國立交通大學

資訊科學與工程研究所

碩 士 論 文

中文主題詞辨識與其應用

Topic Recognition and Its Application to Chinese Texts

研 究 生：潘善均

指導教授：梁婷　　教授

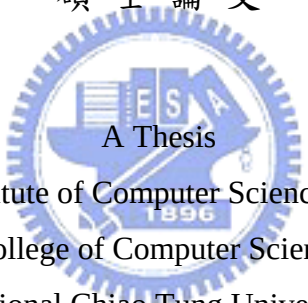
中 華 民 國　九 十 六　年 六 月

中文主題詞辨識與其應用
Topic Recognition and Its Application to Chinese Texts

研 究 生：潘善均　　Student：Shan-Chun Pan

指導教授：梁　婷　　Advisor：Tyne Liang

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 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年六月

國立交通大學

研究所碩士班

論文口試委員會審定書

本校　資訊科學與工程　研究所　潘善均　君

所提論文:

中文主題詞辨識與應用

合於碩士資格水準、業經本委員會評審認可。

口試委員：

指導教授：

所　　長：

中 華 民 國 九十六 年 六 月 二十二 日

致謝

所有的電影片尾曲中都會有工作人員的名字，就如同每本論文的致謝詞一般，細數著每位對論文作者有恩的人。撰寫此篇致謝詞的同時，我也開始回溯這兩年來在交大資工所的回憶，承蒙這兩年來梁婷教授的指導（鞭策？），與建功高中國文教師提供的作文，使我得以完成這本碩士論文。已畢業的鄭守益、蘇傳堯學長，沒有你們的研究生處世之道，我的研究生活一定會相當慘澹吧。黃立泓學長，我們半年後再會，很期待與你再度共事。遠從南非來的曾盛興同學，如果沒有你在壓力大時陪我吃飯，也許我根本無法完成碩士論文。劉正義學長，沒有你那變態般的思想與我一起共演相聲，我的笑容將不會如現在般燦爛。吳典松、朱俊榮、楊哲青學長，感謝你們的經驗之談，朱學長要小心不要鬧太多緋聞喔。龔自良學長，謝謝你在口試前的鼓勵。陳淳齡學姐，感謝你在我每次串門子時不會嫌我太閒而把我趕走。黃慧庭與林彥廷兩位學弟妹們，加油！明年就換你們被鞭策了（笑）。以及圖書館的蔡淑琴小姐，與你一起在圖書館工作是非常快樂的。

已在他界的爺爺奶奶，沒有您，我現在絕對無法高興地去領畢業證書。摯愛的外公外婆，回到家時您噓寒問暖的親情總是能爲我帶來溫暖。爸爸媽媽，實在受您太多恩惠了，此生即使盡心恐也不足爲報。唯一的手足妹妹，感謝你在我耍笨噴太多殺蟲劑差點連自己也殺掉時願意收留我一晚。兩位表弟，能和你們一起玩樂、將來能一起在玉蘭花樹下聊天，是我覺得非常榮幸的事。親親女友，沒有你背後的支持，我甚至無法對未來寄託希望。

師傅鄭世梧，沒有你的教導（？）我將無法瞭解學習日文（？）是很快樂的一件事。三位碩一的室友，海大劉冠驛、阿貓林友涵，我會記得我們一起建立殖民地與抵抗外侮（？）的日子，資優生翁岳暄，我們的風水系統現在也還能運作喔。以及歐他庫林政佐，我會記得與你一起學習開車的日子。小花余思翰，我將懷念與你一起吃飯的日子。Vo 學長，謝謝你的書。Ro 歐陽同學，請小心愛護你的女僕（笑）。

感謝你們，讓我的碩士生活能如此充實快樂。我的碩士生活將到此告一段落了，下一個篇章，不知您是否依舊樂意幫我寫序呢？

中文主題詞辨識與其應用

研究生：潘善均　　指導教授：梁婷

國立交通大學資訊科學與工程研究所

摘要

主題詞辨識是文本理解中一項不可缺乏的工作，它可以釐清文本的核心敘述，進而應用在文章的主題偵測與作文的評分上。本論文首先以重心理論為基礎的方式取得小句重心，再以小句重心作為候選詞，依照主題的各個特徵辨識長句的主題詞，此法並不需任何的訓練語料。最後我們將長句主題詞運用至學生作文的離題偵測上，將小句重心運用至連貫性評量上。我們使用11篇平均字數為1500字的報紙社論文章進行主題詞辨識的驗證，針對包含主題各種特徵的實驗模組加以測試，社論文章的主題詞辨識可達86.84%的正確率，召回率為68.51%。我們另外蒐集95篇400字的學生作文進行主題詞辨識、離題偵測、以及連貫性評量的實驗，學生作文的主題詞辨識可達80.86%的正確率，召回率為71.36%。在離題偵測上，離題文章判別的正確率可達到63.36%，召回率為77.77%。本論文嘗試以長句主題詞來作離題偵測，雖可解決以文章全部詞彙來偵測離題的困難，但尚存有無法解決的問題，例如系統無法辨別學生認知概念上的離題，或者引用新穎的例證而造成系統誤判為離題。

關鍵詞：中文主題辨識、主題特徵、離題偵測、連貫性評量

Topic Recognition and Its Application to Chinese Texts

Student : Shan-Chun Pan Advisor : Tyne Liang

Institute of Computer Science and Engineering

ABSTRACT

Topic recognition is an essential part of document understanding and can help people to quickly understand the core description of the document. It can be applied in topic detection and essay scoring. In this paper, we developed an algorithm to extract the topic from a Chinese sentence. First, we used Centering Theory-based algorithm to center each clauses. Second, we took those centers as candidates and extracted their features to generate a topic in a Chinese sentence. Then, we used those sentence topics to detect off-topic essays, and evaluated essay coherence by clause centers. We collected 11 news editorial articles, each of which contains around 1500 words, as our topic recognition corpus. We also collected another 95 400-words essays written by students to generate sentence topics, detected off-topic essays, and evaluated essay coherence. In our experiment, the precision and recall of topic recognition in editorial articles achieve 86.84% and 68.51%. In students' essays, the precision and recall of topic recognition are 80.86% and 71.36%. In off-topic detection experiment, we can achieve 63.36% precision and 77.77% recall. Our method overcame some problems in using bag-of-words to detect off-topic essays, but still remained some

difficulties that can not be solved. We can not detect the misunderstanding of students' thought, and we also wrongly detected novel ideas given in students' essay as an off-topic sentence.

Keyword: Chinese topic recognition, topic feature analysis, off-topic detection, coherence evaluation.

目 錄

表 目 錄

圖 目 錄

第一章 緒論

1.1 研究動機

文本理解是目前自然語言處理研究中的一個重要的議題，國外重要學術會議如 HLT-NAACL 便有專門爲文本理解開設的研討會—DUC (Document Understanding Conferences)。 而主題辨識則是文本理解中相當重要的一環，不但能節省龐大人力的需求與快速地幫助我們理解文件的核心所在，也在後續對文件建立索引詞以方便搜尋或者分類、對文件的評分、或是「以文找文」的方式瞭解主題的發展脈絡。因此主題辨識除了在文本理解中佔有一席之地外，也對網頁搜尋引擎如Yahoo, Google, MSN等、或是自動作文評分系統如E-Rater[Burstein et al., '01]、或是對新聞事件的追蹤與發展、或是文件自動分類等等，也都是一個重要的議題。

然而主題辨識向來不是一個容易解決的問題，不論是候選詞的篩選、產生與權重的計算，或者是候選詞的特徵抽取都面臨許多的困難。而今，大部分的研究者以文件分類爲目的，故以多篇文章或單一文章爲單位抽取主題，但無法偵測較細部段落上的主題描述。而以作文評分系統爲目的者，則傾向以文章的全部詞彙來偵測文章的合題性，此法缺點是詞彙過多，故稀釋了真正主題描述的詞彙。

本篇論文提出以長句爲單位的主題詞辨識法，並以二階層式的方法辨識長句的主題詞，應用了許多自然語言處理技術如詞性標記與知識本體等自然語言處理工具來分析詞性與詞彙語意概念。我們將文章切分爲長句與小句之後，以長句爲單位抽取各種主題詞的特徵並以這些特徵來幫助辨識與分析主題。

在後續的應用上，我們將主題詞應用在學生作文的離題偵測、結構分析與連貫性評量，期能爲文件主題分析提供一個新的研究方法。

1.2 研究方法簡介

以長句爲主的主題詞偵測法，優點是主題詞的單位適中，可解決以全部詞彙偵測主題詞時造成主題詞稀釋的問題，且也可解決從整篇文章抽取主題而無法偵測細部段落主題的問題，較爲適合作文評分系統上的應用。但缺點便在於應用時，若主題辨識失誤，則整段長句便容易被誤判。雖文章並是不只有一長句，但其容錯率將低於以全部詞彙來偵測主題詞的方法。

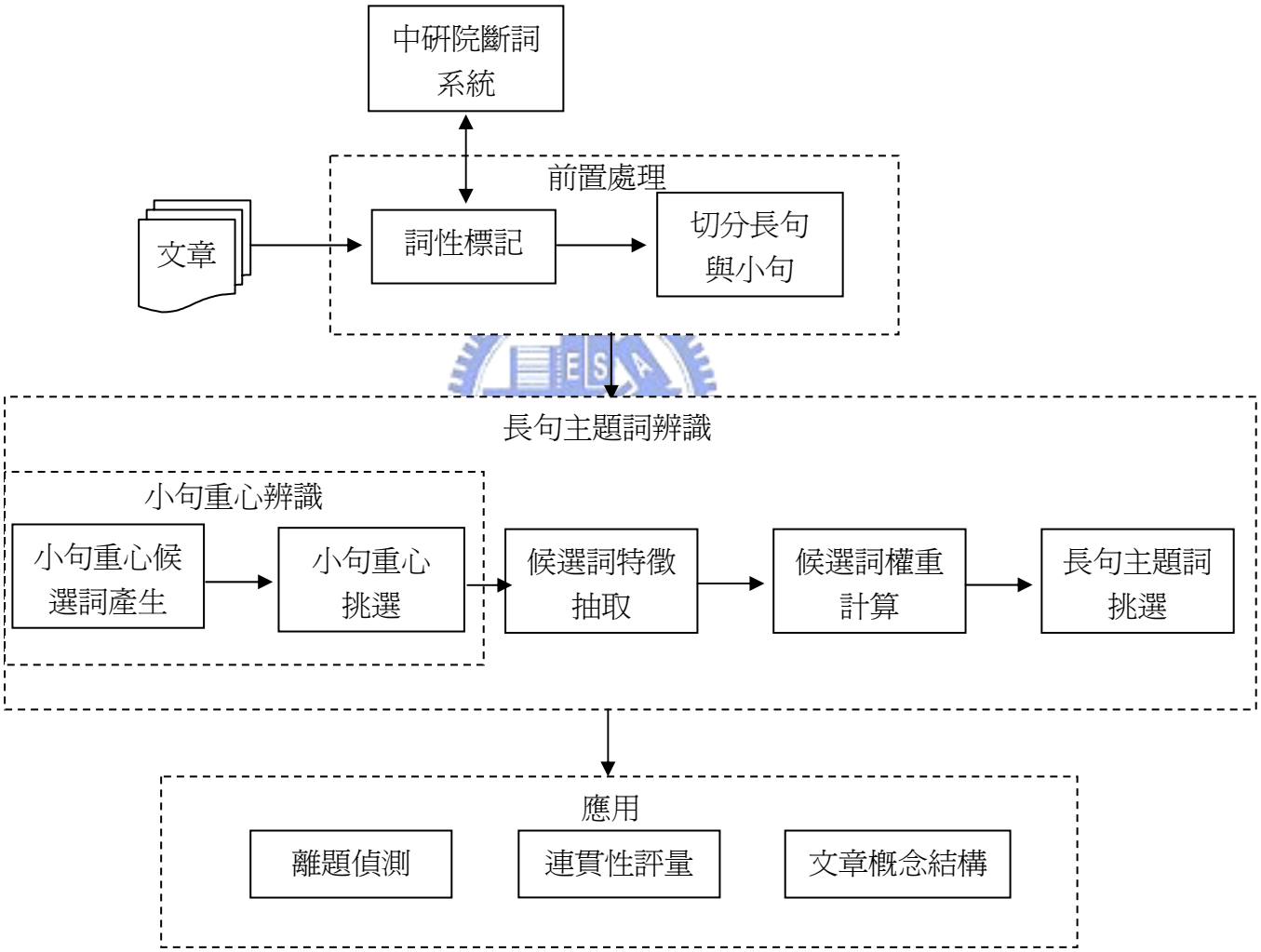


圖 1-1：主題詞辨識與應用流程圖

我們使用標點符號「！？。」將文章分割爲長句，以及標點符號「，；。？！」將文章分割爲小句，如此一篇文章可切分爲數個長句，長句內亦有數個小句。我們將主題視爲是長句的描述中心，目的是找出每個長句的主題，以供後續作文的

離題評量，其流程圖如圖 1-1。

在圖 1-1 中，前置處理時我們將文章先以[中研院 中文斷詞系統]作詞性標記，並將文章切分爲長句與小句。之後我們以二階層式方法辨識長句的主題詞，第一階段先辨識小句的重心，我們先抽取小句重心的候選詞，決定小句的重心後，第二階段以這些小句重心當成長句主題詞的候選詞，抽取這些候選詞的各項特徵，並且依照這些特徵計算其權重，最後挑選出適當的長句主題詞。

在後續應用中，我們將長句主題詞應用在離題偵測與文章概念結構上，將小句重心應用在連貫性偵測上。

第二章 文獻探討

在本章節中，我們將探討國內外關於主題辨識的研究，而我們的後續應用是將主題辨識應用在學生作文的離題偵測(Off-topic Detection)與連貫性評量(Coherence Evaluation)上，因此在本章當中也將介紹關於離題偵測與連貫性評量方面自動作文評分系統的相關研究。

2.1 主題定義

我們認爲中文所指的主題(Topic)，即是一個語篇片段的重心。這個片段可以是整篇文章、或是文章中的長句、小句，端賴研究者依照需求而定。例如以文章爲單位者，Makkonen 等[‘04]將主題視爲是事件(Event)，可由人、事物、時、地等元素構成主題。以語段爲單位定義主題者，曹逢甫[‘95]認爲中文是一種語段取向的語言，主題必須要依據語段才能決定。而以語句爲單位者，Li and Thompson [‘81]認爲中文是一個「主題明顯」(Topic- Prominence)的語言，即語句都必須含有一個不可或缺的元素——「主題」， Yeh and Chen[‘04]認爲主題是「主詞的具現化」(Reification of subject in the real world)，且一個語句的主題總是出現在語句前面。如下例：

例 1：電子股受美國高科技股重挫影響，持續下跌。

徐爲群等[‘04]則認爲中文的「主題」可概括定義爲：「說話者所關注的顯著語意實體 (Salient Semantic Entity, SSE)」，此與重心理論(Centering Theory)[Grosz et al., ‘95]的重心(Center)定義是相似的。重心理論當中重心的定義，是一個詞彙於語句(Utterance)當中，構成這個語句與其他語句之間的連結，則此詞彙稱之爲這個語句的重心，通常重心也被稱爲是主題(Topic)或者是注意中心(Focus) [Miltsakaki and Kukichy, ‘00]。如下例：

例 2：世界各國競相在中國投資設廠，中國儼然成爲全球重要生產國之一。

上述對主題的定義都傾向於選取名詞爲主題，因此我們在選取主題詞時也以名詞爲主，名物化動詞爲輔來挑選主題。

2.2 主題辨識之相關研究

選取主題的基本方法可分爲下列幾種，第一種是最簡單的方法，直接擷取詞頻最高者。第二種則是將第一種方法加以權重計算，例如以 TFIDF 計算權重並擷取權重最高者。第三種是利用知識本體（如 WordNet 的詞彙概念辭典）擷取頻率最高的語義概念當成主題。

我們將目前的研究依擷取主題的單位分成多篇文章、單一文章、單一句子三類來介紹，目前大部分的研究都是以文章爲單位抽取重要的詞彙，做爲文章的主題，較少以句子作爲抽取主題的單位。

2.2.1 多篇文章之主題辨識研究

以多篇文章爲單位來擷取主題的方法，有使用 TFPDF(Term Frequency-Proportional Document Frequency)法計算詞彙權重從而抽取主題[Khoo and Ishizuka, ‘02]，另外也有使用詞彙頻率並加上同義詞與使用 LSI (Latent Semantic Indexing)以構成詞彙群組，計算詞彙的重要性並抽取詞彙作爲主題[Lin ’04]。以及使用概念地圖(Concept Map)的方式，將詞彙先對應至概念地圖，之後計算地圖之間的相似度構成地圖類別，從地圖類別產生初始的第一代主題後，以搜尋引擎搜尋主題，並將傳回的網頁以 SVD(Singular Value Decomposition)計算產生第二代的主題，如此反復計算直至主題不再變動或者代數已超過某一門檻爲止[Leake et al., ‘03]。這些方法的目的是在於事前並不知道文章群聚在一起的原因，而想找出這些文章共同的主題，而當新文章出現時，也可利用找出的主題判斷新文章是否屬於這個文章聚落。

2.2.2 單一文章之主題辨識研究

在單一文章的主題偵測上，可利用 WordNet 將詞彙概念化及權重評量，取每個概念出現頻率，乘以其權重後取其最高者當成主題[Tiun et al., '01]。實驗則從 Yahoo 的商業與經濟目錄中擷取 109 個分類與 202 篇網頁，可將 202 篇 Yahoo 網頁以此法擷取主題後分類至其適當的目錄中，正確率可達 69.8%。

除了利用 WordNet 將詞彙概念化外，也可再加以權重計算，如以 TFIDF 與詞彙所在段落的位置（如標題、第一段、第二段等等），以及詞彙前有無關鍵詞（如 in summary, in conclusion 等關鍵詞）計算詞彙的權重，以此抽取文章的主題[Lin, '98]。韓客松等['00]則在計算詞彙的權重上，考慮了詞彙的位置、長度、詞性標記三個層面，依照詞彙不同的位置設定不同權重，以及越長的詞彙代表資訊含量越多，並且將名詞權重設定最高，名物化動詞次之，其餘虛詞等等權重最低，最後抽取權重最高的數個詞彙代表文本的主題。其以 58 篇文章作爲測試，每篇分別以取 3~9 主題詞計算其主題正確率與召回率，在每篇取 9 個主題詞的情形之下可以達到 69%的正確率與 67%的召回率。

而以機率模型取文章主題者，可將文章與主題當成節點(Node)，設文章內每個詞彙的出現均爲獨立事件，計算詞彙可能的主題機率，最後加總文章內的詞彙對應主題之機率，作爲文章與主題節點相連的機率，從而抽取文章之主題[Chang et al., '02]。

有別於以詞彙權重的挑選方式，文章主題也可用含有「人、事物、時、地」的「事件」來呈現[Makkonen et al. '04]，並將此應用於新聞的主題辨識上。此法在使用 4000 篇新聞作爲訓練語料、4000 篇新聞作爲測試語料時，偵測主題的正確率爲 86.36%，召回率爲 57.58%。

2.2.3 單一句子之主題辨識研究

在句子的主題辨識上，Grosz 等人[‘95]所提出的重心理論(Centering Theory)是一個相當著名的辨識模型，它依照兩兩相鄰句子間的重心候選詞重複與否來判斷句子的重心。而 Yeh and Chen [‘04]將零指代(Zero Anaphora)消解應用在主題辨識上，利用淺層剖析器(Shallow Parser)將句子的主語、賓語、與其他詞彙分辨出來，並以零指代消解之後的先行詞視爲主題，其主題之決定順序爲主題>主語>賓語>其他詞彙。

另外尚有以句型分類方式辨別句子主題，如徐爲群等[‘04]將口語（聊天室的對話）句型以 XST(Extended Sentence Type) 分成陳述句、祈使句、疑問句、感嘆句、功能句等等類型，並進一步將疑問句分成特殊疑問、一般疑問、選擇疑問、附加疑問、反意疑問、零疑問句型，以 HMM(Hidden-Markov Model)與 NBC (Naïve Bayes Classifier)將口語對話自動進行句型分類，或是以人工編寫句型分類規則，產生主題候選詞後以句型分類來辨識主題。使用 58 組對話紀錄，40 組作爲訓練，10 組作爲測試，8 組作爲開發，其正確率介於 65%~75%之間。

目前的研究以句子爲單位抽取主題者較少，而在偵測文章離題時，以文章爲抽取主題的單位似嫌過大，無法細部偵測到文章段落的離題。

2.3 離題偵測之相關研究

離題偵測常是自動評分系統的一部份，而最常見的方法便是計算文章全部詞彙出現在高低分文章的比例來偵測離題。例如 E-Rater[Burstein et al., ‘01]系統的用字分析(word usage)是由訓練語料擷取詞頻高的詞彙，評量與高低分文章之間的用字相似度。E-Rater 每年批改 750,000 份 GMAT 作文，與人類批改者之相似率(Agreement)爲 97%（總分六級分，與人類批改者相差一級分以內）。

而專門用於評量簡答題的 C-Rater[Leacock and Chodorow, ‘03]則藉由分析回

答的邏輯性(例如正向語氣或者否定語氣)，利用知識本體抽取回答的詞彙概念，計算有多少概念符合正確答案，以此判斷回答是正確或是已經偏離題目。C-Rater 於五題簡答題共約 100,000 份回答中與人類批改相似率爲 80%。

Automark[Mitchell et al., '01]則是事先制訂正確與錯誤答案的樣版(scheme template)，利用語句分析器(sentence analyzer)抽取句子的主要詞彙與結構，之後找尋是否有符合模版的結構及概念，以此評量文章是否符合正確答案。Automark 於四題簡答題共 480 份回答中與人類批改者的相關係數(Correlation)爲 0.93~0.96。

在中文作文自動評分系統的離題偵測上，目前也都是以文章的全部詞彙來評量離題，王信智['00]利用向量空間表示法(Vector Space Model)計算專家所提供之範本與作文的相似度，其實驗在小學生科學寫作的 36 篇文章中，與人類批改者的相似率(Agreement)約 80%，其特點是未使用任何訓練語料，而是由專家所提供之範本來評量作文。

而詞彙的概念化也用在中文作文合題偵測上，如蔡沛言['05]、林信宏['06]、粘志鵬['06]等以知網(HowNet) 爲工具，計算文章使用詞彙的「義原」次數，觀察義原屬於高分或是低分文章的特徵，再以不同的分類模組來評量文章，得到不錯的效果。

2.4 連貫性評量之相關研究

連貫性問題雖不比離題問題嚴重，但文章的連貫性是作者邏輯條理是否正確的重要指標之一。有 Miltsakaki and Kukichy['00] ['04]以人工方式標記各語句重心，將重心串成一串主題鏈(Topic Chain)，並以重心理論來評價文章的連貫性，最後將此特徵與 E-rater 結合，可爲 E-rater 增加 3%的正確率。但此法尚未提供一重心辨識法，故需仰賴人工標示重心。

另有 Higgins 等人[‘04]利用人工方式先標明句子的角色（如 Introduction、Main、Conclusion 等），之後標注每一句對題目(Prompt)、主題句(Thesis)、段落(Segment)的關連、以及有無錯別字等，實驗以 890 篇人工標示的作文爲訓練語料，以 SVM (Support Vector Machine)判讀 90 篇作文內的句子對題目、主題句、段落等的關連性，可得 74%的正確率。此法利用 SVM 判讀每句與題目、主題句等關連性之高低，但需大量訓練語料。

我們的方法則是先使用我們小句重心的辨識策略，再採用 Miltsakaki 等人的方式判讀連貫性，一方面不需有訓練語料，另一方面語句的重心也是自動產生，並不需仰賴人工標示。

第三章　語句主題詞萃取

根據[曹逢甫, '95]，中文是一種語段取向的語言，主題必須要依據語段才能決定。我們將長句視爲是文章的語段，採用二階層方式挑選長句的主題詞，第一階段依據重心理論來選取各小句的重心，第二階段則由這些小句重心來決定一個長句的主題，這種由小句重心選取主題的方法可以避免直接面對整個長句而產生太多的主題候選詞。

本章的 3.1 節爲小句重心候選詞的產生，3.2 節由產生的候選詞之中，挑選出小句的重心；3.3 節爲小句重心的實驗結果，3.4 節則描述藉由小句重心來選取長句主題詞的各個挑選特徵，3.5 節爲實驗模組的建立，3.6 節爲實驗結果與分析。

3.1 小句重心候選詞產生

以下爲我們重心候選詞的產生步驟：

步驟 1. 文章經 CKIP 標記，並依標點符號「，。！？；」切分成小句

步驟 2. 將標記爲普通名詞(Na)、專有名詞(Nb)、地方詞(Nc) 的詞彙，以及第一、第二人稱代名詞作爲候選詞

步驟 3. 辨識名物化動詞

步驟 4. 辨識零指代與第三人稱代名詞

圖 3-1：重心候選詞產生步驟

文章需先經過斷詞系統標記各個詞彙的詞性，以便後續的候選詞挑選與零指代的辨識。因此在步驟 1 中，我們利用 CKIP [中研院 中文斷詞系統]標記文章，標記如表 3-1，並依標點符號將文章切分成數個小句單位。

表 3-1：CKIP 部分動詞名詞類別標記

Na	/*普通名詞*/
Nb	/*專有名稱*/
Nc	/*地方詞*/
VA	/*動作不及物動詞*/
VAC	/*動作使動動詞*/
VB	/*動作類及物動詞*/
VC	/*動作及物動詞*/
VCL	/*動作接地方賓語動詞*/
VD	/*雙賓動詞*/
VE	/*動作句賓動詞*/
VF	/*動作謂賓動詞*/
VG	/*分類動詞*/
VH	/*狀態不及物動詞*/
VHC	/*狀態使動動詞/
VI	/*狀態類及物動詞*/
VJ	/*狀態及物動詞*/
VK	/*狀態句賓動詞*/
VL	/*狀態謂賓動詞*/
V_2	/*有*/

我們從報紙的社論文章中蒐集 11 篇社論作爲實驗語料，由實驗語料中觀察連續 30 個小句，發現有 27 小句的重心都是名詞，剩餘 3 小句則沒有重心。因此在步驟 2 中，我們以名詞爲主來產生重心候選詞。另外我們也針對名詞片語只取片語中最後出現的名詞作爲重心候選詞，以及第一、第二人稱代名詞也列爲重心候選詞之一。

雖然這 30 個小句重心並未是名物化動詞，但我們亦觀察到有 4 個名物化動詞存在於這 30 個小句內，而中文語句中，名物化的動詞亦有可能是小句重心的情形，因此步驟 3 中我們參考[馬偉雲, '06]將部分名物化的動詞納入重心候選詞，如下列情形：

1. 將「是」前面的動詞列爲候選詞。

例 2a：打架(VA)　是(SHI)　不(D)　好(VH)　的(DE)

2.　將「的」後面的動詞列爲候選詞。

例 2b：學生(Na)　的(DE)　不(D)　合作(VH)

3.　兩個連續動詞，且第一個動詞爲 VC (動作及物動詞)時，將第二個動詞列爲候選詞。若遇三個以上連續動詞，則以 Bi-gram 方式視爲多次的兩連續動詞。

例 2c：進行(VC)　調查(VE)

以此法在 4 個名物化動詞中共可抽取出 3 個名物化動詞，未抽取到的 1 個乃是由於動詞前面「的」被忽略，造成名物化動詞直接位於名詞後方，而「名詞＋動詞」是一般句型很常見的組合，因此我們無法將此列爲名物化動詞的規則之一。

另外，針對含有第三人稱代名詞與零指代的小句，我們進行代名詞與常見的零指代辨識處理。在步驟 4 中，我們簡單地假設下列動詞：「VA, VAC, VB, VC, VCL, VD, VE, VF, VG, VH, VI, VJ, VK, V_2」前面若無名詞，且亦非步驟 3 的動詞名物化情形，即視爲零指代，但其中我們排除「VH+DE+N」的形容詞用法與動詞「沒有(VJ)」。

我們使用實驗語料的前 200 個小句來評量零指代的辨識情形，200 小句中有 53 個小句被系統辨識爲零指代情形，7 個小句是辨識錯誤。但在後續的 3.3 節當中，我們檢查到這 7 個辨識錯誤，有 6 個均在長句的第一小句，因此將不影響小句重心的判斷。

3.2 小句重心選取

小句重心選取是將重心理論加以改進以適用於中文語句重心的選取。重心理論的原則如下，設目前的第i小句重心爲C_i，前一小句之重心爲C_{i-1}，再前一小句

重心爲C_{i-2}。則依據C_i, C_{i-1}, C_{i-2}之間的關係，共有下列四種情形：

	$C_{i-1}=C_{i-2}$	$C_{i-1}\neq C_{i-2}$
$C_i=C_{i-1}$	延續 Continue	平順遞移 Smooth-Shift
$C_i\neq C_{i-1}$	保留 Retain	粗糙遞移 Rough-Shift

重心模型(Centering Model) [Grosz et al., '95]

第一種情形是$C_i=C_{i-1}$且$C_{i-1}=C_{i-2}$時，表示C_i延續前面兩小句的重心。

第二種情形是$C_{i-1}=C_{i-2}$但$C_i\neq C_{i-1}$時，表示小句重心準備遞移，但尚不知是「平順遞移」或是「粗糙遞移」，須待下小句決定，故「保留」。

第三種情形是$C_{i-1}\neq C_{i-2}$但$C_i=C_{i-1}$時，表示小句重心由C_{i-2}「平順遞移」至$C_{i-1}=C_i$。

第四種情形是$C_{i-1}\neq C_{i-2}$且$C_i\neq C_{i-1}$時，表示小句重心「粗糙遞移」，由C_{i-2}遞移至C_{i-1}後又馬上遞移至C_i。

由於「粗糙遞移」的情形較少，而在中文小句中，若遇$C_{i-1}\neq C_{i-2}$，而C_{i-1}想帶出C_i，兩者並沒有重心候選詞重複或者有零指代與代名詞時，則會誤判爲「粗糙遞移」（狀況 3.3），因此我們根據此重心模型加以改進，將「粗糙遞移」的條件限制較爲嚴格一些，同時將「平順遞移」的條件放寬。詳細情形請參考後面的狀況 4 敘述。

當小句重心候選詞有多個時，以重心模型分成以下四種狀況：

狀況 1. $C_{i-1}=C_{i-2}$ 且

狀況 1.1 $\exists C \in \text{Can}(C_i)$, $C=C_{i-1}$，則$C_i=C=C_{i-1}$。

狀況 1.2 C(3rd-anaphor)$\in$Can(C_i) OR C(Zero-anaphor)$\in$Can(C_i)，則$C_i=C_{i-1}$。

其中Can(C_i)表示C_i的重心候選詞集合，在狀況 1 中我們加入了第三人稱代名詞的辨識與零指代辨識。

狀況 1.1 表示有一個候選詞C與前一小句重心C_{i-1}相同，則C_i「延續」前一小句重心。

狀況 1.2 則表示此小句含有第三人稱代名詞或者零指代，即小句有延續上小句論述的情形，則C_i「延續」上小句之重心。

狀況 2. $C_{i-1}=C_{i-2}$, 且

狀況 2.1 $\forall C \in \text{Can}(C_i)$, $C \neq C_{i-1}$且C(3rd-anaphor)$\notin$Can(C_i) AND C(Zero-anaphor)$\notin$Can(C_i)，則C_i=Can(C_i)。

狀況 2.2 前一小句沒有重心(標示爲E)，或者此小句爲長句中的第一小句，則C_i=Can(C_i)。

狀況 2.1 即是非狀況 1 的情形，C_i所有候選詞均與前一小句重心C_{i-1}不一致，且無第三人稱代名詞與零指代情形，則「保留」C_i所有候選詞Can(C_i)，以待下一小句決定爲「平順遞移」或者「粗糙遞移」，再行決定C_i。

狀況 2.2 前小句沒有重心（即沒有重心候選詞的小句），或者沒有前小句可參考（例如在長句的第一小句時），則「保留」此小句候選詞以待下小句決定。

設前小句重心C_{i-1}未決定，即前小句爲「保留」情形時，此時需根據本小句與前小句之關連，判斷是「平順遞移」抑或「粗糙遞移」，我們分成以下幾種狀況來討論：

狀況 3. $C_{i-1} \neq C_{i-2}$, 且

狀況 3.1：$\exists C_k \in \text{Can}(C_{i-1})$, $\exists C_j \in \text{Can}(C_i)$, $C_k = C_j$，則$C_i=C_{i-1}=C_k=C_j$。

狀況 3.2：C(3rd-anaphor)$\in$Can(C_i) OR C(Zero-anaphor)$\in$Can(C_i) 則$C_i=C_{i-1}$=C, where Freq(C)=MaxFreq(Can(C_{i-1}))。

狀況 3.3：$\forall C_k \in \text{Can}(C_{i-1})$, $\forall C_j \in \text{Can}(C_i)$, $C_k \neq C_j$，且C(3rd-anaphor) $\notin$ Can(C_i) AND C(Zero-anaphor) $\notin$ Can(C_i)，則

$C_i = C_{i-1} = C$, where Freq(C)=MaxFreq(Can(C_i), Can(C_{i-1}))。

MaxFreq($T_{i...}T_k$)：計算$T_{i...}T_k$，k個詞彙於此文章中出現之頻率，並取其頻率最高者。若爲第一、第二人稱代名詞，其頻率視爲 0。

在狀況 3.1 中，C_{i-1}與C_i兩小句重心的候選詞中有在兩句都出現者，則取其爲兩句重心；若有多個候選詞在兩句中都出現，則取其於此文章中出現頻率最高者。重心由C_{i-2}平順遞移至C_{i-1}與C_i。

狀況 3.2 中，由於本小句有第三人稱代名詞或零指代，則取前小句重心候選詞Can(C_{i-1})中出現於此文章最高頻的候選詞當作兩句重心。重心由C_{i-2}平順遞移至$C_{i-1}=C_i$。

狀況 3.3 中，C_{i-1}與C_i皆有多個候選詞，且第i小句不含代名詞或零指代，C_{i-1}與C_i之候選詞亦無重複時，則兩小句重心$C_i=C_{i-1}$=MaxFreq(Can(C_i),Can(C_{i-1})) ，即從兩小句的候選詞中挑選文章詞頻之最高者。重心由C_{i-2}平順遞移至C_{i-1}與C_i。

狀況 4. $C_{i-1} \neq C_{i-2}$, 且

Can(C_{i-1})={C_1}, Can(C_i)={C_2}，$C_1 \neq C_2$，則$C_i=C_2$。

狀況 4 中，C_{i-1}與C_i候選詞均只有 1 個，各爲C_1與C_2，且$C_1 \neq C_2$，則爲「粗糙遞移」。

與重心理論不同的是，在狀況 4 中，我們將「粗糙遞移」的狀況限制的很嚴謹，並將原本應視爲「粗糙遞移」的狀況視爲是「平順遞移」（狀況 3.3）。這是因爲若C_{i-1}已爲「保留」狀況，且此時C_{i-1}想帶出C_i，C_i才是重點，兩小句既無重複的候選詞亦無代名詞與零指代，這種情形按照原始的重心模型（即將狀況 3.3 視爲「粗糙遞移」），便會誤判是「粗糙遞移」，如以下兩個小句：

例 4a. 從外國的經驗來看，

Can =外國, 經驗，$Center_{original}$=經驗，$Center_{improved}$=台鐵

例 4b. 台鐵主要有兩條新的出路：捷運化與觀光化。

Can = 台鐵, 出路, 捷運，$Center_{original}$ =台鐵，$Center_{improved}$ =台鐵

$Center_{original}$代表原始重心模型選取的重心，$Center_{improve}$則是我們的改良法所選取的。4a小句的候選詞爲「外國」與「經驗」，由於是第一小句，依狀況 2.2，保留所有候選詞。4b小句候選詞爲「台鐵」、「出路」、「捷運」，並未含有零指代與代名詞，且候選詞亦未與 4a小句重複，此爲狀況 3.3。若按照原始重心模型，兩小句將獨立各自選取重心，因此 4a小句選取「經驗」，4b小句選取「台鐵」結果如$Center_{original}$所示。

但 4a小句很明顯地是想帶出後面的主題：「台鐵兩條新出路」，而不是真正想講述外國的經驗，縱使 4a小句標示爲「經驗」不能當成錯誤，但因此視爲是「粗糙遞移」並不適當。且比起「經驗」，若能將 4a小句和 4b小句一起標示爲「台鐵」，更爲恰當且更能幫助之後長句主題詞的辨識。因此我們將狀況 3.3 改爲「平順遞移」，兩小句一起選出共同的重心「台鐵」，如$Center_{improved}$。

「粗糙遞移」在語篇當中原本就較少出現，即使判定爲「平順遞移」影響亦不大，且此舉將更有助於之後的長句主題詞辨識，因此我們認爲將狀況 3.3 改爲「平順遞移」並不會造成太多「粗糙遞移」誤判爲「平順遞移」。

另外，在 MaxFreq( )計算詞頻之時，我們會特別對「的」後面的名詞候選詞加權 1.5 倍。常見「的」在中文上的用法可分爲兩種，一種是形容詞修飾，即形容詞＋「的」＋名詞，另一種是所有格用法。

例 6a： 炙熱(VH) 的(DE) 陽光(Na) （形容詞修飾用法）

例 6b： 小明(Nb) 的(DE) 課本(Na) （所有格用法）

這兩者的重心都偏向「的」後面的名詞，而不是在於前面的名詞。

我們以實驗語料 11 篇社論共 1284 小句測試加權「的」後的名詞，結果發現未加權「的」後的名詞共有 204 小句重心錯誤，加權「的」之後可將錯誤的小句數減少至 188 小句。

以下舉一個簡單的小句重心選取實例：

表 3-2 小句重心選取實例

小句編號	小句	Can(C_i)	動作	重心
#1	我們 相信歷任政府及許多台鐵員工都想努力改善台鐵的沉畔，	我們(Ｉ) 政府 員工 台鐵 沉畔	第一小句，保留所有候選詞。(狀況 2.2)	台鐵
#2	然而台鐵以往留下來的包袱實在太大，	台鐵 包袱	僅「台鐵」與上句重複，故兩句重心皆爲「台鐵」。(狀況 3.1)	台鐵
#3	未來前景不明，	前景	僅一個候選詞，直接決定重心爲「前景」。	前景
#4	因此在政府決定及組織的限制下，	決定 組織 限制	無與上句重心重複之候選詞，保留所有候選詞。(狀況 2.1)	台鐵
#5	台鐵的情況一直不易改善。	台鐵 情況	兩句候選詞中選取詞頻最高者：「台鐵」 (「限制」、「情況」雖被加權，但加權後仍爲「台鐵」獲勝) (狀況 3.3)	台鐵

小句#1 爲長句的第一小句，依狀況 2.2，保留所有候選詞。

小句#2 僅「台鐵」與#1 小句重複，依狀況 3.1，將#1 與#2 小句重心標示爲「台鐵」。

小句#3 僅一個重心候選詞，直接決定其重心爲「前景」。

小句#4 並未與上小句有重複的候選詞，依狀況 2.1 因此保留所有候選詞。

小句#5 爲狀況 3.3，「情況」雖在「的」後有被加權，但仍未超越「台鐵」，前一句候選詞「限制」亦被加權但亦超越「台鐵」，因此「台鐵」成爲#4 與#5 的重心。

3.3 小句重心實驗結果與分析

我們蒐集了 11 篇聯合報的社論，內容爲政治和經濟議題，表 3-3 列出此語料的統計資料。1284 個小句當中，有 128 小句無重心候選詞，因此在計算重心選取正確率時，我們會將這 128 小句排除，僅檢視 1156 小句的重心選取結果。

表 3-3：語料統計資訊

語料類型	篇數	字數	標點符號數	長句數	小句數	重心數
社論文章	11	15404	1348	276	1284	1156

我們以原始重心模型來選取小句重心，則得出共 203 句錯誤，正確率爲 82.42%；若以「粗糙遞移」條件嚴格限制的改善方法選取，則可得出 188 句錯誤，正確率爲 83.7%。如下表 3-4：

表 3-4：小句重心選取結果比較表

重心選取方法	正確小句數	錯誤小句數	正確率
原始重心模型方法	953	203	82.42%
嚴格限制「粗糙遞移」條件之改善法	968	188	83.70%

下表 3-5 爲簡單的原始法與改善法比較的例子。由於#4 小句的候選詞不與#3 小句重心相同，爲「保留」情形。#5 小句與#4 小句的候選詞亦無相同且無零指代與代名詞，在未嚴格限制「粗糙遞移」條件時，即判定爲「粗糙遞移」，其重心由各小句候選詞挑選之，造成#4 小句標記成「組織」的錯誤（「限制」於「的」後雖被加權，但加權後頻率未超越「組織」）。

表 3-5：小句重心選取原始法與改善法差異範例

小句編號	小句	候選詞	改善法	原始法
#1	我們相信歷任政府及許多台鐵員工都想努力改善台鐵的沉眛，	我們(I) 政府 員工 台鐵 沉眛	台鐵	台鐵
#2	然而台鐵以往留下來的包袱實在太大，	台鐵 包袱	台鐵	台鐵
#3	而未來前景又不明，	前景	前景	前景
#4	因此在政府決定及組織的限制下，	政府 組織 限制	台鐵	*組織*
#5	台鐵的情況一直不易改善。	台鐵 情況	台鐵	台鐵

經由嚴格限制「粗糙遞移」條件之後，第四與第五小句將由兩個小句的候選詞一起挑選出一個共同的重心：「台鐵」。

但經由改進之後，仍存在一些錯誤，以下爲例：

表 3-6：仍存在的錯誤範例

小句編號	小句	候選詞	動作	重心標記
#1	但真正引人關注的，		無候選詞，標示 E	E
#2	乃是懷抱所謂「五要」兩岸新架構的國民黨主席，	懷抱 兩岸 架構 主席	前小句標示 E，保留所有候選詞。(狀況 2.2)	兩岸
#3	與三合一敗選後大步投向台獨基本教義派懷抱，	零指 敗選 台獨 教義派	含有零指代，取前小句候選詞頻率高者「兩岸」爲前小句與本小句重心。(狀況 3.2)	兩岸
#4	以「廢統」、「終統」搞得人心惶惶、國際社會瞠目結舌的阿扁總統，	廢統 終統 社會 總統	與前小句無候選詞重複，亦無零指代與代名詞。保留所有候選詞。 (狀況 2.1)	總統
#5	雙方南轅北轍的主張，	代名 主張	含有代名詞，選取前小句詞頻最高者「總統」爲前小句與本小句重心。 (狀況 3.2)	總統

縱使#2 小句當中，「主席」於「的」後被加權，但仍未超過「兩岸」，再加上#3 誤判零指代，因此「兩岸」被當成是兩個小句的重心。

在我們描述的重心選取方法當中，當小句含有零指代時，重心將直接跟隨上

一小句的重心，但我們也發現會造成如上表的錯誤範例，因此我們將詳細地評估這個方法所帶來的影響。

在我們的實驗語料 1284 小句中，系統共辨識出 338 個小句有零指代情形，42 小句爲辨識錯誤，74 小句爲消解錯誤（我們將前小句重心視爲是消解的對象）。

辨識錯誤的 42 小句當中有 31 小句位在長句的第一小句，因此系統在判斷小句重心時將直接忽略零指代情形，保留所有候選詞。剩餘 11 個辨識錯誤僅有 2 個會造成長句主題詞判斷的影響，其餘並不因此判斷成錯誤的長句主題詞（我們以 3.5.3 小節的實驗模組Ⅲ來選取長句主題詞）。不影響長句主題詞的判斷標準爲：即使這個小句重心標記錯誤，但整個長句並不因爲這個小句重心錯誤而標記成錯誤的主題詞。

在 74 消解錯誤中，有 52 小句並不影響長句主題詞的正確判斷，僅有 22 小句造成長句主題詞選取的影響。總結爲 1284 小句中零指代辨識與消解會影響長句主題詞判斷者，僅 24 小句，因此我們認爲以我們的零指代辨識方法，以及小句含有零指代時直接承接上句重心的方法對於長句主題詞的影響是有限的，其結果如下表 3-7。

表 3-7：零指代錯誤對長句主題詞之影響評估表

338 小句辨識爲零指代情形	小句數	不影響長句主題詞	影響長句主題詞
零指代辨識錯誤	42	40	2
零指代消解錯誤	74	52	22
總數	116	92	24

3.4 長句主題詞選取

根據[徐爲群 et al., ‘04]，中文語句的「主題」可視爲「說話者所關注的顯著語意實體 (Salient Semantic Entity, SSE)」，此與重心理論的重心定義是相似的。

因此，我們可以說長句主題詞就是小句重心的延伸，以下我們將根據小句重心的選取結果來選取長句主題詞。

依據主題詞的特性，長句主題詞可依頻率特徵、位置特徵、主題一致性特徵、.延伸特徵、概念化特徵、分佈特徵進行萃取。我們依照這些特徵提出一個公式來計算小句重心C_i在長句中的重要性，並從中萃取長句主題詞：

$$TopicWord(S_k) = C_i \text{, where } Weight(S_k, C_i) = \underset{C \in Can(S_k), Can(S_{k-1})}{Max} (Weight(S_k, C))$$

$$Weight(S_k, C_i) = \mathrm{E}(S_k, C_i) \times D(C_i) \quad (1)$$

$$\mathrm{E}(S_k, C_i) = \alpha \times \mathrm{CenterFreq}(S_k, C_i) + \mathrm{P}(S_k, C_i) + \beta \times \mathrm{E}(S_{k-1}, C_i) \quad (2)$$

$TopicWord(S_k)$ 爲長句S_k的主題，C_i爲S_k第i個主題候選詞，即S_k的第i小句的重心；$Can(S_k)$表示S_k的主題候選詞集合。$Weight(S_k, C_i)$ 爲C_i於S_k的重要性權重；$E(S_k,C_i)$爲C_i於S_k的主題重要性；S_{k-1}爲S_k的前一個長句。下表 3-8 列出公式(1)(2)所包含的各個特徵：

表 3-8：長句主題詞特徵一覽表

特徵名稱	符號	使用時的值	不使用時的值
頻率特徵	$\mathrm{CenterFreq}(S_k, C_i)$	C_i在S_k中的重心次數	-- (必須使用)
位置特徵	$\mathrm{P}(S_k, C_i)$	2 或 0 (需依照C_i位置而定)	0
一致性特徵	α	2 或 1 (需依照C_i與S_{k-1}一致性而定)	1
延伸特徵	β	0.5 或 0 (需依照S_k與S_{k-1}之關係而定)	0
概念化特徵	-- (配合頻率特徵使用)	頻率特徵改計算C_i之概念於S_k中的重心次數	維持原本頻率特徵之算法
分佈特徵	$D(C_i)$	Log (SentenceFreq(C_i)+1)	1

以下我們將逐一介紹這些特徵，並描述這些特徵使用時的值是如何訂定的。

3.4.1 .頻率特徵

一個長句的主題詞應以此主題爲中心來陳述，故其頻率可作爲主題識別的依據，因此在這個特徵裡，我們計算這個長句中每個候選詞被選取成爲小句重心的頻率。若單獨使用頻率特徵時，CenterFreq( )值高者即標示爲長句主題詞，若遇主題候選詞 CenterFreq( )值同高時，則我們隨機取其一成爲主題。

3.4.2 位置特徵

長句的主題常在長句的第一個小句就已經出現，故第一小句對於整個長句而言，往往是個很重要的小句。我們抽取實驗語料的前 48 個長句觀察的結果，有 33 長句主題詞在第一小句就已經出現。

因此，我們針對第一小句的重心給予加重$P(S_k,C_i)$，而爲了取得$P(S_k,C_i)$的數值，我們先令公式(1)(2)中的 α =1，β=0，$D(C_i)$ =1（除頻率特徵外，其餘特徵皆不使用），最後取$Weight(S_k,C_i)$ 值最高者爲長句主題詞，並觀察$P(S_k,C_i)$對長句主題詞的影響程度。

我們以上述的 48 個長句來作測試。測試結果如下圖 3-2。若$P(S_k,C_i)=0$，即不考慮爲第一小句重心加權，純粹以頻率特徵則可得 33 長句正確。$P(S_k,C_i)=2$ 時可達最高的正確數，而$P(S_k,C_i)=4$ 時正確的 33 長句也剛好是主題詞在第一小句就出現的長句，表示一般主題詞於小句重心的出現次數均小於等於 4，因此將$P(S_k,C_i)$設定爲 4 就等於直接認定第一小句重心爲主題詞。

由圖 3-2，我們將第一小句的$P(S_k,C_i)$設定爲 2。其數學表示式如下：

$P(S_k,C_i)=2$, if i=1，即C_i爲第一小句之重心。 $P(S_k,C_i)=0$, if i>1，即C_i非第一小句之重心。	(3)

主題位置特徵公式

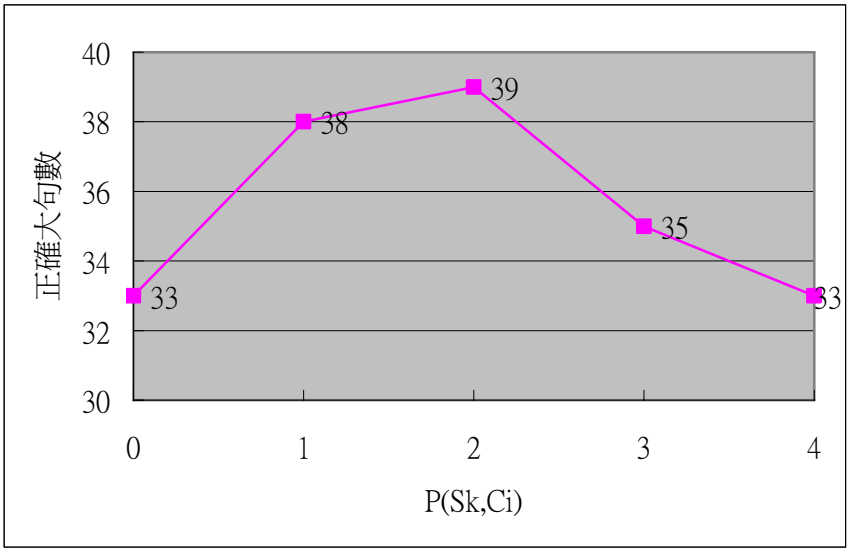


圖 3-2：$P(S_k,C_i)$對 48 長句正確數趨勢圖

3.4.3 主題一致性特徵

當候選詞與上長句主題詞一致時，這個長句往往有繼續上長句的主題來作論述的傾向，因此我們考慮爲與前長句主題詞相同的候選詞作加權計算，此即公式(2)中的α參數。

α爲候選詞C_i與前長句主題詞$TopicWord(S_{k-1})$相同時的加權參數，現欲決定適當的α值，與 3.4.2 小節相同，我們以 48 長句來評估α，取主題詞時不使用除了頻率特徵之外的其餘特徵。其結果如圖 3-3。

根據圖 3-3，我們將α設定爲 2，得公式如下：

$$\alpha=2，\text{if } C_i = TopicWord(S_{k-1})$$
$$\alpha=1，\text{if } C_i \neq TopicWord(S_{k-1}) \quad (4)$$

主題一致性特徵公式

3.4.4 主題延伸特徵

在語段中經由連接詞的句法作用，長句的主題會延續到下一個長句，因此

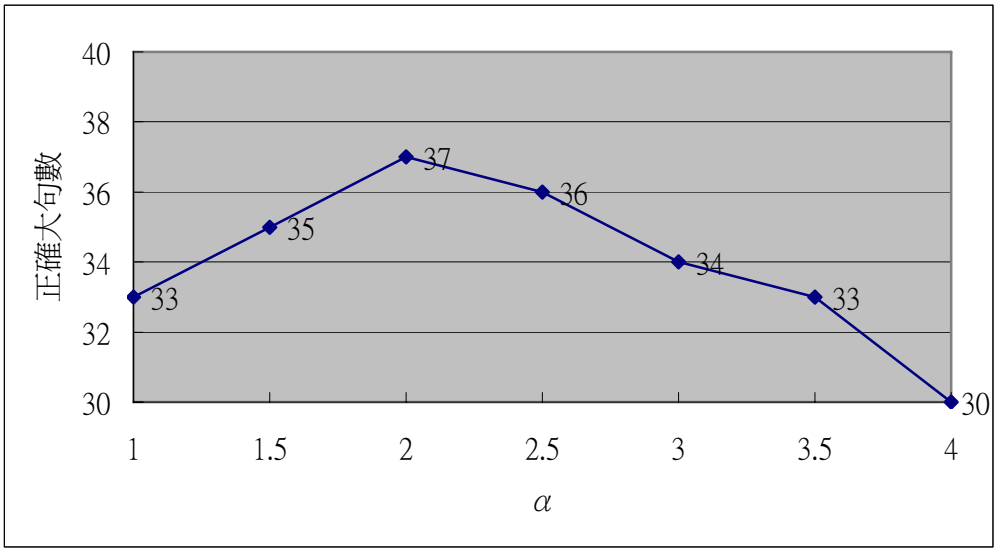


圖 3-3： α 對 48 長句正確數趨勢圖

長句的第一小句含有連接詞時，其涵義往往是承襲上一長句的論述。一般連接詞可分成正向連接詞與反向連接詞，其中正向連接詞有論述延續的傾向，例如「以及」、「同樣」、「比如」、「也就是說」、「尤其」、「或是」、「若是」等等詞彙。我們參考[鄭守益, '06]，抽取 309 個語篇連貫標記的連接詞，以人工方式將其標記爲正向、反向連接詞，其中共有 268 個正向連接詞。

在長句中第一小句含有正向連接詞時，我們將前長句的小句重心也一併納入此長句來計算，但給予減輕權重，以免過度影響本長句的主題詞判斷，其減輕的權重即爲公式(2)中的 β。

爲了評量 β 對主題詞的影響程度，我們一樣以 48 長句來評量 β 值，取主題詞時不使用除了頻率特徵之外的其餘特徵，其結果如圖 3-4。在 β 爲 0，即遇到連接詞亦不考慮合併前長句之結果，共得 33 句正確；β 爲 0.25 時，由於值仍太小，β 並未發揮其功能，保持 33 句正確；當 β=0.5，增爲 35 句；β=0.75 時，下降爲 34 句；β=1，完全計入前長句的結果，則反而造成更多錯誤，爲 32 小句正確。

根據圖 3-4，我們將 β 設定爲 0.5，其公式如下：

β=0.5，if S_k的第一小句含有正向連接詞 β=0， if S_k的第一小句不含正向連接詞	(5)

主題延伸特徵公式

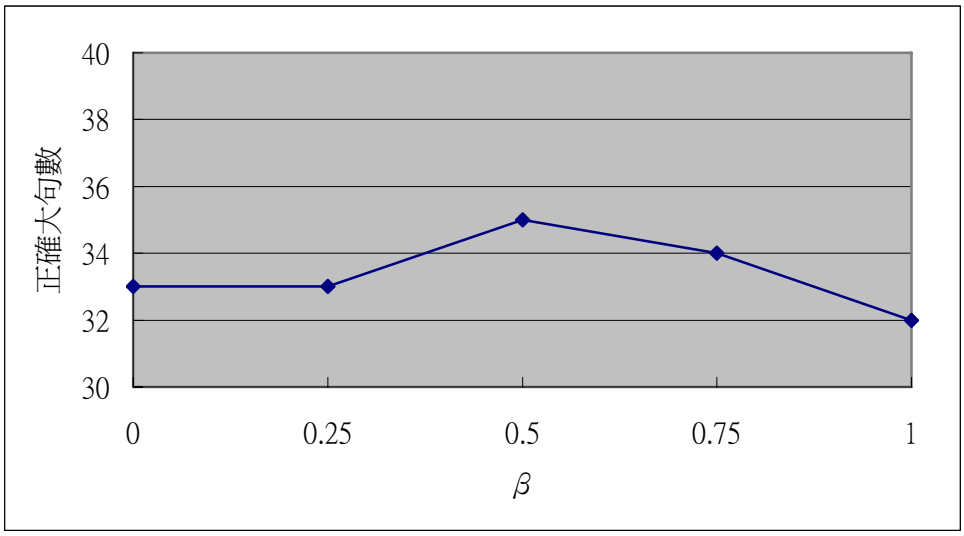


圖 3-4： β 對長句正確數趨勢圖

由圖 3-4，β=1 會造成前長句過度影響此長句，因此其結果反倒不如 β=0。這種情形尤其發生在較短的長句以連接詞承接前面較長的長句時，若 β=1，則在前面的長句極有可能喧賓奪主地將後面的短長句主題詞奪走，而 β=0.5 則可以減緩前長句對此長句的影響。

下表 3-9 爲一個例子：

表 3-9：β = 0.5 與 β =1 的比較範例

長句	小句編號	小句	小句重心	長句主題 β = 0.5	長句主題 β = 1
S_{k-1}	#1	這裡適當的組織方式及人選，	政府	政府	政府
	#2	政府可與各界充分交換意見再做決定；	政府		
	#3	但未來營運除一般公用事業的規範外，	政府		
	#4	皆由受委託的團隊來負責。	政府		
S_k	#5	且台鐵不僅面對高速公路、私家車、以及航空業的競爭，	台鐵	台鐵	*政府*
	#6	很快又有不少生意會被高鐵替代，	台鐵		
	#7	以後的業務必將持續甚至大幅萎縮。	業務		

#5 小句含有正向連接詞「且」，因此我們將納入上個長句的小句重心來作此長句的主題詞候選詞。由於上長句累積過來的小句重心有超過兩個標示爲「政府」，因此「政府」雖然沒有在此長句中出現，在 β=1 依然被錯標成此長句的主題詞。若 β = 0.5，則「政府」將與「台鐵」獲得相同的頻率（2 次），我們優先選擇本句的候選詞「台鐵」爲長句主題詞。

3.4.5 概念化特徵

有時文章作者會以同義詞來表示相同的概念，以避免過度使用某個詞彙造成詞窮情形，此時兩個詞彙的字面上雖然不同，但想描述的意義是相同的。因此在這個特徵中，我們考慮候選詞的概念，以[同義詞林]來標記每個小句重心的概念，同義詞林共有 62642 個詞彙，每個詞彙都有四碼標記，前兩碼爲英文字母，後兩碼爲阿拉伯數字。

而主題陳述時除了使用同義詞避免詞窮外，有時也會針對這個主題的子主題(Sub-topic)來加以詳述，因此我們將實驗語料的 1284 小句重心取同義詞林前三碼標記之，詞彙前三碼相同者則視爲同一個概念，以此輔助計算主題候選詞C_i對於S_k的重要性。使用概念化特徵時，我們將頻率特徵之計算方式改成計算C_i之概念於S_k被當成小句重心的次數。

由於主題講述時有大主題延伸至子主題的概念，大主題會出現在前面，子主題則會出現於後面，故在這個特徵中，當有兩個以上候選詞 Weight( )都爲最高時，我們將以前句的大主題優先，取先出現的候選詞爲長句主題詞。

實驗語料的 1284 長句當中有 128 小句沒有重心候選詞(標示爲 E)，剩餘 1156 小句以同義詞林標記概念，共標記了 728 個小句重心，剩餘 427 小句重心則以人工標示其概念。

3.4.6 分佈特徵

主題有時並不僅僅只出現在同一個長句，其他長句也會有所提及，因此在這個特徵當中，我們在計算主題候選詞C_i的$E(S_k,C_i)$之後，考量這個候選詞位於其他長句的出現頻率，此即公式(2)的 $D(C_i)$ ， $D(C_i)$ 計算公式如下：

$$D(C_i) = \text{Log}\ (\ \text{SentenceFreq}(C_i)+1\ ) \qquad (6)$$

$\text{SentenceFreq}(C_i)$是C_i出現在此篇文章的長句數，例如出現於兩個長句中，則 SentenceFreq()即爲 2。

3.5 實驗模組建立

在這個小節當中，我們將藉由 3.4 節描述的各個長句主題詞特徵，整合成爲五個選取長句主題詞的實驗模組。如下表 3-10：

表 3-10：實驗模組所含特徵一覽表

	頻率特徵	位置特徵	主題一致性特徵	延伸特徵	概念化特徵	分佈特徵
實驗模組 I (基本實驗模組)	V	V				
實驗模組 II (基本+前長句)	V	V	V	V		
實驗模組 III (基本+前長句+概念)	V	V	V	V	V	
實驗模組 IV (基本+前長句+分佈)	V	V	V	V		V
實驗模組 V (全部特徵)	V	V	V	V	V	V

以下我們將詳細描述五個實驗模組。

3.5.1 實驗模組 I

實驗模組I 爲主題頻率特徵、位置特徵，合計共兩個特徵，這兩個特徵僅考

慮S_k長句本身，並未考慮前一長句S_{k-1}。我們將實驗模組I視爲是基本的實驗模組，以下的實驗模組則是將其餘的特徵加入基本實驗模組中，以供實驗結果進行比較。.

3.5.2 實驗模組Ⅱ

實驗模組Ⅱ是基本實驗模組再加上主題一致性與延伸特徵，合計共四個特徵。與實驗模組I的不同之處在於主題一致性與延伸特徵乃將前一長句S_{k-1}列入考量範圍，其中一致性特徵考慮S_{k-1}之主題詞，延伸特徵則將S_{k-1}之主題候選詞也納入S_k之候選詞。

3.5.3 實驗模組Ⅲ

實驗模組Ⅲ爲基本實驗模組再加上一致性特徵、延伸特徵、概念化特徵，合計共五個特徵。與實驗模組 II 不同之處在於加入了概念化的特徵。

3.5.4 實驗模組Ⅳ

實驗模組Ⅳ是由基本實驗模組再加上一致性特徵、延伸特徵、分佈特徵共五個特徵所整合出的模組。其與實驗模組Ⅲ不同之處在於未考慮概念化特徵，但加入分佈特徵。

3.5.5 實驗模組 Ⅴ

實驗模組 Ⅴ 是使用所有主題頻率特徵、位置特徵、一致性特徵、延伸特徵、概念化特徵、分佈特徵共六項特徵的模組。

3.6 長句主題詞選取實驗與分析

在這個小節中，我們將描述各個實驗模組的結果與分析。

3.6.1 語料說明

長句主題詞實驗所使用的語料與 3.3 小節小句重心實驗使用的是一樣的語料，爲聯合報的 11 篇社論。我們以人工方式先行標記這 276 長句的主題詞，作爲評量系統的依據，276 長句以人工方式共標記了 351 個主題詞。其中 2 篇社論共 48 長句已作爲開發語料，我們將剩餘的 9 篇社論共 228 長句作爲社論測試語料，其主題詞有 289 個。

另外我們與新竹市建功高中的國文教師合作，蒐集了 95 篇學生作文以供後續的離題與連貫性評量語料，此作文語料將在 4.1 節詳細描述。我們由此作文語料選出高分作文（13 分以上）22 篇作爲第二份主題詞的作文測試語料，之後使用社論測試語料中表現最好的實驗模組來萃取此作文測試語料的主題詞，以此評量我們的主題詞萃取方法在一般學生作文上的效能。

雖然系統僅能每長句挑出一個主題詞，但爲求謹慎，我們除了衡量正確率(Precision)之外，亦評估其召回率(Recall)，最後計算 F 指標(F-measure)。其計算公式如下：

Precision = #Correct / 總長句數

Recall = #Correct / 總主題詞數

F = 2 * Precision * Recall / (Precision + Recall)

#Correct 表示系統標示正確的長句數量。

3.6.2 實驗與分析

除了五個實驗模組之外，我們作了另外兩個純粹以詞彙來判斷主題詞的系統，詞彙 I 與詞彙 II。詞彙 I 計算每個小句重心候選詞（Na, Nb, Nc 與名物化動詞）於此長句中的出現頻率，小句重心候選詞的產生在 3.1 小節已經詳述過，長句主題詞則取最高頻的小句重心候選詞。

詞彙 II 則計算長句之中的每一個動詞(VA,VB,VC)與名詞(Na, Nb, Nc)頻率，以頻率最高的詞彙當成長句主題詞，不考慮零指代與代名詞。實驗結果如下表 3-11。

表 3-11：長句主題詞實驗結果一覽表

	系統	正確數	錯誤數	正確率 (Precision)	召回率 (Recall)	F 指標 (F-measure)
重心理論	實驗模組 I (基本實驗模組)	188	40	82.46%	65.05%	72.73%
	實驗模組 II (基本+前長句)	194	34	85.09%	67.13%	75.05%
	實驗模組Ⅲ (基本+前長句+概念)	198	30	86.84%	68.51%	76.59%
	實驗模組Ⅳ (基本+前長句+分佈)	196	32	85.96%	67.82%	75.82%
	實驗模組 V (全部特徵)	197	31	86.40%	68.16%	76.20%
詞彙	詞彙-I (小句重心候選詞)	101	127	44.29%	34.94%	39.07%
	詞彙-II (全部詞彙)	89	139	39.04%	30.80%	34.43%

而除了以社論測試語料來作各個實驗模組的評比之外，我們也隨機抽取社論中的長句以比較實驗模組，以便更能了解各實驗模組之間的差異。我們隨機抽樣 40 個長句，共抽樣五次，下表 3-12 爲抽樣的實驗結果。

表 3-12：長句主題詞抽樣實驗結果

抽樣編號	正確率 / 召回率				
	實驗模組 I	實驗模組 II	實驗模組Ⅲ	實驗模組Ⅳ	實驗模組 V
1	75.00%/58.82%	87.50%/68.62%	90.00%/70.58%	87.50%/68.62%	87.50%/68.62%
2	67.50%/54.00%	77.50%/62.00%	82.50%/66.00%	77.50%/62.00%	80.00%/64.00%
3	80.00%/61.53%	80.00%/61.53%	80.00%/61.53%	80.00%/61.53%	80.00%/61.53%
4	82.50%/63.46%	87.50%/67.30%	85.00%/65.38%	87.50%/67.30%	85.00%/65.38%
5	77.50%/59.61%	80.00%/61.53%	82.50%/63.46%	82.50%/63.46%	82.50%/63.46%
AVG	76.50%/59.48%	82.50%/64.20%	84.00%/65.39%	83.00%/64.60%	83.00%/64.60%
VAR	0.33%/0.13%	0.22%/0.12%	0.14%/0.11%	0.20%/0.10%	0.11%/0.07%

以下舉實驗模組 II, Ⅲ, Ⅳ, Ⅴ選取正確但實驗模組 I 錯誤的例子：

表 3-13：實驗模組之結果比較範例 1

長句	小句編號	小句	小句重心	實驗模組 I	實驗模組 II	實驗模組 Ⅲ	實驗模組 Ⅳ	實驗模組 Ⅴ
S_{k-1}	#1	因此台鐵的捷運化也應和台北市區的捷運一樣，	台鐵	台鐵	台鐵	台鐵	台鐵	台鐵
	#2	由政府負擔相當程度的成本，	台鐵					
	#3	不要完全自負盈虧。	台鐵					
S_k	#4	而在捷運化的同時，	捷運	*捷運*	台鐵	台鐵	台鐵	台鐵
	#5	台鐵也可配合進行車站一帶的土地開發，	台鐵					
	#6	活化台鐵土地資產。	台鐵					

在表 3-13 的長句之中，很明顯#4 小句中出現「而」，有承接上一個長句的傾向，而「台鐵」亦爲上長句主題詞。然而在實驗模組I，未加入延伸與一致性特徵時，S_k的第一小句(#4)的重心爲「捷運」，被加重計票($P(S_k,C_i)=2$)的結果使得實驗模組I最後的長句主題詞變成「捷運」。但加入了延伸特徵與一致性特徵後，「台鐵」會被加權計算，則「台鐵」就會正確地標記出來。

表 3-14 則爲實驗模組Ⅲ、Ⅴ正確但實驗模組 II、Ⅳ皆錯誤的實驗結果範例：

表 3-14：實驗模組之結果比較範例 2

小句編號	小句	小句重心	實驗模組 I	實驗模組 II	實驗模組 Ⅲ	實驗模組 Ⅳ	實驗模組 Ⅴ
#1	當然，	E	機關	*一線*	機關	*一線*	機關
#2	這最終正義的實現，	機關					
#3	尚須司法檢調機關爭氣才行；	機關					
#4	台北地院的這樁無效判決，	台北地院					
#5	只是曙光一線，	一線					
#6	還不是大白天！	一線					

由於第一小句重心爲 E，故忽略之。第一小句未含連接詞，「機關」與「一線」

CenterFreq( )各爲 2，故實驗模組 I、II 必須由「機關」與「一線」取一，實驗模組 I 在隨機取一時碰巧抓到正確的「機關」。在實驗模組Ⅳ中，「機關」與「一線」的分佈特徵 D( )計算之後得到相同的值，隨機取一結果「一線」成爲長句主題詞，錯誤地被標上去。

而在加入概念化的實驗模組Ⅲ、Ⅴ中，「機關」與「一線」雖 CenterFreq( )值都是 2，但「台北地院」爲「機關」之一，故「機關」在實驗模組Ⅲ、Ⅴ時 ConceptFreq( )值爲 3，同樣地「台北地院」的 ConceptFreq( )值亦爲 3，我們以前句優先的策略將「機關」標示爲主題詞。

以下我們舉實驗模組I 、II、Ⅲ皆正確但實驗模組Ⅳ、Ⅴ錯誤的例子。在表 3-15 的S_k長句中，延伸特徵將S_{k-1}的主題候選詞「先生」延伸至S_k，但其權重減半，因此於實驗模組I、II、Ⅲ中，可以正確的標示出主題爲「政策」。但「先生」在分佈特徵中的權重比「政策」還要多，故造成實驗模組Ⅳ、Ⅴ將「先生」判定爲主題詞，而非「政策」。

表 3-15：實驗模組之結果比較範例 3

長句	小句編號	小句	小句重心	實驗模組 I	實驗模組 II	實驗模組 Ⅲ	實驗模組 Ⅳ	實驗模組 Ⅴ
S_{k-1}	#1	不禁使人回想起５０年前，	人	先生	先生	先生	先生	先生
	#2	尹仲容先生爲突破當時「積極管理」的禁錮，	先生					
	#3	推出一連串改革開放政策，	先生					
	#4	促成其後台灣３０年的經濟快速成長，	先生					
	#5	也累積了大量資源，	先生					
	#6	卻變成近十年來內耗的本錢。	先生					
S_k	#7	更沒想到，	E	政策	政策	政策	*先生*	*先生*
	#8	５０年後的今天又要重回５０年前的「積極管理」政策，	政策					
	#9	真不知今世何世，	政策					
	#10	怎不教人唏噓！	政策					

根據表 3-11 與表 3-12，我們取最好的實驗模組Ⅲ來進行作文測試語料的主題詞萃取。主題詞作文測試語料共 22 篇高分學生作文，188 長句，903 小句，其中 140 小句無重心候選詞。188 長句以人工標記共 213 個主題詞。其實驗結果如下表 3-16：

表 3-16：實驗模組Ⅲ對作文語料萃取主題詞之實驗結果

正確小句數	錯誤小句數	小句重心正確率	正確長句主題數	錯誤長句主題數	主題正確率	主題召回率	主題 F 指標
604	159	79.16%	152	36	80.86%	71.36%	75.81%

由於學生作文的寫作並未比社論來得嚴謹，也較爲口語化，小句重心候選詞分佈較爲零散，因此小句重心正確率不如社論語料；故以實驗模組Ⅲ來作長句主題詞萃取的結果亦不如社論測試語料。

下一章節我們將描述如何使用長句主題詞與小句重心來偵測作文的離題、連評量貫性、以及文章結構的分析。

第四章 作文主題分析與應用

長句主題詞與小句重心可以用來協助教師作學生作文的離題偵測(off-topic detection)、連貫性評量(coherence evaluation)、以及文章結構分析。離題偵測即是偵測學生的文章有沒有偏離主題，連貫性評量則是評量學生的寫作敘述是否有條理，結構分析則是分析文章的敘述結構。

在這個章節中，4.1 小節是學生作文語料的說明，4.2 小節說明如何應用長句主題詞來作離題偵測，4.3 小節說明如何應用小句重心來作連貫性評量，4.4 小節爲應用長句主題詞分析文章結構。

4.1 學生作文語料

我們與新竹市建功高中的國文教師合作，蒐集了 95 篇學生作文，題目爲：「最上一層樓」，400 字作文，其引導文如下：

> 有位富翁，一日到朋友家拜訪，那位朋友的住屋有三層樓高，雄爲壯麗，氣派非常。特別是走到第三層樓，由此俯瞰四方，感到視野廣闊，景致宜人。他在羨慕之餘，回到家就招來工匠，在自家的地上也要興建樓房，但是他命令工匠，不准先建造第一層及第二層，他要的只是最上面的一層樓。
>
> 看完這則故事，請以「最上一層樓」爲題，寫一篇文章，抒發你的看法和感想。

圖 4-1：作文「最上一層樓」之引導文

作文總分爲 20 分，最高分的作文爲 18 分，分數分佈情形如下表。

表 4-1：作文分數分佈情形

0~3 分	4~6 分	7~9 分	10~12 分	13~15 分	16~18 分
7 人	21 人	15 人	30 人	16 人	6 人

最後，我們以人工方式將其轉爲電子檔，並盡量呈現原始作文的所有文字（包含錯別字）。

另外我們也蒐集中國時報於 2005 年 10 月 31 日刊登的作文基測教室「我發明了一種藥」[1]，共 8 篇作文，滿分爲 6 級分。其中 6 級分者共 4 篇，5 級分有 2 篇，4 級分 1 篇與 3 級分 1 篇。

4.2 作文離題偵測

我們一一檢視「最上一層樓」95 篇作文的教師評語，發現其中共有 33 篇作文含有與離題相關的評語（離題過遠、審題未清、引喻失當、切題不足、與引文無關等等評語），可見離題是學生作文遇到的常見問題。

4.2.1 節我們以實驗模組Ⅲ標記作文主題詞後，以同義詞林作概念化標記，偵測作文離題句數的比例。4.2.2 節則是實驗的結果與分析。

4.2.1 離題偵測方法

由於引導文的字數少，提供的資訊並不足以供系統做離題偵測，一般還需要加上範本文章，範本文章可由專家（或者出題教師）來撰寫，在「最上一層樓」語料中，我們使用得分爲 16~18 分的 6 篇學生作文來做範本文章。

1. 引導文與範本文章經由 CKIP 系統斷詞標記
2. 以同義詞林的概念化標記引導文與範本文章，作爲命題概念範圍
3. 學生作文經由 CKIP 系統斷詞標記
4. 作文以 3.1 小節方法產生小句重心候選詞，並以 3.2 小節方法選取小句重心
5. 作文以 3.5 小節的實驗模組Ⅲ選取長句主題詞
6. 評量作文各長句主題詞是否於命題概念範圍內
7. 計算未在概念範圍內的長句分數，以此評量離題與否

圖 4-2：離題偵測步驟

我們以 CKIP 系統[中研院 中文斷詞系統] 將引導文與範本文章加入詞性標記，此即步驟 1；由於我們必須先找出作文的命題概念範圍，才能偵測作文的離題與否，因此在步驟 2 中，經由斷詞之後，我們以同義詞林標記引導文與範本

[1] http://www.hkhs.tnc.edu.tw/jjean/asp/message/showpaper.asp?messageid=1765&papernumber=1

文章的每一個詞彙，引導文與範本文章的所有詞彙都以同義詞林前三碼標示之，未在同義詞林之內者以原表面詞彙表示之。引導文加上範本文章共 2191 個詞彙，我們將這 2191 個詞彙視爲是此命題的概念範圍，其中未在同義詞林者有 237 個詞彙。

之後的步驟 3 中，我們將學生作文亦以 CKIP 系統加入詞性標記，並以步驟 4 與步驟 5 產生長句主題詞。由於實驗模組Ⅲ已經有同義詞林的概念標記，在步驟 6 中，我們便可將這些長句主題詞視爲作文敘述的主題概念，以此察看是否位於命題概念範圍內。最後在步驟 7 中，統計不在命題概念範圍內的離題長句得分，並以此評量文章是否離題。

長句的得分則是依長句所在段落的重要性而定，我們先將做爲範本的 6 篇作文以三段法區分之，第一段爲承接引文之段落，第二段爲主要論述之段落，第三段爲總結。段落所含長句數如下：

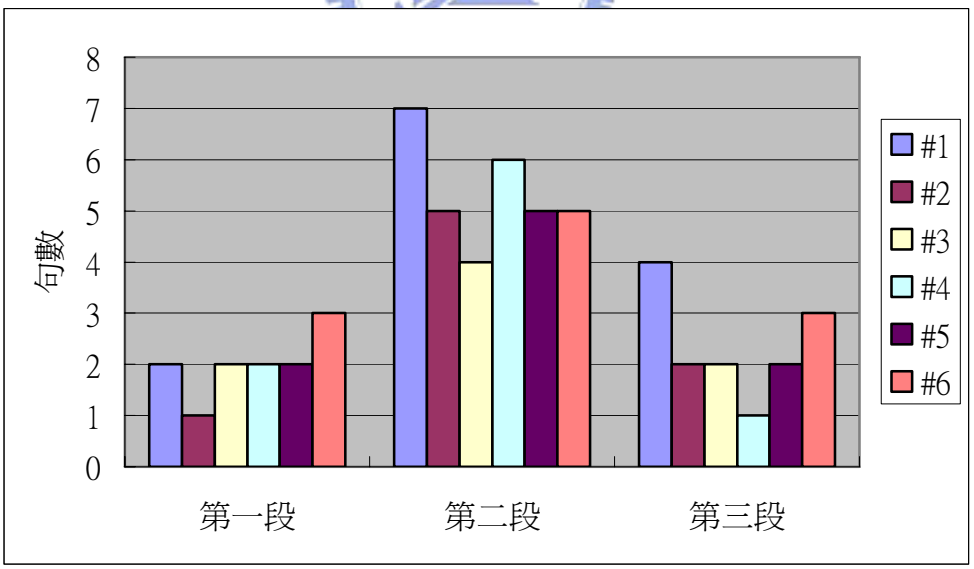


圖 4-3：範本文章之段落所含長句數分佈圖

平均第一段所含句數爲 2 長句，第二段則爲 5.33 長句，第三段爲 2.33 長句，平均每文章有 9.66 長句。我們依此切分段落，將前 2 長句視爲第一段，第 3~7 長句視爲第二段，第 8~9 句及其後視爲第三段。

切分段落後，需根據每個段落對於離題判斷的影響，給予其分數。我們由含有教師離題評語的 33 篇作文中抽取 6 篇離題文章，統計其離題句於段落中的位置，如下圖 4-4：

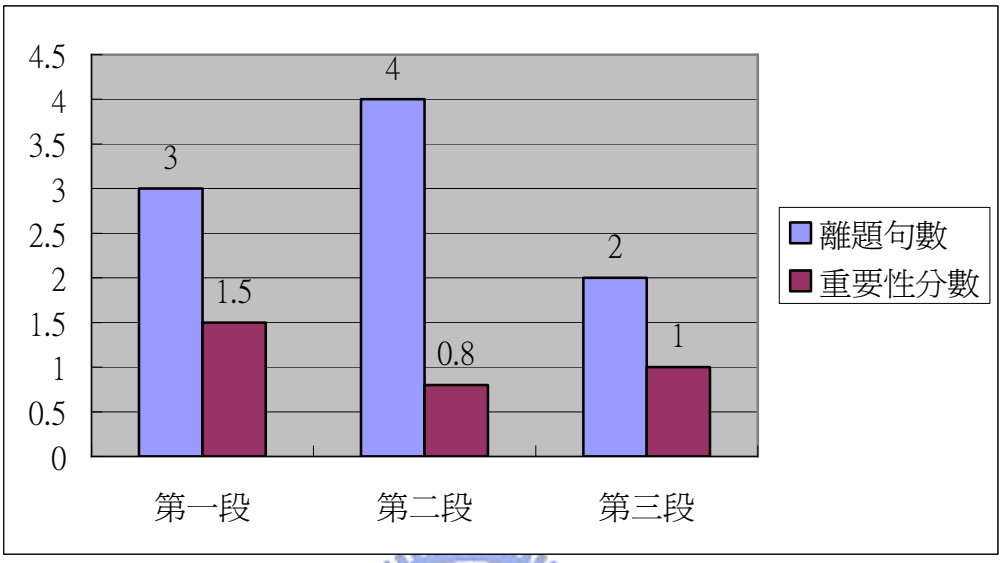


圖 4-4：離題長句分佈圖與重要性分數

6 篇離題文章共有 3 離題長句位在第一段，除以第一段的長句數 2，得到其得分爲 1.5。位於第二段者有 4 長句，除以第二段句數 5，得分爲 0.8。第三段者有 2 個長句，除以第三段句數 2，得分爲 1。

文章被判爲離題的長句將依據其位置計算得分，之後計算離題長句總分與文章長句總分之比例，判斷文章是否離題。如下所示：

$$\text{Off\_Score}(E_i) = \sum_{S \in off\_topic} Score(S)$$

If $\text{Off\_Score}(E_i)/\text{Total\_Score}(E_i) > T$, then E_i is off-topic

E_i爲第i篇作文，$\text{Off_Score}(E_i)$爲E_i的離題總分，S表示E_i的句子，我們統計所有離題的句子得分，加總之後得到E_i的離題總分。$\text{Total_Score}(E_i)$則是E_i所有句子得分的總和。當離題總分比例超過T，我們即判定其離題。

接著我們需決定 T 的值，我們觀察的 6 篇離題文章之全部長句總分爲 50.4，

而離題長句總分爲 9.7，約佔其 19.24%，我們將 T 設定得較小一些，設定爲 0.15。

4.2.2 離題偵測實驗結果與分析

除了 4.2.1 小節所描述的方法之外，我們另外參考[Tiun et al., '01]的作法，分別取概念出現於文章最多次的五、六、七、八、九個名詞(CKIP 標記爲 Na, Nb, Nc)作爲此篇文章的主題，以供實驗結果進行比較。只要其中有一個主題的概念不在命題概念範圍內，我們即視爲此篇文章有離題內容，結果爲以五個名詞當成主題最好，故此法中我們使用五個名詞來代表文章主題。

我們也另外使用文章全部詞彙來評量離題，將所有詞彙概念化之後，參考[蔡沛言, '05][林信宏, '06][粘志鵬, '06]的作法，將引導文與高分 6 篇範本文章的概念範圍視爲「好義原」，計算文章不是「好義原」的詞彙數量，分別以 5%、10%、15%爲門檻值，只要不是「好義原」的詞彙數量超過總詞彙數量 5%、10%、15%，我們即視爲此篇文章有離題內容，結果爲 10%最好，故我們取 10%爲門檻值。

我們將「最上一層樓」95 篇作文扣除觀察用的 6 篇離題作文與作爲範本的 6 篇高分作文之後，剩餘 83 篇作文分別以各個系統判定離題與否，並計算其正確率召回率，其計算公式如下：

正確率(Precision) = 含有教師離題評語篇數 / 系統偵測離題文章篇數

召回率(Recall) = 含有教師離題評語篇數 / 27 (所有含離題評語 33 篇-已用於觀察之 6 篇離題文章)

表 4-2 爲實驗比較結果，以最高頻的五個概念化名詞當成主題，最大的缺點便是抽取出的名詞可能都爲常見的名詞，因此其召回率比起其他二者都低，且僅以少數詞頻高的名詞視爲文章主題詞，只要其中一個不在命題概念範圍內就判定離題，誤判的風險也比長句主題詞方法來得高。

而以全部詞彙概念化來偵測離題，優點便是可以完整地偵測作者使用的所

有詞彙，因此其召回率比起以五個概念化名詞的方法來得高，但缺點便是偵測範圍太廣，以全部詞彙來偵測的結果，誤判的比例也比以五個概念化名詞的方法來得多，而且當文章有某段落離題，其餘大部分名詞都在命題概念範圍內，僅部分離題詞彙集中在此段落時，以全部詞彙的方法將無法偵測到這類的離題情形。

表 4-2：「最上一層樓」語料之離題評量結果

離題偵測方法	文章篇數	含有離題評語篇數	未含離題評語篇數	正確率	召回率	文章平均得分
長句主題評量(4.2.1 小節)	28	18	10	64.28%	66.66%	8.43
五個概念化名詞評量	14	8	6	57.14%	29.62%	8.07
非「好義原」之詞彙數量評量	22	10	12	45.45%	37.03%	8.50
整合型評量	33	21	12	63.36%	77.77%	8.24
所有文章	95	--	--	--	--	9.60
所有文章-範本文章	89	--	--	--	--	9.08
所有文章-範本-觀察用離題文章	83	--	--	--	--	9.28
含教師離題評語文章	33	--	--	--	--	6.76
用於觀察離題之 6 篇作文	6	--	--	--	--	6.16

而 4.2.1 小節的長句主題詞方法，不以整篇文章或者單一詞彙爲單位，而是取範圍適中的長句爲單位，不但可以偵測每個句子的離題與否來提高召回率，也計算離題句數比例，使部分合題長句不在概念範圍內時不致誤判。因此不管正確率或是召回率均比其餘二者來得較佳。

最後我們可將三個方法整合在一起，以 4.2.1 方法爲主，其餘二方法爲輔，文章經 4.2.1 方法判斷離題者，或其餘二方法皆判斷爲離題者，即視爲離題文章。如此比以 4.2.1 方法多了 3 篇離題的正確判斷，在犧牲很少正確率的情形下爲召回率提升了 11.11%。

接下來我們詳細分析 4.2.1 小節所描述方法偵測的離題文章 28 篇，以人工檢驗其結果，如下表所示：

表 4-3：長句主題詞離題偵測實驗結果

共抓出 28 篇作文爲離題				
含有離題教師評語 18	不含離題教師評語 10			
	錯別字過多 1	內容不完整、教師忽略 2	主題詞誤判 4	系統離題誤判 3

4.2.1 小節的系統判斷離題 28 篇作文中共有 18 篇含有離題的教師評語。其餘的 10 篇作文中，有 1 篇評語內有錯別字過多問題，導致主題詞因錯別字關係判定含有離題句；有 2 篇詳加檢視之後，發現文章較短、句數也較少，因此雖有離題語句，但教師也許忽略了，並沒有寫下離題評語，僅留下「內容不完整」的評語。主題詞誤判 4 篇中有 2 篇有離題語句，唯其比例並不會超過 0.15，但由於主題詞判斷失誤，使得離題分數超過 0.15；另 2 篇則無離題語句，但主題詞誤判使得系統也跟著誤判離題。

剩下 3 篇則是真正的系統誤判。我們也詳細檢視了這 3 篇誤判，發現學生是在舉事證時造成系統誤判，學生對於事證的描述相當詳細，因此涵蓋的長句數較多，但事證並不在我們的命題概念範圍內。

至於有 9 篇含有離題評語的作文卻未被系統判斷出來，我們分析如下表 4.4 所示。表 4-4 中，有 3 篇屬於一部份文章離題的情形，例如「第一、二段切要，但第三段末的結語便離題了。」這類的離題分數比例沒有超出 0.15，故沒有被系統判斷出來。

有 3 篇爲系統判斷長句主題詞時判斷錯誤，判斷成常見的名詞，例如「時候」，因此誤判爲概念範圍之內，沒有判斷成離題。而剩下的 3 篇離題作文則是學生認知概念上的誤解，並不是用字上的離題，以下爲例。

表 4-4：長句主題詞未偵測之離題作文分析

共 27 篇含有離題的教師評語			
系統已偵測 18	系統未偵測 9		
	文章一部份離題，但未超過 15% 3	長句主題詞判斷失誤 3	學生認知概念離題 3

例 1a：

> 富翁只蓋三層樓未免太小家子氣了，要蓋也蓋個一百層樓。

主題詞是「富翁」，在命題概念範圍內，但作文的命題重點並不是富翁只蓋三層不蓋一百層，而是富翁沒有第一層、第二層的支撐，怎麼可能蓋起第三層樓。像這類認知概念上的離題便無法被我們的系統所偵測出來。

例 1b：

> 文中的富翁真正想要的是高樓洋房氣派非常華麗的房子？其實富翁真正渴望的是那視野廣闊、景致宜人的風景啊！

主題詞一樣是「富翁」，但都沒有提及基礎的重要，反而誤解成富翁只想要風景而不是華麗的樓房。

而在「我發明了一種藥」的作文語料中，我們將 6 級分的 4 篇作文視爲範本文章，偵測剩餘的 4 篇作文，共偵測到 3 篇作文含有離題，分別爲 5 級分、4 級分與 3 級分，其中 5 級分文章爲誤判，原因在於舉例時誤判離題，且 5 級分文章的長句數過少所致。

4.3 連貫性評量

連貫性不好的作文特徵爲敘述的主題沒有一致、敘述沒有邏輯與條理，如講東之後講西，沒兩下子又回來講東，之後又扯到北，最後回去講西，讓人摸不

透文章的脈絡。

4.3.1 小節我們以文章的重心鏈來評量文章的連貫性，4.3.2 則爲連貫性評量的實驗結果。

4.3.1 連貫性評量方法

我們參考[Miltsakaki and Kukichy, ‘00]的方式，以文章的小句重心組成重心鏈來評量文章的連貫性，並藉著評量小句重心遞移情形，統計其「粗糙遞移」的數量，以了解文章的敘述條理是否跳換不當。步驟如下：

1. 學生作文經由 CKIP 系統斷詞標記
2. 作文以 3.1 小節方法產生小句重心候選詞，並以 3.2 小節方法選取小句重心
3. 將小句重心組成文章重心鏈
4. 以重心鏈評量重心遞移是否有「粗糙遞移」情形
5. 計算「粗糙遞移」次數，以此評量連貫性。

圖 4-3：連貫性評量步驟

我們先將學生作文以 CKIP 系統作詞性標記，如步驟 1；之後以 3.1 小節產生小句重心候選詞與 3.2 小節選取小句重心，如步驟 2。而爲了能夠觀察小句重心的遞移，我們將文章的小句重心組成重心鏈，如步驟 3，重心鏈的組成十分簡單，將文章的所有小句重心依照順序串連起來，便是此篇文章的重心鏈。而我們可以由重心鏈當中觀察重心的遞移情形，如步驟 4。最後計算「粗糙遞移」的次數，並以此來評量文章的連貫性，如步驟 5。

4.3.2 連貫性評量實驗結果與分析

我們統計所有「最上一層樓」95 文章的「粗糙遞移」次數，其結果如圖 4-4。粗糙遞移超過 3 次以上的文章共有 13 篇。

我們檢視這 13 篇作文，發現評語內提及敘述條理不佳者有 3 篇，其餘 10 篇的評語，離題者佔了 7 篇。剩下的 3 篇中，1 篇是結尾草率、用詞不佳，1 篇

爲錯字太多，1 篇是文中未能詳述道理與文章不夠精鍊。

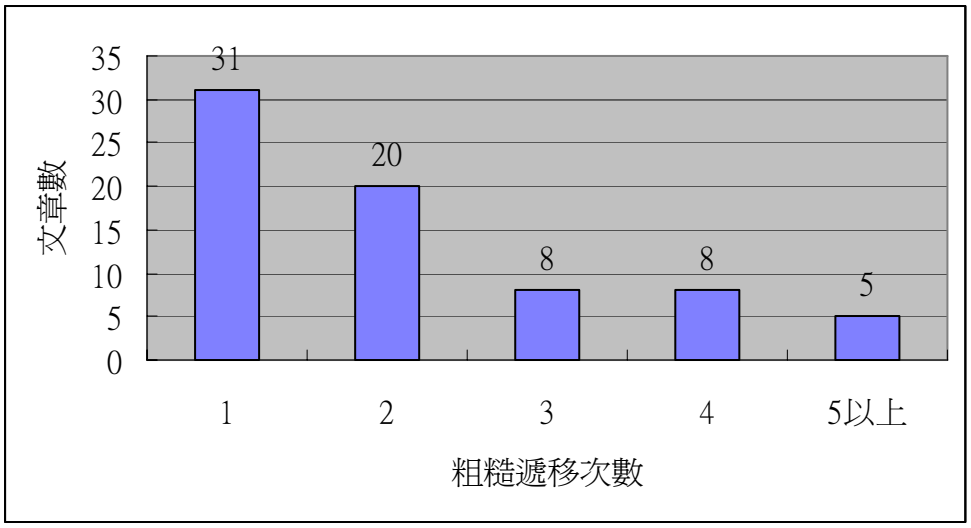


圖 4-4：「最上一層樓」粗糙遞移次數所含之文章數量

我們一一檢視這離題 7 篇，發現這些離題文章會在敘述中途就中間岔開，如此除了有離題問題之外，也因此形成敘述條理中斷的連貫性不佳問題，縱使教師評語僅指出離題情形，但這些離題文本身就有敘述條理被中斷的問題存在。

另外在非離題的 3 篇中，2 篇作文的中段有口語化以及斷句不佳的情形，使得小句過於短促，各小句重心候選詞只有一個，造成重心判斷失誤，其教師評語內亦有「用詞遣詞須待琢磨」、「多引用名言，使文章精鍊」等評語，可見這 2 篇的句子略顯瑣碎；而剩下的 1 篇則是錯字太多，使得 CKIP 斷詞標記時造成錯誤，讓沒有重心候選詞的小句有了錯誤且唯一的候選詞，造成重心判斷失誤，因此增加了「粗糙遞移」的次數。

下圖 4-5 爲被系統判斷爲連貫性欠佳的文章，注意其第二段連貫性欠佳的部分。

這則故事要說的是，小時候，父母沒把你教好，養成智能不足的小孩，以後出社會再厲害也不過是個笑點。還有一啓示是說，大部分智能不足的人，都只會看到別人成功後的樣子，然後虛榮心開張，做出一些沒腦袋的舉動。此富翁只看到第三層的好，就嚷著最上一層，如果我是工匠，早就把他抓去埋了。

沒有白來的**東西**，必是一番努力，從基本建立，致成功的**樓層**，哪有**直街**上去的？不是「疊」起來的不是「**高度**」，沒有**踏腳石**哪裡站得穩？是不會無故有那最高一層**樓**的。
而且**富翁**也不知道，**工匠**很苦的，蓋房子搞設計，要很費神耶！怎麼可以爲了「最上面」而如此捉弄呢？他也是有**家室**有父母的，太過份了！送去**人本**輔導。

或許那第三層樓的風景會迷魅人好了，或許抵抗不了好了，但是，只想蓋第三層野心也太小了吧？這不是變得跟凡人一樣，要蓋嘛蓋到大氣層去，不然叫什麼富翁，況且，最重要的重點是，除非富翁跟朋友住同一個社區，不然**風景**有可能同樣宜人嗎？同樣**視野**遼闊嗎？喔**實尬**，真是太無言了，如果附近是垃圾山，你怎麼蓋都是給人笑，給人笑啦！

圖 4-5：被系統判斷爲連貫性欠佳的文章

而在「我發明了一種藥」的 8 篇文章中，有 1 篇 6 級分文章粗糙遞移爲 4 次，其原因在於此文部分小句過短，因而誤判。

4.4 文章概念結構

長句主題詞除了可作離題偵測外，還可勾勒出文章的結構。我們可利用辨識出的長句主題詞與其概念，畫出文章敘述的結構圖，如下圖 4-6 所示。

長句主題詞：總統→美國→方向→美國→美國→創投→人才→美國→人才

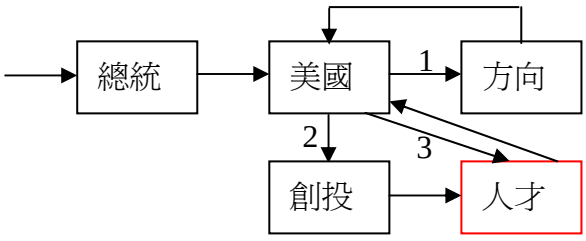


圖 4-6： 社論文章之結構圖

圖 4-6 的社論由美國布希「總統」提出國情咨文開始，敘述「美國」對經濟持續開放的立場，包含對於租稅「方向」朝鼓勵創投之優惠減免措施，故「美國」才能維持蓬勃的「創投」活動。但隨著全球科技「人才」的短缺，「美國」勢必也將會受其限制，故必須提早因應人才不足的問題。我們的科技研發雖已漸受重視，但「人才」的培養則仍待各個政府部門提早整合規劃。

根據圖 4-6，社論文章的結構上有許多重複的主題，例如「美國」與「人才」就在結構圖中重複到訪兩次以上，這顯示社論的文章結構較爲嚴謹。而在學生作文中，高分作文的結構理應要比低分作文嚴謹，我們從「最上一層樓」抽取數篇高分與低分作文來觀察其結構。

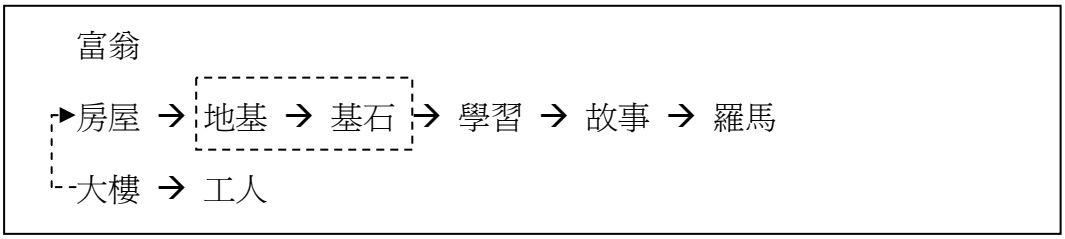


圖 4-7： 18 分作文 A 之結構圖

如圖 4-7 顯示 A 的第一段由「富翁」破題，第二段敘述「房屋」的「地基」與「基石」重要性（「地基」與「基石」概念相同），對應「學習」也是需要基礎的，「故事」就是想告訴我們「羅馬不是一天造成的」，之後第三段回到「大樓」也不是一天就可以蓋好，建築「工人」們沒有基石不可能蓋成，唯有一步步腳踏實地才能達成。其中「大樓」與「房屋」概念相同。

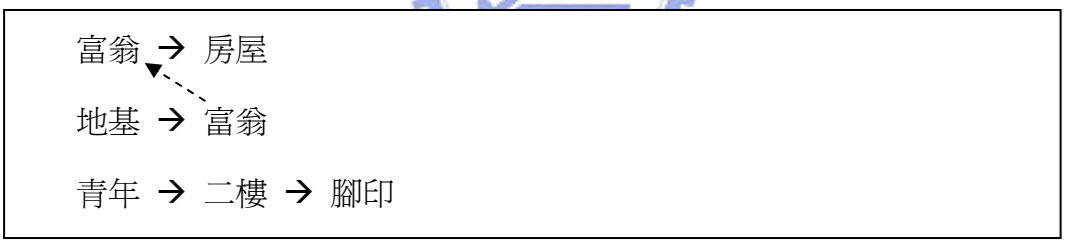


圖 4-8： 17 分作文 B 之結構圖

圖 4-8 顯示作文 B 以「富翁」破題，並講述「房屋」與「地基」的關係，之後回到「富翁」沒有地基不可能蓋成樓房，末段呼籲現代「青年」若想築高樓，絕不可忘記「一、二樓」的重要，要「一步一腳印」地紮實基礎。

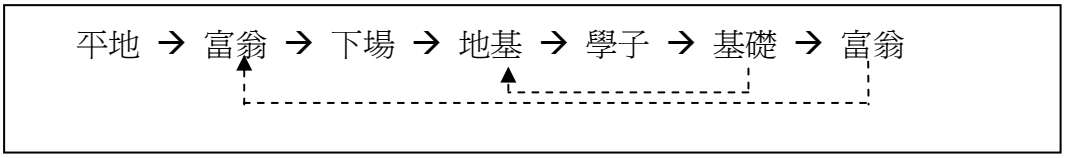


圖 4-9： 17 分作文 C 之結構圖

圖 4-9 中，作文 C 以「萬丈高樓平地起」破題，敘述「富翁」與其「下場」

肯定是高樓塌陷，因爲忘記了「地基」的重要。最後呼籲「學子」，一定要從「基礎」做起，不可如「富翁」想一步登天。

以下我們舉三個低分文章的例子：

長句主題詞：爸→地方→景色→天空→時候→時候

爸 → 地方

景色 → 天空

時候 → 時候

圖 4-10： 2 分作文 D 之結構圖

圖 4-10 中，作文 D 開頭是「爸」養了一隻畫眉鳥，之後與爸回到鄉下「地方」，看見美麗的「景色」與「天空」，以及描述白天與夜晚「時候」的鄉間風景。

意義

叢林

工程 → 讀數字

圖 4-11： 5 分作文 E 之結構圖

圖 4-11 中，作文 E 開頭敘述萬丈高樓平地起的「意義」，並以自己生活在「都市叢林」中，許多的建築「工程」都必須先做好基礎，最後以「讀數字」爲例，必須要先看懂數字，才能作四則運算。

本能

米粒

人

圖 4-12： 3 分作文 F 之結構圖

圖 4-12 中，作文 F 開頭以我們人類的「本能」是要腳踏地上才有安全感，並舉自己在高樓的經驗，看見如「米粒」般的地上景物，便恐懼起來，所以「人」

不能築太高的高樓，否則恐懼便會揮之不去。

我們可以發現高分文章與低分文章的差異，便在於高分文章的結構較爲嚴謹，會重複敘述到重要的主題，以達到前後呼應的效果。而低分文章除了內容乏善可陳之外，也傾向以直敘法來敘述，主題之間沒有重複性，結構較平淡無奇。

我們統計作文結構含有與未含重複到訪主題，並將含重複到訪主題者依照重複到訪主題數與到訪距離細分，例如圖 4-7 中，重複到訪主題數爲 1（僅「房子、大樓」被重複到訪），其到訪距離爲 4（「羅馬」至「房子、大樓」中間相距 4 長句）；而未含重複到訪主題者則依照有無連續主題細分，例如圖 4-10 雖無重複到訪主題，但其「時候」爲連續 2 長句之主題。表 4-5 中，我們可以發現未含重複到訪主題的作文平均分數(8.96 分)，比含有重複到訪主題的平均分數(10.37 分)要低，且 16~18 分的作文 6 篇中 5 篇都含有重複到訪主題，0~3 分的作文則大部分都採平鋪直敘的結構，僅有 2 篇有重複到訪主題。

表 4-5：「最上一層樓」未/含重複到訪主題之比較

文章分數	含重複到訪主題	含重複到訪主題之細分		未含重複到訪主題	未含重複到訪主題之細分	
		到訪主題數=1 (距離<=2 /距離>2)	到訪主題數>1		含連續主題	未含連續主題
0~3 分	2	2 (1/1)	0	5	0	5
4~6 分	8	7 (5/2)	1	13	2	11
7~9 分	5	5 (4/1)	0	10	1	9
10~12 分	17	14 (7/7)	3	13	2	11
13~15 分	6	4 (2/2)	2	10	1	9
16~18 分	5	4 (2/2)	1	1	1	0
文章總數	43	36 (21/15)	7	52	7	45
文章平均分數	10.37	10.11(9.52/10.93)	11.71	8.96	10.29	8.76

進一步分析則可發現，含重複到訪主題數量大於 1 者，表示其結構更爲嚴謹，故其平均分數最高(11.71 分)。而重複到訪主題數等於 1 者，其距離若大於 2，

表示描述主要主題之後，間隔 2 長句以上才回到原本主要主題，這中間的間隔可以是引名言、例證等強調主題，可以是多方取材而造成重複到訪主題的距離增加，故其平均分數(10.93 分)明顯比距離小於 2 的平均分數(9.52 分)要高。

而在未含重複到訪主題者，若含連續主題，其平均分數(10.29 分)明顯比未含連續主題者的平均分數(8.76 分)高，且也比含重複到訪主題但距離小於 2 者(平均分數 9.52 分)稍高一些，表示含連續主題比未含連續主題者的結構嚴謹，且與其說沒兩句話就重回重要主題，不如對重要主題繼續加以強調來得較佳。

在「我發明了一種藥」的語料中，8 篇作文中共 3 篇含有多個重複到訪主題，皆爲 6 級分。而含單一重複到訪主題者有 3 篇，分別爲 6 級分（距離=1）、4 級分（距離=2）與 3 級分（距離=1）。未含重複到訪主題但含連續主題者有 2 篇，皆爲 5 級分。除了到訪距離<=2 中有一篇 6 級分之外，大致上符合我們的假設：多重複到訪主題 > 單重複到訪主題(距離>2) = 連續主題 > 單重複到訪主題(距離<=2) > 無到訪主題且無連續主題。

除了觀察有無重複到訪主題之外，我們還可觀察相鄰主題之間的推導關係。我們將相鄰的三個主題組成三組主題推導，例如主題串 A→B→C 可組成 A→B、B→C 、A→C 三組推導。之後我們統計主題推導的出現次數，如下表 4-6 所示。

表 4-6：「最上一層樓」主題推導之統計次數

全部作文		高分作文 (16 分以上)		低分作文 (3 分以下)	
主題配對	出現次數	主題配對	出現次數	主題配對	出現次數
樓房→地基	11	樓房→地基	4	富翁→身心	2
樓房→人	8	樓房→工匠	2	故事→慾望	2
人→富翁	7	樓房→道理	2	凡事→背後	2
富翁→人	6	事→樓房	2	富翁→美景	1
富翁→樓房	6	下場→地基	2	富翁→職業	1

我們可以發現主題「樓房」後常接另一個主題「地基」，因引導文之本意便

是要學生瞭解建築「樓房」最需要的是底層厚實的「地基」。在全部作文中，「樓房→地基」出現了 11 次，其中 4 次出現在高分作文，而低分作文則完全沒有「樓房→地基」的推導存在。

而在「我發明了一種藥」的語料中，由於作文並未強調要發明何種藥，可以任憑學生自由發揮，未如「最上一層樓」有明確的目的：要學生理解「樓房」與「地基」之間的關係，且語料亦僅有 8 篇作文，因此在表 4-7 中並未產生較有實質意義的主題推導出現。

表 4-7：「我發明了一種藥」主題推導之統計次數

主題配對	出現次數
事情→藥	3
地球→藥	2
藥→病人	2
藥→瘡	2
藥→大地	2

第五章 結論

本論文提出並製作了一個二階段式的主題辨識系統，並且實際將長句主題應用於學生作文之離題偵測與文章結構分析上，將小句重心運用於連貫性評量上。經實驗數據的分析顯示，能正確的辨識出長句主題與小句之重心。本論文從設計、研究方法至實驗完成，可以歸納出幾個主要的成果與貢獻：

1. 針對語料中的主題詞進行研究與觀察，使以後的研究者可以更了解主題詞在實際語料中的特徵，有助於後續研究的進行。
2. 我們所提出的主題辨識特徵，的確可以幫助我們抽取出長句的主題詞。
3. 不需仰賴訓練語料即可實作長句主題詞之辨識。
4. 完成作文離題之偵測與連貫性評量、文章結構分析，希能後續加入中文作文自動批改系統，以評量學生作文寫作能力。

本論文的後續研究有下列幾個方向：

1. 主題特徵之研究

   可以找尋更多的主題特徵，來協助系統辨識主題。

2. 其他語料類型的研究

   可進行其他語料類型之研究，以研究系統在其他類型語料上的效能。

3. 納入自動作文評分系統

   將主題分析納入自動作文評分系統之一，作爲作文合題性、文章結構分析與連貫性評量的子系統。

參考文獻

Burstein, J., Leacock, C., and Swartz, R. (2001). “Automated evaluation of essays and short answers.” *Fifth International Computer Assisted Assessment Conference*, Loughborough University, Loughborough, UK.

Chang, Jeong-Ho., Lee, Jae Won., Kim, Yuseop., and Zhang, Byoung-Tak. (2002). “Topic extraction from text documents using multiple-cause networks.” *Lecture Notes In Computer Science Vol. 2417: Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence*, PRICAI 2002, Tokyo, Japan.

Grosz, B., Weinstein, S., and Joshi, A. (1995). “Centering: a framework for modeling the local coherence of discourse.” *Computational Linguistics vol.21*(2), 203-225.

Higgins, D., Burstein, J., Macru, D., Gentile, Claudia. (2004). “Evaluating Multiple Aspects of Coherence in Student Essays.” *Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL 2004)*, Boston, Massachusetts.

Khoo Khyou Bun., and Ishizuka, M. (2002). “Topic extraction from news archive using TF*PDF algorithm.” *Web Information Systems Engineering(WISE)*, Singapore.

Leacock, C., and Chodorow, M. (2003). “C-rater: Automated Scoring of Short-Answer Questions.” *Computers and the Humanities vol.37*(4), 389–405.

Leake, D., Maguitman A., Reichherzer T. (2003). “Topic Extraction and Extension to Support Concept Mapping.” *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2003)*, Florida.

Li, Charles N. and Thompson, Sandra A., 1981, *Mandarin Chinese – A Functional Reference Grammar*, University of California Press.

Lin, Chin-Yew. (1998) “Assembly of Topic Extraction Modules in SUMMARIST.” *The Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Madison, Wisconsin.

Lin, Sung-Chen. (2004) “Topic Extraction Based on Techniques of Term Extraction and Term Clustering.” *Computational Linguistics and Chinese Language Processing vol.9*(3), 97-112.

Makkonen, J., Ahonen-Myka, H., and Salmenkivi, M. (2004). “Simple Semantics in

Topic Detection and Tracking.” *Information retrieval vol.7*(3), 347-368

Miltsakaki, E, and Kukichy, K. (2000). “Automated Evaluation of Coherence in Student Essays.” *In Proceedings of LREC 2000 Workshop: Language Resources and Tools in Educational Applications*, Athens, Greece.

Miltsakaki, E, and Kukichy, K. (2004). “Evaluation of text coherence for electronic essay scoring systems. ” *Natural Language Engineering Vol.10*(1), 25-55.

Mitchell, T., Russell, T., Broomhead ,P., and Aldridge N. (2002). “Towards robust computerised marking of free-text responses.” *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughboroug University, Loughborouh, UK.

Tiun, S., Abdullah, R., and Kong, T. (2001). “Automatic Topic Identification Using Ontology Hierarchy.” *Lecture Notes in Computer Science : Computational Linguistics and Intelligent Text Processing : Second International Conference*, CICLing 2001. Mexico City, Mexico.

Yeh, Ching-Long and Chen, Yi-Chun. (2004) “Creation of Topic Map by Identifying Topic Chain.” *Proceedings of the 2004 ACM symposium on Document engineering*, Milwaukee, Wisconsin.

曹逢甫 (1995) “主題在漢語中的功能研究 － 邁向語段分析的第一步” 北京語文出版社

梅家駒, 竺一鳴, 高蘊琦, 殷鴻翔 (1997). “同義詞詞林” 東華書局

韓客松, 王永成, 沈洲, 吳芳芳 (2000). “三個層面的中文文本主題自動提取研究” *中文信息學報 Journal of Chinese Information Processing Vol.15*(4), 20-27

王信智 (2000). “應用資訊檢索技術於科學寫作作品評量之探究” 國立台南師範學院 資訊教育研究所 碩士論文

徐爲群, 徐波, 黃泰翼 (2004). “口語對話中的語句主題分析” *中文信息學報 Journal of Chinese Information Processing Vol.19*(4), 89-96

蔡沛言 (2005). “自動建構中文作文評分系統：產生、篩選與評估” 國立交通大學資訊科學系 碩士論文.

林信宏 (2006). “基於貝氏機器學習法之中文自動作文評分系統” 國立交通大學資訊科學與工程研究所 碩士論文.

粘志鵬 (2006). “基於支援向量機之中文自動作文評分系統” 國立交通大學資訊科學與工程研究所 碩士論文.

馬偉雲 (2006). “中文動詞名物化判斷的統計式模型設計” *ROCLING XVIII: Conference on Computational Linguistics and Speech Processing (ROCLING 2006)*, Hsinchu, Taiwan.

鄭守益 (2006). “以語料爲基礎的中文語篇連貫關係自動標記” *ROCLING XVIII: Conference on Computational Linguistics and Speech Processing (ROCLING 2006)*, Hsinchu, Taiwan.

中研院 中文斷詞系統 http://ckipsvr.iis.sinica.edu.tw/