國 立 交 通 大 學

多媒體工程研究所

碩 士 論 文

中 文 情 書 自 動 產 生 系 統

Automatic Chinese Love Letter Generation System

研 究 生：陳智維

指導教授：李嘉晃 教授

中 華 民 國 九 十 七 年 六 月

中 文 情 書 自 動 產 生 系 統

The Automatic System Produced The Chinese Love Letter

研 究 生：陳智維　　Student：Chih-Wei Chen

指導教授：李嘉晃　　Advisor：Chia-Hoang Lee

國 立 交 通 大 學

多 媒 體 工 程 研 究 所

碩 士 論 文

A Thesis

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

中華民國九十七年六月

中 文 情 書 自 動 產 生 系 統

學生 ： 陳智維　　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

本文以數百篇情書做為系統產生的基底，配合中央研究院的平衡語料庫，實做出一個中文情書自動產生系統。首先將情書語料庫和平衡語料庫的文章進行分解、重新整理，再利用處理過後產生的文字片段組合出一篇新的情書文章。將文章中最有意義的關鍵字取出，並以隱藏方式延展擴增成一關鍵字串列，再用填充詞的方法將關鍵字予以串連，以成為一篇全新的情書文章。

產生出來的情書文章無論在取材、句型、創新、文意上都有不錯的效果。本系統仍在發展階段，所以還有改善的空間，但現階段已表達出一般人在寫作時的進行方式與情況。

The Automatic System Produced The Chinese Love Letter

Student ： Chih-Wei Chen Advisor：Prof. Chia-Hoang Lee

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

Abstract

In this paper, we proposed and implemented an automatic system producing the Chinese love letter ,which makes use of hundreds of essays about love letter topic and corporate with Sinica Corpus to produce new love letter. First, rearrange the corpus based on the love letter collection and corpus from Sinica Corpus and then combine the material to produce new love letter. The most meaningful keywords in the article will be retrieved ,and extend the method to hide amplified into a serial keyword, and then filled with the word method will be the keyword link to become a new love letter article.

As our experience with a first simple prototype has shown that this approach could produce the love letters. There is a good result no matter materials, on sentence pattern, innovation, gentle purpose in the love letter article produced out. This system is still during the course of developing, so there is improved space, but has already expressed carrying on the track type and situation while writing of common people at the present stage.

目錄

圖目錄

表目錄

第一章、緒論

1.1 研究動機

自動文本產生在自然語言處理中是一個奧妙且有趣的主題，一般人在寫一篇情書時，會費盡心思的將文章的架構、論點、主旨等文章內容的骨幹先行構思好，並且在寫的過程中，盡量優化用字遣詞且加入修辭，借以來加強讀者對文章的感受程度。

在一些實驗中發現，不同文章中，主題相同的兩個句子，彼此接合後極有機會變成一句語意語法皆通順，且在其他篇文章皆未曾出現過的新句子。其中，又發覺人們在寫文章時，常常會有下一句不知道要寫什麼而進入深思的情況，此時就會在他的腦中搜尋曾經閱讀過的文章，將其可用的句子套到他現在所寫的文章中，這樣的思考方式，與本系統的構想很相近，以上述方法為基礎，將系統給實作出來。

1.2 研究目的與假設

如前文所述，寫作不是件簡單且深奧的事情，但如果能靠電腦自動幫忙產生的話，可以讓現代繁忙的人節省很多時間，而且說不定會創造出一些有趣話語的用法。因此，本文利用了網路上各類的情書，共 446 篇語料庫做為產生文章的基底，再配合中研院的平衡語料庫，利用語料庫的片段詞句去產生新的情書文章，其中創意與有趣的效果是明顯的，目的是讓使用者能夠縮短其寫作的時間外，更能產生意想不到的文章效果。

1.3 論文架構

第一章：前言，描述本文的研究動機、目的，將本文系統的初衷和基本概念做一個介紹。

第二章：相關研究，說明自動文本產生的概念和實行方法，以及已有哪些相關研究。

第三章：系統設計，將前置作業和系統的整體架構做一個完整介紹。

第四章：實驗過程與結果討論，將實作出來的系統做一些比較與討論，分為情書系統與填充字串2個部份。

第五章：系統的結論與展望，將系統做總結和探討系統未來方向。

第二章、相關研究

2.1 內容選擇

內容選擇在文本產生中是一個重要的部份，當一個內容要如何被選擇，就是考驗自然語言產生的技術面，當要決定一個自然語言產生系統在文本產生時該包含哪些內容，這類的系統通常都需要有很大的資料庫來支撐。內容選擇最重要的一點是選出的內容需要考慮到連慣性，讓產生出來的文章能夠看起來通順、連慣。

現今許多內容選擇的系統，都只是將已經知道的部份讓使用者輸入，然後制式的將使用者輸出的選項，套回系統內的文件中，其感覺只是將已知的元件套進資料庫的文章中而已。其中，在相近的內容集合的研究上，Regina Barzilay 和 Mirella Lapata（2003）為一個足球賽事報告產生系統提出一個集合式的內容選擇模型(collective content selection)，他們考慮資料庫中所有項目的子集合，並計算每個子集合的語意關聯性，以分數最佳的子集合當作選擇的內容。

本文的系統，由於題目相近，且情書寫法比較不需要嚴謹的文法限制，因此本系統的內容選擇部分可以比較有彈性。在文章的組成部分，將文章中最重要的關鍵字取出後，這些關鍵字將擴展成為一關鍵字串列，以便作為輸出文章內容的骨幹。

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分得很清楚哪邊為一個字，哪邊為一個詞，所以我們藉助中央研究院的詞庫小組中文斷詞系統[4]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，將情書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 中文作文寫作輔助系統

[余思翰,中文作文寫作輔助系統,2008]，是第一代的文本產生系統。此系統的語料庫是採用國中生所寫的 689 篇作文來做為他的語料庫。此系統將每一篇的作文中的詞與平衡語料庫的詞做頻率的統計，將相同字詞，但排名相差 100 名以上的詞取出，且取前 300 個當做整個作文產生系統的關鍵字，再將這些關鍵字和所有的作文文章去比對，將每篇文章的所有關鍵字取出並組合成一串列，在此稱為關鍵字串列(Keyword List)，所以文章有 689 篇，則關鍵字串列就有 689 串。

如下圖 2-1 所示，使用者可在畫面上選取一串關鍵字串列，系統將關鍵字取出後就會自動把適合的填充詞給一一填入關鍵字之間,則產生了一篇全新的作文文章。

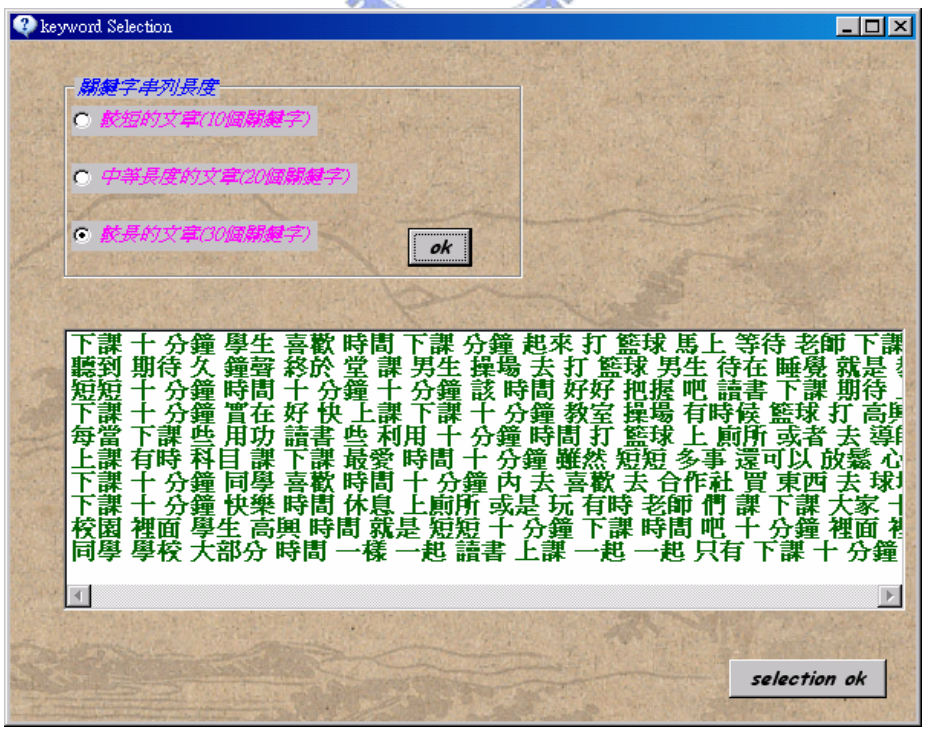


圖 2-1：中文作文寫作輔助系統，關鍵詞選取畫面

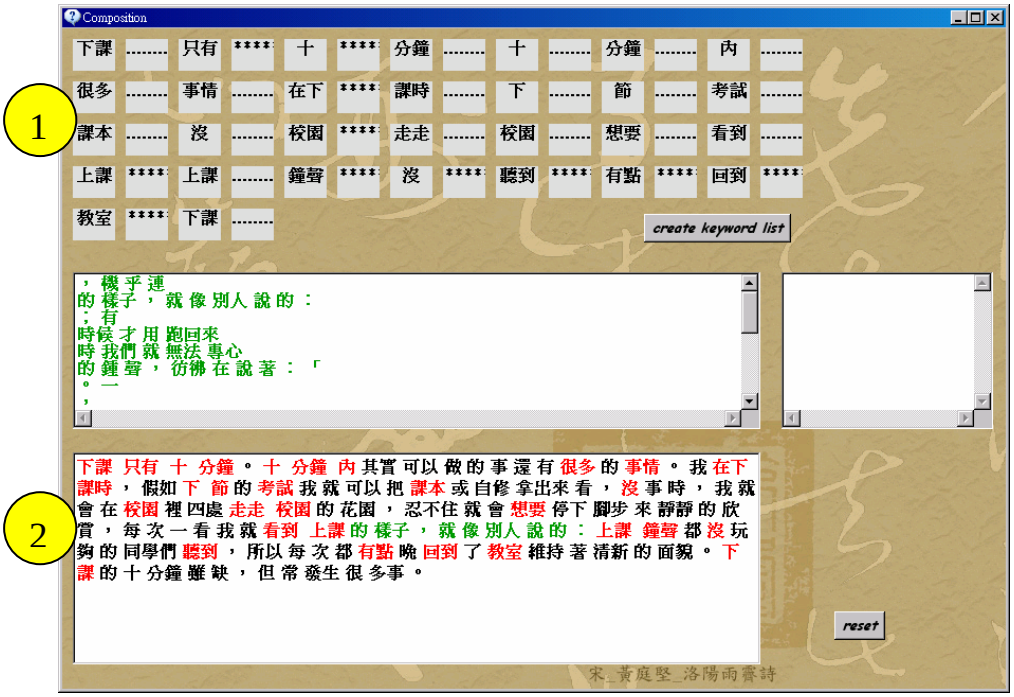


圖 2-2：中文作文寫作輔助系統，文章產生畫面

該中文作文寫作輔助系統尚有幾項缺點：

1. 關鍵字串列尚有贅詞：圖 2-1 所示，關鍵字串列，感覺並不真正像是文章中的關鍵字，串列中還是有贅詞存在，並無準確取出文章中的關鍵詞。
2. 關鍵字串列跟最後產生的文章關聯性低：圖 2-1 所示，關鍵字串列，使用者無法在觀看關鍵字時，就能預想之後文本產生的內容情況。
3. 關鍵字串列內的關鍵字太多：圖 2-2 中①所示，產生出來的關鍵字太多，使用者不易看清操作使用，30 個的關鍵字過多，使用者在操作更換填充詞，需要上下觀看，不夠 User Friendly。
4. 關鍵字串列內的關鍵字與文章長度無關：圖 2-2 中②所示，作文寫作要點，要掌握起、承、轉、合的寫法，雖然關鍵字很多，但產生出來的文章長度不夠，無法掌握寫作文的要點。

後續本系統會將上述的缺失改善，再加入自動產生本文很需要的元素，同義/同意替換與句子修辭。

第三章、系統設計

3.1 概念

早期的概念是文字接龍，是指從已知的字詞中，去決定下一個可接在其後的字詞，長度不一，後接詞句也是可長可短，藉以接出一段句子。自動文本產生，大致可分為二個部份，第一部份是決定最能代表該文章的字詞，有如人的骨頭一般，在此定義其為關鍵字(Keyword)，其中，從每篇文章所挑選出來所有關鍵字，將其串連起來，有如人的骨架，在此定義為關鍵字串列(Keyword List);第二部份是將已決定的字詞中間填入句子，即在關鍵字與關鍵字間填入適合的句子，有如人骨架上的血肉，在此定義其為填充詞，藉此將產生出一完整的文章。

3.2 系統架構

本系統共分為 4 個系統架構，第一部份為前置處理，此處將情書語料庫和平衡語料庫做一些規畫與處理，讓後續系統的進行能更有效率。第二部份為主要系統架構，分成關鍵字產生與隨機產生情書，關鍵字擷取與產生，是利用改良式的 SPLR 去做擷取，抓出文章中最有意義的詞，之後，關鍵字串列產生，與關鍵字串列擴展，並產生隨機的情書文章。第三部份為同義詞系統架構，是將充填詞做變換的部份，讓使用者可隨自己當下情境選取自己想要改成的同義詞，增加填充字詞的變化與靈活度;第四部份為譬喻系統架構，是將所選取的填充詞增加其美感的修辭方式，讓句子更具有深度與文學氣習。圖 3-1 為系統的流程架構圖，其中的每一部份流程與架構將在後續章節中詳細介紹。

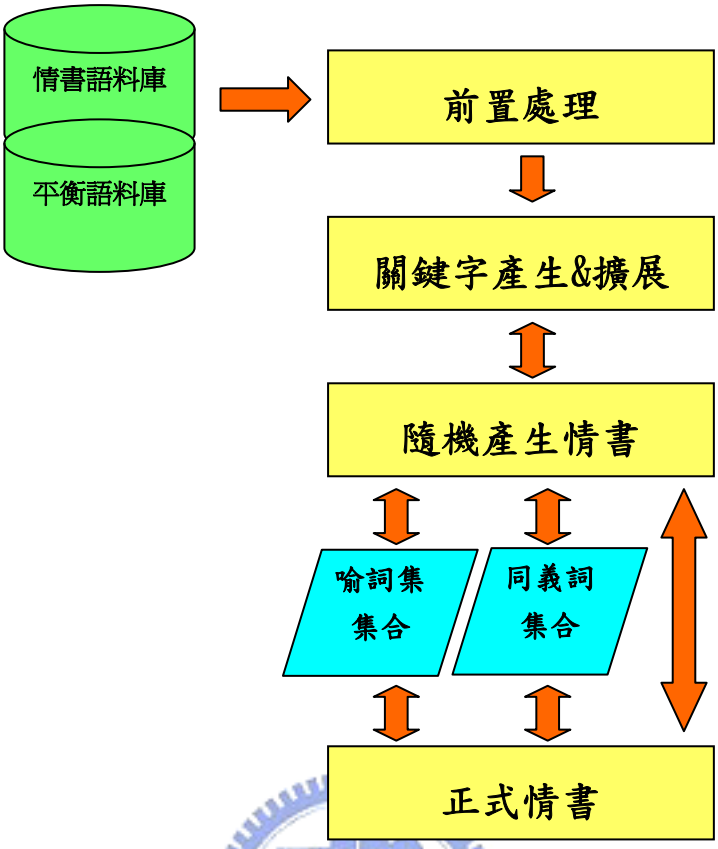


圖 3-1：系統流程架構

3.3 前置作業

本系統採用已經過斷詞處理和詞性標記後的情書語料庫與中央研究院中文詞智識庫小組的平衡語料庫，取用二組語料庫的資料做為系統的基底。其中，前置作業需要處理的資訊主要有 3 部份，第一部份為情書與平衡語料庫斷詞、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。第二部份為同義詞擷取、整理，此處是將情書與平衡語料庫內的名詞與動詞取出，並找出其同義的字與詞。第三部份為喻詞集合建立，建立一個常用喻詞集合，以方便後續修辭子系統的運作。

3.3.1 情書與平衡語料庫斷詞、詞性、結構化

文章中，最能代表一篇文章的字詞，就是名詞(Na)和動詞(VH)，所以將情書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，將每篇情書的文章都做完處理後，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。

例子：

原始文章：

漂亮(VH) 的(DE) 詩函(Na) ：(COLONCATEGORY) 久(VH) 不(D) 通函(VA) ，(COMMACATEGORY) 至(P) 以為(VE) 念(VC) 。(PERIODCATEGORY) 好(VH) 想(VE) 和(P) 你(Nh) 在一起(VH) ，(COMMACATEGORY) 即使(Cbb) 只(Da) 能(D) 靜靜(VH) 地(DE) 看(VC) 著(Di) 你(Nh) ，(COMMACATEGORY) 自己(Nh) 也(D) 心滿意足(VH) 。(PERIODCATEGORY)

結構化：

漂亮(VH) 詩函(Na) 久(VH) 好(VH) 在一起(VH) 靜靜(VH) 心滿意足(VH)……

的(DE)

：(COLONCATEGORY)

不(D) 通函(VA) ，(COMMACATEGORY)至(P) 以為(VE) 念(VC) 。

(PERIODCATEGORY)

想(VE) 和(P) 你(Nh)

，(COMMACATEGORY)即使(Cbb) 只(Da) 能(D)

地(DE) 看(VC) 著(Di) 你(Nh) ，(COMMACATEGORY)自己(Nh) 也(D)

3.3.2 同義詞擷取、整理

同上述的方法，將情書語料庫和平衡語料庫中的名詞(Na)和動詞類(VH、VJ、VC…)擷取出來，共有 80972 筆資料，過濾掉相同的字詞與詞性後，將每一個字詞放入中央研究院文國尋寶記同義反義詞遊戲倒影湖[4]中，可以得到[國語彙詞典]與[同義詞林]這二種同義詞，目前本系統是採用前者[國語彙詞典]來當做同義詞，前者的同義效果比較接近本系統所要求的效果，以此方式擷取到的同義詞，在此定義其為同義詞集合。其運作方法如圖 3-2 所示。

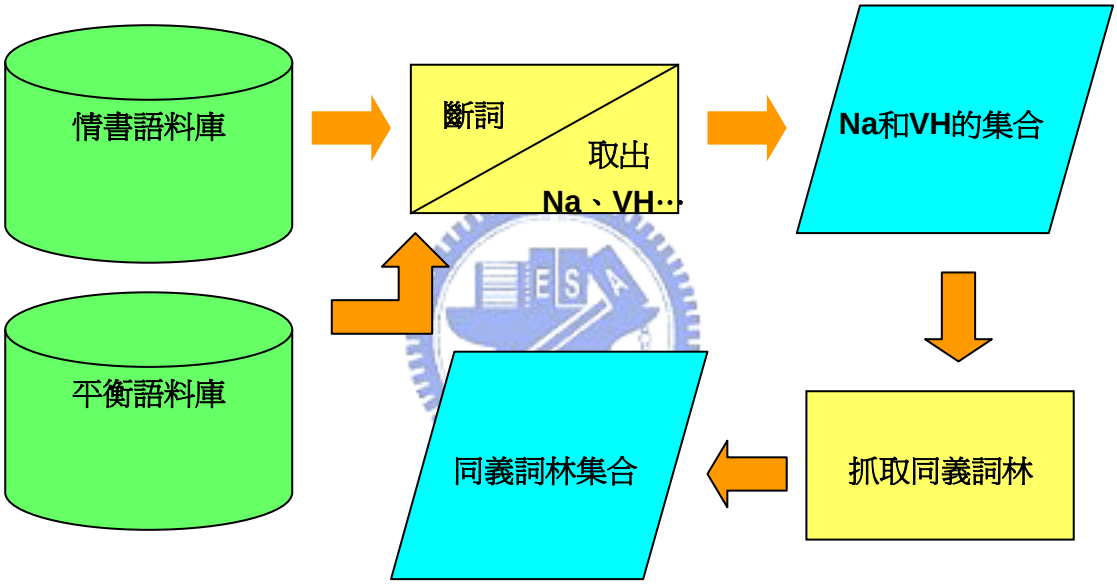


圖 3-2：同義詞流程架構

下表 3-1，表示抓取同義詞後，經結構化後的結果，每一列都代表彼此間具有相同的意思，此結構化後共有 9094 筆有效同義詞資料。

表 3-1：同義詞結構

:事過境遷	世易時移#				
:勇敢	大膽	果敢	英勇#		
:閒扯	閒談	閒聊	閒話#		

:害怕	膽怯	恐懼	懼怕	畏懼	畏怯#
:應酬	寒暄	交際	酬酢#		
:財路	財源#				
:黑幕	內幕#				
:懷胎	妊娠	孕珠#			
:憤懣	憤怒	憤慨	憤恨	怫鬱	怨憤#
:正中下懷	恰如私願#				

3.3.3 喻詞集合建立

參考了教育部和各國中、小學網站上對喻詞的解釋，自行搜集了 14 個最常用的喻詞：好像、就像、竟像、真像、有如、如同、就如、真如、好似、恰似、有若、彷彿、好比、猶如，在此定義 14 個喻詞為喻詞集合。在後續的工作，將採用這些喻詞當基底，透過各類料語庫取出可用的譬喻句子。如下例。

例：

1.好像 花朵絢爛般的 笑容

2.有如 鋼鐵的 意志

3.猶如 小鳥般 自由

3.4 主系統架構

以下介紹主系統的部份，分為 2 大部份，第一部份是 SPLR 擷取關鍵字，介紹關鍵字如何利用改良式 SPLR 的方法進行擷取，取出文章中真正重要的資訊。第二部份是關鍵字串列生成方式，是將第一部份得出來的主要關鍵字，擴展成能組成文章所需要足夠的關鍵字量。

3.4.1 SPLR 擷取關鍵字

下圖 3-3 為 SPLR 在擷取關鍵字的流程架構圖，底下我們將仔細介紹各區塊流程與架構所運行的事務。

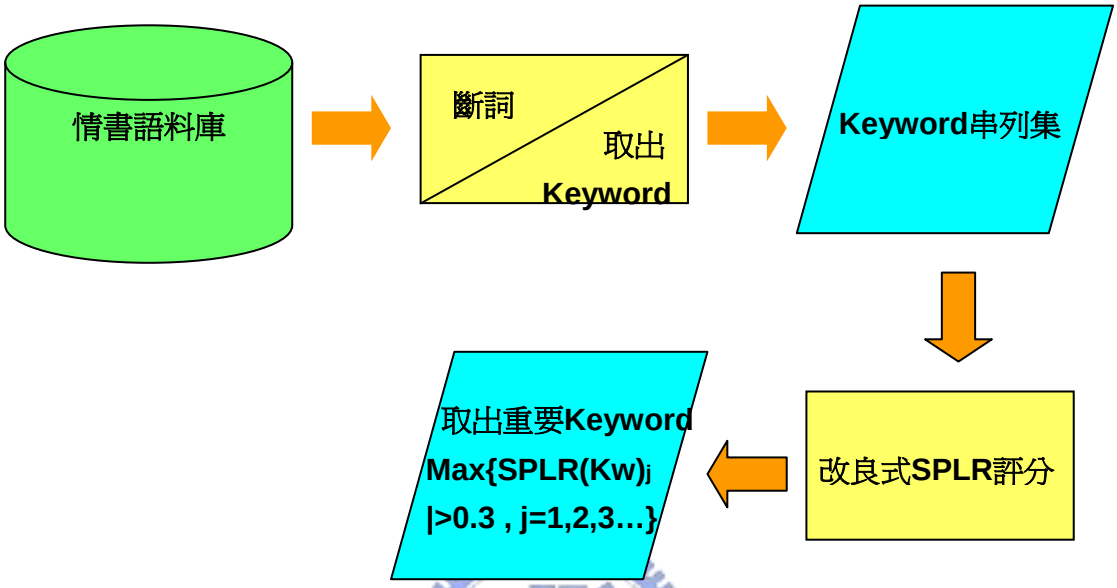


圖 3-3：關鍵字產生流程架構

首先將斷詞好的語料庫當做基底，然後從每一篇情書的文章中，取出其名詞(Na)和動詞(VH)當做主要的關鍵詞，會挑這 2 種詞性，主要是因為這 2 種詞性最有可能是代表一句話的關鍵詞，所以第一步先將名詞(Na)與動詞(VH)取出，形成只包含名詞(Na)與動詞(VH)的一串串列，因為情書語料庫共有 446 篇文章，所以總共有 446 串串列，在此稱這所有的串列為一關鍵詞串列集，如下例子所示。

瘋狂 乳加 巧克力 理想 愛情 人 那樣 時刻 深 榆葉梅開 絢爛 粉色 雲朵…

以下，先定義一些後續公式會使用到的集合，以利後續公式的推導。

關鍵字：

$$KW_i \quad , \quad \forall i = 1,2,3,4 \ldots..$$

關鍵字串列集合：

$$Keyword\_List = \{KW_i, i = 1,2,3,\cdots,n\},$$

n：表示該篇文章關鍵字總數

再來，將已取出的關鍵詞串列，使用 SPLR[5]改良後的方法去計算評分，原本 SPLR 的方法是運用在找 unknown word 的技術上，可以將 unknown word 很精準的取出來；在此概念下得到的啟發，經過多次實驗，將 SPLR 中遞迴式的方法拿掉，所得到的效果最佳。此系統利用改良式的 SPLR 方法，找出關鍵詞串列中，真正重要且能代表此篇文章的關鍵字，再展現出來讓使用者選取，改良式的 SPLR 公式如下(1)為此計算的方法。

$$SPLR = \frac{tf(KW_i)}{Max\,(tf(KW_i\_L), tf(KW_i\_R))}, \quad KW_i\_len > 1 \quad (1)$$

$$KW_i\_L = \left\{ KW_i\_L \,\middle|\, \left| \left\lceil \frac{KW_i\_len}{2} \right\rceil \right| \right\} \quad (2)$$

$$KW_i\_R = (KW_i - \{ KW_i\_L \}) \quad (3)$$

KWi：關鍵詞

KWi_len：關鍵詞字數

KWi_L：擷取關鍵詞左半邊

KWi_R：擷取關鍵詞右半邊

tf：在語料庫中出現次數的頻率

利用上述公式(1)(2)(3)來計算每一個關鍵詞的分數，但有時會遇到一種情況，該篇文章經過 SPLR 評分後，並無高分 1 分的關鍵詞，因此所有的分數最後會做 Normalize，讓分數一定是從 0 到 1 來分佈，做完以上的工作後，再來就是要利用公式(4)將最能代表情書文章的關鍵詞從關鍵詞串列中取出。

$$KW\_SPLR = for\ i=1\ to\ n=5 \left\{ KW\_S \,\middle\|\, Max\left\{ Keyword\_List \right\} \right| > 0.3 \Big\} \quad Keyword\_List - \left\{ KW\_S \right\} \qquad (4)$$

n：看要決定幾個關鍵詞讓使用者看

經過改良式 SPLR 的運算後，將會給予每一個關鍵詞一個分數，關鍵詞所得到的分數將會介於 0～1 分之間，其中，少於 2 個字的關鍵詞將給予 0 分的評分，因為能讓使用者感受的意境不高，之後將會藉由這個得分來擷取最高分的 5 個關鍵詞。除此之外，經過實驗的結果，關鍵詞的 SPLR 評分需大於 0.3 分，才能取出做為情書文章中主要架構上的關鍵詞。

以下舉一個例子來說明改良式 SPLR 的運作方法，以酸甜苦辣和學生會長做例子來介紹。其中，Keyword 表示 tp，tf 表示出現在語料庫中的頻率次數。

例 1：

tp1 = 酸甜苦辣　　tL = 酸甜苦　　tR = 甜苦辣

tf(tp1) = 文章中共出現 10 次

tf(tL) = 文章中共出現 10 次

tf(tR) = 文章中共出現 10 次

$$\text{則 SPLR} = \frac{10}{10} = 1 \quad \text{（1 即為此關鍵字的得分）}$$

例 2：

tp2 = 學生會長　　tL = 學生會　　tR = 生會長

tf(tp2) = 文章中共出現 20 次

tf(tL) = 文章中共出現 40 次

tf(tR) = 文章中共出現 20 次

$$\text{則 SPLR} = \frac{20}{40} = 0.5 \quad \text{（0.5 即為此關鍵字的得分）}$$

利用以上的方法，可以找出每篇文章中的 KW_SPLR 關鍵詞，將 KW_SPLR 內的關鍵詞串可以形成一串列，並且每次隨機產生 10 串關鍵字串列讓使用者選取。每串中的 KW_SPLR 關鍵詞的主要功能是讓使用者可以從關鍵詞串列中體會其中的意境，讓使用者不用看過多的資訊，就可以決定使用者自己想產生的內容。如下圖 3-4 所示為系統產生的關鍵字串列。

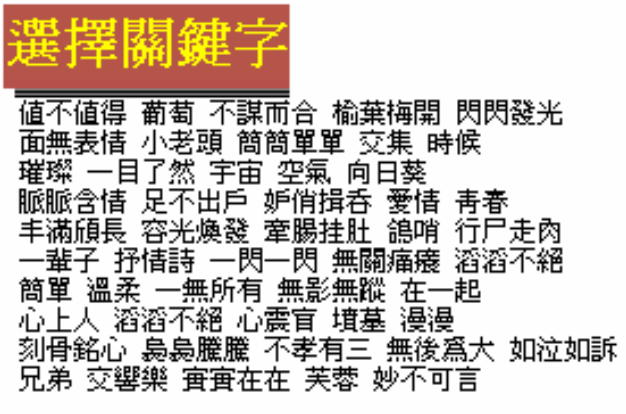


圖 3-4：隨機關鍵詞串列選擇畫面

3.4.2 關鍵字串列生成

此節將介紹，如何將每篇文章中取出的 KW_SPLR 關鍵詞資訊，利用隱含生成的方式，產生整篇情書文章的整體架構。下圖 3-5 為其流程架構。

圖 3-5：關鍵字擴展流程架構

生成的方法，利用上一章節中所取出來的 KW_SPLR 關鍵詞來達成，在 3.3.1 中結構化文章串列時，已將能代表文章的名詞(Na)和動詞(VH)給挑選出來了，再來就利用下面的公式(1)，將 KW_SPLR 關鍵詞與 Keyword List 中其它關鍵詞，一起做生成的動作。

$$Core\_Keyword = \left\{ KW_{i+j} \;\middle|\; \begin{array}{l} i = KW\_SPLR(u)\_Loc \\ |j| \le n \\ 1 \le i+j \le Keyword\_List\_Total \\ ,n \in Z^{+} \end{array} \right\} \quad (1)$$

KWi+j：表示關鍵詞在原文中的位置

Keyword_List_Total：關鍵字串列總數

n：取出 KW_SPLR 關鍵詞前與後各 n 個關鍵詞(例如：n=2)

KW_SPLR(u)_Loc：表示 KW_SPLR 內第 u 元素在原文中的位置

下圖 3-6 表示生成相互連結的情況。

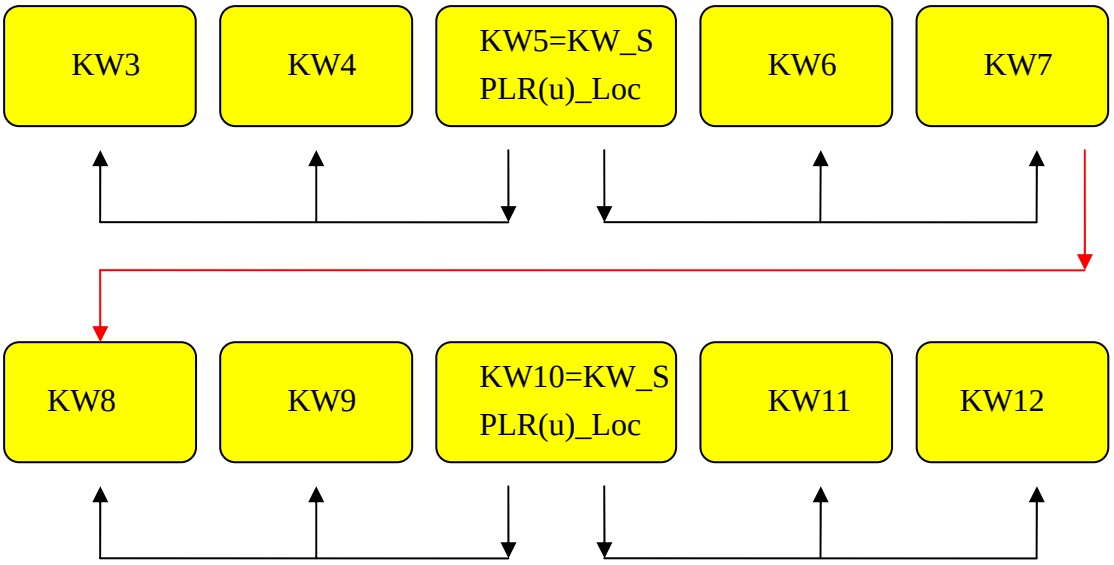


圖 3-6：KW_SPLR 關鍵字生成方式

上圖 3-6 假設 KW_SPLR 其中一個的位置計算出來等於 5，即 KW5。則將 KW5 相鄰位置的前 2 個關鍵詞(KW3，KW4)與後各 2 個關鍵詞(KW6，KW7)取出，形成一個由 5 個關鍵詞所組成的小集合，同樣的 KW10 前後抓出 KW8 KW9 KW11 KW12，4 個關鍵詞所組成的另一個小集合，最後的集合則由 5 組小集合所形成。其中，這樣的產生方式，萬一 KW_SPLR 彼此距離未大於 4 個關鍵詞，則採交集的方式產生，如下圖 3-7 所示。

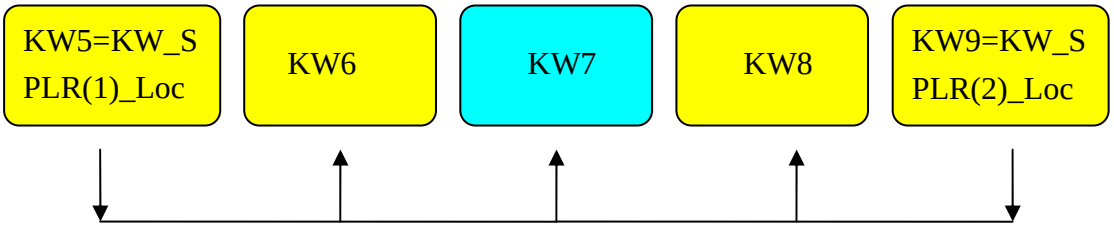


圖 3-7：KW_SPLR 關鍵詞產生重疊

上圖 3-7 藍色區塊，就是二個 KW_SPLR 在生成關鍵詞時，所產生的重疊現象，做法採用交集的方法避免同時抓取到相同的關鍵詞。以上敘述就是此節關於關鍵詞串列生成的方法。

3.4.3 隨機情書產生

將情書文章的主幹都架構好後，再來就是要將候選填充詞填入關鍵詞與關鍵詞中，延伸出一篇情書文章，有如骨幹上的血肉一般，如下圖 3-8 所示，為產生一篇隨機情書的流程架構圖，之後再將詳細介紹其運作方法。

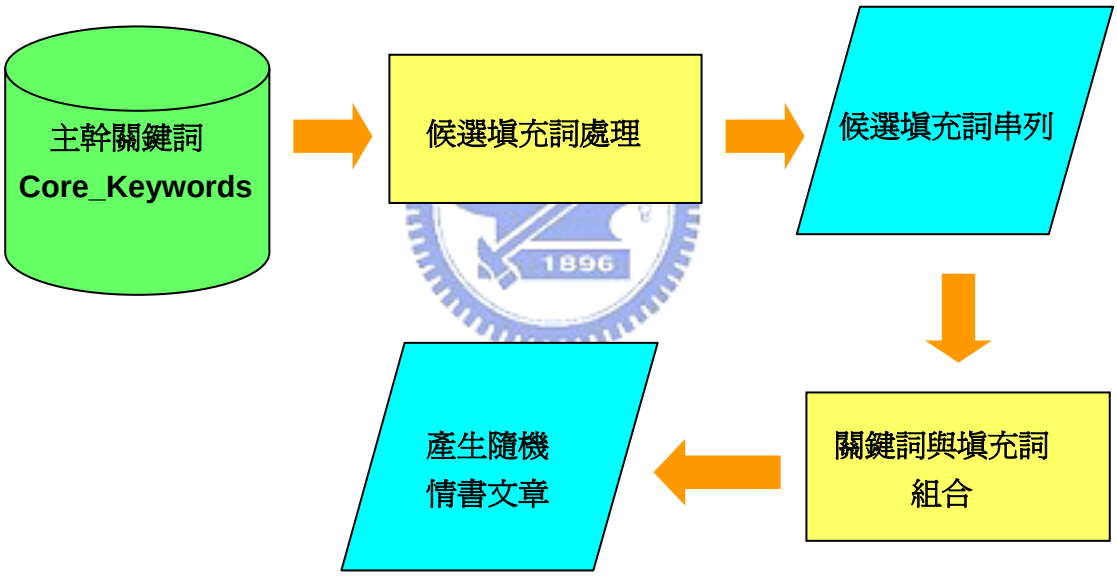


圖 3-8：情書文章隨機產生流程架構

接下來，要先來介紹如何產生可填入在關鍵詞之間的填充詞，在此先定義接在關鍵詞與關鍵詞間的句子稱為填充詞，可允許接在關鍵詞與關鍵詞間的稱為候選填充詞，如果是多數，則稱候選填充詞串列。以下公式(1)為填充詞集合的收集方式，再將所有找到的填充詞收集起來，成為填充詞串列。

$$Candidate(u) = \left\{ C_u = Unit(i,j) \;\middle|\; \begin{Vmatrix} \left| Pre\_KW \cap Word(i,j) \cap Pos\_KW \right| > 0 \\ Word(i,j)\_len \;\middle|\; \leq 30 \end{Vmatrix}, \forall i,j \right\} \quad (1)$$

Unit(i, j)：表示在第 i 篇 j 位置找到填充詞

Pre_kw：指前一個關鍵詞

Pos_kw：指 Pre_word 關鍵詞後 1 個關鍵詞

Word(i, j)：表示可接於 Pre_word 與 Pos_word 間的填充詞

Word(i, j)_len：表示填充詞的最大長度不可超過幾個字

舉例，Pre_KW 為敏感，Pos_KW 為萬眾矚目，則搜尋出來的 Word(i, j)有以下 2 種變化例子。

<1>敏感+出於與生俱來，你正期待我是在一片玫瑰花海中+萬眾矚目

<2>敏感+我知道有一天他會在一個+萬眾矚目

會結合出以上 2 種變化的句子。

利用上述方式，將填充詞一一找出並且收集起來(2)，則成為候選填充詞串列，之後以此基底，則可利用關鍵詞與候選填充詞串列來組成一篇新的情書文章。

$$Candidate(u,n) = \left\{ \begin{matrix} Unit(i,j)\,Unit(i,j+1)\,Unit(i,j+2)\cdots \\ \left| \begin{matrix} Unit(i,j) \\ \in Candidate \end{matrix} \right. \end{matrix} \right\} \quad (2)$$

n：為接在哪一個關鍵詞後的位置

最後，因為利用上述的方式去產生填充詞，所以 Core_keyword 在產生關鍵詞時，當 k 轉變時，有可能會找不到可填充的候選填充詞，則採用下列的公式來解決找不到填充詞的問題。

$$Letter\_Temp = \left\{ \begin{matrix} \text{Pre}\_KW_i \cap Candidate(n) \cap Pos\_KW_{i+1} , \\ \forall i = 1,2,3,4\cdots \end{matrix} \right\}$$

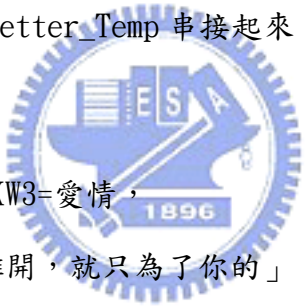
$$Love\_Letter = \quad for \quad i=1 \quad to \quad (p-1)$$

$$Letter\_Temp \qquad (3)$$

Letter_Temp：將關鍵詞與候選填充詞組合起來

p：關鍵詞總數

Love_Letter：將 Letter_Temp 串接起來，成為一篇情書文章



舉例：

KW1=過客，KW2=刻骨銘心，KW3=愛情，

Candidate(1,1)=「中不想離開，就只為了你的」

Candidate(1,2)=「，雖然是最讓你」

Candidate(2,1)=「的」

Candidate(2,2)=「的。沒有」

Candidate(2,3)=「難以忘懷，永生難忘的。」

以此為組合排列的話，共會有 6 種變化，如下：

過客 + 中不想離開，就只為了你的 + 刻骨銘心 + 的 + 愛情

過客 + ，雖然是最讓你 + 刻骨銘心 + 難以忘懷，永生難忘的 + 愛情

過客 + 中不想離開，就只為了你的 + 刻骨銘心 + 的。沒有 + 愛情

過客 + ，雖然是最讓你 + 刻骨銘心 + 的 + 愛情

．．．．．．

以上的例子，可清楚得知情書文章是如何產生的，下面的 2 個例子，即為隨機產生出來的情書文章。如圖 3-9 與圖 3-10。

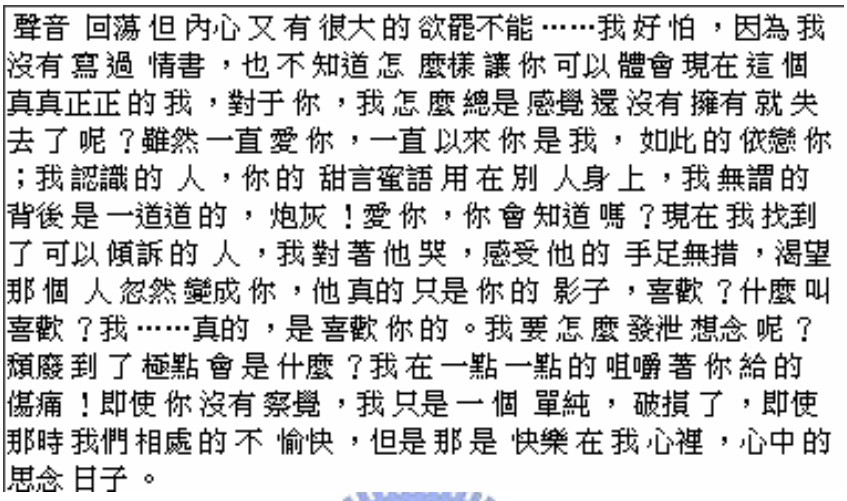

聲音 回蕩 但 內心 又 有 很大 的 欲罷不能……我 好 怕 ，因為 我 沒有 寫過 情書 ，也 不 知道 怎 麼樣 讓 你 可以 體會 現在 這 個 真真正正 的 我 ，對于 你 ，我 怎麼 總是 感覺 還 沒有 擁有 就 失 去 了 呢 ？雖然 一直 愛 你 ，一直 以來 你 是 我 ， 如此 的 依戀 你 ；我 認識 的 人 ，你 的 甜言蜜語 用 在 別 人身 上 ，我 無謂 的 背後 是 一道道 的 ， 炮灰 ！愛 你 ，你 會 知道 嗎 ？現在 我 找到 了 可以 傾訴 的 人 ，我 對 著 他 哭 ，感受 他 的 手足無措 ，渴望 那 個 人 忽然 變成 你 ，他 真的 只是 你 的 影子 ，喜歡 ？什麼 叫 喜歡 ？我 ……真的 ，是 喜歡 你 的 。我 要 怎 麼 發泄 想念 呢 ？ 頹廢 到 了 極點 會 是 什麼 ？我 在 一點 一點 的 咀嚼 著 你 給 的 傷痛 ！即使 你 沒有 察覺 ，我 只是 一 個 單純 ， 破損 了 ，即使 那時 我們 相處 的 不 愉快 ，但是 那 是 快樂 在 我 心裡 ，心中 的 思念 日子 。

圖 3-9：隨機產生情書 1

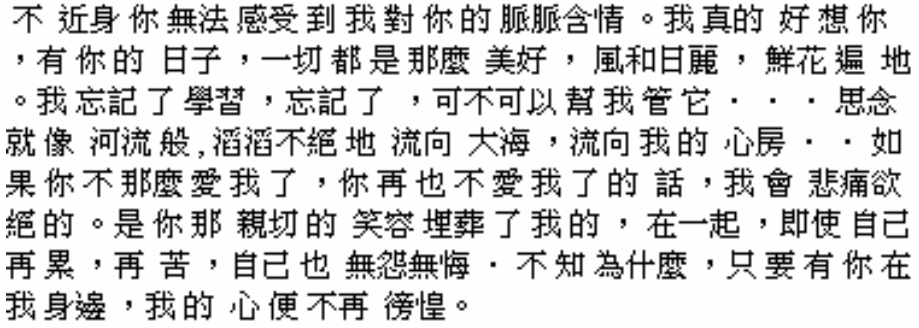

不 近身 你 無法 感受 到 我 對 你 的 脈脈含情 。我 真的 好 想 你 ，有 你 的 日子 ，一切 都 是 那麼 美好 ， 風和日麗 ， 鮮花 遍 地 。我 忘記 了 學習 ，忘記 了 ，可不可以 幫 我 管 它 ． ． ． 思念 就像 河流 般 ,滔滔不絕 地 流向 大海 ，流向 我 的 心房 ． ． 如 果 你 不 那麼 愛 我 了 ，你 再 也 不 愛 我 了 的 話 ，我 會 悲痛欲 絕 的 。是 你 那 親切 的 笑容 埋葬 了 我 的 ， 在一起 ，即使 自己 再 累 ，再 苦 ，自己 也 無怨無悔 ． 不 知 為什麼 ，只 要 有 你 在 我 身邊 ，我 的 心 便 不再 徬惶 。

圖 3-10：隨機產生情書 2

3.5 同義詞架構

同義詞架構系統，為主系統架構中的一個分支子系統，下圖 3-11 表示這個系統運行的流程架構，其中詳細的介紹將在後續一一說明。

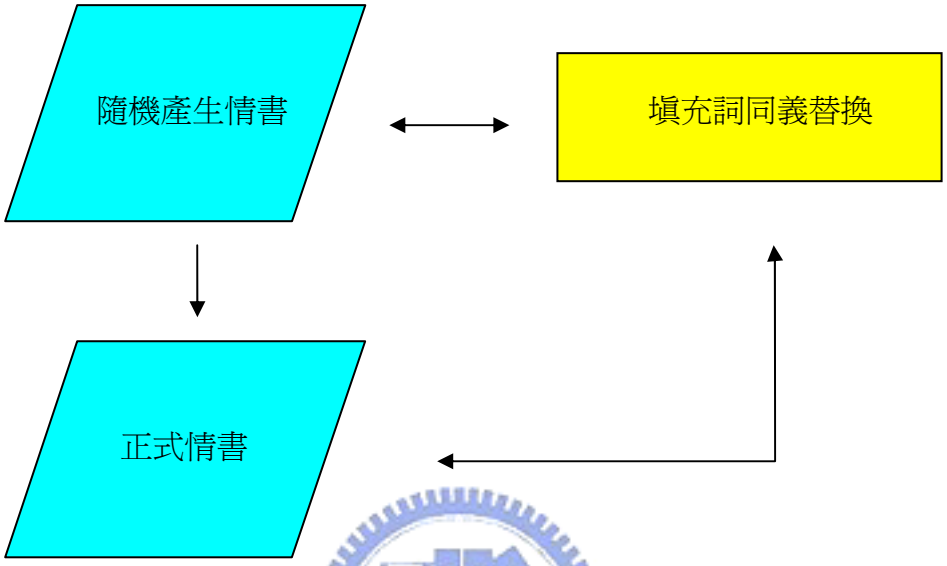


圖 3-11：同義詞替換流程架構

首先，使用者將隨機產生的情書內的填充詞選取出來後，進入填充詞同義替換子系統中，此一子系統將會把使用者能改變的填充同義詞給展示出來，讓使用者可以選取在這一句填充詞中最能代表使用者內心的寫法。其中，填充詞同義替換的方法是將同義詞集合中結構化好的每一筆資料與填充詞內的所有詞比對，當找到相同的詞後，則利用下拉選單給全部一一列出，重覆上述的動作將所有同義的詞都展示出來讓使用者替換。下圖 3-12 為其結果。

圖 3-12：同義替換的情況

3.6 譬喻系統架構

譬喻系統架構，也是主系統的分支子系統之一，其主要的工作是將使用者選出來的填充詞，利用修辭法中的譬喻法，將選出來的填充詞優化，讓句子在情書中感覺更生動。下圖 3-13 表示譬喻系統的流程架構，之後將詳細說明各部份的運作方法。

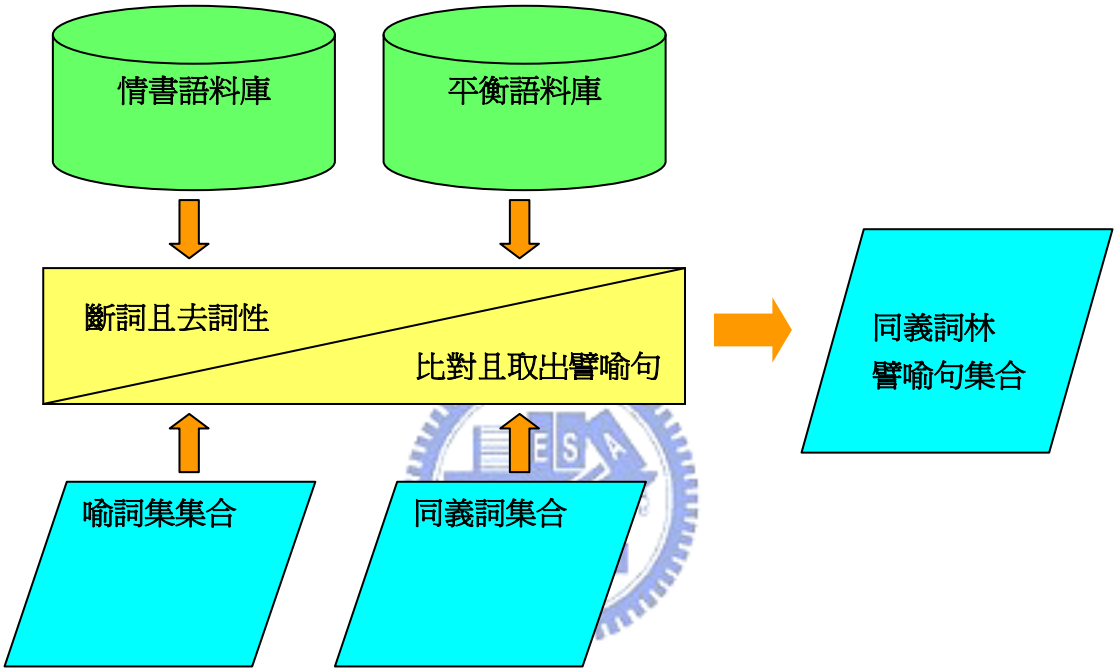


圖 3-13：譬喻流程架構

在 3.3.3 節中已經定義出喻詞集集合，是由 14 個喻詞所組成。接下來，將喻詞集合內的每一個喻詞與同義詞集合內的同義詞，做排列組合，並且利用這些組合在情書語料庫和平衡語料庫中找出符合組合的句子，並將其收集起來，在此定義為同義詞林譬喻句集合。

利用這樣的方式，只要使用者選取填充詞，則譬喻系統會判斷在同義詞林譬喻句集合中是否有其資訊，如果有，則可讓使用者選取要用哪一個譬喻來修辭其填充詞。下面二個例子，為產生後的結果，最上面的字串即為填充詞，下面的幾欄可下拉式選單，即為譬喻修辭的部份。如圖 3-14 和圖 3-15

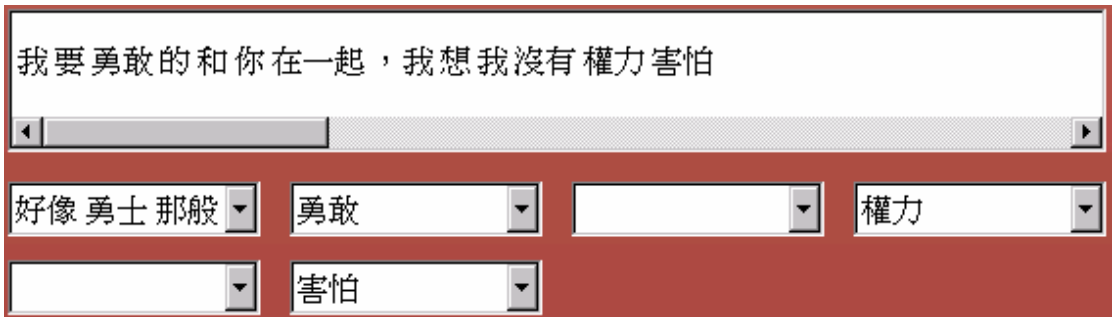


圖 3-14：譬喻系統產生結果 1

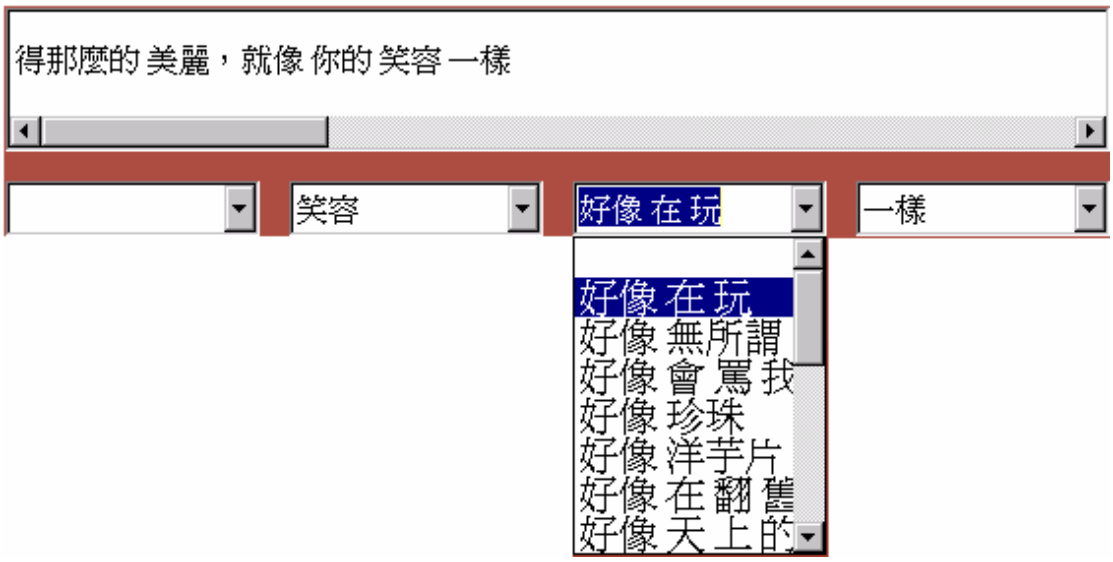


圖 3-15：譬喻系統產生結果 2

上圖 3-14，在同義詞「勇敢」左方的，即為譬喻修辭的結果，「好像勇士那般」。而圖 3-15，在同義詞「一樣」左方的也是譬喻修辭後的結果，「好像珍珠」。在此二例的修飾方式使得句子的感受更生動。

第四章、實驗過程與結果討論

4.1 情書自動產生主系統

情書主系統的部份，將在後 3 節詳細的將系統流程展示出來，並且與中文作文輔助系統做一完整的比較與討論。在 4.1.1 中將描述所使用的語料庫與資料庫，4.1.2 中將詳細的把系統運行的流程說明清楚，4.1.3 中則會與中文作文輔助系統做一討論與比較。下圖為開啟程式後的畫面。

圖 4-1：情書系統首頁

4.1.1 實驗資料

情書產生主系統主要採用的資料為情書語料庫，情書語料庫從網路與各大 BBS 站隨機收集情書共 446 篇做為產生情書文章的基底。

4.1.2 實驗流程

以下，利用圖例來說明主系統流程，流程共分為 3 大部份。第一部份：隨機產生關鍵字串列、修改書寫情書人與接受人;第二部份：隨機產生情書;第三部份：候選填充詞替換。

第一部份、隨機產生關鍵字串列與修改書寫情書、接受人，如下圖中①②所示。

圖 4-2：隨機產生關鍵字串列

第二部份、隨機產生情書，如下圖中③所示。

圖 4-3：隨機產生情書

第三部份、候選填充詞替換，如下圖中④所示。

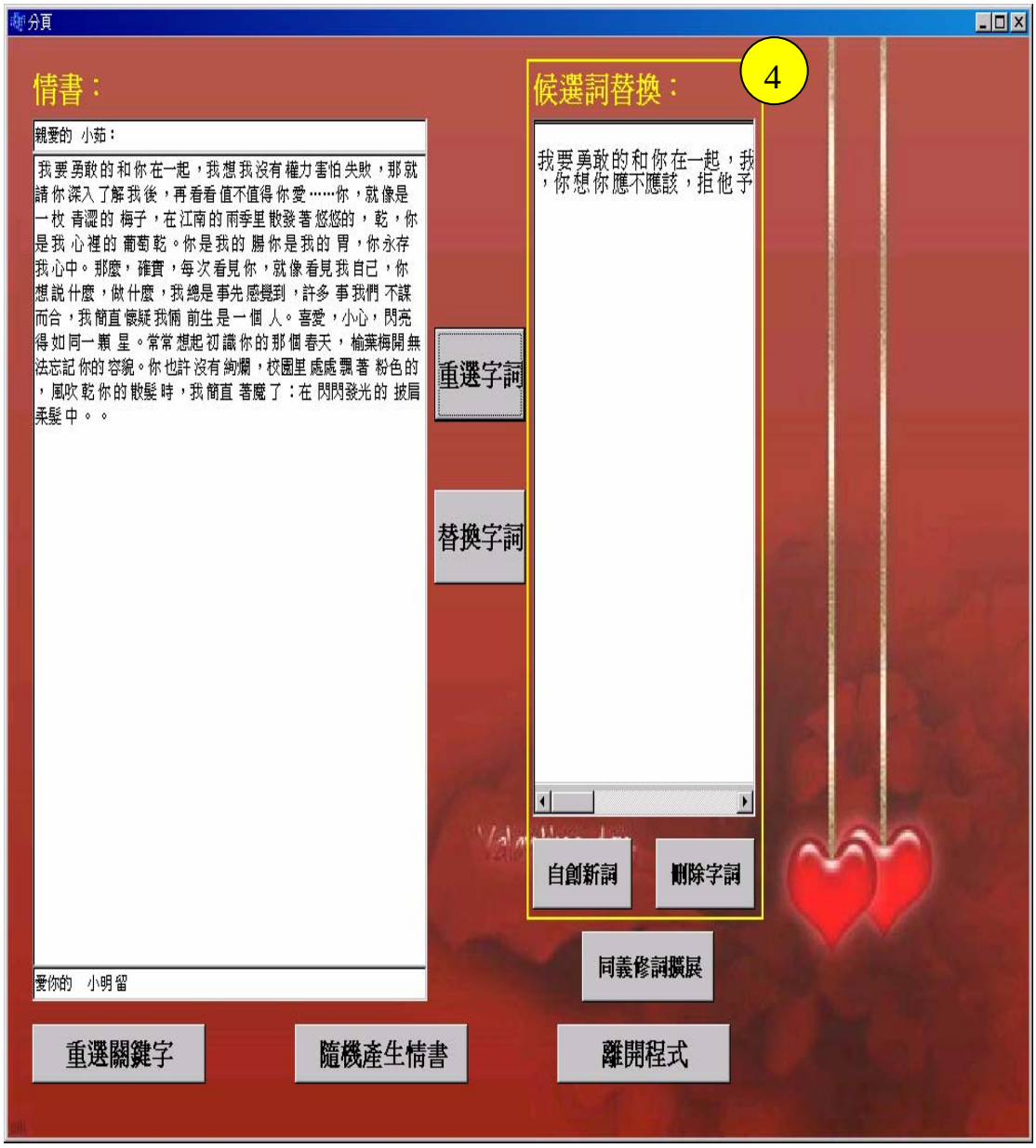


圖 4-4：候選填充詞替換

4.1.3 實驗討論

繼上述的流程，將本系統與中文作文輔助系統做一比較與討論。比較所得的結果如下表所示。

表 4-1：本系統與中文作文輔助系統之比較

	中文作文輔助系統	中文情書自動產生系統
①關鍵字串產生	1. 字串過長，閱讀不易	1. 適當，易閱讀
	2. 關鍵詞過多贅詞，不易了解情境	2. 適當，容易了解情境
②隨機產生文章	1. 因字串限制，文章長度不像作文的長度	1. 情書文章長度適宜
	2. 因關鍵詞過多，使得填充詞修改不易	2. 關鍵詞隱藏，使用者不會覺得文章看起來混亂，而且修改填充詞容易
	3. 文章與填充詞上、下分離，使用不方便，不夠 User Friendly	3. 文章與替換填充詞合併在一起，User Friendly

以下將詳細討論 2 大系統：

①關鍵字串列產生：

1. 關鍵詞的抓取：

●中文作文輔助系統是利用中學生語料庫與平衡語料庫中詞頻的排名去抓取出關鍵詞。利用二種不同語料庫的排名去抓取關鍵字，並無法準確得到關鍵詞。

●本系統採用 SPLR 去給予每一個關鍵字評分，能有效的評比出較能代表此篇文章意境的詞句。

2. 關鍵詞串列的組合：

●中文作文輔助系統是直接利用上述關鍵詞取出後的前 10、20、30 個，

組合成一串關鍵詞串列。

●本系統經由 SPLR 的計算後，再經由 Normalize，取出最高分的 5 個關鍵詞來組成一串關鍵詞串列，這樣組合出來的串列，使用者看起來可以明確感受之後產生出來的文章情況。

②隨機產生文章：

1. 關鍵詞串列擴展：

●中文作文輔助系統直接使用關鍵詞串列的長度為文章骨幹。

●本系統則是經由 3.4.2 的方式，產生出多個隱藏的內含關鍵詞，並以此成為文章的骨幹。這樣的方式，可以讓使用者感覺畫面不混亂，而且使得填充詞的多變性效果更高。

2. 填充詞替換方式：

●中文作文輔助系統將填充詞替換與文章分成上、下 2 區，上區為關鍵詞和填充詞區，下區為文章產生區，使用者需要記住在下區文章內，要換的填充詞前與後的關鍵詞，再到上區找尋該關鍵詞，才能將填充詞來替換。過程較麻煩，且使用者需要花精神去找尋位置。

●本系統是將文章與填充詞替換合併，使用者只需要在文章中點選想替換的填充詞，就可以直接替換。過程簡便，且使用者不需花精神找尋位置。

4.2 填充字串同義替換與修辭系統

此節將介紹二個子系統，同義詞替換系統與修辭系統，在 4.2.1 中將說明使用到的語料庫與資料;在 4.2.2 中將二個子系統運作的流程做一詳細介紹;最後，4.2.3 有二個子系統的討論與比較。

4.2.1 實驗資料

這二個子系統使用到的資料有 2 部份：

1. 情書語料庫，抽取其中的名詞與動詞類，讓同義詞系統去找尋同義詞，並且讓譬喻系統找出與喻詞集合有關聯的句子。
2. 平衡語料庫，同上。

4.2.2 實驗流程

2 個子系統共分為 3 個部份。第一部份：填充詞同義替換;第二部份：填充詞譬喻修飾;第三部份：完成正式情書。其中，第一、二部份已在 3.5 與 3.6 節中介紹過，第三部份則經由上述二部份，經使用者的微調後，完成最後正式的情書文章。下圖為第一、二、三部份運作的流程。

第一與第二部份：

圖 4-5：第一部份與第二部份

⑤分成三部份：

1. 原填充詞：我要勇敢的和你在一起，我想我沒有權力害怕。

2. 同義擴展：大膽、權力、膽怯即是同義替換過的詞。

3. 譬喻修辭：好像勇士那般即是用來修飾大膽的譬喻句。

⑥則是完成後的新填充詞

第三部份、正式情書。

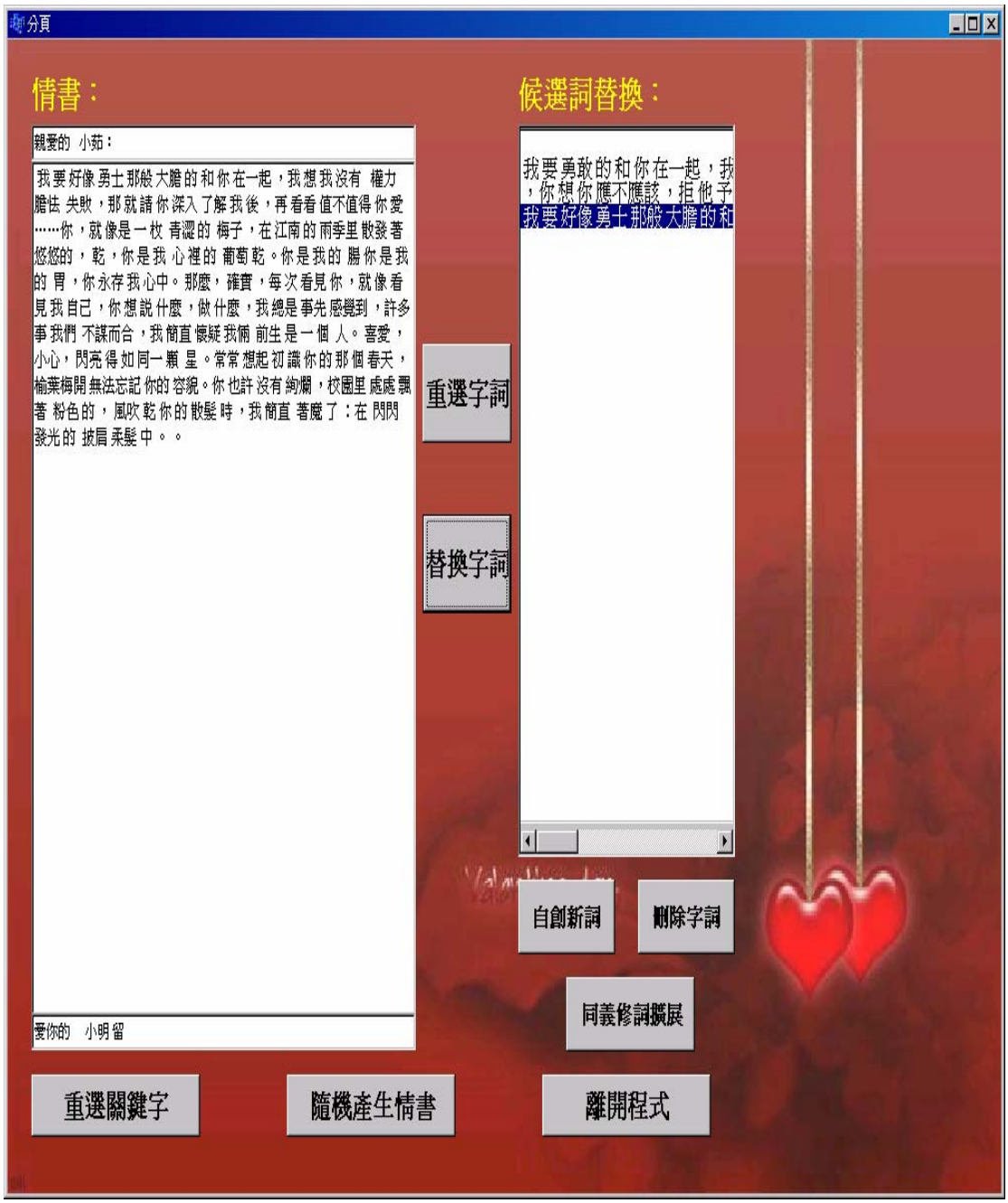


圖 4-6：正式情書

右方則將同義替換過與修飾完成的填充詞加回，左方第一句則可看出已從原文換成使用者希望改動情境了。

4.2.3 實驗討論

接下來，與中文作文輔助系統比較和討論。如下表所示。

表 4-2：子系統比較

	中文作文輔助系統	中文情書自動產生系統
同義詞替換	無此功能	改變填充詞的意境
譬喻修辭	無此功能	使得填充詞更生動

以下，來分別討論二項子系統在附加於情書主系統上的加乘效果。

同義詞系統：

可以讓使用者修正自己所想要當下的意境,讓使用者對於填充詞內的同義詞做替換，以達到符合使用者意境的情況。用以下的 2 個例子做為討論。

1.我要勇敢的和你在一起，我想我沒有權力害怕

勇敢可替換成：大膽 果敢 英勇

權力可替換成：權柄 權利 權益

害怕可替換成：膽怯 恐懼 懼怕 畏懼 畏怯

2.屬於那些雖然曾被背叛過，但依然相信的

背叛可替換成：背離 叛逆 叛離 反叛 倒戈 投降 違背

依然可替換成：仍舊 仍然 如故 依舊

相信可替換成：信賴 信任

以上的這些替換方式，都可以使得句子在閱讀起來，明顯的感覺語氣上的不同，也就能代表了當下使用者的意境。

譬喻修辭系統：

而譬喻修飾，則可以讓使用者對所選的填充詞做譬喻的修飾，可以輔助使用者，讓系統為使用者附加修辭，使得情書文章更加生動。延續上 2 個例子做說明。

1.我要勇敢的和你在一起，我想我沒有權力害怕

在勇敢的前面可加入的修辭有：好像勇士那般

產生：

我要 好像勇士那般 勇敢的和你在一起，我想我沒有權力害怕

2.屬於那些雖然曾被背叛過，但依然相信的

在依然與相信間，可加入的修辭有：有如滴水穿石般去

產生：

屬於那些雖然曾被背叛過，但依然 有如滴水穿石般去 相信的

以上可選擇加或不加的譬喻修辭法，也的確使得情書文章更生動、更活潑。

第五章、結論與展望

5.1 研究總結

本系統經過多次實驗與修正後，已有文章自動產生的具體效果，尤其在抓取文章中代表性的關鍵詞時，經實驗過後，的確較能準確的抓取出關鍵詞，而且利用隱藏生成的方式產生關鍵詞串列架構，讓情書文章能更靈活多變，再附加同義替換與修辭的方式，更能符合中文寫作時的意義，其中，在 1.1 節中所述類似人在寫文章時的情況，本系統模擬了人在寫作時的種種情況，例如：當不知如何寫下一句時，則會從自己看過的書中找尋可往下寫作的句子，當在寫一段句子時，會想用更能符合此句子的用詞，也會想用修辭法來修飾自己寫的句子…等等；該系統已有約 90%的行為是模擬人在寫作的情形，並且經實驗證明有不錯的效果。

5.2 未來工作

有五個方面可以討論。

1. 同義/同意替換：

本系統中，同義只對填充詞中可替換的詞做同義替換，未來，可以直接對填充詞整句來做語義或語意上的判斷，能夠不單只是從資料庫中找尋設定好的同義詞，而是能夠利用句子的語義來替換句子，如此更可加強文章的靈活度。

2. 修辭法：

本系統中，只對填充詞中可替換同義詞採用修辭中的譬喻法裡的明喻，在未來，可加入譬喻法對整句填充詞做修辭，更可以加上其他的修辭方式來修辭，以使得文章具有更強的生動度。

3.語料庫與資料庫的應用：

本系統中，採用的語料庫是平衡語料庫與情書語料庫，在未來，可加入其他種類的語料庫來做更廣大的應用，例如：Blog 日記產生、新聞產生…等等，做各方面不同語料庫的應用。

4.關鍵字讓使用者自行填入：

本系統，採用了 SPLR 的技術從情書語料庫中擷取出重要的關鍵詞讓使用者選用，在未來，可讓關鍵字由使用者輸入，並從語料庫內找出合適可填的詞，同樣的接出完整的情書文章。

5.填充詞以網路為詞庫：

本系統，目前採用 2 個詞庫做為產生文章的主要基礎，在未來，希望利用 Google Search 的方式，將詞庫擴展至整個網路，如此一來，在文章隨機產生時，將具有更靈活的變化

參考文獻

[1] Regina Barzilay & Mirella Lapata, Collective Content Selection for Concept-To-Text Generation, Proceedings of the conference on Human Language Technology and Empirical Methods.(2003)

[2] Kiyotaka Uchimoto & Satoshi Sekine & Hitoshi Isahara, Text Generation from Keywords.(2002)

[3] 余思翰,「中文作文寫作輔助系統」,國立交通大學,碩士論文.(2007)

[4] 張道行,「Automatic Chinese Unknown Word Extraction Using Small-Corpus-Based Method」,國立交通大學,博士論文.(2003)

[5] 中央研究院文國尋寶記同義反義詞遊戲倒影湖
URL:http://www.sinica.edu.tw/wen/Dictionary/sym-asym-demo.html

[6] 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/

[7] 教育部,全球資訊網　URL:http://www.edu.tw/

[8] 東光國小,修辭筆記
URL:http://tkes.tn.edu.tw/tkt056/rhetoric/note.htm

[9] 中研院平衡語料庫 3.1 版