中文混淆字集應用於別字偵錯模板自動產生

Chinese Confusion Word Set for Automatic Generation of Spelling Error Detecting Template

陳勇志 Yong-Zhi Chen, 吳世弘 Shih-Hung Wu
朝陽科技大學資訊工程系
Department of Computer Science and Information Engineering
Chaoyang University of Technology
{9727602, shwu}@cyut.edu.tw

盧家慶 Chia-Ching Lu, 谷圳 Tsun Ku
資訊工業策進會
Institute for information industry
{gaty, cujing}@iii.org.tw

摘要

本研究透過常用字來產生混淆字集，自動產生能夠幫助錯別字偵測的模板，發展華語文錯別字偵測技術。本系統利用辭典爲基礎，使用辭典中的詞彙做爲正面用詞，透過混淆字集自動產生含別字的反面模板，能夠偵測的別字包含同音字、同部首字，並且透過斷詞軟體輔助擷取更正確的反面模板，用以協助華文教師進行大量華文作文的錯別字批改甚至輔助學生進行寫作，最後達到提昇寫作能力之成效。

關鍵詞：模板產生、模板探勘、正反面用語知識庫

Abstract

In this research, we proposed a system that can use automatically generated templates for detecting Chinese spelling error. At first, we use frequently used Chinese characters to produce the Chinese confusion set. Based on a dictionary, our system automatically generated negative vocabulary template with the help of Chinese confusion set. Error types include pronunciation-related errors and radical-related errors. And our system uses word segment to capture more accurately the negative template. We hope that such a system can help the teachers on the checking of students' essays, and also can help students learn to write effectively. Consequently, the students would improve their writing skill.

Keywords: Template generation, Template mining, Pragmatics Knowledge Base.

一、緒論

自民國 95 年起，教育部在國中基本學力測驗中加辦「寫作測驗」隨後列入升學計分，計分標準依據立意取材、結構組織、遣詞造句、錯別字給予 6 個等第的級分，華語學習中的作文能力備受重視。國中基本學力測驗每年約有三十萬學生應試，因此我們可以預見未來將有大量的作文輔助批改與輔助教學的需求，如何應用數位學習的技術來輔助教師批改作文並且幫助學生學習寫作，爲目前普遍研究之議題。

根據寫作測驗的評分依據，錯別字是個重要的評分標準，回顧以往中文錯別字的輔助學生系統的相關文獻有[1]與[2]，這兩篇文獻都是針對中文文章中進行偵錯與訂正的系統，其中[1]是利用替換五筆字型編碼來產生可能的別字，透過每次替換一個編碼即可達到產生多個可能的別字，而五筆字型輸入法主要用於使用簡體中文的中國大陸，五筆字型完全依據筆畫和字形特徵對漢字進行編碼，將漢字筆劃分爲橫 、豎、撇、捺、折五種，把字根或編碼按一定規律分佈在 25 個英文按鍵上。教育部也針對錯別字推出由人工編寫的常用國字辨似[3]，但是常用國字辨似只含有 1477 筆模板，並不敷大量的作文偵錯使用，而我們從書上蒐集常用的正反面用語模板含有 6,701 筆，並且在 2008 年發表中文作文的訂正與更正建議系統[4]，該系統利用學生所書寫的作文蒐集偵錯用模板藉以建立模板偵錯技術的正反面模板偵錯，並且透過統計 Corpus 的 uni-gram、bi-gram 建立語言模型做爲常用語偵錯，由於人工蒐集模板費時耗力且成本過高，所以隨後我們根據 QA 系統中的自動模板產生概念[5]、[6]，利用機器學習之技術並且大量探勘 Corpus 中可能的反面用語模板，最後利用學生所書寫得作文做爲 Training data，於 2009 年 5 月發表了中文作文錯別字偵錯模板自動產生[7]，該系統使用模板擴展演算法來取得大量的模板，並且透過卡方檢定做爲收納模板的檢定公式，但是自動模板產生還是依賴人工蒐集模板中之種子，於是我們使用混淆字集來大量產生蒐集模板用的種子，混淆字集爲一般人容易混用的字之集合，混淆字可能爲同音字、同形字、同部首字等等。

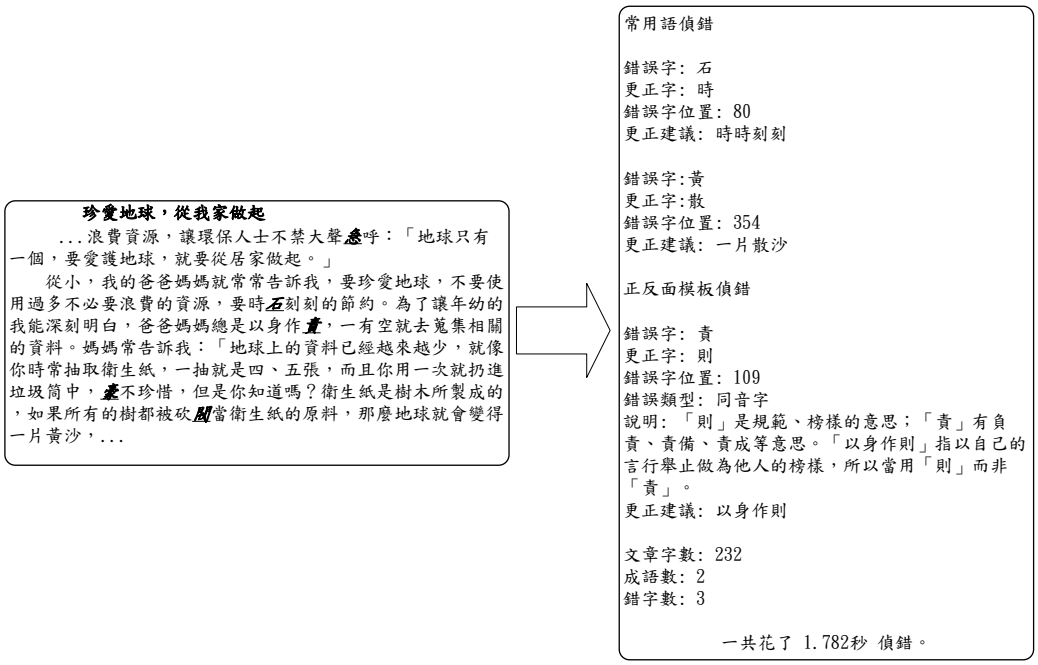
混淆字集是學生容易將正確的字書寫成錯誤的字之集合，根據劉昭麟教授的統計學生作文[8]、[9]，學生書寫的錯別字中同形字佔 30.70%、同音字佔 79.88%、同形同音字佔 20.91%、非同形同音字佔 2.43%，統計結果指出學生書寫的的錯別字大部分來自於同音字錯誤，同音字的混淆字集可以透過字典收集取得而同形字則較不易取得，不過劉教授依據倉頡輸入法發表[10]，利用替換倉頡輸入法字碼來取得同形字與[1]的替換五筆字型編碼相似之處。

二、系統設計與方法

（一）錯別字自動模板產生系統回顧

我們於 2008 於所發表的錯別字偵錯與訂正建議系統[4]是透過一個 Web 介面，讓學生輸入他們所書寫的作文而文章可能含有若干個別字如圖一左邊方塊，經過我們系統兩項功能常用語偵錯與正反面模板偵錯診斷後，常用語偵錯會提供學生的錯誤字、必須更正的字、錯誤字位置和更正建議的資訊，正反面用語偵錯除了常用語偵錯提供的建議之外還會提供詞語的說明如圖一右邊方塊，我們的系統可以偵測出常見的錯別字，並且明確指出學生錯別字在文章的何處，並且給予適當的建議與說明，讓學生瞭解自己何處寫錯字並且從錯誤中學習。

我們在 2009 年 5 月所發表的自動模板產生系統[7]是改進 2008 年發表[4]的系統，2008年的系統所使用的模板必須經由人工蒐集，由人工蒐集模板費時耗力且成本過高，自動模板產生系統確實能夠自動產生大量的偵錯模板不過卻有兩個缺點，1. 產生模板的正、別字種子必須經由人工蒐集。2. 其自動產生的部份模板不具可讀性且不符合詞彙的概念，如圖二。圖二中我們可以看出某些詞彙如“辯護律”、“視辯論”、“電視辯”等並不是完整的詞彙，如“辯護律”可能是從“辯護律師”擷取，這些不完整的詞彙並不適合當作更正建議資訊給使用者參考，因此下面我們根據以上兩個缺點進行改進。



圖一、2008 年所發表系統之偵錯功能

26749	會首長	會首常	7116	清潔隊長	清潔隊常
26750	會給予	會給于	7117	交通隊長	交通隊常
26751	辯論會	辨論會	7118	辯護律師	辨護律師
26752	辯護律	辨護律	7119	視辯論會	視辨論會
26753	的辯論	的辨論	7120	政策辯論	政策辨論
26754	視辯論	視辨論	7121	電視辯論	電視辨論
26755	電視辯	電視辨	7122	公開辯論	公開辨論
26756	半世紀	辦世紀	7123	半個世紀	辦個世紀
26757	半以上	辦以上	7124	一年半的	一年辦的
26758	半個小	辦個小	7125	的另一半	的另一辦

圖二、舊系統所產生的部份模板

（二）混淆字集

我們發表過的偵錯系統所使用的正、別字種子是經由我們所蒐集正、反面用語中擷取其中的正、別字所產生，例如：正反面用語爲“芭蕉”、“笆蕉”，而正、別字種子則爲“芭”、“笆”，由於正、別字種子也是必須經由人工蒐集，同樣也具有費時耗力

成本過高的缺點。根據劉教授的統計[8]、[9]，同音同形字的別字在學生所寫的錯別字中佔有 89.67%，其中同音錯別字高達 79.88%。

我們從字典蒐集所有用字的注音並且將同音同調視爲同音字，如“動”的注音爲“ㄉㄨㄥˋ”，因此“凍”“ㄉㄨㄥˋ”視爲同音字，“東”“ㄉㄨㄥ”視爲不同音字，而我們所蒐集的同音字表共有 1,351 音、15,160 字，如圖三。

漢字字形的構成要素是由筆劃、筆順、偏旁、六書、部首所構成，其中部首是東漢文字學家許慎所著之《說文解字》所創，此後漢字的檢字方式一般皆使用部首，而同部首字含有高相似度，因此我們使用部首資訊來產生同形字，根據《康熙字典》漢字的部首一共有 214 個，我們利用 214 個部首蒐集了 9,752 個漢字並且產生同部首字表，如圖四。

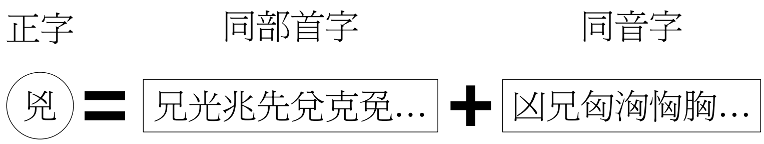
最後自動產生混淆字集方法則是使用正字檢索同音字表、同部首字表，如圖五概念圖利用“兇”即可找出同部首的“兄光兆先兌克免…”與同音字的“凶兄匈洶恟胸…”等字。

1	ㄅㄚ	蚆扒八巴仈叭朳芭疤捌笆粑豝釟吧
2	ㄅㄚˋ	伯罷霸猈弝爸壩灞把耙
3	ㄅㄚˊ	鈸茇拔胈跋菝詙軷魃犮
4	ㄅㄚˇ	鈀把靶
5	ㄅㄚ˙	吧杷琶罷
6	ㄅㄛ	剝嚗波袚玻柭砵缽啵菠碆撥皤蹳鱍帗播襏
7	ㄅㄛˋ	播檗蘗亳擘譒北擗薜簸繴
8	ㄅㄛˊ	爆伯犮襮㧊蔔柏瓝薄泊暴濼鎛帛勃胉挬浡郣脖舶桲
9	ㄅㄛˇ	簸跛蚾
10	ㄅㄛ˙	蔔

圖三、部份同音字表

1	一	一丂丁七三下丈上万丌丑丐不丏丙世丕且丘丞丟並
2	丶	丸凡丹主
3	丿	乂乃久么之尹乍乏乎乒乓乑乖乘
4	乙	乙九乜也乞乣乩乳乾亂
5	亅	了予事
6	二	二于云井互五亓亙些亞亟
7	亠	亡亢交亦亥亨享京亭亮亳亶亹
8	人	人仁什仃仆仇仍今介仄仂仉以付仔仕他仗代令仙仞
9	儿	兀元允充兄光兇兆先兌克免兕兔兒兗党兜兢
10	入	入內全兩

圖四、部份同部首字表

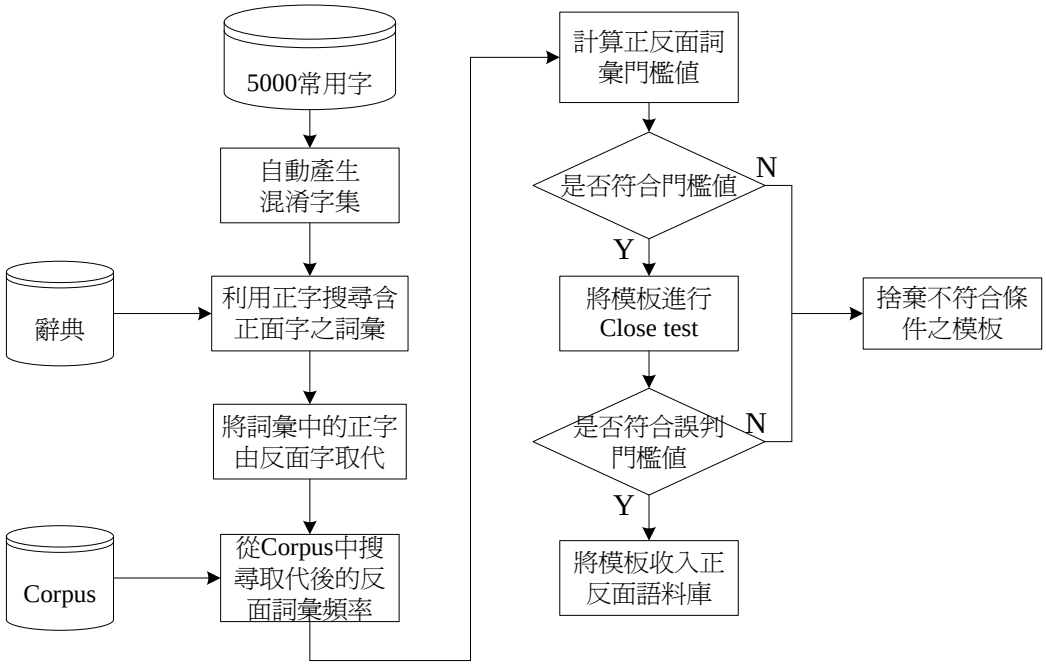


圖五、自動產生混淆字概念圖

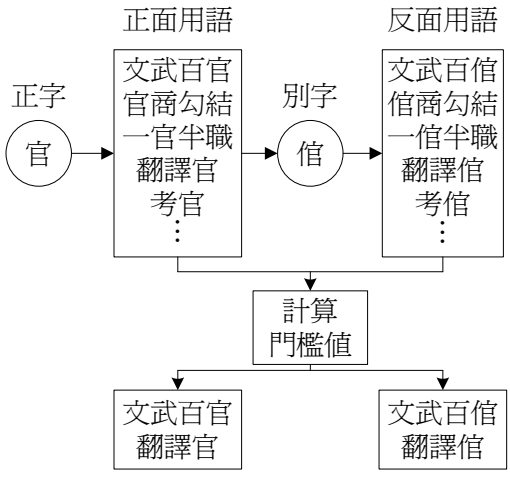
（三）自動化收集模板系統流程

圖六爲我們自動模板產生系統流程圖，我們的自動模板產生系統是基於正面用詞非常頻繁使用，而含別字的反面用語則會被使用很少的條件下，首先我們蒐集由國語推行委員會所公佈的八十七年常用語詞調查報告書[11]中的常用字共 4,998 字作爲正字種子，接著利用這些常用字自動產生同音、同部首的別字，最後將混淆字集輸入至我們的自動模板產生系統。

我們發表過的偵錯系統是使用演算法來產生正面用語模板但會有稍早所提之缺點，於是我們使用現有詞彙的基礎作爲正面用語模板，當系統取得正字後會去檢索辭典是否有包含該正字之詞彙，其中辭典爲教育部所公佈之教育部重編國語辭典修訂本[12]，經由我們濾掉單字詞彙共 145,608 詞，接著我們將檢索之詞彙的正字替換成別字做爲正、反面用語模板，接著到 Corpus 進行頻率統計，統計頻率如果符合門檻值的模板則收集起來進行 Close test，如果符合 Close test 誤判門檻值的模板最後則將此模板收入正反面語料庫，而自動模板產生的概念如圖七。



圖六、系統流程圖

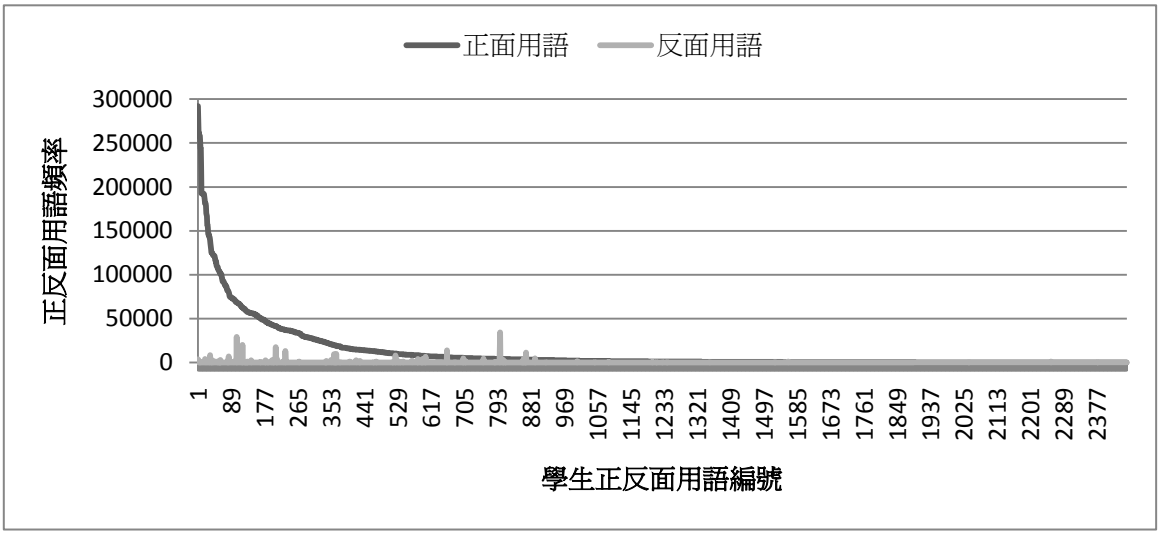


圖七、自動模板產生概念圖

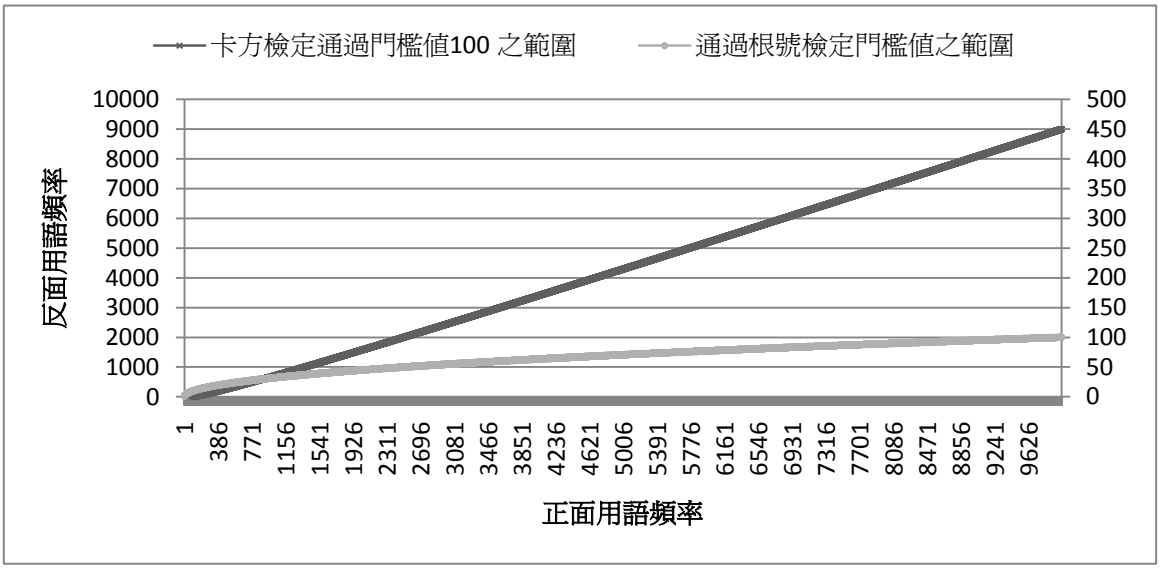
檢定公式方面，我們在 2009 年 5 月所發表的自動模板產生系統[7]是使用卡方檢定來檢定是否收納模板如(1)，其中 E 爲正面用語模板的出現頻率 O 爲反面用語模板的出現頻率，而中文中常有積非成是的用語或通用詞詞彙，爲了避免這樣的情況我們會限定 E > O。

$$X^2 = \frac{(O-E)^2}{E} \qquad (1)$$

隨後我們觀察學生作文中學生所使用的反面用語與教師訂正後的正面用語，發現正面用語的頻率遠大於反面用語如圖八，而卡方檢定公式特性卻只學生正反面用語的頻率分佈不同其卡方檢定特性圖如圖九卡方檢定的門檻值範圍，在門檻值設定爲 100 的條件下，隨著正面用語頻率的提昇而反面用語也呈線性的提昇，也就是圖九上方線段以內的反面頻率皆會通過卡方檢定測試，這與我們從學生所使用的正反面用語有著非常大的差異，所以卡方檢定並不適合用來檢定模板是否收納。



圖八、正反面用語頻率分佈圖



圖九、卡方檢定與根號檢定門檻值分佈圖

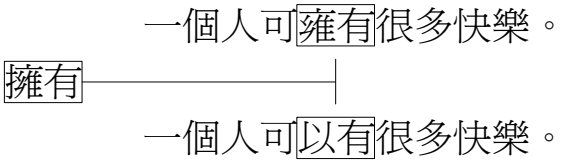
因此我們將否採納該模板的公式修改如(2) (3)，*Cfeq* 爲正面用語的頻率、*Wfeq* 爲反面用語的頻、*Threshold* 爲所有正面用語頻率之平均，而採納模板的條件是正面用語的頻率經過開根號的計算之後必須大於反面用語的頻率，且正面用語必須大於門檻值。使用此公式是依據圖八學生使用的正反面用語之特性所設計，依照每個正面用語得頻率取得相對之反面用語頻率，並且必須符合正面用詞非常頻繁使用，反面用語則被使用很少的條件，根號檢定的特性圖如圖九下方線段，根號檢定能夠針對每一個正面用語頻率去取得他的最佳反面用語頻率之門檻值，最後我們將學生正反面用語頻率共 2455 筆利用(2)公式做頻率分佈分析，其中共有 90.46%的模板符合根號檢定之測試，不符合此檢定測試的模板有“未來”、“爲來”，“已經”、“以經”，“但是”、“但事”等模板，這些模板的前後文資訊不足如果用來當作錯別字訂正模板，則會非常容易引入雜訊。

$$\sqrt{Cfreq} > Wfreq \quad , \quad Cfreq > Threshold \qquad (2)$$

$$Threshold = \frac{\sum_{i=1}^{n} Cvocabulary(i)}{n} \qquad (3)$$

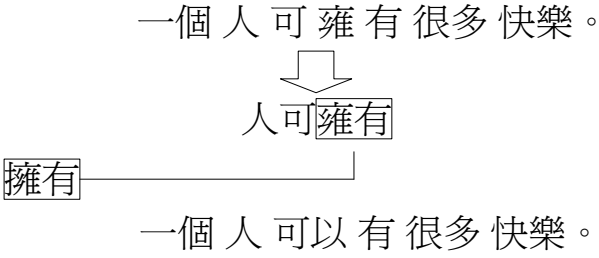
（四）斷詞軟體應用

上述檢定公式與 Close test 處理根據我們實驗與觀察，2 個字的詞彙仍然非常容易造成 False alarm，這個原因是 2 字的詞彙過短容易與其他詞彙發生重疊的現象範例如圖十，如果用一個正面詞彙如“擁有”去擷取反面用語，則會如圖十般將“雍有”、“以有”都收入，但在“一個人可以有很多快樂”這句範例中，“以有”中的“以”字應屬於“可以”這個詞彙。



圖十、詞彙重疊現象範例

斷詞軟體能夠將正確的詞彙斷詞，而含有別字的詞彙則無法正確斷出詞彙如圖十一，因此我們使用這個特性將 Corpus 利用斷詞軟體[13]斷詞，藉以用來擷取更正確的 2 字詞彙模板，我們會將正確斷詞的詞彙移除接著將剩餘單字詞合併用來擷取模板用。最後由我們系統自動產生的模板如圖十二。



圖十一、應用斷詞擷取反面用語模板範例

395	衝擊	衝急	437	絆腳石	伴腳石	879	逼不得已	逼不得已
396	檢視	機視	438	大部分	大不分	880	情非得已	情非得已
397	經濟	經紀	439	手電筒	手電桶	881	逼不得已	逼不得已
398	循環	循還	440	不經意	不經易	882	大勢已去	大勢以去
399	成績	成積	441	不願意	不願易	883	不能自已	不能自以
400	薪水	新水	442	董事長	懂事長	884	迫不得已	迫不得以
401	賺錢	購錢	443	三輪車	三軸車	885	情非得已	情非得以
402	關鍵	關建	444	腦震盪	腦振盪	886	萬不得已	萬不得以
403	老闆	老版	445	辦公室	辨公室	887	逼不得已	逼不得以
404	雖然	隨然	446	成績單	成積單	888	巡弋飛彈	巡曳飛彈

圖十二、經由我們系統所產生的部份模板

三、實驗結果與分析

（一）Corpus 與學生作文

由於統計模板頻率需要大量的語料資料，因此我們蒐集新聞語料庫做爲我們的 Corpus，資料整理如表一。

表一、Corpus 資料整理

資料年份	新聞社	文件數	檔案大小
1998-1999	Chinatimes	38,163	209MB
	Chinatimes Commercial	25,812	
	Chinatimes Express	5,747	
	Central Daily News	27,770	
	China Daily News	34,728	
1998-1999	United Daily News	249,508	320MB
2000-2001	United Daily News	172,421	1.03GB
	United Express	91,958	
	Ming Hseng News	168,807	
	Economic Daily News	463,873	

測試集是從學生作文中拆成兩個部份，一個部份做爲 Close test 用，另一部份則是用來 Open test 用，學生作文我們使用台北市某國中七、八年級考試作文、並且由教師校訂過錯別字共 3264 篇，每篇文章皆輸入成電腦可處理的格式如圖十三，而我們的系統並不處理注音文以及不存在於 Unicode 編碼中之錯字。

最後我們將蒐集到的作文做資料分析如表二，從表格中我們可以看出約 94%的作文用字皆爲常用字的範圍，而表三則爲學生作文的正別字分析同部首的正別字約在 15%同音正別字約 68%，而非同部首同音字約爲 19%，我們的作文統計分析結果與劉教授的分析結果[8]、[9]相近。另外我們也統計學生常犯錯誤之 Top 10 模板如表四。

```
<doc>
<class>七年一班</class>
<number>7</number>
<title>藉口</title>
<score>4.5</score>
<essay>
<p>人，有許多夢想，尼采說：「人因夢想而偉大。」雖然是這麼說，不過光「想」是不會有<revise><wr
<p>你是否曾找過一些<revise><wrong>冠冕唐荒</wrong><correct>冠冕堂皇</correct></revise>的藉口
<p>人非聖賢，誰能無過？知過能改，善莫大焉，摒除藉口，是一個需要決心、毅力、耐心的工程，我常常
<p>燕子去了有再來的時候，<revise><wrong>楊柳估了</wrong><correct>楊柳枯了</correct></revise>
</essay>
</doc>
```

圖十三、作文電子檔的格式

表二、學生作文基本分析

	作文數	平均級分	作文平均字數	平均別字數	常用字比例
Close test essay	2241	3.62	367.12	1.74	94.23%
Open test essay	1023	3.61	420.02	1.94	94.33%

表三、學生作文正別字分析

	正別字同部首比例	正別字同音比例	兩者皆有	兩者皆非
Close test essay	13.82%	70.27%	4.92%	20.81%
Open test essay	16.96%	66.31%	2.85%	19.58%

表四、常犯錯誤之 Top 10 模板

Close essay	正面用語	已經	變得	自己	景象	一旦	寄託	已經	畢竟	而已	根本
	反面用語	己經	變的	自已	景像	一但	寄托	以經	必竟	而己	跟本
Open essay	正面用語	自己	一旦	已經	選擇	煩惱	應該	已經	而已	選擇	後悔
	反面用語	自已	一但	己經	選則	煩腦	因該	以經	而己	撰擇	後侮

（二）實驗設計與評估

我們實驗的比較對象爲[4]所人工蒐集的模板與發表過的偵錯系統[7]所產生的模板，由於二字詞、三字詞、四字以上的詞彙出現頻率差異非常大，因爲我們針對這三組分別計算全部詞彙的平均頻率，依照平均頻率門檻值分別設定爲：2300、500、100，而 Close test 的過濾門檻值經由我們反覆實驗得到 0 爲最佳的設定。

評估的方式是使用 Precision 與 Recall 公式定義如下:

$$\text{Micro Recall} = \frac{\sum(\frac{dr}{r})}{N} \quad (4) \qquad \text{Micro Precision} = \frac{\sum(\frac{dr}{sd})}{N} \quad (5)$$

$$\text{Macro Recall} = \frac{\sum(dr)}{\sum(r)} \quad (6) \qquad \text{Macro Precision} = \frac{\sum(dr)}{\sum(sd)} \quad (7)$$

$$\text{False alarm rate} = 1 - \text{Precision} \quad (8)$$

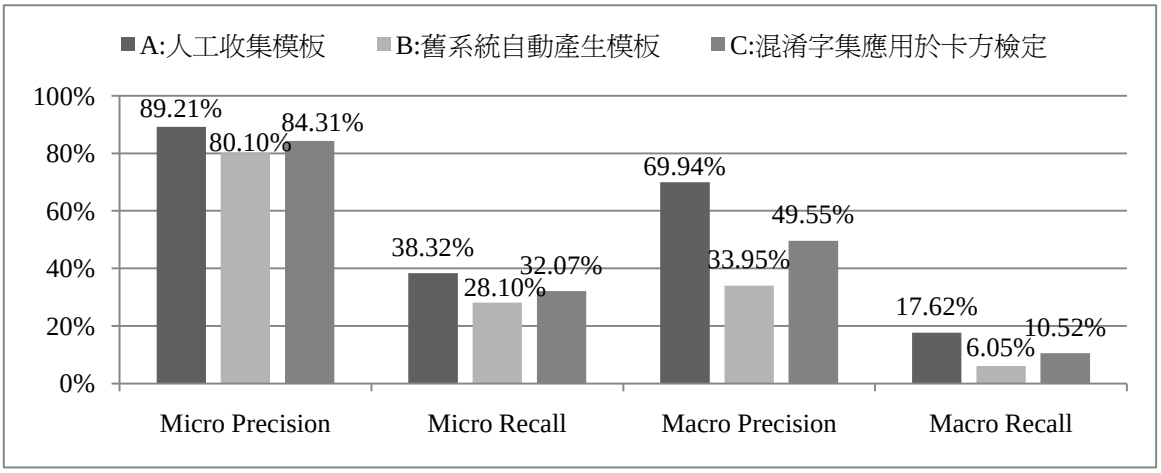
dr 爲每篇文章中偵錯正確的字數，r 爲每篇文章中真正的錯字數，sd 爲 每篇文章中系統偵測出的錯字數，N 爲所有文章的篇數。Micro Precision 與 Micro Recall 是以接近現實生活的偵錯情形，也就是以文章爲單位偵錯效能如何。除此之外還必須考量較

多的樣本，也就是將所有的資料視爲整個大集合，所以我們使用 Macro Recall 與 Macro Precision 來檢視系統的效能。最後我們系統要求的是在維持高 Precision 的情況下來提高 Recall 值，因爲我們不希望給使用者太多 False alarm。

（三）實驗結果

我們設計四組實驗第一組實驗爲混淆字集應用於卡方檢定時的實驗結果，第二組爲混淆字集應用於根號檢定的實驗結果，第三組爲根號檢定加入斷詞後的實驗結果，第四組爲根號檢定所自動產生的模板加入人工蒐集模板之後的實驗結果。

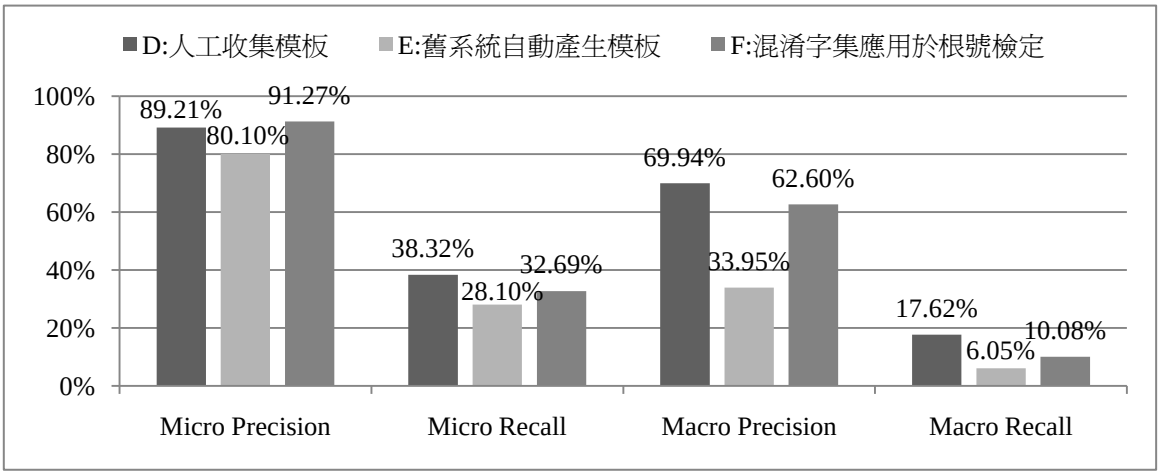
實驗一：混淆字集應用於卡方檢定



圖十四、卡方檢定實驗結果

實驗結果如圖十四，其中 A 組爲人工蒐集的模板共 6,701 筆，B 組爲我們發表過系統所產生的模板共 19,402 筆，C 組爲應用混淆字集後使用卡方檢定之系統產生的新模板共 54,253 筆，其中混淆字採取[11]中的常用字，不在常用範圍的字則不在此範圍。Precision 方面以人工蒐集的模板爲最佳，而 Recall 方面應用混淆字集卡方檢定所自動產生的模板優於過去我們所發表的系統，整體來說則還是以人工蒐集的模板爲最佳，不過自動產生的模板在 Recall 皆都逼近人工蒐集模板的數值。

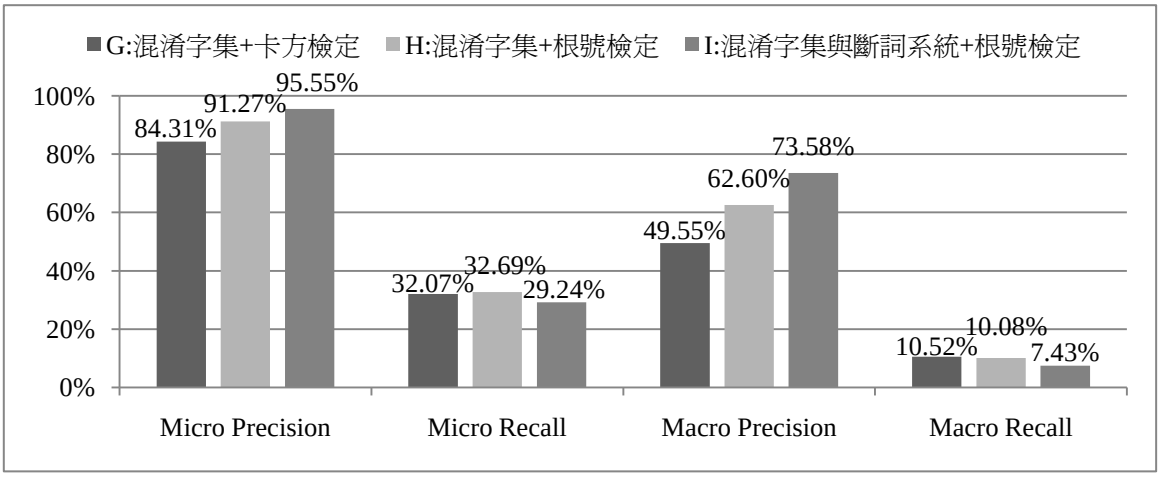
實驗二：混淆字集應用於根號檢定



圖十五、根號檢定實驗結果

實驗結果如圖十五，D、E 組與實驗一的 A、B 相同，F 組爲應用混淆字集後使用根號檢定之後系統產生的新模板共 50,467 筆。與實驗一最大的不同處是根號檢定在 Precision 方面皆比卡方檢定來得優異許多其中以 Macro 提昇最多，在 Recall 方面 Micro 些微提昇 Macro 則是些微下降。由此實驗可以得知過去卡方檢定的檢定模板方式讓許多 noise 進入造成 Precision 的下降，改用根號檢定的檢定方式可以改善以往的缺點。

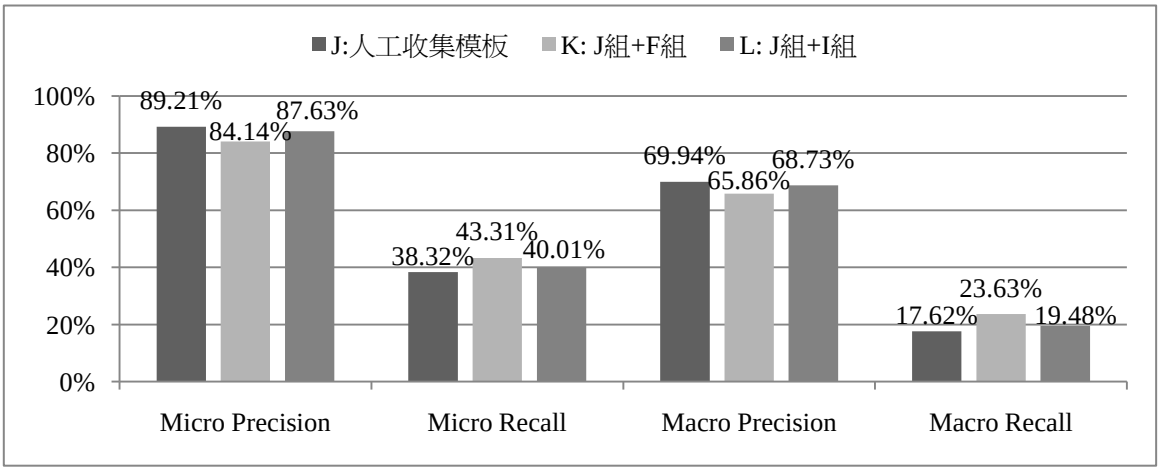
實驗三：加入斷詞系統



圖十六、應用斷詞之檢定比較

實驗結果如圖十六，G 組爲實驗一 C 組之卡方檢定模板，H 組爲實驗二 F 組之根號檢定模板，I 組爲 H 組應用斷詞軟體後自動產生的新模板共 9,013 筆。由於斷詞軟體的使用讓斷詞更準確，因此模板產生數跟前面兩個實驗相比較降低不少，在 Precision 方面可以發現不論 Micro 或 Macro 都比使用斷詞軟體前的模板在準確度有更進一步的提昇，但是在 Recall 部份則是下降 3%左右，這是追求 Precision 所犧牲的地方。

實驗四：混合人工蒐集模板



圖十七、比較混合人工蒐集模板之效能

實驗結果如圖十七，J 組爲實驗一 A 組人工蒐集的模板共 6,701 筆，K 組爲實驗二 F 組根號檢定模板與 J 組人工蒐集的模板混合使用共 57,167 筆，L 組爲實驗三 I 組應用斷詞軟體之根號檢定模板與 J 組人工蒐集的模板混合使用共 15,713 筆。Precision 方面混合人工蒐集模板後的自動產生模板皆下降到人工蒐集模板水平附近，而跟實驗三比較

Recall 方面在 Micro 與 Macro 部份皆有大幅度的提昇，這也表示我們的系統具有可擴充性，如果加入適當的模板能夠使用系統的偵錯範圍更進一步的提昇。

以混淆字爲基礎來產生模板的新系統理當可以掌握 70~80%的錯別字，但是經由我們的實驗卻發現自動產生的模板在 Recall 值方面的提昇非常有限，這個現象我們將在下節做數據分析。

（四）實驗結果分析

用來分析的模板我們使用實驗三中的應用斷詞軟體之根號檢定自動產生的新模板共 9,013 筆做爲分析模板，因爲這組模板比較符合我們當初所預期之在維持高 Precision 的情況下來提高 Recall 值。

1 Precision 方面

利用反面模板來偵測錯誤理當能夠讓 Precision 達成 100%，但是經由我們實驗結果發現卻不是如此，我們將 Open test 中系統偵錯部份造成 False alarm 提出討論如表五。根據[12]“垃圾桶”“垃圾筒”、“奇蹟”“奇跡”、“電線桿”“電線杆”、“銷聲匿跡”“消聲匿跡”，可以得知此四組模板爲通用詞，而“一再”“一在”則牽涉到語意層面再這邊並不適合使用模板的方式來偵錯，“放聲大哭”“放聲大叫”、“不用說”“不用講”、“讀書人”“讀書做”則是我們系統收納到不適合的模板，Precision 無法達到 100%就是上述原因所導致。

表五、部份 False alarm 模板

正面用語	垃圾桶	奇蹟	電線桿	銷聲匿跡	一再	放聲大哭	不用說	讀書人
反面用語	垃圾筒	奇跡	電線杆	消聲匿跡	一在	放聲大叫	不用講	讀書做

2 Recall 方面

我們將學生所書寫正反面用語模板與我們系統所產生的模板做個分析如表六。其中“沒產生到的模板”爲我們系統所沒有產生到的模板，“不在辭典”爲在沒有產生到的模板中其正面用語不在辭典中，“不在 Corpus”爲在沒有產生到的模板中其反面用語不在 Corpus 中，“兩者皆是”爲正面用語不再辭典中同時相對應反面用語也不在 Corpus 中。

從“沒產生到的模板”數值可以發現，絕大部分學生所書寫得模板並沒有被我們自動產生，再從“不在辭典”與“不在 Corpus”數值中相加並且扣除兩者皆有的部份，可得知 Close test essay 有 53.17%的模板與 Open test essay 有 32.97%，是我們的系統無法自動產生出來，因爲我們的系統自動產生模板是基於正確詞彙與 Corpus 曾經有人使用過該反面用語。

至於不存於辭典的詞彙如表七，可以 將這些詞彙加入辭典這樣便可以克服此問題，而不存在於 Corpus 的反面用語則必須蒐集更大量的 Corpus 語料庫，以便能夠蒐集到此類的反面模板。

表六、學生模板與系統模板分析

	沒產生到的模板	不在辭典	不在 Corpus	兩者皆是
Close test essay	91.53%	37.73%	35.64%	20.20%
Open test essay	93.15%	16.27%	23.94%	7.24%

表七、部份未收入辭典之詞彙

佈告欄	蒸飯機	值日生	作業本	辦派對	睡午覺	全班齊心	勤加練習	羞恥心	無厘頭
重拾信心	莽莽撞撞	淘汰	漆彈場	偶像劇	積陰德	融入團體	芬多精	燒炭	拉筋

四、結論及未來工作

根據我們應用混淆字集、根號檢定公式與斷詞軟體，我們能夠省去人工蒐集產生模板用種子的流程，並且能夠產生以詞彙爲基礎的模板如圖十二，改進過去發表過的系統模板如圖二非詞彙基礎的模板，給予使用者更明確的訂正資訊。使用根號檢定公式也經由實驗得知確實能夠比卡方檢定所自動產生的模板有較佳的 Precision，最後藉由斷詞軟體斷詞後的 Corpus 也經由實驗證實能夠更進一步提昇系統的 Precision，而 Recall 部份也能透過持續增加適合的模板來增加偵測率。

在未來我們會蒐集混淆字集所沒辦法產生的模板產生種子，也會持續蒐集更符合學生作文的文章來取代新聞語料庫，詞彙方面則會透過大型詞彙庫或線上資源如：維基百科，來增加我們辭典的詞彙數，最後我們預計使用學生的作文產生含別字之語言模型，利用該語言模型來智慧偵錯以輔助模板偵錯所沒有蒐集到的錯誤模板。

致謝

本研究依經濟部補助財團法人資訊工業策進會「98 年度智慧型網路服務技術與應用計畫(2/4) 」辦理。

參考文獻

[1] Lei Zhang, Chang ning Huang, Ming Zhou, Haihua Pan, *Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm*, Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pp: 248-254, 2000.

[2] Ren, F. , Shi, H., Zhou, Q., *A hybrid approach to automatic Chinese text checking and error correction*, In Proceedings of the ARPA Work shop on Human Language Technology, pp: 76-81, March 1994.

[3] MOE, *Common Errors in Chinese Writings (常用國字辨似)*, Ministry of Education, Taiwan, 1996.

[4] Ta-Hung Hung., & Shih-Hung Wu, *Chinese Essay Error Detection and Suggestion System*. Taiwan E-Learning Forum, 2008.

[5] Cheng-Lung Sung., Cheng-Wei Lee., Hsu-Chun Yen., Wen-Lian Hsu, *An Alignment-based Surface Pattern for a Question Answering System*, the IEEE International Conference on Information Reuse and Integration, pages pp. 172-177, 2008.

[6] D. Ravichandran., & E. Hovy, *Learning surface text patterns for a Question Answering system*, in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 41-47, 2001.

[7] 陳勇志, 吳世弘, 盧家慶, 谷圳, *中文作文錯別字偵錯模板自動產生*, The 13th Global Chinese Conference on Computer in Education, pp. 402-408, 2009.

[8] Chao-Lin Liu, Kan-Wen Tien, Min-Hua Lai, Yi-Hsuan Chuang, Shih-Hung Wu, *Phonological and logographic influences on errors in written Chinese words*, Proceedings of the Seventh Workshop on Asian Language Resources, the Forty Seventh Annual Meeting of the Association for Computational Linguistics, August 2009.

[9] Chao-Lin Liu, Kan-Wen Tien, Min-Hua Lai, Yi-Hsuan Chuang, Shih-Hung Wu, *Capturing errors in written Chinese words*, Proceedings of the Forty Seventh Annual Meeting of the Association for Computational Linguistics, August 2009.

[10] Chao-Lin Liu and Jen-Hsiang Lin, *Using structural information for identifying similar Chinese characters*, Proceedings of the Forty Sixth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 2008.

[11] 國語推行委員會, *八十七年常用語詞調查報告書*, National Languages Committee, Taiwan, 1998.

[12] MOE, *教育部重編國語辭典修訂本*, Ministry of Education, Taiwan, 2007.

[12] CKIP, "Autotag," Academia Sinica, 1999.