國立交通大學

網路工程研究所

碩 士 論 文

非 監 督 式 中 文 寫 作 自 動 評 閱 系 統

An Unsupervised Chinese Automated Essay Scoring System

研 究 生：陳彥宇

指導教授：李嘉晃　教授

中 華 民 國　九 十 六　年 六 月

非 監 督 式 中 文 寫 作 自 動 評 閱 系 統
An Unsupervised Chinese Automated Essay Scoring System

研 究 生：陳彥宇　　Student：Yen-Yu Chen

指導教授：李嘉晃　　Advisor：Chia-Hoang Lee

國 立 交 通 大 學

網 路 工 程 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Network Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

June 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年六月

非監督式中文寫作自動評閱系統

學生：陳彥宇　　指導教授：李嘉晃 博士

國立交通大學網路工程研究所碩士班

摘　　要

寫作自動評閱技術在語文教育領域中是十分重要的輔助工具，然而目前的自動評閱系統均需要一定數量的同主題文章做為訓練資料，因而在使用上有其限制性。本論文提出一個基於文章相似度的中文寫作自動評分系統，此系統不需要同主題文章的人工評分資料，僅需一定數量的測試資料，便可藉文章間的相似度及其表面特徵資訊自動進行評分。另外，本論文亦提出一個演算法可以偵測不合題旨的文章，以避免評分的錯誤。實驗結果允許一級分誤差的正確率可達95%，完全命中的精確率可達52%；可以做為人工評分的參考依據。

An Unsupervised Chinese Automated Essay Scoring System

student：Yen-Yu Chen Advisor：Dr. Chia-Hoang Lee

Institute of Network Engineering
National Chiao Tung University

ABSTRACT

Automated essay scoring is an important research in education domain, but the cost of training is a problem. In this paper, we describe a Chinese automated essay scoring system based on similarity between test essays. This system could grade essays without graded training data. We also describe an algorithm detecting off-topic essays. The adjacent rate in the experiment is about 95%, and the exact rate is about 52%.

目錄

表目錄

圖目錄

一、 緒論

1.1 研究動機

寫作能力是語文教育中相當重要的一環，然而人工評閱所耗費的人力及時間成本十分可觀，文章自動評閱（Automated Essay Scoring, AES）是利用人工智慧技術讓電腦可以模仿人工批改作文的技術，自動評閱技術可以大幅降低評閱作文時的成本，因此在教育測驗、課程教學及心理計量等領域中可說是相當重要的工具。

自動評閱技術在英文寫作方面有著長久的發展，並已經實際應用在大型入學測驗及寫作教學上，如美國教育測驗服務社（Educational Testing Service, ETS）所主辦的 TOEFL、GMAT 均使用自動評閱系統做為閱卷時的輔助工具；其開發的寫作練習軟體 Criterion 可以讓學生練習特定題目的寫作並自動進行評閱。在中文作文的領域上，也已經有數個系統被提出（〔7〕〔8〕〔9〕〔10〕）。

然而，目前的中文自動評分系統皆需要 150~300 篇以上經人工批閱評定分數的同主題文章做為訓練資料，才能正確地進行自動評分；當我們希望系統評閱一個新的題目時，便需要蒐集一定數量該題目的訓練文章，由專業人員評閱這些訓練文章的成績後，才可建立起評分模型，之後電腦才能自動評閱此主題的文章。其訓練過程不僅耗費人力與時間成本，而且使用者必需有蒐集及評閱訓練資料的能力，因此在使用上有很大的限制性。

1.2 研究目的

本研究之目的在建立一個不需要訓練資料的中文自動評分系統。此系統不需要人工評閱過的訓練資料，僅需要一定數量的同主題文章，便可藉由文章的間接特徵資訊與文章之間的相似度自動評閱測試文章之成績，可以有效的降低自動評分系統在訓練階段的成本，亦不需要專業的寫作評閱人員參與。

1.3 論文架構

第一章為緒論，說明本研究的動機與目的；第二章為相關研究，將介紹現有的中文自動評分系統及相關研究資訊；第三章將說明本系統的架構和詳細的演算法流程；第四章為實驗過程與結果，以及與其他系統的比較結果；第五章將說明本論文的結論及未來展望。

二、 相關研究

自動評閱技術在國外發展甚早，1960 年代 Page 利用某些與文章成績有高度相關的統計特徵進行線性回歸而提出了 Project Essay Grader（PEG）系統，可說是自動評閱技術的先驅。由於 PEG 只使用文章的字數、逗號數、罕見詞數等間接特徵，這些特徵雖然與文章成績在統計上有相關性，卻無法代表文章的水準；而文章的直接特徵如內容、組織及文法等資訊並未被加入，雖然評分結果與人工評分的相關係數可以達到 0.78，但寫作者很容易掌握其統計特徵，進而寫出內容明顯不佳卻能被評為高分的文章。

在 PEG 之後，隨著自然語言處理及資料擷取等技術的進步，許多自動評分系統陸續被提出，如 Page〔4〕加入了句子完備性等，Attali 及 Burstein 〔2〕加入了語法正確性與主題符合度等直接特徵；Landauer 等〔3〕利用作品與詞彙的潛在語意關係（Latent Semantic Analysis, LSA）來取得文章間的語意關係，提出了以語意特徵為核心的系統 IEA。以上系統加入了直接特徵的資訊，使得評分結果更接近人工評分，且較難以欺騙。

在中文寫作自動評閱方面，目前提出的系統（〔7〕〔8〕〔9〕〔10〕）主要架構與英文系統相同，均以特徵擷取為基礎，再使用機器學習的方法整合特徵值的資訊以建立評分規則。

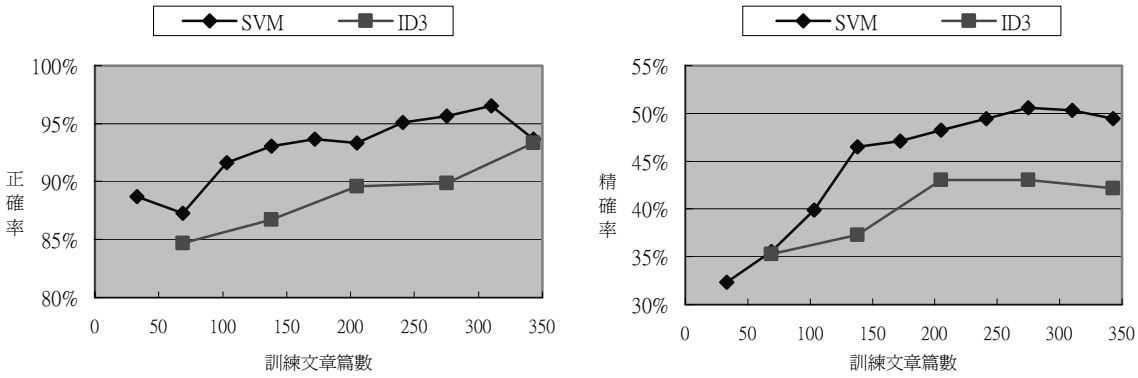
在特徵擷取方面，由於中文的語言特性，遭遇到許多處理英文時不會出現的困難，也造成英文的自動評閱技術無法直接套用在中文寫作上。在目前的系統中，張〔8〕使用了詞數、形容詞數、成語數等表面特徵及譬喻、排比兩項修辭方面的直接特徵；蔡〔10〕使用同主題的訓練文章找出鑑別義原數這項直接特徵；林〔7〕則在鑑別義原數外再加入了單字詞比率、字數、句號數、名詞數及平均段落字數等五項特徵。

在機器學習方面，嘗試過的模型包括貝氏學習機〔7〕、ID3 決策樹〔8〕以及支援向量機〔9〕，這些模型的共通點為必須輸入訓練文章的特徵值和人工評閱成績以建立評分規則，再用這些規則自動評閱測試文章的分數。

然而，這些系統都需要 150~300 篇以上同主題的訓練文章資訊才足以建立起評分規則，因此無法利用在缺乏訓練資料的狀況上。

圖一為 ID3 決策樹〔8〕及支援向量機（SVM）〔9〕在不同訓練文章篇數時，對同

一份包含 346 篇文章的測試資料的正確率及精確率。可見兩個系統在訓練文章不足 150 篇時效果均不盡理想，超過 150 篇後方漸趨穩定。



圖一　現有系統在不同訓練資料數下之正確率與精確率

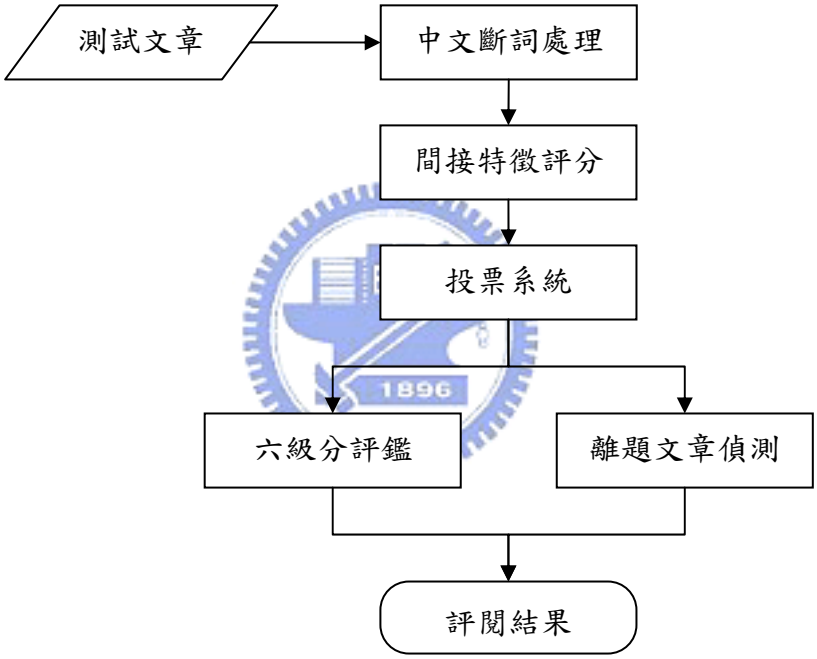
另外，以上的自動評分系統雖然有採用直接特徵為評分項目，但並沒有偵測文章是否合乎主題的機制，因此當不合題意的文章的表面特徵夠好時，便有出現評分錯誤的可能性，使得系統容易被有心人士所破解。

三、　系統設計

中文作文的評分標準，主要分為立意取材、結構組織、遣詞造句及錯別字與格式等項目。本系統僅探討文章取材方面的優劣程度，並未深入文章語意方面之表現。

3.1 節將描述整個系統的架構與流程，其後的四個小節將詳細介紹各個模組的內容。

3.1 系統架構



圖二　系統架構圖

當大量的測試資料進入系統時，系統首先對於每篇測試文章進行中文斷詞處理，將文章切割為詞語的串列；接著根據各篇文章的間接特徵給予一初期分數；以此間接特徵評分之結果為初始狀態，本系統使用一個投票演算法不斷修正評分結果，直到評分結果收斂到穩定狀態為止；在此階段的評分結束後，系統根據其結果建立一個相關詞集以偵測離題文章；最後由測試文章在投票演算法的評分結果，以及歷史資料的成績分配情形決定文章的六級分成績。

3.2 中文斷詞處理

中文斷詞是將連續的中文字轉換為詞的組合，詞為一個或多個中文字的組合，是句法和語意上的最小單位，斷詞在自然語言處理上是相當重要的工作。相對於歐美語系中，詞與詞之間有明確的分隔（空白及標點），中文斷詞處理是困難得多的工作，目前並沒有任何演算法可以達到百分之百的斷詞正確率。

本系統採用長詞優先法，即斷出剩餘字串中最長的有意義字串，如「下課十分鐘」一句，可斷成「[下][課][十][分][鐘]」、「[下課][十][分鐘]」等不同的詞組，而長詞優先法斷詞將將先斷出[下課]，然後再由剩餘的「十分鐘」字串中斷出[十分]，最終結果為「[下課][十分][鐘]」。

長詞優先法的正確率在已知的演算法中雖然並非是最高的，但效果已相當不錯，且本系統僅計算文章間共同出現的詞語，並不探討文章和詞語的語意，因此對斷詞錯誤的容忍度更高。

以前述「十分鐘」為例，雖然長詞優先法會造成斷詞上的錯誤，而使得文意上產生了不通順的狀況。但由共用詞語數的角度來看，若是兩篇文章均寫到「十分鐘」一句，則錯誤的「[十分][鐘]」和正確的「[十][分鐘]」得出的結果均為兩個共用詞語，並不會影響評分結果。

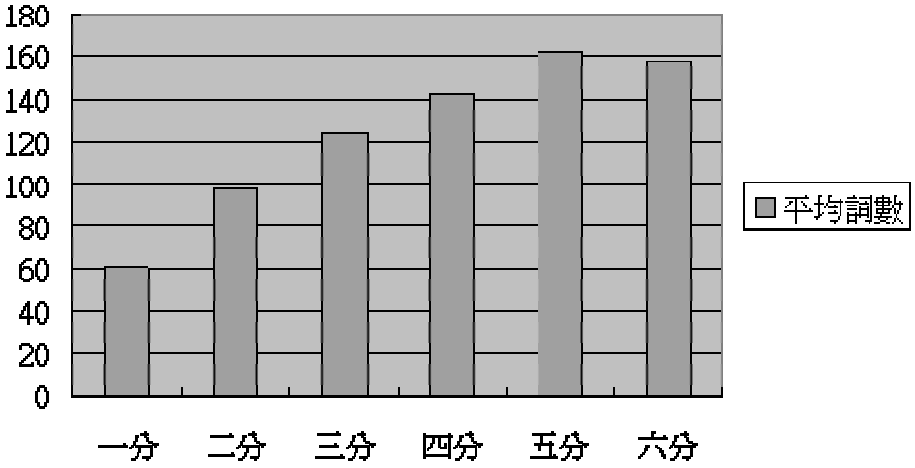
3.3 間接特徵評分

本系統的基本假設為與高分文章越相似的文章，越可能是高分文章，反之則越可能是低分文章。但在系統起始時，測試文章並沒有任何附加資訊，因此我們無從得知測試文章的好壞。為了建立投票演算法的初始狀態，我們需要給定每一篇測試文章一個初期分數；又因為欠缺測試文章的直接特徵資訊，只能根據其間接特徵資訊來給定此初期分數。

本系統選用「相異詞語數」這項表面特徵做為初期評分的依據，即文章所使用的詞語種類數；例如「我最喜歡跟我的朋友玩」，經斷詞後為「[我][最][喜歡][跟][我][的][朋友][玩]」共八個詞，其中[我]出現兩次，相異詞語數為七。若一篇文章的相異詞語數為

100 個，則該文章的初始分數即為 100，以此類推。

觀察人工評閱的結果，發現高分文章中使用的詞數明顯較低分文章為多，因此可以做為間接特徵評分的準則。



圖三　各級分文章之平均相異詞語數

在間接特徵評分的部份，選用不同的間接特徵為依據會產生不同的初始狀態。然而只要選用的間接特徵合理，初始狀態的些許差異在經過投票演算法後，對最終評分的影響會降低到可以忽略的程度。

3.4 投票系統

在間接特徵評分結束後，本系統使用其結果進行投票演算法。此演算法之精神為越相似的文章，分數應該越接近。

3.4.1 基本公式

$$S_{j,t} = \sum_{i \neq j} Sim_{i,j} * Z_{i,j,(t-1)} \qquad (1)$$

$$Z_{i,j,t} = \frac{\left(S_{i,t} - \sum_{k \neq j} S_{k,t} \Big/ (N-1) \right)}{\sigma_t} \qquad (2)$$

$S_{j,t}$：時間為 t 時，文章 j 的分數。

$Sim_{i,j}$：文章 i 與文章 j 的相似度。

$Z_{i,j,t}$：時間為 t 時，文章 i 對於文章 j 的 Z 分數（Z-Score）。

N：文章總數。

σ_t：時間為 t 時，所有文章分數之標準差。

我們將要計算分數的文章稱為目標文章，其他文章稱為參考文章。公式（1）可以分解為三個部份：

第一個部份Σ，將所有參考文章給予的分數加總，但不計算目標文章對本身的給分；第二個部份$Sim_{i,j}$，代表目標文章與參考文章的相似度，在本系統中以兩篇文章的共用詞語數代表；第三個部份為參考文章的Z-Score，代表參考文章的分數與平均值的差異程度。

若一篇參考文章在上個世代的分數高於平均越多，便會給予目標文章越高的正分，反之則給予負分；換言之，所有參考文章均意圖將目標文章吸引至自身的分數。除上標準差是為了讓每個世代的分數收斂在一個範圍中，避免分數隨著時間的增加無止盡地上升。

由於相似度的加權，使得相似度越高的參考文章對於目標文章的影響力越大，目標文章的分數因此往相似度較高的參考文章靠近。這便符合我們的基本精神：若一篇文章與其他高分文章越相似、與其他低分文章越不相似，則此篇文章為高分的機率越高；反之低分的機率越高。

當 t=0 時，系統為初始狀態，所有文章的分數設定為間接特徵評分的結果，亦即該文章的相異詞語數。隨著 t 不斷的增加，文章的分數根據公式（1）更新，當文章數量足夠時，最後評分結果會趨於一個穩定狀態。

舉例來說：若有四篇文章 a, b, c, d。

起始狀態為$S_{a,0}$=1，$S_{b,0}$=1，$S_{c,0}$=4，$S_{d,0}$=6。

文章相似度如下表：

	a	b	c	d
a	-	3	2	3
b	3	-	3	4
c	2	3	-	6
d	3	4	6	-

則在 t=1 時，各文章的分數將更新如下：

$S_{a,1} = Sim_{b,a}*Z_{b,a,0} + Sim_{c,a}*Z_{c,a,0} + Sim_{d,a}*Z_{d,a,0} = 3*(-1.26)+2*(0.16)+3*(1.10) = -0.16$

$S_{b,1} = Sim_{a,b}*Z_{a,b,0} + Sim_{c,b}*Z_{c,b,0} + Sim_{d,b}*Z_{d,b,0} = 3*(-1.26)+3*(0.16)+4*(1.10) = 1.10$

$S_{c,1} = Sim_{a,c}*Z_{a,c,0} + Sim_{b,c}*Z_{b,c,0} + Sim_{d,c}*Z_{d,c,0} = 2*(-0.79)+3*(-0.79)+6*(1.57) = 5.50$

$S_{d,1} = Sim_{a,d}*Z_{a,d,0} + Sim_{b,d}*Z_{b,d,0} + Sim_{c,d}*Z_{c,d,0} = 3*(-0.47)+4*(-0.47)+6*(0.94) = 2.36$

重複以上動作，最後系統將收斂於S_a=-0.70，S_b=0.59，S_c=4.47，S_d=3.80。

3.4.2 相似度計算

在本研究中採用兩文章的共用詞語數做為文章間的相似度，若參考文章與目標文章用到越多相同的詞語，則認為兩篇文章的相似度越高；反之，則認為越不相似。

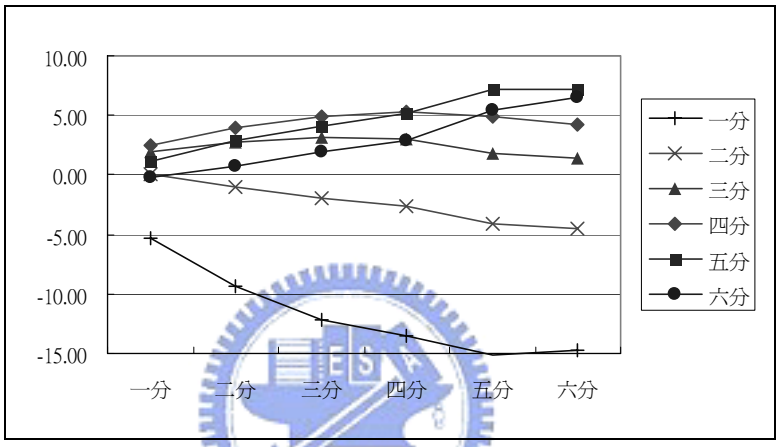
表一　各級分文章間平均共用詞語數表

	一分	二分	三分	四分	五分	六分
一分	13.38	18.64	20.54	21.05	19.80	18.43
二分	18.64	27.00	30.69	31.96	30.80	28.68
三分	20.54	30.69	35.79	37.59	36.70	34.63
四分	21.05	31.96	37.59	39.89	39.78	37.48
五分	19.80	30.80	36.70	39.78	42.08	40.42
六分	18.43	28.68	34.63	37.48	40.42	39.67
Avg.	**18.64**	**27.96**	**32.66**	**34.63**	**34.93**	**33.22**

表一為不同級分的文章之間共用詞語數的平均值，同一篇文章與本身的共用詞語數不計。觀察樣本集中所有文章之間的共用詞語數，普遍來說，大多數的文章與高分文章

的共用詞語數均較低分文章為高，主要原因是高分文章使用的詞語數量較低分文章多，因此詞語出現的機率亦較高；然而成績越高的文章，共用詞語數隨成績上升的幅度越大，而低分文章對不同級分文章的共用詞語數差距相對不明顯。

將表一中每欄的值減去該欄平均後，可以得到圖四的結果。由圖可見一二級分文章的校正後共用詞語數隨成績遞減，代表這些文章與高分文章使用到相同詞語的機會相對較低；三四級分文章呈現中間高兩端低的鐘型，代表與同樣為三四級分的文章用到相同詞語較高，而與高低分文章用到相同詞語的機率較低；五六級分文章的校正後共用詞語數隨文章成績上升，代表這些文章與高分文章使用到相同詞語的機會較高。



圖四　各級分文章間平均校正後共用詞語數

在單一詞語之外，我們亦嘗試加入共用 Bi-word 雙詞組的資訊，以「*[下課][十分][鐘]*」與「*[十分][鐘][的][下課]*」為例，兩者的共用詞語為*[下課]*、*[十分]*、*[鐘]*，共三個，僅使用詞語的相似度即為 3，而共用 Bi-word 數為*"[十分][鐘]"*一個，故在加入 Bi-word 的系統中，相似度為 4。

標點符號不視為詞語，但句首與句尾可以視為 Bi-word 的組成分子。如「*[下課][了][！][同學][們][馬上][衝][出][教室][。]*」與「*[鐘聲][一][響][，][同學][又][走][回][教室][。]*」兩句，便有*"<句首>[同學]"*及*"[教室]<句尾>"*兩個共用 Bi-word。

加入 Bi-word 詞組之資訊可以提供系統些許訓練文章的文法資訊，避免詞袋式評分無視詞語間關聯性及詞語出現位置的問題。實驗結果顯示加入 Bi-word 資訊可以提高系統對高分文章的鑑別力（詳見 4.1.5）。

3.4.3 討論

為了分析使用詞語對文章的影響，以採用共用詞語數為相似度的系統來說，我們可以將公式改寫為：

$$W_{w,j,t} = \sum_{i \neq j} F_{w,i} * Z_{i,j,(t-1)} \qquad (3)$$

$$S_{j,t} = \sum_{w} F_{w,j} * W_{w,j,t} \qquad (4)$$

$$Z_{i,j,t} = \frac{\left(S_{i,t} - \sum_{k \neq j} S_{k,t} \Big/ (N-1) \right)}{\sigma_t}$$

$W_{w,j,t}$：時間為 t 時，詞語 w 對於文章 j 的分數。

$F_{w,i}$：詞語 w 是否在文章 i 中出現。（Binary Value）

$S_{j,t}$：時間為 t 時，文章 j 的分數。

$Z_{i,j,t}$：時間為 t 時，文章 i 對於文章 j 的 Z 分數（Z-Score）。

N：文章總數。

σ_t：時間為 t 時，所有文章分數之標準差。

$W_{w,j,t}$為所有出現詞語w的參考文章對目標文章j的Z-Score總和，又可以分解為詞語的文件頻率與出現文章平均Z-Score的乘積。如果一個詞語出現頻率很低，則此詞語提供之分數必然不高；然而若一詞語有很多篇文章使用，但並沒有特別出現在高分或低分文章的趨勢，則此詞語對文章的影響亦不會太大。

我們並不採用比例式的分數，因為我們假設大多數的測試文章都是正確的描述題目，所以詞語的出現頻率越高，可以視為該詞語和此題目的相關程度越高，故文章頻率亦是一個判斷詞語好壞的參考；其次，在投票系統運作時，我們無法確定文章的好壞，只能做大略的估測，出現頻率高的詞語其文章平均分數較為可靠，而頻率過低的詞語可能出現嚴重的誤差，因此我們必須降低低頻詞語對系統的影響。

由此可見，若是寫作者刻意增加詞語數，但使用的詞語之高低分文章比例不佳（無鑑別力詞語），或是在測試資料中出現頻率過低（非關題旨詞語），並無法提升本系統對

該文章的評價，甚至當使用到負分的詞語時，還會有降低分數的情況。所以雖然普遍來說，本系統的評分和文章的使用詞數有強烈的正相關，但並不會因此而被破解，相反地對於詞語數多的離題文章，有比其他系統更好的效果（見 4.2.3）。

這也說明了本系統在相似度計算上採用累加式計算而非比例式計算的原因。在我們的假設中，若是一個詞語出現文章的成績分布與整個訓練資料的分布相近，或是出現的頻率甚低，我們認為這個詞語對於評分沒有帶來任何資訊，因此我們希望此詞語對文章分數的影響趨近於零。

若採用比例式的相似度（如：將共同詞語數除以兩篇文章的相異詞語數做為文章間的相似度），將使文章分數與使用的詞語數呈反比，便會造成前述的零分詞語對高分文章（Z-Score 大於零、高於平均之文章）有扣分效果，而對低分文章（Z-Score 小於零、低於平均之文章）有加分效果，這兩者均非我們所樂見，因此我們並不使用比例式的相似度計算法。

3.5 六級分評鑑

當系統進入穩定狀態後，我們便可以利用最後的評分結果，評定測試文章的六級分成績。

在此可以有兩種假設，第一種假設是存在某些間接特徵可以代表樣本的水準，而不受到題目的影響，如此我們便可根據此間接特徵值判斷樣本的成績分佈。

舉例來說，若採用相異詞語數為級分區間的標準，而假設根據歷史資料，一級分文章數與相異詞語數未達 68 個的文章數相當。我們在評閱其他樣本時，便可先計算該樣本內相異詞語數未達 68 個的文章數，以做為一級分文章數量的估計值，以同樣的方法求得二至六級分文章數估計值，最後以文章於投票演算法的名次評定成績。

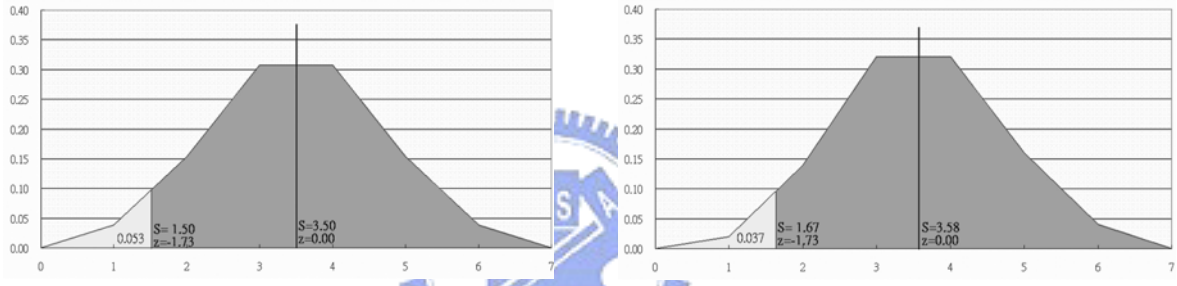
第二種假設是不管任何題目，當樣本的文章數達到一定數量，其分佈就會趨於一致，也就是各級分文章佔測試資料的比例均與歷史資料相近。

在本論文中我們採用第二種假設，我們利用歷史資料的成績分佈，用常態分配計算

各級分的 Z-Score 區間，再視測試文章經投票系統後的最終 Z-Score 所在區間決定其成績。

根據歷史資料，1~5 級分文章佔樣本的累積比率分別為 6.5%、25.1%、55.6%、85.8%、99.0%，換算為標準常態分配之 Z-Score，可得到五個門檻植-1.512、-0.671、0.141、1.070、2.320。因此，在投票系統的最終 Z-Score 低於-1.512 的文章，我們給予一級分的成績，介於-1.512 到-0.671 的文章，則評為二級分，以此類推。

當成績分佈與歷史資料有所差異時，使用 Z-Score 做為評分標準可以有修正效果，當低分文章比例降低時，整個分數分佈將更左偏，在同樣的 Z-Score 區間值設定下，系統判斷為低分的文章比例便自然地降低，反之亦然。



a. 左右對稱分佈之樣本　　b. 低分文章比例降低（左偏分佈）之樣本

圖五　不同分佈之樣本比較圖

以上圖為例，在第一 Z-Score 門檻值設為-1.73 的情況下，左右對稱的樣本有 5.3%的文章會被評為一級分，而在低分文章比例下降之樣本中，同樣的門檻值僅會使得 3.7%的文章被評為一級分。若是採用名次做為評分標準，則不管訓練資料的分布如何，均會使分數最低的 5.3%的文章評為一級分。

採用 Z-Score 為基準的評分結果與實際分佈較為接近，這就是我們採用 Z-Score 而非名次比例做為六級分評鑑的標準的原因。

3.6 離題文章偵測

有一些寫作技巧優秀卻不合題意的文章，這些文章的間接特徵往往十分優異，而直接特徵並不符合題目的要求；當這類文章夾雜在測試文章中進入自動評分系統時，便有可能造成系統的誤判。造成使用者可以靠著這個弱點來欺騙系統，以取得不適當的高

分，因此我們需要一個機制來偵測這些文章，以增強系統的完備性。

寫作所使用的詞語可分為「主題」和「修辭」兩部份，這類離題文章在主題相關的詞語與其餘測試文章普遍不同，故在本系統中不致被評為高分，但在修辭方面卻經常與高分文章用到相同的詞語，因此仍可得到一定的分數，有可能被評為中等文章；而文章的修辭技巧為高等文章與中等文章的主要差異，故不可能將修辭部份的評分由系統中剔除。

在有訓練資料的系統，如 CMU〔5〕和 Allan 等人〔1〕在新主題偵測（New event detection）上所使用的方法，通常是利用訓練資料的直接特徵資訊來處理離題文章的問題，當測試文章與訓練資料有顯著差異時便可認定其為離題文章。我們使用相同的概念來處理無訓練資料的離題文章偵測，我們認為若一篇文章使用的詞語與高分文章有顯著的差異，則此篇文章應為離題文章。此演算法必須設定三個門檻值：詞頻門檻值 x、比例門檻值 y 及詞語門檻值 n。

在所有文章經過投票演算法得到系統評閱的分數後，計算詞語在分數高於平均值的文章中出現的比例 GDF，其值大於詞頻門檻值 x 的詞語，我們認為是和題目相關的。根據實驗的結果，詞頻門檻值 x 定在 15%到 25%之間較為恰當。

然而有些常用詞會廣泛地出現在各種主題的文章中（例：的、可以、因為...），雖然其 GDF 高過詞頻門檻值，但並非與目前的特定主題相關，為了刪去這些詞語，我們由中研院平衡語料庫〔6〕中選取了 876 篇無特定主題的文章，計算所有詞語在該語料庫出現的文章頻率 CDF。

一個詞語要被認定是相關詞語，其 GDF 除了要高過必須詞頻門檻值 x 外，也要大於其 CDF 的 y 倍，才代表和題目有特殊的相關性。經實驗觀察結果，比例門檻值 y 定於 4~8 之間較為適當。

得到相關詞語集後，便可計算每篇文章出現的相關詞語數，若相關詞語數量未超過詞語門檻值 n，代表此文章並未完整描寫題目的意涵，則認定此篇文章為不合題意的文章；反之，則認為是合乎題意的文章。

四、 實驗

4.1 投票系統

4.1.1 實驗資料

本實驗使用的資料為台灣都會區及非都會區的各三所學校國二學生所撰寫的作文，題目為「下課十分鐘」，所有資料共有 689 篇，為了與有訓練資料的系統比較，所有系統均只使用半數的 346 篇做為測試資料，其餘的 343 篇文章則做為監督式評分系統的訓練資料。每篇文章都由二至三名國中國文老師評閱，並給予一至六級分的成績，將所有評分者的成績平均並四捨五入後做為該文章的人工評閱成績。

4.1.2 實驗流程

所有測試文章都先經由長詞優先法進行斷詞處理後，計算各文章的相異詞語數做為文章的初始分數，輸入投票系統運作 50 個迴圈，使文章的分數達到穩定狀態後，最後根據文章的 Z-Score 所在區間評定文章之六級分成績。

在第一次實驗中，僅使用共用詞語數做為文章相似度，在第二次實驗時，則使用共用詞語+Biword 數為相似度。

4.1.3 評鑑方式

本實驗使用正確率（adjacent rate）及精確率（exact rate）兩項指標來評鑑系統效能：

正確率：系統評分與人工評分誤差一分以內之文章數／總文章數

精確率：系統評分與人工評分完全相同之文章數／總文章數

由於現行寫作測驗中，兩名人工閱卷者的評分差距在一分以內皆被視為正常情況，故以誤差一分內的文章比例做為正確率的評斷。

4.1.4 實驗結果

實驗結果如下表所示：

表二　共用詞語數系統之評分結果表

系統評分 / 人工評分	一分	二分	三分	四分	五分	六分	正確率	精確率
一分　(23 篇)	17	5	1	0	0	0	95.7%	73.9%
二分　(64 篇)	9	32	17	6	0	0	90.6%	50.0%
三分　(105 篇)	2	15	46	36	6	0	92.4%	43.8%
四分　(104 篇)	0	2	27	62	13	0	98.1%	59.6%
五分　(46 篇)	0	0	0	23	23	0	100.0%	50.0%
六分　(4 篇)	0	0	0	2	2	0	50.0%	0.0%
合計　(346 篇)	**28**	**54**	**91**	**129**	**44**	**0**	**94.5%**	**52.0%**

表三　共用詞語+Biword 數系統之評分結果表

系統評分 / 人工評分	一分	二分	三分	四分	五分	六分	正確率	精確率
一分　(23 篇)	16	5	2	0	0	0	91.3%	69.6%
二分　(64 篇)	9	33	17	5	0	0	92.2%	51.6%
三分　(105 篇)	1	23	45	32	4	0	95.2%	42.9%
四分　(104 篇)	0	6	29	57	12	0	94.2%	54.8%
五分　(46 篇)	0	0	0	15	30	1	100.0%	65.2%
六分　(4 篇)	0	0	0	1	3	0	75.0%	0.0%
合計　(346 篇)	**26**	**67**	**93**	**110**	**49**	**1**	**94.5%**	**52.3%**

4.1.5 效能比較與討論

表四　各系統效能比較表

	正確率	精確率
投票系統(詞語)	94.5%	52.0%
投票系統(詞語+Biword)	94.5%	52.3%
隨機評分	64.1%	23.9%
完全三級分	78.9%	30.3%
ID3 決策樹	93.9%	42.2%
支援向量機	93.6%	49.4%
貝氏學習機(w/o rules)	93.4%	50.3%
貝氏學習機(with rules)	96.2%	55.8%

ID3 決策樹、支援向量機以及貝氏學習機為〔8〕〔9〕〔7〕中提到的三種機器學習模型，貝氏學習機除了原始模型外，另有加入特殊評分規則的版本。三者使用的特徵均採用〔7〕所使用的六項特徵；訓練資料均為測試資料以外的 343 篇文章，其分數分佈與測試資料相同。

由表四可見，本系統雖然沒有訓練文章的資訊，但正確率和精確率均已超過三個有訓練資料的原始模型，僅略遜於加入規則的貝氏學習機系統。系統的誤差率和專業閱卷者之間的誤差率亦已十分接近，足見本系統具有相當的可信度，可以做為人工評分時的參考依據。

加入共用 Biword 數與僅使用共用詞語數之系統相比，整體效能幾乎沒有差別，惟加入 Biword 資訊得出之成績分佈較接近實際分佈，且對高分文章的鑑別力較佳。僅使用共用詞語數的系統，有過度將文章評為四級分之傾向，使得五六級分文章之表現不佳，然而四級分文章之精確率較高。

4.1.6 分群實驗

為了探討測試文章數量對評分效果的影響，我們將 346 篇測試文章均分為五群，各群的文章數量與成績分佈均相近，再以不同大小的環狀滑動窗選取輸入之測試資料範圍進行多次實驗，計算不同測試文章數時的正確率及精確率。

（例：群數大小設定為 2 時，進行 5 次實驗，測試資料分別為群 1&群 2、群 2&群 3、群 3&群 4、群 4&群 5、群 5&群 1，其餘同理。）

4.1.7 分群實驗結果與討論

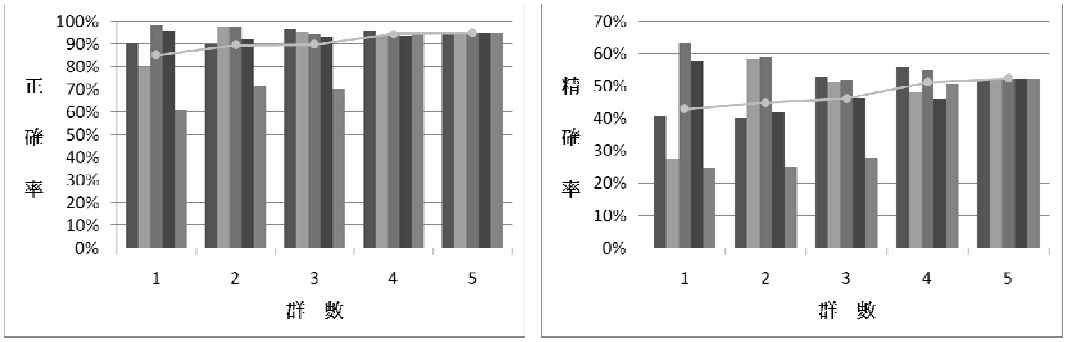
分群測試之結果如下表：

表五　分群測試評分結果表

共用詞語數					
群數	文章數	平均正確率	平均精確率	最低正確率	最低精確率
1	69.2	86.1%	40.5%	65.2%	27.5%
2	138.4	90.0%	46.8%	75.0%	28.6%
3	207.6	90.3%	45.0%	78.5%	34.0%
4	276.8	92.5%	48.3%	86.6%	40.1%
5	346.0	94.5%	52.0%	94.5%	52.0%

共用詞語+Biword 數					
群數	文章數	平均正確率	平均精確率	最低正確率	最低精確率
1	69.2	85.0%	42.8%	60.9%	24.6%
2	138.4	89.5%	44.9%	71.4%	25.0%
3	207.6	89.7%	46.1%	69.9%	27.8%
4	276.8	94.4%	51.1%	93.5%	46.0%
5	346.0	94.5%	52.3%	94.5%	52.3%

詞語+Biword 系統的詳細實驗結果如下圖，直條圖表示個別實驗的正確（精確）率，折線圖表示在各群數下的平均正確（精確）率：



圖六　分群測試結果比較圖　（詞語+Biword）

由表五及圖六可見，本系統的效果受到樣本數多寡的影響，對於加入 Biword 資訊的系統而言，由於有資料稀疏的狀況，此情況更為嚴重。

本系統在樣本數不足 210 篇時，平均正確率和平均精確率均不理想，且可能出現很糟糕的結果，表現並不穩定，但當測試文章數達到 280 篇後，系統效果便趨於穩定，不易出現很低的正確率或精確率。

4.2 離題文章偵測

4.2.1 實驗資料

我們蒐集了 20 篇與「下課十分鐘」無關之文章，內容包括中學國文課文、新聞、小說、散文作品、學術論文、以及流行歌歌詞等，各文章的相異詞語數在 143 到 261 之間。將此 20 篇離題文章加入原本的 346 篇測試文章中做為離題偵測實驗的測試資料。

4.2.2 實驗流程

在投票系統方面，先將 366 篇測試文章輸入投票系統，得到系統評閱的分數，區分出高分文章群後，再由離題偵測系統根據文章使用的詞語及三個可調整的門檻值判斷是否為不合題意之文章。

做為對照組的四個系統不需要其他測試文章，因此直接將 20 篇離題文章做為測試資料輸入系統。

4.2.3 效果比較

表六　各系統離題文章評分結果表

	平均成績	四分以上文章比例	最高成績
投票系統(詞語)	2.50	10%	4
投票系統(詞語+Biword)	2.10	0%	3
ID3 決策樹	3.95	90%	5
支援向量機	3.65	70%	5
貝氏學習機(w/o rules)	5.10	90%	6
貝氏學習機(with rules)	4.50	90%	6

在資料輸入離題偵測系統之前，先觀察投票系統評閱離題文章的結果，由於此演算法基於文章相似度的特性，20 篇離題文章在投票系統結束時多被評為一到三級分；在僅使用詞語的系統中，被評為四級分以上者僅有兩篇，在加入 Biword 的系統中更是一篇也沒有。與其他系統有 70%到 90%被評為四級分以上的結果相比，顯然較為優秀。

4.2.4 評鑑方式

在離題文章偵測部份，本實驗使用兩項指標來評鑑系統的效能：

False negative rate 代表離題文章中，未被偵測出離題之文章比例。

False positive rate 代表被偵測為離題之文章中，人工評閱為二級分以上的文章比例。由於一級分的文章多為不完整的拙劣文章，因此我們也將其視為不合題意文章的一部份。

兩項錯誤率會因門檻值設定的不同而改變，當門檻值越高時，系統越敏感，判斷為離題文章的機率越高，故發生 False negative error 的機率將降低，相對地 False positive error 的發生率將上升；反之，當門檻值降低時，False negative rate 將上升，而 False positive rate 則會下降。

4.2.5 實驗結果與討論

表七顯示根據詞+Biword 系統的結果，在比例門檻值 y=5 時，另外兩個門檻值不同設定下的結果。

表七　離題文章偵測結果表 （粗體代表兩項錯誤率均為 0）

x / n	10%		15%		20%		25%		30%	
	False Negative	False Positive	False Negative	False Positive	False Negative	False Positive	False Negative	False Positive	False Negative	False Positive
2	0.550	0.000	0.450	0.000	0.450	0.000	0.200	0.000	0.100	0.000
3	0.400	0.000	0.350	0.000	0.300	0.000	**0.000**	**0.000**	**0.000**	**0.000**
4	0.400	0.000	0.150	0.000	0.100	0.000	**0.000**	**0.000**	**0.000**	**0.000**
5	0.200	0.000	0.100	0.000	**0.000**	**0.000**	**0.000**	**0.000**	0.000	0.037
6	0.050	0.000	**0.000**	**0.000**	**0.000**	**0.000**	0.000	0.071	0.000	0.097
7	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.000	0.143	0.000	0.268
8	0.000	0.036	0.000	0.034	0.000	0.091	0.000	0.244	0.000	0.417

當比例門檻值 y=5，詞頻門檻值 x=25%，詞語門檻值 n 在 3~5 之間時，系統有相當良好的表現；兩項錯誤率均為 0，代表所有離題文章均被正確的偵測出，且沒有任何二級分以上的文章被誤認，僅分別有 1 篇、1 篇、2 篇一級分文章被認為是不合題意文章。可見此偵測系統能在沒有訓練資料的情況下有效偵測出完全離題之文章。

五、 結論與展望

5.1 結論

在本論文中我們提出一個以文章相似度為基礎的中文作文自動評分系統，可以在沒有訓練資料的情況下自動評閱中文作文的分數。實驗結果可以達到 95%的評分正確率和 52%的評分精確率，代表本系統的評分結果與國中老師十分接近，可以做為批閱文章時的參考依據。本論文亦提出一個能有效偵測離題文章的方法，可以做為測試文章是否符合題意的參考。

5.2 未來工作

本論文僅探討文章在取材方面的表現，並未考慮文章的結構和文法的優劣，希望未來可以將這些資訊整合進系統當中，以達到更好的評分效果。

在投票演算法方面，本系統在文章相似度計算上僅使用簡單的共用詞語數及共用 Bi-word 數計算，未來可以考慮加入其他資訊來獲得更精確的文章相似度。

相似度計算目前以詞為單位，而同義詞對文章造成的影響並不相同，可以考慮以《同義詞詞林》或《知網》等知識庫的資訊來改進，令詞語的影響力更為合理。

參考文獻

〔1〕 J. Allan, R. Papka, and V. Lavrenko. “On-line new event detection and tracking.” Proceedings of SIGIR-98, pages 37–45, Melbourne, Australia, 1998.

〔2〕 Y. Attali and J. Burstein. “Automated Scoring Using With e-rater V.2. The Journal of Technology.” Learning and Assessment, vol.4, no. 3, 2006.

〔3〕 T.K. Landauer, D. Laham, and P.W. Foltz. “The Intelligent Essay Assessor.” IEEE Intelligent System, 15, 27-31, 2000.

〔4〕 E.B. Page. “Computer Grading of Student Prose, Using Modern Concepts and Software.” Journal of Experimental Education, 67, 127-142, 1994.

〔5〕 Y. Yang, T. Pierce, and J. Carbonell. “A study on retrospective and on-line event detection.” Proceedings of SIGIR-98, Melbourne, Australia, 1998.

〔6〕 中研院平衡語料庫 3.1 版

〔7〕 林信宏，「基於貝氏機器學習法之中文自動作文評分系統」，國立交通大學，碩士論文，2006。

〔8〕 張佑銘，「中文自動作文修辭評分系統設計」，國立交通大學，碩士論文，2005。

〔9〕 粘志鵬，「基於支援向量機之中文自動作文評分系統」，國立交通大學，碩士論文，2006。

〔10〕 蔡沛言，「自動建構中文作文評分系統：產生、篩選與評估」，國立交通大學，碩士論文，2005。