主題語言模型於大詞彙連續語音辨識之研究

On the Use of Topic Models for Large-Vocabulary Continuous Speech Recognition

陳冠宇　Kuan-Yu Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
696470203@ntnu.edu.tw

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

摘要

本論文研究使用主題資訊之語言模型(Language Model)。當語言模型用於大詞彙連續語音辨識時，其主要的任務是藉由已解碼歷史詞序列資訊來預測下一個候選詞出現的可能性。傳統的 *N* 連(*N*-gram)語言模型容易受限於模型參數過多的問題，僅能用來擷取短距離的詞彙接連資訊，並不能考慮完整的歷史詞序列之語意資訊。因此，近十幾年來許多研究學者陸續提出各式主題模型(Topic Model)，包括討論文件與詞之關係的機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA)和潛藏狄利克里分配(Latent Dirichlet Allocation, LDA)，以及討論詞虛擬文件與詞關係的詞主題模型(Word Topic Model, WTM)。這些模型主要都是透過一組潛藏的主題機率分布來描述文件與詞、或者詞虛擬文件與詞之間的關係，用以擷取出歷史詞序列長距離的潛藏語意資訊。本論文提出一種新的主題模型，稱之爲詞相鄰模型(Word Vicinity Model, WVM)，它直接地基於語言中詞與詞相互關聯資訊以建構一個機率式的潛藏主題空間，並且透過線性模型結合的方式建立歷史詞序列之主題模型來預測下一個候選詞出現的可能性，藉此輔助傳統 *N* 連語言模型。實驗結果顯示本論文所提出的詞相鄰模型不僅相較大部分主題模型具有較低的模型參數量，同時能對於僅使用三連語言模型的基礎大詞彙連續語音辨識系統也有相當程度的語音辨識率提升。

關鍵詞：主題模型、機率式潛藏語意分析、潛藏狄利克里分配、詞主題模型、詞相鄰模型、大詞彙連續語音辨識。

一、緒論

語言是人與人之間最自然且有效率的溝通方式，不需透過其他的手勢或是動作，就可以讓對方了解我們想要表達的意思。正因爲如此，長久以來我們希望能讓機器聽懂人類的語言、直接與人類對話溝通，開啓了語音辨識的研究。在進行語音辨識時，我們以人類發聲的特性以及考量人耳聽覺感知爲基礎，將數位語音訊號轉換成易於電腦處理的聲學特徵向量(Acoustic Feature Vector)序列。接著，利用機率模型對於所收集到的訓練語音聲學特徵向量建立起聲學模型(Acoustic Model)藉此在測試階段比對測試語句聲之學特徵向量序列，判斷語句中所有可能的音素或詞段落。最後，使用語言模型(Language Model)來估測自然語言中每一個詞彙基於不同上下文之所可能出現的機率分布，用以解決聲學模型的混淆、限制辨識的搜尋空間和評估各個候選詞序列在自然語言中的合理性，因而輸出最有可能之候選詞序列。

當語言模型實際運用於語音辨識時，最主要的方式是從已解碼之歷史詞序列擷取短距離的詞彙接連資訊、或是長距離的語意資訊，據此預測下一個候選詞出現的可能性。在傳統統計式語言模型中，*N* 連(*N*-gram)語言模型[1]是最爲人所知且廣泛地運用於各種自然語言處理領域。*N* 連語言模型嘗試紀錄詞與詞之間同時出現的關係，估測每一個詞在其先前緊鄰 *N*-1 個詞已知的情況下出現的條件機率，並以多項式(Multinomial)分布表示之。但由於詞與詞序列有相當多種排列組合，致使 *N* 連語言模型的參數量相當可觀。*N* 連語言模型常因訓練語料的不足而限制其 *N* 值的大小(通常 *N* 設爲 2 或 3)，以致於它僅能用以計算短距離詞彙接連機率，而缺乏擷取出語句中(或候選詞與歷史詞序列間)所隱含長距離語意資訊的能力。爲了解決 *N* 連語言模型參數量龐大的問題，前人的研究認爲詞序列中每一個詞都有一個其隸屬的詞類別(Word Class)，隸屬於同一個詞類別的詞可能有具有相同的語法角色或相近的語意資訊，透過詞類別資訊可以將 *N* 連語言模型的參數量降低並保有適當的模型預測能力，因而有所謂的類別 *N* 連模型(Class-based *N*-gram Model)[2]。常見的類別 *N* 連模型將每一個詞對應到一個固定的詞類別，但因每一個詞實際上或許並非只有一種語意或是文法角色，所以亦有學者嘗試放寬詞與詞類別的對應，也就是讓一個詞可以隸屬於多個詞類別，爲此提出了聚合式馬可夫模型(Aggregate Markov Model, AMM)[3]。

不論是類別 *N* 連模型或是聚合式馬可夫模型的提出，皆是希望改善 *N* 連語言模型參數量過多的問題。另一方面，近十幾年來許多研究提倡探索在完整的文件或歷史詞序列中所隱含的語意資訊或是語句結構資訊等，以補足 *N* 語言連模型的不足[4, 5, 6]。其主要發展可追溯到早期使用潛藏式語意分析(Latent Semantic Analysis, LSA)的研究[7]，潛藏式語意分析利用線性代數的方法，將文件（或歷史詞序列）與詞投影至一個低維度空間，在這個低維空間中試圖描述文件與詞之間的關係，同時也可解決在高維度的情況下參數量過多和訓練語料量不足的問題。後來，更有所謂的機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA)[8, 9, 10]、潛藏狄利克里分配(Latent Dirichlet Allocation, LDA)[11, 12, 13, 14]及一些延伸方法陸續被提出。基本上，這些方法希望藉由機率模型的使用在低維度的語意空間中找出文件與詞的相關性。不同於潛藏語

意分析，機率式潛藏語意分析與潛藏狄利克里分配皆是機率式的生成模型，藉著對每一篇文件建立機率模型，直接表示文件與詞之間的關係，並且可以描述同義詞或者一詞多義的現象。新近，亦有所謂的詞主題模型(Word Topic Model, WTM)[15, 16]被提出。詞主題模型根據語言中每一個詞在訓練語料出現的資訊，將訓練語料作重新整理與安排，爲語言中每一個詞收集其對應的詞虛擬文件(Word Pseudo-document)，以訓練每一個詞專屬的機率生成模型，最後用來組成文件或歷史詞序列之機率生成模型以預測下一個候選詞出現的可能性。上述這些模型在本論文將統稱爲主題模型(Topic Models)[17]。

本論文提出一種新的主題模型，稱之爲詞相鄰模型(Word Vicinity Model, WVM)，它直接地基於語言中詞與詞的相互關聯資訊，建構出一個機率式的潛藏主題空間；並且透過線性模型結合的方式建立語音辨識中已解碼歷史詞序列之主題模型，用來預測下一個候選詞出現的可能性，藉此輔助傳統 N 連語言模型。本論文的安排如下：第二節將介紹近年來蓬勃發展的主題模型，包含機率式潛藏語意分析、潛藏狄利克里分配、詞主題模型；第三節將闡述本論文所提出之詞相鄰模型，並說明詞相鄰模型與現有各種主題模型之間的差異；第四節則是實驗結果與分析；第五節是結論。

二、主題語言模型相關研究

（一）機率式潛藏語意分析 (Probabilistic Latent Semantic Analysis, PLSA)

潛藏語意分析(Latent Semantic Analysis, LSA)假設文件集中文件與詞的組合存在若干潛藏語意結構成分[5]，藉由線性代數之奇異質分解(Singular Value Decomposition, SVD)可將高維度的文件向量與詞向量共同投影至一個低維度空間，其中每一維度代表某種語意結構成分；文件與詞之間語意的相似度可藉由它們在這個低維度空間的向量距離或者夾角的計算而得。如此一來，不僅可以簡化文件與詞表示方法的複雜度，也可以去除語料集中文件與詞的組合所含有的部分雜訊資訊。機率式潛藏語意分析(PLSA)是由潛藏語意分析延伸發展而來，不同於潛藏語意分析以線性代數的方法尋找語料集中隱含的主要語意結構成分，機率式潛藏語意分析利用機率模型爲每一篇文件建立生成模型，透過一組共享潛藏主題機率分布來描述每一篇文件生成文件中詞的關係[9]。

當機率式潛藏語意分析運用於語音辨識時，會將每一段歷史詞序列 H 視爲是一篇文件並估測其對應的機率式潛藏語意分析模型，用以計算在給定一段歷史詞序列 H 後下一個候選詞 w_i 出現的可能性，其機率式可表示成：

$$P_{\mathbf{PLSA}}(w_i \mid H) = \sum_{k=1}^{K} P(w_i \mid T_k) P(T_k \mid H) \tag{1}$$

其中 T_k 代表一個潛藏主題，具有某種語意結構成分； $P(w_i \mid T_k)$ 是給定潛藏主題 T_k 的情況下，候選詞 w_i 出現的機率； $P(T_k \mid H)$ 是歷史詞序列產生潛藏主題 T_k 的機率。在語音辨識時，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，可在執行語音辨識前就先以期望值最大化演算法(Expectation-Maximization Algorithm)[18]最大化訓練語料發生的機率而求得。另一方面，因歷史詞序列會隨著語音辨識的搜尋過程一直擴展、變動，所以歷史詞序列 H 產生每一個潛藏主題 T_k 的機率 $P(T_k \mid H)$ 需要不斷地被重新估算，同樣地也可以使用期望值最大化演算法來最大化歷

史詞序列發生的機率而得。機率式潛藏語意分析所擷取的長距離語意資訊可以彌補傳統 N 連語言模型在此的不足。在一般語音辨識的使用上，會將機率式潛藏語意分析與傳統 N 連語言模型經由線性插補法(Linear Interpolation)作結合，以提供在歷史詞序列 H 已解碼出的情況下每一個候選詞 w_i 發生的機率：

$$\widehat{P}(w_i \mid H) = (1 - \lambda_{\mathbf{PLSA}}) \cdot P_{N-\mathbf{gram}}(w_i \mid H) + \lambda_{\mathbf{PLSA}} \cdot P_{PLSA}(w_i \mid H) \qquad (2)$$

其中 $P_{N-\mathbf{gram}}(w_i \mid H)$ 就是傳統 N 連語言模型的機率分布，我們可以使用一個介於 0 到 1 之間的可調整參數 $\lambda_{\mathbf{PLSA}}$ 來控制 N 連語言模型與機率式潛藏語意分析模型的權重。

機率式潛藏語意分析的提出讓主題空間的概念得以由線性代數描述轉往機率式模型發展，但機率式潛藏語意分析本身仍然存在著許多問題：首先，它假設在給定某一個潛藏主題的前提下，文件與詞的關係是獨立的。由語意的觀點省視，這樣的假設過度強化了詞與整體文件（或者歷史詞序列）之間的獨立性。其次，隨著我們所收集到的訓練語料集中文件數的增加，機率式潛藏語意分析模型所需的參數也會呈線性增加，有可能會讓模型參數過度符合(Overfitting)訓練語料。一個理想的機率生成模型，對於描述未見過的(Unseen)文件中的詞應具備良好的預測能力。但事實上，機率式潛藏語意分析並沒有具備健全的預測能力，其主要原因在於它對於每一套訓練語料都會產生一組獨特的潛藏主題，並非使用一組全域性的參數描述所有語料。因此當用於估測一篇嶄新文件之主題機率模型時，會受到原始訓練語料的強烈限制。另外，在模型參數的估測過程，機率式潛藏語意分析使用期望值最大化演算法來逼近訓練語料的最大相似度。但以期望值最大化演算法來估測模型參數未必能找到全域最佳(Global Maximum)解，所以模型參數的起始值設定就變得格外重要。過去有研究學者也對訓練起始值提出不少研究討論，諸如多重隨機初始(Multiple Random Initialization)、預先使用非監督式分群(Unsupervised Clustering)或是利用傳統潛藏語意分析找出較好的模型起始值皆是常用的方法[7]，不過使用這些方法卻也會成爲模型訓練過程中一種額外的負擔。最後，當將機率式潛藏語意分析被應用於語音辨識時，需要不斷地使用期望值最大化演算法來對每一歷史詞序列估算其產生潛藏主題分布的機率，但這樣的估算過程事實上是相當耗費時間的，特別在潛藏主題數目龐大時，其所需的運算時間複雜度更是驚人。雖然有學者提出使用漸進式的期望值最大化演算法估測歷史詞序列產生潛藏主題分布的機率，但其能節省的運算時間有限，並且其結果相對地顯得較差。

（二）潛藏狄利克里分配 (Latent Dirichlet Allocation, LDA)

爲了改善機率式潛藏語意分析對於未見過的文件之預測能力以及模型參數量會隨著訓練語料中文件數量的增加而呈現線性成長的缺點，有學者提出了潛藏狄利克里分配[11]。潛藏狄利克里分配的模型詮釋方式與機率式潛藏語意分析不同，並且它可僅以兩組參數 α 與 β 來代表訓練語料的潛藏語意資訊，茲簡述如下。首先，假設訓練語料集 **D** 中共有 M 篇文件，而每一篇文件 d 中有 N_d 個詞，我們先由一組狄利克里分配 α 的參數求得每一篇文件 d 產生所有潛藏主題的機率向量 θ_d ，而文件中每一個詞在每一個潛藏主題 $T_{d,n}$ 下產生的機率分布則由 β 生成。潛藏狄利克里分配對參數的估算是最大化整個訓練語料 **D** 的邊際機率：

$$P_{\mathbf{LDA}}\left(\mathbf{D}\mid\alpha,\beta\right)=\prod_{d=1}^{M}\int P\left(\theta_d\mid\alpha\right)\left(\prod_{n=1}^{N_d}\sum_{T_{d,n}}P\left(T_{d,n}\mid\theta_d\right)P\left(w_{d,n}\mid T_{d,n},\beta\right)\right)d\theta_d \tag{3}$$

對於潛藏狄利克里分配參數的估算，前人提出不少方法諸如變動性貝氏期望值最大化(Variational Bayesian Expectation Maximization)演算法[7, 8]或吉卜森取樣(Gibbs Sampling)[9, 10]等。對於語音辨識中某一已解碼歷史詞序列 H（將 H 視爲一篇文件），潛藏狄利克里分配可以透過變動性貝氏期望值最大化演算法求得最佳參數解或是利用最大事後機率估測法估測其產生所有潛藏主題的機率向量 θ_H ，而當 H 長度足以表達語意資訊後，也可以對 α 進行重新估測。另一方面，若使用吉卜森取樣重估歷史詞序列產生所有潛藏主題的機率，則是結合訓練時對訓練語料集中詞的取樣資訊與對歷史詞序列中詞的重新取樣資訊，以期望逼近潛藏主題在歷史詞序列已知情況下的事後機率。事實上，不論是變動性貝氏期望值最大化法或是吉卜森取樣，在進行重估歷史詞序列的主題分布時都是非常耗費時間的。

（三）詞主題模型 (Word Topic Model, WTM)

不論是機率式潛藏語意分析或是潛藏狄利克里分配皆是希望擷取文件或歷史詞序列中隱含的長距離語意資訊，以彌補 N 連語言模型僅考慮短距離詞彙接連規則之不足。詞主題模型則是希望在建立語言模型時不僅考慮詞彙相鄰資訊，並且透過詞彙間潛藏語意資訊的組合，建立起文件或歷史詞序列之長距離語意資訊[16]。

詞主題模型的特色是透過一組共享的潛藏主題機率分布，爲語言中每一個詞 w_j 建立一個主題模型 $\mathbf{M}_{w_j}$ 。爲達此目的，在模型建立之前，必須從訓練語料中擷取每一個詞出現處其鄰近文字段落內其它詞出現的資訊，並將所有出現處的上下（或左右相鄰）文字段落聚集成每一個詞主題模型對應的訓練文件，稱之爲詞虛擬文件(Word Pseudo-document)。然後，透過一組共享的潛藏主題機率分布，估算每一個詞 w_j 之詞虛擬文件與其它詞 w_i 之共同出現關係；更明確些，即是 w_j 的詞主題模型 $\mathbf{M}_{w_j}$ 產生另一詞 w_i 的機率：

$$P_{\mathbf{WTM}}\left(w_i\mid\mathbf{M}_{w_j}\right)=\sum_{k=1}^{K}P\left(w_i\mid\mathbf{T}_k\right)P\left(T_k\mid\mathbf{M}_{w_j}\right) \tag{4}$$

其中 $P(w_i\mid T_k)$ 是給定潛藏主題 T_k 的情況下，詞 w_i 出現的機率； $P\left(T_k\mid\mathbf{M}_{w_j}\right)$ 是 w_j 的詞主題模型產生主題 T_k 的機率； K 則是潛藏主題總數。

當詞主題模型運用於語音辨識時，就如同機率式潛藏語意分析一般，我們將需要估算在給定了候選詞 w_i 的歷史詞序列 H 後， w_i 出現的機率。在此假設每一個潛藏主題產生候選詞 w_i 的機率 $P(w_i\mid T_k)$ 不隨語音辨識搜尋過程變動，並且每一個詞主題模型也已經由訓練語料求得最佳參數。因此，對於歷史詞序列 H，我們首先將它視爲由一連串的詞所組成的詞串，接著我們將詞串中每一個詞的詞主題模型利用線性插補法的方式結合，以此做爲歷史詞序列的主題模型[15]。

相較於機率式潛藏語意分析需使用期望值最大化演算法在語音辨識搜尋過程中不斷地估測歷史詞序列產生潛藏主題 T_k 的機率 $P(T_k\mid H)$ ，在使用詞主題模型時歷史詞序列

產生潛藏主題的機率 $P(T_k \mid H)$ 可由歷史詞序列中每一個詞 w_j 的詞主題模型產生主題 T_k 的機率 $P(T_k \mid \mathbf{M}_{w_j})$ 線性組合而成，此舉可大大地提升了語音辨識時的搜尋速度。

三、詞相鄰模型 (Word Vicinity Model, WVM)

（一）原理

與詞主題模型作法相似，本論文嘗試透過一組共享的潛藏主題分布，估算訓練語料集中相鄰詞彙間的語意關連性，稱之爲詞相鄰模型。不同於詞主題模型的是，詞相鄰模型直接對訓練語料中任意兩個詞 w_i 與 w_j 的聯合機率 $P(w_i, w_j)$ 透過一組潛藏主題分布所建構的語意空間作機率分解：

$$P_{\mathbf{WVM}}(w_i, w_j) = \sum_{k=1}^{K} P(w_i|T_k)P(T_k)P(w_j|T_k) \tag{5}$$

觀察式(5)與式(4)，我們可以發現詞相鄰模型包括了每一個潛藏主題的事前機率 $P(T_k)$，以及每一個潛藏主題產生每一個詞的機率分布 $P(w_i|T_k)$；而詞主題模型則是對詞彙間條件機率 $P_{\mathbf{WTM}}(w_i|\mathbf{M}_{w_j})$ 透過潛藏主題分布所建構的語意空間作機率分解，故有詞主題模型產生潛藏主題分布的機率 $P_{\mathbf{WTM}}(T_k|\mathbf{M}_{w_j})$ 以及每一個潛藏主題產生每一個詞的機率分布 $P_{\mathbf{WTM}}(w_i|T_k)$。相較之下，詞相鄰模型需要較少的模型參數量，在使用相同的訓練語料下，應會有較佳的模型參數估測表現。

當詞相鄰模型運用於語言模型的使用，諸如用於預測在給定詞 w_j 時另一詞 w_i 發生的可能性（亦即條件機率 $P(w_i \mid w_j)$），我們可以經過適當的機率式轉換，將此條件機率以詞相鄰模型的兩組機率分布 $P(T_k)$ 與 $P(w_i \mid T_k)$ 表示：

$$P_{\mathbf{WVM}}(w_i|w_j) = \frac{P(w_i, w_j)}{P(w_j)} = \frac{\sum_{k=1}^{K} P(w_i|T_k)P(T_k)P(w_j|T_k)}{\sum_{k=1}^{K} P(T_k)P(w_j|T_k)} \tag{6}$$

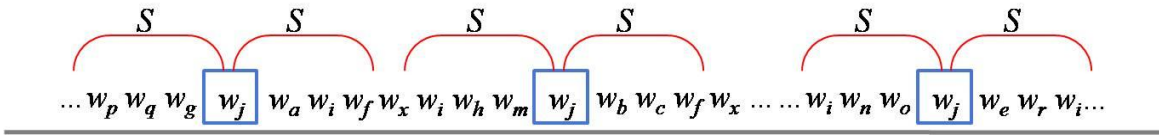
當詞相鄰模型用於語音辨識之語言模型使用時，對於每一個候選詞 w_i 以及其歷史詞序列 $H = w_1, w_2, \ldots, w_{i-1}$，欲估算 w_i 在給定 H 下出現的可能性時，可利用 H 中每一個詞產生 w_i 的條件機率（如式(6)）之線性組合來近似：

$$P(w_i \mid H) \approx \gamma_1 \cdot P_{\mathbf{WVM}}(w_i \mid w_1) + \gamma_2 \cdot P_{\mathbf{WVM}}(w_i \mid w_2) + \cdots + \gamma_{i-1} \cdot P_{\mathbf{WVM}}(w_i|w_{i-1}) \tag{7}$$

其中 $\{\gamma_1, \gamma_2, \ldots, \gamma_{i-1}\}$ 爲線性組合係數。再進一步來看，歷史詞序列 H 產生某一主題 T_k 的機率可經由歷史詞序列中所有詞在潛藏語意空間的分布特性而決定：

$$\begin{aligned} \tilde{P}(T_k \mid H) &= \sum_{j=1}^{i-1} \gamma_j \cdot P(T_k \mid w_j) \\ &= \sum_{j=1}^{i-1} \gamma_j \cdot \frac{P(w_j \mid T_k)P(T_k)}{\sum_{k=1}^{K} P(w_j \mid T_k)P(T_k)} \end{aligned} \tag{8}$$

其中， $P(T_k)$ 與 $P(w_i \mid T_k)$ 爲詞相鄰模型所求得之模型機率分布。因此，當我們將詞相鄰模型用於語音辨識之語言模型的使用時，亦可以如同機率式潛藏語意分析及潛藏狄利克里



圖一、詞相鄰模型訓練框

分配般的方式來表示歷史詞序列與候選詞間長距離的語意相關性：

$$P_{\mathbf{WVM}}(w_i \mid H) = \sum_{k=1}^{K} P(w_i \mid T_k)\tilde{P}(T_k \mid H) \tag{9}$$

值得注意的是，詞相鄰模型對於式(9)其實是透過歷史詞序列中每一個詞與候選詞 w_i 兩兩間在潛藏語意空間上的機率分布關係而計算出，這一點與詞主題模型相似；而機率式潛藏語意分析及潛藏狄利克里分配是將歷史詞序列視爲一整體，計算其與候選詞在潛藏語意空間上的機率分布關係。

在詞相鄰模型的訓練方面，爲了估算語料庫中任意兩個詞 w_i 與 w_j 共同出現的聯合機率 $P(w_i, w_j)$，我們首先須決定一個訓練框 S，用來固定選取每一個詞 w_i 出現時上下（左右相鄰）文段中有哪些的詞彙出現，並估算它們個別與詞 w_i 在訓練語料中會共同出現在此訓練框的次數，以 $n(w_i, w_j)$ 表示，而會有 $n(w_i, w_j) = n(w_j, w_i)$ 的性質。再者，我們假設在討論任意兩兩詞彙間共同出現的關係時，不受到其它詞彙或其它詞彙之間的關係的影響。因此，詞相鄰模型的訓練是以最大化詞典 $\mathbf{V}$ 中任意兩個詞 w_i 與 w_j 在訓練語料共同出現在一定範圍上下文段（或訓練框）的聯合機率 $P_{\mathbf{WVM}}(w_i, w_j)$（參見式(5)）之對數機率值總和 $L_{\mathbf{WVM}}$ 爲目標：

$$L_{\mathbf{WVM}} = \sum_{w_i, w_j \in \mathbf{V}} n(w_i, w_j) \log P_{\mathbf{WVM}}(w_i, w_j) \tag{10}$$

我們藉著使用期望值最大化演算法來進行其中詞相鄰模型機率式之估測。

（二）其他主題模型之比較

在此，我們由圖形模型(Graphical Model)表示、模型參數量多寡、以及於語音辨識時之執行效能等幾個觀點分析與比較各種主題模型之間的關係與優劣，如表一所歸納。

1. 圖形模型表示

首先，藉由圖形模型表示機率式潛藏語意分析(PLSA)。如圖二(a)所示，我們可觀察出機率式潛藏語意分析是先考慮每一篇文件生成每一潛藏主題的機率，接著聯合一組潛藏主題分布分別產生每一個詞的機率，使每一篇文件成爲一個具有預測能力的生成模型。另一方面，詞主題模型(WTM)則收集每一個詞在語料庫中出現位置鄰近處的詞合成對應的詞虛擬文件，考慮每一個詞與詞虛擬文件之間的關係，其圖形模型表示如圖二(b)所示。詞主題模型的圖形模型表示與機率式潛藏語意分析之圖形模型表示非常相似，主要差別在於機率式潛藏語意分析以訓練語料庫中每一篇文件爲模型單位，而詞主題模型則爲語言中每一個詞重新整理其在訓練語料出現資訊而有所謂的詞虛擬文件來作爲模型單位。

表一、各主題模型之比較

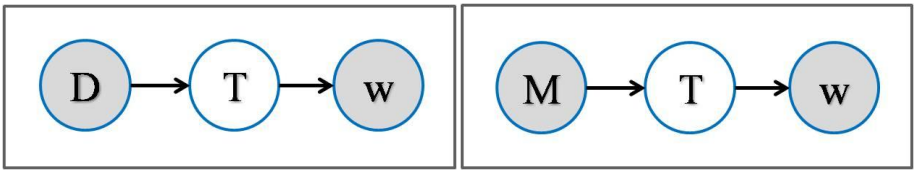
模型	機率式潛藏語意分析	潛藏狄利克里分配	詞主題模型	詞相鄰模型
模型對象	文件與詞	文件與詞	虛擬詞文件與詞	詞與詞
模型參數量	$N \times K + M \times K$	$K + N \times K$	$2 \times N \times K$	$K + N \times K$
用於語音辨識之方式	即時重新估算歷史詞序列之主題分布	即時重新估算歷史詞序列之主題分布	詞主題模型線性結合	機率模型線性結合
速度	中等	慢	快	快

再者，機率式潛藏語意分析在訓練語言模型時，將每一篇文件的語意資訊考慮進模型參數，求取參數的過程中希望最大化每篇文件產生詞的機率，過去研究人員認爲這樣的訓練過程會使得模型參數估測受到訓練語料中文件的限制。如用於未見過的文件其主題分布能符合訓練語料的特質，其模型的預測能力將會有不錯效果；但若用於預測主題偏差較大的未知文件，則可能就無法得到良好的效果。有別於將文件語意資訊直接考慮於模型參數，潛藏狄利克里分配(LDA)的文件主題生成方式僅以兩組參數（α 與 β ）描述，如此可以避免潛藏主題機率參數過度符合訓練語料所收集到的文件，讓生成模型對於預測新文件時更有彈性，潛藏狄利克里分配可表示成一個三層架構的之圖形模型，如圖二(c)所示，其中 α 與 β 是潛藏狄利克里分布模型之所擁有的兩組參數。

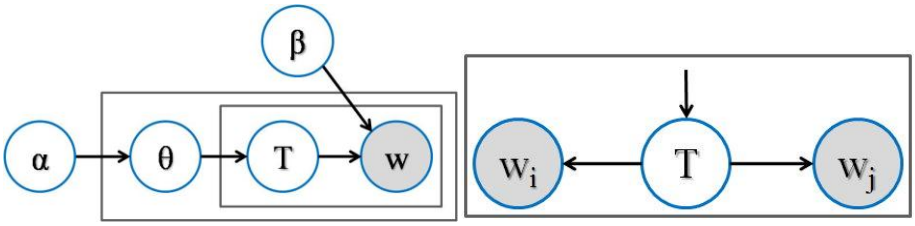
有別於將文件層次語意資訊考慮於模型訓練之中，我們希望模型可以不受文件過度束縛。透過不同層次（詞層次）潛藏主題模型的引入，讓文件生成潛藏主題的過程更具彈性，但卻不至於讓模型演繹過程過度複雜。因此，在本論文我們提出了詞相鄰模型，直接透過估算訓練語料中相鄰詞彙間的語意關連性，建立一組詞彙間共享的主題語意空間，描繪詞與詞之間的相互關係（如圖二(d)所示）。當選定一個潛藏主題後，詞相鄰模型提供一個描述在這個主題下每一個詞出現的可能性的機率分布，而每一個潛藏主題本身有其發生的事前機率。

2. 模型參數量分析

在給定詞典 $\mathbf{V}$（共 N 個詞，$\mathbf{V} = \{w_1, w_2, \ldots, w_N\}$）、訓練語料 $\mathbf{D}$（共 M 篇文件詞，$\mathbf{D} = \{d_1, d_2, \ldots, d_M\}$）和假設潛藏主題數爲 K 的情況下，我們比較各主題模型的模型複雜度。機率式潛藏語意分析有每篇訓練文件產生每一個潛藏主題的機率 $P(T_k|d_m)$ 以及潛藏主題產生詞的機率 $P(w_n|T_k)$，共需 $N \times K + M \times K$ 個參數；詞主題模型則擁有每個詞主題模型產生潛藏主題的機率分布 $P(T_k|\mathbf{M}_{w_n})$ 以及每一個潛藏主題產生每一個詞的機率分布 $P(w_n|T_k)$，共需 $2 \times N \times K$ 個參數；潛藏狄利克里分配則僅需要兩組參數 α 與 β 共 $K + N \times K$ 個參數；最後詞相鄰模型亦僅需 $K + N \times K$ 個參數，分別是每一個潛藏主題的機率 $P(T_k)$，以及每一個潛藏主題產生每一個詞的機率 $P(w_n|T_k)$。當訓練語料中所含的文件數小於詞典大小($M < N$)時，詞主題模型是四個主題模型中參數量最多的，但是若隨著收集的訓練語料越來越多機率式潛藏語意分析的參數量會呈現線性增加，而詞主題模型的參數量是固定的，所以當收集的訓練語料所含的文件數大過詞典大小($N < M$)時，機率式潛藏語意分析會需要最多的參數。潛藏狄利克里分配與詞相鄰模型則不論訓練語料大小，所需的參數量僅和潛藏主題個數與詞典大小有關。



(a) 機率式潛藏語意分析(PLSA)　　(b) 詞主題模型(WTM)



(c) 潛藏狄利克里分配(LDA)　　(d) 詞相鄰模型(WVM)

圖二、主題模型之圖形模型表示

3. 於大詞彙連續語音辨識運用之分析

「即時性」是當今語音辨識技術能否被廣爲使用的關鍵因素，本論文因此對於上述主題模型運用於語音辨識時之執行效能作概略分析。當機率式潛藏語意分析(PLSA)運用於語音辨識時，它最爲人所詬病的是使用期望值最大化法線上估測歷史詞序列的潛藏主題分布；雖然即時估測可以針對歷史詞序列重新計算獲得相對準確的主題分布，但這樣的過程實在過於耗費時間。對照於機率式潛藏語意分析，詞主題模型(WTM)使用線性組合的方式，直接將已解碼的歷史詞序列中每一個詞的詞主題模型線性結合，以此作爲歷史詞序列的主題分布。雖然使用詞主題模型所得到的歷史詞序列主題分布，如同機率式潛藏語意分析一樣，會受到訓練語料集的限制。但實際運用於語音辨識時，詞主題模型可以省去像機率式潛藏語意分析所需耗時的線上主題分布重估，具即時性之優點。另一方面，潛藏狄利克里分配(LDA)雖僅用少量的參數描述訓練語料集之主題分布特性，當直接對觀測到的歷史詞序列重估主題分布，亦可較不受訓練語料庫中文件的限制[11]。但當潛藏狄利克里分配被使用於語音辨識時，一樣遭受重估過於耗時的問題。就我們將語言模型使用於語音辨識實驗所作觀察，潛藏狄利克里分配是四者中最耗時的模型。最後，當詞相鄰模型(WVM)運用於語音辨識時，我們將歷史詞序列視爲由許多詞所組成的詞串，透過適當的機率轉換計算出在給定歷史詞序列中每一個詞下任一個候選詞出現的可能性，再如同詞主題模型以線性插補法的方式結合這些條件機率，以此做爲歷史詞序列的主題模型（參見式(7)）。其過程中雖然需透過一次的機率式轉換（參見式(6)），但是當實際運用於語音辨識時，詞相鄰模型在與機率式潛藏語意分析和潛藏狄利克里分配相較下，仍然擁有較佳的執行速度。

四、實驗結果與分析

（一）實驗設定

在語音特徵擷取部分，我們以梅爾率波器組(Mel-frequency Filter Bank)輸出爲基礎，使用異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)配合最大

化相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)，最後獲得 39 維語音特徵向量。另外，在辨識所需的聲學模型訓練上，考慮了中文語音結構，聲學模型由 22 個 INITIAL 模型、38 個 FINAL 模型（每個中文的音節都是由一個 INITIAL 及一個 FINAL 所組成）及一個靜音(Silence)模型組成，其中 INITIAL 模型會因其右邊可能接的 FINAL 模型種類而進一步細分成 112 個 INITIAL 模型[19]。我們最後總共使用了 151 個隱藏式馬可夫模型(Hidden Markov Models)來作爲這些 INITIAL-FINAL 聲學模型的統計模型。在隱藏式馬可夫模型中，每個狀態則依據其對應到的訓練語料多寡，以 2 到 128 個高斯統計分布來表示，不管男女性別都使用同一套聲學模型。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訓練(Minimum Phone Error, MPE)以期望獲得最佳化聲學模型參數[20]。

本論文實驗使用的訓練與測試語料爲 MATBN 電視新聞語料庫[21]，是由中央研究院資訊所口語小組耗時三年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爲實驗題材，將其中約 25 小時收錄於 2001 年 11 月至 2002 年 12 月期間的語料作爲聲學模型訓練語料，再由 2003 年的收錄語料中定義各約 1.5 小時做爲發展集語料（MATBN 發展集）以及測試集語料（MATBN 測試集），詳細資料集資訊如表二所示。更明確地，我們將由 MATBN 發展集中選定最佳模型參數，並將此參數運用於測試集語料，比較與討論各種主題模型的效能。

另一方面，背景三連語言模型(Trigram Language Model)訓練語料則是來自中央通訊社 2001 年至 2002 年的文字新聞語料，包含了約一億五千萬個中文字，經斷詞後約有八千萬詞。本論文實驗爲語言模型調適，我們由公視廣播新聞語料 2001、2002 與 2003 年的人工轉寫文件中篩選出約三千六百篇報導，約兩百萬個中文字，經斷詞後約有一百萬詞，作爲調適語料。詞典大小約爲七萬兩千詞。採用 SRI Language Modeling Toolkit[22] 訓練實驗所需要的三連語言模型。

論文將主題模型用於調適背景三連語言模型，其方式爲模型插補法。如式(2)所示，調整主題模型與背景三連語言模型影響的權重參數是先由 MATBN 發展集調整至最佳後，再用於 MATBN 測試集。語言模型效能的評估，是透過語言複雜度(Perplexity, PP)以及大詞彙連續語音辨識之辨識字錯誤率(Character Error Rate, CER)來達成。

實驗中我們將詞相鄰模型的訓練框 s 設定爲 2；另外，詞相鄰模型與詞主題模型需要給歷史詞序列中每一個詞的主題模型一個語言模型影響權重 γ_j，我們利用詞與詞之間的距離定義一個指數遞減函數，用來給定每一個詞的主題模型一個語言模型影響權重：

$$\gamma_j = \phi_j \prod_{s=j+1}^{i-1} (1-\phi_s) \tag{11}$$

當 $j = 2,\ldots,i-1$ 時 ϕ_j 是一個介於 0 到 1 的定值，而 ϕ_j 爲 1，並且此遞減函數也會滿足 $\sum_{j=1}^{i-1} \gamma_j = 1$。實驗中，我們將 ϕ_j 設爲 0.6。

再者，如同機率式潛藏語意分析一般，詞相鄰模型與詞主題模型亦可以假設潛藏主題產生每一個詞的機率不隨辨識過程變動，利用期望值最大化演算法調整歷史詞序列的主題機率分布，即調整式(7)或(8)中每一個歷史詞序列中詞的線性組合係數（或語言模

表二、資料集

	MATBN 發展集	MATBN 測試集	NOWnews 測試集
總句數	292	307	13,810
總詞數	16,106	16,494	1,075,409

表三、由詞相鄰模型(32 Topics)中選取出 4 個主題

Topic 8	
詞(word)	權重(weight)
主委陳菊	0.792
發布新聞稿	0.750
總召柯建銘	0.630
副總裁陳師孟	0.625
宜蘭縣長	0.564
副院長賴英照	0.550
立法院黨團	0.519
機要	0.495
中央研究院院長	0.489
聯邦準備理事會	0.469

Topic 13	
詞(word)	權重(weight)
靜脈	1.202
顯微	1.002
切除	0.674
肌瘤	0.668
腦炎	0.618
子宮	0.501
支氣管	0.500
縫合	0.463
割除	0.367
氣管	0.344

Topic 14	
詞(word)	權重(weight)
平均地權	1.306
公職人員財產申報	1.259
土地稅	0.704
菸酒稅法	0.489
財稅	0.457
修正草案	0.446
財政收支劃分	0.428
購併	0.396
暫行條例	0.383
保險法	0.373

Topic 23	
詞(word)	權重(weight)
霍亂	0.752
大腸直腸癌	0.681
沙門氏菌	0.471
口蹄疫	0.337
甲狀腺	0.303
胃癌	0.298
徵狀	0.269
寄生	0.268
皮膚癌	0.267
肺癌	0.234

表四、基礎實驗結果

baseline	MATBN 發展集		MATBN 測試集		NOWnews 測試集	
	CER(%)	PP	CER(%)	PP	CER(%)	PP
Trigram	20.22	667.23	20.08	682.10	null	808.76

型影響權重）。我們將利用指數遞減函數（式(11)）估測歷史詞序列之主題分布的方式以(ED)表示，而期望值最大化演算法估測的方式以(ML)表示之。

（二）實驗結果與分析

首先，我們由 32 個主題數的詞相鄰模型中取出 4 個潛藏主題，並且計算每個詞分別屬於不同潛藏主題時的主題分數(Topic Score)。其中某一個詞 w_i 隸屬於某一個潛藏主題 T_k 時的主題分數定義如下[23]：

$$TS(w_i, T_k) \equiv \frac{\sum_{m=1}^{M} c(w_i, d_m) P(T_k \mid d_m)}{\sum_{m=1}^{M} c(w_i, d_m)(1 - P(T_k \mid d_m))} \tag{12}$$

其中 $c(w_i, d_m)$ 爲詞 w_i 出現在文件 d_m 的次數， $P(T_k \mid d_m)$ 爲文件 d_m 詞產生潛藏主題 T_k 的機率。對於這 4 個潛藏主題，我們分別挑選出主題分數較大的 10 個詞彙，如表三所示。我們可以發現 Topic 8 傾向政黨政治新聞，Topic 13 收集醫學和醫療等相關資訊，Topic 14 關於政府稅收的主題資訊，最後 Topic 23 則是把疾病和病毒等名稱聚集在一起。由此實驗可知，詞相鄰模型本身對於訓練語料亦具備良好的非監督式分群之能力。

接著，我們比較各主題模型之語言複雜度(Perplexity, PP)。語言複雜度最早是由資訊理論發展而來，用來評估一個語言模型的好壞，其幾何意義爲語言模型產生一段文字的機率倒數再取幾何平均數，可視爲語言模型預測詞與詞接連的平均分支度。語言複雜度越小，表示所訓練的語言模型越具有預測詞產生的能力。如表四所示，背景三連語言模型在 MATBN 發展集所得的語言複雜度爲 667.23，而在 MATBN 測試集的語言複雜度爲 682.10。表五是當各種主題模型與背景三連語言模型結合後，作用於 MATBN 測試集的語言複雜度實驗結果；我們可以發現，各主題模型隨著主題數陸續增加語言複雜度也隨之降低。我們亦可由表五觀察到，在不同主題數設定時，不論是詞相鄰模型或是詞主題模型，其語言複雜度表現大都較潛藏狄利克里分配佳（有較低的語言複雜度值），亦較機率式潛藏語意分析好。另外，詞相鄰模型使用期望值最大化演算法估測歷史詞序列中每一個詞的主題模型之語言模型影響權重的方式（即 WVM(ML)）較使用指數遞減函數的方式（即 WTM(ED)，參見式(11)）有較低的語言複雜度值。但若對於詞主題模型而言，使用指數遞減函數的方式估測歷史詞序列中每一個詞的主題模型之語言模型影響權重（即 WTM(ED)）會較使用期望值最大化演算法（即 WTM(ML)）爲佳。另一方面，若比較詞相鄰模型與詞主題模型時，則可發現當潛藏主題數較小時詞相鄰模型(WVM(ML))有最低的語言複雜度，但隨著主題數漸漸增加詞主題模型(WTM(ED))的語言複雜度會快速下降。

當將上述這些主題模型與背景三連語言模型結合時，均能較僅使用背景三連語言模型時有明顯的語言複雜度降低。以最佳實驗設定而言，機率式潛藏語意分析有 23.1%、潛藏狄利克里分配有 21.4%、詞主題模型(WTM(ED))有 26.1%、詞相鄰模型(WVM(ML))有 24.2%的相對語言複雜度降低。

再者，我們比較各種主題模型與背景三連語言模型結合後，運用於大詞彙連續語音辨識時的辨識字錯誤率。如表四所示，在基礎實驗中，MATBN 發展集的字錯誤率爲 20.22%，MATBN 測試集的字錯誤率爲 20.08%。表五展現各主題模型與背景三連語言模型結合後，在不同潛藏主題數設定下的辨識詞錯誤之實驗結果。我們可以觀察到，當主題數設定爲 32 或 64 時各種主題模型可以分別獲得最佳辨識結果。以使用各種主題模型的最低字錯誤率來說，機率式潛藏語意分析有 4.1%、潛藏狄利克里分配有 5.2%、詞主題模型(WTM(ML))有 3.9%、詞主題模型(WTM(ED))有 5.5%、詞相鄰模型(WVM(ML))有 4.0%、詞相鄰模型(WVM(ED))有 5.0%的相對字錯誤率降低。我們發現詞相鄰模型與詞主題模型以線性結合的方式來估算歷史詞序列主題分布的方法運用於大詞彙連續語音辨識時皆可獲得不錯的實驗結果，當潛藏主題數設定爲 64 時詞主題模型(WTM(ED))

表五、各主題模型於 MATBN 測試集之實驗結果

MATBN 測試集	PLSA		LDA		WTM(ML)		WTM(ED)		WVM(ML)		WVM(ED)	
	CER(%)	PP	CER(%)	PP	CER(%)	PP	CER(%)	PP	CER(%)	PP	CER(%)	PP
8 topics	19.26	553.92	19.25	557.02	19.50	539.87	19.17	542.65	19.48	531.63	19.31	546.11
16 topics	19.40	547.03	19.11	550.35	19.44	529.73	19.19	533.41	19.49	528.60	19.18	540.45
32 topics	19.26	535.93	19.06	539.21	19.45	526.61	19.14	521.27	19.27	523.11	19.15	537.17
64 topics	19.29	530.85	19.24	537.03	19.29	523.88	18.98	509.18	19.37	516.75	19.22	530.73
128 topics	19.34	524.60	19.14	536.10	19.39	528.38	19.13	503.74	19.36	519.33	19.23	527.94

表六、不同歷史詞長度(*L*)之詞相鄰模型於 MATBN 測試集之實驗結果(CER(%))

L	1	2	4	8	16
8 topics	19.32	19.25	19.23	19.31	19.31
16 topics	19.24	19.17	19.18	19.18	19.18
32 topics	19.29	19.21	19.13	19.15	19.15
64 topics	19.39	19.28	19.21	19.20	19.22
128 topics	19.44	19.29	19.22	19.23	19.23

更可獲得最低的辨識字錯誤率。

進一步地，我們針對詞相鄰模型進行探討，當詞相鄰模型以線性結合的方式估算歷史詞序列主題分布的方法使用於語音辨識時，我們可以假設一個候選詞的出現僅與前 *L* 個詞有關，用以簡化計算複雜度。表六爲假設候選詞出現的可能性僅與前 *L* 個詞相關時的實驗結果；結果顯示出，不同的 *L* 值皆大約在將主題數設爲 32 或 64 時有最低的字錯誤率。而最低的字錯誤率是設定潛藏主題數共 32 個並且假設預測詞的出現僅與前 4 個詞相關的時候，其字錯誤率約爲 19.13%。值得注意的是，當我們比較 *L*（歷史詞長度）爲 8 與 16 時，辨識字錯誤率僅當潛藏主題數設爲 64 時相差 0.02%，當比較完整的歷史詞序列與僅考慮前 16 個歷史詞時，在不同的潛藏主題數下皆獲得相同的辨識字錯誤率。我們在此推論，詞相鄰模型主要描述訓練語料中任意兩個鄰近詞的共同出現情況，所以距離候選詞較遠的詞僅能扮演輔助的角色，無法對候選詞可能出現與否有決定性的影響。

最後，我們比較各主題模型於同時期(Contemporary)測試文字語料集的語言複雜度。於是，我們收集了與 MATBN 測試集時期相近的今日新聞(NOWnews)文字新聞語料，由 2003 年 1 月至 4 月的新聞中挑選出共約五千六百多則新聞，包含約一萬三千則句子，以此做爲同時期的測試集語料（NOWnews 測試集）。其基礎語言複雜度爲 808.76，詳細語料資訊列於表二。在此，我們將 MATBN 發展集中調定的最佳參數運用於此測試集語料中。表七是各主題模型與背景三連語言模型結合後於 NOWnews 測試集的語言複雜

表七、各主題模型於 NOWnews 測試集之實驗結果(PP)

NOWnews 測試集	PLSA	LDA	WTM(ML)	WTM(ED)	WVM(ML)	WVM(ED)
8 topics	751.89	747.60	702.16	766.72	694.57	766.06
16 topics	742.71	738.61	692.20	762.50	690.77	765.59
32 topics	732.80	732.35	684.71	760.43	682.63	765.79
64 topics	726.17	731.58	679.17	760.62	674.04	761.42
128 topics	716.97	728.29	672.91	762.41	667.99	761.66

度實驗結果。雖說 NOWnews 測試集與語言模型調適語料屬同一時期之新聞語料，但是 NOWnews 測試集是文字新聞，而語言模型調適語料是廣播新聞語音轉寫文字，兩者在主題上有相當的相近度，但在語句或詞彙的使用上會不相同。實驗結果顯示，詞相鄰模型(WVM(ML))與詞主題模型(WTM(ML))對於這些文字新聞的主題預測能力優於機率式潛藏語意分析和潛藏狄利克里分配。相較於機率式潛藏語意分析和潛藏狄利克里分配，詞相鄰模型與詞主題模型希望在建立語言模型時不僅考慮文件或歷史詞序列之長距離語意資訊並且保有鄰近詞的詞彙資訊。故於此實驗中，雖然機率式潛藏語意分析也是以期望值最大化演算法來估測歷史詞序列的主題分布，但語言複雜度仍然高於使用相同方法估測歷史詞序列的詞相鄰模型(WVM(ML))與詞主題模型(WTM(ML))。在最佳的實驗設定下，詞相鄰模型(WVM(ML))可以獲得 17.4%的相對語言複雜度降低，而詞主題模型(WTM(ML))、機率式潛藏語意分析潛藏狄利克里分配分別有有 16.8%、11.3%與 9.9%的相對語言複雜度降低。

五、結論與未來展望

本論文提出一個嶄新的觀點－基於語言中詞與詞的關聯資訊來建構一個潛藏主題空間。我們嘗試於此空間中討論文件與詞之間的關係，提出詞相鄰模型(Word Vicinity Model, WVM)。本論文中討論詞相鄰模型之語言複雜度(Perplexity)，以及運用於大詞彙連續語音辨識時之辨識字錯誤率(Character Error Rate)。結果顯示詞相鄰模型相較於大部分主題模型擁有較低的語言複雜度；當運用於大詞彙連續語音辨識時，詞相鄰模型相較於基礎辨識率亦有 5.0%的相對進步率。未來，我們將研究詞相鄰模型運於不同領域的可用性，以及強化詞相鄰模型運用於大詞彙連續語音辨識時對於歷史詞序列的主題模型估測。

六、致謝

本研究承蒙國科會研究計畫 NSC 98-2221-E-003-011-MY3 、NSC96-2628-E-003-015-MY3、NSC97-2631-S-003-003 的部分補助，僅此致謝。

參考文獻

[1] F. Jelinek, "Up from trigrams! - the struggle for improved language models," in *Proc. of Eurospeech*, 1991.
[2] PF. Brown, VJ. Della Pietra, PV. deSouza, JC. Lai, and RL. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, 18(4):467–479,

December, 1992
[3] L. Saul and F. Pereira, "Aggregate and mixed-order Markov models for statistical language processing," in *Proc. of EMNLP*, 1997.
[4] R. Iyer and M. Ostendorf, "Modeling long distance dependence in language: topic mixtures versus dynamic cache models," Speech and Audio Processing, IEEE Transactions, 1999.
[5] J. Bellegarda, "Statistical language model adaptation: review and perspectives," in *Speech Communication*, 2004.
[6] R. Rosenfeld , "Two decades of Statistical Language Modeling: Where Do We Go From Here?", in *Proc. of the IEEE*, 2000.
[7] J.R. Bellegarda, *Latent Semantic Mapping: Principles and Applications.* Morgan and Claypool, 2007.
[8] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. of* SIGIR, 1999.
[9] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, 2001.
[10] D. Gildea and T. Hofmann, "Topic-based language models using EM," in *Proc. of Eurospeech*, 1999.
[11] D.M. Blei, A.Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, 2003.
[12] Y. Tam and T. Schultz, "Dynamic language model adaptation using variational Bayes inference," in *Proc. of Interspeech*, 2005.
[13] T. Griffiths, "Gibbs sampling in the generative model of Latent Dirichlet Allocation," Technical Report.
[14] T. Griffiths and M. Steyvers, "Finding scientific topics," in *Proc. of the National Academy of Sciences*, 2004.
[15] H.-S. Chiu and B. Chen, "Word topical mixture models for dynamic language model adaptation," in *Proc. of ICASSP*, 2007.
[16] B. Chen, "Latent topic modeling of word co-occurrence information for spoken document retrieval," in *Proc. of ICASSP*, 2009.
[17] M. Steyvers and T. Griffiths, "Probabilistic topic models." In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis.* Hillsdale, NJ: Erlbaum.
[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society*, 1977.
[19] B. Chen, J.-W. Kuo, and W.-H. Tsai, "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription," in *Proc. of ICASSP*, 2004.
[20] S.H. Liu, F.H. Chu, and B. Chen, "Improved MPE-Based Discriminative Training of Acoustic Models for Mandarin Large Vocabulary Continuous Speech Recognition," in *Proc.* of *ROCLING*, 2007. (in Chinese)
[21] 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*, 2005.
[22] A. Stolcke, SRI Language Modeling Toolkit, http://www.speech.sri.com/projects/srilm/
[23] T.-H. Li, M.-H. Lee, B. Chen and L.-S. Lee, "Hierarchical Topic Organization and Visual Presentation of Spoken Documents Using Probabilistic Latent Semantic Analysis (PLSA) for Efficient Retrieval/Browsing Applications," in *Proc. of Eurospeech*, 2005.