

國立交通大學

多媒體工程研究所

碩士論文

自動化中文作文評分與評語回饋系統

Automated Chinese Essay Scoring and Reviews Feedback
System

研究生：林洲銓

指導教授：李嘉晃 教授

中華民國九十八年六月

自動化中文作文評分與評語回饋系統

Automated Chinese Essay Scoring and Reviews Feedback System

研究生：林洲銓 Student：Chou-Chuan Lin

指導教授：李嘉晃 Advisor：Chia-Hoang Li

國立交通大學

多媒體工程研究所



Submitted to Institute of Multimedia 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

中華民國九十八年六月


自動化中文作文評分與評語回饋系統

學生：林洲銓

指導教授：李嘉晃 教授

國立交通大學資訊學院 多媒體工程研究所碩士班

摘要



自動寫作評閱的研究，在自然語言中佔了重要的一環；在中文自動作文評閱研究上，雖然陸陸續續已有相關之研究產生，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學習與評語回饋系統，除了對文章做評分之外，更針對中文寫作評分各個面向，依照使用者輸入的文章，分別給予使用者不同的建議，以便讓使用者了解改進的地方與方向。期望可作為老師批閱或是學生寫作上的輔助工具。

Automated Chinese Essay Scoring and Reviews Feedback System

Student : Chou-Chuan Lin Advisor : Prof. Chia-Hoang Lee

Institute of Multimedia Engineering
College of Computer Science
National Chiao Tung University

Abstract

The research of automated essay scoring is important in natural language processing field. Although some Chinese essay scoring systems have been proposed, all these systems only score the essay in single aspect. They could not provide which aspect the user should enhance to enhance the quality of essay. Thus, we proposed and designed an essay feedback system, which includes an unsupervised learning module for essay scoring and a feedback module for improvement suggestion.

目錄

中文摘要.....	i
英文摘要.....	ii
目錄.....	iii
圖目錄.....	v
表目錄.....	vi
第一章、緒論.....	1
1.1 研究動機.....	1
1.2 研究目的與假設.....	2
1.3 論文架構.....	2
第二章、相關研究.....	3
2.1. 中文系統發展.....	3
2.2. 斷詞與詞性標記.....	4
2.3. 情緒判斷.....	4
2.4. 詞彙的相關程度.....	5
2.5. 非監督式中文寫作自動評閱系統.....	5
第三章、系統設計.....	9
3.1 系統架構.....	9
3.2 前置處理.....	10
3.2.1 作文語料庫斷詞、詞性、結構化.....	10
3.2.2 文章常用詞語組合、高分常用名詞.....	11
3.3 評分系統.....	13
3.4 評語系統.....	14
3.4.1 情緒分析.....	14
3.4.2 分段技巧.....	15
3.4.3 標點符號運用.....	15
3.4.4 注音符號使用.....	16
3.4.5 文章長度.....	16
3.4.6 結構緊密程度.....	16
3.4.7 錯別字判別.....	17
3.4.8 建議補充部份.....	17
3.4.9 佳句讚賞.....	21
3.5 評語整合彙出.....	22
第四章、實驗過程與結果討論.....	23
4.1 評分系統.....	23
4.1.1 實驗資料.....	23

4.1.2 實驗流程	23
4.1.3 評鑑方法	23
4.1.4 實驗結果	24
4.2 評語整合系統	24
4.2.1 實驗流程	24
4.2.2 評鑑方法	24
4.2.3 實驗結果	25
第五章、結論與展望	26
5.1 研究總結	26
5.2 未來工作	26
參考文獻	27



圖目錄

圖 一、評分系統架構圖.....	6
圖 二、投票演算法公式.....	7
圖 三、主要系統架構圖.....	9
圖 四、評分系統結構圖.....	13
圖 五、字句轉換.....	18



表目錄

表 一、各情緒代表詞.....	14
表 二、採詞語特徵之系統評分結果表.....	24
表 三、評語評估結果表.....	25



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) 都曾使用 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。

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)，均已能在評分的準確度上達到不錯的效果，但目前仍少見可針對文章缺失的部份給予適當的評語回饋系統，讓寫作者了解什麼地方是可以再做改進的部份。

1.2 研究目的與假設

本論文之研究目的，在於建立一套可針對作文上不同角度給予建議的評語回饋系統，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的，再利用前半段已完成評分之文章集合，做高分與低分間文章特徵之比對，與一般文章寫作常注意事項的偵測，最後綜合這些文章特徵給予可改進方向，與應注意事項的評語回饋。

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，將簡介近來中文評分系統的發展及相關的研究。

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，首先介紹在中文上系統之發展流程。再來介紹一些相關的研究。

2.1. 中文系統發展

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。陳[6]提出一個非監督式的評分系統構想，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。故葉[8]在 2008 年提出一個針對多面向評分的系統，可分別對立意取材以及結構組織上給予評分，並做出統整得到整合分數。

2.2. 斷詞與詞性標記

斷詞與詞性標記是自然語言處理中基礎且重要的一部份，機器翻譯、資訊擷取、摘要製作及自動作文評分系統等研究都需利用斷詞及詞性標記處理後的結果來進行下一步動作，故斷詞的結果的正確率對研究成果有直接影響。

在中文的句子中，通常不存在有空白這個單元，所以不像英文的句子可以清楚的分隔出單字與單字，所以我們藉助中央研究院的詞庫小組中文斷詞系統[12]來做斷詞與詞性標記的工作，其正確率可達到 95~96%之間 [http://ckipsvr.iis.sinica.edu.tw/]，透過以上的方式，將作文語料庫的文章分隔出字與詞，還有後續工作所需要的詞性。

2.3. 情緒判斷

部落格提供大量具有時間標記的文本，為語言處理所需豐富語料來源。陳信希等三人[11] 針對文本的時間標記特性，將其切分成不同時間域的語料子集合，綜合個別時間域所提供的語料，觀察目標觀點(通常包含意見與情緒)在橫跨時間域的變化，作為觀點趨勢分析的基礎。更利用部落格文本因在網頁呈現方式上所特別能包含的情緒符號，皆用來當作心情或情緒標記，而文本或文句所包含的關鍵字則形成了所謂的特徵值。並以此歸納出喜、怒、哀、樂，四種情緒詞的代表字。

因為一般作文寫作，多以記敘文為主，而此類文章對情緒的抒發是最顯而易見的，所以我們引用其結果，來對文章評判是否在情感上的書寫、描述，有達到表達出情感的起伏。

2.4. 詞彙的相關程度

詞彙的重要程度、詞彙間的相關程度是自然語言處理中的根本問題，在正常的情形下，文件並非僅僅是一系列句子的並排，而是組織完善、有中心意念的文字鋪陳，提供讀者閱讀、欣賞、獲得資訊、或是與作者溝通等等的功能。陳光華 [13] 在 1996 年提出一種以 IDF(Inverse Document Frequency) 為基底，加上各詞彙在文中的相對距離，來運算各辭彙間、詞彙與段落間或段落與段落間的相關程度的方法。

其定義下面兩個式子分別計算名詞與名詞以及名詞與動詞一次共現關係的強度：

$$SNV(N_i, V_j) = IDF(N_i) \times IDF(V_j) / D(N_i, V_j)$$

$$SNN(N_i, N_k) = IDF(N_i) \times IDF(N_k) / D(N_i, N_k)$$

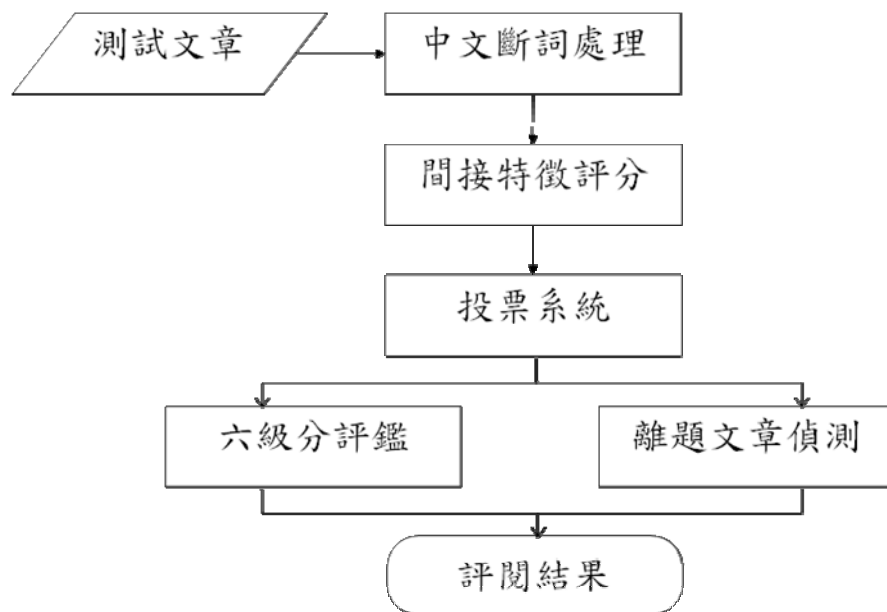
其中 SNV 代表名詞與動詞的共現強度； SNN 則代表名詞與名詞的共現強度；至於 $D(X, Y)$ 表示 X 詞彙與 Y 詞彙之間的距離，目前以 X 與 Y 之間的詞彙數目表示距離； $IDF(A)$ 則是代表詞彙 A 的 IDF 。

2.5. 非監督式中文寫作自動評閱系統

[陳彥宇, 非監督式中文寫作自動評閱系統, 2007] 是本系統採用之評分系統，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。

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。此系統僅探討文章取材方面的優劣程度，未深入文章語意之表現。

下圖為此系統架構圖：



圖一、評分系統架構圖

當大量的測試文章進入系統時，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；接著以相異詞語數量為間接特徵，根據各篇文章的間接特徵給予一初期分數，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，相異詞數為七。若一篇文章相異詞數為100個，則該文章的初始分數即為100；以此間接特徵評分之結果為初始狀態，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，此演算法精神為越相似的文章，分數應該越接近。基本公式如下：

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

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

$S_{j,t}$ ：時間為 t 時，文章 j 的分數(Score)。

$Sim_{i,j}$ ：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。



要計算分數的文章稱為目標文章，其他文章稱為參考文章。基本公式可以分解為三個部份：

- (1) 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。
- (2) $Sim_{i,j}$ ，代表目標文章與參考文章 i 的相似度。
- (3) $Z_{i,j,(t-1)}$ ，代表參考文章 i 的分數與平均值的差異程度。

相似度的算法，則是採用兩文章的共用詞語數做為文章間的相似度。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，反之則給予越多的負分；換言之，所有參考文章均有將目標文章吸引至自身分數的意圖。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。

在此階段評分後，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。



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，用一張系統主架構圖來了解系統整個運作的流程，圖中各個部份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

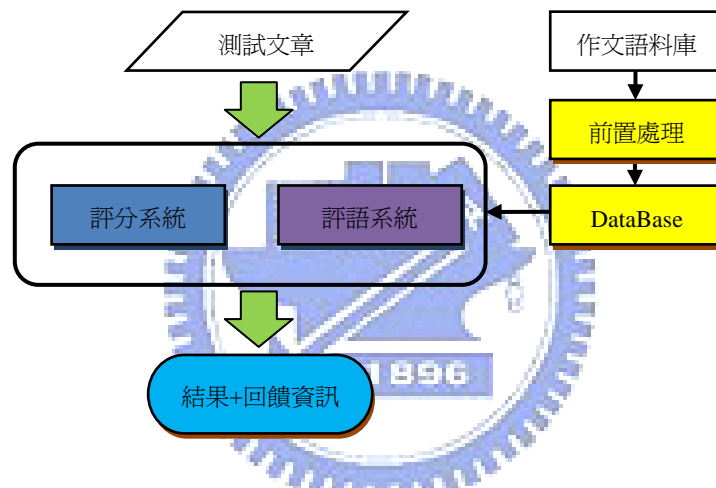


圖 三、主要系統架構圖

本系統裡面共包含 4 個部份：

1. 前置處理：先將作文語料庫做一些規畫與處理，並取出一些評語系統所需的特徵，讓後續系統的進行能更有效率。
2. 評分系統：將測試文章與作文語料庫一起做為評分系統的輸入，放進評分系統得到測試文章的級分評估。
3. 評語系統：針對各評語特徵進行評估與比較，設定各特徵面向的評語結果。
4. 評語整合彙出：將評分系統與評語系統的結果整合輸出。

3.2 前置處理

前置處理的程序主要包含以下兩個部份：第一部份為作文語料庫的斷詞、詞性、結構化，此處是將現有的語料庫的資訊重新匯整成系統所需與可用的資訊，並將這些資訊使用有效率的資料結構方式建構，讓系統在處理後續資料時能更有效率。第二部份為紀錄文章常用詞語組合，以及高分文章常用名詞與頻率。

3.2 1 作文語料庫斷詞、詞性、結構化

文章中，最能代表一篇文章的字詞，就是名詞(N)和動詞(Vt)，所以將作文語料庫內的名詞與動詞全部先取出來做為關鍵字詞的基底，將每篇作文的文章都做處理後，可得到 1 關鍵詞下接 1 填充詞上再接 1 關鍵詞…等等的結構，做好這樣的結構後，在後續的運作上將更方便。如下的例子。

原始文章：

中午(N) 用餐(Vi) 的(T) 時候(N) 都(ADV) 是(Vt) 叫(Vt) 別人(N) 打飯(N) ，(COMMACATEGORY) 打好(Vt) 之(T) 候(N) 之(T) 很多(DET) 男生(N) 在(P) 陽台(N) 上(N) 吃飯(Vi) 聊天(Vi) ……

結構化：

中午(N) 時候(N) 別人(N) 打飯(N) 打好(Vt) 男生(N) 陽台(N)……

用餐(Vi) 的(T)

都(ADV) 是(Vt) 叫(Vt)

，(COMMACATEGORY)

之(T) 候(N) 之(T) 很多(DET)

在(P)

3.2.2 文章常用詞語組合、高分常用名詞

文章分數由一到六級分，評估發現三級分的文章，詞語使用已達一般水準，故文章常用詞語組合的語料庫採用的是三分以上的文章。

首先利用 2.5 節中所提及的評分系統，對整份作文語料庫進行評閱，得到各文章的初閱分數，使用這次評分的結果中三分以上的文章，做為文章常用詞語組合的材料。常用詞語組合分為兩個部份，一個是核心字所構成的組合。有關核心字，將在後續章節介紹，此處僅以作法講解；另一個則是所有提取出的名詞與動詞所構成的組合。

核心字組合：

針對結構化後所提取出來的名詞與動詞，以整份作文語料庫為基底，計算 IDF，

IDF 的定義如下：

$$IDF(A) = \log(P / s(A))$$

其中 P 為所有文章數，s(A) 為含有詞語 A 的文章數目。

然後，將文章以「。」、「！」、「？」區分成各小段落，對照提取出來的詞彙串列，也被切割成幾個部份串列。利用 2.4 節中所提及之方法，計算各串列中詞彙彼此間的關聯程度。此處不必區分兩詞彙的組合，故將公式改寫如下：

$$Correlation(A, B) = \frac{IDF(A) * IDF(B)}{D(A, B)} \quad (1)$$

D(A, B) 的定義為 A 與 B 之間的詞彙數目。



一個詞彙在該段落中的重要性，定義如下：

$$RScore(A_i) = \sum_{i \neq j} Correlation(A_i, A_j) \quad (2)$$

如此，我們可以定義好每個段落中，各詞在該段落的重要程度，接著取出各段落中 RScore 最高的詞，當成是該段落中的核心字。得到一篇文章的核心字串後，以區間為 2 的滑動視窗，從第一個核心字詞開始逐一記錄核心常用詞組合及其出現次數。

範例：

原始字串：

地方 食物 我們 味道 [誘惑]

周圍 環境 氣氛 [引響] 食慾

記得 以前 我們 家人 附近 店家 那裡 氣氛 [羅曼第克] 音樂 周圍 環境 食慾

地方 喜歡 [魚味] 地方 食慾

容忍 環境 地方 [容忍]



核心字串：

誘惑 引響 羅曼第克 魚味 容忍

核心常用詞組合：

[誘惑 引響]、[引響 羅曼第克]、[羅曼第克 魚味]、[魚味 容忍]

一般常用詞組合：

以 3 級分以上文章，提取出的名詞與動詞為基底，同樣設定一個區間為 2 的滑動視窗，由第一個開始逐一紀錄常用詞組合。原始字串如上例，則常用字串如下：

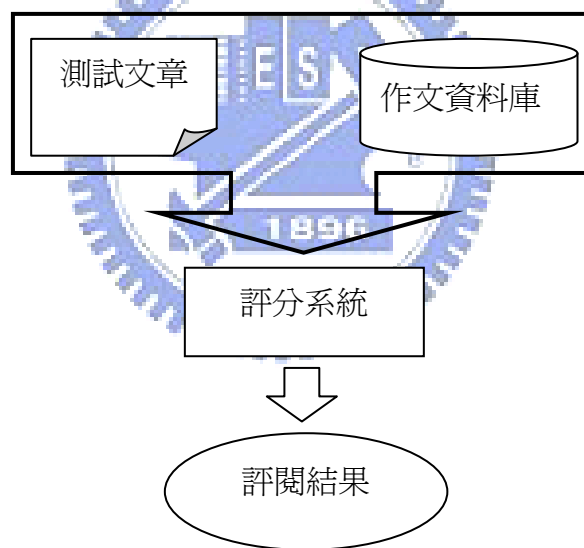
[地方 食物]、[食物 我們]、[我們 味道]……。

高分文章常用名詞提取：

先將六級分文章分成三個部份，第一部份為低分群(1~2分)，第二部份為高分群(5~6分)，全部文章為第三部份(1~6分)。對低分群與高分群的文章，分別統計名詞的出現次數，以兩群中較小的最高出現次數為門檻值，對第三部份提取出高於門檻值的名詞成為 stoplist。接著利用 stoplist 對第二部份做篩選把沒在 stoplist 中出現的詞保留起來，成為高分文章常用名詞。

3.3 評分系統

評分系統採用陳彥宇[6]在2007年提出的非監督式自動評閱系統，來進行評分的動作，因其無法以單篇文章為輸入得到評閱結果，故將原結構稍做修改如下：



圖四、評分系統結構圖

先將待測文章與作文語料庫合併，做為評分系統的輸入文集，以合併後的文集進行投票演算法互相評分直到穩定，以此方式得到各文章的評閱結果。因為使用固定的作文語料庫來合併，對於投票演算法的運行，每篇待測文章都會以同樣的作文語料庫給予評分，因此在評分的標準上，可稱為是同一標準的。

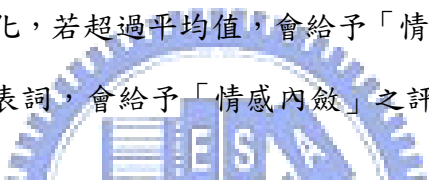
3.4 評語系統

評語系統，一共分成九個不同的面向，來對待測文章進行評估，包含了情緒分析、分段技巧、標點運用、佳句讚賞、錯別字…等九項，將在後續小節逐一介紹。

3.4.1 情緒分析

由於一般中學作文多數為記敘文，在情感流露上，可較為顯見。因此利用陳信希[11]所得到的各情緒代表詞來做為評估基準。

以下表中各情緒代表詞為資料，統計待測文章中使用的情緒詞總個數，再除以待測文章長度做正規化，若超過平均值，會給予「情感豐富」之評語；而使用較少或無使用到情緒代表詞，會給予「情感內斂」之評語。



情緒	代表詞
喜	愛、幸福、可愛、喜歡、謝謝、害羞、感動、寶貝
怒	生氣、討厭、氣死、可惡、幹、煩、罵、哼
哀	哭、痛、鳴、難過、淚、傷、慘、可憐
樂	哈哈、開心、好笑、高興、不錯、加油、很好、好玩

表一、各情緒代表詞

3.4.2 分段技巧

作文常見分段寫作技巧為起、承、轉、合，四個部份。一般一個部份會分配一個段落來負責敘述，也可能因版面問題而將承與轉兩段合併。因此，在寫作分段技巧方面，我們判定若文章段落少於三段，或大於五段，則表示在分段技巧上，需多加強改進。

作文語料庫在建立時，已有特別處理過，在文章的每個段落前標記有「**」。所以以此符號數量為基準，小於三，大於五，均判定為分段技巧有待加強，會分別給予「段落過少」、「段落過多」的評語。

3.4.3 標點符號運用

此面向主要針對段落數為 1，而長度未低於平均文章長度的一半的文章做回饋，假設為標點符號使用不當，造成應分段而未分段的結果。

首先，利用新酷音輸入法的詞庫，統計出常用詞語的使用頻率，可以得到如下格式的資訊：



台北市:48552	到:65361
間:49543	一個:65753
有限公司:49714	...
中心:50424	的:147270
留言:50546	中:148655
簡介:51254	有:153094
時間:51398	我:160394

經觀察發現，設定門檻值 50000 可約略區隔出常用詞語 stoplist。

接著，使用中研院平衡語料庫[14]，統計句號前的常出現詞彙，輔以常用詞語 stoplist，歸納出句號前常用詞彙集。

對於段落數為 1 的文章，系統會進一步搜尋在文章中是否出現句號常用詞彙集中的詞語，除了給予「標點符號使用不佳」的評語外，還加上建議可分段處的位置，讓使用者做為參考。

3.4.4 注音符號使用

經觀察「下課十分鐘」、「用餐時分」兩份作文語料庫發現，仍有少部份學生在寫文章時，會使用注音符號來表示不會寫的中文文字。

範例：

有肉有采 我看了就想吃 可是要進去火ㄍㄨㄛ 裡面 煮一煮才可以吃

在這一部份，系統也將給予提醒的回饋，利用注音符號表，檢查待測文章中是否出現注音符號，若有出現，則會給予「盡量減少使用注音符號」的評語。



3.4.5 文章長度

常見中學作文長度為 600 字，在此，假設三級分以上文章，乃因為內容已達一定水準，所以可以得到三級分以上的成績。因此，以三級分以上的文章平均長度，做為文章長度是否足夠的判別條件，若待測文章長度小於平均長度，則判定長度不足，系統會給予「文章字數過少」的評語。

3.4.6 結構緊密程度

若待測文章長度小於平均長度，且文章得分高於 4 分，代表此文章雖然字數不足，但已包含高分文章所需求的要素，即是以簡單扼要的寫作方式，完成了這篇文章，因此，雖然結構已完整，卻缺乏詳細的敘述，系統會判定為文章結構過於緊密，並給予「文章結構緊密」的評語。

3.4.7 錯別字判別

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。本系統僅利用自行整理的常見錯別字表(以下簡稱錯字表)來做錯別字的判別。錯字表格式如下：

檢只	簡直
幽美	優美
怖製	佈置
...	...

待測文章已是斷詞之後的結果，所以檢查時，即是逐一搜尋錯字表中錯誤部份(左半部)，是否有在待測文章斷詞結果中以詞語出現，最後統計其錯字個數。

範例：這地方的風景，真是太幽美了，檢只是人間仙境。

檢查時，即是在待測文章中搜尋「檢只」、「幽美」、…，並記錄個數。

此例之錯字數為2。



3.4.8 建議補充部份

一句話中，能代表所要闡述的意思的詞語，多數為名詞與動詞。其中可能有一個更是整句話的核心，剩餘的名詞與動詞僅是補充敘述。所以，我們把文章轉換成關鍵字串的方式來處理後續所需的資訊。

首先，從結構化後的待測文章，我們已經可以得到文章的關鍵字串，接著以「。」、「？」、「！」將文章及相對應的關鍵字串切割成小段落，並利用式子(1)(2)來計算各段落中的核心字，如此可以得到待測文章的關鍵字串及其核心字串列。如圖四所示，右邊部份藍色字即為該句子的核心字詞。

吃飯是美天都一定要做的事情，大多數的人每餐都是隨便吃吃，有些實就好了，不會太在意其他的任何觀點，但是有些人就滿在意要在哪裡吃，比較有氣份。* 平時的我，通常是在家中吃飯，雖然在家中吃飯，會感受不到在外吃飯的感覺，可是在家中吃飯還可以一邊看電視一邊吃飯，也是別有一番風味。

美天 事情 吃吃 在意 觀點 在意 哪裡
===
平時 感受 感覺 電視 別有 風味

圖 五、字句轉換

核心字代表著每個句子的重點，而核心字所成的串列，則可代表整篇文章的敘述流程。因此，可藉由觀察核心字串的長度，來了解文章是否已表達完整的流程。建議補充的部份，即是以此為立足點來給予適當的回饋。

以三級分以上文章的核心字串平均長度為基準，對於不足此長度的文章，裁定為文章表達不夠完整，系統會針對關鍵字串做補充的動作，再根據系統新增的地方，挑選出適當的部份來給予回饋。補充的程序，分為三個部份：第一部份為填充核心字串；第二部份為針對新增核心字，擴充關鍵字串列；第三部份為挑選欲給予回饋的串列段落與提示句；以下將詳細介紹各部份內容。

首先，第一部份先利用現有的核心字串，加上前置處理好的核心字常見組合（以下稱核心 bigram）來填充核心字串。填充方式為在每個核心字後面，放入新增的核心字，若不為最後一個核心字，則以其本身及後面一個核心字為一組當成索引，否則僅以其本身為索引，記錄核心 bigram 中前核心字後面，及後核心字前面接的核心字，挑選出最常出現的詞彙，填入兩核心字中做為新增核心字詞。若在核心 bigram 中無法找到常用組合則不做填充的動作，或是找到的最常使用詞出現次數超過全候選字總合的一半，且此時核心字後面接的為此最常使用詞，亦不做填充的動作。

完成一次之後，若長度仍未大於平均長度，則以填充後核心字串，再做一次相同的動作，直接超過平均長度。

範例一：

原始核心字串：

[觀點] [別有]

第一次補充：

針對[觀點]一詞，在核心 bigram 中尋找{[觀點] [A]}、{[B] [別有]}此兩種組合並記錄下 A、B 詞與其出現次數，最後挑選最高出現次數的詞做為填充核心字。接著處理最後一個核心字[別有]，在核心 bigram 中尋找{[別有] [C]}記錄 C 詞及出現次數，同樣取出最高次數的詞來填充。

假設結果為 [觀點] [專題] [別有] [餐廳]。

因為第一次結果未超過平均長度，所以第二次補充即以第一次結果再做一次核心字填充處理，直到完成後結果大於平均長度。



範例二：

原始核心字串：

[溫暖] [做好] [東西] [時候]

第一次補充：

分別針對[溫暖]，在核心 bigram 中查詢{[溫暖] [A]}、{[B] [做好]}此兩種組合並記錄下 A、B 詞與其出現次數，挑選出最高出現次數的詞做為填充核心字；接著針對[做好]、[東西]做相同的處理；[時候]則以最後核心字做另外的處理。補充後結果為[溫暖] [屬於] [做好] [範圍] [東西] [規定] [時候] [那裡]

此結果已超過平均長度，故不需再做第二次的處理。

第二部份為對於已新增之核心字串，利用前處理的一般常用詞組合(以下稱常用 bigram)擴充其前後關鍵字，形成新的一段之關鍵字串列。經計算，平均關鍵字串的長度為 9，故針對每一個新增的字串，最多將擴充到 9 個關鍵字。擴充方式如下：

1. 隨機決定核心字在新字串中的位置。
2. 以核心字為索引，查詢常用 bigram 中核心字前方常出現詞彙，並記錄出現次數，挑選最常出現的詞做為擴充關鍵字，若挑選出來的詞與現有詞相同時，將不做擴充；或常用 bigram 中無法找到擴充關鍵字候選詞時亦然。
3. 以核心字為索引，查詢方式改為核心字後方常出現詞彙，同樣的擴充方法將核心字後方關鍵字串填充完。

範例：

新增核心字：[規定]

假設位置決定為 4，表示核心字前方需擴充 3 個，後方 5 個關鍵字。先對前方做擴充的動作，在常用 bigram 中查詢 {[A][規定]} 的出現次數，並把 A 詞彙記錄下來，最後取出最常出現的詞，與[規定]比對是否相同，不同方將詞加入[規定]前方。假定尋找到的詞彙為[餐廳]，則現有字串更新為[餐廳][規定]。接著以[餐廳]為索引，在常用 bigram 中查詢 {[B][餐廳]} 的出現次數，以此方法逐一擴充，直到挑出的詞與索引詞相同，或是已達到欲擴充詞數為止。後方關鍵字的擴充方式與前方相似，僅改變查詢的索引詞組為 {[索引字][候選字]}。

此例結果可擴充為[東西][覺得][外面][餐廳][規定][喜歡][感覺]。

經過第二部份的擴充，可以得到完整的關鍵字串列表，第三部份將由此結果中，挑選出一段做為建議回饋的依據，來給予使用者可補充內容的資訊。首先，會對每個新增的關鍵字串，再做一次各關鍵字間的關係程度計算，重新擷取出核心字，並以核心字在各段落中與其他關鍵字關係程度的分數，做為依據取出最高

的一個，來做為建議方針。並以該段落中所有的關鍵字做為 term vector 與三級分以上，未超過 100 個字的句子做相似度計算，取出相似度最高的一句做為提示句。相似度的計算方式採用 cosine 相似度算法，範例如下：

關鍵字串列：[天氣] [郊遊] [野餐] [傍晚]

備選句：我們 昨天 傍晚 才 去 了 郊外 野餐 ，現在 好 累。

備選句轉換成向量之後為 $\langle 0, 0, 1, 1 \rangle$ ，與本來的 term vector $\langle 1, 1, 1, 1 \rangle$ 做 cosine 運算。利用 $\cos(A, B) = A \cdot B / |A| |B|$ ，得到此備選句與關鍵字串列的相似度為 $2/2\sqrt{2} = 1/\sqrt{2}$ 。

3.4.9 佳句讚賞

先將前置處理時紀錄的高分常用詞語設定為優美詞彙，並依其使用頻率給予權重。然後對所有文章「。」、「？」、「！」，切割成小句子，以句子為單位，根據使用到的優美詞彙加總其權重，由此得出優美詞句的平均權重值。

接著將待測文章以相同的方法加總其權重。最後挑選出待測文章中高於平均權重，且權重值最高的句子，標示為佳句並列出以供參考。

計算範例如下：

優美詞集 氣氛(0.028) 感覺(0.0025) 全部句子權重平均值(0.03)

句子1：這家餐廳氣氛很棒。

句子1 權重：0.028 < 0.03

句子2：這家餐廳氣氛很好，令人感覺非常的愉快。

句子2 權重：0.028 + 0.0025 = 0.0305 > 0.03

結果：第二句詞語使用佳。

3.5 評語整合彙出

經由評語系統各面向評估後，可以得到待測文章在各方面的回饋，接下來將對回饋做整合，以較完整的方式呈現給使用者。

在各面向中，有部份評估要素是有從屬關係的，如：文章長度與結構緊密程度。因此在整合時，會將上位的評估面向所得到的回饋去除，例如：待測文章長度過短，且得分高於4級分，在評估時，將會得到[文章長度過短]及[文章結構緊密]兩項結果，因為結構緊密程度條件較文章長度嚴謹，即文章長度為結構緊密程度的上位評估面向，所以會將[文章長度過短]此評語去除。

另外，在補充描述部份的回饋，系統將再做一次篩選，方法如下：

1. 若建議方針有在常用名詞表中出現，則建議方針不做更動。
2. 否則，將建議方針去除，改以[可針對主題多加發揮]為建議方向。

這樣做的概念為常用名詞表含有高分文章常用詞彙，而名詞常常是文章能夾帶較多資訊的詞語，所以建議方針若不在常用名詞表中，也間接代表著高分文章並不常出現這樣的寫法，故將方針改為以主題多加發揮。

以上述方式修正評語後，將得到如下方所示最後評語。

評語結果：

情感豐富。文章長度稍短，針對主題多加發揮，可以使文章內容更加豐富。段落安排不恰當，在 [# 報正你去了一次後會讓你有生難忘 還想在去一次 #] 附近可善用標點符號分段。錯別字過多，應多加小心。

第四章、 實驗過程與結果討論

4.1 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「用餐時分」，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 694 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做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，計算各文章的相異詞語數做為文章的初始分數，輸入系統演算法運作，直到文章分數達到穩定後，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

結果如下表所示：

表 二、採詞語特徵之系統評分結果表

系統評分		零分	一分	二分	三分	四分	五分	六分	正確率	精確率
人工評分		零分	一分	二分	三分	四分	五分	六分		
一分	14 篇	7	6	0	1	0	0	0	92.86%	42.86%
二分	34 篇	4	10	5	12	2	1	0	79.41%	14.71%
三分	139 篇	3	8	19	43	59	6	1	87.05%	30.94%
四分	329 篇	3	1	10	55	173	59	28	87.23%	52.58%
五分	140 篇	2	0	0	7	64	49	18	93.57%	35.00%
六分	38 篇	0	0	0	1	12	16	9	65.79%	23.68%
合計	694 篇	19	25	34	119	310	131	56	84.32%	33.30%

4.2 評語整合系統

4.2.1 實驗流程

從作文資料庫中選取三十篇含一到六級分的文章，做為系統的輸入文章，利用評語整合系統得到各篇文章的評語回饋，並請六位使用者幫忙評估系統評語回饋的實用性。

4.2.2 評鑑方法

針對評語通順程度、評語面向準確程度及評語建議良好程度三個方面來做為評估的基準，評語通順程度為評語回饋的語句通順程度；評語面向準確程度為系

統給予的面向是否準確；評語建議良好程度為各面向評語所給予的建議，是否有切中要點。程度的基準分為五個等級，分別以 1~5 來表示，1 為最低，最不準確；5 為最高。最後將六人在各面向評估大於 3 與大於 4 的比例計算出平均值。

4.2.3 實驗結果

表 三、評語評估結果表

結果 \ 面向	通 順 度	準 確 度	良 好 度
平均	4.53	3.81	3.77
>=3	99.33%	95.33%	95.33%
>=4	85.33%	66.00%	59.33%

由結果可以看出，在 3 以上的比例，已有九成以上的滿意度；在 4 以上，在準確度與良好度方面仍可以達到約六成的滿意度。

第五章、結論與展望

5.1 研究總結

在本論文中，我們所提出的中文作文評分及評語系統，有別於以往的評分系統，本系統除給予評估之後的分數外，更針對文章各方面的缺失及優點，給予建議與回饋，讓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各方面能給予建議與改進的回饋，但系統尚未包含中文作文評分的全部標準，仍有未能含括到的部份，如對文章結構上的評論、文意上的探討等，因此希望未來能加入這些方面的考量，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] 葉啟祥,「中文寫作多面向評分系統」, 國立交通大學, 碩士論文.(2008)
- [9] 國中中學學生基本學力測驗推動委員會
URL : <http://www.bctest.ntnu.edu.tw/>
- [10] 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)
- [11] 陳信希、楊昌樺、高虹安,「以部落格語料進行情緒趨勢分析」第十九屆自然語言與語音處理研討會論文集, 2007年九月6-7日, 台灣, 台北, 205-218.

- [12] 中央研究院資訊科學研究所詞庫小組中文斷詞系統
URL : <http://ckipsvr.iis.sinica.edu.tw/>
- [13] 陳光華、陳信希，「文件內容之分析— 語料庫為本的模型」，圖書館學刊，page 95-112. (1996)
- [14] 中研院平衡語料庫 3.1 版

