國立交通大學

資訊科學系

碩 士 論 文

自動建構中文作文評分系統：

產生、篩選與評估

Automated Chinese Essay Scoring System：

Generation、Selection、Evaluation

研 究 生：蔡沛言

指導教授：李嘉晃　教授

中 華 民 國 九 十 四 年 六 月

自動建構中文作文評分系統：產生、篩選、評估

Automated Chinese Essay Scoring System：Generation、Selection、Evaluation

研 究 生：蔡沛言　　Student：Pei-Yan Tsai

指導教授：李嘉晃　　Advisor：Chia-Hoang Lee

國 立 交 通 大 學

資 訊 科 學 系

碩 士 論 文

A Thesis

Submitted to Department of Computer and Information Science

College of Electrical Engineering and Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

In

Computer and Information Science

June 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年六月

自動建構中文作文評分系統：

產生、篩選與評估

學生 ： 蔡沛言　　指導教授：李嘉晃 博士

國立交通大學電機資訊學院　資訊科學研究所碩士班

中文摘要

在本論文中，我們探討概念(Concept)與寫作之間的關係，並以此作為基礎自動建立一套中文作文評分系統。本系統以概念為主要特徵(feature)，系統建立的過程共涉及三個模組：1.Generation-特徵產生 2.Selection-集合的篩選 3.Evaluation-特徵評估。

本系統以知網(Hownet)作為概念空間，並將作文的用詞對應到知網中予以概念化。其次，我們設計一個方法從概念空間中自動有效率的篩選出其中的子集合，並且設計了一個評估函數來計算該集合的鑑別力。根據實驗結果，本系統評閱的正確率可達 92%~95%，可作為協助老師批閱作文時的工具。

Automated Chinese Essay Scoring System：Generation、Selection、Evaluation

Student ： Pei-Yan Tsai　　Advisor：Prof. Chia-Hoang Lee

Department of Computer and Information Science
National Chiao Tung University

Abstract

This thesis explores the relationship between concepts of words and essay writings. Based on their relationship, an automated scoring system can be automatically constructed. The system consists of such three models as Generation model, Selection model, Evaluation model.

The system first transforms essays into the concept space defined by Hownet. Next we develop an efficient method to select subset of concept space to be used for feature subset. Lastly, an evaluation function is designed to evaluate the effectiveness of the subset for grading essays. The experimental results show that our system has achieved 90~95% correctness, it can greatly facilitate human grader.

目錄

圖目錄

表目錄

第一章、緒論

1.1 研究動機

作文的重要性，在國語文能力的培養上，佔了關鍵的地位，因此在求學的各階段中，均相當重視寫作能力的訓練。而除了寫作的語言能力養成外，作文亦可以激發思維的能力。在日常生活中，作文亦扮演著相當重要的工具，舉凡書信、卡片、履歷表、演講稿、公文、報告、婚喪喜慶等文件，都可算是作文的範疇。藉由作文的寫作訓練，可使學生更加理解並靈活應用語言與文字，並訓練學生的思考、理解、推理、創作等能力。

大量的作文批閱，是一個相當耗費時間，也耗費人力的工作。除了耗費的人力和時間之外，維持批閱的公平性也是一大問題，由於需批閱作文的數量龐大，必須由許多老師負責批閱，因此，維持所有作文評分標準達到一致，顯得更加的不易。因此，本研究嘗試建立一套自動化的系統，用來協助閱卷的工作，並且可以解決人工閱卷所造成的評分標準不一致的問題。

英文的自然語言處理研究中， Automated Essay Scoring (AES)系統的發展已有悠久的歷史[1]。在 1999 年 2 月，Analytical Writing Assessment GMAT 已開始使用 e-rater 來協助文章的評分工作[2]。然而，在中文方面，一直沒有出現相關的研究，因此，本論文嘗試建立一套中文的 AES 系統。

1.2 研究目的與構想

本論文的研究目的，就是嘗試建立出一套中文的 Automated Essay Scoring (AES)系統，希望此系統可以在不需要人為介入下，從訓練資料(Training Data，即評閱過的作文)中，完全自動的產生一組特徵(feature)。本系統可直接利用此組特徵對作文評閱，擁有相當高的正確率，此組特徵亦可提供給其他不同的 AES 系統，協助提升其系統效能、評閱的正確率。

1.3 論文架構

第一章為前言，內容主要是介紹本論文的研究動機以及研究目的，提到了作文的重要性以及中文的自然語言處理。第二章為相關的研究，內容包括現今使用的 AES (Automated Essay Scoring)系統，知網(Hownet)，並探討關於寫作基礎的假設。第三章則是本研究最核心的系統介紹，詳細的介紹了整個系統的架構及說明。第四章則是對此研究的實作過程與研究結果報告做討論，包括此系統批閱及正確率，並且用流程圖表現出此系統實驗的先後順序。最後第五章則是描述本論文的研究總結，以及未來尚需繼續研究的工作。

第二章、相關研究與構想

在本章節中，將詳細的介紹本論文提及的相關研究與構想，首先開始先介紹了傳統上的自動作文評分(Automated Essay Scoring，簡稱AES)系統，之後利用知網(Hownet)引進「概念(Concept)」，探討概念(Concept)在寫作中所代表的涵義，以及利用概念來幫助評閱作文的想法。本章的章節安排如下，在2.1節，簡單的介紹傳統的Automated Essay Scoring (AES)系統。在2.2節，介紹了知網(HowNet)，包含知網的結構以及知識描述語言。在2.3節，探討寫作基礎的假設，以及建立系統的想法。最後，在2.4節中，介紹自然語言處理中，有關中文處理的一個問題，中文的斷詞。

2.1 傳統Automated Essay Scoring(AES) 的方法

Automated Essay Scoring (之後簡稱AES)系統，指的是一個自動的作文評分系統，可以使用電腦來評閱作文，並能給予該作文批閱後的分數。一般傳統上AES系統建立的方法，首先，收集由專業批閱人員批閱過的作文當做訓練資料，接著，AES系統設計者從作文中找尋出特徵，比如說，作文字數、標點符號、形容詞數…等，當建立起所有這些找出來的特徵之後，接著使用機器學習(Machine Learning)的方法(如Bayesian Algorithm：***Bayesian Essay Test Scoring sYstem (BETSY)***[3]；或是Neural Network：***Intelligent Essay Marking Systems (IEMS)*** [4]，比較詳細的介紹請參閱論文[5])。這是根據隱含在作文中的特徵，並利用機器學習的演算法，期望能從中學習出專業閱卷人員的行為，經過訓練、學習完成之後，此系統便可以自動的批閱作文，此為一般傳統上AES的方法，其流程圖如下：

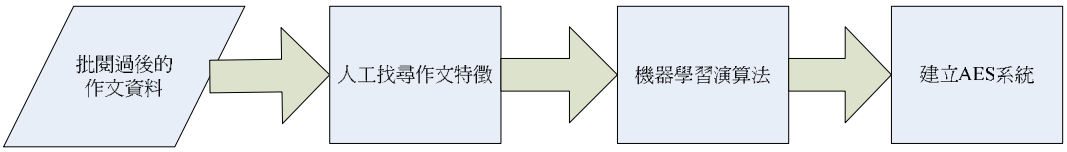

圖 1

2.2 知網(Hownet)

知網是一個包括漢語和英語詞語的知識庫[6]。該知識庫的詞語代表描述物件的概念，並且用以揭示概念與概念之間以及概念所具有的屬性之間的關係。1988 年前後，董振東曾在他的幾篇文章中提出以下的觀點[7]：

(1) 自然語言處理系統最終需要更強大的知識庫的支援。

(2) 關於什麼是知識，尤其是關於什麼是電腦可處理的知識，他提出：知識是一個系統，是一個包含著各種概念與概念之間的關係，以及概念的屬性與屬性之間的關係的系統。一個人比另外一個人有更多的知識是代表他不僅掌握了更多的概念，尤其重要的是他掌握了更多的概念之間的關係以及概念的屬性與屬性之間的關係。

《知網》中兩個主要的概念：「概念」與「義原」。「概念」是對辭彙語義的一種描述，一個詞可以表達好幾個概念。「概念」是用一種「知識表示語言」來描述的，這種「知識表示語言」所用的「辭彙」叫做「義原」。「義原」是用於描述一個「概念」的最小意義單位。

「義原」一方面作為描述「概念」的最基本單位，另一方面，「義原」之間又存在複雜的關係。在《知網》中，一共描述了義原之間的 8 種關係：上下位關係、同義關係、反義關係、對義關係、屬性-宿主關係、部件-整體關係、材料-成品關係、事件-角色關係。可以看出，義原之間組成的是一個複雜的網狀結構，而不是一個單純的樹狀結構。

《知網》中，每一個概念用一個記錄來表示，每筆紀錄共包含下列八個項目：

NO. ： 概念編號
W_C ： 中文的詞
G_C ： 中文的詞性

E_C ： 中文的例子
W_E ： 英文的詞
G_E ： 英文詞性
E_E ： 英文例子
DEF ： 知網對於該概念的定義，我們稱之為一個語義運算式。
DEF 是知網的核心。

一個詞可能有多個描述式，但其中的第一個描述式是對該詞最重要的一個描述式。該描述式呈現了該詞最基本的語義特徵，本系統以每個詞的第一個描述式中的第一個義原作為代表該詞的概念。

在底下的例子一，中文詞「學校」其所代表的概念為 { InstitutePlacel場所:domain={educationl教育},{studyl學習:location={~}},{teachl教:location={~}}}。由 DEF 的內容可看出知網一共用了四個義原來解釋這個概念，分別是 {InstitutePlace|場所}、{education|教育}、{study|學習}、{teach|教:}。本系統採用第一個義原即{InstitutePlace|場所}來表示「學校」這個詞所代表的概念。

在底下的例子二，中文詞「漂亮」所代表的概念爲{beautifull美}。知網只用一個義原{beautifull美}來解釋這個概念，因此我們用{beautifull美}這個義原表示「漂亮」這個詞所代表的概念。其它的中文詞，如「美麗」、「好看」等中文詞其義原皆為{beautifull美}，本研究假設這些詞皆代表著相同的概念。

例子一：

NO.=109046
W_C=學校
G_C=N
E_C=
W_E=school
G_E=N
E_E=
DEF = { InstitutePlacel場所:domain={educationl教育},{studyl學習:location={~}},{teachl教:location={~}}}

例子二：

NO.=073724
W_C=漂亮
G_C=ADJ
E_C=~姑娘，打扮得很~，~臉蛋兒，她一點也不~，~的衣服
W_E=beautiful
G_E=ADJ
E_E=
DEF={beautifull美}

2.3 寫作基礎的假設

從寫作者的角度來觀察，我們認為，學生寫作文時，是先選擇了很多不同的概念(Concept)，然後在撰寫作文的過程中，再把這些不同的概念組織起來。仔細點的說，當寫作者開始寫作，有了作文的題目(Topic)後，便開始構思在作文中要使用哪些概念，決定好之後，再開始用各種不同的寫作技巧、寫作手法來組織這些概念，最後，決定一篇作文好壞的因素，便是根據這些概念、寫作技巧、寫作手法…等等很多不同的因素，而本研究所特別注重的，就是在於這些概念(Concept)的使用。

另一方面，評閱者批閱作文的依據是相當複雜的，其中涵蓋的層面及因素相當廣，關於閱卷者個人方面，其主觀認知、所學背景、批閱習慣，這些都影響批閱結果，而關於作文方面，對於作文的理解，通常主要是以該文是否符合主題、句法是否通順、文章的完整性、組織結構、修辭用字，而這些和寫作者概念的選擇、組織都有密切的關聯。由於要建立出一套系統來完美的捕捉批閱者的行為模式，從而完全的學習批閱者批閱的方式，是相當困難，所以，本論文的焦點在於探討概念(Concept)是否能提供有效的模型給評分系統，批閱的分數也能接近閱卷者的等級。

統計方面來看，在全部的作文資料庫中，我們對每篇作文統計其所使用的概念數量(即義原數量)，並觀察作文所出現的「概念次數」和批閱者對其評分的「分數等級」，兩者之間是否存在任何關聯。圖二中，橫軸表示作文的分數等級，縱軸表示該等級作文所使用的平均概念數。例如，三分的作文有 234 篇，平均每篇使用的概念數為 77 個。

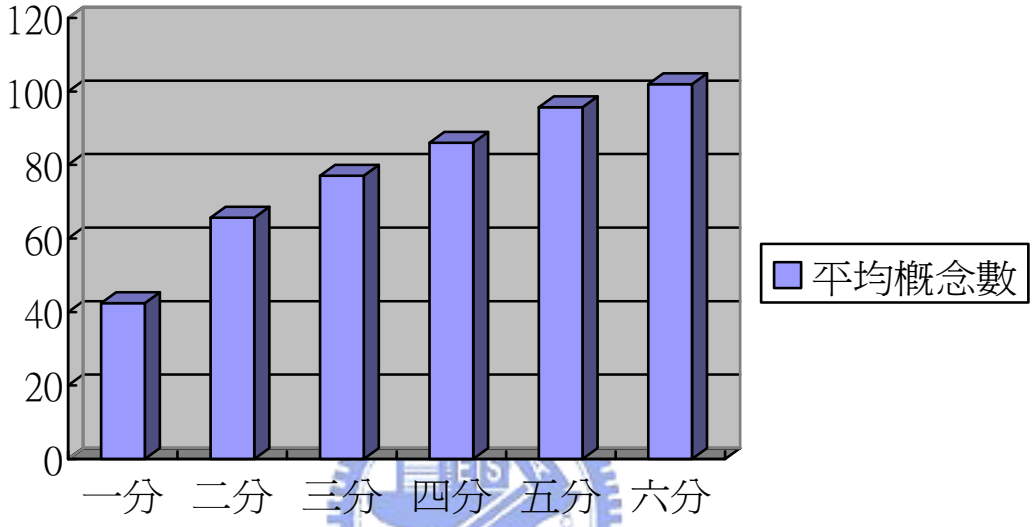

圖 2

從圖二可以觀察到一個現象，其分布圖呈現作文分數等級和其平均概念數有明顯的遞增關係。換句話說，分數等級愈高的作文，其平均所使用的概念數目也愈多。此外，若再進一步分析，在每個等級作文中，找出其中使用概念數量最多的作文以及使用概念數量最少的作文，可以得到底下的圖三。例如在等級二分的 166 篇作文當中，其中使用最多概念數量的一篇作文是總共用了 104 個概念。

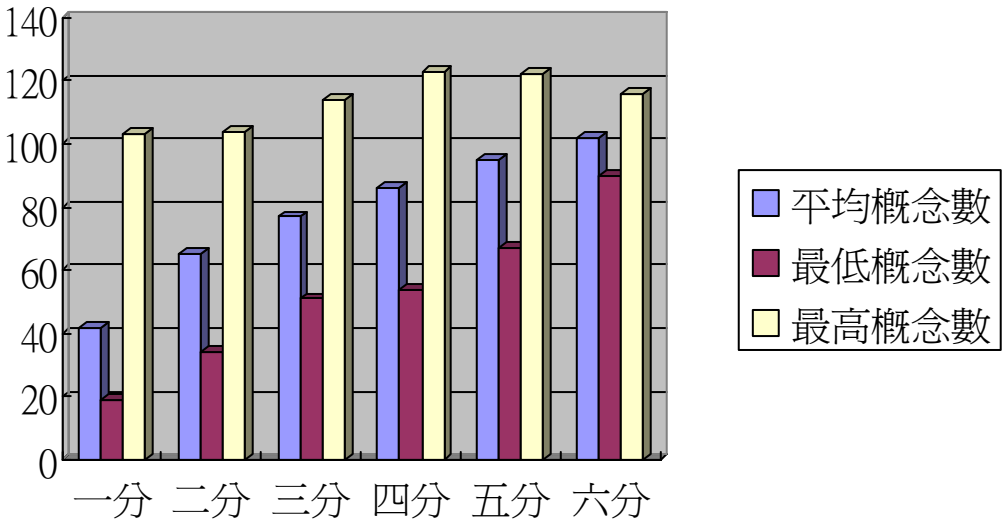

圖 3

從圖二可以觀察到，其分布圖呈現作文分數等級和其平均概念數有明顯的遞增關係。但是從圖三的分布圖可以發現，若只是單純的根據概念的使用數量來評閱作文，結果可能是會相當差的。因為在不同作文分數等級中，都存在概念數量超過 100 的文章，若只是根據概念的使用數量來評閱作文，這些作文皆視為很高分的作文；在六分等級的作文中，存在一篇概念只使用概念數量 90 個的作文，還少於等級五分的平均概念數 95 個。因此，在所有的概念空間中，並不是所有的概念皆是有用的，必須從中挑選出具有鑑別能力的概念子集合。

根據這個觀察，本論文嘗試探討在整個概念(Concept)空間中是否存在有效的概念子集合，能夠用來鑑別出不同等級的作文。這樣子的課題牽涉到概念的產生、子集合的篩選以及如何使用此集合來批閱作文。

2.4 中文斷詞(Word Segmentation)

詞是最小有意義且可以自由使用的語言單位。因此，任何一個語言處理的系統，都必須先能分辨文章中的各個詞才能進行詞性標記、語言分析、資訊擷取等進一步的處理，因此中文自動分詞的工作成了語言處理不可或缺的技術。自然語言處理中，中文與英文最顯而易見的差異，在於中文的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。

在此用一個簡單的中文句子來解釋中文的斷詞：

「今天天氣很好」

而這個句子可能的斷詞有下列數種：

「（今）（天天）（氣很）（好）」……(1)

「（今天）（天氣）（很好）」……(2)

「（今）（天天）（氣）（很好）」……(3)

「（今）（天天）（氣很）（好）」……(4)

「（今天天）（氣很好）」……(5)

……

「今天天氣很好」這句話正確的斷詞為第二句

「(今天）（天氣）（很好）」

其它句的斷詞會造成語意不正確，語法上沒有代表意義。

第三章、系統設計

在本章節中，詳細的介紹整個系統的架構及方法說明，在 3.1 節中，用一個系統架構圖來幫助了解系統執行的架構，架構圖中包含系統的 3 個主要模組，1.Generation - 概念產生、2.Selection -集合選取、3.Evaluation - 評估方式。接著分別用三個小節仔細的說明這三個模組中的執行內容。

3.1 系統架構

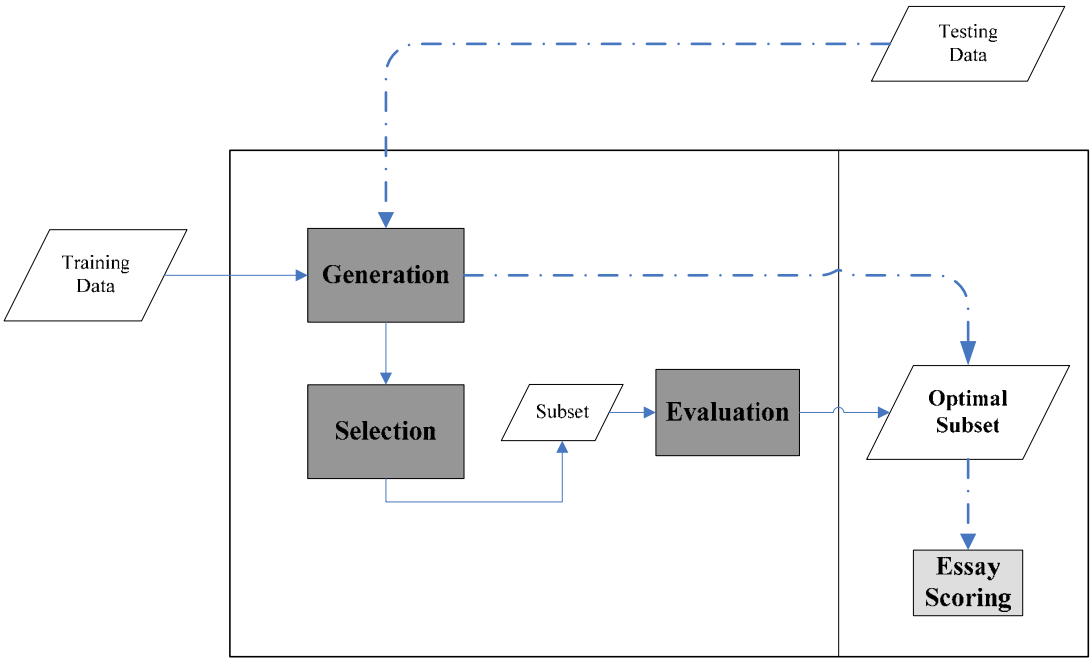

圖 4 系統架構圖

本系統裡面共包含 3 個模組：

1. Generation - 概念產生
2. Selection - 集合選取
3. Evaluation - 評估方式

當 Training Data(訓練資料，即為人工評閱過的作文資料)收集完，系統開始運作。首先第一步，藉由 Generation 模組，對所有作文資料做前置處理工作，包括了中文的斷詞，並將作文概念化(Conceptualize)。第二步，Selection 模

組用設置門檻的方式，從概念空間中篩選出部分子集合(Subset)。第三步，**Evaluation** 模組對子集合評估。最後找出其中最佳的一組子集合(Optimal Subset)，並用此集合作為文章批閱(**Essay Scoring**)的依據。

3.2 Generation - 概念產生

在此模組中，做的是前置處理工作，主要工作內容作有二，首先是中文的斷詞處理，對訓練資料(即作文)做斷詞處理，接著將作文概念化(Conceptualize)，使用的方法是根據知網(Hownet)，將所有斷好的詞轉換為義元。

3.2.1 斷詞處理

本系統的斷詞工具是採用「中央研究院資訊科學研究所詞庫小組中文斷詞系統 1.0 版」。

3.2.2 概念化 (Conceptualize)

完成作文的斷詞工作之後，接著的工作是根據知網(HowNet)，將所有作文中的詞轉換為知網中的義原。用一簡單例子說明，文章中的一句話：「馬上飛奔出去和朋友談天說地」，經過斷詞程序處理之後，成為：

「(馬上)(飛奔)(出去)(和)(朋友)(談天說地)」

共六個詞，根據查詢知網資料庫後，(馬上)這個詞的主義原為{prompt|旋即}、(飛奔)的主義原為{run|跑}、(出去)的主義原為{FuncWord|功能詞}、(和)的主義原為{FuncWord|功能詞}、(朋友)的主義原為{human|人}、(談天說地)的主義原為{talk|談話}，這句話共有六個中文詞，其中包含五個不同的義原。

3.3 Selection - 集合選取

本系統最主要目的是在整個概念(Concept)空間中，找出其中一組子集合，而這組概念集合可以拿來幫助評閱作文，有效的『鑑別』出作文好與壞。而鑑別的依據，就是在等級高的作文中，此集合裡面的義原數出現很多次，相對的，在等級低的作文當中，此集合裡面的義原數會出現較少次，這樣數量上的差異就可以拿來作為鑑別高低分作文之間的差異。

由於作文經過概念化後，產生的概念空間龐大，其所含的義原數量約一千個以上，子集合數量至少為 2^{1000}個，因此要很有效率的從中挑選出能夠適合評閱的一組子集合便顯的相當困難，所以我們設計了一套有效率的挑選方法。首先，針對每個義元計算三個統計數值，也就是H值、L值及F值。然後從這三個數值，再產生三個門檻作為集合選取之用(H/L的下界、F的上界及F的下界)。

3.3.1 義原的特性：H、L、F 的數值

底下對每個概念空間中的義原作詳細的分析並且對每個義原計算下列三個值：

H ： 該義原在高分作文中的出現頻率

L ： 該義原在低分作文中的出現頻率

F ： 該義原在所有作文中的出現頻率

其中高分作文指的是該作文分數等級為 5 分、6 分，低分作文指的是該作文分數等級為 1 分、2 分。

計算方法如下：

H =（該義原在高分作文中出現篇數）/（總作文篇數）

L =（該義原在低分作文中出現篇數）/（總作文篇數）

F =（該義原在所有作文中出現篇數）/（總作文篇數）

此三個值範圍皆介於 0 和 1 之間。

3.3.2 數值門檻的定義

底下我們利用 H、L、F 數值定義三個門檻。門檻設定的目的是允許系統可根據對這三個門檻值做調整，而得到不同的義原子集合，之後系統再利用評估模組篩選出其中最佳的即最具有鑑別力的義原子集合。

1. H/L 值下界的門檻：

前面曾經提到過的義原的鑑別力，就是在高分作文出現次數多，低分作文出現次數少。依數學的觀點來看，即是在高分作文做出出現的頻率較高，低分作文中出現的頻率較低，而之前計算義原的兩個值 H 和 L，就分別代表著該義元在高分作文中的出現頻率及該義元在低分作文中的出現頻率。當(H/L)這個數值大於一(H>L)，即表示此義元在高分作文中佔的頻率相較在此義元在低分作文中所佔的頻率高，也就是此義元在高分作文出現較多次。以(H/L)這個比值當做門檻值，可以定出不同的義原子集合，如果(H/L)值愈高，找出的子集合會愈小。該集合內義原的鑑別力會愈高。如果(H/L)值愈低，找出的子集合會愈大。該集合內義原的鑑別力會愈小。所以在此對此(H/L)值定了一個下界當做門檻，用來找出最佳的(H/L)值。

2. F 值上界的門檻：

F 值代表該義原在所有作文中出現的頻率。F 值愈高，表示此義原在作文中出現頻率愈高。當某個義原有很高的出現頻率，即表示此義原在大部份的作文中皆出現，因此該義原並沒有提供區隔高分作文與低分作文的資訊。因此針對數值 F，定了一個上界當做門檻，過濾掉那些沒有提供資訊的義原。

3. F 下界的門檻：

F 值代表該義原在所有作文中出現的頻率。F 值愈低，表示此義原在作文中出現頻率愈低。當某個義原有很低的出現頻率，即表示此義原在大部份的作文中皆很少出現，因此該義原也無法提供區隔高分作文與低分作文的資訊。因此針對數值 F，也定了一個下界當做門檻，過濾掉那些沒有提供資訊的義原。

3.4 Evaluation - 評估方式

當求出了一組義原子集合，若要以此集合作為評閱作文的依據，則此組集合的鑑別力，理論上來說，是指此集合在將來對測試資料(Testing Data)的鑑別力，並希望此鑑別的能力能夠越高越好。然而實際上的處理並不是如此，我們只能在一開始的訓練資料中，來討論該集合作為評閱作文時的鑑別力，並期許在此訓練資料中找出來最有鑑別力的集合，將來在測試資料中，一樣擁有最佳的鑑別力。

很重要的一點是，究竟該如何表示出一組義原子集合其所代表的鑑別力。很明顯的，此組義原子集合必須要能夠針對不同作文批閱出等級不同的排序。為了達到此目的，本系統設計了三個步驟，第一個步驟，依據此義原子集合針對不同的作文算出它們的全序關係；第二個步驟，挑選出其中的間距點，將此全序關係區分為六群；第三個步驟，依據這六個分群，計算此組義原子集合的鑑別力，最後從中挑出最佳的義原子集合。

步驟 1.(作文全序關係)

假設此義原子集合P共有k個義原，測試資料的作文共有t篇文章。針對每個測試資料中的作文m，計算義原子集合P中的k個義元在該篇作文的出現次數m_i。如此，我們可以得到一組數字集合H：$\{m_i|1<=i<=t\}$。明顯

的，這組數字集合可以由小到大做排序得到$\{n_1,n_2,n_3,\cdots,n_t\}$，而具有全序關係。$m_i$值越高，表示相對應的作文使用此組義原子集合P中的義原愈多。

步驟 2.（分群）

由於步驟一所得的結果僅有全序關係，無法用來批閱作文，因此尚無法評斷該義原子集合 P 的鑑別力。由於訓練資料中作文的分級制度是採用六級分分級，換句話說，即是使用六級的優劣排序，而且可從該資料中求出分級的分布。為了使分群的結果與訓練資料等級的分佈在統計上一致，我們依據訓練資料中作文等級的分佈，依相同比例在數字集合 H 中找出五個邊界(B1,B2,B3,B4,B5)用來區分這六群。

舉例說明：假設考慮中的義原子集合P_1有 250 個義原(k=250)，測試資料計有 300 篇作文(t=300)，其中一分的作文佔了 30 篇，二分的作文佔了 50 篇，三分的作文佔了 70 篇，四分的作文佔了 60 篇，五分的作文佔了 50 篇，六分的作文佔了 40 篇。也就是測試資料作文中等級的分佈為(30,50,70,60,50,40)。依照步驟一，計算得到 300 個由小到大排序的數字。步驟二依訓練作文等級的分佈，挑選出五個邊界。因為等級一分作文 30 篇，所以從排好序的 300 個數字中挑選第 30 個(n_{30})作為邊界一B1，而等級二分作文 50 篇， 所以挑選第 80 (80 = 30 + 50)個(n_{80})作為邊界二B2，以此類推，共可得到 5 個邊界值$B1=n_{30}$，$B2=n_{80}$，$B3=n_{150}$，$B4=n_{210}$，$B5=n_{260}$。

步驟三.（計算鑑別力）

依求出的這五個邊界值 B1,B2,B3,B4,B5，再根據義原子集合 P，可得到訓練資料的作文分級的統計資料。本系統以這五個邊界值作為作文批閱的依據，當一篇作文使用的義元數量小於 B1，即在步驟二的分群中屬於第

一群，系統評閱該篇作文等級為一；依此規則，當一篇作文使用的義元數量介於 B1 和 B2 之間，則認為該篇作文等級為二，其餘以此類推。

底下我們將以同一例子來說明，假設此五個邊界值分別是 6、17、30、47、67。表一顯示依據義原子集合P_1對訓練作文的分級資料。表一中，第一欄的 20、8、2 這三個數字，表示了在所有 30 篇的一分作文中，用了此義元集合P的次數為 0~5 次的有 20 篇，6~16 次的有 8 篇，17~29 次的有 2 篇。其餘以此類推。

邊界\等級	1 分	2 分	3 分	4 分	5 分	6 分
0~5	**20**	15	1	1	0	0
6~16	8	**38**	11	2	0	1
17~29	2	5	**45**	8	2	1
30~46	0	2	10	**27**	3	2
47~66	0	0	3	17	**40**	6
>66	0	0	0	5	5	**20**

表 1 例子 1 相對應的表格

針對每組不同的義原集合 P，可求出此義原子集合 P 所對應的表格 T，以及相對應的五個邊界值 B1、B2、B3、B4、B5。為了計算此組義原集合 P 的鑑別力，設計了一套評估函數計算出義原集合 P 的鑑別力：

邊界\等級	1 分	2 分	3 分	4 分	5 分	6 分
0~5	**2**	1	0	-1	-2	-3
6~16	1	**2**	1	0	-1	-2
17~29	0	1	**2**	1	0	-1
30~46	-1	0	1	**2**	1	0
47~66	-2	-1	0	1	**2**	1
>66	-3	-2	-1	0	1	**2**

表 2 函數的加權表格 H

上面表格 H 是用來幫助了解此評估函數所代表的意義，在表一中，每一格中的數字代表著作文篇數，依其所在表格中位置的不同，皆有著不同含義。在主對角線上的，表示系統對該作文的批閱與原先該作文在訓練資料中所受的批閱完全一致，故給予兩分的加權。遠離主對角線一格，即表示系統的批閱與原先所受的批閱等級相差一分，只給予一分的加權。以此類推，愈遠離主對角線，所能得到的加權分數愈低，當遠離主對角線達三格以上時，給予其負分的加權。加權之後，根據不同等級所佔的比例對加權後的分數做調整，評估函數 S()的公式：

義原集合 P，相對應的表格 T，加權表格 H，則

$$S(P) = \sum_{i=1}^{6} \frac{\sum_{j=1}^{6} H_{i,j} \times T_{i,j}}{\sum_{j=1}^{6} T_{i,j}} \qquad (1)$$

其中$H_{i,j}$指的是在表格H中第i行第j列的欄位

以例子一來計算，該義原集合 P 的分數計算如下：

$$
\begin{aligned}
S(P_1) = {} & (20 \times 2 + 8)/30 + \\
& (38 \times 2 + (15 + 5) + 2 \times (-1))/50 + \\
& (45 \times 2 + (11 + 10))/70 + \\
& (27 \times 2 + (8 + 17) + 1 \times (-1))/60 + \\
& (20 \times 2 + 6 + (-1) + (-2))/50 \\
= {} & 7.2257
\end{aligned}
$$

取得最佳子集合

利用式子(1)評估函數，可以計算出一組義原集合作為特徵時的分數，而該分數變是用來評斷該組義原集合的鑑別力，當分數愈高，即表示該組義原集合愈有鑑別力，最後的目標為在大量的義原當中，篩選出其中一組會使評估函數分數最高，也就是最有鑑別力的子集合。而篩選子集合的方法，便是利用三個門檻值(H/L 的下界、F 的上界及 F 的下界)，藉由設定不同的門檻值，便可得到一組不同的義原子集合，接著利用設計的評估含數計算此組義原集合的分數並紀錄下來，不斷重覆調整三個門檻值，直到篩選的三個門檻範圍都結束，從中挑出評估函數分數最高的義原集合作為系統產生的最佳特徵，此組集合即代表著最具有鑑別力的最佳義原子集合。

第四章、實驗過程與結果討論

在本章節中，4.1 節中說明本系統的實驗資料來源。本實驗共分成兩個階段 1.訓練階段 2.測試階段，分別用兩個流程圖說明其執行步驟。最後一節 4.4 為實驗的數據結果。

4.1 實驗資料

本實驗所使用的作文資料來自台北市立敦化國中，作文的題目是「下課十分鐘」，這些作文輸入成電子檔時保留所有的錯字以及標點符號，不加以修改，以維持學生作文的原貌。經過評分的作文，包含：

1. 整體評價分數
2. 符合主題分數
3. 文章完整性分數
4. 組織架構分數
5. 文句通順分數
6. 修辭用字分數

每一項的評分都是從一分到六分，一分為最低，六分為最高。每篇由二至三名老師評分，取平均分數，若其中有兩名老師的總體分數相差超過兩分，則該篇作文不予列入資料庫，總共得到 693 篇作文。

4.2 實驗流程

本實驗共分為兩大階段，第一階段是系統的訓練，第二階段是系統的測試。在訓練階段中，系統從所收集到的 693 篇作文中，從每個不同等級中亂數挑選一半數量共計 347 篇作文做為訓練資料，訓練階段完成後，系統會產生一組最佳的義原子集合 P 以及五個邊界值，並根據其來評閱作文等級。在測試階段中，系統把訓練資料中沒被挑選的剩餘作文 346 篇作為測試資料，系統並根據訓練階段產生的義原子集合 P 及五個邊界值評閱作文等級，最後比較系統評閱的等級與原先所受的評閱等級差異，來計算系統的正確率。

訓練階段流程：

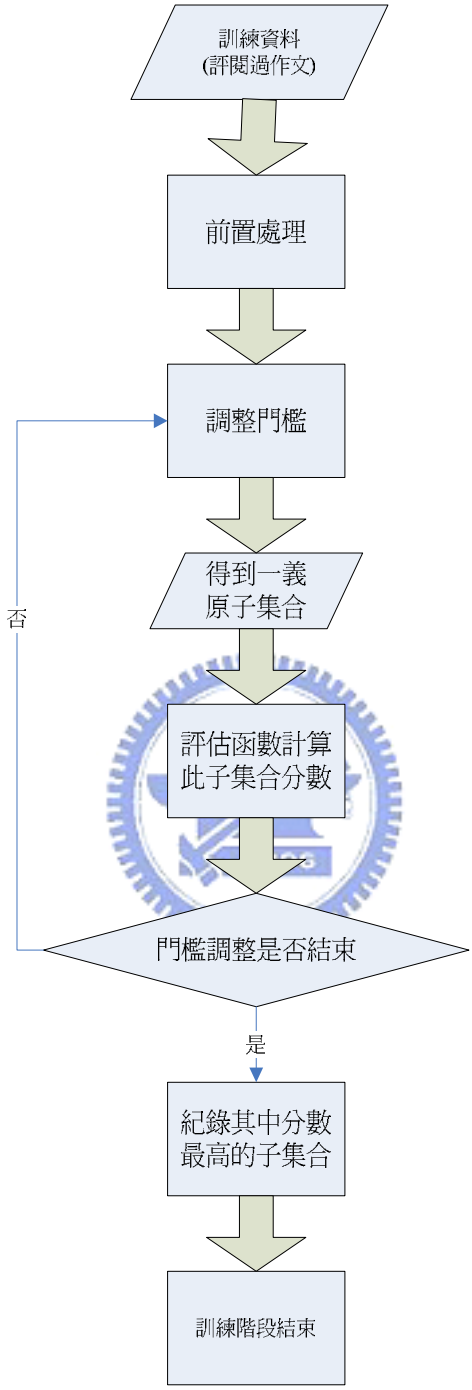

圖 5 訓練階段流程圖

測試階段流程：

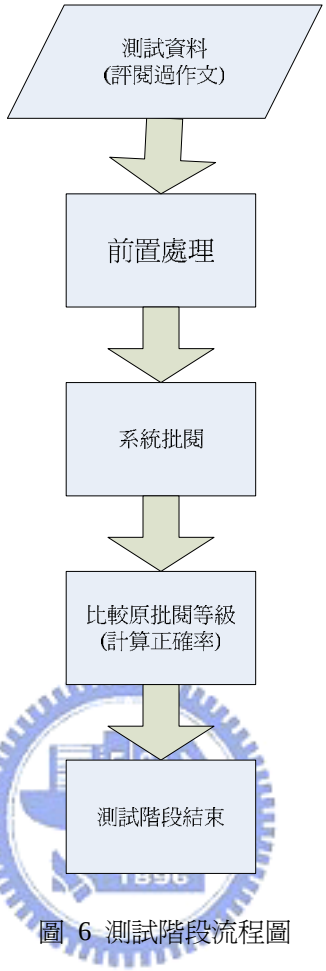

圖 6 測試階段流程圖

4.3 正確率的計算：

從第一階段的訓練得到的結果，包含有一組義原集合 P，以及五個數字 B1、B2、B3、B4、B5，這五個數字分別用來區隔不同分數所該擁有的義元數量，比如說，當一篇作文，對 P 這組義元集合用的次數小於 B1，則認定此篇作文該項評估分數為一分，若次數介於 B1 和 B2 之間，則認定該篇作文此項評分標準為兩分，以次類推，如此一來，第一階段訓練(Training)所得到的結果，便可以拿來批閱未知的作文，預測作文的分數。本次實驗總共計算 3 種正確率 Adjacent、avg. Adj、Exact：

Adjacent： 容許一分誤差的整體正確率

avg. Adj： 容許一分誤差的平均正確率

Exact ： 毫無誤差的精準正確率

定義分別如下：

Adjacent：在全部用來做訓練的作文之中，此系統批閱的分數，和實際老師所批閱的分數，若兩者差距在一分以內，皆視為系統正確的批閱，數學定義為：

Adjacent = 容取一分誤差下正確的批閱數 / 訓練的作文數

由於每位老師的背景知識、主觀認知不盡相同，造成不同的老師對於作文的評分標準也會不同，因此本實驗認為相差一分為可容許的誤差，在這一分的誤差範圍下，皆視為正確的批閱。

avg. Adj：和正確率 Adjacent 很類似，唯一不同的地方在於，Adjacent 計算的是全部作文中正確的比率，而 avg. Adj 是針對不同分數的作文，分別去計算其在容許一分誤差之下的正確率，之後再把此六個正確率平均得到 avg. Adj，數學上的定義：

$$\text{avg. Adj} = \sum_{i=1}^{6} \text{Adjacent}_i / 6$$

定義 avg. Adj 這個正確率，主要目的在於在 Adjacent 正確率考慮的是整體的正確率，而沒有考慮到由於分數的不同，篇數分布的比例也不同，往往高分和低分的篇數較少，中間分數的作文篇數較多，在 Adjacent 的正確率計算中，只計算全部整體的正確率，而沒考慮到不同分數間正確率的關係。然而，我們所期望的結果是，在對不同分數的作文間，都能夠有好的正確批閱，因此，在 avg. Adj 的正確率計算，是先對不同分數的作文，計算其容許一分誤差下的正確率，再對其取平均值得到 avg. Adj 正確率。

Exact： 計算的方式和 Adjacent 唯一不同的地方在於，系統批閱的分數，和實際老師所批閱的分數，兩者必須完全一致，才視為系統正確的批閱，數學上的定義為：

Exact = 不容許誤差下正確的批閱數 / 訓練的作文數

和前兩者計算正確率的方式比較，Exact 計算的方式不容許任何誤差，系統批閱的分數必須和實際老師批閱的分數兩者必須一致，雖然說，容許一分誤差是相當合理的假設(因為不同的老師，會造成評分標準不同，因此容許一分的誤差，可說是合理的假設。)，不過，此 Exact 的正確率計算方式卻可以看出，原本做批閱的老師，和本系統訓練後所得的結果，彼此之間完全一致的程度。

4.4 實驗結果與討論

分別針對不同評分項目做了多次實驗，在每個實驗中，皆從蒐集到所有的 693 篇作文中，在每個不同等級中亂數挑選一半數量共計 347 篇作文做為訓練資料，剩餘的 346 篇作為測試資料，實驗結果如下：

針對作文的整體(holistic)分數：

	Exact	Adjacent	avg. Adj
實驗1	47.09%	93.02%	92.40%
實驗2	45.90%	91.60%	90.80%
實驗3	47.67%	92.27%	91.70%
平均	**46.89%**	**92.30%**	**91.63%**

表 3 實驗結果-整體

針對作文的修辭(Rhetoric)分數：

	Exact	Adjacent	avg. Adj
實驗4	46.50%	95.30%	95.30%
實驗5	48.20%	94.70%	93.60%
實驗6	42.70%	94.50%	88.80%
平均	**45.80%**	**94.83%**	**92.57%**

表 4 實驗結果-修辭

針對作文的主題(Topic)分數：

	Exact	Adjacent	avg. Adj
實驗7	44.60%	92.70%	78.50%
實驗8	42.60%	91.90%	82.80%
實驗9	44.10%	92.50%	86.70%
平均	**43.77%**	**92.37%**	**82.67%**

表 5 實驗結果-主題

本系統對於作文的修辭(Rhetoric)分數、主題(Topic)、整體(Holistic)分數這三個評分項目做實驗。在蒐集的作文資料中，每篇作文由二至三位老師批改，每位老師的批改數量約 50 至 100 篇作文。計算每篇作文的任意兩名老師所評閱分數的差距，Exact 值計算兩個老師對作文給相同評分的百分比，Adjacent 值計算兩個老師對作文批閱差距在一分之內的百分比，可以得到下列表格 6；表格 7 記錄的是表格 3、表格 4、表格 5 中實驗的平均值，代表的涵義為本系統與老師評閱的差距：

	Exact	Adjacent
整體	33.42%	78.14%
修辭	29.72%	74.86%
主題	27.88%	72.45%

表 6 老師之間評閱差距

	Exact	Adjacent
整體	46.89%	92.30%
修辭	45.80%	94.83%
主題	43.77%	92.37%

表 7 系統評閱結果

比較表格 6 與表格 7，本系統在這三項評分項目中，Exact 值與 Adjacent 值皆比老師之間的 Exact、Adjacent 值高出很多，這也代表著本系統的評閱分數具有相當的可信度，可作為老師批閱作文時的參考依據。

第五章、結論

5.1 研究總結

在本論文中，我們提出以概念(Concept)為基礎的作文評分方式，經由統計作文中的概念(Concept)數量，我們觀察發現作文中概念(Concept)的使用數量與該作文分數等級有很大的關係，並以概念為基礎自動的建立了一套作文評分系統。此系統包含了三個模組 Generation、Selection、Evaluation，分別負責概念的產生、子集合的選取、集合的評估方式。

根據實驗結果，本系統對於作文的修辭(Rhetoric)分數、主題(Topic)、整體(Holistic)分數這三個評分項目皆有很高的評閱正確率，在容許一分誤差下的正確率(Adjacent)可以達到 91%~95%，這表示本系統批閱作文的結果與專業人員(老師)批閱的結果相當接近，即兩者之間的誤差範圍很小。因此，本系統提供了一個協助老師評閱作文時的使用工具，尤其是在大量的作文批閱時(如聯考)，更是可以節省大量的人力資源。

5.2 未來工作

本論文提出以概念(Concept)為基礎的自動評閱作文系統，其中對作文概念的轉換是以知網(Hownet)為工具，本系統以每個詞的第一個描述式中的第一個義原作為代表該詞的概念，這樣的優點是處理上較為簡單，但是也因此捨棄了知網中概念之間與義原之間複雜的關係。在未來的工作上，可以以此為方向，利用知網作為工具，找出作文中概念與概念彼此之間複雜的關係，進而提升本系統的能力。

參考文獻

[1] Marti A. Hearst(2000). The debate on automated essay grading.

[2] Jill Burstein. The E-rater Scoring Engine: Automated Essay Scoring With Natural Language Processing. Automated Essay Scoring: A Cross-Disciplinary Perspective (2003). pp. 113-121

[3] Lawrence M. Rudner and Tahung Liang. Automated Essay Scoring Using Bayes' Theorem.(2002)

[4] Ming, P.Y., Mikhailov, A.A., & Kuan, T.L. (2000). Intelligent Essay marking system. In C. Cheers (Ed.), Learners Together,Feb. 2000, NgeeANN Polytechnic, Singapore.

[5] Salvatore Valenti, Francesca Neri and Alessandro Cucchiarelli. An Overview of Current Research on Automated Essay Grading.(2003)

[6] 杜飛龍 (1999),《知網》辟蹊徑，共用新天地——董振東先生談知網與知識共用，《微電腦世界》雜誌，1999 年第 29 期

[7] 劉群,李素建. 基於《知網》的辭彙語義相似度計算