型之排序減損(目標)函數進行討論與比較。這些鑑別式語言模型分別為沒有考慮樣本權重的感知器演算法和全域條件式對數線性模型，與考慮樣本權重的權重式全域條件式對數線性模型和最小化錯誤率訓練。各鑑別式語言模型的排序減損(目標)函數都有其不同的訓練目標，感知器演算法的排序減損函數是使用最小平方誤差的精神，希望排序分數最高的候選詞序列與最低錯誤率詞序列間的誤差平方越小越好；而全域條件式對數線性模型的目標函數則是希望最低錯誤率詞序列的條件機率越高越好；權重式全域條件式對數線性模型與全域條件式對數線性模型的差別即在分母項的樣本權重值 $\omega_{i,W_{i,j}}$，對每條候選詞序列加入其錯誤率或排序作為其樣本權重，以考慮每條候選詞序列對訓練會有不同的影響力，即錯誤率越高或排序越後者，其影響力越小；最小化錯誤率則希望所有候選詞序列的錯誤率或排序的期望值越小越好，由於

$$\sum_{k=1}^{M}\frac{\exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)^{\beta}}{\sum_{j=1}^{M}\exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)^{\beta}}=1 \tag{3-25}$$

因此最小化錯誤率訓練的排序減損(目標)函數其實就是找出一組權重向量可以使錯誤率或排序越小越好，即為最小化錯誤率訓練的精神。

再者，從上述四種鑑別式語言模型的排序減損(目標)函數也可看出，感知器演算法只考慮目前排序分數最高的候選詞序列與最低錯誤率詞序列間的關係；而全域條件式對數線性模型與權重式全域條件式對數線性模型考慮到最低錯誤率詞序列與其他候選詞序列的關係；最小化錯誤率訓練則是考慮所有候選詞序列的錯誤率，而不是只考慮最低錯誤率詞序列的影響。感知器演算法因只考慮目前排序分數最高的候選詞序列與最低錯誤率詞序列越接近越好，沒有考慮到其餘候選詞序列的影響，其一般化(Generalization)的能力可預期的會不若其它三個鑑別式語言模型來的好，而容易導致過度訓練的情況。而最小化錯誤率訓練因考慮所有候選詞序列，因此一般化能力與另外三種鑑別式語言模型相比會較佳，較不受訓練語料影響其在測試語料的表現。

另外，就訓練速度來說，感知器演算法只考慮排序分數最高的候選詞序列與最低錯誤率詞序列間的關係，因此訓練速度最快；最小化錯誤率訓練需考慮所有候選詞序列，因此訓練速度極慢；而全域條件式對數線性模型與權重式全域條件式對數線性模型的訓練速度介於兩者之間。表 3-1 為四種鑑別式語言模型的比較。

第 4 章 語句相關之鑑別式語言模型

本論文嘗試改進傳統鑑別式語言模型在測試過程中，所有測試語句皆使用相同語言模型權重向量的缺點；希望針對每一句測試語句，以線性組合的方式結合由不同訓練語料所訓練的權重參數向量，以期新的語言特徵權重參數向量能更加符合測試語句的特性，其訓練流程如圖 4-1 所示。首先，所有訓練語句的正確轉寫語句先以單連詞向量(Unigram Word Vector)表示並進行分群(Clustering)。我們使用的分群方法為 K 平均演算法(K-means)，假設所有語言模型訓練語句可分為 P 群；接著對 P 群中每一群訓練語句分別訓練其最佳的語言特徵權重參數向量 $\boldsymbol{\lambda}_p$。在語音辨識階段，當有一測試語句 y_k 輸入時，我們可利用 y_k 本身的特性(可為相似度)，選取一組最符合 y_k 特性的特徵權重參數向量作為此測試語句的權重參數向量、或是利用 y_k 的特性(可為相似度或機率)產生 P 個組合係數，利用此組合係數線性結合 P 個語言特徵權重參數向量以產生一新的語言特徵權重參數向量。本論文提出三種產生新的語言特徵權重參數向量的技術：

(1)選取相似度最大權重法：此方法只選取與測試語句最相似的訓練語句群(即相似度最大)所訓練出的語言特徵權重參數向量作為此測試語句的語言特徵權重參數向量。本論文以餘弦值計算測試語句與訓練語句群的相似度，先對每一測試語句 y_k 定義其單連詞向量 $\mathbf{u}_k$，$\mathbf{u}_k$ 為結合 y_k 的前 M 條最佳辨識候選詞序列中每一候選詞序列 $W_{k,j}$ 的單連詞向量 $\mathbf{u}_{k,j}$ 而成：

$$\mathbf{u}_k = \sum_{j=1}^{M} \mathbf{u}_{k,j} \tag{4-1}$$

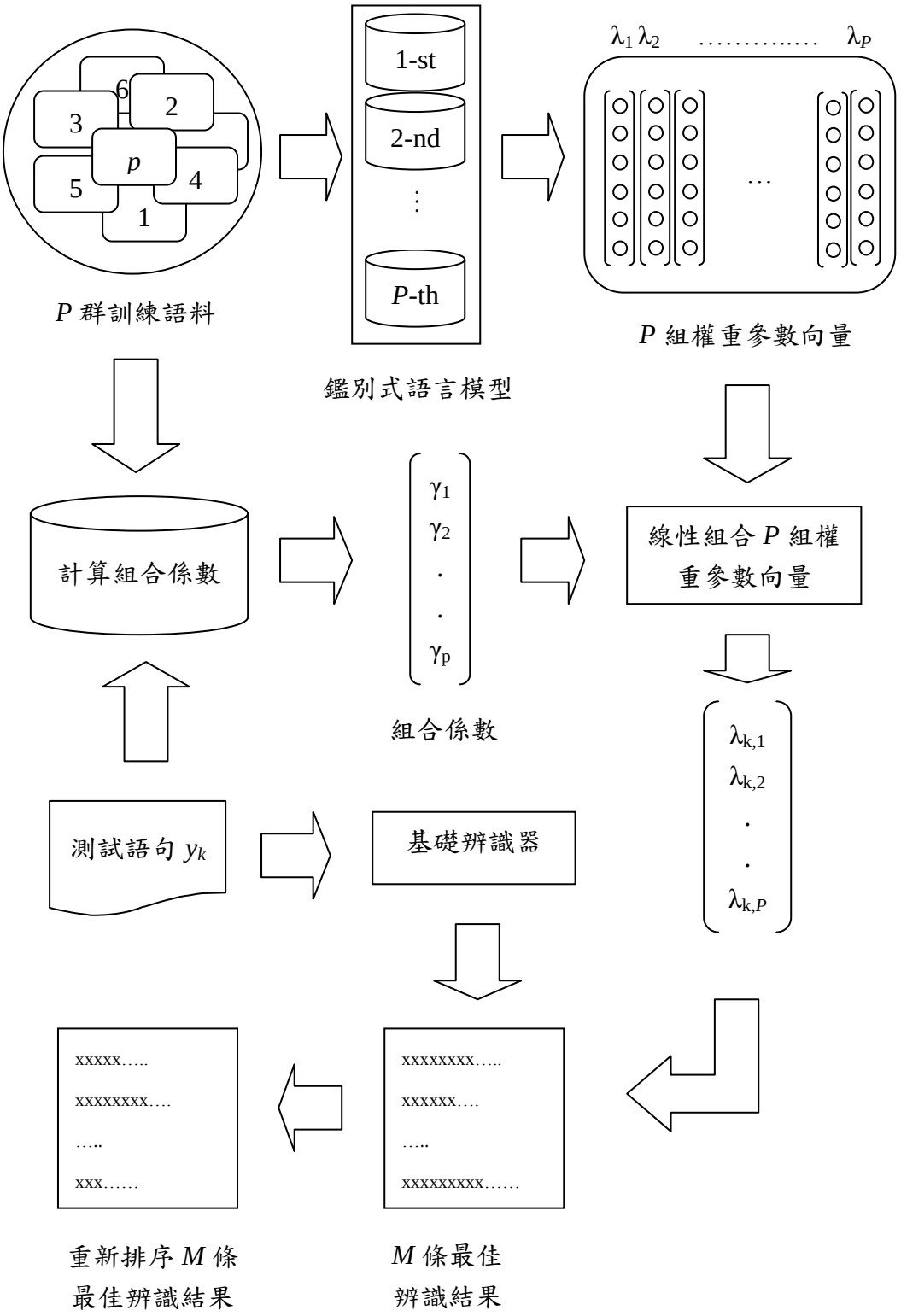

圖 4- 1 語句相關之鑑別式語言模型流程圖

而每一群訓練語句的單連詞向量 $\mathbf{v}_p$ 則是其中所有訓練語句的正確轉寫語句的單連詞向量的質心：

$$\mathbf{v}_p = \frac{\sum_{j=1}^{L_p} \mathbf{v}_{p,j}}{L_p} \tag{4-2}$$

其中 L_p 為 p 訓練語句群中包含的訓練語句句數，則兩單連詞向量的相似度即為兩者間的餘弦值：

$$\cos(\mathbf{u}_k, \mathbf{v}_p) = \frac{\mathbf{u}_k \cdot \mathbf{v}_p}{\sqrt{\mathbf{u}_k^{\,2}}\sqrt{\mathbf{v}_p^{\,2}}} \tag{4-3}$$

因此，對測試語句 y_k 來說，其新的語言特徵權重參數向量 λ_k 即為擁有最大相似度的那群訓練語句所訓練出語言特徵權重參數向量：

$$\boldsymbol{\lambda}_k = \arg\max_{\boldsymbol{\lambda}_p} \cos(\mathbf{u}_k, \mathbf{v}_p) \tag{4-4}$$

(2) 相似度線性組合法：此方法經由計算測試語句與 P 群訓練語句間的相似度而得到組合係數(若測試語句與某一群訓練語句有較高相似度，則此群訓練語句所訓練出語言特徵權重參數向量會有較高的組合係數，亦即在語音辨識會有較大的貢獻)。

與選取相似度最大權重法相同，測試語句與訓練語句群的相似度以餘弦值計算，因此，測試語句 y_k 與 p 訓練語句群的組合係數 $\gamma_{k,p}$ 為：

$$\gamma_{k,p} = \frac{\cos(\mathbf{u}_k, \mathbf{v}_p)}{\sum_{c=1}^{P} \cos(\mathbf{u}_k, \mathbf{v}_c)} \tag{4-5}$$

最後，對測試語句 y_k 來說，其新的語言特徵權重參數向量可表示成為 P 個語言特徵權重參數向量的線性組合：

$$\boldsymbol{\lambda}_k = \sum_{p=1}^{P} \gamma_{k,p} \cdot \boldsymbol{\lambda}_p \tag{4-6}$$

即可求得測試語句 y_k 的語言特徵權重參數向量 $\boldsymbol{\lambda}_k$ 。

(3) 最大機率法：此方法的目標為希望利用最大機率法(Maximum Likelihood, ML)計算最佳組合係數，使測試語句利用此組合係數線性組合的新的語言特徵權重參數向量能有較高的機率，如式(4-7)所示：

$$\mathbf{v}_k^* = \arg\max_{\mathbf{v}_k} P(y_k \mid \mathbf{v}_k) \tag{4-7}$$

也就是我們希望找出最佳組合係數 $\mathbf{v}_k^*$ 能最大化條件機率 $P(y_k \mid \mathbf{v}_k)$ ，而 $\mathbf{v}_k^* = [\nu_{k,1}, \nu_{k,2}, ..., \nu_{k,p}]$ ；因此組合係數的算法則如式(4-8)所示：

$$\nu_{k,p} = \frac{\sum_{j=1}^{M} \exp(\boldsymbol{\lambda}_p \cdot \mathbf{f}(W_{k,j}))}{\sum_{p'} \sum_{j'=1}^{M} \exp(\boldsymbol{\lambda}_{p'} \cdot \mathbf{f}(W_{k',j'}))} \tag{4-8}$$

其中 $W_{k,j} \in \mathrm{GEN}(y_k)$ ；另外，由於特徵向量第 0 維為聲學模型與語言模型機率乘績對數值，與其它維特徵的數值範圍相差過大，而導致計算機率時容易被第 0 維牽制，因此本論文在計算時不考慮第 0 維特徵的影響。

最後，對測試語句 y_k 來說，其新的語言特徵權重參數向量即可表示成為 P 個語言特徵權重參數向量的線性組合：

$$\boldsymbol{\lambda}_k = \sum_{p=1}^{P} \nu_{k,p} \cdot \boldsymbol{\lambda}_p \qquad (4\text{-}9)$$

第 5 章 實驗結果與討論

5.1 實驗架構

在本章中將先介紹台師大之大詞彙連續語音辨識系統。接著說明本論文所使用的聲學模型訓練語料、語言模型訓練語料、調適語料及語音測試語料。

5.1.1 台師大大詞彙連續語音辨識系統

以下將分別介紹台師大大詞彙連續語音辨識系統採用的前端處理(Front-end Processing)、聲學模型、詞典建立(Lexicon Construction)、語言模型以及詞彙樹複製搜尋(Tree-copy Search)等部份[Chen *et al.*2004]。

(一)前端處理

本系統語音特徵抽取使用梅爾倒頻譜係數(MFCC)或是異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)加上最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)[Gopinath 1998; Saon *et al.*2000]兩種不同語音特徵參數。論文中主要使用異質性線性鑑別分析(HLDA)配合最大相似度線性轉換(MLLT)做為語音特徵，並使用變異數正規化法(Cepstral Mean and Variance Normalization, CMVN)加強語音特徵。

(二) 聲學模型

聲學模型部分，我們分別為聲母及韻母建立 INITIAL 與 FINAL 模型。基本的 INITIAL 模型為 22 種，FINAL 模型為 38 種，因為聲母會被右邊相連韻母影響其發音特性，所以我們將 INITIAL 模型細分為 112 種，即右相關聯模型(Right-Context-Dependent Model, RCD Model)，最後加上一個靜音(Silence)模型，

共有 151 個聲學模型。每個模型的狀態有 3 至 6 個不等，每個狀態皆為高斯混合分布，其中使用的高斯分布個數分別為 1 至 128 個不等。此外，這些聲母和韻母共組合成 403 個不同的基本音節(Base Syllables)。本論文的聲學模型首先經由最大化相似度估測(Maximum Likelihood Estimation, MLE)訓練而得，在透過最小化音素錯誤訓練(Minimum Phone Error, MPE)以期望獲得最佳化聲學模型參數。

(三)詞典建立

在中文裡約有 7,000 個單字詞，可由這些單字詞合併產生新詞，本系統根據字詞在語料中的統計特性，以自動化的方式產生新的複合詞(Compound Words)。對於語料中任意相鄰的兩個詞 $w_i w_j$ ，分別計算它們的前向二連(Forward Bigram)機率 $P_f(w_j \mid w_i)$ 與後向二連(Backward bigram)機率 $P_b(w_i \mid w_j)$ ，再以前後向二連的機率幾何平均 $FB(w_i, w_j) = \sqrt{P_f(w_j \mid w_i) P_b(w_i \mid w_j)}$ ，作為詞 w_i 與詞 w_j 是否合併的依據。文字語料先經由一個含有一至四字詞約六萬六千個詞的原始詞典斷詞，再利用上述的計算方法，經過數次的迭代以及不同的門檻值(Thresholds)設定，產生約五千餘個二至十字詞的複合詞。最後將這五千餘個新詞加入原始詞典中，得到一個含有約七萬兩千個詞的新詞典。

(四)詞彙樹複製搜尋

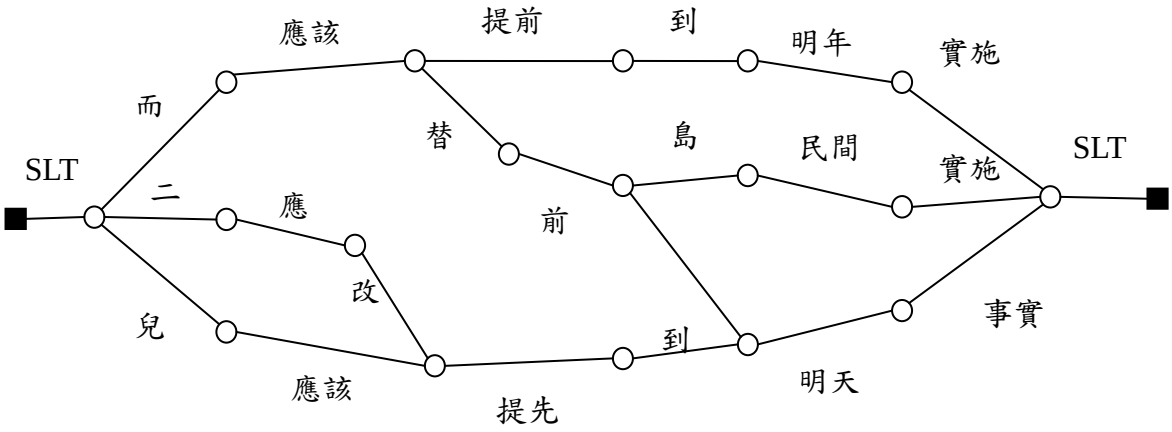

圖 5- 1 詞圖範例

本系統的大詞彙連續語音辨識方法是採用由左至右(Left-to-right)、音框同步(Frame-synchronous)的詞彙樹複製搜尋方式[Aubert 2002]。在詞彙樹中每個分枝(Arc)代表一個 INITIAL 或 FINAL 的隱藏式馬可夫模型，由根節點(Root)到任一個葉節點(Leaf)的路徑代表一個詞或一些發音相同的詞，路徑上的分枝就是代表這個詞或這些詞會使用到的隱藏式馬可夫模型。在每個音框中，若有不完全路徑(Partial Path)已到達葉節點時，代表一個完整詞已可被產生；同時，不同棵詞彙樹複製間已抵達葉節點的不完全路徑，若具有相同的語言模型歷史詞序列，則會進行再結合(Recombination)，保留最大分數者，並以它們的歷史詞序列為標註，產生一棵新的詞彙樹複製，或加入到一棵已存在且具有相同歷史詞序列的詞彙樹複製中。

此外，每個音框會記錄存活的詞彙樹複製葉節點中分數較高者的相關資訊(這些葉節點本身代表著可能的候選詞)，諸如它們的語言模型歷史詞序列、候選詞所對應的開始與結束的音框以及搜尋時聲學模型解碼的分數，然後再依此資訊建立詞圖(Word Graph)，並在詞圖上使用更高階的語言模型，重新進行一次詞圖動態規劃搜尋(Word Graph Rescoring)[Ortmanns *et al.* 1997]，找出最佳的辨識詞序列。在本系統中，詞彙樹複製搜尋階段是使用二連詞語言模型，而在詞圖搜尋階

表 5- 1 實驗語料之統計資訊

	訓練語料	發展語料	測試語料
語料時間長度	23小時	1.5小時	1.5小時
語料句數	30,600句	1,998句	1,997句

段是使用三連詞語言模型。

(五)詞圖搜尋

經過詞彙樹搜尋後，可以產生詞圖(Word Graph or Lattice)，詞圖範例如圖 5-1 所示。每一個分支代表經過裁減所保留的詞段(Arc)，且每一個詞段會記錄其聲學分數。接著我們針對每一個詞段進行維特比搜尋，記錄與其相連(結束時間與 20 目前詞段開始時間相同)且最可能(維特比分數最高)的詞段。詞圖所保留下來的詞段，在聲學上大多是混淆的，所以需要透過語言模型的輔助。由於詞圖已經簡化，搜尋時，可以使用較複雜的語言模型，例如三連詞模型。而完成整個詞圖的重新計分(Rescoring)之後，同樣地從語音結尾的詞段後返最可能詞序列當作最後辨識結果。

5.1.2 實驗語料

(一)實驗語料

實驗語料皆來自公視新聞(Mandarian Across Taiwan─Broadcast News, MATBN)。公視新聞語料是2001年至2003年間由中央研究院資訊所口語小組(SLG])與公共電視台(PTS)合作錄製的，包含內場與外場兩個部分，內場為主播語料(Studio Anchors)，外場則有採訪記者(Field Reporters)語料與受訪者(Interviewees)語料。由於內場主播語料大部分為同一個主播所錄製的因素，為了避免語料的偏差性讓實驗偏向語者相依(Speaker Dependent)，故不採用內場主播

語料；又發現外場的受訪者語料，包含了過多的語助詞，因此實驗語料皆取自外場採訪記者語料。訓練語料取自公視新聞2001年至2002年外場採訪記者語料，共30,600句(約23小時)；測試語料與訓練語料屬於同時期語料，亦取自公視新聞外場採訪記者語料，共1,997句(約1.5小時)；發展語料亦屬於同時期語料，取自公視新聞外場採訪記者語料，共1,998句(約1.5小時)，如表5-1所示。

(二)背景語言模型

對於實驗所使用的背景三連語言模型(Trigram Language Model)，其訓練語料是來自中央通訊社(Central News Agency, CNA)2001年至2002年的文字新聞語料，包含了約一億五千萬個中文字，經斷詞後約有八千萬詞[CNA News]。我們採用SRI Language Modeling Toolkit [SRI]訓練實驗所需要的三連語言模型。

(三)聲學模型語料

聲學模型訓練語料為公視新聞 2001 年至 2002 年外場採訪記者語料，共30,632 句(約 23 小時)，其中包含實驗訓練語料 30,600 句。

5.1.3 語言模型評估

為了評估(Evaluate)語言模型是否能在辨識過程中順利引導辨識器，使其選擇最接近正確轉寫語句的候選詞序列作為辨識結果，以及衡量語言模型實際運用效益多寡，需進行語言模型評估。語言模型評估辨識結果的主要標準為錯誤率(Error Rate)。

錯誤率是將一段語音之正確參照轉寫與語音辨識結果作字串比對所得到之數據。依比對之單位(Unit)不同，可分為字錯誤率(Character Error Rate, CER)與詞錯誤率(Word Error Rate, WER)。

表 5- 2 最大事後機率法實驗結果(CER(%))

α	測試語料字錯誤率	絕對提昇率	相對提昇率
基礎辨識率	16.39	-	-
0.1	15.47	0.92	5.64
0.2	15.10	1.29	7.87
0.3	14.93	1.46	8.90
0.4	14.78	1.61	9.82
0.5	14.70	1.69	10.31
0.6	14.55	1.84	11.23
0.7	14.53	1.86	11.36
0.8	14.51	1.88	11.49
0.9	14.60	1.79	10.90

字串比對可透過動態規劃(Dynamic Programming)方法進行。若以字為對齊(Align)單位來看，若正確字串中的某個字在辨識結果中被取代為錯誤的另一字，稱為替代(Substitution)；若正確字串中存在的某一字並不存在於辨識結果中，而是被移除了，則稱為刪除(Deletion)；與刪除相反，若辨識結果中多出了並不存在於正確字串中的字，則稱為插入(Insertion)。

錯誤率之計算方式如下：

$$\text{Error rate} = \frac{S + D + I}{N} * 100\% \tag{5-1}$$

其中 S 代表替代數，D 代表刪除數，I 代表插入數，而 N 則為正確轉寫語句長度。本論文以字錯誤率作為評估標準。

5.2 基礎實驗結果

本論文實驗所用的基礎辨識器在訓練語料的字錯誤率為 11.26%，在測試語料的字錯誤率則為 16.39%。

5.3 最大化事後機率法(MAP)實驗結果

本節列出最大事後機率法(MAP)的實驗結果，以與傳統鑑別式語言模型及本論文提出的語句相關鑑別式語言模型作比較。本論文使用最大事後機率法中的模型插補法來結合背景語言模型 P_B 與訓練語料所訓練的調適語言模型 P_A，並對完整詞圖中所有候選詞序列作模型插補調適，插補方式如(5-2)所示：

$$P(w_i \mid h_k) = \alpha P_B(w_i \mid h_k) + (1-\alpha) P_A(w_i \mid h_k) \tag{5-2}$$

以 α 調整背景語言模型與調適語言模型間的關係。實驗結果如表 5-2 所示，當 $\alpha = 0.8$ 時可達到最低字錯誤率 14.51%，相對於基礎辨識率 16.39%有 1.88%的絕對提昇率、11.49%的相對提昇率。

5.4 鑑別式語言模型實驗結果

表 5- 3 鑑別式語言模型實驗結果(CER(%))

	訓練語料字錯誤率	絕對提昇率	相對提昇率	測試語料字錯誤率	絕對提昇率	相對提昇率
基礎辨識率	11.26	-	-	16.39	-	-
感知器演算法	6.02	5.25	46.58%	15.71	0.68	4.15
GCLM	9.90	1.36	12.10%	15.53	0.86	5.26
WGCLM (字錯誤率)	9.86	1.40	12.46%	15.44	0.95	5.77
WGCLM (排序)	9.87	1.39	12.33%	15.49	0.90	5.48
MERT (字錯誤率)	10.45	0.81	7.21%	15.33	1.06	6.47
MERT (排序)	10.54	0.73	6.45%	15.40	0.99	6.03

本節將比較並討論本論文介紹的幾種鑑別式語言模型的實驗結果,如表 5-3 所示。由實驗結果可看出,本論文介紹的幾種鑑別式演算法於測試語料的效能以最小化錯誤率訓練為最佳,接著依序為權重式全域條件式對數線性模型、全域條件式對數線性模型及感知器演算法。

從表 5-3 看出,感知器演算法雖在訓練語料達到最佳的效果,但在測試語料效果卻低於其他鑑別式語言模型。為了探究其原因,我們觀察了感知器演算法訓練語料與測試語料字錯誤率隨著訓練次數增加的變化趨勢,如圖 5-2 所示。由圖 5-2 可看出,感知器演算法測試語料字錯誤率在訓練語料達到最低字錯誤率的訓練次數前就已收斂,因此將訓練語料最佳參數權重使用在測試語料上,反而無法讓測試語料達到最佳的效果,而造成此情況的原因應為感知器演算法在訓練過程中容易發生過度訓練的情況,導致訓練後的最佳權重向量過於符合訓練語料,因而使用在測試語料之時,無法發揮對應的效果。

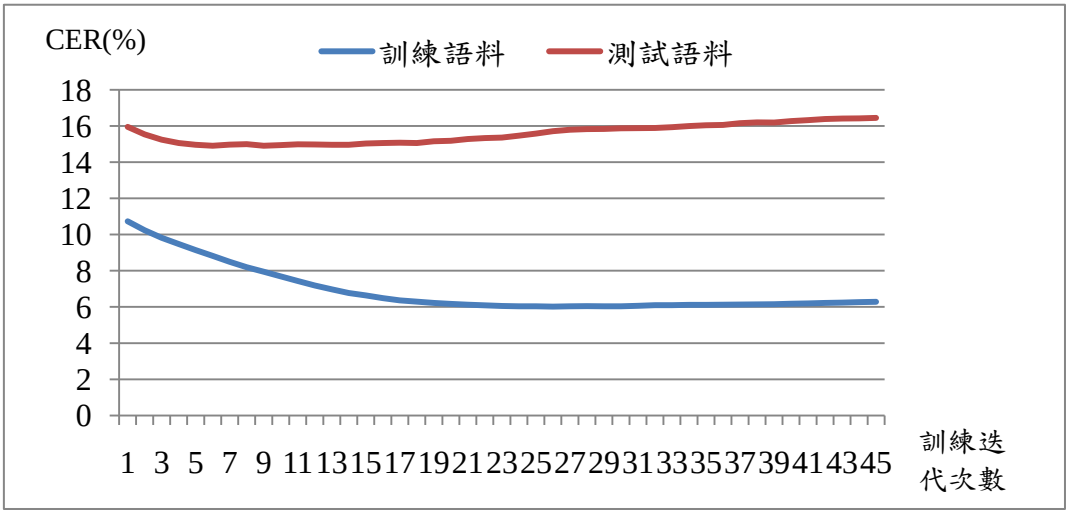

圖 5- 2 感知器演算法訓練語料與測試語料字錯誤率趨勢圖

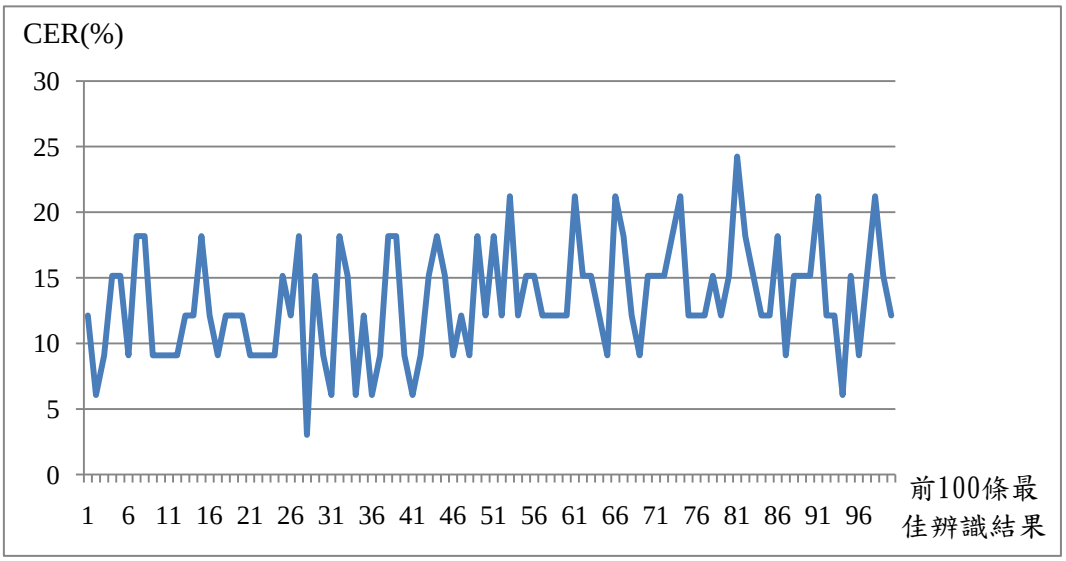

圖 5- 3 訓練語句 100 條最佳辨識結果之字錯誤率

而在比較全域條件式對數線性模型與感知器演算法的實驗結果後發現，全域條件式對數線性模型的實驗結果比感知器演算法有更顯著提昇，探究其原因應為全域條件式對數線性模型在更新權重時，考慮了訓練語句所有的候選詞序列與最低錯誤率詞序列間的關係，而感知器演算法則只有考慮目前訓練權重下，排序分數最高那條候選詞序列與最低錯誤率詞序列間的關係，相對之下全域條件式對數線性模型考慮的更為周全。

在考慮樣本權重的兩種鑑別式訓練模型，權重式全域條件式對數線性模型與最小化錯誤率訓練中，皆使用了兩種樣本權重進行訓練，分別是各候選詞序列的字錯誤率與各候選詞序列根據字錯誤率所做的排序。由實驗結果可看出，不管是權重式全域條件式對數線性模型或最小化錯誤率訓練，皆以字錯誤率為樣本權重的方式會較排序為樣本權重的方式來的有效果，為探究其原因，我們取了其中一條訓練語句的 100 條最佳辨識結果，觀察各個候選詞序列的字錯誤率，如圖 5-3 所示。由圖 5-3 可看出，前 100 條最佳辨識結果中有許多候選詞序列擁有相同的字錯誤率，若以排序為樣本權重，擁有相同的字錯誤率的候選詞序列也會有不同的排序，因而影響了以排序為樣本權重的方式的鑑別能力，導致重新排序結果較差。

而將考慮樣本權重的兩種鑑別式訓練模型與感知器演算法和全域條件式對數線性模型實驗結果做比較，可看出考慮樣本權重的鑑別式語言模型會較未考慮樣本權重的鑑別式語言模型有更好的效果，由此可知，樣本權重對鑑別式語言訓練有一定的幫助，因為訓練時考慮了每條後選詞序列的排序或錯誤率，因此可達到更好的效果。

而將最小化錯誤率訓練與其他方法實驗結果做比較，可看出最小化錯誤率訓練的結果最佳，除了樣本權重的幫助外，從式(3-23)來看，因最小化錯誤率訓練的目標函數並沒有用到最低錯誤率詞序列作為訓練標準，而是考慮所有候選詞序列的影響，因此不像另外三種鑑別式訓練模型容易受最低錯誤率詞序列影響訓練結果，最小化錯誤率訓練可得到更為一般化(General)的權重向量，讓訓練後的最佳權重向量不只符合訓練語料，在測試語料也能達到不錯的效果。

5.5 語句相關之鑑別式語言模型之實驗結果

表 5- 4 語句相關之鑑別式語言模型運用(選取相似度最大權重法)於感知器演算法之實驗結果(CER(%))

分群個數	測試語料辨識錯誤率	絕對提昇率	相對提昇率
2	15.88	0.51	3.09
3	15.96	0.43	2.61
4	15.72	0.67	4.08
5	16.19	0.20	1.21
6	16.03	0.36	2.17
7	15.84	0.55	3.38
8	16.19	0.20	1.21
9	15.95	0.44	2.68
10	15.88	0.51	3.14
原始	15.71	0.68	4.15

表 5- 5 語句相關之鑑別式語言模型(選取相似度最大權重法)運用於全域條件式對數線性模型之實驗結果(CER(%))

分群個數	測試語料辨識錯誤率	絕對提昇率	相對提昇率
2	15.61	0.78	4.74
3	15.64	0.76	4.61
4	15.71	0.68	4.17
5	15.63	0.76	4.63
6	15.61	0.78	4.78
7	15.58	0.81	4.94
8	15.75	0.64	3.88
9	16.03	0.36	2.20
10	15.70	0.69	4.19
原始	15.53	0.86	5.25

本節討論語句相關之鑑別式語言模型之實驗結果，本論文初步使用在感知器演算法與全域條件式對數線性模型上，以下分別討論本論文提出的三種語言特徵權重參數向量產生方式的實驗結果，如表 5-4 至表 5-9 所示。

表 5- 6 語句相關之鑑別式語言模型(相似度線性組合法)運用於感知器演算法之實驗結果(CER(%))

分群個數	測試語料辨識錯誤率	絕對提昇率	相對提昇率
2	15.13	1.26	7.70
3	14.86	1.53	9.34
4	14.77	1.62	9.87
5	15.00	1.40	8.49
6	14.78	1.61	9.82
7	14.70	1.69	10.28
8	14.82	1.57	9.60
9	14.92	1.47	8.99
10	15.11	1.28	7.83
原始	15.71	0.68	4.15

表 5- 7 語句相關之鑑別式語言模型(相似度線性組合法)運用於全域條件式對數線性模型之實驗結果(CER(%))

分群個數	測試語料辨識錯誤率	絕對提昇率	相對提昇率
2	15.52	0.87	5.29
3	15.59	0.80	4.87
4	15.72	0.67	4.06
5	15.48	0.91	5.57
6	15.60	0.80	4.83
7	15.57	0.82	4.98
8	15.73	0.66	4.04
9	16.05	0.34	2.09
10	15.67	0.72	4.41
原始	15.53	0.86	5.25

表 5-4 與表 5-5 比較兩種鑑別式語言模型在選取相似度最大權重法中，其訓練語料分群個數與字錯誤率間的關係、表 5-6 與表 5-7 比較兩種鑑別式語言模型在相似度線性組合法中，其訓練語料分群個數與字錯誤率間的關係、而表 5-8 與表 5-9 則是比較兩種鑑別式語言模型在最大機率法中，其訓練語料分群個數與字錯誤率間的關係，其中表中最後一行皆為傳統鑑別式語言模型的結果。比較語句相關之鑑別式語言模型用在兩種鑑別式語言模型的結果，可看出用選取相似度最

表 5-8 語句相關之鑑別式語言模型(最大機率法)運用於感知器演算法之實驗結果(CER(%))

分群個數	測試語料辨識錯誤率	絕對提昇率	相對提昇率
2	14.86	1.53	9.32
3	14.66	1.73	10.55
4	14.71	1.68	10.24
5	15.06	1.33	8.14
6	14.96	1.43	8.75
7	14.97	1.42	8.66
8	14.93	1.46	8.90
9	14.91	1.48	9.03
10	15.17	1.22	7.46
原始	15.71	0.68	4.15

表 5-9 語句相關之鑑別式語言模型(最大機率法)運用於全域條件式對數線性模型之實驗結果(CER(%))

分群個數	測試語料辨識錯誤率	絕對提昇率	相對提昇率
2	15.53	0.86	5.25
3	15.59	0.80	4.87
4	15.79	0.60	3.64
5	15.97	0.42	2.57
6	16.02	0.37	2.26
7	15.98	0.41	2.52
8	16.14	0.25	1.52
9	16.01	0.38	2.33
10	15.81	0.58	3.53
原始	15.53	0.86	5.25

大權重法的語句相關之鑑別式語言模型無法達到與傳統鑑別式語言模型相抗衡的效果，探究其原因應為選取相似度最大權重法只選取與測試語句最相近的一組特徵權重參數向量作為此測試語句的特徵權重參數向量，而此特徵權重參數向量為分群後的其中一群訓練語句群所訓練，因此訓練語句句數不足，而導致無法訓練出足夠的特徵權重參數以對測試語句產生效果，因此在 5.7 節我們將會提出結

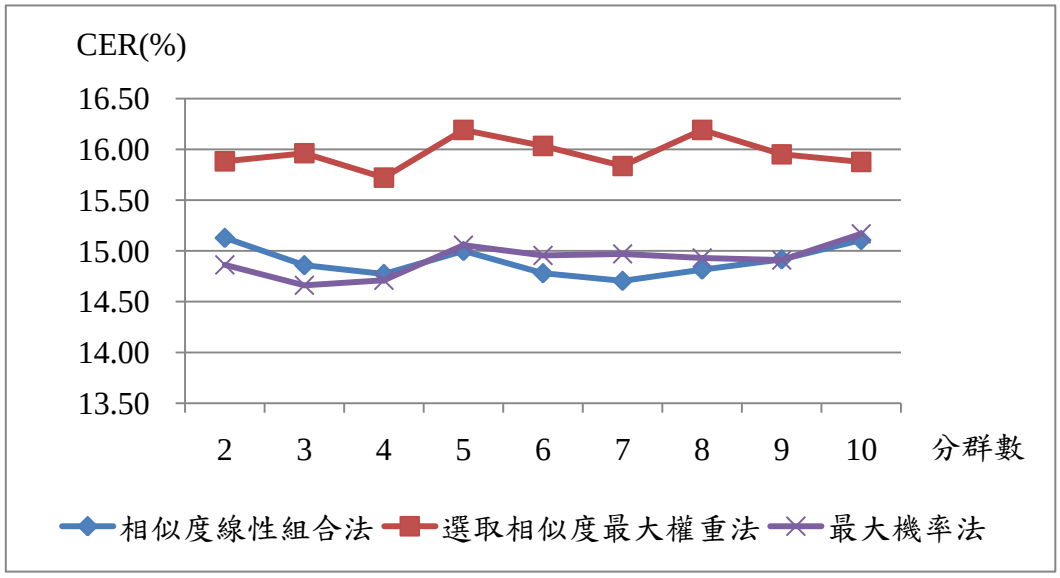

圖 5- 4 語句相關之鑑別式語言模型運用於感知器演算法之比較

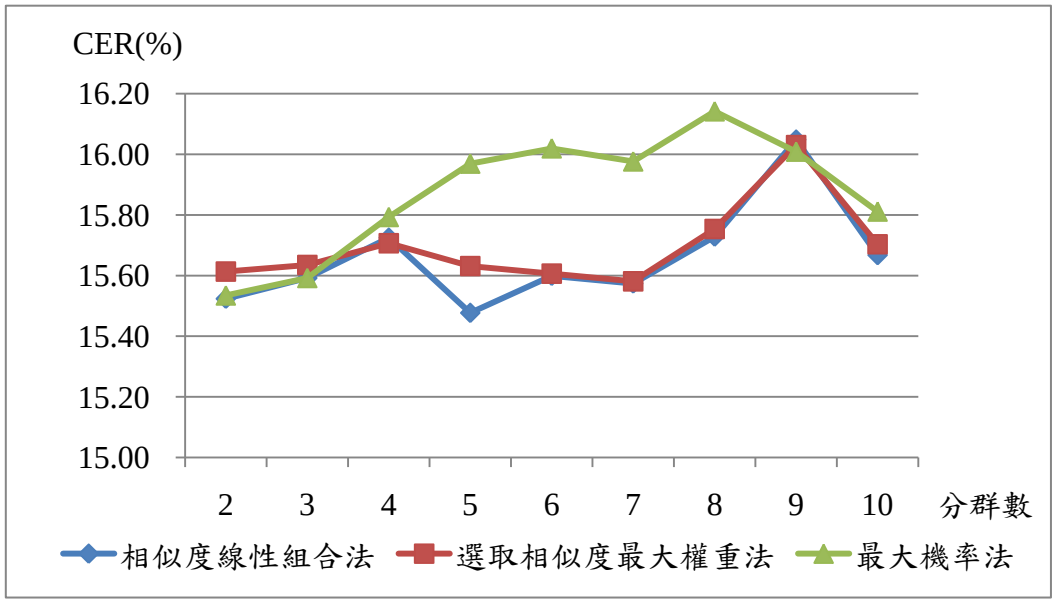

圖 5- 5 語句相關之鑑別式語言模型運用於全域條件式對數線性模型之比較

合語句相關之鑑別式語言模型與傳統鑑別式語言模型所訓練的特徵權重參數向量的方法，以解決語句相關之鑑別式語言模型特徵權重參數不足的問題。

另外，從實驗結果可看出語句相關之鑑別式語言模型與感知器演算法的結合的提昇效果較其與全域條件式對數線性模型的結合為佳；其中運用在感知器演算法的最大機率法更可達到 14.66%的字錯誤率，比前述的加入樣本權重資訊的鑑別式語言模型(權重式全域條件式對數線性模型的 15.44%與最小化錯誤率訓練的

15.33%)更為顯著的錯誤率降低。探究其原因可能為傳統感知器演算法因容易導致過度訓練，而語句相關之鑑別式語言模型因有考慮測試語句的特性來選擇各個訓練語料所產生的語言特徵權重參數向量作組合，所以可以減輕過度訓練的問題。另外，我們也可看出在全域條件式對數線性模型時，若分群數目過多時，可能會導致語句相關之鑑別式語言模型的效能下降；探究其原因應為過多的分群數目導致每一群的訓練語句的句數不足，而無法訓練出具有足夠鑑別能力的權重參數向量，因此同樣會用 5.7 節結合語句相關之鑑別式語言模型與傳統鑑別式語言模型所訓練的特徵權重參數向量的方法來解決此問題。

圖 5-4 與圖 5-5 比較了本論文提出的語句相關之鑑別式語言模型的三種語言特徵權重參數向量產生方式。在感知器演算法時，選取相似度最大權重法的效能最差，而相似度線性組合法與最大機率法兩者在伯仲之間；而在全域條件式對數線性模型時，最大機率法的效能最差，另外兩種方法則有差不多效能。造成感知器演算法與全域條件式對數線性模型的差別原因可能為，感知器演算法因為只訓練排序分數最大與辨識錯誤率最低的候選詞序列間不相同的特徵的權重參數；因此其訓練後的權重參數相對於全域條件式對數線性模型會減少許多，而利用選取相似度最大權重法因為只選取其中一組訓練語句群所訓練的權重參數向量，會更加重這個問題，因此導致感知器演算法在選取相似度最大權重法會有較差的效果。

5.6 結合語句相關之鑑別式語言模型訓練權重與所有訓練語句訓練權重之實驗結果

表 5- 10 結合語句相關之鑑別式語言模型(選取相似度最大權重法)訓練權重與所有訓練語句訓練權重(感知器演算法、分群個數=4)之實驗結果(CER(%))

α	測試語料辨識錯誤率	絕對提昇率	相對提昇率
0.0	15.71	0.68	4.15
0.1	15.48	0.91	5.55
0.2	15.29	1.10	6.73
0.3	15.25	1.14	6.97
0.4	15.12	1.27	7.76
0.5	15.12	1.27	7.72
0.6	15.27	1.12	6.84
0.7	15.36	1.03	6.29
0.8	15.42	0.97	5.92
0.9	15.54	0.85	5.18
1.0	15.72	0.67	4.08

表 5- 11 結合語句相關之鑑別式語言模型(選取相似度最大權重法)訓練權重與所有訓練語句訓練權重(全域條件式對數線性模型、分群個數=7)之實驗結果(CER(%))

α	測試語料辨識錯誤率	絕對提昇率	相對提昇率
0.0	15.53	0.86	5.26
0.1	15.39	1.00	6.08
0.2	15.37	1.02	6.21
0.3	15.40	0.99	6.03
0.4	15.39	1.00	6.12
0.5	15.40	0.99	6.03
0.6	15.42	0.97	5.94
0.7	15.43	0.96	5.86
0.8	15.49	0.90	5.48
0.9	15.56	0.83	5.09
1.0	15.58	0.81	4.94

表 5- 12 結合語句相關之鑑別式語言模型(相似度線性組合法)訓練權重與所有訓練語句訓練權重(感知器演算法、分群個數=7)之實驗結果(CER(%))

α	測試語料辨識錯誤率	絕對提昇率	相對提昇率
0.0	15.71	0.68	4.15
0.1	15.43	0.96	5.86
0.2	15.09	1.30	7.92
0.3	14.91	1.48	9.03
0.4	14.72	1.67	10.20
0.5	14.56	1.83	11.16
0.6	14.33	2.06	12.54
0.7	14.28	2.11	12.87
0.8	14.35	2.04	12.48
0.9	14.51	1.88	11.45
1.0	14.70	1.69	10.28

表 5- 13 結合語句相關之鑑別式語言模型(相似度線性組合法)訓練權重與所有訓練語句訓練權重(全域條件式對數線性模型、分群個數=5)之實驗結果(CER(%))

α	測試語料辨識錯誤率	絕對提昇率	相對提昇率
0.0	15.53	0.86	5.26
0.1	15.42	0.97	5.94
0.2	15.34	1.05	6.40
0.3	15.33	1.06	6.47
0.4	15.29	1.10	6.71
0.5	15.31	1.08	6.58
0.6	15.42	0.97	5.94
0.7	15.47	0.92	5.62
0.8	15.43	0.96	5.86
0.9	15.42	0.97	5.90
1.0	15.48	0.91	5.57

本節將語句相關之鑑別式語言模型所訓練的特徵權重向量$\boldsymbol{\lambda}_k$(考慮測試語句特性)與傳統鑑別式語言模型使用所有訓練語句訓練的特徵權重$\boldsymbol{\lambda}^{all}$(考慮所有訓練語句一般性)作線性組合，得到一組新的權重向量$\boldsymbol{\lambda}_k^{combine}$ ，以解決語句相關之鑑別

表 5- 14 結合語句相關之鑑別式語言模型(最大機率法)訓練權重與所有訓練語句訓練權重(感知器演算法、分群個數=3)之實驗結果(CER(%))

α	測試語料辨識錯誤率	絕對提昇率	相對提昇率
0.0	15.71	0.68	4.15
0.1	15.43	0.96	5.86
0.2	15.21	1.18	7.19
0.3	15.07	1.32	8.03
0.4	14.89	1.50	9.12
0.5	14.72	1.67	10.17
0.6	14.71	1.68	10.26
0.7	14.61	1.78	10.88
0.8	14.58	1.81	11.07
0.9	14.59	1.80	11.01
1.0	14.66	1.73	10.55

表 5- 15 結合語句相關之鑑別式語言模型(最大機率法)訓練權重與所有訓練語句訓練權重(全域條件式對數線性模型、分群個數=2)之實驗結果(CER(%))

α	測試語料辨識錯誤率	絕對提昇率	相對提昇率
0.0	15.53	0.86	5.26
0.1	15.42	0.97	5.92
0.2	15.37	1.02	6.25
0.3	15.35	1.04	6.32
0.4	15.35	1.04	6.32
0.5	15.42	0.97	5.94
0.6	15.44	0.95	5.77
0.7	15.43	0.96	5.83
0.8	15.42	0.97	5.94
0.9	15.51	0.88	5.37
1.0	15.53	0.86	5.22

式語言模型訓練語句不足的問題，並期望兩種不同訓練目標所訓練的特徵權重向量可以幫助語音辨識系統獲到更好的重新排序結果，其線性組合的方式如式(5-3)所示：

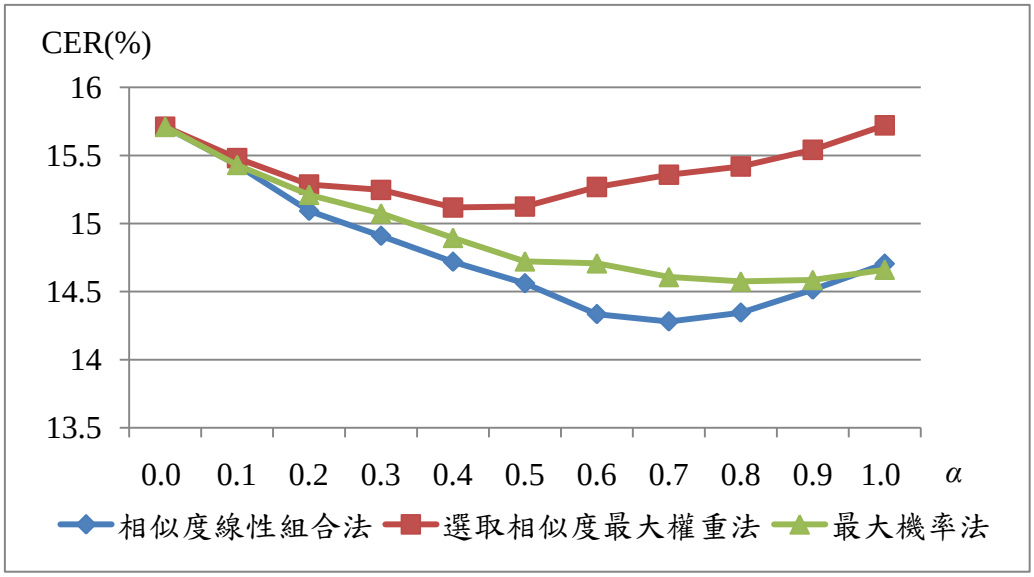

圖 5- 6 結合語句相關之鑑別式語言模型與所有訓練語句訓練權重運用於感知器演算法之比較

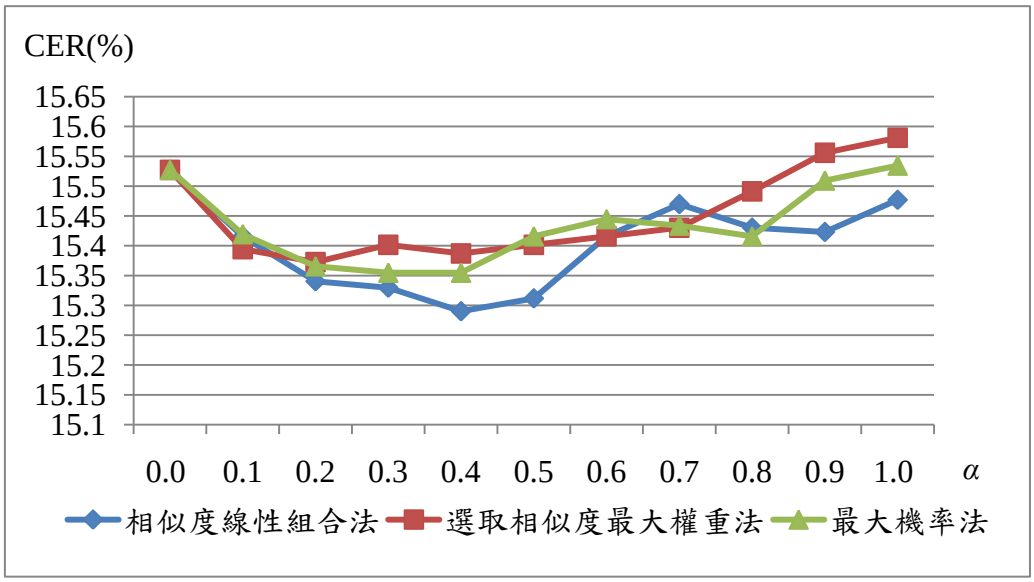

圖 5- 7 結合語句相關之鑑別式語言模型與所有訓練語句訓練權重運用於全域條件式對數線性模型之比較

$$\boldsymbol{\lambda}_k^{combine} = \alpha \cdot \boldsymbol{\lambda}_k + (1-\alpha) \cdot \boldsymbol{\lambda}^{all} \tag{5-3}$$

其中 α 為調整係數，用來控制語句相關之鑑別式語言模型所訓練的語言特徵權重向量 $\boldsymbol{\lambda}_k$ 與使用所有訓練語句訓練的語言特徵權重 $\boldsymbol{\lambda}^{all}$ 的貢獻；在本研究 α 的值是介於 0.1 到 0.9 之間。

表 5-10 至表 5-15 分別為結合本論文提出的語句相關之鑑別式語言模型三種權重參數向量產生方法所產生的特徵權重向量與傳統鑑別式語言模型使用所有訓練語句訓練的特徵權重參數向量的實驗結果，其中語句相關之鑑別式語言模型三種權重參數向量產生方法所產生的特徵權重參數向量皆選取於 5.6 節實驗中，效果最佳的分群數下所訓練出的特徵權重參數向量。表 5-10 與表 5-11 比較感知器演算法與全域條件式對數線性模型在選取相似度最大權重法下，調整係數與字錯誤率間的關係、表 5-12 與表 5-13 比較此兩種鑑別式語言模型在相似度線性組合法中，調整係數與字錯誤率間的關係、而表 5-14 與表 5-15 則是比較兩種鑑別式語言模型在最大機率法中，調整係數與字錯誤率間的關係。

從實驗結果可看出，結合語句相關之鑑別式語言模型所訓練的特徵權重向量與傳統鑑別式語言模型使用所有訓練語句訓練的特徵權重可得到顯著的語音辨識錯誤率降低，其中感知器演算法的相似度線性組合法中，α 取 0.7 時甚至可達到比最大事後機率法的 14.51%更佳的字錯誤率 14.28%。由此可知結合語句相關之鑑別式語言模型訓練權重與所有訓練語句訓練的特徵權重由於訓練目標不同，若將兩者結合則可達到互補的效果，因而達到顯著的提昇。

圖 5-6 與圖 5-7 比較了語句相關之鑑別式語言模型的三種語言特徵權重參數向量產生方式與傳統使用所有訓練語句訓練的特徵權重參數向量結合的實驗結果。從圖 5-6 可看出，在感知器演算法中，相似度線性組合法的效果最佳，接著依序為最大機率法與選取相似度最大權重法；而在全域條件式對數線性模型中，雖然規律較為雜亂，但仍以相似度線性組合法的效果為最佳，與感知器演算法實驗結果相同。

接著觀察最低字錯誤率所取的α值，感知器演算法集中在α取較大數值的後半段，而全域條件式對數線性模型則集中在α取較小數值的前半段。因此可知道感知器演算法中，語句相關之鑑別式語言模型所訓練的特徵權重向量 $\boldsymbol{\lambda}_j$ 佔據較

重要的角色，而在全域條件式對數線性模型則是所有訓練語句訓練的特徵權重 λ^{all} 較為重要。討論其原因應為感知器演算法易有過度訓練的問題，因此語句相關之鑑別式語言模型所訓練的特徵權重向量因為有考慮測試語句本身的特性，所以加重它的重要性可減輕特徵權重過度符合訓練語料的問題；而全域條件式對數線性模型中，所有訓練語句訓練的特徵權重較為重要則可能是因為要減輕訓練語料分群後導致一般化能力下降的問題。

5.7 結合最大化事後機率法與語句相關之鑑別式語言模型實驗結果

表 5- 16 結合最大化事後機率法與語句相關之鑑別式語言模型(選取相似度最大權重法)運用於感知器演算法之實驗結果(CER(%))

分群個數	測試語料辨識錯誤率	絕對提昇率	相對提昇率
2	15.21	1.18	7.19
3	15.30	1.09	6.67
4	15.18	1.21	7.39
5	15.55	0.87	5.11
6	15.35	1.04	6.34
7	15.25	1.14	6.93
8	15.42	0.97	5.92
9	15.49	0.90	5.51
10	15.36	1.03	6.29
原始	15.10	1.29	7.87

表 5- 17 結合最大化事後機率法與語句相關之鑑別式語言模型(相似度線性組合法)運用於感知器演算法之實驗結果(CER(%))

分群個數	測試語料辨識錯誤率	絕對提昇率	相對提昇率
2	14.92	1.47	8.97
3	14.47	1.92	11.73
4	14.22	2.17	13.24
5	14.43	1.96	11.95
6	14.23	2.16	13.20
7	14.09	2.30	14.01
8	14.12	2.27	13.83
9	14.11	2.28	13.92
10	14.07	2.32	14.14
原始	15.10	1.29	7.87

表 5- 18 結合最大化事後機率法與語句相關之鑑別式語言模型(最大機率法)運用於感知器演算法之實驗結果(CER(%))

分群個數	測試語料辨識錯誤率	絕對提昇率	相對提昇率
2	15.20	1.19	7.28
3	15.24	1.15	7.02
4	14.99	1.40	8.53
5	15.23	1.16	7.08
6	15.89	0.50	3.07
7	15.51	0.88	5.40
8	15.65	0.74	4.50
9	14.86	1.53	9.36
10	15.49	0.90	5.46
原始	15.10	1.29	7.87

與上一節相同，將最大化事後機率法的模型差補法所產生的前 M 條候選詞序列利用本論文提出的語句相關之鑑別式語言模型的三種係數估計方法重新排序，由於由上一節的實驗結果顯示，語句相關之鑑別式語言模型運用在感知器演算法的效果較為顯著，因此本節的實驗以感知器演算法為主，實驗結果如表 5-17 至 5-19 所示。

從表中的最後一行可看出，最大化事後機率法與傳統鑑別式語言模型的結合並無法達到原始最大化事後機率法的效果，探究其原因可能為最大化事後機率法所產生的前 M 條候選詞序列已有極低的辨識錯誤率，而傳統鑑別式語言模型只能以重新排序來降低錯誤率的方法，無法獲致更好的效果。另外則可能為感知器演算法容易過度訓練的問題。

而從與語句相關之鑑別式語言模型結合的實驗結果可看出，相似度線性組合法在分群數為 10 群時，可達到比原始最大化事後機率法 14.51%更低的字錯誤率 14.07%，而另外兩種方法雖無法達到原始最大化事後機率法的效果，但皆可達到

表 5- 19 結合語句相關之鑑別式語言模型(選取相似度最大權重法)訓練權重與所有訓練語句訓練權重(感知器演算法、分群個數=10)之實驗結果(CER(%))

α	測試語料辨識錯誤率	絕對提昇率	相對提昇率
0.0	15.10	0.68	4.15
0.1	14.98	0.91	5.55
0.2	14.82	1.10	6.73
0.3	14.75	1.14	6.97
0.4	14.74	1.27	7.76
0.5	14.80	1.27	7.72
0.6	14.85	1.12	6.84
0.7	14.95	1.03	6.29
0.8	15.00	0.97	5.92
0.9	15.03	0.85	5.18
1.0	15.18	0.67	4.08

表 5- 20 結合語句相關之鑑別式語言模型(相似度線性組合法)訓練權重與所有訓練語句訓練權重(感知器演算法、分群個數=4)之實驗結果(CER(%))

α	測試語料辨識錯誤率	絕對提昇率	相對提昇率
0.0	15.10	1.29	7.85
0.1	14.93	1.46	8.92
0.2	14.75	1.64	10.02
0.3	14.55	1.84	11.23
0.4	14.44	1.95	11.93
0.5	14.32	2.07	12.65
0.6	14.17	2.22	13.55
0.7	14.07	2.32	14.14
0.8	14.04	2.35	14.36
0.9	14.01	2.38	14.51
1.0	14.07	2.32	14.14

比基礎辨識率 16.39%或未與最大化事後機率法結合的傳統感知器演算法 15.71%更佳的字錯誤率。

表 5- 21 結合語句相關之鑑別式語言模型(最大機率法)訓練權重與所有訓練語句訓練權重(感知器演算法、分群個數=9)之實驗結果(CER(%))

α	測試語料辨識錯誤率	絕對提昇率	相對提昇率
0.0	15.10	1.29	7.85%
0.1	14.95	1.44	8.82%
0.2	14.79	1.60	9.78%
0.3	14.65	1.74	10.61%
0.4	14.56	1.83	11.14%
0.5	14.46	1.93	11.80%
0.6	14.44	1.95	11.91%
0.7	14.45	1.94	11.82%
0.8	14.54	1.85	11.27%
0.9	14.75	1.64	10.00%
1.0	14.86	1.53	9.36%

相同的，我們也利用式(5-3)將上述的三種係數估計方法中最佳分群數的語句相關之鑑別式語言模型訓練權重與所有訓練語句訓練權重結合，實驗結果如表 5-20 至 5-22 所示。由實驗結果可看出，如之前所述，結合語句相關之鑑別式語言模型訓練權重與所有訓練語句訓練權重可達到更佳的錯誤率降低：在選取相似度最大權重法中可達到 14.74%的字錯誤率、在相似度線性組合法則可達到最佳的字錯誤率 14.01%、而在最大機率法可達到 14.44%的字錯誤率，除了選取相似度最大權重法外，另外兩種組合係數估計方法皆可達到比原始最大化事後機率法 14.51%更低的字錯誤率。由此可知，兩種不同訓練目標的權重結合的確可達到互補的效果。

第 6 章 結論與未來展望

語言模型在人類生活中扮演著極重要的角色，他使語言保有它的規則性，使其更能精確的表達或理解語句。而在語音辨識領域中，語言模型一樣有著舉足輕重的地位。語言模型中最常見的為統計式語言模型，其中最廣泛使用的為 N 連語言模型，此模型以找到對應一語音訊號最可能的詞序列為目標來進行語音辨識。而近年來，一種新興的語言模型被提出，即為鑑別式語言模型；此類語言模型與傳統 N 連語言模型不同點在於希望能直接針對降低錯誤率為目標做訓練，而不是間接的找出機率最高的詞序列。

本論文針對鑑別式語言模型做討論，介紹了鑑別式語言模型的基礎精神，並討論了幾種現有的鑑別式語言模型，探討他們之間的關係，與實驗結果所代表的意義。並提出語句相關之鑑別式語言模型，可針對個別測試語句結合不同權重向量，以改善傳統鑑別式語言模型所有測試語句皆使用相同語言模型權重向量的缺點；再者，我們將語句相關之鑑別式語言模型所訓練的權重參數向量與傳統鑑別式語言模型所訓練的權重參數向量利用插補法結合，以期兩者以不同訓練目標所訓練的權重參數向量能達到互補的效果，並獲得更顯著的效能。實驗結果顯示本論文所提出的語句相關之鑑別式語言模型，在部份的訓練方法下，可相較於僅使用三連語言模型、或使用傳統鑑別式語言模型的基礎大詞彙連續語音辨識系統，有一定的語音辨識率提升；而將語句相關之鑑別式語言模型與傳統鑑別式語言模型結合的方式，更能有相當程度的語音辨識率提升，在部份情況下，甚至可達到比最大事後機率法更佳的效能。

未來希望對語句相關之鑑別式語言模型做進一步的改進，包括嘗試不同的訓練語料的分群方式、或是更佳的組合係數的計算方式；並希望能將語句相關之鑑別式語言模型運用在其它鑑別式語言模型上面，以期能獲致更佳的辨識率提昇；

此外，在鑑別式語言模型的改進方面，期望能考慮不同語言特徵對鑑別式語言模型的幫助，以找出對鑑別式語言模型有益的於延特徵；並且希望對鑑別式語言模型容易被訓練語料與測試語料不相依(Mismatch)而影響辨識率的缺點作更進一步改進。

參考文獻

[Arisoy *et al*. 2010] E. Arisoy, M. Saraclar, B. Roark, and I. Shafran, “Syntactic and sub-lexical features for Turkishi discriminative language models,” ICASSP, 2010.

[Bahl *et al.* 1986] L.R. Bahl, P.F. Brown, P.V. de Souza, and L.R. Mercer, “Maximum mutual information estimation of Hidden Markov Model parameters for speech recognition,” ICASSP, 1986.

[Bahl *et al.* 1983] L. R. Bahl, F. Jelinek and R. L. Mercer, “A Maximum Likelihood Approach to Continuous Speech Recognition,” IEEE Trans. Pattern Analysis and Machine Intelligence, 1983.

[Bellegarda 2005] J. R. Bellegarda, “Latent Semantic Mapping,” IEEE Signal Processing Magazine, Vol. 22. No. 5, pp. 70- 80, 2005.

[Brown *et al.* 1992] P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. “Class-based *n*-gram models of natural language,” Computational Liguistics, 1992.

[Chelba and Jelinek 2000] C. Chelba and F. Jelinek, “Structured language modeling,” Computer Speech and Language, 2000.

[Chen *et al.* 2004] B. Chen, J.-W. Kuo and W.-H. Tsai, “Lightly supervised and data-driven approaches to mandarin broadcast news transcription,” ICASSP, 2004.

[Clarkson and Robinson 1997] P. R. Clarkson and A. J. Robinson, “Language model adaptation using mixtures and an exponentially decaying cache,” ICASSP, 1997.

[CNA News] Central News Agency, http://www.cna.com.tw/.

[Collins and Koo 2000] M. Collins and T. Koo, “Discriminative reranking for natural language parsing,” ICML, 2000.

[Collins 2002] M. Collins, “Discriminative training methods for Hidden Markov Models: theory and experiments with perceptron algorithms,” EMNLP, 2002.

[Collins *et al*. 2005] M. Collins, B. Roark, and M. Sraclar, “Discriminative syntactic language modeling for speech recognition,” ACL, 2005.

[Freund and Schapire 1999] Y. Freund and R. Schapire, “Large margin classification using the perceptron algorithm,” Machine Learning, 277-296, 1999.

[Gao *et al*. 2005] J. Gao, H. Suzuki, and W. Yuan, “An empirical study on language model adaptation,” TALIP, 2005.

[Gildea and Hofmann 1999] D. Gildea and T. Hofmann, “Topic-based language models using EM,” Eurospeech, 1999.

[Gopalakrishnan et al. 1991] P.S. Gopalakrishnan, D. Kanevsky, D. Nahamoo, and A. Nadas, “An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems,” IEEE Trans. Information Theory, vol. 37, no. 1, 1991.

[Huang *et al*. 2010] J.-T. Huang, X. Li, and A. Acero, “Discriminative training methods for language models using conditional entropy criteria,” ICASSP, 2010.

[Juang and Katagiri 1992] B.-H. Juang and S. Katagiri, “Discriminative learning for minimum error classification,” IEEE Trans. Signal Processing, Vol. 40, No. 12, pp. 3043-3054, 1992.

[Jelinek 1999] F. Jelinek, “Statistical methods for speech recognition,” the MIT Press, 1999.

[Johnson *et al*. 1999] M. Johnson, S. Geman, S. Canon, Z. Chi, and S. Riezler, “Estimators for stochastic “unification-based” grammars,” ACL, 1999.

[Kaufmann *et al*. 2009] T. Kaufmann, T. Ewender, and B. Pfister, “Improving broadcast news transcription with a precision grammar and discriminative reranking,” Interspeech, 2009.

[Kobayashi *et al.* 2008] A. Kobayashi, T. Oku, S. Homma, S. Sato, T. Imai, and T. Takagi, “Discriminative rescoring based on minimization of word errors for transcribing broadcast news,” Interspeech, 2008.

[Kuo *et al*. 2002] H.-K. J. Kuo, E. Fosler-Lussier, H. Jiang and C. H. Lee, “Discriminative training of language models for speech recognition,” ICASSP,2002.

[Kuo and Chen 2005] J. W. Kuo and B. Chen, “Minimum word error based discriminative training of language models,” Eurospeech, 2005.

[Lafferty *et al*. 2001] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: probabilistic models for segmenting and labeling sequence data,” ICML, 2001.

[Magdin and Jiang 2009] V. Magdin and H. Jiang, “Discriminative training of *n*-gram language models for speech recognition via linear programming,” ASRU, 2009.

[Magdin and Jiang 2010] V. Magdin and H. Jiang, “Large margin of *n*-gram language model for speech recognition via linear programming,” ICASSP, 2010.

[Oba *et al*. 2010] T. Oba, T. Hori and A. Nakamura, “A comparative study on methods of weighted language model training for reranking LVCSR *n*-best hypotheses,” ICASSP, 2010.

[Och 2003] F. J. Och, “Minimum error rate training in statistical machine translation,” ACL, 2003.

[Ortmanns *et al.* 1997] S. Ortmanns, H. Ney and X. L. Aubert, “A word graph algorithm for large vocabulary continuous speech recognition,“ Computer Speech and Language, Vol. 11, pp.43-72, 1997.

[Povey 2004] D. Povey, “Discriminative Training for Large Vocabulary Speech Recognition,” Ph.D Dissertation, Peterhouse, University of Cambridge, July 2004.

[Rastrow *et al*. 2009] A. Rastrow, A. Sethy, and B. Ramabhadran, “Constrained discriminative training of *n*-gram language models,” ASRU, 2009.

[Ratnaparkhi *et al*. 1994] A. Ratnaparkhi, S. Roukos, and R. T. Ward, “A maximum entropy model for parsing,” ICSLP, 1994.

[Rigazio *et al*. 1998] L. Rigazio, J.-C. Junqua, and M. Galler, “Multilevel discriminative training for spelled word recognition,” ICASSP, 1998.

[Roark *et al*. 2004a] B. Roark, M. Saraclar, M. Collins, and M. Johnson, "Corrective langrage modeling for large vocabulary ASR with the perceptron algorithm," ICASSP, 2004.

[Roark *et al*. 2004b] B. Roark, M. Saraclar, M. Collins, and M. Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," ACL, 2004.

[Roark *et al*. 2007] B. Roark, M. Saraclar, and M. Collins, "Discriminative *n*-gram language modeling," Computer Speech and Language, vol. 21, no. 2, pp. 373-392, 2007

[Roark 2009] B. Roark, "A survey of discriminative language modeling approaches for Large Vocabulary Continuous Speech Recognition," Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods, 2009.

[Rosenblatt 1958] F. Rosenblatt, "A probabilistic model for information storage and organization in the brain," Psychological Review, 65, 386-408, 1958.

[Saul and Pereira 1997] L. Saul and F. Pereira, "Aggregate and mixed-order Markov models for statistical language processing," EMNLP, 1997.

[Singh-Miller and Collins 2007] N. Singh-Miller and M. Collins, "Trigger-based language modeling using a loss-sensitive perceptron algorithm," ICASSP, 2007.

[SRI] A. Stolcke, SRI Language Modeling Toolkit, version 1.5.8, http://www.speech.sri.com/projects/srilm/

[Tam and Schultz 2005] Y. C. Tam and T. Schultz, "Language model adaptation using variational bayes inference," Interspeech, 2005.

[Troncoso *et al.* 2004] C. Troncoso, T. Kawahara, H. Yamamoto, and G. Kikui, "Trigger-based language model construction by combining different corpora," IEICE Technical Report, 2004.

[Warnke *et al*. 1999] V. Warnke, S. Harbeck, E. Noth, H. Niemann and M. Levit, "Discriminative estimation of interpolation parameters for language model classifiers," ICASSP, 1999.

[Xu *et al*. 2009] P. Xu, D. Karakos, and S. Khudanpur, "Self-supervised discriminative training of statistical language models," ASRU, 2009.

[Zhou and Meng 2008] Z. Zhou and H. Meng, "Recasting the discriminative *n*-gram model as a pseudo-conventional *n*-gram model for LVCSR," ICASSP, 2008.

[Zhou *et al*. 2006] Z. Zhou, J. Gao, F. K. Soong, and H. Meng, "A comparative study of discriminative methods for reranking lvcsr *n*-best hypotheses in domain adaptation and generalization," ICASSP, 2006.

[劉鳳萍 2009] 劉鳳萍，"使用鑑別式語言模型於語音辨識結果重新排序，"國立臺灣師範大學資訊工程所碩士論文，2009。

[劉鳳萍等人 2009] 劉鳳萍，陳冠宇，劉家妏，張鈺玫，陳柏琳，"鑑別式語言模型調適法於大詞彙連續語音辨識之研究，" TAAI，2009。

[邱炫盛 2007] 邱炫盛 "利用主題與位置相關語言模型於中文連續語音辨識，" 國立臺灣師範大學資訊工程所碩士論文，2007。

[邱炫盛等人 2007] 邱炫盛，羅永典，陳韋豪，陳柏琳，"使用位置資訊於中文連續語音辨識，" TAAI，2007。