Computational Linguistics and Chinese Language Processing
Vol. 19, No. 4, December 2014, pp. 47-60

使用概念資訊於中文大詞彙連續語音辨識之研究

Exploring Concept Information for Mandarin Large Vocabulary Continuous Speech Recognition

郝柏翰、陳思澄、陳柏琳

Po-Han Hao*, Ssu-Cheng Chen*, and Berlin Chen*

摘要

語言模型是語音辨識系統中的關鍵組成，其主要的功能通常是藉由已解碼的歷史詞序列資訊來預測下一個詞彙為何的可能性最大，以協助語音辨識系統從眾多混淆的候選詞序列假設中找出最有可能的結果。本論文旨在於發展新穎動態語言模型調適技術，用以輔助並彌補傳統 N 連(*N*-gram)語言模型不足之處，其主要貢獻有二。首先，我們提出所謂的概念語言模型(Concept Language Model, CLM)，其主要目的在於近似隱含在歷史詞序列中語者內心所欲表達之概念，並藉以獲得基於此概念下詞彙使用分布資訊，做為動態語言模型調適之線索來源。其次，我們嘗試以不同方式來估測此種概念語言模型，並將不同程度的鄰近資訊(Proximity Information)融入概念語言模型以放寬其既有詞袋(Bag-of-Words)假設的限制。本論文是以中文大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)為任務目標，以比較我們所提出語言模型調適技術與其它當今常用技術之效能。實驗結果顯示我們的語言模型調適技在以字錯誤率(Character Error Rate, CER)評估標準之下，對於僅使用 N 連語言模型的基礎語音辨識系統皆能有明顯的效能提升。

關鍵詞： 語音辨識、語言模型、概念資訊、模型調適

* 國立臺灣師範大學資訊工程學系
Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: {ie965225, boe20211}@gmail.com; berlin@csie.ntnu.edu.tw
The authors for correspondence are Ssu-Cheng Chen and Berlin Chen.

Abstract

Language modeling (LM) is part and parcel of automatic speech recognition (ASR), since it can assist ASR to constrain the acoustic analysis, guide the search through multiple candidate word strings, and quantify the acceptability of the final output hypothesis given an input utterance. This paper investigates and develops language model adaptation techniques for use in ASR and its main contribution is two-fold. First, we propose a novel concept language modeling (CLM) approach to rendering the relationships between a search history and an upcoming word. Second, the instantiations of CLM are constructed with different levels of lexical granularities, such as words and document clusters. In addition, we also explore the incorporation of word proximity cues into the model formulation of CLM, getting around the "*bag-of-words*" assumption. A series of experiments conducted on a Mandarin large vocabulary continuous speech recognition (LVCSR) task demonstrate that our proposed language models can offer substantial improvements over the baseline *N*-gram system, and achieve performance competitive to, or better than, some state-of-the-art language model adaptation methods.

Keywords: Speech Recognition, Language Model, Concept Information, Model Adaptation

1. Introduction

語言模型(Language Models, LM)已被廣泛地使用於語音辨識、機器翻譯、資訊檢索以及文件摘要等各種任務之中，並成為關鍵的組成(Rosenfeld, 2000; Bellegarda, 2004)。在語音辨識任務上，其主要的功能通常是藉由已解碼的歷史詞序列(Word History)資訊來預測下一個詞彙(Upcoming Word)為何的可能性最大，以協助語音辨識系統從眾多混淆的候選詞序列假設(Candidate Word Sequence Hypotheses)中找出最有可能的結果(Furui *et al*., 2012; O'Shaughnessy *et al*., 2013)。最重要也最為常用的語言模型是 *N* 連(*N*-gram)語言模型，諸如二連(Bigram)與三連(Trigram)語言模型。*N* 連語言模型被用來估測每一個待預測詞彙在其先前緊鄰的 *N*-1 個詞彙已知的情況下出現的條件機率；由此可知，*N* 連語言模型是假設每一個詞彙出現的機率僅與它緊鄰的前 *N*-1 個詞彙有關，並以多項式分布(Multinomial Distribution)表示之。然而 *N* 連語言模型仍存在著許多缺點需要改善，至少有三點：(1)*N* 連語言模型限制了 *N* 的大小，僅能擷取短距離的詞彙規則資訊，無法考慮長距離的語句或篇章資訊；(2)當 *N* 增加時不僅會使模型參數量呈現指數性的遞增，造成空間與時間複雜度快速增加，也容易遭遇資料稀疏、無法為每一種詞序列的排列組合估測出準確的機率值的問題；(3)*N* 連語言模型極容易面臨訓練語料與測試語料不匹配(Mismatch)而造成的估測誤差。有鑑於此，近十幾年來有許多動態語言模型調適技術被提出，用以發展有效的語言模型輔助並彌補傳統 *N* 連(*N*-gram)語言模型不足之處。常見的有快取模型(Cache Model)(Kuhn, 1988)，以及源自於資訊檢索領域的主題模型(Topic

Model)(Blei & Lafferty, 2009)等；而主題模型在語音辨識任務的實作上，又以機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA)(Hofmann, 1999)以及其延伸狄利克里分配(Latent Dirichlet Allocation, LDA)(Blei *et al*., 2003)最普遍被使用。

本論文旨在於發展新穎動態語言模型調適技術，用以輔助並彌補傳統 *N* 連(*N*-gram)語言模型不足之處。首先，我們提出所謂的概念語言模型(Concept Language Model, CLM)，其主要目的在於探詢隱含在歷史詞序列中語者內心所欲表達之概念，並藉以獲得基於此概念下詞彙使用分布資訊，做為動態語言模型調適之線索來源。其次，我們嘗試以不同模型架構與估測方式來建立此種概念語言模型，並將不同程度的鄰近資訊(Proximity Information)融入概念語言模型以放寬其既有詞袋(Bag-of-Words)假設的限制。本論文是基於公視電視新聞語料庫來進行中文大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition)實驗，以比較本論文所提出語言模型調適技術與其它當今常用語言模型調適技術之效能。

本論文的後續安排如下：第二節回顧當今常見的語言模型調適技術；第三節介紹本論文提出的概念模型以及不同模型估測方式，並嘗試將鄰近資訊(Proximity Information)融入概念語言模型；第四節介紹實驗語料、實驗設定以及實驗結果分析；第五節比較不同的語言模型；第六節則是結論及未來展望。

2. 常見的動態語言模型調適技術

動態語言模型調適宗旨在於希望在語音辨識過程中動態調整語言模型對於詞彙出現的預測機率，以獲得最好的語音辨識效能。本節將扼要回顧在語音辨識領域常被使用的動態語言模型調適技術。

2.1 快取模型

快取模型(Cache Model)是在二十多年前年首次被提出(Kuhn, 1988)，用在語音辨識過程中動態來輔助或調整 *N* 連語言模型於預測詞彙出現的機率。其基本概念是如果我們講了一些詞彙，則一段時間內這些詞彙再次出現的機率會很高。我們因此可以利用此線索在語音辨識過程中不斷地產生一個語言模型(例如單連快取模型)，並透過線性組合的方式與原始 *N* 連語言模型(例如三連語言模型)結合來動態地調適語音辨識所需的語言模型：

$$\hat{P}_{\text{Trigram}}(w_i \mid w_{i-2}w_{i-1}) = \lambda \cdot P_{\text{Trigram}}(w_i \mid w_{i-2}w_{i-1}) + (1-\lambda)\cdot \frac{n(w_i, H_i)}{|H_i|} \quad (1)$$

其中 $|H_i|$ 代表詞彙 w_i 對應的歷史詞序列 H_i 中的總詞數； $n(w_i, H_i)$ 是 w_i 在 H_i 出現的次數。過去許多研究亦實驗了二連快取(Bigram Cache)模型、三連快取(Trigram Cache)模型等更高階的快取模型，但由於歷史詞序列可能存有許多辨識錯誤資訊，以歷史詞序列來建立模型調適基礎 *N* 連語言模型的效果通常不是很顯著。

2.2 觸發對模型

觸發對模型(Trigger-Pair Model)模型可視為快取模型的延伸(Lau *et al*., 1993; Troncoso & Kawahara, 2005)，其概念簡單來說是由訓練語料來統計出當任一詞彙 w_x 出現後，在同一文件中的一定間隔內會伴隨著另一詞彙 w_y 出現的可能性為何，這種伴隨關係稱之為「觸發對」(Trigger-pair)，其中 w_x 稱之為觸發項， w_y 稱之為被觸發項。觸發項與被觸發項的統計資訊可以藉由訓練語料中，統計、收集兩兩詞序列之間的平均交互資訊(Mutual Information)量多寡或是使用詞頻數(Term Frequency)與反文件頻數(Inverse Document Frequency)的關係來決定是否形成一個觸發對，以及其對應的條件機率 $P(w_y \mid w_x)$。觸發對模型運用於語言模型時，是由待預測詞彙 w_i 對應的歷史詞序列 H_i 中尋找詞彙 w_i 的可能的觸發項 $h_1, h_2, \cdots, h_{L_i}$ (假設歷史詞序列 $H_i = h_1, h_2, \cdots, h_{L_i}$ ，而每一個歷史詞彙 h_l 對於詞彙 w_i 的觸發機率為 $P(w_i \mid h_l)$)，並將這些觸發項分別預測的條件機率 $P(w_i \mid h_l)$ 動態線性組合而成為觸發對模型：

$$P_{\text{Trigger}}(w_i \mid H_i) = \frac{1}{L_i - 1} \sum_{l=1}^{L_i - 1} P(w_i \mid h_l) \tag{2}$$

而式(2)動態產生的觸發對模型亦可再透過線性組合方式與原始 N 連語言模型結合來動態調適語音辨識所需的語言模型(如式(1)的結合方式)。

2.3 主題模型

通常在資訊檢索任務上，主題模型藉由一組潛藏主題分布用來描述“詞彙-文件”共同出現的特性(Blei & Lafferty, 2009)。當主題模型被應用至語音辨識過程時，待預測詞彙 w_i 與其對應歷史詞序列 H_i (在此可視為一篇文件)之相互關係其有一組潛藏的主題分布用來描述歷史詞序列 H_i 與待預測詞彙 w_i 共同出現關係，不再是單純地經由計算 w_i 在 H_i 的出現頻率而估測，而是透過 w_i 出現在不同潛藏主題分布的頻率以及 H_i 產生這些潛藏主題的可能性來決定，是某種程度上的概念比對(Concept Matching)。機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA)(Hofmann, 1999)以及其延伸狄利克里分配(Latent Dirichlet Allocation, LDA)(Blei *et al*., 2003)是最常被使用的主題模型實例。在此舉機率式潛藏語意分析為例來作說明，當其被用至語音辨識來進行語言模型調適時，基於歷史詞序列 H_i 來預測詞彙 w_i 的發生機率可表示為(Gildea & Hofmann, 1999)：

$$P_{\text{PLSA}}(w_i \mid H_i) = \sum_{k=1}^{K} P(w_i \mid T_k) P(T_k \mid H_i) \tag{3}$$

其中 T_k 為某一個潛在主題，而 $P(w_i \mid T_k)$ 與 $P(T_k \mid H_i)$ 分別表示詞彙 w_i 發生在主題 T_k 的機率以及歷史詞序列 H_i 產生此主題的機率。我們假設每一個潛藏主題產生候選詞的機率 $P(w_i \mid T_k)$ 不因詞序列搜尋及拓展過程而變動，可先藉由最大化調適(或訓練)語料發生機率而求得；但由於歷史詞序列在語音辨識之前不能事先決定，而且數量非常多並且會隨語音辨識過程演進而改變，每一個歷史詞序列對於主題分布的權重必須在語音辨識過程使用期望值最大化(Expectation Maximization, EM)演算法(Dempster, 1977)來進行線上

(動態)估測。機率式潛藏語意分析的優點是在決定待預測詞彙 w_i 發生的機率時，不僅會考慮整個歷史詞序列 H_i 的主題分布特性，而且會隨語音辨識候選詞序列搜尋的演進，動態調整詞序列所含有的潛藏主題分布資訊。而式(3)動態地產生的機率式潛藏語意分析模型亦可再透過線性組合方式與原始 *N* 連語言模型結合來動態調適語音辨識所需的語言模型(如式(1)的結合方式)。

另一方面，狄利克里分配擁有與機率式潛藏語意分析相似的數學表示式，可視為後者之延伸，而且狄利克里分配在許多語音辨識任務上都展現了不錯的效用(Tam & Schultz, 2005)。兩個模型間的主要差異在於機率式潛藏語意分析假設其模型參數在參數空間上是固定和未知向量，而狄利克里分配對於模型參數多了先備限制(a Priori Constraints)，認為參數向量本身也是隨機變數，遵循著某種狄利克里分布特性。由於狄利克里分配模型的最佳化較為困難、不容易達到正確的估測，許多近似的估測演算法像是變動性貝氏近似(Variational Approximation)演算法或是吉卜森取樣(Gibbs Sampling)演算法因此被提出來估測狄利克里分配之模型參數(Blei & Lafferty, 2009)。關於主題模型的回顧與近期發展，可以參考(Blei, 2014; Kim *et al*., 2013; Potapenko & Konstantin, 2013)。

3. 概念語言模型

在本論文我們提出所謂的概念語言模型(Concept Language Model, CLM)來實踐語言模型調適，其主要假設是認為每一句的語句都是用來代表語者內心隱含而欲傳達的概念，並藉由語言(及語音)來具體表達相對應的概念。而概念模型最主要的目的則是希望能夠獲取使用者欲表達的概念，並假設在同一概念之中歷史詞序列中所有詞彙以及待預測詞彙具有共同的關係，進而藉此共同關係達到預測詞彙出現機率的目的。在實作上，概念模型會使用(搜尋)與初步語音辨識結果近似同領域文件(或調適語料)內表述的若干概念，用以近似語者內心欲傳達的真正含意，並基於此來建立概念語言模型。在本論文之中，概念模型的建立是分兩個面向來探討，分別是「詞彙」面向與「群聚」面向；以下將依序做介紹。

3.1 以詞彙面向建立概念語言模型

在我們想要表達某一特定概念時，我們常常會利用一組具有代表性的「概念關鍵詞」(Concept Words)來表達我們對事物的看法，而在同一概念底下用來描述事物的概念關鍵詞之間則具有相當高的關聯程度。例如在馬致遠的【天淨沙・秋思】之中，連續使用了「枯藤」、「老樹」等多個名詞串接，並藉由這一組連續的詞彙之組合來描述秋天蕭瑟荒涼的景象。基於此，本論文提出所謂詞概念語言模型(Word-based Concept Language Model, WCLM)，並應用於語言模型調適。在建構詞概念語言模型時，我們期望能夠針對每一語句不同的語言意涵，在調適語料的若干文件中挑選一組具有代表性的概念關鍵詞組 **c**，藉以描述任一對歷史詞序列中所有詞彙與待預測詞彙之間的相依關係，如式(4)所示：

$$
\begin{aligned}
P_{\mathrm{WCLM}}(w_i \mid H_i, W) &= \frac{P(w_i, H_i \mid W)}{P(H_i \mid W)} \\
&= \frac{\sum_{c \in \mathbf{c}} P(w_i, H_i \mid c) P(c \mid W)}{\sum_{c' \in \mathbf{c}} P(H_i \mid c') P(c' \mid W)} \\
&= \frac{\sum_{c \in \mathbf{c}} P(w_i \mid c) \prod_{l=1}^{L_i} P(h_l \mid c) P(c \mid W)}{\sum_{c' \in \mathbf{c}} \prod_{l'=1}^{L_i} P(h_{l'} \mid c') P(c' \mid W)}
\end{aligned} \tag{4}
$$

其 W 代表語者所講語句所欲表達的語言資訊，在此我們先以語音辨識初步(第一階段)所產生的詞圖(Word Graph)(Ortmanns *et al*., 1997)來近似(詞圖包含所有可能的候選詞序列)；而 **c** 代表與 W 所欲表達的語言資訊有關的一組概念關鍵詞組。從式(4)的推導可看出詞概念語言模型欲模型化(紀錄)當某個概念關鍵詞 c 出現的情況下，待預測詞彙 w_i 與其歷史詞序列 H_i 共同出現的關係。同時，考量模型估測之可行性，式(4)進一步假設當某一個概念關鍵詞 c 出現的情況下，待預測詞彙 w_i 與其歷史詞序列 H_i 中任意的詞彙之間是彼此獨立的，也就是所謂的詞袋(Bag-of-Words)假設。而式(4)中 $P(w_i \mid c)$ 與 $P(h_l \mid c)$ 可從調適語料庫裡概念關鍵詞 c 所出現處的鄰近資訊(Proximity Information)，或者說是出現處上下文的詞彙分布而估測得； $P(c \mid W)$ 可透過適當方式計算 W 與 c 之相似度而求得。

實務上，我們首先遭遇到的問題就是「如何挑選具代表性的關鍵詞組？」。為此，本論文在挑選概念關鍵詞時運用了兩階段的挑選方式，如圖 1 所示。在第一階段時，我們利用了在資訊檢索領域之中常使用的虛擬關聯回饋 (Pseudo-Relevance Feedback, PRF)(Baeza-Yates & Ribeiro-Neto, 2011)，並利用基於庫爾貝克—萊伯勒差異量(Kullback-Leibler Divergence, KL-Divergence)之查詢與文件模型化技術(Kullback & Leibler, 1951; Zhai, 2008)，以詞圖 W(含有欲表達的詞彙和語意資訊)為查詢從調適語料的文件集檢索出一組較為相關的文件子集，稱這些文件為虛擬關聯文件(Pseudo-Relevance Documents)，並假設這些文件含有與所欲表達的語言資訊有關的概念。

在第二階段時，我們進一步從虛擬關聯文件子集裡挑選出一組一定數量的概念關鍵詞組，然後藉由這組概念關鍵詞組來量化(機率化)歷史詞序列中所有詞彙與待預測詞彙在此概念關鍵詞組下的共同出現關係。關於概念關鍵詞挑選準則，我們可以基於詞頻與反向文件頻率分數(TF-IDF Score)(Baeza-Yates & Ribeiro-Neto, 2011)。詞頻與反向文件頻率分數是一項常被用於資訊檢索以及文字分析領域中的技術，其公式可以表示如下：

$$
w_{j,m} = \begin{cases} (1 + \log f_{j,m}) \times \log(N / n_j) & \textit{if} \quad f_{j,m} > 0 \\ 0 & \textit{ohterwise} \end{cases} \tag{5}
$$

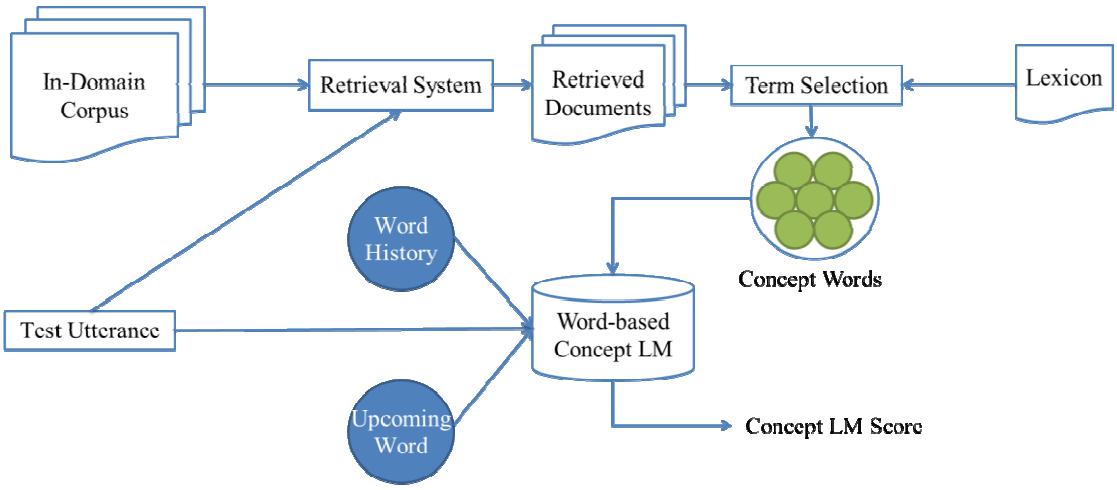

圖 1. 詞概念語言語言模型流程圖

上述的詞頻與反向文件頻率分數主要可分為兩個主要部分：第一部分為 $(1+\log f_{j,m})$，其中的 $f_{j,m}$ 則代表詞彙 w_j 在此文件 d_m 中所出現的次數，稱之為詞頻(Term Frequency, TF)，可以解釋為具越高詞頻的詞彙對文件來講越重要；第二部分為 $\log(N/n_j)$，其中 n_j 之則是代表詞彙 w_j 出現在所有虛擬關聯文件的文件個數，稱之為反向文件頻率(Inverse Document Frequency, IDF)，當某一詞彙出現僅出現在少數的文件之中，則此詞彙越具有獨特性。我們期望透過式(5)能找出具有重要性與獨特性的詞彙做為概念關鍵詞。

3.2 以群聚面向建立概念語言模型

群聚概念語言模型(Cluster-based Concept Language Model, CCLM)假設在調適語料的文件集內之文件可以由一組概念類別 **C** 來表示，藉由語者講的語句所欲表達的語言資訊 W 與這些概念類別的個別關聯程度來獲得語句可能的概念分布，並做為語言模型預測的根據：

$$\begin{aligned} P_{\text{CCLM-1}}(w_i \mid H_i, W) &= \frac{\sum_{C\in\mathbf{C}} P(w_i, H_i \mid C)P(C \mid W)}{\sum_{C'\in\mathbf{C}} P(H_i \mid C')P(C' \mid W)} \\ &= \frac{\sum_{C\in\mathbf{C}} P(w_i \mid C)\prod_{l=1}^{L_i} P(h_l \mid C)P(C \mid W)}{\sum_{C'\in\mathbf{C}} \prod_{l'=1}^{L_i} P(h_{l'} \mid C')P(C' \mid W)} \end{aligned} \tag{6}$$

其中概念類別的求取可透過一般分群演算法諸如 *K*-Means 演算法(Baeza-Yates & Ribeiro-Neto, 2011)而求得；$P(C \mid W)$ 可基於將語言資訊 W 與每一個概念類別 C 表示成向量形式，計算 W 與 C 之(餘弦)相似度而求得；$P(w_i \mid C)$ 代表概念類別 C 預測詞彙 w_i 的單連語言模型機率，可透過最大化相似機率估測而得(Zhai, 2008)。從式(6)的推導可看出群聚概念語言模型欲模型化(紀錄)當某一個概念類別 C 出現的情況下，待預測詞彙 w_i 與其歷史詞序列 H_i 共同出現的關係。

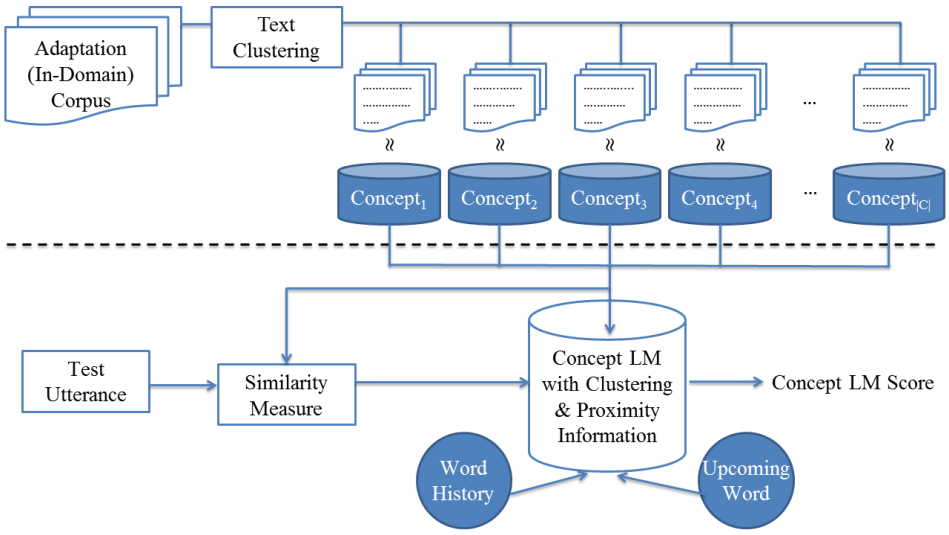

圖 2. 群聚概念語言模型示意圖

我們可以將式(6)中概念類別 C 預測詞彙 w_i 的語言模型延伸成為雙連(Bigram)或者三連(Trigram)語言模型，而可分別得到下面兩個表示式：

$$P_{\text{CCLM-2}}(w_i \mid H_i, W) = \frac{\sum_{C \in \mathbf{C}} P(w_i \mid h_L, C) P(h_1 \mid C) \prod_{l=2}^{L_i} P(h_l \mid h_{l-1}, C) P(C \mid W)}{\sum_{C' \in \mathbf{C}} P(h_1 \mid C') \prod_{l'=2}^{L_i} P(h_{l'} \mid h_{l'-1}, C') P(C' \mid W)} \tag{7}$$

$$P_{\text{CCLM-3}}(w_i \mid H_i, W) = \frac{\sum_{C \in \mathbf{C}} P(w_i \mid h_{L-1}, h_L, C) P(h_1 \mid C) P(h_2 \mid h_1, C) \prod_{l=3}^{L_i} P(h_l \mid h_{l-2}, h_{l-1}, C) P(C \mid W)}{\sum_{C' \in \mathbf{C}} P(h_1 \mid C') P(h_2 \mid h_1, C') \prod_{l'=3}^{L_i} P(h_l \mid h_{l-2}, h_{l-1}, C') P(C' \mid W)} \tag{8}$$

如此一來，概念語言模型可以同時考慮詞彙間出現的先後規則性或者是鄰近資訊(Proximity Information)，可以免除以詞袋(Bag-of-Words)假設的限制。最後，式(4)、式(6)、式(7)與式(8)動態產生的各種不同概念語言模型亦可再透過線性組合方式分別與原始 N 連語言模型結合來動態調適語音辨識所需的語言模型(如式(1)的結合方式)。圖 2 為群聚概念語言模型之示意圖。

4. 實驗設定與結果討論

4.1 實驗語料

本論文的語音辨識實驗是使用台師大所自行研發的大詞彙連續語音辨識系統(所使用詞典大小約為 7 萬 2 千詞)(Chen *et al.*, 2004)以及公視新聞的公視電視新聞語音語料庫(Mandarin Across Taiwan Broadcast News, MATBN)(Wang *et al.*, 2005)。此新聞語音語料庫是由中央研究院資訊所口語小組耗時三年(2001~2003)與公共電視台合作錄製完成。我

們初步地選擇外場採訪記者語料作為實驗題材，將其中約 25 小時收錄於 2001 年 11 月至 2002 年 12 月期間的語料作為最小化音素錯誤(Minimum Phone Error, MPE)聲學模型訓練的語料以建立聲學模型(Acoustic Models)(Liu *et al.*, 2007)。另外，本論文以 2003 年所蒐集的語料中挑選各約 1.5 個小時作為發展集語料(Development Set)以及測試集語料(Test Set)，分別包含了 292 與 307 句的語句；我們以發展集語料來最佳化語言模型訓練所需之參數設定，然後據此作用在測試集語料。

表 1. 語音辨識實驗所使用之發展集語音語料以及測試集語音語料統計資訊

語料	句數	長度(小時)	說話速度
發展集語料	292	約 1.5	8.52 字/秒
測試集語料	307	約 1.5	8.50 字/秒

表 2. 語言模型估測所使用背景文字語料以及調適文字語料統計資訊

語料	詞數	句數
調適語料	約 1,000,000	3,643
背景語料	約 80,000,000	2,068,991

在語言模型的估測上，我們使用自 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)做為背景語料庫用來訓練三連語言模型(Trigram Language Model)，此語言模型是使用 SRI Language Modeling Toolkit (SRILM)(Stolcke, 2000)訓練而得，採用 Good-Turning 平滑化方法來解決資料稀疏的問題。另一方面，我們亦蒐集同為公視電視新聞語料庫中的同領域文件做為調適語料庫，用來估測本論文中所探討的各式做為調適之用的語言模型，總共約三千六百多文句。本論文實驗所使用之語音語料庫以及文字語料庫的扼要統計資訊分別如表 1 與表 2 所示。

4.2 基礎實驗結果

在第一組實驗，我們於表 3 列出基礎語音辨識系統(使用背景三連語言模型，表示成 Background Trigram)以及一些常用語言模型調適技術在測試集語料的語音辨識字錯誤率(Character Error Rate, CER)結果，包括了觸發對模型(記作 Trigger)、機率式潛藏語意分析(記作 PLSA)以及狄利克里分配(記作 LDA)。值得一提的是，機率式潛藏語意分析以及狄利克里分配所使用潛藏主題數目設為 128；而這些語言模型調適技術都是作用在語音辨識的第二階段，也就是詞圖候選詞序列的語言模型重新排列(Word Graph Rescoring)。由表 3 我們可以觀察出三個現象。首先，觸發對模型(Trigger)似乎未能對基礎語音辨識系統的效能有顯著的提升。其次，機率式潛藏語意分析(PLSA)以及狄利克里分配(LDA)獲

得相同的語音辨識效能；相對於基礎語音辨識系統而言，能有約 4.6%的相對字錯誤率降低。第三，狄利克里分配雖使用較複雜的模型參數分布假設與估測演算法，但在我們的實驗裡並沒有獲得比機率式潛藏語意分析明顯較好的成果。

表 3. 語音辨識字錯誤率(%)：分別使用背景三連語言模型以及其它常見語言模型調適技術

Trigram	Trigger	PLSA	LDA
20.08	20.02	19.15	19.15

表 4. 語音辨識字錯誤率(%)：使用不同概念語言模型，包括 WCLM 、CCLM-1 、CCLM-2 、CCLM-3 。

WCLM	CCLM-1	CCLM-2	CCLM-3
19.30	19.26	19.18	19.03

表 5. 語音辨識字錯誤率(%)：群聚概念語言模型(CCLM-1 、CCLM-2 、CCLM-3) 使用不同概念類別(群聚)數目。

概念類別(群聚)數目	CCLM-1	CCLM-2	CCLM-3
8	19.24	19.18	19.03
16	19.26	19.11	19.11
32	19.33	19.24	19.21
64	19.26	19.13	19.09
128	19.37	19.31	19.27

4.3 概念語言模型實驗結果

接著，我們評估本論文所提出兩類概念語言模型的語音辨識效能：詞概念語言模型(記作 WCLM)與群聚概念語言模型(記作 CCLM)。其中群聚概念語言模型因為單連語言模型、雙連語言模型和三連語言模型的使用(參考第三節)可以有三種變形(分別記作 CCLM-1、CCLM-2、CCLM-3)。基於在發展集語料所得出的最佳模型設定，在此 WCLM 共使用 128 個概念關鍵詞，而 CCLM-1、CCLM-2、CCLM-3 所使用的概念類別(群聚)數目分別為 16、8 與 8。它們在測試集語料的語音辨識字錯誤率結果列於表 4。基於表 4 的結果，我們有下列幾個觀察。首先，詞概念語言模型(WCLM)能較基礎語音辨識系統有一定的效能提升(約 3.8%的相對字錯誤率降低)，但其效用較機率式潛藏語意分析(PLSA)以及狄利克里

分配(LDA)來的稍差。其次，群聚概念語言模型在使用雙連語言模型和三連語言模型做為其組成模型(Component Models)時(參考式(7)與式(8))，能達到與機率式潛藏語意分析以及狄利克里分配差不多甚至更好的效果，例如 CCLM-3 能較基礎語音辨識系統有約 5.2%相對字錯誤率降低。值得注意的是我們所提出的詞概念語言模型與群聚概念語言模型僅需要在進行詞圖候選詞序列的語言模型重新排列之前，執行一次文件檢索或者與概念類別(群聚)相似度估算，並不需像機率式潛藏語意分析以及狄利克里分配一樣在詞圖候選詞序列之語言模型重新排列時重新估算其組成模型，所以執行速度上會來得較快。另一方面，我們也嘗試結合詞概念語言模型與群聚概念語言模型，透過線性組合方式同時來調適基礎語音辨識系統所用之背景三連語言模型，而能讓字錯誤率下降至 18.98%。

最後，由於群聚概念語言模型能在上述實驗中獲得相當具競爭力的結果，我們因此進一步觀察它的變形(CCLM-1、CCLM-2、CCLM-3)在測試集語料使用不同概念類別(群聚)數目的表現，如表 5 所示。當我們比較表 4 與表 5 時可以發現，使用基於發展集語料所得最佳概念類別(群聚)數的各種群聚概念語言模型實際上在測試集語料上都有相當好的效能；顯示利用發展集語料所求得的模型(複雜度)參數稍後在測試集語料都能有一致的效能表現。

5. 結論與未來展望

在本論文，我們比較了一些常見語言模型調適技術在中文大詞彙連續語音辨識的效能。此外，我們提出所謂的概念語言模型(Concept Language Model, CLM)，其主要目的在於近似隱含在歷史詞序列中語者內心所欲表達之概念，並藉以獲得基於此概念下詞彙使用分布資訊，做為動態語言模型調適之線索來源。再者，我們嘗試以不同模型架構以及估測方式來實作此種概念語言模型，包括了將不同程度的鄰近資訊(Proximity Information)融入概念語言模型以放寬詞袋(Bag-of-Words)假設的限制。在基於公視電視新聞語料庫所進行的實驗顯示，我們所提出建構在概念語言模型之上的語言模型調適技術與其它當今常用技術相比，都夠達到具競爭性甚至較好的效能。關於未來研究方向，我們希望能結合或使用其它較新穎的模型，諸如遞迴式類神經網路語言模型(Recurrent Neural Network Language Model, RNNLM)(Mikolov *et al*., 2010; Deng & Yu, 2014)，來實現概念語言模型所欲擷取的詞彙和語意使用資訊。同時，我們亦希望能將其它在資訊檢索領域以發展相當不錯的新穎語言模型(Blei, 2014; Chen *et al*., 2004; Kim *et al*., 2013; Zhai, 2008)應用到中文大詞彙連續語音辨識的任務上。

致謝

本論文之研究承蒙教育部－國立臺灣師範大學邁向頂尖大學計畫(102J1A0800)與行政院科技部研究計畫(MOST 103-2221-E-003-016-MY2, NSC 103-2911-I-003-301, NSC 101-2221-E-003-024-MY3 、 NSC 101-2511-S-003-057-MY3 、 NSC 101-2511-S-003-047-MY3 和 NSC 102-2221-E-003-014- MY3)之經費支持，謹此致謝。

參考文獻

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern Information Retrieval: the Concepts and Technology behind Search*, Addison-Wesley Professional.

Bellegarda, J. R. (2004). Statistical language model adaptation: review and perspectives. *Speech Communication*, *42*(11), 93-108.

Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, *1*, 203-232.

Blei, D. M, & Lafferty, J. (2009). Topic models. in Srivastava, A., & Sahami, M., (eds.), *Text Mining: Theory and Applications*, Taylor and Francis, 2009.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Chen, B., Kuo, J.-W., & Tsai, W.-H. (2004). Lightly supervised and data-driven approaches to Mandarin broadcast news transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, 777-780.

Chen, K.-Y., Liu, S.-H., Chen, B., Wang, H.-M., Hsu, W.-L., Chen, H.-H., & Jan, E.-E. (2014). Leveraging effective query modeling techniques for speech recognition and summarization. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Dempster, A. P., Laird , N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, *39*(1), 1-38.

Deng, L., & Yu, D. (2014). *Deep Learning: Methods and Applications*, Foundations and Trends in Signal Processing, Now Publishers.

Gildea, D., & Hofmann, T. (1999). Topic-based language models using EM. In *Proceedings of the European Conference on Speech Communication and Technology*, 2167-2170.

Furui, S., Deng, L., Gales, M., Ney, H., & Tokuda, K. (2012). Fundamental technologies in modern speech recognition. *IEEE Signal Processing Magazine*, *29*(6), 16-17.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceeding of the ACM Special Interest Group on Information Retrieval*, 50-57.

Kim, D.-k., Voelker, G. M., & Saul, L. K. (2013). A variational approximation for topic modeling of hierarchical corpora. In *Proceedings of the International Conference on Machine Learning*.

Kuhn, R. (1988). Speech recognition and the frequency of recently used words: A modified Markov model for natural language. In *Proceedings of International Conference on Computational Linguistics*, 348-350.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79-86.

Lau, R., Rosenfeld, R., & Roukos, S. (1993). Trigger-based language models: a maximum entropy approach. In *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, 45-48.

Liu, S.-H., Chu, F.-H., Lin, S.-H., Lee, H.-S., & Chen, B. (2007). Training data selection for improving discriminative training of acoustic models. In *Proceedings of IEEE workshop on Automatic Speech Recognition and Understanding*, 284-289.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of the Annual Conference of the International Speech Communication Association*, 1045-1048.

Ortmanns, S., Ney, H., & Aubert, X. (1997). A word graph algorithm for large vocabulary continuous speech recognition. *Computer Speech and Language*, 11, 43-72.

O'Shaughnessy, D., Deng, L., & Li, H. (2013). Speech information processing: Theory and applications. *Proceedings of the IEEE*, *101*(5), 1034-1037.

Potapenko, A., & V. Konstantin. (2013). Robust PLSA performs better than LDA. In *Proceedings of the European Conference on Information Retrieval*, 784-787.

Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here? *Proceedings of IEEE, 88*(8), 2000, 1270-1278.

Stolcke, A. (2000). *SRI Language Modeling Toolkit*. Available at: http://www.speech.sri.com/projects/srilm/.

Tam, Y., & Schultz, T. (2005). Dynamic language model adaptation using variational Bayes inference. In *Proceedings of the Annual Conference of the International Speech Communication Association*, 5-8.

Troncoso, C., & Kawahara, T. (2005). Trigger-based language model adaptation for automatic meeting transcription. In *Proceedings of the Annual Conference of the International Speech Communication Association*, 1297-1300.

Wang, H.-M., Chen, B., Kuo, J.-W., & Cheng, S.-S. (2005). MATBN: a Mandarin Chinese broadcast news corpus. *International Journal of Computational Linguistics & Chinese Language Processing*, *10*(1), 219-235.

Zhai, C. X. (2008). Statistical language models for information retrieval: A critical review. *Foundations and Trends in Information Retrieval*, *2*(3), 137-213.