國 立 交 通 大 學

資訊科學與工程研究所

碩 士 論 文

中 文 寫 作 多 面 向 評 分 系 統

Multi-face Automated Chinese Essay Scoring System

研 究 生：葉啟祥

指導教授：李嘉晃 教授

中 華 民 國 九 十 七 年 六 月

中 文 寫 作 多 面 向 評 分 系 統

Multi-face Automated Chinese Essay Scoring System

研 究 生：葉啟祥　　Student：Chi-Hasing Yeh

指導教授：李嘉晃　　Advisor：Chia-Hoang Lee

國 立 交 通 大 學

資 訊 科 學 與 工 程 研 究 所

碩 士 論 文

A Thesis
Submitted to Institute of Computer Science and Engineering
College of Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master

in
Computer Science
Jun 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年六月

中 文 寫 作 多 面 向 評 分 系 統

學生 ： 葉啟祥　　指導教授：李嘉晃 教授

國立交通大學資訊學院 資訊科學與工程研究所碩士班

摘要

自動寫作評閱的研究，在自然語言中佔了重要的一環，尤其是在中文研究上甚是艱難，雖然陸陸續續已有評閱系統之研究產生，但目前的系統皆只針對文章單一方面給分，無法有效提供使用者在寫作技巧上哪方面較微弱之資訊。因此本文提出一個非監督系統，針對中文寫作評分不同面向，分別給予分數以及分數的統整，除了給予使用者在立意取材以及結構組織上的分數外，也根據使用者所寫作的文章給予錯別字回饋的資訊。實驗結果在不同面向上能有相當程度的正確率，在分數統整上，正確率可達到 94%。此外錯別字判斷的正確率能達到 72%，可作為老師批閱或是學生寫作上的輔助工具。

Multi-face Automated Chinese Essay Scoring System

Student : Chi-Hasing Yeh Advisor：Prof. Chia-Hoang Lee

Department of Computer and Information Science
National Chiao Tung University

Abstract

The research of the automated essay scoring is important in the natural language and it is more difficult when applying it on the Chinese language. Although some scoring systems have been proposed, they only score the article in one way. They can not provide the information that which aspect in the article the users should strengthen to improve their writing skill efficiently. Thus, this paper proposed an unsupervised learning system that could grade essays from multi-dimension and give misspell information to the user as well. The experiment shows that the adjacent rate in the overall experiment is about 94% and the misspell judgment rate is about 72%.

目錄

圖目錄

Algorithm

表目錄

第一章、 緒論

1.1 研究動機

各個國家的語言教育，皆脫離不了聽、說、寫、讀這四個方面，而在這四個方面中，尤其以“寫”這一環最為重要；寫作不僅可以培養一個人的表達能力、文學素養，甚至可以激發、訓練一個人的組織與思考以及增進創造、理解等能力。因此在各個語言教育階段中，均重視語言寫作能力的訓練。

但現階段的作文批閱的形式，皆需要耗費大量的人力、物力以及時間，最重要的還是批閱者的主觀不同；但除了批閱者的主觀意識外，另一項重大的問題是批閱者如何在長時間的作業下，還能維持一定的批閱標準。因此單純利用人工來進行作文的批閱，很難達到客觀以及公平性。

在英語批閱研究中，自動作文評分(Automated Essay Scoring ,AES)已經發展許久，甚至已經應用在大行的語文考試中，例如: Graduate Management Admission Test (GMAT) 已使用 E-rater 作為批閱文章的輔助工具[1]。而華語批閱研究中，也針對寫作上由最初所提出的自動建構中文作文評分系統[2]，到之後的貝氏[3]、SVM[4]、修辭[5]、非監督式[6]、結構化[7]等評分系統。

然而目前的中文系統無論是監督式評分系統(指需要一定的篇數且已經過人工評定分數的同一主題文章作為系統的訓練資料)或者非監督式評分系統(無需訓練資料，僅需一定數量的同一主題文章)，皆只針對單一個面向來作評分，並不符合中文寫作的評分標準上[8]所針對的主要四大面向立意取材、結構組織、遣詞造句及錯別字與標點符號等。因此無法有效的反映出學生在寫作技巧上較為薄弱的部份。

1.2 研究目的與假設

本論文之研究目的，在於建立一套不需要訓練資料且可針對作文上不同角度分別給分，最後再結合這些個別分數給予一個總結分數，並且給予錯別字上的回饋。此系統在單一面向評分上是不需要事前藉由人工評定分數來當訓練資料，僅需要一定的同主題文章數，便可藉由文章特徵的資訊、文章間的相似度進行自動評分，而在整合分數上亦不需要訓練資料，藉由分別單一面向所評定出的文章與分數分佈作為參考，進行調整性的分數整合。

1.3 論文架構

第一章為前言，主要內容說明本論文的研究動機以及研究目的。

第二章為相關的研究，將介紹 AES (Automated Essay Scoring)系統的發展。

第三章則是介紹本研究之系統內部架構及流程。

第四章將針對此研究的實驗過程與研究結果以及其他系統之比較做說明。

第五章則是描述本論文的研究總結，以及未來展望。

第二章、相關研究

此章節中，首先先介紹英文作文評閱系統的發展歷史，接著再介紹在中文上系統之發展流程。

早在 1966 年 Ellis Page 就提出的一種簡單的評分方法 PEG (Project Essay Grader)[9]，這也是第一個英文的評分系統，但系統並未包含了 NLP 技術，系統主要包含 training stage 與 scoring stage，training stage 主要是找出依賴文章的間接特徵如: 作文字數、標點符號、形容詞數等，再經由 scoring stage 進行多元迴歸分析，來計算出文章分數。但因為它使用文章的間接特徵，並未加入文章的直接特徵如:結構組織、句法等資訊，所以使用者容易針對這個弱點，撰寫出一篇較長的文章瞞過系統，得到較高的分數。

因此在 PEG 之後，Landauer 提出了一個利用文章的語意關係來進行評分，此系統 Intelligent Essay Assessor (IEA)[9]，主要是利用 LSA 為基底，進而找出文章間的語意資訊進行評分，系統除了評分外，它也會針對文章 grammar、style 與 mechanics 給予回饋資訊。

然而上述兩個系統它們所著重的文章特點不盡相同，PEG 是著重於 style，而 IEA 著重於 content，很難用於大型考試上。於是就有後來的 E-rater 的產生[9]。E-rater 系統已經被美國商業學校入學測驗考試 (GMAT)所使用，在評分的過程中，分為三個階段:結構、組織、內容。在結構部分中主要分析出句法的種類如:不定詞、從屬子句等。在組織部份分析句法的概念如:修辭結構，句子跟句子之前的連接詞等。最後在內容部分，評估文章內所用到的詞語跟主題是否能吻合。文章會由三個階段得到統整的資訊以及些許的回饋。

在中文的自動評閱系統上，是近幾年才陸陸續續有人提出。最早期也是根據文章的表面特徵如:詞語數、成語數等。再加入譬喻以及排比所建立出的評分系統[5]。之後才提出根據同主題文章的訓練，得出能反映文章好壞的直接特徵:義元[2]。以及利用統計的方式，擷取出符合這個主題的結構概念[7]，再針對各篇文章上的結構，比較之間的相似程度進行評閱。

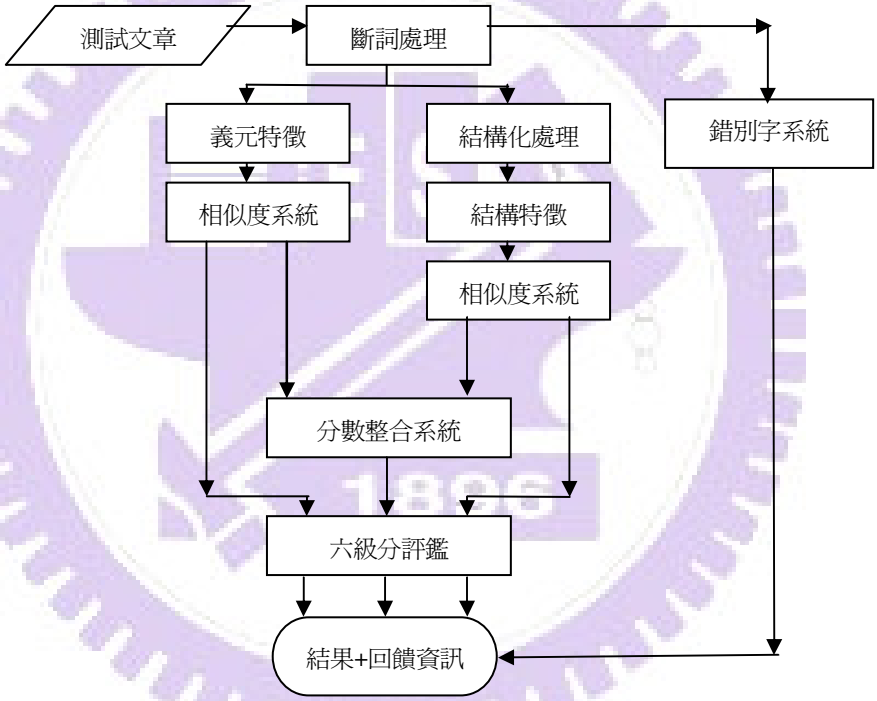
除了利用特徵擷取來評分外，也有人提出利用 Bayesian、SVM 等學習機器來進行評分[3][4]，利用文章的特徵與人工評定好的分數當作訓練資料建立出的機器評分規則，再針對測試文章進行評分。但這些系統都需要一定同主題及人工評定過的文章數當作訓練文章，仍須人工的方式介入。於是一個非監督式的評分系統構想就由陳[6]所提出，根據文章間所共同用到的詞語，來做互相評分的依據，其正確率依然與監督式的系統相差不遠。

以上中文自動評閱雖然分為針對特徵擷取、機器學習、非監督式與監督式等來評分，但都只針對文章寫作上單一角度上來做評分，不像英文系統 IEA、E-rater，可以從不同的角度上評分，並統整分數進而給予回饋資訊。很難反映出使用者在寫作上在那一方面出現問題。

第三章、系統設計

此章節中，將描述整個系統的架構與流程，首先在 3.1 小節中，用一張系統架構圖來了解系統整個運作的流程，圖中各個模組的執行內容將在後續幾小節中做詳細的介紹。

3.1 系統架構



圖一. 系統架構圖

本系統裡面共包含 5 個模組：

1. 斷詞處理
2. 特徵擷取-包含兩個部份:義元特徵、結構化特徵
3. 相似度(評分)系統
4. 分數整合系統
5. 錯別字系統

當一定的文章數資料進入系統時，系統開始運作。首先第一步對每篇文章進行斷詞的處理，再經由特徵擷取取出每篇文章的義元特徵與結構特徵當作後續相似度評分系統的評分依據，藉由相似度內的投票演算法每個時間修正評分的結果直到結果達到穩態為止，在此階段結束後進入分數整合系統，即會根據兩個方向所評出的分數、文章數進行訓練找出一個最佳比例來進行整合。最後這些分數會參考歷史資料的成績分配情況轉換成六級分成績。文章在系統評出分數後也會根據斷詞以及 bigram 資訊進行錯別字判斷。

3.2 中文斷詞處理

斷詞在自然語言上是不可或缺的技術，任何的系統只要是牽扯到語言的都必須先分辨文章中的各個詞才能進行詞性標記、句法分析、資訊擷取等進一步的處理。相對於英文最顯而易見的差異，在於中文語法並沒有空白隔開每一個詞。若斷詞結果不正確，容易造成語意全然的不同，因此中文的自動斷詞成為重要的工作。

目前系統採用的斷詞法乃是長詞優先斷詞法，雖然其正確率在現有的演算法中並不算最好，但效果已達一定的水準，且系統進計算文章間共同出現的材料，並不去探討文章語法，因此對斷詞錯誤的會有較高的容忍度。

在此利用一個簡單的例子說明此演算法:

「下課鐘聲響」.... .....(1)

這個句子可能的斷詞有下列數種詞組:

「（下）（課）（鐘）（聲）（響）」.....(2)

「（下課）（鐘）（聲）（響）」.....(3)

「（下課）（鐘聲）（響）」.....(5)

.....

首先此演算法會針對字串中最長有意義的字串進行判斷，因此先從(1)斷出(下課)，再由剩餘的字串(鐘聲響)中斷出(鐘聲)以及(響)，即為結果。

3.3 詞語特徵

由於系統要針對文章立意取材以及結構組織兩部份做評分，所以我們需要先擷取出所有文章在這兩部份的詞語特徵以及結構特徵來當作相似度系統比較評分的依據，再分別進行評分。因此我們在此小節以及下一小節提到特徵如何擷取。

首先針對詞語特徵的擷取，本系統中是利用文章的義原來當作文章的詞語特徵，我們根據知網(HowNet)將所有文章中的詞語轉化為義原。

在此利用一個簡單的例子說明轉化過程:

「大家的動作由緩慢轉變成快速」

經過斷詞處理後，會變成:

「（大家）（的）（動作）（由）（緩慢）（轉變成）（快速）」

再經過義原轉換的處理，可得知各詞語的主義原如下:

表一. 義原轉換

詞語	主義原
（大家）	{human|人}
（的）	{FuncWord|功能詞}
（動作）	{do|做}
（由）	{FuncWord|功能詞}
（緩慢）	{slow|慢}
（轉變成）	{become|成為}
（快速）	{fast|快}

因此此段話共有七個詞語，但只包含六個不同的義原，而我們即取這六個不同義原當作詞語特徵。

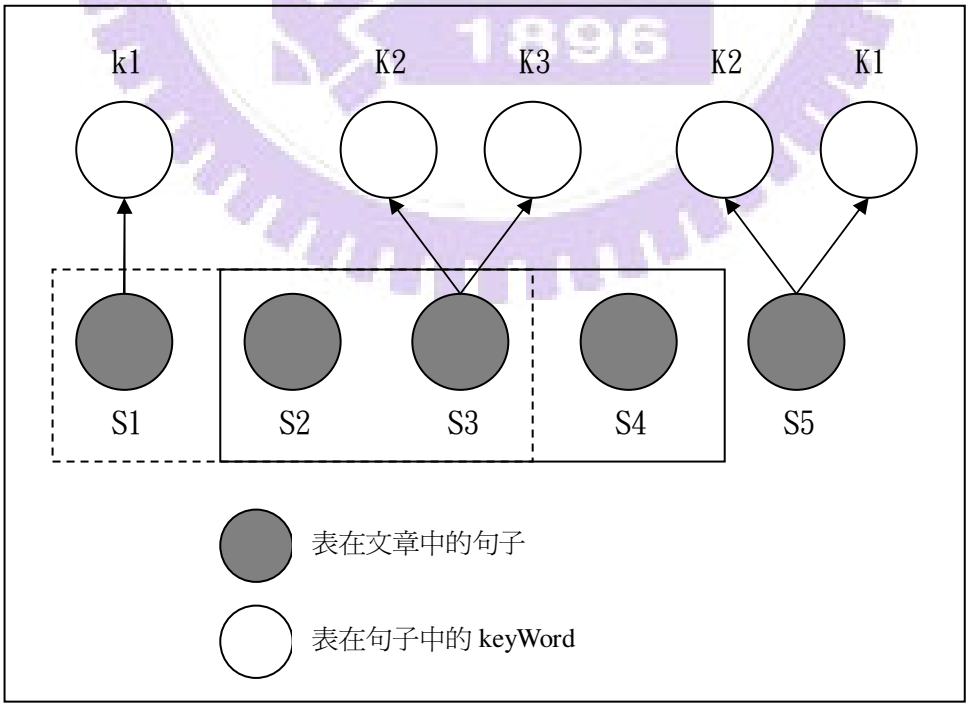
3.4 結構特徵擷取

系統必須先利用全部數量同主題的文章資料建立詞語的非對稱關係矩陣，其主要是觀察詞與詞共同出現的次數以及單獨出現的次數，來得知兩者之間的從屬關係，例如:[合作社]與[麵包]共同出現的次數 100，[合作社]單獨出現的次數 200，[麵包]單獨出現的次數 100 次，表示[合作社]除了跟[麵包]一起出現 100 次外，還跟其他的詞一起出現過;反觀[麵包]只會跟[合作社]同時出現，因此可知[麵包] 應該是[合作社]的子概念，利用這種方式就可來建立出屬於這個主題的概念階層圖，再依據所建立出來的概念圖對每篇文章作結構特徵擷取的處理。

3.4.1 詞語非對稱關係矩陣

針對每篇文章每一句所抽出來的 keyWord 在一個滑動的固定區間裡統計兩兩 keyWord 出現的次數並加以累加記錄。最後除予詞語在全部文章出現的總次數。

在此利用一個簡單圖說明如何針對每篇文章做處理:



圖二. 計算詞語出現關係圖

在上圖中，我們所設定區間為 3，在虛線區間內的句子 S1、S2、S3 中出現 k1、k2、k3 這三個 keyWord，代表它們的關係較為接近，因此分別將兩兩 keyWord 計錄一次，即從每個 keyWord 的角度去看。在滑動區間為實線部份，此時區間內的句子 S2、S3、S4 中只出現 k2、k3 keyWord，所以僅計錄 k2、k3 兩個 keyWord 的部份。

因此詞語非對稱關係矩陣其陣列內容可根據下列公式計算:

$$r_{i,j} = \frac{\sum_{t \in T} \sum_{s \in t} occ(w_i, w_j)}{frequence(w_i)}$$

$r_{i,j}$ ：為陣列第 i 列，第 j 行的內容。

s：文章 t (text)中的句子。

T：指全部文章。

$occ(w_i,w_j)$ ：詞語 i 與詞語 j 是否同時出現。（Binary Value）

$frequence(w_i)$ ：詞語 w_i 在全部文章出現的總次數。

依照上面公式我們就可將圖二轉換成矩陣，即下表:

表二、非對稱關係矩陣

	k1	k2	k3
k1	0	1.5	1
k2	1.5	1	2
k3	2	4	0

如此一來，我們就可以依據矩陣對稱格內容得知兩個詞之間的關係，此部份我們會在下小節提到。

3.4.2 概念圖

由 3.4.1 小節得到的詞語非對稱關係矩陣，為了得知兩個詞語 w_i、w_j 從屬的情況，經由觀察可得知當 $r_{i,j}$ 高於 $r_{j,i}$，表在文章中當 w_i 出現緊接著伴隨 w_j 出現的機率頗高;反之當 $r_{i,j}$ 低於 $r_{j,i}$,表 w_i 出現緊接著伴隨 w_j 出現的機率較低，因為 w_i 也可能跟別的詞出現。例如我們再 3.4 節中一開頭所舉的例子，w_i 為[合作社]；w_j 為[麵包]；$r_{i,j}$ 等於 0.5；$r_{j,i}$ 等於 1，得知當[合作社]出現而[麵包]伴隨出現的機率為 0.5，有 0.5 會跟其他詞出現。因此[合作社]應該為[麵包]的 superordinate。

因此可根據下列演算法，建構出屬於此主題的概念圖:

Algorithm 1： 建立概念圖演算法

$$
\begin{aligned}
&for \quad i=1 \quad to \quad n \\
&\quad for \quad j=1 \quad to \quad n \\
&\quad\quad if \quad r_{i,j} > r_{j,i} + e \quad then \quad a_{i,j} = subordinat\,e \\
&\quad\quad else\ if \quad r_{i,j} < r_{j,i} + e \quad then \quad a_{i,j} = superordina\,te \\
&\quad\quad\quad else \quad a_{i,j} = correlatio\,n \\
&now\_level = 0 \\
&do\{ \\
&\quad for \quad i=1 \quad to \quad n \\
&\quad\quad if\ a_{i,j} \in \{superordina\,te, correlatio\,n, processed\} for\ all\ j \\
&\quad\quad then \quad the\ level\ of\ word_i = now\_level\ ; \\
&\quad for \quad j=1 \quad to \quad n \\
&\quad\quad if \quad the\ level\ of\ word_j\ has\ been\ given \\
&\quad\quad then \quad a_{i,j} = processed\ , for\ all\ i \\
&\quad now\_level = now\_level + 1 \\
&\}while\,(now\_level <= threshold)\,or\,(\forall i \ \ word_i\ has\ been\ given)
\end{aligned}
$$

由上述演算法主要先標示陣列中每個詞的從屬關係,接著判斷是否有一整列

的陣列內容皆不屬於 subordinate，表示這一列所代表的詞是現階段層級最高的，於是標記 now_level，並將詞所對應的行標示 processed，代表已處理過。此演算法會一直做到所有詞語都被標記層級或者已經超過預設的階層數。

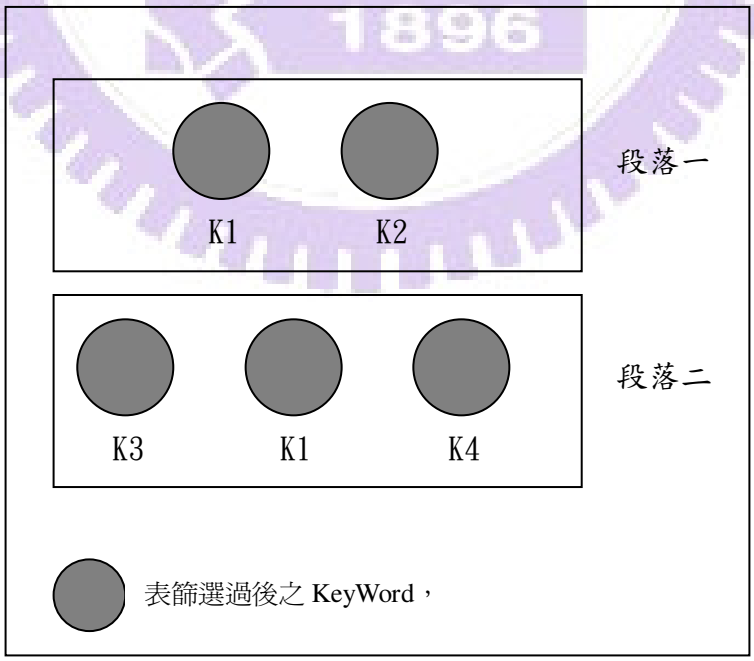
3.4.3 擷取

利用所建立出的概念圖，對每篇文章做結構特徵的擷取，首先必需先篩選每個段落的 keyWord，針對每個段落所得 keyWord 對應其概念圖所在之層級，取最高層級三層範圍內的 keyWord 當作能代表此段意義之主要詞語。

篩選 keyWord 後，隨即將每兩段所取得的主要 keyWord，做 Bi-word 的配對當作此篇文章之結構特徵。在此部份如果我們配對取的越長的話，在相似度系統的比較上，文章跟文章之間就會越難比對到，因此取 Bi-word 使系統能更有彈性。

在此針對篩選過後的 keyWord 說明其特徵擷取:

某篇文章其段落為二，每個段落所篩選之 keyWord 如下所示



圖三、特徵擷取示意圖

因此我們擷取「(k1)(k3)」、「(k1)(k1)」、「(k1)(k4)」、「(k2)(k3)」、「(k2)(k1)」、「(k2)(k4)」等 Bi-word，當作此篇的結構特徵。

3.5 相似度評分系統

此系統的基本假設為:如果一篇文章使用的材料（詞語、結構）與高分文章越相似，或與低分文章越不相似，則此篇文章為高分文章的機率越高；反之，低分的機率越高。

但在系統一開始時，所有的測試文章都不會帶任何附加資訊，因此我們並不知道文章分數以及其分佈圖。所以我們需要在系統一開始時，先針對每篇文章給予一個初始分數，以便於系統能正常運作。

因此在系統一開始會先針對文章的成語數、好義原數、名詞數、句號數這四個間接特徵加總當作文章的初始分數(即公式中 $Z_{i,j,(t-1)}$,t=0)。而這些特徵皆與文章分數有正向關係。

在根據下列公式進行評分:

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

$$W'_{w,j,t} = \sum_{i \neq j} (F_{w,i} - 1) * Z_{i,j,(t-1)} \quad (2)$$

$$S_{j,t} = \sum_{w} F_{w,j} * W_{w,j,t} + \sum_{w} (2F_{w,j} - 1) * W'_{w,j,t} \quad (3)$$

$$Z_{i,j,t} = \frac{\left(S_{i,t} - \sum_{k \neq j} S_{k,t} \Big/ (N-1) \right)}{\sigma_t} \quad (4)$$

$W_{w,j,t}$：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。

在此我們稱欲計算分數的文章稱目標文章且從詞語的角度來觀看(結構同理)。

在公式(1)(2)中主要是計算出所有出現詞語 w 文章對目標文章的權重，亦可說是詞頻與出現文章 Z-score 的乘積。所以當一個詞語的詞頻很低，則此詞語權重必然不高。公式(3)中計算目標文章的分數，其根據詞語 w 出現與否給定分數，如果詞語 w 在目標文章出現則會賦予正向的分數，否則為負向。公式(4)為文章的 Z-Score 表文章分數與平均值的差異度。

當一篇文章在 t-1 時間所得的分數高於平均許多，則目標文章從此篇文章得到較高的分數，反之亦然。在公式中除於標準差是要讓每個 t 都能收斂在一定區間，避免分數的無限成長。由此相似度的加權可得知，目標文章的分數會往相似度越高的文章慢慢趨近。

系統隨著時間 t 不斷的增加，其文章的分數也會不斷的改變，當文章數量達到一定時，其分數會趨於一個穩態的狀態。

3.6 六級分評鑑

當相似度評分系統趨於穩態的狀態後，便利用最後的評分結果，來轉換成六級分成績。我們假設當文章達到一定數量時，其分佈會趨近歷史資料的分佈，也就是說各級分文章佔測試資料比例均與歷史資料接近，於是我們將歷史資料的成績用常態分配來計算出各級分的 Z-Score 區間，再由相似度評分最後的結果所在的區間決定其分數。

由歷史資料 1~5 級分文章佔樣本的累積比率 6.5%、25.1%、55.6%、85.8%、99.0%，轉化成常態分布之 Z-Score 可得到五個門檻-1.5196、-0.6711、0.1405、1.0703、2.3204，Z-Score 低於-1.5196 給予 1 級分，介於-1.5196 到-0.6711 給予 2 級分，以此類推

3.7 分數整合評分系統

在此我們假設評閱者在批改作文時，是根據作文四個面向給予不同的比例，給予統整的分數。所以此階段主要是為了得知在哪個比例下所計算出來的分數是比較符合評閱者。利用 3.4 相似度評分系統中所分別得知文章立意取材與結構組織兩個面向文章分數及其各級分的分佈，進行最佳比例的尋找。本系統中將針對這兩個面向所得到的評分做不同比例的處理。

經上述所言，此系統基本假設為：若是由最佳比例所評定出之分數應與從立意取材、結構組織分別評定出的分數相差不遠；反之，當由最差比例所評定出之分數應該與兩者其一或是兩者的分數差距較大。

因此，系統一開始會針對不同比例先進行整合評分，再將整合出來的各級分文章根據下列所屬的類型，進行誤差權重的計算。

$$P_i = \frac{1}{Pf_{i,w,g} * Pf_{i,s,g}} \quad , if\ G_{i,now} \in G_{i,w} \cap G_{i,s} \tag{3}$$

$$P_i = \frac{|G_{i,w} - G_{i,now}| + |G_{i,s} - G_{i,now}|}{2} * \frac{Pf_{i,w,g} + Pf_{i,s,g}}{Pf_{i,w,g} * Pf_{i,s,g}}$$

$$, if\ G_{i,now} \in G_{i,w} - G_{i,s} \quad or \quad G_{i,now} \in G_{i,s} - G_{i,w} \tag{4}$$

$$P_i = \left(|G_{i,w} - G_{i,now}| + |G_{i,s} - G_{i,now}|\right) * \frac{Pf_{i,w,g} + Pf_{i,s,g}}{Pf_{i,w,g} * Pf_{i,s,g}}$$

$$, if\ G_{i,now} \notin G_{i,w} \cup G_{i,s} \tag{5}$$

P_i ：文章 i 所得到的誤差權重。

$G_{i,now}$ ：文章 i 在現在比例所評定的分數。

$G_{i,w}$ ：文章 i 依比例評定的分數對應立意取材(word)分數的文章集合。

$G_{i,s}$ ：文章 i 依比例評定的分數對應結構組織(structure)分數的文章集合。

$Pf_{i,w,g}$ ：文章 i 在立意取材中所在級分(grade)的篇數。

$Pf_{i,s,g}$ ：文章 i 在結構組織中所在級分(grade)的篇數。

在上述公式中可以得知，當一篇文章經過整合評定出來的分數與立意取材和結構組織評出來皆相同時，其類型落在公式(3)中。其所得到的誤差權重會較低；反之，權重會較高。例如：某篇文章被整合評定為 1 級分，也屬於立意取材和結構組織一級分中，其誤差權重應該經由公式(3)得出。

然而公式(4)、(5)前半項 $|G_{i,w} - G_{i,now}| + |G_{i,s} - G_{i,now}|$ 主要是用來區隔文章

類型落在(4)、(5)內，但文章分數不盡相同的誤差權重。例如:某篇文章(整合評分:1 分;立意取材:1 分;結構組織:6 分)，另一篇(整合評分:1 分;立意取材:1 分;結構組織:3 分)，其誤差權重應該為前者較高。

再由已知的文章誤差權重，經過下列公式，進行目前比例的誤差權重。

$$Error\ ratio = \sum_{i=1}^{6}\left(\sum_{\forall j \in i} P_j\right)^2 \qquad (6)$$

P_i ：文章 i 所得到的誤差權重。

Error Ratio ：目前比例誤差權重。

在公式(6)中，內部 $\sum$ 計算目前比例中每個級分的誤差，而對各個級分誤差值做平方，主要是為了讓差距能夠有所顯著，最後得到目前比例的誤差權重。當對每個比例皆做完誤差權重的計算後，即可得知產生越小權重的比例，應當對文章分數及其分佈為最佳的。

3.8 文章錯別字判斷

文章錯別字主要分作同音異字、異音異字兩種，而通常作文上普遍嚴重的錯誤通常在於同音異字上，因此本系統是針對同音異字來做判斷。

3.8.1 錯別字判斷系統流程

在這部分系統基本假設為:當文章中存在錯別字時，經過斷詞處理後。因無法找到對應的詞語，而會被斷成連續的一字詞。因此系統基本內部的判斷，會針對斷詞後被斷成連續一字詞的做處理，其流程我們根據《國語一字多音審訂表》將連續的一字詞轉換成注音型式後，再將注音型式的連續字由最長比對演算法在《國語一字多音審訂表》(以下簡稱注音表)，中尋找對應的格式。

在此利用一個簡單的例子說明其基本判斷過程:

「 今天天汽好熱 」

因為句子中有錯字，斷詞後會變成:

「（今天）（天）（汽）（好）（熱）」

針對連續一字詞做注音轉換如下:

「 ㄐㄧㄢ　ㄑㄧˋ　ㄏㄠˇ　ㄖㄜˋ 」

最長比對演算法，由最長的字串開始在注音表中搜尋:

「 ㄐㄧㄢ　ㄑㄧˋ　ㄏㄠˇ　ㄖㄜˋ 」 .....(1)

「 ㄐㄧㄢ　ㄑㄧˋ　ㄏㄠˇ 」 .....(2)

「 ㄐㄧㄢ　ㄑㄧˋ 」.....(3)

經過搜尋後，我們經由(3)可在注音表找到對應的組合，變轉換回國字型式「天氣」，剩餘的字皆不做更動，基本判斷最後結果即為「 今天天氣好熱 」。

由上述判斷中，可得知具有一定判斷性效果。但之間未考慮詞與詞之間的關係，可能會對許多原本是對的句子照成誤判的情形。

在此舉一個簡單嚴重誤判例子說明:

「 上課鐘一打 」

斷詞後如下:

「（上課）（鐘）（一）（打）」

針對連續一字詞做處理:

「 ㄕㄨㄥ　ㄧ　ㄉㄚˇ 」.....(1)

「 ㄕㄨㄥ　ㄧ 」.....(2)

系統由(2) 在注音表找到對應的組合為「中醫」，即將其轉換，最後結果為「 上課中醫打 」。

因此我們為了避免這種情形時常發生，即加入了中研院平衡語料庫 bigram 雙詞語的資訊，此資訊可以得知哪些詞語的配對是可能出現的。例如:[上課][鐘]此配對在 Bigram 中有出現；然而[上課] [中醫]在 Bigram 是沒出現過的。當作系統進入基本錯別字判斷的門檻。

其運作過程針對斷詞後的每個詞，進行下述演算法:

Algorithem2 : 錯別字檢查演算法

$$
\begin{aligned}
&If\ (W_{i-1}W_i \in Bigram)\ and\ (W_iW_{i+1} \in Bigram)\\
&\quad then\quad print\ \ Wi\ ;\\
&else\quad if\ (W_{i-1}W_i \in Bigram)\ or\ (W_iW_{i+1} \in Bigram)\\
&\qquad then\quad begin\\
&\qquad\qquad processed\ \ by\ the\ Base-Work\ ;\\
&\qquad\qquad if\ (W_{i-1}W_i^* \in Bigram)\ then\quad print\ \ Wi^*\ ;\\
&\qquad\qquad else\quad print\ \ Wi\ ;\\
&\qquad end
\end{aligned}
$$

W_i ：斷詞後第 i 個欲檢測的詞語。

$W_{i-1}\ W_i$ ：斷詞後第 i-1 個詞與第 i 個詞所構成的 bi-word。

Base-Work ：基本內部的錯別字判斷。

W_i^* ：經過 Base-Word 所跟換的詞。

因此嚴重誤判的例子經過上述演算法後其流程與結果如下:

「 上課鐘一打 」

經斷詞後如下:

「（上課）（鐘）（一）（打）」

首先判斷第一個詞語(上課)，再系統中我們將字首前 bi-word 與字尾後 bi-word 視為 true。

「( )(上課)」=true 且 「(上課)(鐘)」=true

發現在 Bigram 皆可搜尋到 Bi-word，即印出(上課)，再判斷第二個詞:

「(上課)(鐘)」=true 且 「(鐘)(一)」=false.....(1)

在(1)中無法在 Bigram 搜尋到「(鐘)(一)」，因此認為可能有錯誤字產生，隨之進入基本內部的判斷(即針對連續一字詞)。如下:

「ㄓㄨㄥ　一　ㄉㄚˇ」

「ㄓㄨㄥ　一」

根據注音表可找到對應的詞「中醫」後再判斷前半部 Bi-word。

「(上課)(中醫)」=false.....(2)

在轉化後，依然在 Bigram 資訊中搜尋不到如(2)，因此不對詞語做任何更改的動作，即印出(鐘)。以此類推，最後的結果為「上課鐘一打」，不會照成誤判的情況，且對錯誤的字依然仍有效的更正。

3.8.2 系統修正

在 3.7.1 節中，系統只針對每篇文章連續一字詞作處理，但其可能在斷詞時所產生出的結果無法達到預期。如下:

表三、例外狀況

文章	詞語	斷詞
1	「藍球場」	「(藍)(球場)」
2	「藍球」	「(藍)(球)」

再系統中文章 2 所出現的詞語被斷為連續的一字詞，因此會更正為「籃球」，而文章 1 詞語中的(藍)也是錯誤的字詞，但因非斷成一字詞，因此系統不做任何判斷。

為了避免上述這種情況，我們建立成為系統錯字表，其收集系統對每篇文章所改出來的錯別字配對。且再搭配 422 常用錯別字，在系統最後對全部文章做檢測，以避免漏網之魚。

第四章、實驗過程與結果討論

4.1 相似度評分系統

4.1.1 實驗資料

本實驗採用的實驗資料為三所學校之國中二年級學生所撰寫的作文文章，其題目為「下課十分鐘」，這些作文將其輸入成電子檔時保留所有的錯別字以及標點符號，以維持學生所撰寫的原貌。所有的資料共有 689 篇，每篇皆由二到三名的老師所批閱，其分數範圍為一至六級分，再取平均並四捨五入後當作該篇文章的評閱分數。

4.1.2 實驗流程

將所有文章先經過斷詞處理後，再從文章擷取出結構、詞語兩部分的特徵當作系統演算法判斷的依據並分別做相似度評分系統的運作，當系統分別依照兩種特徵進行運作，直到文章分數達到穩態後，最後根據其文章的 Z-Score 分佈區間評定文章六級分的分數。

4.1.3 評鑑方法

本系統之實驗所採取的評鑑方式是針對正確率(Adjacent)以及精確率(Exact)兩項指標當作評鑑系統之效能。

正確率:系統、人工評分之誤差一分內之文章數/文章總數。

精確率:系統、人工評分必須完全相同之文章數/文章總數。

因為不同評閱者的背景知識、主觀認知不盡相同，使得對文章之評分標準也會有所不同。因此本實驗認為誤差一分內皆屬正確之批閱。

4.1.4 實驗結果

其分別針對詞語以及結構上評分，結果如下表所示:

表四、採詞語特徵之系統評分結果表

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	45 篇	29	12	3	1	0	0	91.1%	64.4%
二分	128 篇	20	52	39	16	1	0	86.7%	40.6%
三分	210 篇	2	45	86	67	10	0	94.3%	41.0%
四分	208 篇	0	7	59	107	35	0	96.6%	51.4%
五分	91 篇	0	0	6	44	40	1	93.4%	44.0%
六分	7 篇	0	0	0	2	5	0	71.4%	0%
合計	**689 篇**	**51**	**116**	**193**	**233**	**91**	**1**	**93.0%**	**45.6%**

表五、採結構特徵之系統評分結果表

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	45 篇	16	21	7	1	0	0	82.2%	37.8%
二分	128 篇	12	45	49	17	5	0	82.8%	36.7%
三分	210 篇	4	44	93	60	7	2	93.8%	36.7%
四分	208 篇	2	19	72	77	32	6	87.0%	38.0%
五分	91 篇	0	6	23	29	22	11	68.1%	19.8%
六分	7 篇	0	0	0	4	3	0	42.9%	14.3%
合計	**689 篇**	**44**	**135**	**244**	**188**	**69**	**19**	**85.1%**	**36.7%**

4.2 分數整合評分系統

4.2.1 實驗流程

此部分會使用經由相似度系統分別對兩個面向評分所得到的文章分數、各級分之文章分佈以及文章尚未轉化成六級分之 Z-Score，做為整合評分系統之初始資料。當所有資訊輸入系統進行比例誤差權重的計算，根據系統所得出的最小誤差之比例，進行整合評分，最後再根據文章 Z-Score 分佈區間評定文章六級分的分數。其評鑑方法也是與 4.1.3 小節相同。

4.2.2 實驗結果

由系統得出之最佳比例所進行整合評分，結果如下表所式:

表六、整合評分系統之結果表

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	45 篇	24	17	3	1	0	0	91.1%	53.3%
二分	128 篇	9	59	44	15	1	0	87.5%	46.1%
三分	210 篇	0	45	98	64	3	0	98.6%	46.7%
四分	208 篇	0	8	51	125	24	0	96.2%	60.1%
五分	91 篇	0	0	8	45	37	1	91.2%	40.7%
六分	7 篇	0	0	0	1	6	0	85.7%	0.0%
合計	**689 篇**	**33**	**129**	**214**	**251**	**71**	**1**	**94.2%**	**49.8%**

4.2.3 效能比較與分析

目前各系統之效能如下表所示:

表七、各系統效能比較表

	正確率	精確率
整合評分系統	94.2%	49.8%
詞語相似度系統	93.0%	45.6%
結構相似度系統	85.1%	36.7%
結構系統	82.0%	39.0%
隨機評分	64.1%	23.9%
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%

上述系統主要可分為二種類型:第一種需要一定數量文章做為訓練資料之評分模型。如:貝氏學習機[4]、支援向量機[5]、ID3 決策樹[6]以及結構系統[8]等評分模型。其中貝氏學習機除了原始模型外，另有加入特殊規則的版本。第二種類型即為不需要訓練資料,僅需文章數量達到一定程度方能進行評分之評分模型(即表四中底色為藍色之部分)，其中詞語相似度系統為非監督式評分系統[7]之修改。

由上表可得知，系統在不需要訓練資料下，針對文章兩個面向進行分別評分其效果皆能與使用訓練資料之模型相差不遠。而在整合評分系統上，其正確率及

準確率已超過大部分之模型，僅略差於加入規則的貝式學習機。而且系統批閱之誤差與人工批閱者之誤差已十分接近，其具有一定的可信度，因可以當作人工批閱時之輔助的依據。

4.2.4 文章分批實驗

此部分將針對文章不同資料量來進行整合評分系統之實驗。其主要目的是想得知系統所取出之最佳比例來進行評分是否合理。因此我們將取各級分之文章前100%、90%、80%、70%、60%、50%六種不同的文章數來進行實驗。

4.2.4.1 結果與討論

其分批資料量所得到結果如下表所示:

表八、分批資料量結果表

資料量		詞語	結構	整合	系統比例
100%	689 篇	正確率:0.930 準確率:0.456	正確率:0.851 準確率:0.367	正確率:0.942 準確率:0.498	7:3
90%	620 篇	正確率:0.931 準確率:0.474	正確率:0.858 準確率:0.376	正確率:0.952 準確率:0.495	7:3
80%	551 篇	正確率:0.911 準確率:0.452	正確率:0.853 準確率:0.376	正確率:0.949 準確率:0.481	7:3
70%	483 篇	正確率:0.901 準確率:0.433	正確率:0.845 準確率:0.375	正確率:0.932 準確率:0.468	7:3
60%	414 篇	正確率:0.891 準確率:0.444	正確率:0.860 準確率:0.379	正確率:0.935 準確率:0.454	7:3
50%	343 篇	正確率:0.892 準確率:0.437	正確率:0.866 準確率:0.367	正確率:0.913 準確率:0.449	6:4

表五中其評鑑方式亦是根據 4.1.3 小節相同。由表中可得知系統可根據立意取材及結構組織之部分做最佳比例之分數整合，其效果也達一定的水準。唯有在資料量 50%時，系統所找出的比例為 6:4(表四中底色為藍色之部分所示)，即 60%詞語以及 40%結構，但實際上其最佳比例應為 7:3，但因系統是依據立意取財及結構組織的文章分佈，所算出之誤差權重，因此我們可認為其原因，應為文章資料量過少或者些許文章所得到的誤差權重太大，所造成其結果。

4.3 文章錯別字判斷

我們利用 4.1.1 小節中所提的資料，取各級分前 50%，來做錯別字判斷系統的測試資料。此資料已經由人工的方式找出文章中錯字的部份，其錯別字分類為兩種:第一種同音異字上的錯誤，第二種其餘的錯誤如:多一個字、少一撇、異音異字等錯誤字。

4.3.1 實驗流程與評鑑方式

首先將每篇文章斷詞後，分別輸入原始系統(即尚未修改前之系統)以及修改後之系統進行實驗，其目的是想得知修改過後之系統是否能有效提升正確率。在本系統只針對同音異字做正確率與錯誤率之計算，當作此系統之效能。

正確率:系統與人工得出錯別字之總數/人工得出錯別字之總數。

錯誤率:系統誤判錯字之總數/系統得出錯別字之總數。

4.3.2 實驗結果

其兩個系統所得到之結果如下表所示:

表九、系統尚未修改前之結果表

	結果
人工評定為同音錯別字	286 字
人工評定為其他錯別字	203 字
系統評定為同音錯別字	221 字
系統誤判	4 字
系統之正確率	67.48%
系統之錯誤率	12.66%

表十、系統修改後之結果表

	結果
人工評定為同音錯別字	286 字
人工評定為其他錯別字	203 字
系統評定為同音錯別字	236 字
系統誤判	4 字
系統之正確率	72.72%
系統之錯誤率	11.86%

由表七中可得知，系統經過修正後其效果會有明顯的增加，所得結果足以證明此初階系統之可信度，往後可做為輔助糾正文章錯別字的工具。

第五章、結論與展望

5.1 研究總結

在本論文中，我們所提出的中文寫作多面向評分系統，有別於以往的評分系統，本系統將不再需要事前訓練資料，且不再是以單面向、主觀的方式來給予評分，我們希望在評分上能達到更高的可信度，因此提出了針對多面向來進行評分的方式，根據文章的結構、詞語…等特徵，將各分數加以統整，在上述的實驗中，利用此方法所得之正確率達到 94.2%，證明了多面向整合的評分方式較為客觀，且其實驗結果之整合分數比傳統的評分系統更足以採信；此外，針對錯別字與各特徵都將給予回饋，使用者可藉由此回饋資訊得知在寫作上有哪些需要改進的部份，以提升寫作能力。

5.2 未來工作

本論文目前雖然已針對作文寫作上四個面向當中的立意取材及結構部分上做統整上的評分，且在錯別字判斷中給予回饋的錯誤資訊，但系統尚未包含中文作文評分的全部標準，即少了遣詞造句及錯別字與格式上的分數，因此希望未來能加入這兩個面向上的評分，使系統更能符合人工批閱的模式。

參考文獻

[1] Jill Burstein. “The E-rater Scoring Engine: Automated Essay Scoring With Natural Language Processing.” Automated Essay Scoring: A Cross-Disciplinary Perspective. pp. 113-121, 2003.

[2] 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）

[3] 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）

[4] 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）

[5] 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）

[6] 陳彥宇，「非監督式中文寫作自動評閱系統」，國立交通大學，碩士論文.(2007)

[7] 張道行，「Conceptualization Methodology for Chinese Automatic Essay Scoring」，國立交通大學，博士論文.(2007)

[8] 國中中學學生基本學力測驗推動委員會

URL：http://www.bctest.ntnu.edu.tw/

[9] S. Valenti, F. Neri, and A. Cucchiarelli. “An overview of current research on automated essay grading.” Journal of Information Technology Education, Vol. 2, pp. 319-330, (2003)