3. 實驗設定

3.1 實驗語料

3.1.1 摘要測試語料

論文中分別採用二套測試語料集，進行摘要評估實驗，第一套(DataSet1)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，相關的統計資料如表 3.1 所示：

表 3.1 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

共 200 則廣播新聞，分為自動轉寫(Automatic Transcription)與人工轉寫(Human Transcription)兩種資料集。自動轉寫部分為經師大資工所大詞彙語音辨識器[Chen et al. 2004; Chen et al. 2005]辨識後，再經由自動的斷句處理的結果，辨識字正確率達 85.83%；人工轉寫部分是經由人工處理轉譯成文字的內容。此測試集的自動摘要評估標準答案部分，由三位國立台灣大學文學院大三以上的學生，分別對此 200 則廣播新聞的人工轉寫文件產生人工標註摘要(參照摘要)，摘要的結果可分為依文句重要性排名的句排名形式，與依特定比例重寫的摘要兩種，如下面圖 3.1 所示。此套測試集文件內容與文句個數少，且人工轉寫文件的斷句方

式大多以主播的口語停頓為斷點，如圖 3.2 所示，其文句邊界的定義與以口語停頓時間作為自動斷句依據的語音文件相似；但是，文件中可能會產生部分文句內容不完整的情形，例如“前教育部長楊朝祥”為一文句，若此文句被選出作為摘要，可能會讓瀏覽者無法了解所要表達的內容、主題；又如“限制出境”文句，文句中沒有描述確切主詞，沒有完整的文句語義。

編號：[1]N200108011200-01

請將下列新聞中的每一句依重要性排名，1-最重要、2-次重要、依此類推（排名請用阿拉伯數字）。

（1）桃芝颱風重創花蓮

（2）光復鄉大興村死傷慘重

（4）感觸最多的是在當地送信長達十七年的郵差鄭順發

（7）村子裡頭平常天天見面打招呼的老朋友

（3）一夕之間天人永隔

（5）災後頭一天送信到大興村

（6）鄭順發的心情已經不是複雜兩個字能夠形容

請為本則新聞重寫一個約 34 字左右的摘要：

桃芝風災造成光復鄉大興村的嚴重死傷　老郵差災後送信時感觸良多

圖 3.1 News 98 測試集人工標註摘要範列

編號：[6]N200108011200-10

景文技術學院校產被掏空案

前教育部長楊朝祥

因為涉嫌在教育部任職期間

多次包庇圖利景文集團董事長張萬利

被檢調單位約談並漏夜偵訊近二十個鐘頭後

今天清晨五點

檢察官以涉及弊端罪嫌重大

預定楊朝祥以兩百萬元交保候傳

限制出境

圖 3.2 News 98 測試集中文件內容範例

另外，我們蒐集了第二套摘要測試語料進行摘要結果的評估，使用MATBN(DataSet2)公視新聞語料[Wang et al. 2005]。MATBN 為中央研究院資訊所口語小組[SLG]，耗時三年與公共電視台[PTS]合作錄製完成，每天一個小時的公視晚間新聞深度報導，共收錄了 200 天(約 200 小時)的新聞語料，其中包含 2001 年的新聞 30 小時、2002 年 146 小時及 2003 年 24 小時，所有的新聞語料都有正確的人工轉寫文件。我們由 2001 年 11 月至 2002 年 8 月的新聞中選取其中 205 篇作為自動摘要的測試語料。主要挑選原則為：

1. 文章長度約為字數 500~1000 左右：為了針對不同文件特性進行實驗的比較與分析，我們希望所蒐集的第二套語料能有別於第一套語料的文件內容。因此，第二套語料偏向於選取文章內容較多的文件，以利於與第一套測試語料的結果相互對照，分析不同摘要方法於不同文章長度上的結果。

2. 新聞內容主要以單一語言(國語)為主：我們採用的大詞彙語音辨識器，以單一語言(國語)為主，為了避免其他語言的新聞內容，造成過多的辨識錯誤內容影響摘要結果的分析，我們盡量避免非國語的新聞內容。

3. 採用多元化的新聞內容：為了客觀地比較各種摘要方法的評估結果，我們希望實驗語料能包含不同主題的新聞內容，例如：國際、社會、政治、體育、生活等。

我們根據上述 1~3 點原則來選取適當的新聞文件作為我們的測試語料。第二套測試集(DataSet2)的相關統計資料，如表 3.2 所示：

表 3.2 MATBN 205 則新聞語料相關統計

新聞時間	2001 年 11 月~2002 年 8 月
新聞數	205 則
總長度	7.54 小時
平均每則新聞長度	132.38 秒
總大小(人工轉寫)	約 124 仟字
平均每則新聞大小(人工轉寫)	約 604 字

每則新聞內容主要分為主播、記者、外場訪問三大部分，我們同樣採用師大資工所大詞彙語音辨識器[Chen et al. 2004, 2005]進行辨識。205 則新聞的平均辨識字正確率以及個別的主播、記者與外場訪問三部分的平均辨識字正確率如表 3.3 所示。

表 3.3 MATBN 205 則新聞語料辨識正確率(%)

1.整篇(內外場)	2.主播(內場)	3.記者(外場)	4.受訪者(外場)
61.31	61.31	78.08	4.48

我們聘請三位政大新聞系大三以上同學(具新聞編輯經驗)，對 205 則新聞的人工轉寫文件進行文件斷句及參照摘要標註。首先，先由其中一人獨立完成 205 則文件斷句，之後再由三位評估者分別產生句排名形式與文件重寫摘要兩種結果，如圖 3.3 所示。其中，關於第二套測試語料中人工轉寫文件的斷句，我們僅提供文字內容進行斷句。因為我們不希望主播、記者或受訪者的播報方式及語氣，影響評估者對於文句界界的定義及重要文句的認定；同時，對於文句邊界的定義以呈現文句內容的完整性為主，希望每一文句內容能夠包含一個完整的內容或語意、概念。當然，這樣的斷句方式將會與採用自動斷句方式的語音文件中文句邊界定義差異甚大，也可能因此降低語音文件的摘要正確率。

[編號]：PTSND20011107_1

請將下列新聞中的每一句依重要性排名，1-最重要、2-次重要、依此類推（排名請用阿拉伯數字）。

23 歡迎回到新聞現場關心國際焦點

20 以巴持續傳出流血衝突

1 不過以色列星期二展開另一波撤軍行動退出先前占領的巴勒斯坦政經中心拉瑪拉

2 而根據正在歐洲的以色列外長表示他和以色列總理正在著手新的中東和平計畫

10 以色列星期二開始從巴勒斯坦自治城市拉瑪拉撤軍

6 不過在同一天約旦河西岸還是有零星衝突發生造成六個人死亡

3 自從三個星期前以色列觀光部長被人暗殺以後以色列軍隊就陸續占領六個巴勒斯坦城鎮

4 過去幾個星期在美國的壓力之下以色列已經陸續撤出伯力恆貝特假拉和卡其里雅三個城鎮

5 而在昨天從耶路撒冷北邊的拉瑪拉撤軍後以色列國防部長表示他們還會在拉瑪拉周圍維持一個安全路障

21 拉瑪拉目前是巴勒斯坦當局的重要政治和經濟中心

9 在以色列軍隊和坦克撤離的同時一些巴勒斯坦人也在當地進行示威抗議

7 在以色列結束對拉瑪拉長達兩個星期的占領後目前只剩下節寧和土卡姆仍為以色列軍隊所占領

8 在星期二稍早時西岸的城市節寧有兩名巴勒斯坦激進份子所搭乘的車輛遭到襲擊造成兩人死亡

22 這兩個當中的一人一直受到以色列的通緝

16 另外在靠近那布魯斯附近有三名巴勒斯坦槍手和一名以色列士兵在雙方交戰時喪生

11 以色列和巴勒斯坦新一波的緊張衝突是由於以色列內閣部長在十月十七號遭到暗殺所引發

12 但由於美國擔心中東地區的不穩定會影響到他的反恐聯盟因而向以巴雙方施壓要求停止暴力衝突的發生

13 前往比利時布魯塞爾參加歐洲與地中海國家外長會議的以色列外長裴瑞斯表示他已和總理夏隆建構新的中東和平計畫

17 希望藉此終結雙方一年多來造成超過九百人喪生的流血衝突

14 根據以色列媒體的報導裴瑞斯所提議的新和平方案

15 包括一項停火協議以及以聯合國的議案為基礎和巴勒斯坦人展開談判

18 在聯合國的決議案當中要求以色列撤出一九七六年以阿戰爭中所占領的土地

19 同時建立一個巴勒斯坦國

24 公視新聞陳秋玫編譯

請為本則新聞重寫一個約 180 字左右的摘要：

由於以色列內閣部長在十月十七號遭到暗殺，引發以色列軍隊就陸續占領六個巴勒斯坦城鎮。但美國擔心中東地區的不穩定會影響到他的反恐聯盟，因而向以巴雙方施壓要求停止暴力衝突的發生。以色列星期二開始從巴勒斯坦政經中心拉瑪拉撤軍。正在歐洲的以色列外長裴瑞斯表示，他和以色列總理夏隆建構新的中東和平計畫，包括一項停火協議，及以聯合國議案為基礎，盼藉此終結雙方一年多來的流血衝突。

圖 3.3 MATBN 測試集人工標註摘要範列

3.1.2 訓練語料

本論文中用來訓練機率式摘要模型(例如，隱藏式馬可夫模型、主題混合模型及詞層次主題混合模型)，及建立關聯性模型所需要的同一時期或同一領域文字文件訓練語料庫，是由中央通訊社(Central News Agency, CNA)，從西元 1991 至 2002 年所發佈的文字新聞[LDC]中，分別選出西元 2001 年 8 月至西元 2002 年 10 月，所發佈且型態屬於故事(Type="story")的文字新聞做為文字文件訓練語料庫。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。

語言模型的訓練文字語料是由中央通訊社 2000 與 2001 年所收集而來的文字新聞，約含一萬四千個文字檔案。在本論文中的語言模型，使用 Katz 語言模型平滑技術[Katz 1987]，以 SRI Language Modeling Toolkit[SRI]來做訓練。

聲學模型的訓練語料是由 2001 及 2002 年的 MATBN 新聞語料中篩選出來，其中包含 25.5 小時的訓練集(5,774 句)供聲學模型訓練之用；1.5 小時的評估集(292 句)供辨識評估之用。

3.1.3 台師大大詞彙連續語音辨識系統

台師大大詞彙連續語音辨識系統，分為前端處理(Front-end Processing)、聲學模型(Acoustic Models)、詞典建立(Lexicon Construction)、語言模型(Language Model)以及詞彙樹複製搜尋(Tree-copy Search)等部份。

前端處理：

主要是擷取語音訊號中重要的參數，作為語音辨識所需之聲學特徵。此系統支援梅爾倒頻譜係數(Mel-frequency Cepstral Coefficients, MFCC)以及異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)加上最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)[Gopinath 1998; Saon et al.

2000]兩個不同的語音特徵參數。本論文使用異質性線性鑑別分析配合最大相似度線性轉換做為語音特徵參數。

聲學模型：

使用112個聲母、38個韻母及1個靜音(Silence)，共151個連續密度隱藏式馬可夫模型(Continuous Density Hidden Markov Models, CDHMMs)。每個模型包含有3至6不等的狀態(States)個數，每個狀態皆為高斯混合分佈(Gaussian Mixture Models)，其中每個高斯混合分佈的個數分別為1至128個不等。此外，聲母和韻母共組合成403個不同的基本音(Base Syllables or Toneless Syllables)。

詞典建立：

詞典是先將大量的文字語料經由一個含有一至四字詞約六萬八千個詞的詞典來斷詞，配合字詞在語料中的統計特性，以自動化的方式產生約五千個二至十字詞的複合詞(Compound Words)。最後，語音辨識詞典約七萬二千個一至十字詞。

語言模型

系統中使用雙連詞(Bigram)及三連詞(Trigram)語言模型，是從中央通訊社在2000與2001年間所收集，約一億七千萬個中文字語料作為語言模型的訓練語料。

詞彙樹複製搜尋：

系統是採用由左至右(Left-to-right)且音框同步(Frame Synchronous)的詞彙樹複製搜尋方式[Aubert 2002]，找出可能的辨識結果候選詞建立詞圖。然後，進一步作詞圖搜尋[Ortmanns 1997]，找出最佳的詞序列作為語音辨識結果。系統於詞彙樹複製搜尋階段是採用雙連詞語言模型，詞圖搜尋階段則是使用三連詞語言模型。

3.2 基礎實驗

在本論文中，吾人首先實作數個目前已被使用於文件摘要的方法作為基礎實驗，用來與本論文所提之摘要方法的結果作比較與驗證。基礎實驗中的摘要方法有：向量空間模型(VSM)[Ho 2003]、最大臨界相關(MMR) [Murray et al. 2005]、潛藏語意分析模型(LSA)[Gong et al. 2001]、嵌入式潛藏語意分析模型(eLSA)[黃耀民

2005; 陳怡婷 et al. 2005]、以奇異值分解為基礎之維度化簡法(DIM)[Hirohata et al. 2005]、隱藏式馬可夫模型(HMM)[黃耀民 2005; 陳怡婷 et al. 2005]、主題混合模型(TMM)[Chen et al. 2006; 黃耀民 2005; 陳怡婷 et al. 2005]、重要性分數(Sig) [Hirohata et al. 2005]以及亂數方式抽取重要文句(Random)等。其中，重要性分數方法所採用的計算方式為：

$$Sig(S_i) = \sum_{r=1}^{N_i} [\beta_1 \cdot I(w_r) + \beta_2 \cdot L(w_r)] \quad (3\text{-}1)$$

$I(w_r)$為詞w_r的重要性分數；$L(w_r)$為詞w_r的語言學分數；N_i為文句S_i的長度，即詞的總個數。本論文使用二連語言模型分數。β_1與β_2分別代表重要性分數與語言學分數的權重值。其他基礎實驗的計算方式可參照第二章所述。

我們分別於二套測試語料進行基礎實驗分析，觀察基礎實驗中不同摘要方法於二套測試語料中的摘要結果。對於二套測試語料，我們都將其中的一百篇文件作為發展集(Development Set)，其餘的文件作為測試集(Test Set)。發展集用於調整參數使用，而測試集則使用發展集所找到之最佳參數設定，進行摘要結果之分析。由於某些摘要方法皆有權重參數或是模型參數設定的問題，需要作參數的調整與最佳化，如式(3-1)中的參數β_1與β_2，所以我們先透過發展集來找出每一摘要方法的最佳參數設定，然後於測試集中進行摘要結果的評估與比較。在測試語料DataSet1中，我們以前一百篇文件作為發展集，後一百篇文件為測試集；於測試語料DataSet2中，我們同樣以前一百篇文件作為發展集，而後一百零五篇文件作為測試集。

本論文所採用的評估方法為餘弦評估及Rouge-2評估，Rouge-2評估方法為目前最常用的評估方法，許多摘要研究文獻中均採用此評估方法，關於評估方法詳細的介紹與計算方式，可參考本論文第二章2.3.2小節。其中，餘弦評估方法中所使用的詞頻反文件頻(TF-IDF)，是採用CNA從西元2001年8月至12月所發佈且型態屬於故事(Type="story")的文字新聞文件集中統計而來；同時，餘弦評估是採用三份參照摘要的句排名結果與三份重寫結果進行評估；而Rouge-2評估僅使用參

照句排名形式的摘要結果進行評估。下面將呈現在不同的測試語料及不同的評估方式下的基礎實驗結果。

表 3.4 基礎實驗於 DataSet1 發展集之摘要結果 (餘弦評估)

TD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM	Random
10%	0.4154	0.3726	0.4070	0.3785	0.4196	0.3911	0.4203	0.4263	0.1923
20%	0.4610	0.4061	0.4600	0.4334	0.4584	0.4451	0.4657	0.4707	0.2407
30%	0.5133	0.4757	0.5168	0.4971	0.5282	0.5155	0.5178	0.5293	0.3303
50%	0.6220	0.5923	0.6259	0.6073	0.6271	0.6125	0.6271	0.6282	0.5013
70%	0.6914	0.6791	0.6953	0.6795	0.6950	0.6851	0.6996	0.7029	0.6142
SD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM	Random
10%	0.3344	0.3089	0.3418	0.3267	0.3344	0.3311	0.3415	0.3421	0.1900
20%	0.3725	0.3294	0.3741	0.3685	0.3726	0.3699	0.3789	0.3796	0.2317
30%	0.4194	0.3695	0.4261	0.4150	0.4186	0.4171	0.4281	0.4258	0.2793
50%	0.5136	0.4814	0.5106	0.5113	0.5139	0.5332	0.5159	0.5164	0.4193
70%	0.5836	0.5574	0.5831	0.5756	0.5825	0.5832	0.5847	0.5868	0.5163

3.2.1 DataSet1 基礎實驗

表 3.4 及表 3.5 分別為發展集中人工轉寫文件(TD)與自動轉寫文件(SD)，在餘弦評估與 Rouge-2 評估方法下的摘要結果；表中每一個數據代表不同摘要方法，在不同摘要比例下的摘要正確率。我們透過發展集來調整每一種摘要方法的參數設定，然後於測試文件集使用同樣的參數設定，觀察不同評估方法下每一種摘要方法的摘要結果。表 3.4 及表 3.5 中基礎實驗設定如下：

- 以奇異值分解為基礎之維度化簡法(DIM)：如式(2-22)所示，對於維度 K 值的設定，我們嘗試了 1、3、5、7、9、10 等參數值。發現二種評估方法及不同形式的文件上，將 K 值設定為 1 有最好的結果。[Hirohata et al.

表 3.5 基礎實驗於 DataSet1 發展集之摘要結果 (Rouge-2 評估)

TD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM	Random
10%	0.3888	0.3532	0.3868	0.3571	0.3944	0.3393	0.4157	0.4242	0.0931
20%	0.4310	0.3793	0.4396	0.4179	0.4440	0.3986	0.4591	0.4683	0.1329
30%	0.4708	0.4338	0.4978	0.4666	0.5112	0.4897	0.4947	0.5090	0.2237
50%	0.6104	0.5799	0.6265	0.6024	0.6364	0.6219	0.6527	0.6529	0.4183
70%	0.7369	0.7120	0.7458	0.7283	0.7527	0.7570	0.7845	0.7828	0.5882
SD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM	Random
10%	0.2653	0.2786	0.3043	0.2935	0.2659	0.2972	0.3084	0.3154	0.1215
20%	0.3103	0.2904	0.3325	0.3293	0.3025	0.3344	0.3467	0.3496	0.1470
30%	0.3331	0.2984	0.3573	0.3490	0.3503	0.3597	0.3734	0.3600	0.1702
50%	0.4436	0.4072	0.4486	0.4488	0.4549	0.4870	0.4768	0.4657	0.3177
70%	0.5413	0.5019	0.5444	0.5324	0.5457	0.5632	0.5631	0.5526	0.4382

2006]中亦表示當 K=1 時有最佳的結果。因此，於表 3.4 及表 3.5 中 K 值設定為 1，測試集亦採用相同設定。

- 最大臨界相關(MMR)：如式(2-8)所示，對於權重參數 a，我們嘗試 0.1、0.2、…、0.9 等數值設定，發現在餘弦評估方法下，人工轉寫文件設定為 0.6，及自動轉寫文件設定為 0.9 時，有最好的結果；於 Rouge-2 評估下，人工轉寫文件及自動轉寫文件設定為 0.6 時，會有最佳的結果。此參數設定結果呈現於表 3.4 及表 3.5 中，測試集(如表 3.6、表 3.7)亦採用相同的設定。
- 重要性分數法(Sig)：如式(3-1)所示，實驗中 β_1 與 β_2 值皆設定 0.5。另外，我們嘗試以”詞頻-反文件頻(TF-IDF)”或以式(2-12)的計算方式，來計算詞重要性分數 $I(w_r)$。同時亦分析文句重要性分數(如式(3-1))經由文句長度正規化後的結果。摘要結果顯示，以式(2-12)的方式來計算詞重要性分數 $I(w_r)$，同時不作正規化處理有最佳結果。發展集與測試集均採

用相同設定。

- 隱藏式馬可夫模型(HMM)：如式(2-23)所示。其中λ的設定，我們嘗試0.1、0.2…、0.9，及0.01、0.02、…、0.19等不同的設定，發現當λ=0.05時有最好的結果。測試集亦使用相同設定。
- 主題混合模型(TMM)：如式(2-28)所示。在以主題混合模型進行文件摘要時，我們同時將文件中索引在每一文句S_i中的機率分佈考慮進來，將式(2-28)進一步表示成：

$$P(D|S_i)=\prod_{w\in D}\left[\alpha\cdot\sum_{k=1}^{K}P(w|T_k)P(T_k|S_i)+(1-\alpha)\cdot P(w|S_i)\right]^{n(w,D)} \quad (3\text{-}2)$$

其中α是比重參數。在實驗中，我們嘗試不同α值設定以及不同主題數的摘要結果；以下實驗結果，於餘弦評估方法下人工轉寫文件的α值設定為0.96，主題數K=32，自動轉寫文件的α值為0.98，主題數K=64；於Rouge-2評估方法下，人工轉寫文件α值設為0.95，主題數K=32，自動轉寫文件的α值為0.98，主題數K=16。測試集亦使用相同設定。

首先，除了Random的結果外(由於Random的方式結果不佳，因此我們不作討論與分析)，我們可以將基礎實驗結果分成二部分來看，第一部分為VSM、MMR、LSA、eLSA、DIM及Sig等非機率生成式模型的摘要方法，第二部分為HMM與TMM等機率生成式模型的摘要方法。非機率生成式模型的摘要方法中，在二種評估方法的結果下，我們可以看出MMR於人工轉寫文件有最好的結果；而eLSA於自動轉寫文件的低摘要比例(如10%、20%)結果有較好的正確率，Sig於自動轉寫文件的30%、50%及70%等摘要比例結果有較高的正確率；比較非機率生成式摘要方法於人工轉寫文件與自動轉寫文件上的結果，可以發現在二種評估方法下，VSM、MMR於人工轉寫文件上有很好的摘要結果，eLSA於不同形式的文件集上均有不錯的結果，特別是於自動轉寫文件上的結果，Sig於自動轉寫文件上之摘要結果較於人工轉寫文件上結果有較好的表現。機率生成式模

型的摘要方法中，在二種評估方法下 TMM 於人工轉寫文件的摘要結果均較 HMM 來得好(除摘要比例 70%時 Rouge-2 評估的結果)，於自動轉寫文件的低摘要比例結果，如 10、20%的摘要比例，有較佳的結果。接著，我們比較非機率生成式模型摘要方法與機率生成式模型摘要方法的結果，我們可以觀察出不論是在何種評估方法下，機率生成式模型 TMM 於人工轉寫文件上的摘要正確率較其他方法來得高，而於自動轉寫文件結果中，於低摘要比例 10%、20%時 TMM 亦有最高的摘要結果。

我們進一步分析二種評估方法下的基礎實驗結果。經由觀察人工轉寫文件的餘弦評估結果及 Rouge-2 評估結果，可以看出：雖然機率生成式模型摘要方法在 Rouge-2 評估下正確率呈現小幅度的下降，但是非機率生成式模型的摘要方法(如 VSM)下降的幅度更大；而且當摘要比例為 50%時，HMM 及 TMM 於 Rouge-2 評估下的摘要結果相較於餘弦評估結果有明顯提昇時，VSM、LSM、eLSA 的 Rouge-2 評估結果是往下掉的。觀察自動轉寫文件的餘弦評估結果及 Rouge-2 評估結果，同樣可以看出在二種評估方法下，VSM、MMR 相較於其他方法的結果差異甚大。當然，若分析二種方法的計算方式與原理，Rouge-2 的評估方式或許較餘弦評估來得嚴格一些，因為 Rouge-2 是採用詞雙連(Word Bigram)為評估單位元。但是，機率生成式模型的摘要結果在二種評估方法下，並不如 VSM、MMR、eLSA 有那麼大的差異。因此，我們認為其中一個原因可能是因為 VSM、MMR、eLSA 與餘弦評估均是採用相同的計算方式，因此三種摘要方法於餘弦評估下會有較高的正確率，而於 Rouge-2 評估方式的摘要結果 VSM、MMR、eLSA 等方法才會有較大幅度的下降；另一個原因可能是機率生成式模型的確是一個好的摘要方法，因此，其摘要結果不論於任一種評估方式下都可以得到不錯且一致性的結果。

表 3.6 及表 3.7 為測試集中人工轉寫文件(TD)與自動轉寫文件(SD)，分別在餘弦評估與 Rouge-2 評估方法下的摘要結果；表中每一個數據代表不同摘要方法

表 3.6 基礎實驗於 DataSet1 測試集之摘要結果 (餘弦評估)

TD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM	Random
10%	0.3839	0.3807	0.3895	0.3892	0.3841	0.3679	0.3841	0.3972	0.1896
20%	0.4015	0.3880	0.4113	0.4022	0.4111	0.3897	0.4064	0.4168	0.2173
30%	0.4656	0.4278	0.4813	0.4640	0.4731	0.4574	0.4855	0.4900	0.3203
50%	0.6097	0.5643	0.6218	0.5795	0.6190	0.6011	0.6037	0.6013	0.4917
70%	0.6783	0.6526	0.6864	0.6607	0.6913	0.6633	0.6820	0.6803	0.5886
SD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM	Random
10%	0.3324	0.3194	0.3389	0.3306	0.3338	0.3302	0.3155	0.3171	0.1871
20%	0.3513	0.3277	0.3511	0.3436	0.3546	0.3487	0.3381	0.3438	0.2089
30%	0.4078	0.3770	0.4068	0.3872	0.4117	0.3835	0.3993	0.4058	0.2546
50%	0.4943	0.4620	0.5038	0.4850	0.5039	0.4963	0.5027	0.5047	0.3996
70%	0.5665	0.5349	0.5666	0.5422	0.5603	0.5564	0.5601	0.5627	0.4903

表 3.7 基礎實驗於 DataSet1 測試集之摘要結果 (Rouge-2 評估)

TD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM	Random
10%	0.3400	0.3661	0.3636	0.3794	0.34004	0.3467	0.3714	0.3871	0.1314
20%	0.3432	0.3555	0.3765	0.3767	0.36393	0.3634	0.3892	0.3953	0.1548
30%	0.4123	0.3714	0.4506	0.4141	0.4393	0.4381	0.4669	0.4659	0.2543
50%	0.6064	0.5476	0.6199	0.5623	0.62277	0.6345	0.6308	0.6280	0.4353
70%	0.7257	0.6936	0.7389	0.7216	0.75508	0.7407	0.7660	0.7636	0.5782
SD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM	Random
10%	0.3073	0.3034	0.3225	0.3187	0.3073	0.3144	0.2932	0.3210	0.1204
20%	0.3188	0.2926	0.3186	0.3148	0.3214	0.3259	0.3191	0.3333	0.1392
30%	0.3593	0.3286	0.3637	0.3383	0.3678	0.3428	0.3705	0.3741	0.1679
50%	0.4485	0.3906	0.4547	0.4345	0.4501	0.4666	0.4732	0.4605	0.3163
70%	0.5425	0.4843	0.5459	0.5259	0.5366	0.5538	0.5595	0.5522	0.4343

在不同摘要比例下的摘要正確率。測試集中每一種摘要方法的參數設定與發展集相同。在餘弦評估結果下，由人工轉寫文件與自動轉寫文件的摘要結果可看出：在非機率生成式模型的摘要方法中，eLSA 於摘要比例 10%、20%、30%、50%

時較其他方法有最高的正確率；在機率生成式模型的摘要方法 HMM、TMM 中，TMM 有較好的摘要結果；比較非機率生成式摘要方法與機率生成式方法，TMM 於人工轉寫文件低摘要比例 10%、20%、30%時，均較其他摘要方法來得好，而於自動轉寫文件摘要結果中，eLSA 有最佳的摘要正確率。在 Rouge-2 評估方式下的摘要結果，發現非機率生成式模型的摘要方法中 DIM 於人工轉寫文件低摘要比例 10、20%下有最好的結果，eLSA 於自動轉寫文件低摘要比例 10%、20%、30%時有最佳的結果，而 VSM、MMR 等摘要方法的正確率一樣呈現出大幅下降的趨勢；機率生成式模型的摘要方法中，於人工轉寫文件中 TMM 有較好的摘要結果，並且較其他非機率生成式摘要方法好，於自動轉寫文件中 TMM 在摘要比例 20%、30%時，相較於其他摘要方法有最高的正確率，HMM 於摘要比例 50%、70%有最高的正確率。

從 DataSet1 基礎實驗的發展集結果中，當各種摘要方法的參數均為最佳化設定時，機率生成式摘要模型不論於人工轉寫文件或自動轉寫文件上皆有較佳的摘要結果，同時於二種評估方式下結果呈現一致性且較高的正確率，在測試集人工轉寫文件上亦呈現相同的結果。雖然，機率生成式摘要模型於測試集自動轉寫文件的餘弦評估結果下，HMM、TMM 沒有較非機率生成式模型摘要方法好，但是於 Rouge-2 評估結果 TMM 仍然有很好的結果，特別是在低摘要比例 20%、30%下的摘要結果。我們嘗試找出機率生成式摘要模型於測試集自動轉寫文件上結果不佳的原因，發現 HMM 及 TMM 的摘要正確率於其他參數設定值下亦有很好的結果，甚至較其他摘要方法結果好，這可能是因發展集與測試集不匹配(Missmatch)所造成，同時每一文件中的每一文句模型的分佈均不相同，若將所有文件中的每一文句模型參數設為相同，亦可能無法達到很好的結果。因此，對於機率生成式摘要模型參數設定的方式，仍有很大進步與討論空間。

3.2.2 DataSet2 基礎實驗

除了 Random 的摘要方式外，我們亦將相同的基礎實驗方法，實作於測試語料 DataSet2，觀察於文件內容較多且文句語意較為完整的測式語料 DataSet2 上，不同方法的摘要結果。同樣地，我們先透過發展集來調整每一種摘要方法的參數設定，然後於測試文件集使用同樣的參數設定，觀察不同評估方法下每一種摘要方法的摘要結果：

- 以奇異值分解為基礎之維度化簡法(DIM)：如式(2-22)所示。當 K 值設定為 1 時，於發展集中有最好的結果。
- 最大臨界相關(MMR)：如式(2-8)所示。經由實驗觀察，在二種評估方法下人工轉寫文件及自動轉寫文件之權重參數 a 值均設定為 0.9，會有最好的結果。
- 重要性分數法(Sig)：如式(3-1)所示，β_1 與 β_2 皆設定為 0.5。另外，我們嘗試以詞頻反文件頻(TF-IDF)或式(2-12)的計算方式，來計算詞重要性分數 $I(w_r)$；實驗結果發現採用詞頻反文件頻作為詞重要性分數，同時不作正規化處理，會有最佳的結果。
- 隱藏式馬可夫模型(HMM)：如式(2-23)所示。當 λ 設定為 0.01 時，有最好的結果。
- 主題混合模型(TMM)：實作上與 DataSet1 相同，在以主題混合模型進行文件摘要時，我們同時考慮文件中索引在每一文句 S_i 中的機率分佈，以式(3-2)進行文件摘要的處理。其中，於不同評估方法下人工轉寫文件 α 值設定為 0.99，主題數 K =64；於餘弦評估方法下自動轉寫文件的 α 值設定為 0.98，主題數 K =16；於 Rouge-2 評估下自動轉寫文件的 α 值設定為 0.99，主題數 K =32。

表 3.8 及表 3.9 分別為發展集中人工轉寫文件(TD)與自動轉寫文件(SD)，於餘弦評估與 Rouge-2 評估方法下的摘要結果；

表 3.8 基礎實驗於 DataSet2 發展集之摘要結果 (餘弦評估)

TD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM
10%	0.4492	0.3586	0.4347	0.3878	0.4460	0.4344	0.4427	0.4530
20%	0.5868	0.4893	0.5750	0.5441	0.5854	0.5704	0.5740	0.5891
30%	0.6437	0.5673	0.6353	0.6244	0.6538	0.6320	0.6346	0.6470
50%	0.7413	0.6884	0.7446	0.7307	0.7419	0.7301	0.7376	0.7432
70%	0.7994	0.7697	0.7989	0.7972	0.8016	0.7825	0.7985	0.8023
SD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM
10%	0.4250	0.3161	0.4066	0.3783	0.4229	0.4233	0.4128	0.4248
20%	0.5265	0.4156	0.4993	0.4965	0.5253	0.5167	0.5156	0.5130
30%	0.5671	0.4616	0.5502	0.5461	0.5646	0.5631	0.5623	0.5611
50%	0.6218	0.5437	0.6080	0.6191	0.6170	0.6044	0.6174	0.6212
70%	0.6325	0.5978	0.6276	0.6392	0.6279	0.6185	0.6339	0.6357

表 3.9 基礎實驗於 DataSet2 發展集之摘要結果 (Rouge-2 評估)

TD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM
10%	0.2457	0.2068	0.2927	0.2286	0.2385	0.2434	0.2839	0.3028
20%	0.3920	0.3458	0.4151	0.3774	0.3928	0.3508	0.3995	0.4138
30%	0.4783	0.4216	0.4986	0.4791	0.4933	0.4168	0.4790	0.4875
50%	0.6490	0.6135	0.6756	0.6491	0.6540	0.5771	0.6539	0.6510
70%	0.7877	0.7578	0.8030	0.7874	0.7950	0.6985	0.8019	0.7927
SD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM
10%	0.1947	0.1485	0.2118	0.1966	0.1894	0.1890	0.2141	0.2207
20%	0.2707	0.2112	0.2611	0.2512	0.2707	0.2470	0.2747	0.2788
30%	0.3107	0.2486	0.3033	0.2974	0.3123	0.2860	0.3156	0.3243
50%	0.3970	0.3294	0.3867	0.3951	0.3925	0.3453	0.3959	0.3999
70%	0.4284	0.3921	0.4232	0.4348	0.4232	0.3771	0.4315	0.4348

首先觀察基礎實驗中 VSM、MMR、LSA、eLSA、DIM 及 Sig 等非機率生成式模型的摘要方法之結果。在餘弦評估方法下，VSM 於人工轉寫文件與自動轉寫

文件之低摘要比例 10%、20%時，有最高的摘要正確率，MMR 於摘要比例 30%亦有不錯的結果。在 Rouge-2 評估方法下，eLSA 於人工轉寫文件上有最佳的摘要結果，於自動轉寫文件摘要比例 10%時亦較其他非機率生成式模型方法好，VSM、MMR 於自動轉寫文件摘要比例 20%、30%及 50%時有較高的正確率。觀察機率生成式模型 HMM 與 TMM 的摘要結果，在二種評估方法及不同形式文件的摘要結果中，TMM 於低摘要比例時皆較 HMM 正確率高，同時於大部分的摘要比例下亦較 HMM 結果好。比較非機率生成式模型摘要方法與機率生成式模型摘要方法的摘要結果，機率生成式模型 TMM 無論於人工轉寫文件或自動轉寫文件上，在低摘要比例 10%時二種評估方法下的摘要正確率皆較其他方法好，而且於 Roug-2 評估方法下不同摘要比例的自動轉寫文件結果亦較其他方法好。

表 3.10 及表 3.11 為測試集中人工轉寫文件與自動轉寫文件的摘要結果。於餘弦評估結果下，非機率生成式模型摘要方法中 MMR 於人工轉寫文件及自動轉寫文件的低摘要比例 10%、20%有最好的摘要結果，VSM 於二種形式文件的高

表 3.10 基礎實驗於 DataSet2 測試集之摘要結果 (餘弦評估)

TD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM
10%	0.4916	0.3701	0.4885	0.3930	0.4962	0.4570	0.4835	0.4891
20%	0.6119	0.4821	0.6065	0.5480	0.6201	0.6075	0.6166	0.6233
30%	0.6641	0.5579	0.6545	0.6179	0.6643	0.6541	0.6579	0.6603
50%	0.7382	0.6856	0.7359	0.7198	0.7362	0.7341	0.7338	0.7380
70%	0.7883	0.7620	0.7886	0.7844	0.7877	0.7827	0.7873	0.7914
SD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM
10%	0.4018	0.2936	0.3667	0.3309	0.4034	0.3915	0.3837	0.3871
20%	0.4786	0.3666	0.4624	0.4372	0.4796	0.4691	0.4739	0.4754
30%	0.5191	0.4193	0.5045	0.4933	0.5150	0.5100	0.5110	0.5127
50%	0.5714	0.4982	0.5571	0.5585	0.5640	0.5617	0.5619	0.5647
70%	0.5843	0.5534	0.5774	0.5900	0.5780	0.5763	0.5837	0.5842
TD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM

表 3.11 基礎實驗於 DataSet2 測試集之摘要結果 (Rouge-2 評估)

TD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM
10%	0.2864	0.2131	0.3534	0.2321	0.2915	0.2483	0.3280	0.3350
20%	0.4272	0.3253	0.4646	0.3873	0.4326	0.4081	0.4499	0.4528
30%	0.4921	0.4175	0.5156	0.4778	0.4919	0.4501	0.5012	0.4936
50%	0.6464	0.6214	0.6720	0.6439	0.6448	0.5898	0.6534	0.6536
70%	0.7856	0.7563	0.8028	0.7817	0.7917	0.7062	0.7915	0.7923
SD	VSM	LSA	eLSA	DIM	MMR	Sig	HMM	TMM
10%	0.2044	0.1442	0.1866	0.1622	0.2037	0.1790	0.2008	0.2110
20%	0.2385	0.1832	0.2398	0.2181	0.2412	0.2129	0.2501	0.2618
30%	0.2818	0.2244	0.2762	0.2666	0.2801	0.2475	0.2820	0.2861
50%	0.3660	0.3037	0.3519	0.3509	0.3590	0.3102	0.3622	0.3659
70%	0.4005	0.3683	0.3907	0.4029	0.3937	0.3474	0.4065	0.4093

摘要比例 50%、70%下有最好的摘要結果；與機率生成式模型 HMM、TMM 結果比較，VSM 與 MMR 有較佳的結果。在 Rouge-2 評估方法下，eLSA 於人工要寫文件上有最好的結果，同時亦較機率生成式模型摘要方法來得好；VSM 與 MMR 於自動轉寫文件上有不錯的結果，但是較 TMM 的結果差。由測試集中的結果可以發現，機率生成式摘要模型於餘弦評估結果下並不是最佳的，且於 Rouge-2 評估的人工轉寫文件結果亦較 eLSA 差。但是，其方法於自動轉寫文件 Rouge-2 的評估結果下是最佳的，同時比較表 3.8 至表 3.11 中自動轉寫文件的結果，TMM 於低摘要比例(如 10%)仍較其他方法來得好。

比較 DataSet1 與 DataSet2 二套測試語料的結果。在 DataSet1 的結果中機率生成式摘要模型 TMM 及 HMM，於人工轉寫文件上呈現不錯的摘要結果；同時於發展集自動轉寫文件上亦較其他摘要方法好。此外，eLSA、TMM 於測試集自動轉寫文件上同樣有好的結果。相較於 DataSet2 的結果，TMM 於人工轉寫文件上並沒有呈現出一致且較好的摘要結果，反而是 VSM、MMR 有不錯的結果；

同時，eLSA 於自動轉寫文件上的結果亦顯得較差。由 DataSet2 的結果觀察，不同摘要方法於文件內容較多的測試語料上，似乎 TMM 於自動轉寫文件低摘要比例的摘要結果上有較好的表現，尤其是僅從 Rouge-2 的結果來看，TMM 於自動轉寫文件摘要有最高的正確率。TMM 於 DataSet1 自動轉寫文件上亦很好的結果。

本論文中，我們希望所提出之摘要方法能夠進一步提昇摘要的正確率，特別是對於低摘要比例下的摘要正確率，因為文件摘要的目的即是精簡地表達文件內容的主題與重點，[Kupiec et al. 1995]亦提到 20%摘要比例的摘要內容或甚至更短的內容是有用及有幫助的摘要。