Computational Linguistics and Chinese Language Processing
Vol. 22, No. 1, June 2017, pp. 1-26

當代非監督式方法之比較於節錄式語音摘要

An Empirical Comparison of Contemporary Unsupervised Approaches for Extractive Speech Summarization

劉士弘*、陳冠宇*、施凱文+、陳柏琳+、王新民*、許聞廉*

Shih-Hung Liu, Kuan-Yu Chen, Kai-Wun Shih, Berlin Chen,

Hsin-Min Wang, and Wen-Lian Hsu

摘要

由於網際網路的飛速發展，促成大資料時代的來臨，也因此自動摘要(Automatic Summarization)成為近年來一項熱門的研究議題。節錄式(Extractive)自動摘要是依據事先定義的摘要比例，從文字文件(Text Documents)或語音文件(Spoken Documents)中選取一些能夠代表原始文件主旨或主題的重要語句當作摘要。節錄式摘要可被視為一個資訊檢索(Information Retrieval, IR)的問題，在相關研究中，使用語言模型(Language Modeling)來挑選重要語句之方法，已初步地被驗證在文字與語音文件的自動摘要任務上有不錯的成果。本論文延續此項研究，進一步地提出三個主要的研究貢獻。首先，有鑑於關聯性(Relevance)資訊的概念在資訊檢索領域中已有不錯的發展成果，本論文嘗試結合關聯性資訊來重新估測並建立語句的語言模型，並嘗試使用三混合(Tri-Mixture Model, TriMM)模型，期待得以更精準地描述語句的語意內容，進而提升自動摘要之效能。第二，除了語言模型之外，本論文進一步地嘗試探究機率式檢索模型於語音文件

* 中央研究院資訊所
Institute of Information Science, Academia Sinica
E-mail: journey0621@gmail.com, {kychen, whm, hsu}@iis.sinica.edu.tw
+ 國立臺灣師範大學資訊工程系
Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: {60247065S, berlin}@csie.ntnu.edu.tw

摘要任務上之成效。最後，本論文亦探討不同的語言模型平滑化技術對於語音文件摘要任務之影響。本論文的語音文件摘要實驗語料是採用公視廣播新聞(MATBN)；實驗結果顯示，相較於其它現有的非監督式摘要方法，我們所應用的新穎式摘要方法能提供明顯的效能改善。

關鍵字: 最佳匹配、語言模型、虛擬關聯回饋、關聯模型、節錄式自動摘要。

Abstract

Due to the rapid-developed Internet and with the big data era coming, the automatic summarization research has been emerged a popular research topic. The aim of automatic summarization is in attempt to select important text or spoken sentence to represent the topic (theme) of original text or spoken document according to a predefined summarization ratio. In this study we frame automatic summarizaiton task as an ad-hoc information retrieval (IR) problem and employ the mathematical sound language modeling (LM) framework for extractive speech summarization, which can perform important sentence selection in an unsupervised manner and has shown its preliminary success. The main contribution of this paper is three-fold. First, by the virtue of relevance modeling, we explore several effective sentence modeling formulations to enhance the sentence models involved in the LM-based summarization framework and the first use of tri-mixture model to improve the performance of extractive speech summarization. Second, since the language modeling will suffer from data sparseness problem and the common solution is to adopt smoothing techniques, in this research we investigate three different smoothing approaches to evaluate how they influence the summarization performance. Third, we further apply the well-studied ranking model (BM25) and also its variants in IR community for ranking important sentence in extractive speech summarization. Experiments conducted on public avaiable dataset (MATBN) and the results show that our applied methods have effective summarization performance when compared to the other well-practiced and state-of-the-art unsupervised methods.

Keywords: BM25, Language Modeling, Pseudo-Relevance Feedback, Relevance Modeling, Extractive Automatic Summarization.

1. 緒論 (Introduction)

隨著大資料時代的來臨，眾多的文字及多媒體影音資訊被快速地傳遞並分享於全球各地，資訊超載(Information Overload)的問題也隨之產生。如何能讓人們快速且有效率地瀏覽與日俱增的文字資訊或多媒體影音資訊，已成為一個刻不容緩的研究課題。在眾多的研究

方法中，自動摘要(Automatic Summarization)被視為是一項不可或缺的關鍵技術(Lin & Chen, 2010; Liu & Hakkani-Tur, 2011)。自動摘要之目的在於擷取單一文件(Single-Document)或多重文件(Multi-Document)中的重要語意與主題資訊，藉此讓使用者能更有效率地瀏覽與理解文件的主旨，以便快速地獲得其所需的資訊，避免花費大量時間在審視文件內容。另一方面，語音是多媒體文件中最具資訊的成分之一；如何透過語音(文件)摘要技術來自動地、有效率地處理具時序性的多媒體影音內容，例如：電視新聞、廣播新聞、郵件、電子郵件、會議及演講錄音等(Ostendorf, 2008; Nenkova & McKeown, 2011)，更是顯得非常重要。其關鍵原因在於多媒體影音內容往往長達數分鐘或數小時，使用者不易於瀏覽和查詢，而必須耐心地閱讀或聽完整份多媒體影音內容，才能理解其中所描述的語意與主題，這違反人們講求方便、有效率的資訊獲取方式。

雖然對於含有語音訊號的多媒體影音，我們可透過自動語音辨識(Automatic Speech Recognition, ASR)技術自動地將其轉換成易於瀏覽的文字內容，再藉由文字文件摘要的技術來做處理，以達到摘要多媒體影音或其它語音文件之目的。但就現階段語音辨識技術的發展，語音文件經語音辨識後自動轉寫成文字的結果，不僅存在辨識錯誤的問題，也缺乏章節與標點符號，使得語句邊界定義不清楚而失去文件的結構資訊；除此之外，語音文件通常含有許多口語語助詞、遲疑、重覆等內容，這都使得語音文件摘要技術的發展面臨更多的挑戰。

一般來說，自動摘要研究可從許多不同面相來進行探討，包括了來源、需求、方式、用途以及模型技術，以下將簡述各個不同面相的相關議題(Mani & Maybury, 1999)：

(1) 來源：根據文件來源，可以分為單一文件摘要與多重文件摘要(Cai & Li, 2013)；單一文件摘要是依據事先定義好的摘要比例，選取能夠代表文件的句子當作摘要；而多重文件摘要是收集多篇相似的文件，需要移除文件間彼此冗餘性(Redundancy)的資訊(Carbonell & Goldstein, 1998)，考慮文件描述事件發生的先後順序(Causality)(Kuo & Chen, 2006)，並且確認文件之間的因果關係，經由這些資訊希望能產生有連貫性的文件摘要。

(2) 需求：依據使用者需求不同，摘要內容可區分為具有資訊性(Informative)、指示性(Indicative)、以及評論性(Critical)。具有資訊性的摘要是用來表達文件描述的主旨內容與核心資訊；具指示性的摘要是希望將文件中的主題內容做簡單的描述，並將文件分成不同的主題，例如：政治性、學術性、體育性和娛樂性文件，因此所產生的摘要不要求傳達詳細的原始文件內容；具評論性的摘要提供文件正面與反面的觀點(Positive and Negative Sentiments)(Galley, McKeown, Hirschberg & Shriberg, 2004)。

(3) 方式：可概分為二大類，節錄式(Extractive)摘要與抽象式(Abstractive)摘要(或重寫式摘要)。前者主要是依據特定的摘要比例，從最原始的文件中選取重要的語句來組成摘要；而後者是在完全理解文件內容之後，重新撰寫產生摘要來代表原始文件的內容，其所使用之語彙或慣用語不一定是全然地來自於原始文件，此種摘要方式是最為貼近人們日常撰寫摘要的形式。然而抽象式摘要需要複雜的自然語言處理(Natural Language Processing, NLP)技術，如資訊擷取(Information Extraction)、對話理解(Discourse Understanding)及自然語言

生成(Natural Language Generation)等(Paice, 1990; Witbrock & Mittal, 1999)，因此，近年來節錄式摘要之研究仍為主流。

(4) 用途：依摘要用途可分為一般性(Generic)摘要與以查詢為基礎(Query-focused)的摘要。前者是從整篇文件中萃取出能夠突顯整篇文件全面性主題資訊的語句，期望摘要產生的內容可以涵蓋整篇文件所有重要的主題；後者透過使用者或特定的查詢來產生與查詢相關的摘要。

(5) 模型技術：簡單分成三大類，(i)以簡單的語彙(Lexical)與結構(Structural)特徵做為判斷摘要語句的模型技術(Zhang, Chan & Fung, 2010)，(ii)監督式機器學習(Supervised Machine Learning)以及(iii)非監督式機器學習(Unsupervised Machine Learning)(Liu & Hakkani-Tur, 2011)之模型技術。雖然非監督式機器學習的方法在一般的情況下其效能沒有監督式機器學習方法來的好，但非監督式機器學習方法不需要事先準備大量人工標記的訓練資料，以及具有容易實作(Easy-to-Implement)的特性，仍吸引許多學者進行研究與發展，本論文主要也是探討且比較非監督式機器學習的方式於自動摘要之任務。

綜觀上述各個面向，本論文主要探究一般性、單一文件節錄式語音摘要問題，並比較各式非監督式機器學習模型技術。近年來，各式基於語言模型之非監督式模型技術運用在資訊檢索領域中已呈現卓越的研究成果(Zhai, 2008)，這些技術也初步地被應用於語音文件摘要之研究上(Lin, Yeh & Chen, 2011)，亦獲得一定的摘要成效。本論文將延續此一研究主軸，提出三個主要的研究貢獻。首先，有鑑於關聯性(Relevance)資訊的概念已被應用於資訊檢索領域之中(Zhai & Lafferty, 2001a; Lavrenko & Croft, 2001)，本論文嘗試結合關聯性資訊來重新估測並建立語句的語言模型，並首次使用三混合(Tri-Mixture Model, TriMM)模型，期待得以更精準地估測語句的語意內容，增進自動摘要之成效。當語言模型使用最大化相似度估測時，很可能會遭遇資料稀疏(Data Sparseness)的問題，而使得模型無法準確地估測每一個詞彙真正的機率分佈，也可能因為某些詞彙的條件機率值為零，導致無法準確地計算語句與文件間的相似度。為此，語言模型平滑化技術被提出來減輕上述的現象。過去的研究中顯示，各式平滑化技術的使用時常在各項任務中扮演關鍵的腳色。有鑑於此，本論文首次比較不同平滑化技術對於語音文件摘要任務之影響。最後，除了語言模型的探討之外，我們進一步地提出並使用多種機率式檢索模型於語音摘要任務上。本論文後續安排如下：第二章扼要地介紹現今自動摘要模型技術的相關研究與發展；第三章介紹使用語言模型於節錄式語音摘要任務之原理，然後闡述如何藉助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；第四章介紹多種機率式排序模型並將之應用至語音文件摘要任務中；第五章介紹實驗語料與設定以及摘要評估之方法；第六章說明實驗結果及其分析；最後，第七章為結論與未來研究方向。

2. 自動摘要模型技術 (Techniques of Automatic Summarization)

本論文將過去摘要研究所陸續發展出的自動摘要模型技術大略地歸納成三大類(Mani & Maybury, 1999)：

(1) 以簡單詞彙與結構特徵為基礎之自動摘要模型技術：在 1950 年代，有學者提出使用詞頻(Frequency)來評量每一個詞的重要性與計算文件中每一個語句的顯著性(Significance Factor)(Luhn, 1958)。在實作上，可以對每一個詞進行詞幹分析(Stemming)，將其還原成詞根(Root Form)，同時移除停用詞(Stop Word)的影響並計算實詞(Content Word)的重要性等，最後將語句依其顯著分數進行排序(由高至低)，再根據特定的摘要比例來進行節錄式摘要的產生。後來，有學者利用自然語言分析(Natural Language Analysis)技術對文件結構進行剖析，根據文法結構(Grammar Structure)與語言機制(Linguistic Devices)來決定不同語段的凝聚關係(Cohesion)，例如：首語重複(Anaphora)、省略(Ellipsis)、結合(Conjunction)，或同義詞(Synonymy)、上義詞(Hypernym)等語彙關係(Lexical Relation)，並以此結果進行文件自動摘要。相關研究包括使用語彙鏈(Lexical Chain)(Barzilay & Elhadad, 1997)、宏觀語段結構(Discourse Macro Structure)(Strzalkowski, Wand & Wise, 1998)、修辭結構(Rhetorical Structure)(Zhang *et al*., 2010)等。另有學者在審視 200 篇科技文件後，發現有 85%的重要語句出現在文件中的第一段，7%的重要語句出現在最後一段(Baxendale, 1958)。因此，提出了語句在文件中的位置(Position)資訊是進行摘要語句選取時的一項關鍵線索。

(2) 以非監督式機器學習為基礎之自動摘要模型技術：非監督式機器學習通常將自動摘要任務視為如何排序並挑選具代表性語句之問題，其方法通常是計算出一種摘要特徵供語句排序使用，常見的特徵有：語句與文件相關性(Gong & Liu, 2001)、語句所形成的語言模型生成文件之機率等(Chen, Chen & Wang, 2009)、語句間之相關性(Erkan & Radev, 2004; Mihalcea & Tarau, 2004; Wan & Yang, 2008)、或語句與文件在潛藏主題空間中的距離關係(Lin & Chen, 2009)等。

(3) 以監督式機器學習為基礎之自動摘要模型技術：監督式機器學習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。我們必須事先準備好一些訓練文件以及其對應的人工標註摘要資訊，然後透過各種分類器的學習機制，進行分類模型的訓練。對於尚未被摘要之文件，此類方法將文件裡的每個語句進行二元分類，即可依其結果產生出摘要。此類方法較著名的相關研究包括簡單貝氏分類器(Naïve-Bayes Classifier)(Kupiec, Pedersen & Chen, 1995)、高斯混合模型(Gaussian Mixture Model, GMM)(Murray, Renals & Carletta, 2005)、隱藏式馬可夫模型(Hidden Markov Model, HMM)(Conroy & O'Leary, 2001)、支持向量機(Support Vector Machines, SVM)(Kolcz, Prabakarmurthi & Kalita, 2001; Zhang & Fung, 2007)與條件隨機場域(Conditional Random Fields, CRF)(Shen, Sun, Li, Yang & Chen, 2007)等。監督式模型可同時結合多種摘要特徵來表示每一語句(通常是由上述以詞彙或結構為基礎之摘要方法、或是各式非監督式摘要模型針對語句所輸出的分數或機率值)，綜合各種摘要特徵所形成的特徵向量將被用來做為監督式摘要模型判斷語句是否屬於摘要語句的依據(Lin & Chen, 2009)。

此外，文字文件所要強調的是怎麼說(What-is-said)，而語音文件擁有許多純文字文件所沒有的資訊，通常除了怎麼說，更強調的是如何說(How-is-said)(Penn & Zhu, 2008)，明顯地，語音是多媒體內涵中最具資訊的成分之一，也因此語音文件摘要的相關研究通常從多媒體語音訊號中萃取豐富的韻律資訊(Prosodic Information)來判斷語句的重要性，

如：音調(Intonation)、音高(Pitch)、音強(Power)、語者發聲持續時間(Duration)、語者說話速率(Rate)、語者(Speaker)、情感(Emotion)和說話時場景(Environment)等資訊，這些都是從事語音文件摘要時可以善加利用的語句特徵資訊(Liu & Hakkani-Tur, 2011)。

3. 使用語言模型於語音文件摘要 (Language Modeling for Spoken Document Summarization)

語言模型的研究與發展最早是源自於語音辨識及自然語言處理。語言模型旨在描述語言中的所有詞彙之間共同出現與相鄰資訊的關係。其假設人類語言生成(Human Language Generation)是一個隨機過程，而語言模型就是在模擬如何由詞彙構成片語、語句、段落或者文件之過程的機率模型，故又稱為生成式語言模型(Generative Language Modeling)(Zhai, 2008)。最簡單的語言模型為單連語言模型(Unigram Language Model, ULM)，它不考慮詞彙之間的順序關係，只個別考慮每一個詞本身出現的機率。較為複雜且常被使用的語言模型為 *N*-連語言模型，通常 N 為 2 或 3（即二連或三連語言模型），其考慮兩個詞彙或三個詞彙之間的共同出現與緊連的順序關係。值得一提的是，單連語言模型和 *N*-連語言模型的主要優點之一是：它們僅需使用訓練語料來估測每一個詞本身出現的機率分佈，或者詞彙之間的共同出現與鄰近關係的機率分佈，並不需要額外的人工標記資訊，因此語言模型是屬於基於非監督式機器學習之模型技術。

在過去幾年中，語言模型在資訊檢索任務中已被廣泛地應用且有不錯的實務成效(Zhai, 2008)；但就我們所知，在語音文件摘要的任務上，關於使用語言模型的研究是相對較少的。本論文將藉由語言模型的使用來進行摘要語句選取，其基本方法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)(Chen *et al.*, 2009)。此外，本章第 2 小節我們將闡述如何使用基於關聯性資訊來改進語句模型之估測，使其得以更精準的代表語句的語意內容。

3.1 文件相似度量值 (Document Likelihood Measure, DLM)

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，與其愈相關的語句愈有可能是可用來代表文件主旨或主題之摘要語句。

當給予一篇被摘要文件 *D* 時，文件中每一語句 *S* 的事後機率 *P*(*S*|*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*|*D*)時，我們透過貝氏定理(Bayes' Theorem)將 *P*(*S*|*D*)展開成()：

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

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(Chen *et al.*, 2009)。本論文的研究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， $P(D|S)$ 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 中詞與詞之間是獨立的，並且不考慮每一個詞在文件 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) $P(D|S)$ 可拆解成文件 D 中每一的詞 w 個別發生的條件機率之連乘積：

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

此種方法是為語句 S 建立一個語句模型(Sentence Model) $P(w|S)$， w 是出現在文件 D 中的詞， $C(w,D)$ 是詞 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：

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

在(3)中， $C(w,S)$ 表示詞 w 在語句 S 中出現的次數， $|S|$ 則表示語句 S 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在語句中真正的機率分佈，也可能因為某些詞的條件機率值為零，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，可採用平滑化(Smoothing)技術來達成，常見的平滑化技術包含有 Jelinek-Mercer 平滑化、Dirichlet 平滑化、Add-delta 平滑化 (Zhai & Lafferty, 2001b)，本論文使用 Jelinek-Mercer 平滑化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(Zhai & Lafferty, 2001b)，故 $P(D|S)$ 可進一步地表示成：

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

其中， $P(w|B)$ 是詞 w 在背景單連語言模型 B 中之機率值。

3.2 虛擬相關回饋 (Pseudo-Relevance Feedback)

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)(Chen, Chen, Chen. Wang & Yu, 2014)。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 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，其虛擬關聯回饋示意圖如圖 1 所示。所以本論文針對語句模型調適進行初步研究，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)，接著就要做語句模型的調適估測，底下介紹常見的調適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)。

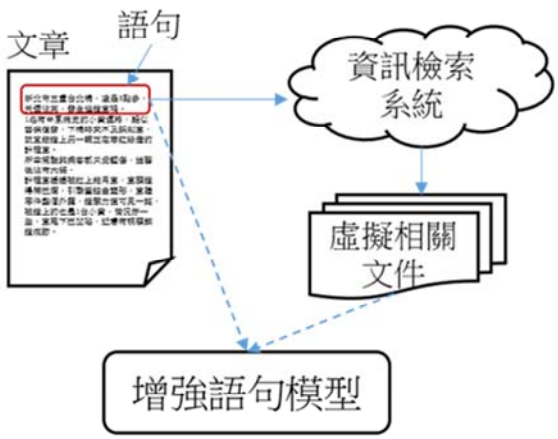

圖 1、虛擬關聯回饋示意圖
[Figure 1. Illustration of pseudo-relevance feedback.]

3.2.1 關聯模型 (Relevance Model, RM)

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)。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 的關聯類別 是非常難以求得的；為此，我們透過虛擬相關回饋(Pseudo Relevant Feedback, PRF)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。更明確地，在實作上我們將虛擬相關文件(最高排序文件)$\mathbf{D}_{\text{Top}}=\{D_1,D_2,\ldots,D_M\}$ 用以代表關聯類別 。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(Lavrenko & Croft, 2001)：

$$P_{\mathbf{RM}}(w,S)=\sum_{D_m\in\mathbf{D}_{\mathbf{Top}}}P(w,S\mid D_m)P(D_m), \tag{5}$$

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：

$$P_{\mathbf{RM}}(w\mid S)=\frac{\sum_{D_m\in\mathbf{D}_{\mathbf{Top}}}\prod_{w'\in S}P(w'\mid D_m)P(w\mid D_m)P(D_m)}{\sum_{D_{m'}\in\mathbf{D}_{\mathbf{Top}}}\prod_{w''\in S}P(w''\mid D_{m'})P(D_{m'})}, \tag{6}$$

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。

3.2.2 簡單混合模型 (Simple Mixture Model, SMM)

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 $\mathbf{D}_{\mathrm{Top}}$ 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(Zhai & Lafferty, 2001a)：

$$\begin{aligned}LL_{\mathbf{D}_{\mathbf{Top}}} = & \sum_{D_m\in\mathbf{D}_{\mathbf{Top}}}\sum_{w\in V}c(w,D_m)\\ & \cdot\log[(1-\alpha)\cdot P_{\mathbf{SMM}}(w\mid S)+\alpha\cdot P(w\mid BG)],\end{aligned} \tag{7}$$

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(7)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：

$$\tau_w^{(l)}=\frac{\alpha\cdot P_{\mathrm{SMM}}^{(l)}(w\mid S)}{\alpha\cdot P_{\mathrm{SMM}}^{(l)}(w\mid S)+(1-\alpha)\cdot P(w\mid BG)}, \tag{8}$$

最大化步驟：

$$P_{\mathbf{SMM}}^{(l+1)}(w \mid S) = \frac{\sum_{D_m \in \mathbf{D_{Top}}} c(w, D_m) \cdot \tau_w^{(l)}}{\sum_{w' \in V} \sum_{D'_m \in \mathbf{D_{Top}}} c(w', D'_m) \cdot \tau_w^{(l)}}, \tag{9}$$

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。

3.2.3 三混合模型 (Tri-Mixture Model)

另一方面，本論文嘗試將三混合模型(Tri-Mixture Model)(Hiemstra, Robertson & Zaragoza, 2004)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 $\mathbf{D}_{\text{Top}}$ 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 $P_{\text{TriMM}}(w|S)$，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(Hiemstra *et al*., 2004)：

$$\begin{aligned} LL_{\mathbf{D_{Top}}} = & \sum_{D_m \in \mathbf{D_{Top}}} \sum_{w \in V} c(w, D_m) \\ & \cdot \log[(1-\lambda-\mu) \cdot P_{\mathbf{TriMM}}(w \mid S) + \lambda \cdot P(w \mid D_m) + \mu \cdot P(w \mid BG)], \end{aligned} \tag{10}$$

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(10)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：

$$\begin{cases} r_{w,D_m} = \dfrac{c(w, D_m) \cdot (1-\lambda-\mu) \cdot P_{\text{TriMM}}(w \mid S)}{(1-\lambda-\mu) \cdot P_{\text{TriMM}}(w \mid S) + \mu \cdot P(w \mid BG) + \lambda \cdot P(w \mid D_m)} \\ e_{w,D_m} = \dfrac{c(w, D_r) \cdot \lambda \cdot P(w \mid D_m)}{(1-\lambda-\mu) \cdot P_{\text{TriMM}}(w \mid S) + \mu \cdot P(w \mid BG) + \lambda \cdot P(w \mid D_m)} \end{cases}, \tag{11}$$

最大化步驟：

$$\begin{cases} \hat{P}_{\mathbf{TriMM}}(w \mid S) = \sum\limits_{D_m \in \mathbf{D_{Top}}} \dfrac{r_{w,D_m}}{\sum_w r_{w,D_m}} \\ \hat{P}(w \mid D_m) = \dfrac{e_{w,D_m}}{\sum_w e_{w,D_m}} \end{cases}, \tag{12}$$

運用此三混合模型來調適語句模型時，可取代原本的語句模型或與之線性結合(linearly interpolation)：

$$\hat{P}(w \mid S) = \gamma \cdot P(w \mid S) + (1-\gamma) \cdot P_{\mathbf{TriMM}}(w \mid S), \tag{13}$$

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(Zhai & Lafferty, 2001a; Lavrenko & Croft, 2001; Hiemstra *et al*., 2004)，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(Chen, Chang & Chen, 2013; Liu *et al*., 2014)，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。

4. 機率式排序模型 (Probabilistic Ranking Model)

在資訊檢索領域(Information Retrieval, IR)中，主要的概念就是設計一個排序模型並利用此模型來將文件做排序。同樣地，我們將節錄式語音文件摘要視為設計一個排序模型，用來排序一篇文件中的每一語句之問題，因此便可應用一些已在資訊檢索領域中發展良好的排序模型於語音摘要任務中，其中最著名的機率式模型即為最佳匹配(Best Matching, BM25)排序模型，我們將陸續介紹最佳匹配排序模型及其延伸。

4.1 BM25

在各式的排序系統中，有學者由機率模型的角度出發，發展出一套簡單且有效地排序計算公式，稱之為 BM25 (Jones, Walker & Robertson, 2000; Robertson & Zaragoza, 2008)：

$$BM25(S, D, B) = \sum_{w \in S} F(w, D) \cdot Sim(w, S) \cdot IDF(w, B) \tag{14}$$

$$F(w, D) = \frac{c(w, D)(k_2 + 1)}{c(w, D) + k_2} \tag{15}$$

$$Sim(w, S) = \frac{c(w, S)(k_1 + 1)}{c(w, S) + k_1(1 - b + b\frac{|S|}{avgsl})} \tag{16}$$

$$IDF(w, B) = \log \frac{B - n(w) + 0.5}{n(w) + 0.5} \tag{17}$$

其中，$c(w,B)$為 w 在文件 D 中的出現次數，B 為背景資訊中所有文件數目，$n(w)$為 w 在背景資訊中出現的文件數目，$|S|$為語句長度，$avgsl$ 為文件 D 中語句的平均長度，k_1、k_2 和 b 為可調的模型參數。

BM25是一個融合語句的詞頻資訊、文件相似度以及反文件頻函數之排序計算公式。在 BM25 的計算公式中，字詞出現在文件 D 的頻率資訊會經由權重函數 $F(w,D)$進行適當

的調整：當參數 k_2 設定為 0 時，則表示 BM25 僅考慮字詞是否有出現於文件當中，而不考慮其出現的頻率，若參數 k_2 的設定不為 0，BM25 將不僅考慮字詞的出現與否，並且進一步地將字詞於文件中出現的頻率資訊做適當的加權；文件相似度 $Sim(w,S)$則用於計算候選文件中與查詢共同出現的詞彙於文件中的重要性，查詢的詞彙在候選文件中亦扮演舉足輕重的角色，若查詢的詞彙共同出現較多次且參數 k_1 的設定不為 0，則表示此篇候選文件應被賦予較高的排序分數；反文件頻函數 $IDF(w,B)$是用於決定每一個詞彙的重要性，也就是加強內容字詞(Content word)的權重，並削弱功能字詞(Function word)的貢獻度。

近年來，有學者將 BM25 運用於意見摘要(Opinion Summarization)研究中，為了符合意見摘要所偏好的語句特性，他們進一步地將 BM25 修改為(Kim, Castellanos, Hsu, Zhai, Dayal & Ghosh, 2013)：

$$BM25_E(S,D,B)=\sum_{w\in S} Sim_E(w,S)\cdot IDF_E(w,B) \tag{18}$$

$$Sim_E(w,S)=\frac{c(w,S)(k_1+1)}{c(w,S)+k_1(1-b+b\frac{|S|}{avgsl})} \tag{19}$$

$$IDF_E(w,B)=\log\frac{|B|-c(w,B)+0.5}{c(w,B)+0.5} \tag{20}$$

其中，$c(w,B)$為 w 在背景資訊 B 中的出現次數，$|B|$為背景資訊所有字詞的次數。比較式(14)與(18)，$BM25_E$ 在對語句進行排序時，省略了考慮查詢詞彙出現頻率的資訊，僅考慮詞彙是否有出現於查詢中；另外，其 $IDF_E(w,B)$的算法是使用字詞 w 在背景資訊 B 中出現的次數，而不是使用字詞 w 在背景資訊 B 中出現的文件數目。

4.2 BM25L and BM25+

當語句很長的時候，文件相似度 $Sim(w,S)$在傳統的 BM25 排序公式(參照式(16))中會變得很小，意即傳統的 BM25 計算公式容易偏好短語句。有鑑於此，有學者提出一個解決方法來平衡語句長度的影響。為了方便解釋此方法，我們將式(16)重新改寫如下：

$$Sim(w,S)=\frac{c'(w,S)(k_1+1)}{c'(w,S)+k_1} \tag{21}$$

其中 $c'(w,S)$為

$$c'(w,S)=\frac{c(w,S)}{1-b+b\frac{|S|}{avgsl}} \tag{22}$$

當重新改寫為式(21)後，有學者提出使用新的文件相似度 *Sim'*(*w*,*S*)，其定義如下：

$$Sim'(w,S)=\begin{cases}\dfrac{(c'(w,S)+\delta)(k_1+1)}{(c'(w,S)+\delta)+k_1} & if\ c'(w,S)>0\\ 0 & c'(w,S)=0\end{cases} \tag{23}$$

其中 δ 為一定值，則新的排序公式為(BM25L)：

$$BM25L(S,D,B)=\sum_{w\in S}F(w,D)\cdot Sim'(w,S)\cdot IDF(w,B) \tag{24}$$

此新的文件相似度不僅保留原有BM25的兩點良好特性，(即當 *c'*(*w*,*S*)=0 時 *Sim'*(*w*,*S*)=0；另外，*c'*(*w*,*S*)與 *BM25L* 皆呈單調遞增，並且 *BM25L* 有漸進最大值(Asymptotic maximal))，同時也因此有了一個正下界(positive lower bound)的特性(即對於 *c'*(*w*,*S*)>0，至少都會有 $(k_1+1)\delta/(k_1+\delta)$)，此特性可以平衡語句長度之影響，不會因為語句過長而影響變大且不會特別容易偏好短語句。

一方面，Lv & Zhai (2011a)發現不只原始 BM25 排序公式會過度懲罰長語句，就連其他的排序公式都會有一樣的情形，因此他們更進一步地提出一般化的方法來解決此問題，也就是要保證只出現一次的詞彙在長語句中至少會有一定的貢獻度，為了達到此目的，他們就在原始 BM25 公式中的文件相似度 *Sim*(*w*,*S*)裡加入一個常數值，且反文件頻函數 *IDF*(*w*,*B*)也有小修改，則新的排序公式為(BM25+，(Lv & Zhai, 2011b))：

$$BM25+(S,D,B)=\sum_{w\in S}F(w,D)\cdot Sim^{+}(w,S)\cdot IDF^{+}(w,B) \tag{25}$$

$$Sim^{+}(w,S)=\frac{c(w,S)(k_1+1)}{c(w,S)+k_1(1-b+b\dfrac{|S|}{avgsl})}+\delta \tag{26}$$

$$IDF^{+}(w,B)=\frac{|B|+1}{n(w)} \tag{27}$$

其中 δ 為一個固定值。

4.3 BM25T

在 4.1 小節所介紹的 BM25 排序公式中有三個需要設定的參數(k_1, k_2, b)，且所有的詞彙共享同一組設定，但其實每個詞彙應該要根據不同的重要性而設計不同的參數值。由於文件相似度 *Sim*(*w*,*S*) 是 BM25 公式中最重要的排序因子，所以參數 k_1 的設計就更顯重要。Lv & Zhai (2012) 認為經長度正規化的詞頻貢獻度應該要與有較高的長度正規化詞頻的文章數成正比，因此他們使用對數邏輯 (Log-logistic)方法來計算每個詞彙所對應不同的

參數 k_1，首先定義一個菁英集(Elite set)C_w，意即所有包含詞彙 w 的語句集合，則詞彙 w 的參數 k_1'之定義如下：

$$k_1' = \arg\min_{k_1} \left(g_{k_1} - \frac{\sum_{S' \in C_w} \left(\log(c'(w, S') + 1) \right)}{n(w)} \right)^2 \tag{28}$$

$$g_{k_1} = \begin{cases} \frac{k_1}{k_1 - 1} \log(k_1) & if \ k_1 \neq 1 \\ 1 & k_1 = 1 \end{cases} \tag{29}$$

其中 $c'(w,S')$與式(22)相同，我們將 k_1 的範圍設定在 0.1 到 10 之間(每次增加 0.1)，透過式(28)我們可找到每一個詞彙 w 的最佳參數 k_1'，將式(28)所求得的參數帶回原始的 BM25 排序公式，便可得到新的排序公式(BM25T，(Lv & Zhai, 2012))：

$$BM25T(S, D, B) = \sum_{w \in S} F(w, D) \cdot Sim^T(w, S) \cdot IDF(w, B) \tag{30}$$

$$Sim^T(w, S) = \frac{c(w, S)(k_1' + 1)}{c(w, S) + k_1'(1 - b + b \frac{|S|}{avgsl})} \tag{31}$$

表 1. 實驗語料統計資訊
[Table 1. The statistics of the dataset.]

	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%

5. 實驗語料及評估方法 (Dataset and Evaluation Method)

5.1 實驗語料 (Dataset)

本論文實驗語料庫為公視新聞語料(Mandarin Chinese Broadcast News Corpus, MATBN)(Wang, Chen, Kuo & Cheng, 2005)，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 1 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(Chen *et al*., 2013)，關於語句 S 的虛擬相關文件(最高排序文件)篇數為 15(也就是$|\mathbf{D}_{\mathbf{Top}}|$=15)。

5.2 評估方法 (Evaluation Method)

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)(Lin, 2003)。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 (單連詞 Unigram)、ROUGE-2 (雙連詞 Bigram) 和 ROUGE-L (最長共同片段 Longest Common Subsequence) 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。

6. 實驗結果 (Experimental Results)

6.1 基礎實驗結果 (Baseline Experiments)

首先，我們比較文件相似度量值(DLM)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)(Penn & Zhu, 2008)、向量空間模型(Vector Space Model, VSM)(Gong &. Liu, 2001)、潛藏語意分析(Latent Semantic Analysis, LSA)(Gong &. Liu, 2001)、最大邊際關聯(Maximal Marginal Relevance, MMR)(Carbonell & Goldstein, 1998)、馬可夫隨機漫步(Markov Random Walk, MRW)(Wan & Yang, 2008)、次模(Submodularity)(Lin & Bilmes, 2010)以及整數線性規劃(Integer Linear Programming, ILP)(McDonald, 2007)。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。馬可夫隨機漫步(MRW)的概念是把文件中的語句視為一個網際網路，文件中的語句代表網路中的節點，節邊權重值是兩個節點之間的語彙相似度，通常是透過節點的內分支度(Indegree)與外分支度(Outdegree)並採用餘弦(Cosine)估測法求得，所以馬可夫隨機漫步主要是依賴較一般化的資訊，例如：有概念性的網際網路，而不是考慮區域性的特徵(例如：每個語句)，因此如果有一個語句跟其他語句很相似的話，則可以代表摘要使之來描述文件中的主旨(Wan & Yang, 2008)。次模是一個貪婪(Greedy) 的語句選取方法，因其滿足次模的特性，意即每選取一語句就會有回報減少(Diminished Return)的效應，因此次模具有一個近似最佳解(Near-Optimal)(Lin & Bilmes, 2010)。整數線性規劃是一個全域(Global)的限制性最佳化(Constraint Optimization)的語句選取方法(McDonald, 2007)。

表 2 為本論文之基礎實驗結果。首先，在 TD 的實驗中，DLM 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。

表 2. 基礎實驗結果
[Table 2. Baseline experiments.]

		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
	MRW	0.412	0.282	0.358
	Submodularity	0.414	0.286	0.363
	ILP	0.442	0.337	0.401
	DLM	0.411	0.298	0.361
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
	MRW	0.332	0.191	0.291
	Submodularity	0.332	0.204	0.303
	ILP	0.348	0.209	0.306
	DLM	0.364	0.210	0.307

相較之下，DLM 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。DLM 與 VSM 皆使用淺層的詞彙(詞頻)資訊，但由於 DLM 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。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，DLM 同樣較優於 LS、LEAD、VSM、LSA 等之摘要方法，但 MMR 的結果則稍微較 DLM 好一點，我們認為這可能是因為 MMR 比較不受到語音辨認錯誤的影響。但 MRW 及次模也可能是受到語音辨識錯誤的影響而造成摘要效能減低，甚至比 DLM 來得差。出乎意料的是原以為 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。

6.2 關聯模型之實驗結果 (Experiments of Relevance Model)

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 15 篇虛擬相關文件。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 3 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 DLM 實驗較好，尤其是三混合模型(TriMM)相較於 DLM 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 3 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(5))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 3、關聯模型之實驗結果
[Table 3. Experimental results of different relevance models.]

		F-score		
		ROUGE-1	ROUGE-2	ROUGE-L
TD	DLM	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	DLM	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**

6.3 平滑化技術於關聯模型之實驗結果 (Experiments of Smoothing Methods for Relevance Model)

語言模型在使用時會遇到資料稀疏的問題，通常的解決方法為替語言模型做平滑化(Smoothing)，我們將探討平滑化技術於語言模型在語音(文字)摘要結果上的影響，在本小節中我們以關聯模型(RM，參考 3.2.1 小節)為例[1]，採用三種不同的平滑化技術於關聯模型中，第一為 Jelinek-Mercer 平滑化，第二為 Dirichlet 平滑化，第三為 Add-delta 平滑化，茲分別如下(Zhai & Lafferty, 2001b)：(i) Jelinek-Mercer 平滑化為最簡單的與背景模型 $P(w|B)$線性結合的平滑化技術，其公式為：

$$P_{JM}(w|S) = \lambda P(w|S) + (1-\lambda)P(w|B) \tag{32}$$

其中 λ 為線性結合參數，在實驗設定中是從 0.1 到 0.9(每次增加 0.1)。(ii) Dirichlet 平滑化主要是根源於貝式平滑(Bayesian Smoothing)而來的，它假設語言模型有個事前(Prior)機率，而此事前機率的分布剛好就是 Dirichlet 分布，因此 Dirichlet 平滑化公式可定義如下(Zhai & Lafferty, 2001b)：

$$P_{Dir}(w|S) = \frac{c(w|S) + \mu \cdot P(w|B)}{|S| + \mu} \tag{33}$$

[1] 我們實驗發現不同的平滑化技術都會對這三種不同的關聯模型有幫助，在摘要成效上也都有明顯的進步，且關聯模型(RM)會有最大的進步。

表 4. 平滑化技術於關聯模型(RM)之實驗結果
[Table 4. Experimental results of various smoothing methods for relevance model.]

Relevance Model (RM)		*F*-score		
		ROUGE-1	ROUGE-2	ROUGE-L
TD	Jelinek-Mercer	0.450	0.336	0.400
	Dirichlet	0.472	0.365	0.428
	Add-delta	0.493	0.386	0.441
SD	Jelinek-Mercer	0.374	0.226	0.321
	Dirichlet	0.401	0.254	0.349
	Add-delta	0.402	0.255	0.347

其中 μ 為 Dirichlet 參數，在實驗設定中的範圍為 1 到 100(每次增加 1)。(iii) Add-delta 平滑化是一個簡單平滑化技術，其原理就是加入一點點值，使沒有出現過的詞彙之機率不為零，其公式定義如下(Lv & Zhai, 2014)：

$$P_{Delta}(w \mid S) = \frac{c(w \mid S) + \delta}{|S| + \delta \cdot |V_F|} \tag{34}$$

其中 δ 為可調參數，在實驗設定範圍為 0.1 到 1(每次增加 0.1)，而$|V_F|$為虛擬相關回饋文件(在此為 15 篇)中不同詞彙的個數。三種平滑化技術於關聯模型(RM)的語音(文字)摘要結果如表 4 所示，無論在 TD 或 SD 的情況下，其中表現最佳為 Add-delta 平滑化，其次是 Dirichlet 平滑化，最差的是 Jelinek-Mercer 平滑化。Add-delta 平滑化表現比較好的原因是因為利用到相關回饋文件中不同詞彙的個數($|V_F|$)的資訊，使之能讓共同出現在語句與相關回饋文件中的詞彙 w 有比較高的機率(相較於沒有共同出現的詞彙)，因此在估測關聯模型時能更具有鑑別能力(區分出重要且共同出現在語句與相關回饋文件的詞彙與一般性且不重要的詞彙)，而讓摘要效能變得更好，尤其是 TD 的情況下相較於 Jelinek-Mercer 平滑化在 ROUGE-2 的絕對進步率有 5%之多，這是相當顯著的。但在 SD 的情況下，雖然 Add-delta 平滑化還是會比 Dirichlet 平滑化及 Jelinek-Mercer 平滑化來得好，但與 Dirichlet 平滑化相比，其摘要效能其實已經相差無幾，可能的原因之一還是因為語音辨識錯誤的影響所造成，在未來研究中，我們將以次詞索引(Subword Indexing)的方式來建立關聯模型以減緩此問題。

6.4 機率式排序模型之實驗結果 (Experiments of Probabilistic Ranking Model)

接著我們將焦點轉移到機率式排序模型上，在本小節中我們將比較各種不同的 BM25 排序模型，包含有原始 BM25(參照式(14))、$BM25_E$(參照式(18))、BM25L(參照式(24))、BM25+(參照式(25))及 BM25T(參照式(30))。其實驗結果如表 5 所示，在 TD 的部分，BM25 的摘要表現已經很不錯，甚至都比其他良好發展的非監督式摘要方法來得好(與表 2 之結果比較) ，$BM25_E$ 因少了一個重要因子來做語句排序，所以可預期它的摘要效能比 BM25 來得差，但超乎預期的是在 ROUGE-2 將近有 16%的差距。BM25L 的提出是為了解決過度懲罰長語句的問題，在本實驗中的 TD 情況下，可看出 BM25L 的摘要效能會沒有比原始 BM25 來的好，其可能的原因是在資訊檢索領域中，確實會有很長文章(Long Document)的出現，懲罰長文章會有一定的效果，但在語音摘要任務中，很長語句(Long Sentence)的出現幾乎是不太可能，因此懲罰長語句就不一定會得到好的摘要效能。BM25+也是為了解決過度懲罰長語句的問題，但更一般化地保證只出現一次的詞彙至少要有個下界，因此 BM25+的摘要效能比 BM25L 能更進一步的提升，與原始 BM25 及 BM25L 相比較，在 ROUGE-2 上分別能有 0.2%及 1.1%的絕對進步。BM25T 從訓練資料中自動學習與詞彙相關的參數 k_1，在原始的文獻裡用於資訊檢索領域中的實驗是相對排序公式來得好(Lv & Zhai, 2012)，但 BM25T 的摘要效能有點出乎意料之外，沒有比 BM25L 與 BM25+好，甚至也會比原始 BM25 排序公式來的差，我們認為此種自動學習參數的方法可能是與資料相關的，或許可以替換另一套訓練資料集來重新學習詞彙相關的參數，但這也是未來的工作之一。另一方面，在 SD 的實驗部分，原始的 BM25 排序公式還是維持一定的水準，與其他非監督式摘要方法相比還是會比較好(參照表 2)，BM25L、BM25+及 BM25T 的優勢就沒那麼大，其摘要效能在 ROUGE-2 上比原始 BM25 都要來得差，確實語音辨識錯誤的影響還蠻大的，原本 BM25L、BM25+及 BM25T 的優點都被錯誤辨識的詞彙所消彌，甚至 BM25L、BM25+與 BM25T 的摘要結果幾乎相差無幾。

表 5. BM25 其及變形之相關實驗結果
[Table 5. Experimental results of BM25 and its variants.]

		F-score		
		ROUGE-1	ROUGE-2	ROUGE-L
TD	BM25	0.484	0.374	0.442
	$BM25_E$	0.352	0.210	0.294
	BM25L	0.480	0.365	0.434
	BM25+	0.486	0.376	0.444
	BM25T	0.463	0.352	0.419
SD	BM25	0.390	0.247	0.338
	$BM25_E$	0.279	0.151	0.250
	BM25L	0.384	0.246	0.337
	BM25+	0.383	0.242	0.335
	BM25T	0.382	0.238	0.332

7. 結論與未來展望 (Conclusions and Future work)

本論文主要有三個研究貢獻，其一為有鑑於關聯性(Relevance)的概念在資訊檢索領域中已有不錯的發展成果，本論文嘗試結合關聯性資訊來重新估測並建立語句的語言模型，並首次使用三混合(Tri-Mixture Model, TriMM)模型，使其得以更精準地代表語句的語意內容，期望可增進自動摘要之效能，實驗結果顯示三混合模型可以有最佳的摘要效能。其二為有鑑於語言模型著重依賴平滑化技術，本論文也是首次比較研究不同平滑化技術所估測得的語言模型對語音文件摘要任務之影響，根據實驗結果 Add-delta 平滑化可以達到最佳摘要效果，所以我們建議關聯模型的平滑化技術應當使用 Add-delta 平滑化來達成。最後為我們首次提出並應用多種機率式資訊檢索排序模型於語音摘要任務上，並且從實驗結果中得知與其他常見的非監督式摘要方法相比較能有不錯的摘要效能。

未來，我們的研究將有三個主要的方向：首先，多種機率式檢索排序模型還是需要經驗去調整不確定的參數，我們將進一步的研究是否可以針對不同的文件或不同的語句給予適當的權重調整，以期獲得更好的摘要成效；第二，目前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，我們希望將非監督式摘要方法所產生的分數視為一種具代表性的摘要特徵資訊並結合於監督式機器學習方法(如條件隨機場域(Conditional Random Fields, CRFs)或深度類神經網絡(Deep Neural Network Learning, DNN)等)中，期望訓練後的模型能夠在文字文件摘要或語音文件摘要上獲得更好的表現。

Reference

Barzilay, R., & Elhadad, M. (1997). Using Lexical Chains for Text Summarization. In *Proceedings of ACL Workshop on Intelligent Scalable Text Summarization*, 10-17.

Baxendale, P. (1958). Machine-made Index for Technical Literature - an Experiment. *IBM Journal of Research and Development*, *2*(4), 354-361.

Cai, X.-Y., & Li, W.-J. (2013). Ranking through Clustering: An Integrated Approach to Multi-Document Summarization. *IEEE Transactions on Audio, Speech and Language Processing*, *21*(7), 1424-1433.

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

Chen, Y.-T., Chen, B., & Wang, H.-M. (2009). A Probabilistic Generative Framework for Extractive Broadcast News Speech Summarization. *IEEE Transactions on Audio, Speech and Language Processing*, *17*(1), 95-106.

Chen, B., Chang, H.-C., & Chen, K.-Y. (2013). Sentence Modeling for Extractive Speech Summarization. In *Proceedings of the International Conference on Multimedia & Expo (ICME)*. doi: 10.1109/ICME.2013.6607518

Chen, B., Chen, Y.-W., Chen, K.-Y., Wang, H.-M., & Yu, K.-T. (2014). Enhancing Query Formulation for Spoken Document Retrieval. *Journal of Information Science and Engineering*, *30*(3), 553-569.

Conroy, J.-M., & O'Leary, D.-P. (2001). Text Summarization via Hidden Markov Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 406-407. doi: 10.1145/383952.384042

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligent Research*, *22*(1), 457-479.

Galley, M., McKeown, K., Hirschberg, J., & Shriberg, E. (2004). Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 669-676. doi: 10.3115/1218955.1219040

Gong, Y., & Liu, X. (2001). Generic Text Summarization using Relevance Measure and Latent Semantic Analysis. In *Proceedings of the 24th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 19-25. doi: 10.1145/383952.383955

Hiemstra, D., Robertson, S., & Zaragoza, H. (2004). Parsimonious Language Models for Information Retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 178-185. doi: 10.1145/1008992.1009025

Jones, K. S., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, *36*(6), 779-808. doi: 10.1016/S0306-4573(00)00015-7

Kim, H.-D., Castellanos, M. G., Hsu, M., Zhai, C., Dayal, U., & Ghosh, R. (2013). Ranking Explanatory Sentences for Opinion Summarization. In *Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1069-1072. doi: 10.1145/2484028.2484172

Kolcz, A., Prabakarmurthi, V., & Kalita, J. (2001). Summarization as Feature Selection for Text Categorization. In *Proceedings of the tenth International Conference on Information and Knowledge Management*, 365-370. doi: 10.1145/502585.502647

Kuo, J.-J., & Chen, H.-H. (2006). Multi-document Summary Generation using Informative and Event Words. *Journal of ACM Transactions on Asian Language Information Processing*, *7*(1), 550-557. doi: 10.1145/1330291.1330294

Kupiec, J., Pedersen, J., & Chen, F. (1995). A Trainable Document Summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 68-73. doi: 10.1145/215206.215333

Lavrenko, V., & Croft, W.-B. (2001). Relevance-based Language Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 120-127. doi: 10.1145/383952.383972

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

Lin, S.-H., & Chen, B. (2010). A Survey on Speech Summarization Techniques. *The Association for Computational Linguistics and Chinese Language Processing Newsletter*, *21*(1), 4-16.

Lin, H., & Bilmes, J. (2010). Multi-document Summarization via Budgeted Maximization of Submodular Functions. In *Proceeding of NAACL HLT*, 912-920.

Lin, S.-H., Yeh, Y.-M., & Chen, B. (2011). Leveraging Kullback-Leibler Divergence Measures and Information-Rich Cues for Speech Summarization. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(4), 871-882. doi: 10.1109/TASL.2010.2066268

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

Liu, Y., & Hakkani-Tur, D. (2011). Speech Summarization. In G. Turand & R. D. Mori (Eds), *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. West Sussex, U.K.: Wiley. doi: 10.1002/9781119992691.ch13

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

Acoustics, Speech, and Signal Processing, 3226-3230. doi: 10.1109/ICASSP.2014.6854196

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, *2*(2), 159-165.

Lv, Y., & Zhai, C.-X. (2011a). When Documents Are Very Long, BM25 Fails! In *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1103-1104. doi: 10.1145/2009916.2010070

Lv, Y., & Zhai, C.-X. (2011b). Lower-bounding Term Frequency Normalization, In *Proceeding of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*, 7-16. doi: 10.1145/2063576.2063584

Lv, Y., & Zhai, C.-X. (2012). A Log-logistic Model-based Interpretation of TF Normalization of BM25. In *Proceedings of European Conference on Information Retrieval (ECIR)*, 244-255.

Lv, Y., & Zhai, C.-X. (2014). Revisiting the Divergence Minimization Feedback Model, In *Proceeding of the 23rd ACM International Conference on Information and Knowledge Management (CIKM)*, 1863-1866. doi: 10.1145/2661829.2661900

Mani, I., & Maybury, M.-T. (1999). *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press.

McDonald, R. (2007). A Study of Global Inference Algorithms in Multi-document Summarization, In *Proceedings of the 29th European Conference on Information Retrieval*, 557-564.

Mihalcea, R., & Tarau, P. (2004). TextRank Bringing Order into Texts. In *Proceedings of Empirical Method in Natural Language Processing (EMNLP 2004)*, 404-411.

Murray, G., Renals, S., & Carletta, J. (2005). Extractive Summarization of Meeting Recordings. In *Proceedings of the Conference of the International Speech Communication Association (Interspeech)*, 593-596.

Nenkova, A., & McKeown, K. (2011). Automatic Summarization. *Foundations and Trends in Information Retrieval*, *5*(2-3), 103-233. doi : 10.1561/1500000015

Ostendorf, M. (2008) Speech Technology and Information Access. *IEEE Signal Processing Magazine*, *25*(3), 150-152. doi: 10.1109/MSP.2008.918685

Paice, C.-D. (1990). Constructing Literature Abstracts by Computer Techniques and Prospects. *Journal of Information Processing and Management*, *26*(1), 171-186. doi: 10.1016/0306-4573(90)90014-S

Penn, G., & Zhu, X. (2008). A Critical Reassessment of Evaluation Baselines for Speech Summarization. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 470-478.

Robertson, S. & Zaragoza, H. (2008). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, *3*(4), 333-389. doi: 10.1561/1500000019

Shen, D., Sun, J.-T., Li, H., Yang, Q., & Chen, Z. (2007). Document Summarization using Conditional Random Fields. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2862-2867.

Strzalkowski, T., Wand, J., & Wise, B. (1998). A Robust Practical Text Summarization. In *Proceedings of AAAI Conference on Artificial Intelligence Spring Symposium on Intelligent Text Summarization*, 26-33.

Wan, X., & Yang, J. (2008). Multi-document Summarization using Cluster-based Link Analysis. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 299-306. doi: 10.1145/1390334.1390386

Wang, H.-M., Chen, B., Kuo, J.-w., & Cheng, S.-S. (2005). MATBN: A Mandarin Chinese broadcast news corpus. *International Journal of Computational Linguistics and Chinese Language Processing*, *10*(2), 219-236.

Witbrock, M., & Mittal, V. (1999). Ultra Summarization: a Statistical Approach to Generating Highly Condensed Non-extractive Summaries. In *Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 315-316. doi: 10.1145/312624.312748

Zhai, C.-X., & Lafferty, J. (2001a). Model-based feedback in the language modeling approach to information retrieval. In *Proceeding of the tenth International Conference on Information and Knowledge Management (CIKM)*, 403-410. doi: 10.1145/502585.502654

Zhai, C.-X., & Lafferty, J. (2001b). A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 334-342. doi: 10.1145/383952.384019

Zhai, C.-X. (2008). Statistical Language Models for Information Retrieval: A Critical Review. *Foundations and Trends in Information Retrieval*, *2*(3), 137-213. doi: 10.1561/1500000008

Zhang, J., & Fung, P. (2007). Speech Summarization without Lexical Features for Mandarin Broadcast News. In *Proceedings of NAACL HLT*, Companion Volume, 213-216.

Zhang, J.-J., Chan, H.-Y., & Fung, P. (2010). Extractive Speech Summarization using Shallow Rhetorical Structure Modeling. *IEEE Transactions on Audio, Speech and Language Processing*, *18*(6), 1147-1157. doi: 10.1109/TASL.2009.2030951