新聞文件摘要之研究

李祥賓　柯淑津

東吳大學資訊科學系

{ms8805, ksj@cis.scu.edu.tw}

摘　要

本文主要以兩種摘要技巧對路透社新聞文件進行自動文件摘要處理，分別為由語句出現的位置來判斷其重要性，以及擴充標題詞彙兩種摘要技巧。我們對文件進行分析探討，找出文件主題通常是佔據了哪些位置，以擷取這些重要位置的句子為摘要。另外，我們認為標題對於文件是相當重要的，因此我們透過 WordNet 找尋標題的相關詞彙，對標題詞彙進行擴充，來找出更多與標題相關的字，增加標題的重要性，進而協助在文件中找尋與標題較相關的摘要語句。在實驗評估方面，我們提出一個以分類方式進行自動摘要評估的想法，並建立了一個分類系統來評估摘要結果。最後，本文提出了一種綜合擴充標題詞彙與重要位置的摘要方法，此方法得到 71.9%分類精確率的實驗結果，相較於 65.6%的基準分類精確率改善了 9.6% ((71.9%-65.6%) / 65.6% = 9.6%)。

1.　簡介

在資訊科技發達的今日，文件已由傳統的書面呈現方式，轉化成數位方式包裝。這些資訊再藉由無遠弗屆的網際網路傳播到世界上各個角落，因此使用者可以輕易地透過網路獲得所需求的資訊。

資訊流通的便利性，雖然帶來了豐富的資源，但同時，也引進了另一個問題：「資訊氾濫」。網路使用者可能都有過這樣的經驗：當瀏覽線上文件時，發現過多的文件讓使用者無法一一詳盡閱讀全文，而只觀看文件的標題又無法掌握到文件的內容，進而判斷出此文件是否符合自己需求。目前新聞網站的線上新聞就是一個相當好的例

子。如果這些網路新聞在總覽時，能適切地提供精簡摘要來協助使用者選擇自己所需求的文件。將有助於降低網路傳輸量，進而提升網路服務品質。

這類的文件摘要工作最先是由專業人員以人工方式來進行，雖然人工摘要的品質相當好，但遇到大量且更新快速的網路新聞，用這樣的方法就顯得緩不濟急。然而，自動文件摘要的技術正是解決這個難題的利器。自動文件摘要技術是擷取文章內重要的訊息出來，經過組合產生較短的摘要，讓使用者可快速地明白這篇文章的主旨，藉此節省使用者的閱讀時間，而能較快地判斷該篇文章是否為自己想要的文件。

過去文獻中，已有許多相關於自動文件摘要的研究，而本研究將針對下列兩種摘要策略進行研究與探討：由位置來判斷文件主題與擴充標題詞彙。並設計實驗來驗證這些摘要策略是否可擷取出品質良好的摘要內容。另外，對於文件摘要的成效評估，本研究提出了一個新的想法，以應用系統來評估摘要結果。我們將摘要結果取代原本文件，進行分類處理。再以分類結果來驗證摘要成效。假若，我們的摘要確實能由文件中擷取重要資訊，相較於用全文來進行分類，在分類效果上應該會不相上下或是有更好的精確率。

本篇文章共分為六節，第二節針對過去有關於文件摘要研究的文獻進行回顧。第三節介紹本文所使用的研究資源，包括路透社新聞語料與含標記詞義訊息的布朗語料庫。第四節主要探討本研究所使用的兩種摘要策略：由位置來判斷文件主題以及擴充標題詞彙。另外，在此節中，我們提出一個以分類系統來評估摘要成效的想法。第五節針對本文提出的摘要方法設計一系列實驗，以路透社新聞語料進行摘要處理，並將摘要結果送交分類系統，再對分類結果進行討論。最後，我們在第六節提出結論以及未來研究方向。

2. 相關研究

過去對於文件摘要的研究，多以單一文本為對象，也就是只針對一篇文章內容來進行摘要處理，應用不同的技巧，來表現出文件中的重要資訊。其中，有些研究透過計算

各詞彙在文件中所擁有的權重(tf × idf)，藉此權重值可找出文件中較具重要性的詞彙，進而擷取出含有重要詞彙的語句來形成摘要(Forsyth and Rada, 1986)。有的研究則是針對文章進行資訊擷取(information extraction)處理，找出文章內的人名、地名、組織名等專有名詞，再對這些專有名詞與新聞內文設定不同的權重，進而擷取出文件中的重要語句(邱中人, 2000)。有些研究則考慮語句在文件中的位置，認為出現在某些特定位置的句子，常常較具重要性，可以直接擷取出來當作摘要(Hovy and Lin, 1997)。

而相對於單一文本的研究，有些學者致力於多文本的摘要研究，他們對報導同一事件的多篇新聞文稿歸納它們的相似處，以及辨識出彼此相異的地方，做成該事件的摘要(McKeown and Radev, 1995; Barzilay, McKeown and Elhadad, 1999; Chen and Huang, 1999)。多文本摘要研究常用的技術，除了上面所談到的慣用於單一文本摘要處理的方法外，McKeown 等人則是為相關於恐怖份子的新聞設定了樣版(template)，樣版用於擷取多篇文章內的資訊，如報導來源、報導日期、事件發生日期、事件發生的情形等等詳細的資訊；這些擷取出的資訊，經判斷與比對處理後，會送至摘要製作模組以產生出一個多文本的摘要(McKeown and Radev, 1995)。

多數的摘要研究在針對文件的重要內容做相關統計時，大部分是以字詞作為處理單位。這些方法大多都是找尋文章內重要的字詞，再依這些字詞來進行進一步的處理，找出重要語句來形成摘要。有的則利用自然語言處理技術，如：片語、暗示字 (cue word)、上下文 (discourse) 處理等來協助辨認文件中的重要語句 (Marcu, 1999)。這些研究認為某些特定片語後面接的句子有某程度的重要性，如：”結論…”，因此包含這些特定片語的句子會依照其上下文關係，經過處理後，被擷取出來形成摘要。

但是，因為字詞的參數空間大，且存在一詞多義與同義字等問題；因此，有些學者認為應該跳脫字詞層次，以字詞所蘊藏的概念或語意來取代字詞本身，彙整成文件的概念主題，進而找出表達文件的重要概念。在 1999 年，Hovy 等人使用 WordNet 同義辭典來做概念之間的相關性聯結(Hovy and Lin, 1999)。Woods 使用片語的分析來組織字詞成為一個概念性的架構(Woods, 1997)。但這類的方法往往需要有強大的語言

知識做為輔助，而這種資源通常是相當不容易獲得的。

3. 研究資源

本文在研究過程中，使用「WordNet」、「路透社新聞語料」與「含標記詞義訊息的布朗語料庫」等資源來進行摘要實驗。對於 WordNet 的介紹，請參見 Miller 等人的文章(Miller et al., 1993; Fellbaum, 1998)。在下面，我們將對後兩項資源逐一介紹。

3-1 路透社新聞語料

路透社新聞語料庫(Reuter Corpus)，由 Lewis 在 1992 年所收集，目前此語料庫是文件分類研究中最常使用的語料庫之一，其內容取自 90 年間的路透社新聞文章，總共含超過兩萬篇標註分類類別的文件(Hayes and Weinstein, 1990; Lewis and Ringuette, 1994)，而從最初的版本演變到現在共有五個版本，各版本的差異在 Yang 的論文中有很詳盡的比較(Yang, 1999)。

本文選擇版本 3 的語料做為我們進行摘要處理時的實驗語料。此版本的路透社新聞語料包含 7789 篇訓練文件與 3309 篇測試文件，共分為 93 種類別。新聞語料中多數的文件被歸類到一種類別，但被歸類在二種類別以上的文件也不在少數。經過統計後，我們得知在訓練語料與測試語料裡，每篇文件的平均類別數分別為 1.23 和 1.24。

在路透社新聞語料中，每篇文件的內容長短不一，有些文件可能只有幾個句子，有些卻可達到十多句以上。經過統計後，在訓練語料與測試語料中，每篇文件的平均句長分別為 6.8 句與 7.3 句；而從整個語料來看，每篇文件的標題與內文平均長度分別為 7.4 和 126.9 個詞彙。

3-2 含標記詞義訊息的布朗語料庫

布朗語料庫(Brown Corpus)是在資訊檢索相關研究上，常被使用的語料庫之一。本文所使用的布朗語料庫，主要詞彙已標記出其所屬的 WordNet 詞義。這個布朗語料庫是由普林斯頓大學認知科學實驗室(Cognitive Science Laboratory)，利用 WordNet 內的詞義架構，以人工方式為此語料內的詞彙進行詞義標記，此語料庫可由

http://www.cogsci.princeton.edu/~wn/ 網址獲得，主要分為 brown1, brown2, brownv 三個部分，詳細的資料列於表 1。

表 1：具標記詞義的布朗語料庫---詳細資訊。

名稱	檔案個數	詞義標記範圍	已標記詞義的詞彙
brown1	103	名詞、動詞、形容詞與副詞	106,725
brown2	83	名詞、動詞、形容詞與副詞	86,414
brownv	166	動詞	41,525
Total	352		234,664

從表 1 內容，可看出這個布朗語料庫共有 352 個檔案。因 WordNet 只包含名詞、動詞、形容詞與副詞這四種詞性，因此，只能針對這四種詞性的詞彙進行詞義標記工作。在整個語料庫中，總共有 234,664 個詞彙被標記上 WordNet 詞義。我們將利用這個布朗語料庫，所提供的詞彙與詞義資訊，來協助本研究進行摘要處理。

4. 研究方法探討

在本節，我們將研究方法區分為摘要與評估兩部分進行討論。在 4-1 至 4-2 小節中，我們說明本文在進行摘要研究時所使用的兩種方法。另外，本研究提出一個以分類方式來評估摘要成效的想法，並針對我們所使用的分類系統，在 4-3 小節中進行詳細的說明。

4-1 由位置來判斷文件主題

在一般文件的內容架構上，文件的主題句通常會佔據著某些特定的位置。因此本研究希望能透過考量位置的重要性來找出文件的主題所在。Edmundson 認為包含標題字的語句，會與文件主題較為相關(Edmundson, 1969)。另外，最重要的語句通常會出現在文件較前面或是較後面的位置。此外，Hovy 等人也認為文件主題常佔據特定位置，他們曾針對電腦相關的新聞文件進行研究，提議將標題字與簡介內的字視為文件的重要詞彙，並應用這些重要詞彙來統計分析出文件的重要位置(Hovy and Lin, 1997)。

本研究將針對訓練語料來進行文件內重要語句位置判定。而重要位置判定的方法運用了 Edmundson 與 Hovy 等人的看法，利用文件的關鍵字來統計分析在文件內的哪一個位置，其語句的內容通常具有較高重要性。

在關鍵字的選取方面，我們運用傳統資訊檢索所常用的 tf × idf 技巧，來找出文件裡的重要詞彙，再從重要詞彙中抽取出一定比例的詞彙來做為關鍵字，而從文件中找出重要詞彙的方法，如公式 1、2 所示。假設一個文件 d，以及出現在 d 中的詞彙 t，我們定義 t 在 d 中所具有的權重值，$W(t, d)$，為 t 在 d 中的出現次數 $tf_{t, d}$ 乘以詞彙 t 本身的重要性 idf_t，再除以 d 中詞彙出現的最高詞頻。接著，我們將權重值較高的詞彙視為文件 d 的重要詞彙。

$$W(t, d) = \frac{tf_{t, d} \times idf_t}{\underset{t}{Max}\ (tf_{t, d})}\ , \tag{1}$$

$$idf_t = \log\ (\frac{T}{df_t} + 1)\ , \tag{2}$$

$tf_{t, d}$ ：文章 d 中此詞彙 t 的字頻，

df_t ：詞彙 t 出現在訓練語料中的文章數，

T ：訓練語料中所有文章總數。

接著，使用這些關鍵字來進行重要位置的判定。我們假設這些關鍵字的重要性比文件內其餘的詞彙高，而且一個重要的語句應該會包含較多的關鍵字。

我們針對給定的一個文件 d，如果詞彙 t 屬於文件 d 的關鍵字，則將擁有權重值 $W(t, d)$。之後，再以句子為單位，計算出文件內句子 s 所擁有的權重值 $Score(s, d)$，計算句子權重值的方法如同公式 3 所示。

$$Score(s, d) = \sum_{t \in s \cap 關鍵字_d} W(t, d)\ , \tag{3}$$

t ：出現在文件 d 中的詞彙。

當我們求出文件中每一個句子的權重值後，依句子的權重值高低就可求得文件中重要位置的排行。我們為訓練語料內的所有文件進行上述的處理，就可以知道每一篇

文件中重要句子的位置分佈。最後，我們統計出每個位置在文件內所擁有的重要性排行。進而利用此結果擷取出重要位置所在的句子以形成摘要。綜合上述，我們將此摘要方法的處理過程加以整理，列出於圖 1。

重要位置摘要方法的演算法

訓練部分

步驟一：利用公式 1 與 2，找出訓練語料內各文件的重要詞彙。

步驟二：將各文件的重要詞彙與標題字結合，成為該文件的關鍵字。

步驟三：利用公式 3，計算出各文件內句子的權重值，並依照權重值來決定各文件內位置的重要性排行。

步驟四：對步驟三整理出的資訊，進行統計與分析，以決定在此訓練語料中重要位置的所在。

測試部分

步驟一：針對文件 d 擷取重要位置所在的句子來形成摘要。

圖 1：重要位置摘要方法的演算法。

4-2 擴充標題詞彙

標題通常與文件主題具有較高的相關程度，我們認為文件中的句子若包含較多標題詞彙，則此句子與文件主題會具有較高的相關程度，因此可擷取作為摘要的內容。但是，標題內容往往因過於精簡，使得我們無法從中獲得足夠的資訊。而且在一篇文章中，作者可能會使用不同的詞彙來表達同一事物或動作，這樣的寫作風格雖增加了詞彙的多樣性，卻使得我們在進行重要詞彙比對過程中，比對到相同詞彙的機率大為降低。對於這樣的現象，我們認為文章作者雖然可能使用多樣性的詞彙來表達同一主題，不過這類相互可替代的詞彙在語意上應該具有高度相關意義。因此，我們希望找出文章作者有可能使用的相關詞彙。

為了找出與標題內容具相關語意的詞彙，我們藉由 WordNet 的豐富語意網絡，利用關聯性指標找出一個詞彙的同義詞與上義詞。另外，我們認為在解釋詞彙詞義的定義中所使用到的詞彙，應該會與此詞彙具有一定的相關程度。因此，我們利用

WordNet 來擴充標題詞彙，透過收集與標題詞彙較為相關的同義詞、上義詞與定義內的詞彙，並將這些詞彙視為標題詞彙，來解決文件內容使用多樣性詞彙所帶來的影響。

在 WordNet 內，有些一詞多義的詞彙會擁有多個詞義。因此在擴充詞彙之前，我們必須知道詞彙的正確詞義，也就是說我們需要先解決詞義歧異的問題。本研究使用了兩個方法來針對標題詞彙進行詞義辨識，分別是以 WordNet 的定義與文件詞彙的重疊性來進行詞義歧異辨識，與利用語料庫詞義出現機率進行詞義歧異辨識。

4-2-1 以 WordNet 進行詞義歧異辨識

這個方法是假設在標題中一個詞彙若具有某項詞義，此詞義所衍生出來的相關詞彙，在文件中出現的機率應該會高過於其它詞義的衍生詞彙。於是我們針對標題內的名詞、動詞與形容詞，使用 WordNet 來找出每個詞彙在不同詞義下所擁有的同義詞與上義詞，以及各詞義在定義內所使用的詞彙，再將三者聯集成一個詞彙集合。接著，將集合內的詞彙與文件內容進行比對，我們把出現重覆性最高的集合所代表的詞義，設定為此詞彙的詞義。圖 2 表示此方法對於詞彙所形成的集合，圖中的詞彙有 n 種詞義，而這 n 種詞義分別有各自的同義詞、上義詞與定義內的詞彙，因此會有 n 個詞彙集合。

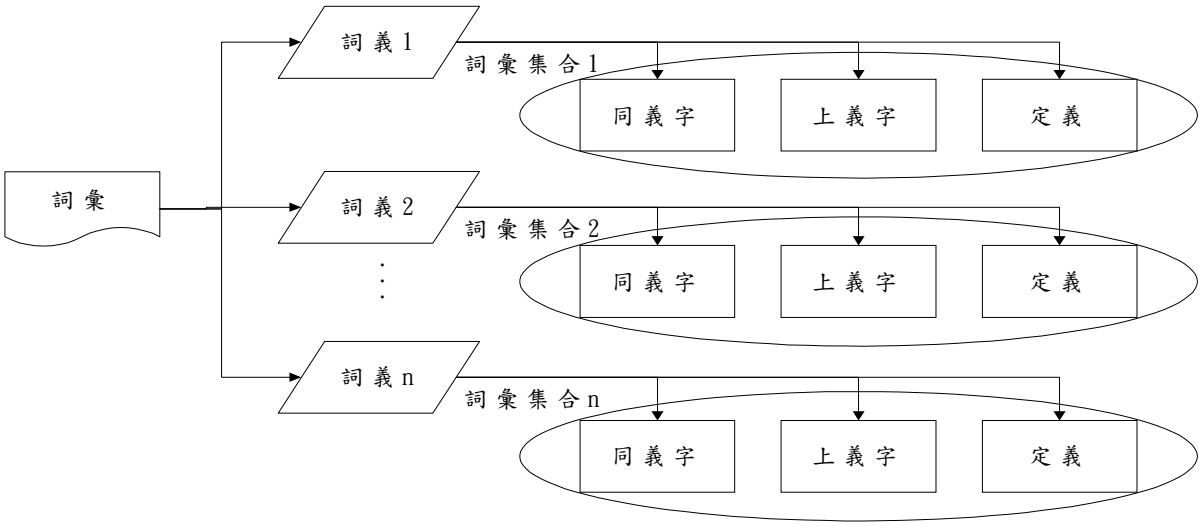

圖 2：詞彙在 WordNet 中所擁有的各種詞義與其相關詞彙集合。

當我們利用 WordNet 對標題詞彙作詞義辨識時，一個詞彙下的所有詞彙集合有可能在文件內出現次數都一樣，使得無法藉著出現次數多寡來判斷正確詞義。此時，

我們便借重第二種詞義辨識方法，利用語料庫詞義出現機率進行詞義歧異辨識。

4-2-2　利用語料庫詞義出現機率進行詞義歧異辨識

第二個詞義辨識方法則是給予一個詞彙在語料庫中最常出現的詞義。本文所使用的語料是已標記詞義的布朗語料庫(Brown Corpus)，本文在 3-2 小節中對此語料庫有詳細的說明。此語料庫內的詞彙使用了 WordNet 來進行詞義標記，我們對這些詞彙進行統計，列出詞彙在不同詞義下出現的機率。因此在辨識詞義的過程中，我們直接挑選出現頻率最高的詞義，來作為這個詞彙出現在標題時所具有的詞義。藉由這樣的處理方式，我們便可輔助以 WordNet 辨識詞義的不足，而提高成功辨識率。

利用上述兩種方法進行詞義辨識後，就可以針對標題詞彙的詞義來擴充其相關的詞彙。詞彙擴充的結果，就是將標題詞彙的上義詞、同義詞與定義內的相同詞性詞彙，擷取出來形成此標題詞彙的相關字。我們將標題詞彙與擴充後的詞彙視為重要詞彙，接著，計算每個句子所擁有的重要詞彙數，作為句子的分數，然後再挑出較高分數的句子來作為摘要。擴充標題詞彙摘要方法的演算法，我們整理於圖 3。

擴充標題詞彙摘要方法的演算法

步驟一：針對文件 *d* 的標題詞彙 *t*，使用 WordNet 找出其詞義的詞彙集合。

步驟二：將詞彙 *t* 的所有詞彙集合與文件 *d* 內的詞彙進行比對，計算各集合的出現次數。

步驟三：計算各詞彙的最高頻詞義與次高頻詞義之比例值 *r*，若 *r* 大於某預先設定的門檻值 *h*，則設定詞彙 *t* 的詞義為出現次數最高的集合所代表的詞義；否則，直接給予語料庫中出現頻率最高的詞義。

步驟四：利用辨識出的詞義，來擴充標題詞彙，收錄該詞義的同義字、上義字與定義內同詞性的詞彙。

步驟五：結合標題詞彙與擴充後的詞彙，作為文件的重要詞彙。

步驟六：計算文件 *d* 內句子 *s* 所擁有的重要詞彙數，挑選出擁有較多重要詞彙的句子作為摘要內容。

圖 3：擴充標題詞彙摘要方法的演算法。

4-3 評估方法

自動摘要的評估是一件相當困難且主觀的工作。一般而言，評估文件摘要成效的方法可分為兩種作法。第一種作法是由公正的第三者對摘要內容進行判斷，決定摘要內容是否恰當，或是涵蓋的資訊是否充足。第二種作法則是將摘要內容應用於其他不同的工作上，觀察此工作的表現來決定摘要的成效(Mani and Bloedorn, 1999)。

本研究提出一個評估的想法，此想法較偏向於 Mani 等人所提出的第二種摘要評估作法。我們針對英文文件所進行的摘要實驗，實驗資料為路透社的新聞語料，這些新聞語料內的文件均已標註其所屬的分類類別。我們認為如果本研究的摘要成果是具有正向意義的，那麼經過摘要處理的文件，因為其摘要內容是由文件中較重要的句子所組成，這些句子所表達的資訊應該足夠代表該文件。因此若將此摘要結果送由分類系統來進行分類處理，相較於用全文來進行分類，在分類效果上應該會不相上下或是有更好的精確率。為因應這個想法，我們需要一個分類系統來協助我們進行摘要的評估工作。

本研究根據 Ker 和 Chen (2000)所提出的分類方法中，挑選了其中一個方法來針對文件全文建立分類系統。首先，我們將語料中的詞彙全部轉成小寫字母，並刪除停用字，再把這些留下來的詞彙予以原形化(stemming)處理，最後所得的詞彙就是特徵字(feature)。接著，我們針對訓練語料中每個類別 c 所擁有的文件，彙總其特徵字 f 來進行統計處理，給予它們應有的權重值 $W(f,c)$，而計算權重值的方式如公式 4, 5 所示。

$$W(f,c) = tf_{\mathrm{f,c}} \times idf_{\mathrm{f}} \text{，} \tag{4}$$

$$idf_{\mathrm{f}} = \log_2\left(\frac{T}{df_{\mathrm{f}}}\right) \text{，} \tag{5}$$

$tf_{\mathrm{f,c}}$ ：特徵字 f 出現在類別 c 的頻率，

T ：類別的總數，

df_{f} ：特徵字 f 出現過的類別總數。

給予一篇測試語料的文件 d，我們在決定所屬的類別時，必須針對文件內所含的各類別特徵字進行權重值的加總，以得到文件在不同類別下的總權重值 $R(c,d)$。接著，統計各類別的總權重值，將文件歸屬於總權重值最高的類別 $Class(d)$，計算方式如公式 6, 7 所示。

$$R(c,d)=\sum_{f\in F_c} tf_{f,d}\times W(f,c)\ ,\qquad (6)$$

$$Class(d)=\arg\max_{c} R(c,d)\ ,\qquad (7)$$

$tf_{f,d}$ ：特徵字 f 在文件 d 出現的頻率，

F_c ：類別 c 的特徵字集合。

由上述的方法，我們建立了一個分類系統。我們將未經過摘要處理的路透社新聞測試語料由分類系統處理，得到精確率為 65.6%的分類效果，我們以此精確率來當作基準精確率，來做為摘要成效比較的基準。

5. 文件摘要實驗

為驗證本研究所提出的文件摘要方法之效果，我們以路透社的新聞語料(版本 3)為實驗資料設計了一系列的實驗，來觀察各摘要方法的成效。首先，我們先介紹資料的前置處理。接著，將分別介紹各種摘要實驗方法，包含實驗設計與實驗結果等，最後對實驗結果進行討論。

5-1 前置處理

在進行摘要處理之前，我們對於實驗資料的前置處理主要有下列幾個步驟：首先必須去除停用字(stopword)，以減少不具重要意義的詞彙。接著是原形化的處理，這步驟主要考量英文詞彙具有不同詞性的轉換。如 relate、relation、relative，這三個詞彙雖具有相似的意義，但在統計處理上會視為不同的詞彙；而原形化能將這類的詞彙變成較短的字根，如 relate，因此可彙整此類詞彙在文件中所佔有的重要性。

5-2 由位置來判斷文件主題

我們運用在 4-1 小節中提出的方法來分析文件的重要位置所在。針對所有訓練語料中

的文件皆挑出五個重要句子，再依照各位置的重要性與出現頻率來進行整理，進而分析重要位置所在。另外，由於路透社新聞語料的文件平均長度約為七個句子，因此我們認為摘要的長度約在三個句子左右較為合適。所以，我們只找出文件內的三個重要位置。我們從統計數據中，找出了三個出現次數較高的位置。按出現次數由高而低地排列下來，分別為第二段第一句、第一段第一句與第三段第一句（簡稱為 P_2S_1，P_1S_1，P_3S_1），如表 2 所示。

表 2：P_1S_1、P_2S_1 與 P_3S_1 出現在文件內的重要性與次數。

句子位置\重要性	一	二	三	四	五	總計
P_2S_1	1584	2227	1303	626	355	6095
P_1S_1	3142	1338	617	412	364	5873
P_3S_1	385	975	1692	1012	631	4695

接著，從表 2 內容，我們發現在總數 7789 篇的訓練語料中，第一段第一句出現在文件內最重要句子的情形就多達 3142 次，依此可明顯判斷出此位置是此新聞語料中的最重要位置。至於第二段第一句與第三段第一句，這兩個位置的句子出現在文件內次重要句子的次數分別為 2227 次與 975 次，明顯地區別了兩者的重要程度，因此次重要位置應屬於第二段第一句。藉由對這三個位置的分析，我們得知在路透社新聞語料的訓練文件中，重要句子出現的位置依排行分別為第一段第一句、第二段第一句以及第三段第一句。

另外，我們由每篇文件提出前五重要句子，總共得到 30800 多個句子；但由於某些文章本身並不足五句，導致我們提出的句子數目比預計的數目少了許多。然而，除了我們上述討論的三個位置總共佔了 16663 個句子外，其餘位置總共出現了約 14200 次，但是這些位置的出現次數都不高，因此我們不對這些位置做進一步的處理。

5-2-1 實驗設計

本研究設計了四組實驗來驗證我們經由統計分析後，所得出的重要位置對於摘要是否

有其正向意義。這些實驗的主要考量為依照不同位置的重要性，分別擷取出不同句數的摘要。

此外，我們考慮在路透社新聞語料中每個類別所擁有文件的數量多寡不一，文章長度也不太一樣（如表 3 所示，類別 acq 文件的平均句長為 5.3 句，而類別 money-fx 文件的平均句長則為 8.9 句）。在這樣的情況下，重要位置分析結果是否適用於不同的分類，我們無法確切得知。於是，本研究針對文件數量較多的三個類別 earn、acq 以及 money-fix，進行相同的實驗，希望能得知在不同類別的文件內，重要句子出現的位置是否一致。綜合上述考量，我們所設定的四組實驗如表 3 所示，其中，實驗一以整個測試語料進行實驗，而其他三組實驗則各自針對不同類別進行相同處理。

表 3：重要位置摘要方法實驗設定。

組別	受測文件	平均文件長度(句)	文件數量(篇)
實驗一	整個測試語料	7.3	3309
實驗二	類別 earn	7.4	1176
實驗三	類別 acq	5.3	776
實驗四	類別 money-fx	8.9	207
註：四組實驗的位置選取設定皆是相同的，分別以$\{P_1S_1\}$、$\{P_1S_1, P_2S_1\}$、$\{P_1S_1, P_2S_1, P_3S_1\}$為摘要內容。			

5-2-2 實驗結果與討論

本研究針對重要位置摘要方法所設計的四組實驗，經過分類系統處理後，所得到的精確率如表 4 所示。由於製作出的摘要包含的句子數量不同，使得摘要佔全文的長度比例也不同。我們對此列出不同實驗設定所製作出的摘要，其內容所佔全文的長度比例，來評估資訊量減少的情形。

從表 4 中，我們可以觀察到實驗二與實驗三所得到的數據，皆高於其它兩組。對此，我們認為這樣的情形是由於製作分類系統時，我們所使用的方法對於文件數量較多的類別會產生較好的分類辨識效果，因此對於數量較多的類別 earn 與 acq 而言，

其分類成效會較高。

表 4：重要位置摘要方法的實驗結果。

重要位置選取	實驗一		實驗二		實驗三		實驗四	
	精確率	長度	精確率	長度	精確率	長度	精確率	長度
P_1S_1	55.4%	13.8%	70.2%	13.5%	76.8%	19.0%	46.4%	11.3%
P_1S_1 與 P_2S_1	66.9%	26.1%	94.4%	26.4%	77.7%	35.4%	65.7%	20.2%
P_1S_1、P_2S_1 與 P_3S_1	65.9%	37.2%	95.2%	39.2%	73.2%	48.1%	67.1%	28.4%
全文	65.6%	100.0%	95.8%	100.0%	76.7%	100.0%	68.1%	100.0%

從實驗一、二與四的結果來看，只選取最重要位置的句子為摘要，此摘要佔全文的長度比例未達三分之一，在這樣的摘要長度下，摘要所能包含的資訊並不充足；因此，得到的精確率與全文的分類成效相比，均有相當大的差距。當摘要內容增加第二與第三重要位置後，摘要長度可達到全文的四分之一至三分之一；這樣的摘要長度可擷取出較充足的資訊，再加上摘要內容是由重要位置的句子所組成；因此，此摘要所得到的分類成效已經接近各組實驗的全文分類精確率。

另外，我們觀察類別 acq 所進行的實驗，發現其摘要長度為全文的三分之一至五分之一時，這樣的摘要所帶來的資訊量是比較充足的，因此實驗成效已經突破了基準精確率。但是增加擷取第三重要位置後，摘要的長度比例增為 48.1%，已相當接近全文的一半；但是由分類精確率反之下降了 4%來看，顯示了過長的摘要並不一定會帶來更多重要的資訊。

除此之外，本研究進行了另一項實驗，我們從測試語料各文件中隨機選取了三分之一的句子來形成摘要，並進行分類評估，得到了 55.7%的分類精確率。此精確率與實驗一的成效相比較，我們可以看出文件內的確有特定重要位置的存在，才會使得由重要位置所形成的摘要在分類成效上比起隨機選取的方式改善了 20.2%左右。

從上述這些實驗中，我們認為對於不同來源的文件，只要經過統計分析，就可以得知其重要位置所在，進而藉此資訊來找出重要句子。而從實驗結果中，我們認為一

個摘要的產生，它的內容若可以包含前兩重要的位置—第一、二段的第一句，在摘要長度與重要性上是較好的考量。

5-3 擴充標題詞彙

擴充標題詞彙主要是以標題內的名詞、動詞與形容詞為對象進行處理，因此我們必須得知詞彙的詞性，才可以擷取這三種詞性的詞彙來進行詞彙擴充。本文使用由麻省理工學院的 SLS(Spoken Language Systems Group)在 1993 年所發表的詞性標記工具(http://www.sls.lcs.mit.edu/sls)，是一種以規則為本(rule-based)的標記詞性方法。

接著，針對標題內的名詞、動詞與形容詞，我們使用了在 4-2-1 與 4-2-2 小節所提出的兩個方法來進行詞義辨識。並藉由辨識出的詞義，來為標題詞彙進行擴充，納入與標題詞彙相關的 WordNet 同義詞、上義詞與定義內容。

5-3-1 實驗設計

首先我們利用 4-2-1 小節所提出的第一個方法來為標題詞彙辨識詞義。我們將這些標題詞彙直接與 WordNet 所擁有的詞彙進行比對，找出標題詞彙的各種詞義。

在這個步驟中，我們先從 23,200 多個標題詞彙中，挑出了 19,600 多個屬於名詞、動詞與形容詞的詞彙。我們利用這些詞彙與 WordNet 進行比對後，總共有 14,600 多個詞彙出現於 WordNet 中。至於那些不包含在 WordNet 內的詞彙，我們經過觀察，發現其中有許多是屬於專有名詞，例如 IBM、COMPAQ 等等；因為這些詞彙未出現於 WordNet 中，我們無法對它們進行詞彙擴充處理。接著，我們從 14,600 多個詞彙中，找出了 69,300 多種詞義，平均一個詞彙有 4.7 種詞義存在。我們藉由 WordNet 內的詞彙網絡，找出了各詞義的同義詞、上義詞以及其定義內與此標題詞彙詞性相同的詞彙，整理成一個一個的詞彙集合。

當各詞義的詞彙集合整理出來後，我們利用這些詞彙集合與文件內容做重覆性比對，計算每個集合中的詞彙被比對到的次數。針對一個標題詞彙，我們選出次數最高的集合所代表的詞義作為此詞彙的詞義。然而，當比對次數最高的集合超過一個的情形，我們便無法判斷此詞彙該歸屬於哪一個詞義。這時我們便利用第二種詞義辨識法

賦予此詞彙在語料庫中最常出現的詞義，以完成對標題詞彙進行詞義辨識的工作。

當成功地判斷出標題詞彙的詞義之後，我們便可依據這些判斷出的詞義，來為各詞彙進行詞彙擴充的工作。詞彙擴充的方式主要是從 WordNet 中，將各已知詞義的同義詞、上義詞與定義內相同詞性的詞彙擷取出來，藉由這些擴充詞彙來補強原有的標題詞彙，增加標題詞彙的影響力。

下面，我們舉個例子做個簡單的說明。表 5 是一篇節錄後的文件，以方形框起來的字是內文中出現的標題字，而以黑底呈現的字則是擴充後的字。我們觀察第七段第一句，此句與標題詞彙比對後，只有一個詞彙 baker 出現在此句；而在擴充詞彙後，我們在此句中找出了 treasury 的上義詞{funds, finances}，以及 shift 的上義詞{modifications}，藉著這些相關詞彙使得第七段第一句變得與標題較為相關。

表 5：詞彙擴充前後的文件範例。

文件位置	內容
標題	U.S. TREASURY'S BAKER SAYS RATE SHIFTS ORDERLY..
P_1S_1	WASHINGTON, April 9 - Treasury Secretary James Baker said that changes in exchange rates have generally been orderly and have improved the prospects for a reduction in external imbalances to more sustainable levels.
P_2S_1	In remarks before the IMF's policy-making Interim Committee, Baker reiterated a Group of Seven statement last night that the substantial exchange rate changes since the Plaza agreement 18 months ago have "now brought currencies within ranges broadly consistent with economic fundamentals."
P_7S_1	Baker also urged the International Monetary Fund's executive board to review possible modifications to the Fund's compensatory financing facility before the annual meeting this fall.
註：{consistent}是 orderly 的同義詞，{change, modification}是 shift 的上義詞，{funds, finances}是 treasury 的上義詞。	

接著，摘要的製作是將經過擴充後的標題詞彙視為文件的重要詞彙，並給予一權重值，藉此對文件內的句子進行重要句子的選取。在這裡，我們將標題詞彙與擴充後的詞彙所擁有的權重值皆設定為 1，將兩者視為具有相同的重要性。在設定好權重值

後，我們利用 4-1 小節的公式 4 去計算文件內每個句子所擁有的權重總值，擷取權重總值較高的句子做為文件摘要。在摘要長度考量上，我們擷取前三分之一的句子（最多三句）作為摘要。

5-3-2 實驗結果與討論

從實驗結果中，我們發現在總數 3309 篇的測試語料文件中，只有 2916 篇文件能成功地以擴充標題詞彙的方法，找出重要句子以形成摘要。其他的 393 篇文件因為標題詞彙與擴充後的詞彙並沒有出現在文件內文裡，使得我們在計算句子權重值時，文件內所有的句子權重值皆為 0，因而無法擷取出任何句子來形成摘要。這說明了擴充標題詞彙摘要方法在某些文件上無法發揮其功效。表 6 列出擴充標題詞彙摘要方法的實驗結果，若以所有 3309 篇文件來計算，摘要結果可得到 61.9%的分類精確率。

表 6：擴充標題詞彙摘要的實驗結果

	文件數量	分類精確率
成功給出摘要的文件	2916	70.3%
無法給出摘要的文件	393	0.0%
所有文件	3309	61.9%

不過，如果我們在評估摘要成效時，只考慮那些有摘要產生的 2916 篇文件，我們可以得到 70.3%的分類精確率。因此，從上述的實驗結果，我們驗證了擴充標題詞彙摘要方法所產生的摘要，其摘要內容的確包含了文件內的重要句子，才能得到較佳的分類成效。更值得一提的是在詞義辨識部分，我們的方法不需繁複的計算與疊代，就可以對標題詞彙找出其最有可能詞義，來協助擴充標題詞彙摘要方法的進行。

5-4 結合擴充標題詞彙與重要位置摘要方法

在 5-3 小節中，我們利用擴充標題詞彙方法來擷取文件內的句子以形成摘要。但在詞彙比對過程中，我們發現並不是所有的文件內容都會出現標題詞彙，這樣的情形使得有些文件因為不能在內容中比對到標題詞彙，而無法擷取出任何句子出來，因此這篇

文件的摘要內容將會是空白的。為了彌補擴充標題詞彙摘要方法的不足，我們參照了 5-2 小節重要位置摘要方法的實驗結果，決定利用此摘要方法來加以輔助。

5-4-1 實驗設計

本研究設計了一組實驗，我們先利用擴充標題詞彙摘要方法來對所有文件進行處理，過程中所使用的各項參數與 5-3 小節中有相同的設定。接著，我們找出摘要內容為空白的文件，以重要位置摘要方法來進行處理，我們直接擷取出文件中的第一段第一句、第二段第一句以及標題來作為該文件的摘要內容。綜合這兩種摘要方法，我們便可以對所有文件擷取出重要句子。

5-4-2 實驗結果與討論

我們製作出的摘要，經過分類系統處理後，得到了 71.9%的分類精確率，相較於 65.6%的基準精確率提升了 9.6%((71.9%-65.6%) / 65.6% = 9.6%)。這顯示出我們結合擴充標題詞彙的摘要方法，並以重要位置為輔助，所製成的摘要，確實包含文件的重要內容，使得評估成效能有大幅度的改善。

6. 結論與未來方向

由第五節各種摘要的實驗結果，我們可以發現重要位置摘要實驗，得到接近基準精確率的實驗結果，說明了文件內的確存在著與主題相關的特定重要位置。另外，我們由實驗結果中，驗證了摘要長度若能達到文件的三分之一，將可提供足夠的重要資訊給予使用者。

在擴充標題詞彙摘要實驗中，我們利用了 WordNet 的特性，藉著它的豐富語意網絡對標題進行詞彙擴充，增加了標題的影響力，找出與標題較相關的語句作為摘要內容。不過在實驗過程中，我們發現這個摘要方法有其不足之處，主要是由於文件的內容不一定有著標題詞彙或其相關詞彙的存在。因此，我們提出了一種綜合擴充標題詞彙與重要位置的摘要方法，藉由結合兩個方法來協助所有的文件都能產生出重要的摘要內容，這樣的作法得到了令人滿意的摘要成效，將分類精確率提升了 9.6%。

未來，我們將試著把本文所提出的摘要方法應用於中文文件上，以測試其強健性

(robustness)；我們認為挑選重要詞彙與重要位置兩個摘要方法，應用於中文上應該會有不錯的摘要成效。不過，由於目前在中文方面，沒有一部類似 WordNet 架構的辭典。因此，可能無法使用擴充標題詞彙方法來對中文文件進行摘要處理。

本文的摘要研究，主要是以英文文件為對象，進行單文本的摘要處理。然而，隨著網際網路的盛行，使得我們可以輕易取得多樣化、多語言的資訊。因此，我們希望在未來，能夠把研究範圍擴展至多文本的摘要處理，甚至多語言摘要處理。

參考文獻

Barzilay R., K. R. McKeown and M. Elhadad, “Information Fusion in the Context of Multi-Document Summarization,” In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999, pp. 550-557.

Chen H. H. and S. J. Huang, “A Summarization System for Chinese News from Multiple Sources,” In Proceeding of the 4th Information Retrieval for Asia Language 1999, pp. 1-7.

Edmundson H. P., “New Methods in Automatic Extracting,” Journal of the ACM, Vol. 16, No. 2, 1969, pp. 264-289.

Fellbaum C., WordNet : An Electronic Lexical Database, The MIT Press, 1998.

Forsyth and Rada, “Adding an Edge in Machine Learning: Applications in Expert Systems and Information Retrieval,” Ellis Horwood Ltd, 1986, pp. 198-212.

Hayes P. J. and S. P. Weinstein, “Construction A System for Content-based indexing of a database of new stories,” In Proceedings of 2nd Annual Conference on Innovation Applications of AI, 1990.

Hovy E. and C. Y. Lin, “Automated Text Summarization in Summarist,” Advances in Automatic Text Summarization, The MIT Press, 1999.

Hovy E. and C. Y. Lin, “Identifying Topic by Position,” In Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP), Washington, DC, 1997.

Ker S. J. and J. N. Chen, “A Text Categorization Based on Summarization Technique,” In Proceeding of NLPIR Workshop of ACL2000, 2000, pp. 79-83.

Lewis D. and M. Ringuette, “Comparison of two Learning Algorithms for Text Categorization,” In Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

Mani I. and E. Bloedorn, “Summarizing Similarities and Difference Among Related Documents,” Information Retrieval, Vol. 1, No. 1, 1999, pp.35-67.

Marcu D., “Discourse Trees are Good Indicators of Importance in Text,” Advances in Automatic Text Summarization, The MIT Press, 1999.

McKeown K. and D. R. Radev, “Generating summaries of multiple news articles,” In Proceedings on the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1995, pp. 74-82.

Miller G. A., R. Beckwith, C. Fellbaum, D. Gross and K. Miller, “Introduction to WordNet: An On-line Lexical Database,” In Proceedings of the fifteenth International Joint Conference on Artificial Intelligence, 1993.

Woods W. A., “Conceptual Indexing: A Better Way to Organize Knowledge,” Sun Labs Technical Report: TR-97-61, editor, Technical Reports, 1997.

Yang Y., “An Evaluation of Statistical Approaches to Text Categorization,” Information Retrieval. Vol. 1, 1999, pp. 69-90.

邱中人，“中文新聞摘要”，碩士論文，清華大學資訊工程系， 2000.