應用語料庫和語意相依法則於中文語音文件之摘要

Spoken Document Summarization Using Topic-Related Corpus and Semantic Dependency Grammar

黃建霖　　謝嘉欣　　吳宗憲

國立成功大學資訊工程系

Email: {chicco, ngsnail, chwu}@csie.ncku.edu.tw

摘要

自動語音文件摘要技術，可應用於資訊的檢索、語意壓縮及資料記錄等方面。目前自動語音摘要存在幾個問題，首先是語音辨識準確率的提升，以及如何對語音內容萃取重要資訊、生成在句法及語意上合理的摘要結果。本論文提出一應用主題相關語料庫和語意相依法則於中文語音文件之摘要。首先，語音文件透過大詞彙連續語音辨識的方法，將語音辨識成文字，並獲得摘要單元斷點、音節以及詞等資訊。語音摘要部份，就語音本質從五個分數去分析，分別為：語音辨識信賴分數、詞重要性分數、語言分數、句法結構分數及語意相依法則分數，而後利用動態規劃搜尋演算法(dynamic programming algorithm, DP)獲得初步摘要結果。最後，為了使摘要語音串接輸出能具平滑特性，我們將摘要語音的有效語音段取出，計算語音頻譜特徵，考慮串聯單元彼此間的流暢度，挑選語音文件中重複的單元以串接生成摘要語音。由實驗結果得知，本研究所提出之自動語音摘要架構與人工摘要結果相比，能有效地萃取重要資訊，串接合成流暢的摘要語音。

1. 簡介

近年來電腦、電信網路、通訊與多媒體等資訊科技的成熟發展，政府為提升行政效率，投入大量人力物力從事資訊化，其中電子公文就是一個很好的例子。現今資訊科技進步，改變了人類溝通方式，也改變了知識的管理和傳承，以及資訊的散播和儲存，對人類社會產生革命性的影響。目前國立故宮博物院、國立歷史博物館等負責保存國家文物的機構，也積極地與產學界合作發展數位典藏計畫，將傳統文化創作的保存工作，利用新科技以資訊化的方式長久保存。此外不乏一般的大型企業、新聞傳播事業等，本身都保有大量累積的資訊。隨著科技的進步，資料型態可能已不再侷限於文字檔案，也包含各式的多媒體影音資料，如：圖片、聲音及影像等。因此許多學者專家研究如何編碼壓縮，研究體積更小、容量更大的儲存媒體，除此之外，文件檢索、摘要一直以來都是研究的重要主題。知識傳授，教育學習以及理念的傳播，透過語音表達是最自然而且直接的方法。自動語音摘要技術對語音資料做語意上的壓縮，目的在於依使用者需求，在大量的資料裡將無用多餘的資訊去除，保留具代表文章意涵的資訊並且自動建構出合乎文法及語意的內容。

自動語音文件摘要研究的主題在於語音辨識、摘要模型以及語音串接。語音辨識雖然仍存在有許多瓶頸，但由於過去學者的努力，已累積有相當的研究成果。目前中文語音辨識研究多以統計式模型的方法為主，應用隱藏式馬可夫模型(hidden Markov model, HMM)，來建立以音節或次音節為基礎的聲學模型，並配合多連語言模型的應用，可將大量連續語音做詞彙的辨識。

摘要部份可分為文字摘要及語音摘要，文字摘要研究主要在於分析文章結構、字詞重要性，一般常見的方法如：分析段落位置、句子長度、以詞頻和反轉文件頻率表示(term frequency * inverse document frequency, tf.idf)計算詞的重要性等[1][2]。相對於文字，語音摘要需要透過自動語音辨識，透過文字分析語意層面的意涵，因此辨識的好壞會對摘要結果產生影響，且因為語音特性像是音高、周期或能量等[3]，可提供音韻上的分析來決定重要語句的選擇。過去日本東京工業大學的研究，就對語音摘要提出了很好的基本概念，透過語音摘要參數的擷取，配合動態規劃搜尋演算法，找尋最佳的詞句組合[4][5]。但方法上缺乏對語意成分的分析理解，且對於關鍵詞的選擇上並不十分合理強健，因此，我們要提出應用語意相依法則和主題相關語料庫的方法於中文語音文件之摘要，同時分析文章中重要資訊的多寡來決定摘要比例，並且利用語音頻譜特性考慮串接的流暢性[6]。

2. 語音自動摘要

本論文提出的自動語音摘要方法，首先，在語音辨識方面，利用最小錯誤鑑別訓練的方法來鑑別容易渾淆的模型，提高語音辨識的正確率[7]。摘要的部份，從五個層面考慮摘要的生成：第一、考慮文章中重要語意的保留，我們使用一組新聞語料知識庫，透過潛藏語意分析找出具代表性的重要文字。第二、語音

辨識正確率會影響文章語意的判斷，為了避免誤判情形的發生，經由計算辨識信賴分數，取辨識可信度較高的詞作為摘要。第三、語音摘要詞與詞之間的串連問關係，可利用語言模型建立。第四、分析文章語意相依的關係，建構合理的語意修飾關聯。第五、配合機率式文法規則，使句子具有文法規則，易於閱讀理解。最後以動態規劃搜尋方法，產生最佳的摘要詞組。此外，為了使串接語音平滑輸出，在串接摘要單元時，必須考慮串接流暢和平滑的程度。因此，在摘要單元串接的選擇上，我們考慮頻譜特性：分別使用頻譜中心(spectral centroid)、頻譜滑動(spectral rolloff)、頻譜變遷(spectral flux)、時域上越零率(time domain zero crossing, ZCR)和梅爾倒頻譜參數(Mel-frequency ceptral coefficient, MFCC) ，找出相鄰差異最小的串接單元以生成平滑之語音輸出。

然而，如何才能從語音文件中萃取出重要的詞句，建構出能夠代表文章意函的內容。本論文分別從語音聲學(acoustics)、語言學(linguitic)，句法(syntax)和語意(semantics)等方向去解決自動語音文件摘要可能面對的問題，一篇語音文件透過特徵參數的計算，可被分析成五個主要的特徵分數，包含有：(1) 語音辨識信賴 (confidence measure, $C(w_m)$) 分數；(2)字詞相對於文章所代表的重要性 (word significance, $R(w_m)$) 分數；(3)語言學結構相鄰 (linguistic trigram, $L(w_m \mid w_{m-2}, w_{m-1})$) 分數；(4)語意相依法則 (semantic dependency grammars, $B_{SDG}(w_{m-1}, w_m)$) 分數；以及(5)機率式文法規則 (probabilistic context-free grammars, $P(S)$) 分數。因為分數值域大小並不相同，所以我們分別計算分數的最大值 Max_{score} 和最小值 Min_{score}，依其不同值域對各分數 X_{score} 做正規化 $(X_{score} - Min_{score})/(Max_{score} - Min_{score})$ 將每一個分數值調整為從 0 到 1 之間。語音文件經過大詞彙連續語音辨識，產生一篇詞長為 M 的轉譯文件 $X = \{w_1, w_2, w_3, ..., w_{M-1}, w_M\}$，辨識資訊包含有次音節的語音斷點資訊。根據摘要比例，系統最後可以獲得長度為 $N = M \times Percentage$ 的摘要結果 $Y = \{w_1, w_2, w_3, ..., w_{N-1}, w_N\}$。

摘要流程如(圖 1)所示，分成下列四個步驟：首先就辨識結果將 stop word 去除，例如：的、及、了等，不具語義表示的詞。再者，因為不同的語音文件可能包含的重要資訊量並不一致，所以摘要壓縮的比例會對摘要結果有很大的影響。因此除了可以依據使用者需求設定摘要比例外，也可以藉由判斷字詞相對於文章所代表的重要性 $R(w_m)$，自動決定摘要比例。第三步驟，則是將語音辨識信賴分數、重要詞語分數、語言學分數和語意相依分數等四種分數作結合，以動態規劃搜尋的方法，尋找可能的串接詞組。

$$S(Y) = \sum_{m=1}^{M} \{\lambda_C C(w_m) + \lambda_R R(w_m) + \lambda_L L(w_m \mid w_{m-2}, w_{m-1}) + \lambda_B B_{SDG}(w_{m-1}, w_m)\} \quad \text{(式 1)}$$

其中，$\lambda_C, \lambda_R, \lambda_L$ 和 λ_B 是代表各個特徵參數的權重(weight)，用以結合這四個分數並且平衡各參數的重要性。

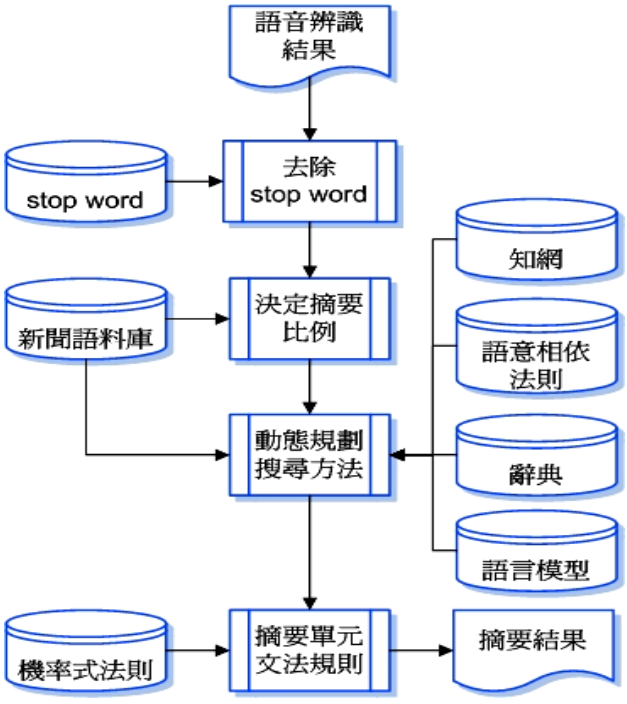

圖 1. 自動語音文件摘要程序

2.1 語音辨識信賴分數

語音摘要需要透過辨識器得到語言上的資訊，但語音辨識可能會產生聲學上和語言學上的辨識錯誤，擾亂最後摘要結果的意義。因此，我們將語音辨識的信賴分數$C(w_m)$引入，目的在於選擇辨識較正確的結果，作為判斷選擇摘要單元的分數之一。統計式語音辨識是基於貝氏法則，信賴分數是估算語音辨識中，給定一串觀測語音序列$X_t = X_1,...,X_t$對於一字串$W_m = W_1,...,W_m$，計算其事後機率(posterior probability) $p(w_m \mid x_t)$。辨識的階段中，我們期望能夠得到最大的事後機率值，也就是能夠有較小的誤差，所以可得下列式子：

$$C(w_m) = \max\, p(w_m \mid x_t) = \max\, p(x_t \mid w_m) \cdot p(w_m) / p(x_t) = \max\, p(x_t \mid w_m) \cdot p(w_m) \qquad \text{(式 2)}$$

其中，$p(w_m)$為語言模型的機率。$p(x_t \mid w_m)$為聲學模型的機率。$p(x_t)$為觀測到聲學特徵的機率。

2.2 重要詞分數

要對辨識結果做語音文件摘要的處理時，首先需要將語音文件內屬於重要的詞語保留下來，而把不具備有表達文章意義的詞與字抽離。我們引用一組標題本文互相對照的新聞語料庫，來輔助判斷辨識結果的詞句是否具有代表性。實驗從公共電視新聞收集 2001 到 2002 年的新聞，整理兩千零六則的新聞報導語料。為了檢索出與摘要文章內容相似的新聞報導語料，我們參照資訊檢索的技術(Information Retrieval, IR) [1]，首先將平行語料的所有文章內容，分別轉換成以詞v_d^w和音節v_d^s為單元的兩個向量，對於所要摘要的語音文件也同樣地做轉換為兩個向量，可表示成$v_d^w = (t_d^{w_1}, t_d^{w_2}, ..., t_d^{w_P})$和$v_d^s = (t_d^{s_1}, t_d^{s_2}, ..., t_d^{s_Q})$。其中，$Q$表示以音節單元為基礎的向量$v_d^s$維度，依據四百零二個中文音節，並考慮詞長為二的所有配對組合，產生維度為$Q = (402 + 402 \times 402) = 162006$的向量。而$P$則表示以詞為基礎的向量$v_d^w$維度，根據辭典內所定義的詞，不考慮虛詞(stop word)的部分，因為虛詞不會影響文章內容意義的檢索，去除用以降低計算的維度，得到結果$P = 28000$。兩向量內的之數值以詞頻和反轉文件頻率表示(term frequency * inverse document frequency, tf.idf)[8]。同時，必須考慮語音辨識$C(w_j)$可能造成的影響，將辨識不好的結果，減低分數。因此每一個索引值的計算方法如下：

$$t_d^{w_j} = C(w_j) \cdot \ln(f_{w_j} + 1) \cdot \ln(N / (df_{w_j} + 1)) \qquad \text{(式 3)}$$

結合兩向量來做文件查詢，利用向量內積的計算，查詢平行語料內所有文章的關聯$R(q,d)$，

$$\begin{aligned} R(q,d) &= \alpha_R \cos(v_q^w S^w, v_d^w S^w) + (1-\alpha_R)\cos(v_q^s S^s, v_d^s S^s) \\ &= \alpha_R \cdot (v_q^w S^{w2} v_d^{wT}) / (\left\| v_q^w S^w \right\| \cdot \left\| v_d^w S^w \right\|) + (1-\alpha_R)(v_q^s S^{s2} v_d^{sT}) / (\left\| v_q^s S^s \right\| \cdot \left\| v_d^s S^s \right\|) \end{aligned} \qquad \text{(式 4)}$$

並且應用參數$\alpha_R = 0.2$來平衡字與音節兩個向量的權重。依據此關聯分數$R(q,d)$，找出一篇文件描述的新聞事件最接近的文章$d^* = \arg\max_d R(q,d)$，之後，以潛藏式語意分析索引使用向量空間的方法[8]，搜尋辨識句子的詞與平行語料標題內的詞，兩者之間存在的關係。

方法說明如(圖 2)所示，首先根據平行語料和辭典，建立一個文章及詞的二維矩陣$A_{t \times d}^w$，維度為2006×5104。經由詞對應於文章以及文章對應詞的關係$(terms \times documents) \cdot (documents \times terms)$，最後可以推導出詞對詞的關聯$AA^T = terms \times terms$。配合奇異值分解方法來達到維度的降低，將共同發生的事件投影到相同的維度上。透過奇異值分解$A_{t \times d} = U_{t \times n} S_{n \times n} (V_{d \times n})^T$，將矩陣分解成三個矩陣$U_{t \times k}$，$S_{k \times k}$和$(V_{d \times k})^T$，其中$n = \min(t,d)$。

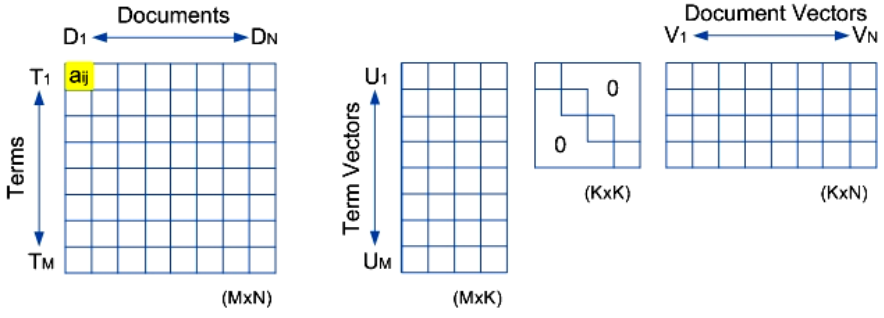

圖 2. 奇異值分解

將取對角矩陣累計變異量之百分之九十作為維度降低的依據$k<n$。矩陣中每一個成分的值，用$tf \times idf$代表。詞對詞的矩陣透過降維度的資訊，來計算$AA^T = US^2V^T$。透過新的詞對詞矩陣便可以得知兩個詞的相似性$P_{LSI}(w_i, w_j)$，其分數計算方法如下：

$$P_{LSI}(w_i, w_j) = \cos(U_i^w S^w, U_j^w S^w) = U_i^w S^{w2} U_j^{wT} \Big/ \left\| U_i^w S^w \right\| \bullet \left\| U_j^w S^w \right\| \quad \text{(式 5)}$$

最後歸納上述的步驟，透過下列程序的計算方法，我們可以從平行語料中萃取重要的資訊$R(w_i)$。其中w_j^{t*}代表平行語料標題內的詞，而W_i是輸入文章經語音識辨後的詞，因此可計算出W_i對於摘要文件的重要性：

$$R(w_i) = \max_j \{P_{LSI}(w_i, w_j^{t*}) \cdot (f_{w_i} + 1) \cdot \ln(N/(df_{w_i} + 1))\} \quad \text{(式 6)}$$

2.3 語言學結構相鄰分數

我們利用三連語言模型$L(w_3 \mid w_1, w_2) = F(w_1, w_2, w_3)/F(w_1, w_2)$，建立詞與詞之間相接的情況，使摘要最後結果更加符合語言學結構[9]。其中$F(\bullet)$表示 frequency count。但為了避免許多詞統計不到 trigram 情況發生，引用 Jelinek et al.所提出的平滑方法(N gram smoothing)[10]，內插 trigram, bigram 和 unigram 等相關機率值。表示方法如下：

$$L(w_3 \mid w_1, w_2) = p_1 \cdot F(w_1, w_2, w_3)/F(w_1, w_2) + p_2 \cdot F(w_1, w_2)/F(w_1) + p_3 \cdot F(w_1)/\sum F(w_i) \quad \text{(式 7)}$$

其中，$p_1 + p_2 + p_3 = 1$表示正數的權重且合為一。

2.4 語意相依法則分數

前面所言，利用重要詞的分數，找到一堆對於文章具有代表性的詞組，並且配合語言學結構相鄰的分數，挑選彼此具有高度相鄰關聯性的詞組。但是這樣的資訊，對於生成一則合理且完整的摘要語句，並不足夠。基於語言學的考量，句子應具備有語法(grammar)和句法上的關係，因此我們對中文語法結構做分析，利用統計機率方法，計算機率式文法規則。語意學研究是字意和句意的描述，藉由語意特徵的探討，可以幫助釐清彼此本質上的意涵。舉例而言，"這顆蘋果(NP) 吃了(V) 那個男人(NP)"的句子可能會令人難以置信。由此可知在語意上，這個句子並不合理，但這並不是因為句法結構所造成的問題。因此，我們引入語意相依法則，從句法和語意相依的概念來解決此問題。

語意相依法則(semantic dependency grammars, SDG)，是透過剖析器(parser)將輸入的詞句，剖析出樹狀的詞性架構，並標記出中心詞(head)所在的位置。以中心詞為基準，考慮其它詞與中心詞的關係。剖析器是將詞句透過斷詞，找到相對應的詞性序列，並且利用動態規劃搜尋的方式，配合機率式文法規則模型，建立對應的語法分析樹和其機率。參考 Collin 在 1996 年提出的相依模型[8]，輸入一句子$\boldsymbol{S}$，可剖析成文法樹結構t，可表示為機率$P(t \mid s)$。並可剖析成 $\boldsymbol{B}$ 個詞(terms of parsing tree)並存在有詞與詞相依的關係 $\boldsymbol{D}$ (dependency relation)。表示如下：

$$P(t \mid s) = P(B, D \mid s) = P(B \mid s) \times P(D \mid s, B) \quad \text{(式 8)}$$

假定每一個相依關係都是獨立的，且剖析後每一個詞W_m，都相依於某一個中心詞h_{w_m}，其相依的關聯可以界定為$R_{w_m h_{w_m}}$。因此，相依關聯可以重新定義成一個集合$\{d(w_i, h_{w_i}, R_{w_i, h_{w_i}})\}$，表示如下：

$$P(D \mid s, B) = \prod_{j=1}^{n} P(d(w_j, h_{w_j}, R_{w_j, h_{w_j}})) \quad \text{(式 9)}$$

在計算兩個詞W_i和W_j，存在相依關係R的機率$F(R \mid w_i, w_j)$時，同一關係可表示如下：

$$F(R \mid w_i, w_j) = C(R, w_i, w_j)/C(w_i, w_j) \quad \text{(式 10)}$$

其中，$C(w_i, w_j)$表示為兩個詞一起出現的頻率，$C(R, w_i, w_j)$表示兩個詞一起出現時存在有的相依關係。且為了避免資料稀疏(sparse data)的問題，進一步地利用知網的知識，將詞轉換成相對應的上位詞(hypernym)，以表示之$H(\bullet)$，得到下列式子：

$$F(R \mid H(w_i), H(w_j)) = C(R, H(w_i), H(w_j))/C(H(w_i), H(w_j)) \quad \text{(式 11)}$$

舉例而言，一句中文"我們遊覽台灣各個景點"，經過斷詞並且對應到相關的上位詞，和中研院 Treebank 內建立的語意關係，配合中研院詞庫小組所提的「中心詞主導原則」(dead-driven principle) [11]，最後可以建

構出如(圖 3)的語意相依網路，得到”我們(first person) 遊覽(tour) 台灣(place) 各個(qValue) 景點(attribute)”。

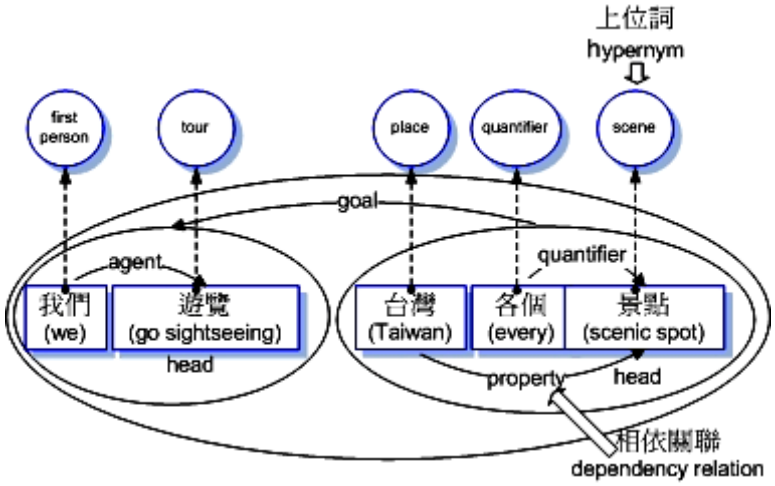

圖 3. 語意相依關聯範例

表 1. 中心詞主導原則

1. 句子(S)和述詞片語(VP)的中心詞皆為述詞(V) 2. 名詞片語(NP)的中心語為名詞(N) 3. 介詞片語(PP)的中心語為介詞(P) 4. 方位詞片語(GP)的中心語為方位詞(Ng) 5. 對等連接詞(XP)的中心語為連接詞(C) 6. XP 的詞類由連接成份決定，連接成份為述詞片語(VP)，則為述詞片語， 連接成份為名詞片語，則為名詞片語(NP)。 - S、VP 的中心語是述詞 - NP 多半以多右方的名詞為中心語 - PP 以介詞為中心語，其論元角色是 DUMMY，成雙岔結構 - GP 以 Ng 為中心語，其論元角色是 DUMMY，成雙岔結構

語意相依法則目的是建立詞組間語意關聯，即使詞組不相鄰，亦可得知詞與詞在語意上修飾的關係。實際上，利用 HowNet 以及統計訓練好的語意相依機率。輸入一詞組，利用 HowNet 將其推展到上位詞的型態[12]，然後判斷詞組間是否有相依的關連。語意相依法則和機率式文法規則的訓練流程如(圖 4)所示：

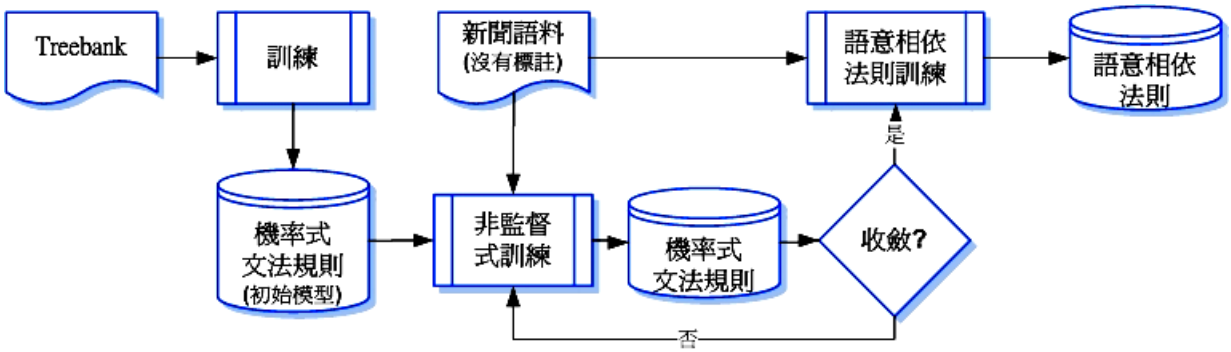

圖 4. 語意相依法則及機率式文法規則訓練流程

我們利用 Treebank 及公視新聞語料進行非監督式的訓練。

$$B_{SDG}(w_a,w_b)=\frac{1}{N_s}\sum_{j=1}^{N_s}\sum_{i}\sum_{k} f_{DR_i^k(w_a,w_b)}(T_i,S^j(w_a,w_b))\times f_{T_i}(S^j(w_a,w_b)) \quad \text{(式 12)}$$

其中，f_{T_i} 表示文法剖析 PCFG 之機率。 $f_{DR_i^k}$ 表示語意相關法則之機率 N_s 表示句子總數。 S^j 表示一個句子包含有 W_a 和 W_b。 T_i 表示針對句子 S^j 可能剖析的文法樹。 k 表示存在的關連性索引。 $D_i = \{DR_i^k(w_a, w_b) | 1 \le k \le N_w - 1\}$ 指長度 N_w 的句子存在相依關係。 考慮訓練語料稀疏的問題(sparse data)，因此使 HowNet 內定義的上位詞(Hypernym)取代原本的詞組：

$$f_{DR_i^k(w_a,w_b)}(T_i, S^j(w_a, w_b)) \cong f_{DR_i^k(H(w_a),H(w_b))}(T_i, S^j(w_a, w_b)) \quad \text{(式 13)}$$

以(圖 3)為例，S^j 為:"我們遊覽台灣各個景點"，$H(\bullet)$ 表示推演到上位詞，如：台灣 $\rightarrow$ place。$f_{DR_i^k(H(w_a),H(w_b))}$ 指 w_a, w_b 存在相依關係 DR_i^k，如：各個 $\underset{quantifier}{\rightarrow}$ 景點。最後，參照(式 11)，(式 13)可由下式計算：

$$f_{DR_i^k(H(w_a),H(w_b))}(T_i, S^j(w_a, w_b)) = F(R^k | H(w_a), H(w_b)) / \sum_{u=1,\, u\neq a}^{N_w} \sum_{v} F(R^v | H(w_a), H(w_u)) \quad \text{(式 14)}$$

在摘要的第三個步驟中動態規劃搜尋程序，每次以兩個詞作為輸入，直接索引在此訓練好的語意相依法則機率值。

2.5 動態規劃搜尋方法

以二維圖形說明動態規劃搜尋方法如(圖 5)所示，橫軸是摘要後的結果，每一個節點表示為一個詞，計算每個節點的分數，並儲存累計分數和回溯路徑指標。縱軸是語音辨識後的結果共有十個詞，經過摘要後為橫軸剩下五個詞。

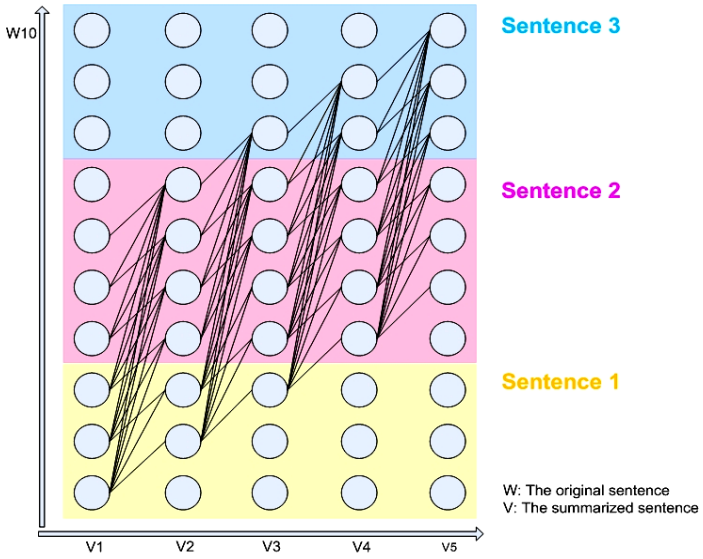

圖 5. 語音摘要使用動態規劃搜尋方法之示意圖

3. 摘要單元串接合成

在挑選最佳的摘要單元之後，為了使摘要的原音重現，可能將原本並非屬於同一時間，也就是非連續發音的語音片段串接合成。不過，如此單元串接可能會影響語音合成後音檔的品質，如聽覺上中斷、跳音、摩擦等不連續情況。因此如何能夠從原本的音檔中，挑選出最適合作為串接的語音片段，使得整體語音可以有流暢平滑的表現。我們考慮語音在頻譜上的特性，並參考[6]中對語音所定義的特徵參數，作為摘要單元選擇的評量依據，以達到最佳平滑程度，串接出自然語音。分別求取五個特徵參數。包含有，頻譜中心(***SC***)、頻譜滑動(***SR***)、頻譜變遷(***SF***)、時域上越零率(***ZCR***)和梅爾倒頻譜參數(***MFCC***)等。將此參數整合之距離定義如下：

$$SSP(w_i, w_j) = \min\{SC(w_i, w_j) + SR(w_i, w_j) + SF(w_i, w_j) + ZCR(w_i, w_j) + MFCC(w_i, w_j)\} \quad \text{(式 15)}$$

如(圖 6)所示，新聞內容經過斷詞以摘要單元為基礎，配合摘要結果來挑選新聞語音內所有的候選語音片段，建立一個詞網格。然後，利用動態規劃搜尋的方式，找到最佳的串接語音。由(圖 6)可知，摘要結果共選出六個摘要詞，其中"耶誕節"及"消費"在原本語音中共出現三個可挑選的串接候選，因此，我們利用動態規劃搜尋的方式串接語音。

新聞內容	歡迎回到新聞現場，來看今年的耶誕消費市場， 每年耶誕節都是美國的消費旺季，而最近幾年， 台灣人過耶誕節的氣氛也越來越濃， 因此耶誕相關的商品消費也跟著旺了起來， 儘管今年台灣籠罩在不景氣的陰影之下， 耶誕節的商機還是很驚人。
斷詞結果	歡迎 回到 新聞 現場，來看 今年 的 耶誕 消費(2-1) 市場(3-1)， 每年 耶誕節(1-1) 都 是 美國 的 消費(2-2) 旺季，而 最近 幾年， 台灣人 過 耶誕節(1-2) 的 氣氛 也 越來越濃， 因此 耶誕 相關 的 商品 消費(2-3) 也 跟著 旺 了 起來， 儘管 今年 台灣(4-1) 籠罩 在 不景氣 的 陰影 之 下， 耶誕節(1-3) 的 商機(5-1) 還是 很 驚人(6-1)。
摘要結果	耶誕節(1) 消費(2) 市場(3) 台灣(4) 商機(5) 驚人(6)

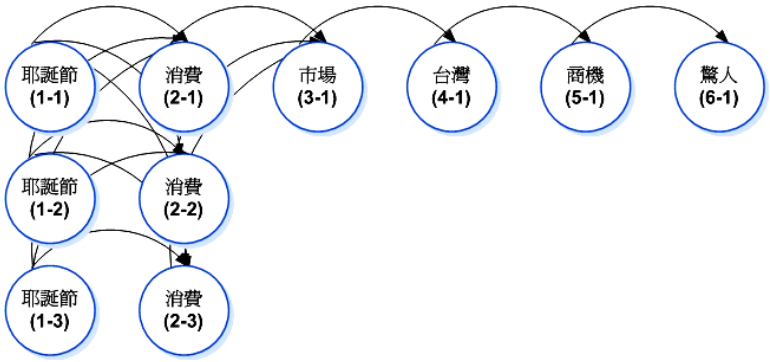

圖 6. 摘要語音串接示意圖

1) 頻譜中心，音訊經過短時域傅立葉轉換，取其頻譜的能量中心。頻譜中心可量測頻譜上特徵，頻心高代表著亮度高、頻率高的訊號。

$$SC(w_i,w_j)=\left\|SC(w_i)-SC(w_j)\right\| \ ; \ SC(w_i)=\frac{1}{F}\times\sum_{t=1}^{F}((\sum_{n=1}^{N}M_t[n]\times n)/(\sum_{n=1}^{N}M_t[n])) \qquad \text{(式 16)}$$

其中，$M_t[n]$ 傅立葉轉換強度；n 頻框索引；t 音訊分頁索引。

2) 頻譜滑動，同樣表示頻譜上特徵，可測量兩單元間的差異，$SR(w_i)=\frac{1}{F}\sum_{t=1}^{F}(0.85\times\sum_{n=1}^{N}M_t[n])$。

3) 頻譜變遷，正規劃相鄰頻譜的平方差，目的在於量測頻譜上的局部變化，$SF(w_i)=\frac{1}{F}\sum_{t=1}^{F}\sum_{n=1}^{N}(N_t[n]-N_{t-1}[n])^2$。其中，$N_t[n]$ 定義在第 t 音訊分頁的正規化傅立葉強度。

4) 時域上越零率，一般用於噪音偵測，在此可知兩單元間，噪音改變程度。$ZCR(w_i)=\frac{1}{F}\sum_{t=1}^{F}(\frac{1}{2}\sum_{n=1}^{N}\left|sign(s_t[n])-sign(s_t[n-1])\right|)$。

5) 梅爾倒頻譜參數，應用語音辨識常用的梅爾倒頻譜參數，共取三十九維，主要是模擬人的聽覺模型，$MFCC(w_i)=\frac{1}{F}\sum_{t=1}^{F}mfcc(f_t)$。

4. 實驗評估

4.1 語音辨識評估

實驗用的摘要語料，收錄自公視晚間新聞共 120 小時，根據標記檔案，取出主播部分四小時三十分鐘，其中三小時做為訓練語料，約 328MB；剩下約一小時三十分鐘，255 則新聞報導作為測試語料，約 166MB。分別計算音節、母音和字元的正確率，正確率的計算有，正確率(accuracy)、插入錯誤(insertion)、刪除錯誤(deletion)以及替換錯誤(substitution)，並且考慮前 N 名辨識結果。其計算式如下：

$$P_{accuracy} = W - I - D - S/W \qquad \text{(式 17)}$$

其中，***W*** 為辨識結果，總字元長度。***I*** 為比較較正確結果多辨識出的字，屬於插入錯誤，***D*** 為比較正確結果少辨識到的字，屬於刪除錯誤。***S*** 為比較正確結果，辨識錯誤的字，屬於替換錯誤。音節正確率為有百分之八十三，字元辨識率約為百分之八十，分析如(表 2)：

表 2. 公視新聞測試語料之正確率

---------- Syllable Results----------				
	ACCURACY	INSERTION	DELETION	SUBSTITUTION
Syllable ,Top 1:	**83.20%**	2.98%	2.03%	11.79%
Syllable ,Top 5:	87.50%	3.09%	2.13%	7.28%
Syllable ,Top 10:	89.02%	3.20%	2.25%	5.53%
---------- Character Results----------				
	ACCURACY	INSERTION	DELETION	SUBSTITUTION
Characters	**80.38%**	2.92%	1.94%	14.76%

4.2 摘要效果評估

利用資訊檢索方式來評估，與原本辨識結果做比較，看是否摘要後結果，能夠充分保留原新聞報導的要旨。隨機選取二十組詞彙作為查詢，依 2.2 節所述之向量模式對測試語料做檢索。由於檢索資料庫數量不大，對於各查詢詞彙所檢索到的文件並不多，因此只取出前十名分數最高的檢索結果。計算其 mean average precision (mAP)[13] 和 raw average precision (rAP)[13]：

$$\text{mAP} = \frac{1}{N_q}\sum_{i=1}^{N_q}\frac{1}{N_i}\sum_{k=1}^{N_i}\frac{k}{rank_{ik}} \quad ; \quad \text{rAP} = \frac{1}{N_q}\sum_{i=1}^{N_q}\frac{N_i}{N} \qquad \text{(式 18)}$$

其中，N_q：查詢的問句數。N_i：對於 q_i 的查詢結果，共有幾篇相關文章。$rank_{ik}$：對於 q_i 的查詢結果，排序第 k 篇相關文章。mAP 可以分析查詢結果，是否有正相關性，也就是前面的文章是相關的，而後面的文章可能相關性較低，mAP 曲線若無跳動的情形，則表示評估查詢的效果好，反之亦然。rAP 則可以判斷在第幾篇文件，文章對於查詢結果相關度的降低。由(圖 7)觀察得知，當摘要比例越高則所含的資訊越高，也就是資訊壓縮越小則語意保留程度越高。但是，當我們做 30%的摘要時，所檢索的前四篇文件與摘要 70%和 50%時的結果很相近。

另外，將測試音檔做人工的摘要記錄後，與自動摘要結果相對照，分別計算其正確率、插入錯誤、刪除錯誤以及替換錯誤等，如(式 17)。同時，實 0 驗各種知識庫所代表的重要程度，以(C_L_W_S)分別代表語音辨識信賴分數、語言學分數、詞重要性分數和語意相依法則分數，考慮各種情況如下圖所示：

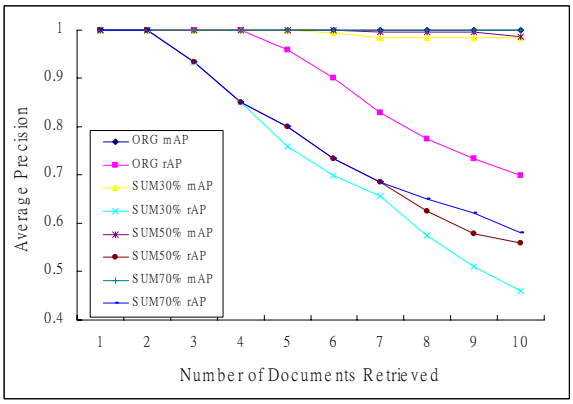

圖 7. 重要資訊檢索的結果

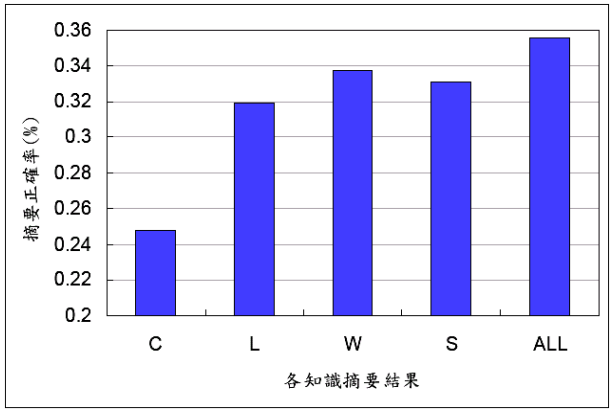

圖 8. 摘要之各分數重要性評估

由實驗結果(圖 8)可知，利用求取關鍵詞的作法(word significance score)效果最為顯著，其次為語意相依法則、三連語言模型，最後是語音辨識信賴分數。

ALL 代表結合四種分數所得到的摘要結果，依據各種知識所代表的重要性程度，設定其權重分別為 C(0.1)、L(0.2)、W(0.4)和 S(0.3)，評估正確率 accuracy 為百分之三十五。詳細的實驗結果如(圖 9)所示。由(圖 9)得知，摘要錯誤較常發生在插入錯誤，其次為替換錯誤和刪除錯誤。由此可知摘要結果的好壞，主觀因素影響較大，插入和替換錯誤較容易發生。

4.3 串接效果評估

串接語音的實驗可由(圖 10)表示，請十位受測者分別針對不同摘要比例評比。受測者先看過原始標準報導，並聆聽報導內容之後。比較摘要後的文字結果和聆聽語音串結效果，是否能表達報導文意及合成語音是否流暢，評比一到十分數，代表從劣到優的表現效果。

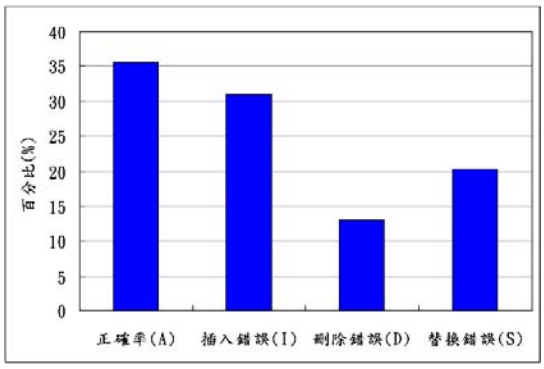

圖 9. 摘要結果正確率評估

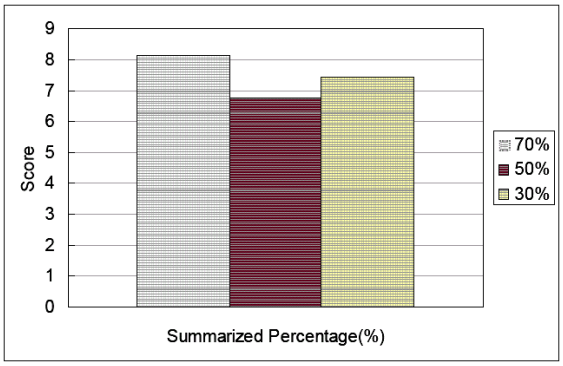

圖 10. 摘要串接及合成結果評估

5. 結論及未來展望

本論文提出新聞語料庫及語意相依法則於中文語音文件摘要，同時對語音串接單元計算頻譜上的特徵參數，利用動態規劃搜尋方法，生成一個兼具語意壓縮和聽覺效果流暢的摘要結果。分析摘要語音文件的聲學、語意和語法等特徵，結合語音辨識信賴分數、詞重要性分數、語言學分數、句法結構分數及語意相依法則分數。摘要單元串結從頻譜上取五項特徵參數，頻譜中心、頻譜滑動、頻譜變遷、越零率以及梅爾倒頻譜參數，決定最佳的語音串接。目前在八成的語音辨識率下，實驗證實系統可以做到良好的語意擷取保留，以及流暢的摘要語音效果。

語音摘要的目的，旨在壓縮語音文件，取出具代表性內容，並且能流暢地將語音串接輸出。以此研究為基礎展望未來，可藉由聯合各種方法，探討如何改善摘要效果：

1) 從摘要語音可分為文體規範式語音和自然口語式語音兩大類。其中，文體規範式語音是指語音內容有事先經過設計，表達內容與書本或文章的格是相近，像是新聞報導。而自然口語式語音則指語音內容無經過設計，表達內容是臨時思考應對，像是對話、訪談等。
2) 分析文章語意，進一步探討應用 Ontology 於摘要。
3) 以新聞語音為例，可將新聞分類並依照不同的新聞類別，抽取出具代表性的關鍵詞，或建立不同新聞類別的句法結構模組，以輔助摘要生成。
4) 分析語音聲學上特性，如：音高、週期和能量等。
5) 藉由網際網路的幫助，可分析因為時間的推進，所產生的新詞、文章用法的表達，和各領域的知識等。

誌謝

感謝國科會支持本研究計畫，計畫編號 NSC90-2213-E-006-088。

參考文獻

[1] Berlin Chen, Hsin-min Wang, Member, IEEE, and Lin-shan Lee, Fellow, IEEE, "Discriminating Capabilities of Syllable-Based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese," IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 10, NO. 5, JULY 2002

[2] Julian Kupiec, Jan Pedersen and Francine Chen, "A Trainable Document Summarizer", Xerox Palo Alto Research Center

[3] Kiyonori Ohtake, Kazuhide Yamamoto, Yuji Toma, Shiro Sado, Shigeru Masuyama,and Seiichi Nakagawa, "NEWSCAST SPEECH SUMMARIZATION VIA SENTENCE SHORTENING BASED ON PROSODIC FEATURES", Toyohashi University of Technology, Japan

[4] Chiori Hori, Member, IEEE, and Sadaoki Furui, Fellow, IEEE, "A New Approach to Automatic Speech Summarization," IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 5, NO. 3, SEPTEMBER 2003

[5] Furui, S.; Kikuchi, T.; Shinnaka, Y.; Hori, C., “Speech-to-Text and Speech-to-Speech Summarization of Spontaneous Speech,” Speech and Audio Processing, IEEE Transactions on , Volume: 12 , Issue: 4 , July 2004, pp. 401 – 408

[6] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” IEEE Transactions on Speech and Audio Processing, vol. 10, No. 5, July 2002.

[7] Biing-Hwang Juang, Fellow, IEEE, Wu Chou, Member, IEEE, and Chin-Hui Lee, Fellow, IEEE, "Minimum Classification Error Rate Methods for Speech Recognition," IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 5, NO. 3, MAY 1997

[8] Christopher D. Manning and Hinrich Schutze, "Foundations of Statistical Natural Language Processing", The MIT Press, 1999

[9] Manhung Siu, Member, IEEE, and Mari Ostendorf, Senior Member, IEEE, "Variable N-Grams and Extensions for Conversational Speech LanguageModeling", IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 8, NO. 1, JANUARY 2000

[10] F. Jelinek and R.L. Mercer, “ Interpolated Estimation of Markov Source Parameters From Sparse Data,” Pattern Recognition in Practice, E.S. Gelsema and L.N. Kanal, Eds., North-Holland Pub. Co., Amsterdam, pp. 381-397, 1980

[11] http://rocling.iis.sinica.edu.tw/

[12] HowNet. http://www.keenage.com/

[13] M. Banko, V. Mittal and M. Witbrock, “Headline generation based on statistical translation, “ in Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, 2000, pp. 318-325.