Computational Linguistics and Chinese Language Processing
Vol. 20, No. 2, December 2015, pp. 65-86

節錄式語音文件摘要使用表示法學習技術

Extractive Spoken Document Summarization with Representation Learning Techniques

施凱文*、陳冠宇+、劉士弘+、王新民+、陳柏琳*

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

Hsin-Min Wang and Berlin Chen

摘要

大量多媒體內容的與日俱增促使自動語音文件摘要成為一項重要的研究議題。其中最為廣泛地被探究的是節錄式語音文件摘要(Extractive Spoken Document Summarization)，其目的是根據事先定義的摘要比例，從語音文件中選取一些重要的語句，用以代表原始語音文件的主旨或主題。另一方面，表示法學習(Representation Learning)是近期相當熱門的一個研究議題，多數的研究成果也證明了這項技術在許多自然語言處理(Natural Language Processing, NLP)的相關任務上，可以進一步地獲得優良的成效。有鑑於此，本論文主要探討使用詞表示法(Word Representations)及語句表示法(Sentence Representations)於節錄式中文廣播新聞語音文件摘要之應用。基於詞表示法及語句表示法，本論文提出三種新穎且有效的排序模型(Ranking Models)。除了文件中的文字資訊外，本論文更進一步地結合語音文件上的各式聲學特徵，如韻律特徵(Prosodic Features)等，期望可以獲得更好的摘要成效。本論文的語音文件摘要實驗語料

*國立臺灣師範大學資訊工程學系
Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: {60247065S, berlin}@ntnu.edu.tw
+中央研究院資訊科學所
Institute of Information Science, Academia Sinica.
E-mail: {kychen, journey, whm}@iis.sinica.edu.tw
The author for correspondence is Berlin Chen.

是採用公視廣播新聞；實驗結果顯示，相較於其它現有的摘要方法，我們所發展的新穎式摘要方法能夠提供顯著的效能改善。

關鍵詞：語音文件、節錄式摘要、詞表示法、語句表示法、韻律特徵

Abstract

The rapidly increasing availability of multimedia associated with spoken documents on the Internet has prompted automatic spoken document summarization to be an important research subject. Thus far, the majority of existing work has focused on extractive spoken document summarization, which selects salient sentences from an original spoken document according to a target summarization ratio and concatenates them to form a summary concisely, in order to convey the most important theme of the document. On the other hand, there has been a surge of interest in developing representation learning techniques for a wide variety of natural language processing (NLP)-related tasks. However, to our knowledge, they are largely unexplored in the context of extractive spoken document summarization. With the above background, this study explores a novel use of both word and sentence representation techniques for extractive spoken document summarization. In addition, three variants of sentence ranking models building on top of such representation techniques are proposed. Furthermore, extra information cues like the prosodic features extracted from spoken documents, apart from the lexical features, are also employed for boosting the summarization performance. A series of experiments conducted on the MATBN broadcast news corpus indeed reveal the performance merits of our proposed summarization methods in relation to several state-of-the-art baselines.

Keywords: Spoken Document, Extractive Summarization, Word Representation, Sentence Representation, Prosodic Feature

1. 緒論

巨量資料充斥著現今的世界，在全球資訊網(World Wide Web)中已存在有數十億篇網頁，並且以指數的倍數持續成長著。為此，人們必須仰賴及時摘要各類資訊的自動化工具，以減緩資訊過載(Information Overload)的問題。這些迫切的需求促使了自動摘要(Automatic Summarization)技術的蓬勃發展(Luhn, 1958)。自動摘要技術可概分為節錄式(Extractive)摘要以及抽象式(Abstractive)摘要。前者主要是依據特定的摘要比例，從原始的文件中選取重要的語句子集(Sentence Subset)，透過該語句子集簡潔地表示原始文件的大致內容；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。雖然抽象式摘要是最為貼近人們日常撰寫摘要的形式，但其涉及深層的自然語言處

理能力(Mitra *et al.*, 1997)，較為困難許多；目前大多數的研究主要集中在節錄式摘要的自動產生(Jones, 1999)。除了傳統的文字文件外，多媒體文件亦迅速地在世界各地傳播，例如語音郵件、會議錄音、電視新聞以及課程演講等；因此，語音文件摘要(Spoken Document Summarization)自然地成為近年來的一項受關注的研究議題。本論文主要探討節錄式語音文件摘要，其目標是基於定量的摘要比例(Summarization Ratio)，從多媒體內容所對應的語音文件中選取能夠表示其內容主題資訊之語句，讓使用者可以迅速地理解多媒體內容的主要意涵。

本論文延續過去學者的經驗、成果以及貢獻，針對語音文件的特性及困難點，進一步地深入研究語音文件摘要方法與特徵使用，希望藉由自動語音文件摘要技術的改進，使自動摘要的結果能更適切地詮釋語音文件內容及主題。本論文提出兩個主要的研究貢獻：(1)基於表示法學習(Representation Learning)技術，本論文提出三種排序模型(即統計式模型、機率式模型以及圖論式模型)，嘗試將表示法學習技術運用於語音文件摘要任務之中。(2)除了利用文件中的文字資訊外，本論文進一步地結合語音文件中的各種韻律特徵，期望增進摘要系統的成效。

本論文的後續安排如下：第二節首先簡介當前主要的文件摘要方法。第三節介紹本論文所探究的表示法學習技術。第四節介紹本論文提出的三種排序模型於表示法學習之技術。第五節簡介語音文件的多種特徵。第六節介紹實驗語料及摘要評估之方法。第七節說明實驗結果及其分析。第八節為本論文之結論與未來展望。

2. 基礎文件摘要方法之簡介

2.1 以文件結構為基礎之摘要方法

摘要單位元(例如詞彙或語句)在文件中的位置資訊(Positional Information)可以作為文件尋找重要語句選取的判斷依據。一般而言，位於段落開頭的摘要單位元通常有較高的重要性及代表性。因此，前導(LEAD)摘要方法是根據摘要比例選取文件前 M %的部分作為摘要(Hajime & Manabu, 2000)。這種摘要的方式簡單且直覺，但僅適用於特定的結構內容，並且文件需遵循某種編排方式。對於某些沒有特定結構編排的文件，或是缺乏文件結構的語音文件內容有可能會不適用。

2.2 以統計值為基礎之摘要方法

A. 向量空間模型(Vector Space Model, VSM)

向量空間模型已被廣泛地應用於資訊檢索中，用以估測使用者查詢(Query)與文件(Document)之間的相關程度(Salton & Lesk, 1968)。節錄式語音文件摘要任務可以被視為是一個資訊檢索的問題，重要語句的選取是以句子與語音文件內容的相關程度而定；亦即將文件內容視為查詢來檢索最相關的 m 個語句。因此，可以將文件表示成向量 $\vec{D}$ ，文件中的每一語句表示為向量 $\vec{S}_t$ ，透過餘弦相似度(Cosine Similarity)計算，就可估測兩者

之間的相似性程度。雖然直接地利用文件中的文字資訊已被證明在許多自然語言處理的相關問題中可以獲得一定的成效，但這樣的方法忽略了隱藏於文字中的語意資訊(Semantic Information)。為了有效地運用這些語意資訊，最早於 2001 年開始有學者提出潛藏語意的概念應用於文件摘要的方法(Gong & Liu, 2001)。

B. 最大邊際關聯法(Maximum Margin Relevance, MMR)

最大邊際關聯法是於 1998 年被提出的自動摘要準則(Carbonell & Goldstein, 1998)。該準則是以遞迴的方式一次一句地挑選可能的摘要語句，其考量的重點不僅是希望所選取出來的摘要語句與文件的關聯性分數要高，更進一步地考慮了候選語句與已被選取的摘要語句之間的重複性分數要低。如此一來，摘要系統選取出的語句不僅可以代表文件中的重要主題，更可以充分的涵蓋文件中的各個面向。直到今日，最大邊際關聯法仍為節錄式摘要方法常使用的重要準則；它也常是一個主要的基礎系統，做為新摘要方法發展時效能比較與評估的依據。

2.3 以機率式模型為基礎之摘要方法

A. 單連語言模型(Unigram Language Model, ULM)

語言模型被廣泛地應用於語音辨識(Speech Recognition)與機器翻譯(Machine Translation)等方面，學者Ponte等人在1998年時將其運用於資訊檢索的問題中(Ponte & Croft, 1998)。由於我們可以將節錄式語音文件摘要任務視為一個資訊檢索的問題，即當給予一篇文件 D 時，希望對文件中的語句 S 依照機率值 $P(S|D)$進行排序；藉由貝式定理的推導，可得(Chen *et al*., 2009)：

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

其中 $P(D)$對於每一語句皆相同，故可忽略。而我們假設每一語句 S 的事前機率 $P(S)$為一個均勻分佈(Uniform Distribution)，因此 $P(S)$亦可忽略。值得一提的是，由於文件中的語句通常較為簡短，不容易建立一個準確的模型來完整地描述每一語句的內容涵意。為此，有研究學者陸續提出各式較為強健性的語言模型，例如關聯模型(Relevance Model)(Lavrenko & Croft, 2001)等，期望可以改善此一問題。關聯模型的優點在於融入關聯文件中的資訊，藉此豐富語句模型使得更準確地表達語句的主題特性，以提升摘要的成效。

B. Okapi Best Match 25 (BM25)

Okapi BM25 是於 1994 年由學者 Robertson 等人所提出的權重計算公式，是現今資訊檢索模型中最著名的機率式檢索模型之一。其權重計算方式主要是將詞頻對文件長度作正規化，有效降低因文件長度不同而產生的檢索誤差(Robertson & Jones, 1976; Robertson &

Walker, 1994; Robertson *et al.*, 1996)。當利用該方法於文件摘要任務時，我們首先對文件的詞序列$D = (w_1 w_2 \dots w_{|d|})$計算出每個詞$w_i$與語句 S 之間的相似性分數，接著將每個詞w_i對於語句 S 的相似性分數進行加權求和，進而得到文件 D 與語句 S 的相似性分數，公式如下：

$$BM25(D,S) = \sum_{w \in D} F(w,D) \cdot Sim(w,S) \cdot log \frac{N}{1 + df_w} \tag{2}$$

$$F(w,D) = \frac{c(w,D) \cdot (k_2 + 1)}{c(w,D) + k_2} \tag{3}$$

$$Sim(w,S) = \frac{c(w,S) \cdot (k_1 + 1)}{c(w,S) + k_1 \cdot \left(1 - b + b \cdot \frac{|S|}{|avgS|}\right)} \tag{4}$$

其中k_1、k_2以及 b 均為自由參數，根據經驗設置，一般$k_1 \in [1.2, 2.0]$、$b = 0.75$；$c(w,S)$是詞w在語句 S 中出現的次數；$c(w,D)$是詞w在文件 D 中出現的次數；而$|S|$是表示語句 S 的長度，$|avgS|$是在文件中所有語句的平均長度；N 是在集合中的文件總數；df_w為在集合中文件包含詞w的篇數。

2.4 以圖論為基礎摘要之方法

A. 詞權重-逆向文件頻率(Term Weight-Inverse Document Frequency, TW-IDF)

詞權重-逆向文件頻率模型是由學者 Rousseau 與 Vazirgiannis 於 2013 年所提出(Rousseau & Vazirgiannis, 2013)。首先，此方法為每一篇文件建立一個有向圖(Directed Graph)，圖中的每一個頂點(Vertex)代表文件中的一個唯獨詞(Unique Word)。如果任兩個詞在文件中曾經相鄰出現，則此兩個頂點可以用每一個邊(Edge)相連，邊的方向表示這兩個詞出現時的先後次序。最後，統計有向圖中每一個頂點的內分支度(In-degree)個數，並與 BM25 模型相結合，即可求得文件與語句間的關聯程度。相較於大多數現存的摘要模型(例如 TF-IDF 與 BM25)，僅考慮每一個詞出現的頻率，詞權重-逆向文件頻率模型基於內分支度個數，重新賦予每一個詞一個權重，進一步地考慮了文字間在文件中的先後次序關係。

B. 馬可夫隨機漫步(Markov Random Walk, MRW)

馬可夫隨機漫步模型的概念是將文件視為一個網際網路，文件中的每一語句代表網路上的一個節點(Node)，而語句之間的相關程度則為節點間邊界(Edge)的權重(Wan & Yang, 2008)。馬可夫隨機漫步模型提出一套遞迴更新的演算法，利用節點間邊界的權重關係不斷地重複更新節點的重要性，最終獲得每一語句的重要性分數。更明確地，語句S_i的重要性分數$SenScore(S_i)$是由相鄰的語句S_j分數的線性組合而得，我們可以將語句間邊界的權重關係以一個矩陣$\widetilde{\mathbf{M}} = \left(\widetilde{M}_{i,j}\right)_{|D| \times |D|}$表示之，則馬可夫隨機漫步模型可以表示為：

$$\widetilde{M}_{i,j} = \begin{cases} \dfrac{sim(S_i, S_j)}{\sum_{k=1}^{|D|} sim(S_i, S_k)} & ,if \sum_{k=1}^{|D|} sim(S_i, S_k) \neq 0 \\ 0 & ,otherwise \end{cases} \quad (5)$$

$$SenScore(S_i) = \mu \cdot \sum_{j \neq i} SenScore(S_j) \cdot \widetilde{M}_{j,i} + \frac{(1-\mu)}{|V|} \quad (6)$$

其中|*D*|為文件 *D* 中語句的個數，$sim(\cdot,\cdot)$為相似度函數，用以計算兩個語句之間的相似程度。

3. 表示法學習(Representation Learning)

3.1 詞表示法(Word Representation)

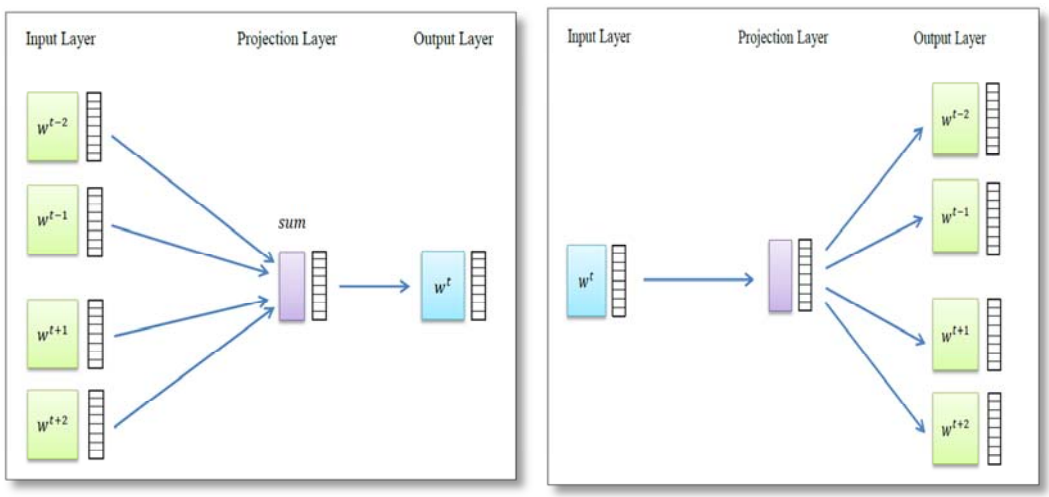
當一種自然語言處理的問題要轉化為機器學習的問題，首先需要找到一種方法將這些語言符號數學化。傳統的自然語言處理中最直觀的方式是採用 One-hot 表示法，即每一個詞皆以一個 *K* 維(通常 *K* 為詞彙的大小)的向量表示之，而此向量中僅有某一個維度為 1，其餘為零。明顯地，此種表示法中任意兩個詞之間彼此互相獨立，意即我們無法計算出任兩個詞之間的相似程度。為了解決上述問題，學者 Hinton 首先於 1986 年提出了一種分散式表示法(Distributed Representation)模型(Hinton, 1986)，藉由訓練將每一個詞重新以一個較低維度的實數向量表示之，透過這個低維度的向量表示法，詞與詞之間的關係可以簡單地透過距離公式(如餘弦、歐式距離)來計算，並依此判斷詞與詞之間語意的相近程度。學者 Bengio 等人於 2003 年提出以前饋式類神經網路(Feed-Forward Neural Network, FFNN)來建立語言模型，在語言模型的建構過程中，每一個詞即會獲得一個低維度的實數向量表示法(Bengio *et al*., 2003)。Google 於 2013 年開發出一套詞表示法工具 word2vec，當中包含連續型詞袋模型(Continuous Bag-of-Words, CBOW)(Mikolov *et al*., 2013a)與跳躍式模型(Skip-Gram, SG)(Mikolov *et al*., 2013b)。據我們所知，這些方法雖然已被使用來解決許多自然語言相關的問題，但卻鮮少被應用於語音文件摘要的任務之中。

A. 連續型詞袋模型(Continuous Bag-of-Words, CBOW)

連續型詞袋模型(CBOW)是由 Mikolov 所提出的兩個經典架構之一，該模型是設法直接地獲得每個詞的向量表示，而不是尋求學習一個統計語言模型(Mikolov *et al*., 2013a)。CBOW 的架構類似於前饋式類神經網路(Feed-Forward Neural Network)，不同之處在於(1)CBOW 移除非線性隱藏層(Non-Linear Hidden Layer)。如此一來，大大的降低了類神經網路模型訓練時間過長的問題，實驗結果顯示，簡化的模型在許多應用中，依然保有優異的性能。(2)每個詞皆共享投影層(Projection Layer)，因此所有的詞皆會投影至相同的位置(即向量相加)，這樣的架構使得詞序列不會影響投影的結果。更明確地，CBOW 的訓練目標是在給定一個詞的上下文(Context)後，期望可以準確地預測該詞的出現，其圖形表示如圖 1(a)所示。該模型不同於傳統的詞袋模型，它是使用連續分散式表示法

(Continuous Distributed Representation)。形式上，給定一詞序列$w_1 w_2 \dots w_T$，CBOW 的目標函數(Objective Function)是要最大化對數機率(Log-Probability)：

$$\sum_{t=1}^{T} logP(w^t | w^{t-c}, \dots, w^{t-1}, w^{t+1}, \dots, w^{t+c}) \tag{7}$$



(a) (b)

圖 1 (a). 連續型詞袋模型之示意圖 ***(b). 跳躍式模型之示意圖***

其中 c 為中間詞w_t的上下文之窗口大小(Window Size)，T 代表訓練語料的長度，且

$$P(w^t | w^{t-c}, \dots, w^{t-1}, w^{t+1}, \dots, w^{t+c}) = \frac{exp(v_{\overline{w}^t} \cdot v_{w^t})}{\sum_{i=1}^{V} exp(v_{\overline{w}^t} \cdot v_{w_i})} \tag{8}$$

其中v_{w^t}為位置 t 詞w的詞表示法，V 是詞彙的大小，$v_{\overline{w}^t}$代表w^t的上下文詞表示法之加權(Qiu *et al.*, 2014)。CBOW 的概念是透過一分佈式假設(Miller & Charles, 1991)，此假設指出具有類似語意的詞會經常出現於類似的上下文之中，因此建議找出可以很好地獲取上下文分佈的w^t詞表示法。

B. 跳躍式模型(Skip-Gram, SG)

跳躍式模型(SG)是由學者 Mikolov 等人於 2013 年時所提出的另一經典架構(Mikolov *et al.*, 2013b)，該模型同樣以簡化的前饋類神經網路來學習詞表示法。更明確地，跳躍式模型與連續型詞袋模型的模型訓練目標恰好相反，跳躍式模型是希望在給定一個詞w後，可以準確地預測其上下文中，詞出現的可能性。訓練的過程中，該模型是使用每一當前詞做為對數線性分類器(Log-Linear Classifier)的輸入，並預測此當前詞一定範圍內的前後的詞，其圖形表示如圖 1(b)所示。當給定一詞序列$w_1 w_2 \dots w_T$後，SG 的目標函數是要最大化對數機率：

$$\sum_{t=1}^{T}\sum_{j=-c,j\neq 0}^{c} logP(w^{t+j}|w^t) \tag{9}$$

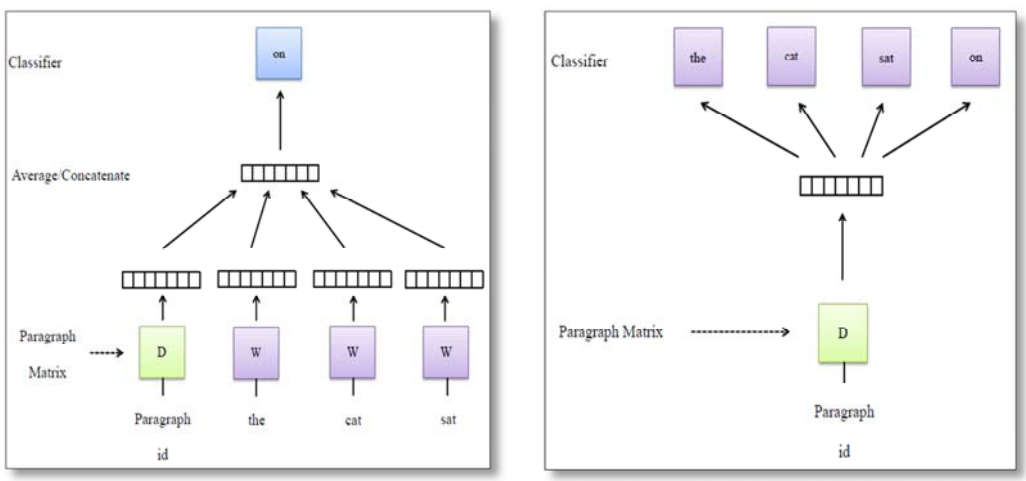
其中 c 為中間詞w_t的上下文之窗口大小(Window Size)，而條件機率(Conditional Probability)經由下式計算：

$$P(w^{t+j}|w^t) = \frac{exp(v_{w^{t+j}} \cdot v_{w^t})}{\sum_{i=1}^{V} exp(v_{w_i} \cdot v_{w^t})} \tag{10}$$

其中w^{t+j}與w^t分別為位置$t+j$及 t 的詞表示法。在 CBOW 與 SG 的實作中皆引入階層軟式最大化法(Mikolov *et al.*, 2013b; Morin & Bengio, 2005)及負例採樣法(Mikolov *et al.*, 2013b; Mnih & Kavukcuoglu, 2013)，以增進訓練過程中參數估測的效能。

3.2 語句表示法(Sentence Representation)

雖然詞表示法已被廣泛使用，但許多自然語言處理的相關任務所需要的是語句的表示法。延續詞表示法的基本模型架構與精神，學者 Le 與 Mikolov 提出兩種學習語句表示法的模型，分別是分散式儲存模型與分散式詞袋模型(Le & Mikolov, 2014)。



(a) (b)

圖 2 (a). 分散式儲存模型之示意圖 ***(b). 分散式詞袋模型之示意圖***

A. 分散式儲存模型(Distributed Memory Model of Paragraph Vector, PV-DM)

分散式儲存模型(PV-DM)類似於連續型詞袋模型。PV-DM 同樣以最大化目標中間詞輸出的機率為目標，其主要差異為：(1)訓練過程中於輸入層(Input Layer)引入一個段落編號(Paragraph ID)，亦即訓練語料中每一語句皆有一個唯一的段落編號。段落編號與一般的

詞相同，亦是先映射成一個向量，即段落向量(Paragraph Vector)。然而段落向量與詞向量的維度相同。在往後的計算中，將詞向量與段落向量串聯作為輸出層軟式最大化法(Soft-max)的輸入。在一個語句或是文件的訓練過程中，段落編號會保持不變，共享相同的段落向量，相當於每次在預測一個詞的機率時，皆利用了整個語句的語意。(2)在預測階段時，給予待預測的語句分配一個新的段落編號，保持詞向量與輸出層軟式最大化於訓練階段所得之參數，重新利用隨機梯度法訓練待預測語句，待收斂完畢後，即得到待預測語句的段落向量，其圖形表示如圖 2(a)所示。

B. 分散式詞袋模型(Distributed Bag-of-Words of Paragraph Vector, PV-DBOW)

分散式儲存模型(PV-DM)是採用詞向量與段落向量的平均或是串聯，進行預測一個詞。分散式詞袋模型(PV-DBOW)則是以段落向量作為輸入，從該向量對應的段落中隨機採樣詞序列作為輸出，該方法減少了輸入層的參數量。類似於跳躍式模型，其圖形表示法如圖 2(b)所示。該模型的概念簡單且僅需少量儲存空間(詞向量與輸出層軟式最大化法於訓練階段所得之參數)。

4. 運用表示法學習於語音文件摘要

4.1 餘弦相似度(Cosine Similarity)

由於向量空間模型簡單、直觀且有效，因此被廣泛地應用於各式自然語言處理的相關研究。藉助於詞表示法模型(例如 CBOW 與 SG)我們可以將文件或語句中所有詞所對應的詞表示法加總後取平均，作為該篇文件或語句的表示法：

$$v_D = \frac{\sum_{w \in D} v_w}{|D|}, \quad v_S = \frac{\sum_{w \in S} v_w}{|S|} \tag{11}$$

其中v_w為詞w的詞表示法，v_D、v_S為代表文件 D 與語句 S 的表示法，$|D|$、$|S|$為文件 D 及語句 S 長度。或是直接藉由語句表示法模型求得文件或語句的向量表示法：

$$v_D = PV_D, \quad v_S = PV_S \tag{12}$$

如此一來，文件 D 及其語句 S 皆有一固定長度的向量表示，其相關性就可藉由餘弦相似度計算而得：

$$Sim(S, D) = \frac{v_S \cdot v_D}{\|v_S\| \cdot \|v_D\|} \tag{13}$$

4.2 馬可夫隨機漫步(Markov Random Walk, MRW)

當結合各式詞與語句表示法模型於馬可夫隨機漫步模型中時，我們首先將文件與語句表示成一個固定維度的向量表示法。接著，我們使用餘弦相似度估測來計算兩兩語句間的相似程度，再透過馬可夫隨機漫步模型所提出的遞迴更新演算法，就可以求得每一語句

的重要性分數。最後，將語句依此分數遞減的方式排列後，根據事先定義的摘要比例來依序挑選語句並作為最後的輸出。

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

我們亦可以運用語言模型(LM)為基礎的方法於節錄式語音文件摘要，其實現的方式是計算文件被每一個語句模型生成的可能性 $P(D|S)$，並依此進行語句重要性之排序。利用詞表示法，我們首先定義一個以詞為基礎的語言模型。在給定一個詞w_i後，詞w_j出現機率為：

$$P(w_j|w_i) = \frac{exp\left(v_{w_j} \cdot v_{w_i}\right)}{\sum_{w_k \in V} exp\left(v_{w_k} \cdot v_{w_i}\right)} \tag{14}$$

接著，透過線性組合(Linear Combination)的方式，可以形成一個複合式的語句語言模型，而文件的生成機率就可以經由下式計算：

$$\mathrm{P}(D|S) = \prod_{w_j \in D} \left[\lambda \cdot \sum_{w_i \in S} P(w_i|S) \cdot P(w_j|w_i) + (1-\lambda) \cdot P(w_j|C) \right]^{c(w_j,D)} \tag{15}$$

其中$P(w_i|S)$為一個權重係數，代表詞w_i出現在語句中的出現的可能性；並且，為了解決資料稀疏的問題，我們透過背景語言模型$P(w_j|C)$對語句模型進行機率平滑化。另一方面，當使用語句表示法時，我們首先為每一個語句 S 建構出一個以語句表示法為基礎的語言模型，用以預測一個詞w_j發生的可能性：

$$P(w_j|S) = \frac{exp\left(v_{w_j} \cdot v_S\right)}{\sum_{w_k \in V} exp\left(v_{w_k} \cdot v_S\right)} \tag{16}$$

其中v_S是以 PV-DBOW 或 PV-DM 所求得的語句表示法。同樣地，文件的生成機率就可以經由下式計算：

$$\mathrm{P}(D|S) = \prod_{w_j \in D} \left[\lambda \cdot P(w_j|S) + (1-\lambda) \cdot P(w_j|C) \right]^{c(w_j,D)} \tag{17}$$

5. 語音文件之各種特徵簡介

文字文件內容除了提供文字訊息作為重要語句選取依據之外，語句中更包含文法(Grammar)、語意(Semantic)以及結構(Structure)等資訊，皆可視為重要的特徵。不同於文字文件，語音文件內容可能因辨識錯誤或語句邊界定義等問題，使得語句文法、語意以及結構等資訊相對較缺乏，但語音文件卻含有豐富的韻律特徵(Prosodic Features)，如語者在說話時發音的長短快慢、語氣的抑揚頓挫以及高低起伏等。因此若將這些語言學、聲韻學以及文件結構等資訊加以善用，相信有助於提升節錄式語音文件摘要的效能。

5.1 韻律特徵(Prosodic Features)

A. 音高(Pitch)

一般語者在敘述一件事情時，會以說話的高低起伏、抑揚頓挫來強調說話的內容以吸引聽者的注意，語者表達自身的感覺使得對方接受到強調的訊息，因此音高可視為一種語音中重要的資訊。

B. 能量(Energy)

能量可用來表示語者說話音量的大小，經常被視為一種可利用的重要資訊。一般語者在特別強調一件事情或是敘述重點時，會刻意地提高音量來表示強調關鍵字或是說話的內容以希望引起聽者的注意。

C. 音框長度(Duration)

類似於語句長度，語句越長所包含的資訊越多，而語句的音框長度代表語者說該語句的時間長度，因此說話時間越長的語句其包含的資訊亦越多。

D. 頻譜峰(Peak)與共振峰(Formant)

共振峰被定義為“聲譜中的頻譜峰”，差異在於母音(Vowel)有共振峰的結構，在母音發音較為清楚的音節(Syllable)，共振峰會較高。共振峰是用來描述聲學共振現象的一種概念，是決定語者特徵的主要因素。在有效頻寬範圍中會有約五個共振峰，從低頻率至高頻率依序排列為第一共振峰(F1)、第二共振峰(F2) 、第三共振峰(F3)、第四共振峰(F4)以及第五共振峰(F5)，而通常以 F1、F2、F3 較為明顯，因此通常以這三個共振峰為代表。若語者在表達某語句較為字正腔圓，希望聽眾可以聽得清楚時，該語句可能為重要語句，共振峰整體來說會較高；若是語者所含糊帶過的語句則可能為非重要語句，其共振峰整體來說會較低。

5.2 詞彙特徵(Lexical Features)

A. 雙連語言模型分數(Bigram Language Model Score)

N 連語言模型(*N*-gram Language Model)是自然語言處理常用到的方法，其假設第 *N* 個詞的出現僅與前面 *N* -1 個詞相關，模型參數則通常藉由最大化相似度估測(MLE)來求得。對一個語句的重要性估測是透過計算在語句中所出現的詞的條件機率之乘積，通常採用二連(Bigram)與三連(Trigram)語言模型。

B. 正規化雙連語言模型分數(Normalized Bigram Language Model Score)

為了避免在計算時因語句長度的影響，透過該語句長度將其雙連語言模型分數進行正規

化並做為另一項特徵。

C. 專有名詞(Named Entities)個數

根據專有名詞詞典(Lexicon)計算語句中的詞與專有名詞詞典重複的數量；其主要想法是含括愈多專有名詞的語句愈可能為重要語句。而專有名詞則包含公司名稱、 地點、人名以及時間等。

D. 停用詞(Stop Words)個數

計算語句中所包含停用詞的數量，如中文詞的“了”、“的”等詞，以及英文詞如“a”、“the”等詞，即使出現的頻率很高，但通常不具有太多資訊，因此在檢索過程中經常被濾除，不列入搜尋的考慮範圍。

5.3 關聯特徵(Relevance Features)

通常為來自不同文件摘要模型所產生的摘要特徵分數，如以統計值為基礎的向量空間模型(Vector Space Model, VSM)、以圖論為基礎的馬可夫隨機漫步(Markov Random Walk, MRW)以及以機率生成模型為基礎的語言模型(Language Model, LM)等。

6. 實驗語料及評估方法

6.1 實驗語料

本論文實驗語料為公視新聞語料(Mandarin Chinese Broadcast News Corpus, MATBN)，由中央研究院資訊所與公共電視台合作錄製整理，其錄製內容為每天一個小時的公視晚間新聞深度報導(Wang *et al*., 2005)。我們選取其中從 2001 年 11 月至 2002 年 8 月共 205 篇的新聞報導，並區分為發展集(185 篇)與測試集(20 篇)兩個部分。全部 205 篇語音文件長度約為 7.5 個小時。我們將語音文件進行人工切音處理，得到真正含有講話內容的音訊段落，再透過自動語音辨識系統進行轉寫，我們稱之為語音文件(Spoken Document, SD)，含有語音辨識錯誤與語句邊界偵測錯誤。此外，我們亦將此 205 篇語音文件透過人工聽寫的方式產生出沒有辨識錯誤的對應文字內容，我們稱之為文字文件(Text Document, TD)。每一篇文字文件皆有三位標記專家所提供的三份摘要結果，我們將此作為語音文件與文字文件的正確摘要答案。透過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的影響。本研究的背景語言模型訓練語料取材自 2001 至 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。此外，本論文蒐集 2002 年中央通訊社的 101,268 篇同時期新聞文件作為詞表示法以及語句表示法的訓練語料以及虛擬關聯文件。我們設定摘要比例為 10%，其定義是摘要字數占整篇文件字數的比例，其詳細的統計資訊如表 1 所示。

表 1. 廣播新聞文件之統計資訊

	訓練集	測試集
紀錄時段	2001/11/07-2002/08/22	2002/01/24-2002/08/20
文件個數	185	20
文件平均持續秒數	129.4	141.3
文件平均詞個數	326.0	290.3
文件平均語句個數	20.0	23.3
文件平均詞錯誤率	38.0%	39.4%

6.2 評估方法

本論文採用 ROUGE 作為文件摘要的評估方式。該方法是計算自動摘要結果與人工摘要之間的重疊單位元(Overlap Units)數目占人工摘要長度的比例。由於該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，且適合用於多份人工摘要的評估。我們使用了較普遍的 ROUGE-1(Unigram)、ROUGE-2(Bigram)以及 ROUGE-L(Longest Common Subsequence, LCS)分數，其中 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。ROUGE-N 是自動摘要和人工摘要之間 N 連詞(N-gram)的召回率，人工標記的參考摘要為一集合 $\boldsymbol{R}$，故 ROUGE-N 計算公式如下(Lin, 2004)：

$$\text{ROUGE}-N=\frac{\sum_{sum\in \boldsymbol{R}}\sum_{gram_N\in sum} Count_{match}(gram_N)}{\sum_{sum\in \boldsymbol{R}}\sum_{gram_N\in sum} Count(gram_N)} \tag{18}$$

其中 sum 為人工摘要集合 $\boldsymbol{R}$ 中的任一個摘要，N 代表詞彙串之連續長度，而 $Count(gram_N)$ 是 N 連詞同時出現於自動摘要與人工摘要的最大數量。ROUGE-L 的計算方式與 ROUGE-N 相似，但前者僅考慮自動摘要與參考摘要的最長共同字串。

7. 實驗結果

7.1 基礎文件摘要之實驗結果

表 2 為測試集中的文字文件(TD)與語音文件(SD)在 ROUGE-1、ROUGE-2 以及 ROUGE-L 評估下的摘要結果；在此我們進行各式的基礎摘要方法的比較，包含前導方法(LEAD)、向量空間模型(VSM)、最大邊際關聯法(MMR)、潛藏語意分析(LSA)、單連語言模型(ULM)、關聯模型(RM)、Okapi Best Match 25(BM25)、詞權重-逆向文件頻率(TW-IDF)以及馬可夫隨機漫步(MRW)。首先在 TD 的實驗中，RM 的摘要效果是所有模型中最佳的，表示使用額外的關聯文件可以有效地彌補語句內容的不足，提高語句的估測能力。其次為 BM25，我們認為在文件摘要的問題中，詞彙的頻率(TF)、反文件頻率(IDF)以及文件長度的正規化(Normalized)是重要且不可或缺的特徵資訊。ULM 無論在 TD 或是 SD 上的摘要成效皆

優於圖論式模型 TW-IDF 與 MRW。TW-IDF 在計算詞頻(TF)時，多考慮了上下文(Context)的資訊，而 MRW 在計算重要語句時，除了使用其它語句的分數之外，亦考慮到語句彼此之間的相關度作為權重來調整，因此兩者效果皆會較僅考慮詞頻的 VSM 為佳。MMR 在進行語句選取時多考慮了冗餘資訊，因此摘要效果較 VSM 佳。

表 2. 基礎實驗於文字文件與語音文件之摘要結果

	文字文件（TD）			語音文件（SD）		
方法	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
LEAD	0.312	0.196	0.278	0.254	0.117	0.220
VSM	0.347	0.228	0.290	0.343	0.189	0.288
MMR	0.365	0.242	0.316	0.360	0.206	0.309
LSA	0.362	0.233	0.316	0.345	0.201	0.301
ULM	0.411	0.299	0.362	0.364	0.218	0.313
RM	0.458	0.345	0.408	0.384	0.236	0.330
BM25	0.422	0.317	0.380	0.394	0.251	0.341
TW-IDF	0.374	0.260	0.317	0.322	0.164	0.270
MRW	0.415	0.296	0.357	0.339	0.194	0.289

LSA 在潛藏語意空間計算文件與語句的餘弦相似度，其結果亦顯示較 VSM 為佳。而 VSM 每個詞彙所構成的向量維度皆為獨立，因此無法得知出文件中詞彙之間的關聯性，使得進行文件相似度的比對時可能造成誤判的情況。

在 SD 的實驗中，BM25 反而超越 RM 成為所有模型中最佳的摘要方法，我們認為這可能是因為 RM 中所使用的語句模型受到語音辨識錯誤的影響，因此降低尋找有效的虛擬關聯文件(Pseudo Relevant Documents)的能力。此外，TW-IDF 與 MRW 的摘要效能皆較 LSA 及 MMR 差，我們認為亦是受到語音辨識錯誤的影響，因一個詞或是一個語句的重要性分數是來自鄰近其它詞或是語句的貢獻。而 LEAD 無論在 TD 或是 SD 上，相較於其它模型皆得到較差的效果，主要原因是 LEAD 僅適用於特殊文件結構，因此若摘要文件不具有某種特殊的結構，其摘要效能就會有所侷限。

7.2 詞表示法與語句表示法於節錄式語音文件摘要之實驗結果

在此我們利用目前兩種最先進的詞表示法—連續型詞袋模型(CBOW)和跳躍式模型(SG)，與最先進的兩種語句表示法—分散式儲存模型(PV-DM) 和分散式詞袋模型(PV-DBOW)之技術來從事語音文件摘要；實驗共分三組來進行，分別結合於餘弦相似度(Cosine Similarity)、馬可夫隨機漫步(MRW)以及文件相似度量值(DLM)的方法作為挑選摘要語句之依據。

表 3. 詞表示法結合於餘弦相似度之摘要結果

	文字文件(TD)			語音文件(SD)		
方法	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
CBOW	0.402	0.280	0.349	0.377	0.228	0.327
SG	0.401	0.265	0.347	0.361	0.214	0.312

首先，我們將詞表示法結合於餘弦相似度(Cosine Similarity)作為選取摘要語句的方法，其結果示於表 3。從實驗結果中觀察到，由於這兩種詞表示法各有著不同的模型結構與學習方式，因此在文字文件(TD)或是語音文件(SD)中，該兩種模型的摘要成效有稍微的差異。根據 TD 的結果顯示，CBOW 的摘要效能較 SG 佳，在 SD 中仍保持相同的情況。儘管該兩種詞表示法皆優於向量空間模型(VSM)與潛藏語意分析(LSA)，卻僅達到詞權重-逆向文件頻率(TW-IDF)差不多的水平，而且在 SD 的情況下的表現 SG 不及單連語言模型(ULM)(表 2)。

表 4. 語句表示法結合於餘弦相似度之摘要結果

	文字文件(TD)			語音文件(SD)		
方法	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
PV-DM	0.429	0.313	0.382	0.387	0.236	0.335
PV-DBOW	0.398	0.277	0.348	0.368	0.227	0.329

同樣地，我們將語句表示法結合於餘弦相似度作為選取摘要語句的方法，其結果示於表 4。在 TD 的結果中，PV-DM 與 PV-DBOW 該兩種語句表示法的摘要效果分別超越 CBOW 及 SG 詞表示法模型(表 3)。PV-DM 摘要成效較傳統的馬可夫隨機漫步(MRW)佳，但較 BM25 差。而在 SD 的結果中，兩種語句表示法的摘要成效比起詞表示法沒有太大的進步，我們認為語句表示法搭配餘弦相似度選取語句的方式亦受語音辨識的影響。

表 5. 詞表示法結合於馬可夫隨機漫步之摘要結果

	文字文件(TD)			語音文件(SD)		
方法	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
CBOW	0.436	0.310	0.384	0.393	0.246	0.346
SG	0.316	0.283	0.351	0.372	0.233	0.325

在第二組實驗中，我們將詞表示法結合馬可夫隨機漫步(MRW)以對語句進行選取，其結果呈現在表 5。從結果中可以觀察到，無論在 TD 或是 SD 上，相較於同樣以詞表示法的技術結合餘弦相似度的方法，使用該方法挑選語句的摘要成效皆優於以餘弦相似度的方式(表 3)。在 TD 實驗中，CBOW 摘要效能較 BM25 差，而 SG 未達到 MRW 的水平。在 SD 實驗中，仍然以 BM25 的摘要效果為佳。

表 6. 語句表示法結合於馬可夫隨機漫步之摘要結果

	文字文件(TD)			語音文件(SD)		
方法	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
PV-DM	0.446	0.343	0.400	0.395	0.253	0.347
PV-DBOW	0.451	0.336	0.398	0.387	0.243	0.337

同樣地，我們以語句表示法結合馬可夫隨機漫步(MRW)對語句進行選取，其結果展示於表 6。從結果中發現到，無論在 TD 或是 SD 上，該方法的摘要成效，顯著地優越以詞、語句表示法結合於餘弦相似度(表 3 和 4)之選取摘要語句方法，亦超越以詞表示法結合於馬可夫隨機漫步的方式(表 5)。在 TD 實驗中，儘管該兩種詞表示法的摘要成效較 BM25 佳，但皆不及關聯模型(RM)。然而於 SD 實驗中，PV-DM 的摘要成效超越所有的傳統文件摘要模型。

表 7. 詞表示法結合於文件相似度量值之摘要結果

	文字文件（TD）			語音文件（SD）		
方法	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
CBOW	0.444	0.329	0.386	0.372	0.221	0.314
SG	0.436	0.323	0.385	0.343	0.197	0.295

在最後一組實驗中，我們探討以詞表示法結合於文件相似度量值(DLM)對語句進行選取，其結果展示於表 7。我們將結果與同樣以詞表示法結合餘弦相似度(表 3)以及馬可夫隨機漫步的方法(表 5)進行比較。從 TD 實驗結果中可以觀察到，文件相似度量值充分地運用詞表示法於文件摘要，表現顯然較佳。我們亦注意到 SG 的摘要成效幾乎接近 CBOW。然而於 TD 與 SD 的實驗中，該兩種詞表示法皆仍不及 RM 的摘要成效。

表 8. 語句表示法結合於文件相似度量值之摘要結果

	文字文件(TD)			語音文件(SD)		
方法	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
PV-DM	0.480	0.375	0.430	0.384	0.240	0.333
PV-DBOW	0.433	0.323	0.384	0.364	0.236	0.321

同樣地，我們以語句表示法於文件相似度量值對語句進行選取，其結果顯示在表 8。從 TD 的實驗結果中可以觀察到，PV-DM 的摘要效能顯著地優於表 2 中所有的傳統文件摘要模型，亦是所有表示法中具最佳摘要效能之模型。我們亦觀察到 PV-DBOW 與表 7 中的詞表示法 SG 有著相同的摘要成效。然而於 SD 中，該兩種語句表示法僅達到 RM 的水平，但皆仍不及 BM25。

7.3 利用聲學特徵結合支持向量機於文件摘要

本論文所使用的語音語料是經由人工切音，不會有語音邊界錯誤的問題，僅須考量語音辨識錯誤於文件摘要的影響，因此文字文件(TD)與語音文件(SD)兩者會有相同的語音邊界，而抽取出的韻律特徵亦會是一致。本論文總共使用 12 種不同的摘要特徵作為支持向量機(Support Vector Machine, SVM)的輸入，可概略分成三大類，分別為詞彙特徵(Lexical Features)、韻律特徵(Prosodic Features)以及關聯特徵(Relevance Features)，詳細的特徵資訊如表 9 所示。

表 9. 實驗採用之各式特徵

韻律特徵(Prosodic Features)	音高(Pitch):最大、最小、平均、差值 能量(Energy):最大、最小、平均、差值 音框長度(Duration):最大、最小、平均、差值 共振峰(Formant):最大、最小、平均、差值 頻譜峰值(Peak):最大、最小、平均、差值
詞彙特徵(Lexical Features)	專有名詞個數(Named Entity) 停用詞個數(Stop Word) 二連語言模型分數(Bigram) 正規化二連語言模型分數(Normalized Bigram)
關聯特徵(Relevance Features)	向量空間模型分數(VSM) 馬可夫隨機漫步分數(MRW) 語言模型分數(LM)

由表 10 中得到，無論在文字文件(TD)或是語音文件(SD)中，韻律特徵(Prosodic Features)相對於其它兩種特徵產生較為顯著的摘要效能，因此韻律特徵比起其它兩種特徵更能夠判斷摘要語句的重要資訊。在 TD 實驗中，詞彙特徵(Lexical Features)在這三種摘要特徵中的表現最差，其原因可能是該特徵描述的是表淺(Shallow)語句性質，包含專有名詞的數量、停用詞的數量以及語句的流暢性，沒有考慮語句的語意內容，因此單憑該特徵無法選取出較正確的摘要語句。此外，關聯特徵(Relevance Features)比起詞彙特徵有較好的摘要成效。在 SD 實驗中得到的結論，與 TD 的結論具一致性，但關聯特徵與韻律特徵之間效果差異較無 TD 來得顯著。

表 10. 單類特徵之摘要結果

	文字文件(TD)			語音文件(SD)		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
韻律特徵	0.452	0.349	0.409	0.363	0.219	0.322
詞彙特徵	0.362	0.237	0.311	0.298	0.176	0.266
關聯特徵	0.389	0.254	0.332	0.355	0.200	0.300

我們進行使用所有摘要特徵於支持向量機器(Support Vector Machine, SVM)之實驗，其結果示於表 11。從實驗結果中可以發現，無論於 TD 或是 SD 中，經過各種面向的考量後，確實可以獲得較好的摘要成效。接著進行探討關聯特徵中使用其它模型分數對摘要效能的影響。因此我們將關聯特徵中的向量空間模型(VSM)、馬可夫隨機漫步(MRW)以及單連語言模型(ULM)的分數，以詞表示法模型摘要之分數作為替換，分別根據於表 3、5 和 7 中最佳的摘要表現，從各表中可以發現 CBOW 的摘要效果始終最佳。

表 11. 結合所有特徵之摘要結果

	文字文件(TD)			語音文件(SD)		
方法	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
所有特徵	0.484	0.384	0.440	0.387	0.247	0.348

同樣地結合所有特徵一併做為支持向量機的輸入，其摘要效能如表 12 所示。從實驗結果中發現到，無論在 TD 或是 SD 中，以詞表示法模型作為關聯特徵，皆使得摘要成效非常顯著，尤其在 TD 中的實驗結果，產生最佳之摘要成效。

表 12. 以詞表示法模型摘要分數為關聯特徵之摘要結果

	文字文件(TD)			語音文件(SD)		
方法	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
所有特徵	0.497	0.406	0.451	0.396	0.254	0.353

我們亦考慮語句表示法模型分數對摘要效能的影響。同樣將關聯特徵中的模型分數替換為語句表示法模型摘要之分數，分別根據於表 4、6 和 8 中最佳的摘要表現，從各表中的結果可觀察到 PV-DM 的摘要效果始終最佳；其摘要成效如表 13 所示。從 TD 的實驗結果中可以觀察到，使用語句表示法模型分數作為特徵之摘要成效較使用詞表示法來得差(表 12)。然而在 SD 中，結合以語句表示法模型分數作為關聯特徵可以達到最佳之摘要效果。

表 13. 以語句表示法模型摘要分數為關聯特徵之摘要結果

	文字文件(TD)			語音文件(SD)		
方法	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
所有特徵	0.487	0.393	0.446	0.385	0.255	0.350

8. 結論與未來展望

過去在自動文件摘要的研究主要仍著重於文字文件摘要，直到 1990 年後期，由於影音多媒體技術的進步與成熟，才逐漸開始有語音文件摘要的研究。文件摘要可分為節錄式摘要與抽象式摘要，本論文旨在探討節錄式中文廣播新聞文件摘要方法。我們提出兩種詞表示法—連續型詞袋模型(CBOW)和跳躍式模型(SG)，以及兩種語句表示法—分散式儲

存模型(PV-DM)和分散式詞袋模型(PV-DBOW)於文件摘要的應用；透過表示法學習的技術能將詞之間、語句之間的關聯性表現出來，用以幫助選取語音文件中重要的摘要語句。經由一連串的實驗分析與討論，證明所提之方法的確可以較其它基礎實驗的摘要方法獲得更高的摘要效能。此外，我們除了利用文字文件的詞彙特徵及關聯特徵之外，亦利用語音訊號中之韻律特徵，希望能對摘要語句的選取提供更多有幫助的資訊。

未來，我們考慮其它先進的詞表示法，如全域向量(Global Vectors, GloVe)，以及希望可以運用詞性(Part of Speech, POS)資訊的詞性表示法(POS Representation)於語音節錄式文件摘要，並且將詞、語句以及詞性表示法結合於其它的語言模型之中，如關聯模型(Relevance Model)(Lavrenko & Croft, 2001)，以進一步地提升摘要成效。

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 104-2221-E-003-018-MY3 和 MOST 103-2221-E-003-016-MY2)之經費支持，謹此致謝。

參考文獻

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3, 1137-1155.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the Annual International ACM 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.

Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval*, 19-25.

Hajime, M., & Manabu, O. (2000). A comparison of summarization methods based on task-based evaluation. In *Proceedings of the International Conference on Language Resources and Evaluation*, 633-639.

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the Annual Conference of the Cognitive Science Society*, 1-12.

Jones, K. S. (1999). Automatic summarising: factors and directions. *Advances in Automatic Text Summarization*, 1-12.

Lavrenko, V., & Croft, W. B. (2001). Relevance-based language models. In *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval*, 120-127.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning*.

Lin, C. Y. (2004). ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out.*

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2*(2), 159-165.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations*, 1-12.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the International Conference on Learning Representations*, 1-9.

Miller, G., & Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1-28.

Mitra, M., Singhal, A., & Buckley, C. (1997). Automatic text summarization by paragraph extraction. In *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization*, 39-46.

Mnih, A., & Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2265-2273.

Morin, F., & Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 246-252.

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval*, 275-281.

Qiu, L., Cao,Y., Nie, Z., & Rui, Y. (2014). Learning word representation considering proximity and ambiguity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1572-1578.

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, *27*(3), 129-146.

Robertson, S. E., Walker, S., Jones, K. S., Hancock-Beaulieu, M., & Gatford, M. (1996). Okapi at TREC-4. In *Proceedings of the Fourth Text Retrieval Conference*, 73-97.

Robertson, S. E., & Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval*, 232-241.

Rousseau, F., & Vazirgiannis, M. (2013). Graph-of-word and TW-IDF: New approach to Ad hoc IR. In *Proceedings of the International Conference on Conference on Information, Knowledge Management*, 59-68.

Salton, G., & Lesk, M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM, 15*(1), 8-36.

Wan, X., & Yang, J. (2008). Multi-document summarization using cluster-based link analysis. In *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval*, 299-306.

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