語言模式在中文文字辨識上的應用

周竝弘　　張俊盛

清華大學資訊科學研究所

摘 要

目前的中文字體辨識的研究，偏向以圖形辨識的方式來解決問題。在本篇論文中，我們引用語言模式來輔助字體辨識系統，使之得到更好的辨識效率。

在運用語言模式的策略方面，我們使用模組化架構來結合影像模式與語言模式，而在語言模式上我們則運用了機率式作法進行斷詞層次的分析，來幫助在影像模式所產生的候選字列中挑出正確字。爲了克服影像處理方面的辨識缺失，我們也運用雜訊通道模式來改善問題。

實驗與分析結果顯示，適當的分析評估作法可以指導我們，試驗不同語言模式技術。而採用適當的語言模式技術，才能對中文字體辨識有最大的助益。本實驗設計、效率評估與錯誤分析的技術，可以提供語音辨識等人機界面研究，引入語言模式時的參考。

一、簡介

目前在中文字體辨識的研究上有正確率不符需求的問題。比如用來掃描處理的印刷物品質稍差，或是在辨識過程中產生了一些雜訊，在在都會影響到整個系統的辨識率。我們將利用語言模式來幫助解決這個問題。

目前所有有關於中文字體辨識的研究，大部分的研究者都將這個問題看做是一般解圖形辨識的問題。一個典型的字體辨識系統的做法，在把一份印刷品掃描以後，經過影像處理過程，然後再與特徵資料庫中的資料比對，而後才決定所處理的字是什麼字。以中文的字體辨識系統而言，由於中文字是一個小型的複雜圖形，而且中文字的特徵分類數量高過於拼音文字以及特殊符號，所以中文的字體辨識系統就沒辦法做得非常精確。在一般的使用狀況中，字體辨識系統所處理的文件，通常是一串串有意義的、互相有關聯的文字所構成的句子。這些句子也就是所謂的自然語言(natural language)，可以用某種文法(grammar) 來加以規範(model)。這樣的文法通常可以利用自然語言處理技術變成計算模式。應用這樣的方式來輔助或驗證人機界面中的影像或語言處理，我們便稱之爲**語言模式**。

將語言模式應用在光學字體辨識，所得到的實驗結果大致會比原來的好。目前在辨識的研究中，有幾篇論文提到應用語言模式來增加人機界面系統的正確率[2,3,10,12]。以字體辨識來說，與只由純影像處理的系統比較，應用語言模式可以提升4% 的正確率。

二、影像模式與語言模式之整合架構

我們將兩種模式整合的方法歸類成三種不同的架構。第一種方法，我們可以將語言模式用來作字體辨識系統的**後編輯**（*post-editing architecture*）。也就是說，把字體辨識系統所做出來的結果（一串句子）當做語言模式的處理資料，然後語言模式便扮演一個語法拼字檢查及改錯器的角色，處理並輸出正確的句子。另一種所謂的**模組化處理**（*modular architecture*）是把語言模式和影像模式分成兩個模式，在影像模組處理結束後，將所有可用的資訊傳給語言模式，而語言模組便用這些資訊來產生最合乎語言現象的結果。最後是**交談式整合**（*interactive architecture*），也就是將語言模式放在影像模式的每一個處理單元中，互相傳遞資訊及資料而達成最佳的結果。

第一種想法就是把語言模式當做是一種影像模式的後編輯處理器。字體辨識系統辨識出來的結果通常都有一些獨立的或是有關聯的錯字。我們可以把這些錯字的改正當做是一個文書處理的問題來解決。比如說，假設影像模式輸出了一個像這樣的句子“將圍形輸入電腦”，則語言模式便檢查這一個句子，將它自動地改成“將圖形輸入電腦”並輸出。

然而這樣的作法在目前可行性與效果都有問題。就目前所知僅有一項初步的中文拼字檢查器的研究與發展計畫[13]。除此之外，這種架構所傳遞的資訊太少，所以在後編輯處理階段時，便無法依照前面的處理過程來作一些決定，所以整體的效能仍然取決於最後改正錯字的極限。此外，就算後編輯器將之改成合於文法的句子，但也不見得就能改成正確的句子。

另外一種應用語言模式的想法，可以由影像模組經語言模組一系列的處理來完成整個字體辨識系統的處理。但是影像模式必須把所有可能的資訊儘量傳給語言模組。

很明顯地，比起後編輯架構，語言模式在模組化處理的架構中可以得到更多的資訊，那麼語言模式就有更大的能力去增加辨識率了。但如果影像模式傳給語言模式的資料中有一些錯誤或雜訊在裡面，那麼它也無法選取正確的字。甚至還會因爲一些錯誤的邊際效應而使得錯誤更加增多。

另外一種架構就是將整個字體辨識問題以解決方法之不同來細分成一個個的小單元，以及把兩種處理模式依特性分割成一個個不同的處理單元。然後我們便在影像模式的小處理單元之間，依照所處理問題的不同，把所處理過的資料傳給符合這種問題的語言模式小處理單元，經過語言模式的小處理單元處理完後再將處理過的資料傳給影像模式的下一個處理單元。這就是交談式整合架構。

這個架構在目前的實行上仍然有困難。首先就是如何在中間作交互運作的問題。另一方面，在每一個影像模式的執行階段中要如何和語言模式的每一個小處理單元互相傳輸資料，以及在兩種模式之間的比重要加減多少，這都是要考慮的問題。除此了這些明顯的缺點之外，交談式整合架構在原則上無疑是一個比較好的處理模式。

在本篇論文中，由於字體辨識處理的影像部分方面是由工研院電通所來技術支援，對於整個影像模式的大概狀況並未涉及太多，所以在這次我們只能選擇採用模組化架構來做實驗及測試。

三、語言模式技術

語言模式在整個字元辨識系統所扮演的角色，一般而言就是利用一個句子中的前後文資訊，來幫助選擇正確的文字。一般來說，在自然語言處理方面所使用的策略，以作法而言，有用法則式和用機率統計等兩種方式；而在分析層次上，則有以字、詞、詞性、句法及語意觀點來做分析以解決問題。

法則式

在自然語言處理的研究中，針對這種問題的處理方式，比較傳統的作法就是定義一組法則來作處理。在目前中文人機界面研究中以法則式爲主要參考架構的，有李琳山的一篇有關於中文語言辨識的論文，裡面用了一些法則式的作法來作句法分析[6]。法則式的作法的優點是一般性較高，用較少的規則就能涵蓋許多的實例。而其缺點是，發展較困難，對於無限制文章的涵蓋能力較差。

機率統計式

第二種技術是在一個龐大的語料庫中統計出所有的語言現象，利用這些統計結果來衡量在某種情況下，某種語言現象會產生的機率。在目前的自然語言處理研究中，以字與字的觀點來解決問題的，有Richard Sproat的中文音轉字以及Richard Sproat與Chilin Shih的統計式斷詞法[8,9]； 以統計詞與詞之間的關係來作分析的，有陳志達的機率式斷詞[2]與Tsuyoshi Kitani的日文手寫後處理[12]；以統計詞性現象來作分析的，有張簡哲輝的手寫中文辨識[3]；AT&T Bell Lab. 李錦輝的一篇有關於自然語言語音處理的論文中，也到了用機率模式來處理語意層次的問題[11]。 所以總括來說，比起法則式的策略，機率統計式的策略顯得比較簡單處理。

在我們中文字體辨識的語言模式中，我們是傾向於使用機率統計式的策略，而在語言層次上，由於字體辨識比較不牽涉到句法和語意的解析，所以在這裡我們運用兩種不同層次的分析：用以詞爲觀點的中文斷詞，與以字跟字間的關係爲主的雙字（或三字）接續表(bigram or trigram)。

斷詞模式中有幾個主要的特性：長詞優先，考慮了詞素的自由性與束縛性，以及句子整體的最佳化。一般來說，除了整體的計算模式簡單之外，其斷詞正確率也比較高。

在經過圖文分割、特徵萃取、群集及分類等過程之後，所輸出的將是一串串的候選字序列。比如目前被掃瞄處理的句子是“檔案延伸名稱”，則在經過影像模式的分類過程後將輸出以下的格式：

檔 2388 儅 2566 棺 2581 楷 2670 枴 2750 倌 2755 棺 2770 侶 2799 楋 2871 褚 2898

寮 2472 索 2554 案 2746 素 2746 素 2832 寮 2896 累 2916 眾 2944 宏 3136 寮 3369

迁 2060 延 2088 建 2950 速 3065 連 3102 逗 3216 逛 3281 遣 3292 進 3351 遢 3386

伸 1522 仲 2351 梓 2786 坤 2808 俾 2847 砷 3029 仰 3065 悼 3160 棹 3210 伴 3318

名 1711 各 2173 乞 3141 色 3191 爸 3328 缶 3412 皂 3438 匕 3444 包 3455 易 3591

稱 2578 稍 2844 稻 3132 枴 3184 榴 3357 櫚 3362 稀 3362 梢 3526 檜 3628 櫚 3685

其中每一行就代表一串候選字序列，而每個候選字之後便是代表此字爲正確字之不可能性的指標。當語言模式得到這些序列時，中文斷詞系統將會先看前兩個候選字序列（在此處便是“檔”行和“寮”行），先在包含詞頻的辭典中找到各個字的頻率，再找“檔寮”、“檔索”、“檔案”、“檔．．”、“儅寮”、“儅索”．．．直到每一個配對都找完爲止。然後把在辭典中有找到的配對及頻率，以及原先每個字的頻率都儲存起來。接下來每一個階段就都只再加入下一個候選字序列來作判斷。以本例而言，

下一個階段便是把“廷”行加進來，除了找出“廷”行每一個字的頻率之外，也找出所有“寮”行和“廷”行的配對中有在辭典的詞頻，並找目前這三個候選字列的所有配對情形（如“檔寮廷”）。如此一個階段一個階段地進行，而我們所得到的是每一個候選字的頻率及在辭典中有成詞的那些詞及詞頻。接著就把所有可能的斷詞結果列出來：

| 檔* | 寮* | 廷* | 伸* | 名* | 稱* |

| 檔案 | 廷* | 伸* | 名* | 稱* |

| 檔* | 寮* | 延伸 | 名* | 稱* |

| 檔* | 寮* | 廷* | 伸* | 名 | 稱 |

| 檔案 | 延伸 | 名* | 稱* |

| 檔案 | 廷* | 伸* | 名稱 |

| 檔* | 寮* | 延伸 | 名稱 |

| 檔案 | 延伸 | 名稱 |

（檔*可以是”檔”行中的所有候選字）

再依照我們前面所說的中文斷詞整體頻率的判定方法把這些斷詞結果賦以斷詞頻率並加上此結果中每個字的錯誤計數，再比較這些數字的大小就可以挑選出我們所要的結果了。在此例中的斷詞結果，經過比較之後 |檔案|延伸|名稱| 是最可能的斷詞結果，則“檔案延伸名稱”便是最後的輸出結果。

四、實驗結果及討論

這次的實驗前半部的影像模式是由工研院電通所所提供的OCR中文字體辨識系統。電通所的這套OCR系統是以5401 個常用字與一般鍵盤上的所有字母及符號爲辨識目標。而整個語言模式系統是在IBM Compatible 486-33 EISA PC 下以Microsoft C 6.0程式語言發展而成的。而詞彙統計資料則使用致遠科技的BDC電子辭典。此辭典包含了約九萬個中文詞，以及在約一百萬字的語料中測試所有詞的出現率，並依照這些詞的出現率把它們分成5個等級。在測試資料方面，我們用了幾個從致遠科技取得的Lotus中文手册及IBM OS-400 中文手册節錄下來的文字片段，除去非中文字資料後，以Microsoft Windows 3.0 中文版中CFShouSong 10~12 pt.的字型輸出至雷射印表機，而列印結果則當做OCR系統的輸入資料。

實驗一

1.首先將測試資料輸入電通所之OCR系統中，並產生一連串的候選字列。

2.其次將步驟1所得的串式輸入機率式斷詞處理器中。處理完後則將所選的字輸出。

3.計算輸出結果的精確率及召回率並與原影像模式的結果作比較。

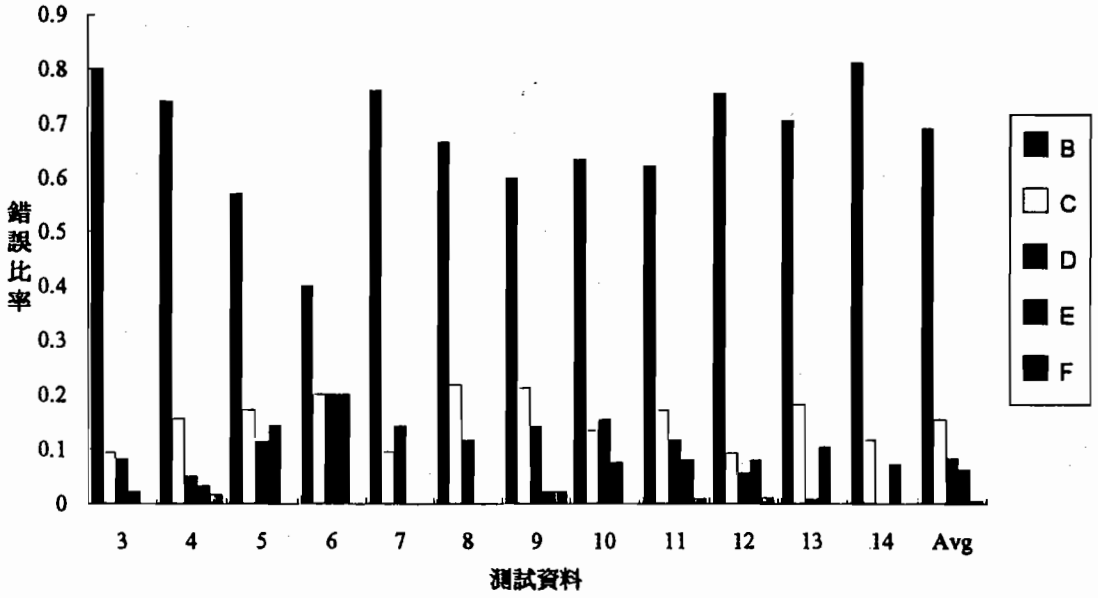
實驗結果

測試資料	總字數 A	影像模式錯誤部份字數B	影像模式正確率 (A-B)/A	語言模式更動字數 C	語言模式更動字中爲B的字數 D	語言模式正確改錯字數 E	語言模式檢錯精確率 D/C	語言模式檢錯召回率 D/B
test3*	295	100	66.10%	78	63	29	87.76%	63.00%
test4	349	68	80.52%	53	42	20	79.25%	61.76%
test5	195	27	86.15%	27	13	6	48.15%	48.15%
test6	188	16	91.49%	17	11	7	64.71%	68.75%
test7	112	18	83.93%	14	9	2	64.29%	50.00%
test8	355	57	83.94%	44	21	9	47.73%	36.84%
test9	283	72	74.56%	77	60	15	77.92%	83.33%
test10	301	49	83.72%	50	33	10	66.00%	67.35%
test11	530	89	83.21%	91	52	16	57.14%	58.43%
test12	451	108	76.05%	91	72	39	79.12%	66.67%
test13	424	115	72.88%	105	82	34	78.10%	71.30%
test14	271	70	74.17%	55	45	12	81.81%	64.29%
總平均	3754	789	78.98%	702	503	199	71.65%	63.75%

測試資料	語言模式修正精確率 E/C	語言模式修正召回率 E/B	語言模式正確字數 F	語言模式整體精確率 F/A	效能增加率 (E-C+D)/A
test3*	37.18%	29.00%	210	71.19%	4.75%
test4	37.74%	29.41%	291	83.38%	2.58%
test5	22.22%	22.22%	160	82.05%	-4.10%
test6	41.18%	43.75%	173	92.02%	0.53%
test7	14.29%	11.11%	91	81.25%	-2.68%
test8	20.45%	15.79%	286	80.56%	-3.94%
test9	19.48%	20.83%	198	69.96%	-0.71%
test10	20.00%	20.41%	249	82.72%	-2.33%
test11	17.58%	17.98%	419	79.06%	-4.34%
test12	42.86%	36.11%	365	80.93%	4.43%
test13	32.38%	29.57%	319	75.24%	2.59%
test14	21.82%	17.14%	202	74.54%	0.74%
總平均	28.35%	25.22%	2963	78.93%	0.00%

所有的實驗結果除了展示整個個程的辨識率之外，並且計算能夠代表所加入語言模式的系統效能的召回率。除此之外，我們在此列出兩種效率指數－檢錯率及修正率。檢錯精確率爲語言模式所更動的字數中原在影像模式部份就錯掉的字的比率，而修正精確率便是語言模式所更正的字數中原在影像模式部份就錯掉的字的

比率；檢錯召回率與修正召回率則是在原影像部份的錯字中，被語言模式偵測到以及改正的比率。精確率可以說明整個系統的效率，而召回率則可以說明語言模式本身的能力。



B：正確字不在候選字列中所產生的錯誤　　C：由於B產生的邊際效應

D：因單字詞的錯誤　　E：因正確詞之頻率較低而選錯詞的錯誤

F：正確詞不在辭典中所產生的錯誤

這個實驗結果指出最嚴重的問題在於候選字列不包含正確字的狀況，下面我們就針對正確字不在候選字列的情形作改進。

實驗二

首先我們將整個OCR系統的處理 過程以下面的圖來表示：

掃瞄資料 -> 影像模式 -> 候選字列 -> 語言模式 -> 結果輸出

而在候選字列的地方出了錯。我們可以想像是測試資料在經過影像模式處理的時候，影像模式產生了一些雜訊而導致一些錯誤的輸出，就像是一串資料經過同軸線傳輸時，由於傳輸線的不穩定

而產生雜訊，致使接收方得到的結果是有錯誤的資料，這就叫做雜訊通道。若是假設雜訊的產生是與原來的資料以及傳輸線的性質有關，一般解決這種問題的方法，就是先統計出怎樣的輸出才有可能是受雜訊影響，以及什麼樣的原資料受到雜訊影響後會變成怎樣的錯誤資料，然後當偵測到有錯誤時，就根據前面的統計資料來判定正確的資料該是如何。

而在自然語言應用的範疇中，也有類似的處理方式[7]。首先我們將5401個字拿進原影像模式中作處理，把得到的候選字列與原測試資料比較，則在同一個位置中，那些不同於原來的字的候選字便是屬於雜訊的範圍，而附屬的錯誤計數便可以代表原來某一個字經過影像模式的處理產生這個雜訊的機率指標，由此我們可以建立一個5401x5401的混淆矩陣(confusion matrix)，而其中每一個記錄便是某一個字可能會認錯成另一個字的指標。

假設一個候選字列內的所有字爲$C_1C_2...C_n$，則我們要求的是對5401個字中每一個字W_i，$P(W_i|C_1C_2...C_n)$爲多少，進而決定哪一個字比較有可能是經雜訊干擾前的正確字，並將之加入候選字列中讓語言模式來挑選。根據貝氏法則(Baye's rule)，

$$P(W_i|C_1C_2...C_n)=\frac{P(W_i\cap C_1C_2...C_n)}{P(C_1C_2...C_n)}=\frac{P(W_i)P(C_1C_2...C_n|W_i)}{P(C_1C_2...C_n)}$$

爲了能做出簡單的證明與實驗，我們做了一些假設及簡化，以便能馬上作分析。首先我們假設每個字在我們的測試環境中絕對會出現，且只針對前五個錯誤計數較小的候選字作由雜訊回復到正確訊號的動作，亦即

$$P(W_i|C_1C_2C_3C_4C_5)=\prod_{j=1}^{5}P(C_j|W_i)$$

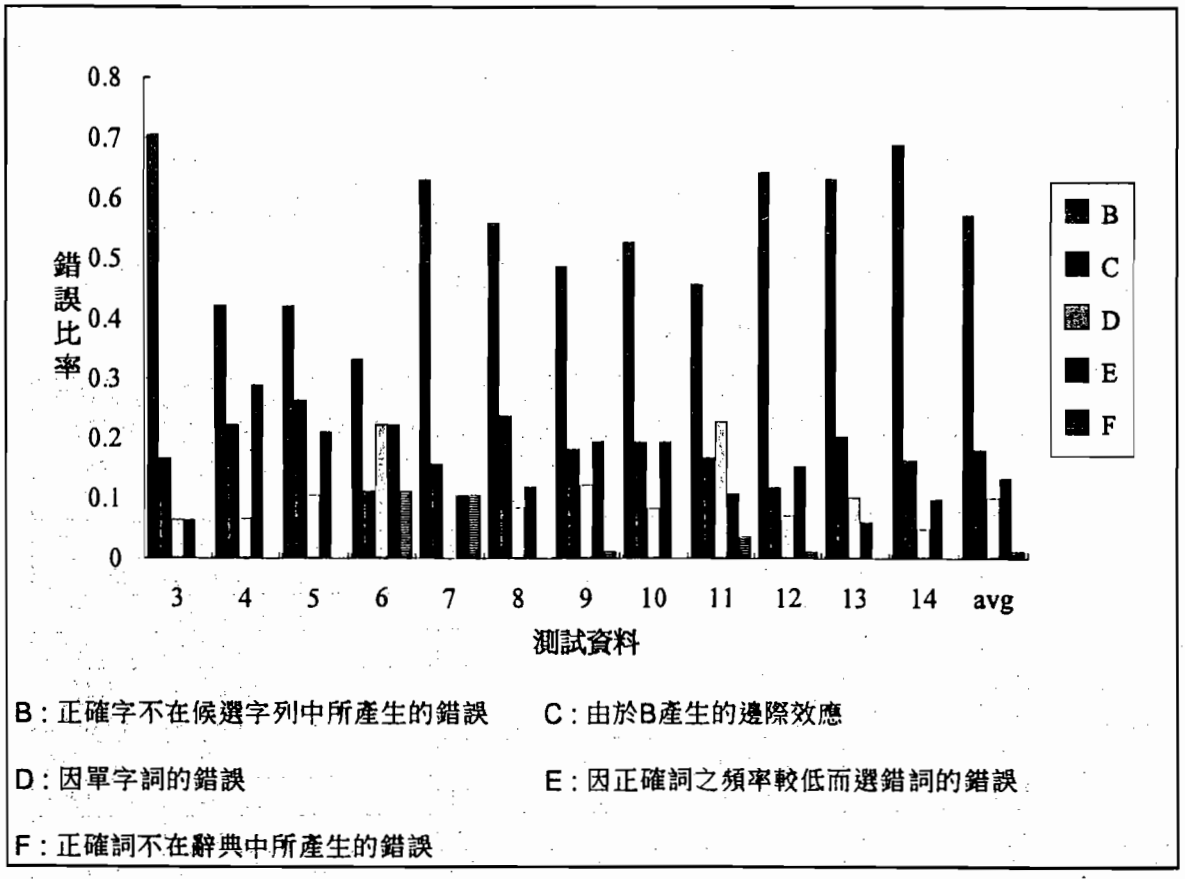
然後至此我們便可以由混淆矩陣中找出這五個字所可能出錯的字，並比較這些錯誤出現率來決定哪一個字要被加入候選字列中。將這些字加入候選字列後，我們再重複前一節所做的實驗。

這次的實驗中，在混淆矩陣方面，我們是將5401個常用字以Microsoft Windows 3.0 中文版的 CFShouSong 12 pt. 字型，經過影像模式處理之後根據結果建立起來的。

測試資料	總字數 A	影像模式錯誤字數 B	影像模式正確率 (A-B)/A	語言模式更動字數 C	語言模式更動字中爲B的字數 D	語言模式正確改錯字數 E	語言模式檢錯精確率 D/C	語言模式檢錯召回率 D/B
test3	295	100	66.10%	96	79	37	82.29%	79.00%
test4	349	68	80.52%	75	58	38	77.33%	85.29%
test5*	195	27	86.15%	33	25	19	75.76%	92.59%
test6*	188	16	91.49%	19	13	13	68.42%	81.25%
test7*	112	18	83.93%	15	10	3	66.67%	55.56%
test8*	355	57	83.94%	55	31	17	56.36%	54.34%
test9	283	72	74.56%	81	45	22	55.56%	62.50%
test10*	301	49	83.72%	54	43	28	79.63%	87.76%
test11*	530	89	83.21%	108	67	44	62.04%	75.28%
test12	451	108	76.05%	105	81	41	77.14%	75.00%
test13	424	115	72.88%	116	90	41	77.59%	78.26%
test14	271	70	74.17%	69	54	24	78.26%	77.14%
總平均	3754	789	78.98%	826	596	327	72.15%	75.54%

測試資料	語言模式修正精確率 E/C	語言模式修正召回率 E/B	語言模式正確字數 F	語言模式整體精確率 F/A	效能增加率 (E-C+D)/A
test3	38.54%	37.00%	217	73.56%	6.78%
test4	50.67%	55.88%	304	87.11%	6.02%
test5*	57.58%	70.37%	176	90.26%	5.64%
test6*	68.42%	81.25%	179	95.21%	10.11%
test7*	20.00%	16.67%	93	89.29%	-1.79%
test8*	30.90%	29.82%	296	83.38%	-1.97%
test9	27.16%	30.56%	201	71.02%	-4.95%
test10*	51.85%	57.14%	265	88.04%	5.65%
test11*	40.74%	49.44%	447	84.34%	0.57%
test12	39.05%	37.96%	367	81.37%	3.77%
test13	35.34%	35.65%	326	76.89%	3.54%
test14	34.78%	34.29%	210	77.49%	3.32%
總平均	39.59%	41.44%	3081	82.07%	2.58%

*測試資料字體爲 CFShouSong 12 pt. 字型



B：正確字不在候選字列中所產生的錯誤　　C：由於B產生的邊際效應

D：因單字詞的錯誤　　E：因正確詞之頻率較低而選錯詞的錯誤

F：正確詞不在辭典中所產生的錯誤

討論

與初步的實驗結果比較，我們可以發現正確字不在候選字列的狀況平均少了11.94%。在相當強的假設與簡化之下，正確率有改善的情形，證明雜訊通道理論的可行性與有效性。

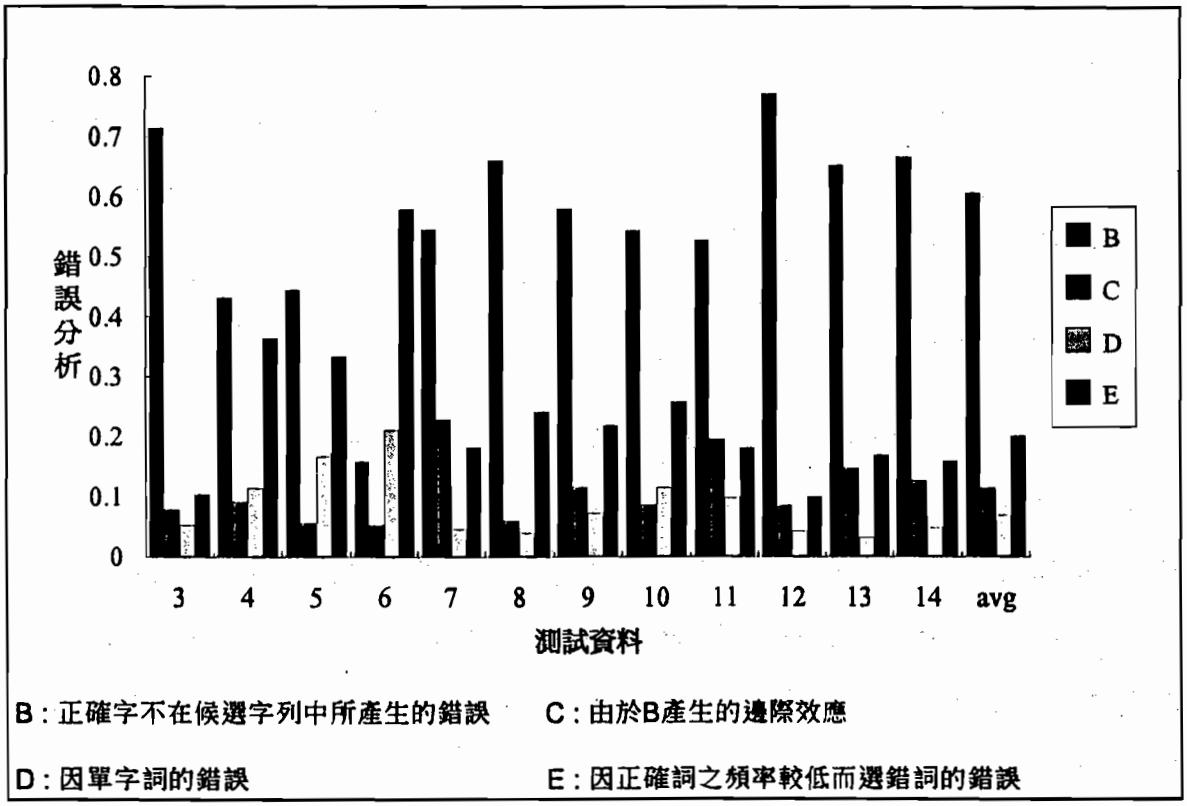
我們可以看到整體的精確率與修正率並沒有提高到一定的水準，顯示仍然有其他的錯誤尚未解決。由錯誤分析表上發現，除了正確字不在候選字列中的錯誤減少之外，其他的錯誤比率都升高了， 尤其是被選者頻率比較高的部分錯誤比率升高了7.05%，實在是造成這次精確率不高的主因。造成這些錯誤比率升高的原因，是我們加入了太多的候選字。但是這種情形仍然有可能可以解決。解決單字詞和錯誤字／詞頻率太高的狀況最有效的方法是應用雙字接續表的作法。

實驗三

在本次的實驗中，我們所加入的是從致遠科技的Lotus中文手册及IBM OS 中文手册所訓練出來的雙字接續表，表內紀錄任兩個字在語料庫中同時出現的機率。

測試資料	總字測試字數 A	影像模式錯誤字數 B	影像模式正確率 (A-B)/A	語言模式更動字數 C	語言模式更動字中爲B的字數 D	語言模式正確改錯字數 E	語言模式檢錯精確率 D/C	語言模式檢錯召回率 D/B
test3	293	100	65.87%	79	62	35	78.48%	62.00%
test4	348	68	80.50%	81	60	47	74.07%	88.24%
test5	191	27	85.86%	30	20	19	66.67%	74.07%
test6	185	16	91.35%	30	14	13	46.67%	87.50%
test7	105	16	84.76%	13	7	4	53.85%	43.25%
test8	354	56	84.18%	44	28	23	63.64%	50.00%
test9	279	72	73.48%	63	38	24	60.32%	52.78%
test10	300	49	83.67%	47	34	27	72.34%	69.39%
test11	529	89	84.61%	88	60	44	68.82%	67.42%
test12	448	108	75.89%	88	76	50	86.36%	70.37%
test13	423	115	72.81%	95	68	38	71.58%	59.13%
test14	269	70	73.98%	55	38	24	69.09%	54.29%
總平均	3724	786	78.89%	713	505	348	70.83%	64.25%

測試資料	語言模式修正精確率 E/C	語言模式修正召回率 E/B	語言模式正確字數 F	語言模式整體精確率 F/A	效能增加率 (E-C+D)/A
test3	44.30%	35.00%	216	73.72%	6.14%
test4	58.02%	69.12%	304	87.36%	7.47%
test5	63.33%	70.37%	173	90.58%	4.71%
test6	43.33%	81.25%	166	89.73%	-1.62%
test7	30.77%	25.00%	87	82.86%	-1.25%
test8	52.27%	41.07%	304	85.88%	1.98%
test9	38.10%	33.33%	210	75.27%	-0.36%
test10	57.45%	55.10%	265	88.33%	4.67%
test11	50.00%	49.44%	457	86.39%	3.02%
test12	56.82%	46.30%	378	84.38%	8.48%
test13	40.00%	33.04%	326	77.07%	2.60%
test14	43.64%	34.29%	206	76.58%	2.60%
總平均	48.81%	44.27%	3092	83.03%	3.76%



B：正確字不在候選字列中所產生的錯誤　　C：由於B產生的邊際效應

D：因單字詞的錯誤　　E：因正確詞之頻率較低而選錯詞的錯誤

與先前的結果比較，我們可以看到，雖然語言模式的更動字數比原來的少，但是所改正的字數卻比原來的高，而且正如我們所預期的，由單字詞所產生的錯誤變少了，而且連帶地由正確字不在候選字列所產生的邊界效應也降低了，這證明雙字接續表在字體辨識上有一定的效能。

而現在我們要面對的問題，首先就是正確字不在候選字列的問題，在這次的實驗中，這種情形仍然相當嚴重，而且由於這種錯誤，導致語言模式檢錯和修正的效果降低，我們必須在雜訊通道的演算法上再多加斟酌；此外，這次的實驗也顯示出雙字接續表仍無法有效地控制高頻錯誤詞（或字對）的出現。關於高頻錯誤詞的問題，除了可以經過方法上的協調來減低錯誤外，或許將來可以在語言模式的解析層次上再提升（如以詞性、語法或語意的觀點作解析等）來設法解決此問題。

六、結論

從這一整個的研究與實驗，我們發現簡單的語言模式應用在中文字體辨識有幾種架構和策略可以使用，但是並不是每種架構與策略都可以有效果。我們這一次使用了模組化處理的架構，而在上一個模組將一切的資訊傳給下一個模組時，同時也將所得的雜訊與錯誤傳下去，而後一個模組也因爲這些錯誤而產生了更多的錯誤。

爲了改善這種錯誤，我們應用了雜訊通道模式來減少這種錯誤，而也正如我們原先的想法，它把正確字不在候選字列的情形減少到一定的程度，這算是在語言模式之外，帶給影像模式的一

種改善策略。但是在展現它的優點之餘暴露了它加了更多雜訊的缺點。所幸這種模式帶給我們的優點仍然大於缺點。

而在語言模式的策略上，我們一開始使用以詞彙機率爲主的斷詞技術，但是在應用了雜訊通道之後，所附加上的不明確度凸顯出了斷詞技術在單字詞及低頻詞上功能的不足。所以我們發現，並非每一種語言模式的策略都可以造成良好的效果。故爲了要得到更好的效果，必須在策略的選擇上花上一番功夫。

下一個步驟所應繼續探討的研究方向，首先雜訊通道作法的加強，使得所加入的正確字比率最高，而所加入的額外雜訊可以減到最低。此外，減少高頻錯誤詞的出現頻率，也是一個重要的工作。利用資料庫處理技術來自動擷取容易混淆詞彙的搭配資訊，並利用這項資訊於語言分析中，可望有效地克服這一類的錯誤。

參考資料

[1] Chang, J. S. and Chou, B. H. (1991). "An Interactive Language Model for Chinese OCR." In *Proceedings of Natural Language Processing Pacific Rim Symposium*, Singapore, 139-141.

[2] Chang, J. S.; Chen, C. D. and Chang, S. D. (1991). "Chinese Word Segmentation through Constraint Satisfaction and Statistical Optimization." In *Proceedings of ROC Computational Linguistics Conference*, Kenting, Taiwan.

[3] Chang Chien, C. H. and Lee, H. J. (1992). "A Marcov Language Model in Handwritten Chinese Chinese Text Recognition Application." Masteral thesis of Institute of Computer Science and Information Engineering, NCTU, Taiwan.

[4] Fan, C. K. and Tsai, W. H. (1987). "Automatic word identification in Chinese Sentence by the Relaxation Technique." In *Proceedings of National Computer Symposium*, National Taiwan University, Taipei, Taiwan, 423-431.

[5] Dechter, R. and Pearl, J. (1988). "Network-Based Heuristics for Constraint-Satisfaction Problems." *Artificial Intelligence* **34**: 1-37.

[6] Lee, L. S.; Chien, L. F.; Lin, L. J.; Huang, J. and Chen, K.-J. (1991). "An Efficient Natural Language Processing System Specially Designed for Chinese Language." *Computational Linguistics* **17** (4): 347-374.

[7] Kernighan, M. D.; Church, K. W. and Gale, W. A. (1990). "A Spelling Correction Program Based on a Noisy Channel Model." In *Proceedings of Computational Linguistics 1990.*

[8] Sproat, Richard (1990). "An Application of Statistical Optimization with Dynamic Programming to Phonemic-input-to-character conversion for Chinese." In *Proceedings of ROCLING III*, 377-390.

[9] Sproat, Richard and Shih, Chilin (1990). "A Statistical Method for Finding Word Boundaries in Chinese Text." *Computer Processing of Chinese & Oriental Languages* **4**, March 1990.

[10] Pieraccini, R. and Lee, C. H. (1991). "Factorization of Language Constraints in Speech Recognition." In *Proceedings of Annual ACL Meeting*, 299-306.

[11] Pieraccini, R.; Levin, E. and Lee, C. H. "Stochastic Representation of Conceptual Structure in the ATIS Task." Manuscript.

[12] Kitani, Tsuyoshi (1991). "An OCR Post-Processing Mathod for Handwritten Japanese Documents." In *Proceedings of Natural Language Processing Pacific Rim Symposium*, Singapore, 38-45.

[13] 施得勝、王良志、陳志達、聶素芬(1992)，“基於統計的中文錯字偵測法”技術草稿。

[14] 工研院電通所技術文件“印刷體中文字辨識系統”，文件編號11180-0025～11180-0232，技術資料編號DW-1041～DW-1048。

附 錄 ： 測 試 資 料 （ 部 份 ）

test3

若要使用
資料
假設分析表格
雙變數
在計算雙變數表格之前
必須先設定表格範圍
設定方式是在工作表中輸入公式和輸入值
若要設定雙變數表格
請決定表格範圍所在位置
決定兩個輸入儲存格的位置
必須在表格範圍之外
然後分別加上標籤便於尋找
在表格範圍左上角的儲存格內輸入公式
這個公式必須用到兩個輸入儲存格
從表格範圍的第一欄
公式下面個儲存格開始
輸入和輸入儲存格
有關的值
在公式右邊的儲存格中
輸入和輸入儲存格
有關的值
若要計算雙變數表格
請選取
資料
假設分析表格
雙變數
在
表格範圍
文字方塊內指定表格範圍
即包含所有公式和輸入值的範圍如
請只含括表格範圍內的公式和輸入值
不要將輸入儲存格包含在內
在
輸入儲存格
文字方塊內指定輸入儲存格
如

輸入儲存格

加强圖表的外觀
本章將說明如何在圖表中加入文字和其他物件
如排在圖表中選定或重新安排物件的位置
如何改變圖表的字型及色彩
以及如何調整圖表的大小
爲什麼要加强圖表的外觀
雖然圖表上的資料點已經精確地反映出了工作表上的數值
但是您可能還想再加强這份資料的視覺效果
使它更生動一些
您可以利用文字說明
特殊字型
色彩
圖形
線條
線條樣式和箭頭等等
加强圖表的效果
建立
中文版的圖表
以及把它加入工作表的詳細資料
請參閱第
章
下圖顯示條是冰淇淋銷售資料條長條圖
本章將利用這個長條圖來展示加入文字和其他物件
以及改變字型
色彩和線條樣式後的效果
在圖表中加入文字
若要在圖表中加入文字
請先將這個圖表顯示在圖表視窗中
當圖表視窗成爲使用中視窗之後
您就可以利用
繪圖
指令
將一些物件
文字
幾何圖表
箭頭和手繪圖表

加入工作表的圖表中

以下將說明如何在圖表的任一處加入說明文字

測 試 結 果 （ 影 像 模 式 部 份 ）

其中( )表辨識錯誤的字

test3

若(要)使用
資料
(值)設分(析)(衮)格
(貸)(蟹)(拽)
在計算(貸)(置)(拽)(衮)格之前
必須先設(巨)(衮)格範(目)
設定(右)(戎)(昱)在工作(衮)中(翰)人(幼)(戎)和(輪)人值
若(更)設定(貸)(蟹)(琅)(衮)格
(誥)(泱)定(衮)格範(圉)所在位置
(泱)定兩個輸人(倔)存格的位置
必須在(衮)格範(圉)之外
(絞)後分別加上(瘡)(鈣)便(始)尋(蛞)
在(衮)格範(目)左上角的儲存格內輸人(幼)(戎)
這個公(戎)必(濱)用到兩個(翰)人儲存格
(往)(衮)格範(目)的第一(捆)
(幼)(戎)(花)面個儲存格開始
輸人和(軔)人儲存格
(右)(闌)的值
在(幼)(戎)右(適)的儲存格中
(翰)人和(輓)人儲存格
有(閏)的值
若要計算(貸)(鏈)(拽)(衮)格
(誥)(遘)(洱)
資(榭)
(侶)設分(析)(衮)格
(貸)(蟹)(勃)
在
(衮)格範(圉)
文字(右)(兔)內指定(衮)格範(圄)
(朋)包含所(石)公(戎)和(翰)人值的範圍
(誥)只含一一(衮)格範圍內的(幼)(戎)和輸人值
不(要)(姥)(輪)人儲存格包含在內
在
(翰)人儲存格
文字(右)(兔)內指定(翰)人儲存格
(翰)人(倔)存格

加強(圃)(裘)的外(硯)
本章將說明如何在(固)(裘)中加入文字和其他物件
如排在(固)(裘)中(遭)定或重新安排物件的位置
如何改(蠻)(固)(裘)的字型及色彩
以及如何調整(圃)(裘)的大小
為什麼要加強(固)(裘)的外(硯)
雖然(固)(裘)上的資料(鍥)已經(猜)確地反映出(束)工作(袁)上的數值
但是您可能(遭)想再加強(遏)份資料的視覺效果
使它(巨)生動一些
您可以利用文字說明
特殊(括)型
色彩
圖形
線條(檬)式和箭頭等等
加強(固)(裘)的效果
中文版的(固)(裘)
以及把(售)加入工作(裘)的詳細資料
(話)參(閏)第
章
(牢)(囤)顯示條是冰淇淋銷售資料條一(良)條圖
本章將利用(遣)個長條(圃)來展示加入文字和其他物件
以及改(肄)(汗)型
色彩和線條(檬)式後的效果
在(固)(裘)中加入文(克)
若要在(固)(裘)中加入文字
請先將(遏)個(圄)(袁)顯示在(固)表視窗中
當(固)(裘)視窗成為使用中視窗之後
您就可以利用
(給)(固)
指令
將一些物件
文(克)
幾何(固)(裘)
箭(預)和(有)(紛)(固)表
加入工作(袁)的圖(裘)中

以(片)將說明如何在(固)(袁)的任一處加入說明文字

測試結果（語言模式部份）

其中( )表正確字不在候選字中的字，〔 〕表經過語言模式後被更改的字，〈〉表正確字在候選字列中但仍然選錯的字。

test3

若〔要〕使用
資料
〔假〕設〔〈公〉〕〔(所)〕〔表〕格
〔〈控〉〕〔(置)〕(拽)
在計算〈貨〉(置)(拽)〔表〕格之前
必須先設(巨)(衾)格範〔圍〕
設定〔(允)〕(戎)(昱)在工作〔表〕中〔輸〕人(幼)〔(文)〕和〔輸〕人值
若〔(置)〕設定〔雙〕(螢)(琅)〔表〕格
(誥)〔決〕定〔表〕格範〔圍〕所在位置
〔決〕定兩個輸人〔儲〕存格的位置
必須在(衾)格範〔圍〕之外
〔然〕後分別加上(瘡)(鈣)便(始)尋〔找〕
在〔表〕格範〔圍〕左上〔〈色〉〕的儲存格〔〈自〉〕〔〈強〉〕〔〈大〉〕(幼)(戎)
這個公〔(文)〕〔〈公〉〕〔(演)〕〔〈周〉〕到〔〈了〉〕個〔輸〕人儲存格
(往)(衾)格範〔圍〕的第一〔(個)〕
〔(外)〕〔(文)〕〔(文)〕〔〈區〉〕個儲存格開始
輸人和〔輸〕人儲存格
(右)(闌)的值
在(幼)〔(文)〕〔〈在〉〕(適)的儲存格中
〔輸〕人和〔輸〕人儲存格
有(閱)的值
若要計算〈貨〉(鏈)〔(下)〕〔表〕格
(誥)(遴)(洱)
資(榭)
〔假〕設分〔(佈)〕〔表〕格
〔〈控〉〕〔(置)〕(勃)
〔〈毛〉〕
〔表〕格範〔圍〕
文字(右)〔(他)〕〔〈由〉〕指定〔表〕格範〔圍〕
〈朋〉包含〔〈了〉〕〔(在)〕公〔(文)〕和〔輸〕人值的範圍
(誥)只〔〈會〉〕一一〔表〕格範圍內的(幼)(戎)和輸人〔〈任〉〕
不〔要〕(姥)〔輸〕人儲存格包含在內
〔〈毛〉〕
〔輸〕人儲存格
〔〈去〉〕〔〈年〉〕〔(九)〕〔(他)〕〔〈由〉〕指定〔輸〕人儲存格
〔輸〕人〔儲〕存格

加強〔圖〕〔表〕的外〔觀〕
本章〔〈冊〉〕說明如何在〔圖〕〔表〕中加入文字和其他〔〈格〉〕件
如排在〔圖〕〔表〕中〔選〕定或重新安排〔〈格〉〕〔〈作〉〕的位置
如何改〔變〕〔圖〕〔表〕的字型及〔〈宜〉〕彩
以及如何調整〔圖〕〔表〕的大小
為什麼要加強〔圖〕〔表〕的外〔觀〕
〔〈邏〉〕〔〈總〉〕〔圖〕〔表〕上的資料〔(鐵)〕已經〔(鬍)〕確地〔〈良〉〕〔〈喚〉〕出〔(直)〕工作〔表〕上的數值
但是您可能〔還〕想再加強〔這〕份資料的視覺效果
使它〔(正)〕〔〈查〉〕動一些
您可以利用文字說明
特殊〔(寨)〕型
色彩
圖形
線條(檬)式和箭頭等等
加強〔圖〕〔表〕的效果
中文版的〔圖〕〔表〕
以及〔〈弛〉〕(售)加入工作〔表〕的詳細資料
(話)參〔閱〕第
〔〈事〉〕
〔(草)〕〔圖〕顯示〔〈像〉〕是冰淇淋銷售資料〔〈係〉〕一〔(員)〕條圖
本章將利用〔(道)〕個長條〔圖〕來展示加入文字和其他〔〈格〉〕〔〈作〉〕
以及〔〈玖〉〕(肄)(汗)型
〔〈宜〉〕〔〈形〉〕和線條(檬)〔〈其〉〕後的效果
在〔圖〕〔表〕中加入〔〈大〉〕〔(支)〕
若要在〔圖〕〔表〕中加入文字
請先將〔這〕個〔圖〕〔表〕顯示在〔圖〕表視窗中
當〔圖〕〔表〕視窗成為使用中視窗〔〈左〉〕後
您就可以利用
〔繪〕〔圖〕
指令
將一些物件
文(克)
幾何〔圖〕〔表〕
箭(預)和(有)〔繪〕〔圖〕表
加入工作〔表〕的圖〔表〕中

以〔(只)〕〔〈冊〉〕說明如何在〔圖〕〔表〕的任一處加入說明文字

混淆矩陣列表（部份）

字列中每個字後的數字表會被認錯爲此字的機率（0最可能，1最不可能，此處只取0~0.3之間的字）。

一,,衰 0.00 八 0.00 哀 0.00
七,,冷 0.00
九,,力 0.25 丸 0.00
二,,三 0.00
人,,入 0.00
刁,,代 0.00
又,,丈 0.06
么,,合 0.00 侖 0.00
亡,,盲 0.00
叉,,丈 0.17
士,,立 0.10
尢,,力 0.00
巳,,呂 0.00
丑,,共 0.02 早 0.00
丰,,字 0.12
五,,共 0.22
分,,公 0.00
切,,坷 0.10
反,,艮 0.20
壬,,乎 0.00 平 0.08 老 0.00
天,,平 0.26
巴,,呂 0.14
幻,,劫 0.00
弔,,罕 0.11
戈,,丈 0.25
斗,,乎 0.15
日,,曰 0.00
曰,,淄 0.00
月,,乃 0.00
木,,俏 0.12
毋,,㚔 0.11
父,,歹 0.26
犬,,丈 0.00
王,,平 0.11
且,,共 0.18
丘,,共 0.00 乓 0.00
付,,忖 0.00
代,,吉 0.00
兄,,兇 0.00