以語料為基礎的中文語篇結構關係自動標記

鄭守益 吳典松 梁婷
國立交通大學資訊科學與工程研究所
{gis93540, gis92807, tliang }@cis.nctu.edu.tw

摘要

語篇分析是文本理解中一項不可缺乏的工作，藉以釐清文章的論題或邏輯結構。本論文提出以語料為主的語篇分析方法，針對並列、承接、遞進、選擇、轉折、因果、條件、解證、目的等九種常見語篇類別，進行表層特徵收集及擴展，並制定標記規則，建立有效的自動標記程序。我們使用中研院平衡語料庫 3.0 版中的報導、傳記日記、散文、信函、評論、說明手冊等文類，共 7265 篇作為探勘語料，進行線索詞、連續詞性序列、特殊標點符號等語篇特徵之探勘。在實驗中，我們使用 100 篇平均字數為 1500 字的報紙社論進行效能評估，在句內的語篇標記部份，正確率可達到 91%，召回率是 95%，篩檢正確率是 98%。另外，在句間的標記部分，正確率可達到 86%，召回率是 93%，篩檢正確率是 95%。我們相信此語篇標記的研究，有助於自動問答、作文評分、閱讀測驗、摘要和簡報系統等應用。

1. 緒論

語篇是指在特定語境下表示完整語義的結構，它可以是一個詞、一個句子、或一群連貫的句子組合，應有一個論題結構或邏輯結構[4]。國外語篇分析的相關研究，多以連貫理論為基礎[6]。在連貫理論中，一個語篇是由許多語篇片段組成，有不同的連貫關係，例如：評估、因果、描述、解釋、排列等等。Wolf 和 Gibson[15]曾發表以語料為主的語篇連貫研究，並提出圖形表示法以描述各種連貫關係的依存現象。相對於連貫理論，許多研究受到知識表徵理論的影響，用線索片語來當作語篇中的重要結構元素[8]，而不強調語用學理論及世界知識。例如 Sadao 和 Makoto [11]利用線索詞、同義詞或片語及句子相似度來自動判斷日文的語篇結構。此外， Grosz 等人[9] 提出所謂的重心理論，探討一段文章的內在結構中，其參照延續性及言談本身特點之間的關聯。

在中文語篇的研究中，黃國文[2]提出語篇特性分為銜接與連貫兩種。銜接關係可以用語法或詞彙連結，而連貫關係則是以語義做為連結。另一方面，胡壯麟[4]將語篇特性分為指稱性、結構銜接及邏輯連接。指稱性及結構銜接都是探討語篇片段中利用詞語或語義的手段來指示語篇之間的關係。相對的，邏輯連接則表示相連的句子或句群之間的連貫關係，分為添加、轉折、因果、時空、詳述、延伸、增強等七種關係。此外，程祥徽和田小琳[1]使用複句及句群作為研究語篇片段關係的單位，將語篇分為並列、承接、選擇、遞進、轉折、因果、條件、

總分、解證、連鎖、目的等十一種關係。文獻上，中文語篇的計算模型甚少被提出。王元凱等人[14]曾提出以一個事件模型來表示中文語篇中語段的發展狀態。藉由時間線的推移將語篇結構成一個個事件，用以表現語義重心的轉移。另外 Chan 等人[7] 以人工方式分析語篇的連貫關係，並制定語篇標記，來協助找出文本中的主題段落作為摘要之候選句。

中文語篇的切分目前尚未有明確的定義，因此在本論文中，我們將依據 Marcu [10]所提出的定義，將語篇片段標記為不重疊的文本片段，並分別對分句間的句內語篇關係(以逗點做為分句的切分界線)和長句間的句間語篇關係(以冒號、句號、問號及驚嘆號為切分界線)提出有效的標記程序。

2. 語篇連貫關係分類

我們依據 [3] 和[15]所提出的複句及句群關係分類來定義語篇片段之間的連貫關係。在本論文中，我們暫不探討沒有明顯表層特徵的總分及連鎖關係，只對如下常見的九種語篇連貫關係提出自動標記程序：

表 1 語篇連貫關係類別

語篇類別	定義
並列關係	指表達幾件相關的事件，但彼此並不構成因果關係，也沒有語氣或語義上的轉折。
承接關係	描述一連續的動作，或是以發生的時間順序來連接的一連串事件，以及依事件發生的空間順序來進行敘述的事件。
選擇關係	含有從幾件事物中進行選擇的語義。
遞進關係	在連續片段中，具有後一個片段比前一個片段的語義層次更進一層關係的語篇視為遞進關係。
轉折關係	指前一片段的語義與後一片段相對或相反。
因果關係	使用兩個或兩個以上的片段來說明事件的原因及其結果。
條件關係	前一片段假設一種情況或提出一種條件，後一片段說明如果實現的話會產生的結果。
解證關係	前一片段提出一種看法、道理、事實、現象，後一片段加以解釋、說明、補充、引申的語篇。
目的關係	前一片段提出一個目的，後一片段說明為了達成這個目的需要做的事。

3. 語篇線索詞探勘

語篇線索詞的探勘是以中研院平衡語料庫 3.0 版中的敘述型語料來進行的，主要步驟為現有線索詞收集、線索詞詞性篩選、成對線索詞組探勘、單一線索詞探勘及輔助特徵探勘。

3.1 線索詞收集與篩選

我們以「現代漢語」[3]所列出的語篇線索詞來當作查詢詞，在探勘語料中進行更多線索詞的收集。由於詞性影響到詞彙的語義或語法角色，因此這些詞在語料中的詞性若為下列的詞性，將不視為線索詞。

{Na, Nb, Nc, SHI, T, VA, VC, VCL, VD, V, VH, VJ, Nf}

3.2 成對線索詞組探勘

成對線索詞組的探勘包括四個主要步驟：

步驟 1：設定抽取線索詞之範圍及位置

從探勘語料中，我們隨機選取 24 個線索詞組進行例句搜尋，共收集 2300 個分句，由其統計分布得到線索詞多為語篇片段中的第 1 到第 12 個詞。

另一方面，從探勘語料的例句統計分布，我們也觀察到成對線索詞組，不論是句內或句間的語篇片段連結距離多為 3，亦即若詞組中的前詞在第一語篇片段，則其搭配詞多在接續的三個片段中出現。

步驟 2：設定線索詞出現位置之權重

語篇中的線索詞是否具有連結功能與其出現的位置有關，因此藉由觀察步驟 1 中線索詞分佈位置的統計資料，我們發現線索詞的分布近似於函數 $\frac{1}{x^3}$ (其中 x 為線索詞的出現位置，$1 \le x \le 12$)，因此我們可以此函數來作為計算線索詞組連結強度時的權重，並進行正規化，使其權重值介於 0 與 1 之間。設線索詞在分句中出現的位置共有 j 個($1 \le j \le 12$)，則由線索詞出現在分句內的位置分佈，我們可得到一正規化常數為 D：

$$D = \sum_{j=1}^{12} \frac{1}{j^3} = 1.2 \qquad (1)$$

因此，若線索詞出現在第 j 個位置，其權重即為

$$w_j = \frac{1}{1.2 j^3} \qquad (2)$$

步驟 3：計算線索詞組之連結強度 k

我們以線索詞組(T_h, T_i)一起出現在語篇片段之間的頻率標準差倍數，作為其連結強度[12]，其中T_h為給定的線索前詞，T_i為T_h的第i個搭配詞，其計算公式如下：

$$k_i = \frac{f_i - \overline{f}}{\sigma} \tag{3}$$

其中搭配詞T_i出現在語篇片段的位置j $(1 \le j \le 12)$之頻率f_i定義為：

$$f_i = \sum_{j=1}^{12} f_{i,j} w_j \tag{4}$$

其平均頻率$\overline{f}$以及標準差σ的計算公式如下：

$$\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f_i \quad ; \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f_i - \overline{f}\right)^2} \tag{5}$$

步驟 4：線索詞組篩選

我們配合 k 值排序並以人工篩選出句內線索詞組 406 組，句間線索詞組 82 組。

3.3 單一線索詞探勘

中文語篇的成對線索詞，有時也可單獨出現，例如：

*例句 1：他(**不但**)吃米飯(A)，**也**吃牛排(B)。*

其中「不但」可以省略。另外中文線索詞在書寫的過程中，常會省略關聯前詞或後詞，如例句 2 中的「如果」和例句 3 中的「所以」常被省略：

*例句 2：(**如果**)我們這麼做，**可能**會導致環境的破壞。*

*例句 3：**因為**情勢如此變化，(**所以**)我們不得不做這樣的決定。*

因此，在語篇連貫關係辨識的過程中，有必要進行單一線索詞的收集及探勘工作。我們以檢驗成對線索詞組的例句進行辨識篩選，收集了 309 個單一線索詞。

3.4 輔助特徵探勘

我們參考 [1]及 [13]的研究，設定了如下四種輔助特徵：

a. 當具有時間詞(Nd)詞性的詞彙，例如：「今天…明天」，出現在連續的語篇片段時，將視這些語篇片段具有「承接關係」。

b. 當具有數詞定詞(Neu)詞性的詞彙，例如：「第一…第二」，出現在連續的語篇片段時，將視這些語篇片段具有「並列關係」。

c. 語篇片段的末尾若出現標點符號「：」，將可判定其次一語篇片段為

「解證關係」。

d. 若相似的語篇片段連續出現時，則可以將這些語篇片段判定為「並列關係」。

4. 辨識及標記執行步驟

我們使用中央研究院的線上中文斷詞系統[1]，進行文本斷詞及詞性標記的工作，並將語篇辨識及標記的工作分為三個階段。其步驟如下表所示：

表 2 語篇連貫關係辨識及標記步驟

階段	步驟	執行動作
一	**成對線索詞組比對**	
	1	以兩個分句為單位進行成對線索詞組比對。
	2	將比對後具有語篇連貫關係之兩個分句合併。
	3	判斷是否已合併完成，並進行語篇連貫關係標記。
二	**單一線索詞比對**	
	1	進行向前連結之單一線索詞比對及合併。
	2	進行向後連結之單一線索詞比對及合併。
	3	判斷是否已合併完成，並進行語篇連貫關係標記。
三	**輔助特徵及特殊單一線索詞比對**	
	1	進行連續 Nd 及 Neu 詞彙之比對及合併。
	2	進行解證關係標點符號之比對及合併。
	3	進行相似句之比對及合併。
	4	進行特殊線索詞之比對及合併。

4.1 成對線索詞組比對

由於成對線索詞組本身即具有排除語篇連貫關係歧義性，及明顯合併範圍和方向之特性，因此我們將其列為第一優先比對的特徵，此階段分為三個步驟：

步驟 1：以兩個分句為單位進行成對線索詞組比對

假設某待處理文本所含之語篇片段數量為 n，語篇連結門檻值為 d，則我們可產生一個長度為 n 的輸入陣列及 $n \times d$ 之比對結果矩陣。若為句間比對，則以每一長句第一分句為輸入之比對片段。將陣列輸入系統，並依序增加 d 值進行成對線索詞比對。

[1] 請參閱網址：http://ckipsvr.iis.sinica.edu.tw/

步驟 2：將比對後具有語篇連貫關係之兩個分句合併。

我們將合併之過程分為兩個部份，第一個部分稱為縱向合併，其遞增變數為語篇連結門檻值d，處理同一片段的合併問題。第二個部分稱為橫向合併，其遞增變數為語篇片段數量n，以處理相鄰片段的合併問題。我們將依循以下規則：

規則 1：縱向合併時，同一片段若可同時與兩個以上之片段形成語篇連貫關係時，只保留距離最小者。

如例句 4 中語篇片段 (B)的線索詞「因為」可與線索詞「所以」連結，但 (C)與(E)卻同時出現該線索詞，因此，我們根據規則 1 取距離最小者，將(B)(C)兩個片段合併。

*例句 4：這種盲識她不覺得有必要去澄清(A)，**因為**知情的人知道真相為何(B)，**所以**她認為夫妻檔是利多於弊(C)，**因為**兩人志趣相投(D)，**所以**不但目標一致(E)，**而且**團結力量一定大(F)。*

規則 2：橫向合併時，相鄰片段若連續形成相同的語篇連貫關係時，則合併成為同一語篇在同一階層。

如例句 5 中語篇片段(C)、(D)、(E)、(F)經依循規則 1 進行縱向連結之後，將以線索詞「或」分別連結成 3 個選擇語篇(CD、DE 、EF)，然後，我們再依規則 2 進行橫向連結，將語篇段落合併在同一階層為一個大的語篇段落。

*例句 5：今年的元宵燈節十分熱鬧(A)，展出的花燈無奇不有(B)，**或**以造型取勝(C)，**或**以作工服人(D)，**或**色彩華麗(E)，**或**聲光迷人(F)。*

規則 3：橫向合併時，相鄰片段若連續形成不同之語篇連貫關係時，以向左合併為原則，合併成為不同語篇不同階層。

如例句 6 中語篇片段(A)和(B) 經依循規則 1 進行縱向連結之後，以線索詞組「不但…還」合併為遞進關係，(A)和(C)又以「因為…所以」合併為因果關係，然後，我們再依規則 3 進行橫向連結，將這兩個語篇段落以不同語篇合併在不同階層。

*例句 6：**因為**她**不但**嘴巴很壞(A)，**還**喜歡打人(B)，**所以**我們都不喜歡她(C)。*

步驟 3：判斷是否已合併完成，並進行語篇連貫關係標記。

若輸入文本已合併為單一語篇段落，則對照語篇連貫關係符號表進行語

篇標記後跳出比對流程，若尚未合併為單一語篇段落，則繼續第二階段之比對工作。

4.2 單一線索詞比對

使用單一線索詞來辨識語篇連貫關係時，需要考慮連結方向、涵蓋範圍以及出現位置等三個問題。因此，我們設計三種屬性：

1. 連結方向

若由線索詞向後連結次一片段，則將此值設為 1，若為向前連結前一片段，則設為-1。

2. 出現位置

線索詞出現的位置可分為兩種，一為出現在語篇片段的前半部份，並在設定的位置門檻值內，則設定為 0；若出現在語篇片段末尾，則設定為 1。至於出現於中間位置的線索詞，我們則忽略不計。

3. 適用片段種類

可同時使用在句內及句間的線索詞，則此值設定為 1；反之若只能使用在句內，則設定為 0。

此階段所指的「單一線索詞」包括成對線索詞組的省略詞，及解證與目的關係中的一般線索詞共 244 個，其屬性值為(-1,0,0)、(-1,0,1)、(1,0,0)。根據我們的觀察，若在句內省略前詞而僅單用後詞，則其連結方向多為向前連結，反之亦然。此外也會有複合線索詞的出現，如例句 7 所示：

*例句 7：他會這麼做(A)，多少**也因為**還愛著你(B)。*

在分句(B)中出現兩個單一線索詞，一個是「也」表示並列關係，另一個是「因為」表示因果關係。因此，我們將以下規則進行第二階段比對及合併：

規則 4：若比對單一線索詞時，同一語篇片段出現兩個以上之候選線索詞，則依以下詞性優先順序決定：

Cbb> Caa> Cab> Cba> D> Da> Dk> P

規則 5：單一線索詞連結時須避免將內含輔助特徵及特殊線索詞之語篇片段合併。如例句 8 中所出現之線索詞「或」，不應合併(A)，因其包含了特殊線索詞「宣示」。

*例句 8：我們建議政府儘快明白**宣示**(A)，**或**為政治、經濟問題(B)，國家永續發展問題，何者才是政府的最大關切？*

規則 6：若向前合併之單一線索詞單獨出現在第一分句，則為句間線索

詞，不與句內連結。如例句 9 之線索詞「然而」。

*例句 9：雲林縣此舉，除了財政拮据之外，還夾雜著對大規模企業「本縣拉屎，他處下蛋」的忿懣與積怨，因此高舉防治污染大旗，以環境保護為名義徵稅。**然而**，純就租稅體制而言，雲林縣此舉並不符合稅制的基本邏輯。*

規則 7：若向後合併之單一線索詞單獨出現在第一分句，則為句內線索詞，不與句間連結，如例句 10 之線索詞「即使」。

*例句 10：**即使**真的應將污染性企業產值列入分配因素考量，亦不應只涵蓋石化工業，高污染產業還有很多。*

4.3 輔助特徵及特殊單一線索詞比對

此階段我們總共設定了四種輔助特徵及兩種特殊線索詞比對，共分為 4 個步驟：

步驟 1：進行連續時間詞及數詞定詞詞彙之比對及合併

如前所述，對連續語篇片段，我們可利用時間詞來輔助辨識承接關係；用數詞標示並列關係。

步驟 2：進行解證關係標點符號之比對及合併

我們以冒號(：)作為輔助解證關係的辨識及標記。

步驟 3：進行相似句之比對及合併

我們採用之前所設計的中文句子相似度計算模組[5]，進行語篇中並列關係的探勘。此模組考量中文相似句中的語義和結構的相似度。從訓練語料中我們抽出 3000 對分句進行測試，部份結果如下：

表 3 中文相似句實驗範例

編號	前分句	後分句	相似值
1	刀魚說生命的顏色是白色的	蚯蚓說生命的顏色是紅色的	1.00
2	久之則漸似矣	久之則愈似矣	1.00
3	法名傳縈	字雪個	1.00
4	能捉的都被捉了	該殺的都被殺了	1.00
5	自一以分萬	自萬以治一	1.00
6	錯開順序	顛倒方向	1.00
7	有一點不凡	有一點叛逆	1.00
8	第一是人文之美	第二是人格之美	1.00
9	先是綠色的葉片	後是白色的花朵	0.84
10	從以前的希特勒、史達林	到近代的馬可仕、哈珊	0.77

由上表觀察，編號 1~8 為並列例句，9~10 為承接例句。我們在實驗中亦發現，相似度大的句子幾乎都為並列結構，只有極少數例句為承接。因此，本系統將相似度高的分句優先判定為並列。我們將相似值的門檻值訂為 0.48，這個數值可以達到資料涵蓋率 80.45%，正確率 83.88%。

步驟 4：進行特殊線索詞之比對及合併

我們在語料中發現有兩種特殊線索詞。這兩種線索詞的共同特性是，都出現在語篇片段的末尾，涵蓋範圍比一般的線索詞要大；不同之處則在於連結的方向，一個向前，一個往後。

第一種為列舉線索詞，此種線索詞的連結方向往前，所連結之語篇連貫關係為並列，屬性值為 (-1,1,0)，僅適用於句內關係的比對，共收錄 5 筆資料。如例句 11 中的「等等」，即可將(C)、(D)、(E)三個語篇片段合併為並列關係。

*例句 11：環保局秘密提前啟用本垃圾場(A)，將垃圾灰燼進場掩埋(B)，原承諾之八十三年元月十五日啟用前對南港居民做簡報(C)，提出污染防治保證書(D)，及有效管理辦法及罰則**等等**(E)，均未兌現(F)。*

第二種為動詞線索詞，此種線索詞的連結方向往後，所連結之語篇連貫關係為解證，屬性值為：(1,1,1)，共收錄 57 筆資料。如例句 12 中的「宣示」，即可將(B) 與(C)、(D)、(E)、(F) 五個語篇片段合併為解證關係。

*例句 12：西方人士說(A)，這份文件**宣示**(B)，一個歐洲關係新時代已開始(C)，各國將不再相互仇恨(D)，轉而建立夥伴關係(E)，並伸出友誼之手(F)。*

4.4 標記範例與說明

我們將輸入的文本自動標記出相應的語篇連貫關係。若某語篇段落內含兩個或以上之語篇片段時，則依規則，標記為樹狀結構。表 4 和圖 1 分別為語篇連貫關係符號表和標記結果範例。

表 4 語篇連貫關係標記符號表

符號	說明
@	做為分隔語篇段落的界線。
( )	標示語篇結構的左右邊界。
|	表示在同一層的語篇片段。
D#,	標示語篇連貫關係之編號。

[]	標示語篇片段的左右邊界。
C#	標示分句語篇片段在整個長句裡的順序。
S#	標示長句語篇片段在整個文章裡的順序。
Theme	標示語篇連貫關係中的第一個語篇片段。
Rheme	以 Rheme 標示語篇連貫關係中的其他語篇片段。

*D8,(Theme: [C1:*尤其是除了金融與企業行為的管理以外，*]|D1,(Theme: [C2:*更是有許多限制與控管是針對個人而來的，*]| Rheme:D4,(Theme: [C3:*例如公司董事與經理人赴大陸投資行為、企業投資的檢舉獎金，*]| Rheme: [C4:*以及開放大陸人士來台灣觀光的管理等等。*]))*

圖 1 語篇標記結果範例

我們將之轉換成樹狀圖，如下所示：

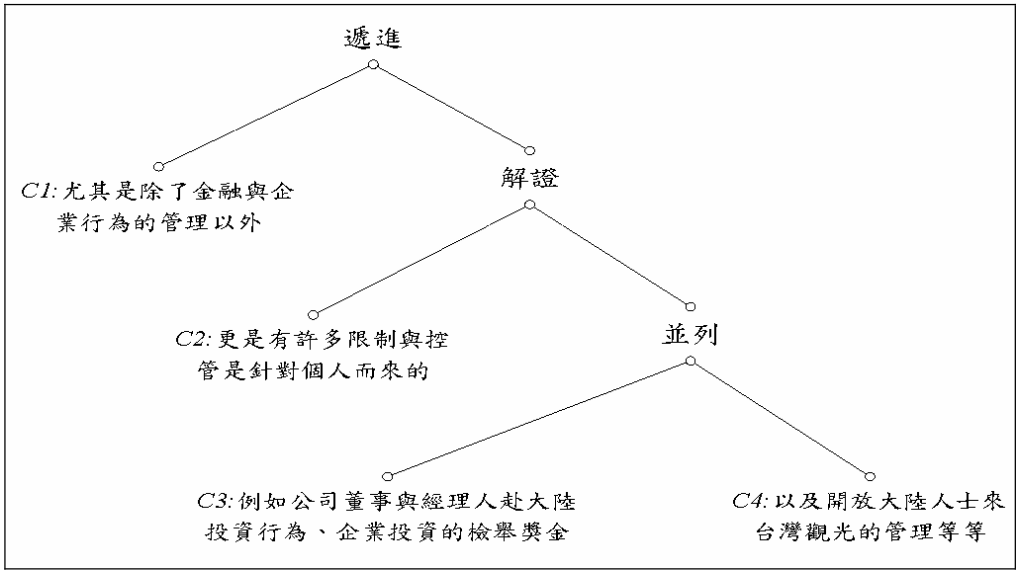


圖 2 語篇標記樹狀結構範例

5. 實驗設計與分析

我們分別從工商時報、中國時報、自由時報、聯合報、及經濟日報等主要的平面媒體電子報，收集 100 篇社論來檢驗本論文所提的標記程序效能，每篇的字數平均約為 1500 字。我們將系統的標記效能依表 5 的標記情況定義為：

表 5 可能的標記情況

	應標記	不應標記
正確標記	a	none
錯誤標記	b	c
未標記	d	e

$$\text{標記正確率 } P=\frac{a}{a+b+c} \tag{6}$$

$$\text{標記召回率 } R=\frac{a}{a+b+d} \tag{7}$$

$$\text{系統篩檢正確率 } FP=\frac{d+e}{c+e} \tag{8}$$

在我們的實驗中，句內標記正確率可達到 91%，召回率是 95%，篩檢正確率是 98%。另外，句間標記正確率可達到 86%，召回率是 93%，篩檢正確率是 95%。

表 6 語篇數量分佈統計表

語篇編號	語篇種類	適用關係	數量	百分比
1	並列	句內	442	17.20%
		句間	65	8.61%
2	承接	句內	99	3.85%
		句間	57	7.55%
3	選擇	句內	85	3.31%
		句間	18	2.38%
4	遞進	句內	521	20.27%
		句間	99	13.11%
5	轉折	句內	703	27.35%
		句間	277	36.69%
6	因果	句內	192	7.47%
		句間	70	9.27%
7	條件	句內	361	14.05%
		句間	14	1.85%
8	解證	句內	136	5.29%
		句間	155	20.53%
9	目的	句內	31	1.21%
		句間	0	0%
總計		句內	2570	100%
		句間	755	100%

另外由表 6 可以看出，社論類的文章使用最多的語篇是遞進與轉折。次多者在句內是並列與條件，而句間則為解證與因果。相對於句間大量使用解證，句內較常使用的是條件語篇。至於在語篇的特徵上，我們也觀察到單一線索詞的使用率不論是在句內或句間的語篇辨識上都是最高的，分別有 77.43%及 75.63%。使用率偏低的特徵在句內是連續的數詞定詞以及最後一個詞使用有列舉涵義的詞，其比例分別只有 0.58%、0.62%。在句間語篇辨識上使用率較少的則是片段最後一個詞使用有解證涵意的線索詞，其比例只有 0.13%。

6. 結論與未來研究

本論文提出並實作一個中文語篇自動標記系統，經實驗數據的分析顯示，能有效地標記出並列、遞進、轉折等九類語篇連貫關係。其後續研究有下列幾個方向：

1. 可利用同義詞或近義詞，搭配連結強度來自動抽取更多的線索詞，以提高系統的資料涵蓋率。
2. 由於語篇的結構有時十分複雜，因此需要找尋更多的輔助特徵，來協助系統標記語篇。
3. 可進行更多位之語篇的定義與研究，以利提高系統的資料涵蓋率。
4. 可利用機器學習及建立語義概念網路的方式，來幫助系統辨識語義的轉折，並可利用統計模型來進行語篇的自動辨識。

參考文獻

[1] 田小琳，”中學教學語法系統提要（試用）”，北京人民教育出版社，1984。

[2] 黃國文編著，”語篇分析概要”，北京商務印書館，1988。

[3] 程祥徽、田小琳，”現代漢語”，台北書林書店，1989。

[4] 胡壯麟，語篇的銜接與連貫，上海外語教育出版社，1994。

[5] 鄭守益,梁婷, “中文句子相似度之計算與應用,第十七屆自然語言與語音處理研討會”, Tainan, Taiwan, 2005 Proceedings of ROCLING XVII pp. 113-124.

[6] Allen, J., Natural Language Understanding, 2nd, Benjamid/Cummings, 1995.

[7] Chan, W. K., Lai, B. Y., Gao, W. J. and T'sou, K., "Mining Discourse Markers for Chinese Textual Summarization." In Proceedings of the 6th Applied Natural Language Processing Conf. and the North American Chapter of the Association for Computational Linguistics. Workshop on Automatic Summarization,Seattle, Washington, 29 April to 3 May, 2000.

[8] Grosz, B. J. and C: L. Sidner, “Attention, intentions, and the structure of discourse”, Computational Linguistics, vol. 12, no. 3, pp. 175-204, 1986.

[9] Grosz, B. J., A. K. Joshi, and S. Weinstein, “Centering: a framework for modeling the local coherence of discourse”, Computational Linguistics, vol. 21, no. 2, pp. 203-225, 1995.

[10] Marcu, D., “The rhetorical parsing of unrestricted texts: A surface-based approach.”, Computational Linguistics 26: 395-448, 2000.

[11] Sadao K., Makoto N. , “Automatic Detection of Discourse Structure by Checking Surface Information in Sentences”, COLING , pp.1123-1127, 1994.

[12] Smadja, F., ”Retrieving collocations from text: Xtract”, Computational Linguistics, 19(1): 143-177, 1993.

[13] Tomohide S. and Sadao K., "Automatic Slide Generation Based on Discourse Structure Analysis", In Proceedings of Second International Joint Conference on Natural Language Processing (IJCNLP-05), Jeju Island, Korea, pp.754-766, 2005.

[14] Wang, Y. K., Y. S. Chen, and W. L. Hsu, “Empirical study of Mandarin Chinese discourse analysis: an event-based approach,” to appear in 10th IEEE Int’l Conf. on Tools with Artificial Intelligence, 1998..

[15] Wolf, F. and Gibson, E.,”Representing discourse coherence: A corpus-based analysis”, Computational Linguistics, 31(2): 249-287, 2005.

中文動詞名物化判斷的統計式模型設計

馬偉雲　黃居仁

中央研究院語言學研究所
ma@iis.sinica.edu.tw　churenhuang@gmail.com

摘要

中文動詞名物化的現象在中文的語法研究上一直是一個重要的課題。而對中文自然語言處理系統來說，自動判別句子當中的動詞是否名物化也是在剖析過程當中不可或缺的技術之一。一個動詞在句子當中所扮演的角色到底是單純的謂語，或是派生名詞，影響剖析結果甚鉅。由於中文動詞名物化時缺乏構形上的變化(zero-derivation)，因此判斷動詞是否名物化就必須仰賴動詞本身的內部語素結構、語意以及上下文方可得知。過去由於語料庫大小限制，欠缺足夠的名物化樣本及其語境可供建立統計式模型，因此前人多利用少數觀察到的語法規則企圖建立規則式模型來判斷名物化。舉例來說，動詞前或後出現“的”時，這個動詞即有很高的可能性是派生名詞。然而，較為複雜的名物化現象仍難以這些簡單的規則就能判定。

本論文是第一個嘗試以統計方式自動判斷中文動詞名物化的研究。利用大規模的帶名物化標記的語料庫，根據不同假設，訓練出各類統計式模型，自動判斷一個動詞在其語境當中是否名物化。實驗結果顯示出，表現最佳的統計模型對於派生名詞的包含率為 71.8%，準確率為 76.6%，F-Score 為 74.1%。我們也針對不同的統計式模型的表現作分析，發現整合派生名詞的動詞來源詞(verbal counterpart)的語法詞類(syntax category)訊息的模型，往往比未包含此訊息的模型表現要來得好。經由實際語料的分析，我們觀察到不同的動詞來源詞的語法詞類不僅僅在扮演謂語角色時語境不同，在扮演派生名詞的角色時，所搭配的語境有時也有極大的差異。這樣的差異性在設計名物化判斷系統上是不可欠缺的關鍵因素之一。

1. 簡介

中文動詞名物化的現象非常普遍[葉美利等 1992]，當一個動詞在句子中轉換成名詞的語法功能時，就稱此動詞名物化或是轉換成派生名詞了。然而，一個動詞要轉換到什麼程度才可稱為派生名詞，目前並沒有統一的看法。根據[詞庫小組技術報告 95-02/98-04]對於動詞名物化的定義。動詞名物化四種最常見的情況是：一.動詞出現在主語位置，如(1)(2)。二.動詞出現在虛化動詞(light verb)的賓語位置，如(3)(4)。三.動詞出現在名詞片語結構的中心語，如(5)(6)。四.動詞出現在名詞前修飾名詞，如(7)(8)。

(1)　打架(VA) [+nom] 是(SHI) 不(D) 對(VH) 的(DE)
(2)　上學(VA) [+nom] 幫助(VC) 我們(Nh) 學習(VC) 知識(Na)
(3)　進行(VC) 調查(VE)[+nom]
(4)　維持(VJ) 清潔(VH)[+nom]
(5)　學生(Na) 的(DE) 不(D) 合作(VH) [+nom]
(6)　他(Nh) 對(P) 國家(Na) 的(DE) 認同(VJ) [+nom]
(7)　主辦(VC) [+nom] 單位(Na)
(8)　吵架(VA) [+nom] 方式(Na)

[葉美利等 1992]提供了派生名詞更細緻的分類，將派生名詞依帶論元與否分為兩大類，並按所帶的論元及其體現將帶論元的派生名詞分成十個小類，且提供了語法表達模式。

自動判斷名物化與否的技術是中文自然語言處理重要的一環，對於句子或動詞片語的剖析更是關鍵。但由於中文動詞名物化時缺乏構形上的變化(zero-derivation)，缺乏英語 -ion、-ment、-ing 等動詞名物化標示，因此無法從構形上判斷，必須仰賴動詞本身的內部語素結構，語意，以及上下文方可得知。過去由於語料庫大小限制，欠缺足夠的名物化樣本及其語境可供建立統計式模型，因此前人多利用少數觀察到的語法規則企圖建立規則式模型來判斷名物化。如[Lin et al., 1997]分析包含派生名詞的名詞片語語法結構，建立一系列的語法規則作為剖析器的參數，當剖析完成時自然就決定了每個動詞名物化與否。這樣的作法好處是將名物化判斷和剖析整合在同一過程當中，考慮的範圍廣而全面，並且最後能夠得到完整的剖析結構。只是[Lin et al., 1997]所定義的包含派生名詞的名詞片語語法結構，必須後接“的”才算是包含派生名詞的名詞片語。並無法全面處理上述[詞庫小組技術報告 95-02/98-04]對於動詞名物化的定義。

相較於[Lin et al., 1997]，本論文的作法將名物化判斷從剖析當中分離出來，名物化判斷之後的結果才送交剖析器作進一步處理，這樣做有兩個理由：1.名物化判斷基本上屬於詞類標記(tagging)的問題，目前主流的剖析技術都是以詞類標記後的結果作為其輸入。2.雖然現今主流的統計式剖析技術，如“機率式上下文無關”(PCFG) 剖析器，理論上只要有大規模的帶名物化標記的語料庫，它也可以訓練出統計式模型來剖析，一併處理名物化判斷的問題。不過，因為影響剖析好壞的原因很多，並不只有動詞名物化這個原因。因此暫時排除剖析這個變因，可以使我們專注在名物化判斷這個主題。除了得到一個高準確率的模型之外，我們也希望能分析出影響動詞名物化的催化因素，提供語言學上的解釋及驗證。特別是在語言學的分析當中，[葉美利等 1992]、[Huang et al., 1994]等都提出轉換後的派生名詞和轉換前的來源詞 (verb counterpart) 有密不可分的關係，在上下文中往往可以找到相對應的論元成分。本論文進一步觀察到針對某些特定來源詞的語法詞類，他們轉換成派生名詞之後，其論元的位置通常也具有共通性。因此，我們設計了區分來源詞語法詞類的模型跟未區分的模型，希望能探究來源詞語法詞類在名物化當中所扮演的角色。

名物化判斷基本上可說是一個詞類標記(tagging)的問題，當我們判斷一個動詞在上下文中已經轉換成派生名詞時，我們可以給予一個通用的名物化詞類標記(如 Nv)取代原本的動詞詞類標記，另一個策略是在其原本的動詞詞類後面附加名物化特徵標記(如(VC)[+nom])。由於詞類標記的技術已經相當成熟，因此在本論文中首先會測試在詞類標記技術當中最廣為使用的隱藏式馬可夫模型(HMM model)，之後為了更進一步掌握動詞本身特性和上下文的因素，我們模仿在詞義區分當中廣為使用的貝氏分類器(Naïve Bayes Classifier)，提出另一種型態的統計式模型，並藉由實驗證明其可行性。在本論文的實驗討論當中，也分析了各類模型所隱含的語言學現象。

2. 系統描述

設計系統之前，首先我們必須先決定什麼是它的輸入以及什麼是它的輸出，才能明確規範出系統所要真正解決的問題。不同的輸入或輸出也會在在評估系統表現上產生極大的差異。若是單一限定在動詞名物化判斷這個主題，並打算測試其表現好壞，直覺上，輸入應是一個已經斷好詞，並且可能已經有一些現成的標記(如詞類標記)的句子，斷詞和標記都是正確無誤的。這樣的輸入經

過名物化模型的判斷，輸出時，系統在某些動詞上取代或標記上名物化的詞類或特徵。

輸入：學生(Na) 的(DE) 不(D) 合作(VH)
輸出：學生(Na) 的(DE) 不(D) 合作(Nv) / 學生(Na) 的(DE) 不(D) 合作(VH)[+nom]

但是以這樣的輸入來評估系統表現顯得不切實際，因為在中文處理當中，我們所面臨真實的輸入是一個未正確斷好詞以及標記的句子。但是若是採取一個連斷詞都不具備的輸入，系統又勢必承接許多斷詞所造成的錯誤，在評估時就難以評估出名物化判斷模組的好壞。因此，我們採取的策略是將斷詞這個變因在系統設計當中剔除。也就是說，在實際的中文處理上，斷詞由另一套斷詞模組負責，和名物化判斷無關，在評估名物化判斷的表現時，我們的輸入就是一句已經正確斷好詞的句子。我們將詞類標記和名物化判斷整併到一個模組當中，原因是這兩者在判斷上實際上密不可分，互為影響，一併設計並且評估它們是比較符合實際的作法。另外，由於在派生名詞的輸出上，附加名物化特徵的呈現方式(如(VH)[+nom])會比只標記名物化詞類(如 Nv)，更多了來源詞的語法詞類訊息(如 VH)，因此我們採用這樣的輸出模式。

輸入：學生 的 不 合作
輸出：學生(Na) 的(DE) 不(D) 合作(VH)[+nom]

我們採用[詞庫小組技術報告 95-02/98-04]對於動詞名物化的定義來設計系統，原因是這套定義十分簡明，並且我們所用來訓練的語料庫的名物化標記亦採用這套定義，如此可以產生具有一致性標注原則的訓練以及測試語料，來產生與評估我們的模型。

3. 統計模型

詞庫小組在 1997 年完成了帶詞類標記的平衡語料庫[詞庫小組技術報告 95-02/98-04]，並且利用附加名物化特徵的方式呈現派生名詞。語料庫的規模為五百萬詞。這樣的規模使得訓練統計式的名物化判斷模型成為可能。

名物化的催化因素大體上可分成兩大類，分別是動詞本身名物化的可能性以及上下文對此動詞名物化的影響力。根據不同假設，這兩大類催化因素又可繼續細分成更細部的催化因素，我們也據此提出兩種隱藏式馬可夫模型以及三種貝氏分類器。

值得一提的是，本論文並沒有使用目前在標記上與分類領域最新的統計式技術，如“條件式隨機場域”(Conditional Random Field)或是“支援向量機器”(Support Vector Machine)等，其原因並非認為它不適用於名物化判斷，而是本論文作為第一個針對統計式名物化判斷的研究，除了想開發出準確的系統之外，我們也想藉由各類的模型探討背後所隱含的語言學現象，找出名物化的催化因素。因此，根據不同假設，我們設計出比較符合直覺的統計式模型，且均以相連機率(Bigram Probability)為運算材料，在實驗分析上，這些機率值可以讓我們深入觀察或驗證名物化的催化因素，並使錯誤分析更容易。以下為本論文所使用的符號:

w_i	the word at position i
v_i	the verb at position i
t_i	the tag at position i
$v_i(t_i)[+nom]$	the nominalized verb and its tag at position i
$v_i(t_i)$	the non-nominalized verb and its tag at position i
$(t_i)[+nom]$	the nominalized-verb's tag at position i
(t_i)	the non-nominalized-verb's tag at position i
$v[+nom]$	a nominalized verb
v	a non-nominalized verb
$c(v_i):\{w_{i-1},t_{i-1},w_{i+1},t_{i+1}\}$	the context of

3.1. 隱藏式馬可夫模型

名物化判斷基本上可說是一個詞類標記的問題，而詞類標記的技術當中最廣為使用的就是隱藏式馬可夫模型。我們提出兩種不同類型的隱藏式馬可夫模型，未區分動詞來源詞語法詞類的模型，稱之為 HMM1，區分動詞來源詞語法詞類的模型，稱之為 HMM2。我們採用傳統的雙連式隱藏式馬可夫模型，以雙連詞類機率作為運算參數。換句話說，這樣的模型定義了判斷標的之上下文為前一個以及後一個語法詞類。

3.1.1 HMM-1

對每一個動詞的上下文來說，當我們假設一個特定的上下文對每一個動詞都具有同樣的名物化催化力時(如任一個動詞，無論它是什麼詞類，只要前接副詞-"地"，幾乎就可以確定此動詞不會轉換成派生名詞，而單純扮演謂語的角色)，我們就可以將所有派生名詞都視為同一個詞類，如 Nv，因此就相當於我們在做詞類標記的問題，只是詞類集多了一個成員-"Nv"。利用典型的隱藏式馬可夫模型，計算

$$\prod_{i=1}^{n}[P(w_i \mid t_i)\times P(t_i \mid t_{i-1})]$$

找出最佳的詞類序列，就可以得到名物化判斷的結果。不過這樣的結果還並不是最後的輸出，因為對於所判斷的派生名詞，我們仍然不知道他們的來源詞語法詞類。一個最簡單的作法，就是取他們在統計上頻率最高的來源詞語法詞類作為輸出。之所以可以這樣做，主要是由於一個動詞的名物化現象絕大多數只發生在它的一個特定動詞詞類身上。以下舉例說明整個流程：

輸入：學生 的 不 合作
經過 HMM 後：學生(Na) 的(DE) 不(D) 合作(Nv)
取頻率最高的來源詞語法詞類為輸出：學生(Na) 的(DE) 不(D) 合作(VH)[+nom]

3.1.2 HMM-2

對每一個動詞的上下文來說，當我們假設一個特定的上下文對不同動詞的詞類具有不同的名物化催化力時，在模型設計上就必須區分出上下文和不同動詞詞類之間的關係。以詞類標記的角度來

看，就相當於詞類集多了許多成員，如“(VC)[+nom]”，“(VA)[+nom]”，“(VH)[+nom]”…等等。利用典型的隱藏式馬可夫模型找出最佳的詞類序列就可得到最後輸出。以下舉例說明整個流程：

輸入：學生 的 不 合作
經過 HMM 後：學生(Na) 的(DE) 不(D) 合作(VH)[+nom]

3.2. 貝氏分類器

名物化判斷除了可以視為是一個詞類標記的問題，事實上它也可視為是一個二元分類的問題。根據本論文第二章所做的系統描述，當輸入是一個動詞時(如：合作)，系統要作的就是根據這個動詞的上下文，判斷這個動詞屬於哪一類-是單純謂語（如：合作(VH)，表示它未名物化且詞類為 VH)，還是派生名詞(如：合作(VH)[+nom]，表示它名物化且來源詞語法詞類為 VH)。

上述的分類法會碰到一個實際上的困難。我們知道上下文是很重要的分類依據，而上下文除了相鄰詞之外，相鄰詞的詞類標記也是很重要的線索，在只有相鄰詞而沒有相鄰詞的詞類標記的情況之下，建立統計模型時必定面臨資料稀疏(data sparseness)的問題。因此，我們將原始的詞類標記和名物化判斷分開處理，也就是說，先將輸入的句子以傳統的隱藏式馬可夫模型決定其每個詞的語法詞類標記，之後再針對其中每個動詞，分析其本身的名物化可能性以及包含相鄰詞詞類標記的上下文，作名物化判斷。這個流程的前半段是傳統的詞類標記問題，後半段是一個二元分類問題，即只有兩個分類類別，包含[+nom]或未包含[+nom]。以下舉例說明整個流程:

輸入：學生 的 不 合作
經過 HMM 後：學生(Na) 的(DE) 不(D) 合作(VH)
經過分類器後：學生(Na) 的(DE) 不(D) 合作(VH) [+nom]

我們提出三種不同類型的分類器，區分動詞來源詞但未區分其語法詞類的模型，稱之為 Classifier-1，未區分動詞來源詞但區分其語法詞類的模型，稱之為 Classifier-2，區分動詞來源詞且區分其語法詞類的模型，稱之為 Classifier-3。之所以設計這三種不同的分類器，主要目的除了希望得到一個最佳表現的分類器之外，也想藉此了解動詞來源詞的語法詞類是否在名物化判斷上扮演重要的角色。另外，跟隱藏式馬可夫模型不同的是，這三個分類器不僅考慮了判斷標的之前後語法詞類，也考慮了判斷標的之前後詞。

Bayes decision rule:
Given a verb v_i, *its tag* t_i and its *context* $c(v_i)$
if $P(v_i(t_i)[+nom] \mid v_i, t_i, c(v_i)) > P(v_i(t_i) \mid v_i, t_i, c(v_i))$, choose $v_i(t_i)[+nom]$
else choose $v_i(t_i)$

$$
\begin{aligned}
&P(v_i(t_i)[+nom] \mid v_i, t_i, c(v_i)) \\
&= \frac{P(v_i, t_i, c(v_i) \mid v_i(t_i)[+nom])}{P(v_i, t_i, c(v_i))} \times P(v_i(t_i)[+nom]) \\
&\cong P(v_i, t_i, c(v_i) \mid v_i(t_i)[+nom]) \times P(v_i(t_i)[+nom]) \\
&= P(c(v_i) \mid v_i(t_i)[+nom]) \times P(v_i(t_i)[+nom]) \\
&\cong \log(P(c(v_i) \mid v_i(t_i)[+nom])) + \log(P(v_i(t_i)[+nom])) \\
&= \log(P(\{w_{i-1}, t_{i-1}, w_{i+1}, t_{i+1}\} \mid v_i(t_i)[+nom])) + \log(P(v_i(t_i)[+nom])) \\
&= \log(P(w_{i-1} \mid v_i(t_i)[+nom])) + \log(P(t_{i-1} \mid v_i(t_i)[+nom])) + \\
&\quad \log(P(w_{i+1} \mid v_i(t_i)[+nom])) + \log(P(t_{i+1} \mid v_i(t_i)[+nom])) + \\
&\quad \log(P(v_i(t_i)[+nom])) \\
&= \log(\alpha P(w_{i-1} \mid v_i(t_i)[+nom]) + \beta P(w_{i-1} \mid (t_i)[+nom]) + \gamma P(w_{i-1} \mid v[+nom])) + \\
&\quad \log(\alpha P(t_{i-1} \mid v_i(t_i)[+nom]) + \beta P(t_{i-1} \mid (t_i)[+nom]) + \gamma P(t_{i-1} \mid v[+nom])) + \\
&\quad \log(\alpha P(w_{i+1} \mid v_i(t_i)[+nom]) + \beta P(w_{i+1} \mid (t_i)[+nom]) + \gamma P(w_{i+1} \mid v[+nom])) + \\
&\quad \log(\alpha P(t_{i+1} \mid v_i(t_i)[+nom]) + \beta P(t_{i+1} \mid (t_i)[+nom]) + \gamma P(t_{i+1} \mid v[+nom])) + \\
&\quad \log(P(v_i(t_i)[+nom]))
\end{aligned}
$$

Classifier-1: $\alpha = 0.8, \beta = 0, \gamma = 0.2$

Classifier-2: $\alpha = 0, \beta = 0.8, \gamma = 0.2$

Classifier-3: $\alpha = 0.5, \beta = 0.3, \gamma = 0.2$

The calculation $P(v_i(t_i) \mid v_i, t_i, c(v_i))$ is similar to $P(v_i(t_i)[+nom] \mid v_i, t_i, c(v_i))$

4. 實驗

為了比較上述模型的表現，本章提出我們的實驗方法。

4.1. 實驗環境

我們以詞庫小組所開發的平衡語料庫作為訓練以及測試的材料，這個語料庫是一個帶語法詞類標記的語料庫，並且利用附加名物化特徵的方式來呈現派生名詞(未附加名物化特徵的動詞即表示此動詞扮演單純的謂語角色)

如：他(Nh) 無法(D) **忍受(VK)** 學生(Na) 的(DE) 不(D) **合作(VH)[+nom]**

若以這個例子作為測試句，系統的輸入和輸出如下：

輸入：他 無法 **忍受** 學生 的 不 **合作**

輸出：他(Nh) 無法(D) **忍受(VK)** 學生(Na) 的(DE) 不(D) **合作(VH)[+nom]**

輸入是一個正確的斷詞結果(因為從語料庫得到)。輸出是帶語法詞類標記並附加名物化特徵的詞串，理想的輸出結果跟語料庫所標注的內容應該完全一樣，比較兩者即可評估模型的表現。表 1 是語料庫的詳細資料，“總詞數”表示語料庫中所有詞的個數總和，“動詞詞數”表示語料庫中所有動詞(即未名物化以及名物化的動詞)的個數總和，“名物化動詞詞數”表示語料庫中所有

名物化的動詞的個數總和，“名物化比率”為“名物化動詞詞數”除以“動詞詞數”的值。

表 1. 平衡語料庫的名物化情形

總詞數(W#)	動詞詞數(V#)	名物化動詞詞數(Nv#)	名物化比率(RofNv)
5821222	1205208	90951	7.5%

我們將平衡語料庫按照主題切分成六大類-文學、生活、社會、科學、哲學以及藝術。在實驗當中我們會個別測試這六大主題及其綜合表現。表 2 則顯示各主題的名物化情形。

表 2. 六大主題的名物化情形

	W#	V#	Nv#	RofNv
文學	939656	203863	5974	2.9%
生活	1061336	212387	14062	6.6%
社會	2330536	486901	47759	9.8%
科學	399934	77828	7739	9.9%
哲學	527602	112703	4869	4.3%
藝術	562158	111526	10548	9.5%

除了按照主題切分之外，我們也按照語式來切分平衡語料庫，在實驗當中我們會個別測試不同的語式-書面語、口語及其綜合表現。表 3 則顯示各語式的名物化情形。

表 3. 不同語式的名物化情形

	W#	V#	Nv#	RofNv
書面語	5193355	1086887	87130	8.0%
口語	627867	118321	3821	3.2%

藉由主題以及語式的分類，讓我們對不同主題和語式的名物化情形有所了解。從系統評估的角度來說，綜合評估之外再加上個別評估可以反映出現實中的不同需求及應用，測試出模型的強健性。

我們將平衡語料庫的 80%當作訓練語料，20%當作測試語料。訓練語料和測試語料當中的主題分佈比率或是語式分佈比率都跟原平衡語料庫相同。也就是說，以文學類為例，它的訓練語料當中，有四倍於其測試語料的文學類訓練語料，以及其餘五類的訓練語料。

4.2. 評量標準

實驗所關注的焦點是動詞名物化的判斷是否正確，我們以派生名詞召回率(recall)、準確率(precision)以及綜合兩者的 F-Score 來評量。

Recall(R)=Nv_Match# / Nv#

Precision(P)= Nv_Match# / Result_Nv#

F-Score(F)=2*R*P/(R+P)

上式的 Nv#是參考語料的名物化動詞詞數，Result_Nv#表示輸出的派生名詞個數，Nv_Match#表示“派生名詞相符”的個數。上述所謂“派生名詞相符”，指的是針對某一個動詞，參考語料和輸出結果都標明它是派生名詞。

4.3. 實驗結果

表 4. 不同主題的 HMM 測試結果

	HMM-1			HMM-2		
	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)
文學	47.7	62.2	54.0	52.3	66.0	58.3
生活	59.4	66.3	62.7	61.3	67.1	64.1
社會	69.4	67.3	68.3	71.6	69.4	70.5
科學	63.4	54.3	58.5	65.8	56.7	60.9
哲學	59.4	58.6	59.0	62.5	59.4	60.9
藝術	64.6	67.8	66.2	67.1	69.2	68.2
綜合	65.9	65.7	65.8	68.3	67.7	68.0

表 5. 不同主題的 Classifier 測試結果

	Classifier-1			Classifier-2			Classifier-3		
	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)
文學	53.6	75.7	62.8	50.7	73.1	59.9	51.2	79.1	62.1
生活	65.0	76.2	70.2	63.7	75.7	69.2	63.8	79.6	70.9
社會	75.9	75.6	75.7	76.3	75.9	76.1	75.7	78.2	76.9
科學	68.9	62.4	65.5	70.3	60.7	65.1	70.0	64.7	67.3
哲學	63.5	68.1	65.7	62.0	66.7	64.3	60.0	69.2	64.3
藝術	71.9	72.2	72.1	71.9	73.5	72.7	71.4	74.7	73.0
綜合	72.3	74.0	73.1	72.3	73.9	73.1	71.8	76.6	74.1

圖 1. 不同主題的 HMM 以及 Classifier 測試結果

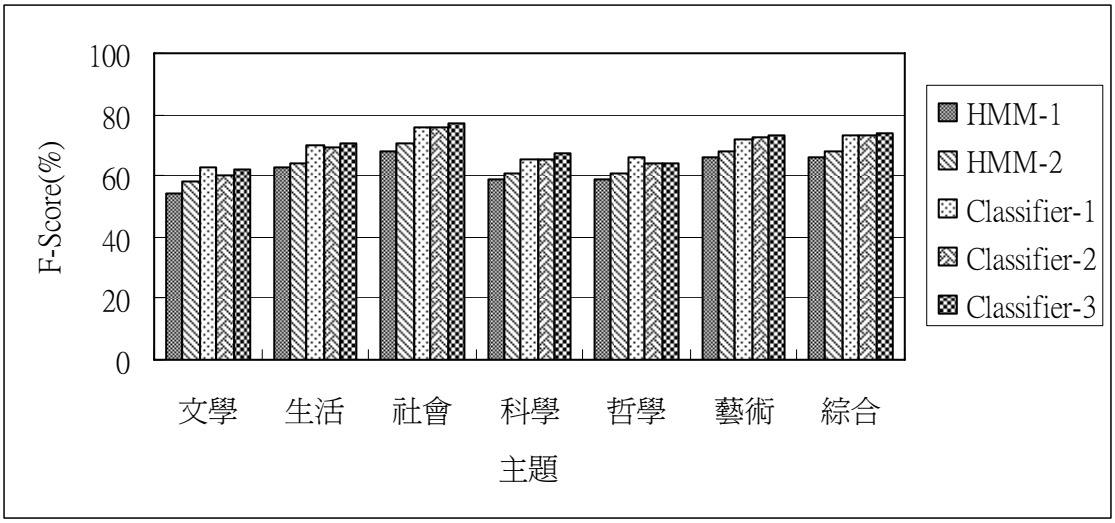


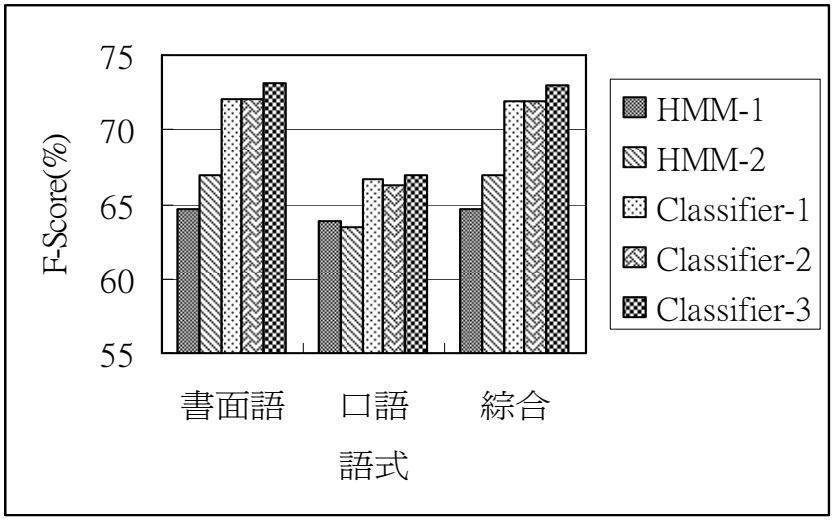
表 6. 不同語式的 HMM 測試結果

	HMM-1			HMM-2		
	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)
書面語	63.8	65.6	64.7	65.9	68.0	66.9
口語	63.4	64.1	63.8	65.0	61.9	63.4
綜合	63.8	65.6	64.6	65.9	67.8	66.9

表 7. 不同語式的 Classifier 測試結果

	Classifier-1			Classifier-2			Classifier-3		
	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)
書面語	69.5	74.7	72.0	70.1	74.0	72.0	69.3	77.4	73.1
口語	65.7	67.8	66.7	64.4	68.2	66.3	64.2	69.7	66.9
綜合	69.4	74.5	71.9	70.0	73.9	71.9	69.2	77.3	73.0

圖 2. 不同語式的 HMM 以及 Classifier 測試結果



4.4. 實驗討論

從主題的角度來看，表 2 顯示出，社會、科學以及藝術的名物化比率較高(分別為 9.8%、9.9%以及 9.5%)。從表 4 和表 5 我們可以大致歸納出名物化比率和測試結果的關係：無論是隱藏式馬可夫模型或是貝式分類器，在測試結果上都顯示出社會、藝術類擁有最佳的測試結果(分占第一、二名)，顯示出名物化比率越高，名物化的判斷越準確。例外的是科學類，雖然名物化比率(9.9%)比生活類(6.6%)高，但是所有模型的測試結果都顯示出科學類的判斷結果輸給生活類，而大致居於第四名的位置，這極可能跟語料的大小有關係(生活類的語料大約是科學類的 2.65 倍)。從計算語言學的角度來看，當我們能夠從訓練語料取得越多的判斷線索或是相關例證時，系統自然較能夠利用這些出現過的資料來判斷標的。因此，對統計式的名物化判斷模型來說，較高的名物化比率以及較大的語料會得到較準確的判斷結果。根據表 3、表 6 以及表 7，我們發現語式也有著一樣的關係:擁有較高名物化比率以及較大語料的書面語比低名物化比率以及較小語料的口語有

著更準確的判斷結果。

由以上的討論出發，若是我們想提升綜合、整體的名物化判斷結果，一個簡單直接的途徑就是增加那些低名物化比率的訓練語料，例如：文學類(名物化比率 2.9%)、口語類(名物化比率 3.2%)，來改善它們的判斷結果，而使綜合的判斷結果也跟著提升。

4.4.1 HMM-1 vs HMM-2

HMM-1 和 HMM-2 唯一的差別在於 HMM-1 未區分動詞來源詞語法詞類，而 HMM-2 區分了動詞來源詞語法詞類(請見 3.1 節)。對每一個動詞的上下文來說，HMM-1 是假設一個特定的上下文對每一個不同的來源詞都具有同樣的名物化催化力，HMM-2 則是假設一個特定的上下文對同一語法詞類的不同來源詞具有同樣的名物化催化力，而對不同語法詞類的來源詞具有不一樣的名物化催化力。

由表 4 及表 6 可以發現，HMM-2 大體上在每一個參考語料下的每一個評量指標(R、P、F)都比 HMM-1 來得高(唯一例外是口語的 P 及 F)。這告訴我們 HMM-2 的假設應該是比較正確的。來源詞語法詞類在名物化判斷上面的確扮演了重要的角色。

這樣的實驗結果也可以從語言學上找到解釋，[葉美利等 1992]以及[Huang et al., 1994]都指出：一個動詞，無論是單純謂語或者是派生名詞，其論元成分往往是一致的。在上下文中可以找到相對應的論元成分。由表 4 的實驗結果和實際的語料觀察，我們進一步地發現某些動詞的語法詞類，他們無論是扮演單純謂語角色或是轉換成派生名詞，相對應的論元成分往往在上下文中也具有相當固定的位置。舉例來說，動作不及物述詞(VA)，當它由單純謂語轉換成派生名詞時，單純謂語的主事者(agent)會從原本單純謂語的前面跑到派生名詞的後面緊鄰位置，當名詞片語的中心語(head word)，作為派生名詞的修飾對象。

如：學生(Na)　示威(VA) → 示威(VA)[+nom] 學生(Na)
　　學生(Na)　違規(VA) → 違規(VA)[+nom] 學生(Na)

另一方面，由於動作不及物述詞在單純謂語時，在語法上不緊鄰任何受詞，所以綜合上述，當我們看到一個 VA，其後緊鄰詞彙“學生”，幾乎可以斷定這個 VA 會轉換成派生名詞當“學生”的修飾語。

動作單賓述詞(VC)又是不同的情況，它在單純謂語時的主事者通常不會成為它在轉換成派生名詞後的修飾對象。

如：學生(Na)　攻擊(VC) 警察(Na) ╳→ 攻擊(VC)[+nom] 學生(Na)

“攻擊(VC)[+nom] 學生(Na)”是不合中文語法的。原因是，VC 在單純謂語時通常會後接一受詞，當我們看到一個 VC，其後緊鄰詞彙“學生”，通常會認為這個 VC 是單純謂語，而學生是它的受詞。也就是說，為了避免歧異，中文似乎不容許 VC 由單純謂語轉換成派生名詞去修飾原本單純謂語時的主事者。

因此，上述的觀察可以解釋為何同樣的上下文(如：後接“學生”)，針對相同來源詞語法詞類的不同動詞(如“示威(VA)”和“違規(VA)”)，有著相似的名物化催化力，而對不同的來源詞語法詞類(如：VA 和 VC)具有不同的名物化催化力。

4.4.2 HMM vs Classifier

由圖 1 和圖 2 可以看出 Classifier 模型均比 HMM 模型來得好。

HMM 的優點是它的目的是求整體最佳詞類序列，能夠將相鄰詞類之間的互相影響都考慮進去。但是缺點是未使用相鄰詞彙的訊息。Classifier 則正好相反，它的優點是它除了使用相鄰詞類作為上下文的元素之外，更把相鄰的詞一併考慮進來，使得所利用的上下文訊息更加豐富，而缺點是它是針對一個個的判斷標的，獨自得到個別的最佳判斷結果。而不是整體的最佳解。

由實驗結果看來，上下文的詞彙提供了名物化判斷更準確的鑑別能力以及更佳的強健性，在名物化判斷上是不可或缺的關鍵特徵。我們目前所提出的 Classifier 模型都只有使用左右相鄰(window size=1)的詞彙和詞類，未來還可再進一步觀察離判斷標的(window size>=2)較遠的詞彙及詞類對名物化判斷的影響。

4.4.3 Classifier-1 vs Classifier-2 vs Classifier-3

由表 5 和表 7 可以發現 Classifier-1 和 Classifier-2 的效果相當類似。

理論上來說，Classifier-1 應該比 Classifier-2 的表現好很多才對，因為在使用同樣上下文特徵的情況之下，Classifier-1 利用了來源詞本身這個更加精細的訊息，而 Classifier-2 卻只有使用來源詞語法詞類這個較粗糙的訊息。但是兩種模型所顯示的結果卻只有極些微的差距。事實上，這樣的結果再次印證了 4.4.1 節的討論。也就是說，一個特定的上下文對同一語法詞類的不同來源詞具有相似的名物化催化力，而對不同語法詞類的來源詞具有不一樣的名物化催化力。所以當我們用來源詞語法詞類代替來源詞本身時，其表現仍然維持幾乎一樣的水準。

因此我們可以推論，來源詞語法詞類可以提供一定程度的鑑別力以及優秀的強健性，來源詞本身則可以提供一定程度的強健性以及優秀的鑑別力，當我們把這兩樣訊息整合在同一個模型 Classifier-3 的時候，即得到了最好的表現結果以及最佳的強健性。

5. 結論

在本論文中，我們提出了兩種 HMM 模型以及三種貝氏分類器作為自動名物化判斷的系統，針對不同模型的表現，我們分析了表現差異的因素並從語言學上的角度加以驗證，其中，最值得注意的是，我們發現派生名詞的動詞來源詞(verbal counterpart)的語法詞類和其語境有密不可分的關係。藉著實際語料的分析，我們發現某些動詞語法詞類，他們無論是扮演單純謂語角色或是轉換成派生名詞，相對應的論元成分的位置在上下文中往往也遵循了某種固定模式。因此，當我們整合這個訊息到模型當中，即會使得系統的表現有顯著提升。表現最佳的統計模型對於派生名詞的包含率為 71.8%，準確率為 76.6%，F-Score 為 74.1%。

6. 未來研究方向

針對自動名物化判斷的未來研究主要有兩大方向，第一個方向是進一步使用範圍更大的語境資訊（如 window size 的加大）或者是更細緻的語法、語意資訊（如使用語意類別）。由於我們已經知道派生名詞的來源詞語法詞類在判斷上可以扮演關鍵角色，我們很好奇派生名詞的來源詞語意訊息(如語意類別)或者是詞構，是否也是重要的判斷因素。

另一個方向是當我們對自動名物化判斷的影響因素已經了解清楚後，最新的統計式技術，如“條件式隨機場域”(Conditional Random Field)或是“支援向量機器”(Support Vector

Machine)等，就可以嘗試來解這樣的問題。藉由最佳化的機器學習技巧，應該可以更準確的求得這些因素相互之間的關係以及使用比重。除了得到更佳的判斷結果之外，也可據此驗證或挖掘出更深入的名物化相關的語言現象。

7. 參考文獻

1. 葉美利、湯志真、黃居仁、陳克健,“漢語的動詞名物化初探—漢語中帶論元的名物化派生名詞, ROCLING V, pp177~193, 1992
2. 洪偉美、黃居仁、湯志真、陳克健,“中文派生詞的構詞規律初探”,第三屆世界華文教學研討會, 1991
3. Huang, Chu-Ren, Meili Yeh, and Li-Ping Chang, “A Corpus-based Study of Nominalization and Verbal Semantics: Two Light Verbs in Mandarin Chinese”, Proceedings of the Sixth North American Conference on Chinese Linguistics. Los Angeles: GSIL, USC. pp. 106-120, 1994.
4. Lin, Koong H.C., Von-Wun Soo, and Sandiway Fong, “Dealing with Nominalizations in Mandarin Chinese Using a Principles and Parameters Parser”, Computer Processing of Oriental Languages 11(3). pp. 291-307, 1998.
5. “中央研究院平衡語料庫的內容與說明”詞庫小組技術報告 95-02/98-04
6. Jane Grimshaw, Argument Structure, the MIT Press, 1990
7. Huang, Chu-Ren. “Mandarin Chinese NP de -- A Comparative Study of Current Grammatical Theories”, Special Publication No. 93 of the Institute of History and Philology, Academia Sinica. 1989.
8. Tsai, Yu-Fang and Keh-Jiann Chen, "Reliable and Cost-Effective Pos-Tagging", Proceedings of ROCLING XV, pp. 161-174, 2003
9. Tsai, Yu-Fang and Keh-Jiann Chen, "Context-rule Model for POS Tagging", Proceedings of PACLIC 17, pp. 146-151, 2003
10. Tsai, Yu-Fang and Keh-Jiann Chen, 2004, "Reliable and Cost-Effective Pos-Tagging", International Journal of Computational Linguistics & Chinese Language Processing, Vol. 9 #1, pp. 83-96, 2004.

大規模詞彙語意關係自動標記之初步研究: 以中文詞網 (Chinese Wordnet) 為例

謝舒凱 Petr Šimon 黃居仁

中研院語言學研究所

{shukai, petr.simon, churenhuang}@gmail.com

Abstract

近年來, 以知識資源爲本的自然處理技術已成爲一種重要的研究取向。對於各種詞彙語意資源之建構, 包括電子辭典 (Lexicon)、同義詞詞林 (Thesaurus)、詞彙網路 (WordNet), 甚至知識本體 (ontologies), 已成爲一個不可抵擋的趨勢。其中, 詞彙網路是在計算語言學相關領域中, 目前最爲普遍利用之一項詞彙語意資源。

然而, 詞彙網路之建構是一項耗時費力之基礎工程。對於世界上許多使用頻度不高的語言而言, 更是一項艱鉅之任務。本文提出一個借力於普林斯頓英語詞網 (Princeton WordNet) 與歐語詞網 (EuroWordNet) 之 bootstrapping 方法, 應用在正在發展的中文詞網詞彙語意關係之自動標記工作上。實驗的結果與初步評估證明, 此法對於詞網建構是一個相當可行的方式。

1 前言

近年來, 以知識資源爲本的自然處理技術已成爲一種重要的研究取向。對於各種詞彙語意資源之建構, 包括電子辭典 (Lexicon)、同義詞詞林 (Thesaurus)、詞彙網

路 (WordNet), 甚至知識本體 (ontologies), 已成爲一個不可抵擋的趨勢。其中, 詞彙網路更已成爲計算語言學相關領域中, 最爲普遍利用之一項標準 (de facto) 詞彙語意資源。

詞彙網路是以同義詞集 (synset), 以及詞彙語意關係 (lexical semantic relation) 所架構出的詞彙知識系統。也就是說, 詞彙網路架構表達的不僅是詞彙本身的概念性知識, 它亦表達了詞彙之間的語意關係。然而, 從普林斯頓英語詞網以及歐語詞網 (EuroWordNet) 的建構經驗來看, 這是一項費時耗力的龐大語言工程。對於經費取得困難、使用頻度較低之語言而言, 建立此項語言資源更爲不易。從詞彙語意與知識表達的角度觀察, 我們認爲不同語言對於**概念原素** (conceptual atoms), 可能有著不同之表達方式, 但是在**詞彙語意關係**的表達上, 則應具有更大程度之「普同性」。因此「借力」於已發展成熟之英語、歐語詞網之語意關係, 以加速新的詞網雛形成形, 就成了一個自然而然的另類選擇。

基於以上動機, 本文之組織如下: 第二節描述目前有關不同之詞彙語意關係所採用之各種自動、半自動判定演算法。在第三節中, 我們提出了一種便捷而有意義的方法與實驗設計。文章之第四節討論實驗結果之評價工作, 最後一節則鋪陳我們的結論與未來的展望。

2 詞彙語意關係之自動判定法

近年來, 關於自動判定或學習詞彙語意關係的研究文獻越來越豐富。爲了重點強調之方便, 本文將之粗分爲兩大走向: 利用單語資源, 與利用雙語或多語資源之研究。

2.1 使用單語資源

此類方法大多以語料庫及網頁資料爲主, 利用詞彙語法模式 (lexical-syntactic patterns) 或是「叢集」(cluster) 來抽取詞彙語意關係。目前這樣的作法之所忽略的兩

個問題是: (1). 所抽取之語意關係不夠全面, 大半部分還是受限在少數之關係上, 例如 is-a, part-of 等等。[1] (2).所需要的學習實例 (seed instances) 太多。最近 (Pennacchiotti and Pantel 2006) 所提的 Espresso 演算法, 欲針對此兩個問題提出新解法, 但是在評價上尚未完整, 仍有待觀察進一步之發展。

2.2 使用雙語或多語資源

這個方向包括使用雙語語料庫 (Diab 2004); 利用同義詞集之雙語對譯 (bilingual correspondences)[2] 直接藉助於以存在之詞網 (通常係普林斯頓詞網)。後者已有相當多之探討 (Pianta, et al 2002; Huang et al 2002, 2003, 2005), 西班牙詞網與義大利之 MultiWordNet 計畫皆是此想法下之實作產物。

3 實驗設計與方法

基於前面之文獻討論, 本文認爲, 著眼於當前多語處理之需求, 在策略上, 應該採取**先求同再求異**。因此, 借力於已存在之多語資源應該是第一步的工作。

3.1 我們提出之 Model: Bootstrapping from Multilingual Wordnets

本文提出的模型, 是基於 (Huang et al. 2002, 2003, 2005) 的擴充版本。先前之文獻, 已就借力於其他成形之詞網的跨語詞義關係預測所涉及之邏輯條件加以闡述, 在 (Huang et al 2003) 中, 並曾針對 210 個中文詞形 (lemma) 做過小規模之試驗與評價。本文則接續之前的基礎, 進一步在**規模**上與**多語擴充**兩個面向上作延伸試驗 。亦即, 在規模上, 我們將目前在中研院中文詞網小組所定義完成之七千多筆

[1]晚近亦有處理所謂 Textual Entailment 之關係。

[2]但是, 在此我們同時必須先理解到, 一組雙語對譯詞不一定是「同義關係」。此外, 它們可能在各自語言系統中與不同的詞/synset 間有不同的詞彙語意關係。

中文同義詞集爲主; 在多語擴充上, 我們將歐語詞網 (Vossen 1998) 亦納入實驗對象。其中包括了德語、法語、捷克語、荷語、西語、義語與愛沙尼亞語等七種歐洲語言。

本文提出之 model 在方法論上有兩個意涵:

- 對於多語之詞網發展, 可提供一個**符應** (correspondence) 與**協作** (collaborative) 架構: 在不同語種之詞網之間建立符應關係, 是多語知識處理工作之一重要環節。我們認爲, 這種符應性應該是表現在詞彙語意關係上, 而非在上層知識本體 (top ontology) 或詞彙翻譯上 (word translations)。此外, 在資料的標誌與管理上, 我們亦採用了全球詞網協會 (Global WordNet Association) 所建議之 XML/Schema 格式, 以便於將來國際詞網網格 (global wordnet grid) 環境架構。

- 另一方面, 此 model 可作爲一個詞網之**快速原型** (rapid prototyping) 發展。加速詞網核心部分建構之過程。

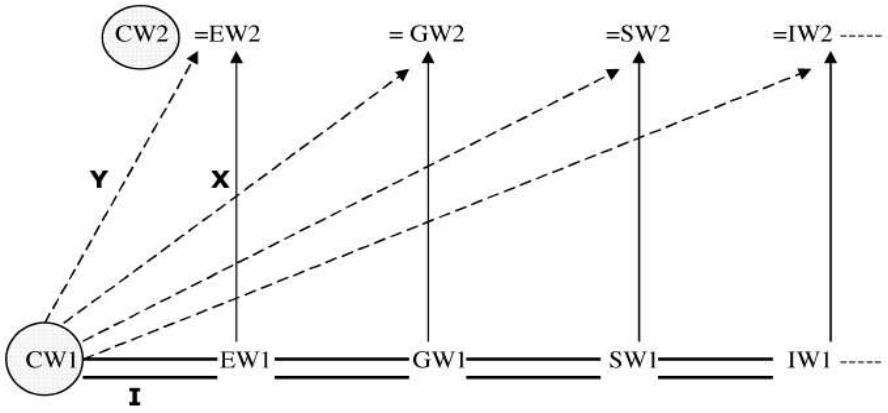


Figure 1: 本文提出之基於多語之大規模詞義關係抽取模型

Figure 1 中展示了這個模型。此模型是 (Huang et al 2002) 之擴充版本。圖中的變項 Y, I, X 間的關係, 可以用簡式來表達 $Y = I + X$ 。中文詞形 $CW1$ 與 $CW2$ 間的詞彙語意關係 Y 可被視爲是 I 與 X 的**機能結合** (functional combination)。最簡單的例子是, 如果 $CW1$ 與 $EW1$ 間的 I 是同義關係, 那麼英語詞形 $EW1$ 與 $EW2$ 間的語意關係 X , 就可被預測爲是 $CW1$ 與 $CW2$ 的關係。同理, 藉由 synset 之中介, 可進一步拓展到其他語種的詞網資源 (如德語、西語等等)。

不過, 就如之前所提到的, 一組雙語對譯詞不一定是「同義關係」。當不同關係 (例如上下位、部分與全體關係等) 涉入時, 我們就需要一組具邏輯性的自動推理規則。表格一則列出了我們所使用的詞彙語意關係邏輯推理規則。[3]

3.2 使用之詞彙資源

本實驗所需要使用的資源包括如下:

- **中文詞網之中頻詞形、詞義資料** 這是目前由中研院語言所中文詞網小組所完成之以中頻詞爲主的詞形(lemma) 及詞義區分資料。[4]
- **中英雙語** synset **對譯資料庫及評價標注** (Huang et al 2003)。此資料庫係由中央研究院與遠見科技股份有限公司共同開發。包括了以 WordNet 1.6 之 99,642 筆同義詞集 (synset) 爲基準之中文對應翻譯。
- **歐語詞網及普林斯頓** WordNet 1.5-1.6 **對應表**

3.3 步驟

- 生成中文詞網之 synset

  因爲正在發展中的中文詞網並無正式之 synset 格式設計, 我們採用 gloss

[3]詳細實例說明請參見 (Huang et al 2005)。

[4]網路展示版請見 http://cwn.ling.sinica.edu.tw

	I	X	Y	Bootstrapped Results
1	HYP	ANT	ANT	{CW1, ANTONYM, CW2}
2	HYP	HYP	HYP	{CW1, HYPONOMY, CW2}
3	HYP	NSYN	HYP	{CW1, HYPONYM, CW2}
4	HYP	HOL	HOL	{CW1, HOLONYM, CW2}
5	HYP	all other LSRs	undecided	?
6	HPO	ANT	ANT	{CW1, ANTONYM, CW2}
7	HPO	HPO	HPO	{CW1, HYPONYM, CW2}
8	HPO	NSYN	HPO	{CW1, HYPONYM, CW2}
9	HPO	MER	MER	{CW1, MERONYM, CW2}
10	HPO	all other LSRs	undecided	?
11	NSYN	ANT	ANT	{CW1, ANTONYM, CW2}
12	NSYN	HYP	HYP	{CW1, HYPERNYM, CW2}
13	NSYN	HPO	HPO	{CW1, HYPONYM, CW2}
14	NSYN	NSYN	NSYN	{CW1, NEAR-SYNONYM, CW2}
15	NSYN	MER	MER	{CW1, MERONYM, CW2}
16	NSYN	HOL	HOL	{CW1, HOLONYM, CW2}
17	HOL	ANT	ANT	{CW1, ANTONYM, CW2}
18	HOL	HYP	HYP	{CW1, HYPONYM, CW2}
19	HOL	NSYN	HOL	{CW1, HOLONYM, CW2}
20	HOL	HOL	HOL	{CW1, HOLONYM, CW2}
21	HOL	all other LSRs	undecided	?
22	MER	ANT	ANT	{CW1, ANTOMY, CW2}
23	MER	HPO	HPO	{CW1, HYPONYM, CW2}
24	MER	NSYN	MER	{CW1, MERONYM, CW2}
25	MER	MER	MER	{CW1, MERONYM, CW2}
26	MER	all other LSRs	undecided	?

Table 1: 詞彙語意關係邏輯推理規則

matching 的方式, 抽出、過濾並作流水編號。[5]

- 依照本文提出之模式從 WordNet 及 EuroWordNet 萃取詞彙語意關係

4 實驗結果與評價

4.1 基本數據與結果

以下則簡列出以 xml 格式標記的部分結果。

```
<synset id="00002517-x">
<gloss>預估費用並承諾以該費用履行合約。</gloss>
<variants>
<variant sense="01">估價</variant>
<variant sense="01">報價</variant>
</variants>
<ILIRelations>
<relation type="SYN" targetID="00692314-v"/>
<relation type="HYP" targetID="01529684-v"/>
<relation type="HPO" targetID="00692437-v"/>
</ILIRelations>
<internalRelations>
<relation type="HYP" targetID="01529684-v"/>
<relation type="HPO" targetID="00692437-v"/>
</internalRelations>
</synset>
<synset id="00002004-x">
<gloss>形容正常的, 只用於疑問句或否定句, 假設其不正常。</gloss>
<variants>
<variant sense="02">對</variant>
</variants>
```

[5]目前已發展了更有意義的編碼方式, 目前的設計純粹爲了實驗目的之故。在此過程中, 剛好附帶的作了詞網品管 Quality Control。包括處理了幾個問題: (1). 不同的中文 synsets 卻有相同的 synset offset; (2). 相同的 gloss 卻有不同的普林斯頓詞網 offset; (3). 類似錯誤的註解。如: 下列註解可能是相同的。

- 將前述對象排除或已知前述對象包含在所屬範圍。常用 “除... 以外”。
- 將前述對象排除或已知前述對象包含在所屬範圍。常用 “除... 之外”。
- 將前述對象排除或已知前述對象包含在所屬範圍。常用 “除..... 外”。

```
<ILIRelations>
<relation type="SYN" targetID="00138021-a"/>
<relation type="x_similar" targetID="00137150-a"/>
</ILIRelations>
<internalRelations>
<relation type="x_similar" targetID="00137150-a"/>
</internalRelations>
</synset>
<synset id="00001749-x">
<gloss>普通名詞。憤怒的情緒。</gloss>
<variants>
<variant sense="05">氣</variant>
</variants>
<ILIRelations>
<relation type="SYN" targetID="05587878-n"/>
<relation type="HYP" targetID="05560878-n"/>
<relation type="HPO" targetID="05588413-n"/>
<relation type="HPO" targetID="05588725-n"/>
<relation type="HPO" targetID="05588822-n"/>
<relation type="HPO" targetID="05588960-n"/>
<relation type="HPO" targetID="05589074-n"/>
<relation type="HPO" targetID="05589169-n"/>
<relation type="HPO" targetID="05589301-n"/>
<relation type="HPO" targetID="05589430-n"/>
</ILIRelations>
<internalRelations>
<relation type="HYP" targetID="05560878-n"/>
<relation type="HPO" targetID="05588413-n"/>
<relation type="HPO" targetID="05588725-n"/>
<relation type="HPO" targetID="05588822-n"/>
<relation type="HPO" targetID="05588960-n"/>
<relation type="HPO" targetID="05589074-n"/>
<relation type="HPO" targetID="05589169-n"/>
<relation type="HPO" targetID="05589301-n"/>
<relation type="HPO" targetID="05589430-n"/>
</internalRelations>
</synset>
```

4.2 評價

對結果之評價上, 我們採取蔡(2002) 所提出之中文詞義關係的判定原則, 並參酌 EuroWordNet 技術報告。爲了便利評估, 我們亦發展了一個簡便之評價系統。Figure 2 是此系統之操作介面。表格三則列出人工評價結果。[6]

[6]因資料龐大, 目前歐語詞網部分僅是目前跑出之部分結果。

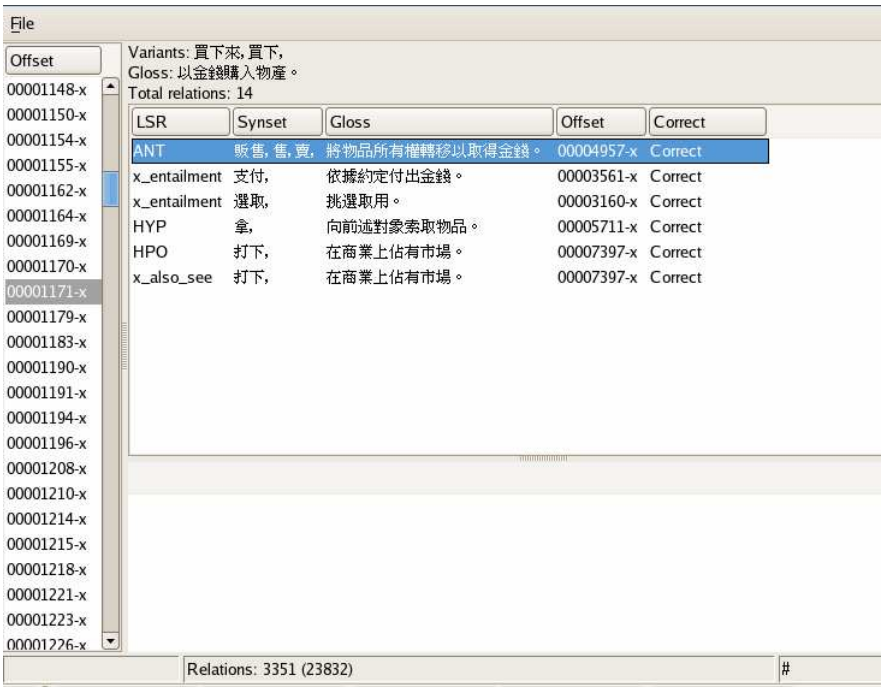


Figure 2: 檢測工具之圖形介面

我們由此得到的準確率如下:

$$Precision\ \ rate_{pwn} = \frac{3057}{3252} * 100\% = 94\%$$

$$Precision\ \ rate_{ewn} = \frac{2673}{2981} * 100\% = 89\%$$

	correct	incorrect	other	total retrieved with Chinese syn.	total retrieved
英語詞網	3057	124	71	3252	23832
歐語詞網	2673	210	98	2981	19711

Table 2: 評價結果

4.3 錯誤分析

從結果上看, 我們可說此法是一個相當可靠的方式。從各項語意關係類型之錯誤分佈來看, 我們的觀察結果大致與先前之小規模試驗相似。預測之正確率由高到低的排列大概都是 { MPT, x _ similair to, x _ pertainym, x _ also _ see ... } ≻ ANT ≻ HYP ≻ HPO ≻ { OTHER RELATIONS }。詳細之數據留待完整實驗結果再行分析。此外, 此種理論模式要面臨到的主要難題, 包括詞彙的缺隔 (lexical gaps) (同一個概念在甲語言中與乙語言所使用的詞彙表達單位不同)、指稱差異 (denotation differences) (對等翻譯存在, 但是概念的抽象度卻不同) 等等, 如何在技術上克服這些非同義關係對應的語言現象, 值得往後進一步細究探討。

5 結論與未來展望

按我們的設想, 詞網之語意關係自動建構可以分由**全域的** (global) 與**區域的** (local) 特徵兩個面向來著手。本文展示了利用全域特徵之可行性, 並由此建立了中文詞網詞彙語意關係網路之雛形。在此基礎上, 我們下一步將細究在地詞網的特徵。包括利用詞彙模式 (lexical patterns) 從辭典釋意 (gloss)、語料庫等資源萃取各種詞彙語意關係, 亦包括利用漢語之構詞語意模式與漢字知識本體資源 (例如 HanziNet/Hantology), 使機器自動學習。我們相信這個成套的半自動詞網自動建構方法, 亦可供其他語種建構詞網參考。

References

[1] 蔡柏生、黃居仁等 (2002). 中文詞義關係的定義與判定原則。*Journal of Chinese Information Processing 16.4*.

[2] Diab, Mona. (2004). The Feasibility of Bootstrapping an Arabic WordNet Leveraging Parallel Corpora and an English WordNet. *Proceedings of the Arabic Language Technologies and Resources*, NEMLAR, Cairo.

[3] Huang, Chu-Ren et al. (2002). Translating Lexical Semantic Relations: The First Step Toward Multilingual Wordnets. *COLING 2002*, Taipei.

[4] Huang, Chu-Ren et al. (2003). Cross-lingual Portability of Semantic Relations: Bootstrapping Chinese Wordnet with English WordNet Relations. *Language and Linguistics 4.3:509-532*.

[5] Huang, Chu-Ren et al (2005). Cross-lingual Conversion of Lexical Semantic Relations: Building Parallel Wordnets.

[6] Pennacchiotti, Marco and Patrick Pantel. (2006). A Bootstrapping Algorithm for Automatically Harvesting Semantic Relations. *LREC 2006.* Italy.

[7] Vossen, P. (ed). (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Dordrecht: Kluwer Academic.

Automatic Learning of Context-Free Grammar

Tai-Hung Chen Chun-Han Tseng
M9430400{71,41}@student.nsysu.edu.tw
Chia-Ping Chen
cpchen@cse.nsysu.edu.tw
Department of Computer Science and Engineering
National Sun Yat-Sen University

Abstract

In this paper we study the problem of learning context-free grammar from a corpus. We investigate a technique that is based on the notion of minimum description length of the corpus. A cost as a function of grammar is defined as the sum of the number of bits required for the representation of a grammar and the number of bits required for the derivation of the corpus using that grammar. On the Academia Sinica Balanced Corpus with part-of-speech tags, the overall cost, or description length, reduces by as much as 14% compared to the initial cost. In addition to the presentation of the experimental results, we also include a novel analysis on the costs of two special context-free grammars, where one derives only the set of strings in the corpus and the other derives the set of arbitrary strings from the alphabet.

Index Terms: context-free grammar, Chinese language processing, description length, Academia Sinica Balanced Corpus.

1 Introduction and Overview

In this paper we study the problem of learning context-free grammar (CFG) [1] from a corpus of part-of-speech tags. The framework of CFG, although not complex enough to enclose all human languages [2], is an approximation good enough for many purposes. For a natural language, a "decent" CFG can derive most sentences in the language. Put differently, with high probability, a sentence can be parsed by a parser based on the CFG.

The main issue with CFG is how to get one. Generally speaking, learning context-free grammar from sample text is a difficult task. In [3], a context-free grammar which derives exactly one string is reduced to a simpler grammar generating the same string. This achieves a lossless data compression. In [4], an algorithm of time complexity $O(N^2)$ for learning stochastic context-free grammar (SCFG) is proposed, where N is the number of non-terminal symbols. This is a great reduction from the inside-outside algorithm which requires $O(N^3)$.

Context-free grammars can be used in many applications. In [5], an automatic speech recognition system uses a dynamic programming algorithm for recognizing and parsing spoken word strings of a context-free grammar in the Chomsky normal form. CFG can

also be used in software engineering. In [6], the components in a source code that need to be renovated are recognized and new code segments are generated from context-free grammars. In addition, since parsing outputs larger and less-ambiguous meaning-bearing structures in the sentence, for high-level natural language processing tasks such as question answering [7] and interactive voice response [8] systems, the design and implementation of CFG can be crucial to their success.

If the goal of learning is to acquire a grammar that derives most sentences in the domain of interest, then a good one is apparently domain-specific. An all-purpose CFG is not likely to be the best since it tends to derive a much larger set than is necessary. We thus propose to learn CFGs from corpus. The basic problem is this: *Given a set of sentences, we want to find a set of derivation rules that can derive the original set of sentences.* Note that there are infinitely many CFGs from which the original set of sentences can be derived. To discriminate one CFG from another, we will consider the costs they incur in deriving the original corpus. The cost functions will be defined shortly. Thus, we are proposing to *find the set of rules that can derive the original language with the minimum cost.*

This paper is organized as follows. Following this introduction and review, we analyze two special cases of CFG and the proposed rules in Section 2. The experimental results are presented in Section 3 followed by discussion and comments. In Section 4, we summarize our work.

2 Mathematical Analysis

2.1 The Cost Functions

There are two different kinds of costs in the description of a corpus by a CFG. The first kind is incurred from the representation of the CFG. A rule in a CFG is of the form

$$A \rightarrow \beta. \tag{1}$$

It consists of a non-terminal symbol A on the left-hand side and a string of symbols β on the right-hand side. The cost of a rule is the number of bits needed to represent the left-hand side and right-hand side. For (1), this is

$$C_R = (1 + |\beta|) \log |\Sigma|, \tag{2}$$

where Σ is the symbol set and $|\Sigma|$ is the number of symbols in Σ.

The second kind is the cost to derive the sentences given the rules. In order to derive a sentence W, the sequence of rules must be specified in the derivation from S [1] to W,

$$S \Rightarrow \alpha_1 \Rightarrow \cdots \Rightarrow W, \quad \text{or } S \stackrel{*}{\Rightarrow} W, \tag{3}$$

where we have adopted the notation defined in [1]. The sequence of rules always starts with one of the S-derivation rules[2],

$$S \rightarrow \alpha. \tag{4}$$

This step results in a derived string α. If there is no non-terminal symbols in α, we are done with the derivation. Otherwise, we expand the left-most non-terminal symbol, say X,

[1] S is known as the sentence symbol or the start symbol.

[2] The Z-derivation rules are those with Z as the left-hand side.

in α by one of its derivation bodies[3]. The process continues until there is no non-terminal symbols in the derived string, which will be the sentence W at that point. To illustrate, suppose we are given the CFG

$$\begin{cases} R_1(S): & S \to XXC \\ & \vdots \\ R_1(X): & X \to AB \\ & \vdots \end{cases}$$

and we want to derive the sentence $W = ABABC$. For this example, one can verify that the derivation sequence is $R_1(S)R_1(X)R_1(X)$, where $R_t(Z)$ represents the tth Z-derivation rule. The cost is

$$C_D = \sum_{k=1}^{m} \log |R(s_k)| = \log |R(S)| + \log |R(X)| + \log |R(X)|, \tag{5}$$

where m is the number of rules in the derivation sequence, s_k is the non-terminal symbol for the kth derivation, and $|R(s_k)|$ is the number of rules in the CFG using s_k as the left-hand side.

Combining (2) and (5), the total cost is

$$C = \sum_{i=1}^{p} C_R(i) + \sum_{j=1}^{q} C_D(j) = \sum_{i=1}^{p} n_i \log |\Sigma| + \sum_{j=1}^{q} \sum_{k=1}^{m_j} \log |R(s_k)|, \tag{6}$$

where p is the number of rules, q is the number of sentences, n_i is the number of symbol tokens in rule i, and m_j is the length of the derivation sequence for sentence j.

2.2 Special-Case Analysis

We will analyze the costs for two special CFGs in this section. The first CFG, which we call the *exhaustive* CFG, uses every distinct sentence in the corpus as a direct derivation body of the start symbol S. The corpus is thus covered trivially. To compute the cost, we first rearrange the sentences in the lexicographic order and then move the repeated sentences to the back. The number of symbols for a rule is simply the number of words of the corresponding sentence n_w, plus 1 (for the start symbol S), and $|\Sigma|$ is the vocabulary size $|V|$ of the corpus plus 1 (again for the start symbol). Thus the rule cost is

$$C_R = n \log |\Sigma| = (n_w + 1) \log(|V| + 1). \tag{7}$$

In this case, each sentence is derived from S in one step, by specifying the correct one out of the $|R(S)|$ rules. Thus the derivation cost for a sentence is

$$C_D = \log |R(S)|. \tag{8}$$

Note that q is generally not equal to $|R(S)|$ as there may be repeated sentences. Combining (7) and (8), the total cost for the exhaustive CFG is

$$C = \sum_{i=1}^{|R(S)|} C_R(i) + \sum_{j=1}^{q} C_D(j) = \sum_{i=1}^{|R(S)|} (n_w(i) + 1) \log(|V| + 1) + q \log |R(S)|. \tag{9}$$

[3]This is also known as the leftmost derivation.

The second case, which we call the *recursive* CFG, uses recursive derivation for S,

$$S \to AS, \tag{10}$$

where the non-terminal A can be expanded to be any word in the vocabulary. Combined with the rule $S \to \epsilon$, this CFG clearly covers any string of the alphabet, Σ^*, which is a much larger set than any real corpus.

The rule cost is significantly smaller in recursive CFG than that of the exhaustive CFG. The only rules are the two instances of S-derivation and the $|V|$ instances of A-derivation, so the rule cost is

$$C_R = n \log |\Sigma|, \tag{11}$$

where n can be $1, 2$ or 3 depending on the rule. The derivation cost, however, is much larger. To derive a sentence W of n_w words, the recursive rule of S and substitution rule of A have to be applied alternatively for n_w times, followed by a final rule of $S \to \epsilon$. Thus the derivation cost for a sentence is

$$C_D = n_w(1 + \log |V|) + 1. \tag{12}$$

Combining (11) and (12), the total cost for the recursive CFG is

$$\begin{aligned} C &= \sum_{i=1}^{2+|V|} n_i \log |\Sigma| + \sum_{j=1}^{q} C_D(j) \\ &= (4 + 2|V|) \log(|V| + 2) + \sum_{j=1}^{q} [n_w(j)(1 + \log |V|) + 1]. \end{aligned} \tag{13}$$

In Table 1 we list the costs of these cases computed on the Academia Sinica Balanced Corpus [9] (ASBC). The exhaustive CFG has a large rule cost (28.1 million bits) and a small derivation cost (4.1 mb). The recursive CFG has an extremely small rule cost (merely 607 bits) and an extremely large derivation cost (88.4 mb). To overall cost is higher for the recursive CFG (88.4 mb) than the exhaustive CFG (32.2 mb). From this table, one can see that there is a trade-off between the rule cost and the derivation cost. In addition, the numbers illustrate the important point that minimizing the rule cost alone will lead to a CFG that is inappropriate.

The exhaustive CFG is too restricted in the sense that it covers only those sentences seen in the learning corpus. The recursive CFG is too broad in the sense that it covers all sentences including the non-sense ones. Our goal is to strike a balance between these two extremes.

2.3 Proposed Rules

The special cases we analyze above do not have the minimum cost of all possible CFGs from which the corpus can be derived. To reduce the overall cost, we start with the initial CFG and then iteratively look for a new CFG rule. The kind of rules we investigate in this study is of the form

$$X \to YZ.$$

The introduction of such a rule to the exhaustive CFG described in Section 2.2 has the following impacts on the cost:

- Each occurrence of YZ is replaced by X, so the total number of symbol tokens in the S derivation rules is reduced.
- $|\Sigma|$ is incremented by 1.
- The derivation cost may or may not change, depending on whether two or more of the S-derivation rules become identical.

Since there are two symbols on the right-hand side, the number of candidate rules is $|\Sigma \times \Sigma| = |\Sigma|^2$, where Σ is the current symbol set. To choose one, we compute the bigram counts of all bigrams and use the bigram with the highest count as the right-hand side of the new rule, whose left-hand side is a new symbol.

3 Experiments

3.1 Data Preparation

We use the ASBC corpus for our experiments. In this corpus, the part-of-speech tag is labeled for each word. On the raw text data, we apply the following pre-processing steps:

1. The punctuation of period, question mark and exclamation mark are used to segment a sentence into multiple sentences.
2. The parenthesis tags are discarded.
3. The part-of-speech tag sequence is extracted for each sentence.

The initial statistics of the data after pre-processing is summarized in Table 2. A total of 229852 sentences are extracted and 203651 of them are distinct. The total number of tokens is 4.84 millions. Note that in the experiments, the symbols are the part-of-speech tags rather than the words for our CFG learning algorithm. This approach focuses more directly on the syntax and alleviates the issue of data sparsity.

3.2 Results

The learning process is an iterative algorithm. We start with the exhaustive CFG introduced in Section 2.2. In each epoch, we

1. compute the bigram counts for each bigram,
2. make a new rule with the bigram of the largest count as the right-hand side,
3. update the alphabet (symbol set), rules and derivations,
4. update the costs.

The representation cost as a function of the number of learned rules is presented in Figure 1. There are three curves in the plot, representing the rule cost, the derivation cost and the total cost. The initial cost is 32.2 million bits, as we show in Section 2.2. As the learning process progresses, the two kinds of cost behave in different ways: the derivation cost stays

constant while the rule cost decreases. The derivation cost is invariant for two reasons: 1) the number of S-derivation rules does not change and 2) there is no ambiguity in expanding non-S symbols, in our current learning scheme. The rule cost reduces because the decrease in the number of tokens in the rules outweighs the increase in the size of symbol set. As a result, the total cost reaches a minimum of 27.7 million bits when the 92nd rule is learned. The cost reduction is 14.0%. After the 92nd rule, the largest bigram count is not high enough for the reduction of the number of tokens to outweigh the increase in the alphabet, so the cost increases. The maximum bigram count is plotted against the epoch (number of rules learned) in Figure 2. From this figure, one can see that the maximum bigram count decreases very fast.

The top-20 rules learned from ASBC are listed in Table 3. In this table, we also include examples of words and sentences from ASBC. In addition, the definition and more examples of the part-of-speech tags are listed in Table 4. From Table 3, one can see that the new symbols (M1, ..., M20) here indeed represents larger phrasal structures than the basic part-of-speech tags. Furthermore, M7 and M9 embed M1, giving evidence for a deep parsing structure. In Figure 3, two sentences in ASBC parsed based on the learned CFG (left) and parsed manually (right) are shown. We can see that the verb phrase (VP) structure of sentence (a) in both parses. For sentence (b), the VP is scattered in two subtrees $M40$ and $M66$. The symbol $M66$ can be identified as a noun phrase (NP).

4 Summary

The construction of a context-free grammar for a specific domain is a non-trivial task. To learn a CFG automatically from corpus, we define a cost function as the number of bits for the representation of CFG and sentence derivation. Our objective is to find a grammar that covers the learning corpus with the minimum cost. We analyze two extreme cases to illustrate the framework. The proposed rules are learned from heuristic bigram counting. The results show that on ASBC corpus, the reduction of cost is 14.0% of the initial cost.

There are other kinds of CFG rules that are not considered in this study, such as the $A \rightarrow B|C$ rules. The candidate set of rules should be enlarged for more descriptive power. Another line of research is to extend the current work to the word level (as opposed to the part-of-speech level). This should be doable at least in a restricted domain. Finally, from the data compression and information theory [10], one can design a different cost function that takes the symbol frequencies into account and achieves further reduction on the number of bits.

5 Acknowledgement

This work is supported by National Science Council under grant number 94-2213-E-110-061. We thank Sheng-Fu Wang and Chiao-Mei Wang for inspirational discussions. We also thank the reviewers for the thorough comments.

Table 1: Costs in bits of exhaustive (G1) and recursive (G2) CFGs.

	rule cost	derivation cost	total cost
G1	28.1m	4.1m	32.2m
G2	607	88.4m	88.4m

Table 2: Initial data statistics for ASBC after text pre-processing. $|V|$ is the vocabulary size, q is the total number of sentences, $|R(S)|$ is the total number of distinct sentences, N_q is the total number of tokens in the corpus, and N_R is the total number of tokens in the distinct sentences.

$\|V\|$	q	$\|R(S)\|$	N_q	N_R
51	229852	203651	4838540	4729276

Table 3: Top-20 rules learned from the ASBC corpus.

X → Y+Z	例子(Y)	例子(Z)	例句
M1 → DE+Na	之	需要	研究計畫之需要
M2 → Na+Na	人際	關係	與你暢談人生、信仰、求學、工作、愛情、人際關係
M3 → Neu+Nf	第一	次	有三成三的大學生第一次有投票權
M4 → Na+D	領域	已	在此一領域已有傑出之研究成果
M5 → D+D	應	不致於	但黃國章應不致於爲此遭人持槍狙擊
M6 → D+VC	應	擬定	政府應擬定合理的都市政策
M7 → Na+M1	計畫	之需要	「台灣與東南亞土著文化與血緣關係」主題研究計畫之需要
M8 → Na+VC	院校	發出	十所大學院校發出四百份問卷
M9 → VH+M1	冷靜	的判斷力	冷靜的判斷力
M10 → DE+Nv	之	研究	在此一領域已有傑出之研究成果
M11 → VH+Na	有效	問卷	有效問卷爲三百八十五份
M12 → P+Na	在	團體	在團體中被依賴
M13 → P+Nc	對	台大	對台大、政大等十所大學院校發出四百份問卷
M14 → Nh+D	他人	一同	和他人一同行動時
M15 → Nep+Nf	這	種	正屬於這種類型
M16 → VC+Na	領導	者	你是屬於領導者型
M17 → Nc+Na	大學	院校	十所大學院校發出四百份問卷
M18 → Dfa+VH	太	遠	回家路途太遠
M19 → D+VH	不	怒	也是不怒而威
M20 → D+SHI	就	是	這就是AB型-牡羊座的一般傾向

Table 4: Selected part-of-speech tags used in the ASBC corpus.

Name	詞性	例子
A	形容詞	特有，一般，主要
D	副詞	應，已，一邊，再，就，不，只有，常常，也，都，必須
DE	語助詞	的，之，地，得
Dfa	副詞（前置）	十分，相當，很，非常，過於，最，太，較，過度，極
Na	名詞	人際，關係，父子，問題，原因，暴動，宗教
Nc	名詞（地方）	出發點，天下，家，醫院，全身，廚房，學校，印尼
Neu	定詞（數量）	一，第一，七十五，兩，幾，十餘，九萬
Nep	定詞	這，其，那，此，其中，什麼，哪
Nf	量詞	個，些，步，歲，次，種，件，位，項
Nh	名詞（代名詞）	我們，對方，自己，別人，大家，雙方
Nv	可當動詞或名詞	研究，存在，否定，演奏，製作，輔導，離婚
P	介詞	於，至，關於，和，譬如，以，將，爲
SHI	及物動詞	是
VH	不及物動詞	普遍，好，疏離，可怕，直接，相當，慈愛
VC	及物動詞	寫好，丟，翻，完成，求，存，做，出

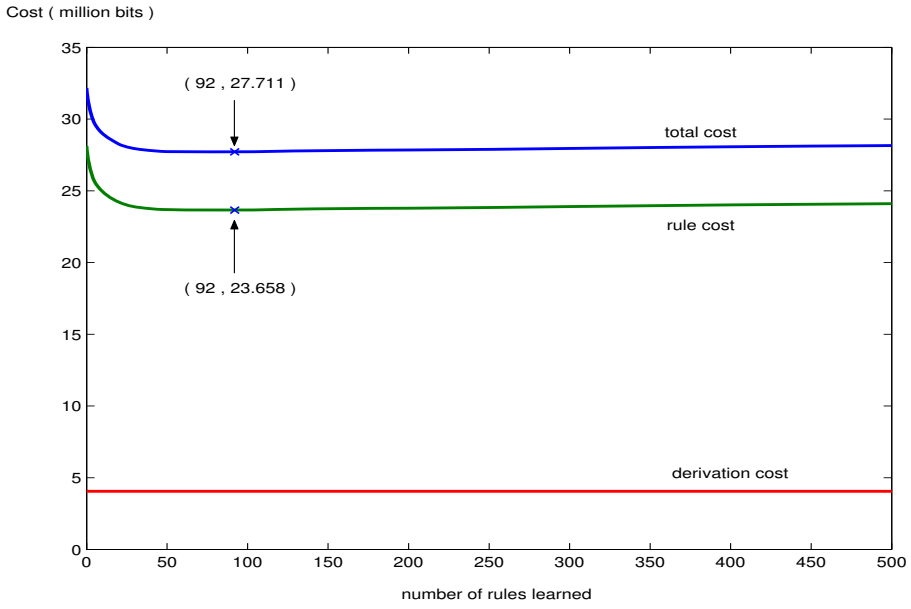


Figure 1: The cost as a function of the number of learned rules.

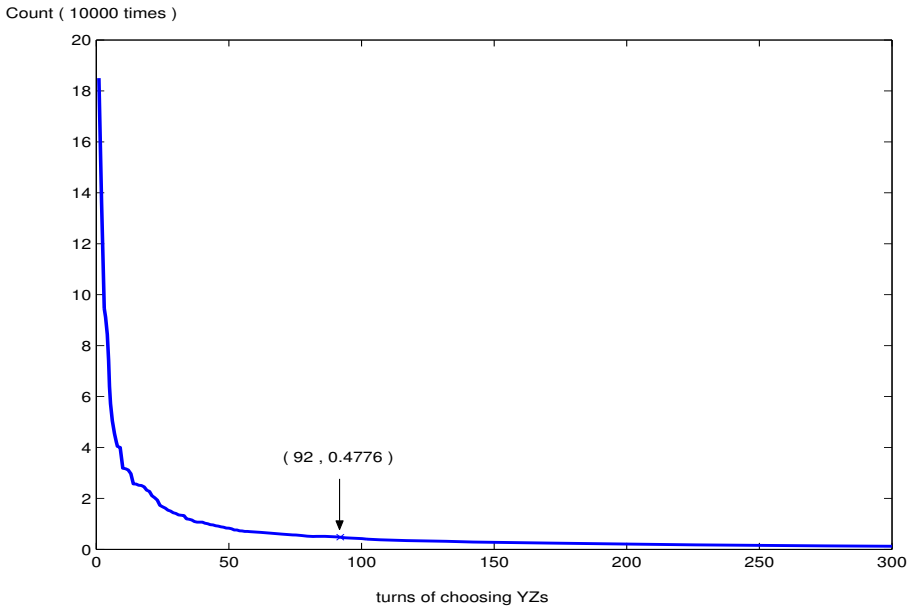


Figure 2: The maximum bigram count as a function of the number of epochs.

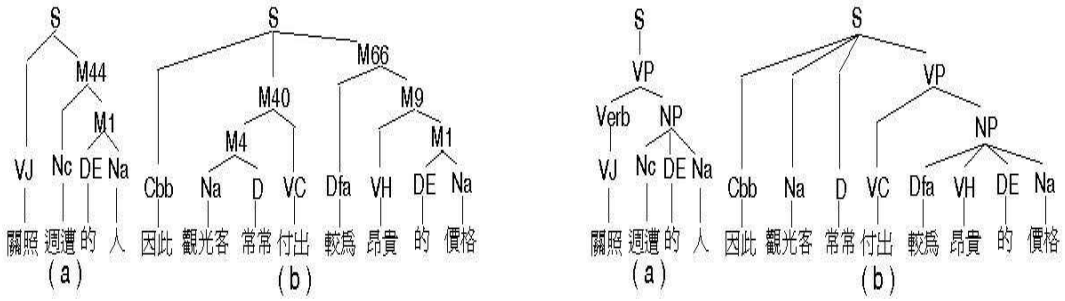


Figure 3: Examples parsed by the learned CFG (left) and parsed manually (right). Here Cbb is conjunctive and VJ is transitive verb.

References

[1] J. E. Hopcroft, R. Motwani and J. D. Ullman, “Introduction to Automata Theory, Languages and Computation”, Addison-Wesley (2001).

[2] D. Jurafsky and J. H. Martin, “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”, Prentice Hall (2000).

[3] John C. Kieffer and En-hui Yang, “Design of context-free grammars for lossless data compression,” Proceedings of the 1998 IEEE Information Theory Workshop, pp. 84-85.

[4] H. Lucke, “Reducing the computation complexity for inferring stochastic context-free grammar rules from example text”, Proceedings of ICASSP 1994, pp. 353-356.

[5] H. Ney, “Dynamic Programming Speech Recognition Using a Context-Free Grammar”, Proceedings of ICASSP’87, pp. 69-72.

[6] Mark van den Brand, Alex Sellink, and Chris Verhoef, “Generation of components for software renovation factories from context-free grammars”, In Working Conference on Reverse Engineering, IEEE Computer Society, WCRE97, pp. 144-153.

[7] C. Yuan and C. Wang, “Parsing model for answer extraction in Chinese question answering system”, Proceedings of IEEE NLP-KE ’05, pp. 238 - 243.

[8] M. Balakrishna, D. Moldovan, E.K. Cave, “Automatic creation and tuning of context free grammars for interactive voice response systems”, Proceedings of IEEE NLP-KE ’05, pp. 158 - 163.

[9] 中央研究院平衡語料庫的內容與說明, http://www.sinica.edu.tw/SinicaCorpus/98-04.pdf.

[10] T. Cover and J. Thomas, “Elements of Information Theory”, John Wiley and Sons (1991).

Improve Parsing Performance by Self-Learning

Yu-Ming Hsieh, Duen-Chi Yang, and Keh-Jiann Chen

Chinese Knowledge and Information Processing (CKIP)
Institute of Information Science,
Academia Sinica, Taipei
{morris, ydc, kchen}iis@sinica.edu.tw

Abstract

There are many methods to improve performances of statistical parsers. Among them, resolving structural ambiguities is a major task. In our approach, the parser produces a set of *n*-best trees based on a feature-extended PCFG grammar and then selects the best tree structure based on association strengths of dependency word-pairs. However, there is no sufficiently large Treebank producing reliable statistical distributions of all word-pairs. This paper aims to provide a self-learning method to resolve the problems. The word association strengths were automatically extracted and learned by parsing a giga-word corpus. Although the automatically learned word associations were not perfect, the built structure evaluation model improved the bracketed *f*-score from 83.09% to 86.59%. We believe that the above iterative learning processes can improve parsing performances automatically by learning word-dependence knowledge continuously from web.

1. Introduction

How to solve structural ambiguity is an important task in building a high-performance statistical parser, particularly for Chinese. Since Chinese is analytic language, words play different grammatical functions without inflections. A great deal of ambiguous structures will be produced by parsers if no structure evaluator is applied. There are three main steps in our approach aim to disambiguate the structures. The first step is to have parser produce *n*-best structures. Secondly, we extract word-to-word association from large corpora and build semantic information. The last one is to build a structural evaluator to find the best tree structure from *n*-best. Formerly, there were some approaches proposed to resolve structure ambiguities. For instances,

- ***to add on lexical dependencies.*** Collins (1999) solves structural ambiguity by extracting lexical dependencies from Penn WSJ Treebank and applying dependencies to the statistic model. Lexical dependency (or Word-to-word association, WA) is one type of semantic information. It is a current trend to add on semantic related information in traditional parsers. Some incorporated word-to-word association in their parsing models, such as the Dependency Parsing in Chen et al. (2004). They take advantage of statistic information of word dependency in the parsing process to produce dependency structures. However, word association methods suffer low coverage for lacking very large tree-annotated training corpora, while checking dependency relation between word pairs.

- ***to add on word semantic knowledge.*** CiLin and HowNet information are used in the statistic model in the experiment of Xiong et al. (2005). Their results prove to solve common parsing mistakes efficiently.
- ***to use re-annotation method in grammar rule.*** Johnson (1998) thinks that re-annotating each node with the category of its parent category in Treebank is able to improve parsing performance. Klein et al. (2003) proposes internal/external/tag-splitting annotation strategies to obtain better results.
- ***to build evaluator.*** Some people re-rank the structure values and find out the best parse (Collins, 2000; Charniak et al., 2005). At first hand, their parser produces a set of candidate parses for each sentence. Later, the reranker finds out the best tree through relevance features. The performance is better that without the reranker.

This paper is going to show a self-learning method to produce imperfect (due to errors produced by automatic parsing) but unlimited amount of word association data to evaluate the *n*-best trees produced by a feature-extended PCFG grammar. The parser with this WA evaluation is considerably superior to those without evaluation.

The organization of the paper is as follows: Section 2 describes how to generate *n*-best trees in a simple way. In Section 3, we account for building word-to-word association and a primitive semantic class as well. As to the design of evaluating model, our probability model, coordination of rule probability and word association probability are presented in section 4. In Section 5 we discuss and explain the experimental data and results. Ambiguities of PoS are to be considered in a practical system. Section 6 deals with further experiment on automatic tagging with PoS. Finally, we offer concluding remarks in section 7.

2. Feature extension of PCFG grammars for producing the *n*-best trees

It is clear that Treebanks (Chen et al., 2003) provide not only instances of phrasal structures and word dependencies but also their statistical distributions. Recently, probabilistic preferences for grammar rules and feature dependencies were incorporated to resolve structure-ambiguities and had great improvements on parsing performances. However, the automatic extracted grammars and feature-dependence pairs suffer the problem of low coverage. We proposed different approaches to solve these two different types of low coverage problems. For the low coverage of extracted grammar, a linguistically-motivated grammar generalization method is proposed in Hsieh et al. (2005). And the low coverage of word association pairs is resolved by a self-learning method of automatic parsing and extracting word dependency pairs from very large corpora.

The linguistically-motivated generalized grammars are derived from probabilistic context-free grammars (PCFG) by right-association binarization and feature embedding (Hsieh et al., 2005). The binarized grammars have better coverage than the original grammars directly extracted from treebank. Features are embedded in the lexical and phrasal categories to improve the precision of generalized grammar. The important features adopted in our grammar are described in the following:

Head (Head feature):	The PoS of phrasal head will propagate all intermediate nodes within the constituent.
Example:	S(NP(Head:Nh: 他)|S'$_{-Head:VF}$(Head:VF: 叫 |S'$_{-Head:VF}$(NP(Head:Nb: 李 四)| VP(Head:VC:撿| NP (Head:Na:球)))))
Linguistic motivations:	To constrain the sub-categorization frame.
Left (Leftmost feature):	The PoS of the leftmost constitute will propagate one–level to its intermediate mother-node only.
Example:	S(NP(Head:Nh: 他)|S'$_{-Head:VF}$(Head:VF: 叫 |S'$_{-NP}$(NP(Head:Nb: 李 四)| VP(Head:VC:撿| NP(Head:Na:球)))))
Linguistic motivation:	To constrain linear order of constituents.
Head 0/1 **(Existence of phrasal head):**	If phrasal head exists in intermediate node, the nodes will be marked with feature 1; otherwise 0.
Example:	S(NP(Head:Nh:他)|S'$_{-1}$(Head:VF:叫|S'$_{-0}$(NP(Head:Nb:李四)|VP(Head:VC:撿| NP(Head:Na:球)))))
Linguistic motivation:	To enforce unique phrasal head in each phrase.

There are two functions in applying the embedded features: one is to increase the precision of the grammar and the other is to produce more candidate parse structures. With features embedded in phrasal categories, PCFG parsers are forced to produce varieties of different possible structures[1]. In order to achieve a better *n*-best oracle performance (i.e. the ceiling performance achieved by picking the best structure from n bests), we designed some different feature-embedded grammars and try to find a grammar with the better *n*-best oracle performance. For instance, "S(NP(Head:Nh:他)|Head:VF:叫| NP(Head:Nb:李四)| VP(Head:VC:撿| NP(Head:Na:球)))". The explanations of feature sets are as follow.

Rule type-1:
Intermediate node: add on "Left and Head 1/0" features.
Non-intermediate node: if there is only one member in the NP, add on "Head" feature.
Example: S(NP$_{-Head:Nh}$(Head:Nh:他)\|S'$_{-Head:VF-1}$(Head:VF:叫\|S'$_{-NP-0}$(NP$_{-Head:Nb}$(Head:Nb:李四)\|VP(Head:VC:撿\| NP$_{-Head:Na}$(Head:Na:球)))))

Rule type-2:
Intermediate node: add on "Left and Head 1/0" features.
Non-intermediate node: add on "Head and Left" features, if there is only one member in the NP, add on "Head" feature.
Example: S$_{-NP-Head:VF}$(NP$_{-Head:Nh}$(Head:Nh:他)\|S'$_{-Head:VF-1}$(Head:VF:叫\|S'$_{-NP-0}$(NP$_{-Head:Nb}$(Head:Nb:李四)\|VP$_{-Head:VC}$(Head:VC:撿\| NP$_{-Head:Na}$(Head:Na:球)))))

Rule type-3:
Intermediate: add on "Left, and Head 1/0" features.
Top-Level node: add on "Head and Left" features. (see example of S$_{-NP-Head:VF}$)
Non-intermediate node: if there is only one member in the NP, add on "Head" feature.
Example: S$_{-NP-Head:VF}$(NP$_{-Head:Nh}$(Head:Nh:他)\|S'$_{-Head:VF-1}$(Head:VF:叫\|S'$_{-NP-0}$(NP$_{-Head:Nb}$(Head:Nb:李四)\|VP(Head:VC:撿\| NP$_{-Head:Na}$(Head:Na:球)))))

[1] The parser adopts an Earley's Algorithm. It is a top-down left-to-right algorithm. So, in parts that have the same non-terminals, we keep only the best structure after pruning, to reduce the load of calculating and thus fasten the parsing speed. Therefore, if we add different features in the Top-Level rules, we'll get more results.

Rules and their statistical probabilities are extracted from the transformed structures. The grammars are derived and trained from Sinica Treebank. Sinica Treebank contains 38,944 tree-structures and 230,979 words. Table 1 shows the number of rule types in each grammar and Table 2 shows their 50-best oracle bracketed *f*-scores on three sets of testing data. The three sets of testing data used in our experiments represent "moderate", "difficult" and "easy" scale of Chinese language respectively. We adopt PARSEVAL measures to evaluate the bracketed *f*-score (BF)[2] as Table 2. A bracket represents the phrasal scope. The reason we don't use labeled *f*-score is that we aim to evaluate the phrasal scope, rather than the effect brought by phrasal category.

Table 1. Numbers of rules for each grammar.

	Rule Type		
	Rule-1	Rule-2	Rule-3
Rule number	9,899	26,797	13,652

Table 2. The 50-best oracle performances from the different grammars.

Testing Data	Sources	Hardness	Rule Type		
			Rule type-1	Rule type-2	Rule type-3
Sinica	Balanced corpus	Moderate	92.97	94.84	96.25
Sinorama	Magazine	Difficult	90.01	91.65	93.91
Textbook	Elementary school	Easy	93.65	95.64	96.81

From the above table, we can observe that the “Rule type-3” outperforms the “Rule type-1” and “Rule type-2”. We adopt the approach used in Charniak et al. (2005) to analyze the *n*-best parse. Table 3 shows the bracketed *f*-score values of different candidate trees. From the result, we observe that the improvement after *n*=5 is slight. Thus the number of ambiguous candidates can be dynamically adjusted according to the complexity of input sentences. For normal sentences, we may consider to take *n*=5 in order to minimize the complexity. For long sentences or sentences with auto PoS tagging should take as large as *n*=50 to raise the ceiling of the best *f*-score.

Table 3. Oracle bracketed *f*-scores as a function of number n of *n*-best parses.

Testing Data	*n*					
	1	2	5	10	25	50
Sinica	91.88	94.39	95.91	96.17	96.25	96.25
Sinorama	86.69	90.44	92.87	93.47	93.86	93.91
Textbook	92.24	95.01	96.21	96.61	96.78	96.81

[2] The harmonic mean of bracketed precision (BP) and bracketed recall (BR), i.e. $BF = \frac{2 * BP * BR}{BP + BR}$

For each candidate tree, its syntactic plausibility is obtained by rule probabilities produced by PCFG parser. Yet, we need semantic related information to help with finding the best tree structure among candidate trees. In the next section, we will see methods to get semantic related information.

3. Auto-Extracting world knowledge

In our experiments, we use a Gigaword Chinese corpus instead of texts from web to extract word dependence pairs. The Gigaword corpus contains about 1.12 billion Chinese characters, include 735 million characters from Taiwan's Central News Agency (traditional characters), and 380 million characters from Xinhua News Agency (simplified characters)[3]. Word associations are extracted from the texts of Central News Agency (CNA). First we use Chinese Autotag System (Tsai et. al., 2003), developed by Academia Sinica, to process the segmentation and PoS tagging of the texts. This system reaches a performance of 95% segmentation ability and 93% tagging ability. Then we parse each sentence[4] in the corpus and assign semantic roles to each constituent. Based on the head word information, we extract dependence word-pairs between head words and their arguments or modifiers. There are three types of the word pairs: (a) head word on the left hand side: (H_W_C, X_W_C); (b) head word on the right hand side: (X_W_C, H_W_C); (c) coordinating structure: (H_W_C, H_W_C). In the word pairs, “H” denotes Head, “W” means word, and “C” refers to PoS tag, “X” refers to any semantic role other than Head role. Figure 1 is an example of extracted word associations. The following illustrates how the automatic knowledge extraction works. We input a Chinese sentence to the parser:

他 叫 李四 撿 球
Ta jiao Li-si jian qiu
He ask L-isi pick ball
"He asked Li-si to pick up the ball."

Here is the sentence after segmentation and PoS tagging:

他(Nh) 叫(VF) 李四(Nb) 撿(VC) 球(Na)

The parser analyzes the sentence structure and assigns roles to each phrase. And then word-pair knowledge of heads and their modifiers are extracted as shown in Figure 1. The processes above are repeated in new data, no matter in Gigaword or texts from the internet. Finally we obtain a great deal of knowledge on words and their relations, and the amount of knowledge is on the increase. Meanwhile the evaluator takes this knowledge is for reference as well.

[3]http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T09
[4]An existing parser is used to produce 1-best tree of a sentence.

Parsing and role assignment:

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)))

Word association extraction:

	Role1	PoS1	Word1	Role2	PoS2	Word2
a	X	Nh	他	H	VF	叫
b	H	VF	叫	X	Nb	李四
c	H	VF	叫	X	VC	撿
d	H	VC	撿	X	Na	球

Figure 1. A sample for word association extraction.

We have 37,489,408 sentences that are successfully parsed and with word association information. And the number of extracted word associations is 221,482,591. The extracted word to word associations that undergo structure analysis and head word assignment are not perfectly correct, but they are more informative than simply taking words on the left and right hand window.

3.1. Coverage rates of the word associations

Data sparseness is always a problem of statistical evaluation methods. We test our extracted word association data in five different levels of granularities. Level-1 to Level-5 represents HWC_WC, HW_W, HC_WC, HW_C, and HC_C respectively. We like to see the bi-gram coverage rates for each level of representation. We divide word association data into ten. Figure 2 shows coverage relationships between five levels and sizes of word association data for three testing data. The extracted word association data are divided into ten layers of different sizes for each level.

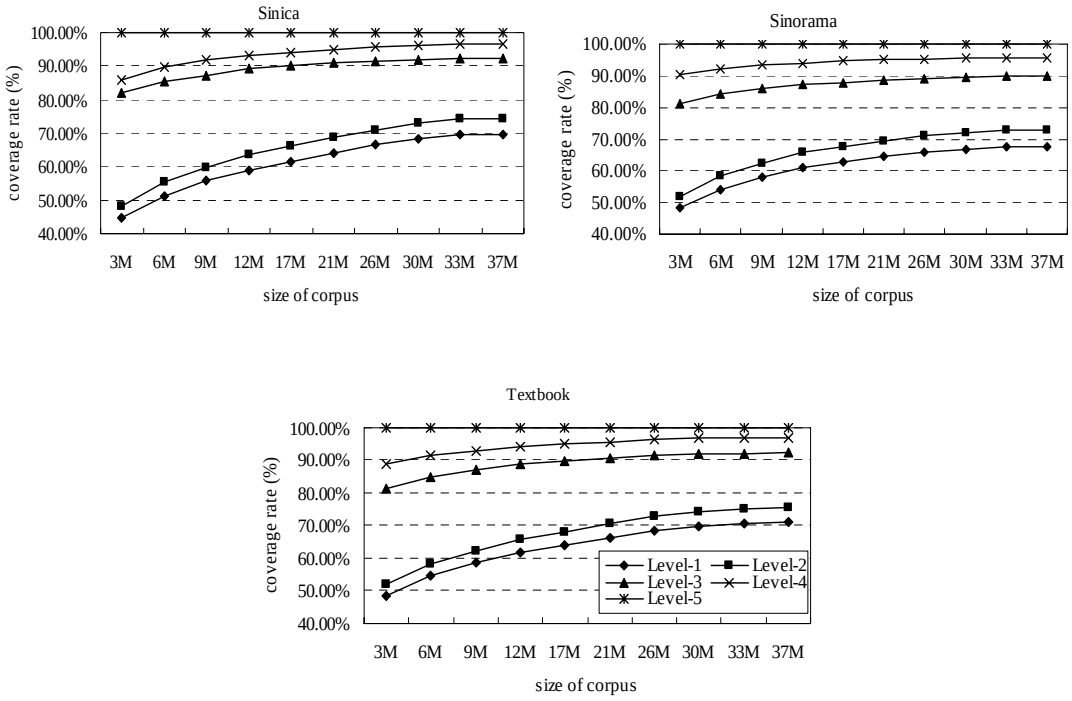


Figure 2. Coverage rates vs. size of Corpus: (a) Sinica; (b) Sinorama; (c) Textbook.

Figure 2 shows that larger data increases the coverage rates, but the coverage of the fine-grained level word associations, e.g. Level-1 (HWC_WC), is about 70%, which are far from saturation. Nonetheless the coverage rate can be improved by reading more texts from web. The coarse-grained level associations, e.g. Level-5 (HC_C), cover the most category bi-gram. But it may not be very useful, since syntactic associations which are partially embedded in the PCFG are redundant. To achieve a better evaluation model, we derived new associations between semantic classes. Criteria for semantic classification are discussed in the following section.

3.2. Incorporating semantic knowledge

In this section, we propose a simple approach to build a semantic-class-based relation for words, and that will be Level-6 (HS_S). Semantic class information is put into Level-6 in order to get high coverage and to avoid redundant syntactic associations in other levels. Besides, we hope to smooth the problem of data sparseness.

The idea is to classify words into their head morpheme. It begins with the transformation of every input "WORD, POS" in the data. We adopt affix database of high frequency verbs and nouns (Chiu et al., 2004) to setup noun and verb classes. There are 34,857 corresponding affixes. As to determinative measures (DM), we refer to the dictionary of measure words, and divide the DMs in the data into thirteen categories, according to the meanings of measure words. The thirteen categories include general, event, length, science, approximate measures, weight, square measures, container, capacity, time, currency value, classification measures, and measures of verbs. Finally we consult parts of speech analyses (CKIP, 1993) and the transformation rules of Figure 3 to build our semantic class. Take "張三, Nb" for example, its semantic class is "PersonalName" in our classification.

```
Notation:     WORD: user input Word
              POS: user input PoS of the word
              CLASS: transformation class of the word
              Affix(WORD): input WORD to find mapping affix from table
              Prefix(WORD): prefix of the WORD
              Suffix(WORD): suffix of the WORD
              DM(WORD): input Word to find DM category
Input:        WORD, POS
Output:       CLASS
Initial Step:
    CLASS=WORD;
    if WORD in affix table then CLASS=affix(WORD);
    if POS is verb or adverb then CLASS=POS+prefix(WORD);
    if POS is noun then CLASS=POS+suffix(WORD);
Mapping Step:
    if POS is non-predicative adjective then CLASS='A'+prefix(WORD);     /* e.g. A */
    if POS is preposition then CLASS='P'+suffix(WORD);   /* e.g. P */
    if POS is SHI then CLASS='SHI';     /* e.g. 是 */
    if POS is V_2 then CLASS='V_2';     /* e.g. 有 */
    if POS is DM or Measure and exist in DM table then CLASS=DM(WORD); /* e.g. DM/Nf */
    if POS is conjunction then CLASS=POS+prefix(WORD);     /* e.g. Caa/Cab/Cba/Cbb */
    if POS is determinative then CLASS=POS;     /* e.g. Nep/Neqa/Neqb/Nes/Neu */
    if POS is pronoun then CLASS=WORD;     /* e.g. Nh */
    if POS is time noun then CLASS='Time';     /* e.g. Nd */
    if POS is Postposition/Place Noun/Localizer then CLASS='Location';     /* e.g. Ng/Nc/Ncd */
    if POS is Proper Noun and is family names then CLASS='PersonalName';     /* e.g. Nb */
    if POS is aspectual adverb, CLASS=POS   /* e.g. Di */
    if POS is pre/post-verbal adverb of degree then CLASS='Df'+suffix(Word) /*e.g. Dfa/Dfb */
    if POS is VD/VCL/VL then CLASS=POS+suffix(WORD)
```

Figure 3. Transformation algorithm.

We estimate the word association coverage rate as the above mentioned. From the results shown in Figure 4, the coverage rate of Level-6 is higher than Level-2, and the problem of data sparseness is indeed moderated.

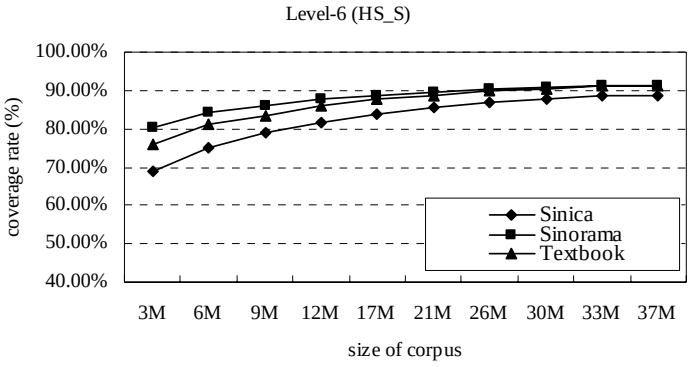


Figure 4. WA coverage rate of Level-6.

Now we have semantic information. How it works with rule probability to find the best structure among the numerous ambiguous candidates will be discussed in Section 4.

4. Building evaluation model

A sentence structure is evaluated by its syntactic and semantic plausibility. The syntactic plausibility is modeled by products of phrase rule probabilities of its syntactic tree. The semantic plausibility is modeled by the word association strengths between head words and their arguments or modifiers. For an input sentence s, the feature-embedded PCFG parser produces *n*-best trees of $\{y_1(s),...,y_n(s)\}$. The evaluating model finds out the best structure according to the rule probability (syntactic) and corresponding word association probability (semantic). Rule probabilities are marked when *n*-best trees are produced. We will estimate word association probabilities in the following formula. In the formula, "Head" means the Head member word association, as HWC, HC, HW. "Modify" means modify or argument member, as in WC, W, C. "freq(Head)" means Head word frequency in the corpus and "freq(Head, Modify)" refers to the co-occurrence frequency of "Head" and "Modify".

$$P(Modify \mid Head) = \frac{freq(Head, Modify)}{freq(Head)} \tag{1}$$

Data sparseness is a common problem in dealing with corpus. A minimal value σ is used to smooth data sparseness, such as $\sigma = \frac{1}{total\ number\ of\ WA\ token}$.

$Value(y_n(s))$ in the formula below means the final evaluation value to each candidate tree.

$$Value(y_n(s)) = \lambda * RuleValue(y_n(s)) + (1-\lambda)WAValue(y_n(s))\,, \quad (2)$$

Where $RuleValue(y_n(s))$ is the rule probability of the sentence and $WAValue(y_n(s))$ is the total word association value in different level *n*. RuleValue and WAValue are normalized, i.e. *(i-min)/(max-min)*. The following shows weighting in different levels and explanation of formula:

$$WAValue(y_n(s)) = \sum_{level=1}^{6} \theta_{level} * WA_{level}(y_n(s)) \quad (3)$$

$$WA_{level}(y_n(s)) = \prod_{all\_word\_association\_for\_y_n(s)} P(Modify \mid Head) \quad (4)$$

After semantic probability collocating with rule probability, we hope to find the best tree $y^*(s)$.

$$y^*(s) = \arg\max Value(y_n(s)) \quad (5)$$

where $y^*(s)$ has the best bracketed *f*-score. We calculate relating λ and θ values from development sets. The development sets are adopted from trees in training data. In evaluation, we substitute λ and θ for every interval of 0.1 from 0 to 1. Then we find out the best results in certain probability. The experiment results will be shown in the following section. Moreover, we justify whether the word associations are reasonable.

5. Experimental results

We evaluated the performance of our evaluating model using the standard PARSEVAL metrics. Hsieh et al. (2005) state that the bracketed *f*-score of short sentence parsing (the length of a sentence is from 1 to 5 words) is over 90% in their experiment. As a result, the following experiments are on sentences more than 6 words. The oracle 50-best bracketed *f*-scores of “Rule type-3" are listed in Table 4.

Table 4. The bracketed *f*-scores of 1-best and oracle performance of 50-best. (sentence length ≥ 6)

Top *n*-best	Testing data		
	Sinica	Sinorama	Textbook
1-best	83.09	77.545	83.195
50-best	90.11	87.445	89.945

To simplify our evaluation model, we try to find the most effective levels of associations first. In turn, the evaluation model uses only one level of association and rule probabilities to select the best structure from n candidates. That is,

$$WAValue(y_n(s)) = WA_{level}(y_n(s)) = \prod_{all\_word\_association\_for\_y_n(s)} P(Modify \mid Head) \quad (6)$$

Figure 5 displays the results of testing data. The best results of Level-1 slightly surpass that of Level-2; results of Level-6 overtake that of Level-3; Level-6 has better performance than Level-5. Therefore, only three levels (Level-1, Level-4 and Level-6) are chosen to be calculated, for dimension reduction.

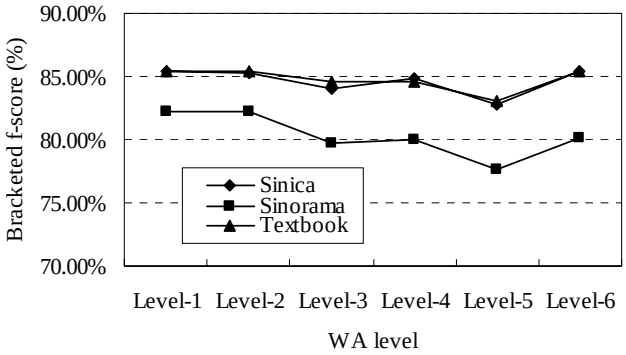


Figure 5. Matching rule with WA value in each level (sentence length ≥ 6).

Finally we use the combination of L1, L4, and L6 associations and rule probabilities to evaluate plausibility of structures. Results of experiments on the three testing data are shown in Table 5.

Table 5. The bracketed *f*-scores of 50-best parses (sentence length ≥ 6)

Models	Testing data		
	Sinica	Sinorama	Textbook
R, L1, L4, L6	86.59	82.81	85.97

In Table 5, we see that semantic information is effective in finding correct structure. If we justify the rationality of WA, about 3.5%~5.2% of the performance is raised. In our experiments, λ =0.7, θ_1=0.7, θ_4=0.3, and θ_6=0.5. In Charniak et al. (2005), the f-score was improved from 89.7% (without reranking) to 91.02% (with reranking) for English; the oracle f-score was 96.8% for n-best in their paper. From the result, we see an improvement in the testing data. With the more data parsed, better word-association values are obtained. This enhances the parsing performance and reaches our goal of self-learning.

6. Further Experiments on Sentences with Automatic PoS Tagging

Perfect testing data was used in the above experiments without considering PoS tagging errors. However, in reality, PoS tagging errors will degenerate parsing performances. The real parsing performances of accepting input from PoS tagging system are shown in the Table 6(1). In this table, "Autotag" mean to markup the best PoS on the segmented data. The naïve approach to overcome the PoS tagging errors is to delay some of the ambiguous PoS resolution for words with lower confidence tagging scores and leave the ambiguous PoS to be resolves at parsing stage. The tagging confidence of each word is measured by the following value.

$$\text{Confidence value} = \frac{P(c_{1,}w)}{P(c_{1,}w) + P(c_{2,}w)}, \tag{7}$$

where P(c1,w) and P(c2,w) are probabilities assigned by the tagging model for the best candidate “c1,w” and the second best candidate “c2,w”.

In Table 6(2), "Autotag with confidence value=1.0" means that if confidence value <= 1.0, we list all possible PoSs for parser to decide. The experimental results, Table 6(2), show that delaying ambiguous PoS resolution does not improve parsing performances, since PoS ambiguities increase structure ambiguities and the PCFG parser is not robust enough to select better syntactic structures.

Table 6. Oracle bracketed *f*-scores of different autotag for parsing:
(1)Autotag; (2)Autotag with confidence value = 1.0.

Top *n*-best		Testing data		
		Sinica	Sinorama	Textbook
(1)	1-best	75.31	72.05	79.27
	50-best	84.09	83.36	87.54
(2)	1-best	73.41	68.34	77.83
	50-best	86.45	83.99	88.83

We then apply our evaluation model to select the best structure from 50-best parses. The results are shown in Table 7. The experiment above takes “Rule type-3” for *n*-best parses. The bracketed *f*-score is raised from the original 73.41% to 79.34%, about 4% of improvement in the Sinica testing data. Sinorama data is improved from 68.34% to 74.78%. Textbook data is from 77.83% to 82.59%. All these results are raised up to 2%~4%. We can see that our evaluating model finds better results than Autotag. In solving the ambiguous POS, our evaluating model produces better tree structures than Autotag.

Table 7. The bracketed *f*-scores in Autotag
with confidence value=1.0 and 50-best parses (sentence length ≥ 6).

Models	Testing data		
	Sinica	Sinorama	Textbook
R, L1, L4, L6	79.34	74.78	82.59

7. Conclusion

Parsers of any language aim to correctly analyze the syntactic structure of a sentence, often with the help of semantics. This paper shows a self-learning method to produce imperfect (due to errors produced by automatic parsing) but unlimited amount of word association data to evaluate the *n*-best trees produced by a feature-extended PCFG grammar. We prove that although the statistical association strengths produced by automatic parsing are not perfect, still the extracted data is reliable enough in measuring plausibility of ambiguous structures. The parser with this WA evaluation is considerably superior to those without evaluation. We believe that the above iterative learning processes can improve parsing performances automatically by learning word-dependence knowledge continuously from web. We also propose an easy method to produce *n*-best of a sentence. First of all we slightly modify the parser to produce *n*-best. Then different feature sets in grammar rules are used to bring forth different results. There is one feature set that covers more structures than the original 1-best. In our experiments, we use 50-best to estimate the efficiency of our evaluating model.

On the other hand, we offer a general syntactic and semantic evaluation model. We input *n*-best parses to our evaluating model. The evaluating model selects the best parse from this set of parses using a rule and semantic probability. The system we described, using the standard PARSEVAL framework, has a bracketed *f*-score of the selected trees, which is 86.59% higher to the original 1-best. Furthermore, ambiguous PoS of a word is also parsed and evaluated on *n*-best. We can see that our evaluating model finds better results than Autotag.

In the future research, we plan to improve the quality of word-association. Three aspects need to done: improving the accuracy of PoS tagger; enhancing the parser's ability to solve common mistakes, such as parsing conjunctive structures; extracting more word associations by reading and parsing text from web. As to the evaluating model, a properly corresponding semantic classifications from coarse to fine-grained category are needed in Level-6.

8. Acknowledgements

This research was supported in part by National Science Council under Grant NSC 95-2422-H-001-008- and National Digital Archives Program Grant 95-0210-29-戊-13-09-00-2.

9. References

Eugene Charniak and Mak Johnson. 2005. Coarse-to-fine *n*-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 173-180, Ann Arbor, MI.

Keh-Jiann Chen, Chu-Ren Huang, Feng-Yi Chen, Chi-Ching Luo, Ming-Chung Chang, Chao-Jan Chen, and Zhao-Ming Gao. 2003. Sinica Treebank: design criteria, representational issues and implementation. In Anne Abeille, (ed.): *Building and Using Parsed Corpora. Text, Speech and Language Technology*. 20:231-248, pp231-248.

Yuchang Chen, Masayuki Asahara, and Yuji Matsumoto. 2004. Deterministic Dependency Structure Analyzer for Chinese. In *Proceedings of the First International Join Conference on Natural Language Processing*, pages 135-140, Sanya City, Hainan Island, China.

Chih-Ming Chiu, Ji-Qing Luo, and Keh-Jiann Chen. 2004. Compositional semantics of mandarin affix verbs. In *Proceedings of ROCLING XVI: Conference on Computational Linguistics and Speech Processing*, pages 131-139, Taipei.

CKIP (Chinese Knowledge Information processing). 1993. The categorical analysis of Chinese. Technical Report no 93-05. Taipei: Academia Sinica.

Michael Collins. 1999. Head-driven statistical models for natural language parsing. PhD thesis, University of Pennsylvania.

Michael Collins. 2000. Discriminative reranking for natural language parsing. *In Machine Learning: Proceedings of the Seventeenth International Conference (ICML 2000)*, pages 175-182, Morgan Kaufmann, San Francisco, CA.

Yu-Ming Hsieh, Duen-Chi Yang, and Keh-Jiann Chen. 2005. Linguistically-motivated grammar extraction, generalization and adaptation. In *Proceedings of the Second International Join Conference on Natural Language Processing*, pages 177-187, Jeju Island, Republic of Korea.

Mark Johnson. 1998. PCFG models of linguistic tree representations. *Computational Linguistics*, 24(4): 613-632.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423-430, Sapporo, Japan.

Yu-Fang Tsi and Keh-Jiann Chen. 2003. Context-rule model for PoS tagging. In *Proceedings of 17th Pacific Asia Conference on Language, Information and Computation (PACLIC 17)*, pages 146-151, COLIPS, Sentosa, Singapore.

Deyi Xiong , Shuanglong Li, Qun Liu, Shouxun Lin, and Yueliang Qian. 2005. Parsing the Penn Chinese Treebank with semantic knowledge. In *Proceedings of the Second International Join Conference on Natural Language Processing*, pages 70-81, Jeju Island, Republic of Korea.

國語雙字語詞聲調評分系統
A Scoring System for Mandarin Tones Uttered in Disyllabic Words

古鴻炎# 孫世諺# 張小芬*
Hung-Yan Gu# Shih-Yan Sun# Hsiao-Fen Chang*

#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 }
http://www.csie.ntust.edu.tw/

摘要

本文對國語雙字詞的聲調發音，研究了一種可行的、語者無關的聲調評分方法，並完成可實際操作的評分系統。評分的處理分為四個步驟，首先作基週偵測；接著作基週軌跡預處理，以更正錯誤的軌跡點，及對基週軌跡作時間與音高的正規化；再者是以距離量測作樣式比對；最後是聲調評分決策，輸出之分數以五等第計分，由高至低為 5 至 1 分。經由測試實驗發現，程式評分和人工評分之間的分數誤差的平均值，已可小於 0.5 分。

關鍵詞： 基週偵測，基週軌跡，音高正規化，聲調評分

1. 前言

國語是一個聲調語言，會因爲聲調的不同而影響語意。不過許多的外國語言，並未以聲調來分辨語意，因此一般外國人在學習國語時，常會因自己母語的說話習慣，而忽略了國語的聲調。因此，外國人和聽障人士學習國語時，正確的聲調發音的學習，是非常重要的[1, 2, 3]。

學習國語的聲調，一般是在課堂上由老師發音帶著學生唸，然後老師再根據學生的發音情況去指導修正。現在隨著資訊科技的進步，語言學習可以透過更多樣的形式來進行，我們期望有一個互動的機制，使國語聲調的學習，可以更爲方便，因此我們就著手研究國語聲調的評分方法，及製作可實際操作的電腦聲調評分系統。

關於國語或閩南語聲調之電腦評量的研究，過去已有一些研究報告被提出[4, 5, 6, 7]。有些研究是以單字的聲調發音來作評量 [4, 5]，但單字發音之方式，和實際溝通時之連續語音的發音方式不同，聲調的特性已經改變很多，例如在唸單字的第三聲時，會發出全上聲(先降再升)，而在語詞或語句裡，第三聲大多只發

前半上，再者兩個都是第四聲的語詞(如”重要”)，第一個四聲的音調下降幅度會明顯地減少。另外，有些研究是以語句整體的聲調發音來作綜合評量[6, 7]，而不管各個組成字的細部聲調發音，這樣的評量方式，我們又覺得太過於籠統。在本研究裡，除了要兼顧語句裡的聲調發音特性，也希望針對各個字作較細部的聲調評量，因此我們選擇以雙字語詞作爲聲調發音練習的單位，並且對兩個組成字分別去作聲調發音的評分。

本研究的聲調評分方法，主要的處理流程如圖 1 所示，第一個方塊是” 基週偵測”，把輸入的語音信號(取樣率 22,050Hz)切成一序列的音框(frame)，音框長度爲 25ms，且每次前進 8ms，然後對各個音框去計算自相關係數和 AMDF 係數[8]，再依這兩種係數的比值去決定出基週的週期長度，較詳細的決定方法可參考我們先前的論文[9]或 Kim, H. Y.等人的論文[10]。圖 1 裡的第二個方塊是”基週軌跡預處理”，預處理的主要工作包含了：(a)倍頻與半頻之頻率更正、(b)連音之分段、(c)時間與音高之正規化等，較詳細的處理方法將在第二節裡說明。圖 1 裡的第三個方塊是”樣式比對”，把事先準備的雙字詞參考發音的基週軌跡拿出，和新輸入的發音的基週軌跡作比對，由於已經作過時間的正規化，所以這裡的比對只是依據特定的距離量測方法來計算距離，較詳細的說明在第三節裡。圖 1 裡的第四個方塊是”評分決策”，這裡所使用的決策方法，並不是語音辨識裡常用的 NN 或 KNN[11]方法，因爲分數是可以作加減計算的，所以我們的決策方法都牽涉到”平均”的觀念，詳細的說明在第四節裡。

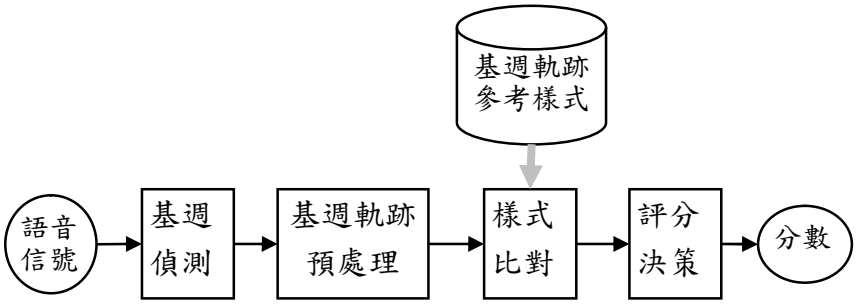


圖 1 聲調評分之主要處理流程

除了研究、探討評分的方法，我們也已經把圖 1 所示的處理流程，製作成一個可實際操作的國語雙字詞聲調的評分系統，系統的畫面如圖 2 所示。使用者可以從圖 2 右上角方塊內點選欲作練習的課程及聲調組合，然後預錄好的老師示範發音之基週軌跡就會被立即偵測出來，並顯示於圖 2 左上方塊內；此時使用者可以按錄音按鈕來練習發音，而其基週軌跡也會被立即偵測並顯示於圖 2 左邊中間的方塊內，如此就可以比較自己的和老師的基週軌跡，在趨勢(向上或向下)及高低的位置是否有差別，也就是經由視覺回饋來幫助學習。進一步使用者也可按圖

2 右下之評分按鈕，讓電腦來作評分，圖 2 裡顯示的是，老師示範的”葡萄”之發音，經電腦評分後分別都得到 5 分，電腦的評分範圍是 1 至 5 分。

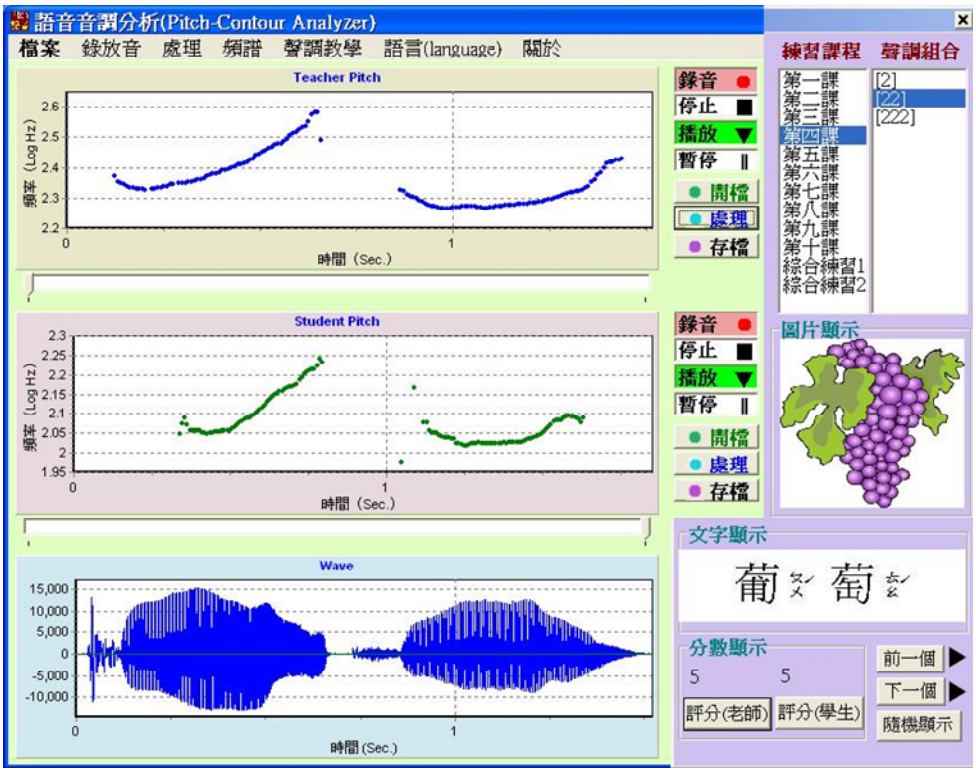


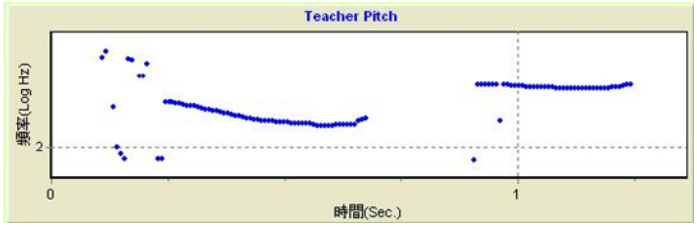
圖 2 國語雙字詞聲調評分系統之畫面

2. 基週軌跡預處理

在圖 1 的基週偵測方塊，並不保證對每個音框都可以偵測到正確的基頻值，例如偵測到真正基頻的倍頻、半頻，或未偵測到，這裡基頻值是以 10 爲底之 log Hz 尺度來表示。再者，有些雙字語詞的發音會因爲 coarticulation 而發生連音的情況，而得到一長條連續的基週軌跡，這就需要作軌跡切割分段的處理。此外，我們希望讓男生、女生之不同使用者，都能來操作我們的評分系統，因此來自不同人、不同音高的基週軌跡，必需先作音高的正規化處理。詳細來看，我們的基週軌跡預處理會依序執行如下的處理項目: (a)音高範圍決定，(b)倍頻、半頻更正，(c)中值法平滑處理，(d)軌跡段落尋找與分段，(e)時間正規化，(f)音高正規化。以下就對各個處理項目作介紹，不過中值法平滑處理[8]，由於是常見的處理方法，我們就省略了，在此使用的是 5 點的中值平滑處理(medium smoothing)。

2.1 音高範圍決定

由於系統並不知道使用者的特性(音調高低的範圍)，所以一開始會把音高值的上、下限設定得較保守，如此畫出來的軌跡圖(以”草蝦”發音爲例)，就會如圖 3(a)所示，顯得平平扁扁的，因此我們設計了音高範圍的動態估計方法，當依據估計的結果來設定音高範圍的上、下限值，就可得到如圖 3(b)所示的對比較強列、較有視覺效果的軌跡圖。除了視覺效果的增進，動態方式決定音高範圍，還可用來濾除一些基週偵測錯誤的雜點，如圖 3(a)左邊散佈的點。



3(a) 固定的音高範圍

3(b) 動態決定音高範圍

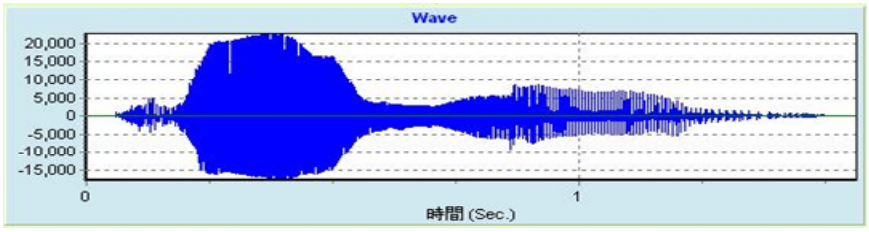
圖 3　音高範圍設定對基週軌跡圖的影響

本文的音高範圍動態估計方法是，基於 histogram 的觀念，把縱軸(log10 Hz 頻率)平均切成寬度爲 0.05 之一序列的區間，然後對各個音框依其基頻值決定所應掉落的區間，接著統計各個區間所收到的音框個數。假設第 n 個區間收到最多的音框數，則接著從 n 往上(n+1, n+2, …)檢查，直到發現有連續三個區間都沒有收到音框，就將此時的三個區間之中間區間所對應的頻率值，設定爲音高值的上限；依此程序從 n 往下(n-1, n-2, …)檢查，也可得到音高值的下限。

2.2 倍頻、半頻更正

圖 1 裡的基週偵測方塊，可能在某一時刻附近的連續數個音框，都偵測到真正頻率的半頻值，如圖 4(a)是”充滿”發音的信號波形，而對應的基週軌跡曲線如 4(b)所示，軌跡曲線有兩個片段發生半頻值之錯誤，前一片段裡有連續 6 點，而後一片段裡有連續 3 點，這樣的錯誤已經無法由下一個處理項目”中值法平滑化”來作更正，並且”充滿”的基週軌跡發生了連音現象，若不謹慎更正半頻值錯誤，則很難正確完成之後第二個處理項目”基週軌跡的分段”，因此我們設計了倍頻、半頻的更正處理方法，執行該方法後，就會得到如圖 4(c)所示的軌跡曲線，即前

述的兩處半頻值錯誤已被更正了。



4(a)　“充滿”的信號波形

4(b)　有半頻錯誤的軌跡曲線

4(c)　更正半頻錯誤後的軌跡曲線

圖 4　有、無半頻錯誤之基週軌跡比較

本文的倍頻、半頻更正方法是，以多數的正確點的力量迫使少數的錯誤點移動到正確位置，實作上則是使用動態規劃的方法。在此令 p(t,1)存第 t 個音框的基週偵測結果(log10 Hz 尺度)，p(t,0)存它的半頻值且 p(t, 2)存它的倍頻值，也就是 p(t,0) = p(t,1) – $\log_{10}2$，p(t,2) = p(t,1) + $\log_{10}2$，而更正方法就是，先計算遞迴公式(1)，

$$D(t,i) = \min_{0\le j\le 2}\left[D(t-1,j) + \left(p(t,i) - p(t-1,j)\right)^2 \bullet 100\right] + c(i) \qquad (1)$$

$$c(i) = \begin{cases} 0, & \text{if } i = 1 \\ 1, & \text{if } i = 0 \text{ or } i = 2 \end{cases}$$

依音框編號 t 由小至大計算，到達最後之音框編號 T 時，再從三種 i 值的 *D(T, i)* 中選出最小者，作 back-tracking 來找出最佳路徑。此公式裡，*D(t, i)*表示從起始點走到音框 t、音高代號 i 之最小累積距離，*c(i)*表示行走半頻、或倍頻點的懲罰距離(penalty)。

2.3 軌跡片段尋找及分段

一個雙字詞的基週軌跡，理想上應該有兩個主要的片段(segment)，再加上一些長度較短的小片段，例如圖 5 是”回去”的波形及其基週軌跡，除了兩個主要片段之外，還存在許多較短的片段，其中被圈起來的是較長者。一個片落的認定是，片段中時間上相鄰的三個軌跡點，設它們的頻率值爲 x, y, z，則 |(z-y) – (y-x)| = |z-2y+x| 要小於一個門檻值，即斜率不可突然大幅度改變，在此設定的門檻值是 0.05。要找出主要的片段，我們首先依時間順序檢查各音框的基頻值，找出各個片段的起點及它的長度，再將所找到的片段按照時間長度作排序，然後就可把兩個最長的片段拿出，作爲雙字詞的兩個組成字的基週軌跡。

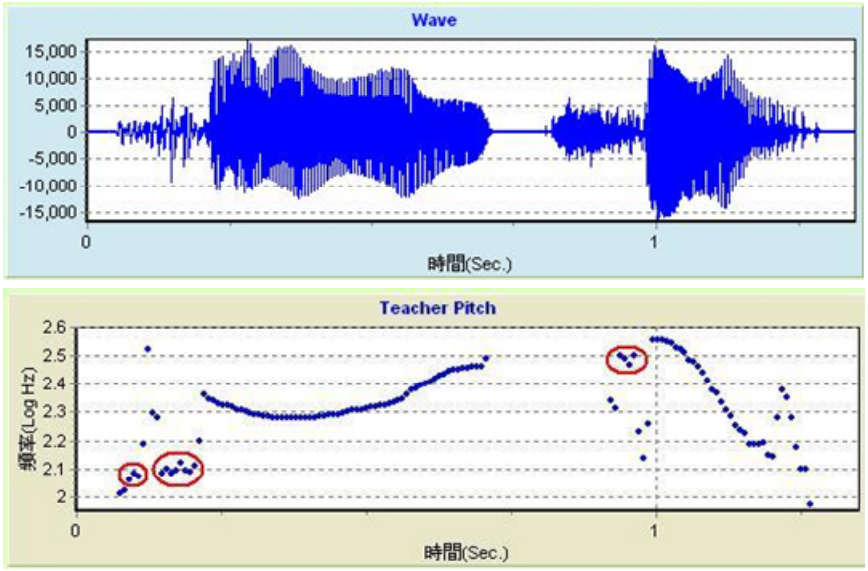


圖 5 “回去”的信號波形、及其基週軌跡

當一個雙字詞的兩個字之間發生連音現象時，如圖 4 顯示的波形和基週軌跡，則時間長度第一與第二的兩個片段，它們的長度差異會很大，例如第一長的超過 400ms 而第二長的卻小於 80ms，此時我們就要捨棄第二長的，而將第一長的軌跡片段作分段的處理。我們的分段方法是，先計算各音框的短時能量值，令其中最大能量值之 rms 值爲 Ex，接著設定 Ex/6 爲 rms 能量值的門檻，去檢查能量曲線，找出能量值低於門檻、並且落於第一長基週軌跡片段之時間範圍內的一個最長的時間片段，若找到了，則依此低能量的片段去對基週軌跡曲線作分段，若不能找到低能量的片段(表示能量曲線起伏不大)，就把門檻升高到 Ex/5、Ex/4，再重複作前述的動作。經由這裡的分段處理後，圖 4(c)裡一長條的基週軌跡曲線，就可切成如圖 6 所示的兩段的紅點片段了。

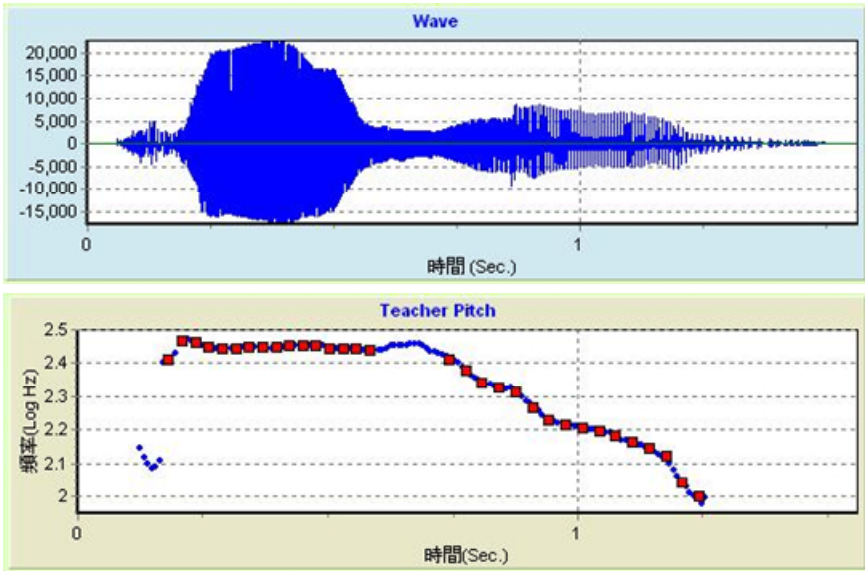


圖 6　基週軌跡分段處理

2.4 時間、音高正規化

一個音節的時間長度，會因爲說話者說話速度的不同，或同一個人不同次說話之速度的不同，而使得音長有長有短。爲了方便後續的處理，在此我們先作時間正規化之處理，將時間長短不一的音節基週軌跡，一律轉換成 16 維度的頻率向量來表示。作法是在一個音節的時間範圍裡均勻放置 16 個音高取樣點，然後以 Lagrange 內插法來求出各取樣點上的音高值[9]。一個時間正規化的例子，如圖 6 所示。

求出一個雙字詞兩音節共 32 點的時間正規化之基週軌跡 $g_0, g_1, ..., g_{15}$ $g_{16}, g_{17}, ..., g_{31}$ 後，接著需考慮音高正規化的問題，因爲不同使用者之間的音高位準差異是很大的，尤其是男、女生之間的音高差異。在此我們考慮了兩種正規化方法，第一種稱爲音節關聯式作法，先求出一個雙字詞發音的平均音高值，公式如下：

$$\mu = \frac{1}{32}\sum_{j=0}^{31} g_j \qquad (2)$$

再將基週軌跡每一點的音高值減去平均值，即令 $\alpha_j = g_j - \mu$，而求得音高正規化後的基週軌跡 $\alpha_0, \alpha_1, ..., \alpha_{15}$ ， $\alpha_{16}, \alpha_{17}, ..., \alpha_{31}$。另外一種音高正規化的方法，稱爲音節獨立式作法，先計算前後音節各自的平均音高值 $\mu 1$ 和 $u2$，接著前後音節的音高值分別減去各自的平均值，即 $\beta_j = g_j - \mu_1$，$0 \le j \le 15$；$\beta_j = g_j - \mu_2$，$16 \le j \le 31$，如此可求得另一種音高正規化的基週軌跡。

3. 樣式比對

3.1 語料準備

在基週軌跡的參考樣式的準備上，需要有標準發音與不標準的發音，這樣才能夠應付各種可能的使用者。因此我們收集了以國語爲母語者所唸的標準音，與外國人、聽障學生唸的非標準發音。參與錄音的語者共 28 人，其中 19 個人的發音作爲參考樣式，包括標準音樣本 4 位，外國人 13 位，聽障者 2 位。剩下的 9 個人的發音則作爲測試語料，包含標準音樣本 3 位，外國人 6 位。標準音樣本是請研究室裡發音標準的研究生，在實驗室裡錄製，外國人的語音樣本是在國際教會錄得，而聽障生的發音是在國小的啓聰班錄得。

雙字詞發音所用的詞彙包含：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，含蓋了 15 種雙字詞的聲調組合，每個詞彙中，第二個字都含有子音聲母，以避免發生連音現象，而造成參考音的基週軌跡樣式的求取錯誤。

3.2 人工評分

對於這些錄到的雙字詞發音，首先要對各個雙字詞的兩個組成字分別作人工評分的工作，有了參考分數之後，才能夠對不同的評分模式作測試及評估。這裡的評分方式是，分成五個等第，最高爲 5 分，最低爲 1 分，其於是 4、3、2 分。我們希望每一個字音所得到的分數，都能夠符合一般人耳的聽覺，因此將一般人所唸的聲調發音定爲 5 分，而將唸的比較差的聽障學生的聲調發音定爲 1 分，以此爲標準，將外國人所唸的發音拿去作人工評分，分數則介於 1 分到 5 分之間。

至於評分者的人選，係由五位實驗室的研究生來擔任，在評分前先聽過正常人與聽障學生的發音，依此爲標準才開始作評分。五位評分者對於同一個字音所給予的分數，我們以多數決來確定最終分數，遇到分數不能決定時，則取五個人評分的中間值作爲客觀分數。

3.3 距離量測

由於雙字詞不論是參考音或是測試音的基週軌跡，都已經先作了時間正規化和音高正規化的處理，使得一個音節的基週軌跡都是固定地表示成 16 維度的頻率向量，因此圖 1 裡的樣式比對方塊，其實只是作頻率向量之間的距離量測。

距離量測的方法中，最常被使用的是幾何距離，此外還有一個常見的是曼哈

頓距離，令 $X = < x_0, x_1, ..., x_{15} >$，$Y = < y_0, y_1, ..., y_{15} >$ 是兩個頻率向量，則幾何距離與曼哈頓距離的定義分別爲:

$$d_e(X,Y) = \sqrt{\sum_{i=0}^{15}(x_i - y_i)^2} \tag{3}$$

$$d_m(X,Y) = \sum_{i=0}^{15}|x_i - y_i| \tag{4}$$

本文研究了這兩種距離量測，觀察它們對於評分正確性的影響。

4. 評分決策

在語音辨識的應用裡，一般的觀念是直接把"**距離**"和"**分數**"連結在一起， 直覺認爲兩者是成正比的關係。但是在本研究裡，距離和分數並不是直接相關的，距離值只是作爲對參考樣式進行排序的依據，而參考樣式的人工評分數值才是作爲評分決策的直接依據。

由於每一種聲調組合的雙字詞的參考發音(19 個)，都已經過人工作評分，所以各個參考發音的兩個組成音節已有參考的分數。我們的評分方法是，將測試的雙字詞發音的前後音節之基週軌跡，和參考發音前後音節的基週軌跡分別作距離量測，再依距離值由小到大作排序，距離越小表示相似的程度越高，然後依據排名在前面的參考發音音節對應的人工評分，使用某一種評分決策方法來決定分數值。

關於評分決策的方法，本文嘗試了二種作法，第一種稱爲"直接平均法"，就是直接取前 M 名的分數來計算出平均值，第二種稱爲"中間平均法"， 先從前 M+2 名的分數中排除最大與最小的分數值，再計算剩下的分數的平均值。

以如下的例子來說，假設測試的雙字詞發音是"熨斗"，它的前後音節和各個參考發音的前後音節分別作距離量測後，再依距離值對前、後音節分別作排序而得到如圖 7 所示的排名次序。如果評分決策方法是**直接平均法**，則"熨"的分數值，當 M 爲 3 時是(5+4+3)/3 = 4；而"斗"的分數值，當 M 爲 2 時是(3+5)/2 = 4，依此可類推其它 M 值時的分數。如果分數決策方法是**中間平均法**，則"熨"的評分方式是，當 M 爲 3 時取距離排名前五名的分數值 5、4、3、5、3，然後刪去最大分數值 5 與最小分數值 3，再取剩下三個分數的平均而得到 4；"斗"的評分方式是，當 M 爲 2 時，取距離排名前四名的分數值 3、5、3、3，然後刪去最大分數值 5 與最小分數值 3，再取剩下二個分數的平均而得到 3，依此可類推其它

M 值時的作法。

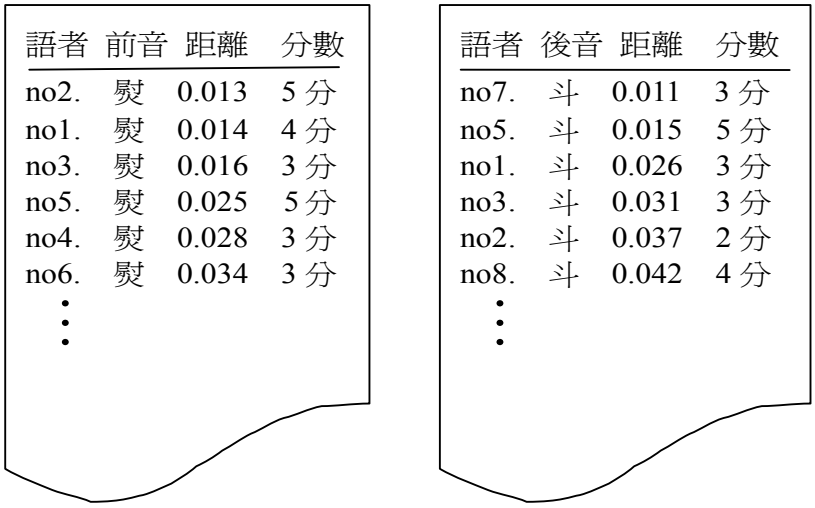


圖 7　前、後音節分別依距離值作排序

5. 評分實驗

5.1 音高正規化方法之比較

首先我們就 2.4 節提到的關聯式與獨立式之音高正規化方法，作評分效果的實驗，評分效果好壞的比較，我們以誤差分數的平均值 AVG，誤差分數的標準差 STD，和最大誤差分數 MAX 等三個統計項目來觀察。關於其它條件，距離量測先採取幾何距離，而評分決策方法先採取 M 值爲 3 之中間平均法。

實驗後，我們得到如表 1 和 2 所示的數值。觀察誤差平均值、誤差標準差、和最大誤差分數等項目在 ALL (四種聲調一起計算)欄下，可發現關聯式的都比獨立式的要好很多，此外就各個聲調分別來看，也是關聯式的都比獨立式的要好，因此以後的評分實驗裡，我們就都採取關聯式的音高正規化方法。

表 1. 關聯式音高正規化方法之評分誤差

	一聲	二聲	三聲	四聲	ALL
AVG	0.462	0.550	0.771	0.462	0.548
STD	0.383	0.462	0.539	0.442	0.469
MAX	1.333	2.000	2.333	1.666	2.333

表 2. 獨立式音高正規化方法之評分誤差

	一聲	二聲	三聲	四聲	ALL
AVG	0.699	0.638	0.993	0.583	0.711
STD	0.709	0.563	0.807	0.443	0.650
MAX	3.000	2.666	3.666	1.666	3.666

5.2 距離量測之比較

除了幾何距離，在 3.3 節裡還提到了曼哈頓距離，因此這裡就對這兩種距離量測作評分效果的比較，關於評分決策的方法，我們先採取 M 值爲 3 之中間平均法。實驗後，我們得到幾何距離的評分誤差情況如表 1 所示，而曼哈頓距離的評分誤差情況則如表 3 所示。

表 3. 曼哈頓距離之評分誤差

	一聲	二聲	三聲	四聲	ALL
AVG	0.444	0.555	0.641	0.486	0.524
STD	0.376	0.503	0.479	0.457	0.460
MAX	1.666	2.000	2.000	1.666	2.000

比較表 1 和表 3 的誤差分數值，可發現第二、四聲時，幾何距離的誤差比曼哈頓距離的小一些，而第一、三聲時則顛倒過來；若就整體來看，即看 ALL 欄，曼哈頓距離在 AVG, STD, MAX 等三個項目，則都比幾何距離的好一些。

5.3 評分決策方法之比較

在第四節裡提到了兩種評分決策方法，即直接平均法和中間平均法，因此這裡就對這兩種決策方法來作比較。每一種決策方法的實驗裡，我們分別對兩種距離量測作實驗，然後把四個聲調的評分誤差放在一起計算出誤差的兩種統計值 AVG 和 STD。此外當使用不同的決策人數、不同的 M 值時，也會得到不同的評分誤差統計數值，所以 M 也是一個需考慮的因素。

實驗後，對於兩種決策方法我們分別得到如表 4 和 5 的數值。在表 4 之直接平均法裡，最小的 AVG 值是 0.514，它是使用曼哈頓距離和決策人數設爲 1 時得到的，次小的是 0.518，則是使用幾何距離和決策人數設爲 2 時得到的，所以使用直接平均法時，決策人數 M 的選擇，會和所使用的距離量測有關係，在目前參考發音爲 19 個的情況下，M 值超過 2 後的評分誤差 AVG 值會持續變大、變差。在表 5 之中間平均法裡，最小的 AVG 值是 0.487，它是使用曼哈頓距離和決策人數設爲 2 時得到的，次小的是 0.490，則是使用幾何距離和決策人數設爲 2 時得到的，所以決策人數 M 的選擇，對於兩種距離量測來說，都是設定爲 2 是

最好的，M 值小於 2 或多於 2 都會使評分誤差 AVG 值變大、變差。如果就這兩種決策方法作比較，中間平均法的評分誤差 AVG 值 0.487，要比直接平均法的 0.514 好 5.3%；此外曼哈頓距離得到的評分誤差 AVG 值，都可以比幾何距離的稍小一些；至於 STD 值，一般來說會隨 M 值的增加而減小。

表 4. 直接平均法之評分誤差

	決策人數 / 誤差統計	1	2	3	4	5
幾何距離	AVG	0.548	0.518	0.543	0.566	0.634
	STD	0.707	0.542	0.474	0.454	0.436
曼哈頓距離	AVG	0.514	0.537	0.535	0.578	0.623
	STD	0.681	0.527	0.457	0.458	0.426

表 5. 中間平均法之評分誤差

	決策人數 / 誤差統計	1	2	3	4	5
幾何距離	AVG	0.529	0.490	0.548	0.640	0.698
	STD	0.623	0.518	0.469	0.426	0.437
曼哈頓距離	AVG	0.522	0.487	0.528	0.649	0.712
	STD	0.585	0.528	0.460	0.439	0.456

6. 結論

國語聲調的評量，尤其是多字語詞之組成字的聲調評量，電腦的評分必需接近人耳的評分，如此，外國人及聽障人士才能有效的使用電腦來學習國語。本文研究、製作了一個國語雙字詞聲調之評分系統，在基週軌跡預處理方面，研究了幾個改進基週軌跡正確性的方法，例如研究了基於 histogram 之音高範圍決定方法，以增進視覺效果；提出了基於動態規劃之倍頻、半頻的更正方法；研究了基於短時能量之基週軌跡分段方法。此外，我們也研究、實驗了不同的音高正規化方法，不同的距離量測方法，不同的評分決策方法，來觀察這些因素對於評分誤差大小的影響，實驗結果顯示，當使用關聯式音高正規化、曼哈頓距離量測、中間平均法之決策方法時，可以得到最小的評分誤差平均值 0.487 分，這相當於滿分 5 分的 9.7%。

目前所使用的測試語料只有 9 個人，還不是很充分，未來可以再錄製更多的語料來作進一步測試。此外，還有一些相關的問題，也可在未來作進一步考慮，例如音域範圍的問題，每個人的音域寬窄不一，並不是本文的音高正規化處理可以直接解決；還有每個人的講話方式(style)，也會有不小的差異，在國語雙字詞的基週軌跡上，就是兩字之間的相對的軌跡高度、斜率的變異問題。雖然本文並未直接去解決這樣的問題，但是我們採取的是間接的方式，來緩和此類問題的影響，那就是收集很多人的雙字詞發音來作爲參考樣式，因此目前收集的 19 人的參考發音，也許尚不足夠，還需收集更多人的。

致謝

感謝國科會計畫的支援，計畫編號 NSC 94-2614-S-019-001

參考文獻

[1] 金東垠，韓籍學生華語聲調錯誤分析與教學研究，碩士論文，國立臺灣師範大學華語文教學研究所，2003。

[2] 張可家、陳麗美，“日本學生學習華語的聲調偏誤分析: 以二字調爲例”，第十七屆自然語言與語音處理研討會(台南)，第 125-139 頁，2005。

[3] 鍾玉梅，「聽障兒童的說話問題」，聽語會刊，第 10 期，第 72-79 頁，1994。

[4] 梅永人，國語聲調電腦評量模式之研究，碩士論文，國立台中師範學院 教育測驗統計研究所，2000。

[5] 黃重光，以自組織特徵映射建立國語聲調電腦評量模式之研究，碩士論文，國立台中師範學院 教育測驗統計研究所，2001。

[6] 李俊毅，語音評分，碩士論文，國立清華大學 資訊工程研究所，2000。

[7] 蔡岳廷、廖嘉新、呂道誠、呂仁園，”台灣閩南語聲調評分系統評估與研究”，第十七屆自然語言與語音處理研討會(台南)，第 227-237 頁，2005。

[8] O`Shaughnessy, D., Speech Communication: Human and Machine, 2’nd ed., IEEE Press, 2000.

[9] 古鴻炎、張小芬、吳俊欣，“仿趙氏音高尺度之基週軌跡正規化方法及其應用”，第十六屆自然語言與語音處理研討會(台北)，第 325-334 頁，2004。

[10] 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.

[11] Rabiner, L. and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, 1993.

一種用於網路電話之遺失封包補償方法
A Packet Loss Concealment Method for Voice over IP

古鴻炎　　陳佳新
Hung-Yan Gu　　Zia-Sin Chen

國立台灣科技大學資工系
National Taiwan University of Science and Technology, Taipei, Taiwan
e-mail: guhy@mail.ntust.edu.tw　　http://www.csie.ntust.edu.tw/

摘要

本文提出一種用於網路電話的遺失封包補償之方法，稱為 TPPWI 法，它首先使用我們研究的基週偵測方法，來偵測遺失封包的前一個與後一個封包內的語音波形的週期長度，再將前一封包與後一封包各分類為有聲或無聲，依照有聲、無聲之四種組合再分成四種補償作法。其中前、後封包均為有聲的情況，我們提出了基週波形的相位同步方法，需佈放之基週個數與長度的決定方法，執行該二方法後，再執行特別的基週波形內差之作法，如此可得到高品質的重建語音。對於前、後封包均為無聲的情況，本文也將前人的作法加以改進。初步的聽測評估顯示，TPPWI 法可以得到比前人的數種方法更好的語音品質。

關鍵詞： 基週偵測，基週波形內差，相位同步，語音品質

1. 前言

隨著網路技術的日益蓬勃發展，硬體設備升級、線路頻寬加大，使得透過網際網路來即時傳輸語音封包變得可能，也就是所謂的 VoIP (Voice over IP)。然而，封包交換網路(packet-switching network)本身並不是非絕對可靠的，語音封包在傳輸的過程中，可能被遺失(network loss)，或可能被延遲送達，如果因此而超過了播放的時間點，接收端可能將之丟棄，形成 late loss，這樣的封包遺失(packet loss)對於 VoIP 所要求的對話即時性和語音品質來說，是一個很大的問題。VoIP 接收端的程式，它的工作就是將收到的語音封包解碼爲可供播放、聆聽的語音波形，因此當語音封包遺失時，若未作補償，就會解碼輸出一段靜音而形成空隙，使得語音波形變得不連續，因而產生讓聽者可察覺的雜訊噪音。

封包遺失的補償方法(Packet Loss Concealment，簡稱 PLC)可以分作兩大類[1, 2]，分別是從傳送端(transmitter)與接收端(receiver)的角度進行考量。傳送端補償方法的目標，在於如何降低封包傳送的失敗率與改善其分佈；而接收端補償方法，則致力於提升重建語音的品質。本文所研究的，是一種在接收端進行補償的

方法。

在接收端作補償的方法可細分爲：插入式(insertion)、內差式(interpolation)與再生式(regeneration)[1, 2]。插入式方法只會參考過去單邊方向(one-sided)的波形資料，以外差(extrapolation)處理方式來重建語音；內差式方法則會同時考慮遺失封包前後雙邊方向(two-sided)的波形資料，以內差處理方式來重建語音；而再生式方法則是從語音編、解碼器著手，重新合成出遺失封包的語音波形。一般而言，內差式補償方法的運算量較大，但是效果也比較好。考慮到今日個人電腦的速度，時脈(clock rate)都在 1 GHz 以上，要執行內差式的補償方法，運算能力是綽綽有餘，因此我們研究了一種內差式的補償方法，希望讓重建後的語音波形，聆聽時盡可能查覺不出有發生過封包遺失，並且希望在封包遺失率很高時（例如頻寬受限之撥接網路），可以得到大幅度的語音品質改進。

接收端內差式補償方法的原理是，參考遺失發生位置前後封包的資料，來產生兼顧前後連續性的重建語音，雖然步驟比較複雜一些，但是相對的效果也會有顯著的提升。相關的方法包括：(a)**相似波形取代法** [3]，利用樣式比對(pattern matching)的技巧，從過去播放過的波形資料中，尋找可以用來取代遺失封包所造成之空隙的合適波形，由於尋找的不一定都是未遺失的封包，也有可能是重建出來的語音波形，因此會有誤差延續(error propagation)的問題；(b)**基週波形複製法** [4]，此方法仰賴的是有效而可靠的基週偵測演算法，以便重複複製基週波形，直到填滿遺失空隙爲止，G.711 [5]內建的封包遺失補償方法便屬於這種；(c)**時域波形伸縮法 [1]**，藉由將過去波形伸展拉長，使之能夠涵蓋封包遺失所造成的空隙，進而達到重建語音的效果，WSOLA (Waveform Similarity Overlap Add)波形相似性疊加法便屬於這種作法 [6]，後來經過 Stenger 等人的修改，提出稱爲 Modified WSOLA 的作法 [7]，使之能夠作爲 VoIP 的接收端補償方法。

接收端補償方法對於遺失封包所重建出的語音信號，其品質好壞的評估可以從如下三項目來檢查，即重建的信號片段與其前後非遺失的信號之間，是否具有(a)振幅連續性(amplitude continuity)、(b)頻率連續性(frequency continuity)、及邊界上的(c)相位連續性(phase continuity)。振幅連續與不連續的對比，可比較圖 1 與圖 2，圖 2 中間重建出的波形，其振幅顯著地比兩邊波形的低；頻率連續與不連續的對比，可比較圖 1 與圖 3；相位連續與不連續的對比，可比較圖 1 與圖 4，圖 4 中間重建出的波形，在右邊界與下一封包交銜接處，出現了相位之不連續。熟知的語音波形重建方法之中，大多數只解決了振幅連續性與封包邊界波形連接的問題，對於頻率與相位連續性的問題考慮不多，或者成效不理想。

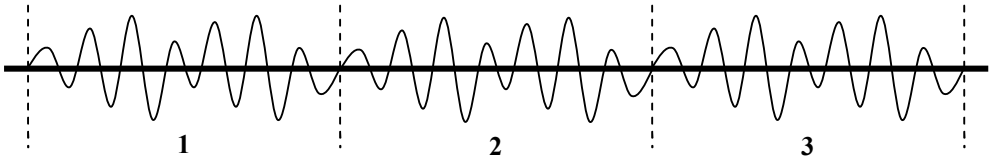


圖 1. 未發生封包遺失的三個連續的語音封包

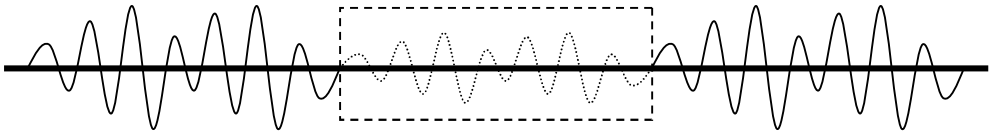


圖 2. 振幅不連續的例子

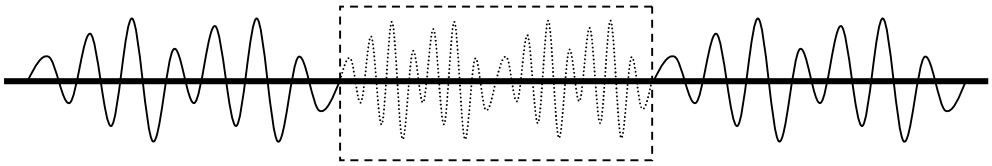


圖 3. 頻率不連續的例子

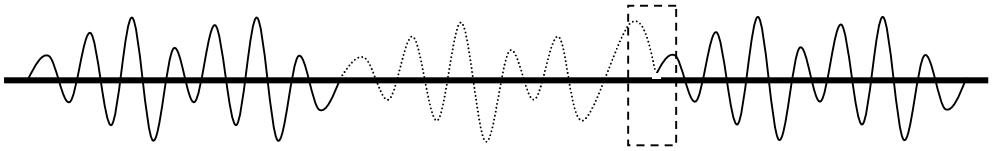


圖 4. 相位不連續的例子

2. 補償方法之架構

我們的接收端補償方法之架構如圖 5 所示，先分別從遺失封包之前的那個封包(簡稱"前一封包")與遺失封包之後的那個封包(簡稱"後一封包")中，去偵測基週長度、及判斷是否爲有聲(voiced)或無聲(unvoiced)的語音，PP (pitch of previous packet)表示前一封包語音的基週長度，PN (pitch of next packet)表示後一封包語音的基週長度；接著，由於前一封包與後一封包的分析爲有聲與無聲的結果共有四種組合，所以分別採取不同的波形重建之處理方式，即 BV (both voiced), PV (previous voiced), NV (next voiced), BU (both unvoiced)等處理方式。

雖然如圖 5 所示的架構，並非本文首先提出，而是參考 Liao 等人的雙邊基週波形複製法(Double Sided Pitch Waveform Replication，簡稱 DSPWR)裡的架構[8]，但是，這樣的架構很明顯地是必需採取的，如此才可能解決頻率與相位之連續性問題。

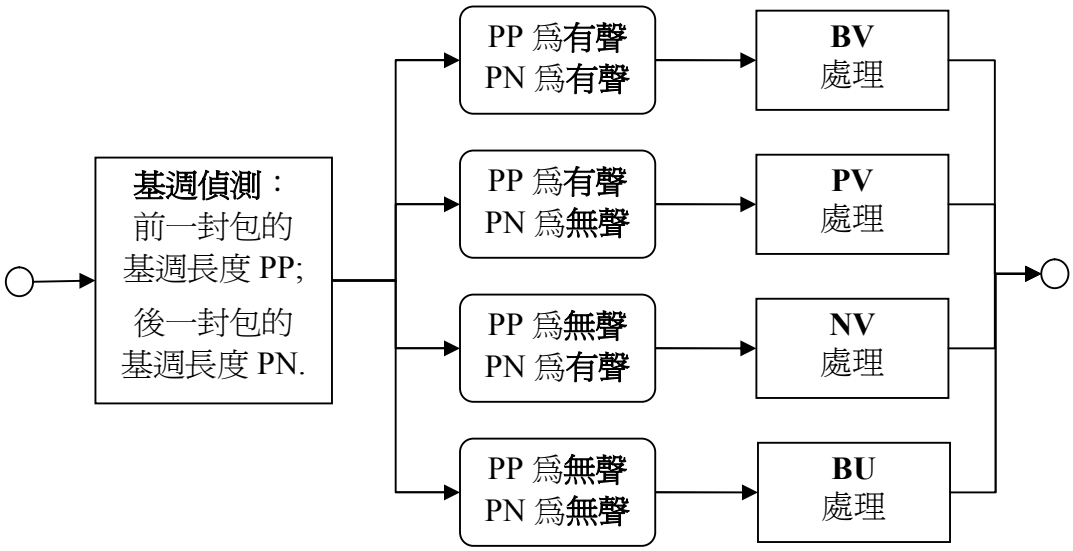


圖 5　補償方法的架構

我們的接收端補償方法稱爲時間比例式基週波形內差法(Time Proportion Based Pitch Waveform Interpolation, 簡稱 TPPWI)，雖然它和 DSPWR 法具有相同的架構，但是細部的處理方法上，則是南轅北轍且效能也有顯著差異，如圖 5 裡的"基週偵測"方塊、及右邊最重要的"BV 處理"方塊，在這兩方塊裡，我們的方法完全不同於 DSPWR 法裡的，且效能也比 DSPWR 法的好很多；在圖 5 裡的"BU 處理"方塊，我們則是改進 DSPWR 裡的作法，以得到更好的語音品質；至於"PV 處理"和"NV 處理"方塊，我們覺得是較少被執行到而較不重要的方塊，所以就直接採用 DSPWR 法裡的作法。

接著，先回顧 DSPWR 法的細部作法，以說明其缺點所在，而我們的 TPPWI 法的細部作法，則在第三、四、五節裡說明。DSPWR 所使用的基週偵測演算法[9]，是以偵測信號波形的峰點(peak)爲基礎，以峰點間的距離作爲基週長度，此種作法的一個明顯缺點是，週期性及週期長度的偵測，準確性不夠高，這會使得重建出的語音的品質降低不少。此外，DSPWR 法在前後封包均爲有聲的情況下，雖然提出一種解決相位不連續的作法，但是該作法有一個嚴重的缺點，就是未考慮音調(pitch)韻律的連續性，對於聲調語言(如漢語各方言)來說，音調的表現是重要的。

DSPWR 法在"BV 處理"方塊(PP 與 PN 都是有聲時)的作法如下，原理是利用相位比對(phase matching)的技巧[10]，來決定 PP 週期個數與 PN 週期個數之線性組合，以便讓基週波形複製填入到遺失封包所殘留的空隙時，能夠讓空隙內的重建波形與後一封包之間的波形相位差最小，藉以舒緩相位不連續的問題，該程序稱爲以基週進行相位比對（Phase Matching using Pitch，簡稱 PMP），最後再使用

線性振幅調整的技巧，來處理振幅連續性的問題。

DSPWR 法在"BU 處理"方塊(PP 與 PN 都是無聲時)的作法，改進了只以前一封包重複的方法[1]，改變成用前一封包的後半段波形與後一封包的前半段波形，來複製填滿遺失封包所產生的空隙，以避免當前一封包裡面包含有聲轉無聲或無聲轉有聲的波形轉變(transition)時，會產生讓人察覺得出來的噪音。

DSPWR 法在"PV 處理"方塊(PP 爲有聲而 PN 爲無聲時)的作法，原理是以 PP 基週波形來複製填滿遺失封包所產生的空隙，並且對重建波形施行線性振幅調整；由於 PN 被偵測爲無週期性，表示後一封包內的波形是隨機跳動之信號，故不存在相位連續性的問題。類似的情況，在"NV 處理"方塊裡，則以 PN 基週波形來複製填滿遺失封包所產生的空隙，並且對重建波形施行線性振幅調整。

3. 基週偵測方法及其實驗

3.1 基週偵測

在網路電話的應用裡，許多標準的語音編碼方法都是設定取樣率爲 8,000Hz，並且封包的長度最常被設定使用的是 20ms [11]，因此本文針對 20ms (160 個樣本點)的情況研究基週長度的偵測問題，所使用之偵測方法係基於波形相似性的量測，從同一個封包的語音樣本中，選取兩段相鄰且等長的訊號 $s(n)$ 與 $s(n+\tau)$ 來計算，計算的方式如公式(1)所示 [12]，計算後會得到一正規化自相關(Normalized Auto-Correlation，簡稱 NAC)函數，其函數值會介於 1 到-1 之間。

$$C_{NAC}(\tau)=\frac{\sum_{n=0}^{L-1}s(n)s(n+\tau)}{\sqrt{\sum_{n=0}^{L-1}s(n)^2}\sqrt{\sum_{n=0}^{L-1}s(n+\tau)^2}}\quad,\quad \tau=\tau_{\min},...,\tau_{\max} \tag{1}$$

$$L=\begin{cases}\tau\quad,\ \tau_{\min}\le\tau\le\frac{N}{2}\\ N-\tau\quad,\ \frac{N}{2}<\tau\le\tau_{\max}\end{cases} \tag{2}$$

其中 $s(n)$ 表示語音訊號樣本，τ 表示可能的基週長度之樣本點數，L 表示 AC 運算的範圍，其數值設定方式如公式(2)所示，N 表示一個封包的語音樣本點數，在 τ 小於等於封包長度之半的情況，L 設定爲與 τ 的數值一樣，而在 τ 大於封包長度之半的情況下，L 的設定則變成封包長度減去 τ 值，例如封包長度 N 爲 160 個樣本點，則當 τ 計數到 120 時，L 將會是 40(即 160 – 120)，亦即計算的是 $s(n)$ 與 $s(n+\tau)$ 兩段波形訊號的前 40 個樣本點之 AC 係數值。

以一段女性聲音的波形爲例(如圖 6 所示)，該波形長度爲 20 毫秒具有 160

個樣本點，若該波形爲"前一封包"，則正規化自相關函數(NAC)之運算將以由右至左的方向來偵測 PP 長度；若該波形爲"後一封包"，則會以由左至右的方向來偵測 PN 長度。以偵測 PP 長度爲例，其步驟如下：

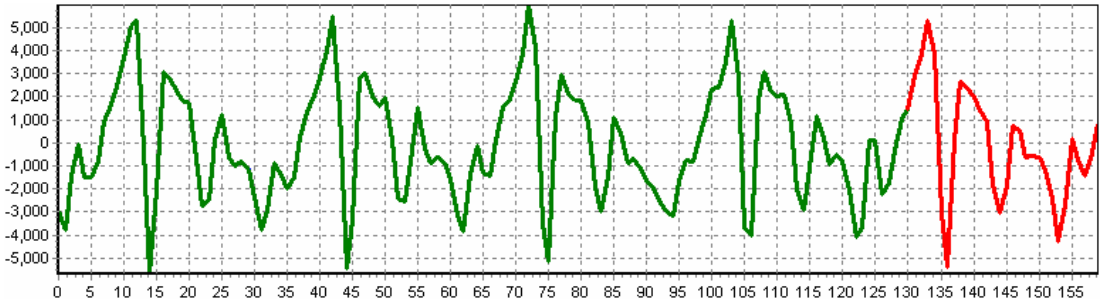


圖 6　女性聲音波形例子

步驟(1.1)：　將 $\tau_{\min}-1$ 到 $\tau_{\max}+1$ 範圍內的數值依序代入公式(1)，並將求得之 NAC 係數值連接起來，繪製成如圖 7 所示的函數圖形。

步驟(1.2)：找出 $\tau_{\min}$ 到 $\tau_{\max}$ 範圍內，NAC 函數的全域最大值之後，將其值乘以 0.8，設定爲局部最大值的門檻值 PeakTH，如圖 7 裡所示的水平線。

步驟(1.3)：由左至右對 NAC 函數大於 PeakTH 的數值作檢驗，檢驗其是否爲局部最大值(反曲點)，若是則記錄其 X 軸座標(即 τ 值)。以圖 7 爲例，局部最大值出現在 30、61 和 91 三個座標點。

步驟(1.4)：令最小 τ 值者爲基週長度，驗證其餘的 τ 值是否爲其倍數。以圖 7 爲例，τ 值最小者爲 30，第 2 小者爲 61，61 位在可允許的兩倍週期範圍內，亦即介於 $(30-5)\times 2=50$ 到 $(30+5)\times 2=70$ 之間，而且第 3 個 τ 值 91，也介於 $(30-5)\times 3=75$ 到 $(30+5)\times 3=105$ 之間，因此判定 30 爲所求的 PP 長度。

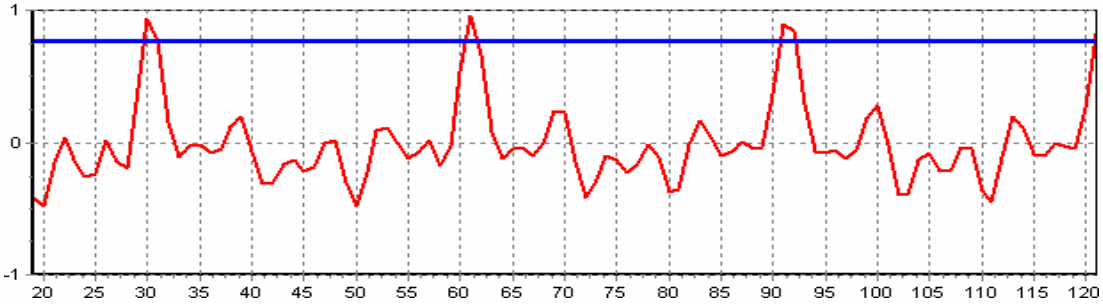


圖 7　NAC 函數圖形

觀察圖 6 的波形，可以確認該封包的 PP 週期長度大致爲 30 個樣本點，而整個封包內大約有 5 個這樣的基週波形。

3.2 基週確認

在 NAC 函數圖形的 PeakTH 水平線之上，由於女性聲音的基週長度較短，因此女性聲音通常會有多個峰點，而男性聲音的基週長度較長，大多只會有一個峰點。在大於 PeakTH 的峰點只有一個的情況下，爲了強化對無聲雜訊與男性有聲聲音的區別能力，本文使用一個波形相似度的判斷：若此峰點的橫座標τ值小於等於 50，則此峰點的 NAC 數值必須大於 0.8(波形相似度 80%)，反之，若τ值大於 50，因爲波形內容的可能變化較多，因此 NAC 係數值只需要大於 0.6(波形相似度 60%)。當 NAC 數值有超過所設定的門檻，才將τ值當作基週長度輸出；對於大於 PeakTH 的峰點超過一個的情況下，採用的相似度門檻值則爲 0.8。

爲了增進基週偵測演算法的正確率，將原先只是偵測前一封包右側與後一封包左側之 PP 與 PN 長度，修改成如下的處理流程：從"前一封包"的右側與左側，分別去偵測左側的基週(PP Left, PPL)與右側的基週(PP Right, PPR)，再從"後一封包"的右側與左側，分別去偵測左側的基週(PN Left, PNL)與右側的基週(PN Right, PNR)；然後，進行基週再確認，即 PPR 與 PPL 比較，PNR 與 PNL 比較，以得到最有可能的 PP 與 PN。

基週再確認的方法，以"前一封包"爲例，若 PPR 爲 0、而 PPL 爲 30，則考慮 PPR 在偵測過程中所繪製的 NAC 函數圖形，檢查τ = 30 的位置附近是否存在 NAC 係數大於 0.6(波形相似度 60%)的局部最大值；若 PPR 不爲 0、而 PPL 爲 0 時，也進行類似的處理步驟，藉以更正有聲誤判爲無聲的情形。若 PPL 與 PPR 皆不爲零，但數值較大者除以數值較小者的比值大於 1.4，則表示兩者數值差異過大，在同一個封包中基週長度不應會發生如此劇烈的變化(例如：PPL 爲 45、而 PPR 爲 30)，此時就尋找 PPR 的 NAC 函數圖形中τ爲 45 的附近，是否存在 NAC 係數值大於 0.6 的局部最大值 PPR'，若未找到則令 PPR'爲零；同樣地，在 PPL 的 NAC 函數圖形中τ爲 30 的附近，找尋 NAC 係數值大於 0.6 的局部最大值 PPL'，若未找到則令 PPL'爲零；接著，計算τ值皆約爲 30 的 PPR 與 PPL'的幾何平均數 GM1，以及τ值皆約爲 45 的 PPL 與 PPR'的幾何平均數 GM2；若 GM1 > GM2，則輸出 PPR 之長度 30，若 GM1 < GM2，則輸出 PPL 之長度 45，藉此更正"週期長度誤判"的情形。

3.3 測試實驗

關於基週偵測的測試實驗，我們使用了一些雙字詞的發音語料，其中女性聲音與男性聲音各有十句、各兩組聲音來源，聲調組合包括：下降(44)、加快(14)、右轉(43)、上昇(41)、取消(31)、前進(24)、停止(23)、啓動(34)、清除(12)、關機

(11)。

實驗的方式是，每個封包都會被當作是"前一封包"與"後一封包"，然後分別由右至左與由左至右去偵測 PP 與 PN 的基週長度，所以基週偵測的次數是封包數目的兩倍。程式偵測出 PP 與 PN 數值後，再以人工方式檢查所得的數值的正確性，結果如表 1 所示。其中"誤判週期"表示基週長度被誤判爲實際週期的倍數，"誤判有聲"表示無聲信號被誤判爲有聲，"誤判無聲"表示有聲信號被誤判爲無聲。

表 1　基週偵測方法之正確率

雙字詞 測試語料	封包 數目	偵測 次數	誤判 週期	誤判 有聲	誤判 無聲	正確率
第 1 位女聲	320	640	2	1	8	98.28%
第 2 位女聲	282	564	0	6	1	98.75%
女聲統計	602	1204	2	7	9	98.52%
第 1 位男聲	265	530	1	7	17	95.28%
第 2 位男聲	249	498	16	5	14	92.97%
男聲統計	514	1028	17	12	31	94.12%
整體統計	1116	2232	19	19	40	96.32%

實驗結果顯示，我們的基週偵測方法對於女性聲音的平均偵測正確率爲 98.52%，對於男性聲音的平均偵測正確率爲 94.12%，整體的平均偵測正確率是 96.32%。男性聲音的基週偵測正確率較低，是因爲男聲基週比較長，在使用公式(1)作計算時，計算範圍會受到限制。以第 2 位男性聲音爲例，他的聲音比較低沈，基週長度經常超過 100 個樣本點，所以他的基週偵測正確率只有 92.97%。

4. 雙邊有聲 BV 之處理

在前、後封包均爲有聲(PP 與 PN 均爲有聲)的情況下，補償方法的好壞，對重建語音的品質影響是最大的。本文的 TPPWI 法，其主要處理步驟是，先對左、右封包的基週波形取得相位同步並作相位對齊，再對修補後的剩餘空隙訂定需要重建的週期個數、與決定各週期的長度，然後進行基週波形內差處理以產生出各週期的波形內容。這裡的基週波形內差之觀念、作法，是啓發自我們先前研究國語語音合成時提出的 TIPW 信號波形合成法[13]，此外由於 BV 處理是補償方法之重點，因此我們的補償方法的名稱 TPPWI，就取自基週波形內差的縮寫。

4.1 基週波形之同步

由於基週偵測方法只決定基週的長度，因此需要進一步確定一個基週內的波形內容。習用的方法是，將前一封包內最接近遺失封包的 PP 個樣本點，作爲前一封包之基週波形(Previous Pitch Waveform, PPW)，而將後一封包內最接近遺失封包的 PN 個樣本點，作爲後一封包的基週波形(Next Pitch Waveform, NPW)。但是，本文考慮到下一步驟裡，要對雙邊封包的基週波形作內差，所以必需先對雙邊基週波形的峰點位置(波峰)取得同步(相位同步)，否則當差異太大時(如波峰對應到波谷)，內差得到的振幅將會趨近於零。對於基週波形取得同步的處理，以"前一封包"爲例，我們方法的步驟如下：

步驟(2.1): 從前一封包中，選取最右邊、長度爲 PP 個樣本點的波形，作爲初始的前一基週波形 PPW'，如圖 8 所示。

步驟(2.2): 找到 PPW' 的波峰，以此點爲分界，將波形切成左右兩段，再將前半段波形複製到後半段波形的後面，成爲前一封包的基週波形 PPW，如圖 9 所示，如此調整後的基週波形之起點將由波峰開始。

步驟(2.3): 比對 PPW'與 PPW 的相位差，然後在遺失封包的開頭，填入所需的樣本點數，以修補成一個同步過的 PPW 基週波形。圖 8 在作相位修補之後，會變成如圖 10 所示的情況。

至於"後一封包"的 NPW 之選取，也可採取相同之方法，以補滿封包邊界的 PPW 與 NPW 之基週波形，如此就可以解決封包邊界波形平順連接的問題。

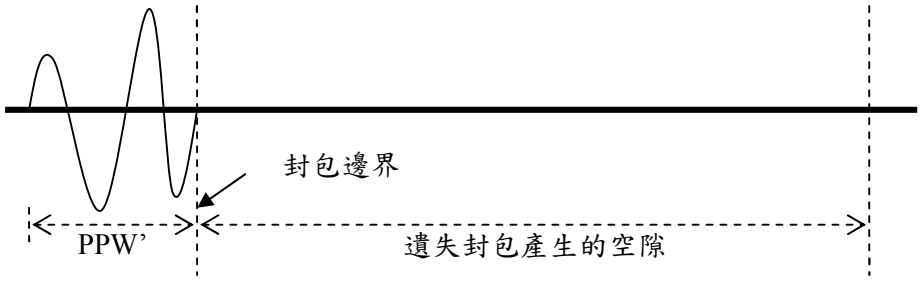


圖 8　前一封包選取出之 PPW'波形

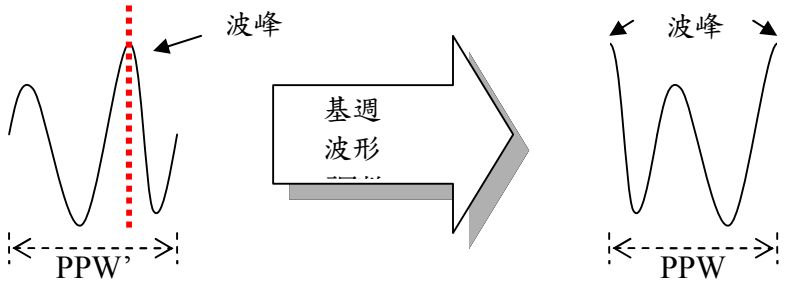


圖 9　從 PPW' 轉換出波峰爲起點之 PPW 基週波形

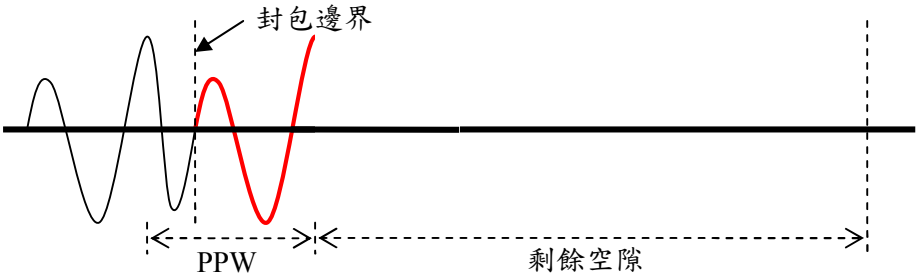


圖 10　補滿封包連接處的基週波形

4.2　基週個數與長度訂定

雙邊封包的基週長度關係共有三種情況，亦即 PP > PN，PP < PN，或 PP = PN，因此剩餘空隙內要填入的重建基週波形(Reconstructed Pitch Waveform, RPW)的個數與長度，決定之方法也稍有不同。令剩餘空隙的長度爲 r，首先就依據 r 來決定需重建出的基週個數，其步驟如下，將 r 除以 PP，得到 PPW 最多可以填入的個數爲 a；將 r 除以 PN，得到 NPW 最多可以填入的個數爲 b；取 a 與 b 的平均再作四捨五入，即爲剩餘空隙內要填入的基週個數，令此數值爲 N_p。

接著，以 PP < PN 及 PP=PN 的情形爲例，訂定 RPW 長度的步驟如下：

步驟(3.1): 以 PP 爲基礎，賦予該 N_p 個 RPW 的初始長度皆爲 PP，如圖 11 所示。將基週長度想像成是"磚塊"數目，如此調整基週長度就相當於調整磚塊的堆積高度。

步驟(3.2): 依據斜率，線性地爲這 N_p 個基週疊上磚塊。以圖 12 爲例， PP 爲 3、PN 爲 7、N_p 爲 3，故斜率爲 4/3，因此這三個 RPW 的長度就依此斜率來作微調，分別是 4/3、(4/3)*2 = 8/3 和(4/3)*3 = 4，四捨五入後分別爲多疊 1 個、3 個與 4 個磚塊。

步驟(3.3): 將線性微調後的 RPW 長度相加，得到填入的總長度 c，以圖 12 爲例，此時的總長度爲 4 + 6 + 7 = 17。接著將 c 與 r 相減，得到差值 d。若 d = 0，表示 RPW 的長度在步驟(3.2)的微調處理後，已經是最佳的組合了。

步驟(3.4): 若 d < 0，表示還需要再疊一些磚塊，此時便以由左至右(由低至高)的順序，逐次對一個 RPW 多加 1 個磚塊，直到 c 等於 r 爲止，多加磚塊的順序是第 1 個 RPW、第 2 個 RPW、第 3 個 RPW、回到第 1 個 RPW…。

步驟(3.5): 若 d > 0 的話，則表示需要拿掉一些磚塊，此時便以相反的順序，每次取下 1 個磚塊，直到 c 等於 r 爲止，取下磚塊的順序是第 3 個 RPW、

第 2 個 RPW、第 1 個 RPW、回到第 3 個 RPW…。

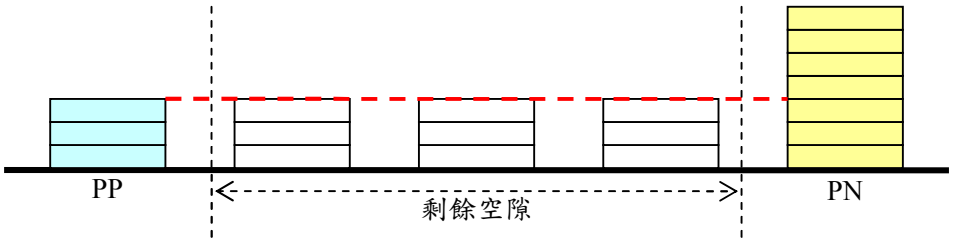


圖 11　填入等長之基週至剩餘空隙

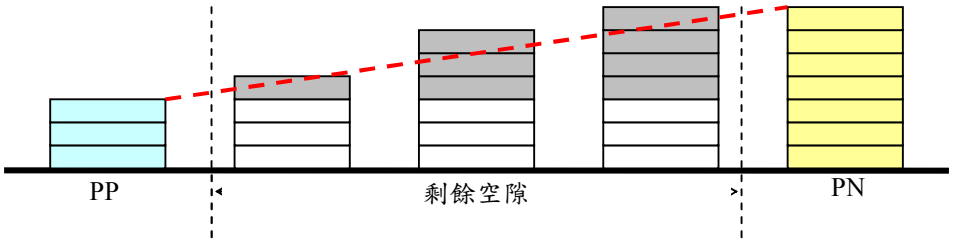


圖 12　線性方式微調後的磚塊數目

至於 PP > PN 的情形，亦可採相同之方法，只要將圖 11 的示意圖反轉過來，變成磚塊堆得較高者在左邊即可。

4.3　基週波形內差

本文作基週波形內差之方法，步驟如下：

步驟(4.1): 雙邊封包的基週波形 PPW 與 NPW 在 4.1 節便已經決定好了，這裡首先依線性時間比例，來決定它們的內差加權值 w1 與 w2，如公式(3)所示，

$$w1 = \frac{r-g}{r} \quad , \quad w2 = \frac{g}{r} \tag{3}$$

其中 r 表示剩餘空隙的長度，g 表示要重建的那個 RPW 的中心點位置。例如在圖 13 中，剩餘空隙內只夠填入一個 RPW，因此計算出來的 $w1 = w2 = 0.5$。另外由公式(3)可知，在剩餘空隙中，離 PPW 較近的 RPW，波形會比較近似於 PPW，相反地，離 NPW 較近的 RPW，其波形會較像 NPW。

步驟(4.2): 分別將 PPW 與 NPW 的樣本點乘以 $w1$ 與 $w2$。例如在圖 13 中，相乘後基週波形的振幅將由-6000 至 5000 的範圍，降爲-3000 至 2500。

步驟(4.3): PPW、NPW 與 RPW 的長度一般來說都不一樣，此時可使用如公式(4)所示的餘弦窗(cos window) ，

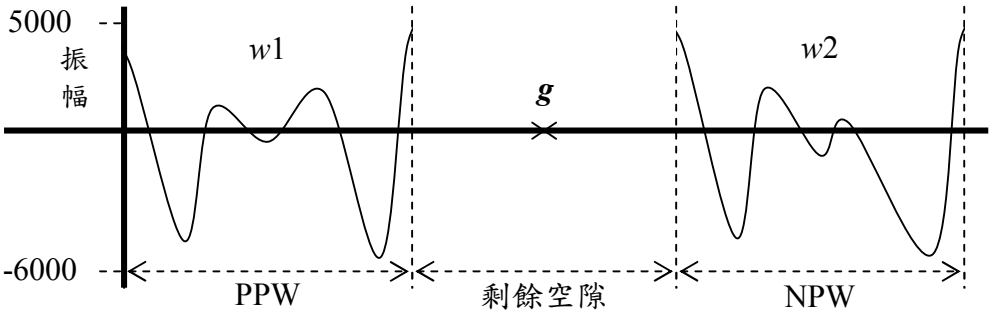


圖 13　加權值 *w1, w2* 與重建基週之中心點 g

$$w(n) = 0.5 + 0.5 \times \cos\left(\frac{2\pi n}{N}\right) \ , \ 0 \le n \le N-1 \qquad (4)$$

來將 PPW 與 NPW 的波形長度作伸展或壓縮，以便調整成和 RPW 的長度相同。這裡我們不能以 resampling 方式來作，因爲單純的 resampling 處理，會破壞原語音的音色[13]。餘弦窗的函數圖形如圖 14 所示，數值介於 0 到 1 之間。

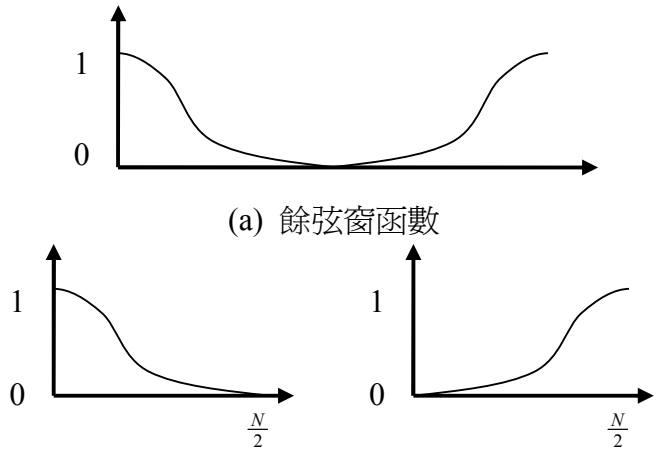


圖 14　餘弦窗的函數圖形

<u>伸展處理</u>**:** 假設要將長度 30 之 PPW 伸展成長度 35，以配合長度 35 之 RPW，那麼就先設定餘弦窗長度 N = 30 × 2 = 60，再將 PPW 的 30 個樣本點乘上左半邊餘弦窗(如圖 14(b)所示)，結果放至 RPW 框的左邊，且與 RPW 框左邊界對齊，如圖 15(a)所示的 PPW-L 波形；接著再將 PPW 乘上右半邊餘弦窗(如圖 14(c)所示)，結果放至 RPW 框的右邊，且與 RPW 框右邊界對齊，如圖 15(b)所示的 PPW-R 波形；然後對重疊的部分作疊加處理。

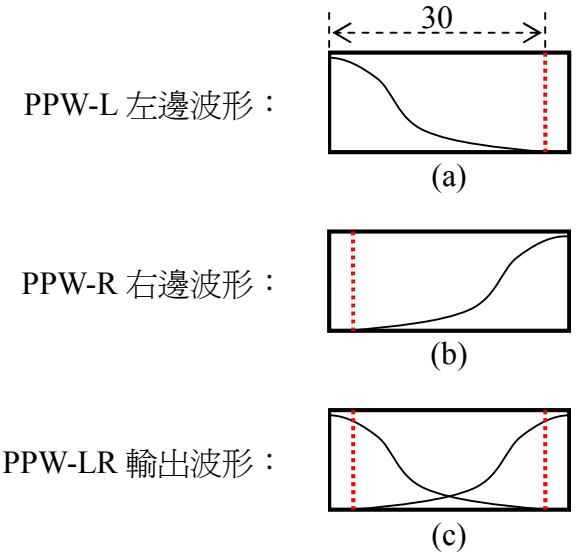


圖 15　以餘弦窗作波形長度伸展

壓縮處理: 假設要將長度 30 之 PPW 壓縮成長度 25，以配合長度 25 之 RPW，則先設定餘弦窗長度 N = 25 × 2 = 50，再將 PPW 的前面 25 個樣本點乘上左半邊餘弦窗，結果放至 RPW 框的左邊，且與 RPW 框左邊界對齊，如圖 16(a)所示的 PPW-L 波形；接著再將 PPW 的後面 25 個樣本點乘上右半邊餘弦窗，結果放至 RPW 框的右邊，且與 RPW 框右邊界對齊，如圖 16(b)所示的 PPW-R 波形；然後對重疊的部分作疊加處理，而得到 PPW-LR 輸出波形。

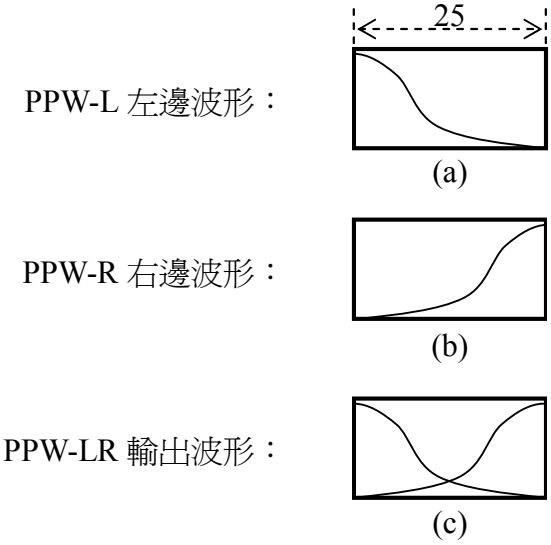


圖 16　以餘弦窗作波形長度壓縮

步驟(4.4): 經過步驟(4.3)的波形伸縮之後，基週波形 PPW、NPW 與 RPW 的長度都變成相同，而且 PPW 與 NPW 的內容在步驟(4.2)已經乘以加權值了，因此只要將它們的每個樣本點相加，即可得到 RPW 波形。圖

17 是將圖 13 的 PPW 與 NPW 基週波形經過振幅加權、長度伸縮與樣本點相加之後，所重建出的 RPW 波形。

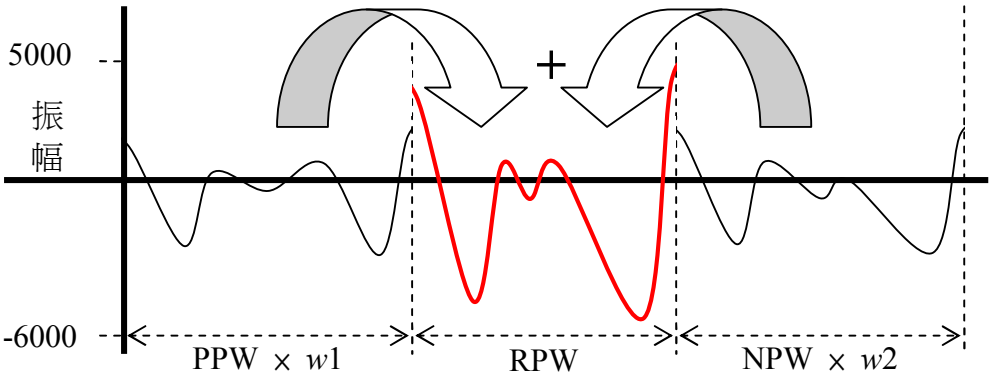


圖 17　重建出的 RPW 波形

5. BU、PV 與 NV 之處理

5.1 雙邊無聲 BU 之處理

當基週偵測方法在遺失封包雙邊的封包內都偵測不到週期時，表示遺失封包的雙邊封包內的波形是無聲信號，此時就以 BU (Both Unvoiced)方式處理。以圖 18 爲例，若該波形爲一個前一封包內的語音波形，隨著時間該波形也由振幅較大且具有週期性的波形，逐漸轉爲振幅較小且無週期性的波形，因爲在前一封包當中，基週偵測是由右至左進行的，因此會偵測不到語音訊號的週期。爲了避免該無聲訊號內可能包含有雜訊的例外狀況，如圖 18 後半段封包的第 2 個四分之一波形所示，需先計算後半段封包的兩個四分之一波形內，各自的峰谷間振幅值；接著，將該二峰谷間振幅值之中數值較大者除以數值較小者，得到一個差異倍數的比值；若該比值大於 1.4，表示可能存在振幅突然變大的雜訊，故只以峰谷間振幅值較小的那段四分之一波形來重複填入兩次到遺失封包內；反之，則表示無聲訊號的變動不大，因此可以拿前一封包的後半段波形來填入遺失封包的空隙內。後一封包亦可以以相同方式來處理。

峰谷間振幅值的差異倍數過大，除了可能存在振幅突起的雜訊之外，基週偵測演方法發生誤判也是可能的原因之一，因此透過上述 BU 方式處理，即使基週偵測錯誤，將有聲波形誤判爲無聲波形時，仍可減小錯誤波形相位不連續所引發的噪音。

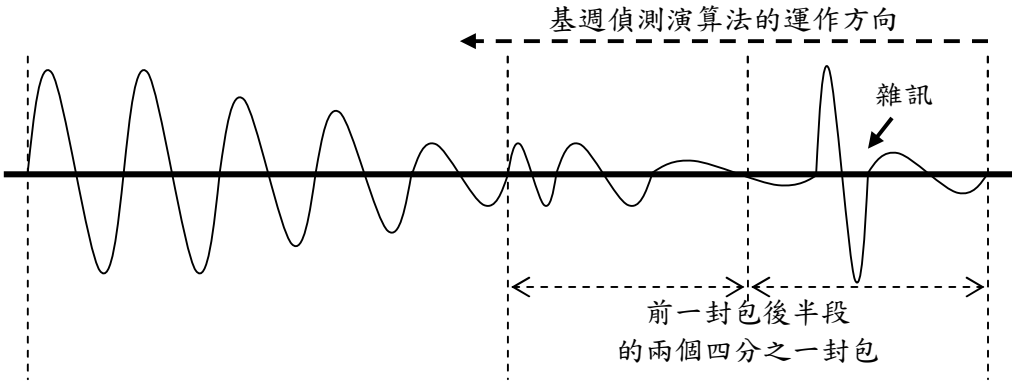


圖 18　週期性逐漸消失的波形例子

5.2　單邊有聲 PV 與 NV 之處理

當 PP 爲有聲、而 PN 爲無聲時，就進行 PV 處理，以 PP 基週波形來複製、填滿遺失封包所產生的空隙，並對重建波形施行前向振幅調整(如圖 19 所示)。

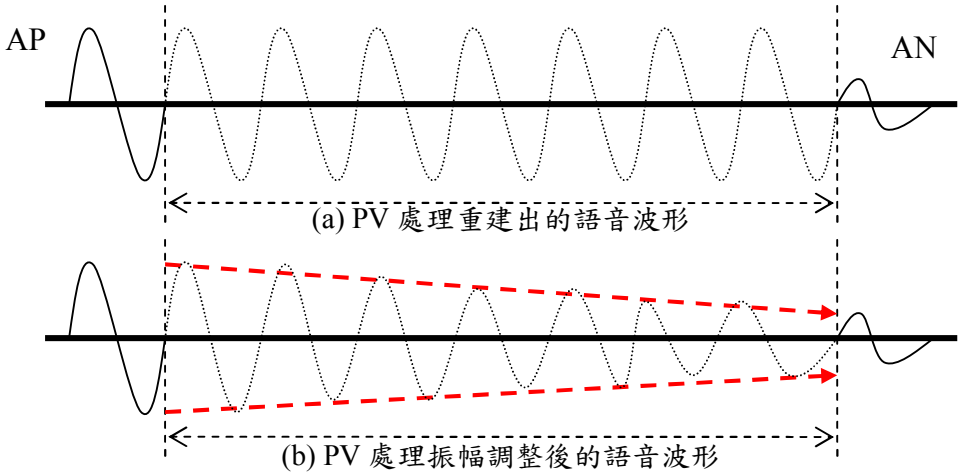


圖 19　前向振幅調整之示意圖

爲了方便說明，我們將前一封包中右邊 PP 個樣本點內的正負峰間振幅值(peak-to-peak amplitude)稱爲 AP，而將後一封包左邊 PP 個樣本點內的正負峰間振幅值稱爲 AN。正負峰間振幅值係指一定範圍內，最高波峰與最低波谷之間的差距。重建波形之前向振幅調整，方法如下：

步驟(5.1):　factor = (AN - AP) / AP / N；

計算振幅需要微調的幅度，其中 AP 與 AN 分別是**前一封包**與**後一封包**中，最鄰近重建波形的 PP 個樣本點內，所求得的正負峰間振幅值；N 爲遺失封包空隙的長度，以樣本點計數。如圖 19 裡，後一封包之無聲信號的振幅較小，故 factor 會是負值。

步驟(5.2):　for（i = 0;　i < N;　i++）　s[i] *= 1 + factor * i；

由左至右逐點調降重建出信號的樣本點振幅值，其中 $s[i]$表示遺失封包內重建出的信號樣本點序列。圖 19 說明了 PV 處理後，信號振幅逐漸由大變小之情形。

當 PP 爲無聲、而 PN 爲有聲時，就進行 NV 處理，以 PN 基週波形來複製填滿遺失封包所產生的空隙，其作法相同於 PV 處理的，只是要顛倒左右的方向。

6. 補償方法之測試實驗

6.1 SNR 訊噪比測試

實驗使用的語音資料，是取自電視連續劇中，兩名演員在餐廳內的對話，取樣率 8000 赫茲、解析度 16 bits、單聲道，時間長度上可以封裝成 20 毫秒爲單位的封包共計 2993 個(59.86 秒鐘)。

在模擬封包遺失然後重建其語音波形的實驗中，封包是否遺失，是以隨機亂數搭配白努力(Bernoulli)模型來決定，分成 10%、30%與 50%三種遺失率的情況來模擬。每種遺失率的情況下，我們都作了三次實驗，也就是對同一種封包遺失率模擬產生出三種不同的封包遺失之序列，然後在接收端補償方法處理之後，分成三次計算原始波形與重建波形的 SNR (Signal-to-Noise Ratio)訊噪比，然後取三次計算結果之平均值作爲實驗數據。

這裡假設每種接收端補償方法只針對封包連續遺失個數小於等於 3 的情況作處理，理由是：(a)當封包連續遺失個數大於 3 時，重建語音的效果已經明顯不好；(b)同時將多個封包放入 Jitter Buffer 內，可能會影響 Jitter Buffer 的處理效能[11]。因此，除了過去封包重複法之外，其餘補償方法在封包連續遺失個數大於 3 時，便只重複前一次播放過的封包。這裡所作的封包遺失補償之實驗，測試了幾種方法，除了基本的過去封包重複法(Packet Repetition[1]，以 Rep 表示)和本論文提出的 TPPWI 法之外，我們也自行將外差式的相似波形取代法(Waveform Substitution based on Pattern Matching [3]，以 PM 表示)、改良式波形相似性疊加法(Modified Waveform Similarity OverLap Add [7]，以 WSOLA 表示) 、與雙邊基週波形複製法(Double Sided Pitch Waveform Replication [8]，以 DSPWR 表示)，寫成可作實驗之程式。

由實驗計算得到的 SNR 值如圖 20 所示，SNR 值的計算公式如下:

$$SNR = 10 \times log_{10} \frac{\sum_{n=0}^{N} [x(n)]^2}{\sum_{n=0}^{N} [x(n) - \tilde{x}(n)]^2} \tag{5}$$

其中 $x(n)$ 是原始信號樣本、$\tilde{x}(n)$ 是重建的信號樣本。

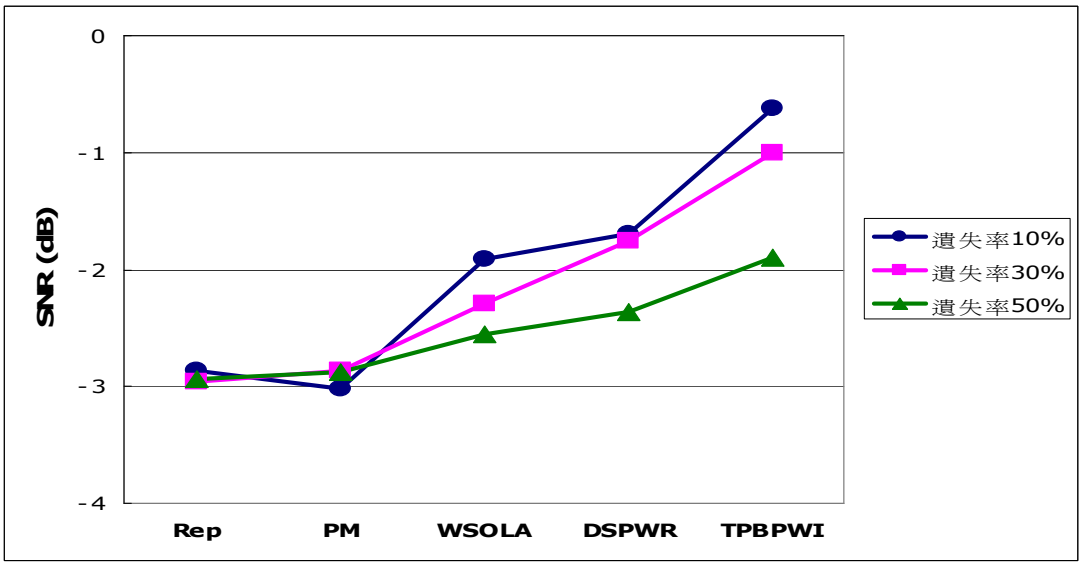


圖 20　重建語音之 SNR

如圖 20 所示的 SNR 的量測結果中，不論在 10%、30%與 50%的封包遺失率之模擬實驗裡，本論文研究之 TPPWI 方法均優於其他四種方法。不過，圖 20 裡重建出的語音波形的 SNR 均爲負數，參考 SNR 的計算公式，這表示在 log10 函數內，分母的數值大於分子，使得相除結果爲分數，而分母會大於分子，表示此時原始語音的樣本點 $x(n)$ 與重建語音的樣本點 $\tilde{x}(n)$ 之間的振幅值差異很大，這意謂兩者的波形存在相位差、峰點位置不同等等的情況。

6.2　聽測評估測試

針對前述五種接收端補償方法，我們參考 6.1 節的 SNR 測試結果，對它們作排名，第一名到第五名的順序是：TPPWI、DSPWR、WSOLA、Rep 與 PM。接著，我們將這五種方法作配對，第一名與第二名、第一名與第三名、第三名與第四名、第三名與第五名共四種組合，每種組合內包含兩個補償方法處理後的音檔，並且是在模擬封包遺失率爲 10%、30%與 50%的情況之下，以相同遺失序列去分別處理得到的，然後讓受測者比較每種組合內的兩個音檔之語音品質差異。評分方式是在受測者不知道組合方式的情況之下，請他們挑選出那一個音檔的語音品質比較好，以及好多少。如果"分不出好壞"的話，就計爲 0 分；"好一些"的

話，就計爲 1 分；”明顯好”的話，就計爲 2 分；”好很多”的話，就計爲 3 分。我們總共邀請了 16 位受測者，年齡介於 23 到 25 歲之間，其中 7 位是我們實驗室的成員，另外 9 位是實驗室成員的友人，不具語音專業背景。測試的語料是唐詩楓橋夜泊，先由女生唸一遍、再由男生唸一遍，封裝成 20 毫秒爲單位的封包，女性聲音封包數目 535 個、男性聲音封包數目 512 個，總音長 20.94 秒鐘。

聽測後，對 16 位受測者的評分求出平均值，詳細數值如表 2 所示，再依表 2 的數值畫圖，得到圖 21 所示的曲線，從此圖可以看到，在 10%的封包遺失率之情況下，TPPWI 比第二名的 DSPWR 好了 1.625 分，接近”明顯好”，此時的補償方法效能排序是：WSOLA < PM < Rep < DSPWR < TPPWI；在 30%的封包遺失率之情況下，TPBPWI 比第二名的 DSPWR 好了 1.375 分，比”好一些”再好一點，此時的補償方法效能排序是：PM < WSOLA < Rep < DSPWR < TPBPWI；在 50%的封包遺失率之情況下，TPBPWI 比第二名的 DSPWR 好了 0.875 分，接近”好一些”，此時的補償方法效能排序是：WSOLA < PM < Rep < DSPWR < TPPWI。

表 2 聽測實驗之平均分數

封包遺失率	PM	Rep	WSOLA	DSPWR	TPBPWI
10%	0.0625	0.625	0	0.75	2.375
30%	-0.125	0.6875	0	1.0625	2.4375
50%	0.8125	1.3125	0	1.5625	2.4375

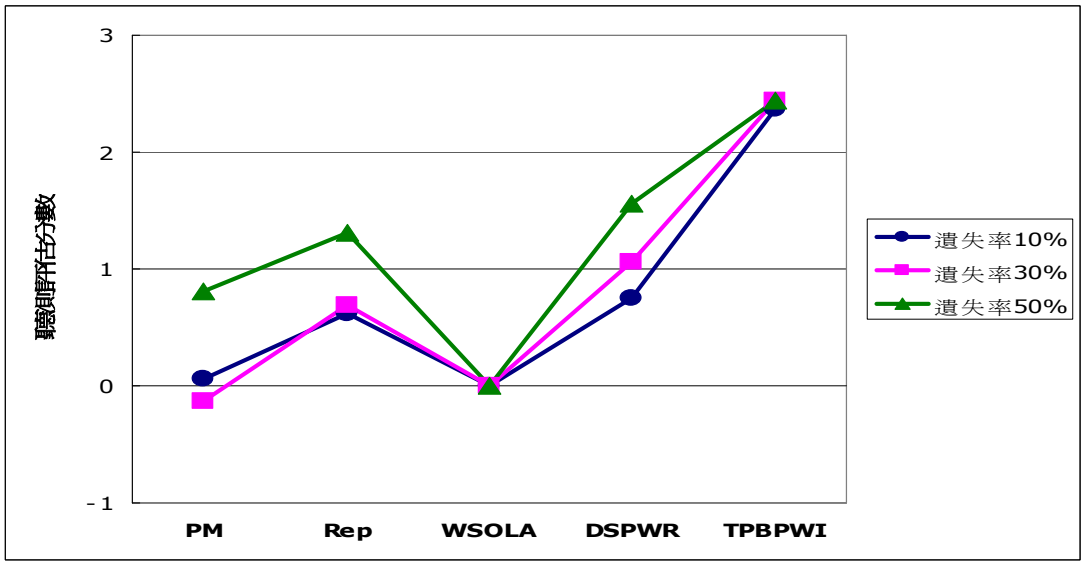


圖 21　聽測實驗之平均評分圖

從這三組數據中我們可以發現，PM 與 WSOLA 的效能比最容易實作的 Rep 略遜一些，原因是它們在搜尋最相似波形的步驟上存在瑕疵，也就是說搜尋範圍

內的波形可能是重建出來的，導致誤差延續，使得它們的補償效果比直接以過去封包重複時還要差。所以 PM 與 WSOLA 兩方法只能夠使用於封包遺失率低於10%的情況。另一方面，隨著封包遺失率的升高，DSPWR 與 TPBPWI 之間的效能差距就越近，顯示在越高的封包遺失率之下，接收端補償方法的效果越難發揮，使得各方法彼此之間的補償效能差異變小。

7. 結論

網路電話的使用愈來愈普遍，但電腦網路常因繁忙或頻寬不夠而造成封包遺失或延遲，使得通話語音的品質衰退、不穩定，因此研究遺失封包之補償方法，來提升通話語音的品質，是有其重要性的。本文研究了一種接收端之遺失封包補償方法，稱爲 TPPWI，它在 SNR 測試和主觀聽測上，都比作比較的 DSPWR、WSOLA、PM、REP 等方法都表現得更好，如在封包遺失率 10%與 30%時，聽測評分會比第二名的 DSPWR 分別好 1.625 分與 1.375 分。

此外，本文研究了一種適用網路電話應用的基週偵測方法，量測正規化的自相關函數，週期長度偵測的正確率還不錯，測試實驗顯示可以達到 96.3%的正確率。在雙邊有聲之 BV 處理，我們提出了不錯的基週波形之同步方法，以及不錯的基週個數與長度的決定方法，再應用先前提出的基週波形內差法，而使得重建出的信號波形具有相位與頻率的連續性，而呈現更高的語音品質。在雙邊無聲之 BU 處理，我們改進了前人的作法，而能夠更有效地避免在無聲信號中，麥克風收錄進來的背景雜訊所引起的噪音。

參考文獻

[1] C. Perkins, O. Hodson and V. Hardman, “A Survey of Packet Loss Recovery Techniques for Streaming Audio”, IEEE Network, Vol. 12 (5), pp. 40-48, 1998.

[2] M. Y. Kim and R. Vafin, “Packet-Loss Recovery Techniques for VoIP”, Technical Report, Royal Institute of Technology (KTH), Sweden.

[3] D. J. Goodman, G. B. Lockhart, O. J. Wasem and W. C. Wong, “Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications”, IEEE trans. Acoustics, Speech, and Signal Processing, Vol. 34, No. 6, pp. 1440-1448, 1986.

[4] O. J. Wasem, D. J. Goodman, C. A. Dvorak and H. G. Page, “The Effect of Waveform Substitution on the Quality of PCM Packet Communications”, IEEE trans. Acoustics, Speech, and Signal Processing, Vol. 36, No. 3, pp. 342-348,

1988.

[5] ITU-T, “A High Quality Low-complexity Algorithm for Packet Loss Concealment with G.711”, Rec. G. 711 Appendix I, 1999.

[6] W. Verhelst and M. Roelands, “An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech”, IEEE ICASSP, Vol. 2, pp. 554-557, 1993.

[7] A. Stenger, K. B. Younes, R. Reng and B. Girod, “A New Error Concealment Technique for Audio Transmission with Packet Loss”, EUSIPCO, 1996.

[8] W.-T. Liao, J.-C. Chen and M.-S. Chen, ”Adaptive Recovery Techniques for Real-Time Audio Streams”, IEEE INFOCOM, Vol. 2, pp. 815-823, 2001.

[9] D. J. Goodman, G. B. Lockhart, O. J. Wasem and W. C. Wong, “Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications”, IEEE trans. Acoustics, Speech, and Signal Processing, Vol. 34(6), pp. 1440-1448, 1986.

[10] R. A. Valenzuela and C. N. Animalu, ”A New Voice-Packet Reconstruction Technique”, IEEE ICASSP, Vol. 2, pp. 1334-1336, 1989.

[11] D. Collins, Carrier Grade Voice Over IP, McGraw-Hill Companies, 2000.

[12] Y. Medan, E. Yair and D. Chazan, “Super Resolution Pitch Determination of Speech Signals”, IEEE trans. Signal Processing, Vol. 39(1), pp. 40-48, Jan. 1991.

[13] Hung-Yan Gu and Wen-Lung Shiu, “A Mandarin-syllable Signal Synthesis Method with Increased Flexibility in Duration, Tone and Timbre Control”, Proceedings of the National Science Council, Republic of China, Part A: Physical Science and Engineering, Vol. 22, No. 3, pp. 385-395, 1998.

基於字詞內容之適應性對話系統

MAGEN: An Adaptive Conversational System based on Terms

朱育德、張嘉惠

國立中央大學資訊工程學系

Email: peterajchu@db.csie.ncu.edu.tw, chia@csie.ncu.edu.tw

摘要

在資訊科技蓬勃發展的今日，資訊化與多元化時代儼然來臨；眾多線上資訊服務的崛起，整合服務與人機互動介面成為矚目焦點。對話系統是發展長久的一項研究，分支眾多，其中一類即是以系統代理為主要目標，是一種目標導向性的對話系統。

本文系統 MAGEN 是一強調適應性的目標導向對話系統，採用以字詞為基礎，以系統訂定的詞類為輔助，捨棄分類器與文法資訊，做為對話之依據，解決以往系統採用分類器、文法資訊所造成擴充性不足，成長性受限的窘境。在此情況下，MAGEN 將適應性充分發揮在三個方面。首先，在對話領域上，由於採用詞庫與詞類的設計，可在僅變動存放在資料庫中的知識庫即達成領域移轉，低門檻、低成本，讓 MAGEN 易於適用各種對話領域。其次，對話過程中，不同的對象會有不同的詞彙用語，透過線上擴增的機制，系統將可學習這些詞彙，下次使用者再度使用這些詞彙時，系統將可有效辨識，達到適應使用者的對話習慣。最後，系統本身核心相當輕量，對話皆以文字方式進行，無須圖形化介面之輔助，因此可輕易移轉到不同平台裝置之上。

為驗證三項適應性，設計有實驗項目，以不同類別之主題、雙回合的方式驗證適應性的情況，並實作三種應用形態的系統，更突顯實際用途上確實存有其經濟價值。

1. 緒論

在資訊科技蓬勃發展的今日，資訊化與多元化時代儼然來臨；在手持式裝置躍上網路通訊平台的那刻起，漸趨輕巧的機身成了大眾追求之目標，顯示空間逐漸被剝奪遂成不可違之宿命，同時，何以將大量資訊與使用者形成良好互動，亦成工程師的一大考驗。在此時空背景之下，語音對話系統，再度成為眾所討論之焦點。

語音對話系統，顧名思義是透過語音進行對話之系統，基本上可以拆解成：語音辨識（Speech recognition），對話系統（Conversational system）以及文字朗讀（Text to speech）等三部份[15]。當然，語音之中夾帶有文字所缺乏的資訊，如：語調、情緒等，這些或可增加對話系統之準確性，但主要之資訊還是來自於文字訊息本身，因此，雖說語音對話系統包含對話系統，但一般來說是將語音對話系統視為是對話系統的一種型態。

一個能以自然語言進行溝通之電腦系統是為對話系統，不論對話的對象是人抑或是系統。對話系統的發展已超過卅年，其間不計其數的成果，但萬流同源，一個對話系統通常都會有三個元件：對話管理員（Dialogue Manager）、知識庫（Knowledge Base）以及自然語言了解（Natural Language Understanding, N.L.U.）[1]，因此常有學者透過這些元件作法之差異，來分類一個系統。

分類對話系統的角度眾多，常見的有：輸入/輸出類型、系統目的、知識庫型態、對話管理員、有無人格特質等，然而，在混合型系統日益增多情況下，分類不再是重點，特質才是受到關注的項目，包括：單一作法是否可同時支援多種語言，能否包容未知詞以及知識庫是否具備成長能力；能力愈強大，即代表此一系統的彈性愈佳，能應付較多情況。

對話系統中，有一類的系統即是以資訊查詢、工作委派等系統代理為目的；其對話目標是為達成一需求，對話過程是以收集資訊為主，而在資訊齊全後，發出任務委派，並等候後端代理人回傳結果。由於其對話是針對特定目的而進行，

所以可稱為目標導向之對話，而以進行此一對話模式所設計之對話系統，可稱為：目標導向對話系統。例如近年來開放式代理人架構（Open Agent Architecture，OAA）[6][16]興起，軟體工程界興起了轉譯與整合這些資訊系統的風潮。由於 Web Service[9]並不涉及使用者介面，僅透過 XML 傳遞執行呼叫與執行結果，也因此，對於使用者而言，即便找到所需之服務，使用上依然存在些困難之處，因此以對話形式做為 Web Service 之人機互動介面，似乎是一個可行的方法。

本篇論文目的在建立一名為 MAGEN 之適應性目標導向對話系統。我們希望 MAGEN 是個高可塑性的系統，在目標導向為前提下，可應用於各種領域，可學習住各種詞彙，可移植到各種平台，深度適應各種情境，而不輕易被時代洪流所淹沒。為此，知識不能以固定的方式儲存，其儲存方式必須能支援系統在運作過程中，從線上所習得的新知識。

以 OAA 架構為例，Web Service 提供的服務眾多而多變，橫跨各種領域，一個對話系統要全然適切其中，除非該對話系統沒有領域限制，亦即具備各領域之詞彙與知識，否則使用者的意圖將無法理解；而若將領域固定，雖是能簡化難度，但同時也大幅降低一個對話系統的存在價值，亦違背我們對適應性的期望。是故，MAGEN 在可支援的對話領域上採取折衷之道：易於增添知識領域，藉此，讓 MAGEN 可以隨需求輕易增加對話領域的知識，而另類的突破領域限制；為此，一個可變動的知識庫更是絕對必要之因素。

自然語言分析器是建構對話系統的根本與基礎，唯有解析來源資訊才能進一步產生回應，而且是即時的回應；過長的回應時間將會讓使用者感到不耐，進而對此系統卻步。因此，選擇適當且適合的自然語言分析技術也是一大重點所在。

2. 相關研究

對話系統的發展已超過卅年[23]，發展出之系統繁雜，焦點與技術各具千秋。概而觀之，可將這些研究分成三種類別：技術導向、擬人導向，以及目標導向等。技術導向的研究並不在提出一個對話系統，而是提出新的技術讓現有之對話系統能夠更強化；從早期的語意分析器（Semantic Parser）如 CKIP system、中

期的機械學習法一直到近來的多重對話序等，這些技術並非對話系統專屬，卻是影響深遠。

擬人導向之系統，是以擬人為目的，著重在知識推論以及人格創造。早期，60 年代，著名的 ELIZA[22]以專家系統為基礎，使用「知識網絡」達成知識推論與對答，深切影響日後人機對談的發展。1991 年，人工智慧領域的 Loebner Prize（www.loebner.net/Prizef/loebner-prize.html）正式開辦；透過杜林測試（Turing Test），尋求最接近真人的系統，十數年來，總有一定數量的系統參加，顯見這部份的發展仍有一定的空間。2005 年的冠軍系統，George（http://chat.jabberwacky.com），由 Rollo Carpenter 所打造，是目前最接近真人的系統；在無法完全了解使用者意圖的情況下，仍能以似是而非的回答與使用者互動，也因此獲得各方好評，不過若是使用者一直無法得到正確的回應，終會失去對話的興趣。

與 ELIZA 同一時期另一個系統則是 PARRY[8]；與 ELIZA 不同，PARRY 捨棄知識了解，僅以關鍵字詞來進行解析，搭配「詞庫集合」，尋找適當字詞，並透過程式組成回應句，講求的是快速反應以及對話進行的流暢度。在擬人導向的發展上，PARRY 的影響力遠不如 ELIZA，卻在事隔廿年之後，深刻影響目標導向類的對話系統。

目標導向的對話是以一個目的為目標，而對話內容即是朝向這個目標而前進，因此常結合多媒體影音、代理人、資訊系統等而成，換言之，對話系統成了一種新的操作介面。既然是以介面為目的，深度知識的推論成了非必要之元素，反而了解使用者之目的與整體的流暢性，才是首重之任務，PARRY 的作法在此處，正式發酵。近年來資訊產業的發展注入新思維，以服務為出發點，以體驗為操作介面，軟體強調親合力與智慧性，加上智慧代理人與手持式系統的風行，引發對話系統大舉朝向目標導向發展。

在分類這些具對話目標的系統上，通常依據「對話控制方式」；J. F. Allen 認為可將系統由簡而難歸納成 Finite-state Script、Frame-based、Sets of Contexts、Plan-based Models 以及 Agent-based Models 等五種類型[2]；McTear 則是進一步

整理統併成三類[17]：有限狀態基礎（Finite State- based），語意框架基礎（Frame-based）以及混合主控（Mixed-initiative）。這些方法並無所謂絕對優劣之分，僅有特質上的差異；依不同的應用情境，選用適當的技術才是重點所在。

- *有限狀態基礎*

顧名思義，這類型的系統背後都存在有一些事先建立之有限狀態圖，以之引導整個對話的進行；使用者依系統發言做適當回答，通常僅是一個短辭或是單字，故有限狀態基礎之系統不需具備強大的自然語言了解能力，系統建置門檻較低，相對的，過於複雜的對話主題，不易繪製有限狀態圖，故不適以之建構。自動銀行系統[17]是此類型中，相當典型的系統，對話流程完全必須依照事先定義好的有限狀態圖進行。在台灣，這類型系統常見於電話語音系統[12][14]，如：醫院的語音掛號系統等，雖使用者僅透過鍵盤傳送訊息，看似與對話有所差異，但就資訊交換的角度來看，確實是一種對話模式。

- *語意框架基礎*

基於語意分析來產生回應句；在最原始的系統中，對話系統僅對句子進行解析，目的在找出主詞，動詞，受詞等部份，然後透過知識庫與文法的輔助來產生回應句；這類型的系統是由使用者來引導對話，彈性遠比有限狀態基礎之系統來得大。飛利浦自動火車時刻資訊系統 [3]是 1995 年由 Aust 等人製作；這是一個以語意框架為基礎的著名系統，用來查詢火車時刻，使用者描述所要搭乘的日期時間、起點、終點，系統則會回傳查詢的結果。

具意圖萃取之智慧型醫療對話查詢系統[4]是台灣成功大學陳銘軍在 2003 年完成的一套整合多項服務的醫療查詢對話系統，整合的服務有掛號諮詢、科別資訊諮詢以及常見問答集諮詢三項大服務；透過分析語意框架內容以及目的偵測來了解使用者目標，然後透過島嶼演算法的方式半自動建立出醫療概念模型。

- *混合主控*

某個角度上來看，混合主控即是有限狀態和語意框架之結合，同時具備引導

對話的資料結構以及語言分析能力，在此種混合式系統下，問與答的界線不再如同前兩種類型般的清晰，且對話的控制權也不再由任一方獨佔，而是共享。

混合主控，此一名詞早在 2000 年以前即已提出[7]；當時的定義其實並不明確，僅言及對話主導權並不完全由單一者控制，而此定義可延伸成對話控制權由多個對話系統共享、由多個對話主題共享、由使用者與系統共享、由多種對話控制方法共享……等等。不論何種共享方式，其對話內容就會如同兩位代理人間的對答一般，因此 McTear 將此類系統改稱為代理人基礎（Agent-based），然而，此名稱卻易與結合真正代理人之對話系統的名稱相混淆，故至今仍有多篇文章採用原始名稱，本文亦然。此類系統由於自由度甚高，對話彈性也相當受到肯定，所以近年來的對話系統大致以此類為主。

ISIS[18]，2004年香港中文大學，Meng等人提出的系統；這是一套線上股市資訊系統，可以查詢股市資料以及代理下單，能接受中、英文，語音上更含括廣東話。作法上採用文法分析，推論目的以及隱含的資訊；功能委派上採用開放式代理人架構，可和其他現成的代理人接軌，達到功能擴充之目的。對話系統雖然可以擴充，但僅限於知識庫，亦即具有未知詞處理和詞彙成長能力，但文法推論規則和對話目的則是固定不變，在對話領域上的移轉具相當的困難度。

基於統計式語意相依關係之對話語句理解系統[24]是另一個成功大學在2004，楊茂柱提出的系統；此系統使用一個新的自然語言理解架構，在此架構中利用語意相依關係幫助語意理解，主要精神是考慮到語句中詞與詞之間可能所隱含之語意資訊而不只是考慮到詞的表徵意義，方法為使用語句結構和語意概念資訊當作語意相依分析之依據。此外還加入對話歷史的觀念，考慮對話語意脈絡，幫助對話理解。在理解過程中使用語意相依圖取代語意框架，避免人為介入定義，提高系統移植性。然而，全系統過度依賴歷史對話，反而使人在領域移轉時顯得心有餘而力不足，僅能拚命收集對話紀錄以餵入分類器中，提高準確性；同時，由於系統相當依賴分類模型，導致對話過程中的知識沒有即時進行回饋，彈性上的缺陷暴露無遺。

智慧型個人助理[19]是一個建構在PDA之上的對話系統，由Nguyen等人在2005年提出；由於是完全建構在PDA之上，系統所受限制較多，比方說，知識庫與對話目標偵測都必須精簡，同時系統彈性也受到相當嚴峻的挑戰，這也是在手持式系統上建構對話系統的最大問題。此系統採用樣板的作法，透過關鍵字詞比對以及較多次問答，取代複雜的邏輯推理以及機率統計運算，達成速度上的優勢以及彌補知識庫精緻化後的缺陷。整體來說，系統的彈性很低，主要受限於知識庫無法成長， 雖有未知詞的處理，但僅限於當下使用，換言之，整個系統完全沒可成長的部份。

2.1 系統比較

表1整理出三種系統和四個評估項目的比較結果。基本上，愈複雜的系統彈性愈好，回應速度愈慢，但請注意，比較是相對性的，實際上，一個對話系統的反應時間都會是在可容忍的範疇之內。對話進行的主導權代表主導對話的走向；在有限狀態基礎和語意框架基礎相當明顯，各為系統與使用者，而在混合起始的系統上則屬於共享。系統必須同時具備有問與答的能力，雖說是共享，卻也不見得是五五平分，通常是系統佔七分，使用者佔三分，換言之，大部分時間仍是由系統在引導整個對話的走向。

由於MAGEN強調適應性，所以這裡不免要歸納整理一下，先前系統在這方面的表現與成果。首先，提到目標導向，直覺必然想到分類器，利用分類器來判斷使用者的目標為何，如同楊的系統[24]一般；然而，分類模型卻會對線上成長造成限制，或許這會是提高辨識率的好主意，但是同時也犧牲了系統的學習性。說到自然語言分析，文法是最常被提及的資訊，ISIS[18]即是立於此點之上，然而產生文法規則並非易事；不論是從歷史對話紀錄萃取文法規則，或是人工自訂，都是偌大工程，尤其是一個嶄新系統要上線時，根本無歷史對話紀錄。在系統平台上的移轉除了要考慮作業系統的差異外，儲存媒體的空間也是一個考驗，例如：要移轉到PDA之上，就不得不考慮知識量與對話流暢性的最大效益比。

表 1：系統比較

	有限狀態基礎	語意框架基礎	混合起始
建構容易度	容易	普通	困難
回應速度	較快	普通	較慢
系統彈性	較低	普通	較高
對話主導權	系統	使用者	共享

3. 系統介紹

本篇系統，MAGEN，是以混合起始為出發點的混合式系統，系統架構如圖1所示，包含對話管理員與知識庫；對話管理員是系統的主體，負責對話的進行，由七個元件組成；知識庫是系統的靈魂，以資料庫為儲存媒體，提供對話管理員了解對話以及產生回應的資料來源。

適應性是MAGEN主要強調的重點，為此，我們選擇以資料庫來儲存知識。由前文的歸納可知，若想讓系統的適應性充分發揮，就必須要：避免使用文法資訊、不要依賴對話紀錄、能避免使用分類器就避免、儘可能讓人能夠直接修正知識內容、有限度的知識成長以及簡化運作的複雜度。扣除這些內容，剩下的就僅有字詞本身。以字詞為基礎，知識成長將變得容易，建構出之腳本將是易於閱讀，當然也就能夠由人所訂定，也因此在移轉對話領域就輕鬆許多。知識以字詞為基礎，此舉引動自然語言解析器的選擇；顯然，基於文法規則的語意分析器（Semantic Parser）不再符合需求，反而是淺層分析器（Shallow Parser）的斷詞系統會是首選。

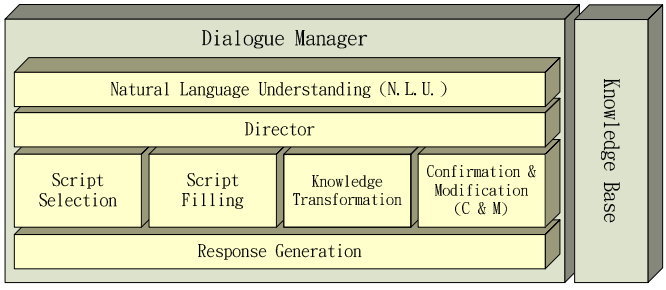


圖 1：MAGEN 系統架構圖

淺層分析器僅對字面做解釋，而忽略背後的文法資訊；相較於語意分析器（Semantic Parser），使用斷詞系統不僅喪失文法相關資訊，並且詞彙散亂，但請注意，這是針對正規文句而言；一般口語並非如同撰文一般注重句法結構，尤以今日網路次文化的盛行，遣詞用語愈趨雜亂無章，在此情況下，額外考慮文法資訊，分析效果不見得會比較好，但較為費時卻是可見的。採用斷詞，捨棄可能錯誤的文法資訊，僅將句子視為字詞堆砌，專注於文字內容之上，反是上上之選。同時，因為斷詞不在乎文法，故當句子中出現多語言夾雜出現，這種令各對話系統棘手的問題亦能一併解決。[11][12]

僅以斷詞來解析自然語言顯然是不足，系統是難以理解世上多變複雜的字詞各代表何種意義，有鑑於此，我們設計了「系統詞類」（TermType）來輔助系統了解各種字詞所代表的概念為何，就如同WordNet裡的Concepts一般，唯詞類是客製化的存在，並非固定不動的各種類別，雖然看似貧乏，卻會是最適切對話領域的存在，讓知識庫不至於過度龐大，除了加快系統反應時間，同時也有效節省知識儲存空間。

添加了系統詞類會不會造成腳本的撰寫與詞彙的學習產生困擾？詞類是詞的概念，比之文法，定義詞的概念應是更加簡單，而有了概念的輔助，相信腳本的撰寫只會更清晰，不會造成太大的困擾才是。因此，建構出來的腳本是相當人性化，不僅易於閱讀，更能輕易異動之；這關鍵的特性，使得領域移轉成了簡單的事情，僅需建構腳本即可，而且MAGEN具有線上成長的特性，無需事先準備過多的字詞知識即可順利運作。

線上成長儼然是目前對話系統備受考驗的地方，而MAGEN也具備此一能力。MAGEN會在對話過程中偵測出新的字詞，並且透過使用者給予定義，更新知識庫，達到學習新字詞的目的；然而，僅如此的話，對話將會相當令人煩悶，尤其在系統尚未習得夠多字詞時。有鑒於此，我們不只設計有學習新字詞的功能，更搭配有推論機制，讓系統先行猜測此未知詞的可能意義，並透過隨後的對話與腳本之記載來確認此推測的正確性；此一作法將可有效降低對話的繁雜度。

最後，為加快系統回應速度，除了實作上的最佳化外，邏輯設計上也有加速的空間。以往的對話系統會經常對於使用者輸入之句子進行目標偵測（Conversational Act Detection）[20]。MAGEN完全不偵測使用者該次輸入目的為何，而採用預設立場的方式，直接認定當次輸入之資訊應當為何；由於，系統與使用者的互動進程皆是透過此一資訊進行，因此，若使用者沒有刻意提供系統錯誤資訊，該預設之狀態不致會錯得離譜。如此作法固然會侷限對話的自由度，無法讓多個主題交互進行，但我們認為「一次一件事」的這種限制，並非太過，屬合理範疇；然此一限制，卻能讓系統執行速度與對話精準度大幅提升，利弊得失，淺顯可見。倒是使用者可以一次將多個資訊告之MAGEN，來省對話回合數。

3.1 知識庫

系統進行對話的依據，以資料庫儲存，可分成用來辨識對話內容的詞庫與引導對話進行的腳本。腳本其實就是對話目的之紀錄，之前的系統需要辨識使用者的需求，而 MAGEN 就是要辨識對話目的，選擇正確的腳本，以引導對話進行，並指示系統該收集何種資訊，相當於有限狀態圖和樣板的角色。

使用資料庫儲存之知識庫具有成長的能力，能夠在系統運作中，透過對答或是操作介面進行成長，達到契合不同使用環境。初始時，這些資料須依靠人工建立少量部份，然而資料量之多寡並不影響系統的能力，僅會使對答流暢性較低，例如：過多未知的字詞就會造成系統不停的進行未知詞概念的確認，但是仍能完成對話目標。知識庫包含詞庫、腳本兩個部份，以下分別介紹如下:

- *詞庫*

詞庫預設有十六個系統詞類(Termtype)，其名稱與範例如表 2 所示；每個詞都會對應到一個或多個類別，例如：再見<結束>，或中央大學<地點/組織>。若詞彙難以歸類，則歸入<其他>，這些詞彙通常表示比較不重要的詞彙，或是比較罕見的專有名詞。<事件>是表示具備動作性質的詞彙，通常也是腳本的觸發關鍵詞。<忽略>則表示該詞屬於可以忽略不計，只是發語詞或是贅詞。除了出

現在詞庫中的十六個詞類外，系統運作過程中尚有一類，<未知>，此類別用來標示未出現在詞庫中的字彙，亦即未知詞；此類詞彙在系統運作過程中，若被確認類別，該詞將會加入詞庫之中，日後再度出現時，就不再是未知詞。

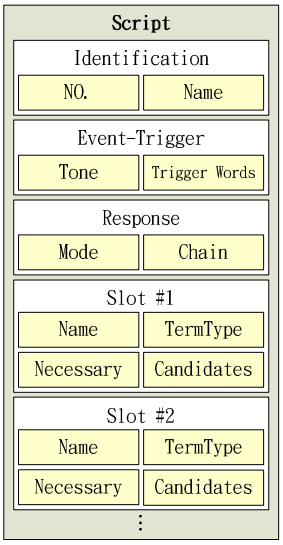
而詞庫裡的字詞不該單從字詞去觀看，應該將之視為「具有詞類的字詞」，例如：我們應該將中央大學<組織>與中央大學<地點>視為兩個不同的「具詞類的字詞」。在這個觀念下，當新增了一個類別<學校>時，應該將中央大學<學校>視為是一個新字詞，而不應認為中央大學<組織/地點>是個錯誤的字詞。

表 2：詞庫類別與範例

問候：唷，早	肯定：是，OK	事件：吃飯，查	數值：19，21.3
情緒：^^，呵呵	否定：不是，否	物品：車，電腦	時間：日，：
結束：Bye，再見	疑問：嗎，？	地點：中央大學	組織：中央大學
忽略：和	驚嘆：吧，喔	人物：趙建銘	其他：反托拉斯

- *腳本*

腳本是對話進行的重要依據，當中紀錄所代表的對話情境，以及所需資訊等，換言之，要進行怎樣的對話，僅需製造出相對應的腳本即可。本文以『』代表腳本，例如：『邀約』代表邀約此一腳本。腳本格式如圖 2。



欄位名稱	說明	值域
Identification	腳本識別資料	
NO.	唯一的編號	01 \ 02 \ 03 …
Name	腳本主題名稱	邀約 \ 訂票 \ …
Event-Trigger	觸發腳本之條件	
Tone	語氣	肯定 \ 疑問 \ 不拘
Trigger Words	觸發關鍵詞	參加 \ 訂購 \ …
Response	填充完成後之指示	
Mode	執行模式	立即回應 \ 不予回應
Chain	連鎖觸發	01 \ 02 \ 03 …
Slot	腳本所需之資訊	
Name	插槽名稱	出發地 \ 日期 \ …
TermType	值型態	地點 \ 日期 \ …
Necessary	必須與否	是 \ 否
Candidates	候選值	華航 \ 長榮 …

圖 2：腳本格式（左）與腳本內容說明（右）

3.2 對話管理員

對話管理員負責邏輯處理，由七個元件所組成，包括 N.L.U、Director、Response Generator 以及四個狀態的處理程式：Script selection、Script Filling、Knowledge Transformation、Completion 等，其運作流程如圖 3 所示。當對話管理員在收到訊息後，首先進行自然語言了解，而後根據 Status 裡頭所描述的狀態來決定走哪個路徑來處理當下的對話狀態，不同的路徑由不同的元件負責，最後由對話產生器依據對話狀態產生回應句。針對自然語言了解部份，分為中文斷詞，字詞類別標示，字詞群組化，未知詞標示等四項工作。

1. 斷詞：根據[5]所簡化之斷詞系統。
2. 字詞類別標示：將斷詞結果，依「詞庫」，內容標示對應詞類；例如，「中央大學」<組織/地點>，表示此字詞可為組織或地點兩種系統詞類，對於可能造成的模糊則以大膽假設及回問使用者來處理歧義。當有詞彙沒有出現在詞庫中時，此一詞彙稱之為未知詞，此時標示成<未知>。

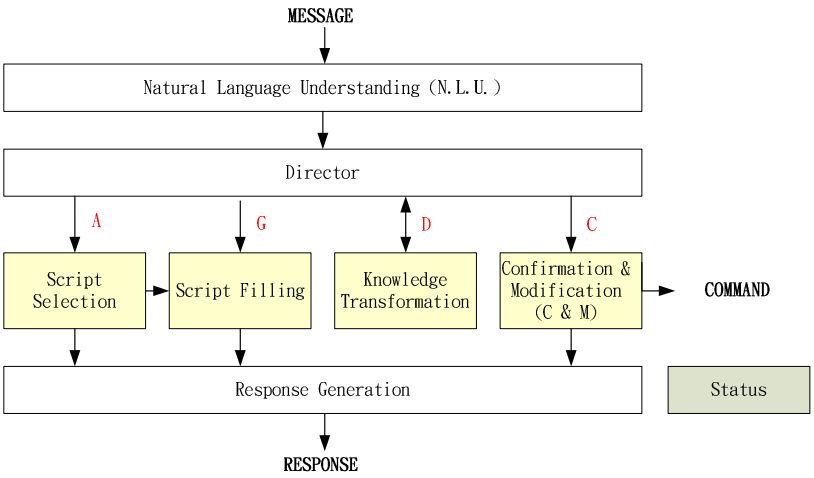


圖 3：對話管理員運作示意圖

3. 字詞群組化：經過斷詞後，句子成了一堆字詞的堆砌，但別忘了，這些字詞原本存在句子中的順序是有其意涵的，相鄰的字詞存在有相當的關聯性，因

此若將這些關係群組化起來，將可提升系統了解所接到之資訊。群組化分成同質群組與異質群組兩部份先後且循環執行，直到沒有任何相鄰字詞可以進行群組化為止。同質群組是將具有同樣詞類的相鄰字詞結合，異質群組則是透過特殊規則將相鄰字詞合併；在尋找可結合之相鄰字詞時，<忽略>一類之詞會被忽視，但仍會保留在群組化的結果之中。舉例來說，「讓我取消訂票行不行？」，經過斷詞與詞類標示之後為：讓<事件/其他>我<人物>取消<事件>訂票<事件>行<肯定>不行<否定>？<疑問>；「取消」與「訂票」這兩個詞首先會在首輪的同質群組時被組合，隨後，「能」與「不能」也在同輪的異質群組被合併，形成：讓<事件/其他>我<人物>取消訂票<事件>行不行<疑問>？<疑問>。因為還有可以群組化的字詞，所以進行第二輪群組化，形成：讓<事件/其他>我<人物>取消訂票<事件>行不行？<疑問>。結束後，已無任何相鄰字詞可以群組化，因此，群組化在第二輪後結束，並輸出結果。

4. 未知詞標示：當有詞彙未出現在詞庫中時，該詞所標示之詞類為<未知>；在群組化的過程中，所有相鄰的<未知>詞亦會被群組化成單詞，這些單詞在每輪群組化結束時，會反覆查詢這些組合後的詞彙是否出現在詞庫之中，若有，則變更詞類。

3.3 對話狀態（Status）

對話管理員以對話狀態記錄目前情況，對各元件而言，狀態就如同是任務指示，標示著目前該執行的任務。對話狀態總共有 19 種(如表 3 所示)，可依其類型分成四類：《Assignment》、《Definition》、《General》、《Complete》。圖 3 中，Director 元件周圍的英文字母，A、D、G、C，即表示各 State 之縮寫。運作時，Director 會根據目前 Status 最上層狀態（Top State）決定將控制權移交給哪個元件，例如：A 就交給 Script Selection，G 就交給 Script Filling。

N.L.U.與狀態並無任何關聯，Director 以及 Response Generation 這兩個元件對於 States 是以偷窺（Peek）的方式，而非使用彈出（Pop），亦不會對狀態產生任何變化，其餘四個元件，則視情況異動狀態。

各種狀態轉換條件如圖 4；此圖為簡化資訊，僅針對四大狀態，不計其參數變化。雖系統運作可用有限狀態圖表示，但不代表 MAGEN 應歸類為有限狀態基礎之對話系統；相關研究中所提有限狀態基礎系統，其有限狀態圖乃應用於腳本，對話僅依該圖進行，不同之對話有不同之狀態圖，MAGEN 將有限狀態圖做為運作基礎，此一狀態圖乃固定不變之存在，是假定使用者所給予之訊息應代表何種涵義，用以引導系統運作，雖同屬有限狀態圖，用途、目標卻大不相同。

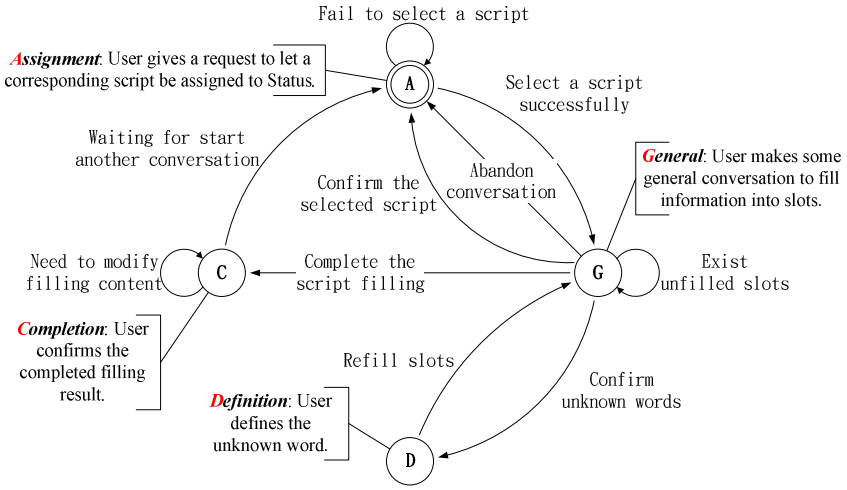


圖 4：狀態變化移轉圖

回應產生(Response Generation)會根據當時的 State 尋找對應之句法，結合狀態資訊，產生回應句。不同之 State 定義有不同的句法格式，如表 3 所示；各變數意義如下：{Script}表目前選定腳本名稱，{ScriptList}表多個腳本的名稱，{Slot}表狀態格式指定之插槽，{Slot.TermType}與{Slot.Name}各表插槽中的詞類限制與識別名稱，{Content}則是整個填充內容所改寫之確認句，而{Trigger word}表示狀態中被保留的字眼。另有一個函數：ASK ( )，此一函數是透過{Slot}的識別名稱、限制詞類以及候選值等資訊改寫成詢問句，如一個型態值為地點的插槽，產生的回應為：集合地點在哪裡？若是此插槽具有候選值，則以插槽名稱配合候選值產生回應句如：火車種類是自強，莒光還是復興？

《5》與《6》僅對用以內部狀態之識別，對使用者而言並無分別，故合用同一回應句法。《12》,《14》與《15》亦屬內部狀態識別，與使用者實際互動之對話狀態為《18》,《19》以及《9》。

表 3：狀態格式暨回應句法明細

編號	類型	情境描述	回應句法
1	A	初始對話	您好，我是 MAGEN。請問您有什麼需求？
2		詢問是否還要繼續對話	還有事嗎？沒事就先這樣囉！
3		結束對話	感謝您的使用，祝您今天愉快。
4		放棄對話	您的問題超出我的理解範圍，或許您該嘗試其他服務。
5		選取失敗	我的回答很有限，你必須問正確的問題才行。
6		選取失敗但保有關鍵詞	
7		多重腳本中選	我不太清楚，您是要進行 {ScriptList} 哪個話題？
8		詢問增添腳本關鍵詞	確認一下，「{Trigger-word}」與「{Script}」有關係嗎？
9		詢問對話目標是否正確	等一下！您確定我們進行的對話目的是：{Script} 嗎？
10		確認是否連鎖觸發主題	嗯，那您接下來要進行 {Script} 嗎？
11	G	提示性填充	ASK（{Slot}）
12		針對未知詞填充	
13		二次提示性填充	請您認真一點！ASK（{Slot}）
14		確認性填充	
15		指示對話目標確認	
16	C	首次確認填充內容	OK！跟您確認一下剛剛的資訊{Content}
17		確認修正後之填充內容	修正後的資訊是： {Content} 請確認。
18	D	確認未知詞之類別	請問「{Term}」是指「{Slot.TermType}」嗎？
19		確認填充內容之類別	確定 {Slot.Name} 是 {Term} 嗎？

3.4 系統運作

MAGEN 的運作與狀態移轉息息相關，因此本文將由圖 4 進行說明。MAGEN 是目標導向之對話系統，因此首先必須偵測使用者之目的，狀態為《Assignment》。目標偵測時，有多種例外狀態，例如：多腳本同時吻合、具有新觸發字須確認添加，這些都會在此狀態中完成；在目標確定之後，狀態移轉至《General》進行腳本填充。

依據所擇定的腳本插槽進行詢問，直到完成後，交由使用者確認；同樣，填充過程中可能會有多種例外狀態，例如：辨識失敗、未知詞干擾等，這些都需做

進一步確認。當然，系統在第一時間中，並不會立即發出確認句，而是採用假設與推論的方式猜測未知之資訊，例如：當系統詢問地點時，使用者回應一未知詞，此時，系統將大膽假設該未知詞意指地點；當然，在無法猜測之時，還是需要透過《Definition》進行未知詞的確認。填充時，未免因腳本設計有誤或其他因素導致對話卡死，單一插槽最多僅嘗試填充五次，超過，將會中止該段對話，當然，使用者也可以隨時提出中止對話之要求，在確認之後即會中止。終止之後，狀態回到《Assignment》，並等候下一個對話主題的開始。

填充過程中，包含有系統的判讀與猜測，因此不可能全然無誤，因此在完成填充之後，狀態移轉為《Completion》，將填充內容改寫成確認句，交由使用者進行確認或是修正。由於填充過程採用詞類為填充之依據，若採用相同策略進行修正，恐難有良好效能，於是我們改採以內容為修正策略。以<肯定>{內容}<否定>{內容}的句法為判定，例如：是今天晚上，不是明天晚上，此時使用者表示已經填充的明天晚上有誤，應以今天晚上取代。

待使用者完成修正，確定填充內容之後，檢視該腳本是否設定有連鎖觸發腳本資訊，若有進行確認，是否要直接進行該話題，若無則回到初始狀態；不論何者，系統將是把狀態移轉至《Assignment》等候下一段對話的開始。

3.5 對話範例

MAGEN 是以 Microsoft® .net framework© 平台實做，可移植至多種裝置之上；表 4 為一完整對話範例，並附加重點說明，加深運作流程與實際對話之聯結。除了將對話核心實作完成之外，更實際完成應用形態。借由開放式元件 DotMSN 的協助，成功與 MSN Messenger 進行溝通。搭配 IBM ViaVoice 與 Microsoft Speech 讓 MAGEN 擁有語音輸入輸出的效果。當然，以 MAGEN 的特性，移植到 Pocket PC 之上更是絕對沒有問題。為了更清楚展示 MAGEN，我們實錄兩段展示影片，可於 http://140.115.155.87/MAGEN/DemoVedio.htm 下載。

表 4：對話範例

User	唷！要不要去吃個飯？
MAGEN	地點在哪裡？
	【說明】「去吃個飯」一詞觸發『邀約』，緊接著首次嘗試填充腳本；成功填充事件插槽為「去吃個飯」。
User	魔法咖哩，西門店
MAGEN	參予的有誰？
	【說明】成功填充地點插槽，接著詢問參予人士。
User	我
MAGEN	時間是什麼時候？
	【說明】代換指稱詞，「我」將被換成「你」再進行腳本填充。
User	週末晚上，如何？
MAGEN	OK！跟您確認一下剛剛的資訊：週末晚上，在魔法咖哩西門店，和你去吃飯。這樣對嗎？
	【說明】『邀約』填寫完成，改寫填充資訊為確認句，並由使用者進行確認與修正的步驟。
User	對！
MAGEN	我會盡快給您答覆。還有事嗎？沒事就先這樣囉！
	【說明】使用者肯定填充內容，加上無連鎖觸發之腳本，故以《2》結束對話。

4. 實驗與分析

評估一個人機互動的對話系統，最直接的方法即是以人來進行評估，然而僅由數人評論，恐有失客觀。未免流於主觀，過程中除了人為觀感外，另有「成功率」、「主題擇定回合數」、「對話回合數」的統計輔助，一來驗證這些實驗對象所言，也提供不同角度之觀點。

「成功率」計算方式為：正確內容之插槽數 / 對話腳本之插槽數，其值愈高愈好。以有四個插槽的『邀約』為例，在修正到最後送出後，仍只有三個插槽填充內容正確，則該對話的成功率為 0.75。「主題擇定回合數」與「對話回合數」皆以「回合數」為計量單位；所謂「回合」，使用者送出一訊息且系統對之產生回應合為一回合。前者表示系統辨識使用者目標所需花費之成本，當然愈低愈好；後者表示對話之經歷，從問候語開始到送出外部命令為止，此數值會與插槽數有高度相關性，但基本而言，數值愈低表示愈快達成目的。

實驗採用透過 MSN Messenger 的方式進行，並事後透過網路問卷回收其滿意度。我們從 MSN 聯絡人中挑選出八位從未接觸過 MAGEN 之對象進行「邀約吃飯」、「邀約看電影」、「留言」、「問題詢問」以及「查詢網站資料」等五個主題之對話。

實驗過程中，八位對象各自獨立進行。知識庫中存有四張腳本以及 1327 個具詞類的字詞（同一字詞，不同類別視為不同之具類別字詞）；這些詞彙是透過人工定義的方式與先建立各種詞類常見之字詞，但刻意忽略與電影相關之字詞。

五個對話主題中，「邀約吃飯」與「邀約看電影」共用『邀約』，進行不同之邀約行為，是故腳本僅有四張；旨在探究能否讓多個近似之主題共享單一腳本，且由於詞庫缺乏電影知識，此情況下，對話是否能夠進行。由表 5 可發現，「邀約看電影」確實是五個主題中，數值表現最差之主題，然整體成功率卻仍有 0.76，顯見此等模糊之主題，系統依舊能夠與使用者進行對話。

「問題詢問」的主旨在查詢系統內部所擁有之知識，而「查詢網站資料」則是針對外部如搜尋網站進行遠端資訊查詢，目的近似，唯所需資訊不同，後者須知搜尋網址，插槽數比前者多一。此二者不論在目標或是所需資訊皆極為相似。由於 MAGEN 棄分類器而以觸發關鍵字與語氣詞偵測使用者之對話主旨，因此在辨識這類主題時，的確會產生困擾並且付出較多之回合數為代價(表 5)。

人機互動，人之情感亦不可少，因此，增加「滿意分數」之評量標準。其值由一至十，以十為佳；表 5 中，此數值與成功率有正向之關係，可見使用者之滿意程度多半與對話之完成性相關。

MAGEN 是個強調高度適應性之系統，因此，實驗的另一個重點即在於學習能力。為此，各對象再次對五個進行另一個循環之對話，此為循環二，其值如表 6。由平均值來看，第二循環之數值皆比第一循環來得好，表示系統與使用者間確實達到某種程度之契合。系統之所以能夠適應使用者之對話，全然是因為字詞的成長所致，而 MAGEN 的字詞成長有兩部份，一者，詞庫，另者，觸發字。在循環一中，總共成功添加 16 個新字詞，以及 6 個觸發字；循環二中，依舊有

16 個新字詞成功添加，而僅有 1 個新的觸發字。

表 5：循環一評估結果

評量 / 主題	滿意分數	成功率	主題擇定回合數	對話回合數
邀約吃飯	8.50	0.95	1.38	10.5
邀約看電影	5.63	0.76	1.63	9.50
留言	9.25	1.00	1.88	4.13
問題詢問	9.13	0.88	2.25	5.13
查詢網站資料	7.25	0.94	4.50	9.25
平均值	7.95	0.91	2.33	7.70

表 6：循環二評估結果

評量 / 主題	滿意分數	成功率	主題擇定回合數	對話回合數
邀約吃飯	9.13	0.98	1.00	6.63
邀約看電影	7.83	0.93	1.75	9.75
留言	9.38	1.00	1.13	3.88
問題詢問	9.38	1.00	1.75	3.75
查詢網站資料	9.00	1.00	1.38	4.63
平均值	8.85	0.98	1.40	5.73

近似的「問題詢問」與「查詢網站資料」兩主題在循環二中的表現明顯優於循環一，尤其是「查詢網站資料」在主題擇定回合數與對話回合數上大幅減少。這是系統成長後的結果嗎？「問題詢問」在循環一中，並沒有增添任何新的觸發字，然而在循環二中，主題擇定回合數卻以 1.75，低於循環一的 2.25，表示系統花費更少的回合數正確辨識出使用者之目的，顯然是使用者的適應性，是使用者有經驗之後，能夠以更快速的方式與系統對話，相同現象也發生於「留言」。「查詢網站資料」的成長性在觸發詞的成長、未知詞數量減少以及使用者適應等三個現象同時發生情形下，是為五個主題中最顯著之主題。

「邀約看電影」不論是「主題擇定回合數」或是「對話回合數」都出現異於其他四個主題之「負成長」現象。究其緣由，一般邀約看電影時，所用之詞通常不會有「電影」一詞，較常直接講述電影片名；此現象讓選擇腳本時產生困擾，

自是少不了需多耗幾個回合。再者，電影相關的字眼多變，詞庫中又相當缺乏，循環二的未知詞數量不僅沒有低於循環一的 6 個字，更增加到 7 個字；額外的未知詞確認，自是讓對話回合數向上攀升。反觀「成功率」與「滿意分數」，在循環二比循環一高的情況而言，字詞確認並不影響使用者對於對話的感覺，同時成功率攀升至 0.93，顯見在循環一的成長後，對話變得更加容易。

5. 結論與未來展望

本篇論文所建構之系統，MAGEN，是一個有目標之對話系統，跟以往的系統不同，以屬於淺層分析器的斷詞系統為出發點，配合上群組化技術、詞類標示等技術，在對話是不合文法、中英字句夾雜的情況下，自然語言辨識能力略勝語意分析器一籌。由於將對話資訊由文法句型回歸到文字內容本身，因此在知識編輯上更加自由、簡單，這使得對話領域的移轉、線上知識成長變得更加簡單，在應用情境與對話風格的適應性上都有傑出表現。

線上知識成長通常需要使用者的一些確認來配合，這很容易讓系統變得很囉唆，可是不足的確認又難以解決模糊不解之處； MAGEN 將所有確認機制完整建立，但搭配推論與猜測的策略，讓某些確認機制備而不用，既可保證不會有模糊地帶且不會有過於囉唆之對話發生。

MAGEN，強調高度適應性，在三個層次中充分展現。首先，以斷詞系統為出發點之下，移轉對話領域變得容易，使得系統得以適應在各個對話領域之中；其次，線上成長搭配詞類推論，系統得以在對話過程中，漸漸適應使用者的語言風格，進而成長；最後，以 Microsoft® .net Framework©實做，搭配三種版本的對話核心，不論在手持式環境，在語音環境，亦或是網路通訊環境，MAGEN 都可以適應其中，達到最大經濟效益。

誌謝

本研究由國科會編號 NSC 94－2524－S－008－002 贊助。

參考文獻

1. J. Allen, “Natural Language Understanding,” *The Benjamin/Cummings Publishing Company*, 1995.
2. J. F. Allen, D. K. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, “Towards Conversational Human-Computer Interaction,” *AI Magazine*, 2001.
3. H. Aust, M. Oerder, F. Seide, and V. Steinbiss, “The Philips automatic train timetable information system,” *Speech Communication*, vol. 17, pp. 249-262, 1995.
4. M. J. Chen, “Intention Extraction for Intelligent Medical Query System,” *National Cheng Kung University, Master dissertation*, Jun. 2003.
5. K. J. Chen, S. H. Liu, “Word Identification for Mandarin Chinese Sentences,” *COLING*, pp. 101-107, 1992.
6. A. E. Cheyer and D.E. Martin, “The Open Agent Architecture,” *Autonomous Agents and Multi-Agent Systems*, 2001.
7. J. Chu-Carroll, “MIMIC: An Adaptive Mixed Initiative Spoken Dialogue System for Information Queries,” *The Sixth Conference on Applied Natural Language*, pp. 97-104, 2000.
8. K. M. Colby, “Artificial Paranoia,” *Artificial Intelligence*, vol. 2, 1971.
9. K. D. Gottschalk, S. Graham, H. Kreger, and J. Snell, “Introduction to Web services architecture,” *New Developments in Web Services and E-commerce*, vol. 41, Nov. 2001.
10. R. Higashinaka, N. Miyazaki, M. Nakano, K. Aikawa, “Evaluating discourse understanding in spoken dialogue systems,” *ACM Transactions on Speech and Language Processing*, vol. 1, pp. 1-24, 2004.

11. R. Kaplan, S. Riezler, T. King, J. Maxwell, A. Vasserman, and R. Crouch, "Speed and accuracy in shallow and deep stochastic parsing," *HLT-NAACL*, 2004.

12. C. J. Lee, E. F. Huang, and J. K. Chen, "A Multi-keyword Spotter for the Application of the TL Phone Directory Assistant Service, " *Workshop on Distributed System Technologies & Applications*, pp. 197-202, 1997.

13. X. Li and D. Roth, "Exploring evidence for shallow parsing," *The Annual Conference on Computational Natural Language Learning*, 2001.

14. Y. C. Lin, T. H. Chiang, H. M. Wang, C. Peng, and C. Chang, "The Design of Mandarin Chinese Spoken Dialogue System," *International Conference on Spoken Language*, vol. 1, pp. 230–233, 1998.

15. M. Lundeberg, J. Gustafson, and N. Lindberg, "The august spoken dialogue system," *Eurospeech*, 1999.

16. D. Martin, A. Cheyer, and D. Moran, "The Open Agent Architecture: a framework for building distributed software systems," *Applied Artificial Intelligence*, vol. 13, pp. 91--128, 1999.

17. M. F. McTear, "Spoken Dialog Technology: Enableing the Conversational User Interface," *ACM Computing Surveys*, vol. 34, pp. 90-169, Mar. 2002.

18. H. Meng, P. C. Ching, S. F. Chan, Y. F. Wong, and C. C. Chan, "ISIS: An Adaptive, Trilingual Conversational System With Interleaving Interaction and Delegation Dialoogs," *ACM Transations on Computer-Human Interaction*, vol. 11, pp. 268-299, Sep. 2004.

19. A. Nguyen, and W. Wobcke, "An Agent-Based Approach to Dialogue Management in Personal Assistants," *International conference on intelligent user interfaces*, pp. 137-144, Jan. 2005.

20. A. Stolcke, E. Shriberg, R. Bates, N. Coccaro, D. Jurafsky, R. Martin, M.

Meteer, K. Ries, P. Taylor, and C. V Ess-Dykema, “Dialog act modeling for conversational speech,” *AAAI Spring Symposium on Applying Machine Learning*, pp. 98-105, 1998.

21. E. Voorhees, “The TREC-8 Question Answering Track Report,” *Eighth Text Retrieval Conference*, pp. 77-82, 1999.

22. J. Weizenbaum, “ELIZA—A Computer Program for the Study of Natural Language Communication between Man and machine,” *CACM*, vol. 10, 1967.

23. Y. Wilks, “Human-Computer Conversation,” *International Workshop on Human-Computer Conversation*, vol. 1, pp. 1-14, Jun. 1999.

24. M. Z. Yang, “Semantic Dependency Based Natural Language Understanding in a Medical Dialogue System,” *National Cheng Kung University, Master dissertation*, Jun. 2004.

一種適用於大量連續語料的語音文句校準方法

簡世杰　　張信常

工業技術研究院 資訊與通訊工業研究所
新竹縣竹東鎮中興路四段 195 號 51 館
{ShihChiehChien and Piosn}@itri.org.tw

摘要

為了使維特比演算法（Viterbi Algorithm）能適用於大量連續語料的語音文句校準，以部分語音文句校準循序進行處裡，是一種較有效率的作法，但如何確保整體搜尋空間的最佳路徑落在部份語音和部分文句所形成的部份搜尋空間集合，以及，如何決定落在部份搜尋空間裡的部分最佳路徑，並且該部分最佳路徑與整體搜尋空間的最佳路徑是重疊的，是實施的關鍵。因此，本文提出一種可靠路徑估測方法估測存在於部分搜尋空間裡的可靠路徑，並藉由可靠路徑估測結果調整部分搜尋空間，以防止最佳路徑可能超出部分搜尋空間的情況。實驗顯示本文方法不但可適用在一般無背景噪音的大量連續語料校準，在高 SNR 背景音樂的情況下也能獲得不錯的結果。

1. 前言

在語音信號處理裡，語音文句校準是常見的前處理工作，其目的在取得語音信號與文句內容之間的對應關係，以進行像是語音辨識的聲學模型訓練或是作為語音合成的合成單元使用。一般而言，這類應用所使用的語料通常都是事先依照需要設計的，並且也常以人工方式進行預處理，以使這些經過設計處理過的語料容易以傳統的維特比演算法（Viterbi Algorithm）進行語音文句校準。不過，對於常見的教學錄音帶或是光碟音軌，這些動輒 5 分鐘以上的連續語料以傳統的維特比演算法來進行語音文句校準，記憶空間和運算時間的耗費是相當大的，並且，當連續語料超過一定長度時，傳統的維特比演算法也就不見得能夠適用了。因此，過去對於這種大量連續語料的處理，通常我們會先採用人工分段，再使用傳統的維特比方法進行細部的校準，但這樣也僅能適用在資料量不大的時候，當資料量大時，譬如要對過去傳統的音訊素材全面的進行數位化和再利用，這時候提供一種適用於大量連續語料的語音文句校準方法，取代人工作業，就是一件相當重要的工作了。

對於大量連續語料的語音文句校準，過去的文獻是設法於連續語料裡取得可信賴的錨點（anchor）以分割語料，將大量連續語料分割成較小的語音片段，並再次取得存在於語音片段裡的錨點，直到這些語音片段得以使用傳統方法進行處理為止[2,3]。其中，幾個重要的模組是這種錨點偵測(Anchor Detection)做法所必備的，包括一個語音辨識器以辨識出可能的文句、一個動態規劃（Dynamic Programming）模組比對識別文句與原始文句以取得一致性的文句、以及一個錨點偵測模組配合一些準則自一致的文句內容裡選出錨點。其中，語音識別器的識別能力和錨點的選擇是影響錨點偵測效果的關鍵所在。對於增強語音識別能力，事前可使用一個文句剖析器

依據給定的文句設定識別器使用的識別詞彙和訓練語言模型，以縮小識別範圍和限定前後文接續關係來提昇識別效果。為使錨點選擇具有可靠性，識別文句與原始文句匹配長度達到一定門檻值的錨點選擇準則是常見的作法。然而，當錨點與錨點間的的語音長度小於文句預估的長度；譬如，語音的音框數小於文句的狀態數，即無法順利完成這些錨點之間的語音文句校準。再者，當重複文句出現，識別文句與原始文句匹配就很容易出現問題，這種情形又特別容易出現在語言教學類型的語料裡，也是以這種錨點偵測方式不易克服的地方。另外一個問題是，不同音訊素材的背景環境或收錄所使用的設備可能是不相同的，在這樣的狀況下，相當於是要以固定訓練環境的聲學模型對不同環境的語料進行語音識別，搭配模型調適或者強健式語音識別技術就是錨點偵測做法所要考慮的，其複雜度和難度可見一斑。

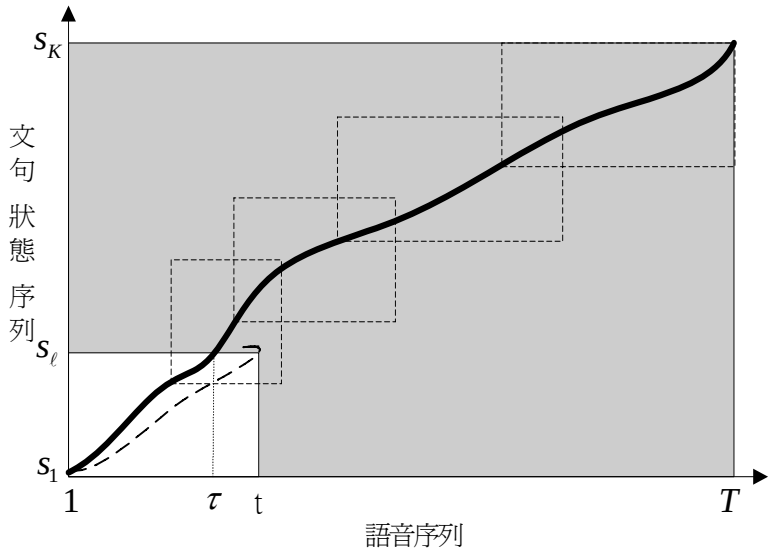
對於已知文句內容的情況下，採用傳統的維特比演算法進行語音文句校準，即便輸入的語音環境可能不同於當時聲學模型的訓練環境，其校準結果通常都仍能具有相當高的準確性。然而，如前所述，傳統的維特比演算法對於大量連續語料的校準有運算速度和記憶體運作上的問題；必須等到所有可能路徑決定以後才能取得存在於整體搜尋空間裡的最佳路徑，難以循序的以部份搜尋結果來進行處理，最大的癥結就在無法確定部分搜尋空間的部份最佳路徑與整體搜尋空間最佳路徑的一致性。因此，本文提出一種可靠路徑估測方法，以找出可能落於部分搜尋空間裡的部分最佳路徑，以部份語音文句校準循序的定出落於部分搜尋空間裡的部分最佳路徑，以使維特比演算法仍能使用於大量連續語料的語音文句校準。實驗顯示該可靠路徑估測方法配合部分語音文句校準，不但可適用在一般無背景噪音的大量連續語料的校準，在高 SNR 背景音樂的情況下也能獲得不錯的結果。

2. 部分語音文句校準問題

對於以部份語音和部分文句對大量連續語料進行語音文句校準，兩個主要問題必須解決：(一) 如何確保整體搜尋空間的最佳路徑落在部份語音和部分文句所形成的部份搜尋空間集合裡；(二) 如何決定落在部份搜尋空間裡的部分最佳路徑，並且該部分最佳路徑與整體搜尋空間的最佳路徑是重疊的。

圖一是一個以部份語音和部分文句進行校準的示意圖，其中灰色區塊為整體搜尋空間，白色區塊表示部分搜尋空間，黑色粗實線表示整體搜尋空間的最佳路徑。在圖一中，s_K 為整體文句狀態序列的最後一個狀態，s_1 為整體文句狀態序列的第一個狀態，s_ℓ 為選定部分文句狀態序列的最後一個狀態，T 為整體語音序列的最後一個音框，t 為選定部分語音序列的最後一個音框，而 τ 則為整體搜尋空間的最佳路徑與選定部分文句狀態序列的最後一個狀態 s_ℓ 所交會對應的語音音框位置。由圖一可以發現，如果自部分搜尋空間的終點 t 才進行回溯取得其最佳路徑，結果將會與整體搜尋結果有相當大的出入（圖一白色區塊的虛線曲線），這也說明：(一) 對於部份語音區間 $[1, t]$ 和部分文句區間 $[s_1, s_\ell]$ 所形成的部分搜尋空間 $\Gamma(t,\ell)$ 並沒有將落在語音區間 $[1, t]$ 的整體搜尋空間的部份最佳路徑完全涵蓋，也就是部分文句區間 $[s_1, s_\ell]$ 是不足的；(二) 對部分文句區間 $[s_1, s_\ell]$ 而言，所使用的部分語音區間 $[1, t]$ 是過長的，適當的語音長度應小於 τ。圖一顯示，最佳路徑在語音序列 τ 之後已超出預設的部分文句區間 $[s_1, s_\ell]$。因此，除非在搜尋的同時可隨時掌握該最佳路徑的語音及文句的適當範圍，否則，也只能不斷擴大部份搜尋空間以保證整體搜尋空間的部份最佳路徑落在該部份搜尋空間裡，最極致的情況就是部份搜尋空間

等於整體搜尋空間，而這也是我們所不樂見的情況。顯然的，如何決定存在於部分搜尋空間裡的最佳路徑，並且使得該部分最佳路徑與整體搜尋空間的最佳路徑是重疊的，是解決上述問題的關鍵。因此，下一節，我們將介紹一個可靠路徑估測演算法來解決上述問題。



圖一：部分語音文句校準示意圖

3. 可靠路徑估測

可靠路徑估測是以維特比演算法為基礎的作法，並且配合最大相似度來施行，因此，以下我們先對基本的維特比演算法和其特性做一簡單描述。

3.1. 維特比演算法

假設由語音序列 $X_T=(x_1x_2..x_T)$ 和文句狀態序列 $S_K=(s_1s_2..s_K)$ 所構成的搜尋空間 $\Gamma(T,K)$，存在一個與 X_T 相對應的最佳狀態序列 $Q_T=(q_1q_2...q_T)$，以最大相似度(ML，Maximum Likelihood)來進行比對時，我們可以在時間t得到對應到文句狀態 s_i 的最佳狀態序列為 $Q_t=(q_1q_2...q_t)$，若將此時的相似度分數定義為

$$\delta_t(s_i)=\max_{q_1,q_2,...,q_{t-1}}\Pr[q_1q_2,...q_{t-1},q_t=s_i,X_t\,|\,\lambda],\ 1\le i\le K \tag{1}$$

則可將t+1時間落在文句狀態 s_j 的相似度分數表示為

$$\delta_{t+1}(s_j)=\max_{1\le i\le K}\left[\,\delta_t(s_i)a_{ij}\,\right]b_j(x_{t+1}),\quad 1\le j\le K \tag{2}$$

其中，λ 為比對所使用的聲學模型，a_{ij} 為狀態 s_i 轉移到狀態 s_j 的轉移機率，$b_j(x_{t+1})$ 為語音 x_{t+1} 在狀態 s_j 的機率分布。不斷反覆Eq.(2)直到語音序列終點T，可得終點T的相似度分數為

$$\max_{1\le k\le K}\left[\,\delta_T(s_k)\right] \tag{3}$$

也就是可由終點T裡選出一個具有最大相似度分數的文句狀態 s_k 與語音序列終點T對應。之後，可由語音序列終點T與文句狀態 s_k 的交會點 $\varphi(T,s_k)$ 回溯至搜尋空間 $\Gamma(T,K)$ 原點 $\varphi(1,s_1)$ 得到最佳

路徑 $path(T,s_k)$。(詳細演算法可參考[1])。

3.2. 維特比演算法特性

維特比演算法以最大相似度來施行，存在有以下兩個特性：

特性一：假設部分搜尋空間 $\Gamma(t,\ell)$ 存在一終止於時間 t 對應到文句狀態 s_i 的路徑 $path(t,s_i)$，該路徑為搜尋空間 $\Gamma(T,K)$ 最佳路徑的一部份。若在 $\Gamma(t,\ell)$ 裡有另一路徑 $path(t,s_j)$ 與 $path(t,s_i)$ 在 $\Gamma(t,\ell)$ 空間裡有一對應到時間 τ 與文句狀態 s_n 交會點 $\varphi(\tau,s_n)$，則由交會點 $\varphi(\tau,s_n)$ 至搜尋空間的原點 $\varphi(1,s_1)$ 之間的最佳路徑 $path(\tau,s_n)$ 必為 $path(t,s_i)$ 的一部份，也必為 $\Gamma(T,K)$ 最佳路徑的一部份。

說明：由於路徑 $path(t,s_i)$ 為 $\Gamma(T,K)$ 最佳路徑的一部份，而 $\varphi(\tau,s_n)$ 又位在 $path(t,s_i)$ 中，因此，由 $\varphi(\tau,s_n)$ 所決定出來的最佳路徑 $path(\tau,s_n)$ 必然是屬於 $\Gamma(T,K)$ 最佳路徑的一部份。

特性二：由 $\varphi(\tau,s_n)$ 所決定出來的最佳路徑 $path(\tau,s_n)$ 必定只有一條，且必定是使得 $\varphi(\tau,s_n)$ 所在位置的相似度分數 $\delta_\tau(s_n)$ 最大的狀態序列 $Q_\tau=(q_1q_2...q_{\tau-1},q_\tau=s_n)$。

說明：維特比算法以最大相似度來施行，每一時間對應到每一狀態的相似度分數都是最大的，且必定僅有一最佳狀態序列與之對應。

3.3. 可靠路徑估測

由於以最大相似度來施行，最佳路徑終點通常都具有較大的相似度分數，因此，如果最佳路徑落在部分搜尋空間 $\Gamma(t,\ell)$ 裡，則在時間 t 由部分文句狀態序列 $S_\ell=(s_1s_2...s_\ell)$ 選出 N 個具有較大相似度分數的狀態，是非常有可能將最佳路徑的終點狀態涵蓋進來。以這些具有較大相似度分數的狀態所回溯出來的路徑就有可能包含 $\Gamma(t,\ell)$ 的最佳路徑。所以，假設這些路徑中包含有最佳路徑，並且這些路徑有共同的交會點，依照特性一，該交會點必定落在最佳路徑裡，並且，由交會點回溯到 $\Gamma(t,\ell)$ 原點的路徑，必定為最佳路徑的一部份，依照特性二，由交會點所回溯的最佳路徑僅有一條，所以可由任一經過交會點的路徑取得該最佳路徑。由於是以 N 個具有較大相似度分數的狀態來取得部分搜尋空間 $\Gamma(t,\ell)$ 裡可能的最佳路徑，我們稱這個取得最佳路徑方式為可靠路徑估測。

可以了解的是，當 N 越大涵蓋最佳路徑終點狀態的可能性就越大；反之，當 N 小時，就不一定保證能涵蓋最佳路徑的終點狀態，尤其當訓練聲學模型所使用的語音語料庫與待估測語音的特性差異很大時，可能就必須加大 N 的數量，以容忍不同的語音環境。另外，由於我們以文句狀態序列來與語音序列進行較準，也就是說，無論是中文的音節、英文的單詞或是存在於語音序列裡的靜音都轉化為狀態序列來表示(譬如：以 3 個聲母狀態和 5 個韻母狀態來表示一個含 8 個狀態的中文音節、以 3 個狀態來表示英文音素及以多個音素來描述一個英文單詞、以及以 1 個可有可無的靜音狀態來表示可能存在於語音序列裡的靜音等，將文句序列轉化為文句狀態序列來表示)，因此，所決定之可靠路徑之端點就不一定是中文的音節端點、英文的單詞端點或靜音處，亦可能是存在於中文音節、英文單詞的內部狀態位置，或是靜音狀態位置。不過，只要能決定出可靠路徑，藉由可靠路徑資訊取得語音序列與文句狀態序列的對應關係，無論是中文的音節端點、英文的單詞端點或是存在於語音序列裡的靜音位置，都是可以決定的。即便是使用者自行標示的文句段落位置，亦可藉由該可靠路徑資訊來取得其所對應的語音序列位置。

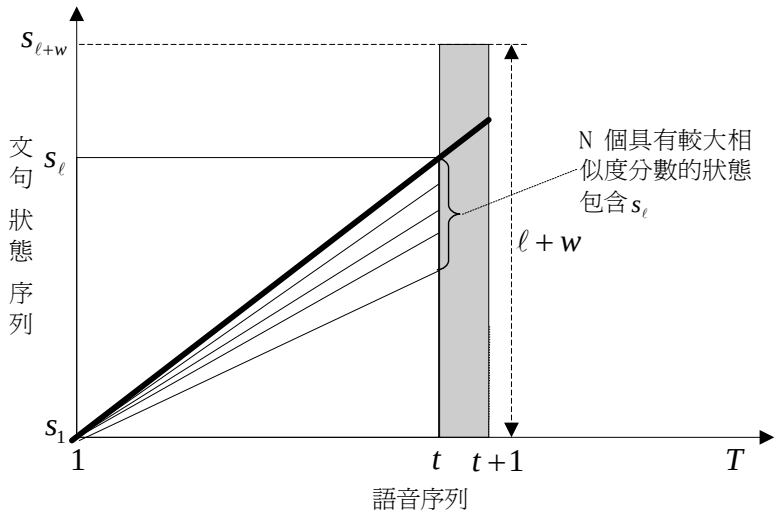
3.4. 部分搜尋空間調整

調整部分搜尋空間的情況共有以下二種：

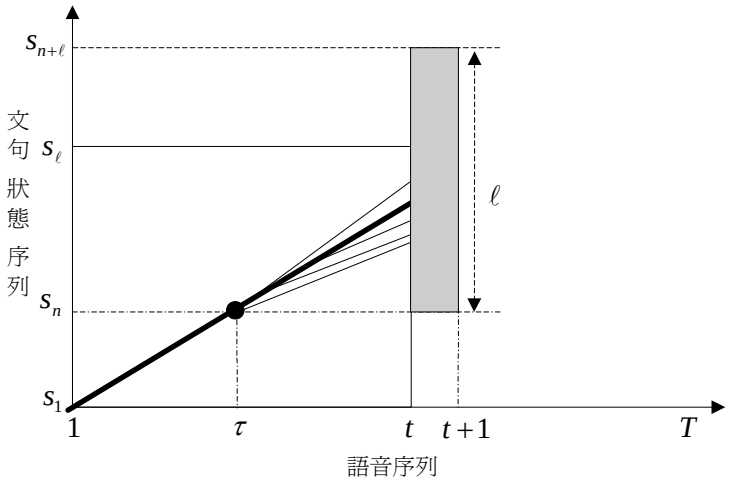
調整一：在部分搜尋空間中未找到可靠路徑，加大部分搜尋空間之文句狀態，以防止可能的最佳路徑在下一時間比對時落在部分搜尋空間外。如圖二，時間 t 所取得的 N 個具有較大相似度分數的狀態集合中包含部分文句狀態序列 $S_\ell = (s_1 s_2 ... s_\ell)$ 的最後一個狀態 s_ℓ，則在下一個時間 t+1 進行比對之前，應加大部分搜尋空間的文句狀態序列為 $\ell + w$，以防止如圖一情況，可能的最佳路徑落在部分搜尋空間之外。

調整二：一旦在部分搜尋空間裡已取得可靠路徑，由可靠路徑終點，我們可重設下一次比對的部分文句序列。圖三是一個在未加大文句狀態序列下，在時間 t 取得可靠路徑，在 t+1 時間保持下一次比對的部分文句序列為 ℓ 個狀態。圖四則是文句狀態序列加大為 $\ell + w$ 後，在時間 t 取得可靠路徑，在去除 n 個文句狀態之後，剩餘的文句狀態 $\ell + w - n$ 大於 ℓ，t+1 時間調整部分文句序列為 $\ell + w - n$ 個狀態。

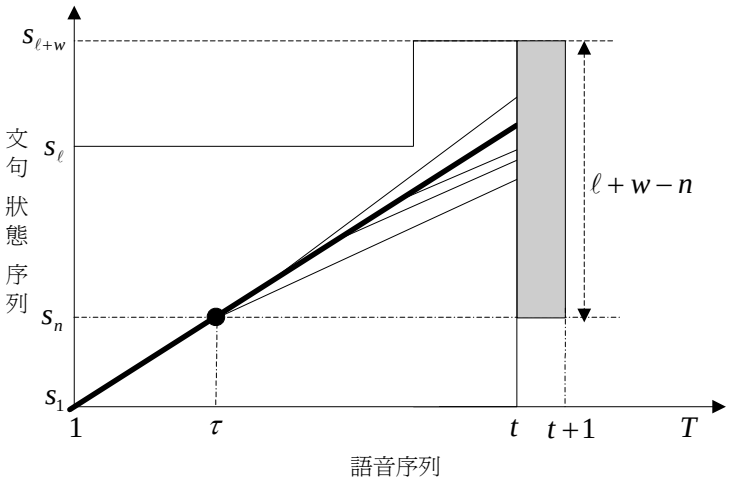
因此，藉由可靠路徑估測和上述搜尋空間的調整，我們就可以循序的以部分語音文句校準完成大量連續語料的語音文句校準工作。圖五是以部分搜尋空間來涵蓋整體搜尋空間 $\Gamma(T, K)$ 的最佳路徑示意圖。以過去傳統作法進行語音文句校準，需要對整體搜尋空間 $\Gamma(T, K)$ 做運算之後才得以決定整體搜尋空間的最佳路徑，以本文所介紹的方法則可以省去如圖五灰色區間的運算量。當整體搜尋空間不大時，本文作法與傳統作法所需的運算量或許差異不大，但是當整體搜尋空間變大，如動輒 5 分鐘以上的教學錄音帶或是光碟音軌，本文作法可節省相當多的運算量，是一種相當有效率的作法。



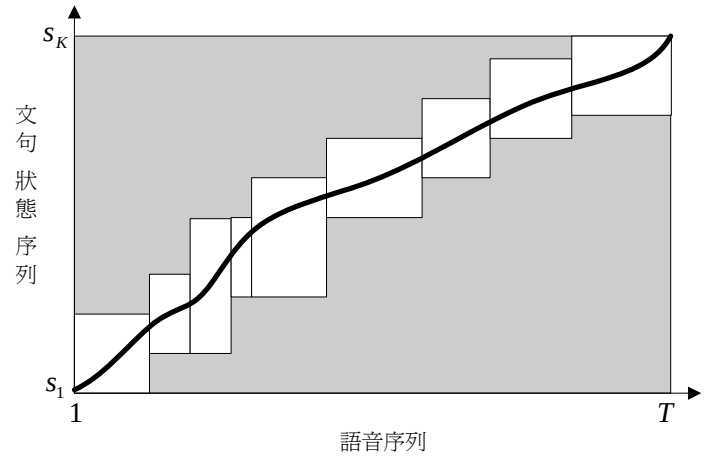
圖二：於語音序列時間 t 選出 N 個具有較大相似度分數的狀態集合，該集合含有部分文句狀態序列的最後一個狀態，於 t+1 時間加大文句狀態序列為 $\ell + w$



圖三：語音序列時間 t 取得可靠路徑，當去除可靠路徑所含 n 個文句狀態序列之後，若剩餘的文句狀態少於 ℓ 個狀態，於 t+1 保持文句狀態序列為 ℓ 個狀態



圖四：語音序列時間 t 取得可靠路徑，當去除可靠路徑所含 n 個文句狀態序列之後，若剩餘的文句狀態 $\ell+w-n$ 大於 ℓ 個狀態，於 t+1 保持 $\ell+w-n$ 個文句狀態



圖五：以部分搜尋空間集合來涵蓋整體搜尋空間 $\Gamma(T,K)$ 的最佳路徑示意圖

3.5. 部分語音文句校準演算法

以下將部分語音文句校準演算法整理如圖六。

1. 取出部分文句狀態序列$S=(s_1 s_2 ... s_\ell)$，並初始文句狀態序列的機率分數$\{\delta_1(s_1),\delta_1(s_2),...,\delta_1(s_\ell)\}$爲極小值；
2. 循序取出每一語音音框t進行下列步驟：
   2.1 依照Eq.(2)，但限制$1\le i\le \ell$和$1\le j\le \ell$，決定每一文句狀態序列的機率分數$\{\delta_t(s_1),\delta_t(s_2),...,\delta_t(s_\ell)\}$，並紀錄其前一語音音框t-1的最佳狀態位置；
   2.2 依照機率分數選出N個具有較大機率分數的文句狀態；
   2.3 以t和N個具有較大機率分數的文句狀態進行回溯，求出路徑的交會點$\varphi(\tau,s_n)$；
   2.4 可靠路徑資訊紀錄和搜尋空間調整：
      2.4.1 若$\varphi(\tau,s_n)$非部分搜尋空間的原點$\varphi(1,s_1)$，則由$\varphi(\tau,s_n)$回溯至$\varphi(1,s_1)$的路徑爲可靠路徑，輸出可靠路徑數據，依照部分搜尋空間調整二，重設部分文句狀態序列，並初始文句狀態序列裡新增加狀態的機率分數爲極小值；
      2.4.2 若$\varphi(\tau,s_n)$爲部分搜尋空間的原點$\varphi(1,s_1)$，則無可靠路徑，若N個具有較大機率分數的文句狀態含部分文句序列最後一個狀態，依照部分搜尋空間調整一，加大部分文句狀態序列，並初始文句狀態序列裡新增加狀態的機率分數爲極小值；
   2.5 若尚有語音信號，重複進行2.1至2.4步驟；
3. 剩餘的部分搜尋空間以部分搜尋空間的語音信號終點和文句狀態序列終點進行回溯，求出其最佳路徑並輸出最佳路徑數據。

圖六：部分語音文句校準演算法

由於機率分數隨著語音信號不斷累積可能會產生溢流（Run-Off）問題，因此，可在重設部分文句狀態序列時，將累積的機率分數進行重設，以避免溢流情況發生。

4. 實驗與結果討論

4.1. 測試語料

我們使用下列幾套語料來驗證上述作法，並以語句邊界的正確率做為語音文句校準的評估。

語料一（DB1）：來自工研院的104自動總機系統[4]所蒐集的人名語音，這些語音都是以8 KHz、16-bit、Mono格式所錄製的電話語音，並且可能夾雜一些背景雜訊，如打字的聲音、話筒撞擊聲或有說話的背景等。我們將其中的751句串接成一約23分15秒的長串語音，共包含2247個音節，句子與句子之間我們以一個分隔符號來作區隔。我們分別以傳統的維特比演算法定出這751句人名語音每一句語音的語音起點和終點邊界，即去除靜音部分，並且求得這些語音邊界對應長串語音的絕對位置作為正確的邊界位置，共包含1502個邊界。

語料二（DB2）：來自教學用的語音光碟[5]，這些語音都是以44.1 KHz、16-bit、Stereo格式儲存。我們取出其中4段音軌，並且將語音格式轉換為8 KHz、16-bit、Mono格式，之後，將這些語音串接成一約23分48秒的長串語音，共包含5175個音節。我們以人工方式將該語音資料分成421句，之後，如語料一的處理方式，插入分隔符號於句子與句子之間，並分別定出每一句語音的語音起點和終點邊界和其所對應長串語音的絕對位置作為正確的邊界位置，共包含842個邊界。

含背景音樂的語料（DB1+MU, DB2+MU)：將上述數語料（DB1, DB2）以語料裡的平均音量為基準，加入對應強度約10 dB SNR的古典樂（四季，維瓦第)，以觀察受背景音樂干擾的語音文句校準情形。同樣使用DB1和DB2所求得的邊界位置為正確答案來進行評估。

4.2. 聲學模型

實驗所使用的聲學模型係自MAT電話語音語料庫[6]訓練而得，語音語料庫的格式為8 KHz、16-bit、Mono。聲學模型共含100個右相關聲母模型、38個左右無關的韻母模型和1個靜音模型。每個聲母模型以3個狀態來描述，韻母模型以5個狀態來描述，靜音模型則僅以1個狀態來描述。聲、韻母模型狀態使用10個混合數，靜音使用64個混合數。

除了上述的聲學模型之外，在進行校準時，我們也使用具有 64 個高斯混合數的語音/非語音模型來進行語音/非語音判斷，並將判斷為非語音的樣本收集起來訓練一個僅有一個混合數的背景模型。將該背景模型與前述靜音模型合併成 65 個混合數，作為部分語音文句校準時吸收靜音和背景雜訊之用。

4.3. 實驗條件設定和語音邊界偵測

不加背景音樂的語料（DB1, DB2)，我們以N=40來估測可靠路徑。加了背景音樂的語料（DB1+MU, DB2+MU)，則以N=80來估測可靠路徑。

部分文句狀態序列長度 ℓ 設為 20 個音節（也就是 180 個狀態，含音節與音節之間的一個靜音狀態)。加大文句狀態序列的序列長度 w 也是使用 20 個音節來防止最佳路徑可能落在部分搜尋空間之外的情況。

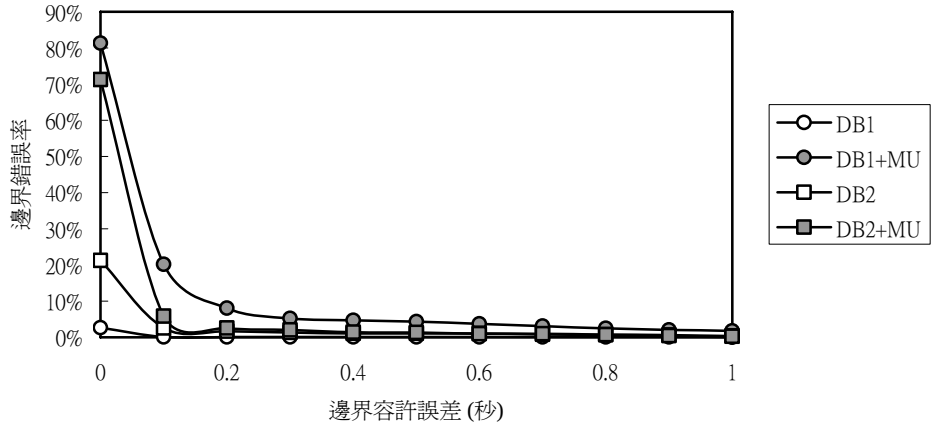
在取得可靠路徑之後，檢驗可靠路徑所對應的文句序列中是否含有分隔符號，若含有分隔符號，我們以下列方式定出前一句語音的終點邊界和後一句語音的起點邊界：(一）若分隔符號所對應的語音位置含有靜音，靜音之前為前一句語音的終點邊界，靜音之後為後一句語音的起點邊界；(二）若不含靜音，則前一句語音的終點邊界和後一句語音的起點邊界為同一邊界。另外，第一句語音的起點邊界和最後一句語音的終點邊界都以不包括靜音部分為其邊界。

4.4. 結果與討論

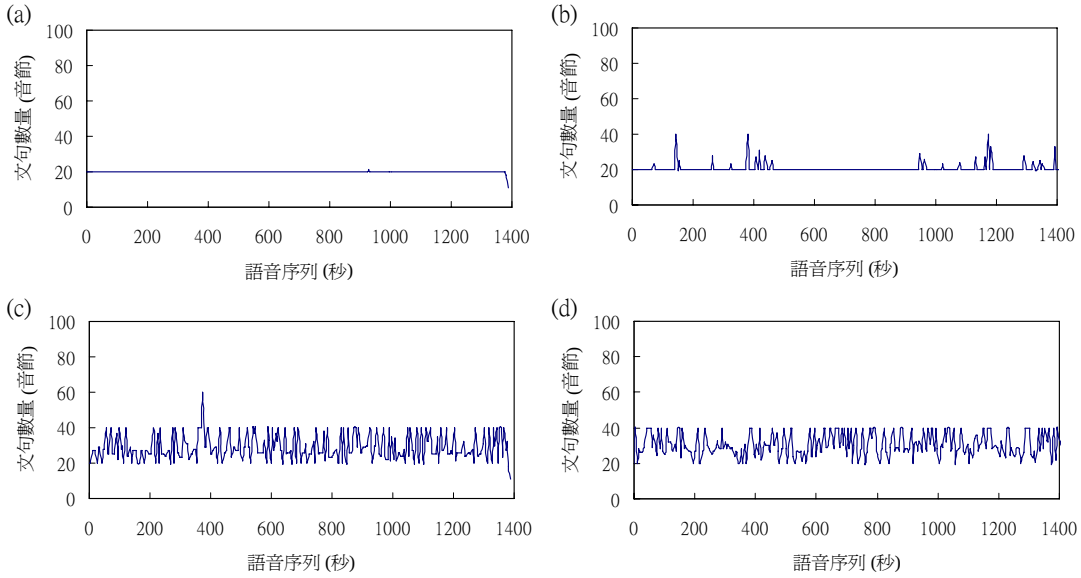
以前述條件進行實驗，4組語料都可以得到正確的邊界數量；DB1和DB1+MU可得到1502個邊界，DB2和DB2+MU可得到842個邊界。圖七顯示4組語料進行語音文句校準之後得到的邊界與正確邊界在不同容許誤差的邊界錯誤率分布情形。不加背景音樂語料DB1和DB2之偵測邊界與正確邊界有相當高的吻合度，尤其是DB1與聲學模型的訓練語料庫有較一致性的語音特性(都是電話語音)，在0.1秒的容許誤差下，可以得到幾近於零錯誤的結果(0.07%)。加入背景音樂語料DB1+MU和DB2+MU之偵測邊界與正確邊界雖然有較大的差異，但是在1秒的容許誤差下，仍然可以得到相當低的邊界錯誤率(DB1+MU：1.73%，DB2+MU：0.24%)。其中，可以注意到DB1+MU的邊界錯誤率比DB2+MU高，這是由於DB1語料庫是由不同說話人的語音串接而成，每一句語音的音量不固定，當加入固定強度的背景音樂時，每一句語音的SNR就不一定是我們所設定的10 dB。而DB2係來自教學用的語音光碟，雖然也有不同說話人的語音，但是出版者為控制其語音品質，DB2裡每一句的音量顯然是較為平均的，這也使得DB2+MU的SNR整體上較接近於我們所設定的10 dB。因此，DB1+MU的語音條件顯然較DB2+MU差一些，也使得其邊界錯誤率較高。

圖八是 4 組語料的比對歷程，在 N=40 的條件下，DB1 和 DB2 每秒約處理 20 個音節，佔整體

搜尋空間的 1%以下。在 N=80 的條件下，DB1+MU 和 DB2+MU 每秒約處理 30 個音節，佔整體搜尋空間的 1.5%以下。兩者皆具有相當高的執行效率。上述實驗都是以一般的電腦系統來執行(AMD 1.0G Hz CPU, Windows 2000)，4 組語料庫都可以在 1 個實體時間內(RT，Real Time)完成切割(N=40 約 0.6 RT 完成，N=80 約 0.98 RT 完成)。



圖七：語音文句校準之後得到的邊界與正確邊界在不同容許誤差的邊界錯誤率分布



圖八：四組測試語料比對歷程

(a) DB1, N=40; (b) DB2, N=40; (c) DB1+MU, N=80; (d) DB2+MU, N=80

5. 結論與未來方向

本文提出了一種可靠路徑估測演算法，以找出可能落於部分搜尋空間裡的部分最佳路徑，以部份語音文句校準循序的定出落於部分搜尋空間裡的部分最佳路徑，以使維特比演算法仍能使用於大量連續語料的語音文句校準。實驗顯示該演算法的穩定性和使用部分校準的效率，其不但可適用

在一般無背景噪音的大量連續語料的語音文句校準，在高SNR背景音樂的情況下也能獲得不錯的結果。

雖然本文提出的演算法可有效處理大量連續語料校準,但是偵測邊界仍會與實際邊界有一些差異，在要求高精度的語料庫處理上仍不免需要人工檢驗，如何進一步提高切割精度；或者，標示出偵測邊界的信心度，以降低人工檢驗和人工校正的負擔，將是我們未來的工作重點之一。

6. 計畫相關資訊

本文係工研院資通所執行經濟部九十五年度前瞻研究專案 5301XS2310 的計畫成果之一。

7. 參考文獻

1. Rabiner L. and Juang B.-H., “Fundamentals of Speech Recognition,” New Jersey, Prentice-Hall International, 1993, pp. 339-340.
2. Robert-Ribes J. and Mukhtar R.G., “Automatic Generation of Hyperlinks Between Audio and Transcript,” Eurospeech, 1997.
3. Moreno P.J., Joerg C., Van Thong J.-M., and Glickman O., “A Recursive Algorithm for the Forced Alignment of Very Long Audio Segments,” ICSLP, 1998.
4. 謝偉強、簡世杰、許志興、張森嘉，”工研院 104 自動總機系統的改進過程”，電腦與通訊，2001 年，第 96 期，pp. 29-34.
5. 康軒文教事業，”TOP945 兒童雙週刊中年級版第 8 期”，http://top945.knsh.com.tw，2003 年.
6. Wang H.-C., “MAT — A Project to Collect Mandarin Speech Data Through Networks in Taiwan,” Computational Linguistics and Chinese Language Processing, Vol. 2, No. 1, pp. 73-89, 1997.

利用聲學與文脈分析於多語語音辨識單元之產生

Generation of Phonetic Units for Multilingual Speech Recognition Based on Acoustic and Contextual Analysis

王士豪 [1]　　黃建霖 [2]　　吳宗憲 [2]

[1] 財團法人資訊工業策進會
Email: shwang@iii.org.tw

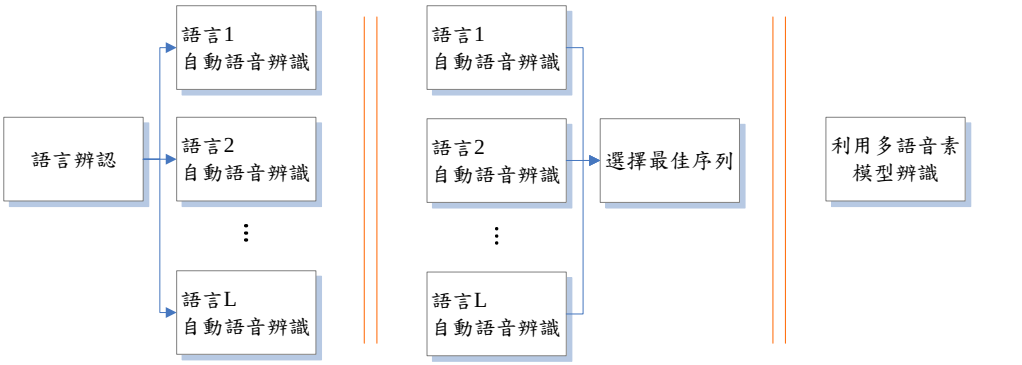
[2] 國立成功大學資訊工程系
Email: {chicco, chwu}@csie.ncku.edu.tw

摘要

由於全球化趨勢之盛行，多語語音常出現於會議紀錄及一般對話等方面。對於會議紀錄及對話系統而言，多語語音自動辨識日顯重要。在多語語音自動辨識中，辨識單元集之定義及選取，將影響辨識之效率及效能。本論文針對中英文利用 IPA 定義之多語語音辨識單元集，考慮前後文相關之三連音模型，並進一步透過對聲學相似度與前後文脈分析，決定一組精簡有效的多語辨識單元。在相似度矩陣分析中，首先我們利用事後機率統計，建立聲學相似度矩陣，然後，基於發音共聲現象的考量，分析語音發音上之相似度。本論文更引入語言超空間相似度之觀念，計算三連音辨識單元前後文脈之關係，建立語言超空間相似度矩陣。最後利用資料融合技術，合併聲學相似度矩陣和語言超空間相似度矩陣，以計算三連音辨識單元間之距離，而後利用向量量化群集方法合併相似性高之三連音辨識單元，建立一個有效的多語語音辨識單元集。本論文以 EAT 中英雙語語料庫作實驗評量，比較所提方法與之前研究方法上的差異與改進。由實驗結果得知，本論文所提出利用聲學相似度與前後文脈分析於多語語音辨識單元集之產生，可提高其辨識效能。

1. 簡介

語音是人類溝通最自然方便的方式，近年來電腦網路多媒體普及，語音在人機互動的介面更是扮演重要角色。自動語音辨識是語音應用技術的重要一環，目前有許多有關語音辨識之應用，如：自動聽寫機、語音文件摘要、語音文件檢索、口述語言對話系統以及語音命令控制等。並且隨著全球化趨勢的來臨，文化交流，商業活動和網路資訊都充斥著多語(multilinguality)的環境及各式各樣的應用，多語自動語音辨識(multilingual speech recognition)顯得相形重要。一般而言，多語自動語音辨識作法上可以歸納成三大類，如(圖 1)所示：



(a) 前端語言辨識方法　(b) 後端最佳序列選取方法　(c) 建立多語音素方法

圖 1. 多語語音辨識三類作法之流程

第一類作法：為在前端處理先利用一個語言辨認 (language identification, LID)方法[1][2]，判斷輸入語音訊號屬於哪種語言，再分別透過單一語言的自動語音辨識器進行轉譯。對於多語語音辨識演算法，前端語言辨認的成效決定了多語語音辨識的好壞，有效的語言辨認方法將使得多語語音辨識的正確率達到和個別單一語言的獨立語音辨識相當之效果。因此，利用先前的語言辨認方法，缺點在於辨識效果會受到 LID 表現影響。

第二類作法：利用個別單一語言的語音辨識器分別產生辨識結果，再經由後端處理選擇最大似然的方

法(maximum likelihood, ML)，決定最佳的多語辨識序列。作法上類似對語音辨識序列做驗證(verification)之處理 [3]；多語語音辨識的表現取決於後端最佳序列選擇之效果。利用選擇最大似然的方法缺點在於，多語辨識的效果會受到 ML 方法的限制，且辨識的多語句型需要另外考慮，切割出語句內不同語言的段落。

第三類作法：藉由定義多語語音辨識單元集[4]，合併個別單一語言之音素模型，來進行多語語音辨識。本論文乃基於此方法，探討如何定義出有效的多語語音辨識單元模型。

在多語音素模型之建立可以歸納為三種方式。首先，我們可以直接合併個別單一語言之音素集，建立多語音素模型，但是這種方法沒有考慮多語音素間參數分享的特性。第二，藉由對照國際音素標準定義，考慮個別單一語言之音素，達到多語音素間參數共用的特性，但是此作法上缺乏資料統計的分析，而是由專家知識決定各音素定義。國際音素標準定義包含有：International Phonetic Alphabet (IPA) [5]、Speech Assessment Methods Phonetic Alphabet (SAMPA) [6] 和 Worldbet [7]等。第三，估計多語語音音素間相似程度，由下而上階層式進行多語音素合併，以定義多語音素集。多語語音音素間相似度的量測，可以利用 Bhattacharyya distance [8] 或者是 Kullback-Leibler (KL) divergence [9]的方法，計算多語音素模型間的距離，決定相似度以定義多語音素集。此作法上，同時考慮多語音素間參數分享的特性，並利用資料統計分析決定音素定義。但是缺點在於計算模型參數間的距離，與實際辨識演算法在執行時，所考慮的聲學相似度(acoustic likelihood)不符。

本論文探討中英文之多語語音辨識之研究，從中英文基本音素作分析。中文可以分為 37 個音素，英文可分為 39 個音素。考慮語音發音共聲的現象(co-articulation)，本論文定義前後文相關之三連音素模型(contextual tri-phone models)，進一步對語音發音相似度作聲學相似度(acoustic likelihood)分析。此外更導入語言超空間相似度分析(hyperspace analog to language, HAL)，考量三連音辨識單元前後文脈之關係，以改善過去單純考量模型參數聲學相似度來量測語音音素間相似度之方式，以決定多語音素模型，符合語音發音中受前後文影響之特性。最後，以資料融合的技術合併定義發音相似的音素。實驗評估，利用自行開發的多語語音辨識系統，使用隱藏式馬可夫模型(hidden Markov model, HMM)，建立以音素為基礎的聲學模型，並配合多語語言模型和多語發音辭典文法樹，進行連續多語語音辨識。

接下來的文章結構將分別探討如下：第二節，探討過去對多語語音辨識之研究。第三節，說明論文方法建立精簡有效的多語音素模型於自動語音辨識之應用。第四節，針對本論文所提方法建立之多語音素模型進行辨識結果評估，實驗並與之前方法比較。第五節是討論說明與結論。

2. 多語語音辨識之音素定義相關研究

多語語音辨識音素定義的方法，主要可分為三種方式：(一)直接結合個別單一語言之音素定義；(二)依據國際音素標準定義，找出個別單一語言之音素聯集；(三)從資料分析的角度(data-driven)，合併個別單一語言之相似音素。現分別介紹如下：

2.1. 直接結合個別單一語言之音素

如(表 1)所示，比照中文和英文單一語言音素的定義。

表 1. 結合中英文音素定義

音素類別	中文	英文
有聲破裂音	b_M, d_M, g_M	b, d, g
無聲破裂音	p_M, t_M, k_M	p, t, k
摩擦音	f_M, s_M, sh_M, h_M, x_M	f, v, th, dh, s, sh, hh
塞擦音	c_M, ch_M, j_M, q_M, z_M, zh_M	ch, jh, z, zh
鼻音	m_M, n_M	m, n, ng
流音	r_M, l_M	r, l
滑音		w, y
前部母音	i_M, v_M, ei_M, er_M	ih, eh, ae, iy, ey
中部母音	an_M, ang_M, en_M, eng_M	ah, uh, er
背部圓唇母音	o_M	ao
背部非圓唇母音	a_M, u_M, ou_M, e_M, ee_M, ai_M, ao_M	aa, uw, ow, ay, oy, aw

將各目標語言的音素合併成一個多語語音辨識音素集合，此方式是較為直覺的多語音素定義方式。(表 1) 內

之音素類別參考 Chomsky 定義[10]，本論文對中文音素記號多以"_M"標籤區隔，分別表示 37 個中文音素定義和 39 個英文音素定義。此方法結合中英兩種語言之音素，建立多語語音辨識之聲學模型。作法上的缺點，在於各目標語言中相似之音素，模型參數無法分享，而且當需要結合的目標語言變多的時候，所需要定義的音素模型會大量隨之增加。

2.2. 以 IPA 為基準定義多語音素

第二種多語音素定義方式是基於專家的知識，將個別獨立的單一語言對應到 IPA 標準的符號定義，藉此各語言間可以分享相同的音素定義。如(表 2)所示是以 IPA 為標準之中英多語音素的定義。

表 2. 以 IPA 為標準之中英多語音素定義

音素類別	IPA 為標準之中英多語音素
有聲破裂音	B, D, G
無聲破裂音	P, T, K
摩擦音	F, S, SH, H, X, V, TH, DH
塞擦音	Z, ZH, C, CH, J, Q, CH, JH
鼻音	M, N, NG
流音	R, L
滑音	W, Y
前部母音	I, ER, V, EI, IH, EH, AE
中部母音	ENG, AN, ANG, EN, AH, UH
背部圓唇母音	O
背部非圓唇母音	A, U, OU, AI, AO, E, EE, OY, AW

(表 2) 內之音素類別參考 Chomsky 定義[10]。如此規則地將中英兩種語言的音素結合，共計有 52 個中英雙語音素定義。作法上可以有效地將部分的中英文音素合併，共享語言間彼此的共同音素，減少語音音素模型的定義和訓練。但此作法的缺點是建構在專家知識的分析，而非從資料特性統計的角度定義。也就是說，直接對照 IPA 定義產生的多語音素集，並沒有考慮到音素模型間頻譜特性。專家知識分析的多語音素集，與最後進行語音辨識，從資料分析角度建立的統計模型計算不一致。因此，採用直接對照 IPA 定義之多語音素模型並不能確實地呈現統計訓練資料上的分佈。

2.3. 量測音素相似度定義多語音素集

除了直接混合多語音素定義，以及利用 IPA 國際標準定義的多語音素，過去研究也曾利用估測三連音素模型間的相似度，以 HMM 模型參數距離計算，利用遞迴方法合併三連音素模型 (triphone)，建構出多語語音辨識的音素集[8][9]。兩個高斯分佈的相似可以利用平均值和變異數函數，來描述彼此的相似程度。利用 Bhattacharyya distance [8]來計算音素模型間的距離 D_{bha} ，表示如下：

$$D_{bha}=\frac{1}{8}(\mu_1-\mu_2)^T\left[\frac{\Sigma_1+\Sigma_2}{2}\right]^{-1}(\mu_1-\mu_2)+\frac{1}{2}\ln\frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \qquad (式\ 1)$$

其中，μ 和 Σ 分別表示音素模型的平均值和變異數向量，T 是轉置矩陣。另外，可以利用 Kullback-Leibler (KL) divergence [9]來決定兩個機率分佈的相似度 D_{KL} 。以 KL-divergence 估算兩個高斯分佈 $N(\mu_1,\Sigma_1)$ 和 $N(\mu_2,\Sigma_2)$ 的相似度，表示如下：

$$D_{KL}=\frac{1}{2}\left(\ln\frac{|\Sigma_1|}{|\Sigma_2|}+\mathrm{tr}\left(\Sigma_1^{-1}\Sigma_2\right)+\left(\mu_1-\mu_2\right)^T\Sigma_1^{-1}\left(\mu_1-\mu_2\right)-d\right) \qquad (式\ 2)$$

不論是以 D_{bha} 或者 D_{KL} 的評估方式，最後利用階層式由下而上進行音素相似度比較，依據音素模型的參數距離來考慮是否合併。其詳細的演算法表示如下：

1. 初始 Initialization:
建立各別單一獨立語言的音素模型。
2. 迴圈 Loop:
計算各音素模型間兩兩彼此的距離，並將最小距離的音素作合併。
3. 結束 Termination:
確認是否達到期望的音素定義個數，或者各音素之間的距離皆大於合併的最小標準。如果是，則結束迴圈；否則繼續進行第二步驟。

利用計算 HMM 模型參數(μ 和 $\sum$)的距離，來決定三連音素的合併，以建立多語音素模型。此方法的缺點與實際辨識演算法在執行時，所考慮的聲學相似度(acoustic likelihood)不符。為了改進上述方法的缺失，本論文提出利用聲學相似度及前後文脈分析，定義有效的多語辨識音素模型，詳細作法將於下節說明。

3. 利用聲學及前後文脈分析於多語音素模型定義

過去直接結合個別單一獨立語言之音素，建立多語音素集合作法上的缺點，在於沒有考量多語音素之間可能的共同音素分享。而直接利用 IPA 標準對照出多語音素集合，則缺乏實際資料統計和訊號特性上的分析。以 Bhattacharyya distance 或 KL-divergence 的作法，在於只有考慮音素模型的參數距離，並沒有針對實際辨識的應用做分析。因此，本論文提出聲學相似度及前後文脈分析自動建立多語音素模型定義，作法如(圖 2)所示。

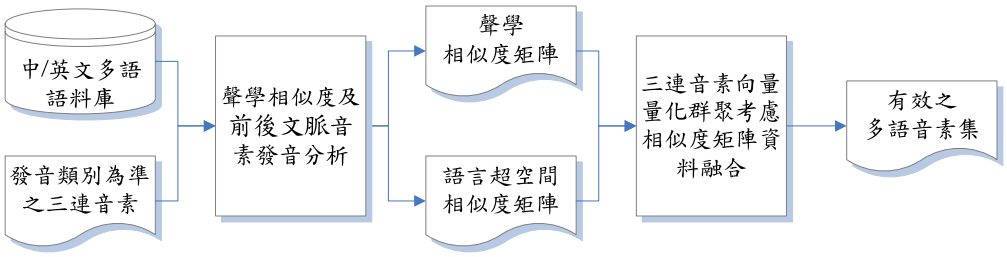


圖 2. 利用聲學相似度及前後文脈分析有效之多語音素集

對於目標語言以中文和英文為例，首先我們先對應 IPA 標準定義出中英文音素之定義，考慮前後音發聲的影響(left and right context dependent)，建立以 IPA 中英標準定義為基礎之三連音素集(triphone)。此外，從資料分析的角度(data-driven)針對聲學相似度及前後文脈做分析。對於聲學音素發音分析，計算各音素之 HMM 模型的聲學相似度(acoustic likelihood, ACL)，並建立聲學相似度矩陣。比較先前計算各音素模型之間參數距離的作法 [8] [9]，利用聲學相似度的做法更符合實際語音辨識上的考量。另外，配合語言超空間相似度分析(hyperspace analog to language, HAL) [11]，從前後文脈的分析，歸納出在連續發音中音素隨著前後文脈產生之發音特性，建立語言超空間相似度矩陣。聲學和語言超空間之相似度矩陣內，每一個音素可以利用向量表示，並計算出音素在空間中彼此相似度關係。最後，利用向量量化(vector quantization, VQ)的方法群聚相似的音素[12]，配合資料融合(fusion)的技術我們可以同時考慮聲學和前後文脈發音的影響，建立有效的多語音素集。

3.1. 聲學相似度分析

對應 IPA 音素標準定義，找出多語音素之聯集。考慮音素發音的位置和發音方法，判斷音素前後文連接的發聲情形，可以定義出 N 個三連音素模型。收集多語語音訓練語料以資料分析的角度，計算各音素在聲學上的相似度，建立聲學相似度矩陣。在聲學相似度矩陣的建立上，論文採用直接校準(forced alignment)的方法，與建立的 HMM 模型進行音素的辨識，利用直接校準方法可以確保參照出一樣的音素個數序列，避免辨識發生插入(insertion)和刪除(deletion)等錯誤的情形。統計語料內第 l 個音素模型與第 k 個音素模型間之相似度，其計算方式為將第 l 個音素所有訓練語音對第 k 個音素模型 ω_k 估算觀測事後機率值 $P(x_l^i \mid \omega_k)$，其中 x_l^i 表示第 l 個音素中之第 i 個訓練資料計算音素之間取對數用距離的方式呈現音素間彼此的關係 $\log(P(x_l^i \mid \omega_k))$，以建立聲學相似度矩陣 $\mathbf{A} = (a_{kl})_{N \times N}$。為建立一個對稱聲學相似度矩陣，我們對其計算對角平均值。

$$a_{kl} = \frac{\frac{1}{I}\sum_{i=1}^{I}\log(P(x_l^i \mid \omega_k)) + \frac{1}{J}\sum_{j=1}^{J}\log(P(x_k^j \mid \omega_l))}{2} \qquad \text{(式 3)}$$

其中，I 和 J 分別為第 l 個音素與第 k 個音素訓練語音之個數。

3.2. 前後文脈之語言超空間分析

本論文針對語言發音上相似度分析，進一步引入文件探勘(text mining)的觀念，模擬在一視窗長度 ℓ 內音素變調的情形。由數個相連音素所呈現語音發聲上的變化，研究引入語言超空間相似度分析(hyperspace analog to language, HAL) [11]，基於發音共聲的行為，如果兩個相同音素在前後文脈類似的情況下，這兩個音素在發音特性上則存在有高的相似性。藉由 HAL 的方法可以從提供音素發聲上潛藏相似度分析，作為判斷音素相似及合併的依據。在 HAL 空間中，與前後文脈相關的三連音素模型，可以利用一個向量來描述與其他音素發音共聲(co-articulation)的現象。每一個向量維度表示目標音素與其他發聲於前後文中音素關聯性的強度，關聯性高則有較高的權重。權重的計算藉由一個長度 ℓ 的觀測視窗，統計語料內各種發聲的情形，所有在觀測視窗內的音素被視為一起出現的共生單元。因此，在視窗內任意兩個音素間的距離 d ，則其權重的計算為 $w=\ell-d+1$。如(圖 3)所示，為 HAL 視窗長度與權重之結構圖。

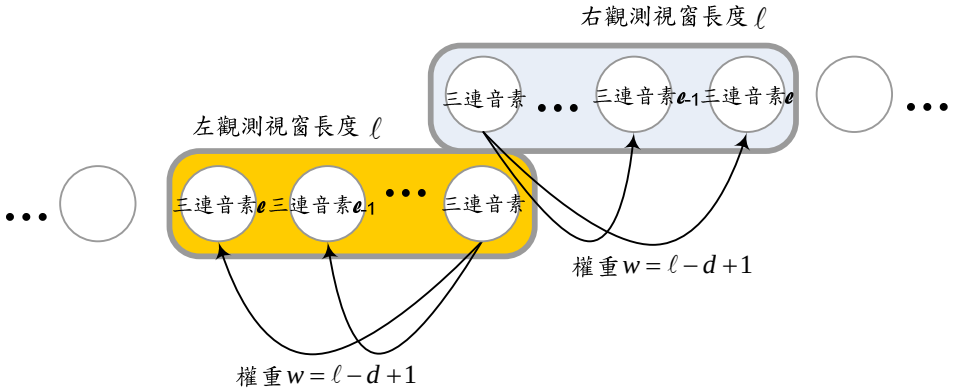


圖 3. HAL 視窗長度與權重之結構圖

在 HAL 空間建立步驟中，多語語音語料被視為一連串的音素發音序列。每串音素序列，藉由移動觀測視窗來計算語料內發音共聲的情形，每次移動一個音素。所建構出來的 HAL 空間是一個 $n\times n$ 大小的矩陣，其中 表示與前後文脈相關的三連多語音素個數。(表 3)是以"**Frank** (sil_F_R, F_R_AE, R_AE_NG, AE_NG_K, NG_K_sil) **早** (sil_Z_AO, Z_AO_sil)"的多語句子為例說明 HAL 空間矩陣 的計算方法，設定觀測視窗為 $\ell=3$ 。

表 3. HAL 空間矩陣

	sil_F_R	F_R_AE	R_AE_NG	AE_NG_K	NG_K_sil	sil_Z_AO	Z_AO_sil
sil_F_R							
F_R_AE	3						
R_AE_NG	2	3					
AE_NG_K	1	2	3				
NG_K_sil		1	2	3			
sil_Z_AO			1	2	3		
Z_AO_sil				1	2	3	

概念上，(表 3)的列向量(raw vector)表示音素與左邊文脈資訊的關聯；另外，(表 3)的行向量(column vector)則表示音素與右邊文脈資訊的關連。因此每一個音素 將由兩個向量維度呈現：

$$h_{l,k}=(v_l,v_k)=\left(\left\langle w_1^l,\ w_2^l,\ \ldots,\ w_N^l\right\rangle,\left\langle w_1^k,\ w_2^k,\ \ldots,\ w_N^k\right\rangle\right) \quad \text{(式 4)}$$

考慮音素發聲受到相鄰音素的影響，以三連音素 $h_{l,k}$ 為中心之空間相似度可以利用與右邊文脈相關之向量 v_l 及與左邊文脈相關之向量 v_k 之描述； w_N^l 和 w_N^k 分別表示利用觀測視窗於 HAL 空間內統計之音素相關權重，l 和 k 分別表示行與列之索引。

在 HAL 空間中，權重之計算需考慮正規化(normalization)因素，本論文利用在資訊檢索中相當重要之參數 $tf\times idf$ (term frequency and inverse document frequency) [13]，重新估計每個向量維度之權重，表示如下：

$$\bar{w}_i = w_i \times \log \frac{N}{C_i} \quad \text{(式 5)}$$

其中，w_i 指在向量 v_l 或向量 v_k 中第 i 個維度之權重；C_i 指在所有向量中，第 i 個維度之權重不為零的向量個數；N 為向量總個數或辨識單元個數。

3.3. 三連音素向量量化群聚考慮相似度矩陣資料融合

經過前兩小節分析，三連音素可以在聲學空間和語言超空間中，用向量的方式表示在空間中的相似程度。本論文分析聲學相似度矩陣 $\mathbf{A} = (a_{kl})_{N\times N}$ 和語言超空間相似度矩陣 $\mathbf{H} = (h_{kl})_{N\times N}$，同時考慮聲學相似度和前後文脈發音的特性，基於前後文脈相關之三連音素模型，合併相似之發音找出最為精簡有效的多語音素模型定義。作法上，參考資料融合的方法[14]，本論文利用加法融合的技術(sum rule)，結合兩相似度矩陣 $\mathbf{A}$ 和 $\mathbf{H}$，將聲學相似度和前後文脈的音素特徵作整合，表示如下：

$$\mathbf{S} = \alpha\mathbf{A} + (1-\alpha)\mathbf{H} = \sum_{l}^{N}\sum_{k}^{N}(\alpha \times a_{l,k} + (1-\alpha)\times h_{l,k}) \quad \text{(式 6)}$$

其中，α 是一個權重因子，負責融合聲學相似度和前後文脈的關聯。針對相似度矩陣 $\mathbf{A}$ 和 $\mathbf{H}$，論文中對其數值正規劃，將聲學和語言超空間相似度矩陣的分數結合，稱知識融合相似度矩陣 $\mathbf{S} = (s_{kl})_{N\times N}$ 為一個對稱矩陣，行 l 與列 k 均表示某一音素與其他音素相似度之向量。為了建立有效精簡的三連音素模型於多語語音辨識之應用，本論文利用向量量化(vector quantization, VQ)的方法[12]，從資料分析的角度(data-driven)，將原本三連音素自動地依據音素相似度分析，合併多語音素定義。向量量化為是一種非監督式的群集分析方法，可以將分散的資料群集成有意義的類別。三連音素在相似度矩陣分析後，可用向量方式表示其空間座標，論文引用[15]在矩陣中，兩向量夾角的計算方法，因此兩音素的相似度計算為 $c(s_l, s_k)$，計算如下：

$$c(s_l, s_k) = \frac{\overline{s_l} \bullet \overline{s_k}}{\left\|\overline{s_l}\right\| \cdot \left\|\overline{s_k}\right\|} = \frac{\sum_{i=1}^{N} s_i^l \times s_i^k}{\sqrt{\sum_{l=1}^{N} s_l^2} \times \sqrt{\sum_{k=1}^{N} s_k^2}} \quad \text{(式 7)}$$

其中，向量 s_l 表示目前相似度矩陣在行索引 l 的音素，向量 s_k 表示相似度矩陣在列索引 k 的音素，全部音素總共有 n 名。本研究利用調整性 k 群聚(modified k-means, MKM)分類方法[16]，定義收斂條件為分群內的資料變異度低於定義之門檻值，則達成分群終止，最後完成論文所提之有效多語音素集，其中收斂條件為：

$$(\sum_{y=1}^{Y}\Delta_y^t - \sum_{y=1}^{Y}\Delta_y^{t-1}) / \sum_{y=1}^{Y}\Delta_y^{t-1} < \theta \quad \text{(式 8)}$$

其中，Δ_y^t 表示在第 t 次遞迴中，Y 群集中第 y 群之集合內個數分數值 $\Delta_y^t = \sum c(s_l, s_k)$，$t = 1,...,t_{\max}$ 表示運算遞迴次數，$t_{\max}$ 指設定之最大遞迴次數，θ 為收斂之門檻值。

4. 實驗評估

為了評估研究方法，論文提出幾項實驗驗證：首先，實驗單獨考慮聲學相似度、語言超空間相似度與本論文所提結合聲學與語言超空間相似度分析之方法，比較其辨識結果。再者，比較與前後文脈獨立之音素集和論文所提與前後文脈相關之音素集在多語辨識準確率的差別。

4.1. 多語語音語料分析

本論文使用的多語語音辨識訓練語料，台灣腔英文(English Across Taiwan, EAT)語料庫，其中包含英文長句,英文短句,英文單詞及中英夾雜句等[17]。從 2004 年 5 月開始收集，至 2005 年 1 月初步完成收集，由師大、交大、清大、成大和台大等五所學校參與語料之錄製收集，經工研院電通所彙整。分別由英語系及非英語系學生錄製，語料依性別做分類，錄製有麥克風語料及電話語料，歸納如下表所列：

表 4. EAT 語料麥克風音檔資料統計

MIC 16khz 16bits 語料				
	英語系		非英語系	
	男性	女性	男性	女性
句數	11,977	30,094	25,432	15,540
人數	166	406	368	224

麥克風語料錄製 16KHz 取樣頻率 16bits 的取樣點音檔，電話語料錄製 8KHz 取樣頻率 16bits 的取樣點音檔，

其中電話語料又可細分為固定式電話(PSTN)語料及行動電話(GSM)語料，電話語料部份是透過 Dialogic 電話語音介面卡，錄得的 8KHz，8Bits，Mulaw 格式的取樣點,經程式轉成 8KHz，16bits，pcm 格式的取樣點；麥克風語料是由個人電腦及麥克風，直接從個人電腦的音效卡錄製 16KHz，16bits 的聲音訊號。最後將所有取樣點以 wav 格式音檔儲存。本論文研究採用麥克風語料部分。

每位語者收錄 80 句語音語料，語料內容設計有英文數字連續語音、英文字母連續語音、中英文混合句、英文單字、片語或句子等，如(表 5)所示。論文主要探討中英夾雜的多語應用，實驗抽取語料內中英文混合句型 (表 5 之 6 和 7)，語料編號#58 至#70 的音檔資訊。

表 5. EAT 語料中多語句型範例

	EAT 語料句型
1	four eight three zero one two nine
2	for instance
3	Safe
4	Silicon Graphics
5	R. S. R. T. E. K.
6	**冠軍家庭 T.V.秀入圍金鐘獎**
7	**幫我查一下 Bryan 的分機**
8	The vote at the September meeting was eleven zero

原本音檔內容皆屬於 raw 格式，因此我們事先對音檔作 dc-offset 及 silence removal 的處理。並且根據英文發音辭典與中文發音辭典，將文字註解轉成音素標記。由於語料內有部份音檔及錄音品質不良，實驗以人工的方式先行校對。最後，論文所採用之實驗語料包含有訓練用中英文混合句型共有 2,018 句，實驗評估測試共有 100 句。

4.2. 音素為基準之自動語音辨識架構

為了評估音素定義的好壞，本論文使用自行開發的多語音素辨識系統，探討多語語音辨識。採用上述之實驗語料中，我們利用 IPA 音素標準定義，找出多語音素之聯集。定義三連音素模型共 N=997 個，訓練語料少於 5 次的三連音素不予考慮。在語音參數擷取的部份，對於輸入的語音訊號計算 26 維的梅爾倒頻譜參數(mel-frequency ceptral coefficient, MFCC)，其中包含 12 階的梅爾倒頻譜參數，加上 12 階的一次微分梅爾倒頻譜參數，以及一階的能量和其一次微分參數，並且對參數做 MVA [18]處理以增加辨識的強健性。

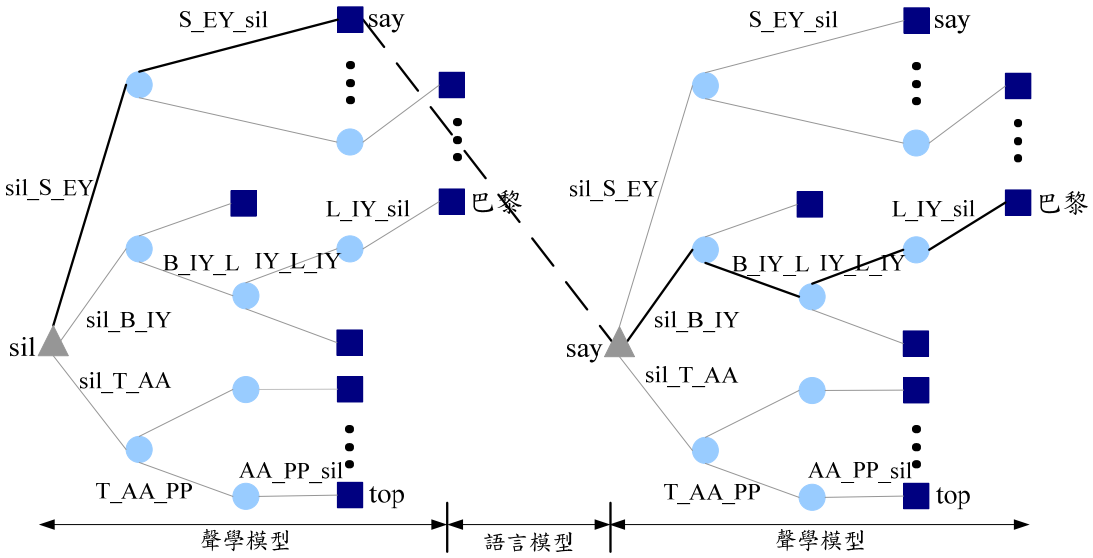


圖 4. 利用樹狀結構發音辭典文法樹於多語語音辨識之架構圖

本論文調整一般語音辨識使用的語言模型(language model)，在計算上利用均等機率(equal probability)的方法[19]，確保可以真正呈現不同多語音素定義的聲學模型(acoustic model)，對多語語音辨識的影響。在多語語音音素辨識，需要依據定義的多語音素結合各個目標語言的發音辭典，建構出一個多語發音辭典。

透過多語發音辭典，可以建構出多語發音之文法樹(grammar tree) [20]。如下(圖 4)所示。在辨識的流程上，每一個分支(arc)表示多語音素定義的 HMM 模型，研究上應用 3 個狀態(state)來描述每一個 HMM 模型，每一個狀態包含有 16 個高斯(mixture)。此多語之樹狀結構發音辭典舉例共有：say(sil_S_EY, S_EY_sil)、巴黎(sil_B_IY, B_IY_L, B_L_IY, L_IY_sil)、top(sil_T_AA, T_AA_P, AA_P_sil)等詞組。本實驗合併英文發音辭典與中文發音辭典，建立包含 29,104 個中英文詞之多語發音辭典。本圖示舉例說明由靜音(silence, sil)為起點，辨識多語語句"**say** (sil_S_EY, S_EY_sil) **巴黎** (sil_B_IY, B_IY_L, IY_L_IY, L_IY_sil)"為例，▲為樹的根節點；──線條表示多語音素也就是訓練的聲學模型，●指音素的節點；■表示葉結點，指出從根節點到此葉結點之發聲音素可能構成的所有多語詞彙；— —表示樹與樹之間連結的語言模型。

4.3. 利用聲學與語言超空間相似度分析群聚三連音素模型

語音辨識可能發生的錯誤有三種型態，分別是插入錯誤(insertion)、刪除錯誤(deletion)以及替換錯誤(substitution)。實驗中音素正確率(accuracy)的計算[21]，方式如下：

$$Accuracy = \frac{len - ins - del - sub}{len} \times 100\% \qquad \text{(式 9)}$$

其中，*len* 為辨識結果，音素序列的長度。*ins* 為比較較正確結果多辨識出的音素，屬於插入錯誤，*del* 為比較正確結果少辨識到的音素，屬於刪除錯誤。*sub* 為比較正確結果辨識錯誤的音素，屬於替換錯誤。分析不同群集條件下的群聚音素個數，利用調整 k 群聚(modified k-means, MKM)分類方法[16]，群聚三連音素模型為有效多語辨識模型。實驗聲學相似度計算(ACL)、語言超空間相似度計算(HAL)及資料融合技術(FUN)等不同方法，在收斂門檻值為 $\theta = 0.01$ 的情況下，實驗不同最大群集數 Y 。(表 6)實驗分析各種不同方法群集之多語音素個數及音素辨識的正確率，如下所示：

表 6. 不同群集數目限制條件下群聚音素個數及辨識正確率 (Y :最大群集數目, $\theta = 0.01$)

	$Y = 8$		$Y = 16$		$Y = 32$	
	正確率	音素個數	正確率	音素個數	正確率	音素個數
ACL	62.22%	161	63.12%	288	64.37%	531
HAL	62.52%	159	64.23%	286	64.57%	530
FUN	64.44%	119	66.07%	260	64.74%	515

實驗考慮聲學相似度矩陣分數計算，利用聲學相似度群聚方法 ACL，在 $Y = 8, 16, 32$ 的情況下，分別可以群聚為 161，288 及 531 個多語音素模型，其音素辨識正確率分別為 62.22%，63.12%及 64.37%。利用語言超空間分析方法 HAL，在 $Y = 8, 16, 32$ 的情況下，分別可以群聚為 159，286 及 530 個多語音素模型，其音素辨識正確率分別為 62.52%，64.23%及 64.57%。利用資料融合方法 FUN，在 $Y = 8, 16, 32$ 的情況下，分別可以群聚為 159，286 及 530 個多語音素模型，其音素辨識正確率分別為 64.44%，66.07%及 64.74%。

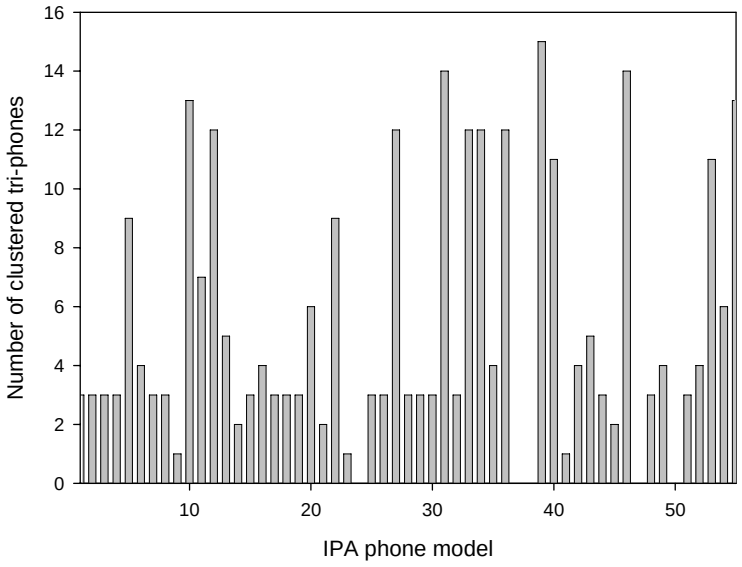


圖 5. 利用聲學相似度及前後文脈分析群集三連音素模型分佈圖，$Y = 16$

利用前後文脈分析方法 HAL 比聲學相似度方法 ACL 有較高的準確率，而同時結合聲學相似度與前後文脈分析 FUN 可以有最佳的辨識效果。當群集數 $Y=16$ 時，論文所提之方法(FUN)可以有最好的辨識效果。因此，論文設定群集分析 $Y=16$，群集三連音素模型，分析如(圖 5)所示。經過分類完後，各個 IPA 定義之音素中，所包含之三連音個數。由上圖可知，以 55 個 IPA 標準定義所產生之 997 個三連音素模型，利用資料融合方法可以合併為 260 個多語音素模型。

4.4. 聲學與語言超空間相似度分析於多語語音辨識

本論文研究探討中文和英文的多語語音辨識應用，實驗首先測試使用單音素模型(monophone)的定義，依據(表 1)和(表 2)等不同標記方法的內容，分別可以定義： (一) 直接結合個別單一語言之音素(MIX)；(二) 以 IPA 為基準定義多語音素之方法 (IPA)。實驗結果如(表 7)所示：

表 7. 單音素模型之多語辨識音素正確率 (括弧內表辨識單元之個數)

========== English Across Taiwan, EAT ==========				
----------Monophone Tree-Search Results----------				
	ACCURACY	INSERTION	DELETION	SUBSTITUTION
IPA phone sets (78)	**55.35%**	13.54%	5.27%	28.92%
Mix phone sets (55)	**55.49%**	21.94%	4.98%	18.06%
========== English Across Taiwan, EAT ==========				

實驗 EAT 語料中，有關中英文混合句型的連續語音分析。定義中文基本音素共有 37 個，英文基本的音素定義有 39 個。由此實驗結果可知，使用直接合併多語音素(MIX) 的辨識正確率為 52.35%，而採用 IPA 標準定義的中英文音素集，正確率為 55.49%。因此具有多語模型參數分享特性的 IPA 標準定義，辨識結果比較直接合併多語音素的方法，在辨識效果上來得顯著。再者，實驗比較原本以 IPA 為基準之三連音素模型定義(triphone sets)，利用語言超空間相似度矩陣群集音素定義(HAL phone sets)，利用聲學相似度矩陣群集音素定義(ACL phone sets)，以及利用資料融合方法於聲學及語言超空間相似度矩陣分析之音素定義(FUN phone sets)，如(表 8)所示：

表 8. 三連音素模型之多語辨識音素正確率 (括弧內表辨識單元之個數)

========== English Across Taiwan, EAT ==========				
----------Triphone Tree-Search Results----------				
	ACCURACY	INSERTION	DELETION	SUBSTITUTION
Triphone sets (997)	**68.07%**	15.87%	4.43%	11.63%
ACL phone sets (288)	**63.12%**	19.73%	4.88%	12.32%
HAL phone sets (286)	**64.23%**	20.67%	4.75%	10.48%
FUN phone sets (260)	**66.07%**	16.94%	4.41%	12.71%
========== English Across Taiwan, EAT ==========				

由實驗結果可知，合併前的三連音素模型(Triphone sets)的多語辨識效果可達 68.07%的正確率。利用聲學相似度矩陣群集音素定義(ACL phone sets) ，在多語辨識效果上可達 63.12%的正確率，而利用語言超空間相似度矩陣群集音素定義(HAL phone sets) ，在中英文多語辨識效果上可達 64.23%的正確率。進一步利用資料融合方法於聲學及語言超空間相似度矩陣群集分析之音素定義(FUN phone sets) ，在多語辨識效果可以提升至 66.07%的正確率。整體而言，採用三連音素的辨識效果比單音素(IPA 或 MIX)定義好。又從語言分析(HAL)效果會較聲學分析(ACL)效果來得顯著，且利用資料融合方法結合聲學相似度及前後文脈分析，對於多語語音辨識可以有明顯的提升。

5. 結論及未來展望

本論文提出應用聲學相似度及前後文脈分析於多語語音辨識之有效音素定義，以 EAT 中英文雙語語料為例。基於 IPA 標準定義之多語單音素集，本研究考慮以發音前後文相依三連音素模型。以此定義，我們分別以聲學相似度及前後文脈分析，音素間相似度高的音素合併，期望找出精簡有效的多語語音辨識音素集。利用音素 HMM 模型，以直接校準方法切音並計算事後機率值，建立聲學相似度矩陣。利用語言超空間相似度分析(hyperspace analog to language, HAL)，找出音素前後發音特性所造成的變音影響，建立語言發音上相似度矩陣。之後，以資料融合方法，同時考慮聲學和語言超空間相似度矩陣。利用向量量化群集分析，找出同一類別之音素定義，建立有效而精簡的多語音素集。實驗證明利用結合聲學和語言超空間相似度矩陣分析方法，可以達到良好的多語連續語音辨識的效果。未來可以將方法應用在單一語言語音辨

識之音素定義研究，也可以進一步分析在更多目標語言下，此方法對於多語語音辨識的效果表現。

參考文獻

[1] Chung-Hsien Wu, Yu-Hsien Chiu, Chi-Jiun Shia, and Chun-Yu Lin, 2006. Automatic Segmentation and Identification of Mixed-Language Speech Using Delta-BIC and LSA-Based GMMs. IEEE Transactions on audio, speech, and language processing, vol. 14, no. 1, pp. 266-276.

[2] Alex Waibel, Hagen Soltau, Tanja Schultz, Thomas Schaaf, and Florian Metze, 2000. Multilingual Speech Recognition. Chapter in Verbmobil: Foundations of Speech-to-Speech Translation, Springer-Verlag.

[3] Rafid A. Sukkar and Chin-Hui Lee, 1996. Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword based Speech Recognition. IEEE Transactions on Speech and Audio Processing, vol. 4, no. 6, pp. 420-429.

[4] Yeou-Jiunn Chen, Chung-Hsien Wu, Yu-Hsien Chiu, and Hsiang-Chuan Liao, 2002. Generation of robust phonetic set and decision tree for Mandarin using chi-square testing. Speech Communication, vol. 38(3-4), pp. 349-364.

[5] Mathews, R. H., 1975. Mathews' Chinese-English Dictionary, Caves, 13th printing.

[6] J. C. Wells, 1989. Computer-Coded Phonemic Notation of Individual Languages of the European Community. J. IPA, 19, pp. 32-54.

[7] James L. Hieronymus, 1993. ASCII Phonetic Symbols for the World's Languages: Worldbet. Journal of the International Phonetic Association.

[8] Brian Mak and Etienne Barnard, 1996. Phone clustering using the Bhattacharyya distance. in Proc. ICSLP, pp. 2005-2008.

[9] Jacob Goldberger and Hagai Aronowitz, 2005. A Distance Measure Between GMMs Based on the Unsented Transform and its Application to Speaker Recognition. in Proc. of EUROSPEECH 2005, pp. 1985-1988, Lisbon, Portugal.

[10] Chomsky, N. and Halle, M., 1968. The Sound Pattern of English. New York: Harper & Row.

[11] Burgess, C. and Lund, K., 1997. Modelling parsing constraints with high-dimensional context space. Language and Cognitive Processes, 12:177–210.

[12] Robert M. Gray and David L. Neuhoff, 1998. Quantization. IEEE Transactions on Information Theory, vol. 44, no. 6, pp. 2325-2383.

[13] G. Salton and C. Buckley, 1988. Term-weighting Approaches in Automatic Text Retrieval. Information Processing Management, vol. 24, no. 5, pp. 513-523.

[14] Josef Kittler, Mohamad Hatef, Robert P.W. Duin, and Jiri MatasOn, 1998. On Combining Classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp. 226-239.

[15] Jerome R. Bellegarda, 2000. Exploiting latent semantic information in statistical language modeling. Proc. IEEE, vol. 88, no. 8, pp. 1279-1296.

[16] Jay G. Wilpon and Lawrence R. Rabiner, 1985. A modified K-means clustering algorithm for use in isolated work recognition. IEEE Transactions on Acoustics, Speech, and Signal Proc., vol. 33, no. 3, pp. 587-594.

[17] English Across Taiwan, EAT [online] http://www.aclclp.org.tw/

[18] Chia-Ping Chen, Jeff Bilmes and Daniel P. W. Ellis, 2005. Speech feature smoothing for robust ASR. in Proc. ICASSP, Philadelphia PA.

[19] Johnston, D., 1997. Statistical Methods for Speech Recognition. The MIT Press, Cambridge, MA.

[20] H. Ney and S. Ortmanns, 2000. Progress in dynamic programming search for LVCSR. Proceedings of the IEEE, vol. 88, no. 8, pp. 1224–1240.

[21] Steve Young, Gunnar Evermann, Mark Gales, Tomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, Phil Woodland, The HTK Book (for HTK Version 3.3).

統計圖等化法於雜訊語音辨識之進一步研究

林士翔[1]　葉耀明[2]　陳柏琳[2]

[1]國立台灣師範大學資訊教育研究所

[2]國立台灣師範大學資訊工程研究所

69308027@cc.ntnu.edu.tw, ymyeh@ice.ntnu.edu.tw, berlin@csie.ntnu.edu.tw

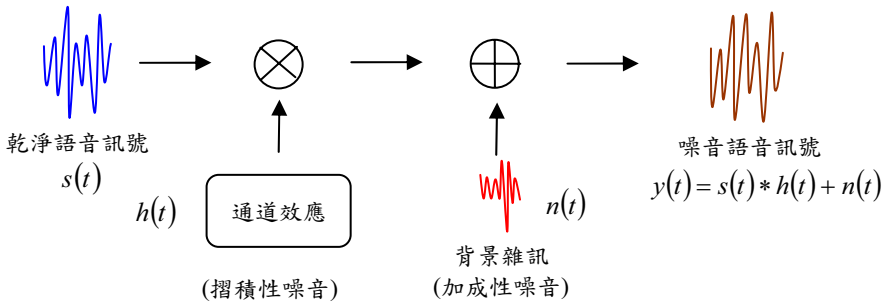
摘要

自動語音辨識系統通常會因語音訊號是否受各種環境雜訊干擾而產生某種程度上的影響。正因如此，語音強健(Speech Robustness)技術的發展長久以來一直被視為一個非常重要的研究領域，過去已有許多方法成功地被提出，可以在抗雜訊上有不錯的效果。其中，統計圖等化法(Histogram Equalization)能有效地補償語音訊號受噪音干擾而所產生的失真情形，因而被公認為非常有效果的方法之一。但前人所提出的統計圖等化法，往往需要大量的記憶體使用空間或是處理器運算時間，本論文探討利用數據擬合(Data Fitting)方法創造一逆函數(Inverse Function)，有效且快速地將測試語句累積密度函數近似至參考分佈的累積密度函數，以正規化雜訊語音特徵，藉由逆函數的使用，能夠節省統計圖等化時所需要的記憶體使用空間以及處理器運算時間。同時，本論文亦探討數據擬合統計圖等化法與時間軸上特徵值移動平均(Moving Average)之結合，來減輕非穩性噪音(Non-stationary Noise)所造成的異常尖峰或波谷的影響。此外，本論文更進一步將所提出的數據擬合統計圖等化法與其他特徵擷取或補償方法進行整合，初步實驗結果證實本論文所提出之方法，能有效補償語音受雜訊干擾所造成的失真情形，進而有效提昇辨識效能。實驗語料庫為由歐洲電信標準協會所發行的AURORA-2 語料，實驗結果初步地顯示數據擬合統計圖等法化確實為一有效的語音強健技術。

1. 序論

現今自動語音辨識 (Automatic Speech Recognition, ASR)系統在語音訊號不受噪音干擾的理想實驗室環境下，可獲得良好的辨識效果，但若要應用至實際日常生活環境中，往往會因為環境中複雜因素的影響，造成訓練環境與測試環境存在環境不匹配(Environment Mismatch)的差異，使得系統辨識效能大幅度降低，環境中複雜因素包括背景噪音(Background Noise)、錄音設備本身產生的噪音或是通道效應(Channel Effect)等。正因如此，語音強健(Speech Robustness)技術長久以來一直被視為重要的研究課題，主要是希望藉由對訊號本身、語音特徵參數或是模型參數做適當的處理與調整，以減緩雜訊干擾的影響、降低訓練環境與測試環境不匹配的情形或提升語音訊號或語音特徵參數本身的強健性，進而提高系統辨識效能。

環境中干擾語音訊號的雜訊可概略分為二種類型：(1)加成性噪音(Additive Noise)和(2)摺積性噪音(Convolutional Noise)。加成性噪音為錄製語音時，原始語音與背景噪音以線性加成(Linearly Additive)的關係同時被收錄進去，例如周遭人聊天的聲音或是機器設備所發出的噪音等；摺積性噪音通常是指語音訊號在經由不同傳輸通道時所產生的通道效應，例如電話線路通道



圖一、雜訊干擾示意圖

效應、麥克風通道效應等。加成性噪音與摺積性噪音對於語音訊號的干擾過程示意圖如圖一所示。

語音強健技術的主要目的就是為了消除不同環境下的差異性以及減輕雜訊對語音訊號的影響，過去已有許多方法成功地被提出，依據方法的本質可概分為以下三種方向[1]：

(1) 語音強化技術(Speech Enhancement)

目的在於提升語音訊號本身的品質，通常是假設語音訊號與雜訊訊號二者在統計上是不相關(Uncorrelated)，希望能由觀察到的雜訊語音(Noisy Speech)重建出乾淨語音(Clean Speech)訊號。常見的技術有頻譜消去法(Spectral Subtraction, SS)[2]、維爾濾波器(Wiener Filter, WF)[3]等。

(2) 強健性語音特徵(Robust Speech Feature)

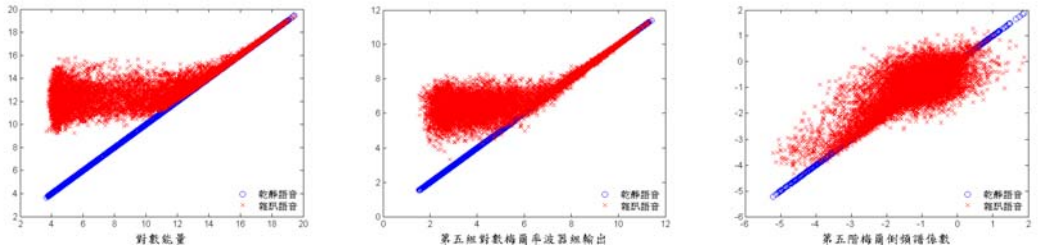
從語音訊號中擷取出較不易受到環境變化干擾而失真的強健性語音特徵參數。常見的技術有倒頻譜平均消去法(Cepstrum Mean Subtraction, CMS)[4]、倒頻譜正規化法(Cepstrum Mean and Variance Normalization, CMVN)[5]等。

(3) 聲學模型調適技術(Acoustic Model Adaptation)

藉由少量的調適語料(Adaptation Data)調整由乾淨語音所訓練而成的聲學模型中的機率分佈參數，如平均值向量(Mean Vector)或共變異矩陣(Covariance Matrix)，期望調適後的模型可以適用於新的環境，以降低環境不匹配的現象。常見的技術有最大事後機率法則(Maximum a Posteriori, MAP)[6]、最大相似度線性回歸法(Maximum Likelihood Linear Regression, MLLR)[7]等。

本論文中將探討的方法是屬於上述第二類強健性語音特徵。目前，倒頻譜平均消去法(CMS)和倒頻譜正規化法(CMVN)已被廣泛的應用且也被成功地證實能有效的提升辨識效果，其分別是針對語音特徵參數第一階動差(Moment)或是第一階動差與第二階動差進行正規化。但因方法本身線性關係的限制，造成只能補償因受噪音干擾所產生的線性失真部份，對於非線性失真部份的補償效果有限。因此許多學者嘗試提出許多不同的補償方法，試圖解決因噪音干擾影響對語音特徵參數所產生的失真情形。例如[8]針對語音特徵參數的第三階動差進行正規化或[9]對語音特徵參數更高階動差進行正規化。此外，近年來亦有學者嘗試將在影像處理中已行之有年的統計圖等化法(Histogram Equalization)應用於語音辨識[10]。

統計圖等化法除了試圖去匹配訓練語料與測試語料之語音特徵參數的平均數和變異數之外，更企圖使訓練語料和測試語料能夠具有相同的統計分佈特性，其作法是藉由將測試語料(Test Speech)的累積密度函數(Cumulative Density Function, CDF) 對應至由訓練語料(Training Speech)所統計出來的參考分佈(Reference Distribution)的累積密度函數，藉由此匹配轉換過程，降低測試



圖二、加成性噪音對語音特徵參數的影響

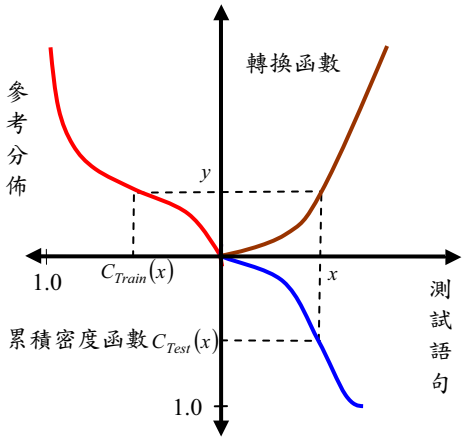
語料與訓練語料由於環境因素影響所造成統計特性不匹配的現象，實驗結果亦證實統計圖等法化對提升辨識效果有很明顯的幫助。另外在[11][12]中，更嘗試將統計圖等化法概念推廣至向量量化編碼(Vector Quantization)，進而應用於分散式語音辨識 (Distributed Speech Recognition, DSR)上，主要是利用統計圖資訊做為向量量化準則，有效解決傳統以距離為量化準則容易受環境噪音影響或是容易形成量化失真(Quantization Distortion)的問題。

雖然統計圖等化法近年來已被廣泛的應用於討論，但仍然有許多可以改善的地方，例如查表式統計圖等化法(Table Look-Up based Histogram Equalization, THEQ)[10]需要將龐大的表格資訊載入記憶體中，方能進行轉換匹配動作，且若要有良好的補償效果，表格所紀錄的點數不能太少，但當表格紀錄點數增加時，同時意謂著需耗費更大量的記憶體使用空間與進行查表轉換的處理器運算時間；又如分位差統計圖等化法(Quantile-based Histogram Equalization, QHEQ)[13,14]，雖然轉換過程不需透過查表動作，只需使用少量的參數即可進行等化動作，但是對每一句待轉換的語句在進行轉換動作前，必須利用格式搜尋(Grid Search)以線上即時運算求取參數，因此所需的處理器運算時間也是相當可觀的。

基於上述原因，本論文提出利用數據擬合(Data Fitting)的概念求算累積密度函數的逆函數，藉由少量的多項式係數與多項式函數的運用，達到具有和統計圖等化法相同的補償效果，同時也探討時間序列上特徵值移動平均法的使用，解決因等化過程中某些特徵值被異常的放大或縮小。本論文後續章節安排如下：第二章將介紹傳統查表式統計圖等化法與分位差統計圖等化法；第三章則介紹數據擬合統計圖等化法及時間序列上特徵值移動平均法的使用；第四章為實驗與討論；第五章結論。

2. 文獻回顧

雜訊干擾會使得語音訊號產生非線性失真，例如加成性噪音對訊號的對數能量影響，在高能量的區域，只有輕微的影響，相反地，在能量強度較低的區域則會有嚴重的失真情形，此一情形，即為造成乾淨語音訊號和雜訊語音訊號二者間統計特性差異的主要原因之一。雜訊干擾對於乾淨語音所造成的非線性失真情形如圖二所示，其中藍色散佈點數的描繪是利用乾淨語音的特徵值當做X軸參考座標值與Y軸參考座標值；紅色散佈點數的描繪是以乾淨語音的特徵值當做X軸參考座標值，雜訊語音的特徵值為Y軸參考座標值，圖片從左至右分別是作用在對數能量(Log Energy)、對數梅爾濾波器組(Mel Filter-Bank)輸出以及梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficient)。由圖中可觀察發現傳統的補償方法諸如倒頻譜平均消去法或倒頻譜正規化法因本身線性特性所限制，可能會使得對於雜訊干擾所造成的非線性失真部份補償效果非常有限。因此，過去幾年有許多學者提出各種方法來補償此非線性失真的情形，其中統計圖等化法為有非常顯著



圖三、統計圖等化法示意圖

補償效果的方法之一[10][15]，下面章節將分述傳統查表式統計圖等化法(THEQ)與分位差統計圖等化法(QHEQ)的概念並分析各方法的優點和缺點。

2.1 統計圖等化法(Histogram Equalization, HEQ)

統計圖等化法假設測試語句之語音特徵參數的統計分佈會和訓練語料語音特徵參數的統計分佈(或稱作參考分佈)是一致的，若以目前較常用的語音特徵參數－梅爾倒頻譜係數(Mel- Frequency Cepstral Coefficients, MFCC)而言，統計圖等化法可以作用在對數梅爾濾波器組輸出[16][17][18]或是梅爾倒頻譜係數[10][19][20]上。統計圖等化法最主要精神可以視為是利用一個轉換函數(Transformation Function)，此函數能將測試語句的語音特徵向量每一維的統計分佈分別轉換至先前已從訓練語句中定義好的對應參考分佈，數學式關係式表示如下[19][20]：假設 x 為測試語句語音特徵向量的某一維特徵參數，且具有機率密度函數(Probability Density Function, PDF) $p_{Test}(x)$，那麼轉換函數 $F(x)$ 可依照下列的數學式將 x 轉換成在訓練語料所對應到的 y，並且讓 $p_{Train}(y)$ 與 $p_{Test}(x)$ 能有式(1)的關係：

$$p_{Train}(y)=p_{Test}(x)\frac{dx}{dy}=p_{Test}\left(F^{-1}(y)\right)\frac{d\left(F^{-1}(y)\right)}{dy} \tag{1}$$

其中 $F^{-1}(y)$ 為 $F(x)$ 的逆函數(Inverse Function)，若上述關係式以累積機率密度函數(Cumulative Probability Density Function, CDF)的觀點表達即為

$$\begin{aligned} C_{Test}(x) &= \int_{-\infty}^{x} p_{Test}(x')dx' \\ &= \int_{-\infty}^{F(x)} p_{Test}\left(F^{-1}(y')\right)\frac{dF^{-1}(y')}{dy'}dy' \\ &= \int_{-\infty}^{y} p_{Train}(y')dy'|_{y=F(x)} \\ &= C_{Train}(y) \end{aligned} \tag{2}$$

其中 $C_{Test}(x)$ 和 $C_{Train}(y)$ 分別為測試語句和訓練語料的累積機率密度函數，y' 為經由轉換函數 $F(x')$ 求得的結果，所以轉換函數 $F(x)$ 會具有下列特性

$$F(x)=C_{Train}^{-1}(C_{Test}(x)) \tag{3}$$

其中 C_{Train}^{-1} 為 C_{Train} 的逆函數，轉換過程如圖三所意示。

在實作上，因為語音特徵參數為一有限集合，所以並無法非常精準估算實際的累積密度函

數，通常實作會使用累積統計圖(Cumulative Histogram)去近似累積密度函數。對於所有訓練語料而言，語音特徵向量的每一維會統計出一個累積統計圖，再依需求將累積統計圖設定為 i 個分位差(Quantiles)，每個分位差區間皆以區間內所有特徵值的平均數(Mean)做為該分位差的代表特徵值，此資訊可被用來當作轉換的參考分佈。對測試語句語音特徵向量的每一維度同樣統計出累積統計圖，也取 i 個分位差，接著每個分位差區間內的特徵值用先前建立好的參考分佈逐一進行轉換取代。一般實作可利用表格查詢的方式進行：首先，以表格方式紀錄參考分佈的累積統計圖資訊，例如記錄成｛區間，特徵值｝；接著，在進行等化(Equalization)過程時，將所有表格載入記憶體中以方便進行查表(Table-Lookup)轉換，故可稱為查表式統計圖等化法(Table Look-Up based Histogram Equalization, THEQ)。但是往往為了要得到良好的辨識效果，使用的分位差個數不可太少，即代表需耗費大量的記憶體空間紀錄表格資訊；並且在進行查表轉換時，也需花費不少的表格搜尋時間。

2.2 分位差統計圖等化法(Quantile-based Histogram equalization, QHEQ)

前一章節所介紹的查表式統計圖等法化為一種非參數(Nonparametric)型態的統計圖等化法，所有的等化動作都是直接根據測試語句的累積統計圖進行，並無需使用任何額外參數，在[19][20]中提出一種參數型態的分位差統計圖等化法，對於語音特徵向量的每一維可利用一轉換函數 $H(x)$ 進行等化動作，數學關係式表示如下

$$H(x)=Q_K\left(\alpha\left(\frac{x}{Q_K}\right)^{\gamma}+(1-\alpha)\left(\frac{x}{Q_K}\right)\right) \tag{4}$$

Q_K 為最後一個分位差值，亦即整句語句中在此一維特徵參數中最大的特徵值；α 和 γ 為轉換函數 $H(x)$ 所需的參數可利用式(5)求得，值得注意的是在對於每一句語句在進行等化過程前，需先將整句語句與參考分佈進行分位差校正(Quantile Correction)，以求得最佳的參數，此校正動作是以最小平方誤差(Minimum Mean Square Error, MMSE)法進行，可以利用格式搜尋法，將一段區間內的 α 和 γ 以等距的數值代入式(5)，找出最佳的 α 和 γ。

$$\{\alpha,\gamma\}=\arg\min_{\{\alpha,\gamma\}}\left(\sum_{k=1}^{K-1}\left(H(Q_k)-Q_k^{train}\right)^2\right) \tag{5}$$

其中 K 為分位差的個數；Q_k 為待轉換語句中第 k 個分位差的特徵值；Q_k^{train} 為訓練語料所統計出的參考分佈中的第 k 個分位差值。

分位差統計圖等化法是經由式(5)計算以求得參數 α 和 γ，接著再利用式(4)一組非線性函數和一組線性函數進行加權平均(Weight Average)，欲使轉換後的語音特徵參數的統計分佈能夠和參考分佈愈相似，亦即受雜訊干擾而形成的非線性失真部份，可藉由 γ 項的使用進行補償。但針對每一轉換語句都必經由式(5)線上即時求得最佳的參數 α 和 γ，因此必須需耗費不少的處理器運算時間利用格式搜尋法做完整的搜尋。

3.統計圖等化法之改良

前面章節所描述的統計圖等化法雖然能有效補償因雜訊干擾而所產生的非線性失真情形，但無論是傳統查表式統計圖等化法或是分位差統計圖等化法，往往在執行等化的過程，需耗費大量的記

憶體使用空間或是處理器運算時間。為了能有效的解決此問題，我們提出利用數據擬合的概念求得累積密度函數的逆函數，藉由少量的多項式函數與多項式係數的使用，達到具有和統計圖等化法相同的補償效果[21]，整體概念和作法敘述如下。

3.1 多項式擬合統計圖等化法 (Polynomial-Fit Histogram Equalization, PHEQ)

當給定一些資料點數(u_i, v_i)，若要以一個函數來描述反應變數(Response Variable) v_i 與解釋變數(Explanatory Variable) u_i 關係，通常可使用迴歸模型(Regression Model)來表示，換句話說迴歸模型可用來解釋在給定 u_i 的情況下，預測 v_i 的可能值為何。通常迴歸公式 $G(u_i)$ 可依係數(Coefficients)組合不同而表示成線性或非線性型式，並且 $G(u_i)$ 係數的選擇影響預測值 $\tilde{v}_i$ 的準確性甚鉅，一般可利用最小誤差平方和 (Minimum Sum of Squares Error)求得，換言之將所有 u_i 分別代入迴歸公式所求得的預測值 $\tilde{v}_i$ 和實際觀測值 v_i 的誤差值平方和必須最小，意謂著經由迴歸模型所預測出的值會跟實際的值較相似，此法又可稱最小平方迴歸法(Least Squares Regression)。假設 $G(u_i)$ 為 M 階的線性多項式函數：

$$\tilde{v}_i = G(u_i) = a_0 + a_1 u_i + a_2 u_i^2 + \cdots + a_M u_i^M = \sum_{m=0}^{M} a_m u_i^m \tag{6}$$

$a_0, a_1, \ldots, a_M$ 為多項式的係數(Coefficients)，則所對應的誤差平方合 E^2 定義成

$$E^2 = \sum_{i=1}^{N} \left(v_i - \sum_{m=0}^{M} a_m u_i^m \right)^2 \tag{7}$$

同理，此概念可延伸為求得參考分佈之累積密度函數的逆函數，我們稱為多項式擬合統計圖等化法 (Polynomial-Fit Histogram Equalization, PHEQ)。對於訓練語料語音特徵向量的每一維皆可利用一個多項式迴歸表示，以特徵值 y_i 為反應變數以及 y_i 對應的累積密度值 $C_{Train}(y_i)$ 為解釋變數，則式(7)可重新表示成

$$G(C_{Train}(y_i)) = \tilde{y}_i = \sum_{m=0}^{M} a_m (C_{Train}(y_i))^m \tag{8}$$

並且誤差平方合 E'^2 定義為

$$E'^2 = \sum_{i=1}^{N} \left(y_i - \sum_{m=0}^{M} a_m (C_{Train}(y_i))^m \right)^2 \tag{9}$$

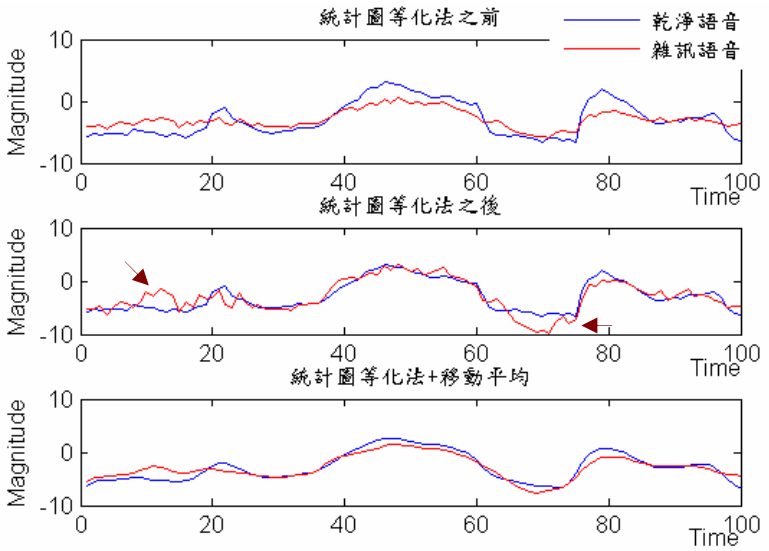
其中 N 為訓練語料中所有音框(Frame)的個數，若要使誤差平方合最小，則所有多項式係數 $a_0, a_1, \ldots, a_M$ 會滿足式(10)的關係，只需透過解聯立方程式，即可求得 $a_0, a_1, \ldots, a_M$ 係數。

$$\frac{\partial E'^2}{\partial a_m} = 0, \forall m = 1 \ldots M \tag{10}$$

在實作上，對轉換語句中語音特徵向量的每一維 $Y = [y_1, y_2, \ldots y_N]$ 而言，每個時間點上，特徵值所對應的累積密度值可利用下列步驟近似而得：

步驟一、將 Y 以遞增的方式做排序得到資料序列 S

步驟二、對於 Y 中的每個特徵值 y_i 而言，可由下式近似其所對應的累積機率密度函數值



圖四、非穩性噪音所造成的異常尖峰或波谷示意圖

$$C(y_i) \approx \frac{S_{pos}(y_i)}{N} \tag{11}$$

其中 $S_{pos}(y_i)$ 為一指示函數，表示特徵值 y_i 在排序後的資料序列 S 中的位置，N 為資料序列中所有音框個數。在訓練階段時，可以利用式(11)求得所有訓練語料的累積密度函數，再利用式(8)、(9)和(10)求得累積密度函數的逆函數的係數；在辨識階段，只需要將測試語句語音特徵向量中的每一維特徵值 y_i 的對應累積密度函數值 $C(y_i)$ 代入先前已於訓練階段中求得的多項式函數(式(8))即可進行等化動作。

因此，本論文所提出數據擬合的使用，能有效地解決傳統統計圖等化法或分位差統計圖等化法需耗費的大量記憶體資源與處理器運算時間的缺點，只需透過少量的多項式係數與多項式函數的運用，便能迅速的將測試語句語音特徵向量每一維的統計分佈轉換至先前已從訓練語句中定義好的參考分佈，並且能擁有和統計圖等化法相同的補償效果。

3.2 時間序列上特徵值移動平均(Moving Average, MA)

雖然統計圖等化法對於補償因雜訊干擾所產生的非線性失真有顯著效果，但值得注意的是，由非穩性噪音(Non-stationary Noise)所造成的異常尖峰(Sharp Peak)或波谷(Valley)，可能造成在等化的過程中，某些特徵值被過度的放大或縮小，此異常情形如圖四所示，最上方的圖為尚未做統計圖等化法前乾淨語音與雜訊語音倒譜頻特徵向量的第二維；中間的圖為做完統計圖等化法後的倒譜頻特徵向量的第二維特徵值，可清楚看見有二個區域被過度強調；最下方的圖為做完統計圖等化法加移動平均後的倒譜頻特徵向量。

移動平均在語音辨識的研究上，已非一個全新的議題，例如[22]利用移動平均的概念提出一種特徵向量正規化(Feature Normalization)的方法，首先對語音特徵向量進行平均消去法(Mean Subtraction)和變異數正規化(Variance Normalization)，接著再利用自動迴歸移動平均(Auto-Regression Moving Average, ARMA)對特徵向量進行正規化的動作，其實驗結果亦證實移動平均的使用對於提升整體辨識率有很大的幫助。依照移動平均所考慮語音特徵來源與時間軸點數不同，可以有下列數種選擇[22]：

- 非因果關係移動平均(Non-Casual Moving Average)

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=-L}^{L} \tilde{y}_{(t+i)}}{2L+1} & if\ L < t \le T-L, \\ \tilde{y}_t & otherwise \end{cases} \tag{12}$$

- 因果關係自動迴歸移動平均(Casual Moving Average)

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=0}^{L} \tilde{y}_{(t-i)}}{L+1} & if\ L < t \le T, \\ \tilde{y}_t & otherwise \end{cases} \tag{13}$$

- 非因果關係自動迴歸移動平均(Non-Casual Auto Regression Moving Average)

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=1}^{L} \hat{y}_{(t-i)} + \sum_{j=0}^{L} \tilde{y}_{(t+j)}}{2L+1} & if\ L < t \le T-L, \\ \tilde{y}_t & otherwise \end{cases} \tag{14}$$

- 因果關係自動迴歸移動平均(Casual Auto Regression Moving Average)

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=1}^{L} \hat{y}_{(t-i)} + \sum_{j=0}^{L} \tilde{y}_{(t-j)}}{2L+1} & if\ L < t \le T, \\ \tilde{y}_t & otherwise \end{cases} \tag{15}$$

其中 $\tilde{y}_i$ 為輸入的語音特徵值，$\hat{y}_i$ 為經由移動平均法後所求得新的語音特徵值，L 表示移動平均項階數（Order of Moving Average）。

3.3 結合等化前與等化後的特徵值資訊

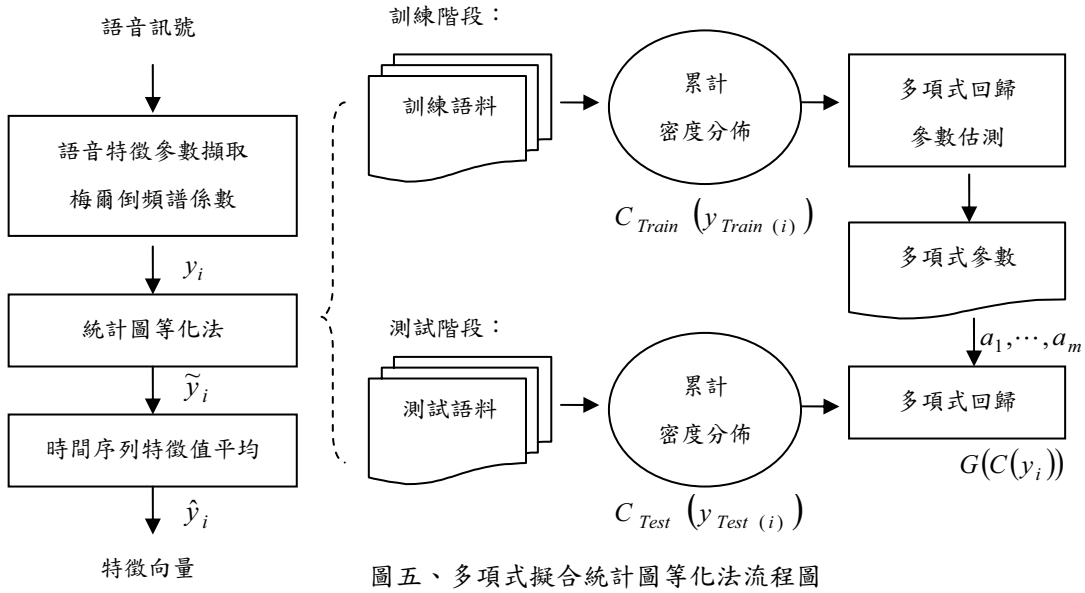
除了上述時間序列上進行特徵值移動平均的方法外，從實際雜訊語音特徵參數中，經由觀察發現等化前與等化後的特徵值，在特徵值受雜訊干擾不嚴重時，原本的特徵值與等化後的特徵值差異不大，相反地，特徵值可能會因等化動作被異常的放大或縮小，進而造成原本語音特徵所帶有資訊被破壞掉，因此可使用加權平均來補償此一異常情形，數學式如式(16)所示。當特徵值未受雜訊嚴重干擾時 $\bar{y}_i \approx \hat{y}_i$，相反地，可依造 α 的設定，決定 $\bar{y}_t$ 較保留等化前或等化後的資訊，此法概念與分位差統計統等法化[19][20]在概念上非常相似，藉由對加權數 α 值的控制，在線性與非線性補償中取得一平衡點，

$$\bar{y}_t = (1-\alpha) \times y_t + \alpha \times \hat{y}_t \tag{16}$$

y_i 為原本的特徵值，$\hat{y}_i$ 為做完等化後的特徵值。α 值的設定，除了可以利用窮舉的方法，逐一調動 α 值，以取得最佳的辨識效果外，亦可使用數據擬合的概念求算而得，誤差平方 E'' 定義如下：

$$E'' = \sum_{i=1}^{N} \left(\left(\bar{y}_{Noisy(i)} - \bar{y}_{Clean(i)} \right)^2 \right) \tag{17}$$

其中 $\bar{y}_{noisy(i)}$ 為雜訊語音第 i 個音框經由式(16)求得後的新的特徵值，$\bar{y}_{clean(i)}$ 為雜訊語音相對應的乾淨語音經由式(16)求得的特徵值，但值得注意的是，式(17)誤差平方計算程程中，為了避免整體誤差被部份異常的錯誤(Outlier)所支配，適當的利用門檻值(Threshold)將異常特徵值排除是須要的。



圖五、多項式擬合統計圖等化法流程圖

4. 實驗與討論

4.1 實驗架構與設定

本論文實驗所使用的語料庫 Aurora-2 是由歐洲電信標準協會(European Telecommunications Standards Institute, ESTI)所發行[23]，其本身為一套含有雜訊的連續英文數字語料，其中雜訊包含八種來源不同的加成性噪音和二種不同特性的通道。加成性噪音包括機場(Airport)、人聲(Babble)、汽車(Car)、展覽會館(Exhibition)、餐廳(Restaurant)、地下鐵(Subway)、街道(Street)及火車站(Train Station)，且依不同訊噪比(Signal-to-Noise Ratio, SNR)各自加入乾淨的語音裡，訊噪比包括 20dB、15dB、10dB、5dB、0dB 和-5dB；通道包含由國際電信聯合會所訂立的二個標準-G.712 和 MIRS。根據測試語料中加入之通道雜訊以及加成性雜訊之種類不同，Aurora-2 分為三組測試群組 Set A、Set B 和 Set C，Set A 所呈現的雜訊是屬於穩性(Stationary)雜訊，Set B 則是非穩性(Nonstationary)雜訊，Set C 除了穩性與非穩性雜訊外，還使用與訓練語料不同的通道效應。

在聲學模型(Acoustic Models)的設定，每個數字模型(1~9 及 zero 和 oh)皆由一個由左到右(left-to-right)形式的連續密度隱藏式馬可夫模型(Continuous Density Hidden Markov Model, CDHMM)表示，其中包含 16 個狀態(State)，並且每個狀態是利用 3 個高斯混合分佈(Gaussian Mixture Distribution)表示。另外靜音模型的部份有二種模型，一個為靜音(Silence)模型包含三個狀態，用來表示語句開始跟結束時的靜音；另一個為間歇(Pause)模型包含六個狀態，表示語句內字與字之間的短暫停止，上述所有聲學模型的訓練與本論文所有的實驗都是使用 HTK 工具套件[24]完成。

在前端處理方面(Front-End Processing)，本論文的基礎實驗是採用梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficients, MFCCs)作為語音特徵參數，取樣音框長度(Frame Length)為25毫秒，音框間距(Frame Shift)為10毫秒，每個音框的資訊是以39維表示，其中包含12維的梅爾倒頻譜係數以及一維的對數能量(Log Energy)，同時會對13維特徵參數取其相對的一階差量係數(Delta Coefficient)和二階差量係數(Acceleration Coefficient)。並且將所提出的多項式擬合統計圖等化法作用在梅爾倒頻譜係數與差量倒頻譜係數上，整體前端處理流程如圖五所示。

表一、多項式擬合統計圖等化法平均字錯誤率實驗結果

		多項式階數			
		3	5	7	9
乾淨語料訓練模式 (Clean-Condition)	所有訓練語料	22.39	21.54	21.08	21.30
	1000組	21.80	21.46	21.13	21.16
	100組	22.68	21.31	20.75	20.55
	10組	23.42	22.20	22.54	23.42
複合情境訓練模式 (Multi-Condition)	所有訓練語料	10.80	10.34	10.43	10.54
	1000組	10.48	10.32	10.40	10.45
	100組	10.73	10.45	10.36	10.45
	10組	11.65	10.61	10.79	11.58

4.1 多項式擬合統計圖等化法實驗

本文第一個實驗是利用多項式迴歸模型描述參考分佈的累積密度函數分佈情形，欲探討使用所有訓練語料與否以及不同多項式階數(Polynomial Order)對於整體辨識效能影響結果如何。其中參考分佈的資訊是由所有訓練語料統計而成的累積統計圖求得，其中累積統計圖所使用的分組組數(Histogram Bins)包括1000組、100組和10組，每一分組皆以組內所有特徵值的平均數做為該組代表特徵值；同時也使用不同階數(Order)的多項式進行等化動作。辨識結果如表一所示，表格內所呈現的數據皆為平均字錯誤率(Word Error Rate, WER)，是由Aurora-2中三組實驗群組(Sets A, B及C)中不同訊噪比(20dB至0dB)的辨識結果加總平均而得。

值得注意的是，當多項式階數結束行為(End Behavior)的特性，使用偶數階數的多項式可能無法滿足累積密度函數結束行為的特性，所以本文不考慮偶數階數的使用，由表一可清楚看到，辨識效能隨著多項式階數增加有所進步，但並非使用階數愈高愈好，因為資料的散佈情形，可能使高階多項式為了要更符合資料分佈情形而造成過度擬合(Overfit)的情形；同樣地，若使用所有訓練語料來求算多項式係數亦會有過度擬合的情形。由於使用7階的多項式迴歸以及100分組組數的累積統計圖有較佳的辨識結果，因此下列所有有關多項式迴歸的實驗將使用7階多項式迴歸，並且參考分佈是利用100組的累積統計圖中統計而得。

4.2 時間序列上特徵值移動平均之使用

本小節將探討移動平均的使用，對於減輕由噪音或等化過程中所造成的異常情形，進而提高辨識能的效果如何，實驗結果如表二所示，表中清楚的呈現無論是使用哪種移動平均法，對於提升多項式擬合統計圖等化後語音特徵的辨識效果皆有明顯的幫助，其中當移動平均項為0時，表示不做任何平均動作，亦即單純使用多項式擬合統計圖等化法所得到的辨識結果。

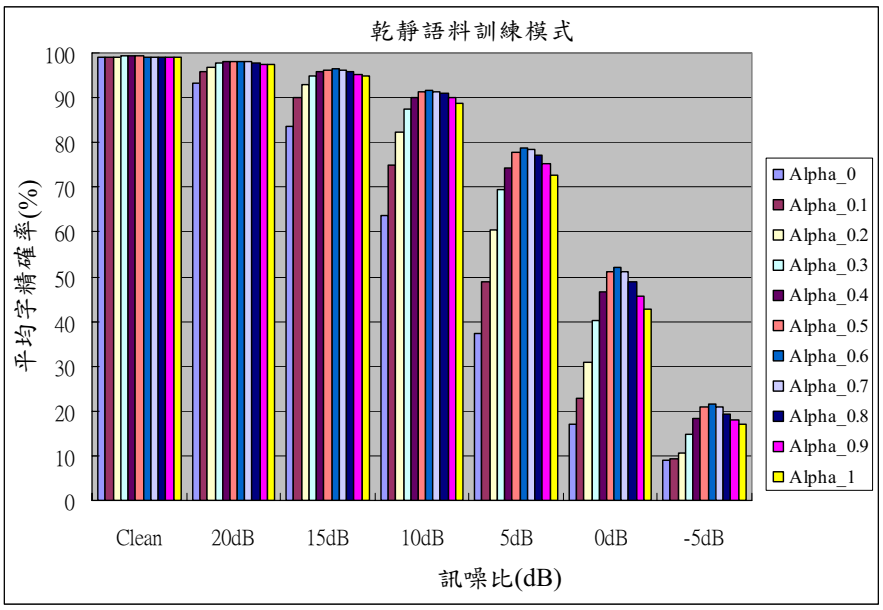
實驗結果和[22]呈現的相同，使用非因果關係自動迴歸移動平均(Non-Casual ARMA)會有較佳的辨識結果，相較於單純使用多項式擬合統計圖等化法而言，乾淨語料訓練模式，字錯誤率可達約 20%的相對進步(Relative Improvement)，對複合情境訓練模式也可有約 5%的相對進步。但是，若移動平均項的階數若使用太高，可能會造成原本帶有鑑別資訊的特徵值，因此被移除掉，使得辨識效果下降。

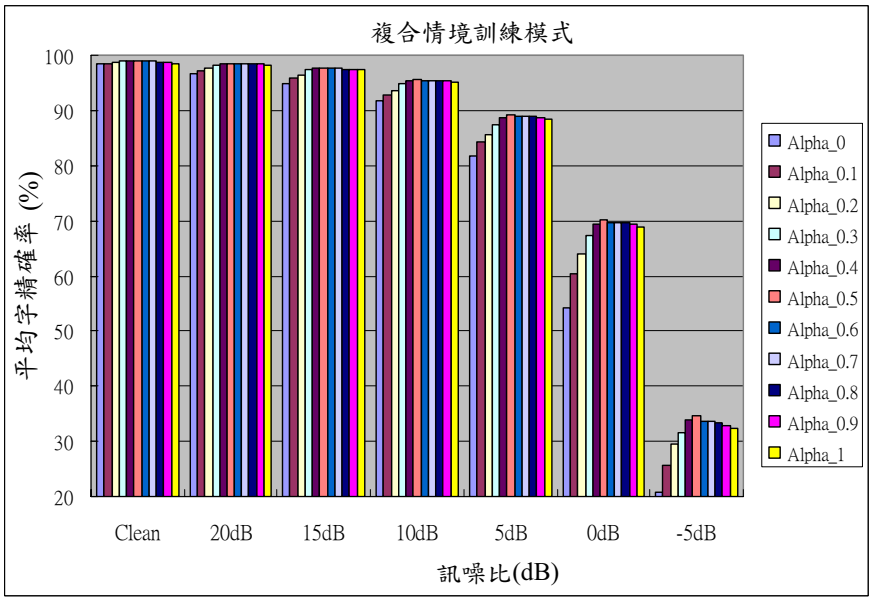
4.3 保留等化前與等化後的特徵值資訊

此小節實驗是根據式(16)，分別以不同的加權值α代入，辨識結果如圖六所示。當$\alpha=1$表

表二、多項式擬合統計圖等化法結合不同移動平均法之平均字錯誤率實驗結果

字錯誤率(Word Error Rate, WER)		移動平均項					
		0	1	2	3	4	5
乾淨語料訓練模式	Non-Casual MA	20.75	17.75	16.83	17.26	18.15	19.66
	Casual MA	20.75	19.23	18.28	17.44	17.12	17.28
	Non-Casual ARMA	20.75	17.83	16.90	16.38	16.99	17.34
	Casual ARMA	20.75	17.93	16.84	19.20	17.44	19.20
複合情境訓練模式	Non-Casual MA	10.36	9.88	9.88	10.24	10.94	11.69
	Casual MA	10.36	10.13	9.74	9.76	9.78	10.12
	Non-Casual ARMA	10.36	9.88	9.78	9.84	9.94	10.11
	Casual ARMA	10.36	9.95	9.71	10.84	9.76	10.68





圖六、保留等化前與等化後的特徵值資訊之平均字精確率實驗結果

表三、本論文所提出的多項式擬合統計圖等化法與與其他正規化補償技巧之比較

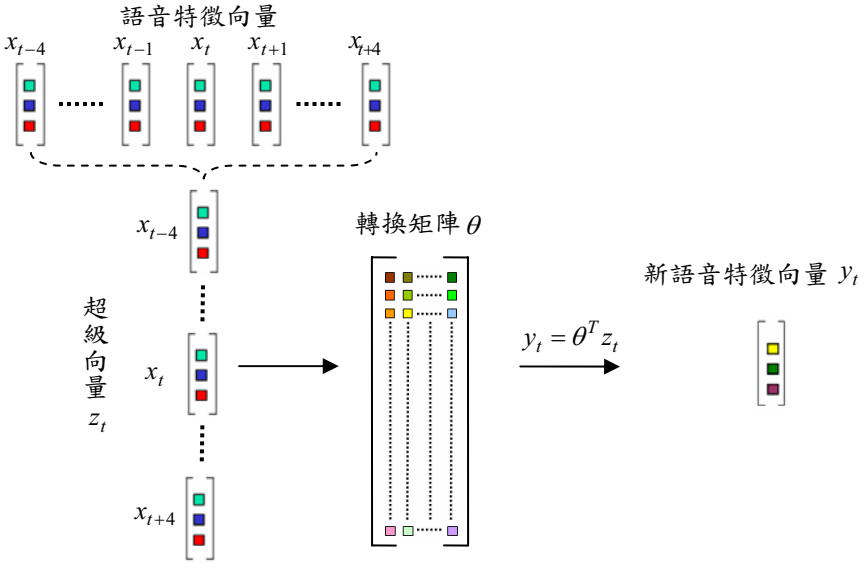
		平均字錯誤率 WER(%)			
		Set A	Set B	Set C	平均
乾淨語料訓練模式	MFCC	41.06	41.52	40.03	41.04
	AFE	38.69	44.25	28.76	38.93
	CMVN	27.73	24.60	27.17	26.37
	MS+VN+ARMA(3)	18.38	16.14	21.81	18.17
	THEQ	19.72	18.57	19.24	19.16
	QHEQ	23.53	21.90	22.36	22.64
	PHEQ	20.98	20.17	21.43	20.75
	PHEQ+MA	16.83	15.10	20.02	16.78
	PHEQ+α+MA	16.19	15.17	19.72	16.49
複合情境訓練模式	MFCC	14.78	16.01	19.33	16.18
	AFE	10.64	10.76	12.85	11.13
	CMVN	12.70	12.45	14.52	12.98
	MS+VN+ARMA(3)	9.49	10.37	10.06	9.95
	THEQ	10.02	10.41	10.34	10.24
	QHEQ	10.20	10.75	10.76	10.53
	PHEQ	9.91	9.41	13.14	10.36
	PHEQ+MA	9.41	9.53	11.21	9.82
	PHEQ+α+MA	9.15	9.08	11.53	9.60

示完全使用多項式擬合統計圖等化法所得到的辨識效果，相反地，當$\alpha=0$時，為未使用多項式擬合統計圖等化法所得到的辨識效果，從圖中可得知隨著訊噪比降低，若只單純使用多項式擬合統計圖等化法可能會有部份特徵值因等化過程而被異常放大或縮小的情況，導致辨識效果降低，因此適當保留等化前的特徵值對辨識效能能有所提昇。

因為此方法跟前面小節所敘述的時間序列上特徵值移動平均皆具有特徵值平滑(Smoothing)的效果，因此若要同時要使用此二種平均法，可能會跟使用高階移動平均項存在著相同的問題，原本帶有鑑別資訊的特徵可能因此被平滑掉，所以本文建議先以式(16)進行等法前與等化後的特徵值加權平均，再搭配式(12)~(15)低階的移動平均使用。

4.4 與其他正規化補償方法之比較

此章節將本論文所提出的方法與其他現有的正規化補償技巧進行比較，包括歐洲電信標準協會的標準前端特徵擷取(Advanced Front-End Processing, AFE)、倒頻譜正規化法(CMVN)、查表式統計圖等化法(THEQ)、分位差統計圖等化法(QHEQ)、3 階移動平均的使用(MS+VN+ARMA)以及本論文所提出的方法(PHEQ)、搭配 3 階非因果關係自動迴歸移動(PHEQ+MA)與採用α設定為 0.6 搭配 1 階非因果關係自動迴歸移動平均(PHEQ+α+MA)，其中查表式統計圖等化法和分位差統計圖等化法的實驗結果分別直接採用[18]和[14]的實驗數據結果，實驗結果如表三所示，本論文所提出的多項式擬合統計圖等化法若與單純的梅爾倒頻譜係數或是歐洲電信標準協會的標準前端特徵擷取或是倒頻譜正規化法都有明顯進步，並且和傳統查表式統計圖等化法或是分位差統計圖



圖七、鑑別性特徵擷取法示意圖

等化法的補償效果不分軒輊，若適當的加入時間序列上特徵值平均的使用(PHEQ+MA與PHEQ+α+MA)，辨識效果則有更顯著的進步。

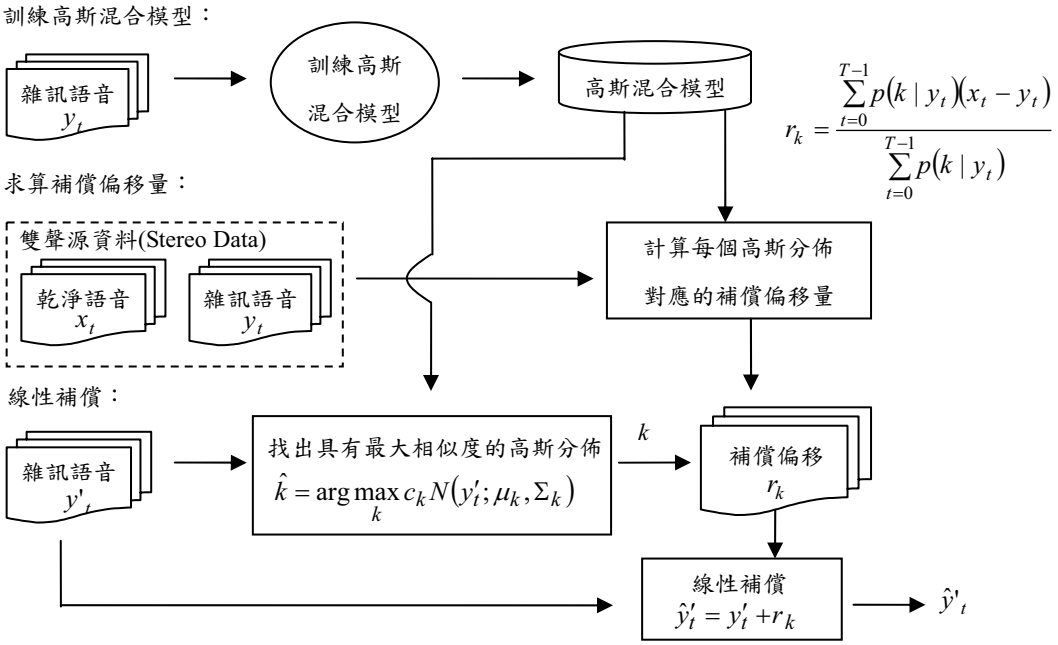
4.5 結合其他特徵擷取或補償方法

最後本論文嘗試將所提出的多項式擬合統計圖等化法與鑑別性特徵擷取法和雙聲源為基礎之分段線性補償(Stereo-based Piecewise Linear Compensation, SPLICE)[25]進行結合。在鑑別性特徵擷取法我們使用異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA) [26]加上最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)[27]並且作用在梅爾對數濾波器組輸出值之後，用來取代傳統梅爾倒頻譜係數擷取過程中需透過離散餘弦轉換(Discrete Cosine Transform, DCT)達到各維度特徵向量部份解相關(Partial Decorrection)的效果，詳細數學推導可參考[28]，整體語音特徵擷取示意圖如圖七所示，對每個時間點 t 的特徵向量，是採用該時間點特徵向量加上前後各取4個時間點特徵向量形成超級特徵向量 z_t (Feature Supervector) ，此特徵向量 z_t 經由異質性線性鑑別分析與最大相似度線性轉換的基底矩陣 θ 進行線性轉換後，可得新語音特徵向量 y_t ，數學式表示如下：

$$y_t = \theta^T z_t \tag{18}$$

最後再以多項式擬合統計圖等化法進行等化動作，實驗結果如表四所示第一列和第二列所示，無論是乾淨語料訓練模式或是複合情境訓練模式都比倒頻譜正規化有更明顯的補償效果，其中以乾淨語料訓練模式而言，與倒頻譜正規化法相比約有25%的相對進步。

另外在與雙聲源為基礎之分段線性補償結合的實驗，因此法有個前題是需擁有雙聲源的語音訊號，正好依照Aurora-2的設定，此雙聲源可從自乾淨語料訓練模式的語音及其對應的複合情境訓練模式的語音而得，整個方法流程如圖八所示，首先雜訊語料須先訓練出一個高斯混合模型，在論文設為512個高斯分佈，對於每個高斯分佈會計算相對應的補償偏移量 r_k ，運算如下[25]：



圖八、雙聲源為基礎之分段線性補償流程圖

$$r_k = \frac{\sum_{t=1}^{N} p(k \mid y_t)(x_t - y_t)}{\sum_{t=1}^{N} p(k \mid y_t)} \tag{18}$$

其中 N 為所有訓練語料音框個數，y_t 為時間點 t 含雜訊的訓練語音特徵向量，x_t 為相對應的乾淨訓練語音特徵向量，k 表示高斯混合模型中第 k 個高斯分佈。而測試語音特徵向量 y'_t 的補償後測試語音特徵向量 $\hat{y}'_t$ 可由下式二個步驟求得：

$$\begin{aligned} \hat{k} &= \arg\max_k c_k N(y'_t; \mu_k, \Sigma_k) \\ \hat{y}'_t &= y'_t + r_{\hat{k}} \end{aligned} \tag{19}$$

第一個步驟是找出 y'_t 和高斯混合分佈中具有最大相似度(Likelihood)的高斯分佈 k ，因為高斯混合模型是經由雜訊語料訓練而成，因此我們可以視每一個高斯分佈是代表某一種類型與訊噪比的噪音，此步驟即找出最相似的高斯分佈來進行補償，接著再以由式(18)所事先求得的補償偏移量進行補償。實驗結果如表四第三列和第四列所示，其中因為Set C是包含與訓練語料不同通道效應的測試語料，在求算補償偏移量時沒有被考慮到，因此補償效果較有限，實驗結果亦證明雙聲源為基礎之分段線性補償結合本論文所提出的方法會比結合倒頻譜正規化有更佳的補償效果。

5. 結論

本論文成功利用數據擬合的方法創造一逆函數，能有效且快速的將測試語句累積密度函數近似至參考資料的累積密度函數，藉由逆函數的使用，成功地改善傳統統計圖等化法或分位差統計圖等化法需要耗費大量記憶體使用空間或是處理器運算時間的缺點，同時也探討時間序列上特徵值移動平均法對於減輕因由非穩性噪音所造成的異常尖峰或波谷及等化過程中造成部份特徵值被過度放大或縮小的異常情形。實驗結果清楚的呈現本論文所提出的特徵值正規化法對噪音環境下的語音有很顯著的幫助。此外，本論文也嘗試和其他特徵擷取或補償方法進行結合，實驗結果亦呈

表四、整合其他特徵擷取或補償方法之實驗結果

		平均字錯誤率 WER(%)			
		Set A	Set B	Set C	平均
乾淨語料訓練模式	HLDA-MLLT+CMVN	21.63	21.37	21.59	21.52
	HLDA-MLLT+PHEQ-MA	15.98	15.96	15.91	15.96
	SPLICE+CMVN	16.34	14.95	21.18	16.75
	SPLICE+PHEQ-MA	13.40	13.41	17.08	14.14
複合情境訓練模式	HLDA-MLLT+CMVN	9.49	9.51	10.40	9.68
	HLDA-MLLT+PHEQ-MA	9.06	8.87	8.55	8.88
	SPLICE+CMVN	10.40	11.00	13.80	11.32
	SPLICE+PHEQ-MA	9.54	10.88	12.18	10.60

現補償效果比倒頻譜正規化法有更顯著的效果，與 HLDA+MLLT 的結合，在複合情境訓練模式下，有最佳的辨識效果，平均字錯誤率達 8.88%；另外與 SPLICE 結合，在乾淨語料訓練模式下，平均字錯誤率達 14.14%。

6. 參考文獻

[1] Y. Gong, “Speech Recognition in noisy environments: A survey,” *Speech communication*, Vol.16, 1995.

[2] S.F. Boll, “Supperssion of Acoutstic Noise in Speech Using Spectral Subtraction,” *IEEE Trans. on ASSP*, Vol.27, No.2, pp.133-120, 1979.

[3] X. Huang, A. Acero and H. Hon, “Spoken Language Processing: A Guide to Theory, Algorithm and System Development,” *Prentice Hall PTR Upper Saddle River*, NJ, USA, 2001.

[4] S. Furui, “Cepstral Analysis Techniques for Automatic Speaker Verification,” *IEEE Trans. on ASSP*, 1981.

[5] A. Viikki and K. Laurila, “Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition,” *Speech Communication*, Vol. 25, 1998.

[6] J.L. Gauian and C.H. Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” *IEEE Trans. on Speech and Audio Processing*, 1994.

[7] C.J. Leggetter and P.C. Woodland, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models,” *Computer Speech and Language*, 1995.

[8] Y. H. Suk, S. H. Choi, H. S. Lee, “Cepstrum Third-order Normalization Method for Noisy Speech Recognition,” *Electronics Letters*, Vol. 35, no. 7, pp. 527-528, April 1999.

[9] C.W. Hsu and L.S. Lee, “Higher Order Cepstral Moment Normalization (HOCMN) for Robust Speech Recognition,” in *Proc. ICASSP 2004*.

[10] S. Dharanipargda and M. Padmanabhan, “A Nonlinear Unsupervised Adaptation Technique for Speech Recognition,” in *Proc. ICSLP 2000*.

[11] C. Y. Wan and L.S. Lee, "Joint Uncertainty Decoding (JUD) with Histogram-Based Quantization (HQ) for Robust and/or Distributed Speech Recognition," in *Proc. ICASSP 2006*.

[12] C.Y. Wan and L.S. Lee, "Histogram-based quantization (HQ) for robust and scalable distributed speech recognition," in *Proc. EUROSPEECH 2005*.

[13] F. Hilger, H. Ney, "Quantile based histogram equalization for noise robust speech recognition," in *Proc. EUROPSEECH 2001*.

[14] F. Hilger et al., "Quantile Based Histogram Equalization for Noise Robust Large Vocabulary Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 14(3), 2005.

[15] A. de la Torre et al., "Non-linear Transformation of the Feature Space for Robust Speech Recognition," in *Proc. ICASSP 2002*.

[16] S. Molau et al., "Histogram Based Normalization in the Acoustic Feature Space," in *Proc. ASRU 2001*.

[17] S. Molau et al., "Feature Space Normalization in Adverse Acoustic Conditions," in *Proc. ICASSP 2003*.

[18] S. Molau et al., "Histogram Normalization in the Acoustic Feature Space," in *Proc. ICASSP 2002*.

[19] J. C. Segura et al., "Cepstral domain segmental nonlinear feature transformations for robust speech recognition," *IEEE Signal Processing Letters*, Vol. 11(5), 2004.

[20] A. de la Torre et al., "Histogram equalization of the speech representation for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 13(3), 2005.

[21] S.H. Lin, Y.M. Yeh and B. Chen, "Exploiting Polynomial-Fit Histogram Equalization and Temporal Average for Robust Speech Recognition," in *Proc. ICSLP 2006*.

[22] C.P. Chen, J. Bilmes and K. Kirchhoff, "Low-Resource Noise-Robust Feature Post-Processing on Aurora 2.0," in *Proc. ICSLP 2002*.

[23] H. G. Hirsch, D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," in *Proc. ISCA ITRW ASR 2000*.

[24] S. Young et al., "The HTK Book Version 3.3," 2005.

[25] L. Deng, A. Acero, M. Plumpe and X. Huang. "Large-Vocabulary Speech Recognition under Adverse Acoustic Environments," in *Proc. ICSLP 2000*.

[26] M. J. F. Gales, "Maximum Likelihood Multiple Projection Schemes for Hidden Markov Models," *Cambridge University Technical Report RT-365,* 2001.

[27] G. Saon et al., "Maximum Likelihood Discriminant Feature Spaces," in *Proc. ICASSP 2000*.

[28] 張志豪, "強健性和鑑別力語音特徵擷取技術於大詞彙連續語音辨識之研究," *國立台灣師範大學資訊工程研究所碩士論文*, 2005.

應用不定長度特徵之條件隨機域於口語不流暢語流修正

Disfluency Correction of Spontaneous Speech using Conditional Random Fields with Variable Length Features

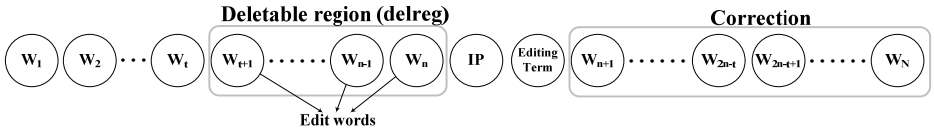
葉瑞峰[2]、吳宗憲[1]、吳維彥[1]
[1]Dept. of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan
[2]Dept. of Computer Science and Information Engineering, Far East University, Tainan County, Taiwan

摘要

針對口語化語音中之不流暢語流(disfluency)現象，本文提出以不定長度特徵之條件隨機域。利用狀態轉移特徵函數、觀測特徵函數以及相對應之參數，針對不流暢語流進行修正。其中觀測特徵函數可整合多種知識來源，包括前後文相關特徵、不流暢相關特徵以及圖樣符合相關特徵。在狀態方面我們使用可變動長度單位，包括詞、字元串集(chunk)以及句子三種不同狀態。在評估上，則使用現代漢語口語對話語料庫(MCDC)做為訓練以及測試語料。其中被修正詞(editing word)錯誤率為 17.3%，相對於 DF-gram、HMM、最大熵以及 N-gram 加校正之混合模型的方法分別降低了 11.7%、8.7%、8%以及 3.9%。在給定中斷點的情況下，被修正詞錯誤率為 6.1%。實驗證明所提之模型優於其他方法，並可有效偵測並修正口述語言中之不流暢語流。

1. 緒論

要應用語音技術於人機介面上，語音辨識則為最重要且核心之技術之一。近十年來，語音辨認技術已臻於成熟且蓬勃發展。目前的語音辨認系統對於朗讀的語音輸入辨認效果極佳，然而要實際應用，必須考慮口語化語音[1]。而口語化語音常會伴隨著非正規化(ill-formed)以及不流暢語流(disfluency)，這些現象會造成目前辨識系統的錯誤率大幅度提高，以至於無法應用於日常生活[2]。而參雜著不流暢語流之辨識後文字，也會使得使用者極不容易閱讀，對使用者造成困擾[3]。編輯不流暢語流結構共可區分為以下四個部份如圖一所示。



圖一 編輯不流暢語流之結構

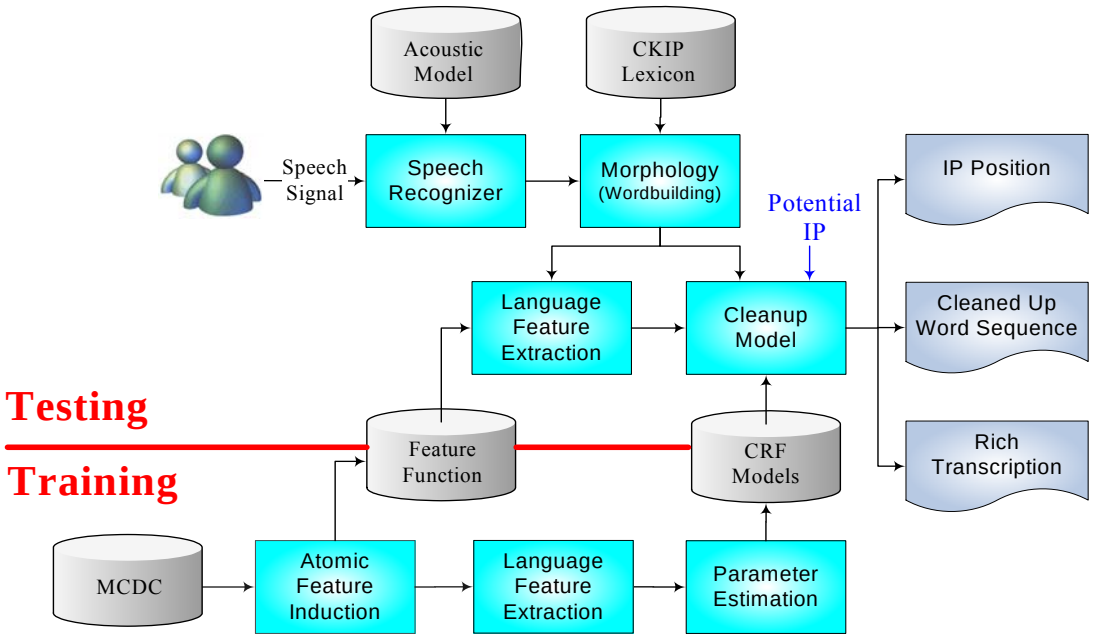
編輯不流暢語流包括三種型別：重複(Repetition)、修正(Repair) 和重開始(Restart)，其定義如下。重複即語者重複語句的某個部份，也就是可刪除區域與修正區域的語句重複。修正即語者將語句的某個部份做修正。也就是可刪除區域將取代修正區域並改變它的意思。重開始:語者將未完成的語句中斷並重新開始另一句。也就是中斷點前面的部分全都是可刪除區域。

相關的研究在國外方面， ISCI 以及 SRI 等國際研究中心利用語言模型以及韻律模型偵測不流暢語流[4]、結合基於詞和詞性的語言模型解決重複[5]和使用隱藏事件語言模型直接對不流暢語流進行統計式分析以及利用不流暢語流語言模型(DF-gram)來預測是否出現不流暢現象[6]以及使

用最大熵模型以及隱藏馬可夫模型修正不流暢語流[7]。John Bear 應用不同知識來源來針對不流暢語流進行偵測及修正[8]，Anand Venkataraman 使用人工訂定之規則來判斷不流暢語流[9]、Matthias Honal 利用噪音頻道(Noisy Channel)的觀念，運用不同特徵訓練出統計模型並以線性組合來修正之[10]。Charniak and Johnson 建立一基於詞性特徵之分類器來預測可被刪除區域[11]。Nakatani and Hirschberg 利用聲學、音韻學以及語言特徵建立一決策樹模型來偵測重複[12]。Snover 等人以及 Joungbum 等人利用轉換學習(Transformation-Based Learning)偵測不流暢語流[13][14]。日本的 Furui 則致力於口語化語音辨識之研究[15]。國內方面，中研院針對不流暢語流的語音特性做分析[16]，台灣大學研究關於不流暢語流中斷點偵測之特徵[17]、交通大學電信工程系則針對自發性中文語音建立辨識系統[18]以及自發性對話語音辨識做研究[19]。近年來，成功大學也投入大量研究能量於口語對話系統中不流暢語音之語音動作型態模型化與驗證[20]以及運用語言模型與校正模型來對編輯不流暢語流做修正[21]。本文則針對編輯不流暢語流提出一利用條件隨機域之修正方法。

2. 系統架構

本文所提之系統架構如下所示:



圖二，不流暢語流修正系統之系統架構

各模組之功能如下:

語音辨識器模組 (Speech Recognizer):利用 HTK 將語音信號進行語音辨識並得到音節網(Syllable Lattice)。其中我們定義 157 個次音節模型(sub-syllable models)以及 11 個填空詞模型(filler models)。構辭模組 (Morphology):將經過語音辨識器所得之音節網對照詞典得到相對應的詞網(Word Lattice)。

語言特徵擷取模組 (Language Feature Extraction):根據特徵，對詞網中語言上的資訊特徵擷取。

子特徵推導 (Atomic Feature Induction):使用語料訓練出最小單元的特徵，稱之為子特徵。

參數估測 (Parameter Estimation): 對於不同的特徵及其相對應的參數值，估計出這些參數。

修正模型 (Cleanup Model): 根據詞絡、可能的中斷點以及擷取之語言特徵配合對應的參數對詞絡作修正。

最後，詞絡經過修正模型修正之後，我們將得到中斷點資訊、修正後結果、以及辨識後結果。而系統流程分為訓練和測試兩部份，分別如下:

訓練部份--首先，從 MCDC 語料中經由子特徵堆導得到子特徵。這些子特徵則成為我們修正模型中的特徵函數。最後則是對語料進行語言特徵擷取並對所有的特徵函數進行參數估測，這些參數即為測試時修正模型之參數。

測試部分--語音訊號經由辨識器進行語音辨識後，得到音節絡，將此音節絡配合詞典經由構辭後會得到詞絡，然後對此詞絡做語言特徵擷取。最後，修正模型則根據詞絡、可能的中斷點以及所擷取之語言特徵，配合訓練所得到的參數模型，找出中斷點資訊、修正後結果、以及辨識後結果並將其輸出。

而修正不流暢語流之流程以式子(1)表示，一語音訊號 X 輸入後，我們要得到其相對應之最佳狀態序列 S，於是引入詞序列 W 此參數，之後在條件獨立的假設下，得到最後的式子，也就是從語音訊號我們可經過辨識器得到詞序列，而後我們從詞序列找到相對應的狀態序列。在本論文中，我們使用不定長度特徵之條件隨機域得到 $P(\mathrm{S}|\mathrm{W})$。

$$\begin{aligned}\hat{\mathrm{S}} &= \arg\max_{\mathrm{S}} P(\mathrm{S}\,|\,\mathrm{X}) \\ &= \arg\max_{S}\left(\sum_{W} P(\mathrm{S}|\mathrm{W},\mathrm{X})\mathrm{P}(\mathrm{W}|\mathrm{X})\right) \qquad (1)\\ &\cong \arg\max_{S}\left(\sum_{W} P(\mathrm{S}|\mathrm{W})\mathrm{P}(\mathrm{W}|\mathrm{X})\right)\end{aligned}$$

3. 不定長度特徵之條件隨機域

條件隨機域

條件隨機域為一種無向圖(Undirected Graphical)的模型，可被用來估算給予一觀測序列，得到相對應的狀態序列其交集的機率分佈。其概念是以隨機域為基礎，加上全域被限制於 X 這個條件，稱之一條件隨機域，X 為觀測序列。正式的來說，我們定義一圖 G=(S,E)為一無向圖，S 為所有點之集合，每個點皆為隨機變數，我們可將某個點 Sv 看成狀態序列 Y 上的某個狀態 Yv。若每個隨機變數 Yv 都遵守馬可夫原則，也就是說給予 X 和所有其他隨機變數 Y{u|u≠v}的條件之下，得到隨機變數 Yv 的機率

$$p(\mathrm{Y_v}\,|\,\mathrm{X},\mathrm{Y_u},u\neq v,\{u,v\}\in \mathrm{V}) \qquad (2)$$

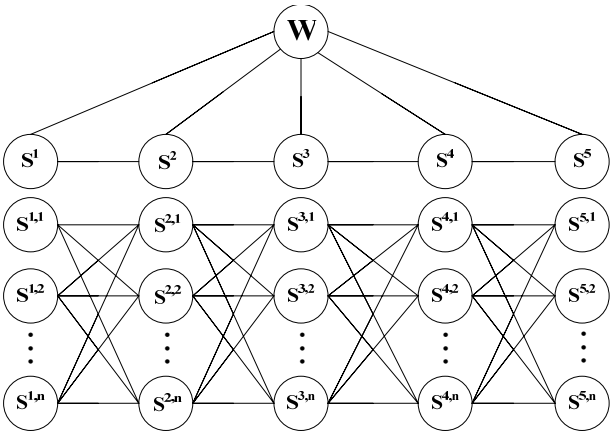
相等於給予 X 和 Yv 的鄰居點的條件之下，得到隨機變數 Yv 的機率，

$$p(\mathrm{Y_v}\,|\,\mathrm{X},\mathrm{Y_u},u=neighbor(v),\{u,v\}\in \mathrm{V}) \qquad (3)$$

則(X,Y)為一條件隨機域。

理論上來說，圖 G 的結構可以是任意的，然而，當用來對序列建模時，最簡單且最普通的圖的結構則為形成一個簡單的一階鏈(First-Order chain)，如下圖所示，其中 W 為觀測值，S 為

狀態序列。



圖三，條件隨機域示意圖

於是在給予觀測序列 X，得到對應狀態序列 S 的機率為:

$$P(S|W)=\frac{1}{Z}exp\left(\sum_t\sum_k \lambda_k f_k\left(s^{(t-1)},s^{(t)},\mathrm{W}\right)+\sum_t\sum_k \mu_k g_k\left(s^{(t)},\mathrm{W}\right)\right) \tag{4}$$

其中 $f_k\left(s^{(t-1)},s^{(t)},\mathrm{W}\right)$ 為整個觀測序列和在狀態序列中位置 t-1 的狀態轉到位置 t 的狀態轉移特徵函數定義為；

$$f_k\left(s^{(t-1)},s^{(t)},\mathrm{W}\right)=\begin{cases}1 & if\ s^{(t-1)}=s\wedge s^{(t)}=s'\\ 0 & otherwise\end{cases} \tag{5}$$

而 $g_k\left(s^{(t)},\mathrm{W}\right)$ 為狀態序列中位置 t 和觀測序列的狀態特徵函數:

$$g_k\left(s^{(t)},\mathrm{W}\right)=\begin{cases}1 & if\ s^{(t)}=s\wedge \mathrm{W}^{(t)}=w\\ 0 & otherwise\end{cases} \tag{6}$$

λ_k 以及 μ_k 為從訓練語料中估出來的參數，Z 是正規化係數，為所有可能的狀態序列的機率總和，其定義如下式所示:

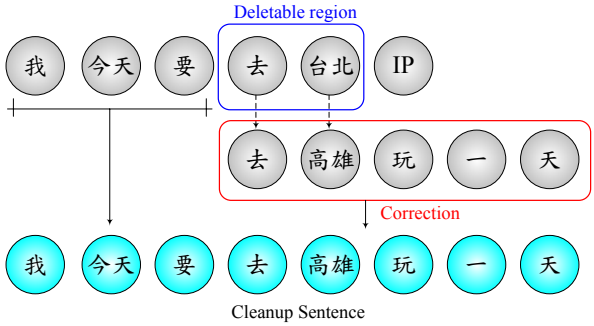
$$Z=\sum_{W,s} exp\left(\sum_t\sum_k \lambda_k f_k\left(s^{(t-1)},s^{(t)},\mathrm{W}\right)+\sum_t\sum_k \mu_k g_k\left(s^{(t)},\mathrm{W}\right)\right) \tag{7}$$

就修正不流暢語流而言，我們可以把觀測序列 X 當作是以詞構成的序列，狀態序列 Y 為以 1 和 0 組成之序列，若在觀測序列中位置 t 的詞所對應的狀態為 0，則代表此詞為一被修正詞，需將其刪除；若為 1，則表示位置 t 的詞為流暢部分，並將其保留。

.

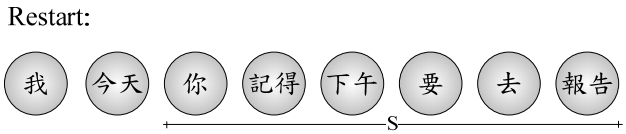
不定長度特徵

在不流暢語流中的"修正"型別中，其發生的情形為語者語句部分有誤，故會修正之。我們從語料觀察到修正時常會存在著圖樣一致性(Pattern Matching)的現象，也就是修正與被修正的兩個部份其句法結構相似，這些圖樣的組成可能是一個片語(phrase)、或是一個字元串集(chunk)，如圖四。於是我們先將某些詞的單元進一步合併為字元串集單元。



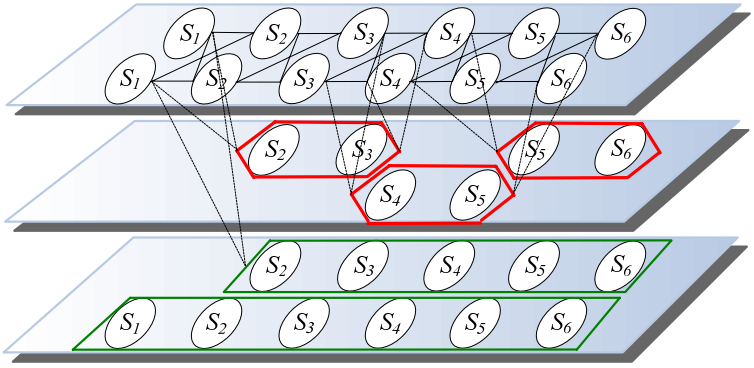
圖四，編輯不流暢語流之修正

而在"重開始"此種型別的不流暢語句中，因為使用者是將原句放棄並重新開始另一語句，故在其結構中幾乎都會有包含另一語句的現象發生，也就是說若一輸入語句包含另一語句，則是"重開始" 此種型別的不流暢語句的機率大增，如圖五。若我們能先將某些詞的單元進一步合併為句子單元，則對於修正不流暢語流之"重開始"型別，將會大大減少被修正辭誤判(False Alarm)的錯誤率。



圖五，編輯不流暢語流之 restart

原本條件隨機域的狀態序列皆以詞為單位，而本論文提出一種以條件隨機域為基礎之方法—不定長度特徵之條件隨機域，對於一輸入之詞序列，於找出最佳路徑狀態序列之前，先將可合併之字元串集及句子合併，觀念似於語音辨識時所做構詞的部份。而狀態部份則從原來的詞，增加為共三層狀態，分別是詞、字元串集以及句子，再根據合併後的狀態絡找出最佳狀態序列，然後依此狀態序列得到最後的修正結果。如下圖六所示:



圖六，不定長度特徵之條件隨機域

在給予觀測序列 W，得到對應狀態序列 S 的機率為:

$$P(S|W)=\frac{1}{Z}exp\left(\sum_{t}\sum_{k}\sum_{p,q}\lambda_k f_k\left(s_p^{(t-1)},s_p^{(t)},\mathrm{W}\right)+\sum_{t}\sum_{k}\sum_{c}\mu_k g_k\left(s_p^{(t)},\mathrm{W}\right)\right) \quad (8)$$

其中 p,q 為層次個數，在此共有詞、字元串集以及句子這三種層次，而每一層次包含兩種狀態，一種為 0，也就是此層次為可刪除區域，修正時需將其刪除；另一種則是 1，即為流暢部分，修正時將其保留。其中 $f_k\left(s_p^{(t-1)},s_p^{(t)},\mathrm{W}\right)$ 為整個觀測序列和在 p 層次狀態序列中位置 t-1 的狀態轉到位置 t 的狀態轉移特徵函數，而 $g_k\left(s_p^{(t)},\mathrm{W}\right)$ 為 p 層次狀態序列中位置 t 和觀測序列的狀態特徵函數。

在特徵函數方面我們共分為三類，分別是上下文相關、不流暢相關以及圖樣符合相關觀測特徵函數，我們一一介紹如下。

上下文相關觀測特徵函數:上下文相關觀測特徵函數乃是根據所觀測點位置之上下文關係取得一觀測範圍並以 N-gram 的概念所建立之特徵函數並以樣板(Template)的形式來表示。

不流暢相關特徵函數:我們從 Apriori 演算法擷取出關聯法則，以此關聯法則為我們的不流暢相關特徵函數。譬如我們找出"去台北=>去高雄"此一關聯法則，我們即可定義觀測特徵函數如下:

$$g_1\left(s_p^{(t)},\mathrm{W}\right)=\begin{cases}1 & if\ s_p^{(t)}=s\wedge W^{t-k}出現"去台北"\wedge W^{t+k}出現"去高雄"\\ 0 & otherwise\end{cases} \quad (9)$$

故若有符合此特徵函數，則"去台北"被刪除機率則大增。所有擷取到的的關聯法則皆可以此方式定義出。

圖樣符合相關觀測特徵函數:因修正(repair)此種不流暢語流會有圖樣符合(Pattern Matching)的情形，故我們訂定了圖樣符合相關觀測特徵函數，例子如下

$$g_2\left(s_p^{(t)},\mathrm{W}\right)=\begin{cases}1 & if\ s_p^{(t)}=s\wedge P^{t-1}=P^{t+1}\wedge P^{t}=P^{t+2}\\ 0 & otherwise\end{cases} \quad (10)$$

也就是當出現連續兩個圖樣其詞性序列一致時，極有可能是不流暢現象。因修正之編輯詞個數集中在 1-2 個，故我們定義時皆以 2 個詞的詞性的組合當做圖樣符合相關觀測特徵函數。我們總共有 37 種詞性，故會組合出 1369 種特徵函數。因填空詞出現在語料中，所以我們加上位移 k 的考慮如下:

$$g_3\left(s_p^{(t)},\mathrm{W}\right)=\begin{cases}1 & if\ s_p^{(t)}=s\wedge P^{t-1}=P^{t+1+k}\wedge P^{t}=P^{t+2+k}\\ 0 & otherwise\end{cases} \quad (11)$$

於是我們利用這些特徵函數配合參數估測得到各權重植，之後利用此修正模型產生與輸入句相對應的狀態序列，並對原句進行修正。

子特徵推導演算法

在建立不定長度特徵之條件隨機域時，我們必須先將不同的層次確認出來。在字元串集(chunk)

部分，我們使用的是 Apriori 演算法[22]，如找出的關聯法則包含互相鄰近的詞，則我們將這些詞組合為一字元串集；在構成句子部份，我們是以動詞配合其必要論元為一組成句子之主要成分來對輸入語句判斷是否包含另一語句。

動詞為一個句子的中心，句子的構成常被視為動詞本身語態的擴展延伸。動詞的歸納整理關係到詞彙知識及句法的表達。於是，我們利用中研院詞庫小組(CKIP)所研究之中文詞類分析[23]，其對於漢語的詞類分析及相對應的詞彙結構提出完整的看法。對於其中每一類動詞皆整理出其必要論元(argument)。我們假設句子的主成分由這些動詞與其必要論元所構成，若一輸入語句之詞性序列 x = [x1, x2, …, xm]和某一動詞以及其必要論元詞性序列 y=[y1, y2, …, yk]存在共同子序列(common subsequence)z，且 z=y，則我們判斷輸入語句包含一句子，範圍從 y1 到 xm。

假設現在有 x, y 兩個序列，若存在一個序列 z 同時為 x 與 y 的子序列，那麼 z 便稱為 x, y 的相同子序列（common subsequence）。舉個例子，假設我們輸入語句"你 今 我 明天 要 去 台北"其詞性序列 x 為[Nhaa, Ndabd, Nhaa, Nddb, Dbab, VA11,Nca]，一動詞 VC1 配合其必要論元之詞性序列 y 為[Nh, VA11,Nca]，我們則可找到其共同子序列 z=y，於是我們判斷[Nhaa, Nddb, Dbab, VA11,Nca]此序列為一句子。

參數估算

對於不定長度特徵之條件隨機域之最大對數似然法參數估測(ML)，我們定義機率為

$$p(s\,|\,W,\Theta)=\frac{1}{Z}exp\left(\sum_{t}\sum_{k}\sum_{p,q}\lambda_k f_k\left(s_p^{(t-1)},s_q^{(t)},\mathrm{W}\right)+\sum_{t}\sum_{k}\sum_{p}\mu_k g_k\left(s_p^{(t)},\mathrm{W}\right)\right) \tag{14}$$

從訓練資料的集合中來估算出使得訓練資料的 log-似然度最大的一組參數 $\Theta=(\lambda_1,\lambda_2,\cdots;\mu_1,\mu_2,\cdots)$。於是我們將條件隨機域的 $p(s\,|\,W,\Theta)$ 代入 log-似然度函數的定義式:

$$\begin{aligned}\mathrm{L}(\Theta)&=\log\prod_{\mathrm{W},s}\mathrm{p}(\mathrm{s}|\mathrm{W},\Theta)^{\tilde{p}(W,s)}\\&=\sum_{W,s}\tilde{p}(W,s)\,log\;p(s\,|\,W,\Theta)\\&=\sum_{W,s}\tilde{p}(W,s)\,log\left(\frac{1}{Z}exp\left(\sum_{t}\sum_{k}\sum_{p,q}\lambda_k f_k\left(s_p^{(t-1)},s_q^{(t)},\mathrm{W}\;\right)+\sum_{t}\sum_{k}\sum_{p}\mu_k g_k\left(s_p^{(t)},\mathrm{W}\;\right)\right)\right)\end{aligned} \tag{15}$$

經過整理之後得到下式

$$\sum_{W,s}\tilde{p}(W,s)\left[\sum_{t}\sum_{k}\sum_{p,q}\lambda_k f_k\left(s_p^{(t-1)},s_q^{(t)},\mathrm{W}\right)+\sum_{t}\sum_{k}\sum_{p}\mu_k g_k\left(s_p^{(t)},\mathrm{W}\right)\right]-\sum_{W}\tilde{p}(W)\,log\,Z \tag{16}$$

之後我們對 log-似然度函數偏微參數 λ_k：

$$\begin{aligned}\frac{\partial L(\Theta)}{\partial\lambda_k}&=\sum_{W,s}\tilde{p}(\mathrm{W},s)\sum_{t=1}^{n}\sum_{p,q=1}^{l}f_k\left(s_p^{t-1},s_q^{t},\mathrm{W}\right)\\&\quad-\sum_{W,s}\tilde{p}(W)p(s\,|\,W,\Theta)\sum_{t=1}^{n+1}\sum_{p,q=1}^{l}f_k\left(s_p^{t-1},s_q^{t},W\right)\\&=E_{\tilde{p}(W,s)}\left[f_k\right]-E_{p(s|W,\Theta)}\left[f_k\right]\end{aligned} \tag{17}$$

若要求得全域最大值則須令式子(17)為零，求得 Θ 解。不過一般來說，這是不可行的，因為將 log-

似然度函數經偏微分後設為零來解出參數Θ，未必為封閉解。故本文採用一些反覆(iterative)形式的技巧取得使 log-似然度最大之參數。因 IIS 具有較快收斂之優點，故本論文是利用 IIS 演算法來進行參數估測。IIS 演算法是以 GIS 演算法為基礎改變而成，其優點為收斂速度較 GIS 演算法快。我們以此為基礎配合 Lafferty 等人提出的動態規劃法來估算參數。

Lafferty 等人[24]觀察到對於一個鏈狀結構的條件隨機域(CRFs)，給予觀測序列 W 所得到狀態序列 s 的條件機率 p(s|W)可簡單的用矩陣的形式來表示。對於觀測序列 W 中的每一個位置 t，我們分別定義了一個 $|\kappa|\times|\kappa|$ 的矩陣隨機變數 $\mathrm{M}_t(W)=\left[M_t(s',s\,|\,W)\right]$，$\kappa$ 為狀態的種類個數。

$$\mathrm{M}_t(s',s\,|\,W)=exp\left(\sum_k\sum_{p,q}\lambda_k f_k\left(s_p^{t-1}=s',s_q^t=s,W\right)+\sum_k\sum_p\mu_k g_k\left(s_p^t=s,W\right)\right) \tag{18}$$

每個 $\mathrm{M}_t(W)$ 可視為表示在時間 t 時，模型中每個轉移的權重。於是我們可以將未正規化的條件機率 $P^*(s\,|\,W)$ 表示為矩陣的連乘積:

$$P^*(s\,|\,W)=\prod_{t=1}^{n+1}M_t\left(s_p^{t-1},s_q^t\,|\,W\right) \tag{19}$$

而正規化係數 Z(W)，只和觀測序列 W 有關，為長度 n+1 時所有可能之狀態序列組合:

$$Z(W)=\left(M_1(W)M_2(W)\cdots M_{n+1}(W)\right)_{\text{start,stop}}=\left[\prod_{t=1}^{n+1}M_t(W)\right]_{\text{start,stop}} \tag{20}$$

故正規化後的條件機率 p(s|W)可表示為:

$$P(s\,|\,W)=\frac{\prod_{t=1}^{n+1}M_t\left(s_p^{t-1},s_q^t\,|\,W\right)}{\mathrm{Z(W)}} \tag{21}$$

我們利用反覆(iterative)的形式，每一回合更新一次參數:

$$\lambda_k=\lambda_k+\Delta\lambda_k \tag{22}$$

$$\mu_k=\mu_k+\Delta\mu_k \tag{23}$$

其中每一回合更新的值

$$\Delta\lambda_k=\frac{1}{S}log\frac{E_{\tilde{p}(s|W,\Theta)}\left[f_k\right]}{E_{\tilde{p}(W,s)}\left[f_k\right]} \tag{24}$$

$$\Delta\mu_k=\frac{1}{S}log\frac{E_{\tilde{p}(s|W,\Theta)}\left[g_k\right]}{E_{\tilde{p}(W,s)}\left[g_k\right]} \tag{25}$$

在(24)式以及(25)式中，S 為某一訓練資料其包含之特徵函數的總數在全部訓練資料中最大的。

$$\mathrm{S} = max_s \left(\sum_t \sum_k \sum_{p,q} f_k \left(s_p^{(t-1)}, s_q^{(t)}, \mathrm{W} \right) + \sum_t \sum_k \sum_p g_k \left(s_p^{(t)}, \mathrm{W} \right) \right) \tag{26}$$

$E_{\tilde{p}(W,s)}[f_k]$ 為特徵函數 f_k 其訓練資料分佈的期望值:

$$E_{\tilde{p}(W,s)}[f_k] = \sum_{W,s} \tilde{\mathrm{P}}(\mathrm{W},s) \sum_{t=1}^{n+1} \sum_{p,q=1}^{l} f_k \left(s_p^{t-1}, s_q^t, W \right) \tag{27}$$

$E_{\tilde{p}(s|W,\Theta)}[f_k]$ 為預估測分佈的期望值:

$$E_{\tilde{p}(s|W,\Theta)}[f_k] = \sum_{W,s} \tilde{\mathrm{P}}(\mathrm{W}) P(s|W) \sum_{t=1}^{n+1} \sum_{p,q=1}^{l} f_k \left(s_p^{t-1}, s_q^t, W \right) \tag{28}$$

Lafferty 等人提出的動態規劃法(dynamic programming)即是利用 $P(s|W)$ 可表示為矩陣 $M_t(W)$ 的形式，故式子(28)可表示為:

$$\begin{aligned} E_{\tilde{p}(s|W,\Theta)}[f_k] &= \sum_{W,s} \tilde{\mathrm{P}}(\mathrm{W}) P(s|W) \sum_{t=1}^{n+1} \sum_{p,q=1}^{l} f_k \left(s_p^{t-1}, s_q^t, W \right) \\ &= \sum_{W} \tilde{\mathrm{P}}(\mathrm{W}) \sum_{t=1}^{n+1} \sum_{p,q=1}^{l} \sum_{s',s} f_k \left(s_p^{t-1} = s', s_q^t = s, W \right) \\ &\quad \times \frac{\alpha_t(s'|W) M_t(s',s|W) \beta_{t+1}(s|W)}{Z(W)} \end{aligned} \tag{29}$$

其中 $\alpha_t(W)$ 以及 $\beta_t(W)$ 為向前(forward)和向後(backward)向量，定義如下:

$$\alpha_0(s|W) = \begin{cases} 1 & \text{if s = start} \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

以及

$$\beta_{n+1}(s|W) = \begin{cases} 1 & \text{if s = stop} \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

其遞迴關係為:

$$\alpha_i(W) = \alpha_{t-1}(W) M_t(W) \tag{32}$$

和

$$\beta_i(W) = M_{t+1}(W) \beta_{t+1}(W) \tag{33}$$

與特徵函數 f_k 相似，特徵函數 g_k 其預估測分佈的期望值為:

$$E_{\tilde{p}(s|W,\Theta)}[g_k] = \sum_{W} \tilde{\mathrm{P}}(\mathrm{W}) \sum_{t=1}^{n+1} \sum_{p=1}^{l} \sum_{s} g_k \left(s_p^t = s, W \right) \times \frac{\alpha_t(s|W) \beta_t(s|W)}{Z(W)} \tag{34}$$

利用這些表示法來算出特徵函數的期望值，我們只需要使用動態規劃法計算每個可能的 $p(s_{t-1}, s_t \mid W)$ 的值，而不用計算整個模型的分布 $p(s_1, \cdots s_n \mid W)$。

4. 實驗與討論

語音辨識器乃採用 HMM Tool Kit (HTK)(http://htk.eng.cam.ac.uk/)，其中定義了 157 個次音節模型以及 11 個填空詞 (filler) 模型。對 TCC-300 (http://rocling.iis.sinica.edu.tw/ROCLING/MAT/Tcc_300brief.htm) 以及 MCDC 語料庫 (http://www.aclclp.org.tw/use_mat_c.php#mcdc)辨識率如表一，

表一，TCC-300 及 MCDC 辨識率

	Acc.	Del.	Sub.	Ins.
TCC-300 (Top 1)	89.51	0.15	9.55	0.79
TCC-300 (Top 3)	89.76	0.15	9.32	0.77
TCC-300 (Top 5)	90.38	0.15	8.82	0.65
MCDC (Top 1)	52.83	7.79	32.35	16.36
MCDC (Top 3)	53.27	7.75	32.32	16.32
MCDC (Top 5)	53.92	7.75	31.42	16.26

依據 MCDC 語料庫中，文字轉寫的標籤(tag) 以 mcdc-01、mcdc-02、mcdc-03 以及 mcdc-05 此四組為訓練語料，mcdc-09、mcdc-10、mcdc-25 以及 mcdc-26 此四組為測試語料。在全部 8 組語料中，包含不流暢語流之語句佔全部之 52.14%，也就是約 2 句對話就有一句具有不流暢語流之現象，顯示出對話時發生不流暢語流之情形在對話中十分常見。訓練語料與測試語料中發生不流暢語流情形如下表所示：

表二，語料中包含之不流暢語流現象之計數

	Repair	Repetition	Restart
訓練語料	107	318	94
測試語料	142	180	246

實驗中所用的比較對象共使用了四個方法分別是以最大熵模型(Maximum Entropy)、結合語言模型與校正模型、以詞為基礎之條件隨機域模型以及不流暢語流語言模型(DF-gram)。系統評估方面，則以 Rich04[25]中的被修正詞錯誤率(edit word error rate)以及中斷點錯誤率(edit IP error rate)作為比較的標準，如以下式子(35)與(36)所示。

$$Error_{EWD} = \frac{n_{M-EWD} + n_{FA-EWD}}{n_{EWD}} \tag{35}$$

$$Error_{IP} = \frac{n_{M-IP} + n_{FA-IP}}{n_{IP}} \tag{36}$$

表三為結合語言模型與校正模型的被修正詞錯誤率，其中在詞性 tri-gram 配合校正模型在 α=0.25 時效果為最佳，故之後和不定長度特徵之條件隨機域比較時我們皆以此為基準。

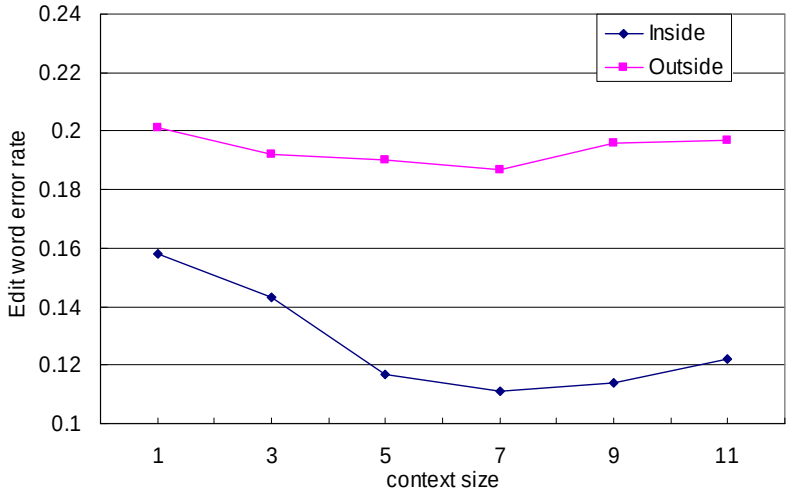
表三，結合語言模型與校正模型不同層次之被修正詞錯誤率

	Human generated transcription (REF)	Speech-to-text recognition output (STT)

	$\frac{n_{M-EWD}}{n_{EWD}}$	$\frac{n_{FA-EWD}}{n_{EWD}}$	$Error_{EWD}$	$\frac{n_{M-EWD}}{n_{EWD}}$	$\frac{n_{FA-EWD}}{n_{EWD}}$	$Error_{EWD}$
2-gram+ alignment	0.17	0.12	0.29	0.40	0.32	0.72
2-gram+ alignment[1]	0.09	0.15	0.24	0.36	0.32	0.68
2-gram+ alignment[2]	0.10	0.21	0.31	0.34	0.54	0.88
3-gram+ alignment	0.16	0.12	0.28	0.31	0.35	0.66
3-gram+ alignment[1]	0.09	0.12	0.21	0.32	0.32	0.64
3-gram+ alignment[2]	0.07	0.16	0.23	0.28	0.30	0.58

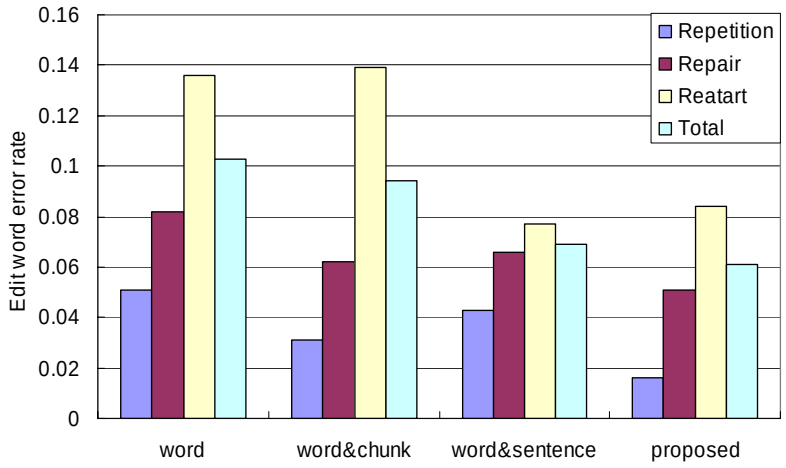
[1]: word class based on part of speech (POS) [2]: word class based on the semantic class

圖七，為使用最大熵模型，再給予中斷點情形下的被修正詞錯誤率，關於特徵函數則採用與條件隨機域相同的特徵，可以看到當 n=3 時為最佳，同樣是因為當上下文觀測範圍長度太短時，所需要的資訊不足；若是太長時，太多的特徵反而會造成混淆。故我們以此 n=3 為基準和我們提出之方法做比較。



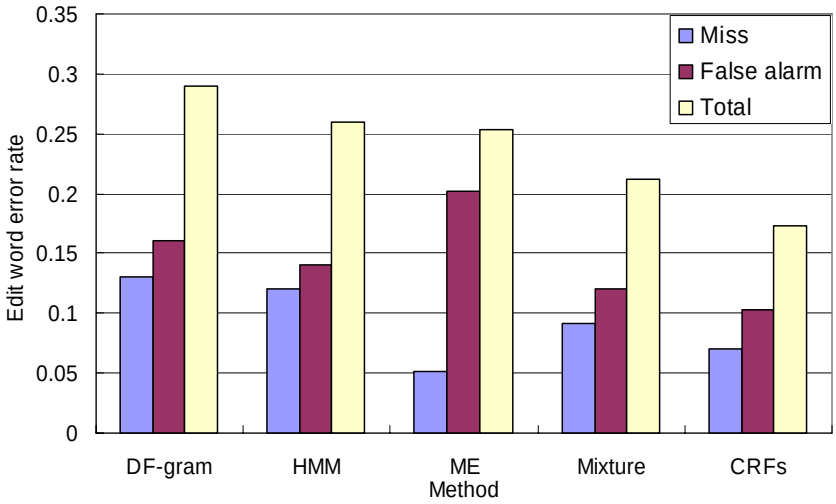
圖七，最大熵模型被修正詞錯誤率

圖八為條件隨機域分別使用不同層次所得之結果，word 代表只使用到詞的單元。word+chunk 即代表多使用了字元串集(chunk)這個單元，最後則是本論文提出的不定長度特徵之條件隨機域。我們可以看到當使用多單元時，不論二個或是三個，其錯誤率皆有降低，且對於修正(Repair)此種不流暢語流型別而言，我們可以從實驗發現使用字元串集這個層次是有幫助的；對於重開始(Restart)而言，使用句子此層次也有相當的助益。由實驗可看出我們所提出之方法對於降低被修正詞錯誤率有顯著的效果。



圖八，使用不同層次之 CRFs 在不同型別之不流暢現象被修正詞錯誤率

圖九為使用我們提出之方法與四種比較系統在文字輸入的被修正詞錯誤率，表四為使用我們提出之方法與四種比較系統在文字以及語音輸入的被修正詞錯誤率實驗證明本論文所提之模型優於其他方法。

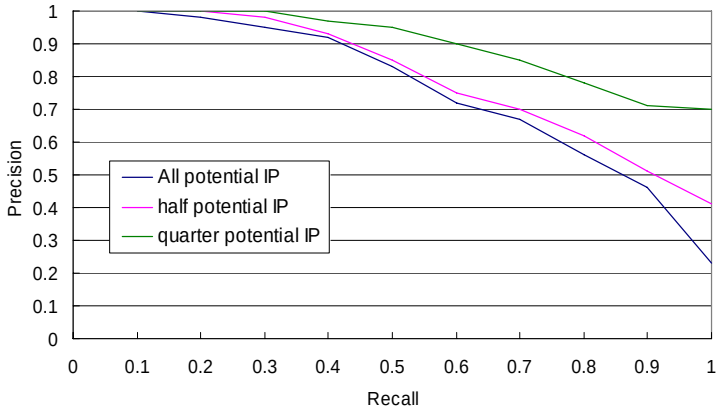


圖九，文字輸入之被修正詞錯誤率比較圖

表四，各種方法於文字與語音之被修正詞錯誤率

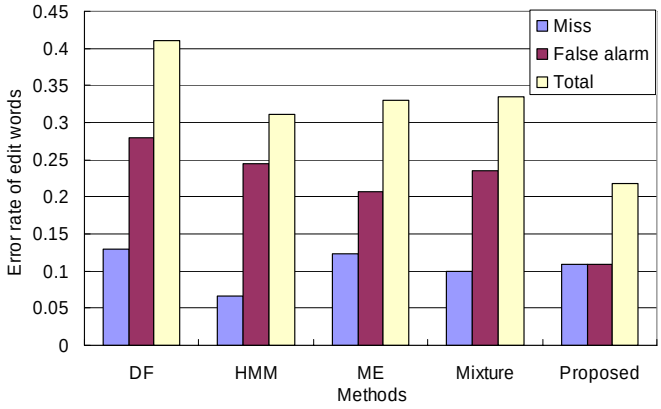
	Human generated transcription (REF)			Speech-to-text recognition output (STT)		
	$\frac{n_{M-EWD}}{n_{EWD}}$	$\frac{n_{FA-EWD}}{n_{EWD}}$	$Error_{EWD}$	$\frac{n_{M-EWD}}{n_{EWD}}$	$\frac{n_{FA-EWD}}{n_{EWD}}$	$Error_{EWD}$
DF-gram	0.13	0.16	0.29	0.37	0.346	0.71
ME	0.05	0.20	0.25	0.14	0.52	0.66
HMM	0.12	0.14	0.26	0.34	0.35	0.68
3-gram+ alignment	0.09	0.12	0.21	0.32	0.32	0.64
Proposed	0.07	0.10	0.17	0.25	0.35	0.60

圖十為在給予不同個數的潛在中斷點下，對中斷點之求全率(Recall)與求準率(Precision)曲線。可以發現如果潛在中斷點越準確的話，對於找到中斷點的效能會越高。



圖十，求全率與求準率曲線

最後，對於複雜的口語不流暢語流也就是在語句中存在有兩個或兩個以上之不流暢現象現象做處理，其結果如圖十一所示。可以發現在複雜的口語不流暢語流情形，所提之方法仍可有效的修正不流暢語流。



圖十一，複雜不流暢語流編輯詞錯誤率比較圖

5. 結論與未來工作

本文提出一不定長度特徵之條件隨機域統計式模型修正不流暢語流，配合觀測特徵函數以及狀態轉移特徵函數找出最佳狀態序列，最後根據最佳狀態序列判斷句中某詞是否須刪除。共有三種不同之狀態單元，分別為詞、字元串集以及句子。而在觀測特徵函數的產生選取上，因人工定義過於耗時且不夠強健，故利用 Apriori 演算法產生特徵函數。由實驗中我們得到所提之方法被刪除詞錯誤率為 17.3%，相對於不流暢語流語言模型、隱藏式馬可夫模型、最大熵模型以及 N-gram 加校正之混合模型的方法分別降低了 11.7%、8.7%、8%以及 3.9%。可以發現所我們所提之方法能夠有效降低被刪除詞錯誤率。同樣的在中斷點錯誤率的實驗中也可得到相同的論證。

以下介紹未來可深入探討與研發之方向:

結合語音參數： 對於修正不流暢語流部份，我們主要抽取之參數皆為語言參數，若未來能夠將語

音參數一起考慮當作特徵函數，應可有效降低中斷點錯誤率。

解決部分辭和音節合併（Contraction）的問題：部分詞及音節合併的情形常出現於自發性口語中且對於語音辨認有一定程度之影響。

6. References

1. Byrne, W., D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J.Psutka, B. Ramabhadran, D. Soergel, T. Ward, and Z. Wei-Jin, "AutomaticRecognition of Spontaneous Speech for Access to Multilingual Oral HistoryArchives," IEEE Trans. on Speech and Audio Processing, Vol. 12, No. 4, pp.420-435, 2004.
2. Kahn, J.G., M. Ostendorf and C. Chelba" Parsing Conversational Speech Using Enhanced Segmentation." Proc. HLT-NAACL, 2004. pp. 125-128.
3. Soltau, H., , B. Kingsbury, , L. Mangu, , D. Povey, , G. Saon, and D. Zweig, " The IBM 2004 Conversational Telephony System for Rich Transcription." In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP '05). (2005), 205-208.
4. Stolcke, A., E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tur, and Y. Lu." Automatic detection of sentence boundaries and disfluencies based on recognized words," In Proc. International Conference on Spoken Language Processing, pages 2247--2250, 1998.
5. Liu, Y., E. Shriberg, and A. Stolcke. "Automatic disfluency identification in conversational speech using multiple knowledge sources," In Proc. Eurospeech, volume 1, pages 957—960, 2003.
6. Stolcke, A. and E. Shriberg. "Statistical language modeling for speech disfluencies". In Proceedings of the International Conference of Acoustics, Speech, and Signal Processing, 1996.
7. Liu, Y., E. Shriberg, A. Stolcke, M. Harper"Comparing HMM, Maximum Entropy, and Conditional Random Fields for Disfluency Detection." Eurospeech 2005.
8. Bear, J., J. Dowding, and E. Shriberg, "Integrating multiple knowledge sources for detecting and correction of repairs in human computer dialog," in Proc. of ACL, 1992, pp. 56–63.
9. Stolcke, A., W. Wang, D. Vergyri, V. R. R. Gadde, and J. Zheng, "An efficient repair procedure for quick transcriptions," in Proc. Intl. Conf. Spoken Language Processing, (Jeju, Korea), October 2004.
10. Honal, M., and T. Schultz, , "Correction of disfluencies in spontaneous speech using a noisy-channel approach," In EUROSPEECH-2003, 2781-2784.
11. Charniak, E. and M. Johnson. "Edit detection and parsing for transcribed speech," In Proceedings of the North American Chapter of the Association for Computational Linguistics annual meeting, pages 118--126, 2001.
12. Nakatani, C. and J. Hirschberg. "A corpus-based study of repair cues in spontaneous speech." Journal of the Acoustical Society of America, pages 1603--1616, 1994.
13. Snover, M., B. Dorr, and R. Schwartz. "A lexically-driven algorithm for disfluency detection". In Proceedings of Human Language Technology Conference / North American Chapter of the

Association for Computational Linguistics annual meeting, 2004.

14. Kim, J., S. E. Schwarm, and M. Ostendorf," Detecting structural metadata with decision trees and transformation-based learning." Proceedings of HLT/NAACL 2004, (2004), 137–144.
15. Furui, S., K. Maekawa, H. Isahara, T. Shinozaki and T. Ohdaira"Toward the realization of spontaneous speech recognition – Introduction of a Japanese priority program and preliminary results –", Proc. ICSLP2000, Beijing.
16. Tseng, S.-C, " Repairs and Repetitions in Spontaneous Mandarin," In Proceedings of Workshop on Disfluency in Spontaneous Speech (DISS 03). Ed. Robert Eklund. Gothenburg Papers in Theoretical Linguistics 90. Pp. 71-74. University of Gothenburg.
17. Lin, Che-Kuang , Tseng, Shu-Chuan , Lee, Lin-Shan, "Important and new features with analysis for disfluency interruption point (IP) detection in spontaneous Mandarin speech", In DiSS-2005, 117-121.
18. 羅應順，自發性中文語音基本辨認系統之建立，國立交通大學電信工程所碩士論文，民國 94 年。
19. 徐文翰，自發性對話語音辨識之初步研究，國立交通大學電信工程所碩士論文，民國 93 年。
20. Wu, Chung-Hsien; Gwo-Lang Yan. "Speech act modeling and verification of spontaneous speech with disfluency in a spoken dialogue system". In Speech and Audio Processing, IEEE Transactions on Volume 13, Issue 3, May 2005 Page(s):330 – 344.
21. Yeh, Jui-Feng and Chung-Hsien Wu."Edit Disfluency Detection and Correction Using a Cleanup Language Model and an Alignment Model,"accepted by IEEE Trans. Audio, Speech, and Language Processing, 2006.
22. Chien, Jen-Tzung,"Association Pattern Language Modeling." IEEE Transactions on Audio, Speech, and Language Processing : Accepted for future publication Volume PP, Issue 99, 2005 Page(s):1-10
23. 中研院詞庫小組技術報告 93-05 中文詞類分析。
24. Lafferty, J., A. McCallum, and F. Pereira. "Conditional random fields: probabilistic models for segmenting and labeling sequence data." In ICML, 2001.
25. The EARS Fall 2004 Rich Transcription Evaluation Plan August 30, 2004。

鑑別性事前資訊應用於強健性語音辨識

丁川偉　吳柏樹　簡仁宗

國立成功大學資訊工程學系

{cwting, bswu, chien}@chien.csie.ncku.edu.tw

摘要

在傳統語音辨識系統中，模型的訓練環境與測試環境不匹配(mismatch)是造成辨識率下降的首要問題，在此議題上，過去文獻已提出許多解決方法，如在語音模型端引入模型參數的不確定性所建立的強健性貝氏預測分類(Bayesian predictive classification)法則，或是調整模型於測試環境的調適方法，如最大事後機率(MAP)調適以及線性迴歸(MLLR)調適，甚至進一步考慮語音模型鑑別性之最小分類錯誤線性迴歸(MCELR)調適等方法。其中，貝氏預測分類法則是將模型參數的不確定性(uncertainty)適當的引入決策法則以達到決策方法的強健性，而參數不確定性反應了雜訊環境及聲學的變異性，它可由事前機率(prior density)來表示，而傳統上貝氏學習則提供了估測並更新參數事前資訊的機制。

為兼顧決策法則的強健性及鑑別性，本論文提出在貝氏預測分類架構下聲學模型及其事前機率模型之鑑別性訓練及更新，我們使用最小分類錯誤(MCE)之鑑別性準則來估測模型參數之超參數(hyperparameter)，並且提出了兩種更新的方法，其一是直接針對隱藏式馬可夫模型平均值向量參數更新其事前統計量；其二是考慮線性迴歸調整，針對迴歸矩陣之事前資訊在最小分類錯誤準則下做更新。在以汽車噪音雜訊語音資料庫為主的評估實驗中，發現使用更新過後的事前機率可以提昇貝氏預測分類之鑑別性，達成強健性語音辨識效能提升之目的。

1. 緒論

語音是人與人之間最直接、最自然溝通的方式，隨著科技和語音辨認技術的進步，讓機械聽懂人類的話，不再是遙不可及的夢想。目前實際應用面中，語音辨識的過程仍舊存在著許多問題，最常遇到的，像是訓練環境與測試環境的不匹配問題，因為語音辨識主要是以樣本比對(pattern recognition)的技術為基礎，若是語音辨識之應用環境與原始樣本之訓練環境不匹配，將會使得辨識率大幅地降低，而這不匹配可能是來自於週遭的環境噪音、傳輸語音的通道不同、或語者不同等，影響語音辨識的因素往往是上述多個失真來源的組合。因此，為了克服語音辨識時不匹配之問題，強健且有效率的補償技術一直是語音辨識極為重要的研究議題。

在此研究領域上，已有許多學者提出不同方法來解決不匹配的問題，我們將之大致分為訊號(signal)空間、特徵參數(feature)空間、以及模型參數(model parameter)空間三類。在第一種方法中，主要以語音強化(speech enhancement)的方式為主，其觀點是將受到環境影響的訊號，透過訊號處理的方式，消減噪音的部份以得到近似乾淨的訊號；第二種方法，與訊號空間的處理觀念類似，都是希望還原原始環境下的特徵參數特性，做特徵參數的補償(compensation)；最後一種則是對已經訓練完成的模型參數做處理，其方法可再細分為兩種：其一是利用新環境所得到的少量語料

將原有之模型參數調適到與新環境接近的方式；另一則是在模型參數中考量其不確定性，以減少新環境中模型變異所造成的影響，進而達到強健性決策的機制。此外在模型的訓練當中，不同模型之間的參數或分佈常會面臨混淆的情況，造成分類錯誤的提升，因此鑑別性(discriminability)的考量也被學者提出引入模型的訓練過程，以期達到更明確之模型並降低分類的錯誤。

在本研究中，主要是在考量參數不確定性的基礎上，希望能夠在鑑別性的分類方法考量下更新其參數的不確定性，以期望進一步達成同時具有鑑別性事前機率的強健性之決策法則。另外在本研究中也將此考量不確定性且具鑑別性的事前機率學習，落實在模型參數的調整，並分為直接對模型參數的調整以及間接對模型參數做調整。在以汽車噪音為主的連續數字語料庫中，都能達到辨識效能的提升。而在本文的編排上，共分為五個小節，除了第一節為緒論外；第二節將簡單介紹模型調適、鑑別式訓練與強健性決策法則之觀念；第三節則為所提出之強健性決策下鑑別性事前資訊學習機制；最後第四、第五兩節則為實驗與結論部份。

2. 文獻探討

2.1 語音調適技術

為了解決在語者或環境所造成的不匹配的變異，學者[4][8]即提出以調適的方法來克服上述所提的問題。調適的精神即在於，以原有的語者獨立聲學模型為初始，運用少量稀疏的特定語者或新環境的調適資料，經由參數的調整，來產生一組新的聲學模型，使它更具有強健性並更適合於測試的語者或環境以達到提升辨識率的效果。調適的技術又可分為特徵向量層面(feature-based adaptation)以及模型層面(model-based adaptation)來調適，在本研究中主要是針對模型層面的調適技術來做說明。

而在模型參數的調適方法中，又可分為模型參數的直接調整與間接調整。在直接調整的技術中，最大化事後機率(maximum *a posteriori*, MAP)[4]為最基本的方法；而間接調整的方法則以線性迴歸轉換為基礎的調適演算法為主。以下將分別介紹之。

2.1.1 直接調整技術

直接調適模型參數的方法，因為是針對模型中各別的參數做調整，所以對於沒有對應調適語料的參數就沒辦法做參數的更新，其示意圖可由 MAP 演算法概念所表示：

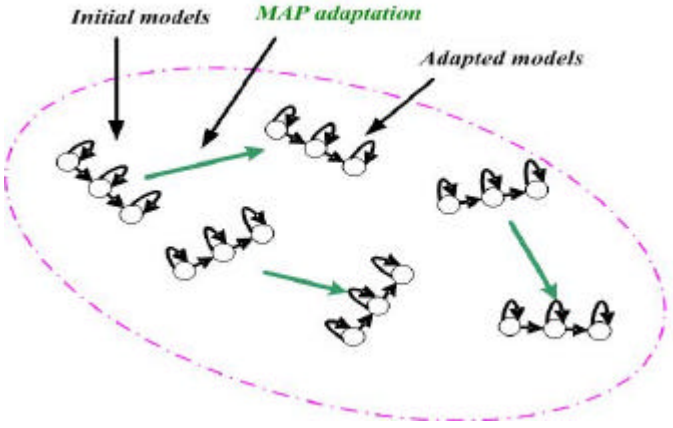


圖 一、最大化事後機率調適方法示意圖

MAP 的基本精神在於結合了事前機率以及調適語料，來估測出新的模型參數。如下式所示，X 代表觀測值，Λ_{MAP} 表示模型參數的最大機率估測值，利用貝氏定理可將事後機率拆解成事前機率與相似度的結合，$g(\Lambda \mid X)$ 及 $g(\Lambda)$ 分別表示參數的事後機率分佈及事前機率分佈，$f(X \mid \Lambda)$ 則是觀測資料 X 的相似度。

$$\Lambda_{MAP} = \arg\max_{\Lambda} g(\Lambda \mid X) = \arg\max_{\Lambda} f(X \mid \Lambda) g(\Lambda) \tag{1}$$

其中在事前機率 $g(\Lambda)$ 的部份，若不考慮事前機率 $g(\Lambda)$ 而將其設為常數函數 (non-informative prior)，則(1)式將會退化成為一般的最大相似度估測法，由此可看出最大化事後機率調適法考慮事前機率的分佈與調整語料的結合，來將事後機率進行最大化的調適。

在最大化事事後機率調適準則下，可表示調適後的模型參數如：

$$\hat{m}_{MAP} = \frac{Nt^2}{s^2 + Nt^2} \bar{\mathbf{x}} + \frac{s^2}{s^2 + Nt^2} r \tag{2}$$

其中 N 為調適語料的觀測資料數，$\bar{\mathbf{x}}$ 為調適語料的平均值，其分佈為 $N(\bar{\mathbf{x}}, s^2)$，事前機率則假設為 $N(r, t^2)$ 之機率分佈。

由此可看出，最大化事後機率調適方法結合了事前機率與調適語料所蘊含的資訊。當訊練語料足夠時，可充分將其使用到調適語料以達成良好的調整效能；而在訓練語料不足時，則是由調適語料所運算出來的模型參數平均值向量來主導，然而在此情況下，如果模型複雜度較高或包含較多的參數，則相對而言也必須使用更多的調適語料來估算出每個新的參數值，這也是為何最大化事後機率調適方法，或是直接調整模型參數的方法在調適語料數量的要求上會比較嚴格的原因。

2.1.2 間接調整技術

在間接調整模型參數的調適方法中，以線性迴歸轉換為基礎的調適演算法為主，此類方法是以轉換整個模型或是定義的相關狀態分群，所以雖然有部分模型或是狀態沒有對應的調適的語料，但是在同一個分群的，就可以分享使用同一個轉換矩陣做參數的調適轉換，因此在調適語料的數量要求便大為降低，其示意圖可由圖二所表示。

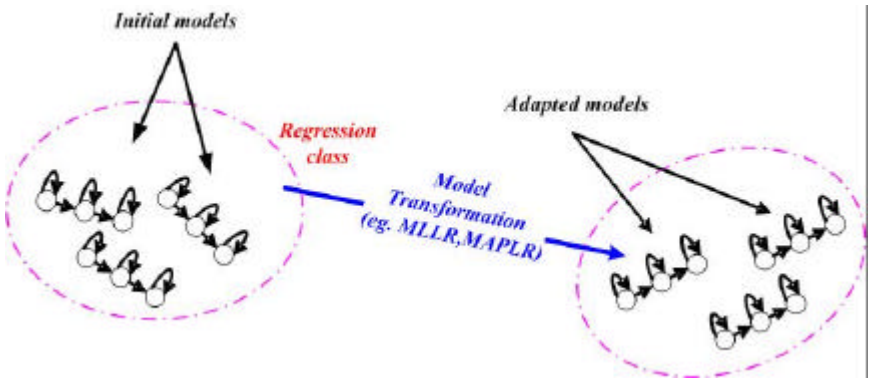


圖 二、迴歸矩陣為基礎的調適方法示意圖

以轉換方式做調適的方法中，最被廣泛使用的即為最大相似度線性迴歸(maximum likelihood linear regression, MLLR)[3][15]調適演算法。其主要為估測同一個馬可夫模型參數群聚(regression class)中的轉換，然後以此轉換對整個群聚內的參數從語者獨立轉換到特定的語者或環境條件下，所以調適語料不需要包含所有模型中的參數，只要在同一個群聚下的轉換有被估測出來，其他同一群聚下的參數就可以一併作調適 而在最大相似度的線性迴歸轉換就是在以最大化相似度的決策法則下做線性迴歸的轉換矩陣估測。

假設在以高斯分佈為狀態機率的隱藏式馬可夫模型中，模型參數 Λ_s，其平均值向量為 m_s 其維度為 $d\times 1$，而經過轉換矩陣 W_s，可以得到調適後的參數 $\hat{m}_s$，其轉換過程我們可以表示為：

$$\hat{m}_s = \mathrm{A}_s m_s + \mathrm{b}_s = \mathrm{W}_s x_s \tag{3}$$

其中，轉換除了是對於平均值向量的線性轉換外，還加上一個偏差的值，所以整體上我們可以整理出 $x_s = [1 \ m_s^T]^T$ 維度為 $(d+1)\times 1$ 且轉換矩陣為 $\mathrm{W}_s = [A_s \ b_s]$，其維度為 $d\times(d+1)$。調適後的狀態機率為：

$$b_s(x) = \frac{1}{(2p)^{d/2}\left|\Sigma_s\right|^{1/2}} \exp\left\{-1/2\left(\mathbf{x}-\mathrm{W}_{r(s)}x_s\right)^T \Sigma_s^{-1}\left(\mathbf{x}-\mathrm{W}_{r(s)}x_s\right)\right\} \tag{4}$$

在整體的訓練中，假設訓練語料為 X，且所有需要估測的轉換矩陣為對應到各別群聚類別的轉換矩陣的集合 $W = \left\{\mathrm{W}_{r(s)}\right\}$，且估測的轉換矩陣必須符合於最大化相似度的法則 $W_{ML} = \arg\max_W p(X \mid W, \Lambda)$。由於調適的模型是假設為隱藏式馬可夫模型，經由EM(expectation maximization)演算法可定義不完整資料對於最大相似度函數的期望值的輔助函數[15]

$$Q(\Lambda, \overline{\Lambda}) = \sum_q p(X, q \mid \Lambda) \log\, p(X, q \mid \overline{\Lambda}) \tag{5}$$

其中 Λ 為目前的參數估測參數值，而 $\overline{\Lambda}$ 是新的參數估測值。透過對輔助函數中之轉換矩陣參數微分並令其為零可求得轉換矩陣之更新運算式

$$w_n^{ML} = \left(\sum_t \sum_{r=1}^{R} \frac{g_{s_r}(t)}{s_{s_r,i}^2} \mathbf{x}_{t,i} x_{s_r}^T\right) \cdot \left(\sum_t \sum_{r=1}^{R} \frac{g_{s_r}(t)}{s_{s_r,i}^2} x_{s_r} x_{s_r}^T\right) \tag{6}$$

2.2 鑑別性訓練法則

在以鑑別性為主的估測方法，最著名的為最小分類錯誤(minimum classification error, MCE)[13]、最大互斥資訊(maximum mutual information, MMI)[16]等方法為主，近年來又有以支持向量機(Support Vector Machine, SVM)的精神為出發點的最大化邊界(Large Margin)[9]分類法則，主要是以求出使得辨識效果最差的邊界，然後以更新模型參數來拉大此分類錯誤的邊界，以找出最佳的分類邊界。以下我們簡單介紹最小分類錯誤之觀念與訓練方式。最小分類錯誤的鑑別性訓練法

則，是由 Juang et al.[14]所提出來的，在此鑑別性訓練方法中主要分為三個部份，一是定義其鑑別性函數間的錯誤分類量測(misclassification measure)，其二是利用損失函數(loss function)來表示其分類正確與錯誤率，第三步驟是在最小化期望損失(expected loss)的目標下估測模型參數。在統計型語音辨識中，鑑別性函數主要以相似度函數(likelihood function)來代表，而參數估測上主要是以最小化訓練觀測資料的期望的損失為目標，來對應到最小化辨識錯誤率的關係。以下我們開始介紹最小錯誤分類法則中的三個步驟：

1) 錯誤分類量測定義為：

$$d_i(X) = -g_i(X;\Lambda) + \left[\frac{1}{M-1}\sum_{j,j\neq k}\exp\left\{\eta g_j(X;\Lambda)\right\}\right]^{\frac{1}{\eta}} \tag{7}$$

其量測主要是在於辨識的錯誤率，C_j 為其混淆類別(confusing classes)或稱競爭類別(competing classes)，$g_i(X;\Lambda)$ 是觀測語料 X 對應到正確類別 C_i 的辨識分數，$\left[\frac{1}{M-1}\sum_{j,j\neq k}\exp\left\{\eta g_j(X;\Lambda)\right\}\right]^{\frac{1}{\eta}}$ 是對應到其他非 C_i 的競爭類別對應到觀測語料 X 的辨識分數，中括號內容的項是對應到 L^{η} norm，當 $d_k(X)>0$ 代表發生分類錯誤，$d_k(X)\le 0$ 代表正確分類。 其中η 是一個正數，改變η 及 M 的值，可以改變(7)中具影響力的競爭類別數量，當 $\eta \to \infty$，則對應到錯誤辨識的分數為競爭類別中最高的類別：

$$d_i(X) = -g_i(X;\Lambda) + \max_{j,j\neq i} g_j(X;\Lambda) \tag{8}$$

此處類別 C_j 是除了類別 C_i 外，和觀察資料 X 相似度最大的競爭類別。

2) 對於某個觀察資料 X，以損失函數來定義分類器的分類風險，能進一步的表示錯誤分類量測與辨識錯誤率的關係，把錯誤分類量測代入由 sigmoid function 所近似的 0-1 損失函數，即

$$l_i(X;\Lambda) = l\left(d_i(X)\right) \tag{9}$$

其中 sigmoid function 是一個對應到值域範圍為[0,1]的連續性函數且具有可微分的特性，sigmoid function 定義如下 $l_i(d_i) = \frac{1}{1+\exp(-\gamma d_i + \theta)}$，其中$\theta$ 常設為 0，而γ 常設為 1 或大於 1，另外因為當 $d_i(X)<0$ 時，代表分類正確，所以 $l_i(d_i)$ 會相對於接近於零，代表沒有辨識錯誤的損失；而當 $d_i(X)>0$ 時，則 $l_i(d_i)$ 會大於零，所以也代表對於 X 有辨識錯誤的損失。

3) 在參數估測上，主要就是尋找能最小化所有觀測資料的期望損失的模型參數，對應於一個觀測資料，其決策上的損失為

$$\ell(X;\Lambda)=\sum_{i=1}^{M} l_i(X;\Lambda)1(X\in C_i) \tag{10}$$

其中$1(\cdot)$為 indicator function，當條件符合時值為 1，否則為 0。在考量所有觀測資料下，整體期望損失為$\ell(\Lambda)=E_X[l(X;\Lambda)]$，最後，即可利用廣義機率遞減演算法(generalized probabilistic decent, GPD)進行疊代運算以實現 MCE 法則[13]。

$$\Lambda_{t+1}=\Lambda_t-e_tU_t\nabla\ell(X_t,\Lambda)\Big|_{\Lambda=\Lambda_t} \tag{11}$$

2.3 強健性決策法則

以統計為主的決策法則，要實現最佳貝氏分類器必須滿足三點：假設觀測資料的空間Ω_x是已知的、其損失函數是給定的、以及假設真正的機率分佈是已知的。但是在實際應用中建構貝氏決策法則時，我們只有有限的訓練資料，並不能代表整體的觀測資料，而且資料對應的真實分佈並非事前可以得知。為了數學上及計算上方便，真實的分佈往往假設為特定的機率分佈，如高斯分佈，也因為現實的情況與理想上不符合，所以在現實狀況中，設計的決策法則會有 1)少量樣本的失真；2)模型假設的失真；3)測試環境訊號的失真；三種假設錯誤所帶來的失真，導致最佳化的決策法則是無法得到，而其中以第三種測試環境與訓練環境不匹配的問題對語音辨識的衝擊最為嚴重。因此強健性的決策法則被提出，其主要是強調當訓練與測試時具有不一致性的條件差異時，仍然能夠保有一定品質的辨識效率，且能減輕在不匹配的情況下辨識效率的損失。所以如何加強語音辨識系統的強健性是近年來重的要研究議題之一。在強健性的決策法則中主要是考慮參數具有不確定性範圍的隨機性(randomness)而非特定的(deterministic)，其中以貝氏預測分類法則(Bayesian predictive classification, BPC) [6]為基礎的方法會在以下說明。

在貝氏預測決策法則[5][10][11]，是在事前假設的不確定性分佈下以平均的方式考慮所有不確定性中的參數點對決策的影響力，而其參數的不確定性也可以視為其參數的事前資訊。所以此分類器主要是考慮所有可能性的參數點對於給定資料的期望值做為決策的依據 在貝氏預測分類器中，我們首先給定模型參數(Λ,Γ)的事前機率，$p(\Lambda,\Gamma\mid j_\Lambda,j_\Gamma)$視為參數的不確定性範圍，且聲學模型參數$\Lambda$和語言模型參數$\Gamma$各自存在於不確定性的範圍$\Omega_\Lambda$和$\Omega_\Gamma$中，當進一步假設聲學模型及語言模型間的獨立性，$p(\Lambda,\Gamma\mid j_\Lambda,j_\Gamma)=p(\Lambda\mid j_\Lambda)\cdot p(\Gamma\mid j_\Gamma)$，所以，其分類器的不確定性是落在參數事前機率分佈下的一個範圍：

$$M_e^*=\{p_\Lambda(X\mid W),p_\Gamma(W)\mid(\Lambda,\Gamma)\sim p(\Lambda,\Gamma\mid j_\Lambda,j_\Gamma);\Lambda\in\Omega_\Lambda,\Gamma\in\Omega_\Gamma\} \tag{12}$$

其模型參數之不確定性可由圖三示意之。

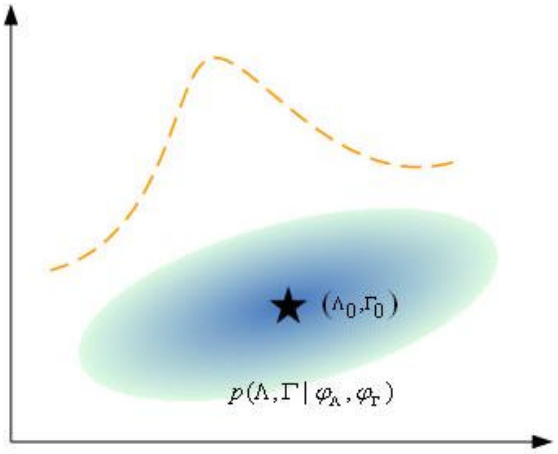


圖 三、貝氏預測法則參數不確定性示意圖

當進一步考慮所有的訓練語料及對應的答案 (X,W) **與給定參數的事前機率** $p(\Lambda,\Gamma\mid j_\Lambda,j_\Gamma)$ **和損失函數**(loss function) $l(W,W')$ **，則其對應的整體決策風險為：**

$$\begin{aligned}\tilde{r}(d(\cdot)) &= E_{(W,X)}E_{(\Lambda,\Gamma)}[l(W,d(X))] \\ &= \sum_{W\in\Omega_W}\int_{X\in\Omega_W} l(W,d(X))\tilde{p}(X\mid W)\tilde{p}(W)dX\end{aligned} \tag{13}$$

其中 $\tilde{p}(X\mid W)$ **為對應到聲學模型的預測分佈密度函數** $\tilde{p}(X\mid W)=\int_{\Omega_\Lambda} p(X\mid W,\Lambda)\,p(\Lambda\mid j_\Lambda)d\Lambda$ **；** $\tilde{p}(W)$ **為其對應到語言模型的預測分佈密度函數** $\tilde{p}(W)=\int_{\Omega_\Gamma} p(W\mid\Gamma)\,p(\Gamma\mid j_\Gamma)d\Gamma$ **，當上式為** 0-1 **的損失函數時，使得上述** $\tilde{r}(d(\cdot))$ **整體風險最小的分類器為：**

$$\tilde{d}(X)=\arg\max_W \tilde{P}(W,X)=\arg\max \tilde{P}(X\mid W)\cdot\tilde{P}(W) \tag{14}$$

就是對應到最到最佳的貝氏預測分類器，而其與嵌入式最大事後機率決策法則最大的不同是，其在辨識時考慮了模型參數的不確定性，而非只是單一的模型參數點。

在貝氏預測分類器的研究上，除了在模型的參數上假設其不確定性外，還有假設聲學模型的不匹配可以透過空間轉換的關係做補償[1][2]**，並且假設模型參數間轉換函數的參數具有不確定性** $M_e^*=\left\{p_\Lambda(X\mid W)\mid\Lambda=V_{J_\Lambda}(\Lambda_0),J_\Lambda\sim p(J_\Lambda)\right\}$ **，其對應的預測密度函數為：**

$$\tilde{p}(X\mid W)=\int_{\Omega_\Lambda} p(X\mid W,\Lambda=V_{J_\Lambda}(\Lambda_0))\,p(J_\Lambda)d\Lambda \tag{15}$$

其中 $V_{J_\Lambda}(\Lambda_0)$ **是模型參數的轉換函數，而此轉換函數是由** J_Λ **所控制。在此類以轉換函數為主的事前機率，主要是在於轉換模型參數的轉換函數其調整參數上。而在此類分類器中，在設計上必須要先假設事前機率的分佈，以及給定事前機率分佈的超參數，以及如何去計算預測密度函數等問題，而前兩個問題是在實作中假設的問題，而第三個問題是如何近似預測密度函數的問題。**

在 BPC **的決策上，當辨識的模型為隱藏式馬可夫模型時，且狀態機率是由混合高斯模型所**

組成，在求解貝氏預測密度函數會包含有狀態序列(state sequence)以及高斯的混合索引(mixture index)的遺失資料(missing data)問題，因此要對 $\tilde{p}(X \mid W)$ 完整的計算其預測密度函數是複雜度十分高且不可實現的。所以在考慮其簡化的近似方法上是一個研究議題，在文獻中，主要有近似貝氏預測分類器(quasi-Bayes predictive classification, QBPC)、維特比貝氏預測分類器(Viterbi Bayesian predictive classification, VBPC)、以及貝氏預測模型補償(Bayesian predictive density based model compensation, BP-MC)三種方式來近似計算 BPC。為考量在實作上與傳統語音模型作最具效率的連結，本研究採用 BP-MC 的方式近似 BPC 分數。

3. 鑑別性事前資訊訓練方法

在語音辨識的系統中，總是存在著很多不同於訓練環境中的影響，其有可能是來自於環境因素的差異，或者是語者發音特性上不同的差異，甚至於訓練或調適語料不足等等在樣本不足上的問題，這些因素均為降低語音辨識器效能的重要因素之一。有鑑於此，在解決這方面的問題上有在參數訊號端做補償的方法，如隨機向量補償或是鑑別性特徵參數擷取；或是在模型端的補償方法，如第二節中所提的強健性決策方法及調適方法。在貝氏預測分類器中，文獻中主要是以貝氏學習的方法來更新其不確定的事前資訊，此外在文獻中，也有學者提出在隨機向量補償的鑑別性不確定估測方法。在此研究中，我們主要是考慮在語音模型端的不確定性事前資訊的鑑別性調整方法，並且同時考慮了在語音模型參數的不確定性，以及調適轉換矩陣兩種不確定性的鑑別性調整方法。

在我們提出的方法上，以物理意義而言，當在不穩定的環境中估測參數，其一定會隱藏很多估測上的錯誤及失真，所以若不考慮參數的隨機性，而在模型參數上以鑑別性的參數估測方式求出對於訓練資料的最佳分類決策參數，雖然能夠表達出對於訓練或是調整語料的鑑別性，但卻未必能對抗存在於測試環境中參數隨機性所造成的失真影響。所以我們會以考慮參數不確定性的強健性方法，來模擬出此不匹配現象下的參數隨機性。極端而言，對於單一的模型，其不確定性的範圍最好能含蓋到所有存在此模型參數的隨機性空間越好，所以對於不確定性的估測和鑑別性的估測，存在有觀念上互相抗衡之處，因此在估測不確定性及加強鑑別性之間存在於一種抉擇(trade off)的關係。

但是[7]指出，雖然不確定性能夠在存在不確定性的環境中展現出其模型的強健性，當選取的不確定性越大，則相對的會造成不同模型間因為其不確定性有重疊而造成分類決策上的混淆，所以進一步的文獻[1][2][12]中，以貝氏學習的方法在分佈估測(distribution estimation)的精神下，學習對於每個模型而言最能表達其不確定性的事前資訊分佈，然而在貝氏學習訓練下的不確定性也仍舊可能存在於與臨近類別間混淆的情況。所以我們會考慮在不確定性事前資訊上考慮其鑑別性的更新，希望此不確定性的更新除了能夠描述整體的參數不確定性外，也能夠減低與其他類別參數不確定性間的混淆所造成的辨識率下滑，以達成具鑑別性的強健性決策法則的效果。我們並將此觀點落實在模型參數調整的技術當中。

3.1 直接調整

在此方法中，我們主要是以最小分類錯誤來實現我們的事前資訊鑑別性方法，不同於模型參數更

新的方法，我們考量了模型參數的不確定性，而且其不確定性是由一個事前機率密度函數 $p(\Lambda \mid j, W)$ 來表示，且其事前機率密度函數可以是由事前的訓練語料，或是對於測試環境的不確定性有所了解而估測出的事前分佈，而對應於考慮模型參數的不確定性的貝氏預估分類器可以表示為：

$$\tilde{p}(X \mid W) = \log \int_{\Omega} p(X \mid \Lambda, W) p(\Lambda \mid j, W) d\Lambda \tag{16}$$

當我們假設模型為隱藏式馬可夫模型時，其參數為 $\Lambda = \{a_{ij}, c_{ik}, m_{ik}, \Sigma_{ik}, i, j = 1,....,M, k = 1,...K\}$，此模型中共有 M 個狀態，其中 a_{ij} 為狀態轉移機率，每一個狀態內是有 K 個混合數的高斯混合模型，其中 c_{ik} 為第 i 個狀態的第 k 個高斯模型的混合索引，m_{ik} 及 $,\Sigma_{ik}$ 為對應到第 i 個狀態第 k 個高斯模型的平均值向量及共變異矩陣。考量 BPC 近似方法，當我們要計算(16)時，因為會遭遇到狀態序列及混合索引等遺失資料的問題

$$\begin{aligned} \tilde{p}(X \mid W) &= \int_{\Omega} p(X \mid \Lambda, W) p(\Lambda \mid j, W) d\Lambda \\ &= \sum_{s,l} \int_{\Omega} p(X, s, l \mid \Lambda, W) p(\Lambda \mid j, W) d\Lambda \end{aligned} \tag{17}$$

所以我們進一步以音框為單位的補償方法來近似此貝氏預測分類器，在此論文中，我們假設只考慮隱藏式馬可夫模型的平均值向量 m_{ik} 的不確定性，其他參數則視為固定不改變。而且假設一個模型中的每一個狀態均存在一個平均值向量，其不確定性事前機率為可定義為如下之高斯分佈

$$\begin{aligned} p(m_{ik} \mid j_{ik}, W) &= \frac{1}{\sqrt{2p}\,|t_{ik}|^{d/2}} \exp \frac{-1}{2} \left\{ (m_{ik} - m_{ik})^t t_{ik}^{-1} (m_{ik} - m_{ik}) \right\} \\ &= N(m_{ik}; m_{ik}, t_{ik}) \end{aligned} \tag{18}$$

其中 $j_{ik} = \{m_{ik}, t_{ik}\}$ 為第 i 個狀態第 k 個混合高斯平均值向量 m_{ik} 的不確定性事前機率超參數。把不確定的影響以音框為單位作補償並且假設其共變異矩陣為對角化矩陣 $\Sigma_{ik} = diag\{s_{ikd}^2\}$ 且平均值矩陣的共變異矩陣超參數也為對角化矩陣 $t_{ik} = diag\{t_{ikd}^2\}$ 時，所以每一個狀態內的每一個混合高斯其貝氏預測密度函數可以分維度各別計算為

$$\begin{aligned} \tilde{f}_{ikd}(x_d) &= \int f(x_d \mid q_{ikd}) p(q_{ikd} \mid j_{ikd}) dm_{ikd} \\ &= N(x_d, m_{ikd}, g_{ikd}^2) \end{aligned} \tag{19}$$

其中 $g_{ikd}^2 = s_{ikd}^2 + t_{ikd}^2$，且對於上式的各別狀態中個別補償過後的高斯混合模型為 $\tilde{p}_i(\mathbf{x}) = \sum_{k=1}^{K} w_{ik} \tilde{f}_{ik}(\mathbf{x})$。在導入鑑別性決策法則上，在此我們的鑑別性函數是定義為貝氏預測密度函數，當給定觀測資料為 $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)$ 後其鑑別性函數可以寫為下式：

$$\begin{aligned} \tilde{g}(X; W, l, j_q) &= \log \tilde{P}_R(X \mid W) = \log \left[p_{\hat{s}_0} \prod_t a_{\hat{s}_{t-1}\hat{s}_t} w_{\hat{s}_t\hat{l}_t} \tilde{f}_{\hat{s}_t\hat{l}_t}(\mathbf{x}_t) \right] \\ &= A_{\hat{s}\hat{l}}^{*} + \sum_t \sum_d \log w_{\hat{s}_t\hat{l}_t} \tilde{f}_{\hat{s}_t\hat{l}_t}(x_{td}) \end{aligned} \tag{20}$$

其中假設 $\hat{s}$ 及 $\hat{l}$ 為對應於正確字串的最佳狀態序列及最佳混合索引序列。且在最小分類錯誤決策法則中，可定義出分類錯誤量測與對應於超參數的期望損失函數 $\ell(X;j)=\frac{1}{M}\sum_{i=1}^{M} l_i(X;j)$，使用廣義梯度遞減的方法對模型平均值向量的超參數做參數更新，對於微分量先做符合於參數限制的事前轉換 $m_{ikd} \rightarrow \tilde{m}_{ikd}=m_{ikd}/g_{ikd}$，$t_{ikd} \rightarrow \tilde{t}_{ikd}=\log t_{ikd}$ 後，對於 $\tilde{m}_{ikd}(n+1)=\tilde{m}_{ikd}(n)-e_n \left.\frac{\partial l_i(X_n;j)}{\partial \tilde{m}_{ikd}}\right|_{\tilde{m}_{ikd}=\tilde{m}_{ikd}(n)}$，則對應於超參數的平均值微分量為：

$$\frac{\partial l_i(X_n;j)}{\partial \tilde{m}_{ikd}}=-g l_i(d_i)(1-l_i(d_i))\sum_{t=1}^{T}\delta(\hat{s}_t-i)\delta(\hat{l}_t-k)\left(\frac{x_{td}}{g_{ikd}}-\tilde{m}_{ikd}\right) \tag{21}$$

其中 $\delta(\cdot)$ 是一個 Kronecker delta function 同理於對超參數的變異數更新量 $\tilde{t}_{ikd}(n+1)=\tilde{t}_{ikd}(n)-e_n \left.\frac{\partial l_i(X_n;j)}{\partial \tilde{t}_{ikd}}\right|_{\tilde{t}_{ikd}=\tilde{t}_{ikd}(n)}$ 其對應的微分量推導為：

$$\frac{\partial l_i(X_n;j)}{\partial \tilde{t}_{ikd}}=-g l_i(d_i)(1-l_i(d_i))\sum_{t=1}^{T}\delta(\hat{s}_t-i)\delta(\hat{l}_t-k)\left[\left(\frac{x_{td}-m_{ikd}}{g_{ikd}}\right)^2-1\right]\left(\frac{t_{ikd}^2}{g_{ikd}^2}\right) \tag{22}$$

，並且對上式之微分量參數轉回原參數空間即可完成鑑別性的超參數更新法則。

得到具有不確定性資訊補償的模型後，進一步，我們在最小分類誤差的鑑別性法則下，給定部分的調整語料來對其不確定性的事前資訊分佈做調整，所以此調整出來的模型會是在以鑑別性為主軸下所調整出來且不同於以貝氏更新方法的貝氏預估分類器，而且因為我們主要是針對於參數的不確定性資訊做調整，所以進一步來說，我們的方法是以強化模型間的鑑別性下的模型參數的不確定性調整，在此研究中我們將此方法命名為 DBPC(discriminative BPC)。

3.2 間接調整

在環境的不匹配狀況下除了以直接對模型參數的調整方法外，還有以間接的調整方法，如以迴歸矩陣為基礎的方法，而模型參數的直接調適，可以細部的針對每一個狀態或是混合高斯做調整，但是當我們的調整語料有所不足時，直接調適則可能遭遇資料量不完全或資料量不足的問題而使得調整效能無法充分彰顯，所以另一個方法是以群聚的方式用轉換的概念做批次調整，這樣可以彌補單獨調整時資料不完整或是緩和調適資料量不足的問題，然而在以轉換矩陣為主的調整方法中，也主要是以特定的(deterministic)的轉換矩陣方法做調整，但是當考慮環境變異性及少量調整資料的問題情況下，轉換矩陣也有其不確定性，而在 LRBPC(linear regression based BPC)中首度以考量轉換矩陣也存在有不確定性的情況下，達成一個具有考慮了轉換矩陣不確定性的決策法則，進一步的，我們是假設當轉換矩陣的不確定性在調整上是以鑑別性的條件下來做學習，以達成在模型參數的轉換中經由鑑別性的轉換矩陣不確定性調整，而達成轉換後的參數也具有鑑別性的不確定行映射。

在 MLLR 中，其主要是找出一組特定的轉換矩陣 $\hat{R}$ 然後透過此轉換矩陣來做參數的轉換以達成調適的目的，所以其決策法則主要是給定原來的模型參數及轉換矩陣下轉換參數決策：

$$\hat{W} = \arg\max_{W} P(X \mid W, \Lambda, \hat{R}) \tag{23}$$

但是當我們進一步考慮其轉換矩陣 $g(R \mid j)$ 也有其不確定資訊時，我們可以線性迴歸貝氏預測分類器的方法平均化的考慮轉換矩陣的不確定性資訊：

$$\tilde{P}_R(X \mid W, \Lambda) = \int p(X \mid W, R, \Lambda) g(R \mid j) dR \tag{24}$$

所以在此我們可以將其線性迴歸貝氏預測分類器，取代(24)中以嵌入(plug-in)特定迴歸矩陣的方法來表示：

$$\hat{W} = \arg\max_{W} \tilde{P}_R(X \mid W, \Lambda) \tag{25}$$

所以針對第 i 個狀態第 k 個混合高斯的平均值轉換可以表示為 $\hat{m}_{ik} = A_c m_{ik} + B_c = R_c x_{ik}$，其中下標 c 代表的是其高斯所對應到的群聚類別(regression class)，且轉換矩陣 $R_c = [B_c \;\; A_c]$ 且延伸的平均值向量為 $x_{ik} = [1 \;\; m_{ik}^T]^T$，當我們進一步假設其轉換矩陣 R_c 為單變數的迴歸矩陣時，則其中 A_c 矩陣為對角化的矩陣 $A_c = diag\{a_{cl}\}$，則代表每個維度的轉換可以各自計算，所以對應到第 d 維的平均值可以表示為 $\hat{m}_{ikd} = a_{cd} m_{ikd} + b_c$ [15]。在此簡化的轉換矩陣中我們另外定義其另定義 $q_c = [a_{c1}, ..., a_{cd}, b_{c1}, ..., b_{cd}]^T$，且考慮以音框為單位的補償型式則

$$\tilde{f}_{ik}(\mathbf{x}_t) = \int f(\mathbf{x}_t \mid q_c, m_{ik}, \Sigma_{ik}) g(q_c \mid j_c) dq_c \tag{26}$$

其中 $f_{ik}(\mathbf{x}_t)$ 為原來 HMM 模型中第 i 個狀態第 k 個混合高斯，而 $\tilde{f}_{ik}(\mathbf{x}_t)$ 為經過以音框為單位的補償後的機率模型，當 HMM 的共變異矩陣為對角化矩陣時 $\Sigma_{ik} = diag\{s_{ikd}^2\}$，則 $f_{ik}(\mathbf{x}_t)$ 可以拆開為每個維度各別考慮，其中 $q_{cd} = [a_{cd} \;\; b_{cd}]^T$。假設 $g(q_{cd} \mid j_{cd})$ 為高斯分佈且不同維度間的轉換參數假設為獨立，則對應到第 d 維的轉換矩陣參數可以表示 $j_{cd} = \{m_{q_{cd}}, \Sigma_{q_{cd}}\}$，所以對應到第 d 維的轉換參數事前機率為：

$$\begin{aligned} g(q_{cd} \mid j_{cd}) &= g(a_{cd}, b_{cd} \mid m_{q_{cd}}, \Sigma_{q_{cd}}) \\ &= \frac{1}{2p} \left|\Sigma_{q_{cd}}\right|^{-1/2} \exp\left\{\frac{-1}{2}\left(q_{cd} - m_{q_{cd}}\right)^T \Sigma_{q_{cd}}^{-1} \left(q_{cd} - m_{q_{cd}}\right)\right\} \end{aligned} \tag{27}$$

其中 $m_{q_{cd}} = \left[m_{a_{cd}} \;\; m_{b_{cd}}\right]^T$ 為第 c 個轉換矩陣的第 d 維的參數的平均值超參數，而

$$\Sigma_{q_{cd}} = \begin{bmatrix} s_{a_{cd}}^2 & s_{a_{cd}b_{cd}}^2 \\ s_{a_{cd}b_{cd}}^2 & s_{b_{cd}}^2 \end{bmatrix} \tag{28}$$

為此維度下轉換參數的共變異矩陣。將(27)式帶回其(26)式整理其積分後可以得到，其線性迴歸

貝氏預測密度函數為：

$$\tilde{f}_{ik}(x_{td}) = N\left(x_{td}; \hat{m}_{x,ikd}, \hat{s}^{2}_{x,ikd}\right) \tag{29}$$

$$\hat{s}^{2}_{x,ikd} = s^{2}_{b_{cd}}\left(1+\frac{s^{2}_{a_{cd}}}{s^{2}_{b_{cd}}}\right)^{2} + m^{2}_{ikd}\left(s^{2}_{a_{cd}} - \frac{s^{4}_{a_{cd}b_{cd}}}{s^{2}_{b_{cd}}}\right) + s^{2}_{ikd} \tag{30}$$

$$\hat{m}_{x,ikd} = m_{a_{cd}} m_{ikd} + m_{b_{cd}} \tag{31}$$

在導入鑑別性決策法則上，同前子小節，鑑別性函數為

$$\begin{aligned}\tilde{g}(X;W,\Lambda,j_q) &= \log \tilde{P}_R(X \mid W,\Lambda) = \log\left[p_{\hat{s}_0}\prod_t a_{\hat{s}_{t-1}\hat{s}_t} w_{\hat{s}_t\hat{l}_t} f_{\hat{s}_t\hat{l}_t}(\mathbf{x}_t)\right] \\ &= A^{*}_{\hat{s}\hat{l}} + \sum_t\sum_d \log f_{\hat{s}_t\hat{l}_t}(x_{td})\end{aligned} \tag{32}$$

其中 $A^{*}_{\hat{s}\hat{l}}$ 為最佳狀態序列及最佳混合數序列所對應的狀態轉移機率的分數。則所對應的分類錯誤量測函數 $d(X;\Lambda,j_q) = -\tilde{g}(X;W_i,\Lambda,j_q) + \tilde{G}(X;\Lambda,j_q)$ 同樣代入 sigmoid function 函數成為損失函數後，利用廣義梯度遞減的方法可對於轉換矩陣參數的不確定性，並且對於參數做符合限制的預先轉換 $j \rightarrow \tilde{j}_c$ 則超參數的更新通式為

$$j_c(n+1) = j_c(n) - e_n \left.\frac{\partial l(X_n;\Lambda,j_q)}{\partial j_c}\right|_{j_c = j_c(n)} \tag{33}$$

上式 j_c 為對應到第 c 個迴歸矩陣的超參數，另外分維度各別微分 $j_{cd} = \{m_{q_{cd}}, \Sigma_{q_{cd}}\}$ 可得

$$\frac{\partial l(X_i;\Lambda,j_q)}{\partial j_{cd}} = a\,l(d)(1-l(d))\left\{-\frac{\partial \tilde{g}(X;W_i,\Lambda,j_q)}{\partial j_{cd}} + \frac{\partial \tilde{G}(X;\Lambda,j_q)}{\partial j_{cd}}\right\} \tag{34}$$

定義 $\Omega_c = \{i,k\}$ 為屬於第 c 個群聚類別的狀態混合高斯索引，然後對於鑑別性函數中的迴歸矩陣的超參數平均值及超參數偏差平均值微分量為

$$\frac{\partial \tilde{g}(X;W_i,\Lambda,j_q)}{\partial \tilde{m}_{a_{cd}}} = -\sum_t\sum_{ik\in\Omega_c} V_t(i,k)\left[\left(\frac{m_{ikd}m_{a_{cd}} + m_{b_{cd}} - x_{td}}{\hat{s}_{x,ikd}}\right)m_{ikd}\right] \tag{35}$$

$$\frac{\partial \tilde{g}(X;W_i,\Lambda,j_q)}{\partial \tilde{m}_{b_{cd}}} = -\sum_t\sum_{ik\in\Omega_c} V_t(i,k)\left[\frac{\left(m_{ikd}m_{a_{cd}} + m_{b_{cd}} - x_{td}\right)}{\hat{s}_{x,ikd}}\right] \tag{36}$$

轉換矩陣的超參數變異數及偏差值超參數變異數的微分量可以更新如下

$$\frac{\partial \tilde{g}(X;W_i,\Lambda,j_q)}{\partial \tilde{s}_{a_{cd}}} = \sum_t\sum_{ik\in\Omega_c} V_t(i,k)\left[\frac{\left(u_{ikd}m_{q_{cd}} - x_{td}\right)^2}{\hat{s}^{2}_{x,ikd}} - 1\right]\cdot\frac{m^{2}_{ikd}s^{2}_{a_{cd}}}{\hat{s}^{2}_{x,ikd}} \tag{37}$$

$$\frac{\partial \tilde{g}(X;W_i,\Lambda,j_q)}{\partial \tilde{s}_{b_{cd}}} = \sum_t\sum_{i,k\in\Omega_c} V_t(i,k)\left[\frac{\left(u_{ikd}m_{q_{cd}} - x_{td}\right)^2}{\hat{s}^{2}_{x,ikd}} - 1\right]\cdot\frac{s^{2}_{b_{cd}}}{\hat{s}^{2}_{x,ikd}} \tag{38}$$

其中 $V_t(i,k)$ 為對應到時間點 t 第 i 個狀態第 k 個高斯發生的事後機率且 $u_{ikd} = [m_{ikd}\ \ 1]$，在以

使用 Viterbi 演算法做最佳狀態及混合數序列的取代時$V_t(i,k)$可以用 Kronecker delta function 取代，並且將對應的前處理轉回原參數空間，所以在此部分，我們提出一個以鑑別性法則來更新其迴歸矩陣超參數的統計量，並命名為 DLRBPC(discriminative LRBPC)，以達成依據給定的調適語料鑑別性的學習其轉換矩陣不確性的範圍。

4. 實驗

4.1 語料庫

連續數字語音資料庫

訓練語料庫共有 1000 句中文連續數字，包括 50 位男生、 50 位女生，每個人發音 10 句，每句語音長度為 3 至 11 個數字長，所有語料皆為麥克風錄音，語音訊號取樣頻率為 8kHZ，語音訊號量化析度為 16 位元。另外我們使用汽車噪音環境下之語料庫(CARNAV98)中的五十公里噪音之語料作為測試語料。其中共有 10 位語者，每人各錄製十五句語料。

汽車噪音語料庫

此組語料庫為實際汽車環境下以遠距離麥克風方式錄得, 如表二所示, 此組語料包含 5 位男生 5 位女生發音的中文連續數字語料, 每人分別在車速 0 公里（怠速路況）錄製 10 句，50 公里（市區路況）錄製 20 句，90 公里（高速公路路況）錄製 30 句，其中每位語者在不同路況下取出 5 句做為調整語料，其餘的做為測試語料，在五十公里語料中有 150 句測試語料，在九十公里語料中有 250 測試語料。所使用的汽車為 TOYOTA COROLLA 1.8（錄製 2 男 2 女）和 YULON SENTRA 1.6 取樣頻率均為 8 k Hz ，以 16 bit 的方式儲存，連續數字長度為 3 至 11 個數字長。其中每種不同速度下的訊號雜訊比為：

表 一、不同環境語料之信號雜訊比

信號雜訊比 SNR（dB）			
	YULON	TOYOTA	平均
0 公里	5.63	10.3	7.96
50 公里	-6.53	0.34	-3.1
90 公里	-10.14	-3.77	-6.96

4.2 語音模型與參數設定

實驗主要是以連續密度隱藏式馬可夫模型(continuous density hidden Markov model, CDHMM)做為 baseline 系統的架構，在辨識噪音連續數字語料時，由於每個狀態所分配到的音框數不固定，所以不限定混合數數量，由程式自動產生，但最大混合數限制為 4 個。在連續數字 HMM 的定義中，每個數字模型固定為 7 個狀態，並且加入語句前後及中間三個靜音狀態，總狀態個數為 73 個。在語音特徵參數求取部份，每個特徵參數皆為 26 維度，其中包括了 12 階的 MFCC，12 階

的 delta MFCC ，1 階的 log energy 以及 1 階的 delta log energy。

4.3 實驗結果

在實驗的部分，我們分別討論 DBPC 及 DLRBPC 的實驗評估在汽車噪音語料庫下的方法評估。在此實驗，我們主要為評估在乾淨環境下所訓練的模型在 50 公里及 90 公里汽車噪下的辨識率，我們首先列出乾淨語料訓練的初始模型在未調適前對於噪音語料的辨識率如下：

表 二、測試環境對於模型調適前的辨識率

語料環境	50km	90km
辨識率	39.21%	32.89%

- **DBPC 之評估實驗**

在此實驗中，初始的語者獨立模型是來自 1000 句的乾淨語料，我們主要是針對環境不匹配下的調適。並且提供 5 至 15 句的調整語料，並且以 2 句為單位。在此實驗中，我們主要是以 MAP 調整及 BPC 的不確定性調整下和我們提出的以最小化分類錯誤法則更新不確定性的方法做比較，以下分別是五十公里與九十公里的汽車噪音環境下的辨識效果。

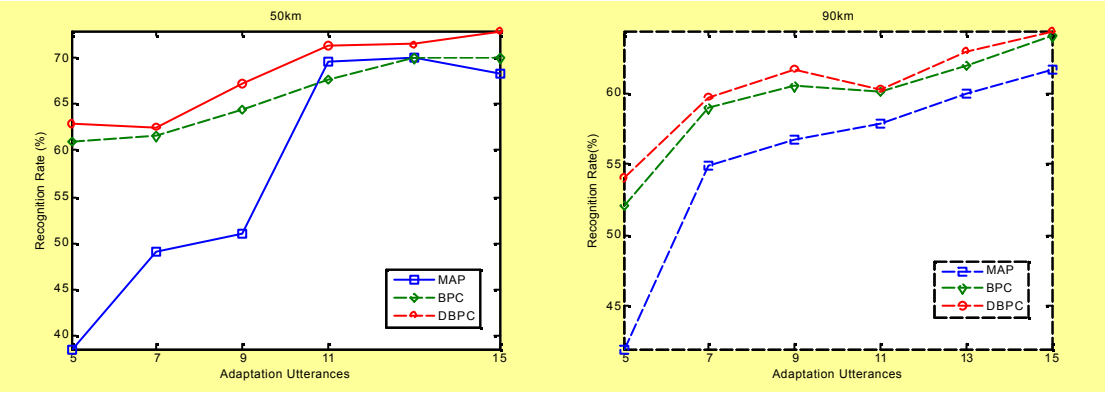


圖 四、DBPC 在五十公里與九十公里噪音語料調適結果

其中發現，雖然有考慮貝氏更新的 BPC 可以在調整資料量不足，以及環境不穩定下表現出相較於的最大事後機率調整的強健性，但是當進一步考慮鑑別性的更新不確定性時，則可以再些微的提升整體的辨識率。

- **DLRBPC 之評估實驗**

在此實驗中，我們首先以 1000 句的連續數字乾淨語料來訓練出一組語者獨立的模型，為了能更實驗在不匹配的環境中調適的效率，我們分別在 50 公里及 90 公里的噪音語料下做辨識，辨識前我們各取其環境下一至五句的語料當成是調適語料，在方法評估上，我們與 MLLR 及 MCELR 和我們提出的鑑別性迴歸貝氏預測分類器做比較。在我們方法中的事前資訊，是以汽車噪音中的 0 公里的 50 句調適語料分 10 位語者來估測出語者相依迴歸矩陣，然後取其樣本平均值及樣本共變異數當成初始的不確定性的事前資訊。以下是在五十公里與九十公里下的調適實驗結果：

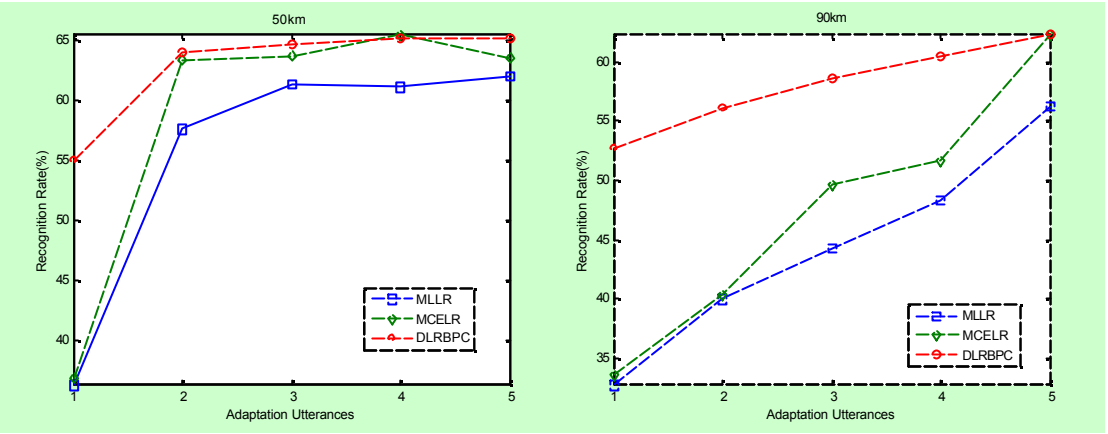


圖 五、DLRBPC 在五十公里與九十公里噪音語料調適結果

在此實驗中，我們曾首先在乾淨語料下以語者為單位對語者獨立的模型估測各別語者的迴歸矩陣，然後取其樣本平均值及樣本共變異數為我們迴歸矩陣的事前不確定性，但是其辨識效果不如我們預期，其可能原因是在乾淨語料中求出的為反應出語者變異的迴歸矩陣事前資訊，不同於描述噪音環境變異下的不確定性，所以我們以汽車噪音中的 0 公里的 50 句調適語料來估測我們不確定性的事前資訊。此外在此實驗中，我們提出的方法，在資料量不足的情況下還能夠保有穩定的辨識水準，主要是在於不確定性的描述及更新能夠符合於測試環境中的不匹配情況且也能代表估測迴歸矩陣的事前資訊。在較複雜的九十公里語料中，因為噪音的問題比 50 公里下嚴重所以參數的不確定性影響將更加劇烈，同時在少量調整語句下，又會有估測錯誤問題，所以明顯的同時考慮鑑別性及不確定性的迴歸矩陣估測可以達成較穩定的辨識效果。

5. 結論與未來展望

在本研究當中，我們提出以鑑別性的法則更新語音模型中不確定事前資訊的評估方法，不同於傳統的鑑別性訓練與調適方法針對模型參數調整，我們考慮在模型參數的超參數來調整，希望在保有維持參數的不確定性的資訊同時，還能兼顧其具有鑑別性的特性，達到更穩鍵之語音辨識效能。在初步的實驗中，我們發現在貝氏預測分類法則中導入對不確定性的鑑別性更新，在噪音環境下，有其提升辨識率的空間，嚴格來說，如同於第二節所言，當初始的不確定性資訊能夠描述參數間不匹配的關係時，則鑑別性的貝氏分類器可以進一步提升辨識率。

在辨識的計算量上，因為我們是使用於 BPMC 的近似架構下，所以整體的辨識流程是融合於傳統的辨識器中，不會有額外的計算量負擔。但是在參數更新上，我們是以最小化分類錯誤為鑑別性法則的出發點，其參數更新的計算量是在求取梯度微分量上及疊代更新的計算，所以主要的計算差異上是在鑑別性的超參數更新上。

在未來研究上，是否可以考慮其他的鑑別性函數來分析不一樣的鑑別性函數下對於不確定性的影響，另一方面，在我們現行所實現的方法，主要是假設在初始環境與調整後的環境間的不確定性都是同為高斯分佈的假設下，但也許這個通道的不確定性會是其他的機率分佈，如何分析模型參數與環境不匹配的參數對合關係，也是可以再進一步分析探討的問題。

參考文獻

[1] J.-T. Chien, "Linear regression based Bayesian predictive classification for speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 11, pp. 70-79, July 2002.

[2] J.-T. Chien and G.-H. Liao "Transformation-based Bayesian predictive classification using online prior evolution," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 399-410, May 2001.

[3] 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.

[4] 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.

[5] Q. Huo and C.-H. Lee, "Robust speech recognition based on adaptive classification and decision strategies," *Speech Communication*, vol. 34, pp. 175-194, 2001.

[6] Q. Huo and Chin-Hui Lee, "A Bayesian predictive classification approach to robust speech recognition," *IEEE Trans. Speech And Audio Processing*, vol.8, no.2, 2000.

[7] Q. Huo and C.-H. Lee, "A study of prior sensitivity for Bayesian predictive classification based robust speech recognition," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 2, 1998, pp. 741-744.

[8] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 161-172, Mar.1997.

[9] H, Jiang, X. Li and C Liu, "Large margin hidden Markov models for speech recognition," *To appear in IEEE Trans. Audio,Speech and Language Processing*, 2006.

[10] H. Jiang and Li Deng, "A robust compensation strategy for extraneous acoustic variations in spontaneous speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 1, January 2002.

[11] H. Jiang and Li Deng, "A Bayesian approach to the verification problem: applications to speaker verification," *IEEE Trans. on Speech and Audio Processing*, vol. 9, vo. 8, pp. 874-884, 2001.

[12] H. Jiang, K. Hirose and Q. Huo, "Improving Viterbi Bayesian predictive classification via sequential Bayesian learning in robust speech Recognition," *Speech Communication*, vol. 28, no. 4, pp. 313-326, 1999.

[13] 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.

[14] B.-H. Juang, and S. Katagiri, "Discriminative learning for minimum error classification", *IEEE Trans. Speech and Audio Processing*, vol. 40, no. 12 , pp. 3043-3054, Dec 1992.

[15] C. J. Leggeter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, pp. 171-185, 1995.

[16] Y. Normandin, R. Cardin and R. De Mori, "High-performance connected digit recognition using maximum mutual information estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 299-311, 1994.

結合韻律與聲學訊息之強健性漢語語者驗證系統

張文杰[2]，陳鼎允[1]，陳子和[2]，曾志仁[1]、廖元甫[1]，莊堯棠[2]

[1]國立台北科技大學電子工程學系

[2]國立中央大學電機工程學系

Email: yfliao@ntut.edu.tw

摘要

在本論文中，我們探討強健式漢語文字特定(text-dependent, TD)與文字不特定(text-independent, TI)語者驗證系統，主要是針對漢語的聲調語言特性，提出潛在韻律分析(latent prosody analysis, LPA)及高斯混合模型(Gaussian mixture model, GMM)兩種方式，分別用來建置每位語者的韻律行為模型及能量與音高軌跡(pitch contour)的動態變化模型。實驗結果顯示在使用 ISCSLP-SRE 語料之漢語文字特定與文字不特定語者驗證實驗情況下，使用韻律訊息(prosodic information)來輔助傳統使用頻譜特徵(spectral features)之語者驗證系統，可有效提升系統效能。

1. 序論

語者驗證在現今的語音處理中為重要的分支研究項目之一 [1] ，目前有相當多的研究不斷地持續發展中。尤其從 1996 年開始，NIST 機構每年都會藉由舉辦語者辨認評估(speaker recognition evaluation, SRE)來提供一個共同的測試平台 [2] ，以促進語者辨認技術演進及各種演算方法的實用性，更讓全世界最新穎的想法得以在競賽裡獲得驗證。相較於外國語言，漢語的語者辨認競賽還在起步階段，在 2006 年舉辦的中文口語語言處理國際會議(ISCSLP)中，首度建立了漢語語言的語者競賽機制 [3] ，讓此領域的研究人員能夠同時在擁有一樣的資源下，透過中文語言資源聯盟(Chinese Corpus Consortium, CCC) [4] 所提供的資料庫，切磋漢語語者的辨認技術與研究。

語者驗證技術在現實生活中可以有許多的應用，例如可以藉由電話連接到銀行或是信用卡等客服中心，並直接透過使用者的聲音來驗證身份以即時提供便利的私人服務。然而使用者若任意使用不同的電話話筒或通道，則會有電話話筒與通道環境不匹配問題，而導致傳統以頻譜特徵為主之語者驗證系統效能降低。為了改善電話話筒與通道不匹配問題，近年來有許多人利用韻律訊息來強化傳統以頻譜特徵為基礎之語者驗證系統 [5-8] 的效能，韻律特徵(prosodic feature)不僅含有語者訊息並已被認定是不易受到電話話筒與通道不匹配的影響，而且在西方語言的研究中亦有很多的文獻證實其效果。因此在本論文中我們將著重在討論如何利用韻律特徵來強化漢語語者驗證系統的效能，主要是考慮到漢語屬於一種聲調(tonal)語言，其本質上依賴聲調的不同來區別出同音異字詞，故韻律特徵對漢語的影響應較西方語言強烈。

一般來說頻譜特徵代表是較短程(short term)且低階層的聲學訊息，都是和發音器官相關的實體線索，其中被廣泛使用的梅爾頻率倒頻譜係數(Mel-frequency cepstral coefficients, MFCCs)是可以擷取並傳達出發音腔道(vocal tract)的分佈；韻律特徵則通常作為聲門資訊(glottic source)的特徵參數，不僅是較長程(long term)且高階的特徵並含有語者本身特殊的訊息，如音高軌跡及音調(intonation)等，因此兩者各是呈現語音訊號中不同的訊息。在韻律訊息改善不匹配問題的方法

中，對於短程韻律方面通常會使用高斯混合模型來統計韻律訊息，能捕捉到如音高與能量的分佈、音高與能量的斜率以及音高與能量的持續時間等韻律特徵，而長程韻律模型則通常有 N-gram 及 discrete hidden Markov model(DHMM) [6] 兩種方法，可以表現出韻律訊息隨時間的長程變化。不過長程韻律模型通常受限於大量語料的需求問題，因為要有充分語料才能有效描述韻律的特性，所以針對這點缺失我們將提出潛在韻律分析方法來得到可靠的韻律訊息。

本文章中，我們會在系統前端的頻譜特徵使用 mean subtraction, variance normalization, and ARMA filtering (MVA) [9] 去除部份通道不匹配的問題，接著語者驗證系統將運用不同模組來整合頻譜與韻律訊息。文字不特定條件下，有三種模組用作語者確認系統的建構，包括目前被視為標準作法的 a maximum a posteriori (MAP)-adapted GMM (MAP-GMM) [10] 、音高與能量之高斯混合模型，以及潛在韻律分析模組。而文字特定則有另外三種模組來構成，包括文字限定的語者高斯混合模型，隱藏式馬可夫模型(hidden Markov model, HMM)以及音高與能量之高斯混合模型。而後端改良型的測試分數正規化(test normalization, T-norm) [11] 則可以對分數作調整。最後我們利用 MIT 林肯實驗室所發展的 LNKnet [12] 軟體做不同模組分數上的結合。

在 MVA 對頻譜特徵的處理主要是將特徵向量作一種正規化，雖然近年來有很多特徵正規化的方法，如 feature warping [13] 及 histogram equalization(HEQ) [14]都能有很好效果，但是 MVA 的良好表現與簡單使用是我們在此優先考量的原因。而文字不特定語者驗證中的 MAP-GMM 是透過通用背景模型(universal background model, UBM)調適出語者個別的高斯混合模型，使每個語者模型所含蓋的聲學特性更具完整性，如此對於文字內容的變異性就能廣泛接納。而韻律特徵由兩方面著手，短程韻律用高斯混合模型對能量與音高軌跡建置其動態變化模型，長程則用所提出的潛在韻律分析更有效地得知韻律行為，其主要是將語者驗證問題轉換為類似文件檢索（document retrieval）的問題，統計出韻律序列的組合並建立韻律空間(prosody space)，再透過 probabilistic latent semantic analysis（PLSA）[15-16] 的空間維度簡化後來呈現語者的韻律模型。

文字特定語者驗證任務對使用者的說話內容是有其限制，所以對語音事件之聲學變化有詳細考慮的隱藏式馬可夫模型是必需的，這樣才能善用系統對使用者先天的限制條件。當然高斯混合模型在頻譜上對語者特性的描述仍是不可或缺的角色，因為用高斯密度函數表示語者的聲學類別仍可反應出語者特性分佈，與隱藏式馬可夫模型分屬不同角度的分析。而在韻律訊息方面，考慮到語料長度的缺乏，僅對短程韻律方面使用高斯混合模型來描述音高與能量軌跡的動態變化。至於系統後端我們也考慮到分數的變化性，來自語者之間說話內容或是長度的不同都會造成影響，且訓練和測試環境的不匹配更是一大主因，所以使用改良式測試分數正規化(modified test normalization, MT-norm)來調整目標語者(target speaker)模型的分數，拉開目標語者與冒充語者(impostor)之間的分佈，進而改善正確率並更簡易產生驗證所用的門檻值。

由於頻譜特徵與韻律特徵是呈現訊號中不同的訊息，所以考慮其之間可能的互補特性，則文字不特定與文字特定語者驗證的不同模組必須整合，而我們是透過多層感知機(multi-layer perceptrons, MLPs)與 development 的測試語料來決定驗證系統的合併方式，在系統求得的分數上作非線性組合，以達到利用韻律特徵來強化漢語語者驗證系統之目的。

本文內容安排如下：第二章節描述在漢語語言裡所使用的各種方法，並討論韻律特徵在系統中的輔助作用；第三節則詳細說明潛在韻律分析的方法；第四節是系統運用在 ISCSLP2006-SRE 的實驗結果；最後則為結論。

2. 文字特定與文字不特定之語者驗證系統架構

圖一及圖二所示分別為文字不特定與文字特定驗證之整合架構，對於文字不特定來說，將有音框(frame)和語者兩種層級一起使用，主要是因為考量到註冊與測試語料數量的關係，在語者層級所需的量遠比音框層級大得多，況且現實狀況中總是只能獲得有限的語料量；反觀文字特定的情況則只是採取音框層級的方式，因為該語料的長度都非常的簡短。

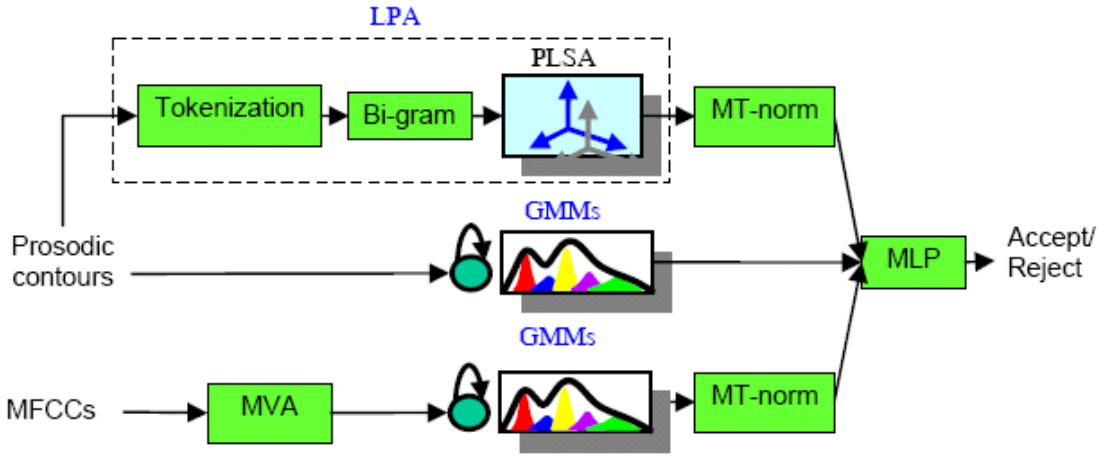
文字不特定任務是由三種不同模組來構成，如圖一所示，首先是高斯混合模型將音高與能量軌跡的動態變化與所提出之潛在韻律分析方法做一合併，完整獲取每位語者的韻律行為，最後則是以 MAP-GMM 完成系統在頻譜特徵的主體。利用 MAP-GMM 取代原本的文字限定語者高斯混合模型，原因在於實際應用情況中不可能要求使用者在註冊時錄製大量的語音，以致於每一個人的訓練語料可能有一些聲學特性沒被涵蓋到，在測試時可能會造成系統效能下降，並且文字不特定的確認是無法限制測試語者說話的內容，所以建立出來的語者模型不僅要能代表該註冊語者的特性，還要能夠涵括在不同聲學情況下的語者變異性。

為了克服電話話筒與通道不匹配的影響，韻律訊息的使用仍是我們主要考量，雖然使用高斯混合模型可以用來統計韻律訊息，但一般只能補捉到音高與能量變化等短程的韻律訊息，其所得到的改善幅度仍然有限，而對於補捉較長程韻律訊息變化的方法通常有 DHMM 和 N-gram 兩種，可是都需使用大量的訓練與測試語料，對此我們提出潛在韻律分析的方法是能在有限的語料情況下得到可靠的韻律訊息。

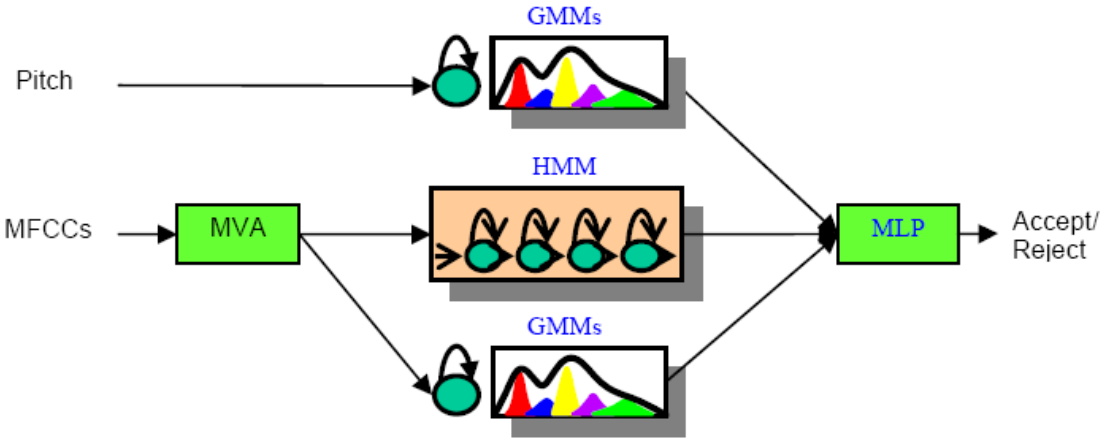
在文字特定條件下，有另外三種模組用作語者驗證系統的建構，如圖二所示，包括獲知音高與能量軌跡動態變化的高斯混合模型、模型化梅爾頻率倒頻譜係數暫態軌跡所用的隱藏式馬可夫模型以及統計梅爾頻率倒頻譜係數分佈的高斯混合模型。一般來說，語者驗證系統都會採用梅爾頻率倒頻譜係數與高斯混合模型的搭配，其中梅爾頻率倒頻譜係數已經將語音的頻譜特徵做了良好描述，然後透過由許多高斯密度函數組成的高斯混合模型來表示語者特性的分佈，而這裡並不如文字不特定中使用 MAP-GMM 來建立語者特定模型，因為藉助通用背景模型補強的聲學特性反而會對文字特定產生困擾，造成語者模型無法針對文字特定任務進行驗證。

另外圖二的文字特定語者驗證中，我們可知測試語者說話內容是有限制性的，它必須符合宣稱語者在系統中註冊語料的語句內容，除此內容外的語句都將一律拒絕，即便是真實語者說出不同樣的內容也是無法接受的，利用這種系統使用上的限制條件，隱藏式馬可夫模型會是更適合用來建立模型的方法，因為隱藏式馬可夫模型對梅爾頻率倒頻譜係數之暫態軌跡可以有詳細的描述，而高斯混合模型中並未考慮到語音事件的聲學變化。

以頻譜特徵為主的系統來說，隱藏式馬可夫模型與高斯混合模型的結合已經可以獲得還不錯的結果，然而在訓練與測試環境不匹配的狀況下，仍需加入不同觀點的韻律訊息來強健系統，因此我們考慮音高與能量軌跡的動態變化，利用有聲音(voiced)的區段中取出每一個音框的對數(log)音高及對數能量，並估計對數音高及能量的一階微分來建立高斯混合模型，而由於韻律特徵是比較不受話筒或通道的影響，所以可以補強原頻譜系統的缺失。



圖一、文字不特定語者驗證方法之方塊圖。



圖二、文字特定語者驗證方法之方塊圖。

最後值得一提的是系統不論特定或不特定的任務，對取自於梅爾頻率倒頻譜係數的特徵向量我們都利用 MVA 去除部份通道不匹配問題，因為頻譜上受通道造成的偏移量相當於時間上的旋積性(convolutional)噪音，而梅爾頻率倒頻譜係數對平均值的削減正可以對抗旋積性噪音下之失真，至於變異數正規化與濾波器的使用則分別可以對抗加成性(additive)與高功率加成性噪音下的失真。在文字不特定語者驗證系統後段的分數方面，更透過改良式測試分數正規化做補償，將同儕語者模型(cohort model set)分數的平均值與變異數來調整目標語者模型的分數，經由減去平均值可以使冒充語者分數的分佈中心移至原點，而除以變異數則能將冒充語者分數分佈之標準差限定為一，這樣的方法不僅可以拉開目標語者與冒充語者之間的分佈進而改善正確率，還能讓決定接受與否的門檻值更容易產生。另外，多層感知機可用來結合各模組之語者的測試分數，進一步把頻譜系統及韻律系統融合，以便強化驗證系統之效能。

3. 潛在韻律分析

在語音訊號中，韻律訊息的動態變化受到各種潛藏因素的影響更甚於語者本身特性，譬如說話速度、情緒轉變以及說話內容等等，因此我們所觀察到韻律軌跡表象的變異量是相當大。而跟西方

語言比較之下可知漢語屬於一種聲調語言，隱藏於內的聲調更是個關鍵的因素，將會大大地影響韻律軌跡的變化。

一般來說，prosody state N-gram 語者模型 [6-8] 已經是以韻律特徵為主之驗證系統所採納的方法，然而想要能夠可靠地估測出此 N-gram 語者模型，則擁有大量的訓練與測試語料通常是先決條件，譬如說知名的 NIST2001-SRE Extended Data Task 中，分別使用 8 句和 2 句約兩分鐘的對話句子作為訓練與測試。然而在我們所提出的潛在韻律分析方法裡，大量語料將不會是必需條件，因為相同的語料庫下已能成功地運用在文字不特定之語者驗證 [8]，且平均來說僅僅只需共兩分鐘及三十秒的訓練與測試語料量，所以我們將嘗試著套用此方法在屬於聲調型的語言，特別是在漢語語言上的表現尚未能明確地得知。

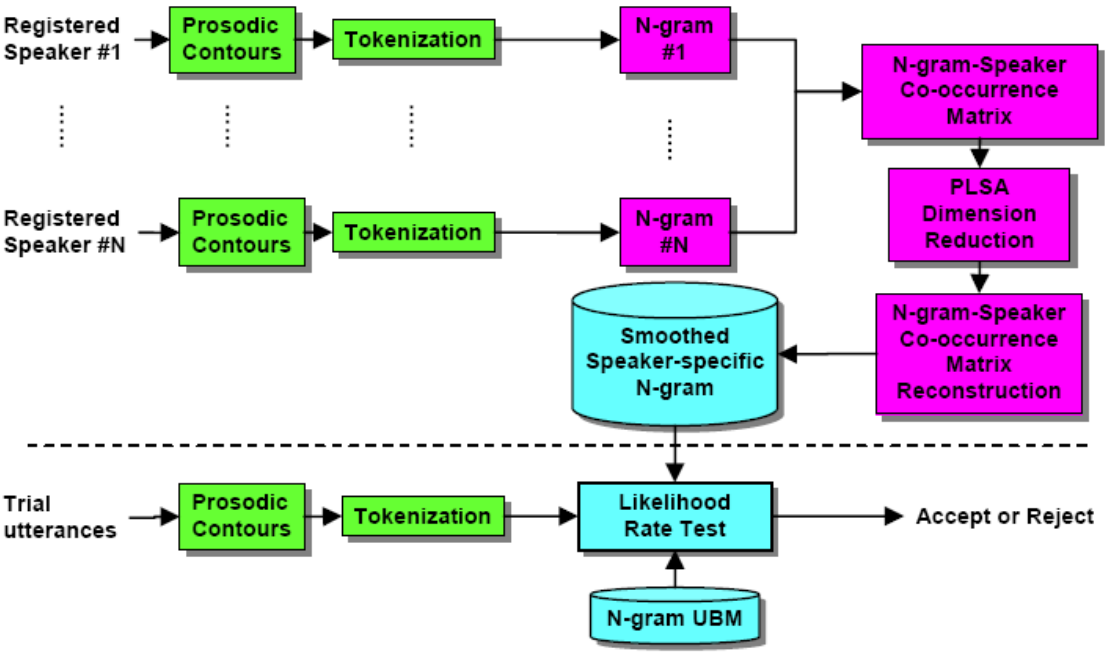
關於潛在韻律分析的基本構想是利用 PLSA 概念找出一個低維度的韻律資訊空間以表示語者的特性所在位置，主要是為了擷取出重要的韻律線索來鑑別語者之間的不同，再者是讓語者特定的韻律狀態 N-gram 語者模型能更可靠的建立。圖三是潛在韻律分析方法在語者驗證應用的方塊圖，首先必須把輸入語句的韻律軌跡經由 Tokenization 自動轉換成韻律狀態序列，並在訓練階段中建立起 N-gram 語者關係矩陣(co-occurrence matrix)，目的是集合每位語者的韻律行為特性來學習韻律狀態資訊和語者之間的相互關係。

圖四則說明 Tokenization 如何自動標記及轉換成韻律狀態序列。由 piece-wise curve fitting 先把每一段傳入的韻律軌跡擷取其韻律特徵向量，且許多鄰近的區段將串連成一個龐大的韻律特徵 supervector，而考量到音節為最小的韻律單位，所以採用五種音節層次的韻律特徵參數，包括一個母音區段的音高斜率(pitch slope)和長度的延長變化(lengthening factor)、兩個母音間的對數能量差和音高跳躍(pitch jump)以及兩個音節之間的暫停長度(pause duration)。此外為了移除語句發音內容（context-information）對韻律變化的影響，必須將韻律特徵參數做正規化的動作，藉由整個訓練語料所統計出來之韻律特徵參數的平均值及標準差，移去任何非韻律特性的影響。於是一個以向量量化為基礎，透過 Expectation- Maximization(EM)演算法訓練好的韻律模型便可以自動地把輸入語句所構成的 supervector 作符號的標記，並且再轉換成一連串的韻律狀態序列。

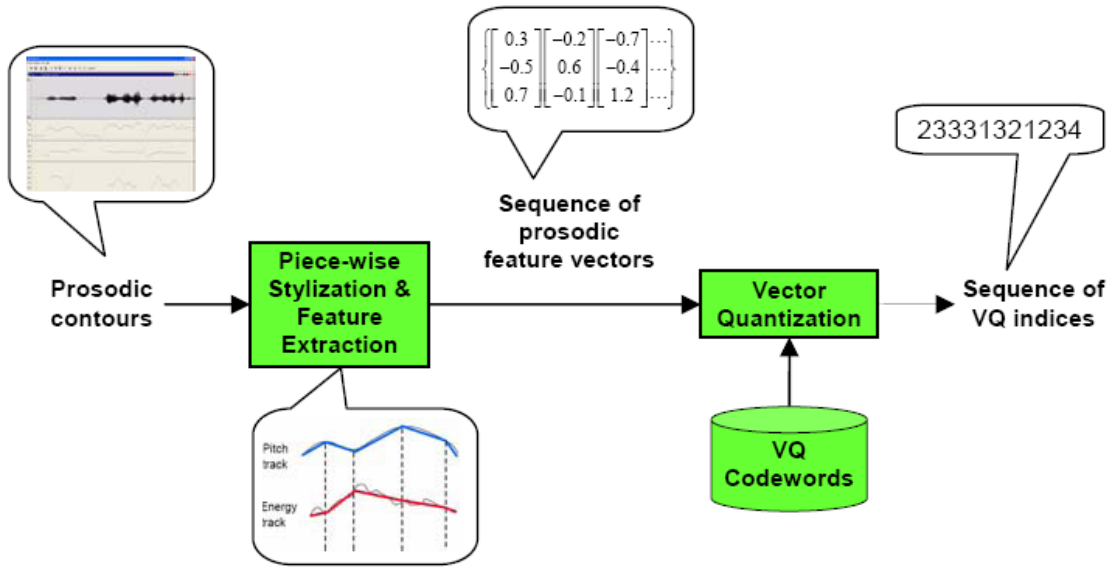
獲得韻律狀態 N-gram 語者關係矩陣後，由於訓練語料與測試語料之資料量的不足，在受此限制之下以韻律訊息所建構出的 N-gram 語者模型可能不夠具有統計特性，沒辦法準確的訓練出代表語者韻律特性的語者模型，所以必須再經過 PLSA 找出降低維度的韻律資訊空間，如圖五所示，而其分解定義如下，

$$P(d_i, w_j) = P(d_i)P(w_j \mid d_i) = P(d_i)\sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i) \qquad (1)$$

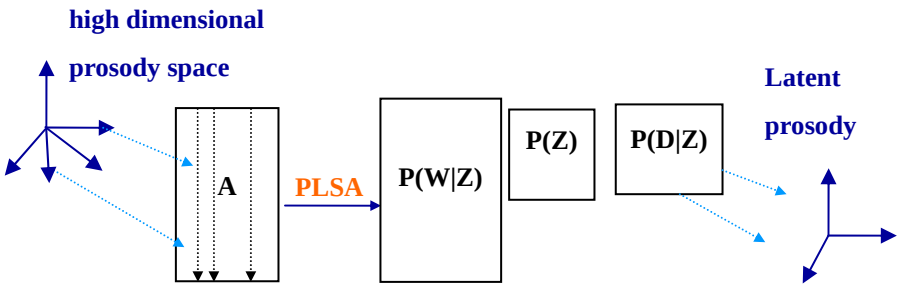
其中 d_i 、 w_i 、 z 所代表的分別是 document、keyword、latent prosody factors， $P(d_i, w_j)$ 所代表的是 document(d_i)與 keyword(w_i)之間的結合機率，且 document 和 keyword 對應到韻律特徵的關係分別為語者及 N-gram term。這樣一來較為可靠的 N-gram 關係矩陣就能藉由幾個少數的特徵韻律向量(eigen-prosody vector)順利重建，達到語者韻律模型平滑化處理之目的。最後在測試階段我們只需將重建空間產生的 N-gram 語者模型和測試語句計算出相似度比值(likelihood ratio)，即可完成驗證的任務。



圖三、潛在語意分析方法輔助系統之方塊圖。



圖四、自動標記及轉換成韻律狀態序列的方塊圖。



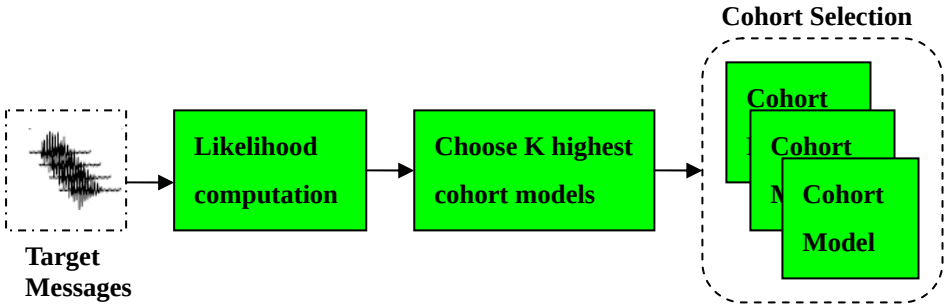
圖五、韻律特徵空間降維。

4. 改良式測試分數正規化

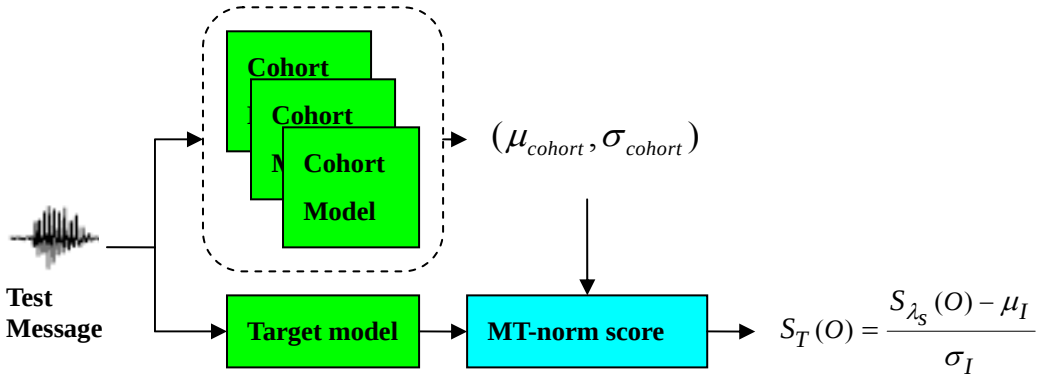
在改良式測試分數正規化的原理是利用一群相似於目標語者模型的同儕語者模型，估計出相似於每位不同目標語者的冒充語者，這和原始測試分數正規化方法有所不同，因為正規化所用的參數不再由同一組同儕語者模型得到，而是針對每個目標語者模型找出各別對應的同儕語者模型，如此才能找出每個目標語者模型真正的冒充語者群，這樣的方式亦可以帶來減少運算量的好處，因為對於和目標語者模型較不相似的同儕語者模型可以不再考慮其影響。而我們的同儕語者模型估測是根據訓練語料對每個語者特定模型量測 log likelihood score 而得，如圖六所示，這和 [11] 中利用距離的估測方式是有所不同的，主要是藉由 log likelihood score 的高低來決定同儕語者模型，並選出前面 K 個同儕語者模型來計算參數。接著每位不同目標語者依據相對應的同儕語者模型計算出分數的平均值與變異數作為調整目標語者分數的參數，其定義如下，

$$S_T = \frac{S_{\lambda_S} - \mu_I}{\sigma_I} \tag{2}$$

其中 $S_{\lambda S}$ 為測試語料與語者模型 λ_s 所計算的 log likelihood score，μ_I 與 σ_I 分別代表測試語料相對於同儕語者模型分數的平均值與變異數，S_T 為經過測試正規化後的分數。從(2)式中看到測試正規化減去 μ_I，這動作可以將冒充語者分數的分佈之中心移至原點，亦即同時拉大目標語者與冒充語者的分數分佈，而除以 σ_I 則可以將冒充語者的分數分佈之標準差限定為一，進而提升正確率。而整個改良式測試分數正規化的架構則如圖七所示。



圖六、同儕語者模型的估測。



圖七、改良式測試分數正規化方塊圖。

5. 漢語之語者驗證實驗結果

5.1. ISCSLP2006-SRE 語料庫

在此語料庫中，不論是文字特定與文字不特定的語者驗證任務，都是來自中文語言資源聯盟所提供的 development 與 evaluation 資料庫，而此語料庫所有的聲音檔案都是採用 8kHz 取樣頻率，且為 16bits 單聲道的 PCM 格式。至於 evaluation 的語料庫中，其真實語者與冒充語者的測試樣本比例為 1 比 20。

5.1.1 文字特定語者驗證之語料庫

development 的資料取自於 CCC-VPR3C2005，語料庫包含了男女性各 5 位的個別資料量，每個人的聲音透過三種不同麥克風通道來獲得，分別用“micl”、“micr”及“micu”三種符號來表示，其中每位語者在分別通道上有五種句子會重複錄製 4 遍，而另外二十一種句子只會各錄製一次。對於 evaluation 的部份則共有 591 位註冊語者，每一位都有相同內容的三個句子，平均用來註冊的語句長度約有 4.5 秒，最後用來試驗的句子共有 11181 個且平均時間長度為 5.2 秒。值得一提的是每句發話開始都有很長的靜音，此外，某些雖然由相同語者所發出但卻為不同語句內容的句子，我們應該視為冒充語者並加以拒絕掉。

5.1.2 文字不特定語者驗證之語料庫

development 的資料取自於 CCC-VPR2C2005-1000，語料庫只包含了 300 位男性語者，每位語者含有兩種語句，分別由電話線(PTSN)及手機通道(GSM)所製成，所以總共有 600 個句子在內。evaluation 的部份則共有 800 位註冊語者，每一位都只會有一句從電話線或是手機通道所提供的語料，平均用來註冊的語句長度約有 36.2 秒，最後用來試驗的句子共有 11800 個且平均時間長度為 15.9 秒。

5.2. 實驗條件

本文中對於所有的頻譜特徵為主之語者驗證系統都用 39 維的梅爾頻率倒頻譜係數作為特徵參數，包括前 13 維倒頻譜係數(包含 C_0)及其差分值 Δ-MFCCs 與二次差分值 Δ2-MFCCs，至於音高與能量的軌跡則是藉由 snack 軟體套件中的 ESPS 音高擷取演算法來求得 [17] ，同時也計算其差分值及二次差分值，最後則將音高與能量連同梅爾頻率倒頻譜係數一併作為使用。

另外在驗證系統的合併方式，我們是運用共有 120 個隱藏節點的多層感知機，將頻譜特徵與韻律特徵在 evaluation 測試語料上所得到的分數作一結合。這部份的步驟是須先把 development 的語料區分為訓練和測試使用，其訓練語料部分用來訓練出各個系統的模型參數，而其測試語料的部份則用來得取每個系統的辨識結果，接著繼續再利用其測試語料的部份建置出多層感知機的各項參數，然後便可將各個系統的融合參數固定套用到之後 evaluation 測試語料所得到的分數上，而多層感知機的各項參數是由實驗數據所得，因此在後面的文字不特定與文字特定語者驗證實驗結果，所呈現的是系統之最佳效果。

語者驗證系統的錯誤率有兩種：一種是錯誤拒絕率(False Rejection Rate, FR)，即正確語者的分數小於門檻值造成拒絕的錯誤率。另一種是錯誤接受率(False Acceptance Rate, FA)，即仿冒語者的分數高於門檻值造成接受的錯誤率。FA、FR 這兩種錯誤率是一種取捨(tradeoff)的關係，若把門檻值提高，則錯誤拒絕率將會提高，而錯誤接受率則會降低；若門檻值降低，則錯誤拒絕率將會降低，而錯誤接受率則會提高。所以系統最後效能的量測主要是透過相等錯誤率(equal error rate, EER)及決策成本函數(decision cost function, DCF)來衡量。

相等錯誤率是一種評估語者驗證系統的方式，所謂的相等錯誤率就是錯誤拒絕率與錯誤接受率相等時的機率值，但在某些特殊的情形中，錯誤拒絕與錯誤接受的後果和重要性並不相等。舉例來說，語者驗證應用在金融交易的情況，為了避免冒領盜用，因此錯誤接受的機率必須減至最低。而決策成本函數則是被定義成一種錯誤機率的加權總和，如下所示。

$$C_{DET} = C_{Miss} \cdot P_{Miss} \cdot P_{Target} + C_{False} \cdot P_{False} \cdot \left(1 - P_{Target}\right) \tag{3}$$

其中 $C_{Miss} = 10, C_{False} = 1, P_{Target} = 0.05$ 。

另外，呈現錯誤拒絕率及錯誤接受率的方式則使用偵測錯誤交易曲線圖(Detection Error Tradeoff Curve, DET Curve)，此種方式是假設目標語者和仿冒語者的對數相似度比分數為兩各個不同的高斯分佈，隨著門檻值的變化表現出相對應錯誤拒絕率及錯誤接受率的曲線變化。

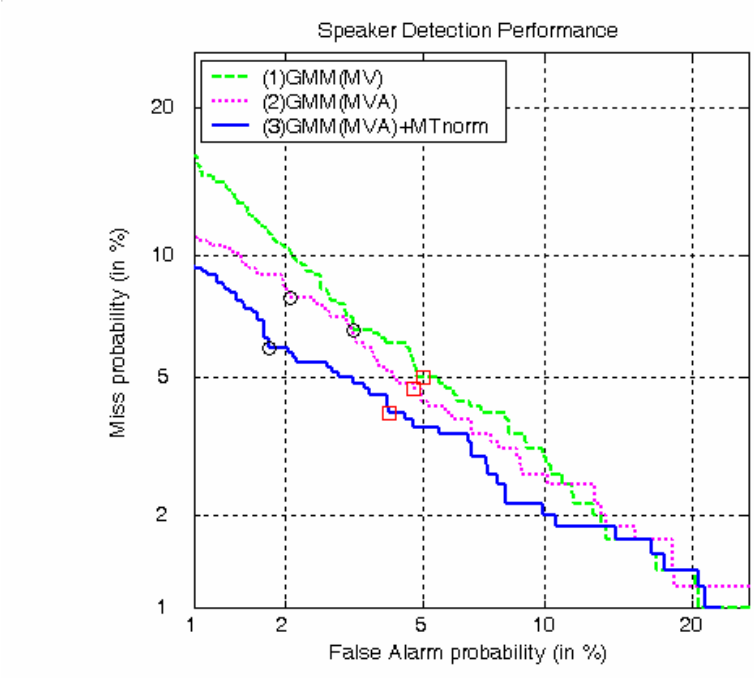
5.3. 文字不特定與文字特定語者驗證實驗結果

5.3.1 文字不特定語者驗證結果

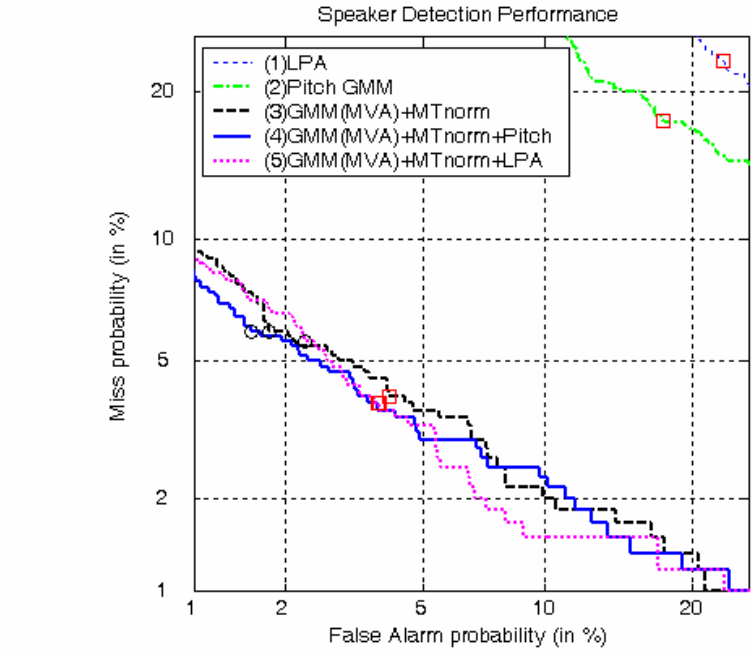
首先，頻譜特徵為主的 MAP-GMM 驗證結果會作為各項模組比較的標準，其通用背景模型一律為所有註冊語者訓練語料集成並用 1024 混合數組成，而語者特定的高斯混合模型是利用其註冊語料向通用背景模型調適得到。針對系統前端有兩種特徵正規化的方法會作為考量，包括 cepstral mean and variance normalization (MV)及 MVA，這兩種方式的結果在圖八可看出，MVA 的效果明顯比 MV 好了許多，因此我們將 MAP-GMM 和 MVA 的組合作為最基本的語者驗證系統。接著系統後端的改良式測試分數正規化則以 320 個相似語者為主，由圖八可看到此方法對驗證統確實有很大影響並大幅改善 MAP-GMM 的結果，由此可知 MVA 與改良式分數正規化法不僅相當有效且是互相補償。所以我們在文字不特定語者驗證中，以 MVA 與改良式分數正規化法和 MAP-GMM 的結合方式，作為頻譜特徵方面最佳的架構，爾後再加入韻律特徵的輔助。

在圖九中我們看到兩種韻律模型化的方法被用來和頻譜特徵最佳之效果做一結合。以高斯混合模型方式來說，該語者特定 64 混合數的韻律模型是直接由其註冊語料訓練而成，而非透過背景模型來調適，其驗證結果的相等錯誤率及決策成本函數分別為 17.7%和 0.223；另外潛在韻律分析方式則使用到 bi-gram 模型及 11 個狀態的向量量化(8 個為音高與能量使用，3 個為 pause segments 所用)，可以讓潛在韻律空間中的文件大小從 112(11*11-9)個維度減少至 30 個，這表示說每位語者其 N-gram 模型的平均參數量可從 112 降到僅僅只有 34.2 維，而其所帶來的好處是大幅的簡化了系統的複雜度，其驗證結果的相等錯誤率及決策成本函數分別為 22.7%和 0.272。

在韻律特徵方面，音高與能量之高斯混合模型及潛在韻律分析的相等錯誤率分別為 17.7%和 22.7%，這樣的結果以強化頻譜特徵為主的輔助角度來看已經是很不錯的，而將這兩種韻律模型化的方法與頻譜特徵最佳結果合併後，分別能讓驗證系統再從相等錯誤率 4.0%及決策成本函數 0.047 改善至 3.8%與 0.045 以及 3.8%與 0.050，可見韻律特徵對頻譜特徵的系統確實是產生輔助的效益。此外，在圖一及圖九的結果可看到文字不特定的語者驗證中，對於韻律特徵的使用並未先做特徵正規化的處理，因為我們主要是先決定好頻譜特徵方面驗證效果最好的架構後，再藉由韻律特徵的輔助作用，用以強化使用頻譜特徵之語者驗證系統的效能，因此沒有討論韻律特徵在未正規化之頻譜特徵下對於系統的影響，相對的文字特定語者驗證也是如此。



圖八、文字不特定語者驗證系統，在頻譜特徵上使用不同前後端處理方式的 DET 曲線圖。



圖九、包括 5 種不同文字不特定語者驗證系統之 DET 曲線圖。

5.3.2 文字特定語者驗證結果

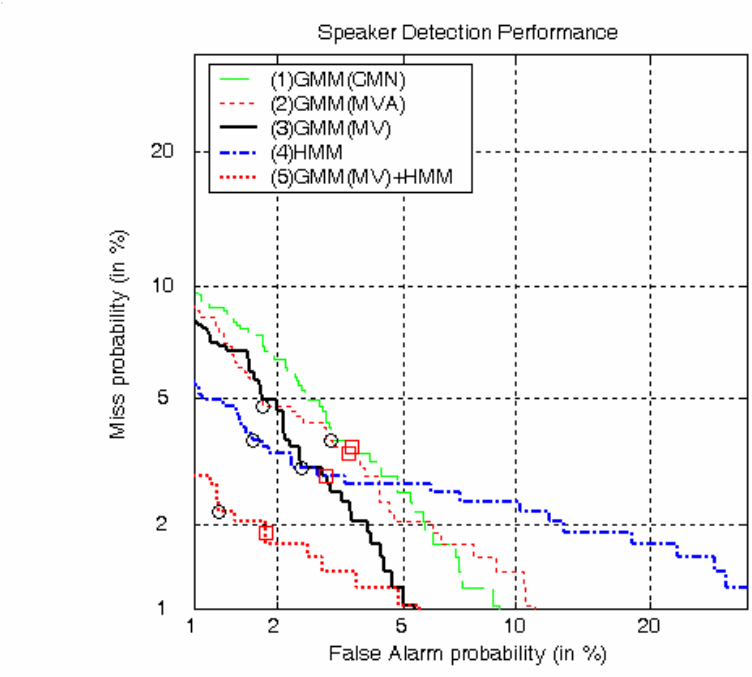
在頻譜特徵的實驗中，針對系統前端處理則考量 cepstral mean normalization(CMN)、MVA 及 MV 三種方式，從圖十的結果可知 MVA 的驗證結果略勝 CMN，然而我們發現在文字特定任務裡，MV 的表現更優於 MVA，這可能是因為在文字特定的語料庫是由麥克風錄製而成，且所有測試

時所用的麥克風特性都在訓練過程中遇過，如此現象在隱藏式馬可夫模型系統中將是不謀而合。

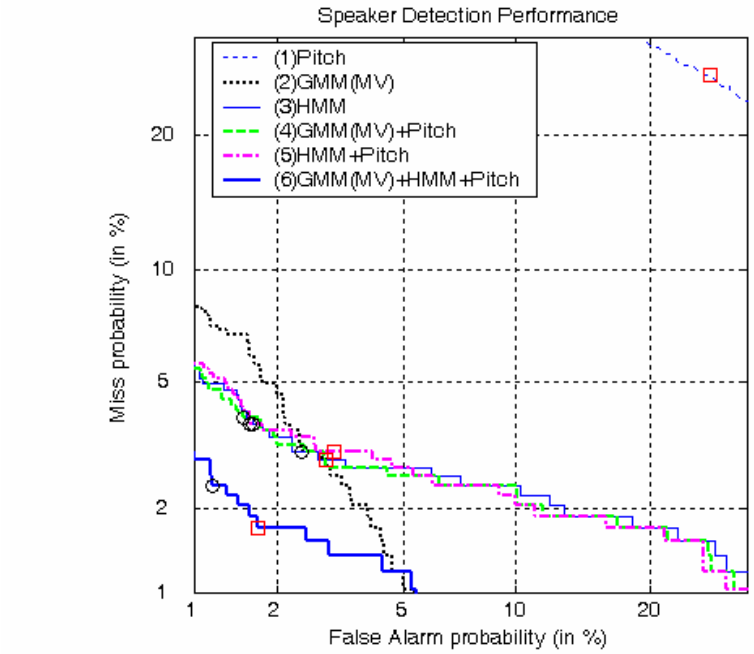
文字特定語者驗證中，有文字限定的語者高斯混合模型及隱藏式馬可夫模型兩種模組被用來建構驗證系統，包含 16 混合數的高斯混合模型及 8 混合數的隱藏式馬可夫模型，而隱藏式馬可夫模型的狀態數目是根據註冊語料中文字的多寡來做調整。從圖十的結果看到，雖然兩種方法結果的曲線趨勢有所不同，但高斯混合模型最佳的相等錯誤率及決策成本函數分別為 2.9%和 0.038，和隱藏式馬可夫模型 2.9%和 0.034 的表現卻是差不多。圖十亦可看見有趣的結果是當這兩個系統結合後會對結果產生強勁的改善，相等錯誤率及決策成本函數分別為 1.9%和 0.023，而這或許就是對這兩種系統的互補性做了最佳的驗證，因為高斯混合模型僅需少量語料便能對每個音框的倒頻譜係數分佈作模型化，反觀隱藏式馬可夫模型多量需求才能仔細描繪出倒頻譜係數的暫態軌跡，可見兩者所長不同於語料量所供應的大小。

當頻譜特徵為主的最佳系統建立好之後，再來就是關於韻律特徵與頻譜特徵的結合，相較於文字不特定的任務，我們只運用了音高及能量的 8 混合數高斯混合模型。圖十一顯示韻律特徵和上面兩套系統的合併結果，與文字限定語者高斯混合模型的結合是相等錯誤率及決策成本函數分別為 2.9%和 0.034，而隱藏式馬可夫模型則是 3.1%和 0.034，雖然由此看到效果沒能有顯著的改善，不過有趣的在於我們將所有的方法結合後，整個系統的錯誤率又再下降些許，如圖十一所示，而這似乎又證實了韻律特徵對頻譜特徵的互補在文字特定的系統上仍舊是很有效果的。

另外在圖二及圖十的結果不難發現，文字特定語者驗證系統並未呈現改良式分數正規化的結果，這是因為改良式分數正規化方法中，不論我們使用多少相似語者的數量，都未能讓系統有所改善。如此的結果和文字不特定語者驗證雖不一致，但由於文字特定語者驗證的測試語料通道特性都是在訓練階段看過的，因此改良式分數正規化無法在文字特定語者驗證上有所貢獻，可能就是測試語料與訓練時通道一致的關係，故針對此語料庫的文字特定語者驗證並不適合使用改良式分數正規化。



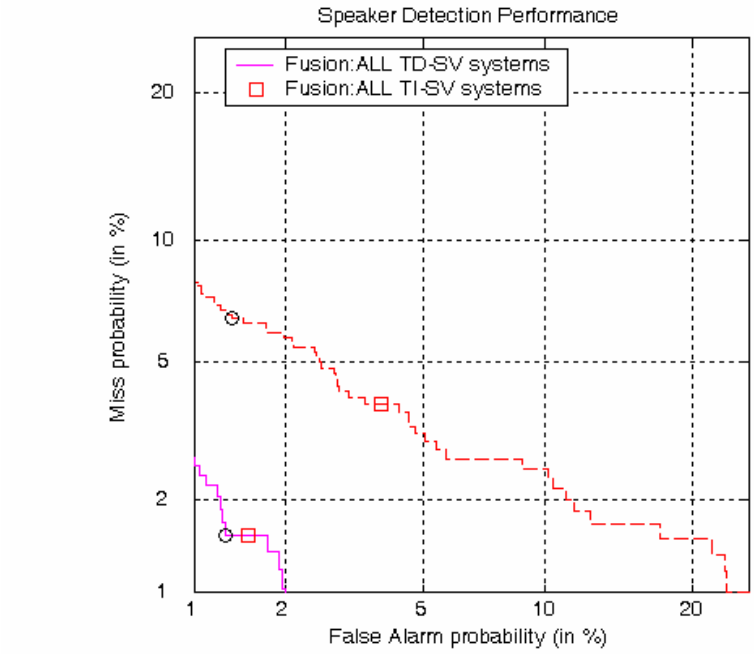
圖十、文字特定語者驗證系統，在頻譜特徵上使用不同處理方式的 DET 曲線圖。



圖十一、包括 6 種不同文字特定語者驗證系統之 DET 曲線圖。

5.3.3 文字不特定與文字特定之語者驗證結果比較

在此我們將融合文字不特定與文字特定語者驗證系統中所有的模組，因為不僅只考慮頻譜特徵與韻律特徵之間的互補特性，不同模組之間的相關性亦可為有效的特徵，所以再藉由多層感知機不同於一般線性組合的方式，將所有模組作全面性合併，結果如圖十二所示。



圖十二、結合系統所有模組的 DET 曲線圖。

此外，表一與表二呈現文字不特定與文字特定語者驗證系統的所有結果。從表一中的比較我們發現韻律特徵的作用必須建構在頻譜特徵的系統上，儘管 22.7%和 17.7%的相等錯誤率及 0.272 與 0.223 決策成本函數不能和頻譜特徵的 4.0%和 0.047 相比，但兩者合併後的相等錯誤率及決策成本函數為 3.8%和 0.045，確實改善了文字不特定語者驗證系統正確率。而表二中更顯示文字特定語者驗證的韻律特徵對於系統強化有很不錯的幫助，讓系統從頻譜特徵最佳結果的 1.9%與 0.023，大幅改善至 1.5%與 0.020。

表一、8 種不同文字不特定語者驗證系統之結果比較。

	ERR (%)	DCF
(1) LPA	22.7	0.272
(2) Pitch GMM	17.7	0.223
(3) GMM(MV)	5.0	0.064
(4) GMM(MVA)	4.7	0.060
(5) GMM(MVA)+MT-norm	4.0	0.047
(6) GMM(MVA)+MT-norm+Pitch	3.8	0.045
(7) GMM(MVA)+MT-norm+LPA	3.8	0.050
(8) Fusion ALL	3.8	0.045

表二、10 種不同文字特定語者驗證系統之結果比較。

	ERR(%)	DCF
(1) Pitch	25.9	0.297
(2) GMM(CMN)	3.6	0.047
(3) GMM(MVA)	3.4	0.041
(4) GMM(MV)	2.9	0.038
(5) HMM	2.9	0.034
(6) GMM(MV)+Pitch	2.9	0.034
(7) HMM+Pitch	3.1	0.034
(8) GMM(MV)+HMM	1.9	0.023
(9) GMM(MV)+HMM+Pitch	1.7	0.023
(10) Fusion ALL	1.5	0.020

5.4. 結果討論

在文字特定語者驗證系統中，如圖二所示，可發現並沒有運用潛在韻律分析方法，這是因為受限於語料量的因數，過於貧乏的語料無法建立有效的韻律特徵，雖然潛在韻律分析在文字不特定上有不錯表現，但對於文字特定語者驗證，仍需藉由調適或其他方式來解決語料問題，而這將是我們往後發展韻律特徵的目標。另外對於潛在語意分析方法輔助系統，如圖三所示，我們先透過自動標記及轉換成韻律狀態序列的方式，如圖四所示，再由 PLSA 作後續處理，這是因為目前尚無

法直接用 PLSA 獲得韻律軌跡對應於語者特性的關係，所以往後針對 PLSA 分析能力的運用將會作改善，以期能更完整保留語者的韻律訊息。

前面的實驗結果顯示，雖然韻律訊息的確能輔助傳統使用頻譜特徵之語者驗證系統，有效提升系統效能，尤其是在文字特定語者驗證系統上，但我們發現與韻律特徵在英文語者驗證系統的貢獻相比，此漢語語料庫的語者驗證，似乎沒能因為屬於聲調語言而突顯出韻律特徵的關鍵性，這可能攸關於些許條件上的差異，好比在此語料庫中的語料量較為簡短，且說話內容是以閱讀句子為主，而非一般的對話形式，造成每位語者的韻律特徵會較有相似之處。此外我們並未針對漢語中的聲調做特別處理，而聲調對於音高軌跡是有極大的影響力，所以韻律特徵的求取並無法很精確獲得。再者，雖然沒有對聲調做正規化，僅利用粗略的韻律特徵卻可對頻譜特徵的系統有不錯的改善，相較於英文中以較高相等錯誤率為基礎的改善來說，韻律訊息在漢語語者驗證的結果應該算是有良好貢獻。

6. 結論

在本文章，我們針對文字特定與文字不特定的語者驗證任務提出我們的語者驗證系統，集合眾多方法之長處來完成，包含了前端特徵正規化的 MVA、後端的改良式測試分數正規化、頻譜特徵的高斯混合與隱藏式馬可夫模型以及韻律特徵的潛在韻律分析與高斯混合模型，而最後的多層感知機更是用來耦合各個相異的驗證系統，使得表現的效果有更進一步的改善。由實驗結果來看，有幾項要點是值得注意的，首先是 MVA 與改良式分數正規化法確實成功地在語者驗證上補償通道不匹配的問題，再者是不論文字特定與文字不特定的情況下，韻律特徵都能與頻譜特徵作完善的結合。而本論文的結果亦顯示了在漢語語者驗證上仍有進一步發掘出關於韻律線索的必要。

7. 致謝

本研究受國科會專題研究計畫及教育部計畫補助，計畫編號分別為 NSC 94-2213-E-027-003 與 A-94-E-FA06-4-4。

8. 參考文獻

1. M. Faundez-Zanuy, E. Monte-Moreno, “State-of-the-art in speaker recognition”, *IEEE Aerospace and Electronic Systems Magazine*, Vol. 20, Issue 5, pp. 7-12, 2005.
2. “NIST - Speaker Recognition Evaluations”, http://www.nist.gov/speech/tests/spk/.
3. The CSLP Speaker Recognition Evaluation (SRE) 2006, http://www.iscslp2006.org/.
4. Chinese Corpus Consortium (CCC), http://www.cccforum.org/.
5. “NIST 2001 Speaker Recognition Evaluation - Extended Data task”, http://www.nist.gov/speech/tests/spk/2001/extended-data/.
6. D.A. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones and B. Xiang, “The SuperSID project: exploiting high-level information for high-accuracy speaker recognition,” *Proc. ICASSP’03*, Vol. 4, pp. 784-787, 2003.
7. A.G. Adami, R. Mihaescu, D.A. Reynolds and J.J. Godfrey, “Modeling prosodic dynamics for speaker recognition,” *Proc. ICASSP’03*, Vol. 4, pp. 788-791, 2003.
8. Z.H. Chen, Z.R. Zeng, Y.F. Liao, and Y.T. Juang, “Probabilistic Latent Prosody Analysis For Robust Speaker Verification,” *ICASSP’06,* 2006.

9. C.P. Chen and J. Bilmes, “MVA Processing of Speech Features”, to appear in *IEEE Trans. on Speech and Audio Processing.*
10. D.A. Reynolds, T. Quatieri and R. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digital Signal Processing*, Vol. 10, pp. 19-41, January 2000.
11. D. Sturim and D.A. Reynolds, “Speaker Adaptive Cohort Selection for Tnorm in Text-Independent Speaker Verification”, *ICASSP’05,*2005.
12. LNKnet Pattern Classification Software, http://www.ll.mit.edu/IST/lknet/.
13. J. Pelecanos and S. Sridharan, “Feature Warping for Robust Speaker Verification,” *Proc. ISCA Workshop on Speaker Recognition - 2001: A Speaker Oddyssey*, June 2001.
14. A. de la Torre, A. M. Peinado, J. C. Segura, J. L. P-C, M. C. Benitez, A. J. Rubio, “Histogram Equalization of Speech Representation for Robust Speech Recognition,” *IEEE Trans. on Speech and Audio Processing*. Vol. 13, pp. 355-366, May 2005.
15. T. Hofmann, “Probabilistic Latent Semantic Analysis,” *Proc. Uncertainty in Artificial Intelligence* 1999, 1999.
16. T. Hofmann, “Unsupervised Learning by Probabilistic Latent Semantic Analysis,” *Machine Learning*, pp. 177-196, 2001.
17. K. Sjölander, “Snack sound toolkit”, http://www.speech.kth.se/snack/.

Personalized Optimal Search in Local Query Expansion

Shan-Mu Lin[1] and Chuen-Min Huang[2]

[1]Dept. of Information Management, National Yunlin University of Technology and Science, Douliou, Taiwan
[2] Dept. of Information Management, National Yunlin University of Technology and Science, Douliou, Taiwan

Abstract

Query Expansion was designed to overcome the barren query words issued by user and has been applied in many commercial products. This treatment tries to expand query words to identify users' real requirement based on semantic computation. It may be critical to deal with the problem of information overloading and diminish the using threshold, however the modern retrieval systems usually lack user modeling and are not adaptive to individual users, resulting in inherently non-optimal retrieval performance. In this study, we propose the LLSF method based on each individual search history to automatically generate specific personalized profile matrix. By which to generate context-based expanded query words. Considering the accuracy of retrieving performance, we process query words re-weighting algorithm to achieve this goal. Finally, the documents list is ranked by the way of stressed density distribution modeling. And the experimental result shows that our framework corresponds to personalization and the performance is very promising.

Keywords: Personalization, Latent Semantic Indexing, Query Expansion, Relevance Feedback, Maximum Entropy Density Function.

1. Introduction

1.1 Research Background and Motivation

The widespread usage of search engines has grown for many years, The searching technique can be used to apply in various aspects, either in World Wide Web or in particular Information Retrieval (IR) database. People can find what they want through it in the world of information overloading. For all of these reasons, search engine now becomes indispensable in the modern lives.

Traditional search engine presents search result based on keyword matching of users' query. It is the simplest method to gather documents associated with specific keywords. However, it's possible for users to acquire undesired results due to inadequate acquisition.. One of these problems is most common users of IR systems type short queries. (Shen et al., 2005) From a sin*gle query, however, the retrieval system can only have very limited clue about the user's information need.*

Mostly, it's hard for users to realize what actual searching requirement is when proceeding searching activities, so there are certain of measure to solve the problem about vague queries issued from users in the past years.

1.2 Research Objective

In this article, we try to provide users more selective query keywords which are related to original query not only for the suggestion but also to help users realizing the real requirement in searching behaviors when users submit too brief query to find out more wanted documents. In addition, we also decide to deliver the decision making authority to users of which documents seem to be more preferred by them for achieving user center approach. Otherwise, in order to provide each user with more personal searching environment and contents, we endeavor to propose several approaches to adapt search results according to each user's information need.

2. Related Work

2.1 Review of Refining Short Query

To solve the problem of low retrieval performance caused by inappropriate query terms, automatic query expansion techniques have been studied for the past 30 years. In a recent study (Jansen et al., 2000,March 1), the number of query terms used by most end users was no more than 2 when searching with a Web search engine, which is even less than that of searching online databases. The same study also pointed out that only 5% of queries were accompanied by any relevance feedback feature.

2.2 Categories of Query Expansion

Query expansion techniques fall into two categories according to the way of implementation. One is to add new terms to an original query before searching, and the other is to formulate a new query on the basis of some retrieved documents of the previous search (Qiu, 1995). While the former is usually called a global or corpus-specific query expansion, the latter is called a local or query-specific query expansion. Global query expansion rely on thesauri that is a manually-built resource, as though WordNet-based (Mandala et al., 1999) provides the relation types include coordination, synonyms, hyponyms and etc for expanding the feature of original query terms.

Local query expansion, which corresponds to feedback retrieval, can acquire relevance information by either user feedback (Robertson & Sparck Jones, 1976; Rocchio, 1971) or system feedback. Query expansion using user feedback based on relevance judgment made by users, brought a significant improvement in retrieval performance (Harman, 1992 June; G. Salton & Buckley, 1990). Another two theories about local query expansion is Local co-occurrence method (HE et al., 2002; ZHANG et al., 2002) and Latent Semantic Indexing (LSi_based) (Deerwester et al., 1990)

2.3 Personal Data Construction

There are several way to gather the user's information for constructing unique data each end user belongs to. One of these approach is to have users describe their general interests. For example, Google Personal asks users to build a profile of themselves by selecting categories of interests. Google's PageRank algorithm can be described as personal web search techniques augmenting traditional text matching with a global notion of "importance" based on the linkage structure of the web. This global notion of importance can be specialized to create personalized views of importance.

User profile data provide information about the users of a Web site. A user profile contains

demographic information (such as name, age, country, marital status, education, interests, etc.) for each user of a Web site, as well as information about the users' interests and preferences. Such information is acquired through registration forms or questionnaires, or can be inferred by analyzing Web usage logs (Eirinaki & Vazirgiannis, 2003, February). Personal profiles can also be combined with the method mentioned above in the context of the Web search to create a personalized version of PageRank for setting the query-independent priors on Web pages. (Teevan et al., 2005). (Liu et al., 2002) used a similar technique for mapping user's queries to categories based on the user's search history.

3. Interactive IR system

Generally in interactive situation, system collects user's intention through designed interactive interface. In principle, every action of the user can potentially provide new evidence to help the system to better infer the user's information need. Thus in order to respond optimally, the system should use all the evidence collected so far about the user. After collection of the user information, how to effectively select and analyze these data is critical to this kind of system.

To retrieve more user demanded results, we carry out Linear Least Squares Fit (LLSF) algorithm to generate personal profile by matrix combination in which the personal searching result will be formed in document-term (DT) matrix, and Singular Value Decomposition (SVD) is used to reduce the dimensions of the original DT matrix. Moreover combining of document-cluster matrix and decomposed DT is to produce the final user profile M matrix. This process is also called Latent Semantic Indexing, which could extract the context-based terms out for expanding personalized query terms.

Simultaneously, as far as possible to promote the retrieval accuracy, relevance feedback of probabilistic model is suitable to be involved in. And in the traditional Retrieval method likes TF*IDF weighting schema, existing problem of mis-weighting could be caused the poor retrieval result. To overcome this defect, we adopt smoothing function of TF*IDF which could be diminished the inadequate weighting result. Finally we try to optimize the result representation, the ranking algorithm is also seen to be critical. For improving the Term Frequency (TF) ranking model, ranking function considering density distribution is brought into our framework.

3.1 Retrieval Method

After word recognition, each document is represented as a bag of words, but it does not mean that every word is a meaningful unit. For subsequently retrieval purpose, we need to set every recognized term a appropriate weight.

Furthermore, when users try to issue single or shorter query for searching, we use traditional keyword matching method to catch documents indicated by the user as relevant and conduct query expansion from these first time extraction documents in which we expect to get a list of longer query words, then we carry out traditional vector space model (VSM) to extract more query-relevant documents for generating more user demanded queries.

Because of the classical term weight model, TF*IDF scheme, usually has mis-weighting problem.

For example, a single document that contains the word "ERP" which only appears one time should not be deemed as relevant to a query containing "ERP" as a longer article that contains 20 occurrences of the word "ERP". On the other hand, we ought not to assume that the longer document is 20 times more relevant. For this reason we prefer a smoothed version of TF and IDF(Croft & Harper, 1979) as listed below:

A common term frequency (TF) expression is then modified:

$$TF = \frac{f(K+1)}{f+KL} \tag{1}$$

where L = the normalized length of document D. If the document is of average length, then L = 1.0. K = a constant, usually set between 1.0 and 2.0. f = specific term occurs in single document. The TF component is designed to increase in value quite modestly as f arises. For instance, if f ,K and L are 1, then TF = 1.0. If f were 9, then TF = 1.8. We can properly avoid the mis-weighting problem of conventional TF through this kind of effort. Smoothing inverse document frequency (IDF) prevents division by zero in the case where a term does not occur in the document collection at all.

$$IDF_t = \log(\frac{N-n_t+0.5}{n_t+0.5}) \tag{2}$$

where N = the size of the collection, n_t = the number of documents containing a given term, t.

3.2 Ranking Result

As noted above, a Boolean search generally returns sets of documents that are unordered, or ordered by certain criteria unrelated to relevance, such as time or date.

Most Web search engines are based on a different technology that ranks search results based upon the frequency distribution, term frequency, of query terms in the document collection. To cite an instance, if a document contains many occurrences of a query term "ERP", this suggests that the document might be highly relevant to a query like "There are many software providers have ERP solutions, and the follow name lists which is one of the ERP providers?"

For this reason, we consider several criteria to consider document ranking score, then we expect document which is more relevant to user's demand will be rank in higher place through sorting specific ranking score. The viewpoint of our criteria separated into four factors between single keyword and individual document. There respectively are similarity, density, term frequency and title appearance.

3.2.1 Similarity

We retrieve documents in VSM-model by comparing similarity information *sim(k,d)* among a keyword k and a document d, then defining a positive threshold value for judging which one passing this value is seen to be relevant. So the single item gets a higher similarity value that we have confidence which one is more relevant to issued query keyword.

3.2.2 Maximum Entropy Density Function

By contrast, conventional ranking technology gives score to documents merely considered term frequency and regardless of the density distribution of specific keyword in subject document. But if

terms stated to be highly concentrated, it maybe mean that some topic is intensely described somewhere. So we carry out Maximum Entropy Function used to examine the density distribution of query keyword k, instead of just term frequency in considering document score. The original equation as formula (3) below, the value of E(K) becomes higher when p(k) in a average value that means probability distribution of k is more steady; E(K) has a lower value when p(k) is extremely in high and low value.

$$E(K) = -\sum_{k \in K} p(k) \log p(k) \tag{3}$$

So the entropy equation is revised to formula (4) (K. F. Jea & P. Y. Hsu, 2000), for ensuring the state between E(K) and p(k) is positive in synchronous up and down.

$$E(K) = -\sum_{k \in K} p(k) \log[1 - p(k)] \tag{4}$$

In physics, the meaning of density is that the degree of object distribution in the unit space. Accordingly considering the keyword density distribution in unit length of document will be more closed to reality and achieve the normalization.

After normalization adjustment, entropy equation is represented as follows formula (5):

$$E_s(K) = -\sum_{i=1}^{n} \frac{p_i(k) \log[1 - p_i(k)]}{S_i} \tag{5}$$

where Pi(x) = the occurrence probability of term k in sentence i, Si = the length of sentence i, n = number of sentences in a document. By this treatment, we can differentiate when document with same term frequency of query keyword, and then rank them by density distribution consideration.

3.2.3 Term Frequency

Although term frequency (tf) is basis to rank the documents, high occurrence of keywords in a document indicates that the weight of this document is remarkable significance. Therefore, we also adopt concept of term frequency to ensure our ranking model. But basic tf weighting method emerges the problem of mis-weighting, likes mentioned before, so we transfer raw $tf(k_i, d_j)$ into normalization according to maximum frequency of any term $Maxtf_j$ in a document d_j.

3.2.4 Title Appearance

When author composes particular topic, title often brings out overall theme or subject within article content. People surf on a search engine or even read news article, using title to decide whether to enter a website or further read an article they are interested is always an obviously evidence that these titles engage their concern. In the other words, if the numbers of query keywords k in a document's title *t* have a higher frequency $f(k_i, t_j)$ means that this article is considered to be more relevant by the user. We formulate an equation of this concept as $W_T * f(k_i, t_j)$, where W_T is a constant we can adjust to determine the weighted stress of this factor.

3.2.5 Rscore Ranking

To sum up these ranking factors, we merge these variables into single equation as formula (6):

$$R_{score} = F(k_i, t_j) * [sim(k_i, d_j) + E_s(K) + \frac{tf(k_i, d_j)}{Maxtf_j}]$$

$$F(k_i, t_j) = \begin{cases} 1, & f(k_i, t_j) = 0 \\ W_T * f(k_i, t_j), & f(k_i, t_j) \geq 1, ...n \end{cases} \quad (6)$$

where $f(k_i, t_j) \in N$, $W_T \in N$, W_T = weighting stress for occurrence of keyword in a title, $f(k_i, t_j)$ = occurrence of keyword k in a title t, $sim(k_i, d_j)$ = similarity value between keywords k and a document d, $E_s(K)$ = sum of each keyword's entropy value in a document d, $tf(k_i, d_j)$ = the frequency of keyword k in a document d, $Maxtf_j$ = the maximum frequency of any keyword k in a document d.

3.3 Query Expansion

3.3.1 Probabilistic Models of Query Expansion

In a probabilistic framework, selecting terms and computing relevance weights are treated as two different problems. This model is used to compute more accurate weight estimates. Consider the term incidence contingency table in Table 1.

Table 1. Term Incidence Contingency Table (Jackson & Moulinier, 2002)

	Relevant	Non-relevant	Total
Containing the term	r	n-r	n
Not containing the term	R-r	(N-n)-(R-r)	N-n
Total	R	N-R	N

where N = the number of documents in the collection, R = the number of relevant documents for this query, n = the number of documents having term t, r = the number of relevant documents containing the term t. The term weight from the equation which we mentioned above, would then be modified to take account of the relevance information as follows:

$$w'_{t,d} = \frac{f(K+1)}{f+KL} \log \frac{(r+0.5)(N-n-R+r+0.5)}{(R-r+0.5)(n-r+0.5)} \quad (7)$$

We utilize this re-expressed formula to re-weight the term within the vector space model when the user explicitly checks the retrieved document seen to be relevant or non-relevant. Subsequently, here address how terms are selected for expanding activity. In table 2 , we can obviously observe the post weighted scores are risen when choosing a list of documents relevant to the topic of "ERP", "PeopleSoft", "SAP" and "Oracle" and non-relevant to "J.D.Edwards", "鼎新", "軟體部" and "Siebel".

Table 2. Comparison of Re-weight activity

	ERP	PeopleSoft	SAP	Oracle	J.D.Edwards	鼎新	軟體部	Siebel
Initial Weight	2.718	3.365	2.792	2.681	5.953	5.302	4.117	4.013
Re-weight	3.714	5.329	4.752	3.253	5.986	4.192	3.390	3.886

This model discussed by Robertson(1990) considers the distribution of scores for relevant and non-relevant documents. The model leads to an "offer weight", the larger the offer weight, the better the candidate, which is used to rank candidate terms.

$$OW_t = r_t \log \frac{(r_t + 0.5)(N - n_t - R + r_t + 0.5)}{(R - r_t + 0.5)(n_t - r_t + 0.5)} \tag{8}$$

This two model proposed by Robertson tightly integrates query expansion using relevance feedback and probabilistic retrieval.

3.3.2 *LLSF Models of Query Expansion*

At the beginning of the second stage expansion, we prefer to take user profile that not only has benefit to provide extra information about personal search intention, but also greatly reduce the falsehood of retrieved result. Furthermore, we adopt algorithm with respect to noise reducing, the Linear Least Squares Fit (LLSF) method proposed by (Liu et al., 2002), to construct matrix as personal user profile.

In order to have one of the matrixes, we first need to introduce our cluster method with respect to Single-Pass Clustering and 2-way K-Nearest-Neighbors (KNN) of Topic Detection and Tracking (TDT).

3.3.2.1 TDT Clustering

1. Detecting New Cluster

A large number of clustering methods were studied in IR research. This section we adopt the TDT proposed by CMU (Tang et al., 1999). One of the two algorithms is Single-Pass Clustering (SPC) for clustering task and the other is 2-way K-Nearest-Neighbors (KNN) for automatic classification. Figure 1. demonstrates SPC flow chart.

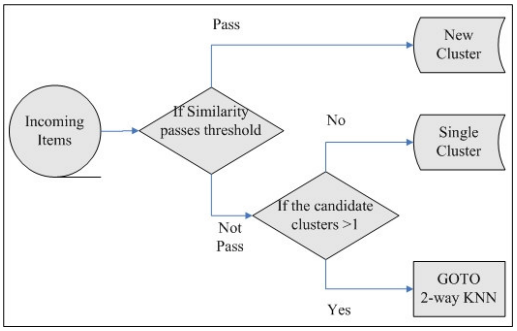


Figure 1. Flow chart of Single-Pass Clustering

Single-pass clustering follow the process as listed below and apply cosine as similarity calculation function: (1) Above all, taking out the first item in document collection as the first cluster. (2) Then take out the second item, calculating the similarity between item and clusters have been created. (3) If there is no similarity passes the threshold, instantaneously letting the incoming item be a new single cluster. (4) If the similarity passes the threshold we just set before, therefore categorize incoming item into appropriate candidate cluster. (5) If step 4 is selected, rescoring the centroid vector space of this cluster. (6) Iterating step 2 to 6, until dealing with entire incoming items.

2. Automatic Classifying

2-way KNN in TDT is used to classify the incoming item into proper classification by computing the relevance score. Which refers to compare objective cluster and else cluster that both take numbers of k Nearest-Neighbors. Objective clusters with respect to documents in this clusters which are prepared for

comparison; else cluster means documents in the clusters which different from objective clusters in the candidate clusters. Formula (9) explains the calculation of relevance score.

$$relevance\_score(\vec{x}, kp, kn, D) = \frac{1}{|U_{kp}|}\sum_{\vec{y}\in U_{kp}} \cos(\vec{x}, \vec{y}) - \frac{1}{|V_{kn}|}\sum_{\vec{z}\in V_{kn}} \cos(\vec{x}, \vec{z}) \quad \textbf{(9)}$$

By picking up the maximum relevance score which specific cluster belongs to, we can estimate this one is suitable to chosen for incoming item.

3.3.2.2 Algorithm to Learn Profile

1. Singular Value Decomposition

With regard to the meaning of SVD, we discuss it as follows. If there is a high dimension data, it can be applied SVD for dimension diminishing. In the linear algebra, SVD has a special characteristic to transform a high dimensional data to lower one. This method is often called matrix decomposition. By this way, high dimension matrix could be reduce to lower one then even achieve rule and noise reduction via selection singular value in diagonal matrix. The potential power of SVD is which can attempt to estimate the hidden structure and discover the most important associative patterns between words and concepts. Figure 2. demonstrates the process of the SVD:

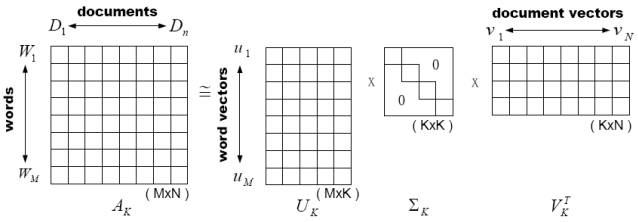


Figure 2. Singular Value Decomposition (SVD)

2. Rank selection

In re-composition process there is a critical point has to be taken notice. The diagonal matrix, we have to select precise rank k for diminishing the noises effectively. And how many rank k we should decide? One of these methods is to observe the singular value when they felled down from violent to smooth, and the previous of the margin value is the best choice. e.g. As the dotted line in the diagram below, there is a margin value k=9 for rank adoption.

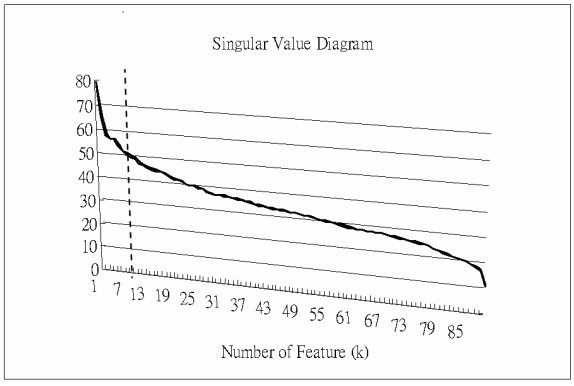


Figure 3. Singular Value Variation Diagram

3.3.2.3 Constructing the User Profile in Matrix Feature

The learning equation mentioned above is concerned with concept of Latent Semantic Analysis (LSA) or Latent Semantic Indexing (LSI) (Deerwester et al., 1990). LSI is a theory for extracting and representing the relationship of words in a large corpus of text by using the co-occurrence of words and a mathematics technique, Singular Value Decomposition (SVD). In addition, there has another statement declared is LSI which could overcome crucial defect happened in searching process. This method projects documents and words to a predefined space, finding out the latent relationship between terms and documents. Even can retrieve the relevant documents when the situation that searching keywords is not appeared.

Given the m-by-n document-term matrix DT and the m-by-p document-cluster matrix DC, the Linear Least Squares Fit method computes a p-by-n cluster-term matrix M. In this step, techniques solving the problem is to employ the concept of Latent Semantic Index (LSI) in which Singular Value Decomposition (SVD) is the mathematical measure to decompose the input matrix. By this measure, DT is decomposed into the product of three matrixes $U_k * \Sigma_k * V_k^T$, where U_k and V_k are orthogonal matrices and Σ_k is a diagonal matrix. After such decomposition, we can straightforward to recompose and combine DC matrix for computing particular matrix M, $M = DC^T * U_k * \Sigma_k^+ * V_k^T$, where Σ_k^+ is the inverse of Σ_k. Figure 6. illustrates the process of learning profile M:

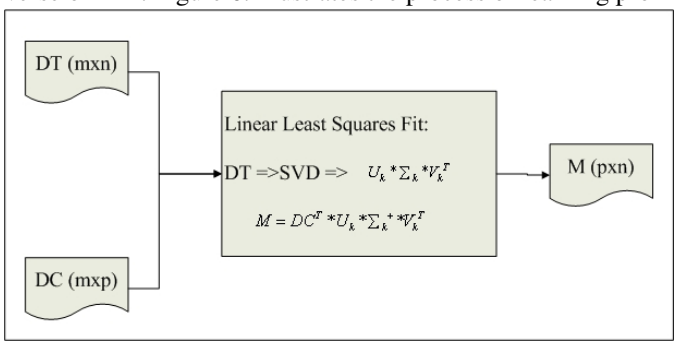


Figure 4**.** Process of learning profile M

We use matrices to represent the user's search histories, clusters of documents and user profiles as following Table 3.

Table 3**.** Document-Term matrix (DT)

Term / Doc	.Net	Exchange Server 2003	Outlook	Palm	PocketPC	Palm OS
D1	5.0222	7.1262	3.3484	0	0	0
D2	0	11.7001	8.4147	0	0	0
D3	4.3711	10.4553	6.8124	0	0	
D4	0	0	0	1.6325	0	0
D5	0	0	0	4.3454	6.5116	3.3619
D6	0	0	0	4.4678	5.6560	2.4723

Matrix DT (m*n) is a document-term matrix, m is the number of documents considered relevant by the user in a user's search history and n is the number of distinct terms occurring in these documents, which is established from user's query and the retrieved relevant documents the user indicates explicitly. The value of DT(i,j) is determined by the modified TF*IDF scheme.

Table 4. Document-Cluster matrix (DC)

Cluster / Doc	Cluster 1037	Cluster 4194
D1	1	0
D2	1	0
D3	0	1
D4	0	1
D5	1	0
D6	0	1

Matrix DC (m*q) is the document-cluster matrix, which is established from the relationships between the clusters and the documents. For each row in matrix DT, there is a corresponding row in the matrix DC. The columns of DC are the set of related clusters. If a row in DT represents a query/document, then the corresponding row in the matrix DC represents the set of clusters related to the query/document. Moreover, if there is an edge between the y-th cluster and the x-th query/document, then the entry DC(x,y) = 1; otherwise it is 0.

Table 5. Cluster-Term matrix M expresses a user profile

Term / Cluster	.Net	Exchange Server 2003	Outlook	Palm	PocketPC	Palm OS
Cluster 314	1.3097	3.2141	2.0575	0	0	0
Cluster 184	0	0	0	1.3215	1.5242	0.7774

We have learned a matrix M (p*n) from DT and DC, which is represented as the user personal profile. In this example, “Cluster 314” and “Cluster 184” are cluster field; “.Net”, “Exchange Server 2003” and “Outlook”…etc are term field.

3.3.2.4 From Profile to Expansion

Following the upper step, we have constructed personal profile in cluster-term matrix format from search’s history and latest relevant documents. Terms in the same cluster means that the relation among them are strongly recognized as Table 5., it can be used for expansion purpose when the one of the query keyword is appeared in this term list. This activity is described in Figure 5.

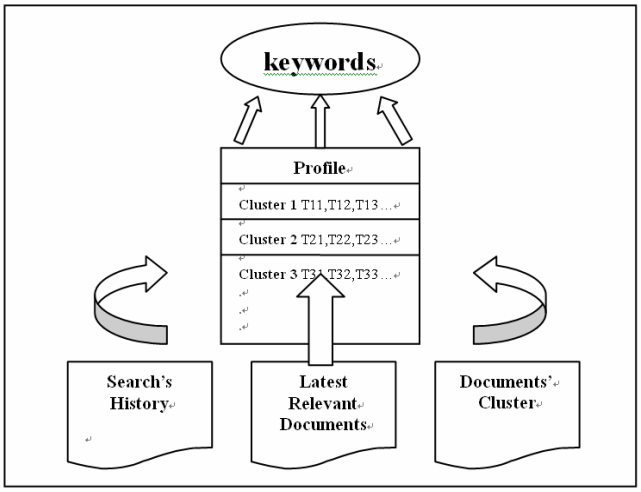


Figure 5. Concept of Profile Expansion

4. PNQES: A Personalized Search Engine

4.1 Design

To experiment with the personalized environment, we create the PNQES search engine. This personalized agent could provide the user a query expansion function which is separated into two stages. Above all, system will automatically catch and parse the query terms when the user has submitted completely. After parsing, search component with VSM-based search going to weight each words in the query according to vector space model (VSM) strategy for retrieving all possible documents related to the original query.

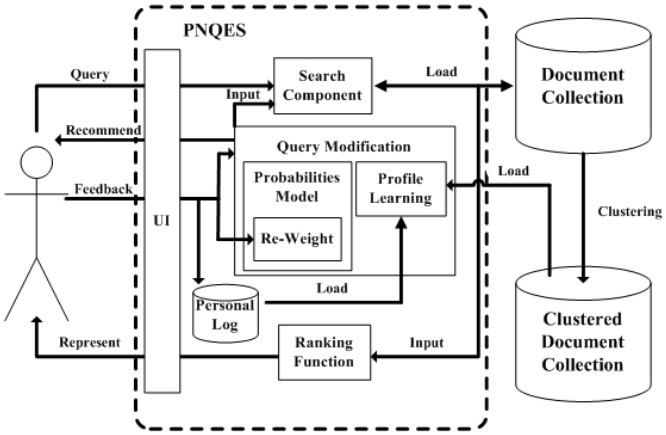


Figure 6. NQES Architecture

Next step, ranking component has considered of several factors which influence the ranking result with similarity, dense distribution, term frequency and title occurrence. Simultaneously this ranked result would be stored in personal log file for later analyzed.

For finding potential search's intention, system will ask the user to respond some feedback, called "Relevance Feedback", with judgment whether documents set is relevant or not while showing the retrieved result on the screen. In the meanwhile, when a user summits this response, Query Modification component adopted probabilistic model is able to give all terms in all relevant items with "Offer Weight" and then output some candidate terms in first stage expansion. In the parallel step, LLSF component combines personal search's history and pre-clustered corpus applied TDT algorithm to construct matrix called personal user profile in second stage query expansion.

Lastly, query issued by the user will be expanded to a number of proper personal keywords via this two stage expansion processes. Figure 6. is illustrated PNQES system architecture.

4.2 Experimental Data Sets

The experiment target where we focus on is enterprise technology reports because most terms contained in are consisted of proper noun in which the experimental result can seem to be more accuracy. Moreover we try to collect data from Website, Taiwan.CNET.com, which contains various documents associated with specific software techniques and hardware information and the reported

date from January 4, 1999 to April 31, 2006.

This corpus has been separated to two main classes, Enterprise Application and News, and the 9 sub-topic. The volumes distribution of these topics is shown in Table 6.

Table 6. Volumes of Corpus

Sub-Topic	Category	Numbers of Article
IT techniques	EA	1000
Special Topic Report	EA	250
Case Study	EA	550
Special Column	EA	1137
Research Report	EA	1511
Enterprise Software	News	6660
Enterprise Hardware	News	4991
Network/Communication	News	2796
3C Product	News	3694
Total Volumes		22988

p.s: EA: Enterprise Application

4.3 Word Recognition

Word segmentation is crucial for the research of information retrieval, especially for Chinese documents. The reason is that there is no word boundary in sentences, which increases the difficulty of this work. In this research, we extract Words with respect to verbs and nouns in “Eighty Thousand dictionary” that is published by Institute of Information Science Academia Sinica, then merge them into another dictionary possessing names of location, Institute and company and gathered by our laboratory. Moreover, we extract terms from a document by principle that treats the long-term has a higher priority than others, when this step is over, next we apply the newest version of word segmentation system developed by Chinese Knowledge Processing Group (CKIP) to pick up the rest terms that Words database can not capture for ascending the precision of recognition.

5. Evaluation of PNQES

5.1 Evaluation Method

5.1.1 Experimental Subject

Because of our experimental corpus is focused on IT related articles, testers have be expected to hold the professional IT knowledge of how to realize which article topic is their demand one. For this reason, we plan to ask 10 users who both are the graduate students and major in the department related to “Computer Science”.

5.1.2 Evaluation Variable

The evaluation step symbols are described as : (1)*QTR*: Query Term Re-weighting, (2)*LLSF*: Linear Least Square Fit Expansion Model, (3)*Baseline*: We execute raw query expansion without query term re-weighting and Linear Least Square Fit Expansion Model analysis, then using NAP (Non-interpolated Average Precision Rate) to evaluate the precision values.

5.1.3 Evaluation Procedure

And then we design an evaluation procedure with regard to retrieval and ranking precision. The brief

evaluation process has listed in table 7.

Table 7. Brief Evaluation Process

Event Number	Evaluation Event
E_01	Precision in baseline event
E_02	Precision in baseline + QTR event
E_03	Average R-precision in TF ranking model
E_04	Average R-precision in Rscore ranking model
E_05	Precision LLSF model based on baseline

5.1.3.1 Description of Evaluation Processes

1. System Initiation:

To actually simulate the real condition, we request tester to input single query to initial the system. Firstly, system will retrieve the keyword related articles by "Boolean AND search" method, and all articles which contain this keyword will be retrieved and ranked in TF ranking approach.

2. Query Expansion:

While possibly related items have been retrieved, the testers will be asked to indicate several articles which they think to be relevant and are explicitly stored into personal log, furthermore the submitted query will be expand from analyzing these relevant items by probabilities model. The testers at will select a number of recommended keywords with scattering issues and then add them to original query list for the following search. The expanded query list will conduct VSM search for retrieval task.

3. Training and Ranking:

In ranking test, we ask tester to interact with our system for 3 time as step 2 for successful training the query list to robustness, so that the after training list is for doing the baseline task. Further, testers who evaluate the TF ranking and Rscore ranking also based on this baseline with the evaluation formulation of ranking function "Average R-precision".

4. Methods Integration and Profile Recording:

The following retrieval tasks we increase one variable "QTR" to our evaluation activity with the ranking function Rscore measure. This step has two objectives, one is to evaluate the usability of variable QTR and the other is to as far as possible make increasing of interaction between tester and system for establishing personal search's history completely.

5. Two Expansion Methods Comparison:

When tester searching behavior has been fully caught, we ask testers to evaluate the results expanded by LLSF model based on baseline to contrast the variation of two models that baseline means rarely using the probabilities model and the latter means a hybrid expanding activity with two expanding model.

5.1.4 Retrieval and Ranking Statistics

First, the evaluation functions we referred is TREC_EVAL method developed by Buckley(1991). We have altered to fit condition of our experiment. Table 8. is the example of one of our scored cards. The rest cards we have appended to appendix A.

Table 8. EVAL Scored Card A

User 1 Queryid (Num) 1 Precision for all relevant documents	
E_01	0.8333
E_02	0.9286
E_03	0.8846

EVAL Scored Card B

User 1 Queryid (Num) 1 Total number of documents over all queries	
Retrieved:	23
Rel_ret:	20
Precision: (in TF ranking)	
At 5 docs:	0.8000
At 10 docs:	0.8000
At 15 docs:	0.7333
At 20 docs:	0.7500
At 30 docs:	0.6667
Average R-Precision:	0.7500
Precision: (in Rscore ranking)	
At 5 docs:	1.0000
At 10 docs:	1.0000
At 15 docs:	0.9333
At 20 docs:	0.9000
At 30 docs:	0.6667
Average R-Precision:	0.9000

p.s: Rel_ret: Retrieved Relevant document

In the experiment, we have recorded retrieved result set and relevant items from evaluation process E_01 to E_05. After recorded, we turn the data to scored card format and draw the bar chart for observing if each effect variable has been added respectively, the result precision will be changed significantly. The evaluation method we have adopted the precision for result retrieval and average R-precision for result ranking. We also have compared whether or not the Rscore Function is significantly better than Term Frequency (TF) ranking method.

5.2 Experiment Result

1. Result of Each Retrieval Methods:

First of all, we examine the baseline and of combining the QTR as E_01 and E_02 to observe the variation of precision. The variation of each variable appended is demonstrated in Figure 19. We can see E_02 significantly outperform the baseline. It is clearly demonstrates that it is worthwhile to combine the QTR to yield higher retrieval precision.

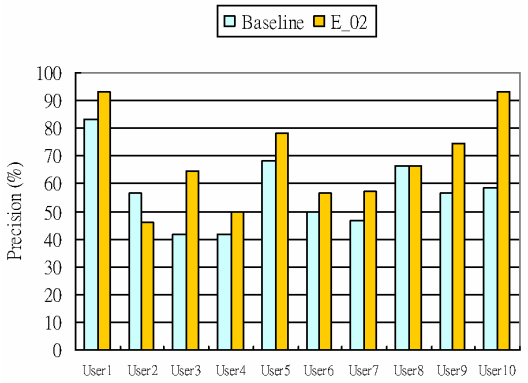


Figure 7. Precision of different combining methods to 10 users

2. Result of Two Expansion Models:

Then add UP to evaluate the personalized search as E_05. Another observation from Figure 8. is that using the UP to revise the expanding terms, this approach gives extraordinary precision value than rarely using classic probabilities model alone. This tends to imply that the personal profile is worthwhile to perform personalized search.

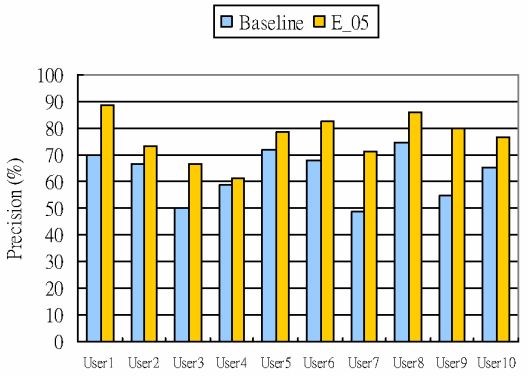


Figure 8. Results of adding the UP analysis

3. Result of Two Ranking Methods:

Distribution of Average R-Precision value presented in Figure 9., applying Rscore Function obviously performs a higher average precision than original TF ranking measure. So we firmly trust that consideration of several factor mentioned in section 3.2 when undertakes ranking task will induce the performance improvement.

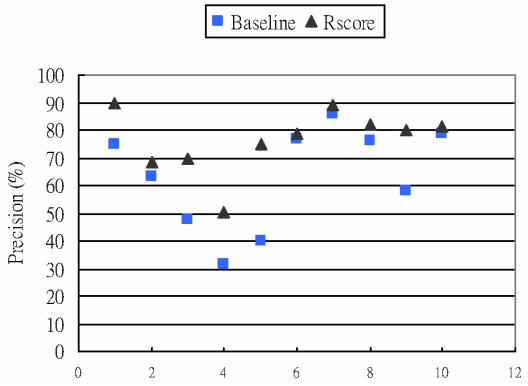


Figure 9. Average R-Precision with two ranking method

6. Conclusion

In this paper, we propose a mechanism which can be intelligent to learn the user's search behavior and provide specific search results for each differentiated end-users. To achieve this purpose, adopting 2 stages query expansion and hybrid density distribution ranking function is our efforts.

Query Expansion activity in first stage we have applied probabilities model which takes that the expanding and weight re-calculating as different parts, both are based on the relevant documents of user's feedback. While in stage 2 expansion, system initiatively combine the personal profile and latest relevant items indicated by the user and transform with respect to LLSF metrics merging procedure to extract out more suggested terms of user-driven's. As the list of documents have been retrieved

completely, so as to show the most relevant items for the user, ranking method we have considered several influence factor to give the appearance priority to each items.

Furthermore we utilize evaluation criteria to prove our PNQES is of feasible and effective. And the result performance has proved this proposed system framework not only could be applied in local database, but also could be well-performed in web-based searching for personalization enhancement.

7. Reference

1. Deerwester, Fumas, Landauer, & Harshman. (1990). *Indexing by intent semantic analysis.* Paper presented at the JASIS.
2. Eirinaki, M., & Vazirgiannis, M. (2003, February). *Web mining for web personalization.* Paper presented at the ACM Transactions on Internet Technology (TOIT).
3. Harman, D. (1992 June). *Relevance feedback revisited.* Paper presented at the Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, New York.
4. HE, H.-z., HE, P.-l., GAO, J.-f., & HUANG, C.-n. (2002). Query expansion based on the context in chinese information retrieval. *Joumal of Chinese Infomation Processing, 16*(6), 32-37.
5. Jansen, B. J., Spink, A., & Saracevic, T. (2000,March 1). Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing & Management, 36*(2), 207-227.
6. Liu, F., Yu, C., & Meng, W. (2002, November 4--9). *Personalized web search by mapping user queries to categories.* Paper presented at the In Proceedings of CIKM, McLean, Virginia, USA.
7. Mandala, R., Tokuanga, T., & Tanaka, H. (1999). *Combining mutiple evidence from different types of thesaurus for query expansion.* Paper presented at the SIGIR.
8. Qiu, Y. (1995). *Automatic query expansion based on a similarity thesaurus.* ETH Zurich.
9. Robertson, S. E., & Sparck Jones, K. (1976). Relevance weighting of search terms. Journal of the *American Society for Information Sciences, 27*(3), 129 - 146.
10. Rocchio. (1971). Relevance feedback information retrieval. In *The smart retrieval system-experiments in automatic document processing* (pp. 313 - 323). Kansas: Prentice-Hall.
11. Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *41*(4), 288 - 297.
12. Shen, X., Tan, B., & Zhai, C. (2005, Augest 15-19). *Contextsensitive information retrieval usingimplicit feedback.* Paper presented at the SIGIR, Salvador,Brazil.
13. Teevan, J., Dumais, S. T., & Horvitz, E. (2005, August 15–19). *Personalizing search via automated analysis of interests and activities.* Paper presented at the SIGIR, Salvador, Brazil.
14. ZHANG, M., SONG, R., LIN, C., MA, S., JIANG, Z., LIU, Y., et al. (2002). *Expansion-based technologies in finding relevant and new information.* Paper presented at the TERC.

以本體論為基礎之新聞事件檢索與瀏覽

許孟淵 [1]，黃純敏 [2]
g9323703@yuntech.edu.tw[1]，huangcm@yuntech.edu.tw[2]
[1] 資訊管理系，雲林科技大學，斗六，台灣
[2] 資訊管理系，雲林科技大學，斗六，台灣

摘要

當前電子新聞的瀏覽，存有以下缺點：(1)新聞文件的瀏覽缺少以事件觀點來加以呈現 (2)電子新聞專輯的內容不包括新聞多文件摘要(Multi-Document Summarization) (3)欠缺社會大眾與網友對於該新聞的評論與看法。當讀者欲全盤掌握新聞內容事實時，須額外找尋數個新聞網站來比較整理，以得到特定新聞事件全貌；此外新聞報導具備前因後果的特性(如白米炸彈客事件到後續處理)，以現今新聞入口網站所提供的瀏覽功能而言，並無法滿足讀者的需求。

本研究主要藉由本體論(Ontology) 理論，提出適性模型來處理電子新聞，提供給讀者更易了解新聞事件發展始末的新聞呈現方式。研究中首先利用事件偵測(Event Detection)與追踪(Event Tracking)之群聚技術，產生新聞事件群集；之後運用自動建構出的新聞本體論及應用到主題地圖(Topic Map)理論的主題地圖索引萃取模型，針對單一事件找出其中蘊含的人、事、時、地、物等主要概念，形成事件中的主要議題及其關聯，並以圖解方式的主題索引地圖來呈現事件中所涵蓋之議題和關聯。本研究另一重點，以建構出的新聞本體論為基礎，找出概念間連結，針對新聞內文中重要字詞加權，擷取出新聞多文件摘要；另外利用新聞本體論結合事件合併演算法，可將相似新聞事件群集做合併處理，便於讀者瀏覽相關新聞事件發展；最後，新聞本體論的重要概念，會被擷取並做為本體論分類概念檢索之用，可讓讀者瀏覽感興趣的新聞人、事、時、地、物概念，節省讀者寶貴閱讀時間，快速找到新聞事件的重點！本研究將上述技術應用到新聞檢索與瀏覽(News Retrieval and View)的目的，是希冀讓讀者在閱讀電子新聞時，能夠了解到新聞事件發展的始末，以及得到更加精確的資訊檢索結果。

本研究之系統評估採公開發佈方式進行系統測試，評估時程為期五天，共回收 72 份問卷。評估結果顯示，本研究確實能增進新聞事件的呈現內容、改善主題地圖呈現之品質，每項系統評估指標都有七成左右受測者能滿意地接受。研究中所提及的新聞事件檢索與瀏覽機制，得到了多數受測者認可。

關鍵字:本體論、主題地圖、事件偵測、事件追踪、新聞多文件摘要、事件合併、改良式新聞檢索

1. 緒論

1.1. 研究背景與動機

現今線上新聞入口網站，除 Google 提供較為完整的新聞事件分群閱讀方式、及 Yahoo！奇摩新聞網提供類似事件閱讀方式的新聞專輯之外，其他新聞網大多只提供基本新聞分類瀏覽。此外，新聞事件有多人同時撰述特性，缺少客觀論點來描述特定事件，使讀者欲客觀了解一新聞事件，須多方閱讀比較，花費在新聞瀏覽搜尋的時間十分可觀。

上述問題，在先前研究(許登傑, 2005; 黃純敏 *et al.*, 2003)以事件分群分類、多文件自動摘要以及主題地圖視覺化新聞呈現等技術，提出對應的解決方案。然而在多文件摘要的可讀性及正確度、以特定事件為基之主題地圖主題及關聯擷取有意義與否，以及資訊檢索的成效上，受限於並非是基於新聞文件語意為主的處理方式，使得呈現效果上還有改善空間。本研究運用新聞本體論(Ontology)及主題地圖(Topic Maps)理論，利用本體論概念間關聯及定義，針對個別新聞事件產生所屬主題索引地圖，提供讀者創新且符合新聞原文語意的新聞關聯知識瀏覽介面。讀者藉由基於新聞本體論之主題地圖幫助，可快速正確地理解事件內包含到的主題及關聯。在文件字詞處理技術及新聞群聚分類的方法論，本研究沿用過去研究成果(許登傑, 2005)，將線上新聞文件依事件相似度分群分類，找出新聞事件群集，並透過新聞本體論之助，剖析出更正確且具內文代表性之多文件摘要內容。隨後透過新聞本體論與主題地圖索引萃取模型，剖析出特定事件之主題地圖，並藉由主題合併機制之輔助，產生事件之完整知識索引介面；利用新聞本體論，將彼此語意相近的新聞事件合併在一起，類似 Yahoo！奇摩新聞網的「新聞專輯」的方式呈現。最後將本體論重要概念擷取出，並顯示成個別新聞類別中重要的新聞人、事、時、地、物概念，可讓讀者針對有興趣的概念加以檢索。有感於讀者在瀏覽新聞事件上的需要，本研究引入社會大眾對於該特定新聞事件的觀感與評論，收集網友對於該事件相關內容電子佈告欄討論及各大新聞網的相關新聞圖片連結，希望讓讀者對於新聞事件，有更加深入且多元化的看法。

1.2. 研究目的與貢獻

本研究目的，乃是利用新聞本體論，達到：

1. 改善新聞多文件摘要的內文代表性及正確性，提升閱讀的流暢度。
2. 結合事件合併演算法，將概念相近新聞事件合併，供讀者在檢索相關事件的方便性。
3. 利用本體論概念語意剖析與連結，有效解決先前研究(許登傑, 2005)中事件主題地圖之主題與關聯不夠相關的缺點，提高讀者閱讀滿意度。

在資訊檢索上，利用新聞本體論包含之人、事、時、地、物，提供更符合使用者語意的資訊檢索、事件專輯瀏覽功能，並納入社會大眾意見，提供特定事件的多元化觀點，強化閱讀深廣度。

1.3. 研究範圍與限制

Yahoo！奇摩新聞網為新聞資料來源，若新聞內文中夾雜英文詞句，將其標記為外來語而不處理。

2. 文獻探討

2.1. 人名辨識

人名是整篇文件的關鍵，已有許多研究(Chang *et al.*, 1994; Chen *et al.*, 1998; Miller *et al.*, 1999; Radev & McKeown, 1998; 李振昌, 1994; 李振昌 *et al.*, 1994; 黃燕萍, 1999; 楊昌樺 & 陳信希, 2004) 投身於人名辨識中。本研究提出一套辨識系統架構，在擷取人名的成效上相當不錯。

2.2. 向量空間模型

以 VSM Model(Salton & McGill, 1983)和相似度計算 Cosine 公式，執行文件分群分類處理。

2.3. 文件分群方法

採用「Single-pass clustering」(Salton & McGill, 1983) ，執行文件分群分類處理。

2.4. 新聞事件偵測與追踪

事件分群處理大致沿用先前研究(許登傑, 2005)事件偵測與追踪系統架構，但因應研究需要做修改。當完成新聞下載或達到設定之單次處理新聞量時，對新聞文件進行字詞處理與文件群聚。

2.4.1 斷句斷詞子系統

包含二個步驟：(1)執行中研院之 CKIP 斷詞和斷句系統 (2)中文詞字過濾器。之後將取回的已逐句逐詞標註詞性之標記文件，剖析器依標記進行判斷，取出斷詞、詞性及其未知詞類別。本研究保留重要名詞和動詞，做為事件主題地圖主題及關聯候選詞。在建置新聞本體論概念時，由於新聞事件特性，會將概念分為人、事、時、地、物，可藉由觀察 CKIP 詞性標註，配合辨識演算法，可取出上述斷詞。本研究採用先前研究(翁頌舜 & 許正欣, 2004; 許登傑, 2005)詞性合併法則，以降低字詞擷取後所產生語意不符問題。

2.4.2 字詞權重計算子系統

本研究使用 TFIDF 公式。考量本研究之斷詞詞性，如主題候選詞包含的專有名詞與一般斷詞，以及詞性類別為人、事、時、地、物的詞類、出現在新聞標題的字詞，會進行特別的字詞加權處理。

2.4.3 事件偵測子系統

將已經由字詞權重處理之新進新聞文件，與既有群集比對相似度後，觀察其是否顯示為新群集文件，其相似度若低於所設定之門檻值則表示此文件表示為新群集，反之則暫時表示為舊群集，再交由時間區間的時間衰退公式，計算出其分數。最後直接將通過相似度與時間區間門檻之文件交由事件追蹤子系統進行處理。新進文件與事件群集間相似度計算採用 Cosine 相似度計算。時間區間之處理主要計算新進文件與候選群集間之新事件信心度，若高於門檻值即視新進文件為新群集，反之則交由事件追蹤子系統，以安排該文件歸屬至適合的候選群集內。時間區間新事件信心度公式如下：

$$score(x) = 1 - \max_{c_i \in window} \{(1 - \frac{k}{m}) \times sim(\vec{x}, \vec{c_i})\} \quad \text{(公式 1)}$$

2.4.4 事件追踪子系統

在找出新進文件之對應候選群集後，此系統會分別計算新進文件與其之間的相關分數，完成計算所有相關分數之後，挑選最大分數之對應候選群集為此新進文件之歸屬群集。本研究採 two-way kNN 法，衡量新進文件所對應目標群集與其他群集間之相關分數。相關分數之計算公式，列示如下：

$$relevance\_score(\vec{x}, kp, kn, D) = \frac{1}{|U_{kp}|} \sum_{y \in U_{kp}} \cos(\vec{x}, \vec{y}) - \frac{1}{|V_{kn}|} \sum_{z \in V_{kn}} \cos(\vec{x}, \vec{z}) \quad \text{(公式 2)}$$

2.5. 語意網

伯納斯李(Berners-Lee & Fischetti, 1999)提出網路資訊架構「語意網(Semantic Web)」，主張將網路上文件有意義的結構化，建立資訊可充分分享與知識重複利用的網路。它能讓本體論除了表達特定領域詞彙定義與詞彙間關係，還包含詞彙和資訊間、以及資源和資源間關係等，能讓電腦交換、搜尋和認同文字意義，提供使用者以語意來搜尋資料而言。

2.6. 本體論(本體知識庫, Ontology)概論

本體論用以定義說明某一特定領域的知識或主題(陳雅絹, 2003)。其內容包含物件(Object)、物件特徵(Property)及物件間關係(Relation)；物件也被稱為概念(Concept)或類別(Class)，用來描述領域中概念；物件特徵則用來描述概念特性；關係則闡述概念間關係。目前語意網上有名的本體知識庫參考資料，如SchemaWeb(Lindesay, 2003)和Swoogle(Finin *et al.*, 2004)，是符合語義網文件格式(如RDF、OWL及DAML+OIL)的資料收集處，收集個人定義本體知識庫，提供使用者或語意網軟體開發者存取。由於本研究新聞資料為中文格式，故採用自訂新聞本體論，做為本新聞事件系統的領域知識。

2.7. 自動建構本體論(Ontology)

在(翁頌舜 & 許正欣, 2004)研究提到，本體論建構方法可分四大類型，分別是以字典為基、以文字分群為基、以關連式法則為基及以知識庫為基之建構法。本研究大致採用(龔俊杰, 2000)關連式法則方式自動建構新聞本體論，但做部分修改，以符合後續階段運作。

2.8. 新聞多文件摘要

新聞多文件摘要可讓節省讀者瀏覽時間、掌握事件重點。在先前研究(戴尚孝, 2003)已取得不錯的多文件摘要研究成果，其步驟包含：(1)斷句與斷詞 (2)群聚語句 (3)形成多文件摘要。但過程並沒有依據文章語意生成摘要，使得摘要內文代表性及語意不足。本研究透過新聞本體論，了解文章描述內容，增加摘要品質。在(吳家威 & 劉昭麟, 2002)研究中，利用 Ontology 幫助摘要系統分析文章語意，其 Precision 和 Recall 的成效較傳統摘要方法為好。

2.9. 主題地圖(Topic Map)

主題地圖可追溯於標準通用標記語言(Standard Generalized Markup Language, SGML)即開始發展，主要用以實現索引與辭典之建構過程，並由國際標準組織制定為 ISO/IEC 13250 標準。此模型包含有三大組合元素，分別為 Topic、Association、Occurrence，這些元素代表了主題、關聯與參照。在先前研究(許登傑, 2005)中，提出植基於主題地圖的新聞事件網頁檢索與瀏覽介面，將新聞事件的主題和關聯，以圖解方式呈現，使得讀者能快速地瀏覽新聞事件中的主題、主題間的關聯關係，以及此新聞主題內容出處(亦即類似書後索引的查找)，可快速呈現新聞事件的重要資訊。主題地圖以圖像與中心發散式加以呈現，將權重越高且越具代表性之關聯式置於主題地圖中央，逐一向外擴展，將事件中最重要的主題突顯出來。主題與關聯具有參照的性質，事件下的每個關聯式皆能透過參照索引功能 — 見(See)，對應出其所屬內文中不同句子位址。主題合併下的關聯式則同樣以參見(See also)功能，對應顯示出此關聯式所屬之不同事件位址。其架構如圖 1 所示：

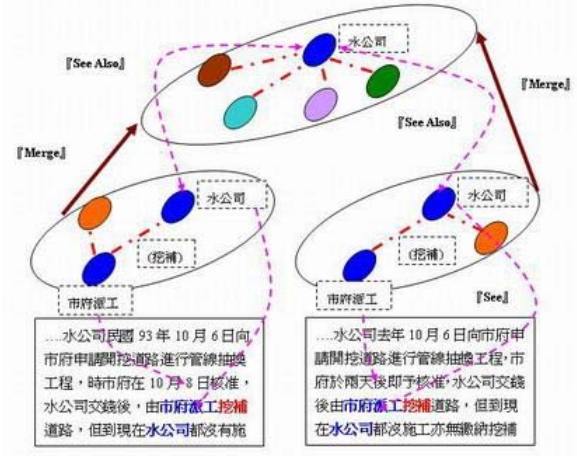


圖 1. 主題地圖中主題與關聯參照示意圖

如圖 1，橢圓形部分是單一新聞事件主題地圖，包含找到的主題，如「水公司」、「市府派工」，主題間關聯「挖補」則串起主題間關係；之後將不同新聞事件間相同主題合併，可查看相同主題下有那些不同的事件內容。經由主題地圖的視覺化呈現技術，吾人可快速掌握住新聞事件重點所在。

2.10. 主題地圖索引合併系統(TMs-merge System)

在「應用 Topic Maps 理論建置知識索引於線上新聞事件檢索研究」(許登傑, 2005)中，透過主題地圖之合併(Merge，亦即「索引的合併」)機制，擴大主題索引的涵蓋範圍，藉以關聯與接續可能相關之事件。而主題地圖索引的合併，是利用 Cosine 相似度公式結合 TDT 分群分類法，如圖 2 所示：

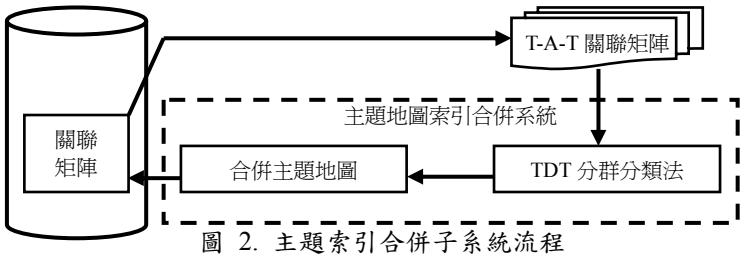


圖 2. 主題索引合併子系統流程

2.11. 中文句結構樹

本研究主題地圖的建構過程中，主要是以名詞篩選出主題候選詞，而以動詞做為主題間關聯。本研究採用先前研究(許登傑, 2005)發展出的中文詞性結構句法則，將重要的主題及關聯候選詞擷取出來，以做為主題地圖中主題和關聯的選取依據。

3. 研究架構

3.1. 系統架構

如圖 3 所示，先從 Yahoo!奇摩新聞網下載新聞文件。接著將原先散落於網路上之各新聞來源的新聞文件，透過相似度比對形成事件群集。新聞本體論自動建構系統產生出新聞本體論後，利用改良式多文件摘要以及主題地圖索引萃取系統，配合新聞事件群集，從事件之新聞文件中產生多文件摘要、萃取其主題地圖知識關聯索引，並將相關主題合併，建構主題地圖間合併關係；這部分是改進先前研究(許登傑, 2005)中無法考量到詞彙語意的缺憾。新聞事件合併處理系統，將語意內容相近事件合併，讓讀者查詢符合某特定主題的新聞事件；新聞事件關鍵字暨重要分類概念檢索詞擷取系統，將產生事件重要關鍵詞、相關新聞圖片連結，加上新聞類別重要人事時地物概念檢索功能，供讀者新聞檢索之用。檢索與瀏覽介面則加入文獻探討中提及的技術，希望提升讀者新聞檢索效能。

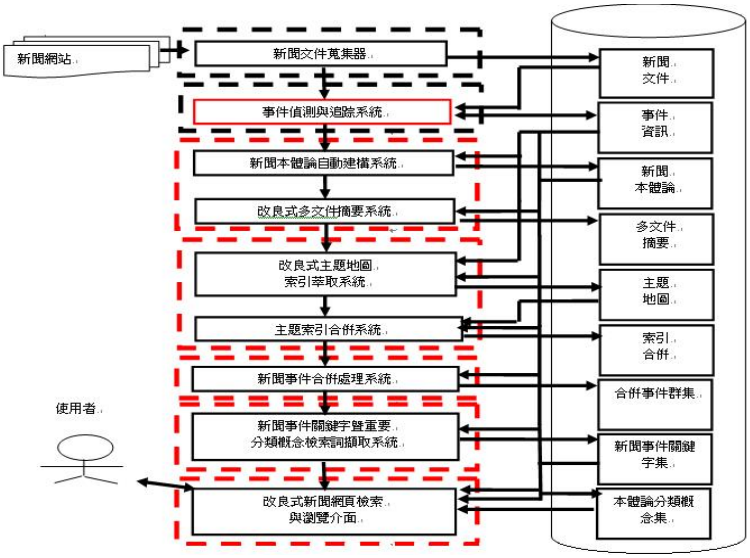


圖 3. 本研究系統架構

3.2. 事件偵測與追踪系統

在完成線上新聞下載、新聞文件的前置處理並儲存至資料庫後，接下來的事件偵測與追蹤系統沿用先前研究(許登傑, 2005)處理的流程，但做部分修改。此系統針對新聞文件進行字詞處理與文件群聚，流程如圖 4，最終產出事件群聚資訊。流程內相關步驟則分述如下。

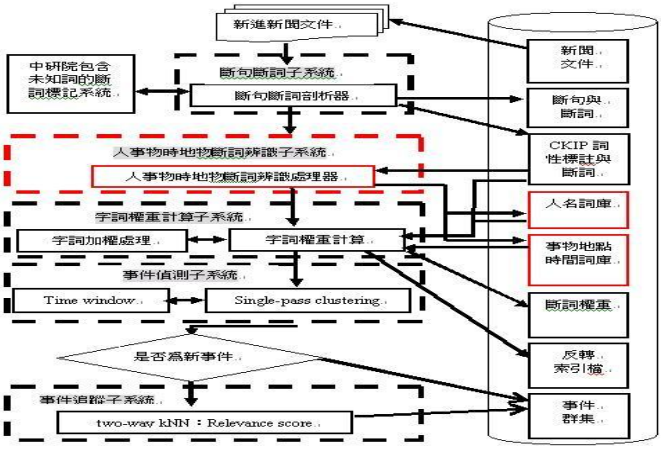


圖 4. 事件偵測與追踪系統流程

3.2.1 人名辨識子系統

執行完文獻探討提及的斷句斷詞子系統，擷取出新聞斷句及關鍵字後，執行此系統以找到新聞中重要人名。首先會將已先經由中研院 CKIP 斷詞處理後所斷出的詞彙做一判斷：和人名無關的詞類(如時間詞、地方詞..等)略過不考慮。接下來，由於人名在 CKIP 的詞性標註中，只會被標註為專有名詞，但並無法單獨判斷為是"人名"的詞性，所以再經由下列人名辨識四流程來加以判斷：

1. 挑出CKIP詞性為"Nb"且未知詞詞性為"CN、EN、MA和xx"的斷詞：根據觀察和實測，人名常出現在詞性標註為專有名詞(Nb)的斷詞裡，其中又以未知詞判斷詞為"CN"(中國人名)、"EN"(歐美譯名)、"MA"(由下而上合併詞)和"xx"(一般詞性)裡最常見。
2. 新聞文件斷詞和當篇新聞未知詞比較：由於CKIP中對於人名並沒有特別的判斷，只會將之歸類為專有名詞(Nb)(CKIP會將之當做未知詞)，故將每一篇新聞文件的前述斷詞和當篇新聞未知詞列表做比對，依據未知詞出現次數和詞性，做特別加權。
3. 運用百家姓詞庫比對：利用以下幾條規則，給予特別加權。當詞性:
   (1) 為"Nb"且為"CN"的斷詞，可能是"中國人名"；再利用百家姓詞庫比對，將含有姓氏的斷詞且長度(介於二至四字詞)給予特別加權，可斷出如"林曉培"。
   (2) 為"Nb"且為"EN"的斷詞，可能是"歐美譯名"；之後將斷詞長度(大於四字詞)的斷詞給予特別加權。可斷出如"凱薩琳麗塔瓊斯"等人名。
   (3) 為"Nb"且為"MA"的斷詞，可能是"日本人名"； 再利用百家姓詞庫比對，將含有姓氏的斷詞且長度(介於三至五字詞)給予特別加權，可斷出如"角川歷彥"。
   (4) 為"Nb"且為"xx"的斷詞，可能是"藝人人名"； 再利用百家姓詞庫比對(，將含有姓氏的斷詞且長度(介於二至四字詞)給予特別加權，可斷出如"侯孝賢"。
4. 挑出大於姓名門檻值的斷詞：整個流程如圖5所示。

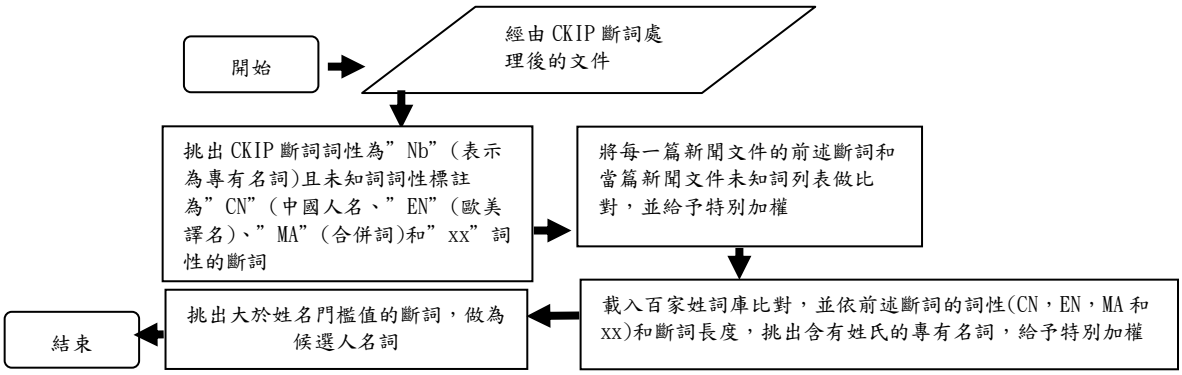


圖 5. 新聞文件人名擷取程序

系統在處理上有所限制，像原住民人名命名規則複雜、沒有一定規則，故不加以處理；外國人英文譯名，在各大新聞網站的新聞報導並沒有統一，也不在處理範圍。在此系統運作完成後，接下來依照文獻探討中的「字詞權重計算」、「事件偵測」、「事件追蹤」等步驟執行，最終產出事件群聚。

3.3. 新聞本體論自動建構系統

3.3.1 物件導向本體論(Object Oriented Ontology)架構

圖 6 為(Lee *et al.*, 2002)所定義的物件導向本體論架構，包含了領域、分類、概念和關係。領域是本體論所要描述的特定標的；分類是領域下的分類項目，會繼承領域的一些特性；概念類似物件，由物件名稱、屬性和操作所構成；關係則類似物件之間具有的三種關係：關聯、概化及聚合關係等。

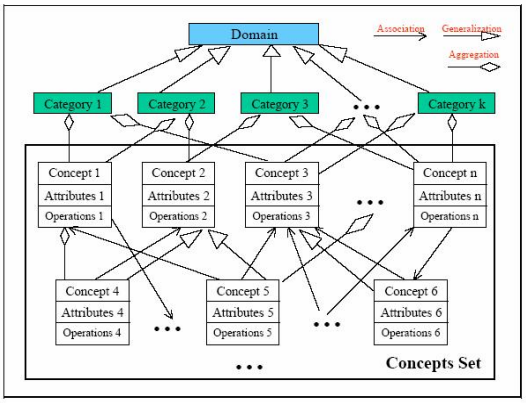


圖 6. 物件導向本體論架構

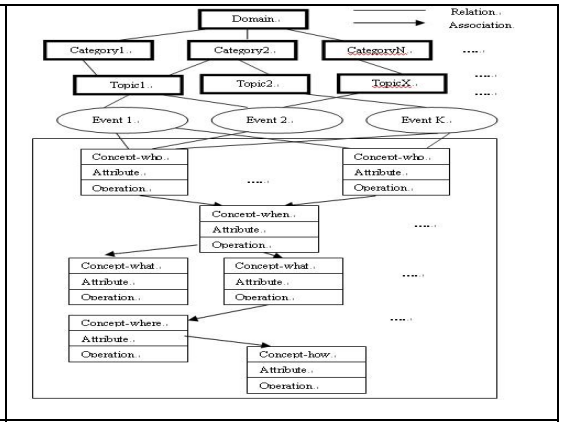


圖 7. 延伸新聞領域物件導向本體論架構

然而此種架構若要對映到新聞本體論的建置，會有所不足。由於新聞領域可用事件觀點來看待相關新聞報導，但圖6無法反應出事件概念；故本研究延伸先前研究(陳雅絹, 2003)提出的Domain Ontology結構，發展出新聞領域物件導向本體論架構。除了原先領域、類別及原先概念層級外，加上主題與事件概念。將經由斷句斷詞、字詞權重與關聯法則所選出之重要名詞建構成新聞本體論中的概念與關係，並把新聞文件人、事、時、地、物等元素納入考量，以反應出新聞文章中重要的主題。圖7為延伸新聞領域物件導向本體論架構。

如圖7所示，將新聞本體論中的概念區分成人、事、時、地、物等分類概念，不但可幫助新聞事件主題地圖主題的瀏覽檢視，更可實作出符合語意的新聞事件分類概念的檢索(如要查找人的話，可以找出和"張錫銘"此分類概念相關的各項主題資訊)。圖8是新聞本體論自動建構的過程。

如圖8所示，此新聞本體論建構暨主題索引萃取系統流程，除沿用先前研究(許登傑, 2005)主題索引萃取流程外，另外加上新聞本體論建構部分，如圖紅色區塊所示。此等改良不但可以產生出事件主題地圖中主題和關聯之外，可產生出新聞本體論，以供用於後續多文件摘要、主題地圖索引合併系統、語意相近新聞事件合併處理擷取所需資料之用。下一節解釋新聞本體論建構理念步驟。

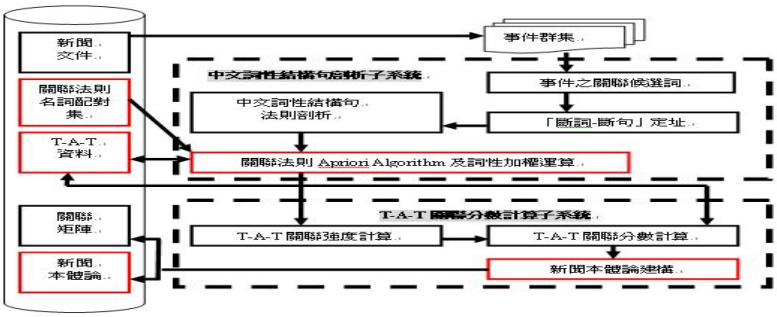


圖 8. 新聞本體論暨主題地圖索引萃取系統流程

3.3.2 新聞本體論建構理念與步驟

理念是使用自然語言處理，亦即中研院的詞性標註和資料探勘的關聯式法則，過濾出重要的名詞與動詞，當作是重要的事件主題地圖主題與關聯候選詞。新聞本體論建構步驟：

1.步驟 1：以事件群集為單位，利用 CKIP 先挑選出有意義的名詞和動詞。所擷取的重要名詞，包含了 CKIP 詞性辨識為普通名詞(Na)、專有名詞(Nb)、人名詞、地方詞(Nc)、位置詞(Ncd)及時間詞(Nd)。動詞則包含 VA、VAC、VB、VC、VCL 及 VD 等；將找出的重要名詞與動詞，依權重高低降冪排序，先挑出動詞以確認概念間關係。

2.步驟 2：使用資料探勘 Apriori Algorithm，挑出兩兩間具有強烈相關名詞，形成關聯法則。再以先前挑選出的動詞為中心，往外找二個名詞串成候選 T-A-T(應用中文句結構樹和詞性合併法則，可參閱文獻探討與本研究 3.5.1 章節的作法)，並配合找出的關聯名詞集合及名詞、動詞對應到的詞性，做適當 T-A-T 加權，再設立門檻值，去除掉主題間較無關聯的 T-A-T 集合，使其每個新聞事件的 T-A-T 集合更具代表性(可參閱本研究 3.5.2 章節作法)。

3.步驟 3：將上述的 T-A-T 做計算與篩選，(應用 T-A-T 關聯強度與分數公式，挑出合適的 T-A-T，可參閱本研究 3.5.3 章節的作法)，挑出要形成新聞事件本體論的 T-A-T 集合。

4.步驟 4：利用新聞類別新聞事件(不小於 2 篇以上新聞的新聞事件群集)的代表性 T-A-T 集合，形成初步新聞本體論(過程可參閱本研究 3.5.4 章節)，並調整本體論的概念。其調整是將主題間較沒有關聯的 T-A-T 集合，左邊 Topic 視為父概念，右邊 Topic 視為子概念，再將 Topic 間關聯，併入到父概念操作(operation)，視為父概念方法之一；把子概念併入到父概念屬性。經上述步驟可形成單一類別中單一特定事件本體論。然而建成的本體論，需經由新聞領域專家檢閱和修改，並調整產生本體論的演算法(如關聯式法則的門檻值設定、T-A-T 關聯強度與分數公式)，反覆測試本體論正確性，直到符合新聞領域專家的期望。

5.步驟 5：產生每一新聞類別的新聞事件本體論資料後，利用事件合併演算法(two-way KNN)，將語意相近新聞事件做事件合併處理，找出新聞事件本體論中主題和相關事件的關聯，以建成完整的新聞領域物件導向本體論。詳細事件合併處理過程，會於本研究 3.6 章節提到。

3.4. 改良式多文件摘要系統

若事件群集內包含兩篇以上新聞文件，即針對該事件產生一篇多文件摘要，協助使用者在極短的時間內判讀事件內容，並快速尋找其所感興趣之新聞事件。本研究之多文件摘要技術大體沿用先前(戴尚孝, 2003)研究，利用 Single-pass clustering 技術群聚語句，但加上了時間性考量、專有名詞加權、新穎性偵測、本體論加權及句子詞彙的關聯字詞加權等，如圖 9：

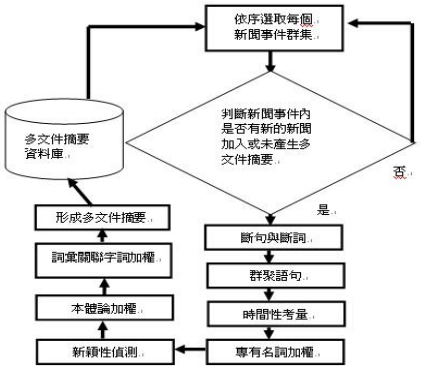


圖 9. 改良式多文件摘要流程

其步驟說明如下：

1. 斷句與斷詞：利用中研院 CKIP 斷詞、本研究自行開發的斷句處理及 CKIP 未知詞偵測，分析出新聞文件包含的關鍵詞彙及新聞語句。
2. 群聚語句：將各文件中描述同一事實的句子群聚起來，而群聚完成後輸出摘要時，是將各句子群集中分別取一句出來即可(不過其句子需經過第 2 步驟後的其它考量及運算)。

句子群聚的方法採用 Single-pass clustering，句子分類的方式則採用 two-way KNN。

3. 時間性考量:加上時間考量後的某事件某句子群集之中的句子，會加重在前面時間及後面時間的句子;而中間句子則依其天數,逐漸加強其時間所佔權重,會影響到句子的 TFIDF。
4. 專有名詞加權：依句子中的專有名詞出現次數，給予特別加權。當句子出現的專有名詞越多，則句子整體的 TFIDF 值也將越高，將提升該句子被選入摘要的機會。本研究使用一個句子中 CKIP 詞性辨識為 Nb 的斷詞做為計算依據。
5. 新穎性偵測：當要形成某事件的多文件摘要時，是將各句子群集中分別取一句出來即可，以避免輸出描述重複事實之語句。其使用的作法是：判斷句子中 name entities 的數量是否通過某個門檻值，也就是判定詞性標註(POS)為 Nb(專有名詞)、Na(普通名詞)、Vc(動詞)的數量是否通過門檻。倘若通過時，則取出此句當成摘要的一部分。
6. 本體論加權：先從單一新聞事件本體論中取出重要的本體論概念(包含人、事、時、地、物等分類概念)後，再判斷語句群集中的句子，其詞彙有出現在新聞事件本體論的任一重要概念上時，則特別予以加權，以突顯出句子裡重要的主題。
7. 句子詞彙的關聯字詞加權：透過事件本體論概念之間相互連結的資訊，以句子中的某特定字詞角度來看，若句子裡的其它詞彙，都和某特定字詞相關，代表它們彼此之間可能在探討某一個特定的事件或訊息(如「張錫銘」此特定字詞來看，「槍擊案」和「制式手槍」可能和它相關)，則依其相關字詞的個數，給予特別加權。
8. 形成多文件摘要：在完成句子群聚後，針對語句的輸出順序是依據該句子在原始文件的相對位址來決定，如公式(3)所示：

   P=句子在原始文件的位置／原始文件總句數　　(公式 3)

   計算所有輸出句子的 P 值後，P 值小的句子會先輸出，而若 P 值相同，則依文件編號順序，最後形成一篇多文件摘要。

3.5. 改良式主題地圖索引萃取系統

大體沿用先前研究(許登傑, 2005)做法，以事件群集為單位，產出代表該事件群集主題地圖索引。此系統運用中文詞性結構句法則與關聯強度計算公式，並依此建立完整索引萃取評估模型，以產出能顯示為主題地圖之關聯式資訊結構矩陣，此矩陣結構由主題 x、關聯 y、主題 z 相關資訊所組成，研究中以 T-A-T 代表關聯式。本研究改良先前研究不足，將新聞本體論建構與 T-A-T 關聯式萃取結合，並利用 Apriori Algorithm，將語意上較不相關 T-A-T 去除，可找到彼此間更具關聯主題。之後利用 T-A-T 關聯式與新聞本體論,即可建置以事件為單位的主題索引地圖建置。流程可參閱圖 8。

3.5.1 中文詞性結構句剖析子系統

利用完成詞性標注之句子與具備詞性之主題、關聯候選詞，透過中文詞性結構句法則的分析，判斷出哪些主題候選詞能經由關聯候選詞而結合成候選 T-A-T。先將該事件群集之所有元素候選詞依權重降冪排序後逐一定址，找出其位於原文件內文之所在短句。以關聯元素為中心，定址之短句上下各另取一句短句，使元素候選詞之定址對象成為三句短句的句組。定址完成後以關聯候選詞為中心，進行中文詞性結構句法則的剖析，找出所有以此候選詞為關聯之 T-A-T，再交由 T-A-T 關聯分數計算子系統加以評估。表 1 將舉例說明以接近優先原則之中文詞性結構句法則剖析後的產出。

表 1.中文詞性結構句法則剖析範例

<table>
<tr><td colspan="3">關聯候選詞「高達(VJ)」所定址之句組</td></tr>
<tr><td colspan="3">目前(Nd) 受難記(Na) 的(DE) 北美(Nc) 票房(Na) 總收入(Na)，</td></tr>
<tr><td colspan="3">高達(VJ) 3億5千480萬美元(DM)，</td></tr>
<tr><td colspan="3">在(P) 北美(Nc) 票房(Na) 排行榜(Na) 上(Ncd) 名列(VG) 第8位(DM)。</td></tr>
<tr><td colspan="3">中文詞性結構句剖析產出(T-A-T 候選關聯式)</td></tr>
<tr><td colspan="3">後面短句含有動詞(V)，判斷爲完整句子結構，不進行主題候選詞組合</td></tr>
<tr><td>總收入</td><td rowspan="4">高達(VJ)</td><td>3億5千480萬美元</td></tr>
<tr><td>票房總收入</td><td>北美</td></tr>
<tr><td>北美票房總收入</td><td>北美票房</td></tr>
<tr><td>受難記北美票房總收入</td><td>北美票房排行榜</td></tr>
</table>

資料來源：(許登傑, 2005) 應用主題地圖理論建置知識索引研究

3.5.2 關聯法則及詞性加權運算子系統

先前研究(許登傑, 2005)的 T-A-T 關聯式，主題間並沒有具備較高的關聯性；故本研究加入關聯法則和詞性加權的判斷機制，進行 T-A-T 關聯式的主題關聯程度計算，步驟如下：

1. 關聯法則的名詞配對判斷：利用資料探勘法則，針對每個新聞事件中已辨識出的人、事、時、地、物等斷詞，執行關聯法則運算，以找出具有高度關聯「人與人」、「人與事」、「人與時」、「人與地」、「人與物」…等依此類推的關聯法詞名詞配對集合。再針對先前的事件 T-A-T 關聯式，依關聯式中主題與關聯法則名詞配對間符合程度，給予不同程度加權。
2. T-A-T 關聯式的主題與關聯詞性加權判斷：經由本研究實驗觀察，發現由名詞詞性為「Na」、「Nb」、「Nc」及「Nd」所構成的主題，以及動詞為「VA」、「VB」、「VC」、「VD」所形成的主題關聯式，會讓讀者感覺到 T-A-T 關聯式的組成較能表達出新聞某一重要的片段資訊；符合以上詞性描述的 T-A-T 關聯式，依其符合程度不同，給予不同程度加權。
3. T-A-T 關聯式的 Support 值及 Confidence 值加權運算：倘若 T-A-T 關聯式中的任一主題符合關聯法則名詞配對集合，則針對原先關聯法則中名詞的 Support 值及 Confidence 值做特別的加權；以資料深勘法則來說，當關聯法則中名詞配對間的支持度(Support)與信賴度(Confidence)越高時，則此關聯式對於讀者而言，應該能得到更具有相關意義的主題閱讀機會；故本研究針對 T-A-T 關聯式中的主題，其原先在關聯法則中所對應到的支持度與信賴度，依其大小給予不同程度的加權，找出更符合事件主題 T-A-T 關聯式。

3.5.3 T-A-T 關聯分數計算子系統

沿用先前研究(許登傑, 2005)提出的 T-A-T 關聯強度公式與 T-A-T 關聯分數公式，以關聯分數評估 T-A-T 之代表性。先以關聯強度公式計算於同一事件群集中擷取之數條 T-A-T，若關聯式存在於標題中，則應適當加權其強度值。關聯強度公式列式如下，藉由類似詞頻的衡量方法，若能求出相對高於其他關聯式的數值，則表示此關聯式於該事件群集中出現之頻率比率越高，即關聯強度越強。

$$association\_strength(T_x, A_y, T_z) = Freq\frac{\{T_x, A_y, T_z\}}{T_x} \times Freq\{T_x, A_y, T_z\} \qquad \text{(公式 4)}$$

公式(4)中，Tx(Tz)為主題候選詞，Ay 為關聯候選詞 y，示為 Tx-Ay-Tz 關聯式，Freq()則表示計算該元素或關聯式之出現頻率。T-A-T 關聯分數之計算，係將該關聯式之關聯強度值乘上其主題與關聯候選詞之權重。公式中有將關聯候選詞納入權重加乘部分，主要考量動詞詞性仍有可能屬於文件之重要關鍵詞，例如關聯詞「謀殺」、「當選」。透過 T-A-T 關聯強度分數的計算，單一事件群集將產生代表其事件之 T-A-T 關聯矩陣，利用矩陣的關聯式資訊可直接呈現出此事件群集之主題地圖分布。T-A-T 關聯分數公式如下。公式(5)中，Tx(Tz)為主題候選詞 x(z)，Ay 為關聯候選詞 y，WTx(WAy、WTz)則為 Tx(Tz、Ay)之權重。

$$association\_score(T_x, A_y, T_z) = association\_strength(T_x, A_y, T_z) \times W_{Tx} \times W_{Ay} \times W_{Tz}$$

(公式 5)

3.5.4 T-A-T 新聞本體論建構

利用此 T-A-T 群集建構成事件本體論，步驟如下：

1. 新聞事件本體論的資料建構來源：一次產生某一特定類別(如政治)的事件本體論。將每個事件裡的代表性 T-A-T 群集取出後，透過本體論門檻值的設定，將低於門檻值的 T-A-T 從本體論概念的候選建構 T-A-T 群集中去除；這些去除掉的 T-A-T 群集將被合併為本體論概念中的屬性(attribute)或操作(operation)。
2. 本體論概念的特徵確認：事件本體論架構包含的人、事、時、地、物等概念特徵，可在之後的本體論加權處理中對於重要的人事時地物做特別加權之用、找出事件中重要的人事時地物等重要關鍵詞等，在此步驟會事先找到每個特定事件所辨識出的人事時地物詞彙，再和被選為本體論概念的 T-A-T 做比對，依此可找出本體論概念的特徵歸屬為何。

3. 計算新聞事件中的本體論概念出現次數與語意連結：一個新聞事件的 T-A-T 群集建成時，有可能會重複相同的主題(如張錫銘－涉入－搶案，搶案－主導－張錫銘)，在此會計算某特定本體論概念在事件中出現的次數、和它相關的其它本體論概念資訊，再透過 T-A-T 中的關聯，可找到兩兩本體論概念間的語意連結；之後依照本體論概念出現的次數，以及是否被歸類為人事時地物等其中之一的概念特徵，設立一個本體論建構門檻值，再將符合門檻值的本體論概念建構成一個初步的事件本體論。
4. 事件本體論的合併處理：由於一個新聞事件大多會針對特定的幾個主題，會有詳細且反覆的描述，倘若能將事件本體論的重要概念抽出，並透過事件間重要概念的比對及合併處理，將可找到彼此間相互有關聯的新聞事件報導，提供給讀者查閱相關新聞事件報導的方便性。本研究提出的事件合併處理機制，結合了已建構好的事件本體論和事件合併演算法(two-way KNN)，可解決先前研究(許登傑, 2005)所提到的事件群聚斷層的問題、事件偵測的時間區間設計問題(可參閱文獻探討)，將彼此間原本獨立的事件群集串連在一起，供讀者瀏覽與檢索之用。新聞事件的合併處理機制，會於論文 3.5 章節詳細說明。
5. 新聞本體論建構完畢：經前四步驟處理，可建構如圖 7 的本體論架構。

3.5.5 主題地圖索引合併系統(TMs-merge System)

在經由改良式主題地圖索引萃取系統的運作後，此時已將個別新聞事件重要的主題及關聯擷取出來，讀者可任意瀏覽它們的內容，快速了解新聞事件的重點所在；不過在不同的新聞事件中，也許會有相同的主題在討論著，在此沿用先前研究(許登傑, 2005)的主題地圖索引合併架構，將原先分散於不同事件的相同主題合併，便於讀者閱讀不同的新聞事件內容。

3.6. 新聞事件合併處理系統

合併處理對象是已經建好的初步新聞事件本體論(參閱 3.3.2 章節)，需要將一個個的新聞事件本體論做分群及分類處理，以呈現出主題和相關事件群集間的關係；所用到的分群技術，採用「Single-pass clustering」，而分類技術則採用修改過後的 two-way KNN 分類法，以衡量候選的新聞事件本體論該歸屬到的新聞事件主題目標群集。詳細過程如下所示：

1. 新聞事件本體論間的相似度矩陣計算：在執行新聞事件本體論的分群處理前，需先計算出新聞事件本體論間的相似度矩陣，以便做分群比較相似度之用。本研究採用 Consine 相似度計算公式，但做了部分修改，如公式(6)所示：

$$Sim(oi,oj) = \Sigma^{M}_{g=1} W_{goi} * W_{goj} / \sqrt{(\Sigma^{M}_{j=1} W_{goi}^{2})} * (\Sigma^{M}_{j=1} W_{goj}^{2}) \qquad \text{(公式 6)}$$

公式(6)中，sim(oi,oj)代表新聞事件本體論 oi 比對新聞事件本體論 oj 的相似度，W_{goi} 為本體論概念 g 在本體論 oi 中的權重(以本體論概念在事件中出現的次數，取代原始 Consine 公式的字詞權重)，W_{goj} 則為本體論概念 g 在本體論 oj 中的權重，M 則代表新聞本體論中本體論概念總數。利用此公式，可計算出新聞事件本體論間的相似度矩陣。

2. 新聞事件本體論分群處理：類似 two-way KNN 演算法執行過程，只是對象換成是某特定新聞類別的事件本體論集合，故不再贅述其流程。
3. 合併後新聞事件本體論的主題命名：從中觀察合併後的事件本體論所描述的內容，實際觀察事件本體論鬥重要且重複的概念，以人工方式為合併事件本體論命名。

3.7. 新聞事件合併處理系統

利用已事先建好的新聞事件本體論，從中取得新聞事件的重要概念，再經過一連串運作，形成新聞事件關鍵字與分類概念檢索資料集，流程如下所述：

1. 取得個別新聞事件本體論的關鍵概念：新聞本體論於建構之初，會計算某一概念在整個新聞事件中出現的次數，故本研究將概念出現次數當作擷取關鍵概念的重要考量因素；將概念出現次數設定門檻值，並取出限制個數的新聞本體論關鍵概念。
2. 產生關鍵概念的 BBS 超連結及相關照片資訊連結：透過 Google、Yahoo！奇摩新聞網、網摯資訊等資料來源，利用 Crawler 抓取網路上關於新聞關鍵字相關討論及照片網址。

3. 依關鍵概念屬性形成本體論分類概念群集：將步驟 1 取得的所有新聞事件本體論重要概念，依其特徵(人、事、時、地及物)，做個概念上的分類群集；設立分類群集門檻值，以擷取出分類概念群集(人、事、時、地及物)中可被擷取為本體論重要分類概念的資料。

3.8. 改良式新聞網頁檢索與瀏覽介面

如圖 10 所示。除了新聞事件主題地圖，更涵蓋社會大眾對於某一特定新聞事件的相關討論、新聞圖片連結、相關新聞、新聞多文件摘要等；本系統另外提供一般新聞關鍵字檢索，以及將新聞事件的重要人、事、時、地、物詞彙擷取出、供讀者檢索之用的新聞本體論分類概念檢索功能等，希望能提供讀者一個內容豐富且具有創新性的新聞事件瀏覽平台。

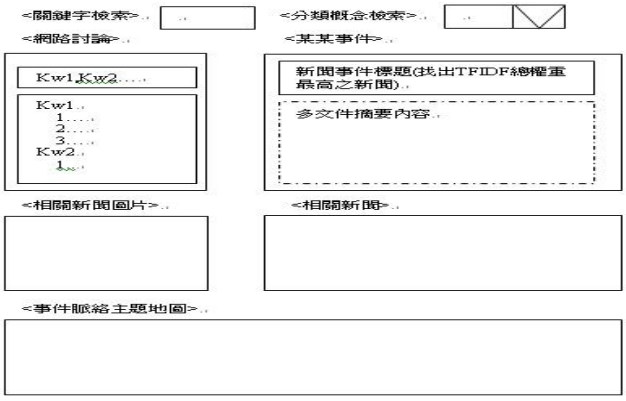


圖 10. 本研究新聞網頁檢索與瀏覽介面架構

4. 系統實作與評估

4.1. 系統開發

本研究之系統介面與功能分為後端與前端兩大部分，第一部分是後端資料處理系統，包含了文件字詞剖析、權重計算、分群分類、主題地圖萃取與合併、擷取新聞候選人名詞、擷取新聞文件事時地物詞、事件合併處理、新聞事件重要關鍵字擷取與自動建置 Ontology 功能等。第二部分為前端網頁檢索與瀏覽介面，包含了分類新聞事件內容瀏覽、關鍵字和分類概念檢索、網友對於該新聞事件的評論、相關新聞圖片、新聞多文件摘要、事件脈絡主題地圖及事件相關新聞文件，如圖 11 所示：

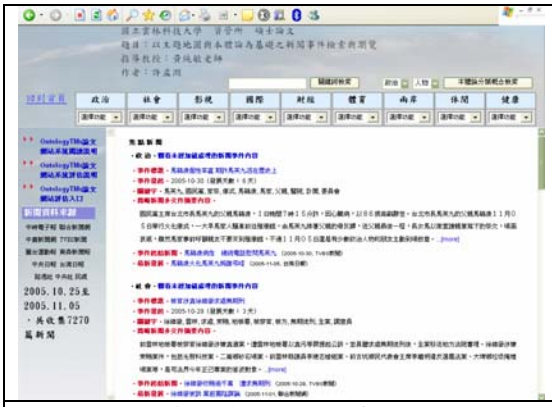


圖 11. 前端系統第一層新聞事件畫面展示圖

圖 12. 前端系統第二層新聞事件畫面展示圖—以政治類別的新聞事件分類瀏覽為例

使用者於畫面上方區塊，可點選不同新聞類別，其中包含三種新聞事件內容呈現功能：「新聞事件分類瀏覽」、「新聞事件主題合併地圖」及「新聞事件合併瀏覽區」;右下方區塊則是資料內容的常駐瀏覽區。網站的三大功能，分別是事件分類新聞瀏覽、主題地圖新聞瀏覽及關鍵詞檢索區(分為關鍵字檢索與分類概念檢索)。如圖 12，讀者可觀看到新聞事件標題、事件發展天數、新聞事件關鍵字、簡略新聞多文件摘要，以及事件起始新聞和最新發展等，讀者可點選其中超連結而進入第三層新聞事件瀏覽畫面。在前端系統第三層顯示畫面左邊地方，除了顯示出該新聞事件重要的詞彙、相關新聞圖片，最特別的地方是本研究加入了網友討論的意見，可見到新聞事件中各方的評論和探討，讓整個事件的呈現更加多元化。在畫面右邊部分，則顯示出新聞事件代表的事件標題、多文件摘要、相關新聞和事件脈絡主題地圖。系統畫面如圖 13，圖 14 則為主題地圖視覺化呈現工具。

圖 13. 前端系統第三層事件內容瀏覽畫面展示圖—以新聞類別「政治」、新聞事件「馬鶴凌個性率直 期許馬英九活在歷史上」為例

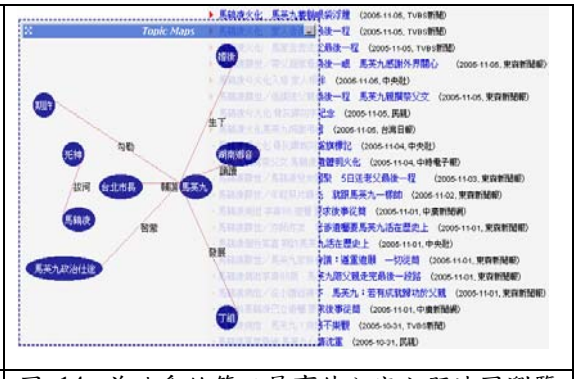


圖 14. 前端系統第三層事件內容主題地圖瀏覽畫面展示圖—以新聞類別「政治」、新聞事件「馬鶴凌個性率直 期許馬英九活在歷史上」為例

4.2. 系統評估

為驗證本研究發展之系統模型確實能改進先前研究(許登傑, 2005)實作出的研究結果，採取讓受測者透過前端瀏覽介面，以線上問卷方式進行評估。評估目標是針對本研究系統運用新聞本體論輔助所產生之新聞事件關鍵字、新聞多文件摘要、主題地圖結果、新聞事件人事時地物字詞辨識程度、新聞事件合併處理、本體論分類概念檢索等項目，期望能取得較先前研究更加進步的成果。

4.2.1 評估資料回收與分析

本研究系統評估採取網站公開發佈的方式進行測試，評估時期共計五天，回收了 72 份問卷。以下則針對回收的問卷內容進行統計分析，描述評估項目的評估結果。

1. 受測者背景資料分析：男女比例大約是一比一，學歷多集中於 21 至 30 歲年齡層學生，佔受測者 95%；大學生與碩士生比例為 13:7；受測者普遍有電子新聞閱讀習慣。
2. 加值處理新聞內容效益分析：本部份包含四個問項，個別是：「您認為新聞事件內容經加值處理後，是否比原事件內容的呈現，更易於理解事件中所出現的關鍵字、事件多文件摘要內容、主題元素所提示的字詞、關聯元素所表達的主題元素間關聯」之中，非常相關的比例與相關的比例相加，約有七成以上受測者感到滿意！
3. 新聞所含人事時地物詞彙擷取效益評估：本部份包含四個問項，個別是：「您認為新聞事件內容的呈現：(1)人名 (2)事物、地點、時間詞 (3)本體論分類概念檢索找出的人名 (4) 本體論分類概念檢索找出的事物、地點、時間詞，其辨識程度」之中，非常相關與相關的比例相加，約有七成以上的受測者感到滿意；而在事物、地點及時間詞的辨識，則約六成受測者感到滿意。
4. 本體論分類概念檢索效益評估：本部份問項是：「您認為新聞事件合併功能所查詢出來的合併事件標題，與相關的新聞事件編號所顯示的事件內容，彼此間的相關程度為何(亦即標題與事件間的相關程度)」，評估結果顯示，經由事件本體論所萃取出的重要概念(如人物、事物、地點及時間)，與相關的新聞事件所對應的事件內容，有七成左右的受測者認為是滿意的。
5. 新聞事件合併效益評估：本部份問項是：「您認為新聞事件合併功能所查詢出來的合併事件標題，與相關的新聞事件編號，所顯示的事件內容、以及新聞事件間，彼此間的相關程度為何」從評估結果可觀察到，應用事件本體論所產生的新聞事件合併處理，在標題與事件間的相關程度，以及被合併的事件間彼此相關程度，約七成左右受測者給予滿意的高度評價。
6. 新聞事件查詢介面觀感評估：大約八成左右受測者對新聞事件檢索與瀏覽機制，給予好評！

4.2.2 與國內各大新聞網站的比較

本研究的新聞事件系統，採用不少以往新聞網站所沒有的技術與概念，讓讀者在閱讀新聞事件上具有更大的便利性及更完整的新聞內容。茲整理如表 2 所示：

表 2. 本新聞事件系統與國內各大新聞網站採用技術之比較

採用技術 / 新聞網站	視覺化呈現(主題地圖)	Ontology 輔助新聞內容	以事件方式瀏覽新聞	新聞多文件摘要	新聞專輯	網友的事件相關討論	個人化自訂新聞瀏覽	RSS 訂閱	新聞內容關鍵字連結	新聞檢索
本新聞系統	●	●	●	●	●	●			●	●
Yam 蕃薯藤					●	●		●		●
Yahoo！奇摩					●	●	●	●	●	●
自由電子報					●	●				●
中時電子報					●	●	●	●		●
udn.com					●	●		●		●
Google 新聞						●	●	●		●

5. 結論與未來研究方向

5.1. 研究成果

本研究延續先前研究成果(黃純敏 *et al.*, 2004; 黃純敏 *et al.*, 2003)，針對主題地圖中主題關聯式間相關程度不高、多文件摘要的正確性及可讀性、新聞關鍵字與事件的相關程度、資訊檢索較不符合語意等，提出有效改善方法。在本體論的幫助下，解決了導致上述問題產生的原因：「受限於並非是基於新聞文件語意為主的處理方式」。鑑於人名、事物、地點及時間詞，在新聞事件(文件)扮演重要角色，本研究開發一套人事物時地物斷詞辨識處理系統，能有效將新聞文件中潛藏人名、事物、地點及時間擷取出來，做為重要新聞主題；本研究亦改進先前研究(陳雅絹, 2003)中新聞 Domain Ontology 結構，發展出改良新聞本體論架構；藉本體論幫助可有效改善先前研究(許登傑, 2005)成果，另外開發出事件合併處理機制與重要關鍵詞擷取系統，增加事件內容的呈現豐富性。

本研究在資訊檢索部分，鑑於以往讓使用者直接鍵入關鍵字搜尋相關新聞事件的方式，似乎無法讓讀者了解到事件發展全貌，因此加入網友對於該事件的討論、相關新聞圖片以及本體論分類檢索的概念，讀者可藉此看到新聞事件不同討論觀點，有更多元化看法；特別是在本體論分類檢索部分，作者將新聞事件中重要的人、事、時、地、物等概念擷取成新聞事件關鍵詞，讀者可依據感興趣的內容做檢索，如對"張錫銘"這個人感到興趣，點選它之後即可看到和它相關的所有事件內容，以及主題地圖的呈現。此一機制能提升讀者在閱讀新聞事件上的便利，快速掌握事件發展脈絡！

本研究觀察到，當新聞事件是圍繞在某一特定人物的相關報導時，如名模林志玲、政治人物馬英九，在事件群聚效果、新聞多文件摘要、Topic Maps 主題和關聯間語意、事件合併的成效等，會呈現較佳的結果。作者推想，由於報導某一特定人物的相關新聞很單純，會圍繞著人物報導。針對此類型新聞事件，本研究的事件相關處理機制都有很良好的效果！反觀如財經、生活和健康等新聞類別的事件內容呈現，則和上述結論相反，可能會有較差的事件處理結果；作者推想，應該是這些類別的「新聞事件」本身持續報導的機會不大，造成事件呈現的結果不佳。

5.2. 未來研究方向

本研究尚存在許多可再精益求精之處；以人名辨識而言，如原住民名字"瓦歷斯．貝林"，受限於原住民名字取法沒有固定規則，成為辨識上限制；再者，由於本研究架構龐大，在本體論開發方法部分，是沿用較舊的關聯法則法，而非熱門的 Formal Concept Analysis(FCA)，後續研究可考慮採用，也許能得到更佳的建構結果。本研究另一限制，在於沒有實作語意推理機制；為增添本體論實用程度，強烈建議未來研究可朝此方向發展。現今本體論發展遇到不少問題。如新聞領域本體論的建構方式與理念，並沒有可依循規範；若能發展出一套可遵循的新聞本體論架構，應能提升新聞本體論間互相交流與應用程度。此外本研究系統，需經長時間後置處理，使得新聞本體論無法做到即時更新。以上問題描述，除了會產生新聞事件瀏覽無法做到一般新聞網站的動態更新外，會引發一個有趣議題：新聞事件發展有其時間性，某些重要概念會隨著事件發展，逐漸降低其重要性；例如以影視圈話題女王許純美舉例，男友從原先的林宗一到後續的邱品叡，為了反應出人名「許純美」此事件本體論的「現今」重要概念，應該將人名「邱品叡」特別加強其權重，而人名「林宗一」重要性應隨著降低，應能得到更符合「當時情況」的新聞本體論。另外為因應讀者閱讀新聞需求，應加入英文新聞文件剖析與處理機制。有些受測者反應本研究的新聞事件資訊量雖然豐富，但對於某些讀者，也許是另類的「資訊過載」，造成閱讀上負擔；可以考慮專注在新聞呈現的質與量如何達成某個平衡點。作者認為，後續研究可從上述方向進行改善。

6. 參考文獻

1. Berners-Lee, T., & Fischetti, M. (1999). *Weaving the web：the original design and ultimate destiny of the world wide web by its inventor* (1st edition ed.). San Francisco: Harper Business.
2. Chang, J. S., Chen, S. D., Ker, S. J., Chen, Y., & Liu, J. S. (1994, June 1994). *A multiple-corpus approach to recognition of proper names in chinese texts.* Paper presented at the Computer Processing of Chinese and Oriental Languages.
3. Chen, H. H., Ding, Y. W., & Tsai, S. C. (1998). Named entity extraction for information retrieval. *Computer Processing of Oriental Languages, 24*, 75-85.
4. Finin, T., Ding, L., Dornbush, S., Doshi, V. C., Java, A., Kolari, P., et al. (2004). Swoogle semantic web search engine. 3.1. Retrieved 08/08, 2006, from http://swoogle.umbc.edu/
5. Lee, C. S., Liao, J. X., & Kuo, Y. H. (2002). *A semantic-based concept clustering mechanism for chinese news ontology construction.* Paper presented at the International Computer Symposium, Taiwan.
6. Lindesay. (2003), V. Schemaweb - rdf schemas directory. Retrieved 08/08, 2006, from http://www.schemaweb.info/default.aspx
7. Miller, D., Schwartz, R., Weischedel, R., & Stone, R. (1999). *Named entity extraction for broadcast news.* Paper presented at the Proceedings of DARPA Broadcast News Workshop.
8. Radev, D. R., & McKeown, K. R. (1998). *Generating natural language summaries from multiple on-line source.* Paper presented at the Computational Linguistics.
9. Salton, G., & McGill, M. J. (1983). *Introduction o modern information retrieval*. New York: McGrw-Hill Co.
10. 吳家威, & 劉昭麟. (2002). 應用本體論設計與建置摘要系統, *民生電子研討會論文集*.
11. 李振昌, 李御璽, & 陳信希. (1994). *中文文本人名辨識問題之研究.* Paper presented at the Proceedings of ROCLING VII.
12. 翁頌舜, & 許正欣. (2004). 於語意網上自動化建構本體論之研究. Paper presented at the 2004 臺灣商管與資訊研討會論文集(光碟片).
13. 許登傑. (2005). 應用主題地圖理論建置知識索引研究. Paper presented at the 2005「開放原始碼」技術與應用研討會, 成功大學數位生活科技研究中心.
14. 陳雅絹. (2003). 基於 ontology 之模糊代理人於中文新聞文件摘要技術之研究. In 國科會.
15. 黃純敏, 郭家良, & 楊顓溥. (2004). *新聞知識管理系統之建構與評估.* Paper presented at the 第十屆資訊管理暨實務研討會.
16. 黃純敏, 戴尚學, & 郭家良. (2003). 新聞事件自動偵測與追踪及多文件摘要系統研究, *中華民國九十二年全國計算機會議*: 教育部.
17. 楊昌樺, & 陳信希. (2004). *以語法分析為輔建立新聞名詞知識庫.* Paper presented at the The Association for Computational Linguistics and Chinese Language Processing.
18. 龔俊杰. (2000). *具物件導向式 ontology 自動建構能力之個人化 xml 資訊服務系統.* 國立成功大學.

以部落格文本進行情緒分類之研究

楊昌樺　陳信希

國立台灣大學資訊工程學系

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

摘要

本文從部落格文本中帶有情緒符號的文句出發，探討人們的溝通行為擴展至網路空間後，如何將情緒表達的需求反映在文字與情緒符號的使用上，並進一步以情緒符號的意涵作為文句表達情緒的分類依據。我們從雅虎奇摩部落格服務取得訓練與測試集，以向量支撐機(SVM)運用文句特徵設計情緒分類器，並藉由各項情緒分類器的實驗數據，研究是否可以利用情緒詞彙解釋人們在部落格中使用情緒符號的偏好與特徵，進而達成對網路空間人們情緒的解讀與分析。

1. 緒論

達爾文於 1872 年發表“The Expression of the Emotions in Man and Animals”一書，他以進化的觀點分析動物和人類在情緒質量上相似與相異之處。之後一百多年來，心理學、腦神經科學、認知科學等領域的學者也投入人類情緒分析的研究，並發展出透過腦部影像、心跳、血壓等生物訊號來判斷人類情緒變化的方法(Dolan, 2002)。情緒狀態的傳遞亦屬於人類溝通行為的一個重點成份，人類可以透過臉部表情、肢體動作、手勢、語言、聲調等訊息來表達自己的情緒狀態，溝通的對象也會根據自身的經驗、以及對表達者的理解來解讀對方的情緒。

情緒解讀的工作也吸引電腦科學領域學者的注意，隨著電腦辨識技術的演進，人類表達出的各項訊息轉換成數位資訊，研究人員採用機器學習等方式，計算這些資訊與各項情緒類別的關係。相關的研究包括：當判斷的類別定義成人類的情緒狀態後，如何挑選適當的資訊作為特徵，以訓練出相關模型來判斷人類的情緒。如 Chuang and Wu (2004)使用文

字及語音兩類特徵資訊對語音句子所隱含的情緒進行辨識，研究結果顯示同時使用兩類特徵，比單獨使用一類特徵對情緒辨識有更好的效果。Pao 等人 (2005)也使用語音特徵建立情緒分類器，開發的工具可協助聽障人士透過語音進行情緒的表達。

近代電腦科學界所興起的一項重大發明－網際網路，除了提供電子資訊分享的平台外，也提供人們不同於以往的溝通介面。追溯自網際網路開始用來傳遞資訊(如電子郵件、電子佈告欄、電子聊天)的年代，透過網際網路傳遞文字，可以在即時即地的基礎上，達成使用者互動溝通的便利性。然而當通訊的對象越來越習慣這項新的通訊方式後，也開始產生了需傳遞彼此情緒狀態的需求。通常在網際網路上傳遞訊息時，會缺乏夠多的前後文資訊來判斷對方的情緒狀態，為了避免因此造成溝通時的誤解，1981年CMU的Scott Fahlmany曾經設計「:-)」和「:-(」兩個可以放在文字中的符號，以象徵性的笑臉和哭臉來代表高興或不愉快的情緒。隨著類似象徵情緒或表情的符號不斷地被網際網路使用者創造與使用，甚至到了90年代unicode要擴充全球通用文字碼時，情緒符號也佔了一席之地，如unicode集合中三個字元「☹」、「☺」、「☻」的十六進位碼分別定為0x2639、0x263a、0x263b。

近年來，因為圖形使用介面的演進，讓更多表情文字得以圖示化的方式呈現，所產生的新興圖示集合通稱為Smiley或Emoticon，這些圖示被大量使用在以網際網路為主的通訊媒介中。而個人化訊息傳遞服務的流行更加速了這種圖示流行的趨勢，如雅虎即時通、微軟MSN Messenger等傳訊軟體，皆提供使用者可在傳訊的介面上使用許多不同情緒符號或是自訂圖示。此類圖形對應到一些表情文字或是自訂的字元順序，讓個人偏好的圖示與動畫得以即時內嵌在通訊的內文中間。Liu等人 (2003)曾以文字與生活常識建立模型來判斷人的情緒，並設計出一個電子郵件介面EmpathyBuddy，根據郵件中每一句話賦予一個Chernoff Faces，這些臉部表情也增進了使用者訊息互動的趣味性。

除了表情與情緒符號的使用外，網際網路使用者也會在既有語言範疇之外，創造新的文字表達規則。所創造的新詞彙或用語，形成網路方言的一部分。這些方言通常出現在以網際網路為主的電子媒介所使用的溝通訊息中，夾雜於正常的語言文句之內。這樣的呈現通常會讓原句子不合於文法、或是讓網路方言本身形成了未知詞彙。高與倪(2003)透過網

際網路文章收集情緒用詞，並將這些詞彙歸納在憤怒、悲傷等情緒分類下。

發佈網路文章的介面、表達方式、語言，隨著更多使用者的參與而不斷地擴充、創新，藉由更多使用者的加入所引發的新創意，也產生了更多新興的溝通方式。這個由許多網際網路使用者彼此互動、溝通所形成的虛擬空間，有個特殊稱號為Cyberspace。最近Cyberspace下，又興起了一個子空間叫Blogosphere。所謂Blogosphere，是指由一種個人導向、可與使用者互動的媒體發佈介面Blog(或稱部落格)，所形成的集合。我們於前文曾敘述過使用者在適應了Cyberspace的溝通方式後，會開始有情緒狀態傳遞的需求。而Blog在開始吸引眾多使用者之後，由於其創立思維更貼近人們表達個人狀態的動機，也吸引更多使用者前來表達自己的觀念、想法、甚至個人情緒。

一些Blog服務提供使用者不同層次的情緒標記選項，例如：Xuite[1]讓使用者選擇個人站台的心情分類，如「甜蜜中」、「憂鬱中」等共15種分類；LiveJournal[2]讓使用者以文章為單位，標記發表文章當時的心情，如sad(悲傷)、mischievous(淘氣)等共118種分類。Mishne(2005)使用由LiveJournal的文本所抽取的特徵，以SVM(Cortes and Vapnik, 2005)訓練以文章為單位的心情分類器。文中顯示實驗成果仍有很大的進步空間，也反映心情分類仍是個相當具有挑戰性的議題。本文從這個角度切入，進一步以文句為單位，觀察個人情緒在Blog文本上反映的現象，這是中文情緒分析的創新嘗試。

本文內容安排如下：第二節介紹帶有情緒表達的Blog文本，第三節敘述以文本資訊進行情緒分類的實驗設計、與實驗結果，第四節總結成果、並指出未來的研究方向。

2. 部落格文本與情緒符號

由於本文探討的個人情緒分析所採用的語料來自 Blog 文本，因此先介紹建立 Blog 文本語料庫相關的議題。過去隨著 WWW 的發展，網站登錄與各家搜尋引擎相繼而起，提供使用者快速便利的資訊搜尋服務。Blogosphere 的發展有點類似，近年來陸續有許多網

[1] http://blog.xuite.net/

[2] http://www.livejournal.com/

站，甚至大廠開始提供 Blog 架設服務，吸引許多 Blog 登錄(Blog Ping)與搜尋服務的開設，如 Technorati[3]提供搜尋網頁介面及 API，Wegblogs.com[4]提供登錄 Blog 的介面，並即時回報世界上最新更新的 Blog 文章。Kolari 等人(2006)的研究即利用 Technorati 及 Wegblogs.com 提供的服務，取回英文文章加以分析。提到最受歡迎的 Blog 架設服務，分別是 Blogspot[5]與 MSN Spaces[6]。其中 Blogspot 文章發佈的介面僅提供基本的文字編輯，而 MSN Spaces 藉由著名的傳訊軟體 MSN Messenger 的附加功能，開始吸引使用者的投入。其發表文章的介面因而引入了 MSN Messenger 中慣用的情緒符號，這些符號可以讓發表 Blog 的使用者加在文章的任何部分。

台灣地區最大入口網站為雅虎奇摩，所提供的 Blog 服務[7]也可以讓使用者在發表文章的介面中，於文章任意部分加入由雅虎即時通軟體引入的情緒符號。雅虎 Blog 服務所提供的情緒符號共 40 種，其圖樣、代碼、意涵如表 1 所示，包括世界通用的笑臉「:)」和哭臉「:(」符號，也有自創的圖示如「@};-」代表一朵橫放的玫瑰花。由於使用者可自行決定要不要在發表的文章中加入情緒符號，因此並不是每篇文章都會含有情緒符號。

為了運用含情緒詞彙的文章來協助本研究的進行，我們首先建立一個前提：

(p1) 使用者情緒的傳遞會反映在這些情緒符號的使用上。

並提出一個假設：

(a1) 使用者伴隨著文本的情緒就是該情緒符號的意涵。

例如下列句子中：

(s1) 一次又一次的這樣~心中莫名的火大..每次他一回來~我都睡不飽 

(s2) 我們也...傻眼了 

(s3) 每次和妳講電話就是莫名的開心 

---

[3] http://www.technorati.com/

[4] http://weblogs.com/

[5] http://www.blogspot.com/

[6] http://spaces.msn.com/

[7] http://tw.blog.yahoo.com/

參照表 1 我們定義(s1)句的使用者情緒為「大哭」、(s2)句的使用者情緒為「驚訝」、(s3)句的使用者情緒為「大笑」。經由這樣的設定，本研究即以雅虎 Blog 服務使用者所發表的文章為材料，分析與 40 種情緒符號(情緒)之間的關係。

2.1 實驗語料庫

為了提供本研究所需的實驗語料，我們透過雅虎 Blog 服務的搜尋引擎，取回 2006 年 6 月底至 7 月初共 47,570 篇 Blog 文章，相關統計資訊如表 2 所示。其中 2006 年 6 月份的文章共 9,668 篇為測試資料集、7 月份第一周文章共 29,528 篇為訓練資料集，兩項資料供實驗一至實驗四使用。7 月份第二周的文章共 8,374 篇為補充訓練資料集，供實驗三至實驗四使用。訓練資料集帶有情緒符號的文章數量為 4,146 篇、測試資料有 1,289 篇、補充訓練資料集有 1,241 篇，比例在 13%~15%之間。如果以去掉 HTML 標記後的 UTF-8 編碼

表 1、雅虎奇摩部落格服務所提供的情緒符號列表

符號	文字	意涵	符號	文字	意涵	符號	文字	意涵	符號	文字	意涵
	:)	微笑		:O	驚訝		0:)	天使		(:|	呵欠
	:(	難過		X-(	生氣		:-B	戴眼鏡		=P~	流口水
	;)	眨眼		:>	得意		=;	再見		:-?	考慮
	:D	開懷		B-)	耍帥		I-)	睡著			偷笑
	;;)	眨眨眼		:-S	擔心		8-)	環顧		=D>	鼓掌
	:-/	疑惑		>:)	邪惡		:-&	不舒服		[-o<	祈禱
	:x	愛意		:((	大哭		:-$	安靜		:-<	嘆氣
	:”>	害羞		:))	大笑		[-(	不說了		>:P	呸
	:p	吐舌頭		:|	呆住		:o)	小丑		@};-	花
	:*	親親		/:)	皺眉		@-)	神智不清		:@)	豬頭

表 2、訓練與測試資料集統計

	文章數	帶有情緒符號文章數	比例	無情緒符號文章平均長度	有情緒符號文章平均長度
訓練資料集	29,528	4,146	14.04%	1,934	1,131
補充訓練集	8,374	1,241	14.82%	1,776	1,096
測試資料集	9,668	1,289	13.33%	1,614	1,058
合計	47,570	5,435	13.87%		

圖 1、Blog 文本[8]、發文介面範例、結構示意圖

長度衡量文章的相對長度，可發現通常沒有使用情緒符號達的文章平均長度較長，且與有使用情緒符號文章的長度有一定程度的差距。這個現象顯示情緒符號在文本中扮演重要角色，有些意涵隱含在符號中。

2.2 部落格使用者心情之反應與文本風格

前文提及以歐美地區使用者為主的 Blog 服務 LiveJournal，提供使用者針對發表文章當時的心情給予一個標記，並搭配一個情緒符號加以表達。本研究的語料來源雅虎 Blog 服務，則提供使用者以平時在網路表達的習慣在文句中加入自己的情緒。圖 1 為本研究所取回的 Blog 文本範例，文本一個頁面中包含有標題、分類、發佈時間、及內文等資訊。使用者在發佈內文時，可自由加入相片、文字、情緒符號，發文的介面也提供變更文字格式、加入連結、變更排版樣式等功能。在這個例子中，發文作者曾使用過「微笑」、「驚訝」、「大笑」、「花」等四種情緒符號，相當程度地表達了其開心的情緒。

如同一般作文時，有諸如記事、抒情、論說、公文等體裁。使用者利用 Blog 發佈文

[8] 文本來自 http://tw.myblog.yahoo.com/jw!IlyqF8KeFRLqmezkFIeTaK5./article?mid=9470

章時，對於內文的呈現，也有許多不同的表達方式，例如：

- 日記、抒發心情、詩賦、創作、評論，可能會引用一些相片。
- 純遊記、遊記加上許多相片、或是純粹只有相片分享。
- 資訊、教學、轉貼，通常有說明圖片和連結，也可能只給連結沒內文。
- 獨白、對話、留言、溝通，可能會擺一幅近照。

不同使用者都會有自己偏好的表達方式。另外，Blog 網站通常會提供預設、或是可自行設定的分類，讓使用者標記所發表的文章。

以上一節介紹的訓練資料集為例，蒐集到的 29,528 篇文章中，使用者自行標記的分類達 10,306 種之多，其使用前 20 名的分類名稱如表 3 所示。其中「未分類資料夾」六字為雅虎 Blog 系統預設的分類，此類文章共有 32.21%，表示約三成使用者沒刻意將文章賦予分類資訊。其他分類除了「心情日記」這四字獲得最多人使用，使用率比例大於 1%外，其餘分類的使用比例皆小於 1%。這些分類字面使用最多的詞彙是「心情」，反映出使用者願意透過 Blog 達到抒發心情的目的。表 4 將所有使用到情緒符號的文章，依照其所屬分類出現次數順序列出前 30 名，可發現使用者會透過部落格，在抒情、日記等這樣的體裁下抒發自己的情感，並使用適當的情緒符號描述自己的心情。

表 3、雅虎 Blog 分類使用比例前 20 名列表

分類	篇數	比例	分類	篇數	比例	分類	篇數	比例
未分類資料夾	9190	32.21%	心情點滴	105	0.37%	心情記事	61	0.21%
心情日記	348	1.22%	文章分享	104	0.36%	心情故事	55	0.19%
歌詞	178	0.62%	笑話	95	0.33%	星座透視	55	0.19%
心情	162	0.57%	生活點滴	84	0.29%	小說	49	0.17%
日記	152	0.53%	文章	74	0.26%	生活	48	0.17%
心理測驗	138	0.48%	蔡依林	73	0.26%	心情寫真	48	0.17%
星座	124	0.43%	心情札記	70	0.25%			

表 4、雅虎 Blog 使用情緒符號文章所屬分類前 30 名列表

分類	篇數	分類	篇數	分類	篇數	分類	篇數
未分類資料夾	1387	心情雜記	16	文章分享	11	自言自語	7
心情日記	84	生活	16	我的日記	11	心情手札	7
日記	35	心情札記	14	心情寫真	9	結婚篇	6
心情	34	心情記事	12	心情筆記	8	笑話	6
生活點滴	26	雜記	12	心理測驗	8	生活 543	6
心情點滴	26	生活雜記	12	我的心情日記	8	分享	6
生活日記	22	生活記事	12			心情日誌	6
心情故事	21	心情分享	11	生活札記	8		

3. 情緒分類實驗

針對從雅虎 Blog 服務所蒐集的文本特性，我們將情緒分析的問題轉換成針對文句特徵來判斷情緒類別的問題。本文採用 SVM 做為建立情緒分類器的核心，SVM 的工具套件為 Fan 等人(2005)所提出的 Libsvm。本節實驗中所提到的**基準值**，即是 Libsvm 在不下任何參數的情況下所做出的分類器，對測試資料集所能作出的分類效能，這種情況下分類器通常會把類別歸給實例最多的一類。參照 Libsvm，本節實驗統一設定的訓練參數為 c=10、g=1.6 訓練出各分類器，之後各小節所回報**內部測試分類正確率**、**外部測試分類正確率**是該分類器分別應用在訓練、測試資料集後所得到的分類效能，而**內測提昇率**和**外測提昇率**則是內部測試分類正確率和外部測試分類正確率分別減去基準值。

SVM 透過各個實例所帶有的特徵向量進行訓練集分類的動作，我們參考第一節所提高與倪(2002)關於網路情緒字眼的研究，使用其蒐集的分類，包括憤怒類、害怕類、悲傷類、同情類、愛與喜悅類、謾罵類，共 2,659 個詞彙作為特徵向量，因此每個文句的特徵有 2,659 個維度，特徵值是有(1)或沒有(0)出現過該特徵詞彙。

我們進一步把訓練、測試集中帶有表情的文句取出。由於有些文句中會連續使用多個相同或不同情緒符號，為了簡化問題排除歧義性，我們只挑出帶有一個情緒符號的文句，也就是沒有分類的模糊性，這些文句以 HTML 的標記<p>為界分隔。如果文句中包含至少一個非零的特徵值，則形成訓練或測試實例。經由這樣的設定，我們分別從訓練資料集和測試資料集中獲得了 4,049 筆訓練實例，以及 1,234 筆測試實例。

3.1 實驗 1－綜合分類器

在這一組實驗中，我們將 4,049 筆訓練資料，直接訓練出一個能標記出 40 種情緒類別的分類器。根據訓練標記數量的分佈，「大笑」類的 342 筆最多，約佔 8.45%；「豬頭」類的 8 筆最少，僅佔 0.2%。訓練出來的分類器直接套用在原始的訓練集，得到的內部測試分類正確率是 55.32%；套用在測試集所得到的外部測試分類正確率，僅有 14.02%。雖然比基準值 8.45%高，卻沒有實際上應用的空間。

表 5、40 類綜合分類器錯誤分析

正確答案	數量	標記答案															
		大笑	開懷	愛意	大哭	微笑	吐舌	驚訝	花	生氣	祈禱	害羞	難過	鼓掌	擔心	疑惑	眨眼
大笑	109	**55**	11	9	15	5	3		1	2	2		1		1	1	1
開懷	81	28	**9**	15	13	4	3		1	2	2	1	1			1	
愛意	90	17	9	**47**	9	1	2			2	2						
大哭	75	17	2	3	**31**	2	1	2	1	5			2		3	2	
微笑	75	22	6	22	8	**3**	1			1	6	2				1	
吐舌	71	23	9	17	9	2				2	1	2	1	1			
驚訝	43	13	3	3	16					1			1		3	1	
花	51	13	2	20	5	1			**2**		5				1	2	
生氣	41	17	1		7	1	1		1	**2**	2		2	1	1		
祈禱	52	5	4	5	17					2	**13**	1	2		1		
害羞	36	9	5	13	4	2			1			**1**					
難過	33	6		3	7			1		1	2		**8**		3	2	
鼓掌	29	9	6	5	4	2					2						
擔心	37	10	1	2	18			2			1		1				
疑惑	28	7	7	2	9		1										
眨眼	30	11	5	6	2					1	3						

表 6、不同類別個數的綜合分類器的效能

分類器	訓練量	測試量	基準值	內測 正確率	外測 正確率	內測 提昇率	外測 提昇率
40 類	4,049	1,234	8.45%	55.32%	14.02%	46.87%	5.57%
32 類	3,877	1,177	8.82%	56.10%	14.70%	47.28%	5.88%
16 類	2,937	881	11.77%	61.94%	19.98%	50.17%	8.21%
8 類	1,860	595	18.38%	68.98%	28.91%	50.60%	10.53%
4 類	1,225	355	27.92%	79.51%	46.76%	51.59%	18.84%
2 類	628	190	54.46%	85.03%	56.84%	30.57%	2.38%

表 7、16 類綜合分類器錯誤分析

正確答案	數量	標記答案															
		大笑	開懷	愛意	大哭	微笑	吐舌	驚訝	花	生氣	祈禱	害羞	難過	鼓掌	擔心	疑惑	眨眼
大笑	109	**56**	12	9	15	5	3		1	2	2		1		1	1	1
開懷	81	28	**9**	15	13	4	3		1	2	2	1	1	1		1	
愛意	90	17	9	**47**	9	1	2			2	3						
大哭	75	18	2	3	**33**	2	1	2	1	5	1		2		3	2	
微笑	75	23	6	22	10	**3**	1			1	6	2				1	
吐舌	71	24	9	17	10	2				2	1	2	1	1	1		1
驚訝	43	13	3	3	16					1			3		3	1	
花	51	13	2	20	5	1			**2**		5				1	2	
生氣	41	17	1		11	1	1	1	1	**2**	2		2	1	1		
祈禱	52	6	4	5	17					2	**14**	1	2		1		
害羞	36	9	5	13	5	2			1			1					
難過	33	6		3	7			1		1	2		**8**		3	2	
鼓掌	29	9	6	5	4	2					2			1			
擔心	37	12	1	2	18			2			1		1				
疑惑	28	8	7	2	9		1						1				
眨眼	30	12	5	6	3					1	3						

為了調整分類器設計的策略，我們先針對綜合分類器表現不佳的原因進行探討，進行錯誤分析。表 5 統計訓練集實例數量分佈前 16 名情緒分類，在測試集所得到的表現，以進行錯誤分析。留意訓練集分佈數量排名，跟測試集排名可能有些許差異，例如「愛意」分類的測試實例有 90 筆，比「開懷」類 81 筆多。表格中對角線上顯示粗體字的格子，表示該分類分到正確類別的數量。標示灰底的格子，表示該分類在綜合分類器處理過後分到最多的地方。表現最好的分類是「大笑」和「愛意」，但正確率僅有 5 成上下。根據對角線往右下角觀察，甚至可以發許多分不到正確類的情況。以斜線標記的格子代表測試後，沒有任何一筆分到該類別，我們可以發現大部分的文句都有機會分到前四類，卻很難歸類到如「驚訝」、「眨眼」等類別。為了進一步了解類別個數是否會影響到分類器表現，我們分別保留排名分類前 32、16、8、4、2 名的分類，來篩選出不同的訓練測試集實例，以這些實例所做出的分類結果如表 6 所示。

表 6 顯示隨著分類篩選數目的減少，訓練量和測試量也隨之下降，但是在基準值、內部測試分類正確率、外部測試分類正確率部份，皆隨著分類難度的減少而提昇，這個結果指出分類類型的減少有助於分類器效能的提昇。另外內、外部測試分類正確提昇率為內、外部分類測試正確率減去基準值，提昇的幅度隨著分類的減少也呈現爬升的狀態，唯獨到了二元分類器時提昇程度僅剩 2.38%。分析其原因是由於前 2 類「大笑」和「開懷」是相當接近的情緒，造成分類器難以區分。另外表 7 類似表 5，列出 16 類綜合分類器分類正確和錯誤的數量統計，各類正確分類數(對角線上) 提昇約 0 到 2 筆不等。雖然觀察到的正確分類所提昇的數量有限，但是綜合表 5、表 7 我們可以發現－儘管文句很容易分類到前四類，但是屬於「難過」類的就不會被標成「開懷」類，屬於「生氣」類的就不會被標成「愛意」類。

3.2 實驗 2－各種二元分類器效能之比較

根據上一節對表 6 的觀察顯示二元分類器提昇效能的潛力，而針對表 5、表 7 的交叉分析，也讓我們推測特定類別特徵之間可能有互斥的表現。因此我們接著將 40 個類別的任兩類的資料，來訓練出共 C_2^{40}=780 個二元分類器。為了符合頁面顯示，我們先將前 16

類配對形成的C_2^{16}=120 個二元分類器的實驗數據歸納在表 8，其中右上角為外部測試分類正確率、左下角為基準值，標示灰底的格子表示外部測試分類正確率比基準值還要差。對角線上的格子沒有特殊意義，例如不需要訓練「大笑 vs.大笑」分類器。

分析表 8 的數據，以正面情緒「大笑」為例，跟「開懷」(56.8%)、「吐舌」(58.9%、小於基準值 66.3%)表示最不容易區分，而跟負面情緒「難過」(78.2%)、「疑惑」(78.8%、小於基準值 79.9%)最容易區分開來。另外，正面情緒更強烈的「愛意」，跟其他負面情緒「大哭」(79.4%)、「擔心」(79.5%)、「難過」(80.5%)、「生氣」(81.7%)都容易區分開來。但是「愛意」跟「微笑」(56.4%)和「花」(64.5%)，除了基準值低外，外部測試分類正確率也無法超越，這顯示跟這兩組正面情緒難以區分開來。

我們另外製作一張類似表 8 的分析表格，首先把相同的灰色區塊複製到表 9，接著在右上三角填入對應二元分類器內部測試分類正確率、左下三角填入對應內部測試分類提昇率(內部測試分類確率減掉基準值)，其提昇率範圍分佈在 13.0%到 43.9%之間，提昇率小於 25%者以粗體字表示。表 9 的數據在訓練後的內部測試階段即可獲得，藉由這樣的觀察我們獲得在內部測試階段可以應用上的經驗法則為：

(h1) 正面情緒和負面情緒比較容易區分開來，其二元分類器所表現的內部測試分類提昇率上升的幅度較高。

(h2) 如果某二元分類器內部測試分類提昇率過低，如「大笑 vs.疑惑」為 14.0%、「愛意 vs.花」為 18.8%，可以預期該分類器在外部測試時也很難有好的表現。

(h3) 如果內部測試後，該類所有分類器平均提昇率過低，如「疑惑」類，則該類在外部測試時，也很難有好的表現。

(h4) 內部測試分類提昇率皆為正數，表示以 SVM 來解決本研究議題有一定程度的強健度，或許有機會再透過訓練資料的增加，進一步提昇正確率。

表 8、二元分類器外部測試正確率與基準值實驗結果

	大笑	開懷	愛意	大哭	微笑	吐舌	驚訝	花	生氣	祈禱	害羞	難過	鼓掌	擔心	疑惑	眨眼
大笑		56.8%	72.9%	69.6%	62.5%	58.9%	69.7%	73.1%	74.7%	72.7%	75.2%	78.2%	76.1%	74.7%	78.8%	73.4%
開懷	54.5%		64.9%	70.5%	56.4%	49.3%	66.1%	63.6%	73.8%	69.9%	68.4%	74.6%	71.8%	76.3%	72.5%	68.5%
愛意	52.1%	52.4%		79.4%	56.4%	62.1%	74.4%	64.5%	81.7%	77.5%	71.4%	80.5%	73.1%	79.5%	78.8%	74.2%
大哭	54.8%	50.4%	52.8%		74.7%	64.4%	63.6%	69.8%	62.9%	70.1%	80.2%	70.4%	75.0%	64.3%	72.8%	79.0%
微笑	64.7%	60.5%	62.7%	60.1%		62.3%	75.4%	62.7%	81.0%	66.1%	66.7%	74.1%	69.2%	77.7%	73.8%	65.7%
吐舌	66.3%	62.2%	64.4%	61.8%	51.8%		66.7%	62.3%	69.6%	70.7%	58.9%	71.2%	56.0%	71.3%	73.7%	65.3%
驚訝	72.2%	68.4%	70.5%	68.1%	58.6%	56.9%		68.1%	61.9%	66.3%	70.9%	53.9%	63.9%	50.0%	63.4%	64.4%
花	70.7%	66.8%	68.9%	66.5%	56.8%	55.1%	51.8%		66.3%	59.2%	60.9%	59.5%	57.5%	70.5%	58.2%	53.1%
生氣	69.7%	65.7%	67.9%	65.4%	55.7%	53.9%	53.0%	51.2%		64.5%	81.8%	63.5%	65.7%	64.1%	62.3%	73.2%
祈禱	67.2%	63.1%	65.4%	62.8%	52.8%	51.0%	55.9%	54.0%	52.8%		68.2%	58.8%	63.0%	61.8%	63.8%	61.0%
害羞	69.7%	65.7%	67.9%	65.4%	55.7%	53.9%	53.0%	51.2%	63.8%	52.8%		79.7%	58.5%	80.8%	62.5%	63.6%
難過	69.9%	66.1%	68.2%	65.7%	56.0%	54.2%	52.7%	50.9%	50.3%	53.2%	50.3%		67.7%	42.9%	55.7%	69.8%
鼓掌	75.5%	72.0%	73.9%	71.8%	62.8%	61.1%	54.3%	56.1%	57.3%	60.1%	57.3%	57.0%		71.2%	50.9%	52.5%
擔心	74.2%	70.6%	72.6%	70.3%	61.1%	59.4%	52.6%	54.4%	55.6%	58.4%	55.6%	55.3%	51.7%		55.4%	67.2%
疑惑	81.2%	78.4%	79.9%	78.1%	70.3%	68.8%	62.6%	64.3%	65.4%	67.9%	65.4%	65.0%	58.4%	60.1%		62.1%
眨眼	73.2%	69.6%	71.6%	69.3%	59.9%	58.2%	51.4%	53.2%	54.4%	57.2%	54.4%	54.0%	53.0%	51.2%	61.3%	

表 9、二元分類器內部測試正確率與內部提昇率實驗結果

	大笑	開懷	愛意	大哭	微笑	吐舌	驚訝	花	生氣	祈禱	害羞	難過	鼓掌	擔心	疑惑	眨眼
大笑		85.0%	89.6%	94.1%	90.0%	87.4%	93.0%	94.0%	94.1%	95.9%	92.3%	94.1%	93.6%	94.6%	95.2%	89.5%
開懷	30.6%		86.5%	93.8%	85.8%	87.0%	92.6%	92.8%	94.0%	93.6%	90.3%	92.8%	91.7%	95.8%	94.8%	88.6%
愛意	37.6%	34.1%		95.1%	84.5%	89.4%	94.6%	87.7%	95.3%	91.9%	86.4%	93.9%	92.5%	97.2%	96.2%	88.9%
大哭	39.3%	43.5%	42.4%		93.0%	91.2%	90.6%	94.8%	92.8%	93.5%	95.1%	89.5%	94.4%	93.5%	93.9%	93.9%
微笑	25.3%	25.4%	**21.7%**	32.8%		85.0%	92.8%	89.1%	94.0%	91.0%	87.8%	91.3%	91.3%	96.4%	94.4%	86.9%
吐舌	**21.1%**	**24.8%**	**24.9%**	29.4%	33.2%		90.8%	92.7%	93.5%	92.4%	88.5%	90.7%	91.2%	94.5%	92.9%	87.3%
驚訝	**20.9%**	**24.2%**	**24.2%**	**22.5%**	34.2%	34.0%		94.2%	92.2%	93.6%	93.6%	88.9%	93.0%	92.0%	90.0%	91.1%
花	**23.3%**	25.9%	**18.8%**	28.3%	32.2%	37.7%	42.3%		95.5%	92.6%	90.0%	93.4%	92.9%	98.1%	95.9%	90.3%
生氣	**24.4%**	28.3%	27.4%	27.4%	38.4%	39.6%	39.1%	44.3%		94.6%	93.0%	93.2%	95.0%	94.4%	89.5%	94.2%
祈禱	28.7%	30.5%	26.6%	30.7%	38.1%	41.3%	37.8%	38.5%	41.8%		93.0%	93.3%	92.8%	96.5%	92.7%	91.8%
害羞	22.6%	24.6%	**18.5%**	29.7%	32.1%	34.7%	40.6%	38.8%	29.2%	40.2%		92.6%	90.8%	94.4%	94.3%	90.1%
難過	**24.1%**	26.8%	25.8%	**23.8%**	35.3%	36.4%	36.2%	42.6%	42.9%	40.1%	42.2%		91.9%	93.2%	91.6%	92.6%
鼓掌	**18.1%**	**19.6%**	**18.5%**	**22.6%**	28.5%	30.2%	38.7%	36.8%	37.7%	32.7%	33.5%	34.9%		95.7%	94.7%	90.7%
擔心	**20.4%**	25.2%	**24.7%**	**23.2%**	35.3%	35.2%	39.4%	43.7%	38.8%	38.1%	38.8%	38.0%	43.9%		93.9%	93.9%
疑惑	**14.0%**	**16.4%**	**16.2%**	**15.8%**	**24.1%**	**24.1%**	27.5%	31.7%	**24.1%**	**24.8%**	28.9%	26.5%	36.3%	33.8%		92.2%
眨眼	**16.3%**	**19.0%**	**17.3%**	24.6%	26.9%	29.1%	39.7%	37.1%	39.8%	34.6%	35.8%	38.6%	37.7%	42.6%	30.9%	

3.3 實驗 3－增加訓練資料量後效能之比較

根據經驗法則(h4)，本實驗引入 2.1 節所介紹的補充訓練資料集，藉此訓練實例可增加至 5,279 筆，各類的訓練實例數量也會有所更動。參照 3.1 節以排名篩選出不同實例集合，所做出的分類結果如表 10 所示。首先從表 6 引入增加訓練量前的分類器效能數據，接著也以不同分類個數套用在增量後的訓練資料，列出了一組外部測試分類正確率參考數據。最後列出增加訓練資料量後，以不同分類個數篩選實例後所作的各項測試，相對提昇率是外部測試分類正確率減去增加訓練量前外部測試分類正確率的數值。與表 6 類似的狀況，是各項評估都隨著分類數量的縮小而呈現上升的趨勢。但是與增加訓練量前的數據比較起來，在 16 類分類器與 4 類分類器的外部測試分類正確率卻降低了。另外二元分類器外部測試分類正確率大幅提昇至 70.85%，是因為「大笑 vs.愛意」比原先「大笑 vs.開懷」還容易區分的關係，因此我們仍需針對二元分類器提昇效能的表現作進一步的分析。

表 10、綜合分類器增加訓練量後之效能

	增加訓練量前		增加訓練量後			
分類器	基準值	外測正確率	基準值	外測正確率	外測提昇率	相對提昇率
40 類	8.45%	14.02%	8.94%	14.26%	5.32%	0.24%
32 類	8.82%	14.70%	9.33%	14.93%	5.60%	0.23%
16 類	11.77%	19.98%	12.35%	19.55%	7.20%	-0.43%
8 類	18.38%	28.91%	18.73%	30.20%	11.47%	1.29%
4 類	27.92%	46.76%	28.59%	45.92%	17.33%	-0.84%
2 類	54.46%	56.84%	53.58%	70.85%	17.27%	14.01%

表 11、二元分類器增加訓練量前後「單類 vs.其他類」平均外部測試正分類確率

分類	增量前	增量後	分類	增量前	增量後	分類	增量前	增量後
小丑	89.06%	88.99%	親親	72.29%	72.56%	神智不清	69.03%	68.58%
大笑	81.34%	81.38%	得意	72.08%	71.60%			
愛意	81.13%	81.25%	嘆氣	72.04%	72.82%	戴眼鏡	69.01%	70.12%
大哭	79.41%	79.28%	驚訝	71.88%	71.85%	難過	68.94%	69.30%
睡著	79.04%	76.96%	花	71.78%	72.03%	邪惡	68.73%	68.40%
微笑	78.20%	77.66%	擔心	70.79%	71.67%	不說了	67.58%	67.22%
開懷	77.88%	78.22%	流口水	70.60%	71.69%	鼓掌	67.26%	67.39%
天使	75.21%	75.44%	安靜	70.54%	70.76%	疑惑	66.59%	68.11%
豬頭	75.16%	74.39%	呆住	70.30%	71.00%	環顧	66.48%	66.20%
吐舌頭	75.15%	75.36%	呸	69.90%	71.24%	皺眉	66.42%	68.16%
生氣	74.60%	74.95%	再見	69.65%	68.25%	偷笑	66.04%	66.44%
呵欠	74.10%	72.49%	眨眼	69.06%	69.64%	考慮	65.04%	66.02%
祈禱	72.96%	71.93%	眨眨眼	69.04%	69.29%			

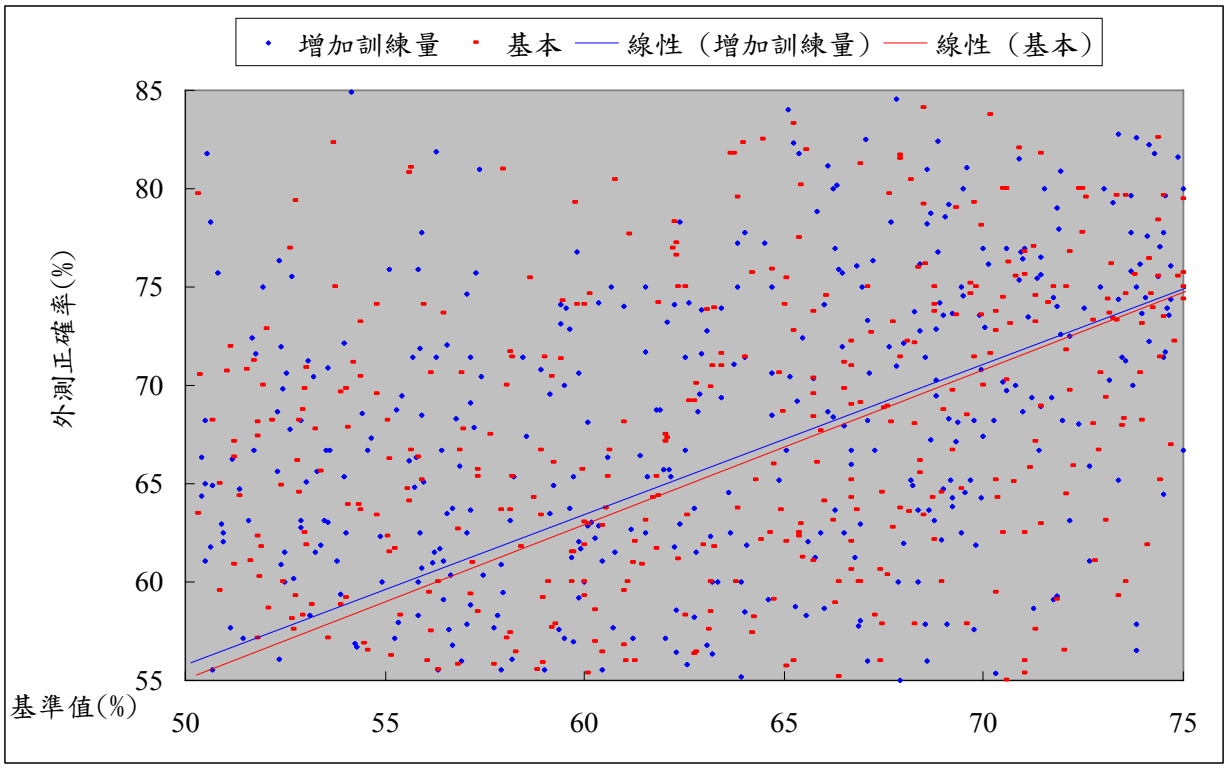


圖 2、增加訓練量前後分類器效能分佈與線性趨勢

我們首先將 40 類別中各個單一類別，與其他 39 個類別所對應的二元分類器之外部測試分類正確率，在增加訓練資料量前後的兩種狀態下取一個平均數值，並依照增加訓練量前的外部測試分類正確率由高到低的順序列出如表 11 所示，其中增加訓練量後外部測試分類正確率下降的分類以灰色網底標示之。由表 11 我們發現 40 個分類中，有 27 個分類的外部分類測試正確率在訓練量增加後上升、13 類下降。接著，我們於圖 2 將 720 個分類器的基準值與外部測試分類正確率的分佈，根據增加訓練量前後的狀況以紅、藍兩種顏色標示，並各列出一個線性趨勢線，趨勢線驗證了隨著分類基準值上昇，外部測試分類正確率也會上升的現象，並且說明增加訓練量對分類器效能的影響是正面的。

3.4 實驗 4—貪婪歸類演算法

透過上文對二元分類器效能的掌握，我們藉此設計一套貪婪演算法，試著將所有情緒歸納到正面或負面情緒類。我們曾在第 1 節提到網路界首先發明的情緒符號是微笑「:)」和難過「:(」，分別用來溝通人們最基本的正面和負面情緒。經由實驗 3 我們得到「微笑

vs.難過」分類器的基準值是 55.09%，內部測試分類正確率是 90.74%，外部測試分類正確率是 75.93%，對情緒具有一定程度的區分能力。因此，我們首先設定正面的種子情緒是「微笑」，負面的種子情緒是「難過」，並依據以下的演算法把其他 38 類情緒歸納進來：

1. 正面情緒集合 P 設成種子情緒如{微笑}，負面情緒集合 N 如{難過}。
2. 從剩下未歸類的情緒中挑出一個情緒 e：
   2.1 訓練「{P∪e} vs. N」分類器，得到內部測試分類提昇率 $increase_P$。
   2.2 訓練「P vs. {N∪e}」分類器，得到內部測試分類提昇率 $increase_N$。
3. 如果 $increase_P > increase_N$，則 e 加入集合 P，反之加入集合 N。
4. 重複步驟 2，直到所有的情緒都被歸類完畢。

歸類後的結果如表 12 所示，另外也挑選兩組種子，包括「愛意 vs.大哭」，以及都是正面的情緒「大笑 vs.開懷」，以觀察不同種子的對歸類結果的影響。前兩組結果可以發現偏正面情緒{微笑、得意、花、愛意}，兩回都可以歸在正面情緒類。相對地，{難過、大哭、嘆氣、不說了、驚訝、生氣}，兩回都歸在負面情緒類。但是，{開懷、大笑}也各有一次歸到相反類別。我們接著應用歸類結果作成正負情緒分類器，分別將(增量後)訓練與測試集原始情緒標記對應成正負類情緒標記，以分類器效能衡量貪婪歸類演算法的實用性，三組種子最後所形成的分類器效能如表 13 所示。

表 12、根據不同情緒種子套用貪婪歸類演算法對正負面情緒歸類後結果

種子	正面情緒	負面情緒
微笑 vs.難過	微笑、天使、睡著、呆住、不舒服、呵欠、邪惡、得意、耍帥、親親、考慮、眨眨眼、疑惑、鼓掌、害羞、花、吐舌頭、愛意、開懷	難過、小丑、安靜、再見、戴眼鏡、豬頭、流口水、皺眉、環顧、嘆氣、呸、偷笑、不說了、神智不清、眨眼、擔心、祈禱、生氣、驚訝、大哭、大笑
愛意 vs.大哭	愛意、睡著、豬頭、戴眼鏡、皺眉、環顧、得意、邪惡、耍帥、呸、神智不清、親親、眨眼、擔心、花、害羞、祈禱、微笑、大笑	大哭、小丑、安靜、再見、天使、流口水、呆住、呵欠、不舒服、嘆氣、考慮、偷笑、疑惑、不說了、眨眨眼、鼓掌、驚訝、生氣、難過、吐舌頭、開懷
(反例) 大笑 vs.開懷	大笑、戴眼鏡、呆住、不舒服、環顧、嘆氣、得意、耍帥、親親、考慮、眨眨眼、疑惑、擔心、害羞、生氣、花、微笑、愛意	開懷、小丑、安靜、再見、天使、睡著、豬頭、流口水、皺眉、呵欠、邪惡、呸、偷笑、不說了、神智不清、眨眼、鼓掌、難過、祈禱、驚訝、吐舌頭、大哭

表 13、種子初始二元分類器與演算法歸類後之正負情緒分類器效能比較

		基準值	內測 正確率	外測 正確率	內測 提昇率	外測 提昇率
微笑 vs.難過	初始分類器	55.09%	90.74%	75.93%	35.65%	20.84%
	正負分類器	50.31%	80.28%	63.45%	29.97%	13.14%
愛意 vs.大哭	初始分類器	53.58%	88.42%	70.85%	34.84%	17.27%
	正負分類器	51.81%	78.71%	57.78%	26.90%	5.97%
大笑 vs.開懷	初始分類器	54.25%	83.10%	56.84%	28.85%	2.59%
	正負分類器	50.41%	78.20%	56.16%	27.79%	5.75%

表 13 顯示原種子所對應的二元分類器分類效能，以及藉由該種子分類經由本節所述演算法學出最後正負情緒分類器的效能，第一組以典型的正反面情緒為種子，最後得到外部測試分類正確率最高，達 63.45%，其各項效能衡量也比其他兩組為高。第三組反例選的都是正面情緒種子，最後得到的效能最低，第二組的效能則介於其他兩組之間。

3.5 小結

實驗 1 我們看到應用在全體文章的 40 類綜合分類器的效能僅有 14.02%，並不理想。篩選成二元分類器後，雖然有 56.84%的效能，但是只能應用在兩種分類所涵蓋的實例。實驗 2 我們進一步探討 780 個二元分類的效能，發現其中帶有正反面情緒分辨意義的分類器效能較高。透過實驗 3 驗證了增加訓練資料量可以提昇效能後，實驗 4 使用貪婪歸類演算法利用世界通用「:)」和「:(」情緒符號為種子，最後得到一個能套用在全部測試實例，且效能達 63.45%的正負面情緒分類器。

4. 結論與未來研究方向

本研究探討人們的溝通媒介擴展到網際網路後，溝通行為中的情緒傳遞需求，也反應在媒介的使用上。部落格作為一個新興的媒介代表，也協助我們觀察到使用者在站台、文章、文句各個層次表達情緒的行為。我們進一步探討中文情緒處理、且由文句層次出發，試著對使用者所表達的情緒加以分類。利用雅虎奇摩提供的部落格服務，取得包含情緒符號的文本來源。並進一步透過情緒分類器的設計，以實驗數據分析問題的難度、效能提昇的方式、以及應用在情緒歸類的方法。

透過實驗章節的討論，我們最後以二元分類作為現階段歸納正面及負面情緒的方法，未來待更多訓練資料的取得，以及更多文本特徵選取的設計後，我們將進一步討論以機器協助分析情緒的效能上限、設計更高效能的多類情緒分類器。本研究以 Blog 文本進行中文情緒分析的率先嘗試，初期以分析人們使用情緒符號的現象使用出發，往後將可以此為基礎，建立對應於語言處理各階段可協助分析的情緒知識庫，幫助了解人類以使用語言來表達複雜情緒的方式。

感謝

本文部分成果由國科會計畫 NSC95-2752-E001-001-PAE.支持，在此致謝。

參考文獻

Ze-Jing Chuang and Chung-Hsien Wu, "Multi-Modal Emotion Recognition from Speech and Text," *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 9, No. 2, pp. 45-62, 2004.

C. Cortes and V. Vapnik, "Support-Vector Network," *Machine Learning*, Vol. 20, pp. 273–297, 1995.

R. J. Dolan, "Emotion, Cognition, and Behavior," Science, Vol. 298, No. 8, pp. 1191-1194, 2002.

Rong-En Fan, Pai-Hsuen Chen and Chih-Jen Lin, "Working Set Selection Using Second Order Information for Training Support Vector Machines," *Journal of Machine Learning Research*, Vol. 6, pp. 1889–1918, 2005.

Pranam Kolari, Tim Finin and Anupam Joshi, "SVMs for the Blogosphere: Blog Identification and Splog Detection," *Proceedings of AAAI 2006 Symposia on Computational Approaches to Analyzing Weblogs*, 2006.

Hugo Liu, Henry Lieberman and Ted Selker, "A Model of Textual Affect Sensing using Real-World Knowledge," *Proceedings of 2003 International Conference on Intelligent User Interfaces*, pp. 125-132, 2003.

Gilad Mishne, "Experiments with Mood Classification in Blog Posts," *Proceedings of Workshop on Stylistic Analysis of Text for Information Access*, 2005.

Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh and Yuan-Hao Chang, "Emotion Recognition and Evaluation of Mandarin Speech Using Weighted D-KNN Classification," *Proceedings of Rocling 2005*, pp. 203-212, 2005.

高台茜、倪珮晶。"華語文網路言論負向情緒用詞檢核軟體研發，"第三屆全球華文網路教育研討會（ICICE2003）論文集，393-402 頁，2003。

MiniJudge: Software for minimalist experimental syntax

James Myers[1]

[1]Graduate Institute of Linguistics, National Chung Cheng University, Minhsiung, Taiwan

Abstract

MiniJudge is free online open-source software to help theoretical syntacticians collect and analyze native-speaker acceptability judgments in a way that combines the speed and ease of traditional introspective methods with the power and statistical validity afforded by rigorous experimental design. This paper shows why MiniJudge is useful, what it feels like to use it, and how it works.

1. Introduction

Linguistics is a science because linguists test hypotheses against empirical data, but this testing is done in a much more informal way than in almost any other science. Theoretical syntacticians, for example, violate protocols standard in the rest of the cognitive sciences by acting simultaneously as experimenter and subject, and by showing little concern with the issues of experimental design and quantitative analysis deemed essential in most sciences. Linguists recognize that their informally-collected data are often inconclusive; controversial native-speaker judgments are commonplace problems in both research and teaching. From my conversations with syntacticians, I get the sense that they would appreciate a tool for collecting judgments more reliably, yet this tool should be one that permits them to maintain their traditional focus on theory rather than method.

This is where MiniJudge comes in. MiniJudge (www.ccunix.ccu.edu.tw/~lngproc/MiniJudge.htm) is software to help theoretical syntacticians design, run, and analyze linguistic judgment experiments quickly and painlessly. Because MiniJudge experiments involve testing the minimum number of speakers and sentences in the shortest amount of time, all while sacrificing the least amount of statistical power, I call them "minimalist" experiments. In this paper I first argue why a tool like MiniJudge is necessary. I then walk through a sample MiniJudge experiment on Chinese. Finally, I reveal MiniJudge's inner workings, which involve some underused or novel statistical techniques. Currently the only implementation of MiniJudge is MiniJudgeJS, which is written in JavaScript, HTML, and the statistical language R (www.r-project.org).

2. Balancing speed and reliability in syntactic judgment collection

Though some readers may wonder why we should bother with judgments when we can simply analyze corpora, judgments and corpus data are actually complementary performance windows into linguistic competence, with their own strengths and weaknesses (see e.g. Penke & Rosenbach, 2004). The

question is how we can extract the maximum value out of judgments in the easiest possible way.

2.1. Experimental syntax

Phillips and Lasnik (2003:61) are entirely right to emphasize that the "[g]athering of native-speaker judgments is a trivially simple kind of experiment, one that makes it possible to obtain large numbers of highly robust empirical results in a short period of time, from a vast array of languages." Even Labov (1996:102), who generally favors corpus data, admits that "[f]or the great majority of sentences cited by linguists," native-speaker intuitions "are reliable." Yet as Phillips and Lasnik (2003:61) also point out, "it is a truism in linguistics, widely acknowledged and taken into account, that acceptability ratings can vary for many reasons independent of grammaticality." Unfortunately, in actual practice linguists don't take the distinction between "acceptability" and "grammaticality" as seriously as they know they should, and their "trivially simple methods" become merely simple-minded (Schütze, 1996).

Since the problem of detecting competence in performance is precisely the problem faced by experimental cognitive scientists every day (e.g., testing vision theories with optical illusions), a reasonable response to the syntactician's empirical challenges would be to adopt the protocols standard in the rest of the experimental cognitive sciences: multiple stimuli and subjects (naive ones rather than the bias-prone experimenters themselves), systematic controls, factorial designs, continuous response measures, filler items, counterbalancing, and statistical analysis. When judgments are collected with these more careful protocols, they often reveal hitherto unsuspected complexity. Recent examples of the growing experimental syntax literature include Sorace & Keller (2005) and Featherston (2005); Cowart (1997) is a user-friendly handbook.

2.2. Minimalist experimental syntax

Full-fledged experimental syntax is complex, forcing the researcher to spend a lot of time on work that is not theoretically very interesting. The complexity of an experiment should actually be proportional to the subtlety of the effects it is trying to detect. Very clear judgments are detectable with traditional "trivially simple" methods; very subtle judgments may require full-fledged experimental methods. But in the vast area in between, a compromise seems appropriate, where methods are powerful enough to yield statistically valid results, yet are simple enough to apply quickly: a minimalist experimental syntax (see Table 1).

Table 1. Defining characteristics of minimalist experimental syntax

Binary *yes*/*no* judgments	No counterbalancing of sentence lists
Experimental sentences only (no fillers)	Maximum of two binary factors
Very few sentence sets (about 10)	Random sentence order
Very few speakers (about 10-20)	Order treated as a factor in the statistics

While conducting a minimalist experiment is much simpler than conducting a full-fledged judgment experiment (an explicit guide is given in Myers 2006), some steps may still be overly complex and/or intimidating to the novice experimenter, in particular the design of the experimental sentences and the statistical analysis. The purpose of the MiniJudge software is to automate these steps.

3. Using MiniJudge

To show how MiniJudge is used, I describe a recent application of it to a morphosyntactic issue in Chinese; for another example, see MJInfo.htm#resultshelp, reachable through the MiniJudge homepage. MiniJudge has also been used to run syntax experiments on English and Taiwan Sign Language, as well as to run pilots for larger studies and to help teach basic concepts in experimental design.

3.1. Goal of the experiment

He (2004) presents an interesting observation regarding the interaction of compound-internal phrase structure and affixation of the plural marker *men*. Part of his paradigm is shown in Table 2, where V = verb and O = object (based on his (2) & (4), pp. 2-3).

Table 2. The VO*men* paradigm of He (2004)

	[+men]	[-men]
[+VO]	*zhizao yaoyan zhe men *make rumor person* PLURAL	zhizao yaoyan zhe *make rumor person*
[-VO]	yaoyan zhizao zhe men *rumor make person* PLURAL	yaoyan zhizao zhe *rumor make person*

He's analysis is not relevant here; the question is simply whether or not his observation about the judgment pattern in Table 2 is empirically correct. As a non-native speaker of Chinese, I have no intuitions myself. When I have informally asked colleagues and students to double-check the judgments, I have received a mixed response, with some ruling out *men* or VO entirely, but this misses the point, since He's claim concerns the ungrammaticality of the VO*men* form relative to all the others. Some speakers shown He's starred and non-starred examples are willing to agree with his judgments, but it's likely that the star pattern has biased them. It may also be that He's generalization works for the few examples he cites, but fails in general. My goal, then, was to use MiniJudge to generate more examples to test systematically on native speakers.

3.2. The MiniJudgeJS interface

MiniJudgeJS is simply a JavaScript-enabled HTML form. Input and output are handled entirely by text areas; generated text includes code to run statistical analyses in R. Like the rest of the MiniJudge family, MiniJudgeJS divides the experimental process into the steps listed in Table 3.

Table 3. The steps used by MiniJudge

I. Design experiment	**II. Run experiment**	**III. Analyze experiment**
Choose experimental factors	Choose number of speakers	Download and install R
Choose set of prototype sentences	Write instructions for speakers	Enter raw results
Choose number of sentence sets	Print or email survey forms	Generate data file
Segment prototype set (optional)	Save schematic survey file	Save data file
Replace segments (optional)		Generate R code
Save master list of test sentences		Paste R command code into R

3.3. Designing the experiment

A MiniJudge experiment begins by choosing the experimental factors. In the case of the VO*men* claim, the paradigm in Table 2 is derived via two binary factors: [±VO] (VO vs. OV) and [±*men*] (with or without *men* suffixation). As noted above, He's observation doesn't relate to each factor separately, but rather to an interaction: the combination of the factor values [+VO] and [+*men*] is claimed to result in lower acceptability, relative to overall judgments for [+VO] and for [+*men*].

The next step is to enter the prototype set of sentences (a pair if one factor, a quartet if two factors). Similar to the example sets shown in syntax papers and presentations, the prototype set serves multiple purposes. Most fundamentally, it helps to make the logic of factorial experimental design intuitive for novice experimenters. Syntacticians are not always aware of the importance of contrasting sentences that differ *only* in theoretically relevant factors, or of the central role played by interactions in many syntactic claims (for further discussion of the relevance of factors and interactions in syntax experiments, see MJInfo.htm#factorial and MJInfo.htm#interact).

Another purpose of the prototype set is that it can be used to help generate further sentence sets that maintain the same factorial contrasts but vary in irrelevant lexical properties. In the case of the present experiment, the claim made in He (2004) says nothing about the particular verb, object, or head that is used. Thus the judgment pattern claimed for Table 2 above should also hold for the sets shown in Table 4 below, regardless of any additional influences from pragmatics, frequency, suffixlikeness (*zhe* vs. the others), or freeness (*ren* vs. the others); the stars here represent what He might predict (lexical content for the new sets was chosen with the help of Ko Yu-guang and Zhang Ning).

Table 4. Extending the VO*men* paradigm of He (2004)

	[+men]	[-men]
[+VO]	*chuanbo bingdu yuan men *spread virus person PLURAL*	chuanbo bingdu yuan *spread virus person*
[-VO]	bingdu chuanbo yuan men *virus spread person PLURAL*	bingdu chuanbo yuan *virus spread person*
[+VO]	*sheji shipin ren men *design ornaments person PLURAL*	sheji shipin ren *design ornaments person*
[-VO]	shipin sheji ren men *ornaments design person PLURAL*	shipin sheji ren *ornaments design person*

MiniJudge partly automates the process of creating new sentence sets by dividing up the prototype sentences into the largest repeating segments and replacing them with user-chosen substitutes. The prototype segments for Table 2 are shown in the first row of Table 5. The user only has to find parallel substitutes for four segments, rather than having to construct whole new sentences while keeping track of the factorial design (Table 5 also shows the segments needed to generate the new sets in Table 4). The segmentation and set generation processes are designed to work equally well in English-like and Chinese-like orthographies. Of course, since MiniJudge knows no human language, it sometimes makes strange errors, so users are allowed to correct its output, or even to generate new sets manually.

Table 5. Prototype segments and new segments for the VO*men* experiment

Set 1 (prototype) segments:	zhizao	yaoyan	zhe	men
Set 2 segments:	chuanbo	bingdu	yuan	men
Set 3 segments:	sheji	shipin	ren	men

After the user has corrected and approved the master list of sentences, it can be saved to a file for use in reports (as I am doing here). In the present experiment, the master list contained 48 sentences (12 sets of 4 sentences each). This is an unusually large number of sentences for a MiniJudge experiment; significant results have been found with experiments with as few as 10 sentences.

3.4. Running the experiment

In order to run a MiniJudge experiment, the user must make three decisions. The first concerns the maximum number of speakers to test. It is possible to get significant results with as few as 7 speakers, but in the present experiment, I generated 30 surveys. As it turned out, only 18 surveys were returned.

The second decision concerns whether surveys will be distributed by printed form or by email. In MiniJudgeJS, printing surveys involves saving the them from a text area and printing them with a word processor. MiniJudgeJS cannot send email automatically, so emailed surveys must be individually copied and pasted. In the present experiment, I emailed thirty students, former students, or faculty of my linguistics department who did not know the purpose of the experiment.

The final decision concerns the instructions, which the user may edit from a default. MiniJudgeJS requires that judgments be entered as 1 (*yes*) vs. 0 (*no*); in the current version, if surveys are to be collected electronically, these judgments must be typed before each sentence ID number. Chinese instructions for the VO*men* experiment were written with the help of Ko Yu-guang.

Surveys themselves are randomized individually to prevent order confounds, as is standard in psycholinguistics. The randomization algorithm, taken from Cowart (1997:101), results in every sentence having an equal chance to appear at any point in the experiment (by randomization of blocks), while simultaneously distributing sentence types evenly and randomly.

Each survey starts with the instructions, followed by a speaker ID number (e.g., "##02"), and finally the survey itself, with each sentence numbered in the order seen by the speaker. Because the speakers' surveys intentionally hide the factorial design, the experimenter must save this information separately in a schematic survey file. This file is meant to be read only by MiniJudgeJS; as an example, the first line of the schematic survey file for the present experiment is explained in Table 6.

Table 6. The structure of the schematic survey information file for the VO*men* experiment

File line:	`01`	`20`	`05`	`01`	`-VO`	`-men`
Explanation:	speaker ID number	sentence ID number	set ID number	order in survey	value of first factor	value of second factor

After completed surveys have been returned, the experimenter pastes them into a text area in any order (as long as each survey still contains its ID number), and pastes the schematic survey information back into another text window. MiniJudgeJS extracts judgments from the surveys and creates a data file

in which each row represents a single observation, with IDs for speakers, sentences, and sets, presentation order of sentences, factor values (1 for [+] and -1 for [-]), and judgments. As an example, the first three lines of the data file for the VO*men* experiment are shown in Table 7.

Table 7. First three lines of data file for the VO*men* experiment

Speaker	Sentence	Set	Order	VO	men	Judgment
1	20	5	1	-1	-1	1
1	45	12	2	1	1	0

3.5. Analyzing the results

For novice experimenters, the most intimidating aspect of psycholinguistic research is statistical analysis. MiniJudge employs quite complex statistical methods that are unfamiliar even to most psycholinguists, yet hides them behind a user-friendly interface. Data from a MiniJudge experiment are both categorical and repeated-measures (grouped within speakers). Currently the best available statistical model for repeated-measures categorical data is generalized linear mixed effect modeling (GLMM), which can be thought of as an extension of logistic regression (see e.g. Agresti et al., 2000).

GLMM poses serious programming challenges, so MiniJudgeJS passes the job to R, the world's foremost free statistical package (R Development Core Team, 2005). R is an open-source near clone of the proprietary program S (Chambers & Hastie, 1993), and like S, is a full-featured programming language. Its syntax is a mixture of C++ and Matlab, and of course it has a wide variety of built-in statistical functions, including many user-written packages. The specific R package used by MiniJudgeJS for GLMM is `lme4` (and its prerequisite package `Matrix`), authored by Douglas Bates and Deepayan Sarkar, and maintained by Douglas Bates. R has a simple GUI interface, and by default, the Windows version nativizes (e.g., in Chinese Windows, menus and basic messages are in Chinese).

However, since R is a command-line program, and its outputs can be unintelligible without statistical training, MiniJudgeJS handles the interface with it. The user merely enters the name of the data file, decides whether or not to test for syntactic satiation (explained below in section 3.5.2), and pastes the code generated by MiniJudgeJS into the R window. After the last line has been processed by R, the code either will generate a warning (that the file was not found or was not formatted correctly), or if all went well, will display a simple interpretive summary report. A much more detailed technical report is also saved automatically; this report is explained, step by step for the novice user, in MJInfo.htm#resultshelp.

3.5.1 A null result?

When the data file containing the 18 completed surveys in the VO*men* experiment was analyzed using the R code generated by MiniJudgeJS, the summary report in Figure 1 was produced. There are three parts: a table showing the number of *yes* judgments for each category, a listing of significant patterns (if any), and a statement about whether there was any significant confound between items and factors (discussion of this last point is reserved for section 4.3.5).

```
Number of YES judgments for each category:

              [+V]     [-V]     Total          V = VO
     [+m]     23       74       97             m = men
     [-m]     89       163      252
     Total    112      237      349

Significance summary (p < .05):

     The factor VO had a significant negative effect.
     The factor men had a significant negative effect.
     Order had a significant negative effect.
     There were no other significant effects.

     The above results do not take cross-item variability into account because
     no confound between items and factors was detected (p > .2).
```

Figure 1. Default results summary generated by MiniJudgeJS for the VO*men* experiment

The negative effects of the [VO] and [men] factors mean that items containing VO or *men* were judged worse, on average. These patterns are also clear from the table showing the number of *yes* judgments (in the total row and total column, respectively). However, as discussed in 3.1, these patterns are not what the empirical claim of He (2004) is concerned with. What we expected to see was a significant interaction between [VO] and [men], but this was not found. Instead, inspection of the technical results file shows that the *p* value for the interaction was 0.89, clearly non-significant.

However, this is not a refutation of He's claim, but merely a null result. Indeed, the number of *yes* judgments trends in the predicted direction: for VO forms, non-*men* forms were judged better than *men* forms by a ratio of almost 4:1 (89/23 = 3.87), about twice as high as the ratio for OV forms (163/74 = 2.20). That is, it was worse to affix *men* to VO forms than to OV forms, just as He claimed.

One possible cause of a null result is a confound with a nuisance variable. A clue to what this nuisance variable might be here is the significant negative effect of order, which means that judgments got worse (i.e., there was a rising probability of judging a form as unacceptable) as the experiment progressed. This shift in judgments suggests that further analysis may be advisable, as described next.

3.5.2 Syntactic satiation

Though MiniJudge factors out raw order effects in its default analysis, it is possible that order also *interacts* with one or more factors. Testing for interactions with continuous variables without a specific theoretical reason may make it more difficult to interpret main effects (see e.g. Bernhardt & Jung, 1979), but MiniJudge offers the option to test for interactions with order because it helps in the detection of syntactic satiation. This is the phenomenon (known informally as "linguist's disease") in which linguistic intuitions are dulled by repeated testing, making it harder to be confident in one's judgments. Following the logic proposed in Myers (2006), MiniJudge tests for satiation by looking for negative interactions with order: early on, the [±F] contrast is strong, but later it's weak.

Snyder (2000) argues that satiation could provide a new window into grammar and/or processing, since different types of syntactic violations differ in whether or not they satiate. Snyder suggests two possible reasons for such differences. On the one hand, satiation may be caused by processing, not

grammar, thus providing a diagnostic for performance effects (a position taken by Goodall 2004). On the other hand, satiability may differ due to differences between the components of competence itself, thus permitting a new grammatical classification tool (a position taken by Hiramatsu 2000).

Although He (2004) makes no predictions relating to satiation, the unexpected null result noted in section 3.5.1 suggests that it may be worthwhile trying out a more complex analysis that includes interactions with order. Running this analysis simply involves telling MiniJudgeJS that we want to test for satiation (by clicking a checkbox), and then pasting the generated code into R. Doing this with the VO*men* data resulted in the two new lines in Figure 2 being added to the significance summary.

```
The interaction between VO and men had a significant positive effect.
The interaction of VO * men with Order had a significant negative effect
  (satiation).
```

Figure 2. New lines in results summary when satiation was tested in the VO*men* experiment

As hoped, factoring out the interactions with order revealed a significant interaction between the factors [VO] and [men]. This shows that the ratio difference seen in Figure 1 is indeed statistically reliable (the detailed report file shows $p = 0.02$), thus vindicating He's empirical claim. This new analysis also detected satiation in the VO*men* effect; it was this interaction with order that had obscured evidence for the VO*men* effect in the default analysis.

This experiment thus not only provided reliable evidence in favor of the empirical claim made by He (2004), but it also revealed three additional patterns not reported by He: overall lower acceptability for VO forms relative to OV forms, overall lower acceptability of *men* forms, and the satiability of the VO*men* effect. Detecting satiation, and the VO*men* effect it obscured, depended crucially on the use of careful experimental design and statistical analysis, and would have been impossible using traditional informal methods. Despite this power, the MiniJudge experiment was designed, run, and analyzed within a matter of days, rather than the weeks required for full-fledged experimental syntax.

4. The inner workings

MiniJudgeJS, as with all future versions in the MiniJudge family, is free and open source. The JavaScript and R code can be modified freely by downloading the HTML file and opening it in a text editor, and both are heavily commented to make them easier to follow. In this section I give overviews of the programming relating to material generation and statistical analysis.

4.1. Material generation

As described in section 3.3, MiniJudgeJS can assist with the generation of additional sentence sets. This involves two major phases: segmenting the prototype sentences into the largest repeated substrings, and substituting new segments for old segments in the new sentence sets.

The first step is to determine whether the prototype sentences contain any spaces. If they do, words are treated as basic units, and capitalization is removed from the initial word and any sentence-final punctuation mark is also set aside (for adding again later). If there are no spaces (as in Chinese),

characters are treated as basic units. Next, the boundaries between prototype sentences are demarcated to indicate that cross-sentence strings can never be segments. The algorithm for determining other segment boundaries requires the creation of a lexicon containing all unique words (or characters) in the prototype corpus. If the algorithm detects that items from the corpus and from the lexicon match only if one of the items is lowercase, this item is recapitalized. Versions of the prototype sentences with "word-based" capitalization is later used when old segments are replaced by new ones.

The most crucial step in the segmentation algorithm is to check each word (or character) in the lexicon to determine whether or not it has at least two neighbors on the same side in the corpus. For example, suppose the prototype set consisted of the sentences "A dog loves the cat. The cat loves a dog." The lexical item "loves" has two neighbors on the left: "dog" and "cat". Thus a segment boundary should be inserted to the left of "loves" in the corpus. Similarly, the right neighbor of "loves" is sometimes "the" and sometimes "a"; hence "loves" will be treated as a whole segment. By contrast, the lexical item "cat" always has the same item to its left (once sentence-initial capitalization is removed): "the". Similarly, the right neighbor of "the" is always "cat". Thus "the cat" will be treated as a segment, and the same logic applies to "a dog". The prototype segments are thus "a dog", "loves", "the cat".

The final phase involves substituting the user-chosen new segments for the prototype segments. This is done using JavaScript's built-in regular expression functions, which only became available with Netscape 4 and Internet Explorer 4.

4.2. Statistical analysis

The statistical analyses conducted by MiniJudgeJS involve several innovations: the use of GLMM, the inclusion of order and interactions with order as factors, the use of JavaScript to communicate with R, the use of R code to extract key values from R's technical output so that a simple report can be generated, and the use of R code to compare by-subject and by-subject-and-item analyses to decide whether the latter is really necessary. In this section I describe each of these innovations in turn.

4.3.1 GLMM

As explained in section 3.5, generalized linear mixed effect modeling (GLMM) is conceptually akin to logistic regression, which is at the core of the sociolinguistic variable-rule analyzing program VARBRUL and its descendants (Mendoza-Denton et al. 2003), but unlike logistic regression, GLMM regression equations also include random variables (e.g., the speakers); see Agresti et al. (2000). One major advantage of a regression-based approach is that no data are thrown away. Moreover, since each observation is treated as a separate data point, GLMM is usually not affected much by missing data, but only if they are missing non-systematically (this is why participants in MiniJudge experiments are requested to judge *all* sentences, guessing if they're not sure).

Though GLMM is the best statistical model currently available for repeated-measures categorical data, it does have some limitations. First, R's implementation of GLMM tests significance using *z* scores, which are reliable only if the number of observations is greater than 50 or so, but in actual practice, 50 judgments are trivial to collect (e.g., 5 speakers judging 10 sentences each). Second, like

regression in general, GLMM assumes that the correlation between the dependent and independent variables is not perfect, so it is paradoxically unable to confirm the significance of perfect correlations. Third, like logistic regression (but unlike ANOVA or ordinary regression), it is impossible to calculate GLMM coefficients and *p* values perfectly; they can only be estimated. Unfortunately, the best way to estimate GLMM values is extremely complicated and slow, so R uses "simpler" yet less accurate estimation methods. Currently, R provides two options for estimating GLMM coefficients: the faster but less accurate penalized quasi-likelihood approximation, and the slower but more accurate Laplacian approximation. MiniJudgeJS uses the latter.

The function in the `lme4/Matrix` packages used for GLMM is `lmer`, which can also handle linear mixed-effect modeling (i.e., repeated-measures linear regression). The syntax is illustrated in Figure 3, which shows the commands used to run the final analyses described above in section 3.5.2. "Factor1" and "Factor2" are variables whose values are set in the R code to represent the actual factors. The use of categorical data is signaled by setting the distribution family to "binomial". The name of the loaded data file is arbitrarily called "minexp" (for minimalist experiment). The first function treats only subjects as random, while the second function treats both subjects and items as random. The choice to test for satiation or not is determined by the user; based on this choice, JavaScript generates different versions of the R code. The choice to run one-factor or two-factor analyses is determined by the R code itself by counting the number of factors in the data file. Both analyses in Figure 3 are always run, and then compared with another R function described below in 4.3.5.

```
glmm1 = lmer(Judgment ~ Factor1 * Factor2 * Order + (1|Speaker), data = minexp,
  family = "binomial", method = "Laplace")
glmm2 = lmer(Judgment ~ Factor1 * Factor2 * Order + (1|Speaker) + (1|Sentence),
  data = minexp, family = "binomial", method = "Laplace")
```

Figure 3. R commands for computing GLMM when testing satiation in a two-factor experiment

4.3.2 Order as a factor

MiniJudgeJS includes order as a factor whether or not the user tests for satiation, to compensate for the fact that MiniJudge experiments use no counterbalanced lists of sentences across subgroups of speakers. List counterbalancing is used in full-fledged experimental syntax so that speakers don't use an explicit comparison strategy when judging sentences from the same set (a comparison strategy may create an illusory contrast or have other undesirable consequences). However, comparison can only occur when the second sentence of a matched pair is encountered. If roughly half of the speakers get sentence type [+F] first and half get [-F] first, then on average, judgments for [+F] vs. [-F] are only partially influenced by a comparison strategy. The comparison strategy (if any) will be realized as an order effect: early judgments (when comparison is impossible) will be different from later judgments. Thus factoring out order effects in the statistics serves roughly the same purpose as counterbalanced lists.

4.3.3 JavaScript as an R interface

JavaScript is much more powerful than many programmers realize. In fact, a key inspiration for MiniJudgeJS was the Logistic Regression Calculating Page (http://statpages.org/logistic.html), a

JavaScript-enabled HTML file written by John C. Pezzullo. Using only basic platform-universal JavaScript, the page collects data, reformats it, estimates logistic regression coefficients via a highly efficient maximum likelihood estimation algorithm, and generates chi-square values and *p* values. Thus a JavaScript-only version of MiniJudgeJS is conceivable, without any need to pass work over to R. Unfortunately, the necessary statistical programming is quite formidable.

Instead, in MiniJudgeJS the role of JavaScript in the statistical analysis is mainly as a user-friendly GUI. Since the statistics needed for a MiniJudge experiment is highly standardized, very little input is needed from the user, but the potential to use JavaScript to interface with R in more flexible ways is there. This would help fix a major limitation with R, whose command-line interface is quite intimidating for novice users, and whose online help leaves a lot to be desired (cf. Fox, 2005).

Of course, using JavaScript as an interface has its limitations, the most notable of which are the built-in security constraints that prevent JavaScript from being able to read or write to files, or to communicate directly with other programs. For example, it's impossible to have JavaScript run R in the background, to save users the bother of copying and pasting in R code. This is why we are currently exploring other versions of MiniJudge. One that has made some progress is MiniJudgeJava, written by Chen Tsung-ying in Java using its own platform-independent GUI tools. Interfacing with R is likely to remain tricky, however, unless we create something like MiniJudgeR, written in R itself, or figure out how to program GLMM directly in JavaScript.

4.3.4 R code to simplify output

GLMM is a high-powered statistical tool, unlikely to be used by people who don't already have a strong background in statistics, and so the outputs generated by R are not understandable without such a background. Since MiniJudge is intended for statistical novices, extra programming is needed to translate R output into plain language. For MiniJudgeJS, the most crucial portion of R's output for GLMM is the matrix containing the regression coefficient estimates and *p* values, like that shown in Figure 4 (from the VO*men* experiment, without testing for satiation). The trick is to extract the estimates (the signs of which provide information about the nature of the pattern) and the *p* values (which indicate significance) in order to generate a simple summary containing no numbers at all.

```
                  Estimate     Std. Error   z value     Pr(>|z|)
(Intercept)       -0.0810381   0.2330613    -0.3477     0.728057
Factor1           -0.8090969   0.0886143    -9.1305     < 2.2e-16
Factor2           -0.9741367   0.0891447    -10.9276    < 2.2e-16
Order             -0.0192680   0.0059976    -3.2126     0.001315
Factor1:Factor2   0.0119932    0.0877194    0.1367      0.891250
```

Figure 4. GLMM output generated by `lmer` for the VO*men* experiment without testing for satiation

Unfortunately, the output of the `lmer` function is a list object, containing only the parameters used to compute the estimates and *p* values, not the values themselves. Thus the R code generated by MiniJudgeJS "sinks" `lmer`'s displayed output to an offline file, and then reads this file back in as a

string (the offline file becomes the permanent record of the detailed analysis). The string is then searched for the string "(Intercept)" which always appears at the upper left of the value matrix. The coefficient is the first value to the right of this, and the *p* value is the fourth value (skipping "<", if any).

If the *p* value associated with a factor or interaction is less than 0.05, a summary line is generated that gives the actual factor name and the sign of the estimate, as in Figures 1 and 2 above. The R code generates the summary table counting the number of *yes* judgments for each category (see Figure 1) directly from the data file itself.

4.3.5 By-subject and by-item analyses

MiniJudgeJS runs both by-subject and by-subject-and-item analyses, but it reports only the first in the main summary unless it finds that the more complex analysis is really necessary. This approach differs from standard psycholinguistic practice, where both by-subject and by-item analyses are always run. A commonly cited reason for always running a by-item analysis is that it is required to test for generality across items, just as a by-subject analysis tests for generality across subjects. However, this logic is based on a misinterpretation of Clark (1973), the paper usually cited as justification.

First, it is wrong to think that by-item analyses check to see if any item behaves atypically (i.e., is an outlier). For parametric models like ANOVA, it is quite possible for a single outlier to cause an illusory significant result, even in a by-item analysis (categorical data analyses like GLMM don't have this weakness). To test for outliers, there's no substitute for checking the individual by-item results manually. MiniJudge helps with this by reporting the by-sentence rates of *yes* judgments in a table saved as part of the offline analysis file; items with unusually low or high acceptability relative to others of their type stand out clearly. In the case of the VO*men* experiment, this table did not seem to show any outliers.

The second problem with the standard justification for performing obligatory by-item analyses, as Raaijmakers et al. (1999) emphasize, is that the advice given in Clark (1973) actually applies only to experiments without matched items, such as an experiment comparing a random set of sentences with transitive verbs ("eat" etc) with a random set of sentences with unrelated intransitive verbs ("sleep" etc). Such sentences will differ in more than just the crucial factor (transitive vs. intransitive), so even if a difference in judgments is found, it may actually relate to uninteresting confounded properties (e.g., the lexical frequency of the verbs). However, if lexically matched items are used, as in the VO*men* experiment, there is no such confound, since items within each set differ only in terms of the experimental factor(s). If items are sufficiently well matched, taking cross-item variation into account won't make any difference in the analysis (except to make it much more complicated), but if they are not well matched, ignoring the cross-item variation will result in misleadingly low *p* values.

Nevertheless, if we only computed models that take cross-item variation into account, we might lose useful information. After all, a high *p* value does not necessarily mean that there is no pattern at all, only that we have failed to detect it. Thus it may be useful to know if a by-speaker analysis is significant even if the by-sentence analysis is not. Such an outcome could mean that the significant by-speaker result is an illusion due to an uninteresting lexical confound, but it could instead mean that if

we do a better job matching the items in our next experiment, we will be able to demonstrate the validity of our theoretically interesting factor. Thus MiniJudge runs both types of analyses, and only chooses the by-subjects-and-items analysis for the main report if a statistically significant confound between factors and items is detected. The full results of both analyses are saved in an off-line file, along with the results of the statistical comparison of them.

The R language makes it quite easy to perform this comparison. The model in which only speakers are treated as random is a special case of the model in which both speakers and sentences are treated as random. This means the two GLMM models can be compared by a likelihood ratio test using ANOVA (see Pinheiro & Bates, 2000). As with the output of the `lmer` function, the output of the `lme4` package's `anova` function makes it difficult to extract *p* values, so again the output is "sunk" to the offline analysis file to be read back in as a string. Only if the *p* value is below 0.05 is the more complex model taken as significantly better. If the *p* value is above 0.2, MiniJudgeJS assumes that items and factors are not confounded and reports only the by-subjects-only analysis in the main summary. Nevertheless, MiniJudgeJS, erring on the side of caution, gives a warning if $0.2 > p > 0.05$. In any case, both GLMM analyses are available for inspection in the offline analysis file. Each analysis also includes additional information, generated by `lmer`, that may help determine which one is really more reliable, including variance of the random variables and the estimated scale (compared with 1); these details are explained in MJInfo.htm#resultshelp.

In the case of the VO*men* experiment, the comparison of the two models showed that the by-subjects-only model was sufficient ($p = 1$). This is unsurprising, given that the materials were almost perfectly matched, and that the by-items table showed no outliers among the sentence judgments.

The final problem with the standard justification for automatic by-item analyses is one that even Raaijmakers et al. (1999) fail to point out. Namely, since repeated-measures regression models make it possible to take cross-speaker and cross-sentence variation into account at the same time, without throwing away any data, they are superior to standard models like ANOVA. To learn more about how advances in statistics have made some psycholinguistic traditions obsolete, see Baayen (2004).

5. Conclusions

MiniJudge, currently implemented only in the form of MiniJudgeJS, is software for theoretical syntacticians without any experimental training who want to collect and interpret judgments quickly and reliably. Though MiniJudgeJS is limited in some ways, in particular in how it interfaces with R, it is still quite easy to use, as testing by my students has demonstrated. Moreover, it is unique, offering syntacticians power that they cannot obtain any other way. Behind this power are original programming and statistical techniques. Finally, MiniJudgeJS is an entirely free, open-source program (as will be all future versions). Anyone interested is invited to try it out, save it for use offline, and contribute to its further development.

6. Acknowledgements

This research was supported by National Science Council (Taiwan) grant NSC 94-2411-H-194-018. MiniJudgeJS is co-copyrighted by National Chung Cheng University. Experimental or programming help came from my research assistants Ko Yu-guang and Chen Tsung-yin. The students in my spring 2006 class *Competence & Performance* helped test MiniJudgeJS and made useful suggestions. John C. Pezzullo and Harald Baayen also provided helpful information on programming and statistical matters. Of course I am solely responsible for any mistakes.

7. References

1. A. Agresti, J. G. Booth, J. P. Hobert, & B. Caffo, "Random-effects Modeling of Categorical Response Data," *Sociological Methodology*, Vol. 30, pp. 27-80, 2000.
2. R. H. Baayen, "Statistics in Psycholinguistics: A Critique of Some Current Gold standards," *Mental Lexicon Working Papers*, Vol. 1, University of Alberta, Canada, 2004, pp. 1-45. www.mpi.nl/world/persons/private/baayen/submitted/statistics.pdf
3. I. Bernhardt & B. S. Jung, "The Interpretation of Least Squares Regression with Interaction or Polynomial Terms," *The Review of Economics and Statistics*, Vol. 61, No. 3, 1979, pp. 481-483.
4. J. M. Chambers & T. J. Hastie, *Statistical Models in S*, Chapman & Hall, 1993.
5. H. Clark, "The Language-as-fixed-effect Fallacy: A Critique of Language Statistics in Psychological Research," *Journal of Verbal Learning and Verbal Behavior*, Vol. 12, pp. 335-359, 1973.
6. W. Cowart, *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Sage Publications, London, 1997.
7. S. Featherston, "Magnitude Estimation and What It Can Do for Your Syntax: Some wh-constraints in German," *Lingua*, Vol. 115, No. 11, pp. 1525-1550, 2005.
8. J. Fox, "The R Commander: A Basic-statistics Graphical User Interface to R," *Journal of Statistical Software*, Vol. 14, No. 9, 2005.
9. G. Goodall, "On the Syntax and Processing of wh-questions in Spanish," in *WCCFL 23 Proceedings*, B. Schmeiser, V. Chand, A. Kelleher, & A. Rodriguez (eds), Cascadilla Press, Somerville, MA, 2004, pp. 101-114.
10. Y. He, "The Words-and-rules Theory: Evidence from Chinese Morphology," *Taiwan Journal of Linguistics*, Vol. 2, No. 2, pp. 1-26, 2004.
11. K. Hiramatsu, *Accessing Linguistic Competence: Evidence from Children's and Adults' Acceptability Judgements*. Doctoral dissertation, University of Connecticut, Storrs, 2000.
12. W. Labov, "When Intuitions Fail," in *CLS 32: Papers from the Parasession on Theory and Data in Linguistics*, L. McNair (ed), University of Chicago, pp. 77-105, 1996.
13. N. Mendoza-Denton, J. Hay, & S. Jannedy, "Probabilistic Sociolinguistics: Beyond Variable Rules," in *Probabilistic linguistics*, R. Bod, J. Hay, & S. Jannedy (eds), MIT Press, Cambridge, MA, pp. 97-138, 2003.

14. J. Myers, "An Experiment in Minimalist Experimental Syntax," National Chung Cheng University ms. Submitted, 2006.
15. M. Penke & A. Rosenbach, "What Counts as Evidence in Linguistics? An Introduction," *Studies in Language*, Vol. 28, No. 3, pp. 480-526, 2004.
16. C. Phillips& H. Lasnik, "Linguistics and Empirical Evidence: Reply to Edelman and Christiansen," *Trends in Cognitive Science*, Vol. 7, No. 2, pp. 61-62, 2003.
17. J. C. Pinheiro & D. M. Bates, *Mixed-Effects Models in S and S-Plus*. Springer, Berlin, 2000.
18. R Development Core Team. "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2005, URL http://www.R-project.org.
19. J. G. W. Raaijmakers, J. M. C. Schrijnemakers, & F. Gremmen, "How to Deal with 'the Language-as-fixed-effect Fallacy': Common Misconceptions and Alternative Solutions," *Journal of Memory and Language*, Vol. 41, pp. 416-426, 1999.
20. C. T. Schütze, *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press, Chicago, 1996.
21. W. Snyder, "An Experimental Investigation of Syntactic Satiation Effects," *Linguistic Inquiry*, Vol. 31, pp. 575-582, 2000.
22. A. Sorace & F. Keller, "Gradience in Linguistic Data," *Lingua*, Vol. 115, pp. 1497-1524, 2005.

基於特製隱藏式馬可夫模型之中文斷詞研究

Chinese Word Segmentation using Specialized HMM

林千翔、張嘉惠

國立中央大學資訊工程學系

Email: pshivp@db.csie.ncu.edu.tw, chia@csie.ncu.edu.tw

摘要

中文斷詞在中文的自然語言處理上，是個相當基礎且非常重要的工作。近年來的斷詞系統較傾向於機器學習式演算法來解決中文斷詞的問題，但使用傳統的作法，隱藏式馬可夫模型在解決中文斷詞的問題上，無法達到較好的斷詞效能（F-measure 約 80%），所以許多研究都是使用外部資源或是結合其他的機器學習演算法來幫助斷詞。本研究的目的是使用「特製化」（specialization）的概念來提升隱藏式馬可夫模型的準確率，我們的作法是給予隱藏式馬可夫模型更多的資訊，在完全不修改模型之訓練及測試過程的前提下，透過兩階段特製化的方式，分別為擴充「觀測符號」，以及擴充「狀態符號」的方式，大大地改善了隱藏式馬可夫模型的斷詞準確性。第一階段中，我們使用長詞優先法，來增加額外的資訊於隱藏式馬可夫模型中，使得模型擁有更多的斷詞資訊做學習。於實驗結果發現，只使用這個最簡單的長詞優先斷詞方法，確實能大幅地提升隱藏式馬可夫模型的效能。而第二階段中，我們則使用詞彙式隱藏式馬可夫模型（Lexicalized HMM）的概念，也就是只根據某些特製詞（specialized words）來做特製化，將狀態做延伸，實驗結果也證明詞彙式隱藏式馬可夫模型可再次提升系統斷詞效能。

1. 緒論

中文斷詞在中文的自然語言處理上，是非常重要的前置處理工作。許多中文的自然語言相關的領域，例如：問答系統、自動摘要、文件檢索、機器翻譯、語音辨識...等，都需要先處理中文斷詞，可見中文斷詞是個相當基礎且非常重要的工作。

所謂的「中文斷詞」就是將一連串的中文「字串」轉換成「詞串」的組合。例如：「我昨天去台北」這個中文句子，透過中文斷詞的處理後變成「我／昨天／去／台北」，也就是將｛我、昨、天、去、台、北｝字串轉成｛我、昨天、去、台北｝的詞串組合。傳統上，處理中文斷詞會遇到的問題，大致可歸納為兩點，一是「歧義性」（ambiguity）問題，二是「未知詞」（unknown word）問題。歧義性問題即是同一個中文字串，於不同的文章當中，存在不同的斷詞結果，因此容易造成斷詞上的錯誤。歧義型態大致上可以分為兩類：

- 交集型歧義（overlapping ambiguity）

  令 x, y, z 代表中文字元所組成的字串，若 x、z、xy 與 yz 皆為辭典中的詞，則 xyz 的組合，於不同的文章中，可能會被斷詞成 xy/z 或 x/yz 等兩種不同的結果，則 xyz 稱為「交集型歧義字串」。例如：「不可以」三個中文字元所組成的字串，辭典中的詞含有「不、不可、可以」，「不可以」所組成的字串，在下列句子中，因其上下文的不同而產生不同的斷詞結果：「不／可以／忘記」、「不可／以／營利／為／目的」。

- 組合型歧義（covering ambiguity）

  令 x, y 代表中文字元所組成的字串，若 x、y、xy 都是辭典中的詞，xy 的組合中，可在不同的文章中，分別被斷詞成 xy 或 x/y，因為詞 xy 是由 x 與 y 等兩個不同的詞所組成，因此 xy 稱為「組合型歧義字串」。例如：「才能」二個字所組成的字串，辭典中的詞有「才、能、才能」，在下列句子中「才能」組成的字串，將產生不同的斷詞結果：「他／才能／非凡」、「只有／他／才／能／勝任」。

另外，「未知詞」則指辭典中未收錄的詞，包含了人名、地名、組織名、人名地名組織名之縮寫、衍生詞、複合詞、數字型態等，由於人類所使用的語言會隨著社會不斷改變，而持續地創造出新的用語，並且詞的衍生現象也非常地普遍，因此新詞會不斷的出現，辭典永遠無法因應新詞產生的速度，所以會出現未知詞問題，斷詞系統必須能夠處理未知詞，才可提高斷詞的正確性。

近年來的斷詞系統傾向於機器學習式（machine learning-based）演算法來解決中文斷詞的問題，例如 Maximum Entropy (ME) [22]、Support Vector Machine (SVM) [2, 6]、Transformation-Based Learning Algorithm (TBL) [11]、Hidden Markov Model (HMM) [2, 11, 23, 25] 等等，並且顯示了使用機器學習式演算法做中文斷詞，確實可以達到很高的斷詞準確率。

本研究使用隱藏式馬可夫模型（Hidden Markov Model, HMM）來解決中文斷詞的問題。雖然已有數篇研究同樣使用隱藏式馬可夫模型來處理斷詞問題 [2, 11, 23, 25]，但使用傳統的作法，隱藏式馬可夫模型在解決中文斷詞的問題上，無法達到較好的斷詞效能（F-measure 約 80%），因此這些研究 [2, 11, 23] 便結合了其他機器學習演算法，以增加斷詞的效能。我們的研究目的是希望只使用隱藏式馬可夫模型當成主要的演算法，並且應用「特製化」（specialization）的概念來提升隱藏式馬可夫模型的準確率。我們的作法是給予隱藏式馬可夫模型更多的資訊，在完全不修改模型之訓練及測試過程的前提下，透過兩階段特製化的方式，分別為擴充「觀測符號」，以及擴充「狀態符號」的方式，大大地改善了隱藏式馬可夫模型的斷詞準確性。

於第一階段中，為了擴充觀測符號，我們使用最簡單也最常被使用的辭典比對式斷詞演算法－「長詞優先法」（maximum matching algorithm），來增加額外的資訊於隱藏式馬可夫模型中，使得模型擁有更多的斷詞資訊做學習。第二階段擴充狀態符號的方式，我們則使用詞彙式隱藏式馬可夫模型（Lexicalized HMM）的概念，也就是只根據某些特製詞（specialized words）來做特製化，將狀態做延伸，來提升系統斷詞的效能。

2. 相關研究

中文斷詞的研究已有相當歷史，但在近幾年仍陸續新的方法提出，底下我們分別就解決歧義性及未知詞兩個問題分別做文獻回顧。

首先就斷詞歧義性問題，M. Li 等人 [9] 於 2003 年的研究中，提出一種非監督式（unsupervised）訓練的方法，藉由訓練 Naïve Bayes 分類器，來解決中文斷詞的交集型歧義問題，實驗結果可達到 94.13% 的準確率。另一方面，解決組合型歧義比解決交集型歧義更加困難，主要的原因是，要解決組合型歧義則需要依賴更多的內文資訊，如句法分析（syntactic）、語意分析（semantic） 以及前因後果的資訊（pragmatic information）等，才能正確的解決這類的歧義問題。1999 年 J. H. Zheng 等人 [26] 使用規則式（rule-based method）的作法來處理組合型歧義，並達到 85 % 的準確率。而 2002 年 X. Luo 等人 [12] 的研究，則是使用類似於自然語言處理領域中解決「詞義消歧」（word sense disambiguation）的問題，來解決組合型歧義問題，該篇研究使用 TF.IDF 權重計算的公式，重新定義新的 TF 與 IDF 的公式，以此方式來解決組合型歧義問題，達到 96.58 % 的準確率。

解決未知詞問題是做中文斷詞的另一個重要步驟，近年來也有數篇研究再處理未知詞問題。中研院陳克建博士（Chen）等人於 1997 年開始，提出了三篇關於解決未知詞問題的研究 [3, 5, 13]，最早於 1997 的研究 [3]，透過統計斷詞語料庫，產生所有單一字元之已知詞的偵測規則。此階段的研究只能偵測出所有的單一字元的結果，並未真正將未知詞擷取出來。2002 年的研究 [5] ，則是使用人工加上一些統計的方法來建立擷取規則，將所有被偵測出屬於未知詞部分的單一字詞，透過擷取規則以合併這些單一字詞而成為未知詞。實驗中測試 1,160 個未知詞，結果達到 89 % 的擷取準確率。另外於 2003 年的研究 [13] 中，同樣做擷取未知詞的研究，該研究中將所有種類的未知詞的構詞方式以 context free grammar 表示出來，並搭配 bottom-up merging algorithm 來解決大部分統計特性低的未知詞擷取問題。實驗效能達到 75 % 的擷取準確率。

其他解決未知詞問題的研究，如 Zhang 等人 [24] 於 2002 年的研究，則使用類似詞性標示（part-of-speech tagging）的作法，稱為「角色標示」（roles tagging），角色指的是在未知詞的組成成分、上下文以及句子中的其他部分，並且依據句子的角色序列來辨識出未知詞。實驗部分針對中國人名以及外國翻譯名等未知詞做測試，並且達到不錯的準確率以及召回率。

近年來的研究主要趨向於機器學習式的方法來處理中文斷詞，例如 Maximum Entropy（ME）[22] 以及 Conditional Random Field（CRF）[20] 等，這些統計式的學習演算法都是轉成字元分類問題（character classification）來處理中文斷詞問題，並且使用了數種類似的特徵，如目前字元、加上前後各一字元、加上前後各兩字元等，來當作模型的屬性。而 C. L. Goh 等人則使用 Support Vector Machine（SVM）[6] 來解決中文斷詞的問題，該篇研究結合辭典比對式方法－長詞優先法，利用長詞優先法的歧義性以及未知詞的資訊，來加強 SVM 的特徵屬性以改善斷詞效能。另外也有使用感知機（Perceptron）[10] 的方法做斷詞，該篇研究認為 Perceptron 方法雖然與 SVM 類似，不過效能卻較 SVM 差一些，但由於其訓練的速度非常快，因此他們系統提出的主要貢獻就是一個速度快且效能不至於差太多的斷詞方法。

有些研究為了加強學習演算法的斷詞效能，則是結合了數個學習模型，採用混合式作法來處理斷詞問題，如 M. Asahara 等人 [1] 以及 N. Xue 等人 [23] 的研究，為了加強 ME 斷詞結果，這兩篇研究則結合了 SVM、CRF [1] 以及 TBL [23] 等作法，使用混合式方法的結果提升斷詞準確率。另外，許多研究也使用隱藏式馬可夫模型來處理斷詞問題。如 HHMM [25] 系統，便使用了五層的隱藏馬可夫模型，根據不同的目的各自訓練出各個模型，最後再整合成斷詞系統。而 HMM+SVM [2]、HMM+TBL [11] 等兩篇研究，則使用隱藏式馬可夫模型的斷詞結果當成是一個屬性，並分別使用 SVM 以及 TBL 來當成主要的演算法做斷詞，以達到較佳的斷詞結果。此兩篇研究於實驗中也列出只使用隱藏式馬可夫模型做斷詞的效能，其 F-measure 的結果分別為 80.4 % 以及 81.4 %。因此，我們發現

隱藏式馬可夫模型需仰賴其他外部資源或是結合其他的學習演算法，才可以達到可接受的斷詞效能。

3. 系統架構

我們提出的系統是以隱藏式馬可夫模型來解決中文斷詞的問題，並且透過兩階段「特製化」（specialization）的方式來加強隱藏式馬可夫模型的斷詞效能。第一階段特製化，我們結合了長詞優先法的結果來增加觀測符號的資訊，以「擴充觀測符號」；第二階段特製化，則是透過詞彙式（Lexicalized HMM）的特製化過程，以「擴充狀態符號」。

因此我們的系統架構主要可分為兩個部分，第一部份：我們稱之為「M-HMM」，也就是結合長詞優先法於隱藏式馬可夫模型中，讓訓練之模型增加斷詞歧義性與未知詞的資訊，藉此以改善隱藏式馬可夫模型處理中文斷詞的正確性；第二部分：我們稱之為「Lexicalized M-HMM」，這部分透過兩種不同的準則（criteria）來決定特製詞（specialized words），並以屬於特製詞之觀測符號做特製化，透過擴充狀態符號而再次加強斷詞準確率。

3.1. 長詞優先法

長詞優先法（Maximum Matching Algorithm, MM）是最簡單也最廣泛使用的辭典比對式的斷詞方法，其斷詞的策略為由句子的一端開始，試著比對出在辭典中最長的詞，當作斷詞結果，接著去除此詞後，剩下的部分繼續做長詞優先法斷詞，直到句子的另一端結束為止。一般來說，如果所使用的辭典夠大，長詞優先法斷詞可達到超過 90 % 以上的斷詞準確率。

長詞優先法依照比對方向的不同又可分為兩種不同的變形，第一種是「正向長詞優先法」（Forward Maximum Matching, FMM），即由句子開頭的第一個字元開始，由左而右逐一掃瞄，比對出在辭典中最長的詞，以當作斷詞的結果，並直到句子的結尾而結束。相反地，另一種長詞優先法的變形則是「反向長詞優先法」（Backward Maximum Matching, BMM），由句子的最後一個字元開始掃瞄，從右至左依序比對辭典中的詞，比對到最長的詞當成反向長詞優先法的斷詞結果，並

直到句子的開頭而結束。

此兩種不同的長詞優先斷詞法，當斷詞的結果不同時，則表示發生交集型歧義，如表 1 中的第二個例子：「即將來臨時」字串，因為「將」可與「即」和「來」結合成｛即將、將來｝等不同的詞，因此屬於交集型歧義字串，正向長詞優先法會斷詞成「即將／來臨／時」，而反向長詞優先則斷詞成「即／將來／臨時」。

表 1　長詞優先法的不同變形

例句	**正向長詞優先**	**反向長詞優先**
即將畢業	即將／畢業	即將／畢業
即將來臨時	即將／來臨／時	即／將來／臨時

另外，由於長詞優先法屬於辭典比對式斷詞方法，只有在辭典中的詞才有可能正確斷出，所以無法解決未知詞問題。當遇到未知詞時，正向長詞優先與反向長詞優先都將斷詞成單一中文字元。例如：「鴻海董事長郭台銘」字串，由於辭典中未收錄｛鴻海、郭台銘｝等詞，因此正向長詞優先法與反向長詞優先法都同樣會斷詞成「鴻／海／董事長／郭／台／銘」。

3.2. BIES 分類問題

利用機器學習式演算法來解中文斷詞的問題時，一般的作法是將中文斷詞問題轉換成分類的問題，而最常被使用的方法就是轉換成字元分類問題（character classification problem），將每個字元都給予其對應的類別，透過字元類別來做分類，這些字元的類別由出現在中文詞當中的特定位置來決定，一個字元的位置可以分為位於詞的開始（beginning）、位於詞的中間 （intermediate）、位於詞的結尾（end）以及由單一字元組成的詞（single-character）等四種類別，因此也稱為「BIES 分類問題」。

理論上中文字元可以存在於中文詞的任何位置上，例如表 2 的例子，字元「中」可以存在於詞的開始（B）、詞的中間（I）、詞的結尾（E）、以及單一字元的詞（S）。所以 BIES 分類所要解決的問題也就是決定每個字元的正確類別。

在中文斷詞的問題上，一旦將欲斷詞字串中的所有字元都已分類完成，則也表示已經斷詞完成，例如：「今天是重要的日子」這個中文字串，利用分類問題將找出每個字元所對應的 BIES 標籤，在此例子中，也就是「BESBESBE」，則相當於是已經斷詞出｛今天、是、重要、的、日子｝等詞出來了，因此原來的中文字串便可以轉換成「今天／是／重要／的／日子」的斷詞結果。

表 2　字元「中」可出現在詞的任何位置

B	中醫
I	國民中學
E	集中
S	在　資料庫　中

3.3. 隱藏式馬可夫模型

隱藏式馬可夫模型可以視為一個雙層的隨機序列，包含了隱藏層的狀態序列（state sequence）和可觀察層的觀測序列（observation sequence）。隱藏層是無法直接觀察得到的，但可以從另一個可觀察的觀測序列之隨機過程的集合觀察得出。因此，隱藏式馬可夫模型是一個馬可夫鏈的機率函數，無法直接觀察的隱藏層就是一個有限狀態的馬可夫鏈，其初始的狀態機率分佈以及狀態之間的轉移機率由狀態初始機率向量 $\prod$ 和狀態轉移機率矩陣 A 來決定，另外還需定義觀測符號機率矩陣 B ，儲存各個觀測符號在不同的狀態下的機率值。

隱藏式馬可夫模型可由 (S, K, N, M, Π, A, B) 等七個元素來表示，底下針對模型相關符號與參數做說明：

- S 表示所有狀態的集合，$S=\{s_1,s_2,...,s_N\}$。
- K 表示所有觀測符號的集合，$K=\{k_1,k_2,...,k_M\}$。
- N 表示模型中所有狀態的個數。
- M 表示模型中所有觀測符號的數目。
- $\Pi=(\pi_i)$ 代表狀態初始的機率向量，$\pi_i=P(q_1=s_i)$，$1\le i\le N$，表示在 $t=1$ 時，狀態為 s_i 的機率，且需滿足 $\sum\pi_i=1$ 的條件。

- $A=[a_{ij}]$ 代表狀態轉移機率矩陣，$a_{ij}=P(q_{t+1}=s_j \mid q_t=s_i)$，$1\le i,j\le N$，表示從狀態 s_i 到狀態 s_j 的機率，且滿足 $a_{ij}\ge 0$ 和 $\sum_{j=1}^{N} a_{ij}=1$

- $B=[b_j(k)]$ 代表觀測符號矩陣，$b_j(k)=P(o_t=v_k \mid q_t=s_j)$，$1\le j\le N$ 和 $1\le k\le M$，表示在狀態為 s_j 時，觀測符號為 v_k 的機率，且滿足 $\sum_{k=1}^{K} b_j(k)=1$。

給定輸入之觀察序列 $O=o_1o_2\cdots o_n$（o_t 表示在時間 t 所對應的觀測符號，且滿足 $o_t\in K$）。隱藏式馬可夫模型的目的就是要選出一個對應於觀測序列之最佳的狀態序列 $Q=q_1q_2\cdots q_n$（q_t 表示在時間 t 所對應的狀態，且滿足 $q_t\in S$），也就是找出 $P(Q_1^n \mid O_1^n)$ 為最大機率值時的狀態序列。

由於在馬可夫基本假設下，第 $t+1$ 的時間狀態只和第 t 的時間狀態有關，與其他任何以前的時間狀態無關，即 $P\{q_{t+1}=s_k \mid q_1,q_2,...,q_t\}=P\{q_{t+1}=s_k \mid q_t\}$，且隨機過程中的機率轉移不隨時間改變，因此 $P(Q_1^n \mid O_1^n)$ 的計算可簡化成：

$$P(Q_1^n,O_1^n)=\prod_{t=1}^{n} P(q_t \mid q_{t-1})P(o_t \mid q_t)=\pi_{q_1}\prod_{t=1}^{n-1} A_{q_t,q_{t+1}}\prod_{t=1}^{n} B_{q_t}(o_t)$$

而取得此最大值的狀態序列 Q_1^n，則是使用維特比（Viterbi）演算法計算得到。

另外於訓練過程中，隱藏式馬可夫模型當初所提出來的方法 [19] 是使用非監督式的學習方法（unsupervised approach）做訓練，也就是從未標示狀態的文件中做訓練，因而稱之為「隱藏式」，訓練的方法則是使用 Baum-Welch 演算法做參數的更新。而近年來許多領域都已發展出大量已標示的語料庫（corpus）可供訓練，隱藏式馬可夫模型同樣可以在已標示狀態的文件中來做監督式（supervised approach）訓練 [14]，訓練過程則直接利用最大概似估計法（maximum likelihood estimation）計算出模型參數則此模型，又可稱為「可見式馬可夫模型」（Visible Markov Model, VMM）或「語言模型」（Language Model）等，但絕大部分的研究仍然稱「隱藏式」馬可夫模型。於我們的系統中，我們使用監督式的方法來訓練模型，在本論文中也直接以「隱藏式馬可夫模型」做系統的說明。

3.4. 特製隱藏式馬可夫模型

隱藏式馬可夫模型的特製化（specialization）概念，最早是由 J. D. Kim 等人於 1999 年與 2000 年等兩篇研究 [7, 8] 所提出來的，之後於 2001 年到 2004 年間，A. Molina & F. Pla 等兩位學者，更是將此概念成功的應用到許多不同的領域上，如詞性標示（part-of-speech tagging）[17, 18]、淺層分析（shallow parsing）[15]、詞義消歧（word sense disambiguation）[16] 等問題上。

特製化的過程是指在不修改隱藏式馬可夫模型的訓練以及測試過程的前提下，透過狀態的延伸使得模型增加更多資訊，以提升模型準確率。其主要的作法就是給予一個特製化函式（specialization function），以產生出新的狀態，特製化的過程以底下式子來說明：

$$f(\langle o_i, q_i \rangle) = \langle o_i, q_i \cdot o_i \rangle$$

$\langle o_i, q_i \rangle$代表某個觀測符號以及其對應的狀態，新的狀態符號經過特製化的過程中，由原來的觀測符號加上原來狀態來產生，此特製化的隱藏式馬可夫模型又稱為「特製隱藏式馬可夫模型」（Specialized HMM）。而如果不將所有的觀測符號所對應的狀態都做特製化，而是只在特定的觀測符號下，才做特製化的過程則稱為「詞彙式的隱藏式馬可夫模型」（Lexicalized HMM），此過程屬於特製化過程的一種特例，又被稱為詞彙化（lexicalization），此過程以底下式子來說明：

$$f(\langle o_i, q_i \rangle) = \begin{cases} \langle o_i, q_i . o_i \rangle & \text{if } o_i \in W \\ \langle o_i, q_i \rangle & \text{if } o_i \notin W \end{cases}$$

其中 W 為特製詞（specialized words），只有屬於特製詞的觀測符號才會做特製化處理，而特製詞的選擇又有許多不同的準則來選取。

3.5. M-HMM

在 BIES 分類問題中，由於一個字元可出現在詞的不同位置，而導至所對應的 BIES 標籤不只一個，一旦類別標示錯誤，連帶會使得斷詞結果錯誤。但此種斷詞歧義性在 HMM 模式下，並無特殊處理方式。由於正向長詞優先與反向長詞

優先在做斷詞時，遇到歧義性的句子會產生不同的斷詞結果，因此如能將正向長詞優先與反向長詞優先的資訊同時加入 HMM 模型中，相當於提供歧義性的資訊，並且長詞優先法屬於辭典比對式斷詞法，雖無法直接提供未知詞的資訊，但可間接的調整辭典大小來反應未知詞多寡。

將隱藏式馬可夫模型（HMM）改成 M-HMM 的過程，主要是將正向長詞優先（FMM）與反向長詞優先（BMM）之斷詞結果(即所得的 BIES 標籤)，與原來的「字元」組成的新的觀測符號，延伸為「字元-FMM-BMM」等三個資訊結合而成的觀測序列。表 3 中以一個例子來針對 M-HMM 訓練以及測試過程做個說明，在訓練階段中，原始的觀測符號序列為「研、究、生、命、起、源」，加入了長詞優先法的資訊後，新的觀測符號序列便被轉換成「研-B-B、究-I-E、生-E-B、命-S-E、起-B-B、源-E-E」。這些中文字元旁的 B、I、E、S 標籤即是由正向長詞優先與反向長詞優先法所標示的，因此新的觀測符號種類相當於增加了 16 倍，在此狀態種類並未做改變。

表 3　M-HMM 的例子

	訓練過程		測試過程	
原始句子	研究／生命／起源		結合成分子	
	觀測序列	狀態	觀測序列	狀態
HMM 訓練測試資料	研-B-B	B	結-B-S	?
	究-I-E	E	合-I-B	?
	生-E-B	B	成-E-E	?
	命-S-E	E	分-B-B	?
	起-B-B	B	子-E-E	?
	源-E-E	E		

3.6. Lexicalized M-HMM

這部分為隱藏式馬可夫模型特製化的第二階段，透過第一階段 M-HMM 的過程，將觀測符號延伸之後，此階段以新的觀測符號來做詞彙化，也就是取特定的觀測符號當成特製詞，來做詞彙式(Lexicalized)的特製化過程。此階段的特製化過程描述如下。對於每一個特製詞中的觀測符號 w_i 及其對應狀態為 s_i，則詞彙

化的過程是新增一個狀態「$s_i \cdot w_i$」，而原本的狀態「s_i」仍由其他觀測符號所擁有。此過程也相當於是將訓練資料中屬於特製詞的觀測符號給予新的類別，而使得新的訓練資料不再只有原來的 B、I、E、S 四個類別。

我們以一個例子來做說明，如表 4，假若觀測符號「生-E-B」、「起-B-B」是屬於特製詞，則經過詞彙化的過程之後，觀測符號 「生-E-B」 以及 「起-B-B」所對應的狀態就被轉換成「B-生-E-B」、「B-起-B-B」了。在觀測符號 「生-E-B」中，原來的狀態 B 便被分割成兩個不同的狀態：一個是由觀測符號「生-E-B」所屬的狀態「B-生-E-B」以及其他未分割的觀測符號（如觀測符號「研-B-B」）之狀態「B」。因此在新的訓練資料中，狀態符號被延伸了。

表 4　特製詞集合｛生-E-B, 起-B-B｝做詞彙化產生新的狀態

觀測符號	原來的狀態	新的狀態
研-B-B	B	B
究-I-E	E	E
生-E-B	B	B-生-E-B
命-S-E	E	E
起-B-B	B	B-起-B-B
源-E-E	E	E

此特製化過程也將牽扯到一個問題：由於隱藏式馬可夫模型的三個主要參數都與「狀態符號」有關，因此這階段的特製化過程，將增加隱藏式馬可夫模型的參數大小，因此計算量也就會跟著增加，而且過多的特製詞不見得能一直提升準確率。所以我們必須根據訓練資料來決定特製詞的大小。

特製詞的選擇方式，我們是使用兩種不同的準則（criteria）來選取，說明如下：

- SWF: (the Words with High Frequency)

  取在訓練資料中屬於最高頻率的觀測符號，當成特製詞。

- SEF: (the Words with Tagging Error Frequency)

  取具有高測試錯誤率 (或稱標示錯誤率) 的詞，當成特製詞。

不論是使用 SWF 或是 SEF 準則來選取特製詞，都需要決定一個門檻值（threshold），此門檻值是決定特製詞的大小，我們會於實驗四中找出最佳斷詞效能的門檻值。

4. 實驗

於系統實驗中，我們使用中研院平衡語料庫第 3.1 版，當成我們實驗的資料。此語料庫，共有 575 萬詞，是第一個已斷好的詞並帶有詞類標記的現代漢語語料庫。我們將其中已斷詞的中文文章來當成我們的實驗對象，並用隨機的方式分成兩個部分，取其中的 80% 當作訓練語料，用來訓練隱藏式馬可夫模型。而剩下的 20% 則當成我們系統的測試語料。斷詞的評估方式則是使用準確率（Precision）、召回率（Recall）以及 F-measure 來驗證斷詞效能，分別定義如下：

$$\text{Precision} = \frac{\text{系統正確斷出的詞數}}{\text{系統斷詞的總詞數}}$$

$$\text{Recall} = \frac{\text{系統正確斷出的詞數}}{\text{真正的詞數}}$$

$$\text{F measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision} + \text{Recall}}$$

由於我們的系統分成 M-HMM 與 Lexicalized HMM 兩部分，因此在實驗的部分，我們也由此兩部分來做實驗。M-HMM 實驗的部份為實驗一、二、三；而 Lexicalized HMM 實驗的部分則為實驗三與實驗四。

4.1. M-HMM 實驗（實驗一、實驗二）

M-HMM 實驗的部分，主要驗證隱藏式馬可夫模型結合長詞優先法之後，在觀測符號中加入更多資訊之前與加入之後的斷詞效能的比較。由於長詞優先法可以提供斷詞歧義性與未知詞等資訊，因此這部分的實驗，我們是先驗證歧義性的斷詞效能，再驗證未知詞資訊多寡之斷詞效能的比較。

實驗一驗證斷詞歧義性效能，方法是取所有訓練資料與測試資料中的所有詞，來當成長詞優先法所使用的辭典的詞（共有 145,608 個詞），使得在測試過程中不會出現未知詞。表 5 為 M-HMM 解歧義性的斷詞效能。其中實驗的基線（baseline）作法為正向長詞優先法（FMM）、反向長詞優先法（BMM），以及只

使用字元資訊當成觀測符號的隱藏式馬可夫模型（HMM）。除了 M-HMM 的實驗結果之外，同時我們也比較只結合正向長詞優先法資訊（FMM+HMM）以及只結合反向長詞優先法資訊（BMM+HMM）的隱藏式馬可夫模型之斷詞效能。

表 5　實驗一：M-HMM 解歧義性的斷詞效能

	FMM	BMM	HMM	FMM+HMM	BMM+HMM	**FMM+BMM+HMM (M-HMM)**
Recall	0.936	0.939	0.812	0.944	0.947	0.957
Precision	0.956	0.959	0.811	0.962	0.965	0.976
F-measure	0.946	0.949	0.812	0.953	0.956	0.967

實驗顯示，隱藏式馬可夫模型只使用字元的資訊時，其斷詞結果只有 0.81 左右，而加入正向長詞優先法與反向長詞優先法之後，系統的斷詞效能 F-measure 由 0.812 大幅地提升到 0.967，並且斷詞結果也勝過正向長詞優先法與反向長詞優先法等兩種基線作法。因此，實驗結果證明了長詞優先法所提供之歧義性資訊的確可提升隱藏式馬可夫模型的效能。

實驗二主要是驗證長詞優先法所使用的辭典，對 M-HMM 斷詞系統的影響，也就是實驗未知詞的斷詞效能。由於辭典是由訓練資料產生，因此實驗時我們將訓練資料隨機分割成兩部分：訓練集合 1（set 1）以及訓練集合 2（set 2），辭典只由訓練集合 1 來產生，藉由調整訓練資料不同的分割比例，以產生出不同的辭典數量，在相同的測試資料下以驗證各自的斷詞效能。實驗結果如表 6 所示。

表 7　實驗二：M-HMM 解未知詞的斷詞效能

	不含未知詞	**未知詞（實驗二）**				HMM
訓練資料比例 (Set1/Set2)	100/0	80/0	60/20	40/40	20/60	0/80
辭典中的詞數	145,608	132,273	116,428	96,780	69,446	0
Set2 中的未知詞數	0	0	17,418	45,212	103,990	All
測試資料中的未知詞數	0	14,415	17,323	22,524	34,573	All
Recall	0.957	0.946	0.946	0.944	0.941	0.812
Precision	0.976	0.951	0.949	0.945	0.934	0.811
F-measure	0.967	**0.948**	0.948	0.945	0.937	0.812

在此實驗的第一個部分，分割比例為 100/0，相當於實驗一歧義性的效能，而未知詞實驗的部分，共實驗 80/0、60/20、40/40、20/60 等分割比例的結果，由未知詞所佔的比例之不同來驗證斷詞效能，而最後一個部分，分割比例為 0/80，代表完全不從訓練資料中建立辭典，也就是測試資料中所有的詞都屬於未知詞，並且在訓練的過程中完全沒有從正向長詞優先法或反向長詞優先法中得到任何資訊，只依賴字元的資訊做斷詞。

實驗二的結果可得知，隨著增加未知詞的資訊，也就是在減少字典的詞數的情況下，M-HMM 的斷詞效能跟著減低，但是降低的幅度並不大，顯見只要有基本詞彙，即可提升 HMM 斷詞效能，但對於未知詞問題，並不能有所做為，因此我們將於實驗三設計 Mask 的實驗來解決此一問題。

4.2. Mask 實驗（實驗三）

由於實驗二是透過減少訓練資料中的詞，來建立長詞優先法所需之辭典的方法以提供未知詞資訊，但是犧牲了長詞優先法的正確性。因此我們引用 Mask 的作法 [21]，在不犧牲訓練資料的詞的前提下，產生具有未知詞資訊的訓練資料。Mask 的概念是讓訓練過程中也有機會碰到未知詞，也就是仿造測試時真正的情形，其作法如下：

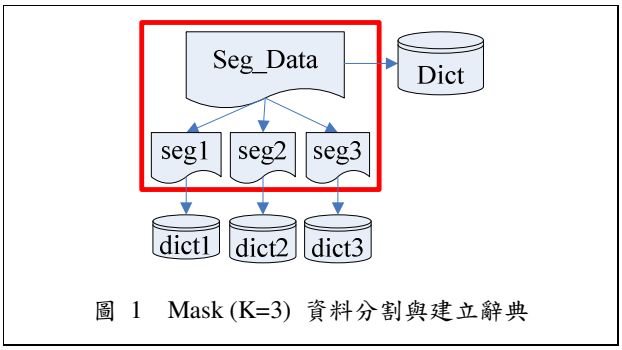


圖 1 Mask (K=3) 資料分割與建立辭典

首先將訓練資料分割成 K 個部分，並且每個部分都建立各自的辭典，因此可產生 K+1 個辭典，如圖 1 所示，此 K+1 個資料便可建立 K+1 個訓練資料。我們先以所有辭典的聯集(Dict=dict1＋dict2+dict3)來標示 M-HMM 所需要的觀測符號，這也相當是原始的訓練資料。接著每次遮住一個部份辭典，也就是產生一個

較小的辭典 Dict-dict(i)，來標示 M-HMM 所需要的觀測符號。在這過程中，有些字詞會因為未知詞的關係，會被錯標成單一字詞 S，但其狀態符號，可以讓 HMM 知道正確的標籤；如果標示結果與原來相同時，則可直接省略，以避免在一個狀態所見到的觀測符號機率不公平的增加，如此重複 K 次最後將此 K+1 個資料形成整個 Mask 的訓練資料。

實驗三為使用 Mask 方法所做的 M-HMM 的實驗，取 Mask K=2 至 K=10 來驗證結果，而 K=1 表示不做分割，也就是沒有使用 Mask 的結果，實驗如圖 2 所示。實驗三結果顯示，使用 Mask 的方法可提供隱藏式馬可夫模型更多未知詞資訊，使得斷詞效能有所提升，並且在 K=2 時，達到最佳的斷詞效能（F-measure = 95.25%）。

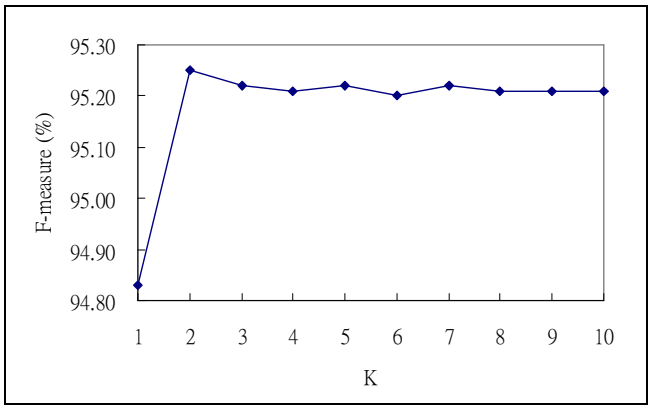


圖 2　實驗三：Mask K=1 至 K=10 的實驗結果

4.3. Lexicalized M-HMM 實驗（實驗四、五）

實驗四是根據 Lexicalized M-HMM 的 SWF 與 SEF 兩種不同的詞彙化策略下，用來調整各自使用的特製詞大小，以找出使得模型能有最佳斷詞效能的門檻值（threshold）。由於這個實驗是用來調整系統用到的特製詞，而不是做斷詞效能的實驗，因此我們只取「訓練資料」來做此實驗。我們將全部訓練資料（佔全部資料 80%）分割成兩部分，依 7 比 1 的比例來分割（分別佔全部資料的 70% 與 10%），其中 70% 的資料（轉換成具有長詞優先法資訊的資料）用來訓練 M-HMM 模型，而剩下的 10% 則當成驗證效能的調整資料（tuning data）。

由於 SWF 為取訓練資料中出現頻率最高的詞當成特製詞，因此我們統計 70%

的資料，取出高頻率的詞做特製詞。而 SEF 為取高測試錯誤率的詞當成特製詞，因此我們先從 70% 的資料建立 M-HMM 模型，並且於調整資料中做測試，根據調整資料中高測試錯誤率的詞做特製詞。取得 SWF 與 SEF 之特製詞後，接著驗證在不同的門檻值下，調整資料的斷詞效能。實驗數據如圖 3 所示。

實驗結果顯示，我們使用 SWF 與 SEF 兩種不同的詞彙化策略，在剛開始取較少的詞當特製詞時，兩者在調整資料下的斷詞效能都有顯著的上升，而 SWF 在取 292 個詞（出現頻率大於 4800 次）時，SEF 取 173 個詞（出現頻率大於 25 次）時，斷詞效能達到最佳結果，並且再繼續隨著特製詞數的增加，斷詞結果便開始往下降，這是因為狀態數增加，使得模型計算量增加而導致準確率下降之緣故。

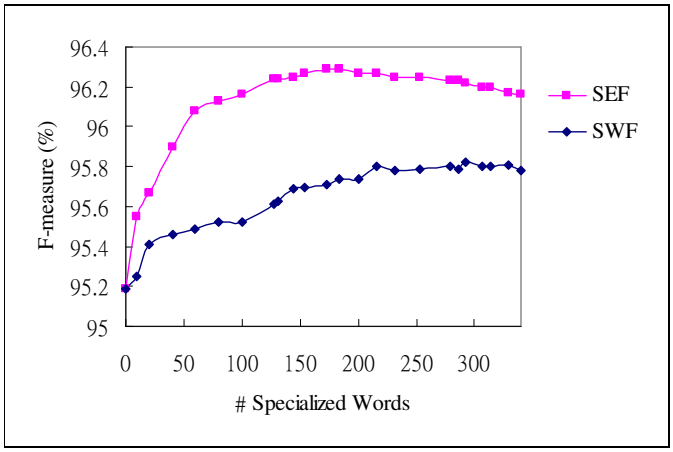


圖 3 實驗四：在不同特製詞大小下，SEF 與 SWF 準則在調整資料下的斷詞效能

實驗五則是測試最佳特製詞結果的 SWF 以及 SEF 準則之 Lexicalized M-HMM 斷詞效能，實驗的設定使用 Mask K=2 之 M-HMM 的設定以及最佳 SWF 與 SEF 特製詞（SWF 為取 292 個詞作為特製詞，而 SEF 則取 173 詞作為特製詞）來做此實驗，並且也與正向長詞優先法（FMM）、反向長詞優先法（BMM）、只使用字元資訊之隱藏式馬可夫模型（HMM）等基線斷詞作法及 M-HMM 的結果作比較，以驗證本系統在狀態延伸前與延伸後的斷詞效能作比較。實驗結果如表 8 所示。實驗結果顯示 Lexicalized M-HMM 不論使用 SWF 或 SEF 準則，其斷詞結果都比 M-HMM 的斷詞效能較好，F-measure 由 0.953 提升到 0.960 與 0.963，

而且使用 SEF 準則與使用 SWF 準則相較之下，SEF 不但特製詞較少且斷詞效能也較好。

表 8　實驗五：Lexicalized M-HMM 的斷詞效能

	FMM	BMM	HMM	M-HMM	**SWF M-HMM**	**SEF M-HMM**
Recall	0.925	0.928	0.812	0.947	0.958	0.963
Precision	0.928	0.930	0.811	0.958	0.962	0.964
F-measure	0.926	0.929	0.812	0.953	**0.960**	**0.963**

5. 結論

在本篇論文中，我們應用隱藏式馬可夫模型之特製化的概念來提升中文斷詞的效能，我們系統的最大的優點，就是完全不需要對隱藏式馬可夫模型的訓練過程以及測試過程做任何修改，只需將訓練資料根據特製化函式來做轉換即可。我們使用兩階段的特製化過程逐步的改良隱藏式馬可夫模型的斷詞效能，在第一階段中結合了長詞優先法的資訊，使得觀測符號增加更多的資訊，於實驗結果顯示，結合長詞優先法在沒有未知詞的情況下，可以大幅地提升隱藏式馬可夫模型的斷詞效能（F-measure: 0.812→0.967），而在有未知詞的情況下，利用 Mask 方式也些微改善斷詞效能（F-measure: 0.948→0.953）。而第二階段使用詞彙式的特製化方式，挑選高錯誤的字元使得狀態增加，實驗也證明能再次提升斷詞效能（F-measure: 0.953→0.963），實驗中發現使用 SEF 準則的結果會比 SWF 準則不但使用的特製詞較小且又能達到更好的斷詞結果。

誌謝

本研究由國科會編號 NSC 94－2213－E－008－020 贊助。

參考文獻

1. M. Asahara, K. Fukuoka, A. Azuma, C. L. Goh, Y. Watanabe, Y. Matsumoto, T. Tsuzuki. “Combination of Machine Learning Methods for Optimum Chinese Word Segmentation,” *In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, pp. 134–137, 2005

2. M. Asahara, C. L. Goh, X. Wang and Y. Matsumoto. “Combining Segmenter and Chunker for Chinese Word Segmentation,” *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp. 144–147, 2003
3. K. J. Chen and M. H. Bai. “Unknown Word Detection for Chinese By a Corpus-based Learning Method,” *In Proceedings of ROCLING X*, pp. 159–174, 1997
4. K. J. Chen and S. H. Liu. “Word Identification for Mandarin Chinese Sentences,” *Proceedings COLING '92*, pp. 101-105, 1992
5. K. J. Chen and W. Y. Ma. “Unknown Word Extraction for Chinese Documents,” *In Proceedings of COLING 2002*, pp. 169–175, 2002
6. C. L. Goh, M. Asahara and Y. Matsumoto. “Chinese Word Segmentation by Classification of Characters,” International Journal of Computational Linguistics and Chinese Language Processing Vol. 10, No. 3, pp. 381-396, 2005
7. J. D. Kim, S. Z. Lee and H. C. Rim. “HMM Specialization with Selective Lexicalization,” *In Proceedings of the joinSIGDAT Conference on Empirical Methods in Natural Lan-guage Processing and Very Large Corpora(EMNLP-VLC-99)*, pp. 121-127, 1999
8. S. Z. Lee, J. I. Tsujii and H. C. Rim. “Lexicalized Hidden Markov Models for Part-of-Speech Tagging,” *In Proceedings of 18th International Conference on Computa-tional Linguistics, Saarbrucken, Germany*, pp.481-787, 2000
9. M. Li, J. F. Gao, C. N. Huang and J. F. Li. “Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Word Segmentation,” *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp. 1–7, 2003
10. Y. Y. Li, C. J. Miao, K. Bontcheva and H. Cunningham. “Perceptron Learning for Chinese Word Segmentation,” *In Proceedings of Fourth SIGHAN Workshop on*

Chinese Language Processing, pp. 154–157, 2005

11. X. Lu. "Towards a Hybrid Model for Chinese Word Segmentation," *In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, pp. 189–192, 2005
12. X. Luo, M. Sun and B. K. Tsou. "Covering Ambiguity Resolution in Chinese Word Segmentation Based on Contextual Information," *In Proceedings of COLING 2002*, pp. 598-604, 2002
13. W. Y. Ma and K. J. Chen. "A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction," *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp. 31–38, 2003
14. C. D. Manning and H. Schutze. "Foundation of Statistical Natural Language Processing," Chapter 9-10. pp. 317-380, 1999
15. A. Molina and F. Pla. "Shallow Parsing using Specialized HMMs," Journal of Machine Learning Research 2, pp. 595–613, 2002
16. A. Molina, F. Pla and E. Segarra. "A Hidden Markov Model Approach to Word Sense Disambiguation," *In Proceedings of the VIII Conferencia Iberoamericana de Inteligencia Artificial, IBERAMIA 2002*, pp. 1-9, 2002
17. F. Pla and A. Molina. "Improving Part-of-Speech Tagging using Lexicalized HMMs," Natural Language Engineering, pp. 167-189, 2004
18. F. Pla and A. Molina. "Part-of-Speech Tagging with Lexicalized HMM," *In proceedings of International Conference on Recent Advances in Natural Language Processing(RANLP2001)*, 2001
19. L. R. Rabiner. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE, Vol.77, No.22*, pp. 257-286, 1989
20. H. H. Tseng, P. H. Chang, G. Andrew, D. Jurafsky, and C. Manning. "A

Conditional Random Field Word Segmenter for Sighan Bakeoff 2005," *In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, pp. 168–171, 2005

21. Y. C. Wu, C. H. Chang and Y. S. Lee, "A General and Multi-lingual Phrase Chunking Model Based on Masking Method," Lecture Notes in Computer Science (LNCS): Computational Linguistics and Intelligent Text Processing, Vol. 3878, pp. 144-155, 2006
22. N. Xue. "Chinese Word Segmentation as Character Tagging," International Journal of Computational Linguistics and Chinese, pp. 29–48, 2003
23. N. Xue and L. Shen. "Chinese Word Segmentation as LMR Tagging," *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp. 176–179, 2003
24. H. P. Zhang, Q. Liu, H. Zhang and X. Q. Cheng. "Automatic Recognition of Chinese Unknown Words Based on Roles Tagging," *In Proceedings of First SIGHAN Workshop on Chinese Language Processing*, pp. 71-77, 2002
25. H. P. Zhang, H. K. Yu, D. Y. Xiong and Q. Liu. "HHMM-based Chinese Lexical Analyzer ICTCLAS," *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp. 187–187, 2003
26. J. H. Zheng and F. F. Wu. "Study on segmentation of ambiguous phrases with the combinatorial type," Collections of Papers on Computational Linguistics. Tsinghua University Press, Beijing, pp. 129-134, 1999

以字串特徵做為文本資料之錯誤偵測

劉吉軒、鄭雍瑋
國立政治大學資訊科學系
jsliu@cs.nccu.edu.tw

摘要

資訊擷取是從自然語言文本中辨識出特定的主題或事件的描述，進而萃取出相關主題或事件元素中的對應資訊。然而資訊擷取的結果會有錯誤情況發生，若單只依靠人工的方式進行錯誤的檢查及更正，將會是耗費大量人力及時間的工作。在本論文中，我們提出一種字串特徵為主的錯誤偵測方法，以資料描述的概念進行字串外表特徵的捕捉與轉換，再透過 C4.5 或 SVM 機器學習分類方法，自動建構適當的二元資料分類模型，進而達到辨別正確與錯誤資料的目的。

實驗結果顯示，本研究所提出的錯誤偵測方法，可以有效偵測出資訊擷取成果中不正確的值組，確保高品質的資訊擷取成果產出，促使資訊擷取技術更廣泛的實際應用。

關鍵詞：錯誤偵測、資料描述、資訊擷取

1. 緒論

資訊擷取(Information Extraction)是自然語言處理中的一個技術，能從自然語言文本中辨識出特定主題或事件描述，進而抽取核心資訊所對應的文字資料，如人、事、時、地、物等，而將原始非結構化的資料轉換成結構化資訊，並彙整成資料庫，

提供進一步的資訊加值處理與應用的可能性。對於大部分的資訊加值應用而言，資料的正確性應是最基本的條件。目前大多數資訊擷取研究皆致力於提升其擷取系統效能及各領域的應用性，但即使是最新的資訊擷取技術的擷取輸出，仍然是包含有某種程度的錯誤。因此，在進行任何資訊加值的應用之前，必須先將擷取結果加以驗證並更正錯誤。一般而言，資訊擷取技術所產生的資料量是相當龐大的，可能包括數以萬計的值組。若要以人工方式進行錯誤偵測及更正，甚至還得比對原始文件做進一步確認，這將會是耗費大量人力及時間的工作。因此，資料驗證所需的成本在部分資訊擷取應用上是一大障礙。可惜的是，現今大部分資訊擷取研究中，較少學者去針對擷取結果，進行驗證、偵測及更正錯誤資料的探討。我們認為，若能針對資訊擷取成果發展出適當的錯誤偵測機制，以確保高品質的資訊擷取成果產出，將可促成資訊擷取技術更廣泛的實際應用。

從資料清理的觀點來看，資訊擷取所產出資料集合的基本特性，是描述某項主題的文本資料。這種特性不利於使用一般的數據分析方法，如統計、分類與分群演算法。文本資料通常得透過文法和語義分析，並且必須對該領域主題事先進行領域知識的定義。這種需特別去定義領域知識的方式，不但會增加開發成本，並且也限制了應用性。為了增強資訊擷取效能與應用性，必須發展較通用的文本資料描述技術。此外，在資料清理的相關領域中，較少學者去針對以中文文字資料進行清理、錯誤偵測等研究。因此，我們希望能夠發展一套針對中文資料錯誤偵測機制，不但能提昇資訊擷取的效用以外，更能對於以中文為主的資料清理技術有所幫助。

2. 相關研究

資料清理(data cleansing)是一種針對資料集合進行識別、移除錯誤與不一致資料，進而改善資料品質的技術 [4]。不論是單一資料來源的資料庫，或是異質性資料來源的資料倉儲，都可以透過資料清理的技術來改善其資料產出的品質。資料清理技術以資料整合分析與資料稽核為主，其目的在於對資料進行分析，取得正確資料的特徵與規則後，便可偵測出異常及矛盾的資料，而指出資料庫中錯誤與不一致的情形。資料描述 (data profiling) 和資料探勘 (data mining) 這兩種相關的技術，對於資料分析有很大的幫助 [3]。資料描述的重點在於描述各個屬性值的資訊，譬如資料型態、長度、資料範圍、值組頻率、空值組出現情況以及字串規則等。資料探勘則是從大量的資料屬性中挖掘出有價值的資訊，藉由統計及人工智慧的技術，將資料做深入分析，找出相關資料的特定規則。

Galhardas和Raman兩位學者分別提出許多相關的資料清理技術[1] [5]，可以解決屬性遺失值(missing value)、雜訊資料(noisy data)、資料完整性、資料一致性等問題，藉此提高資料品質。例如，處理屬性遺失值的方法有忽略法(ignore tuple)、填補法(unknown)、平均法(mean)及線性回歸法(linear regression)等。忽略法就是不理會此資料，或者將此種資料刪除；填補法則是遇到遺失值的屬性欄位補上一特殊的代表值，如unknown；平均法是將所有相關的屬性值加總除上筆數，用此數值填入遺失的屬性欄位；線性回歸法則是利用統計的方法來找出最合適的值填補遺失值。

對於雜訊資料的處理，則有儲存槽法(bin)、叢聚法(clustering)等。儲存槽法的作法是將一連續的資料分割成離散資料，如年齡可分割成20 歲以下、20-40 歲、40-60 歲、60 歲以上等範圍。叢聚法則有分割(partition)、階層(hierarchy)、密度(density)等技術，主要是利用資料間的相似程度來加以分類。資料的一致性則是針對同名異物或同物異名的問題，這種現象通常發生於資料來源多重時。例如「motherboard」，主機板或母板都可代表，此時為了資料的一致，需要選定一代表值。另外在選取資料時，到底那份資料較能完全的符合使用者的需求，則屬於資料完整性的問題。藉由本體論，可判斷資料所包含的特徵個數和本體論所定義的特徵個數比率，計算資料的完整程度，可提供資料選取的一個指標。

在某些資料庫的欄位中，正確值組皆為唯一性的資料，若欄位內有重複值組則屬於錯誤資料。因此，唯一性偵測 (uniqueness detection)技術[6]的目的，在於自動偵測出目標欄位之值組是否為唯一性，以判斷資料的正確與否。這種技術以七個欄位屬性特徵為主，包括 data type、attribute length、whether a default value is specified、whether null is permitted 、distinctness ratio、min/max data length ratio 與 order of distinctness ratio with its relation。從訓練資料中擷取出這七個屬性特徵後，再透過 C4.5 演算法訓練出決策樹，便可透過此決策樹判定該欄位是否為唯一性，達到偵測錯誤資料之目的。

本研究針對資訊擷取系統之輸出資料，提出一種錯誤資料的偵測方式，以篩選

出錯誤的資訊擷取結果，進而搭配人工的查驗與更正，完成資料清理的工作。資訊擷取的結果是文本中的部分字串資料，通常無法適用於處理數值資料的統計方法。另外，欄位中的正確值組通常也包含重複的值組，例如，許多人有同樣的職位，因此，也不適用於唯一性偵測的技術。我們採用資料描述(data profiling)的技術概念，針對中文字串，提出一組字串特徵，以描述資訊擷取出的字串值組，再藉由相同欄位中的正確值組與錯誤值組的描述差異，達到錯誤偵測之目的。

3. 資訊擷取之錯誤偵測

在資料庫領域的資料品質議題上，資料錯誤的情形通常是因為人工輸入疏失及多重資料來源的資料不一致。而資訊擷取結果的資料錯誤情形，則是資訊擷取技術本身在文本資料的辨識與選取上所發生的錯誤，其資料錯誤的形式並不一樣。從資料取得的觀點來看，資訊擷取是針對選定的主題，從大量文本中辨識主題描述的存在、萃取字串資料，建立各屬性(欄位)中的字串集合，最後彙整而成結構化的資料庫資料。在這資料庫中，每一筆資料代表從文本中辨識出的一個主題個體(subject instance)，該筆資料中每一個欄位的值組，則是該主題屬性在文本中所應對應的描述字串。資料的錯誤通常是因為主題辨識、選取字串、對應欄位時發生錯誤。

一個完美的資訊擷取結果資料庫，是經過資訊擷取技術正確而沒有遺漏的辨識、選取、與對應，最後完整的匯集了文本集合中的所有主題資訊。任何不是完美的資訊擷取結果資料庫，就是有錯誤的資料庫。我們將其錯誤或異常情形分為以下四種情況：

1. 遺失個體(missing entities)：資訊擷取技術無法從文本中辨識出存在的主題個體，而造成資料庫中缺失了該主題個體的整筆資料。
2. 遺失值組(missing values)：資訊擷取技術可以從文本中辨識出存在的主題個體，但對於部分的屬性，卻因辨識失誤而忽略了對應字串的選取，造成該屬性在資料庫中的值組缺失。
3. 重複個體(duplicates)：資料庫中存在描述同一主題個體的多筆資料。重複個體的發生，有可能是資訊擷取技術在資料對應輸出時產生錯誤，也有可能是由於在文本集合中，同一主題個體的描述多次出現。
4. 不正確值組(invalid values)：資訊擷取技術可以從文本中辨識出主題個體的存在，但對於部分的屬性，卻因辨識失誤，而在字串選取或對應至欄位時產生錯誤，造成此欄位中的值組是不符合原始文本描述的屬性資訊。

針對一個資訊擷取結果建立的資料庫而言，遺失個體的偵測幾乎是不可能的，因為在不比對文本集合的情形下，並沒有任何的資訊可以判斷該主題個體的缺失。遺失值組的偵測也有同樣的困難，除非在主題資訊的定義上，已知某一屬性為必定存在。在具備此資訊的情形下，遺失值組的偵測是相當容易而直接的。重複個體是主題個體的重複，可以關鍵屬性的值組重複偵測出來。以上三種資料錯誤或異常情形，在偵測上完全是直接可以或不可以，並不具備偵測方法的研究議題。相對的，不正確值組的偵測就必須分析、判斷欄位中每一個值組的正確性，分析方式與判斷的適當性決定偵測的準確性，其結果的好壞差距可能非常大。因此，我們以不正確

值組的錯誤偵測方法為研究目標。給定由資訊擷取結果所彙整的資料庫，對於每一屬性的值組集合，我們將透過資料描述的方法，以機器學習方式建立資料分析與判斷模型，再依此模型去分析、偵測出不正確值組。這些錯誤偵測技術將能有助於降低資訊擷取結果的人工檢驗成本，提升資訊擷取技術之加值應用可行性。

3.1 字串特徵

資訊擷取結果中的屬性值組，是從文本中選取的部份字串。如果這些部分字串是正確的屬性值組，它們應是對應到同一個主題元素，並且通常以一至多個字詞形式呈現，表達同種語意類別的資訊。我們提出一種資料描述的方式，以字串的外表形式上的特徵，做為區別字串類別的依據。我們假設同一個主題元素的正確屬性值組，會有類似的或接近的字串外表特徵。因此，我們可以依據字串特徵的相同或相異，來判斷屬性值組的正確或錯誤。

我們所定義的字串特徵是描述字串的外表特徵，而不考慮其文字意義。依據語言和主題領域的不同，字串可以在字元層級與字詞層級顯示出不同的外表特徵。我們將重點擺在字元層級以及中文字串。同樣的觀念，也可適用於字詞層級與其他語言文本。針對此目的，我們共定義出六個字串特徵：

1. string cardinality (以下簡稱 S_c)：字串中的字元個數。
2. string prefix (以下簡稱 S_p)：字串前 k 個字元，k 是可設定的參數。
3. string suffix (以下簡稱 S_s)：字串後 k 個字元，k 是可設定的參數。
4. string entity (以下簡稱 S_e)：字串的所有字元序列。

5. string numeral (以下簡稱 S_n)：字串是否包含代表數字的字元，輸出結果為 true or false。
6. string format (以下簡稱 S_f)：字串內容所屬的資料型態。

我們以 *SF* 代表六個字串特徵的集合：$SF = \{S_c, S_p, S_s, S_e, S_n, S_f\}$，透過 *SF* 可用來評估資料庫中每個屬性值組(v_i)，$SF(v_i) = (S_c(v_i), S_p(v_i), S_s(v_i), S_e(v_i), S_n(v_i), S_f(v_i))$。以人事異動主題中之單位欄位為例，假設值組內容為「台北市政府」，則其字串特徵為如表一所示。

表一、字串特徵範例

SF(台北市政府) *and* *k*=1					
S_c	S_p	S_s	S_e	S_n	S_f
5	台	府	台北市政府	false	string

3.2 字串特徵之數值轉換

如前所述，我們對於不正確值組的偵測是建立於其異常字串特徵的假設上。進一步的說，我們假設一個欄位中的正確值組會有相同的或相當類似的字串特徵。因此，正確值組的字串特徵會經常出現而甚為普遍及常見。相對的，如果一個值組的字串特徵是少見的，就代表其字串特徵是異常的，也就可能是不正確的值組。這個假設是基於統計學上的多數法則(majority rule)。對於一個欄位中的值組集合，我們對字串的每一個特徵，計算每一個特徵值出現的百分比，再以其百分比轉換成一個適當的數值，做為後續判斷的依據。在資訊擷取結果的資料庫中，任何一個值組會具有六個字串特徵之轉換數值。理想上，一個正確值組的六個字串特徵值在該欄位的值組集合中都是相當常見的，而具備六個較大的特徵轉換數值。如果一個值組的數個特徵轉換數值都是較小的，代表其字串特徵值在該欄位的值組集合中都是較少見的，可以推論其可能的異常或錯誤。

首先，我們定義 $S_j, j \in \{1,2,3,4,5,6\}$,是上一小節中 SF 中各個字串特徵，資料庫中每個值組 v_i 之各別特徵值為 $S_j(v_i)$，其所占的百分比為 $P_{rob}(S_j(v_i))$，轉換成的數值為 $S_j'(v_i)$。要將特徵值出現的百分比轉換成一個適當的數值，有許多可能的方式。在本研究中，我們提出兩種相當直接的字串特徵數值轉換方式，都是將特徵值出現的百分比對應到一個固定等份 w 的區間，而 w 是可設定的參數。我們以 $T(w)$ 代表一個對應函數，將百分比對應到等份依序排列之區間數值，也就是百分比乘以 w 的結果取整數再加 1,但最大不超過 w。譬如 $T(10)$ 便是當百分比值為介於 0% 與 10% 之間時，對應的轉換數值為 1、百分比值介於 10% 與 20%之間時，對應的轉換數值為 2，依此類推到特徵轉換數值最大為 10。

第一種方式為個別百分比轉換，每一個字串特徵值依其出現的百分比個別的進行轉換，特徵轉換數值 $S_j'(v_i)$ 公式為：$S_j'(v_i) = P_{rob}(S_j(v_i)) \cdot T(w)$。第二種方式為累計百分比轉換，先將每一個字串特徵值依其出現的百分比由小到大排列，累加其百分比值後，再進行轉換。以 G 代表特徵值所佔百分比不大於 $P_{rob}(S_j(v_i))$ 的群組，特徵轉換數值 $S_j'(v_i)$ 的公式為：$S_j'(v_i) = \sum_{\forall i \in G} P_{rob}(S_j(v_i)) \cdot T(w)$。

我們以中文姓名屬性值組和字元個數特徵為例說明。假設資料庫中姓名屬性所有值組的字元個數的集合為{1, 2, 3, 4, 5, 6}，各自所佔的百分比為{1.5%, 11%, 79%, 5%, 3%, 0.5%}。假設 w 參數為 10，個別百分比轉換方式得到的結果如表二所示，其中字元個數為{1, 4, 5, 6}的百分比都是介於 0% 與 10% 之間，分別轉換之後的轉換數值都是 1，字元個數為{2}的百分比都是介於 10% 與 20% 之間，轉換之後的轉換數值是 2，字元個數為{3}的百分比都是介於 70% 與 80% 之間，轉換之後的轉換數值是 8。累計百分比轉換方式得到的結果如表三所示，其中字元個數為{6}的百分比最小，轉換之後的轉換數值是 1。接著是字元個數為{1}的百分比，累計字元個數為{6,1}的百分比之後的轉換數值仍是 1。再來是字元個數為{5}的百分比，累計字元

個數為{6,1,5}的百分比之後的轉換數值仍是 1。再來是字元個數為{4}的百分比，累計字元個數為{6,1,5,4}的百分比之後的轉換數值為 2。再來是字元個數為{2}的百分比，累計字元個數為{6,1,5,4,2}的百分比之後的轉換數值為 3。最後是字元個數為{3}的百分比，累計字元個數為{6,1,5,4,2,3}的百分比之後的轉換數值為 10。

表二、個別百分比轉換

$S_c(v_i)$	$P_{rob}(S_c(v_i))$	$S_c'(v_i)$
6	0.5 %	1
1	1.5 %	1
5	3 %	1
4	5 %	1
2	11 %	2
3	79 %	8

表三、累計百分比轉換

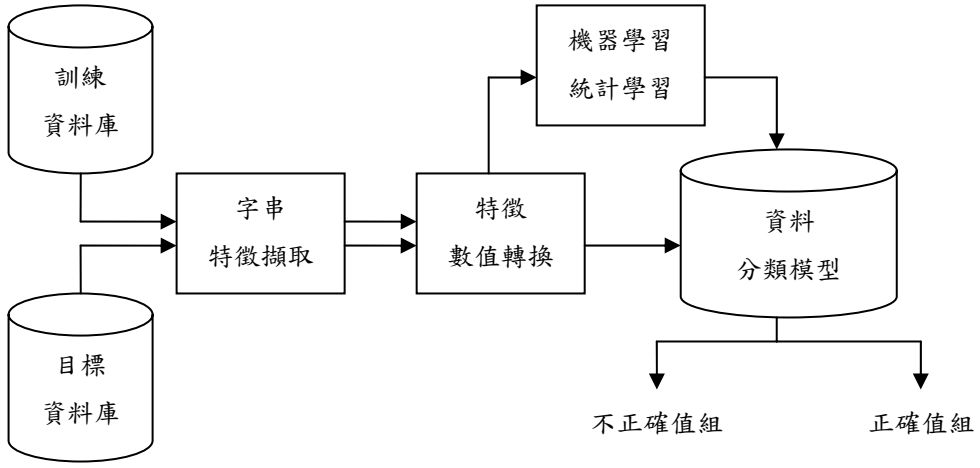
$S_c(v_i)$	$\sum_{\forall i \in G} P_{rob}(S_c(v_i))$	$S_c'(v_i)$
6	0.5 %	1
1	2 %	1
5	5 %	1
4	10 %	2
2	21 %	3
3	100 %	10

如前所述，我們的基本假設是正確的值組會有常見的字串特徵，其特徵出現的頻率較高，轉換後得到的數值較大。因此，特徵轉換數值的大小差異提供了一個推論及判斷值組為正確或錯誤的依據。特徵轉換數值的大小差異愈大時，代表少數值組的異常性愈明確，判斷其為錯誤值組的正確性機會愈高。若一值組集合在某一特徵上並沒有少數與多數的差異，而是平均的分布，則特徵轉換數值的大小差異就會不明顯，使得該特徵不容易做為區別正確與錯誤值組的依據。兩種轉換方式所提供的切割層面不同，對後續判斷與分類的準確性的影響也不同，我們將以實驗進行比較。

3.3 資料規則模型與錯誤偵測架構

資料或值組的錯誤偵測基本上可以視為二元分類的問題，對每一個值組進行正確或錯誤兩個類別的分類。為了自動建立有效的分類模型，我們採用機器學習或統計學習技術，透過含有已知分類結果的訓練資料，歸納出字串特徵及特徵轉換數值的分類規則或分類面。此分類模型即可用於目標資料中，對每一個值組進行分類，

達成判斷、辨識錯誤資料的目的。本研究所提出針對資訊擷取結果的錯誤偵測方法架構如圖一所示。



圖一、錯誤偵測方法架構圖

在自動建立資料分類模型的過程中，我們分別採用監督式機器學習(supervised machine learning)技術中的 C4.5 決策樹，及統計學習理論(statistical learning theory)中的支持向量機(support vector machine)。我們的目的是利用現有的自動學習技術，驗證字串特徵及特徵數值轉換做為錯誤資料偵測依據的成效。採用兩種不同技術的用意在於相互佐證錯誤偵測之成效，同時也驗證字串特徵之方法可有效搭配適當的自動分類學習技術。基於以上的考量，我們在 C4.5 決策樹及支持向量機的技術上，都是直接以公開可取得的軟體為主(本研究使用的 SVM 軟體及參數設定係參考國立台灣大學 LIBSVM 網站[2])，在實驗的過程中，並不做最佳化的調整。因此，相關的實驗結果只用以驗證字串特徵及特徵數值轉換做為錯誤資料偵測依據的成效，而無關 C4.5 決策樹及支持向量機兩者的比較。

4. 實驗評估

本研究選擇政府人事任免公報之擷取結果做為實驗對象，此公報分別以任命或免職的命令，記載政府各部門人事異動情形。政府人事任免公報的範例如圖二所示。

在過去的研究中[7]，我們針對此主題領域，採用型態辨識資訊擷取技術，處理約 20 年份的公報文本，共萃取出超過 10 萬筆人事異動資料，彙集而成包括姓名、組織單位、職位、職等、異動原因和日期等屬性的資料庫。目前大約二分之一的資料(西元 1995 年到 2004 年)已完成人工檢驗與校正，我們以這些資料做為訓練資料與測試資料的來源。

> ・・・
> 任命鄒擅銘為國史館臺灣文獻館簡任第十職等組長。
> 任命吳文慎為交通部臺灣區國道新建工程局人事室簡任第十職等主任，林渭鵬為經濟部水利處人事室簡任第十職等主任。
> 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。
> 行政院國家科學委員會科學工業園區管理科長曹常通另有任用，應予免職。
> ・・・

圖二、政府人事任免公報範例

4.1 評估方法

基本上，我們對資訊擷取結果的錯誤偵測方法是在建立一個適當的二元分類器，以指出資訊擷取結果中的合格資料與異常資料。因此，我們採用標準的 2 x 2 confusion matrix 做為偵測結果的效能指標。表四為以 true positive rate、true negative rate、false positive rate 和 false negative rate 四種量度所組成的 2 x 2 confusion matrix。

表四、2 X 2 confusion matrix

	分類為 正確資料	**分類為 錯誤資料**
原來為 正確資料	number of true positives (TP)	number of false negatives (FN)
原來為 錯誤資料	number of false positives (FP)	number of true negatives (TN)

這四種量度的定義如下，其中目標樣本是指原為正確類別的資料，非目標樣本是指

原為錯誤類別的資料：

1. True positive rate：TP-rate = TP/(TP+FN)，是指目標樣本分類正確的樣本數目比率。
2. True negative rate：TN-rate = TN/(FP+TN)，是指非目標樣本分類正確的非樣本數目比率。
3. False positive rate：FP-rate = FP/(FP+TN) = 1 – TN，是指非目標樣本分類成目標樣本的錯誤樣本數目比率。
4. False negative rate：FN-rate = FN/(TP+FN) = 1 – TP，是指目標樣本分類成非目標樣本的錯誤樣本數目比率。

二元分類器的效能目標是最大化 TP-rate、TN-rate 或最小化 FN-rate、FP-rate。一個理想的二元分類器能夠使 TP-rate 與 TN-rate 為 1。然而，在大多數的實際情形下，TP-rate 與 TN-rate 會是損益平衡關係，當我們對二元分類器進行調整，而能使 TP-rate 增加時，也同時會造成 TN-rate 的降低或 FP-rate 的提升。

4.2 實驗結果

為了能對字串特徵做為錯誤偵測依據的效能做不同層面的評估，本研究進行了四組實驗。第一組「字串特徵數值轉換」實驗是針對本方法所採用的兩種字串特徵數值轉換方式，比較其差異與優劣。第二組「訓練及目標資料範圍」實驗，乃根據訓練及目標資料組成範圍的不同，分為單一年份及合併年份，以觀察訓練資料的範圍，是否影響資料分類模型的表現。第三組「字串特徵數值轉換參數」實驗，將會改變 w 參數(區分的群組數)，以比較及分析其結果。第四組「訓練資料組成」實驗，是調整訓練資料的正反案例筆數，分別為 1:1 到 N:1 的各種組成方式，比較其資料分類模型效能之差異。

4.2.1 字串特徵數值轉換

本實驗中訓練資料的建立是於所有資料中，隨機取樣正反案例各 100 筆，並且隨機取樣三次，得到三組不同的訓練資料。實驗的進行，是以設定 *w* 參數為 100，並且以三組不同的訓練資料分別得到的分類結果，取其平均值，做為整體的實驗結果。表四為兩種字串特徵數值轉換方式的實驗結果，每一組資料中的數據分別代表 TP-rate/FP-rate。二元分類器的效能目標是最大化 TP-rate 和最小化 FP-rate，我們可以發現不論是由 C4.5 或 SVM 所建立的分類模型，「累計百分比轉換」方式的整體效果會比「個別百分比轉換」方式好。兩種轉換方法的差異，在於前者對於特徵值區分群組的切割較細，提供了機器分類法學習出更加細緻的分類規則。

此外，以表五中的各項實驗數據可以發現，對於「姓名」欄位的偵測表現不如其他欄位理想。這是由於該欄位的資料內容變化性較大，造成資料偵測模型較難精準的反映整體資料特性。而「單位」、「職稱」欄位的資料變化程度小於「姓名」資料，因此偵測的效果較好。至於「職等」、「年份」與「總統」欄位的資料內容皆是較為單純，因此透過訓練資料，就能建立出完整的分類模型，而得到非常理想的偵測效能。但是 SVM 分類法仍有些許誤判的情況，這點將在後續的小節說明。

表五、字串特徵數值轉換方式之比較

	C4.5 個別百分比轉換	C4.5 累計百分比轉換	SVM 個別百分比轉換	SVM 累計百分比轉換
姓名	84.91% / 33.57%	90.06% / 29.06%	87.11% / 60.79%	92.12% / 57.19%
單位	84.58% / 13.56%	88.58% / 7.69%	76.54% / 5.99%	79.61% / 4.00%
職等	100.00% / 0.00%	100.00% / 0.00%	86.21% / 5.04%	89.47% / 4.86%
職稱	87.44% / 11.36%	91.13% / 3.81%	80.58% / 2.01%	81.81% / 1.43%
總統	100.00% / 0.00%	100.00% / 0.00%	100.00% / 4.17%	100.00% / 2.58%
年份	100.00% / 2.88%	100.00% / 0.00%	93.96% / 5.00%	97.25% / 3.41%

4.2.2 訓練及目標資料範圍

本實驗建立兩種不同之訓練及目標資料範圍，分別為「單一年份」及「合併年份」，其他實驗參數包括 *w* 參數為 100，訓練正反案例筆數各為 100。「單一年份」

中，訓練資料為西元 2001 年，目標資料則分別為西元 2002 年、2003 年和 2004 年，因此共進行三次個別實驗後，再計算其整體平均之 TP-rate 和 FP-rate。而「合併年份」中，訓練資料與目標資料皆從十年份之資料庫取出(西元 1995 年到西元 2004 年)，而訓練資料為隨機取出三次，分別進行模型建立及資料分類測試，再計算其整體平均之 TP-rate 和 FP-rate。

表六為訓練及目標資料範圍的實驗結果，欄位中的數據仍分別代表 TP-rate/ FP-rate。整體而言，從「合併年份」資料範圍中所建立的分類模型，比從「單一年份」資料範圍中所建立的分類模型能得到較好的偵測結果。這是因為「合併年份」的資料範圍提供了較為全面的資料變化採樣空間，使得學習出的資料分類模型可靠度較高。至於各欄位的實驗結果與上一小節情況相似，對於資料變化程度越大的資料分類的準確度較低，反之亦然。

表六、訓練及目標資料範圍之比較

	C4.5 單一年份	C4.5 合併年份	SVM 單一年份	SVM 合併年份
姓名	95.78% / 37.49%	90.06% / 29.06%	76.33% / 57.29%	92.12% / 57.19%
單位	95.83% / 13.24%	88.58% / 7.69%	90.83% / 21.66%	79.61% / 4.00%
職等	100.00% / 0.00%	100.00% / 0.00%	85.23% / 5.62%	89.47% / 4.86%
職稱	89.62% / 17.45%	91.13% / 3.81%	58.86% / 16.11%	81.81% / 1.43%
總統	100.00% / 0.00%	100.00% / 0.00%	100.00% / 3.98%	100.00% / 2.58%
年份	100.00% / 2.56%	100.00% / 0.00%	95.07% / 4.44%	97.25% / 3.41%

4.2.3 字串特徵數值轉換參數

本實驗的目的是比較*w*參數，也就是特徵值出現的百分比對應的等份區間數目，依序設為5、10、25、50、100等變化下的不同分類表現。根據前面小節的實驗結果，本實驗在字串特徵數值轉換方式上，採用「累計百分比轉換」，而訓練與目標資料範圍，則是採用「合併年份」，訓練的正反案例筆數皆為100，並且隨機取樣三次建立三份測試資料，分別進行模型建立及資料分類測試，再計算其整體平均之

TP-rate和FP-rate。表七為*w*參數變化之實驗結果。

表七、字串特徵數值轉換參數之比較

欄位：姓名					欄位：職稱				
w	TP-rate		FP-rate		*w*	TP-rate		FP-rate	
	C4.5	SVM	C4.5	SVM		C4.5	SVM	C4.5	SVM
100	90.06%	92.12%	29.06%	57.19%	**100**	91.13%	81.81%	3.81%	1.43%
50	87.30%	84.82%	27.50%	51.56%	**50**	91.13%	86.11%	3.81%	6.65%
25	85.82%	72.95%	27.19%	28.75%	**25**	91.13%	89.89%	3.81%	9.33%
10	74.91%	80.17%	27.50%	40.00%	**10**	91.13%	88.12%	3.81%	8.79%
5	71.03%	87.05%	25.93%	49.56%	**5**	91.13%	89.19%	3.81%	11.40%
欄位：單位					**欄位：總統**				
w	TP-rate		FP-rate		*w*	TP-rate		FP-rate	
	C4.5	SVM	C4.5	SVM		C4.5	SVM	C4.5	SVM
100	88.58%	79.61%	7.69%	4.00%	**100**	100.00%	100.00%	0.00%	2.58%
50	94.46%	85.76%	19.01%	6.75%	**50**	100.00%	100.00%	0.00%	2.58%
25	90.06%	89.32%	9.36%	9.00%	**25**	100.00%	100.00%	0.00%	2.58%
10	82.45%	85.70%	4.50%	9.50%	**10**	100.00%	100.00%	0.00%	2.58%
5	81.71%	78.19%	3.94%	3.75%	**5**	100.00%	100.00%	0.00%	2.58%
欄位：職等					**欄位：年份**				
w	TP-rate		FP-rate		*w*	TP-rate		FP-rate	
	C4.5	SVM	C4.5	SVM		C4.5	SVM	C4.5	SVM
100	100.00%	89.47%	0.00%	4.86%	**100**	100.00%	97.25%	0.00%	3.41%
50	100.00%	88.51%	0.00%	3.47%	**50**	100.00%	97.25%	0.00%	3.41%
25	100.00%	88.51%	0.00%	3.47%	**25**	100.00%	97.25%	0.00%	3.41%
10	100.00%	88.64%	0.00%	3.47%	**10**	100.00%	97.25%	0.00%	3.41%
5	100.00%	88.51%	0.00%	3.47%	**5**	100.00%	97.25%	0.00%	3.41%

實驗結果顯示，*w* 參數較大時，所建立的資料分類模型的整體表現較好，至於其差異的大小則因欄位中內容變化的程度而異。例如，「姓名」欄位內容變化程度最大，*w* 參數提高，造成分群數目變多，使得資料分類模型在分類表現上的改善較為明顯。而「職等」、「總統」與「年份」欄位中，獨特值組的數目相當有限，造成該資料集合的特徵數值個數較少，因此當調整 *w* 參數時，不論區間數目增加或減少，

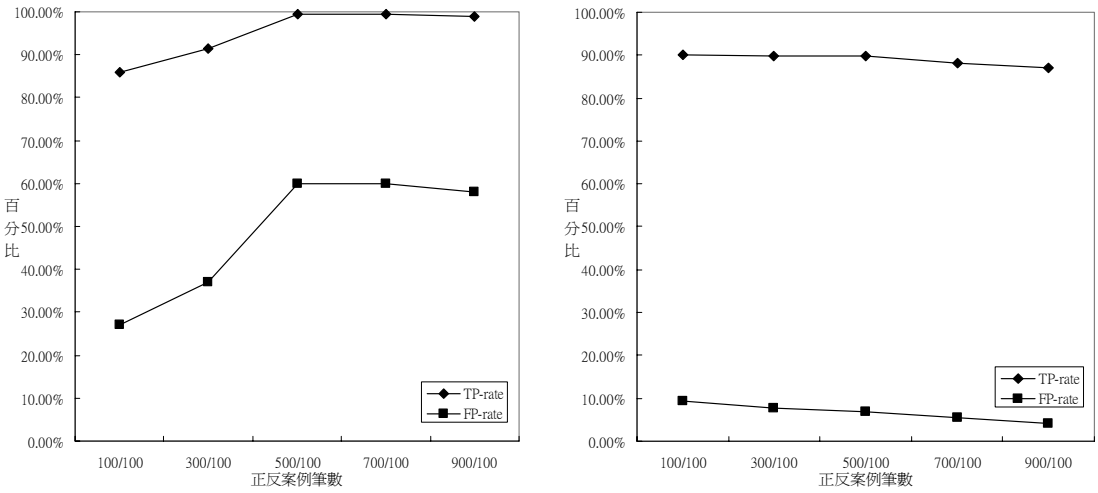
TP-rate 與 FP-rate 皆不受影響。至於「單位」與「職稱」欄位，其獨特值組的數目介於多與少的兩個極端之間，依照合理的推論，其分類表現應具備隨著 *w* 參數提升而改善的現象，但實驗結果所顯示的趨勢並不明顯，可能是本實驗之隨機採樣未能得到較完整之代表性，需要進一步的實驗驗證。另外，實驗結果也大致顯示 TP-rate 與 FP-rate 之損益平衡關係，即 TP-rate 若上升(TP-rate 變好)，則 FP-rate 也隨之上升(FP-rate 變壞)，兩者成反比關係。

4.2.4 訓練資料組成

本實驗是針對訓練資料中正反案例的組成變化，比較其建立之不同資料分類模型之表現。訓練資料的正反案例筆數組成分別為 100/100、300/100、500/100、700/100、900/100，並且隨機取樣三次，建立各種組成之三份測試資料，分別進行模型建立及資料分類測試，再計算其整體平均之 TP-rate 和 FP-rate。另外，*w* 參數設為 25，並使用「合併年份」為資料範圍。由於實驗資料是採用我們之前的資訊擷取成果，其擷取準確率約在 90%以上。因此，我們將正反案例比例逐步調整，從 1:1 到與資訊擷取成果中之正確的比例相同，進而觀察訓練案例比越趨近真實類別比例的表現變化。由於論文篇幅的限制，我們僅呈現較具代表性的實驗結果，包括以 C4.5 決策樹所建立之資料分類模型於「姓名」與「單位」兩個欄位上的表現變化。

如圖三所示，資料分類的表現大致上仍顯示 TP-rate 與 FP-rate 的損益平衡關係。在「姓名」欄位中，隨著正反案例比例的增加，TP-rate 的提升也帶來 FP-rate 的升高。「單位」欄位則是越接近 9:1 的比例，表現越好。「職稱」欄位在 3：1 的比例時，表現較好。而「職等」、「總統」與「年份」變化程度最小的欄位，其表現則是維持不變。每個欄位有不同的結果，是由於各欄位內容變化程度的差異相當大。「姓名」欄位的獨特值組數量非常龐大，所以正確案例筆數為 900 筆時，不但無法建立較完整的資料分類模型，反而造成 C4.5 決策樹所學習出的分類模型偏重於正確案例的一方，TP-rate 變好，但 FP-rate 也大幅增加。「單位」欄位在資料範圍中的獨特值組數

量約 1400 筆左右，所以當正確的訓練資料筆數越接近該數值時，所學習出的資料分類模型會更為正確與完整。



(a) 姓名欄位　　(b) 單位欄位

圖三、訓練資料組成變化對分類表現之影響

「職稱」欄位的獨特值組數量大約在 200 筆左右，所以在 3:1 的案例比時就有不錯的表現，當正確案例筆數增加，表現維持不變。至於「職等」、「總統」與「年份」欄位的資料就非常單純，獨特值組數量少於 30 筆，因此不但不受正反案例筆數的影響，並且整體效果接近完美。

4.3 實驗結果分析與討論

本研究所提出的以字串特徵做為錯誤偵測的方法，在實驗資料中的各項最佳表現如下：「姓名」欄位 TP-rate 約 85%、FP-rate 約 27%，「單位」欄位 TP-rate 約 87%、FP-rate 約 4%，「職稱」欄位 TP-rate 約 91%、FP-rate 約 3%，其他「職等」、「總統」與「年份」等欄位，皆能完美分類資料。實驗結果顯示，字串特徵能夠有效代表部份正反案例資料所展現之特徵，而建構出準確之資料分類模型，並與該欄位的內容特性相當吻合。這是由於字串特徵概念隱含了欄位內容之領域知識元素，再透過特徵數值轉換與 C4.5 決策樹的自動建構出適當的資料分類模型，而能展現有效的錯誤

偵測能力。

我們以表八做為說明。「姓名」欄位的分類模型之主要依據為 string cardinality (S_c)、string prefix (S_p)、string suffix (S_s)，分別是字元個數、字串前 k 個字元以及字串後 k 個字元。這個現象與姓名資料的特性相當吻合，譬如姓名資料的字元個數大都集中於 2 或 3，而字串前 1 個字元也是姓名中的姓氏，因此能夠適度的反映該資料之形式規則。「單位」欄位的決策樹是由這 string cardinality (S_c)、與 string suffix (S_s) 所組成。這個現象也與單位資料的特性相當吻合，不同單位名稱除了長度差異不大以外，最後一個字通常是「局」、「處」與「室」等常見的單位結尾詞。

表八、資料分類模型中之主要字串特徵

姓名	單位	職等	職稱	總統	年份
S_c、S_p、S_s	S_c、S_s	S_p、S_e、S_n	S_s、S_e	S_e、S_f	S_p、S_n、S_f

「職等」欄位則大多為「簡任第十職等」或「警正二階」等字串內容，這些字串大多含有數字類型的資料，因此造成 string numeral (S_n)字串特徵成為該欄位的決策樹主要節點之一。同時，字串開頭與完整字串內容都較為單純，所以 string prefix (S_p) 和 string entity (S_e) 也是該欄位的決策字串特徵。「職稱」欄位之資料內容，最後一個字通常是「員」、「官」與「長」等常見的職稱結尾詞，而「公務人員」與「警察官」的字串出現的次數相當高，因此 string suffix (S_s) 和 string entity (S_e)成為該欄位的主要決策字串特徵。

最後，「總統」欄位的內容只有兩任總統的名字，並且都是字串型態，因此 string entity (S_e) 和 string fromat (S_f)就可以反映出該欄位的特性。「年份」欄位的正確資料皆是數字型態以外，目標資料的年份為 84 年到 93 年，字串開頭為 8 或 9，也正好反映 string prefix (S_p)、string numeral (S_n) 和 string fromat (S_f)三個字串特徵的作用。

5. 結論

本研究提出以字串特徵為主的中文文本資料錯誤偵測機制，並以充分的實驗結果驗證其資料描述能力與錯誤偵測效能。此錯誤偵測機制能以後處理(post processing)的流程步驟，搭配一般的資訊擷取技術，確保高品質的資訊擷取成果產出，促成資訊擷取技術更廣泛的實際應用。另外，我們的研究成果也對於以中文為主的資料清理技術有所幫助，能應用於一般中文的大型商業資料庫上的錯誤偵測。

致謝聲明

本研究成果由國科會計畫 NSC 94-2422-H-004-002 及 NSC 95-2422-H-004-003 提供部分經費支持，特此致謝。

參考文獻

[1] Galhardas, H., Florescu, D., and Shasha, D. An Extensible Framework for Data Cleaning, *INRIA Technical Report*, 1999.

[2] LIBSVM, http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html.

[3] Muller, H., and Freytag, J. C. Problems, Methods, and Challenges in Comprehensive Data Cleansing. *Technical Report HUB-IB-164*, Humboldt University Berlin, 2003.

[4] Rahm, E. and Do, H.-H., "Data Cleaning: Problems and Current Approaches," *IEEE Bulletin of the Technical Committee on Data Engineering*, Vol. 23, No. 4, December 2000.

[5] Raman, V. and Hellerstein, J. M. An Interactive Framework for Data Cleaning, *UC Berkeley Computer Science Division Report No. UCB/CSD00/1110*, September 2000.

[6] 李念秋，資料品質改善之研究：錯誤資料偵測技術之發展與評估，國立中山大學資訊管理系碩士論文，2002。

[7] 翁家緯，以型態辨識為主的中文資訊擷取技術研究，國立政治大學資訊科學系碩士論文，2003。

Learning to Parse Bilingual Sentences Using Bilingual Corpus and Monolingual CFG

Chung-Chi Huang[1] and Jason S. Chang[2]

[1]Dept. of Information Systems and Application/Taiwan International Graduate Program, National Tsing Hua University, HsinChu, Taiwan, u901571@alumni.nthu.edu.tw
[1]Taiwan International Graduate Program, Academia Sinica, Nankang, Taiwan
[2] Dept. of Computer Science, National Tsing Hua University, HsinChu, Taiwan, Jason.jschang@gmail.com

Abstract

We present a new method for learning to parse a bilingual sentence using Inversion Transduction Grammar trained on a parallel corpus and a monolingual treebank. The method produces a parse tree for a bilingual sentence, showing the shared syntactic structures of individual sentence and the differences of word order within a syntactic structure. The method involves estimating lexical translation probability based on a word-aligning strategy and inferring probabilities for CFG rules. At runtime, a bottom-up CYK-styled parser is employed to construct the most probable bilingual parse tree for any given sentence pair. We also describe an implementation of the proposed method. The experimental results indicate the proposed model produces word alignments better than those produced by Giza++, a state-of-the-art word alignment system, in terms of alignment error rate and F-measure. The bilingual parse trees produced for the parallel corpus can be exploited to extract bilingual phrases and train a decoder for statistical machine translation.

1. Introduction

1.1. Background

The amount of information available in English on the Internet has grown exponentially for the past few years. Although a myriad of data are at our disposal, non-native speakers often find it difficult to wade through all of it since they may not be familiar with the terms or idioms being used in the texts.

To ease the situation, a number of online machine translation (MT) systems such as SYSTRAN and Google Translate provide translation of source text on demand. Moreover, online dictionaries have mushroomed to provide access at any time and everywhere for second language learners.

1.2. Motivation

MT systems and bilingual dictionary are designed to provide the services for non-English speakers or to ease learning difficulties for second language learners. Both require a lexicon which can be derived from aligning words in a parallel corpus.

Furthermore, second language learners can benefit by learning from example sentences with translations. By looking at bilingual examples, we acquire knowledge of the usage and meaning of word in context. With word alignment result of a sentence pair, it is much easier to grab the essential

concepts of unfamiliar foreign words in a sentence pair.

For instance, consider the English sentence "These factors will continue to play a positive role after its return" with its segmented Chinese translation "香港 回歸 後 這些 條件 將會 繼續 發揮 積極 作用" shown in Figure 1, where the solid dark lines are word alignment results of them and e, f stand for two sentences in two languages E, F respectively. If we don't know the usage of "play" in the sense of "perform," in this example sentence pair with the help of word alignment, we would quickly understand such meaning and learn useful expressions like "play … role" meaning "發揮 … 作用" in Chinese.

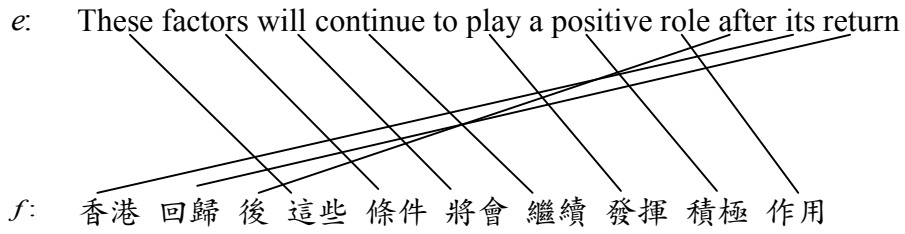


Figure 1. An example sentence pair.

Table 1. The word alignment of the example sentence pair.

i	j	e_i	f_j
1	4	These	這些
2	5	factors	條件
3	6	will	將會
4	7	continue	繼續
5	0	to	ε
6	8	play	發揮
7	0	a	ε
8	9	positive	積極
9	10	role	作用
10	3	after	後
11	1	its	香港
12	2	return	回歸

Table 1 shows the word alignment result of above example sentenece pair. In Table 1 we use 0, and ε to denote the corresponding translation does not exist for a particular word, that is, this word in one language is translated into no words in another and we use e_i, f_j to stand for the words at the position of i, j in sentence e, f respectively.

1.3. Bilingual Parsing

If we look more closely to the example sentence in Figure 1, we would notice that the beginning half "These factors will continue to play a positive role" is translated into the back of the Chinese sentence whileas the ending half "after its return" is translated into the beginning. This phenomenon is very common while translating one language into another. A simple observation is that if one language is SVO-structured and another SOV-structured, the "VO" part of the first language would constantly be reversely translated into "OV" of the second because of the reverse ordering of syntactic structures in "V" and "O" in these languages. We call it inverted word order during translation. More often than inverted cases, we have straight word order such as when "positive role" is translated into "積極 作用". It would occur more frequently if two languages have identical word orientation for a syntactic structure, such as adjectives modifying nouns in English and Chinese noun phrases.

In this paper, we propose a new method of learning to recognize straight and inverted phrases in bilingual parsing by using a parallel corpus and a monolingual treebank. The parallel text will be exploited to provide lexical translation information and project the syntactic information available in the source-language treebank onto the target language. This way we can leverage the monolingual treebank and avoid the difficult problem of inducing a bilingual grammar from scratch. We identify production rules derived from the treebank based on the part of speech information of the source text. This information is simultaneously projected to the target language by exploiting the cross-language lexical information produced by a word-aligning method. The relation of straight or inverted word orders between the syntax of the two languages at all phrase levels can be captured and modeled during the process. At runtime, these production rules are used to parse bilingual sentences, simultaneously determining the syntactic structures and word order relationships of languages involved.

Thus, the proposed model commits to common linguistic labels for words and phrases found in an English treebank, such as NN (noun), VB (verb), JJ (adjective), NP (noun phrase), VP (verb phrase), ADJP (adjective phrase), PP (prepositional phrase). Furthermore, we assume straight and inverted linguistic phenomena, when projected to the target language, should render a reasonable structural explanation of the target language. We extend ITG productions (Wu 1997) to carry out this process of projection. Take word-aligned sentences in Figure 1 for example. It is possible to match the part of speech information of the source language sentence against the right hand sides of the production rules induced from a tree bank and identify the instances of applying specific rules such as $NP \rightarrow JJ\ NN$; $JJ \rightarrow$ "positive" and $NN \rightarrow$ "role." Moreover, by exploiting the word alignment information, it is not difficult to infer that such syntactic structure is also present in the target language with similar rules

such as $NP \rightarrow JJ\ NN$; JJ → ”積極,” and NN → ”作用.” By combining and tallying such information, we are likely to derive ITG productions such as $NP \rightarrow [JJ\ NN]$; JJ → “positive/積極” and NN → “role/作用.” Here, the square bracket pair, “[“ and “]” signifies that a straight synchronous nominal share between English and Mandarin Chinese. Similarly, we would also find out the inverted prepositional phrases like $PP \rightarrow \langle IN\ NP \rangle$; IN → “after/後” and NP → “its return/香港 回歸” where “<“ and “>” indicate cross-language inverted structure. See Figure 3 for more details. Additionally, the occurrence counts of these straight or inverted structures can be tallied and used in estimating the probabilistic parameters of the ITG model.

Intuitively, with rules like those shown in Figure 2 learned from a parallel corpus and a monolingual treebank, we should be able to extend a CYK-style parser to derive bilingual parse tree as shown in Figure 3, where the symbol $\star$ indicates word order of the subtrees in the target language is inverted. According to the theory of ITG, the probability of a bilingual parse tree consists of the lexical translation probability and the probability for the straight or inverted production rules involved.

$$
\begin{aligned}
S &\rightarrow \langle NP\ PP \rangle \\
NP &\rightarrow [NP\ VP] \\
NP &\rightarrow [DT\ NP] \\
VP &\rightarrow [VP\ VP] \\
VP &\rightarrow [TO\ VP] \\
VP &\rightarrow [VP\ NP] \\
NP &\rightarrow [JJ\ NN] \\
PP &\rightarrow \langle IN\ NP \rangle \\
NP &\rightarrow [PRP\$\ NN]
\end{aligned}
$$

Figure 2. Example grammar rules for the sentence pair.

The rest of the paper is organized as follows. We review the related work in the next section. In Section 3, we describe the steps for learning synchronous grammar rules in the form of ITG and the association probabilistic estimation. An implementation of the bottom-up CYK-styled bilingual parser based on ITG is also described in Section 3. Reports on experiments and discussions are covered in Section 4 and 5, respectively.

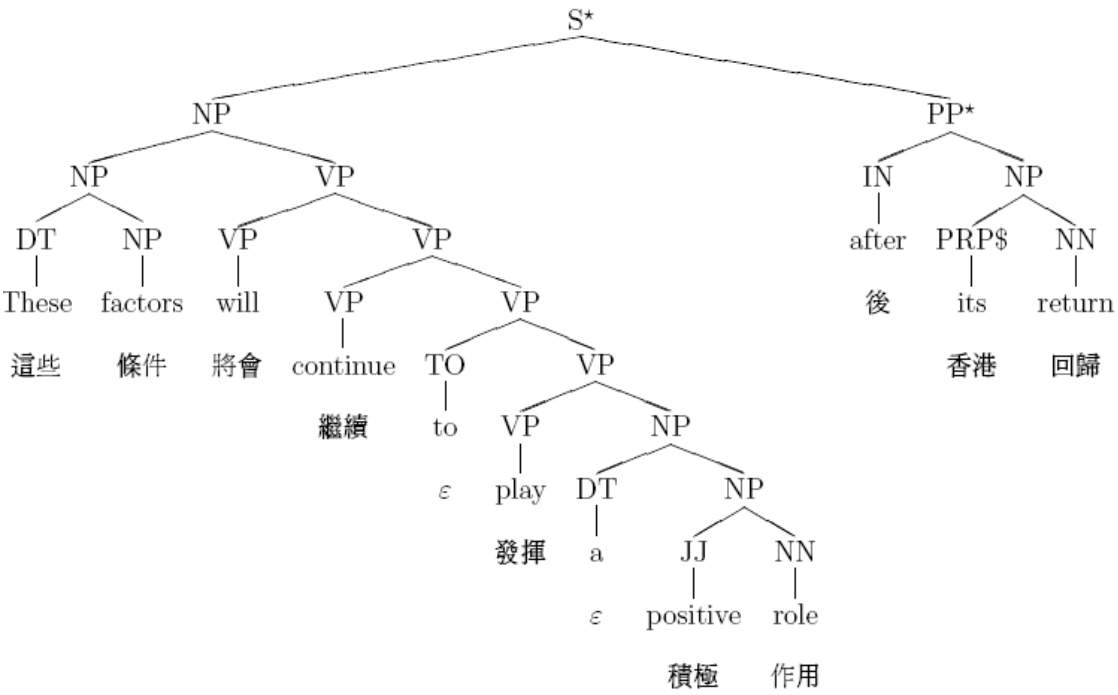


Figure 3. A bilingual parse tree for example sentence pair.

2. Related Works

A statistical translation model (STM) is a mathematical model in which process of human translation from one language into another is modeled statistically. Model parameters are estimated using a corpus of translation pairs with or without human supervision. STMs have been used in various researches and applications including statistical machine translation, word alignment of a sentence-aligned corpus and the automatic construction of a dictionary, just to name a few. For this point of view, a better STM cross language for processing is essential and fundamental for those applications.

Brown et al. (1988) first described a STM, or the alignment of sentence and word pairs in different languages. This and subsequent IBM models are based on noisy channel which converts or translates a sequence of words in one language into another. IBM model 1 can be trained using EM algorithm: starting with a uniform distribution among all translation candidate pairs and ending with convergent probabilities. While IBM model 1 does not utilize position information, the subsequent IBM models take positions into account when modeling for the translation process. (take an English-Chinese sentence pair for example, the first English word more likely translates into the first word in the Chinese sentence)

Another model called Hidden Markov model (HMM) is designed to capture localization effect in aligning the words in parallel texts. Vogel et al. (1996), motivated by the idea that words are not distributed arbitrarily over the sentence positions but tend to form clusters, presented a first-order HMM which makes the alignment probabilities explicitly dependent on the alignment position of the previous word. Nonetheless, Toutanova et al. (2002) pointed out that word order variations (large jumps) between languages seem to be a problem.

Neither IBM models nor HMMs explicitly utilize any linguistic information. However, other researchers have experimented with incorporation of part of speech (POS) information or context-specific features into STM. Exploiting POS tags of the two languages, Toutanova et al. (2002) introduced tag translation probabilities and tag sequences for jump probabilities to improve HMM-based word alignment models in modeling local word order differences. Cherry and Lin (2003) made use of dependency trees of a language to model features and constraints that are based on linguistic intuitions. In contrast, our model which uses POS information and tree structures from a treebank of a language to derive relation of syntax of two languages based on initial word alignments takes into consideration positions and linguistic characteristics such as word order and syntactic structures.

Wang (1998) enhanced the IBM models by introducing phrases, and Och et al. (1999) made use of templates to capture phrasal sequences in a sentence. While flat structures of languages beyond words are being used in above researches, often researchers attempted using nested structures. Those studies can be divided into two approaches according to whether they are linguistically syntax-based or not. Either ways, both approaches try to model structural differences between two languages.

Wu (1997) described an Inversion Transduction Grammar to model translation. However, only a lesser version, bracketing transduction grammar (BTG) with three structural labels *A,B,C* and a start symbol *S,* was experimented to perform bilingual parsing. Nevertheless, BTG accommodates a wide range of ordering variation between languages and imposes a realistic position distortion penalty. In other words, a system with structural-like, or hierarchical-like rules that specify the constituents and the order of the counterparts in both language is good at resolving the word alignment relations within a sentence pair. However, in their experiments, constituent categories are almost not differentiated, and thus their influences on ordering preferences of the counterparts are not taken into consideration. Consequently, very little syntax information is incorporated into the process of bilingual parsing. In contrast to Wu's experiment, we use regular context-free grammar rules in our experiments.

More recently, Yamada and Knight (2001) suggested the syntax differences in languages are really a better way to model translation. In their work, the English sentence goes through a parser to generate a full parse tree. Subtrees of each node are reordered, function words are inserted and finally the tree is linearized to produce the target sentence. The parse tree of an English sentence is generated independently from the target sentence. Although the monolingual parse might be correct, it may be difficult to project the structures onto the target language. Instead, our model has grammar rules that specify bilingual syntactic information including constituent labels and word ordering, which enables us to extend a CYK parser to parse bilingual sentences simultaneously.

Chiang (2005) introduced lexicalized labelless hierarchical bilingual phrase structure to model translation without any linguistic commitment. Since he does not assign any syntactic category to hierarchical phrase pairs, the rules he obtain are not generalized into linguistics-motivated constituents but anchored at certain words. These lexicalized rewrite rules specify the differences in hierarchical structure of two languages without generalization. Therefore, the size of the grammar tends to be very

large (2.2M rules). The rules do not represent some general ideas of languages such as word classes like verb, noun, or adjective, but rather have to do with specific words. In any case, the word classes like verb, noun, and adjective and the phrase categories like verb phrase (*VP*), noun phrase (*NP*) and adjective phrase (*ADJP*) would provide a more general way to reflect the parallel and differences of languages. Chiang also posed the hypothesis that syntactic phrases are better for machine translation (MT) and predicted the future trend of MT is to move towards a more syntactically-motivated grammar. With that in mind, we exploit part-of-speech information and linguistic phrase categories to model the syntactic relation between two languages, which is designed to have a higher degree of generality, unlike Chiang's lexicalized labelless production rules.

In contrast to previous work in STM, the proposed method not only automatically identifies the hidden structural information of two languages but models variations of ordering counterparts within them. Moreover, a much-smaller set of flexible context-free grammar rules obtained from a very large-scale parallel corpus. Syntactic information indicated by those rules is exploited to parse bilingual sentences.

3. The Model

A promising method for learning to parse a bilingual sentence using Inversion Transduction Grammars is based on training on a monolingual treebank and a parallel corpus. We project part of speech information and syntactic structures from a treebank of source language onto target language based on initial word alignment results of a parallel corpus to obtain and estimate the probabilities for ITG rules. During the projection process, word order relationships (*straight* and *inverted*) of shared syntactic constructs between two languages are identified and modeled. At runtime, the derived ITG rules drive a CYK-style parser to construct bilingual parse trees and hopefully lead to better word alignment results at the leaf nodes.

3.1. Problem Statement

The model is aimed at statistically derived ITG rules with probability and making use of those rules for bilingual parsing and word alignments. We focus on the process of bilingual parsing which exploits the syntactic information such as shared syntactic structures and word order relationships in two languages using a parallel corpus and a monolingual treebank.

Problem Statement: Given a sentence-aligned corpus $\mathbf{C} = \{(r, e, f) \mid 1 \le r \le n\}$ where r is the record number of the aligned sentence pair (e, f) and n is the total number of sentence pairs in parallel corpus $\mathbf{C}$, and a grammar $\mathbf{G} = \{lhs \rightarrow rhs \mid lhs \rightarrow rhs \text{ is a grammar rule on } E \text{ side}\}$ derived from a source-language treebank, we extend $\mathbf{G}$ into ITG rewrite rules for bilingual parsing.

For the rest of this section, we describe our solution to this problem. First, we elaborate on our training process for learning synchronous context-free grammar rules in the form of probabilistic estimation for ITG rules in Section 3.2. Then, we describe the implementation of a bottom-up bilingual

parsing algorithm based on ITG in Section 3.3.

3.2. Proposed Training Process

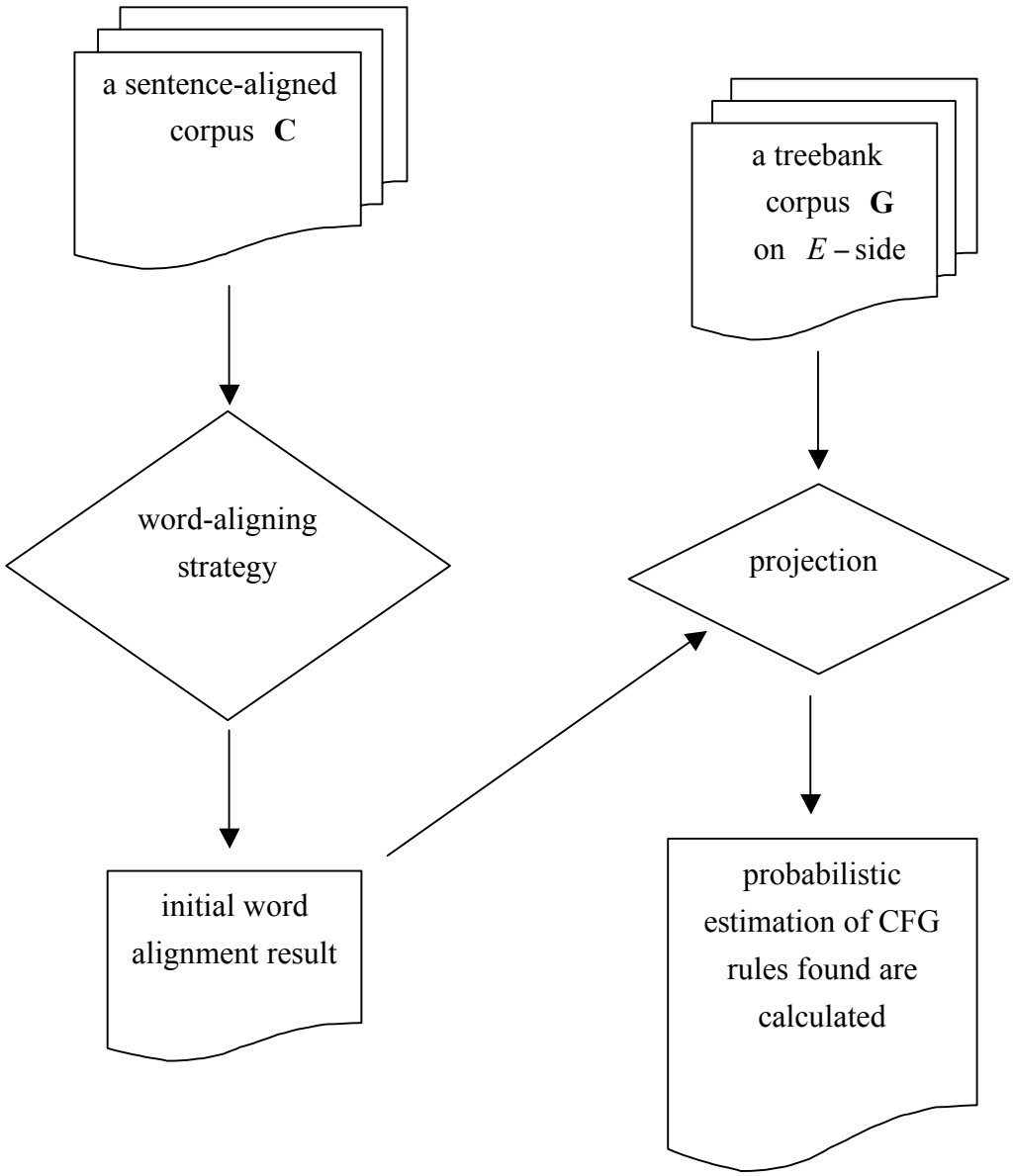


Figure 4: Flowchart of the proposed training process.

The training process can be illustrated using the flowchart in Figure 4.

Given a sentence-aligned corpus $\mathbf{C} = \{(r, e, f) \mid 1 \le r \le n, e \text{ and } f \text{ are an aligned sentence pair}\}$ where r is the record number of the sentence pair and n is the total number of sentence pairs in $\mathbf{C}$, a source-language grammar $\mathbf{G}$, we map part of speech information and syntactic structures of source language onto target language words using word alignment result. During the mapping process, we exploit occurrence of syntactic structures and the differences of word order of the right-hand-side constituents to estimate probabilities. The proposed training process is elaborated as follows.

Table 2. Outline of the training process.

(1)	Tag source-language sentences and segment target-language sentences (Section 3.2.1)
(2)	Apply a word-aligning strategy to obtain word alignment result (Section 3.2.2)
(3)	Apply the algorithm of projecting linguistic information of source language onto target language and estimating related probabilities of grammar rules found (Section 3.2.3)

Table 3. Lemmas and tags for English sentence of sentence pair 193.

position (i)	**lemma (e_i)**	**tag (t_i)**
1	these	*DT*
2	factor	*NNS*
3	will	*MD*
4	continue	*VB*
5	to	*TO*
6	play	*VB*
7	a	*DT*
8	positive	*JJ*
9	role	*NN*
10	after	*IN*
11	its	*PRP*$
12	return	*NN*

3.2.1 Tagging and Segmenting

In the first stage of the training process, for every sentence-aligned pair (e, f) in corpus $\mathbf{C}$, we tag sentence e using a POS tagger and generate $e = (e_1, e_2, \cdots, e_m)$ with tag sequence $(t_1, t_2, \cdots, t_m)$,

where e_i stands for the i[th] word in e with m words and t_i stands for the POS tag of the word e_i. Further, we segment sentence f to obtain $(f_1, f_2, \cdots, f_n)$, where f_j stands for the j[th] word in f with n words.

Take sentence pair whose record number is 193 in Figure 1 for instance. Table 3 shows the lemmatized and tagged result of the English sentence, while Table 4 shows the segmentation result of the Chinese sentence.

Table 4. Segments for Chinese sentence of sentence pair 193.

position (j)	**segments (f_j)**
1	香港
2	回歸
3	後
4	這些
5	條件
6	將會
7	繼續
8	發揮
9	積極
10	作用

The POS information of sentence e will then be projected onto the target language based on word alignments described in next subsection.

3.2.2 Initial Word Alignments

In the second training stage, we obtain a word-aligning set **A** for corpus **C** by applying any existing word-level alignment method.

For notation convenience, we use 8-tuple $(r, i_1, i_2, j_1, j_2, L, rhs, rel)$ to represent that substring pair $(e_{i_1} \cdots e_{i_2}, f_{j_1} \cdots f_{j_2})$ in sentence pair r has $L \rightarrow rhs$ as the derivation leading to the bilingual structure and rel as the cross-language word order relations (straight or inverted) of constituents of rhs. The right hand side, rhs, can be either a sequence of nonterminals or a single terminating bilingual word pair and the word order relation, rel, is either S (straight) or I (inverted). Followings

are some examples using the 8-tuple representation. The tuple $(193,1,2,4,5,NP,DT\ NN,\mathrm{S})$ denotes a straight bilingual noun phrase (these factors,這些 條件) in sentence 193. Similarly, the tuple $(193,10,12,1,3,PP,IN\ NP,\mathrm{I})$ denotes an inverted prepositional phrase (after its return,香港 回歸後). The tuple (193,8,8,9,9,JJ,positive/積極,S) denotes a terminal bilingual adjective (positive,積極) which can be obtained from word alignment result.

Table 5. Some alignments after applying a word-aligning strategy.

# of sentence pair	i	j	e_i	f_j	t_i
406	10	5	in	在	*IN*
406	11	8	overseas	海外	*JJ*
406	12	18	Chinese	中國	*JJ*
406	13	10	community	社區	*NN*

Further take word alignments of the sentence pair specified in Table 5 for example. $\mathbf{A}$ would at least contain entries like (406,10,10,5,5,*IN*,in/在,S), (406,11,11,8,8,*JJ*,overseas/海外,S), (406,12,12,18,18,*JJ*,Chinese/中國,S) and (406,13,13,10,10,*NN*,community/社區,S).

3.2.3 Algorithm for Probability Estimation

In the final stage of the training process, we map the part of speech information and tree structures available in treebank of language E onto language F based on word alignment result.

We exploit following algorithm to identify syntactic structures of E and model the syntactic relation between E and F. The resulting ITG grammar will then be used in a bottom-up CYK parser for parsing bilingual sentences.

The algorithm begins with a set $\mathbf{H}$ initialized as word-aligning result $\mathbf{A}$. Then recursively select two elements from $\mathbf{H}$. If these two tuples have contiguous word sequence on source-language side and exhibit *straight* or *inverted* relation between source and target language during the mapping process, a new tuple representing these two is added into $\mathbf{H}$. In the end, we exploit the occurrence in $\mathbf{H}$ to estimate following probabilities: $\mathrm{P}(L \to [R_1\ R_2])$, $\mathrm{P}(L \to \langle R_1\ R_2 \rangle)$ and $\mathrm{P}(L \to t)$.
In this algorithm, we follow the notation described in section 3.2.1 and use $|\mathbf{W}|$ to stand for the number of entries in set $\mathbf{W}$, $\mathrm{count}(p;\mathbf{Q})$ for the frequency of p in set $\mathbf{Q}$ and δ for the tolerance of *straight*/*inverted* phenomenon within source and target languages.

Algorithm for Probabilistic Estimation

H = A

For $(r,i_1,i_2,j_1,j_2,L,rhs,rel) \in \mathbf{H}, (r,\overline{i_1},\overline{i_2},\overline{j_1},\overline{j_2},\overline{L},\overline{rhs},\overline{rel}) \in \mathbf{H}$ have not yet been considered

If ($i_2 = \overline{i_1} - 1$)

For every $L' \to L\ \overline{L} \in \mathbf{G}$

If ($j_2 + 1 \le \overline{j_1} \le j_2 + \delta$)

$$\mathbf{H} = \mathbf{H} \cup \left\{\left(r, i_1, \overline{i_2}, j_1, \overline{j_2}, L', L\ \overline{L}, \mathrm{S}\right)\right\}$$

If ($\overline{j_2} + 1 \le j_1 \le \overline{j_2} + \delta$)

$$\mathbf{H} = \mathbf{H} \cup \left\{\left(r, i_1, \overline{i_2}, \overline{j_1}, j_2, L', L\ \overline{L}, \mathrm{I}\right)\right\}$$

If ($\overline{i_2} = i_1 - 1$)

For every $L' \to \overline{L}\ L \in \mathbf{G}$

If ($\overline{j_2} + 1 \le j_1 \le \overline{j_2} + \delta$)

$$\mathbf{H} = \mathbf{H} \cup \left\{\left(r, \overline{i_1}, i_2, \overline{j_1}, j_2, L', \overline{L}\ L, \mathrm{S}\right)\right\}$$

If ($j_2 + 1 \le \overline{j_1} \le j_2 + \delta$)

$$\mathbf{H} = \mathbf{H} \cup \left\{\left(r, \overline{i_1}, i_2, j_1, \overline{j_2}, L', \overline{L}\ L, \mathrm{I}\right)\right\}$$

For $\left(r, i_1, i_2, j_1, j_2, L, rhs, rel\right) \in \mathbf{H}$

If ($rhs \ne t$)// t stands for terminating bilingual word pair

$$\mathrm{P}\left(L \to \left[R_1\ R_2\right]\right) = \frac{\mathrm{count}\left(\left(*,*,*,*,*,L,R_1\ R_2,\mathrm{S}\right);\mathbf{H}\right)}{\left|\mathbf{H}\right|}$$

$$\mathrm{P}\left(L \to \left\langle R_1\ R_2\right\rangle\right) = \frac{\mathrm{count}\left(\left(*,*,*,*,*,L,R_1\ R_2,\mathrm{I}\right);\mathbf{H}\right)}{\left|\mathbf{H}\right|}$$

Else

$$\mathrm{P}\left(L \to t\right) = \frac{\mathrm{count}\left(\left(*,*,*,*,*,L,t,\mathrm{S}\right);\mathbf{H}\right)}{\left|\mathbf{H}\right|}$$

Table 6. Some alignments by applying an aligning strategy on corpus **C** .

# of sentence pair	i	j	e_i	f_j	t_i
1	1	1	solemn	莊嚴	*JJ*

1	2	2	ceremony	儀式	*NN*
1	3	3	mark	標誌	*VBZ*
1	4	4	handover	回歸	*NNS*
9	24	6	before	前	*IN*
9	25	5	midnight	午夜	*NN*
62	12	5	provisional	臨時	*JJ*
62	13	6	legislative	立法	*JJ*
62	14	7	council	會	*NN*
249	2	2	will	將	*MD*
249	3	3	strive	致力	*VB*

Consider the word alignment results in Table 6 as an example, the algorithm described above will identity syntactic structures and model syntax relations of languages. The overall projecting process is as follows.

Initially, for sentence pair 1, we have the following in **A** .

(1,1,1,1,1,*JJ*,solemn/莊嚴,S)

(1,2,2,2,2,*NN*,ceremony/儀式,S)

(1,3,3,3,3,*VBZ*,mark/標誌,S)

(1,4,4,4,4,*NNS*,handover/回歸,S)

Table 7. Examples for the algorithm.

# of sentence pair	**rule**	**entry**
9	$PP \rightarrow IN\ NN$	$(9, 24, 25, 5, 6, PP, IN\ NN, \mathrm{I})$
62	$NP \rightarrow ADJP\ NN$	$(62, 12, 14, 5, 7, NP, ADJP\ NN, \mathrm{S})$
249	$VP \rightarrow MD\ VB$	$(249, 2, 3, 2, 3, VP, MD\ VB, \mathrm{S})$

After the first round, we have (1,1,2,1,2,*NP*,*JJ NN*,S), (1,3,4,3,4,*VP*,*VBZ NNS*,S). After the second round, we have (1,1,4,1,4,*S*,*NP VP*,S) where syntactic label *S* means simple declarative clause in linguistic sense.

Table 7 illustrates some derived grammar rules and entries inserted into **H** from sentence pair 9, 62 and 249.

3.3. Bottom-up Parsing

We then describe how we implement a bilingual parser which makes use of syntactic structures and preferences of word order within languages specified by automatically trained ITG rules.

We follow Wu's (1997) definition of $\delta_{stuv}(i)$ to denote the probability of the most likely parse tree with syntactic label i and containing substring pair $(e_{s+1}\ e_{s+2}\cdots e_t, f_{u+1}\ f_{u+2}\cdots f_v)$ in bilingual sentence (e, f).

3.3.1 Implementation

Given sentence $e=(e_1, e_2, \cdots, e_m)$ with tag sequence $(t_1, t_2, \cdots, t_m)$, its corresponding translation sentence $f=(f_1, f_2, \cdots, f_n)$, and a set of probabilities such as $\mathrm{P}(L \to t)$, $\mathrm{P}(L \to [R_1\ R_2])$ and $\mathrm{P}(L \to \langle R_1\ R_2 \rangle)$ associated with ITG, we utilize dynamic programming technique to find the most probable derivation to parse the bilingual sentence (e, f). Basically, we try to calculate the value of $\delta_{0m0n}(\overline{S})$ and backtrack by using following three steps, where $\overline{S}$ is the start symbol.

Step 1: Initial step

$$\delta_{i-1,i,j-1,j}(t_i) = \mathrm{P}(t_i \to e_i / f_j) \qquad \text{for } 1 \le i \le m, 1 \le j \le n$$

$$\delta_{i-1,i,j-1,j}(L) = \mathrm{P}(L \to e_i / f_j) \qquad \text{for } 1 \le i \le m, 1 \le j \le n, L \to t_i \in \mathbf{G}$$

$$\delta_{i-1,i,j,j}(t_i) = \mathrm{P}(t_i \to e_i / \varepsilon) \qquad \text{for } 1 \le i \le m, 0 \le j \le n$$

$$\delta_{i-1,i,j,j}(L) = \mathrm{P}(L \to e_i / \varepsilon) \qquad \text{for } 1 \le i \le m, 0 \le j \le n, L \to t_i \in \mathbf{G}$$

$$\delta_{i,i,j-1,j}(NN) = \mathrm{P}(NN \to \varepsilon / f_j) \qquad \text{for } 0 \le i \le m, 1 \le j \le n$$

Step 2: Recurrent step (bottom-up approach)

We proceed similar to Wu's algorithm. However, we observe that the length of the translation of a substring of source sentence should be bounded. We use the upper and lower bounds of lengths to prune search space and speed up computation. Consequently, $\delta^{[\,]}_{stuv}(i), \delta^{\langle\,\rangle}_{stuv}(i)$ are calculated as below:

If $\frac{1}{ratio} \le \frac{t-s}{v-u} \le ratio$

$$\delta^{[\,]}_{stuv}(i) = \max_{\substack{j \in \mathbf{PJ} \\ k \in \mathbf{PK} \\ s \le S \le t \\ u \le U \le v \\ (S-s)(t-S)+(U-u)(v-U) \neq 0}} \left\{ \mathrm{P}\left(i \rightarrow [j\ k]\right) \times \delta_{sSuU}(j) \times \delta_{StUv}(k) \right\}$$

where **PJ** is the set consisting of possible syntactic labels for substring pair $\left(e_{s+1} \cdots e_S, f_{u+1} \cdots f_U\right)$ and **PK** is the set consisting of possible syntactic labels for substring pair $\left(e_{S+1} \cdots e_t, f_{U+1} \cdots f_v\right)$

Else

$$\delta^{[\,]}_{stuv}(i) = low\_probability$$

If $\frac{1}{ratio} \le \frac{t-s}{v-u} \le ratio$

$$\delta^{\langle\,\rangle}_{stuv}(i) = \max_{\substack{j \in \mathbf{PJ} \\ k \in \mathbf{PK} \\ s \le S \le t \\ u \le U \le v \\ (S-s)(t-S)+(U-u)(v-U) \neq 0}} \left\{ \mathrm{P}\left(i \rightarrow \langle j\ k \rangle\right) \times \delta_{sSUv}(j) \times \delta_{StuU}(k) \right\}$$

where **PJ** is the set consisting of possible syntactic labels for substring pair $\left(e_{s+1} \cdots e_S, f_{U+1} \cdots f_v\right)$ and **PK** is the set consisting of possible syntactic labels for substring pair $\left(e_{S+1} \cdots e_t, f_{u+1} \cdots f_U\right)$

Else

$$\delta^{\langle\,\rangle}_{stuv}(i) = low\_probability$$

Step 3: Reconstructing step

We exploit depth-first-traversal to construct the most probable bilingual parse tree for sentence pair (e, f).

3.3.2 Example Parse

Take sentence pair in Figure 1 for example.

At initial step, we would build the leaf nodes of the bilingual parse tree using probability like P(*DT* → these/這些), P(*NNS* → factors/條件), P(*NP* → factors/條件), ⋯, P(*IN* → after/後), P(*PP* → after/後), P(*PRP*$ → its/香港), P(*NP* → its/香港) and etc.

At recurrent step, we find the most likely derivation of nodes using statistics derived so far. Take nodes in Figure 3 for instance. We will derive (these factors, 這些 條件) as a noun phrase using $NP \rightarrow [DT\ NP]$, an *inverted* prepositional phrase (after its return, 香港 回歸 後) using $PP \rightarrow \langle IN\ NP \rangle$, and a *straight* verb phrase (play a positive role, 發揮 積極 作用) using $VP \rightarrow [VP\ NP]$.

After reconstructing step, the most probable bilingual parse tree of the sentence pair is

constructed. Figure 3 illustrates the tree structures derived for the example bilingual sentences.

4. Experiments

Our model is aimed at capturing shared syntactic structures and preferences in word order between two languages. The context-free grammar rules obtained in training process identity syntactic structures and model relations of syntax of languages involved. These rules can be exploited to produce better word-level alignments and most probable bilingual parse trees since syntactic information is taken into consideration.

In this section, we first present the details of training our model in Section 4.1. Then, we describe the evaluation metrics for the performance of the trained model in Section 4.2. The evaluation results are reported in Section 4.3.

4.1. Training Setting

We used the news portion of Hong Kong Parallel Text (Hong Kong news) distributed by Linguistic Data Consortium (LDC) as our sentence-aligned corpus **C**. The corpus consists of 739,919 English and Chinese sentence pairs. English sentence is considered to be the source while Chinese sentence is the target. The average sentence length is 24.4 words for English and 21.5 words for Chinese. Table 8 and Table 9 show the statistics of number of sentences in this corpus according to sentence length. For monolingual treebank corpus **G**, we made use of PTB section 23 production rules distributed by Andrew B. Clegg (http://textmining.cryst.bbk.ac.uk/acl05/). There are 2,184 distinct grammar rules. The statistics of **G** is shown in Table 10 while Table 11 illustrates some examples of grammar rules in **G**.

Table 8. Statistics on English side.

sentence length	number of sentence	percentage
0~5	93,354	12.6%
6~10	118,513	16.0%
11~15	70,634	9.5%
16~20	66,431	9.0%
21~25	74,813	10.1%
26~30	71,902	9.7%
31~35	63,816	8.6%
36~	180,456	24.4%

Table 9. Statistics on Chinese side.

sentence length	number of sentence	percentage
0~5	146,957	19.9%
6~10	81,716	11.0%
11~15	72,870	9.8%
16~20	90,286	12.2%
21~25	84,802	11.4%
26~30	74,739	10.1%
31~35	57,347	7.7%
36~	131,202	17.7%

Table 10. Statistics of monolingual treebank.

# of constituents on right hand side	# of distinct grammars	percentage
1	106	4.85%
2	418	19.13%
3	752	34.43%
4	553	25.32%
5~	355	16.25%

Table 11. Example grammars in **G** .

grammar rules	
VP → VB	*NP → DT ADJP NNS*
ADJP → RB JJ	*PP → RB IN NP*
VP → TO VP	*VP → MD ADVP VP*
PP → IN NP	*ADVP → ADVP CC ADVP*
NP → DT JJ NN	

As for word alignment, we used bidirectional ranking (BDR) as the word-aligning strategy in training process, which means in a sentence pair, e_i and f_j will be aligned if $j = \arg\max_{i-sw \le q \le i+sw} dice(e_i, f_q)$, $i = \arg\max_{j-sw \le p \le j+sw} dice(e_p, f_j)$ and $dice(e_i, f_j) > \theta_{dice}$ where sw is the winow size (set to 7 in the experiment), θ_{dice} is the threshold for dice (set to 0.002) and $dice(\bar{e}, \bar{f})$ is calculated as

$$\frac{2 \times |link(\bar{e}, \bar{f})|}{|link(\bar{e}, *)| + |link(*, \bar{f})|}$$

where * is the wildcard symbol and $\bar{e}, \bar{f}$ are words in language E, F respectively. Furthermore, in estimating ITG, we consider only fourgram on English side, that is, entries $(r, i_1, i_2, j_1, j_2, L, det)$ in $\mathbf{H}$ satisfy the criterion $i_2 - i_1 \le 3$. For the *straight* case to hold, the two Chinese fragments need to be contiguous or have a function word in-between while they need to be contiguous for the *inverted* case to hold.

Since the pieces have come to together, we follow the steps specified in Table 2 to learn ITG rules. Table 12 shows some of the grammar rules trained and associated estimations.

Table 12. Examples of grammar rules trained and their probabilities.

$L \to R_1\ R_2$	$P(L \to [R_1\ R_2])$	$P(L \to \langle R_1\ R_2 \rangle)$	$count(L \to [R_1\ R_2])$	$count(L \to \langle R_1\ R_2 \rangle)$
$S \to NP\ VP$	0.0107950	0.0009212	145,421	12,409
$VP \to VP\ NP$	0.0109561	0.0005481	147,591	7,383
$PP \to IN\ NP$	0.0031136	0.0007793	41,944	10,498
$VP \to VP\ VP$	0.0035528	0.0003922	47,860	5,283
$NP \to JJX\ NNX$	0.0108844	0.0006971	146,624	9,391
$NP \to ADJP\ NNX$	0.0148228	0.0008140	199,681	10,965

In Table 12 we notice that the adjective-noun structure has much more *straight* cases than *inverted*. In other words, adjectives modify nouns in much the same manner in English and Chinese. In general, the statistics suggests that Chinese, much like English, is SVO with only relatively small number of exceptional cases.

Another point worth mentioning is that the overwhelming predominance of *straight* over *inverted* is not observed in the rule of $PP \to IN\ NP$. For this grammar rule, the *straight* cases like "in August", "在 八月份" and the *inverted* cases such as "before midnight", "午夜 前" are about the same order of magnitude. Consequently, it seems that there is no decisive preference of translation

orientation for prepositional phrases.

4.2. Evaluation Metrics

We evaluated the trained ITG rules based on the performance of word alignment. We took the leaf nodes as word-level alignments and evaluate the proposed model in terms of agreement with human-annotated word alignments.

We used the metrics of alignment error rate (AER) proposed by Och and Ney (2000), in which the quality of a word alignment result $\mathbf{A} = \{(i, j)\}$, where i, j are positions of the sentence pair e, f respectively and $i, j \neq 0$, is evaluated using

$$precision = \frac{|\mathbf{A} \cap \mathbf{P}|}{|\mathbf{A}|}, \quad recall = \frac{|\mathbf{A} \cap \mathbf{S}|}{|\mathbf{S}|} \text{ and}$$

$$AER(\mathbf{S}, \mathbf{P}; \mathbf{A}) = 1 - \frac{|\mathbf{A} \cap \mathbf{S}| + |\mathbf{A} \cap \mathbf{P}|}{|\mathbf{A}| + |\mathbf{S}|},$$

where $\mathbf{S}$ (sure) is the set which contains alignments that are not ambiguous and $\mathbf{P}$ (possible) is the set consisting of the alignments that might or might not exist $(\mathbf{S} \subseteq \mathbf{P})$. For that the human-annotated alignments may contain many-to-one and one-to-many relations. Furthermore, whether a word-level alignment is in $\mathbf{P}$ or $\mathbf{S}$ is determined by human experts who perform the annotation work.

4.3. Evaluation Result

For testing, we randomly selected 62 sentence pairs from the corpus of Hong Kong News. For the sake of time, we only selected sentence pairs in which the length of English and Chinese sentences does not exceed 15. From Table 8 and Table 9, we know the upper bound of 15 would cover approximately 40% of sentence pairs in HKN. We manual annotated the word alignment information in these bilingual sentences. The ratio of $|\mathbf{P}|$ and $|\mathbf{S}|$ of the test data is 1.2.

4.3.1 Baseline

We chose a freely-distributed word-aligning system, Giza++, as the baseline for evaluation. The adopted setting to run Giza++ is IBM model 4, the direction is from English to Chinese same as our model treating English as source language and the alignment units of Chinese are words not characters.

4.3.2 Word-level Evaluation

As preliminary evaluation, we examined whether syntactic consideration would lead to better word-level alignments. Figure 5 shows some alignments produced by the system and Giza++ and Table 13 displays evaluation results on alignments of the test data produced by both systems.

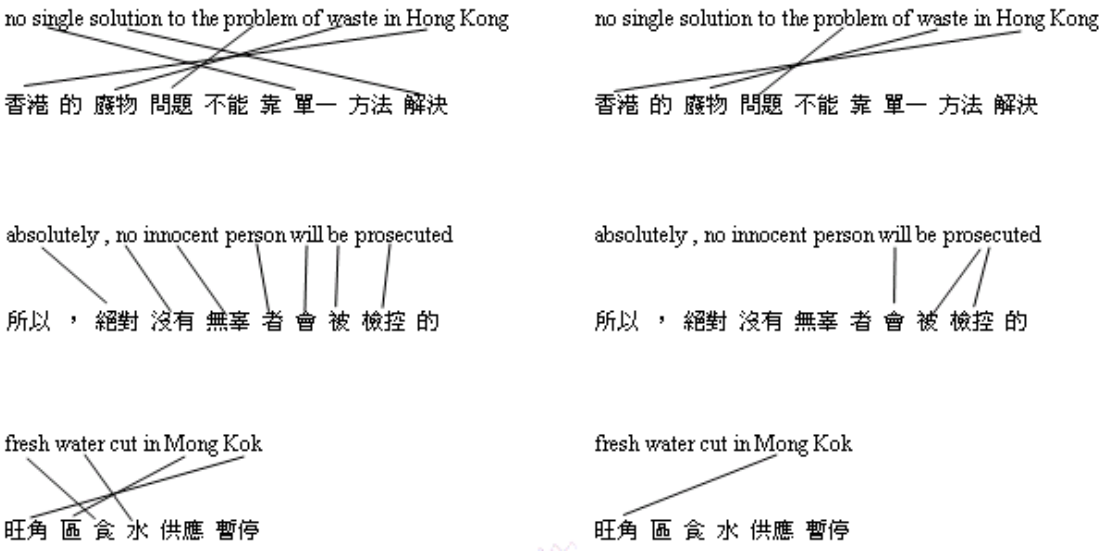


Figure 5. Alignments produced by our system (left) and Giza++ (right).

Table 13. Alignment results of the test data. Our system vs. Giza++.

	Recall	**Precision**	**AER**	**F-measure**
The proposed method	0.55	0.80	0.34	0.65
Giza++	0.37	0.87	0.48	0.52

Table 13 shows that although the precision is 87% for Giza++, the low recall leads to high alignment error rate and poor F-measure. However, our system with lower precision increased recall by 48.6%, which achieved a 29.2% alignment error reduction. From this experiment, we showed the proposed model with ITG rules allows for a wide range of ordering variations with a realistic position distortion penalty, which attributes to significantly better word alignment results.

Since the proposed model takes lexical and syntactic aspects of languages into consideration, the proposed method can be used to improve an existing word-aligning system that utilizes few linguistic information of languages. For that we evaluated the proposed method on top of the alignment results of Giza++, a freely-available state-of-the-art word alignment system. In other words, the **C** and **G** corpora are the same as the previous experiment but we adopted Giza++ as the word-aligning method in the training process.

Figure 6 shows some word alignment results produced by Giza++ with ITG and Giza++. Table 14 shows even better improvement than using the word alignment system along.

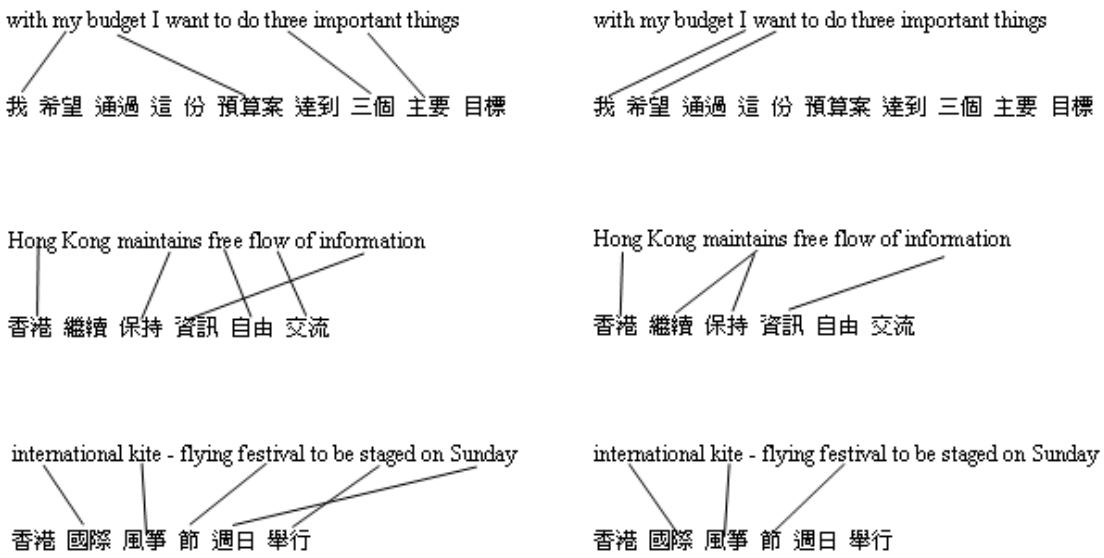


Figure 6. Alignments produced by Giza++ with ITG (left) and Giza++ (right).

Table 14. Alignment results of the test data. Giza++ with ITG vs. Giza++.

	Recall	**Precision**	**AER**	**F measure**
Giza++ with ITG	0.58	0.87	0.30	0.70
Giza++	0.37	0.87	0.48	0.52

The use of ITG results in significant improvement for recall and F-measure of Giza++ by 56.8% and 34.6% leading to substantial alignment error reduction (37.5%) while precision suffers only slightly (0.1%).

4.3.3 Phrase-level Evaluation

We further evaluated base phrases of the generated bilingual parse trees. We take into consideration the correctness of syntactic label and phrase alignment of a base phrase. Table 15 is how we rated a base phrase produced by our method concerning syntactic label and phrase alignment.

Table 15. Points of phrase-level evaluation.

syntactic label	**phrase alignment**	**point**
O	O	1.0
O	X	0.5
X	O	0.5
X	X	0.0

The first row in Table 15 means that if human judges assess the constituent label and alignment of the generated base phrase are both correct, it will be rated as *correct* (1 point). The second row means that if the syntactic label is correct but alignment is not quite right, human judges will rate the base phrase as *partially correct* (0.5 point). However, if the label is wrongly tagged but the phrase

alignment is right, it will also be rated as *partially correct* (0.5 point). In the worse case, the label and alignment are not quite correct, 0 point is given to that base phrase.

The average score of the base phrases generated by Giza++ with ITG was 0.82, showing that our method produced satisfactory result in constituent label of base phrases and alignments in phrase level.

5. Conclusion and Future Work

Improvements of the proposed method and future researches have presented themselves along the way. Currently, we only focus on CFG with two right-hand-side constituents. Nonetheless, in linguistic sense, it is undesirable to divide the structure of $(NP\ CC\ NP)$ into $(NP\ CC)$ and (NP) or (NP) and $(CC\ NP)$ in that it is an indivisible syntactic-meaningful construct. Therefore, one of our future goals is to incorporate grammar rules with more constituents on the right hand side, such as $NP \rightarrow NP\ CC\ NP$, and their related probabilistic estimations into our model. Moreover, to make the structures of the bilingual parse trees more complete and rational, we would include a meaningful label for target-language words translated into no words in the source and grammar rules with the label in the future. It is also interesting to see how produced bilingual parse trees would influence the performance of the actual decoding process of machine translation and facilitate bilingual phrase extraction.

In conclusion, we have presented a robust method for learning ITG rules which specify the syntactic structures and relations of syntax of two languages involved. The proposed method exploits both lexical and syntax information to derive a structural model of the translation process. At runtime, a bottom-up CYK-styled implementation parses bilingual sentences simultaneously by exploiting trained ITG rules. Experiments show that our model consisting of grammar rules with linguistics-motivated labels and preferences of ordering counterparts in languages produces much more satisfying word alignment results compared with a state-of-the-art word-aligning system.

6. References

1. Andrew B. Clegg and Adrian Shepherd. 2005. “Evaluating and integrating Treebank parsers on a biomedical corpus.” In *Association for Computational Linguistics Workshop on software 2005*.
2. Colin Cherry and Dekang Lin. 2003. A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, volume 1, pages 88-95.
3. David Chiang. 2005. “A hierarchical phrase-based model for statistical machine translation.” In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 263-270.
4. Yuan Ding and Martha Palmer. 2005. “Machine translation using probabilistic synchronous dependency insertion grammars.” In *Proceedings of 43rd Annual Meetings of the ACL*, pages 541-548.
5. Wu Hua, Haifeng Wang, and Zhanyi Liu. 2005. “Alignment model adaptation for domain-specific word alignment.” In *Proceedings of the 43rd Annual Meeting of the ACL,* pages 467-474.

6. I. Dan Melamed. 2003. “Multitext grammars and synchronous parsers.” In Proceedings of the 2003 Meeting of the North American chapter of the Association for Computational Linguistics.
7. Franz Josef Och and Hermann Ney. 2000. “Improved statistical alignment models.” In *Proceedings of the 38th Annual Conference of the Association for Computational Linguistics (ACL-00)*, pages 440-447.
8. F Franz Josef Och, C. Tillmann, and H. Ney. 1999. “Improve alignment models for statistical machine translation.” In 1999 *EMNLP*.
9. Kristina Toutanova, H. Tolga Ilhan and Christopher D. Manning. 2002. “Extentions to HMM-based statistical word alignment models.” In *Proceedings of the Conference on Empirical Methods in Natural Processing Language*.
10. Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. “HMM-based word alignment in statistical translation.” In *Proceedings of the 16th conference on Computational linguistics*, volume 2, pages 836-841
11. Wei Wang, Ming Zhou, Jin-Xia Huang, and Chang-Ning Huang. 2002. “Structure alignment using bilingual chunking.” In *Proceedings of the 19th international conference on Computational linguistics*, volume 1, pages 1-7.
12. Wei Wang and Ming Zhou. “Improving word alignment models using structured monolingual corpora.” In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,* pages 198-205.
13. Ye-Yi Wang. 1998. “Grammar inference and statistical machine translation.” Ph.D. thesis.
14. Dekai Wu. 1997. “Stochastic inversion transduction grammars and bilingual parsing of parallel corpora.” *Computational Linguistics*, 23(3):377-403.
15. Kenji Yamada and Kevin Knight. 2001. “A syntax-based statistical translation model.” In *Proceedings of the 39th Annual Conference of the Association for Computational Linguistics (ACL-01)*.
16. Hao Zhang and Daniel Gildea. 2004. “Syntax-based alignment: supervised or unsupervised?” In *Proceedings of the 20th International Conference on Computational Linguistics*.
17. Hao Zhang and Daniel Gildea. 2005. “Stochastic lexicalized inversion transduction grammar for alignment.” In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 475-482.

使用流暢性改善詞組翻譯的統計式機器翻譯

夏敏翔 張耀升 盧文祥

國立成功大學資訊工程研究所

摘要

Peter F. Brown等人提出統計式的機器翻譯後(Statistical Machine Translation)，目前翻譯的基本單位已由單詞轉變成詞組（Phrase），雖然詞組為本的機器翻譯已經可以達到不錯的效果，但使用者仍在詞組閱讀上遇到不順暢的情形，而目前研究機器翻譯的研究領域很少針對詞組流暢化進行探討。我們觀察到譯者在英中翻譯時，常常會加入一些不存在於英文句中的詞彙使句子能夠更加流暢，如果只是簡單的詞對詞翻譯，無法在中文句子顯示這些額外附加的詞彙在中文句子中。有鑑於此，我們提出流暢化詞組機器翻譯 (Fluent Phrase Machine Translation, FPMT）的機率模型來決定加入詞彙後的中文詞組是否流暢，以及使用語料庫(Corpus)和網路搜尋結果(Search Result)來找出附加的詞彙。實驗結果顯示，我們提出的流暢化詞組機器翻譯模型得到的中文詞組，其效能比IBM Model 4的方法佳，可以有效的補回缺少的詞彙，在人工評估上，更顯示我們的方法確實可以提升翻譯的流暢化。

1. 簡介

Brown[2]在1993年提出訊號源通道方法(Noisy-Channel Approach)來完成機器翻譯(Machine Translation)後，現在對於機器翻譯的研究幾乎都是使用統計式的方法，稱為統計式機器翻譯(Statistical Machine Translation)。現在有很多研究都是使用類似的方法，但是使用這類方法的翻譯模型卻有一些相同的問題—以詞為本(Word-based)，因此不管在重新排序或是翻譯階段的會遭遇困難。在重新排序方面，如果以詞為最小單位，對於多詞所組成的詞組，在重新排序後可能會發生錯誤[20]。同樣在翻譯方面，由多個詞所組成的詞組也容易被翻譯錯誤。目前有許多研究都使用詞組為翻譯單位的方法，實驗也證明以詞組為基本單位可以有效的提升結果[3][4][8][13][14][16][20][22][23]。然而，翻譯所得到的中文詞組雖然皆可對應於英文詞組，但是卻會造成翻譯的結果不是很流暢，所以為了增加詞組翻譯的流暢性，譯者都會加入一些不存在英文詞組中詞彙，由表一顯示，有加入額外的詞彙，詞組可以比較通順，所以機器翻譯不是只有單詞的翻譯，還必須補足詞彙使詞組或句子流暢。

機器翻譯有許多種方法，例如基於規則(Rule-based)的機器翻譯、基於實例(Example-based)的翻譯方法和統計式(Statistic-based)機器翻譯。其中，統計式機器翻譯為目前最常用的一種方法，目前有很多研究[14][11][13]，以詞組為本 (Phrase-based) 的方法得到的翻譯品質都比以詞為本的方法要來的好，而我們也將採用統計的方法來完成我們詞組的機器翻譯模型。大部分機器翻譯流程不外乎就是重新排序以及單詞翻譯，但是很少有學者研究翻譯之後的流暢化。翻譯後的句子不一定是通順的，因此在翻譯時，可能需要加入一些詞彙使得句子能更為流暢，所以我們嘗試將這類的字詞可以補回翻譯的結果中。雖然Brown已有提出類似方法，他們在翻譯過程產生一些空字串，這些空詞串可以產生一些不存在英文句中的翻譯詞彙，使翻譯後的結果流暢許多。但是我們覺得，大部份額外加入的詞彙應該是翻譯後決定比較合適，因此我們提出流暢化詞組的機器翻譯(Fluent Phrase-based Machine Translation, FPMT)，對翻譯的詞組再做進一步的詞彙增補，根據我們的實驗結果，我們提出的流暢化詞組機器翻譯模型可以有效的補回缺少的詞彙，其效能比IBM Model 4的方法好，在人工評估上，更顯示我們的方法確實可以提升翻譯的流暢化。

本論文第一節描述詞組的機器翻譯相關工作與一些問題。第二節將描述有關統計式機器翻譯的相關研究與文獻。第三節為我們所提出的詞組翻譯及兩種不同類型的流暢化方法，第四節透過實驗分析，我們提出的方法在詞組內部流暢化得到有效的改善，加入的詞彙確實比Brown提出的插入空字串來的好。第五節是結論以及未來研究方向。

表一、翻譯比較

Translation Methods \ English Phrase	asthma guidelines
Manual Translation	氣喘**治療**指南
Google (http://www.google.com/language_tools?hl=en)	哮喘指南
IBM (http://www-306.ibm.com/software/pervasive/tech/demos/translation.shtml)	氣喘指南
SYSTRAN (http://www.systransoft.com/index.html)	哮喘指南
MojoLingo (http://text.mojolingo.com.tw/)	哮喘指導方針

2. 相關文獻討論

Brown [2]等人首先提出了統計式的機器翻譯方法，有效的從語料庫中獲得所需要的資訊，是目前翻譯研究中最常用的方法。Watanabe[20]等模仿Brown的機器翻譯步驟，但改變了排序時的最小單位，使用數個詞來當成做小單位，稱為詞塊 (Chunk)。Marcu和Wong[13]使用了聯合機率模型(Joint Probability Model)來取代條件機率，並且使語彙模型可以包含詞組的翻譯，使得翻譯上更為精確。Chiang [3]從相對應的詞組結構來學習翻譯的規則，使用正規語言的方法對來源語跟目標語同時處理以達到翻譯的目標。此外還有很多的研究[8][23]都使用詞組為單位，例如 Och [16]跟Yamada [22]等，證明了以詞組為翻譯單位可以有效的提升結果。所以現在的研究幾乎都是使用詞組為基本的單位，也證實了以單詞為基本單位是不夠的。

以語法結構為基礎的統計式機器翻譯跟前面的方法最大的不同是，這類方法會使用語法剖析器(Parser)來獲得語法的剖析樹(Parsing Tree)，再把剖析樹當成輸入進而完成翻譯的動作。Yamada and Knight [21][22]改變了一般的翻譯系統使用字串為輸入的方法，他們使用剖析器，將來源語的句子轉化成樹狀結構，藉此獲得更多的語言學上的資訊。訓練的過程中，獲得所需要的相關資訊並建立了三種表格，分別是重新排序機率表(Reordering Probability Table)、插入字串機率表(Insertion Probability Table)跟翻譯機率表(Translation Probability Table)。當輸入剖析樹時，然後對剖析樹執行重新排序、插入字串和翻譯，然後就輸出目標字串，翻譯的過程中查詢先前建立的三種表格。Ding和Palmer[6]同樣的，使用了文法剖析器來獲取樹狀結構。他們提出同步相依插入文法規則(Synchronous Dependency Insertion Grammars)來完成翻譯，也就是說，當他們對來源語的結構樹執行一個文法規則時，同時，根據相對應的文法規則來產生目標語的樹狀結構。翻譯過程所需要的文法規則，則是在翻譯前先從雙語語料庫中學習，學習的方法是使用他們所提出來的文法規則歸納演算法(Grammar Induction Algorithm)，此演算法簡單的說就是根據某一種語言的結構樹，然後利用詞典來分解相對應的翻譯樹獲得文法規則。以上兩篇都充分的顯示出，在翻譯的過程中使用了語法剖析器都可以提升翻譯的效能。

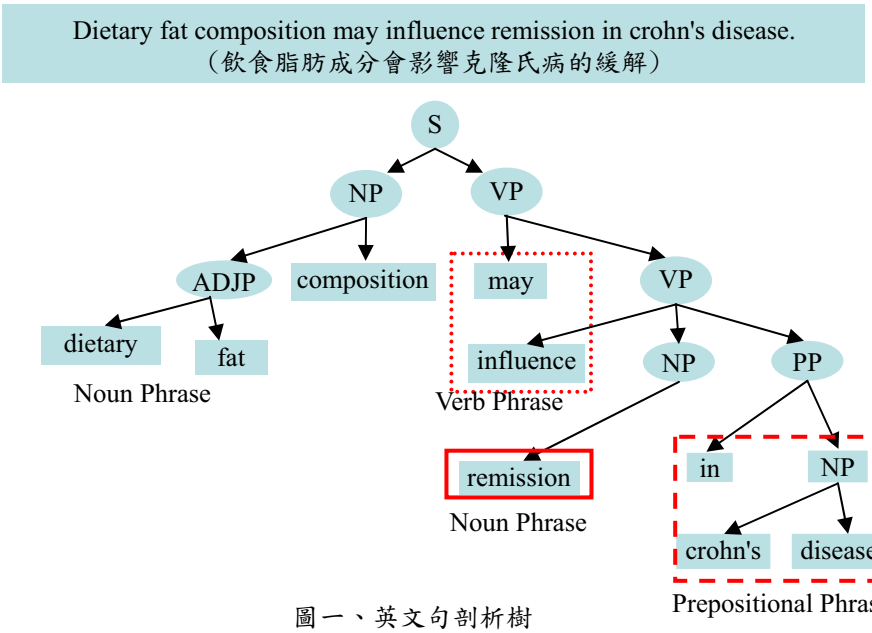
我們在本論文提出的方法也是統計式的機器翻譯，為了增加翻譯的品質，我們使用詞組為翻譯單位。Koehn[12] 提到，如果詞組結構裡包含的字數量太多時，翻譯的效果反而會下降，Och [16]跟Venugopal [19]都有不同的方法來獲得所需要的詞組配對，但是他們使用統計的方法來得到詞組，很可能會使詞組結構過大，因此我們必須找出另一種方法來獲得適當詞組結構。在以句法結構為基礎的統計式機器翻譯中顯示，如果使用剖析器來得到語法剖析樹當成輸入，可以增加翻譯的品質，因此我們使用剖析器來獲得句法剖析樹，再從句法剖析樹中獲得我們需要的詞組結構。我們也發現，如果使用此方法獲得的詞組結構，其包含的單詞數量並不會太多，很適合當成翻譯的基本單位。在重新排序方面，我們也是使用詞組為基本單位；在翻譯方面，也可以學習單詞在不同詞組時的不同翻譯，這樣可以提升翻譯的準確率。模型的推導方面，Brown等人使用貝氏定理將$P(T|S)$代換成$P(S|T)P(T)$，因為$P(T)$可以確保所得到的目標語言是可以符合文法的語句，但是Foster [7]提出直接計算$P(T|S)$的翻譯方式，而且在Och和Ney [15]的實驗結果也証實，如果直接計算最佳化的$P(T|S)$所得到的翻譯效果，會跟使用$P(S|T)P(T)$所得的效果差不多，甚至會更好，所以我們提出的統計式機器翻譯的模型，也是由$P(T|S)$開始推導的，在翻譯過程中使用詞組式語言模型來解決詞義消歧(Word Sense Disambiguation)的問題。本論文最重要的貢獻是提出流暢化詞組機器翻譯機率模型來解決詞組翻譯流暢化問題，現在很多研究都只有處理詞組翻譯，我

們覺得翻譯過後的結果還需要進一步的處理，也就是加入詞彙使結果更加流暢，不同於Brown提出的空字串產生方法，我們覺得應該是翻譯後才決定該加入哪些詞彙。詳細方法將會在第三節中描述。

3. 流暢化的詞組機器翻譯

3.1 詞組擷取

先前已經有很多的研究證實，使用詞組為單位的機器翻譯會比只使用單詞為單位的機器翻譯來的好，所以我們處理英文句的第一步就是必須先取得英文詞組，也就是要將輸入的英文句分割成詞組形式。我們使用史丹福英文語法剖析器（Stanford Parser）[18]來獲得需要的詞組。由圖一可以看見，我們使用一些觀察到的規則，可以從剖析樹（Parser Tree）中輕易的獲得該英文句的詞組，可以清楚看到有許多不同種類的詞組，例如名詞詞組、動詞詞組[1]和介係詞詞組。



圖一、英文句剖析樹

Ding and Palmer[6]提出使用不同語言的剖析器來得到兩棵結構不同的剖析樹，但是在中文方面，目前還缺乏良好的剖析器，而且不同語言的剖析樹因為結構上的差異，很難用相同的文法規則來處理不同結構的剖析樹，所以我們只使用剖析樹中的詞組資訊，並未使用其語法結構。

3.2 詞組翻譯

3.2.1 問題描述

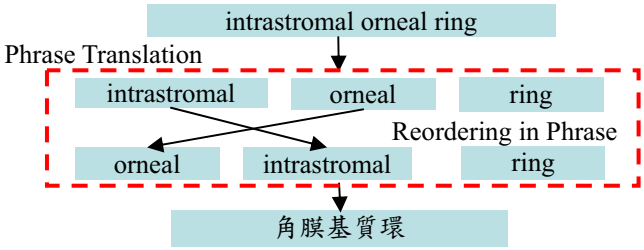
詞組翻譯過程中，並不只是把英文單詞翻譯成中文詞，還必須考慮單詞在詞組中的位置。因此在詞組翻譯的步驟中，我們首先需要完成詞組內的單詞的排序，然後根據所在的位置找出相對應的翻譯。我們知道英文單詞翻譯時，常常會有意義混淆（Ambiguity）現象，例如英文單字"bank"，可以翻譯成"銀行"也可以翻譯成"堤岸"，所以翻譯時必須決定哪一個翻譯才是正確的，也就是詞義消歧（Word Sense Disambiguation）問題，為了解決詞義消歧問題，我們使用詞組基本的語言模型(Language Model)，便可以根據前一個詞來決定最合適的翻譯。

在單詞翻譯前，我們必須先決定每一個單詞在該詞組的位置，由於一些中文詞在詞組中跟英文詞的順序有所不同，若按照原本的順序翻譯，常會發生詞組翻譯錯誤，以圖二為例，"intrastromal orneal ring"在翻譯成中文時，將"intrastromal"以及"orneal"互換後，才會是正確的翻譯，所以作單詞翻譯之前，我們必須先考慮到單詞在詞組中的位置，這樣翻譯成中文時，才會比較通順。當詞組中單詞的位置都決定後，最後步驟為將英文單詞翻成中文，但是並非把英文單詞

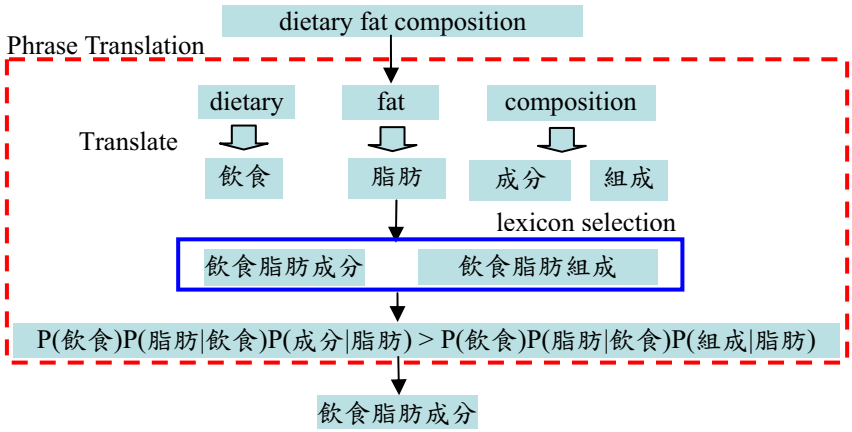
[1] 動詞詞組(Verb Phrase) 在本論文中僅考慮verb, adverb + verb, verb +adverb, 或auxiliary + verb 四種形式。

直接翻譯成中文即可，因為英文翻中文時，常常有意義不明確的情況發生，所以在翻譯時，只要選錯詞，將會造成很大的差異。下圖三，英文單字"composition"在文件中可以翻成"組成"以及"成分"，如果翻譯成"組成"，我們可以發現翻譯後的結果不正確，而為了解決這種詞彙選擇(Lexicon Selection)的問題，我們使用詞組的語言模型來決定該選哪一個詞彙來當成正確的翻譯。

由圖三可以看見，在我們的文件中，"脂肪成份"的機率比"脂肪組成"的機率還大，因此當英文單詞"composition"的前一個文字是"脂肪"時，我們便能知道它的翻譯應該是"成分"會比"組成"合適。所以我們只用簡單的語言模型方法，就可以解決很多的詞彙選擇問題。經過以上兩個步驟，我們便可以將英文詞組翻譯成中文詞組，除了詞與詞之間的順序都正確，還可以選擇最合適的詞彙來當成英文的翻譯。



圖二、詞組內單詞位置差異圖



圖三、單詞翻譯時之詞彙選擇問題

3.2.2 詞組翻譯模型

我們提出詞組的機器翻譯模型跟Brown最大的不同點，在於其使用貝氏定理$P(T|S)$代換成$P(S|T)P(T)$，而我們是直接由$P(T|S)$開始推導，所以我們初始的公式是：

$$C^* = \arg\max_c P(C \mid E) \tag{1}$$

我們使用的翻譯方法首先必須知道英文的詞組結構，所以引入參數S代表我們需要的詞組結構。引入S後，公式可以推導成：

$$C^* = \arg\max_c \sum_S P(S \mid E) P(C \mid S, E) \tag{2}$$

公式(2)中，我們透過史丹福英文語法剖析器（Stanford Parser）以獲得英文詞組結構，而假設此為最佳的一組詞組結構S，所以將$\sum$去除。獲得英文詞組結構後，接下來就是要將這些詞組結構重新排序，因此我們引入參數R表示每一個詞組根據中文結構應該重新排序的對應位置：

$$C^* = \arg\max_c P(S \mid E)P(C \mid S,E)$$
$$= \arg\max_c P(S \mid E)\sum_R P(R \mid S,E)P(C \mid R,S,E) \quad (3)$$

從公式(3)中，我們選擇最佳的一組R，所以將∑去除，我們的模型可以分成三個主要部份，分別為$P(S|E)$、$P(R|S,E)$以及$P(C|R,S,E)$，所以我們將機器翻譯模型可分成三大部份來解釋我們的翻譯模型，分別為詞組產生模型$P(S|E)$、重新排序模型$P(R|E,S)$以及詞組翻譯模型$P(C|R,S,E)$。本篇論文只針對詞組翻譯模型深入探討。

詞組翻譯時，以一個詞組為單位然後翻譯裡面的英文單詞，因此只考慮單詞在詞組中的適當位置，因為英文詞組翻譯成中文時，單詞的位置也會有所變化。詞組翻譯只是針對某一個詞組完成翻譯，並沒有考慮它在中文句子結構中的位置，而翻譯的結果也沒有受到前後詞組的影響，所以在完成詞組翻譯時，我們沒有考慮詞組的位置，根據公式(3)的詞組翻譯模型，我們可以將R省略，因此我們的詞組翻譯模型為：

$$P(C \mid R,S,E) = P(C \mid S,E) \quad (4)$$

不考慮位置R的情況下，詞組翻譯的機率即求$P(C \mid S,E)$，其中C是中文句、E是英文句，S是英文的詞組結構。我們發現，如果要從英文句E直接翻譯成中文句C是一件很困難的事，所以我們利用剖析器得到的詞組結構，將句子翻譯改變成詞組對詞組的翻譯並假設詞組的翻譯是獨立的，而非使用傳統語法結構的分析，故最後使用詞組翻譯如公式(5)：

$$P(C \mid S,E) \cong P(C \mid E)$$
$$= P(sc_1, sc_2, \cdots, sc_m \mid se_1, se_2, \cdots, se_m) \quad (5)$$
$$= \prod_{\substack{sc \in C \\ se \in E}} P(sc \mid se)$$

sc為組成中文句子C的中文詞組，se為輸入英文句子E至史丹福英文語法剖析器得到的英文詞組，最後句子之間的翻譯變成詞組之間的翻譯，當英文詞組都翻譯成中文詞組後，在獲得的詞組最佳排序，就可以完全翻譯成中文句。

假始直接使用$P(sc|se)$計算，所得到的中文詞組不一定是合適的，因為英文翻譯中文會有詞義消歧的問題，所以直接翻譯的結果並不一定是適當的中文詞組。透過貝氏定理轉換$P(sc|se)$，公式變成：

$$P(se \mid sc) = \frac{P(se \mid sc)P(sc)}{P(se)} \quad (6)$$

其中$P(sc)$為利用語言模型中的雙連詞(Bigram)機率來解決詞義消歧的問題，因此使用$\frac{P(se \mid sc)P(sc)}{P(se)}$可以確保得到的中文詞組是比較適當的。我們最終目的不是將英文詞組翻譯成中文詞組，而是選擇一個合適的中文詞組sc，以最接近我們輸入的英文詞組se。公式(6)中的$P(se)$為英文詞組機率，對於可能被選中的中文詞組而言，都是一樣的，所以我們將忽略公式中的$P(se)$。最後我們想要取得最佳的中文詞組sc，機率公式取推導成：

$$P(C \mid S,E) \propto \prod_{\substack{sc \in C \\ se \in E}} P(se \mid sc)P(sc) \quad (7)$$

接著我們定義如何使用$P(se|sc)$選擇出最佳的中文詞組sc。為了得到最佳英文詞組se的中文詞組sc翻譯，我們使用兩個機率公式進行推導。（一）單詞翻譯機率$P(se_{a_i} \mid sc_i)$，其中sc_i是中文詞組sc中的第i個中文詞，se_{a_i}則是sc_i在對應位置a的情況下所對應到的英文詞，以及（二）位置對應機率$P(a \mid l,m)$，其中l是中文詞組的長度，m是英文詞組的長度，a就是在該長度下，中文與英文對應的關係。根據上面的敘述，我們可以將$P(se|sc)$推導成：

$$P(se \mid sc) = P(a_1 \ldots a_l \mid l)\prod_i P(se_{a_i} \mid sc_i) \quad (8)$$

因為中文詞組是由英文詞組加上詞典得到的，所以英文詞組與中文詞組中的詞是屬於一對一

的關係，也就是說英文詞組的長度跟中文詞組的長度是相同的，因此我們只使用參數l來代表詞組的長度。根據上述，我們的詞組翻譯模型就可以推導成：

$$P(C \mid S,E) \propto \prod_{\substack{sc\in C \\ se\in E}} \{P(a_1 \ldots a_l \mid l) \prod_i P(se_{a_i} \mid sc_i) P(sc)\} \tag{9}$$

我們以英文詞組< intrastromal corneal ring>要翻譯成中文詞組<角膜基質環>為例，最後詞組翻譯模型如下：

$$\begin{aligned} & P(se \mid sc)P(sc) \\ &= P(\text{intrastromal corneal ring} \mid \text{角膜基質環}) \times P(\text{角膜基質環}) \\ &= P(213 \mid 3)P(\text{corneal} \mid \text{角膜}) \times P(\text{intrastromal} \mid \text{基質}) \times P(\text{ring} \mid \text{環}) \\ &\quad \times P(\text{角膜})P(\text{基質} \mid \text{角膜}) \times P(\text{環} \mid \text{基質}) \end{aligned}$$

英文詞corneal透過詞典得到某一中文翻譯詞為<角膜>，從之前的訓練樣本中可以計算得到P(corneal |角膜)的機率值，依此類推皆可得到P(intrastromal |基質)及P(ring |環)。因為<角膜>、<基質>及<環>中文字詞，依不同的排列組合而有不同的中文詞組翻譯，故我們計算各中文字詞組合的位置機率，其機率值在訓練樣本已有紀錄，最後再透過雙連詞(Bigram)機率加以計算此中文翻譯詞組是否存在的機率。

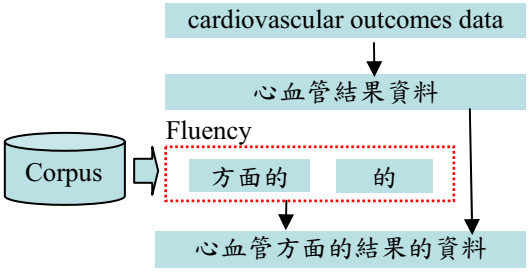
3.3 流暢化

3.3.1 問題描述

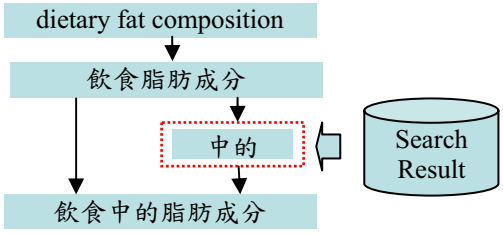
經過3.2詞組翻譯模型處理，可以獲得中文翻譯的詞組，此中文詞組為英文詞組中的單詞重新排序、個別翻譯得之，基本上都可以大略了解此中文詞組翻譯。但是若可以在中文詞組再加入一些詞彙，便可以使得翻譯後的中文詞組更容易閱讀。英文翻譯成中文的過程中，譯者通常會加入一些額外詞彙使得閱讀時更為流暢，然而這些詞彙一般並不會出現英文詞組中，而是為了詞組的流暢性才加入，所以機器翻譯不僅僅是要將詞彙翻譯正確，也要使翻譯出來的詞組可以讓讀者閱讀流暢，所以我們在完成機器翻譯後提出了一種流暢化的方法，嘗試將不存在英文詞組中的字詞，可以順利地加入翻譯後的中文詞組中。所以在本論文我們首先提出中文詞組內部附加詞方法來解決翻譯流暢化問題。

我們使用兩種詞組內部的附加詞方法：語料庫學習以及網路搜尋結果學習。使用語料庫學習方面，由 圖四例子可以看見，英文詞組"cardiovascular outcomes data"如果按照字詞翻譯所獲得的中文詞組為"心血管結果資料"，此詞組雖然可以知道其所要表達的意義，若可以再加上額外的字詞，如"方面"以及"的"，則"心血管方面的結果的資料"比起"心血管結果資料"更容易使人了解其意義。

然而，使用語料庫學習附加詞有其缺陷，由於語料庫大小的限制，只要語料庫中沒學過的詞彙我們將無法找出合適的附加詞，為了補足此缺陷，我們利用網路搜尋結果來補強。圖五的英文詞組"dietary fat composition"翻譯成中文詞組為"飲食脂肪成分"，當使用語料庫時，我們發現並沒有適合的詞可以加入，因此利用網路資源嘗試找出附加詞。使用網路搜尋結果，我們發現在"飲食"以及"脂肪"之間可以加入中文字詞"中的"，而且加入後的詞組"飲食中的脂肪成分"比未加入的詞組"飲食脂肪成分"更為流暢。



圖四、詞組內加入詞彙(使用語料庫)



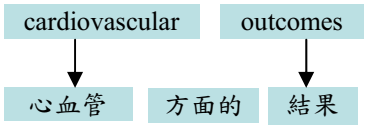
圖五、詞組內加入詞彙(使用網路資源)

3.3.2 流暢化方法

下面將介紹我們利用機率概念推導兩個詞組內部的流暢化計算方法。在使用語料庫方面，我們提出一個簡單的機率公式，利用邏輯變數a檢查可能附加的流暢詞w是否允許加入於中文詞與詞之間(c_i, c_{i-1})，公式如下：

$$P(a \mid c_i, c_{i-1}) = \sum_w P(a \mid c_i, c_{i-1}, w) \times P(w \mid c_i, c_{i-1}) \tag{10}$$

$P(a|c_i, c_{i-1}, w)$計算兩詞之間c_i、c_{i-1}可以插入附加詞w的機率，$P(w|c_i, c_{i-1})$決定兩詞之間c_i、c_{i-1}是否可以加入一個附加詞w。由圖六中為某訓練樣本，我們可以訓練到cardiovascular及outcomes可加入的附加詞<方面的>之機率值，即 $P(\text{方面的} \mid outcomes, cardiovascular)$ 。因為是由英文翻譯中文，故由英文反推找出附加詞會比較符合原本的涵意，再加上流暢化步驟在翻譯之後，所以每一中文字詞c_i即對應一英文詞e_i，故$P(e_i|c_i)= 1$，最後公式推導如：



圖六、附加詞訓練樣本範例

$$P(a \mid c_i, c_{i-1}) = P(a \mid c_i, c_{i-1}, w) \times \sum_w P(w \mid e_i, e_{i-1}) \tag{11}$$

$P(a|c_i, c_{i-1})$就可以使用$P(w|e_i, e_{i-1})$以及$P(a|c_i, c_{i-1}, w)$來計算，然後使用門檻值來決定是否需要加入額外的詞彙。$P(w|e_i, e_{i-1})$用來計算在語料庫中有哪一些詞彙常常出現在單詞c_i, c_{i-1}之間；$P(a|c_i, c_{i-1}, w)$用來計算在c_i, c_{i-1}之間加入w後的出現機率，在計算$P(a|c_i, c_{i-1}, w)$時，我們是利用網路搜尋引擎Google的相關網頁搜尋數目來估計，估計的方法如下：

$$\begin{aligned} P(a \mid c_i, c_{i-1}, w) &= \frac{P(a, c_i, c_{i-1}, w)}{P(c_i, c_{i-1}, w)} = \frac{count(a, c_i, c_{i-1}, w) \big/ N}{count(c_i, c_{i-1}, w) \big/ N} \\ &= \frac{count(a, c_i, c_{i-1}, w)}{count(c_i, c_{i-1}, w)} \end{aligned} \tag{12}$$

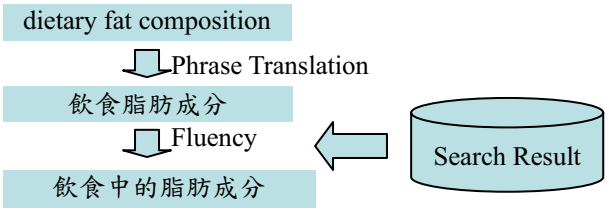
其中N代表Google的所有網頁數目，$count(a, c_i, c_{i-1}, w)$代表將$w$加入$c_i$、$c_{i-1}$中間所形成相鄰的字串"$c_i w c_{i-1}$"(即tri-gram)可以搜尋到的網頁筆數，$count(c_i, c_{i-1}, w)$則是$c_i$、$c_{i-1}$以及$w$三個詞同時出現的網頁筆數，故我們可以得到出現$c_i$、$c_{i-1}$以及$w$三個詞的網頁中其相鄰字串tri-gram所佔的比例。以英文詞組"cardiovascular outcomes data"為例子，我們翻譯出的中文詞組為"心血管 結果 資料"，使用公式(11)計算"心血管"以及"結果"之中是否可加入附加詞：

$$\begin{aligned} P(a \mid \text{結果,心血管}) &= P(\text{方面的} \mid outcomes, cardiovasaular) \times P(a \mid \text{結果,心血管,方面的}) \\ &= P(\text{方面的} \mid outcomes, cardiovasaular) \times \frac{count(\text{"心血管方面的結果"})}{count(\text{"心血管","方面的","結果"})} \end{aligned}$$

當 $P(a \mid \text{結果,心血管})$ 的值大於門檻值時，我們就將附加詞"方面的"加入在"心血管"以及"結果"之中，同理，"結果"以及"資料"中間也可以加入"的"，加入附加詞後的中文詞組為"心血管方面的結果的資料"，比起未加入詞彙時更為流暢。

在使用網路資源方面，因為語料庫的大小限制，我們無法確實的將詞與詞之間的附加詞都找出來，所以我們借用網路上的搜尋結果來幫助我們取得這些附加詞彙。圖七中的英文詞組"dietary fat composition"翻譯成中文後為"飲食脂肪成分"，使用網路搜尋結果可以發現"飲食"以及"脂肪"之間會加入"中的"，則中文詞組"飲食中的脂肪成分"將較為通順，但是在我們的語料庫中"飲食"以及"脂肪"之間並沒有任何辭彙，如果使用語料庫的方法將無法加入任何辭彙，所以我們另外使用網路搜尋結果來幫我們找出合適的辭彙。

圖八是使用網路搜尋結果流暢化步驟的流程圖，最主要的目的就是要找出兩個關鍵中文詞（C_1,C_2）中間可以加入哪些附加詞，使得看起來更為通順。首先我們先把兩個中文關鍵詞送到Google搜尋引擎，取回搜尋結果，然後使用Chien[4]所提出的PAT-Tree-based 的關鍵詞擷取方法找出最常出現的詞彙，然後可以根據這些詞彙再使用關鍵詞找出候選附加詞，再將雙方的候選詞一起送至Google搜尋引擎，算出他的頻率，這樣我們就可以知道在關鍵詞之間，可以加上哪一個附加詞彙可以使他們更為順暢。

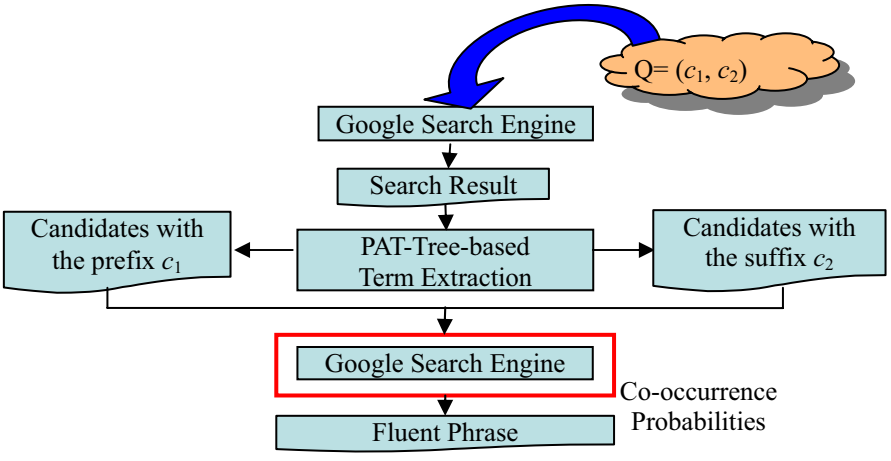


圖七、使用網路搜尋結果加入詞彙

我們使用中文的關鍵字"飲食"(c_1)以及"脂肪"(c_2)作為例子，首先我們使用關鍵字"飲食 脂肪"送至Google取回搜尋結果，然將搜尋結果使用PAT-Tree based關鍵詞擷取方法找出頻率較高的中文詞，再利用關鍵詞"飲食"以及"脂肪"從這些中文詞中找出相關的詞，例如"飲食"(c_1)可以找出產生前綴詞彙c_1的連接詞串"飲食文化"、"飲食中"...等的可能附加候選詞(x_1)如"文化"、"中"，而"脂肪"(c_2)可以找出產生前綴詞彙c_2的連接詞串"飽和脂肪"、"的脂肪"...等的可能附加候選詞(x_2)如"飽和"、"的"。然後將x_1, x_2連接在c_1以及c_2內產生一新詞串，使用公式(13)來計算他們配對後新詞串的機率，如此一來我們就可以知道在"飲食"以及"脂肪"中間可以插入"中" (x_1) "的" (x_2)形成"飲食中的脂肪"一詞就較為通順。我們使用的公式如下：

$$P(x_1,x_2,c_1,c_2)=\frac{P(c_1,x_1,x_2,c_2)}{P(c_1,c_2)}=\frac{count(c_1,x_1,x_2,c_2)}{count(c_1,c_2)} \tag{13}$$

其中我們限制附加詞x_1和x_2就是夾在c_1以及c_2 之間的辭彙，而x_1和x_2可能會有兩種不同的情況，當x_1以及x_2相同時也就是x_1以及x_2是重疊（overlap）的狀況，我們就只考慮x_1即可；若x_1和x_2不同時，將x_1以及x_2連接（concatenate）成一個詞彙，再使用與公式(12)相同的方法，我們也是使用Google網路搜尋引擎所找到的網頁數來估算機率值，然後根據機率值大小再決定是否需要加入詞彙x_1和x_2。



圖八、使用網路搜尋結果流暢化流程圖

4. 實驗

本章節將評估詞組翻譯和流暢化的效能，並且比較我們的方法和IBM Model 4（簡寫IBM4）翻譯模型的結果。首先介紹我們所使用的資料、比較的翻譯模型以及評估方式，接下來就是我們的實驗數據分析。

4.1 實驗資料

我們的訓練語料庫是從國際厚生健康園區網站[1]經由人工收集的雙語語料庫，總共有18752句中英文配對句子，所包含的名詞詞組、動詞詞組以及介系詞詞組的相關數目如表二所表示。訓練翻譯模型前，我們使用GIZA++[9]工具進行訓練，以得到中英文詞組對應的統計式詞典檔。測試資料分為兩部份，分別是內部測試(Inside-test)，其為使用訓練語料庫中的資料做測試；以及外部測試(Outside-test)，此為使用非訓練語料庫內的資料來測試。測試方法為詞組翻譯。測試資料皆為隨機挑選，使用數目分別如表三和表四所列。表四中的外部測試資料各類型詞組是從100句英文中所得來的，本研究是我們在機器翻譯的起步研究，因為準備外部測試資料需要時間以人工方式進行中英文句子的對應，所以目前尚未取得大量的測試資料。

首先我們使用ISI(Information Sciences Institute)自然語言處理小組[10]所提供的解碼器對IBM4進行詞組翻譯評估。評估的方法則使用BLEU[17]，此為一種使用N-Gram的方式來評估機器翻譯的結果效能，使用自動評估的方法比人工評估更加快速方便，由於目前沒有針對流暢化翻譯的自動評估方法，我們則利用人工評分的方式，加以分析。

表二、訓練語料庫中詞組類型及數目

	Number
Noun Phrase	20820
Verb Phrase	10603
Prepositional Phrase	20421

表三、內部測試資料中詞組類型及數目

	Number
Noun Phrase	1000
Verb Phrase	1000
Prepositional Phrase	1000

表四、外部測試資料中詞組類型及數目

	Number
Noun Phrase	170
Verb Phrase	97
Prepositional Phrase	123

4.2 實驗結果

詞組翻譯的實驗的結果分為內部資料測試以及外部資料測試。

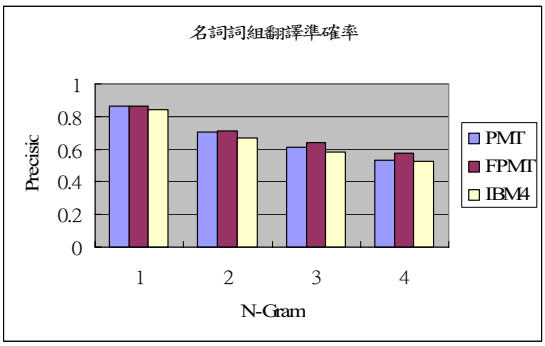
4.2.1 內部測試

我們使用表三中提到的內部資料來做測試。圖九是名詞詞組的N-Gram Precision比較圖，圖十、圖十一分別是動詞詞組以及介系詞詞組的比較圖。PMT為我們的詞組機器翻譯方法，FPMT為我們提出的流暢化詞組機器翻譯方法。圖九是以名詞詞組為測試資料的實驗結果，在1-Gram Precision的數值幾乎一樣，所以我們加入的詞彙可以將英文句子所沒出現的單詞補回中文詞組，而加入這類詞彙使得中文詞組更加順暢，由於其並沒有使準確率下降，故所加入的詞彙幾乎都是正確的。在動詞詞組以及介系詞詞組中，準確率的差別比較小，我們發現在名詞詞組中，需要加詞的情況比動詞詞組以及介系詞詞組來的多，所以4-Gram Precision的數值表現最好，乃因加入附加詞的位置是正確。

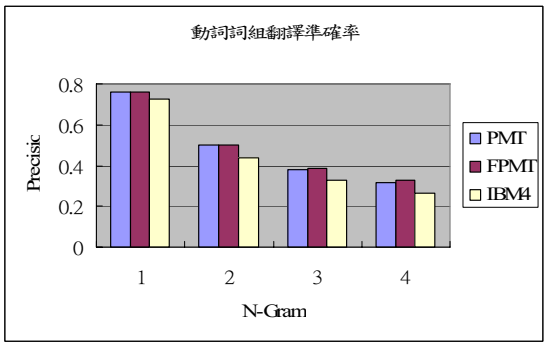
接下來我們為FPMT跟IBM4的分析比較。從圖九可以發現，在名詞詞組中我們所使用的詞組翻譯模型所得到的結果會比IBM4所提出的方法好。從表五的名詞翻譯比較，在單詞翻譯方面，FPMT可以翻譯出較合適的中文詞，"pediatric"在此詞組中，翻譯成"兒童"確實比"小兒科"來的好，而FPMT同時也加入詞彙"的"，使的詞組翻譯除了正確且更加流暢；而表六也顯示，在名詞

詞組"prospective randomized studies"使用FPMT可以正確的將詞彙"的"加入到正確位子，完全符合參考答案，雖然IBM4的系統可以補回詞彙，但是很明顯的位置發生錯誤。

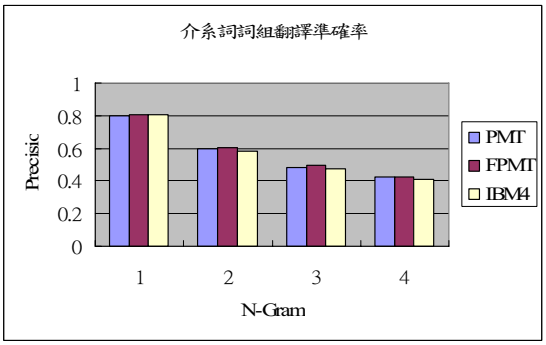
1-Gram-Precision效能評估顯示我們的FPMT翻譯方法在單詞翻譯的準確率確實比IBM4的表現較佳，而其他N-Gram-Precision，除了詞組內部單詞排序的結果表現較佳外，而加入流暢化方法的插入附加詞步驟之後，準確率更有些許的提升。從圖十一中可以發現，介系詞詞組的準確率並沒有提升很多，這是因為在翻譯介系詞詞組時，由於我們的方法一開始便將介系詞視為虛詞(Stopword)處理，所以在詞組翻譯過程中，介系詞會被省略，最後在作流暢化時，會將介系詞當成是詞組與詞組之間的附加詞，而IBM4會將介系詞直接翻譯，如表七所示，導致我們的方法在介系詞詞組的效果較不理想。圖十長度1至4的動詞詞組準確率的差距較大，平均而言FPMT效能較好，我們進一步分析，根據3.1節定義的動詞詞組，其長度不會比名詞詞組或介係詞詞組長，所以能形成的4-Gram數量較少。但是整體而言，翻譯以及排序的結果都可以比IBM4的好。



圖九、名詞詞組翻譯比較圖(內部資料測試)



圖十、動詞詞組翻譯比較圖(內部資料測試)



圖十一、介系詞詞組翻譯比較圖(內部資料測試)

表五、名詞詞組翻譯比較(1)

English Phrase / Translation Method	pediatric mental health problems
Reference Translation	兒童 的 精神 健康 問題
FPMT	兒童 的 精神 健康 問題
IBM4	小兒科 精神 健康 問題

表六、名詞詞組翻譯比較(2)

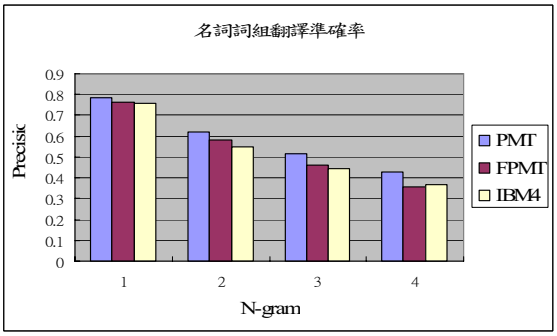
English Phrase / Translation Method	Prospective randomized studies
Reference Translation	前瞻性 的 隨機 研究
FPMT	前瞻性 的 隨機 研究
IBM4	的 前瞻性 隨機 研究

七、介系詞詞組翻譯比較表

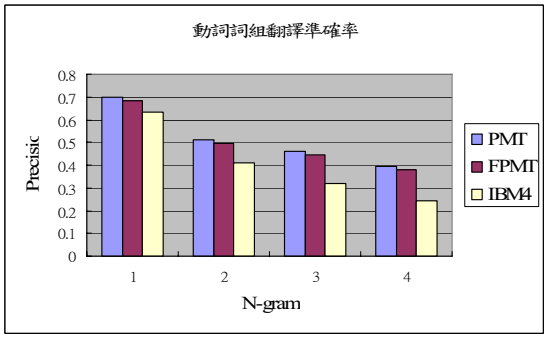
English Phrase / Translation Method	In the polyp prevention trial
Reference Translation	在 息肉 預防 試驗
FPMT	息肉 預防 試驗
IBM4	在 息肉 預防 試驗

4.2.2 外部測試

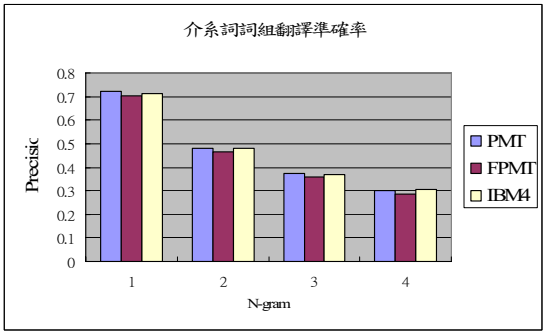
這一小節我們從訓練資料外的文件找出100句中英文配對的句子，並分割其詞組，此分割擷取到的詞組來當成我們的外部測試資料，其資料筆數如表四所示。圖十二是名詞詞組的評估結果，我們發現PMT優於IBM4，而FPMT卻不及IBM4的表現佳，原因有可能是加入的詞彙破壞了原本N-Gram的結構，使得準確率下降，並非表示加入錯誤詞彙，而BLEU是一種以N-gram字串比對的計分公式，當我們多加入的詞彙可能使詞組流暢，但是詞彙如果不在參考答案中出現，反而會使得準確率降低，這也說明BLEU可能無法評比加入詞彙的句子是否流暢。我們將在4.3節針對流暢化的問題予以討論。圖十三為動詞詞組的比較結果，從圖中可得知我們使用的詞組翻譯方法在動詞詞組也有不錯的效果。在動詞詞組領先的幅度比較大，因根據本論文定義的動詞詞組長度一般比較短，所以達到4-Gram的詞組數量非常的少，如此只要有一兩句的4-Gram翻譯的比較好，即可能將差距拉大。在名詞詞組以及介系詞詞組的結果就不會有很大的差距，因為名詞詞組以及介系詞詞組的平均長度皆比動詞詞組長，故可得到的4-Gram數量會比較多，不會因少數的翻譯結果導致影響整體評分。圖十四是介系詞詞組的評估結果，我們的詞組翻譯結果幾乎跟IBM4的結果相同，主要是因為我們的詞組翻譯方法並沒有將介系詞翻譯，例如介系詞"of"可以翻成中文"的"以及"in"翻成"在"等等，而IBM4的方法有翻譯介系詞。雖然我們的方法缺少介系詞的翻譯，但是單詞翻譯以及單詞排序卻比IBM4佳，當在缺少介系詞的情況下，翻譯結果幾乎跟IBM4的相同，若將缺少的介系詞翻譯補回，我們FPMT的表現應當可超越IBM4。



圖十二、名詞詞組翻譯比較圖(外部資料測試)



圖十三、動詞詞組翻譯比較圖(外部資料測試)



圖十四、介系詞詞組翻譯比較圖(外部資料測試)

4.3 實驗討論

從4.2節數據分析，無論在內部測試還是外部測試翻譯效果都有些許的改進，但是在外部測試時，由於BLEU無法評估加入詞彙後的詞組是否通順，導致流暢化的結果使得分數下降，而我們流暢化方法的門檻值並沒有嚴格設定，也造成加入過多的辭彙，這也是分數下降的原因之一。[12]提到，BLEU的參考答案至少要有三組以上才合理，因為英文翻譯成中文時，常常會有語意相同詞不同的情況，例如外部測試的詞組中有一句是<after 16 weeks>，在參考答案中翻譯成<16星期後>，我們的系統翻譯成<16週後>，其實這兩者翻譯結果是屬於同義的，但使用BLEU評分，這種情形卻是錯誤的。

至於詞組內部流暢化的結果，表八是我們使用詞組內部流暢化FPMT方法與IBM4的比較結果，在1-Gram Precision以及2-Gram Precision超越IBM4，其意味就單詞翻譯而言，我們提出的方法比較有效，但是在3-Gram Precision和4-Gram Precision的結果卻不理想，造成的原因有可能為加入的辭彙，破壞了原本3-Gram及4-Gram的結構，所以分數才會下降。觀察翻譯結果後發現，加入的詞彙確實可以使句子更加流暢，以表九說明，詞組英文"diabetes risk"，在參考答案中翻譯成"糖尿病危險"，我們的系統翻譯而且加詞彙後翻譯成為"糖尿病的風險"，若先忽略同義詞的問題，發現加入"的"並不會使詞組更混亂。另一個例子也相同，"no significant differences"參考答案為"沒有顯著差別"，我們的系統答案為"沒有任何顯著差異"，加入詞彙"任何"也使得翻譯通順。

由於加入的詞彙可使得詞組順暢，但這些附加詞幾乎都是參考答案中沒有的，如果將此類詞彙加入參考答案中，而不考慮同義詞的問題，評分結果如表十，將來若把同義詞問題解決，相信應當會有更佳的表現。表十一乃將附加詞彙列入參考答案之後，進行BLEU評估，數據顯示我們提出的方法較IBM4佳。

表八、詞組內部流暢化結果比較(1)

	1-Gram-Precision	2-Gram-Precision	3-Gram-Precision	4-Gram-Precision
FPMT	0.6824	0.4831	0.3530	0.2605
IBM4	0.6803	0.4821	0.3595	0.2734

表九、詞組內部流暢化例子

	diabetes risk	no significant differences
FPMT	糖尿病的風險	沒有任何顯著差異
Reference Translation	糖尿病危險	沒有顯著差別

表十、將附加詞彙列入參考答案之詞組內部流暢化結果比較

	1-Gram-Precision	2-Gram-Precision	3-Gram-Precision	4-Gram-Precision
FPMT	0.6870	0.4916	0.3656	0.2757
IBM4	0.6803	0.4821	0.3595	0.2734

表十一、BLEU 比較(詞組內部流暢化)

	BLEU
FPMT	0.3638
IBM4	0.3407

由上面的實驗可發現，在外部資料測試中，當我們使用的參考答案並沒有解決同義詞的問題時，我們的流暢化方法可以達到一些效果，若同義詞問題可以解決，準確率應當會再提升。由於使用BLEU並不能斷定翻譯結果的好壞，因此我們使用人工評估來評定翻譯的準確率。我們從實驗資料中在隨機抽出100句的名詞詞組、動詞詞組、介系詞詞組並找五位使用者來進行人工評估，將FPMT及IBM4所產生的詞組翻譯，以不固定順序方式隨機排列兩個方法得到的翻譯結果，如此

可避免使用者猜出所列出的翻譯為何種方法產生，以提高公信力。表十二為各詞組的翻譯準確率，表中的正確(Correct)是代表翻譯正確，可接受(Acceptable)則為翻譯結果並非很合適或是有缺詞的情況，但結果仍可以表達其翻譯意義。表十二可以清楚得知，各類型詞組的接受率都可以比IBM4來的好，所以對使用者而言，我們的翻譯方法確實可以達到較好的翻譯品質。表十二可以清楚得知，各類型詞組的接受率都可以比IBM4來的好，所以對使用者而言，我們的翻譯方法確實可以達到較好的翻譯品質。

表十二、人工評估翻譯結果

		Correct	Acceptable
Noun Phrase	FPMT	59%	80%
	IBM4	50%	70%
Verb Phrase	FPMT	41%	70%
	IBM4	24%	52%
Prepositional Phrase	FPMT	38%	65%
	IBM4	37%	60%

從以上的各種實驗數據分析，不論是使用BLEU自動化評估或是人工評估，在各類型的詞組，我們提出的方法皆優於IBM4。最後詞組翻譯流暢度的比較，我們從實驗題目中，取出28題有流暢化的題目進行人工評估，表十三顯示詞組有流暢化的翻譯結果較容易被使用者所接受，其說明我們所加入的詞彙確實可以讓使用者更容易閱讀。

表十三、翻譯流暢度人工評估比較

		Degree of Fluency
28 Test Phrases	FPMT	62%
	IBM4	38%

論文的附錄為測試題目中，流暢化時所加入的詞彙。附錄A為名詞詞組所加入的詞彙，我們發現比較特別的是許多中文名詞詞組應該加入量詞才會通順，一般英文名詞詞組沒有量詞，但是中文翻譯時，在數字後加上量詞才能顯示名詞詞組的完整性以及流暢度，如果只是按照單詞翻譯，所得的結果並不是最正確的。附錄中的資料顯示詞彙大部份加入次數都為1，但這並不表示加入的詞彙是錯誤的，由於測試題目都是隨機選取，所以很多附加詞彙的加入次數只有1次。從附錄中也可以看見加入的詞彙有一些多詞所組成的詞組，這些幾乎都是因為詞典的翻譯錯誤導致的。由於我們使用的詞典是GIZA++產生的，所以詞典裡的翻譯不完全都是正確的，品質不好的詞典會造成學習附加詞彙的錯誤。由於我們的語料庫太小訓練並不足夠，所以很多附加的詞彙可能沒辦法學習到，因此我們未來會收集更多雙語語料加強語料庫的訓練來提升我們翻譯系統的流暢化效能。

5. 結論

本論文針對詞組翻譯部分、單詞位置的不同以及詞義消歧的問題，我們提出流暢化詞組機器翻譯模型，除了改善翻譯的效能，並透過平行語料庫及網路搜尋結果，以提升中文翻譯的流暢度。雖然外部資料實驗數據並沒有提升流暢化後的結果，經由分析其原因，主要為BLEU並不能有效地評估句子的流暢度，所以我們認為句子翻譯品質的好壞，並不適合用N-gram模型加以評估，所以我們透過人工的評估，證明我們的方法確實可以提升翻譯的流暢化，更重要的是翻譯結果能讓使用者可以瞭解詞句表達的涵意。

目前自動的評分方式皆無法有效的評估句子是否流暢，當詞組或句子加入附加詞彙時，往往只會使翻譯效能評估下降，就算加入合適的詞彙還是無法獲得高分數，因此我們希望能找出更好的評分方法，不是單以詞彙來決定分數，仍需考慮結構以及流暢度等問題。

6. 參考文獻

[1] 國際厚生健康園區，http://www.24drs.com/professional/

[2] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguisics*, 19(2), pp. 263-311.

[3] J. S. Chang, D. Yu and C. J. Lee. 2001. Statistical Translation Model for Phrases. *Computational Linguistics and Chinese Language Processing*, 6(2) , pp.43-64.

[4] D. Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 263-270.

[5] L. F. Chien, T. I. Huang and M. C. Chen. 1997. PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval. *Proceedings of the 20th Annual International Conference on Research and Development in Information Retrieval* (*SIGIR*), pp. 50-59.

[6] Y. Ding and M. Palmer. 2005. Machine Translation Using Probabilistic Synchronous Dependency Insertion Grammars. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.541-548.

[7] G. Foster. 2000. A Maximum Entropy Minimum Divergence Translation Model. *Proceedings of the 38rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 45-52.

[8] H. J. Fox. 2002. Phrasal Cohesion and Statistical Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 304-311.

[9] GIZA++, http://www.fjoch.com/GIZA++.html

[10] ISI(Information Sciences Institute), http://www.isi.edu/publications/licensed-sw/rewrite-decoder/index.html

[11] D. Klein and C. D. Manning. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 423-430.

[12] P. Koehn, F. J. Och and D. Marcu. 2003. Statistical phrase-based translation. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp. 127-133.

[13] D. Marcu and W. Wong. 2002. A phrase-based, joint probability model for statistical machine translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 133-139.

[14] F. J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4): 417-449.

[15] F. J. Och and H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 295-302.

[16] F. J. Och, C. Tillmann and H. Ney. 1999. Improved Alignment Models for Statistical Machine Translation. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (*EMNLP/VLC*), pp. 20-28.

[17] K. Papineni, S. Roukos, T. Ward and W. J. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 311-318.

[18] Stanford Parser Parse Visualization Tool, http://ai.stanford.edu/~rion/parsing/stanford_viz.html

[19] A. Venugopal, S. Vogel and A. Waibel. 2003. Effective Phrase Translation Extraction from Alignment Models. *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 319-326.

[20] T. Watanabe, E. Sumita and H. G.. Okuno. 2003. Chunk-based Statistical Translation. *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 303-310.

[21] K. Yamada and K. Knight. 2002. A Decoder for Syntax-based Statistical MT. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 303-310.

[22] K. Yamada and K. Knight. 2001. A syntax-based statistical translation model. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.523-530.

[23] R. Zens and H. Ney. 2004. Improvements in Phrase-Based Statistical Machine Translation. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp. 257-264.

附錄

附錄 A 名詞詞組附加詞彙

加入詞彙	次數	加入詞彙	次數	加入詞彙	次數
的	124	進行	1	律當這樣的	1
名	8	則有	1	充足的	1
種	5	第 2	1	vcu 的	1
有	5	餵養	1	he	1
個	4	了	1	的有老年	1
位	4	住院	1	進行評量由	1
的長期追蹤	3	二	1	外科手術的	1
起	3	篇	1	的生活	1
次	3	的國家	1	與中等	1
是	3	化	1	的專題	1
在	3	器官	1	潛毛症的	1
是否受到	2	劇烈的	1	治療	1
第	2	位再經過	1	未	1
越	2	檢測	1	吸煙的	1
例	2	rs	1	老年	1
科	2	和	1	般	1
性	2	允許	1	發生染色體	1
合作的	1	沒	1	的每天	1
癌外科	1	攝取的	1	方面的	1
原因	1	接	1	的要	1
三分之	1	這項研究	1	年夏	1
位原	1	的藥品	1	的李	1
及	1	友例	1	或他形式	1
近視	1	人因出血而	1	顯示	1
的失	1	極具	1	中的	1
月的	1	的出血	1	是為	1
兒童的	1	院	1	治療的	1
能力的	1	的轉移	1	這種	1
再增加	1	治療過程中	1	來做	1
的正	1	ibs	1	繼續	1
凝血因子	1	用藥	1	氧氣	1
有三分之	1	障礙	1	這種病毒是	1
滴	1	或斷症狀而	1	將	1

加入詞彙	次數
ncet	1
到	1
一種	1
射出	1
限 2	1
這些	1
放棄的	1
星期的	1
發現	1
行為	1
在這方面的	1

附錄 B 動詞詞組附加詞彙

加入詞彙	次數	加入詞彙	次數	加入詞彙	次數
的	12	敏感更	2	事故性處理	1
有	8	受到	1	少要	1
在臨床	6	何	1	規模	1
尚	6	那麼	1	醫療師都應	1
輕度且	6	在一年之內	1	選出的樣本	1
是	5	研究機構的	1	知道	1
可	5	經由皮膚	1	位未	1
進行	4	至警報中	1	目前為止尚	1
御從	4	產生	1	間歇地睡	1
在	4	srs	1	很短暫也很	1
個	3	g 的分配	1	的限制	1
降低並較	3	心臟病	1	經過	1
研究結果	3	年由名	1	左該評估能	1
會	3	成為	1	到 hsv	1
治療的患者	2	接下	1	提出	1
先	2	結果是	1	後的救治	1
受	2	至	1	是那麼常	1
腎臟目標的	2	到	1	病毒	1
得	2	成年	1	為治療	1
非常	2	咳	1	2	1
的團體	2	時	1	應該因	1
二	2	太大的	1	有人	1
介於	2	合理	1	中的	1
可以	2	從九例中	1	化學	1
預防	2	藥物的	1	三	1
公平	2	證明之	1	服用抗癲癇	1
使	2	指令也	1	修正後	1
加	2	不僅	1	過	1
種的	2	的研究	1	小組	1
認為如果	2	由 ct	1	如此	1
應	2	出現不的	1	小腸放射學	1
嚴密的	2	率	1	回	1
使	2	為止這個	1	給	1

加入詞彙	次數
氫化可松	1
地	1
接受疾病的	1
一種生物	1
認為得	1
第 i	1
選	1
其	1
分鐘的	1

附錄 C 介系詞詞組附加詞彙

加入詞彙	次數	加入詞彙	次數	加入詞彙	次數
的	58	即	2	至	1
名	9	治療的	2	的患者其	1
ibs	4	meth	2	沒	1
上	4	他	1	有瀰漫性	1
患者其	4	是否受到	1	兒童心臟	1
位	4	與這些	1	位做	1
在	4	腦膜炎	1	與患者	1
第	4	分 治療	1	起訴及	1
20	3	造	1	中的	1
個	3	所受	1	的要	1
的長期追蹤	3	時提示各	1	使用	1
年	3	的象	1	只有移植前	1
是	3	矯形外科學	1	單獨	1
起	3	發生的	1	將	1
癌	3	劑	1	由的	1
有	3	胃	1	發	1
的類固醇	2	的血中	1	全部	1
仍	2	用	1	次理想的 t	1
進行	2	州	1	改變	1
治療因	2	數	1	為期	1
患者	2	世界上	1	檢查抗	1
性	2	引發	1	抽樣得到	1
藥物	2	方法	1	頭痛的	1
種	2	和	1	更	1
出現了	2	患	1	感染	1
所	2	會的	1	於產生換種	1
發生	2	照	1	患者進行	1
個體	2	乳房接受	1	任與	1
預防	2	項	1	對有	1
srs	2	得到的	1	erd	1
隨訪	2	所報告	1	需要	1
只有	2	月的	1	類型的	1
會	2	位無法手術	1	ncet	1

加入詞彙	次數	加入詞彙	次數	加入詞彙	次數
那些希望	1	介入	1	其	1
40	1	族群的治療	1	術	1
提高對	1	的研究	1	就的知識	1
個歲因	1	的健康	1	時間	1
美國癲癇學	1	公分	1	些	1
科學	1	化的	1	其中	1
位糖尿病	1	的既往	1	困惑	1
例中診斷出	1	直拳加有	1	而罹患殘障	1
與約試驗	1	及呼吸道	1	檢查	1
病例	1	受	1	艾滋病的	1
神下因	1	脊	1	阻	1
臨床腫瘤	1	起為	1	抑制的	1
服用	1	使	1	吐氣	1
年第 36 週	1	的隨機	1	具有	1
全面	1	傳輸頻率	1	共有	1
展開	1	以	1	肩部鈣化	1
是了	1	是否有益	1	一篇研究	1
肥胖者	1	1800 名	1	清楚但無的	1
宿主防禦	1	性過敏症	1	使得	1
濃度	1	的腦血流	1	二	1
包括	1	第 24 次	1		
結果使用	1	內視鏡檢查	1		
助聽器	1	月	1		
類似	1	旦	1		
致	1	日	1		
閉塞情形	1	高	1		
從業人員	1	次	1		
社交恐懼症	1	膽管	1		
用來	1	話則其	1		
例的	1	1800	1		
結果	1	而	1		
發現有	1	必須評估	1		
的藥物	1	一篇有關於	1		
聖路易	1	之間不良的	1		
行動	1	然而	1		

An Evaluation of Adopting Language Model as the Checker of Preposition Usage

Shih-Hung Wu, Chen-Yu Su
Dept. of CSIE, Chaoyang University of Technology, Taiwan, R.O.C.
shwu@cyut.edu.tw, s9427617@cyut.edu.tw

Tian-Jian Jiang, Wen-Lian Hsu
Institute of Information Science, Academia Sinica, Taiwan, R.O.C
Department of Computer Science, National Tsing-Hua University, Taiwan, R.O.C
tmjiang@iis.sinica.edu.tw, hsu@iis.sinica.edu.tw

Abstract

Many grammar checkers in rule-based approach do not handle errors that come from various usages, for example, the usages of prepositions. To study the behavior of prepositions, we introduce the language model into a grammar-checking task. A language model is trained from a large training corpus, which contains many short phrases. It can be used for detecting and correcting certain types of grammar errors, where local information is sufficient to make decision. We conduct several experiments on finding the correct English prepositions. The experiment results show that the accuracy of open test is 71% and the accuracy of closed test is 89%. The accuracy is 70% on TOEFL-level tests.

Keywords: *Language Model, Grammar Checker, English preposition usage*

1. Introduction

Computer-Aided Language Learning is a fascinating area; however, the computer still lacks many abilities of a human teacher, for example, the ability of grammar checking. Technically, it is hard to build a grammar checker that can deal with all types of errors. There are errors caused beyond the knowledge of syntax. For example, to overcome the misusing of prepositions, a system requires more semantic knowledge.

There are three major approaches to implement a grammar checker. The first strategy is the syntax-based checking [Jensen et al., 1993]. In this approach, a sentence is parsed into a tree structure. A sentence is correct if it can be parsed completely. Another choice is the statistics-based checking [Attwell, 1987]. In this approach, the system built a list of POS tag sequences based on a POS-annotated corpus. A sentence with known POS tag sequence is considered as a correct one. The last one is the rule-based checking [Naber 2003], where a set of rules is built manually and used to match against a text. Park et al. proposed an online English grammar checker for students who take English as the second language. This system focuses on a limited category of frequently occurring grammatical mistakes in essays written by students. The grammar knowledge is represented in Prolog language. [Park 1997]

We find that most grammar checkers do not deal with the errors of preposition usage. We suppose that it should be hard to write rules for all of the prepositions. To evaluate this difficulty, we introduce the language model into the grammar-checking task. Since a language model is usually trained from a large training corpus, it may contain many short phrases with prepositions.

The Language Model (LM) is one of the popular natural language processing technology for various applications, like information retrieval, handwriting recognition, speech recognition, and

machine translation. [Jurafsky and Martin, 2000] [Manning and Schutze, 1999] An LM uses short history to predict the next word. Word prediction is an essential subtask of speech recognition, handwriting character recognition, augmentative communication for the disabled, and spelling error detection. An LM can estimate the probability of a sentence. Therefore, it can be a way to distinguish good usages from bad ones of English prepositions.

Figure 1 shows a general architecture of an English grammar checker. An ideal system should consist of both rule-based and language model approaches. Linguistic knowledge of the rule-based system is acquired from domain experts. Statistical knowledge of the language model is gathered from training corpus by programs. In this paper, we design several experiments to assess the ability of the LM on the preposition usage problem.

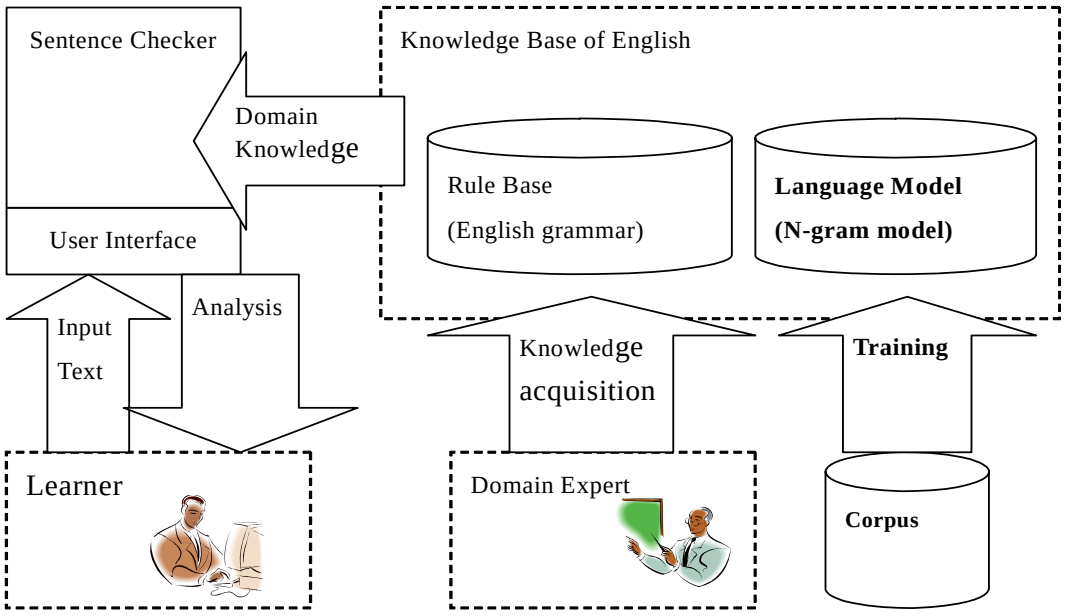


Figure1. The Architecture of a general English Grammar Checker

2. Statistical language model

We briefly restate the notation of N-gram language model. In this model, a sentence is viewed as a sequence of *n* words. The probability of a sentence in a language, say English, is defined as the probability of the sequence.

$$P(w_1^n) \equiv P(w_1, w_2, ..., w_n)$$

That can be further decomposed by the chain rule of conditional probability under the Markov assumption.

$$\begin{aligned} P(w_1^n) &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1^2)...P(w_n \mid w_1^{n-1}) \\ &= P(w_1)\prod_{k=2}^{n} P(w_k \mid w_1^{k-1}) \end{aligned}$$

Since it is not possible to collect all the history, a prefix of size N, as an approximation, is used to replace each component in the product.

$$P(w_n \mid w_1^{n-1}) \approx P(w_n \mid w_{n-N+1}^{n-1})$$

Usually, the N is 1, 2, or 3, are named as unigrams: $P(w_n)$, bi-grams: $P(w_n|w_{n-1})$, and tri-grams: $P(w_n| w_{n-1}w_{n-2})$ model, respectively.

Next step is to estimate the n-gram approximation from corpus. The basic way is called Maximum Likelihood Estimation (MLE), which calculates the relative frequency and is used as the estimation of probability. For bi-gram:

$$P(w_n \mid w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

And, for n-gram

$$P(w_n \mid w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1} w_n)}{C(w_{n-N+1}^{n-1})}$$

where *C* represents the count of each specified n-grams *w* in the corpus. MLE works well for high-frequency n-gram; however, no matter how large the corpus is, there are always some low-frequency n-grams. The frequency might be very low even zero. Some zeroes are really zeroes, which means that they represent meaningless word combinations. However, some zeroes are not really zeroes. They represent low frequency events that simply did not occur in the corpus and might exist in real world. When using n-gram model, we cannot assign a probability to a sequence where one of the component n-gram has a value of zero. An alternative solution is to smooth the probability estimations so that no component in the sequences is given a probability of zero.

2.1 Smoothing methods

To cope with the problem of unseen data, several smoothing methods are developed [Goodman, 2002]; they can be classified as discounting methods and model combination methods. Discounting methods adjust the probability estimators, so that zero relative frequency in the training data does not imply zero relative counts. Model combination methods combine available models (unigram, bi-gram, tri-gram, etc.) by interpolation and back-off. To our knowledge, Good-Turing discounting, absolute discounting and Chen-Goodman modified Kneser-Ney discounting are three of best smoothing methods; therefore, we use them in our experiments. [Chen and Goodman, 1998]

2.1.1 Good-Turing Discounting (GT)

Good-Turing discounting adjusts the count of n-gram from *r to* r^*, which is base on the assumption that their distribution is binomial [Good, 1953].

$$r^* = (r+1)\frac{N_{r+1}}{N_r} \qquad r < M$$

where N_r is types of n-gram occurring r times, and M is a threshold usually smaller than 5. Note that for r=0,

$$r^* = \frac{N_1}{N_0}$$

where N_0 is the number of n-grams that never occurred. The discounted probabilities are thus:

$$P_{GT}(w_1 \ldots w_n) = \frac{r^*}{N}$$

The Good-Turing formula only applies to the situation when r < 5, and need to renormalize to ensure that everything sums to one.

2.1.2 Absolute Discounting (AD)

In the absolute discounting model, all non-zero frequencies are discounted by a small constant discount rate *b*. And all the unseen events gain the frequency uniformly. [Ney et al., 1994]

$$N_0 \cdot P_0 = \frac{1}{N}\sum_{r=1}^{R} N_r \cdot discount\_rate = b \cdot \frac{K - N_0}{N},$$

Where *R* is the highest frequency and *K* is the number of bins that training instances are divided into:

$$K = \sum_{r=0}^{R} N_r, \qquad 0 < b \le 1$$

So the probability is

$$p_{abs}(w_1 \ldots w_n) = \begin{cases} \frac{r-b}{N}, 0 < r \le R \\ b \cdot \frac{K - N_0}{N \cdot N_0}, r = 0 \end{cases}$$

2.1.3 Modified Kneser-Ney discounting (mKN)

The Kneser-Ney discounting model is a back-off model based on an extension of absolute discounting which provides a more accurate estimation of the distribution. Chen and Goodman proposed a modified Kneser-Ney(mKN) discounting model. Instead of using a single discount for all nonzero counts as in KN smoothing, the mKN has three different parameters, D_1, D_2, and D_3 that are applied to n-grams with one, two, and three or more counts, respectively. The formula of mKN discounting is:

$$P_{mKN}(w_i \mid w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i}) - D(c(w_{i-n+1}^{i}))}{\sum_{w_i} c(w_{i-n+1}^{i})} + \gamma(w_{i-n+1}^{i-1}) P_{mKN}(w_i \mid w_{i-n+2}^{i-1})$$

where

$$D(c) = \begin{cases} 0 & if\ c = 0 \\ D_1 & if\ c = 1 \\ D_2 & if\ c = 2 \\ D_3 & if\ c \geq 3 \end{cases}$$

$$D_1 = 1 - 2\frac{N_1}{N_1 + 2N_2} \cdot \frac{N_2}{N_1}$$
$$D_2 = 1 - 3\frac{N_1}{N_1 + 2N_2} \cdot \frac{N_3}{N_2}$$
$$D_3 = 1 - 4\frac{N_1}{N_1 + 2N_2} \cdot \frac{N_4}{N_3}$$

and the gamma is a normalization constant such that the probabilities sum to one.

2.2 Entropy and Perplexity

Entropy is widely used to measure information. The entropy of a random variable X ranges over what are predictable set T (words, letters, or parts-of-speech) can be defined as:

$$H(X) = -\sum_{x \in T} p(x) \log_2 p(x)$$

Perplexity is a variant of entropy. Generally, the perplexity can be defined as:

$$2^{H}$$

Entropy of sequence of words can be defined as:

$$H(w_1, w_2, ..., w_n) = -\sum_{W_1^n \in L} p(W_1^n) \log_2 p(W_1^n)$$

Where $p(W_1^n)$ can be replaced by n-gram models.

3. Experiments

To assess the ability of how LM finds the right preposition, we use various sizes of training sets, and three test sets from three different sources.

3.1 Experiment design

For each original test sentence, we make up some wrong ones, and then calculate the perplexity of the test sentences. The perplexity is the measurement of how well the LM can predict the sentence. The sentence with the lowest perplexity is the most possible sentence with respect to the given LM; we assume that sentence is the correct one.

We conduct the experiments with the SRI Language Modeling Toolkit. [Stolcke, 2002] [http://www.speech.sri.com/projects/srilm/] The first test set comprises 100 sentences that we select from the training set. This test is regarded as a closed test. The second test set is another 100 sentences that we collect from various English literatures outside the training set. This is an open test. In the first two experiments, we focus on only three prepositions: in, on, and at. We fabricate the wrong sentences by replacing the correct preposition with other ones. The third test

set consists of 100 sentences of TOFEL-level questions. We collect these sentences from TOFEL reference books; they contain most of the English prepositions.

The training corpus is selected from LDC Gigaword corpora [LDC 2003]. The Gigaword corpora are very large English newswire text collections. There are four distinct international sources: Agence France Press English Service (AFE), Associated Press Worldstream English Service(APW), The New York Times Newswire Service (NYT) and The Xinhua News Agency English Service (XIE). The total size of the corpora is more than one gigabyte in word counts.

We use the NYT corpus as the training set. The training set sizes in different experiments are different. For bi-gram model, we select the news of the NYT from January 1999 to June 2002 as our training corpus. It consists of 351,427,489 words and is about 1.89 GB. We do not perform any preprocessing and do not remove stop words. For tri-gram model, we select the news of NYT from January 2001 to June 2002. This corpus consists of 156,896,511 words and the size is about 856 MB.

Table 1 The sizes of the training sets for Bi-gram model

Training Set	# of words	MB
nyt200111-200206(8)	69865209	384
nyt200101-200206(18)	156896511	865
nyt199901-200206(42)	351427489	1890

Table 2 The sizes of the training sets for Tri-gram model

Training Set	# of words	MB
nyt200203(1)	9310195	52
nyt200203-200204(2)	18734690	102
nyt200201-200206(6)	52574963	289
nyt200108-200206(11)	97578257	537
nyt200101-200206(18)	156896511	865

3.2 Experiment results

3.2.1 Closed tests

In the first experiment, we select 100 sentences from our training corpus as the test set. We fabricate the wrong sentences by replacing the correct preposition with other prepositions. We calculate the perplexity of the sentences with LMs and check if the sentence with the lowest perplexity is the original one. We do not list the values of perplexity, since it is meaningless for the closed test. In computing perplexities, the model must be constructed without any knowledge of the test set. The knowledge of the test set will make the perplexity artificially low.

Table 3 and 4 shows the accuracy of the first test set on various LMs. In this task, the test accuracy of bi-gram is lower than that of tri-gram; even the training size is doubled. The accuracy for tri-gram converges as the size of training set increasing. In Table 4, we enlarge the training corpus from 1-month news articles to 18-months news articles for the training set, the test accuracy does not increase much. The mKN smoothing method gives the best accuracy 89%.

Table 3. The closed test accuracy of bi-gram models on various training sets.

	Smoothing method		
Training Set	GT	mKN	AD
nyt200111-200206(8)	65%	65%	**73%**
nyt200101-200206(18)	65%	65%	
nyt199901-200206(42)	66%		

Table 4. The closed test accuracy of tri-gram models on various training sets.

	Smoothing method		
Training Set	GT	mKN	AD
nyt200203(1)	80%	86%	88%
nyt200203-200204(2)	80%	85%	
nyt200201-200206(6)	87%	88%	
nyt200108-200206(11)	85%	88%	
nyt200101-200206(18)	85%	**89%**	

3.2.2 Open tests

In the second experiment, the test set is another 100 sentences that we collect from the following five English literary works: (download from Project Gutenberg Online Book Catalog http://www.gutenberg.org/)

1. Amusements in Mathematics by Henry Ernest Dudeney.
2. Grimm's Fairy Tales by Jacob Grimm and Wilhelm Grimm.
3. The Art of War by Sun-Zi.
4. The Best American Humorous Short Stories.
5. The War of the Worlds by H. G. Wells.

Again, we fabricate the wrong sentences by replacing the correct preposition with other prepositions. We calculate perplexities of the sentences with LMs of different sizes and check if the sentence with the lowest perplexity is the original one.

Table 5 and 6 show the accuracy of the second test set. The mKN smoothing method gives the best accuracy 71%.

Table 5. The open test accuracy of bi-gram models on various training sets.

	Smoothing method		
Training Set	GT	mKN	AD
nyt200111-200206(8)	47%	49%	50%
nyt200101-200206(18)	48%	**51%**	
nyt199901-200206(42)	47%		

Table 6. The open test accuracy of tri-gram models on various training sets.

	Smoothing method		
Training Set	GT	mKN	AD
nyt200203(1)	61%	57%	61%
nyt200203-200204(2)	61%	62%	
nyt200201-200206(6)	67%	69%	
nyt200108-200206(11)	68%	69%	
nyt200101-200206(18)	68%	**71%**	

3.2.3 TOEFL-level tests

There is a problem in the setting of the previous two experiments. We do not check if the fabricated wrong sentences are also legal in the real world. Therefore, we collect 100 TOEFL-level single-choice questions from pseudo TOEFL tests. Each sentence has a blank for a preposition. Four candidates are available, but only one is correct. For example:

My sister whispered __ my ear.

(a) in (b) to (c) with (d) on

Then our task is to distinguish which of the following four sentences is correct.

My sister whispered in my ear. (correct)

My sister whispered to my ear. (wrong)

My sister whispered with my ear. (wrong)

My sister whispered on my ear. (wrong)

We also train our LMs with different sizes of training set. We then use the LMs to calculate the perplexities of the four sentences. The system regards the sentence with the lowest perplexity as the correct one. The results in Table 7 show that tri-gram model with mKN smoothing gives the best result even though the training size is much smaller than the one for the bi-gram model.

Table 7. The TOEFL-level tests accuracy of bi-gram and tri-gram model

	Smoothing method	
Training Set	GT	mKN
Bigram model nyt199901-200206(42)	53%	54%
Trigram model nyt200101-200206(18)	69%	**70%**

3.3 Error Analysis

Table 8 shows a part of the test results that the LM gives wrong answers. The system chooses the candidate with the lowest perplexity as the answer; however, in these cases, the candidates with the lowest perplexities are wrong. We manually check these sentences and identify the necessary keyword. We find that, to give the right answer, the system must refer to some words that are not close to the blank. Such long-distance features cannot be learned in a short windows size of two or three; therefore, the tri-gram model cannot give the right answer.

Table 8. Error examples of using the tri-gram model on TOEFL-level tests, where the logprob is the logarithm of n-gram probability and the perplexity is defined as 10^(-logporb/ # of words in the sentence).

No.	Question	choices	correct answer	logprob	perplexity	LM answer
1	It is sometimes difficult to make pleasant conversation ___ people you have just met.	among		-35.5006	343.367	
		to		-33.5936	250.923	v
		for		-36.7712	423.168	
		with	v	-33.6707	254.127	

2	I have no knowledge whatever ___ the sciences.	of	v	-23.0051	751.007	
		to		-23.5853	887.482	
		in		-20.978	419.037	v
		on		-23.8697	963.202	
3	I'm bored ___ staying here.	of	v	-16.5306	2023.56	
		in		-16.7524	2241.21	
		with		-15.1587	1075.83	v
		for		-16.5074	2002.09	
4	He lives ___ 144 Wall Street.	at	v	-17.7392	904.767	
		in		-15.9947	463.223	v
		on		-18.7051	1310.76	
		by		-18.3593	1147.85	
5	We danced ___ the music of Jimmy Dorsey's band.	to	v	-25.8146	738.4	
		with		-26.2586	827.221	
		in		-25.5648	692.674	v
		on		-26.0885	791.996	
6	Write your composition ___ ink.	in	v	-19.8675	9408.04	
		on		-21.0123	15938.9	
		with		-19.4143	7635.85	v
		by		-20.1884	10906.4	
7	In a short while, I'll be free ___ all my worries.	with		-30.8353	635.642	
		of	v	-28.7124	407.583	
		about		-32.6347	926.383	
		to		-27.3856	308.745	v
8	He stopped the car ___ the park.	by	v	-16.5065	228.069	
		in		-13.9978	99.9273	v
		on		-15.4607	161.688	
		to		-14.4585	116.279	
9	That would be ___ my dignity.	beneath	v	-15.0318	320.109	
		under		-14.2876	240.581	
		beyond		-13.86	204.171	v
		above		-15.9486	455.096	
10	The fire began ___ the fifth floor of the hotel, but it soon spread to adjacent floors.	on	v	-40.8443	252.701	
		in		-39.2952	204.872	v
		at		-41.4813	275.47	
		of		-44.1101	393.292	
11	The main office of the factory can	in		-32.6532	109.856	

	be found ___ Maple Street in New York City.	at		-32.1039	101.507	v
		on	v	-33.8979	131.408	
		from		-34.4493	142.26	
12	Conifers first appeared on the Earth ___ the early Permian period, some 270 million years ago.	when		-42.6762	699.972	
		or		-41.3285	569.158	
		and		-39.3227	418.322	v
		during	v	-40.0914	470.714	
13	She'll be here ___ about twenty minutes.	by		-22.3606	1564.51	
		on		-21.2314	1079.08	v
		at		-21.4576	1162.45	
		in	v	-22.4876	1631.24	

4. Conclusions and discussions

In this paper, we report the evaluation of adopting the language model on checking the English prepositions. In our experiments, we assume that a correct sentence has less perplexity than the wrong ones. The experiment results show that tri-gram language model can find most of the correct prepositions. The modified Kneser-Ney smoothing method gives the best accuracy in three test sets. Experiment results show that the accuracy of open test is 71%, the accuracy of closed test is 89%, and the accuracy on TOEFL-level test is 70%. This approach has two advantages, the first one is that it requires only untagged corpus. The second one is that it requires no domain knowledge. Thus, the approach can cooperate with other approaches in the future easily.

To improve the accuracy, the system requires more linguistic knowledge. Other feature-based machine learning approaches, for instance, Maximum Entropy (ME) [Berger et al., 1996], Conditional Random Fields (CRF) [Lafferty et al., 2001] are also promising. They can incorporate more long-distance linguistic features that LM cannot. [Rosenfeld, 1997].

The collection of linguistic features requires more knowledge engineering. In an English grammar textbook of college-level [Eastwood, 1999], the usages of the prepositions are addressed by rules and examples, as listed in Table 9 and 10. To cooperate with the rules, a system requires linguistic resources to recognize the names of different entities such as countries, regions, towns, and time expressions. Moreover, the system still requires templates of specific usages. Table 10 gives many common phrases examples of the three prepositions: in-on-at (used for place only). These "common" phrases might appear in the corpus many times. Since they are short, they will be in the tri-gram model.

Table 9. Rules of preposition usage [Eastwood, 1999]

	Positive and Negative Rules
At	1. Use in (not at) before the names of countries, regions, cities, and large towns. 2. Use in (not at) with seasons, months, and years. 3. Use on (not at) before dates. 4. Without at before 'an hour before', 'a week later', 'two years afterwards' 5. Do not use at to introduce a time expression with ago.
In	1. on a day or date, not in 2. in the morning/afternoon/evening' but 'the following morning', 'the next afternoon', 'the previous evening', etc. 3. When talking about how long something lasts or continues, use for, not in. 4. on/upon doing something, not in 5. made of wool/wood etc., not in 6. in is not used in expressions such as 'the shop is open six days a week.' 'He visits his father three times a year.' 'Bananas cost fifty pence a pound.' 'I drove to the hospital at ninety miles an hour.'
On	1. Do not use a preposition to begin a time expression with next when the point of time is being considered in relation to the present: 'the next morning', 'the next afternoon'. 2. a good/bad thing about someone/something, not on 3. When talking about a particular afternoon, use on. When speaking generally, use in.

Table 10. Common phrase for in, on, and at [Eastwood, 1999]

	Common phrases (place)
In	In prison/hospital In the lesson In a book/newspaper

	In the photo/picture In the country In the middle In the back/front of a car In a queue/line/row
On	On the platform On the farm On the page/map On the screen On the island/beach/coast Drive on the right/left On the back of an envelope
At	At the station/airport At home/work/school At the seaside At the top/bottom of a hill At the back of the room At the end of a corridor

Acknowledgement

This research was partly supported by the National Science Council under GRANT NSC 94-2218-E-324 -003.

References

[Atwell 1987] Eric Atwell, Stephen Elliott, Dealing with ill-formed English text, in: The computational analysis of English : a corpus-based approach / edited by Roger Garside, Geoffrey Leech, Geoffrey Sampson, The computational analysis of English, London ; New York, Longman, 1987.

[Berger et al., 1996] A. Berger, S. A. Della Pietra, and V. J. Della Pietra, A maximum entropy approach to natural language processing, Computational Linguistics, vol. 22, pp. 39-71, 1996.

[Chen and Goodman, 1998] S. F. Chen and J. Goodman, An empirical study of smoothing techniques for language modeling, Technical Report TR-10-98, Computer Science Group, Harvard University, Aug. 1998.

[Eastwood, 1999] John Eastwood, Oxford Practice Grammar, Oxford University Press, 1999.

[Good, 1953] I.J. Good, The population frequencies of species and the estimation of population parameters, Biometrika 40: pp 237-264, 1953.

[Goodman, 2002] Joshua T. Goodman, A bit of Progress in Language Modeling, Technical Report, MSR-TR-2001-72, Microsoft Research, Redmond, 2002.

[Jensen et al.,1993] Karen Jensen, George E. Heidorn, Stpehen D. Richardson (Eds.): Natural language processing: the PLNLP approach, Kluwer Academic Publishers, 1993.

[Jurafsky and Martin, 2000] Jurafsky, D. and Martin, J.H., Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice-Hall, Upper Saddle River, New Jersey, 2000.

[Lafferty et al, 2001] Lafferty, J., McCallum, A., and Pereira, F., Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Paper presented at the ICML-01.

[LDC 2003] LDC, Gigaword Corpora. http://wave.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05

[Manning and Schutze, 1999] Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA: May 1999.

[Naber et al., 2003] Daniel Naber, A Rule-Based Style and Grammar Checker, diploma thesis, University Bielefeld, 2003.

[Ney et al., 1994] Hermann Ney, Ute Essen, and Reinhard Kneser, On structuring probabilistic dependencies in stochastic language modeling, Computer Speech and Language 8: pp1-28, 1994.

[Park et al., 1997] Jong C. Park, Martha Palmer, and Clay Washburn, An English Grammar Checker as a Writing Aid for Students o f English as a Second Language, in Proceedings of the Fifth Conference on Applied Natural Language Processing, 1997. http://acl.ldc.upenn.edu/A/A97/A97-2014.pdf

[Rosenfeld, 1997] Ronald Rosenfeld, A Whole Sentence Maximum Entropy Language Model, In Proc. IEEE workshop on Automatic Speech Recognition and Understanding, Santa Barbara, California, December 1997.

[Stolcke, 2002] Andreas Stolcke, SRILM - An Extensible Language Modeling Toolkit, in Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado, September 2002