國立臺灣師範大學

資訊工程研究所碩士論文

指導教授：陳柏琳　博士

遞迴式類神經網路語言模型使用額外資訊

於語音辨識之研究

Recurrent Neural Network-based Language Modeling with Extra Information Cues for Speech Recognition

研究生：黃邦烜　撰

中華民國　一〇一　年　八　月

摘要

語言模型藉由大量的文字訓練後，可以捕捉自然語言的規律性，並根據歷史詞序列來區辨出下一個詞應該為何，因此在自動語音辨識(Automatic Speech Recognition, ASR)系統中扮演著不可或缺的角色。傳統統計式 *N* 連(*N*-gram)語言模型是常見的語言模型，它基於已知的前 *N*-1 個詞來預測下一個詞出現的可能性。當 *N* 小時，缺乏了長距離的資訊；而 *N* 大時，會因訓練語料不足產生資料稀疏之問題。近年來，由於類神經網路(Neural Networks)的興起，許多相關研究應運而生，類神經網路語言模型即是一例。令人感興趣的是，類神經網路語言模型能夠解決資料稀疏的問題，它透過將詞序列映射至連續空間來估測下一個詞出現的機率，因此在訓練語料中不會遇到未曾出現過的詞序列組合。除了傳統前饋式類神經網路語言模型外，近來也有學者使用遞迴式類神經網路來建構語言模型，其希望使用遞迴的方式將歷史資訊儲存起來，進而獲得長距離的資訊。

本論文研究遞迴式類神經網路語言模型於中文大詞彙連續語音辨識之使用，探索額外使用關聯資訊以更有效地捕捉長距離資訊，並根據語句的特性動態地調整語言模型。實驗結果顯示，使用關聯資訊於遞迴式類神經網路語言模型能對於大詞彙連續語音辨識的效能有相當程度的提昇。

關鍵詞：語音辨識、語言模型、前饋式類神經網路、遞迴式類神經網路

Abstract

The goal of language modeling (LM) attempts to capture the regularities of natural languages. It uses large amounts of training text for model training so as to help predict the most likely upcoming word given a word history. Therefore, it plays an indispensable role in automatic speech recognition (ASR). The *N*-gram language model, which determines the probability of an upcoming word given its preceding *N*-1 word history, is most prominently used. When *N* is small, a typical *N*-gram language model lacks the ability of rendering long-span lexical information. On the other hand, when *N* becomes larger, it will suffer from the data sparseness problem because of insufficient training data. With this acknowledged, research on the neural network-based language model (NNLM), or more specifically, the feed-forward NNLM, has attracted considerable attention of researchers and practitioners in recent years. This is attributed to the fact that the feed-forward NNLM can mitigate the data sparseness problem when estimating the probability of an upcoming word given its corresponding word history through mapping them into a continuous space. In addition to the feed-forward NNLM, a recent trend is to use the recurrent neural network-based language model (RNNLM) to construct the language model for ASR, which can make efficient use of the long-span lexical information inherent in the word history in a recursive fashion.

In this thesis, we not only investigate to leverage extra information relevant to the word history for RNNLM, but also devise a dynamic model estimation method to obtain an utterance-specific RNNLM. We experimentally observe that our proposed methods can show promise and perform well when compared to the existing LM methods on a large vocabulary continuous speech recognition (LVCSR) task.

Keywords: automatic speech recognition, language modeling, feed-forward neural network, recurrent neural network

誌謝

首先，在此感謝我的父母及家人，因為有您們的鼓勵，我才能夠不斷地正視我所遇到的挑戰；因為有您們的支持，我才能自在地追隨自己的理想，並且順利地完成學業。於求學期間的點點滴滴，我都銘記在心，謝謝您們。

感謝指導教授　陳柏琳博士在我研究所時期的教導，無論是在研究方面，抑或是待人處事方面都從老師身上獲益良多。老師總是不厭其煩的給予許多建議和鼓勵，並且提供優質的研究環境，讓我們能無後顧之憂的進行研究。求學過程就如同旅行一樣，每天都能學習到及見識到許多新的事物，兩年的旅程中老師您不僅是研究之師，也是人生之師，誠摯感謝您的諄諄教誨。

感謝口試委員　洪志偉博士及　張道行博士，因為有您們的指導與建議，讓我的論文更臻完整，並且從老師們的研究姿態中，學習到研究的真諦。

感謝實驗室的學長姐，士翔學長、永典學長、家妏學姐、鈺玫學姐、珮寧學姐和紋儀學姐，謝謝你們在學業或生活上給予諸多的建議和幫助。感謝冠宇學長每周都不辭辛勞地與我們討論研究，學長就如同我的第二個指導老師一樣，在我不知所措或是研究沒有進展時，幫助我且啟發我，祝福你在未來的研究生涯能乘風破浪、一帆風順。感謝敏軒學長不只在研究上耐心地給予指導，看著你認真的姿態也時常激勵著我，而在生活或是求學中更是熱心的幫助我。謝謝皓欽、予真和憶文，在研究所的日子裡相互鼓勵與成長，使我的研究生活注入許多動力。也謝謝實驗室的學弟妹，孝宗、逸婷、俊諭、麟傑、柏翰和黃威，因為有你們的加入，使得實驗室增添許多歡笑與熱鬧的氣氛，而在我忙碌之時，也傾注全力協助我，萬分感謝你們。另外也感謝創價學會的學會們，不時地鼓勵我與給予建議，使我徬徨無力或態度消極時，轉換自己的一念，擁有正向、積極的態度去面對。

最後，謹以一句話勉勵自己及分享給各位，人生裡不一定要贏，但是絕對要以不輸的心來完成所有挑戰。

「不論如何，都要洋溢希望，要開朗！煩惱越大，越要咬緊牙關，面帶笑容前進。」

—　池田大作

邦烜　謹誌

目錄

圖目錄

表目錄

第1章　緒論

1.1　研究背景

隨著科技的改變,人們的生活型態也大幅改變,符合人性的智慧型科技因此催生，自動語音辨識(Automatic Speech Recognition, ASR)便是其中代表之一。自動語音辨識的應用回溯到最早在 1920 年發展的玩具狗-Radio Rex，直到現今 iPhone 4S 裡的語音助理 Siri;這些應用的發展就是希望使電腦能夠理解使用者的語音輸入，進而增進生活的便利性和工作的效率。

語音是人與人溝通的基本媒介，如果無法透過語音來對話，取而代之的可能是文字、圖像抑或是肢體語言，但皆無法像語音一樣正確地表達彼此的想法。在對話中，人們藉由語調和詞句，了解對方的情緒與想法等諸多細節，而聽懂對方的話語以及理解對方的想法；這些人所擁有的天賦，是目前資訊科技所無法達到而需研究的。因此自動語音辨識的研究也變得更加重要；其它相關研究有語音合成(Speech Synthesis)、文字轉語音(Test-to-Speech, TTS)、語者辨識(Speaker Recognition)和情緒辨識等。另外，在教育學習方面則有口語對話系統、電腦輔助語言學習(Computer Assisted Language Learning, CALL)和電腦輔助發音訓練(Computer Assisted Pronunciation Training, CAPT)等研究，希望藉由自動辨識給予正向或負向的回饋來輔助學習者的學習。而語音辨識相關研究也常與其他領域結合，如資訊檢索(Information Retrieval)、機器翻譯(Machine Translation, MT)或自然語言處理(Nature Language Processing, NLP)等。由此可知，語音辨識技術的發展與提昇，將帶領人類進入新的生活型態。

本章首先簡介語音辨識的研究內容，接著將說明本論文的研究內容與方向。

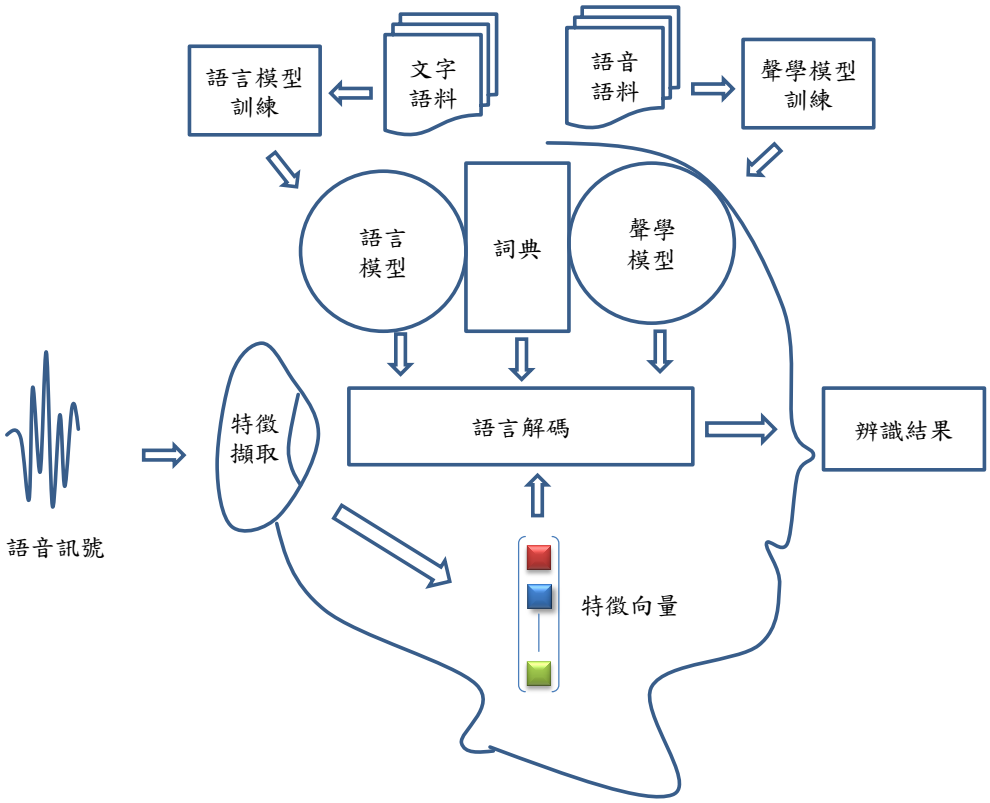


圖 1-1：自動語音辨識流程圖

1.2 語音辨識簡介

科技始終來自於人性，要解決語音辨識的問題，我們可以去探討人類如何接受到聲音或大腦如何進行理解。人類在聽到聲音後，會根據聲音的特性去分析，例如是尖銳的音色或是低沉的音色，再藉由大腦所獲得的記憶去辨別聲音所帶來的資訊，接著作出適當的反應；而當今的語音辨識流程也類似於上述人類的聽覺感受過程。

自動語音辨識系統主要可以分成四個主要的部份，分別是特徵擷取(Feature Extraction)、聲學模型(Acoustic Model)、語言模型(Language Model)與語言解碼(Linguistic Decoding)，透過這些部份才會得到最後辨識結果。當電腦接受到一段

語音訊號，首先透過特徵擷取來處理語音訊號，得到可以代表此段語音訊號的特徵參數；接著，將所擷取的特徵參數轉換成語音特徵向量，以利語音辨識系統使用或分析。另一部分，則使用語音語料和文字語料分別建構出聲學模型和語言模型，用以表示語音與文字之間的對應關係以及代表語言中各種詞彙的出現情形。再根據聲學模型、語言模型、詞典和特徵向量所提供的資訊以進行語言解碼，獲得最後辨識結果。更明確地來說，我們可以將語音辨識的過程透過數學符號來表示，如式(1-1)。

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

輸入一段語音訊號 X 後，自動語音辨識最主要的目的就是找出一段最有可能的對應詞序列 W^*。由於 $P(W|X)$較難以直接計算，因此可利用貝式定理(Bayes' Theorem)做轉換，得到 $p(X|W)$、$P(W)$和 $P(X)$。其中 $p(X|W)$為聲學模型，代表給定一詞序列 W，產生某語音訊號 x 的機率；$P(W)$為語言模型，代表產生某一詞序列 W 的機率；而 $P(X)$為語音訊號 X 的事前機率，對同一語音訊號 X 來說 $P(X)$皆相等，並不會影響排序，所以可忽略不計。因此，可以簡化為 $\arg\max_w p(X \mid W)P(W)$。

接著，本論文將介紹自動語音辨識系統主要的四個部分。

(一)特徵擷取(Feature Extraction)

特徵擷取顧名思義就是從語音訊號中，找出能夠代表此段語音訊號的特徵，而獲取的特徵為了利於電腦做分析及運用。因此表示成特徵向量，常見的方法有線性預測係數(Linear Prediction Coefficients, LPC) [Makhoul, 1975]、感知線性預測係數

(Perceptual Linear Prediction Coefficients, PLPC)[Hermansky, 1990]、異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)[Kumar, 1997]、最大相

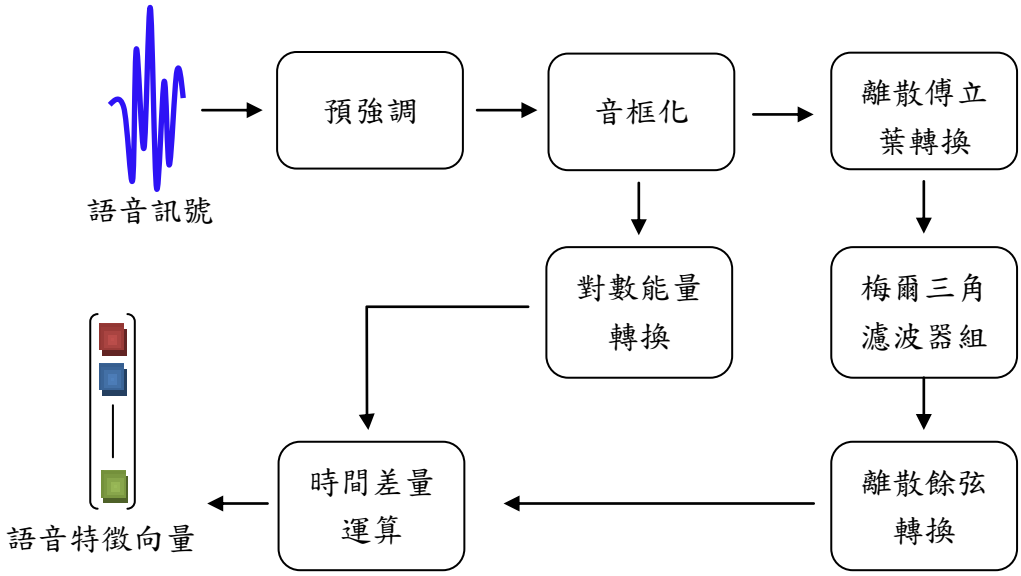


圖 1-2 ：MFCC 特徵擷取流程圖

似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)[Gales, 1998]以及梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficients, MFCC)[Davis and Mermelstein, 1980]等不同的語音特徵參數。其中梅爾倒頻譜係數(MFCC)是由考慮人耳聽覺和發音系統特性發展而成，在不受干擾的情況下，相較於其他方法有較佳的辨識能力，因此有許多強健性語音特徵研究都是針對其發展。

而 MFCC 特徵擷取的過程有預強調(Pre-emphasis)、音框化(Windowing)、離散傅立葉轉換(Discrete Fourier Transform, DFT)、梅爾三角濾波器組處理(Mel-Scaled Triangular Filterbank Processing)、離散餘弦轉換(Discrete Cosine Transform, DCT)、對數能量(Log Energy)運算及時間差量(Time Derivation)運算等程序，流程如圖 1-2 所示。

由於現實生活中，往往會因為環境中諸多複雜因素影響，導致訓練語料會有與測試語料在環境中不匹配(Mismatch)的問題，使得辨識率大幅下降。因此，在取得特徵向量後，便可考慮環境或噪音對語音特徵的影響，進一步對特徵向量進行校正以去除雜訊。

(二)聲學模型(Acoustic Model)

在數字辨識、關鍵詞辨識等小詞彙數辨識任務中，通常以全詞模型(Whole-word Model)當作聲學模型的單位。而在中、大詞彙數辨識任務中，考慮到訓練語料的收集與聲學模型的一般化能力(Generalization Ability)，因此不會使用詞典中的每個詞去建立單獨的聲學模型，取而代之的是使用比詞更小的單位去建立模型，如子詞(Sub-word Unit)單位、音素(Phone)或音節(Syllable)等，接著，利用發音詞典(Pronunciation Lexicon)來串接每一詞彙所對應的每個聲學模型。由於中文的每個字(Character)皆是以一個音節所組成，因此設計中文語音辨識器時會將中文音節再細分為聲母(Initial)及韻母(Final)兩種聲學單位，也可稱為子音(Consonant)和母音(Vowel)，分別建立聲母模型和韻母模型。

因語音是具有時序性的，一般而言皆採用由左至右(Left-to-right)的隱藏式馬可夫模型(Hidden Markov Model, HMM)[Rabiner, 1989]來建立聲學模型。如圖 1-3，它便是一個具有三個狀態的隱藏式馬可夫模型；每個狀態中都會有對每個音框(Frame)所形成的語音特徵參數向量之觀測機率分佈(Observation Probability Distribution)和相對應的狀態轉移機率(State Transition Probability)，用來決定是否要停留在此狀態或是轉移到下個狀態。

而常見的聲學模型訓練方法有最大化相似度訓練法(Maximum Likelihood, ML) [Bahl et al., 1983]、最大化交互資訊(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。

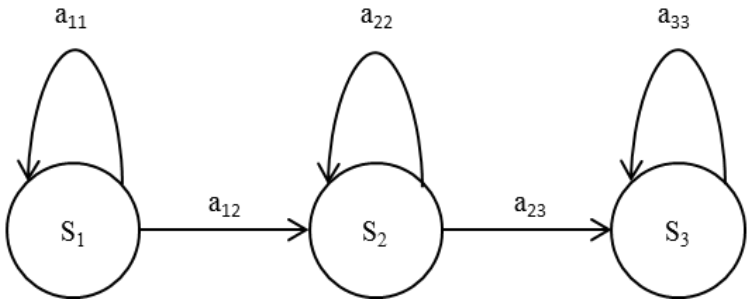


圖 1-3：狀態數為 3 的 Left-to-right 隱藏式馬可夫模型範例

（三）語言模型(Language Model)

不同於聲學模型，語言模型的目的是企圖描述語言的特性，並希望能夠預測語音訊號中的下一個字，由於聲學模型缺少了語句中詞與詞之間的資訊，只能辨識某一段語音訊號與音節、音素或詞的相似程度，有鑑於此，需要利用一個語言模型來估測詞與詞之間的連接關係。在語音辨識過程中，詞彙數是有限的，所以可能得詞序列組合也是可數的。因此，語言模型常使用多項式分佈(Multinomial Distribution)，並且它被用於直接估測詞序列的機率質量函數(Probability Mass Function, PMF)。一段詞序列 W 的事前機率可以用鏈鎖率(Chain Rule)展開，得到條件機率的連乘積：

$$\begin{aligned} P(W) &= P(w_1, w_2, ..., w_m) \\ &= P(w_1)P(w_2 \mid w_1)...P(w_m \mid w_1, w_2, ..., w_{m-1}) \\ &= P(w_1)\prod_{i=2}^{m} P(w_i \mid w_1, w_2, ..., w_{i-1}) \end{aligned} \tag{1-2}$$

這表示每個詞都會與其所看過的歷史詞有關，然而，當 m 較大時，所需的參數空

間將會根據歷史詞序列呈指數成長，因此語言模型很難去估測或儲存所有的資訊。目前最廣泛使用的語言模型為 N 連語言模型，它基於已知的前 N-1 詞來預測下一個詞出現的可能性，即所謂的 N-1 階馬可夫假設(N-1 order Markov Assumption)。

一般為了減少參數量的複雜度，常使用二連詞(Bigram)及三連詞(Trigram)，卻使得語言模型無法獲得更多的歷史資訊或長距離資訊。而另一方面，如果將 N 的大小增加，不僅會造成參數量呈指數倍成長，當實際使用時，在時間、空間複雜度也都會有所影響，並且會導致語言模型招致訓練資料稀疏的問題，需要更多的訓練語料來彌補。換言之，在語言模型中就會有詞的機率為零因而無法正確估測，加上要評估一個詞序列發生的可能性，由於詞的機率是連乘的，當中有一詞之條件機率為零的話會導致詞序列機率也為零，就可能導致辨識錯誤。解決此問題的方法有平滑化(Smoothing)技術及分群(Clustering)，就平滑化技術而言，常見的有 Katz 平滑化法[Katz, 1987]和 Good-Turning 平滑化法[Good, 1953; Chen and Goodman, 1999]等，其概念是將訓練語料中每個詞序列出現的統計次數依照各式比例原則折扣部份次數，再把這些次數以各式比例原則分派給在訓練語料中沒有出現的其它詞序列，以解決機率為零之問題。另外，群集模型(Clustering Model)則是使相似的詞聚集在同一類別，以解決資料稀疏的問題。例如：有兩句話「有個約會在星期五」和「有個約會在星期六」，則我們可以想像「星期五」和「星期六」在此為相似的詞，因此可以視為同一類別。

當語言模型使用於語音辨識時，它不僅可解決聲學混淆的問題與限制辨識的搜尋空間，更重要的是，它能藉由不同上下文或者其它資訊來預測每個詞可能出現的機率分布，並輔助語音辨識器評估各個候選詞序列在自然語言中的合理性，因而找出最有可能之候選詞序列。

（四）語言解碼(Linguistic Decoding)

透過輸入語句的特徵向量在對應聲學模型上之相似度，及所形成詞序列之語言模型機率，我們可以找出最有可能的詞序列。一般我們會使用維特比動態規劃搜尋(Viterbi Dynamic Programming Search)[Viterbi, 1967]，結合聲學相似度和語言模型之機率去有效率地找出可能的詞序列。此外，由於龐大的詞彙量與複雜的語言模型會使得搜尋的空間呈現指數型態的成長，為了降低搜尋的空間複雜度及加快辨識速度，一般會分成兩個階段做處理。第一階段使用聲學模型和較低階的語言模型進行詞彙樹動態規劃搜尋，以及利用語言模型前看(Language Model Look-ahead)[Aubert, 2002]技術、聲學前看與光束剪裁等技術來去蕪存菁，捨棄機率較低的詞序列來產生最佳詞序列和詞圖(Word Graph)。第二階段則是對詞圖重新進行動態規劃搜尋，一般通稱為重新計分(Rescoring)，並且使用更高階的語言模型，結合其它的信心度分數(Confidence Score)來找出 M 條最佳詞序列(M-Best)。由於詞圖所產生的最佳詞序列，並不一定是字錯誤率最低的詞序列。因此有許多研究透過語言模型來來進行重新排序(Reranking)，期望能從 M 條最佳詞序列中找出字錯誤率最低的詞序列，以做為最後的輸出結果。

1.3 研究動機與目的

為了使電腦能理解人類的語言，本論文研究語音辨識中的語言模型，希望藉由語言模型能捕捉語言的規律性。*N* 連語言模型是較常見的語言模型之一，它易於產生且容易使用的特性引發許多研究學者使用。但此語言模型有資料稀疏與缺乏長距離資訊等問題，因此有不同類型的語言模型產生並期望解決這些問題，前饋式類神經網路語言模型(Neural Network-based Language Models, NNLM)則是其中之一。它將歷史詞序列的資訊投影到連續空間，借以解決資料稀疏的問題，但對於長距離資訊的取得仍不盡理想。因此，為了獲得長距離的資訊，有了遞迴式類神經網路語言模型(Recurrent Neural Network-based Language Models, RNNLM)的產生。1994 年有研究[Bengio et al., 1994]指出，遞迴式類神經網路較難取得更長距離的資訊，其理由是當句子越長時，越遠距離的資訊透過機率相乘所得到的值會趨近於零。在本論文，我們嘗試使用額外的資訊來增進遞迴式類神經網路語言模型的預測能力。例如，使用句子和句子間的關聯性或詞與詞之間的關聯性來協助預測下一個詞發生的機率。

1.4 論文貢獻

本論文貢獻有以下幾點：

（一） 本論文透過彙整國內外當今的研究，探索常見的類神經網路語言模型於中文語音辨識之成效，期許能拋磚引玉，提供中文語音辨識研究領域具價值的參考。

（二） 一般前饋式類神經網路語言模型解決了 N 連語言模型資料稀疏的問題，它將歷史詞序列映射到連續空間，以此來估測下一個詞出現的機率為何。但前饋式類神經網路語言模型仍然缺乏長距離資訊，因此有遞迴式類神經網路語言模型的發展，其希望將歷史詞資訊透過遞迴的方式儲存起來以獲得長距離資訊。而遞迴式類神經網路語言模型的確能獲得長距離資訊，但隨著距離越遠，梯度下降法中鏈鎖率的長度就會越長。換句話說，小於 1 的機率值相乘會因此越乘越小，導致回饋給遠距離資訊的權重也較小，最後則缺乏了長距離的資訊，所以本論文分兩部分解決此問題。

i. 本論文提出利用語句關聯以及詞關聯資訊來加以輔助遞迴式類神經網路語言模型；其中，探討了如何表示關聯資訊，並且驗證了關聯資訊的使用能提升語音辨識率與降低語言複雜度。

ii. 由於在過去的研究裡，所有測試語句皆使用由同一份訓練語料所訓練出的單一語言模型，因此本論文希望能夠使測試語句找到符合其特性的語言模型，並透過動態調整語言模型來得到更好的成效。

（三） 本論文使用額外資訊於遞迴式類神經網路語言模型在大詞彙連續語音辨識中有相當程度的改善。

1.5 論文章節安排

本論文接下來的章節安排如下：

第 2 章介紹 *N* 連語言模型與其它不同種類的語言模型，並回顧類神經網路語言模型之相關進展。

第 3 章介紹類神經網路語言模型於自動語音辨識之使用，並且說明類神經網路語言模型相關理論及架構。

第 4 章介紹本論文所使用的詞關聯資訊和語句相關資訊於遞迴式類神經網路語言模型。

第 5 章介紹實驗語料、實驗設定以及實驗結果分析。

第 6 章則是結論及未來展望。

第2章　文獻探討與分析

2.1　*N* 連語言模型

語言模型在自然語言處理中佔有舉足輕重的角色，時常應用於相關領域諸如機器翻譯(Machine Translation)、資訊檢索(Information Retrieval)等研究，其中在自動語音辨識領域更是影響顯著。語言模型主要的功能是擷取語言的特性並預測語句中每一個詞出現的可能性。另外，根據語言的不同以及需求不同，也會發展出不同類型的語言模型。

現今最常見的統計式語言模型，是透過機率模型的建立來描述語言生成的規律性，1948 年克勞德 • 香儂(Claude Elwood Shannon)提出了使用馬可夫鏈於連續英文字母中來產生統計式模型，探討每一個字母的出現與其前 *N*-1 個字母有關，其想法因而發展出了 *N* 連語言模型(*N*-gram Language Model)。更簡單來說，它根據前 *N*-1 個歷史詞序列來預測第 *N* 個詞，以機率的方式來呈現則可以表示成下列式子：

$$P(w_1,\ldots,w_n)=\prod_{i=1}^{n}P(w_i \mid w_1,\ldots,w_{i-1})\approx\prod_{i=1}^{n}P(w_i \mid w_{i-N+1},\ldots,w_{i-1}) \qquad (2\text{-}1)$$

而 $P(w_1,\ldots,w_n)$ 代表一段詞序列，我們可將它拆解成一連串條件機率的連乘積。再經由簡化得到式(2-2)。由於詞序列有相當多種排列組合，使得 *N* 連語言模型的參數量相當龐大。因此，*N* 連語言模型常會限制 *N* 的大小，以至於缺乏長距離的資訊。

$$\prod_{i=1}^{n}P(w_i \mid w_1,\ldots,w_{i-1})\approx\prod_{i=1}^{n}P(w_i \mid w_{i-N+1},\ldots,w_{i-1}) \qquad (2\text{-}2)$$

其中 *N* 連語言模型的訓練方式為最大化相似度估測法(Maximum Likelihood

Estimation, MLE)，它藉由訓練語料中 N 連詞出現的次數(Word Count)來估測 N 連詞的機率分佈，以三連詞為例：

$$P(w_i \mid w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})} \tag{2-3}$$

其中，$C(\cdot)$表示詞序列在訓練語料中出現的次數。

N 連語言模型雖然容易產生以及使用，但仍有許多缺點需改善，如缺乏了長距離的詞彙資訊、維度的詛咒(Curse of Dimensionality)及資料稀疏等問題。而為了改善 N 連語言模型發生的問題，也有許多專家學者提出不同的模型來改進，接下來將介紹其它基於不同層次資訊的語言模型。

2.2 其它基於不同層次資訊之語言模型

依照語言資訊的不同，我們可將語言模型大致分成四類[邱炫盛， 2007]：

(一) 詞相關語言模型：由於傳統 N 連語言模型根據馬可夫假設而僅能獲得短距離的資訊，而為了改進此缺點，這一類型的語言模型試著獲得更長距離的詞彙資訊。如：快取模型(Cache Model)[Kuhn, 1988]、混階層馬可夫模型(Mixed-order Markov Model)[Saul and Pereira, 1997]和觸發對語言模型(Trigger-based Language Model)[Troncoso et al., 2004]等。

(二) 詞類別相關語言模型：使每個詞都有屬於自己的類別，而相同類別的詞則代表具有相似的意義。藉由建立詞與詞之間的關係，找出序列中的詞與欲預測詞之間的關係，因此透過詞類別的加入，原本 N 連語言模型中資料稀疏的問題得以解決。例如：類別 N 連語言模型(Class-based N-gram Model)[Brown et al., 1992]與聚合式馬可夫模型(Aggregate Markov Model, AMM)[Troncoso et al., 2004]等。

(三) 語句結構相關語言模型：此類型的語言模型透過自然語言處理的觀點，使用句法結構來擷取歷史詞資訊中有意義的資訊，使其具有長距離資訊。相關模型如：結構化語言模型(Structured Language Model)[Chelba and Jelinek, 2000]等。

(四) 文件主題相關語言模型：概念類似詞類別相關語言模型，將一篇或一群文件根據主題性建立模型。歷史詞序列可視為尚未完成的文件，假設完成的程度可以表現出某些主題，透過此模型找出其相關的主題性。例如：混合式語言模型(Mixture-based Language Model)[Clarkson and Robinson, 1997]、潛藏語意分析(Laten Semantic Analysis, LSA)[Bellegarda, 2005]、機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA) [Gildea and Hofmann, 1999]和潛藏狄利克里分配(Latent Dirichlet Allocation, LDA) [Tam and Schultz, 2005]等。

此外，還有許多不同的語言模型發展出來，在接下來的章節中本論文將介紹鑑別式語言模型。

2.3 新近所提出之語言模型

2.3.1 鑑別式語言模型

不同於傳統統計式語言模型，鑑別式語言模型的目標為最小化語音辨識錯誤率，由於傳統統計式語言模型所選出的辨識結果通常是發生機率最高的詞序列，而非字錯誤率最低的詞序列。因此鑑別式語言模型希望藉由詞圖(Word Graph or Lattice)來產生 *M* 條最佳辨識候選詞序列(*M*-best list)，將其重新排序(Reranking)來找尋字錯誤率最低的詞序列，而獲得更好的辨識率。由於鑑別式訓練(Discriminative Training)的產生，發展出許多相關研究，鑑別式語言模型初期則應用於機器翻譯(Machine Translation, MT)、自然語言處理與聲學模型等研究。鑑別式語言模型主要可分為兩種研究，其一是以模型訓練方式；其二則是特徵的選用。

- 模型訓練方式

主要是針對目標函數的不同給予不同的學習機制或演算法，此部份常見的鑑別式語言模型有感知器演算法(Perceptron)[Rosenblatt, 1958]、最小化錯誤率訓練(Minimum Error Rate Training, MERT)[Och, 2003]、全域條件式對數線性模型(Global Conditional Log-linear Model, GCLM)[Roark et al., 2007]及權重式全域條件式對數線性模型(Weighted Global Conditional Log-linear Model, WGCLM)[Oba et al., 2010]等。

感知器演算法(Perceptron)的起源是從類神經網路開始發展，在 2002 年，美國學者 Collins[Collins, 2002]將感知器演算法應用於自然語言處理中，並於 2005 年被應用在語言模型調適[Gao et al., 2005]上。感知器演算法是以最小平方誤差法(Least Squared Error, LSE)來作為排序減損函數(Loss Function)，其希望排序分數最高的候選詞序列與最低錯誤率的詞序列之分數差平方後越小越好。然而感知器演算法只考慮了目前排序分數最高的詞序列與最低錯誤率詞序列的關係，因此一般

化能力較差，且會有過度訓練(Over-Training)的問題及未必可找到全域最佳解。反觀其好處則是演算法簡單易操作，並且因為不用考慮樣本權重而有較快的訓練速度。

不同於感知器演算法，最小化錯誤率訓練(MERT)的目標是最小化語音辨識器辨識錯誤率的期望值，也就是說，希望在經過重新排序後，整體的字錯誤率能夠越小越好。其最小化錯誤率訓練中的錯誤率，可以視為一種樣本權重(Sample Weight)的資訊，用於區別每一個候選詞序列對於鑑別式語言模型訓練時的重要性。最小化錯誤訓練是於 2003 年由 Och 所提出且應用在機器翻譯領域裡，而 2008 年時，Kobayashi 等學者[Kobayashi et al., 2008]則將語音辨識領域和此方法做結合。此方法不僅考慮了排序分數最高與擁有最低錯誤率的詞序列，也同時考慮了其他候選詞序列的錯誤率，因此會有較佳的一般化能力，但也因為同時考慮了所有候選詞序列，造成訓練速度較慢。

全域條件式對數線性模型(GCLM)的訓練目標則是希望最低錯誤率詞序列的條件機率越高越好，其概念是於 2007 年 Roark 等學者以有限狀態機(Weighted Finite State Automata, WFSA)實作全域條件式對數線性模型，並且應用於語音辨識結果的重新排序上。由於全域條件式對數線性模型考慮了最低錯誤率詞序列與其它候選詞序列的關係，因此較不會出現過度訓練的問題，而一般化能力則是介於感知器演算法與最小化錯誤率訓練之間。

權重式全域條件式對數線性模型(WGCLM)則是全域條件式對數線性模型的延伸，在 2010 年由 Oba 等學者將樣本權重加入全域條件式對數線性模型進行改進，將每個候選詞序列的分數加上一個不同的權重，以此來表示每一條候選詞序列不同的重要程度，使每個候選詞序列對於訓練有不同的影響力。而錯誤率越高或是排序越後面者，則重要程度就越重、影響力就越大。表 2-1 為各種鑑別式語

言模型之比較，其中 L 為訓練語料的總句數、W_i^R 為最佳候選詞序列中最低錯誤率的詞序列、W_i^* 為排序後分數最高的詞序列、λ 為特徵權重參數向量、函數 $Score$ 為將兩向量內積後的分數、$\omega_{W_{i,k}}$ 則為第 i 句和第 k 句的字錯誤率。

	是否考慮權重樣本	是否考慮 W_i^R	一般化能力	有無全域最佳解	訓練速度
Perceptron	否	是	差	否	快
MERT	是	否	佳	否	慢
GCLM	否	是	略佳	是	慢
WGCLM	是	是	略佳	是	慢

	目標函數
Perceptron	$F_{\text{Perc}}(\lambda)=\frac{1}{2}\sum_{i=1}^{L}\left(Score(W_i^R,\lambda)-Score(W_i^*,\lambda)\right)^2$
MERT	$F_{\text{MERT}}(\lambda)=\sum_{i=1}^{L}\sum_{k=1}^{M}\frac{\omega_{W_{i,k}}\cdot\exp\left(Score(W_{i,k},\lambda)-Score(W_i^R,\lambda)\right)^{\beta}}{\sum_{j=1}^{M}\exp\left(Score(W_{i,j})-Score(W_i^R,\lambda)\right)^{\beta}}$
GCLM	$F_{\text{GCLM}}(\lambda)=\sum_{i=1}^{L}\log\frac{\exp\left(Score(W_i^R,\lambda)\right)}{\sum_{j=1}^{M}\exp\left(Score(W_{i,j},\lambda)\right)}$
WGCLM	$F_{\text{WGCLM}}(\lambda)=\sum_{i=1}^{L}\log\frac{\exp\left(Score(W_i^R,\lambda)\right)}{\sum_{j=1}^{M}\omega_{W_{i,j}}\exp\left(Score(W_{i,j},\lambda)\right)}$

表 2-1：各種鑑別式語言模型之比較

- 特徵的選用

由於傳統的方法是使用候選詞序列中 N 連詞的次數作為特徵，但是此方法缺乏了詞的特性、句法結構或語音訊號的特性。因此 Huang 等學者[Huang et al., 2007][Arisoy et al., 2010][Sak et al., 2010]使用了詞性(Part-of-speech, POS)、句法結構或韻律(Prosody)來作為特徵，期望能夠使用越詳細的資訊來表示候選詞序列。

2.3.2 類神經網路語言模型

2.3.2.1 類神經網路簡介

類神經網路(Neural Networks)起源於人工智慧(Artificial Intelligence)，又可稱為人工類神經網路(Artificial Neural Networks, ANN)。為了讓電腦具備與人類一樣的能力，自 1940 年開始科學家開始模仿神經元(Neuron)的運作模式，認為如果兩個神經元同時被觸發，則它們之間的連結就會獲得增強。從巴伐洛夫的狗與鈴聲的實驗中就可得知，當聽到鈴聲的神經元和看到食物的神經元同時受到刺激時，兩神經元間就會建立起增強的學習關係，此現象也造就了類神經網路的基礎。直到近年來，類神經網路結合了各項領域，如資訊、工商業甚至心理學等都有不錯的成效，其中像是感知器演算法(Perceptron)是第一個實踐出類神經網路的創舉。然而為何類神經網路在近年來興起一股流行呢？許多學者歸類出類神經網路有以下幾點特性：

(一) 具備平行處理的能力

(二) 容忍錯誤(Fault Tolerance)的能力

(三) 擁有學習、圖形辨識、自我調適和結合式記憶(Associative Memory)的能力

(四) 可以解決最佳化和處理一般演算法難以解決之問題

(五) 可以以硬體線路，如超大型積體電路(VLSI Implementation)來實作

(六) 利用了非線性的運算和具有嚴謹的數學基礎

除了前述優點外，類神經網路也有許多機器學習中會遇到的問題，如過度訓

練(Over Training)或訓練不足(Under Training)。若造成過度訓練，可能導致將訓練

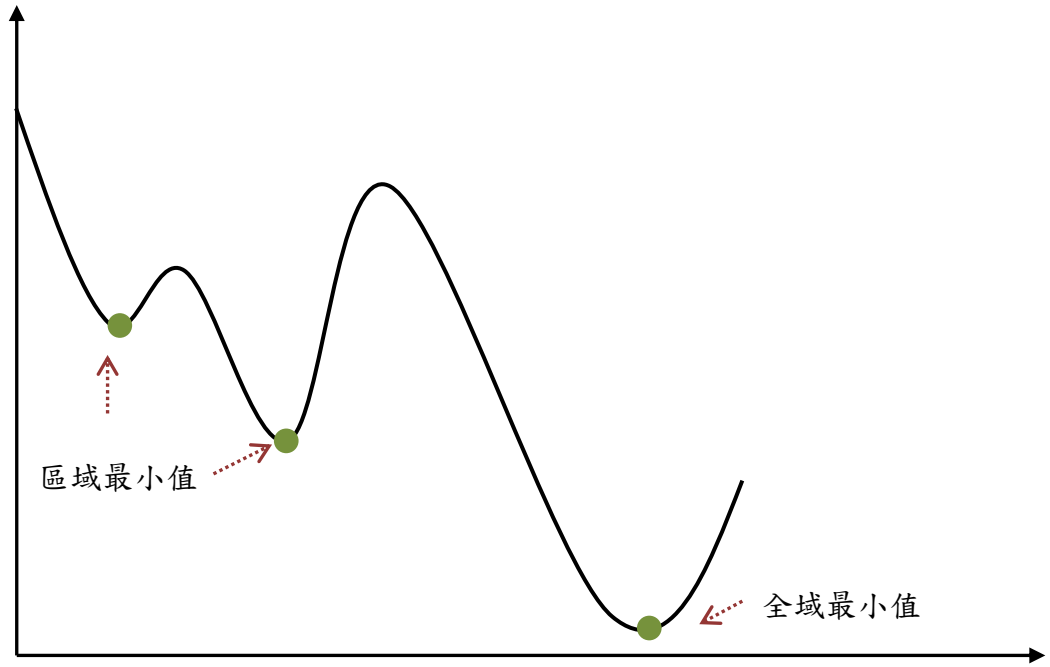


圖 2-1：區域最小值示意圖

資料中的雜訊學習進去，導致未看過的資料較難作預測；反之，若訓練不足也無法做出正確的預測。而隱藏層及隱藏層內神經元數目的取決也會造成一定的影響，一般來說兩層的隱藏層即可處理任何問題[Villiers and Barnard, 1992]。當隱藏層數目太多則複雜度較高，學習時間也相對增加，並且容易找到區域最小值(Local Minimum)，而非最佳解；數目太少則會難以收斂。區域最小值的問題可以圖 2-1 來示意，類神經網路在尋找全域最小值(Global Minimum)所採用的方法為梯度下降法(Gradient Decent Method)，但此法仍有可能會找尋到區域最小值。而除了隱藏層中的神經元太少會導致無法收斂外，訓練資料內有互相矛盾或有極端狀況、訓練資料中的排列順序或學習率(Learning Rate)太大所造成的震盪或誤差容忍度設定太小都可能是無法收斂的原因。

另外根據學習方法的不同，可分下列幾種[Rojas, 1996]：

- 監督式學習網路(Supervised Learning Network)
  - 感知機網路(Perceptron)
  - 倒傳遞式網路(Back-Propagation Neural Network, BPN)
  - 學習向量量化網路(Learning Vector Quantization, LVQ)
  - 機率式神經網路(Probabilistic Neural Network, PNN)
  - 反傳遞網路(Counter-Propagation Network, CPN)
- 非監督式學習網路(Unsupervised Learning Network)
  - 自組織映射圖網路(Self-Organizing Map, SOM)
  - 自適應共振理論網路(Adaptive Resonance Theory Network, ART)
- 聯想式學習網路(Associate Learning Network)
  - 霍普菲爾網路(Hopfield Neural Network, HNN)
  - 雙向聯想記憶網路(Bi-directional Associative Memory, BAM)
- 最適化應用網路(Optimization Application Network)。
  - 霍普菲爾-坦克網路(Hopfield-Tank Neural Network, HTN)
  - 退火神經網路(Annealed Neural Network, ANN)

其中，依其架構可分兩類，分別是前饋式架構(Feed-Forward Network)和遞迴式架構(Recurrent Network)或稱為回饋式架構(Feed-Back Network)，兩者差別在於前者只有從輸入傳遞到輸出，而後者會多增加一個步驟，將上一時間點的資訊傳回給網路。

目前類神經網路主要被用於分類及預測上，在影像辨識方面，如圖案的辨識或雜訊的處理等，而在語音辨識中則有語言模型、語音合成與強健性語音辨識等。另外則是氣象預測、電腦輔助教學、手寫辨識以及超大積體電路的應用。本論文則是探討語音辨識裡的語言模型部份，以下則介紹類神經網路語言模型的演進與改進。

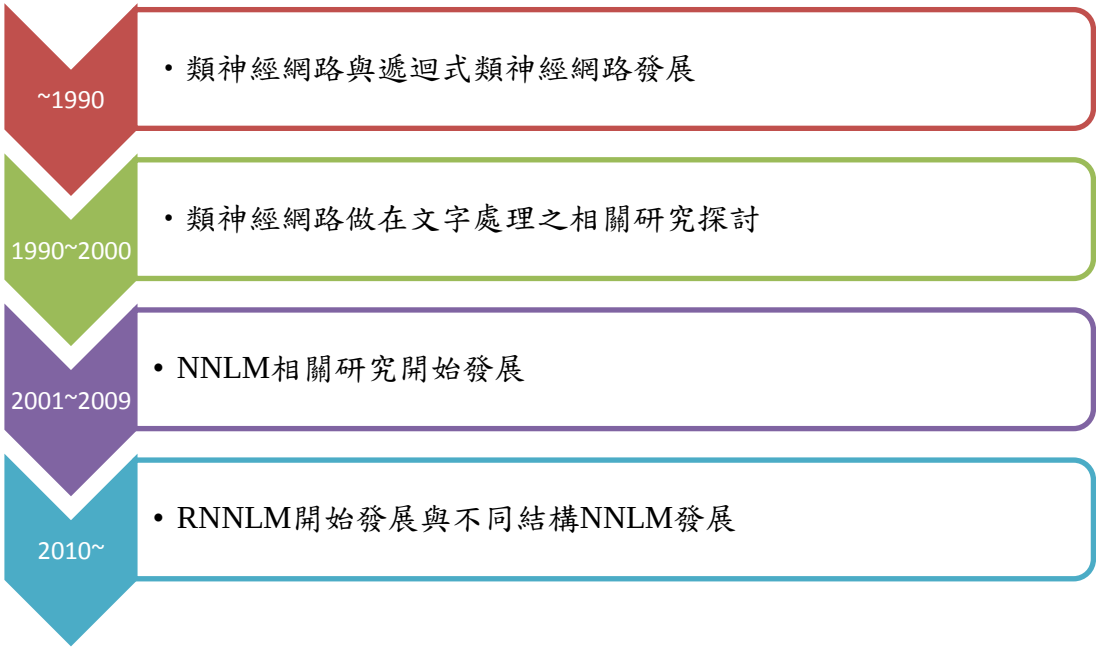


圖 2-2：類神經網路語言模型演進圖

2.3.3 類神經網路語言模型文獻探討

類神經網路語言模型是一種語言模型建立於類神經網路上，透過將詞以連續表示(Continuous Representation)來解決 *N* 連語言模型維度詛咒之影響，另外對未知詞的估測能力也較 *N* 連語言模型來的好。

而從類神經網路出現開始，就有許多研究學者將自然語言處理與類神經網路做結合，Towsey [Towsey et al., 1998]使用遞迴式類神經網路來預測句中下一個詞的詞性。實驗結果顯示，長距離的詞序列仍有較多的誤差，但遞迴式類神經網路

仍能根據前三個詞或更多歷史詞資訊來做預測。

另外值得注意的是，遞迴式類神經網路不同於傳統前饋式類神經網路，它的目標是希望獲得長距離的資訊，但 Bengio 等學者[Bengio et al., 1993, 1994]發現，利用梯度下降法(Gradient Descent Method)於遞迴式類神經網路中，對於學習長距離的資訊是十分困難的。而要獲得長距離資訊必須要具有學習任意時間內的資訊，且擁有抵抗其它資訊干擾的能力。但因為隨著時間變化，距離較遠的資訊會被每一次時間點的輸入資訊所干擾，反而降低了遞迴式結構的好處。

在 1996 年，Lawrence [Lawrence et al., 1996]等學者，調查了數種遞迴式類神經網路。實驗結果中顯示，由艾爾曼網路(Elman Network)建構的遞迴式類神經網路對於學習適當的文法有不錯的成效，可以見得遞迴式類神經網路建構出階層式的網路，幫助了句法或文法上的學習。

時間進入到 2000 年，逐漸有學者應用類神經網路於語言模型上，其中 Xu 和 Rudnicky [Xu and Rudnicky, 2000]比較了類神經網路語言模型和傳統 *N* 連語言模型，實驗結果顯示類神經網路的學習能力的確超越了傳統 *N* 連語言模型。雖然在語言複雜度(Perplexity)上類神經網路語言模型有不錯的成效，但是花費在估測的時間上仍較傳統 *N* 連語言模型來的高許多。

Bengio [Bengio et al., 2001]於 2001 年則是使用了前饋式類神經網路於固定長度的上下文上，透過類神經網路將維度降低，達到比傳統統計式模型還要好的成效，也發現到此方法有益於較長的上下文與有不錯的一般化成果。

Goodman [Goodman, 2001]則將類神經網路語言模型與其他語言模型做比較，發現此模型比混合許多模型的結果來得佳，如快取模型(Cache Model)及類別模型(Class-based Model)，之後 Schwenk [Schwenk et al., 2004, 2005, 2007]將之應用在

語音辨識上，大幅改善了基礎實驗的結果。

但是類神經網路語言模型仍有幾個主要的缺點需要改進，首先是令人詬病的運算複雜度。由於是以詞為單位來訓練模型，加上隱藏層(Hidden Layer)至輸出層(Output Layer)之間的大量運算，造成時間複雜度較高。另外，對於詞的表示方式沒有考慮到其他額外的資訊，如：詞性或聲調等。除此之外，面對 OOV 的問題類神經網路語言模型也沒有一個有效的解決方法，因此，Alexandrescu 等學者[Alexandrescu and Kirchhoff, 2006]將每個詞都各自對應到一個特徵向量，每個向量的維度則代表許多特徵，像是詞性或大小寫等等。如此一來，面對未曾出現過的詞也有辦法找出其特徵向量，並且做到正確的估測。

2009 年，Mikolov 等學者 [Mikolov et al., 2009]使用類神經網路語言模型於屈折語(Inflective Language)的語言上，由於屈折語的詞綴具備有多種意思，因此此種語言視為相當挑戰的任務；而此篇論文利用不同大小的神經元數目和後撤式(Backoff)語言模型做比較，效果相當顯著。

同年，Zamora-Martínez 等學者[Zamora-Martínez et al., 2009]針對類神經網路語言模型龐大的時間複雜度進行了改進。他們將可事先運算好的資料儲存起來，以空間換取時間上的效率，並且提出了階層式(Hierarchy)的概念，將不同高階和低階的類神經網路語言模型做結合，以達到類似快取(Cache)的概念。

隔年，Park 等學者[Park et al., 2010]將類神經網路語言模型做了一點改進，他們將輸入層(Input Layer)加入一個維度來訓練遇到 OOV 之情形，並在估測機率時，使用不同類型的平滑化技術來比較。此外，也加入了一層適應層(Adaptation Layer)，期望加強類神經網路語言模型的適應能力，實驗結果於大詞彙語音辨識有不錯的提昇。

而 Mikolov 等學者[Mikolov et al., 2010]在 2010 年時遂結合語言模型與遞迴式類神經網路，並發展了遞迴式類神經網路語言模型套件供學者下載，實驗結果發現遞迴式類神經網路語言模型有顯著的成效，若結合 *N* 連語言模型則有更進一步的提昇。

2011 年時，Mikolov 等學者[Mikolov et al., 2011]將前一年所提出的遞迴式類神經網路語言模型做了延伸，引入了 Goodman 等學者[Goodman et al., 2001]所提出的概念。將輸出層額外分解出一層類別層，使隱藏層和輸出層間的運算大幅減少，另外也在架構中加入一層壓縮層(Compression Layer)，雖然成效比之前稍差了一點，但卻大幅提升了運算速度。此外 Le 等學者[Le et al., 2011]則在類神經網路語言模型的輸出層做了結構上的改進，利用分群以及決策樹的概念去估測機率。除了在輸出層做改進的研究外，Kang 等學者[Kang et al., 2011]對類神經網路語言模型在輸入層改為由字和詞混合當作輸入做了改進，並應用於大詞彙連續語音辨識。

另一部分，也有研究學者將現行的語言模型和類神經網路語言模型或遞迴式類神經網路語言模型做結合，例如傳統的 *N* 連語言模型[Oparin et al., 2012]、最大熵值法(Maximum Entropy)[Mikolov et al., 2011]與快取語言模型[Zamora-Martínez et al., 2012]等。

而除了 Mikolov，許多學者也將類神經網路語言模型作為比較標準，如 Sarikaya 等日本學者[Sarikaya et al., 2010]將共享混和語言模型(Tie-Mixture Language Modeling)和類神經網路語言模型做比較以及 Mikolov 等學者[Mikolov et al., 2011]將常見的語言模型做逐一比較。因此，我們可以得知類神經網路語言模型之研究越趨重要。

由文獻中我們發現傳統前饋式類神經網路語言模型與遞迴式類神經網路語

言模型有下列缺點：無法有效地獲得長距離資訊、OOV 問題、欠缺適應能力、運算的時間複雜度過高以及詞的表示方式缺少了詞的特性。而許多研究學者針對這些問題進行改進，也可另外分成以下幾類[Oparin et al., 2012]：

- 對架構進行改進

  從微觀來看，可將輸入層或輸出層進行改進，或是新增加幾層來做延伸；巨觀來看則是將整體的架構進行改進，如遞迴式或階層式的方式，抑或是和其他語言模型做結合。

- 對演算法進行改進

  由於類神經網路也是一種模型學習的方法，則可以利用不同的演算法來進行改良，使模型更具有一般性能力或是適應性能力。

在下一章中，將介紹類神經網路語言模型於自動語音辨識之使用。

第3章　類神經網路語言模型於自動語音辨識之使用

本章將透過兩種較常見用於語言模型之類神經網路做介紹，分別是類神經網路語言模型(Neural Network-based Language Modeling, NNLM)和遞迴式類神經網路語言模型(Recurrent Neural Network-based Language Modeling, RNNLM)。

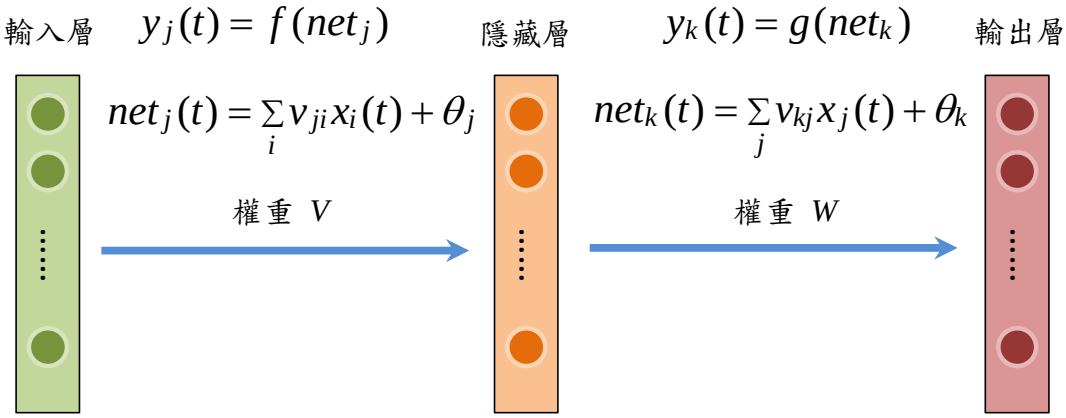


圖 3-1：前饋式類神經網路架構

3.1.1　倒傳遞式類神經網路

3.1.1.1　架構

將神經元彼此連結即可構成類神經網路，類神經網路主要有三層，分別是輸入層(Input Layer)、隱藏層(Hidden Layer)和輸出層(Output Layer)。有時會額外加入一層投影層(Projection Layer)，用來將歷史詞序列的資訊投影至此連續空間，並降低輸入層的維度。輸入層沒有具備運算能力，表示方式是以詞為單位由訓練資料來

給予，依照輸入的資料型態需先做前處理；隱藏層則是用來處理輸入單元傳遞進來的資料；輸出層用來表現網路的輸出結果。層跟層之間的神經元靠著突觸(Synapse)來傳遞訊息。轉化為圖 3-1 則可將各層以向量表示，層之間的突觸則以矩陣來表示。圖 3-1 就是簡單的類神經網路架構，也是前饋式類神經網路。接下來先介紹各層所代表的意義：

- 輸入層與隱藏層

輸入層、隱藏層及輸出層中包含了許多節點，輸入層中的節點以變數 i 表示，隱藏層的節點以變數 j 表示，輸出層的節點則以變數 k 表示。各層的節點結合後可形成一個向量，式(3-1)為各節點在向量中的表示方式。$x_i(t)$表示輸入層中於 t 時間點第 i 個節點，其中 i 為 1 到 N 之間表示輸入層大小，N 為詞彙的數量。

$$x_i(t)=\begin{cases}1 & \text{if 為對應詞的ID}\\ 0 & \text{其他}\end{cases} \tag{3-1}$$

在倒傳遞式類神經網路裡，向量 $\boldsymbol{x}$ 則會透過權重 V 來傳遞，而權重 V 會以矩陣的方式來表示，權重 V 中包含了許多輸入層節點和隱藏層節點間的鏈結權重值。式(3-2)為輸入層各節點傳遞至隱藏層中的節點 j。v_{ji} 是第 j 個隱藏層節點對第 i 個輸入層節點的鏈結權重值，θ_j 為第 j 個隱藏層節點的偏權值，net_j(t)為第 j 個隱藏層節點淨輸入值，$y_j(t)$則為第 j 個隱藏層節點。

$$\begin{aligned} net_j(t) &= \sum_i v_{ji}x_i(t)+\theta_j \\ y_j(t) &= f(net_j) \end{aligned} \tag{3-2}$$

其中， $f(net_j)$為網路的活化函數(Activation Function)，而激發函數大致分為下列幾種：雙彎曲函數(Sigmoid Function)、高斯主動函數(Gauss Action Function)、雙曲線正切函數(Hyperbolic Tangent Function)、片段線性函數(Piecewise-Linear

Function)和雙極性函數(Bipolar Function)。為了保證輸出值能介於 0 到 1 之間，本論文中所使用的是雙彎曲函數，如下面式子所表示：

$$f(x)=\frac{1}{1+e^{-x}} \tag{3-3}$$

- 隱藏層與輸出層

如同輸入層與隱藏層，隱藏層的各個節點會透過權重 *W* 傳遞給輸出層，可以透過式(3-4)來表示：

$$\begin{aligned} net_k(t) &= \sum_j w_{kj}\, y_j(t) + \theta_k \\ y_k(t) &= g(net_k) \end{aligned} \tag{3-4}$$

其中，w_{kj} 是第 k 個隱藏層節點對第 j 個輸入層節點的鏈結權重值，θ_k 為第 k 個隱藏層節點的偏權值，$net_k(t)$為第 k 個隱藏層節點淨輸入值，$y_k(t)$為第 k 個輸出層節點。為了使輸出層各節點的值總和為 1，最後的 $g(net_k)$為軟化最大值活化函數(Softmax Activation Function)，也是轉移函數(Transfer Function)的一種。如下式來表示：

$$g(x)=\frac{e^{x}}{\sum_k e^{k}} \tag{3-5}$$

3.1.1.2 倒傳遞演算法推導

在運算完輸出層的結果之後，便會由誤差函數(Error Function)計算出誤差向量 ***e*** (Error Vector) 以降低輸出值(Output Value)和期望值(Desired Value)間的差距，接著使用梯度下降法求取錯誤函數之最小值傳遞到層跟層之間的權重。而求得誤差向量的方法為期望向量 ***d*** (Desired Vector)減去輸出向量 ***y*** (Output Vector)，其中期望向量在類神經網路語言模型裡可視為下一個字所表示的向量。接著以下為使用梯

度下降法求得隱藏層與輸出層間的權重更新量ΔW之推導。

$$W(t+1)=W(t)+\Delta W \tag{3-6}$$

其中，$W(t+1)$為隱藏層跟輸出層間且時間點為t+1 的權重，$W(t)$為時間點t之權重，而ΔW則是我們想求得的。首先先介紹誤差函數的定義：

$$E(t)=\frac{1}{2}\sum_k e_k^2(t) \tag{3-7}$$

$$e_k(t)=d_k(t)-y_k(t) \tag{3-8}$$

其中，E為誤差函數，其使用的是均方誤差(Mean-Squared Error)，e_k表示誤差向量中第k個維度、d_k表示期望向量中第k維個維度和y_k表示輸出向量中第k個維度。接著利用梯度下降法對權重W偏微分可求得[Boden, 2002]：

$$\begin{aligned}\Delta W &= -\eta\frac{\partial E(t)}{\partial W}\\ &= -\eta\frac{\partial E(t)}{\partial y_k(t)}\frac{\partial y_k(t)}{\partial net_k(t)}\frac{\partial net_k(t)}{\partial W}\\ &= \eta e_k(t)\frac{\partial y_k(t)}{\partial net_k(t)}\frac{\partial net_k(t)}{\partial W}\\ &= \eta e_k(t)g'(net_k(t))\frac{\partial net_k(t)}{\partial W}\\ &= \eta e_k(t)g'(net_k(t))y_j(t)\end{aligned} \tag{3-9}$$

同理，我們也可以推導出輸入層與隱藏層間的權重更新量ΔV。$V(t+1)$為輸入層跟隱藏層間且時間點為t+1 的權重，$V(t)$為時間點t之權重。

$$V(t+1)=V(t)+\Delta V \tag{3-10}$$

$$
\begin{aligned}
\Delta V &= -\eta \frac{\partial E(t)}{\partial V} \\
&= -\eta \frac{\partial E(t)}{\partial y_k(t)} \frac{\partial y_k(t)}{\partial net_k(t)} \frac{\partial net_k(t)}{\partial y_j(t)} \frac{\partial y_j(t)}{\partial V} \\
&= \eta e_k(t) g'(net_k(t)) \frac{\partial net_k(t)}{\partial y_j(t)} \frac{\partial y_j(t)}{\partial V} \\
&= \eta e_k(t) g'(net_k(t)) W \frac{\partial y_j(t)}{\partial V} \\
&= e_k(t) g'(net_k(t)) W \cdot f'(net_j(t)) \sum_i x_i(t) \\
&= e_k(t) g'(net_k(t)) W \cdot net_j(t)(1 - net_j(t)) \sum_i x_i(t)
\end{aligned}
\tag{3-11}
$$

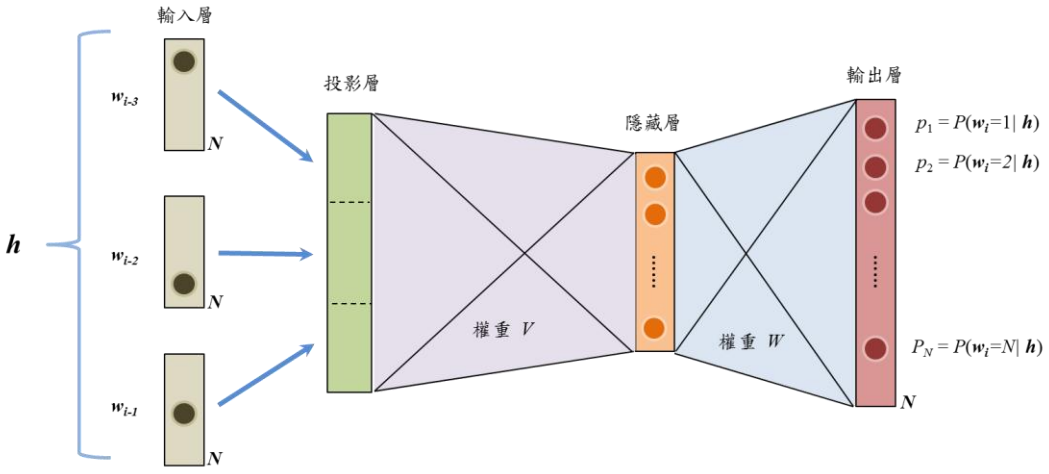


圖 3-2：類神經網路語言模型架構

3.1.1.3 類神經網路語言模型

將類神經網路與語言模型結合的話，則可表示成圖 3-2。輸入層則為欲預測詞的歷史資訊，其中歷史資訊以$\boldsymbol{h}$表示。如圖 3-2所示則代表前三個詞的歷史資訊，因此，此語言模型可視為一個四連的類神經網路語言模型。而每個詞使用one-of-N方式進行編碼，例如詞w_i的表示方式為在長度為N的向量中，只有第i維是1其餘為零。接著將歷史資訊映射到投影層內來進行降低維度的作用，透過圖 3-3可以實際了解到投影層能從投影層權重矩陣有效取得詞的權重資訊。將詞w_{i-3}、w_{i-2}和w_{i-1}透過投影層權重矩陣轉換成投影層，這部份可以了解到，詞序列的結構如果相似，則可使用相同的權重來予以估計。例如有兩段詞序列的語句結構相似，「狗在臥室裡奔跑(The dog runs in the bedroom)」和「貓在房裡走動(A cat walks in the room) 」。因此皆會使用到相同的權重，儘管未在訓練語料中出現，仍可獲得一個較佳的機率預測值，而不須藉由平滑化方法來協助估測下一個詞發生的可能性[Bengio et al. 2000]。不同於N連語言模型因為訓練語料有限，無法完整收集到所有可能的N連詞出現的統計資訊，進而造成資料稀疏的問題，投影層可以接受所有可能的詞序列組合，並且詞序列中的每個詞能獨立貢獻出權重值來估測下一個詞出現的可能性。因此投影層的目的即是將歷史詞序列的資訊映射到連續空間表示，另外，歷史詞序列中詞的順序關係不會改變，所以能讓隱藏層來學習。隱藏層的設計則和原本一樣，而隱藏層中神經元的數目則視為一個參數，需要去自行調整來獲得最佳的預測。輸出層設計也和原本一樣，其大小跟輸入層一樣，是詞彙的數量，不同的是每一維都會有各自預測出來的機率，可表示成$P(x_i \mid h)$。最後取機率最大的當作預測結果，再進行倒傳遞演算法來調整權重。

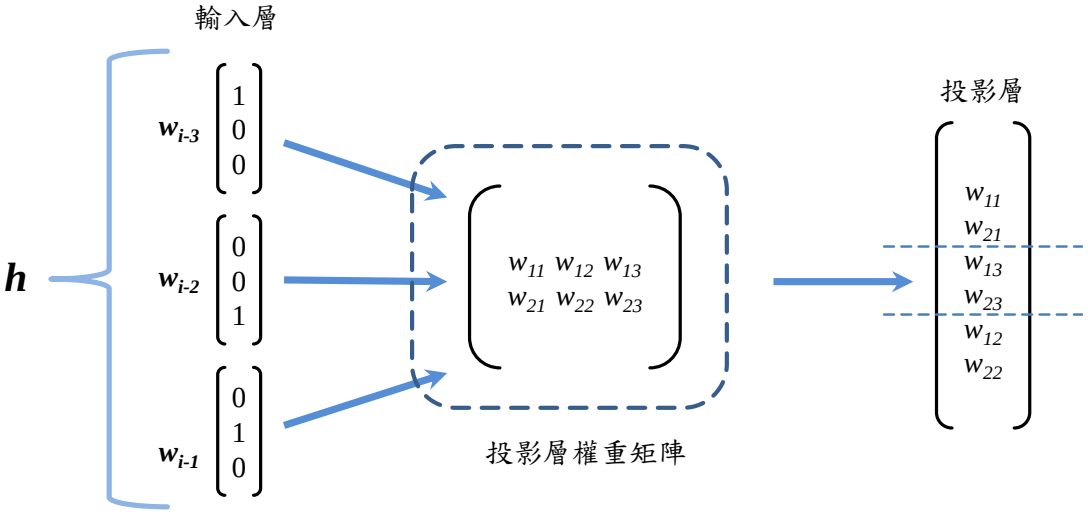


圖 3-3：輸入層映射至投影層過程

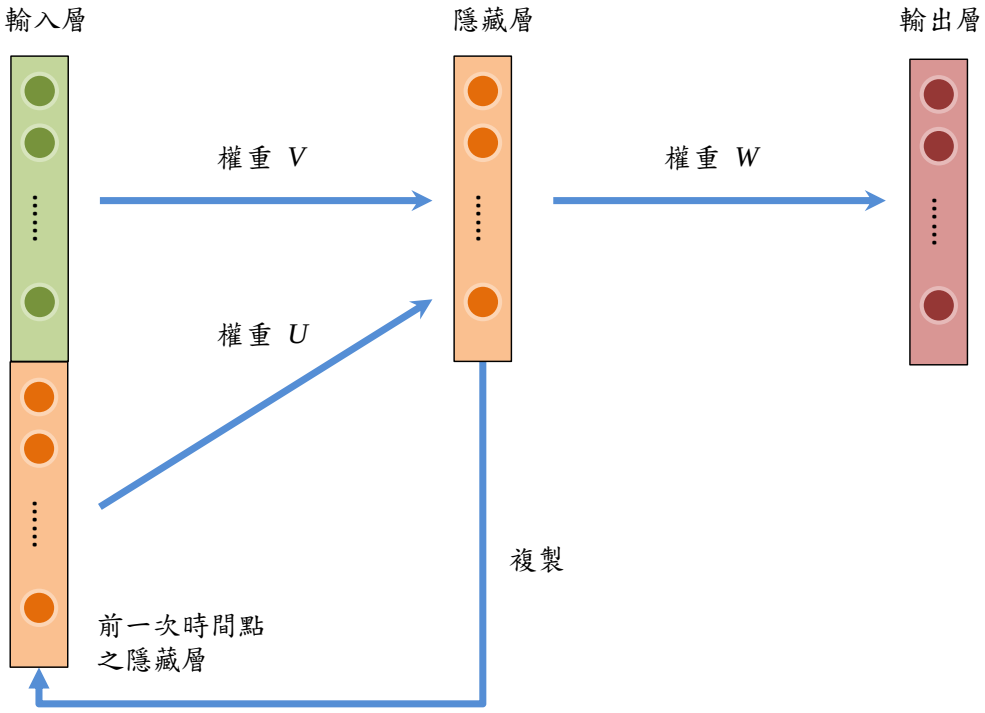


圖 3-4：遞迴式類神經網路架構

3.1.2 遞迴式類神經網路

3.1.2.1 架構

有別於傳統類神經網路，遞迴式的類神經網路更能帶來好的訓練能力，一般常見的是於 1990 年由 Elman 所發展的艾爾曼網路(Elman Networks)[Elman, 1990]。其概念是將隱藏層的輸出當作下一次時間點隱藏層的輸入，而根據不同的需求也有許多不同的網路形成，如喬丹網路(Jordan Networks)[Jordan, 1986]是將輸出層的輸出再傳遞給下一時間點的隱藏層以及雙向遞迴式類神經網路(Bi-directional RNN)[Schuster and Paliwal]利用歷史資訊和未來資訊來做預測，使用的是兩個遞迴式類神經網路來做結合及階層遞迴式類神經網路(Hierarchical RNN)等。本論文則是以艾爾曼網路來進行探討。

遞迴式類神經網路結構可參考圖 3-4，此部分結構是把輸入層加大，且將上一時間點的隱藏層利用暫存複製起來，若以時間方式來階層展開的話，將會更清楚看出其遞迴的概念，如圖 3-5 所示。輸入層、隱藏層和輸出層設計也與之前一樣，但是在前一時間點和目前時間點的隱藏層間會多增加一個權重 U。由於遞迴式類神經網路具有時序處理(Temporal Processing)的能力，一般來評估此類型的網路常會注意它們的穩定性(Stability)、可控性(Controllability)及可觀察性(Observability)。穩定性注重的是隨著時間改變，網路輸出結果需是受侷限的且輸出後的調整量不可過於劇烈，例如網路中輸出的部份或權重。可控性在意的是「是否能夠控制的動態行為」，如果在有限的步驟中，一個初始狀態是可控制至任何期望的狀態，則此遞迴式網路可被稱為具有可控性的。可觀察性關注的是「是否可觀察出控制應用的結果」，如果網路的狀態可以確定從一組有限的輸入或輸出測量，則稱做此網路有可觀察性。而與一般類神經網路不同之處，除了結構上的改變外，演算法部分也有進行調整，接著就介紹時序性倒傳遞演算法

(Backpropagation Through Time)[Werbos, 1990]。

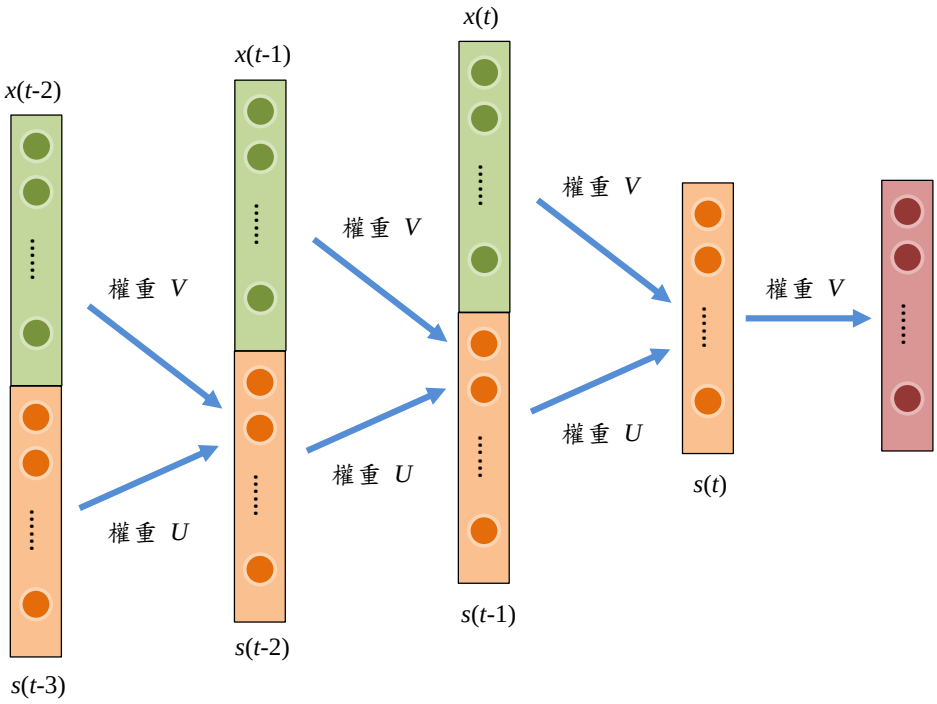


圖 3-5：以時間階層式展開之遞迴式類神經網路架構

3.1.2.2 時序性倒傳遞演算法推導

與倒傳遞演算法不同的地方，遞迴式類神經網路是利用時間的變化來調整權重值，也就是說會調整不只一次且經由不同時間點上的隱藏層資訊來進行調整。如圖 3-5 所示，在時間點 t 所使用的權重是過去時間點所累積的，但利用此方法必須要記錄所有歷史資訊及過去的網路狀態，這將造成記憶體不足和運算量倍增的問題。因此需定義一個變數 τ 當作遞迴的次數[Bengio et al., 1994]，以此來決定想使用多少的歷史資訊，並且忽略掉更早之前的資訊。如前述所提到，假使網路是穩定的話，則權重的更新量將會隨著時間越來越小，這是因為網路倚靠有力的小幅度回饋來增加強度。換句話說，將更早之前的資訊忽略掉並不會造成太大的問題，透過多次的回饋則可彌補此缺點。

其中，誤差函數則增加了時間上的累計，如式(3-12)所示。$E_{total}(T-\tau+1,T)$ 則代表從時間點 $T-\tau+1$ 到 T 的誤差總和，T 為目前的時間點。

$$E_{total}(T-\tau+1,T)=\sum_{t=T-\tau+1}^{T}E(t) \tag{3-12}$$

由於推導過程大致和類神經網路相似，在此就將兩者之間的差別點出。其中隱藏層和輸出層間更新權重 W 的部分和類神經網路相同，式(3-13)則是表示權重 V 與權重 U 在輸入層和隱藏層間的關係。

$$y_j(t)=f(\sum_i v_{ji}x_i(t)+\sum_j u_{jj}y_j(t-1)) \tag{3-13}$$

因此權重 V 與權重 U 的更新量可以由式(3-14)和式(3-15)來求得。

$$\Delta V=-\eta\frac{\partial E_{total}(T-\tau+1,T)}{\partial V}=-\eta\sum_{t=T-\tau+1}^{T}\frac{\partial E(t)}{\partial V} \tag{3-14}$$

$$\Delta U=-\eta\frac{\partial E_{total}(T-\tau+1,T)}{\partial U}=-\eta\sum_{t=T-\tau+1}^{T}\frac{\partial E(t)}{\partial U} \tag{3-15}$$

所以我們可以得到前 τ 次的更新量，並用包含歷史資訊的權重來做預測。

3.1.2.3 遞迴式類神經網路語言模型

遞迴式類神經網路語言模型和類神經網路語言模型主要的差別除了少了投影層、增加了前一時間點的隱藏層外，另一個差別就是輸入層部分。在訓練過程時，輸入層這邊是一次以一個詞來表示並訓練，表示方法則與傳統前饋式類神經網路語言模型相同。它期望透過遞迴的方式獲得長距離之資訊，來做更好的預測，但也有研究學者[Bengio et al., 1993, 1994]指出透過梯度下降法對於學習長距離資訊有一定的困難。我們對權重 V 來舉例，假設目前正計算詞序列中第 T 個詞要回饋給

權重 V 的誤差量，因此可以表示成如式(3-16)所示。t 為詞序列中位置的索引。

$$\frac{\partial E(T)}{\partial V}=\sum_{t\le T}\frac{\partial E(T)}{\partial y_j(t)}\frac{\partial y_j(t)}{\partial V}=\sum_{t\le T}\frac{\partial E(T)}{\partial y_j(T)}\frac{\partial y_j(T)}{\partial y_j(t)}\frac{\partial y_j(t)}{\partial V} \tag{3-16}$$

接著可藉由連鎖率將式(3-16)中，以 $y_j(t)$ 對 $y_j(T)$ 偏微分展開得到式(3-17)。

$$\frac{\partial y(T)}{\partial(t)}=\frac{\partial y(T)}{\partial(T-1)}\frac{\partial y(T-1)}{\partial(T-2)}\cdots\frac{\partial y(t+1)}{\partial(t)}=\prod_{m=t+1}^{T}\frac{\partial y(m)}{\partial y(m-1)} \tag{3-17}$$

因此可以得知，由於式(3-17)等號右邊部份必定會小於 1，且當時間點 t 遠小於目前時間點 T 時，鏈鎖率會不斷的延伸，最終連乘積則會趨近於零。推導結果說明了，遠距離部份的權重更新量只有小幅度的改變，而近距離部份則會有較明顯的影響。所以，遞迴式類神經網路仍然缺乏長距離資訊，但對於中短距離資訊部份，尚可以有效地獲得，也因此許多研究顯示遞迴式類神經網路語言模型的效果好於類神經網路語言模型。

第4章　探索遞迴式類神經網路語言模型之改進

4.1　結合關聯資訊於遞迴式類神經網路語言模型

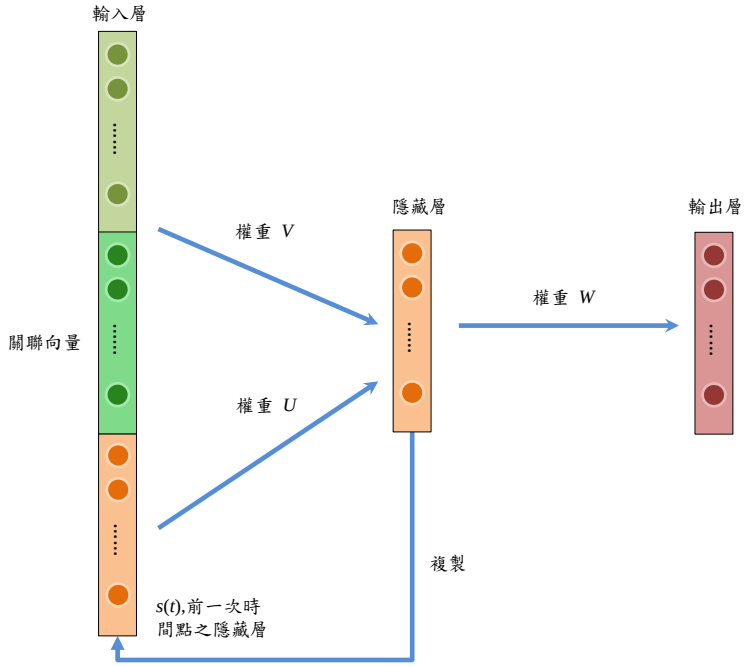


圖 4-1：關聯資訊遞迴式類神經網路架構

傳統統計式 N 連語言模型是容易使用且常見的方法之一，但此模型仍有缺乏長距離資訊與資料稀疏之問題。而使用倒傳遞式類神經網路語言模型能有效解決資料稀疏之問題，可惜在長距離資訊上仍稍嫌不足，因此遞迴式類神經網路語言模型的發展希望能夠取得更多長距離資訊。許多國外研究也顯示遞迴式類神經網路語言模型的確能比倒傳遞式類神經網路語言模型帶來更好的成效，這也是本論文使用遞迴式類神經網路語言模型來探討之原因。遞迴式類神經網路語言模型中回饋的方式是使用時序性倒傳遞演算法，然而，遞迴式類神經網路被證明出此模型無法有效獲得長距離的資訊。因此，本論文透過加入關聯資訊(Relevance Information)來輔助預測下一個詞，其示意圖可以見圖 4-1。而關聯資訊則以向量來表示，大

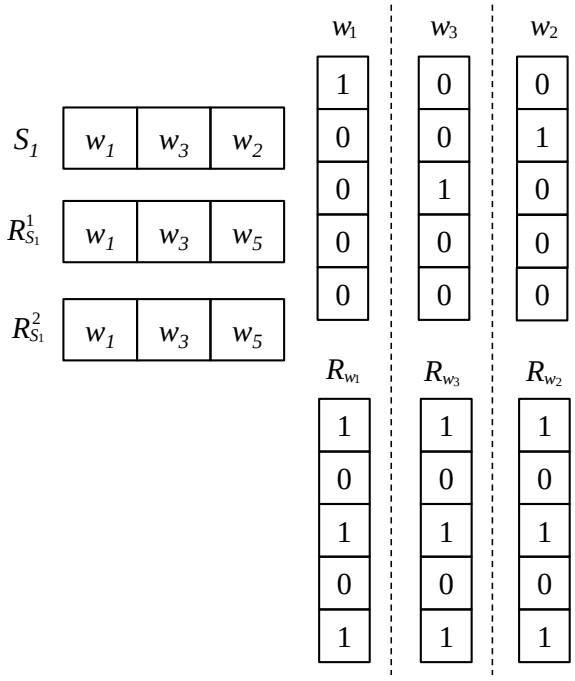


圖 4-2：語句關聯資訊概念圖

小如同原本的輸入層一樣；因此，本論文將輸入層擴增為兩倍，前半段為原本訓練資料的資訊，後半段為對應訓練資料的關聯向量。關聯向量主要又分為兩種，一種為句子間的關聯，稱做語句關聯資訊(Sentence Relevance Information)，另一種為詞跟詞之間的關聯，稱為詞關聯資訊(Word Relevance Information)。語句關聯資訊的產生，是將欲檢索的句子放進訓練語料中進行一次檢索，檢索完可得知對所有訓練語句的關聯分數，根據此關聯分數我們可以決定要使用多少關聯語句來當作關聯向量。圖 4-2 為語句關聯資訊的概念圖，以句子S_1為例，S_1中包含了詞w_1、w_3和w_2。R_1則為對應S_1的語句關聯資訊，但遞迴式類神經網路語言模型是以詞為單位進行訓練，因此賦予每個詞的關聯資訊皆為此句的語句關聯資訊。R_{w_1}、R_{w_3}和R_{w_2}則為詞w_1、w_3和w_2所對應的關聯資訊。詞關聯資訊則是從訓練語料中收集於左右相鄰文段中，相隔一定距離內詞出現的頻率，其結果會得知數個關聯的詞，每個詞的關聯資訊長度也皆不同，其概念如圖 4-3 所示。從圖中可看到，和語句關聯資訊不同之處，為每個詞擁有自己的關聯資訊，若在測試階段時遇到

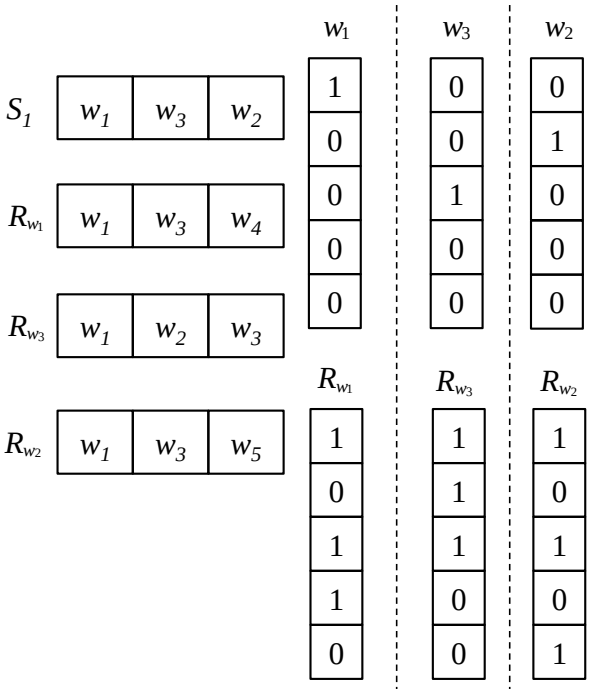


圖 4-3：詞關聯資訊概念圖

未曾看過的詞，則無法進行訓練，所以也不會有關聯資訊的部份。

另外，本論文將關聯資訊以三種不同方式來表示，分別為詞頻、正規化及設定為 1 來進行探討，去看其關聯資訊帶來之影響。以詞頻表示則代表我們使用實際的次數來增加詞與詞之間的關聯性，具有較真實的資訊；以正規化表示使所有關聯資訊的總和為 1，此表示法以較公平的方式來給予值，貢獻大的，也就是次數較多的則值越高，反之，貢獻低的則值較低；而以設定為 1 來表示，則是將有次數出現的維度以 1 來表示，因此此種表示法代表詞與關聯詞之間，其關聯程度均相同。

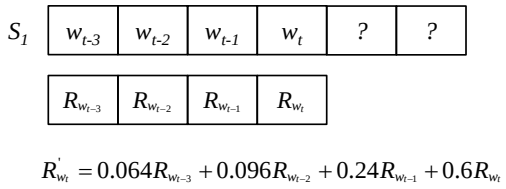
除此之外，也發展了動態詞關聯的方式，由於每個詞所對應的關聯資訊是固定的，因此我們將歷史詞序列中的關聯資訊做結合，得到新的關聯資訊。其中，因為每個詞的歷史詞序列大部分皆不同，所以結合歷史詞序列對應的關聯資訊不易得到相同的關聯資訊。而根據歷史資訊的遠近，分別使用不同權重來做結合，

越遠的歷史資訊則隨著時間越來越小，我們可以用遞迴的方式以式(4-1)來表示。

$$\begin{cases} if \ \ t = 0, & R'_{w_t} = R_{w_t} \\ if \ \ 0 < t \le L, & R'_{w_t} = (1-\alpha)\cdot R_{w_{t-1}} + \alpha \cdot R_{w_t} \end{cases} \tag{4-1}$$

R_{w_t} 為在 t 時間點之原始關聯資訊，R'_{w_t} 為新獲得的關聯資訊，L 為在該語句中的詞數目，α 則為可調控之參數。

一開始時，詞的關聯資訊為原始的關聯資訊；而當時間點大於 0 且小於等於句子長度時，會與所有歷史詞的關聯資訊做線性結合。因此距離越遠的詞，其關聯資訊的權重就越小，相反地，距離越近的詞，其關聯資訊的權重就越大，因而達到動態效果的詞關聯資訊。本論文的 α 值為 0.6。以圖 4-4 為例，句子 S_1 中含有六個詞，假設我們需藉由詞 w_t 的關聯資訊來做下一個詞發生的可能性，而詞 w_t 的關聯資訊可以表示成歷史詞序列所對應的關聯資訊加總，權重部份則取決於詞的距離。



$$R'_{w_t} = 0.064R_{w_{t-3}} + 0.096R_{w_{t-2}} + 0.24R_{w_{t-1}} + 0.6R_{w_t}$$

圖 4-4：動態詞關聯資訊範例

4.2 語句相關之遞迴式類神經網路語言模型

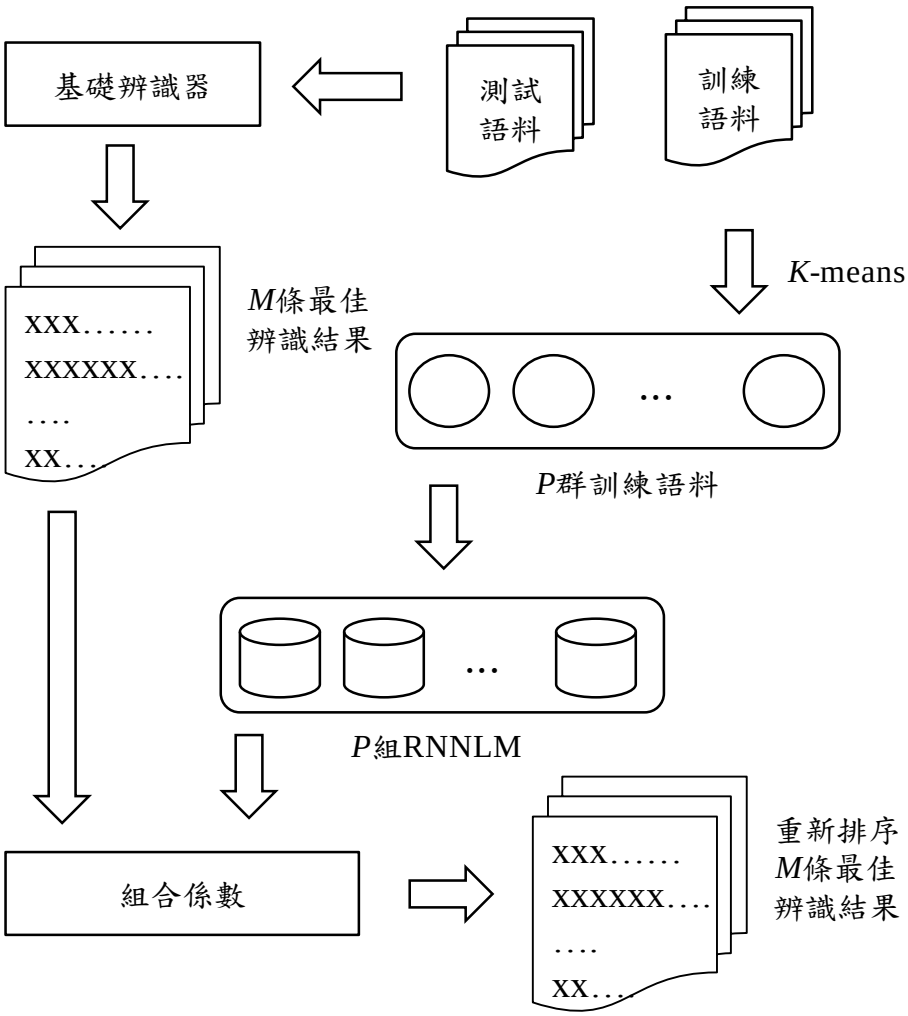


圖 4-5：語句相關之遞迴式類神經網路語言模型流程圖

本節希望藉由動態的語言模型調整方式來輔助估測，因此對於不同測試語句使用不同的遞迴式類神經網路語言模型或是結合不同群的遞迴式類神經網路語言模型。

由於所有測試語句皆使用相同訓練語料訓練出的遞迴式類神經網路語言模

型，因此，本論文希望針對各句測試語句，以線性組合的方式，結合不同訓練語料所訓練出的遞迴式類神經網路語言模型，期望找出較適合各句測試語句的遞迴式類神經網路語言模型。其實作步驟如圖 4-5 所示。一開始，先將所有訓練語句的正確轉寫語句以單連詞向量(Unigram Word Vector)表示，並進行分群(Clustering)。本論文所使用的分群方法為 *K* 平均演算法(*K*-means)[MacQueen, 1967]，在透過分群過後，我們可用各群所分好的訓練語句來訓練各群的遞迴式類神經網路語言模型，假設所有訓練語句可分為 *P* 群。接著，對 *P* 群中每一群訓練語句分別算出各自的特徵權重向量，意即平均向量(Mean Vector)，此部分可由每一群中各句訓練語句的單連詞向量進行加總並取平均求得。可以用式(4-2)來表示：

$$v_p = \frac{\sum_{k=1}^{L_p} v_{p,k}}{L_p} \tag{4-2}$$

其中 $v_{p,k}$ 為第 p 群中第 k 句的單連詞向量， v_p 是第 p 群的平均向量，L_p 為 p 訓練語句群中所含訓練語句之句數。

如此一來，當有一測試語句需要進行估測時，可以利用此測試語句之單連詞特徵向量和所算好的各群特徵權重向量來求取相似度(Similarity)，並選取欲使用的遞迴式類神經網路語言模型或結合不同群的遞迴式類神經網路語言模型。由於我們無法得知測試語句的正確轉寫語句，因此測試語句之單連詞特徵向量皆為各句中 *M* 條最佳辨識結果中的第一名。

本論文主要使用三種選取方式來選取，在計算相似度時，使用的是餘弦值來計算：

$$\cos(U_k, RNNLM_p) = \frac{u_k \cdot v_p}{\sqrt{u_k^{\,2}}\sqrt{v_p^{\,2}}} \tag{4-3}$$

其中 U_k 表示第 k 句測試語句，$RNNLM_p$ 為使用第 p 群訓練語句訓練出的遞迴式類神經網路語言模型，u_k 為測試語句第 k 句中 M 條最佳辨識結果第一名之單連詞向量。以下則介紹三種選取方式：

(1) 選取相似度最大權重法：此方法只選取和測試語句最相似的訓練語句群，也就是所謂相似度最大的。可用下式來表示：

$$RNNLM_{U_k} = \arg\max_{p} \cos(U_k, RNNLM_p) \tag{4-4}$$

因此估測 U_k 的機率便可表示成式(4-5)。

$$P(U_k) \approx P_{RNNLM_{UK}}(U_k) \tag{4-5}$$

其中，$RNNLM_{U_k}$ 代表所挑選出來的遞迴式類神經網路語言模型，因此式(4-4)表示挑選 P 群中餘弦值最大之遞迴式類神經網路語言模型。

(2) 相似度線性組合法：此方法之組合係數是經由計算測試語句與 P 群訓練語句間的相似度來求得，因此如果某測試語句和某群相似度較大則表示該群較符合測試語句之特性，也就是該群會有較大之貢獻。而測試語句與各訓練語句群的組合係數 $\gamma_{k,p}$ 為：

$$\gamma_{k,p} = \frac{\cos(U_k, RNNLM_p)}{\sum_{c=1}^{P} \cos(U_k, RNNLM_p)} \tag{4-6}$$

接著將各群之模型分數用此係數線性組合以獲得新的語言模型分數：

$$P_{RNNLM_k}(w_t) = \sum_{p=1}^{P} \gamma_{k,p} \cdot P_{RNNLM_p}(w_t) \tag{4-7}$$

而 $P_{RNNLM_k}(w_t)$ 為測試語句中估測詞 w_t 經過線性結合之遞迴式類神經網路語言模型分數，$P_{RNNLM_p}(w_t)$ 為估測詞 w_t 第 p 群之遞迴式類神經網路語言模型分

數。

(3) 相似度均勻組合法：

該方法類似相似度線性組合法，將原本組合係數調整成均勻(Uniform)組合係數，因此各群的貢獻度將會相同，則各群均勻組合係數 $\beta_{k,p}$ 為：

$$\beta_{k,p} = \frac{1}{P} \tag{4-8}$$

其中 P 為分群個數。

所以最後結合完的分數可以用下式來表示：

$$P_{RNNLM_k}(w_t) = \sum_{p=1}^{P} \beta_{k,p} \cdot P_{RNNLM_p}(w_t) \tag{4-9}$$

在最後，我們還會將算好的分數結合背景語言模型，以此來得到更好的結果，並將結合完的詞序列分數來做排序。

第5章　實驗架構與結果討論

5.1　實驗架構

5.1.1　臺師大大詞彙連續語音辨識系統

以下將個別介紹臺師大大詞彙連續語音辨識系統採用的特徵擷取、聲學模型、詞典建立、詞彙樹複製搜尋(Tree-copy Search)以及詞圖搜尋等部分。

（一）　特徵擷取

本系統在前端處理中之語音特徵擷取方面，使用了異質性線性鑑別分析(Heteroscedastic Linear Discriminative Analysis, HLDA)[Kumar, 1997]結合最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)[Gales, 1998]兩種不同語音特徵參數。而本論文主要使用異質性線性鑑別分析加上最大相似度線性轉換，獲得 39 維語音特徵向量，接著再使用倒頻譜平均與變異數正規劃(Cepstral Mean and Variance Normalization, CMVN)加強語音特徵。

（二）　聲學模型

在聲學模型部分，由於是處理中文語料，因此我們分別為聲母建立 INITIAL 模型以及為韻母建立 FINAL 模型，基本的 INITIAL 模型為 22 種，FINAL 模型為 38 種。因為聲母會受右邊相連的韻母影響其發音特性，所以再將 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。其中每個模型的中有 3 到 6 個狀態(State)，而每一個狀態為 1 到 128 個高斯分布所組成的高斯混合分布。聲學模型首先經由最大化相似度估測(Maximum Likelihood Estimation, MLE)訓練而得，再透過最小化音素錯誤(Minimum Phone Error, MPE)[Povey, 2004]訓練以期望獲得

最佳化聲學模型參數。

（三） 詞典建立

中文裡大約有 7000 個單字詞，而藉由合併不同的單字詞可以產生新詞。本系統考慮了語料中各個字詞的統計特性，以自動化方式產生新的複合詞(Compound Words)。對於語料中任意相鄰的兩個詞，例如 $w_i w_j$ ，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 $P_b(w_i \mid w_j)$，再由前後向二連機率的幾何平均 $\sqrt{P_f(w_j \mid w_i)P_b(w_i \mid w_j)}$ ，作為詞 w_i 與詞 w_j 是否合併的依據[Chen et al., 2004]。

接著將文字語料從含有一至四字詞約六萬六千個詞的原始詞典進行斷詞，再利用上述的計算方式，經過數次的迭代和不同的門檻值(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

本系統之大詞彙連續語音辨識方法是採取由左至右(Left-to-right)、音框同步(Frame-synchronous)的詞彙樹複製搜尋方法。在詞彙樹中每一個分支(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，路徑上的分支就是代表這個詞或這些詞彙使用到的隱藏式馬可夫模型。進一步來說，我們所用的詞彙樹複製搜尋演算法，在搜尋時每個音框會同時存在數棵詞彙樹複製(Tree Copies)，每個詞彙樹則代表不同的語言模型歷史詞序列(History Word Sequence)。實際上，搜尋時產生的不完

全路徑(Partial Path)如果擁有相同的歷史詞序列會被歸類在同一棵詞彙樹複製裡，以進行隱藏式馬可夫模型狀態層次(State-level)維特比(Viterbi)動態規劃搜尋。

在每個音框裡，假如有不完全路徑已到達葉節點時，表示一個完整詞已可被產生；同時，不同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)，保留較大分數者，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。值得注意的是，我們在實作時並不需要真的建立如此多的詞彙樹複製，僅需建立一棵詞彙樹作為搜尋時路徑展開參考之用即可，並分別記錄搜尋時存活下來的隱藏式馬可夫模型狀態節點的相關資訊。另一部分，因為存下來的隱藏式馬可夫模型的狀態節點會隨著音框呈指數倍成長，因此我們利用光束搜尋(Beam Search)技術，將分數較低的不完全路徑或節點進行剪裁。

此外，根據每個音框中記錄的資訊，例如：語言模型歷史詞序列、候選詞所對應的開始與結束的音框及搜尋時聲學模型解碼的分數，來建立詞圖(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)，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，而在詞圖搜尋階段是使用三連詞語言模型。

（五） 詞圖搜尋與 *M*-最佳結果(*M*-Best)之產生

詞圖為詞彙樹複製搜尋過後所建立的圖，詞圖中的每個分支代表經過裁減所保留的詞段，每個詞段有各自對應的起始音框和結束音框，並會記錄其聲學分數。由於詞圖是已經簡化過的，因此我們在語言模型上可使用較複雜的語言模型，例如三連詞模型、遞迴式類神經網路語言模型或機率式潛藏語意分析模型等。接著將每個詞段進行維特比搜尋，根據音框資訊、聲學分數、歷史詞序列以及從語言模型中計算出的分數查找出多條詞序列。最後挑選分數最高的詞序列當作辨識結果；

亦可以輸出分數前 *M* 高的詞序列進一步做處理，像是藉由鑑別式訓練來找出字錯誤率最低的詞序列，或利用訓練好的語言模型進行重新排序以得到更準確之辨識結果。

短句語料	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
長句語料	句數	長度(小時)
訓練集語料	3643	約 20
發展集語料	292	約 1.5
測試集語料	307	約 1.5

表 5-1：實驗語料統計資訊

5.1.2 實驗語料

本論文使用之實驗語料是來自於公視新聞(Mandarin Across Taiwan-Broadcast News, MATBN)[Wang et al., 2005]。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(PTS)合作錄製，主要內容為國內新聞，其中也包含一小部分國際新聞，共計 197 個小時之語音資訊與其內容標記。在標註過程中，包含雜訊、背景環境、發音不標準、方言、說話者性別、主播、記者、受訪者等資訊。語料內主要可分為內場新聞與外場新聞兩個部分。其中內場新聞為主播語料，外場新聞語料包含有採訪記者(Field Reporters)語料與受訪者(Interviewees)語料。

由於內場新聞的主播語料大部分由同一主播所錄製，為了避免語者相依(Speaker Dependent)之問題導致實驗偏差，因此不使用此部分語料。而外場新聞中，受訪者語料內含過多語助詞和背景音樂，較不適合使用。所以本論文實驗語料使用的是外場採訪記者語料，其使用新聞語料之原因是主播及記者皆有受過專

業的口語訓練，因而在判斷實驗的準確性上不會因為表達不順暢或語助詞太多造成實驗上的誤差。本論文為了實驗上需求，分別使用長句語料與短句語料去探討遞迴式類神經網路模型帶來之影響，因此實驗語料分成兩部分，其一為短句的部分，選自公視新聞 2001 年至 2002 年外場採訪記者，分別為訓練集語料 30,600 句(約 23 小時)、測試集語料 1,997 句(約 1.5 小時)及發展集語料 1,998 句(約 1.5 小時)。其二為長句部分，選自公視新聞 2003 年外場採訪記者，分別為訓練集語料 3,643 句(約 20 小時)、測試集語料 307 句(約 1.5 小時)及發展集語料 292 句(約 1.5 小時) ，如表 5-1 所示。但整體實驗主軸仍以短句語料為主，長句語料只用來探討在遞迴式類神經網路語言模型中長短句的學習差異。

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

另外背景語言模型使用的訓練語料是來自 2001 至 2002 年中央通訊社(Central News Agency, CNA)的文字新聞語料，內含有約一億五千萬個中文字，經由斷詞之後約有八千萬個詞。此語言模型是使用 SRI Language Modeling Toolkit (SRILM)訓練而得，採用 Katz Back-off 平滑化方法[Katz, 1987]來解決資料稀疏的問題。

5.1.3 語言模型評估

在語音辨識中，評估語言模型之辨識結果主要是以計算錯誤率(Error Rate)來做評判。根據美國標準與科技組織所訂立的評估標準(U.S. NIST F.O.M. Metric)，將正確參照轉寫和辨識結果進行字串比對，藉由動態規劃(Dynamic Programming)方式求得最佳字串對齊(Alignment)。依據單位(Unit)不同，可分為字錯誤率(Character Error Rate, CER)與詞錯誤率(Word Error Rate, WER)。在英文中通常使用詞錯誤率當作評估標準，而在中文中以詞錯誤率當作評估標準會產生新詞定義和斷詞問題，因此使用字錯誤率來做為評估標準。

比對過程中，兩字串單元可能會有替代(Substitution)、插入(Insertion)與刪除(Deletion)等錯誤狀況。因此對齊後可以下列公式進行計算辨識字正確率(Accuracy)或字錯誤率(Error rate)：

$$\text{字正確率} = \frac{H - I}{N} \times 100\% \tag{5-1}$$

$$\text{字錯誤率} = \frac{S + I + D}{N} \times 100\% \tag{5-2}$$

其中，H 為兩字串相同(Hit)的單元數量，I 為便是字串插入的單元數量，N 為正確參照轉寫的單元數量。錯誤率則為 1-辨識正確率。

另外一種常見的評估語言模型方式為語言複雜度(Perplexity)。其定義如下，若給定一個語言模型 Γ 以及一段測試詞序列 W，若詞序列的長度非常長，那麼語言模型 Γ 作用於詞序列 W 上的交互熵值(Cross-Entropy)，則可近似為：

$$H_{\Gamma}(W) = -\frac{1}{|W|}\log_2 P_{\Gamma}(W) \tag{5-3}$$

詞序列 W 的聯合機率 $P_{\Gamma}(W)$ 可以將它拆解成一連串條件機率的連乘積。我們將依語言模型的語言複雜度定義為：

$$\text{PPL}(W) = 2^{H_{\Gamma}(W)} \tag{5-4}$$

其代表意義為語言模型 Γ 給予詞序列 W 中每一個詞的(幾何)平均機率值，意可以將複雜度視為語言模型對於詞的預測的平均分支度(Average Branching Factor)。因此，一個語言模型越有效的預測詞序列中每一個出現詞的出現，則語言複雜度分數會越低。

5.2 基礎實驗結果

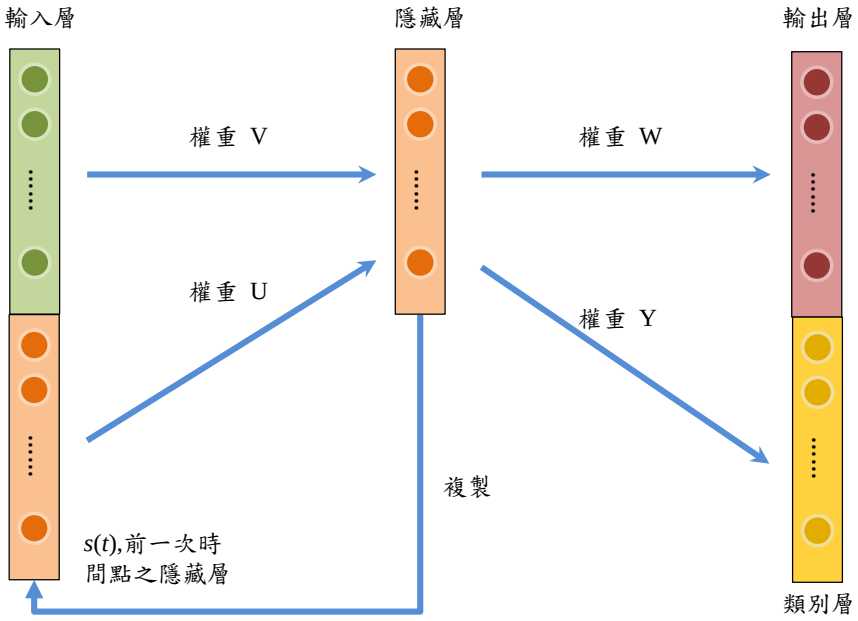


圖 5-1：遞迴式類神經網路語言模型套件架構

實驗的產生是從詞圖產生 100(M=100)句最佳候選詞序列，再經由訓練好的遞迴式類神經網路語言模型得到各句詞序列的分數，並加入聲學模型的資訊以此獲得語音辨識的總分數。透過所選出的第一名詞序列和正確答案去算出 Edit distant 得到字正確率。遞迴式類神經網路語言模型則是使用 Mikolov 等學者[Mikolov et al, 2011]所發展的 Recurrent Neural Network Language Modeling Toolkit 訓練而得，可參考圖 5-1 之架構。本論文所提出的應用關聯資訊則是對其套件做修改，將關聯資訊加入於遞迴式類神經網路語言模型，語句相關之遞迴式類神經網路語言模型則是利用此套件來訓練遞迴式類神經網路語言模型。而實驗的目的，則是希望透過遞迴式類神經網路語言模型來重新排序，以找出字正確率最高的詞序列。另外，實驗中參數的部分，是利用發展集調出的參數來做為測試集的參數。

語言模型設定方面，在訓練遞迴式類神經網路語言模型時隱藏層個數為 100、類別層個數也為 100 、遞迴的次數為 4 且訓練及辨識過程中，句子和句子之間是

獨立的，也就是說上一句的句子和目前訓練的句子是不相關聯的。

	發展集語言複雜度	發展集語言複雜度	發展集語料字正確率(%)	測試集語料字正確率(%)	絕對提昇率(%)	相對提昇率(%)
基礎辨識率(BG)	450.93	459.06	84.73	83.61	-	-
RNN	607.07	623.50	82.31	82.41	-1.2	-7.32
RNN+BG	232.31	236.97	85.67	85.17	1.56	9.52
Oracle	-	-	93.22	92.66	-	-

表 5-2：遞迴式類神經網路語言模型之基礎實驗結果

表 5-2 是關於遞迴式類神經網路語言模型的基礎實驗結果，從語言複雜度的角度來看，遞迴式類神經網路語言模型(RNN)的語言複雜度為 623.5，再看到背景語言模型(BG)的部份，由於訓練語料較多，因此其效果會比遞迴式類神經網路語言模型來的好。而根據文獻中所看到的，遞迴式類神經網路語言模型在獨自使用時效果較不明顯，必須和其他模型做結合。實驗中也可看到背景語言模型結合遞迴式類神經網路語言模型(RNN+BG)效果會來得最好，可以見得，遞迴式類神經網路語言模型具有不錯的成效。另一部分是關於辨識率的實驗結果，基礎辨識率為經由重新計分後的結果，也就是 100 句最佳詞序列中的第一名。但是我們可以看到 Oracle 部分(意即 100 句最佳詞序列中字錯誤率最低的部分)，字正確率可到達 92.66%，正意味著我們仍有很大的進步空間。而這部份的趨勢也和語言複雜度相同，單獨使用遞迴式類神經網路語言模型時，辨識率下降 1.2%，但透過與背景語言模型的結合，其絕對提昇率有 1.56%以及相對提昇率 9.52%。

在本章第 4 節和第 5 節的實驗中，我們將使用背景語言模型結合遞迴式類神經網路語言模型(RNN+BG)的語言複雜度 236.97 和辨識率 85.17%作為 RNN 的基礎辨識率，以此來和我們所提出的方法做比較。

5.3 使用長句語料與短句語料於遞迴式類神經網路語言模型之實驗結果

從表 5-3 中可看出使用短句語料時，必須倚靠背景語言模型才能有好的辨識率，而長句語料則使用較少的背景語言模型分數。由此可以知道，使用長句語料時，遞迴式類神經網路語言模型能提供較多的幫助，這也符合了遞迴式類神經網路可獲取較長距離資訊的精神。反之，在短句語料中，因為內容切的較細而句子較短，所以能提供的資訊較少，導致必須倚靠背景語言模型來提供更多的資訊。

	測試語料字正確率(%)	
RNN 權重參數	短句	長句
0	83.85	80.00
0.1	85.19	80.65
0.2	85.17	80.75
0.3	**85.19**	80.78
0.4	85.18	80.86
0.5	85.09	80.90
0.6	84.92	80.91
0.7	84.77	**81.03**
0.8	84.58	80.95
0.9	84.14	80.91
1	82.41	80.73

表 5-3：使用短句語料與長句語料於 RNN 之差異

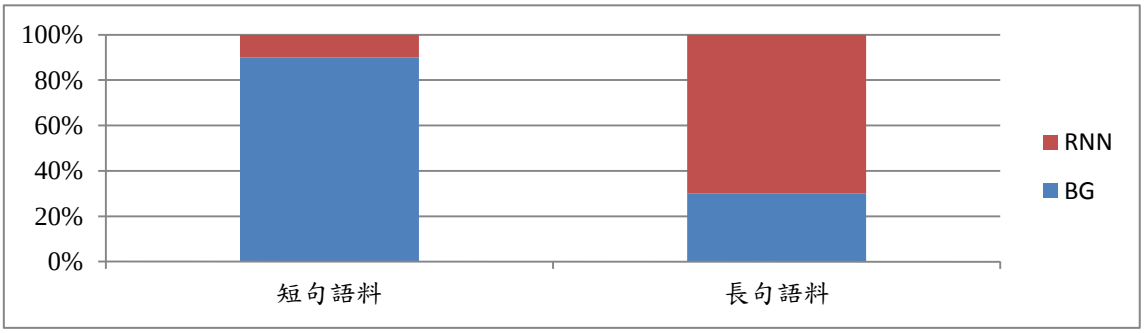


圖 5-2：短句語料與長句語料使用兩種語言模型比例

5.4 結合關聯資訊於遞迴式類神經網路語言模型之實驗結果

首先,我們使用語句關聯資訊來幫助遞迴式類神經網路語言模型做估測,**錯誤! 找不到參照來源。**是結合語句關聯資訊的辨識結果。在訓練模型時，語句關聯資訊是挑選最相關的訓練語句，由於挑選過多的關聯資訊會導致辨識率下降，因此只挑選最相關的部分。其中，因為遞迴式類神經網路語言模型是以詞為單位進行訓練，所以每個詞需要對應到一個關聯向量，此部分的詞關聯向量為此句的關聯向量。而關聯向量的表示方式分別使用了句子中詞出現的次數、將詞出現的次數做正規化及出現該詞的維度設為 1。從實驗中可看到語言複雜度部份，以句中詞出現的次數較好，而辨識率方面，則是使用正規化的表示較好，相較於基礎辨識率 85.17%小幅度的進步了 0.04%，而使用句中次數則下降了 0.08%，以及設定為 1 的方法也下降了 0.14%。

圖 5-3 是在語句關聯資訊中，使用三種表示法的辨識率結果，可看出使用正規化的表示法較好，其餘兩種的辨識率則較 RNN 基礎辨識率來得低。探究其辨識率進步不大的原因應為每一語句內，詞的關聯向量皆為此句的關聯向量，因此

關聯資訊表示方式	發展集語言複雜度	測試集語言複雜度	發展集語料字正確率(%)	測試集語料字正確率(%)	絕對提昇率(%)	相對提昇率(%)
RNN 基礎辨識率	232.31	236.97	85.67	85.17	-	-
句中詞出現的次數	223.63	229.01	85.63	85.09	-0.08	-0.56
正規化	230.51	236.45	85.71	85.21	0.04	0.27
出現該詞則設為 1	226.04	231.19	85.56	85.03	-0.14	-0.95

表 5-4：結合語句關聯資訊之實驗結果

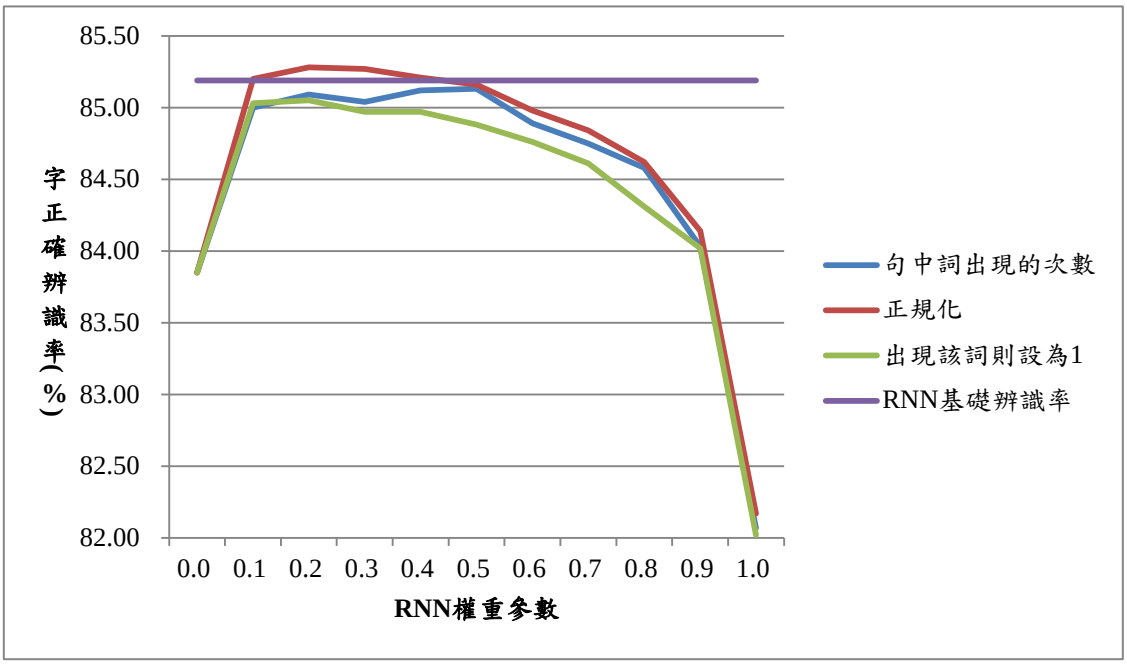


圖 5-3：語句關聯資訊-三種表示法之辨識率比較

關聯資訊表示方式	發展集語言複雜度	測試集語言複雜度	發展集語料字正確率(%)	測試集語料字正確率(%)	絕對提昇率(%)	相對提昇率(%)
RNN 基礎辨識率	232.31	236.97	85.67	85.17	-	-
詞出現的次數	230.05	234.93	85.88	85.36	0.19	1.26
正規化	230.13	234.75	85.83	85.34	0.17	1.16
出現該詞則設為 1	230.12	234.83	85.77	85.40	0.23	1.52

表 5-5：結合詞關聯資訊之實驗結果

關聯向量的重複率就等同於語句中詞的數量。這也成為了辨識率降低的原因之一，另一個原因則是原本輸入的資訊可能被關聯資訊所干擾。因此我們嘗試將關聯資訊切得更細，使用詞關聯資訊來幫助估測。

詞關聯資訊不同於語句關聯資訊之處，就是每一語句中的詞有各自的關聯資訊。詞關聯資訊的產生，是將訓練語料中，詞出現的地方找其相鄰詞作為關聯資

訊，出現相同的相鄰詞則會累加出現次數，本論文則是取左右距離為 3。

表 5-5 使用詞關聯資訊於遞迴式類神經網路語言模型的實驗結果，由於部分詞的詞關聯資訊相當多，包含了關聯詞和非關聯的詞，因此，我們試著去調整詞關聯資訊的使用程度，其中詞關聯資訊的長度是根據發展集中最好的結果來設定。語言複雜度方面，詞出現次數、正規化及出現該詞則設為 1 也都有進步；而字正確率方面，詞出現次數、正規化及出現該詞則設為 1 均有提昇，絕對提昇率分別為 0.19%、0.17%及 0.23%，相對提昇率則有 1.23%、1.16%及 1.52%的進步。圖 5-4 則是詞關聯資訊中，使用三種表示法的辨識率比較，可看到出現該詞設為 1 的辨識結果較好，因為此方法對於每個關聯詞的關聯度較公平，大家皆設定為 1。而詞出現的次數和正規化法，因為每個關聯詞之間的歧異度較高，尤其是詞出現的次數，次數的差距很大，導致有些關聯詞的貢獻被埋沒。

另外此部份實驗結果顯示，使用出現該詞設為 1 且調整詞關聯資訊的長度較好，於是我們進一步去觀察使用此表示法在不同詞關聯資訊長度上的比較，圖 5-5 則是其辨識結果，可看出過多的資訊會導致效果減弱。

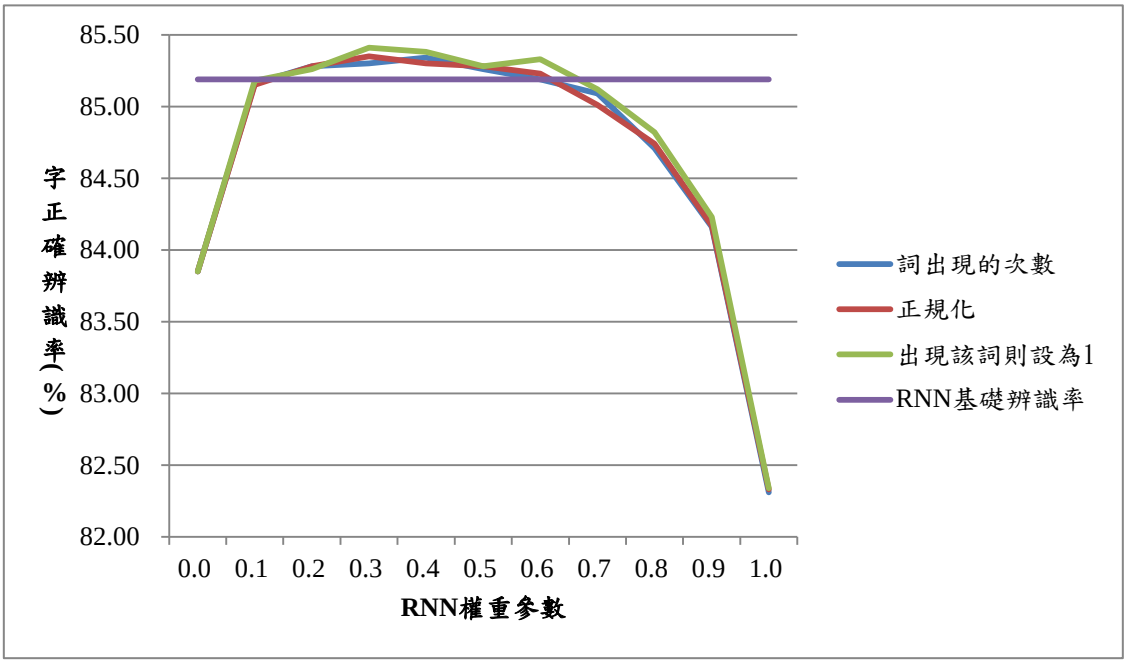


圖 5-4：詞關聯資訊-三種表示法之辨識率比較

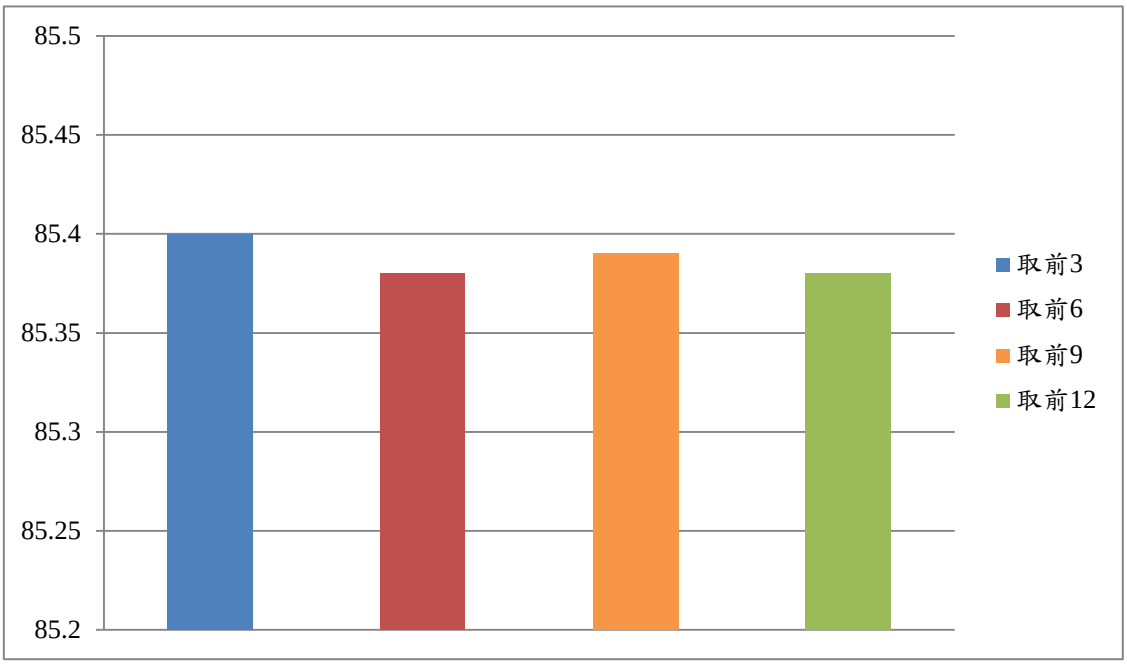


圖 5-5：使用不同長度之詞關聯資訊辨識結果

雖然從實驗結果中得知，使用詞關聯資訊的確提升了辨識率，但我們仍希望可以突破目前的瓶頸。於是我們發現到，雖然詞關聯資訊解決了語句關聯資訊的問題，但是仍有類似語句關聯資訊的缺點存在，其缺點則是語料中的每個詞所對應到的詞關聯資訊仍然一樣，造成在訓練中重複使用同樣的詞關聯資訊。此作法的詞關聯資訊是比較屬於全域的，也就是是針對所有訓練語料中，獲得詞與詞的關聯度。而在訓練遞迴式類神經網路語言模型中，我們也需要區域性的資訊，因為相同的詞在不同的句子可能代表著不同的意思，所以我們希望藉由區域性的資訊來得知上下文或句子中的資訊，使得預測下一個詞時能符合句子中的意思。舉例來說，一篇關於林書豪的報導，我們可得知關聯的詞，像是 NBA、林書豪或是林來瘋。但是就一句話來看，在句子中跟 NBA 相關的可能是其他球隊或是球員。因此，我們希望詞的關聯資訊必須是會變動的，如此一來才能包含全域性的資訊和區域性的資訊。為此，本論文提出了動態詞關聯資訊來做更進一步的改進，此部分是先將所有歷史詞的關聯資訊做結合，其結合依據遠近來給予權重。因此，詞的距離離目前的詞越遠，則該詞的關聯資訊貢獻越小；反之，詞的距離離目前的詞越

關聯資訊表示方式	發展集語言複雜度	測試集語言複雜度	發展集語料字正確率(%)	測試集語料字正確率(%)	絕對提昇率(%)	相對提昇率(%)
RNN 基礎辨識率	232.31	236.97	85.67	85.17	-	-
詞出現的次數	229.72	234.35	85.84	85.26	0.09	0.58
正規化	231.42	236.19	85.83	85.34	0.17	1.14
出現該詞則設為 1	229.98	234.62	85.86	85.34	0.17	1.16

表 5-6：結合動態詞關聯資訊之實驗結果

近，則該詞的關聯資訊越大。表 5-6 則是使用動態詞關聯資訊的實驗結果，結果顯示使用動態詞關聯資訊效果與一般的詞關聯資訊較差一點，辨識率大約為85.34%左右；語言複雜度則較詞關聯資訊好。探究其原因，應為關聯資訊中常包含關聯與非關聯的資訊，因此我們難以準確知道越近距離的詞關聯資訊有較相關的資訊，造成使用的關聯資訊無法正確代表與該詞相關。

根據本論文所提出的結合關聯資訊於遞迴式類神經網路語言模型的確有助於辨識率的提升,但是其效果仍有限且不是那麼的明顯,其原因大概歸類為三種，其一是關聯資訊可能會對輸入層所要傳遞的資訊造成干擾，使得輸入層所要傳遞的資訊減弱，而關聯資訊被成為主要傳遞的資訊；其二是關聯資訊結合輸入層，也可能只是將其表示方式做了延伸，而關聯資訊的表示法可能有更佳的表示方法；其三則是難以準確的決定關聯資訊，導致效果無法彰顯。

5.5 語句相關之遞迴式類神經網路語言模型之實驗結果

本節主要探究利用語句相關之遞迴式類神經網路語言模型之實驗結果，以下表 5-7 至表 5-9 為將訓練語料分兩群之結果，表 5-10 至表 5-12 為四群之結果。

RNN 權重參數	發展集語料字正確率(%)	測試集語料字正確率(%)	絕對提昇率(%)	相對提昇率(%)
0	84.29	83.85	-1.32	-8.89
0.1	85.64	85.20	0.03	0.19
0.2	85.87	85.31	0.14	0.97
0.3	85.86	85.40	0.23	1.52
0.4	85.90	85.41	0.24	1.60
0.5	85.81	85.28	0.11	0.77
0.6	85.70	85.08	-0.09	-0.61
0.7	85.46	84.94	-0.23	-1.55
0.8	85.05	84.57	-0.60	-4.02
0.9	84.60	84.18	-0.99	-6.69
1	82.56	82.38	-2.79	-18.82

表 5-7：選取相似度最大權重法(兩群)之辨識結果

RNN 權重參數	發展集語料字正確率(%)	測試集語料字正確率(%)	絕對提昇率(%)	相對提昇率(%)
0	84.29	83.85	-1.32	-8.89
0.1	85.68	85.10	-0.07	-0.46
0.2	85.88	85.28	0.11	0.75
0.3	85.92	85.24	0.07	0.46
0.4	85.91	85.23	0.06	0.39
0.5	85.78	85.22	0.05	0.36
0.6	85.57	85.15	-0.02	-0.15
0.7	85.43	84.97	-0.20	-1.38
0.8	84.97	84.58	-0.59	-3.97
0.9	84.55	84.12	-1.05	-7.10
1	82.62	82.29	-2.88	-19.43

表 5-8：相似度線性組合法(兩群)之辨識結果

RNN 權重參數	發展集語料字正確率(%)	測試集語料字正確率(%)	絕對提昇率(%)	相對提昇率(%)
0	84.29	83.85	-1.32	-8.89
0.1	85.58	85.08	-0.09	-0.61
0.2	85.94	85.29	0.12	0.82
0.3	85.94	85.28	0.11	0.75
0.4	85.81	85.18	0.01	0.07
0.5	85.74	85.16	-0.01	-0.05
0.6	85.44	85.17	0.00	-0.03
0.7	85.29	84.93	-0.24	-1.60
0.8	84.99	84.69	-0.48	-3.22
0.9	84.52	84.22	-0.95	-6.37
1	82.59	82.27	-2.90	-19.55

表 5-9：相似度均勻組合法(兩群)之辨識結果

其中，表 5-7 為分成兩群後選取相似度最大權重法結合的字正確率，表 5-8 為分成兩群使用相似度線性組合法的字正確率，表 5-9 則為使用相似度均勻組合法的字正確率。此部分我們會額外使用一個全部訓練語料所訓練出的遞迴式類神經網路語言模型來做輔助，稱為遞迴式類神經網路背景語言模型。首先利用三種權重組合方式將各群做結合，接著再將結合完的結果加入遞迴式類神經網路背景語言模型的分數作為輔助，最後才結合背景語言模型。比較表 5-7 至表 5-9 可以看到利用選取相似度最大權重法會有較好的結果，與基礎遞迴式類神經網路語言模型相比則有 0.24%的進步，而選取相似度均勻組合法與相似度線性組合法則略差於基礎辨識率。探究其原因應為資料較偏向某一群，因此使用相似度均勻組合法與相似度線性組合法則較差，只要使用相似度最大的那群就能有較好的效果。

表 5-10 則是分成四群後選取相似度最大權重法結合的字正確率，表 5-11 為分成四群使用相似度線性組合法的字正確率，表 5-12 為使用相似度均勻組合法的字正確率。從將訓練語料分四群來訓練的實驗中，也可得知選取相似度最大權重

RNN 權重參數	發展集語料字正確率(%)	測試集語料字正確率(%)	絕對提昇率(%)	相對提昇率(%)
0	84.29	83.85	-1.32	-8.89
0.1	85.65	85.11	-0.06	-0.44
0.2	85.73	85.29	0.12	0.82
0.3	85.86	85.31	0.14	0.92
0.4	85.87	85.33	0.16	1.06
0.5	85.79	85.29	0.12	0.80
0.6	85.61	85.11	-0.06	-0.41
0.7	85.39	84.95	-0.22	-1.46
0.8	85.15	84.77	-0.40	-2.67
0.9	84.71	84.22	-0.95	-6.42
1	82.75	82.48	-2.69	-18.12

表 5-10：選取相似度最大權重法(四群)之辨識結果

RNN 權重參數	發展集語料字正確率(%)	測試集語料字正確率(%)	絕對提昇率(%)	相對提昇率(%)
0	84.29	83.85	-1.32	-8.89
0.1	85.55	85.12	-0.05	-0.32
0.2	85.73	85.30	0.13	0.89
0.3	85.86	85.31	0.14	0.94
0.4	85.81	85.29	0.12	0.80
0.5	85.77	85.23	0.06	0.43
0.6	85.50	85.15	-0.02	-0.15
0.7	85.35	84.95	-0.22	-1.50
0.8	85.13	84.76	-0.41	-2.79
0.9	84.66	84.27	-0.90	-6.06
1	83.39	82.94	-2.23	-15.02

表 5-11：相似度線性組合法(四群)之辨識結果

RNN 權重參數	發展集語料字正確率(%)	測試集語料字正確率(%)	絕對提昇率(%)	相對提昇率(%)
0	84.29	83.85	-1.32	-8.89
0.1	85.59	85.12	-0.05	-0.36
0.2	85.71	85.30	0.13	0.89
0.3	85.78	85.32	0.15	1.02
0.4	85.69	85.26	0.09	0.60
0.5	85.60	85.23	0.06	0.39
0.6	85.52	85.04	-0.13	-0.85
0.7	85.35	84.94	-0.23	-1.53
0.8	85.05	84.67	-0.50	-3.34
0.9	84.63	84.22	-0.95	-6.37
1	83.31	82.96	-2.21	-14.90

表 5-12：相似度均勻組合法(四群)之辨識結果

法是較佳的。但是跟兩群的實驗相比，雖然與基礎辨識率相比仍有進步，三種選取方法還是較差一點。

探討其原因應為訓練語料不足的關係，由於分群數目提高，則每群中的訓練語料則隨之減少。因此無法訓練出學習能力較佳的遞迴式類神經網路語言模型，導致字正確率的下降。而實驗中也可看出，結合完各群結果後的辨識率仍不好，需要加入遞迴式類神經網路背景語言模型來輔助，以得到更好的辨識率。

5.6 各式語言模型比較與探討

	測試語料字正確率(%)	絕對提昇率(%)	相對提昇率(%)
Baseline	83.61	-	-
2-gram	85.29	1.68	10.25
3-gram	85.23	1.62	9.88
NN	84.75	1.14	6.96
RNN	85.17	1.56	9.52
SR-RNN	85.21	1.60	9.76
WR-RNN	85.40	1.79	10.90
DWR-RNN	85.34	1.73	10.57
Cluster-RNN	85.41	1.80	10.96
PLSA	83.85	0.24	1.46
Perceptron	85.05	1.44	8.79
GCLM	84.11	0.50	3.05
WGCLM	84.63	1.02	6.22
MERT	84.70	1.09	6.65

表 5-13：各種語言模型之實驗結果

表 5-13 為不同語言模型之比較結果，可以分做六個部分來看，第一個部份是基礎字正確辨識率(Baseline)；第二部份是 *N* 連語言模型，分成 2 連語言模型(2-gram)和 3 連語言模型(3-gram)來看，此兩種模型的絕對提昇率，分別是 1.68%和 1.62%的進步，相對提昇率則是 10.25%和 9.88%的進步。可以看到 2 連語言模型較 3 連語言模型來的好，因為 3 連語言模型的參數量大幅增多，雖然可以做出更正確的估測，但是估測錯誤的機會相對也會來的大，導致沒有比 2 連語言模型好。

第三部份為類神經網路語言模型，類神經網路語言模型(NN)不同於鑑別式語言模型，使用的是非線性的方法來做估測，而且解決了資料稀疏的問題，因而絕對提昇率有 1.14%的進步，相對提昇率有 6.96%的進步。遞迴式類神經網路語言模型(RNN)則是繼承了類神經網路語言模型的優點，且能獲得較長距離的資訊，

所以辨識率方面又比類神經網路語言模型進步 0.42%，在絕對提昇率和相對提昇率則為 1.56%和 9.52%。雖然此部份結合背景語言模型後，較 2 連語言模型和 3 連語言模型來的差，但在單獨使用時，也就是不包含背景語言模型的時候，遞迴式類神經網路語言模型會有較好的辨識結果。

第四部份是本論文所提出的方法，表 5-13 中皆為選最好的辨識率來做比較，使用語句關聯資訊的遞迴式類神經網路語言模型(SR-RNN)在效果上的確帶來些許的幫助，其絕對提昇率有 1.6%的進步，而相對提昇率有 9.76%的進步。此部份進一步探討使用更詳細的關聯資訊，可以見得使用詞關聯資訊於遞迴式類神經網路語言模型(WR-RNN)改善了語句關聯資訊的部份，效果上則有 1.79%的絕對提昇率及 10.79%的相對提昇率。但我們發現，使用詞關聯資訊仍不臻完美，其同一個詞所對應的詞關聯資訊皆相同，為了改善這個問題我們進一步將詞關聯資訊做動態的改變(DWR-RNN)。儘管辨識率方面沒有更大的改進，但比起基礎辨識率仍有不錯的效果，絕對提昇率有 1.73%，相對提昇率有 10.57%。接著是將訓練語句分群，再對各群訓練其遞迴式類神經網路語言模型(Cluster-RNN)，由於分群後，測試語句能針對其特性來得到更好的估測，因此辨識率上有不錯的成效，絕對提昇率和相對提昇率分別有 1.8%和 10.96%的進步。

最後兩部分則是主題式模型和四種常見的鑑別式語言模型，機率式潛藏語意分析(PLSA)的辨識率為 85.85%，絕對提升率和相對提升率分別為 0.24%和 1.46%；感知器演算法(Perceptron)雖然一般化能力較差，但在訓練語料和測試語料有高度相關的時候，會有較不錯的效果，因此為四種方法裡面進步最多，絕對提昇率有 1.44%，相對提昇率有 8.79%。全域條件式對數線性模型(GCLM)則在四種模型中表現較差，絕對提昇率只有 0.5%，相對提昇率為 3.05%。權重式全域條件式對數線性模型(WGCLM)則是因為多考慮了樣本權重，因此在辨識率上較全域條件式

對數線性模型來的好，絕對提昇率和相對提昇率則有 1.02%和 6.22%的進步。而最小化錯誤率訓練(MERT)不但考慮了樣本權重，其一般化能力也較佳，所以在此四種鑑別式語言模型中也有不錯的效果，絕對提昇率和相對提昇率分別有 1.09%和 6.65%的進步。因此，我們可以從最後兩部分看出，本論文所提出的方法與遞迴式類神經網路語言模型結合有較佳的效果。

第6章　結論與未來展望

在語音辨識、資訊檢索與自然語言處理領域中，語言模型具有一定的影響力。其中，在語音辨識裡，語言模型更是缺一不可的角色，他提供了語句在自然語言處理中發生的可能性。傳統 N 連語言模型是目前語言模型當中常見的方法之一，但是卻難以捕捉到長距離的語句資訊，加上擁有資料稀疏和維度的詛咒之特性，長期以來一直難以突破。近年來不斷有新型的語言模型被提出，如鑑別式語言模型與類神經網路語言模型，而其中根據國外學者的研究，發現類神經網路語言模型有不錯的成效，不僅能擁有 N 連語言模型的特性也能彌補維度的詛咒之缺點，為語音辨識與語言模型帶來新的視野；然而類神經網路語言模型也仍存在一些缺點，例如缺乏長距離資訊、運算的時間複雜度過高以及詞的表示方式缺少了詞的特性等問題。因此，也有學者針對類神經網路的變形，使用了具有遞迴能力的類神經網路來建構語言模型，而效果也比一般類神經網路語言模型好。

遞迴式類神經網路語言模型所希望的就是具備一般類神經網路語言模型的優點，且能夠提供長距離的資訊。但是有研究[Bengio et al., 1994]提到使用梯度下降法時，由於鏈鎖率(Chain Rule)的關係，當時間越長時機率會越乘越小，導致趨近於 0。使得長距離資訊無法有效取得，因此本論文針對遞迴式類神經網路語言模型做了更進一步的改善，期望使用關聯資訊和動態調整語言模型來輔助機率的估測。從實驗結果中可以看出使用關聯資訊的確能帶來幫助，但是效果仍不夠明顯，其原因應為輸入層或前一時間點的資訊被關聯資訊所干擾，導致成效有限。而實驗中也發現到減少部分關聯資訊能提升辨識率，因此關聯資訊或其他資訊的表示法在未來研究上也是值得注意的部分。另一部分，本論文藉由將訓練語料分群並訓練各群的遞迴式類神經網路語言模型，期望藉由動態的調整語言模型來達到更好的辨識率。此部分實驗結果也顯示分兩群時，使用相似度線性組合法有較

佳的成效。但分成四群時，由於各群中的訓練語料不足，因此無法訓練出學習能力較佳的遞迴式類神經網路語言模型。

在未來的研究裡，可以根據遞迴式類神經網路語言模型無法有效學習長距離資訊之缺點來進行改善，如加入不同的特徵或其他資訊來幫助估測，抑或是針對時序性倒傳遞演算法的缺點進行結構上的改進。而隨著時代的變遷，語言也不斷地在進化，許多以前沒有的詞語也不停出現，因此用不同平滑化的方法來處理 OOV 的問題也是相當重要的議題。另外，與現行的語言模型結合，如主題模型或鑑別式語言模型等，使語言模型更具有一般性能力、適應性能力，甚至鑑別性能力也是將來值得探討的部分。由於鑑別式語言模型的概念和類神經網路語言模型相當的像，差別在於前者是監督式的，後者是非監督式的。而倘若將類神經網路語言模型改良成監督式的方法，則辨識率應該會有更好的提升，期望在未來能將此兩種語言模型做結合，並進一步的獲得更好的辨識結果。

參考文獻

1. 中文部分

[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。

[劉鳳萍，2009] 劉鳳萍，“使用鑑別式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。

[陳冠宇，2010] 陳冠宇，“主題模型於語音辨識使用之改進，”國立臺灣師範大學資訊工程所碩士論文，2010。

[劉家妏，2010] 劉家妏，“多種鑑別式語言模型應用於語音辨識之研究，” 國立臺灣師範大學資訊工程所碩士論文，2010。

[賴敏軒，2011] 賴敏軒，“實證探究多種鑑別式語言模型於語音辨識之研究，”國立臺灣師範大學資訊工程所碩士論文，2011。

2. 西文部分

[Aubert, 2002] X. L. Aubert, “An overview of decoding techniques for large vocabulary continuous speech recognition,” *Computer Speech and Language*, Vol. 16, No. 1, pp. 89-114, 2002.

[Alexandrescu and Kirchhoff, 2006] A. Alexandrescu and K. Kirchhoff, “Factored neural language models,” in *Proc. North American Chapter of the Association for Computational Linguistics*, pp. 1-4, 2006.

[Arisoy et al., 2010] E. Arisoy, M. Saraclar, B. Roark, and I. Shafran, “Syntactic and

sub-lexical features for Turkish discriminative language models," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5538 -5541, 2010.

[Bahl et al., 1983] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," in *Proc. IEEE Transactions on Patten Analysis and Machine Intelligence*, Vol. PAMI-5, No. 2, pp. 179-190, 1983.

[Bahl et al., 1986] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 49-52, 1986.

[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 Linguistics*, Vol. 18, No. 4, pp. 467-479, 1992.

[Bengio et al., 1993] Y. Bengio, P. Frasconi, and P. Simard, "The problem of learning long-term dependencies in recurrent networks," in *Proc. IEEE International Conference on Neural Networks*, Vol. 3, pp. 1183-1188, 1993.

[Bengio et al., 1994] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transaction on Neural Networks*, Vol. 5, No. 2, pp. 157-166, 1994.

[Bengio et al., 2001] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," in *Proc. Advances in Neural Information Processing Systems*, pp. 933-938, 2001.

[Boden, 2002] Mikael Boden, “A guide to recurrent neural networks and back-propagation,” in the Dallas project, 2002.

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

[Chen and Goodman, 1996] S. F. Chen, and J. Goodman, “An empirical study of smoothing techniques for language modeling,” in *Proc. the 34th annual meeting on Association for Computational Linguistics*, pp. 310-318, 1996.

[Clarkson and Robinson, 1997] P. R. Clarkson, and A. J. Robinson, “Language model adaptation using mixtures and an exponentially decaying cache,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 799-802, 1997.

[Chelba and Jelinek, 2000] C. Chelba, and F. Jelinek, “Structured language modeling,” *Computer, Speech and Language*, Vol. 14, No. 4, pp. 283-332, 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,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 10, No. 1, pp. 1-18, 2004.

[Davis and Mermelstein, 1980] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 28, No. 4, pp. 357-366, 1980.

[Elman, 1990] J. L. Elman, “Finding structure in time,” *Cognitive Science*, Vol. 14, No. 2, pp. 179-211, 1990.

[Gales, 1998] M. J. F. Gales “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer, Speech and Language*, Vol. 12, pp.75-98, 1998.

[Gildea and Hofmann, 1999] D. Gildea and T. Hofmann, “Topic-based language models using EM,” in *Proc. 6th European Conference on Speech Communication and Technology*, pp. 2167-2170, 1999.

[Goodman, 2001] J. Goodman, “Classes for fast maximum entropy training,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 561-564, 2001.

[Goodman, 2001] J. Goodman, “A bit of progress in language modeling,” *Computer, Speech and Language*, pp. 403-434, 2001.

[Gao et al., 2005] 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, 2005.

[Hermansky, 1990] H. Hermansky, “Perceptual linear predictive analysis of speech,” *The Journal of the Acoustical Society of America*, Vol. 87, No. 4, 1990.

[Huang et al., 2007] Z. Huang, M. P. Harper, and W. Wang, “Mandarin part-of-speech tagging and discriminative reranking,” in *Proc. Empirical Methods in Natural Language Processing*, pp. 1093-1102, 2007.

[Jordan, 1986] M. L. Jordan, “Attractor dynamics and parallelism in a connectionist sequential machine,” in *Proc. the Eighth Annual Conference of the Cognitive Science Society*, pp.531-546, 1986.

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

[Katz, 1987] S. M. Katz, “Estimation of probabilities from sparse data for the language model component of a speech recognizer,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-35, No. 3, pp. 400, 1987.

[Kuhn, 1988] R. Kuhn, “Speech recognition and the frequency of recently used words: A modified Markov model for natural language,” in *Proc. International Conference on Computational Linguistics*, pp. 348-350, 1988.

[Kneser and Ney, 1995] R. Kneser and H. Ney, “Improved backing-off for m-gram language modeling,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 181-184, 1995.

[Kumar, 1997] N. Kumar, *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*, Ph.D. dissertation, John Hopkins University, Baltimore, 1997.

[Kang et al., 2011] M. Kang, T. Ng, and L. Nguyen, “Mandarin word-character hybrid-input neural network language model,” in *Proc. International Speech Communication Association*, pp. 625-628, 2011.

[Lawrence et al., 1996] S. Lawrence, C. L. Giles, and S. Fong, “Can recurrent neural networks learn natural language grammars?,” in *Proc. International Conference on Neural Networks*, pp. 1853-1858, 1996.

[Le et al., 2011] H.-S. Le, I. Oparin, A. Allauzen, J.-L. Gauvain, and F. Yvon, “Structured output layer neural network language model,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5524-5527, 2011.

[MacQueen, 1967] J. B. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proc. the fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[Makhoul, 1975] J. Makhoul, “Linear prediction: A tutorial review,” *Proceeding of the IEEE*, Vol. 63, No. 4, pp. 561-580, 1975.

[Mikolov et al., 2009] T. Mikolov, J. Kopecký, L. Burget, O. Glembek, and J. Cernocký, “Neural network based language models for highly inflective languages,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4725-4728, 2009.

[Mikolov et al., 2010] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, “Recurrent neural network based language model,” in *Proc. International Speech Communication Association*, pp. 1045-1048, 2010.

[Mikolov et al., 2011] T. Mikolov, S. Kombrink, L. Burget, J. Cernocký, and S. Khudanpur, “Extensions of recurrent neural network language model,” in *Proc.*

IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5528-5531, 2011.

[Mikolov et al., 2011] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J.Cernocky, "Strategies for training large scale neural network language models," in *Proc. the IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 450-455, 2011.

[Mikolov et al., 2011] T. Mikolov, A. Deoras, S. Kombrink, and L. Burget, "Empirical evaluation and combination of advanced language modeling techniques," in *Proc. International Speech Communication Association*, pp. 605-608, 2011.

[Mikolov et al., 2011] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, and J. Černocký, "RNNLM - Recurrent neural network language modeling toolkit," in *Proc. IEEE workshop on Automatic Speech Recognition and Understanding*, pp.16, 2011.

[Och, 2003] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. the 41st Annual Meeting on Association for Computational Linguistics*, pp. 160-167, 2003.

[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," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5126-5129, 2010.

[Oparin et al., 2012] I. Oparin, M. Sundermeyer, H. Ney, and J. L. Gauvain, "Performance analysis of neural networks in combination with *n*-gram language

models," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 5005-5008, 2012.

[Povey, 2004] D. Povey, *Discriminative training for large vocabulary speech recognition*, Ph.D Dissertation, Peterhouse, University of Cambridge, 2004.

[Park et al., 2010] J. Park, X. Liu, M. J. F. Gales, and P. C. Woodland, "Improved neural network based language modeling and adaptation," in *Proc. International Speech Communication Association*, pp. 1041-1044, 2010.

[Rosenblatt, 1958] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Cornell Aeronautical Laboratory, Psychological Review*, Vol. 65, No. 6, pp. 386-408, 1958.

[Rabiner, 1989] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proc the IEEE*, Vol. 77, No. 2, 1989.

[Rojas, 1996] R. Rojas, "Neural networks: a systematic introduction," Springer-Verlag, 1996.

[Roark et al., 2007] 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.

[Shannon, 1948] C. E. Shannon and W. Weaver, *A mathematical theory of communication*, Urbana, University of Illinois Press, 1948.

[Saul and Pereira, 1997] L. Saul and F. Pereira, “Aggregate and mixed-order Markov models for statistical language processing,” in *Proc. the Conference on Empirical Methods in Natural Language Processing*, pp.81-89, 1997.

[Schuster and Paliwal, 1997] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673-2681, 1997.

[Schwenk, 2004] H. Schwenk, “Efficient training of large neural networks for language modeling,” in *Proc. IEEE International Joint Conference Neural Networks*, Vol. 4, pp. 3059-3064, 2004.

[Schwenk and Gauvain, 2005] H. Schwenk and J. L. Gauvain, “Training neural network language models on very large corpora,” in *Proc. Empirical Methods in Natural Language Processing*, pp. 201-208, 2005.

[Schwenk et al., 2007] H. Schwenk, M. R. Costa-jussa, and Jose A. R. Fonollosa, “Continuous space language models,” in *Proc. International Workshop on Spoken Language Translation*, pp. 166-173, 2007.

[Sarikaya et al., 2010] R. Sarikaya, A. Emami, M. Afify, and B. Ramabhadran, “Continuous space language modeling technique,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5186-5189, 2010.

[Sak et al., 2010] H. Sak, M. Saraclar, and T. Güngör, “Morphology-based and sub-word language modeling for Turkish speech recognition,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5402-5405, 2010.

[Towsey et al., 1998] M. Towsey, J. Diederich, I. Schellhammer, S. Chalup, and C. Brugman, "Natural language learning by recurrent neural networks: A comparison with probabilistic approaches," in *Proc. the joint conference on new methods in language processing and computational natural language learning*, pp. 3-10, 1998.

[Troncoso et al., 2004] C. Troncoso, T. Kawahara, H. Yamamoto, and G. Kikui, "Trigger-based language model construction by combining different corpora," *Institute of Electronics, Information and Communication Engineers Technical Report*, Vol. 104, No. 542, pp. 25-30, 2004.

[Tam and Schultz, 2005] Y. C. Tam and T. Schultz, "Dynamic language model adaptation using variational Bayes inference," in *Proc. 9th European Conference on Speech Communication and Technology*, pp. 5-8, 2005.

[Viterbi, 1967] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transaction on Information Theory*, Vol. 13, No. 2, pp. 260-269, 1967.

[Villiers and Barnard, 1992] J. Villiers and E. Barnard, "Back-propagation neural nets with one and two hidden layers," *IEEE Transaction on Neural Network*, Vol. 4, No. 1, pp. 136-141, 1992.

[Werbos, 1990] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, Vol. 78, No. 10, pp. 1550-1560, 1990.

[Wang et al., 2005] 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.

[Xu and Rudnicky, 2000] W. Xu and A. Rudnicky, "Can artificial neural networks learn language models?," in *Proc. International Conference on Speech and Language Processing*, pp. 202-205, 2000.

[Zamora-Martinez et al., 2009] F. Zamora-Martinez, M. J. Castro-Bleda, and S. Espana-Boquera, "Fast evaluation of connectionist language models," in *Proc. International Work Conference on Artificial Neural Networks*, Vol. 5517, pp. 33-40, 2009

[Zamora-Martinez et al., 2012] F. Zamora-Martinez, S. España-Boquera, and M. J. Castro-Bleda, "Cache neural network language models based on long-distance dependencies for a spoken dialog system," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4993-4996, 2012.