第 1 章 緒論

1.1 研究動機與目的

隨著資訊爆炸時代的來臨，人們希望以更高的效率與效能取得資訊，其中自動摘要技術與其後衍生的分類應用，是重要的關鍵技術之一。例如，Google [Google] 利用網頁片段權充摘要、報章雜誌的標題目錄等，這些眾多且未經分析整理的資訊，經過擷取、分析與整理後便成為高質量的資訊，給人們有效的閱讀與吸收。

本論文旨在探討自動摘要模型及分類器模型。摘要後的文件可視為一種特徵選取的前處理，透過前處理可以將重要的資訊摘選出來並減少分類時的運算量，協助分類器做更精確的分類。

1.2 自動摘要

自動摘要技術的目標是依據使用者的需求，將文件縮減濃縮成一或數句重要字句，好讓人們更快速、更方便的得到所需資訊，其優點在於：

節省時間： 使用者不需瀏覽整篇文件即可瞭解文意

加速瀏覽： 在查詢結果中呈現摘要，可方便使用者快速決定所需資訊

協助分類： 摘要過的資訊，可做為分類器的分類依據

節省人力： 自動摘要的產生，不需透過人力介入

自動摘要可以分類如下 [Hovy and Marcu 1998]：

1. 根據形成方式可分類為摘錄式（Extractive）摘要與非摘錄式（Non-extractive or Abstract）摘要。摘錄式摘要是找出文件中重要的字句、段落或章節來組成摘要；非摘錄式摘要則重寫字詞、片語來形成摘要。
2. 根據性質可分類為資訊性（Informative）摘要與指示性（Indicative）摘要。資訊性摘要是從文件中找出所有重要的資訊，科技論文的摘要即為一例；而指示性則偏向於提供文件分類上的資訊，例如圖書館內所用的分類卡。
3. 根據需求可分類為一般性（Generic）摘要與以需求為基礎（Query-based）摘要。

一般性摘要對文件內不同主題視為同等重要；以需求為基礎的摘要則傾向於顯示使用者要求的部份。

4. 根據文件來源可分類為單一文件(Single Document)摘要與多文件(Multidocument)摘要。單一文件摘要是從一篇文件中截取重要資訊；多文件摘要則歸納主題相近的文件共同產生摘要，或指對同一主題但時間先後不同文件進行摘要。
5. 根據語言可分類為單一語言（Monolingual）摘要與多語言（Multilingual）摘要。多語言摘要係從多種語言的文件中產生單一語言的摘要結果，其中牽涉到機器翻譯的技術。

大多數常見的摘要模型原則上可依據其特性分為兩種比對策略。其一，以逐字比對（Literal Term Matching）的方式評估字句與文件的相關性，愈高相關性的字句代表愈重要，這其中以向量空間模型（Vector Space Model, VSM）為代表 [Gong and Liu 2001; 何遠 2003]；其二，以概念比對（Concept Matching）的方式評估，這其中以潛藏語意分析（Latent Semantic Analysis, LSA）為代表 [Gong and Liu 2001; 葉鎮源 2002; 黃建霖 2004; Hirohata *et al.* 2005] 。

本論文針對的是摘錄式、資訊性、一般性、單一文件、單一語言摘要模型做研究，並從逐字比對與概念比對兩個方向作探討，希望能發展出適合的自動摘要模型以供中文自動摘要的產生。

1.3 研究成果

本論文於自動摘要方面，在逐字比對方式上應用隱藏式馬可夫模型（Hidden Markov Model, HMM）做為摘要模型，並分為 HMM-Type1 及 HMM-Type2 二種類型；在概念比對上提出嵌入式潛藏語意分析（embedded LSA）與主題混合模型（Topical Mixture Model, TMM）做為摘要模型；在自動摘要評估上，提出以改良型字錯誤率（modified Character Error Rate, m-CER）為基礎的平均精確度（Mean Average Precision, MAP）評估方式，以解決自動轉寫與人工轉寫文件因斷句不一致，所造成摘要結果無法評估相關的問題。

經由實驗結果顯示，於摘要模型比較上：使用隱藏式馬可夫模型或主題混合模型其結果較其它常見方法有顯著的提升，同時主題混合模型在幾乎所有情況下均較隱藏式馬可夫模型來得佳；於特徵單位比較上：使用雙音節與雙字時，其結果優於使用詞為特徵單位。

最後，我們也研究摘要模型中主題混合模型在文件分類的適用性，並且文件也能預先經由上述摘要模型做前處理。初步實驗結果顯示，主題混合模型分類器較常見 *K*-最近鄰（*K*-Nearest-Neighbor, *K*NN）分類器在分類的效果上有些微的提升。

1.4 章節安排

本篇論文的章節安排如下：

第二章簡介本論文的理論背景，包括向量空間模型、相關評估、潛藏語意分析、馬可夫模型、隱藏式馬可夫模型、統計式語言模型與主題混合模型。

第三章介紹本論文所提出的摘要模型，包括嵌入式潛藏語意分析、隱藏式馬可夫模型-型一、隱藏式馬可夫模型-型二、主題混合模型。

第四章說明本論文的實驗設定，並利用餘弦、ROUGE、平均精準度三種自動摘要評估方法做實驗，對實驗結果做一分析。

第五章概述分類器模型與提出主題混合模型分類器，並介紹實驗語料。在實驗結果上，比較主題混合模型分類器和常見 *K*-最近鄰分類器的實驗結果，並分析自動摘要是否有助於分類器做更精準的分類。

第六章對本論文的主要成果做一總結，並提出結論與未來研究方向。