以語文特徵為基之中學閱讀測驗短文分級

Using Linguistic Features to Classify Texts for Reading Comprehension Tests at the High School Levels

黃昭憲 郭韋狄 李嘉玲 蔡家琦 劉昭麟
Chao-Shainn Huang Wei-Ti Kuo Chia-Ling Li Chia-Chi Tsai Chao-Lin Liu
國立政治大學資訊科學系
National Chengchi University, Taipei, Taiwan
{97753007, 94703041, 96703032, 99753006, chaolin}@nccu.edu.tw

摘要

短文閱讀是中階語文學習者的重要功課，閱讀測驗也是語文測驗中的重要項目。掌握文本的分級技術，是電腦輔助閱讀測驗選題和電腦輔助閱讀的重要基礎。雖然閱讀測驗的文本部分，並不能全然決定整體試題難易度，但是文本部分的分級，仍為一個相關的基石研究。本研究以國內高級中學程度的真實閱讀測驗文本為研究語料，考慮詞彙、句子表面特徵與句法相關訊息等特徵，搭配不同的機器學習技術進行分類工作。我們分析了不同類別文字資訊對於分類正確性的貢獻度，目前最高可達 53.6%的分類正確性。

Abstract

We investigate the issue of classifying short essays based their linguistic issues, for English at the high school levels. A good selection of appropriate essays is crucial for the language learners and for the reading comprehension tests, which is an important type of tests for language competence examinations. Although the text alone does not allow us to judge the difficulty of reading comprehension tests, the capability to identify the levels of high school students for whom the texts were used in the reading comprehension can be an important step toward computer assisted selection of reading comprehension test items. We employed word-level statistics, sentence-level statistics, and syntactic-level information of the text, and applied several machine learning techniques for this text classification problem. Experimental results show that, with the best performing combination of features and learning method, we achieved 53.6% in accuracy.

關鍵字：電腦輔助語文教學、可讀性分級、文件分類、閱讀測驗文本分析

1. 緒論

依照所包含的內容將文字資料加以分級是一個有相當歷史的研究，早在西元 1948 年就有學者研究文章可讀性的論述[21]。這一方向的研究有許多相關的應用，對於語文學習特別有其意義。將讀文適當地分類，可以做為以電腦輔助閱讀的基礎；類似的技術，也可以作為電腦輔助短文評等的基礎。

Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Pages 98-112, Puli, Nantou,Taiwan, September 2010.

廣泛的應用也意味著這一研究方向牽涉到許多研究領域，語言學、認知心理學、教育學與資訊科學家對於這一領域都有所貢獻。從應用面直接接觸的使用者來說，可以考慮語文教育的受教者程度，是母語使用者？還是非母語使用者？一個好的研究工作或者系統開發當然要兼顧上述所應用之目的、相關研究背景和真正使用者的特性來考慮。

早期的研究，基於當時技術與資源的限制，常只考慮文字資料中的詞彙難度、句子數目和句長資訊[19,22]。這樣的作法當然不能滿足實際的需求，就連學理上都有可議之處[17]。國內的學者，考慮比較複雜的詞彙資訊，引入文字的上位詞(hypernymy)與下位詞(hyponymy)資訊來輔助文章可讀性的判讀[24]。除了提昇文字相關資訊的深度之外，更考慮到文字表層之外的資訊，例如文章結構、語意訊息甚至認知機制，如此一來才能比較符合眾人的期待[20]。

本研究以國內高級中學的英文閱讀測驗作為研究對象，依據閱讀測驗文本中的詞彙與語句層次的資訊，來猜測閱讀測驗文本的測驗對象。目前，限於資料來源，測驗的對象只有高級中學的一年級上學期（以下簡稱高一上），高一下、高二上和高二下四個等級。文字資料總共為 845 篇。我們嘗試幾種機器學習(machine learning)為基礎的分類方法，目前最好的正確性，其 F1 measure [30]僅僅達到 53.6%。

不論就規模或成果而言，本研究的現況顯然距離實用還需要一定的努力。然而，目前的實驗數據支持了大多數學者的看法，他們認為只依靠文字表面資訊並不足以正確地將文字可讀性進行分類[17]。此外，由於所使用的訓練語料包含了閱讀文本與測驗題兩大部分，但在本研究中因為技術的缺憾，我們對測驗題的問題深度並沒有進一步的分析，只截取閱讀本文來進行可讀性分級的討論，這是我們未來最可以進行改良的一個方向。我們整體的研究流程，著重於國內學生的真實試題分析，並且分析不同文字、句長、和句法相關深層特徵項目，對於閱讀測驗短文的分級效果應該是最主要的貢獻。

我們在第二節報告所使用的語料來源，說明我們如何對於這一些語料進行前處理。在第三節說明我們使用了哪一些詞彙層次的資訊和抽取該類資訊的方法。在第四節中描述如何抽取相關的句法層次資訊。在第五節中我們報告相關的實驗結果，最後在第六節進行總結。

2. 系統設計

從所要分類的文本中抽取相關的特徵項目是短文分類的首要工作，我們利用圖 1 表示整體的處理程序，詳細的細節於後續節次提供。原始文章經過前處理動作後，同時往兩個方向進行後續工作。一方面透過 Stanford Parser [29]（為行文簡便，除了圖表標題之外，以下將簡稱為**剖析器**）建構出該篇文章所有句子的結構樹，同時從樹中得到句子深度、句法結構的特徵向量；另一方面，依序透過 Stanford POS Tagger 及 Stanford Stemming [28] 得到文章中所有單字的原形，接著再到各字表，CMU 字典[27]和譯典通線上辭典（以下稱為 Dr.eye）[10]統計出各特徵值。我們將在第 0 節中描述語料來源和第 2.2 節中說明所做的前處理工作。

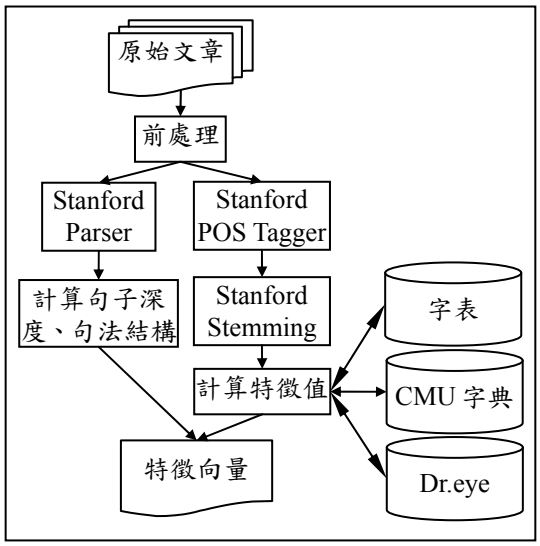

圖 1 特徵項目的抽取流程

2.1 研究背景與語料來源

我們的訓練語料取自於 96 學年度三民版高中英文試題光碟[†]，內部有三個版本由高一上至高二下的四冊語料共十二組資料，如表 1 所示。其中三民陳版本為民國 95 年台大陳凌霞教授主編；三民謝版本為謝國平先生主編；新三民版本為 96 年三民陳的更新版[1]。

表 1 語料來源和其檔案個數

	新三民	三民陳	三民謝	該冊總和	中文提示總數
第一冊	117	47	36	200	142
第二冊	127	64	36	227	45
第三冊	127	48	36	211	151
第四冊	126	45	36	207	14
總和	497	204	144	845	352

由於在高中閱讀測驗題型，常會有一些少用，或是特殊名詞的中文提示，例如：人名、地名或醫學領域的專有名詞等，都可能加上中文字來幫助閱讀者減少閱讀障礙。在前處理時我們將這些中文提示當作一個特徵來記錄，每一個提示當作一筆資訊記錄，接著將此中文字刪除，即可得到沒有中文干擾的全英文語料；統計數據如表 1 最後一行所示，在上學期的試題中，相對於下學期，擁有較多的中文提示總數。我們推測每學年上冊閱讀測驗的目標，偏重學生對於文意的了解，故文章中會有較多的中文提示輔助閱讀。而下學期閱讀測驗的目標，則偏重學生對基礎單字的掌握程度，故給予較少的中文提示。

2.2 前處理

由於人們對於文章的難度常會因為篇幅的大小、單字量的多寡、單字本身字義數目的情況，還有句子的長度而有一定的直覺。大多數我們認為，當句子的長度過於冗長的時候，我們會對該文章的理解能力有所下降，也會較片面的認定，相對於句子長度較短的文章，長篇文章應屬於偏難的程度。另外當文章中出現少見文法的比例高時，我們合理的懷疑該文法是比較困難的，較容易使讀者產生一定的閱讀負擔。

在完成前處理之後，可利用剖析器得到文章中的總句數、各句結構樹深度和文法資訊。利用 Stanford POS Tagger 和 Stanford Stemming 得到包含標點符號的總單字數、不含標點符號的總單字數、標點符號的總個數、平均每句含標點符號的單字數、平均每句不含標點符號的單字數和平均每句的標點符號個數，共六個特徵資訊。我們期望從這些資訊中，可找到適用於中學閱讀測驗短文分類的特徵。

3. 以單字為基礎的特徵值

在英語學習方面，依 97 年國民中小學九年一貫課程綱要的第 5-2-1 條中提到，國中畢業生應能熟習課綱中所標示之 1200 個基本單字，並能應用於日常生活的溝通中[13]。由此可知，單字本身帶有相當程度的分類要素。我們在第 3.1 節中說明使用的字表和其分級模式，第 3.2 節中介紹如何擷取和利用電子字典中的音標資訊，第 3.3 節中說明整合剖析器和 Dr.eye 詞性標記的方法。

[†] http://dcool.com.tw/webexam/newlogin.aspx 類似此類之出題系統

3.1 依據字表進行單字分級

根據大學入學考試中心製訂之字彙表，高中生常用英文字彙共約 7000 字，而在此我們利用不同機構所制定的單字分級字表，來替我們統計出該篇文章中，有多少比例的單字是屬於較難的級別。以此概念為基礎，我們利用國立師範大學舉辦的全國單字大賽所提供的字表（以下稱之為**師大字表**）[12]、全民英檢字表[5]和大考中心字表[4]，如表 2 所示，每個字表各自有不同的級別分布。在做查詢單字級別的同時，我們會特別記錄文章中在字表裡查詢不到的單字個數，該類的單字即等同於 Dale-Chall 所提到的「難字」概念[19]。

表 2 三個字表和其級別分布

字表	級別分布	該級別單字數	該字表總單字數
師大字表	國小 3、4 年級	498	6041
	國小 5 年級	250	
	國小 6 年級	250	
	國中 1 年級	350	
	國中 2 年級	350	
	國中 3 年級	407	
	高中 1 年級	936	
	高中 2 年級	1500	
	高中 3 年級	1500	
全民英檢字表	初級	2184	7853
	中級	2560	
	中高級	3109	
大考中心字表	第 1 級	1775	8976
	第 2 級	1490	
	第 3 級	1472	
	第 4 級	1350	
	第 5 級	1543	
	第 6 級	1346	

從表 2 中，師大字表的級別分布較多，國小的單字量較少，到了高中階段則大幅增加。根據全民英檢的分級標準[7]，可知全民英檢初級相當於國中畢業者，全民英檢中級相當於高中職畢業者，全民英檢中高級相當於大學非英語主修畢業者。此點我們可從表 2 中的單字分布得到印證。

我們利用 Stanford POS Tagger 將每個英文句子中的個別單字標記詞性，如圖 2 左半部分所示。接著再配合 Stanford Stemming 將個別單字做原詞還原，如圖 2 右半所示。透過以上兩步驟，我們可得到文章中所有單字的原形。這些原形的單字即可到指定字表中查詢它的級別。這麼做的目的是因為字表裡的單字為原形，如此處理才能得到最正確的級別字統計資訊。

```
I/PRP          --> I/PRP
liked/VBD      --> like/VBD
playing/VBG    --> play/VBG
basketball/NN  -->basketball/NN
when/WRB       --> when/WRB
I/PRP          --> I/PRP
was/VBD        --> be/VBD
young/JJ       --> young/JJ
./.            --> ./.
```

圖 2 Stanford POS Tagger 和 Stanford Stemming 的範例

最後我們依序將訓練語料與各個字表進行查表的動作。以師大字表為例，我們會記錄該篇文章的單字，對應師大字表中不同單字級別的頻率分布，並將無法在字表中查詢到的單字（或許意味著這是更困難的單字）額外進行計次的動作，我們共可得到 9 個級別的單字頻率與 1 個在字表各級單字中查詢不到的單字數量，總共 10 個特徵值。全民英檢字表和大考中心字表也依上述步驟分別得到 4 個特徵和 7 個特徵。利用上述方法，我們統計出該篇文章在各字表裡的單字級別分布，當高級別的字出現在文章中次數較多時，相對表示他的內容應該是較難以閱讀的。若分級字表夠精準，這種概念可以大幅地提升我們的分類效能。

3.2 以音節數為基礎的特徵值

一般英文的初學者對於多音節的字比較恐懼，在口語對話上是一大壓力，在閱讀上也會造成一定的阻礙，故我們認為單字裡的母音數是一個值得注意的特徵。而長度越長的單字也會給學習者較大的負擔，所以我們一併記錄母音加子音的個數視為單字的長度，當作我們實驗中的特徵值之一。

```
UNIVERSITY
Y  UW2  N  AH0  V  ER1  S  AH0  T  IY0
```

圖 3 UNIVERSITY 在 CMU 字典中的記錄

我們利用 CMU 字典來得到各個單字的音標資訊。圖 3 是單字 UNIVERSITY 在 CMU 字典的標記方式，第一列是英文字，第二列為音標符號。音標符號後面如果有數字 0、1 或 2 的話，則是標記該母音是否為重音，0、1、2 分別代表非重音、主要重音和次重音，而未標記數字的部分即為子音。

我們查詢每個單字的音標資訊，計算出非重音、主要重音和次重音的標記共有多少，即可知該字的音節數目。以 UNIVERSITY 為例，其共有五個音節：UW2、AH0、ER1、AH0 和 IY0，故我們可知 UNIVERSITY 的母音數為 5，母音加子音數為 10。依照上述方法，我們對文章內所有的單字做查詢，可以產生單字 0 個至 7 個音節頻率分布、單字本身母音加上子音 0 個至 16 個的頻率分布，共 25 個特徵值。擁有這些特徵之後，當一篇文章的頻率分布較集中在音節數較高的部分，我們則認為該篇文章的難度較高，而這類的文章，應該較容易出現在高年級的閱讀測驗當中。

3.3 統計字義數目

在閱讀文章時，一個單字的背後常不只有一種詞意，當該單字的詞意越多時，越容易造成讀者閱讀上的障礙，因此當讀者在閱讀外語的閱讀測驗時，文章內部的外語字詞對應到中文字義，大部分都並非是一對一的形式。當一個外語字詞可以被翻譯成多種中文字義時，較容易對讀者造成閱讀障礙。所以我們統計一篇文章中，有多少單字會有這類的情況產生。

開始統計字義數目之前，我們必須先做一些特別的前處理，由於剖析器的詞性標記是以 Penn Treebank [25] 為主，而我們想透過 Dr.eye 來查詢各個單字會有幾種意義。但是 Dr.eye 中的詞性標記並不是以 Penn Treebank 為基礎，所以我們的首要工作便是將剖析器和 Dr.eye 的詞性進行整合。

根據大考中心的高中英文詞類分類表[3]，我們依名詞、動詞、形容詞、副詞、介系詞、代名詞、連接詞和冠詞八大詞性來做分類，如表 3 所示。從剖析器得到的 NN（單數名詞或不可數名詞）、NNS（複數名詞）、NNP（專有名詞，單數）和 NNPS（專有名詞，複數）及 Dr.eye 裡的 n.（名詞）皆屬於八大詞性中的名詞類別。

表 3 Stanford Parser 和 Dr.eye 依八大詞性做整合

八大詞性	Stanford Parser	Dr.eye
名詞	NN, NNS, NNP, NNPS	n.
動詞	MD, VB, VBD, VBG, VBN, VBP, VBZ	vt., vi.
形容詞	CD, JJ, JJR, JJRS	a.
副詞	EX, RB, RBR, RBS, RP, WRB	ad.
介系詞	IN, TO	prep.
代名詞	DT, PRP, PRP$, WDT, WP, WP$, WRB	pron.
連接詞	CC, IN	conj.
冠詞	DT	art.

在表 3 中可發現剖析器的部分詞性標記會分到兩個以上的詞性。這是因為剖析器和 Dr.eye 的詞性為一對多的關係。如 of 在 Dr.eye 中為介系詞，so 在 Dr.eye 中為連接詞，但兩者在剖析器裡都標記為 IN。因此當我們從剖析器抓到某單字的詞性為 IN 時，我們採取的作法是先到 Dr.eye 裡確認該單字是否存在該詞性，如果確實存在才將其詞性做整合，這樣的作法可降低大多數的誤差，但如果某單字在 Dr.eye 中的詞性同時有介系詞和連接詞時，為了能讓實

“divide”
vt. (及物動詞 transitive verb)
1. 分,劃分[(+into/from)]
2. 分發;分享[(+between/among/with)]
3. 分配[(+between)]
4. 【數】除[(+by/into)]
5. 使對立,分裂
6. 使分開,使隔開[(+from)]

vi. (不及物動詞 intransitive verb)
1. 分開
2. 分裂;意見分歧

n. (名詞 noun)
1. 分歧,不和[S][(+between)]
2. 分水嶺[C]

圖 4 Dr.eye 裡的 “divide”

驗繼續進行，現階段我們將其認定連結詞。

接著我們透過查詢 Dr.eye 所記錄的字義數目，由 0 個至 43 個的頻率分布，當文章內所統計的字義數目越多時，表示該單字在文章中所欲表達的「意義」，會產生一定程度的歧異認知，也連帶的使讀者對於該篇文章的內容產生閱讀障礙。最後我們記錄文章中八大詞性的頻率分布，猜測其可能是影響文章分類的因素之一。我們實作的方法如下，當在剖析器標記“divide”為 VB（動詞，原形）時，即可知其對應到八大詞性的動詞，接著再對應到 Dr.eye 裡的 vt.（及物動詞）和 vi.（不及物動詞）。如圖 4 所示，我們可知“divide”為 vt.時有 6 種意義，為 vi.時有 2 種意義，所以“divide”為動詞時共有 8 種意義。

4. 以句子為基礎的特徵值

文章的難度，除了從字詞的層級去看，構成句子的特性也是很大的因素之一。本節介紹以句子為基礎所使用到的相關技術及討論。

文章中的句子，我們可以利用剖析工具來看見它的結構樹。從結構樹中可以知道，一句句子可以切割成許多較小的部分，這些部分之間互相連結，並且結構樹會由上往下拓展，當該結構樹越深時，意味著該句是由許多複雜的句法所構成，並且透過特定的句法規則，來表達出句子真正的意涵。我們利用剖析器把文章中的每句句子都建立出各自的結構樹，透過該樹狀結構，我們可以算出該樹的深度。

以“I liked playing basketball when I was young.”為例，剖析器會為該句子產生一個樹狀的語法結構，如圖 5 所示，我們將樹根定義為第零層，樹根的子樹為第一層，越往下層數字越大。為了讓讀者容易觀看，我們將該樹圖形化，如圖 6 所示。從圖 6 中我們可看出深度最深的葉子節點為在第九層的 young，所以此句子的結構樹深度為 9。

將所欲分類的文章中所有句子依照上述程序執行，並記錄結構樹深度 0 至最深深度的數量（目前的語料最深之深度為 31）；我們將這些數據稱為結構樹深度之頻率分布。

結構樹除了葉子節點之外的內部節點分支，都可視為一組句法結構。我們將標點符號、葉子節點拿掉後，蒐集整個樹狀結構的句法結構。以圖 6 為例，第二層的內部節點「VP」擁有「VBD」、「NP」兩個子節點，那麼就表示在這邊使用到「VP → VBD NP」的句法結構來進行剖析。

我們從旋元佑文法[8]、基礎英文 1200 句[9]、國民中學學習資源網[11] 和教育部委託宜蘭縣建置語文學習領域國中教科書補充資料題庫[14]的評量題庫及資源手冊，蒐集了共七千多句的英文語料。從這些英文語料擷取其句法結構，並進行次數的統計。依次數由高至低排序，稱之為句法頻率。我們一共分析出 985 條文法規則，有趣的是 80%的文法規則出現頻率都偏低，頻率 100 次以上的只有 62 條，1000 次以上的更只有 8 條

```
(ROOT
  (S
    (NP (PRP I))
    (VP (VBD liked)
      (NP
        (NP (VBG playing) (NN basket-
ball))
        (SBAR
          (WHADVP (WRB when))
          (S
            (NP (PRP I))
            (VP (VBD was)
              (ADJP (JJ young)))))))
    (. .)))
```

圖 5 Stanford Parser 的樹狀結構範例

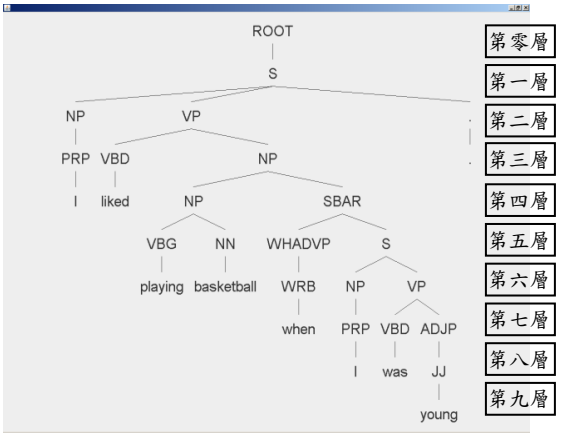

圖 6 將 Stanford Parser 的樹狀結構圖形

而已。

為了記錄短文中所運用到的句法頻率高低，我們必須將這一些句法規則的頻率由低到高切割成數個區間，以便於將此以特徵項目量化。由於上述的極度不均勻分布，等距離裝箱法(equal interval binning)[30] 是不適用的。

因此，我們將同頻率的句法歸為一類，共有 127 種類別的規則，利用等頻率裝箱法(equal frequency binning)[30]，以 21 為間距切割成六等份（第六等份，會有 22 個類別數），前 1/6 的規則，用[0, 1/6]表示，是最常見的規則，(2/6, 3/6]是次常見的規則，…，(5/6, 1]則視為最少見的規則。（目前僅嘗試將類別依照頻率切割為六區域，這並不見得是最佳的選擇，我們也還沒有嘗試切割為其他數量區域的效果。）

有了上述之規範，即可對文章內部進行進一步的剖析，假定在一個句子的結構樹中，我們發現了六條句法規則，其中有兩條是屬於(2/6, 3/6]，有一條是屬於(5/6, 1]，和三條無法查詢到的句法，我們把這些資訊記錄成一個向量：{0,0,2,0,0,1,3}。將文章中的每一個句子都執行上述之處理，即可得到多條向量特徵。

這些特徵向量隱含著該句的難度分數，當句子的句法向量數據集中在較少見的規則時，則表示該句子使用了較大量的罕見句法，這些罕見的句法極有可能是屬於較難或較複雜的文法規則，也意味著該句子的難度偏難，而使讀者較難了解該句所想表達的意思。我們也將此當作文章分類實驗的特徵之一，期盼能對我們的分類效果有所幫助。

除了前面用到的特徵值外，我們進一步蒐集文法特徵值。在此我們以人工的方式進行分析，從不同的詞性標記整理出他們的文法特徵，依序為SBAR(關係代名詞的子句)、SBAQR(一般疑問句)、同時出現SBAQR和SBAR(間接的疑問句)、現在簡單式、過去簡單式、未來簡單式、現在進行式、過去進行式、現在簡單式的被動、過去簡單式的被動、未來簡單式的被動、現在進行式的被動、過去進行式的被動、現在完成式的被動、過去完成式的被動、假設語氣等文法句型的出現次數，我們推測各文法句型的頻率分布，也是影響文章難易度的關鍵要素之一。

5. 實驗過程和分析

我們利用第 2.1 節中所介紹過的語料，以 WEKA [30]來協助文章分類之實驗。實驗的語料已在表 1 中說明。由於高中教育的試題，閱讀測驗考題包含了短文文本與測驗問題兩個部分，受試學生必須閱讀文章，經過思考才能回答問題。由於分析試題部分的難度相當高，因此在本實驗中，我們暫時不考慮閱讀試題對於難易度的影響，只分析文本在語言學上的特徵。將閱讀測驗進行文章與試題的分解，從中只擷取文章的部分來當作整個實驗的訓練語料。

我們也將手邊的語料文章送至SMOG [26]，以獲得文章的可讀性分數。SMOG的計算方法如公式(1)所示。

$$1.043 \times \sqrt{m \times \left(\frac{30}{n}\right)} + 3.1291 \qquad (1)$$

其中m代表三個音節以上的單字個數（number of polysyllables），n代表文章中的句子總數（number of sentences）。

表 4 SMOG 分數平均

	高一上	高一下	高二上	高二下
SMOG 最小值	6.59	7.22	6.75	7.3
SMOG 最大值	17.11	19.88	22.75	22.09
SMOG 平均分數	10.889	11.822	12.554	12.757

從公式中我們可以看出來，SMOG只利用文章中多音節的單字個數和句子總數當作特徵參數。得到這

兩個特徵參數後，SMOG系統即對此篇文章的難易評估出一個分數，而四學期的分數間距和平均如表4所示，我們也將SMOG分數從6至15，以0.5為一個區間，統計出各分數的篇數比例分布，如圖7所示。

從表4和圖7，可以看見試題文章本身的難度，會隨著學期而逐漸變難。表4顯示各學期SMOG平均分數，隨著年級提高。但是圖7顯示在四個學期間，各篇文章SMOG分數的分布曲線，並沒有很明顯的分野，這也代表著我們正在面對一個極有挑戰性的分類問題。

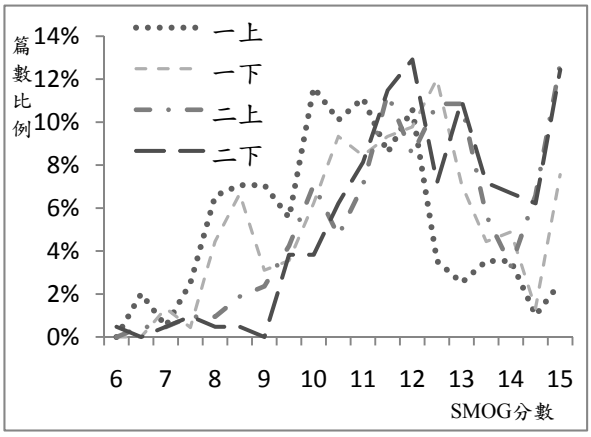

圖 7 SMOG 分數在各冊的篇數分布

5.1 利用單字、句子和分級字表各別進行分類

首先我們對文章內每個字句本身做分析，將所有的特徵值分為三個群組，A 群是以單字為基礎的特徵值，依序為該文章中單字的 8 個音節頻率分布、17 個單字本身母音加上子音的頻率分布，和透過 Dr.eye 所查詢到的 44 個字義數目頻率分布（共 69 個特徵）。B 群則是以句子為基礎的特徵值，依序為該文章之平均句長（包含計算標點符號和不計算標點符號）、平均標點符號數量、結構樹的平均深度和 32 個結構樹深度之頻率分布（共 36 個特徵）。C 群則是以字表為基礎，在本實驗中，共用到了三種不一樣的字表，分別為師大字表 10 個特徵值（Ca 群）、全民英檢字表 4 個特徵值（Cb 群）和大考中心字表 7 個特徵值（Cc 群），藉由這些字表，查出該篇文章中單字在字表各級別的頻率分布。

為了避免文章字數長短，間接影響分類效果，我們將得到的統計數據執行正規化處理後，才進行文章的分類實驗。首先，將 A、B 群中頻率分布的特徵值，音節數量、母音加子音數量和字義數目的數量，除以該文章總字數；且將句子深度頻率分布的次數，除以該文章總句數，以形成該文章句子結構樹不同深度的比重。最後將 C 群所產生的字表各級別頻率分布，也除以該文章總字數。

我們將最原始的五組資料（A、B、Ca、Cb、Cc），分別採用 WEKA 內建分類器，J48 決策樹(decision tree)分類器、LMT 決策樹(decision tree)分類器、ANN 類神經網路分類器（在此我們設定兩個參數，訓練次數(epoch)=500、學習速率(learning rate)=0.3，且只執行一次 ANN 演算法。）和 Ridor 規則(rule)分類器四種分類法，採用 10-fold cross validation 的方式來進行實驗，並且記錄依各類篇數乘上該 *F*-measure，將四類已進行乘積後的數據加總，再除上四類的總篇數，所產生的權重平均 *F*-measure 數據。*F*-measure 公式如下。

$$\text{F-measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \qquad (2)$$

從表 5 中我們可以看見，*F*-measure 的分數落在 0.248 至 0.353 區間，其中最高分是採用全民英檢字表（Cb），且搭配 LMT 演算法進行分類實驗；而平均最低分的組別，是以大考中心字表（Cc）來進行文章分類，而分類效果如此不好的原因，我們推測大考中心所發佈的字表，可能是以高三應屆考生為出發點所建立的，而我們的實驗目標則是高中一年級和二年級，才會

表 5 A、B 和 C 各群實驗數據

組別	各實驗組合之 *F*-measure 分數				
	A	B	Ca	Cb	Cc
J48	0.297	0.27	0.297	0.335	**0.248**
LMT	0.334	0.318	0.3	**0.353**	0.264
ANN	0.278	0.291	0.34	0.323	0.268
Ridor	0.293	0.291	0.307	0.304	0.261

表 6 A、B 和 C 群混合組合實驗數據

	各實驗組合之 *F*-measure 分數									
	A+B	A+Ca	A+Cb	A+Cc	B+Ca	B+Cb	B+Cc	Ca+Cb	Ca+Cc	Cb+Cc
J48	0.281	0.266	0.275	0.293	0.293	0.299	0.274	0.3	0.306	0.312
LMT	0.335	0.345	0.337	0.346	0.344	0.341	0.3	0.348	0.338	0.364
ANN	0.283	0.348	0.33	0.318	0.315	0.319	0.303	0.346	0.347	0.324
Ridor	0.288	0.291	0.323	0.312	0.322	0.356	0.253	0.319	0.341	0.346
	A+B+Ca	A+B+Cb	A+B+Cc	A+Ca+Cb	A+Ca+Cc	A+Cb+Cc	B+Ca+Cb	B+Ca+Cc	B+Cb+Cc	Ca+Cb+Cc
J48	**0.261**	0.303	0.301	0.291	0.307	0.325	0.303	0.295	0.304	0.286
LMT	0.331	0.327	0.35	0.341	0.321	0.35	**0.381**	0.349	0.366	0.358
ANN	0.359	0.309	0.328	0.313	0.33	0.323	0.319	0.352	0.314	0.357
Ridor	0.321	0.329	0.305	0.33	0.31	0.323	0.307	0.299	0.302	0.326

	A+B+Ca+Cb	A+B+Ca+Cc	A+B+Cb+Cc	A+Ca+Cb+Cc	B+Ca+Cb+Cc	A+B+Ca+Cb+Cc
J48	0.305	0.305	0.313	0.302	0.317	0.33
LMT	0.329	0.353	0.369	0.341	0.373	0.358
ANN	0.335	0.335	0.362	0.338	0.333	0.324
Ridor	0.349	0.314	0.329	0.334	0.354	0.362

造成我們分類效果較差。但如果排除 Cc 這個實驗組別，我們發現利用 Ca 和 Cb，以字表為基礎來進行分類實驗，整體效果優於利用 A 群或 B 群的特徵進行分類。或許因為高中閱讀文章的測驗方式，是依賴單字本身的難易程度來進行選題的依據。

5.2混合單字、句子和字表特徵

接下來我們利用混合組合的方式，來進行單字、句子和字表間互相的搭配，得到表 6 實驗結果，由表 6 中可知，這次的分數區間提升為 0.261 至 0.381，並且我們發現數據中最高分的組別是以句子為基礎（B）、配合師大字表（Ca）和全民英檢字表（Cb），採用 LMT 演算法來進行分類，我們得到 *F*-measure 的分數為 0.381，與先前利用單類資訊（表 5）來進行比較，發現 *F*-measure 的分數區間，最小值原本是 0.248 提升至 0.261，最大值從 0.353 提升至 0.381，分類效果進步約 3%左右。我們得知，利用句子本身的結構特徵，搭配查詢字表所得到單字本身的難度分布，對於文章分類有些許的進步。

另一方面當我們綁定相同的字表，與 A 群、B 群互相混合來進行分類時，我們發現帶有 B 群的組別分數，在大部分的情況都贏過帶有 A 群的組別。相比之下，以句子的特徵值對於文章分類的效果，優於以單字的特徵值來進行文章分類。我們由表 5 和表 6 的數據推測，當字表（C 群）的特徵值出現時，幾乎可以取代單字（A 群）的特徵值，此時以單字為基礎（單字的音節頻率分布、單字本身母音加上子音數量的頻率分布，和透過 Dr.eye 所查詢到的字義數目的頻率分布），相較於字表（C 群）對於文章分類的重要性則下降許多。

5.3 由句法頻率分布，協助文章分類

將第 4 節所提的句法頻率分布向量進行正規化動作，對該篇文章，所有的句法向量進行加總，再除以該文章的句法總數目，如一篇文章中有兩條句法向量{0,0,2,0,0,1,3}和{1,1,0,5,5,0,4}，先將兩條向量加總得{1,1,2,5,5,1,7}，再除以總句法數目 22，得到{0.045,0.045,0.091,0.227,0.227,0.045,0.318}（四捨五入至小數點後第三位），此時我們會

表 7 結合句法頻率的實驗數據

	各實驗組合之 *F*-measure 分數					
與 D 群合併	A+B+Ca	A+B+Cb	A+B+Cc	B+Ca+Cb	B+Ca+Cb+Cc	Cb+Cc
J48	**0.251**	0.294	0.294	0.309	0.318	0.309
LMT	0.346	0.342	0.325	0.343	**0.357**	0.345
ANN	0.327	0.339	0.306	0.308	0.351	0.327
Ridor	0.32	0.302	0.302	0.346	0.346	0.326

表 8 混合總數類別的實驗數據

結合 D 群、中文提示字數量和總數類別之特徵值	各實驗組合之 *F*-measure 分數					
	A+B+Ca	A+B+Cb	A+B+Cc	B+Ca+Cb	B+Ca+Cb+Cc	Cb+Cc
J48	0.338	0.314	0.349	0.333	0.331	0.347
LMT	0.412	0.405	0.374	0.423	**0.425**	0.412
ANN	0.37	0.353	0.345	0.352	0.402	0.363
Ridor	0.337	0.36	**0.312**	0.353	0.377	0.341

得到該篇文章七個關於句法頻率的特徵值（在此稱為 D 群）。利用句法在文章中所佔的比重進行實驗，嘗試了解句法是否有助於文章分級的效果。我們選出在第 5.2 節的實驗中，平均分數在前三名的組合（分別為表 6 中的 B+Ca+Cb、B+Ca+Cb+Cc 和 Cb+Cc），再選出由單字和句子的特徵各配合上一個字表（A+B+Ca、A+B+Cb 和 A+B+Cc），共六組實驗組合加入 D 群來進行文章分類，得到表 7 的實驗數據。

從表7可以看見，若將句法頻率的特徵值，加進上面六組實驗中，四種演算法所得的分數區間落在0.251至0.357之間，造成平均分數有些微的下降。這邊可能是因為我們所倚靠的句法來源,是國中程度七千句英文句子,而我們的分類目標是高中的閱讀文章;或許在國中所使用到的句法，相對於高中所使用到的句法過於簡單，或是在國中句法頻率較低，屬於比較難的句法，到高中反而應該是屬於頻率較高，比較常見的句法，導致我們對於句法頻率的統計產生誤差，而使分類效果無法得到提升。

5.4 文章長度的影響

此小節利用文章的總句數、總深度、含標點符號的總單字數、不含標點符號的總單字數和標點符號的總個數（在此稱之為「總數類別」），和各篇文章中文提示字的數量，共六項特徵值，混合至第 0 節的實驗組合來進行實驗，得到表 8 的實驗數據。

這次我們得到的分數落在 0.312 至 0.425 區間,整體的分類效果獲得了上升的趨勢，甚至可以得到 42.5%的正確率。可能在四個學期間，閱讀文章的篇幅長短的確是有極大的影響，再加上中文提示字的因素，迫使整篇文章的難度有所變動。同時在實驗的過程中，影響當初利用字表所產生各級別的頻率分布，進而對於文章分類的效果上升許多。

5.5 觀察八大詞性對文章的分類效果

最後我們將第 0 節實驗中，分數最佳的組合(B+Ca+Cb+Cc+D+總數類別+中文提示數),加入八大詞性分布共八個特徵值，再次進行實驗，我們得到表 9。從表 9 中可以發現，八大詞性分布的特徵值對於 ANN 和 Ridor 兩個演算法甚至有退步的現象。在此我們推論的原因有二種，第一、在第 3.3 節中提到八大詞性的整合，本實驗是利用剖析器和 Dr.eye 兩者的詞性標記來做整合，但是在整合的過程中，會有部分的標記同時出現在兩類的詞性中，這或許是影響分類效能的因素之一。第二、原先我們認為，文章的詞性分布可能會影響整個文章的難易程度，例如當文章中動詞的比例較名詞的比例高時，這種文章對於讀者可能有較大的負擔，又或者當詞性分布較平均時，整篇文章的寫作結構對於讀者負擔較低，但是實驗的結果並沒有反映出我們的期待。

表 9 結合八大詞性頻率分布的實驗數據

B+Ca+Cb+Cc+D+總數類別＋中文提示數之 *F*-measure 分數		
	原始數據	再加上八大詞性分布
J48	0.331	0.349
LMT	0.425	0.425
ANN	0.402	0.346
Ridor	0.377	0.343

表 10 中文和非中文語料篇數

	含中文字	未含中文字
第一冊	74	124
第二冊	26	199
第三冊	34	148
第四冊	12	198
總和	176	669

5.6 含中文語料、未含中文語料個別實驗

表 11 含中文字的語料重作表 5 實驗

組別	各實驗組之 F-measure 分數				
	A	B	Ca	Cb	Cc
J48	0.423	0.364	0.472	0.428	0.382
LMT	0.435	0.404	0.494	0.466	0.363
ANN	0.429	0.396	0.467	**0.52**	0.396
Ridor	**0.353**	0.365	0.364	0.424	0.385

除了表 1 的 839 篇語料外，我們又另外蒐集了 6 篇語料，共 845 篇語料。其中含中文字的語料有 176 篇，未含中文字的語料有 669 篇，如表 10 所示。我們主要想觀察，中文提示字的數量是否會影響文章分類的效果，在此區分為含中文字和未含中文字兩種語料，分別重新操作表 5 和表 6 實驗。含中文字的語料重作表 5 的實驗結果如表 11 所示。將表 11 和表 5 相比，分類效果大幅進步，*F*-measure 的分數區間落在 0.353 至 0.52 之間，其中的最高分是採用全民英檢字表(Cb)，且搭配 ANN 演算法進行分類實驗。

表 12 含中文字的語料重作表 6 實驗

	各實驗組合之 *F*-measure 分數									
	A+B	A+Ca	A+Cb	A+Cc	B+Ca	B+Cb	B+Cc	Ca+Cb	Ca+Cc	Cb+Cc
J48	0.342	0.335	0.406	0.364	0.427	0.349	0.343	0.437	0.443	0.406
LMT	0.4	0.479	0.432	0.458	0.487	0.507	0.402	0.49	0.493	0.493
ANN	0.404	0.406	0.424	0.471	0.457	0.389	0.422	0.475	0.406	0.439
Ridor	0.395	0.375	0.413	0.364	0.381	0.424	0.366	0.462	0.401	0.456
	A+B+Ca	A+B+Cb	A+B+Cc	A+Ca+Cb	A+Ca+Cc	A+Cb+Cc	B+Ca+Cb	B+Ca+Cc	B+Cb+Cc	Ca+Cb+Cc
J48	0.345	0.342	0.342	0.412	0.348	0.394	0.353	0.441	0.391	0.429
LMT	0.46	0.477	0.391	0.449	0.489	0.457	0.485	0.438	0.507	**0.536**
ANN	0.42	0.4	0.364	0.444	0.444	0.44	0.421	0.457	0.436	0.402
Ridor	0.397	0.435	0.355	0.374	0.377	0.45	0.431	0.438	0.369	0.457

	A+B+Ca+Cb	A+B+Ca+Cc	A+B+Cb+Cc	A+Ca+Cb+Cc	B+Ca+Cb+Cc	A+B+Ca+Cb+Cc
J48	**0.33**	0.369	0.353	0.419	0.348	0.371
LMT	0.471	0.47	0.49	0.46	0.465	0.458
ANN	0.448	0.422	0.442	0.48	0.382	0.453
Ridor	0.412	0.387	0.473	0.35	0.424	0.416

含中文字的語料重作表6實驗，結果如表12含中文字的語料重作表6實驗所示。數據最高分是採用三個字表，且搭配LMT演算法來作分類。將表12含中文字的語料重作表6實驗和表11相比，發現*F*-measure的分數區間，最小值由0.353下降到0.33，最大值由0.52上升到0.536，變化幅度不大。但將表12含中文字的語料重作表6實驗和表6相比時，*F*-measure的分數區間，最小值由0.261上升到0.33，最大值由0.381上升到0.536，有明顯的進步。

未含中文字的語料重作表5的實驗結果如表13所示。數據最高分的組別是採用師大字表(Ca)和ANN演算法來進行分類。將表13和表5相比，還是有小幅的進步。但將表13和表11相比，F-measure的分數區間，最小值由0.353下降到0.254，最大值由0.52下降到0.417，退步了不少。未含中文字的語料重作表6的實驗結果如表14所示。數據最高分的組別是以句子為基礎(B)、配合全民英檢字表(Cb)和大考中心字表(Cc)，採用LMT演算法來進行分類，其F-measure的分數為0.414，分數不盡理想。將表14和表6相比，F-measure分數仍有小幅的進步。將表14和表12含中文字的語料重作表6實驗相比，F-measure分數則退步不少。

表 13 未含中文字的語料重作表 5 實驗

組別	各實驗組之 F-measure 分數				
	A	B	Ca	Cb	Cc
J48	0.297	**0.254**	0.344	0.315	0.297
LMT	0.356	0.295	0.378	0.349	0.308
ANN	0.358	0.286	**0.417**	0.372	0.28
Ridor	0.324	0.3	0.351	0.378	0.276

我們比較表5、表6、表11、表12含中文字的語料重作表6實驗、表13、表14後，發現將含中文字的語料另外處理會使分類效果更好，F-measure分數達到了0.536。而未

表 14 未含中文字的語料重作表 6 實驗

	各實驗組合之 *F*-measure 分數									
	A+B	A+Ca	A+Cb	A+Cc	B+Ca	B+Cb	B+Cc	Ca+Cb	Ca+Cc	Cb+Cc
J48	0.265	0.32	0.32	0.28	0.35	0.341	**0.256**	0.386	0.345	0.382
LMT	0.345	0.381	0.401	0.375	0.387	0.366	0.324	0.378	0.4	0.392
ANN	0.327	0.413	0.368	0.339	0.323	0.337	0.267	0.403	0.383	0.366
Ridor	0.334	0.374	0.353	0.323	0.356	0.341	0.317	0.384	0.377	0.381
	A+B+Ca	A+B+Cb	A+B+Cc	A+Ca+Cb	A+Ca+Cc	A+Cb+Cc	B+Ca+Cb	B+Ca+Cc	B+Cb+Cc	Ca+Cb+Cc
J48	0.322	0.327	0.307	0.33	0.316	0.372	0.334	0.356	0.347	0.359
LMT	0.384	0.356	0.335	0.391	0.393	0.393	0.399	0.382	**0.414**	0.385
ANN	0.365	0.352	0.362	0.393	0.382	0.343	0.347	0.349	0.35	0.404
Ridor	0.349	0.36	0.382	0.34	0.335	0.368	0.391	0.333	0.33	0.367

	A+B+Ca+Cb	A+B+Ca+Cc	A+B+Cb+Cc	A+Ca+Cb+Cc	B+Ca+Cb+Cc	A+B+Ca+Cb+Cc
J48	0.352	0.294	0.348	0.38	0.31	0.345
LMT	0.386	0.388	0.369	0.381	0.396	0.4
ANN	0.38	0.379	0.349	0.39	0.353	0.389
Ridor	0.352	0.346	0.367	0.344	0.375	0.334

含中文字語料的分類效果相對上較差了一些，但還是比原來表5、表6的實驗結果來得還要好。我們的結論是含中文字的語料和未含中文字的語料在內容上有一定的分野，這使得兩者個別作實驗的效果比原本混合實驗的效果來得還要好。

表 15 含中文字語料 A、B、C、E 群混合實驗

	各實驗組合之 F-measure 分數									
	E	A+E	B+E	Ca+E	Cb+E	Cc+E	A+B+E	A+Ca+E	A+Cb+E	A+Cc+E
J48	0.34	0.374	0.388	0.383	0.417	0.33	0.354	0.393	0.399	0.386
LMT	0.438	0.478	0.389	0.472	0.476	0.447	0.426	0.501	0.484	0.467
ANN	0.352	0.422	0.375	0.363	0.368	0.412	0.41	0.44	0.454	0.414
Ridor	0.39	0.426	0.373	0.328	0.385	0.298	0.385	0.403	0.386	0.377

	B+Ca+E	B+Cb+E	B+Cc+E	Ca+Cb+E	Ca+Cc+E	Cb+Cc+E	A+B+Ca+E
J48	0.321	0.348	0.345	0.386	0.39	0.428	0.321
LMT	0.45	0.529	0.385	0.461	0.457	0.484	0.459
ANN	0.376	0.4	0.363	0.38	0.418	0.435	0.408
Ridor	0.416	0.381	0.341	0.418	0.361	0.395	0.368

	A+B+Cb+E	A+B+Cc+E	A+Ca+Cb+E	A+Ca+Cc+E	A+Cb+Cc+E	B+Ca+Cb+E
J48	0.36	0.341	0.488	0.408	0.412	0.378
LMT	0.482	0.461	0.482	0.53	0.494	0.465
ANN	0.448	0.378	0.417	0.43	0.43	0.38
Ridor	0.415	0.373	0.423	0.384	0.393	0.384

	B+Ca+Cc+E	B+Cb+Cc+E	Ca+Cb+Cc+E	A+B+Ca+Cb+E	A+B+Ca+Cc+E
J48	0.346	0.383	0.388	0.385	0.306
LMT	0.452	0.522	0.496	0.443	0.482
ANN	0.417	0.427	0.417	0.427	0.415
Ridor	0.345	0.4	0.423	0.385	0.404

	A+B+Cb+Cc+E	A+Ca+Cb+Cc+E	B+Ca+Cb+Cc+E	A+B+Ca+Cb+Cc+E
J48	0.381	0.447	0.403	0.379
LMT	0.474	0.48	0.502	0.472
ANN	0.408	0.46	0.392	0.407
Ridor	0.435	0.404	0.375	0.427

表 16 不含中文字語料 A、B、C、E 群混合實驗

	各實驗組合之 F-measure 分數									
	E	A+E	B+E	Ca+E	Cb+E	Cc+E	A+B+E	A+Ca+E	A+Cb+E	A+Cc+E
J48	0.279	0.33	0.262	0.314	0.323	0.289	0.298	0.314	0.344	0.286
LMT	0.277	0.348	0.308	0.374	0.361	0.307	0.341	0.343	0.384	0.362
ANN	0.265	0.371	0.279	0.336	0.315	0.301	0.339	0.37	0.388	0.365
Ridor	0.27	0.325	0.289	0.345	0.351	0.295	0.299	0.326	0.36	0.36

	B+Ca+E	B+Cb+E	B+Cc+E	Ca+Cb+E	Ca+Cc+E	Cb+Cc+E	A+B+Ca+E
J48	0.307	0.328	0.29	0.356	0.326	0.329	0.262
LMT	0.366	0.371	0.326	0.395	0.378	0.362	0.353
ANN	0.348	0.327	0.299	0.346	0.348	0.343	0.375
Ridor	0.342	0.32	**0.254**	0.376	0.344	0.366	0.315

	A+B+Cb+E	A+B+Cc+E	A+Ca+Cb+E	A+Ca+Cc+E	A+Cb+Cc+E	B+Ca+Cb+E
J48	0.343	0.302	0.337	0.325	0.343	0.288
LMT	0.388	0.34	0.386	0.34	0.374	0.396
ANN	0.37	0.365	0.378	**0.402**	0.384	0.35
Ridor	0.35	0.317	0.387	0.326	0.345	0.353

	B+Ca+Cc+E	B+Cb+Cc+E	Ca+Cb+Cc+E	A+B+Ca+Cb+E	A+B+Ca+Cc+E
J48	0.275	0.329	0.35	0.341	0.292
LMT	0.348	0.37	0.377	0.378	0.35
ANN	0.374	0.307	0.383	0.374	0.386
Ridor	0.338	0.314	0.362	0.372	0.321

	A+B+Cb+Cc+E	A+Ca+Cb+Cc+E	B+Ca+Cb+Cc+E	A+B+Ca+Cb+Cc+E
J48	0.351	0.346	0.326	0.344
LMT	0.38	0.37	0.389	0.393
ANN	0.376	0.406	0.338	0.378
Ridor	0.349	0.368	0.395	0.358

5.7 混合文法特徵值的綜合實驗

最後我們把文法特徵值也當作一個要素，並將其定義為 E 群。除了在第 4 節中所提到共 16 個文法特徵值之外，我們也統計句子中出現 0 到 4 次 VBN 的頻率分布、0 到 6 次 VP 的頻率分布、0 到 3 次 MD 的頻率分布，共 16 個特徵值。利用 Stanford Parser 的片語標記，我們蒐集文章中出現 ADJP(形容詞片語)、ADVP(副詞片語)、CONJP(連接詞片語)的句子數，共 3 個特徵值。從上述可知，E 群總共包含了 35 個特徵值。

我們將E群特徵值加入，和A、B、C群做各種排列組合的分類實驗。中文字語料的實驗結果如表15所示。和表12含中文字的語料重作表6實驗相比，最低值由0.33(J48的A+B+Ca+Cb組合)下降到0.306(J48的A+B+Ca+Cc+E組合)；最高值由0.536(LMT的Ca+Cb+Cc組合)下降到0.53(LMT的A+Ca+Cc+E組合)，皆沒有明顯的變化。

而不含中文字語料的實驗結果如表16所示。和表12含中文字的語料重作表6實驗相比，最低值仍為0.254(J48的B組合和Ridor的B+Cc+E組合)；最高值由0.417 (LMT的Ca+Cb+Cc組合)下降到0.402(ANN的A+Ca+Cc+E組合)，皆沒有明顯的變化。加入E群特徵值後，雖然最低、最高值的組合變了，但實驗結果分數並沒有太大的變化。由於混合E類的實驗並未使分數驟降，我們相信文法特徵值佔有一定的重要性，未來將蒐集更多的文法特徵值做實驗，期待能產生更好的分類結果。

6. 結語

本實驗的目標是將高中一、二年級的閱讀文章，四個學期共 839 篇進行分類。整合所有的實驗結果，利用 36 個以句子為基礎的特徵值，並且結合三份分級字表所產生的

23 個特徵值,再搭配上 7 個句法頻率分布、文章內的 5 個總數類別特徵和 1 個中文提示字數特徵,綜合以上共 72 個特徵,且以 LMT 演算法來進行分類,可以從亂數隨機分成四類的 25%正確率提升至 42.5%正確率。進一步將語料分成含中文字、不含中文字兩部份個別操作實驗,分類可再提升到 53.6%正確率(中文字語料採用三個字表,且搭配 LMT 演算法)。由於分類效果並不是十分亮眼,在此我們將四個學期的語料,從中隨機各選出 25 篇進行商業可讀性分數的比較,除了在第 5 節中所提到的 SMOG 之外,我們還挑選了 Flesch-Kincaid Grade Level (FKGL)[22]和 Automated readability index (ARI)[16]兩個公式來當作我們的指標,將 100 篇依三個不同公式產生之分數進行比較,如圖 8 所示。從圖 8 中三條的分數走勢,都有著極強的共識,主要的原因是 SMOG 與 FKGL 都是利用單字數量和句子數量做為參數來進行計算。而 ARI 除了單字數量和句子數量以外,還多採用文章中字母數量當作參數來進行計算,但從圖 8 中可以發現,ARI 並沒有因為多了一個變數,而與其他兩個公式有太大的不同,整體趨勢大致上是相同的。

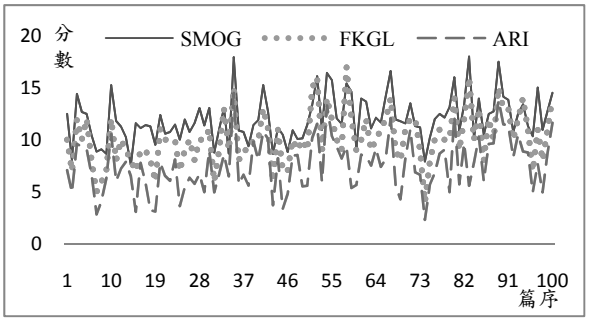

圖 8 三種公式之分數比較

本實驗除了商業公式常用的變數以外,還多了字表、句子結構樹深度和句法結構頻率、常見文法來當作特徵值,期盼能更準確的進行文章分類;但由於本實驗四學期的訓練語料並沒有很明顯的分野(參考圖 7 之分析),原因是在學習的歷程中,常會將低年級程度的試題放在高年級的試題中當作復習考題,以便學生在學習上能夠溫故知新。在未來,我們希望藉由更具辨別度的真實語料,來執行文章分類的工作,甚至進而實現文章分級評分系統的實作。

致謝

本計畫承蒙國科會研究計畫案 NSC-97-2221-004-007、NSC-98-2815-C-004-003-E 與 NSC-99-2221-004-007 補助,謹此致謝。我們感謝匿名評審的寶貴意見,雖然我們一時無法在有限的頁數之內回應所有意見(特別是關於相關研究的評比和增加更多的說明方面),但是仍然在未來工作中,努力進行評審所提出的建議工作。

參考文獻

1. 97 年國民中小學九年一貫課程綱要, http://www.edu.tw/eje/content.aspx?site_content_sn=15326 (最後造訪該網址時間 2010/8/14)
2. 三民學習網,http://www.grandeast.com.tw/englishsite/ (最後造訪該網址時間 2010/8/14)
3. 大考中心英文詞類分類表, http://www.ceec.edu.tw/Research/paper_doc/ce37/6.pdf (最後造訪該網址時間 2010/8/14)
4. 大考中心詞彙分級表, http://www.ceec.edu.tw/Research/paper_doc/ce37/5.pdf (最後造訪該網址時間 2010/8/14)
5. 全民英檢字表, http://www.taiwantestcentral.com/WordList/ (最後造訪該網址時間 2010/8/14)
6. 高照明,計算語言學在華語教學的應用,嘉義大學語言中心演講資料,2008。
7. 時代國際英日語中心, http://www.tilc.tw/test-gept.html (最後造訪該網址時間 2010/8/14)

8. 旋元佑文法，http://tw.myblog.yahoo.com/jw!GFGhGimWHxN4wRWXG1UDIL_XSA--/ (最後造訪該網址時間 2010/8/14)
9. 基礎英文 1200 句，http://hk.geocities.com/cnlyhhp/eng.htm (最後造訪該網址時間 2009/8/27)
10. 譯典通線上辭典 (Dr.eye)，http://www.dreye.com/tw/dict/dict.phtml (最後造訪該網址時間 2010/8/14)
11. 國民中學學習資源網， http://140.111.34.172/teacool/new_page_2.htm (最後造訪該網址時間 2010/8/14)
12. 國立師範大學全國單字大賽字表，http://vq.ie.ntnu.edu.tw/wr01.htm (最後造訪該網址時間 2009/8/27)
13. 教育部全球資訊網，http://www.edu.tw/ (最後造訪該網址時間 2009/8/27)
14. 教育部委託宜蘭縣發展九年一貫課程建置語文學習領域（英語）國中教科書補充資料暨題庫建置計畫，http://140.111.66.37/english/ (最後造訪該網址時間 2010/8/14)
15. 鄧守信，L2 Chinese as an autonomous discipline，*臺灣語言學學會通訊*，第二卷第四期，1-3，2000。
16. Automated readability index. http://en.wikipedia.org/wiki/Automated_Readability_Index (最後造訪該網址時間 2010/8/14)
17. A. Bailin and A. Grafstein, The linguistic assumptions underlying readability formulae: A critique, *Language and Communication*, 21(2), 285-301, 2001.
18. J. Burstein, M. Chodorow, and C. Leacock, Automated essay evaluation: the criterion online writing service, *Artificial Intelligence*, 2004.
19. J. Chall and E. Dale, *Readability Revisited: The new Dale-Chall Readability Formula*. Cambridge, Brookline Books, 1995.
20. S. A. Crossley, J. Greenfield, and D. S. McNamara, Assessing Text Readability Using Cognitively Based Indices, *TESOL Quarterly*, 42(3), 475-493, 2008.
21. R. Flesch, *A New Readability Yardstick*, *Journal of Applied Psychology*, 32(3), 221-233, 1948.
22. Flesch-Kincaid Grade Level. http://en.wikipedia.org/wiki/Flesch-Kincaid_Readability_Test (最後造訪該網址時間 2009/9/09)
23. J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, Derivation of New Readability Formulas （Automated Readability Index, Fog Count and Flesch Reading Ease Formula） for Navy Enlisted Personnel, *Technical Report Research Branch Report 8-75*, 1975.
24. S.-Y. Lin, C.-C. Su, Y.-D. Lai, L.-C. Yang, and S.-K. Hsieh, Assessing Text Readability Using Hierarchical Lexical Relations Retrieved From WordNet, *International Journal of Computational Linguistics and Chinese Language Processing*, 14(1), 45-84, 2009.
25. Penn Treebank II Tags. http://bulba.sdsu.edu/jeanette/thesis/PennTags.html (最後造訪該網址時間 2010/8/14)
26. Simple Measure of Goobledygook (SMOG). http://www.harrymclaughlin.com/SMOG.htm (最後造訪該網址時間 2010/8/14)
27. The CMU Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict (最後造訪該網址時間 2010/8/14)
28. Stanford Log-linear Part-of-Speech Tagger. http://nlp.stanford.edu/software/tagger.shtml (最後造訪該網址時間 2010/8/14)
29. Stanford Parser: http://nlp.stanford.edu/software/lex-parser.shtml (最後造訪該網址時間 2010/8/14)
30. I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2005.