The 2014 Conference on Computational Linguistics and Speech Processing
ROCLING 2014, pp. 3-20

探究新穎語句模型化技術於節錄式語音摘要

Investigating Novel Sentence Modeling Techniques for Extractive Speech Summarization

劉士弘 Shih-Hung Liu, 陳冠宇 Kuan-Yu Chen,
謝育倫 Yu-Lun Hsieh, 王新民 Hsin-Min Wang, 許聞廉 Wen-Lian Hsu
中央研究院資訊科學研究所
{journey, kychen, morphe, whm, hsu}@iis.sinica.edu.tw

陳柏琳 Berlin Chen
國立臺灣師範大學資訊工程學系
berlin@ntnu.edu.tw

摘要

近幾年來，基於語言模型化(Language Modeling, LM)架構之摘要方法已初步在節錄式語音摘要任務上展現具競爭性的效能。在此架構下，對於被摘要文件每一句候選語句之語句模型的建立，可透過虛擬相關回饋(Pseudo Relevance Feedback, PRF)策略來獲得較可靠的參數估測。一般來說，虛擬相關回饋在執行上可分為兩個階段：1)相關資訊(或者說虛擬相關文件)的選取；2)語句模型化與參數重新估測。首先，有別於現有基於語言模型化架構之摘要方法都聚焦在語句模型參數的重新估測，本論文深入探討與應用各種適合於節錄式語音文件摘要的虛擬相關文件選取技術，用以強化語句模型的參數估測。再者，本論文更進一步地考量使用每一語句的非相關性(Non-relevance)資訊對於虛擬相關文件選取的影響。同時，我們亦額外嘗試基於重疊分群(Overlapped Clustering)概念來有效地選取重要的虛擬相關文件。最後，本論文探索使用三混合模型(Tri-Mixture Model)來表示每一語句，期盼其能更精確地表示一句語句之獨特詞彙使用和語意相關資訊。本論文的語音文件摘要實驗語料是採用公開的公視廣播新聞(MATBN)；實驗結果顯示，相較於其它現有虛擬相關文件選取技術，我們所發展的虛擬相關文件選取技術能提供相當不錯的摘要效能改善。

關鍵詞：節錄式自動摘要、虛擬相關回饋、語句模型、非相關資訊、重疊分群

一、緒論

隨著網際網路的普及與蓬勃發展，大量含有語音資訊的多媒體內涵快速地傳遞並分享於全球各地，像是電視新聞、課程演講、會議錄音和語音郵件等。由於這些多媒體內涵透過自動語音辨識(Automatic Speech Recognition, ASR)處理後都可以表示成語音文件及其對應的自動轉寫(Automatic Transcripts)供瀏覽使用，節錄式語音摘要(Extractive Speech Summarization)在過去十年逐漸成為熱門研究議題[17][25]。節錄式語音摘要目標在於根據一定的摘要比例，從語音文件中選取重要語句並組合成摘要，以期能夠扼要的表示語音文件主要的主題或語意資訊；藉此，使用者能迅速地瀏覽大量多媒體內涵並能充分理解其中主題或語意資訊。

當前主流的節錄式語音摘要方法大致可分為三類：(1)基於文件結構或語句位置資訊來選取重要語句；(2)基於特定詞彙或語意資訊之非監督式(Unsupervised)摘要方法；(3)需要使用人工摘要標註來訓練模型之監督式(Supervised)摘要方法。第一類摘要方法大都是簡單地從文件的緒論(Introductoion)或結論(Conclusion)所在段落擷取出若干語句來組成摘要[1]；此類方法僅適用在特定具有一致結構的文字或語音文件上，因此在實際應用上有其侷限性。另一方面，第二類摘要方法通常將自動摘要任務視為如何排序並挑選具代表性(或重要性)語句之問題，其方法通常是以非監督式方式產生出一種語句層次的摘要特徵或重要分數以供語句排序使用。而第二類摘要方法又可進一步地分成三小類：(I)以向量(Vector)為基礎的方法，包含有向量空間模型(Vector Space Model, VSM)[8]、潛藏語意分析(Latent Semantic Analysis, LSA)[8][13]及最大邊際關聯(Maximal Marginal Relevance, MMR)[2]等；(II)以圖(Graph)為基礎的方法，具代表性的有馬可夫隨機漫走(Markov Random Walk, MRW)[30]、詞彙排序(LexRank)[7][23]及最小支配集(Minimal Dominating Set)[28]等；(III)以組合最佳(Combinatorial Optimization)為基礎的方法，包括有次模(Submodularity)[14]以及線性整數規劃(Linear Integer Programming)[22]等。最後，第三類監督式摘要方法通常將自動摘要之任務視為二元分類問題(Binary Classification)，亦即將語句區分為摘要語句或非摘要語句。在訓練階段，必須事先準備好一些訓練文件以及其對應的人工標註過摘要資訊，然後結合各種詞彙、語意或音韻等特徵來表示語句，並且透過各種分類器的學習機制進行摘要(分類)模型的訓練；在測試階段，對於將被摘要之文件，此類方法將文件裡的每一句語句進行二元分類，即可依所設定摘要比例來取摘要語句以產生出摘要。在此類方法中，較著名的包括簡單貝氏分類器(Naïve-Bayes Classifier)[11]、高斯混合模型(Gaussian Mixture Model, GMM)[24]、隱藏式馬可夫模型(Hidden Markov Model, HMM)[6]、支援向量機(Support Vector Machines, SVM)[10]與條件隨機場域(Conditional Random Fields, CRF)[29]等。由於監督式摘要方法所使用的模型在訓練時必須使用一定數量文件及其對應經人工標註過摘要資訊，所以當它們被應用到新的摘要任務或應用領域時，相較上述兩類摘要方法而言，是會耗費許多人力與時間的。值得一提的是，自動摘要研究也可從其它不同面相來進行探討，包括了來源、需求、方式、用途等，有興趣的讀者可參考相關文獻[17][20][21][25]進行更深入的瞭解。

有別於上述的摘要方法，近期有一些基於語言模型化(Language Modeling, LM)架構之摘要方法被提出，並且初步在節錄式文字或語音摘要任務上展現不錯的效能。在此架構下，對於被摘要文件每一句候選語句之語句模型的建立，可透過虛擬相關回饋(Pseudo Relevance Feedback, PRF)策略來獲得更加可靠的參數估測。一般來說，虛擬相關回饋在執行上可分為兩個階段：1)相關資訊(或者明確地說，虛擬相關文件)的選取；2)語句模型化與參數重新估測。本論文同樣也基於語言模型化架構來發展語音摘要方法，其貢獻主要有三方面。首先，有別於現有基於語言模型化架構之摘要方法都聚焦在語句模型參數的重新估測，本論文深入探討與應用各種適合於節錄式語音文件摘要的虛擬相關文件選取技術，用以強化語句模型的參數估測。其次，本論文更進一步地考量使用每一語句的非相關性(Non-relevance)資訊對於虛擬相關文件選取的影響。同時，我們亦額外嘗試基於重疊分群(Overlapped Clustering)概念來有效地選取重要的虛擬相關文件做為語句模型的參數估測之依據。最後，本論文探索使用三混合模型(Tri-Mixture Model)來表示每一語句，期盼其能更精確地表示一句語句之獨特詞彙使用和語意相關資訊。

本論文後續安排如下：第二節首先介紹使用語言模型於節錄式語音摘要任務之原理，然後闡述虛擬相關回饋的觀念及其現有虛擬相關文件選取技術；第三節將介紹本論文提出之新穎式虛擬相關文件選取技術；第四節則介紹現有各種關聯模型，並且說明如何結合語句關聯性資訊來改進語句模型之估測，使其得以更精準地代表語句的語意內容；第五節介紹實驗語料與設定以及摘要評估之方法；第六節說明實驗結果及其分析；

最後，第七節為結論與未來研究方向。

二、使用語言模型於語音文件摘要

在過去近二十年來，各種語言模型在資訊檢索任務中已被廣泛地應用，並且有不錯的實務成效[34]。近期在語音摘要領域，亦開始有一些基於語言模型化架構的非監督式摘要方法被提出。本節將先簡介兩種常見的、基於語言模型化架構的摘要方法：其一為使用語句語言模型生成文件的文件相似度量值(Document Likelihood Measure, DLM)的摘要方法[3]；另外為使用庫爾貝克-萊伯勒離散度量值(Kullback-Leibler Divergence Measure, KL)[13][15]來計算文件模型和語句模型間之距離的摘要方法。接著，我們將闡述如何利用虛擬相關回饋(Pseudo Relevance Feedback)概念來獲得更可靠的語句模型估測，並介紹數個在資訊檢索領域已被發展出的新穎虛擬相關文件選取技術。

2.1、文件相似度量值

我們可以把語音文件摘要任務視為是資訊檢索的問題。一般來說，資訊檢索(Information Retrieval, IR)旨在尋找相關文件(Relevant Document)來回應使用者所送出的查詢(Query)或資訊需求(Information Need)。同樣地，在從事語音文件摘要時，我們可將每一篇被摘要文件視為是查詢，而文件中的每一句語句(Sentence)視為候選資訊單元(Candidate Information Unit)；據此，我們可以假設在被摘要文件中，與文件本身愈相關的語句就愈有可能是可用來代表文件主旨或主題之摘要語句。

當給定一篇被摘要文件 D 時，文件中每一句語句 S 的事後機率 $P(S \mid D)$ 可以用來表示語句 S 對於文件 D 的重要性。當使用語言模型來計算 $P(S \mid D)$ 時，我們透過貝氏定理(Bayes' Theorem)將 $P(S \mid D)$ 展開成[3]：

$$P(S \mid D) = \frac{P(D \mid S)P(S)}{P(D)}, \tag{1}$$

其中 $P(D)$ 是文件 D 的事前機率，由於 $P(D)$ 不影響語句的排序結果，故可省略不討論；另一方面， $P(S)$ 是語句 S 的事前機率，可以使用各式非監督式方法或監督式方法來求得[3]。在此先假設語句的事前機率為一個均勻分布(Uniform Distribution)，所以 $P(S)$ 亦可省略。最後， $P(D \mid S)$ 是語句 S 所形成的語言模型生成文件 D 之機率(或稱作文件相似度)，可以用來表示文件 D 與語句 S 之間的相似關係，如果語句 S 生成文件 D 的機率值愈高，代表語句 S 與文件 D 愈為相似(語句愈能代表文件 D)，即愈有可能是摘要語句。我們可以更進一步地假設文件 D 中詞彙與詞彙之間是獨立的，並且不考慮每一個詞彙在文件 D 中發生的順序關係(即詞袋假設(Bag-of-Word Assumption))，則語句 S 生成文件 D 的文件相似度量值(Document Likelihood Measure, DLM) $P(D \mid S)$ 可拆解成文件 D 中每一的詞彙 w 個別發生的條件機率之連乘積：

$$P(D \mid S) = \prod_{w \in D} P(w \mid S)^{C(w,D)}, \tag{2}$$

此種方法是為語句 S 建立一個語句模型(Sentence Model) $P(w \mid S)$ ， w 是一個出現在文件 D 中的詞彙， $C(w, D)$ 是詞彙 w 出現在文件 D 中的次數。其中，我們可利用最大化相似度估測(Maximum Likelihood Estimation, MLE)的方式來建立每一個語句的語句模型：

$$P(w \mid S) = \frac{C(w,S)}{|S|}, \tag{3}$$

在(3)中， $C(w,S)$ 表示詞彙 w 在語句 S 中出現的次數， $|S|$ 則表示語句 S 所含的總詞數。

值得注意的是，由於語句 S 通常僅由少數字詞所組成，因此容易遭遇資料稀疏(Data Sparseness)的問題，這會使得語句模型使用最大化相似度估測時，不僅可能無法準確地估測每一個詞彙在語句中真正的機率分佈，也可能因為某些詞彙的條件機率值為零，導致語句 S 產生文件 D 的機率值為零。為了減輕上述的現象，本論文使用 Jelinek-Mercer 平滑化(Smoothing)技術藉由使用以大量文字語料訓練而成的背景單連語言模型(Background Unigram Language Model)來調適語句模型[33]，故 $P(D|S)$ 可進一步地表示成：

$$P(D|S)=\prod_{w\in D}[\lambda\cdot P(w|S)+(1-\lambda)\cdot P(w|BG)]^{C(w,D)}, \tag{4}$$

其中， $P(w|BG)$是詞彙 w 由背景單連語言模型所產生的機率值。

2.2、庫爾貝克-萊伯勒離散度量值

語言模型使用於文件摘要的研究中，除了可被用於計算語句生成文件的可能性外，另一種方式為藉由庫爾貝克-萊伯勒離散度量值(Kullback-Leibler Divergence Measure, KL)，來評估文件中每一個語句的重要性。當使用庫爾貝克-萊伯勒離散度量值於摘要任務中，被摘要文件 D 和 D 中的每一個語句 S 都將分別被描述為一個單連語言模型；當相對於被摘要文件 D 的文件模型(Document Model)，語句模型的離散度量值愈小時，則代表語句與文件愈相關，亦即語句 S 愈重要。在此摘要架構下，排序語句重要性的公式如下[15]：

$$KL(D\|S)=\sum_{w\in V}P(w|D)\log\frac{P(w|D)}{P(w|S)}, \tag{5}$$

其中，V (Vocabulary)表示一個由語言裡所有可能的語彙所形成的集合。本論文的研究中，文件模型 $P(w|D)$ 的建立方式與語句模型相同(參照式(3))。當我們更進一步地對(5)作分析時，可以發現當文件模型僅使用最大化相似度估測(MLE)的前提下，採用庫爾貝克-萊伯勒離散度量值所得到的語句排序將與使用文件可能性(Document Likelihood)測量方式(即文件相似度量值)所得到的結果是相同的[4]。

由於使用庫爾貝克-萊伯勒離散度量值時，不僅每一句語句被表示成語句模型，每一篇被摘要文件 D 亦被視為一個文件(機率生成)模型，而文件模型在經由各式語言模型調適與平滑化的技巧下，可以有系統地、適當地調適文件模型的機率分佈；因此相較於文件相似度量值(DLM)只能針對語句模型進行調適，庫爾貝克-萊伯勒離散度量值(KL)能透過不同模型參數估測技術的使用來求取語句模型和文件模型，以獲得更佳的自動摘要效能。

2.3、虛擬相關回饋

通常，文件中的語句僅由少許的詞彙所組成，當語句模型使用最大化相似度估測時，容易遭遇資料稀疏的問題；再者，由這語句 S 中些許的表面詞彙是遠不夠正確估算語句 S 與被摘要文件 D 之間的相似度(或低估了此相似度)，所以藉由背景語言模型進行語句模型之調適為最常見的方法之一(參照式(4))。

為了有效解決語句的資料稀疏及相似度被低估的問題，我們可利用在資訊檢索(Information Retrieval)領域被廣泛應用的虛擬相關回饋(Pseudo Relevant Feedback, PRF)技術來強化語句模型(重新估測或對其做調適)[5]。為此目的，當虛擬相關回饋運用於文件摘要領域中時，會將每一語句 S 當成是一個查詢(Query)，然後輸入到一個資訊檢索系統中，找出一些與語句 S 最可能相關的文件，而這些文件就稱之為虛擬相關文件

(Pseudo Relevant Documents)；一個最簡單的方式即是選取排名最前面(檢索分數最高)的幾篇文件(Top-ranked Documents)。有了這些虛擬相關文件後，就可以利用它們來增進語句模型以解決語句資料稀疏及其相似度低估之問題。

當使用虛擬相關回饋技術時，通常會遭遇到兩個挑戰。第一為如何選取、純化(Purify)虛擬相關文件，意即如何去除虛擬相關文件的冗餘性(Redundancy)和摒除非相關資訊(Non-relevant Information)。第二則是如何有效地運用虛擬相關文件來重新估測語句模型或進一步調適。對於第二個挑戰，已經有許多學者提出各種不同的關聯模型方法，常見的有關聯模型(Relevance Model, RM)[12]和簡單混合模型(Simple Mixture Model, SMM)[32]等(我們將在第四節介紹上述各種關聯模型)。然而對於第一個挑戰的研究，雖亦有少數學者提出一些虛擬相關文件選取方法，但相較於第二個挑戰的研究，仍顯較匱乏。

關於虛擬相關文件的選取方式，過去在資訊檢索領域有學者陸續提出間隔式最高 K (Gapped Top K)選取法[27]、群中心(Cluster Centroid)選取法[27]以及主動式-關聯多元密度(Active-RDD)選取法[31]，其中 RDD 三個英文字母分別代表關聯(Relevance)、多元(Diversity)以及密度(Density)，下面將簡單介紹上述虛擬相關文件選取方法。

間隔式最高 K 選取法就是在最高排序文件中每間隔 J(例如兩個間隔，J 為 2)挑選出 K 個相關文件出來當作是最高排序文件，簡單的例子如下：假設最高排序文件有 10 篇，我們每間隔 2 篇要挑出最高 3 篇出來，則第一、第四及第七篇會被挑選出來當作是最高排序文件。間隔式最高 K 選取法的主要思想是要挑選出具有多元性的文件，但其實此選取方法也是相當不穩定的。群中心選取法則是先將最高排序文件作分群(Clustering)，分群方法可以是任意的，常用的分群方法為 K 中心(K-means)分群法，然後再從分出來的每一群中挑選出一篇最相關的文件，以此構成新的最高排序文件，由分群的觀念可知，群中心選取法旨在選取出具有多元性的文件，與間隔式最高 K 選取法相較，群中心選取法是一個比較穩定的選取方法。主動式-關聯多元密度選取法為同時考量最高排序文件中的關聯性、多元性以及密度性的一種貪婪(Greedy)選取法[31]。以上選取方法常見於資訊檢索領域中，有興趣的讀者可參考相關文獻[5]，本論文是首次將上述方法運用在(語音)文件摘要任務中並做深入探討。

三、新穎式虛擬相關文件選取

基於主動式-關聯多元密度選取法，除了考量到虛擬相關文件(最高排序文件)中的關聯性、多元性以及密度性之外，我們認為非相關性(Non-relevance)資訊(在這裡是指語句的非相關性資訊)也是相當重要的線索，可以用來幫助重新選取更好的虛擬相關文件，因此本論文提出額外考量非相關資訊以改進主動式-關聯多元密度選取法，稱之為主動式-關聯多元密度非相關(Active-RDDN)選取法。另一方面，本論文也提出使用重疊分群(Overlapped Cluster)的概念來幫助重新選取更有效的虛擬相關文件，茲介紹如下：

3.1、主動式-關聯多元密度非相關選取

假設語句 S 已經輸入到資訊檢索系統中並得到了最高排序文件 $\mathbf{D}_{\mathbf{Top}}=\{D_1,D_2,\ldots,D_M\}$，那主動式-關聯多元密度非相關選取則是從最高排序文件中迭代地(Iteratively)同時考慮四種重要因素(關聯性、多元性、密度性以及非相關資訊)來重新選取更具代表性的文件集。具體地說，最高排序文件中的每個候選(Candidate)文件 D_m 都會有著同時考量四種因素的一個線性結合的分數，其選取公式如下：

$$D^* = \underset{D_m \in \mathbf{D}_{\text{Top}} - \mathbf{D}_{\text{P}}}{\arg\max} \Big[(1-\alpha-\beta-\gamma)\cdot M_{Rel}(S,D_m) + \alpha \cdot M_{NR}(S,D_m) + \beta \cdot M_{Diversity}(D_m) + \gamma \cdot M_{Density}(D_m)\Big], \tag{6}$$

其中 $\mathbf{D}_{\text{P}}$ 為已經選入的文件集，$M_{Rel}(S,D_m)$、$M_{NR}(S,D_m)$、$M_{Diversity}(D_m)$及 $M_{Density}(D_m)$分別代表候選文件 D_m 的關聯性量值、非相關資訊性量值、多元性量值、以及密度性量值，而 α、β、γ 為可調參數且其總和為 1(即 $\alpha+\beta+\gamma=1$)。值得一提的是，式(6)與早期用於資訊檢索及摘要領域的經典公式最大邊際關聯(Maximal Marginal Relevance, MMR)相似[2]。另外，關聯性量值 $M_{Rel}(S, D_m)$可定義為文件 D_m 語句 S 的負庫爾貝克-萊伯勒離散度量值(即-$KL(D_m\|S)$)。下面將分別介紹非相關資訊量值、多元性量值、以及密度性量值。

3.1.1、非相關性資訊量值

對於一個語句 S，其非相關性(Non-relevance)資訊通常可以從第一次資訊檢索時排在最後面的一些文件(最低排序(Low-Ranked)文件)來代表，那麼語句 S 的非相關模型 $P(w|NR_S)$ 便可由最低排序文件來估測，而非相關性資訊量值可由下面式子表示：

$$M_{NR}(S,D_m) = KL\big(NR_S \big\| D_m\big), \tag{7}$$

其中 NR_S 表示語句 S 的非相關性資訊，亦即最低排序文件。若非相關性資訊量值越小，表示候選文件 D_m 與對應於語句 S 的非相關性資訊很相像，則不應該被選取當成代表性文件；反之，則表示候選文件 D_m 與語句 S 的非相關資訊離較遠，應該有機會被選取起來當成代表性文件。實際上，虛擬相關文件(最高排序文件)相對於最低排序文件是相當少量的，所以實作上我們可利用全部的文件集來估測語句 S 的非相關模型，更明確的說，就是利用背景語言模型 $P(w|BG)$來當作是語句 S 的非相關模型 $P(w|NR_S)$。

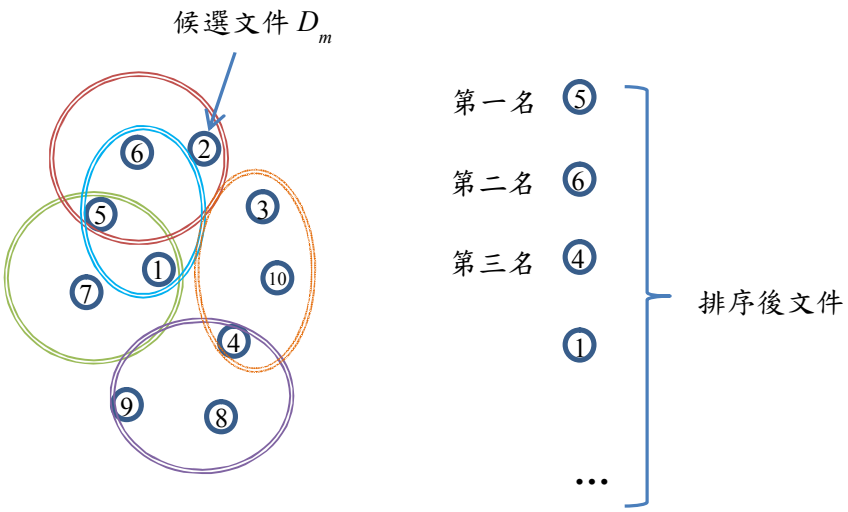
3.1.2、多元性量值

在資訊檢索領域中，多元性(Diversification)的考量已備受矚目，相較於傳統只考量關聯性的查詢使得有太多重複及冗餘查詢結果，額外考慮多元性是希望能提供查詢者多元且多樣性的查詢結果，以滿足不同需求的使用者。而在虛擬相關回饋下，若最高排序文件的重複性文件太多，則會導致後面的關聯模型估測有太多的冗餘資訊，導致模型估測不良，所以多元性量值考量的點就是希望選取過的文件或與已選文件相近的文件不要再被選取進來，亦即去除最高排序文件的冗餘性，使得其後面的關聯模型估測更精準。多元性量值就是從已選的文件 $\mathbf{D}_{\text{P}}$ 中找出與候選文件 D_m 有最小對稱離散度量的值，可由下列表示：

$$M_{Diversity}(D_m) = \min_{D_j \in \mathbf{D}_{\text{P}}} \frac{1}{2}\cdot\Big[KL\big(D_j \,\|\, D_m\big) + KL\big(D_m \,\|\, D_j\big)\Big], \tag{8}$$

3.1.3、密度性量值

另一方面，最高排序文件中的結構(Structural)資訊可以被當成是一個線索來幫助選取代表性文件，其主要目的是希望在考量多元性資訊的同時，也應避免選取到過度極端的文件，因為過度極端文件很可能是錯誤的資訊。為了實現這個想法，我們可以利用計算最高排序文件中的候選文件 D_m 與其他候選文件 D_h 的負平均對稱(Negative Average Symmetric)離散度量值來達成，其公式如下所示：



(a) (b)

圖一、使用重疊分群概念的候選文件選取示意圖，
(a)利用 *k*-NN 為每個候選文件 D_m 找出重疊分群，並計算其重疊分群個數。
(b)依據每一候選文件 D_m 的重疊分群個數做排序。

$$M_{Density}(D_m)=\frac{-1}{|\mathbf{D}_{\text{Top}}|-1}\cdot\sum_{\substack{D_h\in\mathbf{D}_{\text{Top}}\\ D_h\neq D_m}}\left[KL(D_h\parallel D_m)+KL(D_m\parallel D_h)\right], \tag{9}$$

其中$|\mathbf{D}_{\text{Top}}|$為最高排序文件之個數。若負平均對稱離散度量值越大，表示此候選文件 D_m 與最高排序文件中的其他候選文件 D_h 很接近，亦即可能是最高排序文件中心點(密度大的文件)，所以此候選文件 D_m 可能是個重要的文件，應該要被選入；反之，就會離中心點較遠(密度較低)，有可能會是不重要的文件，就不應該被選入。

3.2、重疊分群

本論文提出使用重疊分群(Overlapped Cluster)的概念來重新選取更好的虛擬相關文件(即 $\mathbf{D}_{\text{P}}$)以利接下來各種不同的關聯模型估測。其重疊分群選取方法為一個三步驟的演算法，示意圖如圖一所示，其演算法描述如下：

1. 第一步驟：計算最高排序文件中兩兩候選文件的相似度，在此是將候選文件表達成向量空間模型(Vector Space Model)並使用餘弦相似度(Cosine Similarity)量值來做計算。
2. 第二步驟：利用 *k*-最近鄰居(*k*-NN)來為每個候選文件 D_m 找出 k 個最接近的相關文件，並形成一個群(每個候選文件都會形成一個分群，而此分群中會有 k+1 個文件)。
3. 第三步驟：對於每個候選文件 D_m 都去計算重疊分群的個數(以圖一例子來說明，候選文件編號 5 被三個分群所包圍，所以其重疊分群個數為 3)，並且按照重疊分群個數來對每個候選文件作排序，排序後即可得到新的最高排序文件。

重疊分群的概念在資訊檢索領域中已有些許研究[18]，它是利用最高排序文件的結構資訊來幫助訓練語言模型，而本論文的思想是要利用重疊分群的概念來找出支配(Dominate)文件，若一個候選文件的重疊分群個數很多的話，表示它是很重要的且能夠支配其他候選文件，則應該要被選為代表性文件。本論文是首次將重疊分群的概念用在(語音)文件摘要任務上。

四、各種關聯模型之簡介

當我們透過資訊檢索系統已取得虛擬相關文件(最高排序文件，即 $\mathbf{D_{Top}}$)，或進一步地使用上一節提出之方法來改善虛擬相關文件後(即 $\mathbf{D_P}$)，接下來就要做模型估測，底下介紹常見的模型包含有關聯模型(Relevance Model, RM)、簡單混合模型(Simple Mixture Model, SMM)以及三混合模型(Tri-Mixture Model, TriMM)。

4.1、關聯模型

關聯模型的基本假設是認為每一語句 S 皆是被用來描述一個概念、想法或主題，我們稱之為語句的關聯類別(Relevance Class) R_S。在本論文中，我們的目標是想進一步地模型化關聯類別所代表的資訊，藉此來豐富語句模型所能傳達的語意內容或主題特性。然而，實際上每一語句 S 的關聯類別 R_S 是非常難以求得的；為此，我們透過虛擬相關回饋(Pseudo Relevant Feedback)來尋找與關聯類別可能相關的一些文件，並藉由這些文件來近似關聯類別 R_S。更明確地，在實作上我們將虛擬相關文件(最高排序文件)$\mathbf{D_{Top}}=\{D_1,D_2,...,D_M\}$或透過上一節所介紹的選取方法來產生較佳的虛擬相關文件 $\mathbf{D_P}$ 用以代表關聯類別 R_S。接著，透過檢視詞彙 w 與語句 S 在這些虛擬相關文件中同時出現之關係，可計算出詞彙與語句的聯合機率[12]：

$$P_{\mathrm{RM}}(w,S)=\sum_{D_m\in\mathbf{D}_\mathrm{P}}P(w,S\mid D_m)P(D_m), \tag{10}$$

當我們進一步地假設在給定某一篇虛擬相關文件時，詞彙與語句是獨立的，並且語句內的詞彙也是獨立且不考慮其先後次序(即所謂的詞袋假設)，則透過虛擬相關回饋所估測的語句模型為：

$$P_{\mathrm{RM}}(w\mid S)=\frac{\sum_{D_m\in\mathbf{D}_\mathrm{P}}\prod_{w'\in S}P(w'\mid D_m)P(w\mid D_m)P(D_m)}{\sum_{D_{m'}\in\mathbf{D}_\mathrm{P}}\prod_{w''\in S}P(w''\mid D_{m'})P(D_{m'})}, \tag{11}$$

我們稱之為關聯模型(Relevance Model, RM)。關聯模型的優點在於藉由虛擬相關文件的資訊，可以更清楚地知道語句所蘊含的資訊、所欲表達的內涵，所以相較於傳統使用最大化相似度估測的語句模型，可更準確地表達語句的語意內容或主題特性，以提升摘要的成效。

4.2、簡單混合模型

簡單混合模型的基本想法是假設由虛擬相關回饋技術所得到的虛擬相關文件是相關的且能從最高排序文件中估測比較好的簡單混合模型 $P_{\mathrm{SMM}}(w|S)$，更明確地說，簡單混合模型是假設虛擬相關文件 $\mathbf{D_P}$ 裡的詞彙 w 是源自於二種成分混合模型(Two-Component Mixture Model)，其一為簡單混合模型 $P_{\mathrm{SMM}}(w|S)$，另一為背景語言模型 $P(w|BG)$。簡單混合模型的估測是藉由期望值最大化(Expectation Maximization, EM)演算法來最大化虛擬相關文件的對數相似度(Log-Likelihood)以進行模型的估測，其虛擬相關文件的對數相似度的定義如下[32]：

$$\begin{aligned}LL_{\mathbf{D}_\mathrm{P}} = {} & \sum_{D_m\in\mathbf{D}_\mathrm{P}}\sum_{w\in V}c(w,D_m)\\ & \cdot\log[(1-\alpha)\cdot P_{\mathrm{SMM}}(w\mid S)+\alpha\cdot P(w\mid BG)],\end{aligned} \tag{12}$$

其中 α 為平衡參數，用來控制模型估測時是要比較偏好簡單混合模型或是背景語言模

型，$c(w,D_m)$為詞彙 w 在虛擬相關文件 D_m 的次數，式(12)的最大化可透過期望值最大化迭代更新式來達成：

期望值步驟：

$$\tau_w^{(l)} = \frac{\alpha \cdot P_{\mathrm{SMM}}^{(l)}(w \mid S)}{\alpha \cdot P_{\mathrm{SMM}}^{(l)}(w \mid S) + (1-\alpha) \cdot P(w \mid BG)}, \tag{13}$$

最大化步驟：

$$P_{\mathbf{SMM}}^{(l+1)}(w \mid S) = \frac{\sum_{D_m \in \mathbf{D_P}} c(w, D_m) \cdot \tau_w^{(l)}}{\sum_{w' \in V} \sum_{D'_m \in \mathbf{D_P}} c(w', D'_m) \cdot \tau_w^{(l)}}, \tag{14}$$

其中 l 表示期望值最大化的第 l 次迭代。這個簡單混合模型的估測會加強具有獨特性(Specificity)的詞彙之機率，例如某詞彙沒有在背景語言模型中有好解釋(Well-Explained)則會被加強其機率，這樣使得此模型為更具有鑑別(Discriminant)能力的語句模型；反之，若是沒有獨特性的詞彙，則其機率就會被背景語言模型所吸收。

4.3、三混合模型

另一方面，本論文嘗試將三混合模型(Tri-Mixture Model)[9]用於語音摘要任務。三混合模型可視為是複雜化後的簡單混合模型；它更進一步的假設虛擬相關文件 $\mathbf{D_P}$ 裡的詞彙 w 是源自於三種成分模型(Component Models)，其一為文件模型 $P(w|D_m)$，其二為三混合模型 $P_{\mathrm{TriMM}}(w|S)$，最後為背景語言模型 $P(w|BG)$。三混合模型的估測也是藉由期望值最大化演算法來最大化虛擬相關文件的對數相似度以進行模型的估測，其虛擬相關文件的對數相似度的定義如下[9]：

$$\begin{aligned} LL_{\mathbf{D}_\mathrm{P}} = \; & \sum_{D_m \in \mathbf{D}_\mathrm{P}} \sum_{w \in V} c(w, D_m) \\ & \cdot \log[(1-\lambda-\mu) \cdot P_{\mathrm{TriMM}}(w \mid S) + \lambda \cdot P(w \mid D_m) + \mu \cdot P(w \mid BG)], \end{aligned} \tag{15}$$

其中 λ 和 μ 為平衡參數，用來控制模型估測時是要比較偏好三混合模型或文件模型亦或是背景語言模型，$c(w,D_m)$為詞彙 w 在虛擬相關文件 D_m 的次數，式(15)的最大化可透過期望值最大化迭代更新式來達成：

期望值步驟：

$$\begin{cases} r_{w,D_m} = \dfrac{c(w, D_m) \cdot (1-\lambda-\mu) \cdot P_{\mathrm{TriMM}}(w \mid S)}{(1-\lambda-\mu) \cdot P_{\mathrm{TriMM}}(w \mid S) + \mu \cdot P(w \mid BG) + \lambda \cdot P(w \mid D_m)} \\ e_{w,D_m} = \dfrac{c(w, D_r) \cdot \lambda \cdot P(w \mid D_m)}{(1-\lambda-\mu) \cdot P_{\mathrm{TriMM}}(w \mid S) + \mu \cdot P(w \mid BG) + \lambda \cdot P(w \mid D_m)} \end{cases}, \tag{16}$$

最大化步驟：

$$\begin{cases} \hat{P}_{\mathbf{TriMM}}(w \mid S) = \sum\limits_{D_m \in \mathbf{D_P}} \dfrac{r_{w,D_m}}{\sum_w r_{w,D_m}} \\ \hat{P}(w \mid D_m) = \dfrac{e_{w,D_m}}{\sum_w e_{w,D_m}} \end{cases}, \tag{17}$$

運用此三混合模型來調適語句模型時，庫爾貝克-萊伯勒離散度量值的公式(參照式(5))可進一步地表示成：

$$KL(D \| S) = \sum_{w \in V} P(w|D) \log \frac{P(w|D)}{\gamma \cdot P(w|S) + (1-\gamma) \cdot P_{\text{TriMM}}(w|S)}, \quad (18)$$

其中$0 \le \gamma < 1$，當$\gamma = 0$代表使用三混合模型取代原本的語句模型。

關聯模型、簡單混合模型及三混合模型在資訊檢索領域中已被廣泛應用[12][32][9]，但在摘要任務中卻是相對較少研究的，值得一提的是，雖然關聯模型、簡單混合模型已初步被應用在摘要任務上[4][19]，但三混合模型卻是本論文首次引入到(語音)文字摘要任務中。

五、實驗語料及評估方法

5.1、實驗語料

本論文實驗語料庫為公視新聞語料(Mandarin Chinese Broadcast News Corpus, MATBN)，是由中央研究院資訊科學研究所耗時三年與公共電視台合作錄製並整理的中文新聞語料，其錄製內容為每天一個小時的公視晚間新聞深度報導。我們抽取其中由2001 年 11 月到 2002 年 8 月總共 205 則新聞報導，區分成訓練集(共 185 則新聞)以及測試集(共 20 則新聞)兩部分，其詳細的統計資訊如表一所示。全部 205 則語音文件長度約為 7.5 小時，我們先做人工切音，切出真正含有講話內容的音訊段落，再經由語音辨識器自動產生出的語音辨識結果稱之為語音文件(Spoken Document, SD)，因此語音文件中只包含有語音辨識錯誤之雜訊；另一方面，我們將此 205 則語音文件藉由人工聽寫的方式，產生出沒有辨識錯誤的正確文字語料，我們稱之為文字文件(Text Document, TD)，每則文字文件再經由三位專家標記摘要語句，我們將此標記的人工摘要做為語音文件與文字文件的正確摘要答案。藉由比較語音文件和文字文件的摘要效能，我們可以觀察語音辨識錯誤對於各種摘要方法之影響。本研究的背景語言模型訓練語料取材自2001 到 2002 年的中央社新聞文字語料(Central News Agency, CNA)，並且以 SRI 語言模型工具訓練出經平滑化的單連語言模型，我們假設此單連語言模型為明確度中的非相關資訊之來源。另外，本論文蒐集 2002 年中央通訊社的約十萬則同時期新聞文字文件做為建立關聯模型時的檢索標的[4]，關於語句 S 的虛擬相關文件(最高排序文件)篇數為20(也就是$|\mathbf{D}_{\text{Top}}|$=20)，而經由各種不同虛擬相關文件選取方法的篇數為 3(亦即$|\mathbf{D}_{\text{P}}|$=3)。

5.2、評估方法

自動摘要的評估方法主要有兩種，一為主觀人為評估，另一為客觀自動評估；前者為請幾位測試人員來為系統所產生的摘要做評估，給分的範圍為 1-5 分，後者則是預先請幾位測試者依據事先定義好的摘要比例挑選出適合的摘要語句，系統所產生的摘要句子將與測試者所挑選出的句子計算召回率導向的要點評估(Recall-Oriented Understudy for Gisting Evaluation, ROUGE)[16]。由於主觀人為評估非常耗時耗力，所以目前多數自動摘要方法皆採用召回率導向的要點評估做為文件摘要的評估方式，本論文亦採用此種評估方式。ROUGE 方法是計算自動摘要結果與人工摘要之間的重疊單位元(Units)數目占參考摘要(Reference Summary)長度(單位元總個數)的比例。估計的單位可以是 N-連詞(N-gram)、詞序列(Word Sequences)，如：最長相同詞序列或詞成對(Word Pairs)。由於此方法是採用單位元比對的方式，不會產生語句邊界定義的問題，並且適合於多份人工摘要的評估。其評估的分數有三種，ROUGE-1(Unigram，簡寫為 R-1)、

表一、實驗語料統計資訊

	訓練集	測試集
語料時間	2001/11/07-2002/01/22	2002/01/23-2002/08/22
文件個數	185	20
文件平均持續幾秒	129.4	141.2
文件平均詞個數	326.0	290.3
文件平均語句個數	20.0	23.3
文件平均字錯誤率 (Character Error Rate, CER)	28.8%	29.8%
文件平均詞錯誤率 (Word Error Rate, WER)	38.0%	39.4%

ROUGE-2(Bigram，簡寫為 R-2)和 ROUGE-L(Longest Common Subsequence，簡寫為 R-L)分數，ROUGE-1 是評估自動摘要的訊息量，ROUGE-2 是評估自動摘要的流暢性，ROUGE-L 是最長共同字串，本論文希望觀察摘要的流暢性，因此，實驗數據主要是以 ROUGE-2 分數為主。本論文所設定的摘要比例為 10%，其定義為摘要所含詞彙數占整篇文件詞彙數的比例，也就是以詞彙做為判斷摘要比例的單元。在挑選摘要語句過程中，若選到某語句中的某個詞彙時就已經剛好達到摘要比例，為了保持語句語意完整性，此語句剩下的詞彙也會被挑選成為摘要。

六、實驗結果

本論文主要著重在虛擬相關文件選取方法之發展與改進，是屬於非監督式摘要方法的範疇，因此所比較的對象以非監督式摘要方法為主；除此之外，本論文亦嘗試與現今最被廣為使用的監督式機器學習方法做比較，即支持向量機(SVM)[10]。

6.1、基礎實驗

首先，我們比較庫爾貝克-萊伯勒離散度(KL)與數個非監督式摘要方法之摘要成效，包含有最長語句摘要(Longest Sentence, LS)、首句摘要(LEAD)[26]、向量空間模型(Vector Space Model, VSM)[8]、潛藏語意分析(Latent Semantic Analysis, LSA)[8]、最大邊際關聯(Maximal Marginal Relevance, MMR)[2]以及整數線性規劃(Integer Linear Programming, ILP)[22]。一般來說，文件中長句可能蘊含有較豐富的主題資訊，因此依據文件中語句長度做排序後，依序選取最長語句做為摘要結果是一種簡單的摘要方法。除此之外，也有學者研究發現，文件常以開門見山法的方式來提點出主題，因此文件開頭的前幾個語句經常是具代表性的語句，首句摘要即是以此概念出發，選取前幾句語句來形成整個文件的摘要。最長語句摘要(LS)及首句摘要(LEAD)都僅適用在一部分具有特殊結構的文件上，因此它們的缺點就是有其侷限性。另外，向量空間模型是把文件和語句分別視為一個向量，並使用詞頻-反文件頻(TF-IDF)特徵來計算每一維度的權重值，文件與語句間的關聯性是藉由餘弦相似度量值來估測，當語句分數較高時，則越有機會成為此文件的摘要。潛藏語意分析是在向量空間的假設下更進一步地使用奇異值分解(Singular Value Decomposition, SVD)來找到可能的潛藏語意空間，使之能在考量潛藏語意的情況下進行文件與語句的關聯性量測。最大邊際關聯可視為是向量空間模型的一個延伸，在做語句排序時考量了冗餘性以達到更好的摘要結果。整數線性規劃是一個全域(Global)

表二、基礎實驗結果

		F-score		
		ROUGE-1	ROUGE-2	ROUGE-L
TD	LS	0.225	0.098	0.183
	LEAD	0.310	0.194	0.276
	VSM	0.347	0.228	0.290
	LSA	0.362	0.233	0.316
	MMR	0.368	0.248	0.322
	KL	0.411	0.298	0.361
	ILP	**0.442**	**0.337**	**0.401**
SD	LS	0.181	0.044	0.138
	LEAD	0.255	0.117	0.221
	VSM	0.342	0.189	0.287
	LSA	0.345	0.201	0.301
	MMR	**0.366**	**0.215**	**0.315**
	KL	0.364	0.210	0.307
	ILP	0.348	0.209	0.306

的限制性最佳化(Constraint Optimization)的語句選取方法[22]。

表二為本論文之基礎實驗結果。首先，在 TD 的實驗中，KL 的摘要效果比 LS、LEAD、VSM、LSA、MMR 等非監督式摘要方法來得好些；因 LS 與 LEAD 僅適用於特殊文章結構上，所以若被摘要文件不具有某種特殊的文章結構，其摘要效能就會有限。相較之下，KL 是較具一般性的摘要方法，因此比較不會受限於文章的結構之影響，故摘要效能比 LS 以及 LEAD 來得彰顯。KL 與 VSM 皆使用淺層的詞彙(詞頻)資訊，但由於 KL 是計算語句模型與文件模型之間的距離關係，對於代表語句與文件的語言模型，我們較容易透過各種技術來進行模型的估計與調適，進而獲得較好的摘要成果。MMR 在選取時多考慮了冗餘資訊，所以摘要效果也比 VSM 來得好些。LSA 在潛藏語意空間計算文件與語句的餘弦相似度量值，其結果顯示也會較 VSM 好。整數線性規劃是一個全域選擇方法，所以在 TD 上可以得到最好的摘要效能。另一方面，在 SD 的實驗中，KL 同樣較優於 LS、LEAD 之摘要方法，但 VSM 的結果則稍微較 KL 好一點，我們認為這可能是因為 VSM 比較不受到語音辨認錯誤的影響。原以為 ILP 也會在 SD 中得到最好的摘要效能，結果反而是 MMR 得到最好的摘要效能，可能的原因是 ILP 受到語音辨識錯誤的影響比較大，造成其摘要結果不彰。

通常語音文件主要會有語音辨識錯誤和語句邊界偵測錯誤的問題，但我們有先經人工切音，因此摒除了語句邊界偵測錯誤的問題，藉由比較 TD 與 SD 之實驗結果，我們可以觀察語音辨識錯誤率對摘要結果的影響性。比較各式方法，SD 比 TD 下降了 1.9%~8.8%的 ROUGE-2 摘要效能，由此可知語音辨識錯誤率對摘要效能是有顯著的影響性。為了減緩語音辨認錯誤的問題，在未來我們將嘗試使用音節(Syllable)為單位來建立語句以及文件模型；或利用詞圖(Word Graph)、混淆網路(Confusion Network)來含括更多的可能正確候選詞彙以裨益模型估測；更可利用韻律資訊(Prosodic Information)等聲學線索來輔助減緩語音辨認錯誤對摘要效能的影響。

表三、關聯模型之實驗結果(使用最高排序文件前三篇(Top3))

		F-score		
		ROUGE-1	ROUGE-2	ROUGE-L
TD	KL	0.411	0.298	0.361
	RM	0.450	0.336	0.400
	SMM	0.436	0.325	0.385
	TriMM	**0.457**	**0.350**	**0.404**
SD	KL	0.364	0.210	0.307
	RM	0.374	0.226	0.321
	SMM	0.375	0.221	0.314
	TriMM	**0.379**	**0.228**	**0.325**

6.2、基礎關聯模型之實驗

使用關聯模型於語句模型之建立時，需要做一次的資訊檢索來為每個語句找出虛擬相關文件，由同時期的新聞文字文件(共 101,268 篇)中為每一語句選取出 20 篇虛擬相關文件，但為了要與後續虛擬相關文件選取方法作公平的比較，因此此基礎關聯模型實驗是取前三篇(Top3)來進行關聯模型之估測與相關實驗[4]。由於文件中的語句通常相對簡短，因此當使用最大化相似度估測建立語句模型時，容易遭遇資料稀疏的問題，不容易獲得精準的模型，故我們期望考慮額外的關聯資訊於語音文件摘要，亦即藉由虛擬相關文件來重新估測並建立語句的語言模型，能獲得進一步地摘要成效。重新估測後的關聯模型則可與原本的語句模型相結合或取代之，相結合的參數調整在本實驗中是採用經驗設定(Empirical Setting)。實驗結果如表三所示，在 TD 與 SD 之摘要成效上，使用關聯模型(RM)、簡單混合模型(SMM)及三混合模型(TriMM)皆能比基礎的 KL 實驗較好，尤其是三混合模型(TriMM)相較於 KL 在 TD 及 SD 的 ROUGE-2 結果上能有 5.2%與 1.8%的改進。接著，我們比較不同關聯模型的摘要成效，首先是關聯模型(RM)與簡單混合模型(SMM)的比較，從表三的實驗結果得知關聯模型在 TD 上表現比簡單混合模型來得好，但在 SD 似乎在 ROUGE-1 就沒比簡單混合模型好，不過 SD 的 ROUGE-2 跟 ROUGE-L 都還是比簡單混合模型的效果好。關聯模型的假設是強調詞彙 w 與語句 S 在這些虛擬相關文件中同時出現之關係(參照式(10))來估測模型，而簡單混合模型是強調訓練好的模型能讓有獨特性的詞彙得到更多的機率值因而讓模型具有鑑別能力，兩者皆有其好處。最後，三混合模型(TriMM)因複雜化了簡單混合模型(SMM)，額外多考量文件模型的影響力，因此相較於關聯模型及簡單混合模型能得到更佳的摘要效能，三混合模型相較於關聯模型在 TD 上有明顯的進步，於 ROUGE-2 結果能有 1.4%的改進，但在 SD 上，於 ROUGE-2 結果只有微量的 0.2%改善。

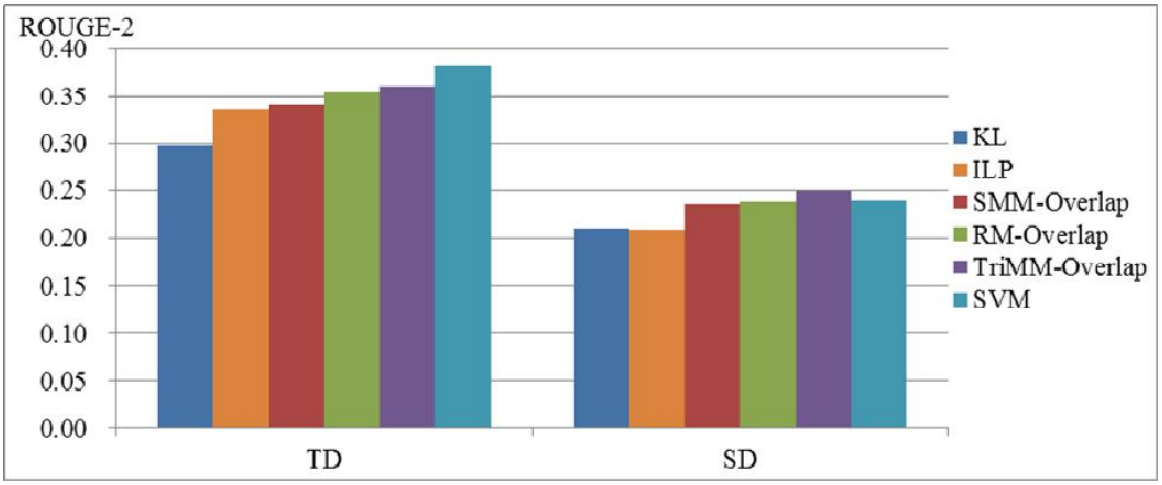
在關聯模型的相關實驗中，語音辨識錯誤也是影響摘要效能非常嚴重，在三混合模型的數據中，SD 比 TD 劇烈下降了 12.2%的 ROUGE-2 摘要效能，在未來研究中，我們認為可以以次詞索引(Subword Indexing)的方式來建立關聯模型以減緩語音辨識錯誤之影響。

表四、各種虛擬相關文件選取方法於關聯模型之實驗結果

F-score		RM			SMM			TriMM		
		R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
TD	Top3	0.450	0.336	0.400	0.436	0.325	0.385	0.457	0.350	0.404
	Gapped *K*	0.451	0.338	0.401	0.433	0.317	0.385	0.454	0.343	0.406
	Centroid	0.449	0.334	0.402	0.439	0.331	0.389	0.456	0.353	0.407
	Active-RDD	0.460	0.341	0.408	0.449	0.342	0.400	0.463	0.355	0.414
	Active-RDDN	0.464	0.352	0.411	0.455	**0.346**	0.405	0.466	**0.367**	0.421
	Overlapped	**0.470**	**0.354**	**0.416**	**0.460**	0.341	**0.410**	**0.471**	0.362	**0.422**
SD	Top3	0.374	0.226	0.321	0.375	0.221	0.314	0.379	0.228	0.325
	Gapped *K*	0.374	0.228	0.322	0.371	0.218	0.313	0.376	0.225	0.315
	Centroid	0.374	0.227	0.314	0.377	0.227	0.320	0.380	0.233	0.328
	Active-RDD	0.379	0.228	0.332	0.378	0.229	0.321	0.388	0.242	0.335
	Active-RDDN	0.383	0.239	0.330	0.380	0.226	0.327	0.391	0.244	0.339
	Overlapped	**0.386**	**0.239**	**0.334**	**0.382**	**0.236**	**0.332**	**0.396**	**0.250**	**0.345**

6.3、各種虛擬相關文件選取方法於關聯模型之實驗

本小節的實驗是由第一次虛擬相關文件(最高排序文件) $\mathbf{D_{Top}}$ 中($|\mathbf{D_{top}}|$=20)再精鍊選取出較佳的虛擬相關文件 $\mathbf{D_P}$ ($|\mathbf{D_P}|$=3)，我們比較所提出兩種新穎的選取方法(即主動式-關聯多元密度非相關(Active-RDDN)和重疊分群(Overlapped))與其他現有選取方法(即間隔式最高 *K* 選取法(Gapped *K*)、群中心選取法(Centroid)以及主動式-關聯多元密度(Active-RDD))於各種關聯模型(RM、SMM 及 TriMM)之摘要效能比較，實驗結果如表四所示，與基礎關聯模型只使用前三篇(Top3)虛擬相關文件的結果相較(參照表三)，大部分透過虛擬相關文件選取方法都會比只使用 Top3 的摘要結果還要來得好，除了 Gapped *K* 無論在 TD 和 SD 中，使用 SMM 與 TriMM 都會有比 Top3 差的摘要效能，因為 Gapped *K* 是一個較不穩定的虛擬相關文件選取方法，在本實驗中有比較差的結果是可以預期的。群中心選取法(Centroid)表現尚可，在 TD 及 SD 中，於各種關聯模型(RM、SMM 及 TriMM)下，比 Top3 及 Gapped *K* 都要來得好。Active-RDD 因在選取虛擬相關文件時同時考量了關聯性(Relevance)、多元性(Diversity)以及密度性(Density)，用於不同的關聯模型訓練時，相對於 Top3、Gapped *K* 以及 Centroid，無論在 TD 或 SD 中，都可以得到更好的摘要結果。Active-RDDN 在多考量了非相關(Non-relevance)資訊的情況下，其實驗結果都會比現有的選取方法較佳，相對於 Active-RDD、Centroid、Gapped *K* 以及 Top3 無論在 TD 或 SD 中，各種關聯模型(RM、SMM 及 TriMM)下都會得到比較好的摘要結果，所以證實非相關資訊確實一個有用的選取線索。最後本論文所提出的重疊分群(Overlapped)選取方法無論在 TD 或 SD 中，於各種關聯模型下(RM、SMM 及 TriMM)都可以得到最好的摘要效果，驗證了重疊分群在利用虛擬相關文件中結構化資訊確實可以找到具代表性的文件以利各種關聯模型的模型訓練或參數估測。



圖二、SVM 與其他非監督式摘要方法之比較

6.4、與監督式模型之比較

除了各式非監督式摘要方法外，本論文亦嘗試比較支持向量機(SVM)於文件摘要之成效，比較的對象有基礎 KL 以及使用重疊分群選取方法於不同的關聯模型中(RM-Overlap、SMM-Overlap 和 TriMM-Overlap)。支持向量機是現今常見的監督式機器學習方法之一，近年來已有學者將其運用至文件摘要領域之中[10]。本論文使用訓練集的 185 篇文件進行支持向量機模型的訓練語料，我們為文件中的每一語句抽取 35 維特徵[13]，包括有韻律特徵(Prosodic Features)、語彙特徵(Lexical Features)、結構特徵(Structural Features)以及基本的模型特徵(Model Features)等資訊，其核心函數設定為半徑式函數(Radial Basis Function)，其中 SVM 的參數設定都是使用預設值。

實驗結果如圖二所示。一如預期地，SVM 與其他各式非監督式模型相比較，在 TD 實驗上(其 ROUGE-2 為 0.383)是表現最好的方法，這是由於監督式機器學習藉由使用人工標注的摘要句子進行模型之訓練，其使用的資訊較非監督式機器學習方法多且正確，因此其摘要的效果也較非監督式機器學習來的好。值得一提的是，使用重疊分群虛擬相關文件選取方法於三混合模型中(TriMM-Overlap)，摘要之成效在 SD 上可比監督式機器學習方法的 SVM 來的好一些，此一實驗結果令人感到驚訝，因為本論文所探討之各式摘要方法僅考慮了文件與語句中的單一種特徵值，即藉由詞彙分佈資訊來挑選語句，而支持向量機不僅使用了 35 種特徵值，更需要使用人工標註的正確答案進行模型的訓練。我們認為，此結果之原因可能是由於支持向量機之摘要技術在語音辨識錯誤的情況下（在此實驗中，訓練集與測試集的詞錯誤率達 40%），未必能真的有效學習分辨摘要與非摘要語句。

七、結論與未來方向

本論文基於語言模型化架構來發展語音摘要方法，其貢獻主要有三方面。第一，有別於現有基於語言模型化架構之摘要方法都聚焦在語句模型參數的重新估測，本論文首次深入探討與應用各種新穎虛擬相關文件選取技術於節錄式語音文件摘要任務，用以強化語句模型的參數估測。第二，本論文更進一步地考量使用每一語句的非相關性(Non-relevance)資訊對於虛擬相關文件選取的影響。同時，我們亦額外嘗試基於重疊分群(Overlapped Clustering)概念來有效地選取重要的虛擬相關文件。第三，本論文探索使用三混合模型(Tri-Mixture Model)來表示每一語句，期盼其能更精確地表示語句之詞彙使用和語意相關資訊。一系列的實驗結果顯示，本論文所提出之方法的確能較其它現有

的非監督式摘要方法有更加的摘要效能表現。

未來，我們的研究將有三個主要的方向：首先，本論文所提出之虛擬相關文件選取方法是建構在向量空間或語言模型空間上，並沒有考慮到語意空間的相似度量，我們將進一步的研究是否可以在潛藏語意空間中來選取較好的虛擬相關文件，以期獲得更好的摘要成效；其次，目前所發展的關聯模型僅運用於重建語句的語言模型，我們將嘗試使用被摘要文件的關聯資訊來重新估測並建立文件的語言模型；最後，我們希望能將非監督式方法所形成的特徵結合於更加複雜且有效的監督式機器學習方法(如 CRF 或深度類神經網絡(Deep Neural Network Learning, DNN)等)中，並融合其它語音文件所獨有之特徵(諸如音韻與語者特徵等)，期望訓練後的模型能夠在語音文件摘要上獲得更好的表現。

致謝

本論文之研究承蒙教育部–國立臺灣師範大學邁向頂尖大學計畫(102J1A0800)與行政院科技部研究計畫(MOST 103-2221-E-003-016-MY2, NSC 103-2911-I-003-301, NSC 101-2221-E-003-024-MY3 、 NSC 101-2511-S-003-057-MY3 、 NSC 101-2511-S-003-047-MY3 和 NSC 102-2221-E-003-014- MY3)之經費支持，謹此致謝。

參考文獻

[1] P. Baxendale, *Machine-made Index for Technical Literature – an Experiment,* IBM Journal of Research and Development, Vol. 2, No. 4, pp. 354–361, 1958

[2] J. Carbonell and J. Goldstein, *The Use of MMR Diversity-based Reranking for Reordering Documents and Producing Summaries,* Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 335–336, 1998

[3] Y.-T. Chen, B. Chen and H.-M. Wang, *A Probabilistic Generative Framework for Extractive Broadcast News Speech Summarization,* IEEE Transactions on Audio, Speech and Language Processing, Vol. 17, No. 1, pp. 95–106, 2009

[4] B. Chen, H.-C. Chang, K.-Y. Chen, *Sentence Modeling for Extractive Speech Summarization*, Proceedings of the International Conference on Multimedia & Expo (ICME), 2013

[5] B. Chen, Y.-W. Chen and K.-Y Chen, Enhancing Query Formulation for Spoken Document Retrieval, Journal of Information Science and Engineering, Vol. 30, No. 3, pp. 553–569, 2014

[6] J.-M. Conroy and D.-P. O'Leary, *Text Summarization via Hidden Markov Models,* Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 406–407, 2001

[7] G. Erkan and D. R. Radev, *LexRank: Graph-based Lexical Centrality as Salience in Text Summarization*, Journal of Artificial Intelligent Research, Vol. 22, No. 1, pp.457–479, 2004

[8] Y. Gong and X. Liu, *Generic Text Summarization using Relevance Measure and Latent Semantic Analysis,* Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 19–25, 2001

[9] D. Hiemstra, S. Robertson, and H. Zaragoza, *Parsimonious Language Models for Information Retrieval*, Proceedings of the international ACM SIGIR conference on Research and development in information retrieval (SIGIR), pp. 178–185, 2004

[10] A. Kolcz, V. Prabakarmurthi and J. Kalita, *Summarization as Feature Selection for Text Categorization*, Proceedings of the International Conference on Information and Knowledge Management (CIKM), pp. 365–370, 2001

[11] J. Kupiec, *A Trainable Document Summarizer,* Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 68–73, 1995

[12] V. Lavrenko and W.-B. Croft, *Relevance -based Language Models,* Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 120–127, 2001

[13] S.-H. Lin and B. Chen, *Improved Speech Summarization with Multiple-hypothesis Representations and Kullback-Leibler Divergence Measures,* Proceeding of the 10th Annual Conference of the International Speech Communication Association (Interspeech), pp. 1847–1850, 2009

[14] H. Lin and J. Bilmes, *Multi-document Summarization via Budgeted Maximization of Submodular Functions*, Proceeding of NAACL HLT, pp. 912–920, 2010

[15] S.-H. Lin, Y.-M. Yeh and B. Chen, *Leveraging Kullback-Leibler Divergence Measures and Information-rich Cues for Speech Summarization,* IEEE Transactions on Audio, Speech and Language Processing. Vol. 19, No. 4, pp. 871–882, 2011

[16] C.-Y. Lin, *ROUGE: Recall-oriented Understudy for Gisting Evaluation.* 2003 [Online]. Available: http://haydn.isi.edu/ROUGE/.

[17] Y. Liu and D. Hakkani-Tur, *Speech Summarization*, in G. Turand R. D. Mori [Ed], Spoken Language Understanding: Systems for Extracting Semantic Information from Speech, Wiley, 2011

[18] X. Liu and W. B. Croft, *Cluster-based Retrieval Using Language Models*, Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR), pp. 186–193, 2004

[19] S.-H. Liu, K.-Y. Chen, Y.-L. Hsieh, B. Chen, H.-M. Wang, H.-C. Yen, W.-L. Hsu, *Effective Pseudo-relevance Feedback for Language Modeling in Extractive Speech Summarization*, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2014

[20] S.-H. Liu, K.-Y. Chen, B. Chen, H.-M. Wang, W.-L. Hsu, *Improving Sentence Modeling Techniques for Extractive Speech Summarization*, ROCLING XXV: Conference on Computational Linguistics and Speech Processing, 2013

[21] I. Mani and M.-T. Maybury, *Advances in Automatic Text Summarization,* Cambridge: MIT Press, 1999

[22] R. McDonald, *A Study of Global Inference Algorithms in Multi-document Summarization*, Proceedings of European Conference on Information Retrieval (ECIR), pp. 557–564, 2007.

[23] R. Mihalcea and P. Tarau, *TextRank Bringing Order into Texts,* Proceedings of Empirical Method in Natural Language Processing (EMNLP), pp. 404–411, 2004

[24] G. Murray, S. Renals, and J. Carletta, *Extractive Summarization of Meeting Recordings,* Proceedings of the Conference of the International Speech Communication Association (Interspeech), pp. 593–596, 2005

[25] A. Nenkova and K. McKeown, *Automatic Summarization*, Foundations and Trends in Information Retrieval, Vol. 5, No. 2–3: 103–233, 2011

[26] G. Penn and X. Zhu, *A Critical Reassessment of Evaluation Baselines for Speech Summarization,* Proceedings of Annual Meeting of the Association for Computational Linguistics, pp. 470–478, 2008

[27] X. Shen and C. Zhai, *Active Feedback in Ad Hoc Information Retrieval*, Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 55−66, 2005

[28] C. Shen and T. Li, *Multi-document Summarization via the Minimum Dominating Set*, Proceedings of the International Conference on Computational Linguistics (COLING), pp. 984–92, 2010

[29] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen, *Document Summarization using Conditional Random Fields,* Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), pp. 2862–2867, 2007

[30] X. Wan and J. Yang, *Multi-document Summarization using Cluster-based Link Analysis,* Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 299–306, 2008

[31] Z. Xu, R. Akella and Y. Zhang, *Incorporating Diversity and Density in Active Learning for Relevance Feedback*, Proceedings of European Conference on Information Retrieval (ECIR), pp. 245–257, 2007

[32] C.-X. Zhai and J. Lafferty, *Model-based feedback in the language modeling approach to information retrieval*, Proceeding of the International Conference on Information and Knowledge Management (CIKM), pp. 403–410, 2001

[33] C.-X. Zhai and J. Lafferty, *A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval,* Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 334-342, 2011

[34] C.-X. Zhai, *Statistical Language Models for Information Retrieval: A Critical Review,* Foundations and Trends in Information Retrieval, Vol. 2, No.3, pp.137–213, 2008

[35] J. Zhang and P. Fung, *Speech Summarization without Lexical Features for Mandarin Broadcast News,* Proceedings of NAACL HLT, Companion Volume, pp. 213–216, 2007