國立臺灣大學電機資訊學院電機工程學研究所

碩士論文

Graduate Institute of Electrical Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

語音文件摘要與語音問答系統之新技術

Advanced Technologies for Spoken Document Summarization and Spoken Question Answering

向思蓉

Sz-Rung Shiang

指導教授：李琳山 博士

Advisor: Lin-shan Lee, Ph.D.

中華民國 103 年 6 月

June, 2014

摘要

本論文研究主題涵蓋語音技術應用之兩大前瞻性方向：語音文件摘要(Spoken Document Summarization)和語音問答系統(Spoken Question Answering)，實驗語料以中文為主，包含語音訊號以及語音辨識轉寫(Transcription)。

語音文件摘要以中文的課程錄音為應用領域，首先使用含有隱藏變數(Hidden Variable)之結構式支撐向量機(Structured Support Vector Machine)。結構式支撐向量機使得整篇文件的資訊可以一同考慮，因此可以加入重複性(Redundancy)的考量，使得有長度限制的摘要能包含更多資訊。此外透過連續句子叢集的隱藏變數，可以考慮到更細微的文件結構，更貼近人工選擇之摘要。

此外，由於督導式的結構式支撐向量機需要訓練資料，並不易在一般的課程系統中使用，因此本論文另外使用非督導式的雙層隨機漫步(Two-layer Random Walk)，除了使用課程錄音轉寫的句子外，另外輔以課程投影片的資訊，因此可以不使用需要大量人工的訓練資料，大大地降低了抽取摘要的前置作業與時間。

最後，本論文就語音問答系統作了初步的研究，主要針對以檢索為基礎(Information Retrieval Based)之模擬陳述問答(Factoid Question Answering)，語音問句使得本論文考慮的問題與一般的純文字問答系統不同。論文中使用問句之前 N 最佳辨識轉寫(N-best Transcription)，透過樹狀條件隨機域(Tree-Structured Conditional Ransom Fields)搭配剖析樹(Parse tree)進行關鍵詞抽取當作查詢指令，再透過搜尋引擎得到相關的網頁。然而，由於語音辨識含有許多辨識錯誤，本論文再利用雙層隨機漫步，對於各個前 N 辨識轉寫之查詢指令所對應的網頁進行重排序(Re-ranking)，讓越有可能含有正確答案的網頁排序提前。最後本論文實作初步的答案生成(Answer Generation)，針對固定的答案類別，利用網頁內資訊以及網頁排序為參數對每個可能的答案進行評分，對將來語音問答系統的研究提供了良好的起點。

致謝

首先要感謝碩士班兩年期間的指導教授—李琳山教授，在我的研究路上用心指導，每個禮拜總是花相當多的時間和學生討論，此外教授提供了如此適合研究的環境，實驗室中大家積極討論研究相關的話題，當有遇到問題，也都能和學長和同學們切磋討論，在如此的環境中，讓我有了研究的動力，探究所面對的問題。而且教授給予每個人相當大的自由度，可以自己開拓想要研究的目標與內容，讓我在碩士班的兩年期間得以找到自己最有興趣的研究主題，並加以深入了解。

感謝研究一路幫助我的大學長李宏毅，從一開始的專題，到碩士班的研究，總是提供我許多意見，更從不吝於指導，即使出國做博士後研究員，也依然對我伸出援手。此外感謝實驗室的學長們，葉青峰在語音辨識上獨撐大樑，當大家有語音辨識的需求時，總是給予幫助；蘇培豪總是很有耐心的和我聊天，不僅是在課業，還是在研究心路上，在我遇到難題時提供意見；溫宗憲總是很有自己的規劃與想法，在我對於研究灰心時，讓我見識他到對於研究的熱忱，而因此能夠繼續前進；胡庭曜對於數學的認真與熱情，讓我看到了對於研究的扎實苦工；傅怡聖學長在我不順心時，總是聽我抱怨和給我意見；詹竣安學長對於研究的好奇心與強大的數學能力，總是令我佩服不已；王祐邦學長在美國求職的經驗，總是不吝於分享給大家。感謝實驗室中的同學們，李昀樵、周伯威、楊子毅總是和我在夜晚的實驗室一同奮鬥，讓實驗室變得如此溫馨，此外還有在實驗室中的每一個人，研究路上有大家不孤單！

當然要感謝我的家人，除了拉拔我一路讀書外，在我的任何抉擇上的給予支持，讓我在課業外無後顧之憂，能夠在我選擇的道路上努力衝刺。最後，再次感謝幫助過我的所有人，因為有了大家，才能有今天的我，謝謝你們！

目錄

圖目錄

表目錄

Part I

背景及目標

第一章 導論

1.1 背景研究

近幾年中，網路資訊發展快速，如今每天全球網路上製造出來的資料量高達 25 億 GB，且同時資料量的成長相當驚人且持續攀升，各類應用如商業、娛樂、資訊查詢、社交等等相繼而生。而面對如此大量的資料量，使得人們難以將資料全部閱讀，諸如每日新聞、瀏覽器搜尋結果、商品評價等等，動輒上千萬的使用者意見讓人眼花撩亂。正因此，方便的資訊檢索系統(Information Retrieval)、摘要系統(Summarization)，以及問答系統(Question Answering)更顯得重要。

隨著網路的急遽發展，人們追求更直覺、便捷、有效率的資訊獲得途徑。在資料量大增的情況下，即使有良好的搜尋檢索系統，所回傳的文字網頁或多媒體量仍遠超過一般人類可閱讀的範圍，而為了讓使用者可以從這些大量的資料中獲得想要的資訊，摘要系統提供精簡化的內容使得使用者可以簡單快速地瀏覽網頁或多媒體內容，並且讓使用者在看過精簡化內容後可以再決定是否要詳細閱覽；使用者甚至不需要閱覽，問答系統從大量的資料中直接擷取使用者想要的答案，僅只提供給使用者問題相關的答案，更加速了使用者獲得資訊的效率。

此外，多媒體與文字相較起來，可以使得使用者方便得到很多資訊，同時也是較為吸引人的一種媒介。在網路多媒體的發展上，語音更是重要的一環，因為語音是人類最自然的基本溝通方式。例如在會議錄音及課程錄音上，單純用純文字的資訊很難提供給使用者表達的重點與傳達的訊息，然而在語音資訊上，隨著語者的抑揚頓挫、輕重緩和，即使沒有文字的轉譯，聽講者卻可以更加容易了解演說內容。近幾年諸如語音信箱、會議錄音到美國史丹佛大學(Stanford University)發展的 Coursera 線上教學網站[1]，亦或是由美國麻省理工(MIT)所發行之課程錄音瀏覽器[2]和電影、餐廳檢索之智慧型手機應用程式[3]等等，各式的多媒體媒

介以大量運用語音資訊技術，快速的發展也證實多媒體資訊對人類不可或缺。而語音文件上因需透過辨識系統而會有辨識錯誤(Recognition Error)的問題產生，辨識錯誤大大的影響做資料檢索、關鍵字抽取、摘要生成等需要文字資訊的研究，也因此，語音資訊的研究比起純文字研究更具有挑戰性。

1.2 本論文研究方向

本論文目標在於解決網路上大量資料所衍生的問題，提出語音摘要系統與語音問答系統的新技術讓使用者更快速方便地獲得資訊，主要研究方向及內容如下。

- 語音文件摘要(Spoken Document Summarization)

  過去的文件摘要研究主要以文字摘要(Text Summarization)為主，意即缺少語音的資訊，而本論文透過加入語音的資訊，主要聚焦於語音文件摘要(Spoken Document Summarization)。本論文以國立台灣大學李琳山教授所開設之數位語音處理概論(Digital Signal Processing; DSP)之課程錄音作為語料，希望藉由結合文字摘要與語音摘要中獨有的特性，例如語者的抑揚頓挫、音量強弱、音調高低等等資訊幫助摘要的選取，而達到比純文字更好的結果。本論文中以兩種演算法實作之：利用加入隱藏變數(Hidden Variable)之結構式支撐向量機(Structured Support Vector Machine)之語音文件摘要，與使用雙層隨機漫步(Two-layer Random Walk)配合課程投影片之語音文件摘要。此語音摘要系統增進語音文件摘要之表現，用以提供學生閱讀，幫助學生學習。如此一來，在沒有聽過完整的課程內容之前，可以先利用語音摘要來對內容略知一二，提供選聽內容的機會。

- 中文語音問答系統(Spoken Question Answering)

  一般傳統的問答系統，大多使用文字問句、文字回答的方式。本論文利用使用者以語音問句輸入，對於網路搜尋引擎進行文件搜索，再對這些搜尋到的文字網頁進行答案的搜尋。本論文利用前 N 最佳結果(N-best)替代單一的語

音辨識結果，使用樹狀隨機域(Tree-structured Conditional Random Fields)配合剖析樹(Parse Tree)，以詞和短語(Phrase)為單位進行搜尋指令生成(Query Formulation)，並且使用雙層隨機漫步(Two-layer Random Walk)對搜尋引擎回傳的網頁進行重排序，讓含有正確答案的網頁能夠排序到越前面。最後，本論文利用網頁排序和查詢指令資訊從網頁文字中抽取答案。

1.3 本論文研究貢獻

本論文的主要貢獻如下：

- 本論文中提出含有隱藏變數之結構式支撐向量機於語音文件摘要，利用結構式支撐向量機可以考慮全域的重複性，而非單獨的考慮單一句子的重要性。此外，由於句子之間具有叢聚的特性，而課程錄音上又有一特性，即叢集句子大多會同時被選作摘要，或是同時不被選擇摘要。此外，若將叢集句子同時選入，也可以增加使用者的可閱讀性，為了模擬此一特性，本論文在結構式向量支撐機中加入了句子叢集之隱藏變數，此隱藏變數的定義在於，訓練資料(Training Data)中不包含這此一資訊，此變數將會從與訓練集中的其他資料整體學習(Joint Learning)得出。
- 本論文提出雙層隨機漫步方法配合課程投影片加強課程錄音之自動摘要系統的效能，由於課程錄音內容通常是照著投影片進行，所以投影片可以視為課程錄音的一種摘要，因此配合課程投影片，課程錄音的辨識轉寫(Transcriptions)可以透過投影片的輔助重新評估每一句子的重要性。透過雙層隨機漫步，投影片內的每一行字可以互相透過相似性找出重要字，辨識轉寫的句子也可以透過重要性互相找出重要句子，而投影片和轉寫句子又可以互相傳遞重要性，因此最後此系統產生之自動摘要可以看作是由課程投影片資訊輔助的摘要。
- 在問答系統中，為了將更精細的文法結構加入查詢指令中，本論文提出樹狀

條件隨機域配合剖析樹結構，將短語(Phrase)加入至查詢指令中，如此一來，可以考慮中文詞與詞組合後造成的語意變化，例如許多書名或是單位名都容易有這個現象，又或者是辨識錯誤透過剖析樹造成的短語容易被濾掉等等，本論文同時利用短語和單詞進行查詢指令生成，在實驗中獲得有效的成果。

- 由於要解決語音辨識使用前 N 最佳結果造成的雜訊增加，在網頁搜尋時回傳的網頁內容也會含有相當多的雜訊，為了解決此一問題，本論文提出使用雙層隨機漫步配合前 N 最佳結果與回傳之網頁內容，目的在於將含有高雜訊或是辨識錯誤的網頁分數降低，以利問答系統做答案之評估。

1.4 章節安排

本論文中第一部分包含第二章的背景知識介紹，首先介紹文字摘要(Text Summarization)，並與語音文件摘要(Spoken Document Summarization)做比較，並簡述摘要系統中的基準方法(Baseline)與評估方法；另外介紹問答系統，包含純文字問答系統和語音問答系統之不同；接下來介紹與本論文有關的機器學習(Machine Learning)方法如向量支撐機(Support Vector Machine)、結構式向量支撐機(Structured Support Vector Machine)和條件隨機域(Conditional Random Fields)，以及圖論(Graph Theory)方法。

第二部分是語音文件摘要部分，包含了第三與第四章：第三章利用含有隱藏變數之結構式支撐向量機(Structured Support Vector Machine with Hidden Variables)配合聲學特徵(Prosodic Features)，結合連續句子組成的文件結構進行摘要抽取；第四章利用雙層隨機漫步(Two-layer Random Walk)配合課程投影片作為輔助做課程錄音摘要。

第三部分為語音問答系統，包含了第五至第七章：第五章主要針對於搜尋引擎為基礎(Information Retrieval based)之問答系統，以樹狀隨機域(Tree-structured Conditional Random Fields; Tree-structured CRF)從問句進行查詢指令生成(Query

Formulation)，第六章為了解決語音辨識上的潛在錯誤而使用之前 N 最佳結果(N-best)中可能含有過多的雜訊(Noise)，因此使用雙層隨機漫步配合前 N 最佳結果進行搜尋網頁之重排序(Re-ranking)，第七章針對搜尋到的網頁內容，利用網頁排序、查詢指令與內容資訊，得到問答系統之答案。

第四部分包含第八章，內容總結本論文提出的方法以及對未來的展望。

第二章 背景知識介紹

2.1 語音文件摘要背景介紹與相關研究

2.1.1 語音文件摘要之分類及應用

純文字摘要系統在過去已受到相當程度的重視與研究[4,5]，其應用目標也有相當大的變化性，而一個文字摘要系統的目的是讓使用者得以更簡單的瞭解文件資訊，系統能將文件資料中重要的資訊擷取出來並傳回給使用者閱讀，意即是將輸入文件精簡化，假定輸入為若干篇文件，輸出則為若干句子或是若干小篇幅的文件。而依照系統用途之不同，文件摘要也可分為不同類型。以下將一一做介紹，如圖 2.1 所示。

- 以輸入文件數目分類

  以文件做分類的摘要類型分為單一文件摘要(Single Document Summarization)及多文件摘要(Multi-Document Summarization)兩種，兩者主要的差別在於產生一篇摘要所需的輸入文件的多寡。在單一文件摘要中，會對每一篇文件產生一篇摘要，如本論文於課程錄音(Lecture)上的實驗，由於每篇文件對應到老師每個禮拜的上課內容，屬於不同主題，因此對每一篇我們都會產生一篇屬於該主題的文件摘要，此屬於單一文件摘要。而多文件摘要則是將多篇文件一起產生一篇摘要，此種方法大多應用於網路摘要系統(Web Summarization)，由於在做網頁搜尋時，使用者常常會給予一個查詢指令(Query)，進而得到一串內容進的網頁排序，然而過多的資訊讓人目不暇給，網路摘要系統大多利用多文件摘要將多數個網頁的內容彙整成一篇摘要提供使用者閱讀，也由於近幾年來網路的發達，多文件摘要也成為近年來較熱門的研究方向。

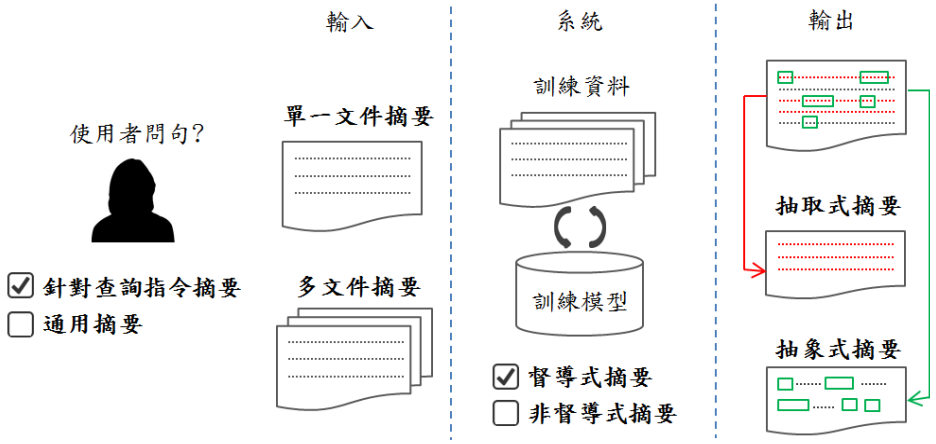

圖 2.1　文件分類摘要示意圖

- 以查詢指令分類

  以查詢指令分類的摘要類型分為通用型摘要(Generative Summarization)，以及針對查詢指令之摘要(Query-focus Summarization)兩種，主要的差別在於是否考慮問句的資訊。通用摘要即為無使用查詢指令的摘要系統，如本論文中的課程摘要即為此類，及目的是讓摘要的內容能夠儘量含括越多的文件資訊。針對查詢指令一般與網路摘要系統結合，如先前所說，使用者給定查詢指令透過網頁瀏覽器作資料搜尋(Information Retrieval)，而針對查詢指令摘要則利用使用者的查詢指令回傳對應之摘要，其目標並非含括所有的文件資訊，只是給使用者與查詢指令相關的資訊，如微軟(Microsoft)所發行之瀏覽器(bing)[6]即提供此種功能，又或者如 2.2 小節將會介紹之定義式問答，使用者透過問句詢問某種東西的定義或是解釋，可以將問題的答案看做是多篇文件針對問句內容所找出的摘要。

- 以摘要形式分類

  此類別分為抽取式摘要(Extractive Summarization)與抽象是摘要(Abstractive Summarization)，抽取式摘要將文件作斷句後選出部分句子作文摘要輸出，此種方式將摘要系統簡化成為如何選出重要句子的問題。而抽象式摘要則是另外製造出若干句子(未必是原文中的句子)作為輸

出，簡而言之抽象式摘要類似於人類產生之摘要，透過自己的想法和思考而輸出一串經自我想法更改過的敘述，也因此抽象式摘要在自然語言處理上一直被認為是相當困難的問題，絕大多數的研究都還是著重於抽取式摘要，如本論文於課程錄音上的實驗均屬於抽取式摘要。

2.1.2 語音文件摘要與文字摘要之不同

語音文件摘要與純文字文件摘要的不同在於語音文件是聲音檔，如課程錄音(Lecture)、會議錄音(Meeting)和新聞錄音(News)等，這些錄音必須透過語音辨識(Speech Recognition)系統將聲音轉寫成文字以方便接下來的研究，也因此就有聲音辨識錯誤(Recognition Error)的產生，而辨識錯誤會對摘要系統造成影響，所以辨識系統的好壞也對於摘要結果有絕對的影響力。再者，語音文件不如純文字文件單純，特別如課程錄音和會議錄音，此類語音文件的語者通常是未經訓練的人，內容較為自發性(Spontaneous)，語句不流暢(Disfluency)，辨識率也會較低。此外，因為是文件是聲音檔，如何斷句也將會是一大問題，如果錯將某一重要句子斷成若干瑣碎句子，則這些句子可能將都不會被選入摘要，又或者是會嚴重影響到挑選出來的摘要的可閱讀性。

如上所述，語音文件摘要要比純文字摘要有更多的挑戰和困難，如語意不流順及斷句皆有研究指出會大大影響摘要效能[7,8]。而過去也有提出使用前 N 最佳(N-best)的辨識結果之方法以解決辨識錯誤的問題[9,10]。但語音文件也有其優勢之處，透過聲學特徵的分析，如語者的語調高低、聲音強弱、快慢等等訊息，我們可以判斷出語者所說的內容是否重要。如課程錄音中，語者語調變高和提高音量的句子，極有可能是在強調某重要的觀念或理論，此種特性可以幫助判斷此句是否適合為摘要，因此在過去的研究中藉由聲學特徵的幫助，語音文件摘要的表現甚至有可能比純文件文字摘要的表現還要理想[11-17]。

2.1.3 傳統語音文件摘要之方法-最大邊際關聯法 (MMR)

最大邊際關聯法(Maximum Margin Relevance)[18,19]，自 1998 年被提出，用於針對問句摘要研究，為一種解目標函數(Objective Function)的貪婪演算法(Greedy Algorithm)。由於其簡單、快速、複雜度低又不需要大量人工標記的特性，一直到今天在各個摘要研究中都仍為抽取式摘要中之基準 (Baseline)。最大邊際關聯法的目標函數如下。

$$S \approx arg\max_{S \in D} \sum_{S_i \in D} rel(S_i, D) - \lambda \sum_{S_i, S_j \in S} red(S_i, S_j) \qquad (2.1)$$

其中S為已選擇之摘要，包含若干句子，D則為某篇文件，S_i代表文件中的單一句子。這個目標函數可以分為兩部分，其中$rel(S_i, D)$代表關聯性分數(Relevance Score)，用於判斷句子S_i與整篇文件D的相似程度，若$rel(S_i, D)$的分數越高代表句子S_i和文件D的相似度越高，因此S_i越有可能是文件中的重要主題，或者是語者一再強調的句子。而$red(S_i, S_j)$代表重複性分數(Redundancy Score)，用於判斷已選擇摘要S的重複程度。而$red(S_i, S_j)$部分的S_i與S_j為已選擇摘要S中的任兩句句子，若分數越高則代表兩句子越像，而最後在對任兩個S中的句子相加，因此可看出S中句子相互的相似程度。此目標函數的目的在於希望由文件D中，在限定的長度之下選出若干句子的集合S，此集合中每句句子與文件的相似度要夠高(關聯性分數高)，且集合S中的句子間相似程度要低，λ為一參數用於調整關聯性分數與重複性分數的權重。此目標函數於摘要系統中被廣泛應用且效果相當不錯，因一般情況下摘要系統希望抽出的句子屬於文件中的重要主題，且所選出的句子又不可以太像，以避免摘要中的內容均為重複性太高的句子，因而讓使用者無法從摘要中得知所有原本文件中的重要資訊。

然而要解此目標函數以達全域最佳(Global Optimal)解並不容易，要窮舉所有可能的句子集合相當耗時且號記憶體，因此替代方案是使用貪婪演算法(Greedy

Algorithm)，先令集合S為空集合，而在每個迭代(Iteration)中選出一個句子能使目前的目標函數最大：

$$\mathrm{MMR}(S_i) = rel(S_i, D) - \lambda \sum_{S_j, S_k \in S} red(S_j, S_k) \tag{2. 2}$$

在第一個迭代中因為集合S為空集合，因此$red(S_j, S_k)$的分數為零，而每次迭代會選擇邊際關聯分數$\mathrm{MMR}(S_i)$最高的句子加入集合S中，而這時$red(S_j, S_k)$項的分數也會因此改變，重複這些步驟直到集合S中的句數或字數達到預定的標準後，此集合便是利用最大關聯邊際法找出的摘要。

2.1.4 語音文件摘要之評估

ROUGE (Recall-Oriented Understudy of Gisting Evaluation)[21]為一被廣泛使用的摘要評估方式，用於判斷自動摘要與人工標記之參考摘要之間的相似度，如果自動摘要與參考摘要的相似度越高，則此自動摘要的品質越佳。此論文中使用ROUGE-N(N=1,2,3)以 ROUGE-L 的評估摘要系統的優劣。而以下將先介紹準確率、召回率以及 F 評估(F-measure)[22]之概念，再介紹 ROUGE 的評估方式。

準確率(Precision)表示有多少比例之自動摘要句子同時包含在參考摘要的標準答案之中，其值越高代表自動摘要產生出的句子越多被包含在參考摘要之正確句子中，也代表正確性越可靠；而召回率(Recall)則代表著參考摘要的標準答案的句子中，有多少比例的句子被包含在自動摘要中，其值越高代表有越多的參考摘要句子被選取至系統之自動摘要中。其定義如下：

$$\text{準確率} = \frac{\text{同時為參考摘要句子與自動摘要句子數}}{\text{自動摘要句子總數}} \tag{2. 3}$$

$$\text{召回率} = \frac{\text{同時為參考摘要句子與自動摘要句子數}}{\text{參考摘要句子總數}} \tag{2. 4}$$

在摘要系統中，摘要長度的設定，會對準確率以及召回率造成不同影響。長度越長會使得準確率下降，但同時召回率會上升，故單一使用準確率或召回率進行摘要評估無法完整顯現系統的整體效能，因此利用 F 評估將兩者作結合，用此評估系統效能較為客觀且完整。

$$\text{F 評估} = \frac{2 \times \text{準確率} \times \text{召回率}}{\text{準確率} + \text{召回率}} \tag{2.5}$$

而 ROUGE 則是使用了 F 評估方式的一種摘要評估方法。在本論文中我們使用了普遍的 ROUGE-N(N=1,2,3)以及 ROUGE-L 分數。ROUGE-N 代表自動產生摘要與人工摘要間 N 連文法(N-gram)的 F 評估，即是以 N 連文法的出現次數之準確率與召回率進行摘要評估。另外 ROUGE-L 的計算方式和 ROUGE-N 十分相似，但只考慮自動摘要與參考摘要的「最長共同子字串(Longest Common Subsequence; LCS)」。在相關的研究中，主要以 ROUGE-1 與 ROUGE-L 被看作是較佳的評估摘要的指標。

2.2 語音查詢指令之問答系統

問答系統(Question Answering; QA)和搜尋檢索系統最大的不同在於，問答系統只針對使用者輸入的問句回傳對應之答案，而不是提供使用者相關的資訊，讓使用者自己找答案。此種系統的優點在於，面對龐大的資料量或者是很多複雜內容的文件，使用者有時候無法從眾多文字中找出對應的答案，因此問答系統透過文字處理，直接將可能的答案回傳，以方便使用者閱覽。

通常問答系統可以根據答案的種類分成三種：模擬陳述問答(Factoid QA)[23,24,25]、定義式問答(Definitional QA)以及複雜式問答(Complex QA)[26,27]。模擬陳述問答的答案是明確且一致的，例如「舊金山在哪個國家？」、「2008 年奧運在哪個城市舉行？」等等；而定義式問答通常是要求問答系統簡述某個定義，

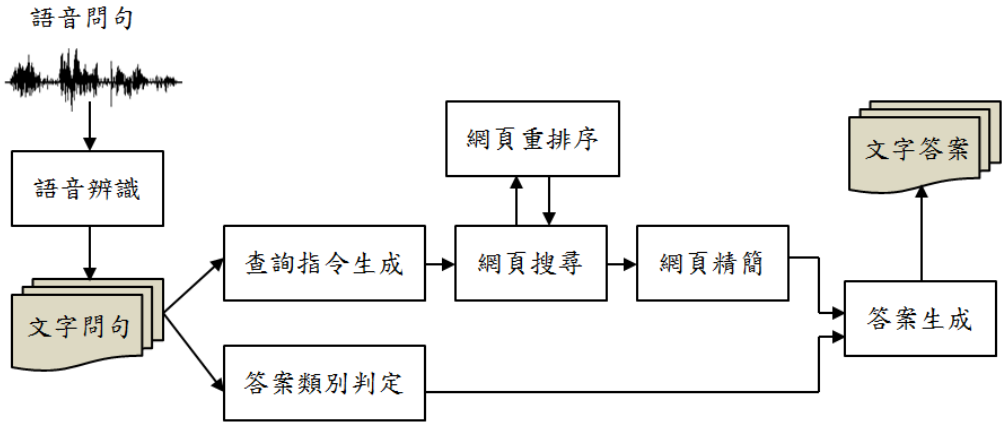

圖 2.2　以搜尋檢索基礎之問答系統之基本架構

其回答由於自然語言的表達方法有很多，所以答案不會一致，但內容應該是相近的，例如「請問語音文件的定義為何？」，其內容和針對查詢指令之文件摘要相仿，對於問句所針對的主題，將回傳只和這些主題相關的內容；複雜式問答系統通常沒有一個標準的答案，往往跟回答者的主觀想法或是立場有關，因此被視為最困難且難以評估的一種問答，諸如「台灣大學附近的推薦美食有哪些？」。本論文討論的問答系統皆採用模擬陳述問答型式。

2.2.1 檢索為基礎之問答系統

這一大類的問答系統利用網路上眾多的資訊，從搜尋引擎中回傳的相關網頁再進行答案的生成，圖 2.2 說明一般搜尋檢索基礎的問答系統[28-31]的系統架構，若是單純文字問答系統不用加入前端語音辨識的部分，系統從問句生成查詢指令(Query)和與答案類別(Answer Type)相關的詞。由於問句通常以自然語言型態表達，必須將問句中和問題內容相關的詞抽取出來，重新組成查詢指令，透過查詢指令可以讓搜尋引擎回傳相關的文件，也由於網頁資料通常含有相當多的不相關資訊，例如網頁的會員系統、廣告等等，進而系統需要對這些相關文件做精簡化的摘要；而答案類別是代表這個問題的答案是屬於哪一種類型，例如「時間」、「人名」、「地名」或是「數字」等等，答案判定往往在問答系統最後決定候選答案時

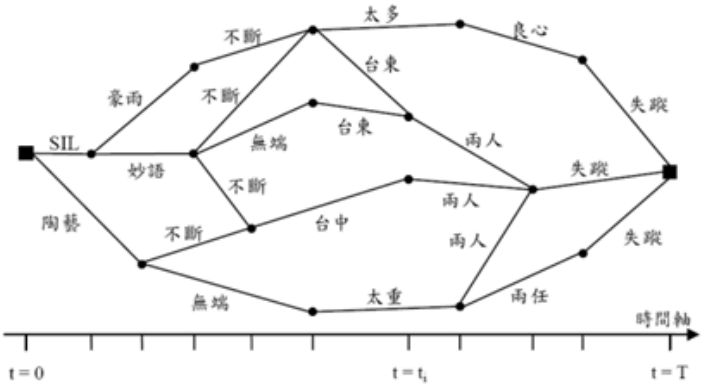

圖 2.3　語音辨識技術的詞格(Lattice)示意圖

具有關鍵性的角色，如果再答案類別判定就錯誤，那麼最後的答案將會錯誤。而具備有相關文件內容與答案類別後，系統針對這些問題做類專有名詞解析(Named Entity Resolution)，若是某字詞的類專有名詞和答案類別一致，那麼此字詞即有可能是此問題的答案。目前許多系統如Google搜尋引擎的部分問句處理、線上版功能的 IBM Watson 機器人[32,33,34]，以及由 TREC 主辦的問答系統競賽[35]都是屬於這一類別，本論文在第三部分的語音問答系統即是基於搜尋檢索基礎之問答系統。

2.2.2 知識為基礎之問答系統

知識為基礎之問答系統[36,37]與檢索為基礎的問答系統最大的不同在於，知識為基礎之問答系統通常有一個分類完整、結構嚴謹的資料庫，也比起含有大量雜訊的網路文件更加精簡，尤其針對專門領域如藥學、生物學的問答系統特別有效，但缺點在於需要花相當多的心力去建構此一資料庫，且建構者需要相關領域的知識。知識基礎之問答系統主要著重在如何將問題劃分到資料庫的結構，例如說「國家」相關的資料庫中，如何將問句「占地最大的國家？」對應到「面積」的項目。目前廣泛被使用的知識為基礎之問答系統包含 Apple 公司在智慧型手機上搭載的 SIRI、IBM Watson、Wolfram Alpha[38]等等。

2.2.3 語音資料之前 N 最佳結果

一般最常見的語音辨識結果，會對輸入的語音檔得到對應的文字檔，此結果被稱作是最佳結果(One-best)，但通常正確的詞不一定會在最佳結果中，卻會在詞格(Lattice)中存在。圖 2.3 為語音辨識技術的詞格示意圖，詞格為在這段聲音訊號中，眾多種可能的辨識假設(Hypotheses)的組成，每一條在詞格上的弧，便代表一個詞，而最佳結果即是在詞格上所能得到最高分數的一條路徑，此分數為語言模型(Linguistic Model)和聲學模型(Acoustic Model)加權總合而來。

為了將更多可能是正確的答案包含在辨識結果中，我們可以考慮更多的辨識結果，而其中前 N 最佳結果，即是考慮前 N 條高分的路徑。通常前 N 最佳結果或是詞格，可以用在資料檢索的研究中，讓未包含在最佳結果中的詞可以被檢索出來，但其代價為雜訊的提高，雖然可能包含了更多正確的詞，同時也可能加入更多辨識錯誤的詞。本論文中使用前 N 最佳結果於問答系統的查詢指令生成，目的在於能夠包含更多的正確辨識結果進入查詢指令中。

2.2.4 語音查詢指令之問答系統

而語音問答系統涵蓋的三種可能，其一是輸入問題是語音，其二是資料庫或知識庫是語音，其三是兩者皆為語音，本論文中的問答系統屬於第一種，即輸入問題是語音，而判別答案的資料庫是純文字的網頁。而當輸入來源是語音時，便會面臨到辨識錯誤的問題，辨識錯誤在語音問答系統中帶來的影響會遠大於語音文件摘要，其原因在於語音文件摘要可利用句子和句子之間的相似度做特徵，或是聲音的抑揚頓挫作聲學特徵，即使有辨識錯誤，重要的句子也仍可以包含在語音文件摘要中，相對地，語音問答系統只要辨識出來的文字有錯誤，就會影響到資料庫檢索出來的相關文件，若是關鍵字被辨識錯誤，甚至會導致對於整個問題的理解錯誤。

為了解決上述的問題，語音問答系統可以使用前 N 最佳結果或是詞格取代最佳結果，如此一來，便可以將多種辨識結果考慮在內，可以讓辨識結果更有可能搜尋或是對應到正確的內容。不過這種做法會使辨識結果含有的雜訊增加，可能讓更多的辨識錯誤詞被考慮，因此如何去降低雜訊造成的影響，也都是語音問答系統必須要考慮的一大難題。

2.3 機器學習 (Machine Learning)

近幾年來，機器學習的方法日趨成熟，無論是影像處理、聲音辨識、社群網路都可以看到其蹤跡，在多媒體上廣為應用。而機器學習的目的，是利用特徵(Features)讓機器亦或者說是電腦學習分類的方法。其中如早期的類神經網路(Neural Network)[39]、條件隨機域(Conditional Random Fields)[40,41,42]、單純貝氏分類(Naïve Bayes Classifier)[43]、支撐向量機(Support Vector Machine)[44,45]以及最近廣泛被使用的深度學習(Deep Learning)[46]等等都屬於機器學習的範疇。

在機器學習領域中可將資料分為兩類，分別是訓練集(Training Set)以及測試集(Testing Set)，兩者分別都有抽取出來的特徵向量(Feature Vector)。訓練集中有人工標註的輸出答案，而測試集中答案未知，需靠電腦由訓練集中訓練出來的分類器(Classifier)來判斷其輸出為何。舉例來說，在本論文中訓練集由若干句子組成，而每個句子都有其特徵諸如該句子的字數、是否含關鍵字、是否有英文、該語句的平均音節(Syllable)發音長度等等，若該句子為摘要句之一，其人工標記的輸出答案為+1，反之則為-1，測試集則為其他若干句子，沒有人工標記的答案。以上的方法均為二元分類法(Binary Classifier)，更為複雜的方法為本論文中使用的結構式支撐向量機(Structured Support Vector Machine)[47-50]，其輸入是含有若干句子的一整篇文件，輸出是被選作為摘要的句子集，使用比二元分類法更為複雜的機器學習方法可以透過結構學習(Structured Learning)與整體學習(Joint Learning)的方式得到。

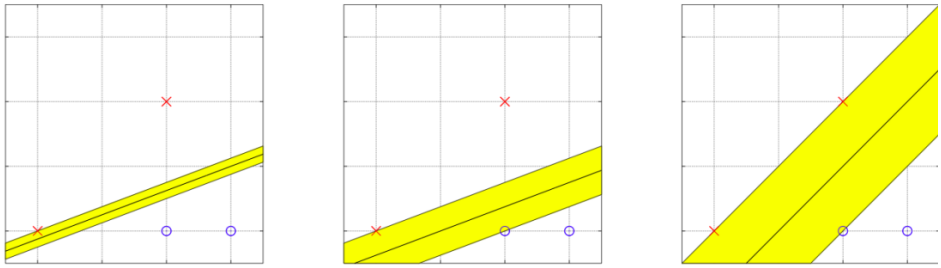

圖 2.4　最大邊際示意圖

而根據訓練資料的有無多寡，機器學習可分為三類：督導式學習(Supervised Learning)、非督導式學習(Unsupervised Learning)以及半督導式學習(Semi-supervised Learning)三種。督導式學習在於訓練資料都有標記好的正確輸出，意即對每一個訓練集中的訓練單位都有人工標記的答案(如二元分類法的+1與-1)；而非督導式學習中訓練資料都沒有標記好的正確輸出，即使沒有正確的標記的答案，仍然可以用基於規則(Rule-based)的方法或是叢集(Clustering)的方式將測試集中的資料做分類。通常督導式的方法因為具有人工標註的輔助，可以得到較好的結果，然而人工標註的取得較為困難且費時費工，因此半督導式的方法是僅利用少量人工標記的答案來幫助學習。

2.3.1 支撐向量機 (Support Vector Machine)

在二分演算法(Binary Classification)中，可以利用簡單的線性模型(Linear Model)對資料做切割分類，然而，這面臨到了許多問題。第一，在無限多個可能的線性模型中，如何選出較好的一個線性平面。第二、資料的分布可能為非線性可分割。支撐向量機起源於最大邊際(Maximum Margin)的觀念，並與正則化(Regularization)和核心(Kernel)映射的觀念做連結，用以解決以上兩個問題，如圖2.4 中為兩類資料的分佈，分別依藍色和紅色表示，而中間的分隔線則為二維平面上的一條線，若資料點離線的距離越遠，表示越低可能分類錯誤，也就是具有較高的信心分數，因此支撐向量機除了要把資料分隔，同時還希望把這些資料點

離分隔線越遠越好，故支撐向量機定義離分隔線最近的資料點為支撐向量(Support Vector)，而由這些支撐向量與分隔線之間的距離，便稱做邊際(Margin)，而支撐向量機的目標就在於將邊際最大化。

在此定義一組高維空間中的超平面(Hyper-plane)：

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b \tag{2. 6}$$

此組超平面的目的在於，將兩種標記的資料區隔開，也就是擁有-1 標記的資料，應該都是落在$f(\boldsymbol{x}) < 0$的區塊，而+1 標記的資料應當在$f(\boldsymbol{x}) > 0$區塊，並且目標在於可以有越大的邊際越好。此外，假定從兩側支撐向量劃分的兩區塊為：

$$\begin{aligned} \boldsymbol{w}^T\boldsymbol{x} + b \geq 1, \text{if y} = 1 \\ \boldsymbol{w}^T\boldsymbol{x} + b \leq 1, \text{if y} = -1 \end{aligned} \tag{2. 7}$$

我們又可以將上公式化簡成：

$$y(\boldsymbol{w}^T\boldsymbol{x} + b) \geq 1 \tag{2. 8}$$

而根據平面之間的距離公式，這兩個區塊之間的邊際為$\frac{2}{\|\boldsymbol{w}\|}$，支撐向量機的目的就是要使得這個邊際值最大，也就是使得$\frac{\|\boldsymbol{w}\|}{2}$最小，同時也要滿足此平面可以將每一筆資料點$(\boldsymbol{x}_i, y_i)$區分在對的區塊。故在此定義支稱向量機的目標函數：

$$\begin{aligned} &\min_{w} \frac{1}{2}\|\boldsymbol{w}\| \\ &subect\ to\ \ y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, \forall \text{i} \end{aligned} \tag{2. 9}$$

此一方程式可以透過二次規畫（Quadratic Programming）求解。

在介紹完支撐向量機的原始型態後，以上例子仍為線性可分割，而當資料為非線性可分割時，必須使用一個轉換函數(Transfer Function)將資料中的所有樣本的值轉換到另一個線性可分割的高維空間，使用對偶形式(Duality Form)與核心映

射(Kernel Projection)可以解決這個問題。

首先，為了解決具有限制條件的最佳化問題(Constrained Optimization Problem)通常比無限制條件的問題困難許多，但數學上有一拉氏乘數法(Lagrange Multiplier)可用以解決有限制條件最佳化之問題。首先將(2. 9)加入拉式輔助函數(Lagrange Function)。

$$\mathcal{L}(\mathbf{w}, \mathrm{b}, \alpha) = \frac{1}{2}\|\boldsymbol{w}\| - \sum_{i=1}^{N} \alpha_i (y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1) \tag{2. 10}$$

其中α_i稱為拉氏乘數(Lagrange Multiplier)，而根據拉氏乘數法，拉氏輔助函數必須要符合$\frac{\partial \mathcal{L}(\mathrm{w},\mathrm{b},\alpha)}{\partial \mathrm{w}} = 0$與$\frac{\partial \mathcal{L}(\mathrm{w},\mathrm{b},\alpha)}{\partial b} = 0$，因此可以得到：

$$\begin{aligned} \frac{\partial \mathcal{L}(\mathrm{w}, \mathrm{b}, \alpha)}{\partial \mathrm{w}} = 0 \Leftrightarrow w = \sum_{i=1}^{N} y_i \alpha_i \boldsymbol{x_i} \\ \frac{\partial \mathcal{L}(\mathrm{w}, \mathrm{b}, \alpha)}{\partial \mathrm{b}} = 0 \Leftrightarrow \quad \sum_{i=1}^{N} y_i \alpha_i = 0 \end{aligned} \tag{2. 11}$$

於是(2. 9)配合(2. 11)得出的限制條件，可以改寫成為：

$$\max_{all\ \alpha_i \geq 0} \left\{ \min_{b,w} \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i (y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1) \right\}$$

$$= \max_{all\ \alpha_i \geq 0} \left\{ \min_{b,w} \frac{1}{2} w^T w - (\sum_{i=1}^{N} \alpha_i\, y_i \boldsymbol{x}_i{}^T)\boldsymbol{w} - \left(\sum_{i=1}^{N} y_i \alpha_i \right) b + \sum_{i=1}^{N} \alpha_i \right\} \tag{2. 12}$$

$$= \max_{all\ \alpha_i \geq 0} \left\{ -\frac{1}{2} (\sum_{i=1}^{N} \alpha_i\, y_i \boldsymbol{x}_i{}^T)(\sum_{i=1}^{N} \alpha_i\, y_i \boldsymbol{x}_i) + \sum_{i=1}^{N} \alpha_i \right\}$$

而上述的推導又可以化簡為以下的矩陣形式：

$$\min_{w} \frac{1}{2}\boldsymbol{\alpha}^T \begin{bmatrix} y_1 y_1 \boldsymbol{x}_1^T \boldsymbol{x}_1 & \cdots & y_1 y_N \boldsymbol{x}_1^T \boldsymbol{x}_N \\ \vdots & \ddots & \vdots \\ y_N y_1 \boldsymbol{x}_N^T \boldsymbol{x}_1 & \cdots & y_N y_N \boldsymbol{x}_N^T \boldsymbol{x}_N \end{bmatrix} \boldsymbol{\alpha} - \mathbf{1}^T \boldsymbol{\alpha}$$

$$subect\ to\ \ \boldsymbol{y}^T \boldsymbol{\alpha} = 0 \text{ and } 0 \leq \alpha, \tag{2. 13}$$

$$w = \sum_{i=1}^{N} y_i \alpha_i \boldsymbol{x_i}$$

上述式子已經簡化成一個二次規劃問題，在數學領域中已被廣泛的研究與探討，而透過二次規劃方程，我們可簡單地求得拉式乘數(Lagrange Multiplier) $\boldsymbol{\alpha}$，並經(2. 11)推得超平面w。

此外，為了解決非線性可分割之情況，必須要經過轉換至較高維度的空間，因此制定一個轉換方式$\phi(x_i) = z_i$，將x_i轉換到高維度的z_i，而希望在較高維的空間中是線性可分割的。經過特徵轉換的對偶形式和(2. 13)幾乎相同，只是將所有的x_i變成做過特徵轉換的z_i。

$$\min_{w} \frac{1}{2}\boldsymbol{\alpha}^T \begin{bmatrix} y_1 y_1 \boldsymbol{z}_1^T \boldsymbol{z}_1 & \cdots & y_1 y_N \boldsymbol{z}_1^T \boldsymbol{z}_N \\ \vdots & \ddots & \vdots \\ y_N y_1 \boldsymbol{z}_N^T \boldsymbol{z}_1 & \cdots & y_N y_N \boldsymbol{z}_N^T \boldsymbol{z}_N \end{bmatrix} \boldsymbol{\alpha} - \mathbf{1}^T \boldsymbol{\alpha}$$

$$subect\ to\ \ \boldsymbol{y}^T \boldsymbol{\alpha} = 0 \text{ and } 0 \leq \alpha, \tag{2. 14}$$

$$w = \sum_{i=1}^{N} y_i \alpha_i \boldsymbol{z_i}$$

此外，一般的現實的狀況是資料大多是不可完全被分割，也就是並不是所有的資料點都會落在應該對應到的超平面切割的空間區塊，為了解決這樣的問題，因此允許有分類錯誤的情況發生，且將此允許錯誤發生的鬆弛變數（Slack Variable）在加入至目標函數：

$$\min_{w} \frac{1}{2}\|\boldsymbol{w}\| + C\sum_{i=1}^{N} \xi_i \tag{2. 15}$$

$$subect\ to\ \ y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \forall \mathrm{i}$$

$$\xi_i \geq 0, \forall \mathrm{i}$$

由上式可以看出，加入鬆弛變數後，$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)$這個限制便可能會小於 1，又甚是會是負的，故可以容許部份分類錯誤。若是樣本差異太大時，會在目標函數中付出$C\xi_i$的代價，此C是一人為設定之參數，當C越大，越不容許分類錯誤的情況。而此式的解法和上述均相同，唯獨多了鬆弛變數，在對偶形式上多了個$\alpha \leq \mathrm{C}$的限制。

$$\min_{w} \frac{1}{2}\boldsymbol{\alpha}^T \begin{bmatrix} y_1 y_1 \boldsymbol{z}_1^T \boldsymbol{z}_1 & \cdots & y_1 y_N \boldsymbol{z}_1^T \boldsymbol{z}_N \\ \vdots & \ddots & \vdots \\ y_N y_1 \boldsymbol{z}_N^T \boldsymbol{z}_1 & \cdots & y_N y_N \boldsymbol{z}_N^T \boldsymbol{z}_N \end{bmatrix} \boldsymbol{\alpha} - \mathbf{1}^T \boldsymbol{\alpha}$$

$$subect\ to\ \ \boldsymbol{y}^T\boldsymbol{\alpha} = 0 \text{ and } 0 \leq \alpha \leq \mathrm{C}, \qquad (2.16)$$

$$w = \sum_{i=1}^{N} y_i \alpha_i \boldsymbol{z_i}$$

2.3.2 結構式支撐向量機 (Structured Support Vector Machine)

結構式支撐向量機(Structured Support Vector Machine)和支撐向量機最大的不同在於，輸入與輸出不在只是針對單一資料做標記分類，它可以使用具結構化的輸入輸出，如鏈狀、樹狀等等，因此，結構向量機可以考慮輸入資料之間的相關性或是輸出標記之間的關係，不再侷限於支撐向量機中每筆資料必須為獨立考慮之限制。舉例來說，在資訊檢索的結構式學習，其輸出可能是排序列表(Ranking List)，以語言剖析為例，則輸出可能是一個剖析樹(Parse tree)，而若是抽取式的摘要，其輸出則是文件中句子的子集合。

首先定義一組鑑別函數(Discriminative Function) $F: \mathcal{X} \times \mathcal{Y} \to \mathcal{R}$：

$$F(x, y; w) = \langle w, \phi(x, y) \rangle \qquad (2.17)$$

以上w為支撐向量機中所欲訓練求得之權重，$\langle\ \rangle$為內積，$\phi(x, y)$為一組特徵向量，鑑別函數的作用在於將結構化的(x, y)對轉換成一組實數。在訓練(Training)階段

時，則希望使得正確標記(x_i, y_i)得到之鑑別函數$F(x_i, y_i; w)$可以大於其他種可能標記之鑑別函數，意即：

$$F(x_i, y_i; w) > F(x_i, y; w), \forall y \neq y_i \tag{2. 18}$$

也就是在透過鑑別函數分數之差異，可以將正確標記和其他標記區分開來。此外方便起見，我們先定義作為之後化簡方程式用：

$$\delta\phi_i(y) = \phi(x_i, y_i) - \phi(x_i, y) \tag{2. 19}$$

而在測試(Testing)階段，目標在於用已知之w與輸入x_i時，得到可以使鑑別分數最高分之標記y^*作為輸出：

$$y^* = arg\max_{y \in \mathcal{Y}} F(x_i, y; w) \tag{2. 20}$$

由上述可知，此結構式向量支撐機，一樣目標在於找到一組超平面最大化訓練集中資料與其他標記資料之間的邊際，於是我們改變(2. 15)中的限制部分：

$$\begin{gathered} \min_{w} \frac{1}{2}\|\boldsymbol{w}\| + C\sum_{i=1}^{N} \xi_i \\ subect\ to\ \langle w, \delta\phi_i(y)\rangle \geq 1 - \xi_i, \forall \mathrm{i}, \forall \mathrm{y} \in \mathcal{Y} \backslash y_i \\ \xi_i \geq 0, \forall \mathrm{i} \end{gathered} \tag{2. 21}$$

此外在上述的條件中，我們對於某一輸入x_i給予相同的鬆弛變數，也就是不管輸出標記y和正確標記y_i的相似程度，皆把每種y視為同等。為了解決此情況，必須要加入減損函數(Loss Function) $\Delta(y_i, y)$: $\mathcal{Y} \times \mathcal{Y} \to \mathcal{R}$，用來判斷某個輸出標記$y$和正確標記$y_i$的差別，其值介於 0 與 1 之間，如果兩者越像，則$\Delta(y_i, y)$的值越小，反之則越大。而減損函數加入目標函數的方式有二，第一種是改變不同標記y所給的懲罰大小：

$$\min_{w} \frac{1}{2}\|\boldsymbol{w}\| + C\sum_{i=1}^{N}\xi_i$$
$$subect\ to\ \langle w, \delta\phi_i(y)\rangle \geq 1 - \frac{\xi_i}{\Delta(y_i, y)}, \forall \mathrm{i}, \forall \mathrm{y} \in \mathcal{Y}\backslash y_i \tag{2. 22}$$
$$\xi_i \geq 0, \forall \mathrm{i}$$

而第二種方式則是在邊際上做調整：

$$\min_{w} \frac{1}{2}\|\boldsymbol{w}\| + C\sum_{i=1}^{N}\xi_i$$
$$subect\ to\ \langle w, \delta\phi_i(y)\rangle \geq\ \Delta(y_i, y) - \xi_i, \forall \mathrm{i}, \forall \mathrm{y} \in \mathcal{Y}\backslash y_i \tag{2. 23}$$
$$\xi_i \geq 0, \forall \mathrm{i}$$

然而在這樣的情況下，必須考慮所有可能的y標記，其數量可能甚至會到指數數量級成長，因此無法如同先前支撐向量機直接對目標函數求解，為了解決過多的限制條件的問題，我們將目標函數調整至只考慮最可能違反分類的標記，也就是考慮能使得以下式子最大化之標記y'：

$$y' = \underset{y\in\mathcal{Y}\backslash y_i}{\operatorname{argmax}}\, \Delta(y_i, y) + F(x_i, y; w) \tag{2. 24}$$

如此一來，針對每一筆輸入x_i，便不需要窮舉所有可能的y組合，只需要將y'加入條件限制集合之中，我們可以進一步更改(2. 23)：

$$\min_{w} \frac{1}{2}\|\boldsymbol{w}\| + C\sum_{i=1}^{N}\xi_i$$
$$subect\ to\ \langle w, \delta\phi_i(y')\rangle \geq\ \Delta(y_i, y') - \xi_i, \forall \mathrm{i} \tag{2. 25}$$
$$\xi_i \geq 0, \forall \mathrm{i}$$

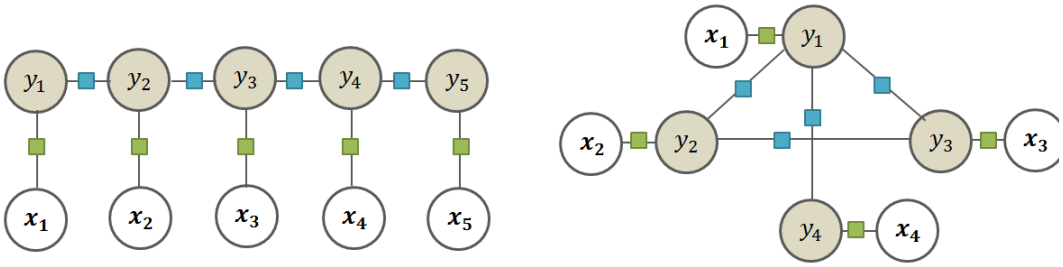

圖 2.5　條件隨機域示意圖

2.3.3 條件隨機域 (Conditional Random Fields)

條件隨機域（Conditional Random Fields; CRF）是一種無向性的圖模型（Graphical Model），對於隨機域內的(x, y)對，找到對應之最大條件機率標記，此種模型被大量的運用在序列標記（Sequential Labeling），其應用領域廣泛，例如自然語言處理（Natural Language Processing）、生物基因序列研究等等。

條件隨機域並沒有一個特定的形式，但最廣泛被使用的是如圖 2.5 左半部所示的直鏈狀條件隨機域（Linear-chain CRF），而如圖 2.5 右半部則是通用型的條件隨機域，每個隨機變數間並沒有順序性，也沒有固定的連結方法。此章節中主要介紹直鏈狀的條件隨機域，假定隨機變數$\boldsymbol{y}$互相有順序性，每一個變數y_t和前一個變數y_{t-1}具有相依性，而每一個變數y_t都有對應到的觀察變數x_t，因此y_{t-1}和x_t對於y_t的標記均有影響力。在此定義此條件隨機域的可能性（Likelihood）：

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(x)} \prod_{t=1}^{T} \phi(y_t, y_{t-1}, \boldsymbol{x}) \tag{2. 26}$$

其中$\phi(y_t, y_{t-1}, \boldsymbol{x})$代表和每一個$y_t$相關所構成的特徵函數（Feature Function），又可以如圖 2.5 中所示劃分為過渡部份(y_t, y_{t-1})與觀察部份$(y_t, \boldsymbol{x})$兩部份：

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(x)} \prod_{t=1}^{T} \phi_t(y_t, y_{t-1}) \cdot \phi_o(y_t, \boldsymbol{x}) \tag{2. 27}$$

每一組的特徵函數$\phi_t(y_t, y_{t-1})$與$\phi_o(y_t, \boldsymbol{x})$皆可以看成是一組參數$\boldsymbol{\lambda}$與特徵向量（Feature Vector）的內積，同時我們再將函數轉換相乘改成相加型式：

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(x)} \sum_{t=1}^{T} \exp\left(\boldsymbol{\lambda_t} \cdot f_t(y_t, y_{t-1}) + \boldsymbol{\lambda_o} \cdot f_o(y_t, \boldsymbol{x})\right) \tag{2. 28}$$

其中$\boldsymbol{\lambda_t}$與$\boldsymbol{\lambda_o}$兩組參數即為條件隨機域中要求解的目標參數，本論文之後將會以$\boldsymbol{\lambda}$代表$[\boldsymbol{\lambda_t}, \boldsymbol{\lambda_o}]$組成的參數集。再者，為了方便求解，先將(2. 28)式轉換成對數（LOG）可能性型式：

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x}) &= \sum_{t=1}^{T} \left(\boldsymbol{\lambda_t} \cdot f_t(y_t, y_{t-1}) + \boldsymbol{\lambda_o} \cdot f_o(y_t, \boldsymbol{x})\right) - \log Z(x) \\ &= \sum_{t=1}^{T} \boldsymbol{\lambda} \cdot f(y_t, y_{t-1}, \boldsymbol{x}) - \log Z(x) \end{aligned} \tag{2. 29}$$

在訓練（Training）的過程中，必須要找到一組參數$\boldsymbol{\lambda}$對於訓練集中的每一筆資料達到條件機率最大化，此目標可以透過高斯牛頓法（Quasi-Newton Method）來達成，也就是對以下的目標函數（Objective Function）$\ell(\boldsymbol{\lambda})$對$\boldsymbol{\lambda}$微分：

$$\ell(\boldsymbol{\lambda}) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k \cdot f_k(y_t^{(i)}, y_{t-1}^{(i)}, \boldsymbol{x}_t^{(i)}) - \sum_{i=1}^{N} Z(\boldsymbol{x}^{(i)}) \tag{2. 30}$$

其中λ_k代表把$\boldsymbol{\lambda}$每一維度分開考慮，$(\boldsymbol{x}_t^{(i)}, y_t^{(i)})$代表了第$i$筆訓練集中的資料。此外，而為了防止過度貼合（Over-fitting）現象產生，在目標函數後加入一正則化項（Regularization Term）：

$$\ell(\boldsymbol{\lambda}) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k \cdot f_k\left(y_t^{(i)}, y_{t-1}^{(i)}, \boldsymbol{x}_t^{(i)}\right) - \sum_{i=1}^{N} Z\left(\boldsymbol{x}^{(i)}\right) - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2} \tag{2. 31}$$

其中σ^2項用來作為比重調整之參數，可以自由調整，此項用來對於數值較大的λ_k

作懲處（Penalty）。接下來，對目標函數微分：

$$\frac{\partial \ell(\boldsymbol{\lambda})}{\partial \lambda_k} = \sum_{i=1}^{N}\sum_{t=1}^{T} f_k(y_t^{(i)}, y_{t-1}^{(i)}, \boldsymbol{x}_t^{(i)}) - \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{y,y'} f_k\left(y, y', \boldsymbol{x}_t^{(i)}\right) p\left(y, y' | \boldsymbol{x}_t^{(i)}\right) - \frac{\lambda_k}{\sigma^2} \tag{2. 32}$$

其中$f_k(y_t^{(i)}, y_{t-1}^{(i)}, \boldsymbol{x}_t^{(i)})$可看作訓練集的特徵向量值，故可直接求得，而$f_k\left(y, y', \boldsymbol{x}_t^{(i)}\right) p\left(y, y' | \boldsymbol{x}_t^{(i)}\right)$則是期望值，故對於所有可能的$y$都要計算。而透過此一微分量，可以套用到高斯牛頓法中，以梯度下降（Gradient Descent）的方式迭代求解。

而測試（Testing）的部份，在權重向量$\boldsymbol{\lambda}$已知的狀況下，目標在於找到一組標記序列可以讓隨機域中的條件機率值最大。假若條件隨機域為直鏈狀，我們可以透過動態規畫（Dynamic Programming）的方法如前向演算法（Forward Algorithm）求得，又或是可以透過近似求法之貪婪演算法(Greedy Algorithm)如維特比演算法（Viterbi Algorithm）[51]求得。

2.4 圖論 (Graph Theory)

透過圖論來計算重要性的方法相當多種，其中最負盛名的是由 Google 早年提出的網頁排名演算法(PageRank)[52]，以及其衍生在非督導式的自動摘要方法中，有使用詞彙特徵排名演算法(LexRank)[53]，是在圖上利用每個句子之間的相似性做分數傳遞(Score Propagation)來抽取重要的句子，本論文中將此種圖論方法以隨機漫步演算法[54]涵蓋之。

2.4.1 隨機漫步 (Random Walk)

在一系統中有許多的節點(Node)，且每個節點都具有各自的分數，隨機漫步演算

法考慮一初始的事前分數(Prior)，對整個系統中每一個節點的分數根據機率分布做分數傳遞。假定系統中有 n 個結果，定義 n 個節點構成的各自之初始機率向量$F^{(0)}$，以及 n 個節點之間的相似度形成的 n 乘 n 矩陣A，經過正規化(Normalized)後可看作為機率傳遞的矩陣。此 n 個節點經過機率傳遞矩陣的作用，可將本身的分數傳遞到其他節點，其傳遞式如下：

$$F^{(t+1)} = (1-\alpha)F^{(0)} + \alpha \cdot A^T \cdot F^{(t)} \tag{2. 33}$$

其中$F^{(t)}$為經過t個迭代後的節點分數向量，A^T代表機率傳遞矩陣A的轉置矩陣(Transpose)，α為一權重分配的參數。此演算法一直重覆進行直到每個節點的分數不再更改，即$F^{(T)} = F^{(T+1)}$之時，此時的節點分數經過彼此相似度傳遞，同時考慮原先事前分數與相似度分數作權重。

此外在此隨機漫步演算法中，必定存有可收斂的節點分數$F^{(T)}$，以下說明之。

$$\begin{aligned} \boldsymbol{F}^{(T)} &= (1-\alpha) \cdot \boldsymbol{F}^{(0)} + \alpha \cdot \boldsymbol{A}^T \cdot \boldsymbol{F}^{(T)} \\ &= [(1-\alpha) \cdot \boldsymbol{F}^{(0)} \cdot \frac{\boldsymbol{e}^T}{n} + \alpha \cdot \boldsymbol{A}^T]\boldsymbol{F}^{(T)} = \boldsymbol{A'}^T\boldsymbol{F}^{(T)} \end{aligned} \tag{2. 34}$$

其中$\boldsymbol{e} = [1,1,\ldots,1]^T$為一都是 1 的 n 維向量。因此此演算法可把$\boldsymbol{F}^{(T)}$看作矩陣$\boldsymbol{A'}$所求得最大特徵值(Eigenvalue)為 1 對應之特徵向量(Eigenvector)。

以此論文中的自動化摘要為例，設定每一個節點即為文件中的一個句子，節點分數的初始值可看作是單一句子的重要性，例如可以用此句子與整篇文件的相似度計算，而傳遞矩陣的分數則是句子之間的相似度，可以透過餘弦相似性(Cosine Similarity)來計算，如果某句子在初始的重要性分數不高，但是它與重要性分數相當高的句子有很大的相似性，則應提高此句子的分數。因此，迭代後的分數可以看做是句子重要性與彼此相似性的結合，前幾高分的句子即可當作抽取式摘要，其後也可把此收斂的句子分數當作最大邊際關聯法(MMR)中的關聯性分數(Relevance Score)，配合重覆性分數提高摘要的效能。

2.5 章節總結

本章對摘要系統與問答系統做簡單的敘述，並且提及幾個在本論文中將會用到的機器學習以及圖論的演算法。在摘要系統的介紹中，先把摘要系統從各個角度分類，並且強調語音文件摘要與純文字摘要之不同，以及在摘要系統中的基準(Baseline)演算法和 ROUGE 的評估方式。在問題系統的介紹中，簡述問答系統實作上大概的分類，並以資料檢索(Information Retrieval)角度的方法做流程的介紹，此外同樣說明語音問答系統和純文字問答系統之不同。最後，由數學的逐步推導，詳細的介紹機器學習領域中的幾個數學模型：支撐向量機、結構式支撐向量機，以及條件隨機域，此外還有以圖論為基礎的隨機漫步演算法。

Part II

語音文件摘要

第三章 利用加入隱藏變數之結構式支撐向量機之語音文件摘要

3.1 簡介

督導式學習的抽取式語音文件摘要大多將問題簡化成二分法問題(Binary Classification Problem)，在過去使用之支撐向量機中，只有單獨考慮單一句子是否重要而該被挑選成摘要，而忽略了摘要整體如重複性(Redundancy)的考慮，或是相連句子的相關性等等。為了將文件整體的資訊一同考慮，本論文使用了結構式支撐向量機(Structured Support Vector Machine)[47-50]，將最大邊際關聯法(Maximum Margin Relevance)[18,19]的重複性考慮加入到目標函數中[55]，同時考慮每一個句子的重要性與重複性，讓重要性與重複性的分數權重參數可以從訓練集中透過整體學習(Joint Learning)得出。此外，在課程錄音的長摘要(30%限制)中，常有連續語句同時被選入摘要的情況，此種狀況可能為語者正在專注講述重要訊息，而在短摘要(10%限制)中，由於能選擇的句數不多，因此在參考摘要中常會出現一連串的句子僅選用一個句子當作代表。而為了將適當的相連句子的特性一同考慮，本論文加入了語句叢集(Cluster)，而由於在參考摘要中並未標記出語句叢集，故被稱做隱藏變數(Hidden Variable)[55]，而本論文提出的方法在於讓結構式向量支撐機自動找出叢集辦法，並且同時選擇出摘要句子。

3.2 督導式模型融入最大邊際關聯法

在過去的研究中，督導式方法大多以統計模型判斷測試資料的好壞，並依據分數高低選取重要句子，然而單純以分數高低選取句子常常會遇到所選出的句子在文字內容上都是非常相似的。為了解決句子重複性的問題，本論文將最大邊際關聯

法引入督導式模型結構式支撐向量機中，在此模型的鑑別函數(Discriminative Function)中加入句子重複性之考量：

$$F(x, y; \boldsymbol{w}) = \sum_{x_i \in y} R(x_i) - \lambda \cdot \sum_{x_i, x_j \in y} Sim(x_i, x_j) \tag{3. 1}$$

其中x_i代表語音辨識轉寫中的一個句子，y代表被選入摘要的句子集合，$R(x_i)$代表句子的重要性，其內容的詳細定義會在 3.5 小節中說明，$Sim(x_i, x_j)$代表句子和句子之間的相似程度，在本論文中使用詞頻和反文件頻(TFIDF)的餘弦相似度(Cosine Similarity)作計算。為了使式子更滿足結構式支撐向量機的形式，我們將(3. 1)中的$R(x_i)$改寫為權重向量和特徵向量之內積的形式：

$$F(x, y; \boldsymbol{w}) = \boldsymbol{w}_0 \cdot \sum_{x_i \in y_i} f_0(x_i) - \lambda \cdot \sum_{x_i, x_j \in y_i} Sim(x_i, x_j) \tag{3. 2}$$

從上述式子，我們可以根據(2. 17)中鑑別函數的形式，重新定義加入句子重複性的權重向量和特徵向量：

$$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{w}_0 \\ \lambda \end{bmatrix}$$

$$\phi(x, y) = \begin{bmatrix} \sum_{x_i \in y_i} f(x_i) \\ -\sum_{x_i, x_j \in y_i} Sim(x_i, x_j) \end{bmatrix} \tag{3. 3}$$

而利用新的權重向量$\boldsymbol{w}'$和特徵向量$\phi'(x, y)$，就可以得到加入句子重複性考量之結構式支撐向量機的鑑別函數。

此外，在前述的鑑別函數中，極有可能在選擇不同標記y時會超過摘要的限制，因此此方法還要滿足摘要系統在選擇字數的限制，在選擇各種可能的標記時，必須要滿足：

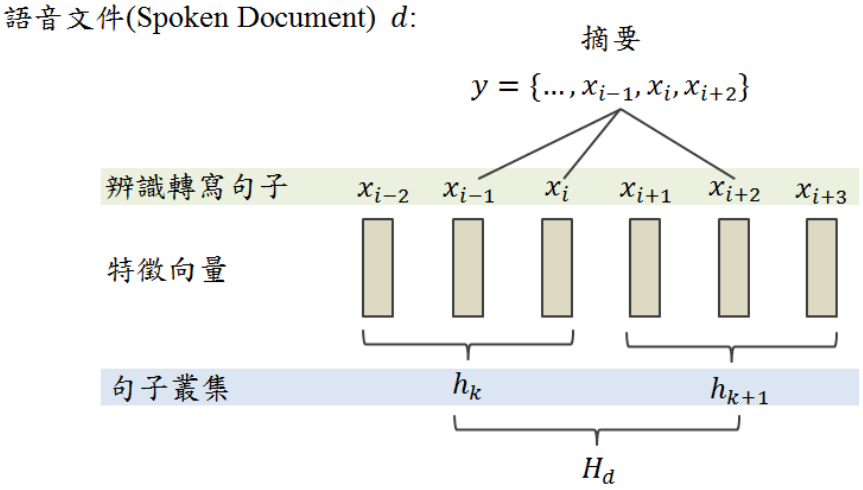

圖 3. 1　本實驗語音文件結構式意圖

$$\sum_{x_i \in y} L(x_i) \leq K \tag{3. 4}$$

其中$L(x_i)$代表句子x_i的長度，本論文句子長度的計算方式是中文字或是英文詞的總數，而K是人為事先定義的摘要總字數限制，本論文中分別使用原文件長度的10%和 30%。

3.3 加入隱藏變數之結構式支撐向量機

此外，由於課程錄音中，大部分的摘要句子有叢集出現的特性，也就是連續或是距離相近的句子常被同時選為摘要，此種現象可以解釋為如果語者想要講解一重要的觀念，則這一個段落都應該會被選入摘要句子，而如果某一段落可能是在用生活化的例子講解課程內容，又或是在講無關課程內容的事，那麼這些段落極有可能會全部被捨棄。為了將此種特性一起加入摘要選句的考量，本論文定義了句子叢集(Cluster)當作輔助，此句子叢集的定義如圖 3. 1 所示是幾個相似的連續句子，在本論文中限定每個叢集大小為 3 至 10 個句子，而由於句子叢集並沒有在訓練集中被標出，所以視為隱藏變數(Hidden Variable)，句子叢集必須由整體學習(Joint Learning)的方式得出。

為了將隱藏變數加入至結構式支撐向量機的鑑別函數中，我們以(3. 1)為基礎加入語句叢集：

$$F(x,y;\boldsymbol{w}) = \sum_{x_i \in y_i} R(x_i) - \lambda \cdot \sum_{x_i,x_j \in y_i} Sim(x_i,x_j) + \sum_{h_k \in H} C(x,y,h_k) + \sum_{h_k \in H} S(h_k) \tag{3.5}$$

其中h_k是句子叢集，H代表整篇文件中所有句子叢集h_k的集合，$R(x_i)$是句子x_i的重要程度，$Sim(x_i,x_j)$是句子之間的相似程度，用來考量摘要的句子重複性，而$C(x,y,h_k)$是代表某一句子叢集h_k在對於整篇文件的代表程度，$S(h_k)$是句子叢集h_k內的相似程度，也就是句子叢集的適當程度。同樣地，此鑑別函數可以化作權重向量與特徵向量之內積。

$$F(x,y;\boldsymbol{w}) = \boldsymbol{w_0} \cdot \sum_{x_i \in y_i} f_0(x_i) - \lambda \cdot \sum_{x_i,x_j \in y_i} Sim(x_i,x_j) + \boldsymbol{w_1} \cdot \sum_{h_k \in H} f_1(x,y,h_k) + \boldsymbol{w_2} \cdot \sum_{h_k \in H} f_2(h_k) \tag{3.6}$$

其中$F_0(x_i)$、$F_1(x,y,h_k)$、$F_2(h_k)$均為特徵向量，將會在 3.5 小節中詳細敘述。再者，此鑑別函數依然符合結構式向量機的定義，整個鑑別函數依然可以看作是權重向量與特徵向量之內積：

$$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{w_0} \\ \lambda \\ \boldsymbol{w_1} \\ \boldsymbol{w_2} \end{bmatrix}$$

$$\phi(x,y,H) = \begin{bmatrix} \sum_{x_i \in y_i} f_0(x_i) \\ -\sum_{x_i,x_j \in y_i} Sim(x_i,x_j) \\ \sum_{h_k \in H} f_1(x,y,h_k) \\ \sum_{h_k \in H} f_2(h_k) \end{bmatrix} \tag{3.7}$$

若是鑑別函數的輸出值越大，便代表此摘要與句子叢集越為理想，反之分數越低代表此摘要和句子叢集不理想。在定義好鑑別函數後，下一個小節要探討如何將其搭配至訓練過程的目標函數中。

3.4 目標函數之定義

根據前一小節的鑑別函數定義(3. 7)，我們同第二章先定義：

$$\begin{aligned} \delta\phi_i(y,H) &= \phi(x_i,y_i,H_i) - \phi(x_i,y,H) \\ subject\ to\ H_i &= arg\max_{H\in\mathcal{H}}\langle w,\delta\phi_i(y,H)\rangle \end{aligned} \tag{3. 8}$$

其中H代表可能的句子叢集方法，H_i代表給定訓練集資料(x_i,y_i)和固定的權重向量$\boldsymbol{w}$的條件下，可以使得鑑別函數分數最大化的句子叢集方法，因此(3. 8)可以看成是使用訓練集資料以及當下最好之句子叢集方法之鑑別函數扣除掉某一標記序列和某一句子叢集方法之鑑別函數。而以下我們微幅調整結構式支撐向量之目標函數：

$$\begin{aligned} &\min_{w}\frac{1}{2}\|\boldsymbol{w}\| + C\sum_{i=1}^{N}\xi_i \\ &subect\ to\ \langle w,\delta\phi_i(y',H')\rangle \geq\ \Delta(y_i,y') - \xi_i,\forall \mathrm{i} \\ &\xi_i \geq 0,\forall \mathrm{i} \end{aligned} \tag{3. 9}$$

其中y'和H'為：

$$y',H' = arg\max_{(y,H)\in\mathcal{Y}\times\mathcal{H}}\Delta(y_i,y) + \langle w,\delta\phi_i(y,H)\rangle \tag{3. 10}$$

其中ξ為第二章說明過的鬆弛變數，$\Delta(y_i,y)$為減損函數(Loss Function)，表達目前產生之標記y和正確答案y_i相差多少，在本論文中定義為：

$$\Delta(y_i, y) = 1 - ROUGE(y_i, y) \tag{3. 11}$$

其中$ROUGE(y_i, y)$為參考摘要與自動摘要之 ROUGE-1 之 F 評估，如果系統產生之自動摘要擁有越高的 ROUGE 分數，代表其和參考摘要所含有的資訊越為相像，因此兩者之間的減損值越小。

然而，理想的狀況下是可以同時產生最佳解的標記序列和句子叢集，但由於計算的複雜度太大導致所需的時間過多，因此本論文使用迭代更新的方式作為替代方案，計算過程中先找到最佳的標記序列，再用此組標記序列找到理想的句子叢集，不斷重複這兩個步驟直到兩者的變化量夠小為止。在找標記序列時，本論文使用貪婪演算法(Greedy Algorithm)，每一次增加一個能使目標函數增加量最大之句子進摘要，直到摘要的限制被滿足；在找句子叢集的過程中，我們使用動態規化演算法(Dynamic Programming)依照句子排列為順序，其中限定每一個句子叢集內的句子數目為 3 到 10 個。而化成(3. 9)式之型式後，其解法就和第二章說明之結構式支撐向量機相同。

在測試(Testing)階段時，利用在訓練階段得到之權重向量$\boldsymbol{w}$，目標在於找到一組標記y^*和句子叢集H^*可以使得鑑別函數最大。

$$y^*, H^* = arg \max_{(y,H)\in\mathcal{Y}\times\mathcal{H}} \langle w, \delta\phi_i(y, H)\rangle \tag{3. 12}$$

同樣地，如果要同時找出標記y^*和句子叢集H^*的最佳解式非常困難的，所以和訓練過程相同，我們將兩者交互迭代求得近似解。

3.5 特徵抽取

此章節為關鍵語句分類器之特徵抽取，對於文件中每個句子抽取其多樣特徵用以訓練分類器。主要的特徵分為六大類，分別為語意特徵(Semantic Features)、相似度特徵(Similarity Features)、韻律特徵(Prosodic Features)、叢集間特徵、叢集內特

徵和其他特徵，將詳細敘述如下。

3.5.1 語意特徵

主題模型是用來表達一文件的主題分佈，在自然語言處理上有相當大的幫助。本論文的實驗中使用了機率式潛藏模型分析模型(Probabilistic Latent Semantic Analysis; PLSA)[56]，並加上潛藏主題亂度(Latent Topic Entropy; LTE)[57]對用語做重要性評估。此模型用於表示一文件的主題分佈，與過去的潛藏語意分析(Latent Semantic Analysis)的目的相當類似，然而，此方法引入了隱藏變數(Latent Variable)，來模擬用語(Term)和文件的共同出現關係，其生成流程定義如下：

1. 根據$p(d_j)$分佈隨機抽樣一個文件
2. 選定文件後，根據$p(z_k, d_j)$抽樣選擇文件表達的主題
3. 選定語意後，根據$p(w_i, z_k)$選擇文件的用語

其中d_j為文件，w_i為用語，z_k為潛藏主題模型。而根據以上流程，經由機率圖型模型中的拆解，可以將某文件d_j產生的用語w_i的共現機率(Joint Probability)寫為：

$$p(w_i, d_j) = p(d_j)p(w_i|d_j) = p(d_j)\sum_{k=1}^{K} p(w_i|z_k)p(z_k|d_j) \qquad (3.13)$$

其中K是人為定義的潛藏主題數目。而透過最大期望值演算法(Maximum Expectation Algorithm)，我們可以藉由持續更新$p(w_i|z_k)$與$p(z_k|d_j)$最大化更新上述共現機率，並將用語和文件的共同出現(Co-occurrence)找出來。

訓練過後之機率式潛藏模型，對文件d_j會有一個以人工定義以K值為其維度的特徵向量$p(z_k|d_j)$，其每一維度為文件d_j屬於某主題的機率。對於每一個主題z_k會有一個以詞典大小為其維度之特徵向量$p(w_i|z_k)$，為該主題中可能出現某個用語的機率。舉例來說，我們可以將z_k當作不同的主題，例如政治、體育、娛樂、

社會等等，若有一篇文件提及「旅外投手王建民」，則該文件之$p(z_k|d_j)$分佈可能會集中於體育主題，在主題體育中$p(w_i|z_k)$可能分佈集中於棒球、籃球、奧林匹克等用語，而非兩岸服務貿易、臺灣獨立等用語。然而事實上我們對每個主題z_k的內容是不知道的，故我們將句子當作文件去作訓練，如此一來每一個句子則會有其主題分佈。本論文中實驗設定主題數量$k = \{16,32,64,128\}$。

此外為了判斷某句子或某文件是否夠具有鑑別度，潛藏主題亂度式引入亂度(Entropy)的概念用於判斷資料分佈的密集程度，在此用於判斷某文件d_j在潛藏主題z上的分佈狀況：

$$E(d_j) = -\sum_{k=1}^{K} P(z_k|d_j) \cdot \log P(z_k|d_j) \tag{3. 14}$$

其值越小代表亂度越低，表示文件集中於某一特定主題，因此，該文件可能含有較多確切的資訊，較為重要。實驗中我們將(3. 14)做為一維之特徵。

3.5.2 相似度特徵

相似度特徵主要呈現每一句子s_i與文件D的相似度，在過去的摘要研究中，相似度的特徵是判斷句子重要性的主要依據。

$$\mathrm{S}(s_j, D) = \frac{1}{|D|} \sum_{s_i \in D} Sim(f(s_i), f(s_j)) \tag{3. 15}$$

其中$f(s_i)$代表某種描述s_i的特徵函數，而$Sim(f(s_i), f(s_j)$是描述$f(s_i)$和$f(s_j)$之間的相似程度，在本論文使用餘弦相似度(Cosine Similarity)計算：

$$Sim\left(f(s_i), f(s_j)\right) = \frac{f(s_i) \cdot f(s_j)}{|f(s_i)| \times |f(s_j)|} \tag{3. 16}$$

此處的$f(s_i)$共有兩種描述方式，第一種為基於用語的特徵(Lexical-based Features)，

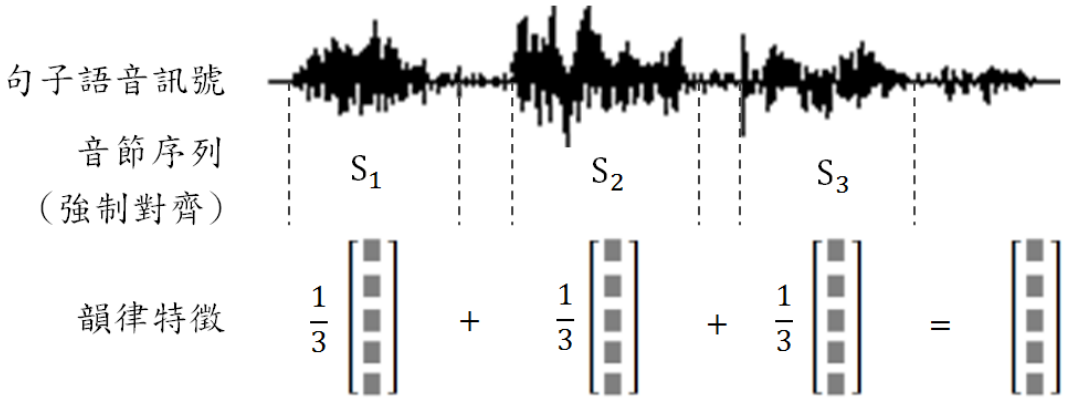

圖 3.2　三音節語句韻律特徵抽取示意圖

該特徵向量的總長數為詞典大小，每一個值為該用語的詞頻(Term Frequency)與反文件頻(Inverse Document Frequency)[58]之相乘；另外一種為前一小結所述之語意特徵，設主題總數 k={16,32,64,128}共四種。若(3. 15)的值越高，則代表句子s_j和整篇文件D越像。

3.5.3 韻律特徵

在過去的研究中[12-17]已證實，韻律特徵可以幫助改善語音文件摘要之表現，本論文中探討單一語者的情況，考慮語者可以透過聲調的抑揚頓挫、音量大小等等表現來強調某重要的概念或句子，故實驗中我們將韻律特徵也一同考慮在內。實驗中我們必須對句子進行強迫性對齊(Forced Alignment)，切成以音節(Syllable)為單位的序列，因過去的研究中韻律特徵須在音節上抽取才有其意義，若小至音素(Phone)大至詞(Word)便會失去其效果，最後我們將每個音節的韻律特徵平均作為該句的韻律特徵，如為三個音節語句之韻律特徵示意圖。而以下將韻律特徵細分為音高、音長、能量和停頓長度等四種[55]。

- 音長特徵(Duration Features)

  因為每個音節的音長本來就不同，或是可能由於語者本身習慣或是發音方式不同而讓其平均發聲長度不一，故本論文對每個音節抽取其音節長度、音節長度對語料中所有該音節之長度正規化、音節與前一音節平均、

與前幾個音節的音長標準差，總共 27 維，細項見表 3. 1。

- 音高特徵(Pitch Features)

  對於每一個音節，本論文抽取了該音節的第一個及最後一個音框(Frame)的音高並對所有的音框做正規化(Normalization)、頭兩個和最後兩個音框的斜率、最高音高、最低音高、平均音高以及其他音高軌跡(Pitch Contour)相關之特徵共 20 維，細項見表 3. 2。

- 能量特徵(Energy Features)

  能量會隨著時間而變化，基於不同的音框和能量軌跡(Energy Contour)，抽出最大、最小、平均能量、能量軌跡斜率、第一個和最後一個音框的能量值，共 13 維的能量特徵，細項見表 3. 3。

- 停頓特徵(Pause Features)

  在自發性(Spontaneous)的語料的句子中，句內的停頓可能代表了語句的不流暢，本論文抽取了停頓長度，以及停頓與停頓前後之音節關係，共 12 維停頓特徵，細項見表 3. 4。

3.5.4 叢集與摘要相關特徵

此類特徵主要是描述句子叢集與句子摘要標記之關係，為(3. 7)中提到的 $f_1(x, y, h_k)$特徵向量。本論文為了描述某些連續或是距離相近的句子常會同時被選作摘要，又或是錄音轉寫中會有一大段直接被捨棄不當作摘要，因此定義以下兩種特徵：

- 叢集內句子之標記純度

  此特徵描述句子叢集中被選入摘要的比例，總共有三維，分別是句子叢集中被選入摘要的句子比例、句子叢集中不被選入摘要的句子比例，以及取前兩種特徵之最大值，如(3. 17)。若值越大，則表示此句子叢集內的句子之摘要分佈越一致。

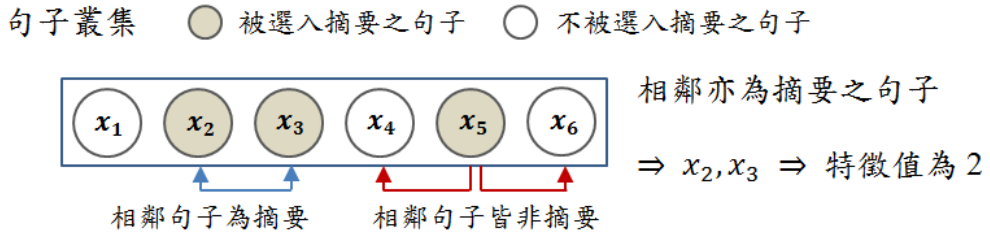

圖 3.3　叢集內句子之標記連續性

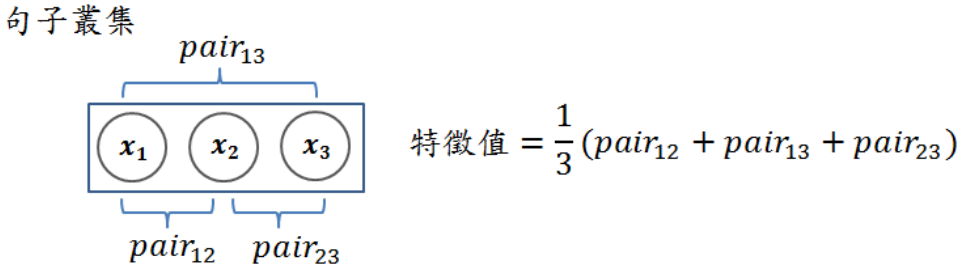

圖 3.4　叢集內句子相似度

$$\max\left(\frac{\text{被選入摘要之句子數}}{\text{句子數}}, \frac{\text{不選入摘要之句子數}}{\text{句子數}}\right) \tag{3.17}$$

- 叢集內句子之標記連續性

  在同一句子叢集內，有多少個被選入摘要的句子，其相鄰的句子也同時被選入在摘要中，如圖 3.3。

3.5.5 叢集內特徵

此類特徵描述句子叢集是否恰當，以下定義兩種特徵：

- 句子叢集內的相似度

  句子叢集內每個句子的兩兩之間的相似度，如圖 3.4，此相似度的計算方式是使用機率式潛藏模型分析模型(PLSA)之餘弦相似度計算。若該值越大，則此叢集內的相似度越高，越可能為適當之句子叢集。

- 句子叢集與整篇文件之相似度

  叢集內的所有句子與整篇文件的相似度，此相似度的計算方式是使用機

率式潛藏模型分析模型(PLSA)之餘弦相似度計算。如果此值越大，該叢集可能含有越重要之意義。

3.5.6 其他特徵

本實驗中抽取了額外的三個特徵，條列於下：

- 語句音節(Syllable)總數：過去在某些語料研究中，句子的長短對於判斷句子重要性相當有效。
- 該語句在文件中的位置：舉例來說，一篇文件中若有 20 個句子，其中第三句的特徵即為$\frac{3}{20} = 0.15$，在課程語料中，語者通常會先對講解的內容做簡介，課程中間也許式數學的推導或是更入的介紹，而最後是對整篇文件總結，故此特徵能幫助判斷該句子可能為何種類型。
- 語句的重要性：此特徵為語句中每一個詞的反文件頻(IDF)相加，越大可能代表該語句中含有較多的關鍵用語。

3.6 實驗基礎設置

3.6.1 實驗語料與辨識

實驗我們使用國立台灣大學由李琳山教授開設之《數位語音處理概論》課程錄音，該課程為單一語者，且課程錄音內容為中英混雜（Code-mixing），主位語言為中文，並夾雜著客位語言英文的詞或片語，通常是專業術語。此課程共有 45 堂，每一堂約一小時，總共語料的長度約為 45.2 小時，共分成 193 個文件，每則文件約 17.5 分鐘。實驗中我們使用辨識轉寫（ASR Transcriptions）並加以斷句（Sentence Segmentation）和斷詞(Word Segmentation)。辨識方面切出其中 12 小時用於訓練聲學模型（Acoustic Model）以及語言模型（Language Model），剩下的 33 小時用於測試，用於摘要之辨識轉寫正確率約為 88%。

1	音節長度
2	音節長度，對句子所有音節之平均長度正規化
3	與前一音節之平均長度
4	與前一音節之平均長度，對句子所有音節之平均長度正規化
5	與前一音節之長度比值
6	下一音節與前二音節之長度比值
7	下二音節與前三音節之長度比值
8	中斷點(語句邊際)前後音節長度比值
9	此音節長度語前兩個音節的音節長度之標準差
10	聲母長度
11	聲母長度，對句子所有音節之平均長度正規化
12	韻母長度
13	韻母長度，對句子所有音節之平均長度正規化
14	尾音長度
15	尾音長度，對句子所有音節之平均長度正規化
16	中斷點前後(最近的停頓點)平均音節長度比值
17	前一音節長度，除以自前一停頓處
18	停頓長度語後一音節長度之乘積
19	停頓長度語前一音節長度之乘積
20	停頓長度語後一音節長度之比值
21	停頓長度語前一音節長度之比值
22	特徵 18 之值，與前兩音節特徵 18 之值的標準差
23	特徵 19 之值，與前兩音節特徵 19 之值的標準差
24	特徵 20 之值，與前兩音節特徵 20 之值的標準差
25	特徵 21 之值，與前兩音節特徵 21 之值的標準差
26	從前一停頓至此音框數量
27	從前一停頓至此音節數量

表 3.1　音長韻律特徵列表

1	聲母第一個音框之基頻，對句子平均基頻正規化
2	聲母最後一個音框之基頻，對句子平均基頻正規化
3	聲母前兩音框基頻差
4	聲母後兩音框基頻差
5	最小基頻，對句子平均基頻正規化
6	最小基頻，對聲母平均基頻正規化
7	最大基頻，對句子平均基頻正規化
8	最大基頻，對聲母平均基頻正規化
9	聲母平均基頻，對句子平均基頻正規化
10	聲母基頻軌跡之斜率
11	基頻音高軌跡之斜率取絕對平均值
12	音高軌跡前段之基頻平均
13	音高軌跡中段之基頻平均
14	音高軌跡後段之基頻平均
15	音高軌跡前段之基頻平均，減去音節平均音高
16	音高軌跡中段之基頻平均，減去音節平均音高
17	音高軌跡後段之基頻平均，減去音節平均音高
18	音高軌跡前段之基頻音高軌跡斜率
19	音高軌跡中段之基頻音高軌跡斜率
20	音高軌跡後段之基頻音高軌跡斜率

表 3.2　音高韻律特徵列表

1	第一個音框之能量，對句子最大音框作正規化
2	最後一個音框之能量，對句子最大音框作正規化
3	字元起始時能量軌跡之斜率
4	字元結束時能量軌跡之斜率
5	能量最小值
6	能量最大值
7	音框平均能量，對句子最大音框作正規化
8	能量軌跡之斜率取絕對平均值
9	中間音框平均能量，對句子最大音框作正規化
10	忠間音框能量之和，對句子最大音框作正規化
11	前段音框能量平均
12	中段音框能量平均
13	後段音框能量平均

表 3.3　能量韻律特徵列表

1	停頓前後之音節長度之比例
2	從前一停頓開始計算之音框數量除以從前一停頓開始計算之音節數量
3	停頓持續時間與後一音節持續時間之乘積
4	停頓持續時間與前一音節持續時間之乘機
5	停頓持續時間與後一音節持續時間之比例
6	停頓持續時間與前一音節持續時間之比例
7	音節與後二音節之停頓持續時間與後一音節持續時間之乘積的標準差
8	音節與後二音節之停頓持續時間與前一音節持續時間之乘積的標準差
9	音節與後二音節之停頓持續時間與後一音節持續時間之比例的標準差
10	音節與後二音節之停頓持續時間與前一音節持續時間之比例的標準差
11	從前一停頓開始計算之音框數量
12	從前一停頓開始計算之音節數量

表 3.4　停頓韻律特徵列表

3.6.2 參考摘要之形成

我們取出其中 40 篇文件用於此摘要抽取實驗，每一篇文件我們請了三位修過該課程之學生做摘要標記，學生被要求聽完該文件之課程內容，並由辨識文本中選出屬於摘要的重要語句，標記的摘要依照長度限制不同又分為兩種版本，分別為短篇（Short）與長篇（Long）摘要，即摘要的限制分別為不可超過文件字數之 10%以及 30%。

3.6.3 實驗配置

此實驗中，為了讓評估更為客觀，我們使用四重交叉驗證(4-Fold Cross Validation)，將 40 篇文件分成四等份，每份各有 10 篇文件，在每一次的實驗中其中兩份共 20 篇文件做為訓練資料，用以訓練督導式模型，一份當做發展資料，用於督導式模型的參數調整，另外一份則是用為測試資料。而每一篇自動生成摘要會依照長篇或短篇選擇該文件的三篇人工參考摘要做比對，以評估效能。

		非督導式方法		督導式方法			
限制	評估	(a) 最長句子	(b) 最大邊際關聯法	(c) 支撐向量機	(d) 結構式支撐向量機	(e) 提出模型(*)	(f) 提出模型
10%	ROUGE-1	0.3815	0.3966	0.4117	0.4315	0.4363	0.4406
	ROUGE-2	0.1778	0.1777	0.1761	0.2162	0.2329	0.2208
	ROUGE-L	0.3754	0.3983	0.4057	0.4229	0.4285	0.4333
30%	ROUGE-1	0.5020	0.5484	0.5372	0.5624	0.5628	0.5657
	ROUGE-2	0.3373	0.3380	0.3354	0.3500	0.3688	0.3627
	ROUGE-L	0.4998	0.5445	0.5335	0.5577	0.5591	0.5616

表 3.5　本論文提出之含隱藏變數之結構式支撐向量機與其它方法之比較
(*為不含叢集內句子之標記連續性之考慮)

3.6.4 評估方式

選用第二章所提及的 ROUGE 的 F 評估（F-measure）做為評估標準，實驗中使用 ROUGE-1、ROUGE-2 做為單連文法（Unigram）和雙聯文法（Bigram）之評估，此外還有 ROUGE-L 做為最長子字串之評估。

3.7 實驗結果與分析

本章結將詳述實驗結果並討論分析。首先介紹實驗中的基準實驗(Baseline)，之後實驗內容將本論文提出之含隱藏變數之結構式向量支撐機分別和三種重要的摘要方法做比較——非督導式的最大邊際法、督導式之二分類法支撐向量機以及結構式支撐向量機。

在文件摘要上，一般會使用一些簡單的方法做為基準，例如在課程錄音上常用之長句優先選取法，或是在新聞錄音中常用之開頭具優先選取法。在課程錄音等自發性(Spontaneous)語料上，由於語者錄音時並無講稿，因此大部分的短句通常帶有語句不流順，或是不含有重要意義，應當不被選入摘要中，反之長句通常含有較多的語義，因此拿長句當作摘要通常有非常好的效果。此基準實驗的結果

可以參照表 3. 5 之(a)欄。

表 3. 5 的(b)欄為第二章所介紹過的最大邊際關聯法(Maximum Margin Relevance)，可以看到此方法明顯比抽取最長句的表現有明顯提升，甚至在 30% 的實驗中，其表現勝過於(c)欄之督導式支撐向量機。推測其原因在於二分類法之支撐向量機只考慮句子的重要性，且在 30%限制時所需選取的句子可能含有較嚴重之句子重複性，因此表現會輸給考量句子重複性之最大邊際關聯法。

此實驗進一步和結構式支撐向量機做比較，也就是本論文方法去除句子叢集做為隱藏變數，只考慮重要性與摘要重複性，其實驗結果見表 3. 5 (d)，可看出此方法和(b)、(c)方法不管在長摘要還是短摘要都有明顯之進步，可推論全域考慮摘要的重複性在摘要抽取中扮演很重要的角色。表 3. 5 (e)與(f)欄是本論文提出之方法，(e)欄為不含 3.5.4 小節之叢集內句子之標記純度特徵，(f)欄是使用所有本論文提出之特徵，實驗結果顯示此方法在(e)與(f)皆比起結構式支撐向量機又有進步，其中(f)欄的 ROUGE-1 與 ROUGE-L 在長摘要與短摘要中皆比(e)欄有進步，而(f)欄在 ROUGE-2 的評估上都比(e)欄稍微退步，推測是因為本實驗使用之減損函數(Loss Function)是與產生出來的摘要標記與訓練集中的人工標記之 ROUGE-1 分數相關，因此在只針對加強 ROUGE-1 的結果下，ROUGE-2 可能因為選取到的句子之雙連文法(Bi-gram)沒有那麼理想，因而表現下降，至於 ROUGE-L 因為只要求句中的順序而不需要連續出現，因此表現並無下降。

3.8 章節總結

此章節中我們提出含有隱藏變數之結構式支撐向量機於抽取式之自動摘要，同時在目標函數中加入重複性和語句叢集之考慮，除了自動化的學習與重複性的比重外，還使此督導式模型自動學習出如何做出連續語句叢集，並且考慮每個叢集的特性選出摘要句。此演算法用於中英文混雜的單一語者語料中，與過去的非督導式或是督導式模型皆有進步。

第四章 使用雙層隨機漫步配合課程投影片之語音文件摘要

4.1 簡介

在一般的課程授課中，講者會以投影片輔助課程的進行，此課程投影片即可以當作是課程錄音之摘要。為了同時考慮課程錄音語句之重要性以及投影片內容所造成之影響，本論文提出雙層隨機漫步(Two-layer Random Walk)[59,60]演算法，使得排序分數(Ranking Score)可以透過課程錄音轉寫句子之間的相似度、錄音轉寫句子與投影片之相似度，以及投影片內句子之間的相似度等機率分佈來進行分數的傳遞，最後排序分數中高分的句子不僅是在課程錄音轉寫中具有重要性，同時也和投影片內容有高相似性，此種句子應該為理想之摘要。實驗結果顯示，配合全英文的課程投影片，在中英混雜(Code-mixing)的課程錄音摘要中，使用配合投影片之雙層隨機漫步演算法之摘要，擁有比單純使用課程錄音之單層隨機漫步演算法更加的表現。

4.2 雙層隨機漫步

在過去的摘要研究中，有一部份的研究[10]使用隨機漫步(Random Walk)演算法，然而此演算法只能考慮錄音轉寫或是純文字的句子與句子的相似性，為了加入投影片的資訊，本論文使用了雙層隨機漫步。此演算法的優點在於，可以將兩種相關異質資料的資訊連結在一起，不僅只是考慮單一資料內的重要性，也考慮從另一種資料得到之重要性，以下將介紹本論文中如何定義兩種異質資料與雙層隨機漫步演算法如何應用在抽取式語音摘要上。

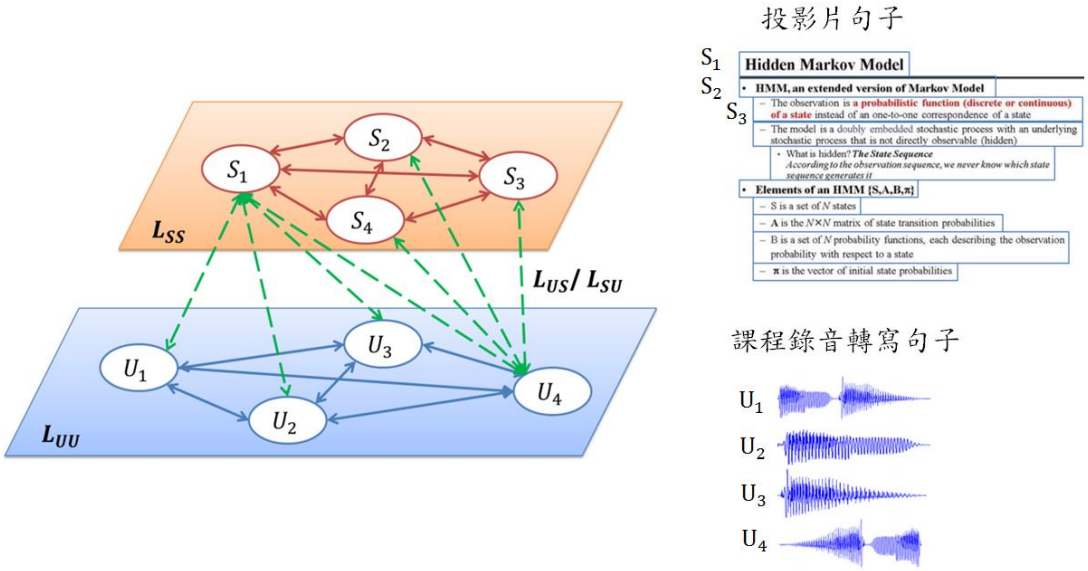

圖 4. 1　本論文中使用之雙層資訊示意圖：錄音轉寫句子以及課程投影片

4.2.1 雙層定義—課程語音辨識轉寫與投影片

本論文定義雙層資訊，分別為課程錄音辨識轉寫層與投影片層，如圖 4. 1 上半部是投影片層，而下半部是錄音辨識轉寫層。在錄音辨識轉寫層中，先利用語音信號檢測(Voice Activity Detection)將語音文件作斷句，而每一句句子在此層中皆有一個對應的節點以及分數U_i，總和所有的節點可形成一分數向量F_U，如果此句子對應之分數越高，則代表此句子越有可能被選作摘要句子，而每一個句子(節點)與句子(節點)之間存在著相似性的分數，這些相似度可以用相似度矩陣L_{UU}的方式表達，如果某一句子和愈多重要的句子的相似性愈高，則此句子的分數也應該被提得愈高，反之則應減少；在投影片層中，投影片的每一行當作一個句子S_i，以分數向量F_S的形式表達，同樣地每一行之間也有相似性的分數，同樣也以相似度矩陣L_{SS}表達。而在這兩層之間也有每一個轉寫句子對應到每一行投影片句子之相似度L_{US}與L_{SU}，讓兩層資料可以互相傳遞。

4.2.2 相似度分數

對於每一個錄音轉寫句子或是投影片句子的相似度矩陣之相似性分數，在本論文的實驗中皆使用餘弦相似度(Cosine Similarity)計算：

$$\mathrm{Sim}(x_i, x_j) = \frac{f(x_i) \cdot f(x_j)}{|f(x_i)| \times |f(x_j)|} \tag{4. 1}$$

其中$f(x_i)$代表對x_i進行特徵抽取或轉換，以下將介紹兩種在實驗中使用的特徵：

- OKAPI/BM25：

  OKAPI/BM25[61]可以視為一種基於詞頻反文件頻(TF-IDF)之權重計算(Term Weighting)的方式，對於每一個詞皆可以算出對應之分數：

$$Score(q_i, D) = IDF(q_i) \times \frac{freq(q_i, D) \cdot (k_1 + 1)}{freq(q_i, D) \cdot k_1(1 - b + b\frac{|D|}{avgl})} \tag{4. 2}$$

  其中D是某一篇文件，$IDF(q_i)$為修改過的反文件頻，$freq(q_i, D)$是某個詞q_i在文件D中的出現次數，$|D|$是文件D的總長度，$avgl$是語料中每篇文件的平均長度，k_1和b是參數，本論文中設定$k_1 = 1.5$、$\mathrm{b} = 0.75$。以下是$\mathrm{IDF}(q_i)$的定義：

$$IDF(q_i) = \log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{4. 3}$$

  其中N代表語料中的文件總數，$n(q_i)$代表某詞q_i總共在幾篇文件中出現，0.5 之常數項是為了防止$n(q_i)$為零時無法取log的狀況。

  而當擁有了每一個詞的對應分數，對每一個句子而言，句子的特徵向量便是一個維度是詞典(Lexicon)大小的向量，只要對應到某個詞便將該維的分數改為 OKAPI/BM25 的分數，否則為 0。

- 機率式潛藏模型分析模型(PLSA)：

  機率式潛藏模型分析模型之詳細描述可參考 3.5.1 小節的說明。每一個句子皆有主題分佈的向量，在此論文的實驗中使用主題數量$k = 32$。

4.2.3 重要性之層內傳遞

層內的的分數傳遞與單層的隨機漫步相同，某一個句子的重要性分數，除了本身的分數外，會由其他相關的句子的重要性傳遞而來：

$$\begin{cases} F_U^{(t+1)} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UU} \cdot F_U^{(t)} \\ F_S^{(t+1)} = (1-\alpha)F_S^{(0)} + \alpha \cdot L_{SS} \cdot F_S^{(t)} \end{cases} \quad (4.4)$$

其中$F_U^{(t)}$代表第t個迭代時辨識轉寫層的分數向量、$F_S^{(t)}$代表第t個迭代時投影片層的分數向量、L_{UU}和L_{SS}分別代表辨識轉寫層內的相似度矩陣以及投影片層的相似度矩陣，而α則是一參數來調整相似度矩陣的貢獻權重。

4.2.4 重要性之層間傳遞

與單層隨機漫步不同，辨識轉寫和投影片資訊必要要能夠互相影響，因此我們必須要調整(4. 4)式：

$$\begin{cases} F_U^{(t+1)} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{US} \cdot F_S^{(t)} \\ F_S^{(t+1)} = (1-\alpha)F_S^{(0)} + \alpha \cdot L_{SU} \cdot F_U^{(t)} \end{cases} \quad (4.5)$$

可以從上述式子看到矩陣傳遞的部份改變為從另外一層傳遞過來，舉例來說，辨識轉寫層的重要性，$\alpha \cdot L_{US} \cdot F_S^{(t)}$的部份是從投影片層分數透過層間的相似度矩陣$L_{US}$傳遞而來；而投影片層的分數，$\alpha \cdot L_{SU} \cdot F_U^{(t)}$部份是從辨識轉寫層透過層間相似度矩陣$L_{SU}$傳遞而來。

4.2.5 結合層內與層間之傳遞

然而4.2.4小節只考慮層間分數傳遞的做法並沒由考慮到4.2.3小節所述的層內分術傳遞，在理想上的雙層隨機漫步，應該是同時考慮兩層間與層內的分數傳遞，因此我們重新改寫(4. 5)：

$$\begin{cases} F_U^{(t+1)} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UU}^T \cdot L_{US} \cdot F_S^{(t)} \\ F_S^{(t+1)} = (1-\alpha)F_S^{(0)} + \alpha \cdot L_{SS}^T \cdot L_{SU} \cdot F_U^{(t)} \end{cases} \tag{4. 6}$$

從上式可以看出在錄音辨識轉寫層的分數F_U的更新方法，先經由L_{SU}矩陣傳遞投影片層的分數到錄音辨識轉寫層，再經由L_{UU}矩陣再做錄音辨識轉寫層內的分數傳遞；在投影片層的分數F_S，會經由L_{US}矩陣先傳遞錄音辨識轉寫層的分數到投影片層，再經由L_{SS}矩陣再做投影片層內的分數傳遞。

而透過(4. 6)不斷更新兩層的分數後，會達到兩層數值的收斂：

$$\begin{cases} F_U^{*} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UU}^T \cdot L_{US} \cdot F_S^{*} \\ F_S^{*} = (1-\alpha)F_S^{(0)} + \alpha \cdot L_{SS}^T \cdot L_{SU} \cdot F_U^{*} \end{cases} \tag{4. 7}$$

此方程式$F_U^{*} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UU}^T \cdot L_{US} \cdot F_S^{*}$可以改寫成：

$$\begin{aligned} F_U^{*} &= (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UU}^T \cdot L_{US} \cdot F_S^{*} \\ &= (1-\alpha)F_U^{(0)} \cdot \frac{\boldsymbol{e}^T}{n} \cdot F_U^{*} + \alpha \cdot L_{UU}^T \cdot L_{US} \cdot F_S^{*} \\ &= (1-\alpha)F_U^{(0)} \cdot \frac{\boldsymbol{e}^T}{n} \cdot F_U^{*} + \alpha \cdot L_{UU}^T \cdot L_{US} \cdot ((1-\alpha)F_S^{(0)} + \alpha \cdot L_{SS}^T \cdot L_{SU} \cdot F_U^{*}) \\ &= (1-\alpha)F_U^{(0)} \frac{\boldsymbol{e}^T}{n} F_U^{*} + \alpha \cdot L_{UU}^T L_{US} \cdot ((1-\alpha)F_S^{(0)} \frac{\boldsymbol{e}^T}{n} F_U^{*} + \alpha \cdot L_{SS}^T L_{SU} F_U^{*}) \\ &= [(1-\alpha)F_U^{(0)} \frac{\boldsymbol{e}^T}{n} \cdot + \alpha \cdot L_{UU}^T L_{US} \cdot ((1-\alpha)F_S^{(0)} \frac{\boldsymbol{e}^T}{n} + \alpha \cdot L_{SS}^T L_{SU}]F_U^{*} \\ &= \mathrm{M} \cdot F_U^{*} \end{aligned} \tag{4. 8}$$

從(4. 8)可以看出雙層隨機漫步的更新式事實上可以化簡成特徵向量的形式，而F_U^*的解也就是M矩陣對應到特徵值(Eigen Value)為 1 時的特徵向量。同樣地，F_S^*的解也可透過同樣的形式求得特徵向量。

當分數傳遞趨於穩定時，本實驗利用收斂的F_U^*分數做為每一個辨識轉寫句子的分數，再依照分數高低依序選取句子當作摘要，直到選取作為摘要的句子到達字數限制為止。

4.3 實驗基礎設置

4.3.1 實驗語料與辨識

本實驗用的課程錄音和第三章的實驗相同，即是由李琳山教授開設之《數位語音處理概論》課程錄音，該課程為單一語者，且課程錄音內容為中英混雜（Code-mixing），主位語言(Host Language)為中文，並夾雜著客位語言(Guest Language)英文的詞或片語，通常是專業術語。此課程共有 45 堂，每一堂約一小時，總共語料的長度約為 45.2 小時，共分成 193 個文件，每則文件約 17.5 分鐘。實驗中我們使用辨識轉寫（ASR Transcriptions）並加以斷句（Sentence Segmentation）和斷詞(Word Segmentation)。辨識方面切出其中 12 小時用於訓練聲學模型用於訓練聲學模型（Acoustic Model）以及語言模型（Language Model），剩下的 33 小時用於測試，用於摘要之辨識轉寫正確率約為 88%。

此外本實驗還使用課程投影片用來輔助產生摘要，總共有 193 張投影片，內容完全以英文表達，且上述每一篇語音文件的辨識轉寫都分別對應到一張課程投影片，其內容由人工轉為純文字形式，每一行的斷句和投影片中表示的相同。

4.3.2 參考摘要之形成

我們取出其中 40 篇文件用於此摘要抽取實驗，每一篇文件我們請了三位修過該

課程之學生做摘要標記，學生被要求聽完該文件之課程內容，並由辨識文本中選出屬於摘要的重要語句，在標記的過程中學生不看課程投影片內容，標記的摘要依照長度限制不同又分為兩種版本，分別為短篇（Short）與長篇（Long）摘要，即摘要的限制分別為不可超過文件字數之 10%以及 30%。

4.3.3 實驗配置

此實驗中，為了讓評估更為客觀，我們使用四重交叉驗證（4-Fold Cross Validation），將 40 篇文件分成四等份，每份各有 10 篇文件，在每一次的實驗中其中一份共 10 篇文件做為發展資料，用於雙層隨機漫步模型的參數調整，另外三份則是用為測試資料。而每一篇自動生成摘要會依照長篇或短篇選擇該文件的三篇人工參考摘要做比對，以評估效能。

4.3.4 評估方式

選用第二章所提及的 ROUGE 的 F 評估（F-measure）做為評估標準，實驗中使用 ROUGE-1、ROUGE-2、ROUGE-3 進行單連文法（Unigram）、雙聯文法（Bigram）和三聯文法（Trigram）之評估，此外還有 ROUGE-L 做為最長子字串之評估。

4.4 實驗結果與分析

本章結將詳述實驗結果並討論分析。本實驗將雙層隨機漫步與基準實驗(Baseline)與最大邊際關聯法(Maximum Marginal Relevance)、單層隨機漫步(Random Walk)做比較，實驗結果見表 4. 1。於前一章節相同，我們使用的基準實驗表 4. 1(a)欄為最長句子抽取方法，在課程錄音中屬於有效且和合理的抽取方式；表 4. 1(b)欄為最大邊際關聯法，其實驗結果和前一章節稍有不同，其原因在於此章節我們對於每一個句子僅使用 OKAPI/BM25 做為特徵向量，其最大邊際法中的重要性也用此特徵和整篇文件做相似性計算得出。

限制	評估	(a) 最長句子	(b) 最大邊際關聯法	(c) 隨機漫步	(d) 雙層 隨機漫步	(e) 雙層 隨機漫步
					OKAIPI/BM25	OKAIPI/BM25 +PLSA
10 %	ROUGE-1	0.3815	0.4005	0.4102	0.4216	0.4223
	ROUGE-2	0.1778	0.1757	0.1992	0.2031	0.2080
	ROUGE-3	0.1369	0.1312	0.1553	0.1588	0.1637
	ROUGE-L	0.3754	0.3938	0.4033	0.4128	0.4139
30 %	ROUGE-1	0.5020	0.5354	0.5316	0.5503	0.5510
	ROUGE-2	0.3373	0.3203	0.3408	0.3548	0.3569
	ROUGE-3	0.2821	0.2635	0.2889	0.2994	0.3015
	ROUGE-L	0.4998	0.5320	0.5285	0.5468	0.5476

表 4.1　雙層隨機漫步實驗結果與其他方法比較表

限制	評估	(a) 雙層隨機漫步	(b) 雙層隨機漫步 +最大邊際關聯法
10%	ROUGE-1	0.4216	0.4257
	ROUGE-2	0.2031	0.2014
	ROUGE-3	0.1588	0.1540
	ROUGE-L	0.4128	0.4177
30%	ROUGE-1	0.5503	0.5463
	ROUGE-2	0.3548	0.3417
	ROUGE-3	0.2994	0.2839
	ROUGE-L	0.5468	0.5424

表 4.2　雙層隨機漫步加上最大邊際關聯法

表 4. 1(c)中的結果為使用單層隨機漫步，可以看出此方法明顯比前兩種的基準實驗以及最大邊際關聯法有許多進步，由其在 ROUGE-2 和 ROUGE-3 的部份。而表 4. 1(d)欄和(e)欄為本論文提出的雙層隨機漫步，(d)欄僅使用 OKAPI/BM25 做為特徵向量，(e)欄為同時使用 OKAPI/BM25 和機率式潛藏模型分析模型(PLSA)做為特徵向量。實驗結果可以看出使用雙層隨機漫步比起單層隨機漫步又有更高的表現，顯示多使用投影片資訊做為輔助可以大幅加強抽取式摘要的表現；而(e)欄稍為比(d)欄有些許進步，可以推論是多加了主題模型所得到更精細的相似度分數，因而表現得以上升。

另外表 4. 2 說明雙層隨機漫步串接最大邊際關聯法的結果，(a)欄結果即為表 4. 1 (d)欄的結果，(b)欄為使用雙層隨機漫步的排序分數當重要分數之最大邊際關聯法，可視為兩種方法的串接。實驗結果在短摘要(10%)時有進步，而在長摘要(30%)時退步，推論可能原因在於短摘要(10%)不適合有句子重複性出現，故使用最大邊際關聯法過濾掉太多相似的句子可以得到幫助。而長摘要(30%)由於句子重複性的限制較低，又因為雙層隨機漫步已經利用投影片資訊盡量將多種主題或資訊包含進摘要，故對於重複性的考慮反而會使得原本重要性分數被打亂，因而造成表現下降。

4.5 章節總結

此章節提出非督導式模型之雙層隨機漫步應用於抽取式摘要，針對於課程錄音語料之特性，語者在課程的行進是搭配課程投影片，所以課程投影片的內容將會是潛藏的摘要，本論文結合投影片之間的重要性傳遞、課程之辨識轉寫之間的重要性傳遞，以及雙層互相傳遞來重新評估某一辨識轉寫句子之重要性。在中英語混雜之課程錄音搭配全英文之投影片實驗中，可以看出與單獨使用課程句子之間的相似性與重要性之單層隨機漫步有明顯的進步。

Part III

中文問答系統

第五章 以序列標號進行查詢指令生成

5.1 簡介

在以下的第五至七章中，本論文將會針對問答系統進行研究與討論。本論文的問答系統屬於第二章曾介紹過的模擬陳述問答(Factoid Question Answering)[23-25]，且使用資訊檢索(Information Retrieval-based)[28-31]為基礎之系統架構，故我們將幾個資訊檢索的搜尋架構拆解，第五章將會介紹如何用問句生成查詢指令，以利搜尋引擎可以利用這個查詢指令，回傳許多相關的網頁文件，而第六章說明如何將回傳網頁重新排序，第七章介紹如何從回傳的網頁中尋找答案。

最基礎的查詢指令生成只由問句中選擇重要的詞串接起來當作查詢指令，不過在許多詞義表達上單詞並不是最好的表達方法，而短語(Phrase)能更精準的表達，一般的書名、機構名就有此種特性，例如「世界衛生組織」、「戰爭與和平」。此外，若是透過單詞去做查詢指令，部份的詞若是沒有被標記成為查詢指令，也會影響到查詢指令的語意表達，例如書名「文學之旅」若是以單詞進行標記，由於「之」大多被看成少意詞(Stop Words)，因此很有可能會被標記成為「文學 旅」當成查詢指令，而此結果將會包含如「文學旅物」、「參旅文學獎」等網頁被當作是相關的網頁回傳。此外，大多數的人名在斷詞系統中並不能完整被斷出，常發生人名被斷開的狀況，如埃及豔后「克莉奧佩托拉」在斷詞中會被標成「克莉奧佩」、「托」、「拉」，若是使用較高階層的短語，就有機會能夠將整個完整的人名同時考慮。

本論文使用樹狀結構之條件隨機域(Tree-structured Conditional Random Fields; Tree-structured CRF)[62-66]結合剖析樹(Parse tree)[67,68]，透過剖析樹的句子拆解，使得短語也能夠被考慮至查詢指令生成中。實驗結果也顯示樹狀條件隨機域做的查詢指令比起單純使用鏈狀條件隨機域的表現更佳，能夠使越多的答案

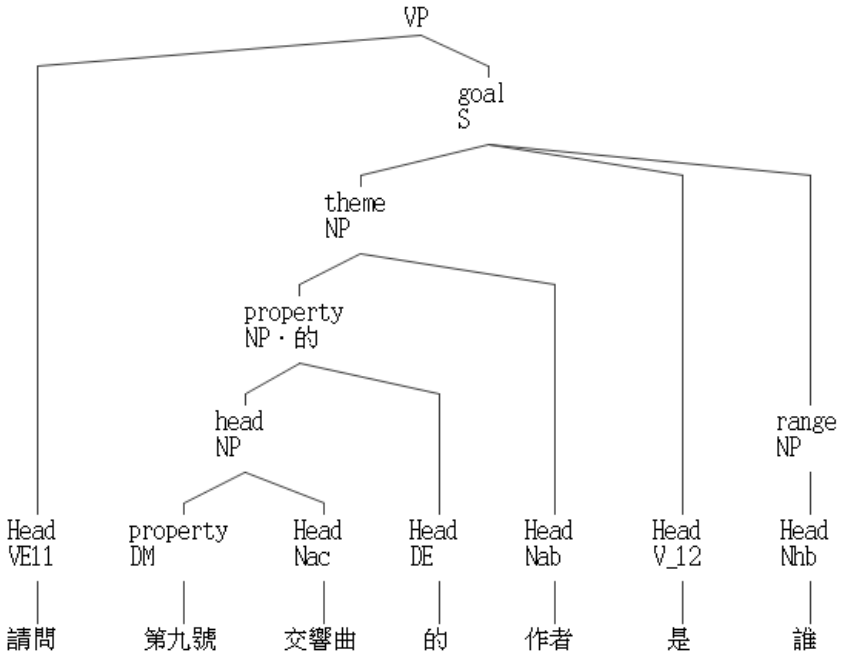

圖 5. 1　剖析樹(Parse tree)圖

能在搜尋引擎回傳的網頁中。

5.2 剖析樹分析自然語言之文法結構

剖析樹(Parse tree)[69,70,71]是一種根據上下文無關文法(Context-free Grammar)[72,73]產生的一種樹狀表示方法，以句子如何由詞組成的方式表達文法結構，其根節點(Root Node)是一整個句子，而葉節點(Leaf Node)則是中文斷詞中的每一個詞，中間的節點則是由其所有子節點所組成之短語或是句子片段，如圖5. 1所示，此樹狀結構表示了句子依照文法的拆解方式，因此可以將短語簡單的抽取出來。而除了前述之文法架構外，部份剖析樹還提供詞性(Part of Speech)標記和語意角色(Semantic Role)[74]的標記，使得句子的文法資訊能夠更加明確。

5.3 使用樹狀條件隨機域之查詢指令生成

本章節說明如何利用樹狀條件隨機域搭配剖析樹進行查詢指令生成，其目的在找

出某個詞或短語的角色。對於每一個句子產生之詞與短語，可以標示為「答案相關」、「查詢指令相關」、「其他」三種標記方法，「答案相關」的詞或短語是可以藉由這些用語判斷出問題所問的目標，例如國家名、人名、時間、數量等等；「查詢指令相關」為句子中欲被選出當作查詢指令之詞或短語；而「其他」為不相干或是不重要的詞或短語。以「數學家高斯是哪個國家的人？」這個問句為例，「高斯」、「數學家」、「數學家高斯」應該為「查詢指令相關」，「國家」為「答案相關」，而其他詞或是短語則為「其他」標記。而被標記為「查詢指令相關」的詞或短語，將會在本章節中被當作室搜尋用的查詢指令，使搜尋引擎回傳相關的文件；而「答案相關」的類別將會在第六章中使用。

5.3.1 樹狀條件隨機域定義

在 2.3 小節中本論文曾描述過直鏈狀隨機域[40,41]，本小節將對其變化之樹狀隨機域進行說明。條件隨機域中，存在著多個隨機變數(Random Variables)以及這些隨機變數所對應之觀察(Observations)，觀察以特徵向量(Feature Vector)的形式存在；在樹狀條件隨機域中，隨機變數y_i以樹狀排列，而和其他條件隨機域相同，每一個隨機變數下皆有其觀察x_i，如圖 5. 2 所示。在每一個(x_i, y_i)中存在著描述x_i與y_i關係的特徵函數，而每一個相連的父子節點(y_i, y_j)也有一組特徵函數描述y_i與y_j若分別是某一個值所代表的關連性。

本實驗利用剖析樹的文法分析資訊與結構，以及其節點可代表短語之特性，將其做為樹狀條件隨機域的隨機變數排列準則，每一個隨機變數便是以圖 5. 1 之剖析樹結構排列，而每一個隨機變數有三種標記方法：「查詢指令相關」、「答案相關」與「其他」，且每一個隨機變數下都有其觀察量，因此，每一個隨機變數的標記不只受到父節點、子節點的標記影響，也受到觀察量的影響。

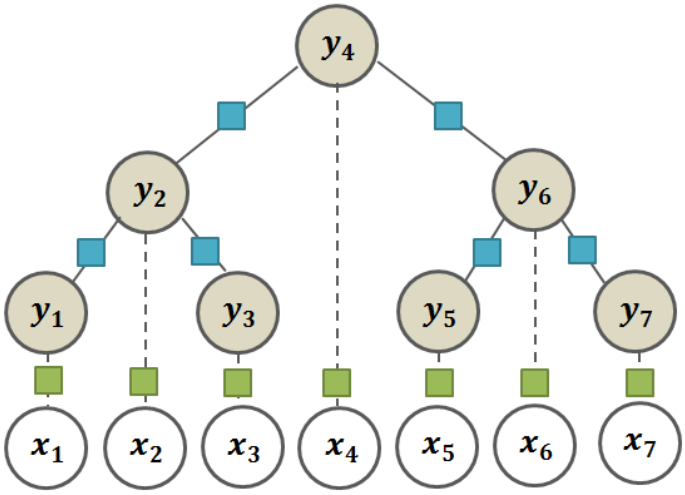

圖 5. 2　樹狀條件隨機域示意圖

5.3.2 目標函數設定

在條件隨機域中，定義如 2.3 小節所述之條件機率：

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(x)} \prod_{t=1}^{T} \phi(y_t, y_t^p, \boldsymbol{x}) \tag{5. 1}$$

其中y_t為欲標記之隨機變數，y_t^p為y_t之父節點之隨機變數，$\boldsymbol{x}$為觀測值(Observation)對應之特徵向量，$Z(x)$是正規化項，使得(5. 1)可以成為機率形式。我們可以將(5. 1)式子中的特徵方程拆解成兩大部份：

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(x)} \prod_{t=1}^{T} \phi_o(y_t, \boldsymbol{x}) \cdot \phi_t(y_t, y_t^p) \tag{5. 2}$$

其中$\phi_o(y_t, \boldsymbol{x})$是隨機變數$y_t$與觀測值$\boldsymbol{x}$之間的特徵方程，$\phi_t(y_t, y_t^p)$是父子節點之間的特徵方程。同樣地，特徵方程可以透過權重向量與特徵向量內積之方式組合而成，且我們透過指數方式使得(5. 2)中的連乘轉成連加，因此改寫如下。

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(x)} \sum_{t=1}^{T} \exp\left(\boldsymbol{\lambda_t} \cdot f_t\left(y_t, y_t^p\right) + \boldsymbol{\lambda_o} \cdot f_o(y_t, \boldsymbol{x})\right) \quad (5.3)$$

其中$\boldsymbol{\lambda_t}$與$\boldsymbol{\lambda_o}$為條件隨機域中欲求得之權重向量，$f_t\left(y_t, y_t^p\right)$為描述隨機變數之間的特徵向量，$f_o(y_t, \boldsymbol{x})$為描述隨機變數$y_t$與觀測值$\boldsymbol{x}$間之關係之特徵向量，其詳細內容會於 5.4 小節中說明。為了方便之後求解，此條件機率可以轉為對數形式：

$$p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{t=1}^{T} [\boldsymbol{\lambda_t} \cdot f_t\left(y_t, y_t^p\right) + \boldsymbol{\lambda_o} \cdot f_o(y_t, \boldsymbol{x})] - \log Z(x) \quad (5.4)$$

在擁有對數化的條件機率後，我們便可以使用如同 2.3 小節所述之高斯牛頓法求解權重向量$\boldsymbol{\lambda_t}$與$\boldsymbol{\lambda_o}$，在此不加贅述。

5.4 特徵抽取

以下章節將介紹在樹狀隨機域中使用的特徵，細分成為詞彙與文法特徵、語意特徵、網路資料特徵以及其他特徵。

5.4.1 詞彙與文法特徵

此類特徵根據詞的特性與句子文法作為分析，實驗中用到特徵如下：

- 詞頻反文件頻(Term Frequency-Inverse Document Frequency; TF-IDF)：詞頻反文件頻是用來對詞進行比重調整的一種方法，在檢索系統中，理想上的字是詞頻高而文件頻低，也就是此詞只出現在少數文件中，而在出現的文件中都有很高的出現次數；而相反地，如果某個詞散佈在大多數的文件中，那個即使在文件中的出現次數很高，對於檢索系統來說是沒有鑑別度的，又或者是此詞在文件中出現的次數太低，應該是文件中不重要的詞，像這些類別的詞之詞頻反文件頻都會偏低。實驗中，我們對於每一個詞皆計算出詞頻反文件頻，而對於每一個短語，分別計算出

最大、最小、平均之組成詞之詞頻反文件頻。

- 詞性(Part of Speech)：

  本實驗使用中研院斷詞系統[75,76]與中研院剖析系統[77]所得到的詞性結果，透過詞性標記，我們可以得到某個詞該被分為「名詞」、「動詞」、「副詞」等等詞性，例如連接詞應該在查詢指令生成中是被視為不重要的，而專有名詞則可能是重要的。

- 剖析樹(Parse tree)：

  本實驗使用中研院剖析系統[77]所得到之剖析樹，除了將剖析樹的結構用到樹狀條件隨機域上，本實驗還使用了語意角色(Semantic Role)，用來檢測某個詞在句子中所扮演的角色，例如短語「黑狗」中「狗」是短語所描述的主體(Head)，而「黑」則是用來形容「狗」的修飾語(Modifier)。

5.4.2 語意特徵

本論文利用潛藏狄氏模型(Latent Dirichlet Allocation)[78]作為主題模型分析，此模型根據機率圖學模型(Probabilistic Graphical Model; PGM)可以將其表示成圖 5. 3。其中W代表詞、Z代表某個潛藏主題、M代表文件總數、N代表該文件中字的總數、K是潛藏主題數目、θ是一個主題對文件的多項分佈(Multinomial topic distribution over document)、α是一個控制θ疏密的超參數(Hyper Parameter)、φ是主題中的文字對主題的多項分佈(Multinomial word distribution over topic)、而β同是控制φ疏密的超參數。此模型為生成模型(Generative Model)，主要用來模擬語料庫D中的每一篇文件d的生成過程：

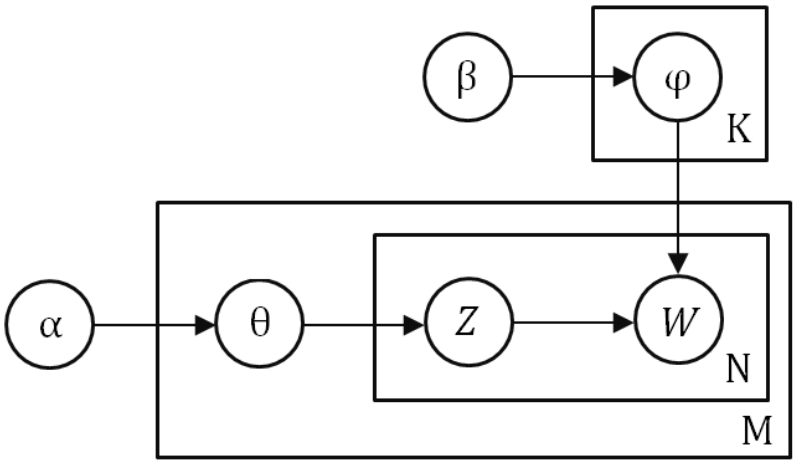

圖 5.3　潛藏狄氏分配

1. 挑選一個主題多項分佈θ_d，$\theta_d \sim Dirichlet(\alpha)$。
2. 對一篇長度為N的文件的每一個詞來說：
   a. 挑選一個潛藏主題Z，$Z \sim Multinomial(\theta_d)$。
   b. 挑選了Z後，便可得該主題下的詞多項分佈φ_{Z_k}，$\varphi_{Z_k} \sim Dirichlet(\beta)$。
   c. 挑選一個詞W，$W \sim Multinomial(\theta_z)$。

潛藏狄氏分配的學習過程，主要是調整其內部參數θ_d與φ_z使其能夠最大化訓練語料庫的機率，其目標函數(Objective Function)可以寫為：

$$P(D|\alpha,\beta) = \prod_{d=1}^{M} \int P(\theta_d|\alpha)\left(\prod_{n=1}^{N}\sum_{k=1}^{K} P(Z_k|\theta_d)P(W_{dn}|\varphi_{Z_k})\right)d\theta_d \qquad (5.5)$$

一般來說，超參數α和β是手動設定，但也有一些數學方法可以在一定迭代次數過後自動化之[79]。解這個最佳化問題，一般遭遇到最大的困難在於將(5.5)展開後會發現θ_d與φ_{Z_k}會出現耦合(Coupling)的情況而無法直接最佳化求解。因此，許多替代方案在過去的幾年中陸續被提出[80,81]，而在本篇論文中使用吉氏取樣程序(Gibbs Sampling)[82]。

而有主題模型後，本論文對每一個詞抽取潛藏主題顯著性(Latent Topic Significance; LTS)和潛藏主題亂度(Latent Topic Entropy; LTE)[57]，定義如下。

$$LTS(w_i, Z_k) = \frac{\sum_{d_j \in D} n(w_i, d_j) \mathrm{P}(Z_k|w_i)}{\sum_{d_j \in D} n(w_i, d_j)[1 - \mathrm{P}(Z_k|w_i)]} \quad (5.6)$$

$$LTE(w_i) = -\sum_{k=1}^{K} \mathrm{P}(Z_k|w_i) \log P(Z_k|w_i) \quad (5.7)$$

，其中w_i為某一個詞。而針對每一個短語，其特徵值為其組成之詞的最大、最小、平均潛藏主題顯著性與潛藏主題亂度。

5.4.3 網路資料特徵

此類特徵運用網路上大量的資料作為特徵，分為以下兩種。

- 搜尋引擎之查詢指令記錄：

  本論文使用搜狗(Sogou)搜尋引擎[83]之使用者之查詢指令記錄，實驗中使用大約三個月期間的查詢指令，大約有千萬筆記錄。此特徵的值為某個詞或短語出現在這些查詢指令集的次數，若此值越高，則代表這個詞或短語越有可能是有效的關鍵字，反之可能這個詞或短語不適合當查詢指令，又或者短語可能是組成錯誤。

- 維基百科(Wikipedia)之條目：

  本論文使用維基百科[84]的條目當作短語是否恰當之判斷依據，此外，維基百科的條目通常為適合當關鍵字的短語，故也適合用來判斷查詢指令。本論文的實驗使用大約七十萬個中文條目。此特徵值為某個詞或短語出現在這些條目的次數，若此值越高，則代表這個詞或短語越有可能是有效的關鍵字，反之可能這個詞或短語不適合當查詢指令，又或者短語可能是組成錯誤。

5.4.4 其他特徵

- 中文斷詞或剖析樹中詞(Word)或短語(Phrase)的長度，以字來計算。

- 語音辨識之信心分數：

  對於語音辨識的結果，可以透過聲學模型(Acoustic Model)和語言學模型(Linguistic Model)的分數以權重方式得到，此分數越高代表這一句話越有可能為辨識正確的句子。

- 疑問詞(Question Words)：

  本實驗中利用人工定義的九個疑問詞當作特徵，如果某個詞符合此定義的疑問詞，則特徵值為 1，反之為 0。九個疑問詞分別為：誰、哪位、何地、何人、何處、哪裡、哪、哪個、哪一國。

- 與疑問詞的距離：

  此距離的計算方式為剖析樹上的樹狀距離，每往父節點或是子節點移動時，距離值會加一。

- 知識圖譜(Knowledge Graph)：

  透過結構化的資訊，知識圖譜將語意相近、上位語、下位語等概念利用樹狀或其他結資訊表達。本實驗中透過中研院廣義知網本體架構[85,86]，抽取其人物相關與地方相關之詞語，總共抽取城市、國家、洲、職業、人物等五大類別，此類的詞分別建立一份分類列表。此特徵值為某一個詞或短語含有這些詞的個數。

5.5 實驗基礎設置

5.5.1 實驗語料與辨識

本實驗使用網路上蒐集的中文問答題庫，經過整理後留下「國家」、「城市」、「人名」三種類別，總共有 189 個問答對(Question-answering pair)，而三個類別的問題數目分別為 38、60、91 個。這些問題由單一語者錄音，總共長度約為 58 分鐘，利用 Google 語音辨識系統得到最佳辨識結果之詞錯誤率(Word Error Rate; WER)

為 12.81%，句子錯誤率(Sentence Error Rate; SER)為 59.61%。實驗中我們使用辨識轉寫(ASR Transcription)並加以斷詞(Word Segmentation)與斷句(Sentence Segmentation)，並且使用前 5 最佳結果之辨識轉寫。

每一個問答對經過斷詞與剖析樹的解析後，分別有人工標記的查詢指令，標記過程中以人看到問句時，會選擇那些詞或是短句當作查詢指令，此外每的問句除了查詢指令的標記外，均會標出和答案類別相關的詞，例如「數學家高斯是哪個國家的人」中的「國家」。

5.5.2 實驗配置

此實驗中由於在樹狀條件隨機域中並沒有調整參數，故我們將資料分為三份，將各種問句的答案類別平均打散至三份，每份 63 個問答對，每次實驗中以兩份共 126 個問答對當作訓練資料，一份共 63 個問答對當作測試資料。

5.5.3 評估方式

本實驗的評估是使用準確率(Precision)和平均準確率(Mean Average Precision; MAP)[87]來計算問句的答案是否有被包含在搜尋到的網頁之中。其中準確率的算法如下。

$$Precision = \frac{\text{含有答案之網頁數}}{\text{搜尋網頁數}} \tag{5.8}$$

而平均準確率(MAP)原是用在資訊檢索系統的評估，加入了文件排序進入考量，舉例來說，如果檢索系統回傳十篇文件，其中三篇為相關文件，如果這三篇文件是回傳的前三篇，則會比起這三篇文件在回傳的後三篇來的理想，然而只是使用準確率來計算，這兩者是相同的，平均準確率(MAP)因此在檢索系統上是十分重要的評估之一，其算法如下：

$$MAP = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{i=1}^{m_j} Precision(R_{ij}) \quad (5.9)$$

其中$|Q|$代表問題的總數，m_j代表問題Q_j經過檢索系統回傳之文件數量，R_{ij}代表回傳之文件從第一篇至第i篇之集合。

5.6 實驗結果與分析

本實驗將樹狀條件隨機域(Tree-structured CRF)與直鏈狀條件隨機域(Linear-chain CRF)進行比較，首先，實驗列出在條件隨機域之訓練與測試錯誤率，結果於表 5.1；再者利用前三、五、十篇回傳之網頁內容進行準確率(Precision)與平均準確率(MAP)的評估，三種類別分開計算實驗數據，並且提供各類別加總之平均數據，實驗結果如表 5.2 與表 5.3。

表 5.1 是直鏈狀的條件隨機域與樹狀隨機域之訓練與測試錯誤率，分別對切成三份之訓練與測試資料進行評估。由於前五最佳結果是經過語音辨識會含有較多雜訊，故在錯誤率上不管在哪種條件隨機域上都較高。

表 5.2 是在純文字之人工轉寫(Manual Transcription)上進行實驗，故實驗結果會比語音辨識後的問句來得好，在實驗結果中可以看出樹狀條件隨機域除了在「城市(city)」類別的準確率表現較差外，其它類別與其他評估方式下皆有大幅提升，可以和本實驗中短語適合作為查詢指令之假設互相呼應。

表 5.3 是在語音辨識的前五最佳辨識轉寫(ASR 5-best Transcription)上的實驗結果。在實驗中所有的類別與評估方式皆可以觀察出有大幅的改進，甚至比純文字的問句上有較大進步的幅度，其原因可能在於前五最佳辨識結果中含有許多雜訊(Noise)，造成有許多詞辨識錯誤，不應該被選為查詢指令，或是這些詞所組成的短語不正確，這些辨識錯誤的詞很可能會造成文法結構的解析不正確，而利用剖析樹的結構或許可以幫忙判斷這種情況，故在雜訊較大的狀況下使用樹狀條件隨機域與剖析樹會有較佳的結果。

	直鏈狀條件隨機域				樹狀條件隨機域(proposed)			
	人工轉寫		前 5 最佳結果		人工轉寫		前 5 最佳結果	
error	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Set1	0.1271	0.1702	0.1594	0.1910	0.1114	0.1219	0.1156	0.1691
Set2	0.1296	0.1766	0.1408	0.1925	0.0913	0.1541	0.1130	0.1642
Set3	0.1478	0.1552	0.1632	0.1797	0.1026	0.1538	0.1301	0.1374
平均	0.1348	0.1673	0.1544	0.1877	0.1077	0.1433	0.1196	0.1569

表 5.1　樹狀條件隨機域與直鏈狀條件隨機域之訓練與測試錯誤於人工轉寫(Manual Transcription)與語音辨識之前五最佳結果(ASR 5-best Transcription)

	直鏈狀條件隨機域				樹狀條件隨機域(proposed)			
type	human	city	country	average	human	city	country	average
ACC@3	0.7372	0.7809	0.7062	0.7362	0.7686	0.7524	0.7684	0.7653
ACC@5	0.6988	0.7771	0.7220	0.7221	0.7435	0.7257	0.7492	0.7417
ACC@10	0.6529	0.7257	0.6508	0.6669	0.6635	0.6743	0.6915	0.6746
MAP@3	0.8412	0.8405	0.8149	0.8326	0.8598	0.8429	0.8234	0.8448
MAP@5	0.8165	0.8295	0.7966	0.8128	0.8376	0.8511	0.8340	0.8392
MAP@10	0.7745	0.8076	0.7807	0.7832	0.8083	0.8295	0.8055	0.8117

表 5.2　樹狀條件隨機域與直鏈狀條件隨機域之比較於人工轉寫(Manual Transcription)之純文字問句。

	直鏈狀條件隨機域				樹狀條件隨機域(proposed)			
type	human	city	country	average	human	city	country	average
ACC@3	0.4609	0.4762	0.4940	0.4746	0.5185	0.5524	0.4821	0.5138
ACC@5	0.4493	0.4800	0.4500	0.4558	0.4914	0.5029	0.4464	0.4793
ACC@10	0.4160	0.4428	0.4321	0.4266	0.4383	0.4600	0.4089	0.4333
MAP@3	0.5658	0.5524	0.5923	0.5716	0.5833	0.6286	0.6235	0.6053
MAP@5	0.5430	0.5739	0.5942	0.5656	0.5715	0.6303	0.6137	0.5968
MAP@10	0.5201	0.5474	0.5614	0.5388	0.5427	0.5906	0.5778	0.5636

表 5.3　樹狀條件隨機域與直鏈狀條件隨機域之比較於語音辨識之前五最佳結果(ASR 5-best Transcription)。

5.7 章節總結

此章本論文提出樹狀條件隨機域(Tree-structured CRF)搭配剖析樹(Parse tree)的方法，進行資訊檢索基礎之問答系統的問句標記與查詢指令生成，此方法可以使得句子之文法結構可以一起被考慮，同時短語(Phrase)可以被當作查詢指令，更有效地使搜尋引擎能回傳相關的文件。實驗結果顯示單純使用問句斷詞之直鏈狀條件隨機域(Linear-chain CRF)產生的查詢指令會比本論文提出之方法差，和本論文之假設互相呼應。

第六章 前 N 最佳結果搜尋網頁結果之重排序

6.1 簡介

為了使正確辨識的句子能夠盡量被考慮，本論文在前一章節使用前五最佳辨識結果轉寫 (ASR 5-best Transcription) 取代一般最佳辨識結果轉寫 (One-best Transcription)，然而，這樣的做法會導致過多的雜訊被考慮至查詢指令中，造成搜尋引擎回傳之網頁含有許多不相關的資訊。為了解決這樣的問題，本章節提出使用雙層隨機漫步(Two-layer Random Walk)[59,60]進行網頁的重排序，不只有網頁間的資訊被考慮，辨識結果也同時被當作重排序的依據。

6.2 前 N 最佳結果應用於語音查詢指令之中文問答系統

在前一章節中，我們使用了語音辨識的前五最佳結果取代最佳結果，以達到越多正確的詞可以被加入至查詢指令中。然而，這樣的做法也可以導致越多的辨識錯誤被考慮進查詢指令中，導致之後我們越難去從這些網頁中去找到正確的答案，因此如何判斷哪些網頁是真正相關的網頁就變成一個重要的課題。

接續前一章節的查詢指令生成，我們對於每一筆辨識結果產生一筆查詢指令，再用這一筆查詢指令從搜尋引擎中檢索出若干網頁。舉例來說，對每一個問句聲音檔我們可以用前五最佳結果得到五筆辨識結果，每一筆辨識結果可以得到對應的查詢指令，而每一筆查詢指令又可以透過搜尋引擎得到若干網頁，如本論文取出十篇，故總共會有五十篇相關的文件。然而，為了讓含有答案的網頁可以在排序的最前端，網頁之間的排序因此成為一個問題，而由於牽扯到語音辨識的錯誤，並非是第一名的辨識結果就是正確答案，也並非搜尋引擎回傳的前幾篇網頁就是正確答案，目標應該是使得越有可能辨識正確的結果產生的網頁越前面，讓含有答案的網頁排序越前面。

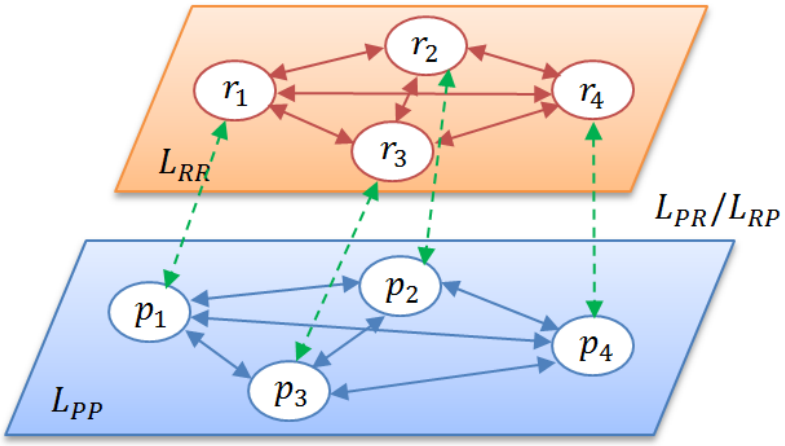

圖 6. 1　雙層隨機漫步(Two-layer Random Walk)示意圖

6.3 以雙層隨機漫步進行前 N 最佳結果與對應網頁之重排序

為了將上述的若干網頁能夠將含有答案的網頁排序到越前面，本章節利用雙層隨機漫步對眾多網頁進行重排序，對於每一筆辨識轉寫結果進行查詢指令生成，各自透過搜尋引擎得到對應之網頁，例如本實驗中以前五最佳辨識結果為例，一個聲音訊號的問句會得到五筆辨識結果，而這五筆辨識結果分別透過前一章節敘述之條件隨機域可以得到適當之查詢指令，進而透過搜尋引擎得到對應之網頁各十篇，因此整個系統將要對這五十篇網頁進行重新排序。

本論文利用語音辨識結果與網頁內容判斷一個網頁是否有越高可能含有正確答案，越可能辨識正確的問句產生之查詢指令與對應之網頁，以及和大部份的網站較為相似的網頁，應該是被認為含有正確答案的網頁。透過雙層隨機漫步演算法，可以使得網頁與辨識結果互相強化分數，以下將定義本論文中如何將這兩種資訊加入考量。

6.3.1 雙層之定義—前 N 最佳結果與搜尋網頁

本論文定義雙層資訊分別為問句之錄音辨識轉寫層與網頁層，如圖 6. 1 上半部

是錄音辨識轉寫層，而下半部是網頁層。在錄音辨識轉寫層中，每一個前 N 最佳結果之辨識句子有一個對應的節點以及分數r_i，總和所有的節點可形成一分數向量F_R，如果此句子對應之分數越高，則代表此句子越有可能是辨識正確的問句，而每一個句子(節點)與句子(節點)之間存在著相似性的分數，這些相似度可以用相似度矩陣L_{RR}的方式表達，如果某一辨識結果和辨識結果信心分數很高的辨識結果的相似性愈高，則此辨識結果的分數也應該被提高，反之則會減少；在網頁層中，每一個網頁當作一個節點p_i，以分數向量F_P的形式表達，同樣地每一個網頁之間也有相似性的分數，同樣也以相似度矩陣L_{PP}表達。而在這兩層之間也有每一個轉寫問句對應到每一個網頁之相似度矩陣L_{PR}與L_{RP}，讓兩層資料可以互相傳遞，在本論文中定義問句對應到每一個網頁之相似度為該網頁是否由此問句的查詢指令產生之，若陳述為真，則該矩陣值為 1，否則為 0。此外，在本論文中的矩陣L_{PP}、L_{RR}、L_{PR}與L_{RP}都已經過正規化(Normalization)轉換。

6.3.2 雙層之資訊傳遞進行重排序

由於在第四章已經談論過雙層隨機漫步，因此對於演算法將不再贅述，透過雙層分數的傳遞，網頁的分數可以透過辨識結果層傳來，而辨識結果層的分數可以透過網頁層的分數傳來，不斷地利用(6. 1)迭代計算出網頁層與辨識結果層之分數後，我們將收斂的網頁層分數由大到小當作網頁新的排序方法。

$$\begin{cases} F_P^{(t+1)} = (1-\alpha)F_P^{(0)} + \alpha \cdot L_{PP}^T \cdot L_{PR} \cdot F_R^{(t)} \\ F_R^{(t+1)} = (1-\alpha)F_R^{(0)} + \alpha \cdot L_{RR}^T \cdot L_{RP} \cdot F_P^{(t)} \end{cases} \tag{6. 1}$$

其中$F_P^{(t)}$和$F_R^{(t)}$分別代表網頁層與辨識層在第t個迭代時的分數向量，L_{PP}和L_{RR}分別代表網頁層與辨識層內之相似度矩陣，與第四章節相同使用 OKAPI/BM25 之餘弦相似度做計算，L_{PR}與L_{RP}代表兩層之間的相似度矩陣，α為一權重參數調整分數傳遞的多寡。若某網頁的分數愈高，代表此網頁可能和其他搜尋到的網頁內

容愈像，也可能是因為此網頁所對應之辨識結果越高分，這兩種情況皆越有可能是含有正確答案的結果。此外，若是辨識結果愈高分，代表此辨識結果和其他辨識結果愈像，或是此辨識結果對應到的網頁皆為高分，這樣的情況下代表此辨識結果越可能是正確的。經過雙層之間的互相傳遞與加強彼此的分數，讓重新排序後的結果能夠包含更多的答案相關網頁。

此外，由於在這邊的雙層隨機漫步要重排序網頁的分數，故我們將定義雙層分數之起使值$F_P^{(0)}$和$F_R^{(0)}$：

$$F_P^{(0)}(p_i) = \frac{1}{rank(p_i) \times nbest(p_i)} \tag{6. 2}$$

$$F_R^{(0)}(r_i) = \begin{cases} 0.01, & if\ webpages(r_i) = \emptyset \\ 1, & otherwise \end{cases} \tag{6. 3}$$

其中p_i是一篇網頁內容，$rank(p_i)$是此網頁在搜尋引擎的回傳清單中排名第幾，$nbest(p_i)$是此網頁是由前 N 最佳辨識結果中的第幾個辨識結果所產生，r_i是一筆辨識結果，$webpages(r_i)$是此辨識結果產生之查詢指令，利用 Google 搜尋的「一字不差」功能所產生的網頁集合，部份辨識錯誤的結果可能因為某些詞不包含在語音辨識的詞典中(Out of Vocabulary; OOV)，造成辨識結果會用其他方法組成詞，使得短語更加模糊，因此在這些狀況下利用「一字不差」功能會有空集合的狀況發生。

6.4 實驗基礎設置

6.4.1 實驗語料與辨識

本實驗使用網路上蒐集的中文問答題庫，經過整理後留下「國家」、「城市」、「人名」三種類別，總共有 189 個問答對(Question-answering Pair)，而三個類別的問題數目分別為 38、60、91 個。這些問題由單一語者錄音，總共長度約為 58 分鐘，

利用 Google 語音辨識系統得到最佳辨識結果之詞錯誤率(Word Error Rate; WER)為 12.81%，句子錯誤率(Sentence Error Rate; SER)為 59.61%。實驗中我們使用辨識轉寫(ASR Transcription)並加以斷詞(Word Segmentation)與斷句(Sentence Segmentation)，並且使用前 5 最佳結果之辨識轉寫。

本章節所用到的查詢指令是由第五章的樹狀隨機域所產生，而對應到該查尋指令之網頁文件同樣也是第五章在透過 Google 搜尋所得到的網頁，每一個搜尋指令有對應到十篇網頁文件，由於一個問句使用前五最佳辨識結果，故一個問句會有相關的五十篇文件。

6.4.2 實驗配置

本論文將資料分為三份，將各種問句的答案類別平均打散至三份，每份 63 個問答對，每次實驗中以兩份共 126 個問答對當作訓練資料，一份共 63 個問答對當作測試資料。

6.4.3 評估方式

本章節的實驗之評估方式如同第五章，使用準確率(Precision)和平均準確率(Mean Average Precision; MAP)。

6.5 實驗結果與分析

本章節將詳述實驗結果並討論分析。實驗針對所有回傳之網頁內含有的答案匹配以準確率和平均準確率評估，並和未經過重排序之雙層隨機漫步重排序後的分數做比較。實驗結果如圖 6.2 以及圖 6.3，分別為準確率與平均準確率之結果，

本章節將詳述實驗結果並討論分析。實驗針對所有回傳之網頁內含有的答案匹配以準確率和平均準確率評估，並和未經過重排序之雙層隨機漫步重排序後的分數做比較。實驗結果如圖 6.2 以及圖 6.3，分別為準確率與平均準確率之結果，

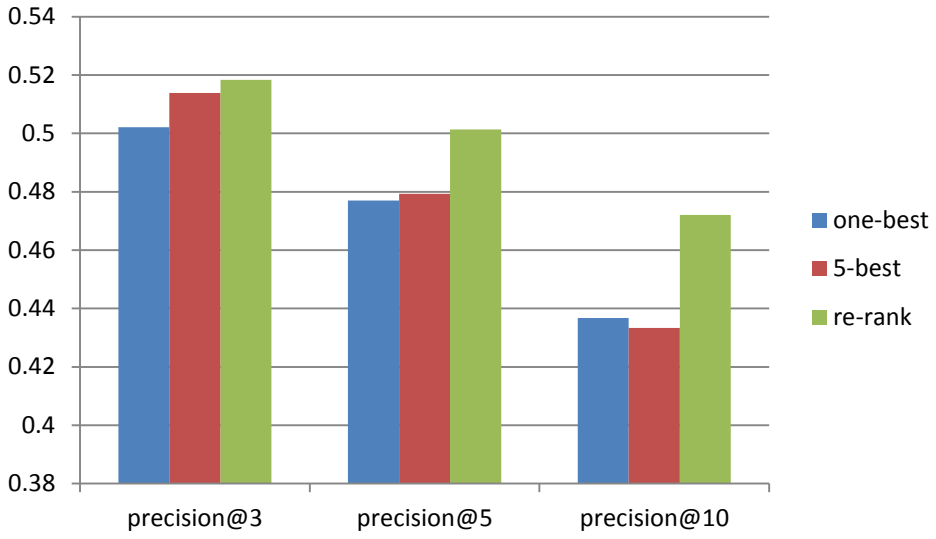

圖 6. 2　網頁排序之準確率(Precision)之比較

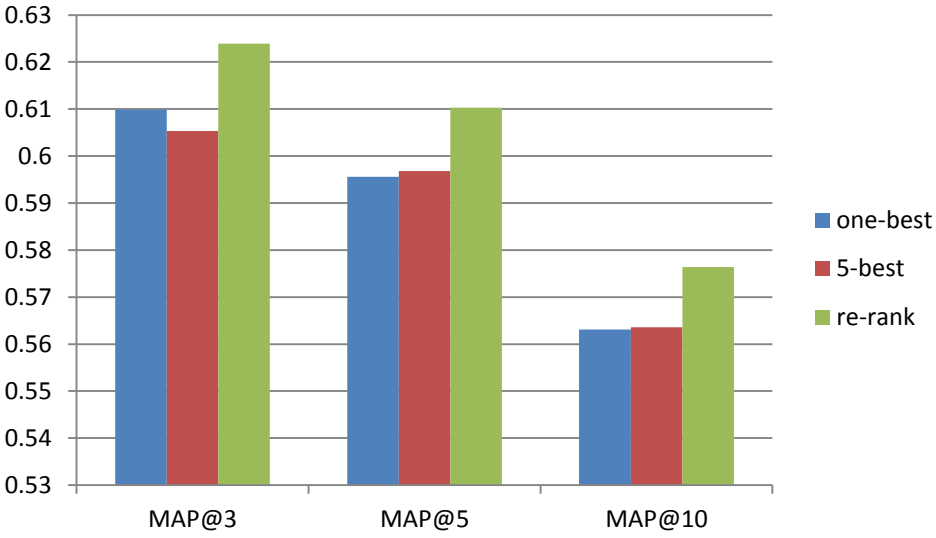

圖 6. 3　網頁排序之平均準確率(Mean Average Precision)之比較

one-best 是僅使用最佳辨識結果產生之查詢指令得到的網頁分數，而 5-best 是使用前五最佳辨識結果之查詢指令之辨識分數，結果可以看出使用前五最佳辨識結果可以得到比最佳辨識結果稍好的答案，其原因可能在於使用前五最佳辨識結果可以使越多的正確辨識的關鍵字被考慮進查詢指令中，因此可以有比較好的表現；

另一方面，由於前五最佳辨識結果含有太多的雜訊，造成搜尋引擎無法準確的找到相關網頁，因此表現的上升有限。Re-rank 是使用雙層隨機漫步後的重排序結果，可以看出明顯比不使用重新排序來的有進步，由其在前五、前十篇的評估，其原因在於網頁重排序讓本來是相關但分數低的網頁可以將排名往前，使得準確率和平均準確率皆有大幅進步。舉例來說，可能由較後排名的辨識結果產生之網頁，其實和排名前面的網頁十分相似，因此其網頁排序應當是要往前的，又或者是某個排名後面的辨識結果和高分的辨識結果非常相似，因此這個辨識結果對應之網頁應該要將排名往前。

6.6 章節總結

此章節提出雙層隨機漫步(Two-layer Random Walk)方法對含有較多雜訊之前五最佳辨識結果轉寫產生之查詢指令與對應之網頁進行重排序，利用辨識轉寫層與網頁層之間互相傳遞分數，讓越有可能辨識正確之轉寫對應之網頁提高分數，也對越多網頁共有之辨識結果提高分數，兩層之間互相分數的傳遞，使得高分的網頁代表越可能是正確辨識產生之查詢指令或是重要的網頁。實驗結果顯示使用雙層隨機漫步之網頁重排序結果明顯比單純使用最佳辨識結果或是前五最佳辨識結果有更大的進步。

第七章 答案種類判定與答案生成

7.1 簡介

在前兩個章節本論文介紹如何對問句抽取查詢指令以及如何對搜尋引擎回傳之網頁做重排序,在此一章節本論文將說明如何從這些回傳網頁中取得問句所對應之答案。首先本論文針對問句抽取答類別，而透過判定之答案類別，再從網頁內容中尋找對應的答案。

7.2 答案種類分析與判別

由於問句對應之答案的種類十分多元,在一般的情況下很難使所有可能的答案都考慮，因此，先將答案種類做判定，例如決定要答案種類是數字、年代、人名、地名等等，再透過答案種類選擇可能的答案，就能將可能的答案數量大幅減少，使得正確率提高。本章節主要探討如何對答案種類進行分類，本論文使用的答案種類有三種:「國家」、「城市」以及「人名」，以下章節將會述敘如何從問句得到答案的種類，以利之後答案之選擇。

7.2.1 支撐向量機進行答案種類分類

為了在抽取答案前先知道答案的種類以利之後答案之選取,本論文使用支撐向量機針對每一筆問句做答案種類之分類，在本論文中總共有三種類別，分別是「國家」、「城市」、「人物」。由於在第二章已經敘述過支撐向量機之運行方法，故在此處不加贅述。

7.2.2 特徵抽取

本章節將描述在答案類別判定中使用的特徵向量。

- 疑問詞：

  本實驗中利用人工定義的九個疑問詞當作特徵，如果某個詞符合此定義的疑問詞，則特徵值為 1，反之為 0。九個疑問詞分別為：誰、哪位、何地、何人、何處、哪裡、哪、哪個、哪一國。

- 類專有名詞：

  在這一個章節中，我們透過史丹佛大學提供之類專有名詞分析服務[88]從句子中抽取類專有名詞，然而這樣的做法在召回率偏低，因此本論文中加入知識圖譜中[85,86]抽取與類專有名詞相關之詞，以及維基百科[84]中的人名清單，如果某個詞對照到這些清單，便也當作是類專有名詞。本論文中使用的類專有名詞為三種：「人名」、「地名」以及「國家名」。

- 答案相關詞之類專有名詞：

  僅針對答案相關詞做類專有名詞分析，同樣也是用「人名」、「地名」以及「國家名」三種類專有名詞。

- 類專有名詞與疑問詞之距離：

  此距離的計算方式為剖析樹上的樹狀距離以及句子中的順序的距離。樹狀距離定義為每往父節點或是子節點移動時，距離值會加一；句子中的順序的距離為兩個詞之間有多少的中文字。

- 答案相關詞之類專有名詞和疑問詞之間的距離。

  僅針對答案相關詞做類專有名詞分析，同樣也是用「人名」、「地名」以及「國家名」三種類專有名詞。距離的計算方式同上。

7.3 答案生成前處理

在每一個網頁內容中，先抽取所有的網頁文字，由於網頁內容中有許多標籤內沒有文字，所以在截取的文字中會含有相當多的空白句，若某一個句子相鄰的五個

句子皆為空白句，則使用該句子做為文件段落的斷點。

而將文件段落抽取後，本論文利用史丹佛大學提供之類專有名詞解析(Stanford Named Entity Recognizer)[88]套件、中研院廣義知網[85,86]，以及維基百科(Wikipedia)[84]中的人名列表作為分析類專有名詞的依據，對於文件中的詞上做出對應之類專有名詞之標記。

7.4 答案生成

本實驗中使用基礎的方法判定答案，在搜尋到的網頁內容中，若是某個潛在的答案出現越多次，則越有可能是正確答案，然而由於網頁常常含有不相關的資訊，有可能是網頁內的其他框架中的文字，或是網頁內的內容只有一小部份為和問句相關，因此顯然某些部份的文字可能和目標問句不相關，因此論文中再加上與查詢指令(Query)的距離當作判斷依據。此外回傳排序越前面的網頁應被視為和查詢指令越為相關，因此評分方式中加入網頁的排序一同考慮：

$$\sum_{i=1}^{|D|}\sum_{j=1}^{|E|}\frac{1}{rank(d_i)*distance(q,entity_j)} \qquad (7.1)$$

其中d_i為一篇文件，$|D|$為抽取出的文件總數，q為第五章使用生成之查詢指令，$entity_j$為類專有名詞，也就是可能的答案，$|E|$為文件中抽取出的類專有名詞數量，$rank(d_i)$代表文件d_i在檢索系統中被回傳的順序，$distance(q,entity_j)$則代表查詢指令q在文件d_i中與類專有名詞$entity_j$之間的距離，其距離的算法為類專有名詞離最近的查詢指令詞或短語相隔幾個字，再對距離進行文件中的段落長度正規化(Normalization)。

7.5 實驗基礎設置

7.5.1 實驗語料與辨識

本實驗使用網路上蒐集的中文問答題庫，經過整理後留下「國家」、「城市」、「人名」三種類別，總共有 189 個問答對(Question-answering Pair)，而三個類別的問題數目分別為 38、60、91 個。這些問題由單一語者錄音，總共長度約為 58 分鐘，利用 Google 語音辨識系統得到最佳辨識結果之詞錯誤率(Word Error Rate; WER)為 12.81%，句子錯誤率(Sentence Error Rate; SER)為 59.61%。實驗中我們使用辨識轉寫(ASR Transcription)並加以斷詞(Word Segmentation)與斷句(Sentence Segmentation)，並且使用前 5 最佳結果之辨識轉寫。

本實驗中使用答案相關詞抽取是從第五章之樹狀條件隨機域中抽取而出，而使用的網頁是從第六章之使用雙層隨機漫步重排序之網頁，每一個問句有對應之五筆辨識結果轉寫，以及五十篇網頁文件。

7.5.2 實驗配置

本論文將資料分為三份，將各種問句的答案類別平均打散至三份，每份 63 個問答對，每次實驗中以兩份共 126 個問答對當作訓練資料，一份共 63 個問答對當作測試資料，此資料為答案種類判定之使用。而由於答案生成目前為非督導式方法且無參數之調整，故直接將所有問句對應之網頁文件、答案種類一同使用。

7.5.3 評估方式

在答案類別的判定上，本論文使用正確率計算有百分之多少的問句對應到的答案類別被標記正確。而答案生成的部份使用前 N 個答案清單中是否含有正確答案之正確率，如果在前 N 的答案候選清單中出現，便算是正確，意即在前 N 個可

	特徵向量	正確率
(1)	疑問詞	74.60%
(2)	+ 知識圖譜中的分類	75.66%
(3)	+ 「答案相關」的詞或短語在知識圖譜分類中出現的次數	92.59%
(4)	+ 距離相關特徵	98.41%

表 7.1　答案種類判定之結果

能答案中，有多少百分比的問句可以從中找到答案，其中在本論文中 N 使用 1、2、3、5、10、20，此外還使用平均倒數排名(Mean Reciprocal Rank; MRR)[89,90]，由於在問答系統中只在乎一個正確答案的排名，故其評估的公式僅使用答案在回傳清單中的順序：

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{7.2}$$

其中$|Q|$為問句的總數，而$rank_i$對第 i 個問句回傳之答案清單，第幾個答案是正確答案。

7.6 實驗結果與分析

7.6.1 答案類別判定

本論文使用向量支撐機對於每一個問句做答案判別，實驗結果如表 7.1。每一列分別是基於前一列再加上指定的特徵向量。第一列為僅使用人工事先定義好的九

個疑問詞當作特徵，第二列中加入以問句中是否有先前描述之知識樹的詞，這兩列的訓練特徵均是沒有加入第五章之樹狀條件隨機域判別之答案相關詞，實驗結果中可以發現兩者無法有效的使答案種類被判別，其原因在於許多疑問詞是使用「哪個」或「哪一個」，因此無法有效的得知答案種類，第二列雖然再加上詞的類專有名詞分類，但是因為在問句中往往會加入許多描述用的詞，例如「數學家高斯是哪一國人？」中會檢測到許多「人」的類專有名詞分類，但是事實上問句想要問的內容為「國家」；而第三列和第四列為分別為限定答案相關詞之疑問詞與距離相關資訊，這兩列為有加入答案相關字的相關特徵，可以明顯看出對答案種類的判別有相當大的幫助。由於在過去的答案相關字產生時，即有加入距離相關資訊，因此並不是所有被偵測到是類專有名詞的詞會被考慮，而是只有在少部份較為可能是答案類別相關的詞會被考慮，因此可以解決之前所述的狀況產生之問題。

7.6.2 答案生成

本論文使用基本的特徵如距離和文件的排序做分析，判斷哪一個潛在的詞可能是問句的答案，實驗結果如表 7. 2，評估方式為前 N 個回傳之答案中含有正確答案之百分比以及平均倒數排名(Mean Reciprocal Rank; MRR)。(a)欄為僅只使用某類專有名詞在文件中的出現次數而不考慮使查詢指令之距離，即：

$$\sum_{i=1}^{|D|}\sum_{j=1}^{|E|}\frac{1}{rank(d_i)} \tag{7.3}$$

其中d_i為一篇文件，$|D|$為抽取出的文件總數，q為第五章使用生成之查詢指令，$entity_j$為類專有名詞，也就是可能的答案，$|E|$為文件中抽取出的類專有名詞數量，$rank(d_i)$代表文件d_i在檢索系統中被回傳的順序。而(b)欄為(7. 1)式之權重方式計算之答案生成法，其距離的計算方式為中間有多少個中文字，而(c)欄則

	(a) count	(b) distance	(c) normalized distance
1	0.3793	0.3851	0.3966
2	0.4310	0.4828	0.4828
3	0.4885	0.5459	0.5345
5	0.5459	0.5977	0.6207
10	0.6436	0.6839	0.6839
20	0.6896	0.7184	0.7184
MRR	0.4532	0.4803	0.4868

表 7.2　答案生成(Answer Generation)

是將距離對文件總長度做正規化。實驗結果可以發現(a)欄由於不考慮與查詢指令之關係，由於大部分的網路文件會有很多不相關資訊，在不考慮與問句相關之查詢指令的狀況下，容易將這些不相關資訊一同考慮，故表現較差。而加入了距離資訊後，(b)欄和(c)欄皆有顯著的進步，此外(c)欄加入了距離的正規化，更能符合每一篇網頁之不同特性，像是某些網頁只使用簡單的詞來表達關係，例如「作者：余光中」，而某些網頁會用許多文字來描述，因此在距離的計算上不能等同視之，因此加入正規化的距離後可以在實驗結果中進步。

7.7 章節總結

此章節對答案種類判定和答案生成進行初步的實驗分析，透過網頁排序與網頁內容與查詢指令之距離進行答案判定，同時在實驗結果中可以看到使用兩者皆對答案生成有正向影響。

Part IV

結論與展望

第八章 結論與展望

8.1 結論

本論文在第二部份中提出了一系列之抽取式語音摘要上之研究。首先，提出督導式之利用結構式支撐向量機加上語句叢集作為隱藏變數，可在訓練階段整體學習(Joint Learning)出語句重要性以及整體摘要重複性的比重，還可以利用語句叢集同時考慮叢集對摘要選取的影響。再者，非督導式之雙層隨機漫步演算法，輔以課程系統附有的投影片，利用隨機漫步的分數傳遞方式，讓語句的分數不僅只是相互之間的相似度作傳遞，而是會加上與投影片的相似程度作傳遞。在本論文中的兩個實驗皆使用 ROUGE 方法對抽取式摘要進行評估，均顯示比起其他演算法有更多的進步，尤其以督導式的方法可以有更好的表現。

在第三部份中，本論文提出以資訊檢索為基礎之中文的模擬陳述問答系統(Factoid Question Answering)。首先本論文先討論查詢指令生成的部份，利用樹狀條件隨機域配合中文剖析樹之結構，找出合理的短語(Phrase)或是詞作為查詢指令，評估方式透過找出的網頁是否含有對應的答案作為計算，分別用準確率和平均準確率(MAP)，也得到大幅的進步。而由於語音文件摘要的特性，本論文中使用前 N 最佳結果取代最佳結果之辨識轉寫，因而造成查詢指令和對應的網頁含有的雜訊過多，因此本論文提出利用雙層隨機漫步得到較佳的網頁排序結果。而最後一個部分則是從利用查詢指令找到的網頁內容中，找出真正對應到問句的答案，此部份為初步的實驗結果，但可能結果看出網頁排序和文件內容與查詢指令相關性對於答案生成有正向的影響，可以在後續實驗中用更精緻的方法應用。

8.2 未來研究方向

在語音摘要中，在第三章和第四章中可以看到文件的結構和投影片資訊對摘要的表現均有幫助，而在本論文中尚未將此兩種資訊結合，例如在結構式支撐向量機中的隱藏變數可以不只考慮連續句子之叢集，也可以考慮每一個句子與投影片的對應，或是將投影片資訊加入至特徵向量中。此外，在課程錄音中，語者通常會反覆強調某一個重要的觀念，或是利用之前提過的概念來加強某一論點，因此除了連續句子的叢集結構外，課程錄音的句子可能含有其他更有用的資訊，例如在語料中一在反覆強調的段落叢集等等。

此外，隨著近幾年網路上的課程影片之普及，如 Coursera 等網站提供大量的科目之授課內容，而由於許多大學都會開設相似的科目，因此使用者可能會花許多時間去聽不同大學開設之同一門課程，如此會降低聽課的效率。而為了解決這個問題，可以結合之前所提過的多文件摘要，將許多相同的課程直接作成摘要，方便使用者可以快速得知授課內容。

在問答系統中，在本論文中每個步驟分開進行，然而可能由於前一步的表現不佳，導致下一步的結果退步，因此可以考慮使用整體學習(Joint Learning)的方式求出答案，讓每一步的錯誤互相影響的程度減低。此外，在答案生成的部份可以考慮更複雜的特徵，如信息抽取(Information Extraction)之關係抽取，還有可以利用網際網路中巨量資料進行文法的分析，例如找尋類似於問句之句法結構，並且從網路資料中找到最相像的答案；在抽取答案時，也可以使用條件隨機域等結構式的學習法找尋答案，文件中每一個出現的類專有名詞有互相有相關性，而非僅用簡單的評分方式；此外，在本論文中先行判斷了答案類別，再依據此類別對該類別之類專有名詞進行評估，也有可能利用此方式時可以不僅是考慮答案類別辨識相同的類專有名詞，可以透過機器學習的方式，有機會讓即使先前答案類別辨識錯誤的問句也可以找到正確答案。

本實驗的問答系統屬於模擬陳述問答(Factoid QA)，而還有另外一大部份的問答系統屬於定義式問答(Definitional QA)，其目標不在於找到一個唯一解的答案，此種問答系統的答案具有多樣性，且有多種表達方式，只要答案內容有說明問句要問的內容，即可以算是正確，此種定義式問答可以使用文件摘要的方式來找到問句對應的答案，類似於針對問句摘要(Query-focused Summarization)之文件摘要，因此可以結合論文中的摘要方法再對此種問答進行研究。

參考資料

[1] http://www.coursera.org/.

[2] T. Hazen J. Glas and R. Barzilay, “Recent progress in the mit spoken lecture processing project,” in Proceedings of 8th Annual Conference of the International Speech Communication Association, 2007.

[3] Jingjing Liu, Scott Cyphers, Panupong Pasupat, Ian Mcgraw, and Jim Glass, “A Conversational Movie Search System Based on Conditional Random Fields,” Interspeech, 2012.

[4] Ani Nenkova, “A survey of text summarization techniques,” in Multimedia, 2001.

[5] Yang Liu and Dikel Hakkani-Tur, “Spoken language understanding. (ch13),” 2005, vol. 40, pp. 1–60.

[6] http://www.bing.com/

[7] Yang Liu and Shasha Xie, “Impact of automatic sentence segmentation on meeting summarization,” in ACL, 2008.

[8] Bin Li Yang Liu, Feifan Liu and Shasha Xie, “Do disfluencies affect meeting summarization: A pilot study on the impact of disfluencies,” in IEEE SLT, 2007.

[9] Shasha Xie Yang Liu and Fei Liu., “Using n-best recognition output for extractive summarization and keyword extraction in meeting speech,” in ICASSP, 2010.

[10] Shasha Xie and Yang Liu., “Using confusion networks for speech summarization,” in ACL, 2010.

[11] Sameer Maskey, “Comparing lexical, acoustic/prosodic structural and discourse features for speech summarization,” in Interspeech, 2005.

[12] Jian Zhang, “Speech summarization without lexical features for mandarin broadcast news.,” in HLT-NAACL, 2007.

[13] ShaSha Xie and Yang Liu, “A comparative study on speech summarization of broadcast news and lecture speech,” in ASRU, 2009.

[14] ShaSha Xie and Yang Liu, “Integrating prosodic features in extractive meeting summarization,” in Interspeech, 2010.

[15] J. Hirschberg, “Communication and prosody: functional aspects of prosody,” in Speech Communication, 2002, pp. 31–43.

[16] H. Chan J. Zhang and P. Fung, “Improving lecture speech summarization using rhetorical information,” in Proceedings of Automatic Speech Recognition and Understanding, 2007.

[17] Gerald Penn Xiaodan Zhu and F. Rudzicz, “Summarizing multiple spoken documents: Finding evidence from untranscribed audio,” in ACL, 2009.

[18] Jaime Carbonell and Jade Goldstein, “The use of mmr, diversity-based reranking for reordering documents and producing summaries,” in ACM SIGIR, 1998.

[19] Shasha Xie and Yang Liu, “Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization,” in ICASSP, 2008.

[20] Chin-Yew Lin, “ROUGE: A package for automatic evaluation of summaries,” in Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)

[21] C. yew Lin, “Rouge: A package for automatic evaluation of summaries,”in *Workshop on Text Summarization Branches Out*, 2004.

[22] DMW Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation,” Journal of Machine Learning Technologies 2 (1), 37-63, 2011.

[23] Silviu Cucerzan and Eugene Agichtein, “Factoid Question Answering over Unstructured and Structure,” in CIKM 2005.

[24] Eduard Hovy , Ulf Hermjakob , Chin-yew Lin, “The Use of External Knowledge in Factoid QA,” In Proceedings of the Tenth Text REtrieval Conference (TREC), 2001.

[25] Cucerzan, Silviu and Agichtein, Eugene, "Factoid Question Answering over Unstructured and Structured Web Content," Paper presented at the meeting of the TREC, 2005.

[26] E. Cabrio, J. Cojan, A.P. Aprosio, B. Magnini, A. Lavelli, and F. Gandon, "QAKiS: an Open Domain QA System based on Relational Patterns," in Proc. International Semantic Web Conference (Posters & Demos), 2012.

[27] John Prager, ”Open-domain question: answering”, in Foundations and Trends in Information Retrieval Volume 1 Issue 2, 2006.

[28] Dragomir Radev, Weiguo Fan, Hong Qi, Harris Wu, and Amardeep Grewal, “Probabilistic Question Answering on the Web,” WWW 2002.

[29] David Azari, Eric Horvitz, Susan Dumais, and Eric Bril, “Web-Based Question Answering: A Decision-Making Perspective,” UAI 2003.

[30] Santosh Kumar Ray, Shailendra Singh, and B.P.Joshi, “World Wide Web Based Question Answering System – A Relevance Feedback Framework for Automatic Answer Validation,” ICADIWT 2009.

[31] Mikhail Ageev, Dmitry Lagun, Eugene Agichtein, “The Answer is at your Fingertips: Improving Passage Retrieval for Web Question Answering with Search Behavior Data,” EMNLP 2013.

[32] Eric Brown, Eddie Epstein, J William Murdock, Tong-Haing Fin, “Tools and Methods for Building Watson,” IBM Research Report RC25356, 2013

[33] D Ferrucci, A Levas, S Bagchi, D Gondek, E T Mueller, “Watson: Beyond Jeopardy!,” Artificial Intelligence199-200, 93-105, Elsevier, 2013

[34] A. Lally, J. M. Prager, M. C. McCord, B. K. Boguraev, S. Patwardhan, J. Fan, P. Fodor, J. Chu-Carroll, “Question analysis: How Watson reads a clue,” IBM

Journal of Research and Development 56(3.4), 2--1, IBM, 2012

[35] Voorhees, E. & Tice, D. "Building a Question Answering Test Collection", Proceedings of SIGIR-2000, July, 2000, pp. 200-207.

[36] Peter Clark, John Thompson, and Bruce Porter, "A Knowledge-Based Approach to Question-Answering," in AAAI, 1999.

[37] Fabio Rinaldi, James Dowdall, Michael Hess, Diego Mollá, Rolf Schwitter, and Kaarel Kaljurand, "Knowledge-Based Question Answering," Lecture Notes in Computer Science Volume 2773, 2003, pp 785-792, 2003.

[38] http://www.wolframalpha.com/

[39] P. Zhou and J. Austin, "Learning criteria for training neural network classifiers," Neural Computing & Applications, Volume 7, Issue 4 , pp 334-342, 1998.

[40] Lafferty, J., McCallum, A., Pereira, F, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," Proc. 18th International Conf. on Machine Learning. Morgan Kaufmann. pp. 282–289, 2001.

[41] Sutton, Charles; McCallum, Andrew, An Introduction to Conditional Random Fields, FnTML 2010.

[42] Rennie, J.; Shih, L.; Teevan, J.; Karger, D, "Tackling the poor assumptions of Naive Bayes classifiers," ICML 2003.

[43] Vladimir N. Vapnik, "The Nature of Statistical Learning Theory", Springer, 1995.

[44] Thorsten Joachims, "Making Large-Scale SVM Learning Practical", LS8-Report, 24, Universität Dortmund, LS VIII-Report, 1998.

[45] Thorsten Joachims, "Learning to Classify Text Using Support Vector Machines", Dissertation, Kluwer, 2002.

[46] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury, " Deep Neural Networks for Acoustic Modeling in

Speech Recognition," in Signal Processing Magazine, 2012.

[47] T. Joachims, "Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning", B. Schölkopf and C. Burges and A. Smola (ed.), *MIT-Press*, 1999.

[48] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. "Support Vector Learning for Interdependent and Structured Output Spaces," ICML, 2004.

[49] T. Joachims. "Learning to Align Sequences: A Maximum Margin Approach, Technical Report," August, 2003.

[50] T. Joachims, T. Finley, Chun-Nam Yu, "Cutting-Plane Training of Structural SVMs," Machine Learning Journal, 77(1):27-59, 2009.

[51]Forney, G.D., Jr., "The viterbi algorithm," *Proceedings of the IEEE* , vol.61, no.3, pp.268,278, March 1973.

[52] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Technical Report. Stanford InfoLab, 1999.

[53] Günes Erkan and Dragomir R. Radev, "LexRank: graph-based lexical centrality as salience in text summarization.", Journal of Artificial Intelligence Research, Volume 22 Issue 1, July 2004.

[54] Pemantle, Robin. "A survey of random processes with reinforcement," Probability Surveys 4, 1--79. doi:10.1214/07-PS094, 2007.

[55] Hung-yi Lee, Yu-yu Chou, Yow-Bang Wang, Lin-shan Lee, "Supervised Spoken Document Summarization Jointly Considering Utterance Importance and Redundancy by Structured Support Vector Machine", Annual Conference of the International Speech Communication Association, 2012

[56] Thomas Hofmann, "Probabilistic latent semantic analysis," in *UAI*, 1999.

[57] Sheng-Yi Kong and Lin-Shan Lee, “Semantic analysis and organization of spoken documents based on parameters derived from latent topics,” in Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, pp. 1875-1889, 2011.

[58] Gerard Salton and Christopher Buckley, “Term-weighting approaches in automatic text retrieval,” Inf. Process. Manage. 24, 5, p513-p523, 1988.

[59] Yun-Nung Chen, and Florian Metze, “Two-Layer Mutually Reinforced Random Walk for Improved Multi-Party Meeting Summarization,” SLT 2012.

[60] Sujay Kumar Jauhar, Yun-Nung Chen, and Florian Metze, “Prosody-Based Unsupervised Speech Summarization with Two-Layer Mutually Reinforced Random Walk,” *IJCNLP 2013*.

[61] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford, “Okapi at TREC-3,” TREC 1994.

[62] Joseph K. Bradley and Carlos Guestrin, “Learning Tree Conditional Random Fields,” ICML 2010.

[63] Wei Lu, Hwee Tou Ng and Wee Sun Lee, “Natural Language Generation with Tree Conditional Random Fields,” EMNLP 2009.

[64] Trevor Cohn and Philip Blunsom, “Semantic Role Labelling with Tree Conditional Random Fields,” CoNLL 2005.”

[65] Jie Tang, Mingcai Hong, Juanzi Li and Bangyong Liang, “Tree structured Conditional Random Fields for Semantic Annotation,” ISWC 2006.

[66] Thomas Mensink, Jakob Verbeek, and Gabriela Csurka, “Tree structured CRF Models for Interactive Image Labeling,” IEEE Trans. Pattern Anal. Mach. Intell. 35(2): 476-489, 2013

[67] Jingjing Liu, Panupong Pasupat, Yining Wang, Scott Cyphers, and Jim Glass, “Query Understanding Enhanced by Hierarchical Parsing Structures,” ASRU 2013.

[68] Wang Hongling and Zhou Guodong, “Semantic role labeling of Chinese nominal predicates with dependency-driven constituent parse tree structure,” Journal of Computer Science and Technology. 28(6): 1117-1126, 2013.

[69] Li Shoushan, Wang Rongyang and Zhou Guodong, “Opinion target extraction via shallow semantic parsing,” AAAI 2012.

[70] Chen Keh-Jiann, Yu-Ming Hsieh, “Chinese Treebanks and Grammar Extraction,” Proceedings of IJCNLP-04, pp560-565, 2004.

[71] Roger Levy and Christopher D. Manning, “Is it harder to parse Chinese, or the Chinese Treebank?,” *ACL 2003*, pp. 439-446.

[72] Hopcroft, John E.; Ullman, Jeffrey D., “Introduction to Automata Theory, Languages, and Computation,” Addison-Wesley. Chapter 4: Context-Free Grammars, pp. 77–106; Chapter 6: Properties of Context-Free Languages, pp. 125–137, 1979.

[73] J. Berstel, L. Boasson. Jan van Leeuwen, ed. “Context-Free Languages,” Handbook of Theoretical Computer Science B. Elsevier. pp. 59–102, 1990.

[74] You, Jia-Ming, Keh-Jiann Chen, 2004, “Automatic Semantic Role Assignment for a Tree Structure,” Proceedings of SIGHAN workshop.

[75] http://ckipsvr.iis.sinica.edu.tw/

[76] Tsai Yu-Fang and Keh-Jiann Chen, "Reliable and Cost-Effective Pos-Tagging," International Journal of Computational Linguistics & Chinese Language Processing, Vol. 9 #1, pp83-96, 2004.

[77] http://parser.iis.sinica.edu.tw/

[78] DM Blei, AY Ng, MI Jordan, “Latent dirichlet allocation,” the Journal of machine Learning research 3, 993-1022, 2003.

[79] Gregor Heinrich, “Parameter estimation for text analysis.”, Technical report,

2004.

[80] Ian Porteous , Arthur Asuncion , David Newman , Padhraic Smyth , Alexander Ihler , Max Welling, “Fast collapsed Gibbs sampling for latent Dirichlet allocation,” In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.

[81] Yee Whye Teh, David Newman, and Max Welling, “A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation.”, In NIPS, 2006.

[82] Casella, George and George, Edward I., "Explaining the Gibbs Sampler," The American Statistician 46, no. 3: 167—174, 1992.

[83] http://www.sogou.com/

[84] http://zh.wikipedia.org/wiki/

[85] Su-Chu Lin, Shu-Ling Huang, You-Shan Chung and Keh-Jiann Chen, “The Lexical Knowledge and semantic representation of E-HowNet,” Contemporary Linguistics, Vol.15, No.2, pp. 177-194, 2013.

[86] Wei-Te Chen, Su-Chu Lin, Shu-Ling Huang, You-Shan Chung, and Keh-Jiann Chen, “E-HowNet and Automatic Construction of a Lexical Ontology,” the 23rd International Conference on Computational Linguistics, Beijng, China, 2010.

[87] R. A. Baeza-Yates and B. A, “Ribeiro-Neto. Modern Information Retrieval,” Addison Wesley, 1999.

[88] Jenny Rose Finkel, Trond Grenager, and Christopher Manning, “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling,” *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, pp. 363-370, 2005.

[89] E.M. Voorhees, "Proceedings of the 8th Text Retrieval Conference," TREC-8 Question Answering Track Report. pp. 77–82, 1999.

[90] D. R. Radev, H. Qi, H. Wu, W. Fan, "Evaluating web-based question answering systems.", Proceedings of LREC, 2002.