以語料為本的中文專有名詞分類

梁婷* 葉政輝 吳典松
國立交通大學
資訊科學學系
*Email: tliang@cis.nctu.edu.tw

摘　　要

正確的專有名詞的語意標示將有助於文件擷取及訊息了解。在本論文中，我們提出一個結合法則和統計方法的分類機制以標示中文文件中中文人名和組織名稱。在人名標示上主要利用人名常見字元來建立中文字元機率模型。組織名稱的辨識則主要建立於專有名詞前後常見詞彙與詞類標記整合。我們以中文平衡語料庫語料做為訓練和測試資料以驗證所提系統中不同模組的標示效能，進而建立最佳的分類程序。

一、　緒論

正確的專有名詞的語意標示無疑地將有助於文件內容的了解，進而促進文件的自動分類、擷取和訊息萃取 [1-12]。在本篇論文中，我們將針對在中文文件中常現的人名和組織名稱提出一個結合法則和統計方法的自動標記系統。

基本上，中文人名是由姓氏加上名字所組成。姓氏有所謂的單姓與複姓，和少數結婚女士冠上夫姓與入贅男士冠上妻姓，故整個姓氏的長度由一到四不等。而名字的部分，常會使用通俗的單字組或雙字組，使得整個中文人名長度範圍在二字元到六字元之間。

中文人名識別的研究多為統計式的機率模型。在[3，6] 的論文中他們首先收集訓練語料庫中扮演姓氏的字元，刪除罕見的姓氏（是、那……等）以減少錯誤的識別。同時根據中文人名姓氏與名字的各種組合，來處理單姓、複姓與冠夫姓的情況。由於姓氏出現複姓的狀況永遠為姓氏，所以複姓模型並沒有判別姓氏的門檻值。冠夫姓模型除了兩個姓氏的組合外，其餘部分皆與單姓模型相同。此外，在[9]的研究中，採用與[3]相同的處理方式，並且再統計非姓氏或名字的字元機率，藉此直接取得門檻值。除了字元的機率計算外，尚利用其他的線索以提供辨識，如人名在同一篇文章中出現次數可能不只一次；前後往往伴隨著職稱（董事長、經理、部長……等）或是敘述類的動詞（表示、說、指出……等）出現。

在組織名稱辨識方面，英文的處理上使用相鄰關鍵詞與特定動詞的差異，如公司名稱後方常會出現 Inc.或是 Ltd.，與慣用特定動詞[1，12]。至於在中文組織名辨識處理上 [3，4]，除了使用相鄰關鍵詞與特定動詞等特徵外，更著重於簡稱的辨別與多個詞彙組成的組織名辨別。利用詞性標記的組合，並搭配關鍵詞的規則，不過前提需公司全名必須至少出現過一次以上。對多個詞彙組成的組織名，則以歸納法則處理。

從以上的相關研究中我們不難發現，若要提高辨識率，除了各類專有名詞的分類模型外，適用規則的提出、關鍵詞的應用以及詞類上扮演的角色皆是有效提高辨識效能的方法。在本篇論文中，我們嘗試使用單一句子內所出現的詞彙與每個詞彙的相鄰詞，輔以文章中詞類的排列順序，作為專有名詞擷取與分類的方法，期望能減少對人工建立規則的依賴。

二、　系統架構與方法

圖 1 為本篇論文所提的系統架構流程。所訓練用的語料為中研院平衡語料庫 version 3.0 並假設中研院平衡語料庫內文章的詞彙斷詞結果正確，並事先以人工將語料中屬於中文人名或是組織名的詞彙標記出來。各模組分述如後小節。

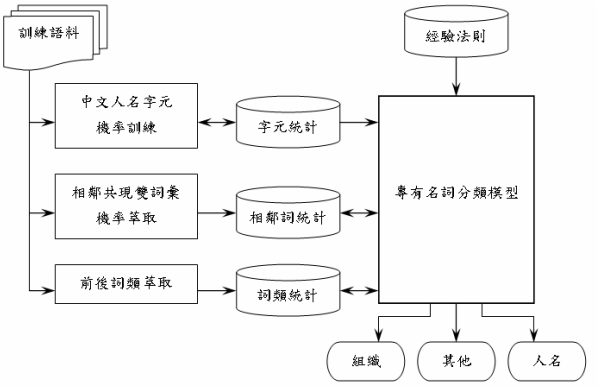

圖 1 ：系統訓練架構圖

2.1 中文人名字元機率模型

中文姓名的模型處理上，分為姓氏與名字兩部分。姓氏為中文人名辨識的主要線索，部分人名則因為冠夫姓情況有可能會出現兩次姓氏，因此在處理之前必須先將詞彙切分為姓氏與名字兩部分。我們可藉著與百家姓氏辭典做字串的比對來分離出姓氏與名字。

由於姓氏與名字在命名上並沒有相依的關聯，且本論文中並不探討人名的性別關係，因此藉由統計每個字元在語料庫中扮演姓氏與名字的機率，我們修改 [3，6] 所提出的模型，將姓氏與名字兩大部分獨立考慮，進而簡化成三條機率公式：

假設中文人名格式：$(C_i'(C_j'))+C_i(C_j)+C_x(C_y)$，其中 $C_i'(C_j')$ 為冠夫姓，$C_i(C_j)$ 為本身的姓氏，$C_x(C_y)$ 為名字。

$$P(C_i)\cong\frac{\sum f'(C_i)}{\sum f(C_i)}\geq threshold_1 \quad (2.1)$$

$$P(C_x)\cong\frac{\sum f'(C_x)}{\sum f(C_x)}\geq threshold_2 \quad (2.2)$$

$$P(C_x)\times P(C_y)\cong\frac{\sum f'(C_x)}{\sum f(C_x)}\times\frac{\sum f'(C_y)}{\sum f(C_y)}\geq threshold_3 \quad (2.3)$$

其中 $\sum f(C_i)$ 表示語料庫中字元 C_i 出現的頻率
$\sum f'(C_i)$ 表示語料庫中字元 C_i 扮演姓氏出現的頻率
$\sum f(C_x)$ or $\sum f(C_y)$ 表示語料庫中字元 C_x 或 C_y 出現的頻率
$\sum f'(C_x)$ or $\sum f'(C_y)$ 表示語料庫中字元 C_x 或 C_y 扮演名字出現頻率

公式 2.1 用來檢視字元 C_i 是否符合單姓人名門檻值 $threshold_1$；公式 2.2 用來檢視字元 C_x 是否符合單名門檻值 $threshold_2$；公式 2.3 用來檢視字元 C_xC_y 是否符合雙名門檻值 $threshold_3$。常出現在中文人名的字元具有較高的機率，也就較能通過各個門檻值的限制。整個演算法流程如圖 2 所示：

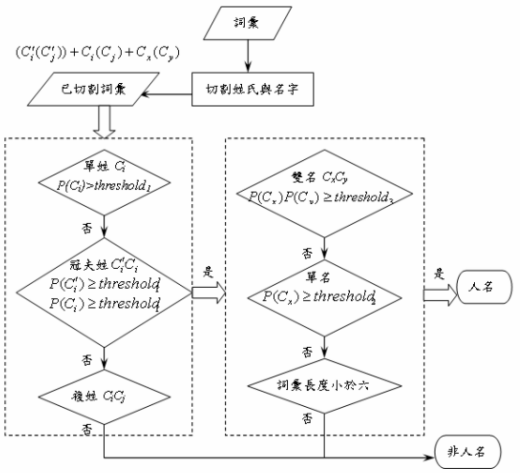

圖 2： 中文人名機率模型

切割姓氏與名字的法則如下：

(1) 詞彙長度為二時，若前一個字元為姓氏，則後詞一個字元為名字；

(2) 詞彙長度為三時，若前兩個字元屬於複姓，則最後一個字元為名字；若前一個字元屬於姓氏，則最後兩個字元為名字；

(3) 詞彙長度為四時，若前兩個字元屬於複姓，則最後兩個字元為名字；若前一個字元屬於姓氏且接續兩個字元屬於複姓，或前兩個字元屬於複姓且接續一個字元屬於姓氏，則最後一個字元為名字；若前一個字元屬於姓氏

且接續一個字元屬於姓氏，則最後兩個字元為名字；

(4) 詞彙長度為五時，若前兩個字元屬於複姓且接續兩個字元屬於複姓，則最後一個字元為名字；若前一個字元屬於姓氏且接續兩個字元屬於複姓，或前兩個字元屬於複姓且接續一個字元屬於姓氏，則最後兩個字元為名字；

(5) 詞彙長度為六時，須前兩個字元屬於複姓且接續兩個字元屬於複姓，則最後兩個字元為名字。

將詞彙切割成姓氏與名字兩部分後，再依據對應的公式 2.1、2.2 或 2.3 計算其門檻值，高於門檻值的詞彙則歸類為中文人名。

以上的辨識模型有以下幾點條件限制：

(1) 處理詞彙的長度須介於二至六之間，否則判定為非中文人名；

(2) 若詞彙的前兩個字元 C_iC_j 或是冠夫姓的位置上 $C'_iC'_j$ 屬於百家姓中的複姓，則該詞彙直接假設為姓氏的開頭，並判別後方的名字是否符合門檻值；

(3) 若有出現冠夫姓 $C'_iC'_j$ 且為單姓 C_i 的情況，則只需考慮原姓氏 C_i 與後方的名字是否能符合門檻值；

(4) 若冠夫姓 $C'_i(C'_j)$ 且為單姓 C_i 單名 C_x 的情況，當辨認不通過時，則將 C_iC_x 視為名字的部分重新檢驗。

2.2 相鄰詞結合模型

專有名詞除了中文人名可以藉由大量語料來訓練字與字之間的結合機率，其他類別的專有名詞，如組織名稱，在此特徵上則不明顯。因此需另外藉由本身前後相鄰的辭彙來判別所屬類別。人名前後常見的詞彙有職稱（如「董事長」、「立委」）、動作句賓動詞（如「指稱」、「報導」）……等。

組織名的前後亦有關鍵詞可以輔助辨識，如公司、集團、銀行……等。在計算相鄰詞彙的集合機率之前，我們先將語料中的中文人名視為一個詞彙 W_P，組織名稱視為一個詞彙 W_O。如此的做法可以調整部分出現頻率低的專有名詞經計算後獲得較高的權重值。

2.2.1 方法一：專有名詞前後共現詞彙

將專有名詞前後相鄰的兩個詞彙合併計算 BF 值，藉此可以強調前後兩個詞彙共同出現的訊息。如下所示：

$$BF(W_{i-1}W_{i+1}) = \log_2 \frac{P(W_{i-1}W_{i+1})}{P(W_{i-1})P(W_{i+1})} \cong \log_2 \frac{\frac{f(W_{i-1}W_{i+1})}{N_2}}{\frac{f(W_{i-1})}{N_1}\times\frac{f(W_{i+1})}{N_1}} \geq \beta \quad (2.4)$$

其中 W_{i-1} 與 W_{i+1} 為相鄰於專有名詞旁的兩個詞彙
$P(W_{i-1})$與 $P(W_{i+1})$為兩個詞彙個別出現的機率
$P(W_{i-1}W_{i+1})$為詞彙 W_{i-1} 出現在 W_{i+1} 之前的機率
$f(W_{i-1})$與 $f(W_{i+1})$為詞彙詞 W_{i-1} 與 W_{i+1} 個別在語料庫中出現的頻率
$f(W_{i-1}W_{i+1})$為語料庫中詞彙 W_{i-1} 出現在 W_{i+1} 之前的頻率
N_1 為語料庫中詞彙總個數，N_2 為語料庫中 trigram 詞彙的總個數
β 為前後共現門檻值

並分別計算出現在人名與組織名前後詞彙的 BF_P、BF_O 值。在分類上藉由比較 BF_P、BF_O 兩個值的大小與是否符合門檻值的限制來辨別該詞彙所屬的類別為何。

$$Class_{BF2}(W_j) = \arg\max\left(BF_P(W_{j-1}W_{j+1}), BF_O(W_{j-1}W_{j+1})\right) \quad (2.5)$$

在辨認之前須檢驗 BF_P 與 BF_O 兩值是否超過 β 門檻值，否則詞彙 W_j 不是專有名詞。因此，若 BF_P 大於 BF_O 且詞彙 W_j 具有姓氏的開頭字元，則判定詞彙 W_j 為中文人名；反之，若 BF_P 小於 BF_O 則判定詞彙 W_j 為組織名稱。

2.2.2 方法二：專有名詞前後共現雙詞彙類別機率

為了減少門檻值調整的問題，我們另外考慮使用前後相鄰的兩詞彙，但是採用純粹的出現類別的機率模型。做法為統計出現在中文專有名詞、組織名稱以及一般詞彙等三種類別，其前面以及後面兩個詞彙所共同出現的類別機率。

$$Class_{Word}(W_i) = \arg\max_{j=0}^{2}\left(Prob_j(W_{i-1}, W_{i+1})\right) \quad (2.6)$$

其中 W_i 為欲辨認的詞彙

W_{i-1} 與與 W_{i+1} 為詞彙 W_i 的前後相鄰兩個詞彙

$Prob_j$ 為兩個詞彙 W_{i-1} 與 W_{i+1} 出現在三種類別前後的機率值

統計訓練語料中專有名詞 W_i的前後相鄰的兩個詞彙組 $W_{i-1}W_{i+1}$在不同的類別中出現的機率值，將雙詞彙組與類別機率建立成對照表以供測試時參照使用。相較於前面做法，此法並不需要訓練額外的門檻值參數；與其他相關的研究比較，此法並不需要事先建立所需的關鍵詞，因此可以減少大量的訓練時間與較少的人力介入。

2.3 前後詞類模型

根據專有名詞前後連續詞類，來預測某位置的詞類為專有名詞的可能性。所討論的專有名詞其分布情況如（表 1)。雖然大部分的專有名詞皆標示為 Nb，也有部分的專有名詞屬於其他的詞類。例如：文建會、交通部等在語料庫中的標記為地方名(Nc)；甚至有些專有名詞標示為動詞（成功(VH)　大學、建國(VA)　中學)。

因此我們就針對標記為名詞類以及動詞類詞彙的專有名詞來討論分類方法。

詞類 \ 頻率	訓練語料庫	中文人名	組織名稱
A	16797	0	8
Na	475356	7	1040
Nb	44187	19689	6837
Nc	124242	3	9712
Ncd	18168	0	9
Nd	46848	1	3
Nh	33705	0	123
VA	41680	0	12
VC	144889	0	48
VH	115572	0	175
VHC	10831	0	7
VJ	34688	0	6

表 1：專有名詞與詞類的頻率分布

我們統計了三類(中文人名、組織名稱、非前兩者)的前後詞類機率值，藉由計算出現在目前詞類以及前方與後方相連詞類的機率值，來評估所屬的類別。所設計的公式如下所示：

$$Fpos_i(class_i|t_1t_2t_3t_4) = \sum_{j=1}^{4}\lambda_j P(class_i|t_1t_2\cdots t_j)$$
$$\text{where } \lambda_j \in [0,1] \text{ and } \sum_{j=1}^{4}\lambda_j = 1 \quad (2.7)$$
$$\text{and } \lambda_1 < \lambda_2 < \cdots < \lambda_j$$

其中 $t_1t_2\cdots t_j$ 為相連的詞類

$Fpos_i(class_i|t_1t_2\cdots t_j)$ 表示連續詞類 $t_1t_2\cdots t_j$ 為類別 i 的機率值

λ_j 為詞類長度的權重值

因為連結詞類個數越多比單一詞類所能給予較多的訊息，故越後面權重所佔的比例須越大（$l_1 < l_2 < l_3 < l_4$)。

另外，為避免句首或句尾的情況而使得機率過於傾向於較短的連續詞類，我們分別計算詞類的前方與後方兩邊權重值，再根據其詞類個數來求得其平均權重值。例如：一段詞類標記為…，t_{k-4}，t_{k-3}，t_{k-2}，t_{k-1}，t_k，$k+1$，$k+2$，t_{k+3}，t_{k+4}，…，若 t_k所在位置為句首，則沒有前面的詞類標記，因此公式 2.7 將只有單一詞類的考量。結合前後詞類的情況下，前詞類所能提供的訊息將較於後詞類來的少。

因此加入前後詞類權重考量的公式如下所示：

假設一片段詞類標記為 $t_1t_2\cdots t_k\cdots t_n$，則

$$POS_i(t_k) = \frac{Fpos_i(t_{k-3}t_{k-2}t_{k-1}t_k)\times N_1 + Fpos_i(t_kt_{k+1}t_{k+2}t_{k+3})\times N_2}{N_1 + N_2} \quad (2.8)$$

其中 N_1 為目前詞類以及前面詞類的總數$(1 \le N_1 \le 4)$

N_2 為目前詞類以及後面詞類的總數$(1 \le N_2 \le 4)$

如此一來，只需透過計算每個類別的 weight 值即可辨別該詞類可能歸屬的類別為何。其計算公式如下：

$$Class_{POS}(t_k) = \arg\max_{i=0}^{2}(POS_i(t_k)) \quad (2.9)$$

其中 i 代表三種類別，分別為中文人名、組織名稱以及其他類別

2.4 經驗法則

經驗法則是由訓練語料觀察取得的，且以加強正確率為主。由於我們並沒有使用辭典式的法則（如常見的專有名詞)，因此會有許多同性質的組織名稱無法以相同的法則擷取出來，像是政府機關全部的院部會。各項經驗法則加如下：

(1) 若 $POS(W_i) \in$ {名詞，動詞} 則 W_i有可能為專有名詞，繼續以後的處理流程；反之，

則不是專有名詞，停止辨識。

(2) 若 $Right(W_i,1)\in$ {院，部}，$Right(W_{i+1},1)\in$ {署，司，局} 且 W_i 與 W_{i+1} 的長度大於二，則 W_i 與 W_{i+1} 歸類為組織專有名詞，並將 W_i 與 W_{i+1} 加入 cache 中。

(3) 若 $W_{i-1}\in$ {行政院，立法院，司法院，考試院，監察院}，$Right(W_i,1)\in$ {會，處} 且 W_i 長度大於二，則 W_i 歸類為組織專有名詞，並將 W_i 加入 cache 中。

(4) 若 $W_2\in$ {官員} 且 W_1 長度大於二，則 W_1 歸類為組織專有名詞，並將 W_1 加入 cache 中。

(5) 若 $W_{i-1}\in$ {國立，省立，縣立，市立，公立，私立} 且 $W_{i+1}\in$ {大學，高中，高職，商職，國中，國小}，則 W_i 歸類為組織專有名詞。

(6) 若詞彙 W_i 存在於 cache 中，則歸類為組織專有名詞。

2.5 綜合各類辨識方法

有鑒於上述各種分類法應用方式不同，故採用階層式的方式， 圖 3 為合併後的管線式模組。

共現雙詞彙模型與前後詞類模型是針對專有名詞識別，兩者皆為類別機率的模型，因此依不同的權重比例將兩項模型結合為一種辨識法（如下公式），用以區分不同類別的詞彙。由於詞彙本身的意義大於詞類能夠提供的訊息，因此在比例上共現雙詞彙模型的權重須要比前後詞類模型來的多。

$$Class(W_j,t_j)=\arg\max_{i=0}^{2}\left(\gamma\times Prob_i(W_j)+(1-\gamma)\times POS_i(t_j)\right) \quad (2.10)$$

其中 W_j 為欲識別的詞彙

t_j 為欲識別詞彙的詞類

γ 為兩項模型結合權重的比例，且 $\gamma>0.5$

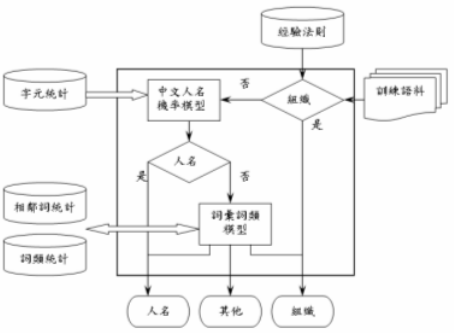

圖 3：分類辨識模型訓練流程

三、 實驗數據與討論

3.1 實驗環境與語料

本實驗架構在 PentiumR III 866，512 MB RAM 與 40GB HDD 的硬體上，作業系統搭配 WindowsR 2000 Server，資料庫利用 SQL Server 2000 來管理。開發軟體為 MicrosoftR Visual Basic 6.0 中文版。

中研院平衡語料庫 3.0 版是一個已標記詞類的現代用語語料庫，內容包括有報導、散文、評論……等 15 種文類，依數據顯示，專有名詞在報導類的文章中出現頻率較其他類別的文章為高，故採用這 6052 篇為整個論文所使用的語料庫。並依照文章出現先後順序，依照三比一的方式分割為訓練語料與測試語料。相關語料數據如表 2 及表 3 所示。

	訓練語料庫	測試語料庫
文章篇數	4539	1513
總句數	87799	27089
相異詞彙數	91703	49161
總詞彙數	2037746	632909
平均詞彙長度	1.694	1.695
已標記檔案大小	27 MB	8.4 MB
去除標記檔案大小	8 MB	2.5 MB

表 2：語料庫數據分析

	訓練語料庫	測試語料庫
長度一(w/o“的”)	716771	221974
長度二	1030048	320629
長度三	135146	41952
長度四以上	33348	10575
詞彙“的”	122433	37779

表 3：語料庫詞彙長度分布

中研院的詞類標記共計有 45 種類別，本篇論文將其粗略分為五類，分別為名詞(N.)、動詞(V.)、形容詞(Adj.)、副詞(Adv.)以及代名詞(Prop.)。其餘的詞類不包括在詞類統計的範圍裡。表 4 為語料庫詞類分布的情形。

	訓練語料庫	測試語料庫
名詞 (N.)	766440 (37.61%)	237024 (37.45%)
動詞 (V.)	527569 (25.89%)	164247 (25.95%)
形容詞 (Adj.)	16797 (0.82%)	5236 (0.83%)
副詞 (Adv.)	204301 (10.03%)	63285 (10%)
代名詞 (Prop.)	33705 (1.65%)	10366 (1.64%)
總計	1548812 (76.01%)	480158 (75.87%)

表 4：語料庫五大詞類分布

	Na	Nb	Nc	Ncd	Nd	Nf
長度一	49360	2301	9898	13533	2081	54988
長度二	368202	14066	78592	4469	35516	2729
長度三	52124	24312	32299	151	4773	23
長度四以上	5670	3508	3453	15	4478	99
總詞彙數	475356	44187	124242	18168	46848	57639
相異詞彙數	32044	13870	9224	218	2219	375

表 5：訓練語料庫名詞詞類頻率分布

	Na	Nb	Nc	Ncd	Nd	Nf
長度一	14670	751	3158	3991	738	16953
長度二	114997	4141	23708	1290	11144	737
長度三	15845	7810	9852	59	1653	12
長度四以上	1768	1166	1102	5	1444	30
總詞彙數	147280	13868	37820	5345	14979	17732
相異詞彙數	16931	5366	4629	179	1380	285

表 6：測試語料庫名詞詞類頻率分布

其中，除了少數的專有名詞(特別是組織名稱)，有標記為動詞類；大部分專有名詞分布的範圍在名詞類裡。表 5 與表 6 為名詞類中各類別頻率分布的情形。

由於中研院平衡語料庫內專有名詞的詞彙在詞類標記上並不只限於 Nb，所以我們事先利用人工分類的方式標記出屬於中文人名以及組織名兩類的專有名詞。在訓練語料內約有 19700 個中文人名以及約有 17980 個組織名稱。此處所指的人名不包括日文姓名(如：宮本武藏)、外文音譯名(如：柯林頓)、暱稱(如：小李)、單純姓氏名(如：陳先生)……等。

3.2 實驗模型設計

3.2.1 中文人名實驗

除了使用(2.1-2.3)三項公式，第四個公式(eq. 3.1)可說明姓氏與名字間並沒有顯著關係性存在著。

$$P(C_i)\times P(C_x)\times P(C_y)=\frac{\sum f'(C_i)}{\sum f(C_i)}\times\frac{\sum f'(C_x)}{\sum f(C_x)}\times\frac{\sum f'(C_y)}{\sum f(C_y)}\geq threshold_4 \quad (3.1)$$

在語料庫中，有許多長度為二的詞彙，第一個字元為姓氏，但卻不是人名。這些詞彙在全部的語料庫內占有 12.45%的比例，而人名在這些詞彙內所佔的比例為 1.34%。若設定較低的門檻值 $threshold_2$ 容易將其誤判為人名，所以初始設定門檻值 $threshold_1$ 與 $threshold_3$ 為零，單一調整 $threshold_2$ 值由 0 至 1 變化，取最高的正確率與召回率平均值(*F-score*)為最佳 $threshold_2$ 門檻值。實驗結果如圖 4　圖 7 所示，細實線代表召回率，細虛線代表正確率，粗實線代表 *F-score*：

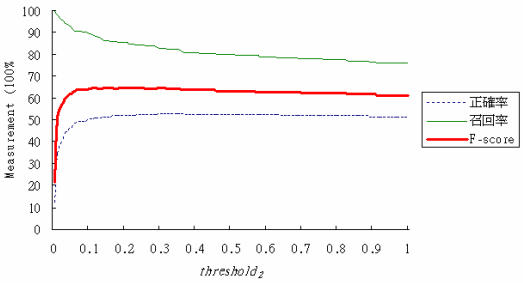

圖 4：門檻值 $threshold_2$ 走勢圖

由圖 4 我們可以看出，透過調整門檻值 $threshold_2$ 來限制須為常用的人名字元，故可以過濾掉部分長度為二且非中文人名的詞彙，但由於尚未限制其他條件，使得大部分長度為三且有姓氏字元開頭的詞彙也一併擷取出來，故正確率最佳只能提升到 52.9%。而召回率因為門檻值逐漸調高，慢慢減低到 75.9%左右。但最佳的 *F-score* 值約有 64.6%。

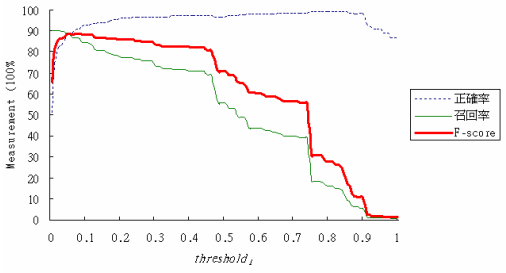

圖 5：門檻值 $threshold_1$ 走勢圖

threshold$_1$ 為限制姓氏出現頻率的門檻值，由圖 5 前段，*F-score* 大幅攀升 22.5%的情況可以清楚指出姓氏在中文人名的辨識上扮演著主要特徵。藉由調整門檻值 *threshold*$_1$ 即可將正確率提升到 89.1%，而召回率仍維持在 88.6%以上。當門檻值限制越高時，表示字元扮演姓氏所需的機率值越高，將使得符合的條件越嚴格。

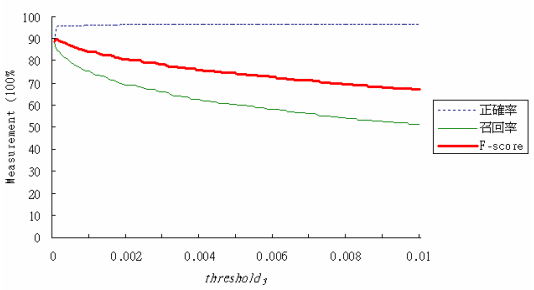

圖 6　門檻值 *threshold*$_3$ 走勢圖

圖 6 表示加上名字出現頻率限制門檻值 *threshold*$_3$的調整曲線，雖然召回率由 88.6%降為 85%，正確率仍由 89.1%上升為 95.2%，此數據表示，有常用的字元來當作名字的命名，才能使得一開始正確率的上升曲線大於召回率的下降幅度，因此可以使用此機率來排除非慣用雙名字集。

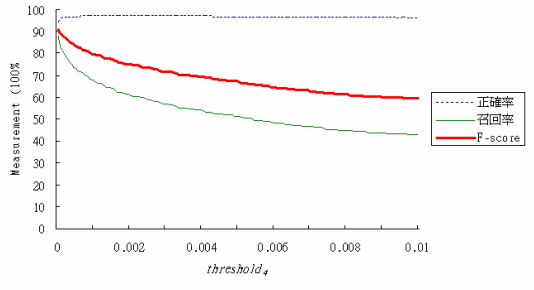

圖 7：門檻值 *threshold*$_4$ 走勢圖

門檻值 *threshold*$_4$代表著姓氏與名字之間共同出現的可能性。由圖 7 我們可以看出使用此門檻值之後，只有一開始正確率有些微的提昇，幅度約為 2.5%，但召回率的降福卻大於正確率的上升的幅度。換句話說，姓氏與名字之間並沒有絕對的關聯性。故後續實驗皆只採用公式 2.1~2.3。

藉由結合前面三種中文人名的辨別方法，我們可以快速的辨認出大部分的中文人名，並減低後續步驟辨識錯誤的可能，以提高整體的效能。將使用訓練後的三個門檻值應用在測試語料的中文人名辨識上，處理 369325 個帶有姓氏詞彙，共有 5992 的中文人名。測試結果：辨識出 5504 個中文人名，其中有 5155 為語料中正確的人名，因此正確率有 93.7%，召回率為 86%。

3.2.2 相鄰詞實驗

根據專有名詞前後相鄰的兩個詞彙共同出現的機率來評估目前詞彙可能的類別，並且討論合併中文人名模型辨識後效能的影響。

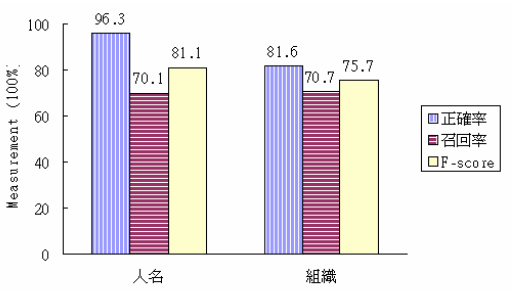

圖 8：前後共現詞彙模型

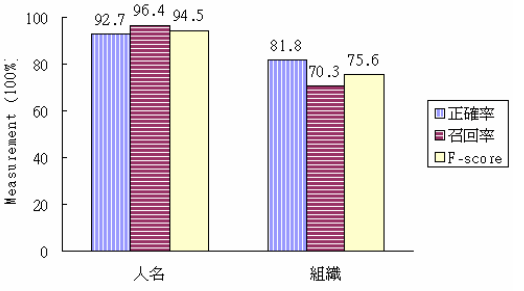

圖 9：中文人名模型＋前後共現詞彙模型

整個訓練語料中前後雙詞彙的組合共有 1095243，我們只取與中文人名或是組織名共同出現的雙詞彙，其組合只剩下 22313。語料中辨識人名的相鄰共現雙詞彙以｛(記者，台北)，(主任，表示)，(總經理，指出)……｝等職稱加上動作句賓動詞類為主；組織名稱的相鄰共現雙詞彙則以｛(台北市，國中)，(彰化縣，國小)，(與，董事長)……｝等地名加

上組織類別，或是伴隨著職稱共同出現。

在圖 8 的測試語料實驗中，即使不使用人名機率模型，對中文人名也能有 96.3%的正確率，而組織名稱也有 81%的正確率。當專有名詞前後共現雙詞彙頻率不高時，則會無法辨別該詞彙的類別，故在這個實驗上的整體召回率普遍不高，約為 70%。

若是事先使用中文人名機率模型的輔助之後，再經由相鄰詞共現詞彙模型的辨識，藉由前後共現詞彙的幫助，將之前無法通過門檻值的中文人名(約 683)順利辨識出來，雖然使得正確率稍微下降為 92.7%，但卻讓召回率一口氣提升到 96.4%。此舉有益於人名的辨識率，但卻對組織辨識無多大的影響。

3.2.3 前後詞類實驗

雖然專有名詞的詞類標記不能藉由語料庫標記 Nb 的詞彙決定，但我們可以藉由詞類以及其相鄰的前後詞類的排列來判別。如:「廿九日(Nd) 到(P) 台灣(Nc) 大學(Nc) 觀賞(VC)」此例，雖然台灣標記為地方名(Nc)，但由於加強前詞類與後詞類的訊息影響，將有助於辨識的結果。

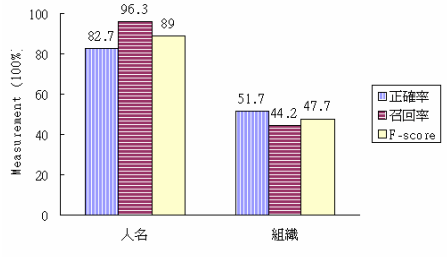

圖 10：前後詞類模型

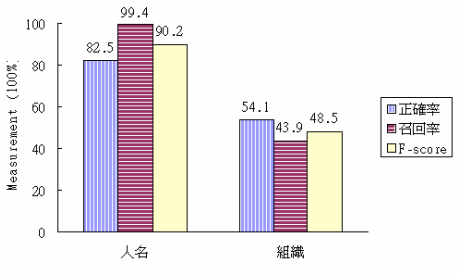

圖 11：中文人名模型＋前後詞類模型

由於中文人名的詞類大部分皆標示為 Nb，且前後的詞類變化幅度不大，不是標記為 Na 的職稱，就是各類動詞，因而能夠涵蓋 99.4%語料中的人名，即使不使用人名機率模型，召回率依舊有 96.3%，正確率維持在 82.5%。組織名稱因為涵蓋詞類範圍較廣，並沒有辦法有效的辨認出來，故使得組織名稱辨識效能無明顯的助益（*F-score* 為 48.5%)。

3.2.4 經驗法則實驗

由於經驗法則是針對組織名稱的專有名詞而設計的，因此正確率遠較其他方法來的高。整個模型在不使用人工經驗法則與暫存機制，組織名稱的辨識即可以達到 87.3%的正確率，但召回率僅能維持在 76.8%。藉由經驗法則的套用，我們可以先行辨識出符合已知條件的組織專有名詞，避免之後的程序誤判，同時可以提高系統執行的效能；若不使用其他的辨識模型而單純使用經驗法則，組織名稱辨識的正確率能夠達到 99.9%，召回率也有 26.2%。但即使經驗法則的召回率有 26.2%的效果，對於整體效能的提升只有 5.1%的影響。

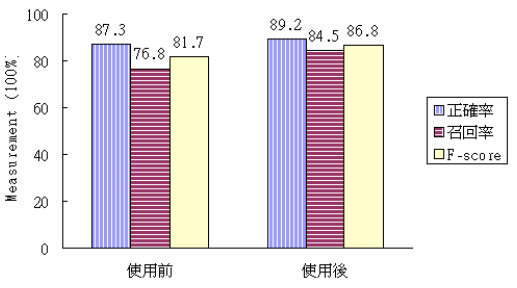

圖 12：經驗法則＋前後共現詞彙模型＋前後詞類模型

四、 結論

在本論文中，我們提出了專有名詞分類的方法，藉由建立專有名詞的前後相鄰共現詞彙與結合了前後詞類模型來達到中文人名以及組織名稱的分類。其中，人名不使用傳統的字元機率即可達到 94.6%的正確率與 93.3%的召回率，若搭配了中文字元機率模型，召回率可達到 99%；至於組織名稱方面，在不使用經

驗法則的情況下，正確率可達 87.3%，而召回率只有 76.8%，若是套用了經驗法則，由於包含了部分的常見組織名稱，故召回率可提升至 88.9%。

與其他相關研究比較中，本篇論文在不使用經驗法則的前提下，整個系統並不需要針對人名或組織名稱前後可能出現關鍵字等建立各項條件與所需相關詞彙表。而只需要在訓練語料中對於已知專有名詞統計前後詞彙與詞類出現的機率即可達到可靠的效能（中文人名與組織名稱的 *F-score* 各為 93.9%與 81.7%）。

統計字元在文章中扮演姓名的機率即可在中文人名辨識上有滿意的 90%以上效能表現，而且有姓氏作為人名的開頭字元，在結合中文斷詞的處理上較易於辨識左邊界，若再搭配如職稱、相關動詞等出現在人名前後的關鍵詞彙，則有助於提高人名的召回率。至於組織名稱辨識方面，由於有縮寫或簡稱的關係，在沒有已知的組織辭典比對下，利用前後共現相鄰詞的幫助下，則有 80%的效能，若文章經過詞性標記，則詞類亦能提供少許的辨識線索。相較於前後共現相鄰詞的做法，採用關鍵詞彙以及 rule-based 方式不僅有著較高的召回率，辨識所需時間亦會減少。但相對的，針對於關鍵詞彙與辭典的維護需要額外人力付出。

五、參考文獻

[1] Alessandro Cucchiarelli and Paola Velardi, (1999), "A Statistical Technique for Bootstrapping Available Resources for Proper Nouns Classification", *Information Intelligence and Systems, 1999*; *Proceedings, 1999 International Conference*, pp. 429-435.

[2] Douglas E. Appelt, Jerry R. Hobbs, John Bear, David Israel, Megumi Kameyama, Andy Kehler, David Martin, Karen Myers, Mabry Tyson, (1995), "SRI International FASTUS System MUC-6 Test Results and Analysis", *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Columbia, Maryland, pp. 237-248.

[3] Hsin-Hsi Chen, Yung Wei Ding, Shih-Chung Tsai, and Guo-Wei Bian, (1998), "Description of The NTU System Used for MET2", *Proceedings of the Message Understanding Conference, Fairfax, VA*, 29 April – 1 May.

[4] Keh-Jiann Chen and Chao-Jan Chen, (2000), "Knowledge Extraction for Identification of Chinese Organization Names", *Proceedings of ACL workshop on Chinese Language Processing*, pp. 15-21.

[5] G. Paliouras, Vangelis Karkaletsis, Georgios Petasis and Constantine D. Spyropoulos, (2000), "Learning Decision Trees for Named-Entity Recognition and Classification", *Proceedings of the Workshop "Machine Learning for Information Extraction", European Conference in Artificial Intelligence, Berlin, Germany*.

[6] Jen-Chang Lee, (1994), "Identification of Proper Nouns in Chinese Texts", *Department of Computer Science and Information Engineering National Taiwan University, Master Thesis*, June.

[7] Jing-Shin Chang, Shun-Der Chen, Ying Chem, John S. Liu, and Sur-Jin Ker, (1991), "A Multiple-corpus Approach to Identification of Chinese Surname-Names", *Proceedings of Natural Language Processing Pacific Rim Symposium, Sungapore*, pp. 87-91.

[8] Jing-Shin Chang and Keh-Yih Su, (1997),

"An Unsupervised Iterative Method for Chinese New Lexicon Extraction", *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 2, NO. 2, pp. 97-148, August.

[9] Heng Ji and Zhensheng Luo, (2001), "A Chinese Name Identifying System Based on Inverse Name Frequency Model and Rules", *Natural Language Processing and Knowledge Engineering (NLPKE) Mini Symposium of the 2001 IEEE International Conference on System, Man, and Cybernetics (SMC2001)*.

[10] Jian-Yun Nie, Marie-Louise Hannan, and Wanying Jin, (1995), "Unknown Word Detection and Segmentation of Chinese using Statistical and Heuristic Knowledge", *Communications of the Chinese and Oriental Languages Information Processing Society*, Vol. 5, pp. 47-57.

[11] Kim-Teng Lua, and Kok-Wee Gan, (1994), "An Application of Information Theory in Chinese Word Segmentation", *Computer Processing of Chinese and Oriental Language*, Vol. 8, NO. 1, pp. 115-124, June.

[12] Silviu Cucerzan, and David Yarowsky, (1999), "Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence", *Proceedings, 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, pp. 90-99.