中文手機新聞簡訊自動摘要

曾元顯
輔仁大學圖書資訊學系, tseng@lins.fju.edu.tw
第十六屆自然語言與語音處理研討會, 台北, 2004/9/2-3, 頁 177-189.

摘要：台灣地區手機的普及率已居全球之冠，國內外產業界陸續開始提供手機新聞簡訊的服務。由於手機螢幕不大，手機上新聞簡訊的自動摘要要求，與一般文獻探討的不同。爲保障訂閱者的權利，其摘要長度必須盡可能接近但不超過指定的字數，如 69 字或 45 字。此指定字數比一般標題長但比長句子還短，而且必須把新聞的重點盡可能完整的呈現出來。由於此摘要是提供給人閱讀，所以還要考慮其可讀性與連貫性等因素。本文提出一套適用於中文手機環境的新聞簡訊自動摘要方法，以降低新聞簡訊服務的營運成本。過去的研究顯示，越高的摘要壓縮比（摘要結果越短），摘要的成效越低，亦即困難度越高。手機新聞簡訊自動摘要，正好屬於高壓縮比、長度有限的極短摘要。本方法的特點在於衡量新聞句子的重要性，並找出句子與標題的相似點，結合成摘要候選句，最後依照其長度比例與相似度排序，供使用者選用。透過 40 篇即時新聞的驗證，顯示從系統提示的第一候選句，即可獲得最佳摘要的比例達 62.5%到 65%。若從系統提示的所有候選句中挑選，可得最佳摘要的比例達 75%到 80%。相對的，系統無法做出好摘要的比例，則約 20%到 25%。

關鍵詞：手機、新聞簡訊、自動摘要、中文、簡訊摘要

壹、導言

根據近一、兩年來報章雜誌的報導 [1]，台灣地區手機的普及率已經超過 100%，普及率居全球之冠。手機帶給人們極爲便利的通訊環境，任何時後、任何地點，都可以與人通訊，其便利性、行動力、易用性比電腦網路更高、更強、更好。然而，電腦網路可傳送大量的資料與數據，且其顯示器畫面較大、解析度較高，因此電腦裡有種類繁多的應用軟體，支援人們日常生活與工作所需的各項活動。相對的，手機內有限的計算能力、記憶容量與顯示器大小，被目前的技術限制了其應用範圍。如何發展與手機特性有關的技術，以釋放手機的便利性、行動力與易用性，便成爲一項產業界與學術界同時都有興趣的研究課題。

電腦網路隨時可以通訊的環境正在改變人們的生活與使用習慣，比起平面紙本新聞，網路新聞縮短了人們取得訊息的時間，但使用者要主動連線上網才能收訊，在各方面都以十倍速進展的時代，訊息流通還不夠快、不夠方便。人們隨身攜帶的手機，只要開機便能接收訊息，才能做到即時、便利的訊息交流。國內外產業界目前已有提供手機新聞簡訊的服務，如聯合線上聯合新聞網 [2]、中央社 [13]、PChome 網站 [4]、新加坡新傳媒新聞公司 [5] 都陸續推出中文手機新聞簡訊的服務。而日本的朝日新聞（日本第二大報、也是全球第二大報），自 1999 年開始透過手機提供新聞，目前該報手機新聞已有 120 萬訂閱户。朝日新聞以低廉的費用爭取手機用户訂閱新聞，用意是爲了讓行動電話使用者熟悉新聞內容，最終目的是希望增加實體報紙的訂閱 [6]。依照這個趨勢，未來很有可能大多數的報社將如同現在提供免費網路新聞一樣，以低廉的手機新聞簡訊提供訂閱者，作爲報社擴大市場、吸引訂户的手段。其前提是，手機新聞簡訊的製作成本必須非常低廉，才足以支撐起這樣便捷的資訊服務。

手機新聞與電腦網路新聞不同之處，在於必須考慮到手機有限的記憶容量與螢幕畫面。通常無法將網路新聞全文傳送到手機上，必須進一步將次要、重複的內容刪除，只留下重點內容，再傳送到手機上。一般手機新聞簡訊的長度，每一則以 69 個全型字爲限，半形字則以 158 字（letter）爲限（如聯合新聞網的限制 [2]），有的 PHS 手機則限制在 45 個全型字。

人工將新聞全文摘要成手機簡訊並非難事，但要嚴格遵守其字數限制，以充分保障訂閱者的權利，顯然會造成人工摘要的額外負擔。在理解內容、摘要新聞的同時，還要計算其字數，會耽擱人工摘要的進度，造成新聞簡訊製作成本的提高。由於電腦計算字數、切割組合文字的速度快，自動化摘要技術的運用，可以降低成本、便利訊息流通、增加手機的應用範圍，促進產業經濟的發展。

本文的目的，在提出一套適用於中文手機環境的新聞簡訊自動摘要方法，以期能降低新聞簡訊服務的營運成本，提升產業界的競爭能力。過去的研究顯示，越高的摘要壓縮比（摘要結果越短），摘要的成效越低 [7-8]。手機新聞簡訊自動摘要，正好屬於高壓縮比、長度有限的極短摘要。顯示研究手機新聞簡訊的自動摘要技術，不僅有實用上的價值，其解決方法回饋於其他類似的問題上，也有學術上的貢獻。

本文組織如下：下一節將介紹自動摘要的相關概念與研究，第三節分析新聞簡訊摘要的特性，第四節說明本文提出的方法與理由，第五節實驗驗證其成效，檢討失敗範例與提出可能的改進之道，最後一節摘要本文重點，並討論其應用與限制。

貳、相關研究

由於全球資訊網路的普及、文字出版的簡易快速，數位文件在近幾年中急速的增加，資訊過載（information overload）問題日趨嚴重，文件自動摘要技術的研發與運用變得不可或缺。產業界像 IBM

[9]、InXight [10]、Megaputer 等 [11] 皆陸續推出相關的產品，學術界近幾年來也積極投入文件自動摘要的研究，以便消除文件中的冗贅，排除次要訊息，協助人工快速消化資訊、管理資訊或降低資料量，以加速數位文件的後續加工處理。

人工摘要可以製作出重點式（informative）摘要、指示型（indicative）摘要、評論型（commentary）摘要。重點式摘要描述文件中重要的內容資訊，節省讀者閱讀全文的力氣，因此有時甚至用來替代原始文件；指示型摘要則提示文件重點項目的存在，提供足夠的資訊讓讀者決定是否應該閱讀原始文件；而評論型摘要是以簡要的形式對原文作評論，除顯露文件的重點外，亦對這些重點提出批判，幫讀者判斷，供讀者參考。人工摘要展現出來的知識處理程度與所需的背景知識，有自動摘要難以比擬之處。然而自動摘要也有其特長，如用以顯示查詢結果、提示比對程度的即時動態摘要等，這些應用導向、需要即時客製、粹取原文型態的摘要，便非常適合電腦自動化的摘要處理。前述手機新聞簡訊的獨特特性，也是文件自動摘要技術適合應用的場合。

自動摘要的作法，大抵可分爲「摘錄」（extraction）與「摘要」（abstraction）兩種。「摘錄」的結果爲文件中重要文句的重組，其作法比較不依賴額外的知識或資源，主要是根據使用者的需求，從文件本身或其他相關的文件中選取重要文句，編輯組合成使用者預期的長度即可。相對的，「摘要」的結果則不限於文件中的文句，其作法需要較多人工準備的資源，如辭典、同義詞庫、詞性標記、語法樹等，經自然語言處理後，自動生成涵蓋原文重點的簡潔文句。由於「摘要」所需資源較多，目前以「摘錄」爲主的研究佔較多數。

自動摘要的成效評估，可分爲直接（intrinsic）與間接（extrinsic）兩種方式。直接的評估需先定義出一組理想的摘要準則或答案，然後跟系統取出的摘要做比較。尤其是給人閱讀的摘要，其評估準則有重點涵蓋率（coverage）、可讀性（readability）、連貫性（coherence）、凝聚性（cohesion）、組織性（organization）及摘要長度等，因此文句中的連接詞（conjunction）、代名詞（pronoun）、前後文照應詞（anaphor）等需做適當的修詞（rhetoric）處理。間接的方式則無須具備理想的摘要答案，而是評估自動摘要的結果在其他相關應用的成效。例如，以問答的方式，比較使用者分別閱讀全文與閱讀摘要後，回答問題的成績來比較自動摘要的成效。或者無須人工直接介入，將原來以全文進行的自動分類或主題檢索的評估，以摘要來取代全文，求出摘要的分類或檢索成效，全自動的比對出各種自動摘要的效果。

近年來自動摘要相關的研究活動，有美國的 SUMMUC [12] 與 DUC [8]，以及日本 NTCIR 的 TSC [13]，其研究對象多以英文、日文的文件爲主。以 DUC 2001 年爲例，單篇文件的 100 字（words）摘要是其兩項評比中的一項，主辦單位提供 30 組、每組 10 篇英文新聞給參賽者，每一篇新聞都有三份人工摘要的答案可供評估比較。大部分參賽者都使用「摘錄」爲主的自動摘要方法。DUC 2001 與 2002 年的評估結果顯示，大部分系統的成效都跟取文件前 100 字的基準方法一樣好，雖沒有人工摘要效果好，但也沒有差太多[7]。DUC 2003 年的評比，則進一步提高難度，以 10 字極短摘要（類似自動擷取標題）的任務，取代 100 字的單篇文件摘要。

日本 NTCIR 的 TSC 2002 年有單篇與多篇文件的摘要評比，文件來源爲 Mainichi 日文新聞的社評與社會新聞。單篇文件分別取原文 20%與 40%的文字量，人工亦做出同比例的摘要，然後再由另一組人工來評斷這些摘要的可讀性與重點涵蓋狀況。八個參加單篇摘要的系統大多採用「摘錄」法，再配合文句編輯與修詞的處理，所有的系統都比只取前數句爲摘要的基準方法好，但沒有人工摘要來得好 [13]。

中文的自動摘要研究，近幾年才開始進行。台大陳信希等研究人員進行了單篇、多篇以及多語言文件自動摘要的探討 [14-18]。在單篇文件的摘要方面，以名詞與動詞來計算主題代表性、距離遠近、共現強度等指標，再結合位置、首次出現、線索詞等資訊，來計算每個句子的分數，最後從分數高者選擇原文件的 10%句子作爲固定比例摘要，並選取 10%-50%的最佳句作爲最佳比例摘要。透過分類任務與回答問卷兩種間接方法的評估，與隨機選取 10%的句子比較，最佳比例摘要的效果比固定比例好，而隨機選句的效果最差。

清大張俊盛根據摘要的形態需求，從關鍵詞首次出現的短句，取得指示型摘要，或從關鍵詞多的各段長句，形成重點式摘要。其方法爲計算文件中各句子的相似度，做階層式叢聚（clustering）後，根據位置、比例，選出指標最高者當作關鍵句，最後再潤飾、結合關鍵句後生成摘要。以 90 則中時電子報的新聞測試，經人工評估爲滿意與尚可的比例，約爲 76%。對光華雜誌的測試結果，大致也達 70%以上。從結案計劃報告的兩則新聞範例中看出 [19]，其摘要的長度分別爲 117 字與 85 字。若取最接近標題的句子作爲摘要，則其長度分別爲 117 字與 40 字。

雲科大黃純敏以內文關連法的 Global Bushy Path（GBP）觀念計算句子權重 [20]，從長度 1000 字以上的網頁中，摘錄長度 100-500 字之間的句子，與人工摘要結果比較，發現 GBP 方法的樂觀重疊率平均可達 89.77%，悲觀重疊率的平均爲 58.25%。

中外的文獻中，跟本文直接相關的計劃或研究並不多見。Banko [21] 與 Kennedy [22] 等人探討如何自動擷取標題，但標題字數太少，其技術不適合本計劃採用。Corston-Oliver 提出一套將 email 訊息濃縮的技術，以便於顯示在手機上。其方法從簡單的字串轉換（如 Monday 轉成 Mon）到複雜的語言處理（如刪除冠詞、刪除母音等）都有，並已運用於 MicroSoft Outlook 的英、法、德、西班牙文版 [23]。Buyukkokten 等人 [24] 與 Yang 等 [25] 則利用文件本身的結構資訊計算摘要，再將結果以樹狀、漸進式的技巧，顯示於 PDA 與手機上。

參、簡訊摘要之特性分析

手機新聞簡訊摘要的要求，與前述文獻探討的摘要，最大的不同，在於不論原文件長度多少，其摘要長度都必須接近但不能超過指定的字數，如前述的 69 字或 45 字。此指定字數在比標題長、比一句長句子短的情況下，必須把新聞的重點盡可能完整的呈現出來。由於此摘要是給人閱讀的，所以還要考慮其可讀性、連貫性等因素。

手機新聞簡訊的特點，是即時（real-time）傳訊。記者在現場採訪的生鮮新聞，常常以每 30 分鐘更新一次的「即時新聞」廣播於網站上。由於網站上的新聞需要閱讀者主動連線瀏覽，此「即時新聞」透過手機主動傳送給訂户，比網站廣播會更有效率。然而「即時新聞」的特點是其長度短，幾乎只有兩三句。如表一的三個例子，文件一含標題只有三句，文件二有兩句，文件三有三句。

從表一的文件範例可知，前述文獻以「摘錄」方法選句子的方式，似乎都不能直接運用於此簡短字數的摘要上，必須要將句子再裁減成較小的單位，才容易處理。

然而單純以逗號「，」來切割句子，造成的「片段」有時在語意上較不完整。黃聖傑 [26] 曾運用連接詞與動、名詞資訊將中文句子切割成較小單位，以應用在多篇文件的自動摘要上。其方法先將文句斷詞，對每個詞彙標上詞性標記（如動詞、名詞、連接詞等），再以自行整理的規則切割長句成「小句」（meaningful unit, MU）。例如：逗點分隔的「片段」，若起始爲「而且」、「且」等詞，則將此「片段」往前合併；若起始詞彙爲動詞，則其主詞應該在前面的「片段」，因此也將此「片段」往前合併。

表一：即時新聞文件範例

文件一	國眾奪下中華電北區 FTTB L2 Switch 採購案【時報-記者莊丙農台北報導 2003/08/14, 11:21:28】國眾電腦宣布取得「中華電信北區分公司 Ethernet- based FTTB L2 Switch 服務系統」採購案，以供中華電信協助中小企業利用寬頻網路發展商機之用。本採購設備包括設置於用户端大樓之遠端超高速乙太網路交換器（GESWr）、超高速數位用户迴路設備（EoVDSL）及維運整體系統所需之網路管理等相關軟、硬體設備，由國眾得標，智邦集團傳易（SMC）、和心光通、飛瑞、安捷倫及浩網等廠商負責提供相關整合產品。
文件二	佼佼訪王貞治，豪華日本行【時報-台北電 2003/09/01, 07:57:33】黃子佼爲訪問王貞治前往福岡巨蛋欣賞日本職棒比賽，雖然 2 天行程緊湊，但佼佼此行可說是「頂級豪華之旅」，除了能親眼目睹日本職棒，專訪職棒明星王貞治，還住在一晚高達 6 萬日幣的飯店裡，且如願吃到頂級的佐賀牛肉壽喜燒。
文件三	中共採購新規定，重擊微軟。【時報-外電報導】中共爲了保護大陸軟體行業，新推出的採購規定中要求政府單位未來僅能購買內裝中國作業系統及應用程式的硬體，要購買非本國軟體系統的政府單位，一律特別呈報。據了解，微軟自去年以來，在大陸業務進展並不順利，儘管微軟大力投資當地，並改組大中華區人事，但在大陸急力扶持國產軟件下，微軟在大陸業務可能遭致命打擊。

以文件二的第二句爲例，按照上述的方法，必須將最後一個片段：「且如願吃到頂級的佐賀牛肉壽喜燒」往前連結（因爲「且」字開頭）。但前一個片段：「還住在一晚高達 6 萬日幣的飯店裡」本身無法當成一個小句的起始（因爲「還」字開頭），必須再往前連結。但前一個片段：「專訪職棒明星王貞治」的「專訪」是動詞，必須繼續往前連結到「除了能親眼目睹日本職棒」。依此做下去，最後可以得到『黃子佼爲訪問王貞治前往福岡巨蛋欣賞日本職棒比賽』（23 個字）、『雖然 2 天行程緊湊，但佼佼此行可說是「頂級豪華之旅」』（25 個字）、『除了能親眼目睹日本職棒，專訪職棒明星王貞治，還住在一晚高達 6 萬日幣的飯店裡，且如願吃到頂級的佐賀牛肉壽喜燒』（53 個字）等三小句。同理，標題本身也可分割成一個小句：『佼佼訪王貞治，豪華日本行』（12 個字）。

將文件分割成小句（MU）後，要組合出預定的長度。由於預定的長度爲 69 字或 45 字，幾乎只能容量一、兩個小句。**爲了能點出文件的大意，使擷取出來的摘要具有畫龍點睛的效果，我們可以選擇標題做爲其中一個小句，剩下的字數再來容納其他小句。**剩下的候選小句應當選擇最長但不超過總長度的小句，以便減少接句造成的文句不通順。以上例而言，選最後一個小句跟標題按原順序結合，可以得到 67 字的摘要：『佼佼訪王貞治，豪華日本行，除了能親眼目睹日本職棒，專訪職棒明星王貞治，還住在一晚高達 6 萬日幣的飯店裡，且如願吃到頂級的佐賀牛肉壽喜燒。』，結果相當漂亮，而且非常接近 69 字，可說是最佳摘要。若要產生 45 字摘要，可將標題及第二長的小句結合，得到 37 字摘要：『佼佼訪王貞治，豪華日本行，雖然 2 天行程緊湊，但佼佼此行可說是「頂級豪華之旅」』。此句比另一小句結合標題得到的 35 字摘要：『佼佼訪王貞治，豪華日本行，黃子佼爲訪問王貞治前往福岡巨蛋欣賞日本職棒比賽』，效果還要好（因爲長度更接近 45 字，且具內容似乎更具互補性）。

上面的作法歸納如下：

一、 將文件斷詞、做詞性標記，按照某些規則分割成小句。

二、 選擇與標題合併後長度最長但不超過預定長度的小句，接在標題後，當成摘要送出。

但上述的作法，有幾個問題：

一、 正確的斷詞、詞性標記與分割小句並不容易，何況新聞從政治、社會、經濟、外交、軍事、科技到生活、運動、娛樂、健康、文學等有各種主題，未知詞、難以預料的語法繁多，如果沒有事先分析完整，分割出的小句，其語意完整度難以保證。

二、 依照上述原則選出的小句並非最佳。例如前例的 45 字摘要，可以選擇標題與文件最後兩個片段做成效果更好的 45 字摘要：『佼佼訪王貞治，豪華日本行，還住在一晚高達 6 萬日幣的飯店裡，且如願吃到頂級的佐賀牛肉壽喜燒。』。（同理，因爲長度更接近 45 字，且具內容更具互補性）

三、 對短文件似乎有用，對長文件如何處理？

上面第一點要處理各種領域的文件，幾乎是目前爲止還在研究的問題。既然第一點跟第二點都顯示上述方法不見得可以得到最佳的結果，我們覺得可以不處理第一點，而直接將文句依逗點斷成「片段」，然後將各個片段與標題結合成候選摘要，再評估哪一個候選摘要最適當。

至於第三點，當文件長度越長而摘要的長度依然固定如此短時，其摘要困難度越高、不同人做的摘要歧異性也越大。想像 5 句中取 2 句的可能組合，與 15 句取 2 句的組合數，兩者差距蠻大的。然而長文件中的每一句，不見得都重要。我們可以仿照重點式「摘錄」選擇重要句子的技巧（考慮關鍵詞詞頻、文句位置、線索詞彙等），先對長文件做出 3 至 5 句左右的摘錄，再從這摘錄中運用上述的技巧獲得最後的簡訊摘要。

肆、本文提出的方法

經過上述分析後，本文提出的方法如下：

步驟一：評估新聞文件每個句子的重要性，取最重要的前 n 句，作爲「候選句」。

步驟二：將上述每個重要句子，與標題結合，做成「摘要候選句」，並記錄其字數與相似度。

步驟三：根據字數與相似度排序摘要候選句，由高到低依序輸出，並提供字數與相似度資訊，方便使用者挑選。

在步驟一中，句子的重要性是以該句子出現的關鍵詞，依下列公式來決定：

$$\sum_{w \in Keywords} (0.5 + 0.5 * tf_w / \max\_tf)$$

其中tf_w爲關鍵詞w在該文件中的詞頻，max_tf爲該文件出現最多次的關鍵詞的詞頻。在此所謂關鍵詞彙，是以Tseng的演算法求出最大重複字串（maximally repeated string）[27]，經濾除停用詞後，得到的重複詞彙（出現多於一次），做爲該文件的關鍵詞彙。此方法假設文件的主題詞彙會重複出現，但並非所有的重複字串都是有用的關鍵詞，它們必須是最長的，或是出現頻率最高的，因此稱爲最大重複字串。例如前兩句中「最大重複字串」出現了二次，而「重複字串」出現了三次，那麼這兩個詞都會被擷取出來。但「大重複字」此字串也出現二次，但因它是「最大重複字串」的完全子字串，所以不會被擷取出來成爲關鍵詞。另外，由於標題在新聞中相當重要，因此我們以 12 萬詞的詞庫對標題做斷詞處理，經停用詞過濾後，將剩下的詞彙都視爲關鍵詞。文件中的每個句子都以上述公式計算其重要性，由大到小排序後，取前n個句子作爲後續處理之用。在此n可視爲使用者指定的候選句子數，亦即，電腦摘要完後，可供使用者選擇的摘要數。

在步驟二中，爲了要讓結合出來的摘要，具有內容一致性、連貫性與互補性，跟標題結合的句子，最好跟標題在內容上有部份重疊，亦即有足夠的相似度。當然相似度最高，則跟標題完全相同，並不恰當。但一般新聞編輯下的標題，很少從文件本身的句子完全複製得來，而是更濃縮、更簡潔的「片段」。其結果是與標題相似的句子，在內容上跟標題就自然而然具有互補性。除此之外，我們也要知道從那個片段開始跟標題做結合。這意味著，不僅要找出句子與標題的相似度，還要找出在哪裡最相似。爲了同時滿足這兩項需求，以動態規劃（dynamic programming）方式比對標題與句子之間的編輯距離（edit distance），亦即相似度，自然成了我們的選擇。

在動態規劃裡，所謂編輯距離是指利用「插入」、「刪除」、與「代換」的動作，將一個字串轉成另一個字串「所需最少的步驟」（或是「所需最少的計算成本」）。一般的動態規劃法可表達如下 [28]：假設有兩字串 A 與 B，長度各爲 n 與 m。將兩字串從頭比對起，則比對到 A 的第 i 個字（以 A[i]表示）與 B 的第 j 個字（以 B[j]表示）的編輯距離爲：

d[i, j] = min(d[i-1, j] + w(A[i], 0), d[i-1, j-1] + w(A[i], B[j]), d[i, j-1] + w(0, B[j]))

其中 min(X, Y, Z)表示取 X, Y, Z 三個數中最小的值，而初始值爲：

d[0, 0] = 0

d[i, 0] = d[i-1, 0] + w(A[i], 0) , 1<= i <= n

d[0, j] = d[0, j-1] + w(0, B[j]) , 1<= j <= m

另外，函數 w(X, Y)的意義爲

w(A[i], B[j]) : 表示將 A[i]代換成 B[j]的計算成本

w(A[i], 0) : 表示插入 A[i]的計算成本

w(0, B[j]) : 表示刪除 B[j]的計算成本

我們以標題A=adc，「候選句」B=**adec**decf爲例，且假設代換、插入與刪除的計算成本都爲 1，則d[i, j]可以表示成矩陣的第i列的第j行，如下：

A \ B	**a**	**d**	**e**	**c**	d	e	c	f
A	0	1	2	3	4	5	6	7
D	1	0	1	2	3	4	5	6
C	2	1	1	1	2	3	3	4

從矩陣最後一列的最後面掃瞄起，發現**第一個距離最低**的地方即是我們想要找的地方，亦即 B 的前四個字 adec 是跟 A 最相似的部份。接句的時候然，就把 B 的前四個字 adec 代換成 A 的 adc，做成「摘要候選句」：adcdecf。

上述方法比對兩字串時，是找出兩字串從頭開始的相似度。但當相似的字串在中間時，則無法如前述方法看出相似的位置。例如A=adc，B=dec**adec**f時，則其距離矩陣爲：

A \ B	d	e	c	**a**	**d**	**e**	**c**	f
a	1	2	3	3	4	5	6	7
d	1	2	3	4	3	4	5	6
c	2	2	2	3	4	4	4	5

結果最相似的片段，距離不是最低。改善的方法，可修改初始條件如下 [28]：

d[0, 0] = 0

d[i, 0] = d[i-1, 0] + w(A[i], 0) , 1<= i <= n

d[0, j] = 0 , 1<= j <= m

亦即比對時，允許從較長字串的任何位置開始比對起。改變後的矩陣如下：

A \ B	d	e	c	**a**	**d**	**e**	**c**	F
a	1	1	1	0	1	1	1	1
d	1	2	2	2	0	1	2	2
c	2	2	2	3	1	1	1	2

當上述動態規劃比對完成，從最後一列的後面掃瞄，找到第一個距離最低的位置後，由於接句必須接在標點符號上以維持可讀性，因此必須左右掃瞄最近的標點符號，找出編輯距離最小的標點符號位置，作爲可能的接句點。

雖然可以找出相似度最佳的片段位置，然而此種相似度僅是一種內容的近似，沒有眞正反映語意的近似，而且相同或極相近的近似點可能有數個，在以長度的適合度爲優先考量的情況下，我們再輔以下列方式微調：

一、若接句以後，超過長度，則接句點試著往後挪，縮短接句的子句，以不超過要求長度的最多子句，與標題連接。

二、若接句以後，比摘要長度還短，則接句點試著往前挪，增長接句的子句，以不超過要求長度的最多子句，與標題連接。

上述調整接句點後，都可從動態規劃比對結果得知其編輯距離。爲了便於比較不同長度句子的相似度，Lopresti 等人 [28] 以公式 exp(edit/(edit-m))將編輯距離轉換成相似度，其中 exp 爲自然指數（natural exponent），edit 爲編輯距離，m 爲標題的字數。雖然此相似度介於 0 到 1 之間，但其間距有時差距太大，不利於比較。例如，標題 15 個字，而編輯距離爲 13、11 與 9 時，相似度分別爲 0.0015、0.0639 與 0.2231。爲縮短其差距，我們**將 m 以 m+n 取代**，其中 n 爲「候選句」的長度，變成：

$$sim = \exp(\frac{edit}{edit - m - n}) = 1 \Big/ e^{\frac{edit}{n+m-edit}}$$

修改後的相似度，其最大值爲 1，最小值爲 1/exp(m/n)。

由於新聞的寫法，以金字塔型方式敘述，細節的描述越後面越詳細，相對的越前面的文字，越像摘要。因此，爲加強前面句子的相似度，優先考量前數句，若原新聞文件的內文超過 k 個句子（後續的實驗中，k 都設爲 3），則非前兩句的句子，其相似度都乘以 0.85，作爲最後的相似度。

在步驟三中，要根據字數與相似度排序摘要候選句。同樣的爲便於比較，先將結合後的字數除以指定的摘要字數，使其轉換成 0 到 1 之間的字數比例。有了「字數比例」與「相似度」後，一個可能比較好的方法，是事先根據此兩度量，人工選出最佳摘要候選句，然後以機器學習技術，學出一套分類器，使其爾後看到某個摘要候選句的字數比例與相似度後，可以決定其是否爲最佳候選句，或是決定其最後的排序。

然而機器學習的效果受訓練個數的影響很大，在訓練資料不易累積的情況下，我們決定先以人工設計規則，有了初步成效，可用來協助獲得訓練資料後，將來再嘗試以機器找出最佳的規則。給定 n 個摘要候選句，我們設計的規則如下：

一、 找出相似度最高的摘要候選句 A 與字數比例最高的摘要候選句 B，若 A 即是 B，則輸出 A，並從摘要候選句中將 A 刪除。

二、 若 A 的相似度大於 B 的 1.25 倍，且 A 的字數比例大於 B 的 0.75 倍，則輸出 A，否則輸出 B，並從摘要候選句中將輸出的句子刪除。

三、 重複步驟一到二，直到沒有任何摘要候選句。

伍、成效評估

我們以 40 篇 2003 年 8、9 月左右的中國時報即時新聞，測試上一節提出的方法。表二是表一中「文件三」的輸出範例。此文件內文只有兩句，與標題組合後，系統依排序結果提示兩摘要候選句供使用者

挑選。在 45 字的摘要中，兩個候選句比較之下，第一句接句的位置有個不太相干的「但」字，閱讀時有突兀感，且其後面的敘述重複標題後半部的內容。相對的，第二句雖然較短，但文句結構與內容都很完整。因此使用者可以選擇第二句當作 45 字的簡訊摘要。至於在 69 字的摘要中，第一候選句在長度與內容上都非常好，直接挑選輸出即可。

表三羅列 40 篇**人工挑選出的最好摘要**。每一篇的第一列爲其標題，第二列與第三列分別爲，針對 45 字與 69 字摘要後，人工選擇出來最佳的候選句。在行方面，第二行的數字表示該摘要候選句的實際字數。倒數第二行的標題列，則顯示該篇文件的內文有幾個句子，在摘要列部份，則顯示該摘要來自系統排序的第幾名候選句。最後一行，則是針對這些組合後的最佳摘要，人工評定其品質，G 代表「佳」、F 代表「普通」、B 代表「差」。

表二：自動摘要範例：文件全文爲表一中的文件三。

45 字 摘 要	排序 1, 相似度=0.8767, 長度比例=0.9333, 共 42 字 中共採購新規定，重擊微軟，但在大陸急力扶持國產軟件下，微軟在大陸業務可能遭致命打擊。 排序 2, 相似度=0.8636, 長度比例=0.8000, 共 36 字 中共採購新規定，重擊微軟，要購買非本國軟體系統的政府單位，一律特別呈報。
69 字 摘 要	排序 1, 相似度=0.8636, 長度比例=0.9130, 共 63 字 中共採購新規定，重擊微軟，儘管微軟大力投資當地，並改組大中華區人事，但在大陸急力扶持國產軟件下，微軟在大陸業務可能遭致命打擊。 排序 2, 相似度=0.8636, 長度比例=0.5217, 共 36 字 中共採購新規定，重擊微軟，要購買非本國軟體系統的政府單位，一律特別呈報。

表三：40 篇人工挑選出的最佳摘要候選句。

篇次	內容		*	品質
1	Title	台鐵計軸器採購下周進行第 11 度招標	2	
	45	台鐵計軸器採購下周進行第 11 度招標，擁有這項產品製造技術的歐洲廠商，已摩拳擦掌準備進場搶標。	1	G
	66	台鐵計軸器採購下周進行第 11 度招標，不限定廠商使用材質，下周公告招標後，等標期約 28 天、審查作業 10 天，最快 10 月中旬可以最低價格進行決標。	1	G
2	Title	台十一線濱海公路山崩，交通中斷	5	
	45	台十一線濱海公路山崩，交通中斷，造成豐濱鄉對外交通完全中斷，民眾必須往台東縣才能找到出路。	1	G
	69	台十一線濱海公路山崩，交通中斷，形成九十度丁坡度，連日來花蓮間歇性豪雨不斷，該地段今天早上九點多終於發生小規模山崩，交通中斷阻斷來往車輛。	1	G
3	Title	台鐵與工會最後協商無交集，中秋是否停駛各說各話	4	
	45	台鐵與工會最後協商無交集，中秋是否停駛各說各話，會員現在也不敢說不上班，只是應付一下主管。	1	G
	61	台鐵與工會最後協商無交集，中秋是否停駛各說各話，工會說，這是台鐵當局的一貫技倆，會員現在也不敢說不上班，只是應付一下主管。	2	G
4	Title	兩岸航空業邁進實質合作時代	1	
	43	兩岸航空業邁進實質合作時代，這項合作也正式宣布兩岸航空貨運開始走入實質合作的經營時代。	1	G
	69	兩岸航空業邁進實質合作時代，將再度齊聚廈門，出席這項兩岸航空業界首度合資的盛會，這項合作也正式宣布兩岸航空貨運開始走入實質合作的經營時代。	1	B
5	Title	高市招商，力邀重量級企業與會	2	
	30	高市招商，力邀重量級企業與會，以及多功能經貿園區的未來遠景。	1	B
	57	高市招商，力邀重量級企業與會，而行程中必定會談到世界大港高雄港和小港機場的海空優勢，以及多功能經貿園區的未來遠景。	1	G
6	Title	雲縣規劃產業聚落，建立招商網路	2	
	32	雲縣規劃產業聚落，建立招商網路，發展各專區內互補特性，相互支援。	1	G
	67	雲縣規劃產業聚落，建立招商網路，並規劃以麥寮自由港區、中科雲林基地及雲林科技工業區發展爲三個相互支援發展的產業聚落，爭取更多企業投資。	1	G
7	Title	中油調高桶裝瓦斯價格	4	
	34	中油調高桶裝瓦斯價格，以二十公斤裝桶裝瓦斯來看，每桶批售價調高八元。	1	G
	64	中油調高桶裝瓦斯價格，爲反應進口成本上漲壓力，中油決定自四日零時起調漲各類液化石油氣產品牌價，調整幅度爲二．六五％至三．九四％。	1	G
8	Title	經濟部：攤販不會就地合法	1	
	29	經濟部：攤販不會就地合法，因此不會有「就地合法」這個問題。	1	B
	54	經濟部：攤販不會就地合法，未來攤販仍須先通過地方政府審核後才能獲得營業許可，因此不會有「就地合法」這個問題。	1	G
9	Title	獅、象四連戰第二役，統一獅將派出威森掛帥	3	
	43	獅、象四連戰第二役，統一獅將派出威森掛帥，親自派遣場務人員前來台北，爲威森整理投手丘。	1	G
	68	獅、象四連戰第二役，統一獅將派出威森掛帥，爭取今天晚間的勝利，統一特別從台南帶著「土坯」前來新莊，賽前將由工作人員親自爲威森整理投手丘。	1	G
10	Title	中華職棒大聯盟，教練護盤，「劉」住勝果	5	
	42	中華職棒大聯盟，教練護盤，「劉」住勝果，戰績繼續保持第一，領先獅隊的勝差拉開爲 1.5 場。	1	F
	69	中華職棒大聯盟，教練護盤，「劉」住勝果，順利終結獅隊最後反撲，拿下 1 次救援成功，距離上次（2000 年 9 月 23 日對牛隊）贏得救援成功，已將近 3 年了。	1	F

11	Title	美國網球公開賽：阿格西驚險闖進 8 強	4	
	33	美國網球公開賽：阿格西驚險闖進 8 強。阿格西遇險，險遭丹特襲擊成功。	2	G
	65	美國網球公開賽：阿格西驚險闖進 8 強；西哥畢竟老江湖，第 2 盤穩中求勝，第 3 盤守住丹特強力攻勢，終於讓小老弟因強攻不破，右腳傷重退賽。	1	G
12	Title	娜姐送吻，小甜甜人氣下滑，克莉絲汀變旺	5	
	40	娜姐送吻，小甜甜人氣下滑，克莉絲汀變旺，克莉絲汀是「一吻成名」，一夕間躍升榜首。	3	G
	50	舌吻事件這兩天在網路上引爆熱烈討論，雖然布蘭妮、克莉絲汀都被娜姐送上香吻，但人氣指數卻呈現兩個極端。	4	G
13	Title	佘詩曼睡一睡，溫碧霞脫一脫，數百萬入袋	6	
	41	佘詩曼睡一睡，溫碧霞脫一脫，數百萬入袋；而溫碧霞則是小脫一下，就賺到四百多萬台幣。	1	F
	67	佘詩曼睡一睡，溫碧霞脫一脫，數百萬入袋，最近港星佘詩曼自稱在床上睡一睡，就有六百萬台幣入袋；而溫碧霞則是小脫一下，就賺到四百多萬台幣。	1	G
14	Title	王識賢求婚很靦腆，張鳳書當老師	3	
	42	王識賢求婚很靦腆，張鳳書當老師，反倒是張鳳書教他，求婚就該在大庭廣眾下告白才有誠意。	2	F
	68	王識賢求婚很靦腆，張鳳書當老師，導演要求他下跪求婚，王識賢靦腆的說人太多，不好意思，反倒是張鳳書教他，求婚就該在大庭廣眾下告白才有誠意。	1	G
15	Title	百慕達銀行在日本開設辦事處	2	
	13	百慕達銀行在日本開設辦事處	1	B
	64	百慕達銀行在日本開設辦事處。Bermuda Global Fund Services Limited 東京辦事處將坐落于東京，並將作爲百慕達銀行旗下全球範圍的 GFS 部門與其日本客户之間的聯繫機構。	1	G
16	Title	東芝公司同意在系統單晶片中使用 ARM 晶片	3	
	45	東芝公司同意在系統單晶片中使用 ARM 晶片，雙方已經通過新的授權協議拓展了彼此間的戰略合作關系。	1	G
	69	東芝公司同意在系統單晶片中使用 ARM 晶片，東芝公司已經同意把 ARM1026EJ-S(TM)晶片用于促成創新的系統單晶片(SOC)應用產品，從而豐富其新一代數碼產品組合。	1	G
17	Title	Inno Micro 在日本經銷並出售 nStor 產品	2	
	31	Inno Micro 在日本經銷並出售 nStor 產品，在日本出售和經銷 nStor 全系列存儲產品。	1	B
	54	Inno Micro 在日本經銷並出售 nStor 產品，日本一家私營整合商和經銷商 Inno Micro 已簽署一份協議，在日本出售和經銷 nStor 全系列存儲產品。	1	F
18	Title	登記列管繳稅營業，攤販將全面合法	2	
	34	登記列管繳稅營業，攤販將全面合法，預估有數十萬攤販可望就地「合法」。	1	G
	64	登記列管繳稅營業，攤販將全面合法，將把全台灣既存和未來可能新增的攤販，全部改以登記制統一管理，預估有數十萬攤販可望就地「合法」。	1	F
19	Title	行動攤販車可在風景區營業	4	
	41	行動攤販車可在風景區營業，甚至還成立加盟總部，鼓勵民眾只要投資數十萬元就可以創業。	1	G
	56	行動攤販車可在風景區營業，包括行動咖啡館、行動彩印店等，甚至還成立加盟總部，鼓勵民眾只要投資數十萬元就可以創業。	1	F
20	Title	輕軌工業擬改採國內標	2	
	30	輕軌工業擬改採國內標，採國內外業者共同承攬但由國內業者主導。	1	G
	57	輕軌工業擬改採國內標，放寬招商「實績」要求，提高國內業者自製率比重至五０％，採國內外業者共同承攬但由國內業者主導。	1	G
21	Title	中共採購新規定，重擊微軟	2	
	36	中共採購新規定，重擊微軟，要購買非本國軟體系統的政府單位，一律特別呈報。	2	G
	63	中共採購新規定，重擊微軟，儘管微軟大力投資當地，並改組大中華區人事，但在大陸急力扶持國產軟件下，微軟在大陸業務可能遭致命打擊。	1	G
22	Title	扶持軟體產業，中共在融資、上市和稅收方面給予優惠措施	4	
	45	扶持軟體產業，中共在融資、上市和稅收方面給予優惠措施，成立風險投資公司，設立風險投資基金。	1	G
	62	扶持軟體產業，中共在融資、上市和稅收方面給予優惠措施，以求二〇一〇年大陸的軟體產業研究開發和生產能力達到或接近國際先進水平。	1	G
23	Title	緊縮房地產業，中共加大力道	3	
	40	緊縮房地產業，中共加大力道，要控制此類項目的建設用地供應量，或暫停審批此類項目。	1	G
	68	緊縮房地產業，中共加大力道，對高檔大户型商品房、辦公大樓與商業性用房積壓較多的地區，要控制此類項目的建設用地供應量，或暫停審批此類項目。	1	G
24	Title	陳總統：中華民國是主權獨立國家	3	
	34	陳總統：中華民國是主權獨立國家，國軍要爲捍衛中華民國主權與領土而戰。	1	G
	66	外傳前總統李登輝指「陳總統只說中華民國是國號，沒有說中華民國是國家」，而陳總統昨天則向三軍官兵強調「中華民國是一個主權獨立的國家」。	3	G
25	Title	明年總統大選，藍綠基本盤皆見鬆動	3	
	45	明年總統大選，藍綠基本盤皆見鬆動，而當年的選民，歷經政黨輪替，如今投票意向已出現明顯改變。	1	G
	64	明年總統大選，藍綠基本盤皆見鬆動，上屆大選支持泛藍的選民，陣腳略微鬆動；而之前支持陳呂配的泛綠選民，也有相當比例出現流失的現象。	1	G
26	Title	競國實業董事會決議配息配股基準日爲 9 月 12 日。	1	
	39	競國實業董事會決議配息配股基準日爲 9 月 12 日，9 月 8 日起至 9 月 12 日停止股票過户。	1	G
	39	競國實業董事會決議配息配股基準日爲 9 月 12 日，9 月 8 日起至 9 月 12 日停止股票過户。	1	G
27	Title	國眾奪下中華電北區 FTTBL2Switch 採購案	2	
	39	國眾奪下中華電北區 FTTBL2Switch 採購案，以供中華電信協助中小企業利用寬頻網路發展商機之用。	1	G
	58	國眾奪下中華電北區 FTTBL2Switch 採購案，由國眾得標，智邦集團傳易（SMC）、和心光通、飛瑞、安捷倫及浩網等廠商負責提供相關整合產品。	1	G

28	Title	亞太電信集團跨足線上遊戲，今年營收約 2500 萬元(2-1)	3	
	36	亞太電信集團跨足線上遊戲，今年營收約 2500 萬元(2-1)，4C 整合的佈局儼然成形。	1	G
	69	亞太電信集團跨足線上遊戲，今年營收約 2500 萬元(2-1)，推出新的娛樂事業群，亞太集團版圖橫跨了電信、網路、通訊、加值內容，4C 整合的佈局儼然成形。	1	G
29	Title	《未上市個股》亞太電信推出「猿人在線」品牌，初期以代理爲主(2-2)。	3	
	41	《未上市個股》亞太電信推出「猿人在線」品牌，初期以代理爲主(2-2)，朝線上遊戲邁進。	2	G
	63	《未上市個股》亞太電信推出「猿人在線」品牌，初期以代理爲主(2-2)，因此結合集團內各式寬頻服務載具與平台的資源，朝線上遊戲邁進。	1	F
30	Title	友達第五代彩色濾光片廠十月起逐步量產，最大月產能 12 萬片	3	
	43	友達第五代彩色濾光片廠十月起逐步量產，最大月產能 12 萬片，使友達有效掌握上游關鍵零組件。	1	G
	64	友達第五代彩色濾光片廠十月起逐步量產，最大月產能 12 萬片，月產能 7 萬片，預估未來每月最大產能 12 萬片玻璃基板，供全球大尺寸面板需求。	1	F
31	Title	中壽投資型商品「一觸得利」狂賣，一周銷售達 13 億元	5	
	40	中壽投資型商品「一觸得利」狂賣，一周銷售達 13 億元，不僅爲業界首創，引發熱賣風潮。	2	G
	57	中壽投資型商品「一觸得利」狂賣，一周銷售達 13 億元，投資標的爲逆浮動+正浮動利率債券，不僅爲業界首創，引發熱賣風潮。	2	G
32	Title	29 日台積電 ADR 收盤價 11.78 美元，較前交易日上漲 0.08 美元。	1	
	42	29 日台積電 ADR 收盤價 11.78 美元，較前交易日上漲 0.08 美元，漲幅爲 0.68%，換算回台股每股價格約 80.54 元。	1	G
	54	29 日台積電 ADR 收盤價 11.78 美元，較前交易日上漲 0.08 美元，較前一交易日上漲 0.08 美元，漲幅爲 0.68%，換算回台股每股價格約 80.54 元。	1	B
33	Title	「美夢成眞」趕戲，葉全眞累壞吊了點滴再上	3	
	45	「美夢成眞」趕戲，葉全眞累壞吊了點滴再上，不顧醫生要她吊點滴多休息的叮嚀，又回棚內拍戲去。	1	G
	59	「美夢成眞」趕戲，葉全眞累壞吊了點滴再上，所以她在打了兩劑粗血管針後，不顧醫生要她吊點滴多休息的叮嚀，又回棚內拍戲去。	1	B
34	Title	八點檔現拍現播，演員連連發病	4	
	43	八點檔現拍現播，演員連連發病，除了中視、華視，其餘三台都以現拍現播的方式，走本土路線。	1	G
	58	八點檔現拍現播，演員連連發病。演員日夜趕戲來趕播出，體力已受考驗，偏偏表演方式更耗費體力，病號、傷兵也因此連連爆發。	2	G
35	Title	周俊三蹲牢房，代價很值得	2	
	35	周俊三蹲牢房，代價很值得，辛苦還是有代價的，讓他獲得 3 萬元的豐厚酬勞。	1	G
	35	周俊三蹲牢房，代價很值得，辛苦還是有代價的，讓他獲得 3 萬元的豐厚酬勞。	1	G
36	Title	佼佼訪王貞治，豪華日本行。	1	
	45	佼佼訪王貞治，豪華日本行，還住在一晚高達 6 萬日幣的飯店裡，且如願吃到頂級的佐賀牛肉壽喜燒。	1	G
	67	佼佼訪王貞治，豪華日本行，除了能親眼目睹日本職棒，專訪職棒明星王貞治，還住在一晚高達 6 萬日幣的飯店裡，且如願吃到頂級的佐賀牛肉壽喜燒。	1	G
37	Title	「棋靈王圍棋入門之旅」活動開跑	3	
	34	「棋靈王圍棋入門之旅」活動開跑，使得圍棋儼然成爲最新的全民益智運動。	1	G
	61	「棋靈王圍棋入門之旅」活動開跑，再加上不久前奪得今年日本本因坊頭銜的旅日棋手張栩效應，使得圍棋儼然成爲最新的全民益智運動。	1	G
38	Title	周末官邸藝文沙龍，王琄邀親子無言的交流	5	
	35	周末官邸藝文沙龍，王琄邀親子無言的交流，激發出親子間的想像力與創造力！	1	G
	60	周末官邸藝文沙龍，王琄邀親子無言的交流，並且藉由各式精心設計的劇場遊戲—模仿、帶領、互動，激發出親子間的想像力與創造力！	1	G
39	Title	故宮德國文物大展，開放展場設計權	3	
	39	故宮德國文物大展，開放展場設計權，舉辦公開說明會，歡迎設計師與建築師前來參與。	2	G
	65	故宮德國文物大展，開放展場設計權，故宮破天荒將公開舉辦展場競圖，預計本月 15 日下午 2 點，舉辦公開說明會，歡迎設計師與建築師前來參與。	1	G
40	Title	藝文界前輩進駐爲豐樂童畫賽暖身	1	
	15	藝文界前輩進駐爲豐樂童畫賽暖身	1	B
	15	藝文界前輩進駐爲豐樂童畫賽暖身	1	B

爲了便於分析，依據最佳摘要的序位及其品質，統計表三的資料，結果列於表四。從表四中可知，由系統提示的第一句，即可獲得最佳摘要的比例達 62.5%或 65%。若從系統提示的所有候選句中挑選，可得最佳摘要的比例達 75% 或 80%。相對的，系統無法做出好摘要的比例，則約 20%到 25%。

表四：序位、品質分析表。左欄 45 字摘要，右欄 69 字摘要。

品質 / 序位	佳	普通	差	品質 / 序位	佳	普通	差
1	26 (65.0%)	2 (5.0%)	5 (12.5%)	1	25 (62.5%)	6 (15%)	4 (10%)
2	5 (12.5%)	1 (2.5%)	0	2	3 (7.5%)	0	0
3	1 (2.5%)	0	0	3	1 (2.5%)	0	0
4	0	0	0	4	1 (2.5%)	0	0
合計	32 (80.0%)	3 (7.5%)	5 (12.5%)	合計	30 (75%)	6 (15%)	4 (10%)

表四顯示表三中有 9 句較差的摘要候選句，分別出現在第 4、5、8、15、17、32、33 與 40 篇文件。其中有 4 句（4、5、8、33）是不當連接詞，如「將再度」、「以及」、「因此」、「所以」等造成的連貫性或可讀性問題。有 2 句（17、32）是接句的位置不當，而造成重複片段的問題。另外，有 3 句（15、40）是重複標題，表示該新聞是一篇難以摘要的新聞，因此得不到適當的候選句。至於總共 10 句的普通摘要，則在語句的連貫性上，比較不那麼流暢，但也還具有可讀性。

要改進不當連接詞的缺失，並非直接以詞庫比對然後加以剔除即可，可能需要更深入的語法剖析或語意理解才行。例如第 36 篇，接句接在「還住在...」，以及「除了能...」，就接得非常好。至於接句位置不當，造成重複的片段，則可在接句時，進一步偵測而加以剔除。而文件本身難以摘要，得不到適當的候選句，則必須根本改變摘要的方法。最後連貫性不流暢，是乃此接句方法造成的根本性問題，除非利用更深入的語法分析，否則簡單的找相似點接句，便可能產生這種現象。

陸、結語

本文提出一套適合手機新聞簡訊的自動摘要方法，其特點在於衡量新聞句子的重要性，並找出句子與標題的相似點，結合成摘要候選句，最後依照其長度比例與相似度排序。透過 40 篇即時新聞的驗證，顯示全自動的摘要製作會產生 1/5 到 1/4 的不好摘要。但若以協助人工摘要的角度來應用，則可大幅減輕人力的負擔，幾乎有六成的機會，使用者只要選第一句送出即可，而有七成五到八成的機會，使用者可從系統提示的候選句中，獲得相當不錯的摘要。

我們曾嘗試完全用人工摘要，然後跟自動摘要比較。為此，我們撰寫了嚴謹的摘要規範，如附件一，供摘要者遵循。初步的比較發現，有時人工摘要與自動摘要的結果不完全相同，但品質都達到相當好的效果，如附件二。直接以機器比對，會認為自動與人工摘要不同，而可能視自動摘要的結果較差。因此，目前還沒有跟人工摘要做比較。

由於摘要結果好壞的認定相當主觀，目前還很難由機器自動比對其效果，而必須仰賴人工評估。在評估成本極大的情況下，我們難以嘗試不同技術與參數，來做多方的比較。

除了 45 字與 69 字的限制外，使用者也可以指定其他接近的字數，例如 80、100 或 120 字，此方法也可以產生具有類似成效的結果。這意味著，其不僅可以運用於手機簡訊，也可以運用於一般新聞的自動摘要，提供使用者畫龍點睛且又具備內容資訊性的摘要。然而依照本方法的設計，若文件的標題或內文撰寫方式，不是簡潔的報導性敘述，例如社論、述評、股市漲跌表、分區氣象圖、條列項目等，其效果可能就會大受影響。

目前大多數的自動摘要方法，以「選句」為主，對於合成句子的「組句」方法，則較少討論。本文討論的合成方式，雖然依賴於事先既有的標題，但近幾年已有標題自動生成的研究，若先自動生成簡短的標題，例如五到十個詞之內的標題（自動生成短標題**應當**比生成長標題容易、效果好），再結合本文的方法，即可做到較高階的句子合成。若能多產生幾組候選標題，每個標題再結合本文方法產生較長句子，則可以產生資訊量較多的合成摘要。如此可將自動摘要方法，從「選句」推向到「組句」的階段。

誌謝：

本研究由威知資訊與國科會研究計劃補助，國科會計劃編號：NSC 93-2213-E-030-007-.

參考文獻：

[1] 王以瑾，”世界第一　台灣手機門號比總人口還多”，ETtoday.com，2002/08/09，http://www.ettoday.com/2002/08/09/339-1337800.htm, accessed on 2003/12/3.

[2] 聯合線上聯合新聞網 egolife 讓生活想像無限－簡訊快遞，http://udn.com/NASApp/LogFriend/UDNSMS/introduction_news.html，accessed on 2003/12/3.

[3] “如何訂閱 中央社股市新聞手機簡訊？”，http://www.suio.com.tw/top/can/can_order_txt.asp，accessed on 2003/12/3.

[4] 陳芸芸，“上網傳簡訊　「哈燒」不麻煩”，自由時報 http://www.libertytimes.com.tw/2003/new/jan/15/today-i1.htm，accessed on 2003/12/3.

[5] 吳顯申，”新加坡今推出手機簡訊中文新聞快訊服務”，中央社，2003-10-01 http://news.yam.com/cna/sports/news/200310/200310010295.html, accessed on 2003/12/3.

[6] 中時行銷：【縱橫網路】媒體挑戰：多元化傳播平台，http://marketing.chinatimes.com/item_detail_page/professional_columnist/professional_columnist_content_by_author.asp?MMContentNoID=4369，accessed on 2003/12/3.

[7] Chin-Yew Lin and E.H. Hovy, “The Potential and Limitations of Sentence Extraction for Summarization,” Proceedings of the Workshop on Automatic Summarization, post-conference workshop of HLT-NAACL-2003, Edmonton, Canada, May 31 - June 1, 2003.

[8] The Document Understanding Conference, http://duc.nist.gov.

[9] IBM, “Intelligent Miner for Text: Summarization Tool”, http:// www-3.ibm.com/software/data/iminer/fortext/summarize/summarize.html.

[10] InXight, http://www.inxight.com/

[11] A List of Summarization Projects, http://www.ics.mq.edu.au/~swan/summarization/projects_full.htm, accessed on 2003/12/3.
[12] The TIPSTER SUMMAC Text Summarization Evaluation : Final Report, Oct., 1999, http://www-nlpir.nist.gov/related_projects/tipster_summac/final_rpt.html, accessed on 2003/12/3
[13] Takahiro Fukusima, Manabu Okumura, and Hidetsugu Nanba, "Text Summarization Challenge 2: Text Summarization Evaluation at NTCIR Workshop3," Proceedings of the Third NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering, Oct. 8-10, 2002, Tokyo, Japan, pp.1-6.
[14] 陳信希，自動摘要方法之研究:單一中文文本之 摘要，行政院國家科學委員會研究計畫，2000，計劃編號：NSC89-2213-E002-064。
[15] 陳信希，多語言資訊檢索與擷取(II)---子計畫 IV：自動摘要方法之研究---多中文文本之 摘要，行政院國家科學委員會研究計畫，2001，計劃編號：NSC89-2218-E002-041。
[16] 陳信希，多語言資訊檢索與擷取(III)---子計畫 IV：自動摘要方法之研究---多語言文本之摘要，行政院國家科學委員會研究計畫，2002，計劃編號：NSC90-2213-E002-045。
[17] Hsin-Hsi Chen, June-Jei Kuo, Tsei-Chun Su, "Clustering and Visualization in a Multi-lingual Multi-document Summarization System," Proceedings of the 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14-16, 2003, pp.266-280.
[18] June-Jei Kuo, Hung-Chia Wung, Chuan-Jie Lin, Hsin-Hsi Chen, "Multi-document Summarization Using Informative Words and Its Evaluation with a QA System," Proceedings of the Third International Conference Computational Linguistics and Intelligent Text Processing, Mexico City, Mexico, February 17-23, 2002, pp.391-401.
[19] 張俊盛，可調式的文件摘要技術之研究(II)，行政院國家科學委員會研究計畫，2001，計劃編號：NSC89-2218-E007-015。
[20] 黃純敏，多語文(中英文)超文件自動摘要與評估，行政院國家科學委員會專題研究計畫成果報告，2001，計劃編號：NSC89-2416-H224-053。
[21] Michele Banko, Vibhu O. Mittal, Michael J. Witbrock, "Headline Generation Based on Statistical Translation," Proceedings of the ACL 2000.
[22] Paul E. Kennedy, Alexander G. Hauptmann, "Automatic title generation for EM," Proceedings of the fifth ACM conference on Digital libraries, 2000 , San Antonio, Texas, U.S., pp. 230-231.
[23] Simon Corston-Oliver, "Text Compaction for Display on Very Small Screens," In Proceedings of the Workshop on Automatic Summarization (WAS 2001), Pittsburgh, PA, USA.
[24] O. Buyukkokten, H. Garcia -Molina, A. Paepcke. 2001. "Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices," The 10th International WWW Conference (WWW10). Hong Kong, China.
[25] Christopher C. Yang, Fu Lee Wang, "Adapting content to mobile devices: Fractal summarization for mobile devices to access large documents on the web," Proceedings of the twelfth international conference on World Wide Web, May 2003, pp.215-224.
[26] 黃聖傑，"多文件自動摘要方法研究"，國立臺灣大學資訊工程學研究所碩士論文，1999 年 6 月。
[27] Yuen-Hsien Tseng, "Automatic Thesaurus Generation for Chinese Documents", Journal of the American Society for Information Science and Technology, Vol. 53, No. 13, 2002, pp. 1130-1138.
[28] Daniel Lopresti and Jiangying Zhou, "Retrieval Strategies for Noisy Text," Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval, April 15-17, 1996, pp. 255-269.

附件一：人工摘要規範表

```
<?xml version="1.0" encoding="Big5" ?>
<!DOCTYPE Summary SYSTEM "guideline.dtd">
<!--
  本文提供一份新聞與摘要的範例，並說明摘要的原則與遵守事項。
-->
<News>
<!--
   ID, Title, Body 均從原新聞文字得來的資訊，本來應該不必修改，但
 1.在 Title 中，如果有其他無義異的符號，如「・」，請將其刪除。
 2.在 Body 中，如果記者報導的前述句（如「中華日報記者程紹菖／台北報導」）
   跟原文沒有用句點「。」斷開來（或者用別的符號、空格斷開來），
   則請加入句點「。」將其斷開。謝謝。
-->
  <ID>chd_eco_19990112_0001</ID>
  <Title>
    政院否認房貸政策又轉彎。
```

```
  </Title>
  <Body>
<!--
    原句爲：
    中華日報記者程紹菖／台北報導針對媒體報導政府提撥的
    斷開後爲：
-->
    中華日報記者程紹菖／台北報導。針對媒體報導政府提撥的
    新台幣一千五百億元優惠低利購屋貸款，在行政院與立法
    院協議之下可能全數改爲不設限，不分購買新舊屋一體適
    用一事，行政院方面昨（十一）日否認相關報導，並表示
    ，行政院並沒有這樣的指示。

    一千五百億元的購屋低利貸款自昨日起開始辦理，其中一
    千兩百億元提供購屋者購置新成屋，另外三百億元則爲首
    次購屋貸款，但有媒體大幅報導，指郵政總局已接到行政
    院的指示，不論新、舊屋都可適用一千五億元貸款，房市
    貸款政策又將轉彎。

    據了解，對於這項報導，代理行政院長的劉兆玄曾與中央
    銀行總裁彭淮南連繫，證實這只是傳聞，行政院相關官員
    向郵政總局查詢時，郵政總局也表示不知有此事，行政院
    方面強調，行政院並沒有這樣的指示。
  </Body>
  <Analysis>
<!--
  Analysis 的部份，爲人工分析得來的資料。
  分析原則先從整理原文中的 proper name 開始，
  再依 5W1H 原則列出對應的 proper names 或簡短的原文描述。
  有了此兩步驟的準備工作後，才進行人工摘要，寫出要求字數的文句。
-->
    <ProperName>
<!--
  1. ProperName 分 人名、地名、機構名，只要原文中有出現就列出。
  2. 人名如果有職稱，則加在屬性裡。
  3. 地名、機構名的全名，列於 tag 之間，作爲 tag 值。
  4. 地名、機構名的簡稱，列於 tag 的屬性裡，作爲 tag 的屬性值。
  5. 如果原文裡只有全名，沒有簡稱，則依個人知識加入常用的簡稱。
  6. 如果原文裡只有簡稱，沒有全名，則依個人知識加入常用的全名。
  7. 如果不知其全名或簡稱，則用原文的名稱作爲 tag 的值，tag 的屬性值可省略。
  8. 如果同義異名詞超過兩個(全名及其簡稱)，則全部條列出來，
     並用原文裡最常用的全名爲 tag 值，而其他異名詞爲 tag 的屬性值。
-->
      <person title="中央銀行總裁">彭淮南</person>
      <person title="代理行政院院長">劉兆玄</person>
      <person title="記者">程紹菖</person>
      <organization abbreviation="">行政院</organization>
      <organization abbreviation="">立法院</organization>
      <organization abbreviation="央行">中央銀行</organization>
      <organization abbreviation="">郵政總局</organization>
      <organization abbreviation="">中華日報</organization>
      <location abbreviation="">台北</location>
    </ProperName>
    <who>
<!--
  在 who 中，通常從 ProperName 裡選出此篇主題的對象即可。
  1. 此對象（主角）可爲主動者（類似主詞）或被動者（類似受詞）。
  2. 如果有次要對象（類似配角），因爲也參與其中，也列於主角之後。
  3. 其他像電影中非主角、配角者，可以不列。(他們已列在 ProperName 中了)
  4. 原文中的同義異名詞可以全部列舉出來。
-->
      <name>行政院</name>
    </who>
    <what>
<!--
  在 what 中描述此主題之現象、事實、事件。
  1. 通常，列出此篇主題的詞彙、片語、或子句。
  2. 優先列詞彙、其次片語、其次子句。
  3. 詞彙、片語、子句以能描述現象、事實、事件爲原則。
```

```
  4. 詞彙、片語、子句盡可能從原文文字得來。
  5. 原文中的同義異名詞可以全部列舉出來。
-->
        <event>房貸政策</event>
        <event>低利購屋貸款</event>
        <event>房貸政策轉彎</event>
        <event>政院否認房貸政策轉彎</event>
      </what>
      <when>
<!--
  在 when 中，以時間日期格式，記載原文描述的主題的時間、日期。
-->
        <date>1999/01/11</date>
      </when>
      <where>
<!--
  在 where 中，通常，從 ProperName 裡選出此篇主題的地名即可。
  1. 地名爲此主題的發生地，若不容易判斷，則從 ProperName 中選出。
     若連 ProperName 裡也沒有（不太可能），則依個人知識加入。
     若無法加入則省略。
  2. 發生地不管有多少個，都全數列出。
-->
        <place>台北</place>
        <place>行政院</place>
        <place>立法院</place>
      </where>
      <why>
<!--
  在 why 中，條列簡述此主題發生的原因、緣起、由來等。
     並簡略說明結果。(有時從結果中亦可看出由來)
  1. 描述方式以從原文文字擷取必要段落爲主。
  2. 如果沒有原因，如單純的運動報導，則描述主題背景（不必從原文來）。
-->
        <reason>媒體報導房貸政策轉彎，政院否認</reason>
      </why>
      <how>
<!--
  在 how 中，條列簡述主題如何進行、進展、結果。
  1. 先從片語、子句描述起，如覺得不足，再多一點描述。
  2. 描述語句最好從原文拷貝修改而來。
-->
        <action>媒體報導房貸政策轉彎</action>
        <action>行政院並沒有這樣的指示</action>
        <action>
          媒體報導政府提撥的新台幣一千五百億元優惠低利購屋貸款，
          在行政院與立法院協議之下可能全數改爲不設限，不分購買
          新舊屋一體適用。
        </action>
        <action>
          代理行政院長的劉兆玄曾與中央銀行總裁彭淮南連繫，證實這只是傳聞
        </action>
      </how>


      <HumanSummary CharLimit="69" char="67">
<!--
  從前面整理的過程中，可得知這篇文章的內容，根據這些資訊，以人工方式
  寫下符合字數之摘要。但盡可能採用原文文字，稍加修改即可。
  1. 屬性 CharLimit="69" 爲此摘要的限制，其值可能爲 69 或 45。
     目前 45 暫不需要。
  2. 屬性 char 爲摘要後的字數，包含標點符號，可用 word 的字數統計功能獲得。
-->
      針對媒體報導政府提撥的新台幣一千五百億元優惠低利購
      屋貸款，可能全數改爲不設限，不分購買新舊屋一體適用
      一事，行政院表示並沒有這樣的指示。
      </HumanSummary>

      <HumanExtract1 CharLimit="69" char="63" CharModified="0">
<!--
```

僅從原文句、子句或小句拷貝組合而來，如有必要，僅作少許詞彙之修改，
如增、刪連接詞等，使文句通順、保留原意，並盡可能符合字數限制。

1. 不管有沒有增刪詞彙，此 tag 表示爲人工摘錄，但**「可以」做少許修改**。
2. 如果有做任何詞彙（如連接詞）的增、刪，則設定加總後的增刪字數於屬性 CharModified="字數" （中文字元的個數），若無，則 CharModified="0"。
3. 屬性 char 爲摘要後的字數，包含標點符號，可用 word 的字數統計功能獲得。
4. 請在註解裡，盡可能說明摘要原則。
   以本例子爲例，摘要原則：
   「從開頭第一句中，刪除中間數個小句而成，沒有修改任何詞彙」。
   所謂「句子」爲「。」、「？」或「！」斷開的敘述，其中被「，」、「；」或「：」斷開者稱爲「小句」。

-->

針對媒體報導政府提撥的新台幣一千五百億元優惠低利購屋貸款，在行政院與立法院協議之下可能全數改爲不設限，行政院並沒有這樣的指示。
</HumanExtract1>

<HumanExtract2 CharLimit="69" char="63">

<!--

僅從原文句、子句或小句拷貝組合而來，<u>不做任何修辭</u>，並盡可能符合字數限制。

請在註解裡，說明摘要原則。
以本例子爲例，摘要原則：「從開頭第一句中，刪除中間數個小句而成」。

-->

針對媒體報導政府提撥的新台幣一千五百億元優惠低利購屋貸款，在行政院與立法院協議之下可能全數改爲不設限，行政院並沒有這樣的指示。
</HumanExtract2>
</Analysis>
</News>

附件二：人工摘要與自動摘要比較表

原文	擺脫股東大舉申讓陰影　世界先進、力晶拚股價。 前陣子受制大股東大舉申報轉讓賣壓蓋頂，股價漲勢遠遜上市DRAM股茂矽、華邦的上櫃DRAM股世界先進及力晶半導體，**由於產業景氣復甦及營運情勢都明顯增強**，<u>原本投資信心動搖的原始股東，已決定暫緩調節</u>，使得力晶及世界先進股價連續飆漲了四根漲停板。 世界先進及力晶半導體兩檔店頭 DRAM 股，前陣子由於分別有華新麗華及新光紡織等原始投資大股東大舉申報轉讓持股的賣壓罩頂，再加上力晶半導體在〇.二五微米製程轉換上表現不順，世界先進的 64MDRAM 遲遲無法量產，使得市場買盤裹足，漲勢遠遠落後於上市 DRAM 股華邦及茂矽。 最近四個交易日，世界先進及力晶半導體頗有急起直追之勢，幾乎天天拉出漲停長紅。 一方面是今年初以來，DRAM 價格再度蠢揚，64MDRAM 美國現貨市場報價已突破十一美元前波高價，使得 DRAM 產業景氣復甦情勢增強，美光股價短短一週飆升四成。 另一方面，力晶半導體在去年十二月終於突破良率瓶頸，最終良率拉高穩固在七成水準，64MDRAM 單月產量並達到三百萬顆水準，已經超越華邦的二百五十萬顆，並且開始轉虧爲盈。 世界先進雖然 64MDRAM 尚未量產，但因爲 16MDRAM 市況同樣熱絡，世界先進持續增加投片，使得十二月營收大幅躍升至十三億六千萬元，該公司內部估算十二月單月盈餘可望超過一億元以上。 加上該公司的 64MDRAM 也準備開始大量投片生產，營運轉機也強勢浮升。 兩家公司基本面明顯增強，讓先前已漸失投資耐心的部份原始投資財團股東，投資信心又逐漸回升，決定暫緩調節腳步，力晶半導體及世界先進得以接連四根漲停長紅，分別突破二十一元與二十四元的前次現金增資溢價心理關卡。
人工摘要	<HumanExtract2 CharLimit="69" char="**61**">擺脫股東大舉申讓陰影世界先進、力晶拚股價，**因產業景氣復甦及營運情勢都明顯增強**，使得力晶及世界先進股價連續飆漲了四根漲停板。</HumanExtract2>
自動摘要	排序 1, 相似度=0.8304, 長度比例=0.9420, 共 **65** 字 擺脫股東大舉申讓陰影世界先進、力晶拚股價，<u>原本投資信心動搖的原始股東，已決定暫緩調節</u>，使得力晶及世界先進股價連續飆漲了四根漲停板。 排序 2, 相似度=0.69258, 長度比例=0.9420, 共 65 字 擺脫股東大舉申讓陰影世界先進、力晶拚股價，使得十二月營收大幅躍升至十三億六千萬元，該公司內部估算十二月單月盈餘可望超過一億元以上。 排序 3, 相似度=0.7146, 長度比例=0.8696, 共 60 字 擺脫股東大舉申讓陰影世界先進、力晶拚股價，世界先進的 64MDRAM 遲遲無法量產，使得市場買盤裹足，漲勢遠遠落後於上市 DRAM 股華邦及茂矽。 排序 4, 相似度=0.6918, 長度比例=0.6957, 共 48 字 擺脫股東大舉申讓陰影世界先進、力晶拚股價，分別突破二十一元與二十四元的前次現金增資溢價心理關卡。 排序 5, 相似度=0.4961, 長度比例=0.7391, 共 51 字 擺脫股東大舉申讓陰影世界先進、力晶拚股價，世界先進及力晶半導體頗有急起直追之勢，幾乎天天拉出漲停長紅。