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。