Computational Linguistics and Chinese Language Processing
Vol. 25, No. 1, June 2020, pp. 29-56

基於端對端模型化技術之語音文件摘要

Spoken Document Summarization Using End-to-End Modeling Techniques

劉慈恩*+、劉士弘#、張國韋*、陳柏琳+

Tzu-En Liu, Shih-Hung Liu, Kuo-Wei Chang, and Berlin Chen

摘要

本論文主要探討端對端(End-to-End)的節錄式摘要方法於語音文件摘要任務上的應用，並深入研究如何改善語音文件摘要之成效。因此，我們提出以類神經網路為基礎之摘要摘要模型，運用階層式的架構及注意力機制深層次地理解文件蘊含的主旨，並以強化學習輔助訓練模型根據文件主旨選取並排序具代表性的語句組成摘要。同時，我們為了避免語音辨識的錯誤影響摘要結果，也將語音文件中相關的聲學特徵加入模型訓練以及使用次詞向量作為輸入。最後我們在中文廣播新聞語料(MATBN)上進行一系列的實驗與分析，從實驗結果中可驗證本論文提出之假設且在摘要成效上有顯著的提升。

Abstract

This thesis set to explore novel and effective end-to-end extractive methods for spoken document summarization. To this end, we propose a neural summarization approach leveraging a hierarchical modeling structure with an attention mechanism to understand a document deeply, and in turn to select representative sentences as

*中華電信研究院巨量資料研究所
Big Data Laboratory, Telecommunication Laboratories, Chunghwa Telecom Co., Ltd
E-mail: hane0131@gmail.com; muslim@cht.com.tw
+國立臺灣師範大學資訊工程研究所
Department of Computer Science and Information Engineering, National Taiwan Normal University
E-mail: hane0131@gmail.com; berlin@csie.ntnu.edu.tw
#台達知識管理部
Delta Management System
E-mail: journey.liu@deltaww.com

its summary. Meanwhile, for alleviating the negative effect of speech recognition errors, we make use of acoustic features and subword-level input representations for the proposed approach. Finally, we conduct a series of experiments on the Mandarin Broadcast News (MATBN) Corpus. The experimental results confirm the utility of our approach which improves the performance of state-of-the-art ones.

關鍵詞：語音文件、節錄式摘要、類神經網路、階層式語意表示、聲學特徵
Keywords: Spoken Documents, Extractive Summarization, Deep Neural Networks, Hierarchical Semantic Representations, Acoustic Features

1. 緒論 (Introduction)

隨著大數據時代的來臨，巨量且多元的資訊透過網際網路快速地在全球各地傳播，資料內容的呈現方式已不侷限於傳統的紙本形式，包含語音及影像的多媒體資訊逐漸取代靜態的文字資訊，如何有效率地閱讀多樣化形式的多媒體資訊，已成為一個刻不容緩的研究課題。此外，在社會逐步行動化的情況下，人手一機已是常態，且伴隨著科技不斷地創新，行動設備不再只能通話和傳遞文本訊息，多媒體訊息如語音及影像等亦能完好地傳遞，更甚於我們能透過聲音及手勢等指令操作設備。

在眾多的研究方法中，自動摘要 (Automatic Summarization) 被視為是一項關鍵的技術，其在自然語言處理 (Natural Language Processing, NLP) 領域中一直都是熱門的研究議題，因其具有能擷取文件重要資訊的特性，在許多應用上更是不可或缺的一項技術，如問答系統 (Question Answering)、資訊檢索 (Information Retrieval) 等。另一方面，語音是多媒體文件中最具語意的主要成份之一，如何透過語音(文件)摘要技術有效率地處理時序資料，更是顯得非常重要。其關鍵在於影音文件往往長達數分鐘或數小時，使用者不易於瀏覽與查詢，而必須耗費許多時間閱讀或聆聽整份文件，才能理解其內容，不符合人們想要快速地獲取資訊之目的。

對於含有語音訊號的多媒體資訊，我們可先經由自動語音辨識 (Automatic Speech Recognition, ASR) 技術將文件轉成易於瀏覽的文字內容，再透過文字文件摘要的技術作處理，以達到摘要語音文件之目的。但因現階段的語音辨識技術仍存在辨識錯誤的問題，也缺乏章節與標點符號，使得語句邊界定義模糊而失去文件的結構資訊；此外，語音文件通常含有一些口語助詞、遲疑、重複等內容，進而使得語音摘要技術的發展更為艱鉅。

本論文主要探討端對端的節錄式語音文件摘要任務常見的自動摘要技術大致上可分為兩種，節錄式 (Extractive) 摘要與重寫式 (Abstractive) 摘要。節錄式摘要方法是本論文的研究重點，其主要會辨別文章中的語句是否具代表性，並依照特定的摘要比例從其中選取作為摘要；重寫式摘要方法則需理解文章後，依文章的主旨重新撰寫摘要，其所使用的詞彙與文法不全然從原文中複製，與人們日常撰寫的摘要較為相似。

常見的語音文件摘要任務主要是分為兩階段，自動語音辨識（Automatic speech recognition, ASR）和自動文件摘要（Automatic document summarization）。當我們得到

一語音文件，自動語音辨識系統會先對語音訊號進行特徵抽取，進而透過預先訓練完成之聲學模型（Acoustic model）和語言模型（Language model）進行語音辨識得到其轉寫文件 （Transcription）。本論文中所使用的語音辨識系統，是採用國立臺灣師範大學資訊工程學系研究所語音暨機器智能實驗室所發展之大詞彙語音辨識器（Large vocabulary continuous speech recognition system, LVCSR）(Chen, Kuo & Tsai, 2004; 2005) 進行自動語音辨識。常見的節錄式文件摘要方法大多是以資料驅動 (Data-driven) 方法為主。其中，又以深度學習 (Deep Learning) 方法發展出的序列對序列 (Sequence-to-Sequence) 架構 (Bahdanau, Cho & Bengio, 2015; Sutskever, Vinyals & Le, 2014)在摘要任務上獲得較多學者的青睞。尤其重寫式摘要被認為是一種序列對序列的問題(Sutskever *et al.*, 2014)，更以此發展出許多方法(Chen, Zhu, Ling, Wei & Jiang, 2016; Chopra, Auli & Rush, 2016; Nallapati, Zhou, dos Santos, Gùlçehre & Xiang, 2016; Paulus, Xiong & Socher, 2017; Rush, Chopra & Weston, 2015; See, Liu & Manning, 2017; Tan, Wan & Xiao, 2017)；而節錄式摘要一般則被視為一種序列標記 (Sequence Labeling) 的問題，對文章中每個語句作標記，標示出其是否為摘要(Cheng & Lapata, 2016; Nallapati, Zhai & Zhou, 2017)。

雖然語音辨識的錯誤對於語音文件摘要任務上會有一定的影響，其主要的影響在於自動轉寫文件中的內文會與人工轉寫結果有差異，進而導致文件摘要系統無法完全準確地理解文件含義，因此使得摘要成效不佳；此外，摘要的呈現亦是一項重要的課題，如何呈現出易於閱讀的摘要，是文件摘要系統中必須學會的重點。而一個良好的摘要表達應該著重於以下四個要素：

- **資訊性（Informativity）**：摘要結果所包含原文件中的資訊程度，應盡可能涵蓋所有重要資訊。
- **文法性（Grammaticality）**：摘要中的語句應符合語言的文法，所得之摘要才易於閱讀；若不符合文法，則會較常被視為關鍵詞擷取（Keyword Extraction）。此要素於重寫式摘要任務上較受關注。
- **連貫性（Coherency）**：此要素所指的是摘要中上下文間的連貫程度，若前後句不存在連貫性，則會類似於畫重點的方式條列出重點，而非根據文件主旨所生成之摘要。此要素於節錄式摘要任務上常被提及。
- **非重複性（Non-Redundancy）**：為了能簡化描述，應避免出現過多重複的詞句或相似的資訊，若重複的資訊太多會影響使用者閱讀。

因此本論文主要會針對上述之資訊性及連貫性兩項要素討論，並嘗試以不同方法避免受到語音辨識錯誤的影響。首先於摘要資訊性部分，本論文發展並改進一個端對端的階層式類神經網路架構，其受益於摺積式類神經網路（Convolutional neural networks, CNNs）之語言模型應用以及遞迴式類神經網路（Recurrent neural networks, RNNs）於自然語言處理領域的優秀表現，使得我們能夠階段式（先語句後全文）地閱讀文件並快速地理解語意；另外我們亦嘗試應用注意力機制（Attention mechanism）更進一步提升模型對於文章的理解度，進而提升摘要資訊性。其次對於摘要連貫性，由於節錄式摘要往

往是挑選較符合摘要語句的結果，因此其通常沒有根據語意進行排序，因此本論文亦嘗試將摘要語句的排序及摘要評估指標應用於強化學習（Reinforcement learning, RL）輔助模型訓練。最後為了避免語音辨識錯誤，我們在模型預測摘要的過程中參考語句的聲學特徵（Acoustic features）及次詞資訊（Subword information），其中前者包含原語音文件中的語音特性，可改善兩階段語音文件摘要系統上，進行摘要時無法參考之原語音特性；而後者則是為了改善前述之詞彙辨識錯誤，因辨識錯誤可能發生在詞彙中的部分區塊，而導致斷詞時無法辨別正確的詞彙，若使用次詞資訊則可以使用周邊資訊推測錯誤的部分其正確的語意。

2. 文獻回顧 (Related Work)

圖 1. 自動文件摘要的分類
[Figure 1. Category of Automatic document summarization]

自動文件摘要方法主要可依照四個面向分類（如圖 1），可依照來源、目的、功能及方法等細分為不同類型：

- **來源**：主要分為單文件與多文件，前者指針對單一文件擷取摘要，後者則是統整歸納多篇主題相近的文件重點產生摘要。多文件摘要通常會與查詢共同進行為以查詢為主之多文件摘要，同時進行檢索與摘要。
- **目的**：可分為一般性和查詢導向，一般性的摘要主要專注在文件中的主要重點；而查詢導向則會根據查詢字串決定其摘要內容，而查詢導向的摘要通常會與多文件摘要同時出現。
- **功能**：大多數摘要是資訊性的，主要專注在產生原文件的簡短版本，能保留其重要資訊；而較少數為指示性和批判性，此二者給予的摘要皆不包含原文的重要內容，前者會指出文件的題目或領域等詮釋資料（Metadata）；而後者則是會判斷整份文件是正面的還是負面的。

- **方法**：此分類方式最為常見，可概分為三種：

  ■節錄式摘要（Summarization by extraction）

  ■重寫式摘要（Summarization by abstraction）

  ■語句壓縮式摘要（Summarization by sentence compression）

  節錄式摘要與重寫式摘要之差異在於其產生摘要的原理不同。節錄式摘要是依據固定之摘要比例(Summarization ratio)，從原文件中選出重要性高的語句、段落或章節簡單組合成摘要。摘要比例是指摘要長度與原文件長度的比例，一般我們通常選用10%的摘要比例，也就是摘要長度為原文件長度的 10%。而重寫式摘要主要會依原文件中的完整概念，重新撰寫出摘要，因此摘要內容中可能還有非原文件中所使用但不影響其語意的詞語。綜上所述，我們可以(Torres-Moreno, 2014)之示例簡單描述節錄式摘要與重寫式摘要的優缺，以學習者為例，一個好的學習者在撰寫摘要時會先閱讀過整篇文章，再以自己的方式撰寫，而得之摘要內容能前後通順且符合文章旨意；而不好的學習者在撰寫摘要時，只會大略看過文章，並且挑選出「可能」重要的語句，組合在一起作為摘要。但此方法得到之摘要可能包含某些不相關的內容，且語句間的銜接可能會有內容不連貫或不通順的情況發生。除了較常見的節錄式摘要及重寫式摘要外，語句壓縮式摘要比較特別一點，主要用於將語句長度縮減，此方法可與節錄式摘要共同使用，而目前通常會將此方法歸類為重寫式摘要的一部分。

本論文主要專注於一般性單文件節錄式摘要的研究。此外摘要亦可針對文件形式分類，如常見的文字文件(Text documents)及包含語音資訊的語音文件(Spoken documents)，針對不同文件形式，所使用的摘要模型細節也應有所變化。文字文件摘要係指一般以文字內容為主的文件產生之摘要，大部分的摘要研究都屬於文字文件摘要；而語音文件摘要則是使用含有語音資訊的文件，通常是透過語音辨識後得到的轉寫文件，其中可能會含有一些語音辨識產生之錯誤，以及口語上無意義的資訊。因此，語音文件摘要會比文字文件摘要更為困難，反之，語音文件包含語音資訊，可以提供摘要方法更多有意義的資訊，能有效地抵銷其辨識錯誤。

此外，有鑒於深層學習的蓬勃發展，現今的技術大多是以端對端的深層類神經網路架構為主。深層學習主要是模擬人類之學習模式，將深層類神經網路架構視為人類大腦神經系統，並輔以大量資料進行訓練，使其能夠學習如何解決該研究問題。其架構中主要學習的是輸入與輸出之間的關係，藉由將不同的輸入樣本投影至相同的空間中，我們即可在該空間中將每個輸入樣本對應至正確的輸出，進而得到正確的結果。因此後續之文獻探討將以端對端之深層學習方法為主。

2.1 節錄式摘要 (Extractive Summarization)

在節錄式文件摘要任務中，我們通常可以將其視為分類問題，因為我們要判斷文件中的語句「是否」為摘要。而分類問題在深層學習技術中是最基本的問題，但是節錄式摘要

任務還是有相當的難度，因為除了簡單的分類外，我們還需理解並解析出文件的重要資訊，才能知道哪些語句有機會成為摘要。

(Cheng & Lapata, 2016) 將節錄式摘要任務視為一種序列標記及排序問題，其方法主要的特色在於使用一階層式編碼器和含有注意力機制(Attention Mechanism)的解碼器。階層式的編碼器有兩層，第一層為摺積式類神經網路(Convolutional Neural Networks, CNNs)，是參考(Kim, 2014)的方法，使用 CNN 計算語句的向量表示；第二層為遞迴式類神經網路(Recurrent Neural Networks, RNNs)，將語句向量做為每個時間點的輸入，而將最後一個時間點的輸出視為文件的向量表示。此作法對於較長的文章而言是相當有效的，因為文章過長時，若單使用一個 RNN，則有可能會遺失掉許多重要的資訊。最後透過另一個 RNN 對每個語句進行標記，並使用預測出的分數進行排序，進而得到最後的摘要成果。此外，(Cheng & Lapata, 2016)還嘗試用節錄式的方法模擬出重寫式摘要，與前述標記語句的不同，主要是從原文件中挑選單詞後組合成摘要句，而生成之摘要相當不符合文法性也不通順，不過關鍵詞彙基本上都能涵蓋。以此得知，(Cheng & Lapata)的方法在語言理解(Language Understanding)及資訊擷取(Information Extraction)有不錯的成效。

除了(Cheng & Lapata, 2016)同時進行節錄式摘要與重寫式摘要的研究外，(Nallapati *et al*., 2017)提出的 SummaRuNNer 亦嘗試生成重寫式摘要。與(Cheng & Lapata, 2016)不同之處在於 SummaRuNNer 在節錄式摘要任務上，並非使用編碼-解碼器架構，僅是單純地建立兩層雙向 RNN 後便判斷語句標記為何。相似之處在於其 RNN 也是階層式的架構，第一層輸入為詞彙向量，第二層則是第一層輸出所得之語句向量。此種作法中使用的參數量較少，因此收斂速度也較為快速。除了節錄式摘要任務外，(Nallapati *et al*., 2017)也嘗試將最後一層預測標記，改為一個簡易解碼器用於重寫式摘要任務。此外，由於摘要任務使用之資料集一般是沒有摘要標記的，(Nallapati *et al*., 2017)提出一種貪婪法對每個語句標記摘要，這個方法能夠找到較好的摘要組合而非只是找單獨比對每句的重要性，亦有許多學者嘗試將此方法用於自身的任務上。

隨著近幾年強化學習(Reinforcement Learning)的熱潮，亦有學者將強化學習應用於節錄式摘要任務上，(Narayan, Cohen & Lapata, 2018a)為了解決前述之節錄式摘要沒有正確摘要標記的情況，因此加入強化學習。其主要架構是改良自(Cheng & Lapata, 2016)，不同之處在於其在第二層編碼器的語句輸入是以倒序方式輸入，因為大多數文件通常會將主旨置於較前面的段落，再加上 RNN 比較容易記得後面時間點資訊的特性，此方式能夠將重要資訊更清楚記得。(Narayan *et al*., 2018a)所使用的強化學習方法，是最基礎的策略梯度(Policy Gradient)，也就是透過計算得之獎勵(Reward)分數與模型訓練梯度加成，使其能夠往我們期待的方向進行訓練。(Narayan *et al*., 2018a)所使用的獎勵分數是使用預測摘要與標準摘要的評估分數，而此方法讓模型收斂速度增加，同時也提升準確度，是一項跳躍性地成長。

然而，對於節錄式摘要任務來說，模型對文件的理解應該要能達到支撐後續分類摘要語句的程度，意即模型所得之文件向量表示應完整涵蓋文件主旨。根據不同的撰寫方式，文件主旨可能分散於文件的不同部分，除去文件主旨的段落，文件的其他部分應為

支持主旨的相關論述。如何讓模型可以準確地理解文件主題呢？(Ren *et al*., 2017)針對此議題提出一個有效的方法，其在產生語句向量表示時，亦將前面的語句以及後面的語句與該句的相關性串接，同時放入一些與該句相關的人工特徵（語句長度、位置等），使得分類時能使用更具語意的語句向量。此方法之架構相當大，但得到之摘要效果也相當不錯。不過從實驗分析可以發現對於摘要結果有較多貢獻的部分大多在於人工特徵上，以此我們可以推論，類神經網路的學習仍需人工特徵輔助方可更加提升成效。

單單只讓類神經網路架構自動學習語句或文件向量表示的效果仍有限，若能加入一些相關的額外資訊輔助訓練，可以讓我們的方法更深入地學習到文件重要資訊。(Narayan *et al*., 2018b)提出在摘要方法中參考文件的標題資訊，可以讓我們的方法更快速地找到文件的主旨，而以此得到的文件向量表示也較能涵蓋文件主旨，因而能提升摘要的成效。而(Narayan *et al*., 2018b)主要用的基本架構是由(Narayan *et al*., 2018a)變化而成，差異在於其將額外資訊向量與語句向量共同用於判斷是否為摘要。此方法更是驗證類神經網路架構有額外資訊輔助能學習更好。

2.2 重寫式摘要 (Abstractive Summarization)

(Rush *et al*., 2015) 是最早將類神經網路架構應用於重寫式摘要的研究，其主要的架構是改良至 (Bahdanau *et al*., 2014) 提出的編碼解碼器 (Encoder-Decoder) 與注意力機制，亦稱之為序列對序列模型，並應用於重寫式摘要任務。注意力機制能讓輸入文件內容與輸出摘要中的文字作一個對應，能找到文件與摘要中詞彙間的關係。(Rush *et al*., 2015) 的架構與 (Bahdanau *et al*., 2014) 不同之處在於其並非使用遞迴式類神經網路作為編碼器與解碼器，而是使用最基本的前向式類神經網路 (Feed-forward Neural Networks) 結合注意力機制作為其編碼器，而解碼器則是基於(Bengio, Ducharme, Vincent & Jauvin, 2003) 提出的 NNLM 變化。此方法在語句摘要 (Sentence Summarization) 任務上得到相當優異的成效，因此也證實類神經網路能夠適用於重寫式摘要任務上。

隨著深層學習的快速發展，遞迴式類神經網路在序列相關任務上的成功亦漸漸廣為人知，因此(Chopra *et al*., 2016) 則提出一個遞迴式類神經網路的編碼解碼器架構，應用於語句摘要任務上。此方法主要是 (Rush *et al*., 2015) 的延伸，其編碼器使用摺積式類神經網路，而解碼器則使用長短期記憶 (Long Short-Term Memory, LSTM) (Hochreiter & Schmidhuber, 1997) 單元作為遞迴式類神經網路的基本單元。LSTM 是 遞迴式類神經網路演變的架構，因其具有三個閘門: 輸入閘 (input gate)、遺忘閘 (forget gate) 及輸出閘 (output gate)，以及一個記憶單元 (memory cell)，所以可以改善消失的梯度(Vanishing Gradient)問題，同時透過不斷更新記憶單元，能保留更多重要資訊，不會隨著時間太長而遺忘以前的資訊。

與此同時，(Nallapati *et al*., 2016) 從 (Rush *et al*., 2015) 和 (Chopra *et al*., 2016) 發想出許多架構，同時也解決許多重寫式摘要潛在的問題。基本的架構是跟(Bahdanau *et al*., 2014) 提出的序列對序列模型相似，同時也加入注意力機制，而與 (Chopra *et al*., 2016) 不同之處則是在於其編碼器與解碼器皆使用遞迴式類神經網路，且使用 (Cho *et al*., 2014)

提出的 Gated Recurrent Unit (GRU) 而非 LSTM，GRU 同樣具有閘門，但是僅有兩個，且沒有額外的記憶單元，但是整體的記憶效果是一樣的，訓練參數量減少很多，可以比 LSTM 更快速地建構和訓練。(Nallapati *et al*., 2016) 中提到在語言生成時會遇到未知詞 (Out-of-vocabulary, OOV) 問題，為了解決此問題，加入 Large Vocabulary Trick (LVT) (Jean, Cho, Memisevic & Bengio, 2014)，此技術是對每小批 (mini-batch) 訓練資料建立單獨的解碼用詞典，因此能夠讓詞典不會太大，同時又能在訓練的時候減少發生未知詞問題。除了基本架構外，還提出三種改良的版本，第一種是在輸入時加入一些額外的特徵，如：詞性、詞頻等；第二種則是在解碼器生成詞彙之前，加入一個控制器，控制解碼器是否要生成新詞或從輸入文件複製，此一機制是參考(Vinyals, Fortunato & Jaitly, 2015)提出的 Pointer Network 架構，當文件中有專有名詞出現時，但解碼器的詞典中可能沒有該詞彙，就需要從輸入資料中複製使用；最後則是將編碼器改成階層式的編碼器，一般的編碼器輸入都是整篇文章的每個詞彙，不考慮語句的分界，而階層式編碼器第一層的輸入一樣是整篇文章的每個詞彙，當遇到每個語句的結尾詞時，就會將此時的輸出向量視為語句的向量表示，並作為第二層的輸入，也就是說，第二層的輸入是文章中的語句，這種方法能夠得到更細部的文件資訊，也使得產生之摘要內容較符合文章主旨。雖然在 (Nallapati *et al*., 2016) 已經有嘗試將 Pointer Network 的想法結合進模型中，但是此種方法過於強硬，因為此控制器得到的結果僅能二選一。

因此 (See *et al*., 2017) 提出的架構能有效的解決此狀況，此篇研究提出的方法是以同時進行產生新詞與選取原有詞彙的動作，最後利用一機率值簡單線性結合兩者所得到的機率分佈，以此得到最終的詞典機率分佈，詞典中包含解碼詞典與輸入文件的詞彙。此外，(See *et al*., 2017)亦提出一種 Coverage 機制，此機制主要是為了解決在語言生成任務上容易出現 OOV 和重複詞的問題，其在每個時間點會將以前時間點得到的注意力分佈加總後作為一 coverage 向量，維度大小為編碼器的時間點數量，而後在當前時間點會參考此向量計算注意力分佈，同時也會將此向量和注意力分佈進行比較，找出每個維度最小值後加總便得到一 coverage 損失，之後會做為訓練時使用的懲罰值，讓模型可以將重複詞的機率降低。此研究所得到的摘要效果比以往的重寫式摘要優異許多，而實驗結果亦顯示摘要成果比較偏向於節錄式摘要，因為複製的比例比生成的比例高出許多，與此同時我們也發現節錄式摘要的成效仍比重寫式摘要更為顯著。

3. 階層式類神經摘要模型 (Hierarchical Neural Summarization Model)

我們將語音文件摘要問題視為一語句分類暨排序問題，以期能依文件主旨選出可能為摘要的語句，且同時能學習到摘要語句間有意義的排序，使得摘要內容能更流暢地表達文件主題及概念。因此，我們提出一基本架構，其中包含一階層式編碼器及一解碼器，亦稱之為語句選取器。階層式編碼器中主要有兩個階層，我們會先針對文件中的語句找到對應的語句表示，再從語句表示中學習到文件中的重要概念，亦可稱為文件表示；最後會將語句表示及文件表示皆放置於語句選取器中，使其能夠根據文件表示及語句表示，辨別及排序摘要句。

此外，為了避免摘要結果受到過多語音辨識錯誤的影響，我們嘗試加入聲學特徵和次詞向量輔助訓練；同時我們亦加入注意力機制和強化學習機制於模型訓練中，以期能增加摘要的資訊性。

3.1 問題定義及假設 (Problem Formulation)

首先我們將語音文件摘要任務定義為一序列標記問題，主要是針對文件中的語句進行摘要的標註。其中摘要類別可分為摘要和非摘要，分別以 1 和 0 表示，因此我們將任務目標定義為最大化類別機率，亦為最大化似然性，並可將目標函式定義為下式：

$$log\, p(\boldsymbol{y}|D,\, \theta) = \sum_{i=1}^{N} p(y_i|s_i,\, D,\, \theta) \tag{1}$$

當給定一文件 D 時，其為一語句序列 $(s_1, \dots, s_n)$，我們的方法會從 D 中選取 M 個語句經由排序後作為其摘要。對於每個語句 $s_i \in D$，我們會預測一分數 $p(y_i|s_i, D, \theta)$ ，作為判定是否為摘要的依據 $y_i \in (0, 1)$。之後會依照語句被視為摘要的分數 $p(y_i = 1|s_i, D, \theta)$ 對所有語句進行排序，取前 M 個語句作為此文件摘要。

對於每個語音文件，我們定義以下幾點假設：

- 語音資訊可透過額外的聲學特徵參考進模型訓練
- 使用字向量可有效改善語句表示的成效並抵銷語音辨識錯誤
- 摘要句可被其他非摘要句解釋
- 強化學習技術可訓練摘要之排序

後續我們會針對上述之假設對模型架構進行不同的改進，且會詳細闡述其動機。

3.2 基本架構 (Basic Architecture)

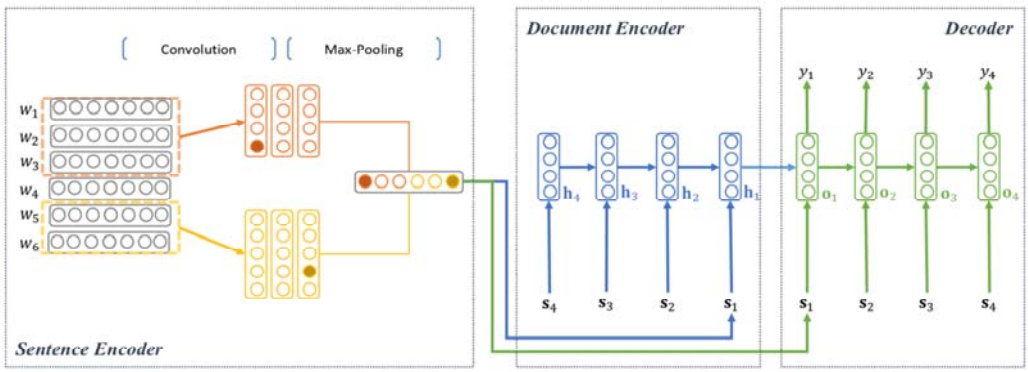

圖 2. 階層式類神經摘要模型 - 基本架構
[Figure 2. Basic architecture]

基本架構中包含一階層式編碼器及一解碼器，亦稱之為語句選取器。階層式編碼器中主要有兩個階層，我們會先針對文件中的語句找到對應的語句表示，再從語句表示中學習到文件中的重要概念，亦可稱為文件表示；最後會將語句表示及文件表示皆放置於語句選取器中，使其能夠根據文件表示及語句表示，辨別及排序摘要句。

3.2.1 語句編碼器 (Sentence Encoder)

我們利用摺積式類神經網路 (Convolutional Neural Networks, CNNs) 將每個不同長度的語句投影至向量空間，能夠得到固定長度的向量表示 (Representation)。在過去的研究中顯示，CNNs 在 NLP 領域的任務中有相當不錯的成效(Cheng & Lapata, 2016; Collobort *et al*., 2011; Kalchbrenner, Grefenstette & Blunsom, 2014; Kim, Jernite, Sontag & Rush, 2016; Lei, Barzilay & Jaakkola, 2015; Zhang, Zhao & LeCun, 2015)。我們使用 1-D 摺積 (Convolution) 並給定寬度 h 的摺積核 (Kernel) K，其定義為每次看 h 個詞彙，類似於 N 元模型 (N-gram) 的概念，可得到特徵圖 (Feature map) f。之後，對每個特徵圖沿著時序使用最大池化 (Max Pooling)，將特徵圖中的最大值視為語句特徵。為了能找到更好的特徵，我們使用多種寬度的摺積核，且每種寬度有多個不同的摺積核，最後將所得到的特徵串接在一起，即為語句的向量表示。

3.2.2 文件編碼器 (Document Encoder)

在文件編碼器中，我們使用遞迴式類神經網路 (Recurrent Neural Networks, RNNs)，將每個文件的語句序列轉換成一固定長度之向量表示，其能夠擷取到文件中的重要資訊。其中為了避免產生消失的梯度 (Vanishing Gradient) 問題，我們選擇使用 GRU (Gated Recurrent Unit) (Cho *et al*., 2014) 作為 RNN 的基本單元。此外，我們參考相關實作，將文件以倒序的方式作為輸入(Narayan, Papasarantopoulos, Cohen & Lapata, 2017; Narayan *et al*., 2018a; Narayan *et al*., 2018b; Sutskever *et al*., 2014)。由於我們使用的訓練語料是以新聞為主，而大多數新聞的主旨通常座落於開頭幾句，因此以倒序方式輸入文章，能使得 RNN 對重要資訊記憶更深。，因此可定義下列算式：

$$\mathbf{h}_i = f^e(\mathbf{h}_{i+1}, \mathbf{s}_i) \tag{2}$$

$$\mathbf{d} = \mathbf{h}_1 \tag{3}$$

其中 $f^e(\cdot)$ 為 RNN，$\mathbf{h}_i$ 是序列中每個時間點經過 RNN 運算後得到的隱藏層輸出，而 $\mathbf{s}_i$ 為語句向量。因輸入方式為倒序，所以每個時間點 $\mathbf{h}_i$ 都會參考後一時間點的輸出 $\mathbf{h}_{i+1}$ 及當前時間點的語句向量 $\mathbf{s}_i$ 。最後為了能得到整篇文章的隱含資訊，我們將最後一個時間點的輸出 $\mathbf{h}_1$ 視為文件向量 $\mathbf{d}$ ，並供之後摘要擷取時使用。

3.2.3 摘要選取器 (Summary Extractor)

我們的摘要選取器主要會將文件中每個語句標示為 1 (摘要) 或 0 (非摘要)。在此部分，我們將會使用另外一個 RNN，其中輸入一樣以語句向量為主，而語句向量同樣是經由語句編碼器所產生。此處與文件編碼器不同之處在於，摘要選取時是以文件的正序輸入，因此可定義成下列方程式：

$$\mathbf{o}_i = f^d(\mathbf{o}_{i-1}, \mathbf{s}_i) \tag{4}$$

$$\mathbf{o}_0 = \mathbf{d} \tag{5}$$

$$\mathbf{y}_i = \text{softmax}(\text{MLP}(\mathbf{o}_i)) \tag{6}$$

其中 $\mathbf{o}_i$ 為隱藏層輸出，$f^d(\cdot)$ 為一 RNN 架構，其輸入包含前一時間點的隱藏層輸出 $\mathbf{o}_{i-1}$ 和當前時間點的語句輸入 $\mathbf{s}_i$。為了在選取摘要時能參考到整篇文章的主旨，我們將初始的隱藏層 $\mathbf{o}_0$ 設定為文件向量 $\mathbf{d}$。此舉可以同時參考局部 (單一語句) 及整體 (文件) 的資訊，因此能更好的辨別語句。最後我們會透過 (6) 計算每個語句的類別 $\mathbf{y}_i$，其中 MLP(·) 為一簡單的前向式類神經網路(Feed-forward Neural Networks) 之後經由一個 softmax 函式得到語句類別的機率 $p(\mathbf{y}_i|\mathbf{s}_i, D, \theta)$ ，並依據 $p(\mathbf{y}_i = 1|\mathbf{s}_i, D, \theta)$ 將每個語句進行排序，依照固定的摘要比例選取排名高的語句作為完整的摘要結果。

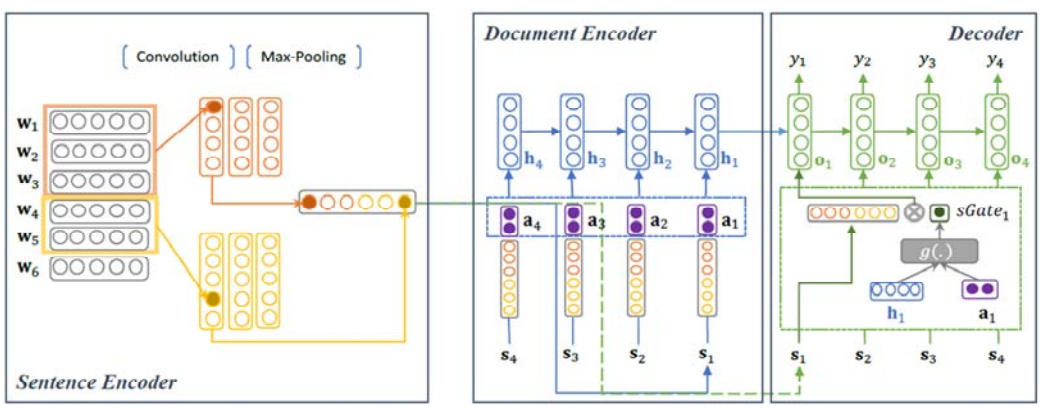

圖 3. 階層式類神經摘要模型 - 結合聲學特徵
[Figure 3. Basic architecture with acoustic features]

3.2.4 聲學特徵 (Acoustic Features)

為了能夠避免摘要結果受到辨識錯誤的影響，我們認為聲學特徵能夠保留每個文件的語音資訊且不受辨識錯誤之影響，因此提出三種方式將聲學特徵與上述架構結合，使得在判斷摘要的時候能夠參考，以得到更好的摘要成果。聲學特徵是以語句為單位，每個語

句會有對應的聲學特徵，因此令聲學特徵向量為 $\mathbf{a}$，我們的方法可定義下列方程式：

$$\mathbf{h}_i = f^{e\prime}(\mathbf{h}_{i+1}, [\mathbf{s}_i; \mathbf{a}_i]) \tag{7}$$

$$\mathbf{o}_i = f^{d\prime}(\mathbf{o}_{i-1}, [\mathbf{s}_i; \mathbf{a}_i]) \tag{8}$$

$$\mathbf{sGate}_i = g(\mathrm{W}_g[\mathbf{h}_i; a_i] + \mathbf{b}_g) \tag{9}$$

$$\mathbf{s}'_i = \mathbf{s}_i \odot \mathbf{sGate}_i \tag{10}$$

$$\mathbf{o}_i = f^{d\prime\prime}(\mathbf{o}_{i-1}, \mathbf{s}'_i) \tag{11}$$

全域向量(Global Embedding)

首先，我們將文件編碼器的輸入語句與其對應的聲學特徵串接，經過編碼後可得到新的文件向量，將 (2) 修改成 (7) 的方程式。我們認為此種做法同時考慮整份文件的聲學特徵，因此我們所得到的文件向量便可包含其聲學特徵，所以稱之為全域向量。

局部向量(Local Embedding)

其次，我們亦嘗試將語句對應的聲學特徵與語句向量串接後，直接用於摘要選取器之輸入，可修改 (4) 為 (8)。此方法使得聲學特徵向量能直接作用於摘要選取時的判斷，卻僅只作用於當前時間點及未來時間點，所以我們稱其為局部向量。

選擇向量(Selective Embedding)

最後一種方式與前面兩種比較不同，我們的想法來自(Zhou, Yang, Wei & Zhou, 2017)的選擇機制 (Selective Mechanism)，其概念主要是希望在生成摘要前可以先進行選擇的動作，預先篩選出可能成為摘要的語句，之後便能找到更準確的摘要。而在本論文中，我們希望透過聲學特徵能預先篩選出可能的摘要句。如 (9) 所示，我們將文件編碼器的輸出$\mathbf{h}_i$及對應之輸入語句的聲學特徵 $\mathbf{a}_i$ 串接，並作為 $g(\cdot)$ 的輸入。$g(\cdot)$ 是一個三層的前向式類神經網路，會得到 $\mathbf{sGate}_i$，其數值範圍在 0~1 之間，可視為語句被選的機率或權重。最後我們將語句 $\mathbf{s}_i$ 和 $\mathbf{sGate}_i$ 相乘後可得到新的語句向量 $\mathbf{s}'_i$，如 (10) 所示，並將其取代 (4) 的輸入 $\mathbf{s}_i$ 如 (11)，因此我們將此種方法稱為選擇向量。

3.2.5 次詞向量 (Sub-word Information)

在語音文件摘要中，常見的語音辨識錯誤大多是因為辨識時將詞彙辨識成同音的其他詞語，而使用此辨識結果進行摘要擷取時，會因為其中的詞彙錯誤導致上下文含義被誤判，因而找不到正確的文件主旨。因此本論文提出使用次詞向量輔助模型學習文件特徵表示以避免詞彙辨識錯誤導致之影響，原本的模型中是使用詞向量為最小單位組成文章，然而詞彙的辨識錯誤亦會影響到斷詞的結果，而語句中的特徵表示較容易受到錯誤的詞向

量影響，因語句中含有的詞彙相對較少；若改以次詞向量進行訓練，同時可以學習到詞彙的語意，亦能減緩受到詞彙錯誤的影響。過去亦有研究(Bojanowski, Grave, Joulin & Mikolov, 2017; Chen, Xu, Liu, Sun & Luan, 2015; Kim *et al.*, 2016)表示使用次詞向量亦能有效地表達文件，且能輔助詞向量訓練。

在本論文中，我們改良基本模型架構，加入一個輔助的語句編碼器（如圖 4），其中的設置與原有的語句編碼器相同。為了方便區隔，我們可將原有的語句編碼器稱為詞階段語句編碼器，而我們使用的次詞向量是字向量，可稱之為字階段語句編碼器。而前述之語句向量為 $\mathbf{s}_i$，我們將其表示為 $\mathbf{s}_i^w$，字階段語句向量則定義為 $\mathbf{s}_i^c$。在此架構中，我們希望能以字向量輔助詞向量訓練語句表示，因此我們定義以下方程式以更好地融合字與詞的向量資訊：

$$\mathbf{s}_i^* = f_s(\mathrm{W}_s^w \mathbf{s}_i^w + \mathrm{W}_s^c \mathbf{s}_i^c + \mathbf{b}_s) \tag{12}$$

其中 $\mathbf{s}_i^*$ 表示詞與字階段的語句向量融合後的語句表示，而W_s^w, W_s^c 和 $\mathbf{b}_s$ 為訓練用之參數，$f_s(\cdot)$ 為一單層的前向式類神經網路，能夠簡單地結合 $\mathbf{s}_i^w$ 和 $\mathbf{s}_i^c$，最後我們可將 (2) 和 (4) 的 $\mathbf{s}_i$ 代換成新的語句向量 $\mathbf{s}_i^*$ 進行摘要選取。

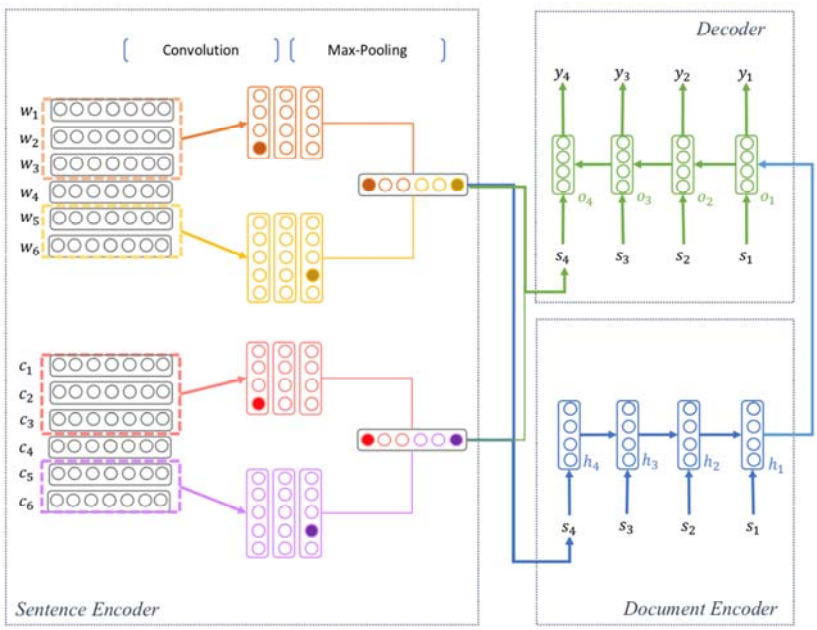

圖 4. 階層式類神經摘要模型 - 結合次詞向量
[Figure 4. Basic architecture with sub-word information]

3.2.6 注意力機制 (Attention Mechanism)

過去曾有學者(Ren *et al.*, 2017)表示，摘要主要是文件的簡短描述，而文件中的其他非摘

要語句則能夠細部地解釋摘要。文件的撰寫可概分為三種可能，第一種「通用到特定(General-to-specific)」所指的是文件開頭便簡短地描述文件內容，之後的內文皆是針對文件的細部闡述；第二種為「特定到通用(Specific-to-general)」，文件中先針對每個重點針對性地討論，最後作總結，因此主旨會落在文件後半部；最後一種則是「特定到通用到特定 (Specific-to-general-specific)」，所指的是文件先做細部討論，然後在中段破題點出文件主旨，之後再繼續討論細部的內容。從這三種情況中，我們可以知道文件中的語句和摘要句都有一定的關聯性，因此要找到摘要，文件中其他語句亦是必不可少。

對於文件摘要任務而言，值得注意的是摘要結果應該盡可能包含更多原文件中重要的資訊。因此，若我們希望摘要能夠包含更多重要資訊，應該要擷取出那些和文件中每個語句都有一定關聯性的語句，所以我們嘗試在我們的架構中加入注意力機制(Attention Mechanism)(Bahdanau *et al.*, 2015)。注意力機制可以找到每個語句與其他句的關聯性，因此我們可以將模型改良成如圖 5 的架構。

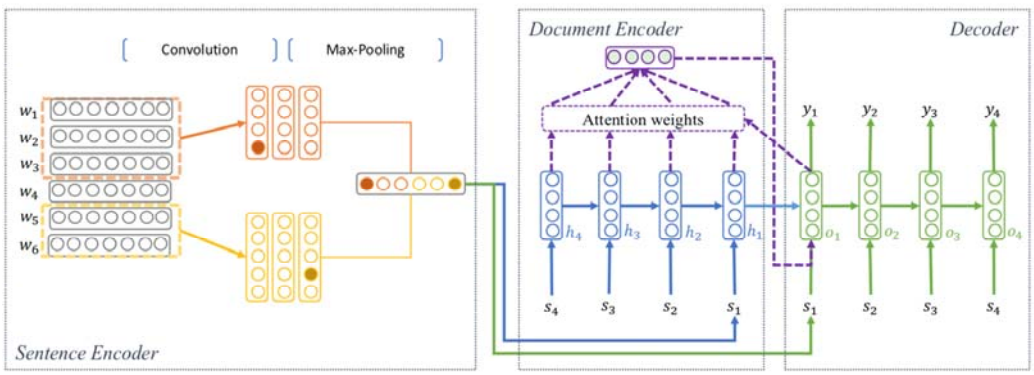

圖 5. 階層式類神經摘要模型 - 結合注意力機制
[Figure 5. Basic architecture with attention mechanism]

為了結合注意力機制，可以先簡單定義我們的摘要任務如下式：

$$p(\mathbf{y}_i|s_i, D, \theta) = m(\mathbf{s}_i, \mathbf{o}_i, \mathbf{c}_i) \tag{13}$$

其中 $\mathbf{c}_i$ 是透過注意力機制計算出的上下文向量，而 $\mathbf{o}_i$ 則是摘要選取器的隱藏層資訊，$m(\cdot)$ 代表整個摘要選取器，此式是表示摘要選取器的目標，主要是要預測語句的摘要類別機率 $p(\mathbf{y}_i|s_i, D, \theta)$。由於我們在摘要選取時結合注意力機制，因此可以重新定義 (4) 為下式：

$$\mathbf{o}_i = f^d(\mathbf{o}_{i-1}, \mathbf{s}_i, \mathbf{c}_i) \tag{14}$$

在每次 RNN 的計算中都會參考前一個時間的隱藏層資訊 $\mathbf{o}_{i-1}$ 、當前的語句向量表示 $\mathbf{s}_i$ 和該語句的上下文向量 $\mathbf{c}_i$，其中上下文向量主要是對文件編碼器的隱藏層資訊 $(\mathbf{h}_1, \dots, \mathbf{h}_n)$ 進行加權：

$$\mathbf{c}_i = \sum_j^N \alpha_{ij} \mathbf{h}_j \tag{15}$$

而 α_{ij} 是文件編碼器的隱藏層向量 $\mathbf{h}_j$ 對應的權重，此權重是透過下式計算：

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_k^N \exp(e_{ik})} \tag{16}$$

$$e_{ij} = a(\mathbf{o}_{i-1}, \mathbf{h}_j) \tag{17}$$

其中 e_{ij} 用來計算語句 $\mathbf{s}_j$ 跟語句 $\mathbf{s}_i$ 的關聯性，若 $\mathbf{s}_i$ 為摘要句，則其跟其他語句都應有一定的關聯性，而非僅跟部分有關。在 (17) 中 $a(\cdot)$ 為一簡單的前向式類神經網路用於計算語句間的關聯性分數，再經過一個 softmax 函數將其轉化為一 0~1 的數值如 (16)。因此在預測摘要語句的機率 $p(\mathbf{y}_i|s_i, D, \theta)$ 時，α_{ij} 能夠反應出語句之間的相關性，因而判定該語句是否被認定為摘要。

3.2.7 強化學習 (Reinforcement Learning)

傳統的摘要模型訓練目標一般都是使用最大似然評估 (Maximum Likelihood Estimation, MLE)，也就是要最大化 $p(y|D, \theta) = \prod_{i=1}^n p(y_i|s_i, D, \theta)$，因此會選擇交叉亂度 (Cross Entropy) 計算損失 (loss)，目標函式可定義為下列方程式：

$$L(\theta) = -\sum_{i=1}^{n} \log p(y_i|s_i, D, \theta) \tag{18}$$

但是此種方法有兩個主要的缺點，第一是因為我們所使用的評估指標與損失函數的定義不同，模型的訓練目標是要最大化似然性，但卻使用 ROUGE 來評估摘要的好壞。其中似然性的定義主要是根據出現的機率決定，而 ROUGE 則是比較模型摘要和參考摘要之間詞彙覆蓋率，兩者的定義完全不同，且大多數評估指標函式是無法進行微分的，因此不適用於訓練參數；第二則是因為我們定義節錄式摘要為語句分類問題，可是其通常被視為單類別分類問題 (One Class Classification, OCC)(Tax, 2001)，主要能被模型學習到的大部分是摘要句，而非摘要句其實不太能辨識（圖 6），因而造成訓練上的困難。

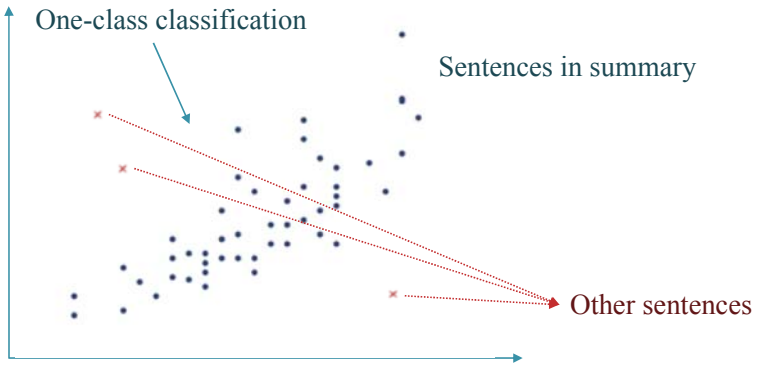

圖 6. 單類別分類問題示意圖
[Figure 6. One-class classification]

因此，我們使用強化學習(Sutton & Barto, 1998) 輔助模型訓練，由於基本的強化學習機制需要獎勵函數 (Reward Function)，此函數主要是用來判斷當前模型所預測的結果是否為正確，若正確則鼓勵訓練，反之則會懲罰。而獎勵函數的設定不像損失函數那麼嚴苛，因此我們使用摘要評估指標 ROUGE 作為獎勵函數，而訓練目標則可改成最小化獎勵期望值：

$$L(\theta) = -\mathbb{E}_{\hat{y}\sim p_\theta}[r(\hat{y})] \tag{19}$$

其中 p_θ 是指 $p(y|D,\theta)$，$r(\cdot)$ 是獎勵函數，而 $\hat{y}$ 是經過取樣 (Sample) 後得到的預測摘要。但是預測摘要 $\hat{y}$ 的可能性有無限多種，我們無法每次訓練都找到所有可能且計算其期望值來調整參數，這是很耗費成本的。因此我們將 (19) 改成 (20)，每次訓練只取一個樣本加速其訓練，並可將梯度 (Gradient) 函式改成 (21)，使其訓練上更為容易：

$$L(\theta) \approx r(\hat{y}) \tag{20}$$

$$\nabla L(\theta) \approx -r(\hat{y}) \sum_{i=1}^{n} \nabla \log p(\hat{y}_i|s_i, D, \theta) \tag{21}$$

4. 實驗結果 (Experimental Results)

4.1 實驗語料 (Corpus)

我們主要使用中文廣播新聞語料庫 (Mandarin Benchmark broadcast news corpus, MATBN)(Wang, Chen, Kuo & Cheng, 2005)。MATBN 是一個公開且常被應用於一些自然語言處理相關的任務上，如語音辨識(Chien, 2015)、資訊檢索(Huang & Wu, 2007)以及自動摘要(Liu *et al.*, 2015; Tsai, Hung, Chen & Chen, 2016)等。此資料集其中有 205 篇廣播新聞文件適用於摘要實驗，我們挑選其中的 20 篇作為測試集，餘下的 185 篇則為訓練集。

資料亦分成兩種，TD 為經過人工標註的文件，而 SD 則為經過自動語音辨識後產生的文件，因此 SD 會有部分的語音辨識錯誤。表 1 是對訓練集及測試集作的一些基本統計資料。此外，語音文件的聲學特徵類型列於表 2 中，是利用 Praat 工具擷取的結果，總計有 36 個特徵。

表 1. 用於摘要之廣播新聞文件的統計資訊[Tsai et al., 2016]
[Table 1. The statistics of MATBN]

	訓練集	測試集
文件數	185	20
每文件平均句數	20	23.3
每句平均詞數	17.5	16.9
每文件平均詞數	326.0	290.3
平均詞錯誤率	38.0%	39.4%
平均字錯誤率	28.8%	29.8%

此外，我們所使用之聲學特徵列於表 2 中，是利用 Praat 工具擷取的結果，總計有 36 個特徵，可簡單分為四種類型介紹：

- **Pitch 音高**：

  當我們在說話時，講到重點的時候，音高就會比較高來吸引注意，反之則會維持相對較低的音高。

- **Energy 能量**：

  能量一般是指語者的說話音量，通常都會被視為一項重要的資訊。當我們要強調某件事情時，除了音高會提高外，音量也會自然地放大，因而能幫助模型分辨重要資訊。

- **Duration 持續時間**：

  持續時間有點類似於一個語句中的詞彙數量，當持續時間越長沒有間斷時，則表示這句話包含的資訊相對較多。

- **Peak and Formant 峰與共振峰**：

  共振峰是頻譜中的峰值，主要用來描述人類聲道內的共振情形。如果聲音比較低沈，則共振峰會比較明顯，聽到的內容亦會較清晰；反之若聲音太過高亢，則共振鋒會比較模糊，同時聽到的內容也會比較模糊難辨。

表 2. 語音文件中每個語句對應的聲學特徵
[Table 2. List of acoustic features in MATBN]

聲學特徵	1. Pitch (min, max, diff, avg) 2. Peak normalized cross-correlation of pitch (min, max, diff, avg) 3. Energy value (min, max, diff, avg) 4. Duration value (min, max, diff, avg) 5. 1st formant value (min, max, diff, avg) 6. 2nd formant value (min, max, diff, avg) 7. 3rd formant value (min, max, diff, avg)

4.2 實驗結果 (Results)

首先本論文先比較過去的摘要方法於我們的資料集上的成效，之後在針對我們提出的架構和不同組合的摘要成果差異。

4.2.1 基礎實驗(Baseline)

過去 MATBN 資料集曾應用在各種不同的摘要方法上，從傳統的摘要方法(VSM, LSA)、非監督式類神經網路架構（SG, CBOW）到監督式類神經網路架構（DNN, CNN）都曾有學者使用。因此我們將過去的研究表現作為本論文比較的基礎實驗，結果列於表 3 中。

表 3. 基礎實驗結果
[Table 3. Results of baseline]

	文字文件			語音文件		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
VSM	0.347	0.228	0.290	0.342	0.189	0.287
LSA	0.362	0.233	0.316	0.345	0.201	0.301
SG	0.410	0.300	0.364	0.378	0.239	0.333
CBOW	0.415	0.308	0.366	0.393	0.250	0.349
DNN	0.488	0.382	0.444	0.371	0.233	0.332
CNN	0.501	0.407	0.460	0.370	0.208	0.312
Refresh [Narayan *et al*., 2018a]	0.453	0.372	0.446	0.329	0.197	0.319

首先我們可以從表中發現傳統的向量空間模型（Vector space model, VSM）在文字文件和語音文件上的效果沒有差異太大，但文字文件的效果仍是比語音文件優異；另外我們可以將 VSM 跟 LSA 作一個簡單的比較，可以發現 LSA 的結果能很明顯的看出文字文件和語音文件的差異，同時也比 VSM 的效果好許多。

接著我們從非監督式類神經網路架構的結果觀察，SG(Skip-gram)和 CBOW 應用於

訓練詞向量的差異其實不大，因此在整體的摘要效果上兩者的差異其實並沒有很大，但 CBOW 相較於 SG 是較優異的，而此二者方法的效能亦超越傳統的向量模型許多。

最後我們針對監督式類神經網路架構作討論，DNN 是最基本的多層類神經網路架構，而 CNN 則是使用摺積式類神經網路架構，Refresh 是與本論文相似的階層式架構。其中在文字文件的效果上，可以很明顯地發現三者都超越了非監督式的方法，尤以 CNN 的效果最好，可能是因為 CNN 比 DNN 更能抓到重要資訊，而參數量又比 Refresh 少，較易於訓練；但在語音文件的成效上，三者都比非監督式的效果差，可能是因為其太過於依賴文件中的詞彙資訊，因而受到語音辨識錯誤的影響較為嚴重，導致其效果較差。

後續章節我們將以 Refresh 的數據與本論文提出之架構進行比較及分析。

4.2.2 階層式類神經摘要模型實驗 (Our models)

在實驗結果分析中，我們前面章節介紹模型時提到的副架構分開實驗，以下會列出不同實驗設置的效果，以及結果討論與分析。

I. 次詞向量

首先，我們先比較詞向量和字向量用於模型中的效果，如下表所示，可以看出單獨使用詞向量的結果在語音文件上的效果反而比單獨使用字向量的時候優異，但在文字文件上反而相反，這樣的結果與我們的假設有些許出入，可能是因為訓練文件中錯誤的字比較集中，因而無法透過周圍的資訊來學習正確的詞彙資訊；此外，若使用融合向量於我們的模型中，在語音文件的結果上可以有很明顯的進步，但在文字文件上僅於 ROUGE-2 有進步，因而我們認為字向量和詞向量之間可能仍有相輔相成的作用。

表 4. 階層式類神經摘要模型-次詞向量結果
[Table 4. Results of our model with sub-word information]

	文字文件			語音文件		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Refresh [Narayan *et al*., 2018a]	0.453	0.372	0.446	0.329	0.197	0.319
詞向量	0.526	0.473	0.520	0.380	0.262	0.370
字向量	**0.544**	0.473	**0.535**	0.363	0.242	0.351
融合向量(詞+字)	0.543	**0.481**	0.533	**0.392**	**0.266**	**0.380**

II. 強化學習

承上所述，我們認為融合向量於語音摘要上有相當大的可能性，因此我們嘗試同時使用融合向量和強化學習於模型上，從表 5 中可以很明顯的看到強化學習於我們的方法中有一定的成效在，不過在文字文件摘要上有比較多的進步，主因可能是在於參考摘要不包含語音辨識錯誤，因此沒有辦法完全解決語音辨識錯誤的影響，若能將聲學特徵亦加入強化學習的獎勵函數中或許能改進此情況。

表 5. 階層式類神經摘要模型-強化學習
[Table 5. Results of our model with reinforcement learning]

	文字文件			語音文件		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Refresh [Narayan *et al*., 2018a]	0.453	0.372	0.446	0.329	0.197	0.319
融合向量	0.543	0.481	0.533	0.392	0.266	**0.380**
融合向量+強化學習	**0.555**	**0.479**	**0.543**	**0.395**	**0.269**	0.379

III. 聲學特徵+強化學習

經過前面兩項實驗比較，我們可以發現融合向量可以解決部分的語音辨識錯誤影響，而強化學習則比較專注於摘要資訊性。因次我們嘗試於模型上結合聲學特徵與強化學習的方法，從表 6 中，我們可以發現在語音文件摘要上，效果比較顯著的是使用局部向量的方式結合聲學特徵；然而在文字文件摘要中，比較好的結果是使用全域向量。因此我們可以推論出聲學特徵對於人類轉寫的文字文件效用不彰，而對於自動辨識的語音文件上，還是有不錯的效果，但可能需要讓聲學特徵直接參與摘要選取的階段才能有效的提升效能。然而，整體的數據上仍是比前面的實驗差了許多，可能是模型上還需作更多細部的調整，或結合其他機制。

表 6. 階層式類神經摘要模型-聲學特徵+強化學習
[Table 6. Results of our model with acoustic features and reinforcement learning]

	文字文件			語音文件		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Refresh [Narayan *et al*., 2018a]	0.453	0.372	0.446	0.329	0.197	0.319
無聲學特徵	0.479	**0.400**	0.469	0.352	0.226	0.342
全域向量	**0.486**	**0.400**	**0.473**	0.350	0.222	0.336
局部向量	0.478	0.399	0.469	**0.384**	**0.264**	**0.370**
全域向量+局部向量	0.464	0.373	0.453	0.350	0.224	0.336
選擇向量	0.448	0.371	0.439	0.350	0.213	0.334

IV. 次詞向量+注意力機制

因前一個實驗結果發現聲學特徵和強化學習共同訓練時效果相對較差，因此我們這次比較結合次詞向量和注意力機制的實驗結果。從表 7 中可以發現同時使用融合向量和注意力機制的效果在文字文件上較為優異，而在語音文件上仍是以詞向量的結果比較好。雖然整體的效果皆比之前的結果好，但可能是因為注意力機制訓練的主要是文件中語句之

間的關聯性，而對於語音文件而言，若辨識錯誤的太多，比較難找到語句間的語意關聯性，因而導致結果相對較差。

表 7. 階層式類神經摘要模型-次詞向量+注意力機制
[Table 7. Results of our model with sub-word information and attention mechanism]

	文字文件			語音文件		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Refresh [Narayan *et al*., 2018a]	0.453	0.372	0.446	0.329	0.197	0.319
詞向量+注意力機制	0.523	0.472	0.519	0.401	**0.290**	**0.392**
字向量+注意力機制	0.535	0.477	0.529	0.368	0.245	0.356
融合向量+注意力機制	**0.567**	**0.496**	**0.557**	**0.402**	0.278	0.389

V. 次詞向量+注意力機制+強化學習

接續前一個實驗，我們加入強化學習機制於訓練中，實驗結果如表 8 所示。從結果可以發現，不管是文字文件還是語音文件，加入強化學習機制後，皆是在輸入為詞向量時會得到較好的效果。這有可能是因為我們的強化學習中獎勵函數使用 ROUGE 分數，而 ROUGE 計算時主要是以詞為基本單位，因而導致在其他情況下結果相對較差。

表 8. 階層式類神經摘要模型-次詞向量+注意力機制+強化學習
[Table 8. Results of our model with sub-word information, attention mechanism and reinforcement learning]

	文字文件			語音文件		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Refresh [Narayan *et al*., 2018a]	0.453	0.372	0.446	0.329	0.197	0.319
詞向量+注意力機制+強化學習	**0.543**	**0.491**	**0.539**	**0.350**	**0.226**	**0.337**
字向量+注意力機制+強化學習	0.525	0.451	0.515	0.342	0.221	0.329
融合向量+注意力機制+強化學習	0.518	0.448	0.502	0.347	0.209	**0.337**

VI. 綜合比較

最後，我們將前述提到之架構做一個綜合比較，實驗結果如表 9 所示。其中我們可以發現當強化學習機制和注意力機制同時使用的情況下，不管是在文字文件還是語音文件上效果都相對較差。此種情況有可能是因為我們的注意力機制主要針對的是摘要資訊性提升，而強化學習中由於使用 ROUGE 分數作為獎勵函數，而 ROUGE 也是計算摘要資訊

性，因此當兩者同時訓練時，雖然都是針對資訊性，但可能因為太過注重而造成反效果。

其次，我們也嘗試結合注意力機制和聲學特徵的應用，如表 9 的最後兩列，由於前面的討論中發現使用局部向量方式結合聲學特徵在語音文件上會有較佳的效果，因此此部分實驗亦採用局部向量。實驗結果顯示加入聲學特徵在文字文件摘要上有些許的提升，但於語音文件摘要中沒有太大的影響，然而跟未加入聲學特徵訓練的實驗數據相比較，我們發現數據其實差異不大，此情況可能是因為此部分的實驗受到注意力機制的影響較顯著，聲學特徵對於此部分實驗不是其訓練的重點，因此沒有顯著的提升。

表 9. 階層式類神經摘要模型-綜合比較
[Table 9. Comprehensive comparison of our models]

	文字文件			語音文件		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Refresh [Narayan *et al*., 2018a]	0.453	0.372	0.446	0.329	0.197	0.319
融合向量+注意力機制+強化學習	0.518	0.448	0.502	0.347	0.209	0.337
融合向量+注意力機制	0.567	0.496	0.557	**0.402**	0.278	0.389
融合向量+注意力機制+聲學特徵+強化學習	0.532	0.455	0.521	0.336	0.220	0.326
融合向量+注意力機制+聲學特徵	**0.569**	**0.507**	**0.561**	0.401	**0.288**	**0.394**

VII. 視覺化注意力

另外，我們亦針對注意力機制中的權重進行分析（圖 7），圖中每個列和行代表代表文件中的語句，每個列的語句標號旁括弧內的數值為 $p(\mathbf{y}_i = 1|s_i, D, \theta)$ ，即該句被辨識為摘要的機率。若該列中每欄的顏色越深，則代表該句和其他句的關聯性越大，則該句也被視為摘要，其中被紅框圈起的列為參考摘要。從紅框的部分看可以很明顯的發現，我們的摘要系統選出的摘要大部分和參考摘要相同，因此可驗證我們的注意力機制於摘要任務上真的有一定的成效。

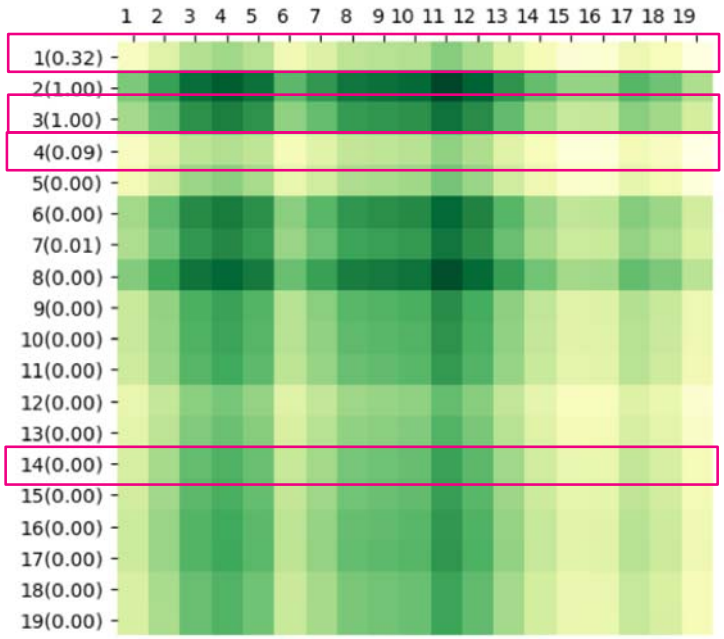

圖 7. 注意力機制權重視覺化
[Figure 7. Visualization of attention weight]

簡單總結整體實驗結果，我們提出之模型架構確實可有效提升語音文件摘要的成效，然而對於避免語音辨識錯誤的影響上，次詞向量和聲學特徵的效果仍有待加強；而注意力機制和強化學習等方法對於文字文件的效果仍比較顯著。因此若要實質性地提升語音文件摘要的成效，我們認為仍須從語音辨識的部分著手，若能不經過轉寫直接擷取摘要，或許更符合語音文件摘要，亦能有較優異的成效。

5. 結論與未來展望 (Conclusion & Future Work)

過去有關自動文件摘要的研究主要仍著重於文字文件摘要；而近年來由於大數據及機器學習技術的蓬勃發展，使得多媒體文件相關研究更為容易，因而逐漸有多媒體文件摘要相關研究的出現。雖然多媒體技術進步快速，但大多數的語音文件摘要方法仍多半由文字文件摘要方法延伸而來。直至近期隨著深層學習技術漸趨成熟，多媒體文件摘要技術也隨之成長。

順應深層學習的浪潮，本論文提出一種階層式類神經網路架構來從事語音文件摘要，同時亦適用於一般文字文件摘要。文件摘要任務可概分為節錄式與重寫式摘要。本論文旨在探討節錄式語音文件摘要方法。其中為了提升摘要資訊性及連貫性，我們加入注意力機制及強化學習技術；另外我們亦嘗試使用聲學特徵及次詞向量於模型訓練中，以避免計算摘要時受到過多語音辨識錯誤影響。經由一系列的實驗分析與討論，首先我們發

現注意力機制和強化學習皆可提升摘要資訊性，但同時使用時效果會相對較差；其次在避免語音辨識錯誤的部分，次詞向量與聲學特徵皆有不錯的成效，尤以次詞向量的效果較為顯著；最後對於摘要連貫性，我們的方法雖然有學習排序，但資料集中的參考摘要不包含排序資訊，因此無法完整地學習到語句間的連貫性。因此透過初步的實驗結果，足以證明我們提出的架構對於語音文件摘要有不錯的成效，但主要都反應於文字內容上，若要實質性的改善語音文件摘要的缺點，仍需更深入的探討。

承上所述，未來的研究我們可以針對幾個面向繼續深入。首先是應用預訓練語言模型於摘要研究上，改善語句或文章的語意表示，由於最近有許多預訓練語言模型已經使用相當大量的資料及高效能的設備進行訓練，且已被證明在許多任務上有相當亮眼的成績，僅需針對應用微調即可，或許可以嘗試進行深入研究；其次是重新整理資料集，因為摘要連貫性對於摘要亦是相當重要的指標，若成本允許，則可以僱請專家幫忙為資料進行重新標註，除了標註摘要語句外，同時亦加入摘要語句的順序，更有利於後續的摘要排序相關研究；再者，節錄式摘要亦可能發生語句間語意重複的情況，然而鮮少學者針對節錄式摘要重複性進行研究，因此為了減少節錄式摘要之重複性，或許可將重寫式摘要研究中常見之減少冗余的機制改良並應用於我們的方法上，應能得到更具意義的摘要結果；最後也最重要的是需要避免語音辨識錯誤影響語音文件摘要效果，從我們的實驗可以得出，現今的方法仍有所侷限，而為了有效地提升語音文件摘要準確性，或許我們能嘗試使用語音特徵如 Fbank 和 MFCC 等作為摘要系統之輸入，應可得到較原始的語音內容，亦能減少遇到辨識錯誤的情況，且因節錄式摘要是進行語句選擇，因此不需再進行轉寫，因而能使得摘要同為語音形式，但此想法需要多加考慮的部分在於難以評估結果正確與否，也相較兩階段的方法難實現，因此較少學者投入這方面的研究，若能實現我們的構想，應可使語音文件摘要技術達到新的高度，亦造福日後的學者們。

參考文獻(References)

Bahdanau, D., Cho, K.H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*.

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

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *TACL*, *5*, 135-146.

Chen, B., Kuo, J.-W., & Tsai, W.-H. (2004). Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing 2004*. doi : 10.1109/ICASSP.2004.1326101

Chen, B., Kuo, J.-W., & Tsai, W.-H. (2005). Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription. *International Journal of Computational Linguistics and Chinese Language Processing*, *10*(1), 1-18.

Chen, X., Xu, L., Liu, Z., Sun, M., & Luan, H. (2015). Joint Learning of Character and Word Embeddings. In *Proc. of IJCAI 2015*, 1236-1242.

Chen, Q., Zhu, X., Ling, Z., Wei, S., & Jiang, H. (2016). Distraction-based neural networks for modeling documents. In *Proc. of IJCAI 2016*, 2754-2760.

Cheng, J. & Lapata, M. (2016). Neural summarization by extracting sentences and words. In *Proc. of ACL*, 484-494. doi: 10.18653/v1/P16-1046

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., …Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proc. of EMNLP 2014*, 1724-1734. doi: 10.3115/v1/D14-1179

Chopra, S., Auli, M., & Rush, A. M. (2016). Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. In *Proc. of NAACL-HLT 2016*, 93-98. doi: 10.18653/v1/N16-1012

Chien, J.-T. (2015). Hierarchical Pitman-Yor-Dirchlet language model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(8), 1259-1272. doi: 10.1109/TASLP.2015.2428632

Collobort, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, M., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493-2537.

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735-1780.

Huang, C.-L. & Wu, C.-H. (2007). Spoken Document Retrieval Using Multilevel Knowledge and Semantic Verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *15*(8), 2551-2590. doi: 10.1109/TASL.2007.907429

Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2014). On using very large target vocabulary for neural machine translation. In arXiv preprint arXiv: 1412.2007.

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modeling sentences. In *Proc. of ACL 2014*, 655-665. doi: 10.3115/v1/P14-1062

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proc. of EMNLP 2014*, 1746-1751. doi: 10.3115/v1/D14-1181

Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2016). Character-aware neural language models. In *Proc. of AAAI 2016*, 2741-2749.

Lei, T., Barzilay, R., & Jaakkola, T. (2015). Molding CNNs for text: Non-linear, non-consecutive convolutions. In *Proc. of EMNLP 2015*, 1565-1575. doi: 10.18653/v1/D15-1180

Liu, S.-H., Chen, K.-Y., Chen, B., Wang, H.-M., Yen, H.-C., & Hsu, W.-L. (2015). Combining Relevance Language Modeling and Clarity Measure for Extractive Speech Summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(6), 957-969. doi: 10.1109/TASLP.2015.2414820

Nallapati, R., Zhai, F., & Zhou, B. (2017). SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *Proc. of AAAI 2017*, 3075-3081.

Nallapati, R., Zhou, B., dos Santos, C., Gülçehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proc. of CoNLL 2016*, 280-290. doi: 10.18653/v1/K16-1028

Narayan, S., Cohen, S. B., & Lapata, M. (2018). Ranking Sentences for Extractice Summarization with Reinforcement Learning. In *Proc. of NAACL 2018*, 1747-1759. doi: 10.18653/v1/N18-1158

Narayan, S., Cardenas, R., Papasarantopoulos, N., Cohen, S. B., Lapata, M., Yu, J., …Chang, Y. (2018). Document Modeling with External Attention for Sentence Extraction. In *Proc. of ACL 2018*, 2020-2030. doi: 10.18653/v1/P18-1188

Narayan, S., Papasarantopoulos, N., Cohen, S. B., & Lapata, M. (2017). Neural extractive summarization with side information. In arXiv preprint arXiv: 1704.04530.

Paulus, R., Xiong, C., & Socher, R. (2017). A deep reinforced model for abstractive summarization. In arXiv preprint arXiv:1705.04304.

Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proc. of EMNLP 2015*, 379-389. doi: 10.18653/v1/D15-1044

Ren, P., Chen, Z., Ren, Z., Wei, F., Ma, J., & de Rijke, M. (2017). Leveraging Contextual Sentence Relations for Extractive Summarization Using a Neural Attention Model. In *Proc. of SIGIR 2017*, 95-104.

See, A., Liu, P., & Manning, C. (2017). Get to the point: Summarization with pointer-generator networks. In *Proc. of ACL 2017*, 1073-1083. doi: 10.18653/v1/P17-1099

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th Advances in Neural Information Processing Systems*, 3104-3112.

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Tan, J., Wan, X., & Xiao, J. (2017). Abstractive document summarization with a graph-based attentional neural model. In *Proc. of ACL 2017*, 1171-1181. doi: 10.18653/v1/P17-1108

Tax, D. M. J. (2001). *One-class classification: Concept learning in the absence of counter-examples.* Unpublished doctoral dissertation, Technische Universiteit Delft.

Torres-Moreno, J. M. (2014). *Automatic text summarization*. Hoboken, New Jersey: John Wiley & Sons. doi: 10.1002/9781119004752

Tsai, C.-I., Hung, H.-T., Chen, K.-Y., & Chen, B. (2016). Extractive Speech Summarization Leveraging Convolutional Neural Network Techniques. In *Proceedings of IEEE SLT 2016*. doi: 10.1109/SLT.2016.7846259

Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer Networks. In *Proceedings of Advances in Neural Information Processing Systems 2015*.

Wang, H.-M., Chen, B., Kuo, J.-W., & Cheng, S.-S. (2005). MATBN: A Mandarin Chinese Broadcast News Corpus. *International Journal of Computational Linguistics & Chinese Language Processing*, *10*(2), 219-236.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Proceedings of Advances in Neural Information Processing Systems 2015*, 649-657.

Zhou, Q., Yang, N., Wei, F., & Zhou, M. (2017). Selective Encoding for Abstractive Sentence Summarization. In *Proc. of ACL 2017*, 1095-1104. doi: 10.18653/v1/P17-1101