利用小波聽覺分頻處理與訊號子空間分解於車內噪音消除*

王駿發[1] 楊宗憲[2] 張凱行[3]

國立成功大學電機工程研所

wangjf@csie.ncku.edu.tw[1] chyang@icwang.ee.ncku.edu.tw[2]

casey019@ms55.hinet.net[3]

摘要 在傳統的訊號子空間語音強化方法(Signal Subspace Speech Enhancement Method)中，其主要是利用噪音能量是均勻分佈於訊號所在的向量空間而語音訊號能量則是分佈於某一子空間的特性，藉由特徵分解(Eigen-Decomposition)來分析出語音訊號及背景噪音，來進行噪音消除。而在車內噪音環境中，噪音能量的分佈在低頻帶為最多延伸到高頻則逐漸較少，單一的訊號子空間的語音強化方法已不能更有效的消除位在低頻帶的背景噪音。本論文提出一個基於人耳聽覺特性的分頻處理，並結合訊號子空間強化方法來克服此一問題。實驗的驗證，則是採用 TAICAR 車內語音資料庫來進行，實驗結果說明本文所提出的方法比起傳統訊號子空間強化法，更適用於車內噪音的消除，低頻噪音的消除也更明顯。

1. 前言

隨著汽車導航系統的日漸普及，除了提供汽車行車資料及娛樂外，藉由結合行動電話的無線通訊功能，更讓汽車儼然已經變成隨時可獲知各種生活資訊的行動中心。在汽車內傳統的人機介面是採用觸碰式螢幕，在行車的狀況下，這樣的介面是不夠安全的，而隨著即時語音辨識技術的日趨成熟，人機介面必定是朝著語音對話的操控方式改進。在行車環境中，充斥各種噪音，對於語音辨識系統而言，這些背景噪音會嚴重地影響辨識結果。因此，一般的辨識系統都需使用手持式或頭戴式麥克風，來促成近距離的錄音，以避免背景噪音的干擾。然而，使用這些錄音設備會對駕駛或者乘客造成不便，所以提供一個在行車環境下能實行遠距離錄音並具有抗噪音能力的麥克風系統，是有其需求。本論文提出一個利用小波聽覺分頻處理與訊號子空間分解來達成車內背景噪音消除的目的。

Ephraim 和 Van-Trees 於 1995 年提出一套基於訊號子空間分解的語音強化系統 [1]，利用噪音能量是均勻分佈於訊號所在的向量空間而語音訊號能量則是分佈於某一子空間的特性，藉由特徵分解來分析出語音訊號及背景噪音，並進一步用一線性估測器來處理得到強化後的語音。由於特徵分解的運算複雜度高，在本論文中採用子空間追蹤(Subspace Tracking)的方式來做特徵分解，這個演算法稱為 PAST (Projection Approximation Subspace Tracking, PAST) [2]，以期能符合即時(Real-Time)的應用。而在車內噪音環境中，噪音能量的分佈在低頻帶為最多延伸到高頻則逐漸較少，在實驗過程中發現，單一的訊號子空間的語音強化方法已不能更有效的消除位在低頻帶的背景噪音。因此，本論文提出一個基於人耳聽覺特性的分頻處理，並結合訊號子空間

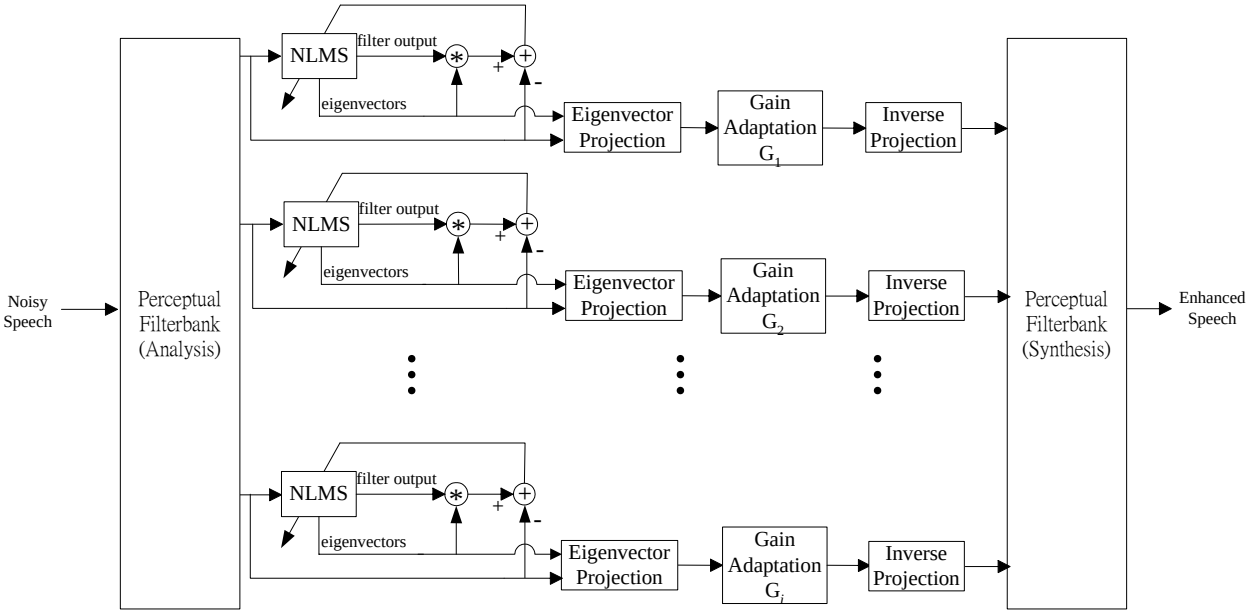
* This work was supported by the National Science Council of the Republic of China, Taiwan, Contract Nos. NSC 92-2213-E-006-022

強化方法來克服此一問題。此一聽覺分頻處理係利用小波轉換(Wavelet Transform)來實現，藉由小波將聲音分解成多個頻帶，而各個子頻帶的分佈則符合人耳聽覺響應的特性，各子頻帶的訊號再經由子空間方法進行噪音消除，再由小波反轉換合成各子頻帶的訊號，進而得到強化後的語音。實驗的驗證，則是採用 TAICAR 車內語音資料庫來進行，實驗結果說明本文所提出的方法比起傳統訊號子空間強化法，更適用於車內噪音的消除，低頻噪音的消除也更明顯。

本論文的章節結構如下：第二節是所提出來的噪音消除系統架構，包含小波聽覺分頻處理、訊號子空間語音強化以及子空間追蹤法之描述；第三節說明實驗結果，包含 TAICAR 車內語音資料庫的介紹以及本文所提出之方法跟其它訊號子空間語音強化法之比較；最後，第四節則是結論與討論。

2. 系統架構

本論文所提出的車內噪音消除系統，如圖一所示。在系統前端，麥克風所錄到的雜訊語音，經由小波聽覺濾波組(Perceptual Wavelet Filterbank)分成數個子頻帶訊號，各個子頻帶則由訊號子空間語音強化來進行噪音消除的處理，而訊號子空間的拆解則是由子空間追蹤法來完成。由子空間追蹤法所估算出來的特徵值(Eigenvalue)，則用以計算各個子頻帶訊號的增益值。語音強化的處理為將子頻帶訊號經過特徵向量(Eigenvector)投影轉換後，由增益值來調整其訊號大小，再經過反轉換來得到強化後的語音訊號。以下各小節則對小波聽覺分頻處理、訊號子空間語音強化以及子空間追蹤法做一描述。



圖一：車內噪音消除系統架構。

2.1. 小波聽覺分頻處理

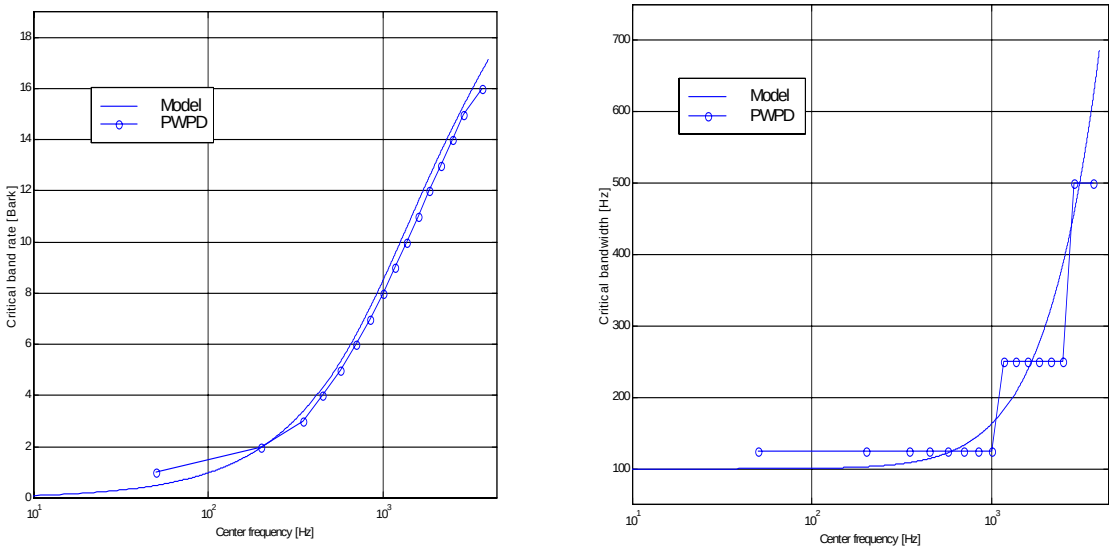
具聽覺感知的小波轉換(Perceptual Wavelet Packet Transform, PWPT)是改良自傳統小波轉換，使語音信號經 PWPT 分解後的各個子頻帶信號的頻寬接近人耳的聽覺響應 [3]，描述人耳聽覺響應的參數主要有巴克頻譜（Bark）以及關鍵頻寬（Critical Bandwidth），表一為人耳聽覺關鍵頻寬的分佈情形。圖二(a)及圖二(b)分別是在 4KHz 內，人耳的聽覺的巴克頻譜及關鍵頻寬曲線圖 [4, 5]。因此，本論文所設計的聽覺分頻處理即朝

此二曲線設計，圖二(a)及圖二(b)內亦標示了利用小轉換逼近巴克頻譜及關鍵頻寬的曲線圖。

由小轉換逼近巴克頻譜及關鍵頻寬是藉由調整小波轉換的樹狀結構來達成。依據表一的關鍵頻寬分佈情形，適當對訊號做高低頻的分解，使得子頻帶訊號的頻率分佈跟關鍵頻寬近似。圖三為所使用的具聽覺感知的小波轉換分解架構圖，其中輸入訊號經五個階段，共 16 次的高低頻分解。

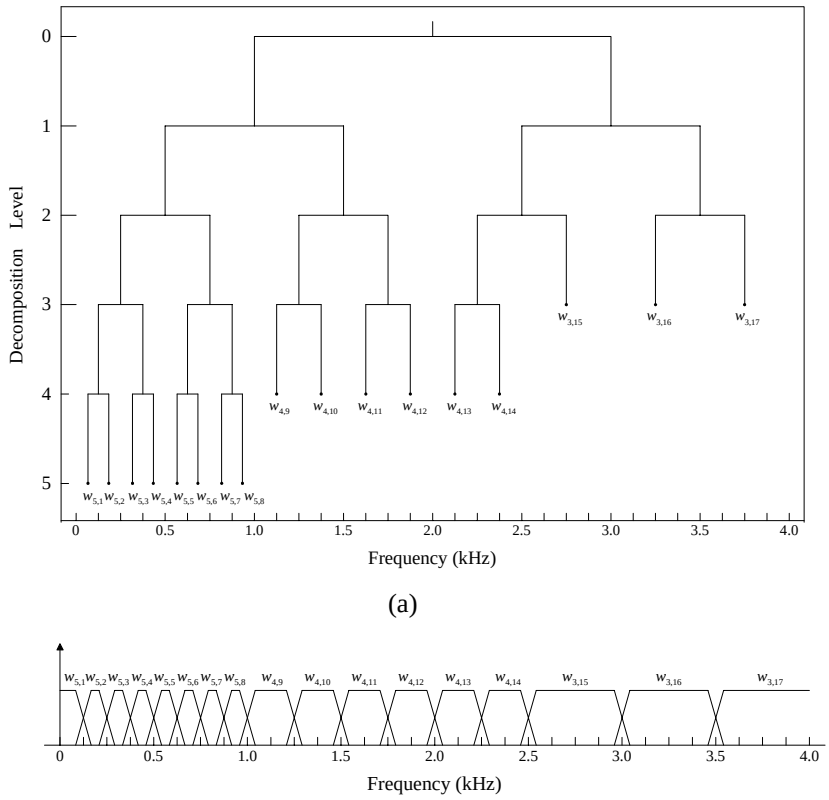
表一：關鍵頻寬分佈。

Critical Band Number	Center Frequency (Hz)	CBW	Lower Cutoff frequency (Hz)	Upper Cutoff Frequency (Hz)
1	50	-	-	100
2	150	100	100	200
3	250	100	200	300
4	350	100	300	400
5	450	110	400	510
6	570	120	510	630
7	700	140	630	770
8	840	150	770	920
9	1000	160	920	1080
10	1170	190	1080	1270
11	1370	210	1270	1480
12	1600	240	1480	1720
13	1850	280	1720	2000
14	2150	320	2000	2320
15	2500	380	2320	2700
16	2900	450	2700	3150
17	3400	550	3150	3700



(a) (b)

圖二：4KHz 內(a)人耳的聽覺的巴克頻譜及(b)關鍵頻寬曲線圖。



(a)

(b)

圖三：(a)聽覺感知分頻的樹狀結構及(b)各子頻帶的頻寬。

2.2. 語音子空間強化

在本論文中採用的語音強化演算法為訊號子空間分析法 [1]。在子空間分析法中，從含有雜訊的語音訊號中的共變異數矩陣(Covariance Matrix)由特徵分解求出其特徵向量及特徵值，接著利用這兩個資訊和一個線性預估器將背景噪音消除。

一個 K 維語音訊號向量 y 其線性模型為

$$y = \sum_{m=1}^{M} s_m V_m \qquad , M < K \tag{1}$$

其中 $s_1, ..., s_M$ 為平均值為零的隨機變數，而 $V_1, ..., V_M$ 為基底。訊號分佈的向量空間其維度為 K，而語音訊號分佈的子空間其維度為 M，$M < K$。式子(1)可以表示為 $y = Vs$，$V \equiv [V_1, ..., V_M]$ 為 $K \times M$ 的矩陣，其秩(Rank)為 M，且 s 為一行向量表示為 $s \equiv (s_1, ..., s_M)^T$。語音訊號的共變異數矩陣 $R_y = VR_sV^T$ 其秩為 M，R_s 為 s 的共變異數矩陣並假設其為正定矩陣(Positive Definite Matrix)。$M < K$ 的性質使得在 K 維語音訊號 y 中，R_y 有 $K - M$ 個特徵值為零，這個對於在以子空間演算法做語音強健中極為重要。

令 w 為 K 維向量表示背景白色噪音，其平均值為零。其共變異數矩陣 $R_w = E\{ww^T\} = \sigma_w^2 I$。白色噪音的共變異數矩陣其秩為 K，也就是說它會佈滿整個歐式空間 R^K 中。因此，對於背景為白色噪音的雜訊語音

訊號，整個 K 維的向量空間由 M 維的訊號子空間及 K 維的噪音子空間所組合而成，其中可以將 $K-M$ 的特徵值為零所對應的子空間去除掉，而剩下的 M 維的雜訊子空間，可以用一線性預估器將其乾淨語音粹取出來。

底下將說明線性預估器的求取，假設雜訊語音為 $Z(n)=Y(n)+W(n)$，$W(n)$ 為 K 維的背景噪音向量，$Y(n)$ 為 K 維的語音向量，n 為訊號音框的索引。令 $H(n)$ 為一 $K\times K$ 的乾淨語音之線性預估器亦即

$$\hat{Y}(n)=H(n)Z(n) \tag{2}$$

則其預估錯誤訊號則為

$$\begin{aligned}\varepsilon(n)&=\hat{Y}(n)-Y(n)\\&=(H(n)-I)Y(n)+H(n)W(n)\\&=\varepsilon_{\mathrm{y}}(n)+\varepsilon_{w}(n)\end{aligned} \tag{3}$$

$\varepsilon_{y}(n)\equiv(H(n)-I)Y(n)$ 代表訊號的失真量，I 為單位矩陣(Identity Matrix)，$\varepsilon_{w}(n)\equiv H(n)W(n)$ 代表噪音的殘餘量。定義訊號失真能量及噪音殘餘能量分別為 $\bar{\varepsilon}_{y}^{2}(n)$ 、 $\bar{\varepsilon}_{w}^{2}(n)$。則訊號失真能量表示為

$$\begin{aligned}\bar{\varepsilon}_{y}^{2}(n)&=tr(E[\varepsilon_{y}(n)\varepsilon_{y}^{T}(n)]),\\&=tr((H(n)-I)R_{y}(n)(H(n)-I)^{T})\end{aligned} \tag{4}$$

且噪音的殘餘能量為

$$\begin{aligned}\bar{\varepsilon}_{w}^{2}(n)&=tr(E[\varepsilon_{w}(n)\varepsilon_{w}^{T}(n)]),\\&=tr(H(n)R_{w}(n)H(n)^{T})\end{aligned} \tag{5}$$

$R_{y}(n)$ 及 $R_{w}(n)$ 分別為乾淨語音訊號及噪音訊號的共變異數矩陣。因要其訊號失真能量最小化，而最小化的情況要限制在噪音能量小於一個很小的常數，因此其最佳的線性預估器定義如下

$$H_{opt}(n)\equiv\arg\min_{H(n)}\bar{\varepsilon}_{y}^{2}(n),\qquad \text{Subject to}:\tfrac{1}{K}\bar{\varepsilon}_{w}^{2}(n)\le\sigma^{2} \tag{6}$$

其中 σ^{2} 是一個正的常數值。求解式子(6)，可利用拉氏乘子法(Lagrange Multiplier)，得到

$$L(H(n),\mu)\equiv\bar{\varepsilon}_{y}^{2}(n)+\mu(\bar{\varepsilon}_{w}^{2}(n)-K\sigma^{2}) \tag{7}$$

及

$$(\bar{\varepsilon}_{w}^{2}(n)-K\sigma^{2})=0,\qquad \mu\ge 0 \tag{8}$$

而 μ 為拉氏乘子。對式子(7)取梯度運算(gradient)，令其為零求解，則線性預估器可得到為

$$H_{opt}(n)=R_{y}(n)(R_{y}(n)+\mu R_{w}(n))^{-1} \tag{9}$$

由特徵值分解，式子(9)可表為

$$H_{opt}(n) = U(n)\Lambda_y(n)(\Lambda_y(n) + \mu\Lambda_w(n))^{-1}U^T(n) \tag{10}$$

特徵分解 $R_y(n) = U(n)\Lambda_y(n)U^T(n)$，$U(n)$ 為特徵向量的矩陣，$\Lambda_y(n)$ 為特徵值矩陣，$\Lambda_w(n)$ 為 $R_w(n)$ 的特徵值矩陣。令 $G(n) = \Lambda_y(n)(\Lambda_y(n) + \mu\Lambda_w(n))^{-1}$，則

$$H_{opt}(n) = U(n)G(n)U^T(n) \tag{11}$$

2.3. 子空間追蹤演算法

子空間語音強健最後的線性預估器須要雜訊語音的特徵分解，其運算複雜度高。所以在本論文中採用追蹤疊代的方式來逼近特徵值，這個演算法稱為 PAST (Projection Approximation Subspace Tracking) [2]。PAST 的演算法用來追蹤子空間的特徵值在許多文獻中被證明是準確且計算複雜度低的。若以子空間演算法，其運算複雜度為 $O(n^3)$，n 為輸入向量的維度，若子空間追蹤方法來做計算，其運算複雜度可以減少至 $O(nr)$，其中 n 為輸入向量的維度，r 為我們需要的特徵值暨特徵向量的的數目。

PAST 演算法其原理為對所給定的成本函數(Cost Function)求取最小值，成本函數的決定與共變異數矩陣有關，

$$J(u(n)) = \sum_{i=1}^{n} \beta^{n-i} \left\| Z(i) - u(n)u^T(n)Z(i) \right\|^2 \tag{11}$$

其中 $u(n)$ 為 K 維的向量，且 $0 \le \beta \le 1$ 為消散係數(Forgetting Factor)，β 可以使成本函數的極值所在會是下面定義的相關矩陣 $R_z(n)$ 中的特徵向量之一。

$$\hat{R}_Z(n) = \sum_{i=1}^{n} \beta^{n-i} Z(i)Z(i)^T \tag{12}$$

定義一個 $J'(u(n))$ 如下

$$J'(u(n)) = \sum_{i=1}^{n} \beta^{n-i} \left\| Z(i) - u(n)u^T(i-1)Z(i) \right\|^2 \tag{13}$$

$J(u(n))$ 和 $J'(u(n))$ 不同在於使用了 $u^T(i-1)$ 代替 $u^T(n)$，直覺的觀察，$J'(u(n))$ 可以被用來近似 $J'(u(n))$。因為語音訊號的統計特性為穩態(stationary)，也就是某個時間區段它變化的很慢所以 $u^T(i-1) \approx u^T(n)$。當 $i << n$ 時，β^{n-i} 會變得很小使得最後 $J'(u(n)) \approx J(u(n))$。我們可以用適應性梯度演算法將 $J'(u(n))$ 取梯度運算迭代求出特徵向量。其 PAST 演算法如下所示。

表二: PAST 演算法。

初始化： $d_i(0)=0, \beta=0.95$

$U(0)=[u_1(0)\,|\,u_2(0)\,|\,\ldots\,|\,u_k(0)]=I_k$

For $n=1,2,\ldots$ do

$Z_1(n)=Z(n)$

For $i=1,2,\ldots,k$ do

$$vi(n)=u_i^T(n-1)Z_i(n)$$

$$d_i(n)=\beta d_i(n-1)+|v_i(n)|^2$$

$$E_i(n)=Z_i(n)-u_i(n-1)v_i(n)$$

$$u_i(n)=u_i(n-1)+T(n)E_i(n)\frac{v_i(n)}{d_i(n)}$$

$$Z_{i+1}(n)=Z_i(n)-u_i(n)v_i(n)$$

end

end

輸出：

$U(n)=[u_1(n)\,|\,u_2(n)\,|\,\ldots\,|\,u_k(n)]$

其中 $d_i(n)$ 為子空間的特徵值。對於乾淨語音以及噪音為無相關的所以在特徵值上的分佈可以寫成

$$\Lambda_z(n)=\Lambda_y(n)+\Lambda_w(n) \tag{14}$$

所以只要求雜訊語音的特徵值且 $\Lambda_w(n)$ 為白色噪音其統計特性已知，可以將 $\Lambda_z(n)-\Lambda_w(n)$ 得到乾淨語音的 $\Lambda_y(n)$ 。一般來說噪音的統計特性不是穩態的。所以要完成估算 $\Lambda_w(n)$ ，通常都假設噪音在某段時間內變化很慢，因為實際一般環境中噪音得變化其實不大(如汽車內、室內冷氣聲…等)。所以在雜訊語音中，前一段的噪音資訊可以存起來給下一個語音段使用以求出 $\Lambda_Y(n)$ 。所以在用追蹤演算法估算 $\Lambda_w(n)$ 時用指數式的方式來疊加平均，以達到正確真實的 $\Lambda_w(n)$ ，其指數式的方式來疊加平均如下

$$\Lambda_w(n)=\beta\Lambda_w(n-1)+U^T(n)W(n) \tag{15}$$

β 為一平滑係數(smoothing factor)且 $W(n)$ 的值從之前的靜音區段選擇代用。所以可以計算噪音的能量在每個語音段之間，其計算出來的噪音能量資訊直接給下個語音段使用。

3. 實驗

對於本文所提出的方法，則是採用 TAICAR 車內語音資料庫來進行實驗的驗證。以下就對 TAICAR 資料庫做一介紹，接著對車內所蒐集的雜訊音檔進行噪音消除的實驗。

3.1. TAICAR 車內語音資料庫

在國外有很多的語音資料庫收集之方法，例如：日本的 CIAIR、歐洲的 SpeechDat 等等，然而在汽車環境下的語料收集卻是很少見，TAICAR 資料庫目標就在於收集汽車環境下的語料以方便各種語音處理技術之開發。例如：噪音補償技術、噪音下動態語音偵測技術、強健型語音辨識技術、語音調適技術等。語料的錄音參考國內執行過的大型計畫「MAT 語料收集」之作法，先由程式從 100 萬字的文字庫中挑選出能夠涵蓋所有國語基本音節的短詞、單字等，並加上英文、數字部分，總共這樣的語料有 360 份。為了實際記錄各種不同路況，錄音時分兩種路段：市區路段以及快速道路路段。市區路段下，時速為 0～50 公里；快速道路則需維持在 70～100 公里。

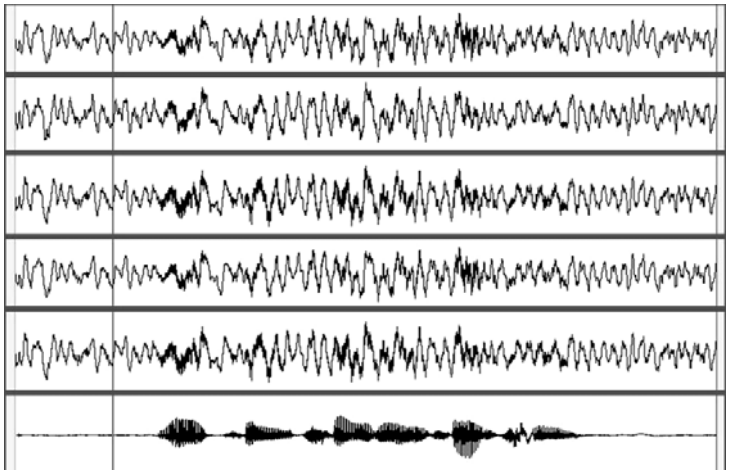
在車內錄音需考慮到便利性，因此以筆記型電腦為錄音的平台，配合上特殊硬體來進行錄音。所用錄音器材計有：

- 筆記型電腦: 負責主要的錄音程式之進行
- PCMCIA 介面之多頻道信號擷取卡：負責擷取多頻道的語音訊號
- 麥克風：1 支指向性(頭戴式，收錄乾淨語音)＋5 支全向性(收錄雜訊語音)：負責語音訊號的輸入
- 車輛：任意

車內的錄音軟體，可同時進行 6 個 channel 錄音，單音取樣: 16KHz , 16bits。在錄音之同時可標記路況、車速、語者性別、基本資料等 [6]。圖四為 TAICAR 車內音檔錄音情況，圖五則為所錄得音檔之時間波形。

(a) (b)

圖四：(a) 車內多麥克風配置及(b)語者與錄音設備。



圖五：車內六個麥克風所錄之時間波形。

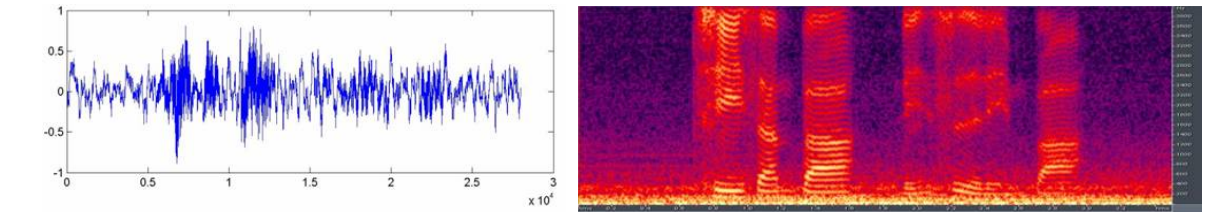
3.2. 效能評估

實驗的驗證以 TAICAR 音檔來做測試，對進行過噪音消除後之音檔進行評分。評分方式採人耳試聽為之(Mean Opinion Score, MOS)，給分等級為：5 為優，4 為好，3 為尚可，2 為略差，1 為不好。以本文所提之方法與另外兩種子空間分解方法作比較，其為子空間分解採用離散餘弦轉換(Discrete Cosine Transform, DCT)及採用 KL 轉換(Karhunen-Loeve Transform, KLT)。計有二十位試聽者給分，給分結果如表三所示。

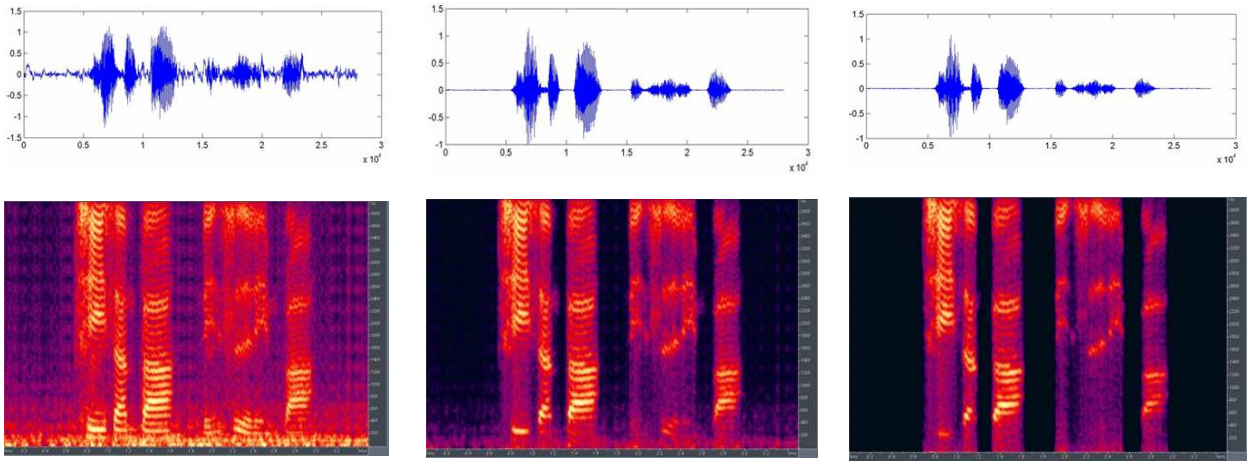
表三: MOS 測試評分。

	TAICAR 音檔		
	待速	市區路段	快速道路路段
DCT	2.6	2.1	1.9
KLT	4.4	4.1	3.8
本論文所提方法	4.2	4.0	3.9

圖六則為雜訊音檔經由上述三種方法進行噪音消除後之波形及頻譜圖。從圖六觀察噪音抑制結果，以 KLT 及本文所提方法皆優於 DCT 的效果，再從低頻帶的噪音消除來看，則是以本文所提的方法為最好。



(a)雜訊音檔波形及頻譜圖



(b) DCT　　(c) KLT　　(d)本文所提方法之結果

圖六：噪音消除結果比較之波形及頻譜圖。

4. 結論

本論文提出一個基於人耳聽覺特性的分頻處理，並結合訊號子空間強化方法來消除車內背景噪音。此一聽覺分頻處理係利用小波轉換來實現，藉由小波將聲音分解成多個頻帶，而各個子頻帶的分佈則符合人耳聽覺響應的特性，各子頻帶的訊號再經由子空間方法進行噪音消除，再由小波反轉換合成各子頻帶的訊號，進而得到強化後的語音。實驗的驗證，則是採用 TAICAR 車內語音資料庫來進行，由 MOS 評分及時間波形和頻譜圖來看，本文所提出的方法比起傳統訊號子空間採用 DCT 及 KLT 等方法，更適用於車內噪音的消除，低頻噪音的消除也更明顯。

5. 參考文獻

[1] Y. Ephraim and H. L. Van-Trees, “A signal subspace approach for speech enhancement,” *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 251–266, July 1995.

[2] B. Yang, “Projection approximation subspace tracking,” *IEEE Trans. Signal Processing*, vol. 43, pp. 95–107, Jan. 1995.

[3] Shi-Huang Chen and Jhing-Fa Wang, “Speech Enhancement Using Perceptual Wavelet Packet Decomposition and Teager Energy Operator,” accepted to appear in *The Journal of VLSI Signal Processing Systems*, Special Issue on Real World Speech Processing.

[4] O. Ghitza, “Auditory model and human performance in tasks related to speech coding and speech recognition,” *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 115-132, 1994.

[5] Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993

[6] Jhing-Fa Wang, Hsien-Chang Wang and Chung-Hsien Yang, “TAICAR - A Collection of In-Car Mandarin Speech Database in Taiwan,” *O-COCOSDA2003 / PACLIC17*, Singapore.

聚集事後機率線性迴歸調適演算法應用於語音辨識
Aggregate *a Posteriori* Linear Regression for Speech Recognition

黃志賢 王奕凱 簡仁宗

國立成功大學資訊工程學系

{acheron,display}@chien.csie.ncku.edu.tw, jtchien@mail.ncku.edu.tw

摘要

在本論文中，我們提出一套由聚集事後機率(aggregate *a posteriori*)為基礎之鑑別式線性回歸(linear regression)轉換矩陣參數調適演算法。在近幾年，由於鑑別式訓練的效果優越，於是出現使用鑑別式訓練法則進行轉換矩陣調適，稱為最小分類錯誤率線性迴歸(minimum classification error linear regression, MCELR)調適演算法。我們認為使用最小分類錯誤率準則進行線性迴歸調適時，若能再進一步考慮線性迴歸矩陣之事前機率分佈，則可以結合貝氏法則之強健性與最小分類錯誤率之鑑別性，以估測出更佳之轉換矩陣用於語者調適上。透過聚集事後機率與鑑別式訓練間之關連及適當之條件簡化，則可得到參數更新之封閉解(close form)型式以加速鑑別式訓練的參數估測。在實驗中，我們使用 TCC300 語料進行語音模型參數之訓練與迴歸矩陣之事前機率分佈之參數估測，而在調適及測試時，則使用公共電視台所錄製之電視新聞語料，進行轉換矩陣估測強健性之評估與其他轉換矩陣參數調適效能之比較，在不同調適語料之實驗結果發現我們提出之聚集事後機率線性回歸可以有效達到鑑別式語者調適的效果。

1. 緒論

在語音辨識的相關研究中，常常需要面對的問題是用於訓練時的語料與測試時語料的語者或環境常常大不相同。每個人的聲學特質都不相同，而不同環境所產生的背景雜訊也都不同。如何有效地將訓練所得的語音模型配合測試時所使用的語料特性進行適當的語者調適，以有效地消除這兩者之間的不匹配情形，是許多學者研究的課題。

語音模型的參數必須在訓練時使用大量語料進行估測，最普遍使用的模型訓練準則為最大相似度估測(maximum likelihood estimate, MLE)[19]，在此種方法中，當模型與所收集之訓練語料的相似度最大時，即可求得在此估測準備下最佳的語音模型參數。由於語音模型參數的估測，有所謂不完整資料(incomplete data)的問題，所以皆利用 EM(Expectation-Maximization)演算法[6]進行理論推導。

除了使用最大相似度作為參數估測準則之外，另一個也常被用於作為參數估測的是基於貝氏理論的最大事後機率(maximum *a posteriori*, MAP)估測法則[8]。貝氏估測法則認為參數為一隨機變數，可以機率分佈表示之。利用根據所給定的訓練語料而使得對應的模型參數之事後機率最大之特性，即可求得基於此方法之最佳參數。在最大事後機率訓練法則之訓練機制下，一般不可直接最大化模型參數之事後機率，而常根據貝氏法則，將之拆解為語料與模型間相似度與模型參數事前機率之組合，所以可利用事前資訊對模型參數加以限制，可以改善訓練資料稀疏所產生的錯誤訓練問題。

除了前述兩者參數估測準則之外，鑑別式訓練(discriminative training)[3]則提供了在模型訓練上的另一種選擇。由較早的 multilayer perceptron(MLP)[17]、learning vector quantization(LVQ)[18]，到近來的最小分類錯誤(minimum classification error, MCE)[11]、最大相互資訊(maximum mutual information, MMI)[20]，有許多不同的理論方法。鑑別式訓練與其它模型訓練方法最大的不同是，除了考慮樣本與本身模型的相似度之外，還額外考慮樣本與其它模型之間的相似度，這種作法可以避免模型訓練時，原本就相似的語音模型產生互相混淆的情況。

Qi Li [15]在 2002 年提出一般化最小錯誤率(generalized minimum error rate, GMER)，由事後機率的角度出發，定義聚集事後機率(aggregate *a posteriori*, AAP)，並將事後機率改寫為具鑑別性形式的誤辨率(misclassification measure)函式。在訓練模型參數上，不使用一般的廣義機率遞減法則(generalized probabilistic descent, GPD)，透過一些條件假設，即可推導出模型參數估測的封閉解形式。

在語者調適的研究上，最廣為使用的有最大相似度線性迴歸(maximum likelihood linear regression, MLLR)調適[7][14]與最大事後機率調適兩大類方法。在本研究中我們將使用前者作為調適的主要架構，透過所估測出之線性迴歸矩陣對語音模型參數進行調適。由於考慮到使用語料量稀少易造成調適效果失準的情況，引入線性轉換矩陣之事前分佈資訊，以強健化調適效能外，也將由鑑別式訓練之角度出發，嘗試找出不同於傳統以貝氏法則為準之最大化

聚集事後機率線性迴歸(aggregate *a posteriori* linear regression, AAPLR)演算法。故我們會針對文獻中所提過之以線性迴歸為主之調適演算法作回顧。除了最大相似度線性迴歸調適演算法之外，主要有最大事後機率線性迴歸(MAPLR)[21]、考慮到漸進式(sequential)學習的近似貝氏線性迴歸(quasi-Bayes linear regression, QBLR)[5]與最小分類錯誤線性迴歸(minimum classification error linear regression, MCELR)[4][9][10]。

我們將提出的語音模型參數調適演算法，使用連續語音辨識系統進行與其他調適演算法的效能評估。接下來，我們先回顧近年來鑑別式訓練的相關研究文獻與前述幾種以轉換為主之語音模型調適演算法及將聚集事後機率應用在鑑別式聲學模型參數估測上的方法。其次，說明我們將一般化最小錯誤率應用在語音模型參數調適及在調適時考慮轉換矩陣的事前機率分佈，最後得到估測的轉換矩陣參數封閉解之相關理論內容。接著說明實驗設定與進行方式並由實驗所得結果進行討論。在結尾部份則簡單歸納本論文的主要重點與結論，並說明未來繼續研究的方向與課題。

2. 鑑別式訓練及線性回歸調整

最大相似度參數估測法則是最普遍用來訓練隱藏式馬可夫模型參數的方法，它利用 EM 演算法估測模型參數非常有效率；最大相似度的缺點是模型參數只利用屬於本身模型的資料來估測，和其它模型的參數估測基本上是獨立的。最小分類錯誤和最大交互資訊，是近來較廣為利用的鑑別式訓練方法，除了訓練語音模型外，還用在語言模型(language model)的訓練上[13]、語者辨識模型訓練、特徵參數擷取。使用鑑別式訓練估測模型參數時，除了本身模型的資料外，還考慮與其它模型參數之鑑別性，所以可以更正確地估測出所需的模型參數內容。在[15][16]中，作者提出了另一種鑑別式訓練方法，稱作一般化最小錯誤率，從事後機率出發，定義與最大事後機率相似的目標函式，並且改寫為鑑別式訓練的形式，以下分別簡介這三種鑑別式訓練法則。

2.1 最小分類錯誤(MCE)訓練法則

在兩個類別 C_1, C_2 的分類器裡，假設 $\mathbf{x} \in C_1$ ，貝氏分類法則定義了最基本的誤辨值函式(misclassification measure)為

$$d(\mathbf{x}) = P(C_2 \mid \mathbf{x}) - P(C_1 \mid \mathbf{x}) \tag{1}$$

上式表示類別 C_1 的觀察資料 $\mathbf{x}$ 被分類器分類到類別 C_2 的可能性，在多個類別的分類器[12]裡，定義誤辨值函式

$$d_k(\mathbf{x}) = \sum_{i \in M_i} \frac{1}{m_k} \left[g_i(\mathbf{x};\Lambda) - g_k(\mathbf{x};\Lambda) \right] \tag{2}$$

其中 $g_i(\mathbf{x};\Lambda)$ 為觀察資料 $\mathbf{x}$ 對類別 C_i 的相似度，Λ 表示所有類別的模型參數，$M_k = \{ i \mid g_i(\mathbf{x};\Lambda) > g_k(\mathbf{x};\Lambda) \}$，代表一群對觀察資料 $\mathbf{x}$ 的相似度比類別 C_k 對觀察資料 $\mathbf{x}$ 相似度更具競爭性的類別集合，即混淆類別(confusing classes)或競爭類別(competing classes)的集合。

式子(2)中，S_k 並非是固定的集合，它隨著模型參數 Λ 和觀察資料 $\mathbf{x}$ 而改變，而且該式在 Λ 不連續[12]，這在最陡坡降法(gradient descent)裡並不適用，因此另外定義了一個連續性的誤辨值公式為

$$d_k(\mathbf{x}) = -g_k(\mathbf{x};\Lambda) + \left[\frac{1}{M-1} \sum_{j, j \neq k} g_j(\mathbf{x};\Lambda)^{\eta} \right]^{1/\eta} \tag{3}$$

其中 η 是一個正數，藉著改變 η 的值，可以改變式子裡具影響力的競爭類別數量，令 $\eta \to \infty$ ，一個極端的誤辨值公式為

$$d_k(\mathbf{x}) = -g_k(\mathbf{x};\Lambda) + g_i(\mathbf{x};\Lambda) \tag{4}$$

類別 C_i 是除了類別 C_k 外，和觀察資料 $\mathbf{x}$ 相似度最大的類別，$d_k(\mathbf{x}) > 0$ 代表發生分類錯誤，$d_k(\mathbf{x}) \leq 0$ 代表正確分類。為了更進一步完成目標函式的定義，把誤辨值公式代入 cost function

$$l_k(x;\Lambda) = l(d_k(\mathbf{x})) \tag{5}$$

cost function 一般為連續性，範圍為[0,1]的函式，最常用於 MCE 的為 sigmoid，

$$l(d_k) = \frac{1}{1 + \exp(-\gamma d_k + \theta)} \tag{6}$$

對於某個觀察資料 $\mathbf{x}$ ，我們可以 cost function 定義分類器的效率為

$$l(\mathbf{x};\Lambda)=\sum_{i=1}^{M} l_i(\mathbf{x};\Lambda)1(\mathbf{x}\in C_i) \tag{7}$$

最後利用廣義機率遞減(generalized probabilistic decent, GPD)演算法進行疊代運算以實現 MCE 法則。

$$\Lambda_{t+1}=\Lambda_t-\varepsilon_t U_t \nabla l(x;\Lambda)|_{\Lambda=\Lambda_t} \tag{8}$$

廣義機率遞減法則是應用很廣的演算法，利用反覆的計算，遞迴得到一收斂的值，缺點是收斂速度慢，而且式中的學習係數ε_t需對應不同的資料特性去調整。更進一步之相關參數估測過程與結果詳見[11]。

2.2 最大交互資訊(MMI)訓練法則

除了最小分類錯誤法則外，最大交互資訊也是普遍利用的鑑別式訓練式法則[1][20]，最大交互資訊較隱性的引入了觀察資料與其它類別的相似度，所以與一般化最小錯誤率較相似，在混合數高的情況下，最大交互資訊能訓練出比最小分類錯誤辨識率更高的模型參數[1]，由於最大交互資訊考慮了觀察資料和所有類別的相似度，因此比最小分類錯誤在實作上難度更高。為了快速計算隱藏式馬可夫模型和觀察資料$\mathbf{x}$的相似度，必須使用 forward-backward 演算法。透過 forward probability $\alpha_j(t)$ 與 backward probability $\beta_j(t)$ 的表示，類別C_m產生觀察資料$\mathbf{X}$的機率可寫為下式

$$P(X\mid C_m,\Lambda)=\sum_{t=1}^{T}\sum_{j=1}^{N}\alpha_j(t)\beta_j(t) \tag{9}$$

定義類別C_m與觀察資料$\mathbf{X}$的交互資訊為

$$\begin{aligned} I_\Lambda(C_m,\mathbf{X}) &= \log\frac{P(\mathbf{X}\mid C_m)}{P(\mathbf{X})} \\ &= \log P(\mathbf{X}\mid C_m)-\log P(\mathbf{X}) \\ &= \log P(\mathbf{X}\mid C_m)-\log\sum_{m'=1}^{M}P(\mathbf{X}\mid C_{m'})P(C_{m'}) \end{aligned} \tag{10}$$

其中$P(\mathbf{X},C_m)$代表類別C_m與$\mathbf{X}$同時出現的機率，即聯合相似度(joint likelihood)。由(10)式可看出，除了觀察資料$\mathbf{X}$與對應類別C_m的相似度之外，還加入了$\mathbf{X}$與其它類別的相似度作為參數估測的考量，因此它屬於鑑別式訓練的一種，以最大交互資訊法則得到的模型參數可使得觀察資料$\mathbf{X}$與類別C_m有較高的相依性，即$I_\Lambda(C_m,\mathbf{X})$較高。與最小分類錯誤相同，最大交互資訊也必須以廣義機率遞減演算法實現，即

$$\Lambda_{n+1}=\Lambda_n-\varepsilon\nabla I_\Lambda(C_m,\mathbf{X}) \tag{11}$$

在這裡以轉移機率、平均值向量、共變異矩陣作說明，而偏微的對象主要是最大交互資訊中的相似度函式

$$\frac{\partial}{\partial\Lambda}\log P(\mathbf{X}\mid C_m,\Lambda)=\frac{1}{P(\mathbf{X}\mid C_m,\Lambda)}\frac{\partial}{\partial\Lambda}P(\mathbf{X}\mid C_m,\Lambda) \tag{12}$$

相似度函式可由 forward-backward probability 表示

$$\begin{aligned} P(\mathbf{X}\mid C_m,\Lambda) &= \sum_{t=1}^{T}\sum_{j=1}^{N}\alpha_j(t)\beta_j(t) \\ &= \sum_{t=1}^{T}\sum_{j=1}^{N}\left\{\sum_{i=1}^{N}\alpha_i(t)a_{ij}\right\}b_j(\mathbf{x}_t)\beta_j(t) \end{aligned} \tag{13}$$

進一步之參數估測過程與結果，請詳見[20]。

2.3 一般化最小錯誤率(GMER)

一般化最小錯誤率是由 *Qi Li* 在 2002 年所提出，以下簡介一般化最小錯誤率的精神和作法。在一個具有M個類別的分類問題裡面，令觀察資料$\mathbf{X}$屬於類別C_m，α_i表示將$\mathbf{X}$分類到類別C_i的動作，則可定義一 loss function 為

$$l(\alpha_i\mid C_m)=\begin{cases}0 & i=m \quad i,m=1,\ldots,M\\ 1 & i\neq m\end{cases} \tag{14}$$

將分類錯誤指定一個單位的 loss，若分類正確則不指定 loss，代表分類錯誤的風險(risk)，且定義對觀察資料 $\mathbf{X}$ 採取動作 α_i 的分類錯誤機率為

$$R(\alpha_i \mid \mathbf{X}) = \sum_{j=1}^{M} l(\alpha_i \mid C_j) P(C_j \mid \mathbf{X}) = 1 - P(C_m \mid \mathbf{X}) \tag{15}$$

$P(C_m \mid \mathbf{X})$ 代表 $\mathbf{X}$ 屬於類別 C_m 的事後機率，貝氏法則告訴我們，令 $P(C_m \mid \mathbf{X})$ 最大可降低分類錯誤的機率，稱作最小錯誤率(minimum error rate, MER)，$P(C_i \mid \mathbf{X})$ 一般以一組定義好的模型參數 λ_i 來計算，即 $P(C_i \mid \mathbf{X}) = P_\lambda(C_i \mid \mathbf{X})$，由於模型參數與類別有一對一的關係，因此簡化表示為 $P(C_i \mid \mathbf{X}) = P(\lambda_i \mid \mathbf{X})$。在訓練方面，首先定義聚集事後機率(aggregate *a posteriori*, AAP)

$$J = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N_m} \frac{P(\mathbf{X}_{m,n} \mid \lambda_m) P_m}{P(\mathbf{X}_{m,n})} \tag{16}$$

$\mathbf{X}_{m,n}$ 代表模型 m 的第 n 個訓練資料，長度為 T_n，即 $\mathbf{X}_{m,n} = \{\mathbf{x}_{m,n,t}\}_{t=1}^{T_n}$，$P_m$ 為類別 m 的事前機率，假設訓練資料分佈為 independent, identically distributed (i.i.d)，因此 $\mathbf{X}_{m,n}$ 與 λ_m 的相似度可表示為 $P(X_{m,n} \mid \lambda_m) = \prod_{t=1}^{T_n} P(x_{m,n,t} \mid \lambda_m)$。為了具有鑑別式訓練的形式，將(16)式改寫為

$$\tilde{J} = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N_m} l(d_{m,n}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N_m} l_{m,n} \tag{17}$$

其中 l 為(6)式 sigmoid function。

$$d_{m,n} = \log P(\mathbf{X}_{m,n} \mid \lambda_m) P_m - \log \sum_{j \neq m} P(\mathbf{X}_{m,n} \mid \lambda_j) P_j \tag{18}$$

為了讓正確類別與競爭類別佔有不同的百分比，在(18)式裡第二項乘上 L，$0 < L \leq 1$，當 $L = 1$ 時，代表正確類別與競爭類別具同樣重要性，同時令(6)式 sigmoid function 內 $\gamma = 1$，$\theta = 0$ 時，$\tilde{J} = J$，$P(\mathbf{X}_{m,n} \mid \lambda_m)$ 為 GMM 函式，為了令 $\tilde{J}$ 為最大，因此對 $\tilde{J}$ 取 gradient 並令為零可得到

$$\begin{aligned} \nabla_{\theta_{mi}} \tilde{J} = & \sum_{n=1}^{N_m} \sum_{t=1}^{T_n} \Omega_{m,i}(\mathbf{x}_{m,n,t}) \nabla_{\theta_{mi}} \log P(\mathbf{x}_{m,n,t} \mid \lambda_{m,i}) \\ & - L \sum_{j \neq m} \sum_{n=1}^{N_j} \sum_{t=1}^{T_n} \Omega_{j,i}(\mathbf{x}_{j,n,t}) \nabla_{\theta_{mi}} \log P(\mathbf{x}_{j,n,t} \mid \lambda_{m,i}) = 0 \end{aligned} \tag{19}$$

其中

$$\Omega_{m,i}(\mathbf{x}_{m,n,t}) = l_{m,n}(1 - l_{m,n}) \frac{c_{m,i} P(\mathbf{x}_{m,n,t} \mid \lambda_{m,i})}{P(\mathbf{x}_{m,n,t} \mid \lambda_m)} \tag{20}$$

$$\Omega_{j,i}(\mathbf{x}_{j,n,t}) = l_{j,n}(1 - l_{j,n}) \frac{c_{m,i} P(\mathbf{x}_{j,n,t} \mid \lambda_{m,i}) P_m}{\sum_{k \neq j} P(\mathbf{x}_{j,n,t} \mid \lambda_k) P_k} \tag{21}$$

為了得到模型參數的封閉解(close-form solution)，這裡假設(20)與(21)式與模型參數獨立，若欲求平均值向量，將(19)式 $\log P(\mathbf{x}_{m,n,t} \mid \lambda_{m,i})$ 對平均值向量取偏微分後可得

$$\nabla_{\mu_{m,i}} \log P(\mathbf{x}_{m,n,t} \mid \lambda_{m,i}) = \Sigma_{m,i}^{-1}(\mathbf{x}_{m,n,t} - \mu_{m,i}) \tag{22}$$

將上式代入(19)式移項後可得平均值向量的解為

$$\mu_{m,i} = \frac{\sum_{n=1}^{N_m} \sum_{t=1}^{T_n} \Omega_{m,i}(\mathbf{x}_{m,n,t}) \mathbf{x}_{m,n,t} - L \sum_{j \neq m} \sum_{n=1}^{N_j} \sum_{t=1}^{T_n} \Omega_{j,i}(\mathbf{x}_{j,n,t}) \mathbf{x}_{j,n,t}}{\sum_{n=1}^{N_m} \sum_{t=1}^{T_n} \Omega_{m,i}(\mathbf{x}_{m,n,t}) - L \sum_{j \neq m} \sum_{n=1}^{N_j} \sum_{t=1}^{T_n} \Omega_{j,i}(\mathbf{x}_{j,n,t})} \tag{23}$$

2.4 線性迴歸語者調適

根據語音模型與語者間之相關性可分為語者獨立(speaker-independent, SI)語音模型及語者相依

(speaker-dependent, SD)語音模型。使用語者相依之語音模型，在辨識時，須先行指定或偵測要使用的語者模型組別，而語者獨立則不須，以此差別看來，語者相依之語音辨識系統，使用上較不便，且需儲存多組語音模型。相對來說，使用語者獨立語音模型時所需要的語音模型數量會較少且模型特性與每一位測試語者均不甚吻合。所以，辨識率會較差。一般而言，使用語者相依語音模型的辨識系統效能會比語者獨立之辨識系統效能高二至三倍[7]。

為了保留兩者優點，一般皆訓練出一組語者獨立的語音模型，取其模型總數量較少的優點，而以此模型為基礎，再利用一些由測試語者所錄得之調適語料，先調適出與該語者語音特性較相符的語音模型，即所謂的語者相依語音模型，可有效提升語音辨識率。不過用於調整的語料一般並不多，容易造成調適語料稀疏的問題，為了解決樣本數不足的問題，做法是將語音模型分群，為每一群的語音模型找出一個參數轉換矩陣，群集內的模型調整只要依照此轉換矩陣即可得到更新後參數。為了得到更新後的轉換矩陣，可以利用不同的法則，較常見的有最大相似度線性迴歸法則，最大事後機率線性迴歸法則，最小分類錯誤線性迴歸法則。

2.5 最大相似度線性迴歸(MLLR)

最大相似度線性迴歸的目標就是，對一群集 s，計算一轉換矩陣 W_s，使得群集內所有調適資料的相似度最大，最大相似度線性迴歸調適演算法的好處在於，調適語料不需要完全涵蓋所有模型，即使沒有調適資料的模型，也可以經由同類別的轉換矩陣進行調適。以調整平均值向量為例，在計算轉換矩陣之前，將平均值向量延展為

$$\xi_s = [1, \mu_1, \mu_2, \ldots, \mu_D]^T \tag{24}$$

其中，D 為向量維度，則更新後的平均值向量為

$$\hat{\mu}_s = \mathbf{W}_{r(s)} \xi_s \tag{25}$$

其中，$r(s)$ 代表狀態 s 所屬迴歸類別，$\mathbf{W}_{r(s)}$ 代表迴歸類別(regression class) $r(s)$ 的轉換矩陣，維度為 $D \times (D+1)$，則透過 EM 演算法，最後可以得到每一個迴歸類別的轉換矩陣之每一列計算方式如下

$$\mathbf{w}_i^T = \mathbf{G}^{(i)^{-1}} \mathbf{z}_i^T \tag{26}$$

$\mathbf{w}_i^T$ 和 $\mathbf{z}_i^T$ 分別代表 $\mathbf{W}$ 和 $\mathbf{Z}$ 的列向量[14]。

2.6 最大事後機率線性迴歸(MAPLR)

由於以最大相似度為主之線性轉換矩陣在計算上十分簡易，所以其應用十分普遍，然而，若調適語料過少，或語料特性不具代表性時，則可能導致得到的轉換矩陣仍舊無法符合測試語者的語音特性，於是，便考慮到引入轉換矩陣的事前分佈資訊。矩陣參數的事前分佈可以在估測轉換矩陣時限制參數可能的調適量，使得參數的估測更具強健性，由文獻實驗可看出，最大事後機率線性迴歸可達到比最大相似度線性迴歸更好的辨識率[21]。

2.7 最小分類錯誤線性迴歸(MCELR)

最小分類錯誤的鑑別式訓練方式在很多應用都能顯示出不錯的效能，不過最小分類錯誤一般以廣義機率遞減演算法實現，並沒有在理論上證明它能收斂到更好的模型，當訓練資料變少時，錯誤的收斂停止點更容易發生，因此將 MCE 應用在模型調適時，使用線性迴歸有其必要。*Chengalvarayan* 在 1998 年提出最小分類錯誤線性迴歸[4]，使用全域性的轉換矩陣並以廣義機率遞減演算法估測矩陣參數，實驗結果顯示出其調適效果比最大相似度線性迴歸演算法好。而在[10]中，更進一步使用多組迴歸類別的轉換矩陣進行調適，在同樣使用廣義機率遞減演算法下，可以有更好的調適效能改進。另外，在[9]中，作者不利用廣義機率遞減演算法實現最小分類錯誤線性迴歸調適演算法，而以一般化調適作法計算轉換矩陣，即轉換矩陣以群集為單位，將最小分類錯誤的目標函式改寫後，可以透過 EM 演算法以封閉解的方式計算轉換矩陣。

3. 聚集事後機率線性迴歸鑑別式調適法

在最小分類錯誤估測法則中，並不考慮類別的事前資訊，且使用廣義機率遞減演算法實現，在調適資料少時，更容易發生錯誤訓練的問題，因此，Beyerlin 將所有模型(語音模型、語言模型)組成一個事後機率的線性組合，利用鑑別式訓練估測出線性組合的係數[2]。由先前所介紹的一般化最小錯誤率[15][16]，從最大事後機率的角度出發，另外定義所謂聚集事後機率(AAP)，並將式子改寫為鑑別式訓練的形式，在所給定的部份假設下，可以得到鑑別式訓練的封閉解，相較於傳統使用的廣義機率遞減演算法，有較快的計算速度，而且不用調整學習速率(learning rate)和步進大小(step size)。由於調適時資料較少，於是將一般化最小錯誤率代入尋找轉換矩陣也應該相當合適。

3.1 聚集事後機率線性迴歸(AAPLR)與最小分類錯誤線性迴歸(MCELR)之關係

考慮到最大事後機率在少量訓練語料下可以得到比最大相似度較正確的模型參數，由前述的一般化最小錯誤率介紹中可以看出，它將事後機率中原本與模型參數無關的 $P(\mathbf{x}_{m,n})$ 表示成與模型相關，即具鑑別式訓練的形式，將原本最小分類錯誤中鑑別式函式為相似度函式改為事後機率函式，可以結合這兩種模型估測方式的優點，並利用封閉解的解法可以快速估測出模型參數，改善以往以廣義機率遞減法則實作時收斂太慢的缺點。

由於語音模型調適時資料量通常較少，因此將一般化最小錯誤率的方式導入將有助於參數的估測，我們將此調適的方式稱為聚集事後機率線性迴歸(AAPLR)調適演算法。為了以 AAPLR 的方式計算轉換矩陣，且加入轉換矩陣的事前資訊可以讓其估測較具強健性，因此將(16)式聚集事後機率改寫為

$$J=\sum_{m=1}^{M}\sum_{n=1}^{N_m}\frac{p(\mathbf{x}_{m,n}\mid\hat{\mathbf{W}}_{r(m)};\Lambda)P_m g(\hat{\mathbf{W}}_{r(m)})}{p(\mathbf{x}_{m,n})}, \tag{27}$$

在繼續推導聚集事後機率線性迴歸演算法前，我們將透過 EM 演算法，發掘最大事後機率線性迴歸與使用最小分類錯誤準則之參數估測演算法之間的差異。給定一語音觀察樣本序列 $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_T\}$，其長度為 T，且存在線性轉換矩陣集合 $\mathbf{W}=\{\mathbf{w}_1,\ldots,\mathbf{w}_R\}$，其中共有 R 組類別。則在給定觀察樣本序列 $\mathbf{X}$ 時，線性轉換矩陣的事後機率可表示如下

$$g(\mathbf{W}\mid\mathbf{X};\Lambda) \tag{28}$$

其中，Λ 表示用於相似度計算之語音模型集合。而上述之事後機率又可以透過貝氏法則轉換如下之相似度與事前機率之組合

$$\begin{aligned} g(\mathbf{W}\mid\mathbf{X};\Lambda)&=\frac{p(\mathbf{X}\mid\mathbf{W};\Lambda)g(\mathbf{W})}{p(\mathbf{X})}\\ &=\prod_{t=1}^{T}\frac{p(\mathbf{x}_t\mid\mathbf{W};\Lambda)g(\mathbf{W})}{p(\mathbf{x}_t)} \end{aligned} \tag{29}$$

此處之 $g(\mathbf{W})$ 代表線性轉換矩陣 $\mathbf{W}$ 之事前分佈機率。再進一步將(29)式對數化可得

$$\log g(W\mid X;\Lambda)=\sum_{t=1}^{T}\log\frac{p(x_t\mid W;\Lambda)g(W)}{p(x_t)} \tag{30}$$

在 EM 演算法中之 E-step 即用於計算以下之輔助函式

$$\begin{aligned} R(\hat{\mathbf{W}}\mid\mathbf{W})&=E\left\{\log\frac{p(\mathbf{X},\mathbf{q}\mid\hat{\mathbf{W}};\Lambda)g(\hat{\mathbf{W}})}{p(\mathbf{X})}\middle|\mathbf{X},\mathbf{W}\right\}\\ &=\sum_{i=1}^{M}\sum_{t=1}^{T}p(q_t=i\mid\mathbf{x}_t,\mathbf{W};\Lambda)\log\frac{p(\mathbf{x}_t,q_t=i\mid\hat{\mathbf{W}};\Lambda)g(\hat{\mathbf{W}})}{p(\mathbf{x}_t)} \end{aligned} \tag{31}$$

其中，$\mathbf{q}=(q_1,q_2,\ldots,q_T)$ 表示給定之觀察序列 $\mathbf{X}$ 之每一時間點所對應之狀態序列。$\hat{\mathbf{W}}$ 表示透過 EM 演算法欲估測之新轉換矩陣參數，而 $\mathbf{W}$ 則是現有透過 EM 演算法在前一次 M 步驟中所估測出之最佳轉換矩陣參數。令 $\gamma_i(\mathbf{x}_t)=p(q_t=i\mid\mathbf{x}_t,\mathbf{W};\Lambda)$ 用以表示在第 t 個時間點，觀察樣本 $\mathbf{x}_t$ 停留於第 i 個狀態之機率，則(31)式之輔助函式可簡單表示為

$$\begin{aligned} R(\hat{\mathbf{W}}\mid\mathbf{W})&=E\left\{\log\frac{p(\mathbf{X},\mathbf{q}\mid\hat{\mathbf{W}};\Lambda)g(\hat{\mathbf{W}})}{p(\mathbf{X})}\middle|\mathbf{X},\mathbf{W}\right\}\\ &=\sum_{i=1}^{M}\sum_{t=1}^{T}\gamma_i(\mathbf{x}_t)\log\frac{p(\mathbf{x}_t,q_t=i\mid\hat{\mathbf{W}};\Lambda)g(\hat{\mathbf{W}})}{p(\mathbf{x}_t)} \end{aligned} \tag{32}$$

在此，我們使用維特比(Viterbi)近似法則來簡化我們的式子。於是，我們使用最佳的狀態序列來取代原有需考慮所有可能性之表示法，同時上述之狀態停留機率 $\gamma_i(\mathbf{x}_t)$ 則簡化如下

$$\gamma_i(\mathbf{x}_t)=\begin{cases}0 & q_t=i\\ 1 & q_t\neq i\end{cases} \tag{33}$$

此外，為了與以下之聚集事後機率比較，我們將使用下列定義之符號重新表示(32)式。使用 m 來表示原有之狀態標示 i，即將狀態視為語音模型類別；使用 $\mathbf{x}_{m,n}$ 取代原有之 $\mathbf{x}_t$。因為原有之觀察樣本 $\mathbf{x}_t$ 在經過維特比解碼器對應出最佳之狀態後，即可以明確知道兩者間之關連。所以用 $\mathbf{x}_{m,n}$ 來表示原有觀察樣本 $\mathbf{x}_t$ 為對應至第 m 類模型之第 n 個觀察樣本。則(32)式可以重新表示為

$$R(\hat{\mathbf{W}} \mid \mathbf{W}) = \sum_{m=1}^{M}\sum_{n=1}^{N_m} \log \frac{p(\mathbf{x}_{m,n}, m \mid \hat{\mathbf{W}}_{r(m)}; \Lambda) g(\hat{\mathbf{W}}_{r(m)})}{p(\mathbf{x}_{m,n})}. \tag{34}$$

其中，$\hat{\mathbf{W}}_{r(m)}$ 表示該線性轉換矩陣是用於轉換第 m 類語音模型參數之用。一般而言，線性轉換矩陣是根據所有語音模型參數中具相似特性之分群結果而分為數個類別，如分為 R 群，被分於同群之語音模型是共用同一組轉換矩陣進行轉換。於是在給定語音模型類別 m 後，即可以透過上述之關係，得到對應之轉換矩陣類別。另外，我們是以 $r(m)$ 表示第 r 類轉換矩陣與第 m 類語音模型之關係。

從另一方面來看，遵循上述變數、標示之定義，則轉換矩陣 $\mathbf{W}$ 之聚集事後機率定義即為(27)式，從(27)式與(34)式比較可知，在使用 EM 演算法對語音模型或是此處所考慮之轉換矩陣之參數進行估測時，是將第(34)式之 $R(\hat{\mathbf{W}} \mid \mathbf{W})$ 針對所欲估測之參數予以偏微分後，而透過封閉解來得到更新的參數內容。而在聚集事後機率的定義式中，則是將各個類別之事後機率全部加總起來，於是在文獻中接下來的推導過程中，才可朝所謂的最小分類錯誤之鑑別式參數估測之同理性進行推導，並經一些假設設定後，得以使用封閉解的方式進行參數內容之更新。

3.2 聚集事後機率線性迴歸(AAPLR)參數估測

接下來將利用(27)式進行模型參數的估測，與一般化最小錯誤率一樣，同樣可將(27)式改寫為一目標函式為

$$\tilde{J} = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N_m} l(d_{m,n}) = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N_m} l_{m,n} \tag{35}$$

$$d_{m,n} = \log p(\mathbf{x}_{m,n} \mid \lambda_m) P_m g(\mathbf{W}_{r(m)}) - \log \sum_{j \neq m} p(\mathbf{x}_{m,n} \mid \lambda_j) P_j g(\mathbf{W}_{r(m)}) \tag{36}$$

其中，$g(\mathbf{W}_{r(m)})$ 為轉換矩陣 $\mathbf{W}_{r(m)}$ 的事前機率分佈，$r(m)$ 代表模型 m 的迴歸類別，$g(\mathbf{W}_{r(m)})$ 為一矩陣版本高斯分佈，稱作 elliptically symmetric distribution 或 matrix variate normal distribution。

$$g(\mathbf{W}_{r(m)}) \propto \left|\Delta\right|^{-\frac{1}{2}} \cdot q\left(\sum_{d=1}^{D} (\mathbf{w}_{r(m)d} - \mathbf{m}_{r(m)d}) \Sigma_d^{-1} (\mathbf{w}_{r(m)d} - \mathbf{m}_{r(m)d})^T \right) \tag{37}$$

$\mathbf{q}$ 為一個 $[0,\infty)$ 的函式，$\mathbf{w}_{r(m)d}$ 和 $\mathbf{m}_{r(m)d}$ 分別代表轉換矩陣和平均矩陣的第 d 列向量，維度為 $1\times(D+1)$，Δ 為一維度 $D(D+1)\times D(D+1)$ 的區塊對角化共變異矩陣(block diagonal covariance matrix)，每一區塊由 $(D+1)\times(D+1)$ 的 Σ_d 組成。為了簡化最佳轉換矩陣，首先將加入轉換矩陣的高斯分佈改寫為一單變量形式如下

$$N(\mathbf{x}_{m,n} \mid \xi_{m,i}, \Sigma_{m,i}) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_{m,i}\right|^{1/2}} \exp\left[-\frac{1}{2}\sum_{d=1}^{D} \frac{(\mathbf{x}_{m,n,d} - \mathbf{w}_{r(m)d}\xi_{m,i})^2}{\sigma_{m,i,d}^2} \right] \tag{38}$$

$\mathbf{w}_{r(m)d}$ 代表轉換矩陣 $\mathbf{W}_{r(m)}$ 的第 d 列向量。將(36)和變更過的高斯分佈(38)代入(35)式得到 AAPLR 的目標函式並對欲求的轉換矩陣第 d 列 $\mathbf{w}_{r(m)d}$ $(d=1,\ldots,D)$ 取偏微分得

$$\begin{aligned}
\nabla_{\mathbf{w}_{r(m)d}} J = & \sum_{m=1}^{M_c}\sum_{n=1}^{N_m} l(d_{m,n})(1-l(d_{m,n})) \\
& \times \left[\begin{array}{l}
\displaystyle\sum_{i=1}^{I_m} \frac{c_{m,i} N(\mathbf{x}_{m,n} \mid \lambda_m, \mathbf{W}_{r(m)})}{P(\mathbf{x}_{m,n} \mid \lambda_m, \mathbf{W}_{r(m)})} \left(\frac{\mathbf{x}_{m,n,d} - \mathbf{w}_{r(m)d}\xi_{m,i}}{\sigma_{m,i,d}^2} \right) \xi_{m,i}^T \\
+2(\mathbf{w}_{r(m)d} - \mathbf{m}_{r(m)d}) \Sigma_{r(m)d}^{-1} \\
\displaystyle - \sum_{\substack{m' \in \mathbf{W}_{r(m)} \\ m' \neq m}} \frac{P(\mathbf{x}_{m,n} \mid \lambda_{m'}, \mathbf{W}_{r(m)}) g(\mathbf{W}_{r(m)})}{\displaystyle\sum_{j \neq m} P(\mathbf{x}_{m,n} \mid \lambda_j, \mathbf{W}_{r(j)}) g(\mathbf{W}_{r(j)})} \\
\times \left[\displaystyle\sum_{i=1}^{I_{m'}} \frac{c_{m',i} N(\mathbf{x}_{m,n} \mid \lambda_{m'}, \mathbf{W}_{r(m)})}{P(\mathbf{x}_{m,n} \mid \lambda_{m'}, \mathbf{W}_{r(m)})} \left(\frac{\mathbf{x}_{m,n,d} - \mathbf{w}_{r(m)d}\xi_{m',i}}{\sigma_{m',i,d}^2} \right) \xi_{m',i}^T \right. \\
\left. +2(\mathbf{w}_{r(m)d} - \mathbf{m}_{r(m)d}) \Sigma_{r(m)d}^{-1} \right]
\end{array} \right]
\end{aligned} \tag{39}$$

令上式為零，移項後可得 $\mathbf{w}_{r(m)d}$ 的解為

$$\mathbf{w}_{r(m)d}\left(\begin{array}{l}\sum_{m=1}^{M_c}\sum_{n=1}^{N_m}\sum_{i=1}^{I_m} l(d_{m,n})(1-l(d_{m,n}))\Omega_{m,i}(\mathbf{x}_{m,n})\frac{1}{\sigma_{m,i,r}^2}\xi_{m,i}\xi_{m,i}^T \\ -2\sum_{m=1}^{M_c}\sum_{n=1}^{N_m} l(d_{m,n})(1-l(d_{m,n}))\Sigma_{r(m)d}^{-1} \\ -\sum_{m=1}^{M_c}\sum_{n=1}^{N_m}\sum_{\substack{m'\in\mathbf{W}_c \\ m'\neq m}}\sum_{i=1}^{I_{m'}} l(d_{m,n})(1-l(d_{m,n}))\omega_{m',n,i}(\mathbf{x}_{m,n})\Omega_{m',i}(\mathbf{x}_{m,n})\frac{1}{\sigma_{m',i,d}^2}\xi_{m',i}\xi_{m',i}^T \\ +2\sum_{m=1}^{M_c}\sum_{n=1}^{N_m}\sum_{\substack{m'\in\mathbf{W}_c \\ m'\neq m}} l(d_{m,n})(1-l(d_{m,n}))\omega_{m',n,i}(\mathbf{x}_{m,n})\Sigma_{r(m)d}^{-1}\end{array}\right)$$

=

$$\left(\begin{array}{l}\sum_{m=1}^{M_c}\sum_{n=1}^{N_m}\sum_{i=1}^{I_m} l(d_{m,n})(1-l(d_{m,n}))\Omega_{m,i}(\mathbf{x}_{m,n})\frac{\mathbf{x}_{m,n,d}}{\sigma_{m,i,d}^2}\xi_{m,i}^T \\ -2\sum_{m=1}^{M_c}\sum_{n=1}^{N_m} l(d_{m,n})(1-l(d_{m,n}))\mathbf{m}_{r(m)d}\Sigma_{r(m)d}^{-1} \\ +\sum_{m=1}^{M_c}\sum_{n=1}^{N_m}\sum_{\substack{m'\in\mathbf{W}_c \\ m'\neq m}}\sum_{i=1}^{I_{m'}} l(d_{m,n})(1-l(d_{m,n}))\omega_{m',n,i}(\mathbf{x}_{m,n})\Omega_{m',i}(\mathbf{x}_{m,n})\frac{\mathbf{x}_{m,n,d}}{\sigma_{m',i,d}^2}\xi_{m',i}^T \\ +2\sum_{m=1}^{M_c}\sum_{n=1}^{N_m}\sum_{\substack{m'\in\mathbf{W}_c \\ m'\neq m}} l(d_{m,n})(1-l(d_{m,n}))\omega_{m',n,i}(\mathbf{x}_{m,n})\mathbf{m}_{r(m)d}\Sigma_{r(m)d}^{-1}\end{array}\right) \tag{40}$$

其中

$$\Omega_{m,i}(\mathbf{x}_{m,n}) = \frac{c_{m,i}N(\mathbf{x}_{m,n}\mid\lambda_m,\mathbf{W}_{r(m)})}{P(\mathbf{x}_{m,n}\mid\lambda_m,\mathbf{W}_{r(m)})} \tag{41}$$

$$\omega_{m',n,i}(\mathbf{x}_{m,n}) = \frac{P(\mathbf{x}_{m,n}\mid\lambda_{m'},\mathbf{W}_{r(m)})g(\mathbf{W}_{r(m)})}{\sum_{j\neq m}P(\mathbf{x}_{m,n}\mid\lambda_j,\mathbf{W}_{r(j)})g(\mathbf{W}_{r(j)})} \tag{42}$$

令(40)式的等號左側為 $\mathbf{w}_{r(m)d}\cdot\mathbf{L}$，$\mathbf{L}$ 為一維度 $(D+1)\times(D+1)$ 的方陣，且令等號右側為 $\mathbf{r}$，維度 $1\times(D+1)$，(40)式變為 $\mathbf{w}_{r(m)d}\cdot\mathbf{L}=\mathbf{r}$，可得轉換矩陣 $\mathbf{W}_{r(m)}$ 的第 d 列為

$$\mathbf{w}_{r(m)d}=\mathbf{r}\cdot\mathbf{L}^{-1} \tag{43}$$

重覆上述步驟，可求得所有迴歸類別的轉換矩陣。

4. 實驗與討論

4.1 實驗環境與語音參數、模型設定

在實驗的硬體方面，我們所使用的是 Pentium 4 2.0GHz 個人電腦，搭配 256MB 的記憶體容量，使用的作業系統為 Windows XP Professional 中文版，並以 Microsoft Visual C++ 6.0 作為軟體開發工具。在語音特徵參數求取部份，使用 HTK 中的 HCopy 指令取出語音特徵參數，每一音框的特徵參數皆為 26 維，其中包括 12 階的 MFCC，12 階的 delta MFCC，1 階 log energy 以及 1 階 delta log energy，關於 HCopy 的詳細介紹使用方法以及求取特徵參數的設定檔和 HTK 的其他命令使用方法等，請參考[22]。本實驗所使用的辨識系統是以連續密度隱藏式馬可夫模型(continuous density hidden Markov model, CDHMM)為架構，以中文之聲母與韻母作為 HMM 之基本單元，在聲母部份使用 3 個狀態表示，而韻母則使用 5 個狀態來表示。同時，混合數數量則根據各個狀態所分配到的音框數量來決定，但最大混合數不得超過 32 個。

4.2 實驗語料

在本實驗中，我們分別使用兩種語料以進行實驗，其一是 TCC300 麥克風語料庫，用於語者獨立之語音模型參數訓練；另一個則是公視晚間新聞語料，用於調適後之辨識效能改進評估。以下是這兩個語料庫的資料說明。

TCC 台大/成大/交大麥克風語音資料庫是由國立台灣大學、國立成功大學、國立交通大學各自擁有之語料庫集合而成，各校錄製之目的是為語音辨認研究，屬於麥克風朗讀語音。其中台大語料庫主要包含詞及短句，內容經過仔細設計，考慮了音節及其相連出現機率，由 100 人錄製而成；成大及交大語料庫主要包含長文語料，文章由中研院提供之 500 萬詞詞類標示語料庫中選取，每篇文章包含數百字，再切割成 3-4 段，每段含至多 231 字，由 200 人朗讀錄製，每人所讀文章皆不相同。在本論文的實驗中，我們所取的部份為交通大學及台灣大學所錄製的音檔。語音訊號取樣頻率為 16kHz，語音訊號量化析度為 16 位元。

公視晚間新聞語料是由中央研究院與公共電視臺共同錄製，主要是公視晚間新聞語音，錄製期間由 2000 年 1 月 11 日到 2000 年 2 月 9 日，總共 120 小時的新聞語料。

4.3 實驗結果與討論

首先，我們使用 TCC300 語料庫進行語者獨立之語音模型訓練，語料數共約 14000 句。訓練所得之語音模型，我們使用 TCC300 中另外未拿來訓練之語料進行測試共 900 句，其語音辨識率為 67.5%。另一份我們所使用之公視晚間新聞語料，初步已整理出三小時語料。所以在接下來的效能測試部份，我們使用此三小時語料進行實驗。我們將使用不同的調適語料量，分別進行最大相似度線性迴歸(MLLR)、最大事後機率線性迴歸(MAPLR)、最小分類錯誤線性迴歸(MCELR)與本論文所提之聚集事後機率線性迴歸(AAPLR)之效能評估。調適之語料量由最短之 2 句，至最長之 30 句，而調適之轉換矩陣類別為 2 類，進行效能之評估，實驗結果如下表所示。

調適方法	調適句數	矩陣類別數	辨識率(%)	調適時間(分鐘)
Baseline	-	-	44.9	-
MLLR	2	2	46.3	2
	5	2	54.1	3
	30	2	56.6	10
MAPLR	2	2	51.5	2
	5	2	54.1	3
	30	2	56.3	10
MCELR	2	2	48.2	2
	5	2	54.1	4
	30	2	56.8	13
AAPLR	2	2	51.5	2
	5	2	54.6	3
	30	2	57.1	11

表一、MLLR, MAPLR, MCELR, AAPLR 在不同調整句數下之辨識率與調整時間比較

首先我們從不同方法的調適效果來比較，可以發現所提出之 AAPLR 與其他調適方法相較，無論給定多少調適語料，均可達到最佳之效能。而與 MCELR 之比較，可以發現最大之效能差距約有 3.3%。另外，由調適時間來比較，可以發現，AAPLR 雖然算是屬於鑑別性調適法則，但是在調適時間上，由於其參數估測有封閉解的存在，可以一次就將調適之最佳參數估測出，所以較同類型之 MCELR 花更短的時間在調適上。另外，由表上可以發現的是，當使用了 30 句調適語料時，所有方法的調適效果並沒有相當大的改進，推測原因應是出在轉換矩陣類別數量上的問題。由於使用之語料數量已不少，但是類別數量還是只有固定在 2 個，過少的轉換矩陣類別數，會使得調適語料無法發揮針對不同模型參數而估測出專屬之轉換矩陣，而失去大量調適語料應有之調適效能改進率。最後，在此初步實驗中，我們直接將 TCC300 所訓練出之語音模型，使用公視語料進行少量語料之調適效能實驗，而未考慮到兩種語料所具備之文句內容與語者分佈的差異。在此實驗結果中，不易區分出調適之效能是來自於針對文句內容的調適效能抑或是來自語者的調適效能。這是在未來我們將會再進行修正之處。

5. 結論與未來工作

在本研究中，我們提出一套具鑑別性訓練特性之快速調適演算法，聚集事後機率線性迴歸(AAPLR)調適演算法。根據最小錯誤率之原則，我們由事後機率出發，定義聚集事後機率函式進行線性迴歸矩陣參數之調適。此調適演算法之優點在於整合最大事後機率的調適演算法則與鑑別式訓練的精神。既可獲得鑑別式訓練必須考量其他類別

與所估測類別參數間之鑑別性法則而得到較傳統最大相似度估測更好的分類正確率，又從推導最終結果之封閉解而可獲得快速調適的效能。在實驗中，我們可以看到無論在任何調適資料量之下，所提出之調適演算法之效能可以比其他同樣基於線性迴歸調適為主之演算法有更好的效能表現。

在本論文中，一般化最小錯誤率中之類別機率以一常數表示，在模型參數估測中較不具參考價值，或許嘗試以真正的機率分佈來代表，可以推導出更完整之結果。此外，我們也將再深入由最基本之理論出發，將此一調適演算法演繹得更加完整。未來我們也將嘗試利用近似貝氏的方法進行理論推導以尋求漸進式調適之效能。此外，除了我們也將增加線性轉換矩陣的類別數，進行更多的實驗以驗證調適效能之外，也要採行先針對訓練與測試語料之文句內容差異進行所謂的 task 調適，以先去除此一因素，再行針對語者調適之效能進行實驗評估，我們也將增加回歸類別數目以及調整語料句數以更有效提高電視新聞語音辨識率。

6. 參考文獻

[1] L. Bahl, P. Brown, P. de Souza and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition", in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 11, April 1986, pp. 49-52.

[2] P. Beyerlin, "Discriminative model combination", in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 1, 1998, pp. 481-485.

[3] P. C. Chang and B.-H. Juang, "Discriminative training of dynamic programming based speech recognizers", *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 2, pp. 135-143, April 1993.

[4] R. Chengalvarayan, "Speaker adaptation using discriminative linear regression on time-varying mean parameters in trended HMM"*, IEEE Trans. Signal Processing Letters*, vol. 5, pp. 63-65, March 1998.

[5] Jen-Tzung Chien, "Quasi-Bayes linear regression for sequential learning of hidden Markov models", *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, pp. 268-278, July 2002.

[6] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society (B)*, vol. 39, pp. 1-38, 1977.

[7] M. J. F. Gales and P. C. Woodland, "Mean and Variance adaptation within the MLLR Framework," *Computer Speech and Language*, Vol. 10, pp. 249-264, 1996.

[8] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observation of Markov chains", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 291-298, April 1994.

[9] X. He, W. Chou, **"**Minimum classification error linear regression for acoustic model adaptation of continuous density HMMs", in *Proc. Int. Conf. Multimedia and Expo (ICME)*, vol. 1, 2003, pp. 6-9.

[10] W. Jian, H. Qiang, "Supervised adaptation of MCE-trained CDHMMs using minimum classification error linear regression**,"** in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing (ICASSP)*, vol. 1, 2002, pp. I-605 - I-608.

[11] B.-H. Juang, W. Hou and C.-H. Lee, "Minimum classification error rate Methods for Speech Recognition", *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 3 , pp. 257-265, May 1997.

[12] B.-H. Juang and S. Katagirl, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol. 40, pp. 3043-3054, December 1992.

[13] H.-K.J. Kuo, E. Fosle-Lussier, H. Jiang and C.-H. Lee, "Discriminative training of language models for speech recognition", in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 1, 2002, pp. I-325-I-328.

[14] C. J. Leggeter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, 1995, pp. 171-185.

[15] Q. Li, B.-H. Juang, "A new algorithm for fast discriminative training", in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 1, 2002, pp. 97-100.

[16] Q. Li, B.-H. Juang, "Fast discriminative training for sequential observations with application to speaker identification", in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 2, 2003, pp. 397-400.

[17] R. P. Lippmann, "An intorduction to computing with neural nets", *IEEE ASSP Mag*., pp. 4-22, April 1987.

[18] E. McDermott and S. Katagiri, "Shift-invariant multi-category phoneme recognition using kohonen's LVQ2," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1989, pp. 81-84.

[19] A. Sankar and C.-H. Lee**, "**A maximum-likelihood approach to stochastic matching for robust speech recognition", *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 3, pp. 190-202, May 1996.

[20] R. Schlüter, W. Macherey, B. Müller and H. Ney, "A combined maximum mutual information and maximum likelihood approach for mixture density splitting", in Proc. *EUROSPEECH*, vol. 4, 1999, pp. 1715-1718.

[21] O. Siohan, C. Chesta, and C.-H. Lee. "Hidden Markov model adaptation using maximum a posteriori linear regression." in *Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, 1999.

[22] S. Young, J. Jansen, J. Odell, D. Ollason, and P Woodland. *The HTK Book (Version 2.0)*. ECRL, 1995.

非監督式學習於中文電視新聞自動轉寫之初步應用

郭人瑋　蔡文鴻　陳柏琳

國立台灣師範大學資訊工程研究所

{rogerkuo, louis, berlin}@csie.ntnu.edu.tw

摘要. 本論文探討非監督式學習於中文電視新聞自動轉寫之初步應用。在聲學模型訓練上，我們提出以發音確認(Utterance Verification)技術來克服訓練語料沒有正確人工轉寫的問題，所謂的發音確認是使用候選詞信心度評估(Candidate Word Confidence Measure)來對某語句及其轉寫進行篩選的動作，用以決定此語句及轉寫是否有足夠的可靠度，進而成為訓練語料。我們先使用大詞彙連續語音辨識器對龐大且無人工轉寫的語料進行自動轉寫，再使用發音確認(Utterance Verification)針對辨識後的語料進行篩選，從中擷取較正確可靠的語料片段，以供聲學模型訓練使用，此舉不僅可大大節省人力成本，在效果上，經訓練過的聲學模型也和單純以人工轉寫結果所訓練出來的模型相距不遠；同時，較正確可靠的文字語料片段，則用於語言模型調適，以增進辨識效能。同樣地，候選詞信心度評估也被應用到非監督式聲學模型調適上，我們初步將它與「最大相似度線性迴歸」(Maximum Likelihood Linear Regression)聲學模型調適技術作結合，以語音辨識所產生之詞圖(Word Graph)作為調適標的。我們以公共電視台的新聞語料為研究題材，結果顯示非監督式聲學模型訓練與調適的結合的確可有效降低字錯誤率(Character Error Rate)，驗證了此作法之可行性。

1　序論

隨著科技快速發展，日常生活中能取得的多媒體影音資訊愈來愈多，如廣播電視節目、演講稿和數位典藏等。這些多媒體資訊早已成為傳統文字資訊外，可供社會大眾廣泛使用之資訊。例如在廣播及電視新聞語音辨識和資訊檢索技術的發展上，近年來已有許多的研究和令人鼓舞的成果陸續被發表出來[1]-[4]。但為了要以語音辨識技術來自動轉寫這些影音資訊，尤其當我們想在新的應用領域建立一套語音辨識系統時，通常必須仰賴大量經由人工轉寫(Manually Transcribed)的語料來供給聲學模型訓練使用，才可達到不錯的辨識效果，但這往往既耗人力又費時間。有鑑於此，近幾年來開始有一些研究，嘗試發展以近乎非監督式(Lightly Supervised)的方式，結合廣播或電視新聞節目對應字幕(Closed Caption)，嘗試從大量龐雜的語料中擷取較可靠的語句片段供聲學模型訓練，使語音辨識系統的準確性更為提升，或能於新的應用領域中迅速建立起新的雛形系統，目前也有一些初步的成果被發表出來[5]-[6]。但是，上述作法的前提為廣播或電視新聞節目對應字幕必須事先提供才能進行。從先前的研究中[7]，我們提出以發音確認(Utterance Verification)的技術來克服訓練語料沒有正確人工轉寫的問題，先使用大詞彙連續語音辨識器對龐大且無人工轉寫的語料進行語音辨識，利用信心度評估(Confidence Measure)對自動轉寫的語料進行篩選，擷取較為正確可靠的自動轉寫語料片段，達到非監督式(Unsupervised)聲學模型訓練的目的。嚴格來說，先前的研究中，所使用的語料都是錄製於數個相同的廣播電台[4]，但仍屬於同一領域內的非監督式聲學模型訓練。因此，本論文嘗試作跨領域的非監督式聲學模型訓練之研究，希望以少量含有正確人工轉寫的廣播新聞語料為基礎[4]，利用非監督式聲學模型訓練方式，建立一個語音辨識雛形系統，處理公共電視台的新聞語音資料(簡稱公視新聞語料或MATBN)[8]-[9]。公視新聞語料為中央研究院資訊所口語小組與公共電視台所合作完成[10]，現在也開始廣為國內各大學及研究機構所使用，如台大、交大、成大等學校都已有初步的研究成果。我們將對公視新聞語料作整理及統計，並定義一些訓練語料及測試語料，對本論文中所提出的方法加以實驗評估。

另一方面，聲學模型調適(Acoustic Model Adaptation)在語音辨識中一直扮演著相當重要的角色，為的就是要補償聲學模型訓練環境與測試環境不匹配所造成的問題，進而提高辨識率。在聲學模型調適方法中最常被使用的調適技術為「最大事後機率(Maximum *a Posteriori,* MAP)」[11]與「最大相似度線性迴歸(Maximum Likelihood Linear Regression, MLLR)」[12]。前者(MAP)視聲學模型參數為一組隨機變數，並為它假設一組對應的事前機率分佈(Prior Distributions)加以限制。當調適語料量多時，調適的效果漸近於「最大相似度估測(Maximum Likelihood Estimation, MLE)訓練」[13]；若調適語料不足時，調適後的模型

參數愈接近原始的模型參數。因此提供了良好的強健性。後者(MLLR)試著為聲學模型中的高斯分佈，因統計特性相近所形成的迴歸群集(Regression Classes)求取一共同的轉換矩陣，再藉由調適語句的參與，使群集內的高斯分佈參數經由轉換矩陣旋轉平移後，對所屬之調適語句得到最大的相似度，就算無調適語料的聲學模型參數，也能藉由共享相同群集的轉換矩陣來做調適，當調適語料不多時，效果則較「最大事後機率」來的顯著。同樣地，聲學模型調適也有監督式與非監督式之分，它們的最大差別在於，後者的調適語句沒有對應的正確人工轉寫(Manual Transcription)，語句的自動轉寫(Automatic Transcription)需先經由一次語音辨識產生，再以此進行聲學模型調適。倘若第一次的轉寫存在大量的錯誤，將連帶地將影響後續的聲學模型調適，錯誤的累積會使辨識率的進展愈來愈受侷限。在應用上，非監督式聲學調適技術較具實際應用價值，但在論文的發表上，大部分仍是以監督式聲學模型調適的實驗為主。故在本論文中，我們將候選詞信心度評估應用於非監督式聲學模型調適的研究，初步配合「最大相似度線性迴歸」的模型調適方法，以信心度評估對自動轉寫的調適語料作適當的加權，並使用語音辨識過程中產生的詞圖(Word Graph)作為聲學模型調適的環境[14]，嘗試研究如何從詞圖的豐富資訊中擷取出適合於非監督式聲學模型調適的資訊，增進聲學模型的準確性。最後，我們也將結合語言模型調適，兩種以最大事後機率(Maximum *a Posteriori*, MAP)為基礎之語言模型調適(Language Model Adaptation)技術：詞頻數混和(Count Merging)和語言模型插補(Language Model Interpolation)[15]-[16]，也初步作為應用自動轉寫用於語言模型調適的方法。

本論文的安排如下：第二節將描述台灣師範大學資工所的新聞語音辨識系統及實驗的語料，第三節將說明我們所使用的發音確認技術，第四節將提出非監督式學習的架構，包含非監督式聲學模型訓練、自動轉寫用於語言模型調適及非監督式聲學模型調適等方法，第五節將報告我們的實驗數據及討論，最後將作總結，並敘述我們正進行的各項研究。

2 新聞語音辨識系統與實驗語料

2.1 台師大資工所新聞語音辨識系統

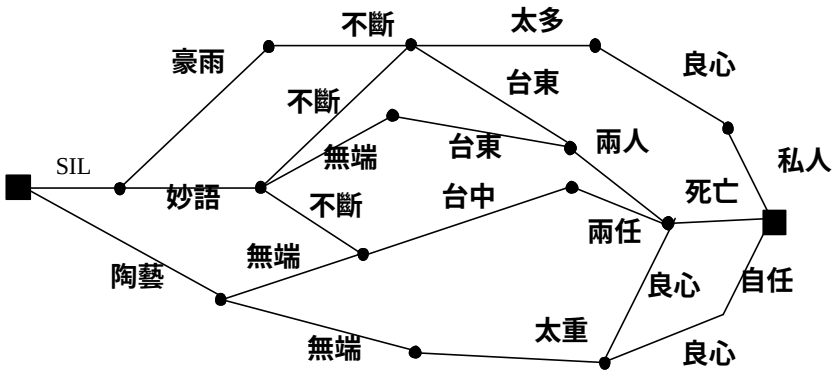
在本節中，我們將扼要介紹台灣師範大學資工所目前所發展的新聞語音辨識系統，它基本上是一套大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition)系統，主要包括前端處理(Front-end Processing)、聲學模型訓練(Acoustic Model Training)、詞典的建立(Lexicon Construction)、語言模型訓練(Language Model Training) 和詞彙樹複製搜尋(Tree-Copy Search)等部分。同時，我們也將介紹與分析本論文中所使用的廣播新聞語料與公視新聞語料。

2.1.1 前端處理與聲學模型訓練

在本論文中我們使用梅爾倒頻譜特徵向量(Mel-frequency Cepstral Coefficients，簡稱MFCC特徵向量)作為語音訊號的特徵參數。在求取MFCC特徵向量時，我們將語音資料切割成一連串部分重疊的音框，每一個音框(Frame)由13維的梅爾倒頻譜特徵加上其一階與二階的時間軸導數(Time Derivatives)所形成的39維特徵向量所組成。其中13維的梅爾倒頻譜特徵是由18個梅爾頻譜上濾波器組(Filter Banks) 的輸出經餘弦轉換求得。同時，為了降低通道效應對語音辨識的影響，我們使用倒頻譜平均消去法(Ceptral Mean Subtraction，簡稱CMS)。另外，在辨識所需的聲學模型訓練上，考慮了中文語音結構，聲學模型由22個INITIAL模型、38 FINAL模型(每個中文的音節都是由一個INITIAL及一個FINAL所組成)及一個靜音(Silence) 模型組成，其中INITIAL模型會因其右邊可能接的FINAL模型種類而進一步細分成112個INITIAL模型[4]。我們最後總共使用了151個隱藏式馬可夫模型(Hidden Markov Models)來作為這些INITIAL-FINAL聲學模型的統計模型。在隱藏式馬可夫模型中，每個狀態則依據其對應到的訓練語料多寡，以2到128個高斯統計分佈來表示，不管男女性別都使用同一套聲學模型。

2.1.2 詞典建立及語言模型訓練

在中文裡約有7000個單字詞，新詞可由此7000個單字詞合併產生，我們可根據字詞在語料中的統計特性，以自動化的方式產生新的複合詞(Compound Words)。新增複合詞的自動產生方式如下面所述：對於語料中任意相鄰的兩個詞(w_i, w_j)，我們分別計算它們的前雙連(Forward Bigram)機率$P_f(w_j \mid w_i)$，與後雙連



圖一、詞彙複製搜尋所產生之詞圖，為所有可能候選詞的簡潔表示。

(Backward Bigram)機率 $P_f(w_i \mid w_j)$，並以前後雙連(Forward and Backward Bigrams)的機率幾何平均 $FB(w_i, w_j) = \sqrt{P_f(w_j \mid w_i) P_b(w_i \mid w_j)}$，作為$(w_i, w_j)$是否合併的依據[7]。文字語料先經由一個含有一至四字詞約六萬八千個詞的詞典來斷詞，然後利用上述的公式，經數次的疊代以及不同的基準閾值(Threshold)設定，產生約五千個二至十字詞的複合詞，使得最後的語音辨識詞典約含有七萬二千個一至十字詞。在語言模型的使用上，我們使用了詞雙連以及詞三連語言模型(Word Bigram and Trigram Language Models)，並以從中央通訊社(Central News Agency)2000與2001年所收集到的約一億七千萬個中文字語料作為背景語言模型訓練時的訓練資料[17]。在本論文中的語言模型使用了Katz語言模型平滑技術[18]，在訓練時是採用SRL Language Modeling Toolkit (SRILM)，它是一套相當方便且容易使用的語言模型研究工具軟體[19]。

2.1.2 詞彙樹複製搜尋

我們發展的大詞彙連續語音辨識方法是採用由左至右(Left-to-right)、音框同步(Frame-synchronous) 的詞彙樹搜尋方式[20]。在詞彙樹中每個分枝(Arc)代表一個INITIAL或FINAL的隱藏式馬可夫模型，由樹根(Root)到任一個樹梢(Leaf)的路徑代表一個詞或一些發音相同的詞，路徑上的分枝就是代表這個詞或這些詞會使用到的隱藏式馬可夫模型。具體來說，我們採用所謂的詞彙樹複製搜尋演算法(Tree-copy Search)，搜尋時每個音框會同時存在數棵詞彙樹複製(Tree Copies)，每個詞彙樹代表不同的語言模型歷史或限制(Language Model History or Constraint)。實際上，搜尋時產生的不完全路徑(Partial Paths)如果擁有相同的語言模型歷史會被歸類在同一棵詞彙樹複製裡，進行隱藏式馬可夫模型狀態層次(State-level)維特比動態規劃搜尋(Viterbi Dynamic Programming Search)。在每個音框中，若有不完全路徑已抵達樹梢時，代表一個完整詞已可被產生；同時，不同棵詞彙樹複製間已抵達樹梢的不完全路徑，若具有相同的語言模型歷史， 則會進行再結合(Recombination)，保留最大分數者，並以它們的語言模型歷史為標註，產生新的一棵詞彙樹複製，或加入到一棵已存在且具有相同語言模型歷史的詞彙數複製中。值得注意的是，在實作時並不需要真的建立如此多的詞彙樹複製，僅需建立一棵詞彙樹作為搜尋時路徑展開參考之用即可，並分別紀錄搜尋時存活下來的隱藏式馬可夫模型狀態節點(也就是不完全路徑目前拜訪到的節點)的相關資訊。另一方面，由於存活的隱藏式馬可夫模型狀態節點可能會隨音框數呈指數倍增加，因此必須以光束剪裁(Beam Pruning)技術適當地剪裁分數較低的狀態節點或不完全路徑。在執行剪裁動作時會同時考量每一個詞彙樹複製內部狀態節點(Internal Node)下涵蓋的可能拜訪樹梢節點代表之所有詞對應的語言模型機率，並以其中最大者當做每一個詞彙樹複製內部狀態節點的語言模型前看分數(Language Model Look-ahead Score)[20]，再加上內部狀態節點本身搜尋時所累積的解碼分數(Decoding Score)當成剪裁比較的依據。在本研究中，我們採用的是詞單連語言模型前看(Word Unigram Language Look-ahead)，對每一個詞彙樹複製內部狀態節點，我們會以其所在分枝(或隱藏式馬可夫模型)之可能拜訪樹梢節點中具最大詞單連語言模型機率，做為該內部狀態節點的語言模型前看分數。此外，在每個音框，我們會紀錄存活的詞彙樹複製樹梢節點中分數較高者的相關資訊(這些樹梢節點本身代表著可能的候選詞)，諸如它們的語言模型歷史、對應候選詞開始與結束的音框以及搜尋時聲學解碼的分數(Acoustic Decoding Scores)，然後再依此資訊建立起一個詞圖(Word Graph)，如圖一所示。並且在這詞圖上使用更高階的語言模型，如

表一、公視新聞語料(NTNU_SA-2)訓練集統計資訊。長度小於了2秒的主播語句在本研究中被排除，男女生語料長度的比約為1:8。

訓練語料(主播部分)	總時間(小時)	句數(句)	平均句長(秒/句)	最長句長(秒)	最短句長(秒)	佔比例(%)	性別
林建成	1.47	422	12.53	55.91	2.01	9.71	男
馬紹	0.13	35	13.30	26.58	5.29	0.86	男
葉明蘭	12.98	2,860	16.34	68.92	2.08	85.85	女
洪蕙竹	0.48	127	13.66	30.94	2.70	3.19	女
蘇怡如	0.06	17	12.58	34.21	5.55	0.39	女
總計	15.12	3,461	-	-	-	100.00	2男3女
平均	-	-	15.73	68.92	2.01	-	-

詞三連、詞四連語言模型等，重新進行一次動態規劃搜尋，找出最佳的詞句。在本研究中，我們在詞彙樹複製搜尋階段是使用詞雙連語言模型，而在詞圖搜尋(Word Graph Rescoring)階段是使用詞三連語言模型。

2.2 廣播及電視新聞語料

本研究所使用的的中文廣播新聞語音語料總共有176小時以上，全是透過收音機收錄，為1998年11月至2004年4月之間由台北地區數家廣播電台所播送之新聞節目。所有的語料都經由人工切割為一則一則的新聞語音檔，每一則新聞均由一個主播所播報，性別上男女都有。某些檔案因錄音的關係，含有相當大的背景雜訊。這些廣播新聞語料僅有少部分有對應的正確人工轉寫，其中有大約4小時語料收錄於1998至1999年，用來作為初始的聲學模型訓練。而電視新聞語料則全為公視新聞語料(MATBN)，為中央研究院資訊所口語小組耗時三年與公共電視台合作錄製完成，預計將收錄220小時的廣播新聞，所有的新聞語料都有正確的人工轉寫以及其它的標註資訊(如：停頓、語助詞、呼吸、強調語氣、反覆、不適當的發音)，所有的人工轉寫與標註均使用DGA&LDC的轉寫器(Transcriber)來完成。每天的新聞約含有二十多則報導，每則報導為一完整主題。除了語音資料，文字語料在其它應用上也有很大的價值(如資訊檢索、主題偵測與文章分段)。公視新聞語料大致上可分內場及外場兩個部份，內場部分主要為主播(Studio Anchors)的語料，外場部分主要為記者(Field Reporters)與受訪者(Interviewees)的語料。經由統計，MATBN2002與MATBN2003共120小時的語料內，只含有五位主播，由於本語料以新聞內容為主，主播不會有大幅度的變動，其中以「葉明蘭」主播的語料佔絕大多數，約85%，使得要在內場中定義出一套較具代表性的訓練及測試語料，顯得有些困難，希望未來能經由國內各相關研究機構及人士的集思廣益與討論，為這套資訊豐富的新聞語料，定義出有實驗價值的訓練及測試語料，作為技術開發的比較平台。我們由MATBN2002與MATBN2003中選擇了內場約16小時的語料作為本實驗的語料(NTNU_SA-2)[21]，包含了約15小時的內場主播語料供訓練與約44分鐘(0.74小時)的測試語料，統計資料如表一及表二所示。訓練語料中，佔有85%語料的主播葉明蘭，也是測試語料內唯一的語者，使得本實驗之聲學模型有著語者相依(Speaker-dependent)的缺失，但本論文強調於完全非監督模式下進行學習，包含聲學模型訓練、聲學模型調適及語言模型調適，相較於初始系統，辨識率上仍有明顯的進步。

3 發音確認

發音確認利用候選詞信心度評估(Candidate Word Confidence Measure) (3.3節介紹)來決定某語句是否予以挑選成為非監督式訓練的語料，候選詞信心度評估包含了候選詞事後機率(Candidate Word Posterior Probability) (3.1節介紹)與聲學信心(Acoustic Confidence Measure) (3.2節介紹)兩個部份。

表二、公視新聞語料(NTNU_SA-2)測試集統計資資訊。長度小於了2秒的主播語句在本研究中被排除，語料共約44分鐘。

測試語料(主播部分)	總時間(小時)	句數(句)	平均句長(秒/句)	最長句長(秒)	最短句長(秒)	佔比例(%)	性別
葉明蘭	0.74	163	16.28	38.50	2.57	100.00	女
總計	0.74	163	-	-	-	100.00	1女
平均	-	-	16.28	38.50	2.57	-	-

3.1 候選詞事後機率

由詞彙樹複製搜尋所產生的詞圖(Word Graph)是存放語音辨識過程中所有可能候選詞(Candidate Word Hypotheses)的簡潔表示[14]，包含了分數較高的樹梢節點相對應的分支，每一分支即代表一個詞，包含起始時間、結束時間及搜尋時聲學解碼的分數。候選詞的事後機率則可利用不同階層的語言模型，利用Forward-Backward演算進行詞圖搜尋(Word Graph Rescoring)，候選詞的事後機率的估測如下[22][23]：

$$CM_{Posterior}(w_{t_s}^{t_e} \mid X) = p(w_{t_s}^{t_e} \mid X) = \frac{p(w_{t_s}^{t_e}, X)}{p(X)} = \frac{\sum_{W_1^{t_s-1}} \sum_{W_{t_e+1}^{T}} p(W_1^{t_s-1} \cdot w_{t_s}^{t_e} \cdot W_{t_e+1}^{T}, X)}{\sum_{W_1^{T}} p(W_1^{T}, X)}, \quad (1)$$

其中$w_{t_s}^{t_e}$為起始時間t_s和結束時間t_e的候選詞；

X為起始時間1及結束時間T的聲學特徵向量序列；

$W_{t_1}^{t_2}$則為起始時間t_1，結束時間t_2的候選詞序列(Candidate Word Hypothesis Sequence)；

$p(w_{t_s}^{t_e} \mid X)$為給定聲學特徵向量序列X，候選詞$w_{t_s}^{t_e}$的事後機率，由於此事後機率也常被用來表示詞的信心度，我們以$CM_{Posterior}(w_{t_s}^{t_e} \mid X)$來表示$w_{t_s}^{t_e}$的事後機率；

$p(W, X)$為候選詞序列W與X的聯合機率，包含了聲學及語言模型解碼(Decoding)分數。

3.2 聲學信心

在另一方面，我們可對$w_{t_s}^{t_e}$求出其聲學信心，設$SUB = \{sub_1, \cdots, sub_{N_w}\}$為$w_{t_s}^{t_e}$內的次詞序列(Subword Sequence)，N_w為SUB內次詞單位(Subword Unit,在本研究中為INITIAL或FINAL)之個數。設sub_i為起始時間$t_{i,s}$，結束時間$t_{i,e}$之次詞，則聲學信心所使用的公式如下：

$$CM_{Acoustic}(w_{t_s}^{t_e}) = \frac{1}{N_w} \sum_{i=1}^{N_w} \frac{2}{1 + \exp\left[-t \cdot LLR(sub_i) + h\right]}, \quad (2)$$

$$where \;\; LLR(sub_i) = \log \frac{p(X_{t_{i,s}}^{t_{i,e}} \mid sub_i)}{\max_{sub} p(X_{t_{i,s}}^{t_{i,e}} \mid sub)},$$

其中$CM_{Acoustic}(w_{t_s}^{t_e})$為給定$X$時，$w_{t_s}^{t_e}$的聲學信心，$t$及$h$分別用來調整指數函數的成長率與平移；

$p(X_{t_{i,s}}^{t_{i,e}} \mid sub)$為在$X_{t_1}^{t_2}$下，$sub$的相似度(Likelihood)；$LLR(sub)$為$sub$與擁有最大相似度的第一名次詞單位之對數相似度比值(Likelihood Ratio)。

3.3 候選詞信心度評估

候選詞信心度評估包含了候選詞事後機率 (3.1)及聲學信心(3.2)，就前者而言，雖然對詞圖上每一候選詞都能求其事後機率，但根據觀察，以愈高階的語言模型進行詞圖搜尋，候選詞之間的事後機率差異愈是懸殊，例如以三連語言模型進行詞圖搜尋時，第一名詞序列(Top1 Word Sequence)中的候選詞往往佔有超過0.95的事後機率，換句話說，語言模型所用的階層(Order)愈高，則候選詞事後機率愈受語言模型所影響，第一名詞序列的事後機率會出奇的高。若以此事後機率作為信心度評估，難免對第一名詞序列產生偏頗，有鑑於此，我們引入信心度比例係數(Confidence Scale Factor)，將原先候選詞事後機率的刻度(Scale)加以調整，使之成為合理的候選詞事後機率。候選詞事後機率經修正後如下：

$$CM^{a}_{Posterior}(w_{t_s}^{t_e} \mid X) = p(w_{t_s}^{t_e} \mid X) = \frac{\left[\sum_{W_1^{t_s-1}} \sum_{W_{t_e+1}^{T}} p(W_1^{t_s-1} \cdot w_{t_s}^{t_e} \cdot W_{t_e+1}^{T}, X)\right]^a}{\sum_{W_1^T} \left[p(W_1^T, X)\right]^a}. \tag{3}$$

公式(3)符號定義與公式(1)相同，其中 a 為信心度比例係數(Confidence Scale Factor)，a 介於 0 與1 之間，表示對聯合機率施以壓縮，使候選詞間的事後機率差異變小。當 a 等於1 時，則表示刻度不變；當 a 等於 0 時，事後機率為均勻機率(Uniform Probability)。其中，$CM^{a}_{Posterior}(w_{t_s}^{t_e} \mid X)$ 為給定聲學特徵向量序列 X，信心度比例係數為 a 時，候選詞 $w_{t_s}^{t_e}$ 的事後機率。

候選詞信心度評估則包含了候選詞事後機率及聲學信心，公式如下：

$$CM(w_n \mid X) = c_1 \cdot CM_{Acoustic}(w_n \mid X) + c_2 CM^{a}_{Posterior}(w_n \mid X), \tag{4}$$

其中 c_1 與 c_2 為權重參數，在以下的非監督式聲學模型訓練中，我們將設 $c_1 = c_2 = 0.5$。

3.4 發音確認

我們提出發音確認(Utterance Verification)之技術，來決定某語句是否予以挑選成為非監督式訓練的語料。發音確認可視為一個決斷函數 $V(X, W, Thr) \in \{accept, reject\}$，根據平均候選詞信心度評估，來決定自動轉寫產生的第一名詞序列 $W = \{w_1, \cdots, w_N\}$ 是否能成為訓練語料。決斷函數 V 定義如下：

$$V(X, W, Thr) = \begin{cases} accept & if \quad \frac{1}{N}\sum_{n=1}^{N} CM(w_n \mid X) \geq Thr \\ reject & otherwise \end{cases}, \tag{5}$$

其中，X 為對應的聲學特徵向量序列，W 為自動轉寫產生的第一名詞序列，Thr 為篩選基準閥值。當平均信心度評估大於篩選基準閥值時，則決斷函數輸出為*accept*，表示 X 值得我們採用為非監督式訓練的語料，W 為其對應的自動轉寫；若輸出為*reject*，則表示不予採用。

4 非監督式學習

4.1 非監督式聲學模型訓練

我們先使用大詞彙連續語音辨識系統(聲學模型由四小時廣播新聞語料來訓練)對十五小時的公視訓練語料(共3,461句)進行自動轉寫，根據辨識結果，每句可再藉由靜音(Silence)切成數個子句，少於五個中文字的子句將被排除，最後有15,473個子句被留下。對每個子句，我們以辨識結果第一名的詞序列當作此子句對應之詞序列，進行發音確認決定此子句是否予以採用。若此子句被留下來作為非監督式聲學模型訓練的語料。則其對應的自動轉寫片段，也將被留下作為自動轉寫用於語言模型調適(4.3節將會介紹)的文字語料。

4.2 非監督式聲學模型調適

大多數的非監督式聲學模型調適僅取第一次辨識所產生的第一名詞序列來做聲學模型調適的依據。然而語音辨識的錯誤可能會對聲學模型調適造成影響，使得調適效果有限[24]。本研究中，我們嘗試使用候

選詞信心度評估為詞圖上的候選詞進行加權，使得每一個候選詞依其信心度評估分數對模型調適都有不同程度的貢獻。但由於計算詞圖上所有候選詞聲學信心的計算量相當大，因此，在聲學模型調適中，我們只使用了事後機率。我們初步地將它與「最大相似度線性迴歸」(Maximum Likelihood Linear Regression, MLLR)聲學模型調適技術做結合。最大相似度線性迴歸的調適技術需先為聲學模型中的高斯分佈加以分群，因統計特性相近而形成的群集稱為迴歸群集(Regression Classes)，根據相似度最大的估測法則對每一迴歸群集求取轉換矩陣，使群集內的高斯分佈參數經此轉換矩陣旋轉平移後，相對應的調適語句能得到最大的相似度，就算調適語料無涵蓋所有的聲學模型，迴歸群集內的高斯分佈也能藉此轉換矩陣來得到調適。

由於非監督式調適沒有正確的人工轉寫，我們須先經由一次的語音辨識來產生語句的相關資訊。實驗中對於測試語句進行非監督式聲學模調適的步驟如下：

1. 測試語句經由詞彙樹複製搜尋(Tree-Copy Search)，產生詞圖(Word Graph)。
2. 利用Forward-Backward演算法在詞圖上進行詞圖搜尋(Word Graph Rescoring)，為詞圖上的每一候選詞求出其對應的事後機率$CM^{a}(w_{t_s}^{t_e}|X)$，其中a為信心度比例係數。
3. 針對每一候選詞語音段落，再使用一次狀態層次(State Level) Forward-Backward演算法，為每一音框(Frame) t及狀態(State) i求其事後機率$g_t(i|w_{t_s}^{t_e})=\Pr(s_t=i|X_{t_s}^{t_e},w_{t_s}^{t_e})$。
4. 最後，將$g_t(i|w_{t_s}^{t_e})$乘上所屬候選詞的事後機率$CM^{a}(w_{t_s}^{t_e}|X)$，並對所有候選詞語音段落加總。可得音框t時，狀態i的事後機率$g_t(i)=\Pr(s_t=i|X_1^T)=\sum_{w_{t_s}^{t_e}} CM^{a}(w_{t_s}^{t_e}|X)\,g_t(i|w_{t_s}^{t_e})$。

重覆上述步驟，收集MLLR模型調適時所需的統計量，並進行MLLR模型調適。

4.3 自動轉寫用於語言模型調適

統計式語言模型(Statistical Language Models)旨在以統計的方法分析及模擬自然語言的規律特性，並以機率量化的方式來決定一個詞串的接受程度。在過去二十年間，一直是語音及語言處理領域中重要的課題。在統計式語言模型中，N連語言模型(N-gram Language Models)是最常被使用的(尤其是二連及三連語言模型)，它主要根據前面的N-1詞歷史(Word History)來決定下一個詞可能出現的機率[25][16]。N-連語言模型的機率表示，通常由最大相似度(Maximum Likelihood Estimation, MLE)來估測，然而，在特定領域下訓練N-連語言模型時，為了解決統計模型訓練時資料稀疏的問題(Data Sparseness Problems)，過去幾年，已經有一些像平滑(Smoothing)或插補(Interpolation)等方法陸續被提出，達到不錯的效果[18]。但另一方面，在處理一些較複雜困難的語音辨識課題上如廣播及電視新聞自動轉寫，由於新聞播報的主題和語言內容的詞彙使用具多變性與時效性，會使得統計式語言模型往往很難做到準確的估測，於是便有了所謂的語言模型調適(Language Model Adaptation)的研究[16]。語言模型調適通常會結合背景文字語料庫(Background Corpus)與測試語音同一時期(Contemporary)或者是同一領域(In-domain)的文字語料庫來訓練出較具強健性的調適後語言模型，以得到較佳的詞接連預測能力，而在過去已有一些不錯研究被發表出來[26]-[27]。在本研究我們嘗試研究使用語音辨識產生的電視新聞自動轉寫用於語言模型調適(Unsupervised Language Model Adaptation)的可行性，直接以非監督式聲學模型性訓練時經發音確認篩選過後的語音片段對應的自動轉寫文字(參見4.1節)，當成同一領域文字語料來做語言模型的調適。我們初步使用兩種常用的語言模型調適技術：語言模型插補(Language Model Interpolation)及詞頻數混合(Count Merging)[15][7]，並比較這兩種由傳統貝氏估測所發展出來的語言模型調適技術。詞頻數混合和語言模型插補的調適公式分別如下(以三連語言模型為例)：

$$\tilde{P}_{Adapt1}(w_i|w_{i-2}w_{i-1})=\frac{m_1\cdot C_{d,Cont}(w_{i-2}w_{i-1}w_i)+m_2\cdot C_{d,Back}(w_{i-2}w_{i-1}w_i)}{m_1\cdot C_{Cont}(w_{i-2}w_{i-1})+m_2\cdot C_{Back}(w_{i-2}w_{i-1})}, \tag{6}$$

及

$$\tilde{P}_{Adapt2}(w_i|w_{i-2}w_{i-1})=g\cdot P_{Cont}(w_i|w_{i-2}w_{i-1})+(1-g)\cdot P_{Back}(w_i|w_{i-2}w_{i-1}). \tag{7}$$

在第(6)式中，$C_{d,Cont}(w_{i-2}w_{i-1}w_i)$與$C_{d,Back}(w_{i-2}w_{i-1}w_i)$分別代表調適語料中及背景訓練語料中的三連減值詞頻(Trigram Discounted Count)，而$C_{Cont}(w_{i-2}w_{i-1})$與$C_{Back}(w_{i-2}w_{i-1})$則分別代表調適語料中及背景訓練語料中的二連詞頻，m_1與m_2則為可調整的權重參數。在第(6)式中，$P_{Cont}(w_i|w_{i-2}w_{i-1})$與$P_{Back}(w_i|w_{i-2}w_{i-1})$分別代表由調適語料及背景訓練語料所估測的三連機率，g為可調整的參數，公式(6)及公式(7)的詳細推導可參考[15]。詞頻數混合是在詞機率估測前，將領域內(In-domain)文字語料與背景(Background)文字語料在詞頻

表三、基礎實驗與非監督式聲學模型調適之語音辨識結果：嘗試改變信心度比例係數α 與計算候選詞事後機率時語言模型的階層。 MLLR(Top1)為傳統只取用第一名辨識結果詞序列來做MLLR調適；MLLR(CM)為引入信心度評估的MLLR調適。字錯誤率減少百分比為相對於無聲學模型調適之字錯誤率。

計算候選詞事後機率時所用的語言模型階層	三連語言模型		二連語言模型	
聲學模型調適	字錯誤率 (%)	相對字錯誤率減少百分比(%)	字辨識率 (%)	相對字錯誤率減少百分比(%)
無	27.67	-	27.67	-
MLLR(Top1)	25.93	6.29	25.93	6.29
MLLR(CM), $\alpha = 1$	25.80	6.76	26.12	5.60
MLLR(CM), $\alpha = 1/4$	25.69	7.16	25.92	6.32
MLLR(CM), $\alpha = 1/8$	25.80	6.76	25.95	6.22
MLLR(CM), $\alpha = 1/12$	25.37	8.31	25.49	7.88
MLLR(CM), $\alpha = 1/16$	25.26	8.71	25.54	7.70
MLLR(CM), $\alpha = 1/20$	25.14	9.14	25.73	7.01
MLLR(CM), $\alpha = 1/24$	25.38	8.28	25.82	6.69
MLLR(CM), $\alpha = 1/28$	25.51	7.81	25.93	6.29

數空間(Frequency Space)上給予權重加總，進而估測機率；而語言模型插補則是估測個別模型之機率後，才根據權重於機率空間(Probability Space)上相加。

5 實驗結果與討論

5.1 實驗環境與非監督式聲學模型調適基礎實驗

我們使用台師大資工所發展的新聞語音辨識系統，並以普遍被使用的梅爾倒頻譜特徵向量作為語音特徵參數。初始的聲學模型由四小時的廣播新聞語料所訓練而成，初始背景語言模型則由從中央通訊社收集的新聞語料訓練而得。這一小節的基礎實驗有三個目的：第一、計算候選詞事後機率時，比較不同階層語言模型帶來的影響；第二、計算候選詞事後機率時，改變信心度比例係數，討論它們對字錯誤率所帶來的影響；第三、比較傳統只取用第一名辨識結果詞序列(Top1)來作調適與使用信心度評估(CM)來作調適的結果。表三為本節基礎實驗的結果。我們在3.3節中曾提到，語言模型所用的階層(Order)愈高，則候選詞事後機率愈受語言模型所影響，第一名詞序列的事後機率會出奇的高。若以此事後機率作為信心度評估，難免對第一名詞序列產生偏頗。這是當語言模型階層愈高，會使得特定詞彙擁有較大的機率(根據訓練語料的特性)，使得聯合機率差距愈趨懸殊，連帶影響候選詞的事後機率。由表三中可見，使用三連語言模型計算候選詞事後機率時，最佳的字錯誤率(25.14%)出現在$\alpha = 1/20$時，若使用二連語言模型時，最佳的字錯誤率(25.49%)出現在$\alpha = 1/12$時。這也驗證了使用較高階語言模型時，將會造成特定詞彙事後機率刻度(Scale)過高的不合理現象，需要使用較小的信心度比例係數加以調整。由於使用三連語言模型明顯比使用二連語言模型要來的好，故往後的實驗中，均使用三連語言模型來計算候選詞事後機率時。信心度比例係數的決定，和系統的語音辨識率有密切的關係，當辨識率較高時，應有較大的信心度比例係數，信任第一階段辨識產生的結果；反之，則信心度比例係數應較小。使用三連語言模型計算候選詞事後機率時，雖然$\alpha = 1/20$時，我們可得較佳的結果，但考量往後的實驗，我們將加上非監督式聲學模型訓練，系統的語音辨識率會再提升，同時為了兼顧一般化，我們在往後的實驗中，信心度比例係數均設為1/16。在表三中，傳統只取用第一名辨識結果詞序列來作調適MLLR(Top1) 之後，可得到6.29%的相對字錯誤率減少百分比。而在引入信心度評估以詞圖資訊來作調適MLLR(Top1)之後，則可達到9.14%的

表四、非監督式聲學模型訓練在使用不同基準閥值下的語音辨識結果。Thr為非監督式聲學模型訓練用以選取語句之基準閥值，MLLR(CM)為引入信心度評估的MLLR調適，α 在此設為1/16。同一列中，MLLR括弧內的數據為相對於無聲學模型調適時字錯誤率減少百分比。最後一列的「監督式訓練」為對照組。

	字錯誤率(%) (相對字錯誤率減少百分比(%))		
	無聲學模型調適	MLLR(Top1)	MLLR(CM)
原來四小時訓練之聲學模型	27.67	25.93 (6.29)	25.26 (8.71)
+ 3.80小時(Thr=0.9)	21.37	21.00 (1.73)	20.97 (1.87)
+11.57小時(Thr=0.8)	20.09	20.00 (0.45)	19.56 (2.64)
+13.30小時(Thr=0.7)	20.25	20.01 (1.19)	19.71 (2.67)
+13.61小時(Thr=0.6)	20.18	19.94 (1.19)	19.59 (2.92)
+13.67小時(Thr=0.5)	20.21	20.01 (0.99)	19.69 (2.57)
+13.70小時(Thr=0.0)	20.32	20.07 (1.23)	19.76 (2.76)
+15.12小時(監督式訓練)	16.26	16.29 (-0.18)	16.47 (-1.29)

相對字錯誤率減少百分比($\alpha = 1/20$)，有近3%的改善，顯示本論文所提出結合信心度評估和詞圖資訊的非監督式聲學模型調適方法，的確能有效的降低字錯誤率。

5.2 非監督式聲學模型訓練實驗結果

我們根據不同基準閥值(Threshold Values)進行語句的篩選，進行非監督式聲學模型訓練。訓練時我們使用HTK Toolkit [28]，進行三次的嵌入式訓練(Embedded Training)。實驗結果如表四所示，在非監督式訓練下，使用發音確認，特別在基準閥值為0.8時，我們可得最佳的字錯誤率20.09%，再經由非監督式聲學調適之後，更可達到19.56%的字錯誤率。由於聲學模型若經監督式訓練為非監督式訓練的上限，但仍有16.26%的字錯誤率，使得非監督式訓練與監督式訓的差距不到4%，說明了非監督式聲學模型訓練有其利用的價值。基準閥值的設定與訓練語料的多寡必須加以妥協，當基準閥愈高，則留下的訓練語料愈少，模型參數則無法有效的估測，如基準閥值為0.9時，錯誤率不降反升；若基準閥值太低，留下的語料雖多，但錯誤標註的語料反而會影響了模型參數估測的正確性。當基準閥值為0.8時，訓練語料總時間還留有11個小時，僅有16%不到的語料被篩除，這表示在0.8之上應存在更佳的基準閥值。在對照組「監督式訓練」的聲學模型中，可發現MLLR的調適反而帶來負面的影響，經過信心度評估的MLLR調適之後，字錯誤率攀升至16.47%，我們嘗試將信心度比例係數α 調小至1/4，則字錯誤率能降低至16.02%，驗證了在高辨識率的系統上應使用較小的信心度比例係數α 。

5.3 語言模型調適實驗結果

5.3.1 自動轉寫用於語言模型調適

我們將語音辨識產生的電視新聞自動轉寫用於語言模型的調適，進行了一些初步的語音辨識實驗。在模型插補的方法中，調適語言模型與背景語言模型的權重各為0.5(公式(7)中之$g = 0.5$)。詞頻數混合的實驗中我們根據訓練語料的大小，調適語言模型與背景語言模型的詞頻數加權比為250:1(公式(6)中之$m_1 = 250$、$m_2 = 1$)。實驗結果如表五所示，我們可觀察出自動轉寫中的詞連接規則資訊對語言模型仍能有不錯的貢獻，如當基準閥值為0.8時，相對於無語言模型調適，詞頻數混合可達到1.74% 的相對字錯誤率減少百分比。雖然以自動轉寫為基礎的適語料過於稀疏(Sparse)，(當基準閥值為0.9僅有約66K個字，即使在篩選閥值為0.0時，也只有約250K的字)，使得整體的字錯誤率下降幅度並不顯著。未來研究中，我們希望能藉由詞圖所提供豐富資訊來加以改善，並且以信心度評估為每一個候選詞的詞頻作加權，俾使詞圖上的候選詞均能對語言模型調適有所貢獻。

表五、自動轉寫用於語言模型調適的語音辨識結果。Thr為非監督式聲學模型訓練用以選取語句之基準閥值，括弧內之數據為相對於無語言模型調適之字錯誤率減少百分比。

聲學模型	調適語料字數	字錯誤率(%) (相對字錯誤率減少百分比(%))		
		無語言模型調適	語言模型插補	詞頻數混合
+ 3.80小時(Thr=0.9)	66,540	21.37	21.85 (-2.25)	21.08 (1.36)
+11.57小時(Thr=0.8)	209,489	20.09	19.97 (0.60)	19.74 (1.74)
+13.30小時(Thr=0.7)	242,630	20.25	20.06 (0.94)	20.27 (-0.10)
+13.61小時(Thr=0.6)	248,701	20.18	20.04 (0.69)	20.06 (0.59)
+13.67小時(Thr=0.5)	249,880	20.21	20.05 (0.79)	20.23 (-0.10)
+13.70小時(Thr=0.0)	250,640	20.32	20.02 (1.48)	20.18 (0.69)

5.3.2 領域內之語言模型調適

在這個實驗中，我們從公視新聞網[8]所收集的2001年與2002年文字語料(約五百萬個中文字)來做為語言模型調適語料，這些語料大多是新聞節目對應字幕(Closed Caption)。本研究訂立兩套調適語言模型來加以實驗：PTS_LM_1由2001年1月至2002年12月的公視新聞網語料所訓練，由於此語言模型訓練語料涵蓋的日期包含測試語料的那五天(2002年8月6日到2002年8月9日及2002年9月26日)，故我們稱之為偏差語言模型(Biased Language Model)。PTS_LM_2則排除2002年8月(含)之後的語料，由2001年1月至2002年7月的語料來進行訓練。其目的主要在於觀察領域內訓練語料的時效性對語言模型調適的影響。我們在此初步以使用監督式聲學訓練的聲學模型(15小時新聞語音資料)來進行實驗；在語言模型插補的方法中，調適語言模型與背景語言模型的權重各為0.5(公式(7)中$g = 0.5$)；在詞頻數混合的方法中，PTS_LM_1的加權比約為20:1(公式(6)中之$m_1 = 20$、$m_2 = 1$)、PTS_LM_2的加權比約為25:1(公式(6)中之$m_1 = 25$、$m_2 = 1$)。結果如表六所示，其中我們也顯示出當使用非監督式聲學模型調適後的辨識結果。在PTS_LM_1下，經過非監督式的聲學模型調適，辨識率可達92.67%(字錯誤率7.23%，語言模型插補)及92.77%(字錯誤率7.33%，詞頻數混合)；PTS_LM_2下，經過非監督式的聲學模型調適的辨識率僅有84.68%(字錯誤率15.32%，語言模型插補)及84.55%(字錯誤率15.45%，詞頻數混合)。由此可見，對電視新聞語音辨識來說，時效性對於語言模型的影響甚鉅。

6 結論與未來展望

本論文探討非監督式的聲學模型訓練與調適於中文電視新聞自動轉寫之初步應用。由實驗結果可觀察出發音確認能有效地挑選較為可靠的語料來進行訓練，節省大量的人力進行人工轉寫，使龐大的語料能被運用，篩選基準閥值的取決，影響了訓練的品質，如何在語料量與信心度評估找到平衡點，仍是一個課題；信心度評估也使得詞圖上更多的資訊能應用在非監督式聲學模型調適上，不再只侷限於Top1辨識的路徑，因此能解決非監督式調適時使用含有錯誤資訊的自動轉寫以及所需調適語料統計量過少的問題，但信心度比例係數α的調整則需考慮辨識率及語言模型的階層。自動轉寫用於語言模型調適能解決新聞辨識主題和語言內容的詞彙使用具多變性的問題，由於資料稀疏，使得字錯誤率的進步並不大，但由於詞圖上含有大量的資訊，我們甚至可根據詞圖上的信心度評估為每一個候選詞的詞頻作加權，俾使詞圖上的候選詞均能對語言模型調適有所貢獻。在寫此論文的同時，我們正將前端特徵值抽取部份，改用更有鑑別力的特徵向量，如線性鑑別分析(Linear Discriminant Analysis, LDA)及異質性鑑別分別(Heteroscedastic Discriminant Analysis, HDA)[29]，也試著將最大交互資訊(Maximum Mutual Information, MMI)訓練[30]與最小音素錯誤(Minimum Phone Error, MPE)訓練[31]等方法結合詞圖(Word Graph)的豐富語音辨識資訊，應用在非監督式聲學模型訓練上，以期能得到更好的語音辨識率。

表六、領域內語言模型調適的語音辨識結果。MLLR(CM)為引入信心度評估的MLLR調適，α 在此設為1/16。

	字錯誤率(%)		
	無聲學模型調適	MLLR(Top1)	MLLR(CM)
無語言模型調適	17.83	17.67	17.51
PTS_LM_1(語言模型插補)	7.46	7.32	7.23
PTS_LM_1(詞頻數混合)	7.47	7.39	7.33
PTS_LM_2(語言模型插補)	15.08	14.93	15.32
PTS_LM_2(詞頻數混合)	15.94	15.72	15.45

誌謝

本研究承蒙國科會「中文語音資訊辨識與檢索之研究」，編號：（91-2218-E-003-002-）及「中文語音資訊摘要技術之研究」，編號（92-2213-E-003-008-）等計畫補助。並感謝中研院口語小組提供公視新聞實驗語料及台大語音實驗室提供廣播新聞實驗語料。另外，也感謝三位審查委員所提供之意見。

參考文獻

[1] P. Beyerlein et al., “Large Vocabulary Continuous Speech Recognition of Broadcast News – The Philips/RWTH Approach,” *Speech Communication*, May 2002.
[2] P.C. Woodland, “The development of the HTK Broadcast News transcription system: An overview,” *Speech Communication*, May 2002.
[3] J. L. Gauvain, L. Lamel, G. Adda, “The LIMSI Broadcast News Transcription System,” *Speech Communication*, May 2002.
[4] B. Chen, H-M Wang, and L-S Lee, "Discriminating Capabilities of Syllable-Based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese", *IEEE Trans. on Speech and Audio Processing*, July 2002.
[5] L. Nguyen, B. Xiang, “Light Supervision in Acoustic Model Training,” in *Proc. ICASSP 2004*.
[6] L. Chen, L. Lamel and J. L. Gauvain, “Lightly Supervised Acoustic Model Training Using Consensus Networks,” in *Proc. ICASSP 2004*.
[7] B. Chen, J. W. Kuo, W. H. Tsai. “Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription,” in *Proc. ICASSP 2004*.
[8] **財團法人公共電視文化事業基金會-公共電視台**. http://www.pts.org.tw/.
[9] H. M. Wang. “MATBN 2002: A Mandarin Chinese Broadcast News Corpus,” in *Proc. SSPR’03*, Tokyo, Japan.
[10] **中央研究院資訊所中文組口語小組**. http://sovideo.iis.sinica.edu.tw/SLG/.
[11] J.-L. Gauvain, C.-H. Lee, “Maximum a Posteriori Estimation for Multivate Gaussian Mixture Observations of Markov Chains,” *IEEE Trans. on Speech and Audio Processing*, April 1994.
[12] M. J. F. Gales and P. C. Woodland (1996). “Mean and Variance Adaptation within the MLLR Framework,” *Computer Speech and Language*, pp.249-264, Vol. 10, 1996.
[13] L. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proceedings of the IEEE*, February 1989.
[14] S. Ortmanns, H, Ney, X Aubert, “A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition,” *Computer Speech and Language*, Vol. 11, 1997.
[15] M. Bacchiani, B. Roark, “Unsupervised Language Model Adaptation,” in *Proc. ICASSP 2003*.
[16] J. R. Bellegarda, “Statistical Language Model Adaptation: Review and Perspectives,” *Speech Communication*, Vol. 42, 2004.
[17] **中央通訊社**. http://www.cna.com.tw/.
[18] S. F. Chen, J. Goodman, “An Empirical Study of Smoothing Techniques for Language Modeling,” *Computer Speech and Language*, Vol. 13, 1999.

[19] A. Stolcke, "SRI language Modeling Toolkit," version 1.3.3, http://www.speech.sri.com/projects/srilm/.
[20] X. L. Aubert, "An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition," *Computer Speech and Language*, January 2002.
[21] **公視新聞語料整理與分析(台師大資工所)**. http://speech.csie.ntnu.edu.tw/MATBN_SetDefinition/.
[22] F. Wessel, R. Schluter, K. Macherey, H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, March 2001.
[23] W. Chou (editor),. B.H. Juang (editor). *Pattern Recognition in Speech and Language Processing*. Chapter 2, CRC Press, 2003.
[24] M. Padmanabhan, G. Saon and G. Zweig, "Lattice-Based Unsupervised MLLR for Speaker Adaptation," in *Proc. ISCA ITRW ASR2000*.
[25] R. Rosenfeld, "Two Decades of Statistical Language Modeling: Where Do We Go from Here," *Proc. IEEE*, 88 (8), 2000.
[26] M. Federico, N. Bertoldi, "Broadcast News LM adaptation Using Cotemporary Texts," in *Proc. Eurospeech 2001*.
[27] W. Kim, S. Khudanpur, "Cross-Lingual Latent Semantic Analysis for Language Modeling," in *Proc. ICASSP 2004*.
[28] S. Young et al.. *The HTK Book*. Version 3.2, 2002. http://htk.eng.cam.ac.uk/.
[29] Nagendra Kumar. *Investigation of Silicon Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*. Ph.D dissertation, Johns Hopkins University, 1997.
[30] P. C. Woodland, D. Povey, "Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition," *Computer Speech and Language*, pp.25-47, Vol. 16, 2002.
[31] P. C. Woodland, D. Povey, "Minimum Phone Error and I-Smoothing for Improved Discriminative Training," in *Proc. ICASSP 2002*.

A Noise Estimator with Rapid Adaptation in Variable-Level Noisy Environments

Bing-Fei Wu, Kun-Ching Wang*, Lung-Yi Kuo
Department of Electrical and Control Engineering
National Chiao-Tung University
Hsinchu, Taiwan, R.O.C.
Corresponding author mail: Kunching@cssp.cn.nctu.edu.tw

Abstract. In this paper, a noise estimator with rapid adaptation in a variable-level noisy environment is presented. To make noise estimation adapt quickly to highly non-stationary noise environments, a robust voice activity detector (VAD) is utilized in this paper and it depends on the variation of the spectral energy not on the amount of that. The noise power spectrum in subbands are estimated by averaging past spectral power values using a time and frequency dependent smoothing parameter, which is chosen as a sigmoid function changing with speech-present probability in subbands. The speech-present probability is determined by computing the ratio of the noisy speech power spectrum to its local minimum. Noise measurement, speech enhancement, spectral analysis, signal process.

1 Introduction

An accurate noise estimator used for speech enhancement in adverse environments is one of most essential parts. Inaccurate noise estimator will result in a perceptually annoying residual noise and speech distortion. In general, noise estimation is usually done by explicit detection of speech detection. This can be very difficult in the case of varying background noise. Furthermore, the background noise is assumed to be related stationary between speech pause. To overcome these problems, the noise spectrum needs to be estimated and updated continuously with a reliable speech detector. Among those algorithms, a recursive averaging is a commonly and easily used approach.

Martin [1] proposed a method which is based on minimum statistic (MS). The noise spectrum estimation is obtained by tracking the minimum of the noisy speech power spectrum over a specific window. To improve the computational complexity of estimating noise spectrum, Doblinger [2] proposed an efficient method. However, it fails to differentiate between a rise in noise power and a rise in speech power. Further, Cohen et al. [3] introduced a MCRA approach to estimate noise power spectrum using a smoothing parameter which is defined as the speech-present probability in subbands. The speech-present probability in subbands of a given frame can be determined by the ratio between the noisy speech power spectrum to its local minimum over a period of 0.5-1.5 sec. Finally, the ratio is compared to a specific threshold value to decide updating noise power or not. In recently, Lin et al. [4] proposed a simple and reliable noise estimation technique. To estimate subband noise adaptively and continuously, the smoothing parameter is adjusted by a sigmoid function. However, a variable-level of noise is not considered in this case. We summarize that the drawback of most methods are slow in adapting to suddenly increase level of noise.

In this paper, a noise estimator with rapid adaptation in a variable level noisy environment is presented. It depends only on the variation of the spectral energy but not on the amount of that. Based on the VAD, a noise estimator updates the noise spectrum fast and accurately even in suddenly increases of noise.

This paper is organized as follow. In order to make the estimator is robust against the time-varying level of noise. The utilized VAD in this algorithm is described in Section II. In Section III, the proposed noise estimation is presented in detail. In Section IV, the performance of the proposed method will be evaluated. Finally, we will discuss experimental results in Section V.

2 Voice Activity Detector

Shen et al. [5] first used an entropy-based parameter for speech detection under adverse conditions. Their experimental results revealed that the spectral entropy of a speech signal differs from that of a non-speech signal. The procedure for calculating a spectral entropy parameter is described as follows.
The short-time Fourier Transform (STFT) of a given time frame $s(n,l)$ is given by,

$$x(k,l) = \sum_{n=1}^{M} s(n,l) \cdot \exp(-j2kn\pi/M), \quad 1 \le k \le M, \tag{1}$$

where $x(k,l)$ represents the spectral magnitude of the frequency component k in l^{th} frame index, and M is the total number of frequency components in FFT ($M = 256$ in the proposed system). The spectral energy of each frame $x_{energy}(k,l)$ is described as follows.

$$x_{energy}(k,l) = \left|x(k,l)\right|^2, \quad 1 \le k \le M/2, \tag{2}$$

Then, the probability associated with each spectral energy component $P_r(m,l)$ can be estimated by normalizing:

$$P_r(k,l) = \frac{x_{energy}(k,l)}{\sum_{m=1}^{M/2} x_{energy}(m,l)}, \quad 1 \le k \le M/2, \tag{3}$$

Following normalization, the corresponding spectral entropy H_l for a given frame is defined as follows.

$$H_l = \sum_{k=1}^{M/2} P_r(k,l) \cdot \log[1/P_r(k,l)], \tag{4}$$

The foregoing calculation of the spectral entropy parameter implies that the spectral entropy depends on the variation of the spectral energy not on the amount of that. Similarly, the spectral entropy parameter is robust against changing level of noise. Fig. 1 illustrates that the VAD can locate the speech-present regions, even in high level of background noise.

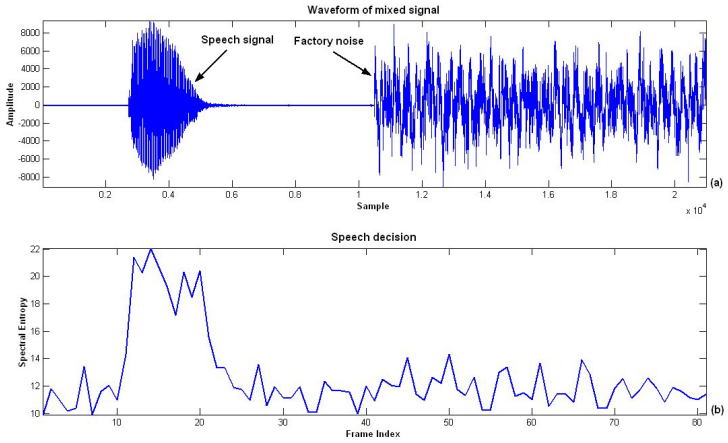


Fig. 1. Speech decision of a mixed signal (including a Factory noise)

3 Proposed Noise Estimator

Fig. 2 presents the flowchart of the proposed noise spectrum estimator. Let assume that noise $d(n)$ and speech $x(n)$ are uncorrelated. The smoothed power spectrum of noisy signal $P(k,l)$ is obtained by

$$P(k,l) = \eta P(k,l-1) + (1-\eta)\left|Y(k,l)\right|^2, \tag{5}$$

where $\left|Y(k,l)\right|^2$ is an estimate of the short-time power spectrum of $y(n)$, given by $y(n) = x(n) + d(n)$. η is a smoothing constant.
Since the noisy speech power spectrum in the speech-absent frames is equal to the noise power spectrum, the estimated noise power spectrum is updated by tracking the speech-absent. To make noise estimation track

speech-absent quickly in highly non-stationary noise environments, a robust voice activity detector (VAD), which depends on the variation of the spectral energy not on the amount of that, is utilized in this section.
First, computing the spectral entropy by Eqs. (1-4) during speech-present frame, a threshold σ is obtained as following:

$$\sigma = c \times E[H_l], \quad 1 \le l \le 5, \tag{6}$$

where c is constant by experiment.
If the spectral entropy value for a given frame is smaller than the threshold σ, then the current frame is regarded as a speech-absent frame. Moreover, the noise estimate is updated according to:

$$\bar{N}(k,l) = \lambda \cdot \bar{N}(k,l) + (1-\lambda) \cdot \left|Y(k,l)\right|^2, \tag{7}$$

where λ is a constant parameter.
Conversely, then the current frame is regarded as a speech-present frame. An algorithm that is suitable for estimating the noise spectrum during speech-present frame is used. First, finding the minimum of the noisy speech spectrum and using the minimum to determine signal-present probability in subbands. The signal-present probability is used to determine a time and frequency dependent smoothing parameter $\alpha(k,l)$, shown as following.

$$\bar{N}(k,l) = \alpha(k,l) \cdot \bar{N}(k,l-1) + (1-\alpha(k,l)) \cdot \left|Y(k,l)\right|^2. \tag{8}$$

To speed up the determination of local minimum of noisy speech spectrum, Doblinger's efficient method is used here [2], which is not constrained by any window length to update noise spectrum estimate.

$$\begin{aligned} &\text{If} \quad P_{\min}(k,l-1) < P(k,l), \\ &\text{then} \quad P_{\min}(k,l) = \gamma \cdot P_{\min}(k,l-1) + \frac{1-\gamma}{1-\beta}(P(k,l) - \beta \cdot P(k,l-1)), \\ &\text{else} \quad P(k,l) = P(k,l), \end{aligned} \tag{9}$$

where $P_{\min}(k,l)$ denote the local minimum of the noisy speech power spectrum and β and γ are constants determined experimentally.
Then, the local minimum is taken to determine speech-present probability $P_{sp}(k,l)$ in subbands, which is similar to that proposed in [3], and the ratio is shown as below:

$$P_{sp}(k,l) = \frac{\left|Y(k,l)\right|^2}{P_{\min}(k,l)}. \tag{10}$$

To improve that a smoothing parameter is produced by comparing the ratio with a fixed threshold value [3], the smoothing parameter is chosen as a sigmoid function changing continuously with speech-present probability in subbands. The smoothing parameter is modified by

$$\alpha(k,l) = \frac{1}{1+e^{-r(P_{sp}(k,l)-T_P(k,l))}}, \tag{11}$$

where $T_P(m,l)$ denotes a adaptive threshold in subbands and is determined during speech-absent frames and shown as following:

$$\begin{aligned} &\bar{N}_{mean}(k,l) = E[\bar{N}(k,i)], \quad i \in \text{all speech-absent frames, up to } l^{th} \text{ frame} \\ &\left|Y(k,l)\right|^2_{mean} = E[\left|Y(k,i)\right|^2], \quad i \in \text{all speech-absent frames, up to } l^{th} \text{ frame}, \\ &T_P(k,l) = \frac{\left|Y(k,l)\right|^2_{mean}}{\bar{N}_{mean}(k,l)} \end{aligned} \tag{12}$$

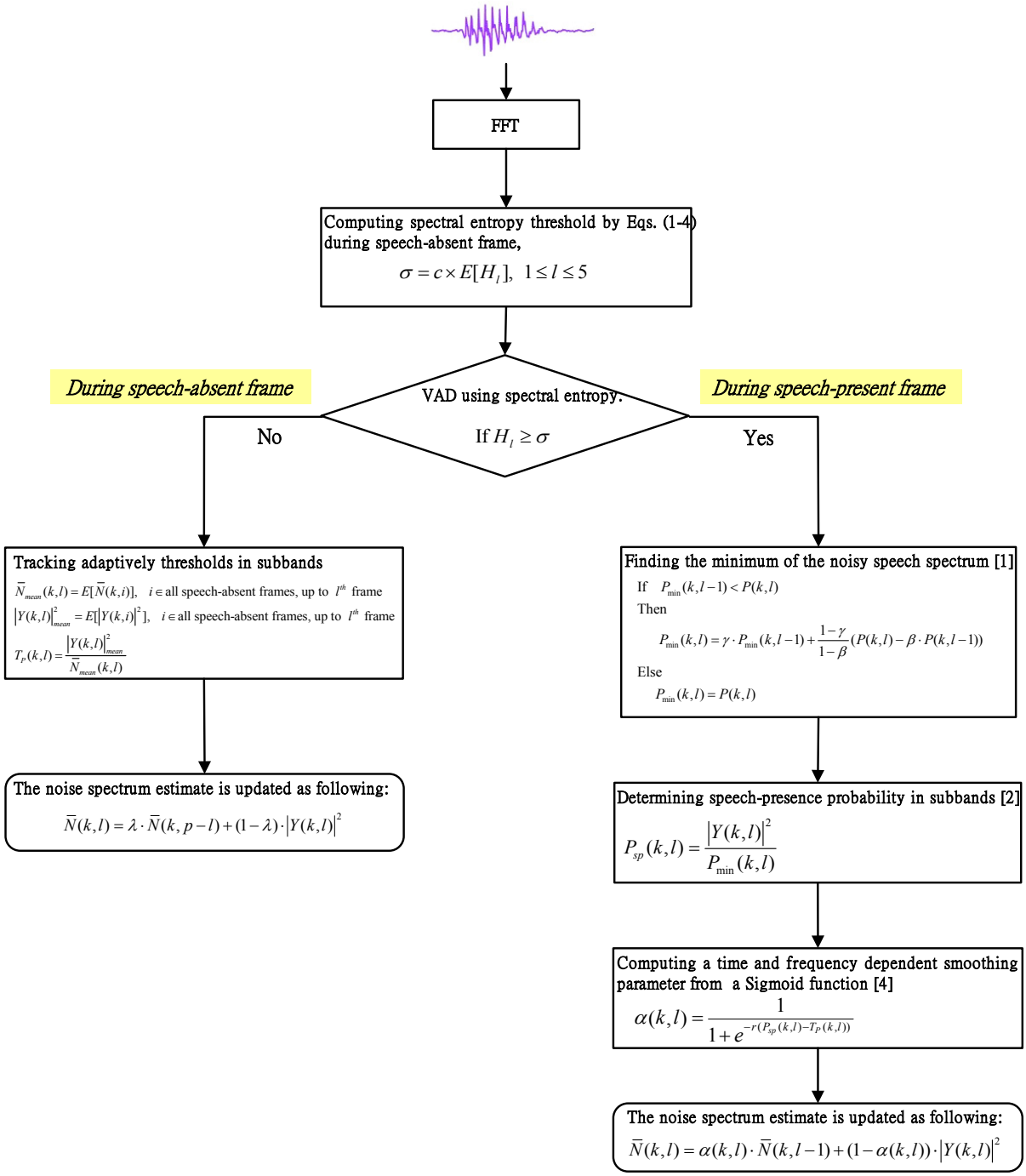


Fig. 2. The flowchart of the proposed noise spectrum estimator

4 Experimental Results

To evaluate the proposed noise estimator, the noisy speech is mixed with a suddenly increase level of Factory noise. Compare with Lin's estimator [4], the results are shown in Fig. 3 and Fig. 4. Fig.3 illustrates the comparison of the proposed noise estimator between Lin's one. Fig.3 (a) shows a phrase "May I Help You ?", which is pronounced from a man in English. It is observed that a background noise suddenly increase in 22000^{th} sample (or 2.75 sec under 8KHz sampling rate). Then, the noise power spectrum is estimated by Lin's noise estimator and a clean speech signal is produced by power spectral subtraction (PSS). The results are displayed in Fig.3 (b). It is observed that the noise is not removed completely later 22000^{th} sample (or 2.75 sec). Fig.3 (c) shows the estimated noise signal from Lin's noise estimator. It is found that the suddenly increase

level of Factory noise is detected in 22000^{th} sample; however, the amplitude of estimated noise is enough large to meet idea noise later 22000^{th} sample. Fig.3 (d) shows the clean signal is generated by the proposed noise estimator and PSS. Compare with Fig.3 (b), the proposed noise can be performed well in suddenly changing level of noise. In Fig.3 (e), the estimated noise signal is produced by the proposed method. Fig.4 shows the spectrograms of a noisy speech signal, an enhanced speech signal of Lin's estimator and that of the proposed estimator, respectively. Similarly, due to the VAD is robust against a changing level of noise, the performance of speech enhancement in the proposed method is better than in Lin's method.

A noisy speech database is generated by applying various segmental SNRs in order to measure the segmental relative estimation error for various types and levels of noise. The segmental relative estimation error (SegErr) is defined by

$$SegErr = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{\omega}[\bar{N}(\omega,m)-N(\omega,m)]^2}{\sum_{\omega}N^2(\omega,m)}. \tag{17}$$

Table I shows the outcomes of the SegErr measured by the proposed estimation method for four noise types with the SNRs range [-5 to 25dB]. The proposed approach is superior to the other methods.

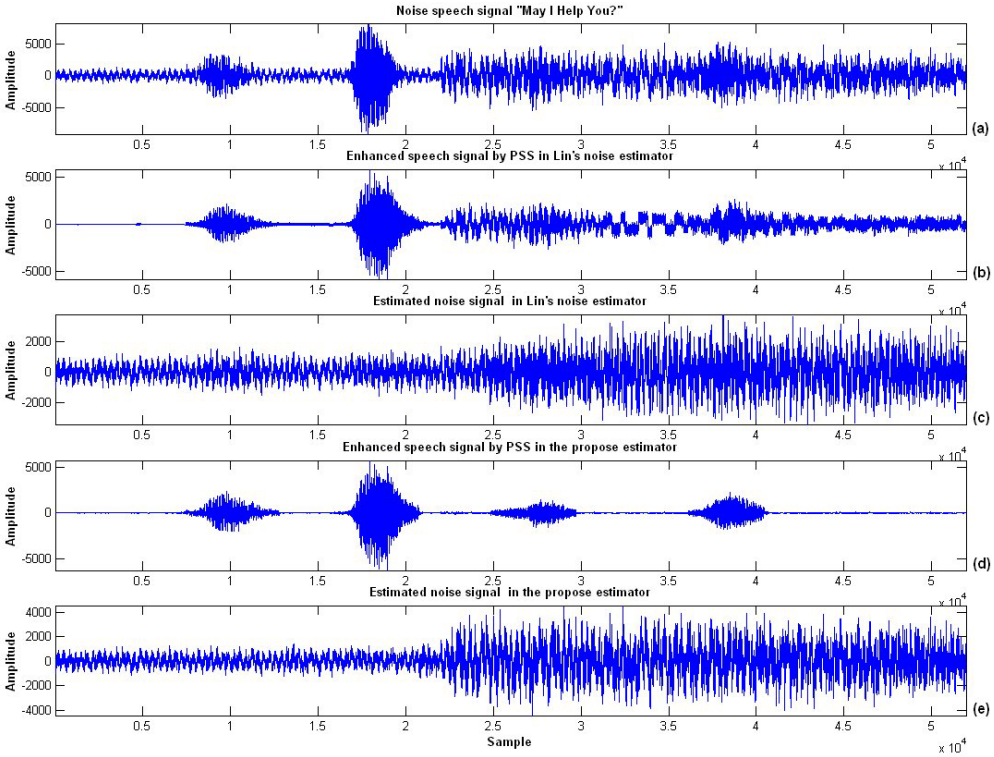
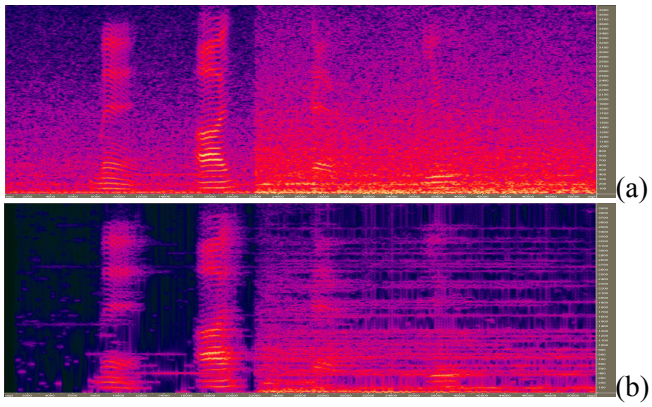


Fig. 3. Waveform of time signal



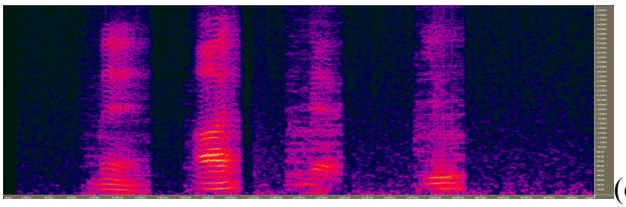
 (c)

Fig. 4. Speech spectrogram (a) noisy speech signal (factory interior noise, suddenly raise in 2.75 sec) (b) speech enhanced with Lin's noise estimator (c) speech enhanced with the proposed noise estimator

Table 1. Example table

Input SegSNR [dB]	Car noise		Factory noise		Babble noise		White noise	
	Proposed	MCRA	Proposed	MCRA	Proposed	MCRA	Proposed	MCRA
-5	0.091	0.132	0.103	0.135	0.115	0.153	0.078	0.095
0	0.085	0.129	0.095	0.132	0.101	0.146	0.068	0.084
5	0.081	0.115	0.086	0.118	0.098	0.127	0.065	0.081
25	0.075	0.108	0.081	0.111	0.095	0.116	0.061	0.079

5 Conclusion

In this paper, a fast noise estimator, which is well suitable for suddenly varying level of noise, is presented. Based on the robust VAD, the speech decision can be determined accurately, and then the proposed algorithm can select the noise spectrum estimation which is suitable for the current frame. Unlike other method [1,3], the adaptation of this time and frequency dependent smoothing parameter does not depend on a specific time window and then updated continuously. Compare with Lin's estimator, the experimental results illustrate that the proposed estimator can remove the noise power spectrum by PSS.

6 Acknowledgement

The authors would like to thank the Promoting Academic Excellence of Universities for financially supporting this research under Contract No.91x104 Ex-91-E-FA06-4-4.

References

[1] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504-512, July 2001.

[2] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," *Proc. EUROSPEECH*, pp. 1513-1516, 1995.

[3] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, Jan. 2002.

[4] L. Lin, W. H. Holmes, and E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement," *Electronics Letters*, vol. 39, no. 9, pp. 754-755, May, 2003.

[5] J.L. Shen, J.W. Hung, and L.S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments,"*Proc. ICSLP-98*, 1998.

A Three-Phase System for Chinese Named Entity Recognition

Conrad Chen

Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu

drchen@csie.nctu.edu.tw

Hsi-Jian Lee

Department of Medical Informatics, Tzu Chi University, Hualien

hjlee@mail.tcu.edu.tw

Abstract. The handling of out-of-vocabulary (OOV) words is one of the key points to a high performance lexical analysis in natural language processing. Among all OOV words, named entities (NE) are the most productive ones. They generally constitute the most meaningful parts of sentences (persons, affairs, time, places, and objects). In this paper, we propose a three-phase "generation, filtering, and recovery" system to address the NER problem. A set of stochastic models is first used to generate all possible NE candidates. Then we treat candidate filtering as an ambiguity resolution problem. To resolve ambiguities, we adopt a maximal-matching-rule-driven lexical analyzer. Last, a pattern matching method is applied to detect and recover abnormalities in the results of the previous two phases.

Pure lexical information is exploited in our system. We get a high recall of 96% with personal names (PER), satisfiable recall of 88%, 89%, and 80% with transliteration names (TRA), location names (LOC), and organization names (ORG), respectively. The overall precision and excluding rate is over 90% and 99%.

1. Introduction

Words are generally the basic unit to process natural languages. However, in Chinese, sentences are composed of string of characters without any delimiters to mark word boundaries. To process Chinese, sentences must be segmented into word sequences first. Most Chinese language processing systems rely on lexicons to recognize words in sentences. Because the number of Chinese words is tremendous, it is impossible to compile all words in a lexicon. Therefore, word segmentation processes often encounters the problem of out-of-vocabulary (OOV) words.

Among all OOV words, named entities are one of the most important sorts. It is impossible to list them exhaustively in a lexicon. They are the most productive type of words. Nearly no simple or unified generation rules for them exist. Besides, they are usually keywords in documents. Named entity recognition (NER) thus becomes a major task to many natural language applications, such as natural language understanding, question answering, and information retrieval.

Many researches have addressed the NE recognition problem in Chinese since 1990. Most of them focused on some specific types as *personal names* [5][13], *location names* [9], *organization names* [10], and *transliteration names* [11]. There are also type-independent approaches of NER. However, most of these approaches need type-dependent data such as role tags. Type-independent approaches can be roughly divided into two major sorts: over-generating & disambiguating [3][12] and over-segmenting & generating [4][8].

Generally speaking, there are two main approaches of the above studies, *rule-based* models and *machine learning* methods. Rule-based approaches could effectively exploit human knowledge and can be tuned conveniently. On the other hand, machine learning approaches, such as *maximum entropy* or *support vector machine*, is more independent from languages and simple to implement. Rule-based approaches is slightly outperform machine learning ones in MUC-7 tests [2].

In our consideration, rule-based approaches are more reasonable than machine learning ones. Boosting performances of rule-based approaches is easier than improving machine learning abilities. Therefore,

rule-based approaches is adopted in this paper, while machine learning methods still could be incorporate in our system under the present framework in future.
A three-phase "*generation, filtering,* and *recovery*" system is proposed to solve NER problem. In the generation phase, stochastic models are responsible for generating all possible candidates of different kinds of named entities in input documents. In the filtering phase, we treat the filtering of false candidates as an ambiguity resolution problem. A maximal-matching-rule-driven lexical analysis is performed to resolve ambiguities caused by false candidates. In the recovery phase, a rule-driven pattern matching method is applied to detect and recover abnormalities in the results of the previous two phases.

2. System Overview

In our system, we try to make use of both the tunability of stochastic models in candidate extraction and the power of lexical analyzers in disambiguation. To implement this idea, we propose a three-phase framework: candidate generation, filtering, and recovery, as shown in Figure 2.1:

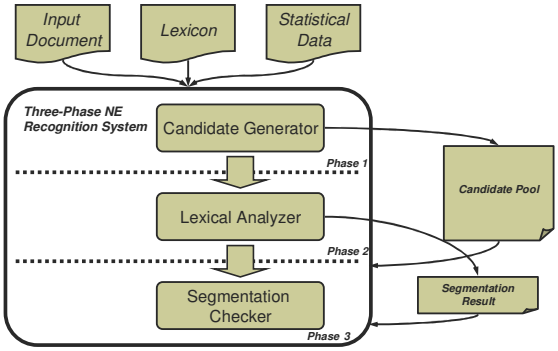


Fig. 2.1. An overview of our system

In the first phase, all possible candidates of various kinds of named entities in the input document are extracted. Notice that this process is inevitably both over-generating and under-generating. Because of the filtering process, the candidate extracting can be tuned to have a higher recall and to sacrifice precision a little for a moment.
Statistical approaches are adopted in the candidate generation phase. The reason is that names are given by people. Therefore, there is no exact answer if a string is a name or not. The only thing can be judged is how likely the string is to be a name. As for computers, to estimate the likelihood of names is basically a fuzzy problem. If a character is more likely to appear in a name, it has a better fuzzy value. The detail of how fuzzy logic and statistic estimation are applied will be discussed later.
The second phase of the system is *false candidate filtering*. How do we verify which candidates are true named entities and which ones are false? False candidates are either a common word or composed of fragments of common words and named entities. The first case has less impact on subsequent applications. The second case usually results ambiguous segmentations. Verification of these candidates could be viewed as an ambiguity resolution problem. If we can judge which segmentation is correct or more proper, we could also verify which candidates are true named entities.
Because of the regularity of lexical choices in modern Chinese, many simple approaches of segmentation ambiguity resolution have good performances. No matter what simple methods it takes, heuristic rules or stochastic estimations, if there are no OOV words, most lexical analysis methods show great precision in ambiguity resolution. That is to say, if we got a high recall in the extraction of NE candidates, most of the segmentation ambiguities caused by false candidates are supposed to be resolved by conventional word segmentation methods. We choose a heuristic approach, which is mainly driven by maximal matching rules, to resolve segmentation ambiguities.
The third phase of the system is *recovery*. The recovery mechanism is used to revive some obviously incorrect results of the first two phases. There are two major target types to be recovered: over-segmentations caused by under-generation and under-segmentations caused by over-generation.

Through the detection of these anomalies, e.g. a succession of single-character words indicating over-segmentations, part of un-extracted named entities could be revived.

3. Candidate Generation

The candidate generator is used to extract all possible named entity candidates in input documents. There are four layers in the candidate generator to handle four sorts of NEs: close-ended NEs, genuine names, whole named entities, and abbreviations.

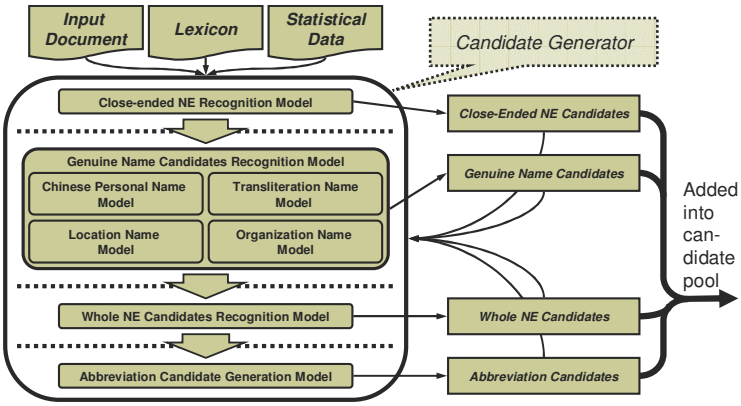


Fig. 3.1. The overview of the candidate generator

Close-ended named entities comprise time and quantity expressions. Since the extraction of close-ended NEs is not the focus of this paper, and previous researches [6] have solved this problem well, a single simplified rule is applied to recognize most of them in our system. The rule is as follows:

$$[\text{“第”}] + (Numerals)^{+} + [Qualifier] + [Unit]$$

This simple rule cannot cover all close-ended NEs, of course. The purpose of this rule is just to prevent unrecognized close-ended NEs affect the performance of the recognition of open-ended ones.
In general, the structure of whole open-ended NEs except for abbreviations can be represented as:

$$[prefixes] + genuine\ name + [suffixes]$$

For example, “台北” is a *genuine name* and “台北市” is a *whole named entity* with *suffix* “市” indicating that “台北” is a city. The handling of prefixes is much similar to that of suffixes, and on the other hand prefixes are much more rarely seen than suffixes. Therefore, for simple implementation, whole NEs with prefixes would not be recognized in our system.
Suffixes generally indicate the type of named entities. There are many types of named entities with different suffixes. Many sorts of them rarely appear in the document. It is not worth to build models for each type of these names. However, suffixes are strong features. It is easier to recognize them, and chances of error recognition are comparatively low. Therefore, a compromised method is adopted that only models for four kinds of genuine names are implemented at present in our system. They are *personal names*, *transliteration names*, *location names*, and *organization names*. These four kinds of genuine name candidates would be used to form various types of NEs with corresponding suffixes. For instance, if a personal name candidate is followed by a publication suffix, they will be recognized as a whole publication name, like:

“余光中”(personal name) + “詩選”(publication suffix) → “余光中詩選”(publication name)

For the same reason above, all NE suffixes are roughly classified into three categories: ones with similar corresponding genuine name types to location suffixes, ones with similar corresponding genuine name types to organization suffixes, and others. The first category covers all location names, racial names, etc. The second one comprises all organization names except for racial names, facility names, publication names, etc. The third one includes feat names, culture names, and so on. Among these three categories, only the first two are addressed by our system. These two categories are called “*location-like NE*” and “*organization-like NE*”. Names belonging to the same category will be

addressed by the same corresponding model. There are two main advantages following this way. First, times spent on designing models and collecting data are saved. Second, confidences brought by suffixes could alleviate the deviation on statistics brought by a compromised approach. The extraction of genuine names and whole named entities will be detailed later.

Open-ended named entities extracted above are used to find possible abbreviations and some rule-recognizable aliases in the abbreviation generation model. Four simple rules are adopted to complete this job:

Rule 1: Take the first characters of genuine name and all suffixes other than typing suffix, and the last character of typing suffix from NE candidates (e.g. "**中**央 **研**究 **院**" → "中研院")

Rule 2: Surnames of personal name candidates (e.g. "**呂**秀蓮" → "呂")

Rule 3: Given names of personal names (e.g. "陳**信安**" → "信安")

Rule 4: *Modifier* + *Surname* or *any character of Given names* (e.g. "陳水扁" → "小陳", "阿水", "阿扁", etc.)

Notice that only abbreviations and aliases with original names appearing in the document could be addressed by our system.

3.1. Statistic Estimation

The recognition of *genuine names* is basically a fuzzy decision problem to computers. There is no exact right or wrong answer for a string to be a name. The only problem is how likely it is. Fuzzy values represent strings' likelihood or properness to be a name. Since Chinese is a character-based language, methods of estimating fuzzy values are generally also character-based. Names are composed of several characters. There are several ways to transform the member characters' fuzzy value to the string's fuzzy value.

Stochastic language models are usually adopted to estimate the likelihood of a candidate to be a named entity. The fundamental principle is that the string with a higher probability or frequency to be a name has a higher fuzzy value or likelihood. There are several ways to estimate the fuzzy value of a string from the statistic data based on characters. These models include Markov models, bi-gram models, unigram models, etc.

Each model has its advantages and disadvantages. Generally speaking, more complex the model is, more precisely it estimate, and more training data it needs. Besides that, the data-sparseness problem is more likely to happen. Since the amounts of features of different types of named entities are varied, each type has its own best-fit model. In this paper, to simplify data collecting and training, unigram models are adopted. Additionally, some supplementary information such as positional feature is exploited to support statistical models.

Generally there are two major ways to estimate fuzzy values of a single character:

Frequency: $freq(typ|c)=counts(typ, c)$

Probability: $prob(typ|c)=counts(typ, c)/counts(c)=freq(typ|c)/counts(c)$

Frequencies stand for differences among naming-characters. They represent popularities of characters to be used in names of some type. If some character is used in more names, it has a higher frequency. If frequencies are used as fuzzy values, a higher recall will be obtained with common names.

Probabilities stand for differences among all characters. They represent possibilities of characters to be used in a name of some type. If some character appears more frequently in names than in common words, it has a higher probability. If probabilities are used as fuzzy values, a higher precision and a higher recall will be obtained with rare names. However, it has a lower recall with common names comparing with using frequencies.

A hybrid statistics is adopted in our system to take advantages of both frequencies and probabilities. With common naming-characters, frequencies are adopted to get a higher recall with common names. With rare naming-characters, probabilities are adopted to complement frequencies' insufficiency with rare names. The resulting model looks like:

$$\mu(typ|c) = Max\{ freq(typ|c), prob(typ|c) \}$$

Data sparseness and reappearances of names make it hard to estimate probabilities. To overcome these difficulties, we propose to use inverse common frequencies to approximate probabilities:

$$icf(c)=1/(freq(common\ word|c)+1)=1/(counts(common\ word, c)+1)$$

Since probabilities are mainly used to estimate the probability of rarely seen events, usually:

$$counts(common\ words,\ c) \approx counts(\sim typ,\ c), where\ counts(typ,\ c) \le 2$$

In this case, *icf(c)* is approximate to *prob(c)*:

$$prob(c) = counts(typ,c)/(counts(typ,c)+counts(\sim typ,c))\ where\ counts(typ,c) \le 2$$
$$\approx 1/(counts(\sim typ,c) + 1) \approx icf(c)$$

Further, we assume that *counts(common word, c)* is in direct proportion to the number of lexicon entries in which the character *c* appears. Under these assumptions, we use inverse lexicon counts to approximate probabilities:

$$ilc(c) = 1/(Num\_of\_Lex\_Entries(c)+1) \approx icf(c) \approx prob(typ|c)$$

Because *ilc(c)* is ranged from 0 to 1, *freq(typ|c)* also needs to be normalized to 0 to 1. The distribution of raw data of *freq(typ|c)* is conformed to Zipf's Law, that:

$$P_n \approx 1/n^a,\ where\ P_n\ is\ the\ frequency\ of\ occurrence\ of\ the\ n^{th}\ ranked\ item\ and\ a\ is\ close\ to\ 1.$$

Values with often seen characters are too high and the distinctions among low frequency characters are not wide enough. Therefore, a logarithm function is taken on the raw data to smooth the distribution curve, and then the result is normalized to 0.1 to 1.

$$freq^*(typ \mid c) = \underset{0.1,1}{Norm}(\log(freq(typ \mid c)))\quad while\ freq(typ|c) \ge 1$$

Notice that the lower bound of *freq*(typ|c)* is set to 0.1, not 0. This is because the meaning of events that appear once is greatly different from the meaning of unseen events.
The final character likelihood model looks like:

$$\mathcal{L}(typ|c) = Max\{\ freq^*(typ|c),\ ilc(c)\ \}$$

Notice that there are two exceptions to this model. With surnames and transliterating characters, likelihoods of unseen events in training data are assigned to zero. This is because generally surnames and transliterating characters are not arbitrarily given. Probabilities of most characters to be surnames or transliterating characters are actually zero. The original model might cause unnecessary over-generation. To prevent this problem, only surnames and transliterating characters appearing in our training data are adopted as possible ones.

3.2. Open-ended Named Entity Extraction

Open-ended named entity extraction models would estimate likelihoods of strings to be some type of named entity from character likelihoods. Unigram models are adopted as the basis of our models. They could be represented as follows:

$$\mathcal{L}'^*(typ|g \bullet s) = \mathcal{L}'(typ|g) \times ConRe(g) \times ConSuf(typ,\ s)$$
where g denotes the genuine name and s denotes the suffix part

If $\mathcal{L}'^*(typ|g \bullet s)$ is over some pre-defined threshold, which is decided by maximizing the *f-measure* of the recall of training data and the excluding rate of lexicon entries, $g \bullet s$ would be recognized as a possible candidate and added into the candidate pool. Each member of the formula is detailed below:

- *ConRe(g)* estimates the confidence could be brought by reoccurrences of the genuine name, which is defined as:

$$ConRe(g) = k^{Reoccurrence(g)}$$
$$k = \begin{cases} 2 & when\ the\ length\ of\ the\ input\ document\ is\ less\ than\ 400\ characters \\ 1+400/LEN(Document) & Elsewhere \end{cases}$$

- *ConSuf(typ|s)* estimates the confidence could be brought by the suffix part. Different types of suffixes could bring different quantities of confidence. One suffix part might comprise many different suffixes. The summation of each member's confidence is computed:

  $$ConSuf(typ,\ s) = Conf(typ_1,\ s_1) + Conf(typ_2,\ s_2) + \ldots + Conf(typ_n,\ s_n) + 1\ where\ s = s_1 s_2 \ldots s_n$$

  Notice that if *s* is empty, i.e., there are no suffix parts, *ConSuf(typ, s)* would be 1.
- The definition of $\mathcal{L}'(typ|g)$ is varied from different types of *genuine names*. As we mentioned before, there are four types of genuine names that would be dealt by our system: *personal names*,

location names, *organization names*, and *transliteration names*. $\angle'(typ|g)$ of different types of names is defined as follows:

- $\angle'(PER|s) = ArgMax\{\angle'(SUR|s1) * \angle'(GIV|s2)\}$ *for every substring s1 and s2, where* $s = s1 \cdot s2$, " $\cdot$ " *denotes the string concatenation, and:*

$$\angle'(SUR|s) = \begin{cases} \ell'(SUR|c1) & \text{when s is constituted of one character} \\ Max\{GAvg(\ell'(SUR|c1),\ell'(SUR|c2)),\ell'(SUR|c1c2)\} & \text{when s is constituted of two characters} \\ 0 & \text{when s is longer than two characters} \end{cases}$$

$$\angle'(GIV|s) = \begin{cases} \ell'(GIV|c1) & \text{when s is constituted of one character} \\ GAvg(\ell'(GIV|c1), \ell'(GIV|c2)) & \text{when s is constituted of two characters} \\ 0 & \text{when s is longer than two characters} \end{cases}$$

GAvg( ) returns geometric means.

- $\angle(TRA|s) = HAvg(\ell'(TRA|c_k))$ *where* $s = c_1 \ldots c_n$ and *HAvg( )* returns harmonic means.
- $\angle'(LOC|s) = \begin{cases} \ell'(LOCL|c1) * \ell'(LOCF|c2) & \text{when s = c1c2} \\ \ell'(LOCL|c1) * \ell'(LOCF|c2) * \ell'(LOCF|c3) & \text{when s=c1c2c3} \\ 0 & \text{elsewhere} \end{cases}$
- $\angle'(ORG|s) = \begin{cases} \ell'(ORGL|c1) * \ell'(ORGF|c2) & \text{when s = c1c2} \\ \ell'(ORGL|c1) * \ell'ORGL|c2) * \ell'(ORGF|c3) & \text{when s=c1c2c3} \\ 0 & \text{elsewhere} \end{cases}$

3.3. Supplementary Mechanism

Besides the above models, there are three supplementary mechanisms designed to relieve over-generation problems of stochastic models:

1. If some candidate is constituted of two multisyllabic words or one multisyllabic word and one often seen monosyllabic word, this candidate would be removed from the candidate pool.
2. If the first or the last character of some three-character-long organization name candidate is a monosyllabic word that often appears adjacent to a name, as "**前**遠東" and "東鼎**與**", this candidate will be removed from the candidate pool.
3. With transliteration names, sometimes a common word might be wrongly attached by a transliteration candidate. In this situation, maximal-matching-rule-driven lexical analyzer cannot filter it out properly.
   A concept called "*team*" based on reoccurrences is introduced to solve the attaching problem. Basically, all substrings of possible transliteration name candidates are also possible candidates. Hence all transliteration name candidates can be grouped into *teams* according to their longest common superstring candidate. For example, a *team* can be represented as:

   $T_{leader=麥可喬丹}$ = {麥可(5), 可喬(5), 喬丹(6), 麥可喬(4), 可喬丹(5), 麥可喬丹(4)}

   Where all appearance times of candidates are marked up, and superstring "麥可喬丹" is called the "*leader*" of the *team*.
   The following algorithm is then applied:
   I. Subtract leader's appearance times from each team member
   II. If the *leader* could be split into candidates with non-zero appearance times after subtraction and multisyllabic common words or frequently used monosyllabic words, discard the *leader* and members whose appearance times being subtracted to zero
   III. Form new teams comprised of remaining candidates with new leaders
   IV. Repeat step I~III, until no candidates could be discarded

4. Lexical Analysis

The lexical analyzer is responsible for verifying candidates generated by the candidate generator. Heuristic rules are adopted to filter out false named entity candidates and resolve ambiguities caused by false candidates. There are six heuristic rules applied in order precedence:

Rule 1: ***Tri-word maximal matching***, which is proposed by Chen & Liu (1992) [1]. The rule follows below three steps:

1. From the segmenting point, look forward for all possible tri-word combinations.
2. Take the first word of the longest sequence of all, segment this word.
3. Move to the next segmenting point.

For example, with the sentence "張大春天天說", "張大春" would be picked instead of "張大" because "張大春 天天 說" is longer than "張大 春天 天".

Rule 2: ***Least number of NEs first***, which would pick the tri-word sequence with the least number of named entities among all sequences of the same length.

Rule 3: ***Most frequently appearing NEs first***, which would pick the tri-word sequence with the most appearing times of component NEs in the input document.

Rule 4: ***Words of even lengths first***, which would choose the sequence with most words of even lengths. There are several exceptions to this rule. First, personal names, transliteration names, and numerical expressions are not concerned in this rule. Second, the often seen monosyllabic words, like "的", "之", "也", etc., are viewed as words of even lengths instead. For example, "**張宇 的 成功**" is regarded as totally having two words of even lengths, one is "成功" and another one is "的", not "張宇". Third, the suffix part of a whole named entity is not considered into the length of it. For example, "**揚昇高爾夫球場**" is viewed as a word of even lengths, not of odd ones.

Rule 5: ***Often seen monosyllabic words first***, which is also proposed by [1], would pick the sequence with the most often seen monosyllabic words.

Rule 6: ***Forward precedence***, which would choose the tri-word sequence with longer forward words. For example, with two ambiguous tri-word sequence "決戰 爭 勝負" and "決 戰爭 勝負", the former would be picked since "決戰" is longer than "決".

In order to measure the performance of our lexical analyzer on ambiguity resolution, the test samples of our system (61 news articles from United Daily News and Central News Agency, which will be further discussed later) are examined. The following measurements are adopted:

- Ambiguous Tri-Word Sequences: *# of all possible tri-word sequences which could not be discriminated by the prior rules*
- Resolved: *# of tri-word sequences which could be filtered by the corresponding rule*
- Errors: *# of correct words which are wrongly filtered*
- Applying Rate: *Resolved / Ambiguous Tri-Word Sequences*
- Accuracy: *1 – Errors / Resolved*

The experimental results are listed in Table 4.1:

Table 4.1. The performance of heuristic rules in ambiguity resolution

	Ambiguous Tri-Word Sequences	Resolved	Errors	Applying Rate	Accuracy
Heuristic Rule 1	81273	78263	265	96.30%	99.66%
Heuristic Rule 2	3010	1935	10	64.29%	99.48%
Heuristic Rule 3	1075	225	5	20.93%	97.78%
Heuristic Rule 4	850	603	20	70.94%	96.68%
Heuristic Rule 5	247	9	1	3.64%	88.89%
Heuristic Rule 6	238	238	49	100.00%	79.41%

5. Recovery

The recovery mechanism would revive obvious incorrect results of segmentations which are not suitable to be solved by priority-style rules. These anomalies mainly comprise two situations: over-segmentations caused by under-generation, and under-segmentations caused by over-generation. The segmentation checker would find suspect segmentation sequences and try to recover them.

To deal with over-segmentations, sequences of three or more seldom used monosyllabic words in a row are suspected. These suspects are checked to see if any fragments of them could constitute NE candidates with $\angle(TYP|s)$ over a predefined suspect threshold of the corresponding type.

For example, since $\angle(TRA|\text{“龐畢度”}) = 0.43 < 0.51$, the candidate threshold of $\angle(TRA|s)$, the string is usually segmented to “龐 畢 度” in the first two phases. This suspect sequence will be detected by the segmentation checker. Because $\angle(TRA|\text{“龐畢度”})$ is larger than the suspect threshold of $\angle(TRA|s)$, which is set to 0.2 in our system, “龐畢度” is added into the candidate list of transliteration names.

With personal names, there is another special case. Let us consider the personal name “陳水扁”. $\angle(PER|\text{“陳水扁”}) = 0.23 < 0.26$, the candidate threshold of $\angle(PER|s)$. However, $\angle(PER|\text{“陳水”})$, which equals 0.54, is larger than the candidate threshold. When this situation happens, the personal name is usually incorrectly segmented into a personal name of two characters and a monosyllabic word, such as “陳水 扁” in this case. To cope with this situation, the following sequence is also viewed as suspects of over-segmentations:

two-character-long personal name candidate + seldom used monosyllabic word

On the other hand, to deal with under-segmentations, segmentation sequences constituted of interlaced appearances of transliteration, location, organization names, and seldom used monosyllabic words, are suspected. These sequences are attempted to be re-segmented into a new sequence containing one more word than the original sequences. For example, if “群中” is incorrectly recognized as a location name, the phrase “台北人群中” would be wrongly segmented into a suspect sequence “台北人 群中”. This sequence would be detected and re-segmented into the right sequence “台北 人群 中”. If the re-segmenting cannot be performed, the original sequence will be kept.

The procedure of segmentation checker is as follows:

1. Check over-segmented sequences
2. Check under-segmented sequences
3. Repeat step 2, until no new suspect sequences appear
4. Check over-segmented sequences again

6. Evaluation

To measure the performance of our system, a corpus which is balanced and well-tagged according to our standard is needed. The most popular standard test corpus, MET-2 data, is biased on some special topics and uses a different tagging standard from ours. Therefore, instead of a standard testing corpus, we obtain 61 articles from United Daily News and Central News Agency as our test bed. These articles are segmented and tagged by our system and corrected manually.

These 61 articles are gathered from five different domains. They are politics, society, business, sports, and entertainment. Because the quantity of politics news and society news is more than others, we obtain three different sub-topics (lawsuit, government, and election) from politics news and two (crime and local) from society news.

Table 6.1 draws the experimental results of our system. Standard measurements are estimated:

$$Recall = (\#\ of\ Ext. - \#\ of\ False)/(\#\ of\ True)$$
$$Precision = 1 - (\#\ of\ False)/(\#\ of\ Ext.)$$

Notice that there are two special columns in the table, *number of words* and *excluding rate*. Because appearing frequencies of NEs are varied in different domains and have a great impact on precision, precision is thus less meaningful. We consider that excluding rate might be a better measurement of over-generation. Excluding rate is counted from:

$$Excluding\ rate = 1 - (\#\ of\ False)/(\#\ of\ Words - \#\ of\ True)$$

It stands for the percentage of non-NEs being correctly filtered by our system.

Table 6.1. Experimental results of our system

Topic	Articles	True	Extracted	False	Words	Recall	Precision	Excluding
Politics 1	7	211	224	34	3460	90.05%	84.82%	98.95%
Politics 2	10	465	444	32	4343	88.60%	92.79%	99.17%
Politics 3	7	158	155	18	2750	86.71%	88.39%	99.31%
Society 1	7	321	317	23	3599	91.59%	92.74%	99.30%
Society 2	10	372	378	39	5423	91.13%	89.68%	99.23%
Business	7	295	289	34	3392	86.44%	88.24%	98.90%
Sports	7	272	226	18	3690	76.47%	92.04%	99.47%
Entertainment	6	196	182	16	2742	84.69%	91.21%	99.37%
Total	**61**	2290	2215	214	29399	87.38%	90.34%	99.21%

Table 6.2 shows the recall of different types of NE. Because we do not focus on automatic classification, one NE might be recognized by many different models, it's hard to judge the precision of each type and only the recalls are listed here.

Table 6.2. Recall of our system with different types of NEs

Topic		PER	TRA	LOC	ORG	ABB	PO	LO	OO	AO	TITLE	MIS
	True	104	2	14	9	6	0	0	2	1	66	5
Politics 1	Detected	102	2	9	5	6	0	0	2	0	61	2
	Recall	98.08%	100.00%	64.29%	55.56%	100.00%	--	--	100.00%	0.00%	92.42%	40.00%
	True	261	0	18	11	9	5	0	148	10	4	0
Politics 2	Detected	250	0	17	5	6	5	0	128	2	2	0
	Recall	95.79%	0.00%	94.44%	45.45%	66.67%	100.00%	--	86.49%	20.00%	50.00%	--
	True	43	0	24	0	2	0	0	42	46	2	0
Politics 3	Detected	43	0	24	0	2	0	0	28	43	0	0
	Recall	100.00%	0.00%	100.00%	0.00%	100.00%	--	--	66.67%	93.48%	0.00%	--
	True	185	1	115	0	30	2	1	2	0	1	0
Society 1	Detected	180	1	98	0	29	1	1	2	0	0	0
	Recall	97.30%	100.00%	85.22%	0.00%	96.67%	50.00%	100.00%	100.00%	--	0.00%	--
	True	135	3	137	13	6	2	8	60	5	0	2
Society 2	Detected	133	2	128	11	6	2	7	43	3	0	2
	Recall	98.52%	66.67%	93.43%	84.62%	100.00%	100.00%	87.50%	71.67%	60.00%	--	100.00%
	True	72	2	65	131	0	0	9	4	5	1	0
Business	Detected	68	2	56	111	0	0	3	4	5	0	0
	Recall	94.44%	100.00%	86.15%	84.73%	0.00%	--	33.33%	100.00%	100.00%	0.00%	--
	True	56	110	23	9	1	4	6	58	3	1	5
Sports	Detected	50	99	19	8	1	0	6	20	2	1	4
	Recall	89.29%	90.00%	82.61%	88.89%	100.00%	0.00%	100.00%	34.48%	66.67%	100.00%	80.00%
	True	126	12	18	4	4	12	1	0	5	12	4
Entertainment	Detected	118	9	16	2	4	6	1	0	5	7	1
	Recall	93.65%	75.00%	88.89%	50.00%	100.00%	50.00%	100.00%	--	100.00%	58.33%	25.00%
	True	982	130	414	177	58	25	25	316	75	87	16
Total	Detected	944	115	367	142	54	14	18	227	60	71	9
	Recall	96.13%	88.46%	88.65%	80.23%	93.10%	56.00%	72.00%	71.84%	80.00%	81.61%	56.25%

Notice that the first five columns (PER, TRA, LOC, ORG, ABB) only include the focused types of our system. Column PER comprise only formal Chinese personal names and personal names with appellations. Other personal names, such as Japanese name "酒井光次郎" and pseudonym "老子", are counted in the column PO instead. Monosyllabic place names without suffixes, like "粵" and "台", are recognized by lexicon matching and counted in the column LO. Government and team names are also recognized by lexicon. They are viewed as OO. All other location names and organization names are included in the column LOC and ORG respectively. Column ABB contains only abbreviations with original reference in the input document, other abbreviations are considered as AO.

7. Conclusions and Future Works

Overall speaking, pure lexical information is employed to recognize named entities in our system. Only statistical features and internal structures of NE are utilized. Our statistical model and heuristic rules are simplified for easy implementation. However, our system gets a satisfied performance, and there are still many rooms for improvement.

First, statistical models could be refined. More training data could be collected. More elaborate candidate generating models could be adopted, such as bi-gram models. More internal features could be exploited, such as positional information of characters. Contextual information, such as word probability of being adjacent to some type of NEs, could be also added into our model.
Second, heuristic rules could be more completed or substituted by other mechanisms. Shortcomings of heuristic rules form an upper-bound barrier of performances. More rules could be introduced to cover the inadequacies of original ones. Other mechanism like statistical approaches could be used to replace rule-driven methods.
Third, more candidate generating models could be added. Many types of NEs have not been addressed in our system. We could find that these NEs occupy a great proportion of true negative errors. If these NEs could be recognized, the recall of our system is supposed to be boosted.
Fourth, more knowledge could be gathered and utilized. The suffix and appellation information used in our system is handcrafted at present. Bootstrapping or machine learning algorithm might help us automatically retrieve these kinds of information from the Internet or corpus. Part-of-speech tagging, syntactic checking and even semantic analysis might also be added into our future system.

References

[1] Chen, Keh-Jiann and S. H. Liu, 1992, "Word Identification for Mandarin Chinese Sentences," Proceedings of COLING-92, Vol. 1, pp. 101-107

[2] Chinchor, Nancy, 1998, "MUC-7 Test Score Reports for all Participants and all Tasks" in Proceedings of the MUC-7.

[3] Chua, Tat-Seng and J. Liu, 2002, "Learning Pattern Rules for Chinese Named Entity Extraction," Proceedings of AAAI/IAAI 2002, pp. 411-418

[4] Goh, Chooi Ling, M. Asahara, Y. Matsumoto, 2003, "Chinese Unknown Word Identification Using Character-based Tagging and Chunking," ACL-2003 Interractive Posters/Demo, pp. 197-200

[5] Ji, Heng and Z. S. Luo, 2001, "Inverse Name Frequency Model and Rule Based Chinese Name Identification," (In Chinese) Natural Language Understanding and Machine Translation, Tsinghua University Press, pp. 123-128.

[6] Mo, Ruo Ping, Y. J. Yang, K. J. Chen, and C. R. Huang, 1996, "Determinative- Measure Compounds in Mandarin Chinese Formation Rules and Parser Implementation," In C. R. Huang, K. J. Chen and B. K. Tsou (Eds.), Readings in Chinese natural language processing, pp. 123-146, Journal of Chinese Monograph Series Number 9.

[7] Sekine, Satoshi, K. Sudo, and C. Nobata, 2002, "Extended named entity hierarchy," Proceedings of the LREC 2002 Conference, pp. 1818-1824.

[8] Sun, Jian, J. F. Gao, L. Zhang, M. Zhou, and C. N. Huang, 2002, "Chinese Named Entity Identification Using Class-based Language Model," Proceedings of the 19th International Conference on Computational Linguistics, Taipei, pp. 967-973

[9] Tan. Hong-Ye, 1999, "Chinese Place Automatic Recognition Research," Proceedings of Computational Language, C. N. Huang & Z.D. Dong, ed., Tsinghua Univ. Press, Beijing, China.

[10] Wu, Xue-Jun, J. B. Zhu, H.Z. Wang, and N. Ye, 2003, "The Application of the Method of Co-Training in Identification of Chinese Organization Names," The 2003 National Joint Symposium on Computational Linguistics (JSCL-2003)

[11] Xiao, Jing, J. M. Liu, and T. S. Chua, 2002, "Extracting pronunciation-translated names from Chinese texts using bootstrapping approach", Nineteenth International Conference on Computational Linguistics (COLING2002), Taipei, Taiwan, Aug 2002.

[12] Yu, Shi-Hong, S. H. Bai, and P. Wu, 1998, "Description of the Kent Ridge Digital Labs System Used for MUC-7," Proceedings of the Seventh Message Understanding Conference (MUC-7).

[13] Zheng, Chen, W. Y. Liu, and F. Zhang, 2002, "A New Statistical Approach to Personal Name Extraction," ICML 2002, pp. 67-74.

主題導向之非結構化文本資訊擷取技術

劉吉軒、翁嘉緯

國立政治大學 資訊科學系

E-mail: jsliu@cs.nccu.edu.tw

Abstract. 資訊擷取(information extraction)是從自然語言文本中辨識出特定主題或事件的描述，進而萃取出相關主題或事件元素的對應資訊，如人、事、時、地、物等。因此，資訊擷取技術能依照需要的主題與事件，自動的解讀自然語言文件，將文件中的原始文字資料轉換成結構化的核心資訊。在本論文，我們提出以型態辨識的方法來處理主題導向的非結構化文本資訊擷取的問題。我們以『總政府人事任免公報』為測試對象，其精確率為98%、回收率為97%，充分印證了本資訊擷取方法處理主題導向之資訊擷取問題的可行性。

1 導論

隨著電腦技術的進步與應用的普及，網際網路與全球資訊網上大量資訊被產生、流動、與保存，這些資訊通常可以說是氾濫而雜亂無章，對我們獲取、吸收、與利用資訊的能力構成嚴肅的挑戰，其結果甚至影響我們工作與生活的產出與品質。目前在網路上獲取資訊最普遍的工具就是利用資訊檢索(information retrieval)或搜尋引擎(search engine)，以使用者提供的關鍵字或索引詞，從大量網頁、文本集合中，調出含有使用者指定的關鍵字或索引詞的子集合。然而不論資訊檢索或搜尋引擎的精準度及回召率(precision and recall)如何，使用者仍然需要自己去進行閱覽與過濾，對於資訊的解讀、吸收、與利用等加值工作仍然需要以人工的方式去完成，常常耗費大量人力與時間或因無暇兼顧而無法得到真正的資訊價值。

資訊擷取(information extraction)是從自然語言文本中辨識出特定主題或事件的描述，進而萃取出相關主題或事件元素的對應資訊，如人、事、時、地、物等[1]。因此，資訊擷取技術能依照需要的主題與事件，自動的解讀自然語言文本，將文本中的原始資料轉換成核心資訊，可供進一步的機器使用及加值處理。資訊擷取技術的研究大約從一九九零年前後開始以英文文字為主要對象，選定數個主題，如恐怖份子攻擊事件、國際企業聯盟合併等，進行先導性的技術發展。一般研究結果認為[2]，資訊擷取以事件描述型式比對(event template matching)為主，再輔以領域語言知識及推理，如字詞、句型、前後指涉分析等，可以達到百分之七十左右的正確率，其問題的困難度不如自然語言處理大，卻具有相當高的實用價值，如情報蒐集與分析等。

一般而言資訊擷取系統的建立大致上可分為兩種方式：知識工程法(knowledge engineering approach) 及自動訓練法(automatically trainable approach)[3]。知識工程法主要是透過人工的方式給定擷取規則，而給定擷取規則的人必須對處理的領域及擷取規則建立的方式有一定程度的瞭解，其處理的範圍與正確性通常取決於擷取規則的充分與適當程度。因此，知識工程法對於人工介入與人力需求的程度較高，其素質及對領域的瞭解，也會對系統表現有非常大的影響。相對而言，自動訓練法並不需要人工介入的方式來建立擷取規則，通常只要將訓練語料做適當的標註，再透過訓練演算法就可以建立擷取規則，但其擷取規則可能產生不小的錯誤率。以發展成本及可攜性而言，自動訓練法似乎是發展資訊擷取系統的一個比較好的選擇，但是當訓練語料不易取得，或是對於資訊擷取系統的正確率有較高的要求時，知識工程法可能較佔優勢。

以資訊擷取的文本對象而言，又可區分為半結構化文本與純文字文本。半結構化文本(如網頁)最主要的特點便是內容含有標籤(Tag)，提供了辨識的依據，同時資訊呈現的方式較有規則性，通常只要掌握住這些標籤與規則，就能進一步的擷取出資訊。相關的研究包括WIEN[4], SoftMealy[5], STALKER[6], IEPAD[7]等。純文字文本則不包含任何的標籤與結構，其內容完全是一長串的文字符號，在處理上無法依賴或藉助於結構特徵，而必須完全針對文字符號的組合去做資訊擷取。相關的研究包括AutoSlog[8], FASTUS[9]等。

另外，文本語言也是資訊擷取技術的重要區別因素。以中文與英文來說，兩者之間最大的不同在於中文詞與詞之間並沒有明顯的界限(如英文字之間的空白)加以區隔，因此許多中文處理的第一個步驟，通常

就是利用詞典，將一個字串中的文字，比對詞典內的詞來當做斷詞的依據。不過因為字組成詞的變化程度相當大，一個句子難免會有許多種斷詞的方式，所以斷詞的錯誤率通常很高。另一個問題則是未知詞的問題，例如專有名詞，包括人名、地名、或組織名，不在詞典中的可能性非常大，而在一般句子中出現未知詞的頻率也很高，這對斷詞的正確率造成嚴重的影響。這些錯誤通常會對自然語言處理中的詞性標註、語法剖析等工作造成相當程度的困難，而使得一般以英文文本為處理對象的資訊擷取技術無法直接適用於中文文本。

隨著科技的進步與資訊數位化的趨勢，數位化之文字資料已呈指數般的大幅成長，而資訊擷取也就成為了近年來相當熱門研究領域。美國政府透過一系列的研討會(Message Understanding Conference, Text Retrieval Conference)為主軸[10]，持續推動資訊擷取研究，並規劃研究主題與實驗。這幾年，國內也有許多學者與研究人員投入資訊擷取領域，分別以網頁或其他半結構性文本為主[11][12]，及以中文純文字文本為主[13][14][15]。

我們的研究目標是發展高度實用的主題導向資訊擷取系統，以中文非結構性文本為擷取對象，並以結果的高度正確為主要考量。在本論文中，我們提出不做斷詞、不做詞性分析，而利用型態辨識的方法，搭配有限狀態自動機的運作機制，來處理中文非結構性文本資訊擷取的問題。我們的中文非結構性文本資訊擷取技術包括擷取模板的建立、多層次擷取型態的給定、及有限狀態自動機的執行工具等模組。擷取模板定義特定主題的構成語意元素，是描述主題的核心資訊，如人事異動的主題必須包括單位組織名稱、相關人員的姓名、及職位、時間等。在多層次擷取型態方面，我們處理中文語句中數個子句共用語意元素的問題。而在系統核心執行工具－有限狀態自動機的執行方面，我們利用演算法，將擷取型態轉換成相對應之有限狀態自動機，自動的進行中文語句的辨識。我們發展出一個主題導向的資訊擷取系統，並以『總政府人事任免公報』[16]為測試對象，蒐集了1981年(民國70年1月)到2003年(民國92年6月)的『總統府人事任免公報』電子檔，共1788期，約10萬個擷取目標，每一個擷取目標為一筆完整的人事異動資料。經過採樣及推測的實驗方法評估，實驗數據顯示98%的精確度與97%的回收率。

本論文以下分為四個章節，第二章描述我們研究的主題與文本特性，第三章提出型態辨識為主的中文資訊擷取方法與機制，第四章為實驗評估與結果討論，第五章為結論及未來的研究方向。

2 人事異動主題與文本

在主題導向的中文非結構性文本資訊擷取研究上，我們認為政府文本是一個相當好的試驗對象。我們從資料面、資訊與知識面、與價值面分別進行分析。首先，政府的官方文件具有下列特性：(1)有長期持續存在的主題，能提供大量的文件做為試驗資料與資料庫建置來源；(2)文件結構與主題描述方式較少變化，可以期望較高的資訊擷取正確率；(3)文件主題間存在關聯性，如人事、組織、考核、獎懲等，可以進一步的提供資訊查詢、資料探勘、與知識擷取研究；(4)政府文件可公開取得，沒有版權問題。而在價值面上，政府文件的精華資訊擷取具有極佳的先導示範作用。

我們選擇政府人事任免公報做為研究初期的試驗對象，此公報從民國三十七年起，週期性出刊(約每週一次)，記載政府各部門人事異動情形，並由總統令公告。這個特定文件只有單一主題，並且是由有限語意元素與句型組成。因此，非常適合做為我們研究起始的、典型的資料領域。政府人事任免公報的範例如下：

總統令 中華民國九十一年五月二十四日
任命鄒擅銘為國史館臺灣文獻館簡任第十職等組長。
任命楊合進為法務部簡任第十一職等權理簡任第十二職等司長。
……
任命鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為簡任第十一職等副處長。
總　　統　陳水扁
行政院院長　游錫堃

總統令 中華民國八十五年六月十三日
經濟部政務次長楊世緘，交通部政務次長蔡兆陽另有任用；財政部政次長王政一，僑務委員會副委員長王能章、張植珊辭職已准；均應予免職。
總　　統 李登輝
行政院院長 連戰

圖一：政府人事任免公報的範例

由初步的分析可知，政府人事公報中的記載分別為有關任命或免職的命令。任命的命令是指派某人到某個職位、機關、階級等，免職的命令是免除某人現有在某機關、階級上的職位。最簡單的任命句型為：任命李大衛為行政院簡任第十三職等參事。最簡單的免職句型為：行政院簡任第十三職等參事李大衛呈請辭職；應予令免。較複雜的任命或免職句型描述多人、共用部分資訊、或個人同時有兩個以上的資訊。例如：任命鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為簡任第十一職等副處長。經濟部政務次長楊世緘，交通部政務次長蔡兆陽另有任用；財政部政次長王政一，僑務委員會副委員長王能章、張植珊辭職已准；均應予免職。我們進一步的分析整理發現，政府人事公報中大約有 20 幾個語意元素，30 幾種句型，部分語意元素與句型顯示於圖二。

部分語意元素代碼：

A – 任命 (appoint)	R – 階級名稱 (rank)
N – 人名 (person name)	T – 職位名稱 (title)
B – 為 (as)	Q – 免職原因 (reason of dismissal)
O – 機關名稱 (organization name)	D – 免職 (dismissal)

句型範例： ANBORT ORTNQD

圖二：政府人事主題之部分語意元素與句型

在所有語意元素中，只有部分語意元素的字詞可以被蒐集而可直接辨識，其他的語意元素則不可能掌握，如人名、機關名等。另外，相關子句間的語意元素可能被省略，需要進行分析，找出對應而補齊。我們的基本想法是以能掌握的關鍵詞部分辨識切割句子，再與已知句型進行型態比對，而依據最接近的句型，詮釋之前未辨識出的語意元素。

3 型態比對模型與機制

資訊擷取為針對特定主題或事件從文本中找到對應於相關觀念或元素的實際資料，如人、事、時、地、物等資訊。從原文到解讀出的核心資訊，需要經過字詞的辨識、語句的分析、描述方式的比對、語意關係的推理、資訊的抽取與對應等步驟。國外相關研究歸納了一套基本流程[2][3]：tokenization (word segmentation) → morphological and lexical processing (part of speech tagging, word sense tagging) → syntactic analysis (full parsing) → domain analysis (co-reference, merging partial results) 。原文先被分解成句子與字詞，並從辭典中找出各字詞的詞類與其他資訊，接著進行各種名字的辨識，包括人名、組織名、日期、幣別等。基於這兩個步驟辨識的結果，各句子被部分的分析，根據句子結構的資訊，確認各重要字詞的意義。這些辨識與分析的結果和已知的主題或事件的可能描述方式進行比對，找出最接近的模型。接著進行同指涉詞的分析，找出前後彼此對應的名詞，再進行必要的推理，確定各字詞的意義與關係。最後，總結所有的資訊，依照選定的模型，確認主題或事件中各觀念或元素適當的對應字詞。

這一套基本流程與步驟提供了研究目標的指引與可行性依據，不過由於人類的語言具有模糊、變動、文化、地域等等的特殊性，從初期的tokenization到中期的syntactic analysis及到最後階段的domain analysis都可能因為這些語言上與主題領域上的特殊性而產生一定程度的錯誤。倘若在處理的前期就產生錯誤的話，就會影響到其後處理的步驟，而這些錯誤的累積也必定會影響到最後的結果。以中文來說，由於中文詞跟詞之間並沒有明顯的界限，斷詞錯誤的情形仍很可能出現，加上中文的結構較為鬆散、多縮寫型式且詞性不易判別，所以相對於英文，其錯誤情形的累積就會更嚴重。

由於我們的目標是高度正確的、具有實用價值的資訊擷取技術，我們決定不採用斷詞及詞性標註的方式來處理原始文句，而藉由分析文本中中文語句的結構、順序及組合方式，以型態辨識的方法確認出中文語句之結構關係，再利用關鍵字詞及其特殊的型態，來推論或擷取出相關的資訊。

3.1 以型態辨識擷取中文資訊之概念形成

一般而言，語言的敘述可以視為多種特定語意元素的組合，而其組合的方式通常具有某種規則或常見的型態，所以只要能掌握到某些特定語意元素常對應之字詞的知識，再加上這些語意元素的組合方式，就可以推論出其它相關語意元素的位置，進而擷取出我們想要的資訊。舉例來說：『總統某某某先生』這個敘述可以區分成三個連續語意元素。假設我們想擷取出”某某某”這個字詞，我們只要將第一個語意元素相對應之字詞”總統”與第三個語意元素相對應之字詞”先生”當作關鍵字，利用前後關鍵字之辨識，再包夾切割出目標字詞的方式，即可擷取出第二個語意元素相對應之字詞”某某某”。

在主題導向的中文資訊擷取中，我們針對含有與主題相關資訊的敘述，將特定語意元素的組合方式稱之為擷取型態(extraction pattern)。以上述的例子而言，我們可以標示一個以三個連續之語意元素代號『TNA』組成之擷取型態，其中T(Title)表示”職稱”，N (Name)表示”姓名”，A (Appellation)表示”稱謂”。透過『TNA』的擷取型態，我們可以從許多文本中擷取出出任過特定職位的人員姓名，例如：(總統、蔣經國)，(總統、李登輝)，(總統、陳水扁)，(首相、邱吉爾)，(首相、柴契爾)等。

3.2 擷取型態中語意元素之辨識與擷取屬性

擷取型態中的語意元素組合，在字詞辨識與擷取的過程中，個別語意元素具有不同的屬性。其中最主要的差異在於直接辨識的難易，各語意元素在不同文本中出現的敘述，其對應之字詞可能變化性較大、不容易掌握，如人名、機關名；也可能變化性較小、容易掌握，如稱謂(先生、小姐等)。通常這些字詞變化性較大的語意元素，也可能會是我們的擷取目標，我們採用的基本方法為：蒐集變化性較小的語意元素所對應之字詞做為關鍵字，利用關鍵字之辨識，以前後包夾目標語意元素的方式，切割擷取出變化性較大的語意元素相對應之字詞。

我們定義了三個語意元素之辨識與擷取屬性，以因應擷取過程，針對擷取型態中個別語意元素的不同處理動作。這三個辨識與擷取屬性為: (1) **EOE**(Extraction Only Element): 屬性為EOE的語意元素表示其相對應之字詞變化多，不能掌握，如人名、機關名等。這類語意元素必須依賴前後關鍵字的辨識，進而切割擷取出相對應之字詞。(2)**ROE**(Recognition Only Element): 屬性為ROE的語意元素表示其相對應之字詞變化少，可以被蒐集而可直接辨識，但此語意元素在主題資訊中不具價值，如前述例子中的”稱謂(A)”及政府人事異動主題中的”為(B)”。其相對應之字詞是用來當做包夾切割目標語意元素字詞的關鍵字，並不需要被擷取。(3)**RTE**(Recognition exTraction Element): 屬性為RTE的語意元素表其相對應之字詞變化少，可以被蒐集而直接辨識，同時，此語意元素也是主題資訊中的重要成分。所以此語意元素相對應之字詞，既是用來當做切割前後其他目標字詞的關鍵字，也是擷取的對象本身。

基本上，由EOE、ROE及RTE這三個屬性的語意元素所組成的擷取型態，對應於較為緊密、精簡或簡短的主題描述方式。例如，擷取型態『TNA』中，語意元素『T』之屬性為RTE，對應之關鍵字為『總統、首相』等，語意元素『N』之屬性為EOE，語意元素『A』之屬性為ROE，對應之關鍵字為『先生、女士』等。在文本中出現『總統陳水扁先生』的語句，經辨識擷取後的結果將為(總統、陳水扁) 。

3.3 模板建立

資訊擷取技術針對文本中特定主題的資訊，進行抽取與對應，而這些抽取出來的文字，必須能完整的對應於主題所需的各部分資訊。倘若只是一味的利用擷取型態來辨識中文語句，而忽略了主題資訊的完整性，那麼擷取出來的結果就會零散而不建全。因此，擷取模板(extraction template)的建立就有其必要性。擷取模板為一組特定語意元素的欄位集合，必須根據不同的擷取主題加以定義。以政府人事異動的主題而言，我們定義的擷取模板包括姓名、組織單位、職位、職等、異動種類、異動原因、異動時間等語意元素，一個擷取目標為文本中有關某一個人員的異動情形，由一個完整的擷取模板來描述。在政府人事公報文本中，我們發現不同擷取目標共用部分語意元素的情形，這是中文前後文句中常見的省略、沿用、及總結等的慣例用法。因此，對於部分擷取目標而言，直接的資訊擷取只能得到不完整的主題資訊。藉由資訊模板的定義與規範，相鄰擷取目標的部分資訊將能互相參照補齊，而建立完整的主題資訊。為了完成這樣一個動作，在建立擷取模板時，就必需針對每一語意元素的主題資訊構成屬性加以定義，主要分成三種：(1)required, 表示此語意元素是必須存在的；(2)optional, 表示此語意元素可能存在，但也可以不存在；(3)context-dependent**,** 表示此語意元素可出現在前後相關的中文語句中。

茲以『任命鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為專員。』為例說明，我們可以觀察到以下特點：(1)人員為人事異動主題中擷取目標的主體，一個擷取目標至少必須具有人員姓名與職稱的存在；(2)語句『黃崇烈為專員』含有一個擷取目標，但除了人員姓名與職位兩個語意元素外，異動種類、組織單位、職等等語意元素並沒有出現在前述的語句中，而是必須由前一個句子中的部分語意元素沿用；(3)職等的語意元素並不是擷取目標的必要元素，有些擷取目標有，但也有擷取目標不具備。因此，在人事異動的主題中，人員姓名及職稱的主題資訊構成屬性為”required”，表示這兩個語意元素必須在擷取目標所在之文句中擷取。組織單位、異動種類與異動原因的主題資訊構成屬性為”context-dependent”，表示這三個語意元素可出現在前後相關的語句中。職等的主題資訊構成屬性為”optional”，表示這個語意元素可能存在，也可能不存在。

3.4 多層次擷取型態

以資訊擷取之觀點而言，主題相關敘述語句中之各語意元素，除了具有連續性的關係之外，更有著前後相關子句『語意元素共用』的關係。而這共用性的關係包羅甚廣，可能為共用的時間、共用的單位、共用的幣值等等。舉例來說，中文語句『總統候選人民進黨陳水扁先生、國民黨連戰先生、親民黨宋楚瑜先生』中，『民進黨陳水扁先生』、『國民黨連戰先生』、『親民黨宋楚瑜先生』均共用了外層字詞『總統候選人』。為了處理『外層語意元素共用』的問題而完整的擷取所有相關資訊，我們設計了多層次的擷取型態，其辨識過程將是從外層共用的部份先行處理，然後再依序從其內層進行下一步的擷取動作。繼續以前述語句『總統候選人民進黨陳水扁先生、國民黨連戰先生、親民黨宋楚瑜先生。』為例，我們可以建立多層次擷取型態為『C{PNA}』來處理此種類型之語句，其中語意元素C為參與活動人員身分，屬性為RTE，並建立其相對應之關鍵字『總統候選人』；”{PNA}”表示可重複出現之內層擷取型態，語意元素P為組織名稱，屬性為RTE，並建立其相對應之關鍵字『民進黨』、『國民黨』及『親民黨』；語意元素 N為人名，屬性為EOE；語意元素A為稱謂，屬性為ROE，並建立其相對應之關鍵字『先生』及『女士』、『小姐』)。因此，經由多層次擷取型態 『C{PNA}』辯識後的擷取結果將為(總統候選人、民進黨、陳水扁)、(總統候選人、國民黨、連戰)及(總統候選人、親民黨、宋楚瑜) 。

3.5 有限狀態自動機

我們以有限狀態自動(finite state automata)做為擷取型態的運作機制。有限狀態自動機的主體在於各狀態間之transition function。我們將擷取型態中各語意元素所代表的辨識資訊轉換成各狀態之transition function。轉換的演算法如圖三所示，其中State的初始值為1 (初始狀態)，『Pattern.length()』可以計算擷取型態中語意元素的總個數，『Pattern.char(I)』表示擷取型態中第I個語意元素，『- Pattern.char(I) 』表示除了『擷取型態中第I個語意元素』的其他語意元素，『＊』表示任何的語意元素，Final_State表最終狀態(接受狀態)，而0-state表拒絕狀態(sink state)。

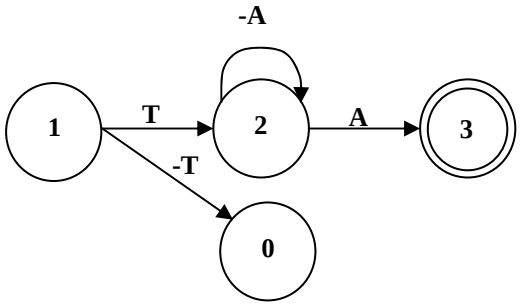
```
0   Begin
1     State = 1
2     For I = 1 to Pattern.length() do
3     If(attribute of Pattern.char(I) is not EOE ) then
4       Generate state transition from “State to (State+1)”,
5       Add transition input as “Pattern.char(I)”
6       If (attribute of Pattern.char(I-1) is not EOE then
7          Generate state transition from “State to 0-state”,
8          Add transition input as “ - Pattern.char(I) ”
9
10      State = State + 1
11    Else
12      Generate state transition from “State to State”,
13      If ( “ -Pattern.char(I+1) ” does not exist)
14         Add transition input as “＊”
15      Else
16         Add transition input as “-Pattern.char(I+1)”
17  End
```

圖三：擷取型態之有限狀態自動機轉換演算法法)

此演算法的基本目標為根據擷取型態及其中各語意元素之辨識與擷取屬性，自動建立一個對應此擷取型態之有限狀態自動機及其中各狀態間的狀態轉移。此擷取型態之辨識與擷取就由其對應之有限狀態自動機執行，而產生適當之輸出結果，包括成功擷取個別語意元素對應之字詞(進入最終狀態及狀態轉移過程之辨識紀錄) 或型態不符合(無法進入最終狀態)。以擷取型態『TNA』為例，語意元素T的屬性為RTE，依據圖三演算法的第4至第5行，將會建立一個從1號狀態至2號狀態的transition，而其transition input為T；而依據演算法的第7至第8行，系統將會建立一個從1號狀態至0號狀態(0-state)的transition，其transition input為-T。而語意元素N的屬性為EOE，依據演算法的第12行，將會建立一個從2號狀態至2號狀態的transition，接著依據演算法的第16行，其transition input為-A。語意元素A的屬性為ROE，依據演算法的

第4至第5行，將會建立一個從2號狀態至3號狀態的transition，其transition input為A。其中最終狀態(Final_State)為3號狀態。圖四為依照此演算法對擷取型態『TNA』所自動建立之狀態轉換表及其相對應之狀態轉換圖。

	Transition Input			
State	**T**	**-T**	**A**	**-A**
1	**2**	**0**		
2			**3**	**2**

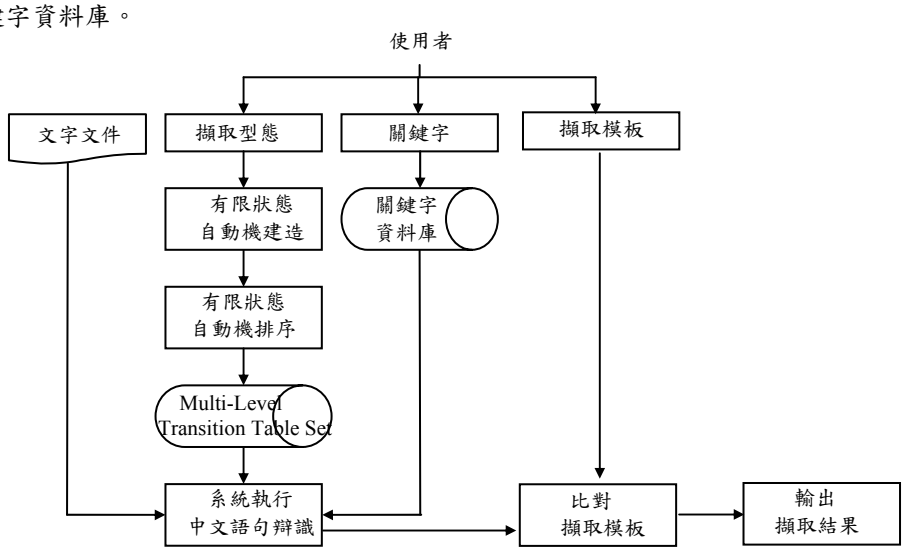


圖四：擷取型態『TNA』之狀態轉換表及圖

此有限狀態自動機的運作過程如下，由1號狀態起始，開始針對目標語句進行辨識。若辨識到語意元素T所對應的關鍵字，則進入2號狀態。反之，若在1號狀態時遇到除了語意元素T外的任何其他字詞，此有限狀態自動機將會進入0號狀態。在2號狀態時，若辨識到語意元素A所對應的關鍵字，則進入3號狀態。在由2號狀態進入3號狀態以前，所遇到的除了語意元素A外的任何其他字詞，將會使此有限狀態自動機以迴圈的方式停留在2號狀態，而這一個迴圈的動作則對應於屬性為EOE的語意元素的字詞蒐集擷取。以擷取型態『TNA』來說，此一迴圈的動作動應於語意元素N對應字詞之擷取。

3.6 系統架構

我們發展的型態比對模型與機制包括語意元素屬性的訂定、關鍵字的蒐集、擷取模板、多層次擷取型態及有限狀態自動機的運作對應等，針對含有特定主題的大量中文文本，進行個別擷取目標主題相關資訊的完整萃取，希望能達成高度正確的資訊擷取，進而匯集成具有實用價值的特定主題結構性資料庫。我們將這些模組整合成一個系統架構(圖五)，其運作流程大致分為三個階段，第一階段為由使用者根據其擷取需求，建立主題領域所需的各項資訊，包括擷取模板的定義、擷取型態的建立、屬於ROE及RTE的語意元素所對應的關鍵字的蒐集與輸入。此階段為擷取主題的訂定與辨識資訊的設置。在第二階段中，由系統將使用者給予之各種擷取型態自動轉換成相對應之有限狀態自動機，接著再透過排序演算法以狀態數目的多少，建立個別有限狀態自動機被執行比對之優先次序。通常，短的擷取型態較籠統，長的擷取型態則較明確。所以，應該以較長的擷取型態優先比對，以避免產生辨識上之錯誤或模糊。排序後之有限狀態自動機被存至Multi-Level Transition Table Set資料庫，而使用者所建立之關鍵字及其對應之語意元素則存至關鍵字資料庫。



圖五：系統架構與執行流程

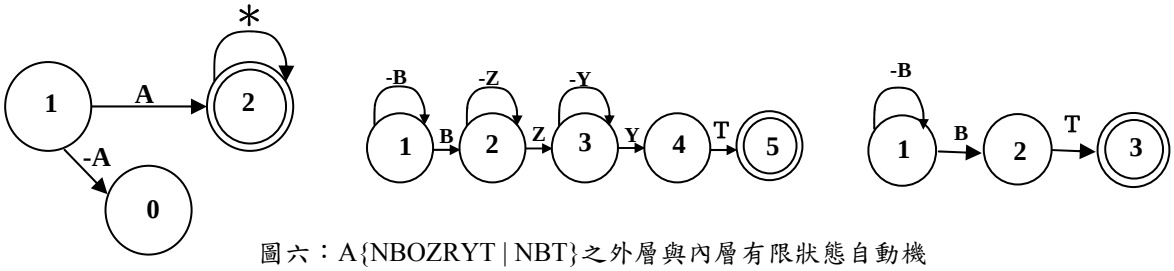
第三階段為系統執行辨識與擷取階段，系統先將輸入文本以句號及逗號切割成個別的句子，再針對每一個句子進行型態辨識。系統從Multi-Level Transition Table Set資料庫中讀取一個Transition Table以執行一個擷取型態之有限狀態自動機，由外而內進行比對與擷取的動作。有限狀態自動機在執行時，系統會將目前遇到的中文字詞與關鍵字資料庫裡的關鍵字詞進行比對(採長詞優先的方式)，經比對確認的語意元素則輸入到有限狀態機中進行狀態的轉移。當有限狀態自動機停留在Final_State，則表示目前處理的中文語句符合此有限狀態自動機所表示的擷取型態。反之，若此中文語句未解譯完就進入0-state或者此中文語句解譯完後最後的狀態不是停留在Final_State，就表示目前比對的型態並不吻合，系統會再讀取下一個擷取型態的有限狀態機繼續比對辨識。

我們以中文語句『任命鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為專員。』為執行範例，說明系統運作過程。在第一階段，使用者所建立的主題領域資訊中與此範例有關的部分顯示於表一。在系統模組中，這些資訊分別存於擷取型態、擷取模板、關鍵字等不同模組，為了呈現上的方便，我們將之顯示於一個表中(表一)。

表一：由使用者建立之部分主題領域資訊

多層次擷取型態	A{NBOZRYT | NBT}	
語意元素	屬性	對應關鍵字
A (異動種類)	RTE	任命、特任、特派、派
N (人員姓名)	EOE	N/A
B (身分賦予)	ROE	為
O(組織單位)	EOE	N/A
Z(職等種類)	ROE	簡任第
R(職等等級)	EOE	N/A
Y(職等稱謂)	ROE	職等
T(職位)	RTE	處長、專員、部長、…

在第二階段，系統建立多層次擷取型態所對應的有限狀態自動機。表一所顯示的擷取型態將會產生一個處理外層中文語句的有限狀態自動機及兩個處理內層中文語句的有限狀態自動機(圖六)。接著在第三階段，當此一擷取型態被選取與語句『任命鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為專員』進行比對時，關鍵字"任命"會先被比對確認，其相對應之語意元素A則輸入至有限狀態自動機進行狀態的轉移，此時有限狀態自動機將會由1號狀態進入2號狀態。由於2號狀態為最終狀態且有擷取的動作，再加上狀態轉移的語意元素為＊(任何的語意元素)，所以最後的擷取結果將為"任命"及"鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為專員"，其中擷取結果"任命"將會被暫存起來，而"鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為專員"將繼續交由內層之有限狀態自動機進行辨識。



圖六：A{NBOZRYT | NBT}之外層與內層有限狀態自動機

接著目前之語句將被切割成兩個子句，其中"鍾萬梅為行政院客家委員會簡任第十二職等處長"經內層有限狀態自動機辨識，同時比對擷取模板後所得到擷取結果如下：任命(A)、鍾萬梅(N)、行政院客家委員會(O)、十二(R)、處長(T)。語句"黃崇烈為專員"則被另一個內層有限狀態自動機所辨識，經比對擷取模板後所得到擷取結果如下：任命(A)、黃崇烈(N)、行政院客家委員會(O)、專員(T)。其中異動種類與組織單位兩個語意元素在擷取模板中所定義的主題資訊構成屬性為"context-dependent"，而職等之語意元素的屬性為"optional" 。因此，以"鍾萬梅"及"黃崇烈"為主體的兩個擷取目標將在這兩個語意元素上，相互參照內外層擷取結果，補上擷取模板所定義的語意元素資訊，既"鍾萬梅"的擷取目標補上外層的異動種類資訊"任命"，而"黃崇烈" 的擷取目標補上外層的異動種類資訊"任命"及前一個擷取目標("鍾萬梅")的組織單位資訊" 行政院客家委員會"，但職等之語意元素之資訊仍然維持空白。這些語意元素之資訊填補動作符合原文之含意。

4 實驗評估

為了有效驗證此一主題導向資訊擷取系統，我們大量蒐集與轉換『總政府人事任免公報』，建立了從1981年(民國70年1月)到2003年(民國92年6月)的『總統府人事任免公報』電子檔，共1788期的實驗文本資料。每一期的人事任免公報為以總統令形式發布的政府各部門人事訊息，長短不一。內容主要為以句號區隔的人事異動命令，每一道人事命令有時只針對一個人，有時則可能牽涉到一、二十人。由於人事異動主要是描述人員職務工作的變更，所以，一個擷取目標就是以每一個人員為主體的相關異動資訊。本研究的實驗文本資料中，約有10萬個擷取目標。而人事命令中有許多精簡、共用、省略的情形，系統必須為每一個擷取目標建立完整的異動資料，包括『人員姓名』、『異動種類(就任/免職)』、『組織單位』、『職等』、『職稱』等。

在系統運作的第一階段，由系統發展者扮演使用者的角色，定義擷取模板，建立了約30種擷取型態、約130個關鍵字。接著，系統經過第二階段的有限狀態自動機建置，於第三階段，對實驗文本資料一一進行辨識擷取。最後，產生約10萬個擷取目標異動資料的輸出。

4.1 評估方式

資訊擷取結果的正確性的檢驗，唯有以人工方式，一一核對每筆擷取資料與原來的相關人事命令是否吻合，同時，也必須確定擷取模板中，每一個語意元素欄位中所填入的擷取字詞是正確的，沒有因誤判而錯置或切割錯誤的情形。我們的實驗結果共有約10萬筆資料的輸出，然而在有限的人力與時間下，我們無法完全核對所有輸出結果。所以，我們以四種採樣方式，選取部份區間的輸出資料進行人工檢驗核對，希望以此推測所有輸出資料的可能評估結果。

我們規劃的四種採樣方式為：(1)連續區間(continuous interval): 涵蓋從1998年1月到2003年6月的所有文本，共有27,541個擷取目標；(2)隨機(不重複)選取(random item): 從1981年1月到1997年12月的文本中，隨機選取1200個命令句；(3)規則間隔區塊(regular block): 從1981年1月到1997年12月的文本中，每隔約4300個句子就選取連續的120個句子為一個區塊，共選取10個區塊，1200個句子；(4) 隨機(不重複)區塊(random block): 從1981年1月到1997年12月的文本中，以連續的120個句子為一個區塊，隨機選取10個區塊。

此外，我們以擷取目標為單位，評估系統在擷取結果上的成效。當系統所擷取出來的各項欄位資訊均符合擷取模板中應該對應的語意元素時，才視為正確的擷取結果。因此，只要有一個欄位發生錯誤，包括不正確的空白、錯置、不正確的字詞切割等，即視為一錯誤的擷取。

在評估的量度方面，我們採用精確度(precision)、回收率(recall)、及F-measure。其中，精確度(p)的算式為正確擷取數與系統擷取數之比率；回收率(r)的算式為正確擷取數與應擷取數之比率；而F-measure的算式為$2pr / (p+r)$，我們採取p與r相同的權重($\beta = 1$)，表示對精確度與回收率採取同樣的重視。

4.2 實驗結果

本研究的實驗結果列於表二。從各項實驗資料可以看出，系統在『總統府人事任免公報』領域約20幾年的資料上，有著不錯的精確度及回收率，印證了以型態辨識的方法應用在主題導向資訊擷取上的可行性。本系統的精確度可達約98%，顯示系統有著高準確度的擷取執行能力；而回收率可達約97%，顯示我們掌握了絕大部份的擷取型態與關鍵字。因此，相信使用者透過擷取模板的給定、擷取型態與關鍵字的建立、系統的運作等執行方式，就能針對特定主題領域的文本，產生具有實用價值的擷取結果。

表二：實驗結果數據

	Continuous Interval (1998-2003)	**Random Item (1981~1997)**	**Regular Block (1981~1997)**	**Random Block (1981~1997)**
應擷取數	27541	2159	2424	1706
系統擷取數	27340	2137	2401	1699
正確擷取數	26916	2102	2370	1671
精確度	98.45%	98.36%	98.71%	98.35%
回收率	97.73%	97.36%	97.77%	97.95%
F-1	98.09%	97.86%	98.24%	98.15%

4.3 結果討論

根據我們的觀察與分析，系統在辨識語句及擷取結果時產生失誤的原因有三種：

(1) 部分擷取型態沒有掌握: 導致某些語句能被其他較鬆的擷取型態對應，而產生錯誤的辨識與擷取，影響到系統的精確度與回收率。另外，也可能使某些語句無法被系統中的任何擷取型態對應，而沒有擷取任何資訊，影響到系統的回收率。

(2) 部分職稱關鍵字沒有掌握: 在擷取模板中，職稱的語意元素是屬於 RTE，是系統依賴以進行比對與辨識的關鍵字之一種。在二十幾年的公報中，我們沒有掌握到所有的職稱，致使部分語句辨識錯誤，影響到系統的精確度與回收率。

(3) 擷取內容含關鍵字: 在擷取模板中，人員姓名及組織單位的語意元素是屬於 EOE，必須依賴緊接其後的 ROE 或 RTE，比對辨識到特定關鍵字後，進行狀態轉移到下一個狀態。假如 EOE 所對應的字詞中含有其後的 ROE 或 RTE 的關鍵字，就會造成辨識上的錯誤，影響系統的精確度與回收率。例如，在人員姓名(EOE)中含有其後的 ROE 關鍵字"為"，或在組織單位(EOE)中含有其後的職稱(RTE)關鍵字，如"審計部審計"或"XXX 委員會委員"等。

因為擷取型態、關鍵字的不足所產生的失誤，可透過補足擷取型態、關鍵字的方式來解決，而這也反應以知識工程法建構的資訊擷取系統，處理的範圍與系統的正確性受到文本領域知識涵蓋程度的直接影響。例如，由系統發展者觀察部分文本而建置的30幾種擷取型態與130個關鍵字在20年的人事任免公報中可能只有少數遺漏，所以，系統可以達成相當高的精確度與回收率。這些系統辨識所需之文本領域資訊的建置，並不需要特定的背景與知識，只要具備一般的中文能力既可勝任。這是因為政府人事任免公報為單一主題的制式文件，語言表達的變化空間相對較小，所以，達成極高精確率與回收率的可能性較大。另外，由"擷取內容含關鍵字"因素所造成的錯誤，則是屬於系統辨識機制如何因應字詞屬性模糊性的問題。我們認為這種辨識問題大致可以用兩種方式改善。第一種是考慮更多的檢查條件，訂定更嚴謹的狀態轉移條件。第二種是以文本領域知識對於擷取內容(EOE)的語意元素加上條件定義，如字數限制。

5 結論及未來研究方向

我們以中文非結構性文本為擷取對象，發展高正確率的主題導向資訊擷取系統，透過擷取模板的建立、多層次擷取型態的訂定、及有限狀態自動機的轉換，系統展現出高度實用的價值。我們採用知識工程的方式來建立資訊擷取系統，雖然在政府人事任免公報領域有足夠的描述與處理能力，但是如果過度依賴使用者來建立擷取型態與關鍵字，就容易造成使用者的負擔，也造成系統可攜性(portability)的不足。未來我們希望透過與使用者互動或是自動學習的方式，建立擷取型態與關鍵字，以提高擷取的正確率與系統的可攜性。另外，我們也將考慮進一步了解與評估中文斷詞與named entity辨識技術對本系統的可能助益。最後，本系統將繼續在公報涵蓋時間的完整性上努力，建置從民國37年第1期到最新出版之公報的全面擷取資料庫。以系統可擴充性(scalability)的角度而言，本系統的辨識機制沒有任何預期的困難，但在辨識知識上，可能必須建置一些新舊文本中可能出現的(系統尚未具備的)辨識型態與職稱關鍵字。

參考文獻

[1] Information extraction: a multidisciplinary approach to an emerging information technology: international summer school, SCIE-97, Frascati, Italy, July 14-18, 1997.

[2] Jim Cowie , Wendy Lehnert. Information Extraction, *Communications of the ACM,* 39 (1), pp. 80-91, 1996.

[3] Applet, D. E. and Israel, D. J. Introduction to Information Extraction Technology. *In Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 1999.

[4] N. Kushmerick, D. Weld, and R. Doorenbos. Wrapper Induction for information extraction. *In Proceedings of the 15th International Joint Conference on AI (IJCAI-97),* pp. 729-737, 1997.

[5] Chun-Nan Hsu and Ming-Tzung Dung. Generating Finite-State Transducers for Semi- Structured Data Extraction from the Web, *Journal of Infromation Systems, Special Issue on Semi-structured Data*, Vol.23, No.8, pp. 521-538, 1998.

[6] I. Muslea, S. Minton, and C. Knoblock. STALKER: Learning Extraction Rules for Semi-structured, Web-based Information Sources. *In Proceedings of AAAI-98 Workshop on AI and Information Integration*, AAAI Press, Menlo Park, California, 1998.

[7] Chia-Hui Chang and Chun-Nan Hsu. Automatic Extraction of Information Blocks Using PAT Trees. *In Proceedings of 1999 National Computer Symposium (NCS-1999)*, Tamking University, Tamsui, Taiwan, 1999.
[8] Ellen Riloff. Automatically Constructing a Dictionary for Information Extraction Tasks. *Proceeding of the Eleventh National Conference on Artificial Intelligence,* pp.811-816, 1993.
[9] D. Applet, J. Hobbs, D. Israel, M. Kameyama, M. Tyson. The SRI MUC-5 JV FASTUS Information Extraction System. *Proceedings of the Fifth Message Understanding Conference*, 1993.
[10] Ralph Grishman, and Beth M. Sundheim. Message Understanding Conference-6 : A Brief History. *Proceedings of the 16th International Conference on Computational Linguistics (COLING 96),* Copenhagen, Denmark, 1996.
[11] 易文韜, 樹狀HTML文件之資訊擷取, 碩士論文, 台大資工, 指導教授: 許永真, 民國86年。
[12] 呂紹誠, 網際網路半結構性資料擷取系統之設計與實作, 碩士論文, 中央資工, 指導教授: 張嘉惠, 民國89年。
[13] 游基鑫, 中文資訊擷取環境建構與同指涉問題之研究, 碩士論文, 台大資工, 指導教授: 陳信希, 民國89年。
[14] 張嘉洋, 古文獻中資訊擷取之研究, 碩士論文, 台大資工, 指導教授: 歐陽彥正, 民國87年。
[15] C.-H. Chang, Information Extraction: A Pattern Mining Approach for Free-Form Text, *Proceedings of 2003 The Joint Conference on AI, Fuzzy System, and Gray System*, Taipei, Taiwan, 2003.
[16] 總統府人事任免公報, *URL :* *www.president.gov.tw/2_report/layer2.html*

Finding Relevant Concepts for Unknown Terms Using a Web-based Approach

Chen-Ming Hung[1] and Lee-Feng Chien[1, 2]

1. *Institute of Information Science, Academia Sinica*
2. *Dept. of Information Management, National Taiwan University*
Taipei, Taiwan

rglly@iis.sinica.edu.tw and lfchien@iis.sinica.edu.tw

Abstract. Previous research on automatic thesaurus construction most focused on extracting relevant terms for each term of concern from a small-scale and domain-specific corpus. This study emphasizes on utilizing the Web as the rich and dynamic corpus source for term association estimation. In addition to extracting relevant terms, we are interested in finding concept-level information for each term of concern. For a single term, our idea is that to send it into Web search engines to retrieve its relevant documents and we propose a Greedy-EM-based document clustering algorithm to cluster them and determine an appropriate number of relevant concepts for the term. Then the keywords with the highest *weighted log likelihood ratio* in each cluster are treated as the label(s) of the associated concept cluster for the term of concern. With some initial experiments, the proposed approach has been shown its potential in finding relevant concepts for unknown terms.

1. INTRODUCTION

It has been well recognized that a thesaurus is crucial for representing vocabulary knowledge and helping users to reformulate queries in information retrieval systems. One of the important functions of a thesaurus is to provide the information of term associations for information retrieval systems. Previous research on automatic thesaurus construction most focused on extracting relevant terms for each term of concern from a small-scale and domain-specific corpus. In this study, there are several differences extended from the previous research. First, this study emphasizes on utilizing the Web as the rich and dynamic corpus source for term association estimation. Second, the thesaurus to be constructed has no domain limitation and is pursued to be able to benefit Web information retrieval, e.g. to help users disambiguate their search interests, when users gave poor or short queries. Third, in addition to extracting relevant terms, in this study we are interested in finding concept-level information for each term of concern. For example, for a term "National Taiwan University" given by a user, it might contain some different but relevant concepts from users' point of view, such as "main page of National Taiwan University", "entrance examination of NTU", "the Hospital of NTU", etc. The purpose of this paper is, therefore, to develop an efficient approach to deal with the above problem.

In information retrieval researching area, extracting concepts contained in one text always plays a key role. However, in traditional way, if the text is too short, it is almost impossible to get enough information to extract the contained concepts. In this paper, utilizing the abundant corpora on the World Wide Web, we attempt to find the concepts contained in arbitrary length of topic-specific texts, even only a single term. For a single term, our idea is that to send it into Web search engines to retrieve its relevant documents, and a Greedy-EM-based document clustering algorithm is developed to cluster these documents into an appropriate number of concept clusters, with the similarity of the documents. Then the terms with the *highest weighted log likelihood ratio* in each clustered document group are treated as the label(s) of the associated concept cluster for the term of concern. To cluster the extracted documents into an unknown number of concept mixtures is important, because it is hard to know an exact number of concepts should be contained in a single term.

Compared with general text documents, a single term is much shorter and typically do not contain enough information to extract adequate and reliable features. To assist the relevance judgment between short terms, additional knowledge sources would be exploited. Our basic idea is to exploit the Web. Adequate contexts of a single term, e.g., the neighboring sentences of the term, can be extracted from large amounts of Web pages. We found that it is convenient to implement our idea using the existent search engines. A single term could be treated as a query with a certain search request. And its contexts are then obtained directly from the highly ranked search-result snippets, e.g., the titles and descriptions of search-result entries, and the texts surrounding matched terms.

The proposed approach relies on an efficient document clustering technique. Usually, in document clustering techniques [1, 3, 4, 8], each text in a training set is transformed to a certain vector space, then begin agglomerated with another one text step by step depending on their cosine similarity [9]. Thus, a proper

transformation from text to vector space, like TFIDF [9], takes the heavy duty of classification accuracy or concept extraction result. However, a good transformation, i.e. feature extraction, needs a well-labeled training data to support; this is not such an easy task in real world. The idea of this paper is to modify the vector space transformation as probabilistic framework.

With the extracted training data from the web, the Greedy EM algorithm [5, 7] is applied in this paper to automatically determine an appropriate number of concepts contained in the given single term through clustering the training documents. This is important while doing relevant concept extraction; otherwise, the number of concepts has to be assumed previously, it is difficult and impractical in real world. After clustering the training documents extracted from the Web into a certain number of mixtures, for each mixture, the representation of this mixture is straightforwardly defined as the term with the highest *weighted log likelihood ratio* in this mixture. With some initial experiments, the proposed approach has been shown its potential in finding relevant concepts for terms of concern.

The remainder of the paper is organized as follows. Section 2 briefly describes the background assumption, i.e. Naïve Bayes, and the modeling based on Naive Bayes. Section 3 describes the overall proposed approach in this paper, including the main idea of the greedy EM algorithm and its application to decide the number of concept domains contained in the training data from the web; in addition, generates keywords via comparing the *weighted log likelihood ratio*. Section 4 shows the experiments and their result. The summary and our future work are described in Section 5.

2. NAÏVE BAYES ASSUMPTION AND DOCUMENT CLASSIFICATION

Before introducing our proposed approach, here introduce a well known way of text representation, i.e., Naive Bayes assumption. Naive Bayes assumption is a particular probabilistic generative model for text. First, introduce some notation about text representation. A document, d_i, is considered to be an ordered list of words, $\{w_{d_{i,1}}, w_{d_{i,2}}, \mathbb{K}\ , w_{d_{i,|d_i|}}\}$, where $w_{d_{i,j}}$ means the jth words and $|d_i|$ means the number of words in d_i. Second, every document is assumed generated by a mixture of components $\{C_k\}$ (relevant concept clusters), for k=1 to K. Thus, we can characterize the likelihood of document d_i with a sum of total probability over all mixture components:

$$p(d_i \mid \theta) = \sum_{k=1}^{K} p(C_k \mid \theta)\, p(d_i \mid C_k, \theta) \qquad (1)$$

Furthermore, for each topic class C_k of concern, we can express the probability of a document as:

$$\begin{aligned} p(d_i \mid C_k, \theta) &= p(< w_{d_{i,1}}, w_{d_{i,2}}, \mathbb{K}\ , w_{d_{i,|d_i|}} > \mid C_k, \theta) \\ &= \prod_{j=1}^{|d_i|} p(w_{d_{i,j}} \mid C_k, \theta, w_{d_{i,z}}, z < j) \end{aligned} \qquad (2)$$

$$p(w_{d_{i,j}} \mid C_k, \theta, w_{d_{i,z}}, z < j) = p(w_{d_{i,j}} \mid C_k, \theta) \qquad (3)$$

Based on standard Naive Bayes assumption, the words of a document are generated independently of context, that is, independently of the other words in the same document given the class model. We further assume that the probability of a word is independent of its position within the document. Combine (1) and (2),

$$p(d_i \mid C_k, \theta) = \prod_{j=1}^{|d_i|} p(w_{d_{i,j}} \mid C_k, \theta) \qquad (4)$$

Thus, the parameters of an individual class are the collection of word probabilities, $\theta_{w_t|C_k} = p(w_t \mid C_k, \theta)$. The other parameters are the weight of mixture class, $p(C_k \mid \theta)$, that is, the prior probabilities of class, C_k. The set of parameters is $\theta = \{\theta_{w_t|C_k}, \theta_{C_k}\}$. As will be described in next section, the proposed document clustering is designed fully based on the parameters.

3. RELEVANT CONCEPTS EXTRACTION

In this section, we describe the overall framework of the proposed approach. Suppose given a single term, T, and its relevant concepts are our interest. The first step of the approach is to send *T* into search engines to retrieve the relevant documents as the corpus. Note that the retrieved documents are the so-called snippets defined in [2]. The detailed process of the approach is described below.

3.1 The proposed Approach

Suppose given a single term, *T*; then the process of relevant-concept extractions is designed as:

Step 1. Send *T* into search engines to retrieve *N* snippets as the Web-based corpus, D_T.

Step 2. Apply the Greedy EM algorithm to cluster D_T into *K* mixtures (clusters), $\{C_k\}_{k=1}^{K}$, where *K* is dynamically determined.

Step 3. For each C_k, k=1 to *K*, choose the term (s) with the highest *weighted log likelihood ratio* as the label (s) of C_k.

3.2 The Greedy EM Algorithm

Because we have no idea about the exact number of concepts strongly associated with each given term, thus for each term it's straightforward to apply the Greedy EM algorithm to clustering the relevant documents into an auto-determined number of clusters. The algorithm is a top-down clustering algorithm which is based on the assumptions of the theoretical evidence developed in [5, 7]. Its basic idea is to suppose that all the relevant documents belong to one component (concept cluster) at the initial stage, then successively adding one more component (concept cluster) and redistributing the relevant documents step by step until the maximal likelihood is approached.

Figure 1 shows the proposed approach and it is summarized in the following.

a) Set *K*=1 and initialize $\theta_{C_k} = 1$ and $\theta_{w_t|C_k}$ straightforwardly by w_t's frequency, for all w_t shown in D_T.

b) Perform EM steps until convergence, then $\theta^i = \{\theta_{w_t|C_k}, \theta_{C_k}\}_{k=1}^{K}$

c) Calculate the likelihood, $L(\theta^i)$.

d) Allocate one more mixture given initial modeling, i.e. $\theta_{K+1} = \{\theta_{w_t|C_{K+1}}, \theta_{C_{K+1}}\}$, described in section 3.2.2.

e) Keep θ^i fixed, and use partial EM techniques, described in section 3.2.3, to update θ_{K+1}.

f) Set $\theta^{i+1} = \{\theta_{w_t|C_k}, \theta_{C_k}\}_{k=1}^{K+1}$. Calculate the likelihood, $L(\theta^{i+1})$.

g) Stop if $L(\theta^{i+1}) < L(\theta^i)$; otherwise, return to c) and set *K*=*K*+1.

3.2.1 Likelihood Function

As described previously, all relevant documents belong to one mixture initially; then check the likelihood to see if it is proper to add a new mixture. Thus, given *K* mixture components, the likelihood of *K*+1 is defined as:

$$L_{K+1}(D_T) = (1-\alpha)L_K(D_T) + \alpha\phi(D_T, \theta_{K+1}) \quad (5)$$

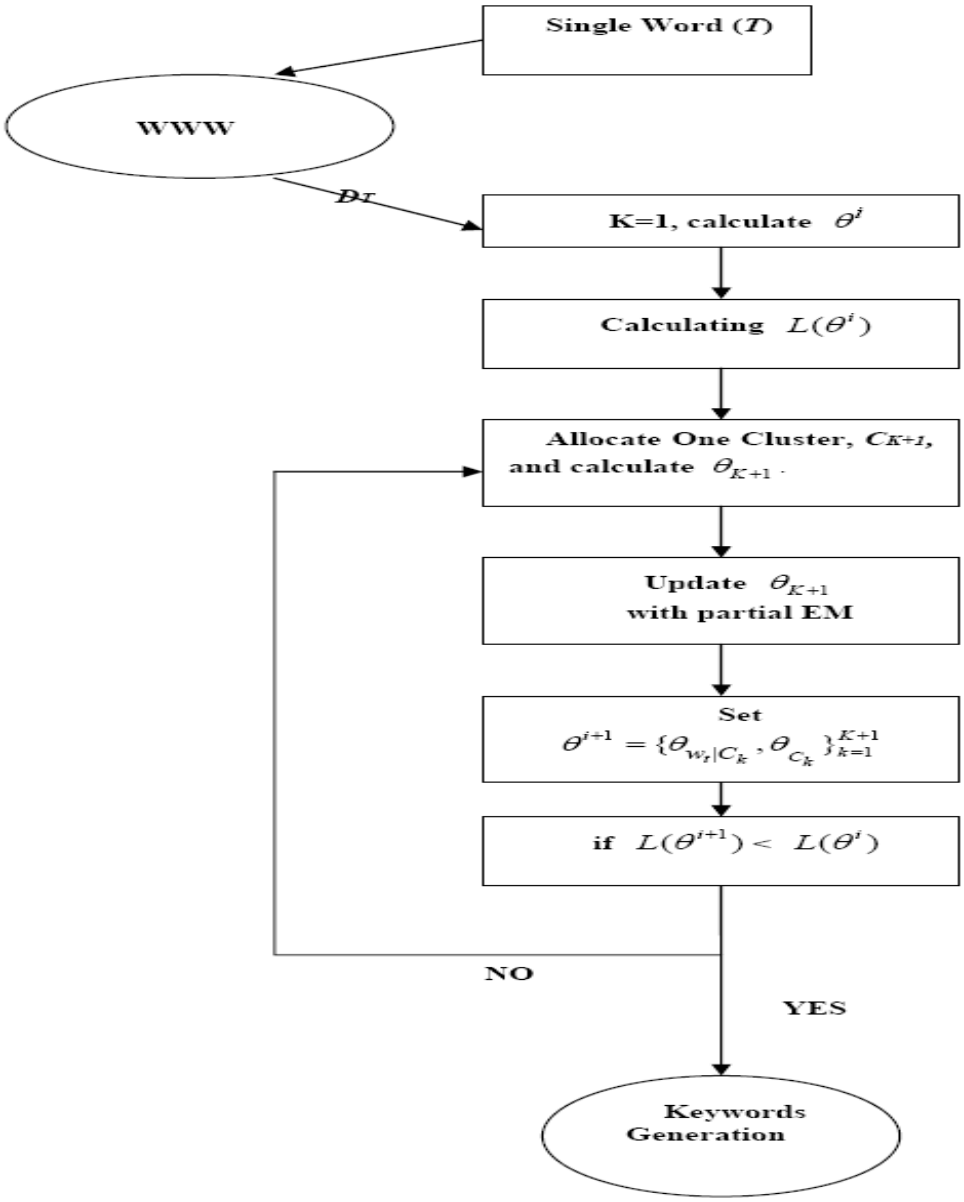


Figure 1. Overall Framework Proposed

with α in (0,1), where $\theta_{K+1} = \{\theta_{w_t|C_{K+1}}, \alpha\}$ is the modeling of newly added mixture C_{K+1} and $\phi(D_T, \theta_{K+1})$ is the likelihood in C_{K+1}. If $L_{K+1}(D_T) < L_K(D_T)$, then stop the allocation of new mixture; otherwise, reallocate a new one.

3.2.2 Initialize Allocated Mixture

In [7], a vector space model, initializing the newly added mixture is to calculate the first derivation with respect to α and to assume that the covariance matrix is a constant matrix. However, in our proposed probability framework, it is much more complicated because of a large amount of word probabilities, $\{\theta_{w_t|C_k}\}$ $\forall w_t$. Thus, we take the approximation of α in [6] as $\alpha = 0.5$ for K=1 and $\alpha = 2/(K+1)$ for $K \geq 2$. The initialization of $\{\theta_{w_t|C_{K+1}}\}$ $\forall w_t$ is randomized to satisfy $\sum_{\{w_t\}} \theta_{w_t|C_k} = 1$.

3.2.3 Update with Partial EM Algorithm

In order to simplify the updating problem, we take advantage of partial EM algorithm for locally search the maxima of $L_{K+1}(D_T)$. A notable property is that the original modeling for k=1 to K are fixed, only θ_{K+1} is updated.

$$\theta_{w_t|C_{K+1}} = \{p(w_t \mid C_{K+1})\}_{t=1}^{|V|} = \left\{ \frac{1+\sum_{n=1}^{|D_T|} N(w_t, d_n)\, p(C_{K+1} \mid d_n)}{|V| + \sum_{s=1}^{|V|}\sum_{n=1}^{|D_T|} N(w_s, d_n)\, p(C_{K+1} \mid d_n)} \right\}_{t=1}^{|V|} \quad (6)$$

$$\theta_{C_K+1} = p(C_{K+1}) = \frac{1+\sum_{n=1}^{|D_T|} p(C_{K+1} \mid d_n)}{(K+1)+|D|} \quad (7)$$

where $|V|$ and $|D_T|$ means the number of vocabularies and the number of documents shown in the D_T respectively.

Since only the parameters of the new components are updated, partial EM steps constitute a simple and fast method for locally searching for the maxima of L_{K+1}, without needing to resort to other computationally demanding nonlinear optimization methods.

3.3 Keyword Generation

In the process, the documents in the training data D_T will be clustered with their similarity into a set of clusters and keywords that can represent the concept of each cluster will be extracted. After clustering the relevant documents into several clusters, the distribution of each cluster in a probabilistic form can be calculated with the data in the cluster by applying the Greedy EM algorithm already described previously.

Next, we have to discover the hidden semantics inside each document cluster. However, retrieving the hidden semantics from a set of documents is a big issue. For convenience, we simply represent the meaning of a cluster with the word that has the highest *weighted log likelihood ratio*[1] among the contained words in this cluster. With this assumption, the "representative" word could be chosen directly by comparing

$$WLR(w_t \mid C_k) = p(w_t \mid C_k) \log\left(\frac{p(w_t \mid C_k)}{p(w_t \mid \overline{C}_k)} \right) \quad (8)$$

for k=1 to K, where $p(w_t \mid C_k)$ means the probability of word w_t in component C_k and $p(w_t \mid \overline{C}_k)$ means sum of the probabilities of word w_t in those clusters except C_k.

4. EXPERIMENTS

In real world, for an unknown term, its associated concepts are what we are interested in; thus, in this section, we will show the experiment results obtained in evaluating a set of test terms. Before the larger amount of experiment, let's preview the experiment of "ATM" to determine the number of retrieved relevant documents. Google (http://www.google.com) is the main search engine which we utilized in the following experiment.

4.1 Appropriate Number of Retrieved Relevant Documents

[1] The sum of this quantity over all words is the Kullback-Leibler divergence between the distribution of words in C_k and the distribution of words in $\overline{C}_k$, (Cover and Thomas, 1991).

We assume that too many retrieved documents will cause noises, but too few won't contain enough information about this unknown term. Thus, the appropriate number of retrieved relevant documents has to be decided. “ATM” in dictionary has six hidden semantics, which are “Automated Teller Machine”, “Asynchronous Transfer Mode”, “Act of Trade Marks”, “Air Traffic Management”, “Atmosphere” and “Association of Teachers of Mathematics” respectively. Table 1 shows the bi-gram extracted concepts via number of retrieved texts.

Table 1: *Extracted concept clusters in “ATM” with respect to different numbers of retrieved relevant terms*

# of training texts	Extracted concept clusters
100	*ATM {card, cell, internetworking, standards}, asynchronous transfer*
200	*ATM {access, information, networking, standards, locations}, credit union, debit cards , safety tips, teller machine, telescope makers*
300	*ATM {applications, cards, fees, networking, surcharges, locations}, adaptation layers, credit union, debit cards, token rings, teller machine,*
400	*ATM {applications, crashes, services, transactions, networking, security}, branch locator, financial institution, personal banking, public transport, telangiectasia mutated*
500	*ataxia telangiectasia, ATM {applications, asynchronous, crashes, products, protocol, resource}*
600	*ataxia telangiectasia, ATM {applications, crashes, protocol, technology}, atmospheric science, communication technology, electronics engineering, network interface, public transport, wan switches, rights reserved*
700	*air traffic, ataxia telangiectasia, ATM {crashes, debit, encryptor, protocol, surcharge, traffic}, atmospheric science, checking account, communication technology, electronics engineering, network interface, public transport*
800	*ataxia telangiectasia, ATM {adapters, crime, debit, protocol, surcharge, usage, cards}, atmospheric sciences, business checking, communication technology*
900	*24 hours, ataxia telangiectasia, ATM {adapters, connections, crashes, crime, debit, industry, protocol, resources}*
1000	*ATM networks, Arizona federal, 24 hour*

Table 1 shows a challenge that choosing the term with the highest *weighted log likelihood ratio* as the label of one concept cluster can not effectively describe its complete semantics appropriately; in addition, for example, “Automated Teller Machine” is composed of many aspects, like security, location, cards, and etc. Thus, concept domain of “Automated Teller Machine” could be figured out while “ATM applications”, “ATM locations”, “ATM surcharges”, and some other aspects associated with “Automated Teller Machine” being extracted. Similarly, “Air Traffic Management” could be figured out while “public transport”, “air traffic” being extracted.

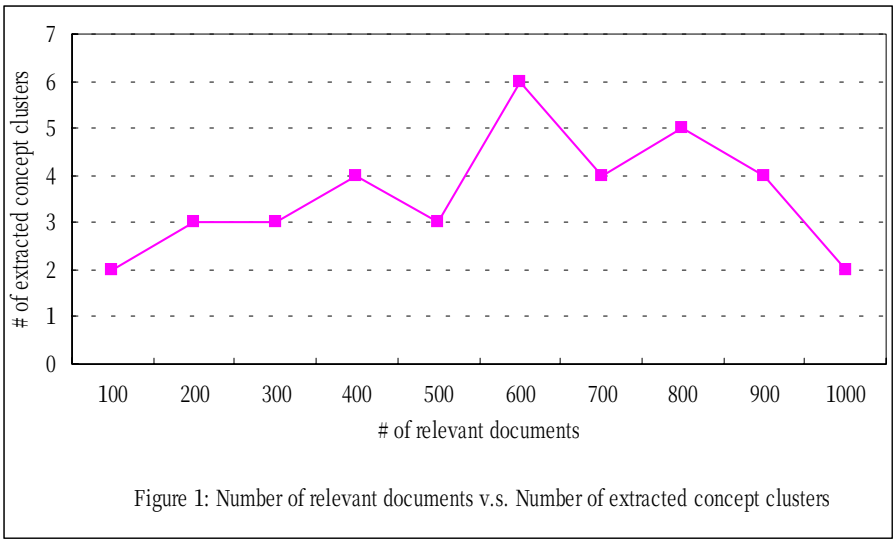


Figure 1: Number of relevant documents v.s. Number of extracted concept clusters

Except the six hidden semantic clusters in ATM, some other concept clusters were also extracted, e.g. “Amateur Telescope Maker” because of “telescope makers” extracted and “Ataxia Telangiectasia Mutated” because of “ataxia teleangiectasia” extracted. One more interesting thing is that the more retrieved relevant documents not necessarily direct to the more extracted concept clusters (Figure 1). This phenomenon is caused from the extra noises extracted from the more relevant documents. The extra noises will not only worsen the performance of the greedy EM algorithm but also generate improper relevant terms from the clustered groups, which will not be considered as “good” categories manually. For each test term, considering the time cost and the marginal gain of extracted concepts, 600 relevant documents were retrieved from Web. Of course 600 relevant documents are not always appropriate for all cases, but for convenience, it was adopted.

4.2 Data Description

The experiments took the "Computer Science" hierarchy in Yahoo! as the evaluation. There were totally 36 concepts in second level in the "Computer Science" hierarchy (as in Table 2), 177 objects in the third level and 278 objects in fourth level, all rooted at the concept "Computer Science". We divided the objects in third-level and fourth-level into three groups: full articles, which were the Web pages linked from Yahoo!'s website list under the Computer Science hierarchy, short documents, which were the site description offered by Yahoo!, and text segments, which were the directory names. We randomly chose 30 text segments from the third-level plus the fourth-level objects. The 30 proper nouns are shown in Table 3.

Table 2: "Computer Science" hierarchy in Yahoo!

algorithms	library and information science
architecture	linguistics
artificial intelligence	logic programming
compression	mobile computing
computational learning theory	modeling
computational sciences	networks
computer vision	neural networks
databases	objective oriented programming
distributed computing	operating systems
DNA-based computing	quantum computing
electronic computer aided design	real time computing
end user programming	robotics
finite model theory	security and encryption
formal methods	software engineering
graphics	supercomputing and parallel computing
handwriting recognition	symbolic computation
human computer interaction	user interface
knowledge sciences	virtual reality

4.3 Relevant-Concepts Extraction

In Section 3, the Greedy EM algorithm is treated as the unsupervised learning method which clusters retrieved relevant documents to extract hidden concepts for each test term.

Table 3: *30 terms from 3rd level and 4th level in Yahoo!’s CS hierarchy*

ActiveX	IRIX	ROADS
CMX	ISDN	RSA
CORBA	Jini	Ray Tracing
Darwin	JXTA	SETL
Ebonics	Mach	SIP
Eponyms	Mesa	Trigonometry
Figlet	PGP – Pretty Good Privacy	VHDL
GNU	PPP	VMS
Hobo Signs	Puns	WAIS
Hurd	QNX	Xinu

Table 4 shows the extracted bi-gram concept clusters for the 30 randomly chosen CS terms; this means that only bi-gram terms in the retrieved documents were extracted. The number of hidden concept clusters in each term was determined automatically by the Greedy EM algorithm.

Table 4: B*i-gram concept clusters for the test terms in Yahoo!'s CS hierarchy*

Test Terms	**Extracted Concept Clusters**
ActiveX	*ActiveX {control, vs, server}*
CMX	*CMX-RTX RTOS, multi-tasking operating, CMX {3000, 5000}, San Jose, Jose BLVD*
CORBA	*application development, C++ software, CORBA {2.2, orbs, servers}, distributed {applications, programming}, IDL compiler, language {IDL, mapping}, object-oriented programming, request broker*
Darwin	*Charles Darwin, Darwin 6.0.2*
Ebonics	*black English, African Americans, Ebonics X-mas*
Eponyms	*medical phenomena, on-line medical, aortic regurgitation, encyclopediaof medical, Firkin Judith, esophageal surgery, historical allusions*
Figlet	*art characters, Figlet {Frank, frontend, package, RPM, tool}, assorted fonts*
GNU	*GNU {aspell, coding, compilers, documentation, desktop}, license GPL, public licenseterms, reference card*
Hobo Signs	*ideogram carved, 45 signs*
Hurd	*Alexander Hurd, Debian developers, Debian Gnu, GNU operating, Hannah Hurd, Hon Lord*
IRIX	*Sgi IRIX, IRIX reinstall, 2.6.5.7 Sgi, 3D graphics*
ISDN	*digital {access, networks, telephone}, Arca technologies, communication standards, copper wire, data {applications, communications, services}, external ISDN, integrated services*
Jini	*Jini technology, Jini Ji*
JXTA	*project JXTA, peer peer*
Mach	*Mustang Mach, disc golf*
Mesa	*Mesa Verde, Mesa Quad*
PGP	*encrypt messages, foaf files, keysigning party, PGP {backend, basics, comments, corp}, ASCII armour*
PPP	*point-to-point protocol, ppp flea.,*
Puns	*French word, Japanese spelling, social sciences, bilingual puns*
QNX	*Microkernel OS, QNX {applications, machine, voyager}, alternative vendor*
ROADS	*{Access, British, Hampton} ROADA, adverse weather*
RSA	*RSA security, RSA lighting*
Ray Tracing	*Computer graphics, Carlo Ray, recursive Ray*
SETL	*set language, ab le*
SIP	*control protocol, {bring, panel, partysip,} SIP, SIP {architecture, application, client, standards}, Jonathan Rosenberg*
Trigonometry	*Trigonometric functions, advanced algebra, Benjamin Bannekers, Banneker's trigonometry*
VHDL	*Asic design, circuit VHSIC, complete VHDL, hardware design, digital logic, verilog simulation, synthesis tool*
VMS	*computational chemistry, shopping cart, administrator authentication, UNIX translation, CCL VMS, equipment corporation*
WAIS	*area informationserver, laws enacted, public laws, WAIS {client, gateway, searching, libraries}, presidential documents*
Xinu	*AMD élan, unix clone, software OS, Xinu {kernel, system}, II internetworking, master distributor, demand paging*

From Table 4, it is encouraging that the proposed approach extracted the main idea for most test CS terms. Taking "Trigonometry" for example, if we have no idea about "Trigonometry", then from "function" and "algebra" in Table 4, there is not difficult to guess that it may be a kind of mathematical functions and

developed by Benjamin Bannekers. Again, our proposed approach caught that “Darwin” is not only a British Naturalist but also the name of graphical software.

Even though the experiment result shows encouraging performance, the result was still bothered by many duplicated and noisy aspects. For example, “CORBA” means “Common Object Request Broker Architecture”; however, “C++ software” and “application development” actually only provide vague or not necessary information about “CORBA”. This was caused by the “too much effort” of the Greedy EM algorithm, which clusters the retrieved mixtures into too many groups.

5. CONCLUSIONS AND FUTURE WORK

We have presented a potential approach to finding relevant concepts for terms via utilizing World Wide Web. This approach obtained an encouraging experimental result in testing Yahoo!’s computer science hierarchy. However, the work needs more in-depth study. As what we mentioned previously, choosing the word with the highest *weighted log likelihood ratio* as the concept of a clustered group after the Greedy EM algorithm does not provide enough representative. In addition, one concept usually contains many domains, e.g. “ATM” contains security, teller machine, transaction cost, and etc. Thus, distinguishing the extracted keywords into a certain concept still needs human intervention. On the other hand, in order to solve the problem of “too much effort” of the Greedy EM algorithm, we need to modify it with another convergence criterion.

References

[1] A. Jain, M. Murty, and P. Flynn. Data Clustering: A Review. In ACM Computing Surveys, 31(3), September 1999.

[2] C. C. Huang, S. L. Chuang and L. F. Chien. LiveClassifier: Creating Hierarchical Text Classifiers through Web Corpora, *WWW* (2004).

[3] E. Rasmussen. Clustering Algorithms. In *Information Retrieval Data Structures and Algorithms*, William Frakes and Ricardo Baeza-Yates, editors, Prentice Hall, 1992.

[4] E. Voorhees. The Cluster Hypothesis Revisited. In *Proceedings of SIGIR* 1985, 95-104.

[5] J. J. Verbeek, N. Vlassis and B. J. A. Krose. Efficient Greedy Learning of Gaussian Mixture Models. *Neural Computation*, 15 (2), pp.469-485, 2003.

[6] J. Q. Li and A. R. Barron. Mixture Density Estimation. In *Advances in Neurarl Information processing Systems* 12, The MIT Press, 2000.

[7] N. Vlassis and A. Likas A Greedy Algorithm for Gaussian Mixture Learning. In *Neural Processing Letters* (15), pp. 77-87, 2002.

[8] P. Willett. Recent Trends in Hierarchic Document Clustering: A Critical Review. In *Information Processing and Management*, 24(5), 577-597, 1988.

[9] T. Joachims. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In *Machine Learning: Proceedings of the Fourteenth International Conference (ICML ’97)*, pp. 143-151.

以自組織映射圖進行計算語言學領域術語視覺化之研究
Visualizing the Terms of Computational Linguistics with Self-Organizing Maps

林頌堅
Sung-Chien Lin

世新大學資訊傳播學系
Department of Information and Communications, Shih-Hsin University
scl@cc.shu.edu.tw

摘要 本論文的研究利用自組織映射圖(SOM)技術將計算語言學相關術語對應到二維圖形，使得術語之間的關係可以在映射圖中加以呈現，提供使用者做爲資訊檢索以及了解重要研究主題的輔助工具。在本論文中，我們所探討的問題有(1)發展SOM技術應用到術語資訊視覺化的方法，(2)評估SOM技術應用到術語資訊視覺化的成效，(3)利用研究結果分析計算語言學中重要的研究主題與主題之間的關係。在SOM技術的應用中，首先從論文資料中抽取重要的術語，接著以術語之間的共現關係做爲基礎，建立每一個術語的特徵向量。再以術語特徵向量做爲輸入資料，進行SOM訓練以及將術語映射到圖形上。對於這項技術在應用上的成效評估，由於映射節點的距離關係在視覺上要需要符合術語間的相關性。因此，我們建議以特徵向量的距離與節點位置的距離之間的相關係數做爲成效評估的標準。最後，對於計算語言學領域的術語所進行的實驗中可以發現大多數相關的術語都可以映射到相鄰近的節點上，而術語所映射節點的位置也可以大致表現主題之間的關係。這個結果表示SOM技術適合應用於術語資訊視覺化。

1 緒論

本論文是一個將計算語言學相關術語(terms)對應到二維圖形的研究，其目的是希望能夠蘊含在術語之間的資訊加以視覺化(visualization)。從論文所抽取出來的術語可以表示研究問題、方法、理論與技術等論文相關的主題，若是針對某一研究領域所發表的論文進行術語抽取並加以統計，所得到的高頻術語便是這個領域的重要主題[1]。因此，這些從論文抽取出來的術語將有助於了解這個領域所發展的研究課題或是進行資訊的檢索。爲了進一步幫助使用者從大量的文件資料庫中搜尋相關的資訊來解決所面對的研究問題以及提供他們對於這個領域研究所產生的知識結構(knowledge structure)有完整的認識，可以將這些術語整理成階層式(hierarchical)組織或網路式(network)組織，來闡明術語之間的關係。在資訊檢索的技術與應用上，索引典(thesaurus)便是將某一特定領域的相關術語與它們之間的關係整理成一個階層式與網路形式的組織[2]。在索引典的結構裡，將每一個術語作爲網路中的節點，而以相關術語之間的關係作爲相應節點之間的連結。近來，許多研究提出各種術語組織的自動化方法，這些方法多以統計的叢集(clustering)技術爲組織術語的方法，將關聯性較強的術語放到相同的集合中，並且利用術語在文句中的共現(co-occurrence)關係作爲術語之間的關聯[3, 4]。利用叢集所形成集合便可以了解術語之間的關聯性，並且在同一集合中的術語往往經常共同出現在主題相關的論文中，因此這些術語集合可以呈現這個研究領域的研究主題。然而，除了利用叢集技術所形成的集合來對於術語之間的關聯進行分析之外，若能夠將術語以及它們之間的關聯呈現在圖形中，提供瀏覽與深入探索，對於檢索相關資訊與分析領域的知識結構勢必更有幫助。

『資訊視覺化』(information visualization)是以二維或三維的圖形來表現一組資料之間的可能關係，目的是輔助人們認知原本的資料間不易察覺的關係，作爲決策判斷或探索新知的依據[5]。在過去，資訊視覺化常被應用於高維的數值資料，然而由於電子文件的數量大幅增加，對於組織大量文件以及方便而有效的全文檢索介面的需求越來越大，已經有許多學者著手進行文字資訊視覺化的探討。文字資訊視覺化的目標是將每一個文字資料對應到圖形上某一位置上的一點，使得文字資料之間的相關程度(relevance)可以用圖形上點與點之間的距離加以表示，兩點間的距離愈近便表示所代表的兩筆文字資料愈相關。使用者便可以直覺地將圖形上表示的距離作爲資料間的關聯，進而了解資料的整體分布情形。因此，在文

字資訊視覺化研究中常見的做法是首先設定文字資料的特徵向量(feature vectors)，再以特徵向量來估算資料兩兩間的相關程度，接著利用映射技術將文字資料對應到圖形上，盡量使圖形上點與點的距離之間的關係保持術語相關程度間的關係。常使用的映射技術有統計導向與類神經網路導向兩類[6]。在統計導向的方法中，將所有資料間的相關程度組合成一個矩陣，每一筆資料對所有資料的相關程度對應到矩陣中的一行與一列，換言之矩陣上的每一個元素便是兩筆資料間的相關程度。接著便利用統計技術，如SVD(singular value decomposition)[7]、PCA (principal component analysis)或是MDS (multidimensional scaling) [6, 8]等，找到一組轉換矩陣將原先的矩陣加以分解與轉換，使得重要的距離訊息得以保留在新產生的矩陣中。而以轉換矩陣作爲將資料映射到圖形的依據。

另一方面，自組織映射圖(self-organizing maps, SOM)則是在應用類神經網路導向的方法到文字資訊視覺化處理中常採用的技術[9]。顧名思義，SOM是一種以資料驅動(data-driven)的非監督式學習(unsupervised learning)方法，利用資料的特徵向量作爲訓練資料，訓練一組排列成方陣的節點，從反覆的訓練過程中讓產生的映射圖反應資料之間的關係[10]。在SOM技術中，每一節點都是一個向量，向量的維度與資料特徵向量的維度相同。在經過多次的訓練過程後，所有的資料都依照其特徵向量與節點的相似程度，映射到某一個節點上，而且節點間愈接近者相似程度愈高。因此，相關程度接近的資料會映射到同一節點或鄰近的節點上，而且所投射節點之間的相對距離可以表示資料的相關程度大小，距離愈大相關程度愈小。SOM的優點包括了可以將高維資料的距離關係，以自組織的型式保留在二維的映射圖中，並且MDS等統計導向方法大多需要極大量的運算資源，且在新增資料時，無法利用先前的計算結果，在實作方面，SOM技術較容易達成。因此，近年來有相當多文字資訊視覺化的研究採用SOM作爲映射技術。

在本論文的研究中，我們嘗試將計算語言學術語的關係視覺化，利用SOM將術語之間的相關程度映射到圖形上。因此，本論文的研究問題包括：(1) 發展SOM技術應用到術語資訊視覺化的方法，(2)評估SOM技術應用到術語資訊視覺化的成效， (3)利用研究結果分析計算語言學中重要研究主題之間的關係。

本論文其餘的章節組織如下，第2節中將簡介SOM技術，並回顧利用SOM技術處理文字資訊的研究；第3節說明本研究如何利用SOM技術，將計算語言學相關術語進行資訊視覺化處理的方法，並提出成效評估的方法；第4節則是對此一研究相關實驗的結果與說明；最後的第5節是本論文的結論與未來進一步研究的建議。

2 相關研究

SOM是一種非監督式的類神經網路[10]，在資料的叢集與視覺化上，應用十分廣泛。SOM的特色包括了它的類神經網路型態(topology)與訓練模式。在SOM中，由一組反映輸入資料的節點所構成，而這些節點排列成矩陣的型態，每一個節點與其他四個節點相連接，此一結構便是所謂的特徵映射圖(feature map)。事實上，每一個節點都代表一個特徵向量，向量的維度與資料項的特徵向量維度相同。在輸入資料之後，便重複訓練過程，調適節點的特徵向量，使得特徵映射圖可以反映輸入的資訊項之間的關係。SOM與一般『向量量化』(vector quantization)在訓練過程中最大的不同是，每次的訓練時，不僅只調適節點中與輸入資料最相接近的特徵向量，而且還同時調適了在特徵映射圖上鄰近範圍(neighborhood)內所有節點的特徵向量。因此，在經過多次的訓練之後，可以使特徵向量接近的資料映射到相同或是鄰近的節點上，使得圖形具有組織化的結構，而且將原本資料在高維特徵向量的距離或接近程度表示到SOM的節點的距離。通常用來衡量節點間距離的方式爲式(1a)中的歐幾里德距離(Euclidean distance)或式(1b)中的Manhattan距離等。

$$d(n_w, n_c) = ((x_w - x_c)^2 + (y_w - y_c)^2)^{1/2} \tag{1a}$$

$$d(n_w, n_c) = | x_w - x_c | + | y_W - y_c | \tag{1b}$$

式(1)中，n_w是特徵向量與某次輸入的特徵向量最接近的節點，在SOM的訓練過程中稱爲『獲勝者』(the winner)，n_c則是映射圖上另一節點，(x_w, y_w)與(x_c, y_c)分別是節點n_w與n_c在映射圖上的座標。在訓練時，每個節點調適的幅度與這個節點跟獲勝者間的距離有關，距離愈近的節點獲得調適幅度愈大；反之，較遠的節點則調適幅度較小。

在SOM訓練的另一項特色是以訓練次數的多寡來控制每次訓練獲勝者的鄰近範圍以及特徵向量調適的幅度，使得隨著訓練次數增加，鄰近範圍與調適幅度愈來愈小，而保證SOM的訓練結果可以收斂。舉例而言，在第τ+1次的訓練中，對某一節點n_c調整的方式如式(2)所示。

$$f_c(\tau+1) \overset{def}{=} f_c(\tau) + h(\tau, d(n_w, n_c))[f_i - f_c(\tau)] \tag{2}$$

式子中，$f_c(\tau)$是表示第τ次的訓練後，節點n_c的特徵向量，f_i是輸入資料的特徵向量，$h(.)$是一個訓練次數τ與節點和獲勝者之間的距離$d(n_w, n_c)$有關的調適函數，爲節點n_c的特徵向量此次訓練的調適幅度，如上所述，當訓練次數愈多，或者距離$d(n_w, n_c)$愈大，$h(.)$所得到的值愈小。

SOM的訓練過程如下。首先，根據輸入資料的數量與特徵向量的維度設定節點的個數與特徵向量的維度，並對每一個節點隨機產生一個特徵向量。在輸入資料後，開始進行多次的訓練。在SOM的每一次訓練中，首先從輸入的資料中隨機選取一個資料，再從節點中選出與訓練資料的特徵向量最相似者，也就是獲勝者。接著如式(2)所示，根據調適函數$h(.)$計算出的調適幅度，調整獲勝者與其鄰近節點的特徵向量，使其愈加相似於訓練的特徵向量。當SOM訓練完成後，便依據術語特徵向量與節點特徵向量的接近程度，將術語映射到圖形上。

在利用SOM技術對文字資料進行叢集或視覺化的研究中，可以依據處理的對象分爲文件與術語兩類。在以文件爲處理對象的SOM研究，大多將輸入的每一筆文件表示成一個以索引詞(index terms)的出現次數爲基礎的特徵向量[11, 12]，因此，索引詞的出現情形較爲接近的文件可以映射到同一節點或鄰近的節點上。爲了使文件的特徵向量可以表示語意訊息，Wermter與Hung利用WordNet的語意階層關係，計數具有相近概念術語的出現次數作爲向量的特徵值，以SOM技術對Reuters新聞語料進行文件分類(text classification)的研究[13]。Kohonen等人則先對術語進行SOM的叢集，使得具有相關語意的術語，映射到同一節點上。再以叢集後的節點作爲基礎，計數節點對應的所有術語出現在文件資料中的次數總和作爲向量的特徵值，作爲資料縮減的技巧來處理極大量的新聞群組(newsgroups)線上文字資料[14]。此外，在文件叢集的應用中，由於以索引詞爲基礎的特徵向量維度非常高，一般的二維映射圖較難表示文件資料間所具有複雜的主題關係，因此，Merkl認爲需要表現出主題間的階層關係，可以利用階層式自組織映射圖(hierarchical self-organizing feature maps)，訓練一組多層的映射圖，使得位置在上層的映射圖之節點表示文件資料中較廣泛的主題，而以下層的映射圖之節點表示較特定概念的主題[12]。

在利用SOM處理術語的研究上，則有Ritter與Kohonen對於英語術語[15]和Ma等人對漢語及日語術語[16]叢集的研究。在術語特徵向量的設定上，Ritter與Kohonen以術語的出現(occurrences)及前後各一個術語的上下文關係(contexts)作爲特徵[15]；Ma等人則利用術語的共現次數爲基礎作爲向量的特徵[16]。

在目前利用SOM技術所進行文字資料叢集或資訊視覺化的研究，其實驗結果可以看出主題相近的文件或術語可以被映射到相同或鄰近的節點，在視覺呈現上，符合人們的認知，這些研究可以證明SOM技術應用於文字資訊視覺化的可行性。然而，從這些研究中卻也可以發現大多數研究在說明實驗結果時，多半以叢集的結果與主題的相關程度進行討論，在客觀的評估方法上也都以傳統資料分類的檢全/檢準(recall/precision)爲標準[16]，甚少討論所得到的實驗結果在不同主題間的關係。但在資訊視覺化的研究中，藉由圖形表示文件或術語之間的分布，是相當重要的目標。在進行這方面的研究時，也應該根據這方面的要求，設計一套合適的評估方法。

3 研究設計

本研究是應用SOM技術的初步研究，因此除了提出術語進行資訊視覺化處理的方法之外，如何評估其結果也是重要的研究問題。此外，在現階段的研究中，本論文採用一般的SOM技術作爲探討的對象，先以一般常用的型態與訓練模式做爲SOM的應用，來了解這項應用的可行性。更爲先進與複雜的技術如階層式自組織映射圖[12]，可在後續的研究中進行。以下首先說明以SOM進行術語資訊視覺化的方法，接著提出評估資訊視覺化成效的方法。

3.1 以SOM進行術語資訊視覺化的方法

在利用SOM進行術語資訊視覺化的方法中，首先進行術語抽取(term extraction)，從輸入的論文題名、摘要與參考文獻的題名等文字資料，抽取出計算語言學領域中重要的中英文術語[1]。判斷一個出現在文字資料中的字串是否是與這文字資料主題相關的術語可以從字串的『單元完整性』(unithood)與『主題代表性』(termhood)的兩方面著手[17]，單元完整性是指做爲術語的字串是否爲語言結構(linguistic structure)上的完整單位，如詞(words)或詞組(phrases)，而主題代表性則是指此一術語能否代表文字資料的主題並與其他主題區別。在本研究中將以統計訊息爲主，配合若干經驗法則(heuristic rules)來達到這兩項要求。首先將論文資料輸入，建立一個PAT-tree資料結構[18]，接著從PAT-tree檢取所有出現在論文資料中的字串，並計算字串在所有論文的出現總次數、字串在論文資料中的平均出現頻次和標準差(standard

deviation)以及字串前後接字的複雜度等統計資訊。其中，字串前後接字的複雜度(如式(3a, b))，加上停用詞(stop words)不能出現在字串首尾的經驗法則，用來檢測字串的單元完整性。

$$C_{1S} \overset{def}{=} -\sum_{a} \frac{F_{aS}}{F_S} \log(\frac{F_{aS}}{F_S}) \tag{3a}$$

$$C_{2S} \overset{def}{=} -\sum_{b} \frac{F_{Sb}}{F_S} \log(\frac{F_{Sb}}{F_S}) \tag{3b}$$

式(3a)和(3b)中，字串*S*的前後接字複雜度分別以C_{1S}和C_{2S}表示，*a*和*b*則代表字串*S*在論文資料中任一個可能的前接字和後接字，F_S、F_{aS}和F_{Sb}分別是字串*S*、*aS*和*Sb*的出現總次數。當字串的前後接字複雜度較小時，表示此一字串需與其前面或後面的某一字串共同構成新的字串，才能表示語法和語意上的一個單元。因此，當前後接字複雜度愈大，愈有可能表示一個完整的術語。而所檢出的高頻字串中，字串首尾經常是介詞、連詞或定詞等停用詞，因此我們過濾掉首尾爲停用詞的字串，使得過濾後的術語句有單元完整性的要求。但停用詞出現在中間的字串，如 “part of speech”，只要出現次數夠多、頻率夠高仍爲重要的術語。在另一方面，字串在所有論文的出現總次數、平均出現頻次和標準差則用來表示術語的主題代表性，出現總次數愈大的術語表示這個術語在領域中常被使用而具有重要意義，術語的平均出現頻次和標準差則可表示這個術語在論文中的使用情形，平均出現頻次愈大的術語，即有可能在許多論文中出現多次，是這些論文的重要術語；而術語的出現頻次標準差較大則表示此術語在某些特定論文出現較多次，對這些論文相當重要。所以這三項統計訊息可以作爲檢驗術語是否符合主題代表性的依據。因此，本研究即整合上述的訊息做爲判斷字串是否爲計算語言學領域中重要術語。

接著，對上述步驟所抽取出來的每一個術語設定一個特徵向量來訓練SOM。爲了產生合適的SOM，相關術語所設定的特徵向量必須相接近。如此一來，當把術語映射到SOM時，相關術語將映射到同一節點上或鄰近的節點中，所形成圖形便具有相關術語的距離將較非相關術語的距離小的特性。本研究以術語對每一個術語的共現關係的估算值做爲這個術語的特徵向量，如式(4)表示術語t_i的特徵向量f_i。

$$f_i = [o_{i,1}, \ldots, o_{i,k}, \ldots, o_{i,N}]^T \tag{4}$$

在式(4)中，假定術語抽取步驟中共得到*N*個術語，因此每一個術語的特徵向量都是一個*N*維的向量。在術語t_i的特徵向量f_i中，第*k*個元素o_{ik}是術語t_i與另一術語t_k共現關係的估算值。當比較術語t_i與t_j的相關程度時，可以比較這兩個術語與其他術語t_k之間的共現情形。一旦當t_i與t_k共同出現在某一些論文資料時，同時t_j也經常出現在這些論文資料時，術語t_i與t_j可能相關於同一個特定的主題，這兩個術語便可能相關。如果t_i與t_j有許多共同的共現術語時，t_i與t_j的特徵向量便很接近而表示兩個術語間具有較大的相關程度。以數學的方式來表示上述的說明，當我們以歐幾里德距離作爲兩個術語特徵向量之間距離的估算方式時，當兩個特徵向量具有愈多相近的元素，在特徵向量所在的*N*維空間的距離愈小，表示兩個術語的相關程度愈大；反之特徵向量之間相異的元素愈多，距離愈大，兩個術語的相關程度便愈小。

在兩個術語t_i與t_k的共現關係上，也就是上述特徵向量f_i中的元素o_{ik}之值，可以利用近來資訊檢索常使用的『隱含語義分析』(latent semantic analysis, LSA)技術[19]來進行估算，使得某些相關術語卻較少共同出現的問題可以減輕。其估算方法如下，我們首先建立『術語-文件矩陣』(term-document matrix)，以每一個抽取出來的術語對應到矩陣中的一行(row)，矩陣中的每一列(column)則對應到一筆論文資料，在矩陣中第*i*行第*p*列的元素，其值爲第*i*個術語在第*p*筆論文資料中出現的次數。接著對於『術語-文件矩陣』進行奇異值分解(singular value decomposition)，求得一組維度較小的新術語向量。比方說新向量的維度爲δ，新的術語向量組便是所有維度爲δ的向量組中，內積的估算值與原先『術語-文件矩陣』的內積誤差最小的向量組之一，術語間共現關係便以這組向量兩兩之間的向量內積值作爲估算值。而且對於缺乏共同出現的術語，此一共現關係的估算方法具有適當的補償效果，使得相關術語的特徵向量較爲接近。因此，本研究所產生的特徵向量可以作爲SOM技術的輸入，所得到的結果將比由『術語-文件矩陣』所估算的共現關係爲佳。

接下來，便對於每一個術語所產生的特徵向量進行SOM訓練。本研究中所採用的調適函數如式(5)所示，

$$h(\tau, d(n_w, n_c)) = e^{-\frac{\tau \times [d(n_w, n_c)^2 + 1]}{\alpha}} \tag{5}$$

在式(5)中，α是一個預設的參數值，用來控制訓練次數和獲勝者鄰近範圍中進行調適的節點數量。如同第二節中所提到的，對於某一節點n_c，調適函數*h*(.)所產生的調適幅度與訓練次數τ和這個節點與獲勝者n_w之間的距離$d(n_w, n_c)$有關。在本研究中採用歐幾里得距離做爲$d(n_w, n_c)$的計算方式。在式(5)中，可以發現在每次訓練中，愈接近『獲勝者』的節點($d(n_w, n_c)$值愈小)，獲得的調整幅度愈大，愈遠離則幅度愈

小；而『獲勝者』是調整幅度最大的節點。而且隨著訓練次數增加，調適的節點數量以及調適幅度都愈來愈小。因此，可以保證在經過多次的訓練之後，所產生的SOM會收斂。

3.2 資訊視覺化成效的評估

利用SOM技術進行資訊視覺化的目的是希望當資料被映射到圖形上時，它們的關係仍然可以盡量保持原先在高維特徵向量之間的關係，如此一來，可以從SOM產生的圖形認知原先的資料關係。也就是說，假設任何兩對術語(t_1, t_2)和(t_3, t_4)，每一個術語的特徵向量分別是f_1、f_2、f_3和f_4，如果在特徵向量上的距離關係是$d(f_1, f_2) > d(f_3, f_4)$。在經過術語資訊視覺化的過程後，我們希望當術語映射到節點n_1、n_2、n_3和n_4時，可以發現n_1、n_2、n_3和n_4在圖形的位置上，其歐幾里得距離具有$d(n_1, n_2) > d(n_3, n_4)$的關係。

所以，在比較應用SOM進行資訊視覺化的成效時，可以先計算出每一對術語在特徵向量的距離，在將術語映射到圖形後，再以所映射的節點計算術語在圖形上的距離，最後再計算這兩種距離的相關係數(correlation coefficients)，做爲資訊視覺化成效的評估標準，相關係數較小，表示SOM的結果較不理想；相關係數愈大，則表示SOM所產生的圖形保留愈多原先在高維特徵向量上的關係，可以從圖形上認知術語的叢集以及分離的關係，進而探索研究主題彼此之間的關係。

4 結果與討論

本論文以第一屆(1988)到第十四屆(2001) ROCLING研討會的235篇論文資料做爲分析計算語言學主題的素材，從這些論文的題名、摘要及參考文獻的題名中，抽取重要的術語，並將術語的關係視覺化。進行術語抽取時，本論文字串出現總次數的閾值設定爲20次，平均頻次和標準差的總和設爲2.5，前後接字的複雜度則設爲0.5，結果共得到229個術語。

接著將所抽取出來的229個術語，利用LSA技術估算彼此間的共現關係，建立各個術語的特徵向量。最後以術語的特徵向量進行SOM訓練，在本研究中，我們以20×20個節點進行實驗，測試訓練次數以及第3節式(4)的參數α之影響結果。在實驗中，參數α分別設定爲250、150、 50和25，每一個不同的α值，進行三次試驗，記錄訓練次數0 (初始)、10、50、100與200等各次的相關係數。取三次試驗中第200次訓練獲得較佳結果的試驗，也就是τ=200時相關係數最大者，進行比較。實驗的結果所產生的相關係數，如表1各欄所示。

表1 以自組織映射圖進行術語資訊視覺化的實驗結果

訓練次數τ	α=250	α=150	α=50	α=25
0	0.07	0.06	0.07	0.08
10	0.54	0.52	0.44	0.24
50	0.36	0.52	0.44	0.34
100	0.30	0.49	0.42	0.32
200	0.29	0.50	0.41	0.32

從表1的結果，我們可以看到幾個現象。(1) 初始的時候，映射圖仍未組織化，術語映射到圖上的各個節點上，其距離與特徵向量的距離無關，因此，相關係數不高，各欄均在0.06至0.08之間，顯示此時除了少數的相關術語映射到相同的節點上，大多數的術語的相關程度未能映射到圖形中。(2) 經過幾次訓練之後，映射圖上節點的特徵向量已經依照某種規則排列，此時的實驗結果獲得較大的相關係數，顯示若干相關的術語已經被映射到鄰近的節點中，比方說，以α=150一組的數據爲例，在訓練次數超過10次之後，相關係數約爲0.50。(3) 各欄的資料也表示，訓練次數相當大時，本研究提出的SOM技術可以收斂。(4) 如第3節中所提到參數α可以控制調適的節點數量，α值愈大，調適的節點數量愈多。從實驗中，我們發現α值過大，在訓練的過程中較不穩定；但較小的α值，卻很容易收斂到較爲次佳的結果。在本研究的實驗中，以α值爲150所得到的結果，較令人滿意。(5) 然而必須加以說明的是在SOM的訓練模式中，是以輸入的特徵向量對映射圖進行組織化，並不是對資訊視覺化的評估條件進行最佳化。因此，相關係數並不會呈現單調遞減的情形。而且，相關係數雖然可以提供客觀的評估標準，然而所得到的結果還需要進一步呈現來加以詮釋，才能看出SOM技術運用在術語資訊視覺化的成效。

因此，除了以相關係數來衡量資訊視覺化的成效之外，最爲重要的仍是經實際產生的映射圖所表達的訊息，我們將上面實驗中所得到較佳的結果之一，α值爲150、訓練次數50次所得到的映射圖，呈現在

圖1中。從圖1中，我們可以發現大多數相關的術語都被映射到同一節點或是鄰近的節點上，比方說。在映射圖下方，所包括的術語大多與語言學研究相關，如最左邊的 “syntax”、“functional”、“syntactic” 、“semantic” 、“lexical” 、“semantics” 、“lexicon” 以及“verb”。以及較右邊的“剖析”、“名詞”、“結構” 、“語法” 、“動詞” 、“詞類” 、“語意” 以及“詞彙”。又如在橫軸的16，縱軸10到12的地方可以發現這裡的術語都與語言模型的研究相關，如“bigram”、“language model”、“language modeling” 、“language models” 、“clustering” 、“class based”以及“n gram”。因此，在映射圖上可以發現主題相關的術語會形成叢集，我們可以依據圖1的相關術語分布情形，將幾個較大主題叢集表示成圖2。

除了相關的術語會映射在相近的節點上，從圖1與圖2也可以顯示在映射圖上距離很接近的主題具有相關性，比方說，『機器翻譯』(machine translation)相當接近於『剖析器與文法規則』(parser and grammars)與『語法與語意』(syntax and semantics)的研究，表示語法、語意、文法規則以及剖析器經常應用在機器翻譯的研究。『語音處理』中各個主題，包括『語音合成』(speech synthesis)、『語音辨認』(speech recognition)、『語言模型』(language models)等主題，彼此間也很接近。另外，映射圖上方的『斷詞』(word segmentation)、『未知詞偵測』(unknown word detection)與『詞類標示』(part-of-speech tagging)等相鄰近的情形，可以推測這些主題之間有相關性。圖形上『資訊檢索』(information retrieval)相關的主題，除了『斷詞』以及『語言模型』之外，還有『摘要』(summarization)。整體的圖形看來，偏左偏下的部份與語言學研究相關，而右上則是各種的技術應用與系統製作的研究，如『資訊檢索』和『語音處理』等各種主題便在圖形的右方。

然而，由於術語的數目相當龐大，特徵向量的維度也相當高，事實上，也有若干的術語映射結果並不理想，比方說，的“pat”與“tree”等術語所表示的PAT-tree是資訊檢索中重要而常用的技術[18]，但在這個映射圖上並沒有和位於橫軸19，縱軸17處的『資訊檢索』主題相鄰。此外，整個圖形中最明顯的現象是中英文同義或相關的術語雖然在圖形上它們的位置已經相當接近，但仍然可以認爲是分離。比方說，圖1中分布在圖形橫軸12到18，縱軸8到10處的三個同義的術語，“語音辨認”、“語音辨識”和“speech recognition”。這個現象表示即便我們利用參考文獻的題名做爲輸入資料以及LSA來進行補償，但中英文的資料仍然有區別，在論文資料中缺乏共現關係，使得中英文同義或相關的術語在圖形上相近但無法映射到同一節點上。

5 結論

本論文的研究利用自組織映射圖(SOM)技術將計算語言學相關術語對應到二維圖形，使得術語之間的關係可以在映射圖中加以呈現，提供使用者做爲資訊檢索以及了解研究領域的重要主題的輔助工具。在本論文中，我們所探討的問題有(1)發展SOM技術應用到術語資訊視覺化的方法，(2)評估SOM技術應用到術語資訊視覺化的成效，(3)利用研究結果分析計算語言學中重要的研究主題與主題之間的關係。在SOM技術的應用中，本研究首先從論文資料中利用字串出現的統計訊息以及經驗法則，抽取重要的術語。接著以術語之間的共現關係做爲基礎，建立每一個術語的特徵向量，以特徵向量之間的距離表示術語的相關性，愈相關的術語，特徵向量間的距離愈小。再以術語特徵向量做爲輸入資料，進行SOM訓練並將術語映射到圖形上，使得特徵向量距離相近的術語映射到同一節點或鄰近的節點上。如此一來，利用術語在映射圖上的分布情形，便可以輔助使用者認知研究主題之間的關係。對於這項技術的成效評估，我們建議將特徵向量的距離與節點位置的距離進行相關係數的計算，以所得到的相關係數大小做爲成效評估的標準。最後，對於計算語言學領域，以ROCLING論文集的論文資料做爲研究對象，進行術語資訊視覺化的實驗。在經過若干次的訓練之後，映射圖逐漸組織化。因此，術語特徵向量的距離與所映射節點位置的距離之相關係數增加。並且從實際產生的圖形中可以觀察出，大多數相關的術語都可以映射到相鄰近的節點上，在映射圖上所形成的叢集與計算語言學的主題相關，而這些叢集在圖形上的位置也可以確實地表現計算語言學主題之間的關係。值得一提的是，本研究所提出的方法並不會因爲論文資料數量增多，而增加時間與記憶體等計算資源的需求。由於這方法需要較多計算資源的階段是在SOM的訓練過程。而我們以術語的特徵向量做爲SOM的訓練資料，在特定領域中，術語的數目極爲有限，其數量並不會隨論文數目增多而快速成長，而且可以在術語抽取的階段，藉由參數的設定，只選取出現次數較多並較重要的術語，所以可以控制所需的計算資源，因此這個方法具有可升級性(scalability)。這些都顯示了SOM技術應用到術語資訊視覺化的可行性。

在進一步研究的建議上，除了進一步了解SOM在術語視覺化的能力與極限之外，比方說在產生映射圖之後，可以進一步自動將節點再度叢集、歸類，使得使用者更能解讀領域內的主題與趨勢。此外，階層式自組織映射圖等更先進的映射圖型態與技術可以表現出術語的概念階層(conceptual hierarchy)，將可以提供更有效的資訊組織工具。而如何應用本研究發展出來的成效評估方法，使得SOM更有效率、結果

更有用將也是發展的目標之一。再者，可以利用術語與論文之間的關係，產生論文的特徵向量，將論文映射到節點上，根據論文發表的年代，觀察研究主題的發展趨勢，對於了解研究領域的知識結構將有幫助。

參考文獻

[1] 林頌堅, “基於自然語言處理技術的研究主題抽取與分析,” Proceedings of ROCLING XV, pp. 231-256.
[2] P. Srinivasan, “Thesaurus Construction,” Information Retrieval—Data Structures & Algorithms, edited by W. B. Frakes and R. A. Baeza-Yates, Prentice-Hall, Inc., pp. 161-218.
[3] H. Chen, T. Yim, D. Fye, and B. Schatz, “Automatic Thesaurus Generation for an Electronic Community System,” Journal of the American Society of Information Science, Vol. 46, No. 3, pp. 175-193.
[4] Y-H. Tseng, “Automatic Thesaurus Generation for Chinese Documents,” Journal of the American Society for Information Science and Technology, Vol. 53, No. 13, pp.1130-1138.
[5] S. K. Card, J. D. Mackinlay, and B. Shneideman “1 Information Visualization,” Readings in Information Visualization— Using Vision to Think, Morgan Kaufmann, pp. 1-34.
[6] S. Huang, M. O. Ward, and E. A. Rundensteiner, Exploration of dimensionality reduction for text visualization. Technical Report TR-03-14, Worcester Polytechnic Institute, Computer Science Department, 2003.
[7] T. K. Landauer, D. Laham, and M. Derr, “From Paragraph to Graph: Latent Semantic Analysis for Information Visualization,” Proceedings of the National Academy of Science of the USA, Vol. 101, pp. 5214-5219.
[8] A. Flexer, “On the Use of Self-organizing Maps for Clustering and Visualization,” Intelligent Data Analysis, Vol. 5, pp. 373-384.
[9] X. Lin, “Visualization for the Document Space,” Proceedings of IEEE Visualization 1992, pp. 274-281.
[10] T. Kohonen, Self Organizing Maps, Springer Verlag.
[11] X. Lin, D. Soergel, and G. Marchionini, “A Self-organizing Semantic Map for Information,” Proceedings of SIGIR 1991, pp. 262-269.
[12] D. Merkl, “Exploration of Text Collections with Hierarchical Feature Maps,” Proceedings of SIGIR 1997, pp. 186-195.
[13] S. Wermter and C. Hung, “Selforganizing Classification on the Reuters News Corpus,” Proceedings of COLING 2002.
[14] T. Kohonen, S. Kaski, K. Lagus, and T. Honkela, “Very Large Two-Level SOM for the Browsing of Newsgroups,” Proceedings of ICANN 1996, pp. 269-274.
[15] H. Ritter and T. Kohonen, “Self-organizing Semantic Maps,” Biological Cybernetics, 61, pp. 241-254.
[16] Q. Ma, M. Zhang, M. Murata, M. Zhou, and H. Isahara, “Self-organizing Chinese and Japanese Semantic Maps,” Proceedings of COLING 2002.
[17] K. Kageura and B. Umino, “Methods of Automatic Term Recognition-A Review,” Terminology, Vol. 3, No. 2, pp. 259-289.
[18] G. H. Gonnet, R. A. Baeza-Yates, and T. Snider, “New Indices for Text: PAT Trees and PAT Arrays,” Information Retrieval—Data Structures & Algorithms, edited by William B. Frakes and Ricardo Baeza-Yates, Prentice-Hall, Inc., pp. 66-101.
[19] S. Deerwester, S. T. Dumais, G. W. Furnas, Thomas K. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis,” Journal of the American Society for Information Science, 41(6), pp. 391-407.

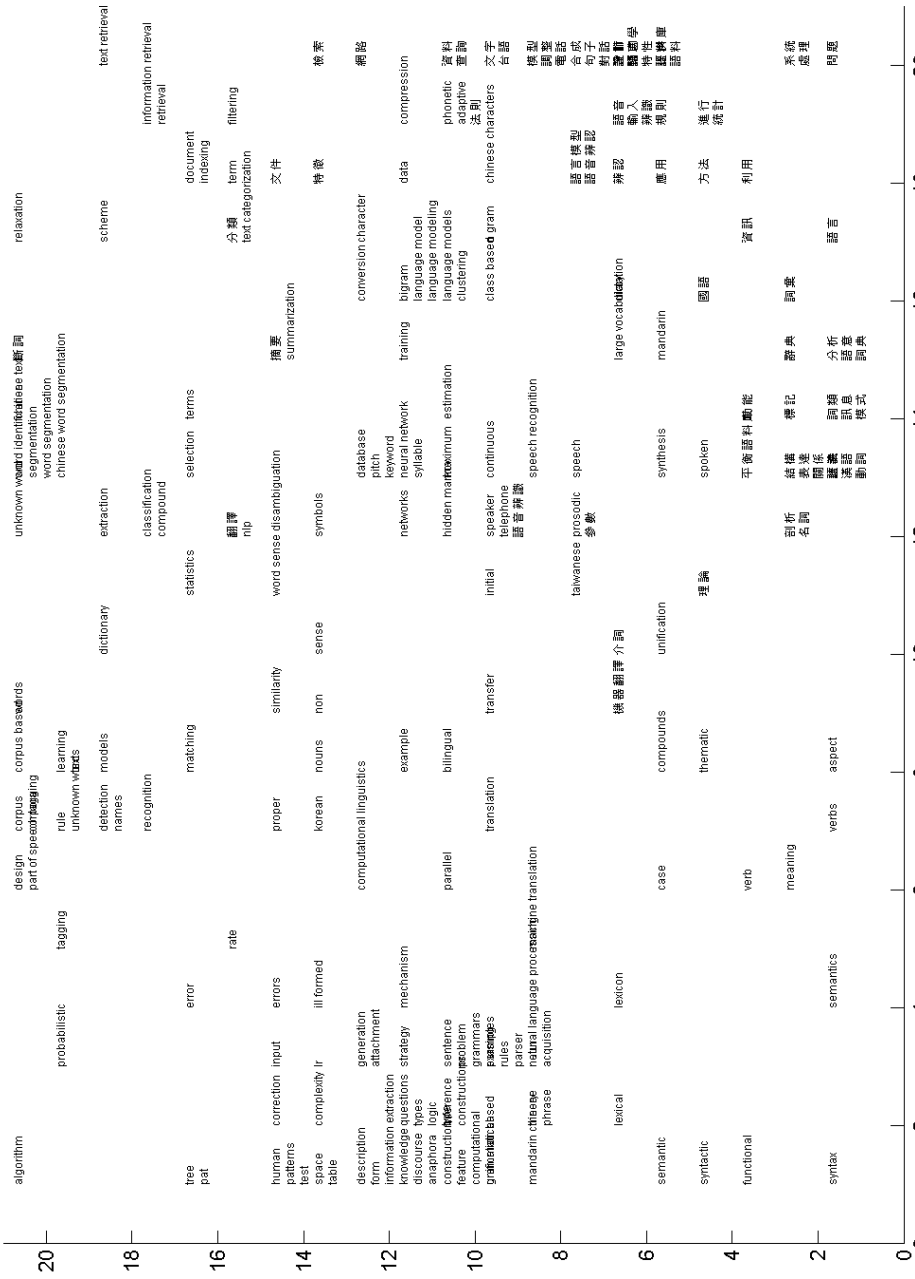


圖1 利用SOM技術對計算語言學術語進行資訊視覺化的結果

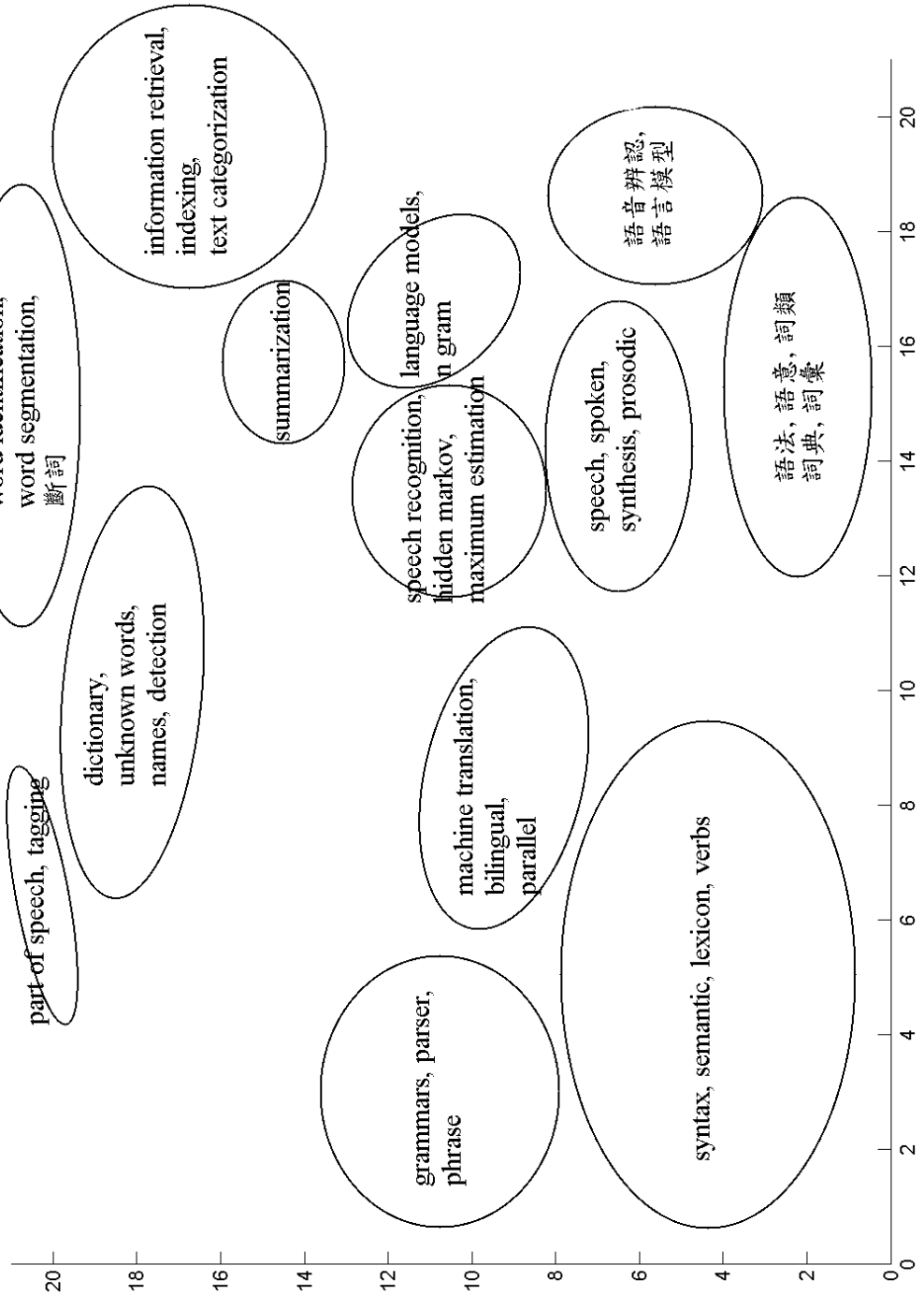


圖 2　計算語言學術語主題叢集的分布情形

Applying Meaningful Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem

Jia-Lin Tsai, Tien-Jien Chiang and Wen-Lian Hsu
Institute of Information Science, Academia Sinica,
Nankang, Taipei, Taiwan, R.O.C.
{tsaijl,tmjiang, hsu}@iis.sinica.edu.tw

Abstract. Syllable-to-word (STW) conversion is a frequently used Chinese input method that is fundamental to syllable/speech understanding. The two major problems with STW conversion are the segmentation of syllable input and the ambiguities caused by homonyms. This paper describes a meaningful word-pair (MWP) identifier that can be used to resolve homonym/segmentation ambiguities and perform STW conversion effectively for Chinese language texts. It is designed as a support system with Chinese input systems. In this paper, five types of meaningful word-pairs are investigated, namely: noun-verb (NV), noun-noun (NN), verb-verb (VV), adjective-noun (AN) and adverb-verb (DV). The pre-collected datasets of meaningful word-pairs are based on our previous work *auto-generation of NVEF knowledge in Chinese* (*AUTO-NVEF*) [30, 32], where NVEF stands for noun-verb event frame.

The main purpose of this study is to illustrate that a hybrid approach of combining statistical language modeling (SLM) with contextual information, such as meaningful word-pairs, is effective for improving syllable-to-word systems and is important for syllable/speech understanding. Our experiments show the following: (1) the MWP identifier achieves ***tonal*** (syllables with four tones) and ***toneless*** (syllables without four tones) STW accuracies of 98.69% and 90.7%, respectively, among the identified word-pairs for the test syllables; (2) by STW error analysis, we find that the major critical problem of tonal STW systems is the failure of homonym disambiguation (52%), while that of toneless STW systems is inadequate syllable segmentation (48%); (3) by applying the MWP identifier, together with the Microsoft input method editor (MSIME 2003) and an optimized bigram model (BiGram), the tonal and toneless STW improvements of the two STW systems are 25.25%/21.82% and 12.87%/15.62%, respectively.

Keywords: syllable-to-word, contextual information, top-down identifier, n-gram model.

1. Introduction

More than 100 Chinese input methods have been developed in the past [1, 17, 12, 5, 18, 10, 19, 16, 4, 28, 11, 20]. Their underlying approaches can be classified into four types:

(a) Optical character recognition (OCR) based [5],
(b) On-line handwriting based [19],
(c) Speech based [4, 10], and
(d) Keyboard based, such as syllabic-input-to-character [27, 16, 2, 14, 15, 22]; arbitrary codes based [8]; and structure scheme based [11]. The major goal of these syllable input systems is to achieve high STW accuracy, but syllable understanding is rarely considered [16].

Currently, the most popular method for Chinese input is syllable based (or phonetic/pinyin based), because Chinese people are taught to write the corresponding phonetic/pinyin syllable of each Chinese character in primary school. Basically, each Chinese character corresponds to at least one syllable. Although there are more than 13,000 distinct Chinese characters (of which 5,400 are commonly used), there are only 1,300 distinct syllables. The homonym (homophone) problem is, therefore, quite severe when using a Chinese phonetic input method [5]. As per [26], each Chinese syllable can be mapped from 3 to over 100 Chinese characters, with the average number of characters per syllable being 17. Therefore, *homonym disambiguation* is a critical problem that requires the development of an effective syllable-to-word (STW) conversion system for Chinese. A comparable problem for STW conversion in English is word-sense disambiguation (WSD).

There are two conventional approaches for STW conversion: the *linguistic approach* based on syntax parsing, semantic template matching and contextual information [18, 22, 16, 28, 15]; and the *statistical approach* based on the *n*-gram model where *n* is usually 2 or 3 [12, 10, 11, 20, 21, 13, 27]. Although the linguistic approach requires considerable effort in designing effective syntax rules, semantic templates or contextual information, it is more user-friendly than the statistical approach (i.e. it is easier to understand why such a system makes a mistake) [16]. On the other hand, the statistical language model (SLM) used in the statistical approach requires less effort and has been widely adopted in commercial systems. However, the power of the statistical approach depends on the training

corpus [10] and the SLM pays little attention to syllable understanding [16]. Following the work of [12, 18, 10, 28, 11, 15, 13], a better approach to STW conversion is to integrate both linguistic knowledge (such as contextual information) and statistical approaches (such as an n-gram model). We believe that our research proves the efficacy of such an integrated approach.

According to previous studies [5, 28, 11, 20, 9], besides homonyms, correct *syllable-word segmentation* is another crucial problem of STW conversion. Incorrect syllable-word segmentation directly influences the conversion rate of STW. For example, consider the syllable sequence "yi1 du4 ji4 yu2 zhong1 guo2 de5 niang4 jiu3 ji4 shu4" of the sentence "一度(once)覬覦(covet)中國(China)的(of)釀酒(making-wine)技術(technique)." According to the CKIP lexicon [6], the two possible syllable-word segmentations are:

(F) "yi1/du4ji4/yu2/zhong1guo2/de5/niang4jiu3/ji4shu4"; and
(B) "yi1/du4/ji4yu2/zhong1guo2/de5/niang4jiu3/ji4shu4."

(We use the forward (F) and the backward (B) longest syllable-word first strategies [3], and "/" to indicate a syllable-word boundary).

Among the above syllable-word segmentations, there is an ambiguous syllable-word section: /du4ji4/yu2/ (/{妒忌}/{于,圩,余,於,盂,俞,娛,魚,愉,渝,腴,萸,隅,畬,愚,榆,瑜,虞,逾,漁,諛,餘,輿}/); and /du4/ji4yu2/ (/{妒,杜,肚,度,渡,鍍,蠹}/{覬覦,鯽魚}/), respectively. In this case, if the system has the contextual information that the pairs "技術(technique)-覬覦(covet)" and "一度(once)-覬覦(covet)" are, respectively, meaningful noun-verb (NV) and adverb-verb (DV) word-pairs, then the ambiguous syllable-word section can be effectively resolved and the word-pairs "技術(technique)-覬覦(covet)" and "一度(once)-覬覦(covet)" of this syllable sequence can be correctly identified.

For the above case, if we look at the Sinica corpus [6], the bigram frequencies of "覬覦(covet)-中國(China)" and "於(at)-中國(China)" are 0 and 24, respectively. Therefore, by using a bigram model trained with the Sinica corpus, the forward syllable-word segmentation would conclude that the following word segmentation /於/中國/, will be incorrect. In fact, if we use Microsoft Input Method Editor 2003 for Traditional Chinese (a trigram like STW product), the syllables of the above example will be converted to "一度(once)繼(continue)於(to)中國(China)的(of)釀酒(making-wine)技術(technique)." It is widely recognized that unseen event ("覬覦-中國") and over-weighting ("於-中國") are two major problems of SLM systems [10, 11]. Practical SLM is either a bigram or a trigram model. As the above case shows, the meaningful word-pairs (or contextual information) "技術(technique)-覬覦(covet)" and "一度(once)-覬覦(covet)" can be used to overcome both the unseen event and over-weighting problems of SLM-based STW systems. In [29], we showed that the knowledge of noun-verb event frame (NVEF) sense-pairs and their corresponding NVEF word-pairs (NVEF knowledge) are useful for effectively resolving word sense ambiguity with an accuracy of 93.7%. In [28], we showed that a NVEF word-pair identifier with pre-collected NVEF knowledge can be used to obtain a tonal (syllables with four tones) STW accuracy of more than 99% for the NVEF related portion in Chinese.

The objective of this study is to illustrate the effectivness of meaningful noun-verb (NV), noun-noun (NN), verb-verb (VV), adjective-noun (AN) and adverb-verb (DV) word-pairs for solving Chinese STW conversion problems. We conduct STW experiments to show that the ***tonal*** and ***toneless*** STW accuracies of conventional SLM models and the commercial input products can be improved by using a meaningful word-pair identifier without a tuning process. In this paper, we use ***tonal*** to indicate the syllables input with four tones, such as "niang4(釀) jiu3(酒) ji4(技) shu4(術)," and ***toneless*** to indicate the syllables input without four tones, such as "niang(釀) jiu(酒) ji(技) shu(術)."

The remainder of this paper is arranged as follows. In Section 2, we propose the method for *auto-generating the meaningful word-pairs* in Chinese based on [30, 32], and a *meaningful word-pair identifier* to resolve homonym/segmentation ambiguities of STW conversion in Chinese. The meaningful word-pair identifier is based on pre-collected datasets of meaningful word-pairs. In Section 3, we present our STW experiment results and analysis. Finally, in Section 4, we give our conclusions and suggest some future research directions.

2. Development of the Meaningful Word-Pair Identifier

To develop the meaningful word-pair (MWP) identifier, we selected Hownet [7] as our system's dictionary because it provides knowledge of Chinese words, word senses and part-of-speeches (POS). The Hownet dictionary used in this study contains 58,541 Chinese words, among which there are 33,264 nouns, 16,723 verbs, 8,872 adjectives and 882 adverbs.

In this system's dictionary, the syllable-word for each word is obtained by using the inverse phoneme-to-character system presented in [15], while the word frequencies are computed according to a fixed-size United Daily News (UDN) 2001 corpus. The latter is a collection of 4,539,624 Chinese sentences extracted from

articles on the United Daily News Website [25] from January 17, 2001 to December 30, 2001. Table 1 shows the statistics of the number of articles per article class in this UDN 2001 corpus.

Table 1. The number of articles per article class in the training corpus.

article class	大陸 China	地方 Local	社會 Society	股市 Stock	政治 Politics	科技 Science	旅遊 Travel
# of articles	90	26,843	136	19,699	133	5,870	6,183
article class	消費 Consumption	財經 Financial	國際 World	運動 Sport	影視 Entertainment	醫藥 Health	藝文 Arts
# of articles	12498	23,563	7,404	12,404	18,674	5,653	9,989

2.1 Generating the Meaningful Word-Pair

In [32], we propose an *AUTO-NVEF* system to auto-generate NVEF knowledge from in Chinese. It extracts NVEF knowledge from Chinese sentences by four major processes: (1) Segmentation checking; (2) Initial Part-of-Speech (IPOS) sequence generation; (3) NV knowledge generation; and (4) NVEF knowledge auto-confirmation. The details of the four processes can be found in [32]. Take the Chinese sentence "音樂會(concert)/現場(locale)/湧入(enter)/許多(many)/觀眾(audience members)" as an example. For this sentence, *AUTO-NVEF* will generate two collections of NVEF knowledge: 現場(locale)-湧入(enter) and 觀眾(audience members)-湧入(enter). In [32], we reported that *AUTO-NVEF* achieved 98.52% accuracy for news and 96.41% for specific text types, which included research reports, classical literature and modern literature. In addition, it automatically discovered over 400,000 NVEF word-pairs in the UDN 2001 corpus.

Using *AUTO-NVEF* as the base, we extended the system into a meaningful word-pair (MWP) generation called AUTO-MWP. The steps of AUTO-MWP are:

Step 1. Use *AUTO-NVEF* to generate NVEF word-pairs for the given Chinese sentence. *AUTO-NVEF* adopts a *forward=backward* maximum matching technique to perform word segmentation and a *bigram-like* model to perform POS tagging [32]. If no NVEF word-pairs are generated, go to Step 3.

Step 2. According to the generated NVEF word-pairs and the word-segmented sentence with POS tagging from Step 1, the auto-generation methods of meaningful NN, VV, AN and DV word-pairs are:

(1) *Generation of NN word-pair*. When the number of generated NVEF word-pairs is greater than 1, this sub-process will be triggered. If the nouns of two generated NVEF word-pairs share the same verb, the two nouns will be designated as a meaningful NN word-pair. Take the generated NVEF word-pairs of 現場(locale)-湧入(enter) and 觀眾(audience members)-湧入(enter) for the sentence "音樂會(concert)現場(locale)湧入(enter)許多(many)觀眾(audience members)" as examples. The noun 現場(locale) and the noun 觀眾(audience members) are designated as a NN word-pair because the two nouns share the same verb 湧入(enter) in this sentence.

(2) *Generation of VV word-pair*. When the number of generated NVEF word-pairs is greater than 1, this sub-process will be triggered. If the verbs of two generated NVEF word-pairs share the same noun, the two verbs will be designated as a meaningful VV word-pair. Take the generated NVEF word-pairs 年底(the end of year)-預定(prearrange) and 年底(the end of year)-完成(complete) for the sentence "全部(whole)工程(construction)預定(prearrange)年底(the end of year)完成(complete)" as examples. The verb 預定(prearrange) and the verb 完成(complete) are designated as a VV word-pair because the two verbs share the same noun 年底(the end of year).

(3) *Generation of AN word-pair*. For each noun of a generated NVEF word-pair, if the word immediately to its left is an adjective, the noun and the adjective are designated as one AN word-pair. Take the generated NVEF word-pair 觀眾(audience members)-湧入(enter) for the word-segmented and POS-tagged sentence "音樂會(N)現場(N)湧入(V)許多(ADJ)觀眾(N)" as an example. Since the word immediately to the left of 觀眾(audience members) is an adjective 許多(many), the adjective 許多(many) and the noun 觀眾(audience members) are designated as a AN word-pair.

(4) *Generation of DV word-pair*. For each verb of a generated NVEF word-pair, if the word immediately to its left is an adverb, the verb and the adverb are designated as one DV word-pair. Take the generated NVEF word-pair 物價(price)-維持(maintain) for the word-segmented and POS-tagged sentence "物價(N)大抵(ADV)維持(V)平穩(ADJ)" as an example. Since the word immediately to the left of 維持(maintain) is an adverb 大抵(ordinarily), the adverb 大抵(ordinarily) and the verb 維持(maintain) are designated as a DV word-pair.

Step 3. Stop.

Table 2 shows the number of generated NV, NN, VV, AN and DV word-pairs obtained by applying AUTO-MWP to the *UDN* 2001 corpus. The frequencies of all the generated meaningful word-pairs were computed by the *UDN* 2001 corpus. Note that the frequency of a meaningful word-pair is the number of sentences that contain the word-pair *with the same word-pair order* in the *UDN* 2001 corpus. Table 3 shows fifteen randomly selected NV, NN, VV, AN and DV word-pairs and their corresponding frequencies in the generated MWP datasets for the *UDN* 2001 corpus.

Table 2. The number of generated NV, NN, VV, AN and DV word-pairs obtained by applying AUTO-MWP to the *UDN* 2001 corpus.

NV	NN	VV	AN	DV	Total
430,698	533,780	220,022	138,055	111,879	1,434,434

Table 3. Fifteen randomly selected examples of meaningful NV, NN, VV, AN and DV word-pairs and their corresponding frequencies from the generated MWP datasets for the *UDN* 2001 corpus.

NV	NN	VV	AN	DV
大學-附設/118	全國-比賽/83	開放-投資/541	對外-交通/206	即將-舉行/188
人選-決定/35	偶像-明星/103	接受-訪問/1483	資深-釣友/103	幾乎-都是/390
路線-是/96	市府-人員/107	擔心-造成/124	最高-榮譽/129	再度-成為/144

2.2 Meaningful Word-Pair Identifier

We developed a NVEF word-pair identifier [28] for Chinese syllable-to-word (STW) and achieved a tonal STW accuracy of more than 99% on the NVEF related portion. This NVEF word-pair identifier is based on the techniques of *longest syllabic NVEF-word-pair first* (LS-NVWF), *exclusion-word-list* (EWL) checking and pre-collected NVEF knowledge. By modifying the algorithm of this identifier in [28], we obtain our meaningful word-pair (MWP) identifier, (Figure 1). In Figure 1, the MWP data is a mixed collection of all auto-generated meaningful NV, NN, VV, AN and DV word-pairs. As shown in the figure, if the MWP identifier only uses one of the meaningful NV, NN, VV, AN or DV word-pair datasets, it will naturally become an MNV, MNN, MVV, MAN or MDV word-pair identifier.

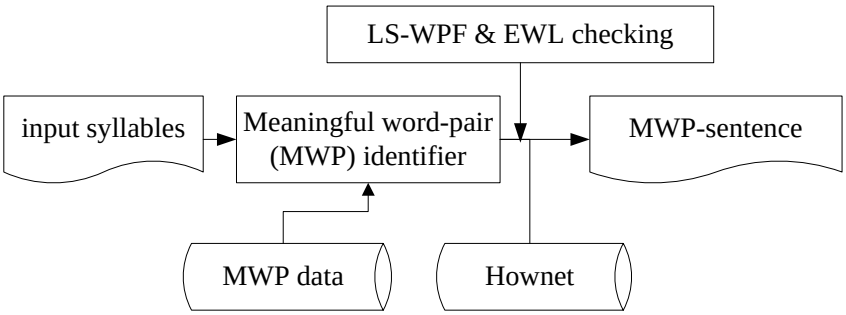


Figure 1. A system overview of the meaningful word-pair (MWP) identifier.

The algorithm of the MWP identifier is as follows:

Step 1. Input tonal (with four tones) or toneless (without four tones) syllables.

Step 2. Generate all possible word-pairs found in the input syllables. Exclude certain NV word-pairs based on EWL checking [28]. Appendix A lists all of the exclusion words used in this study. Note that our meaningful word-pairs include monosyllabic nouns/adjectives/adverbs and monosyllabic verbs, except "是(be)" and "有(has/have)" that are dropped in this Step.

Step 3. Word-pairs that match a meaningful word-pair in the generated MWP data are used as the initial MWP set for the input syllables. From the initial MWP set, select a key word-pair and its co-occurring word-pairs to be the final MWP set. Conflicts are resolved using the *longest syllabic word-pair first* (LS-WPF) strategy. If there are two or more word-pairs with the same condition, the system triggers the following processes.

(1) The word-pair with the greatest frequency (the number of sentences that contain the word-pair *with the same word-pair* order in the *UDN* 2001 corpus) is selected as the key word-pair. If there are two or more word-pairs with the same frequency, one of them is randomly selected as the key word-pair.

(2) The word-pairs that co-occur with the key word-pair in the *UDN* 2001 corpus are selected.

(3) The key and co-occurred word-pairs are then combined as the final MWP set for Step 4.

Step 4. Arrange all word-pairs of the final MWP set into a *MWP-sentence* as shown in Table 3. If no word-pairs can be identified from the input syllables, a null MWP-sentence is produced.

Table 3. An illustration of an MWP-sentence for the Chinese syllables "yi1 ge5 wen2 ming2 de5 shuai1 wei2 guo4 cheng2(一個[a]文明[civilization]的[of]衰微[decay]過程[process])." (The English words in parentheses are included for explanatory purposes only.)

Process	Results	Pair freq.
Step.1	yi1 ge5 wen2 ming2 de5 shuai1 wei2 guo4 cheng2 (一 個 文 明 的 衰 微 過 程)	
Step.2	The meaningful word-pairs found: 文明(wen2 ming2)-過程(guo4 cheng2)/NN pair 文明(wen2 ming2)-衰微(shuai1 wei2)/NV pair	 3 1
Step.3	The key meaningful word-pair: 文明(wen2 ming2)-過程(guo4 cheng2)/NN pair The co-occurred word-pair: 文明(wen2 ming2)-衰微(shuai1 wei2)/NV pair	
Step.4	MWP-sentence: yi1 ge5 文 明 de5 衰 微 過 程	

Table 3 is a step by step example that illustrates the four processes of our MWP identifier for the Chinese syllables "yi1 ge5 wen2 ming2 de5 shuai1 wei2 guo4 cheng2(一個[a]文明[civilization]的[of]衰微[decay]過程[process])." When we used MSIME 2003 to convert the same syllables, the output was "一個(one)聞名(famous)的(of)衰微(decay)過程(process)." Obviously, the over-weighted bigram "聞名-的(wen2 ming2-de5)" causes an STW error in MISIME 2003, which uses a statistical language model (SLM) with a trigram-like Chinese input product [24]. If we use the MWP-sentence shown in Step 4 to directly replace the corresponding characters of the MSIME 2003 output in this example, the error converted word "聞名(famous)", caused by the over-weighting of MSIME 2003, becomes the correct word "文明(civilization)."

3. The STW experiment

To evaluate the STW performance of our MWP identifier, we define the STW accuracy, STW improvement, identified character ratio (ICR) by the following equations:

STW accuracy =
# of correct characters / # of total characters. (1)

STW improvement (STW error reduction rate) =
(accuracy of STW system with MWP – accuracy of STW system)) / (1 – accuracy of STW system). (2)

Identified character ratio (ICR) =
# of characters of identified MWPs / # of total characters in testing sentences. (3)

3.1 Closed Test Set and Open Test Set

We use the inverse translator of the phoneme-to-character system in [15] to convert a test sentence into a syllable sequence. We then apply our MWP identifier to convert this syllable sequence back to characters and calculate its STW accuracy and identified character ratio by Equations (1) and (2). All test sentences are composed of a string of Chinese characters.

In following experiments, the training/testing corpus, closed/open test sets and the collection of MWPs were:

Training corpus: We used the *UDN* 2001 corpus mentioned in Section 2 as our training corpus. All knowledge of word frequencies, meaningful word-pairs, MWP frequencies was auto-generated and computed by this corpus.

Testing corpus: The *UDN* 2002 corpus was selected as our testing corpus. It is a collection of 3,321,504 Chinese sentences that were extracted from articles on the *United Daily News* Website [25] from January 1, 2002 to December 30, 2002.

Closed test set: 10,000 sentences were randomly selected from the *UDN* 2001 corpus as the *closed test set*. The

{minimum, maximum, and mean} of characters per sentence for the closed test set were {4, 37, and 12}.

Open test set: 10,000 sentences were randomly selected from the *UDN* 2002 corpus as the *open test set*. At this point, we checked that the selected open test sentences were not in the closed test set as well. The {minimum, maximum, and mean} of characters per sentence for the open test set were {4, 43, and 13.7}.

Meaningful word-pair data: By applying our AUTO-MWP on the *UDN* 2001 corpus, we created 430,698 NV, 533,780 NN, 220,022 VV, 138,055 AN and 111,879 DV word-pairs as the MWP testing data.

In this study, we conducted the STW experiment in a progressive manner. The results and analysis of the experiment are described in Sub-sections 3.2, 3.3 and 3.4. Appendix B presents two STW results that were obtained from the experiment.

3.2 STW Experiment for the MWP Identifier

The purpose of this experiment is to demonstrate the tonal and toneless STW accuracies by using the MWP identifier with the generated meaningful NV, NN, VV, AN, DV and (NV+NN+VV+AN+DV) datasets, respectively. Note that the symbol (NV+NN+VV+AN+DV) stands for a mixed collection of all auto-generated meaningful NV, NN, VV, AN and DV word-pairs.

Table 4a. The results of the tonal STW experiment for the MWP identifier with NV, NN, VV, AN, DV and (NV+NN+VV+AN+DV) word-pairs.

	Closed	Open	Average (identified character ratio)
NV	99.08%	98.70%	98.90% (21.69%)
NN	98.54%	98.30%	98.43% (34.56%)
VV	98.25%	97.25%	97.81% (14.64%)
AN	97.41%	96.83%	97.14% (10.07%)
DV	98.07%	97.45%	97.80% (9.46%)
(NV+NN+VV+AN+DV)	98.69%	98.20%	98.46% (46.67%)

Table 4b. The results of the toneless STW experiment for the MWP identifier with NV, NN, VV, AN, DV and (NV+NN+VV+AN+DV) word-pairs.

	Closed	Open	Average (identified character ratio)
NV	91.53%	90.03%	91.01% (24.46%)
NN	91.41%	89.82%	90.92% (27.79%)
VV	88.80%	86.96%	87.67% (12.20%)
AN	88.00%	86.04%	86.89% (10.67%)
DV	88.98%	86.51%	88.03% (10.03%)
(NV+NN+VV+AN+DV)	91.33%	89.99%	90.70% (38.63%)

From Tables 4a and 4b, the average tonal and toneless STW accuracies of the MWP identifier with the MWP (NV+NN+VV+AN+DV) data for the closed and open test sets are 98.46% and 90.70%, respectively. Between the closed and the open test sets, the differences of the tonal and toneless STW accuracies of the MWP identifier with the (NV+NN+VV+AN+DV) data are 0.49% and 1.34%, respectively. These results strongly support our belief that meaningful word-pairs can be used as application independent knowledge to effectively convert Chinese STW on the MWP-related portion.

3.3 A Commercial IME System and A Bigram Model with MWP Identifier

We selected Microsoft Input Method Editor 2003 for Traditional Chinese (MSIME 2003) as our experimental commercial IME system. In addition, a bigram model called BiGram was developed. The BiGram STW system is a bigram model using Lidstone's law [23], as well as forward and backward longest syllable-word first strategies. The system dictionary of the BiGram is comprised of CKIP lexicon and those unknown words found automatically in the *UDN* 2001 corpus by a Chinese word auto-confirmation (CWAC) system [31]. All the bigram probabilities were calculated by the *UDN* 2001 corpus.

MSIME 2003, which uses a statistical trigram-like model [24], is one of the most widely available input methods. Table 5a compares the results of MSIME 2003, and MSIME 2003 with the MWP identifier on the closed and open test sentences. Table 5b compares the results of BiGram, and BiGram with the MWP identifier on the

closed and open test sentences. In this experiment, the STW output of MSIME with the MWP identifier, or BiGram with the MWP identifier, was collected by directly replacing the identified meaningful word-pairs from the corresponding STW output of MSIME or BiGram. From Table 5a, the tonal STW improvements of MSIME and BiGram by using the MWP identifier are 25.25% and 12.87%, respectively. Meanwhile, from Table 5b, the toneless STW improvements of MSIME and BiGram by using the MWP identifier are 21.82% and 15.62%, respectively.

Table 5a. The results of tonal STW experiments for closed and open test sentences, using MSIME, BiGram, MSIME with MWP identifier and BiGram with MWP identifier.

Identified-word	MSIME [a]	BiGram [a]	MSIME + MWP [b]	BiGram + MWP [b]
MWP portion	96.87%	97.29%	-	-
Overall	95.05%	96.27%	25.25%	12.87%

[a] STW accuracies of the words identified by the Microsoft Input Method Editor (MSIME) 2003 and the BiGram
[b] STW improvements of the words identified by the MSIME 2003 with the MWP identifier and the BiGram with the MWP identifier

Table 5b. The results of toneless STW experiments for closed and open test sentences, using MSIME, BiGram, MSIME with MWP identifier and BiGram with MWP identifier.

Identified-word	MSIME [a]	BiGram [a]	MSIME + MWP [b]	BiGram + MWP [b]
MWP portion	89.40%	89.23%	-	-
Overall	86.94%	85.47%	21.82%	15.62%

[a] STW accuracies of the words identified by the Microsoft Input Method Editor (MSIME) 2003 and the BiGram
[b] STW improvements of the words identified by the MSIME 2003 with the MWP identifier and the BiGram with the MWP identifier

To sum up the results and observations of this experiment, we conclude that the MWP identifier can achieve better MWP-portion STW accuracy than the MSIME 2003 and BiGram STW systems. The results show that the MWP identifier can help both MSIME 2003 (trigram-like) and BiGram (bigram base) systems to increase their performances to achieve 96.30%/96.75% of tonal STW accuracies and 89.79%/87.74% of toneless STW accuracies, respectively. Furthermore, the results indicate that the meaningful word-pairs, or contextual information, can be used to effectively overcome the unseen event and over-weighting problems of SLM models in Chinese STW conversion.

3.4 Error Analysis of STW Conversion

We examine the Top 300 cases in the *tonal* and *toneless* STW conversion from the open testing results of BiGram with the MWP identifier and classify them according to the following three major types of error (see Table 6):

(1) ***Unknown words***: For any NLP system, unknown word extraction is one of the most difficult problems [31]. Sinc proper names are major types of unknown words, we classify the cases of unknown words into two sub-types and calculate their corresponding percentages, as shown in Table 6.
(2) ***Inadequate syllable segmentation***: When an error is caused by word overlapping, instead of an unknown word problem, we call it *inadequate syllable segmentation.*
(3) ***Homophones***: These are the remaining errors.

Table 6. Three major error types of tonal/toneless STW conversion.

Types	Sub-Types	Percentage within this type (%)	Examples	Overall Percentage (%)
Unknown word	Proper names	50 / 50	蘿拉、戴維絲、美邦	11.7 / 10.5
	Other cases	50 / 50	筍農、腋下	
Inadequate syllable segmentation			Tonal cases: 經/嘗試；經常/是、 這些/原油/是；這些/元/郵市、 Toneless cases: 意思/是；以/四史、 的/國家；德國/家	35.9 / 50.8
Homophone			Tonal cases: 需/須、心形/新型、美股/每股 Toneless cases: 主義/注意、雖是/隨時、提醒/體型	52.4 / 38.7

From Table 6, we make the following observations:

(1) **The percentages of unknown word errors for the tonal and toneless STW systems are similar**. Since the unknown word problem is not specifically a STW problem, it can be easily taken care of through manual editing or semi-automatic learning during input. In practice, therefore, the tonal and toneless STW accuracies could be raised to 98% and 91%, respectively. However, even though unknown words of the first error type have been incorporated in the system dictionary, they could still face the problems of *inadequate syllable segmentation* or *failed homophone disambiguation*.

(2) **The major error types of tonal and toneless STW systems are different**. To improve tonal STW systems, the major targets should be cases of failed homophone disambiguation. For toneless STW systems, on the other hand, cases of inadequate syllable segmentation should be the focus for improvement.

To sum up the above observations, the bottlenecks of the STW conversion lie in the second and third error types. To resolve these issues, we believe one possible approach is to extend the size of MWP data to increase the identified MWP character ratio. This is because our experiment results show that the MWP identifier can achieve better tonal and toneless STW accuracies than those of BiGram and MSIME 2003 on the MWP-related portion (see the examples given in Appendix B).

4. Conclusions and Directions for Future Research

In this paper, we have applied a MWP identifier to the Chinese STW conversion problem and obtained a high degree of STW accuracy on the MWP-related portion. All of the MWP data was generated fully automatically by using AUTO-MWP on the *UDN* 2001 corpus. The experiments on STW conversion in [28] and on WSD in [29], as well as the STW experiments in this study, demonstrate that meaningful word-pairs (i.e. contextual information) are key linguistic features of NLP/NLU systems.

We are encouraged by the fact that MWP knowledge can achieve tonal and toneless STW accuracies of 98.46% and 90.70%, respectively, for the MWP-related portion of the testing syllables. The MWP identifier can be easily integrated into existing STW conversion systems by identifying meaningful word-pairs in a post-processing step. Our experiment shows that, by applying the MWP identifier together with MSIME 2003 (a trigram-like model) and BiGram (an optimized bigram model), the tonal and toneless STW improvements are 25.25%/21.82% and 12.87%/15.62%, respectively.

Currently, our approach is quite basic when more than one MWP occurs in the same sentence (Step 3 in Section 2.2). Although there is room for improvement, we believe it would not produce a noticeable effect as far as the STW accuracy is concerned. However, this issue will become important as we apply the MWP knowledge to parsing or speech understanding.

The MWP-based approach has the potential to provide the following information for a given syllable sequence: (1) better word segmentation; and (2) MWP-sentence including the information of five types of MWPs. Such information will be useful for general NLP and NLU systems, especially for syllable/speech understanding and full/shallow parsing. According to our computations, the collection of MWP knowledge can cover approximately 50% of the characters in the *UDN* 2001 corpus.

We will continue to expand our collection of MWP knowledge to cover more characters in the *UDN* 2001 corpus. In other directions, we will try to improve our MWP-based STW conversion with other statistical language models, such as HMM, and extend it to other areas of NLP, especially Chinese shallow parsing and syllable/speech understanding.

5. Acknowledgements

This project was supported in part by *Chinese Multimedia Information Retrieval System* (中文多媒體資訊擷取) under an excellent Grant AS-91-TP-A09, Research Center for Humanities and Social Sciences, Academia Sinica,and National Science Council under a Center Excellence Grant NSC93-2752-E-001-001-PAE. We would like to thank Prof. Zhen-Dong Dong for providing us with the Hownet dictionary.

References

[1] Becker, J.D. Typing Chinese, Japanese, and Korean. *IEEE Computer* 18(1), pp. 27-34, 1985.

[2] Chang, J.S., S.D. Chern and C.D. Chen. Conversion of Phonemic-Input to Chinese Text Through Constraint Satisfaction. *Proceedings of ICCPOL'91*, pp. 30-36, 1991.

[3] Chen, C.G., K.J. Chen and L.S. Lee. A model for Lexical Analysis and Parsing of Chinese Sentences. *Pro-*

ceedings of 1986 International Conference on Chinese Computing, Singapore, pp. 33-40, 1986.
[4] Chen, B., H.M. Wang and L.S. Lee. Retrieval of broadcast news speech in Mandarin Chinese collected in Taiwan using syllable-level statistical characteristics. *Proceedings of the 2000 International Conference on Acoustics Speech and Signal Processing*, 2000.
[5] Chung, K.H. *Conversion of Chinese Phonetic Symbols to Character*s. M. Phil. thesis, Department of Computer Science, Hong Kong University of Science and Technology, Sept. 1993.
[6] CKIP. Technical Report no. 95-02. *the content and illustration of Sinica corpus of Academia Sinica*. Institute of Information Science, Academia Sinica, http://godel.iis.sinica.edu.tw/CKIP/r_content.html, 1995.
[7] Dong, Z. and Q. Dong. *Hownet*, http://www.keenage.com/, 1999.
[8] Fan, C. and P. Zini. Chinese Character Processing system based on character-root combination and graphics processing. *Document Manipulation and Typography, Proc. of the Int. Conf. on Electronic Publishing, Doc. Manipulation and Typography*, Nice (France), Cambridge University Press, 1988.
[9] Fong, L.A. and K.H. Chung. Word Segmentation for Chinese Phonetic Symbols. *Proceedings of International Computer Symposium*, pp. 911-916, 1994.
[10] Fu, S.W.K, C.H. Lee and Orville L.C. A Survey on Chinese Speech Recognition. *Communications of COLIPS* 6 (1), pp.1-17, 1996.
[11] Gao, J, J. Goodman, M. Li and K.F. Lee. Toward a Unified Approach to Statistical Language Modeling for Chinese. *ACM Transactions on Asian Language Information Processing* 1(1), pp. 3-33, 2002.
[12] Gu, H.Y., C.Y. Tseng and L.S. Lee. Markov modeling of mandarin Chinese for decoding the phonetic sequence into Chinese characters. *Computer Speech and Language* 5(4), pp.363-377, 1991.
[13] Ho, T.H., K.C. Yang, J.S. Lin and L.S. Lee. Integrating long-distance language modeling to phonetic-to-text conversion. *Proceedings of ROCLING X International Conference on Computational Linguistics*, pp. 287-299, 1997.
[14] Hsu, W.L. and K.J. Chen. The Semantic Analysis in GOING - An Intelligent Chinese Input System. *Proceedings of the Second Joint Conference of Computational Linguistics*, Shiamen, pp. 338-343, 1993.
[15] Hsu, W.L. Chinese parsing in a phoneme-to-character conversion system based on semantic pattern matching. *Computer Processing of Chinese and Oriental Languages* 8(2), pp. 227-236, 1994.
[16] Hsu, W.L. and Y.S. Chen. On Phoneme-to-Character Conversion Systems in Chinese Processing. *Journal of Chinese Institute of Engineers*, 5, pp. 573-579, 1999.
[17] Huang, J.K. The Input and Output of Chinese and Japanese Characters. *IEEE Computer* 18(1), pp. 18-24, 1985.
[18] Kuo, J.J. Phonetic-input-to-character conversion system for Chinese using syntactic connection table and semantic distance. *Computer Processing and Oriental Languages* 10(2), pp. 195-210, 1995.
[19] Lee, C.W., Z. Chen and R.H. Cheng. A perturbation technique for handling handwriting variations faced in stroke-based Chinese character classification. *Computer Processing of Oriental Language*s 10(3), pp. 259-280, 1997.
[20] Lee, Y.S. Task adaptation in Stochastic Language Model for Chinese Homophone Disambiguation. *ACM Transactions on Asian Language Information Processing* 2(1), pp. 49-62, 2003.
[21] Lin, M.Y. and W.H. Tasi. Removing the ambiguity of phonetic Chinese input by the relaxation technique. *Computer Processing and Oriental Languages* 3(1), pp. 1-24, 1987.
[22] Lua, K.T. and K.W. Gan. A Touch-Typing Pinyin Input System. *Computer Processing of Chinese and Oriental Languages*, 6, pp. 85-94, 1992.
[23] Manning, C. D. and Schuetze, H. *Fundations of Statistical Natural Language Processing*, MIT Press, pp.191-220, 1999.
[24] Microsoft Research Center in Beijing, http://research.microsoft.com/aboutmsr/labs/beijing/, 2003
[25] On-Line United Daily News , http://udnnews.com/NEWS/
[26] Qiao, J., Y. Qiao and S. Qiao. Six-Digit Coding Method. *Commun. ACM* 33(5), pp. 248-267, 1984.
[27] Sproat, R. An Application of Statistical Optimization with Dynamic Programming to Phonemic-Input-to-Character Conversion for Chinese. *Proceedings of ROCLING III*, pp. 379-390, 1990.
[28] Tsai, J.L. and W.L. Hsu. Applying an NVEF Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem. *Proceedings of 19th COLING 2002*, Taipei, pp.1016-1022, 2002.
[29] Tsai, J.L, W.L. Hsu and J.W. Su. Word Sense Disambiguation and Sense-based NV Event-Frame Identifier. *Computational Linguistics and Chinese Language Processing* 7(1), pp.29-46, 2002.
[30] Tsai, J.L, G. Hsieh and W.L. Hsu. Auto-Discovery of NVEF word-pairs in Chinese. *Proceedings of ROCOLING XV*, pp.143-160, 2003.
[31] Tsai, J.L, C.L. Sung and W.L. Hsu. Chinese Word Auto-Confirmation Agent. *Proceedings of ROCOLING XV*, pp.175-192, 2003.
[32] Tsai, J.L, G. Hsieh and W.L. Hsu. Auto-Generation of NVEF knowledge in Chinese. *Computational Linguis-*

tics and Chinese Language Processing 9(1), pp.41-64, 2004.

Appendix A. Exclusion Word List

I. Monosyllabic exclusion words

/之/的/不/與/兩/再/以/了/較/就/次/得/於/已/把/都/太/一/某/最/內/均/原/由/被/全/初/及/將/該/總/塊/項/和/二/從/三/凡/尚/前/十/極/番/元/件/甚/因/甲/向/才/四/本/若/先/便/五/粒/常/卅/後/左/曾/竟/廿/八/支/六/著/首/剛/應/篇/能/七/終/依/位/暫/共/須/中/九/時/可/俱/整/謹/宜/邊/往/批/夥/在/唔/年/諸/略/束/特/磅/

II. Polysyllabic exclusion words

/所以/不能/不會/是否/之間/終於/不必/唯一/西方/恐怕/連續/必須/不妨/大家/不得/一旦/初步/據說/看來/全面/臨床/無數/依法/國立/過度/突然/通常/一同/單一/大力/純粹/大都/當然/種種/大概/國有/順便/總是/不再/默默/無不/那麼/黑白/個人/四處/自行/恰好/終究/最佳/一心/十分/甚為/私立/一起/可以/多元/所有/依然/現成/正好/針對/一般/難怪/等到/到底/應該/貿然/獨家/原先/根據/微微/不勝/國產/整整/衷心/好些/安然/慈善/為什麼/一下子/一塊兒/非正式/

Appendix B. Two tonal and toneless STW results used in this study (The pinyin symbols and English words in parentheses are included for explanatory purposes only)

I.

Tonal STW results for the Chinese tonal syllable input “yi3 li4 gong1 ke4 guan1 ka3” of the Chinese sentence “以利攻克關卡”

Methods	STW results	Identified MWP word-pairs
MWP	**以利攻克關卡**	攻克(V)-關卡(N)/NV (key MWP); 以利(V)-攻克(V)/VV
MSIME	以利公克關卡	
MSIME+MWP	**以利攻克關卡**	
BiGram	以利公克關卡	
BiGram+MWP	**以利攻克關卡**	

Toneless STW results for the Chinese toneless syllable input “yi li gong ke guan ka” of the Chinese sentence “以利攻克關卡”

Methods	STW results	Identified MWP word-pairs
MWP	**以利攻克關卡**	攻克(V)-關卡(N)/NV (key MWP); 以利(V)-攻克(V)/VV
MSIME	以理工科關卡	
MSIME+MWP	**以利攻克關卡**	
BiGram	以利公克關卡	
BiGram+MWP	**以利攻克關卡**	

II.

Tonal STW results for the Chinese tonal syllable input “you2 qi2 zai4 cheng2 shou2 qi2 dao4 gu3 bao3 shi2 lv4 bu4 jia1” of the Chinese sentence “尤其在成熟期稻穀飽實率不佳”

Methods	STW results	Identified MWP
MWP	**尤其　成熟　稻穀飽實**	尤其(ADV)-成熟(V)/DV (key MWP); 稻穀(N)-飽實(V)/NV
MSIME	尤其再成熟期稻穀保十率不佳	
MSIME+MWP	**尤其**在**成熟**期**稻穀飽實**率不佳	
BiGram	尤其在成熟期稻穀保時率不佳	
BiGram+MWP	**尤其**在**成熟**期**稻穀飽實**率不佳	

Toneless STW results for the Chinese toneless syllable input “you qi zai cheng shou qi dao gu bao shi lv4 bu jia” of the Chinese sentence “尤其在成熟期稻穀飽實率不佳”

Methods	STW results	Identified MWP
NVEF	**尤其　成熟　稻穀飽實**	尤其(ADV)-成熟(V)/DV (key MWP); 稻穀(N)-飽實(V)/NV
MSIME	尤其再成熟期稻穀保十率不佳	
MSIME+MWP	**尤其**再**成熟**期**稻穀飽實**率不佳	
BiGram	尤其在成熟期稻穀保時率不佳	
BiGram+MWP	**尤其**在**成熟**期**稻穀飽實**率不佳	

語料庫統計值與全球資訊網統計值之比較：以中文斷詞應用為例

林筱晴　陳信希

國立台灣大學資訊工程學系

hclin@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw

摘要. 近年來全球資訊網(World Wide Web，簡稱 Web)快速成長，不同來源、不同領域、不同媒體的資訊透過網路傳遞到使用者手上。Web 除了扮演資訊傳播的角色外，也可以被視為是一個超大的資料集，提供語料庫為基礎－統計導向方法(Corpus-Based Statistics-Oriented Approach)所需要的統計值。本文以中文斷詞應用為例，由傳統語料庫和全球資訊網中，取得運用 word-based n-gram model 解斷詞歧義時所需要的統計值，藉以比較傳統語料庫和全球資訊網的差異。在第一組實驗，我們假設完全沒有未知詞，運用傳統語料庫的統計值最佳，其次依序為 Google 為基礎、AltaVista 為基礎、和 Openfind 為基礎。在第二組實驗，我們針對指定實體辨識，地名和組織名這兩類有不錯的效能。在第三組實驗，我們整合斷詞系統與指定實體辨識模組，全球資訊網統計值比傳統語料庫的統計值好。在最後一組實驗，我們將傳統語料庫和全球資訊網混合在一起，以全球資訊網統計值解決未知詞問題，再以語料庫統計值解斷詞歧義性，實驗顯示具有最佳的斷詞效能。

1. 緒論

在統計式自然語言處理(statistical natural language processing)，語言模型的設計、和統計值的來源是兩個實驗成功不可缺少的要件。因為語言是“活的”(live)，在日常生活中不斷有新的辭彙、和新的用法產生，系統必須能及時反應新的語言現象，因此使用具有時效性的資源非常重要。對於傳統語料庫而言，資料量規模固定、內容領域變動小、時效性較弱是其缺點，但優點是可以先加上標記(tagging)，增加附加價值，同時可以直接透過程式，精確掌握所需要的統計資訊。相對地，全球資訊網(World Wide Web，簡稱 Web)擁有十分龐大的資訊量、收集各種不同種類的文件、動態性等優點，但缺點是沒有加上語言標記，通常需要透過搜尋引擎(search engine)取得統計資訊，容易受到搜尋引擎本身設計上的限制(例如，文件索引的方式、查詢詞彙處理等)。本文將 Web 視為一個資料量龐大、且具時效性的語料庫，研究如何利用網路上的資訊來訓練統計式語言模型，並與傳統語料庫比較。近年來，運用 Web 於自然語言處理，有些相關論文發表。Zhu 和 Rosenfeld (2001)運用 Web 改善 trigram model，Computational Linguistics 期刊(2003) 也發行專刊探討這個課題，Resnik 和 Smith (2003) 將 Web 視為平行語料庫，提供翻譯模型所需要的雙語句子。Keller 和 Lapata (2003) 說明由語料庫和 Web 上所擷取的英文 bigram 統計值是有關聯性，但顯而易見這項理論在中文上，由於斷詞的問題，不見得就成立。

中文斷詞在中文自然語言處理上是個基本的工作，許多自然語言處理應用都以斷詞作為前置處理，例如機器翻譯、問答系統、自動摘要等。歧義性是中文斷詞系統第一個必須解決的問題，由於中文字串可能有多種不同的斷詞組合，斷詞系統必須選出其中最好的一種斷詞方式。另外，受限於辭典覆蓋度的問題，未知詞處理也是必要的工作。而指定實體(named entities，簡稱 NE)是常見的未知詞，一般斷詞系統都會輔以指定實體辨識模組，提出策略自動辨識出人名、地名、和組織名等的存在(Chen, Ding 和 Tsai, 1998; Chen, Yang 和 Lin, 2003)。本文以中文斷詞系統為例，以統計式語言模型作為基礎，將 Web 統計資訊應用在中文斷詞上。透過對搜尋引擎查詢，所傳回的網頁數(page count)，模擬統計式模型所用到的詞頻，藉此比較以傳統語料庫和以全球資訊網為基礎的差異。

本文第 2 節首先說明中文斷詞所用到的語言模型，以及所需的統計值，傳統語料庫和全球資訊網如何提供，在這裡我們假設完全沒有未知詞問題的存在。第 3 節討論指定實體辨識，我們運用可能比例測試(likelihood ratio test)方法，判斷某一字串是否為指定實體。第 4 節嘗試將斷詞和

指定實體辨識整合，以彰顯傳統語料庫和全球資訊網的特點。第 5 節提出結論與未來可能的研究。

2. 中文斷詞

2.1. 實驗材料

中文斷詞實驗所用到的辭典，是從中研院平衡語料庫(ASBC, 1995)擷取。我們將語料庫中的詞，直接收錄在辭典內，共 145,929 個詞。此外，我們將平衡語料庫切分成兩部分，作為訓練語料庫(5,386,820 個詞)、以及測試語料庫(586,698 個詞)。前者用以訓練實驗的語言模型，後者作為實驗測試的標準答案。

詞綴為具有衍生性的附著語素，可根據構詞律組合成詞。前綴為「附加於別的成分之前構成詞」，例如：大卡車、大騙子中的「大」屬於前贅詞。後綴為「附加於別的成分之後構成詞」，例如：犧牲者、造物者中的「者」屬於後贅詞。在前綴/後綴詞表中，我們共收錄 1,135 個前綴詞，以及 1,419 個後綴詞。

2.2. 演算法

2.2.1. Word-based N-gram Model

本文採用 word-based bigram model 作為斷詞系統的語言模型，對中文字串 $c_1c_2c_3...c_n$，查完辭典後可能會得到一種以上的組合，歧義性分析就是從這些不同的組合中，挑選出最可能的詞串 $\hat{W} = w_1w_2...w_m$，使得機率值 $\prod_{i=1}^{m} P_r(w_i \mid w_{i-1})$ 為最大，也就是說：

$$\hat{W} = \arg\max_{W} P_r(W \mid C) \approx \arg\max_{W} \prod_{i=1}^{m} P_r(w_i \mid w_{i-1})$$

$P_r(w_i \mid w_{i-1})$ 代表在前一個字是 w_{i-1} 的情況下，w_i 出現的機率。這個數值可以轉換成頻率，以最大可能估計(maximum likelihood estimation, MLE)計算如下：

$$P_r(w_i \mid w_{i-1}) \approx P_r(W_{i-1}W_i) / P_r(W_{i-1}) = \frac{Count(w_{i-1}w_i)/N}{Count(w_{i-1})/N} = \frac{Count(w_{i-1}w_i)}{Count(w_{i-1})}$$

$Count(w_{i-1}w_i)$ 是詞 w_{i-1} 和 w_i 相鄰出現的次數，Count(w_{i-1})是詞 w_{i-1} 出現的頻率。採用傳統語料庫策略，詞頻從中研院平衡語料庫訓練取得。採用全球資訊網策略，詞頻是以搜尋引擎(如：Google)所傳回的網頁數來模擬。檢索時查詢關鍵字(search term)前後會加上雙引號（“ ”），代表必須是“exactly match”。

由於統計資料的稀疏性(sparseness)問題，不論是透過訓練語料庫、或是全球資訊網，所得到的 $Count(w_{i-1}w_i)$ 數值，會出現某些相鄰詞組沒有共同出現過的情形，因而發生「零機率」，即 $P_r(w_i \mid w_{i-1}) = 0$，在實驗中我們將其機率值設定為一個極小的數值($10^{-12}$)。

除了 word-based bigram model 外，我們也同時規劃 word-based tri-gram model 的實驗。我們想知道詞串的掃描方向，對於整個實驗結果的影響，所以除了由左到右掃描字串外，也由右而左掃描詞串，作了 reverse bigram model 的實驗。

2.2.2. Prefix/Suffix Rule

對於某個詞串 AB，若 A、B、AB 皆收錄在辭典內，則會有「A　B」和「AB」兩種不同的斷詞組合，在解歧義性階段會選擇其中之一，作為斷詞結果。例如：「使用者」，因為辭典包含「使用、者、使用者」等詞彙，所以會產生「使用　者」和「使用者」兩種不同的組合。我們觀察中

研院平衡語料庫，這種包含前綴/後綴的詞彙，大部分都被斷成單一詞，而非兩個詞，所以增加 prefix/suffix rule 的策略：給予詞串 AB，查前綴/後綴詞表，發現 A 是前綴(或 B 是後綴)，且 AB 收錄於辭典中，則我們只建議「AB」這單一詞為候選者。例如「使用者」的「者」，查表發現屬於後綴詞，則系統就將之斷成「使用者」，而不會有拆開成兩個字的情形發生。

2.3. 結果與討論

2.3.1. 實驗結果

在第一組實驗，有三種語言模型：bigram model、reverse bigram model、和 tri-gram model；兩種策略：Dict-only 策略和 Dict+preffix/suffix rule 策略；全球資訊網：Google、Openfind、和 AltaVista，共有 11 個實驗，結果如表 1 所示。

首先我們固定語言模型為 bi-gram model，以及字串掃描方向為由左到右，以觀察傳統語料庫和全球資訊網之差異，不管在 Dictionary only 或 Dictionary and prefix/suffix rule 策略，bi-corpus 的 F-measure 都最好，其次依序為 bi-google、bi-altavista、和 bi-openfind。當字串掃描方向改為由右到左，這四種方法的順序不變，由右到左策略比由左到右的效能好。比較 bi-gram 和 tri-gram models，使用傳統語料庫和全球資訊網，bi-gram model 都比 tri-gram model 好。

表 1. 歧義性分析結果

	Dictionary Only			Dictionary and Prefix/Suffix Rules		
	precision	recall	F-measure	precision	recall	F-measure
bi-corpus	96.30%	95.39%	95.84%	97.09%	95.17%	96.12%
bi-google	94.76%	94.42%	94.59%	95.83%	94.24%	95.03%
bi-openfind	94.67%	93.42%	94.04%	95.39%	93.75%	94.56%
bi-altavista	94.70%	93.71%	94.20%	95.70%	94.16%	94.92%
rev-corpus	96.75%	95.08%	95.91%	97.00%	94.92%	95.95%
rev-google	94.87%	93.86%	94.36%	95.92%	94.30%	95.10%
rev-openfind	94.84%	93.56%	94.19%	95.52%	93.87%	94.69%
rev-altavista	94.82%	93.80%	94.31%	95.80%	94.23%	95.01%
tri-corpus	96.59%	94.64%	95.61%	96.66%	94.63%	95.63%
tri-google	93.88%	93.24%	93.56%	95.20%	93.83%	94.51%
tri-altavista	93.82%	93.13%	93.47%	95.04%	93.67%	94.35%

2.3.2. 實驗討論

我們將斷詞錯誤分成四種類型。

1. 合併：表示在標準答案中應該為分開的幾個詞，但系統卻將這些字合併為一個詞彙。例如標準答案應為「有　一些」 (例: 外頭　有　一些　矮　房子)，系統答案卻斷成「有一些」。
2. 拆開：表示在標準答案中應該為一個詞彙，但系統卻將此詞彙拆開成數個詞。例如標準答案應為「商學所」，系統答案卻斷成「商學　所」。
3. 搶字：表示兩個相鄰的辭彙，其中一個詞彙的部分中文字元被另一個詞彙搶走形成它的一部分。例如標準答案應為「法務　部門」，系統答案卻斷成「法務部　門」。
4. 其他：不屬於以上三種型態的錯誤，通通歸於其他。例如標準答案應是「及　第二組」，系統答案卻斷成「及第　二組」；另外，又如標準答案應是「原　裝設　於」，系統答案卻斷成「原裝　設於」。

根據這四種錯誤型態分類，我們計算不同模型的錯誤個數，如表 2 所列。

表 2. 錯誤型態之分佈

	Dictionary Only				Dictionary and Prefix/Suffix Rules			
	合併	拆開	搶字	其它	合併	拆開	搶字	其它
bi-corpus	12,136	3,805	1,244	402	14,931	942	835	84
bi-google	12,133	6,811	2,496	508	14,894	2,225	2,518	518
bi-openfind	14,828	3,936	3,915	718	14,967	1,702	3,896	743
bi-altavista	14,675	5,280	2,541	742	14,895	2,162	2,518	783
rev-corpus	14,852	1,330	1,438	109	15,066	744	1,482	112
rev-google	14,584	5,363	2,409	507	14,917	2,051	2,402	529
rev-opefind	14,834	3,798	3,572	710	14,975	1,674	3,557	735
rev-altavista	14,669	5,138	2,378	723	14,896	2,117	2,357	747
tri-corpus	14,961	787	2,318	153	15,125	527	2,319	150
tri-google	14,547	6,556	2,835	1,329	14,848	2,507	2,843	1,307
tri-altavista	14,578	6,125	3,230	1,447	14,848	2,326	3,239	1,447

由表 1 所列的實驗結果可知，無論採用何種語言模型，最後的表現都是傳統語料庫的方法最佳。這是因為傳統語料庫方法，所用的訓練語料庫是斷過詞的語料，在訓練階段掌握真正的詞頻。相對地，全球資訊網的統計值則是用搜尋引擎傳回的網頁數來估算，而網頁數直觀上僅代表「有多少個網頁包含此特定的詞」，是資訊檢索中所稱的「文件頻率」，並不是真正的詞頻。此外，全球資訊網的網頁，沒有像傳統語料庫已經人工斷詞標記，頻率會有誤差。例如，我們對 Google 下「門聯」這個查詢，會發現包含「澳門聯網」這個字串的網頁，也被計算在網頁數中，但事實上「澳門聯網」正確斷詞為「澳門　聯網」，所以不能計算在「門聯」的網頁數內，因此以網頁數來替代詞頻會有誤差。

由表 2 可觀察到傳統語料庫方法發生搶字類型的錯誤情況，明顯比全球資訊網方法來得少。字串 ABC 若可以斷成 A/BC 或是 AB/C，則會有相鄰兩詞彙互相搶字的情況，根據 word-based bigram model，這兩種斷詞組合的機率值為 $count(ABC)/count(A)$，及 $count(ABC)/count(AB)$。在全球資訊網方法，這兩個數值的分子部分相同，都是字串 ABC 的網頁數，所以分母部分決定了這兩個數值之間的大小關係。如果 A 屬於高頻字，則 $count(A)$ 數值很大，使得 $count(ABC)/count(A) < count(ABC)/count(AB)$，則斷詞系統會選擇 AB/C 作為斷詞結果，反之亦然。這種斷詞模式往往造成錯誤的結果，例如：將「後來/的」誤斷成「後/來的」、將「是/故意」誤斷成「是故/意」。相對於傳統語料庫方法，因為統計值皆來自於經過斷詞的訓練語料庫，因此分子部分的 $count(ABC)$ 在這兩種斷詞組合下，實際上為 count(A　BC)和 count(AB　C)此二個不同的數值，所以這種錯誤情況的發生較少。

統計值來自 Google 的實驗，拆開類型的錯誤最多，這是因為 Google 採用查辭典方式建立索引，所以發生「substring 的網頁數比 superstring 的網頁數來得少」這種情形，造成 $P_r(w_{i-1} \mid w_i) = \frac{Count(w_{i-1}w_i)}{Count(w_{i-1})}$ 的數值大於 1。當發生這種情況時，斷詞結果會傾向於拆開成多個詞彙，例如含「原住民」的網頁數為 170,000，而「原住」的網頁數只有 9,990，因此在例子「南島語族系的原住民」中，會被誤斷詞為「原住　民」。

實驗採用 prefix/suffix rule 策略可以減少拆開類型的錯誤，因為我們會把所有包含前綴/後綴的詞彙斷成單一詞彙，而不會拆成兩個詞彙。雖然這個策略同時也可能增加合併類型的錯誤，將標準答案中應該分開的兩個相鄰詞彙，合成單一詞彙。例如：標準答案為「性/騷擾」，系統誤斷為「性騷擾」，但整體斷詞系統的效能是提升了。我們從錯誤答案中也發現，有些字串在中研院平衡語料庫中，並沒有一致性的斷詞，例如「小河」在「隨著　屋　後　小河　的　潮　漲　潮　落」和「我　和　弟弟　溜到　附近　的　小　河　玩　水」，這兩個句子就有不同的斷詞標記。

我們仔細比較三個搜尋引擎的差異，Google 在文件層次、和查詢層次皆參考內建辭典，會有 substring 沒有收錄到辭典中，所以網頁數會較 superstring 少的情形發生。而 AltaVista 因為沒有查詞的動作，所以與檢索詞彙字串比對，成功的網頁就會被傳回。以「巴拿馬」和「巴拿」的例子來說明：對 Google 查詢「巴拿馬」，傳回的網頁數為 89,900；但如果查詢「巴拿」，則會發現它的網頁數為 11,200，比「巴拿馬」的數值來得少。對 AltaVista 查詢「巴拿馬」，得到的網頁數為 393，查詢「巴拿」所得到的網頁數為 9,560，大於「巴拿馬」的數值，與「substring

的限制性比 superstring 弱，應有較多的網頁包含」這個假設吻合。

Openfind 顯示的網頁數和實際傳回的網頁數不符合，例如查詢「中文詞知識庫計畫」，會出現「Openfind 找到 10 篇相關網頁」的訊息，但實際上只顯示了 7 個網頁連結。另外，Openfind 會偵測使用者的查詢流量，若是查詢動作過於頻繁，會限制此使用者的查詢，網頁會出現「很抱歉，系統偵測您的查詢狀況異常，已限制查詢。」的訊息，因此不適合對 Openfind 作大量的查詢。

Google 所顯示的網頁數只是個估算的數值，例如我們對 Google 搜尋「電腦」會傳回數值 4,110,000，搜尋「網際網路」則會傳回數值 322,000。這是由於 Google 為分散式查詢，為了爭取時間效率，當一台機器搜尋結束之後，即立刻回傳資訊給使用者，因此造成網頁數不精確的結果。

總結，傳統語料庫方法因為可以精確計算出詞頻，相較於全球資訊網方法受限於搜尋引擎的限制，造成網頁數並不準確，因此在斷詞解歧義實驗有較佳的結果，不過差距約在 1%-1.5%間，全球資訊網方法仍可媲美於傳統語料庫方法。

3. 指定實體辨識

3.1. 實驗資源

由於中研院平衡語料庫，沒有對指定實體作特別的標記，而是將其切分成連續的辭彙，例如「塞凡尼克國際公司」被標記為「塞凡尼克　國際　公司」，因此這個語料庫不適合用來做為評估指定實體辨識效能的素材。本階段的實驗採用 MET-2 測試語料(MUC, 1998)，作為指定實體辨識的測試文件。

這階段實驗所用到的辭典，仍然是由中研院平衡語料庫所擷取的辭彙組成，但做了小部分的更改。我們檢查 MET-2 測試語料所有答案(即指定實體)，如果指定實體已被收錄在辭典中，則從辭典中刪除這些詞彙。主要的目的是「使辭典與答案形成互斥(mutual exclusive)的關係」，以精確地評估系統效能。

指定實體關鍵詞集，收錄中文人名、地名、與組織名等三種類型的指定實體的關鍵詞。其中，中文人名部分共收錄了 387 個中文姓氏，例如：陳、李等等。地名部分收錄了 32 個關鍵詞，例如：市、港等等。組織名則收錄了 851 個關鍵詞，例如：黨、工會等等。我們根據指定實體關鍵詞集，由查詞典處理後的句子中，找出可能的指定實體候選詞，再由後續步驟判斷是否為指定實體，並確認其邊界。

有一些詞語出現頻率極高，例如「的、了」，這類的詞稱為停用字(stopword)。實驗中採用的停用字集，共收錄了 1,332 個停用字。當我們掃描詞串時，查詢此停用字集，就可得知哪些詞屬於停用字。

3.2. 可能比例測試(Likelihood Ratio Test)

3.2.1. 基本演算法

假設指定實體的組成成分間形成 collocation，要判斷 w^1 和 w^2 之間的關聯性，我們採用可能比例測試(Manning and Schutze, 1999)如下：

假設一. $P(w^2 \mid w^1) = p = P(w^2 \mid \neg w^1)$

假設二. $P(w^2 \mid w^1) = p_1 \neq p_2 = P(w^2 \mid \neg w^1)$

$$p = \frac{c_2}{N} \quad p_1 = \frac{c_{12}}{c_1} \quad p_2 = \frac{c_2 - c_{12}}{N - c_1} \quad \text{（假設 } p_1 > p_2 \text{）}$$

其中，c_1 表示 w^1 出現的頻率，c_2 表示 w^2 出現的頻率，c_{12} 表示 w^1 和 w^2 同時出現的頻率，N 則為語料庫內的辭彙總數。假設一代表 w^1 和 w^2 是獨立的，假設二則代表 w^1 和 w^2 有關聯性。

假設機率分布是 binomial distribution，計算可能比例(likelihood ratio) λ 的 log 值。而 $-2\log\lambda$ 的數值是呈現 χ^2 機率分配，當自由度等於 1 時，我們將信賴指標設定為 99%，其臨界值為 2.71。如果我們將所有參數值都代入公式，計算之後得到 $-2\log\lambda$ 值大於 2.71，則代表"接受假設二"，意即是 w^1 和 w^2 是有關聯的。

假設統計值來自檢索全球資訊網回傳的網頁數，令中文字串 $w = c_1c_2...c_k$ 為等待檢驗的指定實體候選字串，為了增加檢索字串的限制性，我們所取的子字串為原來字串頭尾各去除一個字元，亦即 $w_{left} = c_1c_2...c_{k-1}$，和 $w_{right} = c_2c_3...c_k$，則參數 $c_1 = pc(w_{left})$、$c_2 = pc(w_{right})$、$c_{12} = pc(w)$、N=搜尋引擎所收錄網頁總數，其中 $pc(word)$ 代表 $word$ 的網頁數。有了這些參數值，套進上述公式計算 $-2\log\lambda$ 的值。如果大於 2.71，則代表通過公式檢驗，我們認定字串 w 為一指定實體。如果小於 2.71，則表示非指定實體。舉例來說，令 w 為「艾特納保險公司」，則 w_{left} 為「艾特納保險公」，w_{right} 為「特納保險公司」。我們藉由搜尋引擎得到此三個字串的網頁數，並計算 $-2\log\lambda$ 值。

基本演算法如後：先運用辭典把句子切分出各種可能的詞組合，然後掃描詞串。若是掃描到的某個詞，屬於指定實體的關鍵字，則表示此詞串聯其前/後相鄰的幾個詞可能組成一個指定實體，因此驅動指定實體辨識的檢驗。如果屬於中文人名姓氏，則我們往後檢查最多兩個字，計算字串 $w_iw_{i+1}w_{i+2}$ 是否通過公式檢驗。若是某個詞 w_i 屬於地名、或是組織名的關鍵詞，則我們最多往前檢查五個詞，計算此串聯字串是否通過公式檢驗。我們對於通過公式檢驗的字串，再根據關鍵詞的類別，給予不同的語意標記。

例如，字串「酒泉衛星發射中心」，經過查辭典之後，得到「酒 泉 衛星 發射 中心」。掃描詞串發現「中心」是地名的關鍵詞，我們串聯前面五個詞(即「酒泉衛星發射中心」)，將此字串的各個相關參數值，即 w「酒泉衛星發射中心」、w_{left}「酒泉衛星發射中」、與 w_{right}「泉衛星發射中心」，在搜尋引擎所得到的網頁數套進公式檢驗。

對於外國人名部分，是以連續出現 n 個以上的單字詞，所串聯的字串來做判斷。如果此連續單字詞所組成的字串，能夠通過公式檢驗，我們就視之為外國人名，將字串加上人名的標示。

3.2.2. 修改演算法

根據原來公式的定義，當 $pc(w)$、$pc(w_{left})$、和 $pc(w_{right})$ 有一值為零時，則就不會通過檢驗。但是我們發現可能有種情況：檢索某字串 w，回傳的網頁數大於零($pc(w) > 0$)，但是其子字串的網頁數 $pc(w_{left})$、或 $pc(w_{right})$ 卻等於零。如果依照原始公式，這種情形的字串就會被遺漏，所以將原來檢驗過程，做了點小小的修改：當 $pc(w) > 0$ 時，若是 $pc(w_{left})$、或 $pc(w_{right})$ 的值等於零時，則直接通過檢驗，不需要計算 $-2\log\lambda$ 的數值。

3.3. 實驗與結果

3.3.1. 實驗一：字典與答案為互斥

表 3 列出第一組指定實體辨識實驗的結果，系統效能以 F-measure 表示，altavista_1 和 google_1 是以原始未修改的公式檢驗，altavista_2 和 google_2 是以修改後的公式檢驗。本實驗所用的辭典，與 MET-2 測試語料中的指定實體集為互斥關係，表示測試語料中的每個答案，都不會收錄於辭典內。由於一些指定實體 w 的網頁數大於零，但 w_{left} 或 w_{right} 卻等於零，因此無法被辨識出。實驗驗證 altavista_2 和 google_2 的效能，都比對應的 altavista_1 和 google_1 高。另外，一些停用字出現頻率太高，因而導致系統辨識出的指定實體邊界錯誤，例如：「在太原衛星發射中心」，"在"是多餘的。因此，當我們掃描詞串時，碰到停用字就停止。altavista_3 和 google_3 是以修改後的

公式，加上判斷是否包含停用字的結果。修正後的公式，加上停用字的判斷，有助於提升指定實體的辨識，所以之後的實驗以 altavista_3 和 google_3 為基礎。本方法可以找出，如「李塵風」、「西昌衛星發射中心」、和「國際衛星通信組織」等指定實體。

表 3. 指定實體辨識實驗一之結果

	PER	LOC	ORG	Total
altavista_1	34.60%	12.98%	51.41%	32.12%
altavista_2	34.60%	12.54%	52.81%	32.50%
altavista_3	43.80%	16.95%	62.58%	39.44%
google_1	47.03%	10.22%	21.37%	22.06%
google_2	48.70%	10.02%	40.42%	29.18%
google_3	55.84%	16.72%	54.65%	37.84%

3.3.2. 實驗二：辭典內增加收錄國名和省名

實驗一地名辨識不好的主要原因之一，是許多地名並沒有包含常見的關鍵字，例如國名：印度、巴西等等，所以沒辦法根據關鍵字來驅動這些地名的辨識。修改的方式是：在辭典內增加收錄世界各國的國名，以及大陸各省的省名，這部分在地名來說是比較屬於 closed set。有了這方面的資訊之後，我們可以對文件中的國名與省名作標記，以增加地名的辨識效能。文件中也常常出現地名縮寫，例如：英、法、德等等，我們在辭典內增加收錄一些國名縮寫，因此部分國名縮寫可以被辨識出來，但像是「中」或「以」這類國名縮寫，本身屬於常常出現的停用字，則沒被收錄在辭典內。

表 4 中的 altavista_4 和 google_4 表示以修改後的公式，和增加收錄地名的辭典所實驗的結果，altavista_5 和 google_5 表示增加國名縮寫的標記，altavista_5_n 和 google_5_n 表示：我們檢查連續出現 n 個以上的單字詞是否通過檢驗，通過視之為外國人名，藉此觀察對人名辨識的影響。表 4 顯示增加收錄國名和省名，以及標記部分國名縮寫後，能提高辨識效能。增加外國人名的判斷，在連續 5 個單字詞判斷時分數為最高。AltaVista 在組織名辨識效能較 Google 佳，而 Google 辨識人名和地名的效能則較 AltaVista 佳。

表 4. 指定實體辨識實驗二結果

	PER	LOC	ORG	Total
altavista_4	48.30%	74.29%	62.71%	67.45%
altavista_5	48.30%	77.45%	62.71%	69.54%
altavista_5_2	23.16%	76.40%	62.71%	59.39%
altavista_5_3	40.27%	77.48%	62.71%	67.34%
altavista_5_4	47.93%	77.65%	62.71%	69.28%
altavista_5_5	50.91%	77.45%	62.71%	69.81%
altavista_5_6	48.45%	77.45%	62.71%	69.57%
google_4	61.40%	75.61%	55.15%	67.77%
google_5	61.40%	78.94%	55.15%	70.02%
google_5_2	30.07%	77.95%	55.15%	60.12%
google_5_3	51.25%	78.91%	55.15%	67.92%
google_5_4	58.12%	78.99%	55.15%	69.38%
google_5_5	63.91%	78.94%	55.15%	70.32%
google_5_6	61.21%	78.94%	55.15%	70.00%

3.3.3. 實驗三：混合使用 Google 和 AltaVista 的統計值

本實驗結合此兩搜尋引擎的優點：當我們要辨識一詞串是否為人名、或地名時，採用 Google 搜尋回來的網頁數代入公式檢驗。如果要辨識是否為組織名，所套入公式的參數值，則是採用 AltaVista 所傳回的網頁數。表 5 中的 mix_1 表示以修改後的公式，以及混合使用 Google 和 AltaVista 的統計值後的實驗結果，mix_2 表示增加國名縮寫的標記後的實驗結果。mix_2_n 表示我們檢查連續出現 n 個以上的單字詞是否通過檢驗。

表 5. 指定實體辨識實驗三結果

	PER	LOC	ORG	Total
mix_1	60.30%	75.47%	62.69%	69.56%
mix_2	60.30%	78.80%	62.69%	71.69%
mix_2_3	50.00%	78.76%	62.69%	69.47%
mix_2_4	56.49%	78.85%	62.69%	70.90%
mix_2_5	62.43%	78.80%	62.69%	71.92%

3.3.4. 實驗四：刪除測試資料內網頁數等於 0 的答案

以全球資訊網為基礎的指定實體辨識，效能的好壞取決於網路上是否有收錄此詞彙的資訊。如果沒有，搜尋結果的網頁數就會是零，不能通過檢驗，因此遺漏此答案。我們想知道在最佳情況下，此演算法的效能，即每個測試文件中的指定實體都可以找到對應的統計值，所以將原來測試文件根據 Google 和 AltaVista 做了兩份不同的修改，把網頁數等於 0 的指定實體從原始文件中刪除，以此修改後的文件當做系統的輸入資料。表 6 中的 altavista_6 和 google_6 是以修改後的演算法，配合增加收錄地名的辭典所得到的結果，altavista_7 和 google_7 表示增加國名縮寫的標記，altavista_7_n 和 google_7_n 表示我們檢查連續出現 n 個以上的單字詞是否通過檢驗。

表 6. 指定實體辨識實驗實驗四結果

	PER	LOC	ORG	Total
altavista_6	67.53%	80.96%	78.16%	78.74%
altavista_7	67.53%	84.75%	78.16%	81.18%
altavista_7_4	66.67%	84.81%	78.16%	80.94%
altavista_7_5	70.29%	84.75%	78.16%	81.42%
google_6	64.74%	77.55%	58.93%	70.58%
google_7	64.74%	81.06%	58.93%	72.93%
google_7_4	61.16%	81.12%	58.93%	72.26%
google_7_5	67.50%	81.06%	58.93%	73.27%

表 6 顯示 AltaVista 的實驗結果比 Google 好，原因之一是在 AltaVista 的實驗，我們從測試文件刪除較多的指定實體，所以要辨識的指定實體個數 AltaVista 實驗比 Google 實驗少。另外，某些詞彙在 Google 的三個統計值($pc(w), pc(w_{left}), pc(w_{right})$)，雖然都不等於零，但套進公式計算之後的結果，卻不能通過檢驗。例如「新華社」等詞彙，因而造成 Google 無法辨識出這些指定實體。對應之下，AltaVista 中幾乎不發生這種情形，唯一例外的例子是「羅俏」這個詞。

3.4. 分析與討論

3.4.1. 測試資料之分析

在 MET-2 測試語料中，全部共有 1,301 個指定實體，單字詞因為只包含一個字元，所以無法套用公式。不包含重複，且可供檢測的指定實體共有 384 個，我們分析這 384 個指定實體，結果如表 7 所示。

表 7. 測試資料分析

	通過檢測		未通過檢測		網頁數等於 0		substring page count 較 superstring 少	
	AltaVista	Google	AltaVista	Google	AltaVista	Google	AltaVista	Google
PER	70	97	36	9	35	9	1	27
LOC	108	112	16	12	16	6	3	84
ORG	92	122	62	32	62	22	9	121
Total	270	331	114	53	113	37	13	236

就 AltaVista 而言，未通過的指定實體總共有 114 個，而網頁數等於零的有 113 個，猜測檢驗不通過的原因，可能是來自於搜尋引擎所收錄的網頁中沒有包含此詞彙的資訊。如果搜尋結果的網頁數大於零，幾乎大部分都通過檢驗。唯一的例外就是「羅俏」這個人名，雖然 AltaVista 可以搜尋到有關這個詞的網頁，可是這些數值代入公式後，並不通過檢驗。就 Google 而言，未通過的指定實體總共有 53 個，而網頁數等於零的只有 37 個，這表示有 16 個指定實體雖然在 Google 上可以搜尋到相關網頁，可是這些數值代入公式，並不會通過通過檢驗，例如「新華社」。

進一步分析網頁數等於零，以及通不通過檢定，和搜尋引擎內部的設計間的關係。考慮原始檢驗的三個重要參數：$w = c_1 c_2 ... c_k$ 、 $w_{left} = c_1 c_2 ... c_{k-1}$ 、及 $w_{right} = c_2 c_3 ... c_k$ ，我們故意將組合成指定實體的辭彙頭尾各去除一個字元，例如「艾特納保險公司」，則 w_{left} 為「艾特納保險公」，w_{right} 為「特納保險公司」。在 substring 的結合性比 superstring 弱的假設下，w_{left} 和 w_{right} 的網頁數應該大於 w 的網頁數，可能比例測試就建立在這個假設之下。AltaVista 完全用字串比對，上述假設是成立的，其主要的問題是網頁數等於零的情況太多。猜測的原因是，地名和組織名都比較長。相對的，Google 經查詞處理，前述的假設可能就不成立，造成比例測試不通過，但其網頁數等於零的情況較少。

3.4.2. 錯誤分析

因為某些指定實體本身包含停用字，例如「美國航空航天局」和「楊天」中的「天」是停用字，在這種情況下系統就無法正確地辨識出。某些外國人名的錯誤：「塞萬提斯」會變辨識成「萬提斯」，因為「萬」是中文人名姓氏。部分錯誤是因為指定實體太長，如「中國衛星發射測控系統部」會被斷詞成「中國 衛星 發射 測 控 系統 部」，「部」是組織名的關鍵字，我們最多只往前找五個詞，因此「衛星發射測控系統部」，就會被辨識為組織名，屬於左邊界錯誤。

有些錯誤是因為網頁數等於零所引起的，所以會造成 MIS 或 INC 的錯誤。像是把「美國艾科斯達衛星公司」，辨識為「斯達衛星公司」。有時候網頁數等於零不只會造成 INC 錯誤，同時也會增加 SPU 錯誤，例如「香港亞太通信衛星公司」，會被辨識為「香港(LOC) 亞太通信衛星公司(ORG)」；「中國空間技術研究所」，會被辨識為「中國(LOC) 空間技術研究所(ORG)」。有些指定實體的網頁數雖然大於零，可是卻不會通過公式檢驗，會造成 MIS 錯誤，如「新華社」和「歐洲宇航局」等。

我們進一步分析各種錯誤情形。首先是 INC 錯誤，這類型錯誤就是指系統辨識出的指定實體邊界錯誤，例如「西昌衛星發射中心」，辨識為「抵運西昌衛星發射中心」。造成這種錯誤的原因，包括指定實體本身包含停用字、網頁數等於零、指定實體長度太長等因素。

總結，造成遺漏錯誤的原因有：

1. 指定實體本身包含停用字。
2. 網頁數等於 0。
3. 指定實體本身沒有關鍵字，像是『法塔赫』武裝、五角大樓等，這類型的指定實體因為缺乏關鍵字，所以無法驅動指定實體辨識。
4. 有些國名的縮寫沒辦法辨識出，例如「中」、「以」、「日」等。
5. 有些指定實體的關鍵字，並沒有收錄在關鍵字集中，因此不會驅動指定實體的辨識工作，例如「日本科技廳」中的「廳」。
6. 外國人名不一定會被斷詞為連續的單字詞，因此無法藉由連續單字詞的策略來辨識。

多餘錯誤發生的原因有：

1. 網頁數等於零。
2. 通過公式檢驗，但實際上並不是專有名詞，例如「私人公司」。
3. 人造衛星的名稱「亞洲二號」，系統將「亞洲」標記為地名。
4. 連續單字詞的標記，常常會標記錯誤為人名。

由本實驗可以說明：全球資訊網的查詢結果，和搜尋引擎的內部設計有關。藉由關鍵字驅動辨識工作，透過搜尋引擎得到網頁數，套入可能比例測試中，可以判斷一個字串是否為指定實體。

4. 斷詞與指定實體辨識的整合

4.1. 實驗資源

CTS 語料是由華視新聞內容所收集而成，文本與中研院平衡語料庫採用相同的斷詞標準與詞性標記。指定實體，會被標記為連續詞彙，例如「美商奇異公司」，被斷為「美商/奇異/公司」，但是在自然語言處理應用，我們希望斷詞的結果是將所有具有「完整語意」的詞彙切分出邊界，即我們希望得到的是「美商奇異公司」，因此我們對於人名、地名、和組織名這三種型態的指定實體，以人工標記的方式，將這些被拆開的專有名詞，組合起來成為單一的詞彙。用此修改過的語料作為測試語料，觀察增加未知詞偵測，對斷詞結果所帶來的影響。

4.2. 實驗步驟

首先經過查字典的步驟，找出各種候選詞彙，然後掃描詞串，進行指定實體辨識。如果某些相鄰的詞串經辨識出為屬於人名、地名、或是組織名，表示我們偵測到未知詞，則將這些詞彙組合起來成為單一詞。最後進行解歧義性步驟，找出此句子的最佳斷詞組合。

4.3. 結果與討論

表 8 為實驗結果，1_corpus、1_google、1_altavista 表示僅以辭典斷詞，未加上指定實體辨識模組的實驗。2_google_uw、2_altavista_uw，則是辨識出未知詞之後，才進行解歧義的實驗。2_google_corpus_uw 則是以 Google 統計值辨識出未知詞，這些辨識出來的指定實體，將輸入句子切分成幾個小句子，即「S1 UW1 S2 UW2 S3 …」，我們再分別以語料庫統計值，對這些小句子 S1、S2 等進行歧義性分析。

完全使用辭典，亦即未加指定實體辨識，與第 2 節的實驗結果一致，傳統語料庫方法的效能 92.38%，仍然比全球資訊網方法 91.21% 和 90.31%好。當考慮指定實體，整合進 Google 和 AltaVista 為基礎的斷詞中，效能分別增加 2.36%和 2.11%，首先超越純傳統語料庫的方法 1.19% 和 0.04%。再考量傳統語料庫統計值，解決斷詞歧義性的效能，比全球資訊網統計值佳的現象，我們先以全球資訊網偵測未知詞的存在，再以傳統語料庫統計值來解歧義性，得到的結果為最佳

94.66%。

表 8. 增加未知詞偵測的斷詞結果

	Recall	Precision	F-Measure
1_corpus	94.03%	90.79%	92.38%
1_google	93.07%	89.43%	91.21%
1_altavista	92.45%	88.27%	90.31%
2_google_uw	94.11%	93.02%	93.57%
2_altavista_uw	93.42%	91.45%	92.42%
3_google_corpus_uw	95.02%	94.31%	94.66%

測試文件共包含 7,169 個詞，其中有 224 個未收錄在系統辭典內的詞彙。這些未知詞，包含普通名詞「等壓線」、「國際約」，或是動詞「拘提」、「收賄」，以及專有名詞「華視晚間新聞」、「魏政賢」等。我們利用 web 資訊，可以辨識出未知詞，例如「魏政賢」、「大傑旅行社」、「中山足球場」等。但是對於其他型態的未知詞，則無法切分出正確邊界，因此「等壓線」會被斷為「等壓　線」，造成斷詞錯誤。由於中文人名辨識模組，只能找到長度為 2 或是 3 的人名，因此「張趙惠朱」無法正確的辨識出來，造成錯誤。另外由於某些指定實體中含有停用字，所以無法辨識成功，例如「彭南雄」斷為「彭 南 雄」。

5. 結論

本論文應用全球資訊網的統計值於自然語言處理上，並以中文斷詞為例。解歧義性實驗顯示傳統語料庫方法，得到的結果優於全球資訊網方法，但兩者差距不大。針對人名、地名、和組織名三種類型設計出一套指定實體辨識，實驗顯示辨識成功與否，取決於搜尋引擎的內部結構，和收錄的網頁內容。如果指定實體曾出現在網路上，則它的網頁數大於零，有很大機會通過公式檢驗，被成功地辨識出來。由於全球資訊網資訊量龐大、且具及時性，對指定實體辨識有很大潛力。

最後將斷詞和指定實體辨識系統整合，運用指定實體辨識模組偵測未知詞，予以正確的邊界切分。實驗顯示在原本的斷詞系統中加入未知詞偵測，對於斷詞效能是有助益的。由於傳統語料庫統計值解決歧義性問題的效能，略勝於全球資訊網統計值，而全球資訊網統計值可以用於指定實體辨識，因此結合雙方面的優點，先利用全球資訊網統計值偵測未知詞，再利用傳統語料庫解歧義性，使斷詞系統得到最佳的表能。

由於傳統語料庫資訊量不夠龐大、不具即時性、且需要大量人工標記、耗時，相較之下全球資訊網擁有資訊量龐大、即時性、且取得容易等優勢，本文提出全球資訊網方法解決中文斷詞問題，不需要太多語言知識，只需要透過搜尋引擎介面得到網頁數，視之為詞頻應用於統計模型上，實作容易。實驗顯示全球資訊網資訊，在自然語言處理上是有用的。

註謝

本文部分成果為國科會計畫 NSC 93-2752-E-001-001-PAE 補助。

參考文獻

[1] Hsin-Hsi Chen, Yung-Wei Ding and Shih-Chung Tsai (1998). “Named Entity Extraction for Information Retrieval.” *Computer Processing of Oriental Languages, Special Issue on Information Retrieval on Oriental Languages* 12(1), 1998, 75-85.

[2] Hsin-Hsi Chen, Changhua Yang and Ying Lin (2003). "Learning Formulation and Transformation Rules for Multilingual Named Entities." *Proceedings of ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models*, July 12, Sapporo, Japan, 2003, 1-8.

[3] CKIP (1995). A Description to the Sinica Corpus. *Technical Report 95-02, Academia Sinica, Taipei*.

[4] F. Keller and M. Lapata (2003). "Using the Web to Obtain Frequencies for Unseen Bigrams." *Computational Linguistics* 29(3), 459-484.

[5] Christopher D. Manning and Hinrich Schutze (1999). Foundations of Statistical Natural Language Processing, *MIT Press*, 1999.

[6] MUC (1998). *Proceedings of 7th Message Understanding Conference*, Fairfax, VA, 29 April - 1 May, 1998, http://www.itl.nist.gov/iaui/894.02/related_projects/ muc/index.html.

[7] P. Resnik and N. A. Smith (2003). "The Web as a Parallel Corpus." *Computational Linguistics*, 29(3), 349-380.

[8] X. Zhu and Ronald Rosenfeld (2001). "Improving Trigram Language Modelling with the World Wide Web." *Proceedings of the International Conference on Acoustics Speech and Signal Processing*.

Pronominal and Sortal Anaphora Resolution for Biomedical Literature

Yu-Hsiang Lin and Tyne Liang
Department of Computer and Information Science
National Chiao Tung University
Hsinchu, Taiwan
Email: gis91534@cis.nctu.edu.tw; tliang@cis.nctu.edu.tw;

Abstract. Anaphora resolution is one of essential tasks in message understanding. In this paper resolution for pronominal and sortal anaphora, which are common in biomedical texts, is addressed. The resolution was achieved by employing UMLS ontology and SA/AO (subject-action/action-object) patterns mined from biomedical corpus. On the other hand, sortal anaphora for unknown words was tackled by using the headword collected from UMLS and the patterns mined from PubMed. The final set of antecedents finding was decided with a salience grading mechanism, which was tuned by a genetic algorithm at its best-input feature selection. Compared to previous approach on the same MEDLINE abstracts, the presented resolution was promising for its 92% F-Score in pronominal anaphora and 78% F-Score in sortal anaphora.

1 Introduction

Anaphora resolution is one of essential tasks in message understanding as well as knowledge discovering. For example recognizing biomedical relations among biomedical entities from research literature like MEDLINE database requires anaphora resolution for those mentioned entities from texts.

There are different types of anaphora to be solved like pronominal, sortal (definite), zero, event, and coreference anaphora. In biomedical literature, pronominal anaphora and sortal anaphora are the two common anaphora phenomena. Pronominal anaphora is that mentioned entity is substituted by the pronoun. Sortal (definite) anaphora occurs in the situation that a noun phrase is referred by its general concept entity. Definite noun phrases are noun phrases stating with demonstrative articles, such as those, this, both, each and these or starting with a definite article.

Generally identifying antecedents of an anaphor can be handled by using syntactic, semantic or pragmatic clues. In past literature, syntax-oriented approaches for general texts can be found in [Hobbs, 76; Lappin and Leass 94; Kennedy and Boguraev 96] in which syntactic representations like grammatical role of noun phrases were used.

On the other hand more information other than syntactic information like co-occurring patterns obtained from the corpus was employed during antecedent finding in [Dagan and Itai, 90]. Information with limited knowledge and linguistic resources for resolving pronouns were found in [Baldwin, 97]. In [Denber, 98, Mitkov, 02], more knowledge from the outer resource like WordNet was employed in solving anaphora. Similarly WordNet together with additional heuristic rules were applied for resolving pronominal anaphora in [Liang and Wu, 04] which animacy information is obtained by analyzing the hierarchical relation of nouns and verbs in the surrounding context learned from WordNet.

In biomedical literature, it was found that sortal anaphors are prevalent in the texts like MEDLINE abstracts [Castaño et al., 02]. To deal this type of anaphora, Castaño et al. [02] used UMLS (Unified Medical Language System) as ontology to tag semantic type for each noun phrase and used some significant verbs in biomedical domain to extract most frequent semantic types associated to agent (subject) and patient (object) role of SA/AO-patterns. The result showed SA/AO-pattern could gain increase in both precision (76% to 80%) and recall (67% to 71%). In [Hahn et al., 02], a center list mechanism was presented to relate each noun to those nouns appearing in a previous sentence anaphora. Gaizauskas et al. [03] presented a predefined domain rules for ensuring co-referent between two bio-entities so that implicit relations between two entities could be recognized.

In this paper, the anaphora resolution for biomedical literature is achieved by employing UMLS ontology and syntactic information. The proposed system identifies both intra-sentential and inter-sentential antecedents of anaphors. In addition, anaphora resolution for unknown words has concerned in this paper by using headword mining and patterns mined from PubMed search results. Determining semantic coercion type of pronominal anaphor is done by SA/AO patterns, which were pre-collected from GENIA 3.02p corpus, a MEDLINE corpus annotated by Ohta et al. [02]. The final set of antecedents finding is decided with a salience grading mechanism, which is tuned by a genetic algorithm at its best-input feature selection. Compared to previous approach on the

same MEDLINE abstracts, the presented resolution is promising for its 92% F-Score in pronominal anaphora and 78% F-Score in sortal anaphora.

2 The Presented Resolution

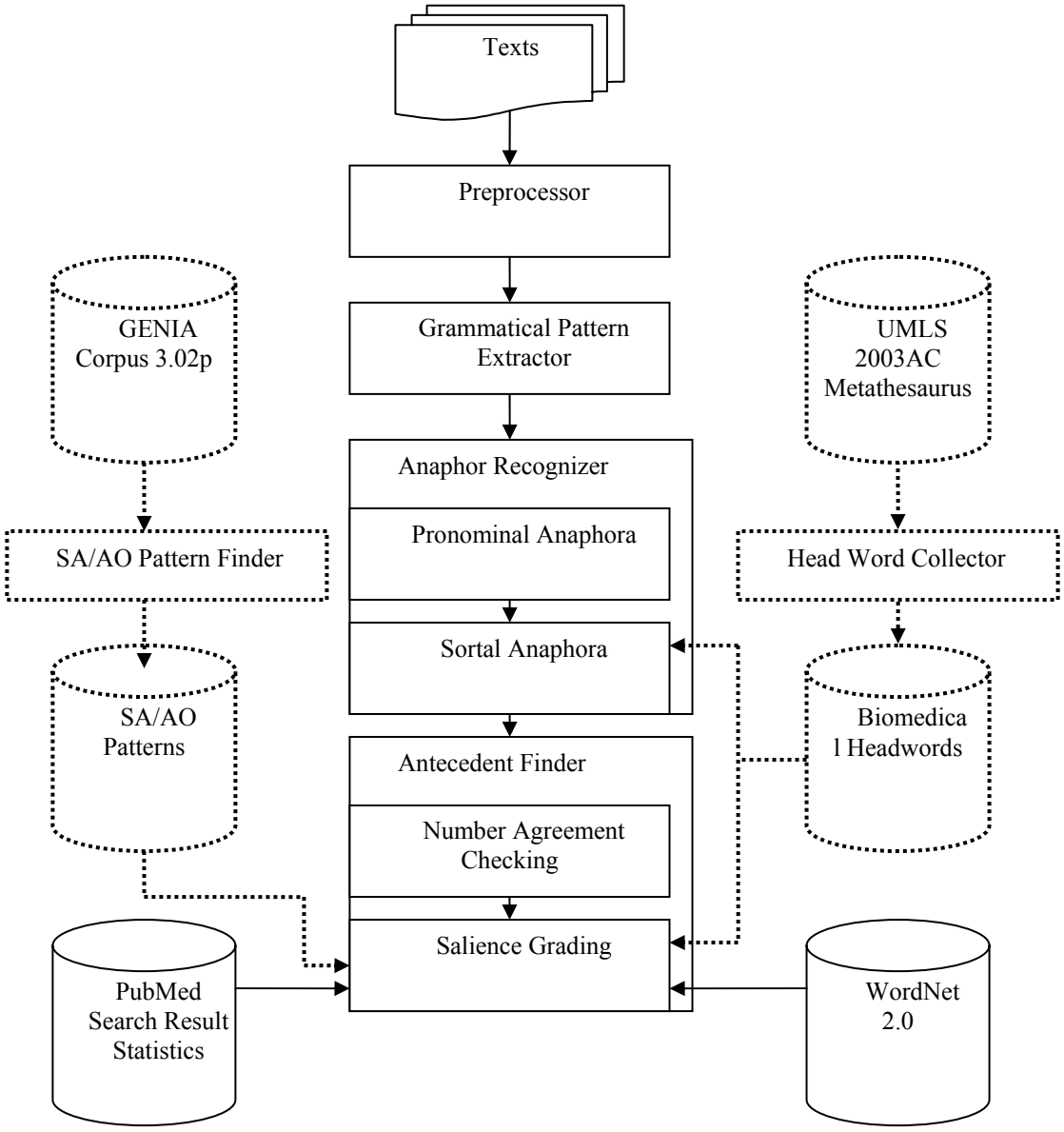


Figure 1: Architecture overview.

Figure 1 is the presented overview architecture which contains background processing, including SA/AO patterns and headword collection, indicated with dotted lines and foreground processing, including preprocessor, grammatical pattern extractor anaphor recognizer, and antecedent finder, indicated with solid lines.

2.1 SA/AO Patterns Collection

In this paper we used co-occurring SA/AO patterns obtained from GENIA corpus for pronominal anaphora resolution. Then we tag subjects and objects with UMLS-semantic type tags. Each SA/AO pattern is scored by the scoring function (Eq. 1). The antecedent candidates are concerned if their scores are greater than a given threshold.

$$score(type_i, verb_j) = \frac{frequency(type_i, verb_j)}{frequency(verb_j)} \times \frac{1}{No.\ of\ types(verb_j)} \quad (1)$$

The following is a pattern extraction example:

Example1:
<NFATp> <binds> to two sites within the kappa 3 element
UMLS semantic type of NFATp: Amino Acid, Peptide, or Protein
Extracted pattern: <Amino Acid, Peptide, or Protein> <bind>

2.2 Headword Collection

For unknown words, we need to predict their semantic types of the word. In [Pustejovsky et al., 02], they use the righthand head rule (the head of a morphologically complex word to be the righthand member of that word) to extract headwords to be subtype of the semantic type in UMLS. Table 1 is an example for headword 'receptor' which changes other noun phrase which were tagged with different semantic into 'Amino Acid, Peptide, or Protein'. 'Adhesion' is tagged with 'Acquired Abnormality, Disease or Syndrome' but 'adhesion receptor' becomes the tag of 'Amino Acid, Peptide, or Protein' by addition of 'receptor'.

Table 1: Example with righthand rule.

Noun Phrase	Semantic Type
Adhesion	Acquired Abnormality, Disease or Syndrome
adhesion receptor	Amino Acid, Peptide, or Protein
Contraction	Pathologic Function
Contraction receptor	Amino Acid, Peptide, or Protein
Estrogen	Steroid, Pharmacologic Substance, Hormone
estrogen receptor	Amino Acid, Peptide, or Protein
Dopamine	Organic Chemical…
dopamine receptor	Amino Acid, Peptide, or Protein

We collected all UMLS concepts and their corresponding synonyms, and then selected headwords for each semantic type (super-concept). For example, concept 'interleukin-2' has synonyms 'Costimulator', 'Co-Simulator', 'IL 2', and 'interleukine 2'. We collected 'interleukin', 'costimulator', 'simulator', 'IL', and 'interleukine' as headwords for 'interleukin-2'. Then, we found semantic types of 'interlukin-2' is 'Amino Acid, Peptide, or Protein' and 'Immunologic Factor'. We assigned synonym headwords of 'interleukin-2' into both semantic types. Eq. 2 was designed to score each headword for each semantic type. The scoring function smoothes the semantic type size.

Headword scoring function:

$$w_{i,j} = \frac{w_i}{Max\ \ c_j} \times \frac{1}{tw_i} \quad (2)$$

$w_{i,j}$: score of word i in semantic type j
w_i : count of word i in semantic type j
Max c_j : Max count of word k in semantic type j
tw_i : count of semantic types that word i occurs in

2.3 Preprocessor

After input untagged documents, we go through POS tagging and NP Chunking these preprocessing will give us more information about the documents.

2.4 Grammatical Function Extraction

Grammatical function is defined as creating a systematic link between the syntactic relation of arguments and their encoding in lexical structure. For anaphora resolution, grammatical function is an important feature of salience grading. We extended rules from Siddharthan [03], from following rules 1~4 to rules 1~6.

Rule 1: Prep NP (Oblique)
Rule 2: Verb NP (Direct object)
Rule 3: Verb $[NP]^{+}$ NP (Indirect object)
Rule 4: NP (Subject) [",[^Verb] appositive),"|Prep NP]* Verb
Rule 5: NP1 Conjunction NP2 (Role is same as NP1) Conjunction]
Rule 6: [Conjunction] NP1 (Role is same as NP2) Conjunction NP2

Rule 5 and rule 6 were presented for dealing those anaphors that have plural antecedents. We use syntactic agreement with first antecedent to find other antecedents. Without rules 5 and 6, 'anti-CD4 mAb' in Example 1 will not be found when resolving 'they''s antecedents.

Example 1:
"Whereas different anti-CD4 mAb or HIV-1 gp120 could all trigger activation of the ..., they differed…"

3 Anaphora Resolution

Anaphor and antecedent recognition are the two main parts of the anaphora resolution system. Anaphor recognition is to recognize the target anaphora by filtering strategies. Antecedent recognition is to determine appropriate antecedents with respect to the target anaphor.

3.1 Anaphora Recognition

Noun phrases or prepositional phrases with 'it', 'its', 'itself', 'they', 'them', 'themselves' and 'their' are considered as pronominal anaphor. 'it', 'its', and 'itself' are considered as anaphor which has singular number of antecedent, others are considered as anaphor which has plural number of antecedents. Relative pronouns 'which' and 'that' are also pronominal anaphors but these anaphors can use a simple rule, point to the nearest noun phrase or prepositional phrase, to find its antecedent or point to the relative clause behind when paired with a pleonastic-it.

Noun phrases or prepositional phrases with 'either', 'this', 'both', 'these', 'the', and 'each' are considered as candidates of sortal anaphors. Noun phrases or prepositional phrases with 'this' or 'the+ singular noun' are considered as anaphors which have singular antecedent. Anaphor with plural number of antecedents are shown in Table 2.

Table 2: Number of Antecedents

Anaphor	Antecedents #
Either	2
Both	2
Each	Many
They, Their, Them, Themselves	Many
The +No.+ noun	No.
Those +No.+ noun	No.
these +No.+ noun	No.

3.1.1 Pronominal Anaphora Recognition

Pronominal anaphora recognition was done by filtering out pleonastic-it. Following rules are used to recognize pleonastic-it instances.

Rule1: It be [Adj|Adv| verb]* that

Example 2:
"It is shown that antibody 19 reacts with this polypeptide either bound to the ribosome or free in solution."

Rule 2: It be Adj [for NP] to VP

Example 3:
"However, it is possible for antidepressants to exert their effects on the fetus at other times during pregnancy as well as to infants during lactation."

Rule 3: It [seems|appears|means|follows] [that]*

Example 4:
"It seems that the presence of HNF1 sites in liver-specific genes was favoured, but that no counter-selection occurred within the rest of the genome."

Rule 4: NP [makes|finds|take] it [Adj]* [for NP]* [to VP|Ving]

Example 5:
"Furthermore, the same experimental model makes it possible to image lymphoid progenitors in fetal and adult hematopoietic tissues."

3.1.2 Sortal Anaphora Recognition

Sortal anaphora recognition was done by filtering those sortal anaphor, which have no referent antecedent or which have antecedents but not in the defined biomedical semantic types. Following two rules are used to filter out those un-target anaphors.

Rule 1: Filter out those noun phrases or prepositional phrases if they are not tagged with the following UMLS classes.

Amino Acid, Protein, Peptide, Embryonic Structure, Cell Biomedical Active Substance, Organism, Functional Chemical, Bacterium, Molecular Sequence, Chemical, Nucleoside, Cell Component, Enzyme, Gene or Genome, Structural Chemical Nucleotide Sequence, Substance, Organic Chemical, Pharmacologic Substance, Organism Attribute, Nucleic Acid, Nucleotide.

Rule 2: Filter out proper nouns with capitals and numerical features.

3.2 Number Agreement Checking

Number is the quantity that distinguishes between singular (one entity) and plural (numerous entities). It makes the process of deciding candidates easier since they must be consistent in number. All noun phrases and pronouns are annotated with number (singular or plural). For a specified pronoun, we can discard those noun phrases whose numbers differ from the pronoun. With singular antecedent anaphor, plural noun phrases are not considered as possible candidates.

3.3 Salience Grading

Salience grade for each candidate antecedent is assigned according to Table 3. Each candidate antecedent is assigned with zero at initial state.

Recency is a feature about distance between an anaphor and candidate antecedents. The closer between an anaphor and a candidate antecedent, the more chance the anaphor points to this candidate antecedent. For grammatical role agreement, if we use same entity in the second sentence and in the same role, it is easy for readers to identify which antecedent that the anaphor points to, so an author might use anaphor instead of full name of the entity. In addition to role agreement, subjects and objects are important role in sentence, which may be mentioned many times and writer might use an anaphor to replace a previously mentioned items. Singular anaphors may only point to one antecedent, while plural anaphors usually points to plural antecedents. For the feature of semantic type agreement, when we mention entity the second time, it is common for us to use its hypernym concept. Therefore such feature will receive high weights at salience grading.

Table 3: Salience grading for candidate antecedents.

Features	Score
Recency	0-2
Subject and Object Preference	1
Grammatical Role Agreement	1
Number Agreement	1
Longest Common Subsequence	0-3
Semantic Type Agreement	-1 if not or +2
Biomedical Antecedent	-2 if not or +2

3.3.1 Antecedent and Anaphor Semantic Type Agreement

For pronominal anaphora, we collected coercion semantic type between verb and headword by GENIA SA/AO patterns, and we generalized subjects and objects by using UMLS semantic types. For a pronoun, we tagged the pronoun with coercion semantic types on the basis of SA/AO pattern.

Sortal anaphoras are dealt by checking semantic agreement between anaphor and antecedent. So, all noun phrases and prepositional phrases will be tagged in advance by following steps.

(1) UMLS type check
(2) The Antecedent contains the headword in the anaphor's semantic type.
(3) If there is no headword found in antecedent then check {anaphor, antecedent} pair by using PubMed

For {anaphor, antecedent} pair {The nmd mouse mutation, of a second site suppressor allele}, we created query1 :<anaphor: "The nmd mouse mutation", antecedent: "of a second site suppressor allele"> and query2: <antecedent: "of a second site suppressor allele">. Queries are used to query from PubMed website and Eq. 3 was used to score the antecedent for semantic type agreement.

$$Score = -1 + \left\lceil \frac{Query\ \ pages\ \ from\ \ query\ 1}{Query\ \ pages\ \ from\ \ query\ 2} \times 10 \right\rceil \times 0.3 \qquad (3)$$

3.3.2 Longest Common Subsequence (LCS)

The use of the LCS exploits the fact that the anaphor and its antecedents are morphological variants of each other (e.g., the anaphor "the grafts" and the antecedent "xenografts") [Castaño, 02]. We score each anaphor and candidate antecedent as follows:

If total match between a anaphor and its candidate antecedents
    then salience score = salience score + 3
Else if partial match between a anaphor and its candidate antecedents
    then salience score = salience score + 2
Else if one antecedent match its anaphor hyponym by WordNet 2.0
    then salience score = salience score + 1

3.3.3 Antecedent Selection

We search noun phrases or prepositional phrases in range of two sentences preceding the anaphor. We count salience grader scores for each noun phrase. Antecedents are selected by using best fit or nearest fit strategy.

(1) Best Fit: select antecedents with the highest salience score that is greater than threshold
(2) Nearest Fit: Select the nearest antecedents whose salience value is greater than a given threshold, and find candidate antecedents from the anaphor to the two sentences ahead

We have identified the number of antecedents for its corresponding anaphor. If an anaphor is identified to have plural antecedents, we will use following steps to choose antecedents.

(1) If the number of antecedents is identified, set the highest number of noun phrases or prepositional phrases to the anaphor.
(2) If the number of antecedents is unknown, find those noun phrases and prepositional phrases that are greater than a given threshold and they have the same patterns as the top-score noun phrase or prepositional phrase.

3.3.4 Feature Selection

Feature selection for salience grading was implemented with a genetic algorithm which can get the best features by choosing best parents to produce offspring leave local maximum by mutation.

In the initial state, we chose features (10 chromosomes), and chose crossover feature to produce offspring randomly. We calculated mutations for each feature in each chromosome, and found about two features to be mutated in each generation. Max F-Score was used to evaluate each chromosome and top 10 chromosomes were chosen for next generation. The algorithm terminated if two contiguous generations did not increase the F-score.

3.4 Experiments and Analysis

The test corpus, Medstract, was adopted from (http://www.medstract.org/), containing 32 MEDLINE abstracts and 83 anaphora pairs (26 pronominal and 57 sortal pairs). For pronominal anaphora, we tagged another 103 MEDLINE abstracts (103-MEDEDLINSs) corpus which contains 177 pronominal anaphora pairs.

From the experimental results in Table 4, best fit strategy performed better than the nearest first strategy. In addition, the features selected by the genetic algorithm indicated that syntactic features affect pronominal anaphora, and semantic features will impacts on both sortal and pronominal anaphora.

Table 4: System result with best-first and nearest-first algorithm for Medstract.

	Best Fit		Nearest Fit		[Castano et al., 2002]	
	Sortal	Pronominal	Sortal	Pronominal	Sortal	Pronominal
Total Features	64.08%	88.46%	50.49%	73.47%		
Genetic Features	F5~F7	All-{F5}	F5~F7	All-{F2,F5}	F4~F6	F4, F6, F7
	78.26%	92.31%	61.18%	79.17%	74.4%	75.23%

F1: Recency, F2: Subject and Object preference, F3: Grammatical role Agreement, F4: Number Agreement, F5: Longest common subsequence, F6: Semantic type Agreement, F7: Biomedical Antecedent

The impact of each feature was also concerned and verified with the same corpus. Syntactic features (F1~F4) play insignificant roles in sortal resolution but they are useful for pronominal anaphora resolution. Sortal anaphora resolution are sensitive to semantic features (F5~F7), semantic type agreement plays an important role in sortal anaphora resolution. In addition to UMLS, headwords and PubMed search results were used to determine semantic type agreement between anaphor and antecedents. Table 5 shows F3 increases F-score in pronominal anaphora but drop F-score in sortal anaphora. Medstract and 103-MEDLINEs results show semantic type match is important in both sortal and pronominal anaphora. Table 6 shows F-score when removing headword and PubMed query result. Headword features show improvement in F-score because the semantic type of new words become precisely. PubMed query results improved little in F-score may because we only use co-occurrence information was concerned.

Table 5: Impact of each feature in pronominal and sortal.

	Medstract		103-MEDLINEs
	Sortal	Pronominal	Pronominal
All	64.08%	88.46%	85.88%
All – Recency (F1)	61.05%	73.08%	79.10%
All - Subject or Object preference (F2)	65.96%	88.00%	84.18%
All - Grammatical Role Match (F3)	72.00%	80.77%	80.79%
All - Number Agreement (F4)	64.65%	81.48%	85.88%
All – LCS (F5)	48.00%	92.31%	86.44%
All – Semantic Type Match (F6)	44.04%	88.46%	77.40%
All - Biomedical Antecedent (F7)	38.26%	59.26%	61.02%

Table 6: Impact of headword and PubMed.

	With Headword	Without Headword
With PubMed	78%	59%
Without PubMed	76%	58%

4 Conclusion

In this paper, pronominal and sortal anaphora which are common phenomenal in biomedical texts are discussed. The pronominal anaphora processing was achieved by syntactic and semantic features, while sortal anaphora was tackled by semantic features. For new biomedical entities to UMLS, we solve the entities semantic agreement by using headword mining and patterns mine from PubMed query results. Experiment results showed the proposed strategies indeed enhance the resolution in terms of higher F-Score.

Acknowledgement
This research is partially supported by MediaTek Research Center, National Chiao Tung University, Taiwan.

5 References

[1] Breck Baldwin, "CogNIAC: high precision coreference with limited knowledge and linguistic resources," *In Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, 1997, pp. 38-45.

[2] José Castaño, Jason Zhang, Hames Pustejovsky, "Anaphora Resoution in Biomedical Literature," *In International Symposium on Reference Resolution,* 2002

[3] Ido Dagan and Alon Itai, "Automatic processing of large corpora for the resolution of anaphora references," *In Proceedings of the 13th International Conference on Computational Linguistics (COLING'90), Vol. III, 1-3,* 1990.

[4] Michel Denber, "Automatic resolution of anaphora in English," *Technical report, Eastman Kodak Co.* , 1998.

[5] Udo Hahn and Martin Romacker, "Creating Knowledge Repositories from Biomedical Reports:The MEDSYNDIKATE Text Mining System, "*In Pacific Symposium on Biocomputing, 2002*

[6] J. Hobbs, "Pronoun resolution," *Research Report 76-1, Department of Computer Science, City College, City University of New York, August 1976*

[7] R. Gaizauskas, G. Demetriou, P.J. Artymiuk and P. Willett, "Protein Structures and Information Extraction from Biological Texts: The PASTA System," *In Bioinformatics 2003*

[8] Christopher Kennedy and Branimir Boguraev, "Anaphora for everyone: Pronominal anaphora resolution without a parser," *In Proceedings of the 16th International Conference on Computational Linguistics*, 1996, pp.113-118.

[9] Shalom Lappin and Herbert Leass, "An Algorithm for Pronominal Anaphora Resolution," *Computational Linguistics*, Volume 20, Part 4, 1994, pp. 535-561.

[10] Tyne Liang and Dian-Song Wu, "Automatic Pronominal Anaphora Resolution in English Texts," *In Computational Linguistics and Chinese Language Processing Vol.9, No.1, 2004, pp. 21-40*

[11] Ruslan Mitkov, "Robust pronoun resolution with limited knowledge, " *In Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference Montreal, Canada.* 1998, pp. 869-875.

[12] Ruslan Mitkov, "Anaphora Resolution: The State of the Art," *Working paper (Based on the COLING'98/ACL'98 tutorial on anaphora resolution)*, 1999.

[13] Ruslan Mitkov and Catalina Barbu, "Evaluation tool for rule-based anaphora resolution methods," *In Proeedings of ACL'01, Toulouse*, 2001.

[14] Ruslan Mitkov, Richard Evans and Constantin Orasan, "A new fully automatic version of Mitkov's knowledge-poor pronoun resolution method," *In Proceedings of CICLing- 2000, Mexico City, Mexico*.

[15] T. Ohta, Y. Tateisi, J.D. Kim, S.Z. Lee and J. Tsujii. "GENIA corpus: A Semantically Annotated Corpus in Molecular Biology Domain." *In the Proceedings of the ninth International Conference on Intelligent Systems for Molecular Biology (ISMB 2001) poster session. pp. 68. 2001.*

[16] James Pustejovsky, Anna Rumshisky, José Castaño, " Rerendering Semantic Ontologies: Automatic Extensions to UMLS through Corpus Analytics," *LREC 2002 Workshop on Ontologies and Lexical Knowledge Bases,* 2002.

[17] J. Pustejovsky, José Castaño, J. Zhang, B. Cochran, M. Kotecki, ” Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations.,” *In Pacific Symposium on Biocomputing, 2002*

利用自然語言處理技術自動產生英文克漏詞試題之研究

王俊弘† 劉昭麟† 高照明‡
†政治大學資訊科學系 ‡台灣大學外國語文學系
{g9124, chaolin}@cs.nccu.edu.tw zmgao@ntu.edu.tw

摘要

電腦輔助產生試題系統的研究熱潮正方興未艾，其研究目的在於節省人力以建置大規模的題庫，並進一步支援網路學習、成效評估與適性化測驗。受惠於來自網際網路上充裕的文字資源，吾人可以利用既有的語料來產生涵蓋各種不同主題的克漏詞試題，以增加題庫的多樣性。另一方面，由於電腦輔助產生試題系統減少人為的干預，也得以保持試題隱密性。我們提出一個詞義辨析的演算法，利用詞典與 selectional preference 所提供的資訊，分析試題的答案的詞義，並以 collocation 為基礎的方法篩選誘答選項。實驗結果顯示我們的系統可在每產生 1.6 道試題中，得到 1 道可用的試題。

Key Words

試題編寫工具與方法論、自動化產生試題、電腦輔助語言學習、詞義辨析、自然語言處理、collocations、selectional preferences

1 緒論

電腦輔助產生試題（computer-assisted item generation，CAIG）可提供題庫（item pools）所需求的各種特質，近年來已廣泛地引起國內外研究學者的重視 [2][6]。利用電腦高速運算的能力，電腦輔助產生試題的系統可產生大量且多樣化的試題，以提供評估學生學習能力的試題來源，也因而減輕了確保試題隱密性的問題 [11]。此外，隨著網路資料量的快速成長，我們可以搜尋並篩選網路上的文字資源作為試題的句子，有效率地產生大量的試題。在這篇論文中，我們即利用自然語言處理（natural language processing）的技術，從網路上的文字資源中有效率地產生克漏詞測驗試題（cloze item）。

自然語言處理的技術提供許多可行且有效的方式以產生英文克漏詞測驗試題。其中一種作法，是以句型樣版為基礎（template-base）的方法建立句子 [3]，或採用較複雜的方式以諸多前置條件來建立句子 [2]。另一套截然不同的演算法則是採用現有的語料庫，如 LDC <http://www.ldc.upenn.edu/> 與 OTA <http://ota.ahds.ac.uk/>，或自行建立的語料庫，從中選取合適的句子以產生試題。前者的方法提供了特定句型且語境容易得到控制的測驗試題，但相對地，一些檢查句子是否合理的複雜機制所需的成本要較後者來得高 [16]。因此，我們可以嘗試利用網路上提供大量的文字檔案，並嚴格過濾其中的文字，挑選出高品質的句子以供我們產生克漏詞測驗試題。測驗編撰者便能夠以相當低的代價從這些產生的試題中挑選合適可用的測驗試題。

一些研究學者已致力於應用自然語言處理的技術來構成語意完整的句子，以便產生多選題（multiple-choice）型式的克漏詞測驗試題。（為了簡單起見，以下用「克漏詞試題」或「試題」代表多選題型式的克漏詞測驗試題。）Johns [7] 與 Steven [17] 使用 concordance 與 collocation 的概念從一般化語料庫中產生試題。Coniam [1] 藉由統計標記化語料庫中詞的詞頻（word frequency）以產生特定類型的試題。在我們 2003 年的工作中，以網路為介面的環境來側寫與評估學生英文能力的情況下，則是利用網路上的英文語料為主要的克漏詞試題的來源 [4][18]。

然而，目前為止僅有少數進階的自然語言處理的技術被套用在產生試題上。例如，許多英文詞通常有多種詞義，而測驗編撰者通常希望在試題中測驗某一詞的特定的使用方法。在這種情況下，盲目地使用關鍵詞比對（keyword matching）的方法－如 concordancer，也許會導致我們得到一連串毫無用處的句子，因而提高人員在後續篩選試題的工作量。此外，要組成一道克漏詞試題不僅僅需要一個語意完整的句子。圖 1 顯示了一道克漏詞試題的範例，挖掉一個詞的句子稱為**題幹**（**stem**），被挖掉的詞即是該試題唯一的**答案**（**key**），而其他三個選項稱為**誘答選項**（**distractor**）。

1. My sister is _____, that is, I am going to be an uncle soon.
(A) supposing (B) assigning
(C) expecting (D) scheduling

圖1　一道英文克漏詞試題的範例

給定一個句子，我們仍需要誘答選項來組成一道試題。誘答選項的選擇影響到試題的**難易度**（**item facility**）與試題的**鑑別度**（**item discrimination**），是一項重要的工作 [12]。因此，誘答選項的選擇也需要更謹慎的策略，若只考慮以詞頻作為挑選誘答選項的依據 [1][4]，顯然不符合實際的需求。為了消弭這類型的缺失，我們使用了詞義辨析（word sense disambiguation）的技術，從語料庫中挑選含有指定詞義的答案的句子，並利用統計 collocation 與 selectional preference [10] 的技術來挑選誘答選項。實驗評估的結果顯示我們的方法能夠建立令人滿意的品質的試題，我們並實際運用系統產生的試題在大一程度的英文課程隨堂測驗之中。

我們在第 2 節概述產生克漏詞試題的流程，並在第 3 節解釋語料庫的準備方法，旁及詞典的介紹與使用。在第 4 節中我們詳述將詞義辨析應用到系統中以選擇句子，並且在第 5 節探究將 collocation 與 selectional preference 的應用套在產生誘答選項的策略上。對於系統的評估與相關的應用將在第 6 節提出。

2　系統架構

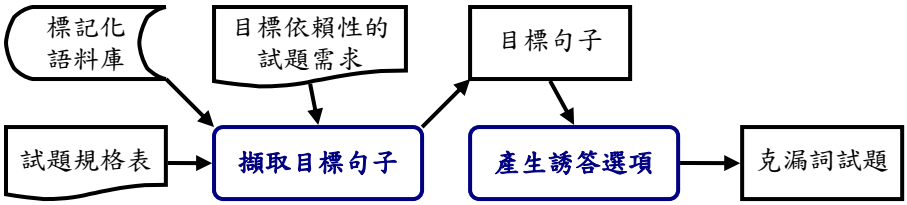


圖2　系統架構圖

圖 2 是自動產生克漏詞試題的系統架構圖。產生克漏詞試題涉及兩個主要的步驟，分別實作於兩個子系統中。**擷取目標句子**（**target sentence retriever**）子系統在測驗編撰者的要求與目標依賴性的試題需求的雙重限制下，從**標記化語料庫**（**tagged corpus**）中擷取克漏詞試題所需的句子。透過**試題規格表**（**item specification**）的介面，測驗編撰者可輸入試題的答案，並指定答案的詞性與詞義。圖 3 顯示試題規格表的介面。我們的系統將嘗試依試題規格表的要求以產生所需的試題量。**目標依賴性的試題需求**（**target-dependent item requirements**）具體地指定所有針對特殊測驗目標所設計的試題，必需遵循的一般化原則。舉例而言，在台灣大學入學考試（College Entrance Examinations）中，一道克漏詞試題所包含的詞數大致介於 7 個到 28 個詞之間 [18]，而測驗編撰者可依循這樣的傳統來建立測驗用的試題。此外，我們的系統允許測驗編撰者也可以在不指定答案的情形下，要求我們的系統產生特定數量且特定詞性的試題。

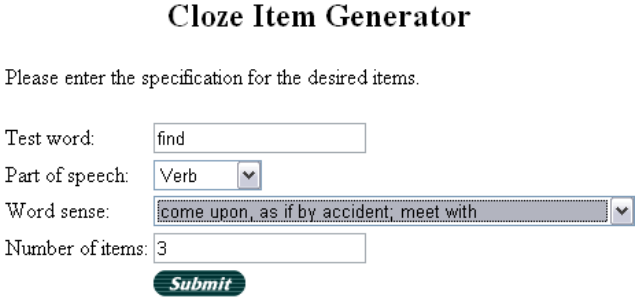


圖3　試題規格表的介面

在取得目標句子（target sentence）後，下一個步驟是由**產生誘答選項**（**distractor generator**）子系統考慮詞頻排名、collocation 與 selectional preference 等參考條件來篩選誘答選項。如果無法找到足夠的誘答選項（一般情形下是 3 個）以滿足答案與題幹的限制，系統會放棄這個目標句子而重新啟始整個產生克漏詞試題的程序。

3 語料來源與詞典

在搜集語料的工作上，我們利用網路爬梳器（web crawler）從 Taiwan Journal <http://taiwanjournal.nat.gov.tw>、Taiwan Review <http://publish.gio.gov.tw/FCR/fcr.html> 與 China Post <http://www.chinapost.com.tw> 定期抓取最新的文章。這些線上期刊與新聞報導提供高品質且更新速度快的文字資源，拼詞錯誤率極為罕見。目前在我們的語料庫中，共有 163,719 個句子，其中包含了 3,077,474 個詞次（word token）與 31,732 個詞型（word type）。

一份 HTML 格式的網頁文件含有各式各樣的多媒體內容。我們需要從混合了標題、主體文字、圖片與影音檔案的網頁內將主體文字的部份擷取出來，而擷取出的文字段落需經由斷句的步驟以切裁成個別的句子，作為試題題幹的來源。我們使用 Reynar 開發的 MXTERMINATOR 工具以完成斷句的工作。MXTERMINATOR 實驗於 Brown 與 Wall Street Journal 等著名的語料庫中有大約 97.5%的斷句正確率 [15]。我們接著對每個句子斷詞，以利於我們接下來對個別的詞加註有用的標記。

為語料庫中的詞標記各種資訊可提高產生克漏詞試題的效率。我們利用 Ratnaparkhi 的 MXPOST 工具為語料庫中的詞標記詞性，MXPOST 遵循 Pen Treebank 詞性集的標準 [13]。在標記詞性後，我們依每個詞的詞性標記其應有的詞根（lemma）。舉例而言，若 *classified* 的詞性是 *VBN*，我們標記其詞根為 *classify*；若詞性是 *JJ*，則標記成 *classified*。另外，我們也使用 Lin 的 MINIPAR [8] 標記句子中的慣用語。MINIPAR 能夠偵測 *arrive at* 與 *in order to* 等不可分的慣用語。這對於偵測可分的慣用語而言是不足夠的，我們也將極積尋求較佳的替代方案。

然而，使用 MINIPAR 最主要的目的在於它提供了局部語法剖析（partial parse）的功能，我們將之應用於產生克漏詞試題的系統中。一個句子中的詞與詞之間若有語法上的關係將會被 MINIPAR 所偵測，利用這些關係可輔助我們施行詞義辨析。為求簡便，對於詞 *w* 而言，句子中其他詞與 *w* 有語法上的關係，稱之為 *w* 的**信號詞**（**signal word**）或簡稱為**信號**（**signal**）。

由於歷屆大學入學考試英文科的克漏詞試題大都是以動詞、名詞、形容詞與副詞為測試標的 [18]，我們目前也著重於產生以這四種詞性作為答案的克漏詞試題，因此只對句子中屬於這些詞性的詞施行詞義辨析。對於動詞而言，其信號詞包括它的主詞、受詞與修飾它的副詞；對於名詞而言，其信號詞包括修飾它的形容詞或是將之視為主詞或受詞的動詞；舉例而言，在 *Jimmy builds a grand building* 一句中，*build* 與 *grand* 都是名詞 *building* 的信號詞；對於形容詞與副詞而言，其信號詞包括它們所修飾的詞與修飾它們的詞。

若產生試題的過程中需要詞典定義詞的資訊時，系統將訴諸於機器可讀取的電子詞典。當我們需要詞的詞義、同義詞與例句等資訊施行詞義辨析時，我們藉由 WordNet <http://www.cogsci.princeton.edu/wn/> 的輔助；當我們需要動詞、名詞、形容詞與副詞的類別（class）以統計 selectional preference 與 collocation 的程度時，我們仰賴 HowNet <http://www.keenage.com\verb> 的定義。

4 擷取目標句子

在圖 2 中，**擷取目標句子**子系統從語料庫中擷取高品質的句子。一個被視為候選目標句子（candidate target sentence）的句子必需包含指定的答案與詞性。藉由先前利用 MXPOST 對詞所標記的詞性，我們可以輕易地達成上述的要求。在有了候選目標句子後，試題產生器（item generator）需要決定答案的詞義是否符合需求。我們施行詞義辨析的演算法是建構在 selectional preference 的觀念上。

4.1 廣義的 Selectional Preference

利用 selectional preference 的輔助以施行詞義辨析的著眼點在於，在一般情形下，句子中某一特定詞的詞義，會受到句子中其他詞的種類所限制。Selectional preference 與詞義辨析之間的密切性可用一個簡單的例子加以說明，當名詞 *chair* 出現在句子 *Susan interrupted the chair* 中時，應是指一個人而並非傢俱 [10][14]。因此我們可以觀察句子中與某個多義詞（polysemous word）有語法關係的信號詞來猜測多義詞在該句子中所扮演的詞義。

我們可以仰賴 HowNet 的定義，將動詞對其主詞及受詞的偏好性，延伸到克漏詞試題的答案（詞性可能是動詞、名詞、形容詞或副詞）對其信號詞的偏好性。在統計 selectional preference 的強度時，以「詞對類別」的方式統計，令 w 與 π 分別代表詞與類別（定義在 HowNet 中），以 $f_v(w,\pi)$ 表示 w 與 π 共同參

與語法關係 v 的頻率，且 π 為 w 的信號詞的類別，並以 $f_v(w)$ 表示 w 參與語法關係 v 的頻率，不計其信號詞的類別。我們將 w 與 π 的 selectional preference 的強度以式子 (1) 表示：

$$A_v(w,\pi) = f_v(w,\pi) / f_v(w) \quad (1)$$

當語料庫中出現 w 參與 v 的情形下，不論 w 的信號詞為何，$f_v(w)$ 皆累加 1 次。令 s 為 w 在 v 關係下的信號詞，並以 $\Pi(s) = \{\pi_1, \pi_2, \ldots, \pi_y\}$ 表示 s 的類別集合，當語料庫中出現 w 與 s 共同參與 v 且 $\pi \in \Pi(s)$ 時，則 $f_v(w,\pi)$ 的計數加上 $1/y$ 。表 1 顯示 3 個英文的動詞 *eat*、*tell* 與 *find* 與其受詞 HUMAN、FOOD 兩個類別的統計資料。由表 1 可知，動詞 *eat* 對其受詞的偏好性，明顯地傾向類別 FOOD，與動詞 *tell* 恰好形成對比。

表1　Selectional preference 的部份統計資料

v = 動詞對受詞		動詞		
		eat	*tell*	*find*
類別	HUMAN	0.047	0.487	0.108
	FOOD	0.441	0.005	0.057

4.2　詞義辨析

我們藉由 4.1 節介紹的廣義的 selectional preference 與詞典 WordNet 的輔助，辨析候選目標句子中答案的詞義。為了避免造成混淆，本小節將以「關鍵詞」代表候選目標句子中欲施行詞義辨析的詞。若是關鍵詞在 WordNet 中僅具有一種詞義，詞義辨析演算法將會指派其唯一的詞義給予關鍵詞。反之，若是關鍵詞擁有多種詞義，以一個候選目標句子 *They say film makers don't spend enough time developing a good story* 為例說明詞義辨析的演算法。我們欲對句子中的動詞 *spend* 作詞義辨析，在 WordNet 的定義下，*spend* 有兩種詞義：

1. (99) spend, pass – (pass (time) in a specific way; “How are you spending your summer vacation?”)
2. (36) spend, expend, drop – (pay out; “I spend all my money in two days.”)

第一個詞義為 *pass (time) in a specific way*，第二個詞義是 *pay out*。WordNet 對每個詞義所包含的資訊包括 (Ⅰ) **標頭詞**（**head words**），由一個或數個詞組成，這些標頭詞共享該詞義；(Ⅱ) 該詞義所專屬的**例句**，展示該詞義的獨特用法。在之後關於詞義辨析的作法的討論中，每當提及關鍵詞的詞義的標頭詞時，我們不考慮關鍵詞為其本身某個詞義的標頭詞。因此，動詞 *spend* 的第一個詞義僅有 1 個標頭詞：*pass*，而第二個詞義有 2 個標頭詞：*extend* 與 *drop*。

一個對候選目標句子中動詞 *spend* 施行詞義辨析的直覺的方法，是以 *spend* 的標頭詞取代其在句子中的地位。正確詞義的標頭詞與其他詞義的標頭詞相較之下，套用在候選目標句子中應有較合理的語意。反之，若是語料庫中極少出現標頭詞取代後的句子，該標頭詞所屬的詞義就不太可能是關鍵詞的詞義。這項直覺引領我們計算一個詞義在標頭詞部份所應得的分數，也就是我們所表示的 Ω_t 。

此外，我們可以比較不同詞義的例句與候選目標句子中，*spend* 的上下文（context）的相似程度，在這裡所指的上下文，與 *spend* 的信號詞有密切關係。再次以句子 *They say film makers don't spend enough time developing a good story* 為例，我們可以比較 *spend* 在候選目標句子中主詞（makers）與受詞（time）的類別，是否與 *spend* 在例句中主詞與受詞的類別相同。若 *spend* 的第一個詞義的例句提供一個與 *spend* 在候選目標句子中較近似的上下文，則第一個詞義獲得較高的分數，反之則由第二個詞義獲得較高分。這項直覺引領我們得到詞義在例句部份所應得的分數，也就是我們將在下面介紹的 Ω_s 。

假設關鍵詞 w 在 WordNet 的定義下有 n 個詞義，令 $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ 是關鍵詞 w 的詞義集合。假設關鍵詞 w 的詞義 θ_i 在 WordNet 中有 m_i 個標頭詞。（注意我們並不考慮 w 本身為 θ_i 的標頭詞。）我們使用集合 $\Lambda_i = \{\lambda_{i,1}, \lambda_{i,2}, \ldots, \lambda_{i,m_i}\}$ 代表關鍵詞 w 的詞義 θ_i 的標頭詞集合。

當我們使用 MINIPAR 對一關鍵詞所屬的候選目標句子 T 作語法分析時，可以得到關於關鍵詞的信號詞的資訊。假設 w 在句子 T 中有 $\mu(T)$個信號詞。我們令集合 $\Psi(T,w) = \{\psi_{1,T}, \psi_{2,T}, \ldots, \psi_{\mu(T),T}\}$ 代表 T 中 w 的信號詞集合。同時，以 $\nu_{k,T}$ 代表 T 中 w 與 $\psi_{k,T}$ 的語法關係，並且以 $\Gamma(T,w) = \{\nu_{1,T}, \nu_{2,T}, \ldots, \nu_{\mu(T),T}\}$ 代表 w 與 $\psi_{k,T}$ 之間的語法關係的集合。

就關鍵詞 w 的第 i 個詞義 θ_i 而言，其第 j 個標頭詞 $\lambda_{i,j}$ 所獲得的分數，是利用式子 (1) 計算 $\lambda_{i,j}$ 與 T 中每個信號詞 $\psi_{k,T}$ 的 selectional preference 的強度，再求其平均值。

$$\frac{1}{\mu(T)}\sum_{k=1}^{\mu(T)} A_{v_{k,T}}(\lambda_{i,j},\psi_{k,T})$$

因此詞義 θ_i 所獲得的分數，即是 θ_i 所有標頭詞所獲得的分數的平均值。

$$\Omega_t(\theta_i \mid w,T) = \frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{\mu(T)}\sum_{k=1}^{\mu(T)} A_{v_{k,T}}(\lambda_{i,j},\psi_{k,T}) \tag{2}$$

我們用式子 (2) 計算候選目標句子 T 中關鍵詞 w 的詞義 θ_i 在標頭詞部份所獲得的分數。注意 selectional preference 的強度 $A_{v_{k,T}}(\lambda_{i,j},\psi_{k,T})$ 與標頭詞分數 Ω_t 兩者的數值皆落於 0 到 1 的範圍之內。

既然 WordNet 對許多詞義有提供例句，我們可以比較詞義的例句與候選目標句子之間的語境相似程度去判別候選目標句子中的關鍵詞的詞義。我們利用關鍵詞的信號詞決定句子的語境。令 T 與 S 分別為 w 所屬的候選目標句子與 w 的詞義 θ_i 的例句。我們將以下列 3 個步驟，計算詞義 θ_i 在例句部份所獲得的分數 Ω_s。如果一個詞義有多個例句，我們將用式子 (3) 對 θ_i 的每個例句計算分數，最後並對個別分數的總合值作平均。

計算 $\Omega_s(\theta_i \mid w,T)$ 的步驟

1. 分別求得關鍵詞 w 在 T 與 S 中的信號詞集合 $\Psi(T,w)$ 與 $\Psi(S,w)$，以及它們與 w 的語法關係集合 $\Gamma(T,w)$ 與 $\Gamma(S,w)$。

$$\Psi(T,w) = \{\psi_{1,T},\psi_{2,T},\ldots,\psi_{\mu(T),T}\}$$
$$\Psi(S,w) = \{\psi_{1,S},\psi_{2,S},\ldots,\psi_{\mu(S),S}\}$$
$$\Gamma(T,w) = \{v_{1,T},v_{2,T},\ldots,v_{\mu(T),T}\}$$
$$\Gamma(S,w) = \{v_{1,S},v_{2,S},\ldots,v_{\mu(S),S}\}$$

2. 我們尋找 T 中 w 的信號詞 $\psi_{j,T}$ 與 S 中 w 的信號詞 $\psi_{k,S}$ 使得 $v_{j,T} = v_{k,S}$，假設 $\psi_{j,T}$ 在 HowNet 中有 $n_{j,T}$ 個類別，以集合 $\Pi(\psi_{j,T}) = \{\pi_{j,T,1},\pi_{j,T,2},...,\pi_{j,T,n_{j,T}}\}$ 表示，而 $\psi_{k,S}$ 在 HowNet 中有 $n_{k,S}$ 個類別，以集合 $\Pi(\psi_{k,S}) = \{\pi_{k,S,1},\pi_{k,S,2},...,\pi_{k,S,n_{k,S}}\}$ 表示。對於 $\Pi(\psi_{j,T})$ 內的每個類別 $\pi_{j,T,l}$，逐一比對 $\Pi(\psi_{k,S})$ 中是否存在類別 $\pi_{k,s,m}$ 使得 $\pi_{j,T,l} = \pi_{k,S,m}$。每比對一組相同的類別，將累計分數 $M(\theta_i,T)$ 加上 $1/n_{j,T}$。

```
M(θ_i, T) = 0;
mark all v_{j,T} ∈ Γ(T, w) as unmatched;
for(j = 0; j < μ(T); j++)
    for(k = 0; k < μ(S); k++)
        if ((v_{j,T} unmatched) and (v_{j,T} = v_{k,S}))
        {
            mark v_{j,T} as matched;
            for(l = 0; l < n_{j,T}; l++)
                for(m = 0; m < n_{k,S}; m++)
                    if (π_{j,T,l} = π_{k,S,m})   M(θ_i, T) = M(θ_i, T) + 1/n_{j,T}
        }
```

3. 式子 (3) 度量關鍵詞在候選目標句子的信號詞與在例句中的信號詞，當與關鍵詞有相同語法關係的情形下，所得到的平均分數。

$$\Omega_s(\theta_i \mid w,T) = \frac{M(\theta_i,T)}{\mu(T)} \tag{3}$$

候選目標句子 T 中的關鍵詞 w 的詞義 θ_i 所獲得的分數，是由式子 (2) 所計算的 $\Omega_t(\theta_i \mid w,T)$（標頭詞所獲得的分數）與式子所 (3) 計算的 $\Omega_s(\theta_i \mid w,T)$（例句所獲得的分數）加總而得，若至少存在一個詞義的分數超過我們在式子 (4) 所選定的門檻值（threshold），候選目標句子 T 中的關鍵詞 w 就會被指派給分數最高的詞義。反之，若 $\Omega_t(\theta_i \mid w,T)$ 與 $\Omega_s(\theta_i \mid w,T)$ 的加總值太小，表示演算法的結果可信度不高，系統將不會作出任何草率的決定，並挑選其他候選目標句子，針對其中的關鍵詞重新起始詞義辨析的運作。我們將在 6.1 節闡明並討論選用不同的門檻值對詞義辨析的正確率所造成的影響。

$$\arg\max_{\theta_i \in \Theta_i} \Omega_t(\theta_i \mid w,T) + \Omega_s(\theta_i \mid w,T) \tag{4}$$

5 產生誘答選項

克漏詞試題中的誘答選項影響了學生幸運猜中答案的可能性。若試題中含有明顯可看出不可能是答案的誘答選項，學生也許能夠在不知道答案的情況下得知正確的答案。因此，我們需要選擇可以訴諸於填補這漏洞的誘答選項，並必需避免同一試題中有多重答案的情形發生。

有一些可想到的方法與可供選擇的方案是容易參考且實作的。答案的反義詞是一項選擇，但一般情形下學生將會忽視之。而誘答選項的詞性必需與答案一致，否則學生將很容套用基本的文法知識僅依詞性去選擇答案，而不需知道整個句子的語意。我們也可以考慮文化背景的影響。華文語系背景的學生較易受到具有相同中文譯詞的英文詞彙所干擾。雖然學習策略在大部份時間是有作用的，學生也許會發現要分辨有相似中文詞義的英文詞彙是困難的。因此，文化背景依賴性（culture-dependent）的策略，可將具有與答案相似中文詞義的英文詞作為誘答選項的考慮，但需準備具公信力的中英文雙語詞典與中文同義詞詞典。

為了有系統地產生誘答選項，我們使用詞的詞頻排名作為初步篩選誘答選項之用[12][18]。假設我們產生一道答案的詞性為 ρ 的試題，且詞性 ρ 在辭典中有 n 個詞次，而答案的詞頻排名在 n 個詞次中排第 m 個。我們從 n 個詞次中，在 $[m-n/10, m+n/10]$ 的詞頻排名範圍內隨機挑選詞作為候選誘答選項。我們限制誘答選項的詞頻排名需與答案相近。接著檢驗這些候選誘答選項與題幹的**合適度**（**fitness**），以從中篩選誘答選項。合適度是由候選誘答選項的類別與題幹中其他詞的類別的 collocation 值而決定，詞的類別同樣定義在 HowNet 中。

在第 3 節曾提及，我們已對語料庫中的詞標記上它們的信號詞。句中的某一詞若擁有較多的信號詞，通常該詞對句子的語意貢獻較多，所以也應當在選擇誘答選項時佔有較重要的地位。既然我們並非真正檢視整個目標句子的語意，一個相對上挑選誘答選項較安全的方法，是選擇很少與重要的詞一起出現在句子中的詞。

令 $T = \{t_1, t_2, \ldots, t_q\}$ 代表題幹的詞的集合（亦即不包括答案在內）。我們從 T 中依下列兩個條件過濾出**重要詞**：(Ⅰ) 詞性為動詞、名詞、形容詞或副詞之一者，且 (Ⅱ) 該詞擁有兩個（含）以上的信號詞或被某一子句（clause）所修飾。令 $T' \subset T$ 為句子 T 中重要詞的集合，$T' = \{t_1', t_2', \ldots, t_{q'}'\}$，我們計算候選誘答選項 κ 與每個重要詞的 pointwise mutual information，並求其平均值。假設 $C = \{S_1, S_2, \ldots, S_{\aleph}\}$ 為語料庫中所有句子的集合，$\Pi(\kappa)$ 與 $\Pi(t_i')$ 分別為候選誘答選項 κ 與重要詞 t_i' 的類別集合，我們定義 $\Pr(\Pi(\kappa))$ 為語料庫中存在 S_η 包含一詞 w 且 $\Pi(w)$ 與 $\Pi(\kappa)$ 的交集不為空集合的比例；同理 $\Pr(\Pi(t_i'))$ 為語料庫中存在 S_η 包含一詞 χ 且 $\Pi(\chi)$ 與 $\Pi(t_i')$ 的交集不為空集合的比例。

$$\Pr(\Pi(\kappa)) = \frac{1}{\aleph}\{\sum_{S_\eta} 1 \mid S_\eta \text{ contains } w \text{ and } \Pi(w) \cap \Pi(\kappa) \neq \phi\}$$

$$\Pr(\Pi(t_i')) = \frac{1}{\aleph}\{\sum_{S_\eta} 1 \mid S_\eta \text{ contains } \chi \text{ and } \Pi(\chi) \cap \Pi(t_i') \neq \phi\}$$

除此之外，定義 $\Pr(\Pi(\kappa),\Pi(t_i'))$ 為語料庫中存在 S_η 同時包含 w 與 χ 且 $\Pi(w)$ 與 $\Pi(\kappa)$ 的交集、$\Pi(\chi)$ 與 $\Pi(t_i')$ 的交集皆不為空集合所佔的比例。

$$\Pr(\Pi(\kappa),\Pi(t_i')) = \frac{1}{\aleph}\{\sum_{S_\eta} 1 \mid S_\eta \text{ contains } w, \chi \text{ where } w \neq \chi \text{ and } \Pi(w) \cap \Pi(\kappa) \neq \phi \text{ and } \Pi(\chi) \cap \Pi(t_i') \neq \phi\}$$

候選誘答選項 κ 與題幹的適合度的定義如式子 (5)。

$$f(\kappa) = \frac{-1}{q'} \sum_{t_i' \in T} \log \frac{\Pr(\Pi(\kappa),\Pi(t_i'))}{\Pr(\Pi(\kappa))\Pr(\Pi(t_i'))} \tag{5}$$

若 $f(\kappa)$ 的值高於 0.3，則可被接受成為誘答選項。為了讓較低的 collocation 得到較高的分數，我們將整個式子加上一個負號。設定門檻值為 0.3 的原因是基於式子 (5) 對 220 筆訓練資料統計後得知，這份訓練資料搜集自 1992 年到 2003 台灣的大學入學考試英文科的克漏詞試題。

6 評估與應用

6.1 詞義辨析

詞義辨析在自然語言處理的研究中是一項被廣泛探索與討論的課題 [10]。不同的演算法在異質的環境下使用相異的評估方法，使得詞義辨析的正確率落於 40%到 90%的大範圍內 [14][19]。主觀地比較不同演算法之間的優劣，若不依賴像 SENSEVAL 一個共同比較的基礎，並非一項簡單的工作，因此在本論文中只回報我們的實驗結果。

表2　詞義辨析正確率

關鍵詞的詞性	基準	門檻值 = 0.4	門檻值 = 0.7
動詞	38.0%(19/50)	57.1%(16/28)	68.4%(13/19)
名詞	34.0%(17/50)	63.3%(19/30)	71.4%(15/21)
形容詞	26.7%(8/30)	55.6%(10/18)	60.0%(6/10)
副詞	36.7%(11/30)	52.4%(11/21)	58.3%(7/12)

實驗材料是從語料庫中選取 160 個句子，針對每個句子選定其中一個關鍵詞作詞義辨析。這些關鍵詞包含了 50 個動詞、50 個名詞、30 個形容詞與 30 個副詞，這 160 個關鍵詞在 WordNet 的定義中，詞義數量介於 2 個（如名詞 *verification* 與形容詞 *frightened*）到 19 個（如動詞 *have*）之間－亦即測試用的關鍵詞皆屬於多義詞，每個關鍵詞平均有 4.85 個詞義。我們將這 160 個關鍵詞交由系統施行詞義辨析，並由人工判斷詞義辨析的正確率，正確率顯示於表 2。

基準（**baseline**）一欄所顯示的正確率，是當我們總是選擇該關鍵詞最常見的詞義，而一個關鍵詞最常見詞義則是仰賴 WordNet 所提供的資訊。其右兩欄顯示我們套用式子(4)，設定不同的門檻值（分別是 0.4 與 0.7）所得到的正確率。如同我們在 4.2 節所提到的，當門檻值提高時，將會留下較多的關鍵詞無法做詞義辨析（因主觀認定詞義辨析的可信度不足以採信），因此門檻值的選擇直接影響了詞義辨析的正確率。不令人意外地，選用較高的門檻值會得到較高的正確率，但同時卻增加了退回率（rejection rate）。所幸語料庫可以不斷擴充以容納更多的句子，我們可專注於提高詞義辨析的正確率，在退回率上稍作犧牲。

我們注意到 WordNet 並非對每個詞的詞義都提供例句，當一個詞義沒有任何例句時，這個詞義將得不到任何例句方面所提供的分數，也就是 Ω_s 得到 0 分。在我們目前相當倚重 WordNet 提供的例句的情形下，所造成必然的結果是：系統不傾向將一個詞分派給沒有例句的詞義。這點在我們目前的設計中是一項明顯的缺失，但事實上，這個問題在對於產生試題的系統並非嚴重且無解的。一個多義詞常用或重要的詞義通常會配有一個以上的例句，所以詞義辨析的問題並不常發生。若我們想要完全避免這個問題，可以客製化 WordNet，使得常見的詞的所有詞義皆有專屬的例句。

6.2 產生克漏詞試題

圖 4 顯示圖 3 給定的條件的輸出。當系統依需求而產生一些試題後，測驗編撰者可以從中挑選最佳的數道試題以供實測。

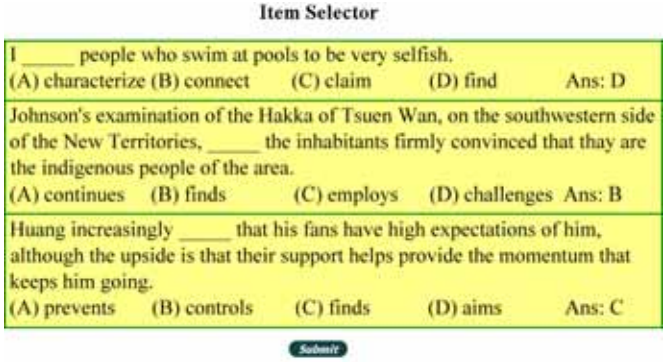


圖4　依圖 3 的規格所產生的試題

在評估階段我們要求試題產生器（item generator）產生 200 道試題。為了模擬真實情況下測驗題型的分布，我們對於以動詞、名詞、形容詞與副詞為答案的試題，個別分配不同的題數。參考先前的研究成果 [18]，我們從 1992 年到 2003 年台灣的大學入學考試的英文克漏詞試題中，歸納出答案的詞性不外乎 4 種，所佔的比例分別為：動詞 35%、名詞 30%、形容詞 20%與副詞 14%。因此，我們依相似的比例選用 77 題動詞、62 題名詞、35 題形容詞與 26 題副詞做為答案以評估系統在擷取目標句子上的效能。

在評估的過程裡，我們檢查產生的試題的答案，其詞義與詞性是否能夠符合需求。我們使用式子 (4) 並設定門檻值為 0.7，對試題中的答案作詞義辨析。實驗結果顯示在表 3，事實上，結果與表 2 的差距並不大。因為在標記詞性的正確率相當高的情況下，表 3 的實驗結果主要仍受到詞義辨析的正確率的影響。不論對於何種詞性的答案而言，在每產生小於 2 道試題之中，就有 1 道試題的答案能符合所要求的詞義與詞性。

此外，我們亦依照 [18] 對四種詞性的試題所歸納的比例，由系統產生 200 道克漏詞試題以檢驗誘答選項的品質，並判斷這些產生的試題是否能確保 4 個選項中僅有 1 個是正確選項（即答案本身）。由於試題是由語料庫中任意篩選出來的句子，所以我們是由人工判斷試題是否有唯一的答案，而以四個選項只有一個可以作為答案的題目來計算正確率。表 4 顯示，我們的系統在大多數情形下，能夠符合多選題的基本要求。

由於詞義辨析的工作相當具有挑戰性，而試題產生器能回傳大量的候選試題供測驗編撰者挑選，我們認為這套系統足以實際用於教師準備測驗卷的輔助工具。

表3　擷取目標句子的正確率

試題類型	答案的詞性	試題數量	目標句子的正確率
克漏詞試題	動詞	77	66.2%
	名詞	62	69.4%
	形容詞	35	60.0%
	副詞	26	61.5%
		總結	65.5%

表4　產生誘答選項的正確率

試題類型	答案的詞性	試題數量	誘答選項的正確率
克漏詞試題	動詞	64	90.6%
	名詞	57	94.7%
	形容詞	46	93.5%
	副詞	33	84.8%
		總結	91.5%

6.3 更多的應用

我們已將自動產生試題的系統實際應用於政治大學大一英文課程的隨堂測驗，並整合試題產生器到網路英文學習系統的環境中 [4]。在這個系統中，我們有兩項主要的子系統：測驗編撰系統與線上測驗系統。使用測驗編撰系統，測驗編撰者可以從圖 4 的介面中挑選試題，並將試題存放到測驗卷中，之後可依個人需求編輯測驗卷中的試題，包題幹、答案、誘答選項與正確選項等，測驗編撰者可對試題獲得最大的控制權。當測驗卷編輯完成後，輸入測驗卷的標題即可製成一份線上測驗卷。使用線上測驗系統，學生可透過網路進行線上測驗，並能夠立即獲得系統回報的成績（如果測驗編撰者有開放此功能）。學生的作答情形會記錄在學生模型（student modeling）中，系統將利用這些資料分析試的題難易度與試題的鑑別度。

為了支援不同題型的克漏詞測驗，我們也開發了產生慣用語（idiom）試題與片語（phrase）試題的系統。圖 5 描繪了這部份功能的輸出情形。更進一步地，我們的系統提供學生英文聽寫能力的測驗 [5]。在可見的未來內，我們計劃擴展我們的系統以支援全方面英文學習的需求（聽、說、讀、寫），並適性學生的能力以加強我們系統 [9]。

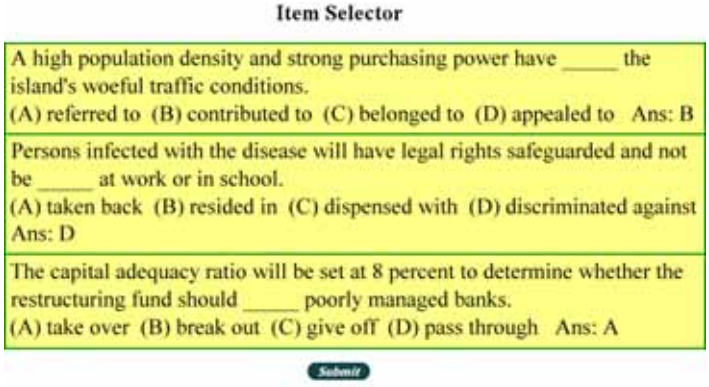


圖5　英文片語試題的範例

7 結論

在本論文中，我們提出了以自然語言處理技術為基礎的方法，可依特定需求而產生克漏詞試題，對於測驗編撰者在編寫測驗卷上有相當的助益。藉由將詞義辨析的演算法加入產生試題的流程中，使得產生的克漏詞試題能夠包含具有指定詞義的答案。詞義辨析本身並不是件容易的工作，在自然語言處理的相關研究中已研究了數年。對於一個目標句子中的答案而言，雖然我們的方法並不能對其做到最佳的詞義辨析，但是在產生試題的工作上已提供一個重要的幫助。畢竟，語言學家與心理學家普遍認為，詞義辨析仍需要來自上下文甚至是整個段落的資訊，而並非僅僅是單一個句子，而在單一句子做詞義辨析並非不可能，只是有較高的難度 [10]。

我們也提出一個新的策略來挑選克漏詞試題中的誘答選項。利用 collocation 為基礎的方法與詞頻統計資料，我們能挑選與答案具有相似挑戰性的誘答選項，並確保產生的試題中，有 90%以上的試題，答案是唯一或最佳選項。

既然測驗編撰者可以要求我們的系統傳回大量的克漏詞試題，並從中挑選數道最合其意的試題以編入測驗卷中，我們並不需要建立一個完美的電腦輔助試題編寫系統。目前我們的系統能夠做到每產生 1.6 道試題中，就有 1 道能夠用於實測的試題。然而，我們意圖考慮較深入的語言特徵來改進詞義辨析的正確率，以增進我們的系統的效能，並可望在不久的將來更新我們的實驗結果。

我們的研究仍可朝三個大方向持續發展。其一是句子的分析從語法的層面深入到語意的層面，檢驗句子是否含有完整的語義，藉以提高試題的品質；其次是從學生作答的情形中，歸納出控制試題難易度的因子，期許系統能大致猜得試題的難度；最後，我們希望能由英文教師的觀點，檢視系統的效能，比較教師在使用本系統後，是否在出題效率上有顯著的提升。

致謝

我們感謝多位不具名的評審對本文原稿的指正和建議，雖然因為篇幅和一些其他的限制使得我們暫且不能完全依照評審的建議加入新的材料，不過我們會在未來參照評審的珍貴建議加強論文的內涵。本研究承蒙國家科學委員會資助之研究案 91-2411-H-002-080 和 92-2213-E-004-004 的部分補助，謹此致謝。

參考文獻

[1] D. Coniam, A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests, *Computer Assisted Language Instruction Consortium*, **16** (2–4), 1997, 15–33.

[2] P. Deane, K. Sheehan, Automatic item generation via frame semantics, Education Testing Service: http://www.ets.org/research/dload/ncme03-deane.pdf (2003).

[3] I. Dennis, S. Handley, P. Bradon, J. Evans, S. Nestead, Approaches to modeling item generative tests, in: *Item Generation for Test Development* [2] 53–72, 2002, 53–72.

[4] Z.-M. Gao, C.-L. Liu, A Web-based assessment and profiling system for college English, *Proc. of the 11th Int'l Conf. on Computer Assisted Instruction*, 2004, CD–ROM.

[5] S.-M. Huang, C.-L. Liu, Z.-M. Gao, Toward computer assisted learning for English dictation, *Proc. of the 2003 Joint Conf. on Artificial Intelligence, Fuzzy Systems, and Grey Systems*, 2003, CD–ROM.

[6] S. H. Irvine, P. C. Kyllonen (Eds.), *Item generation for test development* (Lawrence Erlbaum Associates, 2002).

[7] T. Johns, http://web.bham.ac.uk/johnstf/timcall.htm.

[8] D. Lin, Dependency-based evaluation of MINIPAR, *Proc. of the Workshop on the Evaluation of Parsing Systems in the 1st Int'l Conf. on Language Resources and Evaluation*, 1998,.

[9] C.-L. Liu, Using mutual information for adaptive student assessments, *Proc. of the 4th IEEE Int'l Conf. on Advanced Learning Technologies*, 2004, to appear.

[10] C. D. Manning, H. Schütze, *Foundations of statistical natural language processing* (MIT Press, 1999).

[11] A. Oranje, Automatic item generation applied to the national assessment of educational progress: Exploring a multilevel structural equation model for categorized variables, Education Testing Service: http://www.ets.org/research/dload/ncme03-andreas.pdf (2003).

[12] C. J. Poel, S. D. Weatherly, A cloze look at placement testing, *Shiken: JALT (Japanese Assoc. for Language Teaching) Testing & Evaluation SIG Newsletter*, **1** (1), 1997, 4–10.

[13] A. Ratnaparkhi, A maximum entropy part-ofspeech tagger, *Proc. of the Conf. on Empirical Methods in NLP*, 1996, 133–142.

[14] P. Resnik, Selectional preference and sense disambiguation, *Proc. of the Applied NLP Workshop on Tagging Text with Lexical Semantics: Why, What and How*, 1997, 52–57.

[15] J. C. Reynar, A. Ratnaparkhi, A maximum entropy approach to identifying sentence boundaries, *Proc. of the Conf. on Applied NLP*, 1997, 16–19.

[16] K. M. Sheehan, P. Deane, I. Kostin, A partially automated system for generating passage-based multiple-choice verbal reasoning items, paper presented at the Nat'l Council on Measurement in Education Annual Meeting (2003).

[17] V. Steven, Classroom concordancing: vocabulary materials derived from relevant authentic text, *English for Specific Purposes*, **10** (1), 1991, 35–46.

[18] C.-H. Wang, C.-L. Liu, Z.-M. Gao, Toward computer assisted item generation for English vocabulary tests, *Proc. of the 2003 Joint Conf. on Artificial Intelligence, Fuzzy Systems, and Grey Systems*, 2003, CD–ROM.

[19] Y. Wilks, M. Stevenson, Combining independent knowledge sources for word sense disambiguation, Proc. of the Conf. on Recent Advances in NLP, 1997, 1–7.

An Infrastructure for Creating Web Automation Applications

Wen Heng Yen[1], Hao-Ren Ke[2], and Wei-Pang Yang[3]

[1]Degree Program of Electrical Engineering and Computer Science,
National Chiao Tung University,
1001 Ta Hsueh Rd., Hsinchu, TAIWAN 30050, R.O.C.
helio@lib.nctu.edu.tw
[2]University Library, National Chiao Tung University,
1001 Ta Hsueh Rd., Hsinchu, TAIWAN 30050, R.O.C.
claven@lib.nctu.edu.tw
[3]Dept. of Computer & Information Science, National Chiao Tung University,
1001 Ta Hsueh Rd., Hsinchu, TAIWAN 30050, R.O.C.

Dept. of Information Management, National Dong Hwa University,
1, Sec. 2, Da Hsueh Rd., Shou-Feng, Hualien, TAIWAN 97401, R.O.C.
wpyang@cis.nctu.edu.tw

Abstract. With the growth of the World Wide Web (WWW), many people nowadays spend a lot of time performing various tasks with browsers, most of these tasks repetitive and tedious. Many applications are created to reorganize and simplify the usage of Web resources, which are called Web automation applications. This paper proposes an infrastructure to create Web automation applications, the WIS (Web Integration Solution) system. WIS integrates a diversity of tools and technologies to provide a software environment for developing Web automation services. Developers can use this tool to create various Web automation applications. We show the versatility of WIS by implementing three exemplary services. The first is a metasearcher system that provides a single search interface for several indexing and abstract databases. The second is a cataloguing tool to simplify the task of retrieving bibliographic data from the Web; this tool is also integrated with a book-recommendation system for libraries. The third is a Web site checking system that periodically checks if some critical Web sites are working correctly, no matter how complex the check procedure is.

Keywords: User Surrogates; Web Agent; Web Automation

1 Introduction

The World Wide Web (WWW) provides a vast amount of information and plentiful services, and continues to grow at a staggering rate. Because of its explosive growth, people are spending more and more time on the Web, performing various tasks, many of these tasks repetitive and tedious. The following is some typical scenarios:

- I daily check several sources, including online news, Usenet newsgroups, and bulletin board systems (BBS) for some specific topics.
- Some WWW sites recommend several books to me, but I don't know where the best places to retrieve them are.
- I want to buy a second-hand notebook that fits some requirements. There are several places where I could find them and I don't want to miss any good buy.

From the above scenarios, it can be concluded that: (1) many tasks are repetitive; (2) a user may not know where to start surfing in the Internet; (3) many tasks involve several resources in different sites. What we need in these scenarios are agents, or Web automation applications, which act as our surrogates to perform these laborious tasks. A good surrogate should have the following benefits: (1) let nontechnical users be able to exploit the information available on the Web without being overwhelmed by technical detail; (2) free users from repetitive browsing tasks; (3) reformat and recombine the information from various Web sites to best fit a user's task.

There are many applications that use data from the Web, but they are usually developed for specific purposes. Instead of creating from scratch every time we need a new type of Web automation application, it will be helpful to have a tool specialized for the creation of such kind of applications. This paper (1) identifies the

common needs of Web automation applications; (2) creates an infrastructure, which is called WIS, that integrates these essentials together in a single tool suitable for the creation of a wide range of Web automation applications; (3) provides technological solutions for the challenges imposed by Web automation tasks; (4) constructs some applications to show the feasibility of this architecture.

2 Related Work

We define Web automation as user surrogates operating on existing Web resources to simplify users' tasks. Many applications have been created to act as user surrogates, and we discuss some related works to clarify the scope of what we mean by Web automation.

Search is a very common task in the Web. A metasearcher is an application that helps users perform search on multiple search engines. It provides a single interface for the multiple target search engines, and lets the user search these targets simultaneously with the same query. Many types of metasearchers exist today, with different capabilities (search ability, results display, and so on) and purposes. The Metacrawler [1] is a metasearcher that searches general-purpose Web search services such as Lycos and Google. Our previous works, the Unisearch [7] and VUCS, search different target collections: the first abstract online databases, and the second online public access catalogs of multiple libraries. Metalib [15] is a library portal that lets institutions manage hybrid information resources under one umbrella. The resources that may be managed by MetaLib include library catalogs, reference databases, digital repositories, and subject-based Web gateways. Metasearchers are not necessarily to work only in the server side; some are created as desktop applications. The Copernic Agent [2] is an example of such a desktop metasearcher application. As source Web sites tend to change over time in many aspects, profiles of Web sites need to be updated to keep them accessible. Installed Copernic Agents keep up to date by downloading these profiles periodically from the Copernic Web site.

Price comparison sites collect data from various online stores and produce a report for the buyer about the interested good. BestWebBuys [14] covers several kinds of products such as books, music, video, electronics and bikes, and uses dozens of stores as sources. BestBookDeal [13] focuses only on books, searching and comparing prices among 61 online bookstores to make sure that the buyer gets the best price. It claims to get real-time information about books, pricing, shipping cost, shipping time, sales tax and availability, saving the user from searching every online bookstore.

Many business systems are available for transforming the Web browser from an occasionally informative accessory into an essential business tool. Business organizations that have previously been unable to agree on middleware and data interchange standards for direct communication are agreeing on communication through HTTP and HTML, which needs human intervention (Figure 1). The need of manual operation may become highly inefficient when a lot of transcription or copy-and-paste operations are part of the daily job. The goal of the Web Interface Definition Language (WIDL) [6] is to enable automation of interactions with HTML/XML documents and forms, accepting the Web to be utilized as a universal integration platform without efficiency problems.

WIDL uses the XML standard to define interfaces and services, mapping existing Web content into program variables, allowing the resources of the Web to be made available for integration with business systems. It brings to the Web what is similar in IDL concepts that were implemented in standards such as CORBA for distributed computing. WIDL describes and automates interactions with services hosted by Web servers on intranets, extranets and the Internet.

3 The WIS Platform

3.1 Common Needs of Web Automation Applications and Solutions

The main concept of Web automation is to reuse existing Internet services. For example, we can put online bookstore services, library online services and inter-library loan services together to create a powerful solution for providing books of any kind; and empowered by Web automation, it will not have only descriptions and links such as traditional portal sites, but also really cooperating together. Continuing this example, which may be called a book agent, the user uses the metasearcher capability of the agent to search every potential provider of the book, no matter the provider is a library or a bookstore. The agent may create two groups of providers having the wanted book, one of libraries and one of bookstores. If the user selects the group of libraries, the data is passed to an inter-library loan system to acquire the book. If the bookstore group is selected, the agent may propose the best buy by comparing the prices and conveniently send the request for the user. As we can see in the example, entire services are reused, creating an altogether new experience for acquiring a book.

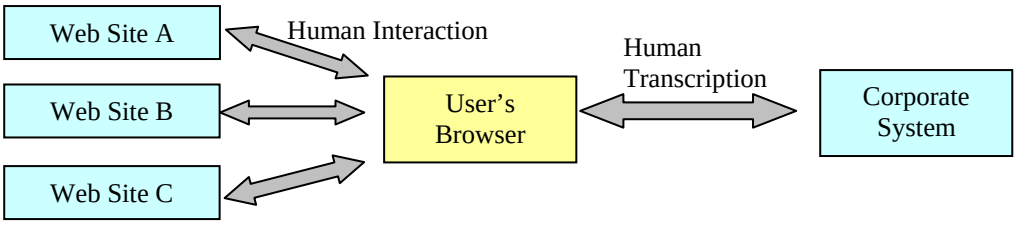


Fig.1. The need for Web automation

To reuse services, the first issue that needs to be solved is interoperability. Many protocols have been proposed for interoperation. The increasing interest in the metasearcher area introduces various protocols, such as Z39.50, OAI, and OpenURL. But there are always services that do not support these protocols, since their orginal purpose is for interaction with human users, not machines. The result is that the only thing we may be sure about Web services is that they use well-known Web technologies that a competent Web browser will surely do its job. The most common is the HTTP protocol and the HTML presentation language. Additional mechanisms such as cookies, security and scripts make the picture of the common interoperation even more complicated. But to meet the goal that the ideal Web automation tool needs to be of common purpose, these are the only things that can be relied.

By choosing to only use common Web technologies, a problem arises: the user interface of Web services usually changes with the time. For example, many sites have advertisements that by their nature change frequently. For the human user it is not a big problem because he/she can understand their meaning and skip them. But it is difficult for a computer to really "understand" the interface of a Web service. Intelligent solutions are hard to be designed for general purpose. WIDL instead provides a language to describe the position of the required data with more chance to skip unwanted changes. The WIS infrastructure uses the WIDL solution, with some modifications.

Parallel processing is a common need of Web automation applications. In a typical metasearcher for example, when the user submits a query, the query is dispatched to various sources at the same time, so that every source can do its job in parallel with others. When a source terminates its job, the results are returned and the agent can do further processing while there may still be sources performing the request.

From the metasearcher example, it seems that only the main process needs to have the privilege of creating other process. But there may be cases in which a process may need to create sub-processes, and messaging between any one of the processes is needed. The WIS system supports any arbitrary arrangement of processes due to its various scopes. Scopes are to be discussed in detail in the next sections.

In a Web automation application, the need to interoperate with various sources at the same time may consume the resources of a server significantly. In a three-tier architecture, the business-logic tier accomplishes the Web automation tasks (Figure 2). For example, in our previous work, the VUCS system, the automation component is implemented as a DCOM object. When a user performs a search, the interface program requests the Web automation DCOM object, which interoperates with the target resources. To scale up to a large number of users, computing power can be increased by adding more servers in the business-logic tier; but the network may eventually become the bottleneck of the system when multiple users are performing Web automation tasks in the server.

The solution provided by WIS for this problem is to permit Web automation tasks being performed at the client side, which reduces the load in the business-logic tier. From Figure 3 we can see that WIS replaces the position of the Web browser. It is especially efficient when users are widely spread in the network because the Web automation task will mainly spend local network bandwidth, saving the bandwidth of the server. But for this architecture to work, some issues need to be solved.

By removing the Web automation task from the business-logic tier, a new problem arises. When in the business logic tier, the Web automation component could easily work with other components and interact with the data tier. For example, in our book recommendation system, the Web automation's task involves reading hyperlinks stored in the database, extracting the metadata from the online bookstores, and writing the new information back to the database. WIDL is mainly designed to run at the server, integrated with the enterprise system, and any protocol could be used since it needs the supplement of a traditional programming language. But with the Web automation task moved to the client side, there needs to be a way to keep the interaction with the enterprise system. Web service technology plays this role by exporting functionality from the enterprise system to WIS. Whenever the Web automation needs support from the enterprise system, it acquires interface information by the WSDL [10] (Web Service Description Language) document and then with the definition, functions at the server side can be called by using SOAP [9] (Simple Object Access Protocol) messages. Thus the Web automation task running on WIS can access whatever function it needs from the server, from anywhere in the network.

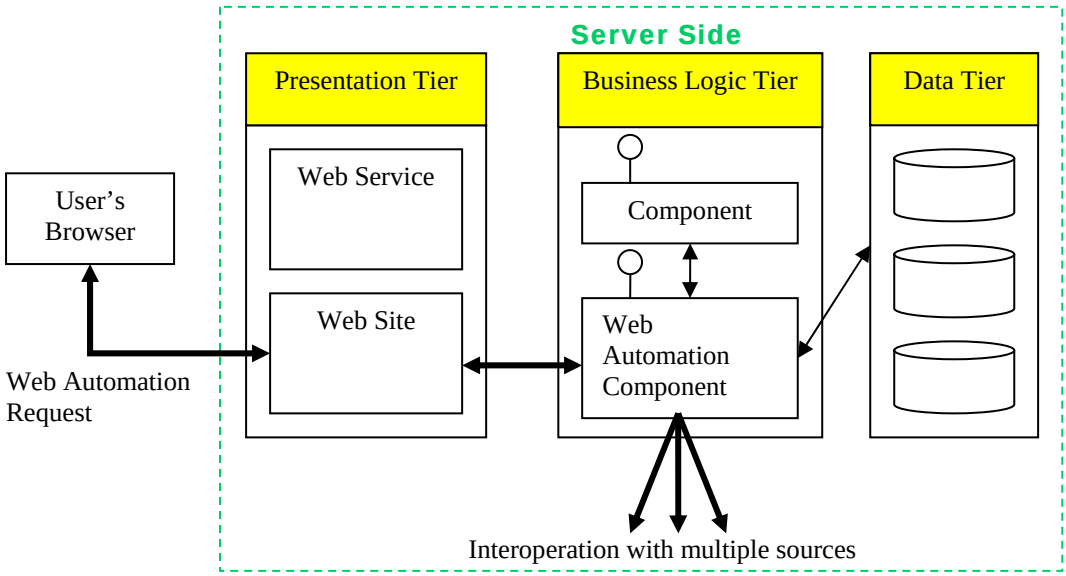


Fig.2. Web automation in a three-tier model

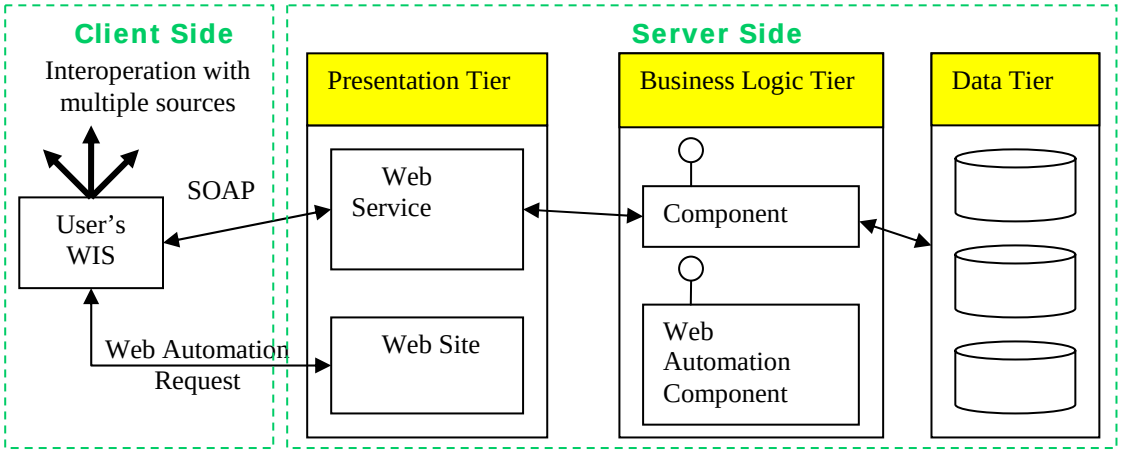


Fig. 3. Three-tier model of WIS

The business logic issue in the WIS architecture is solved, but there is still problem with the interface. In the traditional three-tier model, modifications to interfaces can be easily done at the presentation tier. When the user requests a Web automation task from the server, an entire new page is passed back to the browser whenever the server finishes it. But when entire Web automation tasks are running at the client side, how can WIS create whole new pages to show the progress or changes, which varies a lot from application to application? A solution would be to go back to the two-tier model, where clients are designed for specific applications, making the business logic work tightly with the presentation in the desktop. With this approach, WIS will come with different tailored interfaces for different applications, and the maintenance nightmare of updating hundreds or thousands of desktops comes back. Another solution is to use HTML interfaces instead of hard-coded interfaces, which are downloaded from the server. To refresh the interface, the Web automation process can request the server to construct a new page for it according to the parameters given and send it back. In this way WIS can keep its generality and compliance with the three-tier model. Better performance can be achieved by using DHTML, preventing the need of the round-trip every time the interface needs a refresh. The Web automation process can notify the user about any change without having to reload entire pages from the server. Any kind of report can be created by this way.

Many Web automation tasks are repetitive and need to be executed periodically. In the Website checking application, the check task is performed by intervals determined by the site manager. The solution provided by WIS comes from DHTML, which provides timed function call.

Data returned from sources needs further processing before presenting them to the user. Users' requests to the Web automation application need processing before sending to sources. In a metasearcher for example, the query given by the user is translated to a format that the target resource accepts, which may be different for

every resource. Results from every resource may have many differences, such as different date formats and lack or availability of some fields. Reranking will need even further computing. The WIDL, which plays only the role of an interface definition language, let this work to the complementary programming language. Because we did not figured out a single tool that can process so many variations in data processing of different automation applications, WIS uses a common programming language and allows the developer using JavaScript or Java applets to perform the transformation task.

With the architecture described above, WIS moves the Web automation task from the server to the client, distributing even more the workload without affecting the convenience of a counterpart browser. WIS substitutes the browser while still having the advantage of being a common client for the various applications. There is no need to have a special version of WIS for every application since it is designed with the concern of supporting any application, keeping the original idea of a common client that simplifies the deployment of new applications.

3.2 The WIS Infrastructure

A WIS application is composed of several pieces working together to accomplish the Web automation needs, which are called WIS components. WIS components are located at the server and downloaded to the WIS client whenever a user starts a Web automation application. Every WIS component is acquired by a URL, so WIS components does not necessarily need to be in a single server; it can be located in different locations, adding more management possibilities.

WIS components can be divided in two distinctly different categories: WIS pages and WIS profiles.

WIS pages are similar to Web pages, with the difference that it contains code proprietary to WIS and cannot be displayed in ordinary Web browsers. Its main function is to provide interface to the user. The first WIS component that a WIS client gets from the server is the WIS application main page. It provides a starting point for the Web automation application, which can do initializations such as downloading WIS profiles. WIS pages can contain HTML, DHTML, JavaScript or Java applets.

WIS profiles are documents written in XML and can contain WIS defined profiles or profiles specifically defined for a specific Web automation application. Profiles specified by the developer may contain any information such as the setup parameters for the automation application. WIS proprietary profiles are understood by WIS and consist of two types: WIS surfing process and WIS extraction definition. The WIS surfing process (Figure 4) defines the process of interoperation with a resource. The WIS extraction definition tells WIS how to get desired data from pages returned by sources and uses part of the WIDL definition.

```
<AutoProcesses StartURL="starting url">
  <ScriptCode>Session context code</ScriptCode>
  <ScriptCodeGlobal>Global context code</ScriptCodeGlobal>
  <AP>Page context code
    <SP>Frame context code</SP>
    <SP>Frame context code</SP>
    …
  </AP>
  <AP>Page scope code</AP>
  …
</AutoProcesses>
```

Fig. 4. Structure of a WIS Surfing Process

A typical Web automation application works as follows. The user first selects a Web automation application by giving the URL of the WIS application main page. Once the application is downloaded, the main page performs initializations, usually downloading the profiles needed. After initialization is finished, the user can interact with the Web automation application interface. The Web automation application will then create WIS sessions as many as needed to accomplish the required task. Every WIS session performs interoperation with a target resource according to the description in the WIS surfing process. A WIS session can also be used to load other WIS pages from the Web automation application server and display to the user.

The programming language used by WIS is JavaScript. But the stateless nature of Web applications and the various WIS components creates different regions of code that have different concerns, which are called programming contexts. There are two programming contexts that originate from the type of the WIS component: WIS pages context and WIS surfing process context. Like JavaScript in Web pages, the developer can expect to have anything that an ordinary browser would provide for scripting, and the lifetime of entities such as functions, variables and objects declared here are only within the page. These two contexts are

essentially stateless, since every time the page is reloaded or substituted, user defined entities disappear. WIS provides two more contexts to keep persistent entities. The WIS session context has the lifetime of a WIS session, which supplements the WIS surfing process. When an entity should persist between pages of a surfing process, it can be placed at the WIS session context. Code in the WIS surfing process context can access the persistent entities defined in the WIS session context. Another context available is the WIS global context, which exists until the Web automation application terminates. The WIS global context is accessible by all other contexts. Table 1 summarizes the different program contexts.

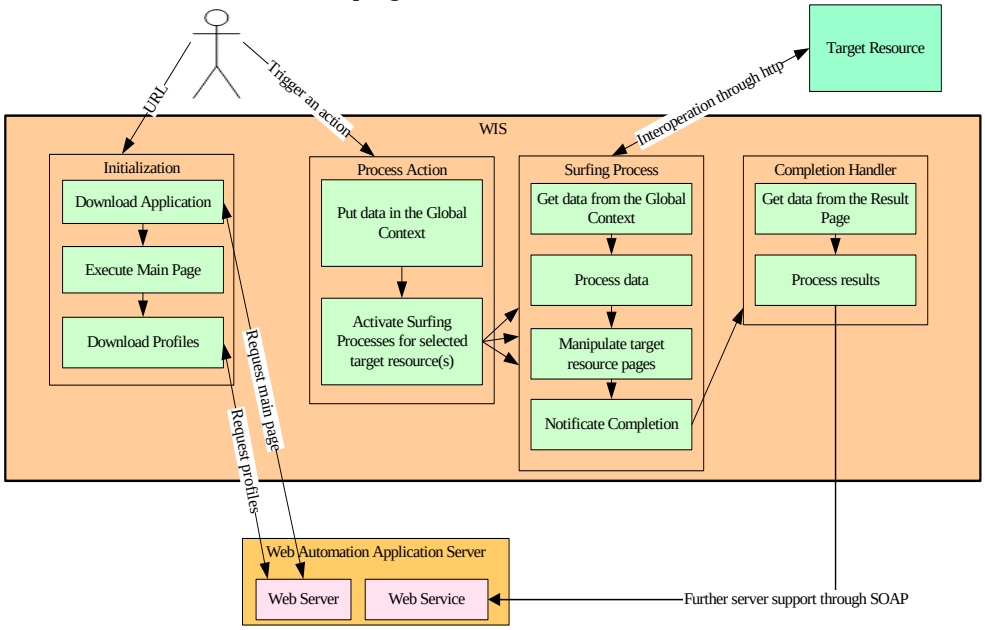


Fig. 5. Typical workflow of an application in WIS

Table 1. Summary of programming contexts

Context type	Page	Surfing process	Session	Global
Purpose	Automation application	Resource interaction	Session persistence of entities	Global persistence of entities
Lifetime of entities	Within page	Within page	Within session	Within application

Table 2. Objects in WIS

Property	Description
GlobalCO	Global context
UniDoc	Provide access to WIS profiles
WMultiWB	Provide management facilities for a collection of WIS sessions
WAutoWB	WIS session
WMessage	Log messages from the Web automation application
WBExt	Miscelaneous tools are provided by this object
Soap	Support for Web services

WIS profiles are documents written in XML. A single WIS application may consist of many WIS profiles. For convenience, these WIS profiles can be divided in small pieces for development and management convenience. WIS profiles can be placed anywhere. Putting them in the client-side can save download time, especially if there are hundreds of profiles. But frequent updates may be difficult if there are many clients. It is more convenient to have it stored at the server-side and be accessed every time a Web automation application needs.

Multiple profiles can be unified to a single document for easy access from the Web automation application. The UniDoc object does this job. It identifies the <UniDocInclude> element that is substituted by the demanded profile defined in the Src attribute. In this manner, profiles are embedded into parent profiles, forming a single profile for the Web automation application. Parts of the profile then can be obtained by using DOM or XPath.

WIS exposes various functionalities through an API that is called by JavaScript. Table 2 gives a list of objects provided by WIS.

3.3 Data Extraction from HTML Pages

WIS adopts WIDL for data elements extraction from HTML pages, but with some enhancements.

When a user wants to find something in a HTML page that has changed, he/she first identifies unchanged parts, such as titles, which were associated with the wanted data in past versions. Whenever the unchanged part is found, it is very likely that the desired part of the page is located nearby. The <REGION> element with the SINGLE attribute tells WIS what probably will not change in the HTML page, and thus can serve as a reference point.

Object references in WIDL uses an object model to provide access to elements and properties of HTML. To access a child property or element of the parent, the dot operator is used. WIS introduces another operator, the parent operator (^), which returns the parent of the element. It is useful with the reference element. For example, in an online bookstore we may find that the element with the text "ISBN" is the most likely to not change, and the rest of demanded data surrounds it. The following defines it as a reference element:

```
<REGION NAME="RefEle" SINGLE="li['ISBN']" />
```

The parent operator returns the element <UL> containing the <LI> element with the ISBN, which is also the parent of other variables. To get other variables such as title and author, we can use the parent operator with the reference element:

```
<VARIABLE NAME="Title" REFERENCE="RefEle^li ['Title'].text" />
<VARIABLE NAME="Author" REFERENCE="RefEle^li ['Author'].text" />
```

If the title is in a higher level of the document object hierarchy, the variable definition can be obtained through a series of parent operators:

```
<VARIABLE NAME="Title" REFERENCE="RefEle^^^text" />
```

WIDL provides several types of indexing, but only one type can be used at the same time. There are occasions that a single indexing method is not sufficient to match wanted elements and multiple conditions must be given to filter out undesired elements. WIS solves this by providing multiple indexing, with every condition separated with a comma. For example, the following defines a variable that has the class attribute with value "small" and contains the text "Price":

```
<VARIABLE NAME="Price" REFERENCE="td[class='small','Price'].text" />
```

4 WIS Example Applications

4.1 Unisearch 2

The Unisearch [7] is metasearcher system for online databases provided in CONCERT [20]. It performs the translation of queries and dispatches queries to various sources, but does not combine the returned results. Using WIS to create Unisearch 2 still accomplishes the requirements of Unisearch plus some improvements. Exploiting WIS's data extraction capability, Unisearch 2 improves Unisearch by organizing the results and then returning the organized results to the user.

In WIS, the Unisearch application makes connections from the client side, thus respecting the access policies and statistic mechanisms of sources. Depending only on HTTP, its creation does not need the cooperation of source providers. Figure 6--Figure 8 show the Unisearch 2.

4.2 Book Recommendation System for Library

A library has to know which books are really needed by readers before acquiring them into the library's holdings. Traditionally, a reader obtains metadata about a desired book from a source, such as online bookstores, and recommends it by passing the information to the librarian through many different ways and formats, such as emails or slip of papers. No matter which way, errors and losses in the metadata provided seems to be inevitable, forcing a librarian to check the data. After the check, the librarian needs to type the metadata into the library automation system (Figure 9).

With the recommendation system (Figure 11) described in this Section, the metadata is passed directly from machine to machine, increasing efficiency and convenience (Figure 10). No more manual transcriptions are needed.

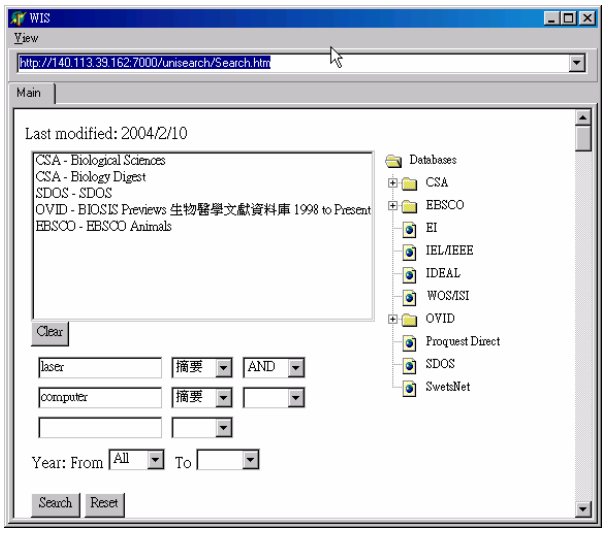


Fig. 6. Unisearch 2 search interface

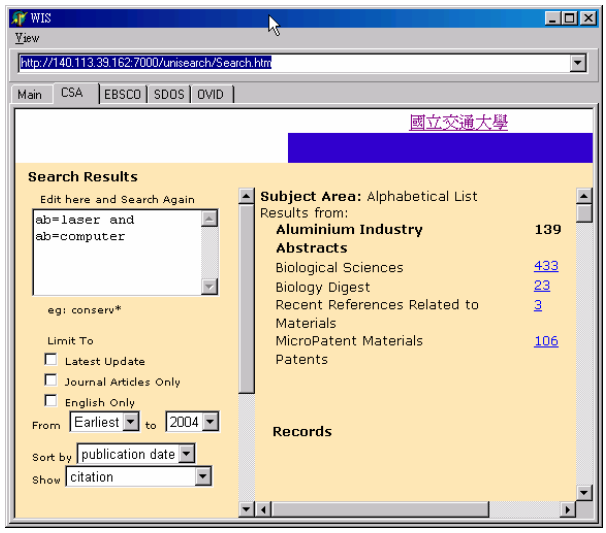


Fig. 7. Results returned by Unisearch 2

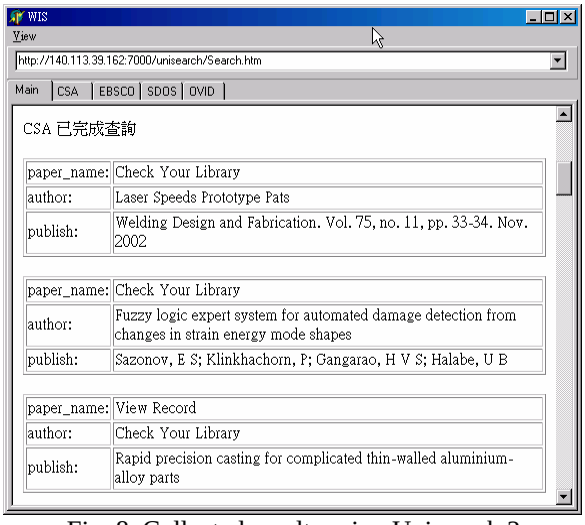


Fig. 8. Collected results using Unisearch 2

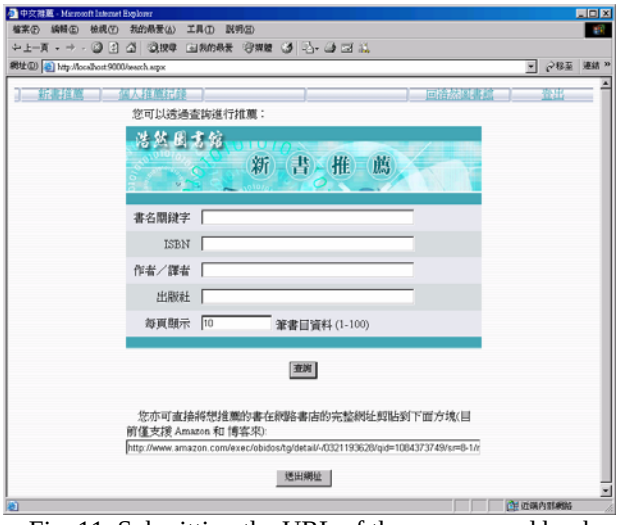


Fig. 11. Submitting the URL of the recommend book

4.3 Web Site Checking System

Many Web sites need to keep online 24 hours, without any interruption; otherwise business opportunities may be lost and users unsatisfied with the poor service quality. For simple sites, a connectivity check to the first page may be sufficient. But nowadays many sites have dynamic pages with complex code behind and a plenty of functions. A check to the first page is then no more sufficient; the manager of the site may have to go through many steps until he/she can be sure that everything operates normally. Manipulations such as login test and search test may be required.

The Web site checking example checks two sites at the moment. One is the Taiwan mirror site of IDS (http://www.csa.com.tw), which is a mirror site of multiple online abstract databases produced by CSA and serves Taiwan users. Another is MyLibrary@NCTU (http://mylibrary.e-lib.nctu.edu.tw), which provides information about various resources in National Chiao Tung University plus personalization services. The test for the mirror site of IDS is a search procedure, and for MyLibrary is database browsing verification and a login procedure.

The manager first accesses the Web automation application using WIS and defines the interval between the tests. Whenever a step fails during the test process, a Web service is invoked to report the error to a server that will save the message to the database and send an E-mail to the responsible system manager. A centralized log server is beneficial, especially if we want to make the test from different subnets to ensure that there is no problem when difference in location could affect functions of the site such as user access control.

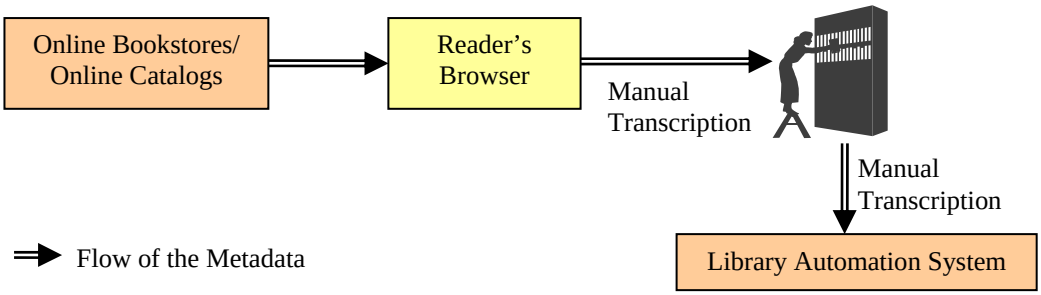


Fig. 9. Conventional book recommendation process

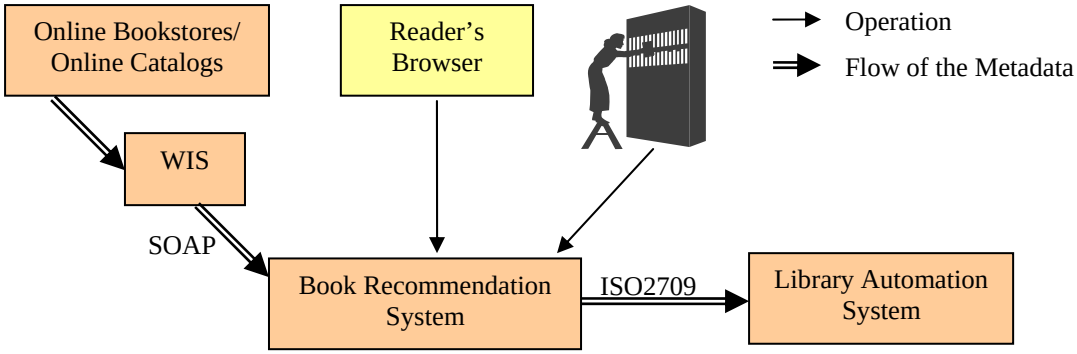


Fig. 10. Automated book recommendation process

5 Conclusion and Future Work

Based on the common needs of Web automation applications, this paper proposes an infrastructure to create Web automation applications, without being limited to any specific kinds of applications. A general Web automation creation solution should consider the follow issues, and Table 3 shows how these design considerations are supported by WIS.

- Interoperability: the core concept of Web automation applications is to reuse existing Web resources. From the interoperation aspect, Web automation applications can adopt standard interoperation protocols (such as OAI and Z39.50), or rely only on the common HTTP protocol. Standard interoperation protocols are not supported by many Web resources; on the other hand, using the common HTTP protocol gives access to all Web resources, but problems such as complexity of Web sites and volatility of the interfaces need to be solved.
- Parallel processing: Web automation applications usually need to interact with several Web resources at the same time.
- Server-side or client-side Web automation execution: execution in server side has the advantage of easier deployment, maintenance and management, but puts limit to performance scalability. Execution in client side gives the contrary.
- Flexible presentation: a flexible interface is necessary so that the tool can fulfill different design needs.
- Integration with corporate systems: Web automation applications may need to work with enterprise servers.
- Scheduling: Web automation tasks are usually repetitive and thus may need scheduled execution.
- Integrated development environment: modern application creation tools provide integrated debugging, wizards and others.
- Intelligent tools: Web automation applications are to replace human labor, so intelligence is a desirable characteristic of an agent.

WIS is an initiative to make the creation of Web automation applications easier, but there are still much more functions to be incorporated into WIS.

An integrated development environment (IDE) is a good target for the next step, which gives the possibility of massive creation of Web automation applications. The software industry has benefited much from IDEs, and so can be the field of Web automation applications, pushing us to the era of "service reuse". There are several issues that can be considered in an IDE for Web automation application creation:

- Integrated debugger for easier troubleshoot of applications.
- Authoring tool: editors and GUI tools with drag-and-drop capability.
- Wizards: there are some related works [3][5] that emphasizes the creation of Web automation tasks by learning the surfing steps from the user's interaction with the source. But it is doubtful that the computer can realize everything that the user has done because the complexities of Web applications, which may contain scripts with dynamic content that are not so easy to be detected. Thus implementing this facility as a wizard is a viable solution, which serves as a starting point for the creation phase. The developer can then make changes to the generated code to fix the incorrect parts.

Web automation applications are related with agents to work in place of human interaction, which in many cases need some sort of intelligence. For example, intelligent metasearchers should analyze returned results and organize them by reranking and summarizing. These intelligent features are usually designed for specific situations, a characteristic that keeps them out of the API set of WIS. Intelligent tools of common purpose for Web automation applications are a matter of future research.

Table 3. Design considerations and WIS support

Feature	Support by WIS
Interoperability	Needs HTML and HTTP only; data extraction facility
Parallel processing	Multiple session, multiple context
Server-side or client-side	Mixes the advantages of both server and client approaches
Flexible presentation	HTML interfaces; round-trip free by using DHTML
Integration with corporate systems	Uses Web Services
Scheduling	Enable
Integrated development environment	None
Intelligent tools	None

References

[1] E. Selberg, O. Etzioni, "The Metacrawler Architecture for Resource Aggregation on the Web", IEEE Expert, pp.11-14, Jan.-Feb. 1997.

[2] M.W. Spalti, "Finding and Managing Web Content with Copernic 2000", Library Computing, Westport, pp. 217-221, Volume 18, no. 3, September 2000.

[3] A. Sugiura, K. Yoshiyuki, "Internet Scrapbook: Automating Web Browsing Tasks by Programming-by-Demonstration", Computer Networks and ISDN Systems, Volume: 30, Issue: 1-7, pp. 688-690, April 1998.

[4] C. H. Tseng, S. S. Huang, H. R. ke, and W. P. Yang, "Information Extraction for Documents with Common Structure in Virtual Union Catalog Systems," Journal of Internet Technology, vol. 2, no. 1, pp. 59-68, 2001.

[5] B. Krulwich, "Automating the Internet - Agent as User Surrogates," IEEE Internet Computing, Volume: 1, Issue: 4, pp 34-38, July-Aug. 1997.

[6] M.G.. Wales, "WIDL: Interface Definition for the Web", IEEE Internet Computing, Volume 3, Issue 1, Jan.-Feb. 1999.

[7] W.H. Yen, M.J. Hwang, H.R. Ke, "Integrated Search of Digital Library", Proceedings of 2000 Taiwan Area Network Conference, pp.484-491, October 2000.

[8] WIDL, http://www.w3.org/TR/NOTE-widl-970922.

[9] SOAP, http://www.w3.org/TR/SOAP/.

[10] Web Services Description Language, http://www.w3.org/TR/wsdl.

[11] Microsoft Scripting Technologies, http://msdn.microsoft.com/scripting/.

[12] Meta-Search Engines, http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/MetaSearch.html.

[13] BestBookDeal, http://www.bestbookdeal.com.

[14] BestWebBuys, http://www.bestWebbuys.com.

[15] Metalib, http://www.exlibrisgroup.com/metalib.htm.

[16] A Gentle Introduction to XML, http://www.tei-c.org/P4X/SG.html.

[17] Scripting Technologies, http://msdn.microsoft.com/scripting/default.asp.

[18] "Document Object Model (Core) Level 1", World Wide Web Consortium, http://www.w3.org/TR/2000/WD-DOM-Level-1-20000929/.

[19] "Document Object Model (Core) Level 2", World Wide Web Consortium, http://www.w3.org/TR/2000/REC-DOM-Level-2-Core-20001113/.

[20] Consortium on Core Electronic Resources in Taiwan, http://www.stic.gov.tw/fdb/index.html.

現代漢語複合動詞之詞首詞尾研究

邱智銘，駱季青，陳克健

中央研究院資訊科學研究所詞庫小組

henning@hp.iis.sinica.edu.tw, airport@iis.sinica.edu.tw, kchen@iis.sinica.edu.tw

摘要

一般探討現代漢語複合動詞(compound verbs)，不外乎提到動補結構(verb-result)、動賓結構(verb-object)、並列結構(verb-verb)以及偏正結構(adverb-verb)四種類型，但卻鮮少提及複合成分(modifier-modifee)與其複合後所形成的動詞語意及與句法之互動。本文主要是探討複合動詞(以[詞庫小組1993]標記後的複合動詞)的組成，此可分爲兩類來分析[Fab,2001]：一類爲區分前後述詞的關連性及本身屬性(predicator-predicator relation)，區分複合動詞的核心述詞(Head)以及輔助述詞(verbal satellite)；另一類爲區分述詞及論元的關連性及本身屬性(predicator-argument relation)。藉由中央研究院現代漢語平衡語料庫(Sinica Corpus)找出衍生性強的詞首(morpheme-initial)及詞尾(morpheme-final)，在這兩類中的所扮演的語意角色(semantic role)及複合後的語法功能。

1. 緒論

關於現代漢語的複合動詞，本文根據構詞情形(morphological rules)將複合動詞分爲兩大類：第一類爲N-V或V-N結構的複合動詞，即傳統的主謂結構(subject-verb)以及動賓結構(verb-object)。第二類爲V-V結構的複合動詞，此類動詞包含了傳統的四種分類：動補結構(verb-result)、偏正結構(adverb-verb)、並列結構(parallel verb-verb)以及連動結構(serial verb-verb)。本文依三個外部結構將其分爲以下的內部結構：.

外部結構	N-V	V-N	V-V
內部結構	instrument+verb	verb+subject	manner+verb
		verb+object	aspectual+verb
		verb+locus	subevent+verb
			verb+parallel
			verb+serial verb
			verb+result

內部結構爲本篇文章所提出的分類，以核心述詞(head)爲主，輔助述詞(verbal satellite)爲副，旨在探討複合動詞內部語意的關係。

本篇中的複合動詞是取自於五百萬詞的平衡語料庫中，複合動詞的詞類標記(tagging)爲動詞、其詞首或詞尾出現頻率3次以上，且不出現在辭典中(詞庫小組八萬目辭典)。將所選取出來的複合動詞，建立一詞彙庫，以詞首、詞尾爲索引，把出現於詞首或詞尾的單字詞，再用人工作做語意及結構的區分及分類。經過人工檢視後，得到常用複合動詞之詞首735個(918個詞義)、詞尾282個(300個詞義)。所列的複合動詞先做語意的區分，語意區分後再依結構類型V-V、V-R、A-V及V-O分類[1]。

[1] 目前複合動詞的詞彙庫屬於剛建立的階段，只將內部結構作四類跟語意的區別，此四類型(type)個別包含的複合動詞例子(token)爲：V-V (726 個)、V-O(262 個)、A-V(877 個)、V-R(3314 個)。

此詞彙庫建立(http://turing.iis.sinica.edu.tw/affix/)，是來研究現代漢語中詞首/詞尾[2]與其他詞(輔助述詞或體詞)的共現(collocation)情形，更進一步的去探討所選出之複合動詞(共現)的結構類型、論元關係(argument structure)及其語法功能(grammatical function)，幫助推演出複合動詞詞類的給予，以當作現代漢語機器判別未知詞、斷詞、合分詞以及剖析句結構樹中心語的參考依據。目前複合動詞語彙庫的建立只做到外部結構區分，內部語意結構區分與詞類的關係爲未來持續進行的目標。

2. 複合動詞表層結構與內部語意

雙字複合動詞在現代漢語爲一常見的現象，其複合動詞大致上可分爲兩類，一類爲V-N或N-V結構的複合動詞，另一類爲V-V結構的複合動詞；根據兩複合成分(morpheme)的組合情形，可歸納出結構與語法、語意的依存關係，以下就以三種複合動詞結構爲出發點，來探討內部語意、表層結構與論元結構的關連。

2.1 N-V結構複合動詞

常見的N-V結構複合動詞就是語法學家所說的主謂結構(subject-predicate)性的複合動詞，其N-V結構的第一個名詞成分通常爲其複合動詞之論元結構主詞的部件詞(part meronym)，例如：眼小、鼻歪、腿短，等等，不過這類所謂的主謂結構並不在我們所收的N-V結構複合動詞中，因其語法行爲並不能說明其結構爲一分詞單位：

1. 他很心*痛*(VH)。
2. ＊ 他很鼻歪→ 他鼻很歪。

而我們所分析的N-V結構複合動詞，根據其第一個名詞成分與第二成份核心語的互動關係，可歸納出其語意角色爲工具格(instrument)，例如：槍殺、水洗、火烤，等等，其第一個名詞成分都是核心語動詞在執行動作時所用到的媒介，基本上，這類的N-V結構複合動詞可還原使用介詞型式的句型「用N來V」，例如：

3. 我們來*火烤*這塊肉。→我們用*火*來*烤*這塊肉。

現代漢語複合動詞新詞衍生中，這類N-V結構的衍生性很強，其實在非謂形容詞中，此一類型的複合結構已常見，例如：紙糊的、肉做的，等等，下一節即會討論詞類與語法行爲的改變。

2.2 V-N結構複合動詞

現代漢語的V-N結構複合動詞可分兩小類探討，第一小類爲verb-subject的依存關係，類似主謂結構的倒置；第二小類爲verb-locus的依存關係；最後一小類爲verb-object的依存關係，即所謂的動賓結構。

首先我們先從verb-subject類著手討論，這類複合動詞的核心語通常爲動作性不及物動詞，例如：走人、跑馬，等等，所複合的動詞語意稍有轉變(meaning shift)；這類複合動詞不像主謂結構的複合，其名詞成分與複合動詞的論元角色並無部件-整體(meronym-holonym)關係，且複合後的動詞與核心語的語法行爲類似，都是爲動作性不及物動詞，例如：

4. 我們走人。

[2] 本文探討的詞首詞尾與屈折詞綴(inflectional affix)及衍生詞綴(derivational affix)不同；語法詞綴有固定獨立的語法功能，且不影響緊鄰成分的語法類別，故一律切分開，例如：「了」；衍生詞綴衍生性高，可以用規律預測產生，多半會改變所附加成分的語法類別，例如：「-化」。理論上，詞頭詞尾與相關附著成分合爲一複合詞，可視爲一分詞單位，然而詞綴是個小範圍的詞集。

第二小類的verb-locus 語意結構，所表達的是核心詞與處所間的互動關係，例如：跳車、跳樓。若核心詞與固態有型體的名詞複合，其名詞所表示的是爲一起始點(source)，若核心詞與物質的名詞複合，所表示的是實體爲一目標地 (target)。例如：

5. 有人*跳樓*了
6. 我們去哪裡*泡溫泉*？

例句(5)「樓」所表示的爲核心語「跳」的一起始點，然而，在例句(6)當中，「溫泉」，爲一液態無形體物質，所以爲核心語「泡」的一目標地。

最後一類爲verb-object的複合動詞，也就是傳統分類中的動賓結構複合動詞。此類核心語類似動作單賓述詞，後接的賓語可爲一終點(goal)或客體(theme)，例如：打水、吃醋。

總括而言，V-N結構複合動詞雖有語意分類上的差別，但基本上來說，其複合後的語法功能，大都呈現動作性不及物動詞，但若核心語本身有類雙賓功能，可能會影響V-N結構複合動詞的論元結構，例如verb-locus 語意結構的複合動詞：泡醋、浸酒，其複合後的表現出動作類單賓述詞：

7. 我們將檸檬*泡醋*。

「泡醋」爲一動作類單賓述詞，「檸檬」爲一終點(goal)，表現出與一般的verb-locus類與verb-object類不同語法層次。

2.3 V-V結構複合動詞

現代漢語V-V結構複合動詞是一很複雜現象，傳統上是將其分爲三小類，即爲偏正結構(adverb-verb)、並列結構(verb-verb)以及動補結構(verb-result)，但往往由於難以釐清核心語與輔助動詞，所以不好歸類其複合動詞屬於何種結構，接下來就將各結構內部語意特性提出探討，以釐清兩單字動詞結合後的語意及語法的關係。

2.3.1 偏正結構複合動詞

偏正結構類最基本的形式爲狀態述詞(stative)加上一核心語，此狀態述詞可視爲核心語的一輔助語意角色-狀態副詞(manner)，例如：錯寫、暗殺、亂說；基本上，這類的複合動詞都爲本身核心語的下位詞(Hyponym)，其狀態述詞可視爲提供核心語一較精確的語意行爲(background)，而主要的語法行爲還是取決於核心語本身(foreground)。例如：

8. 他(暗)殺了總統。

其複合結果並不影響其論元結構的改變(他：agent；總統：goal)。

另外兩類的輔助動詞則指涉核心語內部事件結構，一類爲動貌詞(aspectual verb)，另一類爲次動詞(subevent verb)。動貌詞顧名思義即是描述一事件的時間內貌，可強調一事件的開始(initial)、繼續(resumption)以及結束(final)，例如：開罵、續看，

9. 老師準備要*開罵*了！
10. 妳不*續借*這本書嗎？

例句(9)中，「開-罵」表示「開始-罵人」，但複合後「開罵」這個述詞的論元角色並沒有承襲核心語「罵」這個動詞，由於焦點是在「罵」這個事件的開始所以複合後論元結構只剩老師這一角色(agent)。例句(10)的「續借」雖然沒有改變論元數目，但客體(theme)這論元卻有其語意限制(semantic restriction)，客體必須爲舊訊息：

11. *妳不*續借*一本新書嗎?

總之，動貌詞類的複合動詞，不能完全從核心語來判斷其語意及論元結構。

另一類輔助動詞涉核心語內部事件結構爲次動詞(subevent verb)複合動詞，此類複合動詞的核心語即爲輔助動詞的事件整體詞，例如：鋪築、燒製，等等，

12. 他們用很大的鵝卵石*鋪築*路面。

例句(12)中，「鵝卵石」是屬於輔助動詞「鋪」的論元結構，本句可詮釋爲：「他們鋪大的鵝卵石來築成路面」，所以「鋪」可視爲達成「築」的次事件；當「鋪」完了「鵝卵石」，即是「築」成道路之時。不過，有時次動詞取代核心語的語意(semantic extension)，例如： 鋪(築)道路、燒(製)陶。

其實狀態述詞(stative)的偏正結構與次動詞(subevent)的偏正結構都是要提供核心語一較精確的語意行爲(background)，而主要的語法行爲還是取決於核心語本身(foreground)，所以一個可說是 stative adverb-verb 結構，一個則是dynamic adverb-verb結構。

2.3.2 並列結構複合動詞

先前提到的次動詞偏正結構，常常會被誤判爲並列結構(verb-verb construction)，事實上，現代漢語的並列結構是指單獨兩事件，因時序上(temporal)的關係，分爲事件並列結構(temporal parallel construction)以及事件連動結構(temporal serial construction)。

事件並列結構常常是兩事件爲近義詞(synonym)或反義詞(antonym)，不過反義詞兩事件的複合，常會形成一上位名詞，例如：愛恨、生死、進退，等等，然而，近義詞兩事件的複合，會保留兩事件的論元結構，形成事件並列複合動詞，例如：推析、拿持。

13. 他並沒有仔細*推析*這問題。
14. 這時忽然看見一個小孩兒，*拿持*琉璃瓶。

例句(13)、(14)中的「推析」、「拿持」都是包含了兩個近義詞的事件，而且時序上，並無先後順序，複合動詞中兩個近義詞的事件，都是具有相同的論元結構，故此爲事件並列結構(temporal parallel construction)。

事件連動結構(temporal serial construction)，常跟次動詞偏正結構(dynamic adverb-verb)有分類上的困難，不過可以確定的是，偏正結構的輔助動詞在時序上是與核心語平行，然事件連動結構是具有兩個核心語，且有先後時序關係，例如：購閱，裝送，

15. 乘客們紛紛*購閱*本報特刊。
16. 誰要*裝送*貨物？

很明顯的是，「購閱」一定是先「購」書、才「閱」書，而「購書」不包含在「閱書」的事件中，所以事件連動結構爲一時序相關結構(temporal-related construction)。

2.3.3 動補結構複合動詞

動補結構爲現代漢語使用頻率較高的一種複合動詞，一般認爲動補結構有使成的意味(causative reading)，原因爲第一事件動作會造成另一事件或狀態的發生，然在本章節，依補語成分區分爲兩種類型，一爲一般狀態補語的動補結構(verb-result)，另一爲虛化(grammaticalized)成分較高的補語，在傳統上被歸類於動相詞(phase markers)，然而這類的補語則帶有起使的意味(inchoative reading)。以下就這兩種動補結構加以闡述。

2.3.3.1 一般狀態補語的動補結構

一般狀態補語的動補結構可分爲兩種情形，一是論元結構中的主語狀態改變(verb + subject change-of-state)，另一則是賓語狀態改變(verb + object change-of-state)。主語狀態改變的例如：

17. 阿甘跑累了！

此複合動詞的主事者爲「阿甘」，狀態改變的也爲「阿甘」，可是說是主事者致使自己變胖(causer=causee)。同樣的，是賓語狀態改變也是有使成(causative)的意味，但主事者對賓語狀態的改變處與於主導地位(causer≠causee)，例如：

18. 唐先生*打破*了他太太的花瓶。

所以例句(18)的主事者「唐先生」致使賓語「花瓶」成爲「破」的狀態。然而如何得知動補的補語所描述的是主語還是賓語？這必須探討補語本身與核心語的互動關係。在例句(17)中，「跑」跟「累」間有著時間持續 (duration)的關係，換言之，並不是核心語跟補語同時發生(co-exist)，而是「跑」了一段時間，結果才變「累」，這樣的動補結構，焦點在最後的狀態(result)上，而通常焦點在狀態的動補複合詞，就論元結構，當然在句構表面也只會出現被描述狀態的那個論元，即爲主語，例如：

19. 小弟怎麼*吃胖*了？

「吃胖」這動補結構的焦點在「吃」的狀態上，所以主語即爲被描述的「小弟」，然而，「吃」什麼東西並不重要，所以賓語(goal)不需要出現。總體而言，一般狀態補語的動補結構的區分在於核心語與補語之間的時間持續(duration)的關係，若兩者間有著時間持續的關係，則語意焦點會移往補語上，當然語意焦點的轉移也會反映在論元結構以、詞類與句法上，下一節會就此點加以說明。

2.3.3.2 時相詞補語的動補結構

時相詞補語在動補結構中雖然扮演的是補語的角色，但不像狀態補語一樣，有著明顯的結果狀態，然而所補述的卻是核心語本身的事件面貌，這類的補語除了包括傳統的時相詞：好、完、到、掉、光，等外，語意虛化的方向詞：上、下、過、起、開、回、進、出，也是屬於這類，這類的補語保有了一般狀態動補結構的狀態轉變的(change-of-state)語意，但無使成(causative)的意味，例如，

20. 飯菜冷掉了！

時相補語並不像動貌詞(aspect marker)，動貌詞是屬於完全虛化的詞彙，像句末「了」，可在任何語境下使用，表「狀態改變」，有著起使(inchoative)的功能，例如，

21. 飯菜冷了！

例句(21)中的「了」，只表文法詞「狀態改變」的功能，並沒慘雜本身詞彙語意在其中，然而像例句(20)中的接尾詞「掉」，雖然也表「狀態改變」的功能，但其與核心詞的結合性，還須取決於其語意的限制(semantic restriction)，例如，

22. 車子*壞掉*了！
23. *車子*好掉*了！

根據語料顯示，與「掉」結合的核心語，都是負面(negative)的狀態詞，例如：爛掉、臭掉、呆掉，等等，所以例句(23)中的「*好掉」則無法成爲一複合詞。由此可知，時相詞在與核心語結合成複合動詞時，是必須通過語意規則的測試，並不能像文法詞能廣泛使用，因爲除了語法功能外，時相詞還保有本身某些語意特性。

2.4 小結

本節就現代漢語V-V結構複合動詞分爲三小類探討，即爲偏正結構(adverb-verb)、並列結構(verb-verb)以及動補結構(verb-result)，從討論中可整理出一些V-V結構動詞複合的現象，就語意及功能相似性方面，偏正結構與動補結構有著對稱性的相似：動貌詞的偏正結構與時相詞的動補結構、狀態述詞的偏正結構與狀態補語的動補結構；就結構相似性方面，事件連動結構(temporal serial construction)，常跟次動詞偏正結構(dynamic adverb-verb)有分類上的困難。

偏正結構的動貌詞(aspectual verb)和動補結構的時相詞(phase marker)都是屬於虛化程度較高的詞素，都是有著狀態轉變的(change-of-state) 功能，例如：

24. 一般來講，我在電影開拍半年前就會準備跟導演*開聊*。
25. 反倒是這個來自路易斯安那州的美國仔居然跟我*聊開*。

雖然例句(24)、(25)中，詞首的「開」與詞尾的「開」都表核心語的起使功能(inchoative)，但還是有些微的差別(nuance)，例如：

26 *我準備跟那導演*聊開*。
27 ？慢慢的*開聊*後，我才發現這個人並不壞。

從例句(26)、(27)中可發現，「開」當作詞首時，是有著明顯(overt)起始功能的狀態轉變，然而「開」當作詞尾時，有著與「起來」一樣的功能，並無明顯(covert)起始功能的狀態轉變，例如：

28. 我一進門，就已經看到他們*聊開(/起來)*了！

在狀態述詞的偏正結構與狀態補語的動補結構方面，雖然輔助詞都是狀態動詞，同時也並沒有虛化現象，但在複合後整體的語意表現上，卻有不同的詮釋，例如，

29. 我一直*錯寫*這個字。
30. 我一直*寫錯*這個字。

例句(29)所以表達的是「我一開始就不知道這個字的正確寫法，到現在才知道，字是錯的」；而例句(30)則須靠上下文判斷，有可能是「一開始就不知道這個字的正確寫法」，也有可能是「知道正確寫法，但一直寫出不正確的字」。

如上所述，可看出，動貌詞的偏正結構與時相詞的動補結構在語意上較接近，但語法功能上不同，然而狀態述詞的偏正結構與狀態補語的動補結構在語法功能上較接近，但在語意詮釋上較不同。

事件連動結構(temporal serial construction)與次動詞偏正結構(dynamic adverb-verb)，結構表面都是V-V結構，不僅機器判定會有困難，人工在檢視時也常有決定上的困擾，因此，若能建立共現性較高的詞彙庫，即能輔助結構、語法以及語意上的判定。

3. 複合動詞內部結構與詞類的關係

詞庫小組的述詞分類可分爲12大類[CKIP1993][3]，然而在衍生性強的複合動詞詞類該如何給予？是否能找出複合動詞的內部結構(核心動詞)與詞類之間的關聯？基本上，複合動詞爲核心結構(endocentric construction)，即能從核心動詞判斷其詞類爲何；若不爲核心結構(exocentric construction)的複合動詞，又該如何判斷其詞類？本章節就上一章節所討論的複合動詞表層結構與內部語意的關係來探討複合動詞結構與詞類的關係。

[3] 此十二類可分爲「動作述詞」VA, VB, VC, VD, VE, VF,VG七類；「狀態述詞」爲VH, VI, VJ, VK, VL五類。

3.1 核心結構的複合動詞

所謂核心結構的複合動詞爲複合動詞的詞類可由核心詞相同，此即複合動詞的論元結構與核心動詞的論元結構一樣。以下就三類結構的複合動詞，提出其核心詞詞類與複合後的詞類相同：

(1) N-V結構(instrument+verb) →(火)烤=VC （動作單賓述詞）
(2) V-N結構(verb+subject) →跑(馬) =VA（動作不及物述詞
(3) V-N結構(verb+locus)→ 跳樓=VA （動作不及物述詞）
(4) 偏正結構(manner + verb) → (暗)殺=VC （動作單賓述詞）
(5) 偏正結構(aspectual + verb) →(續用)(VC) （動作單賓述詞）
(6) 偏正結構(subevent + verb) → (鋪)築(VC) （動作單賓述詞）
(7) 並列結構(verb+verb)→ 拿持=VC （動作單賓述詞）
(8) 連動結構($verb_{t1}$+$verb_{t2}$)→ (裝)送=VD （動作雙賓述詞）
(9) 動補結構(verb + object change-of-state)→弄(髒)=VC（動作單賓述詞）
(10) 動補結構(verb + phase complement)→吃 (掉)=VC（動作單賓述詞）

以上九類的複合結構動詞，只要清楚其內部結構及其語意，皆可從核心詞的詞類來決定複合詞的詞類，此爲核心結構(endocentric construction)複合動詞，常用的詞首詞尾核心詞及其語意可從詞庫小組所整理的「現代複合動詞詞頭詞尾表」得知[4]。

3.2 非核心結構的複合動詞

非核心結構的複合動詞的部分可分「可歸納出規則的結構」、「個別詞綴」與詞類的關係，以下分別論述之。

3.2.1 可歸納出規則的結構與詞類的關係

不爲核心結構(exocentric construction)的複合動詞且可歸納出規則的結構與詞類的關係可分五類，如下：

(1) V-N結構(verb+object) → 打水[5]= VA （動作單賓述詞”打”→動作不及物述詞）
(2) V-N結構(verb+material)→ 泡醋=VB （動作單賓述詞”泡”→動作類單賓述詞）
(3) 動補結構(verb + subject change-of-state)→吃胖=VH（動作單賓述詞”吃”→狀態不及物述詞）
(4) 動補結構(verb+directional complement)→說出去=VB (動作句賓述詞”說”→動作類單賓述詞)

這五類的詞類是從結構與複合後所表達的事件結構來決定，不像核心結構(endocentric construction)複合動詞，可從核心詞來給予詞類，大致上，這五類結構所以呈現的都是VA（動作不及物）、VB(動作類單賓)及VH（狀態不及物）這三類述詞。

3.2.2 個別詞綴與詞類的關係

「個別詞綴」與詞類的不規則關係散佈在核心結構與非核心結構中，這必須要靠人工檢閱來分析及論元結構與詞類的關係，例如：防滑、欠揍、耐摔，等等，所呈現的都是V-N結構的述語，但其所現出的事件結構都是狀態不及物類述詞，所以「防-」「欠-」和「耐-」這三個詞首就可被列在特殊規則中，此類的詞綴與詞類的關係還需要人工來詳加探討。

---

[4] 詳見http://turing.iis.sinica.edu.tw/affix/

[5]V-N結構是否爲一非核心結構(exocentric construction)需取決於複合後是否爲一分詞單位；若不爲一分詞單位，例如：吃(VC)麵包(Na)，則不屬於本文討論的複合動詞。

3.3 小結

根據本詞彙庫統計，各類型複合動詞在詞類的分佈，V-V結構的複合動詞中的詞類VC佔58.76%，V-O結構的詞類VA佔62. 07%，A-V結構的詞類VC佔50%，以及V-R結構的詞類VC佔60.64%，如下表：

結構 / 詞類分佈	VV	VO	AV	VR
A		0.38%		
D		0.38%		
VI	0.14%	0.38%	0.68%	0.04%
VL	0.28%			0.25%
VB	0.41%	2.68%	0.80%	7.16%
VK	0.69%		1.71%	0.47%
VG	0.97%		2.51%	5.53%
VF	1.66%		1.71%	0.61%
VD	2.62%	0.38%	3.65%	2.35%
VJ	3.59%	1.92%	4.91%	4.77%
VE	5.24%		4.91%	1.63%
VA	8.69%	62.07%	13.70%	9.76%
VH	16.97%	20.31%	15.41%	6.79%
VC	58.76%	11.49%	50.00%	60.64%

由此表可得知各詞類在四類結構的分佈情形，進而從詞類中得知在現代漢語複合動詞的語法功能，也提供了複合動詞詞彙衍生(productivity)所對應的詞類。

4. 結論

現代漢語的複合動詞可分三個結構語意關係來探討：N-V結構複合動詞、V-N結構複合動詞，與V-V結構複合動詞。N-V結構複合動詞即爲(instrument+verb)的語意結構； V-N結構複合動詞可分三小類：(verb+subject)、(verb+object)、(verb+locus)；V-V結構複合動詞可分四小類：(manner+verb)、(verb+ parallel verb)、(verb+ serial verb)，和(verb+result)。複合動詞內部結構與詞類的關係又可分三方面探討：核心結構(endocentric construction)、非核心結構(exocentric construction)以及個別詞綴與詞類的關係；如何從電腦自動化區分核心及非核心結構，無法單純從表面結構區分，因此複合動詞內部語意分析及結構與詞類的關係，爲此提供了一套輔助規則。

致謝

本論文由國科會數位典藏國家型科技計畫-語言典藏計畫-子計畫20世紀現代漢語語料庫與句法結構資料庫，計劃編號：NSC 93-0201-29-戊-3-6.2.2 支援部分研究經費。感謝中研院詞庫小組李祥賓先生設計詞首詞尾字檢索及下載介面http://turing.iis.sinica.edu.tw/affix/。

參考文獻

[1] Chao, Y.-R.1968. *A Grammar of Spoken Chinese*. Berkeley and Los Angeles: University of California.

[2] CKIP. 1993. *An Introduction to Sinica Corpus*. CKIP Technology Report 93-05. IIS, Academia Sinica.

[3] Comrie, B.1976. *Aspect*. Cambridge: Cambridge University Press.

[4] Fab, Nigel. 2001. *Compounding*. In Spencer & Arnold M. Zwicky (eds), the Handbook of Morphology. Oxford: Blackwell Publishers.

[5] Feng, Sheng-li 馮勝利(2002) 漢語動補結構來源的句法分析語言學論叢2002:178-208, Beijing University.

[6] Koontz-Garboden, Andrew and Levin, B. 2004. *The morphological typology of change of state event encoding*. To appear in the Proceedings of the fourth Mediterranean Morphology Meeting. Catania, Italy.

[7] Li, Charles & Sandra Thompson. 1981. *Mandarin Grammar: A functional Reference Grammar*. Berkley: The University of California Press.

[8] Lin, Fu-wen. 林甫雯(1990): 漢語的述補式複合動詞 清華大學語言所碩士論文

[9] McIntyre, Andrew. 2002. *Argument structure and event structure : Resultatives and Related Constructions*. In Workshop 'Complex Predicates, Particles and Subevents', Universität Konstanz

[10] Tang, Ting-chi. 湯廷池 (1992) : 漢語詞法句法四集 台北: 學生書局

[11] Wolters, Maria. 1997. *Compositional Semantics of German Prefix Verbs*. In Proceedings of the 35th Meeting of the Association for Computational Linguistics and the Seventh Meeting of the European Chapter of the ACL, Madrid, Spain

語法規律的抽取及普遍化與精確化的研究
Grammar Extraction, Generalization and Specialization

謝佑明　　楊敦淇　　陳克健

中央研究院資訊科學研究所
{morris, ydc, kchen}@iis.sinica.edu.tw

摘要. 相較於傳統PCFG的CNF處理，在本篇論文中，我們提出二元化句法規則產生模式。並且深入探討其語法普遍化與精確化方法對中文剖析器的影響。實驗設計從中研院中文句結構樹中依不同的語法抽取原則，抽取出不同的語法規律集合，來剖析三份測試語料並評估結果。觀察結果試著去尋找出有效的語法普遍化及精確化方法，得到覆蓋率高且精確的句法規則，以加強中文剖析器的剖析效能。剖析精確率的實驗結果，從基本普遍化語法的81.45%增加到精確化語法的86.14%。**(關鍵詞:覆蓋率、語法歧義、句法剖析、語法抽取)**

1　緒論

自然語言處理的過程中，句法剖析(parsing)是一個核心處理過程。在過去研究中，剖析器(parser)利用從樹庫(treebank)中訓練出的probabilistic context-free grammar(以下簡稱PCFG)，對句子剖析是很常用的技術。在英文的部份，因為有大量的英文樹庫資料，利用PCFG剖析英文句子都會有不錯的效果，現有資料顯示約可至九成，還進一步的做到詞彙化剖析(lexicalized parsing)[6]。相對於有限的中文句結構樹庫，非詞彙化剖析(unlexicalized parsing)是一個研究的開始。在本篇論文中，研究如何從有限的中研院中文句結構樹庫(Sinica Treebank)中，抽取最佳的PCFG，使得抽取出的語法規律有較佳的覆蓋率(coverage)及較低的語法歧義(ambiguity)。我們同時建立一個符合需求的中文剖析器，從中文句結構樹庫抽取出不同的語法規律集合，來剖析三份測試語料並評估結果。從這些實驗中，觀察結果試著去尋找出有效的語法普遍化(generalization)及精確化(specialization)方法，得到覆蓋率高且精確的句法規則，以加強中文剖析器的剖析效能。

在最後章節中，我們也探討到未來進一步的研究與實驗方向，如何整合句法與語意訊息讓剖析器有效解決句法結構歧義的問題。

1.1　中文句結構樹庫簡介

中央研究院中文句結構樹庫由中研院詞庫小組於2000年開始建制。目前的版本是 2.0 (9個檔案)，其中包含有 38,944 句結構樹與240,979詞。每一個句結構樹都有標註 "詞、細詞類、語法結構與語意角色" 訊息。一般看到的句結構樹是只有標註語法結構訊息，較少有語意角色訊息。Chen等[4]提到了語意角色的定義與考量，想利用單純句法限制去定義中文的關係是困難的。在中文句結構樹中，特別的是加上了語意的訊息，意指在瞭解到每一個成員(constituent)與其它成員的關係。舉例如下：

他　叫　李四　撿　球.
Ta　jiao　Li-si　jian　qiu.
"He asked Lisi to pick up the ball."
S(agent:NP(Head:Nhaa:他)|Head:VF2:叫|goal:NP(Head:Nba:李四)|theme:VP(Head: VC2: 撿|goal:NP(Head:Nab:球)))

圖1. 他叫李四撿球

圖1表示完整的中文句結構樹內容，標示了詞組結構(phrase structure)規則及語法與語意關係。對於句中標示每個詞的語法詞類意義，詳細定義與例子說明可參照[1]的技術報告手冊。

1.2 研究方法

二個操作策略：(1)將語法規律普遍化 及 (2)將語法規律精確化。普遍化的結果是增加語法覆蓋率，但同時可能產生的副作用就是降低了精確性與增加歧義性。一般的普遍化方法是放鬆詞類限制及增加不同的詞組規律，反之精確化的方式是增加詞類及特徵(feature)限制，使語法規律更精確。由於普遍化和精確化的操作互為制衡，常有過與不及的情形發生，如何拿捏控制找到有效的普遍化及精確化的方法是本論文探討的重點。

在語法規律精確化方面，Charniak[2,3]與Johnson[8]都有相同的結論，就是，對於非詞彙化PCFG在每個詞組節點加上子節點成員或是上層節點的特徵訊息，對剖析的精確率會有所提升。Zhang等[12]也提到，在詞組節點上加入上層詞組詞類與出現位置順序同樣也會對剖析結果有一些改進。Klein和Manning[9]與Collins[6]則提到對CFG加上中心語，也是有所幫助。

我們研究的方法就是試著從訓練語料中，依不同的語法抽取原則與語言學特性，抽取出不同的PCFG，再對測試語料進行剖析與結果評估。觀察結果並調整語法抽取原則，再次產生PCFG。從反覆的實驗中，找出較佳的句法規則來。

在普遍化的方法中，詞類限制放鬆在 Chen和Hsieh[5]有基本的研究結果。在語法覆蓋率增加方面，本文提出二元化句法規則產生方法，3.2節有詳細說明。在精確化方面，我們嘗試了不同的附加特徵，其中以中心語及最左成份特徵對剖析結果有最明顯的改進。其它特徵的實驗結果也會在第6節討論中說明。

2 系統設計架構

整個系統流程包含句法抽取、剖析與結果評估三大部份，圖2表示了整個的系統操作流程。對於每個部份的處理內容與方式，將分三小節來說明。

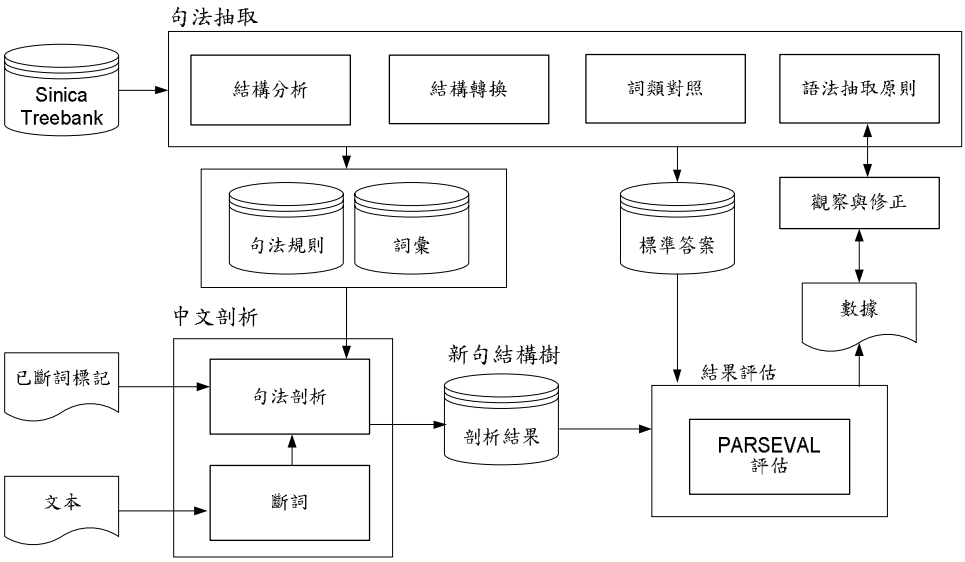


圖2. 系統架構圖

2.1 句法抽取

主要動作就是從樹庫中依語法抽取原則抽取出我們要的句法規則，並進行部份規則的調整與計算出相關的機率值，目的在供剖析器作為中文剖析的依據。說明如下：

- 從中研院中文句結構樹庫整理出完整且可以處理的樹
- 依語法抽取規則自動抽取出我們要的句法規則

- Charniak[3]提到很直覺的統計公式，對每條句法規則計算其機率值。

$$\hat{P}(N^i \rightarrow x^j) = \frac{C(N^i \rightarrow x^j)}{\sum_k (N^i \rightarrow x^k)}$$

- 建立虛規則列表，以解決Top-Down模式的中文剖析問題。如：起始符號為 S 的問題。加入 S' → VP、S' → NP…等。
- 去除 A → A的規則，避免出現遞迴(recursive)的情形，導致剖析器出現問題。

2.2 中文剖析器

PCFG的句法分析中，最廣泛採用的為LR與Chart演算法。而Earley's Algorithm是在英文剖析中，最常用被拿來實作的一個演算法[7]。它的處理特性為上到下(top-down)、預測(predictive)與左到右(left-right)，當應用在中文時，我們保有原有的predict、scan與complete這三個分析動作，同樣是分析到有節點到S'為止。為了符合我們對文本輸入格式[1]與產生句結構樹結構[2]的需求，對剖析器增加了資料輸入格式分析與處理虛規則的能力。

2.3 評估模型

最常用來評估樹結構括號(bracket)與詞組標記(label)好壞的評估模型是PARSEVAL[7]]。如果樹結構括號有錯，詞組詞類標記位置也會是錯的。所以，當括號的精確率愈高，連帶詞組標記的精確率也會愈高。我們採用如下評估項目與公式如下：

- 詞組標記精確率 LP(Labeled Precision)

$$\mathrm{LP} = \frac{\#\,\text{label correct constituents in parser's parse of testing data}}{\#\,\text{label constituents in parser's parse of testing data}}$$

- 詞組標記召回率LR(Labeled Recall)

$$\mathrm{LR} = \frac{\#\,\text{label correct constituents in parser's parse of testing data}}{\#\,\text{label constituents in treebank's parse of testing data}}$$

- 詞組標記效能評估LF(Labeled F-measure)

$$\mathrm{LF} = \frac{\mathrm{LP} * \mathrm{LR} * 2}{\mathrm{LP} + \mathrm{LR}}$$

- 括號精確率BP(Bracketed Precision)

$$\mathrm{BP} = \frac{\#\,\text{bracket correct constituents in parser's parse of testing data}}{\#\,\text{bracket constituents in parser's parse of testing data}}$$

- 括號召回率BR(Bracketed Recall)

$$\mathrm{BR} = \frac{\#\,\text{bracket correct constituents in parser's parse of testing data}}{\#\,\text{bracket constituents in treebank's parse of testing data}}$$

- 括號效能評估BF(Bracketed F-measure)

$$\mathrm{BF} = \frac{\mathrm{BP} * \mathrm{BR} * 2}{\mathrm{BP} + \mathrm{BR}}$$

另外，我們增加幾個觀察變數，以加強瞭解句法規則有用程度。說明如下：
RC-Type：句法規則覆蓋率－類型(type coverage of rules)
RC-Token：句法規則覆蓋率－頻率(token coverage of rules)
PA：句子可以剖析的比率

[1] 輸入格式為： 他(Nh) 叫(VF) 李四(Nb) 撿(VC) 球(Na) ，(COMMACATEGORY)
[2] 輸出格式為：#1.1..[0] S(NP(Head:Nh:他)|Head:VF:叫|NP(Head:Nb:李四)|VP(Head:VC:撿|NP(head:Na:球)))#，(COMMACATEGORY)

PC：句子剖析中，樹結構全對的比率
LF-1：僅對可剖析的句結構樹進行 LF 的評估。
BF-1：僅對可剖析的句結構樹進行 BF 的評估。

3 語法普遍化

Chen和Hsieh[5]抽出的句法規則，它是保有完整的詞組規則，我們稱為長詞組規則。對於長詞組規則的低覆蓋率問題，我們歸納出幾個可以對語法普遍化的方法，除了將詞類做粗細的變化，也試著將長詞組規則進行二元化的轉換。

3.1 詞類簡化

規律右邊詞類簡化意義在使一條規律能通用在更多不同的詞彙上，換句話說一個詞能夠應用到更多條句法規則上。詞類愈細，可以用到的句法規則就愈少，反之就愈多。而過度的詞類簡化會造成精確率的下降，太細的詞類，雖然提升剖析精確率，卻也降低語法覆蓋率，使得無法剖析的句子增加。詞類簡化例子如圖3所示，至於詞類簡化的原則，可以參照附錄1的詞類對照表。

原始詞組規則	S **à** agent:NP time:Dd manner:Dh Head:VA4
細詞類(Sinica Treebank詞類)	S **à** NP Dd Dh VA4
粗詞類(斷詞標記詞類)	S **à** NP D D VA
最粗詞類	S **à** NP D D V

圖3. 詞類簡化例子

從Chen和Hsieh[5]所做的實驗結果來看，很明顯地證明上述的看法。從細詞類層級到最粗的單一詞類層級，抽出的語法覆蓋率從82%增到92%；但是語法歧義也從每一詞類平均適用規律132條增加到835條；而且對訓練語料的剖析正確率不升反降平均從70%降到62%。剖析結果精確率和句法規則的覆蓋率與精確性有相當程度的互動關係。因此，本論文探討如何提高句法規則的覆蓋率與精確性以增加剖析效能。其中面臨的一個最重大問題就是，句法規則愈細，可能因訓練資料量的不足，而無法產生高覆蓋率且精確的語法規律。於是，對長詞組規則的簡化是有必要的。

3.2 二元化句法規則

從過去的研究中可以知道，句法規則的操作直接影響到剖析的結果，因此句法規則的呈現方式在句結構樹中是一個可調整的變數。長詞組規則其機率值較低，且覆蓋率也會較低。在縮短詞組規則的研究中，Manning[7]提到Chomsky normal form(CNF)的處理方式，主要在增加剖析器的剖析速度。經由CNF處理後的語法與原來的語法是完全相同的，並不會改變句法規則的覆蓋率。因而，我們提出的二元化句法規則方法，其擁有CNF增加剖析速度的特性，又有增加語法覆蓋率與句法延伸性[3]。其語法上的意義在於，同一詞組下的附加成份(adjuncts)在二元化的語法規則中可以任意出現或不出現。

CNF句法的基本定義格式：A**à**a 或 A**à**B C。A、B與C為非終端符號(non-terminal)，a為終端符號(terminal)。'**à**' 左邊的成員我們稱之為LHS(left-hand side)，'**à**'右邊成員稱之為RHS(right-hand side)。在RHS中，最多為二個成員。非終端符號在句結構樹中為詞組詞類，如VP、NP、S…等；終端符號為詞的詞類，如Nh、Nb…等。簡單的CNF在英文的做法是分析每條詞組規則，從最左開始，反覆向右，兩兩切割規則，產生新的非終端節點，並標記Rx，直到每條規則都符合上述的定義。二元化句法規則與CNF的差異，可以從表1得知。

[3] 所謂延伸性，如表 1(b)的二元化句法規則所示，VP**à**D VP 與 VP**à**D VA 這二條句法規則中的 D 即擁有延伸性。如 S **à** NP D D D VA 在二元化後，也是只有這二條句法規則，D 是可多可少的成份。

表格1. CNF與二元化句法規則比較

原始句法規則：(以斷詞標記詞類層級為例)　S **à** NP D D VA	
(a) CNF 結果	(b) 二元化句法規則
S **à** NP R0 R0 **à** B0 R1 R1 **à** B0 B1 B0 **à** D B1 **à** VA	S **à** NP S S **à** D S S **à** D VA

二元化句法規則的產生方式與CNF類似，不同處是在分析每條詞組時，從最右節點開始向左，兩兩切割詞組，不僅考慮到區域(local)的新二元化成員與詞組詞類，也考慮到原本整體(global)詞組詞類與中心語訊息。新規律的左邊成份Rx的操作，R必須維持原有詞組名外，x部份表示俱有附加特徵區別多樣化，在不同的語法抽取原則下，x也會選有不同的特徵值，而抽出的句法規則也就不同，這都會影響句法剖析的結果。舉例說明，我們對現有中文句結構樹庫中每一顆樹進行二元化的轉換，例1為原來的句結構樹；例2為二元化不加任何特徵；例3 Rx為詞組加中心語特徵；例4 Rx為詞組加成員特徵。

S(NP(Nh:我們)|D:常常|D:一起|VA:上學)	(例1)
S^{-} (NP^{-} (Nh:我們)| S^{+} (D:常常| S^{+} (D:一起|VA:上學)))	(例2)
S^{-}_{VA} (NP^{-}_{Nh} (Nh:我們)| S^{+}_{VA} (D:常常| S^{+}_{VA} (D:一起|VA:上學)))	(例3)
$S^{-}_{NP,S}$ (NP^{-}_{Nh} (Nh:我們)| $S^{+}_{D,S}$ (D:常常| $S^{+}_{D,VA}$ (D:一起|VA:上學)))	(例4)

註：S-為終節點(原節點)，S+中間節點(虛節點)。

另外，在二元化句法規則中，我們建立特徵的觀念。特徵會因語法抽取的定義不同而有所改變，一般用在語法精確化的處理上。如果不加任何特徵只處理詞類的部份，就是最原始的語法普遍化。第4節將討論如何運用有語法支持的特徵。

4　語法精確化

上節討論的語法普遍化，雖然覆蓋率會有所提升，但句法剖析的精確率卻因語法歧義的增加而降低。語法精確化目的在提高句法剖析的精確率，句法規則覆蓋率又不會降低太多。主要研究內容為語法精確化的操作方式與特徵的選擇。

4.1　詞類精確化

在斷詞標記的詞類中，中研院詞庫小組[1]對詞類有完整的定義與說明。基本上相同語法行為的詞被歸為同一詞類，詞類精確化目的是針對某些詞類在句法剖析時易造成混淆的情況，進行更細的分類，使句法剖析有較好的結果。例如從樹庫我們觀察到DE詞類在PCFG中的問題。在中文詞中，詞類標記為DE的有“得、地、之、的”四種。而從各別的語法特性，進一步瞭解到不應該在PCFG的規則中混在一起。根據觀察與分析得到可以分成三組：”的、之”、“得”與“地”。因此，我們將DE這類精確化後成DE、DE1與DE2，舉例說明如下：

的、之：炙熱**的**太陽、言下**之**意。 得：說**得**容易做**得**難 地：高興**地**唱歌

4.2 特徵限制

本論文研究了許多不同組合的特徵限制，並歸納出提了二個最重要的特徵限制：中心語訊息與詞組成份訊息，目的都在使語法歧義減少，增加剖析器對歧義結構的鑑別能力。在特徵的限制上，中心語是很重要的訊息。在中文句結構樹中，每一種詞組NP、VP、S、GP…等都各自不同中心語詞類，其語法規律也不同。因此，將詞組的中心語詞類加入二元化句法規則是非常很重要的，相信對於句法剖析的精確率會有所提升。例子5說明了中心語加入的結果。

句結構樹：VP(D:終於|VC:到|Di:了)

二元化：VP^{-}_{VC}(D:終於|VP^{+}_{VC}(VC:到|Di:了)) (例5)

另外，參考Johnson[8]提到在詞組上加入子節訊息能有較佳的結果，於是，我們提出二元化過程中，對每個二元化詞組加入了最左成份的詞類訊息。第5節實際的實驗結果顯示中心語及最左成份的特徵對語法規率精確性的確有幫助，且不會降低太多語法覆蓋率，對整體的剖析結果有顯著的改善。

5 實際操作

本實驗方式在圖2中有一個語法抽取原則處理部份，依不同的方法，產生不同的句法規則，剖析器再依此句法規則進行測試語料的剖析，並評估剖析的結果。反覆的實驗與觀察，找出最好用的語法抽取原則出來。

訓練語料與測試語料的介紹：中研院中文句結構樹庫的訓練語料統計情形，包含37,889句數、235,669詞數與6.22每句平均詞數訊息。我們從不同領域中取三份測試語料(Sinica、光華雜誌與南一課本)進行實驗。表說明了這三份語料的句長分佈、句數與難易程度。

這三份測試語料句法難易程度有所不同。Sinica語料與訓練語料句法較為相近的，程度屬適中。南一課本為小學正式上課教材，句子不會太長，程度屬簡單。光華雜誌為市面上出版書目，句法嚴僅，程度屬較難。

表2. 三份測試語料內所包含句子的資訊

測試語料	難易度	0-5 詞句數	6-10 詞句數	11 詞以上句數	總計
Sinica	適中	612	385	124	1,027
光華	較難	428	424	104	956
南一	簡單	1,159	566	25	1,750

5.1 詞類普遍化

Chen和Hsieh[5]的實驗顯示一但語法覆蓋率超過90%，增加訓練資料量，很難快速的增加語法的覆蓋率。因此，在有限的訓練語料下，操作語法規律的普遍化是有其必要性。對於詞類普遍化的做法，本文3.1節中已有相關實驗的結果討論。在我們的實驗中，詞類對照的部份，同樣參照附錄1的對照原則，例子如圖3所示。分別對這三份測試語料實作的結果如表3所示，以最粗詞類層級的語法覆蓋率98.303%為最高，但其剖析效能只有74.29%，且最粗詞類並無法對詞性有效區分；另外，細詞類層級的語法覆蓋率約91%，明顯較其它二組低，剖析效能為76%，也不是最好的；因此，粗詞類層級(斷詞標記詞類)的實驗結果有其操作空間，往後實驗目標是將長詞組規則二元化以提高語法覆蓋率，並試著從語法精確化的過程中，在覆蓋率減少的容許範圍內，提升剖析效能。

表3. 詞類普遍化實驗

	細詞類層級			粗詞類層級			最粗層級		
	Sinica	光華	南一	Sinica	光華	南一	Sinica	光華	南一
RC-Type	63.015	73.194	71.806	74.681	82.154	79.877	83.269	86.581	86.330
RC-Token	86.358	91.195	92.928	93.223	96.033	96.506	97.289	98.303	98.679

PA	96.88	98.22	98.06	99.32	99.69	99.66	100	100	99.49
LF	76.84	75.11	82.42	82.80	77.82	84.04	81.22	74.29	80.97
BF	82.19	81.22	86.70	87.92	84.27	88.56	86.00	80.87	85.47
LF-1	79.31	76.48	84.05	83.37	78.07	84.33	81.22	74.29	81.38
BF-1	84.83	82.69	88.41	88.52	84.53	88.86	86.00	80.87	85.91
PC	46.25	40.48	57.89	58.13	46.44	63.66	59.69	39.85	61.14

5.2 二元化句法規則vs. 長詞組規則

從前面章節瞭解二元化句法規則的優點，而長詞組規則為語法抽取的最原始規則。有關二元化的例子，可參照3.2節說明，本節就實作與結果進行討論。

實驗結果如表4所示。二元化句法規則擁有99%的高語法覆蓋率，對於測試語料的句子，幾乎都可以剖析出來，但是，LF-1與LF反而下降了約0.5%到1.8%。句法規則不加任何特徵限制，則將可以剖析出所有的樹，PA值大於99.9%，但過度普遍化卻造成了剖析的精確率的下降。接下來的研究就在於如何去提升剖析的精確率。

在剖析速度的分析方面，二元化句法規則的剖析時間明顯的較長詞組規則快，這是句法剖析過程中，二元化句法規則可以比長詞組規則較快去過濾掉不必要的句法候選規則，減少運算量。

表4. 長詞組規則與二元化句法規則實驗結果

	(a)長詞組規則			(b)二元化句法規則		
	Sinica	光華	南一	Sinica	光華	南一
RC-Type	74.681	82.154	79.877	96.154	94.657	94.761
RC-Token	93.223	96.033	96.506	99.590	99.362	99.548
PA	99.32	99.69	99.66	99.90	99.90	100
LF	82.80	77.82	84.04	81.45	76.16	83.50
BF	87.92	84.27	88.56	87.71	83.58	88.38
LF-1	83.37	78.07	84.33	81.53	76.24	83.50
BF-1	88.52	84.53	88.86	87.79	83.67	88.38
PC	58.13	46.44	63.66	52.29	42.36	60.00

5.3 特徵限制：中心語

二元化過程中，我們對詞組部份加入了中心語特徵，能保有原詞組的語法特性，又可避免句法規則的過度普遍化。舉例來說：”一輛大型玩具機車”的長句法規則為NP à DM A Na Na，在二元化後為 NP^{-} à DM NP^{+}，NP^{+} à A NP^{+} 與 NP^{+} à Na Na，加入中心語特徵限制後為 NP^{-}_{Na} à DM NP^{+}_{Na}，NP^{+}_{Na} à A NP^{+}_{Na} 與 NP^{+}_{Na} à Na Na。從表5的結果中知道，在加入中心語特徵後，語法覆蓋率下降0.3% 左右，而剖析的精確率確提升1.4% 左右。證明了在二元化句法規則中加入中心語的特徵是有用的。

表5. 中心語特徵限制實驗結果

	中心語		
	Sinica	光華	南一
RC-Type	95.824	94.124	94.456
RC-Token	99.273	99.026	99.334
PA	99.61	99.90	99.94
LF	82.62	77.50	84.52
BF	88.65	84.47	89.02
LF-1	82.95	77.58	84.57
BF-1	89.00	84.56	89.08
PC	60.18	45.40	64.80

5.4 特徵限制：最左成份

另一個特徵限制的實驗，是在詞組上加註左邊成份的特徵。這個特徵的意義是說，成份和成份之間有順序及搭配的限制。例如上節例句‘一輛大型玩具機車’中的附加成份定量詞(DM)、形容詞(A)及名詞修飾語(Na)有順序上的限制，通常定量詞一定是在最外層。同前一節的例子，在詞組節點加入最左成份特徵限制後的二元化句法規則為中心語特徵限制後為 $NP^{-}_{Na,DM}$ à DM $NP^{+}_{Na,A}$， $NP^{+}_{Na,A}$ à A $NP^{+}_{Na,Na}$ 與 $NP^{+}_{Na,Na}$ à Na Na。表6分別顯示不同的實驗結果：(a)單獨加左成份與(b)加中心語和左成份。

從表6(b)與表4(b)來看語法覆蓋率，因特徵限制的增加，減少至98.5%左右。剖析的精確率則增加2%到4%不等。觀察測試語料的特性，在Sinica增加4.2%，其它光華與南一測試語料分別為2.88%與1.91%。這說明了訓練語料與Sinica測試語料句法結構較為相似，其它則不然。表6(a)與表6(b)的差異在中心語加入與否，從數據看來，加入中心語的PC值有些許提升，整體剖析效能則差不多。

表6. 最左成份實驗結果

	(a) 左成份			(b) 中心語 + 左成份		
	Sinica	光華	南一	Sinica	光華	南一
RC-Type	95.633	94.939	94.900	93.977	92.988	92.557
RC-Token	99.236	99.106	99.345	98.602	98.500	98.810
PA	99.90	99.79	100	99.42	99.79	99.94
LF	85.62	79.82	85.74	85.64	79.04	85.41
BF	89.75	85.64	89.52	89.84	85.29	89.42
LF-1	85.70	79.98	85.74	86.14	79.21	85.46
BF-1	89.83	85.82	89.52	90.37	85.47	89.47
PC	64.65	48.54	65.83	65.73	49.06	66.23

5.5 詞類精確化：”DE”處理

從4.1的說明中，知道DE可分為三組語法特性。這樣的特徵限制，屬於區域限制，目的在於降低詞類DE造成的混淆。在我們實作中，為了鑑別不同的詞所代表的 DE，因此給予DE不同的編號。例如：“的、之“ 標示為DE、 “得” 標示為DE1、”地”標示為DE2。我們設計三種實驗方式 (a)單獨在二元化後加入DE特徵；(b)依續5.3節的研究，再加入DE特徵；(c)依續5.4的研究，再加入DE特徵。實驗結果如表7所示。表7(a)與表4(b)的實驗差別在DE特徵限制的部份，剖析精確率在光華語料中有0.55%的提升，其它則不太明顯。

表7(c)有最佳的剖析結果，表7(b)相較於表5的數據，也只是微幅的提升。因此，從數據看來，DE的加入，或多或少對剖析結果還是有影響的。

表7. ‘DE’ 實驗結果

	(a) DE			(b) 中心語+DE			(c) 中心語+左成份+DE		
	Sinica	光華	南一	Sinica	光華	南一	Sinica	光華	南一
RC-Type	96.161	94.368	94.611	95.829	93.920	94.352	93.982	92.840	92.478
RC-Token	99.590	99.330	99.536	99.273	98.994	99.322	98.602	98.470	98.798
PA	99.90	99.90	100	99.61	99.90	99.94	99.42	99.79	99.94
LF	81.33	76.71	83.59	82.71	77.82	84.62	85.63	79.43	85.50
BF	87.78	84.19	89.00	88.87	84.89	89.67	89.97	85.73	89.98
LF-1	81.41	76.79	83.59	83.04	77.90	84.66	86.13	79.60	85.55
BF-1	87.87	84.28	89.00	89.21	84.98	89.72	90.49	85.91	90.03
PC	52.29	42.47	60.23	60.27	45.40	65.03	65.73	49.06	66.57

6 討論

在實作語法精確化的過程中，我們嘗試了不同的附加特徵，其中以第5節探討的最左成份及中心語特徵對剖析結果有最明顯的改進。在本節我們將一一討論其它的實驗結果。

語法抽取規律：詞組角色 "Head/head" 的保留。不同於圖3的例子，我們對於詞組角色為Head/head的部份進行保留並抽取規律與實驗。從結果得知，雖然整體剖析效能提高1%左右，但是語法覆蓋率卻降低。因此，在我們第5節的實驗中，以覆蓋率高的詞組角色 "Head/head" 不保留為實驗基礎。

二元化詞組的變化：原節點不加特徵限制。原節點的意思說明請參考3.2節，在我們最早的實驗中，對於原節點的部份，並不加入任何特徵限制訊息，保有原來詞組詞類，只專注在虛節點的處理上。後來改進對原節點同步處理時，發現整體的剖析效能明顯地提升。因此，我們在實驗過程中，談到語法精確化部份，都是同步對原節點與虛節點處理。

二元化方向。從左向右二元化，在實際剖析上，很明顯地剖析速度比較慢，主要原因是剖析器運用為由上到下的句法規則連結，而剖析方向為從左到右，對一個要處理的詞類，向右二元化的句法規則產生的規則候選集較向左句法規則來的多，剖析運算量也比較多，這就是我們不採用的原因。另外一個二元化的方向，是以中心語為開始點，先向右二元化，再向左二元化。從實際剖析結果來看，並沒有比向左二元化來的明顯較好，且建立的方法又繁雜許多，於是不用採它。

特徵限制：加入上層詞組詞類訊息。Johnson[8]與Zhang[12]的內容中提到加入上層詞組詞類訊息可以提升剖析效能。我們實作將上層詞組詞類加入特徵限制後，因語法規律更為精確，覆蓋率明顯地下降，LF-1值確實有些許提高，但從整體剖析效能來看影響不太。因此，該方法並不適合我們使用。

角色普遍化。中文句結構樹的角色，是一個可以嘗試普遍化之處，我們試著觀察哪些角色可以代替詞類或詞組用在語法規律上。以time為例，在time:Dd或time:NP的語法規律相近，可以用time代替達到角色普遍化目的。從剖析結果看來，每一個動作的語法覆蓋率都有些許提升，合併使用多個角色普遍化時提升更多，但是整體對於剖析效能卻影響不太。此外，如果選錯角色進行普遍化，卻造成剖析效能明顯下降。

7 結論與未來研究

本論文裡，我們建構一完整的實驗平台，其中包含符合語料輸入輸出格式與處理需求的中文剖析器。剖析器所參考的句法規則，從最原始的長詞組規則到二元化句法規則，從沒有加入任何特徵的語法普遍化到加上特徵限制的語法精確化，每一個動作都有相關的實驗結果評估。從實驗數據得知，建構二元化句法規則的方法與特徵的選擇，對中文剖析有直接的影響，而二元化句法規則有比長詞組規則較好的剖析結果，剖析速度也相對快些。當句法規則愈精確化，其句法規則覆蓋率會降低一些與剖析效能會有所成長；句法規則的普遍化增加了語法歧義，可以利用其它的知識來幫忙解決結構歧義的問題。You和Chen[11]的研究說明了詞與詞相關訊息(word-association)的應用，不僅僅是可以對剖析好的樹結構進行語意角色的指定，也可以用來幫忙解決歧義句結構的問題。因此，在剖析器的設計上，可以進一步的將詞與詞相關的訊息整合進來，也就是用PCFG來提供剖析器句法結構的選擇，再利用句法規則的機率和詞與詞相關的機率來指導結構的挑選。

中心語加左成份似乎並沒有比僅加左成份更好，主要是中心語限制降低了語法的覆蓋率，相信在更大的訓練語料下，有中心語特徵的語法規律會比不採用中心語特徵的語法規律表現得更好。並且經由中心語的限制可以增加剖析的效率，因為和輸入句不相關的規律，也就是說語法規律指定的中心語詞類不在句中出現，此規律可以提早刪除。

從實驗的結果可以得到另一個結論，特定領域的句法規則學習是有必要的，不同領域的文章會有其專門句法表示，比如，新聞類、財經類或教材課文類的文本，句法上彼此會有些許的不同；最好的方法就針對特定領域處理，比如要剖析財經類文章，就用從財經領域抽取的句法規則來剖析，以取得較好的剖析結果。另外，Verdú-Mas、Calera-Rubio與Carrasco[10]提出一個平滑技術(smoothing techniques)是將二種較粗與較細的句法規則混合使用，以提升剖析的能力，這是我們未來可以參考實驗的方向。

致謝

本論文的研究得到國科會計畫編號 NSC93-2752-E-001-001-PAE 及 NSC93-2422-H-001-0004 的部分補助。

參考文獻

[1] 中央研究院詞庫小組, "中文詞類分析（三版）." CKIP Technical Report No.93-05.

[2] Charniak, E. and Carroll, G., "Context-sensitive statistics for improved grammatical language models." In *Proceedings of the 12th National Conference on Artificial Intelligence*, AAAI Press, pp. 742-747, Seattle, WA, 1994,

[3] Charniak, E., "Treebank grammars." In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pp. 1031-1036. AAAI Press/MIT Press, 1996.

[4] Chen, Feng-Yi, Tsai, Pi-Fang, Chen, Keh-Jiann, and Huang, Chu-Ren, "Sinica Treebank." *Computational Linguistics and Chinese Language Processing*, 4(2):87-103, 2000.

[5] Chen, Keh-Jiann and Hsieh, Yu-Ming, "Chinese Treebanks and Grammar Extraction." the *First International Joint Conference on Natural Language Processing* (IJCNLP-04), March 2004.

[6] Collins, Michael, "Head-Driven Statistical Models for Natural Language parsing." Ph.D. thesis, Univ. of Pennsylvania, 1999.

[7] Manning, Christopher D. and Schutze, Hinrich, "Foundations of Statistical Natural Language Processing." the MIT Press, Cambridge, Massachusetts, 1999.

[8] Johnson, Mark, "PCFG models of linguistic tree representations." *Computational Linguistics*, Vol.24, pp.613-632, 1998.

[9] Klein, Dan and Manning, Christopher D., "Accurate Unlexicalized Parsing." *Proceeding of the 4lst Annual Meeting of the Association for Computational Linguistics*, pp. 423-430, July 2003.

[10] Verdú-Mas, Jose L., Calera-Rubio, Jorge and Carrasco, Rafael C., "Smoothing Techniques for Tree-k-Grammar-Based Natural Language Modeling." *IbPRIA* , pp. 1057-1065, 2003.

[11] You, Jia-Ming and Chen, Keh-Jiann, "Automatic Semantic Role Assignment for a Tree Structure."the *Third SIGHAN Workshop on Chinese Language Processing*," July 2004.

[12] Zhang, Hao, Liu, Qun , Zhang, Kevin, Zou, Gang and Bai, Shuo, "Statistical Chinese Parser ICTPROP." Technology Report, Institute of Computing Technology, 2003.

附錄1 Syntactic Category Mapping

Level2	Level3	Level4	Nc*	Nc	N	V_2	V_2	V
Caa	Caa	C	Ncd*	Ncd	Ncd	VA*	VA	V
Cab	Cab	C	Nd*	Nd	N	VA2	VAC	V
Cba	Cba	C	Nep	Nep	Ne	VB*	VB	V
Cbaa	Cbb	C	Neqa	Neqa	Ne	VC1	VCL	V
Cbab	Cba	C	Neqb	Neqb	Ne	VC*	VC	V
Cbba	Cbb	C	Nes	Nes	Ne	VD*	VD	V
Cbbb	Cbb	C	Neu	Neu	Ne	VE*	VE	V
Cbca	Cbb	C	Nf*	Nf	N	VF*	VF	V
Cbcb	Cbb	C	Ng	Ng	Ng	VG*	VG	V
D*	D	D	Nh*	Nh	N	VH*	VH	V
Dab	Da	D	Nv1	Nv	N	VH16	VHC	V
DE	DE	DE	Nv2	Nv	N	VH22	VHC	V
Dfa	Dfa	D	Nv3	Nv	N	VI*	VI	V
Dfb	Dfb	D	Nv4	Nv	N	VJ*	VJ	V
Dk	Dk	D	P*	P	P	VK*	VK	V
I	I	I	T*	T	T	VL*	VL	V
Na*	Na	N	V_11	SHI	V	DM	DM	DM
Nb*	Nb	N	V_12	SHI	V	Di	Di	D

A Resolution for Polysemy: the case of Mandarin verb *ZOU* (走)

Yaling Hsu Meichun Liu
National Chiao Tung University
yealings@ms38.hinet.net mliu@mail.nctu.edu.tw

Abstract

In this paper, we propose a procedural schema as a model of cognitive processing of word senses, which can be viewed as a derivational resolution of polysemy. Previous researches, such as Frame-Based Lexicon by Fillmore [4] and Lexical Semantics by Cruse [2], are all concerned with word senses, but what is still missing is a holistic resolution of polysemy. Therefore, in this paper, we focus on the cognitive process from word form to word senses, based on corpus-based procedural resolution. In this way, we hope to provide an overall discussion and a computerizable way of solving multiplicity of semantic usages of a single word form. A case study of the Mandarin verb *ZOU* (走) is presented and used as an illustration.

1 Introduction

Since ‘two or more semantic elements may be expressed in a single monomorphemic lexical item’ (Bybee [1]), to understand the meaning of a word in a particular utterance, we need to resort to ‘cognitive structures, knowledge of which is presupposed for the concepts encoded by the words’ (Fillmore [4]). According to Fillmore [4], we know word senses are not related to each other directly, but only by way of their links to common background frames and indications of the participant roles associated with such frames (i.e., Frame elements). However, when we turn to semantic multiplicity of a single word form in Mandarin, such as *ZOU* (走), the highlighted core elements of frames may not be enough to help us distinguish the different meanings of the single monomorphemic lexical item. The problem can be the spelled out as follows:

a) Different senses of a single lexical item may have similar participant roles in the general terms and similar patterns of expressing these elements. Therefore, if we only depend on the information of core frame elements, how could we tell the different senses and in what way we can tell the non-prototypical senses from the prototypical one? Since the process of sense selection is under the force of many interacting factors (Bybee [1]), a reliable source of clues is collocational patterns that reveal lexical as well as grammatical associations of words. To fully utilize corpus data, we will look at Colloconstruction (a term adopted from Stefanowitsch and Gries’ idea with some modification), i.e., clause-internal, morpho-syntactic patterning characteristic of each sense, to further distinguish semantic polysemy.

b) With the postulation of Colloconstruction, we may still encounter ambiguous cases where two different senses may share similar frame elements and similar Colloconstrucions. Thus, next in our cognitive resolution, we propose ‘Contextual Dependence’ as another disambiguation factor which depends on discourse-level patterning across sentences, and we will have a detailed discussion in the following sections.

2 Cognitive resolution

The resolution model proposed here intends to simulate human cognitive process of detecting word senses. As Cruse [2] describes, a lexical form may well be associated with an unlimited number of possible senses, but **these are not all of equal status** (bold is added by us)…every lexical form has at least one relatively well-utilized sense. Our resolution is based on Cruse’s observation and the assumption made by cognitive linguists that each word has at least one cognitively most salient meaning, the prototypical sense. First, a single word form within a clause comes into our cognitive system, and then according to the salient frame-evoked elements, we might easily get one sense from the word form (as shown by the arrowed line ‘a’ in Figure 1 below.). In most cases, readers tend to start with the predominant sense with the highest frequency count (we will discuss this in the following sections). However, some words may have two different senses that share similar participant roles and surface patterns, and then we need an efficient mechanism to detect the different sense while probing into the underlining frame. In these cases, we need to go through the next step—identifying Colloconstruction (as shown in the following) to get more information to help delimit the different senses. Colloconstruction provides information regarding morpho-syntactic patterns within a particular construction which consists of frequently co-occurring lexemes. Still, in some cases, Colloconstructions are not distinct enough to disambiguate. There might be another sense which requires similar core elements in a similar Colloconstruction as the more prototypical sense does. Then, we have to go into the next step - finding ‘Contextual Dependence’, i.e., discourse-pragmatic variables commonly associated with a given sense. The resolution formula is schematically

represented in Figure 1 below. Assuming that the most prototypical and thus more frequently used sense is easier to detect, the process starts with checking the highlighted frame elements and the high frequently use for identifying the prototypical sense. As shown in Figure 1, the path with the arrowed line ‘a’ represents this shortest route – frame element checking. The paths with the arrowed lines ‘b’ and ‘c’ represent the additional efforts required for identifying less prototypical senses.

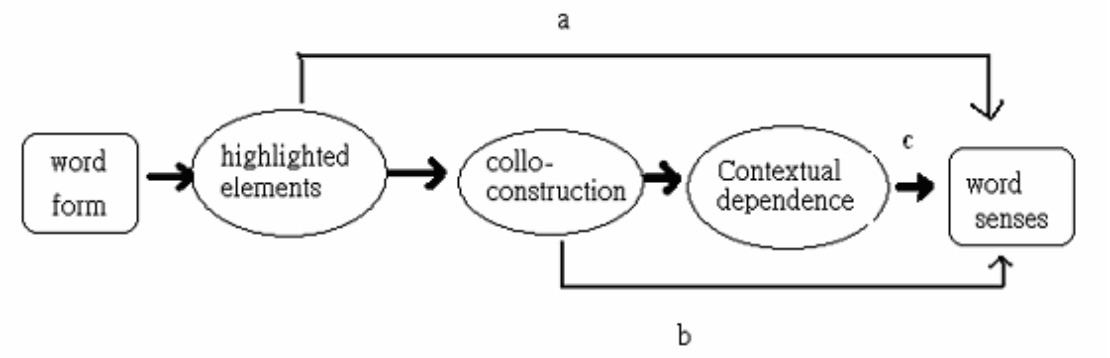


Fig. 1. Cognitive resolution

3 The different senses of Mandarin verb ZOU(走):

In this paper, we use the Mandarin verb ZOU(走) to test and illustrate our cognitive resolution. First, we will distinguish the different senses of the verb by frame conceptions (adopting the frame definitions in FrameNet II with little modifications), as shown in Table 1.

Table 1. The four main senses of Mandarin verb ZOU(走)’[1]

Sense	**Frame**	**Frame Elements**	**Frequency (Total: 200)**
Sense1: walk/go	Self_motion	Area, goal, path, source, self-mover, duration	135 (67.5%)
Sense2: move	Motion	Area, goal, path, source, theme	10 (5%)
Sense3: visit	Arriving	Area, goal, self-mover	9 (4.5%)
Sense4: leave	Path_shape	Path, path_shape, road, self_mover, duration	46 (23%)

As we can see in Table 1, all the senses are in different frames with some shared core frame elements. In what way, then, can we identity these different senses by their frame elements? Besides, how do we distinguish the different senses when they are composed of the same pattern with the same instantiated frame elements? In order to provide an overall analysis of semantic polysemy, we propose fluid routes for cognitive resolution.

4 Frame-based sense distinction

As we mentioned above, the predominant sense goes through fewer steps since it is cognitively more accessible. Take the verb ZOU(走) in Chinese as an example. Among the four possible senses, sense 1 occurs most frequently (as shown in Table 1) and denotes a specific sensory motor action that is assumed to be cognitively salient and prototypical. Sense 2 can be viewed as extended from sense 1 in that the human action of moving by walking is broadened to denote the moving of entities in general. While sense 1 and sense 2 are both motional and they share a number of core frame elements, the two meanings can be easily distinguished in terms of the semantic attributes of participate roles. That is, sense 1 is associated with human or animate self-mover, and sense 2 is associated with inanimate moving entities or ‘theme’, as exemplified in the examples (1) and (2) below.

(1) Sense 1
Self-mover [animate] <*< Distance
他也不知道究竟走了多遠，終於在一個荒僻的大山下面，發現了一個山洞。
ta1 ye3 bu4 zhi dao4 jiu4jing4 zou3 le duo1yuan3, zhong1yu2 zai 4yi1ge4 huang1pi4 de da4shan1 xia4mian4, fa1xian4 le yi1ge shan1dong4
He also not know actually walk LE how far, finally at one-CL desolate DE great mountain under, find LE one-CL cave

‘He also did not know how far he walked actually, and finally under the desolate great mountain, he found a cave.’

(2) Sense 2
Mover [inanimate] <*< Distance
大約１３０分鐘，火車走了約２００公里，我們到了統一前東德的第三大都市德勒斯登。
da4yue1 130 fen1zhong1, huo3che1 zou3 le yue1 200 gong1li3, wo3men dao4 le tong3yi1 qian2 dong1de2 de di4san1da4 du1shi4 de2le41si1deng1
about 130 minutes, train walk LE about 200 kilometers, we come LE unify before east German DE third big city Deluxe
‘About 130 minutes, the train walked about 200 kilometers; we came to the third metropolis, Deluxe, of ex- east German.’

Figure 2 below is meant to capture the details of the sense derivational process of the word form ZOU(走), and we will see that sense 1 and sense 2 are distinguished in the first step. Semantic information of their frame elements is utilized to process these two senses in cognition.

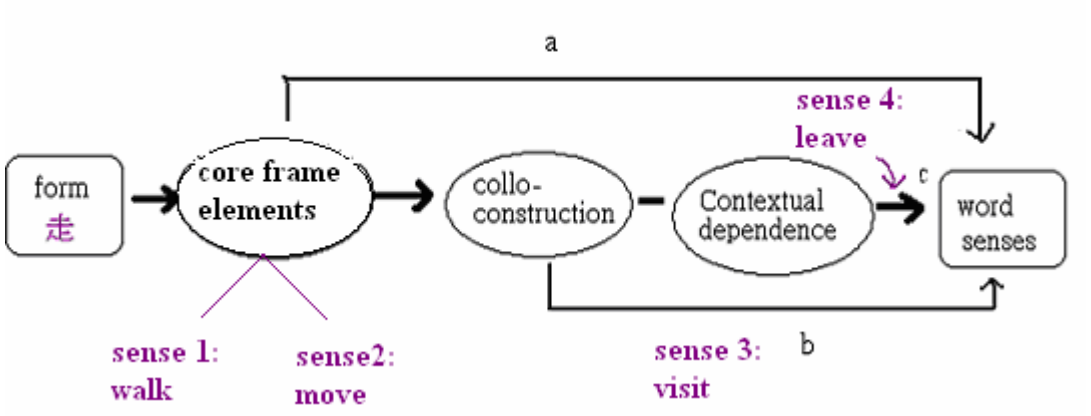


Fig. 2. The cognitive resolution of Mandarin verb ZOU(走)

5 Colloconstruction

Colloconstruction refers to a combination of lexical and grammatical collocations. It is used to identify the specific morpho-syntactic sequences of lexical items. Colloconstruction may help detect crucial collocational elements and constructional features when the word form is semantically compatible with the construction. When semantic information of participant roles is insufficient, word senses can only be detected with a careful examination of lexical and grammatical collocations. For example, sense 1 and sense 3 of ZOU(走) can only be distinguished when colloconstructions are taken into consideration. Consider the following uses of sense 3:

(3) Self-mover <*< path
民眾欲見南仁山區的森林生態系，只消走一趟南迴公路即可
min2zhong4 yu4 jian4 nan2ren2 shan1qu1 de sen1lin2 sheng1tai4xi4, zhi3 xiao1 zou3 yi1tang4 nan2hui2 gong1lu4 ji2ke3
people want see Nan-Ren mountain area DE forest ecosystem, only just walk once south highway all right
‘If people want to take a look around the forest ecosystem of the Nan-Ren mountain area, they may have a visit of the south highway’

(4) Self-mover <*< goal
臺灣本島的小朋友來說，要看看恐龍標本，只要走一趟科博館就可以了。
tai2wan1 ben3dao3 de xiao3peng2you3 lai2 shuo1, yao4 kan4kan4 kong3long2 biao1ben3, zhi3 yao4 zou3 yi1tang4 ke1bo2guan3 jiu4 ke3yi3 le
Taiwan insular DE kids come say, want see dinosaur specimen, only have to walk once science museum all right LE
‘For insular kids in Taiwan, if they want to see the dinosaur specimen, they may visit of the science museum all right’

(5) (CNI/self-mover) <goal<*
在小人國內走一遭，彷彿自己便是童話中的巨人格列弗。
zai4 xiao3ren2guo2 nei4 zou3 yi1zao1, fang3fu2 zi4ji3 bian4 shi4 tong2hua4 zhong1 de ju4ren2

ge2lie4fu2
in Lilliputian inside walk once, like oneself is nursery tale inside DE giant Grief.
'Visiting in Lilliputian, one may imagine themselves as is being the giant Grief in the nursery tale'

In these utterances, the delimiting phrases *yitang* (一趟) or *yizao*(一遭) combined with a Location are crucial indicators of the 'visiting' sense of ZOU(走). In other words, the word form ZOU(走) and the phrases *yitang* (一趟) or *yizao*(一遭) co-construct a specific sequence commonly associated with the sense of 'visiting'. Exactly, in what way can Colloconstruction help? The answer is: when frame-based semantic roles fail to disambiguate. Let's consider the following utterances which contain the uses of sense 1:

(6) Self-mover <*< path
接著是訓練他們走路。走斜坡，走臺階；走平路，也走不平的路。
jie1zhe shi4 xun4lian4 ta1men zou3lu4. zou3 xie2po1, zou3 tai2jie1; zou3 ping2lu4, ye3 zou3 bu4ping2 de lu4
then is train them walk. walk slope, walk step, walk even, also walk not even DE road
'Then, train them walk, walk slope, walk step, walk even road, and also walk uneven road'

(7) Self-mover <*< goal
我對她笑一笑走開了。仰頭一看，才知道走到一排松樹下，
wo3 dui4 ta1 xiao4yi1xiao4 zou3kai1 le, yang3tou2 yi1kan4, cai2 zhi1dao4 zou3dao4 yi1pai2 song1shu4 xia4
I to her simile walk away. Faced upward a look, just know walk to a line pine tree under
'I gave a smile to her and walked away. Then, I faced upward taking a look and found that I had walked under a line of pine trees.'

(8) (CNI/self-mover) < goal < *
往英國花園的西南邊走，是一個舊市區（Vielle Ville），
wang3 ying1guo2 hua1yuan2 de xi1nan2 bian1 zou3, shi4 yi1ge jiu4 shi4qu1 (Vielle Ville)
toward British garden DE southwest side walk, is an old downtown (VielleVille)
'Walk toward the southwest side of the British garden, there is an old downtown (VielleVille)'

As we can see, the examples (6)-(8) above contain uses of sense 1 and the core frame elements (Self-mover, Area, and Goal) are similar to those of sense 3 (examples in (3)-(4)). To detect the differences between these two senses, we need to pay attention to their collocational features. Here, the colloconstruction [ZOU(走) +*yitang* (一趟) /*yizao*(一遭)] help to identify the occurrence of sense 3. Therefore, as we proposed above, Colloconstructions might be the anchor for the derivational senses. In this case, the adjunct *yitang* (一趟), *yizao*(一遭) help to anchor sense 3 in a commonly recognized construction, taking the following NP as a destination (Goal) and then the sense 'visit' is derived. This resolution conforms to the perspective of Emergent Grammar, as Firth [5] contended that usage patterns of lexical forms can best be examined by looking at 'the company' they keep. However, given the dynamic nature of word usage, collocational associations alone may not be flexible enough to distinguish subtle differences of the senses of a word. Therefore, we need to take another step, looking into contextual dependence to obtain the overall resolution for polysemy.

6 Contextual dependence

The word form ZOU(走) has another sense —sense 4 'leave'—as shown in examples (9)-(11). Initially, we take the first step and test whether sense 4 can be derived only by utilizing information of core frame elements. Let's consider the following utterances:

(9) Self-mover<* (sense 4: leave)
於是大夥兒便分頭走了，帶著滿腔的興奮。
yu2shi4 da4huo3er2 bian4 fen1tou2 zou3le, dai4zhe man3qiang1 de xing4fen4
hence a group of people separately walk away LE, bring full De excitement
'Hence, the group of people walk away separately filled with excitement.'

(10) Self-mover<* (sense 4: leave)
一部摩托車，沒有腿的騎士，遠颺了…走了…

yi1bu4 mo2tuo1che1, mei2you3 tui3 de qi2shi4, yuan3yang2le…zou3le…
one-CL motorcycle, no leg DE knight, far wary LE…walk LE
‘A motorcycle, carrying a knight without legs, moved far away … leave …’

(**11**) Self-mover<* (sense 1: walk)

我在滿街水兵和軍官們中間走著，聽他們用熟悉的粗話互相笑鬧著、喧囂著，一直來到碼頭邊

wo3 zai4 man3 jie1 shui3bing1 han4 jun1guan1 men zhong1jian1 zou3zhe, ting1 ta1men yong4 shou22xi1 de cu1hua4 hu4xiang1 xiao4nao4zhe, yi1zhi2 lai2dao4 ma3tou2 bian1
I in full street soldiers and military officerses centre walk ZHE, hear them use familiar obscene language each other laugh ZHE, make hullabaloo ZHE, until arrive at wharf side.
‘I walk in the street full with the soldiers and the military officerses, hearing them use the familiar obscene language, laugh to each other, make hullabaloo, and has been arriving at the wharf.’

Relying solely on core elements, it would be difficult to tell the differences between instances of sense 4 ‘leave’ (as in examples (9), (10)) and the use of sense 1 ‘walk’ (as in example (11)), because they show the same highlighted elements and the same associated constructions. At first glance, in terms of Colloconsruction, we may find an anchor for sense 4 - the verb-final ‘了’, which has quite distinct distributions with clauses containing either ‘walk’ or ‘leave’, as we can see in the statistics in Table 2:

Table 2. Collocate frequencies for the ZOU LE(走了) construction as the meaning of ‘walk’ and ‘leave’[1].

Sense	‘leave’	‘walk’
Co-occur with verbal LE (了)	28/42 (57.14%)	7/53 (13.2%)
without verbal LE(了)	14/42 (42.86%)	46/53 (86.8%)

Although, as shown in Table 2, the possible anchor ‘了’ indeed has a higher frequency of occurrence in clauses compatible with the sense of ‘leave’, we still have to explain how people distinguish the two senses in the fewer cases where both senses have the same Colloconstruction - co-occurring with ‘了’ to form a colloconstruction - [walk + 了+ duration] or [leave +了+ duration] (such as examples (12)-(13)). Moreover, how can we deal with utterances with a bare ‘了’ and no other constructional anchors can be found (such as example (14))?

(**12**) sense **1**: walk
(CNI/self-mover) <*< duration

一路跟蹤而進，有時岔路上兩邊都有腳印，只得任意選一條路。走了好半天，山洞中岔路 不知凡幾

yi1lu4 gen1zong1 er2jin4, you3shi2 cha4lu4shang4 liang3bian1 dou1you3 jiao3yin4, zhi3dei3 ren4yi4 xuan3 yi1tiao2lu4. zou3le hao3ban4tian1, shan1dong4zhong1 cha4lu4 bu4zhi1 fan2ji3,
all road follow and enter, sometimes branch road above two sides have footprints, only can choose one road. walk LE good half-day, cave inside branch road not know how much
‘All the way we follow the footprints and walk into the caves, sometimes both sides of branch road all left the footprints and we just can choose one road arbitrarily. We walked a long time, we met uncountable branch roads on our way ,’

(**13**) sense 4: leave
(CNI/self-mover) <*< duration

我翻身坐了起來，怔了一怔，清醒了許多，問道：「走了多久？」「不知道，我下午開始陪他，後來看書看得睏了，就睡著了，起來就沒看到他了

wo3 fan1shen1 zuo4le qi3lai2, zheng1leyi1zheng1, qing1xing3le xu3duo1, wen4dao4: “zou3leduo1jiu3?” “bu4zhi1dao4, wo3 xia4wu3 kai1shi3 pei2ta1, hou4lai2 kan4shu1 kan4 de2 kun4le. jiu4 shui4zhao2 le, qi3lai2 jiu4 mei2kan4dao4ta1le
I turn body sit LE up, stun LE one stun, wide awoke LE many, ask: “walk LE how long?”
“don’t know, I afternoon start accompany him, then read books read DE feel asleep, then sleep LE, get up then not see him LE.
‘I turned over and sat to get up, and was being stunned a while, and when I wide awoke, I asked:" How long did he leave?" "I don't know, I started to accompany him in the afternoon, and then read the book and I felt asleep, and then imperceptibly I fall asleep; when I got up, I did not see him.’

(**14**) sense 4: leave

林昭良三下兩下就把麵吃完，就跟他們一窩蜂走了。
lin2zhao1liang2 san1xia4liang3xia4 jiu4 ba3 mian4 chi1wan2, jiu4 gen1 ta1men yi1wo1feng1 zou3le.
lin2zhao1liang2 immediately then BA noodles eat over, then with them an onrushing crowd people leave LE
'Zhao-Liang Lin finished the noodles immediately, and then left with them blindly.'

Comparing examples (12) and (13) above, sense 1 and sense 4 are almost identical in surface structure as they share the following:
Shared core frame elements: path, self_mover, duration
Shared syntactic pattern: (CNI/self-mover) <*< duration
Shared colloconstruction: [* + 了 + duration] (* represents the verb)

To distinguish the two senses, additional information from the larger context is needed. Each sense is believed to display certain features of Contextual Dependence. Here, Contextual Dependence refers to both foregrounded and backgrounded factors that are contextually linked with a given sense. In other words, we derive the sense 'leave' or the sense 'walk' from contextually bounded elements across clause boundaries. We make inferences on the basis of identifiable sense relations. For example, in example (12), the preceding sequence provides a valuable clue – the mention of *jiao yin* (**腳印**) 'footprint', which helps infer to the sense of 'walk'. A semantic link is established since the definition of 'walk' is 'the act of traveling by foot' (from WordNet)[2]. The clear mention of 'footprint' thus motivates a contextually appropriate reading of ZOU(**走**). In examples (13) and (14), the sense of 'leaving' is motivated by contextual sequences referring to 'appearing/disappearing', 'seeing/not seeing' or 'finishing/departure'. For example, *mei kan dao ta le* (**沒看到他了**) '(he) is no longer seen', and *mian chi wan* (**麵吃完**) 'finished eating the noodles', both are related to the concept of disappearance. The semantic distance or proximity of contextual variables can be obtained if an independently motivated hierarchy of semantic categories is available. In practice, a valuable resource would be databases such as SUMO. The contextually salient features can be readily identified if a close link in the SUMO hierarchy can be established. In our proposed resolution, discourse-level factors may be utilized with a clear measure of their semantic relations. According to Hopper and Thompson [6], 'users of a language are constantly required to design their utterances in accord with their own communicative goals and with their perception of their listeners' needs.' We also believe that communicative goals will often be realized with semantically coherent sequences.

7 Conclusion

In the previous sections, we present a preliminary model of the cognitive process for detecting word senses. Given the principle of economy and mechanisms in prototype theory, we assume that not all the senses of a word have equal weights and require exactly the same procedure for sense derivation. Therefore, three modules are called upon in a sequence when needed. The first module focuses on frame-based information regarding participating frame elements and their expressions. The second module identifies colloconstructions that go beyond the expression of core arguments and look for detailed lexical as well as grammatical association patterns. The last module deals with contextually dependent cues that are semantically or ontologically related to the target word.
In sum, the proposed resolution schema could be viewed as the cognitive procedure drawn upon when multiple senses are present in a single word form, as illustrated with the case of Mandarin verb ZOU(**走**). For further research, an automatic procedure may be established that makes use of frame-based semantic analysis and ontological hierarchy such as Sumo. A comprehensive investigation of Mandarin lexical semantics is under way (Liu [7][8]) and a bilingual ontological wordnet (Sinica BOW) is also available (Huang et al [9]). With useful tools, the cognitive procedure may offer a workable model to develop a computer system dealing with polysemy resolution. This model aims to integrate lexical semantics, corpus-based morphosyntax, and discourse analysis to provide a procedural and holistic solution. We also hope that this resolution can be applied in languages besides Mandarin.

Notes

1. The data used in Table 1 and Table 2 in this paper is from the Sinica Corpus (http://www.sinica.edu.tw/SinicaCorpus/). And the numbers are based on the randomly 200 utterances found in the corpus. The total occurrences of ZOU(**走**) in Sinica Corpus is over 2000. The examples cited in this paper are also from Sinica Corpus.
2. The definition is adopted from WordNet 2.0

References

[1] Bybee, Joan L. Morphology: A study of the relation between meaning and form. Amsterdam: John Benjamins, 1985.
[2] Cruse, D.A.. Lexical semantics. Cambridge University Press,1986
[3] Stefanowitsch, Anatol and Gries, Stefan th. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics 8:2* Amsterdam: John Benjamins, pp.209-243, 2003.
[4] Fillmore, Carles J. and Atkins, Beryl T. Frames, fields, and contrasts: *Toward a Frame-based Lexicon*, ed. by Lehrer, Adrienne. L. Erlbaum Associates, 1985.
[5] Firth, J.R. *Papers in Linguistics*. London: Oxford University Press, 1957.
[6] Hopper, Paul J. and Thompson, Sandra A. Transitivity in grammar and discourse. *Language, vol.56, Number 2*, pp.251-299, 1980
[7] Liu, Mei-chun. Mandarin Verbal Semantics: a corpus-based approach. Taipei: Crane Publishing Co. 2002.
[8] Liu, Mei-chun, and Chun Edison Chang. From Frame to Subframe: evidence from Mandarin verbs of conversation. Proceedings of the Fifth Chinese Lexical Semantics Workshop. Singapore. 2004.
[9] Huang, Chu-Ren, Ru-Yng Chang, Shiang-Bin Lee. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon. Portugal. 2004.

On-line Resources

FrameNet II: http://www.icsi.berkeley.edu/~framenet/
Sinica BOW: http://BOW.sinica.edu.tw/
SUMO: http://ontology.teknowledge.com/
WordNet: http://www.cogsci.princeton.edu

Too Good to Be True: A Case Study of "*Zui Hao Shi*"

Hsiang-nan (Gustav) Chou

Graduate Institute of Linguistics
National Taiwan Universuty
r92142009@ntu.edu.tw

Abstract. This paper aims to investigate the polysemy and multifunctionality of the expression "最好是". It has been observed in recent years that "最好是" has two different meanings and semantic-pragmatic functions. The first function is to express speaker's suggestion or expectation to the hearers to reach the optimal outcome. It is noted as the deontic optative meaning by Bybee (Bybee et al. 1994). As this meaning expresses expectation to the events in a hypothetical world, the deontic "最好是" also functions as a conditional marker (Traugott 1983). The other meaning of "最好是" is the epistemic meaning. The epistemic meaning of "最好是" performs the indirect speech act to show the speakers' denial or disbelief brought forth by the interlocutor. The paper is to argue that the epistemic meaning of "最好是" derives from the deontic meaning. This semantic change is motivated by subjectification of the semantic implication of deontic meaning, which consists of implicature of "not yet done" and "too good to be true". The data for this paper consists of three main sources: on-line corpus (Academia Sinica Balanced Corpus), Internet (google and yahoo), and conversation data. The three databases consists of different types of discourse (written and spoken) and different levels of formality. The observation from the data shows that the process of semantic change of "最好是" follows the path of semantic change proposed by Traugott and Dasher (2002). The epistemic meaning of "最好是" derives from the deontic meaning as the result of subjectification of the semantic implications (Traugott 1999). The distribution of data also points out that the epistemic "最好是" is informal and requires an interactional context while the deontic "最好是" appear much more frequently in formal context and written discourse. At last, according to the data, it is proposed that the epistemic "最好是" should be established as an epistemic formula. This formulaic form of "最好是" functions as verbal irony. It serves as an option for politeness strategy (Brown and Levinson 1987) to soften and counter direct criticism, complaints, and disbelief.[1]

1 Introduction

It has been observed that the expression "最好是" has two different meanings and semantic-pragmatic functions. One function is to show speaker's expectation or suggestion toward an event or an action. This function has the deontic meaning of expressing wish and desire. By using this function speakers show their intended hearers what is necessary to be done or to be possessed to achieve the optimal outcome. This function is illustrated below:

(1) a. **最好是**能早期診斷，以便得到最佳的治療效果。(Academia Sinica Balanced Corpus)
b. 他也建議自助旅行**最好是**有兩人以上同行，可互相照應。(Academia Sinica Balanced Corpus)

It is observed that both examples in (1) are suggestions or expectations for the hearers. In (1a) it is suggested or expected that for the best outcome of medical treatment it is best to diagnose the problem early. In (1b) it is suggested that to backpack one better have two or more companions to go with so they can look after one another during the trip. From the two examples it is readily shown that what is suggested or expected can be either an action or a property.

[1] I would like to thank Professor Shuanfan Huang and Professor I-wen Su for their opinions on the path of semantic change and the classification of "最好是". I am also grateful to Drew Tseng and Sharon Chen, two diligent students who helped me collect spoken data around the campus. The analysis is never complete without the data they collect. Last I'd like to thank Alvin Chen, Claire Wu and Weilun Lui for their comments on the observation and how the data should be approached.

The second function of “最好是” is as an epistemic formula. It functions to sarcastically deny or reject a proposition brought up by the interlocutors (2a) or existent in the situational context (2b). The use of this function is frequently observed among the young population (college students and high school students).

(2) a. A: 你這麼厲害，你應該也要去參加演講比賽的 B: **最好是**

b. (on seeing a dog sleeping with it belly up)**最好是**狗這樣睡覺

This paper is to argue that the epistemic formula “最好是” derives from its deontic meaning of suggestion and expectation toward hearer’s obligation. The epistemic meaning emerges from on-line communication not only due to the rise of subjectivity (Traugott 1999) but also due to the integration of the well-entrenched concept of “too good to be true”. Moreover, from a pragmatic-cognitive perspective this paper will attempt to examine that the multifunctionality of the expression “最好是” is based on mechanisms such as conditionality and politeness principle.

The data for this study consist of three different sources. The first source is the data collected from Academia Sinica Balanced Corpus. As the concordance is unable to recognize “最好是”, the search is done by keying in “最好”. Of all the 498 tokens 64 of them are “最好是”. The second source is the web search in google.com. The web search generated more than 6 million results. The first one hundred results are taken for this study. The third source is the personal notes of daily conversation. The notes contain 42 instances of “最好是”. The conversation data consists of mainly conversations between university students. The data collection and classification is done with the help of two other students in National Taiwan Normal University. In the later discussion, the data from different sources will be noted. This notation of data is first to recognize the source of data. Also the recognition of the source would also be able to indicate the interactional nature of the expression “最好是”.

The paper is organized as follows. Section 2 reviews the approaches and mechanisms which are crucial to the study of semantic change. Section 3 categorizes the expression “最好是” into two types: deontic “最好是” and epistemic “最好是”. Section 4 proposes that the semantic change of “最好是” involves not only subjectification but also semantic implication. Section 5 is the analysis of the distribution of two types of “最好是”. Section 6 draws a conclusion to the paper.

2. Overview of Semantic Change

In this section two topics will be covered. First is the unidirectionality of the semantic change of modality and subjectivity. The second is Hopper’s (1991) approach to semantic change.

2-1 The Unidirectionality of Semantic Change

Traugott and Dasher (2002) stated that semantic change has a lot to do with two aspects of language. The first is modality, the second is subjectivity.

Modality, according to Kiefer (1994) “consists in the relativization of the validity of sentence meanings of a possible world.” (p.2515) Modality can generally be separated into two categories: deontic modality and epistemic modality. Deontic modality concerns mainly about obligation or compulsion. Lyons (1977) identified the characteristics of deontic modality as concerned with the possibility and necessity of the actions performed by a morally responsible agent. It describes the state-of-affair that will obtain when the action in question is performed and it typically proceeds or derives from either outer or inner compulsions.

Another kind of modality presented in Traugott and Dasher (2002) is the advisability. Advisability is the sense that the action to be performed by the agent is not only normatively wished but also be profitable to the agent. They discovered that the advisability modality plays an important role in the development of modals in English and Japanese.

Bybee and her colleagues (1994) examined deontic modality in a more detailed fashion. They divided deontic modality into two parts: agent-oriented modality and speaker-oriented modality. Agent modality includes a. Oligation, b. Necessity, c. Ability, d. Desire. The speaker-oriented modality includes a. imperative b. prohibitive, c. optative, d. hortative, e. admonitive, and f. permissive modality. What should be noted is the optative modality. Optative modality expresses speaker’s wish and hope toward a hypothetical world. As will be seen in the later discussion it will be shown that the deontic meaning of “最好是” is mainly optative modality.

Epistemic modality, on the other hand, is largely concerned with speaker’s knowledge and belief.

According to Traugott and Dasher (2002), epistemic expressions are used to express speaker’s commitment to the truth of the proposition. For example, the sentence “John must be tired.” is a strong epistemic expression for it reflects the speaker’s belief of John being tire is firm. A weak epistemic reading

can be exemplified by the sentence "John may be tired." In this sentence the state of belief of John being tired is not so high in degree.

Traugott (1989), examined the process of semantic change of English words *allow* and *evidently*. She concluded with the general process of semantic change. It is shown in (i)

(i) main verb > premodal > deontic > weak epistemic > strong epistemic

This general process of semantic change is later confirmed in Traugott and Dasher (2002). The semantic change of modals *must* and *ought to* are examined. They concluded that the process of semantic change is that epistemic meaning derives from deontic meaning. The process is unidirectional. In other words, the process can not be reversed.

Another key issue in semantic change is subjectification. According to Lyons (1982), subjectivity refers to the way in which languages provide expressions of the attitudes and beliefs of a locutionary agent. Traugott (1989) identifies three tendencies of grammaticalization, in which meanings may change from propositional to textual or to expressive meanings. The third tendency she identifies is cited below:

Tendency Ⅲ: "Meanings tend to become increasingly based in the speakers subjective belief state/attitude toward the proposition." (1989:p.35)

In Traugott (1999), she defines the process of subjectification as how meaning is increasingly based on the subjective belief and attitude toward what is being said and what is being said. In other words, subjectivication is the process in which meanings tend to encode or externalize the speaker's perspectives and attitudes within the hypothesized world rather than the real world.

In Traugott and Dasher (2002), it is concluded that as the deontic meaning changes to epistemic meaning, the subjectivity will rise at the same time. In this way, the process of semantic change can be summarized as in (ii):

(ii) epistemic meanings derive from deontic meanings, meanwhile the subjectivity becomes higher

It should also be noted that unidirectionality is also true with subjectivity. So in the process subjectivity only gets higher, not vice versa.

2-2 Hopper's Approach to Semantic Change

The path of semantic change identified by Traugott and Dasher (2002) is well supported by diachronic data. What about semantic changes in languages or meanings that are present but not available in written documents? There are many languages do not have a written forms or documents of their languages. Also even for languages that have written forms the new meaning may only exist in spoken data. According to Hopper (1991), synchronic approach to semantic change or grammaticalization is possible based on the tendency from cross-linguistic observation. He also proposes several principles to deal with grammaticalzation. One of the principles is the "layering" principle, which states that the new meanings and the old meanings in the process of semantic change can co-exist. New meanings do not immediately replace old ones. In this way the cognitive-pragmatic reconstruction of synchronic spoken data is likely to draw some clues.

3. The Taxonomy of "最好是"

From the collected data, two main meanings of "最好是" are observed. The two meanings will be presented and discussed respectively in section 3-1 and 3-2.

3-1 Deontic Meaning:

The first meaning of "最好是" observed in the data is the deontic meaning. The deontic meaning is used to express one's will. The function of this meaning is for one to express one's wish, desire, and suggestion to a certain issue or proposition. Here the issues or propositions are the desired situations or conditions that the speakers seek for the profit of theirs or the addressees. The following are some examples:

(3) a. 這裡有沒有能在平常白天打球的朋友，**最好是**混雙 (google.com)
b. 徵集有關特洛的故事，**最好是**有歷史依據的！ (google.com)
c. **最好是**明天會放晴，這樣我們就可以去九份玩了 (personal notes)
d. **最好是**可以嫁一個有錢的老公，這樣以後就不用愁了 (personal notes)
e. 寫自傳**最好是**手寫，除較具親和力外，人事主管也偏向透過字跡對求職者態度 (google.com)
f. 參觀清真寺，穿著更要保守（**最好是**長袖與過膝長裙）(Academia Sinica Balance Corpus)

g. 烹調蔬菜時，應該迅速烹煮，**最好是**鍋蓋一次便煮好，最忌常常開鍋蓋 (Academia Sinica Balanced Corpus)

In example (3a), the speaker is looking for a teammate to play tennis. Also he expects the teammate to be of the opposite sex. This wish is revealed by the linguistic coding of using "最好是" before the desired condition. In (3b) the speaker is looking for stories about the Trojan War. Here he is not only asking for a story but also expecting the story to be one based on real history. Again his wish is specified by the use of "最好是". In these two examples, instead of integrating the desired condition into the main clause, the speakers choose to separate the desired condition from the main clause and add the expression "最好是". In this way the desired condition is highlighted. In (3c) and (3d), both speakers expect a desired optimal condition (a clear day, a wealthy husband) for their wish to come true. The use in (3c) and (3d) differ from (3a) and (3b). What is different is that in (3a) and (3b) "最好是" is used to directly code the desired entity or condition. However, in (3c) and (3d), "最好是" is used to code the premises for the desired condition to come true. The use of "最好是" in (3c) and (3d) are more like conditional markers. This function of conditional marker will be discussed latter in 3-1.1.

In examples (3e), (3f), and (3g) the expression "最好是" performs to give suggestion. One of the deontic meanings listed in Traugott and Dasher (2000) is advisability. Advisability includes the sense that the action sought of is not only normatively wished but also beneficial to the one to carry it out. Take (3f) for example. It tells the addressee it is not only necessary to wear conservative clothing when visiting a mosque, but it would be best or beneficial for visitors to wear long sleeves and long skirts to avoid troubles.

The difference between the function of expressing wish/desire and the function of giving suggestion is the difference of degree. The expression of wish and desire, according to Bybee et al (1994) is a subcategory of speaker-oriented deontic modality named optative modality. On the other hand, the function of giving suggestion is of the agent-oriented modality of obligation and necessaity, in which social or physical need would compel the agent (in the case the addressee) to perform the predicate actions. When expressing wish and desire, the hope (subjectivity) that the desired condition to true is usually very strong. This is because the desired condition is often beneficial to the speaker himself. On the other hand, when giving a suggestion the hope for the desired condition is not so strong in comparison to wishing. The difference in degree here is due to the fact that when giving suggestions the desired condition may not be directly profitable to the speakers themselves but to the addressees. As the speakers are not the ones benefited from the accomplishment of the desired situation, the motivation and the hope for it to be true will not be high.

There are also instances of deontic "最好是" is used for the function of threatening. As the following example:

(4) 妳**最好是**快點說，不然你就完蛋了 (personal notes)

This threatening function can be viewed as a peripheral type of advise/suggestion. It can be interpreted as that the speaker suggests the addressee to fulfill the premises (in the case, the talking) or something very bad will happen to him. The same as the instances of suggestion, the cases of threaten function to give advice for the benefit of the addressee.

3-1.1 Conditionality and Desirability of "最好是"

It has been mentioned in that the deontic meaning of "最好是", expressing wish and desire, behaves like a conditional marker. Here are some more examples:

(5) a. 我贊成學生可以選校長，但**最好是**推派學生代表參與遴選過程。 (Academia Sinica Balanced Corpus)

b. 想保持魔鬼身材嗎？**最好是**放輕鬆 (google.com)

Traugott (1983) stated that many lexical sources can become conditional markers. One of the sources is modality, especially optative modality that expresses wish and desire. Conditionals are about the hypothetical worlds. It is true that imagined hypothetical worlds are often ones that are wished for by the speakers. This is why optative modality can be motivated to be a conditional marker. As we see in example (5a) and (5b), both of the two instances put forward a desired condition (students being able to vote for university principal, keeping slim). These conditions are the imagined or hypothesized world wished for by the speakers. For the imagined or hypothesized world to come true some premises have to be done in the first place. The clauses with "最好是" provide the premises for the desired condition to be fulfilled. The conditional marker function of deontic "最好是" arises as it is to provide the premises for the hypothesized scenarios wished for.

Furthermore, it is noteworthy that when the deontic meaning of "最好是" is used, the expected/suggested premises is not yet fulfilled. The desired outcome is, therefore, far from being accomplished. It can be viewed as a kind of the "predictive" conditionals, which predicts that if a desired/undesired action is carried out or a desired/undesired condition is fulfilled, the desired/undesired consequence will take place. Clancy et al

(1997) observes that in American English, Japanese, and Korean, children less than three years old are given warnings or advices in the reasoning process cited below:

(iii) It is *desirable* that p will happen. If "not p" happens, it will lead to *undesirable* consequences.

Akatsuka and Strauss (2000) also states that speaker's stance of desirability is how people understands the various usages of conditional utterances in daily lives. It is through the reasoning process of described in (iii) that people understand the conditionality in utterances.

In this case of deontic modality of "最好是" desirability is crucial to the functions the expression performs. Take (5b) for example, the line of reasoning can be recorded in the format similar to (iii) and it will look like the one in (6)

(6) It is desirable that p (one being relaxed) will happen. If p happens, it will lead to the desirable consequence (keeping slim).

Example (4) can also be put in the same line of reasoning as (iii). The outcome is as (7):

(7) It is desirable that p (speaking quickly) will happen. If "not p" happens, it will lead to the undesirable consequence (being done for).

In this way the inclusion of undesirability is the difference between the pragmatic functions of advising and threatening. We can see that the use of "最好是" as a conditional marker often involves premises and outcomes. For example, in (5b) the premise is "to relax" and the outcome is "keeping slim". The premises are always what are required for the desirable outcome. If the outcome is the desirable one, then "最好是" functions to give advise and suggestion. If the outcome is the undesired one, "最好是" would function to be a threat.

3-2 Epistemic Meaning:

As have mentioned, epistemic meaning is largely concerned with the knowledge state or subjective belief. It is mainly about the speaker's evaluation or judgments on the truth of the proposition. The epistemic meaning of the expression "最好是" performs the indirect speech act of rejection or denial to the proposed proposition. It shows that speakers are not holding the evaluated propositions to be desired, rightful, or true.

The following are some examples:

(8) a. A:你今天頭髮捲捲的，好可愛，好像混血兒 B:**最好是**，是泰勞混印地安人吧！(personal notes)
b. A:喔你們在幽會喔 B: **最好是**在幽會 (personal notes)
c. A: 不預習也可以survive B: **最好是**不預習也行 (personal notes)
d. (responding to a previous article)**最好是**那ㄇ好啦 我柴不幸你這套勒 (google.com)
e. A:都沒有地方游泳 B:那你在家裡的浴缸游啊! A:**最好是** (personal notes)

Example (8a) shows that B does not take A's proposition of her being a person with mixed ethnicity to be true. At least she does not think that her curly hair is symbolic of a typical hybrid (European-Asian, for example). That is why after she sounded the denial with "最好是" she added another comment. That comment shows that she opposes the proposition brought up by A. Example (8b) and (8c) are similar ones. In both examples the denied proposition are repeated after the expression "最好是". It shows that in this kind of context it is the proposition brought about by the interlocutors (secret dating; survive the course without previewing the material) that are denied, not other elements of the previous statement. It is through this kind of instantiation that the negative reading of "最好是" can be inferred. (8d) shows that the proposition denied can be not only a single proposition but also a whole article. (8e) is another convincing instance that "最好是" is used to deny the truth or validity of a previously proposed idea. Most of the instances of epistemic "最好是" take the formulaic-like form in the observed daily conversations.

4. The Process of Semantic Change of "最好是"

In Traugott (1989, 1990) and Traugott & Dasher (2002) the unidirectionality of semantic change is proposed. Using examples like *allow* and *evidently*, Traugott (1989) discovered the direction of change of these words. Both these words go through the stages as illustrated in (iiii).

(iiii) deonitc meaning > object epistemic > strong epistemic

Note that not all the stages have to take place for the process to be complete. The general pattern of the change, as noted by Traugott and Dasher (2002), is that epistemic meaning derives from deontic meaning, not vice versa. Meanwhile subjectivity increase as the epistemic meaning derives from the deontic meaning.

In the case of "最好是". It starts out to have a deontic meaning. The deontic "最好是" functions as a conditional marker to give advises or suggestions. These suggestions and advises aim to guide the hearers to achieve the desired optimum. Also as conditionals the suggestions are given in the hope that the optimum

would come true in the hypothetical world. It is clear then at the time a speaker uses "最好是" the required premises (actions, properties) is not yet available and the desired outcome not yet accomplished. In other words, there is this implicature that the situation is irrealis. Akatsuka (1985), in the discussion of conditional and counterfactual reasoning, states that the conceptual domains of realis and irrealis have to do with one's epistemic evaluation. These two domains affect speaker's evaluation of the realizability of an event. In this way, subjectively one is capable of using this implicature to show that he knows that the event is not true. Hopper and Traugott (2000) point out that in early stages of grammaticalization the implicatures often become part of the semantic meaning of a form. In this case "最好是" the implicature of "not yet ture" or "not done" is clearly the sources of the epistemic use of "最好是" as a means to show disbelief. Meanwhile in the process as subjectivity of the speaker becomes higher the meaning will move toward the speaker's strong belief or disbelief of the event. In this case the semantic implicature is strengthen by the sujectification of meaning in the change from deontic "最好是" to epistemic "最好是".

The other source of semantic implication is the well-entrenched concept of "too good to be ture". In Traugott (1989), she suggested that the shift from deonitc meaning to epistemic meaning is done through the conventionalization of the conversational implicature. She stated that this conversational implicature is used in speaker's attempt to regulate communication with others. Levinson (2000) provided a more comprehensible definition of conversational implicature. For Levinson, the conversational implicature is a default inference "…that captures our intuitions about a preferred or normal interpretation." (p. 11). Then, what is the implicature that leads the epistemic meaning to a negative one? It is the cognitive factor that leads to the negative reading. As Langacker (1987) and Johnson (1987) pointed out, cognitive machenisms are often involved in the process of semantic change. They both proposed that the integration of familiar information to make sense of the new experience is a very basic process. Here the integrated concept is a well-entrenched one "too good to be true". The expression "最好是" often denotes an optimal condition which is desired by the speaker. However, everyone knows that the optimum is often hard to reach. For example, it is impossible to form an optimal rule without exception. Also it would be impossible for everything to go smoothly the way one expects. If anything can go wrong, it will. This implicature is best illustrated by (9):

(9) 當然婚姻在一起，我們最好是每天生活，能在一起快快樂樂的，但是這是不可能的！(Academia Sinica Balanced Corpus)

Therefore the optimum would often be related to those tasks that are impossible. It becomes predictable then, that when one proposes something that looks perfect, it is usually impossible. As the implicature becomes more deeply rooted in one's subjective belief, the conceptual connection between optimum and impossible is thus linked and integrated. In this way, when one proposes something that is optimal one is actually proposing something impossible. When the concept of "too good to be true" in integrated into the interlocutors, they would automatically connect the optimal meaning with disbelief, especially when the optimal proposition sounds untrue, undesired, or invalid to the interlocutor (which is the case with the epistemic "最好是"). Take (10) for example:

(10). A:你是不是整天都在做報告? B: **最好是** (personal notes)

In (10), the proposition brought up by A, to work on research papers all day long, would sound to B (and most others) to be very good, but impossible (or even exaggerating). Therefore B would see the optimal proposition of working on research papers all day long as untrue. With the concept "too good to be true" integrated to his mind, B would automatically treat the incoming material as not true and assign the negative meaning of disbelief/denial to the proposition to the expression of "最好是".

Overall, we can see that the motivation of "最好是" is mainly semantic implicature. It is implied that when one uses "最好是" the suggested qualification is not fulfilled and the desired outcome therefore not reached. The other semantic implication is that the outcome brought up by "最好是" is often too hard to reach in real life. When these two implicatures are "semanticized" to add new meaning of "最好是", the new meaning of disbelief or denial emerges in order to express speaker's subjective evaluation.

5. "最好是" as an Epistemic Formula

This part of the paper discusses the property of "最好是" as an epistemic formula. It will also be discussed why among so other possible collocations with "最好" it is in "最好是" that epistemic emerges. The third part of the analysis will draw reference to the politeness theory (Brown and Levinson 1987)

5-1. The Epistemic Formula "最好是"

The end point of semantic change or grammaticalization is often that a lexical item becomes grammaticalized and becomes a discourse marker. Discourse markers, according to Schffrin (1986), are "sequentially dependent elements which brackets units of talk" (p.31). They have lost their lexical meaning during the process of grammaticalization. Their functions are primarily discourse-oriented, such as turn-taking, topic-management, or discourse organizing.

This, however, is not the case with the epistemic "最好是". It is obvious that thought the meaning is altered the lexical meaning of epistemic "最好是" still exists. It is more appropriate to call it an epistemic formula. Both Bolinger (1976) and Fillmore (1967) noted that a large portion of language is memorized, automatic and rehearsed rather than created, generated, or freely put together. Coulmas (1979) termed these automatically produced parts of language as "routine formula". They are lexically and syntactically unchangeable groups of words. They are situationally-bound utterances to perform pragmatic functions such as greeting (e.g. good morning) or politeness (e.g. thank you).

Judging from these criteria, the epistemic "最好是" looks fit as an epistemic formula. First, the lexical meaning of denial or disbelief is fixed. Second it always occupies the clause-initial position. The situations in which they are used is when an optimal proposition is brought forth by the interlocutor that is untrue, invalid or undesired for the evaluation of the speaker.

Of all the 46 tokens of epistemic "最好是", 31 of them are used alone without the repetition of the denied proposition. Two theories provide convincing explanation for the formation of the formula. First, Givon (2001) states that reduced expressions are favored when the speaker is biased. The more the speaker is biased the more reduced the form will be. Here "最好是" serves as a good example. As the speakers are biased not to believe the possibility and probability of the proposition they would choose the minimal form. Second, Traugott (1995) and Traugott and Dasher (2002) proposes that In on-line communication (in which the instances of epistemic "最好是" are observed) the speakers invite their interlocutors to make inferences (invited inference) on their subjective evaluation of the current speech situation. Meanwhile hearers make the most effort to infer what is meant by the speakers. As long as the invited inference is semanticized it is predictable, the new meaning can be used for most informativeness with minimal linguistic coding. In the case of "最好是", once the negative reading is established from the interaction, the denied or rejected propositions no longer have to be repeated.

5-2 Why "最好是"?

From the previous analyses it is clear that the semantic change of "最好是" is the result of the semantic implication of "最好" and subjectification. However, there are many other possible collocations with "最好". The following are examples of the most frequent collocations with "最好" from the Academic Sinica Balanced Corpus:

Table 1. Collocations with "最好"

詞	Token
是	63
能	50
不要	47
的	29

In Table 1. are the four most frequent collocations with "最好" the tokens are the times of their appearances immediately following "最好" (最好是, 最好能, 最好不要 etc.). The ones that do not follow immediately "最好" are excluded for the purpose to see why only "最好是" undergoes semantic change. It is very likely that the reason lies in the different kinds of components following those words. From the data collected from the Academia Sinica Balanced Corpus, it is possible to look into the types of components these words introduce. First look at the three words "能", "不要", and "的".

The collocation of "最好能" always introduces a verb phrase (VP) as we can see in (11):

(11) a. 室內上課外，務必能進行戶外教學，勉強在校園進行之，但**最好能**真正在田野裡進行教學，最真實有效。 (Academia Sinica Balanced Corpus)

b. 而且天然鈾有用盡的一天，**最好能**有代替的核燃料。(Academia Sinica Balanced Corpus)

c. **最好能**立法通過一些保護條文，以確保「情色文學」的地位。(Academia Sinica Balanced Corpus)

The verb "能" is a copula verb denoting ability. From the data it is observed that all the instances of "最好能" are accompanies with a verb. The meaning of "最好能" is then the expectation that some action is to be

taken for the desired optimum. The meaning of the verb "能" then restricts the kind of proposition that follow it to only those related to actions i.e. verb phrases.

"最好不要" shows a similar patter with "最好能". 46 of the 47 instances takes the construction of "最好不要+VP". For instance:

(12) a. 因此不欲人知的事**最好不要**存在電腦檔案中，或在網路上傳送。(Academia Sinica Balanced Corpus)

b. 有一些菜是喜宴不能用的，如：鰱魚（結婚一次就好，**最好不要**連續。）(Academia Sinica Balanced Corpus)

There is also an instance of "最好不要" in the clause-final position:

(13) 有位業主向建築師詢問能否採用開放空間設計，建築師告訴他**最好不要**(Academia Sinica Balanced Corpus)

When placed in the clause-final position like the one in (13), the VP that is omitted following "最好不要" can be found in the preceding clause. The components that follow "最好不要" are always VPs in the Academia Sinica Balanced Corpus.

Now look at the third collocation "最好的". All the instances in the corpus of "最好的" are followed by noun phrases (NPs). As shown in (14):

(14) a. 爲了選取最好的角度拍攝下牠**最好的**神態，胡教授在樹叢中等了好長時間。(Academia Sinica Balanced Corpus)

b. 有一成多的民眾則分別認爲專職人員、教師是**最好的**交通導護人選。(Academia Sinica Balanced Corpus)

c. 例如表現**最好的**一%學生，就有選擇進入前一%的學校(Academia Sinica Balanced Corpus)

It appears that the three collocations "最好能", "最好不要", "最好的" are biased in the components that they introduces. "最好能" and "最好不要" always introduce VPs. "最好的" introduces NPs.

On the other hand, "最好是" can introduce a wider variety of components. For instance, it can introduce a full clause like (15) cited below:

(15) 如果你有機會選擇什麼時候現身的話，考慮一下時機；**最好是**父母最近沒有什麼重大事件需要憂慮(Academia Sinica Balanced Corpus)

It is also able to introduce VPs like (16a) and (16b).

(16) a. 我們**最好是**這個禮拜以內決定，我好給旅館打電話定房間。(Academia Sinica Balanced Corpus)

b. 便當的價格應該不是同一價格，**最好是**分爲好幾種價格，讓學生選擇(Academia Sinica Balanced Corpus)

"最好是" can introduce the desired property as in (17a) and (17b).

(17) a. 一要說本國的故事。二**最好是**寓言式的。(Academia Sinica Balanced Corpus)

b. 適用於野外活動的圖鑑**最好是**攜帶式的(Academia Sinica Balanced Corpus)

Like "最好的", "最好是" can introduce NPs.

(18) a. **最好是**美國頂尖學府學位(Academia Sinica Balanced Corpus)

b. 「將來行政院長**最好是**通才，多年來老是財經內閣，總要找個人不是財經的。」(Academia Sinica Balanced Corpus)

c. 砧板**最好是**松木製的「立砧」（樹幹橫切取材）(Academia Sinica Balanced Corpus)

Here it is obvious that "最好是", among other frequent collocations, can introduce more variety of components. It is in this sense semantically and syntactically more general than other collocations. As speakers use "最好是" it includes the functions of other collocations. It is why among the many collocations "最好是" it chosen to undergo semantic change.

5-3 "最好是" as Verbal Irony.

It has been discussed in 3-2 that the epistemic "最好是" functions to perform the indirect speech act. The form "最好是" would look like the speaker see the proposition brought up by the interlocutor or in the situational context to be desirable. The actual meaning is that the proposition is to the speakers as untrue or undesirable. This function can be viewed as an ironical function. According to Sperper and Wilson (1995), verbal irony is "invariably the rejecting and the disapproving kind (p.237)". The speakers of uses verbal irony to disassociate themselves from the proposition echoed and indicate that they do nit hold the proposition to be true. Sperper and Wilson put forth that there are three requirements in understanding verbal irony. First is to recognize the speech as echoeic. Second is to identify the source of echoed opinion. Third is to recognize the

speaker attitude as rejection and disapproving. We can use these three criteria to examine "最好是" as verbal irony. Let's look at examples (8b) and (8c):

8b. A: 喔你們在幽會喔 B: **最好是**在幽會 (personal notes)

8c. A: 不預習也可以survive B: **最好是**不預習也行 (personal notes)

In both examples the rejected proposition is echoed. In (8b) the assumption of the secret dating that is echoed. In (8c) it is the opinion of being able to survive the course without previewing the material that is echoed. The propositions in both examples are from the other interlocutors. The propositions from the other interlocutors in both are rejected by the sentence containing "最好是".

Besides verbal irony, the epistemic meaning of "最好是" has another pragmatic function. It is recognized by Brown and Levinson (1987) that indirect speech and verbal irony are both strategies of politeness. Politeness is way to soften or to counter the effect face-threatening acts (FTAs). Indirectness can save face by allowing speakers to avoid responsibility for the potentially face-damaging interpretation of the utterance. By using indirect speech the speech act is not directed to the the hearer as the speaker do not really commit to the utterances. Also by using irony to express criticism, disapproval, and complaint can be thought of as a softening a threat to the positive face of the hearer. It is also noted that the use of indirectness and irony is often among intimates or close friends. "最好是" in this sense, is also able to soften negative feelings. It is not a direct criticism or rejection such as "不對", "不好", or "不行". It is also observed that of the tokens which "最好是" is used as verbal irony 38 of them are used between classmates and friends and the other one is used between mother and child. These observations shows that "最好是" as verbal irony to perform indirect speech act of denial, rejection or disbelief is a politeness strategy used among close friends or intimate individuals.

6. The Distribution of "最好是"

The distribution of the deontic and epistemic "最好是" in different data collections is shown in Table 2.

Table 2. Distribution of "最好是"

	Academia Sinica	google.com	Personal Notes
Deontic (%)	63(100%)	93(93%)	3(7.1%)
Epistemic (%)	0	7(7%)	39(92.9%)

Of all the data, the data from Academia Sinica Balanced Corpus consists of mainly written discourse, which is the most formal set of data. The data from google.com contains a variety of sources and is regarded as with mixed formality. The personal notes are all face-to-face interaction data among peers (only one of the token is between mother and son). It is deemed the least formal set of data. From the distribution data it can be concluded that the epistemic "最好是" takes place mainly in interaction situations. On the other hand, the deontic meaning of "最好是" enjoys wider distribution in all three sets of data. The distribution is in itself capable of showing the nature of the different situations. In daily, face-to-face discourse, the exchange of ideas is often very rapid. As there are exchanges of ideas there would inevitably be confrontations. This rapid pace of discourse and potential of ideational confrontation would promote the use of the short epistemic formula "最好是". On the other hand, as there is often not need of seeable change of ideas in written course (ideas often go unidirectionally from the writer to the reader), no confrontation would take place and therefore not necessay to use the epistemic formula "最好是". As to the deontic "最好是", the function of expressing wish, desire or suggestions are universal no matter what discourse type it would be. Therefore the deontic "最好是" enjoys a wider distribution.

7. Summary and Conclusion

The process of semantic change of the expression "最好是" confirms the process proposed by Traugott and Dasher (2002). From the data collected from spoken discourse and written corpus, it is observed that different layerings of meaning co-exist in contemporary Mandarin Chinese. The epistemic meaning of "最好是", the one expressing speaker's denial and disbelief toward a proposition, derives from it's deontic meaning, which is mainly the optative modality of expressing wish and desire. This process is motivated by the semantic implication of "not yet ture" and "too good to be true". The process is completed by subjectification that makes the usage move toward the speaker's subjective evaluation of the proposition. It is through these two processes that the negative meaning (denial and disbelief) rather than the positive (strong belief) of the epistemic "最好是" would come about. The distributional data show that the epistemic "最好是" is strongly

interaction-oriented. It mainly takes place in conversations when exchange of propositions and confrontations are available.

As “最好是” is to denote a desired condition in the hypothesized world, it can be used as a conditional marker in its deontic sense. This deontic meaning and conditionality brings the assumption of a hypothetical world. It is this conditionality that allows speakers to grasp the implicatures that would motivate the semantic change.

The epistemic “最好是” can be used as an epistemic formula. It can be used under the situation in which a proposition, which is viewed by the speaker as not true or invalid, is proposed to deny and show speaker’s disbelief. The usage of “最好是” in isolation as an epistemic formula is the invited inference. It is through the invited inference that hearers can understand speaker’s intention of expressing subjective evaluation. As long as the inference is semanticized and predictable, the meaning of epistemic “最好是” as denial or rejection is then stable and isolated use is understood by other hearers. As a formula it also is a politeness strategy owing to its nature of indirect speech act and verbal irony. As an indirect speech act it allows the speaker to not directly commit to the utterances that aim to criticize or to reject. As verbal irony it softens the strong negative feeling of direct rejection, criticism, and disbelief.

To sum up, the epistemic meaning of “最好是” emerges from the deontic use. The conditionality expressed in the deontic function contains implicature that the desired outcome is not yet achieved and is too good to be true. As subjectivity rises the epistemic function of expressing speaker’s denial and disbelief takes place. The use of “最好是” to express subjective evaluation is then stabilized through invited inference. The epistemic meaning is also used as a formula in situations which requires the expression of disbelief and denial. The present study confirms the unidirectionality of semantic change and investigates the semantic-pragmatic properties of such shift in meaning.

References:

Academia Sinica Balanced Corpus. Academia Sinica, Taipei (http://www.sinica.edu.tw/SinicaCorpus/)

Akatsuka, Noriko 1985. Conditionals and the epistemic scale. Language 61(3): 625-639

Akatsuka, Noriko M. and Susan Strauss 2000. Counterfactual reasoning and desirability. In Couper-Kuhlen, Elizabeth and Bernd Kortmann, eds., Cause Condition Concession Contrast. Mouton de Gruyter, Berlin: 205-234

Bolinger, Dwight 1976. Meaning and memory. Forum Linguisticum 1: 1-14.

Brown, Penelope and Stephen C. Levinson 1987. Politeness: Some Universals in Language Use. Cambridge: Cambridge University Press

Bybee, Joan L., Revere Perkins, and William Pagliuca. 1994. The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World. Chicago: University of Chicago Press.

Coulmas, Florian 1979. On the sociolinguistic relevance of routine formulae. Journal of Pragmatics 3: 239-266

Fillmore, Charles 1976. The need for a frame semantics with linguists. SMIL, Skriptor, Stockholm: 5-29

Givon, Talmy 2001. Givon, T. (2001) Syntax. Amsterdam: Benjamins.

Hopper, Paul J. 1991. On some principles of grammaticalization. In Traugott, Elizabeth C. and Bernd Heine eds., Approaches to Grammaticalization. Amsterdam: John Benjamins.

Hopper, Paul J. and Elizabeth C. Traugott 1983. Grammaticalization. Cambridge: Cambridge Universitu Press.

Johnson, Mark 1987. The Body in the Mind: The Bodily Basis of Meaning, Reason and Imagination. Chicago: University of Chicago Press.

Kiefer, Ferenc 1994. Modality. In Asher and Simpson, vol. V: 2515-2520

Langacker, Ronald W. 1987. Foundations of Cognitive Grammar, volume 1: Theoretical Prerequisites. Stanford, CA: Stanford University Press.

Levinson, Stephen C. 2000 Presumptive Meanings: The Theory of Generalized Conversational Implicature. MIT

Lyons, John 1977. Semantics. Cambridge: Cambridge University Press.

Lyons, John 1982. Deixis and subjectivity: Loquor, ergo sum? In Robert J. Jarvella and Wolfgang Klein, eds., Speech, Place, and Action: Studies in Deixis and Related Topics. New York: Wiley.

Schiffrin, Deborah 1986. Discourse Markers. Cambridge: Cambridge University Press.

Sperer, Dan and Deirdre Wilson 1995. Relevance: Communication and Cognition. Balckwell Publishe

FUNCTIONAL DISTINCTION BETWEEN *ZAI* (在) AND *ZHENGZAI* (正在) IN MANDARIN
Evidence from Collocations

Lin, Tsi-chun Liu, Mei-chun

National Chiao-Tung University

tsichung@yahoo.com.tw mliu@mail.nctu.edu.tw

Abstract

Both zai and zhengzai are progressive markers in Mandarin Chinese, and by the principle of economy, there should be some differences in these two progressive markers. With the Sinica Corpus on-line tools, a significant difference is found in the collocation of adverbial adjuncts with the use of zai and zhengzai. This paper discusses three types of adverbials to distinguish these two markers: modality adverbs, time adverbs, and manner adverbs. Zhengzai cannot co-occur with [+iterative] adverbs and adverbs without a specific time reference. It mainly indicates the progression of an on-going event at a given specific time point. On the other hand, zai not only indicates the on-going process but can also signal the progression of repeated event as habitual- progressive.

1 Introduction

In Mandarin Chinese, both *zai* and *zhengzai* are progressive markers which present an internal interval of a durative situation, and often have the connotations of activity and temporary imperfectivity associated with non-statives [3]. But, when talking about imperfective aspect markers in Mandarin Chinese, most studies just represent *zai* as a progressive marker ([2], [5], [6], [7], [11], [17], and [18]), while fewer studies indicate that there are other representations of progressive markers ([3] , [4], and [15]). Although mentioning both *zai* and *zhengzai* as progressive markers, previous studies treat them almost as interchangeable and only focus on the discussion of *zai* with few detailed distinction between these two forms. Given that these two markers are morphologically distinct, some questions will have to be asked: do they encode the same grammatical, semantic and pragmatic information? Do they occur in completely the same contexts or they have different contextual constraints? If the two markers are functionally identical, we may wonder why Chinese has two different and redundant progressive representations. Given the principle of economy in language, our assumption is that there must be some fine-grained distinctions between *zai* and *zhengzai*. In this paper, we adopt a corpus-based approach, since corpus data provide a wealth of grammatical associations that may help delimit the key semantic distinctions, as successfully shown in recent studies on lexical semantics (cf. [9], [13], and [14]). This paper thus aims to explore the semantic distinction between *zai* and *zhengzai* by examining their collocational patterns in discourse.

2 Methodology

Most of our data come from the Sinica Corpus. Since *zai* has a variety of senses, we only look at the instances of *zai* that are compatible with *zhengzai* as aspectual markers. As a result, there are 2000 utterances with *zai* and 696 utterances with *zhengzai*.

With the searching tools and POS tags in Sinica Corpus, we can readily retrieve the frequency counts of neighboring categories which precede or follow *zai* and *zhengzai* for further analysis.

3 Finding and Discussion

The data in Sinica Corpus reveal that the top ten grammatical categories before and after *zai* and *zhengzai* do vary, as shown in Tables la, lb and 2a, 2b.

Table 1a. Top-10 Categories before *zai*

詞類	左 5	左 4	左 3	左 2	左 1	合計	比例
Na	303	304	285	338	413	1643	17.31
D	**179**	**161**	**177**	**260**	**748**	**1525**	**16.06**
，	220	227	254	213	98	1012	10.66
Nh	83	97	106	177	246	709	7.47
VH	94	92	82	71	38	377	3.97
Nf	61	67	68	98	18	312	3.29
SHI	25	31	38	63	134	291	3.07
VE	75	54	58	59	5	251	2.64
Nc	45	67	47	51	28	238	2.51
。	50	68	55	38	25	236	2.49

Table 1b. Top-10 Categories after *zai*

詞類	右 1	右 2	右 3	右 4	右 5	合計	比例
Na	20	333	270	299	304	1226	12.86
VC	762	91	78	68	94	1093	11.46
，	3	389	295	147	153	987	10.35
D	71	35	121	238	239	704	7.38
VA	353	31	37	34	43	498	5.22
Nh	4	153	103	110	102	472	4.95
VE	282	30	33	47	61	453	4.75
VH	142	57	77	85	90	451	4.73
。	1	121	137	68	47	374	3.92
Nep	3	116	36	44	27	226	2.37

Table 2a. Top-10 Category before *zhengzai*

詞類	左 5	左 4	左 3	左 2	左 1	合計	比例
Na	107	102	113	124	204	650	19.85
，	54	62	77	117	70	380	11.61
D	**51**	**43**	**34**	**26**	**59**	**213**	**6.51**
Nc	25	36	46	42	63	212	6.48
Nh	20	19	23	13	88	163	4.98

Nd	23	16	26	25	52	142	4.34
VH	35	40	36	24	2	137	4.18
。	26	32	36	23	18	135	4.12
VC	38	39	26	15	14	132	4.03
Nf	12	25	18	30	13	98	2.99

Table 2b. Top-10 Categories after *zhengzai*

詞類	右 1	右 2	右 3	右 4	右 5	合計	比例
Na	14	105	132	169	138	558	17.37
VC	290	56	59	31	30	466	14.51
，	0	65	73	66	71	275	8.56
VH	81	13	26	30	31	181	5.64
Nc	37	37	25	29	28	156	4.86
VA	57	24	28	17	14	140	4.36
D	24	11	13	45	45	138	4.30
。	0	19	45	33	27	124	3.86
P	47	9	6	12	18	92	2.86
VE	34	10	9	12	23	88	2.74

(Na = common noun; Nc = place word; Nd = temporal word; Nf = measure; Nep = demonstrative determinative; Nh = pronoun; VA = active intransitive verb; VC = active transitive verb; VE = active verb with a sentential object; VH = stative intransitive verb; D = adverb; P = preposition; SHI = *是*)

The result shows that although there is no significant distinction in the categories after *zai* and *zhengzai* from figure 1b and 2b, there are indeed differences in the categories before the two markers as shown in Tables 1a and 2a. The three categories commonly found after *zai* and *zhengzai* are the same and their percentages are close—no more than five percent. On the other hand, from Tables 1a and 2a, we found an interesting difference in the preceding positions. The three higher-frequency categories preceding the two markers are categorically the same but with different rankings, among which adverbs show a significant distinction. Adverbs preceding *zai* is 10 percent more than that of *zhengzai*. It may suggest that adverbs could be an important indicator to distinct *zai* and *zhengzai*. In the following, we mainly focus on three types of adverbs to characterize their relations with *zai* and *zhengzai*: modality adverbs, time adverbs, and manner adverbs.

Both *zai* and *zhengzai* can occur with modality adverbs such as *yiding* (一定), *keneng* (可能), yexu (也許), *haoxiang* (好像), and so on.

(1) 他 **可能** 在 躲避 什麼
ta keneng zai duobi sheme
he possibly ZAI avoiding something
"He is possibly avoiding something."

(2) 她 想了 一下, **似乎** 在 找 台詞
ta xiangle yixia sihu zai zhao taici
she think a while seem ZAI search what to say
"She thought for a while, seemingly to be searching what to say."

(3) 他的 手 **可能** 正在 摸 黑黑 的 機油

tade shou keneng zhengzai mo heihei de jiyou
his hands may ZHENGZAI group dark and black engine oil
"His hands may be groping in the dark the black engine oil."

(4) 技術性 反彈 **似乎** 正在 醞釀
jishuxing fantan sihu zhengzai yunniang
technical rebound seem ZHENGZAI ferment
"The technical rebound seems to be fermenting."

However, when the modality adverb contains the feature [+ iterative], indicating a repetition or recurrence of an event, such as *you* (又) and *zai* (再) [6], the marker *zhengzai* is not allowed to co-occur.

(5) 真的嗎? 你 **又** 在/*正在騙人
*zhendema ni you zai/*zhengzai qianren*
really you again ZAI/*ZENGZAI decive people
"Really? You are deceiving people again."

The above example indicates that the event with *zhengzai* can not be viewed as a repetition or a recurrence pertaining to a previous reference event. Thus, *zhengzai* is [-iterative], constrained by semantic and contextual factors, but *zai* is free to be used with [+iterative] events.

According to Givon [16], time adverbs can be classified into three sub-groups: temporal adverbs, frequency adverbs, and aspectuality adverbs. Both *zai* and *zhengzai* can co-occur with temporal adverbs, such as *xianzai* (現在), *muqian* (目前), *zuijin* (最近) and the like.

(6) 你們 知道 那個 男孩 **現在** 心中 在 想 什麼 嗎?
nimen zhidao nage nanhai xianzai xinzhong zai xiang sheme ma
you know that boy now in mind ZAI think what
"You know now what that boy is thinking in mind."

(7) 明知, 她 **這時** 在 做 什麼
mingzhi ta zheshi zai zuo sheme
know perfectly well she at this time ZAI do what
"Knew perfectly well what she is doing at this time"

(8) **現在** 正在 施工 中
xianzai zhengzai shigong zhong
now ZHENGZAI construction under
"Now it is under construction"

(9) 因為 **那時** 孩子 正在 傷心
yinwei nashi haizi zhengzai shangxin
because at that time child ZHENGZAI sad
"Because at that time the child is being sad."

But, there are some limitations of *zhengzai*. It can only occur with temporal adverbs referring to a specific time point or a short period of time with a clear reference point as *youyitian* (有一天), *xianzai* (現在), *zuijin* (最近) and so on. It cannot occur with a durational time adverb without a specified reference point, such as *meitian* (每天), *shiwunianlai* (十五年來), *yibeizi* (一輩子), etc.

(10) 大人們 會 比 **現在 每天** 在 /*正在 用 的人 還要 懂 嗎
*darenmen hui bi xianzai meitian zai/*zhengzai yong deren haiyao dong ma*
adults will than now everyday ZAI/*ZENGZAI use people more understand
"Will the adults understand more than those who use everyday?"

(11) 他 不是 個 騙子, 就是 個 沒有 感覺 **一輩子** 都 在 /*正在 說謊 的 白痴。
*Ta bushi ge pianzi jiushi ge meiyou ganjue yibeizi dou zai/*zhengzai shuohuang de baichi*
he either a liar or a without feeling lifetime all ZAI/*ZENGZAI tell a lie DE idiot
"He is either a liar or an idiot who telling a lie for a lifetime without feeling."

From this, we know that *zhengzai* indicates what is on-going at a specific time reference, locating the event in the time axis and contributing contrastive and attitudinal features to the sentence [3].

With frequency adverbs such as *changchang* (常常), *zhongshi* (總是) and aspectuality adverbs as *luxu* (陸續), *buduan* (不斷), *zhengzai* are NOT allowed to appear, either, since these frequency adverbs signal a progressive aspect in the habitual sense, i.e. without a specific time reference [16] or indicate repetitive-progressive that extends over an unspecified period of time.

(12) frequency adverbs
基金會 也 **常常** 在/*正在 廣告 大腸癌，
*jijinhui ye changchang zai/*zhengzai guanggao dachangai*
foundation also often ZAI/*ZENGZAI advertise Colon Cancer
"Foundation is also often advertising Colon Cancer."

(13) aspectuality adverbs
目前 全國 各 運動 單項 協會 **陸續** 在/*正在 召開
會員 大會 進行 理監事 及 理事長 改選。
*Muqian quanguo ge yundong danxiang xiehui luxu zai/*zhengzai zhaokai*
currently national each exercise single-item association continuously ZAI/*ZENGZAI hold
huiyuan dahui jinxing lijianshi ji lishizhang gaixuan
general meeting carry on supervisor and director re-election
"Currently each national association of single-item exercise is continuously holding the general meeting to carry on the re-election of the supervisor and the director."

The shorter form *zai* can co-occur with the above frequency adverbs, since it is compatible with the feature [+iterative] in that the progressive event can be repetitive. It not only represents simple progressive as an on-going event but can also signal habitual-progressive with the use of durational adverbs.

With regard to manner adverbs, both *zai* and *zhengzai* may occur with a variety of manner adverbs. But, there is a significant constraint in terms of the position of the manner adverb. *Zhengzai* cannot take a preceding manner adverb as shown in (14), while *zai* can occur both with preceding and following manner adverbs as (14) and (15).

(14) 畫家 **悠閒地** 在/*正在 寫生
*huajia youxiandi zai/*zhengzai xiesheng*
painter leisurely and carefree ZAI/*ZENGZAI draw from nature
"The painter is leisurely and carefree drawing from nature"

(15) a. 整天 都 在 **不停的** 叫
zhengtian dou zai butingde jiao
all day ZAI continuously cry
"(she) is crying continuously all day."

b. 聽見 小雞 正在 **唧唧的** 吵鬧
tingjian xiauji zhengzai jijide chaunau
hear chicken ZHENGZAI peep make noise
"Hears the chicken is peeping and making noise."

Manner adverbs typically characterize the way or means the event is carried out. Since *zhengzai* signals the overlapping of an on-going event with a specific time point, which, when substantiated, is supposed take up the slot immediately preceding *zhengzai*. Thus, a manner adverb cannot take the pre-aspectual position that may be occupied by a time reference. It then ended up only in the post-aspectual position immediately preceding the verb, a slot that will not block the expression of reference time. On the other hand, *zai* is free from a specified time reference and may take a pre- or post-aspectual manner adverb. But the scope of modification differs with different positions of manner adjuncts. When a manner adjunct occurs after *zai* and immediately before the verb, it is event-internal, modifying the single instance of the predicated event. However, when a manner adjunct occurs before *zai*, it is event-external, modifying the relation of the predicated event with some other constituent.

In sum, *zhengzai* requires a time reference, indicating the on-going process pertaining to a specific time point. It is a semantically and pragmatically stronger form to represent progressive event [8]. It cannot be used to express repetitive-progressive (*He is repetitively hitting the ball*), continuous-progressive (*He continued hitting the ball*) or habitual progressive (He is always hitting the ball). But *zai* can occur with [+iterative] and [-iterative] events without a specified time reference. The distinction between them is that *zhengzai* only indicates **deictic progressive** (tensed aspect), while

zai is compatible with other types of progressives.

4 Conclusion

This paper discusses the distinction between *zai* and *zhengzai* with evidence from their collocational patterns. It is found that the use of adverbial adjuncts with *zai* and *zhengzai* represents a significance difference. *Zhengzai* is more constrained in semantic and pragmatic specifications. It cannot co-occur with [+iterative] adverbials indicating repetition of an event or adverbials without a specific time reference. *Zhengzai* indicates an on-going progressive event at a specific time, signaling temporally deictic aspectuality. On the other hand, *zai* is less restricted in marking all kinds of progressive perspective.

In Chinese, there are other markers which can also indicate the progressive or imperfective aspect, such as *zheng* (正) or *zhe* (著). Thus, in further studies, we can compare *zai*, *zhengzai* and other progressive markers to come to a complete picture of the imperfective marking system. In addition, this paper does not exhaust all types of adverbial collocations. There are other types of adverbs which do not collocate with *zhengzai* , such as negative adverbs. It can be reserved for a follow-up study in the future. Moreover, discourse-level constraints on the use of the two markers would be another interesting area to look further into.

References

[1] B. Corrie. *Aspect*. Cambridge: Cambridge University Press, 1976.

[2] Carlota S. Smith. Aspectual Viewpoint and Situation Type in Mandarin Chinese. *Journal of East Asian Linguistics* 3, pp. 107-46, 1994.

[3] Carlota S. Smith. *The Parameter of Aspect.* Dordrecht: Klumer, 1991.

[4] Chao-mao Huang. Riyu dongci ' *te i-ru* xing' de yufa gongneng: yu Hanyu *zhengzai* yu *zhe* de duibi (日語動詞「*te i-ru* 形」的語法功能--與漢語「正在」「著」的對比). *Jingwen jishu xueyuan xuebao* 11, pp.125-136, 1990.

[5] Charles N. Li and Sandra A. Thompson. *Mandarin Chinese: A functional reference grammar, ch. 6: Aspect.* Berkeley: university of California Press, 1981.

[6] Chouncey C. Chu. *A Discourse Grammar of Mandarin Chinese.* New York: Peter Lang Publishing, 1998.

[7] C-Y. Chen. Aspectual features of the verb and the relative positions of the locatives. *Journal of Chinese Linguistics* 6, pp. 76-103, 1978.

[8] L. Zhang. *A contrastive study of aspectuality in German, English, and Chinese*. Peter Lang Inc.: New York. [Berkeley insights in linguistics and semiotics 19], 1995.

[9] Li-li Chang, Keh-jiann Chen and Chu-ren Huang. Alternation Across Semantic Field: A Study of Mandarin Verbs of Emotion. *International Journal of Computational Linguistics and Chinese Language Processing* 5 (1), pp. 61-80, 2000b.

[10] Lillian Meei-jin Huang. *Aspect: A general system and its manifestation in Mandarin Chinese*. PhD dissertation, Rice University, 1987.

[11] Lillian Meei-jin Huang and W. Davis. Philip. An Aspectual System in Mandarin Chinese. *Journal of Chinese Linguistics* 17, pp. 128-66, 1989.

[12] Mari Broman Olsen. *A Semantic and Pragmatic Model of Lexical and Grammatical Aspect*. New York & London: Garland Publishing, 1997.

[13] Mei-chun Liu. From Collocation to Event Information: The Case of Mandarin Verbs of Discussion. *Language and Linguistics* 4.3, pp. 563-585, 2003.

[14] Mei-chun Liu. *Mandarin Verbal Semantics: A Corpus-based Approach 2nd ed.* Crane Publishing, 2002.

[15] Søren Egerod. Aspect in Chinese. Carl Bache, Hans Basbøll, Carl-Erik Lindberg (eds.), *Tense, Aspect and Action: Empirical and Theoretical Contributions to Language Typology (Proceedings of seminars on Verbal Semantics at Odense University in 1986 and 1987.)*. Berlin: Mouton de Gruyter, pp.279-310, 1994.

[16] T. Givon. *English Grammar: A function-based introduction.* Amsterdam and Philadelphia: John Benjamins, 1993.

[17] Wolfgang Klein, Ping Li and Henriette Hendriks. Aspect and Assertion in Mandarin Chinese. *Natural Language & Linguistic Theory* 18, pp. 723-770, 2000.

[18] Yasuhiro Shirai. Where the progressive and the resultative meet: Imperfective aspect in Japanese, Chinese, Korean and English. *Studies in Language* 22, pp. 661-92, 1998.

[19] Yi-fen Luo. *Imperfective Aspect Marker "Zai","Zhe" In Mandarin Chinese:A New Look At An Old Problem* (論漢語未完成貌動貌詞"在"與"著"). MA dissertation, NTHU, 1995.

Web Resource

Sinica Corpus: http://www.sinica.edu.tw/ftms~bin/kiwi1/mkiwish

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

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。為保障訂閱者的權利，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。手機帶給人們極為便利的通訊環境，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，便成為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 萬訂閱戶。朝日新聞以低廉的費用爭取手機用戶訂閱新聞，用意是為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，作為報社擴大市場、吸引訂戶的手段。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 個全型字為限，半形字則以 158 字（letter）為限（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，大抵可分為「摘錄」（extraction）與「摘要」（abstraction）兩種。「摘錄」的結果為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」為主的研究佔較多數。

自動摘要的成效評估，可分為直接（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、日文的文件為主。以 DUC 2001 年為例，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」為主的自動摘要方法。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，文件來源為 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為摘要的基準方法好，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%句子作為固定比例摘要，並選取 10%-50%的最佳句作為最佳比例摘要。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。其方法為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，經人工評估為滿意與尚可的比例，約為 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]，其摘要的長度分別為 117 字與 85 字。若取最接近標題的句子作為摘要，則其長度分別為 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%，悲觀重疊率的平均為 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戶，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」，若起始為「而且」、「且」等詞，則將此「片段」往前合併；若起始詞彙為動詞，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戶端大樓之遠端超高速乙太網路交換器（GESWr）、超高速數位用戶迴路設備（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】黃子佼為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】中共為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。

以文件二的第二句為例，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（因為「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（因為「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『黃子佼為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。由於預定的長度為 69 字或 45 字，幾乎只能容量一、兩個小句。為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為其中一個小句，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，黃子佼為訪問王貞治前往福岡巨蛋欣賞日本職棒比賽』，效果還要好（因為長度更接近 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，因為長度更接近 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，幾乎是目前為止還在研究的問題。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，作為「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 為關鍵詞 w 在該文件中的詞頻，max_tf 為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），做為該文件的關鍵詞彙。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，因此稱為最大重複字串。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為關鍵詞。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，將剩下的詞彙都視為關鍵詞。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 個句子作為後續處理之用。在此 n 可視為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，為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。為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，長度各為 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）的編輯距離為：

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，而初始值為：

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)的意義為

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 為例，且假設代換、插入與刪除的計算成本都為 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，則其距離矩陣為：

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，作為可能的接句點。

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，沒有真正反映語意的近似，而且相同或極相近的近似點可能有數個，在以長度的適合度為優先考量的情況下，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。為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 為自然指數（natural exponent），edit 為編輯距離，m 為標題的字數。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，而編輯距離為 13、11 與 9 時，相似度分別為 0.0015、0.0639 與 0.2231。為縮短其差距，我們將 **m** 以 **m+n** 取代，其中 n 為「候選句」的長度，變成：

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

修改後的相似度，其最大值為 1，最小值為 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，為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 都設為 3），則非前兩句的句子，其相似度都乘以 0.85，作為最後的相似度。

在步驟三中，要根據字數與相似度排序摘要候選句。同樣的為便於比較，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，可以決定其是否為最佳候選句，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。每一篇的第一列為其標題，第二列與第三列分別為，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：文件全文為表一中的文件三。

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為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，為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，調整幅度為二．六五％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，為威森整理投手丘。	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為威森整理投手丘。	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為 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，並將作為百慕達銀行旗下全球範圍的 GFS 部門與其日本客戶之間的聯繫機構。	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，對高檔大戶型商品房、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，國軍要為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為 9 月 12 日。	1	
	39	競國實業董事會決議配息配股基準日為 9 月 12 日，9 月 8 日起至 9 月 12 日停止股票過戶。	1	G
	39	競國實業董事會決議配息配股基準日為 9 月 12 日，9 月 8 日起至 9 月 12 日停止股票過戶。	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，初期以代理為主(2-2)。	3	
	41	《未上市個股》亞太電信推出「猿人在線」品牌，初期以代理為主(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，初期以代理為主(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，不僅為業界首創，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，投資標的為逆浮動+正浮動利率債券，不僅為業界首創，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，漲幅為 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，漲幅為 0.68%，換算回台股每股價格約 80.54 元。	1	B
33	Title	「美夢成真」趕戲，葉全真累壞吊了點滴再上	3	
	45	「美夢成真」趕戲，葉全真累壞吊了點滴再上，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	「美夢成真」趕戲，葉全真累壞吊了點滴再上，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，使得圍棋儼然成為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，使得圍棋儼然成為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	藝文界前輩進駐為豐樂童畫賽暖身	1	
	15	藝文界前輩進駐為豐樂童畫賽暖身	1	B
	15	藝文界前輩進駐為豐樂童畫賽暖身	1	B

為了便於分析，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>
<!--
    原句為：
    中華日報記者程紹菖／台北報導針對媒體報導政府提撥的
    斷開後為：
-->
    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為不設限，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，另外三百億元則為首
    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，為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，作為 tag 值。
  4. 地名、機構名的簡稱，列於 tag 的屬性裡，作為 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，則用原文的名稱作為 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為 tag 值，而其他異名詞為 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）可為主動者（類似主詞）或被動者（類似受詞）。
  2. 如果有次要對象（類似配角），因為也參與其中，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、事件為原則。
```

```
  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. 地名為此主題的發生地，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為主。
  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為不設限，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" 為此摘要的限制，其值可能為 69 或 45。
     目前 45 暫不需要。
  2. 屬性 char 為摘要後的字數，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，可能全數改為不設限，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 表示為人工摘錄，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 為摘要後的字數，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。
     以本例子為例，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」為「。」、「？」或「！」斷開的敘述，其中被「，」、
     「；」或「：」斷開者稱為「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為不設限，
    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，不做任何修辭，並盡可能符合字數限制。
  請在註解裡，說明摘要原則。
    以本例子為例，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為不設限，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，原本投資信心動搖的原始股東，已決定暫緩調節，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○．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，並且開始轉虧為盈。 世界先進雖然 64MDRAM 尚未量產，但因為 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，原本投資信心動搖的原始股東，已決定暫緩調節，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。

Using the Web as Corpus for Un-supervised Learning in Question Answering

Yi-Chia Wang[1], Jian-Cheng Wu[2], Tyne Liang[1] and Jason S. Chang[2]

1. Dep. of Computer and Information Science, National Chiao Tung University, Taiwan, R.O.C.
2. Dep. of Computer Science, National Tsing Hua University, Taiwan, R.O.C.

rhyme.cis92g@nctu.edu.tw

Abstract. In this paper we propose a method for unsupervised learning of relation between terms in questions and answer passages by using the Web as corpus. The method involves automatic acquisition of relevant answer passages from the Web for a set of questions and answers, as well as alignment of wh-phrases and keywords in questions with phrases in the answer passages. At run time, wh-phrases and keywords are transformed to a sequence of expanded query terms in order to bias the underlying search engine to give higher rank to relevant passages. Evaluation on a set of questions shows that our prototype improves the performance of a question answering system by increasing the precision rate of top ranking passages returned by the search engine.

1. Introduction

It was noted that people have submitted longer and longer queries to the Web search engines. Recently, users have started to submit natural language queries instead of a list of keywords. It has encouraged many researchers to develop question answering systems which specifically aim at natural language questions, such as AskJeeves (www.ask.com) and START (www.ai.mit.edu/projects/infolab/).

For typical question answering systems, document/passage retrieval is the most significant subtask. In this step, the QA system breaks a natural language question into a set of keywords, uses keywords to query a search engine, and returns documents or messages that are related to the queries for further processing. However, the keywords in questions usually are not very effective in retrieving relevant passages. Consider the question "*Who invented glasses with two foci?*" Typically, we will send the keywords "*invented glasses two foci*" to a search engine to retrieve documents or passages. Submitting such keywords to AltaVista, we got irrelevant information about astronomy or physics rather than the inventor "*Benjamin Franklin*" of bifocal glasses. Intuitively, if we include the phrase "*inventor of*" or "*bifocal*" in the query sent to the search engine (*SE*), we are likely to retrieve passages with the answer.

We present the system *Atlas* (Automatic Transform Learning by Aligning Sentences of question and answer), which automatically learns the transforms from wh-phrases and keywords to n-grams in relevant passages by using the Web as corpus. The transformed query should be more likely to retrieve passages that contain the answer. For instance, consider the natural language question "*Who invented the light bulb?*" Using the keywords in the question directly, we end up with the keyword query, "*invented light bulb*," for a search engine such as Google. We observed that such a query has room for improvement in terms of bringing in more instances of the relevant answer. Our experiment indicates that the proposed method will determine the best transforms for the wh-phrase "*who invented*" including "*inventor of*", "*was invented*", and "*invented by*". On the other hand, the best transforms discovered for the keyword "*bulb*" include "*light bulb*" and "*electric light*." Intuitively, these transforms used together will convert the question into an expanded query for Google, "*("was invented" || "invented by") ("electric light" || "light bulb")*" which is more effective in retrieving relevant sentences in the top ranking summaries returned by the search engine, such as "*The light bulb was invented by an illuminated scientist called Thomas Edison in 1879!*". One indicator of effective query is the precision rate at *R* document retrieved (P_R), the percentage of first *R* top ranking Web pages (or summaries) which contain the answer. Another indicator is the mean reciprocal rank (MRR) of the first relevant document (or summary) returned. If the *r*-th document (summary) returned is the first one with the answer then the reciprocal rank is 1/*r*. Our goal in this study is exploration of methods that will automatically learn the transforms that convert natural language questions to queries with high average P_R or MRR.

The rest of the paper is organized as follows. In Section 2, we survey the related work. In Section 3, we describe our method for unsupervised learning of transforms for question and answer pairs which are automatically acquired from the Web and how we use the aligned result for effective query expansion in the QA

system. The experiment and evaluation results are given in Section 4. In the last section, we conclude with discussion and future work.

2. Related Work

Extensive work on question answering has been reported in the many literature (Buchholz et al., 2001; Harabagiu et al., 2001; John et al., 2002; Shen et al., 2003). In this study, we focus on learning the transforms that can be used to convert questions into effective queries in order to retrieve relevant passages.

Hovy et al. (2000) utilized hypernyms and synonyms in WordNet to expand queries for increasing recall. However, blindly expanding a word to its synonyms sometimes causes undesirable effects. As for hypernyms, it is difficult to determine how many hypernyms a word should be expanded. In contrast to this approach, our method learns query transforms specific to a word or phrase based on real-life questions and answer passages.

In a recent study most closely related to our method, Agichtein et al. (2004) described the *Tritus* system that learns transforms of wh-phrases such as "*what is*" to "*refers to*" by using FAQ data automatically. Our method learns transforms for wh-phrases as well as keywords from the web. Tritus system uses heuristic rules and thresholds for term and document frequency to learn transforms, while we rely on a mathematical model method for statistical machine translation. Shen, Lin and Chen (2003) proposed a method that is similar to the *Tritus* system for the why question.

Recently, Echihabi and Marcu (2003) presented a noisy channel approach to question answering. Their method also involves collecting answer passages from the web and aligning words across a question and relevant answer passages. However, they require full parsing of the sentences and complicated decision of making a "cut" in the parse tree to determine whether to align word, syntactic, or semantic categories. Our simple method is also based on alignment but it does not require full parsing and perform alignment at the surface levels of words and n-grams.

In contrast to previous work on query expansion for question answering, we propose a method that learns query transforms for all phrases in a natural language question automatically on the Web.

3. Method for Learning Question to Query Transforms

In this section, we present an unsupervised method for QA which automatically learns transforms from wh-phrases and keywords to answer n-grams by using the Web as corpus.

3.1 Problem Statement

Given a set of natural language questions *Q*s and answer terms *A*s, we obtain a collection of passages that contain the answer *A* to the question *Q* via some search engine *SE*. From the collection of answer passages *AP*s, our goal is to discover a set of transforms *T* that can be applied to wh-phrases and keywords in *Q* in the hope that the transformed queries are more effective in retrieving passages containing *A*.

3.2 Procedure for Learning Transforms

This subsection illustrates the procedure for learning transforms *T* from wh-phrases and unigrams in *Q* into bigrams in *AP*. The reason why we decide to use bigrams in *AP* is that bigram contains more information than unigram and is more effective in retrieving relevant passages. On the other hand, we break *Q*s into unigrams following the standard approach in IR.

(1)	Automatically collect pairs of *Q* and *AP* from the Web for training. (Section 3.2.1)
(2)	Select frequent wh-phrases. (Section 3.2.2)
(3)	Apply the alignment technique to the collected material. (Section 3.2.3)

Fig.1. Procedure for learning transforms

3.2.1 Collecting Training Material from the Web

In the first step of the learning process (see Figure 1), we retrieve a set of (*Q*, *A*, *AP*) pairs from the Web for training purpose where *Q* stands for a natural language question, and *AP* is a passage containing keywords in *Q* and the answer term *A*. The data gathering process is described as follows:

1. For each (*Q, A*) pair in the given collection, we extract keywords *K* of *Q*, say, $k_1, k_2, \ldots, k_n$.
2. Submit ($k_1, k_2, \ldots, k_n, A$), as a query to *SE*.
3. Download the top *M* summaries that are returned by *SE*.
4. Retain only those summaries containing *A*. See Table 1 for details.

Table 1. An example of converting a question (*Q*) with its answer (*A*) to *SE* query and retrieving answer passages (*AP*)

(*Q*, *A*)	***AP***
What is the capital of Pakistan? Answer:(*Islamabad*)	Bungalow For Rent in *Islamabad*, Capital Pakistan. Beautiful Big House For …
($k_1, k_2, \ldots, k_n, A$)	*Islamabad* is the capital of Pakistan. Current time, …
capital, Pakistan, Islamabad	…the airport which serves Pakistan's capital *Islamabad*, …

3.2.2 Selecting Frequent Wh-phrases

In the second step, we produce a set of high frequency phrases that characterize different question categories. We follow the method proposed by Agichtein et al. (2004). The method simply involves computing the frequency of all n-grams in *Q*s and filters out those with small counts. We will treat the wh-phrases (*QP*s) as a token in the subsequent steps. However, we differ from their approach in that we are not limited to n-grams of function words. For instance, we derived "*in what year*", "*who wrote*", etc. More examples of wh-phrases are listed in Table 2.

Table 2. An example of wh-phrases that are used

Wh-words	**Wh-phrases *QP*s**
What	"what is the", "in what year", "what was", …
Who	"who was the", "who wrote", …
Which	"which country", "with which", …
⋮	⋮

3.2.3 Learning Question to Query Transforms

In the third step, we use word alignment techniques originally developed for statistical machine translation to find out relation between wh-phrases or keywords in *Q* and n-grams in *AP*. We use the Competitive Linking Algorithm proposed by Melamed (1997) to align (*Q*, *AP*) pair. We proceed as follows:

1. Perform Part of Speech (POS) tagging on both *Q* and *AP* in the collection. (See Table 3 and 4)
2. Replace all instances of *A* with the tag <ANS> in *AP*. For example, the answer "*Islamabad*" in *AP* for the question "*What is the capital of Pakistan?*" is replaced with <ANS>. (See Table 4.) The purpose of <ANS> is to avoid data sparseness while counting bigrams in the following step.
3. Segment *Q* into unigrams or *QP*s and eliminate unigrams with low counts. We denote the remaining unigrams as $q_1, q_2, ..., q_n$. (See Table 5)
4. Segment *AP* into bigrams and eliminate bigrams with small term frequency (*tf*) or very large document frequency(*df*). We denote the remaining bigrams $a_1, a_2, ..., a_m$. (See Table 6)
5. For all *i*, *j*, calculate log likelihood ratio (LLR) of q_i and a_j. (See Table 7)
6. Eliminate candidates with a LLR value lower than 7.88. (See Table 8)
7. Sort list of (q_i, a_j) by decreasing LLR value. (See Table 8)
8. Go down the list and select a pair if it does not conflict with previous selection.
9. Stop when running out of pairs in the list.
10. Produce the list of aligned pairs for all *Q*s and *AP*s.
11. Select top *N* bigrams, $a_1, a_2, ..., a_r$, for every wh-phrase or unigram q_i in alignment pairs. (See Table 9)

Table 3. Part of Speech of *Q*

***Q* word**	**Lemma**	**Position**	**POS**
What is the	what be the	1	*
capital	capital	2	nn
of	of	3	in
Pakistan	Pakistan	4	np
?	?	5	.

Table 4. Part of Speech of *AP*

***AP* word**	**Lemma**	**Position**	**POS**
Most	most	1	rbt
of	of	2	in
Pakistan	Pakistan	3	np
rainfall	rainfall	4	nn
is	be	5	bez
scarce	scarce	6	jj
.	.	7	.
Islamabad	<ANS>	8	np
,	,	9	,
the	the	10	at
capital	capital	11	nn
of	of	12	in
Pakistan	Pakistan	13	np
since	since	14	in
1963	1963	15	cd
,	,	16	,
and	and	17	cc
Rawalpindi	Rawalpindi	18	np
,	,	19	,
are	be	20	ber
both	both	21	abx
located	locate	22	vbn
on	on	23	rp
the	the	24	at
Pothowar	Pothowar	25	np
Plain	Plain	26	nn

Table 5. Wh-phrases and unigrams in *Q*

Question words	Number of occurrence
what is the	14,571
capital	7,513
of	29,673
Pakistan	135

Table 6. Bigrams in *AP*

N-gram in AP	Number of occurrence
most of	368
of Pakistan	54
Pakistan rainfall	1
rainfall be	4
be scarce	2
scarce .	1
. <ANS>	8574
<ANS> ,	16665
, the	10227
the capital	1690
capital of	1669
of Pakistan	54
Pakistan since	1
since 1963	3
1963 ,	58
, and	9994
and Rawalpindi	3
Rawalpindi ,	5
, be	3718
be both	77
both locate	2
locate on	174
on the	4868
the Pothowar	2
Pothowar Plain	2

Note: The entries in the shaded area are eliminated for their low counts

Table 7. Combination of q_i and a_j

q_i	a_j	Number of co-occurrence	LLR
what is the	most of	82	754.72
capital	most of	34	293.2
Of	most of	127	1118.59
Pakistan	most of	1	1.27
what is the	of Pakistan	47	614.3
Of	of Pakistan	49	580.78
capital	of Pakistan	43	602.61
Pakistan	of Pakistan	44	990.37
…	…	…	…

Table 8. Alignment results

q_i	a_j	LLR
of	, and	27227.9
capital	capital of	21194.2
what is the	, is	7443.56
Pakistan	of Pakistan	990.37

Table 9. Examples of transforms selected from alignment results for N=3

Wh-phrase and Keyword in *Q*	Bigram in *AP*	Alignment counts
what is the	is the	1,254
what is the	in the	503
what is the	, the	242
capital	capital of	545
capital	capital city	241
capital	state capital	236
Pakistan	Pakistan ,	39
Pakistan	of Pakistan	21
Pakistan	in Pakistan	6

3.3 Runtime Transformation of Questions

At run time, *Q* is broken into wh-phrases and keywords which are converted to a sequence of query terms according to transforms based on the alignment results described in Section 3.2 in order to give higher ranks to passages that contain the answer for specific *SE*. See Table 10 for example of the conversion process of the question "*Who invented light bulb?*"

Table 10. An example of transformation from question into query

Question		
Who invented light bulb?		
Wh-phrase	**Keywords**	
Who invented	light	bulb
Transform wh-phrase and keywords		
was invented invented by	electric light light bulb	electric light light bulb
Expanded query		
Boolean query: ((was invented)OR(invented by))AND((electric light)OR(light bulb))		
Equivalent Google query: ("was invented" || "invented by") ("electric light" || "light bulb")		

4. Experiments and Evaluation

4.1 Training Data Set

Our data training data set were collected from http://www.quiz-zone.co.uk.We use 3,581 distinct (*Q*, *A*) pairs for automatically retrieving *AP* from the search engine Google. For each *Q*, top 100 summaries returned by Google are downloaded. See Table 11 for details of the training corpus.

Table 11. The training corpus

Training data set	Distinct (*Q*, *A*)	Distinct (*Q*, *AP*)
Quiz-Zone	3,581	99,697

4.2 Alignment Results

We choose the top 2 (N=2) bigrams for each *QP* or keyword in alignment results. Table 12 lists examples of *QP* or keyword and its two corresponding transformed bigrams.

Table 12. Parts of alignment results

QP* or Keyword in *Q	**Bigram in *AP***	**Alignment count**
invent	be invent	175
invent	invent by	43
who wrote	be bear	94
who wrote	he write	87
capital	capital of	545
capital	capital city	241

4.3 Evaluation Results

We used a test set of ten questions which are set aside from the training corpus. Table 13 shows the keyword queries and the expanded queries based on the transforms learned from the Web. We evaluated the expanded query by the mean reciprocal rank (MRR) and the precision rate at ten summaries returned by Google. For comparison, we also evaluated Google without applying query transforms. During experiment, the ten batches of returned summaries for the ten questions were evaluated by two human judges. As we can see in Table 14, using keywords from the natural language questions directly to query Google resulted in an MMR value of 0.48. However, when using expanded queries provided by the Atlas system, we had an MMR of 0.70, a statistically significant improvement. The average precision rate was improved slightly from 40% to 47%. The experimental results show that the Atlas system used in conjunction with the search engine Google outperforms the underlying search engine itself.

Table 13. Test questions

Q	**Keyword query for Google (GO)**	**Expanded query for Google (AT+GO)**
What is the capital of Pakistan?	capital Pakistan	("capital +of" || "capital city") Pakistan
What became the 50th state of the America?	became 50th state America	("+to become" || "leader +of") "50th state" "United State"
Who had a hit in 1994 with "Zombie"?	hit 1994 "Zombie"	("number one" || "hit +in") 1994 "Zombie"
In which year did Coronation Street begin?	year Coronation Street begin	("was found" ||"was born") Coronation Street ("began +in" || "began on")
In "The Simpsons", what is the name of Ned Flanders wife?	"The Simpsons" name Ned Flanders wife	"The Simpsons" ("name +is" || "name +of") Ned Flanders wife
In mythology, who supported the heavens on his shoulders?	Mythology supported heavens shoulders	"+in Greek" "+of +his" supported heavens shoulders
Which Saint's day is on March 1st?	Saint day March 1st	"+is +a" Saint "St" March 1st
What is the largest city in Switzerland?	largest city Switzerland	("largest country" || "second largest") Switzerland
Who directed the Oscar-winning film "The English Patient"?	directed Oscar-winning film "The English Patient"	("directed +by" || "+and directed") Oscar-winning film "The English Patient"
Which country was once ruled by Tsars?	country once ruled Tsars	("country +is" || "country +in") "ruled +by" Tsars

Table 14. Evaluation results

Performances	MRR	Precision (%)
AT+GO (*Atlas* expanded query for Google)	0.70	47
GO (Direct keyword query for Google)	0.48	40

5. Conclusions and Future Work

We show that our method clearly provide means for learning transformation from a natural language question to a query by applying statistical word alignment technique. The method involves automatically acquiring relevant passages from the Web for a set of questions and answers, aligning phases across from questions to answer passages in order to create phrase transforms that involve wh-words as well as content words. Evaluation on a set of questions shows that our prototype in conjunction with a search engine outperforms the underlying search engine used alone.

Many future directions present themselves. For example, the patterns learned from answer passages acquired on the Web can be extended to include longer and more effective n-grams to further booster the MMR value or average precision rate. Additionally, an interesting direction to explore is creating phrase transforms that contain the answer extraction patterns. These answer extraction patterns can be learned for different types of answers. Yet another direction of research would be to provide confidence factors for ranking the likelihood of many candidate answers extracted using patterns.

In summary, we have introduced a method for learning query transforms that improves the ability to retrieve passages with answers using the Web as corpus. The method involves finding query transformations based on techniques borrowed from training a noisy channel in machine translation study. We have implemented and thoroughly evaluated the method as applied to a set of more than 4,000 questions. We have shown that the method can be used with a search engine as an effective component in a question answering system.

References

[1] Abdessamad Echihabi, Daniel Marcu. A Noisy-Channel Approach to Question Answering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp.16-23, July 2003.

[2] Buchholz, Sabine. Using grammatical relations, answer frequencies and the World Wide Web for question answering. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*.

[3] Eugene Agichtein, Steve Lawrence, Luis Gravano. Learning to find answers to questions on the Web. In *ACM Transactions on Internet Technology (TOIT)*, Volume 4, Issue 2, pp.129-162, 2004.

[4] Harabagiu, S., D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Buneascu, R. Gîrju, V. Rus and P. Morarescu. FALCON: Boosting Knowledge for Answer Engines. In *Proceedings of the 9th Text Retrieval Conference (TREC-9)*, pp.479–488.

[5] Hovy, E., Gerber, L., Hermjakob, U., Junk, M., and Lin, CY. Question answering in Webclopedia. In *Proceedings of the TREC-9 Question Answering Track*, pp.655-672, 2000.

[6] John M. Prager, Jennifer Chu-Carroll, Krysztof Czuba. Use of WordNet Hypernyms for Answering What-Is Question. In *Proceedings of the TREC-2002 Conference (TREC 2002)*.

[7] Melamed, I. Dan. A Word-to-Word Model of Translational Equivalence. In *Proceedings of the 35st Annual Meeting of the Association for Computational Linguistics*, pp.490-497, 1997.

[8] **沈天佐, 林川傑, 陳信希. 以網際網路內容為基礎之問答系統 "Why" 問句研究**. In *Proceedings of Rocling 2003*, pp.211-229, 2003.

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。

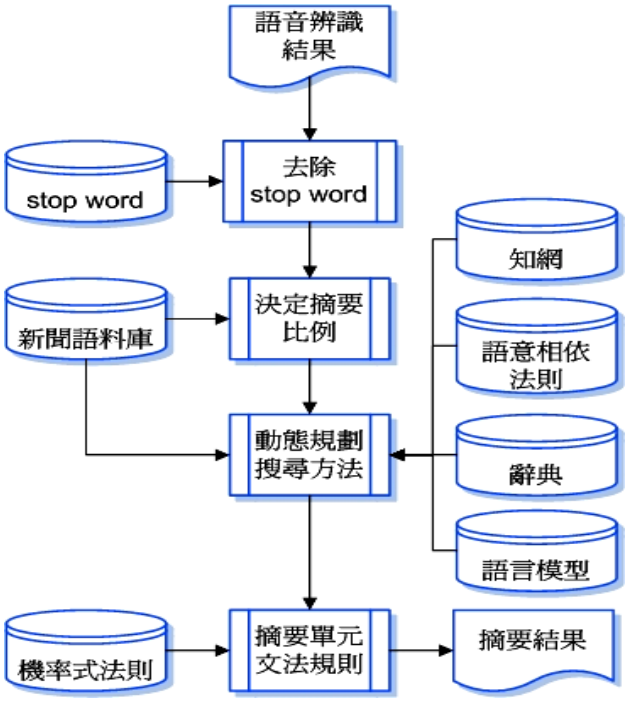


圖 1. 自動語音文件摘要程序

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：

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

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：

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

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$R(q,d)$，

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

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$n = \min(t,d)$。

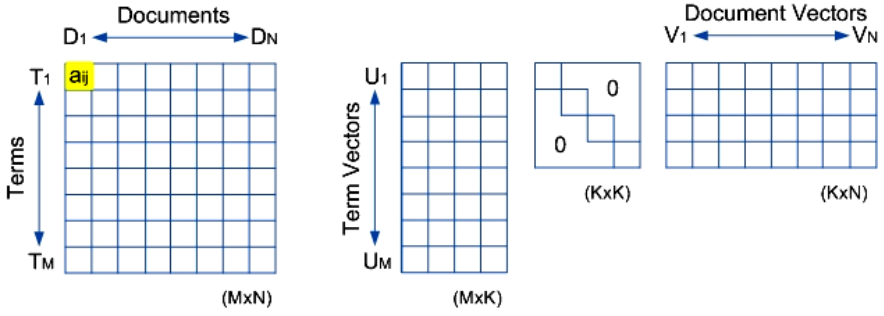


圖 2. 奇異值分解

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)”。

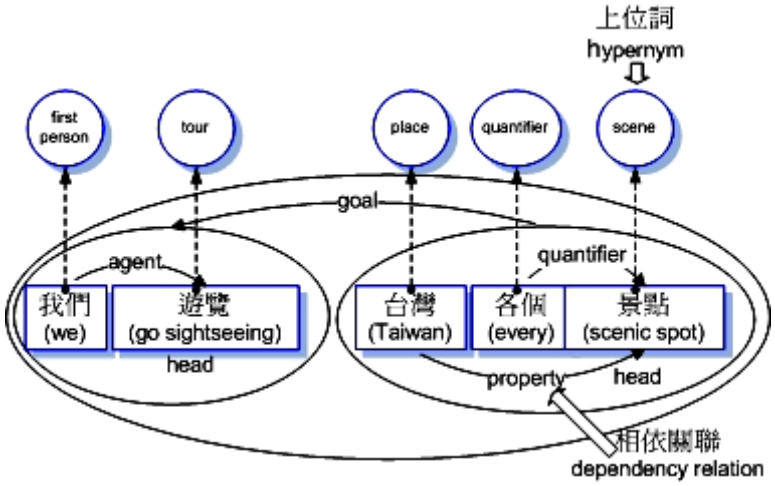


圖 3. 語意相依關聯範例

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：

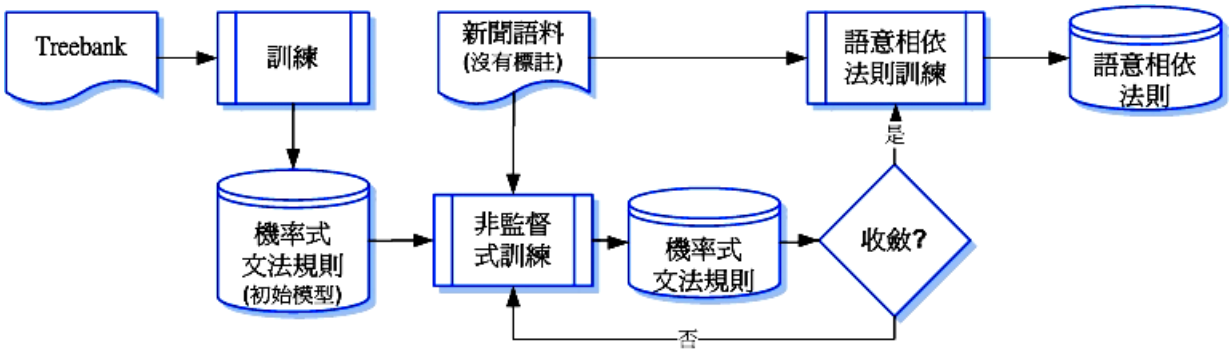


圖 4. 語意相依法則及機率式文法規則訓練流程

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。

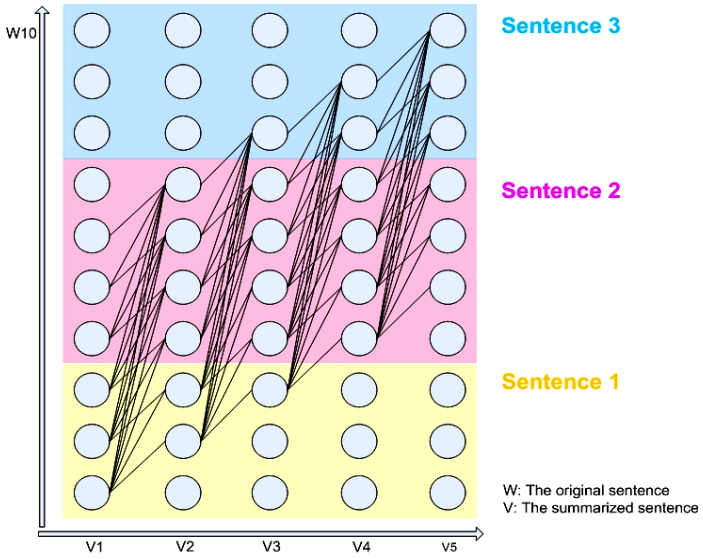


圖 5. 語音摘要使用動態規劃搜尋方法之示意圖

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)

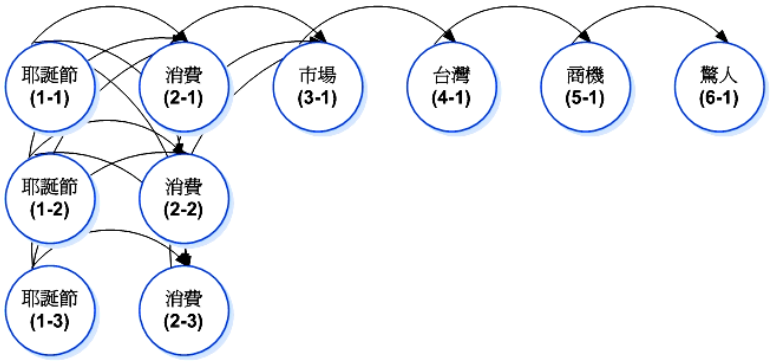


圖 6. 摘要語音串接示意圖

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：

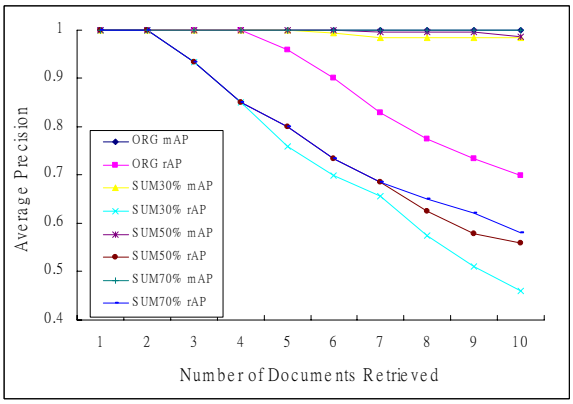


圖 7. 重要資訊檢索的結果

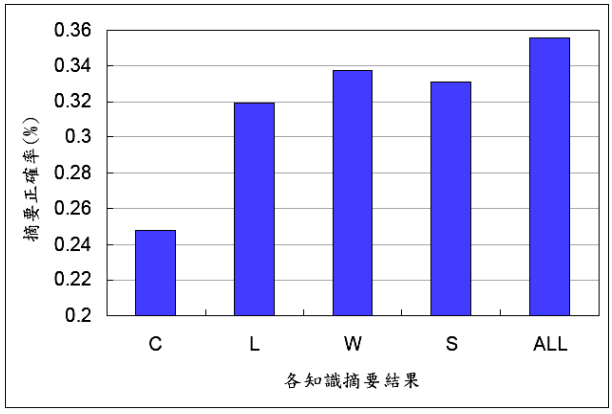


圖 8. 摘要之各分數重要性評估

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。

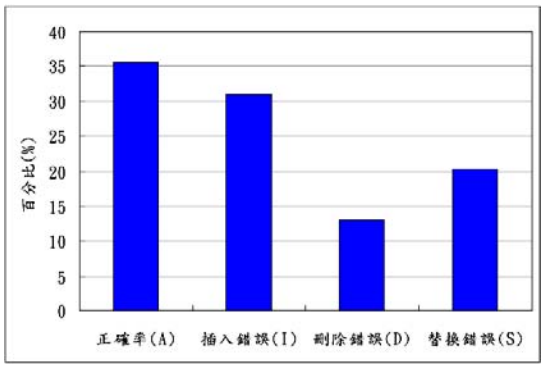


圖 9. 摘要結果正確率評估

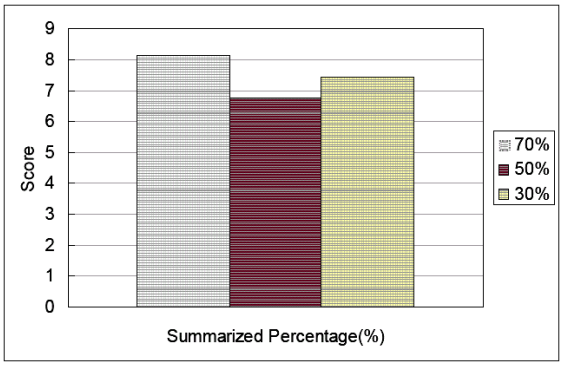


圖 10. 摘要串接及合成結果評估

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.

具相關資訊回饋能力之貝氏混合式機率檢索模型

Using Relevance Feedback in Bayesian Probabilistic Mixture Retrieval Model

簡仁宗 楊敦淇
國立成功大學資訊工程學系
Email: jtchien@mail.ncku.edu.tw

摘要

本篇論文提出新穎之相關回饋(Relevance Feedback)方法並應用於混合式機率檢索系統(Mixture Probability Model)以提昇檢索效能。相關資訊回饋法以往最常用的技術是查詢句擴充法（Query Expansion），本回饋方式是架構在以混合式機率模型為主的檢索系統上，為了加強檢索效能，我們是在查詢句擴充法中，強調不同查詢詞的重要性，所以提出查詢詞權重重調整（Query Term Reweighting）技術；此外，我們也利用檢索出來的前 *N* 名文件和資料庫的每份文件個別重調成新的文件語言模型，以提供較好的文件語言模型提供檢索時使用。在查詢字權重之重調整部分以最佳相似度（Maximum Likelihood）為估測準則，而文件語言模型之調整部分先後以最佳相似度與最佳事後機率（Maximum *a Posteriori*）為估測準則供我們對照比較，並使用了 EM（Expectation Maximization）演算法去估測出適當的參數。實驗結果顯示使用資訊回饋及貝氏語言模型調整可有效提升文件檢索正確率。

1. 簡介

目前資訊檢索的型態大致可分為[1]：布林式(Boolean)檢索，類神經網路(Neural Network)檢索，向量式(Vector-Based)檢索以及機率式(Probability-Based)檢索等；以上數種檢索式中，目前在搜尋引擎上較為廣泛使用的為布林式檢索，目前常被使用的 Google 搜尋引擎根據網站上的檢索方式說明[19]，整個過程便是從布林運算發展，以比對字串為主的檢索。

資訊檢索的領域裡，有一種能有效地提昇效能的方法稱為相關資訊回饋（Relevance Feedback），它是使用前一次檢索所得到的文件分數中，找出檢索分數較高的前 *N* 篇或是適當的 *N* 篇文件，從其中擷取可用的資訊回饋加入下一次遞迴的檢索中，增強檢索所需要的資訊；其概念是假設某些和查詢句相關的文件檢索後排名很前面，但是某些相關文件（Relevant Document）語意上雖相似，但是也許內容出現了問題，例如：查詢詞出現的比較少，因此檢索的排名會比較後面，所以利用排名前面的相關文件去想辦法拉抬排名於後的相關文件。在過去常用於資訊檢索的相關回饋方式主要為查詢句擴充和查詢詞權重再調整。

一般使用者在搜尋引擎所下的查詢句通常都不長，因此提供的資訊並不多；另外，相關回饋於資訊檢索之研究大部分都是針對向量模型檢索系統，對於以機率為主的 *n*-gram 語言模型檢索系統，只能使用查詢句擴充法來提昇檢索效能，但是觀察整個檢索流程，發現將每一份文件視為一個語言模型時，裡面能提供的資訊其實也不多，會造成不同文件之間的混淆，假若能利用前一次遞迴檢索出排名較高的數篇文件去調整資料庫中的文件，與它們相關的文件提供較多的資訊，與它們不相關的文件便提供少一點的資訊，那麼在下一次遞迴的檢索中，便能減少一些文件與文件之間混淆的程度，而達成有效的自動檢索過程；此外在一些檢索系統上會用到的查詢詞權重的觀念若能引進來，將這些參數額外地加到混合式 *n*-gram 檢索架構中輔助原本的語言模型計算分數，並利用回饋的資訊去重調整權重，如此應可加強一些重要字的分數以提昇檢索效能。所以我們以混合式機率檢索架構為主，於此架構上使用相關資訊回饋。除了沿用先前的查詢句擴充方式外，我們嘗試在檢索式中針對每個查詢詞加入權重的參數，將前一次遞迴檢索分數最高的 *N* 篇文件去做查詢詞的權重重調整，期望以這 *N* 篇文件內的分布情形，去調整出每個查詢詞的重要程度，此外，針對文件內提供資訊過少的問題，我們使用最佳事後機率（Maximum *a Posteriori*）法則將這 *N* 篇文件和資料庫裡的每一份文件調成新的文件混合語言模型，利用這 *N* 篇文件模型適當的補充資訊予資料庫內的文件。

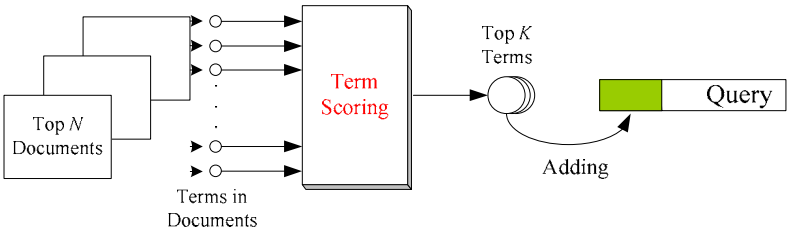
2. 相關研究

2.1 相關資訊回饋

使用者提供給檢索系統的查詢句中，通常句子的長度都偏短，如此能提供之資訊便相對的減少，容易造成檢索時產生混淆的情況[3]。為了此類問題的解決，有很多研究朝著上下文分析，語意分析等自然語言處理以及文件內容標記的定義，如 XML 上來發展。而利用前一次檢索所獲得之相關文件來調整查詢句，對於檢索效果

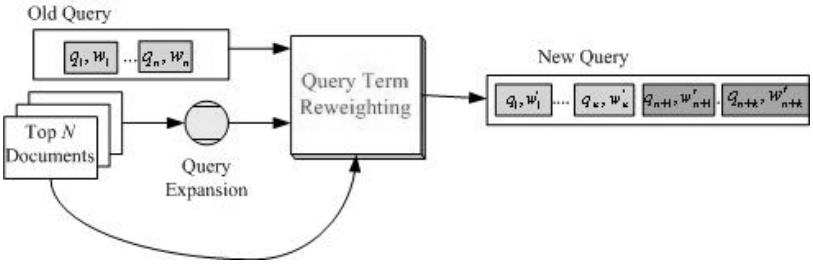
也有相當程度之改善。

在現有的檢索模型架構中，查詢句擴充的目的就是為了能從相關文件內多找些同主題中常會出現的詞，以補充查詢句過短之缺點，所以當查詢句的長度越長，查詢句所含之資訊越多，查詢句擴充所能提供的效果就越有可能降低。其架構如圖一所示：



圖一、查詢句擴充架構圖

查詢詞的權重調整以前一回檢索排名較高之數篇文件裡面的詞分布情形為依據，為了強調某些常出現的詞而設計的計算式，以便在向量等檢索模型的表現中會更趨近同主題文件。架構如下圖所示：



圖二、查詢詞權重重調整架構圖

相對於查詢詞 q 有一個對應的權重 w，經過查詢句擴充及權重之調整後，除了新增詞於查詢句之外，原本的權重也被更新過了。

2.2 向量檢索模型之資訊回饋

以向量檢索模型而言，針對查詢句 Q 和文件 d，利用每個字的出現次數以及在文件間分布的情形去算出特徵向量 $\mathbf{q}$ 和 $\mathbf{d}$，計算查詢句和文件的相似度以內積（Inner Product）運算為主，在此情形下，回饋資訊必然以向量的型態去調整查詢句的向量從 $\mathbf{q}$ 到 $\tilde{\mathbf{q}}$，目前常見之向量型態的資訊回饋略舉兩例[8]：

Rocchio：
$$\tilde{\mathbf{q}} = \mathbf{q} + \frac{\beta}{|V|}\sum_{\mathbf{d}_i \in V}\mathbf{d}_i - \frac{\gamma}{|U|}\sum_{\mathbf{d}_j \in U}\mathbf{d}_j \tag{1}$$

Ide dec-hi：
$$\tilde{\mathbf{q}} = \mathbf{q} + \sum_{\mathbf{d}_i \in V}\mathbf{d}_i - \max_{\mathbf{d}_i \in U}\mathbf{d}_j \tag{2}$$

β 和 γ 是經實驗所找出的經驗值，V 是指和查詢句 Q 相關的文件群，U 是指和查詢句 Q 不相關的文件群，其方法是利用找出相關與不相關的文件群來改善查詢句向量 $\mathbf{q}$，以提昇下一次遞迴的檢索效能。

3. 使用相關資訊回饋於貝氏混合式機率檢索

3.1 *N*-gram 模型的建立

N-gram[6]模型在自然語言處理中是常見的技術，應用的範圍很廣，有資訊檢索、語音辨識、光學文字辨識和文件分類等方向。本論文的主架構混合式機率檢索即是 *n*-gram 模型於資訊檢索上的應用，我們首先針對 *n*-gram 的建立方法與評估作概略的介紹。

語言模型主要的功能是在評估一段文句出現的機率，假設有一查詢句 Q 其長度為 T 並且是由一段詞序列 $q_1, q_2, \ldots, q_T$ 所組成，則 Q 出現的機率可以寫成：

$$\begin{aligned} P(Q) &= P(q_1, q_2, \ldots, q_T) = P(q_1)P(q_2 \mid q_1)\cdots P(q_T \mid q_1, q_2, \ldots, q_{T-1}) \\ &= \prod_{t=1}^{T} P(q_t \mid q_1, q_2, \ldots, q_{t-1}) \end{aligned} \tag{3}$$

但是此種方法的計算量與空間使用量太大而無法實現，為解決這個問題所以有 *n*-gram 模型的產生，在 *n*-gram 模型中，它是假設一個詞出現的機率只跟前面 *n*-1 個詞有關，因此（3）式可以近似為

$$P(Q) = P(q_1, q_2, \ldots, q_T) \cong \prod_{t=1}^{T} P(q_t \mid q_{t-n+1}^{t-1}) \tag{4}$$

其中 q_{t-n+1}^{t-1} 代表 $q_{t-n+1}, q_{t-n+2}, \cdots, q_{t-1}$ 詞序列如此一來使用 *n*-gram 可以大量節省計算時間與記憶體，讓實用性大為提高。而建立 *n*-gram 機率模型 $P(q_t \mid q_{t-n+1}^{t-1})$ 的基本式如下：

$$P(q_t \mid q_{t-n+1}^{t-1}) = \frac{c(q_{t-n+1}^{t})}{c(q_{t-n+1}^{t-1})} = \frac{c(q_{t-n+1}^{t})}{\sum_{q_j} c(q_{j-n+1}^{j})} \tag{5}$$

其中 $c(q_{t-n+1}^{t})$ 代表 q_{t-n+1}^{t} 在訓練文集中出現的次數

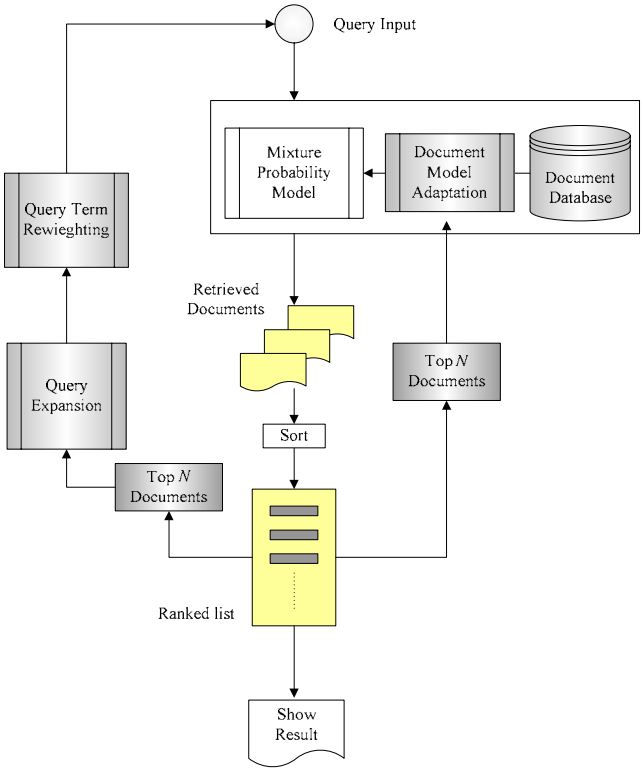
3.2 混合式機率檢索模型

混合式機率檢索是以 *n*-gram 模型為主的資訊檢索技術。此架構原出於[11]，被稱為隱藏式馬可夫模型（Hidden Markov Model）[13]，可是因為此架構只有單一狀態，故稱為"混合式機率模型"較為適當。裡面包含了 $P(q_i \mid d_j)$和 $P(q_i \mid q_{i-1}, d_j)$這種相對於文件 d_j 的 Uni-gram 和 Bi-gram，並且為了表示出查詢句 Uni-gram 和 Bi-gram 一般分布的情形而引入了一個背景語料(*Corpus*)，這個背景語料的語言模型 $P(q_i \mid Corpus)$ 和 $P(q_i \mid q_{i-1}, Corpus)$ 是由大批的文件集合依照（5）所算出來的。而相似度的量測 $P(Q \mid d_j)$，即是由查詢句 Q 中每個詞，循序計算的機率值，累計的結果即可視為其相關程度，則查詢句 Q 相關於文件 d_i 的機率表示如下：

$$\begin{aligned} P(Q \mid d_j) = & \left[\lambda_1 P(q_1 \mid d_j) + \lambda_2 P(q_1 \mid Corpus)\right] \\ & \times \prod_{t=2}^{|Q|} \begin{bmatrix} \lambda_1 P(q_t \mid d_j) + \lambda_2 P(q_t \mid Corpus) \\ + \lambda_3 P(q_t \mid q_{t-1}, d_j) + \lambda_4 P(q_t \mid q_{t-1}, d_j) \end{bmatrix} \end{aligned} \tag{6}$$

關於混合式 *n*-gram 檢索模型之推導與詳細內容可參考[17]。

本論文的重點便是在混合式機率檢索模型內加入回饋的機制，下方圖三為本論文檢索模型之新穎回饋流程架構，而其中的研究，主要在相關資訊回饋方式有三個改進方向：*查詢句擴充*、*查詢詞權重重調整*和*文件模型調整*（Document Model Adaptation）；於前一回的檢索及文件排序完成後，先使用 Top N 的文件做查詢句擴充、查詢詞權重重調整之後更新 Query 後,於檢索開始時調整文件模型。而每一程序之處理過程，將於接下來之各節做詳細的描述。



圖三、加入相關資訊回饋機制的檢索流程

3.3 查詢句擴充

過去的研究中，都顯示出了查詢句擴充的效果，並發現查詢詞的挑選，應該以接近檢索時的計算式為主，如此較有機會補充適合此檢索系統的查詢詞；在本論文裡，因為採用語言模型的方式檢索，所以擴充詞挑選的

方式便簡單地利用字詞相對於文件的機率，因此我們根據下式為選擇判斷式，對每個出現於排名前 N 名文件 $\{\hat{d}_1,\cdots,\hat{d}_N\}$ 內的詞做排名，並找出名次最高的前幾個詞加入查詢句中：

$$\sum_{j=1}^{N} P(q_t \mid \hat{d}_j) P(\hat{d}_j) P(Q \mid \hat{d}_j) \tag{7}$$

其中 q_t 是出現在前 N 名文件內的詞，$P(q_t \mid \hat{d}_j)$ 為在文件 $\hat{d}_j$ 的 Uni-gram 的機率；$P(\hat{d}_j)$ 為事前（Prior）機率，是文件 $\hat{d}_j$ 長度（詞數）於 N 篇文件長度總和的比例；$P(Q \mid \hat{d}_j)$ 就是前一次遞迴的檢索中，查詢句 Q 和文件 $\hat{d}_j$ 比對的分數。

3.4 查詢詞權重重新調整

在向量檢索的相關資訊回饋中有一種常被應用且變化的 Rocchio 公式，其焦點放在正面的例子（相關文件），而忽略了負面例子的加入（不相關文件）。從這裡，我們可以得到一個想法，有些查詢詞因為有其重要的代表性，所以在某些相關文件中出現的次數比較多，使得這些相關文件的排名會比較前面，但是這些查詢詞在其他的相關文件出現次數比較少，於是使得這些文件就會被排名比較後面；假設我們能夠對這些查詢詞適當地分配一權重，於式子中可改變每個查詢詞所提供之資訊，比較重要的詞給予較高的權重，相反地，對於不重要的詞給予較低的權重，如此，期望對於這些含有具代表性查詢詞比較少的相關文件在計算對查詢句之相似分數時能夠有所提昇。我們把簡化語言模型的機率檢索式來看：

$$P(Q \mid d_j) = \prod_{t=1}^{|Q|} P(q_t \mid d_j) \tag{8}$$

若可以從式子中抽取出一個因子 κ_t 代表查詢詞 q_t 的權重，在第一輪的最初檢索過程中，初始的查詢句因為沒有其他資訊介入，所以每個 κ_t 可視做 1，對原結果不受任何影響，即：

$$P(Q \mid d_j) = \prod_{t=1}^{|Q|} \kappa_t P(q_t \mid d_j) \tag{9}$$

至於下一輪的更新若以 $\kappa_t + \Delta\kappa_t$ 表示，因為在使用者所下的查詢句中，經過調整出來的權重必有一定的代表性，當權重值高時，則此查詢詞可看做檢索之關鍵，在此情形下，舊的查詢詞權重也應該在下一回的回饋程序中保留，並加上一更新權重值 $\Delta\kappa_t$ 以做調整，每一個查詢詞的權重更新值依據此分配的量做正規化，得到以下更新後的機率值：

$$\hat{\kappa}_t = \frac{\kappa_t + \Delta\kappa_t}{\sum_{k=1}^{|Q|} (\kappa_k + \Delta\kappa_k)} \tag{10}$$

其中 $\sum_{t}^{|Q|} \hat{\kappa}_t = 1$，而 $|Q|$ 意指查詢句的長度。令 D 是一個集合，裡面是檢索分數中排名前 N 名的文件，$D = \{\hat{d}_1, \hat{d}_2, \cdots, \hat{d}_N\}$，$K$ 也是一個集合，裡面是每個查詢詞相對應的權重，$K = \{\kappa_1, \kappa_2, \cdots, \kappa_{|Q|}\}$。因為檢索式之目的是為了提昇與查詢句相關文件的分數，以便增加相關文件與非相關文件之差別，我們為每個查詢詞加入相對應的權重也是為了這個原因，所以我們必須找出一組適當的權重，而這組權重是確定可以提昇與相關文件之相似度：

$$\hat{K} = \arg\max_{K} P(Q \mid D, K) \tag{11}$$

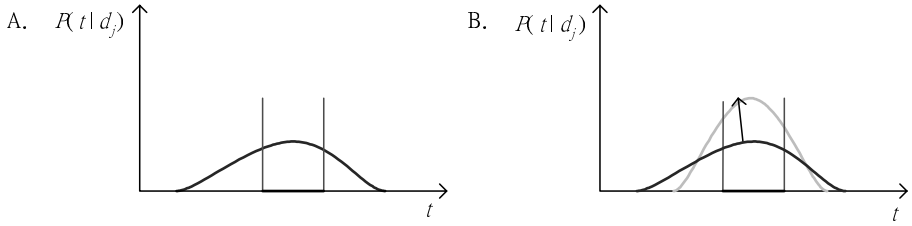
其中能觀察到的資料是我們所使用的語言模型與查詢句 Q，這是不完整的資料集（Incomplete Data），但是權重為未知的參數，如此只能想辦法去近似出適當的權重，所以我們以最佳相似度估測（Maximum Likelihood Estimation，MLE）為標準（Criterion），使用 EM 演算法[5]的步驟去推估出新的估測值 $\hat{\kappa}_t$ 的公式如下：

$$\hat{\kappa}_t = \frac{\sum_{j=1}^{N} \frac{\kappa_t P(q_t \mid \hat{d}_j)}{\sum_{k=1}^{|Q|} \kappa_k P(q_k \mid \hat{d}_j)}}{\sum_{l=1}^{|Q|} \sum_{j=1}^{N} \frac{\kappa_l P(q_l \mid \hat{d}_j)}{\sum_{k=1}^{|Q|} \kappa_k P(q_k \mid \hat{d}_j)}} = \frac{\sum_{j=1}^{N} \frac{\kappa_t P(q_t \mid \hat{d}_j)}{\sum_{k=1}^{|Q|} \kappa_k P(q_k \mid \hat{d}_j)}}{N} \tag{12}$$

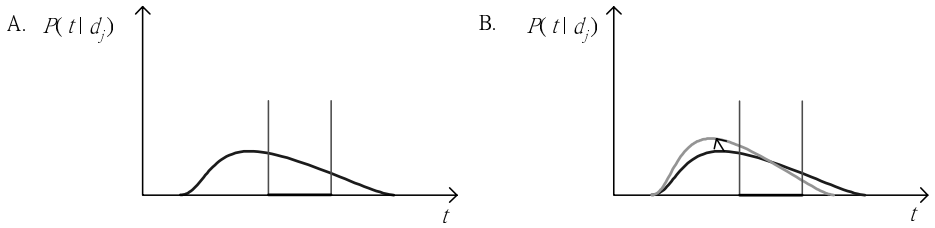
$\hat{\kappa}_t$ 值將會依照（12）遞迴地被訓練出來。

3.5 貝氏混合式機率模型調整

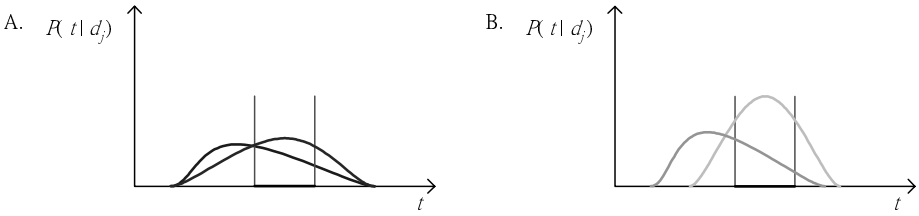
對於文件模型，我們嘗試利用前一回檢索所得到排名前 N 名文件的文件語言模型來補充目前查詢句正要比對的文件 d_j 其文件模型的資訊，使得文件 d_j 的語言模型能適用於目前查詢句的檢索。而調整文件語言模型之目的以下圖來說明：（水平軸：詞彙 t；垂直軸：在文件 d 內的詞彙 t 機率 $P(t \mid d_j)$。）



圖四、與查詢句相關文件的語言模型假想圖，(A)未經過相關資訊回饋的調整和(B)經過相關資訊回饋的調整



圖五、與查詢句不相關文件的語言模型假想圖，(A)未經過相關資訊回饋的調整和(B)經過相關資訊回饋的調整



圖六、相關與不相關文件的模型重疊假想圖，(A)未經過相關資訊回饋的調整和(B)經過相關資訊回饋的調整

在這些語言模型的機率分布裡，水平軸上出現框線的間隔，即代表查詢詞出現的範圍，所以此間隔與曲線圍起來之區域可以說是查詢句於文件模型內可能會用到的機率值，圖五是文件和查詢句相關的情形，圖六是和查詢句不相關的文件機率分布情形，圖七為兩種文件模型對於查詢句的機率分布重疊的比較。當原始的兩篇文件其分布情形為圖五-(A)和圖六-(A)，重疊之後得到圖七-(A)，並由圖七-(A)可得知，對於目前的查詢句經過比對計算得到的分數差距比較小，這有可能會造成檢索排名出現問題。若是在前一回檢索出來的結果，前 N 名的文件其資訊是可以利用的，意即可用來調整每份文件的模型，經過調整後得到圖五-(B)，圖六-(B)以及將這兩種調整過的文件模型分布重疊後得到圖七-(B)的情形，結果是相關與不相關之語言模型差異度變大了，如此一來就可以減少語言模型的模糊情形，並且有利於相關文件檢索分數的提昇。

假設資料庫裡的文件 d_j 和相對應的文件語言模型，可以和排序列表（Ranked List）內排名前 N 名的文件一起作用，來產生出新的文件語言模型，此時即是把排名前 N 名的文件語言模型和文件 d_j 的文件語言模型混合成一個新的文件語言模型，我們將導入權重參數作語言模型的合併，原（8）在加入了相關資訊回饋的機制，於第二次以及之後的遞迴檢索過程，將會變成

$$\tilde{P}(Q \mid d_j) = \prod_{t=1}^{|Q|} \tilde{P}(q_t \mid d_j) \tag{13}$$

（13）之變換過程描述於下：

1.令 M_j 為一個相對應於 $D_j = \{d_j, \hat{d}_1, \cdots, \hat{d}_N\}$ 之權重參數（Mixture Weight）集合,裡面放置相對應的合併權重參數，$M_j = \{m_{j,0}, m_{j,1}, \cdots, m_{j,N}\}$，$m_{j,0} + \sum_{k=1}^{N} m_{j,k} = 1$，$D_j$ 和 M_j 皆是針對文件 d_j 用到的資訊。而 $m_{j,k}$ 意指混合數 k 之是語言模型權重，會隨著文件 d_j 有所不同。

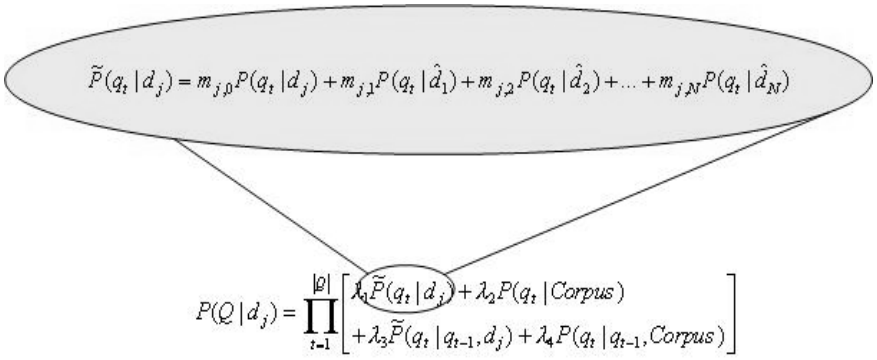
2.因為查詢詞是和文件內容相關，而文件內容和文件模型相關，若是文件模型產生文件內容之機率能更適當，則一個和此文件相關之查詢句 Q，使用此文件模型產生出來的機率也會更適當；所以文件語言模型權重的訓練過程即是以文件內容為估測之主要內容，目的是要針對文件內容 d_j，利用回饋的資訊去調出最適當的文件語言模型，因此最後依照我們所選擇之標準去找出適當的合併參數。

綜合上述三點及原本的想法，我們得到以下的式子：

$$\tilde{P}(q_t \mid d_j) = m_{j,0} P(q_t \mid d_j) + m_{j,1} P(q_t \mid \hat{d}_1) + m_{j,2} P(q_t \mid \hat{d}_2) + \ldots + m_{j,N} P(q_t \mid \hat{d}_N) \tag{14}$$

此式對原檢索模型之影響如圖八，圖中$\{\lambda_1,\ldots,\lambda_4\}$是混合式檢索模型的參數，是在建立混合式機率模型時就已經計算好了，這裡是強調在做模型參數$P(q_t \mid d_j)$之調整。若將d_j當成$\hat{d}_0$,結合前N名文件$\hat{d}_1,\hat{d}_2,\cdots,\hat{d}_N$，於是此式可轉換為

$$\tilde{P}(q_t \mid d_j) = \sum_{k=0}^{N} m_{j,k} P(q_t \mid \hat{d}_k) \tag{15}$$



圖七、混合式機率模型之調整

A.最佳相似度估測

若使用最佳相似度（Maximum Likelihood，ML）估測法則，其最佳參數M_j^{ML}計算如下[17]

$$M_j^{ML} = \arg\max_{M_j} P(Q \mid D_j, M_j) \tag{16}$$

在此$\sum_{k=0}^{N} m_{j,k} = 1$為一限制（Constraint）。我們必須執行有限制的最佳化（Constraint Optimization），利用文件本身與回饋之文件調整出一個更符合該文件之模型出來，因為參數 M_j未知，已知的觀察資料為文件集合與語言模型集合，資料並不完全，所以依照 EM 演算法去推出合併參數的式子，其結果如下

$$m_{j,k}^{ML} = \frac{\sum_{t=1}^{|Q|} \frac{m_{j,k} P(q_t \mid \hat{d}_k)}{\sum_{l=0}^{N} m_{j,l} P(q_t \mid \hat{d}_l)}}{\sum_{v=0}^{N}\sum_{t=1}^{|Q|} \frac{m_{j,v} P(q_t \mid \hat{d}_v)}{\sum_{l=0}^{N} m_{j,l} P(q_t \mid \hat{d}_l)}} = \frac{\sum_{t=1}^{|Q|} \frac{m_{j,k} P(q_t \mid \hat{d}_k)}{\sum_{l=0}^{N} m_{j,l} P(q_t \mid \hat{d}_l)}}{|Q|} \tag{17}$$

B.最佳事後機率估測

雖然使用最佳相似度為標準可估出一組參數去調整文件的語言模型，不過以最佳事後機率（Maximum *a Posteriori*，MAP）為標準，和 ML 比起來，在估測過程中多加入了事前機率通常是有助於在稀疏（Sparse）資料條件下的估測[7]。在進行MAP 的推導之前，我們定義所需的參數 Ω_j如下，$\Omega_j = \{m_{j,k}, P(q_t \mid \hat{d}_k), 0 \le k \le N, 1 \le t \le |Q|\}$

$$\Omega_j^{MAP} = \arg\max_{\Omega_j} g(\Omega_j \mid Q, D_j) = \arg\max_{\Omega_j} P(Q \mid D_j, \Omega_j) g(\Omega_j) \tag{18}$$

其中$g(\Omega_j)$是參數 Ω_j的事前機率，我們假設為 Dirichlet 機率分佈，而 j 是指目前查詢到第 j 篇文件 d_j。事前機率$g(\Omega_j)$如下所示

$$g(\Omega_j) \propto \prod_{k=0}^{N} m_{j,k}^{v_{j,k}-1} \prod_{t=1}^{|Q|} P(q_t \mid \hat{d}_k)^{l_{j,k,t}-1} \tag{19}$$

$v_{j,k}$ 和 $l_{j,k,t}$ 是 Dirichlet 機率分佈的 Hyperparameter。k 代表回饋文件的編號，k=0 時，代表資料庫裡正被查詢到的文件，$k = 1,\cdots,N$ 為前一回找出排名前 N 的文件。

針對混合參數$m_{j,k}$推導出來的結果為

$$m_{j,k}^{MAP}=\frac{\sum_{t=1}^{|Q|}\frac{m_{j,k}P(q_t\mid\hat{d}_k)}{\sum_{l=0}^{N}m_{j,l}P(q_t\mid\hat{d}_l)}+(v_{j,k}-1)}{\sum_{v=0}^{N}\left[\sum_{t=1}^{|Q|}\frac{m_{j,v}P(q_t\mid\hat{d}_k)}{\sum_{l=0}^{N}m_{j,l}P(q_t\mid\hat{d}_l)}+(v_{j,v}-1)\right]} \tag{20}$$

其語言模型參數 $P(q_t\mid\hat{d}_k)$ 部分，其最後推導結果如下

$$P^{MAP}(q_t\mid\hat{d}_k)=\frac{n_{t,k}\left(\frac{m_{j,k}P(q_t\mid\hat{d}_k)}{\sum_{l=0}^{N}m_{j,l}P(q_t\mid\hat{d}_l)}+(l_{j,k,t}-1)\right)}{\sum_{v=1}^{|Q|}\left[n_{v,k}\left(\frac{m_{j,k}P(q_v\mid\hat{d}_k)}{\sum_{l=0}^{N}m_{j,l}P(q_v\mid\hat{d}_l)}+(l_{j,k,v}-1)\right)\right]} \tag{21}$$

其中 $n_{t,k}$ 為詞組 q_t 在第 k 個混合數出現的次數。

C. Hyperparameter 的初始化與更新方式

在 Hyperparameter 的初始化的部分，我們參考[10]並採用以下的公式做初始化

$$v_{j,k}^{(0)}=1+\varepsilon\cdot\overline{m}_{j,k} \tag{22}$$

$$l_{j,k,t}^{(0)}=1+\varepsilon\cdot\overline{P}(q_t\mid\hat{d}_k) \tag{23}$$

其中 $\overline{m}_{j,k}$ 及 $\overline{P}(q_t\mid\hat{d}_k)$ 的計算方式是將訓練資料估測出來的最佳相似度值 $m_{j,k}^{ML}$ 及 $P^{ML}(q_t\mid\hat{d}_k)$，進行取平均值的運算而得到的。$0<\varepsilon<1$ 是一個加權的係數，目的是去調整事前資料的權重。而 Hyperparameter 的更新公式是根據 Dirichlet 事前機率分布是屬於 Conjugate Prior 的特性推導出如下的結果[10]

$$v_{j,k}^{new}=v_{j,k}^{old}+\sum_{t=1}^{|Q|}\frac{m_{j,k}P(q_t\mid\hat{d}_j)}{\sum_{l=0}^{N}m_{j,l}P(q_t\mid\hat{d}_l)} \tag{24}$$

$$l_{j,k,t}^{new}=l_{j,k,t}^{old}+\frac{m_{j,k}P(q_t\mid\hat{d}_j)}{\sum_{l=0}^{N}m_{j,l}P(q_t\mid\hat{d}_l)} \tag{25}$$

把前一次遞迴算出 Hyperparameter 的 $v_{j,k}^{old}$ 及 $l_{j,k,t}^{old}$，用此公式更新到 $v_{j,k}^{new}$ 及 $l_{j,k,t}^{new}$。

4. 實驗

4.1 實驗環境-斷詞工具與實驗文集說明

為了將本論文方法實現在中文新聞資訊檢索系統中，首先必須製作了一套詞典，這個詞典之功用是為了能將文件裡得句子斷成更小的詞單位，同時將每一個斷出來的詞轉成詞典中對應的編號，詞典中主要部分是來自 CKIP (Chinese Knowledge Information Processing)中文詞庫[18]，主要是利用國語日報辭典中約四萬目詞的原始資料加以分類，並且附加部分的語法及語意訊息在其中，但在本論文，只使用到詞出現的頻率，並無使用到語法與語意資訊，我們只有取出其中一、二、三、四詞的部分作為基本辭典。

實驗過程所使用的文集為 TDT2（Topic Detection and Tracking Phase 2），是由 LDC（Linguistic Data Consortium）所收集的新華社新聞文件。總計有 11,161 篇西元 1998 年 1 月 1 日到 6 月 30 日的新聞，總共有 20 個主題，1,183 篇新聞文件，剩下 9,978 篇無標出主題文件，從其中取出 4,815 篇來算出檢索模型中所需要的背景語言模型，由於我們所使用之新華社新聞並無經過分類（國際、政治、財經及體育…等類別），為了保持背景語言模型之平衡性，針對每月每日之新聞以隨機方式抽出，平均每月取八百篇文件。標明主題之 1,183 篇文件與尚未使用無標主題的 5,163 篇文件合併為本實驗中的測試文集，總共 6,346 篇。而實驗測試時所需之查詢句，

為了模擬使用者使用檢索之情形，於是從標明主題之 1,183 篇文件中，挑選出 102 篇新聞文件的標題來當做查詢短句樣本，其平均長度約為 17 個字。

4.2 檢索效能的評估方法

Non-Interpolated Average Precision Rate（NAP）以單一數值來作效能評估，是文件檢索效能相當普遍的評估方式，其式子如下：

$$NAP = \frac{\sum_{i=1}^{N} \frac{i}{Rank}}{N} \tag{26}$$

舉例來說，在檢索出來的文件中實際相關的文件被排名在第一名、第二名、第四名及第六名，則 NAP 的值為 0.854 ($NAP = \frac{\frac{1}{1}+\frac{2}{2}+\frac{3}{4}+\frac{4}{6}}{4} = 0.854$)。

4.3 實驗結果

關於實驗結果表達所用的符號，以 QE 代表 Query Expansion，QTR 代表 Query Term Reweighting，MA 代表 Model Adaptation，而 ALL 代表 QE+QTR+MA。本實驗基礎架構為混合式機率檢索，以不加入任何回饋方式之檢索正確度作為我們比較的基本系統（Baseline），並使用 NAP 做評估量測。

A.不同資訊回饋法在檢索效能之影響

本部分實驗取前一回檢索分數排名於前 6 名之文件（*N*=6）做回饋，在查詢句擴充裡，每次找出分數最高之前 6 個詞（*K*=6），並對查詢句做比對，刪除重複部分，剩下的詞便可加入查詢句成為一擴充之新查詢句。我們得到基本系統的 NAP 為 66.4%，不同資訊回饋法得到的文件檢索 NAP 如下表所示

表一、不同資訊回饋演算法之實驗結果比較

回饋方式	QTR	MA	QE	QTR+QE	QTR+MA	QE+MA	ALL
NAP (%)	66.1	72.1	72.2	72.3	77.7	78.3	81.1

本實驗比較各資訊回饋演算法之效果，從圖表中可以看出各方法對於檢索之準確度皆有提昇。在查詢詞權重重調整方面，比較 QTR、QTR+QE 與 QTR+MA 這三組實驗發現，詞權重之調整雖然單獨使用之改善不甚明顯，但是若有較好的語言模型調整，則檢索效果的提昇會更顯著。此外，我們對各 QTR、QE 與 MA 這三種方式做不同的合併，亦有不同之提昇效果顯現出來，而全部合併時，檢索之準確度提昇最多。接下來，比較文件模型調整時使用 ML 與 MAP 之效果，我們以 QTR、QE 與 MA 三者合併之實驗來比較。

表二、混合式機率檢索模型調整使用 ML 及 MAP 之實驗結果比較

回饋方式	ALL (ML)	ALL (MAP)
NAP (%)	81.1	82.4

從上表中可以看出在改用 MAP 去調整文件模型後，其檢索效果的確有改進。

B.不同資訊回饋量之影響

本部分實驗將設定不同之查詢詞增加數與回饋時的文件數，並且針對文件模型調整的 ML 與 MAP 做比較，同樣地，實驗是以三者合併(ALL)的效果來觀察。

表三、不同資訊回饋量之實驗結果比較

回饋方式	ML(*K*=10, *N*=6)	ML(*K*=10, *N*=10)	ML(*K*=6, *N*=10)	MAP(*K*=6, *N*=10)
NAP (%)	80.7	81.3	82	83.5

表中 *K* 是指查詢句擴充裡，算完詞分數後所挑選的詞數量；*N* 是指使用於回饋程序中的文件數。我們很明顯的看出，增加詞的個數，造成不適當的詞加進查詢句的機會提昇，如此會造成檢索效果的降低；而提昇回饋之文件數，可以補充更多的資訊於系統內，並進一步地提高檢索效能。

C. MAP 調整之不同參數初始比較

在貝氏的文件模型調整中，對於 Hyper parameter 的初始，需使用一係數 ε 做調整，其範圍 $0 < \varepsilon < 1$，我們針對不同的 ε，以合併 QTR、QE 與 MA 的實驗結果找出可能最佳值。

表四、不同 ε 值之實驗結果比較

ε	0.2	0.5	0.8
NAP (%)	82.44	82.41	82.38

由表中的結果可看出，不同的係數雖然結果不同，但相差量是很少的，不過在其他的實驗比較中，仍以 $\varepsilon = 0.2$ 為主。

D.較短查詢句與較長查詢句之比較

在這一小節裡，我們將實驗樣本的長度分成兩群，每一樣本大於 15 個中文字的分成一群，小於或等於十五個字的分成另一群，以觀察不同長度對實驗結果之影響，同樣以 QTR、QE 與 MA 合併的實驗觀察。

表五、長句與短句之實驗結果（NAP（%））比較

平均長度(字)	基本系統	本論文方法
12.64	63.4	84.5
20.47	68.7	80.9

從表中看出，較長查詢句可用的資訊量包含較多，所以在基本系統可表現較好，但是同時參雜了一些多餘字出現，所以在回饋之後的效果容易低於較短的查詢句，不過，在這些實驗中發現到，檢索最後效果的好壞不在於查訊句的短或長，使用者所下的查詢句，其意思的表達是否明確，才是檢索效果的關鍵。

E.臺灣電子報之實驗結果

本小節實驗目的在做一組對照的結果，其資料來源為 YAHOO 奇摩網站上搜得之電子報，作為被查詢的新聞文件其範圍從西元 2002 年 1 月 25 日至 5 月 21 日與西元 2002 年 11 月 12 日至西元 2003 年 1 月 11 日總共有 7,800 篇，並取出 110 新聞文件之標題作為查詢句之樣本，背景語言模型為 CKIP 平衡語料庫。

表六、臺灣電子報實驗結果

回饋方式	基本系統	QTR	MA	QE	QTR+QE	QTR+MA	QE+MA	ALL	ALL (MAP)
NAP (%)	85.3	84.0	89.5	91.1	90.4	91.6	92.8	93.7	93.8

從實驗結果可以看出兩種實驗文集的差異，這是因為實驗資料只利用詞典去斷詞，並無做其他的處理，並且兩種文集之書寫表達方式有很多差異的存在，交互影響所造成的結果。雖然如此，但本方法之效果大致上的表現是差不多的。

5. 結論與未來研究方向

本論文於混合式檢索的架構上研究相關資訊回饋的效果，並證明我們的方法可使檢索之最後效能提昇許多；此外，在實驗中發現到，檢索系統的關鍵有二：文件模型的好壞與相關文件的回饋數；當我們調出較好的文件模型時，對於查詢詞權重、查詢句擴充或者是合併來檢索，結果都會更加優秀；相關文件的回饋數量增加時，能夠補充的資訊也會相對的增加，這有助於檢索效果的提昇。文件模型的調整方面，我們從實驗中發現了 MAP 這種加入事前資訊的準則比使用 ML 的準確度多出 1.33%左右，是可以進一步地調出更好的文件模型。在查詢句擴充方面，我們從實驗發現到"回饋有如兩面刃"這個事實，當加入查詢句的詞無法控制時，便有可能出現不適當的詞加入查詢句，使得原本句子的表達走樣，這在自動的回饋裡是不可避免的現象，所以在一個實踐的系統中必須讓使用者能夠自行判斷與干涉，如此或許才可確實將使用者想要閱讀的文件或網頁拉抬其檢索的分數。

未來，在查詢句擴充中，可嘗試不同之挑選詞的方式。我們亦可改良本回饋方式，以配合加入潛在語意資訊與增加混合式檢索系統的混合數，以提昇查詢句所能提供的資訊，使其有可能再次提高系統檢索的能力。另外，對於查詢句與文件來說，這兩者便是檢索的主角，我們目前檢索的實驗中，對於這兩者相似度的計算，就只是利用到詞頻的變化，若可以加入自然語言處理的相關技術，針對這兩者做語意、語法等結構的分析，使檢索時能夠使用之資訊量增加，並進而改良本論文內相關資訊回饋的方法，也將是提昇檢索效能之方向。一般而言，文件模型的好壞影響著檢索效能，不管是以 ML 或是以 MAP 方式去調整文件，最後都可以有明顯的改善，這說明了文件模型若有更好的調整方法，則檢索系統便有機會提供給使用者更好的搜尋結果。

參考文獻

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley Longman, pages 118-123, May 1999.

[2] Claudio Carpineto, Renato De Mori, Glovanni Romano and Brigitee Bigi, "An Information-Theoretic Approach to Automatic Query Expansion", *ACM Transactions on Information Systems*, Vol.19, No. 1, pages 1–27, January 2001.

[3] Claudio Carpineto, Glovanni Romano and Vittorio Giannini, "Improving Retrieval Feedback with Multiple Term-Ranking Function Combination", *ACM Transactions on Information Systems*, Vol. 20, No. 3, pages 259–290, July 2002.

[4] Berlin Chen, Hsin-min Wang, and Lin-shan Lee, "An HMM/N-gram-based Linguistic Approach for Mandarin Spoken Document Retrieval", *In Proceedings of the 7th European Conference on Speech Communication and Technology* (*Eurospeech*2001), Aalborg Demark, Sept. 2001.

[5] A.P. Dempster, N.M. Laird, and D.B Rubin, "Maximum Likelihood From Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society*, Vol. 39, No. 1, pages 1-38, 1977.

[6] Jelinek Frederick, *Statistical Methods for Speech Recognition*, The MIT Press, Cambridge, Massachusetts, 1997.

[7] Jean-Luc Gauvain and Chin-Hui Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observation of Markov Chains", *IEEE Transactions on Speech And Audio Processing*, Vol. 2, No. 4, pages 291-298, April 1994.

[8] Donna Harman, "Relevance Feedback Revisited", *In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1-10, 1992.

[9] Xuedong Huang, Alex Acero and Hsiao-Wuen Hon, *Spoken Language Processing-A Guide to Theory, Algorithm, and System Development*, Microsoft Research, Prentice Hall PTR, pages 73-132, 2001.

[10] Qiang Huo and Chin-Hui Lee, "On-Line Adaptive Learning of the Continuous Density Hidden Markov Model Based on Approximate Recursive Bayes Estimate", *IEEE Transactions on Speech And Audio Processing*, Vol. 5, No. 2, pages 161-172, March 1997.

[11] David R. H. Miller, Tim Leek and Richard M. Schwartz, "A Hidden Markov Model Information Retrieval System ", *In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 214-221, 1999.

[12] Jay M. Ponte and W. Bruce Croft, "A Language Modeling Approach to Information Retrieval", *In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275-281, 1998.

[13] L. Rabiner and Biing-Hwang Juang, "An introduction to hidden Markov models", *IEEE Signal Processing Magazine*, Vol. 3, Issue: 1, pages 4 –16, Jan 1986.

[14] S. E. Robertson, S. Walker, and M. Beaulieu, "Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive Track", *In Proceedings of the 7th Text Retrieval Conference* (*TREC*-7), pages 253-264, 1999.

[15] S. E. Robertson, S. Walker, "Okapi/Keenbow at TREC-8", *In Proceedings of the 8th Text Retrieval Conference* (*TREC*-8), pages 151-162, 1999.

[16] F. Song and W. Bruce Croft, "A General Language Model for Information Retrieval", *In Proceedings of the 8th International Conference on Information and Knowledge Management* (*CIKM*99')*,* ACM Press, pages 93-96, 1999.

[17]李建志,"應用混合式機率模型於新聞資訊檢索之研究", 碩士論文, 成功大學資訊工程學系, 2002.

[18] CKIP, http://godel.iis.sinica.edu.tw, 中央究院資訊科學研究所詞庫小組。

[19] Google 搜尋說明,http://www.google.com.tw/intl/zh-TW/help.html。

藍芽無線環境下中文語音辨識效能之評估與分析

Performance Evaluation and Analysis of Mandarin Speech Recognition over Bluetooth Communication Environments

陳銀城[1] 譚旦旭[1] 王新民[2] 蔡偉和[2]
國立台北科技大學電機系[1]
中央研究院資訊科學研究所[2]
E-mail: thtan@ntut.edu.tw; whm@iis.sinica.edu.tw; wesley@iis.sinica.edu.tw

摘要

本論文探討語音辨識技術於藍芽(Bluetooth)無線環境下之效能。我們分別在藍芽實際與模擬使用環境下，應用 TCC-300 語料庫及 HTK 軟體，進行一系列語者無關(Speaker Independent)的語音辨識實驗。此外，為彌補通道效應之影響，我們亦引用若干強健性技術以提升辨識率。

為評估藍芽實際使用環境下之語音辨識效能，我們將 TCC-300 語料庫轉錄成室內使用環境 0 公尺、4 公尺以及走廊使用環境 50 公尺三個藍芽操作環境語料庫，此語料庫可提供語音辨識或其他相關語音處理研究之用。實驗結果顯示，在訓練環境與測試環境完全匹配情況下，測試距離為 0、4 與 50 公尺所獲得之音節辨識率分別為 55.82%、53.54%、以及 42.74%，辨識率隨著距離增加而下降，而且遠低於在原來的 TCC-300 語料庫進行相同測試所得之 69.25%的辨識率。另外，在環境不匹配的情況下，辨識率更是大幅度地下滑。本論文即針對辨識效能衰退原因進行探討，並提供可能的改進方向。另一方面，無論是重新收集大量藍芽實際使用環境的訓練語音，或是將原始訓練語音轉錄成藍芽實際使用環境的訓練語音，均非常耗費時間及人力，有鑑於此，我們提出一套模擬藍芽實際使用環境的系統，可以自動將訓練語音模擬至藍芽實際使用環境，進而訓練出可以模擬藍芽實際使用環境的語音辨識模型。以此模擬模型辨識藍芽語音的辨識率與前述環境匹配情況下所得辨識率之差距分別為 0 公尺之 5.18% (55.82% - 50.64%)、4 公尺之 5.6%(53.54% - 47.94%)、以及 50 公尺之 14.22% (42.74% - 28.52%)，初步證實此系統具有模擬藍芽實際使用環境的實用價值。值得注意的是，本研究進行大語彙語音辨識實驗，在語音控制等實際應用上，通常指令數量相當有限，其辨識率將遠高於本論文實驗結果，據此，藍芽無線環境下之語音辨識對家庭自動化等應用應深具潛力。

一、簡介

經過多年的發展，目前已有許多語音辨識系統被開發出來[1-3]，並成功地應用在人類日常生活中，例如語音輸入鍵盤、聲控手機、聲控家電、聲控玩具，語音下單等，這些應用系統大幅提升了我們的生活品質。

1998 年 5 月，Intel、Ericsson、Nokia、IBM 以及 Toshiba 等公司共同成立藍芽聯盟(Bluetooth Special Interest Group, Bluetooth SIG)，研議制訂一種兼具低功率、低成本優勢的短距離無線通訊標準，此即藍芽通訊協定的由來。藍芽使用 2.4 GHz (2.402~2.480 GHz)之免費頻帶，利用跳頻(Frequency Hopping)技術以避免同頻帶之干擾。藍芽不像紅外線傳輸會受到須在視線範圍(Line of Sight, LOS)內直線連線之限制，且其價格已下降至合理門檻，因此，逐漸成為 PDA、手機、筆記型電腦之標準配備。藍芽可分別使用 ACL (Asynchronous Connection-Less)及 SCO (Synchronous Connection-Oriented)通道傳送數據及語音信號。

Bluetooth SIG 於 2003 年 11 月正式頒佈 Bluetooth 1.2 版標準，除與現行的 1.1 版標準向下相容外，1.2 版增加了可降低同頻干擾的適應性跳頻（Adaptive Frequency Hopping, AFH）技術以及 Extended SCO(eSCO)通道，eSCO 為具備錯誤偵測及重傳能力之語音通道，可提升聲音訊號傳輸品質。1.2 版相關產品預計 2004 年 12 月上市。

家庭及辦公室自動化是人類一直追求的目標，也陸續有各式各樣的系統被開發出來[4]，惟多數仍存在有線的束縛，且控制方式仍多採手動按鍵，為改善上述缺點，提升自動化系統的品質，整合無線及語音辨識技術的研究已日益受到重視。目前探討無線通訊環境下之語音辨識的研究大都針對 GSM 環境[5, 6]，探討藍芽環境下之語音辨識的研究則尚在萌芽階段。Bawab 等人利用分散式語音辨識(Distributed Speech Recognition, DSR)架構以及藍芽 ACL 通道，針對特徵參數傳輸之封包遺失現象提出內插法(Interpolation)來改善辨識率[7]。Nour-Eldin 等人利用藍芽 SCO 語音傳輸通道，探討 802.11 與藍芽系統之同頻干擾所導致封包遺失對語音辨識效能的影響，並修正藍芽 CVSD(Continuous Variable Slope Deltamodulation, CVSD)解碼器改善辨識效能[8]。上述兩組團隊的研究均以模擬為主，並未在藍芽實際使用環境下進行實證。為深入探討此一問題，本研究先在藍芽實際使用環境下錄製語料庫，進行相關實驗。接著，依據藍芽規範及實際使用環境條件，建構一個模擬藍芽實際使用環境的系統，可以自動將訓練語音模擬至藍芽實際使用環境，進而訓練出可以模擬藍芽實際使用環境的語音辨識模型。

本論文第二節介紹藍芽無線技術。第三節敘述 TCC-300 語料庫及藍芽實際使用環境與藍芽模擬環境下語音資料庫之建立。第四節針對藍芽實際使用環境與模擬環境進行語音辨識效能的評估。第五節為結論，並探討本研究未來可改善的方向。

二、藍芽

藍芽是一種低功率(1 mW ~ 100 mW)、短距離(10 m ～ 100 m)無線通訊技術，可以讓內嵌藍芽模組的各種電子裝置(例如資訊家電、手機、電腦等)形成一個無線個人區域網路(Personal Area Network, PAN)。一個藍芽設備至多可以同時連結另外七個藍芽設備，每個藍芽裝置均可擔任 Master 或 Slave 的角色。由於藍芽傳輸不受方向的限制，故能取代現有的紅外線裝置，同時也能解決裝置間纜線過多的問題。此外，藍芽使用全球通用的 ISM(Industrial, Scientific, and Medical)頻帶中的 2.402 ～ 2.480 GHz 頻段，並將 ISM 頻段切割成 79 個頻道以供跳頻之用，每個頻道頻寬為 1 MHz，跳頻速率每秒 1600 次，由於此頻帶毋須申請使用執照，所以有助於藍芽裝置的普及化。

藍芽使用 SCO 連結(Synchronous Connection-Oriented Link)傳輸語音，SCO 連線屬電路交換的同步傳輸型態，每一條 SCO 支援 64 Kbit/s 的語音通訊，一旦建立 SCO 通道，Master 和 Slave 即可直接發收 SCO 封包，進行單點對單點的對稱連線服務。SCO 連結使用之封包如表 1 所示，包括 HV 系列及 DV 封包，皆未包含 CRC，且不允許重傳，本論文採用的封包格式為未具備通道編碼的 HV3。

基本上，藍芽封包涵蓋 Access Code、Header 以及 Payload 三個部份。但依照封包種類的不同，如圖 1 所示，又可分為下列三種情況：

1. 若封包須傳送資料，則封包內將包含 Access Code、Header 以及 Payload 三個部份。
2. 若不傳送資料，則封包將只含 Access Code 和 Header 兩個部分。
3. 若封包沒有 Header 時，則只剩 Access Code，例如 ID Packet。

圖 2 為 SCO 封包中之 HV3 封包，其 Payload Data 經過 Whitening 之後 Payload 固定為 240 bits，資料傳輸率為 8 KBytes/sec。

表 1. SCO 封包格式

Type	Payload Header (Bytes)	User Payload (Bytes)	FEC	CRC	Time Slot Interval	Symmetric Max. Rate(KB/s)
HV1	NA	10	1/3	No	2	64.0
HV2	NA	20	2/3	No	4	64.0
HV3	NA	30	No	No	6	64.0
DV	1 D*	10+(0 ~ 9) D*	2/3 D*	Yes D*	V**	64.0+57.6 D*

註: * 代表只有數據資訊部分，** 代表只有語音資訊部分

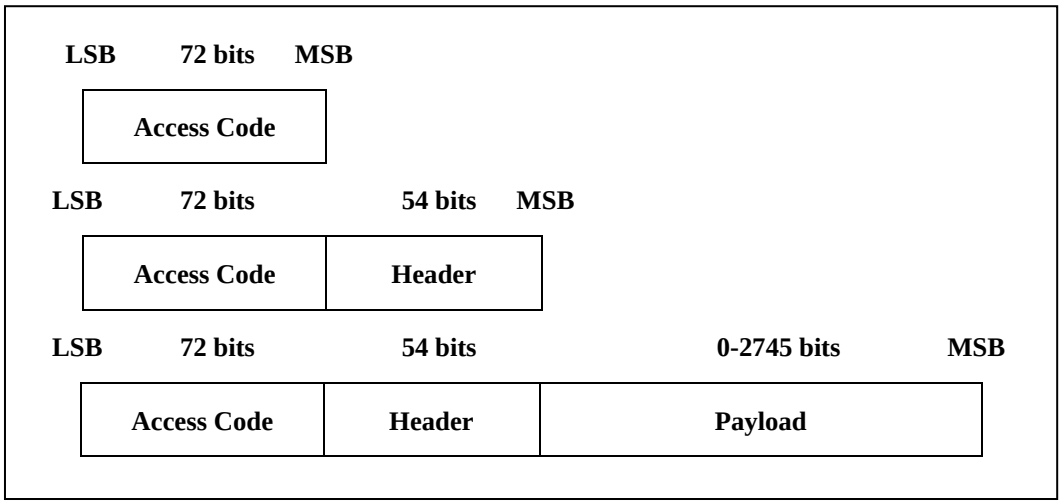


圖 1. 藍芽封包標準格式

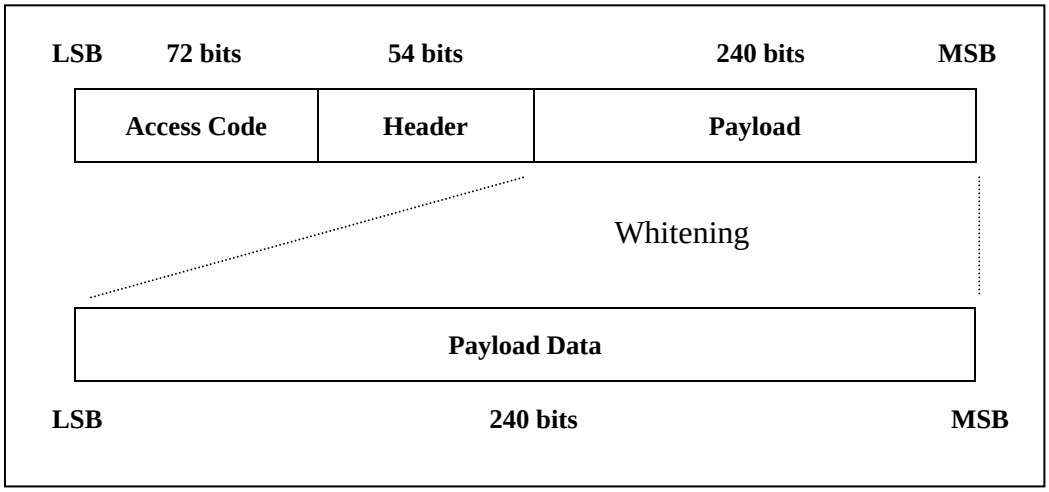


圖 2. HV3 Packet 的格式

三、藍芽無線語音資料庫之建立

本研究共使用三種語音資料庫，包括中華民國計算語言學學會授權使用之 TCC-300 麥克風語料庫[9]、由 TCC-300 轉製之藍芽實際使用環境語料庫及藍芽模擬使用環境語料庫，分別說明如下。

3.1 TCC-300 麥克風語音資料庫

TCC-300 為一麥克風朗讀語料庫，係由台灣大學、成功大學、交通大學三所學校各收集 100 位語者之語料集合而成，總數為 300 位，總音節數為 332,708，取樣頻率為 16 KHz，量化解析度為 16 bits。為符合藍芽規範，我們將取樣頻率由 16 KHz 降低到 8 KHz。

3.2 藍芽實際使用環境語音資料庫

我們利用圖 3 之錄音系統錄製藍芽實際使用環境之語料，其中藍芽封包為不具通道編碼(Channel Coding)之 HV3 封包。本系統應用 CSR 藍芽開發模組 Casira[10]之 BlueChat 程式建立 Audio Type SCO 連線以傳送語音資料。傳送端電腦內之 TCC-300 語料經由 USB 傳輸線傳至藍芽裝置，完成語音編碼程序後，再透過無線通道將訊號傳送至接收端。接收端的藍芽完成解碼動作後，利用 USB 傳輸線將語料送至電腦中錄製藍芽語料庫。

為達同步錄音之目的，我們利用網路 TCP/IP 通訊協定控制錄放音的動作。當系統開始錄音時，傳送端電腦的放音同步程式會透過網路傳送一個同步訊息至接收端電腦，當接收端電腦的錄音同步程式偵測到”錄音”訊息時，則開始錄製語料。語料錄製完畢後，錄音同步程式會回傳”錄音完畢”訊息至傳送端電腦，並結束錄音程序。

我們將 TCC-300 語料庫轉錄成室內使用環境 0 公尺、4 公尺以及走廊使用環境 50 公尺三個藍芽操作環境語料庫。實際錄製環境請參考附錄之圖 A1~A3。

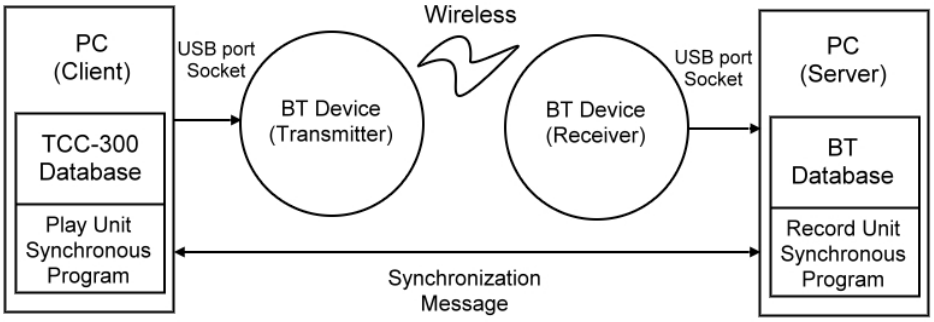


圖 3. 建立藍芽實際使用環境語音資料庫之錄音系統

3.3 藍芽模擬環境之語音資料庫

我們建立圖 4 所示之藍芽環境模擬系統。為符合藍芽規範，此模擬系統採用內插器(Interpolator)將語音訊號之取樣頻率從 8 KHz 提高至 64 KHz，接著使用連續變化斜率增量調變(Continuous Variable Slope Deltamodulation, CVSD）[11]進行語音編碼。調變器部份，我們使用藍芽規範所定義的高斯頻率移鍵(Gaussian Frequency Shift Keying, GFSK)[12, 13]技術。在通道部份，我們以萊斯衰減(Rician Fading)[14]模擬藍芽通道環境。由於藍芽工作於短距離低速環境，因此假設通道衰減為緩慢衰減(Slow Fading)[15]，各個封包經過通道時均會受到不同程度靜態衰減(Static Fading)的影響，換言之，同一個封包內的所有位元均受到相同的衰減效應。解碼端透過解調變、語音解碼、降低取樣率(Down-sampling)、以及低通濾波

器還原語音後，即完成藍芽模擬系統語音資料庫的建立。

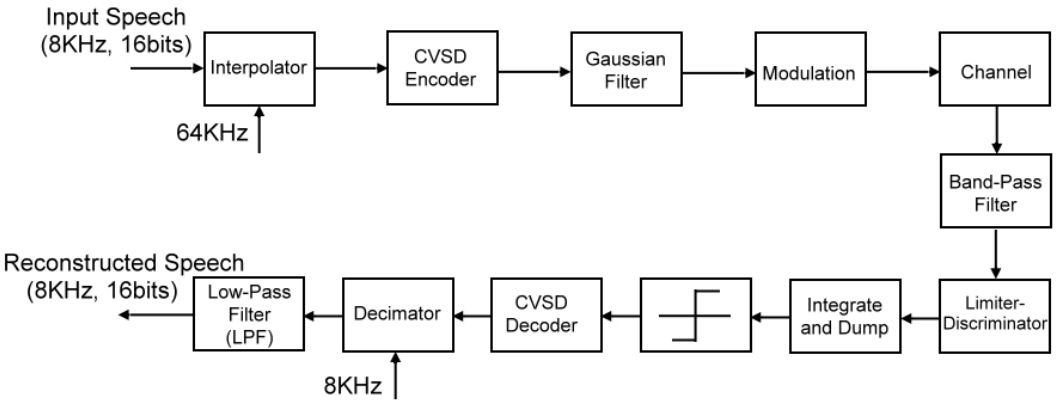


圖 4. 藍芽模擬系統

四、實驗結果及討論

4.1 基準實驗

本論文中，語音特徵參數採用 12 階梅爾頻率倒頻譜係數(Mel-scale Frequency Cepstrum Coefficients, MFCCs)、能量(Energy)及它們的一階和二階迴歸係數(Delta and Delta-delta Coefficients)，擷取特徵參數之相關變數設定如表 2 所示。

表 2. 擷取特徵參數之相關變數設定

音框長度（ms）	20
音框位移長度（ms）	10
Filter-Bank 階數	18
Cepstrum 階數	12
特徵向量維度	39
Window	Hamming

本論文採用已廣泛應用於語音相關研究的 HTK[16]軟體，進行由左至右(Left to Right)的隱藏式馬可夫模型(Hidden Markov Model, HMM)之訓練、辨識效能分析與測試。我們採用次音節模型(Sub-Syllable Model)作為語音聲學模型，其聲母模型的狀態數(state)為 2~3 個，韻母模型為 4 個狀態，辨識正確率之計算方式如下：

$$\text{音節辨識正確率} = \frac{N - D - S - I}{N} \times 100\%$$

其中 N 為所有辨識音節數，D 為刪除型錯誤，S 為替代型錯誤，I 為插入型錯誤。表 3 為模型訓練與測試相關變數之設定。TCC-300 共收集 300 位語者的語音，我們取其中 240 位語者之 21,844 句語料進行模型訓練，另外 30 位語者之 2,480 句語料進行測試。根據上述相關變數設定，我們得到之基準實驗的音節辨識率為 69.25%。

表 3. 模型訓練與測試相關變數之設定

聲學模型種類	Sub-Syllable Model
訓練模型人數	240 人
測試人數	30 人
所有狀態數	785
所有高斯混合數	9,056

4.2 藍芽實際使用環境下語音辨識效能之評估與分析

本節中我們在藍芽實際使用環境下評估其辨識效能，實際使用環境包括室內環境(0 公尺、4 公尺)與走廊環境(50 公尺)三種，其訊號傳輸路徑均為可視線(Line of Sight)，其中 0 公尺距離的定義如附錄圖 A3 所示。模型訓練與測試的條件與基準實驗完全相同，我們同時考慮環境匹配與不匹配之情況，其測試結果如表 4 所示。

表 4. 藍芽實際使用環境下之辨識效能

Training / Test	Baseline	BT = 0 m	BT = 4 m	BT = 50m
Baseline	**69.25%**	52.18%	50.55%	46.29%
BT = 0 m	45.59%	**55.82%**	54.07%	50.54%
BT = 4 m	44.46%	54.87%	**53.54%**	50.35%
BT = 50 m	22.93%	33.28%	38.71%	**42.74%**

(一) 根據表 4 結果，我們發現在室內短距離環境中(0、4 公尺)，其辨識率的差異不大，而在走廊環境(50 公尺)則因長距離的衰減及人群穿梭阻擋效應，造成辨識率的大幅下降。

(二) 由表 4 得知在匹配條件下，0、4、50 公尺的辨識率分別為 55.82%、53.54%、42.74%，4 公尺環境的模型對 0 公尺及 50 公尺的不匹配實驗，可發現其效能為 54.07%及 38.71%，分別趨近匹配環境下 55.82%、42.74%的辨識率，意味著短距離的模型(例如 4 公尺)可做為通用模型。

(三) 觀察表 4 第 3 欄，以 4 公尺之模型測試 0、4、50 公尺語料之辨識效能分別為 54.07%、53.54%、38.71%，可知 0 公尺之 54.07%優於 4 公尺之 53.54%及 50 公尺之 38.71%，此現象是由於 0 公尺測試語料受到通道效應及距離之影響較小，所得語料較為乾淨，因此其辨識率優於 4 公尺及 50 公尺測試語料所得之結果。同理，若以 50 公尺之模型測試 0、4、50 公尺語料，其辨識效能趨勢亦與 4 公尺模型所得結果一致。

(四) 使用訓練自一般麥克風語音的模型辨識藍芽實際使用環境的語音，辨識率從 69.25%下降至 0 公尺的 45.59%、4 公尺的 44.46%以及 50 公尺的 22.93%。因此，一般麥克風語音的辨識模型，並不能適用於藍芽無線環境下的語音辨識。

接著我們嘗試使用頻譜消減法(Spectral Subtraction, SS)[17]、倒頻譜平均值消去法(Cepstral Mean Subtraction, CMS)[18]及倒頻譜平均值正規法(Cepstral Mean Normalization, CMN)三種通道補償方法以及它們的混合模式來降低環境不匹配之效應，實驗結果如表 5 所示。由於 SS 係一種即時的背景雜

訊消除方法，而本論文不考慮背景雜訊的因素，因此除了單獨使用 SS 未見補償效果外，其餘方法之辨識率均有 5%以上的提升，其中又以 Baseline(SS+CMN)所獲得之辨識率 53.03%、51.91%、32.56%最佳，較未經補償方法(Baseline)之結果 45.59%、44.46%、22.93%分別有 7.44%、7.45%、9.63%的提升。另外，比較實際藍芽使用環境 4 公尺通用模型所得之辨識率可知，Baseline(SS+CMN)與 4 公尺模型所獲得結果分別有 0 公尺之 1.04% (54.07%-53.03%)、4 公尺之 1.63%(53.54%-51.91%)、50 公尺之 6.15%(38.71%-32.56%)的差距，證實補償方法可效地操作在藍芽短距離使用環境中。

表 5. 藍芽實際使用環境下之強健辨識效能

Training / Test	Baseline	Baseline (SS)	Baseline (CMS)	Baseline (CMN)	Baseline (SS+CMS)	Baseline (SS+CMN)
Baseline	69.25%	NA	69.55%	69.58%	NA	NA
BT = 0 m	**45.59%**	44.21%	52.53%	52.55%	52.87%	**53.03%**
BT = 4 m	**44.46%**	42.95%	51.39%	50.88%	51.18%	**51.91%**
BT = 50 m	**22.93%**	21.67%	30.28%	28.13%	30.21%	**32.56%**

4.3 藍芽模擬環境下語音辨識效能之評估與分析

本實驗使用的語料庫為依據圖 4 之模擬架構利用軟體轉檔的 TCC-300 語料庫(所有語料不經過 SCO 通道)。模型訓練與測試的條件與前面的實驗相同，也是採用相同的 240 位語者之 21,844 句語料進行模型訓練，30 位語者之 2,480 句語料進行測試。本實驗亦使用二種通道補償方法(以 CMS、CMN 表示)降低環境不匹配之效應。調變方法為 GFSK，以 Rician 模擬通道特性，BER 為 0.5%。

我們首先測試藍芽 CVSD 編碼對語音辨識的影響，實驗結果如表 6 所示，藍芽 CVSD 編碼在無通道效應下並不會造成辨識率的明顯下降。我們進一步將模擬藍芽實際使用環境的語音辨識模型對 0、4 及 50 公尺藍芽實際使用環境錄製之測試語音進行辨識，結果如表 7 所示，比較第 1 及 2 欄，我們發現光是使用藍芽 CVSD 編碼來模擬藍芽實際使用環境是不夠的，模擬模型對於藍芽實際使用環境錄製之測試語音的辨識率並未優於原來的麥克風語音模型。如表 7 第 3 欄所示，加上解碼端 LPF 後得到的模擬模型可以讓辨識率顯著提升，表示該模擬方式更接近藍芽實際使用環境，對於麥克風語音的測試辨識率自然因 mismatch 加大而下降。若我們進一步考慮調變(GFSK)及通道效應(Rician)，如表 7 第 4 欄所示，測試 0、4、50 公尺藍芽實際使用環境語料得到之辨識率分別為 50.64%、47.94%及 28.52%，較麥克風語音模型分別提升 5.05%(50.64%-45.59%)、3.48%(47.94%-44.46%)及 5.59% (28.52%-22.93%)。此外，比較藍芽實際使用環境 4 公尺通用模型所得之辨識率可知，模擬系統(CVSD+LPF+Rician+GFSK)與實際 4 公尺模型的辨識率差距分別為 0 公尺之 3.43% (54.07%-50.64%)、4 公尺之 5.6%(53.54%-47.94%)、50 公尺之 10.19%(38.71%-28.52%)的差距，初步證實此模擬系統具有模擬藍芽實際使用環境的實用價值。如表 7 第 5、6 欄所示，加上通道補償方法(如 CMS 及 CMN)後，辨識率可以再略微提升，但並未明顯改善。

表 6. 僅考慮 CVSD 因素之語音辨識效能

Training / Test	Baseline	藍芽 CVSD
Baseline	69.25%	67.9%
藍芽 CVSD	67.65%	68.03%

表 7. 藍芽模擬系統之辨識效能

Training / Test	Baseline	CVSD	CVSD+ LPF	CVSD+LPF+ Rician+GFSK	CVSD+LPF+ Rician+GFSK (CMS)	CVSD+LPF+ Rician+GFSK (CMN)
Baseline	69.25%	67.9%	60.17%	49.34%	58.99%	59.23%
BT = 0 m	**45.59%**	45.97%	50.15%	**50.64%**	**50.94%**	**51.07%**
BT = 4 m	**44.46%**	44.35%	49.07%	**47.94%**	**48.4%**	**49.17%**
BT = 50 m	**22.93%**	21.14%	26.22%	**28.52%**	**29.94%**	**29.88%**

五、結論與未來展望

目前藍芽無線環境下之語音辨識尚處萌芽階段，本研究首先在藍芽實際使用環境下應用 TCC-300 麥克風語料庫及 HTK 軟體，進行一系列語者無關(Speaker Independent)的語音辨識實驗。接著，我們依據藍芽規範建構一套模擬系統，以探索藍芽環境下影響語音辨識效能之因素，提供這方面研究一個辨識效能之參考標準。此外，為彌補通道效應之影響，我們亦引用若干強健性技術以提升辨識率。根據實驗的結果，我們有以下結論:

(一) 本研究錄製了三套藍芽實際使用環境語料庫，包含 0、4 公尺室內環境以及 50 公尺走廊環境之語料，這些語料庫可提供藍芽環境語音辨識相關研究使用。

(二) 本研究所建立模擬系統(CVSD+LPF+Rician+GFSK)之辨識率，與實際匹配環境下的差距分別有 0 公尺之 5.18% (55.82% - 50.64%)、4 公尺之 5.6%(53.54% - 47.94%)、以及 50 公尺之 14.22% (42.74% - 28.52%)，我們推測此差異可能與路徑損失(Path Loss)、瑞雷衰退(Rayleigh Fading)等因素所致之封包遺失(Packet Loss)有關，這點尚待進一步實驗釐清。

(三) 通道補償方法包括 CMS、CMN，可改善藍芽傳輸環境之辨識效能。

(四) 本研究探討大語彙連續語音音節辨識，即使是在實際匹配環境下的辨識率 55.82%(0 公尺)、53.54%(4 公尺)、以及 42.74%(50 公尺)，似乎也無法令人滿意，但在實際自動化應用上需用的語音控制指令數量並不是太多，若考慮以一般慣用之關鍵詞(Keyword)為辨識對象，辨識率將可大幅提升，例如[19]所開發之小量字彙(9 個關鍵詞)藍芽語音辨識系統，辨識率可達 90%以上，因此，我們認為藍芽語音辨識之應用是具有潛力的。

本研究未來可朝以下方向發展及改進：

(一) 加入路徑損失(Path Loss)、瑞雷衰退(Rayleigh Fading)等因素，以建立更精確的模擬系統，俾可取代實際系統以免除需於實際使用環境下錄製大量藍芽訓練語料之困擾，有效節省人力與時間。

(二) 開發強健性通道補償方法改善辨識效能。

(三) 加入錯誤更正碼，改善通道效應，或結合來源編碼及通道編碼以得更好的辨識效能。

(四) 由於藍芽適用於短距離室內通訊，因此可考慮進行車內環境中的語音辨識研究。

參考文獻

[1] 張照煌，語音辨識技術應用之發展趨勢，中央研究院計算中心通訊，第十四卷，第七期，1998。Available: http://www.ascc.net/nl/87/1407/04.txt

[2] http://www.szjtvoice.com/products/3.htm(捷通語音技術開發公司)

[3] http://www1.vghtpe.gov.tw/asrOpen/ASROpen_index.htm

[4] http://home-automation.org/

[5] C. Mokbel, *et al*., "Towards Improving ASR Robustness for PSN and GSM Telephone Applications," *Speech Communication*, pp. 141-159, 1997.

[6] J. M. Huerta and R. M. Stern, "Distortion-Class Modeling for Robust Speech Recognition under GSM RPE-LTP Coding," *Speech Communication*, pp. 213-225, 2001.

[7] Z. A. Bawab, I. Locher, J. Xue, and A. Alwan, "Speech Recognition over Bluetooth Wireless Channels," in *Proc. Eurospeech*, pp. 1233-1236, 2003.

[8] A. H. Nour-Eldin, H. Tolba, and D. O'Shaughnessy, "Automatic Recognition of Bluetooth Speech in 802.11 Interference and the Effectiveness of Insertion-Based Compensation Techniques," in *Proc. ICASSP*, pp. 1033-1036, 2004.

[9] http://rocling.iis.sinica.edu.tw/ROCLING/MAT/index_cf.htm

[10] CSR, Casira Quick Start Guide, CSR Inc., 2001.

[11] Bluetooth Special Interest Group. Available: http://www.bluetooth.com

[12] A. Soltanian and R. E. Van Dyck, "Physical Layer Performance for Coexistence of Bluetooth and IEEE 802.11b," in *Proc. Virginia Tech. Symposium on Wireless Personal Communications*, June 2001.

[13] R. Steele, *Mobile Radio Communications*, John Wiley & Sons Inc., 1996.

[14] J. Proakis, *Digital Communication*, New York: McGraw-Hill, 2001.

[15] A. Conti, D. Dardari, G.. Pasolini, and O. Andrisano, "Bluetooth and IEEE802.11 Coexistence: Analytical Performance Evaluation in Fading Channels," *IEEE Journal on Selected Areas in Communications*, pp. 259-269, 2003.

[16] S. Young, *et al.*, The HTK Book, Version 3.0, July 2000.

[17] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. on Acoustics Speech and Signal Processing*, pp. 113-120, 1980.

[18] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing*, 全華書局。

[19] 王新富，語音辨識技術於藍芽通訊環境之應用研究，碩士論文，國立台北科技大學電機工程系碩士班，台北，2004。

附錄

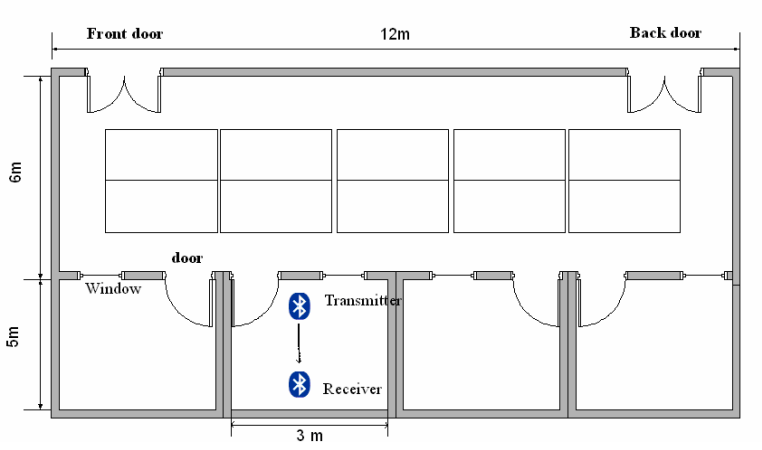


圖 A1. 綜合科館 416 實驗室空間配置圖(0~4 公尺)

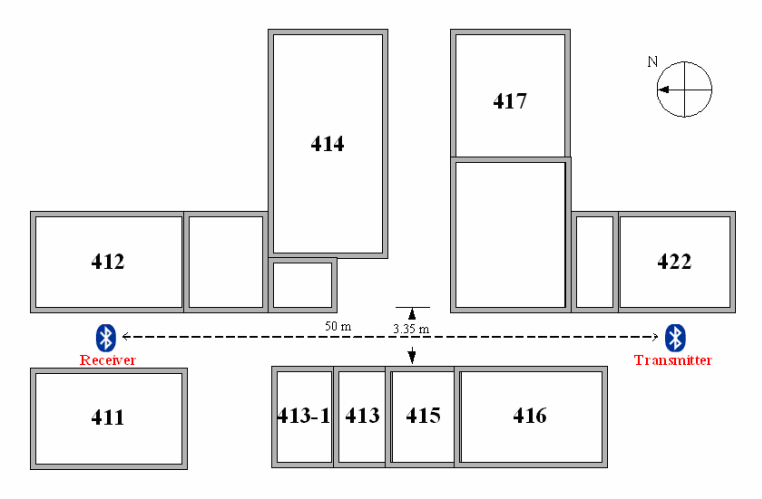


圖 A2. 綜合科館 4 樓空間配置圖(走廊環境 50 公尺)

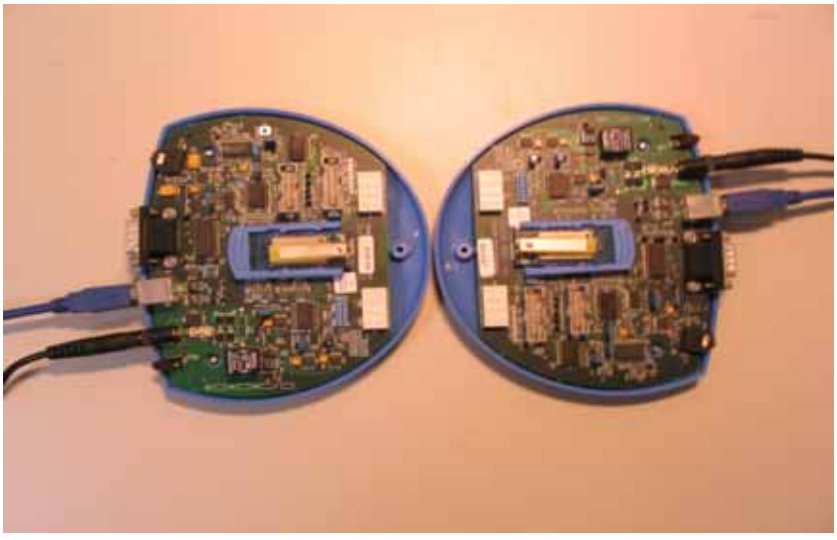


圖 A3. 0 公尺距離的定義

華台雙語發音變異性之語音辨識研究及 PDA 之應用

呂道誠 [1, 3]，謝鴻文 [1]，李勇憲 [2]，劉仲英 [1]，許鈞南 [3]，江永進 [4]，呂仁園 [2]

1．長庚大學電機工程研究所
2．長庚大學電資訊工程究所
3．中央研究院資訊所
4.清華大學統計研究所

E-mail: rylyu@mail.cgu.edu.tw, TEL:886-3-2218800ext5967
daucheng@iis.sinica.edu.tw, TEL:886-2-27883799ext2104

Abstract. 本篇論文提出一種方法來有效的處理華台雙語同時存在於同一句話的語音辨識問題。主要的核心可分為三部分；一.聲學模型：此部分是用一個共同的標音系統，使相同的發音的標音在不同語言上能夠做語料的分享，而且在語音特徵擷取上也加上聲調的參數，以減少華字與音節間的混淆。二.發音模型：此部分是結合了以專家知識為主的發音辭典與實際上語料分析結果而成變異發音，前者是統計了的華台雙語辭典的華字對音節發音機率，找出一個華字在辭典上所有可能的發音；而後者是將音節的辨識結果做成發音對華字的混淆機率。第三部份是將華字直接嵌入在語言模型中，作為搜尋的節點。之後用唐詩300首的實驗，其針對目前台灣地區華台夾雜的語句，以及發音變異性的問題，都能確實降低一成五到兩成的漢字相對錯誤率。最後將此技術移植到PDA上，也做了相關的應用。

1 簡介

華語是目前世界上使用人口最多的一種語言之一，數目超過十億[1]以上，最主要的分佈是在中國大陸和台灣，然而中國大陸的國土廣大、地理位置的阻隔、或時間的演變、人口的遷徙與外來語的影響，使得華語產生了許多的變化，在不同的省分人們所講的話雖然都是華語但之間會有些明顯的差異。如北京話、上海話、廣東話、四川話、閩南話等等，以[2]來說，這些話彼此的關係是介於語言與方言之間，因為一種語言是一個國家或一群種族所說的共同話語，而方言是屬於地方性的語言，彼此之間的差異性並不大，很容易理解，然而上述所說的那五種話都屬於華語的分支，但變化又比方言複雜；以數字一到十做為例子，以上五種話的發音各不相同，相互間難理解，因此我們把這些話統稱為省話(因為大部分都是以省分為稱呼)。雖然這些話所發的音不盡相同，但還好，這些省話都有個共同的書寫系統與發音特性，其系統就是"漢字"，發音是以音節為單位，而每個音節都能對應到一個漢字。所以在做語音辨識的人就是在處理"音"與"字"的問題。

語音辨識的目的就是要把人所說的音轉化成文字，因此要做廣義華語的語音辨識，就有點像做多國語言的辨識了。就如之前所說，華語其實包含了所多的省話，有統一的漢字系統，但各個發音都不盡相同，以之前在做多語語音辨識的研究來說，[3]是先將語言的種類辨識出來，然後再用那一種的辨識引擎來辨識那個發音是哪一個字；而[4]是一次用多種語言的辨識引擎來做辨識，看哪一種於言的哪一個字機率最大，另一種[5]是用一個單一的標音方式來將多種語言的聲學模型作結合，也是一次將特定語言的字給辨識出來。而目前在台灣大部分的語音辨識研究都是以華語[6]為主閩南語[7]次之，客語是幾乎沒有，所以本篇論文就是要處理華語和台語雙語的語音辨識研究。

華台雙語除了其字體同樣是漢字以外，聲調也是其一特色。以語音學來說，華語有五種聲調，台語有7種聲調。而以音節的單位來說，中文不帶聲調音節有約400個，而帶調音節約1300個；以一萬三千個中文常用字來說，平均每33個字會對應到一個不帶調音節；而每10個字會對應到一個帶調音節。所以如果在做語音辨識不把聲調的特性也考慮在內的話，會造成嚴重的錯誤，如"睡覺"與"水餃"的混淆。而台語的情況更是嚴重。在[8][9]也證明了華語和粵語加入聲調的特徵會使得整體辨識率上升。因此既然語音辨識就是要把音轉成文字，那麼聲調的問題一定要處理。

對於一個漢字除了因為有不同的省話而產生不同的發音之外，個人的發音習慣、地域不同或上下文連音的發聲耦合(co-articulation)也會影響到一個漢字的發音，這種情況我們稱之為發音的變異性(pronunciation variation)。最明顯的幾個例子為：在台灣大部分的人都常常把捲舌音發成不捲捲舌

音，或台語的入聲音發不出來，我想這都是受到本身母語的影響；因為如果說話者本身是以台語為母語，則台語在語音學中並沒有捲舌的音素，因此在每次遇到捲舌音時，就已相近的不捲舌音來替代。相反的，如果本身是以華語為母語的說話者，因為在他成長的過程中，並沒有受到入聲音的訓練，因此不能正確的發出台語的入聲音。所以在這一方面的情形，我們也要考慮在內， [10]利用決策樹與發現規則的方法來做改進；而[11]是將音節的混淆程度利用機率的方法表現，來做語者的調適；在[12]更是發現其實並不是所有的發音都會有相同的規則，而是在部分的情況下某些發音才會改變，這些文章說明了發音變異性的問題是值得注意且必須重視解決的，這個問題會在未來會越來越多，因為語言的發音一直都在轉變。

以下為本論文的章節安排。首先在第二節將介紹同時處理華台雙語語音辨識所遇到的困難點，並針對這些問題我們提出一個有效的方法來解決。而在第三、四節中我們詳細的介紹國台雙語聲學模型以及發音模型的作法。而後第五節透過實驗測試來驗證此方法的優點。再來第六節並將此技術移植到PDA手機上做了一些相關的實際應用，最後是總結與未來的展望。

2 問題定義與解決方案

我們都知道，做雙語或多語的語音辨識比做單一語言來的困難；難就在難於兩種語言本身並沒有統一的發音標記方法、語言本身發音的不同，連帶的所用的文字不同、語料分配不平均的問題、要花許多的時間在瞭解語言，如文法與結構上的差異性大等等。然而目前全球化的速度越來越快，國與國的邊界越來越模糊，也造成語言和語言之間的相互影響越來越密切，一個人因為環境的音素同時會說多種語言的情形也越來越普遍。在台灣也不例外，有將近70%[13]以上的人口在台灣會同時使用華台雙語來作為日常生活中的交談，電視上的連續劇也常常會出現華台夾雜的對白，因此由以上的情形看來，在台灣多語的語音辨識也變的日漸重要了。

然而如果要同時處理華台雙語夾雜的語音，一些相關的問題必須要解決。在[14]提出用樹狀的辭典搜尋法能將以音節[15]為基礎的中文連續語音辨識改進為以漢字為基礎，而且效果快速且能提高辨識率，因此我們採用其特性做為依據，將原本兩階段辨識先辨識音節在轉成漢字的方式，改為直接用一階段的方法來辨識；然而此方式要同時處理兩種語言的話會遇到以下的問題：1. 因為我們不能限制說話者什麼時候講華語或台語，因此在語言模型的設計上必須要用開放式的架構來接受所有可能得情況，這樣來說，不但會增加搜尋空間，而且也會造成空間上不必要的浪費。2.在辭典上如何的有效整合兩種語言的發音？或如何將台語本身的南腔北調問題或是語者本身發音變異性的問題都反映在內？因為我們也不可能把所有可能的發音詞句都納入到發音辭典中，因為這樣也會導致發音的混淆。3.在語言模型中如何整合華台雙語本身的文法或詞結構不同的問題？

在這裡我們提出的方法是用一階段搜尋方式來做華台雙語大詞彙的語音辨識，其是利用華台雙語都是對應到同一漢字書寫系統的特性來解決這個問題。當然這個方法也可以擴大到整個以漢字為書寫系統的語言上，如上海話、廣東話、四川話等等。不管說話者說的是華語還是閩南語或是上海話，語音辨識引擎都是將音轉成漢字，如<圖一>所示。最右邊的語言模型上用樹狀結構的漢字當作搜尋的節點，一來可以加速說尋的速度，二來在做華台雙語的辨識上在不用考慮到語言的問題，因為最後的輸出就是漢字，而不是音節。關鍵就在於<圖一>中間的發音模型；此模型記載了一個漢字所有可能的發音，不管是辭典上有的或是發音變異而產生的，都有其相對應的機率，因此這樣的架構下才能讓使用者在一句話中可以任意的說出華語或台語，而不會增加語言模型的負擔。而聲學模型的架構是不變的，只是在這裡有考慮到華台雙語統一標記與聲調的部分。底下我們就將聲學模型與發音模型做詳細的解說。

<圖一>. 三層次的語音辨識示意圖

3 聲學模型

在華台雙語的辨識中，聲學模型要有效的整合華語和台語的語料，以即要考慮到避免在發音模型中造成一個漢字與對應發音的混淆，因此我們提出了兩大方向來處理聲學模型：

3.1 福爾摩沙標音系統(Formosa Phonetic Alphabet)

由於聲學模型是透過語料所產生的，因此如何有效的利用語料使所產生的聲學模型更加強健是一個要考慮的問題。其二，本論文是做語音辨識，而不是語言辨識，不管語者是發台語"阿"的的音，還是華語"阿"的的音並不是重點，因此我們提出了一種標音系統能夠將目前在台灣主要的三種語言：華語、台語以及客語的發音都納入其中，稱為"福爾摩沙標音"系統[16]，簡稱ForPA，其有效的整合此三種語言的發音，一方面讓語料能夠充分的分享，另一方面也可讓音素的分佈更加均勻。因為用ForPA，在華語的音素有37個，而台語有56個，聯集共有63個，而交集的就有32個；相同的標音符號彼此分享語料，所以交集音素的部分在華台語裡面能過獲得雙倍的語料，就某種程度來說這是好的，因為相同的音素有更多的語料可以拿來訓練。在[17]證明了利用此標音方式將華台雙語所訓練出來的聲學模型在相同複雜度的語言模型下其辨識率比單一語言來的好。

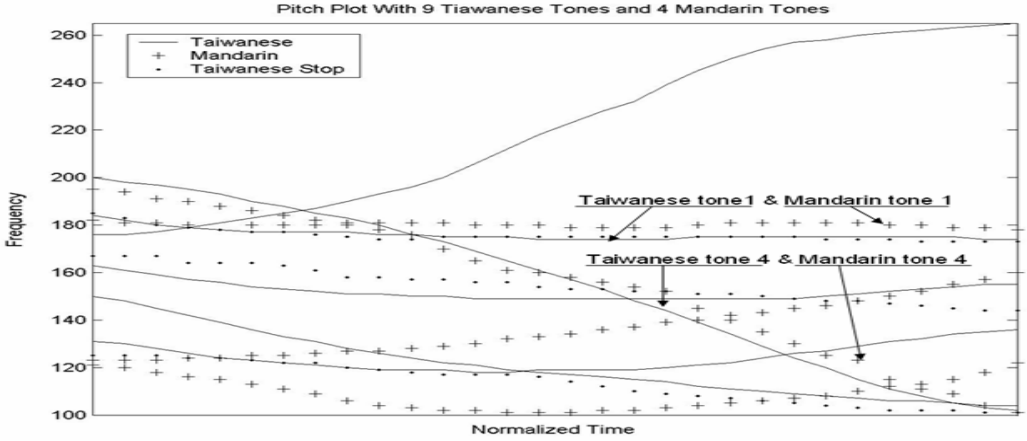
3.2 短時間聲調特徵擷取(short-time tonal feature vectors)

我們都知道，華語一個字的發音結構可拆成基本音結語與聲調，相同的基本音節結合不同的聲調其意義不相同，因此聲調也是一個特徵用來有效的區分字與字的分別。而一般目前在做語音辨識所用的特徵參數都是屬於短時間的梅爾倒頻係數(Mel-frequency cepstral coefficients)，因此為了要將聲調的係數結合梅爾倒頻係數，本論文所採用的聲調特徵取法是用短時間的正規化自動相關函數[18]所得到的。其研究顯示，大部分人的說話頻率介於65Hz到600Hz之間，為了使我們所取的聲調在一個音框中得到比較正確的頻率，通常在一個音框內希望能夠有三次以上的週期，所以在以16k取樣頻率的麥克風的語料在短時間上我們以40ms為一個聲調的音框。

由於聲調具有和諧的特性，為了避免所取的聲調數值變成原本頻率的倍數或一半，我們把自動相關函數所得到的頻率，在一個音框中最多設15個候選值，然後再利用動態規劃(Dynamic Programming)的方法將每一次的發音取得較佳的數值，這樣能夠有效的避免突然的雜訊，但缺點就是比較花時間。

另一個問題就是聲調只有出現在有週期訊號的音框中，如母音或韻母之類的部分，而子音或聲母的部分是沒有辦法得到聲調的數值。為了填補子音部分的聲調特徵參數，[19]提出了用指數函數的方式將聲調與聲調之間的空白連接起來，其中也做了五點平均，以平滑連接的部分，使聲調曲線看起來更加自然。

最後就是每個人的說話頻率不相同，訓練或是測試語料中有男生也有女生，為了使的聲調特徵參數沒有所謂的偏差(bias)，因此我們將其參數作了正規化(normalization)。由<圖二>可看出台語的一、二、三聲調的斜率幾乎相同，如果不作聲調的正規化，我們將不能容易的區分出不同人聲調的一、二、三分別。而正規化的方法有以動態時間為單位、每句話為單位、以男生女生類別為單位、或以每個人為單位，這些方法我們都有做實驗比較，結果是以每句話為單位的正規化效果最佳，因此我們將每句話求出一個最大值的聲調當作分母，而將一句話中每個音框的值都除以分母，而得到聲調特徵參數。



<圖二>. 華台雙語聲調示意圖 (由100人語料所統計而得聲調曲線)，可以看出台語聲調(實線)有三條線其斜率幾乎是相互平行的的。

4 雙語字典

在這個部分，將介紹連接聲學模型與語言模型之間的發音模型的作法。發音模型其實是一個漢字與音節的對應表，一邊是漢字，一邊是所有可能的音節，中間是個漢字可能發聲的機率，如<圖一>所示。我們的目的就是要找出一個漢字對華語和台語所有可能的發音，二來就是利用統計的方法來計算其相對的機率。依照華語發音的不同特性如：省話的不同或是腔調口音的影響，此發音模型將以辭典統計與實際語料發音的變異性這兩個方向來進行。

4.1 專家知識的方法(knowledge-base approach)

華語的字體是漢字；相同的漢字在不同的省話其發音是不完全相同的，而一種省話又有其他的方言存在，就是所謂的腔調或口音。以台灣的閩南話為例，就有分宜蘭腔、漳州腔、泉州腔、鹿港腔等等，這就是我們常說台語的南腔北調。此情形為一個漢字可以"合法"的對應到許多種的發音，在這裡所謂的"合法"是從各種辭典中找出一個漢字所有可能的發音；就以"長"這個字為例，其華語因為其意義的不同可發成/zhang3/、/chang2/的音，而台語可發為/dng5/、/dinnunn4/；此外台語本身因一個字在一個詞中的位置不同而產生的變調問題[20]也要考慮在內的話，"長"也可發成/dng1/ /dinnunn5/的音。所以光一個"長"字就有六種發音，這樣的發音就是由一些語言學專家所著的辭典統計得來的，而這種以辭典或專家知識為基礎的漢字發音我們稱為多種發音(multiple pronunciations)。

4.2 實際資料的方法(data-driven approach)

另外我們也考慮到一個漢字的發音，因個人的發音器官構造的問題、上下文發音的牽連、外來語或母語的影響，而造成原本要發的音變成以相似音來代替，最明顯的例子就是大部分的人在說"輕輕的"發音通常會以"親親的"發音來取代；原本/ng/的發音會變成/n/，我想大家都能體會。再來就是因說話速度的快慢，也會影響到發音。"這樣子"在說話速度快的時候常常會變成"降子"的發音，由三個發音變成兩個發音，這種情形稱之為發音的刪減。在[21]提出華語的發音是以音節為單位，因此大部分的發音變異性通常是取代，而較少刪減或插入的問題，因此在這裡只討論發音取代的問題。另一個現象就是目前台灣特有的華語發音，台灣國語。這種現象就是所謂的母語影響發音的問題，以ForPA來說，華語和台語的基本音素有交集的部分，但也有彼此特有的音素，因此如果從小就習慣以台語為說話的語言，則在發華語特有的音節時某些音就發不太出來，而以台語的相似音節代替，如"吃飯"就會發成"粗犯"，"阿扁"會變成"阿bieng3"，這就是由於台語本身沒有/ㄔ/與/ㄧㄢ/的聲母與韻母。而外來語的影響在台灣出現的比較少，這種其實也是母語影響的其中一個例子。

以上說了這些例子，也許是有規則，也許是沒有規則可循的，這些規則也是要靠觀察實際的語音整理統計而得的。但實際上我們又不可能用人工方式一句句的聽看看有沒有發音變化，因此在這裡我們用音節矩陣來統計出這些發音的變異性。方法是將評估的語料透過基礎的辨識引擎(baseline recognizer)，將帶聲調的音節辨識出來，之後再利用動態規劃(dynamic programming)的方法找出標準發音與辨識結果的發音做單音節文法(one-gram)的對位(alignment)。如此我們就可以得到以帶聲調音節為單位的相似音節矩陣，透過這個矩陣，再與原本的漢字作結合，我們就能得到實際上以語料庫為基礎的漢字發音變異機率。

有了多種發音(multiple pronunciations)與發音變異性(pronunciation variation)的漢字對音節的發音機率，將兩個機率用相同的權重將他們合併起來就成了本論文所用的發音模型。一方面可承接聲學模型，另一方面在語言模型中只要用漢字當搜尋的節點，就能處理以漢字為基礎的華語發音料，因為不管是什麼發音，都有個適當的發音機率對應到漢字，當然在這裡也要顧慮到一個漢字有太多的發音的時候，就會變成贅贅，反而造成辨識上的困擾，因此在實際上還是要做修剪，以達到最佳的辨識效果。

5 實驗

5.1實驗環境設定

5.1.1語料

本論文所用的華台雙語訓練與測試語料如<表一>所示。訓練語料語料共有100人(50男，50女)，共將近22.5個小時的16k取樣頻率、辦公室環境的麥克風語料。評估語料是為了產生發音模型的另外20人語料。而測試語料是為了測試發音模型的10人的華台雙語語料，這三種語料是相互獨立的。因為本論文是為了測試一句話中同時存在華台雙語情形，因此我們用了唐詩300首做為我們的劇本。每個人各念了100句以台語和以華語為主的詩句，其中有50句是字正腔圓且單一語言的，如<表一>中的MtestR，其代表了以華語為主的標準測試語料。另外50句是使用者以華語或台語為主，但一句話中可以穿插夾雜任何其他另一種語言，自然而然的說出來，如"芙蓉帳暖度春宵"可能會發成這樣的音/(fu2 long2(沒捲舌) zang4(沒捲舌) nuan3) (do3 cun2 siau1)/，前面說華語後面說台語，或者可以非標準的發音來說，如"蓉"與"帳"都沒有發捲舌音，其完全看說話者的喜好或平常講話習慣；如<表一>中的TtestS，其代表了以台語為主的口語化測試語料。

	訓練語料 100 人		評估語料 20 人		測試語料 10 人			
語料編號	Mtrain	Ttrain	Mevl	Tevl	MtestR	TtestR	MtestS	TtestS
人數	100		10	10	10			
句數	43078	46086	1000	1000	250	250	250	250
時間（小時）	11.3	11.2	0.28	0.28	0.14	0.14	0.13	0.14
每句平均音節數	2.7	1.9	2.5	2.6	5.9	5.9	5.9	5.9

<表一>. 實驗用的華台雙語訓練、評估與測試語料一覽表

5.1.2**聲學模型**

本實驗是以隱藏式馬可夫(Hidden Markov)模型為主的聲學模型中，每個聲學模型以聲母與右相關帶聲調韻母為單位，其狀態數目分別為3個與4個。相同標音的聲韻母分享彼此的語料。特徵參數共有42個，包含了12個梅爾倒頻係數，一個以對數為單位的正規化能量，再加上聲調的係數；之後再取一階與二階差分係數當作本實驗的語音特徵參數。所使用的工具為[22]

5.1.3**發音模型**

在發音模型中，首先針對多語發音的問題，我們從本實驗室的華台客福爾摩沙辭典中統計出一個漢字所有可能的發音機率，而建立了第一個發音辭典為K-Lexicon。之後再用基礎的華台雙語辨識引擎(辨識率：華語64%,台語也是64%)來辨識另外的10人的評估測試語料，建立了漢字發音變異性的混淆矩陣。再從中統計其發音變異機率，而建立了重實驗或語料中而得的發音變異辭典D-Lexicon。

而在語言模型中，我們就用以樹狀結構為基礎的詞彙搜尋網路，其詞彙量為3223，每個詞彙的節點以漢字為單位。因此不用為了語言的轉換而另外的設計語言模型，這樣在不增加語言模型複雜度的情形下，有效的解決多語發音與發音變異性的問題。

5.2 **實驗結果**

本次的實驗室為了測試發音模型在華台雙語中的重要程度，因此我們設計了一套以唐詩300首詞句的測試語料，用了兩套的發音模型，其一就是只用K-Lexicon的多語發音模型，其二是將K-Lexicon結合D-Lexicon的發音變異辭典，在辭典中每個漢字所對應的發音總和為1.0，因此以K-Lexicon來說，平均每個漢字有2.5個發音，而D-Lexicon來說，有1.3個發音，整合之後(K+D)-Lexicon發音辭典，平均有2.7個發音，這是因為我們將發音機率小於0.1以下的都刪除，之後平均的加回到剩下的發音上，以減少一個漢字因太多的發音而造成辨識上的混淆。

實驗的結果列於<表二>，針對唐詩300首的測試語料，我們有用了兩種發音模型，分別是K-Lexicon與(K+D)-Lexicon。整體而言(除了華語的標準發音MtestR)，漢字的錯誤率，使用第二種發音模型比第一種來的好。尤其是用在口語式的語料上(MtestS, TtestS)，有20.1%與15.1%相對錯誤下降率。而華語的標準發音語料為何其相對錯誤率反而是上升的呢？原因應該是因為其語料本身的發音就比較正確，沒有過多的發音變異性，以就是一個漢字就對應到一個標準華語發音，但在發音模型上反而有發音變異性的機率存在(一個漢字對應到2.7個發音)，這樣的發音機率會干擾到辨識，而造成漢字的辨識率不升反降。

語料編號	MtestR	TtestR	MtestS	TtestS
以 K-Lexicon 的 漢字錯誤率[%]	5.9	30.5	3.9	39.8
以(K+D)-Lexicon 的 4 漢字錯誤率[%]	6.1	28.2	3.1	30.7
相對的漢字錯誤減少率[%]	-3.4	11.2	20.1	15.1

<表二> 唐詩300首測試語料用於K-Lexicon 與 (K+D)-Lexicon 的漢字錯誤率

6 PDA的應用

從上一節的結果知道，這樣的方法確實能解決部分的問題，因此我們把這樣的技術移植到PDA上，再做實際上的應用，而我們也成功的將此方法移植到XDA II[23]與HP H5550[24]的機子上，底下是其應用與實際上所遇到的問題和解決的辦法。

6.1 應用

6.1.1語音搜尋mp3播放機：

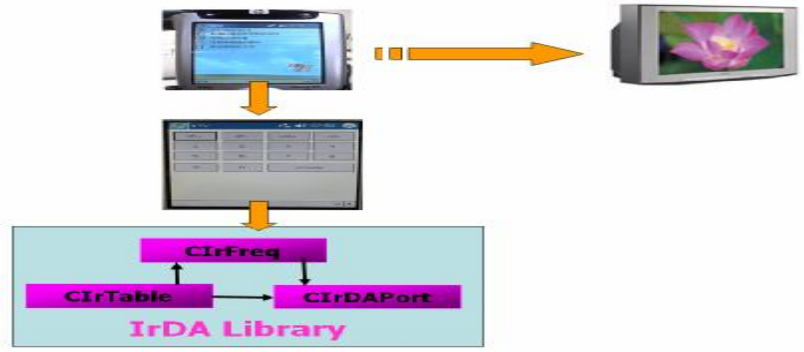
主要的功能就是讓使用者減少搜尋MP3的時間，畢竟有時自己喜愛的歌一多，要找起來的費時許多，在這分秒必爭的時代裏，這將帶給人們不少的便利，所以操作介面以簡易操作、方便使用為主，輸入一段語音聲波後，經多語辨識引擎後，直接以Window Media Player播放之。如<圖三>所示。

<圖三> 語音搜尋mp3播放機

6.1.2聲控資訊家電：

IrDA是通過紅外線進行數位信號交換的技術，IrDA數據傳輸技術被推薦使用在高速、短距離，點對點的無線數據傳輸場合，如：掌上電腦、數位相機等。自1994年以後，IrDA數據傳輸技術已經在超過30億電子的產品得到使用，包括PC、筆記型電腦、掌上電腦、印表機、數位照相機、行動電話、PDA等設備。

由於要控制紅外線與電視台名稱頻率對照表，以及各家廠商的紅外線頻率設定，我們以三個class來操作它們，以達到系統的強健性。如<圖四>所示。



<圖四> 聲控電視示意圖

由於各家廠牌紅外線規格不同，下面只舉了Sony廠牌的頻率規格，如<表三>所示。

Brand	Length	Type	HeadP	HeadS	1Pulse	1Space	0Pulse	0Space	Space
Sony	15	Pulse	2200	550	1100	550	550	550	23000
P.S：All numbers are time in us (micro seconds)									

<表三> Sony紅外線規格

6.1.3**手機人名撥號：**

手機人名播號其實主要是要減少使用者尋找的時間而且可以Hands Free與Eyes Free，若你在開車時，可以增加很多方便的地方。這個主要只是將連絡簿與名字跟PDA的連絡人做個連接，我們以兩個class來實作完成它。如<圖五>所示。

<圖五> 手機撥號示意圖

6.2 問題與解決方法

在整個將PC之辨識核心移植至PDA上，遇到許多程式方面的問題，條列如下：

問題1：WinCE並沒有支援在PC上的string.h這個檔案，我們在PDA上處理字串時，並沒有在PC上來得便利。

解決方法：我們採用傳統c語言中char的方式或是MFC提供的CString與TCHAR來處理我們的大量字串。

問題2：在我們處理漢字多音的對照表時，如果單純用字元指標去處理文字會有問題發生，常常在列印或顯示時，出現一堆莫名奇怪的記憶體暫存資料。

解決方法：使用字元陣列(給定大小值)並給它初始值，便解決這個難以查覺的問題。

問題3：在建立搜尋網路的地方，WinCE所支援之Standard Template Library(STL)的函式庫不是很齊全，所以讓我們在建立網路的過程中，部份演算法及函式都不能使用。

解決方法：儘量利用有提供之演算法與函式來達到我們的要求。

問題4：在PDA上，由於沒有支援ifstream.h、iostream.h、ofstream.h等檔案，在讀寫檔時也是非常不方便。

解決方法：我們用傳統c語音的FILE以及MFC提供之CFile來處理檔案讀寫的功能。

問題5：在聲音處理部份，雖然在PDA與PC上都是控制最底層的WaveInOpen、WaveInStart等函式，但由於PC版的是由Borlnad C++ Bbuilder 6.0所製作完成的，跟PDA的Embeded Visual C++ 4.0是不同，所以聲音處理上，我們也花了許多時間下去研究。

解決方法：將錄放音整個改寫成另一種新的版本，這種格式的錄放音對於移植性有更大的空間。

問題6：最後在做辨識部份的地方，由於資料結構太複雜，故我們盡量使用有STL支援的演算法，但前面有說過，演算法功能的不足，使得我們在做用上與操作上也花了很多心思在改變它處理的方式。

解決方法：改變其比大小與排序的運算子(operator)與函式指標(function pointer)，以達成我們想要的功能。

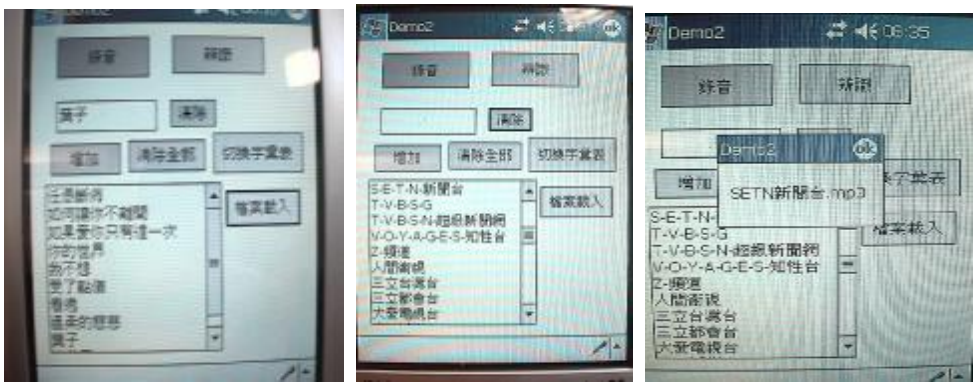
6.3 結果

PDA實驗結果列於<表四>，所用的聲學模型與PC上用的相同；而測試語料是用PDA所錄製的， 共有442句，分為歌名、電視台、人名與歌名四大類。而每一類用三種語言來做測試。如第二行的歌名類測試語料種共錄製了32句，其中有10句是用國語發音；20句是用台語發音；而2 句是用英語發音。每句的平均時間長度約在兩秒左右。而關鍵詞彙量就是所謂的語言模型詞彙量。

辨識率平均來說都有不錯的結果。在小於100句關鍵詞彙量的辨識結果都有九成以上，而在兩百句的詞彙量也有8成8。對於人名來說，在辨識上本來就比較難，所以辨識只到84%。辨識時間上，平均每秒的語料所辨識的時間也會因為詞彙量的變大而變慢，這是可以理解的。之後我們再仔細的分析辨識時間成分，大約平均有三分之一的時間花在語音檔轉特徵參數的轉換上，而剩下的三分之二時間花在計算機率與維特比搜尋上。

種類	測試語料 (國語,台語,英語)	平均每句 秒數	關鍵詞彙量	辨識率	平均每秒語音 辨識的秒數
歌名	32(10,20,2)	2.344	32	96.88%	16.31
電視台	79(30,25,24)	2.431	79	94.93%	23.57
人名	103(68,20,15)	1.569	103	84.47%	35.02
歌名	228(102,75,51)	2.087	228	88.60%	90.92

<表四> PDA上的辨識結果



<圖六> Mp3操作介面圖、電視台名稱載入以及辨識結果

7 結論與未來展望

本篇論文提出了一個在以華台雙語為基礎的語料，利用整合式發音模型，有效解決多語發音與發音變異性的問題。實驗也證明了此方法的確適合用於雙語或多語的語料，同時也能補償因發音不標準而產生的辨識錯誤問題。其二也將此技術成功的用於PAD上。相關應用也說明了此方法能夠很簡單的加入一些英文的詞彙，做有效的辨識，而不必再重新訓練英文語料。

對於未來的目標我們很希望能夠將個別的單詞句辨識強化成連續多詞彙的辨識，這樣才能真正的利用此技術用於實際上應用。第二對於發音模型的研究，未來希望能找出更好的方法來統計一個漢字的適當發音數目，以瞭解發音的個數與辨識率之間的關係，因為我想太多不必要的發音或單一發音，對辨識率都不好。第三就是對於PAD上的研究，希望能夠更徹底一點，已解決目前辨識數度慢的問題。目前我們的語音長度為3.5秒，從一開始錄音進去，經過整個辨識核心後，平均辨識率約為40秒，速度慢雖是它的缺點，但是辨識結果幾近八、九成是正確的。在速度上，也許硬體會加快或記憶體加大，這都有助於解決速度上方面的缺失。至於為何會那麼慢？跟float point 及fixed point運算這也是關係很大。

參考

[1] http://www.sinica.edu.tw/ioe/staff/c9-1-28.htm

[2] W. H. Tsai, “Automatic Identification and Indexing of Chinese Multilingual Spoken Messages,” Ph.D. Dissertation, department of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan, ROC, 2001.

[3] P. Dalsgaard, O. Andersen, H. Hesselager, B. Petek "Language-Identification using Language-Dependent Phonemes and Language-Independent Speech Units", in Proceedings of the International Conference on Spoken Language Processing, Philadelphia, USA, October 1996.

[4] A Study of Multilingual Speech Recognition, F. Weng, H. Bratt, L. Neumeyer, and A. Stolcke, SRI International, 1997

[5] Waibel, A. (2000) Multilinguality in Speech and Spoken Language Systems. Geutner, P.; Tomokiyo, L.M.; Schultz, T.; Woszczyna, M. Proceedings of the IEEE: Special Issue on Spoken Language Processing, Vol.: 88 Issue: 8, pp. 1297 -1313

[6] Bo-ren Bai, Berlin Chen, Hsin-min Wang, Lee-feng Chien, and Lin-shan Lee, "Large-Vocabulary Chinese Text/Speech Information Retrieval Using Mandarin Speech Queries," in Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP98), Singapore, Dec. 1998, pp. 284-289.

[7] Ren-yuan Lyu. Chi-yu Chen, Yuang-chin Chiang and Min-shung Liang (2000) Bi-lingual Mandarin/Taiwanese(Min-nan), Large Vocabulary, Continuous Speech Recognition System Based on the Yong-yong Phonetic Alphabet., ICSLP2000,Oct. 2000, Beijing, China

[8] Hank, Huang C.H., Frank Seide ''Pitch Tracking and Tone Features for Mandarin Speech Recognition''. In Proc. ICASSP, 2000

[9] Tan Lee, Wai Lau, Y. W. Wong and P.C. Ching, "Using tone Information In Cantonese Continuous Speech Recognition," ACM Transactions on Asian Language Information Processing, Vol. 1, pp. 83 - 102, 2002

[10]Mirjam Wester, "Pronunciation Modeling for ASR-knowledge-based and Data-driven Methods," Journal of Computer Speech and Language 17(2003), pp. 69-85, 2003

[11]Lee, Kyung-Tak / Melnar, Lynette / Talley, Jim (2002): "Symbolic speaker adaptation for pronunciation modeling", In PMLA-2002, 24-29.

[12]Liu, Yi and Pascale Fung, “Partial change accent models for accented Mandarin speech recognition.” In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, U.S. Virgin Islands, December, 2003.

[13]http://teach.erdc.cyc.edu.tw/~chinese/newfile9.html

[14]Berlin Chen, Hsin-min Wang, and Lin-shan Lee, "Improved Spoken Document Retrieval by Exploring Extra Acoustic and Linguistic Cues," the 7th European Conference on Speech Communication and Technology (Eurospeech 2001), Demark, September 2001.

[15]Hsin-min Wang, "Experiments in Syllable-based Retrieval of Broadcast News Speech in Mandarin Chinese," Speech Communication, 32(1-2), pp. 49-60, Sept. 2000.

[16]Liang M.S., R.Y. Lyu, Y.C. Chiang "An efficient algorithm to select phonetically balanced scripts for constructing corpus" NLP-KE, Beijing 2003

[17]Dau-Cheng Lyu, Bo-hou Yang, Min-Siong Liang, Ren-Yuan Lyu, Chun-Nan Hsu "Speaker Independent Acoustic Modeling for Large Vocabulary Bi-lingual Taiwanese/Mandarin Continuous Speech Recognition", SST, Melbourne, 2002

[18]Paul Boersma ''Accurate Short-Term analysis of the Fundamental Frequency and the Harmonics-To-Noise Rate of A sampled Sound'', 1993.

[19]Dau-Cheng Lyu, et al, "Large Vocabulary Taiwanese (Min-nan) Speech Recognition Using Tone Features and Statistical Pronunciation Modeling" In Proc. EuroSpeech, Switzerland, 2003.

[20]R. Y. Lyu, Z. H. Fu, Y. C. Chiang, H. M. Liu "A Taiwanese(Min-nan) Text-to-Speech(TTS) system Based on Automatically Generated Synthetic Units", the 6th International Conference on Spoken Language Processing (ICSLP2000), Oct. 2000, Beijing, China

[21]Mingkuan Liu, Bo Xu, Taiyi Huang, Yonggang Deng,Chengrong Li, "Mandrain Accent Adaptation Based on Contest-Independent/Context-Dependent Pronunciation Modeling," In Proc. ICASSP,

2000
[22] Steve Yang et al. Hidden Markov Model Toolkit V3.1, Cambridge University Engineering Department, 2002
[23]http://www.my-xda.com/
[24]http://www.search4hardware.com/10/p_10_246_HP_h5550.html

以語音辨識與評分輔助口說英文學習

[1]陳江村 [1]羅瑞麟 [1]張智星 [1,2]李俊仁

[1]國立清華大學 資訊工程系
新竹市光復路二段 101 號

[2]中華電信研究所
桃園縣楊梅鎮民族路 5 段 551 巷 12 號

E-mail : { jtchen, roro, jang}@wayne.cs.nthu.edu.tw, cjlee@cht.com.tw

摘要：

本論文提出利用音訊處理和語音辨識的技術，進行英文語音評分。依英文發音的特有性，進行評分系統的各個模組之設計、製作及實驗，期許建立一套合理的英文語音評分系統。

本論文結合三項技術 — 「說話驗證」、「語音訊號切割」和「英文語音評分」達成輔助口說英文之學習。「說話驗證」利用說話驗證的可信度評估，依此拒絕文句內容(context)不正確的評分語句。「語音訊號切割」提供一個將語音訊號切割出每個音素時間區段的方法，以預先訓練好的英文發音聲學模型當作切割依據，爾後經由語音辨認技術，以合適的聲學模型切割出正確的發音區段。

「英文語音評分」為評分系統的核心，使用的評分方式是比較標準語音和評分語音的相似度。本文採用四個評分參數 — 音量強度曲線、基頻軌跡曲線、發聲急緩變化及 HMM 對數機率差異進行語音相似度評分。經由實驗，對於一個合理的評分系統，我們得到音量強度曲線的權重為 7.45%，基頻軌跡曲線的權重為 22.40%，發聲急緩變化的權重為 17.24%，HMM 對數機率差異的權重為 52.91%，經由實驗證實本系統之語音評分與人工評分具有約 60%的正相關性。

關鍵詞:

語音辨識、語音評分、HMM、聲調辨識、Viterbi Search、音量強度、音高向量、梅爾式刻度倒頻譜、Forced Alignment、Downhill Simplex Search

1 前言

由於近年來電腦計算能力的提昇以及語音辨識技術的進步，語音處理在我們日常生活上的應用與日俱增，如語音辨識、語音合成、語者識別等等。其中，在跨國界的語言學習中，以電腦輔助使用者進行非母語學習(CALL, Computer-Assisted Language Learning)已受到相當重視，各方也紛紛投入相關的研究[10][11][18][15][20]。

電腦輔助發音訓練(CAPT, Computer-Assisted Pronunciation Training)可視為是語音辨識和圖形比對(Pattern Matching)兩項技術的結合。本論文研究主題，包含「說話驗證」、「語音訊號切割」以及「英文語音評分」三個部份，希望融合目前語音辨識和圖形比對的技術，對使用者進行公正的語音評分。

在語音評分系統中，如果能先濾除內容和標準語音完全不同的評分語音，可以使整個語音評分系統更具公信力。本論文運用了可信度評估的技術來達成說話驗證(Utterance Verification)。確保了評分語音內容的正確性後，對於評分語音我們使用 HMM(Hidden Markov Model)切割出每個音素(phoneme)的時間區段，使用高辨識率的 HMM 聲學模型可確保切割出來的音素區段有一定的可信度及正確率。在英文語音評分部份，我們利用標準語音資料來進行一種較為主觀的評分方式，主要使用圖樣比對(Pattern Matching)的方法，根據四個評分參數：音量強度曲線(Magnitude)、基頻軌跡曲線(Pitch Contour)、發聲急緩變化(Rhythm)以及 HMM 對數機率差異(HMM Log-Likelihood)，將評分語音和標準語音的資料逐音素地來做比較，以期找出評分語音和標準語音的差異程度。

2 相關研究

1997 年時，C. Cucchiarini、H. Strik 及 L. Boves 以荷蘭語為主，定義了 Total Duration of Speech no/plus Pause、Mean Segment Duration、Rate of Speech 以及 Global Log-Likelihood，經由類似的實驗後得出 Global Log-Likelihood 對於人類主觀評分占較重的比重[19]。1999 年 L. Neumeyer、H. Franco、V. Digalakis 和 M. Weintraub 以法語語料庫進行實驗，採用 HMM Log-Likelihood、Normalized Acoustic、Segment classification、Segment Duration、Timing 當作其實驗的評分參數，經由實驗後得出了 Normalized Acoustic 在評分系統和語言專家給予的分數中，其相關性高於 Segment Duration[10]。

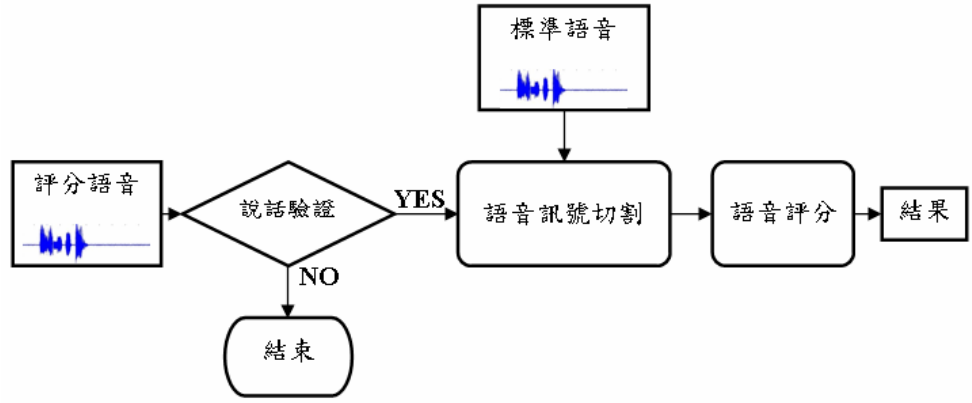
至於英文的語音評分，2002 年清華大學的李俊毅以梅爾倒頻譜、Magnitude 及 Pitch 三種評分參數觀察對英文語音評分的影響，其實驗發現梅爾倒頻譜參數對英文語音評分的重要性最大，另外他也將各個特徵的差異程度轉換成分數，以回饋給使用者參考[15]。2004 年陳江村和張智星等人利用了 HMM 和 GMM 分別對中文的發音和聲調進行評分，並以 Downhill Simplex Search 進行了評分系統參數的最佳化，以求達到和中文專家一致

的評分標準[20]。

接下來的論述中，首先我們提出實作「說話驗證」的方法，包含聲學模型相似度排名、驗證系統的建立及驗證系統的可靠性等。接著是「語音訊號切割」。這部份包含隱藏式馬可夫模型(Hidden Markov Model)的訓練和以維特比演算法(Viterbi algorithm)為基礎的語音訊號切割技巧。再來是「英文語音評分」，其中提到了關於評分參數的擷取、評分參數正規化、圖樣比對流程、評分機制的建立等，並設計實驗以求出各評分參數在英文語音評分中的權重，以符合人類專家對英文語句好壞的看法。最後是總結及今後研究工作的展望。

3 英語評分系統架構

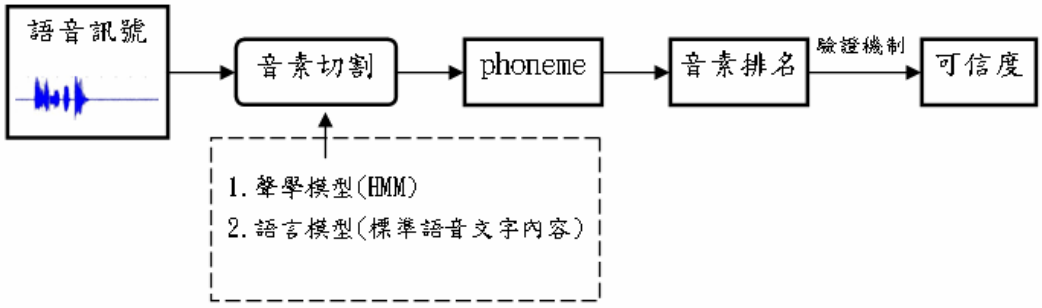
在此英文語音評分系統中，首先以說話驗證做為第一道檢視關卡，爾後以聲學模型來對標準語音及評分語音切割音素的時間區段，再將這些資訊送至英文語音評分系統的核心，利用各種評分參數，逐音素地比較評分語音和標準語音的差異程度。本文所提之英文語音評分系統架構流程，如圖表 1所示。



圖表 1 英文語音評分系統流程圖

3.1 說話驗証

所謂的說話驗證(Utterance Verification)，就是我們可以針對不同的評分語音產生判斷數值，並依此而對該評分語音內容的正確性做出判斷[1]。此說話驗證流程如圖表 2 所示，當驗證系統接收到語音訊號後，分別對每個音素進行語音辨識，之後再依辨識結果的機率值排名並配合驗證機制給予最後的可信度值。



圖表 2 說話驗證系統流程圖

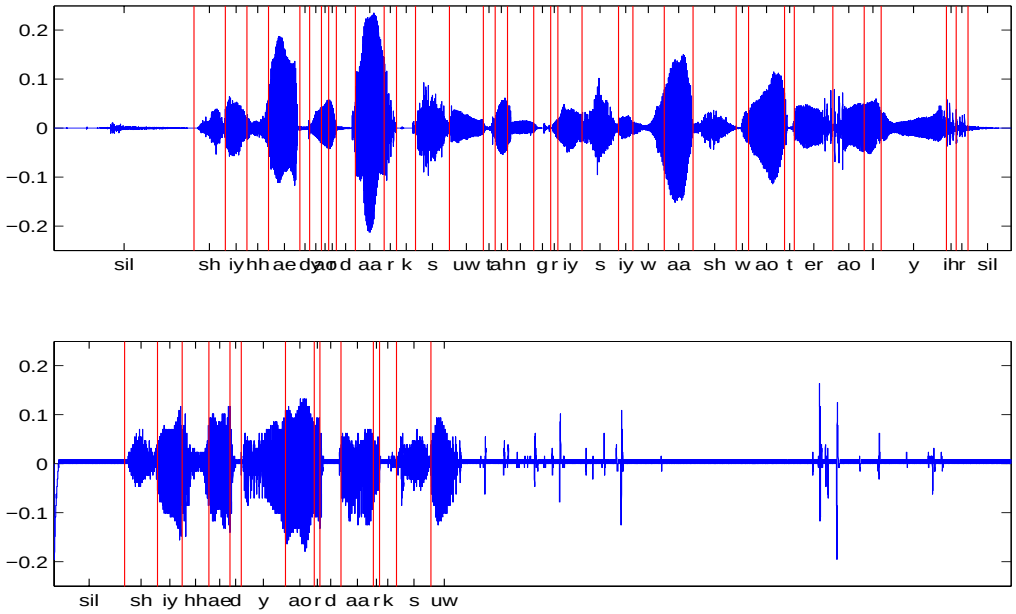
3.1.1 音素切割

這裡切割用的技術，並不是用 Viterbi Decoding 中常見的 Forced Alignment，而是使用 beam search 中 pruning 的方式，將語音盡可能地依序切割出每一個音素。在這種情況下，評分語音切割後，如果原來的內容和標準語音相當類似，則經由切割後產生音素的數量將接近甚至等同於標準語音音素的數量。相反地，若亂講的評分語音中只有前 n 個音素和標準語音相同(後幾個音素完全不同)，則經由 pruning 後的音素也大約等於 n。舉例來說，如果標準語音為「she has your dark suit in greasy wash water all year」、評分語音為「she has your dark suit」，則經由語音辨識後，在評分語音中所能切割出來的音素數量是 15，如圖表 3。

對於沒有切割出來的音素，我們則將其可信度值設為 0，如此一來可以增加驗證系統的區別性，使得和標準語音內容完全不相同的評分語音，其可信度值變得相當低。

圖表 3 為兩個語音經由語音訊號切割後產生的不同結果。上半部的語音內容等同於標準語音內容，因此

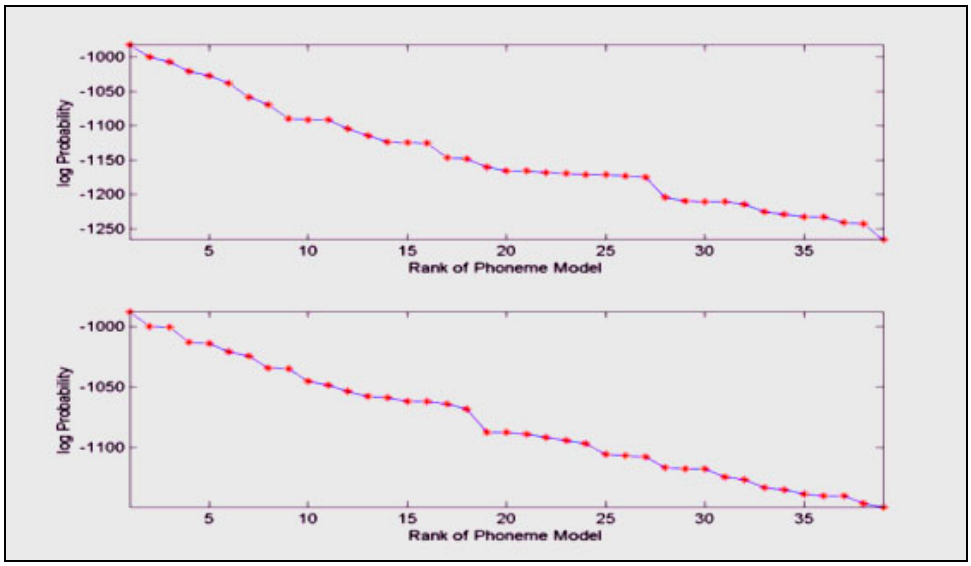
切割出來的音素很完整，而下方的語音內容和標準語音內容不盡相同，因而辨識程式將樹狀網路展開至節點 uw，就無法再繼續。圖表 3之音素符號是採用 CMU Phone Set 表示法[21]。



圖表 3　說話驗證的語音訊號切割比較圖

3.1.2 音素排名

切割語音訊號得到音素時間區段後，首先以每個音素對 39 個 phone models 計算對數機率[21]，並以排名的順序得到相對應的可信度值。機率排名的示意圖如圖表 4：



圖表 4　　音素機率排名

上下兩個機率分佈表示不同的音素經由辨識程式求得 39 個對數機率的結果，由圖表 4可以看出，對於不同的音素，即使排名同樣是第二名，可是和第一名的對數機率差距卻不相同，會造成這樣的原因在於有些音素的發音相似，而有些音素的發音差異則相當大[16]，因此我們對於上方圖中的音素，可解釋成其第一名和第二名 phone model 的發音很接近，造成對數機率的差距相當小。 而在下方圖中的音素，也許在我們 39 個 models 中，只有一個 model 的發音和該音素接近，因此更加突顯了其第一和第二名的對數機率差距。

3.1.3 驗證機制

經由語音訊號切割之後，產生的結果可能有兩種情況：一種是部份的語音訊號已經成功切割出時間區段的音

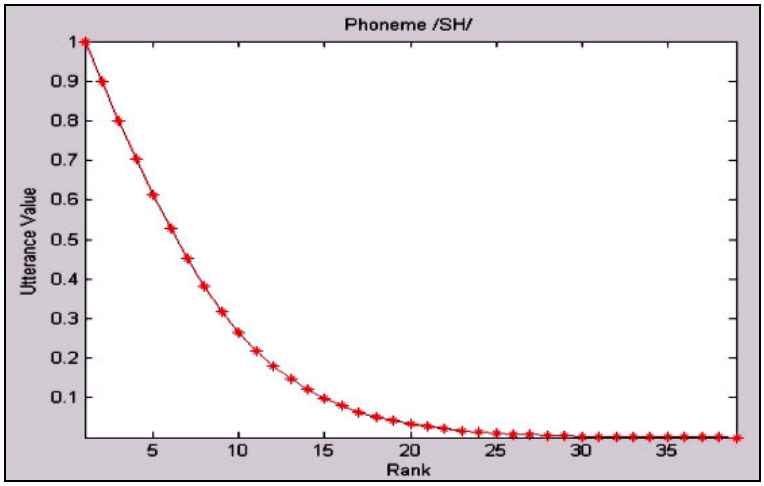
素，另一種則是語音訊號的後半部可能沒有辦法切割出音素。而在這一節討論的驗證機制，主要是針對前者的情況，也就是如何將音素的排名正規化，得到一個合理的數值。

在 Sukkar 和 Lee 於 1996 年發表的論文[17]中提到，音素的對數機率以及對所有音素的對數機率排名，和驗證系統的可信度值是有很大影響的。基於以上的前提，我們將 Sukkar 和 Lee 所提出求取可信度值的式子改寫並以下列公式表示：

$$value_{pho} = \frac{2}{1+\exp\left(\alpha \cdot \left(Rank_{pho}-1\right) \cdot \dfrac{\log P_{Rank_{pho}}}{\log P_{Rank_1}}\right)}$$

$\exp(x)$ 表示 e^x，即自然對數的 e 的 x 次方。$Rank_{pho}$ 和 $\log P_{Rank_{pho}}$ 分別表示該音素在 39 個 models 中的排名及對數機率值，1 表示第一名，α 為我們調整的參數值。由此公式可得知，當某音素相對於 39 個 models 的排名為第一名時，該音素的可信度值為 1。

圖表 5 表示對於「SH」這個音素之語音區段藉由上述的公式可將其對應於 39 個 models 所產生的對數機率及名次換算成可信度值。從圖中可以看出，當名次在第 10 名左右時，可信度值已經降至 0.2 了。



圖表 5　　音素 **SH** 的排名與可信度值的關係

另外由於音素間發音的差異性，因此我們在評斷可信度值時，不能單純地以排名來做比較。舉例來說，音素「OW」〔*o*〕和「S」〔*s*〕比對完 39 個 models 後同樣都得到第二名的結果，但是對於「OW」而言，其第一名是「AO」〔ɔ〕，而「S」音素的第一名是「T」〔*t*〕，則我們可以很明顯地看出「OW」和第一名的對數機率差距較小，也因此可信度值應該要比較高才合理。因此在上述公式中，我們將排名的差異再乘上對數機率的比例差異，如此一來就會使得每個音素的可信度值受到排名及對數機率的影響。最後經由計算得到的可信度值介於 0 和 1 之間。

當計算出句子所有成功切割的音素可信度值之後，利用每個音素的時間長度占句子時間長度的百分比作為權重，即可推導得出一句語音訊號的可信度值。以下是設定的公式：

$value_{sen} = 100 \cdot \sum_{n=1}^{N} \dfrac{len(pho_n)}{len(sentence)} \cdot value_{pho_n}$，$N$ 為一單字中評分音素的數量，$len(x)$ 表示 x 的時間長度。

至於有些單字可能其中的一些音素沒有辦法經由語音訊號切割產生，對於這些音素，我們就直接將其 $value_{pho}$ 設為 0。最後乘上常數 100 代表我們將說話驗證系統的結果定義在 0 至 100 之間。

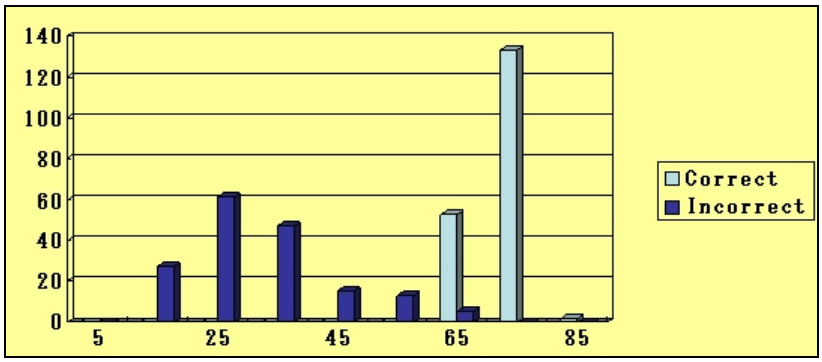
3.1.4 說話驗證實驗結果

對於在實驗中求出的門檻值(threshold)而言，如果語音訊號得到的可信度值高於門檻值，則我們稱「此句語音訊號的內容和標準語音訊號的內容相同」這句話是相當可靠的，也就表示我們可以放心地針對這句語音訊號進行評分。相反的，則表示這句話和標準語音的內容不相同，因此我們也就停止讓兩句不相同的語音進行後續的評分動作。對此我們蒐集兩部份的實驗語料：

1. Correct： 取 168 句說話內容相同的語音訊號當作標準語音內容，這部份語音檔案長度總和約為 9 分 10 秒。
2. Incorrect：取 168 句內容不等於標準語音內容的語料，這部份語音檔案長度總和約為 7 分 31 秒。其中一部份內容和標準語音完全不相同。另一部份則是語音訊號內容「部份相同」於標準語音內容。在此我們定義一句

話中若存在連續 3 個單字以上和標準語音內容相同，但並不是完全相同，即為「部份相同」。

實驗用的語料其音訊格式皆為 PCM，音訊取樣頻率為 16 kHz，位元解析度為 16 bits，所有的實驗語料皆為單聲道。接著將上述兩部份各 168 句的實驗語料經由說話驗證系統得到對應的可信度值，而後再統計、分析這些可信度值即求得驗證系統的門檻值。圖表 6 為求取門檻值的實驗結果分佈圖，橫軸為可信度值的範圍，縱軸為可信度值處於該範圍內的語音訊號個數。



圖表 6　說話驗證求取門檻值實驗結果分佈情況

我們以「型別 I 錯誤率(Type I error, False Reject)加上型別 II 錯誤率(Type II error, False Accept)為最小」作為尋找門檻值的前提。根據實驗結果，我們發現 Correct 中的語料其最小可信度值為 63.21，而在 Incorrect 可信度值大於 60 的語料中最接近 63.21 的可信度值為 61.59，因此我們將說話驗證系統的門檻值設定成 62.40(即兩者的平均)，如此可達到型別 I 錯誤率為 0%，型別 II 錯誤率為為 1.19%。

經由上述實驗計算求出門檻值後，我們另外準備一組內含 Correct 及 Incorrect 各為 168 句的測試語料，其中 Correct 語料的語料長度總和約為 7 分 27 秒，Incorrect 語料的語料長度總和約為 8 分 57 秒。將這些語料以門檻值為 62.40 的實驗結果，其型別 I 錯誤率為 7.14%，型別 II 錯誤率為為 0.60%。

3.2 語音訊號切割

「語音訊號切割」模組的功能乃是將標準語料及評分語料切割出音素發音的區段。其作法是以預先訓練好的英文發音聲學模型，切割出語料中之正確的音素發音區段。以下章節將分成「聲學模型的訓練」和「利用語音辨識來進行語音訊號切割」這兩部份來介紹。

3.2.1 聲學模型 HMM 的語料

實作語音訊號切割之前，我們必須先產生聲學模型，才能針對各種不同的語音進行切割動作。本論文中我們設計了兩種不同的聲學模型：一個是臺灣人口音的聲學模型，一個是外國人標準語音的聲學模型。

首先針對母語為英文的聲學模型，我們使用 TIMIT 語料來加以訓練。語料內容為 2,342 句平衡語料，由 438 位男性、192 位女性，共 630 人錄製，每人分配錄製 10 句，故共有 6,300 句語音。依 TIMIT 的建議取其中 4,620 句、語料長度總和約為 3 小時 49 分 10 秒的語音訊號作為母語為英文的聲學模型訓練，另外 1,680 句、語料長度總和約為 1 小時 23 分 51 秒的語音，則作為外在測試檔(Outside Test)。

另一方面針對母語為國語的聲學模型，我們請 33 位學生，其中包含了 23 位男性、10 位女性，依 TIMIT 的資料錄製 7,026 句平衡語料，我們取其中的 4,684 句、語料長度總和約為 4 小時 11 分 3 秒的語音作為母語為中文的聲學模型訓練，而另外的 2,342 句、語料長度總和約為 1 小時 57 分 43 秒的語音作為外在測試檔。
上述語料的音訊格式皆為 PCM，取樣頻率為 16 kHz，位元解析度為 16 bits。

3.2.2 聲學模型設計

英文中每一個音節可能由一個或數個音標所組成，而每一個音標都會對應到一個音素，而聲調、重音和破音(multiple pronunciation)的問題，在目前的聲學模型設計中則暫時忽略。TIMIT 的字典有 62 個音素，由於華人對於一些音素不像外國人念得那麼準確，再加上訓練語料不足下，如果我們減少訓練 model 的個數，則可使每個 model 的訓練語料取樣數目增多。鑑於上述兩個原因，我們將原先 TIMIT 設計的 62 個音素刪減成 40 個音素(含靜音 SIL 音素)。在本章中我們使用的聲學模型和音素是一對一對應的。舉例來說，“school”這個單字，其 KK 音標為〔*skul*〕，以我們設計的聲學模型來說，就是「S」＋「K」＋「UW」＋「L」。表格 1 是我們所設計的 40 個聲學模型與 KK 音標對照表：

表格 1　40 個聲學模型與 KK 音標對照表

模型	音標	模型	音標	模型	音標	模型	音標	模型	音標
AA	ɑ	D	d	IH	ɪ	OW	o	TH	θ
AE	æ	DH	ð	IY	i	OY	ɔɪ	UH	ʊ
AH	ʌ	EH	ɛ	JH	ʤ	P	p	UW	u
AO	ɔ	ER	ɝ	K	k	R	r	V	v
AW	aʊ	EY	e	L	l	S	s	W	w
AY	aɪ	F	f	M	m	SH	ʃ	Y	j
B	b	G	g	N	n	SIL	sil	Z	z
CH	ʧ	HH	h	NG	ŋ	T	t	ZH	ʒ

我們使用以下三種原則來對 TIMIT 的 62 個 model 做刪減的動作，各 model 後面刮弧內的英文字其底線部份即表示該 model 的發音。

- 替換：將發音相似的音素使用一個 model 代替。例如：**AXR** (butter〔ɚ〕) → **ER** (bird〔ɝ〕), **NX** (winner〔ɲ〕) → **N** (noon〔n〕)
- 分解：將一個 model 拆開成兩個以上的 model 來組成。例如：**EN** (button〔n̩〕) → **AH** + **N** (〔ʌ n〕), **ENG** (Washington〔ɪ ŋ〕) → **IH** + **NG** (〔ɪ ŋ〕)
- 刪除：將許多設定細微的暫停音素刪除。例如：**PAU**、**EPI**

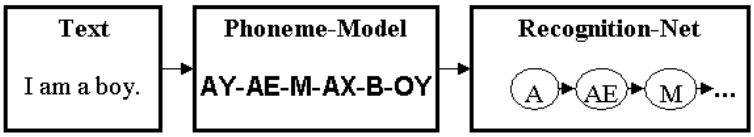
3.2.3 聲學模型訓練

在訓練語料的特徵擷取中，我們以 20ms 為單位取音框，每隔 10ms 重取音框。每一音框採用 39 維特徵向量當作聲學模型之特徵參數[7]，該特徵向量包含了 12 維的 MFCC 和一維能量參數，以及其一階微分、二階微分。

聲學模型訓練部份，我們採用隱藏式馬可夫模型(Hidden Markov Model, HMM)[6]作為聲學模型。在 monophone 的 model 下， 每個 model 有 3 個 state，每個 state 則使用了 18 個 mixtures，以 HTK(Hidden Markov Model Toolkit)進行訓練。

3.2.4 訓練結果

語音訊號切割的主要目標即是希望能夠將連續的英文語音句子，其中包含了標準語音和評分的語音，切割成獨立的音素，如此一來我們才可以針對每一段句子中的音素和標準語音中的每一個音素做比較。在此我們使用強迫對應(Forced Alignment)[6]的方式將語音訊號切割成各個音素的時間區段，以利評分機制的運作。在前處理的過程中，我們利用內含 127,102 個英文單字的 CMU 字典(Dictionary from Carnegie Mellon University)對各單字標音並建立各自獨立的辨識網路[21]。如下圖：



圖表 7 語音訊號切割前處理流程示意圖

完成前處理動作後，我們可繼續進行語音訊號切割的流程，首先將一語音訊號經過端點偵測後再經由特徵擷取，取出語音中的特徵，然後將這些特徵參數透過聲學模型(隱藏式馬可夫模型)及語言模型(辨識網路)，利用維特比演算法(Viterbi algorithm)即可找出最相似的音素，並得知各音素的時間區段。

關於實驗測試語料的部份，我們使用了 1,680 句母語為英文的語音檔案，其語料的長度總和約為 1 小時 23 分 51 秒，以下我們簡稱為 N-Wave (Waves from Native-Speaker)。另外使用了 2,342 句母語為國語的語音檔案，語料的長度總和約為 1 小時 57 分 43 秒，以下簡稱為 T-Wave(Waves from Taiwanese)，來做 Outside Test。實驗用的語料其音訊格式皆為 PCM，音訊取樣頻率為 16 kHz，位元解析度為 16 bits。

在聲學模型這個部份，我們訓練出了兩個聲學模型：一個是由以英文作為母語的使用者所錄製的訓練語料產生的聲學模型，以下我們簡稱為 N-HMM(HMM trained from Native-Speaker)，另一個則是由臺灣人所錄製的訓練語料所產生的，以下我們簡稱為 T-HMM(HMM trained from Taiwanese)。

關於實驗的方式，我們分別對每一句語音訊號和已知的語音內容文字作 Forced Alignment，再由產生的結果對每個單字及音素判斷其時間區段的切割是否正確。

為了比較兩個聲學模型所產生的影響，我們對語料(N-Wave, T-Wave) 和聲學模型(N-HMM, T-HMM)作交叉實驗。表格 2 列出音素切割正確率的實驗結果：

表格 2 語音訊號切割實驗結果

實驗方式 / 項目	N-Wave /N-HMM	N-Wave /T-HMM	T-Wave /N-HMM	T-Wave /T-HMM
實驗語料音素總數	58,282	58,282	81,229	81,229
切割後正確音素總數	58,253	57,142	77,293	80,230
音素時間正確率	99.95%	98.04%	95.15%	98.77%

在判斷音素時間正確率的部份，對於 N-Wave 而言，由於所有的語料 TIMIT 都有提供標音檔，因此我們可比對切割出來的時間點和標音檔，若相差在 0.1 秒以內(5 個音框)，則我們稱此音素的時間為正確。而對於 T-Wave 而言，由於並沒有經過人工標音，因此我們只在龐大的語料中取樣 10%進行人工判斷，只要該區段人耳聽起來相差不大，則我們稱該音素的時間為正確。

由表格 2 的實驗結果可知，在不同的聲學模型下，Forced Alignment 的音素時間區段都非常準確。表格 3 則是 N-Wave、T-Wave 透過大詞彙辨識的方式，經由 N-HMM、T-HMM 所得出的辨識率，其中詞彙內容為 2,342 句英文句子。

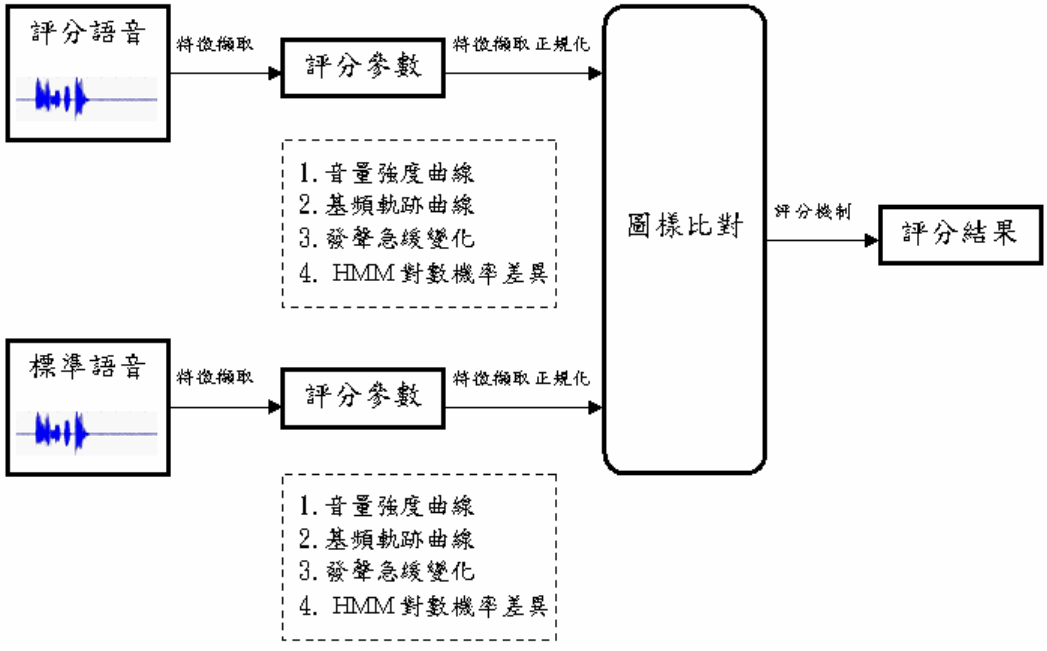
表格 3 英文語音辨識率

實驗方式 / 項目	N-Wave /N-HMM	N-Wave /T-HMM	T-Wave /N-HMM	T-Wave /T-HMM
實驗語料句子總數	1,680	1,680	2,342	2,342
辨識正確句子總數	1,650	622	1,997	1,425
句子辨識率	98.21%	37.02%	85.26%	60.85%

由表中的結果我們可以發現，對於相同語料，N-HMM 的辨識率皆高於 T-HMM，這就表示當我們以 N-HMM 為聲學模型來對語音訊號求取對數機率時，所得到的對數機率值其可信度會高於 T-HMM。根據此實驗結果，在接下來的章節中，我們將會以 N-HMM 當作我們評分比對的聲學模型。

3.3 英文語音評分

圖表 8 為評分系統流程圖，我們將就評分參數擷取、圖樣比對方式和評分機制建立分別作介紹。



圖表 8 評分系統流程圖

3.3.1 評分參數擷取

除了音量強度曲線、基頻軌跡曲線為評分參數外[15]，我們也採用了 HMM 對數機率差異和發聲急緩變化這兩項評分參數。在 Forced Alignment 的同時，我們可以得到每個音素對應於聲學模型的對數機率(HMM log-Probability)[10][11]和各音素的時間區段，這就是所謂的 HMM 對數機率差異和發聲急緩變化這兩項評分參數。

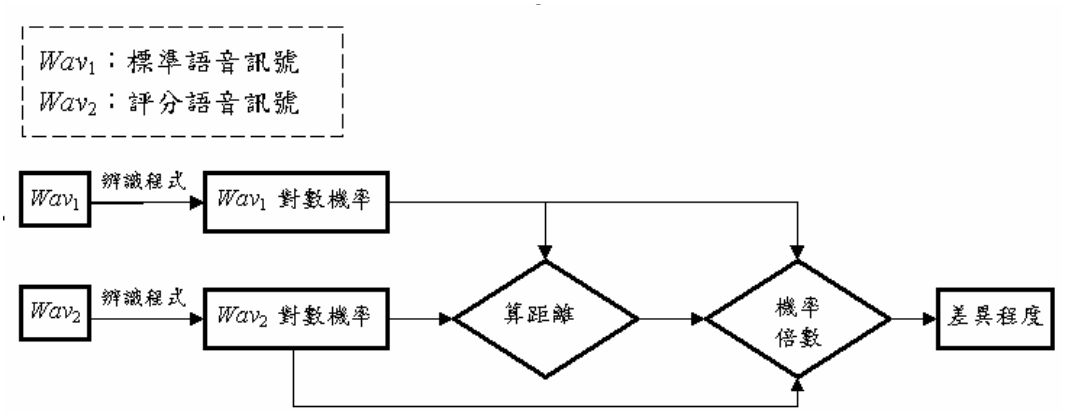
3.3.2 圖樣比對方法

在前三個評分參數中，我們使用不同的正規化方法如內插法、線性平移和線性縮放[15]，如表格 4。而 HMM 對數機率差異則採用較為不同的比對方法，在以下說明。

表格 4 各評分參數採用的正規化及距離算法

評分參數	正規化方法	距離算法
音量強度曲線	內插法、線性縮放	Euclidean Distance
基頻軌跡曲線	內插法、線性平移	Euclidean Distance
發聲急緩變化	無	Euclidean Distance

為了計算 HMM 對數機率的差異，我們先以 N-HMM(HMM trained from Native Speaker)求出標準語音訊號及評分語音訊號中每個音素的對數機率，若對數機率值愈大，表示該音素的發音愈接近聲學模型。圖表 9 為 HMM 對數機率差異比對的流程圖：



圖表 9 HMM 對數機率差異比對流程圖

由於機率值是絕對的，不容易從數值直接作比較，因此我們設計了機率倍數來修正對數機率的差異值，當兩語音的對數機率絕對值皆小於 1050 時，機率倍數的變化趨勢較小。當兩語音的對數機率絕對值皆大於 1050 時，機率倍數的變化趨勢較大。關於機率倍數我們定義以下的公式：

$$Const = \begin{cases} \left\lceil \dfrac{|\log - probabilit\,y|}{350} \right\rceil, & 0 \le abs_{\log} \le 1050 \\ 3 + \min\left(1, \left\lceil \dfrac{|\log - probabilit\,y|}{1400} \right\rceil\right), & abs_{\log} > 1050 \end{cases}$$

$$factor_P = (Const_{stard})^2 + (Const_{Evaul})^2$$

當算出標準語音和評分語音的 $Const$ 值後，再經由平方相加即可得到機率倍數 $factor_P$，將此機率倍數乘上兩語音訊號對數機率的差距就是我們發音特徵的差異程度。

3.3.3 評分機制建立

在音素層次，我們由四種評分參數得到不同的分數，再往上由單字(word)和句子(sentence)層次作評分，就可以得到最後評分的結果，以下則分四個層次作介紹。

評分參數層次：對於每個音素中評分參數的分數，我們設定以下的公式[15]：

$$score_{fea} = \frac{100}{1 + a \cdot (dist)^b}$$

由這個公式我們就可以將兩音素間某個特徵的差異程度轉成 0 到 100 之間的分數，只要設定好兩組的 $dist$ 及對應的 $score_{fea}$，即可從中求出 a 和 b，接著所有的距離也將可以計算出對應的分數。

音素層次：當計算出每個音素中四項評分參數的分數後，利用四項特徵對於英文語音評分系統所占的權重加總後即可得到每個音素的分數。以下是設定的公式：

$$score_{pho} = w_1 \cdot score_{fea_1} + w_2 \cdot score_{fea_2} + w_3 \cdot score_{fea_3} + w_4 \cdot score_{fea_4} ,$$

w_1、w_2、w_3、w_4 分別代表四個評分參數的權重。經由下一節的實驗，我們可以求出這四項權重，也可以由權重的比例得知四項評分參數對於英文評分的重要性。

單字層次：得知每個音素的得分後，以每個音素占單字的時間為權重，即可求出句子中每一個單字的分數，以下為設定的公式：

$$score_{word} = \sum_{n=1}^{N} \frac{len(pho_n)}{len(word)} \cdot score_{pho_n}$$ ，其中 N 為一單字中評分音素的數量，*len(x)*表示 *x* 的時間長度。

句子層次：由於單字的時間長短會影響人耳對於一句話的關注點，因此我們也是以單字的時間為權重來計算出一句語音訊號最後得到的分數。以下為定義的公式：

$$score_{sen} = \sum_{n=1}^{N} \frac{len(word_n)}{len(sentence)} \cdot score_{word_n}$$ ，其中 N 表示句子中單字的總數，*len(x)*表示 *x* 的時間長度。

4 實驗結果

得到四個評分參數中各音素的差異程度後，我們依所佔的比例求出一個句子的平均差異程度，即可代入以下的公式：

$$score = w_1 \cdot \frac{100}{1 + a_1 \cdot (dist_1)^{b_1}} + w_2 \cdot \frac{100}{1 + a_2 \cdot (dist_2)^{b_2}} + w_3 \cdot \frac{100}{1 + a_3 \cdot (dist_3)^{b_3}} + w_4 \cdot \frac{100}{1 + a_4 \cdot (dist_4)^{b_4}}$$

其中 $a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4$ 為差異程度轉成分數的參數，w_1, w_2, w_3, w_4 為四個評分參數的權重，而 $dist_1, dist_2, dist_3, dist_4$ 表示標準語音和評分語音訊號在比對後其四項評分參數的距離，再經由以下的實驗，即可求得各參數值。

在語料訓練部份我們收集 200 組語料，每一組的語料分別包括一句標準語音和一句評分語音，每句語音長度為 5 秒、音訊格式為 PCM、音訊取樣頻率為 16 kHz、位元解析度為 16 bits。其中標準語音的語料長度總和約為 12 分 51 秒，評分語音的語料長度總和約為 18 分 39 秒。接著請外語所老師協助我們對每一句評分語音作主觀的評分，之後再統計實驗中每一句語音人為評分的平均分數。同樣的，按照訓練語句的作法，我們也收集了 200 組語句作為測試用。

將這 200 組訓練語料透過評分系統評分，則每組評分語音都會得到四個特徵對應的差異程度 $dist_1, dist_2, dist_3, dist_4$。收集了這些差異程度和對應的分數後，使用 Simplex Downhill Search，就可以找出 $a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4$ 和四個評分參數的權重 w_1, w_2, w_3, w_4。

經由上述的實驗，我們得到音量強度曲線的權重為 7.45%，基頻軌跡曲線的權重為 22.40%，發聲急緩變化的權重為 17.24%，HMM 對數機率差異的權重為 52.91%。

接著我們將 200 句測試語句的人工評分結果分成三個等級：Bad(0~59)、Average(60~79)、Good(80~100)，另外也把將 200 句測試語句的系統評分結果依此分成三個等級。最後再統計每個句子的人工評分和系統評分後，就可以得到表格 5的結果：

表格 5 人工評分和系統語音評分的關係對照表

人工評分 / 系統評分	Bad	Average	Good
Bad	28	17	7
Average	20	27	20
Good	10	11	63

其中橫軸表示人工評分的等級項目，縱軸表示系統評分的等級項目，表格中的數字則表示相對的語句數目。從表中我們可以明顯地看出來，對角線的數目都比同一列、同一欄的數目高，這就表示在經由 Simplex Downhill Search 調整各參數之後，我們的評分系統和人工評分已有一定的正相關性，約 (28+27+63) / 200 = 59%。

5 結論

「說話驗證」對評分語音進行初步的評估，若可信度夠高，接下來的評分才具有可信度。「語音訊號切割」則是以 Forced Alignment 得到每個音素的時間區段。經由實驗結果我們可以知道，使用辨識率較高的聲學模型，其 Forced Alignment 的音素切割時間將更為準確。「英文語音評分」包括評分參數的擷取、圖樣比對方法的設計和評分機制的建立等三個部份。藉由實驗我們可以知道，「HMM 對數機率差異」在英文語音評分中所代表的重要性最高，而「音量強度曲線」則是最低。

語音評分的運用相當廣泛且實用，配合未來技術的成熟，不只可作為英語學習的工具，之後的台語、客語評分學習也將是台灣地區重要的研究之一。

參考資料

[1] 鐘林，“漢語語音辨別說話驗證”，北京清華大學碩士論文，民國 91 年

[2] 楊永泰，“隱藏式馬可夫模型應用於中文語音辨識之研究”，中原大學碩士論文，民國 89 年

[3] 陳柏琳，“中文語音資訊檢索－以音節為基礎之索引特徵、統計式檢索模型及進一步技術”，台灣大學博士論文，民國 90 年

[4] 呂道誠，“不特定語者、國台雙語大詞彙語音辨識之聲學模型研究”，長庚大學碩士論文，民國 90 年

[5] G.S. Ying, L.H. Jamieson and C.D. Michell, A probabilistic approach to AMDF pitch detection, Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on Volume: 2 , 1996 , Page(s): 1201-1204 vol.2

[6] Steve Young, The HTK Book version 3, Microsoft Corporation, 2000

[7] Lawrence Rabiner, B.H Juang, Fundamentals of speech recognition, Prentice Hall, 1993

[8] J.D., J.G., J.H. and L.H., Discrete-Time Processing of Speech Signals, Prentice Hall, 1993

[9] Giuliano Monti, Mark Sandler, Mnophonic transcription with autocorrelation, Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italy, December 7-9, 2000

[10] L. Neumeyer, H. Franco, V. Digalakis and M. Weintraub, Automatic scoring of pronunciation quality, 1999

[11] H. Franco, L. Neumeyer, Y. Kim and O. Ronen, Automatic pronunciation scoring for language instruction, Proc. Int. Congress on Acoustics, Speech and Signal Processing(ICASSP), 1997

[12] J.-S. Roger. Jang, C.-T. Sun, and E. Mizutani, Neuro-Fuzzy and Soft Computing, Prentice Hall, 1996

[13] 高名揚，“以聲音內容為主的音樂資料庫檢索系統的加速方法”，清華大學碩士論文，民國 90 年

[14] J. T. Tou and R. C. Gonzalez, Pattern Recognition Principles, Addison-Wesley Publishing Company, 1974

[15] 李俊毅，“語音評分”，清華大學碩士論文，民國 91 年

[16] Gies Bouwman and Lou Boves, Utterance Verification based on the Likelihood Distance to Alternative Paths, Department of Speech, University of Nijmegen, The Netherlands, 2002

[17] Rafid A. Sukkar and Chin-Hui Lee, Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword based Speech Recognition, IEEE Transactions on Speech and Audio Processing, VOL. 4, No. 6, November 1996

[18] Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, and Patti Price, Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech, 1996

[19] C. Cucchiarini, H. Strik and L. Boves, Automatic Evaluation of Dutch Pronunciation by Using Speech Recognition Technology, Department of Speech, University of Nijmegen, The Netherlands, 1997

[20] Jiang-Chun Chen, Jyh-Shing Roger Jang, Jun-Yi Li and Ming-Chun Wu, “Automatic Pronunciation Assessment for Mandarin Chinese”, Proc. Int. Conf. on Multimedia And Expo (ICME), 2004

[21] http://www.speech.cs.cmu.edu/sphinx/doc/phoneset_s2.html

The Design of Computer Multimedia Material for English Language Learning

Yu liang Ting
Oriental Institute of Technology
Ff031@mail.oit.edu.tw

Yaming Tai
Yuan Ze University
yaming@saturn.yzu.edu.tw

Abstract: This paper addresses the design of multimedia presentation of language learning material for Freshman English writing course at a University in Taiwan. The material is used to assist teacher's instruction, and serves as the preliminary step regarding the implementation of instructional strategies in the language teaching course for digital learning materials. Feedbacks from students are collected to reflect learners' preferences over the current status of multimedia learning. The purpose of this paper is to demonstrate an initial step in achieving the effective learning by coordinating external and internal stimulus to the learners' cognitive activities.

1 Introduction

The advance of Information Technology (IT) has varied the ways of learning in recent years. One important application of information technology is distance learning, which has become a prevalent way of learning. Research has focused on the comparison of learning performance between traditional learning and digital learning. However, due to the cost-effectiveness of digital learning and the widely spread of the Internet, it is foreseen that digital learning will become one of the important method in the future learning activities. In general, the applications of IT in the field of digital learning can be classified into learning material development and learning management system. The study focuses on the material development related to the instructional and learning theory, and it serves as the preliminary discussion related in achieving the effective learning through well-designed digital learning material.

As the development of IT, many areas of English learning have employed computers as learning tools. In the area of English writing, the most common use of the computer technology is mediating communication such as E-mail, chat, or MUDS and MOOS (Sokolik, 2001). By employing the notions of negotiation of meaning from second language acquisition, the goal of students' writing is having students engage in real communication through the Internet. Some researchers found that the computer is a helpful tool; nevertheless, there are still some other aspects that computers can provide to facilitate students' writing. The area of how to use the computer to help learners develop and elaborate their specified cognitive representation for their second language writing is still under explored. Not knowing how to start writing and what to write are always two major challenges novice writers face. The use of information technology can help English writing teachers to produce multimedia materials to facilitate their students in their writing processes. The study is intended to use the computer as a tool to integrate teaching materials through the use of multimedia to motivate students by presenting second language writing in its more complete communicative context. In addition, with regard of schema theory, the use of media can help students relate their existing schemata and employ their prior background knowledge in their writing process.

In the study, theories related the above goal are discussed in the following section of literature review. It consists of three parts: the first one is the fundamental instruction and learning theories, and language learning theory. The second one is the design of multiple external representations in facilitating instruction and existing theories in achieving effective learning. The third one is previous study done by the author related to the effects of various ITs upon the development of learning materials and learners' learning activities. In light of these reviews, this study focuses on the implementation of instructional strategies in the development of digital learning materials, and the first step is the integration of classroom curriculum into digitalize and programmed computer material. The preliminary design philosophies and related program in the implementation are presented and discussed in the section of material design. Then, the designed material was used in a semester course, and students' feedbacks were collected to reflect current status of digital learning material upon college student's learning, followed by the conclusion as the end of this paper.

2 Literature Review

2.1 Language learning and Instruction

The theory regarding how people learn the foreign language primarily base on two assumptions. The first one focuses on the cognitive science, and discusses the mental process of learners with their internal representations and their structures related to the external language stimuli. The second one originated from the cultural psychology, which treated the learning process as a social and cultural process. Frawley (1997) proposes non-conscious, conscious, and meta-conscious processing. The first two processes recognize the learning process as a cognitive activity, and the third one fulfills the role of human learning as an internal mental process and external social-cultural interaction. Zahner et al (2000) described the language learning is a deliberate and controlled activity, not only involves the non-conscious and conscious processes but also situated it the meta-conscious stage. It invokes learners' inter-mental process socially and culturally.

The trend of teaching English writing has changed from product-oriented to process oriented. Unlike the traditional paradigm, which focuses on evaluating students' writing, the process approach emphasizes on the writer's whole writing process (Kroll, 2001). From a process perspective, the composing process is a recursive, exploratory and generative one (Silva & Matsuda, 2002). The procedure includes stages of generating ideas, structuring, drafting, reviewing, evaluation, and focusing (White & Arndt, 1991). Therefore, incorporating writing strategies is an important issue in the classroom of process approach. However, the process writing approach is time consuming (Harmer, 2001).

2.2 Multiple External Representations

Ainsworth, Bibby and wood (1997) made a study with performance of estimation task, the learning process is carried out by two different cases. The first one learning material has two External Representation carrying the same information for children to learn, and the second one simultaneously display two dimensions of information. For example, a mathematic variables relationship is to be stated in both equation and picture.

Anisworth's Multiple External Representation has the functions of complement, constrain, and construct. The complementary information is needed when a single representation is not sufficient to present all the information needed. The algebra relationship between mathematic variables is an example. The relationship can be presented as mathematic equation in text format or graphic image. For a problem solving study, Tabachneck et al found that, in an algebra problem, each representation was associated different strategies. Multiple representations stimulate learners to exercise multiple strategies. The weakness of each strategy is overcome by switching the strategies during the problem solving process.

Moreover. the function of constrain is achieved by using one representation to constrain the interpretation of a second representation. By doing so, the learner can develop a better understanding of a domain. One of the examples is using the inherent properties of a presentation to constrain the ambiguity of another one. Stenning and Oberlander (1995) presented a study regarding that the graphical representation in general is more specific than the narrative one in describing the relative position of two objects. Hence, the narrative representation can be constrained by the second graphical one.

The third function is construction of deep understanding and thinking. Usually, for an abstract idea or notion, it is very difficult for learner to build his internal metal entity (or representation) for further construction of concepts or procedure knowledge. Dienes (1973) proposed that same concept expressed in varying way can help learner building such mental entities by providing perceptual variability.

Since this study is to implement the instructional strategies with multiple representations in the multimedia material to activate the deep thinking in the learners, the studies of multiple representations are needed to be included. One of the major functions in multiple representations is the construction of thinking, and it can be achieved by the abstraction and extension of the relations between representations in learner's internal metal exercise or instructor's direct teaching (Ainsworth , 1999). The challenge then will be the role of multiple presentations during the learner's conceptual construction process. Too many external representations will make the learner passive in forming his own mental presentation and hurdle his further construction of existing concept with new one. However, too few representations may not make learner form the corresponding entity or a wrong entity. Contingency theory (Wood & Wood) suggests that the support of presentation should be based on the learner's feedback and performance during the learning activities. Scaffolding strategy proposes correspondingly that the level of support should fade out gradually when the learner can achieve a cognitive linking of representation. The representation can be removed when the learner can construct his own internal representation without much external stimulus.

2.3 Effecst of Information Technologies

The advance of computer network and related information technology makes the Web-based learning as a new learning style. This new type of learning model is pushing educational scholars to redefine the field of instructional material design and related theories of learning, especially regarding the mode of presenting material as well as the interaction between learners and materials. Among the above issues, the varieties of presentation modes are increased by the advance of related information technologies. The presentation modes in the study include classroom instructor-lead mode, streaming video mode, and programmed mode.

The challenge of integrating curriculum and implementing instructional strategies into computer multimedia material primary is the gap between the instructor and the program expert regarding the knowledge to teach and technical limitation of authoring tool. The author of this study (Ting, 2004) proposed a model addressing the relationship model among instructor, programmer, and learning content to reflect this challenge, and also use a practical case to examine the effects of various information techniques applied in the authoring process, with the interview results and participating in the material development process, Ting presented several phenomena observed. For instance, the instructor mentioned that he removed "situation-based" lecture used in the classroom when he was filmed for the streaming video mode materials. The "situation-based" lecture was a reinforcement of his teaching based on the learners' responses during the lecture. Since there were no responses in the filming process, he got no clues or intentions to give these situation-based talking, especially giving example(s) for some proposition(s). Therefore, the learners of streaming video mode might not get the chance to learn various examples of propositions if they needed them. The deletion of the knowledge content when the instructor cooperated with the programmer in developing programmed mode materials, for example Authoreware. Since the content needed to be transferred and interpreted by the programmer, if a concept was difficult for the programmer to present with the programming tool, it might be removed due to the constraint of resources.

As to the learners' feedback to the digital learning material, Ting (2003) found that significant differences are found due to the different modes of presenting materials. Participants from three different backgrounds (university students, vocational school students, and enterprise employees) are invited to participate in the study. The significant differences of learners' satisfaction toward three learning modes are learning style (presentation mode) and time arrangement. In the survey of learning style, classroom instructor-leading mode has the best satisfaction, followed by programmed mode, then streaming video. According to the follow-up interviews with some participants, their responses were that they had been used to the traditional classroom learning mode, and had direct eye contact with the teacher even there were no Q/As between them. The participants in the other two E-learning modes responded that they were unfamiliar with the learning situation in which the instructor did not show up. It infers that digital learning has not been popular among these participants. The second highest satisfaction of learning mode was the programmed one, and the reason was the good interaction between the material and learners.

2.4 Summary

The primary components in designing the digital learning applications are: learner motivation, learner interface, content structure and sequencing, navigation and interactivity (Allen, 2003). The goal of digital learning material for sure is to assist learner to achieve best learning performance through effective instructional strategies implemented in the learning material. Several related theories have been revealed in the above section, and their concern is to design a best learning environment. Such environment not only creates the external stimulus in activating learner motivation into the engagement of learning, but also strengthens the internal cognitive processes in forming learner's propositions and procedure knowledge.

This study will base on the above notions to implement related theories into a language learning course, "Freshman English Writing". Due to the constraint of available resource, this paper will present a preliminary result in integrating the multimedia materials which are used separately during the instruction activities.

3 Material Design and Development

Before this study, the writing materials used to assist the teacher's instruction are textbook and graphics in the hard-copy format, and video and audio tapes played by video/audio player. In this study, all of the above materials are digitalized through appropriated equipments and transformed into appropriate formats for the storage and later manipulation by the computer authoring program. The integration and sequencing of learning material is programmed through the Authoring tool: Macromedia Authorware. Authorware is chosen due to its

strength in variables setup and logic controls in fulfilling the needs of multiple flow paths for supporting adaptive learning, and such characteristics are contrast to those in Flash or Director, which is good at rich multimedia presentation and graphic motion control.

Functions of learners' interaction with material includes the content browsing and quiz taking, and they are made through some information technologies to make the material more attractive to students. The program is described logically to demonstrate how the fundamental sequencing mechanism is implemented. Type of responses available in Authorware 6.0 are Button, Hot Spot, Hot Object, Target Area, Pull-down Menu, Conditional, Text Entry, Key-press, Tries Limit, Time Limit, and Event. Integrating the above user responses with predefined control logic, the implemented material can provide fundamental passively adaptive learning for individual learner. One of the goals in future study is to include the learner profile into the sequencing and navigation of learning activities actively and dynamically.

In this study, the material was designed for "Freshman English Writing" course for English-major students at a university in northern Taiwan. Because this project was carried out in the second semester of an academic year, the researchers used the already-used textbook, Developing Composition Skills by Mary K. Ruetten, as a major source of material development in order not to confuse students' learning. The content of three chapters, which represent three writing genres, digitalized as computer assisted language learning materials. Every genre includes phases of prewriting activities, readings, writing step explanations, language usage, and writing assignments. In addition to the revised textbook contents, some extra writing tasks were also inserted. The goal of these writing tasks is to provide learners more authentic materials as well as more authentic contexts to write. For the purpose of illustration, some examples of the chapter related to analyze a process are illustrated below.

Because the reading in this chapter is about how to do a library searching operation on a computer, the prewriting activities the researcher designed are asking students to do some library searches at their school library and list the procedure step by step. The purpose of these activities is to help students relate to their schemata for their future writing. The activities are shown in Figure 1 and Figure 2.

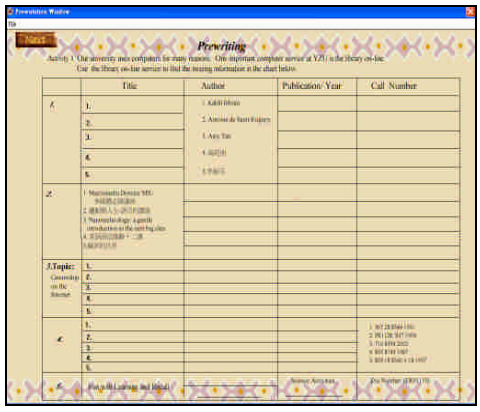


Figure 1. Presented learning activities for CALL

Figure 2. Presented learning activities for CALL

In order to motivate students' learning, the model essay for students to analyze the writing styles, the content of one of the essays are divided into different sections for students' comprehension check. In addition, the sentence order of the other model essay was scrambled for students to find the right process of the action. Furthermore, students can choose to listen to the texts being read by recorded voice.

With regard of writing assignment, the researchers designed a series of activities as prelude to the process analyzing writing. This was done due to the fact that students always complained about not knowing what to write. These activities offered students some vocabulary words and ideas for their writing assignment. In this case, they needed to write some procedures of traditional Chinese or Taiwanese wedding. One of the activities was done as an information gap activity. Students read some gapped texts related to the traditional western wedding and listen to the recorded voice to read the complete version of the text. Then they have to fill in the gaps by selecting words from the word bank offered on the computer screen. In this way, they may learn some vocabulary words related to wedding (see Figure 3). The text was adopted from the Cultural dictionary, and has been revised. They also can try to find some websites to provide them information about Chinese or Taiwanese wedding traditions (Figure 4). Moreover, they could click the viewing buttons to view some video clips of a Taiwanese wedding (Figure 5). When the student learned several marriage related vocabulary and customs, they were required to write the essay. The material here is intended to promote students' interest in the wedding procedure, and motivate their intentions in using rich and various terminologies in their writing.

Figure 3. Portal for supplementary CALL material - "wedding"

Figure 4. The narration of multimedia learning material

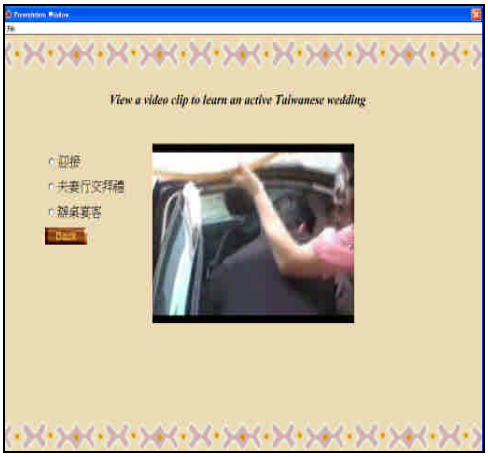


Figure 5. The video clips of CALL material

As shown in the Figure 5, the buttons are made with the Authorware program to support the function of sequencing program.

4 Preliminary Survey Result

Subjects

Two classes with 24 and 25 freshman students majoring in English were presented with the experimental material. . The learning content is Freshman English Writing course presented in the multimedia form described in the previous section. Due to the limitation of time and available computers, all learners are seated as in the computer lab with computer projector during the learning activity. The instructor (author of this study) used a computer connected to the projector to present the CALL material. All learners were encouraged to ask questions as a normal classroom course, and they also had their own textbooks for reference.

Instrument

Learners' feedbacks toward the designed learning material were surveyed by a Likert questionnaire as shown in Appendix 1. The topic is divided into four categories: attractiveness of presentation mode, helpfulness of presentation media, personal preference toward presentation style, and overall acceptance. There are five questions for each category, and each question has 5-point scale. The total questions are 20.

Survey Results

The preliminary summary result of learners' feedbacks upon the designed CALL material is shown in the Figure 6, and it is obvious that the attractiveness of multimedia presentation gains the best positive feedback from the students. As to the helpfulness in assisting learning and the style of learning method play fair around the average role in the overall gauge of learning activities for language learning. The results may be expected to be further improved by the inclusion of adaptive interaction between learner and material.

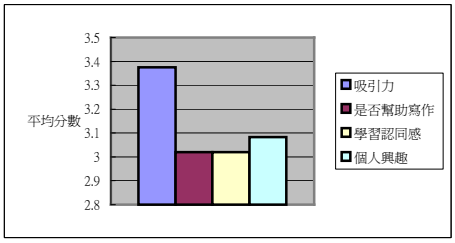


Figure 6. Summary result of learners' feedback upon the designed CALL material

5 Conclusions

The role of computer has changed from the facilitator of learning into activate learners into deep thinking. However, misunderstanding regarding the attractiveness of learning material and interactivity of learner with material is the richness of display and plenty of menu buttons in the user interface. As a matter of fact, the effectiveness of learning is best achieved through the construction of learner's internal reflection and thinking due to the external multimedia stimulus, which should be able to activate learners into the engagement of deep thinking and constructing their own mental entities and corresponding representations and concepts. This paper reveals some fundamental theories related to the up-to-date instruction and learning theory, especially those related to the Computer Assisted Language Learning. In addition to the challenge faced by computer multimedia material designers, how the learning material should play as an effective one is clarified from the perspective of multiple external representations. Followed with studies related with the effect of various ITs upon the development of learning material, learners' feedback upon different technologies are also briefed.
Along with the above literature review, this paper addressed the design of multimedia presentation of language learning material for an English Writing course offered at an University in Taiwan. The material was used to assist teacher's instruction, and serve as the preliminary step regarding the implementation of instructional strategies in the language learning course for digital learning materials. Feedback from students were collected to reflect learners' preferences over the current status of multimedia learning
Further study will concentrate on three issues in sequences. The first one is the integration of Authorwaring program with the instructional strategies. The next step will be the incorporation of dynamic learner profile into the navigation and sequencing of learning material. The third step is to examine the effect of the above function's performance from the perspective of social and cultural interaction between learners and learning material. It is expected to achieve the effective learning by coordinating external and internal stimulus to the learners' cognitive activities through well-designed digital learning material.

Acknowledgements

This study's financial support is given by Yuan-Ze University in developing the computer assisted teaching materials, and research grant is also partly provided by Oriental Institute of Technology. Appreciation is give to research assistants of Ssu-Chia Chen (陳思嘉) at Yuan-Ze university and Chun Ting Chen (陳俊廷) at Oriental Institute of Technology for their helps in the graphics, and computer program presented in this paper.

References

Allen, M. W. (2003) *Michael Allen's Guide to e-learning*, John Wiely & Sons, Inc., New Jersey.
Ainsworth, S. E., Bibby, P. A., & Wood, D. J. (1997) Evaluating principle for multi-representational learning environments. *7th EARLI Conference*, Athens.
Ainsworth, S. (1999) The Functions of Multiple Representations, *Computer & Education*, 33(2/3), p. 131-152.
Dienes, Z. (1973) The six stages in the process of learning mathematics. Slough, UK, NEFR-Nelson.

Frawley W. (1997) *Vygotsky and cognitive science: Language and the unification of the cognitive and social mind*. Cambridge: Harvard University Press.
Harmer, J. (2001) The Practice of English Language Teaching. Longman.
Kroll, B. (2001) Considerations for Teaching an ESL/EFL Writing Course. In *Teaching Englihsh as a Second or Foreign Language*, edited by M. Celce-Murcia. Heinle & Heinle.
Ruetten, M. K. (2003) Developing Composition Skills. Heinla & Heinle.
Silva, T. & Matsuda, P. K. (2002) Writing. In *An Introduction to Applied Linguistics*, edited by N. Schmitt. Arnold.
Sokolik, M. (2001) Computers in Language Teaching. In *Teaching Englihsh as a Second or Foreign Language*, edited by M. Celce-Murcia. Heinle & Heinle.
Stenning, K., & Oberlander, J. (1995) A Cognitive Theory of Graphical and Linguist Reasoning: Logic and Implementation. *Cognitive Science*, 19, 97-140.
Ting, Y. L. (2004) "The Comparative Study of Two Types of Information Technology Applied in the Development of E-Learning Materials," *Proceedings of the 2nd International Conference on Information Technology for Application*, Jan. 08~11, Harbin, China.
Ting, Y. L. (2003) The Effectiveness of Different Presentation Modes of E-Learning Materials in Learning Performance, *International Conference on Computers in Education*, Hong Kong, Dec 2003, pp. 990-994.
White R. & Arndt, V. (1991) Process Writing. Pearson Education Ltd.
Zahner, C. Fauverge, A. & Wong, J. (2000) Task-based language learning via audiovisual networks The LEVERAGE project. (In Warschauer M. & Kern R. Ed.) *Network-based Language Teaching: Concepts and Practice*, Cambridge University Press.

Appendix

	題　目	非常同意	同意	普通	不同意	非常不同意
1	本課程所使用電腦教學工具來呈現教材的教學方式，是吸引我的	5	4	3	2	1
2	本課程的電腦教學教材能引起我的興趣	5	4	3	2	1
3	我不會對本課程的電腦教學教材的內容感到枯燥乏味	5	4	3	2	1
4	我的注意力會被電腦教學教材所吸引	5	4	3	2	1
5	本課程的電腦教學教材是生動有趣的	5	4	3	2	1
6	本課程的電腦教學教材，更能協助我寫出好的文章	5	4	3	2	1
7	使用電腦呈現多媒體教材的學習方式，有效的增進我的英文寫作能力	5	4	3	2	1
8	本課程的電腦教材內容，使我更進一步瞭解寫作技巧	5	4	3	2	1
9	課堂所使用的電腦多媒體，更能激發我思考英文寫作的相關知識	5	4	3	2	1
10	英文寫作能力的培養，能透過本課程的電腦教材，提供有效的訓練內容	5	4	3	2	1
11	使用電腦多媒體教材的學習方式，對英文寫作是適當的	5	4	3	2	1
12	我能接受以電腦多媒體教材的學習方式，來學習英文寫作	5	4	3	2	1
13	對學習英文寫作而言，以電腦多媒體來呈現教材的學習方式是適當的	5	4	3	2	1
14	電腦多媒體教材的學習方式，應該用來替代傳統教學方式	5	4	3	2	1
15	對未來英文寫作或相關課程，希望能以電腦多媒體方式呈現教材內容	5	4	3	2	1
16	整體而言，我喜歡電腦多媒體教材的學習方式	5	4	3	2	1
17	電腦多媒體教學方式，無法增進我對英文寫作的滿意情形	5	4	3	2	1
18	對本課程的電腦教學內容，滿足我對英文寫作練習的期望	5	4	3	2	1
19	我無法接受英文寫作課程，以電腦多媒體教學方式進行	5	4	3	2	1
20	我對本課程的電腦教學感到滿意	5	4	3	2	1

Collocational Translation Memory Extraction Based on Statistical and Linguistic Information

Jia-Yan Jian
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road, Hsinchu, Taiwan
d914339@oz.nthu.edu.tw

Yu-Chia Chang
Inst. of Information System and Applictaion
National Tsing Hua University
101, Kuangfu Road, Hsinchu, Taiwan
u881222@alumni.nthu.edu.tw

Jason S. Chang
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road, Hsinchu, Taiwan
jschang@cs.nthu.edu.tw

Abstract. In this paper, we propose a new method for bilingual collocation extraction from a parallel corpus to provide phrasal translation memory. The method integrates statistical and linguistic information for effective extraction of collocations. The linguistic information includes parts of speech, chunks, and clauses. With an implementation of the method, we obtain first an extended list of collocations from monolingual corpora such as British National Corpus (BNC). Subsequently, we exploit the list to identify English collocations in Sinorama Parallel Corpus (SPC). Finally, we use word alignment techniques to retrieve the translation equivalent of English collocations from the bilingual corpus, so as to provide phrasal translation memory for machine translation system. Based on the strength of chunk and clause analyses, we are able to extract a large number of collocations and translations with much less time and effort than those required by N-gram analysis or full parsing. Furthermore, we also consider longer collocation pattern such as a preposition involved in VN collocation. In the future, we plan to extend the method to other types of collocation.

Keyword. Bilingual Collocation Extraction, Collocational Translation Memory, Collocational Concordancer

1 Introduction

Example-based machine translation (EBMT), a corpus-based MT method, has been recently suggested as an efficient step toward automatic translation (Nagao, 1981; Kitano, 1993, Carl, 1999, Andrimanankasian et al., 1999; Brown, 2000). Under the approach, systems exploited examples similar to input and adjusted the translations to obtain the result. Translations are preprocessed and stored in a translation memory which serves as an archive of existing translation for MT system to reuse. Nowadays, there have been a number of transducers applied to convert sentences in bilingual corpus into translation patterns, which can be further exploited as a translation memory, such as Transit[1], Deja–Vu[2], TransSearch[3], TOTALrecall[4], and so on.

A problem that most MT system may encounter is the collocational translation if the system intends not to literally translate the input text. This smaller syntax unit not only facilitates a more native-like translation, but also enhances the performance of recent EBMT system. Elastic collocation structure provides more flexibility in handle translation pattern as in "…not yet to **take** what he wants **into consideration**…"

[1] *Transit* (http://www.star-group.net/eng/software/sprachtech/transit.html)
[2] *Deja–Vu* (http://www.atril.com/)
[3] *TransSearch* (http://www.tsrali.com/)
[4] *TOTALrecall*(http://candle.cs.nthu.edu.tw/Counter/Counter.asp?funcID=1)

2 Extraction of Collocational Translation Memory

Using valuable linguistic information—chunk and clause analyses, we can retrieve Verb-Noun collocations from a large corpus (i.e. BNC) with good quality and quantity. We further use this collocation type list to identify the concise collocational instances in a bilingual corpus (i.e. SPC). We also use word-alignment technique to extract the matching translation of verb and noun respectively, so as to obtain phrasal translation memory. The detailed approach is described in this section:

2.1 Chunk and Clause Information Integrated

CoNLL-2000[5] shared task considered text chunking as a process that divides a text into syntactically correlated parts of words. With the benefits of chunk information, we can chunk the sentence into smaller syntactic structure which facilitates precise collocation extraction. It becomes easier to identify the argument-predicate relationship between each chunk, and save more time to extract as opposed to full parsing. Take a passage in CoNLL-2000 benchmark for example:

Confidence/B-NP in/B-PP the/B-NP pound/I-NP is/B-VP
widely/I-VP expected/I-VP to/I-VP take/I-VP another/B-NP
sharp/I-NP dive/I-NP if/B-SBAR trade/B-NP figures/I-NP for/B-PP
September/B-NP

Note: I-NP for noun phrase words and I-VP for verb phrase words. Most chunk types have two different chunk tags: B-CHUNK for the first word of the chunk and I-CHUNK for the other words in the same chunk.

The words in the same chunk can be further grouped together (as in Table 1). With chunk information, we can extract the target VN collocation, "take … dive" from the text by considering the last word of each adjacent VP and NP chunks. We built a robust and efficient chunker from the training data of the CoNLL shared task, with over 93% precision and recall[6].

Table 1: Chunked Sentence

Sentence chunking	Features
Confidence	NP
in	PP
the pound	NP
is widely expected to ***take***	VP
another sharp ***dive***	NP
if	SBAR
trade figures	NP
for	PP
September	NP

[5] CoNLL is the yearly meeting of the SIGNLL, the Special Interest Group on Natural Language Learning of the Association for Computational Linguistics. The shared task of text chunking in CoNLL-2000 is available at http://cnts.uia.ac.be/conll2000/.

[6] We built the chunker from shared CoNLL-2000 training data and evaluate the result with the test data provided by CoNLL-2000. The precision and the recall are both 93.7%.

In some cases, only considering the chunk information is not enough. For example, the sentence "…the attitude he had towards the country is positive…" may cause problem. With the chunk information, the system extracts out the type have towards the country as VP + PP + NP, yet this one is erroneous because it cuts across two clauses. To avoid this case, we further take the clause information into account.

With the training data from CoNLL-2001, we built an efficient clause model based on HMM to identify the clause relation between words. The language model provides sufficient information to avoid extracting wrong VN collocation instances. Examples show as follows (additional clause tags will be attached):

(1) ….the attitude *(S** he has ***S)* toward the country

(2) *(S** I think *(S** that the people are most concerned with the question of *(S** when conditions may become ripe. ***S)S)S)*

As a result, we can avoid the verb from being combined with the irrelevant noun as its collocate (as in (1)) or extracting the adjacent noun serving as the subject of another clause (as in (2)). When the sentences in the corpus are preprocessed with the chunk and clause identification, we can consequently assure high accuracy of collocation extraction.

Log-likelihood ratio : LLR(x;y)

$$LLR(x;y) = -2\log_2 \frac{p_1^{k_1}(1-p_1)^{n_1-k_1}(1-p_2)^{n_2-k_2}}{p^{k_1}(1-p)^{n_1-k_1}\ p^{k_2}(1-p)^{n_2-k_2}}$$

k_1 : of pairs that contain x and y simultaneously.
k_2 : of pairs that contain x but do not contain y.
n_1 : of pairs that contain y
n_2 : of pairs that does not contain y
$p_1 = k_1 / n_1$
$p_2 = k_2 / n_2$
$p = (k_1+k_2) / (n_1+n_2)$

2.2 Extraction of Collocation Types

A huge set of collocation candidates can be obtained from BNC, via the process of integrating chunk and clause information. We here consider three prevalent Verb-Noun collocation structures in corpus: VP+NP, VP+PP+NP, and VP+NP+PP. Exploiting Logarithmic Likelihood Ratio (LLR) statistics, we can calculate the strength of association between each two collocates. The collocational type with threshold higher than 7.88 (confidence level 95%) will serve as one entry in our collocation type list.

2.3 Extraction of Collocation Instances

We subsequently identify collocation instances in the Sinorama Parallel Corpus (SPC) matching the collocation types extracted from BNC. Making use of the sequence of chunk types, we again single out the adjacent structures: VP+NP, VP+PP+NP, or VP+NP+PP. With the help of chunk and clause information, we thus find the valid instances where the expected collocation types are located, so as to build a collocational concordance. Moreover, the quantity and quality of BNC also facilitate the collocation identification in another smaller bilingual corpus with better statistic measure.

2.4 Extracting Collocational Translation Equivalents in Bilingual Corpus

When accurate instances are obtained from bilingual corpus, we continue to integrate the statistical word-alignment techniques (Melamed, 1997) and dictionaries to find the translation candidates for each of the two collocates. We first locate the translation of the noun. Subsequently, we locate the verb nearest to the noun translation to find the translation for the verb. We can think of collocation with corresponding translations as a kind of translation memory (shows in Table 2).

Table 2: Examples of collocational translation memory

English sentence	Chinese sentence
If in this time no one shows concern for them, and directs them to correct thinking, and teaches them how to express and release **emotions**, this could very easily leave them with a terrible personality complex they can never resolve.	如果這時沒有人關心他們，引導他們正確思考，教他們表達、渲洩**情緒**，極易在人格成長上留下一個打不開的死結。
Occasionally some kungfu movies may appeal to foreign **audiences**, but these too are exceptions to the rule.	偶爾有一些武打片對某些外國**觀眾**有吸引力，但也是個案。

3 Implementation and evaluation

We extracted VN collocations from the BNC which contains about 4 million sentences, and obtained 631,638 VN, 15,394 VPN, and 14,008 VNP collocation types with an implementation of the proposed method. We continued to identify 26,315VN, 3,457 VPN, and 4,406 VNP collocation instances in SPC and generated eligible translation memory via word-alignment techniques. The implementation result of BNC and SPC shows in the Table 3, 4, and 5.

Table 3:The result of collocation types extracted from BNC and collocation instances identified in SPC

Type	Collocation types in British Nation Corpus (BNC)	Collocation instances in Sinorama Parallel Corpus (SPC)
VN	631,638	26,315
VPN	15,394	3,457
VNP	14,008	4,406

Table 4:Examples of collocation types including a given noun in BNC

Noun	VN types	VN instances
Language	320	945
Influence	319	880
Threat	222	633
Doubt	199	545
Crime	183	498
Phone	137	460
Cigarette	121	379
Throat	86	246
Living	79	220
Suicide	47	134

Table 5:Examples of collocation instances extracted from SPC

VN type	Example
Exert influence	That means they would already be exerting their influence by the time the microwave background was born.
Exercise influence	The Davies brothers, Adrian (who scored 14 points) and Graham (four), exercised an important creative influence on Cambridge fortunes while their flankers Holmes and Pool-Jones were full of fire and tenacity in the loose.
Wield influence	Fortunately, George V had worked well with his father and knew the nature of the current political trends, but he did not wield the same influence internationally as his esteemed father.
Extend influence	The cab extended its influence into the non-government sector, funding research by the Cathedral Advisory Commission and the Royal Society for the Protection of Birds.
Reflect influence	The general standard of farming was good, reflecting the influence of the sons who had attained either a degree or a diploma in agriculture before returning home.
Diminish influence	To break up the Union now would diminish our influence for good in the world, just at the time when it is most needed.
Gain influence	In general, women have not benefited much in the job market from capitalist industrialization nor have they gained much influence in society outside the family through political channels.
Counteract influence	To try and counteract the influence of the extremists, the moderate wing of the party launched a Labour Solidarity Campaign in 1981.
Reduce influence	Whether the curbs on police investigation will reduce police influence on the outcome of the criminal process is not easy to determine.
Show influence	Ellis and Shepherd (1974) first drew attention to this but a number of experiments by Young and his colleagues have failed to show any influence of age of acquisition of words on dichotic listening (Young and Ellis , 1980) or tachistoscopic hemifield asymmetry (Ellis and Young , 1977 ; Young and Bion , 1980b) even when it is the age at which words are first read rather than heard that is under investigation.

As for each collocation type, we randomly selected 100 test sentences for manual evaluation. A human judge, who majored in Foreign Languages, assessed the result of the matching translation. The evaluation was done by judging whether the corresponding collocational translation is valid or not. The three levels of quality were set: satisfactory translation, approximant translation (partial matching), and unacceptable translation. The examples of each level are shown in Table 6.

Table 6: Three levels of quality of the extracted translation memory

Level of quality	English sentences	Chinese sentences
satisfactory translation	Thus when Chinpao Shan put out its advertisement last year, looking for new people to develop its related **enterprises**, the notice frankly stated "Southern Taiwanese preferred.	去年，金寶山在發展**關係企業**徵招新人的廣告上，就坦白指明「本省籍南部人優先」。
approximant translation	Ah-ying relates that "Teacher Chang" friendly and easy-going, is always there to answer her **questions**. She even goes to him for answers when her friends have legal questions.	阿英表示，「張老師」親切隨和，只要有不懂的事，都去問老師，就連朋友有法律上的**問題**，也去請教他。
unacceptable translation	Said one observer, "If I can speak bluntly, the mainlanders are robbing graves of their treasures and smuggling them away, and the situation is bad. In reality, though, it is Taiwan that is behind it all committing the **crime**.	「說得不好聽，大陸近年來盜墓、文物走私情形嚴重，台灣其實是背後的劊子手！」有人這樣**認為**。

The evaluation result indicates an average precision rate of 89 % with regard to both satisfactory and approximant translation memory (shows in Table 7).

Table 7: Experiment result of collocational translation memory from Sinorama parallel Corpus

Type	The number of selected sentences	Translation Memory	Translation Memory (*)	Precision of Translation Memory	Precision of Translation Memory (*)
VN	100	73	90	73	90
VPN	100	66	89	66	89
VNP	100	78	89	78	89

The average precision of translation memory: 72.3%
The average precision of translation memory (*): 89.3%

(*) stands for the numbers of translation memory which includes approximant translation.

4. Discussion and limitation

Collocation, a hallmark of near native speaker, is an important area in translation yet has long been neglected. Traditional machine translation tends to translate input texts word by word, which easily leads to literal translation. Therefore, even with abundant vocabulary from dictionary and grammar rule-based model systems still fail to generate fluent translation into a target language. For example, with the lack of collocational knowledge, machine translation system may recognize take as “na” (i.e. take away) and medicine as “yao” (i.e. medicine) in Chinese respectively. Thus, systems are inclined to literally translate take medicine into “na yao” (i.e. take away the medicine), and probably result in odd translation or mistranslation. We suggest that machine translation system take collocational translation memory into consideration for improved translation quality. The notion of collocation is also consistent with Example-Based Machine Translation (EBMT).

Due to the limitation of word-alignment technique, our method may incorrectly recognize some matching translation. We need better word-alignment to align translations more correctly. Moreover, the expansion of bilingual corpora can also increase the precision of retrieving collocational translation memory. It enables us to obtain enough counts for each collocate (i.e. verb and noun in VN collocation) in the target language so as to increase the reliability with the LLR statistics, which in turn eradicates the anomalous collocational translation memory.

5. Application: Collocational Concordance—TANGO

With the collocation types and instances extracted from the corpus, we built an on-line collocational concordance called TANGO for looking up collocation instances and translations. A user can type in any English words as query and select the expected part of speech of the accompanying words. For example in Figure 1, after query “influence” is submitted, the result of possible collocates will be displayed on the return page. The user can even select different adjacent collocates for further investigation. Moreover, using the technique of bilingual collocation alignment and sentence alignment, the system will display the target collocation with highlight to show translation equivalents in context. Translators or learners, through this web-based interface, can easily acquire the usage of each collocation with relevant instances. This bilingual collocational concordance is a very useful tool for self-inductive learning tailored to intermediate or advanced English learners.

Figure 1 Web-based Collocational Concordance

6. Conclusion

In the field of the machine translation, the Example-Based Machine Translation (EBMT) exploits existing translations in the hope of producing better quality in translation. However, the importance of collocational translation has always been neglected and hard to be dealt with. We propose the collocational translation memory — to provide a better translation method, intending to solve some problem encountered by literal translation. With satisfactory precision rates of collocation and translation extraction, we hope collocational translation memory will path ways to more applications in translation and computer assisted language learning.

Acknowledgements

This work is carried out under the project “CANDLE” funded by National Science Council in Taiwan (NSC92-2524-S007-002). Further information about CANDLE is available at http://candle.cs.nthu.edu.tw/.

Reference

[1] Andriamanankasina, T., Araki, K. and Tochinai, T. 1999. Example-Based Machine Translation of Part-Of-Speech Tagged Sentences by Recursive Division. Proceedings of MT SUMMIT VII. Singapore.

[2] Brown, R. D. 2000. Automated Generalization of Translation Examples. In Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000), pp. 125-131. Saarbrücken, Germany, August 2000.

[3] Carl, M. 1999. Inducing Translation Templates for Example-Based Machine Translation, Proc. of MT Summit VII.

[4] Melamed, I. D. 1997. A Word-to-Word Model of Translational Equivalence. Proc. of the ACL97. pp 490-497. Madrid Spain, 1997.

[5] Kitano, H. 1993. A Comprehensive and PracticalModel of Memory-Based Machine Translation. Proc.of IJCAI-93. pp. 1276-1282.

[6] Nagao, M. 1981. A Framework of a MechanicalTranslation between Japanese and English byAnalogy Principle, in Artificial and Human Intelligence, A. Elithorn and R. Banerji (eds.) North-Holland, pp. 173-180, 1984.

A New Two-Layer Approach for Spoken Language Translation

Jhing-Fa Wang, Shun-Chieh Lin, and Hsueh-Wei Yang

Department of Electrical Engineering, National Cheng Kung University
wangjf@csie.ncku.edu.tw

Abstract. This study proposes a new two-layer approach for spoken language translation. First, we develop translated examples and transform them into speech signals. Second, to properly retrieve a translated example by analyzing speech signals, we expand the translated example into two layers: an intention layer and an object layer. The intention layer is used to examine intention similarity between the speech input and the translated example. The object layer is used to identify the objective components of the examined intention. Experiments were conducted with the languages of Chinese and English. The results revealed that our proposed approach achieves about 86% and 76% understandable translation rate for the Chinese-to-English and the English-to-Chinese translations, respectively.

1 Introduction

With the growing of globalization, people now often meet and do business with those who speak different languages, on-demand spoken language translation (SLT) has become increasingly important (See JANUS III [6], Verbmobil [9], EUTRANS [3] and ATR-MATRIX [1]). Currently, there are two main architectures of SLT: conventional sequential architecture and fully integrated architecture [1]. For the sequential architecture, a spoken language translation is composed by a speech recognition system followed by a linguistic (or non-linguistic) text-to-text translation system. In the integrated architecture, acoustic-phonetic models are integrated into translation models in the similar way as for speech recognition.

Recently, an integrated architecture based on stochastic finite-state transducer (SFST) has been presented in [3,4]. The SFST approach integrated three models in a single network where the search process takes place. The three models are Hidden Markov Models for the acoustic part, language models for the source language and finite state transducers for the transfer between the source and target language. The output of this search process is the target word sequence associated to the optimal path. Fig. 1 shows an example of the SFST approach. λ denotes the empty string. The source sentence "*una habitación doble*" can be translated to either "*a double room*" or "*a room with two beds*". The most probable translation is the first one with probability of 0.09.

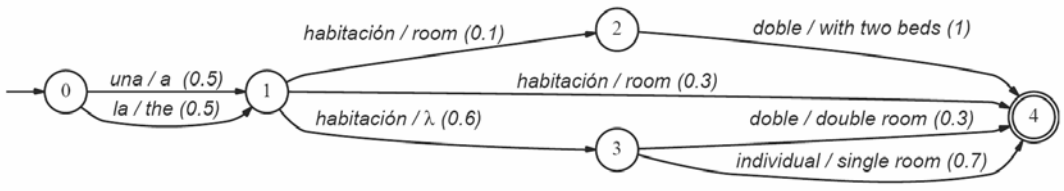


Fig. 1. Examples of the stochastic finite-state transducer

However, when the training data of SFST is insufficient, the results obtained by the sequential architecture are better than the results obtained by the integrated architecture [4]. In addition, word reordering is still a thorny problem in SFST which is based on statistical-based translation methods [5]. Therefore, we propose adopting example-based approaches for better integration. Such the adopted approach does not require the database to be as large as in SFST and can utilize word mappings between source-target language of a chosen translated example for word reordering [2,8]. In this paper, we further propose a new two-layer approach for the example-based spoken language translation. First, we develop translated examples and transform them into speech signals. Second, to properly retrieve a translated example by analyzing speech signals, we expand the translated example into two layers: an intention layer and an object layer. The intention layer is used to examine intention similarity between the speech input and the translated example. The object layer is used to identify the objective components of the examined intention.

The rest of this paper is organized as follows. Section 2 discusses the proposed two-layer approach. Score normalization is presented in Section 3. The experimental results are given in Section 4. Concluding remarks are finally made in Section 5.

2 The Proposed Two-Layer Approach

Referring to Fig. 2, the first step of the proposed two-layer approach is to expand translated examples, which have intention components and object components. After expanding the translated examples, the second step is to adapt the two-layer search plan composed of an intention layer and an object layer. At last, measurement modification is used to modify similarity measurement between the intention layer and the object layer. This study further discusses *translated example expansion*, *two-layer search plan adaptation*, and *measurement modification*.

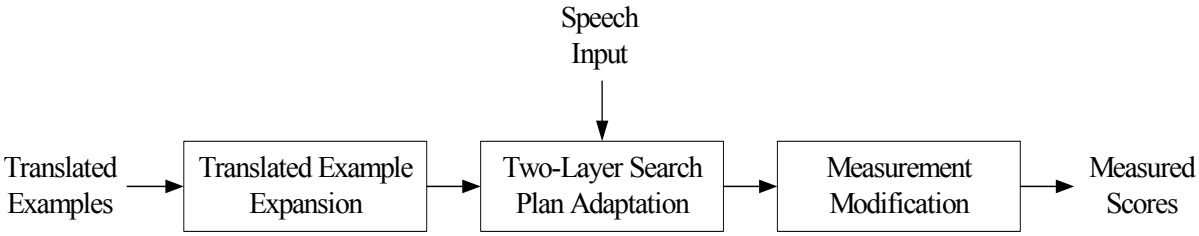


Fig. 2. Framework of the proposed two-layer approach

2.1 Translated Example Expansion

The process of translated example expansion is to group similar translated examples and compare their differences for expanding objects. Table 1 shows an example of fours pairs of grouped translated examples. For these grouped translated examples, the similar constitutes "*Is … still available for …*" ↔ "… 還 有 … 嗎" are defined into an intention sequence translation, which would conduct the meaning of a translation. And the differences compared with the intention sequence are regarded as expanded objects.

Table 1. Fours pairs of grouped translated examples

	Translated examples	Word mappings
1	Is room service still available? ↔ 還 有 客房 服務 嗎?	⟨Is↔嗎, room↔客房, service↔服務, still↔還, available↔有⟩
2	Is breakfast available for tomorrow? ↔ 明天 有 早餐 嗎?	⟨Is↔嗎, breakfast↔早餐, available for↔有, tomorrow↔明天⟩
3	Is laundry service still available? ↔ 還 有 洗滌 服務 嗎?	⟨Is↔嗎, room↔洗滌, service↔服務, still↔還, available↔有⟩
4	Is a single room available for tonight? ↔ 今晚 有 一間 單人房 嗎?	⟨Is↔嗎, a↔一間, single↔單人, room↔房, available for↔有, tonight↔今晚⟩

For example, a new expanded translated example, denoted by *ExTrans*, derived from the translated examples in Table 2 is shown below.

Table 2. An example of expanded translated example

Expanded translated example: *ExTrans*	
The intention sequence translation:	
Is ⟨V^1⟩ still available for ⟨V^2⟩ ?	
↔⟨V^3⟩ 還 有 ⟨V^4⟩ 嗎?	
where ⟨V^1⟩=⟨room service, breakfast, laundry service, a single room⟩,	
⟨V^2⟩=⟨tomorrow, tonight⟩,	
⟨V^3⟩=⟨客房 服務, 早餐, 洗滌 服務, 一間 單人 房⟩,	
⟨V^4⟩=⟨明天, 今晚⟩	
Object translations:	
⟨V^1⟩↔⟨V^4⟩	⟨V^2⟩↔⟨V^3⟩
⟨room,service⟩↔⟨客房,服務⟩	⟨tomorrow⟩↔⟨明天⟩
⟨breakfast⟩↔⟨早餐⟩	⟨tonight⟩↔⟨今晚⟩
⟨laundry,service⟩↔⟨洗滌,服務⟩	
⟨a,single,room⟩↔⟨一間,單人,房⟩	

where *ExTrans* comprises an intention translation, and six object translations. The six object translations are "*room service*↔客房 服務," "*breakfast*↔早餐," "*laundry service*↔洗滌 服務," "*a single room*↔一間 單人房," "*tomorrow*↔明天," and "*tonight*↔今晚".

2.2 Two-Layer Search Plan Adaptation of Expanded Translated Examples

After expanding translated examples, each translated example has two parts: an intention part and an object part. While measuring the speech signals of *i-th* translated example v_i, the speech signals of v_i need to be redefined two layers $v_i = \{v_i', v_i''\}$, where v_i' is an intention layer component of v_i and v_i'' is an object layer component of v_i. Each two-layer searching plan is generated by the translated example and the speech input and the object layer is used to identify the objective components of the examined intention. In terms of searching for an optimal path of states through the two-layer search plan, the issue now is to measure the pair (s, v_i') of a fixed number, says N_i, of v_i'.

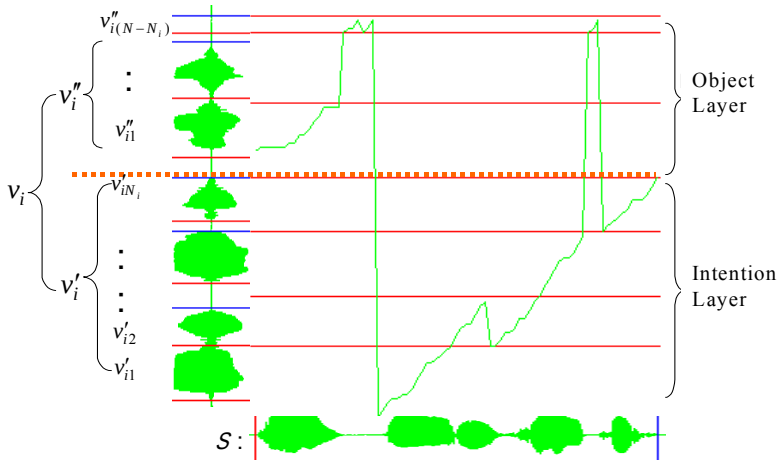


Fig. 3. The proposed two-layer search plan

2.3 Measurement Modification

After adapting the two-layer search plan, another problem is how to measure the similarity of pair (s, v'_i) while adjudging the object frames of v''_i for identifying the other object patterns. Referring to Fig 4., given two similarity measurement scores of pair (s, v_i) and pair (s, v_j), the scores used for comparing the two pairs are D'^{*}_i and D'^{*}_j, where D'^{*}_i is the similarity measurement of pair (v'_i, s) and D'^{*}_j is the similarity measurement of pair (v'_j, s).

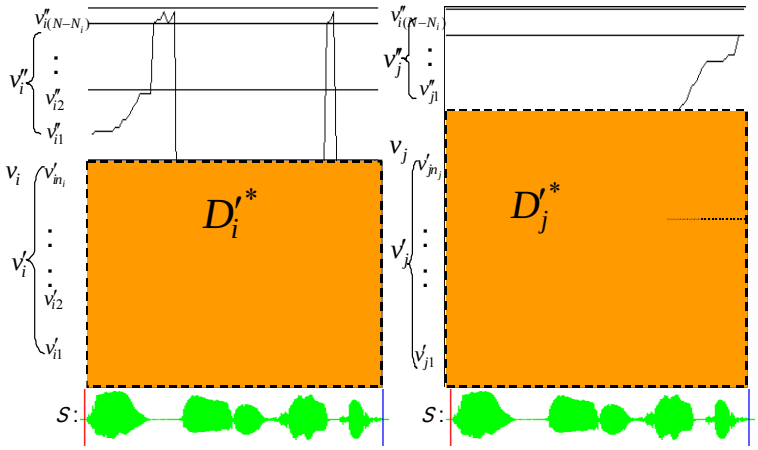


Fig. 4. Search results of various translated examples

For the modification of similarity measurement between the intention layer and the object layer, there are two additional types of search paths in this research: 1) paths between v'_i and v''_i and 2) paths within v'_i or v''_i. For the paths between v'_i and v''_i, a search block Z in the object layer, which will be referred to a score skip level block, contains more than one path connected by $node_{start}$ (or $node_{end}$). And D'^{*}_i is computed in the intention layer. (See Fig. 5)

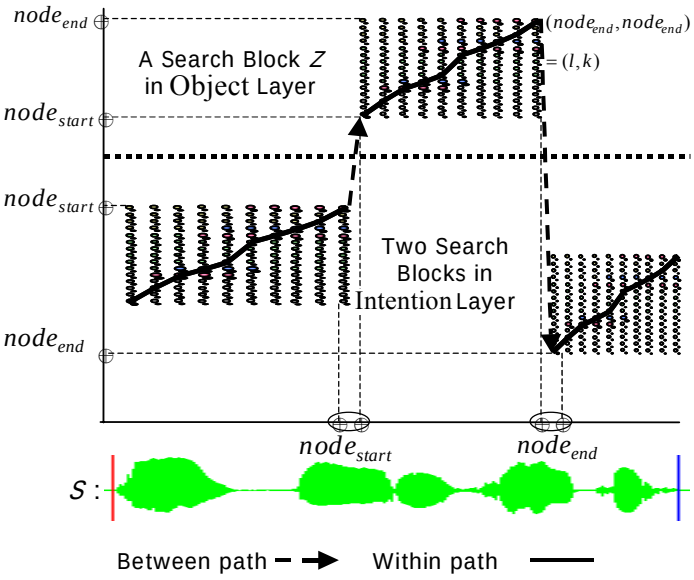


Fig. 5. Additional types of two-layer search paths

3 Score Normalization

The intention sequence in the translated example is an important identification part, where the intention sequence would conduct the meaning of a translation. Therefore, the dissimilarity measurement of the part of the intention sequence is used to rank all the translated examples. However, the cumulative measured dissimilarity score is propagated to the length of the intention sequence. In this study, a length-conditioned weight concept is adopted to compensate this defect. The normalized measured dissimilarity ($\Delta(s, v_i')$) is determined as follows:

$$\Delta(s, v_i') = \partial^{w_{v_i,s}} \tag{1}$$

where ∂ is a weight factor, $w_{v_i',s} = \left(\left\|\left\|v_i'\right\| - \left\|s\right\|\right\|\right) \cdot \left\|s\right\|^{-1}$. The weight ∂ is decided by an interval [1.0, 2.0]. Experimental analysis shown in Fig. 6 indicates that the interval ∂, which yields the most accurate retrieval results, is $[1.3-\delta, 1.3+\delta]$. Therefore, the ∂ is set to 1.3 in this study.

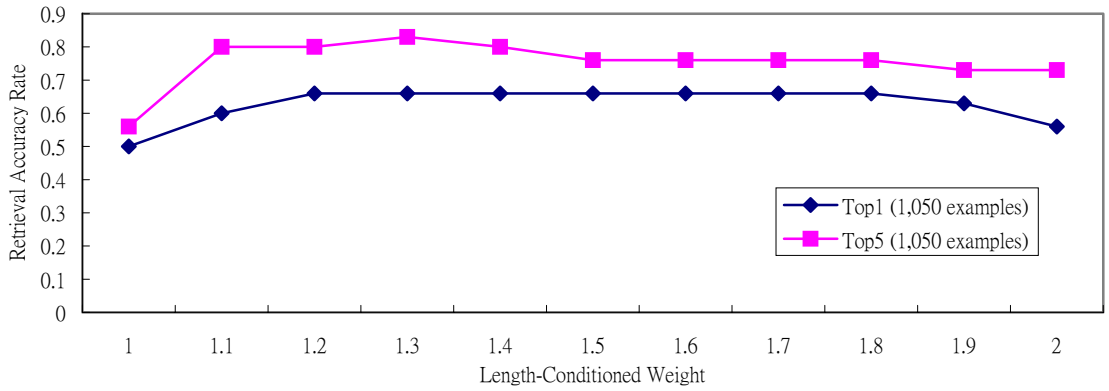


Fig. 6. Retrieval accuracy rate comparison with different setting of weight ∂

4 Experimental Results

4.1 The Task and the Corpus

This study built a collection of English sentences and their Chinese translations that frequently appear in phrasebooks for foreign tourists. Because the translations were made on a sentence-by-sentence basis, the corpus was sentence-aligned after being collated. Table 3 lists a summary of the corpus used in the experiments. The corpus comprises two parts: a training set of 11,885 translated examples for the training phase, and a test set of 105 translated examples for the translation phase (the test set differs from the training set).

Table 3. Basic characteristics of the collected translated examples

		English	Chinese
Training:	Translated Examples	11,885	
	Lexicons	80,699	66,915
	Vocabulary Size	6,278	5,118
	Average number of lexicons	6.79	5.63
Test:	Sentences	105	
	Lexicons	673	641

In order to evaluate the system performance, a collection of 1,050 utterances from the 11,885 examples were speaker-dependent trained, and 105 additional utterances of each language were collected by using one male speaker (Sp1) for inside testing and by using two bilingual male speakers (Sp2 and Sp3) for outside testing. All the utterances were sampled at an 8 kHz sampling rate with 16-bit precision on a Pentium® IV 1.8GHz, 1GB RAM, Windows® XP PC.

4.2 Translation Evaluations

For the spoken language translation system, we found that the recognition performance of 39-dimension MFCCs and 10-dimension LPCCs was close. Therefore, we adopted 10-dimension LPCCs due to their advantages of faster operation. Speech feature analysis of recognition was performed using 10 linear prediction coefficient cepstrums (LPCCs) on a 32ms frame that overlapped every 8ms.

When input speech is being translated, a major sub-problem in speech processing is determining the presence or absence of a voice component in a given signal, especially the beginnings and endings of voice segments. Therefore, the energy-based approach, which is a classic one and works well under high SNR conditions, was applied to eliminate unvoiced components in this research. The measurement results were divided into four parts: the dissimilarity measurement of linear prediction coefficient cepstrum (LPCC)-based (baseline), the baseline with unvoiced elimination (+unVE), the baseline with the score normalization (+ScN), and the combination of unVE and ScN considerations with the baseline (All). A given translated example is called a match when it contained the same intention as the speech input. The reason for adopting this strategy was that objects could be confirmed again while a dialogue was being processed, while wrong intentions could cause endless iterations of dialogue. The experimental results for proper translated example retrieval are shown in Table 4 and Table 5.

Table 4. Average retrieval accuracy of baseline and the improvement in English-to-Chinese(E2C) Translation

	1		2		3		4	
	Baseline		+unVE		+ScN		All	
Example Size	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
150	0.53	0.66	0.63	0.86	0.66	0.86	0.8	1
250	0.53	0.66	0.63	0.86	0.66	0.86	0.8	1
350	0.53	0.63	0.6	0.83	0.66	0.86	0.76	0.96
450	0.53	0.63	0.6	0.83	0.63	0.83	0.76	0.93
550	0.5	0.6	0.6	0.8	0.6	0.8	0.76	0.93
650	0.5	0.56	0.6	0.76	0.6	0.8	0.76	0.9
750	0.46	0.5	0.56	0.73	0.56	0.76	0.73	0.86
850	0.43	0.5	0.53	0.7	0.53	0.73	0.73	0.83
950	0.43	0.46	0.53	0.7	0.5	0.66	0.7	0.83
1050	0.4	0.43	0.46	0.66	0.46	0.66	0.66	0.8

Table 5. Average retrieval accuracy of baseline and the improvement in Chinese-to-English(C2E) Translation

	1		2		3		4	
	Baseline		+unVE		+ScN		All	
Example Size	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
150	0.46	0.6	0.63	0.8	0.6	0.76	0.76	1
250	0.46	0.6	0.6	0.76	0.6	0.73	0.76	0.96
350	0.46	0.56	0.6	0.76	0.56	0.7	0.73	0.93
450	0.43	0.56	0.56	0.73	0.53	0.66	0.7	0.9
550	0.43	0.53	0.56	0.7	0.53	0.63	0.7	0.86
650	0.43	0.53	0.53	0.7	0.5	0.6	0.66	0.83
750	0.4	0.5	0.53	0.66	0.5	0.6	0.63	0.8
850	0.4	0.5	0.5	0.66	0.46	0.56	0.63	0.8
950	0.4	0.46	0.46	0.63	0.46	0.56	0.6	0.76
1050	0.36	0.43	0.46	0.6	0.43	0.56	0.6	0.7

Based on the developed translated examples, when the example or vocabulary size increases, more examples would possibly lead to more feature models and more similarities in speech recognition, thus causing false recognition results and lower retrieval accuracy. Additionally, multiple speaker dependent results were obtained using three speakers. The first speaker's feature models were used to perform tests on the other two speakers, and the results are shown in Table 6. The experimental results show that although the feature models were trained by Sp1, the retrieval accuracy of Sp2 and Sp3 was only reduced by 10 to 15 percent.

Table 6. Average retrieval accuracy in multiple speaker testing

			Example Size (Speech features of Sp1)									
		(Top5)	150	250	350	450	550	650	750	850	950	1050
All	Sp1	E2C	1	1	0.96	0.93	0.93	0.9	0.86	0.83	0.83	0.8
		C2E	1	0.96	0.93	0.9	0.86	0.83	0.8	0.8	0.76	0.7
	Sp2	E2C	0.9	0.86	0.83	0.8	0.76	0.73	0.73	0.7	0.66	0.66
		C2E	0.83	0.83	0.8	0.76	0.73	0.73	0.7	0.66	0.63	0.63
	Sp3	E2C	0.83	0.8	0.76	0.76	0.73	0.7	0.7	0.66	0.66	0.63
		C2E	0.76	0.76	0.73	0.73	0.7	0.66	0.66	0.63	0.6	0.6

A bilingual evaluator was used to classify the target generation results into three categories [10]: Good, Understandable, and Bad. A Good generation needed to have no syntactic errors, and its meaning had to be correctly understood. Understandable generations could have some syntactic errors and variable translation errors, but the source speech had to be conveyed without misunderstanding. Otherwise, the target generations were classified as Bad. With this subjective measure, the percentage of Good or Understandable generations for the Top 5 was 86% for English-to-Chinese (E2C) translation and 76% for Chinese-to-English (C2E) translation. The percentage of Good generations for the Top 1 was 60% for E2C translation, compared to 56% for C2E translation. We examined the translated examples in a specific domain and found that 100% translation accuracy could be achieved. In other words, translation errors occurred only as a result of speech recognition errors, such as word recognition errors and segmentation errors. Besides, these results also indicate that C2E performed worse than E2C. This difference may occur because Chinese is tonal, whereas English is not; thus, it is harder for C2E translation to obtain an appropriate translated example.

5 Conclusions

In this work, we have proposed a new two-layer approach for example-based spoken language translation. According to the proposed approach, the translated example can be properly retrieved by measuring the speech signals on the intention layer and the object layer. Experiments using Chinese and English were performed on Pentium® PCs. The experimental results reveal that our system can achieve an average understandable translation rate of about 81%. By collecting more speech databases, the system also applies speaker-dependent or speaker-independent HMM to the proposed two-layer approach for more robust speech translation.

References

[1] ATR Spoken Language Translation Research Laboratories research, http://www.slt.atr.co.jp/

[2] M. Carl. Inducing Translation pattern for Example-Based Machine Translation. In *Proc. of the 7th Machine Translation Summit*, pp.617–624, 1999.

[3] F. Casacuberta, D. Llorens, C. Martinez, S. Molau, F. Nevado, H. Ney, M. Pastor, D. Pico, A. Sanchis, E. Vidal, J. M. Vilar. Speech-to-Speech Translation Based on Finite-State Transducers. In *Proc. of 26th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.613 –616, 2001.

[4] E. Vidal. Finite-State Speech-to-Speech Translation. In *Proc. of 22nd IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.111-114, 1997.

[5] H. Ney, S. Nießen, F. J. Och, H. Sawaf, C. Tillmann, and S. Vogel. Algorithms for Statistical Translation of Spoken Language. *IEEE Transaction on Speech and Audio Processing*(8), pp.24-36, 2000.

[6] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld and P. Zahn. JANUS III: Speech-to-Speech Translation in Multiple Languages. In *Proc. of 22nd IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.99–102, 1997.

[7] Rabiner, L. and B. H. Juang. Fundamentals of Speech Recognition. Prentice-Hall, Inc., 1993.
[8] J. Liu and L. Zhou. A hybrid model for Chinese-English machine translation. In *Proc. of IEEE International Conference on Systems, Man, and Cybernetics*, pp.1201-1206, 1998.
[9] Wahlster, W. Verbmobil: Foundations of Speech-to-Speech Translation. New York: Springer-Verlag Press, 2000.
[10] K. Yamabana, K. Hanazawa, R. Isotani, S. Osada, A. Okumura and T. Watanabe. A Speech Translation System with Mobile Wireless Clients. In *Proc. of the Student Research Workshop at the 41st Annual Meeting of the Association for Computational Linguistics*, pp.119–122, 2003.

ROCLING XVI
結合統計與語言訊息的混合式中英雙語句對應演算法

林語君　　高照明

台灣大學資訊工程學系　　台灣大學外國語文學系

sbbjun@gmail.com　　zmgao@ntu.edu.tw

摘要. 本文結合包括句長、標點符號、數字時間詞、原文詞，雙語辭典等各種語言訊息，以動態規劃演算法，找出群組句對應，再階段式的運用以上應用，將所得到的多句對多句的對應作更細微的分割。根據評估，運用兩階段式的動態規劃演算法，再加上以上語言訊息，可以達到近95%的召回率（recall）的情況下，達到80%以上的精確率（precision）。

1　導言

自然語言處理研究中，機器翻譯是其中一項重要主題，從50年代開始研究就從未間斷，其中固然有小規模發展，但機器翻譯的整體效果始終沒有太大的突破。這樣的瓶頸讓研究者得到了一個打破現有思考的結論：長期以來由語言結構分析以及人工建置的翻譯模組的方式，要完全掌握所有自然語言翻譯特徵，有技術上以及複雜性上的困難。90年代左右，發展方向轉爲從一個龐大的雙語語句對應資料庫中，搜尋與擬翻譯語句相關連的雙語翻譯句對。再從這些句對當中以自動或半自動的方式得到翻譯知識與規則。這樣的架構能得出較爲自然、較具延展性的翻譯語句，使機器翻譯的發展出現了另一條可能方向。

在越來越多的研究學者肯定平行語料庫（Parallel Corpus）在機器翻譯上的潛在價值下，伴隨著幾個重要的研究議題，首先是雙語句對資料庫的建立。這個資料庫必須包含各個領域、必須具有足夠龐大的資料量。其中平行語料庫又有不同層次的對應單位，尤以句對應（Sentence Alignment）爲主要的可應用（Utilizable）對應，可以說是最主要的機器翻譯參考結構。平行語料庫往下可以繼續往細部擷取出詞組對應及詞對應，作爲機器翻譯上更小的翻譯單元。

但句對應的平行語料庫得來不易，人工對應不僅昂貴，資料量的不足更會直接影響到應用平行語料庫的品質。本系統目標在於大量且自動的取得中英句對應的平行語料庫，除了綜合至目前爲止的句對應模型（Sentence Alignment Model）之外，更針對中文以及英文的特殊屬性做不同參數以及演算法的調整，如中英句長、標點符號、數字時間詞、原文詞，再以各種不同的英漢／漢英字典以及Hownet等各模組的互相配合，以動態規劃演算法以及更進一步的遞迴階段分割方式產生較小的句群對應，並在縮小對應的區塊時盡量兼顧對應的正確率。

1.1　文獻回顧

雙語句對應的研究開始於90年代初期。Gale 與Church (1991) 及Brown 等 (1991) 觀察到長句的翻譯對應句一般而言較長，而短句的翻譯句通常較短。他們利用句長的關連性配合動態規劃或EM演算法得到96%以上的正確率。Gale 與Church (1991) 及Brown 等 (1991)兩者最大的差別是前者透過人工先得到先驗機率（prior probability）而後者利用EM演算法得到相關的參數。Wu (1994) 及Xu and Tan (1996)以句長爲主結合一個包含日期及數字等訊息小的辭典得到96%的正確率。以句長爲基礎的統計方法的優點是不需要語言知識及辭典就可以運作。缺點是如果語料中含有豐富的多對多的句對應關係，或是翻譯的語料中有增添或刪減的現象發生就會造成正確率大幅下降。前述幾項研究由於大都採用議會的紀錄，例如Gale與Church (1991) 及Brown 等（1991）用加拿大國會Hansard英法平行語料，Wu (1994)則利用香港立法局議會質詢與答詢的中英平行語料，由於是口語紀錄所以句子較短，且不少是一對一對應。Gale 與Church (1991)統計Hansard語料80%以上是一對一的對應關係，罕有多對多的對應關係或增添或刪減的情形發生，所以以句長爲主的統計方法得到很好的效果。但McEnery and Oakes (1996)以Gale 與Church (1991)

的方法做實驗卻顯示此種演算法的正確率對不同的文類與語言會產生很大的差異。例如波蘭文英文平行語料的正確率因文類不同介於於100%與64.4%，而他們所實驗的中英新聞平行語料更低於55%,這證明單純以句長關連性顯然無法得到高正確率。

另一個不需要辭典的方法是Kay and Röscheisen (1993) 以詞彙的頻率（去除低頻的詞及高頻的詞）及在文章中出現的分佈，建立可能的詞對應表及句對應表並不斷的修正，以relaxation方法達到收斂。與Gale 與Church (1991) 及Brown 等 (1991)方法一樣，Kay and Röscheisen (1993)的方法只有在一對一的情形佔絕大多數時才會有好的效果。此外此種方法過度重視詞頻，文章的長度太短會造成正確率的大幅下降。這個演算法另一個實做上的問題是處理十分耗時，無法快速處理大量語料。

另外Melamed (1997a)提出Smooth Injective Map Recognizer (SIMR) 利用統計和同源詞(cognate)，正確率高於Gale 與Church (1991)，但我們以光華雜誌做初步實驗發現正確率仍然只有60%左右。

以統計爲主的方法不管是以長度，詞頻及詞彙內部分佈，或geometric，在正確率及強健性方面似乎都不理想，因此使用雙語辭典似乎是提昇正確率所必需，但如果只以雙語辭典找句對應效果也不理想，原因是翻譯的基本單位在很多時候並不是詞，而是詞組或結構，因此Catizone et al. (1989)提出結合辭典與統計訊息。Haruno and Yamazaki (1996)比較純統計式，辭典，與混合式三種方法，發現混合式在精確率precisionz 召回率recall介於91.6% 到 97.1%之間，比採純統計式或只用辭典的方式好。Utsuro 等 (1994)也採取辭典與句長爲主的混合法，但錯誤率介於4.6% and 21.6%。顯示即使採用混合法正確率也隨著語料的不同與演算法細節的不同而有相當大的差異。

語言訊息除了雙語辭典還有其它的訊息可以用來找句對應，例如Yeh (2002)發現在光華中英平行語料中標點的訊息有助於找到句對應。本文的目的在於探討混合式的句對應演算法包含哪些訊息及這些訊息應該如何組合起來才能達到最佳的句對應效果。

2 系統架構與流程

我們的系統大體架構如圖1所示。輸入雙語平行語料，最後輸出其中的雙語群組句對（一個群組句對可爲一句對一句／一句對多句／多句對多句的雙語組合，但是不包含零句對一句或零句對多句的情況，也就是說所有的句子都會被納入某個群組句裡）。

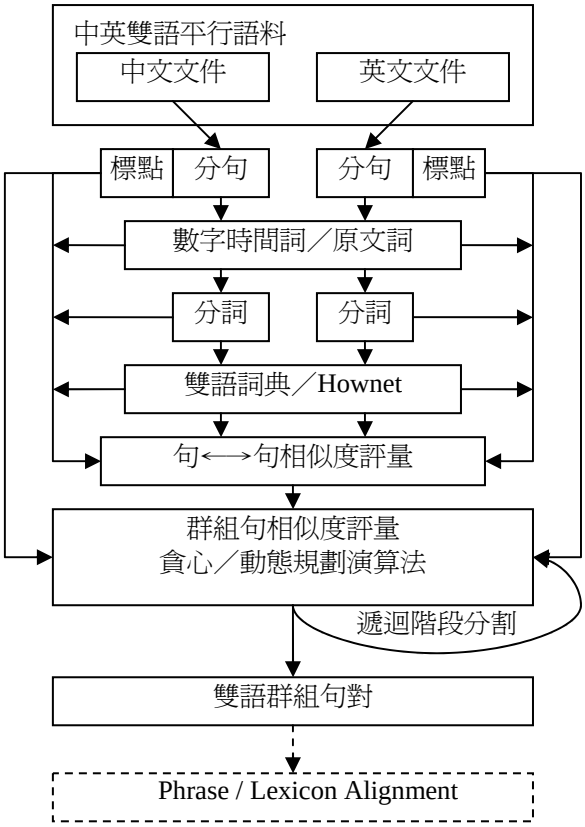


圖. 1. 系統架構與流程

2.1 分句

中文句子的界線並不清楚，句點與逗點都有可能是句子的界線。如Gao (1997)所指出句點事實上是一個比句子還要大的言談單位(discourse)，裡面可能包含數個句子，而逗點有時是一個句子，有時只是詞組。我們的分句主要將原始文件切割成最小的句對應處理單位，但因句對應模組所採用的是可以接受句子多對多的群組句對應演算法，所以並不一定要對分句做太嚴格的處理，也就是說，錯誤的分句仍可以在之後的群組句對應演算法中，有機會與前後語句結合，成爲一個適當的「句單位」。

因仰賴之後的多對多句對應演算法，最佳的句分割就成爲「在句子不會失去完整的意義情況下，盡量地細微分割。」因此，實作上就成了以下原則：

一、 以標點符號作爲句子爲單位，將句號、問號、驚嘆號、分號前後的區塊拆分開來，但在英文方面需注意是否該句點爲單字的一部份，如Mr.等之縮寫。英文句的切割我們使用Shlomo Yona Lingua::EN::Sentence的Perl模組（http:// search.cpan.org / ~shlomoy / Lingua-EN-Sentence-0.25 / lib / Lingua / EN / Sentence.pm）。

二、 如果有兩個太長的句子（兩個超過十個中文詞或六個英文單字的句子）以逗號連結起來，則視爲兩個句子。

三、 另外，對一些對句對應有幫助的符號做特殊的處理，如將發言句、引用句等的前後冒號及引號作爲單獨的句子（一個句子僅擁有幾個符號），如此可以強制句對應演算法將這些有特殊意義的句子視爲「錨」（Aligning anchor），賦予極高的對應權重，作爲一個重要的對應參考點。

2.2 分詞與辭典查詢

因中英文的詞組結構迥異，故分別採用不同的分詞方法。在這裡我們一樣不採用嚴格的分詞法，因分詞的主要目的爲取得該句的對應句中，所有可能的對應翻譯詞，以供之後的句與句之間相似度的衡量，所得出來的相似度衡量值爲一相似概值，故分詞的準確度，並不會大幅度的影響分句的結果。

2.2.1 分詞

英文方面因爲有比較明顯的分詞符號（空白），比較沒有分詞上的困難。

在中文分詞問題方面，我們採用的是「所有可能翻譯詞」。舉「把他的確實行動作了分析」這個句子爲例，「的確」、「實行」、「動作」、「分析」，以及跟上述分詞有重疊的「確實」、「行動」，因爲在漢英字典裡有對應翻譯，故全部予以採用（但如其中的高頻率詞不予採用，述於下節）。

2.2.2 字典查詢

分詞的目的在於找到對應翻譯詞，故可同時採用查詢不同的多個字典，如一般辭典以及Hownet裡的特殊辭典。但要如何避免因爲擴張翻譯詞而造成的錯誤碰撞（false hit）呢？我們採用一個頻率辭典來完成，頻率高於某值的詞被視爲stop word，不列入翻譯對應詞的衡量。

對於英文的翻譯詞查詢，我們利用stemming得到詞的原型，在此我們採用Ron Savage的Lingua-EN-Infinitive之Perl模組（http://search.cpan.org / ~rsavage / Lingua-EN-Infinitive-1.08 / lib / Lingua / EN / Infinitive.pm）。

2.3 雙語句之間的相似度評量

這個階段將分句階段的中文各句以及英文各句，連同各句包含的詞對應，以及各句句中的標點符號，數字詞，原文詞，一起作爲雙語句關連評量的參數，計算出一個任意句對關連值來。這些句對關聯值（一個表）將爲下一階段的群組句對相似度評量的最重要參數。

2.3.1 標點符號

這裡的標點符號並非單指句末的標點符號，而是句中的「所有重要標點符號」所形成的序列（在這裡我們將非逗號及非句號之標點符號視爲「重要」），句與句之間的標點符號相似度即爲這個序列的相似度，實作方法爲：

$$P_i = \{p_{i1}, p_{i2}, p_{i3}, \ldots\ldots, p_{in}\}$$
$$Q_j = \{q_{j1}, q_{j2}, q_{j3}, \ldots\ldots, q_{jm}\}$$
$$punc\_sim_{i,j} = \mathrm{LCS}(P_i, Q_j) \times W_{punc\_sim}$$

以上P_i代表句子$_i$中的標點符號，以Q_j代表句子$_j$中的標點符號，按照出現順序所排列而成的標點符號序列。LCS函數爲最長共同子序列（Longest Common Sequence）。W_{punc_sim}爲標點符號相似度參與句與句相似度衡量的比重。最後，$punc_sim_{i,j}$爲這兩個序列的相似度的加分值。

2.3.2 數字詞與時間詞

數字詞與時間詞爲跨語言中的重要共通語意部分，字典往往沒有這些詞。在這裡，我們額外地從中英文句對中，偵測並抽取出數字以及時間的資料，如果有相同的數字以及時間記錄，則將該句對設定爲一個極高權值的對應「錨」。但，太簡單的數字不予考慮，在這裡我們忽視1, 2, 3三個對應中英文。

2.3.3 原文詞

另外一個更具高權值對應的部分爲原文詞，這種在譯文中保留原始語言文字的特性極常出現，一旦偵測到即有極高的可信度。我們的作法即爲在中文語料中偵測英文單字、反之亦然，額外地我們也將中文全形英數字以及標點符號做一個簡單的轉換，提高原文詞比對率。

對於一個句對中比對到一個原文詞，則該句對亦被設定爲一個極高權值的「錨」。

2.3.4 句對相似度評量

綜合以上原則以及各項參數，我們可以得出

$$\mathrm{eval}\left(sen_i^{\mathrm{CH}}, sen_j^{\mathrm{EN}}\right) =$$

$$\frac{\left(\#\left\{ word_{i,k}^{\mathrm{CH}} = trans\_word_{j,l,m}^{\mathrm{EN}} \;\middle|\; \begin{array}{l} \text{for } i = 1..\#sen^{\mathrm{CH}} \\ \text{for } j = 1..\#sen^{\mathrm{EN}} \\ \text{for } k = 1..\#word_i^{\mathrm{CH}} \\ \text{for } l = 1..\#word_j^{\mathrm{EN}} \\ \text{for } m = 1..\#trans\_word_{j,l}^{\mathrm{EN}} \end{array} \right\} + \right.}{\#word_i^{\mathrm{CH}} + \#word_j^{\mathrm{EN}}}$$

$$\left. \#\left\{ word_{i,k}^{\mathrm{EN}} = trans\_word_{j,l,m}^{\mathrm{CH}} \;\middle|\; \begin{array}{l} \text{for } i = 1..\#sen^{\mathrm{EN}} \\ \text{for } j = 1..\#sen^{\mathrm{CH}} \\ \text{for } k = 1..\#word_i^{\mathrm{EN}} \\ \text{for } l = 1..\#word_j^{\mathrm{CH}} \\ \text{for } m = 1..\#trans\_word_{j,l}^{\mathrm{CH}} \end{array} \right\} \right)$$

$$+ \; punc\_sim_{i,j}$$

$$+ \; \infty \times if\_share\_數字詞_{i,j}$$

$$+ \; \infty \times if\_share\_時間詞_{i,j}$$

$$+ \; \infty \times if\_share\_原文詞_{i,j}$$

我們假設所有編號從1開始，以 sen_i^{CH} 來代表編號*i*中文句，以 $\#sen^{\mathrm{CH}}$ 來代表中文句數，以 $word_{i,j}^{\mathrm{CH}}$ 代表句子 sen_i^{CH} 的分詞後的編號第*j*個詞，以 $\#word_i^{\mathrm{CH}}$ 代表句子 sen_i^{CH} 的分詞後的詞個數，以 $trans\_word_{i,j,m}^{\mathrm{CH}}$ 代表詞 $word_{i,j}^{\mathrm{CH}}$ 的編號第*m*個翻譯詞（故爲英文），以 $\#trans\_word_{i,j}^{\mathrm{CH}}$ 代表詞 $word_{i,j}^{\mathrm{CH}}$ 的翻譯詞數目。並以*i..j*代表一個從*i*至*j*的迴圈（各一次）。英文的情況則將以上代號之CH改成EN。 $punc\_sim_{i,j}, if\_share\_數字詞, if\_share\_數字詞, if\_share\_原文詞$ 爲前述之句對應評量調整參數，後三者布林變數，該詞存在則爲1，不存在則爲0。

對於檢驗某 $word_{i,j}^{\mathrm{CH}}$ 是否等於 $trans\_word_{i,j,m}^{\mathrm{EN}}$，兩者均爲中文詞，除了完全相同的比對方式以外，另可使用部分比對（Partial Match）的方式：如果兩詞中有兩個以上（包含兩個）部分字元相等，則視爲兩詞相等。

2.4 群組句與群組句之間的相似度評量

這可以說是雙語語料庫中配對句對應的最後一個階段。在這個階段裡，有四個輸入參數：

一、 上個階段中文各句以及英文各句的所有句對的相似值（即爲一表格）。
二、 句子的句長。
三、 句子的句末標點符號。

輸出則爲本系統的最終所求：中英文文章的「句群對應」，「句群」爲一個或一個以上的句子的群組，換言之句群對包含傳統「一對一」句對應、「一對多」句對應、「多對一」句對應、以及「多對多」句對應。

一般來說，一個標點符號結構相似，「句譯」比「意譯」來的多的平行語料庫，群組句對應的正確結果應多爲「一對一」。相對的，在標點符號結構差異大，或者平行語料庫屬於粗略、大概的翻譯，或者翻譯多採意譯的方式的平行語料庫，則正確的群組句對應爲「一對多」、「多對多」較多。

在一個群組句裡有多句的情況，邊界的判斷就成了正確率的關鍵，也就是該群組不可以無限制的膨脹，包含了過多的句子。更正確的說法應該是，我們必須加上一個群組膨脹的「懲罰值」（Penalty），來避免這種情況。因爲，如果在沒有懲罰的情況下，一個群組包含了越多的句子，則該群組與其他群組的相似值則會無限制的上升，造成最後所有的句子會變成一個群組的窘境。各項實作方法敘述下。

以下爲三個群組句對應相似度評量的參數說明。並在之後接著說明群組句對應的建立模組，也就是本系統的核心演算法。

2.4.1 句長

一個群組句的句長定義爲：

$length_{gi}$ = 群組句*i*裡的所有句子的句長總和。

群組句對應的句長影響值定義爲：

$$eval\_length_{gi,gj} = -\left|length_{gi} - length_{gj}\right| \times W_{length}$$

W_{length}爲總句長相異度參與群組句之間相似度衡量的比重。$eval_length_{gi,gj}$爲兩個群組句*i*以及*j*之間的句長的相異度在群組句對應評量中的影響值（懲罰值）。

2.4.2 句末標點符號

一個群組句的句末標點符號定義爲：

$punc_{gi}$ = 群組句*i*裡最後一個句子的句末標點符號

群組句對應的句末標點符號影響值定義爲：

$$eval\_punc_{gi,gj} = \begin{cases} -W_{punc}, & \text{if } punc_{gi} \neq punc_{gj} \\ 0, & \text{if } punc_{gi} = punc_{gj} \end{cases}$$

W_{punc}爲句末標點符號的相異在群組句之間相似度衡量上的比重。$eval_punc_{gi,gj}$爲兩個群組句*i*以及*j*之間的句末標點符號相異在群組句對應評量中的影響值（懲罰值）。

2.5 群組句對應建立模組

運用以上群群組句對應相似度評量的原則以及參數，我們以動態規劃來完成最佳化的群組對應系統。在這裡我們有兩個模組：動態規劃演算法、以及以多重不同的句子規模，重複運用動態規劃演算法，來達到漸漸增加群組對應解析度的遞迴階段動態規劃。在介紹動態規劃之前，我們先討論較爲直觀的貪心演算法。

2.5.1 貪心演算法

貪心法（Greedy Algorithm）基本原則爲，參考已經算出來的中英各句子之間的評量係數值，來完成句子對應的工作。

首先選取「一對一」作爲iteration的base result，指定中文句集合以及英文句集合各一個可移動的指標，每一回合的iteration嘗試延伸任意一個指標至下一個中文（或英文）的句子，並把這一個句子加入評量對應句組，重新評量對應句組的對應評量係數，並減去一個懲罰係數（敘述於下），以作爲避免產生指標無限延伸的結果。當兩指標都無法經由延伸而提高評量係數的時候，一組中英對應句組於是產生。

如此重複此貪心演算法，直至任一指標使用完所有的句子爲止。

懲罰係數爲額外囊括一個額外句子進入目前句對應組所必須付出的代價，我們將之定義如下：

$$penalty(S) = f_S \times avg_{matced\_words\_rate}$$

f_S 爲句子的語意單元個數，$avg_{matced\_words\_rate}$ 爲平均一個正確的中英句對應中，相同語意單元（lexical unit）的數量除以句子語意單元個數的平均。這個平均值將影響句對應組是糾結成群的模糊對應，還是零碎的精準對應的關鍵值。此平均值可爲手動統一指定（按照字典的完整程度來調整），或由程式統計判斷（在此一個正確的中英句對應的定義爲，句對應的評量係數明顯高於相鄰句對應組的情況，則該對應句的*matched_words_rate*則可作爲調整 $avg_{matced\_words\_rate}$ 的因子）。

此演算法的優點爲速度快、效率高，缺點是當如果有任何一次選擇對應句組的時候遭遇錯誤分配，則該句之後的句對應演算情形將不樂觀。故本演算法適合處理正確句對應句數應盡量接近「一對一」爲原則的文章。

2.5.2 動態規劃演算法

爲了避免Greedy Algorithm在句對應演算中，因每次產生的對應句組僅爲該次句對應評量的最佳組合，而非整體文章的對應組最佳組合，故我們使用動態規劃演算法來解決總體最佳化的問題。

我們假設句編號從1開始，以 sen_i^{CH} 來代表編號*i*中文句，以 $\#sen^{\mathrm{CH}}$ 來代表中文句數。對於群組句，以 $sen_{i,j}^{\mathrm{CH}}$ 代表一個擁有編號*i*至*j*所有中文句的中文群組句，並以*i..j*代表一個從*i*至*j*的迴圈（各一次）。英文的情況則將以上代號之CH改成EN。

首先建立 $DP\left(1..\#sen^{\mathrm{CH}}, 1..\#sen^{\mathrm{EN}}\right)$ 的兩維動態規劃表，對於累積至目前爲止的最佳句對應組考量評量係數的計算方式，採取

$$DP(i,\ j) = \max\left\{\begin{array}{l} DP(p,q) \\ \quad + eval(sen^{\mathrm{CH}}_{p+1..i}, sen^{\mathrm{EN}}_{q+1..j}) \\ \quad + eval\_length_{sen^{\mathrm{CH}}_{p+1,i-1}, sen^{\mathrm{EN}}_{q+1,j-1}} \\ \quad + eval\_punc_{sen^{\mathrm{CH}}_{p+1,i-1}, sen^{\mathrm{EN}}_{q+1,j-1}} \\ \quad - \sum penalty(sen^{\mathrm{CH}}_{p+1..i-1}) \\ \quad - \sum penalty(sen^{\mathrm{EN}}_{q+1..j-1}) \\ \quad , \text{for } 0 \le p < i, 0 \le q < j \end{array}\right\}$$

*eval*爲上一階段之各單句對應的評量函數之表格查詢。e*val_length,* e*val_punc*爲前述之群組句對應衡量時之句長、以及句末標點符號的影響值（懲罰值）。*penalty*懲罰函數與貪心法定義之懲罰函數相同。

最後，$DP\left(\#sentence^{\mathrm{CH}}, \#sentence^{\mathrm{EN}}\right)$則爲最佳整體對應組之評量係數。由此評量係數回溯組合成對應的最佳整體句對應分配。

此演算法的優點爲可解決中途的句對應的錯誤所造成的全體錯誤，某些不合理的分配將會在其它回合的評量當中獲得矯正。本演算法擁有較爲可觀的計算雜度以及額外的空間。但可以不受斷句的相關性的影響，再配合更精良的判斷評量權值以及屬性，更可大幅提高句組對應的正確率。

2.5.3 遞迴階段動態規劃

這個步驟是句對應演算法改良的一個重點。我們以上述的DP加上可參數化的機制以後，對文章不只作一次的DP句對應，而是以不同的參數設定，執行兩次或兩次以上。如此一來可以根據不同的文章和考量規模作出優化的調整。

對評量的中英文文章作第一次的DP處理時，所用的參數和分數的懲罰等可以放寬，目的是增加區塊對應的正確率，但同時也會使得區塊容易因爲區塊邊界的模糊而呈現脹大的現象。把這些初步的小區塊再分別送至第二階段的DP處理，第二階段的DP處理參數則會採取比較嚴格、採重罰的形式，鼓勵句子分句單獨化，在第一階段正確區塊對應的前提下，這樣子冒險的假設可以在比較安全的環境下，得出正確率高又不令人失望的對應。

這樣的Iterative Process要執行幾次可以透過自動判斷的方式進行。在某些篇幅十分壯觀，經過兩次的DP處理後，仍然存在著大區塊對應的情況下，可以將參數定義的更加的嚴格，送往第三次、甚至是第四次的分割處理。但也有可能因爲該對應區塊因爲確實應該對應在一塊，這樣的情況下就不應該無限制的增加參數的嚴格性，導致因爲處理太過於苛求細緻而產生錯誤的對應。

3 效能評量與雙語庫

我們的效能評量採用Pierre Isabelle (1996) 的Sentence Alignment的Evaluation Metric。對於兩組句對應的找出「雙語對應空間」上的面積交集（圖. 2）：

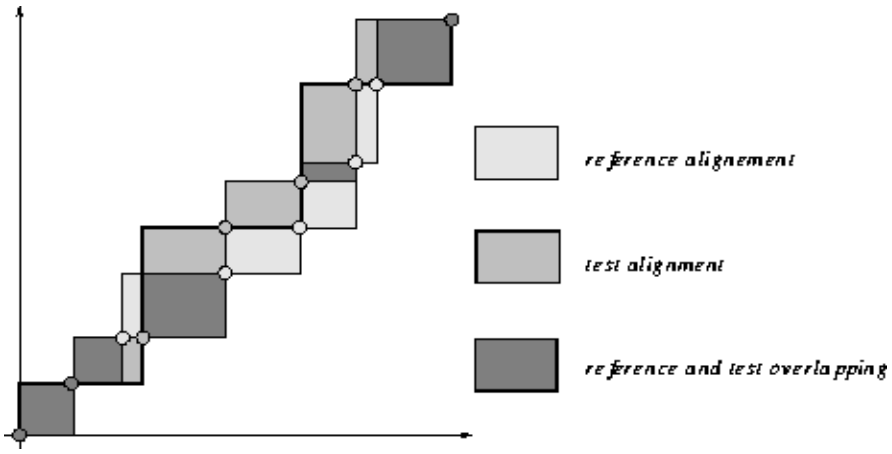


圖. 2. 句對應recall以及precision之圖示

將兩軸視爲兩語言的字元串，則句對應的表示方式則爲其個別語言之句子，在兩軸上所佔據的範圍，延伸在對應空間上則會交出一個長方形。則我們可將兩組不同的文章句對應用以上的方式表示，利用其交集來計算精確率（precision，即交集面積除以正確句對應的所有面積）以及召回率（recall，即交集面積除以評估的句對應的所有面積）。

運用上述的不同的演算法組合，對此平行語料庫產生出來的句對應分配，與「正確」的句對應分配進行評估。該正確句對應分配的產生方法，以人工的方式來產生。

因人工產生之正確句對應組合需要人力介入，故我們採用兩組不同的雙語語料庫：其一爲光華雜誌2001年1月單月的雙語語料來作評估。該月雙語語料庫專文共24篇，共有54336中文字，37195英文字。另外一組爲以隨機的方式選擇各式雙語文件做評估，如美國知音等，以避免系統針對光華雜誌做最佳化的調整。

4 群組句對應評估

因爲群組句對應的建立模組與其他幾個評估變因較爲獨立，故可提前評估。如此在以下的評估中，可以直接採用較高正確率的一個模組，以簡化評估之變因數。

在這裡我們測試貪心演算法（Greedy Algorithm），動態規劃（Dynamic Programming）的優劣，以及遞迴階段式動態規劃（Iterative DP）是否有對動態規劃造成助益，呈現更好的結果。除此變因外，採用所有其他的輔助參數。

Table 1. 評估結果

Recall (%) Precision (%)	Greedy	DP	Iterative DP (2 stages)	Iterative DP (3 stages)
Sinorama	73.37	96.47	93.01	86.79
(200101)	70.82	72.40	81.26	83.83
Random	84.18	98.17	94.70	91.12
	73.92	76.81	83.23	84.65
Average	78.78	97.32	93.86	90.45
	72.37	74.60	82.25	83.24

由評估結果，我們可以看出，兩階段的Iterative DP總體效能最好，但與三階段的Iterative DP來比較，則所勝不多。貪心法速度快，但容易造成一對句配對錯物導致之後的句子配對的連環錯誤。DP則因爲解決了這樣的問題，做整體性的規劃，故在效能上有明顯的提升。相較於DP對於Iterative DP的更進一步的語句裁剪，導致Recall的下降，Precision上升，由上表可知，超過了兩階段的DP之後，群組句對應已經到了一個穩定的階段，加深DP的階段數並無法提高整體的效能。

5 實驗結果與分析

總和以上評估，可得出以下最佳的句對應組合：

- 群組句對應模組
  - 2 Stages Iterative DP
- 句對應參數
  - 採用廣泛的字典翻譯詞以及Stop list
  - （中文翻譯詞的部分比對（Partial Match））。
  - 句中重要標點符號序列相似度
  - 共同數字詞、時間詞、原文詞之對應錨
- 群組句對應參數：
  - 句長相異度
  - 句末標點符號要求相同

翻譯詞的比對是否運用中文Partial Match（本系統採用兩個字以上），因爲在效能上沒有太大的改善，故可以斟酌使用，或是運用更複雜的Partial Match。

系統在召回率率方面表現良好，運用單階段動態規劃將可使召回率幾近100%，但因動態規劃本身的多對多對應的性質，故常有對應正確，群組句的成員句數量過大的情況，導致精確率略微偏低。運用多階段的動態規劃將可以在犧牲些許的召回率的情況下，相當程度的提高精確率至80%以上的水準。

6 結論與未來的研究方向

本系統綜合各種現存已知的不同的自然語言訊息，協助單階段、甚至多階段的動態規劃演算法，在可以處理多對多群組句對應的情況下，得出一個極高召回率、並且擁有可接受程度的精確率的一套句對應系統。

因爲中英文屬於不同語系，相較於其他Sentence Alignment較多著眼於同一語系的雙語語料庫來說，因爲在語言結構上有著明顯的差異，故在譯文中常運用意譯的方式，整體效果無法呈現相同的高效能。又中文缺乏明顯的詞的分界，功能詞的數量多而且用法多變，更加深了自動中英文句對應上的難度。

本系統總結出的最佳效能多對多群組句對應演算法組合，仍有一些可以改善的重點：

- 未知詞比對
  加入專有名詞的檢索與翻譯詞的比對，並建立未知詞資料庫，提供不同的未知詞相似權值，可視爲不同的對應錨（Anchor）。
- 翻譯詞比對：
  因中文詞與英文詞在翻譯詞上較難出現相同詞的特徵下，採用完全比對的方式往往會造成相關句之間的評量的權值不足，被誤判爲沒有相關，故可在翻譯詞上的比對上嘗試以下改良：
  - 詞語意類別：
    可以嘗試採用詞語意類別（參考Ker and Chang (1997)），如car, train等字詞屬於transportation的category，可以額外在字詞比對上提供其他的相似訊息。
  - 更精準的部分比對：
    採用統計的方式來支援字詞的部分比對，提供完全比對之外的相似訊息。
- 加入統計式的詞對應相似度模型（如Melamed (1997b)）
  針對語料庫做統計式的詞分析，可以用來支援字典涵蓋詞的不足。

致謝

本文得到國科會計劃NSC91-2411-H-002-080 「詞彙語意關係之自動標注—以中英平行語料為基礎」資助及清華大學劉顯親教授與張俊盛教授共同主持之國科會計劃NSC92-2524-S007-002 「前瞻性數位語言學習中心CANDLE之研發：應用（雙語）語料庫及電腦化學習之支援」分項子計畫之資助特此致謝。

參考文獻

Brown, P. et al. (1991) “Aligning Sentences in Parallel Corpora.” In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, pp.169-176, Berkeley, Ca.

Catizone, R. et al. (1989) “Deriving Translation Data from Bilingual Texts.” In Proceedings of the First International Lexicon Acquisition Workshop, Detroit, Michigan.

Chang, B., Danielsson, P. and Teubert, W. (2002) "Extraction of Translation Unit from Chinese-English Parallel Corpora," *COLING-02: The First SIGHAN Workshop on Chinese.Language Processing*.

Gale, W. and Church, K. (1991) “A Program for Aligning Sentences in Bilingual Corpora.” In Proceedings of the Annual Conference of the Association for Computational Linguistics, pp. 177-184.

Gao, Z.-M. (1997) Automatic Extraction of Translation Equivalents from Parallel Corpora. Ph.D. Thesis. University of Manchester Institute of Science and Technology.

Haruno, M. and Yamazaki, T. (1996) “High-Precision Bilingual Text Alignment Using Statistical and Dictionary Information.” Proceedings of the Annual Conference of the Association for Computational Linguistics, pp.131-138.

Isabelle, P. and Simard, M. (1996). "Propositions pour la représentation et l'évaluation des alignements de textes Parallèles. ”

Kay, M. and Roscheisen, M. (1993) “Text-Translation Alignment.” Computational Linguistics, Vol. 19, No 1, pp. 121-142.

Ker, S. J. and Chang, J. S. (1997) “A Class-based Approach to Word Alignment.” C*omputational Linguistics*, Vol. 23, No. 2, pp. 313-343.

Le, S., Youbing, J., Lin, D., and Sun, Yufang 2000 “Word Alignment Of English-Chinese Bilingual Corpus Based on Chunks”, In *Proc. 2000 EMNLP and VLC*, pp. 111-116.

Melamed, D. (1997a) “A Portable Algorithm for Mapping Bitext Correspondence.” In Proceedings of the 35th Annual Conference of the Association for Computational Linguistics, Madrid.

Melamed, D. (1997b) “A Word-to-Word Model of Translational Equivalence.” In Proceedings of the 35th Annual Conference of the Association for Computational Linguistics, Madrid.

McEnery, O. and Oakes, M. (1996) “Sentence and Word Alignment in the CRATER Project.” In Thomas and Short (eds.) Using Corpora for Language Research, pp. 211-231. New York: Longman.

Michel, S.and Plamondon, P.1(996) “Bilingual Sentence Alignment: Balancing Robustness And Accuracy,” In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA-96)*, pp. 135-144, Montreal, Quebec, Canada.

Tiedemann, J. (1998) “Extraction of Translation Equivalents From Parallel Corpora ” In *Proceedings of the 1th Nordic Conference on Computational Linguistics*, Center for Sprogteknologi, Copenhagen.

Utsuro, T. et al. (1994) “Bilingual Text Matching Using Bilingual Dictionary and Statistics.” In Proceedings of International Conference on Computational Linguistics, pp. 1076-1082, Kyoto.

Wu, D. (1994) "Aligning A Parallel English- Chinese Corpus Statistically With Lexical Criteria," In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM.

Xu, D. and Tan, C. L. (1996) “Automatic Alignment of English-Chinese Bilingual Texts of CNS News.” In Computational Linguistic Archive cmp-lg/9608017.

Yeh, Chih-Cheng. (2002) Using Punctuation Marks for Bilingual Sentence Alignment. MA Thesis. National Tsing Hua University.

中央研究院中英雙語知識本體詞網 (Sinica BOW)：結合詞網，知識本體，與領域標記的詞彙知識庫

張如瑩　黃居仁

中央研究院語言學研究所

ruyng@gate.sinica.edu.tw　churen@gate.sinica.edu.tw

Abstract. 論文中介紹了由中研院語言所和資訊所從語言工程之角度建置的中央研究院中英雙語知識本體詞網（The Academia Sinica Bilingual Ontological Wordnet, Sinica BOW）[1]，希冀能滿足電腦更多需求以處理更多變的問題。它以WordNet[2]爲基礎，加上台灣地區所使用的中文，搭配領域以及SUMO（Suggested Upper Merged Ontology，建議上層共用知識本體），並以WordNet1.6版offset爲延伸的識別碼作爲媒介連結了領域詞彙庫和領域知識本體。該系統爲了便於使用者查詢，藉由多元、友善的介面，將功能切割爲詞網、知識本體以及索引三個主要單元，提供跨語言資訊轉換、詞義的區分與詞義關係的連結、語言資訊與概念架構（知識本體）的連結以及使用領域等訊息。文中並提到，隨著WordNet版本的變更，爲保留語言資訊演變，便以版本的比對，我們以WordNet所提供的1.6版和1.7.1版之名、動詞的單詞義和多詞義對應資料爲基礎，輔以Sinica BOW現有資訊，將資料分爲單詞義、多詞義對應資料以及1.7.1版新增詞形資料，透過同義詞集、詞形以及詞類訊息，經進一步自動和人工處理後，將Sinica BOW和WordNet1.7.1版結合，並提供版本對應與比對，建立之後Sinica BOW與WordNet各版本結合的處理模式。.

1　簡介

知識經濟時代的來臨，全球資訊網的誕生，擴大人們接收資訊的觸角，於是乎能否掌握知識往往成爲致勝關鍵，以關鍵詞檢索已無法滿足使用者的需求，面對此現象，全球資訊網的提議者 Tim Banners-Lee 提出語意網（Semantic Web）願景，希冀電腦能真正理解人們的需求，爲達成此願景他一並提出以XML（Extensible Markup Language）、RDF（Resource Description Framework）加上URI（Uniform Resource Identifier）以及知識本體（Ontology）解決知識呈現[3]。除此之外，更有多位語言學專家提出其所面臨的挑戰應必須包括處理表達知識的基礎－語言。而語言工程的處理訊息應囊括語意，甚至概念上的處理，這是全球皆有的共識。然而其中，中文是台灣本土語言，相關訊息的研究我們必然責無旁貸，而英文爲世界共通的語言，是達成國際接軌的必經途徑，中英文語言訊息的處理和銜接是必要的任務。

語言學界、計算語言學界爲達語意網的願景作了不少努力，中央研究院中英雙語知識本體詞網（The Academia Sinica Bilingual Ontological Wordnet, Sinica BOW）便則以WordNet爲基礎，企圖建立中英雙語的基礎知識架構。然而隨著語言變遷，WordNet不斷進行版本修正，我們希望藉由Sinica BOW與WordNet各版本的結合，建置並保留完整中英雙語的語言訊息。這篇論文簡介Sinica BOW以及它與WordNet1.7.1版結合的模式。在第二節中我們將針對Sinica BOW製作動機、系統主要使用資源與架構以及現有系統開放的功能進行簡介。第三節描述如何運用WordNet提供的訊息及系統現有的資訊試著將Sinica BOW與WordNet1.7.1版結合。因與WordNet1.7.1版本的結合，Sinica BOW而新增的功能則在第四節中陳述。最後，則是結論以及未來工作。

2 中央研究院中英雙語知識本體詞網（Sinica BOW）

Sinica BOW是由中央研究院語言所（文獻語料庫）和資訊所（詞庫小組）合作建置，企圖從語言工程的角度，以台灣地區的語言使用爲經驗基礎，整合語言和語言，語言和概念以及語言和領域的資訊，甚至是跨語言間的訊息。近期目標爲將以建立完整精確的中英對譯資料庫及檢索介面，作爲數位典藏知識國際化的基礎；並逐步建立各領域之雙語領域辭典，以作爲該領域／典藏雙語控制詞彙的參考標準，及具領域判斷能力資訊檢索之依據；建立帶領域標記之雙語辭典及檢索介面，以加值成爲知識加值雙語電子辭典。未來則希望建立精確的領域知識架構，以作爲高加值知識產業的基礎；建立完整的知識本體架構，做爲下一代網路（如「語意網」）之語意骨幹以及建立以知識爲經緯的中英雙語訊息交換平台，作爲多語知識處理的憑藉爲目標。相信藉由它提供的訊息將有助於含領域專門知識加上語意資訊之領域詞彙庫的建立、人工或自動翻譯、語言學習、自然語言處理、異質性系統間資料之交換與處理以及自動推論。[4]

2.1 主要資源和架構

Sinica BOW主要使用的資源包含WordNet、ECTEC（English- Chinese Translation Equivalents Database）以及SUMO（Suggested Upper Merged Ontology，建議上層共用知識本體）。

「WordNet是結合辭典和知識本體的文字百科全書」在自然語言處理、資訊檢索等的相關研究中常見其身影。[5]1985年普林斯頓大學認知科學實驗室以現代心理語言學理論所述的人類詞彙記憶爲啓發，作出語意式電子字典－WordNet（http://www.cogsci.princeton.edu/~wn/），他以每個同義詞集表達一種詞彙概念，將其區分爲四種英文詞類：名詞、動詞、形容詞、副詞，並以二十幾種詞義關係組織同義詞集。[2]因WordNet的出現，使電腦擁有更豐富的信息可處理各式問題。

由中研院資訊所詞庫小組所建構的ECTEC是以WordNet爲基礎，經由現有英中或中英電子辭典的詞形對應，替每個同義詞集的詞義找出可能相對應的中譯詞組，再經由人工檢驗。尋找對譯的過程中，盡可能的以詞彙而非描述性短語表達，目的在於讓每個同義詞集都有最適當的一至三個左右的中文對譯。其中有5%特殊領域詞彙暨無法在現有電子資源中找到，翻譯者也無法填入適當的中譯，藉於此，我們又花了兩年的時間參考特殊領域辭典將其完成。[6]

SUMO則是由IEEE標準上層知識本體工作小組所建置，其目的在於促使自然語言處理、資訊檢索、自動推論以及資料互通性等工作的進行。知識本體類似於字典或詞彙表，但訊息更豐富，以便於電腦處理其內容。知識本體以格式化的方式表達概念（Concept）、關係（relation）以及公理（axioms）。上層知識本體是將一般性、後設性（meta）、摘要性以及哲學類的概念指出，所以特殊領域的概念可由其中的概念所涵蓋，但特殊領域概念的知識本體則期許由各領域自行制訂。[7][8]（Niles and Pease, 2001）依據SUMO2002年版資料，我們將系統介面以及概念節點進行中文化，其涵蓋11大類的概念，每大類又分爲二至五個類別，總共囊括3,912個概念。日前SUMO已經與WordNet1.6以及2.0版本結合，且以同義（synonymy）、上位（hypernym）、體例（instantiation）這三種類別顯示同義詞集和SUMO概念間的對應關係，例如：同義詞集cell（細胞）與細胞概念（cell）是同義。hockey（曲棍球）屬於運動概念（sport），兩者間的關係爲上位，也就是說運動涵蓋hockey（曲棍球）。China（中國大陸）屬於國家（nation）這概念的體例。[9]

除此，我們以「中國圖書分類法」爲基準，並參考各知識分類與實際研究經驗，提出：包含九大類的知識分類（Knowledge Content），涵蓋427個領域。另外，並因應語言資源特性加入下列語言使用（Language Usage）的各類訊息：專名（說明文字符號的指涉）（Proper Name）、語體（說明文字符號的使用）（Genre/Strata）、各種語言／詞源（Language/Etymology）、各國地名（Country Name）。領域階層的建立在於替不同詞義中的詞彙項目區別其使用的領域，例如：stock作「股票」和「家畜」兩個不同解釋時，分屬於財政學裡的資本以及動物學的脊椎動物學。加註領域信息可降低詞彙歧異性，增加資料交換時的互通性，輔助領域詞彙庫之建構等。

Sinica BOW透過WordNet1.6 offset延伸所產生的識別碼作爲媒介，進行串連，將每個資源以及各類訊息連結。因WordNet1.6 offset延伸的識別碼可獲得原本WordNet存在的詞類、解釋、英文例句、同義詞集、各同義詞集間的詞義關係及其所屬詞彙。而SUMO概念與WordNet的連結，使得可透過該識別碼獲取詞義與概念搭配的訊息。以WordNet爲基礎所建置的ECTED與針對WordNet同義詞集的各詞彙項目所給予的領域值，也是透過該識別碼獲取。如果是特殊領域詞彙庫，加上相對應的Sinica BOW識別碼，也可保留原始

資源的資料庫格式和WordNet連結。又，領域知識本體則是在SUMO某些概念下進行延伸發展。每個特殊領域詞彙庫中的詞彙一樣具有所屬的概念，其所屬概念可能是SUMO或特殊領域知識本體的某一概念，特殊領域詞彙庫和領域知識本體的結合，使得透過該識別碼又串起所有的訊息。Sinica BOW的資源和架構如圖 1所示。由於透過WordNet可以和同是以WordNet爲基礎架構所建置的其他語系WordNet資源加以連結，例如：EuroWordNet[10]，以此作基礎架構可編製成多語的詞彙網路，成爲多語環境中所需之語言知識結構的基礎資料。

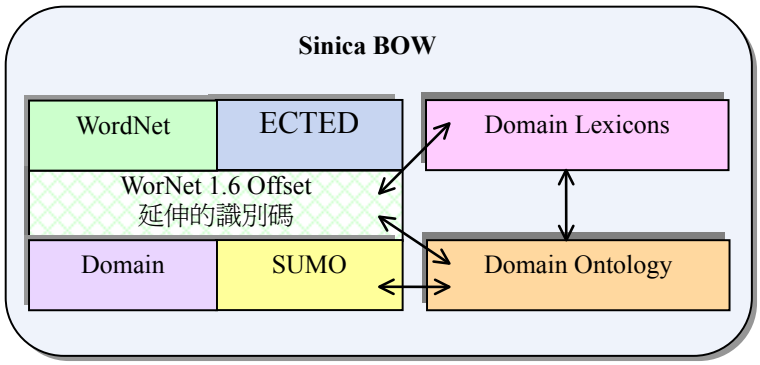


圖 1. Sinica BOW主要資源及架構[11]

2.2 現有功能

Sinica BOW的主要資源和架構中，使用者可獲得跨語言資訊轉換、詞義區分與詞義關係連結、語言資訊與概念架構（知識本體）連結以及使用領域的訊息。爲了讓使用者透過友善介面便很容易地可擷取所需訊息，系統將介面切割爲詞網、知識本體以及索引。

1. 詞網

又分成「中文查詢」、「英文查詢」以及「專門領域」三個單元。「中文查詢」和「英文查詢」中使用者可輸入中或英文詞形，系統將顯示詞形搭配詞類在各資源中的情況，包括出現與否以及分部頻率，所比對的資源是以具有代表性且容易獲取電子檔者爲考量。點選其他各資源的超連結都將連結該資源原始或相關網頁進行進階查詢。但點選「WordNet1.6英中對譯」將出現該詞形所有的詞義，每個詞義以表格呈現該詞義下的所有訊息，包括所屬的領域、詞類、解釋、翻譯、同義詞集、各詞義關係詞、SUMO概念以及英文例句。使用者可再點選任一詞彙項目進行再查詢。

目前的領域訊息主要有三個來源1.藉由領域階層的中英文值找出詞形相符之中譯或同義詞集的詞彙項目，再透過WordNet本身同義和上位詞義關係，假設獲得之同義詞集和上位詞屬於該領域。[12]2.透過遠見科技提供的電子辭典，針對其中擁有領域的詞形推論至WordNet[13]，以及3.線上「使用者建議領域」功能所得之訊息。綜合上述三項結果，目前針對同義詞集之中譯詞彙總共有18,562筆資料有領域值，針對同義詞集之詞彙項目總共有24,229筆資料有領域值，這相當於有14,396筆同義詞集具領域訊息。

「專門領域」可查詢其他領域詞彙庫與WordNet連結的訊息。除了繼承領域詞彙庫中原始的領域訊息，其他則保留WordNet和中譯等訊息。目前可在線上查到中研院資訊所詞庫小組財經辭典和WordNet連結的資訊。

2. 知識本體

由「SUMO」和「領域知識本體」這兩單元組成。都可以以中、英詞形或概念查詢，並以樹狀結構呈現各概念間的關係。提供的訊息包括以詞形查詢所屬概念、概念的定義、公理、詞彙和概念連接之關係以及相連結的WordNet之所有訊息，其中概念訊息皆已中文化。

此外，我們曾嘗試建置「唐詩三百首知識本體」和「魚類知識本體」兩個主題的領域知識本體。我們試著從唐詩三百首所使用的「植物」、「動物」、「人造物」主題詞彙，利用WordNet詞義和詞義關係等訊息的驗證並加以連結，在SUMO的基礎架構下進而建置唐詩三百首領域知識本體。系統介面可顯示加入WordNet語意訊息後知識本體以及詞彙分佈的差異，唐代的知識架構因此可被驗證，例如：有袋類動物並不出現在唐代，有關鳥類或有翅膀的昆蟲詞彙則廣泛分布於詩中，這正符合唐代時興的飛天觀念。[14]

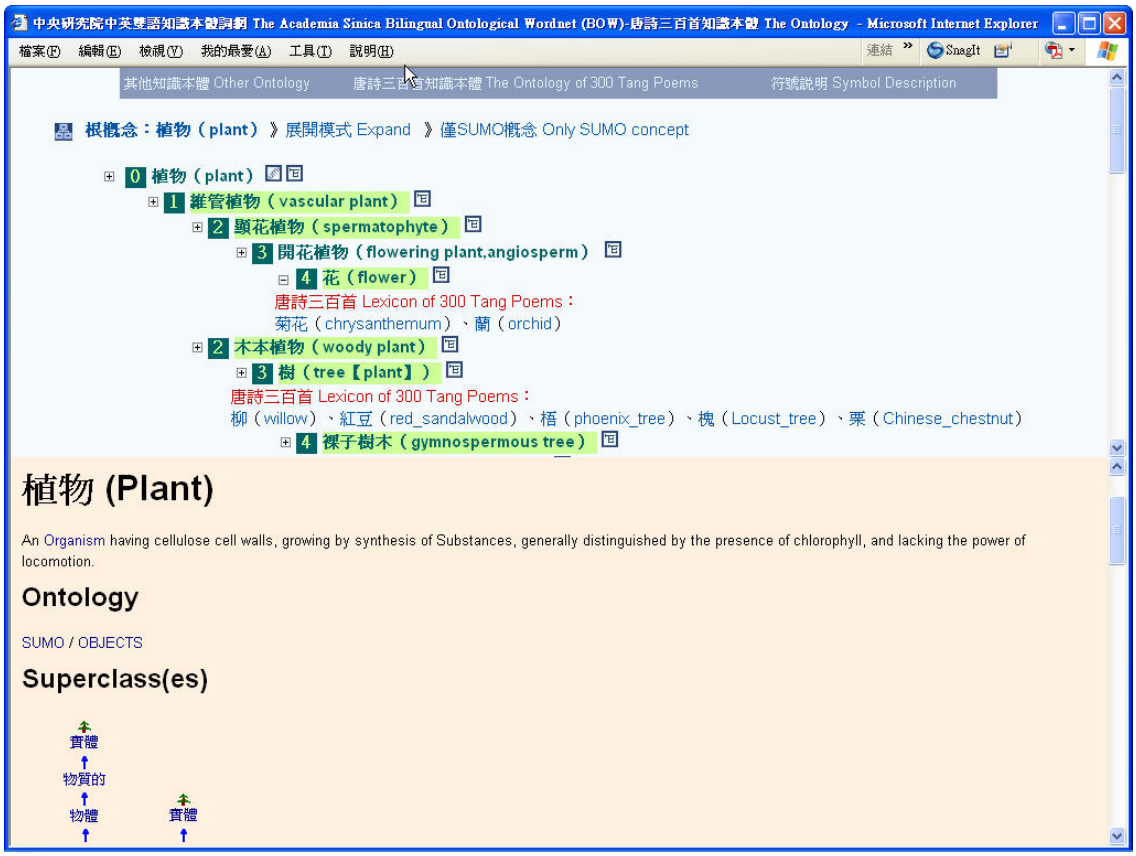


圖 2. 經WordNet語意驗證的植物類唐詩三百首知識本體

另外，跟「臺灣動物相之典藏－臺灣魚類資料庫」計畫（網址：http://fishdb.sinica.edu.tw/）的合作，蒐集領域相關後設資料（metadata）和分類系統，並以WordNet輔助分析上下位關係，最後，以該領域的所有詞彙觀察知識本體是否完整，而完成了魚類知識本體的初步結果。

3. 索引

索引窗口的建立在於讓使用者以字母、字首字尾、詞類、來源、頻率、領域、概念以及綜合（來源、詞類、頻率）的限制條件從事進一步資訊查詢。系統將列出符合條件的詞彙列表，使用者可再進一步查詢對應SUMO、WordNet以及在各資源分佈情況。使用者可從符合條件的詞表中，進一步分析所具有之共同特性和差異。

接下來，我們將針對Sinica BOW與WordNet1.7.1版結合之步驟與提供的新功能進行簡介。

3 Sinica BOW 結合WordNet1.7.1資料

語言會隨著時空的變化而有所變更，WordNet也隨其變更不斷更新、修正其資料。直至2004年止最新推出授權的爲2.0版本。Sinica BOW針對1.7.1版本進行先導測試性實驗，試圖作爲之後Sinica BOW與各版WordNet資料結合時處理模式的根據。我們企圖保留、追蹤並呈現其中的演變，以豐富語言訊息，且這對於作進一步的語言變遷研究有十分的助益。

3.1 WordNet1.7.1以及和WordNet1.6對應原始資料

WordNet1.7.1版本於2001年公布，WordNet 1.7.1詞義順序和1.6版本並不相同，原本詞義第一順位的並不見得仍保持爲第一順位。有很多的新詞形和新詞義，因此詞義重新整理過，1.6版中被察覺到的重複詞義已經被刪除。1.7.1版本無法和1.6版的同義詞集間之差異進行直接對應，針對此問題WordNet提供兩版本間名詞和動詞的對應，其中名詞和動詞又細分爲單詞義的（Monosemous）和多詞義（Polysemous）的對應。對應準則和方法爲：

1. 1.6或1.7.1版資料庫中唯一的名詞或動詞則忽略。
2. 兩版本資料庫中，若爲名詞和動詞的單詞義之sense_key（同義詞集中的某一詞彙項目）和synset_offset則直接對應。論文中將分別以1.6to1.7.1.noun.mono 以及1.6to1.7.1.verb. mono表示這類對應的結果。
3. 1.6版和1.7.1版兩個資料庫中若爲名詞和動詞的多詞義，WordNet則以不同的啓發方式去評估1.6和1.7.1版本之間詞義的相似性，並針對每個比對結果給予分數。針對每個詞彙（word），每個1.6版詞義將和1.7.1版詞義的相同詞彙進行比對，1.7.1版詞義(s)有比較高分數的將被視爲最佳的對應結果。文中將這些詞義對應的結果以1.6to1.7.1.noun. poly 以及1.6to1.7.1.verb. poly示之。WordNet在進行對應時採用的啓發方式包括：sense keys比較、同義詞集詞彙項目（term）相似度以及相關樹位置（relative tree location）（也就是上位詞比較）。比對時，註解（Glosses）的資訊不被參考，因爲她們常更新。對應分數從0到100，表示經由啓發對應之可信度，分數越高可信度越高。90或100是最主要的詞義對應。其中，96%的名詞對應分數高過90，94%的動詞對應分數高過90。分數和資料分布如下：

表 1. WordNet提供1.6和1.7.1版多詞義名詞與動詞對應的分數和資料分布情況

1.6to1.7.1.noun.poly				1.6to1.7.1.verb.poly			
分數	筆數	1.6 同義詞集數	1.7.1 同義詞集數	分數	筆數	1.6 同義詞集數	1.7.1 同義詞集數
100	30,072	22,757	22,757	100	12,690	7,905	7,905
90	4,605	3,130	3,143	90	3,167	1,793	1,796
80	412	291	297	80	321	187	189
70	230	190	187	70	149	121	120
60	145	100	103	60	79	48	50
50	19	16	16	50	41	39	39
40	43	38	38	40	47	41	42
30	48	45	44	30	25	24	23
20	135	120	122	20	120	96	102
0	32	29	0	0	46	41	0
總和	35,741	26,560	26,484	總和	16,685	10,037	9,969

表 2. WordNet提供1.6和1.7.1版單詞義名詞與動詞對應的資料分布情況

項目 資料類型	詞數	1.6.同義詞集數	1.7.1同義詞集數
1.6to1.7.1.noun.mono	80,163	49,870	49,946
1.6to1.7.1.verb.mono	5,278	3,933	3,936

WordNet1.6版有116,317名詞詞義，其中115,904可對應到1.7.1版WordNet。殘餘WordNet1.6表達唯一詞義的413名詞詞義。在1.6to1.7.1.noun.poly中被對應的多詞義名詞全部有35,741詞義，在1.6to1.7.1.noun.mono中對應的單詞義名詞有80,163個。

WordNet1.6版有24,169動詞詞義，其中21,963可對應到1.7.1版WordNet。殘餘WordNet1.6表達唯一詞義的2,206動詞詞義。在1.6to1.7.1.verb.poly中被對應的多詞義動詞全部有16,685詞義，在1.6to1.7.1.verb.mono檔案中對應的單詞義動詞有5,278個。（WordNet, 2001）

表 3. WordNet提供1.6和1.7.1版名詞與動詞對應的資料分布情況

項目 資料類型	對應1.7.1總詞義／1.6原詞義	1.6.同義詞集數	1.7.1同義詞集數
1.6to1.7.1.noun	116,317 ／115,904	65,946	66,039
1.6to1.7.1.verb	24,169 ／21,963	12,113	12,065
1.6to1.7.1.noun+verb	140,486／137,867	77,987	78,041

3.2 進一步處理WordNet提供的WordNet1.7.1和1.6版對應訊息

根據WordNet提供的資料特性，爲了便於與Sinica BOW資料的結合，我們分別針對WordNet提供的「單詞義」以及「多詞義」資料透過暨有的1.6版本中譯資訊的輔助作進一步的處理。又，因1.7.1版新增詞形並不會有比對訊息，當然更無法從WordNet1.6暨有的中譯訊息中獲得任何輔助，因此1.7.1版中新增的英文詞形又另作處理。這三類資料的個別處理方式與結果如下：

1. WordNet 1.6版中對應到WordNet 1.7.1版的「單詞義」資料

不論名詞或動詞，假設直接繼承1.6版本的翻譯，且不再進行人工檢查。

2. WordNet 1.6版中對應到WordNet 1.7.1版的「多詞義」資料

(1) 不論名詞或動詞，由於WordNet對應多詞義資料的方式是以兩版同義詞集中的某一詞彙項目進行比對，並給與對應分數，若對應分數爲100分，則假設1.7.1版本中對應的該同義詞集下的所有中譯詞組，並不需要變更，直接繼承1.6版相對應的翻譯。

(2) 其他對應分數小於100的多詞義資料，則保留原WordNet提供的對應模式，也就是以兩版同義詞集中的某一詞彙項目比對，我們並提供1.6版本的中譯，再針對1.7.1版本中該詞彙項目是否仍適用於自動對應到之1.6版中譯詞組進行人工檢視，若否，則變更爲校正值。

表 4. WordNet提供其對應分數小於100的多義詞人工校對結果

項目 資料類型	總筆數	1.6沒中譯	1.7.1有校正	結合後有中譯	結合後沒中譯
1.6to1.7.1.noun.poly.70-20	1,064	3	109	1,064	0
1.6to1.7.1.noun.poly.80	412	0	9	412	0
1.6to1.7.1.noun.poly.90	4,605	0	53	4,605	0
1.6to1.7.1.verb.poly.70-20	461	7	61	459	2
1.6to1.7.1.verb.poly.80	321	2	8	319	2
1.6to1.7.1.verb.poly.90	3,167	18	48	3,158	9

3. WordNet 1.7.1新英文詞形

我們依據1.6版本將1.7.1版本中新增的英文詞形另外找到1.7.1版本中相對應的同義詞集，其中新增了18,969英文詞形，若去掉大小寫因素則新增18,461英文詞形，共佔13,708個1.7.1版的同義詞集。對於這類相對於1.6版本而新增的英文詞形，我們假設新英文詞形在1.6版本中找不到對應的同義詞集，於是我們直接進行人工檢驗，由人工新增相對應的中譯詞組。人工翻譯的結果，找不到中譯的同義詞集仍有17筆，多爲特殊專有名詞（名詞14筆，形容詞3筆）。

3.3 綜合處理對應訊息的結果

接著我們試著將上述步驟所產生的資料綜合，其中若1.7.1版本的中譯內容沒有校正值，則假設繼承1.6版本中相對應的同義詞集之中譯詞組。所有單詞義或多詞義之名詞或動詞的對應資料類型中，只要針對同一個同義詞集進行對譯，不論針對該同義詞集下哪一個英文詞彙項目的中文對譯，都視爲該同義詞集的中譯詞組的成員。

表 5. 經自動及人工檢驗後的WordNet提供兩版本對應對應資料以及1.7.1版新增詞形的中譯詞組統計

項目 資料類型	有中譯的同義詞集筆數	有中譯的同義詞集百分比	產生的中譯詞形	產生的WN1.6對應WN1.7.1數
1.6to1.7.1.noun.mono	459,94	41.3530%	55,922	80,163
1.6to1.7.1.noun.poly.70-0	500	0.4495%	745	620
1.6to1.7.1.noun.poly.80	298	0.2679%	473	412
1.6to1.7.1.noun.poly.90	3,143	2.8259%	4,636	4,604
1.6to1.7.1.noun.poly.100	22,758	20.4616%	27,131	30,072
1.6to1.7.1.verb.mono	3,904	3.5101%	5,847	5,278
1.6to1.7.1.verb.poly.70-0	368	0.3309%	547	461
1.6to1.7.1.verb.poly.80	190	0.1708%	339	321
1.6to1.7.1.verb.poly.90	1,797	1.6157%	2,688	3,167
1.6to1.7.1.verb.poly.100	7,906	7.1082%	9,036	12,691
1.7.1新增詞形的同義詞集	13,708	12.3248%	17,258	
總和	85,619	76.9796%	96,378	137,789

表 6. WordNet1.6版Sinica BOW資訊和1.7.1版利用WordNet提供對應資料以及新增詞形產生中譯詞組之訊息的比對

版本 項目	WordNet1.6	WordNet 1.7.1
同義詞集數	99,642	111,223
有中譯的同義詞集數	99,642	85,619
沒中譯的同義詞集數	0	25,604
同義詞集的中譯總詞彙數	149,780	128,873
平均每個同義詞集的中譯詞數	1.5032	1.15587
同義詞集的英文詞總數	17,401	195,817
平均每個同義詞集的英文詞數	1.7463	1.7606
平均每個英文對應的中譯數	0.8608	0.6581
中譯總詞形數	109,982	96,378
同義詞集英文總詞形數	122,864	140,488

綜合WordNet提供的兩版對應資料所做的中譯資料校正，以及針對1.7.1版新增詞形所對應的同義詞集進行人工對譯，上述兩大類所產生的中譯資料如上表 5和表 6所示，發現尚有25,604筆同義詞集尚未有任何中譯。除了多詞義資料在人工校對時即無法找到適合對譯的，其他我們假設可能是WordNet1.6和1.7.1 版並未針對形容詞和副詞進行對應。於是我們透過1.6版和1.7.1版本的資料再作第二次的進一步比對處理：

1. 若兩版本的同義詞集完全相等，則中譯詞組直接繼承，但發現並無此類同義詞集。
2. 具有相同詞形，且詞類相同的資料，就給與1.6版本相對應的同義詞集的中譯詞組。這類資料有15,942筆同義詞集，佔尚缺中譯資料的62.2637%，相對於1.6版本的33,791筆同義詞集。
3. 其餘則繼續比對，只要具有相同詞形，一樣給與1.6版本相對應的同義詞集的中譯詞組，找到9,645筆這類同義詞集，佔尚缺中譯資料的37.6699%，相對於1.6版本的48,121筆同義詞集。

經過前面三步驟的檢驗，除了在1.7.1版本中針對新增的英文詞形進行人工檢閱增加中譯詞組時就無法找到適當中譯的17筆同義詞集，其他所有同義詞集都已找到中譯詞組。爲確保中譯資料的正確性，將再針對第二和第三步驟所產生的資料進行人工檢驗，並保留經人工確認後1.7.1版對應的1.6版同義詞集之訊息。而以此產生的訊息，除了兩版本間同義詞集的對應，我們也假設兩版本其中對應到同義詞集詞彙項目也互相對應，但我們並不另外提供類似WordNet的對應分數之訊息。

4 Sinica BOW 結合WordNet1.7.1新增的功能

Sinica BOW因與WordNet1.7.1結合而擁有更豐富的訊息，它除了保留原來的以WordNet1.6版offset延伸之識別碼爲媒介的架構與功能，還新增了查詢各版本資訊、版本間對應以及版本比對的功能。

1. 查詢各版本

使用者可以中或英文的詞形進行查詢，系統將會出現該詞彙及所屬詞類在各版的分佈情況，包括是否出現在該版本中以及綜合其他資源統計出的出現頻率層次（核心詞彙、參考詞彙以及一般詞彙）。點選某一版次後，系統以表格格式出現在該版中該詞形所有的詞義，以及每個詞義下的所有訊息。

2. 版本間的對應

每個詞義中還內含對應的資訊，包括兩版本間同義詞集的對應，以及版本間同義詞集之詞彙項目（或中譯詞）的對應，若原本查詢1.6版本，點選對應的1.7.1版，則畫面上、下視窗將同時出現兩版本該詞義的所有訊息。同樣的，若查詢在1.6版同義詞集詞彙項目對應1.7.1版本中哪一個同義詞集中哪一個同義詞集詞彙項目，點選後便將兩版本的相關訊息出現在同畫面的上、下視窗中，便於使用者直接比較。

3. 版本比對

在WordNet1.6或1.7.1版本中即使相對應的同義詞集但可能同義詞集詞彙項目、中譯詞組或各詞義關係的成員不盡相同，因此當使用者查詢某詞彙的各詞義資訊時，同時在該詞義訊息表格中顯示同義詞集詞彙項目、中譯詞組或各詞義關係的成員是否有所變更。若有所變更的，會在該詞彙項目旁出現不同標記，畫面上可清楚顯示該詞彙項目是否存在於另一版本相對應的詞義中。譬如：1.6版的computer作自動計算的機器的解釋時，中譯爲”“電腦”和 “電子計算機”，但1.7.1版爲“電腦 ”、“計算機 ”、“資料處理器”、“訊息處理系統”、“電腦 ”以及“電子計算機 ”。系統顯示結果如下圖 3：

圖 3. Sinica BOW版本對應比對之功能

5 結論

Sinica BOW以WordNet爲基礎，在其資訊上融合中英跨語言資訊轉換，使用者可以中英文的詞形查詢，獲得詞義區分與詞義關係連結、語言資訊與概念架構（知識本體）連結以及所屬領域的資訊。除此，我們在上層共用知識本體SUMO下延伸領域知識本體，並以WordNet1.6版Offset所延伸的識別碼串連上述訊息及各類領域知識本體和領域詞彙庫。並以多樣化友善介面讓使用者從詞形、領域、詞類、概念、出處等角度切入，交錯查詢各類綜合訊息。

論文中並針對WordNet提供的1.6版和1.7.1版單詞義及多詞義的名、動詞對應，以及利用現存於Sinica BOW中WordNet1.6版的各類訊息，輔助進行Sinica BOW與WordNet1.7.1版的結合，並新增各版本資訊查詢、版本間同義詞集以及同義詞集詞彙項目對應以及版本間同義詞集詞彙項目、中譯詞及各詞義關係成員的版本比對功能。希冀以此嘗試作爲與WordNet2.0版甚至與其他版本擴增的基礎模式。

未來，除了以此模式繼續融合WordNet其他具有代表性之新版本的資訊，也希望與透過相同架構，以不變更各資源原格式爲原則，進一步結合其他資源，甚至是完整的中文詞網，藉以豐富並建構成完整的知識網路，作爲電腦多語訊息處理基礎架構。

References

[1] 中央研究院中英雙語知識本體詞網 The Academia Sinica Bilingual Ontological Wordnet (Sinica BOW)，http://BOW.sinica.edu.tw

[2] WordNet, http://www.cogsci.princeton.edu/~wn/

[3] Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001.<http://www.sciam.com/2001/0501issue/0501berners-lee.html>

[4] 黃居仁、張如瑩、蔡柏生（民93）。「語意網時代的網路華語教學: 兼介中英雙語知識本體與領域檢索介面」（Chinese Language Education and the Developing Semantic Web:An Introduction to Chinese-English Bilingual Ontology Interface）。在黃一農、張寶塔總編輯、羅鳳珠執行編輯，資訊與社會叢書系列之三：語言文學與資訊科技（卓越計畫）（頁443-467）。新竹市：清華大學出版社出版。

[5] 張俊盛，「喬治米勒在聖塔菲 WordNet：結合辭典和本體論的文字百科全書」，科學人雜誌（2004年5月），<http://www.sciam.com.tw/forum/forumshow.asp?FDocNo=444&CL=16>。

[6] Chu-Ren Huang,. Elanna I. J. Tseng, Dylan B. S. Tsai, and Brian Murphy. Cross-lingual Portability of Semantic relations: Bootstrapping Chinese WordNet with English WordNet Relations. Language and Linguistics. 4.3, pp.509-532. 2003.

[7] Suggested Upper Merged Ontology, http://www.ontologyportal.org/

[8] Niles, I., and Pease, A. "Toward a Standard Upper Ontology". In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001). Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19, 2001.

[9] Niles, I., and Pease, A. "Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology". In proceedings of the IEEE International Conference on Information and Knowledge Engineering. (IKE 2003). Las Vegas, Nevada, June 23-26, 2003.

[10] EuroWordNet: Building a multilingual database with wordnets for several European languages., http://www.illc.uva.nl/EuroWordNet/

[11] Chu-Ren Huang, Ru-Yng Chang, and Shiang-Bin Lee. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon. Portugal. 26-28 May, 2004.

[12] Chu-Ren Huang, Xiang-Bing Li, and Jia-Fei Hong. Domain Lexico-Taxonomy: An Approach Towards Multi-domain Language Processing. The 1st International Joint Conference on Language Language Processing (IJCNLP-04) Asian Symposium on Natural Language Processing to Overcome Language Barriers. Sanya City. Hainan Island, China. 25-26 March ,2004.

[13] Echa Chang, Chu-Ren Huang, Sue-Jin Ker, and Chang-Hua Yang. Induction of Classification from Lexicon Expansion: Assigning Domain Tags to WordNet Entries. 19th COLING 2002 Post-Conference Workshop--SEMANET: Building and Using Semantic Networks Processing. Center of Academia

Activities, Academia Sinica. Taipei. Taiwan. September 1, 2002.
[14] Chu-Ren Huang, Feng-ju Lo, Ru-Yng Chang, and Sueming Chang. Reconstructing the Ontology of the Tang Dynasty: A pilot study of the Shakespearean-garden approach. 4th International Conference on Language Resources and Evaluation (LREC2004) Workshop on Ontologies and Lexical Resources in Distributed Environments (OntoLex 2004). Lisbon. Portugal. 29 May, 2004.
[15] WordNet: 1.6 to 1.7.1 Sense Mapping Package, ftp://ftp.cogsci.princeton.edu/pub/wordnet/1.7.1/WNsnsmap-1.7.1.tar.gz

以語法分析為輔建立新聞名詞知識庫

楊昌樺　陳信希

國立台灣大學資訊工程學系

chyang@nlg.csie.ntu.edu.tw; hh_chen@csie.ntu.edu.tw

摘要. 本文針對新聞文件中出現之具名實體──人名為出發，以語法分析選定同位語結構為線索，輔助建立新聞知識庫。我們採用中文新聞文件為測試題材，經由分詞詞性標記和具名實體辨識處理。在標記人名及其前後文詞性資訊後，統計常出現在人名標記之前的同位語結構。由演算法統計後的結果篩選出常用於敘述新聞人物之稱謂 (title)，藉以建立新聞領域中人物實體之知識本體雛型，以協助文件解析及機器翻譯等應用。

1　緒論

新聞媒體每天針對不同的「人、事、時、地、物」，產生許多報導和敘述，尤其在網路化的環境更是突破了時效與場合的限制，新聞文件得以電子化的型式讓大眾選閱。人們可以隨時隨地獲得最新的訊息，並追蹤感興趣事件之後續發展。由於網際電子化新聞具有上述的時空便利性，成為發展自然語言處理的基礎，使得我們能以最快的速度收集新聞文件，進而建立豐富的語料庫。近來新聞語料提供了許多系統及研究所需要的知識來源，如文件摘要系統(Chen *et al*., 2003; Lee and Ker, 2001)、跨語言資訊檢索(NTCIR-4 Test Collections, 2004)[1]、中文語料庫的建構(Ma *et al*., 2002)等等。本文擬針對新聞文件所出現之具名實體，以語法分析選定同位語結構為線索，輔助建立新聞知識庫。

從知識本體 (Ontology)[2] 的角度來看，新聞文件敘述人們這個主要實體 (entity) 在各個領域(domain)(如：國際、政治、經濟、娛樂)所參與的事件(events)。在這項前提之下，如何從新聞文件中找出實體，成為處理新聞文件最基本的工作。新聞文件中傳遞資訊的對象是一般的大眾，因此具名實體(Named Entity)的描述通常採用完整的名字，而少用指涉的方式。例如，記者使用「台東東河鄉的果農王金發先生」來取代「隔壁的老王」。既然"具名"是新聞文件中敘述之實體所具備的重要特性，因此我們可套用具名實體辨識(Named Entity Recognition; 以下簡稱NER)模型，來找出新聞文件中的人名、地名、組織名──這些名稱通常代表了新聞文件中最重要的實體。

從語法分析的角度來看，由新聞文件可歸納出許多具名實體所形成的名詞片語(NP; noun phrase)。以新聞文件中常出現的一個樣版(template)，「某人在什麼時候表示…」為例，我們節錄2004/6/30的三則新聞：

(東森新聞報)　日本首相**小泉純一郎**6月29日表示…
(中國日報)　新黨主席**郁慕明**30日表示…
(聯合新聞網)　聯發科發言人**喻銘鋰**昨天表示…

這三則新聞所描述的對象，除了人物的姓名之外，姓名前面還附上其隸屬的組織名與稱謂。組織名加上稱謂形成了另一個名詞片語，與接在後面的人名形成語法上所稱的同位語 (apposition)架構，在中研院詞庫小組所建構中文句結構樹資料庫[3]敘述的基本原則[4]中，同位語與名詞中心語等義，因此與名詞中心語構成雙岔結構。若以人名當作主要的名詞中心語，我們可以從中央研究院中文句結構樹 (以下簡稱

[1] http://research.nii.ac.jp/ntcir-ws4/data-en.html

[2] 「知識本體」參照自 http://bow.sinica.edu.tw/

[3] http://rocling.iis.sinica.edu.tw/ROCLING/Treebank/Treebank_cf.htm

[4] http://godel.iis.sinica.edu.tw/CKIP/treebank/index.html

TreeBank)中觀察到很多 形成雙岔結構的同位語 現象，如取出字串 (s1)、(s2)兩個具有相同同位語結構 NP→NP + Nba[5] (正式專有名詞)的部分剖析樹如下：

(s1) 經濟部次長江丙坤
NP(apposition:NP (property:Nca:經濟部|Head:Nab:次長)
|Head:Nba:江丙坤
)

(s2) 三重市長蔡火石
NP(apposition:NP (property:Nca:三重|Head:Nab:市長)
|Head:Nba:蔡火石
)

(s1)和(s2)這兩個句子片段的剖析皆滿足以下兩個條件(c1)、(c2)：

(c1) LHS(left-hand side)的NP和Nba為同位語、(c2) LHS的NP繼續剖析為Nca(專有地方名詞)+ Nab(個體名詞)；

並且具有性質(*1)：

(*1) 兩個連續標記成Head的名詞具有同位語法架構。

同樣的情況也出現在TreeBank中「瑞典好手貝爾格斯壯」、「琉森藝術家羅芙布姆」等字串的剖析中。另外，在下面兩個同樣從TreeBank取出語法結構為NP→NP + Nba的例子(s3)和(s4)字串中：

(s3) 湖人主力史卡特和魔術
NP(apposition:NP (property:Nba:湖人|Head:Nad:主力)
|Head:Nba(DUMMY1:Nba:史卡特|Head:Caa:和|DUMMY2:Nba:魔術)
)
(s4) 清太祖努爾哈齊
NP(apposition:NP (Head:Nba:清太祖)
|Head:Nba:努爾哈齊
)

(s3)第一個NP同位語分別剖析成「湖人」和「主力」，其詞性分別為Nba(專有名詞)和Nad(抽象名詞)。由「主力」開始連續標記成Head，因而具有性質(*1)連續同位語架構。不過與(s1)和(s2)相較之下只符合條件(c1)，並沒有符合(c2)。(s4)的「清太祖」、「努爾哈齊」連續標記成Head可視為滿足性質(*1)，但是「清」和「太祖」並沒有分開[6]，因此與(s1)和(s2)相較之下條件(c1)和(c2)無法成立。

然而(s4)的例子仍具有語法剖析上的歧義性，根據CNS14366[7]中文分詞標準，分詞的單位的性質(*2)是：

(*2) 具有獨立意義且扮演固定詞性的字串。

因此，我們也需藉由對Head是否扮演固定詞性進行分析。換句話說，我們需觀察滿足性質(*1)的第一個Head，是否也滿足性質(*2)。為了簡化討論範圍，本文不考慮(s4)這種情形的例子。

[5] 關於詞性次分類的定義參照自(中文詞庫小組, 1993)

[6] 使用 CKIP 自動斷詞系統(http://godel.iis.sinica.edu.tw/CKIP/r_content.htm)處理類似的字串「清太宗皇太極」，會將「清太宗」分詞成「清」和「太宗」。

[7] http://www.sinica.edu.tw/~ndaplib/channels/dlm_paper/0910-05.pdf，CNS14366 國家標準 (Huang, 1998)。

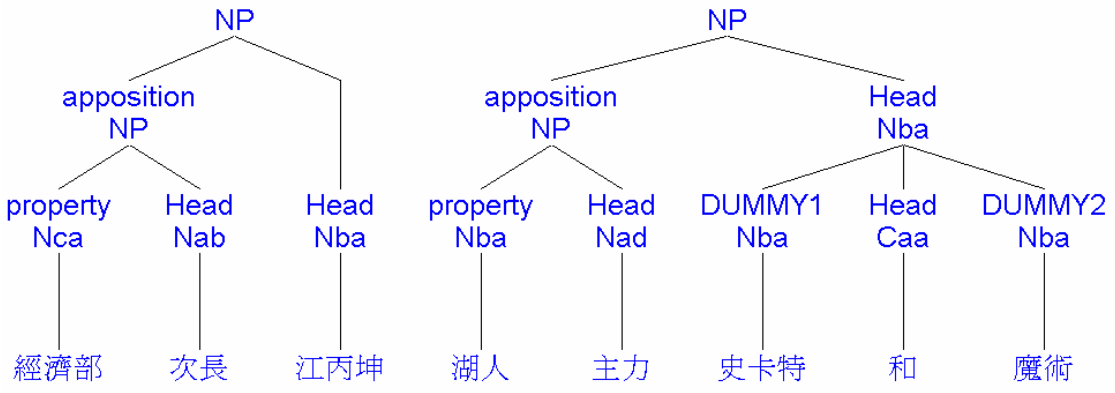


圖1 字串「經濟部次長江丙坤」和「湖人主力史卡特和魔術」的剖析樹

對照(s1)與(s3)的剖析樹，如圖1所示。可發現在人名之前的名詞片語所構成的兩個單位中，後者具有稱謂的語意特性。根據這樣的觀察，我們提出兩個假設(r1)及(r2)。在包含稱謂的名詞片語，其前後兩個單位：

(r1) 後者為比較廣泛的概念，不是專有名詞；
（次長、市長、主力）

(r2) 前者為特定的實體，通常是專有名詞。
（經濟部、三重、湖人）

(r2)應用的對象的是同位語名詞片語組中的第一個詞，在語法框架下具property特性。(r1)應用的對象是同位語名詞片語組的第二個詞，這一類詞彙也是NER模型中常用來偵測人名的線索(Chen *et al.*, 1998)。這類詞彙除了會出現在人名之前，也會出現出現人名之後。但是在中文句結構的分析中，人名之後所出現的稱謂會與該人名形成名詞片語，而不是上述分析中所出現之雙岔結構同位語現象，如下列兩個從TreeBank取出的名詞片語組：

NP(property:Nba:胡適 |Head:Nab:先生)；
NP(property:Nbc:周 |Head:Nab:老師)。

同樣的現象在新聞人名的表達上則較少出現，而且格式較為固定，通常是敘述是不需要具名或是不適合具名的新聞人物，如節錄2004/8/10的兩則新聞：

(東森新聞報) 王姓**民眾**在花蓮縣…
(中時電子報) 台北市某國小的徐姓男**老師**涉嫌…
(聯合新聞網) 台南市廢五金業馬姓**少東**…

這兩則新聞針對描述對象所使用的表達方式，除了沒有列出完整的姓名外，也不適用(r1)及(r2)所提出的假設。

本文分析出現在人名標記之前的同位語，篩選出常用於敘述新聞人物之稱謂結構，內容共分成四節，第一節描述目標與語言現象，第二節提出方法、實驗題材，並作評估。第三節顯示所自動擷取的知識庫和瀏覽介面。最後第四節作一總結。

2 研究方法與實驗

根據(r1)和(r2)這兩個假設，在不使用中文剖析器(Parser)的情況下，我們設計以下字串檢驗的方法。由於沒有高準確性之中文剖析器輔助，我們將觀察的對象調整成三連字串(trigram)ABC，在C為人名的前提下，測試問題(q1)及(q2)是否滿足：

(q1) A是常出現的名詞且符合(r2)；

(q2) 當A、C成立時，B滿足(r1)的可能性。

我們使用大量未經人工標記之新聞文件，經由分詞與詞性標記系統處理後，用finite-state的方式取得三連字串(trigram)──依序輸入字串Ａ、Ｂ及Ｃ，當state滿足C為人名且A為名詞時，即取出此trigram作為候選字串。接著設計演算法(詳見2.3節)，篩選候選字串，留下高頻率之字串集合。

2.1 實驗新聞語料

實驗所使用的語料來自三個新聞來源，分別是中央社從2002/1/1到2002/12/31共24,342則新聞、中時新聞從2002/1/1到2002/12/31共82,606則新聞、以及聯合新聞從2002/04/09到2003/06/26共100,617則新聞[8]。新聞的統計資訊如表1所示，詞數為經分詞與詞性標記系統處理(詳見2.2節)後得之。統計顯示中央社每則新聞平均詞數最多、聯合新聞則數最多、中時新聞平均單則新聞詞數最少。合計的語料詞數規模為1 : 2.33 : 3.86。

表1 新聞文件數量與詞數分季統計表

	中央社		中時新聞		聯合新聞	
	新聞則數	詞數	新聞則數	詞數	新聞則數	詞數
2002第一季	7,663	3,340,600	20,114	7,879,821		
2002第二季	6,500	3,048,492	21,514	6,946,475	21,694	8,429,921
2002第三季	5,290	2,901,751	20,436	6,528,886	30,961	13,054,258
2002第四季	4,889	2,676,842	20,542	6,577,950	28,912	13,522,185
2003第一季					10,215	6,048,806
2003第二季					8,835	5,159,618
合計	24,342	11,967,685	82,606	27,933,132	100,617	46,214,788
數量比例	1	1	3.39	2.33	4.13	3.86
平均詞數(比)	491.65 (1)	338.15(0.69)	459.3139			

2.2 分詞與詞性標記系統

本實驗使用台大自然語言處理實驗室分詞與詞性標記系統(以下簡稱Tagger)，處理上述所有新聞文件。該系統核心採取Tigram Markov模型，整合入NER[9]模組(Chen *et al*., 1998)，其中本實驗著重之人名辨識核心是在系統中是(Chen and Lee, 1996) 所提出的架構。該系統針對以句子為單位之字串如「財政部次長林宗勇表示，」產生如下之分詞與詞性標記結果：

財政部(Nc) 次長(Na) 林宗勇(Nb_PERSON) 表示(VE) ，(COMMACATEGORY)

下節敘述將Nc、Na、 Nb_PERSON等詞性，或具名實體標記統稱為「詞性碼」。

[8] 聯合新聞語料日期之收集較前兩家晚一至二季，可保留偵測新人名之擴充性。

[9] NER 辨識系統之使用可參照 http://nlg.csie.ntu.edu.tw/

2.3 演算法

實驗新聞語料經由Tagger處理後，採取計算詞頻的方式，來篩選新聞人物常使用的同位語稱謂 。其演算法詳述如下：

(1) 從Tagger結果中找出任何9-gram[10]，令其為XXXABCXXX，當其中的mid-trigram(亦即ABC)滿足第一字A的詞性碼第一碼為N，且第三字 C 的詞性碼為Nb_PERSON。

(2) 蒐集每一種tri-gram與其詞性配對<A, tagA, B, tagB, C, tagC>形成集合S，並統計其出現頻率 $freq_{<A, tagA, B, tagB, C, tagC>}$ 。

(3) 統計第二詞及詞性配對<B, tagB>出現的頻率$freq_{<B, tagB>}$。

(4) 篩選B詞長度大於2，tagB第一碼為N，且$freq_{<B, tagB>}$大於β [11]。

(5) 人工檢驗選出之B詞是否為稱謂，通過檢驗之B詞集合為S_B。

(6) 取子集合S'={s ls屬於S, 且s中對應的<B, tagB>屬於S_B、$freq_{<A, tagA, B, tagB, C, tagC>}$＞1}。以通過檢驗之稱謂篩選S並保留其詞頻>1之資訊。子集合說明篩選之同位語片語除人名 C前出現常見之稱謂B外，同樣的同位語片語ABC必須出現兩次以上。)

2.4 評估分析

表2列出三家媒體選出的稱謂字數分別為65、103及125，比例是1 : 1.58 : 1.92。中央社新聞篩選出的稱謂正確率達100%、中時新聞為97.01%、聯合新聞為93.60%。我們發現篩選的稱謂詞數，跟新聞語料分詞的數量，呈現正相關，但不是以等比增長。這顯示新聞人物的稱謂不會毫無限制地增加，應該會限定在一個範圍內。

第2.3節演算法(1)-(5)步驟，找出所有trigram ABC中限定A、C的詞性之後，B的詞性碼第一碼為N的詞。(因此B在這樣的限定下tag可能是Na、Nb、等合法的名詞。)接著檢驗問題(q2)，我們觀察B是否符合(r1)為廣泛的稱謂概念(通常是Na)。觀察由中央社篩選出的65個名詞列出如表3所示，列出的65個詞在Tagger皆標記成Na，因此假設(r1)大致成立。若將演算法步驟(4)的β設成50，則會篩選出102個詞，其中B不是Na或意思有誤的有8組，分別是：

第67名的「批評」	(Na;	頻率99)、
第71名的「要求」	(Na;	頻率89)、
第75名的「支持」	(Na;	頻率81)、
第78名的「政務委員」	(Nb;	頻率70)、
第83名的「土耳其」	(Nc_LOCATION;	頻率65)、
第87名的「國小」	(Nc;	頻率61)、
第90名的「颱風」	(Na;	頻率59)、
第97名的「行政長官」	(Nb;	頻率53)，

其中除「政務委員」、「行政長官」應是Tagger的錯誤，其他三種錯誤有動詞名物化如「批評」、地名如「土耳其」、事物擬人化如「颱風」等，系統正確率因此降至94.12%。

[10] 取 9-gram 可保留觀察前後文之擴充性。

[11] β 為本實驗自訂之參數，β=100 次，要求稱謂字串出現頻率要超過 100 次。

表2 自動篩選稱謂之結果統計表

	自動選出 B字(比)	人工檢驗 正確字數	正確率	正確稱謂 詞頻第一名	正確稱謂 詞頻第二名	錯誤稱謂 詞頻第一名
中央社	65(1)	65	100.00%	立委(3,656)	總統(3,119)	
中時新聞	103(1.58)	100	97.01%	局長(2,742)	立委(2,414)	選區(166)
聯合新聞	125(1.92)	117	93.60%	總經理(5,671)	董事長(5,599)	分析(824)

表3 中央社新聞稱謂篩選統計表

稱謂	頻率	稱謂	頻率	稱謂	頻率	稱謂	頻率	稱謂	頻率
立委	3656	立法委員	618	執行長	352	副院長	198	大使	143
總統	3119	副主席	560	國務卿	351	資政	194	里長	141
市長	1991	縣長	549	院長	344	司長	187	首相	138
主席	1836	候選人	511	會長	335	負責人	185	參議員	134
秘書長	1155	處長	491	召集人	316	副總統	184	秘書	132
主委	1102	教授	481	外長	295	經理	184	總幹事	125
黨主席	1094	領袖	454	祕書長	291	行政院長	179	委員長	125
局長	1087	代表	452	署長	277	校長	178	鄉長	120
部長	1038	總經理	443	國防部長	270	領導人	175	議員	112
董事長	1014	次長	420	夫人	263	總裁	173	研究員	107
主任	1009	議長	410	市議員	260	專家	151	常委	105
發言人	966	理事長	400	顧問	235	副總理	147	官員	105
總理	734	委員	363	檢察官	214	組長	146	記者	102

表4 新聞人物同位語結構字串篩選統計表

		中央社		中時新聞		聯合新聞	
篩選數(比)		2,399(1)		3,927(1.64)		5,463(2.28)	
前50名	1.專有地名	16	32%	14	28%	14	28%
	2.專有組織名	4	8%	6	12%	5	10%
	3.普通組織名	10	20%	9	18%	4	8%
	4.普通名詞	16	32%	17	34%	24	48%
	5.斷詞錯誤	4	8%	4	8%	1	2%
	6.無法判斷					2	4%
正確字串頻率第一名		台北市長馬英九		台北市長馬英九		美國總統布希	
正確字串頻率第二名		美國總統布希		美國總統布希		台北市長馬英九	
正確字串頻率第三名		中國國民黨主席連戰		法務部長陳定南		國家主席江澤明	

表4列出三家媒體篩選出的同位語結構字串數分別為2,399、3,927、5,463，比例是1：1.64：2.28，數量上依然跟新聞語料分詞的數量呈現正相關。雖然倍數不是以等比增長，但是較稱謂筆數的比例來的放大，顯示同樣一種稱謂可能可以成功套用在多筆結構上，如「民進黨立委蔡同榮」、「國民黨立委陳學聖」等。

我們分別針對三家新聞媒體語料，抽出出現頻率前50名同位語結構字串，針對trigram的A作型態的判斷，共分成六種類型：

(1) 專有地名：　　如台北、美國等。

(2) 專有組織名：　　如中國國民黨、民進黨等。

(3) 普通組織或地名(Nc)：　如總統府、行政院等。

(4) 普通名詞(Na)：　　如國家(主席)、法務(部長)等。

(5) 斷詞錯誤：　　如國民(黨主席)、(親)民黨(副主席)。

(6) 無法判斷：　　如橘子(董事長)──橘子標記成名詞，其實是專有組織名、鴻海(董事長)──鴻海標記成人名，其實是專有組織名。

結果顯示各家媒體使用專有名詞的部分(1)+(2)的比例約為40%，使用普通名詞的比例約為50%，錯誤及無法判斷的比例在10%以下。

回到問題(q2)，我們觀察A是否符合(r2) 為特定的實體，由實驗結果顯示僅有 40%的機會為專有名詞，許多Na及Nc都可以用來限定稱謂，例如用「集團」(Na)限定「董事長」、用「總統府」(Nc)限定祕書長。

另外，使用三連字串的設計來偵測字串本身即有限制，如果是以名詞片語「政治大學(Nb)新聞系(Nc)教授(Na)」為例，其同位語架構中的A，即包含了開頭的Nb和Nc兩部份，因此需要擴充2.3節演算法的步驟(2)，蒐集四連字串才能取得。以中央社的新聞為例，以擴充後的演算法，透過四連字串可選出861筆資料，茲列出前十名如表5所示。使用五連字串可選出308筆資料，前十名如表6所示。使用六連字串[12]可選出112筆資料，前十名如表7所示。

針對使用不同長度的n-gram所獲得的結果顯示，使用的n-gram越大，獲得的詞數就越少，而且所需要的計算時間隨之增長；若欲透過增加語料庫規模來增加獲得的詞量，也會增加計算的時間。表5、表6及表7列出結果中標示錯誤的地方，是由於Tagger將「副總」分成一個詞，以及沒有收錄「(親)民黨」詞彙所造成。

綜合以上三個表格所提供之觀察，發現同位語元素Head B仍是緊鄰在人名之前，因此驗證了本文所提出演算法之可行性，但是property A的長度限制會是問題所在。同樣的問題也說明了設計一個良好的中文剖析器會遇到的困難──因為無法保證同位語結構中A能允許的詞數能有多長。

表5. 中央社新聞人物同位語結構四連字串前十名列表

正確	<A_1, tagA_1, A_2, tagA_2, B, tagB, C, tagC>								次數
√	中共	Nb	國家	Na	主席	Na	江澤民	Nb_PERSON	319
√	中央	Nc	委員會	Nc	祕書長	Na	林豐正	Nb_PERSON	149
√	中共	Nb	國家	Na	副主席	Na	胡錦濤	Nb_PERSON	147
√	中央	Nc	委員會	Nc	秘書長	Na	林豐正	Nb_PERSON	139
√	發展	Na	委員會	Nc	主委	Na	陳健治	Nb_PERSON	103
√	政策	Na	委員會	Nc	執行長	Na	曾永權	Nb_PERSON	93
√	高雄	Nc_LOCATION	市議會	Nc	議長	Na	黃啟川	Nb_PERSON	79
√	臺北	Nc_LOCATION	市長	Na	候選人	Na	李應元	Nb_PERSON	68
X	黨團	Na	副總	Na	召集人	Na	秦慧珠	Nb_PERSON	67
√	高雄	Nc_LOCATION	市長	Na	候選人	Na	黃俊英	Nb_PERSON	64

[12]由於 2.3 節演算法步驟(1)從語料庫中保留了 9-gram，因此本實驗可偵測字串之最長長度，是使用六連字串的所偵測出的 AAAABCXXX。

表6. 中央社新聞人物同位語結構五連字串前十名列表

正確	<A_1, tagA_1, A_2, tagA_2, A_3, tagA_3, B, tagB, C, tagC>										次數
√	組織	Na	發展	Na	委員會	Nc	主委	Na	陳健治	Nb_P	103
√	中國國民黨	Nb_O	中央	Nc	委員會	Nc	祕書長	Na	林豐正	Nb_P	62
√	國民黨	Nb_O	中央	Nc	委員會	Nc	秘書長	Na	林豐正	Nb_P	56
√	中央	Nc	政策	Na	委員會	Nc	執行長	Na	曾永權	Nb_P	54
√	民進黨	Nb_O	臺北	Nc_L	市長	Na	候選人	Na	李應元	Nb_P	52
√	中國國民黨	Nb_O	中央	Nc	委員會	Nc	秘書長	Na	林豐正	Nb_P	38
√	中國	Nc_L	大陸	Nc	國家	Na	主席	Na	江澤民	Nb_P	33
X	立法院	Nb_O	黨團	Na	副總	Na	召集人	Na	秦慧珠	Nb_P	32
√	白宮	Nb_O	國家	Na	安全	Na	顧問	Na	萊斯	Nb_P	31
√	經濟	Na	研究	Na	中心	Nc	主任	Na	林毅夫	Nb_P	30

表7. 中央社新聞人物同位語結構六連字串前十名列表

正確	<A_1, tagA_1, A_2, tagA_2, A_3, tagA_3, A_4, tagA_4, B, tagB, C, tagC>												次數
√	中國國民黨	Nb_O	組織	Na	發展	Na	委員會	Nc	主委	Na	陳健治	Nb_P	55
√	國民黨	Nb_O	組織	Na	發展	Na	委員會	Nc	主委	Na	陳健治	Nb_P	34
X	民黨	Nb	立法院	Nb_O	黨團	Na	副總	Na	召集人	Na	秦慧珠	Nb_P	30
√	中國	Nc_L	經濟	Na	研究	Na	中心	Nc	主任	Na	林毅夫	Nb_P	30
√	國民黨	Nb_O	中央	Nc	政策	Na	委員會	Nc	執行長	Na	曾永權	Nb_P	28
√	中國國民黨	Nb_O	中央	Nc	政策	Na	委員會	Nc	執行長	Na	曾永權	Nb_P	23
√	行政院	Nb_O	農業	Na	委員會	Nc	主任	Na	委員	Na	范振宗	Nb_P	19
√	美國	Nc_L	聯邦	Na	準備	Na	理事會	Na	主席	Na	葛林斯潘	Nb_P	17
X	民黨	Nb	立院	Nb_O	黨團	Na	副總	Na	召集人	Na	秦慧珠	Nb_P	14
X	民黨	Nb	立法院	Nb_O	黨團	Na	副總	Na	召集人	Na	李鴻鈞	Nb_P	13

3 知識庫整理與視覺化瀏覽介面

目前由演算法自動擷取出來的各報社人名知識庫 (如中央社的2,399筆資料)，透過知識本體介面對這些資料進行瀏覽，可以協助系統或使用者了解與整理所擷取的知識。圖2及圖3分別使用中央社新聞及中時新聞，以樹狀結構的方式，對照兩個新聞資料庫所收錄的人名實體及事件關係。由左側視窗Root「新聞人物」，首先列出通過檢測之新聞人物稱謂成為children(trigram的B部分)(如中央社的65筆資料)，接著列出該稱謂的所有property(trigram的 A部分)，圖示中同一個node的children以顯示五筆為上限。右側上半部列出具有稱謂人物後面常出現之事件，例如點選「民進黨」後出現之動詞片語如所列。下半部分則列出點選稱謂下特定property所代表人物後面常出現之事件。

以圖2為例，「立委」人物的後面常出現的動詞依序有有「表示」、「指出」、「質詢」、等等，與立委分類中的「民進黨立委」，後面常出現動詞的順序大致相符，顯示中央社報導的立委行為有一定的模式。圖3同樣是以「立委」為例，但中國時報的動詞順序便有些許不同，不過大致上仍是相符。另外觀察的window size若放大到2，也可計算出常出現的動詞片語有「昨天召開」、「昨日表示」、等等。

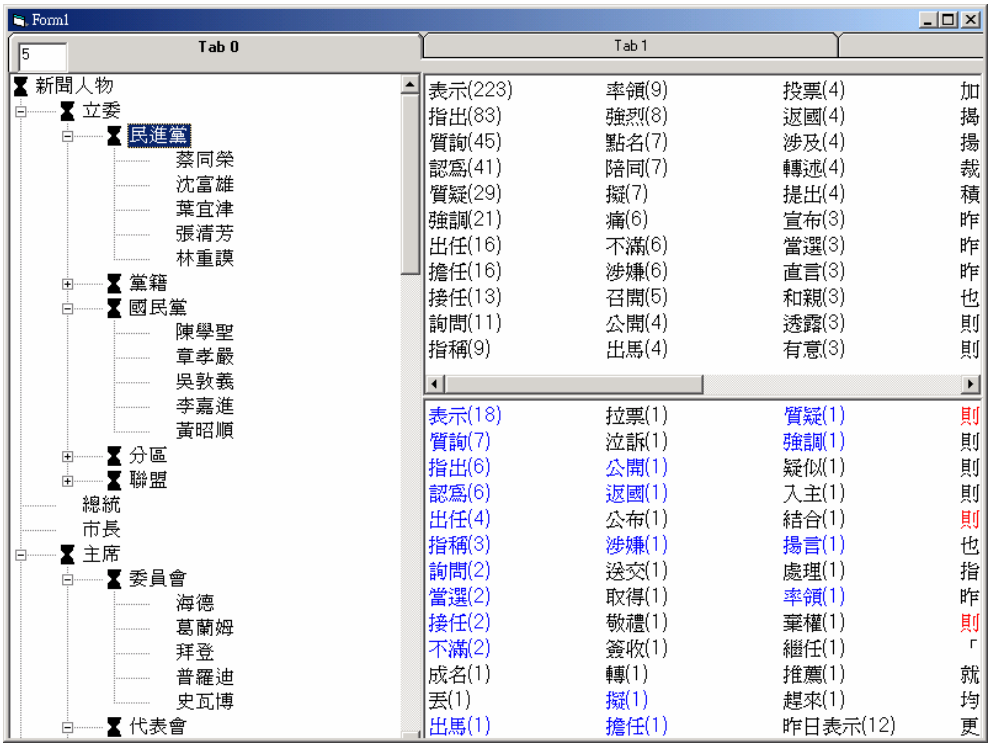


圖2 瀏覽中央社新聞人名之知識庫介面

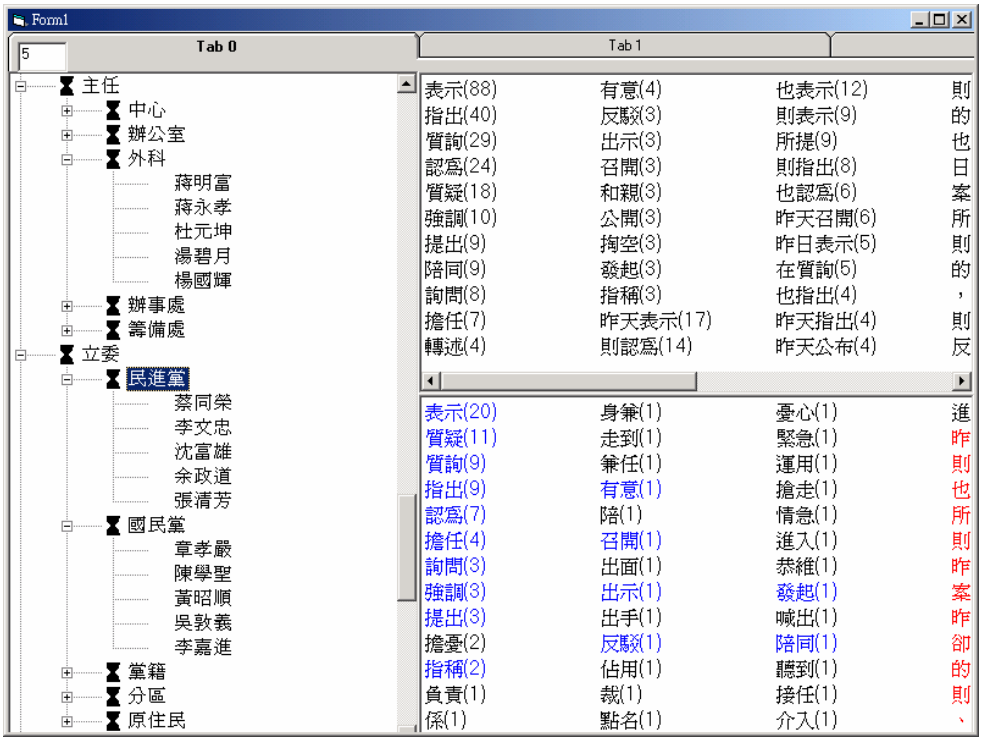


圖3 瀏覽中時新聞人名之知識庫介面

4 結論

新聞媒體透過網路環境突破時空的限制，隨著時間不斷的推移提供更多更新的文件。一個與時並進的自然語言系統，也可以利用網路新聞的這項特質，獲取最新的字詞統計資訊，以延伸出各種形式的應用，如對n-gram機率的估計、衡量(單語或跨語言)字詞間的相關度、建立同義詞集或知識本體(Ontology)等等。

透過本文敘述的方法所建立的人名實體與事件關係的雛型，將可以協助新聞文件的解析和機器翻譯的應用。例如在一個敘述多個實體的事件報導中，運用本知識庫所建立的概念關係，可以挑出適合的指涉人物，以解決指涉問題。例如提及「立委」和「主任」之後，誰是接著敘述「涉嫌...」事件所指涉的對象。另外，如英文的NER及剖析系統技術都較中文純熟，如果能用同樣的機制針對英文新聞建立知識庫，將可對不同語言，但屬於新聞領域知識庫所包含的實體及事件，透過連結的方式，建立跨語言的對應關係。

註謝

本文部分成果為國科會計畫NSC 93-2752-E-001-001-PAE補助。

參考文獻

Hsin-Hsi Chen, June-Jei Kuo, Sheng-Jie Huang, Chuan-Jie Lin and Hung-Chia Wung (2003). "A Summarization System for Chinese News from Multiple Sources." Journal of the American Society for Information Science and Technology, 54(13), pp. 1224-1236.

Hsin-Hsi Chen, Yung-Wei Ding and Shih-Chung Tsai (1998). "Named Entity Extraction for Information Retrieval." Computer Processing of Oriental Languages, Special Issue on Information Retrieval on Oriental Languages, 12(1), pp. 75-85.

Hsin-Hsi Chen and Jen-Chang Lee (1996). "Identification and Classification of Proper Nouns in Chinese Texts." Proceedings of 16th International Conference on Computational Linguistics, Copenhagen, Denmark, August 5-9, pp. 222-229..

李祥賓、柯淑津 (2001)，「新聞文件摘要之研究」，第十四屆計算語言學研討會論文集，pp. 23-42，台南成功大學。

馬偉雲、謝佑明、楊昌樺、陳克健 (2001)，「中文語料庫構建及管理系統設計」，第十四屆計算語言學研討會論文集，pp. 175-191，台南成功大學。

中文詞知識庫小組 (1993) ，「中文詞類分析(三版)」。

The Construction of a Chinese Named Entity Tagged Corpus: CNEC1.0

Cheng-Wei Shih, Tzong-Han Tsai, Shih-Hung Wu,
Chiu-Chen Hsieh, and Wen-Lian Hsu

Institute of Information Science, Academia Sinica
{dapi,thtsai,shwu,gladys,hsu}@iis.sinica.edu.tw

Abstract. In order to build an automatic named entity recognition (NER) system for machine learning, a large tagged corpus is necessary. This paper describes the manual construction of a Chinese named entity tagged corpus (CNEC 1.0) that can be used to improve NER performance. In this project, we define five named entity tags: PER (person name), LOC (location name), ORG (organization name), LAO (location as organization), and OAL (organization as location) for named entity categories. In addition, we propose a special tag, DIFF (Difficulty), to annotate ambiguous cases during corpus construction. A, corpus-annotating procedure, a tagging tool, and an original corpus are also introduced. Finally, we demonstrate a part of our manual-tagged corpus.

1 Introduction

Named entity recognition (NER), which includes the identification and classification of certain proper nouns in a text, is an important task in information extraction. It is useful in many natural language processing systems for document indexing and managing data with named entities [Tsai et. al 2004]. Since numerous new proper nouns are generated every day, it is not enough for an IR system to index names from Internet documents or refer to gazettes. Therefore, NER has become an important method for information processing in recent years.

Machine learning (ML) is one of the most popular methods in NER, due to its easy maintenance and portability [Tsai et. al 2004]. Typical machine learning approaches applied in NER include the Hidden Markov Model (HMM) [Bikel et. al 1997], Support Vector Machine (SVM) [Asahara 2003], and Maximum Entropy (ME) [Borthwick 1998]. No matter which approach is used, a tagged named entity corpus with clear annotating criteria is needed in the training phase of building an NER system. However, constructing such a corpus is a labor-intensive task, so few researchers have focused on it. The Automatic Content Extraction program (ACE) executed by the Linguistic Data Consortium (LDC) [Http://wave.ldc.upenn.edu/] annotates seven common entities in English, simplified Chinese, and Arabic. Meanwhile, the IREX (Information Retrieval and Extraction Exercise) [Http://nlp.cs.nyu.edu/irex/] defines 8 kinds of named entities (NE) in Japanese [Sekine and Isahara 2000], and the shared task in CoNLL 2002 and 2003 [Erik 2002] [Erik et. 2002] develops the NER system using four types of NE in English, German, Dutch, and Spanish. However, as none of these methods focus on traditional Chinese, there is an urgent need for a traditional Chinese NE corpus and NE annotating standards to support an automatic Chinese NER system like Mencius [Tsai et. al 2004].

The categories of named entities defined by Message Understanding Conferences (MUC) are the names of persons, organizations, locations, temporal expressions and number expressions [Grishman and Sundheim 1996]. Since temporal and number expressions, such as “the past year” and “40 percent”, are generally used as adjectives to describe other entities, we disregard them and focus on the annotation of person names, organization names, and location names as NEs. We separate organizations and locations into four non-overlapping categories to accommodate common Chinese usage. We also propose a temporary tag, “Difficulty”, to represent named entities that are ambiguous.

The remainder of this paper is organized as follows. Section 2 discusses the main issues of labeling named entities. Section 3 introduces all the NE categories used in CNEC1.0. Section 4 describes our annotation procedure and environment. Finally, in Section 5, we present our conclusion and the direction of future research.

2 Named entities annotation issues

The applications of a corpus determine the kinds of entities to be tagged. In our project, the NER system should support information extraction, question answering, and information retrieval of new documents. The following three issues should be considered before tagging Chinese named entities.

2.1 Proper nouns

We think named entities should be proper nouns. Therefore, each named entity should denote a unique object, so words without uniqueness should not be annotated. For example, in the sentence “他把車停放在停車場/He parked his car in the parking lot.”, we cannot be sure which parking lot he parked in. Therefore, the term “停車場/parking lot” will not be labeled

2.2 Inner feature and Outer feature

Generally, a named entity can be determined in three ways: viewing its literal meaning, checking the context, and semantic understanding as shown in the following example:

“台北市長馬英九/ Mayor of Taipei City, Ma Ying-Jeou” (1)

“台北車站人潮洶湧/Taipei Main Station is crowded.” (2)

“我聽說遠東搬家了/I heard that Yuan-Dong has moved out.” (3)

Obviously, we can easily determine that the term “馬英九/ Ma Ying-jeou” in Sentence (1) is a person name according to the position of the title “台北市長/ Mayor of Taipei City”. “台北車站/Taipei Main Station” in sentence (2) can be identified as a location name because of the term “車站/Station”. However, sometimes we cannot identify or judge a word as a proper noun by its literal meaning. For example, in Sentence (3), we do not know if the term “遠東/ Yuan-Dong” represents a person or a company. These kinds of inaccuracies are due to abbreviations or borrowings. For this reason, two types of feature are used to classify the recognition modes of named entities：the inner feature and the outer feature. In the above examples, sentence (1) is the outer feature type, while sentence (2) can be classified by its inner features. In our work, we do not limit the types of features annotators apply during tagging, but if a named entity cannot be identified by both inner and outer features, as in sentence (3), we ask annotators not to mark it. This eliminates confusing terms and keeps the corpus as clear as possible

2.3 Maximum and minimum semantic unit matching

Named entities are occasionally nested or appear next to one another in a text. In some cases we can combine them to form a larger entity because they may describe the same object. Therefore, determining the boundary of named entities is an important issue. We have found that different named entities have a corresponding annotation policy, which can be classified as maximum and minimum semantic unit matching. Minimum semantic unit matching is recommended for named entities such as person names and location names because these entities singly represent a unique item. For example, the sentence “台北縣板橋市/Ban-Qiao City, Taipei County” is tagged as two place names because “台北縣/Taipei County” and “板橋市/Ban-Qiao City” both denote specific independent entities. On the other hand, named entities such as organizations should apply the maximum unit policy. The term “台北市環保局/Department of Environmental Protection, Taipei City Government” cannot be separated into “台北市/Taipei City” and “環保局/Department of Environmental Protection” for retaining the original meaning.

3 Named entity categories

We propose five target named entity categories for annotating the "unique identifiers" of entities, including organizations, persons and locations, as well as one function tag. These are shown in Table 1 *and explained* in the following sub-sections. For practical purposes, we began our experiment with these NE tags.

Table 1. Tag set of the NE corpus

DIFF	**Difficult problem**
PER	**Person name**
LOC	**Location name**
ORG	**Organization name**
LAO	**Location as Organization**
OAL	**Organization as Location**

3.1 Difficult problem - DIFF

Diff (Difficult problem) is designed to identify problems such as nested, ambiguous, or poorly defined NEs. It addresses controversial items in Chinese named entity identification.

In our opinion, DIFF is essential for identifying ambiguous cases. Named entities that are difficult to classify are isolated from others for data cleansing to ensure that the content of the corpus is clear. DIFF entities will become the future expansion direction of our NER processing domain.

3.2 Person name - PER

Traditionally, the structure of Chinese person names follows the principle that the surname (one or two characters) is placed before the person's chosen names (one or two characters). In our research, the annotation of person names follows this principle. But some entities with "person" meaning as Diff tag such as nicknames, incomplete Chinese person names, foreigners' names and pronouns, are marked as Diff. These exceptions are discussed below.

3.2.1 Nicknames

Nicknames are not only given to people, but are sometimes given to pets or even objects like toys and vehicles. Because of their uniqueness, we mark nicknames as DIFF within a context, as the following example shows.

[小炳<DIFF>] 疼愛的女兒 [央央<DIFF>]

[xiao-bing<DIFF>] loves his daughter [yang-yang<DIFF>]

3.2.2 Incomplete Chinese person names

Following the full name principle, an incomplete Chinese person name indicates that the surname or chosen name may be omitted. For instance, "Shin, Cheng-Wei" is a full person name, but we sometimes only use the chosen name "Cheng-Wei" to address the person. Another example of an incomplete person name is a surname that follows a title or an appellation such as "李先生(Mr. Lee)". Both of these cases are tagged as DIFF.

[陳總統<DIFF>] 上午前往日本北海道遊玩

[President Chen<DIFF>] went to Hokaido for sightseeing this morning.

3.2.3 Foreigners' names

Chinese NER has difficulty dealing with foreigners' names because of the following name constructions: direct translation, Japanese person names, and Korean person names. First, direct translation cannot normally meet the principle of Chinese person names. For example, the Chinese translation of "Mel Gibson" is "梅爾吉勃遜(Mei-er-ji-bo-xun)", which obviously doesn't meet the naming rule. Second, Japanese mostly uses Chinese characters for person names, but, there isn't a surname or composite first name as in Chinese person names. For example, in "日本首相(Japanese prime minister)[小泉純一郎](Junichiro Koizumi)", Koizumi is his surname

and Junichiro is his first name. These names don't match the Chinese person naming principle. Finally, we treat Korean person names as PER because most of them match the Chinese naming rules, e.g. "李英愛 /Lee-Ying-Ai".

As most foreigners' names may cause confusion in Chinese NER, we use DIFF tag as a temporary solution in CNEC 1.0 to solve the problem.

[梅爾吉勃遜<DIFF>] 導演受難記名利雙收

[Mel Gibson<DIFF>] has achieved both fame and wealth by directing the movie "The Passion of the Christ."

3.2.4 Pronouns

Some NER researchers claim their systems can handle pronouns as named entities for person names. However, because of the "uniqueness" of NEs, pronouns are beyond our scope and we do not annotate them as NE tags.

3.3 Location names - LOC

Basically, a location name pinpoints a place's geographical position on an accurate map or in other reference material. Proper names like "Hyde Park", "New York Art Theater" or "Berlin Wall" are suitable for NER, but terms like "a park", "a theater", or "a wall" are not. So the main purpose in tagging location names is to recognize an existent location in geographic.

A location name included in another compound word such as "西班牙海鮮飯"(Spanish seafood rice) is an issue in NE annotation. In Chinese, a term's noun and adjective forms are the same, In this case, "Spanish" and "Spain" are translated as the same Chinese word(西班牙). Therefore, we suggest a syntactic frame: insert "de（的）" between a possible location name and the other words close to (A de(的) B) to solve such cases.

Chinese Word 1:	西班牙海鮮飯
In English:	Spanish seafood rice
[A de B] in Chinese:	[西班牙][的][海鮮飯]
In English:	[Seafood Rice] [of] [Spain]

Maximum and minimum semantic unit selection (section 2.3) is regarded as a Chinese segmentation problem. For example, in the phrase "美國(United States)德州(Texas)奧斯汀(Austin)", there are no spaces between the words in written Chinese. Location follows minimum semantic unit matching to tag the example as [美國(United Sate)<LOC>]、[德州(Texas)<LOC>]、[奧斯汀(Austin)<LOC>]. In addition to the basic tagging rule, several location types have to be labeled

3.3.1 Roads, sections, and addresses

Address' location information should be marked from country name, state, city, road (Boulevard, avenue, street, etc.,) to section. Other information in an address is excluded.

[忠孝東路四段<LOC>] 100號

[Zhong-Xiao East Road, Section 4<LOC>], No. 100

Sometimes a road section can be described in a city-section or area-section. In this case the road name and its description should be tagged separately.

[西濱快速道路<LOC>] [嘉義<LOC>] 段

[Xi-Bin Express Way<LOC>] [Jia-Yi<LOC>] Section

3.3.2 Location abbreviations

Location abbreviations are terms composed of two or more place names in a single entity. The other type of location abbreviation is multi-name expression containing conjoined modifiers. For instance, 中(Taichung)彰(Changhua)投(Nantou)地區(area) is a common way to describe the location of three neighboring cities in Taiwan. We suggest that such cases should be tagged as DIFF.

[桃竹苗<DIFF>] 地區連日豪雨

[Tao-Chu-Miao<DIFF>] area it has been raining torrentially for a couple of days.

Tao-Taoyuan; Chu: Hsihchu; Miao: Miaoli

3.3.3 World place names

World place names can be divided into two sets: locations written in a foreign language and translated location names (written in Chinese). A translated name such as "紐約/New York" can be considered as a location name in CNEC 1.0. But an original name, like Tokyo, should be tagged as DIFF. The following two sentences demonstrate the criteria we set.

[密蘇里州(Missouri State)<LOC>] 的 [聖路易市(St. Louis City)<LOC>]

[Missouri<DIFF>]州的 [St. Louis City<DIFF>].

#[St. Louis City<LOC>] in [Missouri State<LOC>]

3.4 Organization names - ORG

In general, organizations include companies, government bodies, institutes, and other organized groups. We define an organization as having the ability to execute plans and projects. The tagging of ORG has to apply maximum semantic unit matching. A typical case is shown below.

[裕隆汽車三義廠(Yulon Motor Sanyi plant)<ORG>]

The maximum semantic unit is used because this entity cannot be separated into "裕隆汽車(Yulon Motor)" and "三義廠(Sanyi plant)". "Sanyi plant" cannot be tagged as a LOCATION according to the uniqueness characteristic of an NE. Like location names, some ambiguity may occur in the following cases.

3.4.1 Organization abbreviations

Organization abbreviation tagging follows the rules for location abbreviations described in Subsection 3.3.2. For example,

[國親新(Guo-Qin-Xin)<DIFF>] 議員下午到[健保局(BNHI)<ORG>]

Guo-Qin-Xin councillors went to the Bureau of National Health Insurance this afternoon.

Guo: Kuomintang; Qin: People First Party; Xin: New Party

3.4.2 Foreign organization names

As in Section 3.3.3, an original organization name is regarded as DIFF in CNEC 1.0, but a translated name is tagged as a location name.

[惠普<ORG>] 與 [微軟<ORG>] 的共同秘密

The secret of [HP<ORG>] and [Microsoft<ORG>]

3.4.3 Groups, bands, crowds, and teams

Through the tagging process, we have found that most Chinese people have a problem tagging similar concepts like groups, bands, crowds and teams as organizations. Hence, we give a definition of an organization to help annotators determine if a term is an ORG. “An organization has five fundamental parts: a founder, capital, structure (departments, section, class, etc.,), a hierarchy (chief, director, dean) and employees.” According to this definition, we do not mark a term as an organization if it doesn’t have an organized structure.

3.5 LAO and OAL

Location as Organization (LAO) and Organization as Location (OAL) are proposed for semantically meaningful NE in Chinese NEC. In some cases, “location” represents an organization-like role to make decisions or to perform some duties. Compare the following two examples from the China Times news corpus:

1. “[台北市政府<ORG>]同意[總統府<OAL>]前的集會遊行”

[Taipei municipal government<ORG>] agreed to the protest in front of the [presidential plaza <LOC>].

2. [總統府<ORG>]發表一中一台政策

[Presidential Office <LAO>] announced “one China, one Taiwan policy”.

The above examples show that the term “總統府(presidential plaza / presidential office)” has a double meaning in Chinese: location and organization. The first term “總統府” is obviously a place name, but, the second “總統府” is an organization. This use of the same term to indicate a place name and an organization name, we call it “borrowing”, is common in Chinese. Sometimes we cannot sure if a location entity is a real location name or just a borrowing. In order to avoid confusion, we separate the LAO tag and OAL tag in location and organization names.

Differentiating between a borrowing and a true NE depends on an entity’s category., By deciding which category an entity belongs to in common usage, we can tell whether it is a borrowing, or not. For example, a country’s name can refer to its geographical position, but it may also be used as an organization name, as in: “中國昨日警告美國停止對軍售/China warned the United States yesterday to stop selling advanced arms to Taiwan.” Obviously, in this case we can tell the country names are borrowings and should be annotated as LAO as follows:

[中國<LAO>] 昨日警告 [美國<LAO>] 停止對 [台灣<LAO>] 軍售

[China<LAO>] warned the [United States<LAO>] yesterday to stop selling advanced arms to [Taiwan<LAO>]

OAL can be identified in the same way:

這輛巴士有到 [行政院<OAL>]

This bus goes by [the Executive Yuan<OAL>]

4 Manual Annotating Process

The most famous corpus in Chinese NER is MET-2 made by MUC [Chinchor, 1998]. However, it only contains single domain data and is not large enough for building a machine-learning-based NER system [Tsai et. al 2004]. We, therefore, collected over a million sentences without any annotations from the online United Daily News (UDN) and China Times for the period December 2002 to December 2003 as raw data. The sentences extracted from raw data recorded in XML format shown in Figure 1.

```
- <Sentence id="id383442" text="所以成了男女朋友">
    <TagGroup type="NameTagging" />
    <Log type="human-tagging" duration="0" />
  </Sentence>
- <Sentence id="id180560" text="中華文化重視家庭、家族關係">
    <TagGroup type="NameTagging" />
    <Log type="human-tagging" duration="0" />
  </Sentence>
- <Sentence id="id1122061" text="樂透頭獎加碼活動將於今日暫時劃下休止符">
    <TagGroup type="NameTagging" />
    <Log type="human-tagging" duration="0" />
  </Sentence>
- <Sentence id="id103554" text="甚至播放視訊檔案等多媒體用途">
    <TagGroup type="NameTagging" />
    <Log type="human-tagging" duration="0" />
  </Sentence>
```

Fig. 1. Original corpus

We chose high school students as annotators and gave them basic training before they performed the annotations. The training process was:

1. All the students attended courses about the project, including an introduction to named entity recognition, segmentation, and parts-of-speech tagging.
2. Students took a qualifying test to select participants for the tagging task.
3. Participants had to acquaint themselves with the annotating criteria we suggest ed and the operation of the tagging tool program. (Figure 2 shows the interface of the tagging tool.) The participants were then divided into three groups.

Each group was given the same sentence set containing 21,000 randomly extracted sentences; 13,208 from the UDN and 8,892 from the China Times. Table 2 shows the distribution of sentences. Participants were asked to finish the tagging task with the tagging tool program in two weeks. Figure 3 shows a tagged XML file in which the annotations are marked. Tagging results were then collected from each group and checked for consistency.

Fig. 2. Tagging tool

Table 2. Raw sentences count for each domain

大陸新聞	中時電子報	638	文化藝術	聯合新聞網	1362
生活消費	中時電子報	226	本日焦點	聯合新聞網	1019
地方新聞	中時電子報	143	生活消費	聯合新聞網	848
社會新聞	中時電子報	883	地方新聞	聯合新聞網	1662
政治新聞	中時電子報	791	重點新聞	聯合新聞網	2305
重點新聞	中時電子報	2987	旅遊休閒	聯合新聞網	608
旅遊休閒	中時電子報	318	財經產業	聯合新聞網	2303
財經產業	中時電子報	672	意見評論	聯合新聞網	519
意見評論	中時電子報	868	資訊科技	聯合新聞網	598
資訊科技	中時電子報	647	影視娛樂	聯合新聞網	1332
影視娛樂	中時電子報	719	醫療保健	聯合新聞網	652

Fig. 3. Tagged corpus. The marked area shows the annotations the participants made: “pos” means the named entity’s starting location in the sentence and “len” is the length of the NE. The label “name” indicates what kind of NE the word is.

5 Conclusion and Future work

In this paper we describe the construction of a tagged corpus for Chinese NER. We define the criteria of Chinese NE tagging, and design a standard tagging procedure for NE corpus annotation. We also demonstrate an annotator training procedure and the statistics of the corpus. The resulting corpus, CNEC 1.0, can be used to improve the performance of Mencius, our Chinese NER system. We do not use some ambiguous entities, labeled as DIFF, that involve issues such as abbreviations, cross-language loanwords and borrowings for training the NER model. As these entities need to be re-classified, advanced annotation will be executed in the next version of CNEC.

References

[1] D. Bikel, S. Miller, Richard Schwartz and Ralph Weischedel: “Nymble: a High-Performance Learning Name Finder” Proceedings of the Fifth Conference on Applied Natural Language Processing, 1997, pp. 194-201.

[2] M. Asahara and Y. Matsumoto: “Japanese Named Entity Extraction with Redundant Morphological Analysis”, HLT- North American Chapter of the Association for Computational Linguistics, Edmonton, Canada, 2003,
[3] A. Borthwick, J. Sterling, E. Agichtein, R. Grishman, “Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition”, Sixth Workshop on Very Large Corpora, 1998.
[4] S. Sekine, R. Grishman, H. Shinnou “A Decision Tree Method for Finding and Classifying Names in Japanese Texts”, Sixth Workshop on Very Large Corpora, 1998.
[5] F. Erik Sang T.K., "Introduction to the CoNLL-2001 Shared Task: Language-Independent Named Entity Recognization", Proceeding of the 6th Conference on Natural Language Learning 2002 (CoNLL 2002), pp. 155-158
[6] F. Erik Sang T.K., Meulder, F.D., "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition", Proceeding of the 7th Conference on Natural Language Learning 2003 (CoNLL 2003)
[7] R. Grishman, B. Sundheim, “Message Understanding Conference - 6: A Brief History”, Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, June 1996.
[8] N. Chinchor, "MUC-7 Named Entity Task Definition(Version 3.5)" The 7th Message Understanding Conference (MUC), 1998
[9] Tsai, T.H., Wu, S.H., Lee, C.W., Shih, C,W., Hsu, W.L.: “Mencius: A Chinese Named Entity Recognizer Using The Maximum Entropy-Based Hybrid Model” Computational Linguistics and Chinese Language Processing, Vol.9, No.1, 2004, pp.65-82.

應用機率式句法結構與隱含式語意索引於情緒語音合成之單元選取

陳俊甫 夏啟峻 吳宗憲

國立成功大學資訊工程學系
{cama, shiacj, chwu}@csie.ncku.edu.tw

摘要

在人機溝通介面中，語音逐漸扮演著重要的角色。然而，傳統電腦語音缺乏情緒特性，使得電腦與人的互動機能嚴重降低。因此，使電腦合成出帶有不同情緒特性的電腦語音是本文主要目的。在本論中，對於語料式情緒語音合成系統主要的問題，分為下列四項研究重點：1）根據不同情緒，設計一套平衡語料庫，並利用自動單元切割技術，生成基本合成單元；2）提出修正式可變長度單元機制，將機率式句法模型概念導入，決定單元長度與單元合適性；3）有別於一般聲學上的單元失真度計算，應用隱含式語意索引的概念，針對單元的語意失真度進行量度；4）最後，應用動態規劃與自動斷句預測，挑選出單元並合成情緒語音。在實驗中，首先針對中文斷句預測的正確率做比較；接著，對於語音合成的結果，觀察合成語音與實際語音在參數上的差距。並利用主觀式的評估方式，分別進行自然度 MOS 測試，情緒鑑定測試與理解度測試，本文提出之方法，在合成的自然度與情緒的表現上，皆有不錯的表現。

1. 緒論

現階段的語音合成技術已達成熟階段，但是關於帶有情緒特性的電腦語音合成技術，卻還處於起步的狀態。在語音合成方面國外最具代表性的研究機構為 AT&T[1]與微軟亞洲研究院，其相關的研究發展[2][3][4]皆有顯著的成果，微軟亞洲研究院更是投入大量的人力與物力來支持語音科技的相關研究；台灣在 80 年代開始中文語音方面的研究，如台大、清大、交大、成大、工研院電通所、交通部電信所、中研院等，都積極投入研究工作並累積了大量的研究成果。就中文而言，發音一般是以詞為基礎，音節僅能包含子母音相接連的連音變化方式，對於音韻的變化是比較不足的；另一方面，以詞為發音基礎的中文，語者情緒的變化也會呈現在詞的層級上。因此我們這套 Corpus-based 語音合成系統，是以詞（word）與音節（syllable）共同為最基本的合成單元，據此設計四種情緒的語料，以合成目標情緒語音。接著，利用兩階段語音切割機制，生成基本的合成單元。為了量測切割出來的語音片段的合適性，以其在對應的音節序列的統計模型的觀測機率，作為評量標準，來確定切割單元是否正確。由於中文語句是以詞為音韻基礎，本文提出一套修正式可變長度單元挑選機制，採用詞序列語音段為合成單元，再加上中文語音合成常用的基本單元－音節－為最小的合成單元，讓語音在串接時能有保有最原始的音韻與韻律節奏，達到情緒語音合成的目標。為決定候選詞序列單元的合適性，本文針對目前失真度定義上的不足與人類構句與發音時連音的型態，利用機率式句法結構（PCFG，Probabilistic Context Free Grammar）[5]，模擬最符合人類原始連音構句模式的單元挑選機制，並運用隱含式語意索引（LSI , Latent Semantic Indexing）量度合成單元在語意結構上的失真度。最後整合 1.）修正式可變長度單元挑選機制，2.）隱含式語意索引文法結構距離，3.）聲學參數失真度，計算出各種合成單元序列的失真度，再以動態規劃的方式，快速找出最佳的合成單元序列，合成出帶有情緒特性的電腦語音，系統流程如圖 1 所示：

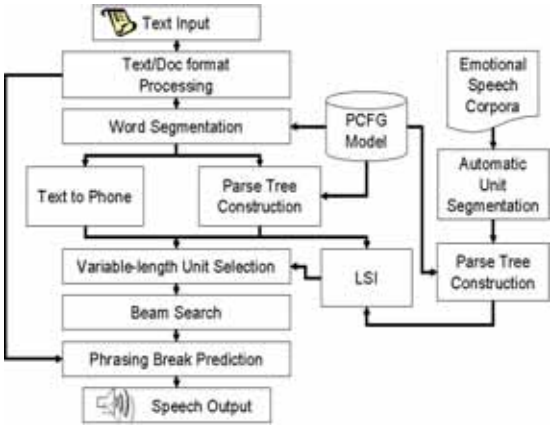


圖 1：情緒語音合成系統流程

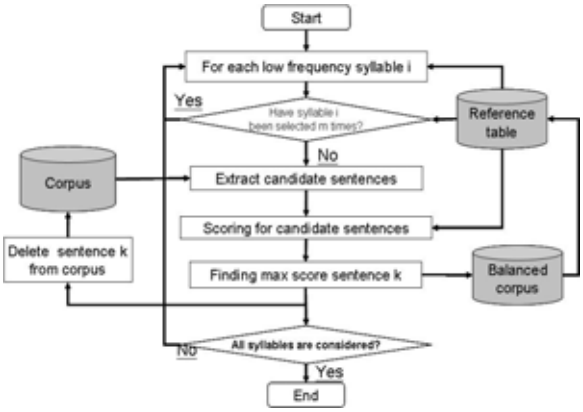


圖 2：平衡語料挑選流程圖

2. 平衡語料庫與合成單元之生成

2.1 平衡語料庫

根據研究人類情緒的文獻探討，本研究中結合 Russell[6]的 Dimensional Approach 與 Prototype Emotion Approach[7]的分類法，以四大類情緒為目標合成的語音之情緒：高興，生氣，悲傷，中性。為錄製收集語音合成系統所需之合成單元，必須先設計一套包含各發音單元及涵蓋大多數中文常用詞彙的文字平衡語料庫；從大量的文字語料（包括新聞、小說，短文等，合約兩萬字）中，根據從平衡條件所制定的計分原則加以挑選；由於用文字來表達情緒有不同程度的差別，本研究希望可以合成出明顯且強烈的電腦語音，為增加文句對情緒表達的程度[8][9][10]，用以下列準則修飾文句，使文句帶有部分情境及誘導目標情緒的特性，藉此增加情緒的表達；

(1) 插入可增強目標情緒的句子，例如：我好難過、太棒了…等；
(2) 加入可增強目標情緒的修飾辭，例如：非常、好、實在、耶…等；
(3) 保持句子的長度及音韻節奏來增強或保持語者的情緒狀態；
(4) 修改影響情緒表達的文句或是多餘的贅語。

語料的完整性與平衡特性是一套 Corpus-based 語音合成系統的基礎，本研究採用修正式可變長度單元挑選機制，所以在語料的設計上，需要包含較多的中文詞，所以本研究以常用詞為主要收集對象，根據詞頻計算每個候選文句的分數；此外，在需要收集到所有音節的考量情況下，音節的計分也被納入，據此我們定義了挑選平衡句的計分方式。計分條件（對第 j 個句子計分）

音節（Syllable）的出現頻率： $\left(\prod_{k=1}^{C\_S_j} fs_{j,k}\right)^{1/C\_S_j}$ (1)

以個別音節的頻率（Uni-gram）計算之。其中 $fs_{j,k}$ 表示第 j 個句的中的第 k 個音節的頻率，C_S_j 則表示第 j 個句子的音節個數。

詞（Word）的出現頻率： $\left(\prod_{k=1}^{C\_W_j} fw_{j,k}\right)^{1/C\_W_j}$ (2)

以每個詞的林率計算之。其中 $fw_{j,k}$ 表示第 j 個句的中的第 k 個詞的頻率，C_W_j 則表示第 j 個句子的詞數。

計分方式： $Score_j = \left(\prod_{k=1}^{C\_S_j} fs_{j,k}\right)^{1/C\_S_j} \times \left(\prod_{k=1}^{C\_W_j} fw_{j,k}\right)^{1/C\_W_j}$ (3)

結合以上兩種計分。

平衡語料挑選的流程圖如圖 2，整個架構分為兩層，外層用來保證將所有音節收入，內層迴圈用以將所有句子計分，挑選出分數最高的句子，每選出一句平衡句都必須要更新計分用的參考表，並且從原始語料將其刪除，終止條件為滿足所需的音節出現次數。

2.2 合成單元之生成

在本研究中，提出一套自動的語音單元切割與確認方法，來加快合成單元的標記作業以及提供驗證的方式。但是由於自動找邊界的結果，並非都是令人滿意的[11]，因此需要做一些調整，才能讓單元切割達到不錯的結果。因此，我們利用兩階段語音切割模組，第一階段利用隱藏式馬可夫模型進行初步預測斷點，第二階段利用語音參數的特性，根據觀察與測試利用規則將斷點調整在正確的位置上。最後，利用機率模型做比對，確認單元正確性。

1.隱藏式馬可夫模型：利用強迫路徑之隱藏式馬可夫模型進行語音切割，在參數的設定上，我們使用 26 維的參數，包括 12 階 MFCCs、12 階的△MFCCs、能量差（delta energy）以及能量差之差（delta delta energy）值，而在模型個數上，總共包括 150 個次音節模型。對於先前所錄製的聲音語料，根據其文字內容，利用已知音節型態（syllable type）所對應的隱藏式馬可夫模型，使用 Viterbi 演算法，搜尋每個對應模型狀態外轉的位置，據此找出每個音節的邊界位置。

2.斷點調整：經由觀察的結果，發現邊界切割錯誤主要有幾種：第一，子音開頭。可能受到前一個音節的母音結尾連音的影響，或是子音前的靜音部分的干擾，可能造成斷點錯誤。第二，母音結尾。母音結尾可能因為在頻譜上的落差太大，造成外轉提前發生，導致母音還未結束時就被切割下來。因此，本研究利用能量（energy）以及過零率（zero crossing rate）當作調整的觀測參數[11]，對於每一種不同的音節型態，設定多個規則來微調。

3.單元確認：根據前一步驟所述可將連續語音切成最基本的合成單元，然而並非每個單元在錄置的過程中，都是正確無誤的，本節的目的是要測試各語音切割結果跟相對應的音節型態是否一致，也就是測試此語音單元在對應的音節模型內的機率高低，如果機率高，表示此單元的結果是較為正確的，反之，則單元較不正確。

$$\begin{cases} P(X \mid \lambda_i) < threshold_i & \rightarrow \text{reject} \\ Otherwise & \rightarrow \text{accept} \end{cases} \tag{4}$$

其中，λ_i 代表相對於觀測資料 X 的音節模型，$threshold_i$ 代表此音節型態所對應的臨界值。

3. 修正式可變長度單元合成機制

3.1 可變長度單元合成機制

從一個大量的語料庫中挑選出合適的合成單元已經被證明確實有助於提升合成系統的品質[4][12]，而單元的型態包括音素（Phoneme）、雙音（Diphone）、半音節(Demi-Syllable)、音節(Syllable)、不定長度的單元(Non-Uniform Unit)等。就中文而言，如果能找到較長詞來當合成單元，當然是一個比較好的選擇，因為這樣的合成單元內，已經包含了本身的音韻，因此在串接的自然度上有一定的效果提升。過去，可變長度單元的挑選機制主要是以詞為基礎。對於每一個可能出現的詞或是音節，去搜尋所有可能的組合方式，找出一組最佳的詞序列。例如：

中 國 人 是 聰 明 的 民 族

就這個句子而言，他所可能衍生出來的可能組合性有很多：

(1)中國人 是 聰明 的 民族　　(2)中國人 是 聰明 ***的民族***

(3)中國人 是 聰明的民族　　(4)中國人 是 聰明的 民族

(5)中國人 ***是聰明*** 的民族　　(N)……

但是，其中有許多的組合是不符合中文音韻的組合，例如「***的民族***」「***是聰明***」，而且若要搜尋所有可能的組合，所要耗費的時間跟空間複雜度太龐大。因此我們提出了一套新的可變單元長度挑選機制，主要考慮兩個觀點。第一，模擬人類構句的方式，根據中文發音的音韻與斷句，我們可以找到合適的合成單元。由於人類構句的方式，是先將單音節（syllable）組合成詞（word），再將多個詞組合成長詞或專有名詞，進一步組合成片語、句子，如圖 3 所示。

因此，我們可以根據這樣的想法，將不適合的組合性去除，並可根據不同階層上的組合方式，進行階層式的單元挑選。第二，除了聲學上的失真度之外，語意結構上的失真度也該被考量。根據中文語音學的觀點，同一個詞或是同一個音節，在不同的語句結構中，它們在聲學參數上的表現會不一樣，舉例來說：

(A). 例一：　漂亮的雙殺，化解了滿壘的危機　(44.3ms)

　　墾丁的風景還是一樣漂亮　(65.3ms)

這個例子中，同樣的詞，位在不同的詞性的長詞中，明顯的，兩個詞的音長是不同的。

(B). 例二：　在院子裡栽種了好多鮮豔的花　(39.1ms)

　　我的手藝真是越來越高超，花招也變多了　(28ms)

這個例子中，「花」作為不同詞性之用，在音高的變化上也有不同的結果。根據這兩個想法，本研究提出一個修正式可變長度單元挑選機制，利用機率式句法結構轉譯器（probabilistic syntactic parser），將中文句轉換成一個階層式樹狀語意結構，這棵樹上的每一個終端節點，代表的是一個詞。而每一個非終端節點，代表了一種可能的長詞組合，如圖 4 所示。這樣的作法有以下幾種優點：1.)可移除不適當的長詞組合；2.)利用樹狀結構，挑選出適合的合成單元；3.)可根據語意結構，量測單元間的語意失真度。

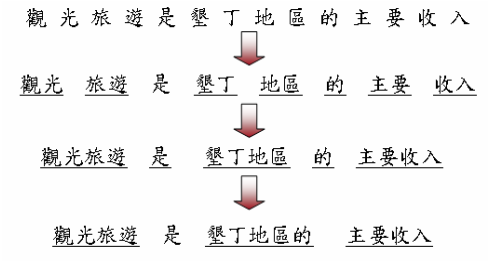


圖 3：人類構句過程模擬示意圖

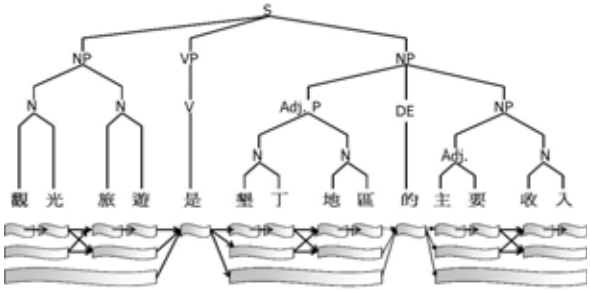


圖 4：中文文法樹範例

3.2 中文文法機率模型

首先，我們需要一個語意剖析器來處理中文文句，並建立對應的語意樹狀結構，本研究利用機率式句法結構（PCFG , Probabilistic Context Free Grammar）來對中文句進行剖析[5]。所謂的機率式句法結構是由句法結構（CFG, Context Free Grammar）衍生而來，以機率的觀點來看語言模型，更可藉由

賦予句法結構 CFG 的規則機率，使得機率式句法結構能夠更正確的模擬口述語言，使語意混淆度降低。機率式句法結構的觀念，與語音辨識中隱藏式馬可夫模型的概念類似，同樣的是想要找出若給定一個文法G，從起始符號N_0開始，產生一串詞序列$W_{1,T}=w_1,w_2\cdots w_T$的機率值：

$$P\left(S\overset{*}{\Rightarrow}W_{1,T}\,|\,G\right) \tag{5}$$

其中，箭號⇒表示衍生的意思，而箭號上方的星號*則表示所有衍生的路徑。這項機率值是由所有合法的衍生規則組合而成，每條規則的機率則是預先由訓練語料中估算求得。假設有一條規則是$A\to\alpha$，則此規則的機率求法為：

$$P\left(A\to\alpha_j\,|\,G\right)=\left[C\left(A\to\alpha_j\right)\right]\Big/\left[\sum_{i=1}^{m}C\left(A\to\alpha_i\right)\right] \tag{6}$$

其中，$C(\bullet)$代表的是每條規則出現的次數，m表示α_i的所有可能性，或所有由A衍生出來的規則個數。本研究採用中研院詞庫小組所定義的 Tree-Bank 文法規則以及相對應的機率為 PCFG 模組的原始模型，擷取部分作為本研究中的文法規則。在此我們導入 Chomsky Normal Form，目的是簡化說明 PCFG 模組以及本研究提出的文法結構距離量測。假設每個非終端項只能分為兩個非終端項的組合$N_i\to N_j+N_k$或是一個終端項（terminal term）$N_i\to w_l$，且其所有可能性的機率和為 1：

$$\sum_{j,k}P\left(N_i\to N_jN_k\,|\,G\right)+\sum_{l}P\left(N_i\to w_l\,|\,G\right)=1 \tag{7}$$

根據這套文法規則G，如圖 5，從起始符號N_0開始，推衍產生一串詞序列$W_{1,T}=w_1,w_2\cdots w_T$的機率值為：

$$P\left(N_0\overset{*}{\Rightarrow}w_1w_2\cdots w_T\,|\,G\right)=\sum_{i}\left(P\left(N_i\overset{*}{\Rightarrow}W_{m,n}\,|\,G\right)P\left(N_0\overset{*}{\Rightarrow}W_{1,m-1}N_iW_{n+1,T}\,|\,G\right)\right) \tag{8}$$

式(8)中，$P\left(N_i\overset{*}{\Rightarrow}W_{m,n}\,|\,G\right)$我們稱之為內部機率（Inside Probability），代表的是一個非終端項N_i被推成詞序列$W_{m,n}=w_m\cdots w_n$的機率值，我們將此機率值表示為$\beta_i\left(m,n\,|\,G\right)$。根據 Chomsky Normal Form 的表示式，一個非終端項只能被分為兩個非終端項的組合，以遞迴的寫法表示成：

$$\begin{aligned}P\left(N_i\overset{*}{\Rightarrow}W_{m,n}\,|\,G\right)&=\beta_i\left(m,n\,|\,G\right)=\sum_{j,k}\sum_{d=m}^{n-1}P\left(N_i\to N_jN_k\,|\,G\right)P\left(N_j\overset{*}{\Rightarrow}W_{m,d}\,|\,G\right)P\left(N_k\overset{*}{\Rightarrow}W_{d+1,n}\,|\,G\right)\\&=\sum_{j,k}\sum_{d=m}^{n-1}P\left(N_i\to N_jN_k\,|\,G\right)\beta_j\left(m,d\,|\,G\right)\beta_k\left(d+1,n\,|\,G\right)\end{aligned} \tag{9}$$

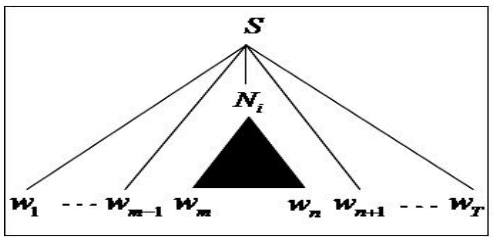


圖 5：機率式句法結構示意圖

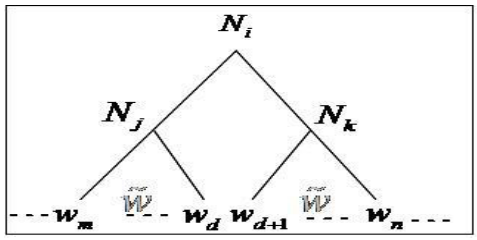


圖 6：單元內部機率

由於本研究只需要在建立樹狀結構的過程中，取分數最高的一棵樹，因此我們將式(9)改寫，在所有可以建出一棵樹狀結構的可能中，挑選出分數最高的當作輸出的機率值，如下表示：

$$\begin{aligned}\hat{\beta}_i\left(m,n\,|\,G\right)&=P\left(N_i\overset{\max}{\Rightarrow}W_{m,n}\,|\,G\right)=\max_{\substack{j,k\\ m\le d<n}}\left(P\left(N_i\to N_jN_k\,|\,G\right)\times P\left(N_j\overset{\max}{\Rightarrow}W_{m,d}\,|\,G\right)P\left(N_k\overset{\max}{\Rightarrow}W_{d+1,n}\,|\,G\right)\right)\\&=\max_{\substack{j,k\\ m\le d<n}}\left(P\left(N_i\to N_jN_k\,|\,G\right)\hat{\beta}_j\left(m,d\,|\,G\right)\hat{\beta}_k\left(d+1,n\,|\,G\right)\right)\end{aligned} \tag{10}$$

式(8)中的$P\left(N_0\overset{*}{\Rightarrow}W_{1,m-1}N_jW_{n+1,T}\,|\,G\right)$，我們稱為外部機率（Outside Probability），代表的是由起始符號N_0推出詞序列$W_{1,m-1}=w_1\cdots w_{m-1}$與$W_{n+1,T}=w_{n+1}\cdots w_T$，且兩詞序列中夾著$N_j$的機率值，同樣的，把外部機率表示成$\alpha_j\left(m,n\,|\,G\right)$。由於非終端項$N_j$可能位於上一層非終端項$N_i$推導出的規則中的左項或右項。因此，將式子寫為所有可能的規則與詞斷點的機率和。

$$
\begin{aligned}
&P\left(N_0 \overset{*}{\Rightarrow} W_{1,m-1}N_jW_{n+1,T} \mid G\right) = \alpha_j\left(m,n \mid G\right) \\
&= \sum_{i,k}\left(\begin{array}{l}\sum_{d=n+1}^{T_q}\left(P\left(N_i \to N_jN_k \mid G\right)\times P\left(N_0 \overset{*}{\Rightarrow} W_{1,m-1}N_jW_{d+1,T} \mid G\right)P\left(N_k \overset{*}{\Rightarrow} W_{n+1,d}\right)\right) \\ +\sum_{d=1}^{m-1}\left(P\left(N_i \to N_kN_j \mid G\right)\times P\left(N_k \overset{*}{\Rightarrow} W_{d,m-1}\right)P\left(N_0 \overset{*}{\Rightarrow} W_{1,d-1}N_jW_{n+1,T} \mid G\right)\right)\end{array}\right) \\
&= \sum_{i,k}\left(\begin{array}{l}\sum_{d=n+1}^{T_q}\left(P\left(N_i \to N_jN_k \mid G\right)\alpha_i\left(m,d \mid G\right)\beta_k\left(n+1,d \mid G\right)\right) \\ +\sum_{d=1}^{m-1}\left(P\left(N_i \to N_kN_j \mid G\right)\beta_k\left(d,m-1 \mid G\right)\alpha_i\left(d,n \mid G\right)\right)\end{array}\right)
\end{aligned}
\tag{11}
$$

同樣的，由於我們只需要求取最高分樹的那棵樹狀結構，因此我們將式(11)改寫為：

$$
\begin{aligned}
&\hat{\alpha}_j\left(m,n \mid G\right) = P\left(N_0 \overset{\max}{\Rightarrow} W_{1,m-1}N_jW_{n+1,T} \mid G\right) \\
&= \max_{j,k}\left(\begin{array}{l}\max_{n+1\le d\le T_q}\left(P\left(N_i \to N_jN_k \mid G\right)\hat{\alpha}_i\left(m,d \mid G\right)\hat{\beta}_k\left(n+1,d \mid G\right)\right), \\ \max_{1\le d\le m-1}\left(P\left(N_i \to N_kN_j \mid G\right)\hat{\beta}_k\left(d,m-1 \mid G\right)\hat{\alpha}_i\left(d,n \mid G\right)\right)\end{array}\right)
\end{aligned}
\tag{12}
$$

由於本研究採用不固定長度的單元挑選機制，系統選用的候選合成單元不是音節而是詞序列，所以對於內部機率的剖析，須考慮所要的合成單元，此單元在剖西的過程中，不能再被切割。因此，我們需要求出一個由非終端項 N_i 推導出詞序列 $W_{m,n} = w_m \cdots w_n$ 且包含詞序列（合成單元）$\tilde{w}$ 的共同機率值，因此我們必須求得 $P\left(N_i \overset{*}{\Rightarrow} W_{m,n}, \tilde{w} \mid G\right)$，如圖六所示。

$$
\begin{aligned}
&P\left(N_i \overset{*}{\Rightarrow} W_{m,n}, \tilde{w} \mid G\right) = \gamma_i\left(m,n,\tilde{w} \mid G\right) \\
&= \sum_{j,k}\left(P\left(N_i \to N_jN_k \mid G\right)\times \sum_{d=m}^{n-1}\left(\begin{array}{l}\gamma_j\left(m,d,\tilde{w} \mid G\right)\beta_k\left(d+1,n \mid G\right)\delta\left(m,d,\tilde{w}\right) \\ +\beta_j\left(m,d \mid G\right)\gamma_k\left(d+1,n,\tilde{w} \mid G\right)\delta\left(d+1,n,\tilde{w}\right)\end{array}\right)\right)
\end{aligned}
\tag{13}
$$

$$
\delta\left(m,n,\tilde{w}\right) = \begin{cases}1, & \text{if } \tilde{w} \text{ is a substring of } W_{m,n} \\ 0, & \text{otherwise}\end{cases}
\tag{14}
$$

同樣的，由於我們只需要求取最高分樹的那棵樹狀結構，因此我們將之改寫為：

$$
\begin{aligned}
&\hat{\gamma}_i\left(m,n,\tilde{w} \mid G\right) = P\left(N_i \overset{\max}{\Rightarrow} W_{m,n}, \tilde{w} \mid G\right) \\
&= \max_{\substack{j,k \\ m\le d<n}}\left(\begin{array}{l}P\left(N_i \to N_jN_k \mid G\right)\hat{\gamma}_j\left(m,d,\tilde{w} \mid G\right)\hat{\beta}_k\left(d+1,n \mid G\right)\delta\left(m,d,\tilde{w}\right), \\ P\left(N_i \to N_jN_k \mid G\right)\hat{\beta}_j\left(m,d \mid G\right)\hat{\gamma}_k\left(d+1,n,\tilde{w} \mid G\right)\delta\left(d+1,n,\tilde{w}\right)\end{array}\right)
\end{aligned}
\tag{15}
$$

4. 文法結構距離與情緒語音合成

4.1 文法結構距離

在前一節提到，同樣的單元在不同的語意結構上，會有不同的表現，因此本研究設計了一套測量文法結構距離的方法，主要是根據機率式文法結構所產生出的語法樹，藉由隱含式語意索引，計算單元在不同語意結構上的差距。

4.1.1 文法結構樹向量化

由於每個句子可以由一棵文法結構樹來表示，而一棵樹只會由少數幾個規則所構成，會有稀疏資料（sparse data）的問題。因此，為了解決這個問題，並且求出單元在不同文法結構樹上的關係，因此我們採用資訊檢索中向量空間比對（Vector Space Model）的方法。將樹結構的比對，視為向量的比對。

將所有的文字語料轉換成規則向量，儲存在一個維度為 $R\times Q$ 的文法結構資訊矩陣 $\mathbf{\Phi}_{R,Q}$。其中 R 代

表整個 PCFG 模型G中文法規則的個數，Q代表語料庫中句子的個數。

$$\mathbf{\Phi}_{R\times Q}=\begin{bmatrix}\phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,Q}\\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,Q}\\ \vdots & \vdots & \ddots & \vdots\\ \phi_{R,1} & \phi_{R,2} & \cdots & \phi_{R,Q}\end{bmatrix} \tag{16}$$

矩陣中每個元素$\phi_{r,q}$代表著第r條規則在第q個句子S_q中所佔的重要性。因此本研究中定義$\phi_{r,q}$的估計法如下：

$$\phi_{r,q}=(1-\varepsilon_r)P\left(Rule\ r:N_i\rightarrow N_jN_k,W_{1,T},\tilde{w}\,|\,G\right) \tag{17}$$

其中，等號右側第二項代表的是該條規則佔該句語法結構的比重，該項可以寫為：

$$P\left(Rule\ r:N_i\rightarrow N_jN_k,W_{1,T},\tilde{w}\,|\,G\right)=C\left(N_i\rightarrow N_jN_k,W_{1,T},\tilde{w}\right)\Big/\sum_{a,b,c}C\left(N_a\rightarrow N_bN_c,W_{1,T},\tilde{w}\right) \tag{18}$$

而第一項是用來度量該條規則在語料中的鑑別性是否足夠，當作矩陣中該元素的權重，利用量度文字亂度（Entropy）的方法，量度某條規則在該語料中是否具有鑑別性：

$$\varepsilon_r=-\frac{1}{\log Q}\sum_{q=1}^{Q}\left(\frac{C\left(N_i\rightarrow N_jN_k,W_{1,T_q}^{(q)}\right)}{\sum_{a=1}^{Q}C\left(N_i\rightarrow N_jN_k,W_{1,T_a}^{(a)}\right)}\log\frac{C\left(N_i\rightarrow N_jN_k,W_{1,T_q}^{(q)}\right)}{\sum_{a=1}^{Q}C\left(N_i\rightarrow N_jN_k,W_{1,T_a}^{(a)}\right)}\right) \tag{19}$$

其中$W_{1,T_q}^{(q)}=w_1^{(q)}\cdots w_{T_q}^{(q)}$表示語料庫中第$q$個句子，$T_q$表示該句的長度，而$C\left(N_i\rightarrow N_jN_k,W_{1,T_q}^{(q)}\right)$則表示文法規則$N_i\rightarrow N_jN_k$出現在第$q$個句子的次數。

4.1.2 中文文法結構距離

由於語意樹結構矩陣十分的龐大，在計算上也非常耗時，本研究導入資訊檢索上的隱含式語意索引技術（LSI, Latent Semantic Indexing），不僅可以找出規則間的隱含關係，更可達至大幅降低向量維度的目標，隱含式語意索引是由奇異值分解後，由奇異值矩陣上決定要保留的變異比例，藉此決定所需的維度，再將所有的向量透過轉換矩陣，投射到教低維度且較有鑑別能力的空間上，且可以有效保留住規則與語意樹的關係。數值運算如下所示，本研究中所則保留 98%的變異量：

$$\mathbf{\Phi}_{R\times Q}=\begin{bmatrix}\phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,Q}\\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,Q}\\ \vdots & \vdots & \ddots & \vdots\\ \phi_{R,1} & \phi_{R,2} & \cdots & \phi_{R,Q}\end{bmatrix}=\mathbf{T}_{R\times n}\mathbf{S}_{n\times n}\left(\mathbf{D}_{Q\times n}\right)^{\mathrm{T}} \tag{20}$$

where $n=\min(R,Q)$

$$\tilde{\mathbf{\Phi}}_{R\times Q}=\mathbf{T}_{R\times d}\mathbf{S}_{d\times d}\left(\mathbf{D}_{Q\times d}\right)^{\mathrm{T}}\ \text{where}\ d<n,\ d=\min_k\left(\sum_{i=1}^{k}\lambda_i\Big/\sum_{i=1}^{n}\lambda_i\right)>98\% \tag{21}$$

經過奇異值分解後，我們可以利用$\mathbf{T}_{R\times d}$矩陣，將兩個句子的文法結構向量投射到較低維度的向量空間做比對，假設要合成的目標語句是$\mathbf{x}$，而包含的所需的合成單元$\tilde{w}$的候選語句為$\mathbf{y}$，則利用上述方法，定義文法結構距離：

$$SyntacticCost(\mathbf{x}^{(\tilde{w})},\mathbf{y}_q^{(\tilde{w})})=-\log\left(\hat{\gamma}_0\left(1,T_q,q,\tilde{w}\,|\,G\right)\times\frac{\left(\left(\mathbf{T}_{R\times d}\right)^{\mathrm{T}}\times\mathbf{x}^{(\tilde{w})}\right)\bullet\left(\left(\mathbf{T}_{R\times d}\right)^{\mathrm{T}}\times\mathbf{y}_q^{(\tilde{w})}\right)}{\left\|\left(\mathbf{T}_{R\times d}\right)^{\mathrm{T}}\times\mathbf{x}^{(\tilde{w})}\right\|\times\left\|\left(\mathbf{T}_{R\times d}\right)^{\mathrm{T}}\times\mathbf{y}_q^{(\tilde{w})}\right\|}\right) \tag{22}$$

4.2 情緒語音合成

根據前幾張的介紹，本研究提出一套基於語意結構的可變長度單元挑選機制，決定了合成單元的選擇基準，而且考慮單元在不同語意結構上的關係，定義出語意失真度，本節將定義其他聲學上的失真度，並利用中文斷句預測，在長句中，加入自然且合理的停頓，以符合中文發音該有的韻律跟節奏。

4.2.1 聲學失真度

（一） 頻譜斜度

藉由量測連續兩個合成單元間在各頻譜間的不連續性，計算音節間失真度。首先，將合成單元語音在串接點的前後三個框架做 256 點的 FFT（Fast Fourier Transform）轉換，轉成各頻譜的能量。接著，將轉換出來的頻譜範圍，分成 k 個頻帶。針對各個不同的頻帶，利用線性迴歸（Linear Regression）的

方法，在連續的三個框架間，求出一條迴歸曲線；最後，量測連續兩個合成單元在每個頻帶上，其迴歸曲線的斜率差值。

$$SD(u_n, u_{n+1}) = \sum_{i=1}^{k} w(i)[\Delta u_{n+1}(i) - \Delta u_n(i)]^2 \tag{23}$$

（二） 音高與能量

連續的兩個合成單元，利用 Autocorrelation 的方法，求取其平均基週，藉由量測此兩單元間的基週差異，訂定一音高失真度的量測。同樣的，利用計算兩單元的漢明能量，訂定一能量失真度。在這兩個失真度之間，取一個權重，取其總和，定義為音韻失真度。

$$PD(u_n, u_{n+1}) = w_{Fo} C_{F0}(u_n, u_{n+1}) + w_{ene} C_{ene}(u_n, u_{n+1}) \tag{24}$$

4.2.2 整句元網格最佳路徑搜尋

根據上述的失真度定義，我們可以將本研究中的失真度[13]，分為以下兩種：

音節失真度：利用單元與單元在不同的語意結構造成其發音語音韻的不同，並利用機率式句法結構與隱含式語意索引，定義出語意失真度，作為音節失真度。

$$C_C = w_{SD} SD(u_n, u_{n+1}) + w_{PD} PD(u_n, u_{n+1}) \tag{25}$$

音節間失真度： 利用語音在聲學上連續的特性，分別利用頻譜斜率、音高與能量，量測連續兩個合成單元在這些參數上的差異，定義為音節間的失真度。

$$C_S = SyntacticCost(\mathbf{x}^{(\tilde{w})}, \mathbf{y}_q^{(\tilde{w})}) \tag{26}$$

其中 $\tilde{w}$ 定義為單元 u_n 相對應的中文描述。根據這兩個失真度的定義，我們便可得到下面的式子：

$$\hat{u}_{1:N} = \arg\min_{u_{1:N}} \left(C_S(u_0, u_0') + C_C(u_0, u_1) + C_S(u_1, u_1') + C_C(u_1, u_2) + \cdots + C_C(u_{N-1}, u_N) + C_S(u_N, u_N') \right) \tag{27}$$

因此，我們利用動態規劃演算法，在一連處的候選單元序列中，求得一個失真度總和最小的合成單元序列。但是，由於當語料大或是句子太長，會導致搜尋的空間過大，使的時間複雜度太高，因此我們利用 Beam Search，來限定路徑，減少搜尋時間。

4.3 中文音韻詞組預測

對於中文發音中韻律與節奏的產生，停頓佔了扮演了一個很重要的角色，他不只可以避免語意上的扭曲（semantic ambiguity），更可以增加中文句音韻的效果。由於中文是一種單音節詞的語言，通常在一個單詞的音節間，不會有停頓的產生，因此如何在這些詞與詞之間，找出停頓的位置與長度，便是這節的重點。首先，我們可以將音韻詞組預測的預測，視為一種自動學習的問題，也就是說，如何從小量的資料訓練中，找出預測音韻詞組的位置跟長度的。本研究利用分類與回歸樹（CART，Classification and Regression Tree）的方法來達成。我們設計一套問題集，包含了關於詞性，詞長，以及詞性對的的相關問題，利用根據一小量的訓練資料，自動訓練出一棵決策樹，其中，每個非終端節點代表的是一個問題，而每個葉節點才代表著一種停頓類型，如圖 7 所示。本研究中，我們間停頓的種類分為三類：A.)沒有停頓（No break）：在詞與詞間沒有停頓、B.)次停頓（Minor break）：詞與詞之間有一個小的無聲區、C.)主要停頓（Major break）：詞與詞之間有一長停頓。當一個測試資料進來，根據這棵決策樹，我們就可以判斷在兩個詞之間，是否有停頓的產生以及停頓是屬於何種停頓。

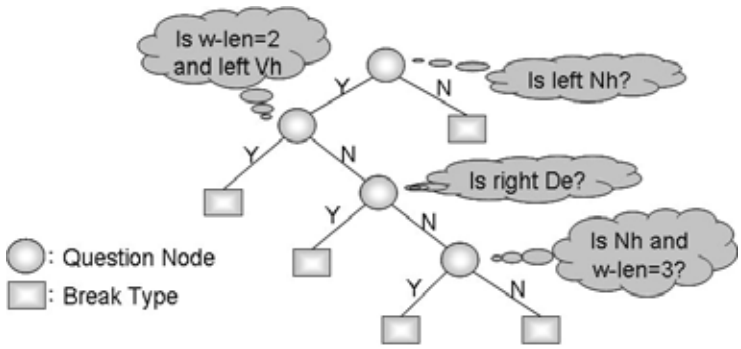


圖 7：中文音韻詞組預測 CART 示意圖

5.實驗結果與討論

5.1 中文音韻詞組預測

本研究中，人工標記了 400 句中文句，其中三百句當作訓練語料，一百句為測試語料。在一百句的測試語料中，總共有 574 個音韻斷點，每種斷點的預測結果如下表所示。由表 1 可以發現，在大部分的音韻斷點中，多數是屬於不在詞之間插入停頓，而主要停頓的次數也是較少的。而利用本方法所得到的正確率，平均都可達到 80%以上。

表 1：中文斷句預測結果

	No break	Minor break	Major break
No break	243	12	6
Minor break	24	156	8
Major break	18	15	65

5.2 合成語音參數比照

為了觀察合成語音與原始語音在平均基週、平均能量以及音節長度之差別，依據本研究中所訂定的四種情緒，錄製語料外的情緒語句，並利用本研究之合成系統，合出相同中文文句之語音，相互比較，圖 8、圖 9 分別是悲傷與生氣對照圖。我們可以發現，在基週、能量、音長上，兩條曲線在大部分的音節是相近的。但可以發現，音長的曲線上，有某些音節例如生氣句的第 13、14 個音節"可以"，高興句的第 17、18 音節"終於"，其音長與能量原始語音有差距，主要是因為沒找到對應的長詞，因此都以單音節的方式合成，所以會造上差異。

5.3 主觀式評估與聽覺實驗

（一）自然度評估（Naturalness Evaluation Test）：本研究採用平均鑑定分數（Mean Opinion Scores , MOS）作為評估之標準，這種評估方式將合成語音輸出的自然度與情緒表達度分為優良（Excellent），良好（Good），尚可（Fair），差（Poor），極差（Unsatisfactory）五個等級，分別給予 5 至 1 不等的分數。測試人員在聽過合成的語音後，以所感覺到的自然度與情緒表現度評分。測試是由合成系統根據基本合成單元長度與語意失真度的使用與否，合成同樣的中文句，做對照實驗。對於每種情緒，合成十個句子，選擇十位大學及研究生（8 為男性，2 位女性），聆聽並根據自己所感受的語音自然度打分數，最後取一個平均。實驗中，比較三套系統（A）、（B）、（C），在合成語音自然度上的差異。（A）系統是利用單一音節為合成單元之合成系統、（B）系統為可變單元長度，但沒有加入語意失真度、（C）系統為本研究所提之系統。由表 2 結果可瞭解，利用本研究所提出的方法，進行單元的挑選，在自然度的表現上，相較於利用單音節的方式，所合成的語音，有相當大改進，在挑選過失真度上，若加入語意失真度，會使的挑選出的語句，在中文音韻上，更符合目標句所要表達的。

（二）可理解度評估（Intelligibility Evaluation Test）：本實驗的目的，是希望探討利用本實驗提出的方法所合成的語音，在可理解度上，是否達到實用的階段，並做相關比較。實驗部分，要求受測者，將所聽到的中文結果，以聽寫的方式寫出來，計算與原始文字的異同，計算其聽寫正確率。同樣的，用前一節所提到的（A）、（B）及本研究中所實作之系統，分別進行實驗。對於每個系統，四種情緒各產生十個句子，讓受測者聽寫。每個受測者平均聽寫了 1632 個音節。由圖 10 可以看出，雖然三套系統，平均都有不錯的理解度：（A）83%，（B）89.5%，（C）96.5%，但是本系統之方法，仍較一般可變單元長度之方法高。這結果顯示，本系統在可理解度以及實用性上是足夠的。

（三）情緒鑑定評估（Emotion Identification Test）：本實驗利用本研究提出之系統，針對四種情緒，各合成十個語音範例，隨機播放，讓受測者決定聽到的語音是何種情緒，由此判定系統在情緒的表現的程度。實驗由受測者，分別聆聽生氣、快樂、悲傷、中性語音，但是先不告知所聽的合成語音是何種情緒，讓受測者依據自己的聽覺，判斷並記錄，表 3 是情緒句子範例。圖 11 顯示情緒鑑定評估在各情緒的正確率：高興 83%，中性 70%，悲傷 93%，生氣 92%。由圖可以看出，中性與高興情緒被誤判的機率明顯的比生氣與悲傷高，主要是因為錄音員在錄製後兩組語料時，在情緒表達上較為強烈，因此被誤判的機率，相對降低，除此之外，也由於本研究所提出的方法，在自然度與可理解度上的提高，因此情緒鑑定的結果也較佳。

表 2：自然度實驗結果

項目	自然度		
系統	(A)	(B)	(C)
快樂	3.2	3.5	4.1
中性	2.7	3.25	3.6
悲傷	3.01	3.2	3.85
生氣	2.85	3.15	3.7

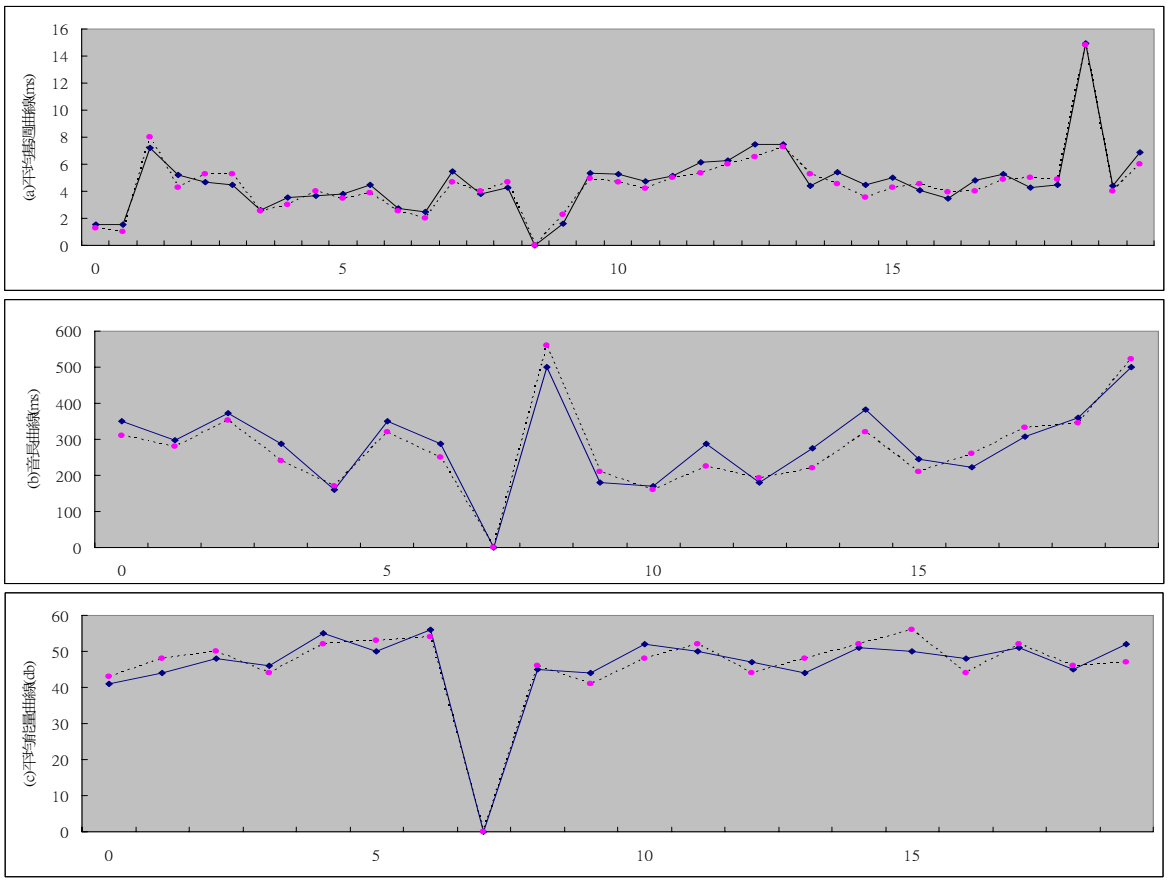


圖 8 快樂對照語句：我今天真的好高興，因為我的著作終於問世了。

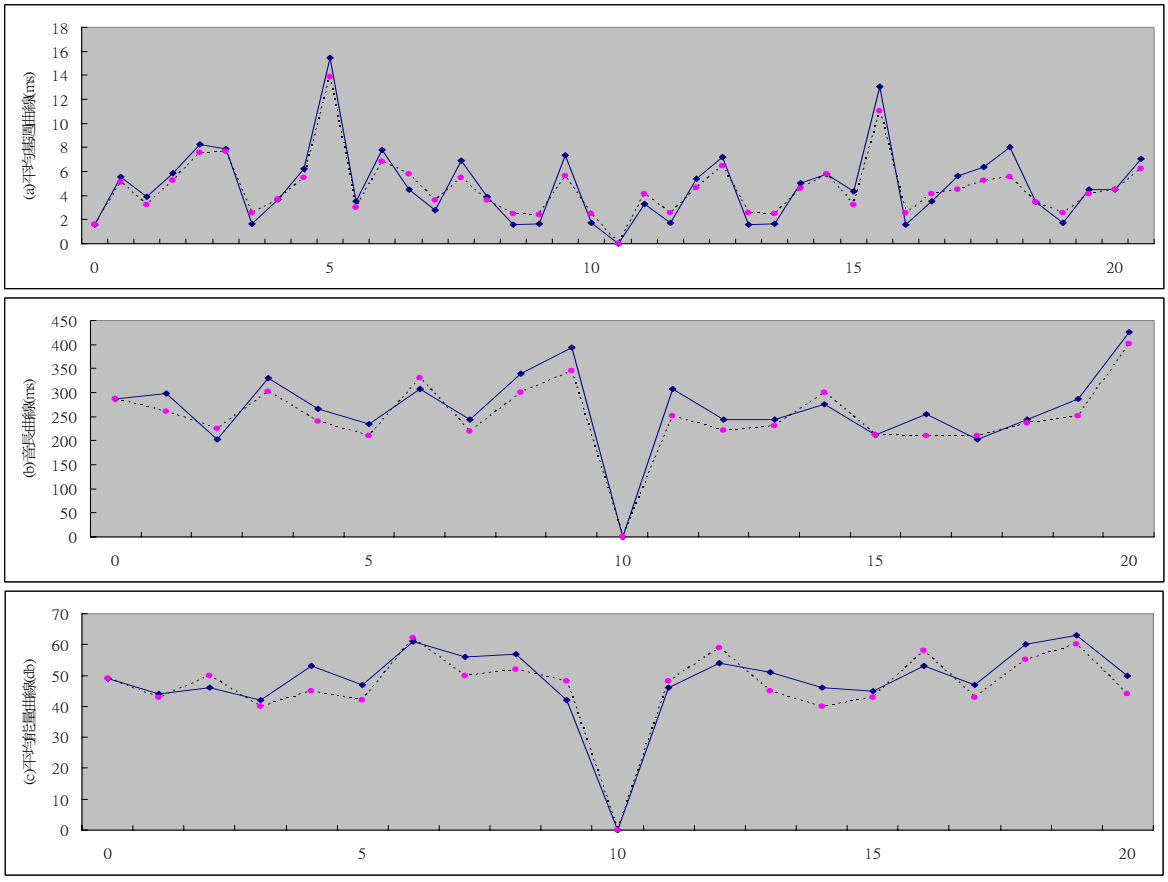


圖 9 生氣對照語句：這次旅行業者太過分了，居然可以不顧旅客安全。

表 3：情緒鑑定之例句

編號	情緒範例句
001	我今天真的好高興，我的書有不錯的銷售成績。
002	因為他的過世，我現在跟家屬一起哭泣擁抱。
003	這次的期末考再考不好，妳就完蛋了。
004	政府為了打擊犯罪，成立了聯合執行小組，顯示對於犯罪打擊不遺餘力。

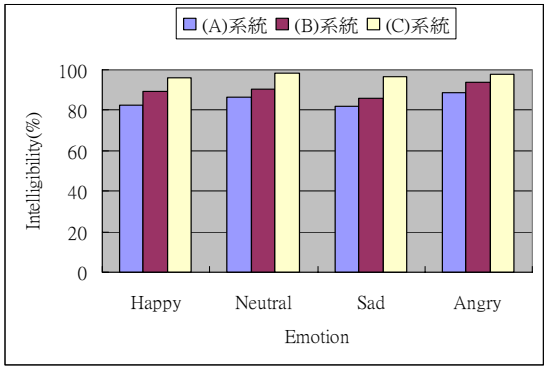


圖 10：理解度實驗結果直方圖

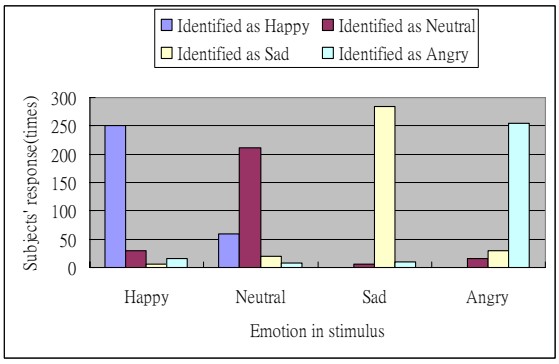


圖 11：情緒鑑定評估結果直方圖

6. 結論與未來展望

本論文中，提出了一個新的情緒語音合成系統的架構，此架構中，利用一套平衡語料挑選機制，設計並產生一套最小但包含足夠音節及常用詞資訊之情緒語料。除此之外，單元挑選方法中，不需要使用音韻模型數字化的去預測句子的音韻參數，相反的，將人類發音、構句的特性，與中文語音音韻、重音等現象，利用機率式句法結構將句子解構成一個樹狀結構，並利用樹狀結構上階層式的關係，選擇適當、合理的合成單元，在這個過程中，保留了原始語音的音韻特性。進一步，運用隱含式語意索引，計算出合成單元之間的語意失真度，挑選最適合的單元。

透過實驗評估，合成語音品質有不錯的表現，但仍有下列問題有待改進：1.)在語音的切割上，情緒語音的切割結果，相較於中性語音，有較差的斷點位置，主要是因為我們利用隱藏式馬可夫模型進行斷點切割時，並未考慮情緒因素。2.)語料式合成系統，若能收集足夠的語料，期能有較好的合成表現。3.) 在機率式句法結構模型中，會出現新詞問題（OOV）與文法規則不足（OOR）的問題，需要提出一套自動修正的方法，才能避免類似的問題。

參考文獻

[1] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, “The AT&T Next-Generation TTS System,” in *Proc. of ICSLP’98*, Sydney, Australia, pp. 931-934,1998

[2] Jon Rong Wei Yi, *Corpus-Based Unit Selection for Natural-Sounding Speech Synthesis*, Ph.D. thesis, Massachusetts Institute of Technology, 2003

[3] T. Dutoit, *Text, Speech and Language Technology. vol.3: An Introduction to Text-to-Speech Synthesis.*, Kluwer Academic Publishers, Dordrecht, 1997

[4] W. J. Wang, W. N. Campbell, N. Iwahashi and Y. Sagisaka, “Tree-based Unit Selection for English Speech Synthesis,” in *Proc. of ICASSP’93*, Minneapolis, MN, vol.2, pp. 191-194, Apr. 1993

[5] X. Huang, A. Acero and H. W. Hon, *Spoken Language Processing*, pp. 133-190, Prentice Hall, 2001

[6] J.A. Russell, “Measures of Emotion,” in R. Plutchik and H. Kellerman (Eds.), *Emotion Theory, Research, and Experience*. pp. 83-111, Academic Press, N.Y., 1989

[7] A. Iida, “A Study on Corpus-based Speech Synthesis with Emotion,” Doctor of Media and Governance thesis, Graduate School of Media and Governance, Keio University, Sep. 2002

[8] R. Carlson, G. Granstrom and L. Nord, “Experiments with Emotive Speech, Acted Utterances and Synthesized Replicas,” *Speech Communication*, vol. 2, pp.347-355, 1992

[9] N. Frijda, *The emotions*, Cambridge University Press, N.Y., 1986

[10] L. K. Guerrero, P. A. Andersen and M. R. Trost, “Communication and Emotion: Basic Concepts and Approaches,” in P. A. Andersen and L. K. Guerrero (Eds.), *Handbook of Communication and Emotion: Research, Theory, Applications, and Contexts*, pp. 3-27. Academic Press, San Diego, 1998

[11] C. C. Kuo, C. S. Kuo, J. H. Chen and S. C. Chang, “Automatic Speech Segmentation and Verification for Concatenative Synthesis,” *in Proc. of Eurospeech’03*, Geneva, Switzerland, 2003

[12] M. Chu, H. Peng, H. Y. Yang and E. Chang, “Selecting Non-uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer,” *in Proc. of ICASSP’01*, vol. 2, pp.785-788, Salt Lake City, Utah, U.S.A., 2001

[13] C. H. Wu and J. H. Chen, “Automatic Generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis,” *Speech Communication*, vol.35, pp.219-237, 2001

仿趙氏音高尺度之基週軌跡正規化方法及其應用
A Pitch-Contour Normalization Method Following Zhao's Pitch Scale and Its Application

古鴻炎# 張小芬* 吳俊欣#
Hung-Yan Gu# Hsiao-Fen Chang* Jiun-Hsin Wu#

#國立台灣科技大學資工系 *國立台灣海洋大學
#National Taiwan University of Science and Technology, *National Taiwan Ocean University
e-mail: { guhy@mail.ntust.edu.tw , joanne@ntou.edu.tw }
www: http://www.csie.ntust.edu.tw/

摘要

本文研究一種音節基週軌跡的正規化方法，將音高由絕對式的赫茲(Hz)尺度轉換至相對式的趙氏(趙元任)音高尺度，以便解決跨人之音高位準和音域差異的問題。在對數赫茲尺度上，這個方法先對各個語者求出個人之平均音高和音高標準差，再用以作個人內部的音高正規化處理，接著依據跨人之音高分佈所呈現之分佈參數，作音高尺度之對應而轉換至所模仿之趙氏音高尺度。為了驗證所提方法的效能，我們收集了聽障與耳聰學生的語詞發音，比較未作、有作音高正規化處理之基週軌跡曲線的差別，結果顯示所提之音高正規化方法，的確可大幅減低跨人之音高位準、音域差異的影響。此外，我們也以這個音高正規化方法，來分析、探討聽障學生的聲調發音，在人耳評分上的一些現象。

1. 前言

許多語音處理相關的應用裡，都需要量測出語音信號內攜帶的基週(pitch)資訊，例如國語語音辨識之研究，需要基週資料來作聲調辨識[1, 2]；文句翻語音(text-to-speech, TTS)之研究，需要訓練語句的基週資料來建立基週軌跡的產生模型[3, 4]；語言學家研究各種語言的特性，也需要擷取出語音信號裡的基週資料來作分析[5, 6]。量測語音信號內的基週數值的問題，過去已有許多人作過研究及提出不錯的量測方法[7, 8, 9]，一般來說，程式自動量測出的基週數值，已有很高的可信度(雖然偶而還是會有量測錯誤)。

然而不同的語音處理之應用裡，各自對基週資料在後續處理上的需求，是不太一樣的，例如聲調辨識之研究，需要考慮使用者個人的音高位準和音域的問題；TTS 之基週軌跡模型，縱使不需考慮多個語者的音高差異問題，也仍需考慮同一語者在不同天所唸的訓練語句，音高位準存在變異的問題；趙元任先生提出的五度制調值標記法[5]，則是假設作調值標記的人，會對不同語者的音高、音域差異，作機動式的調適。

由前面提到的應用例子可知，對聲調語言的語音信號作處理，大多會牽涉到音高正規化的問題，不過，並不是一種應用裡發展出的音高正規化方法，就可直接使用於另一個語音處理的應用裡。以前我們研究 TTS 之基週軌跡產生模型，曾經發展了一種以語句爲單位作整體調整的音高正規化方法[4]，雖然由合成語音之聽測實驗，顯示此正規化方法之效能很不錯，不過它並未解決跨人之音高、音域差異的問題。再者，本論文將

要考慮的語音處理應用是，聽障與耳聰學生的跨人聲調的分析，無可避免地需要解決不同學生之間音高位準與音域寬窄差異的問題。音高位準的差異是很明顯的，因爲語者包含男、女學生，且同性別的學生中也包含了處於不同發育變音階段的人；至於音域指的是基週數值的變化範圍，這個也會因人而變，我們說某人說話像唱歌(音域較寬)，而某人說話很平淡(音域較窄)，音域就是這種感覺的一個重要的聲響(acoustic)線索。然而不管男、女生，講話抑揚明不明顯，我們都聽得懂他們講的話，這意味一般人都具有機動地調適音高位準、音域寬窄差異的能力。反觀電腦程式，要如何讓它也具有這樣的能力？

因此，我們開始思索如何把赫茲(Hz)尺度(scale)之絕對音高轉換成相對音高，並且音高尺度希望能夠仿傚趙元任先生的五度制(稱爲趙氏音高尺度)作法，藉以解決跨性別、年齡之音高正規化的問題，就我們所知，文獻上還未看過電腦自動作音高正規化以轉成趙氏音高的方法。後來我們發展了一種正規化的作法，這個作法精簡說來是，在量測出一個音節的一序列的基週資料後，由於序列的長度不一，因此要先作音長正規化的處理，然後將赫茲值之音高轉到對數(log)赫茲值之音高，接著依據各人的音高平均值與標準差值，作各人內部的音高正規化處理，再依跨人的音高值分佈的統計，取得尺度轉換之參數，而據以轉換至趙氏音高尺度。基週量測，音長正規化，及音高正規化的較詳細作法，在第二、三節中說明，之後第四、五則說明所提之正規化方法應用於聽障與耳聰學生之跨人聲調分析的結果。

2. 基週量測與音長正規化

在作音高正規化之前，有幾個相關的前置處理步驟要作，如圖 1 之整體處理流程所示，從一個詞彙的錄音檔案，取樣率 22,050Hz，16bits/sample，取出語音樣本，首先作

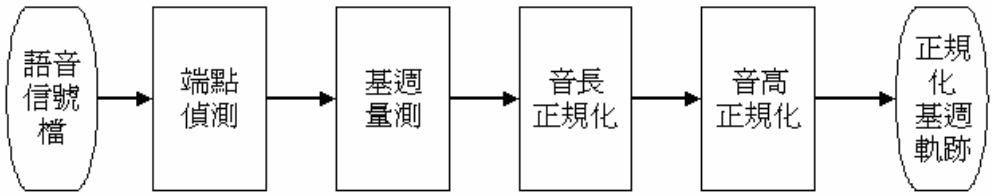


圖1 整體之處理流程

端點偵測，以決定各個音節的起點與終點；將起點與終點之間的語音樣本，切割成一序列的音框(frame)，以量測各音框內信號的基週數值，音框長度爲 25ms 且相鄰音框重疊 12.5ms；由於各音節的有聲(voiced)音框數量有多有少，所以接著要作音長正規化之處理，以便用固定 16 維(dimensions)之頻率向量來表示一個音節的基週軌跡；之後，進行主要的音高正規化處理，以轉換成仿趙氏音高尺度所表示之基週軌跡。

關於圖 1 裡的處理方塊的細部說明，”音高正規化”將於下一節說明，本節就對”基週量測”和”音長正規化”來作進一步說明，較基本的”端點偵測”，則可參考語音處理之文獻[9, 10]，不過有一點需注意的是，詞彙發音時，兩音節之間可能會有共發聲(coarticulation)現象，而使端點偵測發生錯誤，這時就需要人工來作更正。

2.1 基週量測

依據端點偵測得到之起點與終點，將兩點之間的語音樣本，切割成一序列的音框，再對各音框來作基週量測。音框內的信號可能是無週期性的(如無聲子音部分)，我們要能夠偵測出來，並直接設定基週值爲一個特殊值；若爲週期性的，則要量測出週期的長度值。

我們所用的偵測方法如下。先計算自相關(autocorrelation)函數 $R(k)$及平均振幅差距(average magnitude difference)函數 $M(k)$，計算公式分別是

$$R(k)=\frac{1}{N}\sum_{n=0}^{N-1}x(n)x(n+k)\quad,\quad k=k_{min},...,k_{max} \tag{1}$$

$$M(k)=\frac{1}{N}\sum_{n=0}^{N-1}\left|x(n)-x(n+k)\right|\quad,\quad k=k_{min},...,k_{max} \tag{2}$$

其中 $x(n)$ 表示語音信號樣本，N 表示一個音框內的樣本點數，k 表示可能的基週長度之點數。我們先設定最低及最高的音高值分別爲 60 及 750 赫茲，再據以設定 k_{min} 及 k_{max} 的數值。在公式(1)中，當 $k=0$ ， $R(0)$ 就是短時能量，我們也要計算此值，然後採用 Kim 等人提出的判斷規則[11]，用以判斷本音框內的信號，是無週期性或是有週期性的信號，兩個規則的內容是：

規則(1): if $\underset{k_{min}\le k\le k_{max}}{Max}\ (R(k)) < R(0)/4$ then not periodic

規則(2): if $\underset{k_{min}\le k\le k_{max}}{Max}\ (M(k)) / \underset{k_{min}\le k\le k_{max}}{Min}\ (M(k)) < E$ then not periodic

其中 E 爲門檻值，我們依實驗的結果設定其值爲 2.1。通過判斷規則後，再去計算基週數值，計算方式如下[11]：

$$f=\frac{22{,}050}{\underset{k_{min}\le k\le k_{max}}{Max}\ (\frac{R(k)}{M(k)+1})} \tag{3}$$

其中 22,050 是取樣率。

2.2 音長正規化

每個人講話的速度，或同一人在不同時間的說話速度都會不相同，導致音節內有聲音框的數量也呈現出有的較多有的較少，但是爲了方便作音高正規化之處理，所以在此先作音長正規化，以便用固定 16 維(dimensions)之頻率向量來表示一個音節的基週軌跡。

我們的作法是，在時間軸上均勻放置 16 個音高之取樣點，然後以內差的方式來求出各點上的音高。設本次分析的音節裡有 NF 個有週期性之音框，則第 k 個音高取樣點放置的時間位置(以音框爲單位)是：

$$t_k=\frac{NF-1}{16-1}\times k,\quad k=0,\ ...,\ 15 \tag{4}$$

以圖 2 爲例，設時刻 t_k 時的音高值 g_k 是所要求取的，則我們可取 t_k 前後各兩個音框所量測出的音高值來作 Lagrange 內差，而求得 g_k 的值。實作上，令圖 2 裡的 i 表示第 i 個音框，它的值由 t_k 決定，即 $i=\lfloor t_k\rfloor$ -1 (取整數後減 1)，若 i 的值小於 0，則改設 i 的值爲 0； f_i 表示第 i 個音框內量得的音高值；接著使用如下之 Lagrange 內差公式[12]

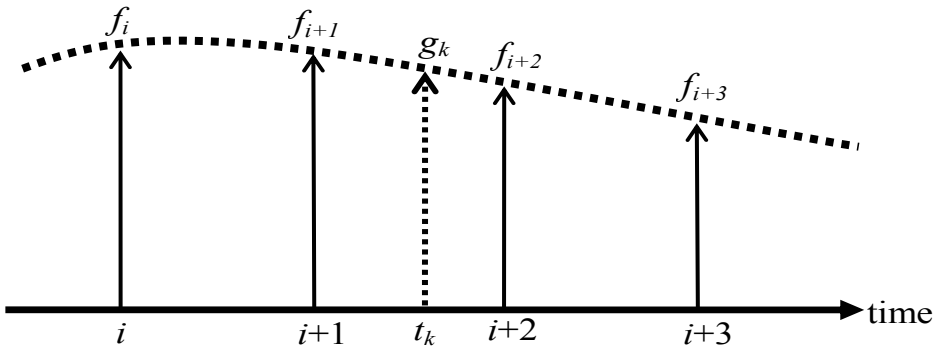


圖 2 音高內差示意圖

來求 g_k 之值：

$$g_k = \sum_{n=0}^{3} L_n(t_k) \cdot f_{i+n} \quad , \quad L_n(t) = \prod_{\substack{j=0 \\ j \neq n}}^{3} \frac{t-(j+i)}{(n+i)-(j+i)} \tag{5}$$

其中 i 就是圖 2 裡由 t_k 決定之 i 值。

3. 音高正規化

求出一個音節的 16 點時間正規化之基週軌跡 $g_0, g_1, \ldots, g_{15}$ 後，接著就將音高值轉換至對數尺度，即令 $h_k = \log_{10}(g_k)$, k=0, …, 15，這樣做是要仿傚人耳對音高的知覺，就如同音樂裡以十二平均律來排列音階，或語音辨識裡常用的梅爾(mel)尺度。然後，我們按照圖 3 的流程來作音高正規化之處理，圖 3 的"個人平均音高"方塊，分別對各個

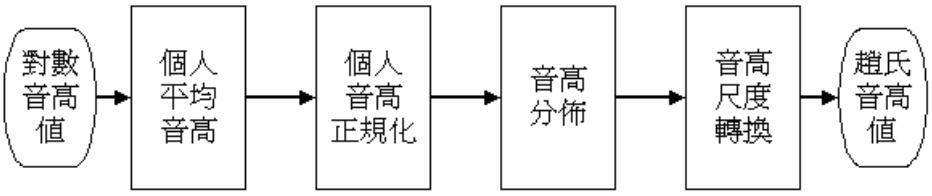


圖3 音高正規化之處理流程

語者的音高資料去算出各個語者個人的平均音高 μ 和音高標準差 σ，計算公式如下:

$$\mu = \frac{1}{NS \times 16} \sum_{n=0}^{NS-1} \sum_{k=0}^{15} h_k^n \tag{6}$$

$$\sigma = \left[\frac{1}{NS \times 16} \sum_{n=0}^{NS-1} \sum_{k=0}^{15} \left(h_k^n - \mu \right)^2 \right]^{1/2} \tag{7}$$

其中 NS 表示本次分析之語者所唸的音節個數，h_k^n 表示第 n 個音節的第 k 個基週軌跡點上的音高值。

求得一個語者的平均音高和音高標準差後，接著在圖 3 的"個人音高正規化"方塊，就依如下公式

$$\alpha_k^n = \left(h_k^n - \mu\right)/\sigma \tag{8}$$

來作個人內部的音高正規化處理，α_k^n表示第 n 個音節的第 k 點上的內部音高正規化後的值。h_k^n 扣掉 μ 就可把絕對式音高轉變成相對式音高，而解決了不同語者之間音高位準差異的問題；接著除以σ，則用以解決語者之間音域差異的問題。

做完每個語者的個人內部音高正規化處理，接著在圖 3 的"音高分佈"方塊，我們把所有語者唸的所有音節之音高資料α_k^n放在一起，作 Histogram 處理，由於大於 5(5 倍標準差)或小於 -5 的數值幾乎沒有，因此我們把α_k^n數值大於 5 的只當作一個區間，小於 -5 的也只當作一個區間，而在 5 與-5 之間，則切割成 100 個區間，如此我們得到如圖 4 所示的音高分佈圖，由此圖可看出，音高數值主要出現於橫軸第 20 至 80 區間之

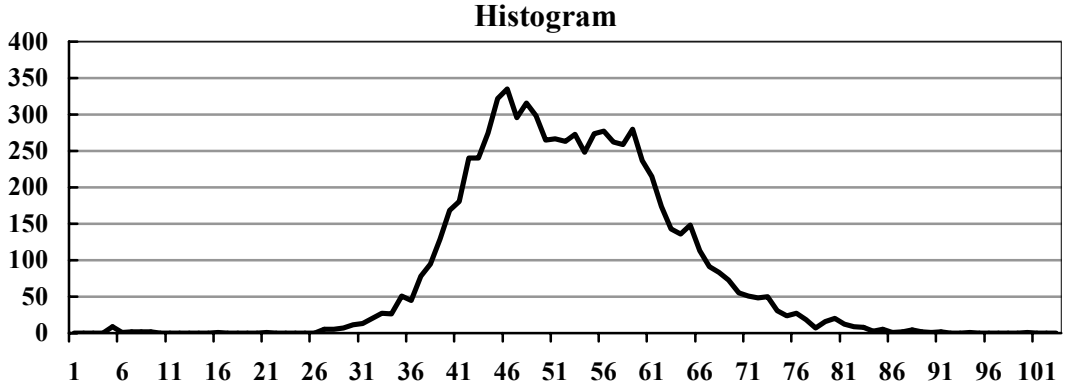


圖 4 跨人之正規化音高值分佈

間，也就是α_k^n的數值主要出現於-3 至+3 之間，並且分佈圖呈現出雙峰的現象，我們覺得這是可以解釋的，因爲國語裡第一聲的基週軌跡音高很高且相當平坦，所以造就右邊的峰，而第三聲大多唸前半上且音調較低，且詞尾位置的第二聲，音調也較低，所以形成左邊的峰。

依據圖 4 的音高值分佈，接著在圖 3 的"音高尺度轉換"方塊，我們考慮把α_k^n的數值 0 對應至趙氏音高尺度的中間音高，即數值 3，這是加 3 的關係。至於數值範圍的對應，趙氏音高尺度的數值範圍是 1 至 5，而我們的α_k^n的數值範圍主要是在-3 至+3 之間，範圍差距不大，因次就不作數值範圍的減縮調整，如此我們決定使用如下之音高尺度轉換公式：

$$\beta = \begin{cases} -3, & \text{if } \alpha < -5 \\ \alpha + 3, & \text{if } -5 \le \alpha \le 5 \\ 9, & \text{if } \alpha > 5 \end{cases} \tag{9}$$

來把個人內部正規化後之音高值 α_k^n 轉換至趙氏尺度的音高值β_k^n。公式(9)中，α值小於-5 或大於 5 的特殊情況，是用以處理聽障生所唸音節，有時會音調過高或過低。

4. 本方法之實驗驗證

對於第二、三節說明之仿趙氏尺度的音高正規化方法，在此節我們以實驗的結果來檢驗它的效用。這裡參與實驗的發音語者共計 14 名耳聰學生，均係說話清晰且國語發音標準者，其中 9 名爲國小生 5 名爲國中生，性別上則是男生 10 名女生 4 名。選用的發音詞彙爲: '冰棒'、'熨斗'、'帆船'、'肥皂'、'小熊'、'草地'、'崖谷'、'愛心'、'卡車'、'燈塔'、'西瓜'、'圍巾'、'照片'、'拼圖'、'氣球'等共 15 個，滿足雙字詞的所有聲調組合，且詞中第二字都含有子音聲母，以方便程式自動作音節切割。各詞彙以隨機出現的方式讓每人唸三遍，同時錄音存檔，之後選擇其中一遍聽起來聲調最準確且聲波振幅較佳者來作本節的實驗。

如果只作圖 1 中前三個方塊的處理，即不作音高正規化，且把各個聲調的基週軌跡分別集合起來繪圖，則我們得到如圖 5, 6, 7, 8 所示的四個聲調之基週軌跡帶狀圖，其中較粗且有菱形標記的實線爲平均的基週軌跡曲線。接著如果進行音高正規化的處理，則圖 5, 6, 7, 8 之基週軌跡帶狀圖，會分別變成如圖 9, 10, 11, 12 所示的情況。比較圖 9 至圖 12、和圖 5 至圖 8 這兩組基週軌跡帶狀圖，整體而言，音高正規化之處理可以讓軌跡線條由鬆散變得較爲集中；再者，可觀察到第一、二聲的軌跡，在作過音高正規化的圖 9、圖 10 之間，有較明顯的**音調高度**、與**軌跡彎曲度**的差異，而在圖 5、圖 6 之間，高度與彎曲度的差異則較不明顯；此外，觀察第三、四聲的軌跡，在作過音高正規化的圖 11、圖 12 之間，有較明顯的軌跡斜率的差異，而在圖 7、圖 8 之間，軌跡斜率的差異較不明顯。所以，本論文研究之音高正規化方法，的確可用以大幅減低個人音高位準、音域寬窄等因素所造成的影響。

對於作過音高正規化的圖 9, 10, 11, 12 裡，各聲調的軌跡線條，仍然表現出一些寬度的帶子形狀，而不是十分密集，這種情形是可以解釋的，我們所錄的 15 個雙字詞發音，各聲調都會均勻地出現在詞頭和詞尾，並且就同一個聲調而言，出現在詞頭的一般來說會比出現在詞尾的音調較高(即下傾現象)，所以，雖然是同一個聲調，音調高低仍然會存在著明顯的變異。

5. 聽障生之語詞聲調分析

這裡參與實驗的聽障生人數、年齡、性別都和耳聰生一樣，所以一共有 14 人， 9 名國小生加 5 名國中生，而性別上則是男生 10 名女生 4 名。參與的聽障生，都是語言學習前失聰，包括中度 2 名、重度 6 名，極重度 6 名，且都有接受聽障班或聽障資源班之聽覺口語訓練。在此選用的發音詞彙和耳聰生的完全一樣，並且各詞彙也是以隨機出現的方式讓每人唸三遍，同時作錄音存檔，之後選擇其中一遍聽起來聲調最準確且聲波振幅較佳者來作本節的實驗。

聽障生所唸的雙字詞發音，基本上也是按照圖 1 的流程來作處理，不過在圖 3 的〝個人音高正規化〞方塊，我們認爲聽障生的音高標準差值，有很大的可能性是不正確的，因爲聽障生的一種發音障礙情形是，分不清四個國語聲調的相對音調高低，因此，我們在處理聽障生的個人音高正規化時，音高標準差值是直接設定爲 14 個耳聰生的音高標準差的平均值。作完音高正規化的計算後，各個音節的基週軌跡已是在趙氏音高尺度上

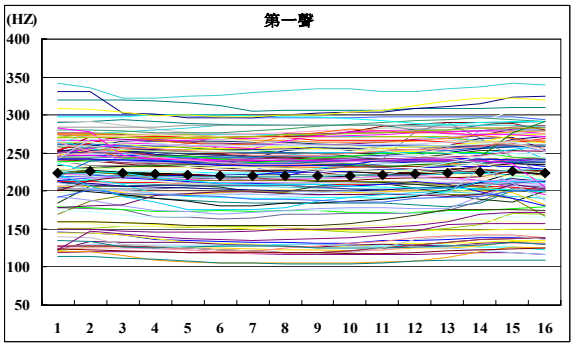


圖 5 第一聲基週軌跡(未音高正規化)

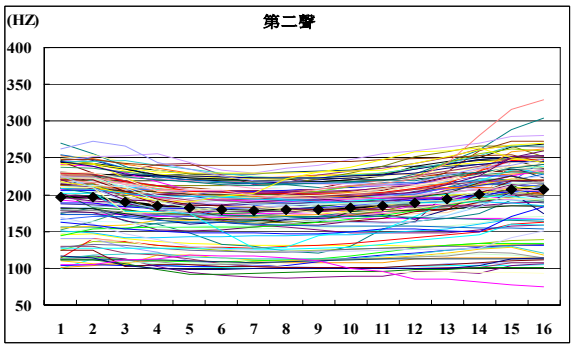


圖 6 第二聲基週軌跡(未音高正規化)

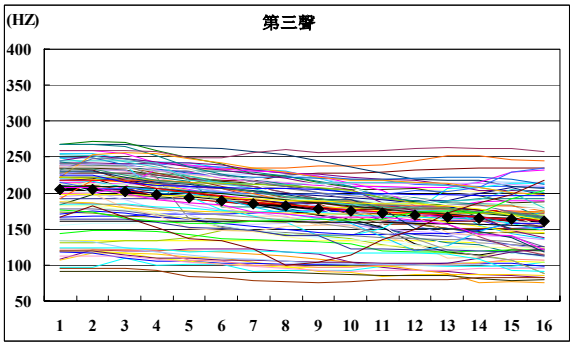


圖 7 第三聲基週軌跡(未音高正規化)

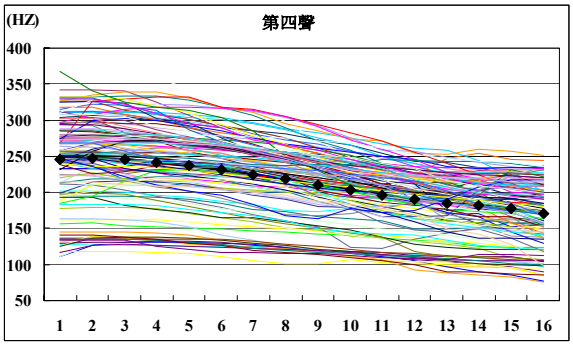


圖 8 第四聲基週軌跡(未音高正規化)

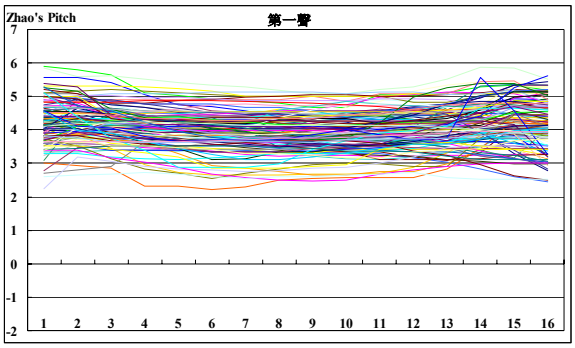


圖 9 第一聲基週軌跡(音高正規化後)

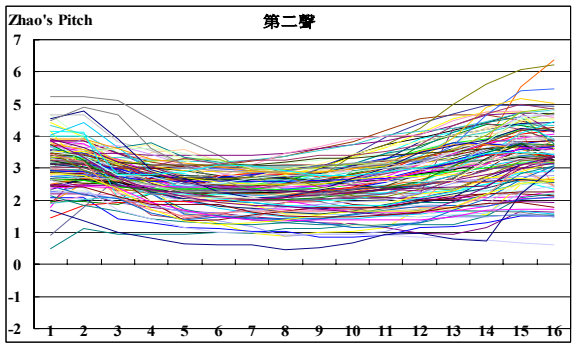


圖 10 第二聲基週軌跡(音高正規化後)

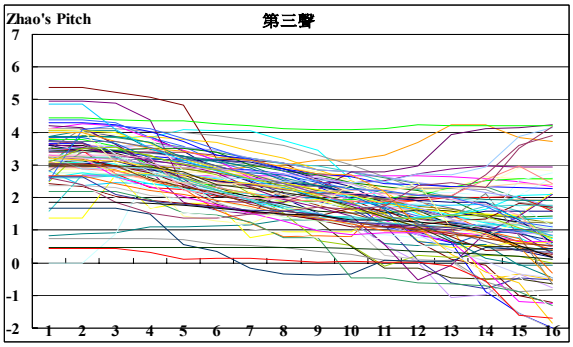


圖 11 第三聲基週軌跡(音高正規化後)

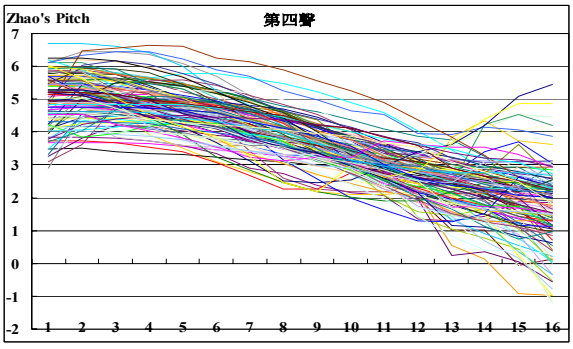


圖 12 第四聲基週軌跡(音高正規化後)

的數值，接著把各個聲調的基週軌跡分別集合起來，再依據詞頭與詞尾之發音位置分成兩組，各自去計算平均的基週軌跡，用以分析各個聲調的基週軌跡是否和詞頭、詞尾之位置有相關性。我們依照聲調及詞內發音位置，分別把聽障生和耳聰生的平均基週軌跡畫在同一圖裡，以方便作比較，結果得到如圖 13 ~ 圖 20 所示之平均基週軌跡圖形。

由圖 13 ~ 圖 20 可觀察得知，就耳聰生而言，第一、三、四聲的基週軌跡形狀，和詞內的發音位置並無明顯的相關性，但是第二聲的基週軌跡形狀則和發音位置有非常明顯的關係存在，即在詞頭位置所發的第二聲，後端具有大幅度的上揚趨勢，而在詞尾位置所發的第二聲，後端上揚的趨勢很和緩，反而前端下降的趨勢更爲明顯。此外，對於第一、二、三、四聲都共同具有的現象是，同一個聲調的兩個發音位置，在詞頭位置的都會具有比較高的音高。就聽障生而言，各個聲調的基週軌跡形狀，都和詞內發音之位置沒有明顯的相關性。此外，聽障生對於兩個發音位置，在詞頭位置也會表現出比詞尾位置較高一些的音高。

另一方面，從圖 13 ~ 圖 20 也可觀察到耳聰生和聽障生的不少差異的地方。例如第二聲的基週軌跡(圖 15、16)，耳聰生的兩個位置的軌跡都會先下降再上揚，反觀聽障生的軌跡則都只有下降，而無上揚的趨勢，在軌跡的尾端，耳聰生和聽障生的軌跡甚至於呈現完全相反(一個上升一個下降)的走勢。再者，觀察第三聲的基週軌跡(圖 17、18)，聽障生和耳聰生之間也有明顯的差異，聽障生的軌跡較爲平緩，下降速度明顯地比耳聰生的要緩慢很多，以致於聽障生本身的的第二聲與第三聲的基週軌跡，無論在軌跡形狀和音高方面，幾乎是很難加以區分。這與潘奕陵[13]與鍾玉梅[14]之研究相呼應，即聽障兒童發音方面構音的轉換速度太慢，由此可見，聽障學生在第二聲和第三聲的發音最無法區辨，也就是第二聲、第三聲的發音最爲困難。

從人耳對聽障生唸的聲調所作的評分，也發現聽障生四個聲調得分會因聲調不同而有差異存在。人耳評量係由修習教育學程的三位研究生擔任，在評量前均接受過講習，講習重點除說明評量標準與計分方式，也實際就語音樣本進行演練，確認給分標準與評鑑項目之後才分開進行評分工作，每位受試之各個語詞聲調得分，給分標準必須符合三人中至少有兩人給分相同，若任何一項得分不一致時，需三人同時再聽一次，重新確定真正得分。從評分結果我們得到，平均分數上第二聲和第三聲的分數最低，而第一聲和第四聲的分數較高，這樣的評量結果，與過去有關聽障學生之聲調研究的結果是一致的[15, 16]。所以，人耳評分的結果確實可由圖 13 ~ 圖 20 之音高正規化的基週軌跡曲線來作聲學分析上的佐證。第一聲和第四聲的得分會較高許多，這個也可從圖 13、14 和圖 19、20 的基週軌跡曲線來分析，因爲發第一聲時，聽障生和耳聰生的軌跡都相當平坦且音高夠高，雖然在軌跡的左右兩端，聽障生和耳聰生的軌跡仍存在一點差異；至於唸第四聲時，聽障生和耳聰生的軌跡都呈現出劇烈的下降走勢，音高的高低落差至少是縱座標的兩個單位以上，所以可讓人耳感覺出是第四聲，不過聽障生的軌跡的下降幅度明顯地是比耳聰生的少許多。

6. 結論

本文研究提出一種跨性別、年齡的音高正規化的方法，將音高由絕對式的 Hz 尺度轉換至相對式的趙氏(趙元任)音高尺度，如此用以解決不同語者之間音高位準和音域差

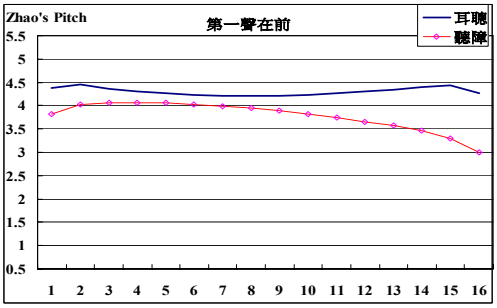


圖 13 第一聲在詞頭的平均基週軌跡

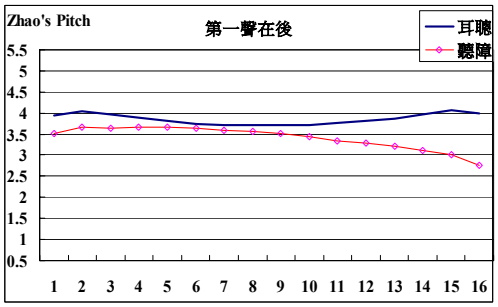


圖 14 第一聲在詞尾的平均基週軌跡

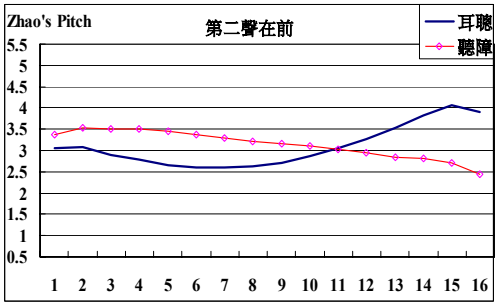


圖 15 第二聲在詞頭的平均基週軌跡

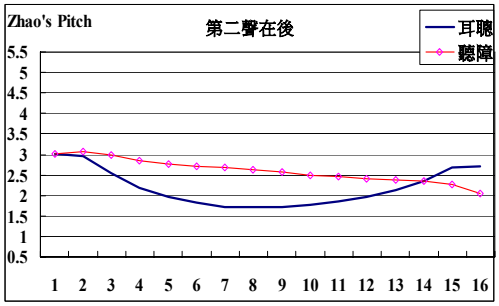


圖 16 第二聲在詞尾的平均基週軌跡

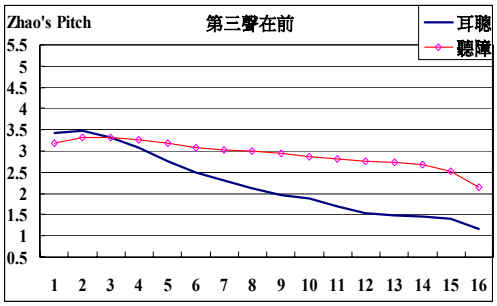


圖 17 第三聲在詞頭的平均基週軌跡

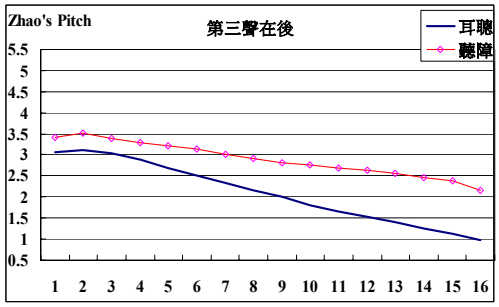


圖 18 第三聲在詞尾的平均基週軌跡

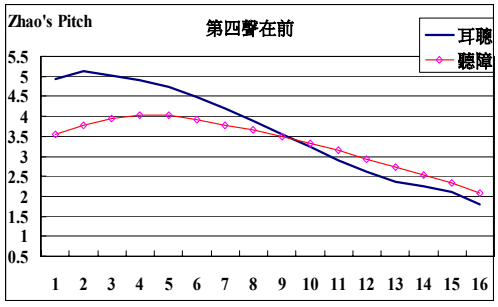


圖 19 第四聲在詞頭的平均基週軌跡

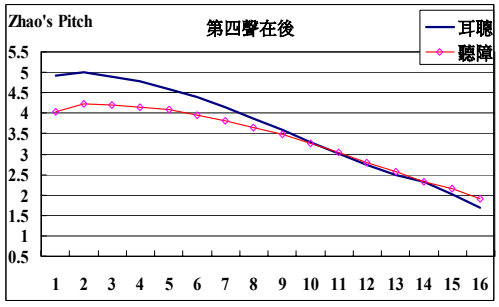


圖 20 第四聲在詞尾的平均基週軌跡

異的問題。趙氏音高尺度，是語言學家分析聲調語言最常用之音高尺度，但是過去只能靠人耳去調適、標記受測者講話的音高，難免有時會不甚精確，因此我們所提出的方法，也許仍有缺點，但至少是一種客觀、電腦自動作音高分析的方法。這個方法的處理流程也許看來很簡單，不過它的效能的確已經經過實驗的驗證。

對於所提出的仿趙氏音高正規化方法，我們以耳聰學生所錄的語詞發音，來比較未正規化前和正規化後的基週軌跡曲線，發現音高正規化處理可以讓軌跡線條由鬆散變得較爲集中，並且第一、二聲之間的軌跡差異，在音調高度和軌跡彎曲度兩項上，會變得更爲明顯；第三、四聲之間的軌跡斜率的差異，在作過音高正規化後，會變得更明顯而容易分辨。此外，我們也以所提之方法來對聽障生和耳聰生的基週軌跡作處理，之後再比較聽障生和耳聰生的軌跡差異，發現音高正規化後的軌跡，的確可用以解釋人耳對聽障生聲調評分的一些現象。

參考文獻

[1] Yan, W. J., J. C. Lee, Y. C. Chang and H. C. Wang, “Hidden Markov Model for Mandarin Lexical Tone Recognition”, IEEE trans. ASSP, Vol. 36(7), pp. 988-992, July 1988.

[2] Gu, H. Y. and Lin-Shan Lee, A Study on a few Relevant Problems about Machine Dictation of Mandarin Speech, Ph.D. Dissertation, Department of Computer Science and Information Engineering, National Taiwan University, Taiwan, Jan. 1990.

[3] Chen, S. H., S. H. Hwang and Y. R. Wang, “An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech”, IEEE trans. Speech and Audio Processing, Vol. 6(3), pp. 226-239, 1998.

[4] Gu, H. Y. and C. C. Yang, "A Sentence Pitch Contour Generation Method Using VQ/HMM for Mandarin Text-to-speech", International Symposium on Chinese Spoken Language Processing (ISCSLP2000), Beijing, pp. 125-128, 2000.

[5] 趙元任，中國話的文法，台北: 敦煌書局，1981。

[6] Tseng, C. Y., An Acoustic Phonetic Study on Tones in Mandarin Chinese, Special Publications No. 94, Institute of History & Philology, Academia Sinica, Taiwan, 1990.

[7] Rabiner, L., M. Cheng, A. Rosenberg and C. McGonegal, “A Comparative Performance Study of Several Pitch Detection Algorithms”, IEEE trans. ASSP, Vol. 24, pp. 399-418, Oct. 1976.

[8] Medan, Y., E. Yair and D. Chazan, “Super Resolution Pitch Determination of Speech Signals”, IEEE trans. Signal Processing, Vol. 39(1), pp. 40-48, Jan. 1991.

[9] O’Shaughnessy D., Speech Communications: Human and Machine, 2nd ed., IEEE Press, 2000.

[10] Rabiner, L. and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, 1993.

[11] Kim, H. Y., *et al.*, “Pitch Detection with Average Magnitude Difference Function Using Adaptive Threshold Algorithm for Estimating Shimmer and Jitter”, Proc. of the 20th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society, Vol. 6, pp. 3162 -3164, 1998.

[12] Stoer, J. and R. Bulirsch, Introduction to Numerical Analysis, 2nd ed., New York: Springer-Verlag, 1993.

[13] 潘奕陵，聽覺障礙者語詞及句子層次的說話清晰度之知覺分析，碩士論文，特殊教育研究所，國立台灣高雄師範大學，1998。

[14] 鍾玉梅，「聽障兒童的說話問題」，聽語會刊，第 10 期，第 72-79 頁，1994。

[15] 張蓓莉，「聽覺障礙學生說話清晰度知覺分析研究」，特殊教育研究學刊，第 18 期，第 53-78 頁，2000。

[16] 張淑品，國中重度聽障學生與耳聰學生國語單元音與聲調的聲學比較分析，碩士論文，特殊教育學系，國立台灣師範大學，1999。

致謝

感謝國科會計畫支援，計畫編號: NSC 91-2520-S-019-002.

基於反轉檔查找與最佳片段選取演算法的中文語音合成系統

林政源 謝明峰 陳冠廷 張智星

國立清華大學資訊工程學系

{gavins, pacific, marco, jang}@cs.nthu.edu.tw

摘要

本論文主要是解決以大量語料庫為基礎的語音合成的兩個問題，其一是搜尋比對大量語料庫非常費時，其二是從不同語句所取出的片段語音檔來加以接合，因為韻律參數的不一致，會使聽者明顯感覺不自然。因此，我們提出了反轉檔查找技巧來解決搜尋時間的問題，為求整體句子的自然韻律表現，我們提出了最佳片段選取演算法來達成這個目標，而對於PSOLA在調整音長表現可能不佳的情形，我們改以WSOLA方式實作。在搜尋比對時間與MOS評分的實驗中，我們均獲得到了不錯的成果。

1 系統簡介

近年來，隨著電腦科技不斷的蓬勃發展，中文文字轉語音 (TTS, Text-To-Speech) 的合成系統也慢慢朝向由單音節為主的合成單元架構轉變成以大量語料庫 (large corpus-based) 為主的合成單元架構。這方面的研究目前有 Heo-Jin Byeon 的 Event-Driven f0 Weighting[5], 大陸學者 Min Chu 等人的 Domain Adaptation[1]的方法, Ivan Bulyko提出的BMM models[6]以及台大周福強博士的decision trees方法[10]等。

一般而言， 採用大量語料庫的系統，其合成品質較單音節為主的系統來的好。因為它的方法是直接從語料庫擷取所需要的片段進行接合，所以在韻律表現上會較自然，也因為如此，在聲音方面所需調整的地方就會不太多，這也避免了聲音經過調整後而造成音質破壞的疑慮。然而，採用大量語料庫的做法也有其缺點，以下列出二個常見的問題:

1. 輸入文句需要和大量語料庫作比對：
   文句經過斷詞以後，再去語料庫找尋可能的詞句片段，並取出後加以接合，然而若演算法設計不當則會讓比對時間相對費時，所以發展一個有效率的演算法來縮短比對時間對系統的效能是非常重要的。
2. 詞句片段之間的韻律參數差異性問題：
   從不同語句所取出的片段語音檔來加以接合，因為韻律參數的不協調，會使聽者明顯感覺不自然。

有鑑於兩種缺點的考量，本論文採用反轉檔查找技巧來降低比對時間，而以動態規劃演算法來尋找最佳的接合片段使其合成自然度提升。這兩種方法將在第三以及第四節中論述。

2 系統架構說明

本論文所建立的中文語音合成系統架構將如下圖表示：

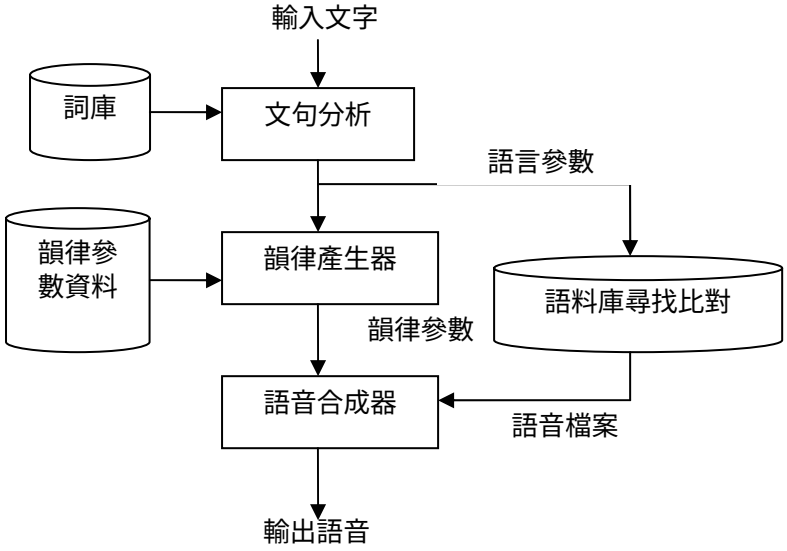


圖1. 中文文字轉語音系統流程圖

此系統主要分為四大類：

1. 文句分析：將所輸入的文字加以分析，得到音節以及詞的語言參數。
2. 韻律產生器：將語言參數轉換成語音合成所需要的韻律參數，而韻律產生器所需要的參數資料，是以類神經網路來獲得。
3. 語音合成器：根據韻律參數，將語料庫中所得到的語音檔案加以調整。
4. 語料庫搜尋比對：這是本論文最重要的一環，主要是將分句分析的結果和語料庫作比較查詢，並找出最適當的語音檔案當作輸出。

2.1 文句分析

當文句輸入時，第一步驟就是針對此文句做分析，以得到其語言參數，如此才可進一步的得到韻律參數，合成出所需要的語音。而所謂語言參數，又可以分為音節的語言參數和詞層的語言參數。文句分析的系統如下圖所表示。

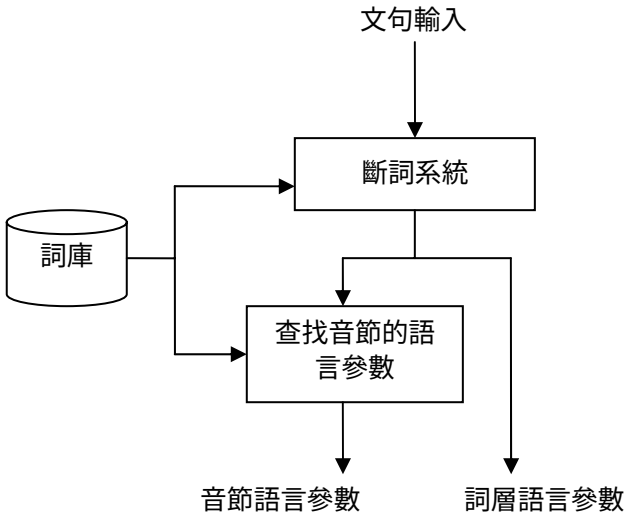


圖2．文句分析系統流程圖

在文句分析中，斷詞的處理為最重要的部分。而本篇論文的斷詞方法是根據一個大詞庫 (中研院漢語平衡語料庫，Sinica Corpus 3.0 [18]，共有130,757個詞。)，然後輸入語句再比對此大詞庫來進行查找。在斷詞的研究上，也有相當多的方法[12][11][16]，我們為了系統的效率則採用長詞優先法，再以各種構詞的方法補足其詞庫不足的缺點。

2.2 韻律產生器

語音合成系統的關鍵技術就在於韻律的變化是否平順自然。而韻律的變化包括音調高低起伏、音量的大小變化、每個音節的長短及停頓這三個部分。而韻律產生器大致上有幾種方法：規則法、統計法、類神經網路法[8][17]。目前大多數的實驗結果以類神經網路為較佳，故本論文採用其方式來製作韻律產生器。在類神經訓練的實作方面，輸入是音節和詞層的語言參數，輸出是韻律參數。而音節語言參數包括本音節的聲母、韻母、音調，下個音節的聲母、音調等。詞層語言參數包括本詞詞長、下一個詞的詞長和本音節在本詞的位置、本詞和下詞之間的標點符號。輸出的韻律參數為音節間的停頓時間、聲母長度、韻母長度、音節的韻母平均能量、基頻軌跡。其中基頻軌跡參數是以正交化展開的前四階係數[13]表示。

而基本的單層類神經網路函式運算時，輸入參數有 m 個輸入參數 $i_{1\sim m}$ 和 n 個輸出參數 $o_{1\sim n}$，而輸出參數和輸入參數的關係為：

$$o_j = f(\sum_{k=1}^{n} i_k \cdot w_{k,j} + b_j)$$

$w_{k,j}$ 為第 k 個輸入神經元到第 j 個輸出神經元的加權值，b_j 則是偏差值(bias)，其關係圖如下圖：

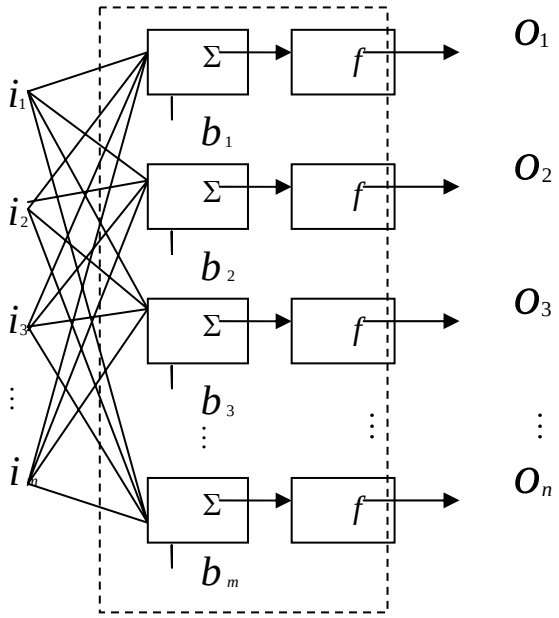


圖3 單層類神經網路結構圖

而 $f(x)$ 可以自行定義，本論文使用最簡單的線性轉換函數(Linear Transfer Function)，即：

$$f(x) = x$$

2.3 語音合成器

語音韻律參數主要包括音調高低軌跡、語音長度、音量大小三部分。而在調整音調方面，常用的方法有大致上分為兩種：Sinusoidal Modeling [4]與PSOLA (Pitch Synchronous Overlap and Add) [2]，雖然Sinusoidal Modeling方法的合成音質與PSOLA 相當，但是基於執行效率的考量以及音高調整幅度通常不大，我們採用後者(PSOLA)作為基本校正音調的方法。而調整音長的方法，本論文所採用的是WSOLA (Overlap-add Technique Based on Waveform Similarity)方法[9]。採用此方法的理由是：我們的語音合成單元大部分來自大量語料庫，所以無法對每一個語音檔作基週標位的修正，然而使用PSOLA方法調整音長時，基週標位須十分準確，否則音長調整後的音質會有雜音的現象，但是使用WSOLA方式調整時，並不需要基週標位的資訊，WSOLA是用AMDF (Average Magnitude Difference Function)[3] 進行音框比對，藉以找出最適合的音框作波形疊加。所以WSOLA單就調整音長而言，其效果較PSOLA為佳。

2.4 語料庫搜尋比對

大量語料庫的語音合成系統必須能在大量的資料中，找到需要的片段加以接合。而在實作上，會遇到以下的問題：

1. 需要有較大的儲存空間：

   目前大部分的Embedded的系統較不適合採用此方法實做TTS系統，所以目前以大量語料庫為基礎的TTS都是在PC的硬體上執行居多。由於硬碟容量日漸增大、語音壓縮的技術也不斷地改良，所以用PC實作的TTS系統，也會更加的普及與實用。
2. 搜尋語料庫中所需的片段：

   如何從大量的語料庫中，快速地去找到最適合的片段來接合，是這方面設計最需要解決的問題。我們將採用反轉檔的查找技巧來克服這個問題，即使語料庫的大小增為兩倍，我們的搜尋時間也不會線性成長兩倍，甚至兩者的時間差異極小。然而，除了搜尋時間的問題之外，我們希望找到的片段長度越長越好，這樣韻律的表現最為自然，所以我們提出建立最長詞數表的方法並搭配之前的反轉檔可以很快的找到所需要的最佳片段。這些方法將在第三節中加以闡述。
3. 片段與片段之間差異過大：

   從大量語料庫中所取出來的片段，會受到前後音、句子節奏韻律和個人情緒的影響，造成片段與片段之間韻律參數有極大的差異，會讓聽者明顯感受出是由不同片段組合而成的語音。例如，片段的平均音高或者音量差異過大、片段間的接合不連續等。在本論文中，我們提供了另一個基於動態規劃演算法為基礎的最佳片段選取法來解決這方面的問題，這將在第四節會加以說明。

3 反轉檔與最長連續詞數表

3.1 反轉檔查找

反轉檔的目的是在改變原本語料庫的資料結構以減少搜尋時間。若原來的資料是經常變動的，就不適合採用反轉檔 (因為每次的變動都要再重建反轉檔)。而在本系統中，大量語料庫的資料是固定的，因此可以使用反轉檔的技巧來進行查找的動作。首先我們必須先將文句編號，再進行斷詞的動作。下面是語料庫中有的文句：

表1　語料庫中的句子

語料庫中的句子	1.母親 真 偉大 2.我 母親 是 老師 3.當 老師 是 他 一生 的 夢想 4. ….

根據以上的句子，我們可以建立以下的反轉檔：

表2 反轉檔範例

母親	<1>,<2>
老師	<2>,<3>
夢想	<3>
是	<2>,<3>
⋮	⋮

然而，反轉檔只存入出現的句子編號，並不能讓我們快速的找到本詞和下一個詞的關係，因此在反轉檔中我們要存入這個詞在句子中，所連接的下一個詞。因此反轉檔變成以下格式：

表3 反轉檔範例二

母親	<1,真>,<2,是>
老師	<2, Φ >,<3,是>
夢想	<3, Φ >
是	<2,老師>,<3,他>
⋮	⋮

不過，由於語料庫的資料量龐大，一個詞往往會重覆出現多次，所以必須再加入一個數值，就是下一個詞出現在該詞反轉檔的哪一個位置，而範例反轉檔如下：

表4 反轉檔範例三

母親	<1,真,1>,<2,是,1>
老師	<2, Φ ,0>,<3,是,2>
夢想	<3, Φ ,0>
是	<2,老師,1>,<3,他,1>
⋮	⋮

上例的第一列第二行：<2,是,1> 表示 『母親』這個詞是在出現在大量語料庫中的第2句，它的下一個詞是『是』，而『是』這個詞是出現在它反轉檔的第 1 位置（『是』在大量語料庫中可能有好幾個），所以我們根據上例來看，『是』的第 1 個位置所擺放的是 <2,老師,1>，如此我們又可以繼續追蹤下去。建立這個反轉檔的資料結構之後，就可以很快的找到在語料庫中，每個詞的下一個詞的反轉檔位置，系統也就可以快速的找到輸入文句中任何一個詞開始，接下來連續最長的文句了。

3.2 建立最長連續詞數表

當語音合成系統輸入文句時，需要立刻從大量語料庫中找到相同的文字片段，並加以取出。而找到片段的原則是每個片段的字數越多越好。在同個片段中，是由人在同一時間所錄下的連續語音，所以一定是最自然的，在語音合成的觀點上，當然是越自然越好。因此在搜尋時，以找到最長的片段為優先。然而，當語料庫十分龐大時，在尋找比對的所花的時間甚巨。又因為希望能有單一最長的片段，無法使用由左往右找出最長連續片段的方式來進行。而當找到最長片段並取出時，還要從剩下的文句繼續比對語料庫，再找出次長的片段，如此反覆進行，計算量將會非常大。在這種情況之下，我們可以先建立以下表格：若輸入的文句可以被斷詞系統斷出 N 個詞，而我們要找到從每個詞 S_n 開始，和資料庫中最長的連續詞數。其中 $1 \le n \le N$ 。例如欲輸入下列文句：

母親 明年 將 離開 台南 前往 嘉義

而在語料庫中有下列相關語句，括號部分是與輸入文句相同的部分：

表5 與輸入文句有關的語料庫範例

語料庫中和輸入文句相關的句子	1. (母親) 真 偉大 2. 他 (明年 將 離開) 台北 3. (離開 台南) 後 的 生活

	4. (前往 嘉義) 的 路 很 遙遠

可以得到以下的表格：

表6 每個詞開始最長連續詞數表格

輸入文句的斷詞	母親 S_1	明年 S_2	將 S_3	離開 S_4	台南 S_5	前往 S_6	嘉義 S_7
從本詞起最長的連續詞數	1	3	2	2	1	2	1

根據此一表格，就可以找到最長連續片段。再將最長連續片段從輸入文句中取出來，而被取出來的詞在其表格中的數字補上0，以上表為例，會得到以下表格：

表7 取出片段後的連續詞數表格

輸入文句的斷詞	母親 S_1	明年 S_2	將 S_3	離開 S_4	台南 S_5	前往 S_6	嘉義 S_7
從本詞起最長的連續詞數	1	0	0	0	1	2	1

而表格內還未成為0的詞，就是剩下來仍需從大量語料庫取出的句子。在所有的數字還未變成0之前，仍需要重覆取出數字最大的部分做處理。本例到最後會斷成以下句子:

（母親）（明年 將 離開）（台南）（前往 嘉義）

4 最佳片段選取演算法

在大量語料庫中尋找所需的片段之後，我們並非直接拿來作語音檔的接合，因為這樣的接合會造成不自然的韻律，這裡可能的問題大致上有兩種：

1. 最佳片段選取問題：
   因為片段是從句子中所取出來的，所以同一種詞在不同句子中所表現的韻律就會有所不同，例如聲音的音量、音高或音長等。所以我們必須在這個詞所有出現的句子中，找到最適合的片段來合成。
2. 片段與片段之間的接合問題：
   即使找到最適合的片段來合成，但還是存在前後片段的接合不協調的問題。所以我們也必須在每一句的所有片段，找到它們最佳的組合方式來克服這個問題。

事實上，關於以上的兩個問題，可以同時以我們所提出的動態規劃演算法來解決。以下就是我們針對每一個句子的最佳選取片段演算法:

1. 我們制定狀態機率 (State Probability) 來定義 較佳的可能片段，通常保留前三名可能的片段，即每一個詞皆有三個候選片段。
2. 我們制定狀態轉移機率 (Transition Probability) 來定義片段之間可能組合的選擇。
3. 最後，根據前兩者累積的機率值，由最大機率值的片段回溯找出最佳可能的組合路徑。

4.1 狀態機率

首先，關於第一項算出狀態機率，我們考量到即使選出最佳的片段後，仍需要參考韻律參數而加以改變才作接合的處理，所以我們定義其機率的計算應該要根據音高和音長的差異來制定，也就是希望需要調整的韻律不要與原先的韻律差別太大，使得經由改變音高音長而破壞音質的情況降低。所以我們制定了以下的公式:

$$dist_i(j) = \sum_{j=1}^{M}\sum_{k=1}^{N} \frac{\left|Pitch(j,k) - TrainPitch(k)\right|}{Pitch(j,k)} + \frac{\left|Duration(j,k) - TrainDuration(k)\right|}{Duration(j,k)}$$

有 M 個候選詞，一個詞有 N 個音節，所以 $1 \le j \le M$ ， $1 \le k \le N$ ， $Pitch(j,k)$ 為第 j 個候選詞的第 k 個音節的平均基頻軌跡，單位為 Hz ， $TrainPitch(k)$ 是第 k 個音節經過韻律產生器所產生出來的平均基頻軌跡。而 $Duration(j,k)$ 為第 j 個候選詞的第 k 個音節的音長， $TrainDuration(k)$ 是經過韻律產生器的音長， $Duration$ 為聲母和韻母音長的加總，單位為秒。而結果 $dist_i(j)$ 表示句子之中第 i 個詞語的第 j 個候選詞的距離值。

計算出每個候選詞的dist值之後，只保留前3個最小距離的候選詞，而狀態機率就以其距離的倒數成正比，它的定義如下:

$$Sprob_i(j) = \left(\frac{\frac{1}{dist_i(j)}}{\sum_{k=1}^{3} \frac{1}{dist_i(k)}} \right)$$

上例公式說明：第i個詞語的第j個候選詞片段它的狀態機率值公式。

4.2 狀態轉移機率

至於狀態轉移機率，我們考慮的因素並不再以候選片段的音高或音長當作特徵了，因為前後片段的音高或音長本來就會不一樣。我們這邊所考量的方向是希望所選出來的候選片段組合，有最通順的接合品質，不會讓聽者感到整句話是由幾個詞分開唸出來的效果。而能達成最好的接合品質，就是去觀察每個候選片段的前一個音或者它的後一個音，然後跟其他候選片段來作比較。

舉例來說，假設某一個句子有個5個詞片段：〔今天〕〔去〕〔台北〕〔買〕〔衣服〕，句子中的每個片段皆有3個候選詞。如何決定〔去〕這個片段的第一個候選詞與〔今天〕這個片段的哪一個候選詞有最佳的接合效果，我們所採用的是相近音查表法來決定。例如〔去〕這個片段的第一個候選詞，它在原句子中其前面所接的字是〔點〕（例如原句是：她點去了這個痣），此時我們便可利用此資訊去查詢相近音表的〔點〕與〔天〕的相似程度，這裡的相似程度是以音節中的韻母來比較，而上例的韻母則是‘ㄧㄢ’。

另外，我們也需要判斷這三個〔今天〕的候選片段它們後面所接的音，然後與〔去〕作相似音的比對。例如，共有3個候選詞〔今天〕，第一個〔今天〕後面所接的字為〔我〕，第二個〔今天〕後面所接的字為〔天〕，而第三個〔今天〕後面所接的字為〔氣〕，如此我們將會給予第三個候選詞〔今天〕有較高的狀態移轉機率。而這裡的相似程度是以音節中的聲母來比較，而上例有較高的狀態移轉機率的聲母則是‘ㄑ’（因為〔去〕的聲母也是‘ㄑ’。）

至於如何決定相似程度的高低，則可採用兩種方式來計算，一種就是直接採用rule based的國語聲韻母分類表[15]，另一種則是利用語音辨識的技巧去統計那些聲韻母的發音最為相近，由於rule based所定義的分類表較難以具體描述其相似度的高低，所以我們採用後者的方式去統計聲韻母的發音，建立了一個相近發音查詢表，包含兩種統計模式，一是聲母相近發音的統計，另一個則是韻母相近發音的統計。所以狀態移轉機率的公式定義如下:

$$Tprob_i(j_1, j_2) = \left(\frac{similarTable(j_1\_nextword,\ j_2\_prevword)}{\sum_{k=1}^{3} similarTable(j_1\_nextword,\ k\_prevword)} \right)$$

上例公式說明：第 i 個詞語中的第 j_1 個候選片段到第 $i+1$ 個詞語中的第 j_2 個候選片段的狀態轉移機率值，而similar Table表示相近發音查詢表，可查詢聲韻母之間的距離。這裡的聲母距離指的是第 j_1 個候選片段的下一個字的聲母和第 j_2 個候選片段首字的聲母作比較，而韻母距離則是指第 j_2 個候選片段的上一個字的韻母和第 j_1 個候選片段尾字的韻母作比較。

4.3 累積機率以及回溯最佳路徑

累積機率的方式本質上是採用乘法，但是電腦的精確度是有限位數，所以我們將累積機率的方式改為加法，因此最後機率值會再取ln。所以第 i 個詞語的第 j 個候選片段的累積機率值其定義為 $P(i,j)$ ：

$$P(i,j)=\max_{k}\left(P(i-1,k)+\ln(Tprob_i(k,j))\right)+\ln(Sprob_i(j))$$

初始值 $P(1,j)=\ln(Sprob_1(j)),\quad for\quad j=1\ \ to\ \ n$ (在本論文,n = 3)

$B(i,j)$ 紀錄了第 i 個詞語中的第 j 個候選片段，它的前一個詞語與它有最佳的接合效果的候選片段位置：

$$B(i,j)=\arg\max_{k}\left(P(i-1,k)+\ln(Tprob_i(k,j))\right)$$

所以經由最高累積機率的 $P(lastword,j)$ 值回溯到 $P(1,j)$，參考相對應的 $B(i,j)$ 紀錄，即可以找出最佳的接合片段組合。

5 實驗結果

本論文所採用的大量語料庫為台北科技大學黃紹華教授[19]所提供的語料庫，共655個語音檔（句子），總共包含35085音節，錄音時間為9300秒，取樣頻率為20 KHz，16位元編碼。在採用這些語料檔案之前，我們有人工修正過的每個音節的子音和母音的位置標示，以配合後面的斷詞分析與韻律訓練。

首先，在類神經訓練方面，我們所使用的演算法為倒傳遞演算法(Back Propagation Algorithm)，使用訓練語料中的455個語音檔，測試語料為200個語音檔，所得到的實驗數據如表8所示。而計算誤差的公式為均方根誤差(Root Mean Square Error, RMSE)，算式如下：

$$RMSE=\sqrt{\frac{\sum_{t=1}^{N}(T(t)-S(t))^2}{N}}$$

N 為總共的個數，$1\le t\le N$ ，$T(t)$ 為訓練出來的結果，而 $S(t)$ 為正確的結果。

表 8 類神經訓練韻律參數的內外部測試結果

	整體語料庫資料	內部測試	外部測試
基頻軌跡平均	平均145.87H,標準差23.13Hz	RMSE 19.1Hz	RMSE 20.1Hz
聲母長度	平均56.3ms,標準差44.5ms	RMSE 16.5ms	RMSE 17.6ms
韻母長度	平均141.7ms,標準差52.2ms	RMSE 33.7ms	RMSE 38.2ms
停頓長度	平均16.8ms,標準差50.2ms	RMSE 38.4ms	RMSE 38.7ms
聲音能量	平均65.18dB,標準差6.23dB	RMSE 4.46dB	RMSE 5.04dB

我們從測試語料挑選某一個語句，並觀察其韻律參數的結果：

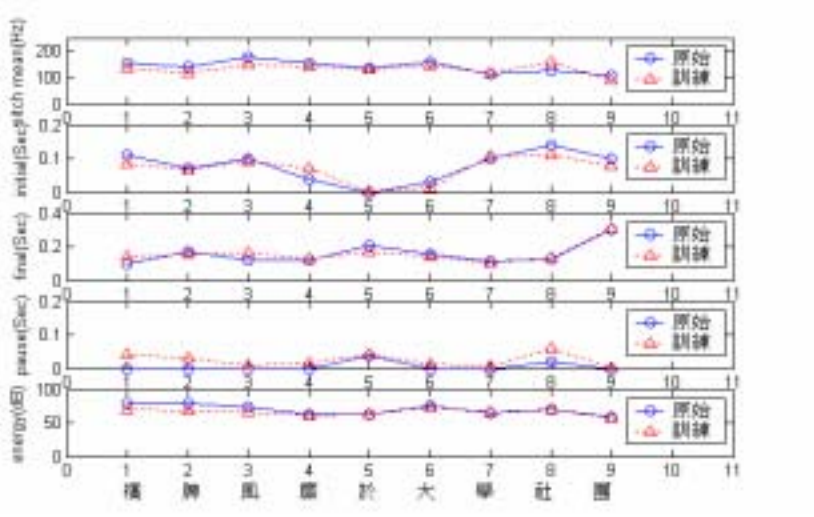


圖4 外部測試所得的韻律參數和原本韻律參數比較圖

實驗結果顯示，此類神經訓練法所得的韻律參數確實可以給予後面的語音合成器採用。此外，為了證實我們採用的反轉檔查找法能有效的找尋所需要的候選片段，我們以互相關比對法(Cross-correlation)[14] 來作為比較。測試語料為75任意句子，2005個音節，共1291個詞，平均每句26.7個音節、17.2個詞，執行時間如下：

表 9 演算法時間比較表

使用演算法	執行時間
互相關比對法	5955.47秒
反轉檔	65.96秒
互相關比較法＋最長連續詞數表	173.53秒
反轉檔＋最長連續詞數表	3.82秒

由此可知，使用反轉檔和最長連續詞數表，可以加速找到輸入文句和大量語料庫中對應的片段，和最慢的單純使用互相關比較法快了1500倍之多。

除了加速找尋所需的片段後，我們對每個詞保留前三名候選片段，然後再使用動態規劃演算法配合相近音查表，便可求出最佳的候選片段組合。為了證實此方法的確可行，我們採用MOS (Mean Opinion Score) [7]的評分方式，針對我們所提的方法以及隨意取出任一候選片段來實驗。另外，在大量語料庫的前提下，為了證實WSOLA調整音長方面的能力會優於PSOLA，我們在實驗過程中也加入這方面的比較。對於合成語句所花的時間，我們也作了統計，實驗平台為Pentium IV 1.6 GHz，執行環境為WINDOWS XP + MATLAB。測試語料為任意20句，參與合成語句的聽力測驗總共為10人，以下為其實驗的結果：

表 10 合成語句的MOS值與其計算時間統計

使用演算法	平均MOS	平均花費時間
任意取出候選片段組合＋PSOLA	2.6	6.3 秒
任意取出候選片段組合＋WSOLA	2.8	8.1 秒
最佳片段選取演算法＋PSOLA	3.3	7.8 秒
最佳片段選取演算法＋WSOLA	3.5	9.7秒

WSOLA在音長調整方面的合成品質確實勝過PSOLA，這是因為PSOLA音質的好壞取決於基週標位(Pitch Mark)的正確性，而大量語料的資料量通常很大，在實作上較難掌握每個音節的正確基週標位，而WSOLA並不需要基週標位的資訊。但是，若以系統效率而言，WSOLA所花費的時間則會較久，較不適合real time的系統設計。而最佳片段選取演算法的確大大的改善了原先使用隨意片段選取方法的音質。

6 實驗結果

在本論文中，我們已經實作了一個完整的TTS系統，此系統是以大量語料庫為基礎，並且配合我們所提出的反轉檔查找法與最佳片段選取演算法，使得此系統提升了在大量語料庫的搜尋速度之外，也保有較貼近自然的人聲，另外，也實驗證實了WSOLA對此系統的在合成音質方面的貢獻。

7 References

[1] Chu Min, Li Chun, Peng Hu, Chang Eric, "DOMAIN ADAPTATION FOR TTS SYSTEMS", *ICASSP* 2002

[2] F. Charpentier and Moulines, "Pitch-synchronous Waveform Processing Technique for Text-to-Speech Synthesis Using Diphones," European Conf. On Speech Communication and Technology, pp.13-19, Paris, 1989

[3] G.S. Ying and L.H. Jamieson and C.D. Michell, "A probabilistic approach to AMDF pitch detection", Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on Volume: 2 , 1996 , Page(s): 1201-1204 vol.2

[4] George E.B, Smith M.J.T., "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model", IEEE Transactions on Speech and Audio Processing, 1997

[5] Heo-Jin Byeon, Yung-Hwan Oh, "An event-driven f0 weighting for prosody control in a large corpus-based TTS system", Signal Processing Letters, IEEE 2004.

[6] I. Bulyko, M. Ostendorf and J. Bilmes. "Robust Splicing Costs and Efficient Search with BMM Models for Concatenative Speech Synthesis", in *Proceedings of ICASSP*, 1:461-464, 2002.

[7] ITU-T, Methods for Subjective Determination of Transmission Quality, 1996, Int. Telecommunication Unit.

[8] S. Haykin,"Neural Networks – A Comprehensive Foundation," Macmillan College Publishing Company, 1994

[9] Werner Verhelst and Mark Roelands"An Overlap-Add Technique Based on Waveform Similarity For High Quality Time-Scale Modification of Speech" In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 554--557, Minneapolis, USA, apr #"27--30 1993

[10] 周福強, "以語料庫為基礎之新一代中文文句翻語音合成技術", 國立臺灣大學電機工程學研究所博士論文, 1998.

[11] 唐大任,"中文斷詞器之研究", 國立交通大學電信工程系碩士論文, 2001.

[12] 朱怡霖, "中文斷詞與專有名詞辨識之研究", 國立臺灣大學資訊工程學研究所碩士論文, 2001.

[13] 王逸如,"對基週軌跡做向量量化之線性預估語音編碼 ", 國立交通大學電信研究所碩士論文, 1886.

[14] 謝明峰, "使用大量語料庫的中文語音合成系統實作", 國立清華大學資訊工程所碩士論文, 2004.

[15] 郭智超,"以音節為基礎之中文語音文件檢索系統的研究", 國立清華大學資訊應用所碩士論文, 2003.

[16] 鍾綸,"用於語音合成的中文斷詞分析 ", 國立清華大學資訊應用所碩士論文, 2004.

[17] 黃紹華, "中文文句翻語音系統中韻律訊息產生器之研究", 國立交通大學電子研究所博士論文,1995.

[18] http://rocling.iis.sinica.edu.tw/ROCLING/corpus98/wordlist.htm

[19] http://214lab.ee.ntut.edu.tw/

Improved prosody module in a Text-to-Speech system

Wen-Wei Liao and Jia-Lin Shen
Research Center, Delta Electronics, Inc.
wei.liau@delta.com.tw, lynn.shen@delta.com.tw

Abstract

The newly-developed prosody module of our text-to-speech (TTS) system is described in the paper. We present two main works on it's establishment and improvement. On the basis of potential factors influencing prosody parameters, inclusive of duration, pitch and intensity, the prosody model is built as groundwork of this module which is superior to the former rule-based one in generation of natural prosody. In addition, due to the current model's flaw in prediction of the pitch contour, we further employ an technique named "Soft Template Mark-up Language"(STEM-ML) to improve the smoothness of intonation which has the crucial influence on the naturalness of synthetic speech.

Results of the evaluation indicate that the new prosody model is precise enough to predict reliable prosody parameters' values and with the STEM-ML technique, the prosody module can further yield 14.75% reduction in the root mean square (RMS) error of the predicted pitch contour.

1. Introduction

In consideration of severe limitation in the resource afforded by some applications in need of speech response, we choose to develop one storage-saving TTS system which has functioned successfully in our spoken dialogue system. Accordingly, the acoustic inventory used in our system is simply composed of about four hundred base syllable units whose duration and pitch contour will be modified with the algorithm called Pitch-Synchronous Overlap-Add (PSOLA) [1][12] in the synthesizing phrase.

In order to produce natural-sounding synthetic speech, the generation of prosody plays a key role and is a difficult issue yet. Outperforming rule-based method [13][14] which was employed in our system previously, the newly-built statistical model based on sum-of-products approach with key factors affecting prosody [7][8][9][10][11] can predict more accurate values of prosody parameters. And in general, the intonation which is characterized by the pitch contour seems more crucial to the naturalness and intelligibility of synthesized speech in comparison with other prosody elements such as duration, intensity etc [6]. Nevertheless, the pitch contour generated by our current prosody model is still short of smoothness. As a result, we further concentrate our work on this problem. Based on the F0 (fundamental frequency) mean value predicted by the current prosody model, an technique named STEM-ML [2][3][4][5] is adopted to overcome this shortcoming. In the evaluation phrase, we prove that this technique can help to reduce the difference between the predicted and observed pitch contours, which means that a more natural intonation is achieved.

The paper is organized as follows. In the chapter 2, we present the prosody modeling in our system, The chapter 3 reports STEM-ML technique and the result of implementation. The conclusion is described in the chapter 4.

2. Prosody modeling

In general, prosody mainly consists of duration, pitch, intensity of the spoken unit which is one syllable in terms of Mandarin. Besides, the break between units is one of it's important elements as well. Therefore, one utterance's prosody can be regarded as the elaborate composition of these four perceivable characteristics. And the variation in prosody stem from a lot of factors in different dimensions which can be observed in the real speech corpus such as the syllable's position in the sentence, lexical tone even the speaker's emotion and so on. Furthermore the complex interactions between factors further lead to another difficulty in designing the prosody model. As a result, in addition to inferring the reliable factors influencing the prosody, to model the interactions between factors intelligently is also a challenge in this work. .

2.1 Modeling

2.1.1 Base model and sub-models

The potential factors affect one characteristic simultaneously and have additive, multiplicative or repulsive interactions . Thus, it's troublesome to derive their eventual combined effect on the characteristic. However, for the purpose of assuring that the basically reasonable value for the characteristic can be preserved, one major factor in possession of dominant influence are elected to build the base model while the remaining minor factors take charge to constitute sub-models. In other words, under this framework, the base model provides fundamental value for the characteristic and sub-models act on this base value (BV for short) through the mechanism modeling their interaction to obtain the ultimate characteristic value (CV for short).

2.1.2 Ratio of characteristic value to base value (RCB)

In order that this concept of modeling can be put into practice concretely, the training sample for sub-models, namely the CV of each syllable has to be normalized by it's corresponding BV beforehand. Thus, pre-processed CV is computed as follows.

$$RCB = \frac{CV}{BV} \tag{1}$$

2.1.3 Mechanism

In brief, the ultimate objective of the mechanism devised here is to make combined effect of minor factors quantized to one RCB value used as the multiplier of the BV. The interactions of minor factors are modeled by the approach of sum-of-products and the predicted CV is computed as follows.

$$\begin{aligned} \hat{CV} &= RCB_{comb} \times B_i \\ RCB_{comb} &= \sum_{i}^{SMN} \sum_{j}^{i} C_{ij} S_i^{m_{ij}} S_j^{n_{ij}} \end{aligned} \tag{2}$$

where

Bi is the parameter of the base model for the characteristic i and

SMN is the numbers of sub-models for the characteristic i and

Si is the parameter of the sub-model i and

Cij is a coefficient associating the sub-model i and sub-model j and

mij and nij represent the stress of sub-model i and sub-model j respectively.

2.1.4 Factors

We infer seven potential factors crucial to the characteristics in prosody. Those are listed and described briefly as below.

- **Base syllable (BS)**

  408 identities

- **Lexical tone (LT)**

  4 lexical tones and one neutral tone

- **Left and right context tones (LRCT)**

  175 levels: 25(bi-tone) + 125(tri-tone)

- **The syllable's position in the word and the syllable number of one word (SInW)**

  15 levels: 1+2+3+4+5 (longest word length)

- **The word's position in the phrase (WInP)**

  4 levels: $WInP = \frac{WordIndex}{WordNumberOfPhrase} \times 4$

- **Right context break (RCBk)**

  4 levels: inter-syllable pause, inter-word pause, comma, period

- **Right context initial (RCIt)**

  32 identities

Accordingly., four kinds of base models and seven kinds of sub-models will be established in light of these factors.

2.2 Estimation

2.2.1 Corpus

Recorded by a single female speaker, the speech corpus contains 3657 sentences (70000 syllables;about 7 hours) with moderate intonation and constant speaking rate. In terms of linguistics ,the properly-designed one has enough coverage to tackle diverse variability of prosody. Among these sentences, around 3200 ones are used as training data and the rest of them are reversed for the purpose of evaluation. The syllable boundaries in the waveform are further calibrated manually after aligned by the automatic speech recognizer.

2.2.2 Objective function

The distortion rate (DR) is defined to measure the precision of predicted value.

$$DR = \left| \frac{O - P}{O} \right| \quad (3)$$

where

O is the occurrence's CV and

P is the predicted CV.

Accordingly, the objective function is defined as average DRs of all occurrences in the training data.

$$O = \frac{1}{N}\sum_{i} DR_i \tag{4}$$

where N is the number of training samples.

2.2.3 Approach

- **Model**

Both base models and sub-models have only one parameter. The parameters of base models and sub-models are calculated as the average of observed occurrences's CVs and RCBs which correspond to them in the training corpus respectively.

$$\mu = \frac{1}{oN}\sum_{i}^{oN} o_i \tag{5}$$

where

μ *is the parameter of the model and*

oi is observed occurrence whose value is either RCB or CV depending on whether the model is a sub-model or base model and

oN is the number of occurrences.

- **Coefficients and Stress**

Firstly, the initial values of coefficients and stress are calculated by means of linear least square error and given value 1 respectively. And furthermore beginning with the initial values, Levenberg-Marquardt algorithm [15][16] with numerical differentiation is employed to find the optimal values of these parameters with the goal of minimizing the objective function O defined in (4).

2.3 Characteristic model

In this section ,the characteristic models, inclusive of duration, pitch and intensity are discussed in terms of the related factors and precision. And as for the break characteristic, we straightforwardly give each type of break an empirical length instead of building the model.

2.3.1 Duration

This characteristic means the time for which one syllable endures in the utterance. Since the boundaries between syllables are demarcated precisely by hand in our speech corpus, it is straightforward to calculate the syllable's duration.

- **Factors**

  Major *BS*

  Minor *1. LRCT 2. SInW 3. WInP 4. RCBk 5. RCIt*

- **Speaking rate**

Each syllable's duration in the corpus needs to be normalized by the utterance's speaking rate (SR)

which is estimated as:

$$SR = \frac{1}{SylN} \sum_{i}^{SylN} \frac{D_i}{\overline{D}_{BSi}} \tag{6}$$

where

D_i *is duration of one syllable (named Si),* $\overline{D}_{BSi}$ *is average duration of base syllable corresponding to Si in the corpus and* $SylN$ *is the number of syllables in one utterance.*

2.3.2 Pitch

Pitch here means the one syllable‘s pitch contour which is depicted with F0 (fundamental frequency) computed at a constant frame rate. In our task, this characteristic is discussed in two separate aspects, namely the pitch contour ‘s F0 mean (FM for short) and F0 shape. The former can leave the each syllable’s pitch contour in a proper level and the later considerably concerns it’s smoothness.

In this chapter, we only concentrate discussion on the F0 mean. In the other hand, one technique named STEM-ML is adopted to deal with F0 shape. This work will be reported in next chapter.

- **Factors**

  Major *LT*

  Minor *1. BS 2. LRCT 3. SinW 4. WinP 5. RCBk*

- **FM rate**

Each syllable’s FM in the corpus needs to be normalized by the utterance’s FM rate (FMR) which is estimated as:

$$FMR = \frac{1}{SylN} \sum_{i}^{SylN} \frac{F_i}{\overline{F}_{Tonei}} \tag{7}$$

where

F_i *is FM of one syllable,* $\overline{F}_{Tonei}$ *is average FM of Tonei in corpus and* $SylN$ *is syllable number in one utterance.*

2.3.3 Intensity

This characteristic means one syllable’s volume in one utterance. We measure one syllable’s intensity with it’s power. The power can be estimated as below.

$$Power = \log 10 \left(\frac{\sum_{i} X_i^2}{N} \right) \tag{8}$$

where

Xi and N are the sample value and number of samples respectively.

- **Factors**

  Major *LT*

  Minor *1. BS 2. LRCT 3. SInW 4. WInP 5. RCBk*

■ **Power rate**

Each syllable's power in the corpus needs to be normalized by the utterance's power rate (PR) which is estimated as:

$$PR = \frac{1}{SylN} \sum_{i}^{SylN} \frac{P_i}{\overline{P}_{Tonei}} \tag{9}$$

where

P_i *is power of one syllable (named Si),* $\overline{P}_{Tonei}$ *is average power of Tonei in the corpus and* $SylN$ *is syllable number in one utterance.*

2.4 Evaluation

The evaluation set consists of 300 stentences, exclusive of the sentence in the training set and the precision of the characteristic models are evaluated with DR defined in (3). The results are shown in the Table 1.

Model	Precision
Duration	11.35%
Pitch	5.6%
Intensity	1.98%

Table 1. The preciosion of characteristic models.

3. Soft Template Mark-up Language (STEM-ML)

The prosody model developed in the previous chapter establishes the groundwork for the prosody module of our TTS system. However, since it merely aims at assuring the accuracy of F0 mean without putting emphasis on the F0 shape, the predicted pitch contour lacks smoothness. For the sake of this drawback , we proceed to employ an model devised by Kochanski, G. P. et al. and called STEM-ML that is abbreviated from "Soft Template Mark-up Language".

It is a tagging system which computes the pitch contour in light of a set of tags serving to interpret the variation in the pitch contour more humanly. In order to make the artificial pitch contour closer to the real one, the mechanism of model has to comply with the constraints actually existing in the human uttering process. Thus, each tag concretely takes effect by imposing constraints on prediction of the pitch curve.

As a result, the pitch curve is eventually generated by the model on condition that those constraints come to a compromise. In fact, such compromise can be considered to be the result of tradeoff between two events with reversal interaction, namely effort and error. The effort term stands for physiological energy consumed in the uttering processing and the error one means the communication error rate caused under the current effort. Obviously, they behave contrary to each other. With more effort, the uttering can achieve more accurate expression on words while the error results from little effort spent on uttering. In conclusion, the model can be also thought to predict the pitch curve with the goal of minimizing the sum of effort and error caused in the uttering process.

3.1 Model

3.1.1 Soft templates

Soft templates consists of pitch contours of four lexical tones (tone1,tone2,tone3 tone4) and the neutral tone (graphed in Figure 1).Since the syllable's tone shape varies considerably due to the affection from syllables nearby, five templates aren't apparently equal to express such variability . However, the adjective, "Soft" significantly implies that their shapes are allowed to change properly (see Figure 2). Consequently, these templates with the elastic property can form smoother pitch contour.

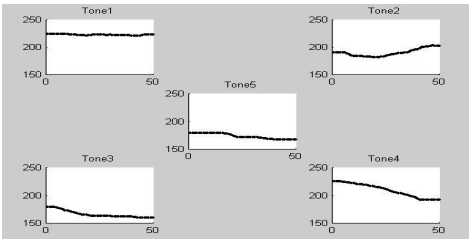


Fig 1. 5 tone templates.

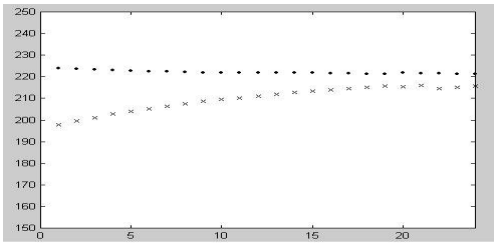


Fig2. A example of how one syllable is effected by it's neighbor. Succeeding to Tone3, the original shape of Tone1 template (dot line) is bended under control of the model and turns out to be the one (cross line) with tilt in the front part.

3.1.2 Tags

The tags function as adjustable parameters of the model. Each kind of tag governs the pitch curve's variability in one certain dimension. For instance, the tag ***smooth*** determines the permissible velocity of change in pitch values and the priority over one pitch curve's shape and F0 mean is dependent on the tag ***syllable-type*** **.** Thus, the tags have the critical influence on the generated pitch curve's look and should be given proper values so that the one can has good quality. The estimation of tags will be reported in the section 3.3. 10 kinds of tags in total are used in our work as listed below.

max, min, base, range, add, slope, smooth, pdroop, adroop syllable-type, syllable-strength

Moreover, to account for the more detailed pitch curve's variation inside one word, the tag ***syllable-strength*** is specially given a distinct value depending on the syllable's position inside the word. As the case for the sub-model **SInW**, this actually leads to 15 kinds of ***syllable-strength*** tags considered in the model.

3.2 Calculation of pitch contour

Based on the templates and tags, the process of calculating the pitch curve mainly includes two steps.

Step1

The first step purposes to prepare the plain templates assembling a prototype of the pitch curve.

1. Select the templates according to each syllable's tone among five basic templates as mentioned above.

2. The templates have to be modified to conform to the desired duration and F0 mean predicted by the prosody model.

Step2

In this step, the tags start to be applied in the calculation along with ready templates. The constraints on generation of the pitch curve are realized by translating the tags to a number of conditional equations with pitch instants (F0) as unknown variables to be solved. One tag can brings in one equation or one group of equations. For example, the ***slope*** tag which controls the pitch's increasing or decreasing rate in the phrase level yields the equation $Pt+1 - Pt = S$ where P and S are the pitch variable and the ***slope*** tag's value respectively. These joint conditional equations can be written as the form $Ax = b$ where A is matrix with rows composed of the coefficients in the left-hand side of all equations and x is a vector containing the unknown variables and the b is a vector with elements consisting of the right-hand side of all equations .Consequently, the pitch values of the curve are the solution of the algebraic problem $Ax = b$.

Furthermore, the calculation proceeds in the order of phrase level and the syllable level. Riding on the phrase's pitch curve solved firstly, the syllable's one is calculated . The process in the phrase level aims at deciding the trend of the whole resultant pitch curve which is finally obtained in the syllable level. Step2 is illustrated in Figure 3.

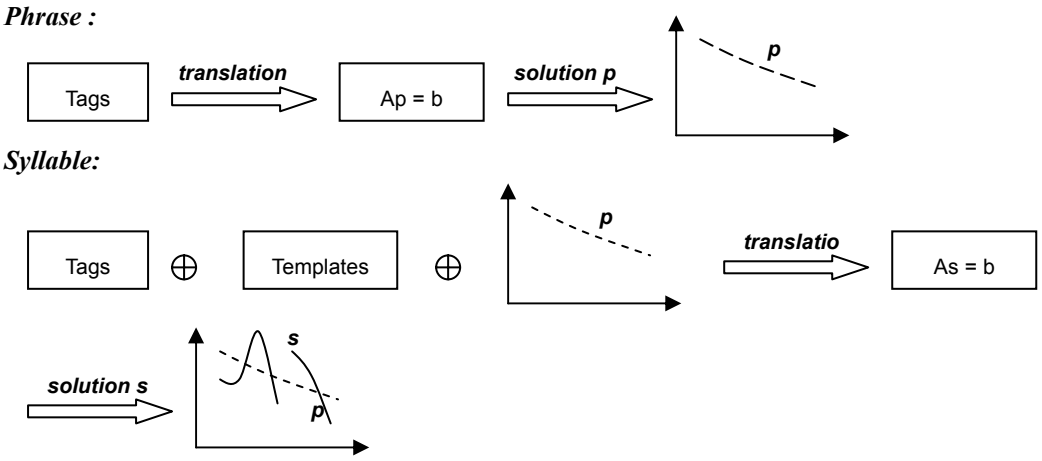


Figure 3. The procedure for calculating pitch contour which is carried out in the order of the phrase and syllable levels .

A real case for the syllable's pitch curve (dot line) and phrase's one (dash line) generated by the model is plotted in Figure 4 .

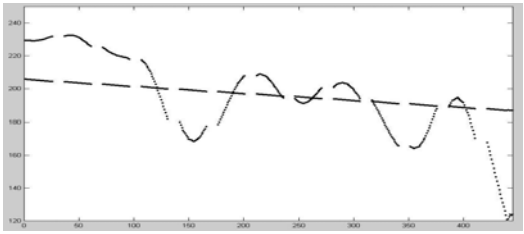


Fig 4. A example of the pitch contour generated by the model.

3.3 Estimation of tags

3.3.1 Approach

We estimate the tags by data fitting with the objective to minimize root mean square (RMS) error of the predicted F0 in comparison with the observed F0 in the data. The development data set composed of 300 sentences is designed to cover enough occurrences for each kind of tag and templates. Similarly, Levenberg-Marquardt algorithm with numerical differentiation is employed in this task. In addition, the number of pitch samples per syllable in the data is normalized to a constant and the syllable's un-voiced position is excluded.

3.3.2 Results

The process of minimization ends in RMS error that is equal to 16.16 (Hz) One example of fitting results is shown in Figure5.

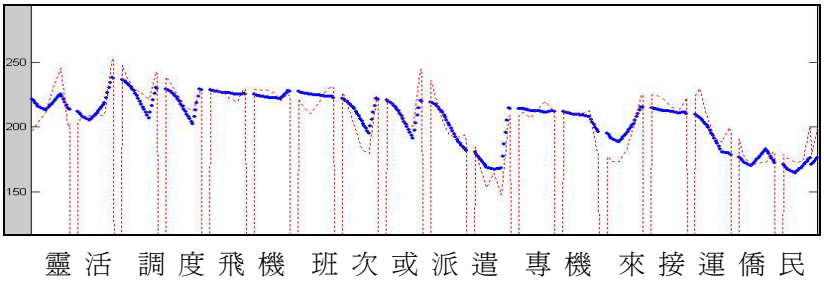


Fig5. A example of one utterance's simulated pitch curve (dot line) along with the real one (dash line) in the data-fitting result.

3.4 Evaluation

The evaluation data set is the same to one in the chapter 2 and the prosody model is used as the baseline of this task. In the baseline, the templates are unvaried in the shape but shifted to have the F0 mean predicted by the prosody model. The accuracy of the pitch contour generated by the model is measured by the RMS error of predicted F0 .The result is shown in the Table 2.

Prosody model (baseline)	19.46 (Hz)
Prosody model + STEM-ML	16.59 (Hz)

Table 2. The RMS F0 error of the pitch contour generated by the prosody model and prosody model + STEM-ML.

The result indicates that based on the prosody model, this technique can further reduce 14.75% RMS error of F0 in the predicted pitch contour.

4. Conclusions

In this paper, we successively report two works on the development of the prosody module in our TTS system, Firstly, the prosody model based on the framework of base models and sub-models and sum-of-products approach has been proven to have the capability of predicting reliable prosody parameters' values. Furthermore, the employment of the STEM-ML technique further bring in the improvement in the smoothness of the intonation which the prosody model originally lacks

In order to raise the accuracy of the prosody model, the refinement of the mechanism in the modeling should be necessary . Besides, we consider expanding the types of STEM-ML tags defined in

our system to generate more natural and lively intonation.

References

[1] Moulines, E. and Charpentier, F. Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. *Speech Communication* 9, 453-467, 1990.

[2] Kochanski, G. P. and Shih, C., "Prosody modeling with soft templates," accepted by *Speech Communication.*

[3] Kochanski, G. P. and Shih, C., "Automatic modeling of Chinese intonation in continuous," in Proceedings of EUROSPEECH 2001, pp.911-914.

[4] Grep P. Kochanski and Chilin Shih, "Stem-ml: Language independent prosody description," in *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, 2000.

[5] Chilin Shih and Greg P. Kochanski, "Chinese tone modeling with stem-ml," in *ICSLP*, Beijing, China, 2000

[6] Plumpe, M., Meredith, S. Which is more Important in a concatenative Text To Speech System – Pitch, Duration or Spectral Discontinuity ?, *Proceedings of the third ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan, Autralia, Nov. 25-29, 1998

[7] Van Santen, J. P. H. Assignment of segmental duration in text-to-speech synthesis. *Computer, Speech and Language*, 8, 1994.

[8] *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach*, Richard Sproat, editor, Kluwer Academic Publishers, 1998.

[9] J. van Santen, "Prosodic modeling in text-to-Speech synthesis", *Proceedings of EuroSpeech'97*, KN-19,Rhodes 1997.

[10] Febrer, A.; Padrell, J.; & Bonafonte, A. 1998. Modeling phone duration: Application to Catalan TTS. *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves, Australia, 43-46.

[12] K.M. Law and Tan Lee, "Cantonese text-to-speech synthesis using sub-syllable units", in *Proceedings of the 7th European Conference on on Speech Communication and Technology*, Vol.2, pp.991 - 994, Aalborg, Denmark, September 2001.

[13] L.S.Lee, C.Y. Tseng, and M. Ouh-Young, "The synthesis rules in a chinese text-to-speech system", *IEEE trans. Acoust., speech, signal Processing,* Vol. 37, pp. 1309-1320, 1989.

[14] 許文龍,"使用時間比例基週波形內差之國語語音合成器",國立台灣科技大學電機工程研究所,民國 84 年.

[15] K Levenberg, "A method for the solution of certain problems in least sqrares," *Quart. Applied Math*., vol. 2, pp. 164-168, 1944.

[16] D. Marquardt, "A algorithm for least-squares estimation of non-linear parameters," *SIAM J. Applied Math*, vol. 11, pp.431-441, 1963.

The Construction and Testing of a Mandarin Emotional Speech Database

Tsang-Long Pao, Yu-Te Chen, Jhih-Jheng Lu, Jun-Heng Yeh

Department of Computer Science and Engineering, Tatung University, Taipei

E-mail: tlpao@ttu.edu.tw, d890600S@mail.ttu.edu.tw, g9106001@ms2.ttu.edu.tw

Abstract. Researches in speech synthesis and speech analysis are underpinned by the databases they used. The performance of an emotion classifier relies heavily on the qualities of the training and testing data. A good database can make researches in these fields achieving better results. Hearing-impaired people are poor in presenting their emotions in speech. We want develop a computer-assisted speech training system that can help to teach them to present their emotions similar to normal people. In this paper, we present a way to build a Mandarin emotional database, including the process of collecting data, arranging data, clips naming rules, and a listening test. Then we construct a computer-assisted speech training system to help in teaching the hearing-impaired people presenting their emotion in their speech correctly by analyzing the emotion in their speech and those in the database using KNN and M-KNN techniques.

Keywords: Emotional speech database, Emotion evaluation, Emotion Radar chart, M-KNN

1. Introduction

The performance of an emotion classifier relies heavily on the quality of emotional speech data and the similarity of it to real world samples. As mentioned in [1], there are three different categories of emotional speech: acted speech, elicited speech, and spontaneous speech. In this section we will describe the ways to obtain these three kinds of speech data.

In acted speech recording, actors are invited to record utterances, where each utterance needs to be spoken with multiple emotions. The method is adopted by most researches because it can get large amount of data in a short time and the data is undistorted. For general use, we should invite speakers with different age, gender, even with different social or educational background if possible. And if we hope the emotion in the data to be more obvious, we could invite professional actors.

We can also collect the clips that contain utterances with specific emotion in a film. We must avoid the background noise including music, surrounding noise, and other people's voice. This method takes quite a lot of time in viewing the content of films.

In elicited speech recording, the Wizard-of-Oz (WOZ) is used. The WOZ means using a program that interacts with the speaker and drives him into a specific emotion situation and then records his voice. This method needs a good program that can induce the participator to say something in our expected emotion state. So how to design such a program may not be easy.

In spontaneous speech recording, the real-world utterances that express emotions are recorded. Although data got from this method has the best naturalness, it is the most difficult because we need to follow the speaker. When he or she is in some emotion state, his voice is recorded immediately. This method will face many problems. For examples, we must hide our recording device in order to make the speaker without any pressure to present his real emotion. Furthermore, we also cannot assure the environment is quiet. Generally speaking, the method is generally infeasible.

2. Mandarin Emotional Speech Database

In our research, five emotions are investigated: anger, happiness, sadness, boredom, and neutral. We invite 18 males and 16 females to simulate five emotions. A prompting text with 20 different sentences is designed. The length of each sentence is from one word to six words the sentences are meaningful so speakers could easily simulate them with emotions. During the recordings process, speakers are asked to try their best to simulate each

emotion. And speakers can simulate one sentence many times until they are satisfied what they simulated. Finally, we obtained 3,400 emotional speech sentences. After the recording procedure, a listening test is held to evaluate these recorded sentences.

It is very important to use speech with unambiguous emotional content for further analysis. This can be guaranteed by a listening test [2], in which listeners evaluate the emotional content of a recorded sentence. Moreover, we can understand the performance of human in emotion recognition.

We perform the listening test in a three-pass procedure. First, we delete the speech data that is very hard to identify its emotional content. After this process, 1,178 sentences are remained. Then, the remaining sentences are evaluated by three speakers. The sentences with the same agreement are remained. After the stage, 839 sentences are remained. Finally, we invite 10 people whom did not have their speech data in the 839 sentences to take part the final listening test.

The results of the listen test are shown in Fig. 1. We can see the recognition results of the 10 evaluators in the figure and the confusion matrix in Table 1. The results reveal that people are good in recognizing anger (89.56%), sadness (82.76%), and neutral state (83.51%), but are less confident for happiness (73.22%), and boredom (75.16%)

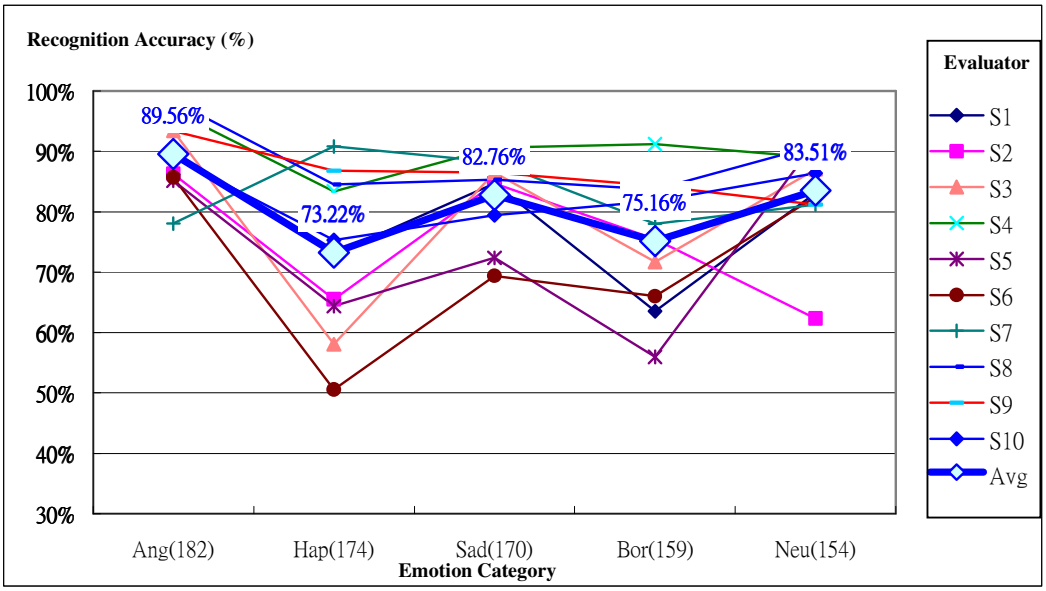


Fig. 1. Recognition results of 10 evaluators.

Table 1: Confusion Matrix of Human Performance.

	Anger	Happiness	Sadness	Boredom	Neutral	None of above
Anger	89.56%	4.29%	0.88%	0.77%	3.52%	0.99%
Happiness	6.67%	73.22%	3.28%	2.36%	13.56%	0.92%
Sadness	2.94%	1.00%	82.76%	9.29%	3.29%	0.71%
Boredom	1.26%	0.44%	8.62%	75.16%	13.65%	0.88%
Neutral	1.69%	0.91%	1.56%	12.27%	83.51%	0.06%

Table 1 shows the human performance confusion matrix. The rows and the columns represent simulated and evaluated categories, respectively. For example, first row says that 89.56% of utterances that were portrayed as angry were evaluated as angry, 4.29% as happy, 0.88% as sad, 0.77% as bored, 3.25% as neutral, and 0.99% if none of above. We can see that the most easily recognizable category is anger (89.56%) and the poorest recognizable category is happiness (73.22%). And we can find that human sometimes are confusing in differentiating anger from happiness, and boredom from neutral.

Table 2 shows the statistics of 10 evaluators for each emotion category. We can see that the variance for anger and sadness are less than for the other emotions. It means that human are better in understanding how to recognize anger and sadness than other emotions.

Figure 2 shows the percentage of remained sentences with different lengths for each emotion. We can see that the shortest sentence (only single word) is least remained in most emotions, especially in neutral. It means that we should avoid too short sentences when we make the prompting text in the future because emotions are hard to be recognized by human if the sentence is too short.

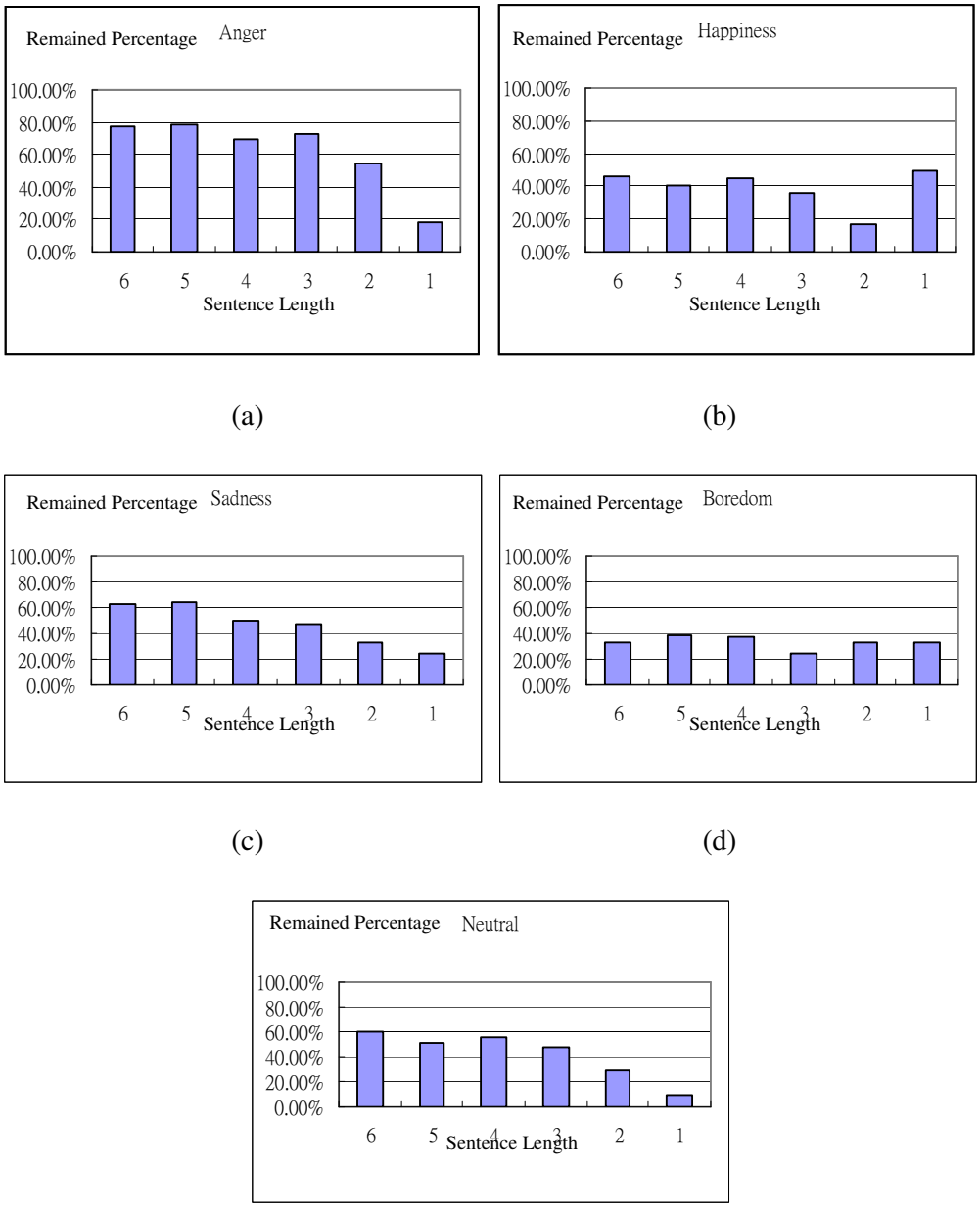


Fig. 2. Percentages of remained sentences with different lengths

Table 2: Evaluator's statistics in each category.

Category	**Mean**	**S.T.D**	**Median**	**Maximum**	**Minimum**
Anger	89.56%	6.11%	89.29%	98.35%	78.02%
Happiness	73.22%	13.36%	74.14%	90.80%	50.57%
Sadness	82.76%	6.92%	85.00%	90.59%	69.41%
Boredom	75.16%	10.87%	76.73%	91.19%	55.97%
Neutral	83.51%	8.37%	84.74%	91.56%	62.34%

For further analysis, we only need the speech data that can be recognized by most human. So we divide speech data into different dataset by their recognition accuracy. We will refer to these data sets as D80, D90, D100, which stand for recognition accuracy of at least 80%, 90%, and 100%, respectively, as listed in Table 3.

Table 3: Datasets and their sizes

Data set	D80	D90	D100
Size (number of sentences)	570	473	283

3. Emotion Recognition and Emotion Evaluation

3.1 Emotion Recognition

Figure 3 shows the block diagram of our emotion recognition system. We calculate the MFCC as the emotional feature from each input data [3]. Then, the speech is classified by pattern classification method (K-NN)[4]. K-NN will find the k neighbors nearest to the new sample from the training space based on some suitable similarity or distance measure methods.

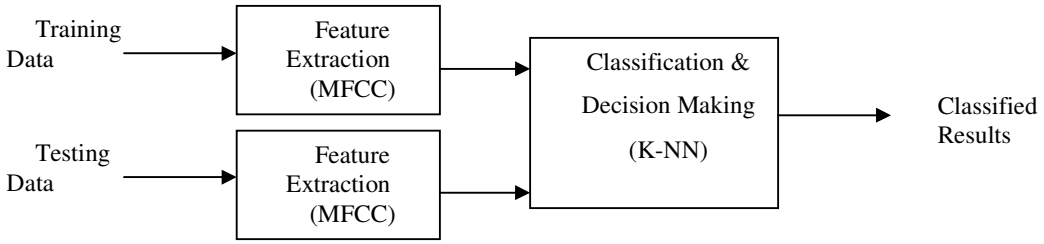


Fig. 3. Block diagram of emotion recognition

Figure 4 shows the recognition accuracy with different values of K of the K-NN classifier where we choose 70% and 30% of the speech data for training and testing, respectively. We notice that while the value of K increased, the recognition accuracy of happiness begins to drop gradually. The higher average recognition accuracy exist in K=1 and K=3. Therefore, we choose K=1 for the K-NN classifier.

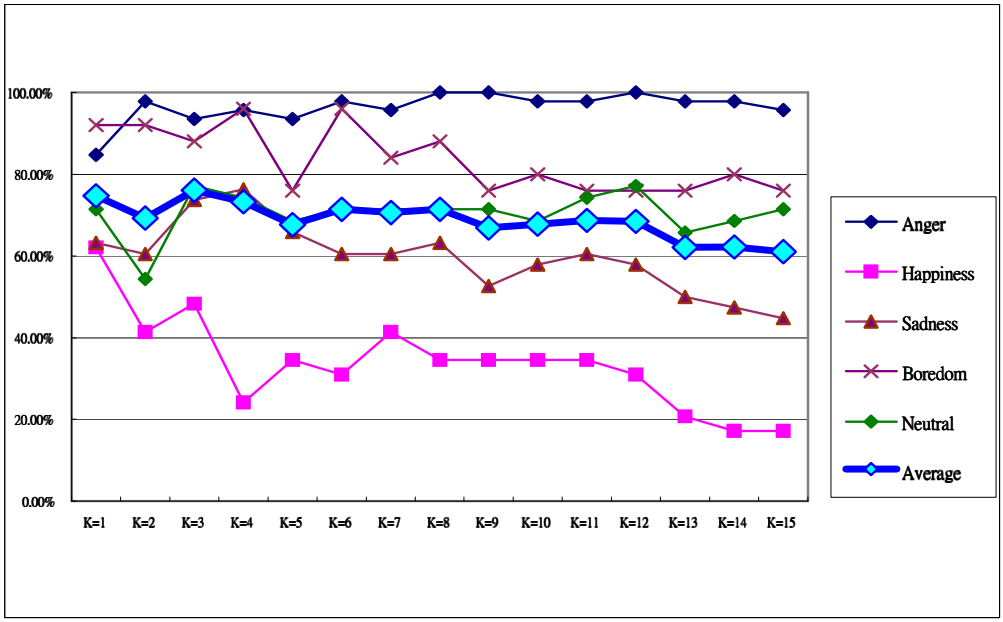


Fig. 4. Recognition rate of K=1 to 15 using KNN classfier.

Table 4 shows the confusion matrix of our recognition system that the value of K is set to one. The rows and the columns represent original and recognized emotion categories, respectively. For example, first row says that 39 sentences that belong to angry were recognized as angry, 6 sentences as happy, 1 sentence as sad, and 0 for the rest. And the recognition accuracy of anger is 84.78%. We can see that our system do better in recognizing anger and boredom. The mean recognition rate is 74.69%.

Table 4 Confusion matrix of our system

	Anger	**Happiness**	**Sadness**	**Boredom**	**Neutral**	**Recognition rate**
Anger	39	6	1	0	0	84.78%
Happiness	7	18	1	0	3	62.07%
Sadness	4	3	24	4	3	63.16%
Boredom	0	0	0	23	2	92.00%
Neutral	1	2	1	6	25	71.43%

3.2 Emotion Evaluation

We use K-NN to classify input testing data, so we modify the K-NN method to be our evaluator in the emotion evaluation stage. We called the method “M-KNN”. Figure 5 shows the block diagram of the evaluation system.

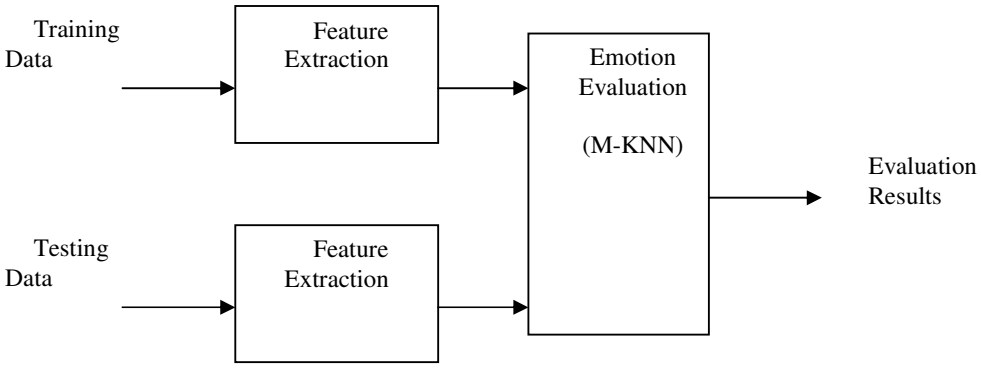


Fig. 5. Block diagram of emotion evaluation

3.3 Emotion Radar Chart

An emotion radar chart is a chart with multi-axes. Each of the axes stands for one category of emotion. In our system, it just looks like a regular pentagon as shown in Fig. 6. We need to measure the distance of a testing data to each category to plot its radar chart. Thus a modified version of KNN (M-KNN) is needed.

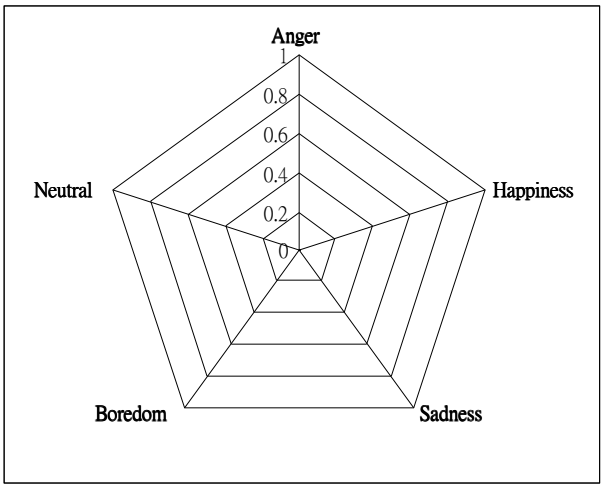


Fig. 6. Emotion radar chart

The M-KNN is based on the KNN technique. It calculates the K-nearest neighbors' distances in each class to the input testing data. We set the value of K to 1 corresponding to the K in emotion recognition. Figure 7 shows the M-1NN method.

After the calculation of M-KNN, we will get five distances from five emotion categories. We take inverse of each distance, and base on these inverses of distances to plot a radar chart. For example, Table 5 list the calculation result of an input testing speech with angry emotion. And Fig. 8 shows the radar chart according to Table 5.

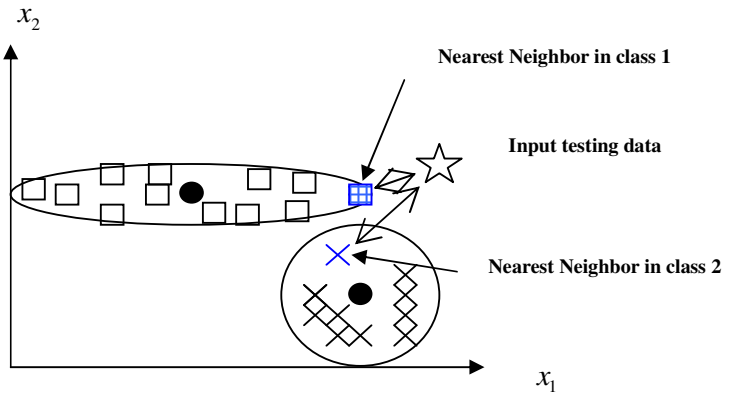


Fig. 7. M-1NN: The computation of the distance to the nearest neighbor in each class

Table 5. Distance measured by M-KNN

Emotion	Anger	Happiness	Sadness	Boredom	Neutral
Distance	12.325	19.31	23.14	27.868	22.83

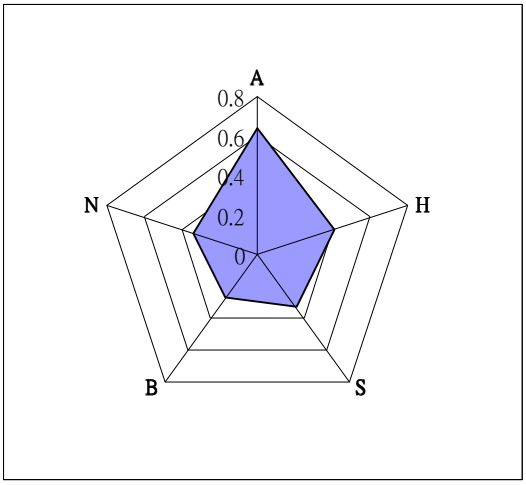


Fig. 8. Emotion radar chart of a speech with anger emotion.

From Fig. 8, we can find that this input data is closed to anger. It means the intensity of anger is greater than the other emotions. An unambiguous emotion should close to one emotion and far away the other emotions similar to the one shown in Fig. 9.

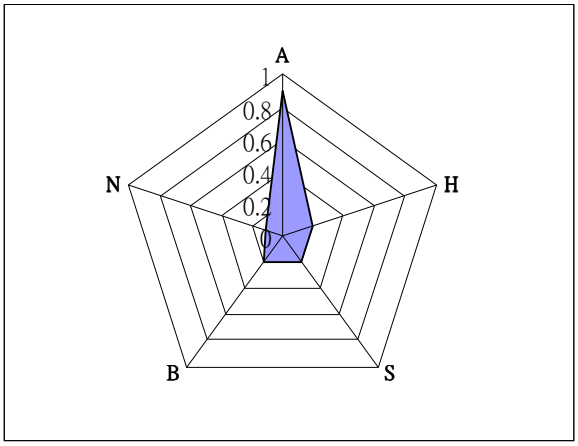


Fig. 9. Emotion radar chart of speech with unambiguous emotion

3.4 System Interface

Figure 10 shows the interface of our system. A user can record his or her voice by pressing the Record button. The user can hear the speech within the selected range by pressing the Play button. Finally, the user can see the evaluation result of his or her speech by pressing the Eva button.

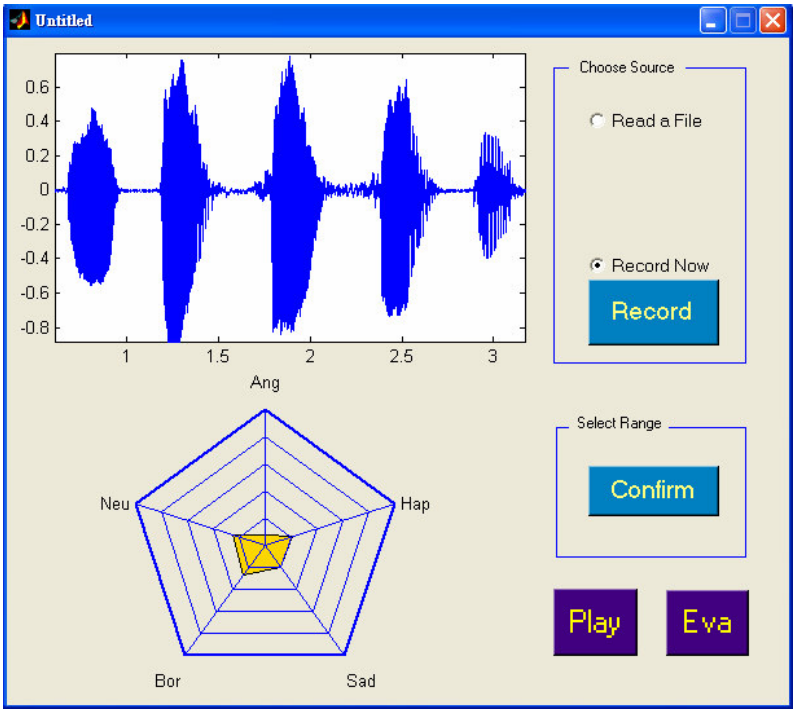


Fig. 10 System interface

When teaching the hearing-impaired people, teacher could ask him or her to say something with certain emotion. Hearing-impaired people can easily understand what emotion is presented by the emotion radar chart. Guiding by the teacher, they can try and improve the naturalness of their emotion expression through the use of the emotion radar chart.

4. Conclusions

In this paper, we build a Mandarin emotional speech database for research in this field. We also propose an emotion recognition and evaluation system. For hearing-impaired people, it could provide an easier way to learn how to speak more naturally.

We will continue to get more speech data into our database, and improve the recognition accuracy of the emotion recognition system. We also want to make the emotion evaluation more effectively. Furthermore, friendlier interface to hearing-impaired people is needed to be designed.

5. Acknowledge

A part of this research is sponsored by NSC 93-2213-E-036-023.

6. Reference

[1] Raquel Tato, Rocio Santos, Ralf Kompe, "Emotional Space Improves Emotion Recognition", *Man Machine Interface Lab, Advance Technology Center Struttgart*, Sony International (Europe) GmbH.

[2] Inger Samso Engberg, Anya Varnich Hansen , "Documentation of the Danish Emotional Speech Database", *Department of Communication Technology Institute of Electronic Systems,* Aalborg University, Sep. 1996

[3] Bo-Syong Juang, "Automated Recognition of Emotion in Mandarin", *Department of Engineering Science, National Cheng Kung University Master Thesis*, Jun 2002

[4] Maleq Khan, Qin Ding, William Perrizo, "k-Nearest Neighbor Classification on Spatial Data Streams Using P-Trees", *Computer Science Department, North Dakota State University*.

Detecting Emotions in Mandarin Speech

Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh and Jhih-Jheng Lu

Department of Computer Science and Engineering, Tatung University, Taipei

tlpao@ttu.edu.tw, d8906005@mail.ttu.edu.tw, g9106004@ms2.ttu.edu.tw

Abstract. In this paper, a Mandarin speech based emotion classification method is presented. Five primary human emotions including anger, boredom, happiness, neutral and sadness are investigated. For speech emotion recognition, we select 16 LPC coefficients, 12 LPCC components, 16 LFPC components, 16 PLP coefficients, 20 MFCC components and jitter as the basic features to form the feature vector. Two text-dependent and speaker-independent corpora are employed. The recognizer presented in this paper is based on three recognition techniques: LDA, K-NN, and HMMs. Results show that the selected features are robust and effective in the emotion recognition at the valence degree in both corpora. For the LDA emotion recognition, the highest accuracy of 79.9% is obtained. For the K-NN emotion recognition, the highest accuracy of 84.2% is obtained. And for the HMMs emotion recognition, the highest accuracy of 88.7% is achieved.

1 Introduction

Various opinions of emotions proposed by more than 100 scholars are summarized in a classical article [1]. Research on the cognitive component focuses on understanding the environmental and attended situations that gives rise to emotions; research on the physical components emphasizes the physiological response that co-occurs with an emotion or rapidly follows it. In short, emotions can be considered as communications to oneself and others [1]. Emotions consist of behaviors, physiologic changes and subjective experience as evoked by individual's thoughts, socio-cultures and so on.

Emotions are traditionally classified into two main categories: primary (basic) and secondary (derived) emotions [2]. Primary or basic emotions generally could be experienced by all social mammals (e.g. humans, monkeys, dogs, whales) and have particular manifestations associated with them (e.g. vocal/facial expressions, behavioral tendencies, and physiological patterns). Secondary or derived emotions are the combination or derivation from primary emotions.

Emotional dimensionality is a simplified description of basic properties of emotional states. According to Osgood, Suci and Tannenbaum's theory [3] and subsequent psychological research [4], the computing of emotions is conceptualized as three major dimensions of connotative meaning, arousal, valence and power. In general, the arousal and valence dimensions can be used to distinguish most basic emotions. The emotions location in arousal-valence space is shown in Fig. 1 [3], which results in a representation that is both simple and capable of conforming to wide emotional applications.

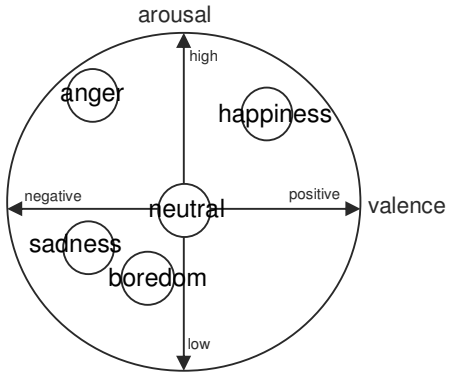


Fig. 1. Graphic representation of the arousal-valence theory of emotions

Table 1. Emotions and speech relations

	Anger	Happiness	Sadness	Fear	Disgust
Speech Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much faster
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy, chest	Breathy, blaring tone	Resonant	Irregular voicing	Grumble chest tone
Pitch changes	Abrupt on stressed	Smooth, upward inflections	Downward inflections	Normal	Wide, downward terminal inflects
Articulation	Tense	Normal	Slurring	Precise	Normal

There are numerous literatures that indicate emotion on the signs within the psychological tradition and beyond [1-2, 5-13]. The vocal cue is one of the fundamental expressions of emotions [1-2, 5-9, 11, 13]. All mammals can convey emotions by vocal cues. Humans are especially capable of expressing their feelings by crying, laughing, shouting and more subtle characteristics from speech. In ordinary conversation, the emotive cues communicate readily arousal. The communication of valence is believed to be by more subtle cues, intertwined with the content of the speech.

An important research is accomplished by Murray and Arnott [2], whose result particularizes several notable acoustic attributes for detecting primary emotions. Table 1 summarizes the vocal effects most commonly associated with five primary emotions. Classification of emotional states on basis of the prosody and voice quality requires classifying the connection between acoustic features in speech and the emotions. Specifically, we need to find suitable features that can be extracted and models it for use in recognition. This also implies the assumption that voice carries abundant information about emotional states by the speaker.

To estimate a user's emotions by the speech signal, one has to carefully select suitable features. All selected features have to carry information about the transmitted emotion. However, they also need to fit the chosen model by means of classification algorithms. A large number of speech emotion recognition methods adapt prosody and energy related features. For example, Schuller *et al.* chose 20 pitch and energy related features [14]. A speech corpus consisting of acted and spontaneous emotion utterances in German and English language is described in detail. Accuracy in the recognition of 7 discrete emotions (anger, disgust, fear, surprise, joy, neutral, sad) exceeded 77.8%. As a comparison, the similar judgment of human deciders classifying the same corpus at 81.3% recognition rate was reported. Park *et al*. used pitch, formant, intensity, speech speed and energy related features to classify neutral, anger, laugh, and surprise emotions [7]. The recognition rate is about 40% in a 40-sentence corpus. Yacoub *et al*. extracted 37 fundamental frequency, energy and audible duration features to recognize sadness, boredom, happiness, and cold anger emotions in a corpus recorded by eight professional actors [15]. The overall accuracy was only about 50%. But these features successfully separated hot anger from other basic emotions. However, in this experiment, the accuracy obtained from a 15 emotions recognition result is only 8.7%. The accuracy is 63% for male voice and 73.7% for female voice. Tato *et al.* extracted prosodic features, derived from pitch, loudness, duration, and quality features [19] from a 400-utterance database. The most important results achieved are for the speaker-independent case and three clusters (high = anger/happy, neutral, low = sad/bored). The recognition rate is close to 80%. However, the recognition accuracy of five emotions is only 42.6%. Kwon *et al.* selected pitch, log energy, formant, band energies, and Mel frequency spectral coefficients (MFCC) as the base features, and added velocity/acceleration of pitch to form feature streams [12]. The average classification accuracy was 40.8% in a SONY AIBO database. Nwe *et al.* proposed the short time log frequency power coefficients (LFPC) accompanying MFCC as emotion speech features to recognize 6 emotions in a 60-utterance corpus involving 12 speakers [13]. Results show that the proposed system yields an average accuracy of 78%.

According to the experimental results stated previously, the vocal features related prosody and energy that were extracted from time domain seem not stable in distinguishing all primary emotions. Furthermore, the prosodic features between female and male are obviously intrinsic in speech. Simple speech energy feature calculation method is also unconformable to human auricular perception.

In this paper, we make efforts on searching for an effective and robust set of vocal features from Mandarin speech to recognize emotional categories rather than modifying the classifiers. The vocal characteristics of emotions are extracted from a spontaneous Mandarin corpus. In order to surmount the inefficiency of conventional vocal features in recognizing anger/happiness and boredom/sadness valence emotions, we also treat arousal and valence correlated characteristics to categorize emotions in the emotional discrete categories. Several systematic experiments are presented. The characteristic of the extracted features is expected not only facile, but also discriminative.

Table 2. Utterances of Corpus I

Emotion \ Sex	Female	Male	Total
Anger	75	76	151
Boredom	37	46	83
Happiness	56	40	96
Neutral	58	58	116
Sadness	54	58	112
Total	280	278	558

Table 3. Utterances of Corpus II

Emotion \ Sex	Female	Male	Total
Anger	36	72	108
Boredom	72	72	144
Happiness	36	36	72
Neutral	36	36	72
Sadness	72	35	107
Total	252	251	503

The rest of this paper is organized as follows. In Section 2, two testing corpora are addressed. In Section 3, the details of the proposed system are presented. Experiments to assess the performance of the proposed system are described in Section 4 together with analysis of the results of the experiments. The concluding remarks are presented in Section 5.

2 The Testing Corpora

An emotional speech database, Corpus I, is specifically designed and set up for speaker-independent emotion classification studies. The database includes short utterances coveting the five primary emotions, namely anger, boredom, happiness, neutral, and sadness. Non-professional speakers are selected to avoid exaggerated expression. Twelve native Mandarin language speakers (7 females and 5 males) are employed to generate 558 utterances as described in Table 2. The recording is done in a quiet environment using a mouthpiece microphone at 8k Hz sampling rate.

All native speakers are asked to speak each sentence in the chosen five emotions, resulting in 1200 sentences. First, we eliminated the sentences involved excessive nose. Then a subjective assessment of the emotion speech corpus by human audiences was carried out. The purpose of the subjective classification is to eliminate the ambiguous emotion utterances. Finally, 558 utterances were selected over 80% human judgment accuracy rate. In this paper, utterances in Mandarin are used due to an immediate availability of native speakers of the languages. It is easier for the speakers to express emotions in their native language than in a foreign language. In order to accomplish the computing time requisition and bandwidth limitation of the practical recognition application, e.g. the call center system [15], the low sampling rate, 8k Hz, is adopted.

Another corpus, Corpus II, was obtained from [17]. Two professional Mandarin speakers are employed to generate 503 utterances with five emotions as listed in Table 3. The sampling rate is down-sampled to 8k Hz.

3 Emotion Recognition Method

The proposed emotion recognition method has three stages: feature extraction, feature vector quantization and classification. Base features and statistics are computed in feature extraction stage. Feature components are quantized as a feature vector in feature quantization stage. Classification is made by using various classifiers based on dynamic models or discriminative models.

3.1 The Selected Features

Fig. 2 shows the block diagram of feature extraction. In pre-processing procedure, locating the endpoints of the input speech signal is done first. The speech signal is high-pass filtered to emphasize the important higher frequency components. Then the speech frame is partitioned into frames of 256 samples. Each frame is

overlapped with the adjacent frames by 128 samples. The next step is to apply Hamming window to each individual frame to minimize the signal discontinuities at the beginning and end of each frame. Each windowed speech frame is then converted into several types of parametric representation for further analysis and recognition.

Most effective features in speech processing are found in the frequency domain. The speech signal is more consistently and easily analyzed spectrally in the frequency domain than in the time domain. And the common model of speech production corresponds well to separate spectral models for the excitation and the vocal tract. The hearing mechanism appears to pay much more attention to spectral magnitude than to phase or timing aspects. For these reasons, the spectral analysis is used primarily to extract relevant features of the speech signal in this paper.

In base feature extraction procedure, we select 6 features, which are 16 Linear predictive coding (LPC) coefficients, 12 linear prediction cepstral coefficients (LPCC), 16 log frequency power coefficients (LFPC), 16 perceptual linear prediction (PLP) coefficients, 20 Mel-frequency cepstral coefficients (MFCC) and jitter extracted form a frame. LPC provides an accurate and economical representation of the envelope of the short-time power spectrum of speech [18]. For speech emotion recognition, LPCC and MFCC are the popular choices as features representing the phonetic content of speech [19-20]. LFPC is calculated from a log frequency filter bank which can be regarded as a model that follows the varying auditory resolving power of the human ear for various frequencies [13]. The combination of the discrete Fourier transform (DFT) and LPC technique is PLP [21]. PLP analysis is computationally efficient and permits a compact representation. Perturbations in the pitch period are called jitter, such perturbations occur naturally during continuous speech.

3.2 Feature Vector Quantization

To further compress the data for presentation to the final stage of the system, vector quantization is performed. The division into 16 clusters is carried out according to the Linde-Buzo-Gray (LBG) algorithm. The vector f_n is assigned the codeword c_n^* according to the best match codebook cluster z_c using

$$c_n^* = \arg\min_{1 \le c \le C} d(f_n, z_c) \tag{1}$$

For a speech utterance with *N* frames, the feature vector Y_1 with 16 parameters is then obtained as

$$Y_1 = [c_1^* c_2^* ... c_N^*] \tag{2}$$

In another simple vector quantization method, we treat the mean feature parameters corresponding to each frames as a feature vector Y_2. Therefore, another feature vector Y_2 with 81 parameters is then obtained.

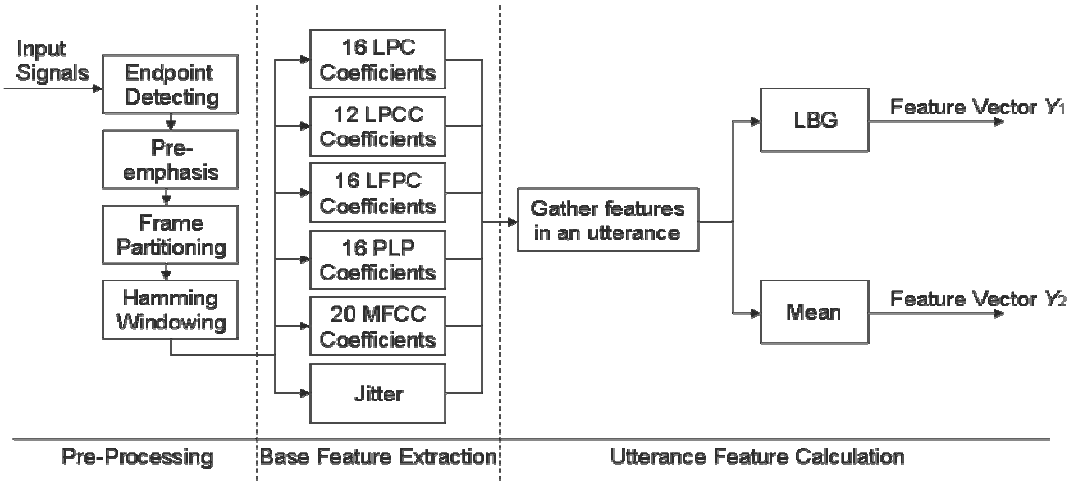


Fig. 2. Block diagram of the feature extraction module

3.3 Classifiers

Three different classifiers, linear discriminate analysis (LDA), k-nearest neighbor (K-NN) decision rule, and Hidden Markov models (HMMs), are selected to train and test these two testing emotion corpora with the extracted features from Corpus I. In K-NN decision rule, there are three nearest samples closest to the testing sample. In HMMs, our experimental studies show that a 4-state discrete ergodic HMM gives the best performance compared with the left-right structure. The state transition probabilities and the output symbol probabilities are uniformly initialized.

4 Experimental Results

The selected features in Section 3.1 will be quantified as the LBG feature vector Y_1 and the mean feature vector Y_2. Then the feature vectors will be trained and tested with three different classifiers, which are LDA, K-NN and HMMs. All these experimental results are validated by the leave-one-out (LOO) cross-validation method.

4.1 The Experimental Results Using the Conventional Prosodic Features

In [9], Kwon *et al.* drawled a two-dimensional plot of 59 features ranked by forward selection and backward elimination. Features near origin are considered to be more important. By imitating the ranking features method as [9], the speech features extracted from Corpus I are ranked by forward selection and backward elimination in Fig. 3. The experimental results of this Mandarin experiment and Kwon's show that the pitch and energy related features are the most important components for the emotion speech recognition in both Mandarin and English. We select the first 15 features proposed by [9] from Corpus I to examine the efficiency and stability of the conventional emotion speech features. The first 15 features are pitch, log energy, F1, F2, F3, 5 filter bank energies, 2 MFCCs, delta pitch, acceleration of pitch, and 2 acceleration MFCCs. Then the feature vector Y_2 and K-NN are used.

The accuracy rate of confusion matrix using conventional emotion speech features is shown in Table 4. The overall average accuracy rate of five primary emotions is 53.2%. As most previous surveyed experimental results and discussion, the pitch and energy related features extracted form the time domain confuse in anger and happiness valence emotions. The reason is that anger and happiness are close to each other in the pitch and energy related speech features; hence the classifiers often confuse one for the other. This also applies to boredom and sadness.

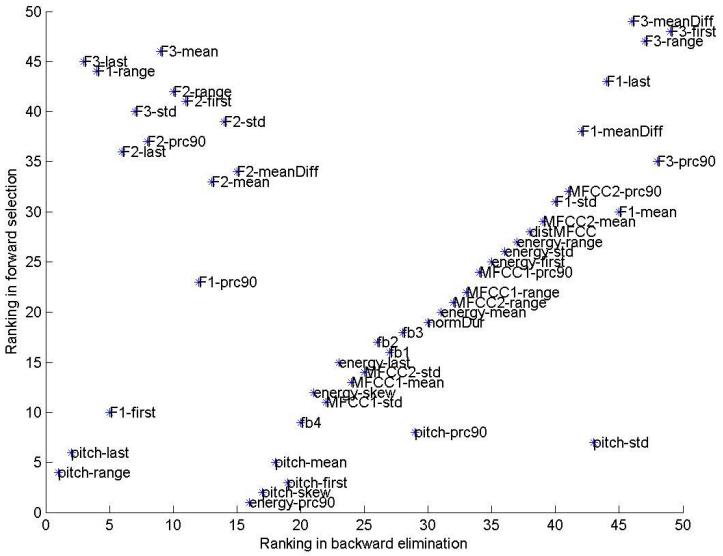


Fig. 3. Conventional emotional speech features ranking

Table 4. The experimental result of conventional prosodic features

Accuracy (%)	Anger	Boredom	Happiness	Neutral	Sadness
Anger	**59.5**	1.1	34.4	4.4	2.6
Boredom	0	**46.8**	1.1	20.4	31.7
Happiness	32.4	2.5	**58.7**	4.2	2.2
Neutral	9.4	7.7	8.7	**52.1**	22.1
Sadness	1.7	29.4	2.4	17.6	**48.9**

Table 5. The experimental result of anger and happiness recognition

Accuracy (%)	LDA		K-NN		HMMs	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
Anger	93.1	93.4	93.7	91.6	93.9	92.6
Happiness	87.7	91.2	90.4	92.8	91.2	93.5
Average	90.4	92.3	92.0	92.2	92.5	93.0

Table 6. The experimental result of boredom and sadness recognition

Accuracy (%)	LDA		K-NN		HMMs	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
Boredom	89.5	90.5	89.7	92.1	90.5	94.3
Sadness	92.2	87.6	93.5	90.4	93.2	90.9
Average	90.8	89.0	91.6	91.0	91.8	92.6

4.2 Experimental Results of Valence Emotions Recognition

The prosodic features as pitch and energy related speech features are failed to distinguish the valence emotions. The selected features in Section 3.1 will be quantified as the LBG feature vector Y_1 and the mean feature vector Y_2. Then the feature vectors will be trained and tested in Corpus I with three different classifiers, which are LDA, K-NN and HMMs. All the experimental results are validated by the LOO cross-validation method. According to experimental results shown in Table5 and 6, by applying the set of our selected emotion speech features, three recognizers are undoubted to separate the anger and happiness which most previous emotion speech recognizers are always confuse in this emotion cluster. In addition, as shown in Table 5 and 6, the high

and stable accuracy rate of various recognizers with two feature vector quantization methods provides the appropriateness to distinguish the emotions at the valence degree.

These pairwise emotions, anger and happiness, are considered to be close to each other at the valence degree with the similar prosody and amplitude. So do boredom and sadness. Conventional speech emotion recognition method suffers the infectiveness and instability in emotion recognition, especially involving emotions at the same valence degree. On the contrary, the proposed selected features solve the problem and obtain high recognition accuracy. The set of selected features are not only suitable for various classifiers but also effective for the speech emotion recognition.

4.3 Experimental Results of Corpus I and Corpus II

Table 7 and 8 show the accuracy of five primary emotions classified by various classifiers with two feature vector quantified methods in Corpus I and II. The different classifiers have different ability and property, and then we have the different recognition rates in each classifier or quantization method.

According to the experimental results shown in Table 7 and 8, the accuracy overall five primary emotions, which are anger, boredom, happiness, neutral and sadness, is approximately equivalent with the same classifier. In addition, the accuracy of two feature quantization methods of LBG and mean is quite close to each other in the same conditions. This shows that the set of the selected speech features is stable and suitable to recognize the five primary emotions in various classifiers with different feature quantization methods. By this high recognition rate of the experimental results in Corpus I and II, the selected features are proofed to be efficient to directly classify five primary emotions of arousal and valence degree simultaneously rather than only arousal degree.

Table 7. Experimental result of five emotion classes in Corpus I

Accuracy (%)	**LDA**		**K-NN**		**HMMs**	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
Anger	81.5	80.4	82.3	84.8	86.4	86.7
Boredom	80.3	79.8	84.9	82.3	89.1	88.4
Happiness	76.5	72.3	79.5	82.1	82.3	83.6
Neutral	78.4	80.5	80.4	81.2	84.5	90.5
Sadness	82.5	81.3	91.2	89.1	92.4	92.3
Average	79.8	78.8	83.6	83.9	86.9	88.3

Table 8. Experimental result of five emotion classes in Corpus II

Accuracy (%)	**LDA**		**K-NN**		**HMMs**	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
Anger	82.4	76.2	83.2	84.5	90.2	91.4
Boredom	78.9	80.2	81.5	80.9	84.3	86.7
Happiness	81.4	77.8	86.4	82.5	87.5	88.1
Neutral	76.5	79.8	84.1	83.2	90.3	86.0
Sadness	80.3	76.5	86.0	87.5	89.5	91.5
Average	79.9	78.1	84.2	83.7	88.3	88.7

Two different corpora are involved to validate the robustness and effectiveness of the selected features that the conventional speech emotion recognition method is difficult to accomplish. As the relative experimental results shown in Table 7 and 8, the overall recognition rates of both corpora are similar. The proposed selected features solve the thorny problem and obtain a high accuracy recognition rate. The set of selected features are not only suitable for various classifiers but also effective for the recognition outperform in different corpora.

5 Conclusion

In conventional emotion classification of speech signals, the popular features employed are fundamental frequency, energy contour, duration of silence and voice quality. However, some recognizers employing these

features confuse in the recognition of the valence emotions. In addition, these features employed in different corpora reveal the instable recognition results of the same recognizer.

In this paper, we use 16 LPC coefficients, 12 LPCC components, 16 LFPC components, 16 PLP coefficients, 20 MFCC components and jitter as featuers, and LDA, K-NN, HMMs as the classifiers. Presentation of the selected feature parameters is quantified as a feature vector using LBG and mean methods. The emotions are classified into five human primary categories. The emotional category labels used are anger, boredom, happiness, neutral and sadness. Two Mandarin corpora, one consisting of 558 emotional utterances employed 12 native speakers and the other consisting of 503 emotional utterances employed 2 professional speakers, are used to train and test in the proposed recognition system. Results show that the proposed system yields the best accuracy of 88.3% for Corpus I and 88.7% for Corpus II to classify five emotions.

According to experimental outcomes, we attain a high accuracy rate to distinguish anger/happy or bored/sad emotions that have similar prosody and amplitude. The proposed method can solve the difficult of recognizing the valence emotions using the set of extracted features. Moreover, the recognition accuracy of the experimental results of Corpus I and II shows that the selected speech features are suitable and effective in different corpora for the speech emotion recognition.

Further improvements and expansions may be achieved by using one or more of the following suggestions:

A possible approach to extract non-textual information to identify emotional state in speech is to apply various different and known feature extraction methods. So we may integrate other features into our system to improve emotion recognition accuracy. Besides, recognizing the emotion translation in real human communication is an arduous challenge in this field. We will try to find out the point where the emotion transition occurs

6 Acknowledge

A part of this research is sponsored by NSC 93-2213-E-036-023.

References

[1] P.R. Kleinginna and A.M. Kleinginna, “A Categorized List of Emotion Definitions with Suggestions for a Consensual Definition,” Motivation and Emotion, pp. 345-379, 1981.

[2] I. Murray and J.L. Arnott, “Towards the Simulation of emotion in Synthetic Speech: A review of the Literature on Human Vocal Emotion,” *Journal of the Acoustic Society of America*, pp. 1097-1108, 1993.

[3] C.E. Osgood, J.G. Suci and P.H. Tannenbaum, *The Measurement of Meaning*, University of Illinois Press, pp. 31-75, 1957.

[4] A. Mehrabian and J. Russel, *An Approach to Environmental Psychology*, Cambridge MA: MIT Press, pp. 192-203, 1974.

[5] A. Pasechke and W.F. Sendlmeier, “Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements,” In SpeechEmotion-2000, pp.75-80, 2000.

[6] C.D. Park and K.B. Sim, “Emotion Recognition and Acoustic Analysis from Speech Signal,” Proceedings of IJCNN, pp. 2594-259, 2003.

[7] C.H. Park, K.S.Heo, D.W.Lee, Y.H.Joo and K.B.Sim, ”Emotion Recognition based on Frequency Analysis of Speech Signal,” International Journal of Fuzzy Logic and Intelligent Systems, pp. 122-126, 2002.

[8] H. Holzapfel, C. Fügen, M. Denecke and A. Waibel, “Integrating Emotional Cues into a Framework for Dialogue Management,” Proceedings de International Conference on Multimodal Interfaces, pp.141-148, 2002.

[9] O.W. Kwon, K. Chan, J. Hao, T.W. Lee , “Emotion Recognition by Speech Signals,” Eurospeech, pp.125-128, 2003. [10][13] P. Ekman, *Handbook of Cognition and Emotion*, New York: John Wiley & Sons Ltd, 1999.

[11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J.G. Taylor, “Emotion Recognition in Human-Computer Interaction,” IEEE Signal Proc. Mag., 18(1), pp. 32-80, 2000.

[12] R.W. Picard, *Affective Computing*, MIT Press, Cambridge, pp. 178-192, 1997.

[13] T.L. Nwe, S.W. Foo and L.C. De Silva, “Speech Emotion Recognition Using Hidden Markov Models,” *Speech Communication*, pp. 603-623, 2003.

[14] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov Model-based Speech Emotion Recognition,” Proceedings of IEEE-ICASSP, pp. 401-405, 2003.

[15] S. Yacoub, S. Simske, X. Lin, J. Burns, “Recognition of Emotions in Interactive Voice Response Systems,” Eurospeech, HPL-2003-136, 2003.

[16] R.S. Tato, R. Kompe, J.M. Pardo., “Emotional Space Improves Emotion Recognition,” ICSLP, pp. 2029-2032, 2002.
[17] 張柏雄, “中文語音情緒之自動辨識,” master thesis of Engineering Science department, National Cheng Kung University, 2002.
[18] J.F. Kaiser, *Discrete-Time Speech Signal Processing*, pp.11-99, Prentic Hall PTR, 2002.
[19] B.S. Ata, “Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification,” *Journal of the Acoustical Society of America*, pp.1304-1312, 1974.
[20] S.B. Davis and P. Mermelstein, “Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences,” IEEE Transactions on Acoustics, Speech, and Signal Processing, pp. 357-366, 1980.
[21] H. Hermansky. “Perceptual linear predictive (PLP) analysis of speech,” *Journal of the Acoustical Society of America*, pp.1738-1752, 1990.