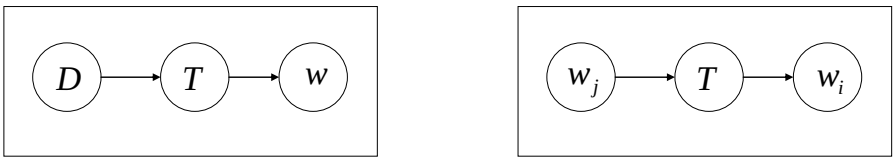
第4章 詞主題混合模型與位置相關語言模型

本章主要提出以詞為模型單位的詞主題混合模型(Word Topical Mixture Model, WTMM)，包括其模型特徵、詳細訓練方式，並與其他模型，如機率式潛藏語意分析與觸發對語言模型作模型的分析與比較[Chiu and Chen 2007]。我們希望透過詞主題混合模型，同時使用詞主題資訊以及長距離的詞關聯。除了詞主題資訊外，我們亦提出了詞位置資訊。我們認為，詞在文件或語句出現的位置，亦是可用的資訊。我們嘗試將詞位置資訊與*N*連詞模型的詞彙資訊與潛藏語意分析模型的語意資訊作整合，建立位置相關語言模型(Position-Dependent Language Model)，如位置性*N*連模型(Positional *N*-gram Model)以及位置性機率式潛藏語意分析(Positional PLSA)等，並與混合主題式語言模型及潛藏語意分析作比較。

4.1 詞主題混合模型(Word Topical Mixture Model)

4.1.1 詞主題混合模型

傳統的 N 連詞模型只能捕捉到短距離的詞彙資訊，對於辨識過程複雜的大詞彙連續語音辨識系統已經不敷使用，所以近年來有許多模型及方法被提出，嘗試補充*N*連詞模型不足的地方，例如機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA)模型及觸發對語言模型(Trigger-based Language Model)。機率式潛藏語意分析使用文件中隱藏的主題資訊，而觸發對語言模型嘗試放寬詞與詞的距離限制。機率式潛藏語意分析透過機率式架構，並使用期望最大化法估測隱藏的主題機率分布；觸發對語言模型直接建立歷史詞及辨識詞的關係，所以能夠捕捉長距離的關聯性且具有直接的詞預測能力。根據機率式隱藏主題以及長距離之詞關聯性這兩種特性，我們提出了詞主題混合模型(Word Topical Mixture Model, WTMM)，希望透過此模型，能夠結合這兩種特性。



(a) (b)

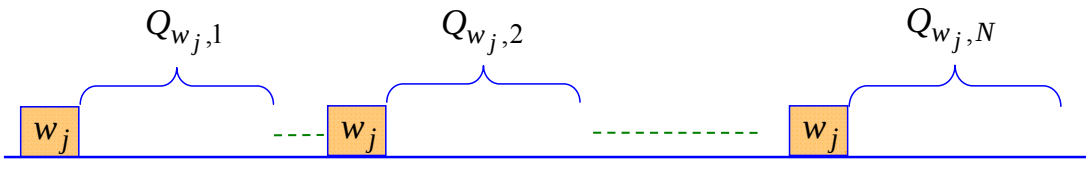
圖 4-1 圖形模型 (a) 機率式潛藏語意分析 (b) 詞主題混合模型

詞主題混合模型定義為：

$$P(w_i \mid M_{w_j}) = \sum_{k=1}^{K} P(w_i \mid T_k) P(T_k \mid M_{w_j}) \tag{4-1}$$

M_{w_j} 代表詞 w_j 的模型，$P(w_i \mid T_k)$ 是主題 T_k 產生詞 w_i 的機率，$P(T_k \mid M_{w_j})$ 是模型 M_{w_j} 產生主題 T_k 的機率，K 是可能的主題數。詞主題混合模型的模型架構類似於機率式潛藏語意分析，差別在於機率式潛藏語意分析是以訓練語料的每一文件為模型單位，詞主題混合模型則是以詞典中的每一詞為單位，兩者的圖形模型差異如 圖 4-1 所示，D 代表文件，T 代表隱藏主題，w 代表詞。

我們如何訓練模型呢？首先我們需要得到詞模型 M_{w_j} 的訓練觀測值 Q_{w_j}。假設對於某個詞 w_j 而言，在固定的窗長度 W 下，跟隨在詞 w_j 的第一個詞序列 $Q_{w_j,1}$，第二個詞序列 $Q_{w_j,2}$，一直到第 N 個詞序列 $Q_{w_j,N}$，是與詞 w_j 相關的，所以我們可以收集這詞序列，然後串連起來，形成詞模型 M_{w_j} 的訓練觀測值 Q_{w_j}，



$$Q_{w_j} = Q_{w_j,1}, Q_{w_j,2}, \cdots, Q_{w_j,N}$$

圖 4-2 詞主題混合模型訓練圖

如 圖 4-2 所示。然後我們可以針對所有的詞模型的對應訓練觀測值進行相似度最大化法訓練：

$$\begin{aligned}\log L_{\mathrm{Q}_{TrainSet}} &= \sum_{Q_{wj}} \log P\left(Q_{w_j} \mid M_{w_j}\right) \\ &= \sum_{Q_{wj}} \sum_{w_n \in Q_{wj}} n\left(w_n, Q_{w_j}\right) \log P\left(w_n \mid M_{w_j}\right)\end{aligned} \tag{4-2}$$

$n\left(w_n, Q_{w_j}\right)$表示詞 w_n 出現在訓練觀測值 Q_{w_j} 的次數。我們可以透過期望值最大化法求得參數：

E-step

$$P\left(T_k \mid w_n, M_{w_j}\right) = \frac{P\left(T_k \mid M_{w_j}\right) P\left(w_n \mid T_k\right)}{\sum_{l=1}^{K} P\left(T_l \mid M_{w_j}\right) P\left(w_n \mid T_l\right)} \tag{4-3}$$

M-step

$$\hat{P}\left(T_k \middle| M_{w_j}\right) = \frac{\sum_{w_s \in Q_{w_j}} n\left(w_s, Q_{w_j}\right) P\left(T_k \middle| w_s, M_{w_j}\right)}{\sum_{w_l \in Q_{wj}} n\left(w_l, Q_{w_j}\right)} \tag{4-4}$$

$$P\left(w_n \mid T_k\right) = \frac{\sum_{w_j} n\left(w_n, Q_{w_j}\right) P\left(T_k \mid w_n, M_{w_j}\right)}{\sum_{w_l} \sum_{w_{n'} \in Q_{wl}} n\left(w_{n'}, Q_{w_l}\right) P\left(T_k \mid w_{n'}, M_{w_l}\right)} \tag{4-5}$$

$n\left(w_n, Q_{w_j}\right)$是詞 w_n 在詞模型 w_j 訓練觀測 Q_{w_j} 出現的次數， $P\left(T_k \mid w_n, M_{w_j}\right)$是給定詞模型 w_j 與詞 w_n ，主題 T_k 發生的機率。

如何將詞主題混合模型用於語音辨識呢？辨識詞 w_i 可視為一是長度為 1 的觀測值，而辨識過程中的歷史詞序列 $H_{w_j} = w_1, w_2, ..., w_{i-1}$ 是由一連串的詞所組成，所以我們可以直覺地使用線性組合的方式，將歷史詞模型預測辨識詞的詞主題混合模型整合，形成一個複合式詞模型(Compoisite Word Model)，並定義為歷史詞

序列模型：

$$
\begin{aligned}
P(w_i \mid H_{w_i}) &= \sum_{j=1}^{i-1} \alpha_j P(w_i \mid M_{w_j}) \\
&= \sum_{j=1}^{i-1} \alpha_j \sum_{k=1}^{K} P(w_i \mid T_k) P(T_k \mid M_{w_j}) \\
&= \sum_{k=1}^{K} P(w_i \mid T_k) \sum_{j=1}^{i-1} \alpha_j P(T_k \mid M_{w_j}) \\
&= \sum_{k=1}^{K} P(w_i \mid T_k) P'(T_k \mid M_{H_{w_j}})
\end{aligned}
\tag{4-6}
$$

α_j 表示每一個歷史詞模型的權重，其中 $\sum_{j=1}^{i-1} \alpha_j = 1$，且 $\alpha_j \geq 0$，。經推導後，複合式詞模型的參數 $P'(T_k \mid M_{H_{w_j}})$ 是原有詞模型參數 $P(T_k \mid M_{w_j})$ 的線性組合。所以即使辨識過程中的歷史詞序列 H_{w_i} 一直在改變，歷史詞序列模型仍可以透過先訓練好的詞模型組成，如 圖 4-3 所示。複合式詞模型與觸發對語言模型相似，都是找出歷史詞與辨識詞之間的關係，差別在於我們認為每一個歷史詞都是一個模型，並用一個潛藏的機率分布表示此關係，進而表現詞與詞關係，而觸發對模型沒有定義明確的機率分布。

在訓練詞模型時，固定長度的窗內的詞序列會被當作是訓練觀測值。當窗長度 W 增加時，訓練觀測值也會增加，所以我們可以使用一些統計測量的方法來減少訓練觀測值，例如交互資訊(Mutual Information, MI)與前向後向二連機率

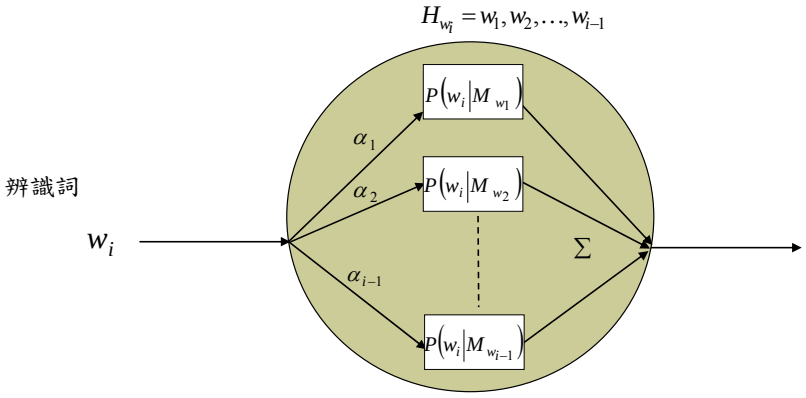


圖 4-3 複合式詞主題混合模型示意圖

(Forward-Backward Bigram, FB)測量：

$$Score_{MI}(w_j, w_i) = \log \frac{P(w_j, w_i)}{P(w_j)P(w_i)} \tag{4-7}$$

$$Score_{FB}(w_j, w_i) = \sqrt{P_f(w_i \mid w_j) P_b(w_j \mid w_i)} \tag{4-8}$$

其中的聯合機率 $P(w_j, w_i)$ 與條件機率 $P(w_j \mid w_i)$ 或 $P(w_i \mid w_j)$ 都是使用長度 l 為視窗內的詞配對次數所估測出：

$$P(w_j, w_i) = \frac{n_l(w_j, w_i)}{T} \tag{4-9}$$

$$P(w_j \mid w_i) = \frac{n_l(w_j, w_i)}{\sum_w n(w, w_i)} \tag{4-10}$$

$n_l(w_j, w_i)$ 是詞 w_j 與詞 w_i 在同一個視窗共同出現的次數。T 是窗內總詞數。根據這兩個分數，我們可以對視窗內所有的訓練詞配對進行排名，然後選擇保留一定比例的語料，再進行詞模型訓練。我們認為這樣不僅可以加快訓練速度(訓練觀測值變少)，亦可能增進模型的效果(模型與觀測值更相關)。

4.1.2 詞主題混合模型與其他模型之比較

我們可由幾個方面比較詞主題混合模型、機率式潛藏語意分析模型及觸發對語言模型，如 表 4-1 所示。首先，詞主題混合模型與觸發對語言模型都是建立歷史詞與預測詞的關係，而機率式潛藏語意分析模型強調的是預測詞與整個歷史詞序列的關係；而模型的使用上，詞主題混合模型與觸發對語言模型都是先從訓練語料訓練好後，於辨識過程中直接使用(線性結合歷史詞模型)，而機率式潛藏語意分析模型除了從訓練語料訓練好之後，還需要於辨識過程中調整主題權重(最大化歷史詞序列相似度)。主題模型建立方面，詞主題混合模型與潛藏語意分析模

表 4-1 詞主題混合模型、機率式潛藏語意分析與觸發對語言模型之比較

模型	詞主題混合模型	機率式潛藏語意分析	觸發對語言模型
模型關係	預測詞與歷史詞	預測詞與歷史詞序列	預測詞與歷史詞
模型估測	事先訓練	線上估測	事先訓練
主題模型	明確分布	明確分布	不明確分布
模型參數	$V \times T + T \times V$	$V \times T + T \times D$	至多$V \times V$
預測能力	是	否	是

型都會使用一個明確的主題機率分布來模擬隱藏的詞或文件主題，而觸發對語言模型則沒有。模型參數方面，詞主題混合模型的參數量是$V \times T + T \times V$，潛藏語意分析模型是$V \times T + T \times D$，觸發對語言模型則是至多$V \times V$，V是詞典大小，T是主題數，D是訓練文件數。預測能力方面，因為詞主題混合模型與觸發對語言模型都是建立歷史詞與辨識詞的關係，我們認為有直接的預測詞能力，潛藏語意分析模型則是透過最大化歷史詞序列相似度，找出其中的主題分布，再藉由與詞分布結合，計算詞機率，我們認為是以間接方式來預測詞。

4.2 位置相關語言模型(Position-Dependent Language Model)

過去已經有許多文章主題或詞類別相關語言模型被提出，例如以文件主題為模型單位的混合主題式語言模型、潛藏語意分析、機率式潛藏語意分析等等，潛藏語意分析透過線性代數方法對潛藏語意建立模型；機率式潛藏語意分析使用機率方式對潛藏語意建立模型；混合主題式語言模型透過分割語料達到不同主題擁有不同訓練語料。除了文章主題之外，詞類別相關語言模型，如*N*連類別模型、聚合式馬可夫模型以及詞主題混合模型亦被提出，用來建立詞與詞的關聯。此外，透過機率式潛藏語意分析建立模型的語者資訊[Akita and Kawahara 2004]或是使用最大熵值法整合詞位置資訊應用於機器翻譯等等亦被提出[Foster 2000]。雖然已經有許多語意資訊的模型或方法被提出，但是位置相關的資訊卻鮮少被提及。本節針對文件或語句的位置資訊，做初步的分析，並提出了位置相關語言模型，如位置性*N*連詞模型(Positional *N*-gram Model)與位置性機率式潛藏語意分析(Positional PLSA)等。

4.2.1 位置資訊的呈現

我們首先說明位置資訊(Position Information)為何。對於某些具有特殊用途的文件或語音而言，我們認為在結構上是有一致性的。文件方面，以學術論文為例，格式通常會是論文摘要、論文簡介、文獻回顧、方法介紹、實驗分析、結論及未來展望，最後是參考文獻。語音方面，以電視或廣播新聞報導為例，順序常會是主播問候語及開場白、主播對新聞事件介紹、記者訪談及註解。這邊的註解，指的是新聞報導告一段落時，記者可能會使用的語句，例如，「記者某某某台北報導」或是「鏡頭交還棚內主播」等結尾用語。這樣的位置資訊，可以視為是一種題材的主題資訊且是區域性的主題資訊，區域性指的是我們認為在每個章節段落之中，有特別的資訊可以使用。

表 4-2 文件層次之位置性樣式詞

D1： 您、公布、宣布、轉到、鏡頭、下面、專題報導、晚安、接下來
D2： 蠻、大概、米酒、那麼、我、我們、念書、了解、裡面、珊瑚
D3： 比方說、譬如說、上面、或者、大概、老師、身分、裡面、小孩子
D4： 公視、編譯、採訪、瑤、蕙、綾、諭、煌、保羅
D1~D4 共同出現： 就是說、這樣子、這邊

表 4-2 是文件層次之位置資訊樣式詞，使用的語料是調適語料SetMAT。呈現方式是先將語料中的每一篇文件等分切成四部分，然後將每篇文章的相同段落合併，再用詞頻數(Term Frequency, TF)與反文件頻數(Inverse Document Frequency, IDF)將每個段落裡較具代表性的詞選出來。詞頻數是調適語料SetMAT的詞頻數，反文件頻數則是SetMAT語料加上背景CNA語料統計得到。SetMAT是公視廣播新聞語料，主要內容是主播或記者報導新聞的過程，偏向口語。我們可以發現，第一段(D1)主要都是一些開場白或是一些連接詞，例如「晚安」、「接下來」、「轉到」等等。第二段(D2)是主要新聞事件的內容，包含一些內容詞(Content Words)，例如「米酒」、「珊瑚」等等。第三段(D3)也是新聞的內容或是一些承接前幾段的說法，例如「上面」、「比方說」、「老師」等等。第四段(D4)則是新聞報導的結束，例如「採訪」、「編譯」及記者名字等等。還有一些詞是每一個段落都出現的，例如「就是說」、「這樣子」等等口語用詞。我們觀察到，相似題材的語料在不同的段落的確有其樣式存在，例如在公視廣播新聞語料中，第一段與第四段比較有一致的特色，而中間段落因為是新聞事件則較分散。

另一種位置資訊是語句(Sentence)之中詞的位置。這種資訊跟語句的結構有關，且與詞性有關。例如一個語句中可能包含了主詞、動詞、受詞、形容詞、副詞、連接詞等，這些詞性的組合及順序，往往受限於語言本身的文法設定，例如在英文中，主詞加動詞加受詞的主動句型，或受詞加被動詞加介系詞加主詞的被動句型，或是副詞加動詞加名詞這種倒裝句型，中文也有其文法概念，我們希望

表 4-3 語句層次之位置性樣式詞

S1： 不過、今年、他們、包括、另外、由於、根據、記者、對於、雖然
S2： 仍、文、可能、股、長、國、著
S3： 四、很、相當、高雄、得、最、給、達、過
S4： 之外、之後、方式、以上、編譯、調查、認為、處理、報導
S1~S4 共同出現： 了、上、也、不、中、他、台灣、民眾、在、多、年

能夠使用這樣的位置資訊。

表 4-3 是語句層次的位置資訊樣式詞，使用的語料為SetET。SetET是東森新聞文字語料，語句較具有文法特性。呈現的方式是先將語料以每一句的形式表示，例如透過逗點、問號、句號等標點符號斷句，再將每一語句等分成四份，然後選出前一百名詞頻數較高的詞，最後每一段落只保留僅在某一段落出現詞。我們可以發現，於第一段(S1)，名詞、連接詞或時間副詞是最常出現的詞性，如「今年」、「不過」、「記者」等。第二段(S2)及第三段(S3)則是較不明顯，大部分是副詞，如「仍」、「很」、「相當」等。第四段(S4)則多是動詞或時間副詞，如「編譯」、「調查」、「之後」等。第一到第四段共同出現的部分則表示這些詞可能是有多種詞性，例如、「台灣」、「民眾」等，可以在句子前面當主詞使用，亦可能在句子後面當受詞，或是「不」、「也」等副詞可以在不同位置修飾形容詞或動詞。

我們希望能將這兩種層次的位置資訊應用於語言模型之中。我們初步嘗試建立了位置性 *N* 連詞模型與位置性潛藏語意分析等模型。

4.2.2 位置性*N*連詞模型(Positional *N*-gram Model)

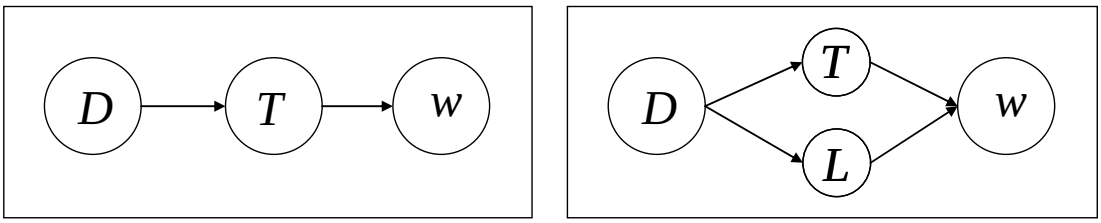
由前一節可以知道，結構化的文件之中，會有位置相關的資訊可以使用，如果文件群有相似的文件結構，可以收集其統計資訊。一般而言，文件分為簡介、事件、結論等，其中所佔的比例可能不太一樣。假設我們的語料結構相似，所以我們初步地將文件個別依長度做等分切割，再將位置相同的文件段落合併成新的文件集。接著分別針對新的文件集訓練 *N* 連詞模型，例如三連詞模型

$P(w_i \mid w_{i-2}, w_{i-1}, L_j)$，$L_j$ 表示不同的段落。而為了降低因為分割語料而產生資料稀疏問題，我們可以對每個段落的語言模型做平滑化，或是直接結合不同段落產生的語言模型：

$$P_{Pos}(w_i \mid w_{i-2}, w_{i-1}) = \sum_{j=1}^{L} \lambda_j P(w_i \mid w_{i-2}, w_{i-1}, L_j) \tag{4-11}$$

L 是總共分段數，λ_j 是段落 L_j 對應的權重。我們亦可以加入未分段的一般語言模型當作背景平滑化。語句位置相關模型亦是類似的作法，我們可以先將語句分段，再分別訓練 N 連詞模型。

要如何在辨識過程中使用位置資訊？有兩種方式，一種是直接使用結合的位置相關模型，如式 (4-11)，然後根據不同的歷史詞序列調整其權重 λ；另一種是如果是詞圖重計分(Word Graph Rescoring)階段，我們會知道第一名(Top 1)序列的長度，與歷史詞序列長度的比例，判斷目前欲辨識詞位置，然後使用固定位置的語言模型。如果是語句的位置資訊，我們假設靜音位置代表語句的開始或結束，則可以將前一個靜音位置到目前詞視為一個未完成語句，再找出其中的詞性，判斷其可能所屬的句型，再透過訓練語料不同句型的平均語句長度來決定目前語句完成度，再決定目前詞位於語句的位置為何。除此之外，也可如式 (4-11)直接結合不同語句位置的模型，再根據最大化未完成語句相似度以調整權重。



(a) (b)

圖 4-4 圖形模型 (a) 機率式潛藏語意分析 (b) 位置性機率式潛藏語意分析

4.2.3 位置性機率式潛藏語意分析(Positional Probabilistic Latent Semantic Analysis)

我們亦嘗試將位置資訊整合於機率式潛藏語意分析，我們稱為位置性機率式潛藏語意分析(Positional Probabilistic Latent Semantic Analysis, Positional PLSA)。圖 4-4 是機率式潛藏語意分析與位置性機率式潛藏語意分析的圖型模型示意圖。D 表示文件，T 表示主題，L 表示位置，w 表示詞。我們可以發現，機率式潛藏語意分析與位置性機率式潛藏語意分析的差別在於位置資訊L被加入。假設給定文件D時，位置L與主題T是無關的，而且給定位置L與主題T時，詞w與文件D無關。所以我們可將位置性機率式潛藏語意分析表示成：

$$P_{PosPLSA}\left(w_i\middle|M_{H_{w_i}}\right)=\sum_{j=1}^{S}P\left(L_j\middle|M_{Hw_i}\right)\sum_{k=1}^{K}P\left(T_k\middle|M_{Hw_i}\right)P\left(w_i|T_k,L_j\right), \tag{4-12}$$

$P\left(L_j\middle|M_{Hw_i}\right)$是給定歷史詞序列$H_{w_i}$，位置$L_j$的機率，$P\left(T_k\middle|M_{Hw_i}\right)$是給定歷史詞序列$H_{w_i}$，主題$T_k$的機率，$P\left(w_i|T_k,L_j\right)$是給定位置$L_j$與主題$T_k$時，詞$w_i$的機率。於語音辨識過程中，與機率式潛藏語意分析相同，我們保留詞機率$P\left(w_i|T_k,L_j\right)$，而需要即時調整位置機率$P\left(L_j\middle|M_{Hw_i}\right)$與主題機率$P\left(T_k\middle|M_{Hw_i}\right)$。同樣地，位置$L_j$可以是固定的。當位置固定，即$P\left(L_j\middle|M_{Hw_i}\right)$是 1，其餘位置機率為 0，會變成傳統的機率式潛藏語意分析，差別在於是詞機率是由使用不同位置的語料所求得。

4.3 實驗結果與分析

4.3.1 詞主題混合模型

我們首先進行不同訓練視窗大小與不同辨識歷史詞長度的實驗。在實驗中，辨識時的歷史詞模型權重α先採用平均分布。使用SetMAT語料於發展集結果如 表 4-4、表 4-5、表 4-6、表 4-7 所示。我們可以發現，當歷史詞模型權重α採用平均分布，不管訓練窗長度是 1、3、5、10，其結果都是隨歷史詞序列變長而變差。這可能是因為歷史詞序列會有辨識錯誤，會使用到不正確的詞模型。或是因為對於新聞報導而言，內部的主題差異可能也很大，產生歷史與目前辨識主題不同的情況。根據實驗結果，發現在訓練視窗長度等於 5，歷史詞序列等於 3 的時候，在 64 及 128 主題數時較好，如 圖 4-5 所示。此外，我們也根據固定的歷史詞長度來觀察不同的訓練視窗長度，如 圖 4-6、圖 4-7、圖 4-8、圖 4-9 所示。我們發現，歷史詞序列固定時，訓練視窗為 3 的時候會有較佳的結果。而且我們亦觀察到，如果歷史詞序列等於 1，其訓練視窗長度及模型主題數不需太大。反之，歷史詞序列越長時，訓練視窗長度需要越長，且模型亦要變得複雜，才能捕捉到更多的資訊。最後，我們採用訓練視窗長度為 3 的模型，再進一步分析。

表 4-4 SetMAT 訓練窗為 1 之詞主題混合模型於發展集字錯誤率(%)結果

歷史詞長度	16	32	64	128
1	19.67	19.67	19.70	19.79
3	20.04	19.91	19.79	19.95
5	19.97	19.99	19.92	19.92
10	20.07	20.09	20.12	20.06
20	20.10	20.20	20.18	20.10
50	20.10	20.13	20.23	20.09
∞	20.12	20.08	20.24	20.08

表 4-5 SetMAT 訓練窗為 3 之詞主題混合模型於發展集字錯誤率(%)結果

歷史詞長度	16	32	64	128
1	19.88	19.91	19.89	19.76
3	19.93	19.86	19.83	19.70
5	20.00	19.98	19.86	19.86
10	20.08	20.05	19.93	19.88
20	20.14	20.07	20.06	20.06
50	20.24	20.16	20.11	20.13
∞	20.23	20.18	20.10	20.13

表 4-6 SetMAT 訓練窗為 5 之詞主題混合模型於發展集字錯誤率(%)結果

歷史詞長度	16	32	64	128
1	19.88	20.00	19.86	19.92
3	19.95	20.00	19.83	19.74
5	20.03	20.06	19.87	19.87
10	20.10	20.07	19.99	19.99
20	20.19	20.12	20.04	20.06
50	20.25	20.20	20.12	20.08
∞	20.24	20.20	20.10	20.10

表 4-7 SetMAT 訓練窗為 10 之詞主題混合模型於發展集字錯誤率(%)結果

歷史詞長度	16	32	64	128
1	19.96	20.03	19.92	19.86
3	20.01	20.01	19.97	19.87
5	20.10	20.12	20.00	19.99
10	20.14	20.17	20.06	20.07
20	20.17	20.20	20.08	20.08
50	20.19	20.22	20.05	20.08
∞	20.18	20.20	20.05	20.08

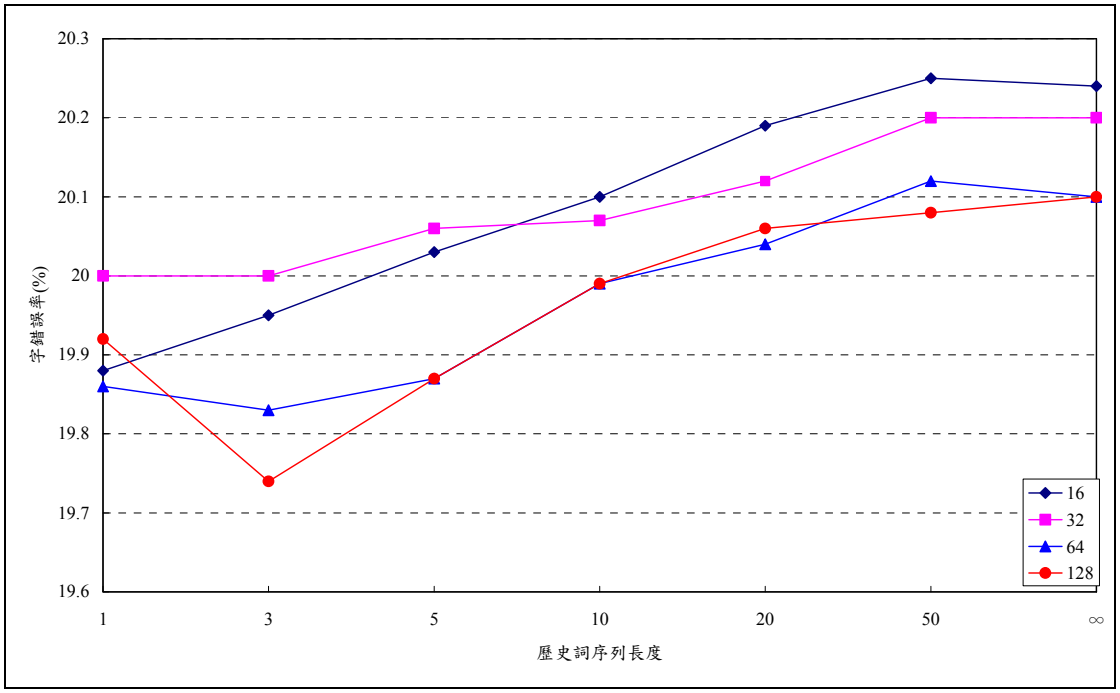


圖 4-5 SetMAT 訓練窗為 5 之詞主題混合模型於發展集字錯誤率(%)結果

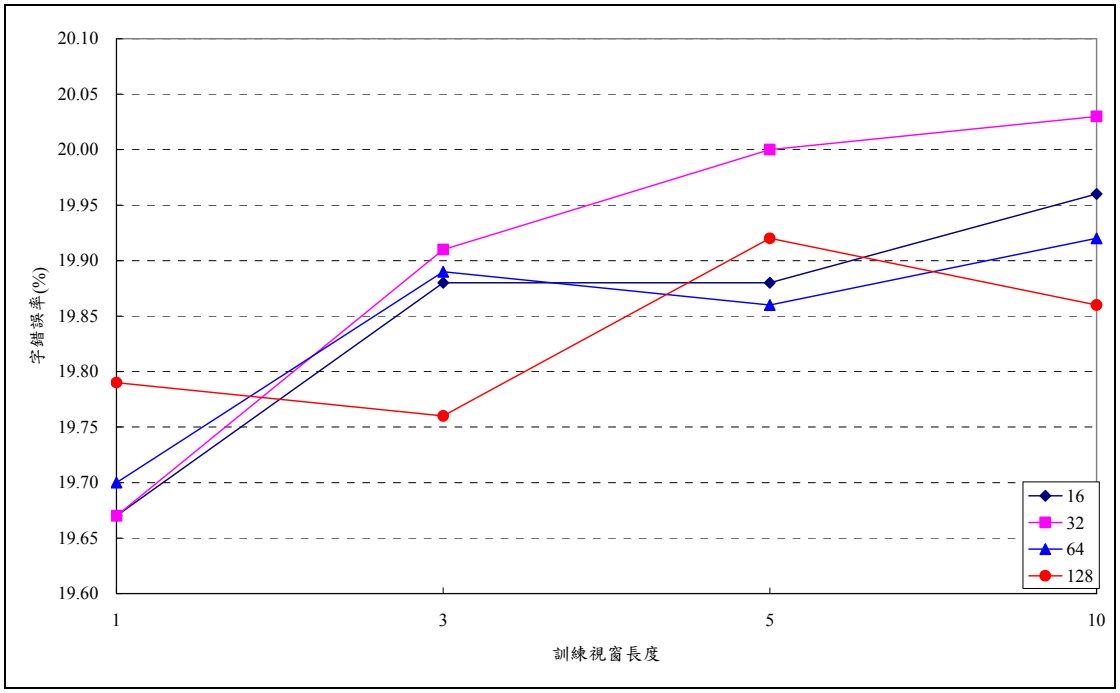


圖 4-6 SetMAT 歷史詞序列長度為 1 之詞主題混合模型於發展集字錯誤率(%)結果

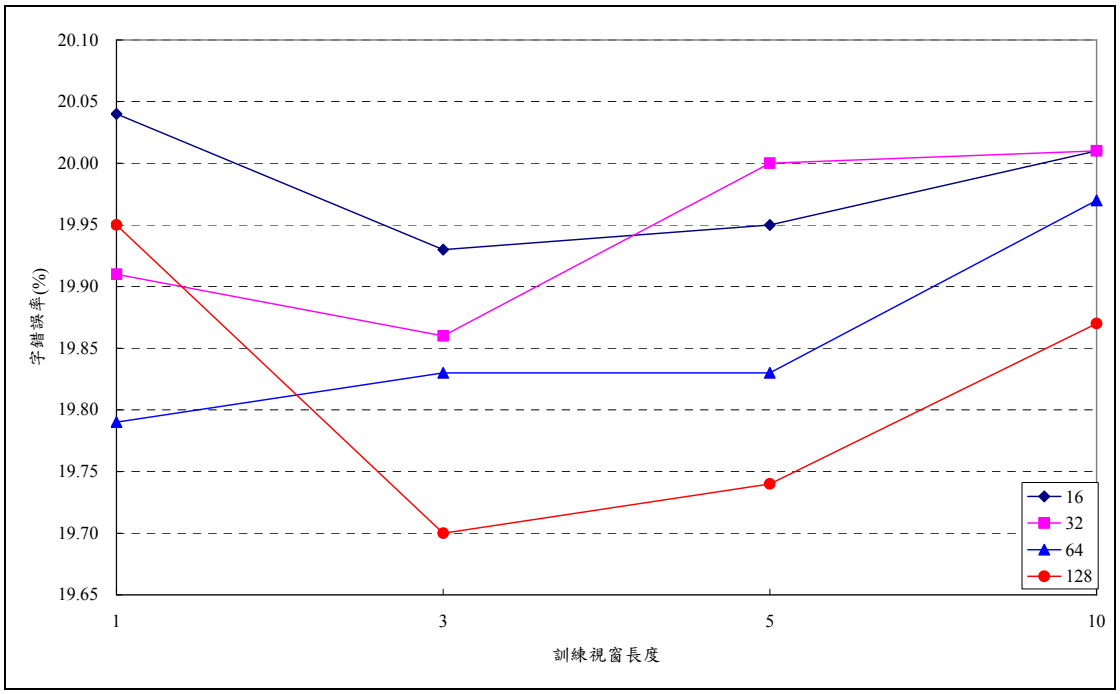


圖 4-7 SetMAT 歷史詞序列長度為 3 之詞主題混合模型於發展集字錯誤率(%)結果

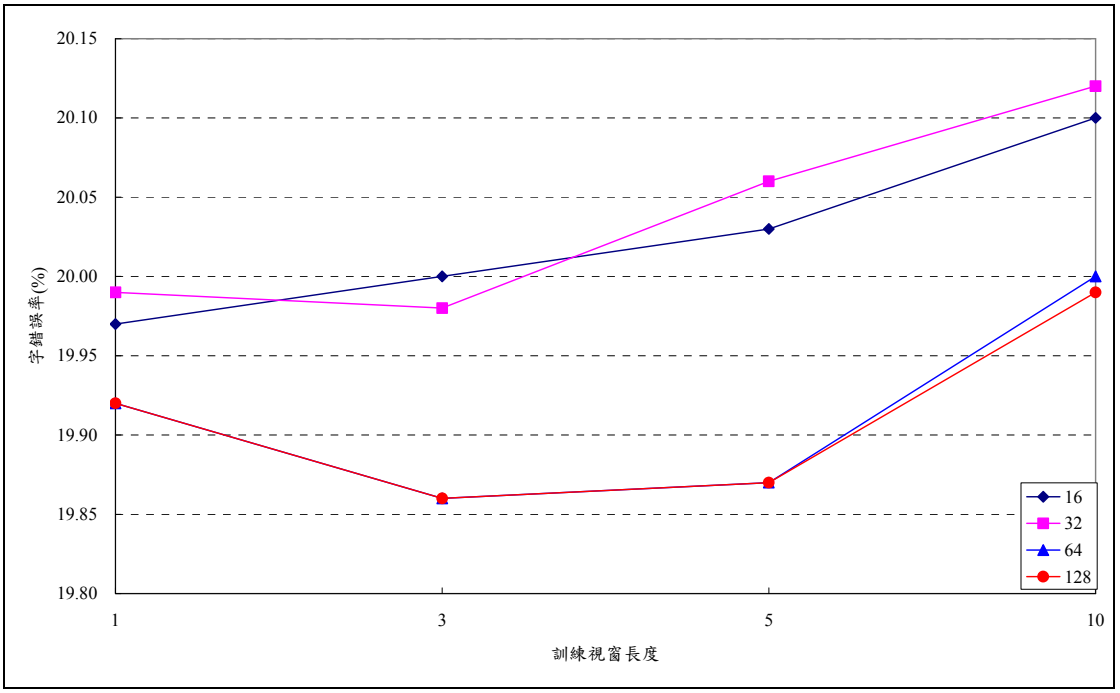


圖 4-8 SetMAT 歷史詞序列長度為 5 之詞主題混合模型於發展集字錯誤率(%)結果

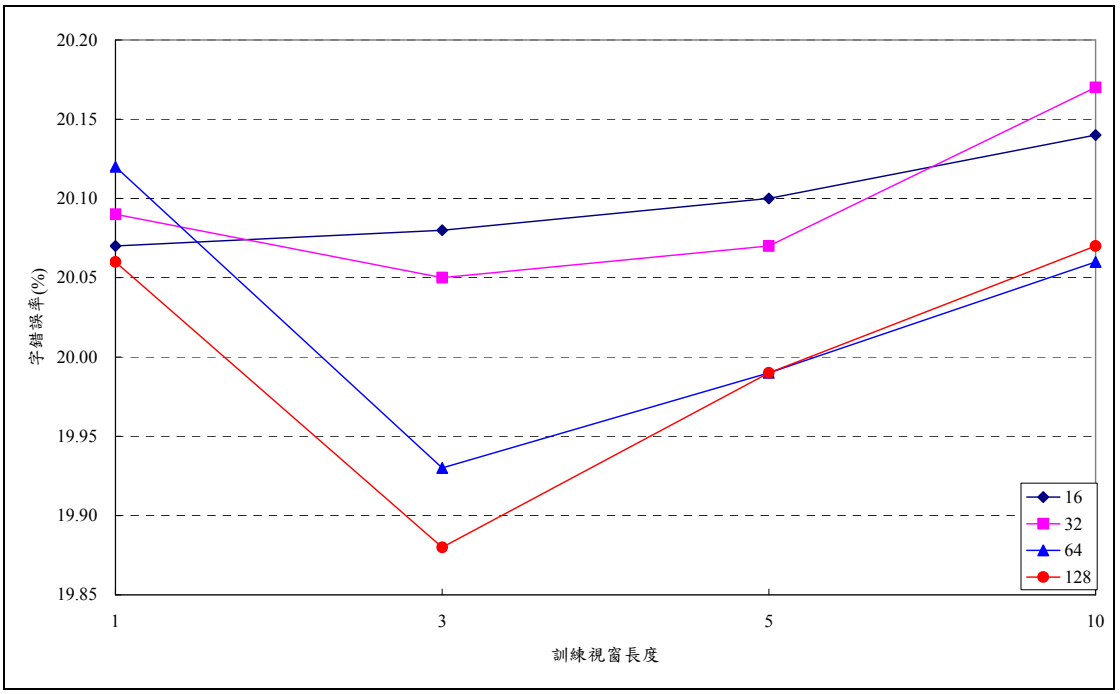


圖 4-9 SetMAT 歷史詞序列長度為 10 之詞主題混合模型於發展集字錯誤率(%)結果

因為歷史詞序列錯誤及長度的影響，我們初步嘗試使用指數遞減(Exponential Decayed)的方式來設定α值，如表 4-8 與圖 4-10 所示。權重越高，表示越早時間的詞影響越小。我們可以發現，當我們採用遞減權重時，在歷史詞長度為 5 與遞減權重設為 0.3 時能有最好的結果，而平均分配時，則會是越來越差。

表 4-8 SetMAT 詞主題混合模型指數遞減權重實驗於發展集字錯誤率(%)結果

歷史詞長度	權重					
	平均分配	0.1	0.2	0.3	0.4	0.5
1	19.76	19.76	19.76	19.76	19.76	19.76
3	19.70	19.69	19.69	19.63	19.63	19.67
5	19.86	19.71	19.68	19.55	19.65	19.65
10	19.88	19.71	19.68	19.65	19.64	19.65
20	20.06	19.71	19.68	19.65	19.64	19.65

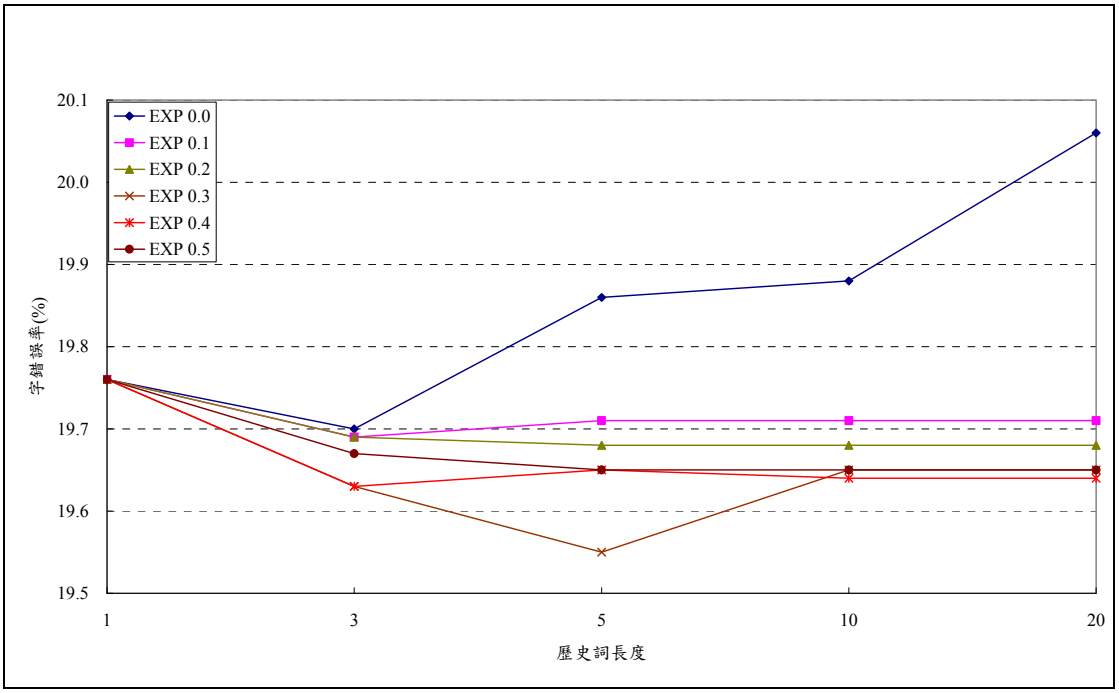


圖 4-10 SetMAT 詞主題混合模型指數遞減權重實驗於發展集字錯誤率(%)結果

然後我們嘗試比較詞主題混合模型(WTMM)、機率式潛藏語意分析(PLSA)與觸發對語言模型(Trigger-based LM)的效果。我們可以發現，詞主題混合模型在字錯誤率方面能夠比機率潛藏語意分析與觸發對語言模型好，表示透過隱藏詞主題來表現詞關聯的確能有助於語音辨識。而語言複雜度方面，則是以使用交互資訊的觸發對語言模型為最好。這可能是因為詞主題混合模型的詞機率分布會透過隱藏主題，分給更多相似的詞，而觸發對語言模型則不會。雖然語言複雜度較低，但是對於字錯誤率來說，未必較好，反而是具有生成能力的詞主題混合模型來得有效。而使用詞頻數與反文件頻數的觸發對語言模型效果不如預期[Troncoso *et al.* 2004]，因為原始作法是使用從辨識文件的最佳*N*序列抽出的觸發對作調適，而我們是使用另一份調適語料，加上題材是新聞語料，變化較大所致。

最後，我們探討不同的訓練資料選取方式對於詞主題混合模型的影響。我們採用了交互資訊及前向後向二連機率之幾何平均兩種分數。我們先根據不同的訓練窗找出訓練詞配對，將這些詞配對根據這兩種分數排序，然後選擇不同比例的訓練配對，結果如 圖 4-12 所示，X軸表示不同比例的訓練資料，Y軸是使用訓練資料對應的字錯誤率。我們可以觀察到，在前向後向二連機率之幾何平均這種選擇方式(FB)下，於 16 與 128 個主題數時，使用 70%的訓練資料能夠幾乎達到跟全部使用的效果，這表示我們能夠減少 30%的訓練資料。這 30%訓練資料對於模型來說可能是沒有幫助的，例如兩個詞不太相關，亦可能是因為這 30%資料與測試語料不相關的緣故。

表 4-9 詞主題混合模型、機率式潛藏語意分析與觸發對語言模型於發展集結果

語言複雜度	16/50K	32/80K	64/400K	128/900K
WTMM	520.31	510.26	507.13	499.3
PLSA	540.52	533.07	527.82	519.71
Trigger-Based LM (MI)	514.34	499.63	481.86	467.62
Trigger-Based LM (TFIDF)	627.59	614.7	559.84	501.01
字錯誤率(%)	16/50K	32/80K	64/400K	128/900K
WTMM	19.80	19.76	19.69	19.55
PLSA	20.13	20.06	19.99	19.95
Trigger-Based LM (MI)	20.13	20.09	20.07	19.99
Trigger-Based LM (TFIDF)	20.69	20.63	20.63	20.25

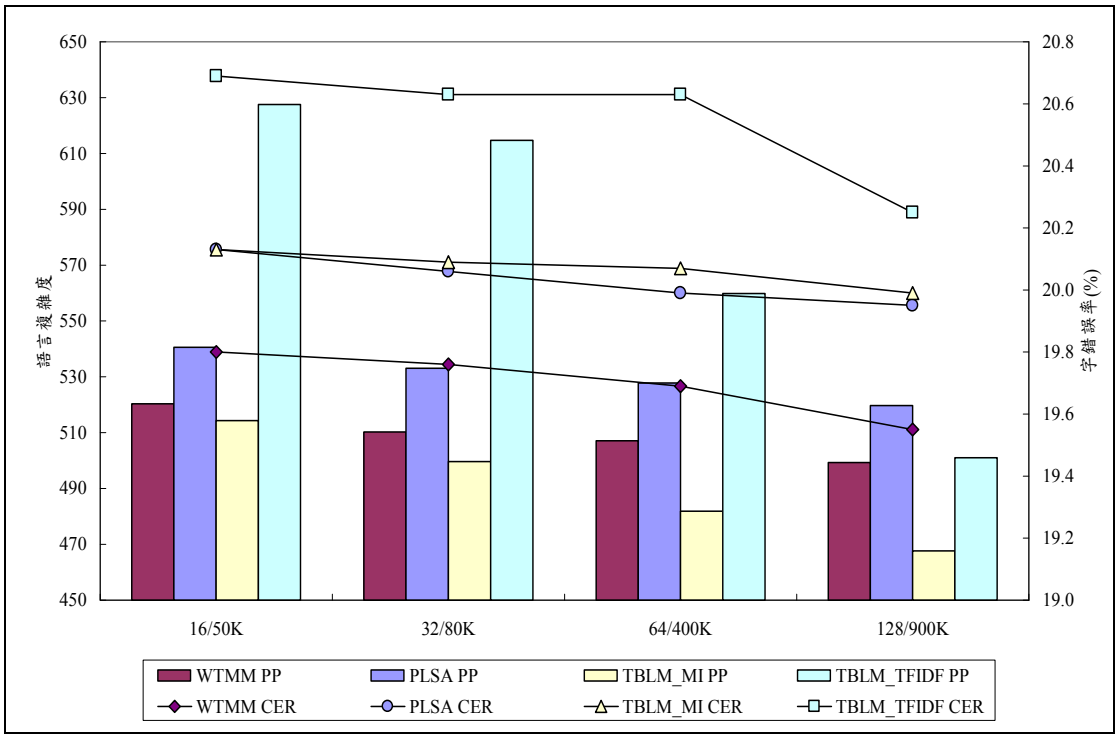


圖 4-11 詞主題混合模型、機率式潛藏語意分析與觸發對語言模型於發展集結果

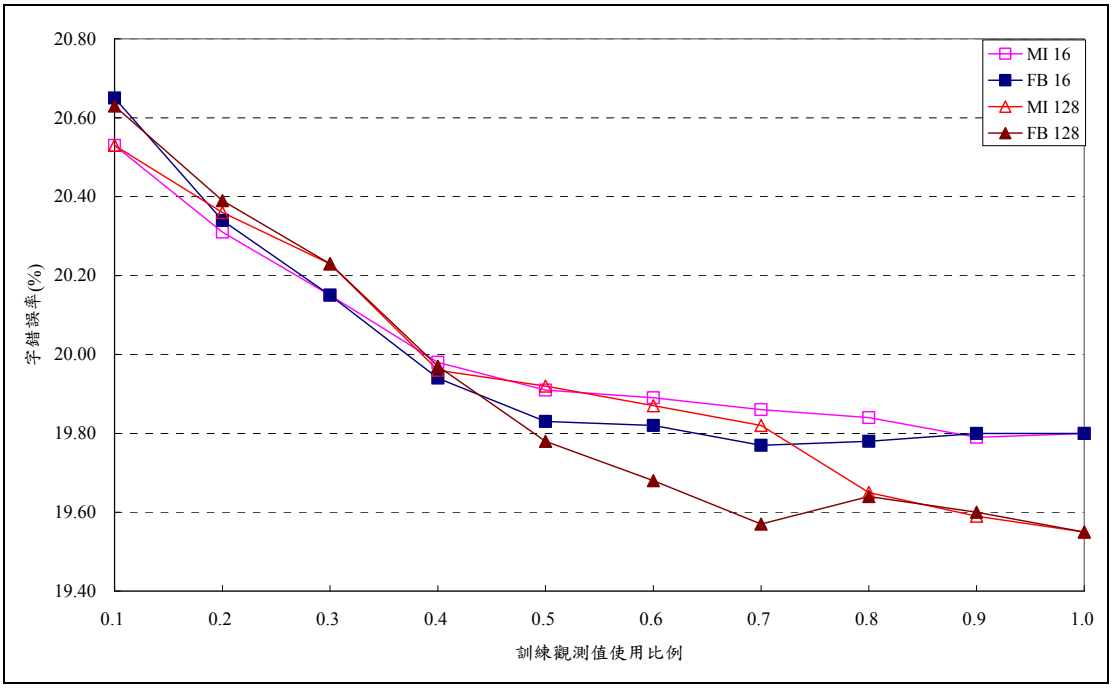


圖 4-12 詞主題混合模型使用不同比例及選擇方式訓練資料於發展集結果

4.3.2 位置相關語言模型

我們初步地探討詞在文件位置的資訊。先根據每一文件大小將語料等分成不同的段落，再收集每文件所屬段落訓練*N*連詞模型，其中*N*連詞模型亦經過Katz平滑化法。我們首先比較歷史詞序列與第一名序列(Top 1)長度，透過其比例，找出目前辨識詞在文件位置，再使用其位置所屬的*N*連詞模型，我們稱為決定位置性*N*連詞模型(Determined Positional *N*-gram Model)，如 表 4-10 與 表 4-11 所示。然而，其結果未如預期，這可能是因為測試語料的文件結構與調適語料仍有一些差異，所以段落的資訊沒有辦法完整顯示。而這個問題，我們可以於辨識過程中透過動態地結合不同位置性*N*連詞模型解決，結果如 表 4-12 與 表 4-13 所示，我們呈現了使用二連及三連詞模型的結果，位置數 1 表示使用未分段平滑化*N*連詞模型。實驗結果顯示，分段數在 2 到 4 之間時，會有較佳的結果，當分段數太多時，反而變差。這可能因為測試文件結構或長度與調適語料不一致的關係。在SetET時，位置性相關*N*連詞模型的確會比未分段來的好，然而在SetMAT時，改善幅度較小，這可能是因為SetMAT語料量太少的關係，所以分段訓練平滑化*N*連詞模型的機率分布都很類似。於評估集中，使用SetET語料，位置數等於 4 時，三連詞模型字錯誤率由 19.34%降為 19.08%，相對改進 1.34%。語言複雜度由 544.04 降為 482.20，相對改進 11.36%。使用SetMAT語料，則由 19.23%降為 19.08%，相對改進 0.78%，語言複雜度由 434.46 降為 399.67，相對改進 8%。

表 4-10 SetET 與 SetMAT 之決定位置性 *N* 連詞模型於發展集結果

位置數	SetET	SetMAT
二連	字錯誤率(%)	字錯誤率(%)
1	19.89	19.51
2	19.90	19.64
3	20.04	19.50
4	20.04	19.75
8	20.15	19.76
16	20.24	19.94
三連	字錯誤率(%)	字錯誤率(%)
1	19.65	19.46
2	19.80	19.67
3	19.89	19.58
4	19.93	19.71
8	20.13	19.81
16	20.26	19.95

表 4-11 SetET 與 SetMAT 之決定位置性 *N* 連詞模型於評估集結果

位置數	SetET	SetMAT
二連	字錯誤率(%)	字錯誤率(%)
1	19.65	19.23
2	19.55	19.16
3	19.65	19.37
4	19.80	19.33
8	19.81	19.59
16	19.86	19.83
三連	字錯誤率(%)	字錯誤率(%)
1	19.34	19.23
2	19.36	19.09
3	19.41	19.26
4	19.61	19.29
8	19.79	19.62
16	19.88	19.85

表 4-12 SetET 與 SetMAT 之位置性 *N* 連詞模型於發展集結果

位置數	SetET		SetMAT	
二連	字錯誤率(%)	語言複雜度	字錯誤率(%)	語言複雜度
1	19.89	540.10	19.50	439.16
2	19.61	481.35	19.50	400.16
3	19.62	481.13	19.51	405.78
4	19.67	481.16	19.54	409.76
8	19.68	485.00	19.60	421.76
16	19.76	490.97	19.63	434.58
三連	字錯誤率(%)	語言複雜度	字錯誤率(%)	語言複雜度
1	19.65	507.30	19.46	426.59
2	19.44	451.39	19.50	387.95
3	19.52	451.13	19.44	392.23
4	19.42	451.86	19.52	395.27
8	19.54	459.94	19.57	403.86
16	19.67	471.54	19.60	417.60

表 4-13 SetET 與 SetMAT 之位置性 *N* 連詞模型於評估集結果

位置數	SetET		SetMAT	
二連	字錯誤率(%)	語言複雜度	字錯誤率(%)	語言複雜度
1	19.65	576.04	19.23	447.55
2	19.54	510.36	19.05	405.34
3	19.48	509.77	19.18	410.70
4	19.55	510.55	19.11	414.65
8	19.37	512.51	19.22	426.20
16	19.42	517.71	19.38	439.88
三連	字錯誤率(%)	語言複雜度	字錯誤率(%)	語言複雜度
1	19.34	544.04	19.23	434.46
2	19.18	481.70	19.08	392.48
3	19.18	481.00	19.11	397.32
4	19.08	482.20	19.08	399.67
8	19.14	488.19	19.19	408.54
16	19.26	501.87	19.35	423.13

我們可以進一步分析位置性*N*連詞模型與 混合主題式語言模型。如 圖 4-13 所示，給定訓練語料文件集，混合主題式語言模型是根據文件的主題做分群，每一群由許多相似的文件所組成，位置性*N*連詞模型是根據文件的段落做分割，再將同一段落收集起來，我們認為這些段落是相似的。相較之下，混合主題式語言模型需要額外的分群技術，而位置性*N*連詞模型需要相似的文件結構，段落的主題性才能突顯出來。我們認為這兩種不同的分割方式能夠補捉住不同的文件資訊。

圖 4-14、圖 4-15、圖 4-16 與 圖 4-17 是採用三連模型的位置性*N*連詞模型與混合主題式語言模型的語言複雜度的比較，P表示位置性*N*連詞模型，M表示混合主題式語言模型，分割數是指分群數或是分段數。我們發現，位置性*N*連詞模型與混合主題式語言模型皆能改善原始的三連詞模型，而使用混合主題式語言模型的結果則比位置性*N*連詞模型稍微好一點，特別是在分割數大於 4 的情況更明顯。我們認為是因為位置資訊是針對每一篇文件做切割，而文件結構本身可能就沒有這麼多的段落。而主題資訊是對文件分群，所以相似的仍會集中於一群，在字錯誤率方面亦是如此。圖 4-18、圖 4-19、圖 4-20 與 圖 4-21 是採用三連模型的位置性*N*連詞模型與混合主題式語言模型的字錯誤率的比較。我們可以直接比較圖例實心點與空心點的分布，實心點代表位置性*N*連詞模型，空心點代表混合主題式語言模型。

混合主題式語言模型
(Mixture-based Language Model)

位置性*N*連詞模型
(Positional *N*-gram Model)

圖 4-13 混合主題式語言模型與位置性 *N* 連詞模型示意圖

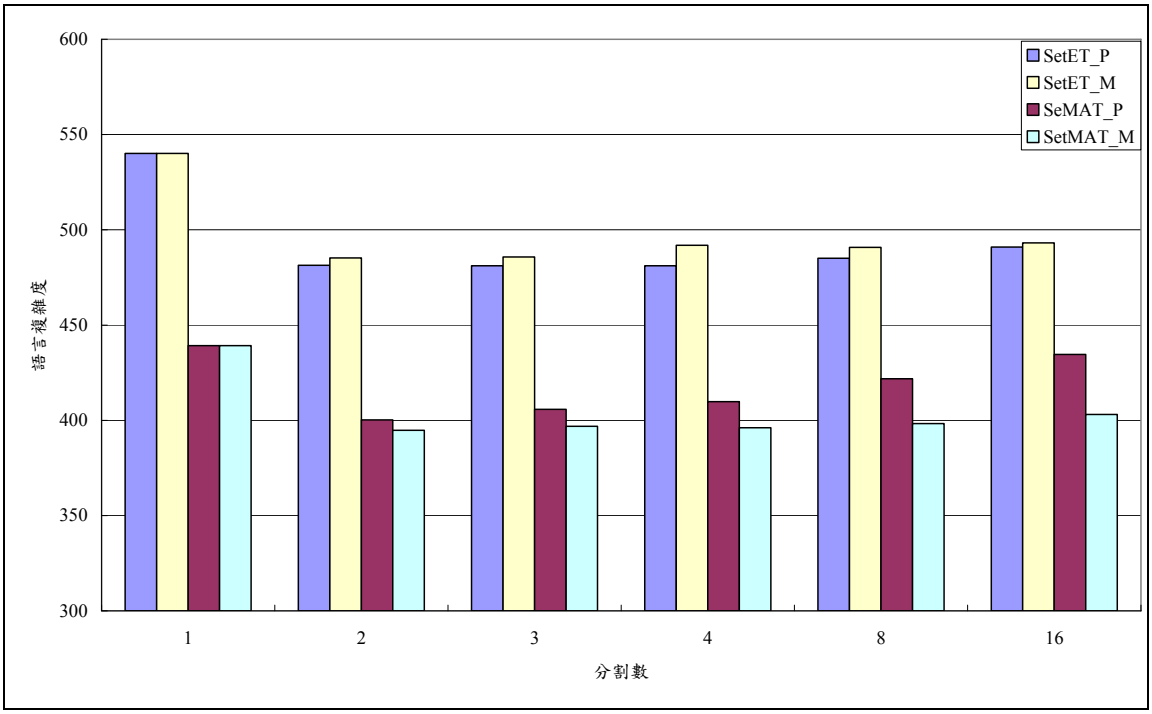


圖 4-14 位置性二連詞模型與混合主題式語言模型於發展集語言複雜度結果

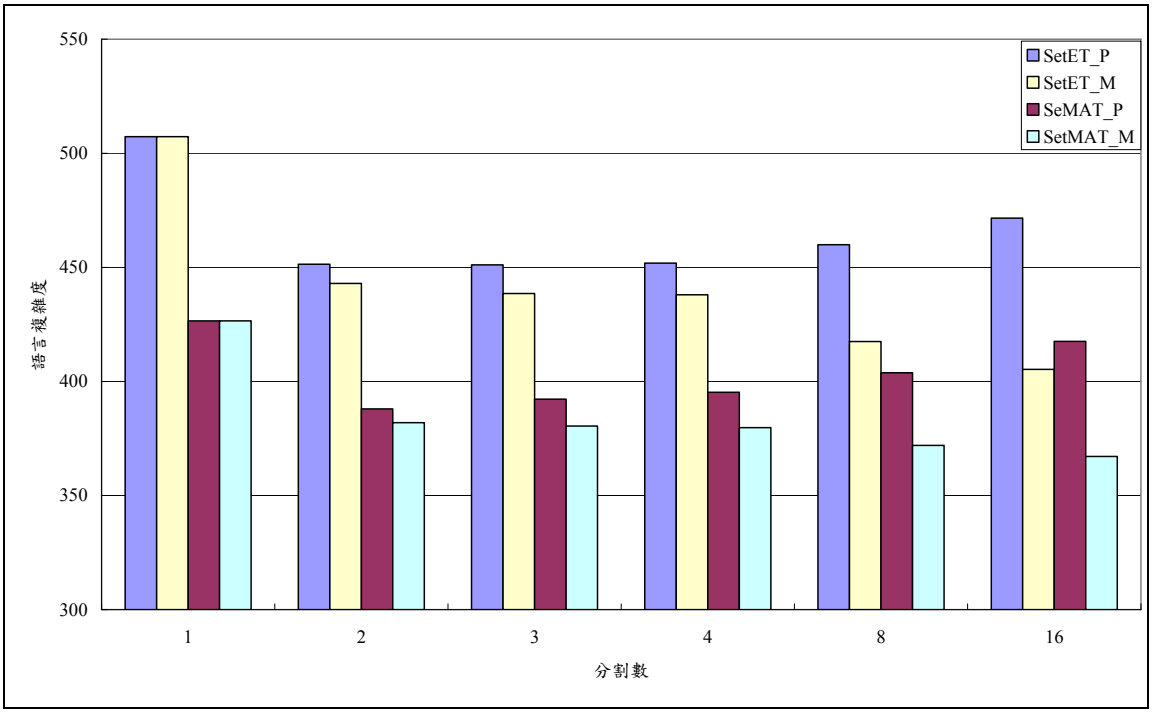


圖 4-15 位置性三連詞模型與混合主題式語言模型於發展集語言複雜度結果

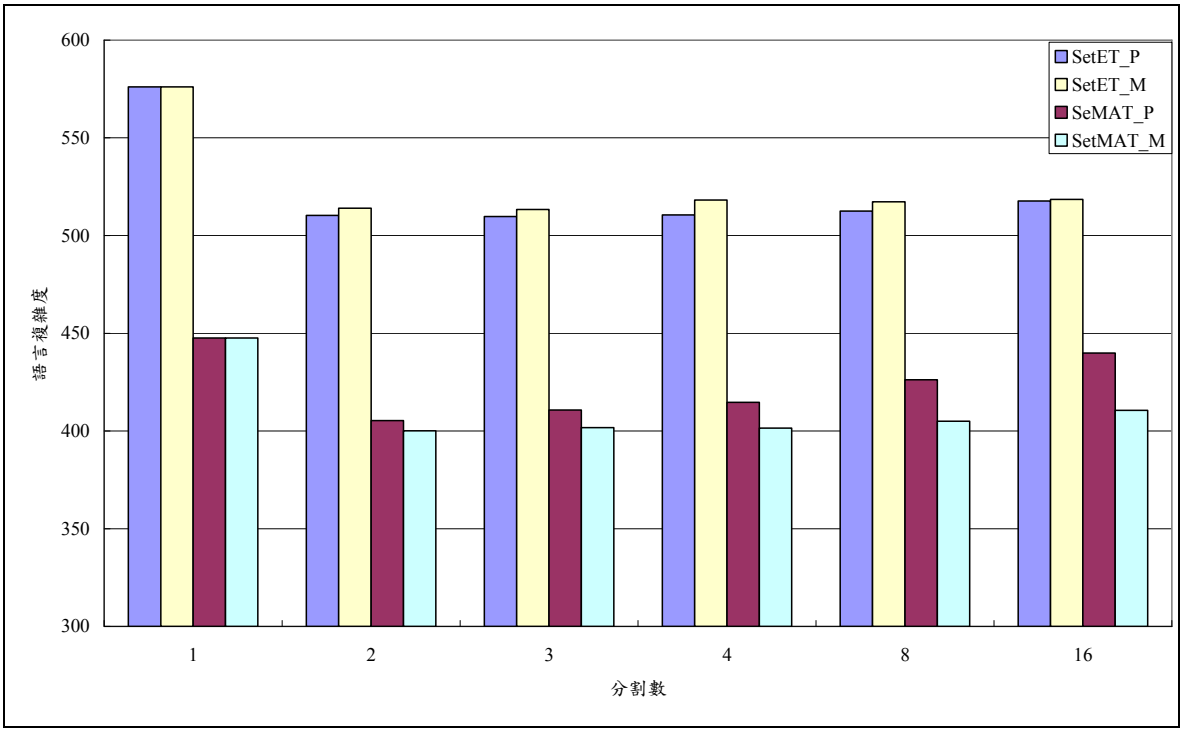


圖 4-16 位置性二連詞模型與混合主題式語言模型於評估集語言複雜度結果

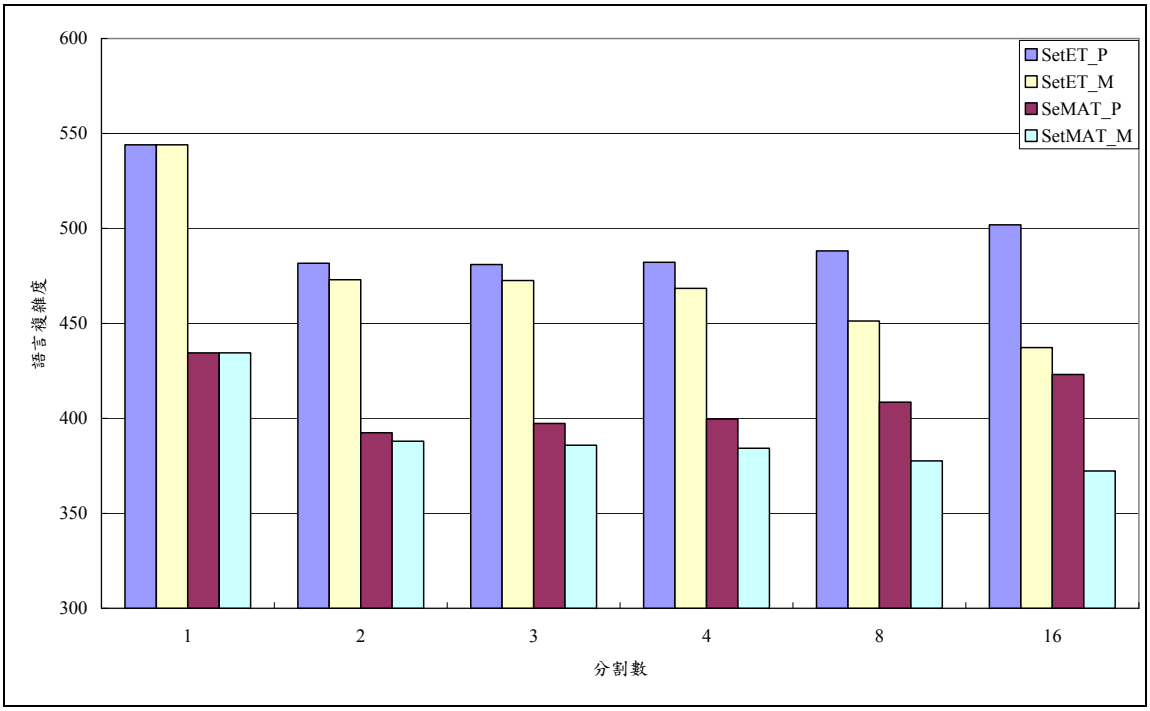


圖 4-17 位置性三連詞模型與混合主題式語言模型於評估集語言複雜度結果

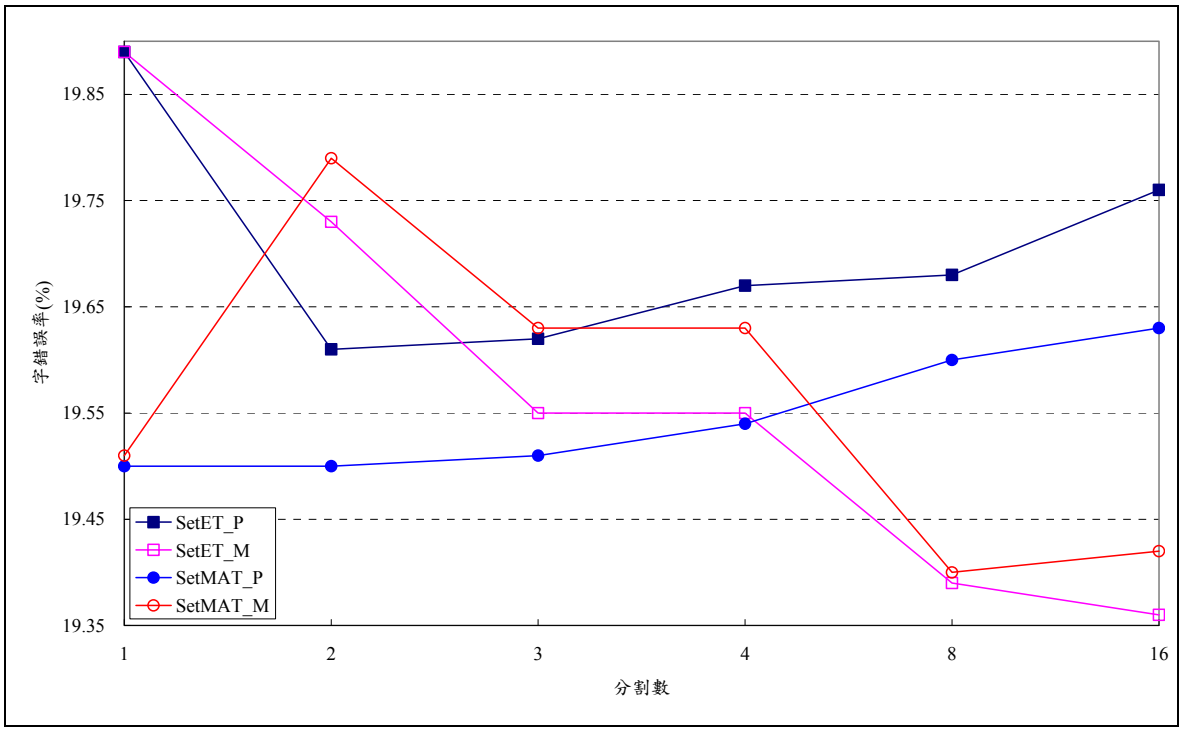


圖 4-18 位置性二連詞模型與混合主題式語言模型於發展集字錯誤率(%)結果

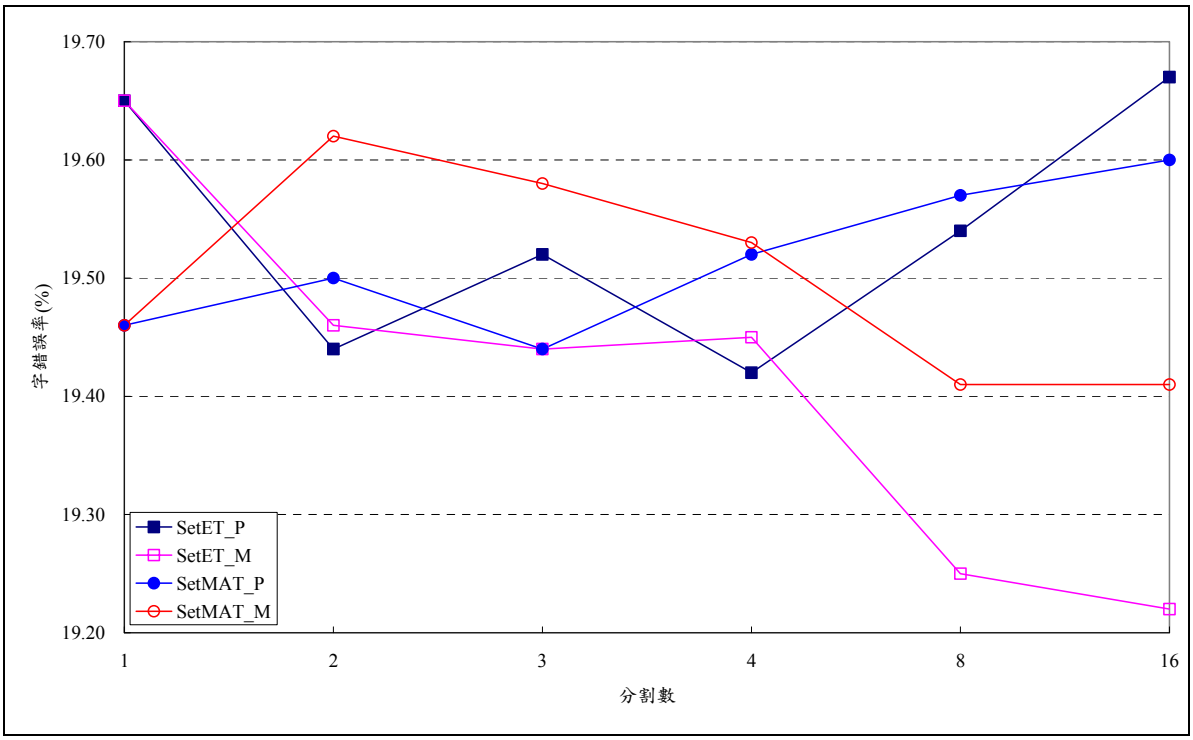


圖 4-19 位置性三連詞模型與混合主題式語言模型於發展集字錯誤率(%)結果

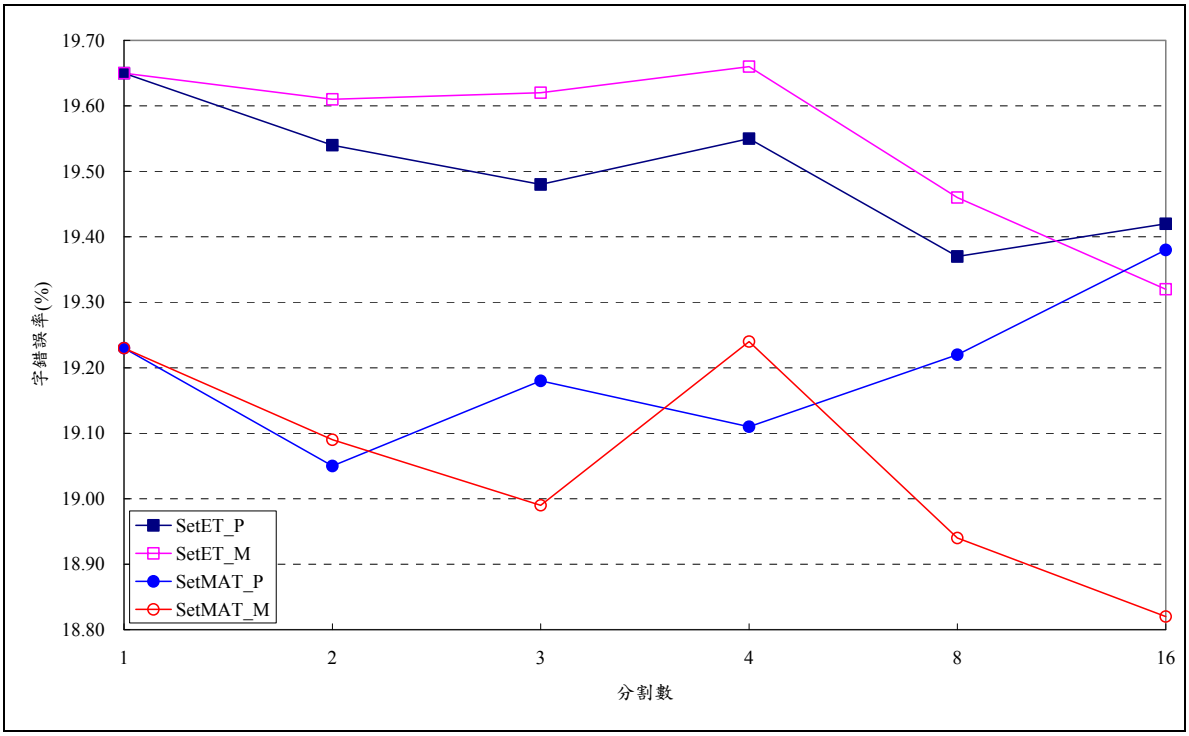


圖 4-20 位置性二連詞模型與混合主題式語言模型於評估集字錯誤率(%)結果

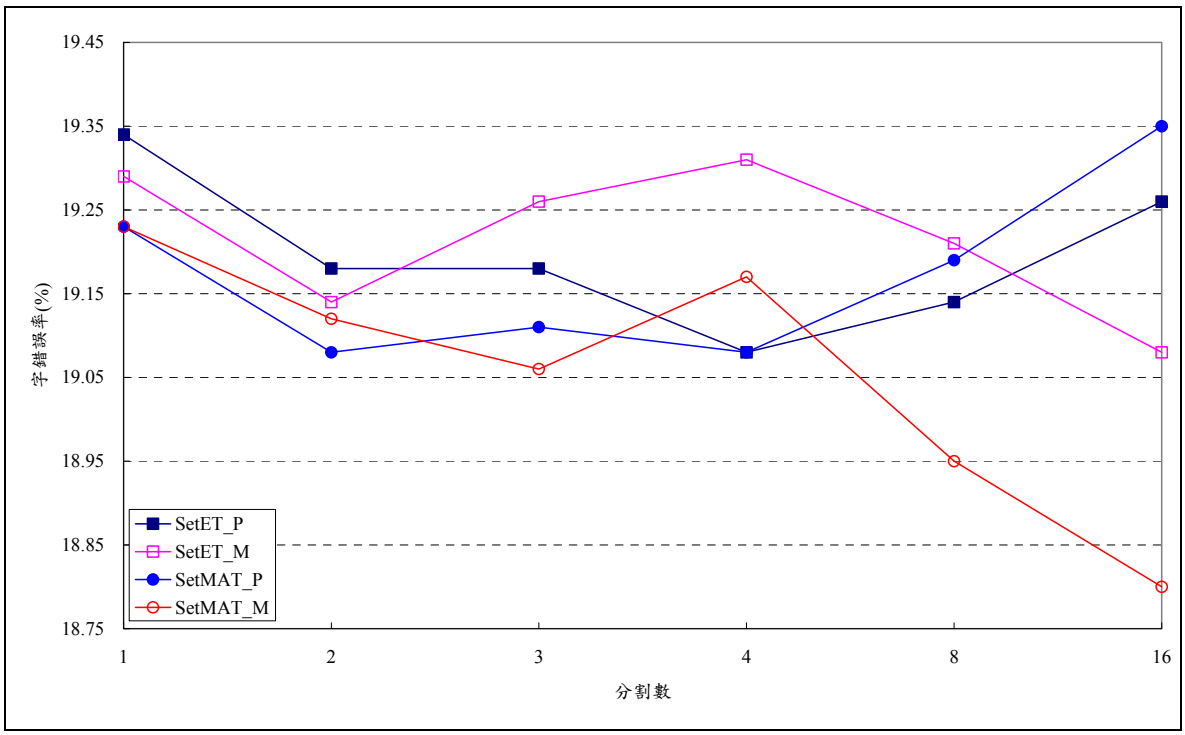


圖 4-21 位置性三連詞模型與混合主題式語言模型於評估集字錯誤率(%)結果

此外，我們將位置資訊加入機率式潛藏語意分析當中，建立位置性機率式潛藏語意分析(Positional PLSA)，如 表 4-14、表 4-15、表 4-16 與 表 4-17 所示。位置數等於 1 是原始的機率式潛藏語意分析。我們可以發現，於SetET中，相同主題數的情況下，當位置數增加，模型效果的確提升。然而在SetMAT中則效果變差，原因跟機率式潛藏語意分析相同，我們認為是SetMAT語料較少，所以訓練出來的詞機率分布 $P(w|L,T)$ 較差。位置增加，效果變好的原因則在於模型的複雜度增加。如果我們比較相似模型複雜度的模型，例如位置性機率式潛藏語意分析的位置數為 2，主題數為 8，與機率式潛藏語意分析主題數為 16 相比，效果會略差。主要原因是因為位置性機率式潛藏語意分析的模型參數量是 $V\times L\times T+D\times(L+T)$，而機率式潛藏語意分析是 $V\times T+D\times T$，V 是詞典大小，L 是位置數，T 是主題數，D 是訓練文件數。所以位置數為 2，主題數為 8 的位置性機率式潛藏語意分析參數為 $16\times V+D\times 10$，而主題數為 16 的機率式潛藏語意分析參數為 $16\times V+D\times 16$。此外，當位置數等於 3 或 4，主題數等於 8 時的位置性機率式潛藏語意分析亦不如主題數為 16 的機率式潛藏語意分析，我們認為是因為權重 $P(L|D)$ 與 $P(T|D)$ 在即時更新時參數量大小的影響。圖 4-22 是機率式潛藏語意分析與位置性機率式潛藏語意分析示意圖。

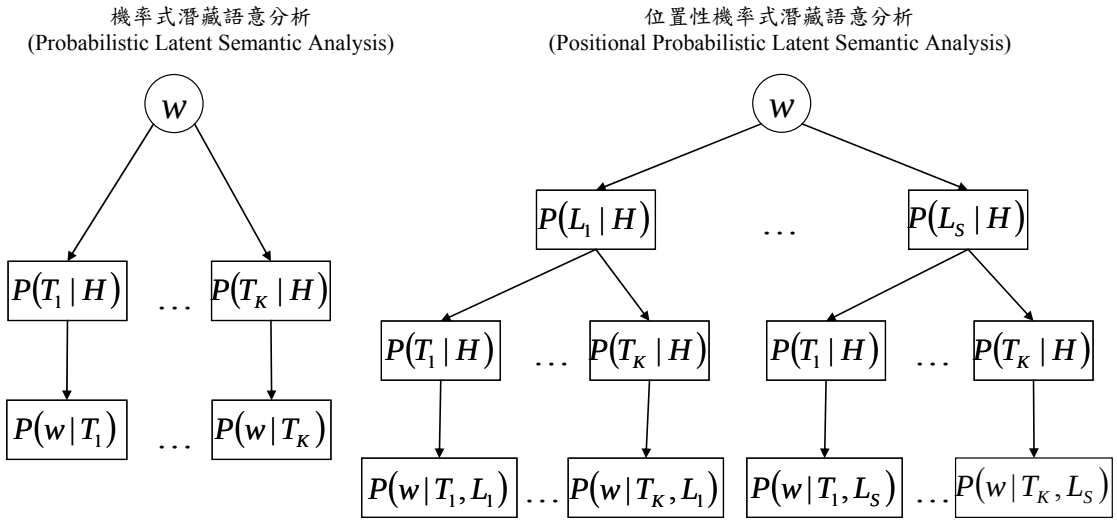


圖 4-22 機率式潛藏語意分析與位置性機率式潛藏語意分析示意圖

表 4-14 SetET 之位置性機率式潛藏語意分析於發展集結果

字錯誤率 (%)		主題數		
		8	16	32
位置	1	20.17	19.81	19.77
	2	19.96	19.75	19.75
	3	19.95	19.74	19.75
	4	19.93	19.78	19.80
語言複雜度		主題數		
		8	16	32
位置	1	579.35	558.49	543.26
	2	556.54	541.58	530.68
	3	551.52	533.87	525.79
	4	546.29	529.16	521.69

表 4-15 SetMAT 之位置性機率式潛藏語意分析於發展集結果

字錯誤率 (%)		主題數		
		8	16	32
位置	1	20.15	20.13	20.06
	2	20.13	20.18	20.14
	3	20.11	20.18	20.07
	4	20.08	20.10	20.16
語言複雜度		主題數		
		8	16	32
位置	1	549.19	540.52	533.07
	2	541.27	532.40	527.00
	3	534.87	529.84	524.05
	4	540.74	547.55	549.69

表 4-16 SetET 之位置性機率式潛藏語意分析於評估集結果

字錯誤率 (%)		主題數		
		8	16	32
位置	1	19.71	19.57	19.63
	2	19.65	19.58	19.49
	3	19.52	19.50	19.47
	4	19.48	19.45	19.42
語言複雜度		主題數		
		8	16	32
位置	1	606.51	585.44	575.07
	2	585.31	569.65	558.64
	3	577.90	562.89	554.52
	4	570.88	557.26	553.35

表 4-17 SetMAT 之位置性機率式潛藏語意分析於評估集結果

字錯誤率 (%)		主題數		
		8	16	32
位置	1	19.76	19.77	19.60
	2	19.76	19.57	19.63
	3	19.73	19.68	19.68
	4	19.69	19.75	19.68
語言複雜度		主題數		
		8	16	32
位置	1	563.70	554.07	545.14
	2	555.97	546.27	538.73
	3	547.90	544.28	537.77
	4	552.22	554.66	557.70

4.4 本章結論

本章主要提出了詞主題混合模型(Word Topical Mixture Model, WTMM)與位置相關語言模型(Position-Dependent Language Model)。詞主題混合模型，以詞為模型單位，透過潛藏的主題分布及機率架構建立詞與詞的關係。於語音辨識中，針對歷史詞序列建立複合式詞模型，能夠表示歷史詞與辨識詞的關聯。我們描述了詳細的模型訓練方式及採用統計方法選擇部分訓練語料，如交互資訊及前向後向二連詞機率等。我們亦從一些角度來分析詞主題混合模型與機率式潛藏語意分析及觸發對語言模型的異同。實驗結果顯示，使用詞主題混合模型的效果能夠與機率式潛藏語意分析及觸發對語言模型相同，甚至更好。而使用統計方法選擇語料亦能夠改善訓練速度而不致於降低模型效果。

詞在文件中的位置資訊可視為文件的樣式，而詞在語句中的資訊則能表現詞性的用法。我們提出了位置相關語言模型，將詞位置資訊整合至現有的模型，如*N*連詞模型和機率式潛藏語意分析等，並分別提出了位置性*N*連詞模型(Positional *N*-gram Model)及位置性機率式潛藏語意分析(Positional Probabilistic Latent Semantic Analysis, Positional PLSA)。透過位置資訊的加入，我們能夠增進語音辨識正確率，並且比使用原始模型效果來的好。表 4-18 是詞主題混合模型與位置相關語言模型的分析，V 是詞典大小，T 是主題數，n 是*N*連數，L 是位置數，D 是訓練文件數。

表 4-18 詞主題混合模型與位置相關語言模型之分析

	模型層次	模型使用	模型複雜度
詞主題混合模型	詞類別	事先訓練	$V \times T + T \times V$
位置性 N 連詞模型	文件位置	可即時調適	$V^n \times L$
位置性機率式潛藏語意分析	位置+主題	需即時調適	$V \times L \times T + (T + L) \times D$