國立交通大學

資訊科學與工程研究所

碩 士 論 文

基於支援向量機之中文自動作文評分系統

Automated Chinese Essay Scoring System Based on Support Vector Machine

研 究 生：粘志鵬

指導教授：李嘉晃　教授

中 華 民 國　九 十 五　年 六 月

基於支援向量機之中文自動評分系統

Automated Chinese Essay Scoring System Based on Support Vector Machine

研 究 生：粘志鵬　　Student：Chih-P'eng Nien

指導教授：李嘉晃　　Advisor：Chia-Hoang Li

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

June 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年六月

基於支援向量機之中文自動作文評分系統

學生 ： 粘志鵬　　指導教授：李嘉晃 博士

國立交通大學電機資訊學院　資訊科學與工程研究所

中文摘要

在本論文我們提出一種基於「特徵義原空間」的方法，並使用支援向量機理論模型來自動建立一套中文作文評分系統。另外，本論文也對中文作文評分系統，有可能遭遇到的後門攻擊，提出了一種基於 ***Google Search*** 的偵測模組，負責避免鑽評分機制漏洞的文章，造成系統嚴重的評分錯誤。本論文提供了一套可作為協助老師評分作文時所使用的工具，而根據實驗結果，本系統評分的精準正確率，可以達到 *55.20%*，而在容許一分誤差下的正確率，可以達到 *96.82%*。

Automated Chinese Essay Scoring System Based on Support Vector Machine

Student ： Chih-Peng Nien Advisor：Prof. Chia-Hoang Lee

Institute of Computer Science and Engineering

College of Computer Science

National Chiao Tung University

Abstract

We propose a kind of method based on “the original space of characteristic definition”, and use Support Vector Machine (SVM) to set up an Automated Chinese Essay Scoring System (ACESS) in this thesis. In addition, we also discuss the probably existed attacks on ACESS and propose a detection model based on Google Search to responsible for avoiding the system from critical scoring mistakes caused by the article of premeditatedly attacking systematic shortcomings.

The ACESS we proposed in this thesis can be as a tool used while helping the teacher to score the composition. According to the experimental results, the perfect exact rate of this system performance can be up to 55.2% and the adjacent rate under permitting a point of errors can be up to 96.82%.

目錄

圖表

表格

第一章、緒論

1.1 研究動機

半導體教父*張忠謀*曾在《中文優勢論》中提到,『在大中國市場興起的時代，中文能力已成為台灣吸引外資的人才優勢』;日本經濟大師*大前研一*也說,『台灣人具備的語文優勢，全世界沒得比』。這兩位前瞻性的當代人物不約而同地對於未來中文能力都有著相當的重視,也道出未來台灣的經濟利基就在於自身的中文語文能力上。

隨著數位時代的來臨，由電視取代閱讀，電腦取代書寫，已漸漸成為一種全球化的趨勢，加上數位影音紀錄的發明與進步，語言表達能力的重要性更是逐漸抬頭。具備良好的中文能力不只在日常生活扮演重要的溝通工具，也是所有學問的基礎，更是每個人在未來都用得到的能力，不僅僅數學、理化等學科才應用得到，就連將來求職撰寫履歷、面試對談等也都需要具備良好的表達與寫作能力，才能夠替自己的競爭力加分。

作文的能力就是在培養組織與思考能力，訓練自身表達能力的重要推手，藉由不斷地練習寫作，不僅能增進學生閱讀吸收能力，亦可增進文藝欣賞及創造的能力，更可將所學的零星片斷知識串連整合，藉此提昇自身的中文能力，並且能夠正確地表達個人思想與溝通能力。

然而作文所面臨到的最大問題在於，批改大量文章不僅需耗費大量的人力，同時也耗費許多的時間，最重要的是由於個人主觀不同，可能造成評分有所差異性而增加評分難度。因此單純的人工閱卷很難維持公平、公正、及客觀性。有鑒於西方英文的自動作文評分系統已經有長久的發展，其效果也已被大眾所接受，如英文托福考試作文***(AES**、**e-rater** 及 **IEA)***[1][2][3]，也改以電腦來取代人工閱卷，因此本研究特別針對中文作文處理，發展一套加強型的自動化的中文作文評分系統，不僅改善以往的評分系統，更提供一個標準化的平台，藉此用來協助老師評分，並且改善由人工閱卷所造成的評分標準不一致問題。

1.2 研究目的與構想

目前的評分系統利用「詞袋式」(即假設詞語與詞語之間相互獨立，彼此並沒有關係)的自然語言處理的方式，由於詞語之間相互獨立，使得目前系統能有效地擷取所需的特徵，進而對文章作評分。然也正因為詞袋式的模型，使得系統暴露了一個很嚴重的漏洞，牛頭不對馬嘴的文章更可能因為此漏洞而造成系統嚴重的誤判(即本應低分的文章，因此系統的漏洞而被評為高分文章)。關於此問題，將在本論文第三章 3.6 節中討論。

因此本論文的研究目的在於，有效地建立一個加強型的自動化的中文作文評分系統，能試圖攔截針對傳統詞袋式自然語言處理的攻擊。之後再經由分析訓練集資料後，能自動且客觀地從中產生一組特徵*(attributes)*，而本系統最後可利用此組特徵經由支援向量機（*Support Vector Machine*）訓練產生的預測模型，對中文作文做出標準化的評分，且最後的評分效果需與老師人工評分不能有很大的偏差，亦即需有著相當高的準確率。

1.3 論文架構

第一章為緒論，內容為介紹本論文的研究動機，提到目前世界對中文能力的重視，以及作文能力提昇的重要性；最後介紹本論文研究目的，及整個論文架構。第二章為相關研究與本論文所用到的理論模型，包括《知網》、中文斷詞、支援向量機分類理論以及特徵選取的方法。第三章則是本研究的系統設計，包括核心的系統介紹，並詳細地介紹整個系統的架構及說明。第四章為實驗過程與結果討論，一開始藉由比較系統評分的等級與其它評分系統相比評分等級差異，來計算系統的正確率，最後分析不同的訓練集大小對於支援向量機在作文評分分類上的比較。第五章為結論。

第二章、相關研究與理論

在本章節中，將詳細的介紹本論文提及的相關研究與理論。2.1 節介紹了《知網》，包含《知網》的結構以及義項。2.2 節介紹自然語言處理中，有關中文處理中最基本且重要的一個問題，中文的斷詞。2.3 節介紹了本論文所採用的分類理論-支援向量機*(Support Vector Machine)*的原理探討。最後，2.4 節介紹了目前關於特徵選取的兩種常見方法，文件頻率*(Document Frequency)*及互訊息*(Mutual Information)*，以及本實驗結合這兩個方法，建立系統所需的「特徵義原空間」的系統建立想法。

2.1 《知網》*(HowNet)*

《知網》（英文名稱為 *HowNet*）是一個以漢語和英語的詞語所代表的概念為描述物件，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為基本內容的常識知識庫[4]。*1988* 年前後，*董振東*先生曾在他的幾篇文章中提出以下的觀點[5]：

（1）自然語言處理系統最終仍需要強大的知識庫來支援。

（2）知識乃是一個系統，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。

（3）在建立知識庫的問題上，他提出應首先建立一種可以被稱為知識系統的常識性知識庫，藉著用通用的概念來描述物件，建立並描述這些概念之間的關係。

（4）而關於由誰來建立知識庫的問題，他提出：首先整個知識庫的框架應由知識工程師來設計，並建立常識性知識庫的原型。最後在此基礎上再朝向專業性知識庫延伸和發展。而所謂專業性知識庫或稱百科性知識庫主要靠專業人員來完成。這裡很類似於通用的詞典由語言工作者編纂，百科全書則是由各專業的專家編寫。

《知網》的研究和建立是實踐上述觀點的努力。

《知網》中最主要是由「概念」與「義原」兩個定義所組織而成。其中，概念是表示對辭彙語義的一種描述，且此種描述是由一種知識表示語言來表達，而這種知識表示語言所用的辭彙就叫做義原。

義原除了為描述一個概念所代表的最小意義單位。在義原與義原彼此之間又存在著複雜的關係。《知網》共定義了 8 種義原之間的關係：*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*。所以義原之間的組成可看成一個複雜的網狀結構，而不是一個單純的樹狀結構。

在《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，我們稱之為義項。*DEF* 是《知網》的核心。

一個詞語可能有多個義項，但其中的第一個義項對於該詞語來講是最重要的一個義項。該義項呈現了該詞語最基本的語義特徵，而每一個詞語的第一義項中的第一個義原稱為「主義原」，在本系統中也用以作為代表該詞語的概念。

為了更了解《知網》，我們從底下的例子來說明，中文詞「**下課**」其所代表的「概念」為 {*cease*|*停做*:*content*={*study*|*學習*},*domain*={*education*|*教育*}}。由 *DEF* 的內容可看出《知網》一共用了三個義原來解釋這個概念，分別是{ *cease*|*停做*}、{ *study*|*學習*}、{ *education*|*教育*}。而第一個義原即{ *cease*|*停做*}當作主義原來表示「**下課**」這個詞語所代表的概念。

例子：

NO.= 103222
W_C= 下課
G_C=V
E_C=
W_E=finish class
G_E=V
E_E=
DEF={cease|停做:content={study|學習},domain={education|教育}}

而交通大學碩士*蔡沛言*先生，在其發表的論文《自動建構中文作文評分系統：產生、篩選與評估》[6]中，運用《知網》中的「主義原」當作「概念」，提出以「概念」為基礎的作文評分方式。

2.2 中文斷詞*(Word Segmentation)*

所謂的詞乃是具有最小有意義，並且可以自由使用的語言單位。在所有自然語言處理的系統上，首要步驟都必須先能分辨文章中的各個詞，然後才能進行詞性標記、語言分析、資訊擷取等進一步的處理。

因此，中文自然語言的處理，在中文自動斷詞的工作處理上就成了最根本且重要的技術。然而中文的自動斷詞工作並沒有如英文處理一般容易，在自然語言處理中，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。而倘若斷詞結果不正確，更會造成語意上全然的不同，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，考慮以下的句子：

「今天天氣很好」

而這個句子可能的斷詞結果有下列幾種：

「（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」這句話正確的斷詞應為第二句

「(今天）（天氣）（很好）」

而其它句的斷詞結果都會造成語意上不正確，語法上也沒有代表意義。

2.3 支援向量機*（Support Vector Machine）*

支援向量機[7]是由 *Vapnik* 在 *1995* 年所提出來的，在近幾年被廣泛用於解決各種分類的問題上，逐漸已成為機器學習*(Machine Learning)*領域中極為熱門的一種方法。

本論文對支援向量機分類的理論作最初步的探討[8]，而有關更詳細深入之資訊可另行再參考有關支援向量機的相關著作。

2.3.1 原理介紹

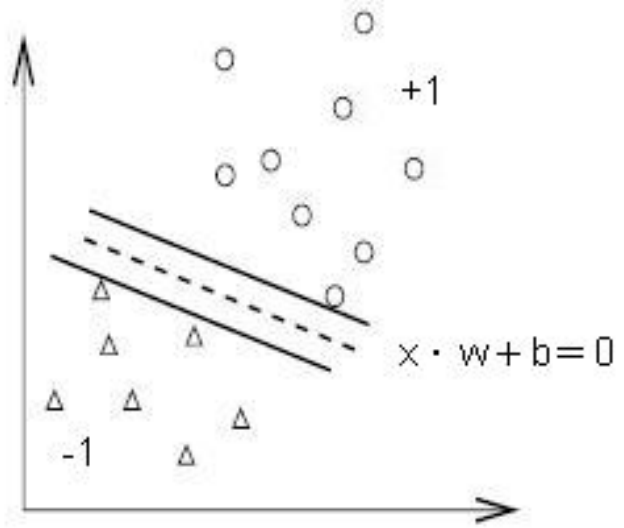
有關支援向量機分類方法的基礎定義：

x_i：*vector*，表示一筆資料中的各個屬性，$x_i \in R^n$，$i=1,\ldots,l$

y_i：*Label* 為+1 或-1，表示分類的兩種類別，$yi \in \{\pm 1\}$，$i=1,\ldots,l$

f：*decision function*，決定函數，$f: R^n \rightarrow \{\pm 1\}$

支援向量機分類理論的判斷，就是在給予一筆資料 x_i 時，判斷該筆資料是屬於哪一個類別(+1 或-1)。而分類原理是，在已給定的訓練資料群中，找出一個超平面(*hyperplane*)，目的是將這些訓練資料區分開來，見圖表 *1*。



圖表 1　超平面示意圖

支援向量機的分類方法也就是藉由找到相對應的 w 和 b，來把資料切成兩半，得到所謂的預測模型。而測試資料時，只要根據決定函數 $f: x \cdot w + b$ 的值來做分類，若 f 的值大於 0，則該筆資料屬於+1，反之若 f 的值小於 0，則該筆資料歸屬於−1。

而這樣的 w 和 b 的可能性組合，有上千種甚至更多，所以支援向量機的問題在於如何找到一個最合適的超平面。而此問題，可藉由尋找一個擁有最大區域的分離超平面，使得資料能有效的分開，而這種方法更可有效地降低測試錯誤。而此時支援向量機就必須滿足下列限制條件：

$$x_i \cdot w + b \geq +1 \qquad \forall y \in \{+1\} \qquad (1)$$

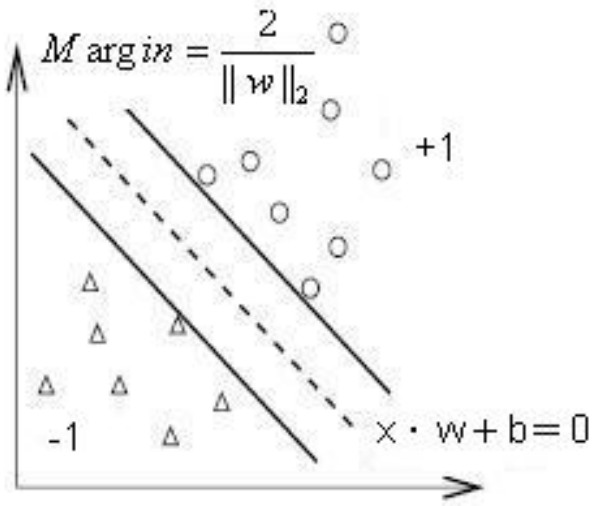
$$x_i \cdot w + b \leq -1 \qquad \forall y \in \{-1\} \qquad (2)$$

組合等式(1)與等式(2)，可得到下式

$$y_i(x_i \cdot w + b) - 1 \geq 0 \qquad \forall i \qquad (3)$$

在等式(1)中，對於適當的 w 與 b 使得 $x_i \cdot w + b = 1$ 成立，則位於此平面上的點 x_i 到原點的垂直距離為 $|1-b|/\|w\|_2$。同樣地在等式(2)中，對於適當的 w 與 b 使得 $x_i \cdot w + b = -1$ 成立，則位於此平面上的點 x_i 到原點的垂直距離為 $|-1-b|/\|w\|_2$。因為此兩平面互相平行，因此定義 ***Margin*** 為此兩平面的距離為 $2/\|w\|_2$。

下圖表 *2* 表示一個擁有最大區域的分離超平面的示意圖。



圖表 2　最佳超平面示意圖

當使 *Margin* 最大時產生，亦當使$\|w\|_2^2$最小化時，即可得到最佳的超平面。此時該問題可轉化為解凸面最佳化的問題*(Convex Optimization Problem)*，即在線性不等式的條件*(Linear Inequality Constraints)*下，對二次函數*(Quadratic Function)*求最小化，亦即得到下列式子

minimize　　$\frac{1}{2}\|w\|_2^2$

subject to　　$yi(xi \cdot w + b)-1 \geq 0$　　$\forall\, i=1,\ldots,l$

而上述式子為二次函數求極值的問題，可藉由 *Lagrangian Theorem* 來幫助解決，得到下式：

Lagrangian：$L_p(w,b,\alpha) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{l}\alpha_i y_i(x_i \cdot w + b) + \sum_{i=1}^{l}\alpha_i$

Lagrangian Multipliers：α_i, $i=1,\ldots,l$ 為正數

將得到的$L_p(w,b,\alpha)$對 *w* 及 *b* 作偏微，

再來分別令$\frac{\partial L_p(w,b,\alpha)}{\partial w} = 0$及$\frac{\partial L_p(w,b,\alpha)}{\partial b} = 0$，得到下列兩個條件式：

$$w = \sum_i \alpha_i y_i x_i \tag{4}$$

$$\sum_i \alpha_i y_i = 0 \tag{5}$$

將式子(4)代回原 *Lagrangian*，可得到其對偶格式(*Dual Form*)，且是一個最大化的問題：

$$\textit{maximize} \qquad L_D(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

$$\textit{subject to} \qquad \alpha_i \geq 0 \quad , \forall\, i=1,\ldots,l \quad \textit{and} \quad \sum_i \alpha_i y_i = 0$$

而對線性可分離的資料而言，可以從對偶格式求出使得最佳化問題最大的 α_i，再加上式子(5)的條件，最後由式子(4)中求出 w。

接下來再利用最佳化理論的 *Karush-Kuhn-Tucker(KKT)*條件，而式(3)的 *KKT* 條件求出如下，

$$y_i(x_i \cdot w + b) - 1 \geq 0, \forall i \tag{6}$$

$$\alpha_i \geq 0, \forall i \tag{7}$$

$$\alpha_i \cdot (y_i(x_i \cdot w + b) - 1) = 0, \forall i \quad \textit{and} \quad \alpha_i \geq 0 \tag{8}$$

透過 *KKT* 條件(8)可得知，若(6)中的不等式限制*(Inequality Constraints)*不等於零時，則 *Lagrange Multipliers* α_i 必為零；若(6)中的不等式限制為零時，則 *Lagrange Multipliers* $\alpha_i \neq 0$。而在求出的 α_i 之中，$\alpha_i \neq 0$ 對應的那些資料稱為支持向量*(Support Vector)*，會使得等式(3)成立。

在訓練的過程中 w 已經可以求出，而要求出 b，要再利用 $\alpha_i \neq 0$ 的資料，透過 *KKT* 條件(8)即可以求出。當把 w 跟 b 求出後，就可以找到一個具有最大分離區域的超平面。

2.4 特徵抽取方法的相關研究

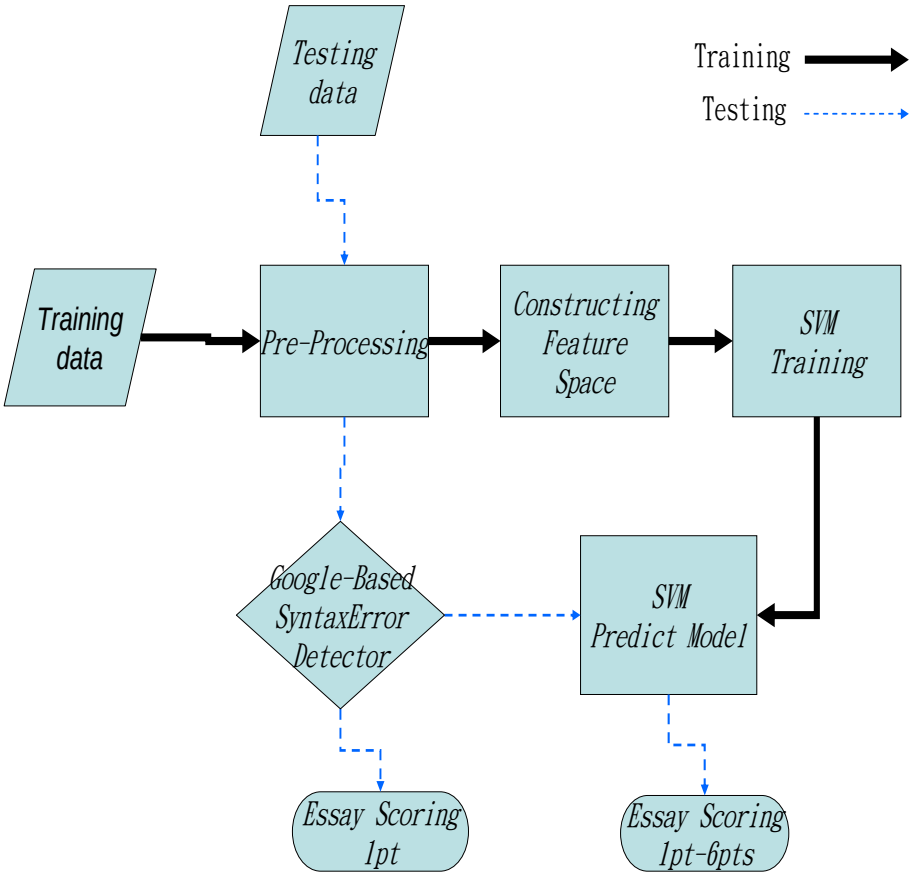
特徵抽取在分類方法中常扮演著相當重要的角色,傳統上常見的兩種特徵抽取方法為文件頻率*(Document Frequency)*、互訊息*(Mutual Information)*[9][10]。其中,文件頻率 *DF* 的方法首先計算每份文件中每一個詞語的 *DF* 值(一個詞語的 *DF* 值即為在所有訓練資料中包含該詞語的文件數),然後將 *DF* 值低於某一門檻的詞語從原始的特徵空間中移除掉,進而達到降低維度和特徵抽取的目的。而互訊息 *MI* 的方法則是觀察每一個詞語與各個類別之間的彼此相關度,當一個詞語與某一個類別的 *MI* 值表現越大,則說明兩者關聯越緊密,亦即此詞語對於該類別來說,有著相當高的依賴性。同樣地,在最後特徵選取時,再將 *MI* 值低於某一門檻的詞語從原始的特徵空間中移除,達到降低維度和特徵抽取的目的。

總結上述兩種方法,不管是文件頻率 *DF* 和互訊息 *MI*,其最大的貢獻,在對去除特徵空間中的雜訊問題,都有不錯的表現。根據*黃飛燕*小姐等的「中文文本分類中特徵抽取方法的比較研究」[9]中指出,對於中文來講,由於原始特徵空間的維度會比英文多出許多,所以可能出現的低頻詞語會更多,而這些低頻詞語會減低系統的效能,因此可先利用文件頻率 *DF* 去掉部份低頻詞語,然後再使用互訊息 *MI* 方法,這樣會使特徵抽取的效果更進一步提高。該文章中,並進一步地透過實驗證明了這種「組合特徵抽取」的有效性。而在本論文中我們也基於此優點,採用「組合特徵抽取」的概念而將此兩種方法整合,亦即先用文件頻率 *DF* 的方法去除原始特徵空間中的噪音(即低頻詞語),再利用互訊息 *MI* 的方法來挑選特徵,並藉此建立本系統所需的「特徵義原空間」*(Characteristic Definition Space)*。

第三章、系統設計

本章節將詳細介紹整個系統的架構及實作細節。3.1 節先用一個系統架構圖來描述整個系統的運行流程，包括系統訓練資料部分及系統測試資料部分。在此架構圖中，總共描述了系統的 5 個主要模組，包括 1. ***Pre-Processing – 資料前置處理***、2. ***Constructing Feature Space – 建構特徵義原空間***、 3. ***SVM Training – SVM 訓練階段***、 4. ***SVM Predicting Model – SVM 預測模型***、 5. ***Google-Based SyntaxError Detector – 語法錯誤偵測***。接下來的 5 個小節將會詳述這 5 個模組的執行內容。

3.1 系統架構



圖表 3　　系統架構圖

本系統包含 5 個模組：

1. *Pre-Processing – 資料前置處理*
2. *Construct Feature Space–建構特徵義原空間*
3. *SVM Training-SVM 訓練階段*
4. *SVM Predicting Model-SVM 預測模型*
5. *Google-Based SyntaxError Detector-語法錯誤偵測*

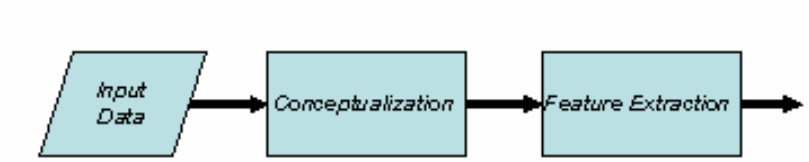
系統的主要流程如圖表 *3*，分為系統訓練資料部分及系統測試資料部分。

系統訓練資料部分首先收集訓練資料(*Training data*，即事先由老師人工評分過的文章)後。第一步先經過「***Pre-Processing***」模組，對所有作文資料做前處理的工作，包括中文斷詞、作文概念化及抽取系統所需表面特徵(如文章字數、文章段落數、文章所使用成語數、完整句子數及好義原個數)。第二步，「***Constructing Feature Space***」模組會為每一個分數類別(一分至六分)建立個別的「特徵義原空間」*(Feature Space)*。第三步，「***SVM Training***」模組會先根據特徵義原空間的資料建立 *6* 個主要訓練特徵，最後加上文章的 *5* 個表面特徵，共 *11* 個特徵全部作為 *SVM* 訓練時所需的一組特徵*(attributes)*。最後，「***SVM Predicting Model***」模組係根據 *SVM* 訓練後所產生的預測模型，作為未來文章批閱*(Essay Scoring)*的依據。

系統測試資料部分，第一步首先測試資料一樣先經過「***Pre-Processing***」模組，對所有作文資料做前處理的工作，包括中文斷詞、作文概念化及系統所需表面特徵抽取(如文章字數、文章段落數、文章所使用成語數、完整句子數及好義原個數)。第二步，「***Google-Based SyntaxError Detector***」模組會根據 *google-search* 的資料，來偵測是否為蓄意鑽系統漏洞的文章。若「是」則直接評為低分文章(即一分)；反之，若「否」則繼續進行下一個步驟。第三步，「***SVM Predicting Model***」模組會先根據之前訓練資料所建立的「特徵義原空間」*(Feature Space)*資料，一樣產生 *6* 個主要特徵，最後加上文章的 *5* 個表面特徵，共 *11* 個特徵供 *SVM* 預測模型做文章最後批閱*(Essay Scoring)*。

3.2 *Pre-Processing* – 資料前置處理

「*Pre-Processing*」模組主要作資料前置處理的工作，包括中文斷詞、作文概念化及系統所需表面特徵抽取(如文章字數、文章段落數、文章所使用成語數、完整句子數及好義原個數)。此模組主要由「***Conceptualization*** *－概念化*」及「***Feature Extraction*** *－特徵抽取*」兩個小模組所構成。其中，第一個小模組，負責將輸入資料(即訓練資料及測試資料的作文)作中文斷詞處理，接著將作文資料概念化，使用的方法係根據知網*(HowNet)*，將所有斷好的詞語轉換為義原；第二個小模組，主要負責抽取系統所需表面特徵。資料的處理流程見圖表 *4*。



圖表 4　資料前處理流程圖

以下分別用兩個小節，詳述這兩個小模組的執行內容。

3.2.1 *Conceptualization* – *概念化*

◆ 中文斷詞處理

本系統所採用的中文斷詞工具係根據「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」。[11]

◆ 作文概念化

在完成作文的斷詞工作之後，緊接著的工作係根據知網*(HowNet)*，將所有作文中的詞語轉換為知網中所描述的義原。以下用一個簡單的例子來說明此轉換過程，考慮文章中的一句話：

「大家的動作由緩慢轉變成快速」

在經過中文斷詞程序處理之後，成為：

「（大家）（的）（動作）（由）（緩慢）（轉變成）（快速）」共七個詞語，根據查詢知網資料庫後，各詞語的主義原如下所式：

（大家）這個詞語的主義原為{human|人}、

（的）的主義原為{FuncWord|功能詞}、

（動作）的主義原為{do|做}、

（由）的主義原為{FuncWord|功能詞}、

（緩慢）的主義原為{slow|慢}、

（轉變成）的主義原為{become|成為}、

（快速）的主義原為{fast|快}，

這句話共有七個中文詞語，而其中包含六個不同的「義原」。

3.2.2 *Feature Extraction－特徵抽取*

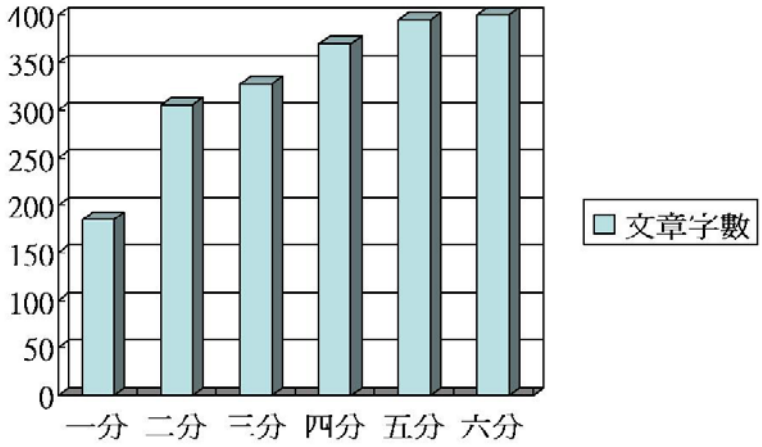
負責萃取系統所需表面特徵，包括「文章字數」、「文章段落數」、「文章所使用成語數」、「完整句字數」及「好義原個數」，以下分別說明及描述此五項表面特徵的抽取過程。

首先定義高分作文群乃指該作文分數等級為五、六分，而低分作文群乃指該作文分數等級為一、二分。

◆ **文章字數特徵**

通常在老師所評分的文章中，「高分作文」在文章所使用總字數上會比「低分作文」更多，這也與一般作文基本要求有關，不足規定字數的文章並不會獲得過高的評價。

觀察所有訓練文章中(見圖表 *5*)，作文文章所使用總字數分佈，高分群與低分群有相當的鑑別力。高分群的平均文章使用總字數為 *398* 個字，而低分群的平均文章使用總字數為 *245* 個字。



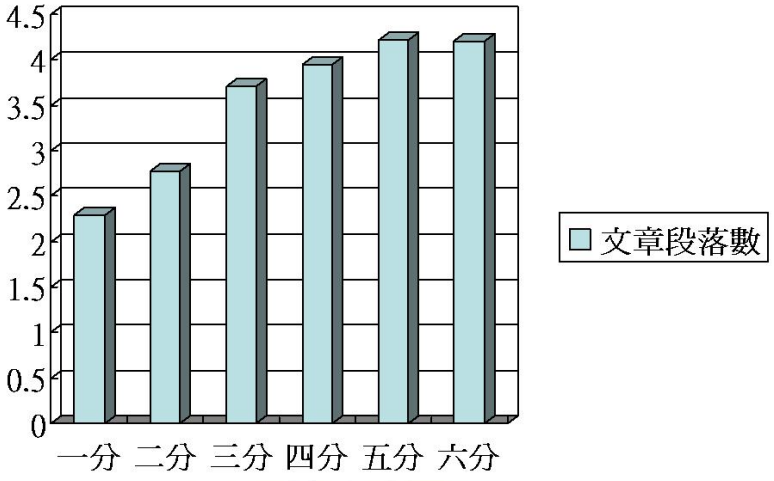
圖表 5　訓練資料中文章字數分佈

基於此項觀察，因此將「文章所使用總字數」納入系統評分的表面特徵中。

◆ **文章段落數特徵**

雖然在中國古代的各種書籍典冊上，文章通常不分段，但是一篇未分段的文章可想而知，閱讀起來勢必格外吃力。因此在現今的命題作文基本假設上，皆要求需對文章作基本分段，而一般作文常見的分段法則有「三段法」為結構*引言*、*正文*及*總結*與「四段法」為結構*起*、*承*、*轉*、*合*。藉由文章的分段效果，可使文章層次看起來較清楚，也較容易使讀者掌握文意。

觀察所有訓練文章中(見圖表 *6)*，作文文章所採用分段數分佈，高分群與低分群亦有相當的鑑別力。高分群的平均文章採用分段數為 *4.209* 段，而低分群的平均文章採用分段數為 *2.537* 段。



圖表 6　訓練資料中文章段落數分佈

基於此項觀察，因此將「文章所採用分段數」納入系統評分的表面特徵中。

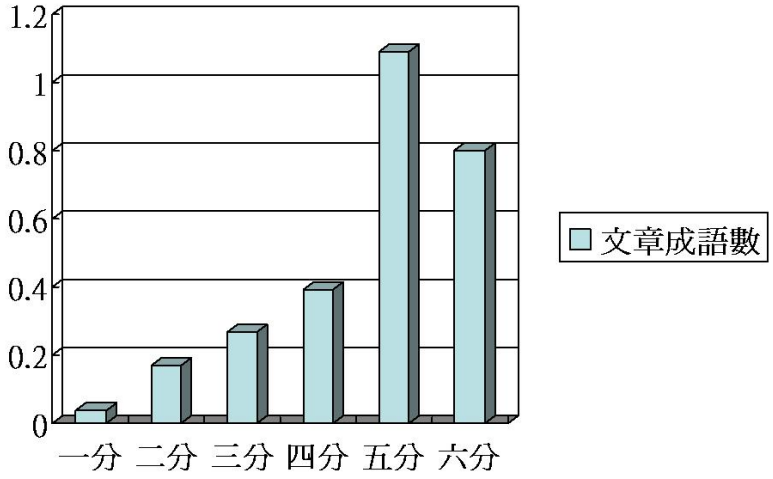
◆ **文章所使用成語數特徵**

中國歷代文化悠久，其文學領域更是廣闊，浩如煙海，而最能代表中國文學精髓的，說來簡單，就是我們常說的「成語」。

中國成語有它的特點，而且大都有經有典可查，而大部份更是文學作品上的名言錦句，它非但有很強的概括力，而且更能反映在生活的遭遇上或作行為道德的見證。

就講述某件事實來說，可能用了幾十個字還說不明白道理，但只要適當地用一句成語，往往就非常出色，使人完全領會，在寥寥數字裡卻寓意著貼切而深刻的道理。因此成語的掌握，對於一篇作文來講，也有著極為正面的加分作用。

觀察所有訓練文章中(見圖表 7)，作文文章所使用的成語比例上，高分群與低分群亦有相當的鑑別力。高分群的平均文章使用成語數為 *0.945* 個，而低分群的平均文章使用成語數為 *0.218* 個。



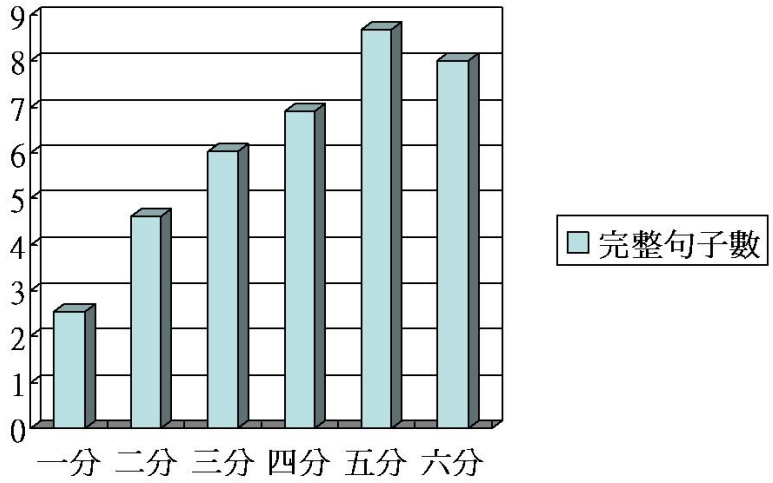
圖表 7　訓練資料中文章所採用成語數分佈

基於此項觀察，因此將「文章所採用成語數」納入系統評分的表面特徵中。

◆ **完整句子數特徵**

對閱讀者來講，適當的句子結尾更能方便讀者閱讀。試想當一段文章從頭到尾都沒有結束符號的話，真的很難讓人能一口氣閱讀完。

觀察所有訓練文章中(見圖表 *8*)，作文文章所使用的完整句子數上，高分群與低分群亦有相當的鑑別力。高分群的平均文章使用完整句子數為 *7.8* 句，而低分群的平均文章使用完整句子數為 *3.5* 句。



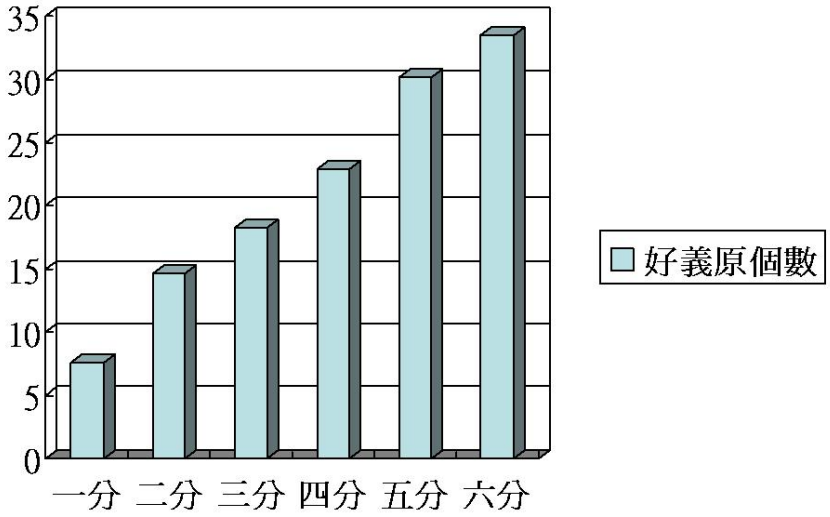
圖表 8　訓練資料中文章所採用完整句子數分佈

基於此項觀察，因此將「文章所採用完整句子數」納入系統評分的表面特徵中。

◆ **好義原個數特徵**

首先對「好義原」的定義作說明，所謂的好義原係指在訓練資料中，收集某些特定義原，而這些義原在文章高分群裡出現頻率甚高於在文章低分群的出現頻率。而本系統從訓練資料中所收集的好義原數量共計 *229* 個。

觀察所有訓練文章中(見圖表 *9*)，作文文章所使用的完整句子數上，高分群與低分群亦有相當的鑑別力。高分群的平均文章使用好義原個數為 *31.81* 個，而低分群的平均文章使用好義原個數為 *11.12* 個。



圖表 9　訓練資料中文章所採用好義原個數分佈

基於此項觀察，因此將「文章所採用完整句子數」納入系統評分的表面特徵中。

3.3 *Constructing Feature Space* – 建構特徵義原空間

首先定義「特徵義原空間」*(Feature Space)* 為收集某些特定義原的集合空間，而這些被收集的義原，在某種程度上即作為該類別的特徵代表群。而在此模組下，主要目的即是對一分至六分的分數類別各建立所屬的特徵義原空間，而每一個特徵義原空間中包含了與這個分數類別(即一分至六分)關聯緊密的若干個主要特徵義原。

從上述定義得知，建構特徵義原空間乃是收集與該類別關聯緊密的主要特徵義原，因此我們利用互訊息 *MI*[9][10]來衡量每一個義原與該類別的關聯程度，透過計算訓練資料中出現的每一個義原與每一個分數類別的互訊息 *MI* 來確定此義原是否屬於該類別的特徵義原空間。

首先經由文件頻率 *DF*[9][10]對文章中每一個義原做統計，若該義原的 *DF* 值低於系統所設立的門檻植，則將此義原去除，留下來的義原再去做互訊息計算。對於每一個義原 *t* 與分數類別 *c*，則計算 *t* 與 *c* 的互訊息 *MI* 可以利用下面的公式來計算：

$$MI(t,c) \approx \log \frac{A \times N}{(A+C) \times (A+B)}$$

A = 包含該義原 *t* 且屬於分數類別 *c* 的文章總數

B = 包含該義原 *t* 但不屬於分數類別 *c* 的文章總數

C = 屬於分數類別 *c* 但不包含該義原 *t* 的文章總數

N = 所有訓練資料的文章總數

同樣地，系統事先設定一個門檻值，當 *MI(t,c)* 值高於此門檻值時，便可以認為該義原 *t* 屬於分數類別 *c*，即將該義原 *t* 加入於分數類別 *c* 的特徵義原空間。在計算每一個義原與各個分數類別的互訊息 *MI* 後，即可建立系統所需的六個(分別為一分至六分)特徵義原空間。

而值得注意的是，利用此方法建立的特徵義原空間，某些義原可能同時屬於多個分數類別的特徵義原空間(例如「Advantage|利」一義原，同時屬於四分、五分及六分的特徵義原空間)。表格 *1* 描述了六個特徵義原空間間互相的重覆關係，舉六分的特徵義原空間來看，其所包含的義原共有 *139* 個，其中，與一分的之間有 *15* 個重覆，與二分的有 *29* 個重覆，與三分的有 *52* 個重覆，與四分的有 *87* 個重覆，與五分的則有 *110* 個重覆。

	義原個數	一分	二分	三分	四分	五分	六分
一分	77	-	20	33	35	42	15
二分	151	20	-	44	80	83	29
三分	207	33	44	-	91	125	52
四分	308	35	80	91	-	203	87
五分	334	42	83	125	203	-	110
六分	139	15	29	52	87	110	-

表格 1　特徵義原空間重覆關係圖

關於本論文所定義的「特徵義原空間」可以這樣解讀，特徵義原空間的目的是希望在真正分類之前為各個分數類別建立起該類別的特徵知識，而這些特徵知識的具體表現即為特徵義原空間。

3.4 *SVM Training* – *SVM* 訓練階段

有了上述所建構的六個分數類別特徵義原空間後，此模組第一份工作在於建立系統所需的 *6* 個主要特徵。在此我們引入常用的 *TF-IDF* 公式，負責計算文章中的義原特徵權重值，針對文章 *d* 中出現的義原 *t*，我們計算義原 *t* 的特徵權重 *W(t,d)* 值，而使用的是一般常用的 *TF-IDF* 公式如下：

$$W(t,d)=\frac{tf(t,d)\times\log(N/n_t+0.01)}{\sqrt{\sum_{t\in d}[tf(t,d)\times\log(N/n_t+0.01)]^2}}$$

tf(t,d) = 義原 *t* 在文章 *d* 中的出現頻率

N = 訓練資料中的文章總數

n_t = 訓練資料中出現義原 *t* 的文章總數

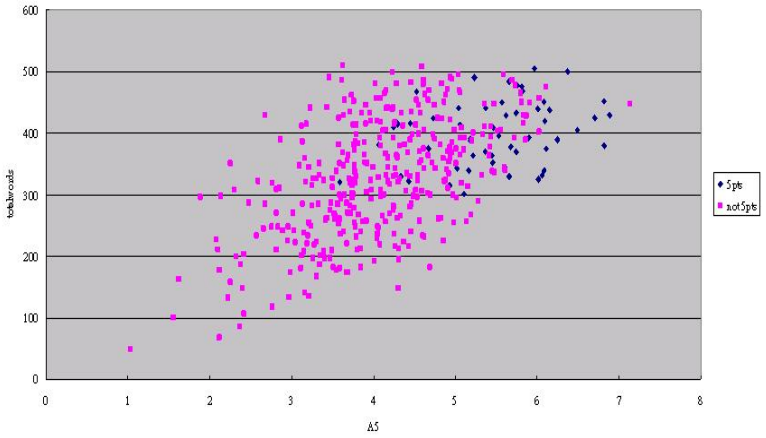
接下來，我們建立系統最主要的 *6* 個特徵 *(A1,A2,…,A6)*，初始值設為 *0*。再來根據文章中的每一個義原 t，看義原 t 屬於哪一個「特徵義原空間」，則系

統的主要特徵*(A1,A2,…,A6)*，跟著加總該義原 *t* 的特徵權重 *W(t,d)*值，直到文章中所有義原處理完畢，系統的 *6* 個主要特徵即生成。

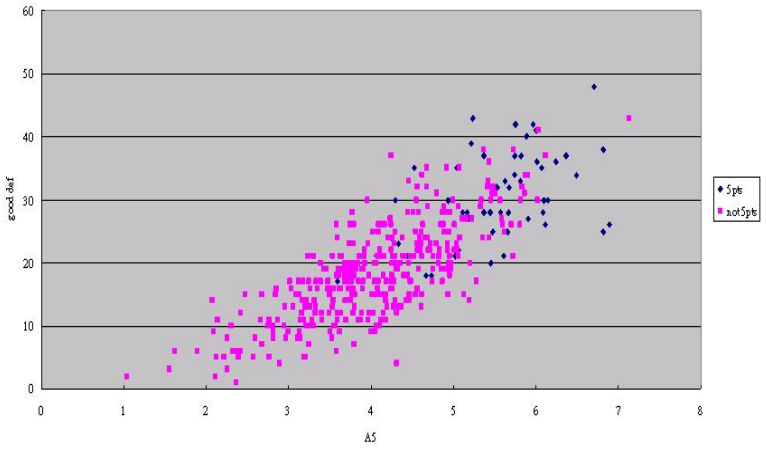
以下我們舉個例子來說明此方法，考慮一文章 *d* 中共有 *3* 個義原分別為 *t1*、*t2* 及 *t3*，而其特徵權重值分別為 *w1*、*w2* 及 *w3*，而 *t1* 屬於一分的特徵義原空間，*t2* 屬於二分、三分、四分的特徵義原空間，*t3* 屬於四分、五分的特徵義原空間。一開始系統的主要特徵 *A1=0, A2=0, A3=0, A4=0, A5=0, A6=0*。因為 *t1* 只屬於一分的特徵義原空間，因此只有 *A1* 的值需作變更，系統的主要特徵值變為 *A1=w1, A2=0, A3=0, A4=0, A5=0, A6=0*。而 *t2* 屬於二分、三分及四分的特徵義原空間，因此 *A2*、*A3* 及 *A4* 的值皆需作變更，系統的主要特徵值變為 *A1=w1, A2=w2, A3=w2, A4=w2, A5=0, A6=0*。最後 *t3* 屬於四分及五分的特徵義原空間，因此 *A4* 及 *A5* 需作變更，系統的主要特徵值變為 *A1=w1, A2=w2, A3=w2, A4=w2+w3, A5=w3, A6=0*。而此最後的 *6* 個特徵值即為系統所需主要特徵。

最後一步，有了 *6* 個主要特徵值*(A1,A2,…,A6)*後，再加上之前建立的 *5* 個表面特徵(包括文章字數、文章段落數、文章所使用成語數、完整句子數及好義原個數)，總合而成系統最後需要的 *11* 個訓練特徵。

為了更了解各特徵對於文章分數類別的鑑別性，我們將各特徵以二維的方式表達，底下僅擷取二張示意圖來解釋。圖表 *10* 及圖表 *11* 分別為訓練資料中，五分文章與其它非五分文章的「*A5* 值」相對「文章字數」二維特徵關係圖及五分文章與其它非五分文章的「*A5* 值」與「好義原個數」二維特徵關係圖。其中，標為藍色的點為五分文章在二維空間中的相對位置，而標為紅色的點為其它非五分文章在二維空間中的相對位置。在可允許的誤差下，藍色的點與紅色的點在二維空間下，已存在著某種程度的分離性。



圖表 10　五分與非五分二維特徵對應圖之一



圖表 11　分與非五分二維特徵對應圖之二

因此，有了這 *11* 個訓練特徵後，接下來直接經由支援向量機作訓練動作，在此我們使用的支援向量機分類器是由台灣大學*林智仁*老師所提供的 *LIBSVM-- A Library for Support Vector Machines*[12]，實驗中的 *kernel function* 採用了 *RBF* 函數，並採用交叉比對*(Cross Validation)*的方法。

3.5 *SVM Predicting Model* – *SVM* 預測模型

此模組是整個系統最後的評分依據，主要的預測模型係根據之前的支援向量機訓練模組而來。而測試資料在這個模組下，一樣經由與「*SVM Training – SVM 訓練階段*」模組同樣的特徵建立方式，依據訓練資料的特徵義原空間資訊，產生 *6* 個主要的系統特徵*(A1,A2,…,A6)*，再加上 *5* 個表面特徵(包括文章字數、文章段落數、文章所使用成語數、完整句子數及好義原個數)，總合而成最後系統所

需的 *11* 個特徵，供支援向量機預測模型作最後的評分。

3.6 *Google-Based SyntaxError Detector* – 語法錯誤偵測

此模組是為改善「詞袋式」的自然語言處理方式所帶來的缺點。以往的中文作文評分系統係採用詞袋式的語言處理方式，因此常見的問題為，斷好詞後的文章，隨機亂調其排列順序，對於最後的評分並沒有影響。亦即，有可能出現一篇不通順的文章，但其經由系統評分後，仍得到很高的分數。考慮以下的片段文章：

「 條 …
問 …
地 …
解 噹噹 來，
愉快 說 愉快 的 諄諄 的 時間 著 — —
面對 ，
一段 不斷 座位 是 十 分鐘 惟有 說上 分鐘 ，
纏 一樣 點 這時 著 印記 ，
此 總是 著 同學 、
學生 的 上課 則 。
的 對於 的 像是 需 ，
馬上 的 只有 來 大過 在 。
的 校園 及格 堂 三五 是 教誨 ，
的 便是 的 渲染 在 工具 了 ；
…」

上篇文章片段係由，原始資料中高分(六分)作文的某篇文章，在斷好詞後隨機改變其排列順序，而這樣的一篇文章在經由系統評分後，仍然得到很高的分數，此乃由於其義原的統計數量分佈並不會因為這樣的隨機排列而改變。因此，對於最後的評分與原本通順的文章相比結果仍相同一致。而這樣的結果，對於系統來說存在著某種程度上的攻擊後門。

更極端的例子，由於以往的中文作文評分系統，係採用「好義原」的分佈來做為主要特徵，考慮以下片段文章：

「 涼亭 一路 東張西望 開懷 ，
不約而同 放眼 操場 可怕 大街 進進出出 最多 。
動向 好處 光顧 選美 ，

低頭　人來人往　壞事　大大小小　喜歡　重要　表面　。
果汁　不停　嬉鬧　電梯　影子　，
漫長　口水　生命力　佩服　歡樂　選擇　妨礙　。
光臨　自由自在　抬頭　找著　階段　過癮　大大小小　！
朝氣蓬勃　厲害　喜悅　時針　，
…　」

由人的眼睛來看，這的確像是隨意亂湊的文章，前後文無法相配合，但是這樣的一篇文章在交由以往使用「詞袋式」的中文作文評分系統評分下，確出乎意料地得到了相當高的評分(經測試為五分)，分析此問題得到，因為以往的中文作文評分系統，主要的評分特徵為文章中好義原的分佈比例，而在此篇文章中盡其所能的寫了很多個詞語，大大增加了好義原的出現比例，因此系統的評分就出現了盲點。

而此模組的目的，即是為了防止這樣的缺點而建立。基礎的假設是，經由自然語言處理上常用的 *bi-gram* 來偵測詞語與詞語之間是否為正常搭配，然而 *bi-gram* 最大的問題在於，其背後需要一足夠大的語料庫才能建立理想的 *bi-gram* 對應表，而這樣的語料庫往往很難實現。因此，本實驗嘗試以 *Google* 為基礎，試圖建立以 *Google-Search* 為架構的 *bi-gram*，而演算法的設計分兩層如下：

第一層：文章先經過中文斷詞後，考慮兩個相鄰的詞語對，首先一開始由本地端的語料庫所建立的 *bi-gram* 表中查詢，若找到則繼續下一個相鄰的詞語對；反之，則將這樣的詞語對，經由 *google* 去作查詢，檢查回傳的 *search-result* 數，若小於系統預設的門檻值，則儲存此詞語對及回傳的 *search-result* 數，初步列為有可能語法錯誤的候選對，而未儲存的詞語對則判斷為合理的配對。接著再繼續處理下一個相鄰的詞語對。

第二層：逐一檢查由第一層所儲存的候選對，並掌握以下兩個原則：

原則一：如果候選對的 *search-result* 數為零，則判定該候選對為語法有問題。

原則二：若不為零，則檢查是否為連續的相鄰詞語對，若是，亦判定此候選對為語法有問題。

在此用一例子來說明此演算法，考慮以下句子，

「不約而同　放眼　操場　可怕　大街　進進出出　最多　。」

而在經由第一層判斷的結果共有 *6* 個候選對，分別為：

1. 不約而同 放眼 *16*
2. 放眼 操場 *2*
3. 操場 可怕 *0*
4. 可怕 大街 *20*
5. 大街 進進出出 *0*
6. 進進出出 最多 *9*

其中，儲存格式為「詞語 *I*　詞語 *II*　*search-result* 數」

而此模組第二層將逐一檢查此 *6* 個候選對，第一個候選對「不約而同 放眼」其 *search-result* 數雖不為 *0*，但其與第二個候選對，為連續的相鄰詞語對，符合原則二，故判定為語法有問題。同理，檢查所有候選對，發現此 *6* 個候選對皆屬於語法有問題。

表格 *2* 為部份測試資料的結果，可以看出鑽系統漏洞的文章與一般正常的文章相比都有很大比例的語法錯誤被偵測到。其中，*#1*、*#2*、*#3* 及*#4* 為正常的文章，由訓練作文高分群中選出；而 *Rand(#1)*、*Rand(#2)*、*Rand(#3)* 及 *Rand(#4)* 為隨機重新排列正常文章中的詞語位置；*All Rand I* 及 *All Rand II* 為隨機選擇若干詞語排列產生而成的文章。

	Article	Detected in first layer	Detected in second layer
鑽系統漏洞的文章	All Rand I	208	169
	All Rand II	192	154
	Rand(#1)	47	18
	Rand(#2)	63	33
	Rand(#3)	41	16
	Rand(#4)	56	24
正常文章	#1	7	0
	#2	16	1
	#3	6	2
	#4	12	2

表格 2　*Google-based SyntaxErr* 測試結果

上述的文章在經由原有中文評分系統測試後，皆獲得高分的成績(至少四分)。然而蓄意鑽系統漏洞的文章，在經由此模組下，會因為過多的語法有問題處，而被系統判斷為有問題的文章，直接判定為屬於低分(一分)的文章。如此，在系統的保護上，又多增加了一層防護，避免使用者藉此鑽中文作文評分系統的漏洞，對系統造成嚴重的錯誤評分。

第四章、實驗過程與結果討論

本章節將介紹本實驗的過程與結果來證明本系統的有效性。4.1 節首先說明本系統所使用的實驗資料來源。4.2 節介紹本實驗的系統訓練流程及系統測試流程。4.3 節介紹本實驗所採用的評價方法數學公式。最後一節 4.4 為本實驗的數據結果。

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，為了維持學生作文最原始面貌，所以保留所有原始特徵，包括原有的錯字以及標點符號，不加以修改。經過評分的作文都是從一分到六分，而一分為最低，六分為最高。每篇作文係經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 689 篇作文。

4.2 實驗流程

本實驗流程共分為兩大階段，第一階段是系統訓練，第二階段是系統測試。在訓練階段中，系統從所收集到的 689 篇作文中，從每個不同分數類別(一分至六分)中隨機挑選約二分之一的數量共計 343 篇作文做為訓練資料，而在訓練階段完成後，系統會產生 6 個分數類別的「特徵義原空間」，最後導出「*SVM Predicting Model – SVM 預測模型*」模組來供最後的作文評分。而在測試階段中，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，系統會根據訓練階段產生的 6 個「特徵義原空間」資訊並加上表面特徵的資訊來產生一組特徵 *(attributes)*，供「*SVM Predicting Model – SVM 預測模型*」模組來做最後的作文評分，最後比較系統評分的等級與其它評分系統相比評分等級差異，來計算系統的正確率。

4.3 評價方法

本實驗主要採用的評價方法是準確率 *P(Precision)*、召回率 *R(Recall)*及 *F1* 值等常用的評價方法，而其數學公式表示如下：

j 分的準確率：$P_j = (l_j / m_j) \times 100\%$

其中，l_j 為 *j* 分評分正確的文章數，m_j 為系統實際評分為 *j* 分的文章數

j 分的召回率：$R_j = (l_j / n_j) \times 100\%$

其中，l_j 為 *j* 分評分正確的文章數，n_j 為 *j* 分實際所包含的文章數

j 分的 *F1* 值：$F1_j = \frac{P_j \times R_j \times 2}{P_j + R_j}$

最後並引入總體平均*(Macro)*和個體平均*(Micro)*兩種計算準確率、召回率及 *F1* 值的方法。具體定義如下：

總體平均準確率：$MacroP = \frac{1}{n}\sum_{j=1}^{n} P_j$

總體平均召回率：$MacroR = \frac{1}{n}\sum_{j=1}^{n} R_j$

總體平均 *F1* 值：$MacroF1 = \frac{MacroP \times MacroR \times 2}{MacroP + MacroR}$

個體平均準確率：$MicroP = \sum_{j=1}^{n} l_j \Big/ \sum_{j=1}^{n} m_j$

個體平均召回率：$MicroR = \sum_{j=1}^{n} l_j \Big/ \sum_{j=1}^{n} n_j$

個體平均 *F1* 值：$MicroF1 = \frac{MicroP \times MicroR \times 2}{MicroP + MicroR}$

本實驗在最後分析不同的訓練集大小*(Training Size)*對於支援向量機在作文評分分類上，另外採用兩種評價方法為 *Adjacent* 及 *Exact*：

Adjacent ：容許一分誤差的整體正確率

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 = 容許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，因此在此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。

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。

4.4 實驗結果與討論

在本實驗中，從蒐集到所有的 *689* 篇作文中，在每個不同等級中隨機挑選約二分之一數量共計 *343* 篇作文做為訓練資料，剩餘的 *346* 篇作為測試資料，實驗結果與原有評分系統比較如下：

針對各分數類別的準確率 ***P(Precision)***：

系統 / 準確率	原始評分系統	SVM-base評分系統
P_{1pt}（一分準確率）	*	73.08 %
P_{2pts}(二分準確率)	34.57 %	68.29 %
P_{3pts}(三分準確率)	37.14 %	49.12 %
P_{4pts}(四分準確率)	44.26 %	51.35 %
P_{5pts}(五分準確率)	47.05 %	58.49 %
P_{6pts}(六分準確率)	*	*

* ：系統未分到此類別

表格 3　實驗結果準確率比較

針對各分數類別的召回率 ***R(Recall)***：

系統 / 召回率	原始評分系統	SVM-based評分系統
R_{1pt}（一分召回率）	0.00 %	82.61 %
R_{2pts}(二分召回率)	38.89 %	43.75 %
R_{3pts}(三分召回率)	52.00 %	53.33 %
R_{4pts}(四分召回率)	36.00 %	54.80 %
R_{5pts}(五分召回率)	26.67 %	67.39 %
R_{6pts}(六分召回率)	0.00 %	0.00 %

表格 4　實驗結果召回率比較

針對各分數類別的 *F1* 值：

系統 / F1值	原始評分系統	SVM-based評分系統
$F1_{1pt}$（一分F1值）	0.00 %	77.55 %
$F1_{2pts}$(二分F1值)	36.60 %	53.33 %
$F1_{3pts}$(三分F1值)	43.33 %	51.14 %
$F1_{4pts}$(四分F1值)	39.70 %	53.02 %
$F1_{5pts}$(五分F1值)	34.04 %	62.63 %
$F1_{6pts}$(六分F1值)	0.00 %	0.00 %

表格 5　實驗結果 *F1* 值比較

本系統與原有的評分系統相比皆有很高的效能提昇，其中，表格 *3* 在系統準確率 ***P(Precision)*** 方面，分數類別二分至五分分別有著 *34.57%→68.29%*、*37.14%→49.12%*、*44.26%→51.35%* 及 *47.05%→58.49%* 的效能提昇，其中在二分的效能提昇更有將近一倍的顯著表現，而整體提昇幅度在 *7.09% ~ 33.72%*；而表格 *4* 在系統召回率 ***R(Recall)*** 方面，分數類別一分至五分有著 *0%→82.61%*、*38.89%→43.75%*、*52%→53.33%*、*36%→54.80%* 及 *26.67%→67.39%* 的效能提昇，而在一分及五分的效能提昇更是有大幅度的成長表現，而整體提昇幅度在 *1.33% ~ 82.61%*；而表格 *5* 在系統 *F1* 值方面，分數類別一分至五分有著 *0%→77.55%*、*36.6%→53.33%*、*43.33%→51.14%*、*39.7%→53.02%* 及 *34.04%→62.63%* 的效能提昇，而在一分及五分的效能提昇更是有大幅度的成長表現，而整體提昇幅度在 *7.81% ~ 77.55%*。

實驗結果的總體平均準確率(*MacroP*)與個體平均準確率(*MicroP*)比較：

	Original System	SVM-based System
MacroP	40.75 %	60.07 %
MicroP	38.46 %	55.20 %

表格 6　實驗結果平均準確率比較

實驗結果的總體平均召回率(*MacroR*)與個體平均召回率(*MicroR*)比較：

	Original System	SVM-based System
MacroR	25.59 %	50.31 %
MicroR	38.46 %	55.20 %

表格 7　實驗結果平均召回率比較

實驗結果的總體平均 *F1* 值(*MacroF1*)與個體平均 *F1* 值(*MacroF1*)比較：

	Original System	SVM-based System
MacroF1	31.43%	54.76 %
MicroF1	38.46 %	55.20 %

表格 8　實驗結果平均 *F1* 值比較

本系統與原有的評分系統相比皆有很高的效能提昇，其中，表格 *6* 在系統總體平均準確率(*MacroP*)與個體平均準確率(*MicroP*)方面，有著 *40.75%→60.07%* 及 *38.46%→55.2%* 的效能提昇；而表格 *7* 在系統總體平均召回率(*MacroR*)與個體平均召回率(*MicroR*)方面，有著 *30.71%→60.37%* 及 *38.46%→55.2%* 的效能提昇；而表格 *8* 在系統總體平均 *F1* 值(*MacroF1*)與個體平均 *F1* 值(*MacroF1*)方面，有著 *35.02%→60.22%* 及 *38.46%→55.2%* 的效能提昇。

本實驗最後另外對支援向量機模型在處理中文作文評分分類上，對於不同的 *Training Size* 做實驗比較分析，對照組為不同老師的評分比較，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分分數的差距，*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 *9*。

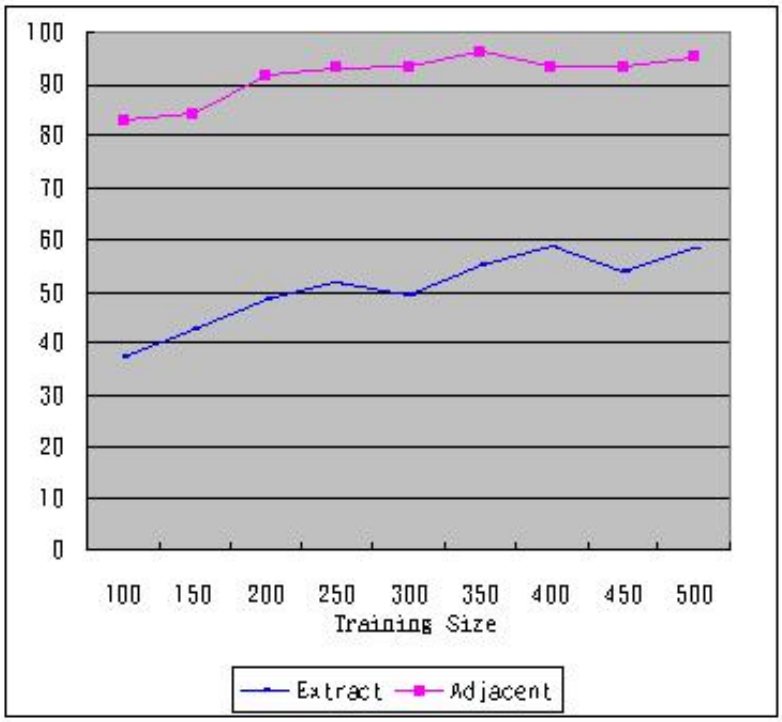
	Exact	*Adjacent*
整體	33.42 %	78.14 %

表格 9　老師之間評分差距

而不同 *Training Size* 實驗流程分別從蒐集到所有的 *689* 篇作文中隨機挑選 *100,150,…,500* 篇作文做為訓練資料，剩餘的篇數作為測試資料。表格 *10* 顯示其實驗結果。

Training Size	*Exact*	*Adjacent*
100	37.44 %	82.91 %
150	42.46 %	84.36 %
200	48.45 %	91.75 %
250	51.72 %	92.87 %
300	49.22 %	93.52 %
350	55.06 %	96.13 %
400	58.39 %	93.36 %
450	53.81 %	93.22 %
500	58.06 %	95.16 %

表格 10　不同訓練集大小正確率比較



圖表 12　不同訓練集大小正確率折線圖

圖表 *12* 為將表格 *10* 的實驗結果畫成折線圖來看，當訓練集大小大於總資料的二分之一時，其系統的 ***Exact*** 值已有 *50%* 以上的表現，而系統的 ***Adjacent*** 值也有超過 *90%* 的表現，再跟表格 *9* 的老師之間評分表現做比較，發現本系統在這兩項評分項目中，***Exact*** 值與 ***Adjacent*** 值皆比老師之間的 ***Exact*** 值與 ***Adjacent*** 值高出很多，這也代表著本系統的評分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

本論文針對中文作文評分系統，提出一種基於「特徵義原空間」的方法，並使用支援向量機理論模型來做系統分類。此方法是透過為每一個分數類別建立各自專屬的特徵義原空間，並藉此作為重要特徵的合併動作，這種方法可以看成是從訓練語料庫中自動統計出每一個分數類別所屬的特徵知識，並將其用於特徵抽取上。另外，本論文也對中文作文評分系統，有可能遭遇到的後門攻擊，提出了一種基於 *Google Search* 的偵測模組，負責避免鑽評分機制漏洞的文章，造成系統嚴重的評分錯誤。

透過這種「組合特徵抽取」的方法，結合傳統文件 *DF* 頻率與互訊息 *MI* 並基於特徵義原空間的方法結合起來，對於改善中文作文的評分分類有著顯著的效果。在系統整體的精準正確率*(Extract)*上，可以達到 *55.20%*，比原有的評分系統 *38.46%*，提高了 *16.74%* 的系統效能。而在容許一分誤差下的正確率*(Adjacent)*上，可以達到 *96.82%*，比原有的評分系統 *94.46%*，提高了 *2.36%* 的系統效能。而這也正表示本系統評分作文的結果與專業人員(老師)評分的結果相當接近，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分作文時所使用的工具，尤其是在大量的作文評分時(如聯考)，更是可節省大量的人力資源。

5.2 未來工作

本論文提出以「特徵義原空間」為基礎的自動作文評分系統，其中對作文概念的轉換是透過知網*(HowNet)*為工具，在本系統中所使用的義原，僅運用了知網描述式中的第一個義原來作為該詞語的概念，這樣捨棄了知網中概念與義原之間複雜的關係。在未來的工作上，可以此為方向，找出知網中概念與義原彼此之間的關係，進而對文章語意更了解，增進系統在作文評分的效能。

參考文獻

[1] Hearst, M.. The debate on automated essay grading. *IEEE Intelligent Systems* (2003), 15(5), 22-37, IEEE CS Press.

[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] Thomas K Landauer, Darrell Laham, Peter W. Foltz. Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor. *Automated Essay Scoring: A Cross-Disciplinary Perspective* (2003), pp. 87-112

[4] 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

[5] 劉群, 李素建. 基於《知網》的辭彙語義相似度計算

[6] 蔡沛言. 自動建構中文作文評分系統：產生、篩選與評估(2005)

[7] V. Vapnik,「Statistical Learning Theory, John Wiley and Sons,」New York, 1998

[8] 藍紹緯. 基於平滑支向機之網站入侵系統(2004)

[9] 黃飛燕等. 中文文本分類中特徵抽取方法的比較研究[J].中文信息學報, 2003,18(1):26-32

[10] 趙明生等. 中文文本分類中的特徵選擇研究[J].中文信息學報, 2003,18(3):17-23

[11] 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
URL：http://ckipsvr.iis.sinica.edu.tw/

[12] LIBSVM-- A Library for Support Vector Machines
URL：http://www.csie.ntu.edu.tw/~cjlin/libsvm/