中文文本中長句的主題詞辨識與其應用

梁婷
國立交通大學資訊工程與科學系
tliang@cis.nctu.edu.tw

潘善均
國立交通大學資訊工程與科學系
panshannjiun.cs94g@nctu.edu.tw

陳冠熙
國立交通大學資訊工程與科學系
cyberetro@gmail.com

***摘要*—主題詞辨識可釐清文本的核心敘述，是文本理解中的一項重要工作。本論文提出一個中文文本語句主題詞之辨識法。此法乃利用重心理論並考量中文語句結構的特性，使之可應用於作文評分中的離題偵測、主題連貫性和語義概念結構分析。我們以長句中每小句重心為基礎，分別就其頻率、位置、主題延伸性、前後句之一致性、主題概念等特徵，進行權重設計以辨識每一完整長句的主題詞。從實驗結果顯示此法在長句主題詞的辨識上，在 9 篇報紙社論可達 86.84%正確率，在 22 篇高中生作文可達 80.86%正確率。另在 83 篇高中生作文的離題偵測實驗上，所提的方法也較前人所用的詞彙方法得到較佳的偵測，辨識效果可達 63%正確率。**

***關鍵詞*—中文文本、主題詞辨識、離題偵測、主題特徵、重心理論**

1.緒論

主題詞辨識是文本理解中一項不可缺乏的工作，可為訊息的中心陳述。以往主題詞辨識多應用於文件分類問題上，從單篇或多篇文章中以統計方法抽取關鍵詞做為文件內容的主題描述。如以一袋子詞與權重計算，抽取重要的詞彙當成主題詞 [9] [10]。另有將人、事、時、地、物等當成向量元素，以「事件」來呈現主題[11]。其它有[7]所提出的重心理論(Centering Theory)，從連續兩個英文句子間，判斷句子的重心，抽取主題。[16]亦將重心理論利用於中文文本中的零指代消解。然而前述所提方法並不考慮文章中句間的邏輯結構是否呈現主題敘述的連慣性和段落間主題論述首尾呼應的一致性，是以無法直接應用於作文評分的自動化。

眾所皆知，作文能力是評量語文學習的一項重要指標，也是邏輯推理、記憶與組織、創造等能力的顯示。因此有效的作文評分自動化或作文分析工具的技術開發日顯重要。一篇作文是否離題、主題論述是否連貫是作文評分的重要參考指標。目前常見的離題偵測是以計算文章全部詞彙出現在高低分文章的比例來評量，如英文作文評分系統 E-Rater[6]，或中文作文評分系統 [1][3]。至於主題論述連貫性則是檢驗作者邏輯條理是否正確的重要指標之一。[13]以人工方式標記各語句重心，將重心串成一串主題鏈，並以重心理論來評價文章的連貫性。 [8]先標明句子的角色，使用SVM(Support Vector Machine)來標注每一句對題目(Prompt)、主題句(Thesis)、段落(Segment)的關連、以及句子有無錯別字等來評量文章的連貫性。

在本論文中，我們針對可應用於偵測中文文章主題敘述連慣性和離題性，提出以中文小句重心為基礎的長句主題詞的辨識法。此乃是基於在處理中文文章結構上，我們視每一長句具有一完整的主題論述。中文長句定義為以「。！？」為長句結束符號。一個中文長句句法結構上可包含數個小句，因此在進行辨識長句主題詞的辨識時，我們考量小句重心在文章中的頻率、位置、延伸性、一致性、概念等特徵。此種長句主題詞辨識法，不僅可偵測文章段落上的主題，也可應用於主題概念的推演結構分析。

2.二階段式的長句主題詞辨識法

我們採用二階段方式挑選長句的主題詞，第一階段依據重心理論來選取長句中各小句的重心，第二階段則由這些小句重心來決定一個長句的主題詞，這種由小句重心選取主題的方法可以避免較長的句子有太多的主題候選詞。

2.1 小句重心辨識

2.1.1重心候選詞萃取

小句重心辨識是基於重心候選詞的萃取。文章經 CKIP[1]標記系統，並依標點符號「，。！？；」切分成小句。而後將標記為普通名詞(Na)、專有名詞(Nb)、地方詞(Nc) 的詞彙，以及第一、第二人稱代名詞作為候選詞。若遇連續的名詞組，則取最後的名詞作為候選詞。由於中文常有出現名物化動詞的情形，故我們參考[2]，將部分的名物化動詞列入候選詞，如「是」前面的動詞、「的」後面的動詞，及物動詞(VC)後面的動詞。我們也參考[15]，針對中文常出現的零指代與第三人稱代名詞進行辨識。

2.1.2重心候選詞選取

我們參考[7]的重心模型來進行小句重心選取，並依照中文的特性加以改進。重心模型如表 1 所示，其中 C_i 代表第 i 小句重心，C_{i-1} 則是 C_i 前一小句的重心，C_{i-2} 則是 C_{i-1} 前一小句的重心，C 是 C_i 的候選詞。$Can(C_i)$表示 C_i 的重心候選詞集合，C(3rd-anaphor)代表第三人稱代名詞，C(Zero-anaphor)代表零指代。

表 1：重心模型

Case 1	延續：$C_{i-1}=C_{i-2}$ 且
1.1	$\exists C \in Can(C_i)$, $C=C_{i-1}$，則 $C_i=C=C_{i-1}$
1.2	(3rd-anaphor)$\in Can(C_i)$ OR C(Zero-anaphor)$\in Can(C_i)$，則 $C_i=C_{i-1}$。
Case 2	保留： $C_{i-1}=C_{i-2}$, 且
2.1	$\forall C \in Can(C_i)$, $C \neq C_{i-1}$ 且 C(3rd-anaphor)$\notin Can(C_i)$ AND C(Zero-anaphor)$\notin Can(C_i)$，則 $C_i=Can(C_i)$。
2.2	前一小句沒有重心(標示為 E)，或者此小句為長句中的第一小句，則 $C_i=Can(C_i)$。
Case 3	平順遞移：$C_{i-1} \neq C_{i-2}$, 且
3.1	$\exists C_k \in Can(C_{i-1})$, $\exists C_j \in Can(C_i)$, $C_k = C_j$，則 $C_i=C_{i-1}=C_k=C_j$。
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($Can(C_{i-1})$)。
3.3	$\forall C_k \in Can(C_{i-1})$, $\forall C_j \in Can(C_i)$, $C_k \neq C_j$，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 k 個詞彙 $T_{i...}T_k$，於此文章中出現之頻率，並取其頻率最高者。若為第一、第二人稱代名詞，其頻率視為 0。
Case 4	粗糙遞移 $C_{i-1} \neq C_{i-2}$, 且 $Can(C_{i-1})=\{C_1\}$, $Can(C_i)=\{C_2\}$，$C_1 \neq C_2$，則 $C_i=C_2$。

[1] CKIP: http://ckipsvr.iis.sinica.edu.tw/

在 Case 4 中，我們將「粗糙遞移」的狀況限制的很嚴謹，並將原本應視為「粗糙遞移」的狀況視為是「平順遞移」(Case 3.3)。這是因為若 C_{i-1} 已為「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。「粗糙遞移」在文章當中原本就較少出現，即使判定為「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。

2.1.3小句重心實驗

我們蒐集了 11 篇內容為政治和經濟議題的聯合報社論。11 篇社論共包含 15404 字數，共 276 長句數，平均每一長句有 4.65 小句。在計算重心選取正確率時，我們排除無重心候選詞的 128 小句，僅人工檢視 1156 小句的重心選取結果，得到 83.70%正確率。

2.2 長句主題詞辨識

2.2.1長句主題詞特徵

對任一長句中每小句重心，我們分別就其頻率、位置、主題延伸性、前後句之一致性、主題概念等特徵，進行重心 C_i 在長句 S_k 中的重要性權重 $Weight(S_k,C_i)$ 計算如公式(1)，選取最高權重值的重心詞做為長句主題詞。

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

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

其中 $Can(S_k)$表示 S_k 的主題候選詞集合；特徵符號說明見表 2。

頻率特徵是考量一個長句的主題詞應為中心陳述，故小句重心頻率可作為主題識別的依據。位置特徵是考量主題詞常在長句的第一個小句出現。一致性特徵是考量長句的主題往往與前一長句呼應。而延伸特徵是考量當長句的第一小句含有正向連接詞時，其涵義往往是承襲上一長句的論述。故我們使用[5]所列舉的正向連接詞，遇第一小句含連接詞時將上一長句之主題候選詞也納入考慮，做為延伸特徵。至於概念特徵是考量主題詞可以同義詞來表示相同的概念，以避免過度使用某個詞彙。我們以同義詞詞林[4]來標記小句重心的概念。

表 2：長句主題詞特徵表

特徵名稱	符號	使用時的值
頻率	$\mathrm{CenterFreq}(S_k,C_i)$	C_i 在 S_k 中的重心數
位置	$\mathrm{P}(S_k,C_i)$	C_i 在第一小句為 2，否為 0
一致性	α	C_i 與前句主題詞一致則為 2，否為 1
延伸	β	第一小句含正向連接詞為 0.5，否為 0
概念化	$\mathrm{CenterFreq}(S_k,C_i)$	概念化後之重心數

2.2.2長句主題詞實驗

長句主題詞實驗使用的語料有兩組，第一組是與 2.1.2 小句重心實驗相同的社論語料。我們使用其中 2 篇來作為開發語料，剩餘 9 篇作為測試語料，社論測試語料共 228 長句，以及 289 個包括同義詞的主題詞。第二組是 22 篇高中學生的作文語料，共 188 個長句(詳細數據將在第三段敘述)。

在本論文中我們對每一長句僅挑出一個主題詞，系統以辨識正確率進行評估，其計算公式如下：

正確率 = 系統正確辨識主題詞的長句數/總長句數

為比較主題詞辨識成效，我們設計五種實驗模組，分別以重心理論和一般詞彙方法來做比較，如表 3。其中模組 I 為基本模組僅考慮頻率、位置特徵，模組增加句間的一致性和延伸特徵，模組 III 則是再增加考慮主題詞的概念化特徵。

模組 IV 是以長句中出現頻率最高之小句重心候選詞，做為長句的主題詞。模組 V 是以長句中出現頻率最高之名詞或動詞為主題詞。實驗結果證實所提的以小句重心理論模組確實較辭彙模組得到較高的辨識。此外，所用的五種特徵都是正向影響因子，因此我們以模組 III 做為後續應用於作文評量的工具。在第二組的 22 篇作文測試語料上，對 188 個長句進行主題詞辨識，得到 80.86%的正確率。

3.主題辨識應用

3.1 作文語料分析

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。在本論文中，學生作文語料為新竹市一所市立高中國文考試的 95 篇學生作文；題目為「最上一層樓」，約 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，雄為壯麗，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」為題，寫一篇文章，抒發你的看法和感想。

圖 2：作文「最上一層樓」之引導文

95 篇學生作文作文分數分佈情形如表 4。表 5 是我們以實驗模組Ⅲ對作文語料萃取主題詞之實驗結果，並將此應用到作文離題和連貫性偵測。

表 3：社論語料之長句主題詞辨識實驗結果

	模組	正確數	錯誤數	正確率
重心理論	模組 I	188	40	82.46%
	模組 II	194	34	85.09%
	模組 III	198	30	86.84%
詞彙	模組 IV	101	127	44.29%
	模組 V	89	139	39.04%

表 4：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	21	15	30	16	6

3.2 離題偵測

我們以長句主題詞來偵測離題，並藉由作文的引導文與範本文章來判定文章的長句是否有離題。在「最上一層樓」語料中，我們抽取 6

篇高分文章做為範本與 6 篇離題文章作為系統開發語料。萃取主題詞並以同義詞林擴充，視為合題與離題詞庫範圍。對每一篇作文 E_i 我們計算每一長句的分數值 Score(S_i)。長句的分數是考量主題詞所在的位置 P(S_i)與主題詞是否在離題詞庫或合題詞庫來計算 M (S_i)。公式(3)(4)分別定義每篇文章的離題分數 Off_Score(E_i)和總分分數 Total_Score(E_i)。若 Off_Score(E_i)/Total_Score(E_i)大於門檻值(在本論文中，門檻值設定為 0.15)，即判別為離題文章。

$$\text{Off\_Score}(E_i) = \sum_{S \in off\_topic} Score(S) \qquad (3)$$

$$\text{Total\_Score}(E_i) = \sum_{S \in Ei} Score(S) \qquad (4)$$

$$\text{Score}(S_i) = P(S_i) \times M(S_i) \qquad (5)$$

其中

$$P(S_i) = \begin{cases} 1.5 & \text{if } i <= 2 \\ 0.8 & \text{if } 3 <= i <= 7 \\ 1 & \text{if } i >= 8 \end{cases}$$

$$M(S_i) = \begin{cases} 1 & \text{if } S_i \text{ 的主題詞是在離題詞庫} \\ 1.2 & \text{if } S_i \text{ 的主題詞是在合題詞庫} \\ 0.9 & \text{其他} \end{cases}$$

在離題實驗中我們以其他的 83 篇「最上一層樓」文章作為測試語料，其中 27 篇含有離題評語。離題實驗結果如表 5 所示。其中，除了長句主題詞的方法外，我們另外參考[14]的作法，抽取文章中頻率最高的五個概念化名詞為文章主題詞。文章若不含有名詞在此集合內，則判定為離題。另外我們也參考[3]等人的作法，以文章全部詞彙的概念來判定離題，只要離題詞彙的比例超過一定的值，即視為離題。

以五個概念化名詞當成主題詞[14]，缺點是抽取出的名詞可能都為常見名詞，因此其召回率比起其他二者都低。而以全部詞彙來偵測離題，其優點是可完整地偵測作者使用的所有詞彙，缺點便是偵測範圍太廣，且當大部分名詞都在命題概念範圍內，僅部分離題詞彙集中在某段落時，以全部詞彙的方法將無法偵測到這類的離題情形。相對的以所提的長句主題詞來判斷離題，便可解決這些問題，故其正確率與召回率較其他二者高。

檢驗以長句主題詞為判斷離題依據，系統判斷有 27 篇作文，其中 10 篇是誤判。誤判的原因包括 1 篇是錯別字過多； 2 篇是文章較短且句數較少，雖有離題語句，但教師並沒有寫下離題評語，僅留下「內容不完整」的評語。另外有 4 篇因主題詞判斷失誤使其誤判；另 3 篇則是因學生在舉事例時描述詳細，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，因此誤判。致於 10 篇含有離題評語卻未被系統判斷出來的原因有 4 篇是屬於一部份文章離題的情形但離題比例沒有超出 0.15。另外由於訓練語料較少故離題詞彙量少，若不在範圍內將不會被辨識出。

3.3 連貫性評量

我們參考[12]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。

在「最上一層樓」作文中，粗糙遞移超過 3 次以上的文章共有 13 篇，其中評語內提及敘述條理不佳者有 3 篇。其他 10 篇的評語，離題者佔 7 篇，而這些離題文是中途岔開離題，故的確有連貫性問題。剩下的 3 篇中，1 篇是結尾草率、用詞不佳，1 篇為錯字太多，1 篇是文中未能詳述道理與文章不夠精鍊。

3.4 文章概念結構

我們利用長句主題詞與其同義詞概念，進行文章敘述結構的分析。圖 3a 與圖 3b 分別為一篇高分與低分文章的主題詞詞串。我們可以

發現高分文章的結構較為嚴謹，會重複敘述到重要的主題詞，以達到前後呼應的效果；而低分文章傾向以直敘法來敘述，主題詞之間沒有重複性，關連性低。

房子→地基→基石→學習→故事→羅馬→大樓(與房子同概念)→工人

圖 3a： 高分作文 A 之長句主題詞串

意義→叢林→工程→讀數字

圖 3b： 低分作文 E 之長句主題詞串

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如表 6 所示。重複到訪主題數大於 1 者，表示其論述前後呼應，故其平均分數最高(11.71 分)。到訪主題數等於 1 且其主題詞距離若大於 2，表示間隔 2 長句以上才回到原本主要主題，這中間的間隔有引名言、例證，以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。

表 5：「最上一層樓」離題實驗結果

離題偵測方法	偵測結果篇數	正確篇數	未含離題評語篇數	正確率	文章平均得分
五個概念化名詞	14	8	6	57.14%	8.07
全部詞彙	22	10	12	45.45%	8.50
長句主題詞	27	17	10	62.96%	8.56

表 6：「最上一層樓」作文 未/含重複到訪主題之比較

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

除了觀察有無重複到訪主題之外，我們統計相鄰主題之間的推導，發現高分作文中常見主題「樓房」後接主題「地基」，此種主題推導與引導文題意吻合，要學生瞭解建築「樓房」最需要的是底層厚實的「地基」。然而此一主題配對卻未見於低分作文中。

4.結論

本論文提出一個中文長句主題詞辨識方法，從實驗顯示此方法能有效地應用於學生作文的離題偵測與文章結構分析。此外，所提出的改良式的小句重心方法也可應用於作文連

貫性評量。在未來研究上，我們期望能擴充作文語料和評分機制，建立一個有效的中文自動評分並引導學生寫作機制，以增加寫作技巧。

5.參考文獻

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

[2] 馬偉雲, “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, 2006.

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

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

[5] 鄭守益, “以語料為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, 2006.

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

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

[8] D. Higgins, J. Burstein, D. Macru and G. Claudia, “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, 2004.

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

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

[11] J. Makkonen, H. Ahonen-Myka and M. Salmenkivi, “Simple Semantics in Topic Detection and Tracking.” Information Retrieval vol.7(3), 347-368, 2004.

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

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

[14] S. Tiun, R. Abdullah and T. Kong, “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, 2001.

[15] D. Wu and T. Liang, “Improving Chinese Pronominal Anaphora Resolution Using Lexical Knowledge and Entropy-based Weight.” Journal of the American Society for Information Science and Technology, 59(13) (2008, November), pp. 2138-2145, 2009.

[16] C. Yeh and Y. Chen, “Creation of Topic Map by Identifying Topic Chain.” Proceedings of the 2004 ACM symposium on Document engineering, Milwaukee, Wisconsin, 2004.