第四章　實驗分析

4.1　新詞擷取

為證明本文所提出的方法，實驗的部份使用實際的新聞事件來套入本文提出的新詞自動擷取方法。表4-1~表4-5 為五個實際的新聞事件，粗體字即為各個新聞事件被擷取出來的字詞。可以看出新聞標題裡的關鍵字大多都能被提取出來，且正確率也有一定的水準。參數的部份 R_n 設定為1而M則設定為2。

表4-1　新詞擷取實驗結果(1)

<table>
<tr><td colspan="4">區委書記批發61頂官帽南充開審“書記賣官”案
南充"賣官區委書記"親戚也不打折庭審當庭翻供
四川開審全省最大的賣官案區委書記賣官61次
17日再審"南充賣官書記"
鄂蜀兩官員因涉嫌貪汙受賄面臨審判
四川省最大買官賣官案區委書記賣官61次
四川開審最大買官賣官案區委書記賣官61次
川最大「官帽批發商」受審
賣官「賺」二百五萬蜀區書記受審
賣官帽61頂斂財200萬(圖)
四川賣官書記批發61頂烏紗
南充賣官區委書記親戚也不打折庭審當庭翻供</td></tr>
<tr><td colspan="3">擷取出的新詞</td><td>新詞數量</td><td>正確率</td></tr>
<tr><td>[書記] => 22</td><td>[批發] => 2</td><td>[官帽] => 3</td><td rowspan="3">9</td><td rowspan="3">100%</td></tr>
<tr><td>[南充] => 5</td><td>[開審] => 2</td><td>[賣官] => 65</td></tr>
<tr><td>[四川] => 5</td><td>[最大] => 4</td><td>[受審] => 1</td></tr>
</table>

表4-2　　新詞擷取實驗結果(2)

<table>
<tr><td colspan="4">北京正陽門城樓發現裂縫管理人員稱暫無危險
古建專家：北京市明年將大修正陽門城墻
北京正陽門城樓發現達6米裂縫明年將大修
象征北京城的明建築正陽門城樓發現裂縫(組圖)
北京文物局回應正陽門裂縫事件稱其不會塌(圖)
正陽門裂縫不是結構問題明年將開始首次大修
北京正陽門城樓發現長達六米的裂縫明年將大修
正陽門城樓發現裂縫正陽門城樓發現裂縫
正陽門城樓發現裂縫
北京正陽門城樓發現裂縫明年將迎來大修
文物古建專家解釋墻體裂縫：正陽門城墻塌不了
象征北京城的明代建築正陽門城樓發現裂縫(圖)</td></tr>
<tr><td>擷取出的新詞</td><td></td><td></td><td>新詞數量</td><td>正確率</td></tr>
<tr><td>[北京] => 20
[文物] => 1
[正陽門城] => 11</td><td>[裂縫] => 41
[正陽門] => 15
[古建專家] => 1</td><td>[大修] => 7
*[明年將] => 4</td><td>8</td><td>75%</td></tr>
</table>

表4-3　　新詞擷取實驗結果(3)

雙布要打開和平“**路線圖**” **巴以**和談：**美國**重口惠 **希望**在**任期內**幫巴**建國** 布什**希望巴勒斯坦**4年內**建國** **美國**靜觀**巴以**走向 **布希**第二個**任期內**將堅持中東**和平路線圖**計劃 時評："後**阿拉法特**時代"是否中東**和平**的機會 [東方時空]永遠的戰士——**阿拉法特**(多圖) 綜述：以色列謹慎應對**阿拉法特**去世 **布希**與布萊爾表示支援**巴勒斯坦建國** [網友原創] 巴“三駕馬車”能否撿起橄欖枝 **布希**希望**巴勒斯坦**4年內**建國**				
擷取出的新詞			新詞數量	正確率
[和平] => 2	[巴以] => 1	[美國] => 1	10	90%
[希望] => 2	[建國] => 5	[布希] => 3		
[路線圖] => 1	*[任期內] => 1	[巴勒斯坦] => 2		
[阿拉法特] => 1				

表4-4　　新詞擷取實驗結果(4)

<table>
<tr><td colspan="4">環球時報：推動中日關係“再次正常化”
日傳媒報道指不明來歷潛艇可能是中國核潛艇
日本確定闖領海潛艇屬中國海軍核潛艇
外交部表示核潛艇闖日本水域事件已解決
日外相稱中共為潛艦入日本海域道歉
中國聲稱東海潛艇事件已妥善解決
日解除海上警備行動並考慮向中國抗議
日本確認中國潛艦侵犯領海正式向中國抗議
美智庫專家指潛艦事件凸顯中國戰略野心
衛星確認侵日領海的潛艦返抵中國海軍基地
神秘潛艇現身日本海日媒體借機炒作向中國抗議
日本徹夜追蹤神秘潛艇</td></tr>
<tr><td colspan="3">擷取出的新詞</td><td>新詞數量</td><td>正確率</td></tr>
<tr><td>[潛艇] => 23
[領海] => 3
[潛艦] => 6</td><td>[中國] => 29
[事件] => 2
[確認] => 1</td><td>[日本] => 13
[解決] => 1</td><td>8</td><td>100%</td></tr>
</table>

表4-5　　新詞擷取實驗結果(5)

費盧傑之役未竟**摩蘇爾**硝煙又起 伊拉克武裝繼續抵抗**美軍** 新華社通訊：**費盧傑**大逃亡 **美軍**進剿"**死亡**三角" **費盧傑**發現化武實驗室 **費盧傑**發現爆炸物和化武實驗室**巴格達**南戰火烈 伊政黨呼籲推遲**大選** 伊選舉委員會：選期不改 伊臨時政府稱**大選**將如期舉行 **扎卡維**挪窩**摩蘇爾**？ **扎卡維**組織稱對17名伊安全部隊成員被殺負責 **巴格達**三起襲擊事件至少32人**死亡**				
擷取出的新詞			新詞數量	正確率
[美軍] => 1	[死亡] => 1	[大選] => 1	8	100%
[費盧傑] => 5	[摩蘇爾] => 1	[巴格達] => 1		
[扎卡維] => 1				

接下來本文提出的方法將與Yih-Jeng Lin and Ming-Shing Yu [14]所提出的方法比較其字詞擷取的數量與正確率。由於Yih-Jeng Lin and Ming-Shing Yu提出的方法中取出net frequency>1的字詞為正確字詞，換句話說也就是取出所有重複過的字詞。由於計算的方式不同，為了公平起見，相對於本文提出的方法，將設定參數R_n=1，這也是本文提出的方法中最低的門檻值，而設另一個參數M=2。實驗將使用十個不同的新聞事件作為測試，並將實驗結果交由中文語言專家來判斷兩個方法的正確率。表4-6與表4-7為新聞事件一跟二的實驗結果，有*號的字詞為經過中

文語言專家所認定是不合理的，表4-8為所有十個新聞事件的實驗結果。N為實驗中總共擷取出來的新詞，R則為新詞的正確率

$$N = \text{total number of detected new words}$$

$$R = \frac{\text{number of correctly detected new words}}{\text{total number of detected new words}}$$

表4-6　　實驗結果比較(1)

拒認一中不提九二共識扁"十裁示"欺天下
拒認"一中"不提"九二共識" 扁拋出"十點裁示"
陳水扁建議兩岸共同商定軍事緩衝區
拒認一中不提九二共識扁“十裁示”巧言欺天下
國民黨稱陳水扁“重大談話”是騙選票的空話
陳水扁發表“重大談話”稱“四不一沒有”不改變
陳水扁提出“十點裁示”拒認一中不提九二共識
扁又在兩岸關繫上玩兩面手法
臺媒言論：扁又在兩岸關繫上玩兩面手法
扁對在野黨“釋利”？ 國親嗤之以鼻：騙票手段
陳水扁再發表"重大談話" 國民黨：騙選票的空話
國民黨："十項重要裁示"不過是陳水扁慣用伎倆
中國官員：台灣接受一中原則兩岸即可復談
翁松燃：陳總統談話恐仍難獲對岸善意回應
張五岳：總統十項裁示向國際展現誠意
美中台／扁國安會議 10 項裁示美表歡迎並籲兩岸重啟對話
扁提軍事緩衝區北京智庫：善意誠意待觀察台辦冷臉對
主軸未變談何善意陳水扁圖謀“ 海峽行為準則”
陳水扁提出“ 十點裁示” 拒認一中不提九二共識

陳水扁籲兩岸放棄大規模毀滅性武器

扁倡兩岸設軍事緩衝區

用「陽光政策」對抗中國

扁國安會議十項裁示邱義仁立即透過管道轉知美國

國台辦稱台灣若接受一中原則兩岸即可復談

台灣・阿扁：禁用大殺傷力武器兩岸應設軍事緩衝區

扁允裁軍 10 萬不搞核武

扁十點裁示，邱義仁：美國可做台、中調人，指美有問題可與

國台辦：台灣接受一中兩岸可復談

阿扁吁兩岸禁用大殺傷武器

多維追擊: 陳水扁倡議兩岸軍事緩衝區

Method	The detected words	N	R(%)
The proposed method	總統/一中/*十項/中國/台灣/台辦/兩岸/拒認/阿扁/武器/*扁倡/*扁提/美國/裁示/發表/善意/誠意/談話/*可復談/邱義仁/國民黨/陳水扁/*欺天下/九二共識	24	79.17
Yih-Jeng Lin and Ming-Shing Yu	總統/中國/台灣/兩岸/阿扁/武器/*扁倡/美國/發表/善意/誠意/*籲兩岸/*議兩岸/*十裁示/邱義仁/國民黨/國台辦/陳水扁/*欺天下/*十點裁示/*十項裁示	21	66.67

表4-7　　實驗結果比較(2)

深圳對<時差七小時>及相關問題進行調查處理
深圳對《時差七小時》及相關問題進行調查處理
深圳調查《時差 7 小時》事件李意珍家屬退出公司
深圳調查《時差 7 小時》事件
深圳市對"妞妞事件"及相關問題的調查處理意見
只有漂亮話的"妞妞事件"處理意見
《時差七小時》主角李倩妮之父做檢討
高官迫童看電影市委查辦
中國網民斥深圳市委為李意珍開脫
四重角色:25 歲妞妞的“非凡”人生
深圳通報處理意見責令妞妞父親管好家屬
深公布《時差七小時》調查
涉包庇下屬經濟犯罪李意珍事件燒到黃麗滿?
深圳通報處理妞妞事件
深處理《時差七小時》及相關問題
「妞妞」父親 6 日回應媒體監督
深圳市委處理李意珍違紀
《時差七小時》事件簿
深圳市委副書記李意珍。（網上圖片）
深圳市委處理妞妞事件(圖)
深圳調查《時差七小時》事件
涉嫌“促銷”女兒電影及家屬擁巨款深圳市委副書記獲“從輕發落”
深圳處理副書記之女主演電影事件要求廉潔自律
李意珍指控互聯網造謠絕地反擊？

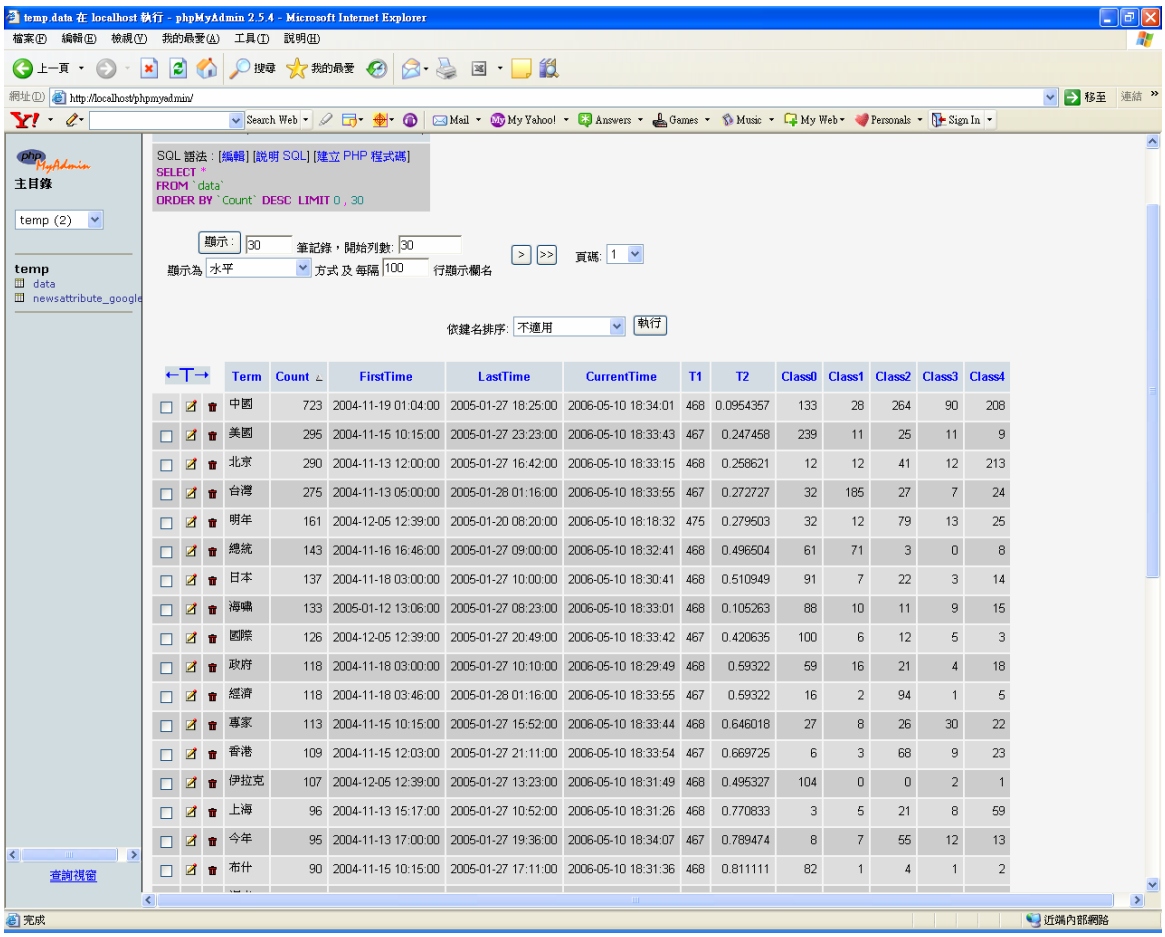
中國媒體勸深圳市委副書記李意珍早下台 亞洲時報：妞鈕事件燒到黃麗滿? 多維追擊：中國網民斥深圳市委為李意珍開脫			
Method	The detected words	N	R(%)
The proposed method	中國/父親/市委/妞妞/事件/家屬/深圳/處理/媒體/電影/調查/李意珍/副書記	13	100
Yih-Jeng Lin and Ming-Shing Yu	父親/妞妞/家屬/處理/媒體/電影/李意珍/深圳對/處理意見	9	77.78

表4-8　　實驗結果字詞數量與正確率之比較

Training news	Number of titles	The proposed method		Yih-Jeng Lin and Ming-Shing Yu	
		N	R(%)	N	R(%)
News(1)	30	24	79.17	21	66.67
News(2)	28	13	100	9	77.78
News(3)	33	25	84	16	56.25
News(4)	44	17	94.12	15	53.33
News(5)	18	5	100	5	60
News(6)	4	2	100	1	0
News(7)	40	17	76.47	13	61.54
News(8)	18	11	45.45	4	25
News(9)	5	6	33.33	3	33.33
News(10)	14	7	71.43	5	20
Average performance		12.7	78.397	9.2	45.39

4.2 中文新詞詞庫

本實驗收集了 79 天(2004/11/9~2005/1/28) 總共 8310 個新聞事件包含 168699 個新聞標題，平均每個新聞事件大約有 20 個標題。從這 8310 個新聞事件中一共擷取出 14463 個的中文字詞，其中有 5257 個新詞曾經在不同的新聞事件中出現，重複出現次數最高的字詞為'中國'總共出現在 723 個新聞事件中。如圖 4-1 所示為所擷取出來的新詞使用 mysql 資料庫來典藏。除了紀錄了所擷取出來的新詞以外，如表 4-9 所示還記錄了每一字詞在每一類別出現的次數，並隨時更新字詞的T_1與T_2值。藉由分析每一字詞的T_1與T_2值將可以把一些錯誤或過時的字詞濾除，並提高新聞類專業詞庫的正確率。



Term	Count	FirstTime	LastTime	CurrentTime	T1	T2	Class0	Class1	Class2	Class3	Class4
中國	723	2004-11-19 01:04:00	2005-01-27 18:25:00	2006-05-10 18:34:01	468	0.0954357	133	28	264	90	208
美國	295	2004-11-15 10:15:00	2005-01-27 23:23:00	2006-05-10 18:33:43	467	0.247458	239	11	25	11	9
北京	290	2004-11-13 12:00:00	2005-01-27 16:42:00	2006-05-10 18:33:15	468	0.258621	12	12	41	12	213
台灣	275	2004-11-13 05:00:00	2005-01-26 01:16:00	2006-05-10 18:33:55	467	0.272727	32	185	27	7	24
明年	161	2004-12-05 12:39:00	2005-01-20 08:20:00	2006-05-10 18:18:32	475	0.279503	32	12	79	13	25
總統	143	2004-11-16 16:46:00	2005-01-27 09:00:00	2006-05-10 18:32:41	468	0.496504	61	71	3	0	8
日本	137	2004-11-18 03:00:00	2005-01-27 10:00:00	2006-05-10 18:30:41	468	0.510949	91	7	22	3	14
得嘯	133	2005-01-12 13:06:00	2005-01-27 08:23:00	2006-05-10 18:33:01	468	0.105263	88	10	11	9	15
國際	126	2004-12-05 12:39:00	2005-01-27 20:49:00	2006-05-10 18:33:42	467	0.420635	100	6	12	5	3
政府	118	2004-11-18 03:00:00	2005-01-27 10:10:00	2006-05-10 18:29:49	468	0.59322	59	16	21	4	18
經濟	118	2004-11-18 03:46:00	2005-01-28 01:16:00	2006-05-10 18:33:55	467	0.59322	16	2	94	1	5
專家	113	2004-11-15 10:15:00	2005-01-27 15:52:00	2006-05-10 18:33:44	468	0.646018	27	8	26	30	22
香港	109	2004-11-15 12:03:00	2005-01-27 21:11:00	2006-05-10 18:33:54	467	0.669725	6	3	68	9	23
伊拉克	107	2004-12-05 12:39:00	2005-01-27 13:23:00	2006-05-10 18:31:49	468	0.495327	104	0	0	2	1
上海	96	2004-11-13 15:17:00	2005-01-27 10:52:00	2006-05-10 18:31:26	468	0.770833	3	5	21	8	59
今年	95	2004-11-13 17:00:00	2005-01-27 19:36:00	2006-05-10 18:34:07	467	0.789474	8	7	55	12	13
布什	90	2004-11-15 10:15:00	2005-01-27 17:11:00	2006-05-10 18:31:36	468	0.811111	82	1	4	1	2

圖4-1　　新聞類專業詞庫

表4-9　新聞類專業詞庫之欄位說明

表格名稱	data			
放置資料	新聞相關資料			
欄位名稱	欄位設定	是否允許 NULL	備註	說明
Word	varchar(8)	No		字詞
count	int(11)	No		字詞總共出現的文件數
firsttime	datetime	No		第一次出現的日期
latesttime	datetime	No		最後一次出現的日期
currenttime	datetime	No		現在時間
T1	int(11)	No	天數	(Latesttime - firsttime)/count
T2	float	No		Currenttime-Latesttime
Class0	int(11)	No	字詞出現在該類別次數	0. 國際
Class1	int(11)	No	、	1. 台灣
Class2	int(11)	No	、	2. 財經
Class3	int(11)	No	、	3. 科技
Class4	int(11)	No	、	4. 兩岸

4.3 字詞淘汰

本文 3.3 節中提出利用字詞出現的時間點來分類字詞的可信度，並藉此淘汰過時的字詞。透過分析 8310 個新聞事件所擷取出來的 14463 個字詞，其中 T2=0 的字詞共有 9330 個，如表 3-17 所示。當 T1 或 T2 值很小時，字詞的可信度至少都會被歸類為中等，所造成的結果就是 T1 或 T2 值只要很小，就不會被淘汰。但事實上 T1 值很小代表字詞在短時間之內曾經出現過，因此 T1 值很小的字詞理所當然不應該被淘汰。但 T2 值很小代表字詞平均生命週期很短，乍看之下似乎這類字詞也一定不該被淘汰，實際上 T2 值很小的字詞有一部份是曾經在某一段時間出現一次或多次，但過了那段時間之後卻不再出現，這一類的字詞假如一直不被淘汰將是不合理的。因此對於 T2=0 的字詞在本章節被另外提出來分析。經由分析這 9330 個字詞，其分布狀況如表 4-10。假如將 T1 值分為短、中、長，則當 T1 值介於 50~78 的字詞將被淘汰掉。

表4-10　　字詞(T2=0)分布狀況

字詞數量		T1		
		Short(0~22)	Moderate(23~49)	Long(50~78)
T2=0	0	3037	3151	3142

表4-11　　T1 與 T2 分布範圍

Linguistic variable	T1	T2
Short	0~4	0~5
Moderate	5~9	5~10
Long	20~78	10~38

表4-12　　字詞可信度分布與相對應之次數

字詞數目		T1		
		S	M	L
T2	S	592	389	754
	M	528	536	582
	L	603	741	408

剩下的 5133 個字詞將被套入本文 3.3 提出的方法來淘汰可信度最低的字詞。表 4-11 為根據實驗數據所訂定之 T1，T2 的短、中、長分布。在套入本文提出的方法之後，將可求得表 4-12 的結果。將表 4-12 與表 3-17 互相對應則有 408 個字詞是屬於可信度非常低的字詞，因此在這 5133 個字詞中將淘汰 408 個字詞。

4.4　新詞分類

實驗所採用的新聞資料總共有五大類分別為國際、台灣、財經、科技與兩岸。理論上經由國際類新聞所擷取出來的字詞應該被歸類為國際類，而台灣類的新聞所擷取出來的新詞則該被歸類為台灣類字詞，以此類推。但經由觀察每一類別出現次數前二十名的字詞(表 4-13)，可以發現有些字詞例如”中國”在每一個類別都出現過，而字詞”台灣”也在不同的類別出現過，這一類的字詞似乎無法直接經由字詞的來源來分類。

表4-13　　各類別常見字詞

新聞類別	出現次數前二十名之字詞
國際	1 美國 2 中國 3 伊拉克 4 國際 5 日本 6 海嘯 7 布什 8 布希 9 英國 10 大選 11 印尼 12 總統 13 政府 14 襲擊 15 美軍 16 人死 17 舉行 18 印度 19 死亡 20 發生
台灣	1 台灣 2 立委 3 民進黨 4 總統 5 國民黨 6 陳水扁 7 選舉 8 連戰 9 政治 10 兩岸 11 國親 12 馬英九 13 民黨 14 台聯 15 黨團 16 中國 17 李登輝 18 台北 19 北市 20 游揆
財經	1 中國 2 經濟 3 明年 4 香港 5 銀行 6 美元 7 今年 8 億元 9 公司 10 市場 11 億美元 12 北京 13 央行 14 投資 15 企業 16 去年 17 外資 18 台股 19 增長 20 發展
科技	1 中國 2 我國 3 病毒 4 專家 5 發現 6 微軟 7 網路 8 手機 9 電腦 10 衛星 11 學生 12 科技 13 明年 14 教育 15 研究 16 科學家 17 北京 18 全球 19 今年 20 英特爾
兩岸	1 北京 2 中國 3 上海 4 廣東 5 事故 6 發生 7 兩岸 8 河南 9 死亡 10 中共 11 胡錦濤 12 浙江 13 煤礦 14 人死 15 明年 16 台灣 17 重慶 18 四川 19 香港 20 爆炸

表 4-14 為詞庫中出現次數前十名的字詞與其在各個新聞類別出現的次數。可以看出字詞”中國”出現在財經類與兩岸類的次數相對較多，不同的字詞出現在各類別的比重皆不相同，或許也可以經由這個方式來對字詞做分類，但每一個領域的新聞資料量並不相同,因此如果單單依照表 4-14 每一字詞出現在各領域的次數來分類的話，可能不是很客觀，因為也許國際類的新聞資料來源比較多，每一個字詞出現在國際類新聞的次數也會相對的拉高，因此必須利用另一個有效的方法來對字詞做分類。

表4-14　　常見字詞出現在各類別次數

出現次數		新聞類別					total
		國際	台灣	財經	科技	兩岸	
常見新聞字詞	中國	133	28	264	90	208	723
	美國	239	11	25	11	9	295
	北京	12	12	41	12	213	290
	台灣	32	185	27	7	24	275
	明年	32	12	79	13	25	161
	總統	61	71	3	0	8	143
	日本	91	7	22	3	14	137
	海嘯	88	10	11	9	15	133
	國際	100	6	12	5	3	126
	政府	59	16	21	4	18	118

表4-15　　正規化之後的次數

	國際	台灣	財經	科技	兩岸
中國	9	8	10	10	9
美國	10	4	6	6.66	3
北京	1	6	8	7.77	10
台灣	3	10	7	4.44	7
明年	3	6	9	8.88	8
總統	5	9	1	0	2
日本	7	2	5	1.11	4
海嘯	6	3	2	5.55	5
國際	8	1	3	3.33	1
政府	4	7	4	2.22	6

Hahn-Ming Lee、Chih-Ming Chen[16]等人提出一有效方法來解決這類問題。其提出的方法可以將同一類別不同字詞的出現次數正規化，表4-15為經由正規化之後得到的次數，其中字詞在每一類別出現的次數皆被正規化為0至10次，因此可以看出字詞”中國”經過正規化之後在每一個類別的出現次數皆很高，因此可以判定”中國”並不特別屬於某一類別，其他字詞例如”政府”原本在表4-14看起來應該是屬於國際類的字詞，但經由正規化之後反而應該比較可以被歸類為台灣類的字詞。

4.5 實驗討論

4.5.1 新詞擷取

由於部份網路新聞媒體對於同一新聞事件使用了完全相同的標題，因此實驗中使用的新聞標題已經事先過濾掉完全一樣的新聞標題，避免影響實驗的正確性。根據實驗結果，可以看出本文提出的方法無論是新詞的擷取數目或新詞的正確率皆比 Yih-Jeng Lin and Ming-Shing Yu 提出的方法還要好。新詞的數目與正確率能夠比 Yih-Jeng Lin and Ming-Shing Yu 提出的方法好，主要應歸功於新詞自動擷取方法中的步驟一與步驟四，由於語言使用習慣的關係，常常會出現經由兩個或多個字詞所組成的字詞例如片語，這類多字詞由於使用習慣的關係其詞頻通常很高，因此容易被當成是單一正確的字詞而不是片語，而本文提出的方法，依靠每個新聞標題的訂定者對於語言使用習慣的不同，將這類片語拆解成正確的字詞，並利用正確的字詞有效的過濾掉大部份詞頻較高的片語。由於標題數的不同，因此造成實驗結果中使用不同的新聞事件其新詞的正確率有很大的落差，標題數過少確實會影響新詞的數目與正確率，這是很合理的結果。因此當系統運作時必須避免套入標題數過少或過多的新聞事件，以確保新詞的正確率。

當使用本文提出的方法時，藉由改變參數R_n與 M 所產生出來的中文字詞，其數量與正確率將受影響，當兩個參數R_n與 M 皆設的很高時，正確率將會非常的高，但新詞的數量將銳減，反之如果兩個參數R_n與 M 皆調的很低時，新詞的數目提高但卻犧牲了正確率。在本文實驗中為求客觀比較，因此將參數設成R_n=1，R_n=1 是最低的門檻值因此實驗中所擷取出來的新詞是最多但正確率也是最差的，藉由調高參數R_n的值將有效的提高系統的正確率。由於每一新聞事件所產生的標題數量不盡相同，熱門的新聞動輒上百個新聞標題，而比較冷門的新聞往往只有十來個標題。因此參數R_n與 M 應當隨著標題數而有所調整，如何設定合理的參數值將是一個值得探討的題目。

4.5.2 字詞淘汰

由於中文字詞的特性，出現頻率較高的辭彙並不全然都是正確的字詞，有可能是片語或者根本就是一錯誤的字詞，因為新聞標題的訂定者必須用最簡短的字數把新聞事件描述出來，往往使用錯誤的文法，但當這些錯誤的字詞出現的頻率很高時，仍舊會因為詞頻很高而被當成正確的新詞擷取出來。而有許多字詞只是一時流行或只在某一時段被使用，或許過了一兩年即將消失並不再出現，也或許每隔好幾年才出現一次，這類的字詞或許是正確的，但因為使用頻率太低，並不應該永遠收藏在詞庫裡並佔據詞庫空間。

考慮時間的因素並用來淘汰錯誤與不合時宜的字詞，也能避免過多無用的字詞佔據詞庫的空間，也避免因為詞庫的過大而影響中文斷詞系統的效率。新聞資料包含新聞發生的時間點，透過分析字詞出現的時間點有助於建立不同層級的詞

庫，可以將熱門的字詞與冷門的字詞加以分類。將能更有效率的應用在中文語言處理。為了驗證本文的淘汰詞庫中過時的字詞的功能，實驗將所擷取出來的 14463 個字詞透過分析其 T1、T2 分佈的狀況，訂定出模糊推論的參數值，並成功的將每一個字詞其可信度分為五個部份，其結果顯示出此方法確實可行，附錄一為被歸類為可信度相當低的字詞，由於實驗只使用大約兩個半月新聞事件來做分析，所以可以看出附錄一中那些被歸類為可信度很低的字詞，其實字詞的正確率還是很高，那是由於所使用的新聞事件只有兩個半月的份量，因此被歸類為可信度很低的字詞其 T1 與 T2 實際上並沒有很大，相信隨著時間的增加，此方法所產生出來的結果將能更加合理。未來的研究將朝訂定方法中較佳的模糊推論參數值為發展。也或許還有更佳的方法可以利用時間這個因素來提升詞庫的正確率，這些都將是未來值得研究的課題。

4.5.3 字詞分類

新聞資料包含許多領域例如政治、藝術、運動等等，Google new 提供的新聞服務將各個領域的新聞分類，有助於建立不同領域的專業詞庫。本文提出的方法所建立的新聞類專業詞庫，除了是屬於新聞類相關性較高的詞庫，其中每一個字詞在經由 Hahn-Ming Lee、Chih-Ming Chen 等人提出的正規化方法處理之後將能做更詳細的分類。現今熱門的網路搜尋引擎追求更精準的搜尋結果，因此建立不同領域的專業詞庫將也能提升搜尋引擎的精確度，例如當處理國際類新聞時，加強國際類字詞的權重，一定能達到更好的效果。同樣的因為實際套入實驗的新聞資料並不夠多，所以實驗結果可能還無法讓人信服，但隨著時間的增加所統計的結果將會更多且也更客觀，真正的效能將隨著資料量的增加而慢慢顯現。

4.5.4 舊有詞庫的應用

舊有詞庫的應用必須選定 entropy 值來判定是否應當把字詞加入新聞類專業詞庫，entropy 值的訂定必須經由大量實驗來決定合理的數值。也許有更好的訂定方法有待未來的研究者研究。藉由利用舊有詞庫來擷取在單一新聞事件中詞頻較低的字詞，對於建立新聞類專業詞庫將有很大的幫助，一個正確並且充足的新聞類專業詞庫，將能提昇自然語言處理的正確性，尤其是詞庫式斷詞系統，必須依賴詞庫的品質與數量才能正確的處理新聞資料。本文建立的新聞類專業詞庫利用舊有詞庫彌補不足，將能成為一更完整的詞庫。