# 應用詞彙、語法與語料規則於中文手寫句辨識之校正模組

# Revision for Recognizing Chinese Handwritten Sentences Based on Lexical, Syntactical and Corpus Rules

張道行　Tao-Hsing Chang
周嘉彬　Chia-Bin Chou
蘇守彥　Shou-Yen Su
國立高雄應用科技大學 資訊工程系
Department of Computer Science and Information Engineering
National Kaohsiung University of Applied Sciences
changth@cc.kuas.edu.tw
papperkut@hotmail.com
shouyen@gmail.com

劉建良　Chien-Liang Liu
國立臺灣師範大學 資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
clliu@mail.nctu.edu.tw

## 摘要

離線手寫中文文字辨識有使用者書寫字跡的變異和文字書寫字體不明顯等問題，造成辨識系統難以辨識其特徵而影響正確性。本論文的研究目的是利用特定領域主題語料中呈現的詞彙、語法及語料規則提高離線手寫中文文字辨識率。本文提出了一個三階段方法來達成目標。首先、利用詞彙優先概念，從候選字中挑選語料庫中的詞彙為辨識結果。第二、查看候選字中是否出現特定的文法組合，並以該組合的候選文字為辨識結果。第三、將剩下相鄰兩個未決定的候選字集組成字串，並和事先由語料庫所產生收錄的雙字組比對，若候選字中存在雙字組則以做為辨識結果。實驗結果顯示本文所提方法可有效的提高辨識率，由單一字頻決定法的 0.45 提升至 0.85。

## Abstract

Recognition of off-line handwritten Chinese character had been an important problem. Because of the variation and vagueness derived from different users' handwritings, it was hard to recognize handwriting characters via statistical features obtained from database. The purpose of this study is to use lexical, syntactical and corpus rules for increasing the accuracy mentioned above. Our methods could be divided into three phases. First, we used lexical rule "multi-syllable words priority" to predict some characters of a sentence from candidate characters. Second, neighbor several candidate characters in which particular grammar patterns appear will be treated as the characters of the sentence. Finally, two adjacent

candidate characters will be regarded as a string. The strings which occur in a corpus frequently will be used to be the characters of the sentence. To contrast approach “highest frequency priority”, experimental results shown that the accurate rate of Chinese handwriting character recognition could be effectively increased from 0.45 to 0.85.

關鍵詞：中文手寫辨識，離線辨識，詞彙規則，語法規則，語料規則

Keywords: Handwritten Chinese character recognition, offline recognition, lexical rules, syntactic rules, corpus rules. .

## 一、緒論

由於資訊科技發展，許多文件都有數位檔案的版本供資訊系統擷取使用。而如何將傳統紙張文件回溯建立可辨識內容的數位文件檔案，成為重要的問題，因為若以人工手動輸入方式將資料輸入建檔，將耗費龐大的人力和時間成本。因此許多研究提出自動轉換傳統紙張文件為數位資料的方法，其中文字辨識(Character Recognition)是自動辨識文件內容的解決方法之一。文字辨識技術主要可分成兩類：一為即時文字辨識(On Line Character Recognition , OLCR)，二為光學文字辨識(Optical Character Recognition , OCR)。即時文字辨識採用手寫過程的時間序列特性，主要用於即時的手寫輸入文字辨識。光學文字辨識則是一種離線(Off-line)文字辨識的方式，適合已存在之文件的後續數位化處理。

光學文字辨識的處理流程大略分成五個程序[8]：首先，文字文件影像輸入，將文字文件以影像檔案方式儲存。第二，影像前處理。將影像檔案做二值化、細線化、文字切割、正規化等，此階段會產生單一字的字元影像，並利用一些方法消除同字的外形變異以及影像失真所造成辨識錯誤。第三，特徵擷取。擷取影像特徵並記錄做為辨識的依據。第四，分類器與辨識模型。先經由分類器學習將已知字元影像正確分類至所屬文字，再透過辨識模型將先前記錄的字元影像特徵與分類器進行運算，找出最相似的文字。第五，輸出文字辨識結果。

其中分類器與辨識模型程序雖然已有許多不同的辨識方法[7][9][10]被提出，但由於大多採單一字元候選字機率最高者為結果的基本架構，因此仍會有發生辨識錯誤的可能。若是能利用詞彙、語法或語意關係等語言線索，設計一個有效的校正模組，或許可以進一步提升傳統文字辨識系統的效能。例如以傳統辨識模組辨識「今天」兩個手寫字，並假設辨識模組對每個字元影像都會產生三個較可能的候選字元。其中第一個影像辨識結果依機率高低分別為「金」、「今」、「會」。而第二個影像辨識結果分別為「天」、「大」、「夫」。傳統辨識模組會根據各字元最可能的候選字選出「金」和「天」。但若從詞彙的角度來看，則僅只有「今天」這組字元組合是有意義的詞彙。這說明了若運用像詞彙這類的語言線索，應可進一步提升傳統文字辨識系統的效能。

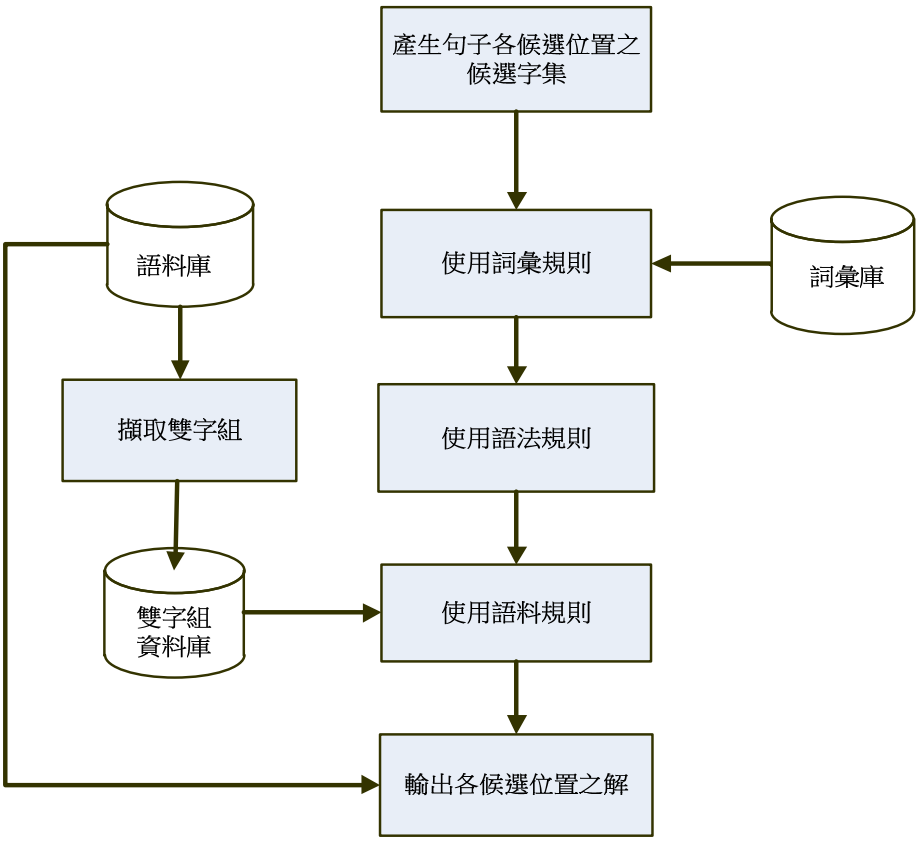
許多研究也提出類似的觀點用於提高辨識效果或是校正錯別字的方法。[1]提出一套兩階段手寫中文辨識系統，其在第二階段以一個上下文後處理器修正第一階段候選字選取的結果。[2]則提出一個判別正確字的設計來改善印刷體中文字辨識結果，這個設計主要利用詞語訊息來判別正確字是否包含在候選字集內。[3]提出一個方法來更正文件中的錯誤字，主要是利用在主題文章中某些詞彙會有重覆出現的特性(例如：主題為學校時，經常會出現學生、教室、老師…等詞彙)，更正錯誤字。[4]則發表一個錯別字偵錯與訂正建議的系統，可偵測文章中的錯誤字。

本文的基本假設是，在特定主題語料中，某些文字相鄰出現的情況相當頻繁，因此

二字元之 OCR 影像產生的候選字元可兩兩配對，由每一配對在語料中所出現的機率、或是其語法規則的合理性，可推測正確的結果。因此，本文將提出一個三階段方法，分別運用詞彙規則、語法規則與語料規則，自傳統辨識模組提供的候選字中挑選正確的字。在第一階段中，我們利用詞彙規則從可能的字元組合中挑選詞彙庫出現的詞彙做為挑選字。接著針對在第一階段中未確認的字元影像序列，尋找是否出現符合語法規則的詞性組合，並以該詞性組合所對應的字元組合作為對應字元影像的挑選字。最後針對兩個相鄰未確認的候選位置，形成一組字元組合，而以在語料中共發生機率最高的字元組合作為對應候選位置的挑選字。

## 二、系統架構

圖一為本論文所提方法的架構圖。首先，每一處理句內的每個字元影像會有一個代表正確字的空間，稱為候選位置，例如一個句子經掃描分析後有八個字元影像，則表示該句有八個候選位置。接著對每個候選位置都會產生一組候選字集，也就是該字元影像經辨識後產生的候選字集合。這些候選字集將先使用詞彙規則來進行詞彙挑選。詞彙規則是藉由詞彙庫提供詞彙，並以符合規則的字元組合作為對應候選位置的系統解。接著使用語法規則來找出符合規則的詞性組合，並以符合規則之詞性組合所對應的字元組合做為對應候選位置的系統解。最後使用語料規則挑選雙字組。雙字組是從語料庫中所擷取並收錄於雙字組資料庫，因此存在於雙字組資料庫的字元組合做為對應候選位置的系統解。經由三個規則的挑選流程後，最後若有候選位置仍無法確認系統解，則以語料字頻決定。經由上述流程處理，各候選位置均產生系統解，也就是每個字元影像都有系統所判斷的對應字。



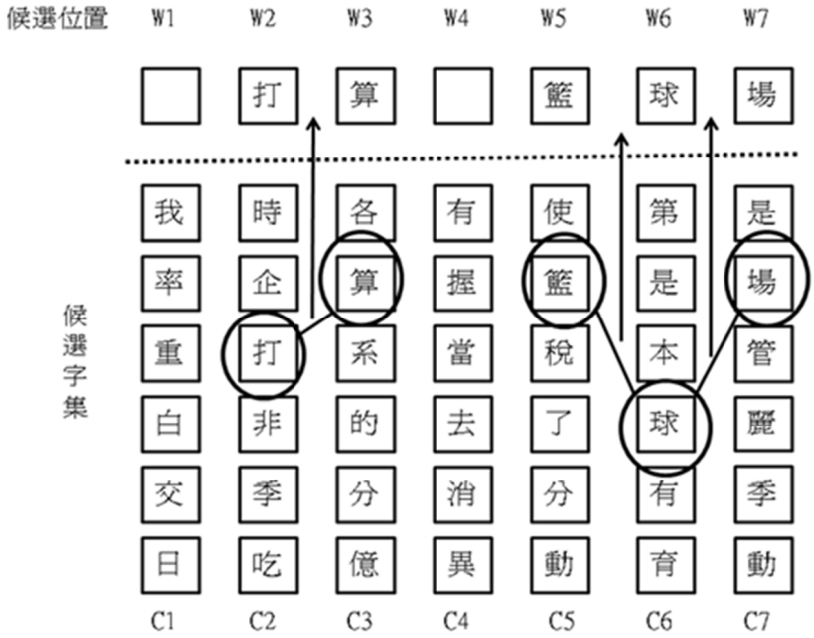

圖一、系統架構圖

## 三、規則使用

### （一） 詞彙規則

由相鄰候選位置之候選字集內的字元產生各種字元組合，再將所有的字元組合與詞彙庫中的詞彙進行比對，若有字元組合與詞彙相同，則將此字元組合作為對應之候選位置解。此方法我們稱為詞彙規則。圖二說明如何運用詞彙規則可以找出數個連續相鄰位置的解。

圖二以一個包含七個字元影像的句子為例，有七個候選位置及其所屬候選字集，依序分別為 C1、C2、...至 C7，且假設每個候選字集內皆有六個候選字元。候選位置 W1 的候選字集 C1＝｛我、率、重、白、交、日｝，候選位置 W2 的候選字集 C2＝｛時、企、打、非、季、吃｝，以此類推。首先根據詞彙規則，在候選位置 W2 與候選位置 W3 兩個相鄰候選位置中，將 C1 與 C2 兩個相鄰候選字集內的候選字元形成 36 組字元組合，並將這 36 組字元組合與詞彙庫中的二字詞彙進行比對。在,這些字元組合中，僅有字元組「打算」可在詞彙庫中找到，因此候選位置 W2 的解為「打」、而候選位置 W3 的解為「算」。另外，根據詞彙規則，在候選字集 C5、C6、及 C7 三個相鄰候選字集中，可將相鄰的候選字元形成 216 組字元組合，並將兩百十六組字元組合與詞彙庫中的三字詞彙進行比對。在這些字元組合中，僅只有字元組「籃球場」可在詞彙庫中找到。因此分別將「籃」「球」「場」做為候選位置 W5、W6、及 W7 的解。




圖二、詞彙規則範例

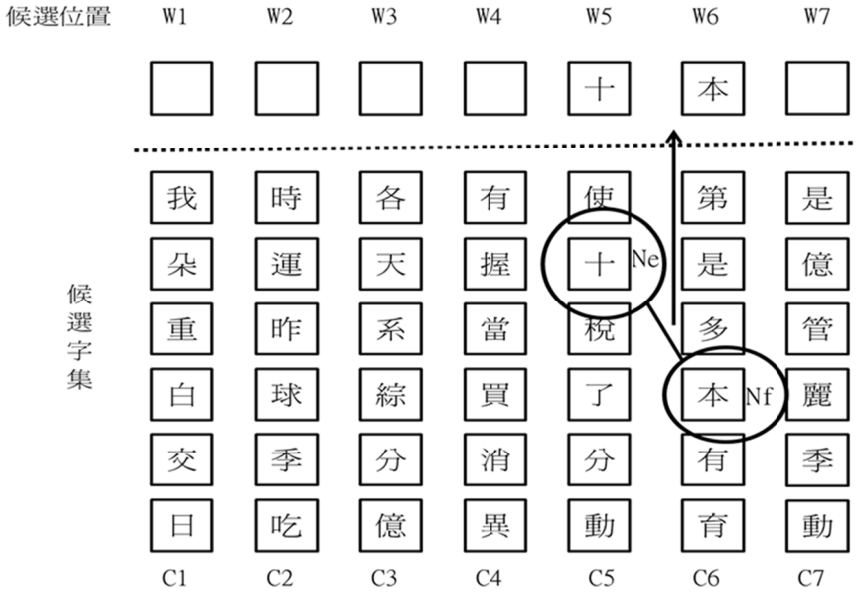
由圖二可以發現，根據詞彙規則，可在 C5、C6 及 C7 三個相鄰候選字集挑出三字詞「籃球場」，也可在 C5 及 C6 挑出二字詞「籃球」。其中，「籃球」與「籃球場」皆為符合詞彙規則的字元組合，此時發生有兩組以上解皆符合規則、應選擇何者為解的困擾，本文稱為規則衝突。由於規則衝突有許多類型，前述衝突本文稱為詞彙包含衝突。

在詞彙包含衝突中，字數較多的詞將作為正確解。因為三個相鄰候選字集中，隨機

挑選一組符合詞的字元組合的可能性，遠低於兩個相鄰候選字集中，隨機挑選一組符合詞的字元組合的可能性。換句話說，三字詞作為解的可能性遠大於二字詞隨機結合另一個單字詞的可能性。因此本文優先挑選字數較多的詞做為候選位置解。這樣的概念也被用於中文斷詞，稱為「長詞優先法」[5]。

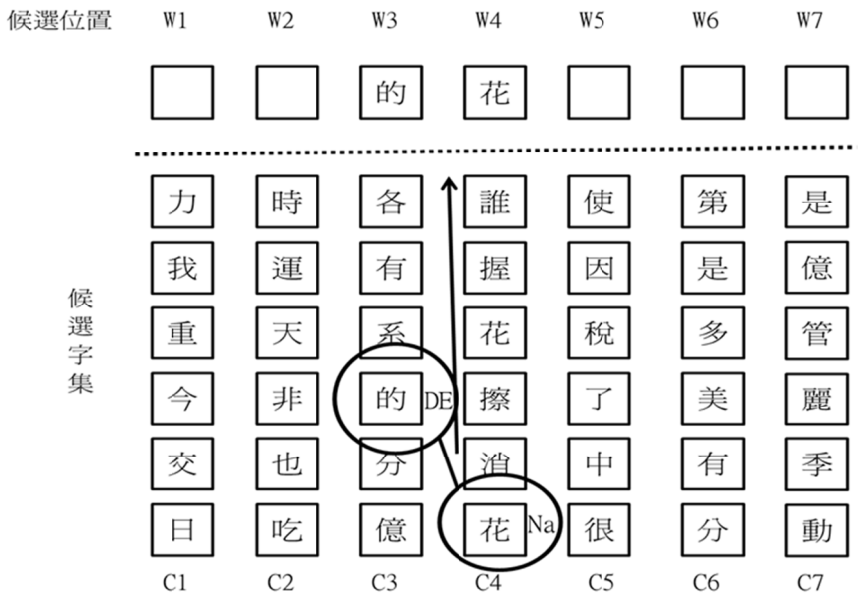
## （二） 語法規則

詞性是語法的基本單位，根據[6]的定義，中文總共有三十六種詞性。在語料中可觀察到許多詞性組合常常發生，例如詞性 Ne 是數詞(如：三、六、千...等字之詞性)，詞性 NF 代表量詞(如：個、對、片…等字之詞性)。由於描述數量的句子出現次數非常頻繁，因此若在兩個相鄰的候選字集中出現詞性組合「Ne-Nf」，則此詞性組合相對應的字元組合應該被挑選作為對應候選位置的解。本文將以兩種常見的詞性組合做為語法規則。第一種是詞性 Ne 後緊接著出現詞性 Nf 的組合，記為「Ne-Nf」規則。圖三說明了一個使用「Ne-Nf」規則的情境。




圖三、使用 Ne-Nf 規則之範例

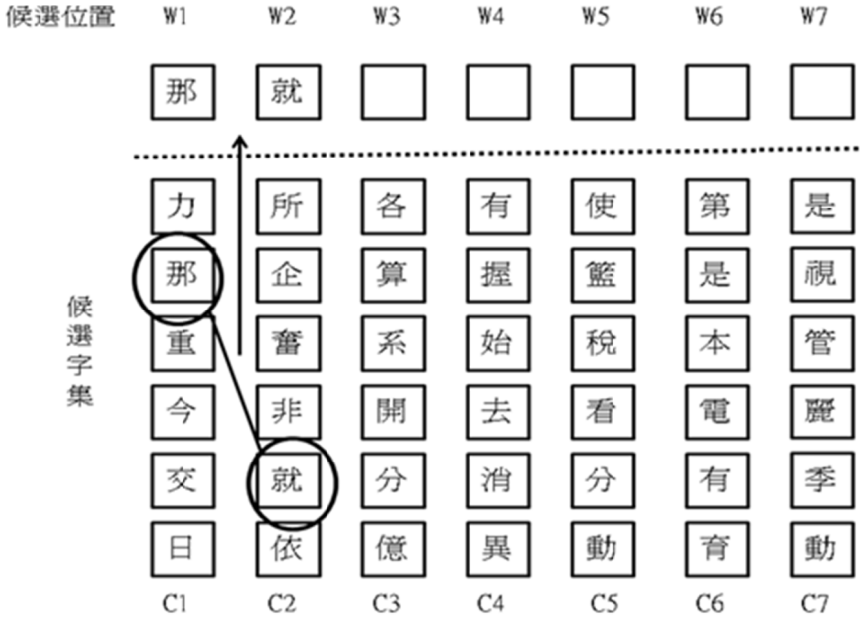
第二種是是詞性 DE 後緊接著出現詞性 Na 的組合，記為「DE-Na」規則。在中文常見的字元「的」，其常見作用有兩類，一種是類似中文所有格的概念，例如：「我的手」；另一種是用於形容詞與名詞間，例如「藍藍的花」。前述的例句，中文「的」之詞性是 DE，「花」與「手」的詞性皆為 Na。由於這類詞性 DE 的字後緊接出現一個詞性 Na 的詞之現象在語料中非常普遍，因此在候選子集群中若出現「DE-Na」規則，則可推測候選位置的解應為詞性 DE 與詞性 Na 所對應之字元組合。圖四說明了一個使用「DE-Na」規則的情境。




圖四、使用 DE-Na 規則之範例

## （三） 語料規則

語料中可觀察到有些單一字元和另一個單一字元頻繁地一起出現，例如「那就這麼決定吧！」中的「那就」、「我是這裡主管。」的「我是」等二字元組合。此種字元組合本文稱為「雙字組」。圖五為使用雙字組「那就」挑選候選字的範例。與二字詞不同，雙字組沒有明顯的語意，只是單純地以高頻率出現在語料中的兩個單字組合。這種現象在特定主題或領域(domain-specific)的語料中特別容易出現。本文以下列程序取得雙字組。首先將語料庫中的一個句子視為一個處理單位，將句中的兩個相鄰字元形成一字組。接著統計所有字組在語料中出現頻率。最後出現頻率超過門檻設定值的字組則收錄至雙字組資料庫。若在收錄過程中發現該字組是一組二字詞或符合語法規則的字元組合，則排除收錄。



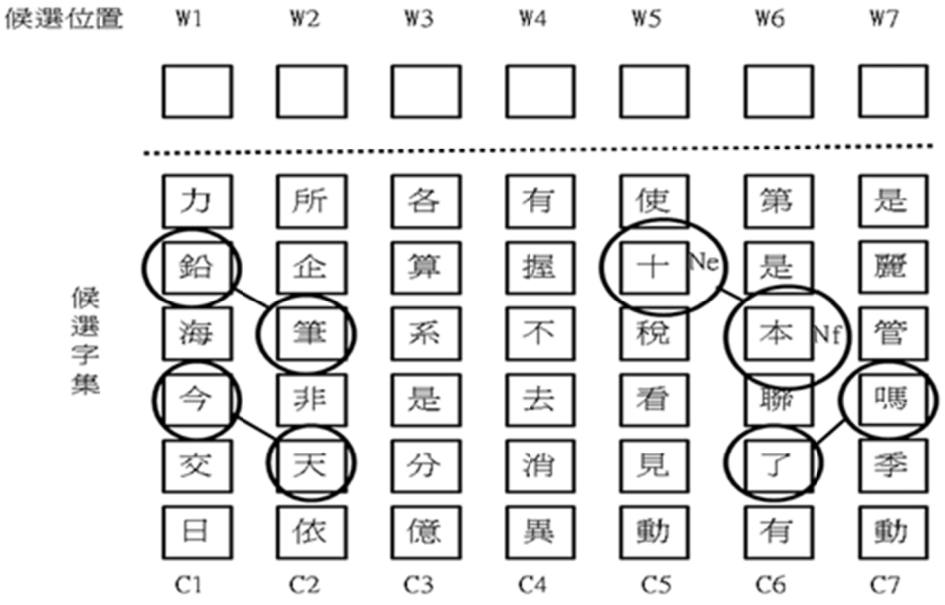

圖五、語料規則範例

有了雙字組資料庫，便可將候選字集群中兩兩相鄰候選位置之候選字集內的候選字元，形成許多字組，並將字組依序與雙字組資料庫進行比對。若該字組被收錄，則將此雙字組做為候選位置的解。

## 四、規則衝突處理

三之一節曾提到，在相鄰候選字集中挑選字元時若有兩組以上解皆符合規則，會發生應選擇何者為解的問題，稱為規則衝突。圖六列舉部分規則衝突的情況。在規則衝突可分「相同規則衝突」與「相異規則衝突」。相同規則衝突是指在兩個或三個相鄰候選位置之候選字集中，出現兩組或兩組以上的字元組符合同一規則的挑選條件。而相異規則衝突是指在兩個或兩個以上的相鄰候選字集中，出現兩組或兩組以上的字元組合符合兩個不同規則。若相同規則衝突或相異規則衝突在全部的位置重疊，這種情況稱為「全部位置重疊衝突」，如圖六之候選位置 W1 及 W2；若只有若干個位置重疊，這種情況則稱為「部份位置重疊衝突」,如圖六之候選位置 W5 至 W7。

由於有許多可能造成衝突的規則組合，以下三小節分別以三種規則為發生衝突的規則之一，討論規則衝突時的處理方法。




圖六、規則衝突範例

### （一） 詞彙規則的衝突

在詞彙規則造成的規則衝突分為相同規則衝突與相異規則衝突。詞彙規則之相同規則衝突是指在候選字集內的候選位置出現兩組或兩組以上的詞彙組合，造成詞彙與詞彙之間衝突。在詞彙規則衝突上，不論是全部位置衝突或部份位置衝突，對詞彙規則的相同規則衝突皆採取詞頻較高的詞彙作為對應之候選位置的系統解。

詞彙規則之相異規則衝突有兩種，分別是語法規則衝突及語料規則衝突。當詞彙規則與語法規則發生衝突時，則是以詞彙規則為優先。當詞彙規則與語料規則發生衝突時，若詞彙長度大於 2，則詞彙規則優先。而一個二字詞與一個雙字組之間的衝突，其解決方式是針對二字詞及雙字組各自設定一組門檻值，以下列三種情況來判斷優先規則。首

先，若二字詞的頻率超越門檻值，則使用詞彙規則。第二，若二字詞的頻率未超過所設定的二字詞門檻值，而雙字組的出現頻率超過雙字組所設的雙字組門檻值，則使用語料規則。最後，當二字詞和雙字組皆未超過各自的門檻值，則以詞彙規則為優先。

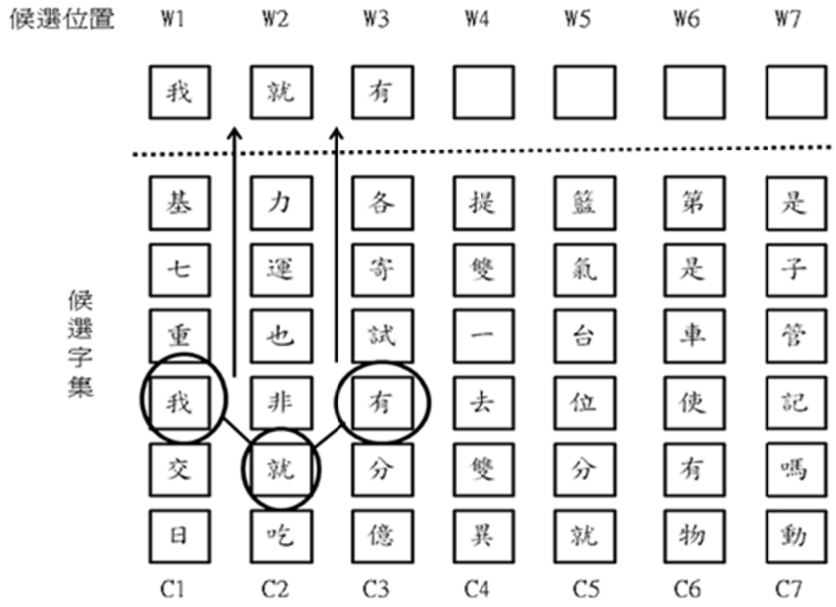
## （二）語法規則的衝突

語法規則所包含的規則衝突也可以分為相同規則衝突與相異規則衝突。語法規則之相同規則衝突解決方式是比較個別衝突位置之兩個字元，以單一字頻較高的字元做為對應之候選位置的系統解。

語法規則可能與詞彙規則及語料規則發生相異衝突。與詞彙規則的衝突處理已於四章一節說明。而與語料規則之衝突解決方式是，若雙字組的發生頻率超過所設定的門檻值，則優先以雙字組做為系統解，否則以符合語法規則的字元組合做為系統解。而不論全部位置重疊及部份位置重疊衝突，皆採取相同處理方式。

## （三） 語料規則的衝突

語料規則與其他規則產生的衝突已在先前兩小節討論。對於語料規則的相同規則衝突，不論相同位置衝突及部份位置衝突，皆以發生頻率較高的雙字組做為對應位置的解。在部份位置衝突情況中，有時候會出現衝突雙字組共用一個字元的現象，這種情形我們不視為規則衝突，而是同時將兩個雙字組做為三個對應位置的解。圖七以範例說明了這種情形的細節。當雙字組「我就」與「就有」在候選位置 W2 上同時與候選字「就」成為雙字組，此時不視為發生語料規則衝突，而是直接以「我就是」做為 W1 至 W3 的解。




圖七、語料規則之部分衝突位置字元相同範例

## 五、 實驗

### （一） 實驗環境

本文使用的特定主題語料是由國立臺灣師範大學心理與教育測驗研究發展中心所提供之國民中學九年級學生寫作作品。寫作篇數共計 1234 篇，寫作題目為「用餐時刻」。該語料每篇作品平均字數為 349 個字，平均每句字數為 9 個字。本實驗的測試資料是從全部寫作文本中進行 5 次隨機挑選，每次隨機挑選 200 篇寫作做為一組測試資料集。在挑選的過程中，對於先前已被挑選過的寫作文本，將不再挑選。測試資料集之後將以「資料集」簡稱之，所取得的五次資料集亦分別稱為資料集 1 至資料集 5。

本文先進行兩項實驗以便後續效能評估，第一項以純字頻挑選系統解的效能建立測試基準，第二項則測試本文所提方法必須設定的各項門檻值對正確率的影響。之後效能評估都分別對資料集 1 至資料集 5 測試，並將這 5 次測試得到的正確率取平均值做為該項效能評估之正確率。另外，本文模擬以光學辨識字元影像產生候選字集的方式進行實驗，亦即對每句中的每個已知正確字，加上隨機從字典中選取九個字形成該字之候選位置的候選字集。當之後同一個字再度出現在其他句子中，仍使用相同的候選字集，也就是本實驗不考慮正確字不在候選字集中的情形。

第一項實驗是藉由「中央研究院中文分詞詞典」所提供每個候選字元發生頻率，將每個候選字集中出現頻率最高者的字元來做為對應候選位置的解。此方法以「純字頻挑選」簡稱之，可以說明在沒有本文所提方法下，以最簡單的純字頻法可達成的正確率。表一為純字頻挑選之實驗結果。此 5 組資料集的平均正確率為 45.8%。由結果得知，若無使用任何校正方法，純字頻挑選方法的效能相當不理想。

表一、純字頻挑選方法實驗結果

| | 資料集 1 | 資料集 2 | 資料集 3 | 資料集 4 | 資料集 5 | 平均值 |
|---|---|---|---|---|---|---|
| 總測試字數 | 69051 | 63144 | 66155 | 67153 | 61723 | 65445 |
| 正確挑選字數 | 31796 | 28523 | 30349 | 30492 | 28873 | 30006 |
| 錯誤挑選字數 | 37255 | 34621 | 35806 | 36661 | 32850 | 35449 |
| 正確率百分比 | 46.0% | 45.1% | 45.8% | 45.4% | 46.7% | 45.8% |

本文的第二項實驗將設定 4 組門檻值分別對 5 個資料集進行測試。4 組門檻值分別簡稱為「設定集 1」、「設定集 2」、「設定集 3」、「設定集 4」，每個設定集包含兩個值，分別是二字詞出現次數的門檻值(以下簡稱 T 值)、以及雙字組出現次數的門檻值(以下簡稱 B 值)。設定集 1 的參數設定值為 T=212，B=106。設定集 2 的 T=106，B=52。設定集 3 的 T=920，B=52。設定集 4 的 T=52，B=920。這些數值是依據所有二字詞及雙字組在語料中出現頻率排序後，位於全體 25%、50%、75%的頻率值所設定。實驗結果如表二所示。

表二、各種門檻值組合對各資料集的選字正確率

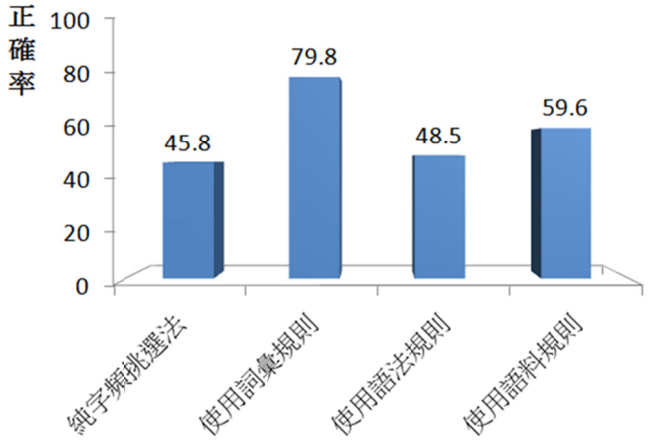
| 資料集 / 門檻設定值 | 資料集 1 | 資料集 2 | 資料集 3 | 資料集 4 | 資料集 5 | 25%, |
|---|---|---|---|---|---|---|
| T=212, B=106 | 84.8 | 84.6 | 85.0 | 84.4 | 85.5 | 84.8 |
| T=106, B=52 | 84.7 | 84.5 | 85.0 | 84.4 | 85.6 | 84.9 |
| T=920, B=52 | 80.8 | 80.3 | 80.7 | 80.5 | 81.5 | 80.8 |
| T=52, B=920 | 84.4 | 84.3 | 84.5 | 84.1 | 85.0 | 83.7 |

由表二可知，設定集 1 與設定集 2 的實驗結果的正確率差異並不大，代表同時調低二字詞與雙字組門檻值影響不大。經檢視資料後發現，雖然某些出現次數較少之二字詞及雙字組在調低門檻值後可以被詞彙與語料規則使用後正確校正，但較少出現次數也代表用來做為正確解的錯誤風險增加。在兩相抵消之下，其正確率差異並不大。

設定集 3 的門檻值相較於設定集 1，調高了二字詞門檻值但降低雙字組門檻值。其造成的影響為詞彙規則較少使用、而增加使用語料規則的次數，實驗結果顯示設定集 3 的正確率較低，代表使用語料規則取代詞彙規則會造成效能下降。而設定集 4 的門檻值相較於設定集 1，則是調低了二字詞門檻值但提高雙字組門檻值。其造成的影響為增加錯誤風險較高的詞彙規則的使用次數、而提高語料規則的信賴度。實驗結果顯示設定集 4 的正確率較低，但較設定集 3 為高。這說明了詞彙規則的正確性較語料規則的正確性為高，應該優先使用詞彙規則，然而適當取得兩個門檻值的平衡，才能使效能接近最佳。五之二節將採用設定集 2 的門檻值做為評估依據。

## （二） 效能評估

本小節的實驗分為「單項規則效能分析」與「架構效能評估」，目的是測試個別規則對正確率的影響。單項規則效能分析是針對只使用單一規則的效能。而由於規則執行後並非全部候選位置均能得到系統解，因此未有系統解的候選位置則利用純字頻挑選法產生解。由於此實驗僅針對單一規則進行測試，在挑選文字的過程中不會發生規則衝突，因此無需設定門檻值來處理規則衝突。

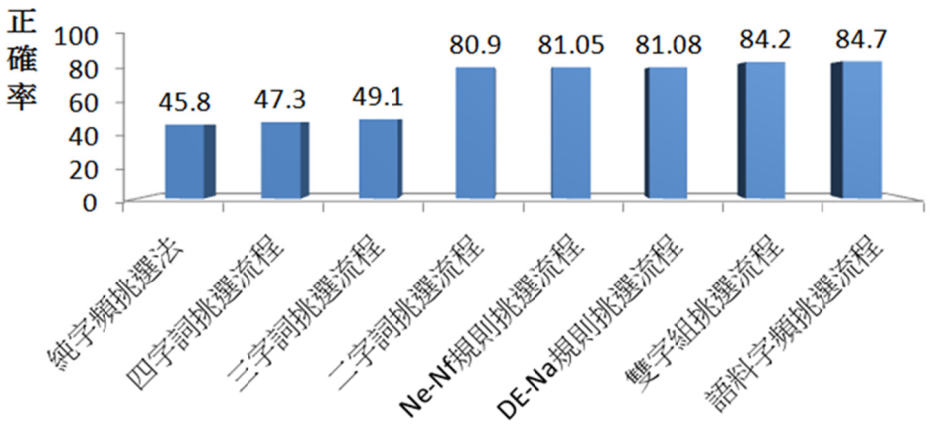



圖八、單項規則效能分析

由圖八所示，以使用詞彙規則與純字頻挑選法比較，使用詞彙規則挑選文字，其正確率可從 45.8%提升到 79.8%。使用該規則可提升 34%的正確率。在單項規則效能分析之實驗結果中，該規則提升的正確率最高，主要原因是中文句所包含的字元大多是由詞彙組成，且詞是句子的語意基本單元。因此使用詞彙規則可以有效提高正確率。

語法規則與純字頻挑選法比較，使用語法規則挑選文字，其正確率可從 45.8%提升到 48.5%。使用該規則可提升 2.7%的正確率。在此實驗中該規則所能提升的正確率最低。經檢視該規則的實驗結果發現，使用語法規則確實可以找到正確的字元組合，但由於實驗語料沒有大量出現符合該規則的句子片段，因此導致正確率提升有限。

再以語料規則與純字頻挑選法比較，使用語料規則挑選文字，其正確率可從 48.8%提升到 59.6%。相較純字頻挑選法，使用該規則可提升 13.8%的正確率。該規則對於提升正確率已有明顯效果，但低於詞彙方法的提升率。主要原因是當語料規則發生相同規則衝突，本研究是挑選出現次數較多的雙字組做為挑選字。這種解決方法並不保證所挑選的雙字組是正確解，所挑選的雙字組也可能造成挑選錯誤。本文所提之三類規則所提升的正確率差異性很大，也與本文所設計的規則優先順序相當符合。




圖九、各流程執行後正確率變化

圖九則顯示依序使用第二節所述本文所提架構各階段執行結束後正確率之變化。在五之一節中已說明使用純字頻挑選法之實驗正確率為 45.8%。在各階段文字挑選處理後仍未確認的候選位置，則仍使用純挑選字頻法進行字元挑選。

詞彙規則階段因長詞優先以及詞長不同分為四字詞、三字詞及二字詞挑選流程。由圖九實驗結果可知，加入二字詞挑選程序後才會大幅提升正確率，主要原因為二字詞佔多字詞的比例相當高，相較之下四字詞與三字詞出現比例偏低，因此能處理的字數有限。特別說明的是，圖九中，使用詞彙規則的正確率比在圖 8 中要稍高，其因是圖 9 之實驗有設定門檻值解決二字詞或雙字組的規則衝突，因此正確率與圖 8 有所差異，此亦顯示使用門檻值有些微提高正確率的效果。

語法規則緊接在詞彙規則使用後加入。語法規則在圖九中也分為兩個子程序，先使用 Ne-Nf 規則再加入 DE-Na 規則，其正確率從百分之 80.9%提升至 81.8。很明顯的其正確率提升效果有限，原因如圖八之實驗討論。最後語料規則也分為兩個子程序，先使用雙字組規則再加入語料字頻挑選法，正確率由 81.08%提升至 84.7%。與圖八的實驗結果比較，此處的語料規則能提升的正確率較低，顯示最後實施的語料規則可校正的部

分有許多與前兩項規則重複，因此有些語料規則可校正部分已先由前兩類規則正確校正，因此正確率提升有限，但仍具有提升正確率的效果。

## 六、結論

本文是利用特定領域語料特性提出三類規則做為挑選候選字依據以提高離線手寫文字辨識正確率。在詞彙規則中，由於詞是句子的語意基本單元，藉由這個特性辨識模型可以利用詞彙規則來挑選多字詞的詞彙，並做為對應位置的校正結果。而在數個連續相鄰的候選字集中，欲隨機組合出一組符合詞彙規則的字元組合之可能性相當低，因此以符合詞彙規則的字元組合做為對應之候選位置的解，較有可能正確挑選。在辨識模型辨識文字的結果中，可以使用詞彙規則來進一步校正辨識字，以提高辨識的效果。

使用語法規則可以從候選字集群中，找出符合規則的詞性組合。由於傳統辨識模型是以模型計算機率值較高的候選字做為辨識結果，因此其選出的相鄰字間並不一定符合語法規則，會出現整句不符語法的現象。加入語法規則可以在候選字集群中找出符合規則的的字元組合做為解，藉由解決整句不符語法問題改善候選字挑選流程。有別於詞彙規則與語法規則，語料規則相當依賴特定領域語料的特性，以頻繁相鄰出現的字元組合來做為對應位置的系統解。由於字元組合資訊是由語料蒐集，可表現語料使用語言的偏好，因此對於辨識準確度會有不錯的改善效果。

雖然本文已經有效提升光學文字辨識的正確率，但還有一些部分可進一步研究。首先，本文提出的詞彙規則會發生相同規則衝突與相異規則衝突。對於詞彙規則所發生的相同規則衝突，本文是以詞頻較高的詞彙做為對應候選位置的正確解，而針對詞彙規則所發生的相異規則衝突，本文是優先以較頻繁的二字詞來做為解。這種以發生頻率做為衝突解決依據的方法所造成的錯誤佔校正失敗相當大的比例。如何更有效解決規則衝突問題值得進一步研究。另外，本文所提方法可考慮提前於文字辨識的第二階段結合，提早進行候選字篩選，如此可產生更為可靠的候選字集，使得挑選正確解的可能性更高。

## 誌謝

作者感謝國立臺灣師範大學心理與教育測驗研究發展中心提供語料供研究使用。

## 參考文獻

[1] 李宜靜，2009，“中文字印刷體影像文字辨識之研究”，義守大學，碩士論文。

[2] 謝尚琳，1999，“用於辨正印刷中文字辨識結果的實用設計”，國立臺灣大學，博士論文。

[3] 曾元顯，2004，“應用於資訊檢索的中文 OCR 錯誤詞彙自動更正”，中國圖書館學會會報，72 期，頁 23~31，6 月。

[4] Yong-Zhi Chen, Shih-Hung Wu, Chia-Ching Lu and Tsun Ku, “Chinese Confusion Word Set for Automatic Generation of Spelling Error Detecting Template”, The 21th Conference on Computational Linguistics and Speech Processing , Taichung, Taiwan ,September 1-2, pp.359-372, 2009.

[5] Keh-Jiann Chen and Shing-Huan Liu, “Word identification for Mandarin Chinese sentences”, In Proceedings of COLING-92, Nantes, France,pages 101-107, 1992.

[6] CKIP, “Analysis of Syntactic Categories for Chinese”, CKIP Tech. Report#93-05, Sinica, Taipei, 1993,

[7] Mingrui Wu, Bo Zhang and Ling Zhang, “A neural network based classifier for handwritten Chinese character recognition”, 15th International Conference on Pattern Recognition, Barcelona, Spain, September 3-8, pp.2561-2564, 2000.

[8] Zhi-guo He and Yu-dong Cao, “Survey of Offline Handwritten Chinese Character Recognition”, Computer Engineering, Vol.34, No.15, pp.201-204, 2008.

[9] Feng-Jun Guo, Li-Xin Zhen, Yong Ge and Yun Zhang, “An Efficient Candidate Set Size Reduction Method for Coarse-Classifier of Chinese Handwriting Recognition”, Proceedings of the 2006 conference on Arabic and Chinese handwriting recognition, College Park, MD, USA, September 27-28, pp.152-160, 2006.

[10] Hairong Lv, Wenyuan Wang, Chong Wang and Qing Zhuo, “Off-line Chinese signature verification based on support vector machines”, Pattern Recognition Letters, Vol.26, Issue 15, pp.2390-2399, 2005.