中文語音文件摘要使用主題混合模型

陳怡婷 游斯涵 李家豪 陳柏琳
台師大資工所 台師大資工所 台師大資工所 台師大資工所
g93470070@csie.ntnu.edu.tw u91316@ice.ntnu.edu.tw koichi700503@yahoo.com.tw berlin@csie.ntnu.edu.tw

摘要

本論文探討摘錄式中文語音文件摘要方法。摘錄式摘要是將文件摘要視為一個排序的問題，透過對文句、片語或詞的重要性或與文件的相關性作排名，再根據特定摘要比例依序選取重要的文句、片語或詞之組合。與其他作法不同的是，我們將被摘要文件中的每一個文句視為一個機率模型，透過機率生成方式來估測被摘要文件與其文句間的相關程度，並以此作為重要文句選取的依據。我們提出以詞層次主題混合模型來從事中文語音文件摘要。詞層次主題混合模型的概念與我們過去所提出的文句層次主題混合模型相似，是透過 K 個潛藏主題的混合模型來估測文件與每一文句的相關程度。但特別的是，我們係將文句中的每一個詞視為一個機率生成模型，每一個詞成為包含了 K 個潛藏主題的混合模型。如此，文句中每一個詞對於其文件中其它詞的關係即可經由 K 個潛藏主題來表示。我們在中文語音廣播新聞語料庫上實作了一系列的實驗，實驗結果顯示出我們所提出之詞層次混合模型的確優於其它摘要模型。

關鍵詞：文件摘要、摘錄式、語音文件、詞層次主題混合模型。

1. 前言

網際網路已成為現代人獲取資訊的主要來源，大量的資料傳遞與分享於全球各地，而資料內容的呈現方式也因影音多媒體技術的成熟與不斷地創新，逐漸以豐富生動的多媒體影音與動畫取代了靜態的文字內容。具時序性的多媒體影音內容，往往長達數分鐘或數小時。例如廣播(或電視)新聞、演講錄音、語音郵件等，不易瀏覽及查詢，使用者往往需要耐心地聽完或看完整筆檔案，才能了解內容所表達的主題或重點，這對於現今資訊爆增且講求效率的新時代而言是極不符合需求且不方便的。再者，由於同一時期的廣播(或電視)新聞內容重覆性高，往往不同的檔案卻包含著相似的內容，使用者常因此需要花大量的時間來過濾與判別多媒體影音資訊內容。

自動語音文件摘要的目的即在於解決上述種種的問題，針對這些大量的時序性多媒體影音文件或檔案以自動的方式進行重要資訊及主題的擷取，以摘要結果作為文件或檔案的呈現或是索引使用，讓使用者可以快速且方便地了解影音文件或檔案的要點與主題資訊。

本論文主要探討摘錄式(Extractive)中文語音文件摘要方法，摘錄式摘要是依據特定摘要比例從原始文件選取重要的文句、片語或詞來組成摘要。目前自動文件摘要的相關研究技術多以摘錄式摘要為主[1]。摘錄式摘要其實可視為一種詞、文句、或段落重要性的排序問題，當某個詞、文句、或段落對於整篇文件的重要性越高或是相關性越大，則優先被選取作為文件摘要的內容。目前一般常見的摘錄式自動摘要方法大致上可分為三大類，分別是(1)以文句結構或位置為基礎、(2)以索引特徵統計值為基礎、(3)以機率模型為基礎等三大類。

以文句結構或位置為基礎的摘要方法是依據文句內字、詞或文句本身位於的文章中的位置來決定其重要性[2][3]，位於重要段落的字、詞或文句給予較高的權重，例如第一段、最後一段或位於章節標題為「簡介、目的、結論」的文句可被視為重要文句並選取成為摘要。此種摘要方式簡單且直覺，但僅適用於具有特定章節結構或是遵循著某種編排方式的文件，而語音文件的自動轉寫(Automatic Transcripts)往往沒有很明顯的章節結構資訊，因此此類方法並不能完全適用。

而以索引特徵統計值為基礎的摘要方法可分為依文句相似度方式[4][5][6]、依潛藏語意概念表示方式[3][5][7]、依文句特徵值分數的方式[3][8]及依分類器分類[9][10][11]方式來從事摘錄式文件摘要。以索引特徵統計值為基礎的語音摘要方法通常是根據摘要索引特徵的統計值，如詞頻(Term Frequency)、詞重要性分數(Significance Score)、語言學分數(Linguistic Score)、辨識信心度(Recognition Confidence Measure)及聲韻特徵(Prosodic Features)，依不同的方式來計算文句的重要性[8]。例如可將被摘要文件中每一文句及文件本身均以一 L 維向量表示，向量的每一維度代表某個索引特徵(可以是詞、字或音節等單位)在文句或是文件中的統計值，以計算文句與文件的相似度或相關度[4][5][6]。每一文句分別依其與被摘要文件之向量表示式的相似度或相關度大小作排名，並依摘

要比例選取出重要之文句。此外，若我們希望選取出重要的摘要文句並可以概括整篇文件的不同主題性，則可在計算文句重要性時考慮文句內容或主題重覆性的問題[5][6]。而潛藏語意概念分析(Latent Semantic Analysis, LSA)摘錄模型是先將文件以“索引－文句”矩陣表示，然後透過奇異值分解(Singular Value Decomposition, SVD)將文件投射到一低維度的潛藏語意空間，並假設每一奇異值及其對應的奇異向量(Singular Vector)代表一潛藏主題或概念(奇異值越大越重要)，且文件中每一文句可由右奇異矩陣轉置的行向量表示。接著，依奇異值大至小，從所對應的右奇異向量(右奇異矩陣轉置的列向量)中選出有最大對投影量的文句作為文件的摘要[5]。近期也有許多基於潛藏語意概念模型的延伸研究與應用陸續被提出[6][7]。而以文句特徵值分數的摘錄方式亦可將被摘要文件中的每一文句視為一連串索引特徵(例如詞或音節等)表示[8]，並以文句中各索引特徵的統計值(如詞頻、反文件頻等統計資訊)、語言評估值(如對類專有名詞或是不同詞性的詞給予不同的分數)、或是其它聲韻特徵值經加權後的累加值作為文句的重要性分數，以此分數做為文句選取的依據。以分類器結果為依據的摘錄方法，是採用不同的摘要特徵，如詞頻、語言評估值、及其他語音特徵，經由一個標註有摘要資訊的文件集訓練出分類器，然後使用此分類器將被摘要文件中所有文句分為屬於摘要內容與不屬於摘要內容二類，被分類成屬於摘要內容這類的文句即可作為摘要內容。目前也已有學者提出以 SVM[11]、Logistic Regression[11]、GMM[6]等不同分類器應用於摘錄式摘要的文句分類問題上。

再者，近些年亦有以機率生成模型為基礎的摘要方法被提出[7][12][13]，其方式是將文件中每一文句視為一個機率生成模型，用以估測文件中所有索引特徵發生在每一文句模型的可能性(Likelihood)，並以此作為每一文句與文件間的相關程度排名的依據。隱藏式馬可夫模型(Hidden Markov Model, HMM)[7][13]及主題混合模型(Topical Mixture Model, TMM)[7][12]均屬於此類的摘要方法，其中主題混合模型可視為隱藏式馬可夫模型的一個特例，進一步將文句視成一個成包含有 K 個潛藏主題的混合模型來估測文件與文句的相關性，可以達到所謂概念比對(Concept Matching)的目的。除上述所描述三大類重要文句的選取方法外，我們亦可進一步對於重要文句進行縮減[8]，像是刪除文句中無意義或不重要的索引特徵。

本論文我們研究以機率生成模型為基礎的摘要方法，提出詞層次主題混合模型(Word Topical Mixture Model, w-TMM)應用於中文語音文件摘要。詞層次主題混合模型的概念與我們過去所提出的主題混合模型[12]相似，均是透過 K 個潛藏主題分佈來估測被摘要文件與每一文句的相關程度。不同之處在於主題混合模型是將文件中的每一文句視為一個機率混合模型，而詞層次主題混合模型是直接將文句中的每一個詞視為一個包含了 K 個潛藏主題的混合模型。如此，文句中每一個詞對於被摘要文件中其它詞的關係即可由 K 個潛藏主題來表示。我們在中文語音廣播新聞語料庫上實作了一系列的實驗，實驗結果顯示出我們所提出之詞層次混合模型的確優於其它摘要模型。

本論文接下來的安排如下：第二節將介紹我們所提出的摘要模型與訓練方式；第三節將呈現相關的實驗設定、實驗結果及分析，第四節為結論及未來展望。

2. 摘要模型

2.1 背景簡介

詞層次主題混合模型的概念是由主題混合模型(Topical Mixture Model, TMM)[7][12]延伸而來。主題混合模型是隱藏式馬可夫模型的一個特例，與隱藏式馬可夫模型[13]同樣是將每一文句視為一個機率生成模型，每個文句有其不同機率分佈。而主題混合模型進一步將文句機率生成模型看成包含有 K 個潛藏主題的機率分佈。其中每一潛藏主題 T_k 分別由一個單連語言模型 $P(w|T_k)$ 所表示，而每一文句 S_i 對於每一潛藏主題 T_k 有不同的權重 $P(T_k|S_i)$。當給定一篇語音文件 D 時，文件 D 與每一文句 S_i 的相關程度可被表示為：

$$P(D|S_i)=\prod_{w\in D}\left[\sum_{k=1}^{K}P(w|T_k)P(T_k|S_i)\right]^{n(w,D)} \quad (1)$$

其中 $n(w,D)$ 為詞 w 於文件 D 中出現的次數，$P(w|T_k)$ 為詞 w 出現於某一個主題 T_k 的機率，$P(T_k|S_i)$ 為文句 S_i 對某一個主題 T_k 的權重值，$P(w|T_k)$ 與 $P(T_k|S_i)$ 機率值均可經由期望值最大化(Expectation-Maximization, EM)訓練所產生。由式(1)可看出，即使文件 D 中的大部分詞並未出現於文句 S_i 中，當文件 D 中的詞出現在某一個主題的機率 $P(w|T_k)$ 與文句 S_i 產生此主題的機率 $P(T_k|S_i)$ 之乘積越大，代表文件 D 與文句 S_i 愈相關，此即可以達到所謂概念比對的目的。

2.2 詞層次混合模型(w-TMM)

相同的概念可進一步延伸到詞層次主題混合模型(Word Topical Mixture Model, w-TMM)，我們可以將每個詞視為包含 K 潛藏主題的機率模型，每一個詞對於不同的潛藏主題有不同權重值，同時在不同的潛藏主題下每一個詞的機率分佈亦所有不同。因

此，我們可以透過詞與這些不同潛藏主題之間的關係，來建立詞與詞之間的關聯性，並且藉由這樣的關連性來計算文句與文件的相關程度。

每一個詞 w_j 與另一個詞 w 之間的關係可透過 K 個潛藏主題來表示:

$$P(w|M_{w_j}) = \sum_{k=1}^{K} P(w|T_k)P(T_k|M_{w_j}) \quad (2)$$

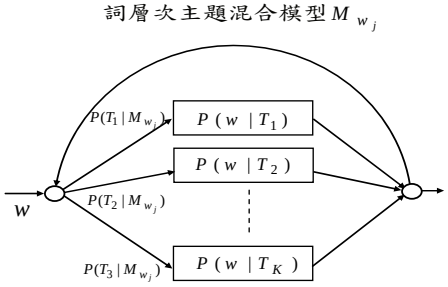
$P(w|M_{w_j})$ 表示詞 w_j 的詞模型 M_{w_j} 生成詞 w 的機率，其中 $P(w|T_k)$ 為一個單連語言模型，表示詞 w 於潛藏主題 T_k 中的機率分佈，$P(T_k|M_{w_j})$ 為詞 w_j 對於潛藏主題 T_k 的權重值。機率分佈 $P(w|T_k)$ 及 $P(T_k|M_{w_j})$ 可經由期望值最大化訓練的方式來估測，據此來描述詞 w_j 與詞 w 間的關係。

由於詞層次主題混合模型是將每一文句 S_i 中的每一個詞 w_j 均視為包含了 K 個潛藏主題的混合模型，每一個詞 w_j 於文句 S_i 中都有一個權重值 $\alpha_{j,i}$，用以表示詞 w_j 於文句 S_i 中的重要性與貢獻程度，$\alpha_{j,i}$ 可以根據 w_j 於文句 S_i 出現的次數來決定。對於一篇被摘要語音文件 D，文件 D 與文件 D 中的任一文句 S_i 的相關程度，可由文句 S_i 中所有的詞 w_j 的主題混合機率模型生成文件 D 中所有詞 W 的機率來決定:

$$P(D|S_i) = \prod_{w\in D}\left[\sum_{w_j\in S_i}\alpha_{j,i}\sum_{k=1}^{K}P(w|T_k)P(T_k|M_{w_j})\right]^{n(w,D)} \quad (3)$$

其中 $\alpha_{j,i}$ 滿足 $\sum_{w_j\in S_i}\alpha_{j,i}=1$，$P(w|T_k)$ 表示文件中詞 W 於潛藏主題 T_k 中的機率分佈，$P(T_k|M_{w_{i,j}})$ 為詞 w_j 對於潛藏主題 T_k 的權重。依據文句 S_i 中每一個詞 w_j 於文句 S_i 的權重 $\alpha_{j,i}$ 以及每一個詞模型 M_{w_j} 生成文件 D 中詞 w 的機率，可得到文句 S_i 生成文件 D 中詞 w 的機率，最後以文句 S_i 生成文件中所有詞機率的連乘積來表示文件 D 與文句 S_i 的相關程度。

由於我們可以從網絡上得到與被摘要語音文件(例如廣播新聞)同一時期(Contemporary)或同一領域(In-domain)的文字新聞文件集 $\{D_c\}$，這些新聞文件通常都會有一句人工產生的標題，它可被視為是一個非常簡潔的文件摘要資訊，於是可以將文件集中每篇文件 D_C 與其標題 H_C 對應關係 (D_C,H_C) 作為詞層次主題混合模型的訓練使用。我們利用新聞文件集中的對應關係 (D_C,H_C)，作為標題 H_C 中每一個詞 w_j 的主題混合模型 M_{w_j} 與對應的文件 D_C 中每一個詞 w 的關聯性，以訓練詞典中所有詞層次主題混合模型 M_{w_j} 的機率分佈 $P(w|T_k)$ 及 $P(T_k|M_w)$：



圖一、詞層次主題混合模型。

$$\hat{P}(w|T_k) = \frac{\sum_{D_C} n(w,D_C)P(T_k|w,H_C)}{\sum_{D_C'}\sum_{w_n\in D_{C'}} n(w_n,D_{C'})P(T_k|w_n,H_{C'})} \quad (4)$$

$$\hat{P}(T_k|M_{w_j}) = \frac{\sum_{D_C}\sum_{w\in D_C} n(w,D_C)P(M_{w_j}|w,M_{H_C})P(T_k|w,M_{w_j})}{\sum_{D_{C'}}\sum_{w'\in D_{C'}} n(w',D_{C'})P(M_{w_j}|w',M_{H_{C'}})} \quad (5)$$

其中 $P(T_k|w,M_{H_C})$ 可表示為:

$$P(T_k|w,M_{H_C}) = \frac{P(w|T_k)\left[\sum_{w_j\in H_C}\alpha_{j,C}P(T_k|M_{w_j})\right]}{\sum_{l=1}^{K}P(w|T_l)\left[\sum_{w_j\in H_C}\alpha_{j,C}P(T_l|M_{w_j})\right]} \quad (6)$$

而 $P(T_k|w,M_{w_j})$ 可表示為:

$$P(T_k|w,M_{w_j}) = \frac{P(w|T_k)P(T_k|M_{w_j})}{\sum_{l=1}^{K}P(w|T_l)P(T_l|M_{w_j})} \quad (7)$$

$P(M_{w_j}|w,M_{H_C})$ 可表示為:

$$P(M_{w_j}|w,M_{H_C}) = \frac{\alpha_{j,C}P(w|M_{w_j})}{\sum_{w_l\in H_C}\alpha_{l,C}P(w|M_{w_l})} \quad (8)$$

我們以文字新聞訓練文件集來訓練所有的詞 w_j 的主題混合模型，然後測試文件中的每一文句均共享這些詞層次主題混合模型，用以估測文件與文句之間的相關程度。圖一為詞層次主題混合模型的概念圖。

3. 實驗設定

3.1 實驗語料

實驗所使用之中文語音文件蒐集自 News 98 新聞網 2001 年 8 月 1 日至 8 月 24 日中午 12:00 到 13:00 的 FM 廣播新聞，共 200 則廣播新聞，分為自動轉寫 (Automatic Transcripts) 與人工轉寫 (Manual

Transcripts)兩部分 [4]。自動轉寫部分為師大資工所大詞彙語音辨識器辨識後之結果，其辨識字正確率達 85.83%[14]；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，分別對這 200 則廣播新聞的人工轉寫文件產生人工摘要，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[4]。

本論文中用來訓練機率式摘要模型（例如，主題混合模型及詞層次混合模型）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)[15]從西元 1991 至 2002 年所發佈的文字新聞中，選出西元 2001 年 8 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採用中央通訊社 2000 與 2001 年所收集而來的文字新聞，約含一萬四千個多個文字檔案。

3.2 評估方式

自動文件摘要的評估至今仍未有一套有系統且一致性的方法，許多文獻所使用之評估方法不一。目前所使用的評估方法大略可分為主觀評估與客觀評估二類。主觀評估是指以人的主觀判斷作為評估自動摘要好壞的依據；客觀評估通常以摘要正確率作為判斷好壞的準則，係採用某種評估法來自動計算自動摘要與人工摘要之間的相似(相關)程度做為摘要的正確率。客觀評估仍需經由人工來產生人工摘要，其中可能亦包含著某種程度上的主觀因素，因此大多使用多份人工摘要來評估。

我們所採用的人工摘要結果包含有摘錄式摘要(或稱為句排名式摘要)及非摘錄式摘要(或稱為重寫摘要)二種。前者可依不同的摘要比例而產生不同長度的摘要內容，例如 10%、20%、30%、50% 等摘要比例；後者則是根據特別摘要比例（例如 20%~30%）針對文件內容所重寫的摘要內容，其長度、內容固定且無法依摘要比例作任意的更改。本論文採用二種客觀評估的方法，分別為餘弦(Cosine)評估[4]及 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)評估[16]，我們於下面小節作詳細說明。

3.2.1 餘弦評估

此評估方法是以計算自動摘要與人工摘要結果的相關度作為評估標準。假設 $A_D(m\%)$ 代表對某篇文件 D 之自動摘要 $m\%$ 摘要比例的結果、$E_{z,D}(m\%)$ 代表第 z 個人對文件 D 以摘錄方式選取 $m\%$ 摘要比例的結果、$R_{z,D}$ 代表第 z 個人對文件 D 重寫摘要結果，則對所有 $|\{D\}|$ 篇新聞的自動摘要正確率被定義為[4]：

$$ACC(m\%) = \frac{1}{|\{D\}|}\frac{1}{Z}\sum_{D\in\{D\}}\sum_{z=1}^{Z}\frac{SIM\left(A_D(m\%),E_{z,d}(m\%)\right)+SIM\left(A_D(m\%),R_{z,d}\right)}{2} \quad (9)$$

其中，Z 為產生人工標準答案的人數，$SIM(\cdot,\cdot)$ 為評估自動摘要與人工摘要結果的餘弦值，公式如下：

$$SIM\left(A_D(m\%),E_{z,d}(m\%)\right)=\frac{\vec{V}_{A_D(m\%)}\bullet\vec{V}_{E_{z,d}(m\%)}}{\left|\vec{V}_{A_D(m\%)}\right|\left|\vec{V}_{E_{z,d}(m\%)}\right|} \quad (10)$$

上式中 $\vec{V}_{A_D(m\%)}$ 與 $\vec{V}_{E_{z,d}(m\%)}$ 分別為自動摘要結果 $A_D(m\%)$ 與人工摘錄式摘要結果 $E_{z,D}(m\%)$ 的向量表示式。

3.2.2 ROUGE 評估

ROUGE 是召回率導向(Recall-Oriented)自動摘要評估方式[6]。此方法是計算自動摘要結果與人工摘要結果的重疊單位元次數，單位元可為 N-連詞(N-gram)、詞順序(Word Sequences)或詞對(Word Pairs)。由於此方法是採用單位元比對的方式，不會產生文句邊界定義的問題，並且適合於多份人工摘要的評估。以詞 N-連單位元（如詞雙連、詞三連、…等）的評估為例，其計算公式如下：

$$ROUGE-N=\frac{\sum_{S\in S_H}\sum_{g_n\in S}C_m(g_n)}{\sum_{S\in S_H}\sum_{g_n\in S}C(g_n)} \quad (11)$$

其中 S_H 為人工摘要結果集合，S 為個別的人工摘要結果，g_n 表示單位元，$C_m(g_n)$ 表示自動摘要結果與人工摘要的重疊（Overlapping）單位元次數，$C(g_n)$ 表示人工摘要結果的單位元次數。目前有許多自動摘要方法皆採用此法作為文件摘要的評估[16]。

本論文中使用詞雙連（即 Word Bigrams）為評估單元（即 ROUGE-2 評估方法），同時僅使用摘錄式人工摘要作為摘要正確率的評估。

表一、各種摘要模型之比較，使用餘弦評估。

摘要比例	VSM	LSA	MMR	SIG	HMM	TMM
10%	0.3596	0.3339	0.3612	0.3440	0.3647	0.3675
20%	0.3895	0.3566	0.3940	0.3816	0.3929	0.3967
30%	0.4428	0.3986	0.4414	0.4302	0.4447	0.4480
50%	0.5409	0.5034	0.5407	0.5465	0.5453	0.5467
70%	0.6027	0.5716	0.6024	0.6042	0.6045	0.6052

表二、各種摘要模型之比較，使用 ROUGE-2 評估。

摘要比例	VSM	LSA	MMR	SIG	HMM	TMM
10%	0.2845	0.2755	0.2875	0.2760	0.2989	0.3043
20%	0.3110	0.2911	0.3218	0.3190	0.3295	0.3345
30%	0.3435	0.3081	0.3493	0.3491	0.3670	0.3688
50%	0.4565	0.4070	0.4668	0.4804	0.4743	0.4753
70%	0.5482	0.4930	0.5595	0.5653	0.5633	0.5631

4. 實驗結果

4.1 基礎實驗結果

我們分別以向量空間模型(VSM)[4]、最大臨界相關(MMR)[6]、潛藏語意分析模型(LSA)[5]、文句特徵值分數方法(SIG)[3]以及機率式模型－隱藏式馬可夫模型(HMM)[7]與主題混合模型(TMM)[7][12]為實驗的基礎實驗。我們以詞為索引單位進行不同摘要比例下的摘要結果比較與分析，表一與表二為使用不同摘要方法的實驗結果，表一為採用餘弦評估方式的結果，表二為使用 ROUGE-2 的評估結果。其中，最大臨界相關(MMR)方法中的權重參數設定為 0.6 [6]，主題混合模型(TMM)為潛藏主題數 8 的摘要結果。

由表一、表二的實驗結果，我們可以看出相較於其他摘要方法，機率式摘要模型(HMM 及 TMM)有較高的摘要召回率，其次為目前較常被使用的 MMR 及 SIG 的摘要方法，而 LSA 摘要結果則顯得較差。

4.2 詞層次混合模型實驗結果

再者，我們探討詞層次混合模型(w-TMM)於語音文件摘要應用的摘要結果。我們以由中央通訊社(CNA)發佈的文字新聞中與測試語料同一時期新聞作為訓練語料。由於這些新聞文件通常都會有一句人工產生的標題，可將文件集中每篇文件 D_C 與其標題 H_C 對應關係 (D_C, H_C) 使用於詞層次主題混合模型的訓練使用，如式(4)-(8)所示。首先，我們將訓練

表三、詞層次主題混合模型(w-TMM)於不同潛藏主題數之結果，使用餘弦評估。

摘要比例	2	4	8	16	32	64
10%	0.3648	0.3646	0.3632	0.3631	0.3652	0.3643
20%	0.3941	0.3944	0.3922	0.3958	0.3948	0.3951
30%	0.4488	0.4488	0.4489	0.4491	0.4513	0.4509
50%	0.5467	0.5473	0.5456	0.5487	0.5501	0.5500
70%	0.6061	0.6063	0.6065	0.6095	0.6084	0.6099

表四、詞層次主題混合模型(w-TMM)於不同潛藏主題數之結果，使用 ROUGE-2 評估。

摘要比例	2	4	8	16	32	64
10%	0.3034	0.3048	0.3041	0.3000	0.3074	0.3082
20%	0.3315	0.3340	0.3322	0.3348	0.3337	0.3366
30%	0.3710	0.3703	0.3696	0.3648	0.3659	0.3692
50%	0.4744	0.4740	0.4712	0.4751	0.4755	0.4755
70%	0.5614	0.5609	0.5604	0.5623	0.5615	0.5618

表五、詞層次主題混合模型(w-TMM)使用檢索文件訓練下，於不同潛藏主題數之結果，使用餘弦評估。

摘要比例	2	4	8	16	32	64
10%	0.3630	0.3603	0.3648	0.3694	0.3738	0.3619
20%	0.3932	0.3925	0.3962	0.4000	0.4037	0.3897
30%	0.4540	0.4478	0.4534	0.4552	0.4583	0.4521
50%	0.5451	0.5491	0.5482	0.5475	0.5468	0.5524
70%	0.6084	0.6075	0.6079	0.6080	0.6105	0.6094

表六、詞層次主題混合模型使用檢索文件訓練下，於不同潛藏主題數之結果，使用 ROUGE-2 評估。

摘要比例	2	4	8	16	32	64
10%	0.3013	0.2997	0.3053	0.3152	0.3193	0.2986
20%	0.3282	0.3273	0.3302	0.3379	0.3437	0.3209
30%	0.3731	0.3641	0.3703	0.3694	0.3716	0.3713
50%	0.4732	0.4741	0.4730	0.4700	0.4676	0.4759
70%	0.5632	0.5619	0.5619	0.5571	0.5606	0.5572

語料中的全部文件，共一萬四千多個檔案，與其對應關係 (D_C, H_C) 用於訓練詞層次主題混合模型所需之機率值 $P(w|T_k)$ 及 $P(T_k|M_w)$ 。在使用詞層次主題混合模型於語音文件摘要時，我們會同時考慮每一個詞 w 實際上出現於文句的機率分佈 $P_{ML}(w|S_i)$，並將其與式(3)結合成為：

$$P(D|S_i) = \prod_{w \in D} \left[\lambda \cdot \sum_{w_j \in S_i} \alpha_{j,i} \sum_{k=1}^{K} P(w|T_k) P\left(T_k \middle| M_{w_j}\right) + (1-\lambda) P_{ML}(w|S_i) \right]^{n(w,D)} \quad (12)$$

然後，根據式(12) 的機率值來進行文句的排名。表

三與表四分別為在不同評估方式下之詞層次主題混合模型之結果，可看出不同潛藏主題數摘要結果的召回率。

從實驗結果中，我們可以看出在餘弦評估下（表三）的摘結果與表一中所列的機率式模型結果相似，摘要正確率並沒有提昇；而在 ROUGE-2 方法的評估下，詞層次主題混合模型的摘要正確率較主題混合模型有些微的提昇，這樣的結果顯示出詞層次主題模型適用於語音文件摘要中，並且可提昇摘要召回率。

因此，我們進一步考慮詞層次主題混合模型的效率及訓練文件集與測試語料之間的關聯性，以期能提昇詞層次主題混合模型摘要召回率。由於訓練語料共包含有一萬四千多個檔案，大量的檔案數使得詞層次主題混合模型訓練花費很多時間，因此我們希望減少訓練文件數來提高訓練時的效率，同時考慮僅以與測試語料有關聯之訓練文件來訓練詞層次主題混合模型，希望提昇對測試語音文件中文句所包含的詞模型之估測。

我們使用資訊檢索系統來實行上述的二點考量，透過檢索得到與每篇測試語音文件最相關的前 *N* 篇訓練語料中的文件作為訓練文件集，不僅可減少訓練文件數，亦可考慮到訓練文件集與測試語音文件之間的關聯性。本論文中所使用的資訊檢索系統是採用隱藏式馬可夫檢索模型來進行文件檢索[17]。我們分別以 200 篇測試文件為查詢，來檢索訓練語料中的前 *N* 篇相關文件作為詞層次主題混合模型的訓練文件集。實驗中 *N* 設定為 3，每篇測試文件取回前 3 名相關文件，共 600 篇的訓練文件用以訓練詞層次主題混合模型。訓練後的詞層次主題混合模型摘要結果列於表五、表六。相較於表三、表四的摘要結果，表五、表六中所呈現的結果有顯著的提昇。因此，我們可以發現訓練文件集對於詞層次主題混合模型的影響，若可以找到一個好的訓練文件集來訓練詞層次主題混合模型，那麼詞層次主題混合模型則可對摘要召回率有所提昇。此外，經由實驗結果證明我們所提出之詞層次主題混合模型優於其它摘要摘型。

5. 結論與未來展望

多數的摘要方法均屬於逐字比對的方式，也就是說在摘錄重要文句時，是以文句中所包含的文字內容與文件的文字內容的相關程度來決定。而潛藏語意分析模型與主題混合模型是屬於概念比對的方式，即是以文句所表達的概念與文件中的主題概念的相關程度來決定文句的重要性，不管文字內容是否相同，因此概念比對的方式更適合應用於文件摘中。中文文字中往往很多不同的文字皆可表達相同的概念，即使用字遣詞不同亦能表示同一個主題，所以概念比對的摘要模型所產生的摘要將更相似於人工摘要結果。在本論文中我們延伸主題混合模型的方法，更進一步考慮每個詞的不同主題性，並且找出詞與詞之間的相關性，並且運用於語音文件摘要中。我們實作不同摘要方法於中文語音廣播新聞語料上，實驗結果顯示出機率式摘要模型可達較高的摘要正確率，而當選用適當的訓練文件集時，所提出之詞層次主題混合模型的摘要結果又較隱藏式馬可夫模型及主題混合模型有較好的召回率。

6. 參考文獻

[1] I. Mani and M. T. Maybury (Eds.), *Advances in Automatic Text Summarization*. Cambridge. MA: MIT Press, 1999.

[2] P.B. Baxendale, “Machine-Made Index for Technical Literature-An Experiment”, *IBM Journa*l, pp. 354-361, October 1958.

[3] Makoto Hirohata, Yousuke Shinnaka, Koji Iwano and Sadaoki Furui, “Sentence Extraction-Based Presentation Summarization Techniques and Evaluation Metrics”, in *Proc. ICASSP 2005*.

[4] Y. Ho, *An initial study on automatic summarization of Chinese spoken documents*, Master Thesis, National Taiwan University, July 2003.

[5] Y. Gong and X. Liu, “Generic text summarization using relevance measure and latent semantic analysis,” in *Proc. ACM SIGIR 2001*, pp. 19-25.

[6] Gabriel Murray, Steve Renals, Jean Carletta, “Extractive Summarization of Meeting Recordings”, in *Proc. Eurospeech 2005*.

[7] 陳怡婷、黃耀民、葉耀明、陳柏琳,“中文語音文件自動摘要之摘要模型”, 「第十屆人工智慧與應用研討會」, December 2-3, 2005.

[8] S. Furui, T. Kikuchi, Y. Shinnaka, C. Hori, “Speech-to-Text and Speech-to-Speech Summarization of Spontaneous Speech”, *IEEE trans. speech and audio processing*, 12(4), July 2004.

[9] S. Maskey, J. Hirschberg, “Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization”, in *Proc. Interspeech 2005*.

[10] K. Koumpis, S. Renals, “Automatic Summarization of Voicemail Messages Using Lexical and Prosodic Features”, *ACM trans. Speech and Language Processing* 2(1), February 2005.

[11] X. Zhu, G. Penn, “Evaluation of Sentence Selection for Speech Summarization”, in *Proc* the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP-05), pp. 39-45. September 2005.

[12] B. Chen, Y.M. Yeh, Y.M. Huang, Y.T. Chen, “Chinese Spoken Document Summarization Using

Probabilistic Latent Topical Information," in *Proc. ICASSP 2006*.

[13] Y.T. Chen, S. Yu, H.M. Wang, and B. Chen, "Extractive Chinese Spoken Document Summarization Using Probabilistic Ranking Models," in *Proc. ISCSLP 2006*.

[14] B. Chen, J.W. Kuo, W.H. Tsai, "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription," in *Proc. ICASSP 2004*.

[15] Central News Agency (CNA), http://210.69.89.224/search/hypage.cgi?HYPAGE=login.htm

[16] C.Y. Lin, "ROUGE: Recall-oriented Understudy for Gisting Evaluation," 2003, http://www.isi.edu/~cyl/ROUGE/.

[17] B. Chen, H.M. Wang, L.S. Lee, "A Discriminative HMM/N-gram-based Retrieval Approach for Mandarin Spoken Documents," *ACM Trans. Asian Language Information Processing* 3 (2), 2004.