國立臺灣師範大學
資訊工程研究所碩士論文

指導教授： 葉耀明　博士
陳柏琳　博士

以字句擷取為基礎並應用於文件分類
之自動摘要之研究

Research on Sentence Extraction-based Automatic Summarization Applied to Document Classification

研究生：黃耀民　撰

中華民國　九十四　年　六　月

摘要

摘錄式（Extractive）摘要旨在於從原始文件中依據摘要比例自動選取一些重要的字句、段落或章節，並按順序將其形成簡潔摘要。大多數常見的摘要模型原則上可依據其特性分為兩種比對策略。其一，以逐字比對（Literal Term Matching）的方式評估字句與文件的相關性，這其中以向量空間模型（Vector Space Model, VSM）為代表；其二，以概念比對（Concept Matching）的方式評估，這其中以潛藏語意分析（Latent Semantic Analysis, LSA）為代表。

基於這些觀察，在本研究中我們提出數種自動文件摘要的改進方法。在逐字比對上，研究隱藏式馬可夫模型（Hidden Markov Model, HMM），並對其兩種變化（型一及型二）做廣泛的探討。於隱藏式馬可夫模型-型一：視文件為一生成模型（Generative Model），對於每個索引都有一對應的機率分佈，文件與文件中每一字句的相關性，是藉由字句的所有索引，被文件模型生成相似值（Likelihood）的連乘積來決定，換句話說當字句含有較高的相似值，則其與文件的相關性就越高；於隱藏式馬可夫模型-型二：則視文件中每一字句為一機率生成模型，文件中每一字句與文件的相關性，是藉由文件被字句生成的相似值來決定，並且文件中各字句可依據其所產生的相似值作排序。另一方面，在概念比對上，提出兩種摘要模型，分別為嵌入式潛藏語意分析（embedded LSA）與主題混合模型（Topical Mixture Model, TMM）。於嵌入式潛藏語意分析：文件與文件中每一字句同時參與潛藏語意空間的建構，並且字句的重要性可經由適當評估在潛藏語意空間內，其向量表示式與文件的相關性而得；於主題混合模型：文件中每一字句被分別表示成一混合模型，並由 K 個潛藏主題分佈及其相對應特定文件的事後機率所組成，文件中每一字句與文件相關性，即可藉由文件中索引發生在潛藏主題及字句產生各別主題的機率值來評估。我們在中文語音廣播新聞語料庫上執行了一系列的實驗，實驗結果顯示使用隱藏式馬可夫模型或主題混合模型其結果較其它常見方法有顯著的提升，同時主題混合模型在幾乎所有情況下均較隱藏式馬可夫模型

來得佳。

最後，我們也研究摘要模型中主題混合模型在文件分類的適用性，並且文件也能預先經由上述摘要模型做前處理。初步實驗結果顯示，主題混合模型分類器較常見 *K*-最近鄰（*K*-Nearest-Neighbor, *K*NN）分類器在分類的效果上有些微的提升。

關鍵字：摘要、潛藏語意分析、隱藏式馬可夫模型、主題混合模型、
K-最近鄰分類器

Abstract

The purpose of extractive summarization is to automatically select a number of salient sentences, passages, or paragraphs from the original document according to a target summarization ratio and then sequence them to form a concise summary. Most of the conventional summarization models in principle can be characterized by two matching strategies. One is to measure the relevance between the sentence and document by the literal term matching, as exampled by the Vector Space Model (VSM); while the other is to measure such relevance by the concept matching, as exampled by the Latent Semantic Analysis (LSA).

Based on these observations, in the study, we propose several improved approaches for automatic document summarization. For literal term matching, the Hidden Markov Model (HMM) is investigated, and of which two variants (HMM-Type1 and HMM-Type2) were extensively studied. In the HMM-Type1, the document is viewed as a generative model which has the probability distribution over each indexing term. The relevance between the document and each sentence in the document is measured by the product of the likelihoods of all indexing terms of the sentence generated by the document model. In other words, the sentences with higher likelihood scores are more likely to be relevant to the document. In the HMM-Type2, each sentence in the document is viewed as the probability generative model instead. The relevance between the document and a given sentence is measured by the likelihood that the document is generated by the sentence, and the sentences in the document can be ranked according to the respective likelihood scores they generate. On the other hand, for concept matching, two summarization models, i.e., the embedded Latent Semantic Analysis (embedded LSA) and the Topical Mixture Model (TMM), were proposed. In the embedded LSA, the document to be summarized is

also involved in the construction of the latent semantic space, and then the importance of each sentence can be properly measured by the proximity of its vector representation to that of the document in the latent semantic space. While in the TMM, each sentence in the document is respectively taken as a mixture model consisting of *K* latent topical distributions and their corresponding document-specific posterior probabilities. The relevance between each sentence in the document and document itself is then measured based on the likelihoods of the indexing terms of the document observed in the latent topics and the probabilities that the sentence generates the respective topics. A series of experiments has been conducted on the Mandarin broadcast news speech and the experimental results show that the performance achieved by using either the HMM or the TMM is significantly better than that of the conventional approaches, and the TMM further outperforms the HMM in almost all conditions.

Finally, we also study the possibility to adapt the TMM summarization model to document classification, for which the documents also may be preprocessed beforehand by the summarization models mentioned above. The initial experimental results show that the TMM classifier yields slightly superior performance when compared to the conventional *K*-Nearest-Neighbor classifier.

Keywords：Summarization, Latent Semantic Analysis, Hidden Markov Model, Topical Mixture Model, *K*-Nearest-Neighbor Classifier

誌謝

首先感謝父母及所有家人的支持，能夠順利完成學業、拿到碩士學位，全靠你們栽培與照顧。家人的支持是我最大的後盾與動力，除了供應我生活花費，更讓我無後顧之憂、能全力鑽研於知識的學習。

其次，我要感謝葉耀明老師與陳柏琳老師，您們嚴謹的治學態度、淵博的知識、廣闊的胸懷，讓學生印象深刻、受益匪淺。從老師的指導中學到正確做研究的方法與態度，及解決與分析問題的技巧；此外，也感謝碩一時林順喜老師的指導，從您身上學到了不少發現問題及做研究的方法。

接著，感謝口試委員鄭永斌博士、葉慶隆博士與陳永昇博士對於學生論文的殷切指正，使我得以整理我的思維，彌補不足之處，使論文更臻完善。

再來感謝韋金、書維、立德，與你們同窗的時光是研究中美好回憶；感謝人瑋、文鴻、信維、惠銘、志豪、成韋、怡婷、燦輝、士弘、炫盛，每次與你們討論都能開闊我的視野，從中學習到不同領域的知識；感謝善泰學長、士傑學長、德義、耀才、漢昇、淑琼、慧雯、燕玲、美君、怡萍、宜珍，在學習路途上的陪伴，豐富了研究的生活；謝謝明正學長、耀明學長、金麟學長、秉鋒學長、榮泰學長、嘉漢學長、恬欣學姐，平日的諸多指教，並學習你們勤奮的精神，成為督促自己的動力之一；謝謝慶全、繡如、立平、俊卿、才業、志彬、士翔、宜達、鼎元、智翔，和你們相處是件美好的事。

研究之路，有太多令人難忘的回憶，我將一點一滴收藏在人生的百寶箱中，時時回味、細細品嚐。

謹將這本論文獻給所有關心我的人。

謝謝大家 耀民謹誌

目錄

表目錄

圖目錄

第 1 章 緒論

1.1 研究動機與目的

隨著資訊爆炸時代的來臨，人們希望以更高的效率與效能取得資訊，其中自動摘要技術與其後衍生的分類應用，是重要的關鍵技術之一。例如，Google [Google] 利用網頁片段權充摘要、報章雜誌的標題目錄等，這些眾多且未經分析整理的資訊，經過擷取、分析與整理後便成為高質量的資訊，給人們有效的閱讀與吸收。

本論文旨在探討自動摘要模型及分類器模型。摘要後的文件可視為一種特徵選取的前處理，透過前處理可以將重要的資訊摘選出來並減少分類時的運算量，協助分類器做更精確的分類。

1.2 自動摘要

自動摘要技術的目標是依據使用者的需求，將文件縮減濃縮成一或數句重要字句，好讓人們更快速、更方便的得到所需資訊，其優點在於：

節省時間： 使用者不需瀏覽整篇文件即可瞭解文意

加速瀏覽： 在查詢結果中呈現摘要，可方便使用者快速決定所需資訊

協助分類： 摘要過的資訊，可做為分類器的分類依據

節省人力： 自動摘要的產生，不需透過人力介入

自動摘要可以分類如下 [Hovy and Marcu 1998]：

1. 根據形成方式可分類為摘錄式（Extractive）摘要與非摘錄式（Non-extractive or Abstract）摘要。摘錄式摘要是找出文件中重要的字句、段落或章節來組成摘要；非摘錄式摘要則重寫字詞、片語來形成摘要。
2. 根據性質可分類為資訊性（Informative）摘要與指示性（Indicative）摘要。資訊性摘要是從文件中找出所有重要的資訊，科技論文的摘要即為一例；而指示性則偏向於提供文件分類上的資訊，例如圖書館內所用的分類卡。
3. 根據需求可分類為一般性（Generic）摘要與以需求為基礎（Query-based）摘要。

一般性摘要對文件內不同主題視為同等重要；以需求為基礎的摘要則傾向於顯示使用者要求的部份。

4. 根據文件來源可分類為單一文件(Single Document)摘要與多文件(Multidocument)摘要。單一文件摘要是從一篇文件中截取重要資訊；多文件摘要則歸納主題相近的文件共同產生摘要，或指對同一主題但時間先後不同文件進行摘要。
5. 根據語言可分類為單一語言（Monolingual）摘要與多語言（Multilingual）摘要。多語言摘要係從多種語言的文件中產生單一語言的摘要結果，其中牽涉到機器翻譯的技術。

大多數常見的摘要模型原則上可依據其特性分為兩種比對策略。其一，以逐字比對（Literal Term Matching）的方式評估字句與文件的相關性，愈高相關性的字句代表愈重要，這其中以向量空間模型（Vector Space Model, VSM）為代表 [Gong and Liu 2001; 何遠 2003]；其二，以概念比對（Concept Matching）的方式評估，這其中以潛藏語意分析（Latent Semantic Analysis, LSA）為代表 [Gong and Liu 2001; 葉鎮源 2002; 黃建霖 2004; Hirohata *et al.* 2005] 。

本論文針對的是摘錄式、資訊性、一般性、單一文件、單一語言摘要模型做研究，並從逐字比對與概念比對兩個方向作探討，希望能發展出適合的自動摘要模型以供中文自動摘要的產生。

1.3 研究成果

本論文於自動摘要方面，在逐字比對方式上應用隱藏式馬可夫模型（Hidden Markov Model, HMM）做為摘要模型，並分為 HMM-Type1 及 HMM-Type2 二種類型；在概念比對上提出嵌入式潛藏語意分析（embedded LSA）與主題混合模型（Topical Mixture Model, TMM）做為摘要模型；在自動摘要評估上，提出以改良型字錯誤率（modified Character Error Rate, m-CER）為基礎的平均精確度（Mean Average Precision, MAP）評估方式，以解決自動轉寫與人工轉寫文件因斷句不一致，所造成摘要結果無法評估相關的問題。

經由實驗結果顯示，於摘要模型比較上：使用隱藏式馬可夫模型或主題混合模型其結果較其它常見方法有顯著的提升，同時主題混合模型在幾乎所有情況下均較隱藏式馬可夫模型來得佳；於特徵單位比較上：使用雙音節與雙字時，其結果優於使用詞為特徵單位。

最後，我們也研究摘要模型中主題混合模型在文件分類的適用性，並且文件也能預先經由上述摘要模型做前處理。初步實驗結果顯示，主題混合模型分類器較常見 *K*-最近鄰（*K*-Nearest-Neighbor, *K*NN）分類器在分類的效果上有些微的提升。

1.4 章節安排

本篇論文的章節安排如下：

第二章簡介本論文的理論背景，包括向量空間模型、相關評估、潛藏語意分析、馬可夫模型、隱藏式馬可夫模型、統計式語言模型與主題混合模型。

第三章介紹本論文所提出的摘要模型，包括嵌入式潛藏語意分析、隱藏式馬可夫模型-型一、隱藏式馬可夫模型-型二、主題混合模型。

第四章說明本論文的實驗設定，並利用餘弦、ROUGE、平均精準度三種自動摘要評估方法做實驗，對實驗結果做一分析。

第五章概述分類器模型與提出主題混合模型分類器，並介紹實驗語料。在實驗結果上，比較主題混合模型分類器和常見 *K*-最近鄰分類器的實驗結果，並分析自動摘要是否有助於分類器做更精準的分類。

第六章對本論文的主要成果做一總結，並提出結論與未來研究方向。

第 2 章 相關文獻探討

自動摘要不是一種新興觀念，在 1950 年代至 1960 年代學者們就已開始在這方面的研究。因為當時尚無較大的文字語料集，在自然語言處理上也無較為成熟的統計模型，再加上電腦的計算能力及記憶體的容量也有所限制，因此當時的研究重點在於精簡的流程處理，著重於下列技術 [Luhn 1959；Edmundson 1969] ：

- 文字所在的位置：位於重要段落的字句佔有較高的權重，如第一段或者位於標題如『簡介、目的、結論』的段落，被視為重要
- 語彙的隱含：字句中包含重要字詞，如『重要的、艱難的』等主題句
- 位置：每一段落的第一句和最後一句被視為重要

雖然上述的方法有效，然而它們非常依賴於特別的寫作格式與風格。例如利用第一段形成摘要，僅在新聞及新聞雜誌類型的文件中適用。是以本研究試圖能發展對不同文件類型皆能通用的自動摘要模型，並不專注於特定文件的寫作方式與風格；換言之，本論文希望所探討的摘要模型，能經過一些處理（訓練）進而能自動獲得這方面的資訊，如構成摘要的重要語彙。

在回顧自動摘要模型上，可以發現其技術裡的許多重要觀念來自於資訊檢索（Information Retrieval, IR），此外資訊檢索上許多成功的檢索模型，也被驗證同樣適用於自動摘要上，如向量空間模型、潛藏式語意分析模型等 [Gong and Liu 2001; 葉鎮源 2002; 何遠 2003; 黃建霖 2004; Hirohata *et al.* 2005] 。

資訊檢索處理的問題是如何依使用者的問句（Query）從大量的文件中找出相關（即符合使用者需求）的文件；而自動摘要常常假定使用者的需求為 『看看文件中的最重要的部分是什麼?』 來找出與文件最相關的字句。資訊檢索與自動摘要的比較，可由圖 2.1 所示。

以下小節介紹幾個自動摘要中常用且來源於資訊檢索的觀念。

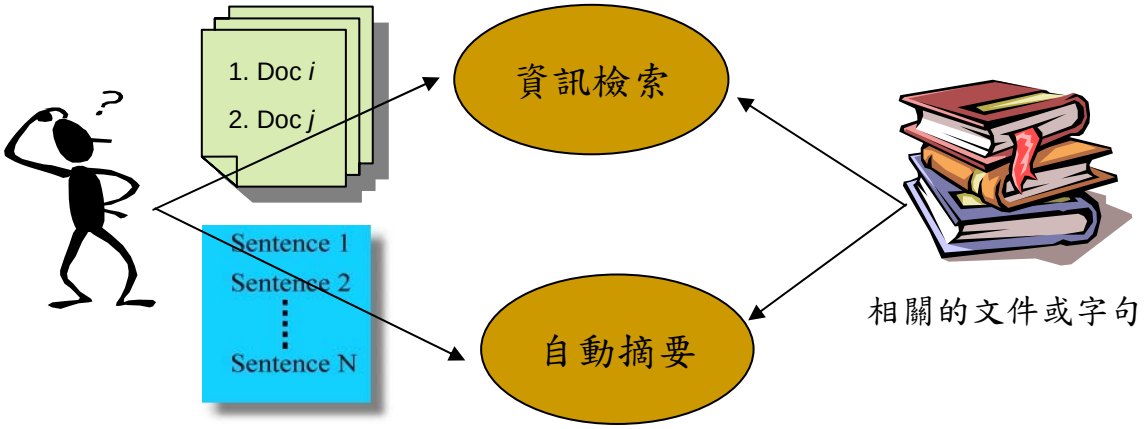


圖 2. 1 資訊檢索與自動摘要比較圖

2.1 向量空間模型（Vector Space Model, VSM）

在資訊檢索（Information Retrieval, IR）領域中，向量空間模型是個典型的檢索模型 [Baeza-Yates *et al.* 1999]。其將每一篇文件 d_j 與問句 q 視為一 T-維的向量（T 是索引特徵的總數）：

$$\vec{d_j} = (w_{1,j}, w_{2,j}, ..., w_{T,j}) \tag{2. 1}$$

$$\vec{q} = (w_{1,q}, w_{2,q}, ..., w_{T,q}) \tag{2. 2}$$

向量的權重 $w_{i,j}$ 代表索引特徵 i 在文件 d_j 的權重，其計算常使用 詞頻-反文件頻（Term Frequency-Inverse Document Frequency, TF-IDF）乘積來表示。詞頻統計其出現的頻率來決定其重要性，越常出現愈重要，其值經由正規化算出；反文件頻用以決定一索引特徵是否具有鑑別力，如一索引特徵在每一篇文件都存在（如中文：的、了），則應降低其權重，上述討論可由以下數學式來表示：

$$w_{i,j} = tf_{i,j} * idf_i = \frac{freq_{i,j}}{\max_h freq_{h,j}} * \log\frac{N}{n_i} \tag{2. 3}$$

其中

$freq_{i,j}$ ：索引特徵 i 在文件 d_j 中出現的次數

N ：文件集總數

n_i ：索引特徵 i 在文件集中出現的文件數

最後每一篇文件 d_j 與問句 q 經由估測兩向量的餘弦(Cosine)值來決定其相關性：

$$sim(d_j, q) = \frac{\overrightarrow{d_j} \cdot \vec{q}}{\left|\overrightarrow{d_j}\right| \times \left|\vec{q}\right|} = \frac{\sum_{i=1}^{T} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{T} w^2_{i,j}} \times \sqrt{\sum_{i=1}^{T} w^2_{i,q}}} \quad (2.4)$$

近年來有學者應用向量空間模型於自動摘要上，其拿整篇文件做問句（Query）去檢索文件中的每一字句，得到一相關度排名（句排名），並依摘要比例將字句摘錄出來形成摘要 [何遠 2003]。

2.2 相關評估（Relevance Measure, RM）

Gong 提出使用相關評估的方法來產生摘要 [Gong and Liu 2001]，其方法主要以向量空間模型為基礎，試圖找出文件中不同主題的重要字句為標的，其步驟如下：

1. 將文件 D 斷句，$D = \{S_1, S_2, ..S_i.., S_N\}$，這些字句 S_i 用來組成候選句 S
2. 對於每一字句 S_i 產生詞頻（Term-Frequency, TF）向量 $\overrightarrow{S_i}$，以及對於整篇文件 D 的詞頻向量 $\overrightarrow{D}$
3. 對於 S 中每一字句 S_i，估測 $\overrightarrow{S_i}$ 與 $\overrightarrow{D}$ 之間的相關分數（餘弦分數）
4. 選取最大相關分數的字句 S_k，並將其置於摘要中
5. 將 S_k 自 S 中移除，並將 S_k 中所含的字詞自文件 D 中移除；並重新計算向量 $\overrightarrow{D}$
6. 如摘要的字句達到摘要比例的量則終止運算，否則回到步驟 3 執行

本方法在步驟 4 中，選取文件中最大的相關分數的字句，代表其含有文件的主要意涵。為了使相關分數所選取到的摘要可覆蓋整篇文件的主要主題，是以在步驟 5 去除第 k 句中所含的字詞，讓接下來所選取的字句與第 k 句具有最小重覆，此傾向於所摘要的字句間具有最小的重覆。

2.3 潛藏語意分析（Latent Semantic Analysis, LSA）

潛藏語意分析 [G. Furnas *et al.* 1988; Bellegarda 2000] 是基於線性代數方法為核心的模型，包括了奇異值分解（Singular Value Decomposition, SVD）與維度約化（Dimension Reduction）兩個處理過程。LSA 的應用非常廣泛，諸如同義詞建構、判斷字詞與字句間的關係、跨語言語言模型調適（Language Model Adaptation）[KIM *et al.* 2004]、與自動摘要 [Gong and Liu 2001; 葉鎮源 2002] 等。

2.3.1 索引與字句矩陣

在進行奇異值分解之前，要將文件轉換成 索引-字句矩陣（Term-Sentence Matrix）。假設一篇文件中不同的索引字或詞有 M 個，此外文件可斷句成 N 句。

所以 索引-字句矩陣 A 的維度是 $M \times N$，矩陣中每個元素 w_{ij} 的值，可使用對數-熵（Log-Entropy）來計算 [Bellegarda 2000; Giles *et al.* 2003] ：

$$w_{ij} = l_{ij} \times g_i \tag{2.5}$$

l_{ij} 代表索引 i 在字句 j 的對數權重，g_i 代表索引 i 的熵權重：

$$l_{ij} = \log\left(1 + f_{ij}\right) \tag{2.6}$$

$$g_i = 1 - \varepsilon_i \tag{2.7}$$

其中

f_{ij}：索引 i 在字句 j 中出現的次數

p_{ij}：索引 i 在字句 j 中的機率值 $= \dfrac{f_{ij}}{\sum_{j=1}^{N} f_{ij}}$

ε_i：索引 i 在字句中的正規化熵值 $= -\dfrac{1}{\log N}\sum_{j=1}^{N} p_{ij} \log p_{ij}$，即 $0 \le \varepsilon_i \le 1$。$\varepsilon_i$ 越接近 0 代表索引 i 在越少的字句中出現，越具有鑑別力

N ：字句數

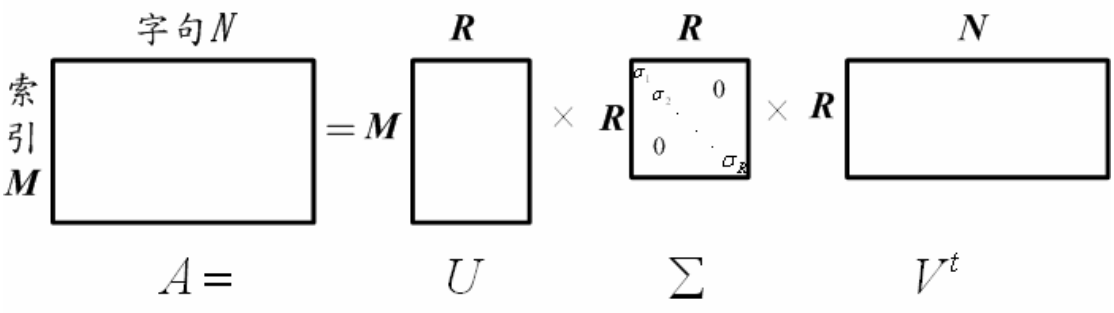


圖 2.2 奇異值分解圖示

使用 Log-Entropy 權重的方式在大部份潛藏語意分析為基礎的實驗中皆有不錯的效果 [Berry and Browne, 1999; Bellegarda 2000]。

2.3.2 奇異值分解（Singular Value Decomposition, SVD）

建立好索引-字句矩陣之後，便可進行奇異值分解：

$$A = U \sum V^T \tag{2.8}$$

其中 $\sum$ 是 $R \times R$ 維的對角奇異值矩陣，U 是 $M \times R$ 維的左奇異向量（Left Singular Vector）矩陣，V^T 是 $R \times N$ 維的右奇異向量（Right Singular Vector）矩陣，奇異值分解如圖 2.2 所示。

經過奇異值分解後，索引和字句都被投影到新的空間，稱為潛藏語意空間（LSA Space），此空間的維度是 R 維，R 小於 M 與 N。換句話說，奇異值分解透過降維的方式，將在高維度（M 與 N）內互不相關的索引與字句，投影到低維度的同一空間內，如此即可於潛藏語意空間估測其相關性。

2.3.3 潛藏語意分析摘要模型

Gong 提出應用潛藏語意分析於摘要模型上 [Gong and Liu 2001]，其方法如下：

1. 將文件 D 斷句，$D = \{S_1, S_2, ..S_i.., S_N\}$，這些字句 S_i 用來組成候選 S，設 k=1
2. 由 D 建立 索引×字句矩陣 A
3. 對 A 進行奇異值分解，在右奇異向量（Right Singular Vector）矩陣 V^t 中，

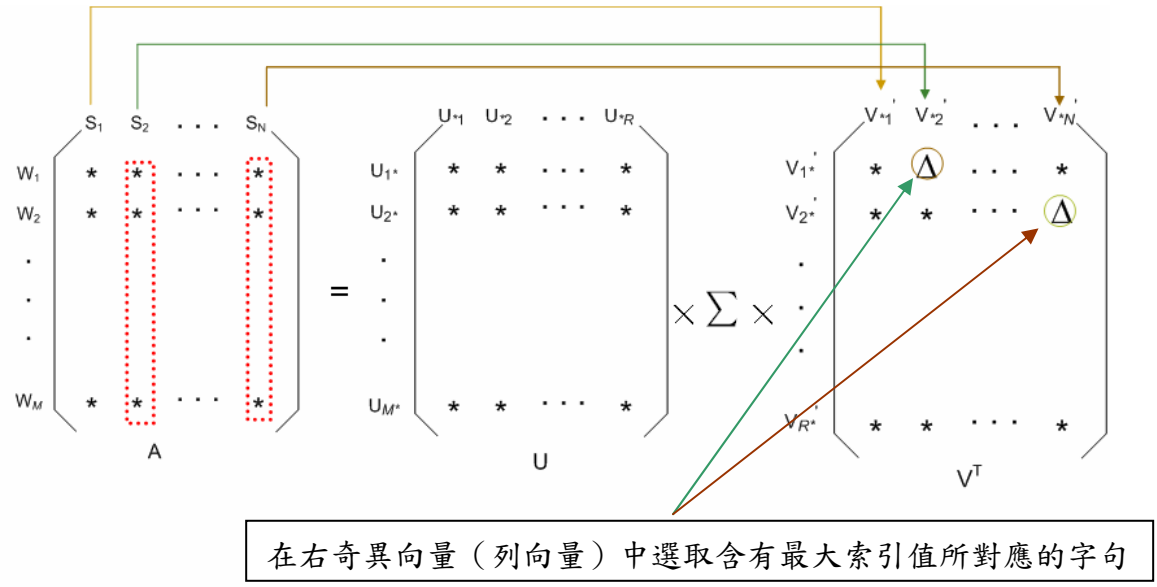


圖 2.3 潛藏語意分析摘要模型示意圖

每一字句 S_i 可由 V^t 中的行向量 $[v_{i1}, v_{i2}, ..., v_{iR}]^T$ 表示

4. 在 V^t 中選取第 k 個右奇異向量（列向量）
5. 由上述向量中，選取含有最大索引值所對應的字句，將其加入摘要中
6. 如 k 達到摘要比例的量則終止運算，否則將 k 加 1，並執行步驟 4

此方法假設，每一奇異值分別代表一概念或主題。是以奇異值所對應到的列向量（V^t 中某一列），用以描述各字句所能表達的概念或主題。因奇異值矩陣 $\sum$ 是經由遞減排序，是以第 k 個所選取的列向量，代表第 k 名重要的概念或主題，而其含有最大索引值所對應的字句就代表第 k 名重要的字句；且每一右奇異向量是相互獨立，是以所選取的字句間具有最小的重覆，如圖 2.3 所示。

A 代表原索引-字句矩陣，$\sum$ 是 $R \times R$ 維的對角奇異值矩陣、U 是 $M \times R$ 維代表索引在此語意空間的表示法、V^T 是 $R \times N$ 維代表字句在此語意空間的表示法。如在 V^T 的第 1 個右奇異向量（列向量），以第 2 個索引值為最大，是以將其所對應原始文件 D 中的第 2 句加入摘要；同理，第 2 個右奇異向量，加入第 N 句。

2.4 馬可夫模型（Markov Model）

隱藏式馬可夫模型是由馬可夫模型演變而來，根據 [Rabiner *et al.* 1989] 馬可夫

模型之相關定義，如下所示：

定理 1：若隨機過程（Stochastic Process）$\{S_t, t \geq 0\}$ 中，第 $t+1$ 的時間狀態只和第 t 的時間狀態有關，並與之前的時間狀態無關：

$$p\{S_{t+1} = s_{t+1} \mid S_0 = s_0, S_1 = s_1, ..., S_t = s_t\} = p\{S_{t+1} = s_{t+1} \mid S_t = s_t\} \quad (2.9)$$

則稱這個隨機過程為一階馬可夫鏈（First Order Markov Chain），此乃馬可夫模型中最簡單的模型。

一階馬可夫鏈在 N 個狀態下，可用三個元素來表示 (S, A, Π)

- S 表示所有狀態的集合，$S = \{s_1, s_2, ..., s_N\}$，其中 N 為狀態的個數
- $A = (a_{ij})$ 代表狀態轉移機率矩陣，$a_{ij} = p\{S_{t+1} = s_j \mid S_t = s_i\}$，$1 \leq i, j \leq N$ 表示從狀態 i 跳到狀態 j 的機率，且必須滿足 $a_{ij} \geq 0, \sum_{j=1}^{N} a_{ij} = 1$
- $\Pi = (\pi_i)$ 代表狀態初始的機率向量 $\pi_i = p(S_1 = s_i)$，$1 \leq i \leq N$ 表示在 t=1 時，狀態為 i 的機率，且需滿足 $\sum \pi_i = 1$ 的條件

若馬可夫鏈中每一時間的可能狀態均來自一有限集合 $S = \{s_1, s_2, ..., s_N\}$，則稱之為有限狀態馬可夫鏈（Finite State Markov Chain）。

定理 2：若隨機過程 $\{S_t, t \geq 0\}$ 的轉移機率 a_{ij} 不隨時間改變，也就是說滿足性質：

$$p\{S_{t+1} = s_j \mid S_t = s_i\} = p\{S_2 = s_j \mid S_1 = s_i\} = a_{ij} \quad (2.10)$$

則稱為穩定型之有限狀態馬可夫鏈（Stationary Finite State Markov Chain）。

滿足上述定理 1 與定理 2 的隨機過程即可稱之為馬可夫鏈或具有馬可夫之性質。

2.5 隱藏式馬可夫模型（Hidden Markov Model, HMM）

隱藏式馬可夫模型最早是由 Baum 和 Petrie 在 1966 年所發展出來 [Baum *et al.* 1966]，其植基於統計的機率模型，並於近十幾年來逐漸被廣泛應用，概因其擁

有豐富的數學架構及基礎能夠成功地解決所欲處理的問題。

目前，除了被廣泛應用在語音辨識（Speech Recognition）[Rabiner *et al.* 1989] 、自然語言 [Thede *et al.* 1999] 處理，甚至被應用於影像處理（Image Processing）之分析 [Aas *et al.* 1999] 與網路通訊 [Salamatian *et al.* 2001] 上。

根據 [Rabiner *et al.* 1989]對離散型隱藏式馬可夫模型之定義為：它是一個雙層隨機程序，包含了隱藏的狀態層和可觀察的輸出層；隱藏層無法直接觀察，但可從另一能產生輸出序列之輸出層觀察得出。

隱藏式馬可夫模型在 N 個狀態下，可用四個元素來表示(S,Π, A,B)

(一) 符號的表示意義如下：

- S 表示所有狀態的集合，$S=\{s_1, s_2, ..., s_N\}$，其中 N 為狀態的個數
- $V=\{v_1, v_2, ..., v_M\}$代表可觀察輸出的集合，其中 M 為所有可能輸出符號的數目
- $O=\{o_1, o_2, ..o_t.., o_T\}$表示可觀察的輸出序列，$o_t$ 代表在時間 t 下，對於任一狀態所有可能產生的觀察輸出符號，且需滿足$o_t \in V$

(二) 模型的參數為：

- $\Pi=(\pi_i)$代表狀態初始的機率向量$\pi_i = p(S_1 = s_i)$, $1 \le i \le N$ 表示在 t=1 時，狀態為 i 的機率，且需滿足$\sum_{i=1}^{N} \pi_i = 1$的條件
- $A=(a_{ij})$代表狀態轉移機率矩陣，$a_{ij} = p\{S_{t+1} = s_j \mid S_t = s_i\}$, $1 \le i, j \le N$ 表示從狀態 i 跳到狀態 j 的機率，且必須滿足$a_{ij} \ge 0, \sum_{j=1}^{N} a_{ij} = 1$
- $B=\{b_j(k)\}$ 代表可觀察輸出矩陣 $b_j(k) = p\{o_t = v_k \mid S_t = s_j\}$, $1 \le j \le N$, $1 \le k \le M$，表示在狀態為 j 時，v_k 的發生機率，且滿足 $\sum_{k=1}^{K} b_j(k) = 1$

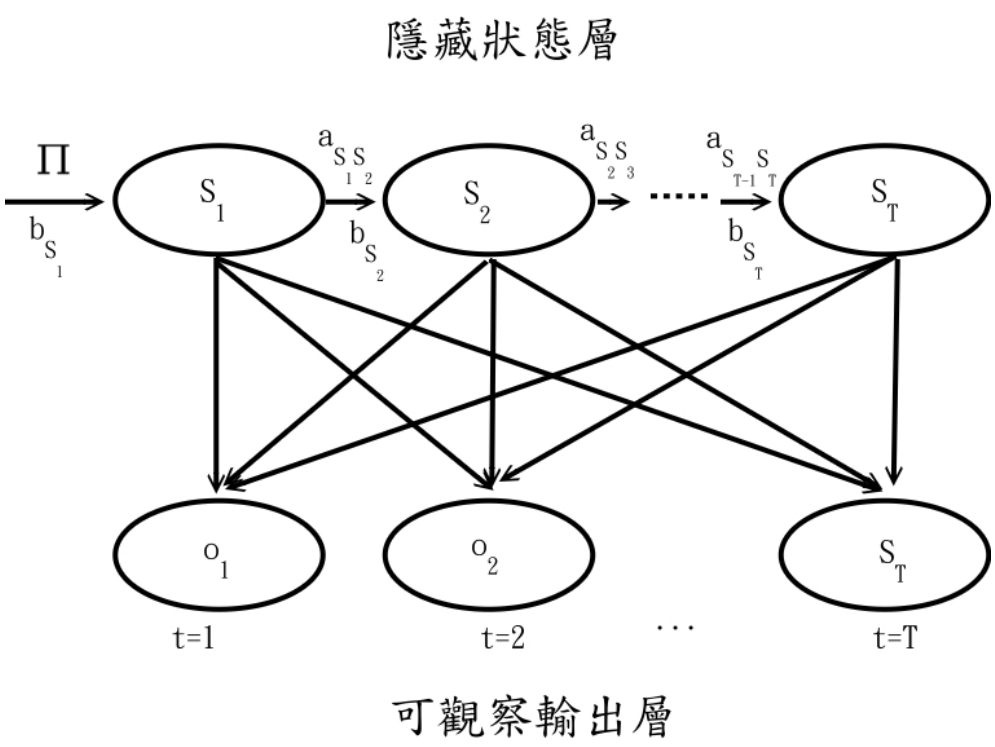


圖 2.4 隱藏式馬可夫示意圖

當適當地決定Π、A 和 B 時，隱藏式馬可夫模型的產生過程、運作方式如下：

1. 根據起始狀態機率分佈Π決定 S_1
2. 設定 t=1。
3. 由 $b_{S_t}(k)$ 的機率分佈產生 o_t
4. 由狀態轉移機率矩陣 $a_{S_tS_{t+1}}$ 的機率分佈決定 S_{t+1}
5. 設定 t=t+1，當 t<T 時回到步驟 3，否則結束程式。

上述的步驟，可由圖 2.4 所示。

2.6 統計式語言模型（Statistical Language Model, SLM）

以統計式語言模型（Statistical Language Model, SLM），來觀察字詞間可能相接的情形，已被廣泛於語音辨識器上 [Rosenfeld 2000; Siivola *et al.* 2001]，給定一長度為 n 之詞串 W，$W = w_1, w_2, ..., w_n$，要估測 W 的機率，$P(W)$，可以利用連鎖律（Chain Rule）將其分解：

$$
\begin{aligned}
P(W) &= P(w_1)P(w_2,...,w_n \mid w_1) \\
&= P(w_1)P(w_2 \mid w_1)P(w_3,...,w_n \mid w_1,w_2) \\
&= \prod_{i=1}^{n} P(w_i \mid w_1,...,w_{i-1}) \\
&= \prod_{i=1}^{n} P(w_i \mid h_i)
\end{aligned}
\tag{2. 11}
$$

其中 h_i 是詞 w_i 的歷史詞串（history），$h_i = w_1,...,w_{i-1}$。

假設 $|V|$ 為詞典大小，則式(2.11)中 $P(w_i \mid h_i)$ 的 w_i 與歷史詞串 h_i 之參數量為 $|V|^i$，此為一極其龐大的計算量而無法估測，勢必要做簡化。是以 *N*-連語言模型廣泛的被使用來處理這個問題，N 連語言模型是帶入 $N-1$ 階馬可夫模型假設，即假設詞 w_i 的出現只與其前面 $N-1$ 個詞有關聯，而與 $N-1$ 個詞以前的詞沒有關聯，所以式(2.11)可以改寫成：

$$
P(W) = \prod_{i=1}^{n} P(w_i \mid h_i) = \prod_{i=1}^{n} P(w_i \mid w_{i-N+1},...,w_{i-1}) \tag{2. 12}
$$

如三連語言模型（Tri-gram Language Model）可表示成

$$
P(W) = \prod_{i=1}^{n} P(w_i \mid w_{i-2}, w_{i-1}) \tag{2. 13}
$$

要估測式(2.13)中的 $P(w_i \mid w_{i-2}, w_{i-1})$ 可使用最大相似度估測法（Maximum Likelihood Estimation, MLE）得到：

$$
P(w_i \mid w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})} \tag{2. 14}
$$

$C(w_{i-2}, w_{i-1}, w_i)$ 與 $C(w_{i-2}, w_{i-1})$，分別為 w_{i-2}, w_{i-1}, w_i 同時出現的次數與 w_{i-2}, w_{i-1} 同時出現的次數

2.7 主題混合模型（Topical Mixture Model, TMM）

主題混合模型最早由 [Chen *et al.* 2004b; Chen 2005] 所提出並使用於語音文件檢索上。在傳統資訊檢索上，給定一使用者查詢 Query $Q = q_1 q_2 .. q_n .. q_N$，一文件 D_i 可根據其機率 $p(D_i \mid Q)$，得到相關程度的排名，經貝式定理可表示為：

$$
p(D_i \mid Q) = \frac{p(Q \mid D_i)p(D_i)}{p(Q)} \tag{2. 15}
$$

$p(Q|D_i)$ 是文件 D_i 產生查詢 Q 的機率，$p(D_i)$ 是文件 D_i 相關的事前機率，$p(Q)$ 是查詢 Q 的事前機率（Prior Probability）。對於所有文件來說 $p(Q)$ 是相同的且不影響文件的排名，是以可省略。於外，估計 $p(D_i)$ 的機率仍然未知，是以可進一步簡化假設 $p(D_i)$ 是均勻分佈（Uniform Distribution），也就是對於所有的文件是相同的 [Miller *et al.* 1999]。如此便可藉由 $p(Q|D_i)$ 來近似 $p(D_i|Q)$。

另一方面，假設查詢 $Q = q_1 q_2 .. q_n .. q_N$ 中，每個查詢項的發生互為獨立事件，因此估測 $p(Q|D_i)$ 可視為查詢 Q 中每一查詢項 q_n 於文件 D_i 機率分佈的連乘積，數學式如下：

$$p(Q|D_i) = \prod_{n=1}^{N} p(q_n|D_i) \tag{2.16}$$

在此研究中，每一篇文件 D_i 可被詮釋為混合模型（Mixture Model），模型中定義 K 個潛藏主題，各由一個主題單連語言模型（Topical Unigram）所表示，且每一潛藏主題在各文件都有不同的權重。換句話說，每一篇文件可以產生許多主題，每個主題都有相對應的單連語言模型，因此查詢 Q 與每一文件 D_i 的相關程度，可進一步改寫為：

$$p(Q|D_i) = \prod_{n=1}^{N} \sum_{k=1}^{K} p(q_n|T_k) p(T_k|D_i) \tag{2.17}$$

$p(q_n|T_k)$ 指特定潛藏主題 T_k 產生查詢項 q_n 的機率，$p(T_k|D_i)$ 是潛藏主題 T_k 在文件 D_i 的權重，且須滿足 $\sum_{k=1}^{K} p(T_k|D_i) = 1$ 的限制。總結來說，主題混合模型的主題單連語言模型，$p(q_n|T_k)$，是經由整個文件集訓練而來，且每一潛藏主題 T_k 在各文件 D_i 都有其所屬的權重，$p(T_k|D_i)$，如圖 2.5 所示。

在主題混合模型中，不同於逐字比對（Literal Term Matching），如向量空間模型，是以查詢 Q 中每一查詢項 q_n 在文件 D_i 出現的次數做計算，在主題混合模型中是以 q_n 發生在主題 T_k 與文件 D_i 產生主題 T_k 的機率來表示。是以即使查詢項並未出現在文件 D_i 中，經由主題混合模型還是可以給予 $p(Q|D_i)$ 較高的值，而達到概念比對的目的。

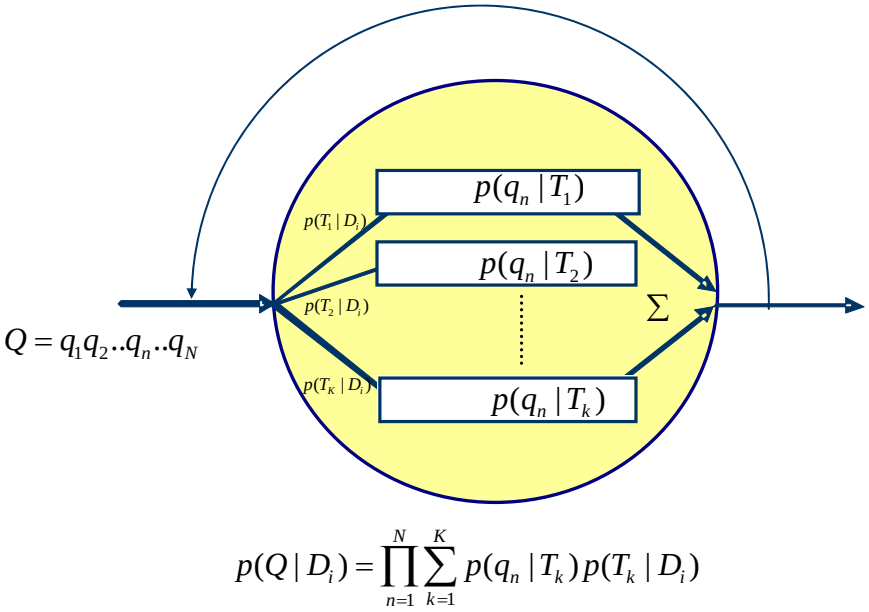


圖 2. 5 主題混合模型示意圖

2.7.1 主題混合模型訓練

在訓練時，*K*-means 演算法 [Ball and Hall 1967; Duda and Hart 1973] 被用來事先切割整個文件集為 K 個潛藏主題。因此，對於每一潛藏主題，其初始的主題單連語言模型可用主題所包含的文件來估測；而其在每一文件 D_i 的權重，可由與中心 C_k 的鄰近程度來估計，如下所示：

$$p(T_k \mid D_i) = \frac{R(\overrightarrow{D_i}, \overrightarrow{C_k})}{\sum_{r=1}^{K} R(\overrightarrow{D_i}, \overrightarrow{C_r})} \tag{2. 18}$$

其中 $R(\overrightarrow{D_i}, \overrightarrow{C_k})$ 代表利用餘弦估測文件 D_i 與中心 C_k 的距離，如下所示：

$$R(\overrightarrow{D_i}, \overrightarrow{C_k}) = \frac{\overrightarrow{D_i} \cdot \overrightarrow{C_k}}{\left\|\overrightarrow{D_i}\right\| \times \left\|\overrightarrow{C_k}\right\|} \tag{2. 19}$$

2.7.1.1 主題混合模型訓練—監督式

更進一步來說，主題單連語言模型與其在各文件的權重，可使用期望值最大化（Expectation-Maximization, EM）演算法來優化此二者的機率分佈 [Dempster *et al.* 1977]。給定一訓練集，如每一查詢 Q 均有與其相關文件的資訊，則主題混合

模型可迭代更新，利用下面三個公式：

$$\hat{p}(q_n \mid T_k) = \frac{\sum_{Q \in [TrainSet]_Q} \sum_{D_i \in [Doc]_{R\ to\ Q}} n(q_n, Q)\, p(T_k \mid q_n, D_i)}{\sum_{Q \in [TrainSet]_Q} \sum_{D_i \in [Doc]_{R\ to\ Q}} \sum_{q_s \in Q} n(q_s, Q)\, p(T_k \mid q_s, D_i)} \tag{2. 20}$$

$$\hat{p}(T_k \mid D_i) = \frac{\sum_{\substack{Q \in [TrainSet]_Q \\ st.\ D_i \in [Doc]_{R\ to\ Q}}} \sum_{q_s \in Q} n(q_s, Q)\, p(T_k \mid q_s, D_i)}{\sum_{\substack{Q \in [TrainSet]_Q \\ st.\ D_i \in [Doc]_{R\ to\ Q}}} |Q|} \tag{2. 21}$$

$$p(T_k \mid q_n, D_i) = \frac{p(T_k \mid D_i)\, p(q_n \mid T_k)}{\sum_{l=1}^{K} p(T_l \mid D_i)\, p(q_n \mid T_l)} \tag{2. 22}$$

其中，$[TrainSet]_Q$ 是查詢範例的訓練集合，$[Doc]_{R\ to\ Q}$ 是與特定查詢範例 Q 相關的文件集合，$n(q_n, Q)$ 是每一查詢項 q_n 出現在查詢範例 Q 的次數，$|Q|$ 是查詢範例 Q 的長度，$Q \in [TrainSet]_Q \ \ st.\ D_i \in [Doc]_{R\ to\ Q}$ 表示查詢範例 Q 滿足 D_i 在文件集中是與其相關的條件，$p(T_k \mid q_n, D_i)$ 是在查詢項 q_n 與文件 D_i 出現的條件下潛藏主題 T_k 發生的機率。

2.7.1.2 主題混合模型訓練—非監督式

如果訓練資料集，沒有與使用者查詢 Q 相關文件的資訊，則可將每一文件 D_i 視為與自己相關，用以訓練主題混合模型，經由簡單的更改式(2.20)-(2.22) 得到：

$$\hat{p}(q_n \mid T_k) = \frac{\sum_{D_i \in [D]} n(q_n, D_i)\, p(T_k \mid q_n, D_i)}{\sum_{D_i \in [D]} \sum_{q_s \in D_i} n(q_s, D_i)\, p(T_k \mid q_s, D_i)} \tag{2. 23}$$

$$\hat{p}(T_k \mid D_i) = \frac{\sum_{q_s \in D_i} n(q_s, D_i)\, p(T_k \mid q_s, D_i)}{|D_i|} \tag{2. 24}$$

[D]代表整個文件集，$|D_i|$ 是文件 D_i 的長度，$n(q_n, D_i)$ 是查詢項 q_n 出現在文件 D_i 的次數，$p(T_k \mid q_n, D_i)$ 是在查詢項 q_n 與文件 D_i 出現的條件下潛藏主題 T_k 發生的機率。

第 3 章 自動摘要模型

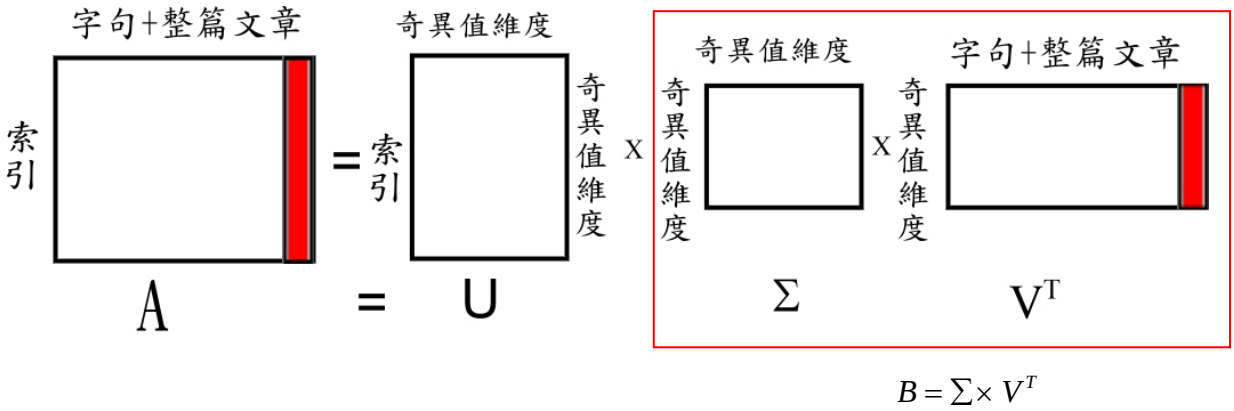
如何才能從文件中自動擷取出重要的字句，以之做為整篇文件的摘要，這是自動摘要所要探討的問題，本論文提出嵌入式潛藏語意分析模型、隱藏式馬可夫模型、主題混合模型等自動摘要模型，茲將其分別說明如下各小節。

3.1 嵌入式潛藏語意分析（Embedded LSA）模型

基於對潛藏語意分析與向量空間模型的探討，本論文提出嵌入式潛藏語意分析模型，其將每一字句與整篇文件共同投影到潛藏語意空間，最後藉由向量空間模型，估測各字句與整篇文件的相關性，演算法如下：

1. 將文件 D 斷句，$D = \{S_1, S_2, ..S_i.., S_N\}$
2. 由文件 D 建立 索引×字句矩陣 A，並將整篇文件嵌入到矩陣的最後一行
3. 對 A 進行奇異值分解，得到左奇異向量矩陣 U、奇異值矩陣 $\sum$ 與右奇異向量矩陣 V^t
4. 在右奇異向量矩陣 V^t 中，最後一行向量即為整篇文件在語意空間的表示法，其餘行向量即為原始文件中各字句在語意空間的表示法，將 $\sum$ 與 V^t 相乘得到各字句與整篇文件在潛藏語意空間的投影（$B = \sum \times V^T$）
5. 將 B 的最後一行（即整篇文件的投影）與 B 中的其他行向量（各字句），以向量空間模型表示，並進行餘弦相關度估測，得到一句排名
6. 依摘要比例，將句排名所對應的字句，摘錄形成摘要

如圖 3.1 所示，紅色部份即為所嵌入的整篇文件，*矩陣 B* 最後一行向量即為整篇文件的投影，將其與其他行向量（字句）做餘弦相關度估測後，得到一句排名，用以依摘要比例摘錄形成摘要。



$$B = \sum \times V^T$$

圖 3.1 嵌入式潛藏語意分析模型示意圖

3.2 隱藏式馬可夫模型-型一（HMM-Type1）

近年來有學者提出 HMM/N-gram based Model 用於中文語音文件檢索上 [Chen *et al.* 2004a]。延伸其應用，視文件為一機率生成模型（Probabilistic Generative Model），對於每個索引都有一對應的機率分佈，文件與文件中每一字句的相關性，是藉由每一字句的所有索引在文件發生的相似值（Likelihood）來決定，也就是說當字句的索引在文件的機率分佈值連乘積越高，則字句與文件的相關性就越高，如圖 3.2 所示，數學式如下：

$$p(S_i|D) = \prod_{w \in S_i} p(w|D) \approx \prod_{w \in S_i} \left[\lambda p(w|D) + (1-\lambda) p(w|Corpus) \right] \quad (3.1)$$

其中 $p(w|D)$ 為文件 D 產生索引 w 的機率值，並與一更大語料庫做平滑化（Smooth），$p(w|Corpus)$。

演算法如下：

1. 將文件 D 斷句，$D = \{S_1, S_2, ..S_i.., S_N\}$
2. 計算文件 D 的單連語言模型
3. 對文件 D 中各字句 S_i 估測 $p(S_i|D) \approx \prod_{w \in S_i} \left[\lambda p(w|D) + (1-\lambda) p(w|Corpus) \right]$

   機率值，並依此做排序形成一句排名
4. 依摘要比例，將句排名所對應的字句，輸出形成摘要

假設在一篇文件中的索引，其重要性皆相同，愈長的字句其分數愈低，是以

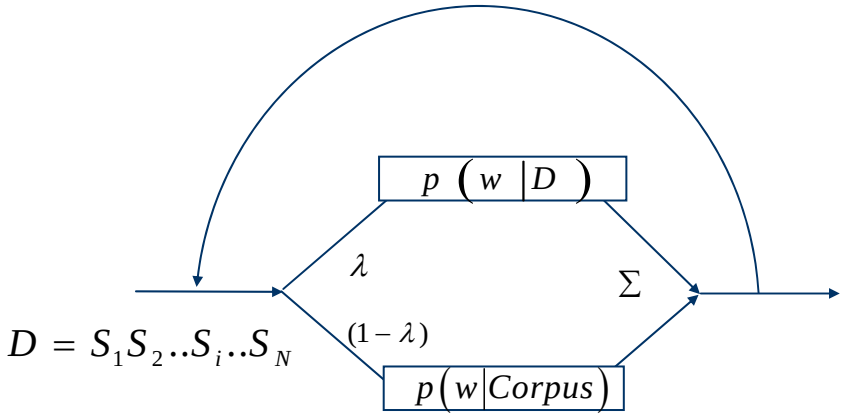


圖 3.2 隱藏式馬可夫模型-型一示意圖

在估測文件產生每一字句的機率 $p(S_i \mid D)$ 時，以每一字句長度分之 1 為次方對分數開根號（正規化），$\sqrt[|S_i|]{p\left(S_i\middle|D\right)}$，以避免句長影響到選取摘要字句的正確性。

此外，對於每一個文件，將文件 D 視為與自己相關，則參數 λ 與文件 D 產生各索引 w 的機率值可藉由期望值最大化演算法 [Dempster *et al.* 1977]，自動調整參數與訓練模型，數學式如下所示：

$$\hat{\lambda} = \frac{\sum_{w \in D} E\left(w, D\right)}{\left|D\right|} \tag{3.2}$$

$$\hat{p}(w \mid D) = \frac{E\left(w, D\right)}{\sum_{w \in D} E\left(w, D\right)} \tag{3.3}$$

$$E\left(w,D\right) = n\left(w,D\right)\frac{\lambda\, p\left(w\middle|D\right)}{\lambda\, p\left(w\middle|D\right) + \left(1-\lambda\right) p\left(w\middle|Corpus\right)} \tag{3.4}$$

其中 $\left|D\right|$ 是文件 D 的長度，$n\left(w,D\right)$ 是索引 w 出現在文件 D 的次數。

更進一步來說，文件 D 中每一字句 S_i 可利用與其相關的字句 $\widehat{S}_i$（可由字句 S_i 與一斷句後的文件語料庫，經由餘弦估測其相關度，最後再選取最相關的字句組成 $\widehat{S}_i$），做字句擴充（Sentence Expansion），如下所示：

$$p\left(\widehat{S}_i\middle|D\right) = \prod_{w \in \widehat{S}_i}\left[\lambda\, p\left(w\middle|D\right) + (1-\lambda)\, p\left(w\middle|Corpus\right)\right] \tag{3.5}$$

3.3 隱藏式馬可夫模型-型二（HMM-Type2）

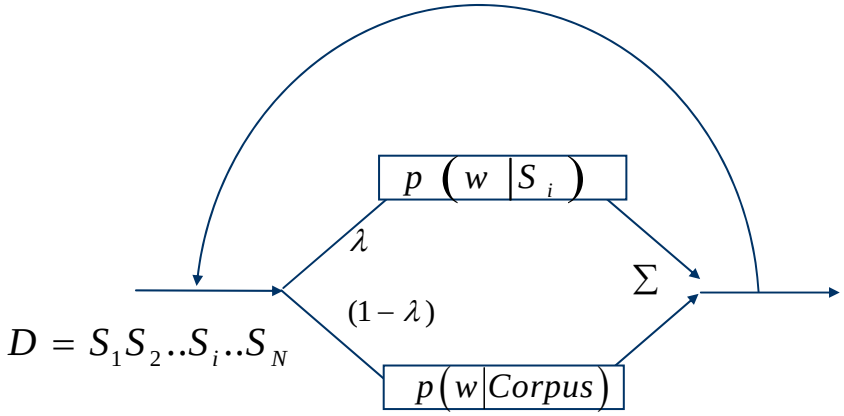


圖 3.3 隱藏式馬可夫模型-型二示意圖

同隱藏式馬可夫模型-型一的概念，當一篇文件進來時，視文件中每一字句為一機率生成模型，對於每個索引都有一個對應的機率分佈，文件中每一字句與文件的相關性，是藉由文件的所有索引在每一字句發生的相似值來決定，如圖 3.3 所示，數學式如下：

$$p(D|S_i) = \prod_{w \in D} p(w|S_i) \approx \prod_{w \in D} \left[\lambda p(w|S_i) + (1-\lambda) p(w|Corpus) \right] \quad (3.6)$$

其中 $p(w|S_i)$ 為文件中字句 S_i 產生索引 w 的機率值，並與一更大語料庫做平滑化， $p(w|Corpus)$。

演算法如下：

1. 將文件 D 斷句， $D = \{S_1, S_2, ..S_i.., S_N\}$
2. 對文件 D 中每一字句 S_i，計算其單連語言模型
3. 對文件 D 中各字句 S_i 估測 $p(D|S_i) \approx \prod_{w \in D} \left[\lambda p(w|S_i) + (1-\lambda) p(w|Corpus) \right]$ 機率值，並依此做排序形成一句排名
4. 依摘要比例，將句排名所對應的字句，輸出形成摘要

此外，對於每一個文件，將文件中每一字句 S_i 視為與文件 D 相關，則參數 λ 與每一字句 S_i 產生各索引 w 的機率值可藉由期望值最大化演算法 [Dempster *et al.* 1977]，自動調整參數與訓練模型，數學式如下所示：

$$\hat{\lambda} = \frac{\sum_{w \in D} E\left(w, S_i\right)}{|D|} \tag{3. 7}$$

$$\hat{p}(w \mid \mathrm{S}_i) = \frac{E\left(w, S_i\right)}{\sum_{w \in D} E\left(w, S_i\right)} \tag{3. 8}$$

$$E\left(w, S_i\right) = n\left(w, S_i\right) \cdot \frac{\lambda\, p\left(w \middle| S_i\right)}{\lambda\, p\left(w \middle| S_i\right) + (1-\lambda)\, p\left(w \middle| Corpus\right)} \tag{3. 9}$$

其中$|D|$是文件 D 的長度，$n\left(w, S_i\right)$是索引 w 出現在字句 S_i 的次數。

更進一步來說，因每個觀測（Observation）文件 D 中，皆含有模型 S_i 的資訊，是以可去除文件 D 中模型 S_i 的字詞，做字句移除（Sentence Removal），如下所示：

$$p(D \mid S_i) = \prod_{w \in D \wedge w \notin S_i} \left(\lambda p(w \mid S_i) + (1-\lambda) p\left(w \middle| Corpus\right)\right) \tag{3. 10}$$

3.4 主題混合模型（Topical Mixture Model, TMM）

根據 2.7 節關於主題混合模型的討論，給定一使用者查詢 Query $Q = q_1 q_2 .. q_n .. q_N$ ，一文件 D_i 可根據其相關程度做排名，$p(D_i \mid Q)$，經由推導後可由式(2.17)表示：

$$p(Q \mid D_i) = \prod_{n=1}^{N} \sum_{k=1}^{K} p(q_n \mid T_k)\, p(T_k \mid D_i)$$

延伸其應用於自動摘要模型上，將使用者查詢 Q 視查詢為一文件 D，一標題 H_i（標題可視為某一字句）可根據其相關程度做排名，$p(H_i \mid D)$，類同於 2.7 節的推導，最後可仿照式(2.17)表示成：

$$p(D \mid H_i) = \prod_{n=1}^{N} \sum_{k=1}^{K} p(q_n \mid T_k)\, p(T_k \mid H_i) \tag{3. 11}$$

也就是說，將原先以文件為模型，轉為以標題為模型。

於此以標題為模型主題混合模型中，可得到主題單連語言模型，如 $p(q_n \mid T_k)$，與其在各標題的權重，如 $p(T_k \mid H_i)$。

在訓練時，如果文件集已含有文件與標題相對應的資訊，如在一般新聞網站

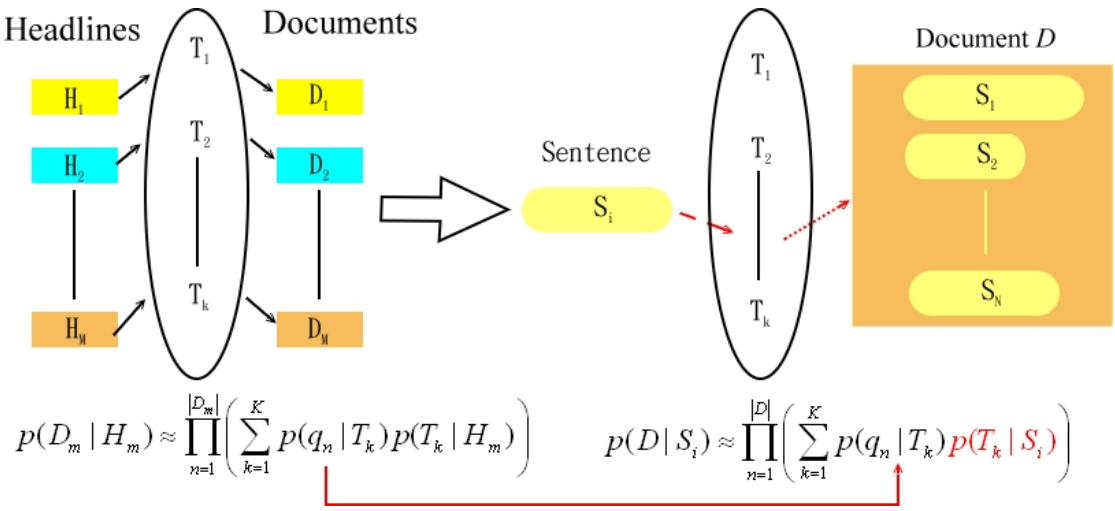


圖 3.4 主題混合模型示意圖

的新聞文件通常皆含有標題，則可藉由每一篇新聞的文件與其所相對應的標題來訓練。相對應的標題可使用當篇新聞的標題（本研究使用），也可由相關的新聞構成標題集，接著藉由式(2.20)-(2.22) 將 Q 轉為 D、D_i 轉成 H_i 做監督式訓練，以優化主題單連語言模型與其在各標題的權重。透過此訓練過程來學習標題（可視為字句）產生文件的流程。

在訓練時，如果文件集並無文件與標題相對應的資訊，則可將每一標題視為與自己相關，也就是將文件以標題取代，並藉由式(2.23)-(2.24) 將 Q 轉為 D、D_i 轉成 H_i 來進行非監督式訓練。

經由訓練過後，使用主題單連語言模型來代表主題的資訊。考慮如下情況，給定一使用者查詢文件 $D = q_1 q_2 .. q_n .. q_N$，文件中每一字句 S_i 可根據其相關程度做排名，$p(S_i \mid D)$，類同於 2.7 節的推導，最後可仿照式(2.17)表示成：

$$p(D \mid S_i) = \prod_{n=1}^{N}\sum_{k=1}^{K} p(q_n \mid T_k)p(T_k \mid S_i) \tag{3.12}$$

此機率，$p(D \mid S_i)$，即為主題混合模型在自動摘要的模型，其中主題單連語言模型由式(3.11)以標題為模型的主題混合模型訓練得之，如 $p(q_n \mid T_k)$，是以目前尚不知 $p(T_k \mid S_i)$ 的機率值，於此可利用原以標題為模型的主題混合模型，所得到的主題資訊，在摘要時即時迭代更新 $p(T_k \mid S_i)$，來估測每一字句 S_i 產生整篇文件 D 的機率，如圖 3.4 所示。

進一步來說， $p(T_k \mid S_i)$ 的初始值，可用下式估計：

$$p(T_k \mid S_i) = \frac{R(\overrightarrow{S_i}, \overrightarrow{T}_k)}{\sum_{r=1}^{k} R(\overrightarrow{S_i}, \overrightarrow{T}_r)} \tag{3. 13}$$

其中主題 T_k 是由原以標題為模型的主題混合模型而來，$R(\overrightarrow{T}_k, \overrightarrow{S_i})$ 代表利用餘弦估測字句 S_i 與主題 T_k 的距離，如下所示：

$$R(\overrightarrow{T}_k, \overrightarrow{S_i}) = \frac{\overrightarrow{T}_k \cdot \overrightarrow{S_i}}{\left\|\overrightarrow{T}_k\right\| \cdot \left\|\overrightarrow{S_i}\right\|} \tag{3. 14}$$

得到 $p(T_k \mid S_i)$ 的初始值之後， $p(T_k \mid S_i)$ 可藉由非監督式訓練，視每一字句 S_i 與自己相關，即時迭代更新得之，如下所示：

$$\hat{P}(T_k \mid S_i) = \frac{\sum_{q_s \in S_i} n(q_s, S_i)\, p(T_k \mid q_s, S_i)}{\left|S_i\right|} \tag{3. 15}$$

$$p(T_k \mid q_s, S_i) = \frac{p(T_k \mid S_i)\, p(q_s \mid T_k)}{\sum_{l=1}^{K} p(T_l \mid S_i)\, p(q_s \mid T_l)} \tag{3. 16}$$

$\left|S_i\right|$ 是字句 S_i 的長度， $n(q_s, S_i)$ 是查詢項 q_s 出現在字句 S_i 的次數， $p(T_k \mid q_s, S_i)$ 是在查詢項 q_s 與字句 S_i 出現的條件下潛藏主題 T_k 發生的機率。

在實作上，額外考慮每一查詢項在各字句中的重要性，是以式(3.12)可進一步延伸為：

$$p(D \mid S_i) = \prod_{n=1}^{N} \left(\lambda\, p(q_n \mid S_i) + (1-\lambda) \sum_{k=1}^{K} p(q_n \mid T_k)\, p(T_k \mid S_i) \right) \tag{3. 17}$$

$p(q_n \mid S_i)$ 為字句 S_i 產生查詢項 q_n 的機率， $p(q_n \mid T_k)$ 可由以標題為模型的主題混合模型訓練得之， $p(T_k \mid S_i)$ 可經由非監督式訓練即時迭代更新得之。

演算法如下：

1. 訓練以標題為模型的主題混合模型 $p(D \mid H_i) = \prod_{n=1}^{N} \sum_{k=1}^{K} p(q_n \mid T_k)\, p(T_k \mid H_i)$ ，得到主題單連語言模型用以代表潛藏主題的資訊

2. 將文件 D 斷句， $D = \{S_1, S_2, ..S_i.., S_N\}$

3. 由式(3.17) 估測 D 在每一字句 S_i 的機率值， $p(D|S_i)$：計算各字句 S_i 的單連語言模型，如 $p(q_n|S_i)$，與查詢項 q_n 發生在潛藏主題及字句產生各別主題的機率值，$\sum_{k=1}^{K} p(q_n|T_k)p(T_k|S_i)$。並依此機率值做排序，形成一句排名

4. 依摘要比例，將句排名所對應的字句，輸出形成摘要

此外，對於每一個文件，將文件中每一字句 S_i 視為與文件 D 相關，則參數 λ 與每一字句 S_i 產生各索引 w 的機率值可藉由期望值最大化演算法 [Dempster *et al.* 1977]，自動調整參數與訓練模型，數學式如下所示：

$$\hat{\lambda} = \frac{\sum_{w \in D} E(w, S_i)}{|D|} \tag{3.18}$$

$$\hat{p}(w|S_i) = \frac{E(w, S_i)}{\sum_{w \in D} E(w, S_i)} \tag{3.19}$$

$$E(w, S_i) = n(w, S_i) \cdot \frac{\lambda p(w|S_i)}{\lambda p(w|S_i) + (1-\lambda)\sum_{k=1}^{K} p(w|T_k)p(T_k|S_i)} \tag{3.20}$$

其中 $|D|$ 是文件 D 的長度，$n(w, S_i)$ 是索引 w 出現在字句 S_i 的次數。

更進一步來說，因每個觀測（Observation）文件 D 中，皆含有模型 S_i 的資訊，是以可去除文件 D 中模型 S_i 的字詞，做字句移除（Sentence Removal），如下所示：

$$p(D|S_i) = \prod_{w \in D \wedge w \notin S_i} \left(\lambda p(w|S_i) + (1-\lambda)\sum_{k} p(w|T_k)p(T_k|S_i) \right) \tag{3.21}$$

第 4 章 實驗語料與實驗

4.1 實驗語料

實驗語料蒐集自 News 98 新聞網 [News 98] ，包含 2001 年 8 月 1 日至 8 月 24 日中午 12:00~13:00 的 FM 廣播新聞，相關統計資料如表 4. 1 所示：

表 4. 1 News 98 廣播新聞相關統計

新聞時間	2001 年 8 月 1 日~2001 年 8 月 24 日
新聞數	200 則
總長度	1.61 小時
平均每則新聞長度	28.96 秒
總大小（人工轉寫）	約 31 仟字
平均每則新聞大小（人工轉寫）	約 157 字

這 200 則廣播新聞，有自動轉寫（Automatic Transcription）與人工轉寫（Human Transcription）兩種資料集，自動轉寫部分包含兩個辨識結果，相關資訊如表 4. 2 所示：

表 4. 2 自動轉寫相關資訊

語音辨識自動轉寫	
經詞圖（Word Graph）搜尋後的辨識結果，以 SP_WG 代表	加上非監度式語者調適（+MLLR）的辨識結果，以 SP_Adapt 代表
辨識率為 84.11%	辨識率為 84.64%

自動摘要評估的標準答案部分，請三位國立台灣大學文學院大三以上的學生，分別對 200 則廣播新聞的人工轉寫做摘要，摘要的結果分為句排名（Extraction）與重寫（Abstraction）兩種 [何遠 2003]。

(a) 以句排名標準答案為例：

（1）經發會兩岸組達成鬆綁戒急用忍共識

（2）行政院也將大幅開放兩岸經貿政策

（5）一向主張戒急用忍的經建會主委陳博志今天表示

（7）如果國內企業需要大陸市場來成為世界領導廠商的話

（8）他絕對樂觀其成

（9）不過陳博志也強調

（3）經發會分組會議達成的鬆綁共識不是毫無設限

（4）應該有總量管理的觀念

（6）他也認為對大陸投資不應該太多

此表示第一句重要性第 1、第二句重要性第 2、第三句重要性第 5 … 以此類推。此種人工標準的優點在於可容易的可選出任意比例的摘要結果做為標準答案。

(b) 以重寫標準答案為例：

經發會達成鬆綁戒急用忍共識，行政院也將大幅開放兩岸經貿政策。經建會主委陳博志表示鬆綁並非毫無設限，應有總量管理觀念。

此種人工標準答案可讀性較佳。

此外為了訓練隱藏式馬可夫模型與主題混合模型的參數，將此 200 則廣播新聞按時間先後順序分為發展集（Development Set）及測試集（Test Set）兩部分。其中發展集用來訓練參數，而測試集用以算出最後的摘要結果，並與其他摘要模型做比較，發展集與測試集的相關資訊如表 4.3 所示：

表 4.3 發展集與測試集大小

語料庫	新聞數
發展集（Development Set）	100 則
測試集（Test Set）	100 則
總共　200 則	

4.2 斷詞與統計式語言模型

斷詞（Tokenization）是利用詞典，將一個字串中的文字，比對詞典內的詞來當做斷詞的依據。舉例來說，就是將下列中字句：

“經發會兩岸組達成鬆綁戒急用忍共識”

斷為

“經發會”“兩岸”“組”“達成”“鬆綁戒急用忍”“共識”

在斷詞演算法，本實驗以長詞為優先，如“鬆綁”“戒急用忍”與“鬆綁戒急用忍”皆為詞典內之詞的話，則斷詞演算法將輸出“鬆綁戒急用忍”。

經由第 3 章的討論，在隱藏式馬可夫模型與主題混合模型中，需要訓練語料庫，用以得到統計式語言模型與文件-標題查詢範例的訓練集合，本實驗採用中央通訊社（Central News Agency）[中央通訊社] 在西元 2001 年 08 月且型態屬於故事（type="story"）的文字新聞做為訓練語料庫，每一篇新聞皆含有文件與標題兩部份，其內容包括國內外及大陸文教、交通、社會、財經、國會、影劇、醫藥衛生、體育及地方新聞，相關資訊如表 4.4 所示：

表 4.4 中央社文字語料相關統計

新聞時間	2001 年 8 月
新聞數	14,178 則
總大小	約 709 萬字
平均每則新聞長度	約 500 字

在訓練統計式語言模型時，一般第一步是將語料做斷詞，再用斷詞過的語料來訓練語言模型，最後得到以詞為基礎的語言模型。此種做法的優點在於字句的主旨由詞的詞義而非字的字義組成，因此以詞為基礎的語言模型可以得到以字為基礎的語言模型，更豐富的語言資訊。

然而斷詞也有其不足之處：其一，因為字組成詞的變化程度相當大，是以一個字句難免會有多種斷詞方式；其二，未知詞（Out-of-Vocabulary, OOV）的問

題，例如專有名詞、地名、人名等，不在詞典裡的機率通常很高。這兩個不足之處是目前以斷詞為基礎的研究所必需解決的問題。

語言模型的訓練部份本實驗使用 SRI 國際（SRI International）語音科技及研究實驗室（Speech Technology and Research Laboratory）所提供的 SRI 語言模型工具箱（SRI Language Modeling Toolkit）[SRI Toolkit]。 SRI 語言模型工具箱是一套用來訓練或應用統計式語言模型（Statistical Language Model）的工具程式組合。

本實驗所使用的命令參數如下：

ngram-count -text 斷詞後的文字語料 -lm 輸出語言模型檔

語言模型的平滑化（Smoothing）演算法則是使用預設的 Good-Turning 演算法。

4.3 自動摘要評估

本實驗的評估方式有三種：第一為餘弦（Cosine）評估：藉由計算自動摘要與人工摘要的相關度為標準；第二為 ROUGE（Recall-Oriented Understudy for Gisting Evaluation）評估 [Lin 2003]：其為一召回率導向的自動摘要評估方式；第三為平均精確度（Mean Average Precision）評估：藉由計算在相同摘要比例下，自動摘要與人工摘要的平均精確度為標準，然因自動轉寫與人工轉寫斷句的不同，造成自動摘要與人工摘要的字句無法判斷是否相關（即所摘要出的字句，在人工句排名內），是以本論文提出以改良型字錯誤率（modified Character Error Rate, m-CER）為基礎的相關判斷準則，只要符合此一準則就視自動摘要的字句與人工摘要的字句是相關而予以計算精確度，茲將其分別說明如下各小節。

4.3.1 餘弦評估

在此自動摘要的評估方式上，是以計算與人工摘要結果的相關度做為標準。由於人工摘要的產生有兩種。第一種為摘錄式(Extractive)摘要(或稱為句排名方式)：對於每則新聞中的字句依其對於整則新聞的重要性作排名，因此很容易的可選出

任意比例的摘要結果，如 20%、30%、50%、70% 等摘要比例；第二種為非摘錄式（Non-Extractive or Abstractive）摘要（或稱為重寫摘要的方式）：其摘要比例就無法作任意的修改。

假設 A_d 代表對某篇文件 d 自動摘要結果、$E_{n,d}$ 代表第 n 個人對文件 d 以摘錄式摘要結果、$R_{n,d}$ 代表第 n 個人對文件 d 重寫摘要結果、以及 $m\%$ 為可能的摘要比例，則對所有新聞自動摘要的正確率被定義為 [Orasan 2002; 何遠 2003]：

$$ACC(m\%) = \frac{1}{D}\frac{1}{N}\sum_{d=1}^{D}\sum_{n=1}^{N}\frac{SIM\left(A_d(m\%), E_{n,d}(m\%)\right) + SIM\left(A_d(m\%), R_{n,d}\right)}{2} \quad (4.1)$$

其中，$SIM(\cdot,\cdot)$ 為評估自動摘要與人工摘要結果的相關度公式，此公式使用向量空間模型做為相關度評估的方法：將自動摘要與人工摘要結果以向量形式表現，並計算向量間夾角的餘弦值來得到相關度，例如式中的 $SIM(A_d(m\%), E_{n,d}(m\%))$ 可表示成：

$$SIM\left(A_d(m\%), E_{n,d}(m\%)\right) = \frac{\vec{V}_{A_d(m\%)} \cdot \vec{V}_{E_{n,d}(m\%)}}{\left|\vec{V}_{A_d(m\%)}\right|\left|\vec{V}_{E_{n,d}(m\%)}\right|} \quad (4.2)$$

上式中 $\vec{V}_{A_d(m\%)}$ 與 $\vec{V}_{E_{n,d}(m\%)}$ 分別為自動摘要結果 $A_d(m\%)$ 與人工摘錄式摘要結果 $E_{n,d}(m\%)$ 的向量表示式，而向量表示式中的每一維度則是代表著某一個索引特徵 t，t 在摘要中的重要性（或稱權重）$w(t)$，是以它在摘要中出現的詞頻（Term Frequency, TF）與反文件頻率（Inverse Document Frequency, IDF）乘積來表示:

$$w(t) = \left(\frac{f_t}{\max_h f_h}\right)\log\left(N/n_t\right) \quad (4.3)$$

上式中 f_t 為索引特徵 t 在摘要中出現的次數，$\log(N/n_t)$ 是反文件頻率，N 是所有廣播新聞則數，n_t 是索引特徵 t 出現的廣播新聞則數。

4.3.2 ROUGE 評估

ROUGE（Recall-Oriented Understudy for Gisting Evaluation）是召回率導向的自動

摘要評估方式 [Lin 2003]。此評估計算自動摘要與人工摘要重疊單位元的次數，單位可為 *N*-連語言模型（*N*-gram Language Model）、詞順序（Word sequences）與詞對（Word pairs）。

如 *N*-連語言模型單位元的評估，如下所示：

$$ROUGE-N=\frac{\sum_{S\in\{人工摘要標準答案\}}\sum_{N-連語言模型\in S} Count_{match}(N-連語言模型)}{\sum_{S\in\{人工摘要標準答案\}}\sum_{N-連語言模型\in S} Count(N-連語言模型)} \quad (4.4)$$

$Count_{match}$ 代表 *N*-連語言模型共同出現在自動摘要與人工摘要標準答案的最大次數，在人工摘要標準答案部份本實驗僅使用句排名。此外本實驗以 ROUGE-2 ，即雙連語言模型為評估標準。

4.3.3 平均精確度評估

在此摘要評估方式上，是以不同摘要比例 *m*% 下的平均精確度（Mean Average Precision, MAP）來評估。假設 A_d 代表對某篇文件 *d* 自動摘要的句排名、$E_{n,d}$ 代表第 *n* 個人對文件 *d* 做的句排名以及 *m*% 為可能的摘要比例（用以保留前 *m*% 分數較大的字句來計算平均精確度），平均精準度計算公式如下：

$$\text{mAP}(\text{m}\%)=\frac{1}{D}\frac{1}{N}\sum_{d=1}^{D}\sum_{n=1}^{N}\frac{1}{R_d}\sum_{s=1}^{R_d}\frac{s}{r_{d,s,n}} \quad (4.5)$$

R_d 為 A_d (m%) 與 $E_{n,d}$ (m%) 相關的字句個數，$r_{d,s,n}$ 為 A_d (m%)與 $E_{n,d}$ (m%) 相關的第 *s* 句從 A_d 前頭（Top）數來的位置。

在自動轉寫上，因斷句的不同而無法與人工轉寫的字句做一對一對應，是以無法決定所摘要出的字句在人工標準答案的字句上，也就是說式(4.5)中無法決定相關，是以本論文定義相關判斷準則來解決此一問題。

首先介紹 Levenshtein Distance（LD），LD 為二個字串之間最短的編輯距離。最早由俄國科學家 Vladimir Levenshtein 在 1966 所提出 [Levenshtein 1966]，又稱為編輯距離（Edit Distance）。假設 *I* 是測試字串，*R* 是正確字串，並以單字

（Character）為索引。LD 演算法計算字串 *I* 轉換為 *R* 所需的刪除、插入、替代編輯次數。例如：

- *I* 是"陳水扁" 且 *R* 是 "陳水扁" 則 LD(*I*,*R*)=0 ，因在此範例中無任何轉換
- *I* 是"陳水篇" 且 *R* 是 "陳水扁" 則 LD(*I*,*R*)=1 ，因在此範例中只需一個替代（改變篇到扁）就足以將 *I* 轉換到 *R*

是以經由 LD 演算法，可以得到<插入，刪除、替代、正確>的個數，且愈大的 LD，代表字串的相異度愈大。目前 LD 演算法，已大量運用於拼字檢查 [黃上銘等 2003]、語音辨識、DNA 分析等應用上。

接著定義改良型字錯誤率（modified Character Error Rate, m-CER）如下，假設 *I* 是測試字串、*R* 是正確字串，並以單字（Character）為索引，<插入，刪除、替代、正確>為經由 LD 演算法所計算出的個數，則

$$m-CER=\frac{(\text{插入}+\text{刪除}+\text{替代})-(diff)}{\text{R的字數}} \tag{4.6}$$

$$diff=\begin{Bmatrix}\text{句長差} & \text{if 句長差} \le \text{正確個數}\\ \text{正確個數} & \text{if 句長差} > \text{正確個數}\end{Bmatrix} \tag{4.7}$$

其中 *diff* 主要用來解決，因句長差（$\text{abs}(|\text{R}|-|\text{I}|)$）所產生的錯誤個數。

例如：

I ：臺北電信網路展今天開幕

R ：臺北電信網路展今天開幕一連展出四天

LD = <插入，刪除，替代、正確> = < 6,0,0,11>

如不使用句長差，則原字錯誤率高達 $CER=\frac{(6+0+0)}{11}=\frac{6}{11}=0.5454$

然而經由計算 *diff* 後，則 *m-CER* =0.0

經由上述討論，式(4.5)評估自動轉寫與人工轉寫，字句相關的定義如下：

假設 A_d 代表對某篇文件 *d* 自動摘要的句排名、$E_{n,d}$ 代表第 *n* 個人對文件 *d* 做

的句排名、$m\%$ 為可能的摘要比例、$m-CER_{\min}$ 代表 $A_d(m\%)$ 中的某一句 I，對 $E_{n,d}(m\%)$ 每一字句 R ，計算 *m-CER* 所得的最小值

如果 $m\text{-}CER_{\min}$ <0.3 則 相關

4.4 實驗結果

基礎實驗（Baseline）有三個，分別為 VSM（向量空間模型）、RM（相關評估）與 LSA（潛藏語意分析），本論文提出的模型計有 embedded LSA（嵌入式潛藏語意分析）、HMM-Type1（隱藏式馬可夫模型-型一）、HMM-Type2（隱藏式馬可夫模型-型二）、TMM(D|T)（主題混合模型，監督式訓練）、TMM(T|T)（主題混合模型，非監督式訓練）。

根據台大與中研院近年來在中文語音文件檢索（Spoken Document Retrieval）領域的研究成果 [Chen *et al.* 2002]，若使用音節（Syllable）或字（Character）為單位的索引特徵（Indexing Feature）組合，往往可以有比詞更好的檢索表現。因此，本實驗將探討不同的特徵單位（Feature Unit）對自動摘要的影響。首先，以所帶資訊的豐富程度來看，顯然詞所帶的資訊量是最多的，其次是字，然後為音節 [Lee *et al.* 2003]，如表 4.5 所示：

表 4.5 音節、字、詞所帶資訊的比較

單位	所帶資訊的比較
音節（Syllable）	低
字（Character）	中
詞（Word）	高

此外在音節中，因為中文的同音異義字（Homonym）很多，一個音節可能對應好幾個字，所以音節的混淆度（Ambiguity）很大。然而雖然音節所帶的資訊少、混淆度大，但是如果將兩個音節連結起來成為一個雙音節（Syllable Pair），那麼所帶的資訊就變得豐富很多 [Lee *et al.* 2003] ，綜合上述討論，本實驗除了使用單詞為特徵單位來做自動摘要外，並將實驗使用以雙字、雙音節組合為特徵

單位，來觀察自動摘要的正確率。

其中雙字、雙音節都是互相重疊（Overlapping）的。例如一字句由 10 個音節所組成：

$$(S_1, S_2, ..., S_{10})$$

則其雙音節特徵即為

$$(S_1, S_2), (S_2, S_3), ..., (S_9, S_{10})$$

同樣地互相重疊的雙字，定義類似。

4.4.1 餘弦評估

在本小節所有表格中，TMM 欄內的數字代表所使用的潛藏主題數。如表 4. 6 TMM P(D|T) 64 即代表使用 64 個潛藏主題數。此外 HMM 與 TMM 的數據，均使用期望值最大化演算法，自動調整參數與訓練模型。

由表 4. 6~表 4. 8 以單詞為特徵單位的結果顯示，在低摘要比例（20%）時，TMM 與 HMM-Type2 其結果不論在人工轉寫或自動轉寫上，均優於其他摘要模型，並以 TMM 為較佳。進一步比較表 4. 6（人工轉寫）、表 4. 7（自動轉寫 SP_WG）可發現，所有模型的正確率均下降，這可由兩個方面來說明：其一，因為為語音辨識會產生錯誤的辨識結果，錯誤結果當然會影響摘要的正確率；其二，人工轉寫的斷句與標準答案（句排名）是一致的，而語音辨識結果的斷句目前主要是以靜音（silence）的長度為標準，因此最後斷開成的字句會與標準答案的字句有所差異，所以也會導致摘要正確率的下降。此外比較表 4. 7、表 4. 8（自動轉寫 SP_Adapt）可發現，隨著辨識率的上升，其結果也有所提升。

表 4.6 摘錄方法比較(評估方式：餘弦，特徵單位：詞，人工轉寫)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 64	TMM P(T|T) 32
20%	0.4473	0.4551	0.3599	0.4439	0.3639	0.4577	0.4631	0.4654
30%	0.5071	0.5108	0.4413	0.4983	0.4132	0.5278	0.5296	0.5265
50%	0.6365	0.6269	0.6098	0.6334	0.5918	0.6260	0.6302	0.6287
70%	0.6961	0.6956	0.6864	0.7051	0.6729	0.6951	0.6960	0.6952

表 4.7 摘錄方法比較(評估方式：餘弦，特徵單位：詞，自動轉寫 SP_WG)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 64	TMM P(T|T) 64
20%	0.3573	0.3586	0.2971	0.3652	0.3112	0.3719	0.3771	0.3750
30%	0.4074	0.4052	0.3487	0.4020	0.3570	0.4160	0.4188	0.4226
50%	0.5397	0.5130	0.4863	0.5102	0.5098	0.5285	0.5279	0.5289
70%	0.5839	0.5608	0.5540	0.5747	0.5738	0.5782	0.5777	0.5771

表 4.8 摘錄方法比較(評估方式：餘弦，特徵單位：詞，自動轉寫 SP_Adapt)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 32	TMM P(T|T) 32
20%	0.3630	0.3603	0.3138	0.3726	0.3206	0.3712	0.3718	0.3738
30%	0.4269	0.4195	0.3592	0.4142	0.3697	0.4234	0.4243	0.4230
50%	0.5413	0.5132	0.4937	0.5176	0.5057	0.5296	0.5311	0.5302
70%	0.5874	0.5632	0.5558	0.5788	0.5760	0.5775	0.5791	0.5783

由表 4.10~表 4.11 以雙音節為特徵單位的結果顯示，在低摘要比例（20% 與 30%）時，TMM 與 HMM-Type2 其結果不論在人工轉寫或自動轉寫上，均優於其他摘要模型。此外比較表 4.10、表 4.11 可發現，隨著辨識率的上升，其結果也有所提升。

表 4.9 摘錄方法比較(評估方式：餘弦，特徵單位：雙音節，人工轉寫)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 4	TMM P(T|T) 4
20%	0.4485	0.4425	0.3688	0.4457	0.3613	0.4618	0.4618	0.4525
30%	0.5176	0.5075	0.4578	0.5159	0.4352	0.5254	0.5283	0.5246
50%	0.6326	0.6300	0.6101	0.6413	0.5997	0.6386	0.6387	0.6380
70%	0.7054	0.7066	0.6882	0.7116	0.6692	0.7007	0.7014	0.7018

表 4.10 摘錄方法比較(評估方式：餘弦，特徵單位：雙音節，自動轉寫 SP_WG)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 16	TMM P(T|T) 16
20%	0.3665	0.3686	0.3054	0.3567	0.3266	0.3811	0.3811	0.3833
30%	0.4128	0.4121	0.3608	0.4038	0.3783	0.4167	0.4150	0.4138
50%	0.5287	0.5125	0.4931	0.5268	0.5017	0.5318	0.5287	0.5329
70%	0.5851	0.5782	0.5640	0.5881	0.5633	0.5889	0.5879	0.5887

表 4.11 摘錄方法比較(評估方式：餘弦，特徵單位：雙音節，自動轉寫 SP_Adapt)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 16	TMM P(T|T) 16
20%	0.3633	0.3675	0.2966	0.3646	0.3276	0.3833	0.3833	0.3833
30%	0.4157	0.4079	0.3555	0.4136	0.3824	0.4296	0.4268	0.4273
50%	0.5299	0.5185	0.4917	0.5315	0.5119	0.5391	0.5371	0.5407
70%	0.5854	0.5793	0.5625	0.5880	0.5694	0.5912	0.5899	0.5908

由表 4. 12~表 4. 14 以雙字為特徵單位的結果顯示，向量空間模型為基礎的摘要模型（VSM 與 RM）有較佳的結果。此外在高摘要比例（70%）時，embedded LSA 其結果均優於其他摘要模型。

表 4. 12 摘錄方法比較(評估方式：餘弦，特徵單位：雙字，人工轉寫)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 32	TMM P(T|T) 64
20%	0.4596	0.4538	0.3781	0.4381	0.3458	0.4460	0.4531	0.4489
30%	0.5191	0.5255	0.4487	0.5165	0.4343	0.5107	0.5211	0.5059
50%	0.6407	0.6335	0.6049	0.6430	0.5895	0.6386	0.6399	0.6390
70%	0.7008	0.7024	0.6803	0.7064	0.6669	0.7020	0.7011	0.7012

表 4. 13 摘錄方法比較(評估方式：餘弦，特徵單位：雙字，自動轉寫 SP_WG)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 64	TMM P(T|T) 4
20%	0.3862	0.3879	0.2974	0.3544	0.3231	0.3708	0.3741	0.3737
30%	0.4223	0.4101	0.3586	0.4109	0.3829	0.4168	0.4173	0.4207
50%	0.5331	0.5176	0.5037	0.5293	0.5118	0.5253	0.5258	0.5265
70%	0.5857	0.5729	0.5611	0.5884	0.5657	0.5819	0.5816	0.5821

表 4. 14 摘錄方法比較(評估方式：餘弦，特徵單位：雙字，自動轉寫 SP_Adapt)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 64	TMM P(T|T) 64
20%	0.3784	0.3779	0.3009	0.3508	0.3234	0.3749	0.3794	0.3759
30%	0.4335	0.4240	0.3671	0.4106	0.3864	0.4331	0.4357	0.4304
50%	0.5354	0.5189	0.5044	0.5299	0.5104	0.5297	0.5302	0.5319
70%	0.5882	0.5766	0.5551	0.5920	0.5710	0.5869	0.5869	0.5881

4.4.2 ROUGE 評估

在本小節所有表格中 TMM 欄內的數字，代表所使用的潛藏主題數。此外 HMM 與 TMM 的數據，均使用期望值最大化演算法，自動調整參數與訓練模型。

由表 4. 15~表 4. 17 以單詞為特徵單位的結果顯示，在低摘要比例（30%）時，TMM 與 HMM-Type2，其結果不論在人工轉寫或自動轉寫上，均優於其他摘要

模型，並以 TMM 為較佳。此外在高摘要比例（50% 或 70%）時，embedded LSA 其結果均優於其他摘要模型。

表 4. 15 摘錄方法比較(評估方式：ROUGE，特徵單位：詞，人工轉寫)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 64	TMM P(T|T) 32
20%	0.3697	0.3890	0.2722	0.4046	0.1797	0.3969	0.4070	0.4098
30%	0.4272	0.4623	0.3701	0.4480	0.2175	0.4694	0.4736	0.4676
50%	0.6142	0.6302	0.5892	0.6511	0.4824	0.6110	0.6161	0.6136
70%	0.7221	0.7503	0.7046	0.7666	0.6357	0.7277	0.7293	0.7260

表 4. 16 摘錄方法比較(評估方式：ROUGE，特徵單位：詞，自動轉寫 SP_WG)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 64	TMM P(T|T) 32
20%	0.2208	0.2284	0.1906	0.2758	0.1485	0.2538	0.2590	0.2581
30%	0.2532	0.2648	0.2259	0.2694	0.1762	0.2804	0.2817	0.2766
50%	0.3992	0.3905	0.3565	0.4059	0.3304	0.4004	0.3987	0.3994
70%	0.4665	0.4542	0.4405	0.4831	0.4268	0.4714	0.4694	0.4697

表 4. 17 摘錄方法比較(評估方式：ROUGE，特徵單位：詞，自動轉寫 SP_Adapt)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 32	TMM P(T|T) 32
20%	0.2291	0.2306	0.2041	0.2831	0.1588	0.2609	0.2612	0.2618
30%	0.2745	0.2833	0.2382	0.2849	0.1789	0.2862	0.2889	0.2840
50%	0.4060	0.3929	0.3697	0.4133	0.3248	0.4049	0.4060	0.4052
70%	0.4680	0.4527	0.4441	0.4873	0.4304	0.4707	0.4722	0.4720

由表 4. 18~表 4. 20 以雙音節為特徵單位的結果顯示，在低摘要比例（20% 或 30%），TMM 與 HMM-Type2，其結果不論在人工轉寫或自動轉寫上，大致均優於其他摘要模型。此外在高摘要比例（50% 或 70%）時，embedded LSA 其結果均優於其他摘要模型。

表 4. 18 摘錄方法比較(評估方式：ROUGE，特徵單位：雙音節，人工轉寫)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 4	TMM P(T|T) 16
20%	0.3834	0.3902	0.2941	0.4168	0.1352	0.4273	0.4273	0.4151
30%	0.4493	0.4664	0.3671	0.4848	0.2224	0.4843	0.4864	0.4801
50%	0.6110	0.6284	0.5623	0.6492	0.4728	0.6310	0.6310	0.6334
70%	0.7395	0.7605	0.6828	0.7646	0.6176	0.7398	0.7408	0.7421

表 4. 19 摘錄方法比較(評估方式：ROUGE，特徵單位：雙音節，自動轉寫 SP_WG)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 4	TMM P(T|T) 16
20%	0.2513	0.2590	0.1990	0.2559	0.1730	0.2812	0.2812	0.2826
30%	0.2790	0.2874	0.2216	0.2863	0.1883	0.2933	0.2919	0.2884
50%	0.4041	0.3988	0.3548	0.4248	0.3170	0.4142	0.4135	0.4154
70%	0.4770	0.4779	0.4465	0.4909	0.4087	0.4852	0.4849	0.4855

表 4. 20 摘錄方法比較(評估方式：ROUGE，特徵單位：雙音節，自動轉寫 SP_Adapt)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 4	TMM P(T|T) 16
20%	0.2503	0.2574	0.1962	0.2674	0.1701	0.2809	0.2809	0.2809
30%	0.2839	0.2837	0.2281	0.3030	0.1902	0.3068	0.3027	0.3027
50%	0.4066	0.4065	0.3560	0.4314	0.3245	0.4235	0.4228	0.4253
70%	0.4780	0.4830	0.4458	0.4917	0.4095	0.4843	0.4841	0.4846

由表 4. 21～表 4. 23 以雙字為特徵單位的結果顯示，在低摘要比例（20%）時，向量空間模型為基礎的摘要模型（VSM 與 RM）有較佳的結果。此外在高摘要比例（50% 或 70%）時，embedded LSA 其結果大致均優於其它摘要模型。

表 4. 21 摘錄方法比較(評估方式：ROUGE，特徵單位：雙字，人工轉寫)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 64	TMM P(T|T) 64
20%	0.4132	0.4194	0.2910	0.4091	0.1474	0.4037	0.4110	0.4110
30%	0.4733	0.4979	0.3553	0.5000	0.2343	0.4787	0.4706	0.4712
50%	0.6271	0.6452	0.5575	0.6506	0.4559	0.6418	0.6402	0.6401
70%	0.7361	0.7627	0.6817	0.7586	0.6047	0.7470	0.7403	0.7472

表 4. 22 摘錄方法比較(評估方式：ROUGE，特徵單位：雙字，自動轉寫 SP_WG)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 64	TMM P(T|T) 4
20%	0.2767	0.2804	0.1735	0.2569	0.1673	0.2518	0.2549	0.2578
30%	0.2811	0.2765	0.2065	0.2910	0.1934	0.2814	0.2880	0.2840
50%	0.4134	0.4099	0.3604	0.4215	0.3298	0.4122	0.4126	0.4143
70%	0.4772	0.4800	0.4403	0.4921	0.4143	0.4783	0.4773	0.4786

表 4. 23 摘錄方法比較(評估方式：ROUGE，特徵單位：雙字，自動轉寫 SP_Adapt)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T) 64	TMM P(T|T) 4
20%	0.2685	0.2680	0.1852	0.2589	0.1641	0.2626	0.2651	0.2626
30%	0.2905	0.2931	0.2215	0.2971	0.1969	0.2979	0.3057	0.2979
50%	0.4160	0.4111	0.3647	0.4252	0.3244	0.4170	0.4178	0.4190
70%	0.4798	0.4863	0.4339	0.4953	0.4157	0.4829	0.4829	0.4829

4.4.2 平均精確度評估

在本小節所有表格中，每一欄的 HMM 與 TMM 數據又可分成三列：第一列為在發展集中找的較佳參數，第二列為直接在測試集找的較佳參數，第三列為使用期望值最大化演算法，自動調整參數與訓練模型所得的數據。

例如表 4. 24 中 HMM-Type2 0.40 / 0.95 ，分別代表 λ 為 0.40(發展集)與 0.95

（測試集）；TMM P(D|T) 0.40,16 / 0.95,32 / 64 分別代表 λ 為 0.40 和潛藏主題數為 16（發展集）、λ 為 0.95 和潛藏主題數為 32（測試集）與潛藏主題數為 64。

由表 4. 24~表 4. 26 以單詞為特徵單位的結果顯示（HMM 與 TMM 使用期望值最大化演算法，即第三列數據），在低摘要比例（20% 或 30%）時，TMM 與 HMM-Type2，其結果不論在人工轉寫或自動轉寫上，均優於其它摘要模型，並以 TMM 為較佳。

此外，比較 HMM 與 TMM，於發展集與測試集的數據（即第一列、第二列數據）顯示，因發展集與測試集不一致使得參數的變化不一，在發展集較佳的參數拿到測試集使用時，其結果有所降低。另一方面，期望值最大化演算法所得到的結果，大致與測試集所得到的較佳結果不分軒輊。

表 4. 24 摘錄方法比較(評估方式：MAP，特徵單位：詞，人工轉寫)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1 0.90 / 0.75	HMM-Type2 0.40 / 0.95	TMM P(D|T) 0.40,16 / 0.95,32 / 64	TMM P(T|T) 0.35,08 / 0.95,16 / 32
20%	0.3517	0.3606	0.2653	0.3756	0.1828 0.1928 0.1828	0.3739 0.3877 0.3856	0.3739 0.3888 0.3906	0.3772 0.3881 0.3989
30%	0.5139	0.5185	0.4260	0.5132	0.3158 0.3269 0.3055	0.5264 0.5498 0.5435	0.5264 0.5500 0.5504	0.5283 0.5494 0.5486
50%	0.7456	0.7575	0.7214	0.7518	0.6191 0.6203 0.6135	0.7550 0.7637 0.7621	0.7591 0.7626 0.7589	0.7568 0.7644 0.7591
70%	0.8328	0.8437	0.8201	0.8490	0.7382 0.7403 0.7307	0.8427 0.8487 0.8428	0.8443 0.8489 0.8466	0.8437 0.8502 0.8448

表 4. 25 摘錄方法比較(評估方式：MAP，特徵單位：詞，自動轉寫 SP_WG)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1 0.70 / 0.80	HMM-Type2 0.70 / 0.10	TMM P(D|T) 0.55,64 / 0.10,64 / 64	TMM P(T|T) 0.55,32 / 0.10,64 / 32
20%	0.2683	0.2683	0.2317	0.3050	0.1928 0.2011 0.2011	0.3000 0.3406 0.3272	0.3122 0.3439 0.3289	0.3122 0.3472 0.3306
30%	0.3853	0.3787	0.3430	0.3956	0.3028 0.3067 0.3058	0.4100 0.4592 0.4459	0.4308 0.4597 0.4457	0.4292 0.4607 0.4450
50%	0.6845	0.6706	0.6240	0.6876	0.6088 0.6089 0.6036	0.6978 0.6910 0.6959	0.7025 0.6928 0.7003	0.7020 0.6964 0.6957
70%	0.7674	0.7542	0.7364	0.7690	0.7392 0.7379 0.7289	0.7706 0.7716 0.7771	0.7747 0.7747 0.7818	0.7730 0.7742 0.7774

表 4. 26 摘錄方法比較(評估方式：MAP，特徵單位：詞，自動轉寫 SP_Adapt)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1 0.70 / 0.95	HMM-Type2 0.45 / 0.4	TMM P(D|T) 0.55,64 / 0.55,64 / 64	TMM P(T|T) 0.40,04 / 0.40,04 / 32
20%	0.2742	0.2769	0.2350	0.2994	0.1944 0.2028 0.2044	0.3050 0.3050 0.3117	0.3117 0.3050	0.3117 0.3117
30%	0.4100	0.4097	0.3505	0.3996	0.3078 0.3089 0.3067	0.4322 0.4339 0.4359	0.4322 0.4312	0.4355 0.4356
50%	0.6930	0.6861	0.6485	0.6828	0.6093 0.6082 0.6007	0.7061 0.7059 0.7045	0.7070 0.7052	0.7090 0.7030
70%	0.7749	0.7612	0.7474	0.7708	0.7455 0.7341 0.7312	0.7870 0.7872 0.7843	0.7871 0.7856	0.7878 0.7842

由表 4. 27~表 4. 29 以雙音節為特徵單位的結果顯示(HMM 與 TMM 使用期望值最大化演算法，即第三列數據)，在低摘要比例（20%）時，TMM 與 HMM-Type2，其結果不論在人工轉寫或自動轉寫上，均優於其它摘要模型，並以 TMM 為較佳。此外 HMM-Type2 與 TMM，其在發展集所得的參數 λ 均為 1.00，是以均退化為僅使用每一字句產生索引特徵的機率值， $p(w|S_i)$ 。

表 4. 27 摘錄方法比較(評估方式：MAP，特徵單位：雙音節，人工轉寫)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1 0.10 / 0.05	HMM-Type2 1.00 / 1.00	TMM P(D|T) 1.00 / 1.00 / 04	TMM P(T|T) 1.00 / 1.00 / 16
20%	0.3753	0.3725	0.2994	0.4081	0.1761 0.1782 0.1475	0.4369 0.4153	0.4369 0.4156	0.4369 0.4036
30%	0.5459	0.5389	0.4450	0.5721	0.3383 0.3406 0.3118	0.5794 0.5658	0.5794 0.5678	0.5794 0.5611
50%	0.7550	0.7582	0.6752	0.7733	0.6382 0.6440 0.6125	0.7834 0.7752	0.7834 0.7769	0.7834 0.7786
70%	0.8370	0.8491	0.7854	0.8506	0.7285 0.7386 0.7253	0.8552 0.8508	0.8552 0.8516	0.8552 0.8508

表 4. 28 摘錄方法比較(評估方式：MAP，特徵單位：雙音節，自動轉寫 SP_WG)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1 0.20 / 0.20	HMM-Type2 1.00 / 0.30	TMM P(D|T) 1.00 / 0.30,16 / 04	TMM P(T|T) 1.00 / 0.20,02 / 16
20%	0.3122	0.3133	0.2367	0.2883	0.2322 0.2306	0.2983 0.3467 0.3500	0.2983 0.3467 0.3500	0.2983 0.3467 0.3522
30%	0.4326	0.4209	0.3472	0.4148	0.3742 0.3623	0.4270 0.4584 0.4642	0.4270 0.4584 0.4642	0.4270 0.4554 0.4579
50%	0.7272	0.7141	0.6328	0.7137	0.6507 0.6317	0.7053 0.7288 0.7295	0.7053 0.7297 0.7287	0.7053 0.7281 0.7272
70%	0.7949	0.7853	0.7287	0.7931	0.7328 0.7264	0.7909 0.7893 0.7907	0.7909 0.7894 0.7898	0.7909 0.7885 0.7894

表 4. 29 摘錄方法比較(評估方式：MAP，特徵單位：雙音節，自動轉寫 SP_Adapt)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1 0.05 / 0.05	HMM-Type2 1.00 / 0.30	TMM P(D|T) 1.00 / 0.30,16 / 04	TMM P(T|T) 1.00 / 0.20,02 / 16
20%	0.2942	0.2969	0.2267	0.2775	0.2317 0.2322	0.3092 0.3375 0.3350	0.3092 0.3375 0.3350	0.3092 0.3350 0.3350
30%	0.4268	0.4175	0.3399	0.4148	0.3861 0.3575	0.4448 0.4609 0.4639	0.4448 0.4609 0.4623	0.4448 0.4584 0.4571
50%	0.7180	0.7107	0.6217	0.7187	0.6612 0.6256	0.7102 0.7250 0.7318	0.7102 0.7256 0.7282	0.7102 0.7269 0.7300
70%	0.7953	0.7951	0.7297	0.8009	0.7518 0.7211	0.7972 0.7969 0.7979	0.7972 0.7961 0.7960	0.7972 0.7957 0.7969

由表 4. 30~表 4. 32 以雙字為特徵單位的結果顯示(HMM 與 TMM 使用期望值最大化演算法，即第三列數據)，向量空間模型為基礎的摘要模型（VSM 與 RM）有較佳的結果。此外在低摘要比例（20%）時，TMM 其結果大致均優於其它摘要模型；在高摘要比例（50%）時，VSM 其結果均優於其它摘要模型。

表 4. 30 摘錄方法比較(評估方式：MAP，特徵單位：雙字，人工轉寫)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1 0.05 / 0.05	HMM-Type2 1.00 / 0.95	TMM P(D|T) 1.00 / 0.9,04 / 64	TMM P(T|T) 1.00 / 0.9,64 / 64
20%	0.3983	0.3992	0.2924	0.4017	0.2049 0.1742	0.3944 0.4003 0.3917	0.3944 0.4103 0.4125	0.3944 0.4103 0.4053
30%	0.5683	0.5744	0.4435	0.5696	0.3981 0.3550	0.5653 0.5753 0.5599	0.5653 0.5786 0.5612	0.5653 0.5683 0.5567
50%	0.7798	0.7760	0.6944	0.7701	0.6504 0.6401	0.7748 0.7735 0.7714	0.7748 0.7758 0.7679	0.7748 0.7740 0.7726
70%	0.8520	0.8598	0.7883	0.8490	0.7367 0.7288	0.8488 0.8508 0.8498	0.8488 0.8513 0.8479	0.8488 0.8488 0.8483

表 4. 31 摘錄方法比較(評估方式：MAP，特徵單位：雙字，自動轉寫 SP_WG)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1 0.05 / 0.25	HMM-Type2 1.00 / 1.00	TMM P(D|T) 1.00 / 1.00 / 64	TMM P(T|T) 1.00 / 1.00 / 64
20%	0.3122	0.3089	0.2217	0.2833	0.2333 0.2650 0.2322	0.3106 0.2872	0.3106 0.2956	0.3106 0.2839
30%	0.4232	0.4104	0.3482	0.4152	0.4080 0.4239 0.3742	0.4421 0.4113	0.4421 0.4163	0.4421 0.4080
50%	0.7282	0.7049	0.6478	0.7102	0.6693 0.6728 0.6632	0.7132 0.7144	0.7132 0.7161	0.7132 0.7153
70%	0.8059	0.7984	0.7407	0.7936	0.7660 0.7565 0.7473	0.7977 0.7893	0.7977 0.7873	0.7977 0.7892

表 4. 32 摘錄方法比較(評估方式：MAP，特徵單位：雙字，自動轉寫 SP_Adapt)

摘要比例	VSM	RM	LSA	embedded LSA	HMM-Type1 0.05 / 0.25	HMM-Type2 1.00 / 1.00	TMM P(D|T) 1.00 / 1.00 / 64	TMM P(T|T) 1.00 / 1.00 / 64
20%	0.2950	0.2867	0.2150	0.2750	0.2272 0.2711 0.2350	0.3056 0.2933	0.3056 0.3072	0.3056 0.2939
30%	0.4245	0.4120	0.3339	0.4146	0.3890 0.4261 0.3853	0.4471 0.4309	0.4471 0.4390	0.4471 0.4293
50%	0.7215	0.7043	0.6416	0.7085	0.6677 0.6738 0.6461	0.7110 0.7184	0.7110 0.7196	0.7110 0.7186
70%	0.8068	0.8026	0.7366	0.7997	0.7711 0.7577 0.7424	0.8020 0.7962	0.8020 0.7972	0.8020 0.7995

4.5 綜合比較

由表 4. 33 以單詞為特徵單位的結果顯示（使用期望值最大化演算法），對於主題混合模型而言，隨著潛藏主題的增加其結果有所提升。

表 4. 33 不同潛藏主題個數比較(評估方式：MAP，特徵單位：詞，人工轉寫)

摘要比例	2	4	8	16	32	64
20%	0.3856	0.3856	0.3856	0.3856	0.3856	0.3906
30%	0.5435	0.5441	0.5441	0.5418	0.5428	0.5504
50%	0.7609	0.7618	0.7614	0.7603	0.7575	0.7589
70%	0.8427	0.8435	0.8441	0.8425	0.8436	0.8466

由表 4. 34~表 4. 36 的結果顯示（使用期望值最大化演算法），不論在人工轉寫或自動轉寫上，雙音節為特徵單位的結果大致均較其它特徵單位來得好，且模型中以 HMM-Type2 及 TMM 為較佳。此外，雙音節、雙字其結果均較以單詞為特徵單位來得好。

表 4. 34 不同特徵單位比較(評估方式：MAP，摘要比例：20%，人工轉寫)

特徵單位	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T)	TMM P(T|T)
W	0.3517	0.3606	0.2653	0.3756	0.1828	0.3856	0.3906	0.3989
S2	0.3753	0.3725	0.2994	0.4081	0.1475	0.4153	0.4156	0.4036
C2	0.3983	0.3992	0.2924	0.4017	0.1742	0.3917	0.4125	0.4053

表 4. 35 不同特徵單位比較(評估方式：MAP，摘要比例：20%，自動轉寫 SP_WG)

特徵單位	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T)	TMM P(T|T)
W	0.2683	0.2683	0.2317	0.3050	0.2011	0.3272	0.3289	0.3306
S2	0.3122	0.3133	0.2367	0.2883	0.2306	0.3500	0.3500	0.3522
C2	0.3122	0.3089	0.2217	0.2833	0.2322	0.2872	0.2956	0.2839

表 4. 36 不同特徵單位比較(評估方式：MAP，摘要比例：20%，自動轉寫 SP_Adapt)

特徵單位	VSM	RM	LSA	embedded LSA	HMM-Type1	HMM-Type2	TMM P(D|T)	TMM P(T|T)
W	0.2742	0.2769	0.2350	0.2994	0.2044	0.3117	0.3050	0.3117
S2	0.2942	0.2969	0.2267	0.2775	0.2322	0.3350	0.3350	0.3350
C2	0.2950	0.2867	0.2150	0.2750	0.2350	0.2933	0.3072	0.2939

4.6 隱藏式馬可夫模型與主題混合模型進一步實驗

對於隱藏式馬可夫模型-型一，可由式(3.5) 做句子擴充。首先每一測試文件與表 4. 4 語料庫，藉由向量空間表示法估測餘弦分數，並依分數大小選取前 50 名最相關的文件。接著由測試文件中的各字句，與此前 50 名文件的各字句依同樣方式，找到前 k 句相關的字句組成 $\widehat{S_i}$ ，以做句子擴充。

由表 4. 37 可知，以文件為模型時，擴展觀測（Observation）的字句，對自動摘要正確率有顯著的提升。

表 4. 37 HMM-Type1 字句擴充比較(評估方式：MAP，特徵單位：詞，人工轉寫)

摘要比例	λ=0.9	λ=0.9 k=23(加入 23 句)
20%	0.1828	0.3353
30%	0.3158	0.5168
50%	0.6191	0.7528
70%	0.7382	0.8313

對於隱藏式馬可夫模型-型二，可由式(3.10) 做字句移除。

由表 4. 38 可知，以字句為模型時，移除觀測（Observation）中模型本身的字句。於人工轉寫文件上，有助於自動摘要正確率的提升；然而在自動轉寫文件上，因辨識錯誤及斷句不同，抵銷了其結果，但是可以發現在辨識率上升時（SP_Adapt 較 SP_WG 好），其結果有所提升。

表 4. 38 HMM-Type2 字句移除比較(評估方式：MAP，特徵單位：詞，人工轉寫)

摘要比例	人工轉寫		SP_WG		SP_Adapt	
	HMM-Type2 λ=0.4	HMM-Type2 λ=0.4 Sentence Removal	HMM-Type2 λ=0.7	HMM-Type2 λ=0.7 Sentence Removal	HMM-Type2 λ=0.45	HMM-Type2 λ=0.45 Sentence Removal
20%	0.3739	0.3844	0.3000	0.2906	0.3050	0.3067
30%	0.5264	0.5420	0.4100	0.3800	0.4322	0.4128
50%	0.7550	0.7635	0.6978	0.6910	0.7061	0.6997
70%	0.8427	0.8457	0.7706	0.7604	0.7870	0.7774

對於主題混合模型，可由式(3.21) 做字句移除。

由表 4. 39~表 4. 40 可知，以字句為模型的主題混合模型，在移除觀測（Observation）中模型本身的字句。於人工轉寫文件上，有助於自動摘要正確率的提升；然而在自動轉寫文件上，因辨識錯誤及斷句不同，抵銷了其結果，但是可以發現在辨識率上升時（SP_Adapt 較 SP_WG 好），其結果有所提升。

表 4. 39 TMM p(D|T) 字句移除比較(評估方式：MAP，特徵單位：詞，人工轉寫)

摘要比例	Txt		SP_WG		SP_Adapt	
	TMM P(D|T) 0.40, 16	TMM P(D|T) 0.40,16 Sentence Removal	TMM P(D|T) 0.55,64	TMM P(D|T) 0.55,64 Sentence Removal	TMM P(D|T) 0.55,64	TMM P(D|T) 0.55,64 Sentence Removal
20%	0.3739	0.3828	0.3122	0.2872	0.3117	0.2967
30%	0.5264	0.5420	0.4308	0.3783	0.4322	0.4078
50%	0.7591	0.7651	0.7025	0.6891	0.7070	0.6962
70%	0.8443	0.8466	0.7747	0.7594	0.7871	0.7734

‘

表 4. 40 TMM p(T|T) 字句移除比較(評估方式：MAP，特徵單位：詞，人工轉寫)

摘要比例	txt		SP_WG		SP_Adapt	
	TMM P(T|T) 0.35,08	TMM P(T|T) 0.35,08 Sentence Removal	TMM P(T|T) 0.55,32	TMM P(T|T) 0.55 ,32 Sentence Removal	TMM P(T|T) 0.40,04	TMM P(T|T) 0.40,04 Sentence Removal
20%	0.3772	0.3828	0.3122	0.2922	0.3117	0.3067
30%	0.5283	0.5420	0.4292	0.3806	0.4355	0.4128
50%	0.7568	0.7662	0.7020	0.6905	0.7090	0.6992
70%	0.8437	0.8472	0.7730	0.7608	0.7878	0.7801

在主題混合模型中，進一步比較初始化 $p(T_k \mid S_i)$ 時，使用均勻分佈（Uniform）是否較好；換句話說，即假設字句 S_i 產生每一潛藏主題的機率皆一致，即

$$p(T_k \mid S_i) = \frac{1}{K} \tag{4. 8}$$

並且比較 S_i 迭代次數，即式(3.15)、(3.16) 的執行次數。

由表 4. 41 可知，使用均勻分佈在低摘要比例（20%）下，有不錯的表現。

表 4. 41 TMM p(D|T) 初始方式比較(評估方式：MAP，特徵單位：詞，人工轉寫)

摘要比例	1 次	100 次	Uniform 1 次	Uniform 100 次
20%	0.3828	0.3828	0.3844	0.3828
30%	0.5420	0.5424	0.5420	0.5424
50%	0.7651	0.7627	0.7640	0.7636
70%	0.8466	0.8454	0.8458	0.8463

4.7 本章小結

經由實驗結果顯示，於摘要模型比較上：使用隱藏式馬可夫模型或主題混合模型其結果較其它常見方法有顯著的提升，同時主題混合模型在幾乎所有情況下均較隱藏式馬可夫模型來得佳；於特徵單位的比較上：使用雙音節與雙字時，其結果優於使用詞為特徵單位。

此外對於隱藏式馬可夫模型-型一使用字句擴展，其結果有顯著的提升。對於隱藏式馬可夫模型-型二與主題混合模型，使用字句移除，在人工轉寫文件上均有效提升摘要結果；然而在自動轉寫上，因辨識錯誤及斷句不同，抵銷了其結果，但是可以發現在辨識率上升時，其結果有所提升。

最後在主題混合模型的初始化使用均勻分佈，在低摘要比例下，有不錯的表現。

第 5 章 自動摘要於文件分類上之應用

自動摘要可視為去除網頁雜訊（如版權宣告、廣告）的技術；也可視為一種特徵抽取的方式，將網頁重要的字句摘錄出來，用以訓練分類器來預測新網頁所屬的類別 [Shen *et al.* 2004] ，此外以摘要後的文件做索引，也可加速計算速度與減少儲存空間。

本章以自動摘要為基礎，提出主題混合模型分類器，用來做文件分類，並與 *K*-最近鄰（*K*-Nearest-Neighbor, *K*NN）做比較，並實驗在不同摘要比例下的結果。

自動文件分類，是根據文件內容或標題決定其類別的流程，其目的在於對文件進行分門別類的加值處理，讓文件易於管理、利用。例如，新聞文件可依其內容，給予「政治」、「經濟」、「社會」等類別資訊，方便使用者瀏覽取得其所需資訊。而類別的制定一般以使用者自定為主，傳統上的分類工作都藉由人工，然而在文件繁多、類別日益增大的情況下，此一工作變得日益艱難，是以經由文件分類就可節省大量的人力及提升分類的效率。

在進行文件分類時，需要瞭解文件的內文大意才能據此給予類別，這需要相當高階的知識處理，然而目前自然語言理解的技術，尚無法讓電腦瞭解任意的自由字句。 因此電腦在做文件分類時，常將文件分解成一個個語意較小的單位， 通常為文件的關鍵詞彙，再從這些詞彙與類別中找出對應的關係。有時分類的問題，簡單到只要文件的某個欄位中出現什麼特徵詞， 就分到什麼類別去。但在一般情況下，在進行文件分類前首先必須歸納出分類時的規則，如此電腦才能據以執行。而在類別龐大時，分類規則就難以用人工分析而得。所以，電腦在進行自動分類之前必須加以訓練，使其自動學習出人工分類的經驗與知識，此一流程即機器學習（Machine Learning）所欲達到的目標。

5.1 分類（Classification）與分群（Clustering）

文件分類（Document Classification）是將文件依據其內容指定為一個或多個事先

定義好的文件類別的過程。而文件分群（Document Clustering）為將相似的文件放置同一群中，且讓不相似的文件在另一群；然而分群無法給予所得到的類別（Class）一個綜合性的描述。文件分群與文件分類最主要的差異在於：文件分類是擷取文件特徵並與文件類別的特徵作比較，再依照其相關程度進行分類；而文件分群亦是擷取文件特徵並進行比對，但文件分群並不需要事先定義文件類別，而是依照各文件之間之相關程度進行分群。

本論文後續章節將繼續對文件分類的相關議題進行探討。

5.2 特徵抽取

在文件分類上常使用特徵抽取，以得到代表某一類別的詞彙，經由這些有鑑別力的詞彙我們可以加速運算、提高正確度及避免過度學習 [Joachims 1998]。

常見的方法有文件頻門閥值（Document Frequency Thresholding）[Yang *et al.* 1997] 、互斥資訊量（Mutual Information，MI）、條件式互斥資訊量（Conditional Mutual Information，CMI）[Wang *et al.* 2004]，敘述如下：

文件頻門閥值

計算每一詞在訓練語料庫中的文件頻率，並設定門閥值用以移除低於其值的字詞，其假設在於較少出現的字詞，對分類預測較無資訊，且不影響整體準確度；此方法也可移除那些干擾字詞。

互斥資訊量

計算每一詞 t，與一類別 c 之 MI 值

$$\mathrm{MI(t;c)}=\log\frac{p(t,c)}{p(t)\,p(c)}\approx\log\frac{A\times N}{(A+C)\times(A+B)} \tag{5.1}$$

其中 A 是 t 與 c 共同出現的次數，B 是 t 出現 c 不出現的次數，

C 是 c 出現 t 不出現的次數，N 是總文件數。

條件式互斥資訊量：$I(F_k;C\,|\,F_1,...,F_{k-1})$

其中 C 為某一類，F 為特徵值，經由 CMI 我們可以得到 JMI（Joint Mutual

Information）值為

$$I(F_1,...,F_k;C) = I(F_1,...,F_{k-1};C) + I(F_k;C \mid F_1,...,F_{k-1}) \quad (5.2)$$

其計算步步驟如下

開始	選取最大 MI 之項為特徵值
迭代	設已選取 *k*-1 特徵值，使得 JMI (Joint Mutual Information)值最大，選取使 CMI 最大的特徵值加入之，使 JMI 值最大

5.3 分類器（Classifier）

分類器，是一種文件歸納的處理過程，用以決定某個文件屬於某個分類[Sebastiani 2002]。對於某一類別 $c_i \in C$，給定某一文件 d_j，分類器計算對於每一類的類別狀態值（Categorization Status Value, CSV），不同類型的分類器，其 CSV 值域會有所不同。

$$CSV_i : D \rightarrow [0,1] \quad (5.3)$$

5.3.1 空間向量模型（Vector Space Model, VSM）

此模型將文件用一多維空間之向量來表示，藉由計算新進文件與分類規則向量，兩向量之間的相關程度函數，餘弦（Cosine），得到每一類別的 CSV 值，再經排序得到類別相關程度排名，最後可藉由設定門閥值判斷文件是否符合特定類型之文件，或者使用相關度最高的類別當作所屬類別。

5.3.2 單純貝式（Naïve Bayes, NB）模型

NB 模型假設索引單位之間相互獨立，在新進一篇新文件 $\vec{d}$ 時，估計給定特徵值下每個類別 C_i 的機率，以得到 CSV 值

$$p(C = c_i \mid \vec{d}) = \frac{p(\vec{d} \mid C = c_i)\, p(C = c_i)}{p(\vec{d})} \quad (5.4)$$

對於 $p(C = c_i)$ 簡化假設其為均勻分佈（Uniform distribution），也就是對於全

部的類別是相同的， $p(\vec{d})$ 不影響結果是以可以省略，此外在估計類別為 c_i 的情況下產生文件 $\vec{d}$ 的機率， $p(\vec{d} \mid C = c_i)$ ，假設各特徵值間是互相獨立的，是以可進一步簡化為：

$$p(\vec{d} \mid C = c_i) = \prod_{w \in \vec{d}} p(w \mid C = c_i) \tag{5. 5}$$

5.3.3 *K*-最近鄰（*K*-Nearest-Neighbor, *K*NN）

*K*NN 分類器在學習階段並不會像 Naive Bayes 分類器會產生或記錄每個類別的特徵，相對地只是簡單的將訓練文件中每筆資料以適當的表示法予以儲存，如此便完成其訓練工作。當有一筆測試資料集中的文件資料需要進行分類時，*K*NN 分類器會將欲進行分類的文件資料與所有訓練資料集中的文件資料逐一計算相關度，找出 *K* 筆最相近的訓練資料，再依據這 *K* 個訓練資料所屬的類別，來決定此測試資料最後所屬的類別 [V. Tam *et al.* 2002] 。*k*NN 法的過程可以下列公式表示

$$y(q, c_i) = \sum_{d_j \in kNN} sim(q, d_j) y(d_j, c_i) \tag{5. 6}$$

其中 $y(q, c_i)$ 代表類別 c_i 對新進文件 q 的 CSV 值、$y(d_j, c_i) \in \{\text{true}, false\}$ 用以表示文件 d_j 是否屬於類別 c_i ，而 $sim(q, d_j)$ 表示測試文件 q 與訓練文件 d_j 之問的相關程度（可利用餘弦或其它公式），$d_j \in KNN$ 代表與測試文件 q 最相關的 k 筆文件。

5.3.4 分類器比較

綜合上述，整理如下表所示

表 5. 1 分類器比較

分類器	優點	缺點
空間向量模型	容易計算、快速分類	準確度較不佳
單純貝式模型	計算容易、快速分類	簡化假設使準確度不足
*K*NN	訓練資料少時仍有不錯之分類準確度	訓練資料增加會造成算速度過慢

5.4 主題混合模型分類器

由 2.7 節關於主題混合模型的討論，由式(2.17)可得：

$$p(Q \mid D_i) \approx \prod_{n=1}^{N} \sum_{k=1}^{K} p(q_n \mid T_k) p(T_k \mid D_i)$$

由此模型，我們可得到機率值 $p(q_n \mid T_k)$ 與 $p(T_k \mid D_i)$。如果將每一潛藏主題視為一類別，並於分類時即時迭代更新，得到某一新進文件 N 的機率值，$p(T_k \mid N)$，代表類別的 CSV 值。最後由最大 $p(T_k \mid N)$ 值，所對應的類別代表所推薦的類別，即可完成分類流程，詳述如下：

訓練階段：

設定潛藏主題大小為類別的種類 K

給定一文件集，內含文件 D_i，並且每一文件已事先得知其所屬類別 T_k

(a) 初始化

$$p(T_k \mid D_i) = \begin{cases} 0.99999999 & \text{if Di} \in \mathrm{T_k} \\ \dfrac{0.00000001}{\text{K-1}} & \text{if } \mathrm{D_i} \notin T_k \quad \text{(close to 0)} \end{cases}$$

在類別內的文件，其文件產生類別的機率 $p(T_k \mid D_i)$ 為接近 1 的值，否則設為很小的值

$p(q_n \mid T_k)$ ，由主題單連語言模型而來

(b) 迭代

使用非監督式訓練式(2.23)、(2.24)，迭代更新 $p(q_n \mid T_k)$ 與 $p(T_k \mid D_i)$

測試階段：

新進一篇新文件 N

(a) 初始化

預設新文件 N 產生各類別的機率是均勻分佈，$p(T_k \mid N) = \dfrac{1}{\mathrm{K}}$

(b) 迭代

$$\hat{P}(T_k \mid N) = \frac{\sum_{q_s \in N} n(q_s, N)\, p(T_k \mid q_s, N)}{|N|} \tag{5. 7}$$

$$p(T_k \mid q_s, N) = \frac{p(T_k \mid N)\, p(q_s \mid T_k)}{\sum_{l=1}^{K} p(T_l \mid N)\, p(q_s \mid T_l)} \tag{5. 8}$$

其中 $p(q_s \mid T_k)$ 由訓練階段所得的主題單連語言模型而來

(c)決策

由 $p(T_k \mid N)$ 的大小，找最大的值所對應的類別，視為新進文件 N 的類別

5.5 實驗設定

本實驗使用東森新聞做語料庫 [東森新聞報]，相關統計如表 5.2 所示：

表 5. 2 東森新聞語料相關統計

	發展集	測試集
類別	共十類 {政治、財經、社會、地方、兩岸、國際、生活、綜藝、資訊、運動}	
新聞時間	2003 年 1 月	2003 年 2 月
新聞數	5533 則	4632 則

在文獻中可發現 *K*NN 為目前在分類結果上較佳的分類器之一 [Yang *et al.* 1999]，是以在基礎實驗上使用 *K*NN 與所提出主題混合模型分類器做一比較。在 *K*NN 分類器中，使用向間向量模型來表達每一文件，並使用餘弦估測相關度。在實驗時，將發展集由向量空間模型模型做摘要，再利用固定測試集做測試，以觀察自動摘要對於分類器正確率的提升是否有所助益。

在實驗評估方面，使用 MicroF 以及 MacroF 值同時呈現分類的效果，其計算方式如下：

$$\text{MicroF}=\frac{2\times\sum_{i=1}^{C}TP_i}{2\times\sum_{i=1}^{C}TP_i+\sum_{i=1}^{C}FP_i+\sum_{i=1}^{C}FN_i} \quad (5.9)$$

$$\text{MacroF} = \frac{1}{C}\sum_{i=1}^{C}\frac{2\times TP_i}{2\times TP_i+FP_i+FN_i} \quad (5.10)$$

其中 C 是類別總數，i 代表某一類別，而 TP_i（True Positive）、FP_i（False Positive）、FN_i（False Negative），分別代表 ：是類別 i 而且也正確分為類別 i 的文件數、不是 i 類卻分為 i 類的文件數、是 i 類卻沒有分為 i 類的文件數。

由於 MicroF 是全部文件一起累加統計，不分類別，因此容易受到大類別（佔大多數文件）表現好壞的影響。相對的，MacroF 考慮每個類別的成效後再做平均，因此容易受到大量的小類別影響。將兩種平均數據都報告出來，可以瞭解大多數文件的分類效果（MicroF），以及大多數類別的分類效果（MacroF）。

5.6 實驗結果

在表 5. 3~表 5. 4 的 1~25、26~50 等，代表選擇 K 個文件的出處，如 26~50 代表由相關度排名介於 26~50 名的文件。

由表 5. 3~表 5. 4 實驗結果顯示：

1. 當 K 值選取愈大時，MicroF 與 MacroF 均有變好的趨勢，但過大時因雜訊的增多，其結果慢慢變差。
2. 經由自動摘要過後，其 MicroF 與 MacroF 值大都降低，這與預期結果相反，這可能是因為在做摘要的過程當中，因只保留對文件重要的資訊，而沒有考量與類別的關係，也就是說，有可能將對類別有高鑑別性的字句給去除，是以造成結果的下降。

表 5. 3 *K*NN 於測試集 MicroF 值

K	25				50		100
摘要比例	1~25	26~50	51~75	76~100	1~50	51~100	1~100
0.2	0.5782	0.5365	0.5214	0.4989	0.5868	0.5402	0.5827
0.3	0.5786	0.5412	0.5276	0.5060	0.5885	0.5443	0.5872
0.5	0.5883	0.5501	0.5348	0.5140	0.6013	0.5538	0.595
0.7	0.5874	0.5570	0.5432	0.5179	0.5984	0.5548	0.5978
1.0	0.5954	0.5745	0.5475	0.5352	0.6056	0.5715	0.6066

表 5. 4 *K*NN 於測試集 MacroF 值

K	25				50		100
摘要比例	1~25	26~50	51~75	76~100	1~50	51~100	1~100
0.2	0.5692	0.5251	0.5038	0.4817	0.5771	0.5206	0.5699
0.3	0.5712	0.5306	0.5113	0.4840	0.5788	0.5237	0.5732
0.5	0.5827	0.5378	0.5197	0.4942	0.5921	0.5365	0.582
0.7	0.5820	0.5452	0.5278	0.4988	0.5898	0.5360	0.5855
1.0	0.5891	0.5644	0.535	0.5182	0.5967	0.5563	0.595

由表 5. 5~表 5. 6 實驗結果顯示

1. TMM 於摘要比例 1.0（即不做自動摘要）迭代 1 次時，比 *K*NN 略差。但經由多次迭代後，不論在 MicroF 與 MacroF 均較 *K*NN 分類器來得好，且迭代的次數愈多，其結果愈見明顯。
2. TMM 在摘要後，其結果不論在 MicroF 與 MacroF 大致均較 *K*NN 分類器來得好，且迭代的次數愈多，其結果愈見明顯。
3. 經由自動摘要過後，*K*NN 與 TMM 分類器，其 MicroF 與 MacroF 值大都降低。

表 5. 5 TMM 與 *K*NN 分類器，於測試集 MicroF 值比較

摘要比例	*K*NN	TMM 1 次迭代	TMM 50 次迭代	TMM 100 次迭代
0.2	0.5868	0.5939	0.5954	0.5961
0.3	0.5885	0.5939	0.5941	0.5961
0.5	0.6013	0.6015	0.6051	0.6066
0.7	0.5984	0.5987	0.6058	0.6060
1.0	0.6056	0.6015	0.6077	0.6101

表 5. 6 TMM 與 *K*NN 分類器，於測試集 MacroF 值比較

摘要比例	*K*NN	TMM 1 次迭代	TMM 50 次迭代	TMM 100 次迭代
0.2	0.5771	0.5834	0.5871	0.5886
0.3	0.5788	0.5843	0.5864	0.5891
0.5	0.5921	0.5909	0.5962	0.5985
0.7	0.5898	0.5885	0.5978	0.5983
1.0	0.5967	0.5920	0.6006	0.6033

5.7 本章小結

文件自動摘要的目的，是將文件縮減濃縮成重要字句，並去除冗餘的訊息。此外，摘要後的文件，因資料量較少，也可提升後續文件處理的效率。基於這樣的觀察，自動摘要的技術，可能有助於文件的自動分類。然而經由實驗，結果並未如預期，自動摘要雖然提升了自動分類文件的效率，卻因損失一些分類資訊，使分類文件的精確度降低。

初步實驗結果顯示，主題混合模型分類器較常見 *K*-最近鄰（*K*-Nearest-Neighbor, *K*NN）分類器在 MicroF 與 MacroF 分類結果上，有些微的提升。

第 6 章 結論與展望

6.1 結論

本論文於自動摘要方面，在逐字比對方式上應用隱藏式馬可夫模型（Hidden Markov Model, HMM）做為摘要模型，並分為 HMM-Type1 及 HMM-Type2 二種類型；在概念比對上提出嵌入式潛藏語意分析（embedded LSA）與主題混合模型（Topical Mixture Model, TMM）做為摘要模型；在自動摘要評估上，提出以改良型字錯誤率（modified Character Error Rate, m-CER）為基礎的平均精確度（Mean Average Precision, MAP）評估方式，以解決自動轉寫與人工轉寫文件因斷句不一致，所造成摘要結果無法評估相關的問題。

經由實驗結果顯示，於特徵單位比較上：使用雙音節與雙字時，其結果優於使用詞為特徵單位；於摘要模型比較上：使用隱藏式馬可夫模型或主題混合模型其結果較其它常見方法有顯著的提升，同時主題混合模型在幾乎所有情況下均較隱藏式馬可夫模型來得佳。

此外對於隱藏式馬可夫模型-型一使用字句擴展能有效增進摘要正確率；對於隱藏式馬可夫模型-型二與主題混合模型中做字句移除，在人工轉寫文件上均有效提升摘要結果，然而在自動轉寫上，因辨識錯誤及斷句不同，抵銷了其結果，但是可以發現在辨識率上升時，其結果有所提升。

另一方面，在主題混合模型的初始化使用均勻分佈，在低摘要比例下有較佳的結果。

最後本論文提出主題混合模型分類器，初步實驗結果顯示，主題混合模型分類器較常見 *K*-最近鄰（*K*-Nearest-Neighbor, *K*NN）分類器在 MicroF 與 MacroF 分類結果上，有些微的提升；然經過自動摘要前處理後，此二者的分類結果均略顯降低，進一步的來說，自動摘要雖然提升了自動分類文件的效率，卻因損失一些分類資訊，使分類的結果降低。

6.2 未來展望

在摘要模型上：

1. 依據文件屬性，動態結合屬性資訊（如新聞開頭、結論段落為重要字句）
2. 探討混合不同摘要模型的方法
3. 結合更多自然語言方面的資訊，如詞性（Part of Speech, POS）
4. 結合語音聲學上特性，如：音高、能量等
5. 結合分類資訊進行摘要，使自動摘要能具體幫助分類器做分類

在分類器上：

1. 對主題混合模型，進行特徵抽取，以選取具有代表性的字詞
2. 探探與其它分類器結合的可能性

自動摘要技術，在資訊爆炸時代裡愈顯重要，在與其它相關科技結合後，如語音辨識，我們可能擁有手機留言自動摘要服務、一台能自動摘要廣播新聞的數位收音機等，這些都將可能於未來逐步實現。

參考文獻

[東森新聞報] http://member.ettoday.com/newsflash/index.php

[中央通訊社] 中央通訊社（Central News Agency），
http://210.69.89.224/search/hypage.cgi?HYPAGE=login.htm

[葉鎮源 2002] 葉鎮源，『文件自動化摘要方法之研究及其在中文文件的應用』，國立交通大學資訊科學研究所，2002 年碩士論文。

[何遠 2003] 何遠，『中文口語文件自動摘要之初步研究』，碩士論文，國立臺灣大學電信工程學研究所，2003。

[黃上銘等 2003] 黃上銘、劉昭麟、高照明，『適性化線上英語聽寫測驗系統之研究』，2003 人工智慧模糊系統及灰色系統聯合研討會論文集 (TAAI'03)，CD-ROM。台灣台北，4-6 December 2003。

[黃建霖 2004] 黃建霖，『應用平行語料和語意相依法則於中文語音文件之摘要』，碩士論文，國立成功大學資訊工程學系碩士班 , 2004。

[Aas *et al.* 1999] K. Aas, L. Eikvil and R. B. Huseby, “Applications of hidden Markov chains in image analysis”, The Journal of Pattern Recognition Society, 32, pp.703-713, 1999.

[Baeza-Yates *et al.* 1999] Ricardo Baeza-Yates and Berthier Ribeiro-Neto，”Modern Information Retrieval” , 1999, pages 27-30，Addison Wesley.

[Ball and Hall 1967] Ball G.H., Hall D.J., “A Clustering Technique for Summarizing Multivariate Data”, Behavioral Science, 1967, Vol. 12, 153-155.

[Baum *et al.* 1966] Baum L.E., T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains", Annals of the Institute of Statistical Mathematics, 37, pp.1554-1563, 1966.

[Bellegarda 2000] J. R. Bellegarda., “Exploiting latent semantic information in statistical language modeling”. Proceedingsof the IEEE, 2000, 88(8):1279–1296.

[Berry and Browne, 1999] Berry, M.W. and Browne, M., “Understanding Search Engines: Mathematical Modeling and Text Retrieval”, Philadelphia SIAM, 1999.

[Chen *et al.* 2002] Berlin Chen, Hsin-Min Wang and Lin-shan Lee, “Discriminating Capability of Syllable-based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese”, IEEE Trans. on Speech and Audio Processing, Vol.10, No5, July 2002, pp. 303-314.

[Chen *et al.* 2004a] Berlin Chen, Hsin-min Wang, Lin-shan Lee, “A Discriminative HMM/N-Gram-Based Retrieval Approach for Mandarin Spoken Documents,” ACM Transactions on Asian Language Information Processing (ACM TALIP), Vol. 3, No. 2, June 2004, pp. 128-145.

[Chen *et al.* 2004b] Berlin Chen, Jen-Wei Kuo, Yao-Min Huang, Hsin-min Wang, “Statistical Chinese Spoken Document Retrieval Using Latent Topical Information” the 8th International Conference on Spoken Language Processing (ICSLP 2004), Vol. II, 1621-1625, Jeju island, South Korea, October 4-8, 2004.

[Chen 2005] Berlin Chen, “Exploring the Use of Latent Topical Information for Statistical Chinese Spoken Document Retrieval,” accepted for publication in Pattern Recognition Letters, 2005. (SCI Expanded, EI)

[Dempster *et al.* 1977] 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, Vol. 39, No. 1, 1-38.

[Duda and Hart 1973] Duda R.O., Hart P.E., 1973, “Pattern Classification and Scene Analysis”, John Wiley Sons.

[Edmundson 1969] Edmundson H.P., 1969. “New Methods in Automatic Extraction.” Journal of the ACM 16(2) pp 264-285.

[G. Furnas *et al.* 1988] G. Furnas, S. Deerwester, S. Dumais, T. Landauer, R. Harshman, L. Streeter and K. Lochbaum, “Information retrieval using a singular

value decomposition model of latent semantic structure," in The 11th International Conference on Research and Development in Information Retrieval, Grenoble, France: ACM Press, 1988, pp. 465--480.

[Giles *et al.* 2003] Giles, J.T., Wo, L., Berry, M.W. (2003), "GTP (General Text Parser) software for Tex mining." In Bozdogan, H. (Ed.). Statistical Data Mining and Knowledge Discover, , Boca Raton, FL CRC Press.

[Gong and Liu 2001] Yihong Gong，Xin Liu，"Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis"，ACM SIGIR，2001.

[Google] http://www.google.com

[Hirohata *et al.* 2005] Makoto Hirohata, Shinnaka Y., Iwano K. and Sadaoki Furui, "Sentence Extraction-based presentation summarization techniques and evaluation metrics", ICASSP05.

[Hovy and Marcu 1998] Eduard Hovy and Daniel Marcu "Automated Text Summarization Tutorial" , COLING/ACL 1998.

[Joachims 1998] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features.", In Procedings 10th European Conference on Machine Learning (ECML'98)，1998.

[KIM *et al.* 2004] WOOSUNG KIM and SANJEEV KHUDANPUR "Lexical Triggers and Latent Semantic Analysis for Cross-Lingual Language Model Adaptation", ACM Transactions on Asian Language Information Processing, Vol. 3, No. 2, June 2004, Pages 94–112.

[Lee *et al.* 2003]Lin-shan Lee, Yuan Ho, Jia-fu Chen, Shun-Chuan Chen, "Why is the Special Structure of the Language Important for Chinese Spoken Language Processing? -Examples on Spoken Document Retrieval, Segmentation, and Summarization", 2003.

[Levenshtein 1966] V. I. Levenshtein, "Binary codes capable of correctingdeletions,

insertions, and reversals," Cybernetics and Control Theory, Vol. 10, No. 8, pp. 707-710, 1966.

[Lin 2003] Lin, C.-Y. 2003, "ROUGE: Recall-oriented understudy for gisting evaluation.", http://www.isi.edu/~cyl/ROUGE/

[Luhn 1959] H. P. Luhn , "The Automatic Creation of Literature Abstracts." IBM Journal of Research and Development 159-165.

[Miller *et al.* 1999] Miller, D. R. H., Leek, T., Schwartz, R. "A Hidden Markov Model Information Retrieval System. In Proceedings of ACM SIGIR Conference on R&D in Information Retrieval",214-221.1999.

[News 98] http://www.news98.com.tw/

[Orasan 2002] Constantin Orasan, "Issues in Evaluation of Automatic Summarization", http://www.clg.wlv.ac.uk/papers/orasan-report-02.pdf , 2002.

[Rabiner *et al.* 1989] Rabiner, L. R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, Vol.77, No.22, pp.257-286, 1989.

[Rosenfeld 2000] Rosenfeld, R. 2000, "Two decades of statistical language modeling: Where do we go from here? ", Proc. of the IEEE, 88:1270-1278, August.

[Salamatian *et al.* 2001] Kave Salamatian, Sandrine Vaton, "Hidden Markov modeling for network communication channels", ACM SIGMETRICS 2001.

[Sebastiani 2002] Farbrizio Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, Vol.34, No.1, March 2002, pp.1-47.

[Shen *et al.* 2004] Dou Shen, Zheng Chen, Hua-Jun Zeng, Benyu Zhang, Qiang Yang, Wei-Ying Ma, Yuchang Lu, "Web-page Classification through Summarization", The 27th Annual International ACM SIGIR Conference (SIGIR'2004), 2004.

[Siivola *et al.* 2001] Vesa Siivola, Mikko Kurimo, Krista Lagus, "Large Vocabulary Statistical Language Modeling for Continuous Speech Recognition in Finnish",

EUROSPEECH 2001.

[SRI Toolkit] SRI International, “SRILM - The SRI Language Modeling Toolkit”, http://www.speech.sri.com/project/srilm/

[Thede *et al.* 1999] S. M. Thede and M. P. Harper. “A second-order Hidden Markov Model for part-of-speech tagging”, ACL1999.

[V. Tam *et al.* 2002] V. Tam, A. Santoso and R. Setiono, “A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization.”, In Proceedings of the 16th International Conference on Pattern Recognition, ICPR 2002, Vol. 4 Quebec, Canada, August 2002, pages 235-238.

[Wang *et al.* 2004] Gang Wang, Frederick H. Lochovsky, Qiang Yang, “Feature Selection with Conditional Mutual Information MaxiMin in Text Categorization”, CIKM’04, November 8–13, 2004.

[Yang *et al.* 1997] Y. Yang, and Pedersen J. O., “A comparative study on feature selection in text categorization”, Proceedings of the 14th International Conference on Machine Learning ICML97, pages 412-420, 1997.

[Yang *et al.* 1999] Yiming Yang and Xin Liu, “A Re-Examination of Text Categorization Methods,” Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1999, Pages 42 – 49.