鑑別式語言模型於語音辨識結果重新排序之研究

Exploiting Discriminative Language Models for Reranking Speech Recognition Hypotheses

劉家妏　Chia-Wen Liu
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
697470171@ntnu.edu.tw

林士翔　Shih-Hsiang Lin
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
896470017@ntnu.edu.tw

陳柏琳　Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
berlin@ntnu.edu.tw

摘要

任何語言都有許多潛在的規律性，若能擷取與分析這些特性，電腦就能進一步被用來理解人類語句或自動產生能表達某種語意的語句。統計式語言模型(Statistical Language Models)透過機率模型的建立來描述語言生成的規律性，其模型參數可由大量的文字語料庫所訓練而成。語音辨識常使用 *N* 連(*N*-gram)語言模型，它以估測每一個詞在其先前緊鄰 *N*-1 個詞已知的情況下出現的條件機率，來判斷語音辨識結果的可能性；但因其訓練並不是以降低語音錯誤率爲目標，導致在語音辨識效能表現上有所侷限。有別於傳統 *N* 連語言模型，近年來有許多直接以最小化語音辨識錯誤率爲目標的鑑別式語言模型(Discriminative Language Model)被提出。本論文介紹了多種基於不同訓練精神的鑑別式語言模型，並比較與討論它們在中文大詞彙連續語音辨識上的表現。另外，我們提出語句相關之鑑別式語言模型，改進了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特徵權重參數向量的缺點，讓不同測試語句擁有各自的組合係數來線

Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Pages 30-49, Puli, Nantou,Taiwan, September 2010.

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，以期新的權重參數向量能更加符合測試語句的特性。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，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，能有相當程度的語音辨識率提升。

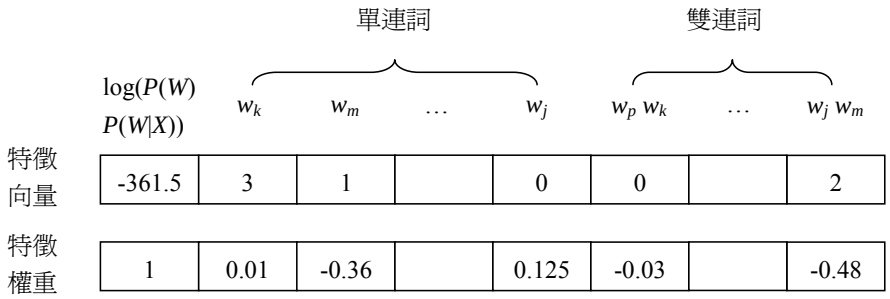
關鍵詞：語音辨識、語言模型、鑑別式語言模型、重新排序。

一、緒論

語言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」能力就是語音處理領域長期以來努力的目標。爲了讓電腦擁有此能力，首先要做到的便是如何讓它能夠「聽」懂使用者的語音輸入；而將語音訊號轉換成文字的過程，須透過自動語音辨識(Automatic Speech Recognition, ASR)來達成。爲此，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。在傳統統計式語言模型中，最爲被廣泛運用的爲 N 連(*N*-gram)語言模型[1]。 N 連語言模型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；並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有最高機率的詞序列(Word Sequence)做爲最後語音辨識結果。

然而，藉由使用傳統 N 連語言模型所找出機率最高的詞序列未必是最佳(錯誤率最低)的語音辨識結果；在詞圖或前 M 條最佳辨識候選詞序列中，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。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語言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爲目標的鑑別式語言模型[2, 3, 4, 5, 6, 7]被提出。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爲訓練目標；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爲最後的輸出，因此降低辨識錯誤率。

本論文將介紹各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；另外，本論文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出語言模型特徵權重參數向量的方法(或稱之語句相關之鑑別式



圖一、特徵向量與特徵權重向量

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特徵權重參數向量之缺點。本論文的安排如下：第二節將介紹多種鑑別式語言模型的訓練準則與幾種近年常見的鑑別式語言模型；第三節則會介紹本論文提出的語句相關之鑑別式語言模型；第四節則是實驗結果與分析；第五節是結論。

二、常見鑑別式語言模型介紹

(一)、鑑別式語言模型與訓練之基礎

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爲目標，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爲例，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作內積後而得的分數能爲最高。以下將對鑑別式語言模型訓練定義其參數[3, 4, 5, 6, 7]：

(a) 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爲 $\mathrm{GEN}(x_i)=\{W_{i,j}\}$，其中 j 介於 1 到 M 之間。

(b) 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 爲訓練語料句數；W_i^R 爲 $\mathrm{GEN}(x_i)$ 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 爲測試語料之句數。

(c) 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，每個語言特徵皆爲一個將候選詞序列 $W_{i,j}$ 對應到實數值之函數。$f_0(W_{i,j})$ 爲基礎語言特徵，本論文定義爲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爲候選詞序列 $W_{i,j}$ 中，各 N 連詞出現的次數；本論文使用單連詞

(Word Unigram)與二連詞(Word Bigram)出現的次數做爲其餘的語言特徵。

(d) 每一語言特徵定義都有其對應的權重參數值，爲一 $D+1$ 維的參數向量 $\lambda=[\lambda_0,...,\lambda_D]$，每一語言特徵及其對應的權重參數值如圖一所示意。候選詞序列 $W_{i,j}$ 的排序分數則定義爲特徵權重向量 $\boldsymbol{\lambda}$ 與特徵向量 $\mathbf{f}(W_{i,j})$ 之內積：

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

則排序分數最高的候選詞序列 W_i^* 即爲重新排序結果：

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

鑑別式語言模型的訓練目標，即是求得一組最佳權重參數解，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(1)的評分機制，從測試語句 y_k 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^*，將其作爲重新排序後的輸出結果。以下將介紹幾種常用的鑑別式語言模型。

(二)、一般鑑別式語言模型

1、感知器演算法(Perceptron Algorithm)

感知器[8]最初是被應用在人工類神經網路(Artificial Neural Networks)領域，它是一種二元分類器，把輸入 $\mathbf{v}$ (實數值向量)映射到輸出值 $f(\mathbf{v})$ 上(二元數)，如式(3)所示：

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

其中 $\mathbf{w}$ 實數權重向量，b 爲一常數，表偏移量。此方法可視爲一種最簡單形式的前饋式(Feed-Forward)人工神經網路[8]。感知器演算法[9]亦可以視爲條件式隨機域(Conditional Random Fields, CRFs) [10, 11]的一種變形，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)，其排序減損函數的觀念爲最小平方誤差(Least-Squares Error, LSE)[12]，以鑑別式語言模型訓練言，若感知器演算法的目標函數表示成：

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

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 model $\boldsymbol{\lambda}$ to choose W_i^* from GEN(x_i)
    For $d = 1...D$
      $\lambda_d = \lambda_d + \eta \cdot \left(f_d(W_i^R) - f_d(W_i^*)\right)$, where η is the size of the learning step.

圖二、感知器演算法[12]

則最佳權重向量$\boldsymbol{\lambda}^*$即爲滿足$\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}} F_{Perc}(\boldsymbol{\lambda})$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。爲了求得$\boldsymbol{\lambda}^*$，我們可採用梯度下降法(Gradient Descent Method)將式(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(5)所示：

$$\frac{\partial F_{Perc}(\boldsymbol{\lambda})}{\partial \lambda_d} = \sum_{i=1}^{L} \left(Score(W_i^R, \boldsymbol{\lambda}) - Score(W_i^*, \boldsymbol{\lambda})\right)\left(f_d(W_i^R) - f_d(W_i^*)\right) \tag{5}$$

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，感知器演算法常採取隨機近似法(Stochastic Approximation)[3]；相對於梯度下降法同時對所有訓練語句計算權重參數更新量，隨機近似法使用增量式(Incremental)的方法計算更新量，也就是將式(4)對每一句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參數調整量：

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

於是，隨機近似法可視爲最佳化個別訓練語句的排序減損函數$\hat{F}_{Perc}(i, \boldsymbol{\lambda})$：

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

而感知器演算法對每一句訓練語句的每一維特徵權重參數更新式可表示成：

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

其中η爲學習步調常數。關於權重調整量亦有學者提出直接以計算語言特徵向量的差值來更新特徵權重參數[4]，即$\hat{\lambda}_d^i = \lambda_d^i + \left(f_d(W_i^R) - f_d(W_i^*)\right)$，本論文亦使用此方式來更新特徵權重參數，其演算法如圖二所示。

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[13]：

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

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

2、全域條件式對數線性模型(Global Conditional Log-Linear Model, GCLM)

全域條件式對數線性模型早先在自然語言處理領域[14, 15]被提出，並在 2007 年被第一次應用在語音辨識重新排序問題上[4]；它利用權重參數向量 $\boldsymbol{\lambda}$ 及每一句語音訊號 x_i 的所有候選詞序列 $W_{i,j} \in \mathrm{GEN}(x_i)$ 來定義一個條件分佈：

$$p_\lambda(W_{i,j} \mid x_i) = \frac{1}{Z(x_i, \boldsymbol{\lambda})} \exp\left(Score(W_{i,j}, \boldsymbol{\lambda})\right) \tag{10}$$

其中 $Z(x_i, \boldsymbol{\lambda}) = \sum_{j=1}^{M} \exp\left(Score\left(W_{i,j}, \boldsymbol{\lambda}\right)\right)$ 代表所有候選詞序列的指數排序分數總和，爲一正規化的常數。全域條件式對數線性模型希望給定一訓練語句 x_i，其最低錯誤率詞序列 W_i^R 的對數條件機率能越大越好。因此，根據式(10)的定義，全域條件式對數線性模型的目標函數可表示成：

$$F_{GCLM}(\boldsymbol{\lambda}) = \sum_{i=1}^{L} \log p_\lambda(W_i^R \mid x_i) = \sum_{i=1}^{L} \log \frac{\exp\left(Score(W_i^R, \boldsymbol{\lambda})\right)}{\sum_{j=1}^{M} \exp\left(Score(W_{i,j}, \boldsymbol{\lambda})\right)} \tag{11}$$

我們亦可將式(11)視爲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，式(10)所表示的目標函數其實跟其它鑑別式訓練方法非常相像：例如，它與最大化交互資訊估測(Maximum Mutual Information Estimation, MMIE)[16]的目標函數有相同的表示式；而最小分類錯誤(Minimum Classification Error, MCE) [17]則可看做此方法的延伸。

在實作時，爲了避免過度訓練(Overtraining)，我們可在式(11)加上一權重參數的零均值高斯事前機率(Zero-Mean Gaussian Prior Probability)[10, 18]：

$$F_{GCLM}(\boldsymbol{\lambda}) = \sum_{i=1}^{L} \log \frac{\exp\left(Score(W_i^R, \boldsymbol{\lambda})\right)}{\sum_{j=1}^{M} \exp\left(Score(W_{i,j}, \boldsymbol{\lambda})\right)} - \frac{\|\boldsymbol{\lambda}\|^2}{2\sigma^2} \tag{12}$$

其中係數 σ 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(Development Set)估算其值。而最佳權重參數需符合 $\boldsymbol{\lambda}^* = \arg\max_{\lambda} F_{GCLM}(\boldsymbol{\lambda})$，因爲 $F_{GCLM}(\boldsymbol{\lambda})$ 爲一凸函數(Convex Function)，所以可求得 $F_{GCLM}(\boldsymbol{\lambda})$ 的全域最佳解(Globally Optimal Solution)。爲此，我們可以利用梯度下降法將此目標函數對每一維權重參數 λ_d 偏微分後即可求得權重參數向量 $\boldsymbol{\lambda}^*$ 每一維的調整量：

$$\frac{\partial F_{GCLM}(\boldsymbol{\lambda})}{\partial \lambda_d} = \sum_{i=1}^{L}\left[f\left(W_i^R\right) - \sum_{k=1}^{M} \frac{\exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M}\exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)} f_d\left(W_{i,k}\right)\right] - \frac{\lambda_d}{\sigma^2} \tag{13}$$

於是對每一維特徵分別更新其權重參數，以求得最佳權重參數向量：

$$\hat{\lambda}_d = \lambda_d + \eta \cdot \sum_{i=1}^{L}\left[f\left(W_i^R\right) - \sum_{k=1}^{M} \frac{\exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M}\exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)} f_d\left(W_{i,k}\right)\right] - \frac{\lambda_d}{\sigma^2} \tag{14}$$

(三)、考慮樣本權重(Sample Weight)之鑑別式語言模型

1、權重式全域條件式對數線性模型(Weighted Global Conditional Log-Linear Model, WGCLM)

權重式全域條件式對數線性模型[19]是將全域條件式對數線性模型加入樣本權重作延伸。因此，它的目標函數定義爲：

$$F_{WGCLM}(\boldsymbol{\lambda}) = \sum_{i=1}^{L} \log \frac{\exp\left(Score\left(W_i^R,\boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M}\omega_{i,W_{i,j}}\exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)} \tag{15}$$

我們可發現與全域條件式對數線性模型不同點在於，每一條候選詞序列 $W_{i,j}$ 會因其樣本權重 $\omega_{i,W_{i,j}}$ 不同而有不同的重要性。另外，與全域條件式對數線性模型相同的是，爲了避免在參數調整過程中發生過度訓練現象，我們可加入零均值高斯事前機率於式(15)而得：

$$F_{WGCLM}(\boldsymbol{\lambda}) = \sum_{i=1}^{L} \log \frac{\exp\left(Score\left(W_i^R,\boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M}\omega_{i,W_{i,j}}\exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)} - \frac{\|\boldsymbol{\lambda}\|^2}{2\sigma^2} \tag{16}$$

因此，最佳權重參數向量需符合 $\boldsymbol{\lambda}^* = \arg\max_{\lambda} F_{WGCLM}(\boldsymbol{\lambda})$。同時，從式(16)我們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；於是，可以求得 $\boldsymbol{\lambda}^*$ 的全域最佳解。與全域條件式對數線性模型相同的是，透過梯度下降法將此

目標函數對每一維權重參數 λ_d 偏微分後即可求得權重參數向量 $\boldsymbol{\lambda}$ 的調整量：

$$\frac{\partial F_{WGCLM}(\boldsymbol{\lambda})}{\partial \lambda_d} = \sum_{i=1}^{L}\left[f\left(W_i^R\right) - \sum_{k-1}^{M} \frac{\omega_{i,W_{i,k}} \exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M} \omega_{i,W_{i,j}} \exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)} f_d\left(W_{i,k}\right) \right] - \frac{\lambda_d}{\sigma^2} \tag{17}$$

因此，我們可對每一維特徵分別更新其權重，以求得較佳權重參數向量 $\boldsymbol{\lambda}$ ：

$$\hat{\lambda}_d = \lambda_d + \eta \cdot \sum_{i=1}^{L}\left[f\left(W_i^R\right) - \sum_{k=1}^{M} \frac{\omega_{i,W_{i,k}} \exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)}{\sum_{j=1}^{M} \omega_{i,W_{i,j}} \exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)} f_d\left(W_{i,k}\right) \right] - \frac{\lambda_d}{\sigma^2} \tag{18}$$

2、最小化錯誤率訓練(Minimum Error Rate Training, MERT)

最小化錯誤率訓練[20, 21]的排序減損函數被定義爲：

$$F_{MERT}(\boldsymbol{\lambda}) = \sum_{i=1}^{L}\sum_{k=1}^{M} \frac{\omega_{i,W_{i,k}} \exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right) - Score\left(W_i^R,\boldsymbol{\lambda}\right)\right)^{\beta}}{\sum_{j=1}^{M} \exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right) - Score\left(W_i^R,\boldsymbol{\lambda}\right)\right)^{\beta}} \tag{19}$$

經運算後，我們可很技巧地將式(19)中的 $\exp\left(Score\left(W_i^R,\lambda\right)\right)^{\beta}$ 項消去，使排序減損函數化簡爲：

$$F_{MERT}(\boldsymbol{\lambda}) = \sum_{i=1}^{L}\sum_{k=1}^{M} \omega_{i,W_{i,k}} \frac{\exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)^{\beta}}{\sum_{j=1}^{M} \exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)^{\beta}} \tag{20}$$

若式(20)中的變數 $\omega_{i,W_{i,k}}$ 設爲每一條候選詞序列 $W_{i,k}$ 的錯誤率時，則此排序減損函數可視爲前 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爲最小化前 M 條最佳辨識候選詞序列錯誤率的期望值。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。另外，從式(20)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向量需符合 $\boldsymbol{\lambda}^* = \arg\min_{\lambda} F_{MERT}(\boldsymbol{\lambda})$；因此，我們可對式(20)的每一維權重參數 λ_d 偏微分而可求得權重參數向量 $\boldsymbol{\lambda}$ 的權重調整量：

表一、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
感知器演算法	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

$$\frac{\partial F_{MERT}(\boldsymbol{\lambda})}{\partial \lambda_d} = \sum_{i=1}^{L}\sum_{k=1}^{M} \omega_{i,W_{i,k}} \cdot \beta \cdot \exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)^{\beta} \cdot \frac{\sum_{j=1}^{M} \exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)^{\beta}\left(f_d\left(W_{i,j}\right) - f_d\left(W_{i,k}\right)\right)}{\left(\sum_{j'=1}^{M} \exp\left(Score\left(W_{i,j'},\boldsymbol{\lambda}\right)\right)^{\beta}\right)^2} \tag{21}$$

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：

$$\hat{\lambda}_d = \lambda_d - \eta \cdot \sum_{i=1}^{L}\sum_{k=1}^{M} \omega_{i,W_{i,k}} \cdot \beta \cdot \exp\left(Score\left(W_{i,k},\boldsymbol{\lambda}\right)\right)^{\beta} \cdot \frac{\sum_{j=1}^{M} \exp\left(Score\left(W_{i,j},\boldsymbol{\lambda}\right)\right)^{\beta}\left(f_d\left(W_{i,j}\right) - f_d\left(W_{i,k}\right)\right)}{\left(\sum_{j'=1}^{M} \exp\left(Score\left(W_{i,j'},\boldsymbol{\lambda}\right)\right)^{\beta}\right)^2} \tag{22}$$

(四)、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(或目標)函數進行討論與比較。其中，沒有考慮樣本權重的有感知器演算法和全域條件式對數線性模型；而考慮樣本權重的有權重式全域條件式對數線性模型和最小化錯誤率訓練。各種鑑別式

語言模型的排序減損(目標)函數都是基於不同的訓練目標：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 $\omega_{i,W_{i,j}}$ ，對每一條候選詞序列加入其錯誤率或排序作爲其樣本權重，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

$$\sum_{k=1}^{M}\frac{\exp\left(Score\left(W_{i,k},\lambda\right)\right)^{\beta}}{\sum_{j=1}^{M}\exp\left(Score\left(W_{i,j},\lambda\right)\right)^{\beta}}=1 \tag{23}$$

所以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參數向量可以使排序錯誤越小越好，即爲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)的能力可預期地會不若其它三種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。表一爲四種鑑別式語言模型的比較。

三、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參數向量的缺點；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地符合測試語句的特性。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爲 *K* 平均演算法(*K*-means)，假設所有語言模型訓練語料中的語句可分爲 *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。在語音辨識階段，

當有一測試語句 y_k 輸入時，我們利用 y_k 本身的特性(即 y_k 與每一群訓練語句的相似程度)，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個新的語言特徵權重參數向量。組合係數可經由計算測試語句與 P 群訓練語句的相似度而得(若測試語句 y_k 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)。本論文以餘弦值(Cosine Value)計算測試語句與訓練語句群的相似度；我們先對每一句測試語句 y_k 定義其單連詞向量 $\mathbf{u}_k$，$\mathbf{u}_k$ 爲結合 y_k 的 M 最佳辨識候選詞序列中每一條候選詞序列 $W_{k,j}$ 的單連詞向量 $\mathbf{u}_{k,j}$ 而成：

$$\mathbf{u}_k = \sum_{j=1}^{M} \mathbf{u}_{k,j} \tag{24}$$

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：

$$\mathbf{v}_p = \frac{\sum_{j=1}^{L_p} \mathbf{v}_{p,j}}{L_p} \tag{25}$$

其中 L_p 爲 p 訓練語句群中包含的訓練語句的句數。因此，測試語句 y_k 與 p 訓練語句群的組合係數爲：

$$\gamma_{k,p} = \frac{\cos_{k,p}}{\sum_{c=1}^{P} \cos_{k,c}} \tag{26}$$

其中，$\cos_{k,p}$ 即爲 $\mathbf{u}_k$ 與 $\mathbf{v}_p$ 的餘弦值：

$$\cos_{k,p} = \frac{\mathbf{u}_k \cdot \mathbf{v}_p}{\sqrt{\mathbf{u}_k^{\,2}}\sqrt{\mathbf{v}_p^{\,2}}} \tag{27}$$

最後，對測試語句 y_k 來說，其新的語言特徵權重參數向量 $\boldsymbol{\lambda}_k$ 可表示成爲 P 個語言特徵權重參數向量的線性組合：

$$\boldsymbol{\lambda}_k = \sum_{p=1}^{P} \gamma_{k,p} \cdot \boldsymbol{\lambda}_p \tag{28}$$

表二、訓練語料與測試語料之統計資訊

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

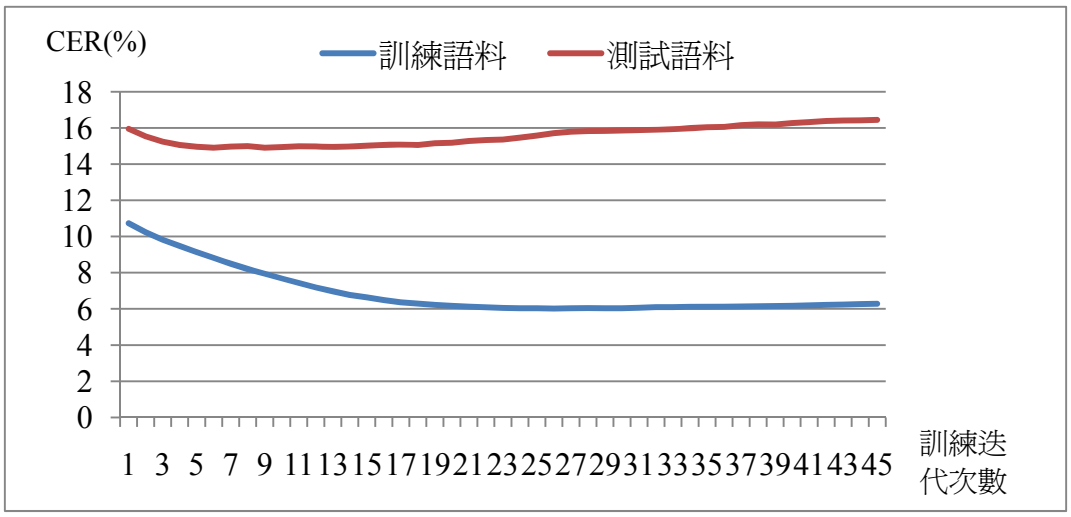
表三、基礎實驗結果(CER(%))

	訓練語料 辨識錯誤率	絕對 提昇率	相對 提昇率	測試語料 辨識錯誤率	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

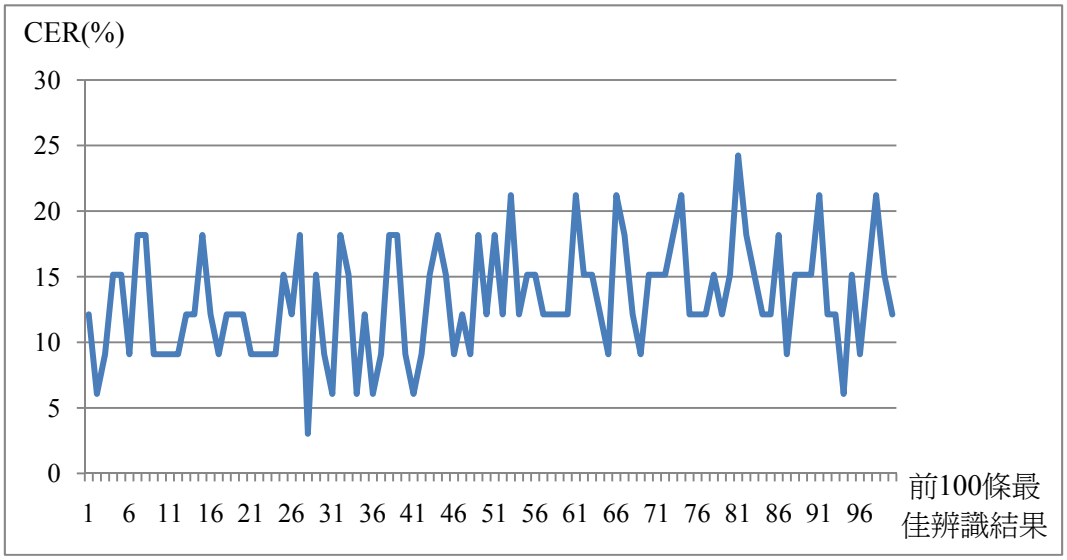
四、實驗結果與討論

(一)、實驗語料

本論文語音辨識實驗所用的語音語料來自公視新聞(Mandarian Across Taiwan─Broadcast News, MATBN)[22]。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與台灣公共電視台合作蒐集而成；每一篇新聞報導包含內場與外場兩個部分：內場爲主播(Studio Anchors)語料，外場則有採訪記者(Field Reporters)語料與受訪者(Interviewees)語料。另一方面，對於實驗所使用的背景三連語言模型(Background Trigram Language Model)，其訓練語料則是來自中央通訊社 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。我們採用 SRI Language Modeling Toolkit[23]訓練實驗所需要的三連語言模型。而訓練鑑別式語言模型所需的語音語料是取自公視新聞，共 30,600 句，約 23 小時。最後，語音辨識之測試語料與鑑別式語言模型訓練語料屬於同時期(亦是取自公視新聞)，共 1,997 句(約 1.5 小時)，如表二所示。



圖四、感知器演算法訓練語料與測試語料字錯誤率趨勢圖



圖五、訓練語句 100 條最佳辨識結果之字錯誤率

(二)、基礎實驗結果

首先，我們比較本論文所介紹的幾種鑑別式語言模型訓練演算法的基礎實驗結果，如表三所示。本論文實驗所用的基礎辨識器(僅使用背景三連語言模型)的字錯誤率(Character Error Rate, CER)，在訓練語料爲 11.26，在測試語料則爲 16.39。由表三可看出，本論文所介紹的幾種語言模型訓練演算法當中，以最小化錯誤率訓練(MERT)在測試語料有最佳的效能，接著依序爲權重式全域條件式對數線性模型(WGCLM)、全域條件式對數線性模型(GCLM)及感知器演算法(Percepton)。

從表三亦可看出，雖然感知器演算法在訓練語料可達到最佳的語音辨識率，但當它使用在測試語料時，其效果卻是較其它鑑別式語言模型來得低；這顯示出

感知器演算法的一般化(Generalization)能力較其它鑑別式語言模型來得差。爲了探究其原因，我們進一步觀察了感知器演算法在每一次訓練迭代時，訓練語料與測試語料字錯誤率表現之變化趨勢，如圖四所示。由圖四可看出，對於測試語料而言，感知器演算法會在較少的訓練迭代次數時就會達到最低字錯誤率，相較於它在訓練語料上的表現。因此，我們若將由訓練語料所獲得的最佳參數權重使用在測試語料上時，反而無法讓測試語料達到最佳的語音辨識效果。造成此情況的原因應爲感知器演算法在訓練過程中容易發生過度訓練(Overtraining)的情況，導致訓練後的最佳權重參數向量過於符合訓練語料；使得當訓練好的語言模型的權重參數向量被使用在測試語料時，並無法發揮對應的效果。

另一方面，當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時，我們可以發現全域條件式對數線性模型較感知器演算法能有更顯著的語音辨識率提昇。探究其原因，我們推測應爲全域條件式對數線性模型在更新權重時，考慮了訓練語句中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。相較之下，全域條件式對數線性模型在訓練時考慮的層面應更爲周全。

接下來，我們比較兩種考慮樣本權重的鑑別式訓練模型，也就是權重式全域條件式對數線性模型與最小化錯誤率訓練；它們皆使用了兩種不同的樣本權重來進行訓練，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，皆以字錯誤率爲樣本權重的方式會較排序爲樣本權重的方式來的有效果。爲探究其原因，我們取了訓練語料其中一句訓練語句的前 100 條最佳候選詞序列來觀察每一條候選詞序列的字錯誤率，如圖五所示。由圖五可看出，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，若以排序爲樣本權重，擁有相同的字錯誤率的候選詞序列也會有不同的排序，因而影響了以排序爲樣本權重的方式的鑑別能力，導致重新排序結果較差。若將上述兩種考慮樣本權重的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，則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，因爲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。

再者，當我們比較最小化錯誤率訓練與其他方法的實驗結果時，可看出最小化錯誤率訓練獲得較佳的語音辨識率。其原因除了是樣本權重的幫助外，還可能是由於最小化錯誤率訓練的目標函數(參考式(20))並沒有用到最低錯誤率詞序列作爲訓練標準，而是考慮所有候選詞序列的影響，不像其它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。因此，最小化錯誤率訓練可得到更具一般化的權重向量，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。

表四、語句相關之鑑別式語言模型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
5	15.00	1.39	8.49
10	15.11	1.28	7.83
原始	15.71	0.68	4.15

表五、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
5	15.48	0.91	5.57
10	15.67	0.72	4.41
原始	15.53	0.86	5.26

(三)、語句相關之鑑別式語言模型

本節討論語句相關之鑑別式語言模型之實驗結果，初步分別以感知器演算法及全域條件式對數線性模型做爲鑑別式語言模型爲基礎來建立語句相關之鑑別式語言模型。表四與表五比較兩種鑑別式語言模型所使用訓練語料的分群個數與在語音辨識時字錯誤率之間的關係，它們的最後一行皆爲傳統鑑別式語言模型的結果。比較這兩種語句相關之鑑別式語言模型，可看出訓練語料分群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爲佳。探究其原因可能爲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，所以可以減輕傳統感知器演算法容易過度訓練的問題。另外，我們也可看出分群數目過多時，可能會導致語句相關之鑑別式語言模型的效能下降；探究其原因應爲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。

(四)、結合不同測試語句訓練權重與所有訓練語句訓練權重

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 $\boldsymbol{\lambda}_j$ 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 $\boldsymbol{\lambda}^{all}$ 作線性組合，得到一組新的權重向量 $\boldsymbol{\lambda}_j^{combine}$，以期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(29)所示：

表六、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(感知器演算法、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
0.1	15.41	0.98	6.01
0.2	15.18	1.21	7.37
0.3	15.06	1.33	8.11
0.4	14.96	1.43	8.71
0.5	14.86	1.53	9.36
0.6	14.79	1.60	9.78
0.7	14.81	1.58	9.63
0.8	14.74	1.65	10.04
0.9	14.88	1.51	9.23

表七、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(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
0.1	15.32	1.07	6.56
0.2	15.07	1.32	8.05
0.3	14.94	1.45	8.84
0.4	14.77	1.62	9.89
0.5	14.59	1.80	11.01
0.6	14.45	1.94	11.82
0.7	14.46	1.93	11.77
0.8	14.56	1.83	11.14
0.9	14.87	1.52	9.25

$$\boldsymbol{\lambda}_j^{combine} = \alpha \cdot \boldsymbol{\lambda}_j + (1-\alpha) \cdot \boldsymbol{\lambda}^{all} \tag{29}$$

其中α爲調整係數，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$\boldsymbol{\lambda}_j$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$\alpha$的值是介於 0.1 到 0.9 之間。

表六與表七爲使用感知器演算法的實驗結果，而表八與表九則爲使用全域條件式對數線性模型的實驗結果。由表六與表七可看出，對於使用感知器演算法的

表八、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(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
0.1	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

表九、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(全域條件式對數線性模型、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
0.1	15.43	0.96	5.88
0.2	15.37	1.02	6.23
0.3	15.38	1.01	6.14
0.4	15.39	1.00	6.12
0.5	15.42	0.97	5.94
0.6	15.37	1.02	6.21
0.7	15.41	0.98	5.97
0.8	15.45	0.94	5.72
0.9	15.56	0.83	5.09

實驗結果，當分群個數爲 10 且α設爲 0.6 時字錯誤率爲 14.45，可達到相對辨識率 11.82%的顯著提昇效果，比起單獨使用所有訓練語句訓練權重的字錯誤率 15.71 或單獨使用不同測試語句訓練權重的字錯誤率 15.11 都要好的效果。而從表八與表九則可看出當分群個數爲 5 且α設爲 0.4 時，全域條件式對數線性模型的實驗結果爲 15.29，可達到相對辨識率 6.71%，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可達到互補的效果，因而達到顯著的提昇。

另外，我們若觀察在最低字錯誤率所設的α值時，可發現對於感知器演算法α值爲 0.6，全域條件式對數線性模型爲 0.4；也就是說，在感知器演算法中，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 $\boldsymbol{\lambda}^{all}$ 較爲重要。探究其原因應爲感知器演算法易有過度訓練的問題；因爲語句相關之鑑別式語言模型所訓練的特徵權重參數向量因爲有考慮測試語句本身的特性，所以加重它的重要性似乎可以減輕特徵權重參數向量過度符合訓練語料的問題。另一方面，對於全域條件式對數線性模型，所有訓練語句訓練的特徵權重則顯的較爲重要，可能是因爲要減緩訓練語料分群後導致訓練資料量不足的問題。

六、結論與未來展望

在語音辨識領域中，語言模型有著舉足輕重的地位，傳統的 *N* 連語言模型以找到語音訊號對應機率最高的詞序列爲目標；近年來新興的鑑別式語言模型以直接降低錯誤率爲訓練目標。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，探討它們的表現與實驗結果所代表的意義。我們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。從實驗結果可發現，此方法可達到比所有測試語句皆爲相同權重向量更好的效果，因其考慮了不同的測試語句的特性，給予其不同的權重向量。未來，我們將研究語句相關之鑑別式語言模型不同的組合係數計算方法，以其達到更加的效果。

參考文獻

[1] F. Jelinek, *Statistical Methods for Speech Recognition*, the MIT press, 1999.

[2] J.-W. Kuo, B. Chen, "Minimum Word Error Based Discriminative Training of Language Models," in *Proceedings of International Conference on Spoken Language Processing*, pp. 1277-1280, 2005.

[3] J. Gao, H. Suzuki, and W. Yuan, “An empirical study on language model adaptation,” *ACM Transactions on Asian Language Information Processing*, Vol. 5, No. 3, pp. 209–227, 2006.

[4] B. Roark, M. Saraclar, and M. Collins, “Discriminative n-gram language modeling,” *Computer Speech and Language*, Vol. 21, No. 2, pp. 373-392, 2007.

[5] M. Collins, T. Koo, “Discriminative reranking for natural language parsing,” *Computational Linguistics*, Vol. 31, No. 1, pp. 25-70, 2005.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley, 2001.

[7] M. Collins, “Parameter estimation for statistical parsing models: Theory and practice of distribution-free methods,” in H. Bunt, J. Carroll, G. Satta (eds.), *New Developments in Parsing Technology*, 2004.

[8] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain,” *Psychological Review*, No. 6, pp. 386-408, 1958.

[9] M. Collins, “Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1-8, 2002.

[10] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proceedings of International Conference on Machine Learning*, pp. 282-289, 2001.

[11] F. Sha, F. Pereira, “Shallow parsing with conditional random fields,” in *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 134-141, 2003.

[12] T. M. Mitchell, *Machine Learning.* The McGraw-Hill Companies, 1997.

[13] Y. Freund and R.E. Schapire, “Large margin classification using the perceptron algorithm,” *Machine Learning*, Vol. 37, No. 3, pp. 277-296, 1999.

[14] A. Ratnaparkhi, S. Roukos, and R.T. Ward, “A maximum entropy model for parsing,” in *Proceedings of International Conference on Spoken Language Processing*, pp. 803-806, 1994.

[15] M. Johnson, S. Geman, S. Canon, Z. Chi, and S. Riezler, “Estimators for stochastic “unification-based” grammars,” in *Proceedings of The annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 535–541, 1999.

[16] 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,” in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 49-52, 1986.

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

[18] 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,” in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 141-144, 2006.

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

Proceedings of International Conference on Acoustics, Speech and Signal Processing, pp. 5126-5129, 2010.

[20] F.J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of The annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 160-167, 2003.

[21] K. Akio, O. Takahiro, H. Shinichi, S. Shoei, I. Toru, and T. Tohru, "Discriminative rescoring based on minimization of word errors for transcribing broadcast news," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1574-1577, 2008.

[22] H.-M. Wang, B. Chen, J.-W. Kuo, and S.-S. Cheng, "MATBN: A Mandarin Chinese broadcast news corpus," *International Journal of Computational Linguistics & Chinese Language Processing*, Vol. 10, No. 2, pp. 219-236, 2005.

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