中文辭彙岐義之研究—斷詞與詞性標示

彭載衍　　張俊盛

國立清華大學資訊科學研究所

摘要

目前電腦應用在中文處理方面，對於斷詞已能達到相當高的正確率（95%以上），然而在中文詞性標示的基礎研究上，仍未有相當的研究及令人滿意的結果。分析其原因，不外乎中文詞性訂定尚無標準；中文句法較複雜，變化較大，想要以法則分析法來運作似乎不太容易；還有缺乏良好的含有詞性及頻率的電子詞典。

目前我們已擁有含有詞性及頻率的電子詞典，且捨棄傳統法則分析的方法，改以機率式的方法，來作詞性的標示。在這個系統裡，我們用了幾個模型，並且分析比較了它們的結果，以期達到最好的效果。此外，對於部份的未知詞，詞長爲一或二的，我們也做了處理；還有中文姓名的部份，也在我們討論範圍內。

在本篇的後面部份，我們作了正確率的評估與錯誤的分析，以利我們了解什麼是發生錯誤的主因，尋求改進之道。

一、簡介

在英文或其他西方語言中，並沒有斷詞的問題，然而在中文的處理上，斷詞卻是首先面臨的問題。什麼是中文斷詞？所謂中文斷詞是指將輸入的中文句子，依據語意及文法結構，切割文句至以詞爲基本單位的工作。文句經過斷詞後，形成以詞爲基本單位的句子，再進一步的分析是將句內的每個詞標上詞性，然而詞性如何正確地標示上去，正是我們所要探討的。

中文斷詞的研究，在近幾年來已相當成熟。歸納以往所使用的方法，基本上可以分成兩大類。第一類是法則式的斷詞方法，第二類是以統計數據爲基礎的機率式作法。在法則式的作法中，有何[1]、陳[2,3]和李[4]。第二類有蔡[5]的鬆弛法、Sproat 和 Shih[6]、張[7]、江和蘇[8]等人。

中文詞性標示處於起步階段，已有陳[9]，張[10]，張簡和李[11]等人做過這方面的研究。然而詞性標示的研究在英文發展甚早，到目前爲止，可說是到達相當成功的地步，正確率已高達百分之九十六以上。所發展出來的方法相當多，基本上亦可分成兩大類。第一大類是法則式的方法，第二大類是機率式的方法。

法則式的方法發展甚早，在1971年 Greene and Rubin [12] 發展出 TAGGIT，在Brown corpus 上測試的結果，正確率達77%。一般認爲法則式的方法正確率不夠高，然而 Brill [13] 在1992年提出了一個新方法，正確率高達95%，說明了法則式的作法亦可達到高正確率。

機率式的作法一般包含兩個不同的機率值，一個是字本身所能貢獻的機率值（lexical probability），另一個是前後文字所能提供的機率值（contextual probability），大多數的公式都是由這兩個所組成的。Leech[14]的CLAWS，採用詞性的二元接續（bigram）與字本身所擁有的詞性分配頻率。從文獻上得知CLAWS對測試資料達到96.7%的正確率。Church[15]採用的是詞性的三元接續表（trigram），方法與Leech差不多。此外尚有DeRose[16]發展的VOLSUNG，其用的方法與CLAWS的類似，但若加上片語（idiom）的處理，在LOB corpus上可達99%的正確率。林和蘇[17]等人，利用學習（learning）和結合（merging）詞性二元接續表與三元接續表的方法來處理詞性標示，正確率亦達95%以上。

二、斷詞與詞性標示的機率式方法

在這一節裡，首先介紹斷詞的方法，接著是詞性標示，然後是斷詞與詞性標示整合在一起的機率模式。此外，尚有未知詞與中文姓名的處理。

機率式的斷詞方法

假設輸入系統的中文字串爲$C_1C_2...C_n$，輸出爲詞串$W_1W_2...W_p$（$p<=n$），如何決定輸出的詞串是倚賴一類似0階的馬可夫模式（Markov Model）(2.3)，計算出各種詞串組合的機率值，找出擁有最大值的詞串，此時詞的組合爲在機率模式下的最佳組合，也就是被輸出的詞串。

公式的推導如下：

$$P(W_1,W_2,\ldots,W_p|C_1,C_2,\ldots,C_n) \qquad -(2.1)$$

$$\approx P(W_1,W_2,\ldots,W_p) \qquad -(2.2)$$

$$\approx P(W_1)P(W_2)\ldots P(W_p) \qquad -(2.3)$$

$$=\prod_{i=1}^{p} P(W_i)$$

其中 $P(W_i)$ 表示句子內第 i 個詞的出現機率值

式子 (2.1) 表示在給定中文字串 $C_1C_2\ldots C_n$ 下，詞串 $W_1W_2\ldots W_p$ 的出現機率。當這個機率值高時，表示此狀況下的詞串出現的機會較大，通常正確的斷詞結果，即是機率模式下的最佳組合，因此我們能以機率的模式來解決斷詞的問題。

然而式子 (2.1) 中的條件機率，需要統計相當大的語料才可以得到，在實做上幾乎是不可能的，因此需要做一些假設，以得到一個較簡單的機率模式，能在電腦上快速運作。

首先，假設詞在句子內的出現機率與詞在整個大語料內的出現機率並無不同，那麼式子 (2.1) 就可以化簡爲式子 (2.2)。此外，我們再假設詞與詞間的出現並無關連，彼此間是互相獨立的，那麼式子 (2.2) 可再進一步被化簡爲式子 (2.3)。至此我們將斷詞問題轉變成找出一詞串 $W_1W_2\ldots W_p$ 使得式子 (2.3) 擁有最大值。因此機率式的斷詞模式可被寫成一個如下公式 (2.4)。

$$Max\{\prod_{i=1}^{p} P(W_i)\} \qquad -(2.4)$$

在眾多的詞串中，滿足式子 (2.4) 的詞串即是此模式下斷詞的輸出結果。

斷詞與詞性標示的解決方法

假設斷詞完後的詞串爲 $W_1W_2...W_k$，所要標示的對應詞性爲 $T_1,T_2,...,T_k$。詞性標示所用的公式是使用詞性二元接續表的模式。表示成式子(2.5)。

$$\prod_{j=1}^{k} P(T_j|W_j)P(T_j|T_{j-1}) \qquad -(2.5)$$

現在我們將說明如何將斷詞與詞性標示合在一起成爲一個單獨的模式。若將斷詞視爲事件 A，發生的機率爲 P(A)；詞性標示視爲事件 B，發生的機率爲 P(B)。事件 B 是發生在事件 A 之後，因此事件A與B皆要發生的機率爲此兩機率之積P(A)P(B)。從機率的觀點來看，斷詞與詞性標示合在一起的公式爲式子 (2.3) 與式子 (2.5) 的合併，成爲下面的式子 (2.6)：

$$\begin{aligned}&\prod_{i=1}^{k} P(W_i)P(T_i|W_i)P(T_i|T_{i-1})\\ &=\prod_{i=1}^{k} P(W_i)\frac{P(W_i,T_i)}{P(W_i)}P(T_i|T_{i-1})\\ &=\prod_{i=1}^{k} P(W_i,T_i)P(T_i|T_{i-1})\end{aligned} \qquad -(2.6)$$

其中 $P(W_i,T_i)$ 表示第 i 個詞 W_i 與其詞性爲 T_i 的出現機率。T_0 是句子開頭的標示。

機率式的斷詞詞性標示可被表示爲底下的公式：

$$Max\{\prod_{i=1}^{k} P(W_i,T_i)P(T_i|T_{i-1})\} \qquad -(2.7)$$

在詞性標示的處理上，句子結束這個訊息亦有助於標示的正確性，因此我們加入句子末尾的標示（記爲 $T_{\&}$）於公式（2.7）內。將式子（2.7）改爲式子（2.8）。

$$Max\{(\prod_{i=1}^{k} P(W_i,T_i)P(T_i|T_{i-1}))P(T_{\&}|T_k)\} \quad -(2.8)$$

未知詞的解決模式

在斷詞與詞性標示的系統裡，詞典所收錄的詞受到記憶體大小限制和執行速度的要求，不可能將所有的詞納入。即使在系統設備允許的狀況下，想要將所有詞納入，亦需花費相當大的成本。此外，詞是具有時間性的特徵，古代所使用的詞有許多在今日已不被使用，當代依需求、流行或其他因素，亦創造出許多新的詞彙。還有姓名、譯名、專有名詞等，這一類名詞數量相當多，須要常常做更新以符合需求。因此，詞典是無法收錄所有的詞，僅能依據使用者的需求和系統的要求，適當地涵蓋所需求的辭彙。

在這裡我們所稱的未知詞是指詞典未收錄的詞。既然無法要求詞典擁有所有的詞，而未知詞會影響系統的斷詞與詞性標示的正確性，因此，未知詞的處理是必須的。未知詞的種類很多，如：譯名、姓名、專有名詞等。它們被使用的程度依文章的性質而異，所使用的單字在頻率方面也有差異，因此，這些是需要分開討論的。因爲所得到的資料有限，我們只將姓名的部份在下一節特別討論，其餘的未知詞不再細分類，將在底下討論。

根據統計，文章中出現的詞長度爲一或二的佔了絕大部分，所以我們將詞長爲二的未知詞作爲首先解決的問題，詞長超過二的未知詞暫不討論。

假設未知詞 W 分別由左邊的字 M_1 與右邊的字 M_2 所組成，表示爲 $W = M_1M_2$ ，它的詞性爲 T 。因爲一個詞所擁有的詞性不只一個，對未知詞而言，每種詞性似乎都有可能，但是以名詞、動詞、形容詞居多。所以我們只針對這三類詞性做估計，估計名詞、動詞、形容詞出現的可能機率。

估計的公式如下：

$$P(T|M_1,M_2) \approx P(T|LM = M_1 \& RM = M_2)$$
$$\approx P(T|LM = M_1) \times P(T|RM = M_2) \qquad -(2.9)$$

其中 LM 表示左邊的字， RM 表示右邊的字。

$$P(T|LM = M_1) = \frac{P(LM = M_1 \& POS = T)}{P(LM = M_1)}$$
$$P(T|RM = M_2) = \frac{P(RM = M_2 \& POS = T)}{P(RM = M_2)}$$

其中 POS 表示詞 W 的詞性標示，

$P(LM = M_1 \& POS = T)$ ， $P(RM = M_2 \& POS = T)$,
$P(LM = M_1)$ ， $P(RM = M_2)$

分別可由對詞典的統計而來。

整個計算未知詞 W 的公式可被表示爲：

$$P(W,T) = P(M_1M_2,T)$$
$$\approx P(UNW)P(M_1M_2|UNW)P(T|M_1M_2 \cap UNW)$$
$$= P(UNW)P(M_1M_2|UNW)P(T|M_1M_2) \qquad -(2.10)$$

其中 UNW 表示未知詞；P(UNW) 爲未知詞的出現頻率，可以從測試的語料估計得到；$P(M_1M_2|UNW)$ 是未知詞爲 M_1M_2 的機率，可從底下的估計法求得；$P(T|M_1M_2)$ 是未知詞 M_1M_2 詞性爲 T 的機率，可從式子 (2.10) 估計求得。

$P(M_1M_2|UNW)$ 的估計方法。

$P(M_1M_2|UNW) \approx P(LM = M_1|D)P(RM = M_2|D)$

其中 D 表示詞典。

以式子 (2.10) 代入上一節的斷詞詞性標示系統，實驗顯示的確能改善詞性標示中未知詞所引起的問題。

中文姓名的處理

完整的中文姓名是由姓的部份與名的部份所共同組成，如：張無忌。張是姓，無忌是名。因爲在文章中姓與名不一定要成對出現，如：張先生，只出現姓的部份：無忌孩兒，只出現名的部份。因此我們分別給予姓和名不同的詞性，'fn' 和 'gn' 來反應上述的情況。

中文的姓可分成單姓和複姓，名可分爲單名和複名，這些區別在機率的模式底下所使用的的公式差異不大，不同的地方在所使

用的統計資料不同，底下所介紹的公式，是參考張 [18] 所得到的。

單姓的情況：

假設 W 是一單字，其爲單姓的機率爲：

$$P(W,fn) = P(FN1)P(W|FN1) \qquad -(2.11)$$

其中 $P(FN1)$ 是指語料中出現單姓的機率，可由統計語料得到。 $P(W|FN1)$ 是指詞 W 的標示爲 fn ，在單姓底下的條件機率。

複姓的情況：

$W = C_1C_2$ ，其爲複姓的機率爲：

$$P(W,fn) = P(FN2)P(W|FN2) \qquad -(2.12)$$

其中 $P(FN2)$ 是指語料中出現複姓的機率，可由統計語料得到。 $P(W|FN2)$ 是指詞 W 的標示爲 fn ，在複姓底下的條件機率。

接著將介紹名字的機率公式。

單名的機率公式爲：

$$P(W,gn) = P(GN1)P(W|GN1) \qquad -(2.13)$$

其中 $P(GN1)$ 是指語料中出現單名的機率。 $W = C_1$

複名的機率公式爲：

$$P(W,gn) = P(GN2)P(W|GN2)$$
$$= P(GN2)P(C_1C_2|GN2)$$
$$\approx P(GN2)P(LG = C_1|LG)P(RG = C_2|RG) \qquad -(2.14)$$

其中 $P(GN2)$ 是指語料中出現複名的機率。 $W = C_1C_2$ ： LG 表示複名的第一字， RG 表示複名的第二字。

從式子 (2.11) 到式子 (2.14) ，是以一個機率的方法來解決中文姓名部份，與前面的斷詞詞性標示是一致的，能合併成爲一個系統。

三、實驗結果與錯誤分析

一個模型的好壞，除了理論的嚴密性外，還需要實驗的驗證。因此，在這一節中，我們將列出實驗的結果，並且分析錯誤的成因，進而對模型做一些修正，以期得到較好的結果。

基線模式

在利用詞性二元接續表斷詞與詞性標示之前，我們先介紹一個基線 (base line) 的斷詞與詞性標示模式，作爲詞性接續表模式的比較對象，使得實驗結果較爲客觀。這個基線模式是這樣的：斷詞與詞性標示所用的機率公式爲

$$MAX\{\prod_{i=1}^{i=P} P(W_i, T_i)\} \qquad -(3.1)$$

其中 P 爲句內的總詞數

基線模式是在詞典內挑選詞 W 的詞性 T ，使得 P(W,T) 是在詞 W 下的最大頻率值。這樣的作法，並不考慮詞性與詞性間的接連關係，與斷詞的作法相類似。

斷詞與詞性標示結果及分析

從實驗的結果可看出，基線模式的正確率爲 70.8% ，召回率爲 71.4% ，這樣的結果並不令人滿意。剛才已提過，這個基線模式並未充分利用有助於結果正確的訊息—詞性間的接連關係。例如'然後按 Enter 鍵 ' 在基線模式底下標示爲：

```
|然後|按|Enter|鍵|
|adv |p | np   |nc|
```

若考慮詞性間接連的關係，標示的結果成爲：

```
|然後|按|Enter|鍵|
|adv |v | np   |nc|
```

'按'的介詞機率最高，因此在基線模式中被標示爲介詞。在考慮詞性接連關係的模式中，標示結果變成爲動詞。然而標示爲介詞是錯誤的，動詞才是正確，所以使用詞性間接續關係的機率模式應會有許多改進。

本實驗所用的是詞性二元接續表的模式，僅考慮與前一個詞的接連關係，其他的關係並不考慮。實驗的結果列於表一，其中檔案一至檔案八爲訓練資料，檔案九爲測試資料，明顯地看出檔案九並不因未參與詞性二元接續表的建立而結果表現較差，其正確率與召回率仍有不錯的表現。因此，詞性二元接續表在類似的文章中是具有好的轉移性，也能在它們中使用同樣的接續表。

因爲接續表有好的轉移性，所以在統計與分析方面，我們並不將檔案九分開，而是全部合在一起，表中所列的平均值部份亦包含檔案九。

比較基線模式與詞性二元接續表模式的結果，在平均表現上，後者正確率較前者多了 5.0% ，召回率多了 5.4% ，這點說明了二元接續表模式確實有助於詞性標示，大約能提升百分之五左右。

	總字數 A	總詞數 B	標示後總詞數 C	標示正確詞數 D	標示正確字數 E	詞的正確率 D/C %	詞的召回率 D/B %	字的正確率 E/A %
檔案1	3275	2137	2161	1792	2818	82.9	83.9	86.0
檔案2	3177	2076	2124	1551	2365	73.0	74.7	74.5
檔案3	3412	2177	2199	1564	2465	71.1	71.8	72.2
檔案4	3340	2118	2137	1651	2623	77.3	78.0	78.5
檔案5	2667	1721	1753	1299	2020	74.1	75.5	75.8
檔案6	4701	3044	3049	2418	3777	79.3	79.4	80.3
檔案7	4605	3021	3063	2292	3571	74.8	75.9	77.5
檔案8	3210	2150	2196	1571	2381	75.1	73.1	74.2
檔案9	3091	2018	2056	1583	2445	77.0	78.4	79.1
	31478	20462	20738	15721	24465	75.8	76.8	77.7

表一 ： 使用詞性二元接續表的標示結果

錯誤分析

以下是從測試資料中摘選出一些錯誤的例子，例句中底線表示詞性標示的主要錯誤。

(1) |檔案|名稱|後面|加上|兩|個|字|元|
|nc |nc |nc |v |q |cl|nc|nc|
(2) |包含| 1-2-3 |的 |群|組|視|窗|
|v | np |ctm|cl|nc|v |nc|

(3) |您 |可以|按照|姓|氏|或|識別|碼|來|排序|
|pron|aux|v |nc|nc|cj|v |cl|v |nc |
(4) |本|章|將 |說明|如何|使用|資料庫|函數|
|cl|nc|aux|v |adv|v |nc |nc |
(5) |使|用於| 1-2-3 |舊|版本|
|v |v | np |a |nc |
(6) |然後|再 |加上|工作|表|表|名|
|adv |adv|v |nc |v |v |nc|

在錯誤分析方面，除了機率模式本身標示錯誤外（例句（5）(6)），還有其他非模式本身所產生的錯誤，我們分析的結果可歸納成兩類。第一類是詞典內未含有的詞所產生的錯誤，這類的詞我們稱爲未知詞（例句（1）（2））；第二類是詞典內雖含有詞但並不含有正確標示的詞性，我們稱這類詞性爲未知詞性（例句（3）(4)）。這兩項因素對於標示的正確性有絕對的影響，機率式的斷詞詞性標示模式，若未加上特殊的處理，並無法對未知詞和未知詞性做正確的標示，因此它們將造成絕對性的錯誤，影響甚巨。平均而言，未知詞有 4.9% ，未知詞性有 4.3% ，兩者合起來共佔9.2%。系統標示的錯誤率爲 24.2% ，因未知詞與未知詞性所產生的錯誤爲 9.2% ，因此，系統本身所產生的錯誤爲 24.2% - 9.2% = 15.0% 。

在未知詞分析上，我們對它做了更進一步的分析，分成單字詞、雙字詞、三字詞、四字詞四類，將各種所佔的詞數與比例求出，列於表二。從表二中看出未知詞的種類絕大部分是屬於單字詞或是雙字詞，兩者合起來共佔了 92% 。從這個訊息得知，未知詞的處理首要在單字詞與雙字詞上。

	缺少詞數 A	單字詞數 B	雙字詞數 C	三字詞數 D	四字詞數 E	單字詞百分比 %	雙字詞百分比 %	三字詞百分比 %	四字詞百分比 %
總計	993	505	410	75	3	50.9	41.3	7.6	0.3

表二 ：未知詞的統計

未知詞處理結果及分析

前面已說明過未知詞佔了測試資料的 4.9% ，這一部份使得系統無法對它處理，除了本身詞與詞性的標示錯誤外，對於接連的詞與詞性亦造成影響，因此所發生的錯誤將大於其所佔的比例。對未知詞的處理可分爲兩部份，第一部份是未知詞中比例最高的單字詞，第二部份是雙字未知詞。首先介紹單字未知詞的處理。

單字未知詞的處理結果

在原始系統處理過程中，單字未知詞無法與鄰近的字結合成詞，因此會被標示爲單字詞，但並無給予詞性，所以我們要做的工作是給未知單字詞一個詞性。

詞性標示的過程中必須用到二元接續表，因此我們利用它來協助標示單字未知詞的詞性。只要給未知詞每種詞性一個相同頻率值，那麼系統所決定的詞性將僅由其前後的詞性與後面的詞所決定，這個詞性滿足機率公式的最大值。說明如下：

若 W_u 爲單字未知詞，爲句內第 i 個詞

設定 $P(W_u, T_i) = C$ $\qquad$ $n(T_i) = s$

其中 C 爲常數，$n(T_i)$表示詞性 T_i的個數，s 爲所有詞性總數。

實驗的結果顯示，加上單字未知詞的處理後，正確率變爲 77.2%，召回率變爲 78.1%，與先前結果比較，正確率增加了 1.4%，召回率增加了 1.3%。

雙字未知詞的處理結果

這個部份是用先前所提的雙字未知詞方法，再加上單字未知詞的處理。平均而言，正確率爲 78.6%，召回率爲 76.7%，與僅單字未知詞的處理結果比較，加上雙字未知詞後，正確率增加了 1.4%，然而召回率卻下降了 1.4%，這樣的結果並不令人滿意。

召回率的提高必須靠標示正確的詞才能提升；然而正確率的增加卻並不一定代表著標示正確的詞增加。召回率的下降代表著有過多的詞被錯誤地合成雙字詞。因此，在修正的作法裡，我們設法減少二字詞的合成，減少犯錯的機會。從實驗的經驗得到單字詞中詞性爲介詞、方位詞、連接詞、量詞等詞性都較少有二字合成詞的現象，所以我們將具有這類詞性的單字詞排除在二字詞的合成範圍內，以期減少錯誤的詞合成。

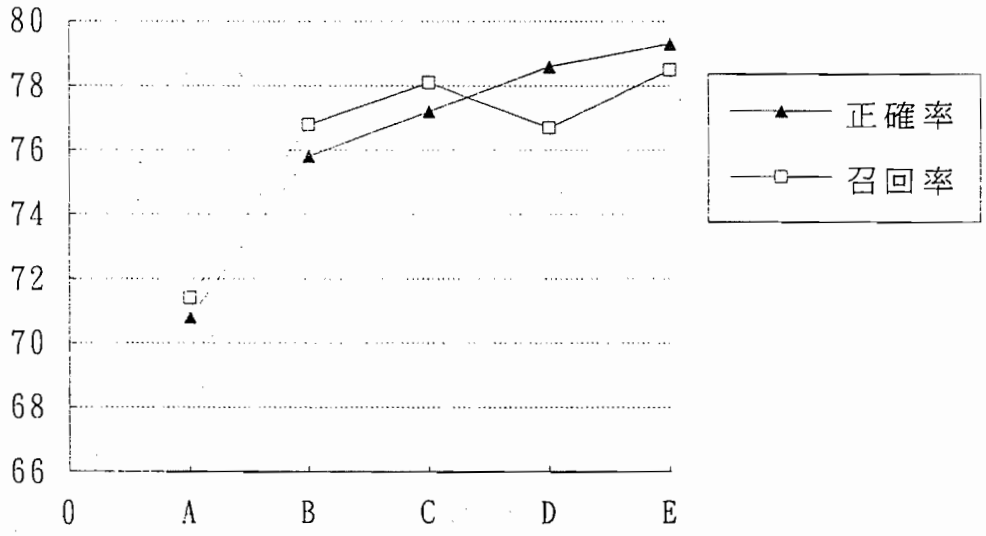
實驗顯示，正確率較未修正前多了 0.7%，召回率多了 1.8%，這點說明了改進的方法確實有效。與沒有未知詞處理的二元接續表模式相比較，得到整個未知詞的處理，在正確率上增加了 3.5%，召回率上增加了 1.7%。

綜合前面實驗的結果，繪成圖一，以方便觀察正確率與召回率的變化。經過一連串的改進後，正確率較基線模式增加了 8.5% ，召回率增加了 7.1% 。

表三所列的是測試資料的分析表，有未知詞、未知詞性及詞所含有的詞性數目等項的百分比。

未知詞	未知詞性	一個詞性	二個詞性	三個詞性	四個詞性	五個詞性	七個詞性
4.9%	4.3%	51.9%	26.7%	8.0%	1.6%	2.6%	0.03%

表三 ： 測試資料分析表



	基線模式 A	二元接續表 B	接續表+單字未知詞 C	接續表+單字未知詞+雙字未知詞 D	接續表+單字未知詞+雙字未知詞(修改) E
正確率	70.8	75.8	77.2	78.6	79.3
召回率	71.4	76.8	78.1	76.7	78.5

圖一 ： 正確率與召回率一覽圖

中文姓名處理結果及分析

姓名的測試資料是新聞類的語料，共有三個檔案，大約三千五百個詞。實驗結果將姓與名分開來，各別計算其正確率與召回率。

表四所列的是姓、名標示正確與錯誤的數目。表五是根據表九內的數據，所求出姓與名的正確率和召回率。姓的正確率達到86.7% 召回率更高達 93.3% ，這數據說明姓的處理是相當不錯的。名的正確率是 78.3% 召回率是85.5%，對於名的處理也有不錯的結果。表六顯示姓與名在測試資料中分別所佔的比例，平均上姓佔 2.4% 、名佔 2.2% ，共是 4.6%。除了從處理姓名的觀點來分析結果外，我們亦須從整體的標示系統來看姓名的處理效果如何？在表七中顯示當系統加上姓名的處理後，姓的部份能增加 2.0% ，名的部份能增加 1.4% ，總共是 3.4%。

	姓總數 A	名總數 B	標示正確姓數 C	標示錯誤姓數 D	標示正確名數 E	標示錯誤名數 F	標示總合姓數 G=C+D	標示總合名數 H=E+F
檔案一	27	21	26	1	17	3	27	20
檔案二	27	26	27	5	22	5	32	27
檔案三	30	29	25	6	26	10	31	36
	84	76	78	12	65	18	90	83

表四 ： 姓名標示正確錯誤數統計表

	姓正確率 A/G	姓召回率 A/B	名正確率 C/H	名召回率 C/B
檔案一	96.3	96.3	85.0	81.0
檔案二	84.4	100.0	81.5	84.6
檔案三	80.6	83.3	72.2	89.7
	86.7	93.3	78.3	85.5

表五 ： 姓與名的正確率與召回率

	總詞數 I	姓所佔的比例 A/I %	名所佔的比例 B/I %
檔案一	777	3.5	2.7
檔案二	1390	1.9	1.8
檔案三	1285	2.3	2.3
	3452	2.4	2.2

表六 ： 姓與名在測試資料中所佔的比例

	姓標示正確所佔的比率 C/I %	姓標示錯誤所佔的比率 D/I %	名標示正確所佔的比率 E/I %	名標示錯誤所佔的比率 F/I %	姓標示淨正確比率M= C/I-D/I%	名標示淨正確比率 N= E/I-F/I%	姓名標示淨增加比率 M+N %
檔案一	3.3	0.1	2.2	0.4	3.2	1.8	5.0
檔案二	1.9	0.1	1.6	0.4	1.8	1.2	3.0
檔案三	1.9	0.5	2.0	0.8	1.4	1.2	2.6
	2.3	0.3	1.9	0.5	2.0	1.4	3.4

表七 ： 姓名標示對測試資料正確姓的改進

四、結論

由於機率式的中文詞性標示研究還處於剛開始的階段，在這方面並沒有許多的論文可以參考，所以只能從英文在這方面的研究得到一些想法；然而中文與英文間存有很大的差異，在英文處理不錯的詞性標示方法，運用到中文卻未有相同的效果。從實驗的結果看出，利用詞性二元接續表機率式的標示模式，較基線模式改進了百分之五。雖然結果並非理想，但是這卻也告訴我們，在中文詞性標示的領域尚有許多研究的空間，值得努力改進的地方還有很多。

中文的未知詞與英文的未知詞不同。在英文方面，英文詞的分界明顯，因此不會造成詞與詞間的混淆，未知的程度僅在詞性以上的階層。中文未知詞會使得系統無法辨識這個詞，所以中文未知詞的處理比英文的還要困難。本篇對單字與雙字未知詞做了些簡單的處理，結果顯示正確率增加了 3.5% ，召回率增加了 1.7% ，使得系統的正確率與召回率接近百分之八十。

另外，我們利用一些姓名的頻率資料，加入標示系統內。結果顯示姓的召回率達到 93.3% ，名的召回率達 85.5% ，這點證明以機率模式方法來處理中文姓名是可行的，且有很好的效果。

參考資料

[1] 何文雄，1983，中文斷詞的研究，國立台灣工業研究技術學院（碩士論文）。

[2] 陳克健，陳正佳，林隆基，1986，中文語句分析的研究—斷詞語構詞，TR-86-004，Nankang : Academia Sinica (技術報告)

[3] Chen, K .J. et al. Word Identification for Mandarin Chinese Sentences , Proceedings of COLING-92 ,14th Int. Conference on Computational Linguistics , pp. 101-107, July 23-28 ,1992.

[4] C. K. Fan and W. H. Tsaai ,1987, Automatic Word Identification in Chinese Sentence by the Relaxation Technique ,Proc. of National Computer Symposium, pp. 423-431,National Taiwan University,Taipei,Taiwan.

[5] Lee, H. J. et al. Rule-Based Word Identification for Mandarin Chinese Sentences - A unification Approach . Computer Processing of Chinese and Oriental Languages . Vol 5 , no 2, pp. 97-118 , March 1991.

[6] Richard Sproat and Chilin Shih , 1990 , A Statisstical Method for Finding Word Boundaries in Chinese Text, Computer Processing of Chinese & Oriental Languages, Vol. 4 , March 1990.

[7] Jyun-Sheng Chang, Chi-Dah Chen ,and Shun-Der Chen, Chinese Word Segmenttation through Constraint Statisfication and Statistical Optimization , In Proceedings of ROC Computational Linguistics Conference, pp. 147-165, 1991.

[8] T. H. Chiang,T. S. Chang,M.Y. Lin,and K. Y. Su,1992, Statistival Models for Word Segmentation and Unknown Word Resolution , ROCLING V , pp.147-175.

[9] 陳志達，991，中文斷詞與詞性標定，國立清華大學資訊科學研究所，碩士論文。

[10] J.S. Chang, T.Y. Tseng, Y. Cheng, H.C. Chen, S.D. Cheng, S.J. Ker, and J.S. Liu, A Corpus-Based Statistical Approach to Book Indexing , Applied Natural Language Processing ,1992.

[11] 張簡哲輝，1992，馬可夫語言模式于手寫中文辨識之應用，國立交通大學資訊工程研究所，碩士論文。

[12] Greene and Rubin ,1971, Automatic Grammatical Tagging of English Technique report , Department of Linguistics , Brown University, Providence,Rhode Island.

[13] Brill ,1992, A Simple Rule-Based Part of Speech Tagger , ACL , pp .152-155.

[14] G. Leech,R. Garside, and E. Atwell ,1983 ," The Automatic Grammatical Tagging of the LOB corpus", ICAME News 7, pp .13-33.

[15] Church K. W. , 1988 , A Stochastic Parts Program and Noun Phrase Paser for Unrestricted Text , Second Conference on Applied Natural Language Processing, pp . 136-143.

[16] DeRose , S. J. , 1988 , Grammatical Category Disambiguation by Satistical Optimization, Computational Lingustics 14, 31-39.

[17] Y.C. Lin, T.H.Chiang, and K.Y. Su , Discrimination Oriented Probabilistic Tagging , R.O.C. Computational Linguistics Conference ,pp.85-96,1992.

[18] Chang, J.S. et al. A multiple-corpus Approach to Identification of Chinese Surname-Names , Journal of Computer Processing of Chinese and Oriental Languages , Feb .1993.