實證探究多種鑑別式語言模型於語音辨識之研究

Empirical Comparisons of Various Discriminative Language Models for Speech Recognition

賴敏軒 [1], 黃邦烜 [1], 陳冠宇 [2], 陳柏琳 [1]
[1] 國立臺灣師範大學資訊工程學系
{698470623, 699470204, berlin}@ntnu.edu.tw
[2] 中央研究院資訊科學研究所
kychen@iis.sinica.edu.tw

摘要

傳統語言模型(Language Models)是藉由使用大量的文字語料訓練而成，以機率模型來描述自然語言的規律性。*N* 連(*N*-gram)語言模型是最常見的語言模型，被用來估測每一個詞出現在已知前 *N*-1 個歷史詞之後的條件機率。此外，傳統語言模型大多是以最大化相似度爲訓練目標；因此，當它被使用於語音辨識上時，對於降低語音辨識錯誤率常會有所侷限。近年來，有別於傳統語言模型的鑑別式語言模型(Discriminative Language Model)陸續地被提出；與傳統語言模型不同的是，鑑別式語言模型是以最小化語音辨識錯誤率做爲訓練準則，期望所訓練出的語言模型可以幫助降低語音辨識的錯誤率。本論文探究基於不同訓練準則的鑑別式語言模型，分析各種鑑別式語言模型之基礎特性，並且比較它們被使用於大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)時之效能。同時，本論文亦提出將邊際(Margin)概念引入於鑑別式語言模型的訓練準則中。實驗結果顯示，相較於傳統 *N* 連語言模型，使用鑑別式語言模型能對於大詞彙連續語音辨識有相當程度的幫助；而本論文所提出的基於邊際資訊之鑑別式語言模型亦能夠進一步地提升語音辨識的正確率。

關鍵詞：語音辨識、鑑別式語言模型、邊際、訓練準則

一、緒論

在人與人的互動當中，語音是最自然且直接的表達方式之一。透過語音，人們可以彼此溝通，傳達想法、感受以及情緒。因此，我們期望能讓電腦具備與人溝通的能力，能爲生活帶來便利性。要達到此目標，我們必須先對使用者輸入的語音訊號進行辨識；待轉換成文字後，再對文字所欲表達的語意作理解，進而做出最適當的動作來回應使用者。將語音訊號轉換成文字的過程，可以透過自動語音辨識(Automatic Speech Recognition, ASR)技術來完成。在自動語音辨識的過程中，我們必須先將語音訊號做特徵擷取(Feature Extraction)，保留語音訊號中的聲學特性(Acoustic Characteristics)，並轉換成能使電腦容易處理的聲學特徵向量(Acoustic Feature Vector)；利用這些聲學特徵向量，我

們可以爲不同的音素(Phoneme)分別建立聲學模型(Acoustic Model)，進而產生可能的候選詞序列(Candidate Word Sequences)。另一方面，我們也必須收集大量的文字訓練語料，用以統計自然語言中各種詞序列的出現情形，並藉此訓練語言模型(Language Model)。傳統語言模型是收集各種詞彙出現在自然語言中的詞頻數，經由最大化相似度估測(Maximum Likelihood Estimation, MLE)來建立語言模型。例如，*N* 連(*N*-gram)語言模型[1]是估測每一個詞在其前面緊鄰 *N*-1 個歷史詞序列已知情況下的條件機率；它可協助語音辨識器從所產生的候選詞序列中，選取機率最高(最可能)的詞序列做爲最後的語音辨識結果。

利用傳統語言模型(例如*N*連語言模型)所選出的語音辨識結果通常是發生機率最高的詞序列，但未必是最佳(錯誤率最低)的；換句話說，在候選詞序列中其實有可能存在著其它擁有較低錯誤率的詞序列可以做爲語音辨識器的輸出。於是，我們希望能透過使用更多其它語言特徵，以及候選詞序列所提供的資訊，並經適當訓練的語言模型將所有候選詞序列做重新排序(Reranking)，以輸出擁有較低錯誤率的語音辨識結果。近年來，有許多學者採用鑑別式訓練(Discriminative Training)的概念來訓練語言模型以幫助重新排序。與傳統語言模型不同，鑑別式語言模型(Discriminative Language Model)[2, 3, 4]是以最小化語音辨識錯誤率爲訓練目標，藉由一組預先定義的語言特徵以及所對應的特徵權重參數，將所有候選詞序列(存在於詞圖或 *M* 條最佳辨識候選詞序列)重新計分(Rescoring)或重新排序(Reranking)，期望使具有最低錯誤率的候選詞序列能擁有最高的分數(排序)，並且做爲最後的輸出結果。

本論文延續我們先前對於鑑別式語言模型之研究[5, 6]，探究基於不同訓練準則的鑑別式語言模型，分析各種鑑別式語言模型之基礎特性，並提出將邊際(Margin)概念引入於鑑別式語言模型的訓練準則中。本論文的安排如下：第二節將介紹近年來常見的、基於不同訓練準則的鑑別式語言模型；第三節將說明本論文所提出基於邊際資訊之鑑別式語言模型；第四節是實驗結果與分析；第五節則是結論與未來展望。

二、鑑別式語言模型介紹

(一)、鑑別式語言模型訓練之定義

一般來說，鑑別式語言模型是以最小化辨識錯誤率爲訓練目標，希望對基礎語音辨識器(Baseline Speech Recognizer)所產生的候選詞序列(如前 *M* 條最佳辨識結果)作重新排序，使得具有較低辨識錯誤率的候選詞序列能擁有較高的排序。而重新排序的依據則是以基礎語音辨識器的辨識分數做爲基礎，並加上額外定義的語言特徵向量，藉由前述兩者與其對應的特徵權重參數向量做內積後的語言模型分數來進行排序，使得前 *M* 條最佳辨識候選詞序列中最低錯誤率的詞序列能擁有最高的語言模型分數。以下將對鑑別式訓練所需的參數做定義：

(a) 給定一句語音訊號 x_i，其經由基礎辨識器所產生的 *M* 條最佳候選詞序列集合爲 $GEN(x_i)=\{W_{i,j}\}$，其中 *j* 爲1到 *M* 之間。

(b) 將訓練語料視爲 $\{x_i, W_i^R\}$的集合，其中 *i* 的值介於1到 *L* 之間， *L* 爲訓練語料的總句數； W_i^R 爲語音訊號 x_i 在其對應 *M* 條最佳候選詞序列中最低錯誤率之詞序列。

(c) 對於每一條候選詞序列定義一組 $D+1$ 維的特徵向量 $f_d(W_{i,j})$，其中 d 是從 0 到 D 之間； $f_0(W_{i,j})$ 爲基礎辨識器所產生的分數，即爲聲學模型與 N 連語言模型的對數機率(Log Probability)分數總和，在此我們使用三連(Trigram)語言模型；而其它維度 d ，可分別表示每一條候選詞序列 $W_{i,j}$ 中各種 N 連詞出現的次數(視爲一種語言特徵)，以 $f_d(W_{i,j})$ 來表示，本論文所定義各種可能的語言特徵爲單連詞(Word Unigram)與雙連詞(Word Bigram)。

(d) 定義一組 $D+1$ 維的特徵權重參數向量 $\lambda=[\lambda_0,\lambda_1,\ldots,\lambda_d,\ldots,\lambda_D]$，其中每一個特徵權重參數 λ_d 分別對應於每一個語言特徵 $f_d(W_{i,j})$。

因此，候選詞序列 $W_{i,j}$ 的重新排序分數可表示爲：

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

而經由重新排序後分數最高的候選詞序列 W_i^* 即做爲最後的輸出結果：

$$W_i^*=\arg\max_{W_{i,j}\in GEN(x_i)} Score(W_{i,j},\lambda) \tag{2}$$

鑑別式語言模型的訓練在於求取最佳的特徵權重參數向量 λ，期望使得測試語句的前 M 條最佳辨識候選詞序列中最低錯誤率的詞序列能在式(2)擁有最高的分數。

(二)、常見的鑑別式語言模型

鑑別式語言模型早期大多都使用在其它的應用領域上：例如，機器翻譯(Machine Translation, MT)、自然語言處理(Natural Language Processing, NLP)等。近十年來，陸續有許多學者將各種基於不同訓練準則的鑑別式語音模型介紹到大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)來使用。鑑別式語言模型的訓練可分成三個面向來探討，分別爲訓練語料、訓練準則與特徵。以下將介紹常見的各種鑑別式語言模型，並將它們依其訓練準則區分爲以下四類：最小化平方誤差、最小化錯誤率期望值、最大化對數條件機率、以及考量語句之間彼此之關係。

1、最小化平方誤差

感知器演算法(Perceptron)[7]早期是被應用在人工類神經網路(Artificial Neural Network)領域中；在 2002 年，美國學者 Collins[8]將感知器演算法應用在自然語言處理領域中。感知器演算法可視爲是最大化熵值法(Maximum-Entropy, ME)或條件式隨機域(Conditional Radom Fields, CRF)[9, 10]的一種變形。感知器演算法以最小平方誤差(Least Squared Error, LSE)[11]爲觀念，透過最小化其訓練目標函數(Training Objective) $F_{\text{Perc}}(\lambda)$ 以求得最佳的特徵權重參數向量 $\hat{\lambda}$ ：

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

爲了求得 $\hat{\lambda}$ ，我們可以利用梯度下降法(Gradient Descent Method)將 $F_{Perc}(\lambda)$ 的每一個維

1 Initialize all parameters in the model, i.e. $\lambda_0 = 1$ *and* $\lambda_d = 0$ for $d = 1,..., D$

2 For $t = 1...T$ where T is the total number of iterations

3 For each training sample $(\mathrm{x}_i, W_i^R), i = 1,..., L$

4 Use current model λ to choose the W_i^* from $GEN(x_i)$

5 For $d = 1,..., D$

6 $\hat{\lambda}_d = \lambda_d + \eta \cdot \left(f_d(W_i^R) - f_d(W_i^*)\right)$ where η is the size of the learning step

圖一、感知器演算法[12]

度特徵權重參數 λ_d 分別做偏微分，由於 $F_{\mathrm{Perc}}(\lambda)$ 可能存在許多局部最佳解(Local Minimum Solutions)，而使用梯度下降法並無法保證可求得全域最佳解(Global Minimum Solutions)。因此，感知器演算法採取隨機近似法(Stochastic Approximation)，即對每一句訓練語句的每一維特徵權重參數分別求最佳解，求得每一維特徵權重參數的調整量：

$$\hat{\lambda}_d = \lambda_d - \eta \cdot \left(Score(W_i^R, \lambda) - Score(W_i^*, \lambda)\right) \cdot \left(f_d(W_i^R) - f_d(W_i^*)\right) \tag{4}$$

其中 η 爲學習步調常數(Learning Step Size)。除了式(4)此種特徵權重參數更新式之外，也有學者提出省略 $Score(W_i^R, \lambda) - Score(W_i^*, \lambda)$ 項，將更新式簡化爲 $\hat{\lambda}_d = \lambda_d + \eta \cdot \left(f_d(W_i^R) - f_d(W_i^*)\right)$ 來更新特徵權重參數。其演算法如圖一所示。

2、最小化錯誤率期望值

(1)、最小化錯誤率訓練(MERT)

最小化錯誤率訓練(Minimum Error Rate Training, MERT)是在 2003 年由學者 Och[13]提出，並且運用在機器翻譯(Machine Translation)領域中；在 2008 年由 Kobayashi 等學者[14]將最小化錯誤率訓練方法介紹到語音辨識領域中使用。應用於語音辨識時，其訓練準則定義成最小化基礎語音辨識器所產生的 M 條候選詞序列之錯誤率期望值，藉此找出一個最合適的語言模型特徵權重向量：

$$F_{\mathrm{MERT}}(\lambda) = \sum_{i=1}^{L} \sum_{k=1}^{M} \frac{\omega_{W_{i,k}} \cdot \exp\left(Score(W_{i,k}, \lambda) - Score(W_i^R, \lambda)\right)^{\beta}}{\sum_{j=1}^{M} \exp\left(Score(W_{i,j}) - Score(W_i^R, \lambda)\right)^{\beta}} \tag{5}$$

其中 $\omega_{W_{i,k}}$ 爲候選詞序列 $W_{i,k}$ 的錯誤率(Error Rate)；而 β 爲一平滑化參數。透過進一步的數學推導，我們可以將式(5)中 $\exp\left(Score(W_i^R, \lambda)\right)^{\beta}$ 項提出而簡化成：

$$F_{\mathrm{MERT}}(\lambda) = \sum_{i=1}^{L} \sum_{k=1}^{M} \frac{\omega_{W_{i,k}} \cdot \exp\left(Score(W_{i,k}, \lambda)\right)^{\beta}}{\sum_{j=1}^{M} \exp\left(Score(W_{i,j}, \lambda)\right)^{\beta}} \tag{6}$$

再將式(6)針對每一維特徵權重參數 λ_d 做偏微分，得其調整量爲：

$$\hat{\lambda}_d = \lambda_d +$$

$$\eta \cdot \sum_{i=1}^{L}\sum_{k=1}^{M} \omega_{W_{i,k}} \cdot \beta \cdot \exp\left(Score\left(W_{i,k},\lambda\right)\right)^{\beta} \cdot \frac{\sum_{j=1}^{M} \exp\left(Score\left(W_{i,j},\lambda\right)\right)^{\beta}\left(f_d\left(W_{i,k}\right) - f_d\left(W_{i,j}\right)\right)}{\left(\sum_{j'=1}^{M} \exp\left(Score\left(W_{i,j'},\lambda\right)\right)^{\beta}\right)^2} \quad (7)$$

其中η爲學習步調常數。我們可以將最小化錯誤率訓練中錯誤率$\omega_{W_{i,k}}$視爲一種樣本權重(Sample Weight)資訊，用來區別每一個候選詞序列$W_{i,k}$對於鑑別式語言模型訓練時的重要性。

3、最大化對數條件機率

(1)、全域條件式對數線性模型(GCLM)

早期全域條件式對數線性模型(Global Conditional Log-linear Model, GCLM)被應用在自然語言處理領域中；2007年Roark等學者[4]以有限狀態機(Weighted Finite State Automata, WFSA)實作全域條件式對數線性模型於語音辨識結果的重新排序上，並且與感知器演算法進行比較。

全域條件式對數線性模型是希望在給定一句語音訊號x_i與所對應的 M 條最佳候選詞序列$GEN\left(x_i\right)$時，其中擁有最低辨識錯誤率的詞序列其對數條件機率可以越大越好，亦即最大化下列訓練目標函數：

$$F_{\mathrm{GCLM}}(\lambda) = \sum_{i=1}^{L} \log \frac{\exp\left(Score\left(W_i^R,\lambda\right)\right)}{\sum_{j=1}^{M} \exp\left(Score\left(W_{i,j},\lambda\right)\right)} \quad (8)$$

爲了避免過度訓練(Overtraining)，我們可以在目標函數$F_{\mathrm{GCLM}}(\lambda)$中加上一個權重參數的零均值高斯事前機率(Zero-Mean Gaussian Prior Probability)項：

$$F_{\mathrm{GCLM}}(\lambda) = \sum_{i=1}^{L} \log \frac{\exp\left(Score\left(W_i^R,\lambda\right)\right)}{\sum_{j=1}^{M} \exp\left(Score\left(W_{i,j},\lambda\right)\right)} - \frac{\|\lambda\|^2}{2\sigma^2} \quad (9)$$

因爲$F_{\mathrm{GCLM}}(\lambda)$爲一凸函數(Convex Function)，因此可以求得全域最佳解(Globally Optimal Solution)，爲求得最佳特徵權重參數向量$\hat{\lambda}$。將式(7)針對每一維特徵權重參數λ_d做偏微分，得其調整量爲：

$$\hat{\lambda}_d = \lambda_d + \eta \cdot \sum_{i=1}^{L}\left[f\left(W_i^R\right) - \sum_{k=1}^{M} \frac{\exp\left(Score\left(W_{i,k},\lambda\right)\right)}{\sum_{j=1}^{M} \exp\left(Score\left(W_{i,j},\lambda\right)\right)} \cdot f_d\left(W_{i,k}\right)\right] - \frac{\lambda_d}{\sigma^2} \quad (10)$$

(2)、權重式全域條件式對數線性模型(WGCLM)

不同於全域條件式對數線性模型(GCLM)，Oba 等學者[15]在 2010 年提出將樣本權重加入全域條件式對數線性模型進行改良，爲每一個候選詞序列的分數加上一個不同的權重，用來表示每一條候選詞序列不同的重要程度，此方法稱爲權重式全域條件式對數線性模型(Weighted Global Conditional Log-linear Model, WGCLM)。換句話說，每一個候選詞序列$W_{i,j}$都會有一個相對應的樣本權重$\omega_{W_{i,j}}$；根據不同的樣本權重來表示每一個候選詞序列對於語言模型訓練的不同重要性。其訓練目標函數可表示爲：

$$F_{\mathrm{WGCLM}}(\lambda)=\sum_{i=1}^{L}\log\frac{\exp\left(Score\left(W_i^R,\lambda\right)\right)}{\sum_{j=1}^{M}\omega_{W_{i,j}}\exp\left(Score\left(W_{i,j},\lambda\right)\right)} \tag{11}$$

同樣地，爲了避免在調整參數的過程中，發生過度訓練的問題，我們也可以加入一個零均值高斯事前機率項於權重式全域條件式對數線性模型的訓練目標函數中：

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

將式(12)針對每一維特徵權重參數λ_d做偏微分，得其調整量爲：

$$\hat{\lambda}_d=\lambda_d+\eta\cdot\sum_{i=1}^{L}\left[f\left(W_i^R\right)-\sum_{k=1}^{M}\frac{\omega_{W_{i,k}}\exp\left(Score\left(W_{i,k},\lambda\right)\right)}{\sum_{j=1}^{M}\omega_{W_{i,j}}\exp\left(Score\left(W_{i,j},\lambda\right)\right)}\cdot f_d\left(W_{i,k}\right)\right]-\frac{\lambda_d}{\sigma^2} \tag{13}$$

值得一提的是，樣本權重的$\omega_{W_{i,j}}$設計也是一個值得研究的議題，通常我們可以將每一個候選詞序列本身的錯誤率當成其樣本權重。

4、考量語句之間彼此之關係

(1)、輪轉雙重鑑別式模型 (R2D2)

全域條件式對數線性模型(GCLM)是期望最低錯誤率詞序列的對數條件機率能夠越大越好；Oba 等學者等針對全域條件式對數線性模型提出改良方法，在訓練目標函數中考慮了訓練語句所有候選詞序列彼此之間的關係，因而有所謂的輪轉雙重鑑別式模型(Round-Robin Dual Discrimination Model, R2D2)[16]。輪轉雙重鑑別式模型可以視爲是全域條件式對數線性模型(GCLM)的一種延伸；它因爲考量了兩兩候選詞序列彼此之間的關係，使得輪其擁有較好的一般化能力。同時，類似於權重式全域條件式對數線性模型(WGCLM)，輪轉雙重鑑別式模型也使用了樣本權重：

$$F_{\mathrm{R2D2}}(\lambda)=\sum_{i=1}^{L}\log\left\{\sum_{j'=1}^{M}\sum_{j=1}^{M}\frac{\exp\left(\sigma_1\omega_{W_{i,j}}\right)\exp\left(Score\left(W_{i,j},\lambda\right)\right)}{\exp\left(\sigma_2\omega_{W_{i,j'}}\right)\exp\left(Score\left(W_{i,j'},\lambda\right)\right)}\right\} \tag{14}$$

其中，σ_1與σ_2爲實驗參數。相同的，將式(14)針對每一維特徵權重參數λ_d做偏微分，得其調整量爲：

$$\hat{\lambda}_d = \lambda_d + \eta \cdot \sum_{i=1}^{L} \left\{ \frac{\left[\sum_{j'=1}^{M} \sum_{j=1}^{M} f_d(W_{i,j}) A \right] - \left[\sum_{j'=1}^{M} \sum_{j=1}^{M} f_d(W_{i,j'}) A \right]}{\sum_{j'=1}^{M} \sum_{j=1}^{M} A} \right\}, \quad (15)$$

$$\text{where } A = \exp\left(Score(W_{i,j}) - Score(W_{i,j'}) + \sigma_1 \omega_{W_{i,j}} - \sigma_2 \omega_{W_{i,j'}} \right)$$

(三)、鑑別式語言模型之特性

首先，感知器演算法(Perceptron)是以最小平方差爲其精神，希望排序分數最高的候選詞序列與最低錯誤率之詞序列(亦即參考詞序列)的分數差平方後越小越好；然而感知器演算法只考慮目前排序分數最高的詞序列與最低錯誤率詞序列之間的關係，因此其一般化(Generalization)的能力不是很好，很容易就會有過度訓練(Over-Training)的問題。

相較於感知器演算法，最小化錯誤率訓練(MERT)的精神是希望語音辨識所產生的候選詞序列其錯誤率期望值越小越好。因此，在訓練的過程中，它不僅僅考慮分數最高與擁有最低錯誤率的詞序列，更同時考慮了其它候選詞序列，故其會有較佳的一般化能力。但也因爲同時考慮了所有候選詞序列，導致在訓練的速度上相對較慢。

全域條件式對數線性模型(GCLM)的訓練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；因爲全域條件式對數線性模型考慮到最低錯誤率詞序列與其它所有候選詞序列的關係，因此其一般化的能力會比感知器演算法來的好，比較不會有過度訓練的問題出現。

權重式全域條件式對數線性模型(WGCLM)是全域條件式對數線性模型(GCLM)之延伸，差別在於權重式全域條件式對數線性模型的分母項多考慮了樣本權重$\omega_{W_{i,j}}$，目的是讓每條候選詞序列對於訓練有不同的影響力。我們可以用每一條候選詞序列的錯誤率(或排序位置)來當作此樣本權重；錯誤率越高或排序越後面者，其重要程度就越重、影響力就越大。

輪轉雙重鑑別式模型(R2D2)與權重式全域條件式對數線性模型(WGCLM)類似，其目標函數期望每候選詞序列彼此之間的對數差異越小越好；輪轉雙重鑑別式模型亦考慮了樣本權重$\exp(\sigma_k \omega_{W_{i,j}})$，不同的候選詞序列會因其本身的錯誤率(或排序位置)對於模型的訓練有不同程度的影響。因爲輪轉雙重鑑別式模型考慮了每一個候選詞序列與其它候選詞序列之間的關係，所以訓練的過程亦較其它鑑別式語言模型耗時。

三、基於邊際資訊之鑑別式語言模型(MDLM)

近年來有許多學者針對鑑別式語言模型提出了不同的觀點與做法。例如，爲了讓鑑別式語言模型的訓練更有效率，在感知器演算法中融入了訓練語句不同錯誤率程度的資訊中[17]；另外，也有學者額外地將候選詞序列的文法結構與各種詞性的出現頻率等語言特徵加入鑑別式語言模型使用，讓鑑別式語言模型在對候選詞序列進行重新排序時，可以

參考詞序列所含豐富的語言相關資訊[18]。在本論文的研究裡，我們提出了考慮邊際(Margin)資訊的概念[19, 20]於鑑別式語言模型之訓練資料選取；對於每個訓練語句嘗試以其每一個候選詞序列各自的辨識錯誤率爲基礎，動態地來決定訓練資料(亦即候選詞序列)是否選取以用於模型訓練。在此，我們將先回顧邊際估測法則，接著說明本論文所提出的基於邊際資訊之鑑別式語言模型。

(一)、邊際估測法則

基於邊際資訊的資料選取方法目的是選取對於鑑別式模型訓練較具重要性的訓練資料，期望在模型訓練的過程中不僅可以降低訓練的時間，亦希望能夠得到較好的模型參數，提升辨識的正確性。例如，最大邊際估測法則(Large-Margin Estimation, LME)[19]、柔性邊際估測法(Soft-Margin Estimation, SME)[21]皆是基於邊際資訊的估測法則中典型的代表。

當最大化邊際估測法使用於鑑別式語言模型時，其基本精神是希望拉大參考(或是錯誤率最低)候選詞序列(Reference Word Sequence) W_i^R 與其它可能候選詞序列 $W_{i,j}$ 之排序分數的差異，讓參考候選詞序列 W_i^R 的分數較其它可能候選詞序列 $W_{i,j}$ 愈大愈好；通常我們將此分數的差異稱爲“分離邊際(Separation Margin)”：

$$\tau(x_i) = Score(W_i^R) - \max_{W_{i,j} \neq W_i^R} Score(W_{i,j}) \tag{16}$$

其中 $Score(W_i^R)$ 爲參考詞序列的重新排序分數； $Score(W_{i,j})$ 爲某一個候選詞序列的重新排序分數。由式(16)可知，若 $\tau(x_i) > 0$ ，表示使用目前的鑑別式語言模型對於語句 x_i 所對應的 M 條候選詞序列進行排序時，可以賦予詞序列 W_i^R 最高的排序分數，我們可以視爲沒有辨識錯誤發生(爲理想狀況)；反之，若 $\tau(x_i) < 0$ ，則表示正確(或是錯誤率最低)的候選詞序列之排序分數不是所有候選詞序列中最高的，因此經重新排序的語音辨識輸出將不是最佳的結果。在最大邊際估測法的訓練過程中，我們首先爲訓練語料 $\{x_1, x_2, \ldots, x_L\}$ 定義一組支援集(Support Set)：

$$S_{\text{LME}} = \{x_i \mid 0 \leq \tau(x_i) \leq \varepsilon\} \tag{17}$$

ε 是一個正實數，可以用來控制支援集中所包含的訓練語料個數，最大化邊際估測法的目標函數就可定義爲最大邊際估測法的訓練目標是希望最大化支援集中的最小分離邊際[19]：

$$F_{\text{LME}}(\lambda) = \min_{x_i \in S_{\text{LME}}} \tau(x_i) \tag{18}$$

由式(18)可知，訓練時所選取的訓練語料是原本使用重新排序可以正確地選出參考候選詞序列之訓練語句，而其它的訓練語句則被排除在訓練之外。在理論上，經過最大邊際估測法訓練後，對於訓練語料的分離邊際應變大，代表語言模型更具有一般化的能力。另一方面，由於大詞彙連續語音辨識系統是複雜而且其所提供辨識率尚未達完美，因此實際在鑑別式語言模型的訓練語料中，被定義於支援集裡的訓練語句個數將會是非常有限的，這會使得鑑別式語言模型調整後對整體的辨識率提升非常有限(這個問題在當最大邊際估測法被使用於聲學模型估測時，亦曾被討論過)。

爲了解決最大邊際估測法僅考慮支援集的資訊於鑑別式模型訓練的缺失，柔性最大邊際估測法則(Soft-Large Margin Estimation, S-LME)[22]被提出來改善此一問題。柔性最大邊際估測法則不在僅將重新排序可以正確地選出參考候選詞序列之訓練語句納入考量，它對訓練語句另外定義了一組錯誤集(Error Set)：

$$\varphi = \{x_{i'} \mid \tau(x_{i'}) < 0\} \tag{19}$$

結合支援集與錯誤集，柔性最大邊際估測法爲最大化下列目標函數[22, 23]：

$$F_{\text{S-LME}}(\lambda) = \min_{x_i \in S_{\text{LME}}} \tau(x_i) - \sigma \cdot \frac{1}{|\varphi|} \sum_{x_{i'} \in \varphi} \delta(x_{i'}) \tag{20}$$

也就是除了支援集所提供的鑑別性資訊外，還加入了平均錯誤估測於模型的目標函數中。在式(20)中，σ是一個正實數，用來控制平均錯誤估測對於訓練鑑別式模型時的影響性；$\delta(\cdot)$是錯誤函數，通常被定義爲[22]：

$$\delta(x_i) = \max_{W_{i,j} \neq W_i^R} \left(Score(W_{i,j}, \lambda)\right) - Score(W_i^R, \lambda) \tag{21}$$

即分離邊際的負數。

最大邊際估測法則只考慮了“與分離邊際較近”(參照式(17))且重新排序可以正確地選出參考候選詞序列之訓練語句，如此不僅忽略了分離邊際附近的其它資訊，亦會導致訓練語句數量不足，最終使得訓練出來的鑑別式模型一般化能力不足；有別於最大邊際估測法，柔性邊際估測法(Soft Margin Estimation, SME)則是藉由考慮條件的放寬，將那些辨識錯誤(亦即參考候選詞序列之重新排序分數不是最高)在一定範圍內的訓練語句也一併列入考量，來彌補訓練語句上的不足。

柔性邊際估測法則的訓練目的如同最大邊際估測法則一樣，希望最大化訓練語料中的最小分離邊際，差別是柔性邊際估測法則在定義支援集時的條件比較彈性，加入了一個鬆弛變量(Slack Variable)ξ：

$$S_{\text{SME}} = \{x_i \mid -\xi \leq \tau(x_i) \leq \varepsilon\} \tag{22}$$

其中ξ爲一個大於零的實數，其表示那些辨識錯誤的訓練語句若其分離邊際大於$-\xi$也會在語言模型訓練時列入考量。柔性邊際估測法爲最大化下列目標函數：

$$F_{\text{SME}}(\lambda) = \min_{x_i \in S_{\text{SME}}} \tau(x_i) \tag{23}$$

(二)、基於邊際資訊之鑑別式語言模型(MDLM)

由上一節的簡介可知，過去考慮邊際概念於鑑別式語言模型時，通常只考慮參考詞序列與最佳候選詞序列之間的關係(參照式(16))。本論文嘗試將分離邊際的概念定義爲參考詞序列與每一個候選詞序列之間的關係；並且更進一步地，在定義支援集時，同時考慮每一訓練語句其參考候選詞序列與最高錯誤率候選詞序列的錯誤率差值。在多考慮每一個候選詞序列與參考詞序列的關係後，不僅可以解決訓練語料不足的問題，更可以改進鑑別式語言模型的一般化能力。我們所提出的鑑別式模型主要的目的是希望將參考(錯誤率最低)候選詞序列與其它候選詞序列彼此間的分離邊際越大越好。如此一來，可以

1 For $t = 1...T$ where T is the total number of iterations

2 For each training sample $(x_i, W_i^R), i = 1...L$

3 $v_j = 0,\ j = 1...N$

4 For $1 \le j \le k \le n$ where n is the N-best

5 $if\left(Score\left(W_{i,j}^R\right) - Score\left(W_{i,k}\right) < \tau\right)$

6 $v_j = v_j + 1$

7 $v_k = v_k - 1$

8 $\hat{\lambda}_{\mathrm{d}} = \lambda_d + \eta \cdot \sum_{j=1}^{N} v_j f_d\left(W_{i,j}\right)$

圖二、基於邊際之鑑別式語言模型演算法

降低重新排序候選詞序列時的混淆程度，進而提升鑑別式語言模型的效果。本論文所提出的基於邊際之鑑別式語言模型(Margin-based Discriminative Language Model, MDLM)的演算法如圖二所示。

首先，我們將分離邊際定義爲：

$$\tau_{\mathrm{MDLM}}\left(W_{i,j}\right) = Score\left(W_i^R, \lambda\right) - Score\left(W_{i,j}, \lambda\right) \tag{24}$$

若$\tau\left(W_{i,j}\right) > 0$，表示使用目前的鑑別式語言模型可以賦予詞序列$W_i^R$有較$W_{i,j}$高的排序分數，反之，若$\tau\left(W_{i,j}\right) < 0$，則表示參考(辨識錯誤率最低)候選詞序列之排序分數較詞序列$W_{i,j}$低，因此辨識器的輸出將不會是最佳結果。接著，我們定義一組支援集(Support Set)：

$$S_{\mathrm{MDLM}} = \left\{W_{i,j} \mid \tau_{\mathrm{MDLM}}\left(W_{i,j}\right) \le \gamma_i\right\} \tag{25}$$

其中，γ_i是每一個訓練語句的判別量；在本論文中，我們將它定義爲：

$$\gamma_i = \exp\left(\alpha\left(\max_j \omega_{W_{i,j}} - \omega_{W_i^R}\right)\right) \tag{26}$$

α是一個實驗常數；$\omega_{W_i^R}$爲W_i^R的辨識錯誤率；$\omega_{W_{i,j}}$爲候選詞序列$W_{i,j}$的辨識錯誤率；所以γ_i是隨著訓練語句的不同而有所變動，當參考候選詞序列W_i^R的錯誤率遠小於錯誤率最高的候選詞序列$W_{i,j}$時，γ_i的值應愈大。至此，不同於過去考慮邊際概念的鑑別式語言模型，本論文所提出的基於邊際之鑑別式語言模型考慮每一訓練語句其參考候選詞序列與其它所有候選詞序列的關係進行資料選取。並且在選取的過程中，考慮了訓練語句各自的辨識錯誤率。最後，結合式(24)、(25)與(26)，我們將基於邊際之鑑別式語言模型的目標函數定義爲：

$$F_{\mathrm{MDLM}}(\lambda) = \frac{1}{2}\sum_{i=1}^{L} \sum_{\substack{W_{i,j} \in GEN(x_i) \\ \& W_{i,j} \in S_{\mathrm{MDLM}}}} \left(\tau_{\mathrm{MDLM}}\left(W_{i,j}\right)\right)^2 \tag{27}$$

利用梯度下降法將此目標函數對每一維權重參數λ_d做偏微分可求得其調整量，每一維特徵權重向量的更新式爲：

	有無考慮樣本權重 $\omega_{W_{i,j}}$	有無考慮 W_i^R	一般化能力	訓練速度
Perceptron	無	有	差	快
MERT	有	無	佳	慢
GCLM	無	有	略佳	慢
WGCLM	有	有	略佳	慢
R2D2	有	有	略佳	很慢
MDLM	無	有	略佳	慢

表一、鑑別式語言模型之間的比較

$$\hat{\lambda}_d = \lambda_d - \eta \cdot \sum_{\substack{W_{i,j} \in GEN(x_i) \\ \& W_{i,j} \in S_{\text{MDLM}}}} \left[\left(\tau_{\text{MDLM}}(W_{i,j}) - \gamma_i \right) \cdot \left(f_d(W_i^R) - f_d(W_{i,j}) \right) \right] \tag{28}$$

事實上，基於邊際資訊之鑑別式語言模型(MDLM)目標函數與感知器演算法有些相似，皆是考慮最小平方誤差。然而，感知器演算法是期望排序分數最高的候選詞序列與參考詞序列之分數差異越小越好；而基於邊際之鑑別式語言模型不僅考慮排序分數最高的候選詞序列，更以分離邊際爲基礎，考慮了更多參考詞序列與其它候選詞序列之間的關係，因此不會像感知器演算法會有過度訓練的問題。再者，由於我們將邊際的資料選取概念稍作改良，使得參與訓練的資料變多，因此不像先前簡介的其它各式運用邊際資訊的鑑別式語言模型，容易遭遇訓練資料太少的問題。

值得一提的是，基於邊際之鑑別式語言模型也可以如同感知器演算法一般，將式子(28)中的 $\tau_{\text{MDLM}}(W_{i,j}) - \gamma_i$ 省略，將更新式子簡化成：

$$\hat{\lambda}_d = \lambda_d + \eta \cdot \sum_{\substack{W_{i,j} \in GEN(x_i) \\ \& W_{i,j} \in S_{\text{MDLM}}}} \left(f_d(W_i^R) - f_d(W_{i,j}) \right) \tag{29}$$

透過此更新式以及迭代的訓練，期望能求得最佳的特徵權重向量。

利用表一來說明本論文引入的基於邊際資訊之鑑別式語言模型(MDLM)與其它鑑別式語言模型之間的比較關係。其中，考慮樣本權重的好處是能將訓練語句根據其樣本權重的不同，對於模型的訓練影響程度有不同的影響，而非每一個訓練語句都佔有相同的權重。在第四節實驗結果觀察，全域條件式對數線性模型(GCLM)與權重式全域條件式對數線性模型(WGCLM)來比較，權重式全域條件式對數線性模型多加入了樣本權重，其結果優於全域條件式對數線性模型，可推測樣本權重的考量對於模型訓練上的確是有正向幫助。

語料	句數	長度(小時)
訓練集語料	30,600	約 23
發展集語料	1,998	約 1.5
測試集語料	1,997	約 1.5

表二、實驗語料統計資訊

接著，最小化錯誤率訓練(MERT)的目標函數中沒有根據最低錯誤率詞序列W_i^R爲目標參考去做訓練，使得在訓練的過程當中訓練語句不會過度訓練去適合這些最低錯誤率詞序列，因此最小化錯誤訓練會有較佳的一般化能力；反觀感知器演算法(Perceptron)，會因爲過度訓練(Overfitting)，使得模型的一般化能力較差。

在訓練的速度上，則因爲各式方法著重的訓練目標不同，而有不同的時間複雜度。感知器演算法在訓練的過程當中，只考慮正確(或是錯誤率最低)的候選詞序列與排序分數最高的詞序列之間的關係；最小化錯誤率訓練、全域條件式對數線性模型、權重式全域條件式對數線性模型在訓練的過程中，考慮了正確(或是錯誤率最低)的候選詞序列與其它候選詞序列之間的關係;輪轉雙重鑑別式模型(R2D2)是考慮了所有候選詞序列之間彼此的關係。因此在訓練的過程中，輪轉雙重鑑別式模型所需的時間複雜度最高，相較下，感知器演算法訓練時所花費時間最少。

四、實驗結果與討論

(一)、實驗語料

本論文實驗語料取自公視新聞(Mandarian Across Taiwan-Broadcast News, MATBN)[24]。公視新聞語料是 2001 年至 2003 年間由中研院資訊所口語小組(SLG)與公共電視台(PTS)合作錄製，包含了內場新聞與外場新聞兩個部分。其中內場新聞爲主播語料，外場新聞語料包含有採訪記者(Field Reporters)語音語料與受訪者(Interviewees)語音語料。

由於內場主播語料大部分來自於同一主播所錄製，爲了避免語者相依(Speaker Dependent)現象造成實驗偏差，故不採用內場主播語料；外場受訪者語料，則是包含許多語助詞與背景音樂，所以也沒有採用；因此，本論文的實驗語料選取自外場採訪記者語料。訓練集語料、測試集語料及發展集語料皆選取自公視新聞 2001 年至 2002 年外場採訪記者，分別爲 30,600 句(約 23 小時)、1,997 句(約 1.5 小時)及 1,998 句(約 1.5 小時)。如表二所示。

背景語言模型爲三連語言模型(Trigram Language Model)，採用 Katz Back-off Smoothing 平滑化方法來解決資料稀疏的問題。其訓練語料來自 2001 年至 2002 年中央通訊社(Central News Agency, CNA)的文字新聞語料，包含了約一億五千萬個中文字，經過斷詞後約有八千萬詞。此語言模型是使用 SRI Language Modeling Toolkit(SRILM)[25]訓練所得。

我們以基礎辨識器[26]配合背景三連語言模型於完整詞圖搜尋(Word Graph Rescoring)的最佳結果做爲基礎辨識率(Baseline)，它在訓練集語料、發展集語料以及測試集語料的

各式鑑別式語言模型	訓練集(%)	發展集(%)	測試集(%)
Perceptron	8.20	14.14	14.99
MERT	10.48	14.27	15.33
GCLM	10.61	14.62	15.88
WGCLM	10.38	14.39	15.39
R2D2	8.76	13.39	14.23

表三、各式鑑別式語言模型的基礎實驗結果

辨識字錯誤率(Character Error Rate)分別爲 11.26%、15.27%與 16.39%。並且，我們挑選基礎辨識器產生的前 100 條最佳($M = 100$)的辨識結果，做爲鑑別式語言模型的訓練與測試語料。

(二)、各式鑑別式語言模型的實驗結果

首先，我們比較各種不同的鑑別式語言模型應用於語音辨識結果之重新排序，各種方法的辨識字錯誤率如表三所示。我們可以由表三觀察到，如同先前提到的，感知器演算法(Perceptron)在訓練語料上的表現的確優於其它各種鑑別式語言模型，但在測試語料的表現則無法有相同的效果。而在測試語料的實驗結果，以輪轉雙重鑑別式語言模型(R2D2)的效果最爲顯著，這也說明了在訓練的過程中，考量較多有關候選詞序列彼此之間關係的資訊對模型的訓練是有正面幫助的，會使得模型有較好的一般化能力。

(三)、基於邊際資訊之鑑別式語言模型相關實驗結果

本論文所提出之基於邊際資訊之鑑別式語言模型(MDLM)於語音辨識之實驗結果如表四所示。實驗中，我們比較了四種不同支援集的定義方式：

- 動態型(MDLM-D)： $S_{\text{MDLM-D}} = \{W_{i,j} \mid \tau_{\text{MDLM}}(W_{i,j}) \le \gamma_i\}$
- 正確分類動態型(MDLM-CD)： $S_{\text{MDLM-CD}} = \{W_{i,j} \mid 0 \le \tau_{\text{MDLM}}(W_{i,j}) \le \gamma_i\}$
- 固定型(MDLM-F)： $S_{\text{MDLM-F}} = \{W_{i,j} \mid \tau_{\text{MDLM}}(W_{i,j}) \le \rho\}$，其中 ρ 是一個正實數
- 正確分類固定型(MDLM-CF)： $S_{\text{MDLM-CF}} = \{W_{i,j} \mid 0 \le \tau_{\text{MDLM}}(W_{i,j}) \le \rho\}$，其中 ρ 是一個正實數

其中，正實數 ρ 設定爲 5。首先，動態型的鑑別式語言模型在訓練集的辨識字錯誤率爲 6.09%，測試集的辨識字錯誤率爲 14.10%，而固定型的鑑別式語言模型在訓練集的辨識字錯誤率爲 5.18%，測試集的辨識字錯誤率爲 13.91%。因此，不論在訓練集或測試集，使用固定型的方式定義支援集似乎較考慮錯誤率資訊於資料選取的方式要好。值得一提的是，我們認爲考慮錯誤率於資料選取應該有助於挑選具鑑別性的訓練語料，但如何設

	訓練集(%)	發展集(%)	測試集(%)
MDLM-D	6.09	13.37	14.10
MDLM-CD	6.69	13.38	14.20
MDLM-F	5.18	13.25	13.91
MDLM-CF	5.49	13.17	13.98

表四、基於邊際之鑑別式語言模型實驗結果

	訓練集(%)	發展集(%)	測試集(%)
MDLM-D	5.97	13.34	13.96
MDLM-CD	7.01	13.38	14.00
MDLM-F	5.56	13.35	13.78
MDLM-CF	5.86	13.30	13.87

表五、考慮多條正確(錯誤率最低)的詞序列與樣本權重於基於邊際之鑑別式語言模型之實驗結果

計判別量 γ_i，是一個值得研究的題目，也是本論文在未來將繼續研究的問題。

接著，如果我們更進一步地限制僅考慮分離邊際大於 0 的候選詞序列於模型訓練中(即正確分類動態型(MDLM-CD)與正確分類固定型(MDLM-CF))，實驗結果如表四所示。正確分類動態型的鑑別式語言模型在訓練集的辨識字錯誤率爲 6.69%；在測試集的辨識字錯誤率爲 14.20%。而正確分類固定型的鑑別式語言模型在訓練集的辨識字錯誤率爲 5.49%，測試集的辨識字錯誤率爲 13.98%。同樣地，使用固定型的方式定義支援集較考慮錯誤率資訊於資料選取的方式要好。另外，值得注意的是，僅考慮分離邊際大於 0 的候選詞序列於模型訓練，並不會得到較好的結果。分析其原因，可能由於分離邊際小於 0 的候選詞序列表示其排序分數大於參考(錯誤率最低)候選詞序列，如果我們將這些候選詞序列皆捨去不考慮，則鑑別式語言模型在訓練的過程中，可能會無法適當地調整模型參數，以使得參考候選詞序列之分數高於其它候選詞序列，如此，將會使得語言模型的鑑別性較差。

相較於其它各式語言模型(參照表三)，我們所提出的基於邊際資訊之鑑別式語言模型(包含四種不同的支援集定義方式)不論是在訓練集、發展集以及測試集皆有最低的辨識錯誤率。由此可知，使用邊際資訊於資料選取的鑑別式語言模型，的確保留了富有鑑別性資訊的訓練語句於模型訓練的過程中，並且也不會容易地遭受一般化能力不足的問題。

接著，我們更進一步的探討在訓練過程中，參考詞序列的個數對實驗的影響。在前人的研究中，使用辨識器所產生的候選詞序列中錯誤率最低的詞序列作爲訓練時的正確答案，會較使用正確的人工轉寫(Manual Transcription)詞序列有更佳的實驗結果[27]。然

而，在辨識器所產生的候選詞序列中，往往會存在數條辨識率相同且最低的候選詞序列。通常，我們會隨機選取其中分數最高的一條，當成訓練時的正確答案。在此，我們嘗試將這些擁有最低辨識錯誤率的候選詞序列皆當作參考候選詞序列，以用於模型的訓練過程中。另外，在更新特徵權重參數時，我們將詞序列的排列順序當作一種樣本權重，使得每一條詞序列對於鑑別式模型的影響程度有所不同，我們認爲，排序越前面的詞序列，應該對模型有較重要的影響性，因此我們將樣本權重定義爲詞序列排列順序的倒數之差。實驗結果如表五所示，我們可以發現，加入考慮多條正確(錯誤率最低)的候選詞序列與樣本權重後，測試集皆可以獲得更好的辨識結果，其中以固定型的支援集(MDLM-F)定義方式可以獲得 13.78%的辨識字錯誤率，是最佳的實驗結果。值得探討的是，訓練集與發展集皆沒有因爲考慮多條正確(錯誤率最低)的候選詞序列與樣本權重而獲得較佳的辨識結果。我們認爲，可能的原因是因爲模型的一般化能力增加，故雖然在訓練集上沒有正向的幫助，但當使用於測試集時，相較於先前的實驗(參照表四)，的確可以獲得較好的辨識結果。

五、結論及未來展望

語言模型不論是在機器翻譯、資訊檢索、語音辨識等領域中，都扮演一個不可或缺的重要角色。在語音辨識中，語言模型能輔助解決聲學模型的混淆，估計一段語句在自然語言中發生的可能性。*N* 連語言模型是最常被使用的，但它僅能捕捉到短距離的詞彙規則資訊，近十幾年來，鑑別式語言模型陸續被提出，並且廣泛的使用在於各個領域之中；在語音辨識的領域中，它提供了一個新的視野，以直接降低語音辨識錯誤率爲訓練目標。本論文針對常見的鑑別式語言模型做了一系列的討論與實驗比較，在我們的實驗中，各式鑑別式語言模型確實可以更進一步地輔助 *N* 連語言模型，降低辨識的錯誤率。

未來，我們將繼續研究基於不同訓練準則之鑑別式語言模型，並著重於探討各式語言特徵加入於鑑別式語言模型使用[28, 29, 30]。另外，除了加入各種語言特徵外，我們也有興趣於特徵選取對於鑑別式語言模型的影響。目前大部分鑑別式語言模型所使用的語言特徵數量皆非常龐大；面對這麼一群龐大的語言特徵，我們期望發展出一套特徵選取的方式，期望可以降低鑑別式語言模型訓練過程的時間需求，更希望進一步地改善鑑別式語言模型而獲得更好的辨識結果。

參考文獻

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

[2] M. Collins, and T. Koo, “Discriminative reranking for natural language parsing,” *Computational Linguistics*,” Vol. 31, No. 1, pp. 25-70, 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, M. Collins and M. Johnson,” Discriminative n-gram language modeling,” *Computer Speech and Language*, Vol. 21, No. 2, pp. 373-392, 2007.

[5] J.-W. Liu, S.-H. Lin and B. Chen, “Exploiting discriminative language models for reranking speech recognition hypotheses,” in *Proceedings of ROCLING XXII:*

Conference on Computational Linguistics and Speech Processing (ROCLING 2010), pp. 30-49, 2010.

[6] B. Chen and C.-W. Liu, "Discriminative language modeling for speech recognition with relevance information," in *Proceedings of the International Conference on Multimedia and Expo*, 2011.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1574-1577, 2008.

[15] 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.

[16] T. Oba, T. Hori and A. Nakamura, "Round-robin discriminative model for reranking ASR hypotheses," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 2446-2449, 2010.

[17] T. Oba, T. Hori and A. Nakamura, "An approach to efficient generation of high-accuracy and compact error-corrective models for speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1753-1756, 2007.

[18] E. Arisoy, M. Saraclar, B. Roark and I. Shafran, "Syntactic and sub-lexical features for Turkish discriminative language models," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 5538-5541, 2010.

[19] H. Jiang, X. Li and C. Liu, "Large margin hidden markov models for speech recognition," *IEEE transactions on audio, speech, and language processing*, Vol. 14, No. 5, pp. 1584-1595, 2006.

[20] Y.-T. Lo and B. Chen, "A comparative study on margin-based discriminative training of acoustic models," in *Proceedings of ROCLING XXII: Conference on Computational Linguistics and Speech Processing (ROCLING 2010)*, pp. 65-85, 2010.

[21] J. Li, M. Yuan and C.-H. Lee, "Soft margin estimation of hidden markov model parameters," in *Proceedings of International Conference on Spoken Language Processing*, pp. 2422-2425, 2006.

[22] H. Jiang and X. Li and C. Liu, "Incorporating training errors for large margin HMMs under semidefinite programming framework," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 629-632, 2007.

[23] V. Magdin and H. Jiang, "Large margin estimation of n-gram language models for speech recognition via linear programming," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 5398-5401, 2010.

[24] 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.

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

[26] B. Chen, J.-W. Kuo and W.-H. Tsai, "Lightly supervised and data-driven approaches to mandarin broadcast news transcription," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 777-780, 2004.

[27] B. Roark, M. Saraclar and M. Collins, "Corrective language modeling for large vocabulary ASR with the perceptron algorithm," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 749-752, 2004.

[28] L. Shen, A. Sarkar and F. J. Och, "Discriminative reranking for machine translation," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 177-184, 2004.

[29] E. Arisoy, B. Ramabhadran and H.-K. J. Kuo, "Feature combination approaches for discriminative language models," in *Proceedings of Annual Conference of the International Speech Communication Association*, 2011.

[30] L. Wang, J. Lin and D. Metzler, "A cascade ranking model for efficient ranked retrieval," in *Proceedings of Annual International ACM SIGIR Conference*, pp. 105-114, 2011.