國 立 交 通 大 學

資 訊 科 學 與 工 程 研 究 所

碩 士 論 文

以語料為基礎的中文語篇連貫關係自動標記

Corpus-Based Coherence Relation Tagging in Chinese Discourse

研 究 生：鄭守益

指導教授：梁婷 博士

中華民國 九十五 年 六 月

以語料為基礎的中文語篇連貫關係自動標記

研究生：鄭守益　　指導教授：梁婷

國立交通大學資訊科學與工程研究所

摘要

語篇分析是文本理解中一項不可缺乏的工作，以釐清文章的論題或邏輯結構。因此，本論文乃以語料為主的方法，針對語篇的表層特徵進行收集及探勘，並制定相關的規則，以及提出一套有效的中文語篇自動標記程序。我們使用中研院平衡語料庫 3.0 版作為探勘的語料，計有報導、傳記日記、散文、信函、評論、說明手冊等文類，共 7265 篇。分別針對並列、承接、遞進、選擇、轉折、因果、條件、解證、目的等九種語篇類別，進行線索詞和連續詞性、特殊標點符號等輔助特徵的探勘。在我們的實驗中，使用 100 篇平均字數為 1500 字的報紙社論進行效能評估，在句內的標記部份，正確率可達到 91%，召回率是 95%，篩檢正確率是 98%。另外，在句間的標記部分，正確率可達到 86%，召回率是 93%，篩檢正確率是 95%，。

我們相信藉此語篇標記的研究，有助於將其應用在問答系統、作文評分系統、自動摘要和自動投影片產生系統之上。

關鍵詞：中文、連貫關係、特徵分析、語篇標記、詞彙探勘

Corpus-Based Coherence Relation Tagging in Chinese Discourse

Student : Shou-Yi Cheng Advisor : Tyne Liang

Institute of Computer Science and Engineering

ABSTRACT

Discourse analysis plays an important role of document understanding and is crucial for clarifying the proposition and logical structure of the document. Therefore, this thesis is aimed to built a automated Chinese discourse tagging system by collecting and expanding the coherence feature of discourse base on corpus study and to design the corresponding rules. We used the written documents from Sinica Balance Corpus 3.0 as our mining corpus. It includes 7265 articles covering news, biographies, essays, letters, commentary and illustration manuals. We mine individually cue term, continuous POS tag and peculiar punctuation marks for nine types of rhetorical relations of Chinese discourse, that includes Coordinate, Continue, Option, Forward, Disjunctive, Cause and Effect, Conditions, Elaboration and Goal. In our experiment, we used 100 news editorial articles, each of which contains around 1500 words(1424~1558), as testing corpus. The precision, recall and filtration precision of intra sentence tagging achieve 91%, 95% and 98%. On the other hand, the precision, recall and filtration precision of inter sentence

tagging achieve 86%, 93% and 95%.

Keyword: Chinese, Coherence relation, surface feature analysis, discourse tag, cue term mining.

誌　　謝

本篇論文能夠順利完成，首先要感謝我的指導教授 梁婷博士，在百忙之中撥冗悉心指導，給予許多啟發，感念此恩將永銘在心。感謝所上師長們對我的教誨與提攜，讓我得以一窺管理學術的堂奧。另外，還要特別感謝口試委員陳信希教授、李嘉晃教授及劉美君教授細心審閱全文，對於未臻詳盡之處，給予精闢指正與寶貴之意見，使整篇論文更加完整並契合主題。此外，更要謝謝我的家人，尤其是我的老婆青貞的支持、關懷與體諒，陪我走過這段充滿壓力與痛苦艱熬的論文寫作期間。最後，我要感謝所有實驗室的學長姐、同學以及學弟妹，同窗的情誼是永遠無法抹滅的記憶，你們都是我最好的朋友。

僅以此篇論文獻給我最愛的家人，及曾經關愛、協助過我的人，再次謝謝您們!

目　　錄

表　目　錄

圖　目　錄

第一章　諸論

1.1 研究動機

隨著電腦可處理的語料數量急速成長，自然語言處理技術的開發不論是在資訊的擷取、知識庫建立的自動化、和語言學習等應用上都日顯重要。是以麻省理工學院在 2001 年元月/二月的科技評論中便將自然語言處理列為未來改變世界十大資訊科技之一。

在眾多的自然語言處理技術中，文本理解是能否正確分析及處理語料的重要基礎之一。文本理解的層次小從詞彙、句子，大到段落甚至整篇文章，其最重要的關鍵便在於能否正確的掌握語義的脈絡。以往中文文本處理多以句子為主[林孝璘 '01]，但以句號作為結尾的中文句子有時會錯用，而妨礙表達。例如根據[楊遠 '62]的統計，在五千篇《中國學生週報》文稿中，其中逗點到底者佔 55%，用得不精確者佔 40%，用得正確者 5%。另一方面，由於中文句子中重要的語法項可以刪略[曹逢甫 '95]，因此，當句子單獨地解釋會有歧義，但由上下文裡處理，歧義就消解。

另外，有些語法學家如程祥徽等人['89]發現有時單一句子無法自足，必須由前後銜接連貫的一組句子，定義為句群，才可以解決某些語法問題。句群中有一個明晰的中心意義，因此語篇結構分析可從某個主題引導的一些述題之間的關係來探討。

由上可知，語篇分析是文本理解中一項不可缺乏的工作，以釐清文章的論題或邏輯結構。根據[胡壯麟 '94]，語篇是指在一定語境下表示完整語義的結構，它可以是一個詞、一個句子、或一群連貫的句子組合，

應有一個論題結構或邏輯結構。例如在下面的範例中：

例句 1：雖然天山這時並不是春天(A)，但是有哪一個春天的花園能比得過這時天山的無邊繁花呢(B)？

例句(A)、(B) 小句之間形成轉折的語篇連貫關係，我們可以從這樣的關係來推論作者對於天山的景色這個論旨，含有先貶後褒的用意。

從實際的應用上，正確的語篇結構分析不僅有助於問答系統對解釋型或敘述型答案的辨識，亦有助於作文評分系統中語義連貫的判別[Burstein et al., '98]和文章結構完整性的檢驗[Anthony and Lashkia, '03]。其它的應用還有自動摘要[Chan et al., '00]和自動投影片產生系統

[Tomohide et al., '05]，利用分析語篇的結構找出論文中各個小句之間的連貫關係，抽出關鍵的主題句作為投影片的內容。

由於目前關於中文語篇的研究資料及標記語料皆十分缺乏，對於需要大量已標記語料進行分析及設計語言模型的資訊科學研究者而言，研究中文語篇是一件很困難的工作。因此，本論文的研究目的，乃以語料研究為主的方法，針對語篇的表層特徵進行收集及探勘，並制定相關的規則來建置一自動化的中文語篇自動標記系統，以協助研究者在標記語料時，可以節省大量的人力。

1.2 相關研究

綜觀國外語篇分析的相關研究，許多計算模型都以連貫理論(coherence theory)為基礎[Allen '95; Russell et al., '95]。在連貫理論中，一個語篇是由許多語篇片段(discourse fragment)所組成，有不同的連貫關係，例如：評估、因果、描述、解釋、排列等等。劍橋大學的 F. Wolf 和 MIT 的 E. Gibson[2005]在計算語言學期刊便發表一篇以語料研究為

主的語篇連貫探討，並提出以圖形來表示各種連貫關係的依存現象。相對於連貫理論，許多研究受到知識表徵理論(knowledge representation theory)的影響，用線索片語(cue phrase)來當作語篇中的重要結構元素[Grosz et al., ’86; Hirschberg et al., ‘93]，而不強調語用學理論及世界知識(world knowledge)。例如 Sadao 和 Makoto [‘94]利用線索詞、同義詞或片語及句子相似度來自動判斷日文的語篇結構。

此外， Grosz 等人[‘95] 提出所謂的重心理論(Centering Theory)，探討一段文章的內在結構中其參照延續性(referential continuity)及言談本身特點之間的關聯。該理論有兩項重要的論點： (1) 一段文章或語篇中最重要的訊息，應被視為語篇整體的重心。(2) 在一段連續性的交談中，重複出現的主題或訊息，應被視為語篇整體的重心。重心理論也被應用在指代消解[Chen et al. ‘05]和作文評分上[Miltsakaki ‘00]。

至於中文語篇的研究多為語言學理論的分析。許多學者除了研究句子之外，也試圖提出更大的研究單位，來進行中文語法研究。田小琳['84]提出句群的概念，並定義為有一個明晰的的中心意思，前後銜接連貫的一組句子，把句群和語素、詞、短語、句子並列，正式承認句群也是語法單位。周國正['93]則認為從語法的角度來看，句群應該是某一語言片段中包含帶有一定語法標記例如指示代名詞、句間線索詞語等。若純以意義關係結合為一的語言單位可稱為篇章句群，屬於篇章學的範圍，並非語法單位之一。曹逢甫['95]直接引進語篇概念，他認為：研究漢語應該要有「大句子」的觀念，如此才能解釋主題提昇之後的所謂「雙主語句」。否則一個句子有兩個主語，這是句子取向的語言所不能出現的形式。

黃國文[‘88]提出語篇特性分為銜接與連貫兩種。銜接分為語法與詞彙兩種表層結構，而連貫指的則是語篇片段之間的語義連結。他並將連

貫關係分為並列、對應、順序、分解、分指、重複、轉折、解釋、因果等九種關係。胡壯麟[‘94]則並未強調應將語篇分為銜接及連貫，而是將語篇特性分為指稱性、結構銜接及邏輯連接三種。指稱性及結構銜接都是探討語篇片段中利用詞語或語義的手段來指示語篇之間的關係，他們的不同在於，指稱性的詞語及其所指的對象是相同的，但結構銜接則並不一樣。而邏輯連接則表示相連的句子或句群之間的連貫關係，分為添加、轉折、因果、時空、詳述、延伸、增強等七種關係。程祥徽、田小琳[‘89]使用複句及句群作為研究語篇片段關係的單位，將語篇分為並列、承接、選擇、遞進、轉折、因果、條件、總分、解證、連鎖、目的等十一種關係。

而中文語篇的計算模型則較少被提出。Wang 等人[’98]提出以一個事件模型來表示中文語篇中語段的發展狀態。藉由時間線的推移將語篇結構成一個個的事件，用以表現語義重心的轉移。在過程中使用四種知識以協助事件模型的語義推論，用以解決指代及省略的問題，並應用到數學運算式的問答系統中。另外 Chan 等人[‘00] 以人工方式分析語篇的連貫關係，並制定語篇標記，來協助找出文本中的主題段落作為摘要之候選句，以解決中文自動摘要的問題。

1.3 系統概觀

我們先參考已有的語篇研究資料，並收集各項可資辨認語篇的表層特徵，然後再利用大量語料作為研究對象，進行各項特徵的分佈統計，並以人工進行特徵的篩選及探勘。在觀察及研究的過程中，逐步收集各項比對及合併的規則，並以各項特徵對於語篇辨認的重要性，制定優先順序。本系統之辨識過程分為三個階段，第一階段是針對成對的線索詞組進行比對及合併，第二階段則針對單一出現的線索詞進行比對及合

併，第三階段再處理輔助特徵及特殊單一線索詞的比對及合併，如圖 1-1 所示：

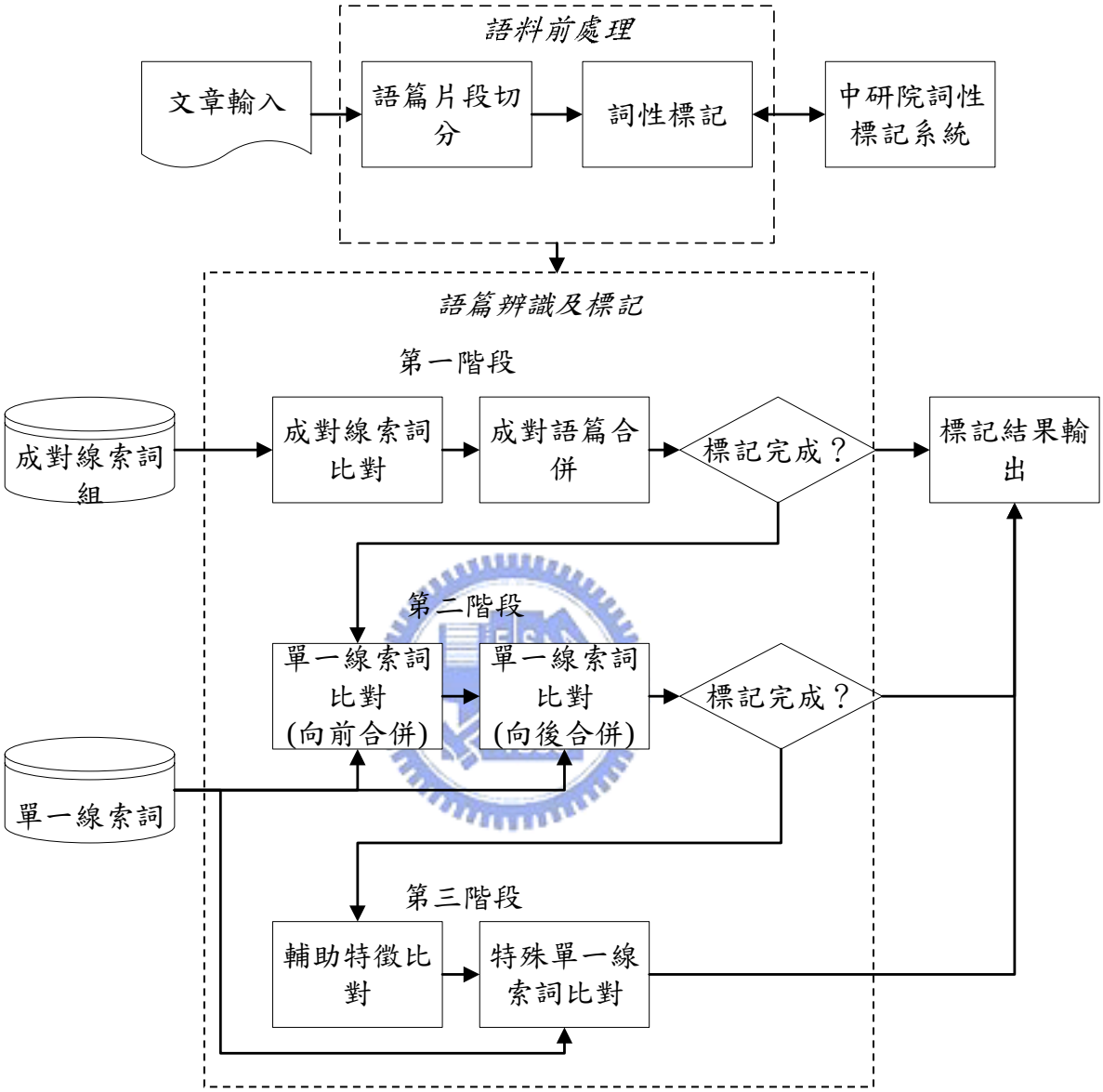

圖 1-1 語篇辨識及標記流程圖

第二章　以文本為主的語篇研究

2.1 切分片段

語篇乃在一定語境下表示完整語義的結構，因此它可以是一個詞、一個句子、或一群連貫的句子[胡壯麟 ’94; 黃國文 ‘88]。語篇單位的切分依照研究者對於語篇的研究需求而有所不同，目前研究者普遍認為語篇片段應為不重疊之文本片段[Marcu ‘00]。Hirschberg 和 Nakatani [‘96]將其定義為韻律單元(prosodic units)，Grosz 和 Sidner[‘86]定義為目的單元(intentional units)，Lascarides 和 Asher[‘93]、Longacre[‘83]及 Webber 等人[‘99]定義為片語單元(phrasal units)，Wolf 和 Gibson[‘05]定義為分句單元(clause units) 以及 Hobbs[‘85]定義為句子單元(sentences)。

由於中文語篇尚未有何種切分片段較為適合做為自動辨識及標記的研究，因此我們將同時以分句(clause)以及長句(sentence)作為切分片段。我們以逗點(，)做為分句的切分界線，長句則以冒號、句號、問號及驚嘆號為切分界線(：。？！)。此外我們對於語篇連貫關係作用於分句間則稱之為句內關係，若作用於長句間則稱之為句間關係。

2.2 語篇連貫關係分類

依據程祥徽與田小琳[‘89] 所提出的複句及句群關係分類來定義語篇片段之間的連貫關係，並參照 Wolf 和 Gibson [‘05]的理論，我們在本論文中，排除沒有明顯表層特徵的總分及連鎖關係，將語篇連貫關係分為如下九類：

表 2-1 語篇連貫關係分類

語篇分類	程&田['89]	Wolf & Gibson['05]	適用種類
並列	並列	Similarity Contrast	句內，句間
承接	承接	Temporal Sequence	句內，句間
選擇	選擇	未定義	句內，句間
遞進	遞進	Elaboration	句內，句間
轉折	轉折	Violated expectation	句內，句間
因果	因果	Cause-effect	句內，句間
條件	條件	Condition	句內，句間
未定義	總分	Generalization	
解證	解證	Attribution Example Elaboration Generalization	句內，句間
未定義	連鎖	未定義	
目的	目的	未定義	句內

其定義分別說明如下：

1. 並列關係：

如 Wolf 與 Gibson ['05]所定義的相似句與對照句的概念(Similarity and Contrast)。指表達幾件相關的事件，但彼此並不構成因果關係，也沒有語氣或語意上的轉折。這種語篇連貫關係在使用上可以不需使用線索詞，例如：「紅的像火，粉的像霞，白得像雪。」也可以用「一方面...另一方面」、「第一...第二」等線索詞，例如：「一方面我們要承擔這樣的責任，另一方面也必須爭取屬於我們的權利。」

2. 承接關係：

描述一連續的動作，或是以發生的時間順序來連接的一連串事件

(Temporal Sequence)，以及依事件發生的空間順序來進行敘述的事件。此類語篇使用線索詞的比例不高[程祥徽與田小琳 ‘89]，有時使用時間名詞或「於是」、「接著」等線索詞，例如：「他先是看了我一眼，接著便怒氣沖沖的走了出去。」

3. 選擇關係：

含有從幾件事物中進行選擇的語義，常用的線索詞有「或者...或者」、「要嘛...要嘛」，例如：「另外有一些人，用他們畢生的勞動鑽研，或者為群眾建造了若干房屋，或者培育出植物的新品種...。」

4. 遞進關係：

我們將凡是連續片段中，具有後一個片段比前一個片段的語意層次更進一層關係的語篇歸類為遞進關係。更進一層可以是範圍更大，數量更多，程度更深等等。常用的線索詞有「不但...而且」、「不只...也」，例如：「她不但鋼琴彈得好，唱歌也很好聽。」

5. 轉折關係：

指前一片段的語義與後一片段相對或相反。常用的線索詞有「雖然...但是」、「盡管...然而」，或僅在後段使用線索詞，例如：「雖然天山這時並不是春天，但是有哪一個春天的花園能比得過這時天山的無邊繁花呢？」

6. 因果關係：

指使用兩個片段來說明事件的原因及其結果，前一片段說明原因，後一片段說明結果。常用的線索詞有「因為...所以」或僅在後段使用「因此」、「因而」，例如：「因為水面上生長白蘋，所以就叫做白蘋湖。」

7. 條件關係：

可分為兩種情況來討論，第一種就是語氣中含有假設成份，前一片段假設一種情況，後一片段說明如果實現的話會產生的結果。常用的線索詞都是成對出現，像是「要是...就」、「假使...就」，例如：「假如我是個舞蹈家，我就要盡情的跳舞。」

第二種情況乃前一片段提出一種條件，後一片段則說明在這種條件下會產生的結果。常用的線索詞有「只有...才」、「除非...才」，例如：「只有用最真誠的演技及誠意，才能感動無數的觀眾。」

8. 解證關係：

此關係可對應 Wolf 與 Gibson [‘05]所定義的四種語篇連貫關係，分別是屬性、範例、詳述以及歸納(Attribution, Example, Elaboration and Generalization)，其中的詳述關係指的是不包含有更進一層的意味，而只是將前一片段所描述的事件作補充說明的範例而言。只要前一片段提出一種看法、道理、事實、現象，而後一片段加以解釋、說明、補充、引申的語篇，我們通稱為解證關係。常用的線索詞有「也就是說」、「所謂」、「宣示」，例如：「她的這種惡意的行為，就是所謂的欺善怕惡。」

9. 目的關係：

這種語篇連貫關係只出現在句內[程祥徽與田小琳 ‘89]，前一片段提出一個目的，後一片段說明為了達成這個目的需要做的事。常用的線索詞有「為了」，例如：「為了便利山區的農業發展，開展山區物候觀測是必須的。」

2.3 語料使用

我們採用的語料是如表 2-2 所列的，中研院平衡語料庫 3.0 版中的

敘述型語料來研究書面形式的語篇連貫關係。

表 2-2 語料研究資料數量統計表

文類	文章篇數	句數	詞數	百分比
報導	5594	104093	2926400	77.00%
傳記日記	399	6466	2600	5.49%
散文	15	501	441200	0.21%
信函	436	16206	67100	6.00%
評論	86	4468	520600	1.18%
說明手冊	735	18800	105400	10.12%

2.4 語篇線索詞研究

根據黃國文['88]的研究指出，語篇的連貫關係可以利用詞彙之間邏輯連接關係來判定，而所使用的詞彙稱為線索詞。這種詞彙指的是可以表示兩個或更多的語篇片段之間的某種邏輯關係，並可藉此辨識出這些片段是屬於哪一類的語篇連貫關係的詞。為觀察語篇線索詞在語料庫中的各項特性，我們採行以下步驟：

步驟一、現有線索詞收集

步驟二、線索詞詞性篩選

步驟三、成對線索詞組探勘

1.設定抽取線索詞之範圍及位置

2.設定線索詞出現位置之權重

3.計算線索詞組之連結強度 k

4.線索詞組篩選

步驟四、單一線索詞探勘

步驟五、輔助特徵探勘

圖 2-1 語篇線索詞探勘流程

2.4.1 現有線索詞收集

我們先以「現代漢語」[程祥徽、田小琳 ’89]這本書裡所列出的語篇線索詞來當作種子，進行更多線索詞的收集。

2.4.2 線索詞詞性篩選

詞的詞性可以影響到詞彙的語義或語法角色，因此我們假定語篇線索詞的詞性也有可能具有某些特定傾向，於是我們從平衡語料庫中將各類線索詞的可能的詞性抽出如下：

Caa, Cba, Cbb, D, Da, DE, Dfa, Dfb, Dk, Na, Nb, Nc, Ncd, Nd, Neqa, Neqb, Nes, Neu, Nf,P, SHI, T, VA, VC, VCL, VD, VE, VG, VH, VJ, VK, VL

圖 2-2 種子線索詞詞性列表

由於有些詞根據他的語法角色及出現的位置不同，會具備有許多的詞性。例如：「也」當成語助詞時並不具備有連接分句，並構成連貫關係的功能，如例句 1；但在例句 2 中，「也」當成副詞時，便構成並列關係。

*例句 1：有骨氣，哲學家當如是**也**(T)。*

*例句 2：不僅武功一流，內秀**也**(D)十分了得。*

表 2-3 列出的詞性是我們認為可以排除的詞性：

表 2-3　排除詞性表

詞性	排除原因與範例

Na 一般名詞	一般名詞並不具備連接分句或句子的功能。 例如：此**等**情懷見乎其自明本志令、短歌行**等**文字。 只得向顧客賠聲**不是**
Nb 專有名詞	專有名詞並不具備連接分句或句子的功能。 例如：**如**一面幫著城把東西拿進來 第二次會談是七月二十五日在京**都**的**都**飯店進行
Nc 地方詞	地方名詞並不具備連接分句或句子的功能。 例如：申辦公元二千年夏季奧運會的世界五大名**都**
SHI 是	是並不具備連接分句或句子的功能。 例如：所以休閒活動本身**就是**倫理性的。
T 語助詞	語助詞並不具備連接分句或句子的功能。 例如：有骨氣，哲學家當如是**也**！
VA 動作不及物動詞	動作不及物動詞並不具備連接分句或句子的功能。 例如：壯士一去兮不復**還**！
VC 動作及物動詞	動作及物動詞並不具備連接分句或句子的功能。 例如：我母親前半年在公車站牌**等**車時，
VCL 動作接地方賓語動詞	動作接地方賓語動詞並不具備連接分句或句子的功能。 例如：他說我們大眾不是軍隊，可以**越**二級報告
VD 雙賓動詞	雙賓動詞並不具備連接分句或句子的功能。 例如：其他商家借了巨款而終於無力償**還**，借出的商家便大方地一筆勾銷
VG 分類動詞	分類動詞並不具備連接分句或句子的功能。 例如：人生本**若**夢，又何必辛苦規劃工作？
VH 狀態不及物動詞	狀態不及物動詞並不具備連接分句或句子的功能。 例如：於是只能退而求**其次**
VJ 狀態及物動詞	狀態及物動詞並不具備連接分句或句子的功能。 例如：不要馬上又**接著**長途的開車。
Nf 量詞	並不具備連接分句或句子的功能。 例如：他今年 12 **才**。

由上列資料，我們將收集的各語篇之線索詞詞性重新整理。

2.4.3 成對線索詞組探勘

我們利用 Smadja [‘93]所提出用來抽取英文共現詞彙的 *Xtrac*t 為基礎，設計一個抽取連接分句片段或長句片段的線索詞組的改良式演算法，稱為 *CoXtract*。此演算法將產生一個可以量化具連結兩個語篇片段的線索詞組之連結強度值 *k*，我們用來協助將候選詞彙進行排序。並藉由已知的線索詞作為觀察標的，以人工的方式，抽取出更多的語篇線索詞組，並在過程中針對 *k* 值的效用進行評估。

由於解證及目的這兩種語篇並沒有成對的線索詞[程祥徽、田小琳 ’89]，因此我們在探勘成對線索詞時，將排除這兩種語篇連貫關係。以下是探勘步驟：

1. 設定抽取線索詞之範圍及位置

   下圖為線索詞所連結的分句或長句的示意圖：

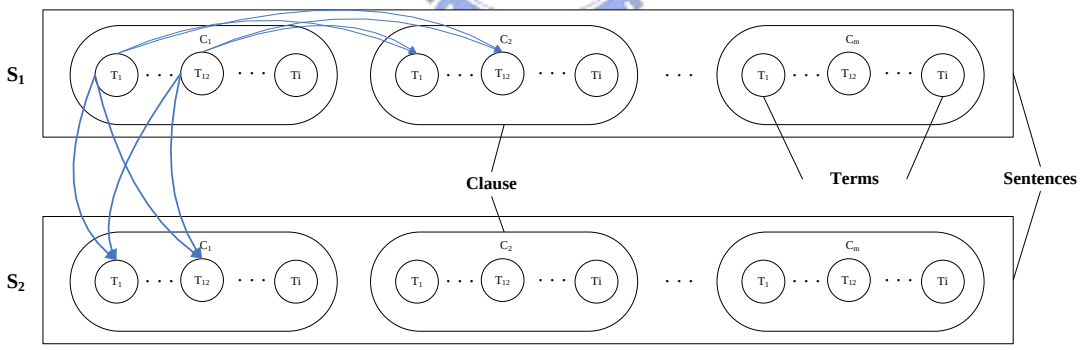

圖 2-3 語篇連結示意圖

我們從「現代漢語」中選出已知的線索詞組進行語料觀察，並設定以下兩項資料的抽取門檻值，分別是線索詞出現在分句內之位置及成對線索詞的平均涵蓋範圍。

在計算線索詞可能出現在分句內的位置時，我們隨機選取 24 組線索詞及 1150 組例句共 2300 個分句片段。接著計算各分句的平均

詞數，並假設成對線索詞出現的位置皆在分句的前半部分，統計結果，我們以平均詞數 24 除以 2 作為觀察線索詞分布的抽取門檻值，我們的統計結果如下：

表 2-4 線索詞分布位置觀察結果

線索詞位置	例句數	百分比	累計百分比
1	1942	84.43%	84.43%
2	193	8.39%	92.83%
3	49	2.13%	94.96%
4	43	1.87%	96.83%
5	27	1.17%	98.00%
6	24	1.04%	99.04%
7	8	0.35%	99.39%
8	6	0.26%	99.65%
9	1	0.04%	99.70%
10	4	0.17%	99.87%
11	2	0.09%	99.96%
12	1	0.04%	100.00%

由以上的結果，我們為求能盡量觀察更多的資料，因此在抽取候選詞組時，將抽取線索詞的分布位置門檻值設為 12。

另一方面，計算成對線索詞的平均涵蓋範圍時，分為句內及句間兩種類型來分別統計。我們以人工選取句間線索詞組數 21 個，例句 625 句；句內線索詞組數 22 個，例句 1037 句。觀察之後，我們將抽取線索詞的句內及句間連結門檻值均設為 3，選取的線索詞及統計的結果分別如下表所示：

表 2-5 句間線索詞連結範圍統計結果

連貫關係	關聯前詞	關聯後詞	前詞詞性	後詞詞性	平均距離	例句數
並列	一方面	另一方面	Cbb	Cbb	2.32	35
	既然	而	Cbb	Cbb	2.82	33
承接	目前	未來	Nd	Nd	3.09	51
	當時	現在	Nd	Nd	3.26	19
選擇	或	或	Caa	Caa	2.76	46
	或	更	Caa	D	2.76	33
遞進	不只	也	Da	D	2.75	20
	不但	也	Cbb	D	3.22	25
	不僅	而	Cbb	Cbb	2.92	24
轉折	雖然	但是	Cbb	Cbb	3.28	32
	雖然	不過	Cbb	Cbb	3.09	33
	雖然	還	Cbb	D	2.79	34
因果	由於	因此	Cbb	Cbb	2.78	49
	由於	所以	Cbb	Cbb	3.28	36
	由於	於是	Cbb	Cbb	3.05	21
假設	如果	將	Cbb	D	2.5	28
	即使	也	Cbb	D	2.97	31
	即使	都	Cbb	D	2.83	29
條件	不論	也	Cbb	D	2.58	24
	只要	就	Cbb	D	2.97	30
	只要	都	Cbb	D	2.74	27
總計資料					2.89	625

表 2-6 句內線索詞連結範圍統計結果

連貫關係	關聯前詞	關聯後詞	前詞詞性	後詞詞性	平均距離	例句數
並列	一方面	一方面	Cbb	Cbb	1.23	80
	既	也	Cbb	D	2.45	22
承接	目前	未來	Nd	Nd	1.97	29
	首先	其次	Cbb	Cbb	1.98	22
選擇	或者	或	Caa	Caa	2.64	32
	或是	或是	Caa	Caa	2.38	47
	或是	或	Caa	Caa	2.63	57
遞進	不但	也	Cbb	D	2.38	58
	不但	而	Cbb	Cbb	2.43	69
	不僅	而	Cbb	Cbb	2.27	60
轉折	雖然	但	Cbb	Cbb	1.98	53
	雖然	但是	Cbb	Cbb	3.54	39
	雖然	還	Cbb	D	3.89	37
因果	由於	因此	Cbb	Cbb	2.79	33
	因為	因此	Cbb	Cbb	2.98	40
	因為	才	Cbb	Da	2.6	58
假設	如果	則	Cbb	D	2.64	58
	如果	可能	Cbb	D	2.84	50
	即使	也	Cbb	D	3.04	50
條件	只有	就	D	D	2.61	46
	只要	可以	Cbb	D	2.15	55
	只要	會	Cbb	D	2.17	42
總計資料					2.53	1037

接著我們以上述兩個抽取門檻值來建立線索詞組候選資料，共抽取句間 207867 組及句內 209525 組候選詞組。

2. 設定線索詞出現位置之權重

由於線索詞在分句或句子中的位置，出現在越前面的位置，就越有可能具有判斷兩片段是具有連結性的功能。因此在判斷線索詞是否具有連接功能時，必須針對出現在不同位置，給予不同的權重。例如在例句 3 中之「或是」具有連接兩個分句的功能，但在例

句 4 中則無此功能。

*例句 3：你可以選擇今天就回家，**或是**(Caa)明天再回家。*

*例句 4：今天的晚餐很豐富，你可以選擇要吃便當**或是**(Caa)燒烤。*

我們藉由表 2-4 中所計算的線索詞分布數據作圖之後發現，其分布近似於函數 $\frac{1}{x^3}$，如下圖所示：

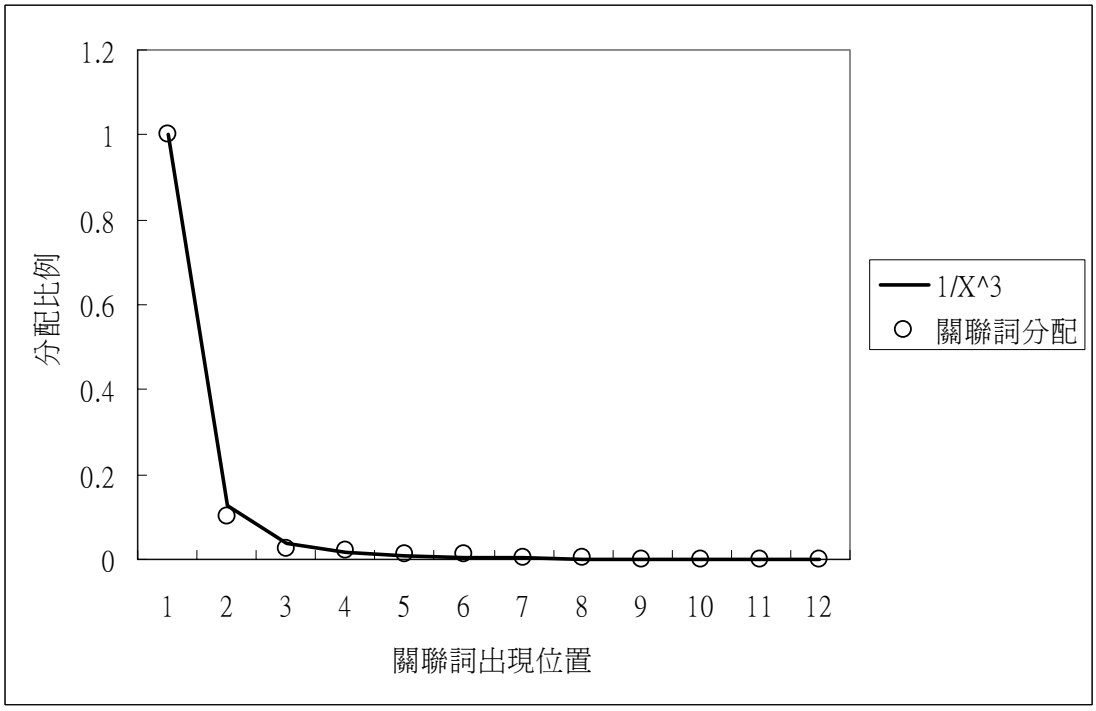

圖 2-4 線索詞分布位置

由上圖之數據並進行正規化使其權重值介於 1 於 0 之間。設線索詞在分句中出現的位置共有 j 個，由線索詞出現在分句內之位置的觀察結果，設$1 \le j \le 12$。則：

正規化常數為 D，$D = \sum_{n=1}^{12} \frac{1}{n^3} = 1.2$ (2.1)

若線索詞出現在第 j 個位置，則其權重 $w_j = \frac{1}{1.2 j^3}$ (2.2)

3. 計算線索詞組之連結強度 k

我們可定義詞彙之間的共現度為(T_h, T_i, d)。其中T_h是給定的關聯前詞，T_i則是T_h的共現關聯後詞，根據線索詞連結範圍觀察結果所設之門檻值，其出現位置共區分為兩種，第一種為出現在T_h所在分句片段的次三分句片段範圍內；第二種出現在T_h所在長句片段的次三長句片段範圍內，且出現位置在距離 d 之內，依據線索詞出現位置的觀察結果設 d 為[1,12]。我們以(T_h, T_i)這對詞彙一起出現在語篇片段之間的頻率的標準差倍數，作為其連結強度[Smadja '93]，其公式如下：

$$k_i = \frac{f_i - \overline{f}}{\sigma} \tag{2.3}$$

其中共現關聯後詞T_i出現在詞窗 [1,12]的頻率 f_i 定義為：

$$f_i = \sum_{j=1}^{12} f_{i,j} w_j \tag{2.4}$$

其平均頻率 $\overline{f}$ 以及標準差σ的計算公式如下：

$$\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f_i \quad ; \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f_i - \overline{f}\right)^2} \tag{2.5}$$

4. 線索詞組篩選

我們配合 k 值排序並以人工觀察，篩選出句內線索詞組 406 組，句間線索詞組 82 組。

2.4.4 *k* 值觀察

我們利用所篩選之線索詞組分別觀察，*k* 值對於抽取出的各種語篇連貫關係線索詞組正確率的影響。我們的觀察指標有：

1. 正確詞組累計

   當 *k* 值由小到大變化時，可以抽取出的正確線索詞組數量的變化趨勢。

2. 涵蓋例句累計

   當 *k* 值由小到大變化時，可以抽取出的正確線索詞組所涵蓋之例句數量的變化趨勢。

3. 詞組平均正確率

   當 *k* 值由小到大變化時，可以抽取出的正確線索詞組與全部詞組的平均比例變化趨勢。

4. 詞組平均涵蓋率

   當 *k* 值由小到大變化時，可以抽取出的正確線索詞組所涵蓋之例句數量與全部詞組所涵蓋的數量之平均比例的變化趨勢。

由圖 2-5 及圖 2-6 可以看出，*k* 值對於句內之線索詞組有較好的鑑別度。在 *k* 值為 0.8 時，線索詞組平均正確率及詞組平均涵蓋率可達 92% 及 94%，其數量累計分別為 76%及 93%，且其變化趨勢已呈現一收斂狀態。反觀句間線索詞的情形則明顯的鑑別度較差，在 *k* 值為 0.8 時，線索詞組數量累計可達 87%及 90%，但平均正確率及詞組平均涵蓋率卻分

別只剩 63%及 65%，且其變化趨勢尚呈現起伏的波動狀態。我們根據所收集的涵蓋資料量推論，由於句間大部分皆使用單一線索詞作為語篇結構的連結，因此在例句數量不足的情況之下，使得 *k* 值的變化趨勢波動幅度變大，甚至無法收斂。

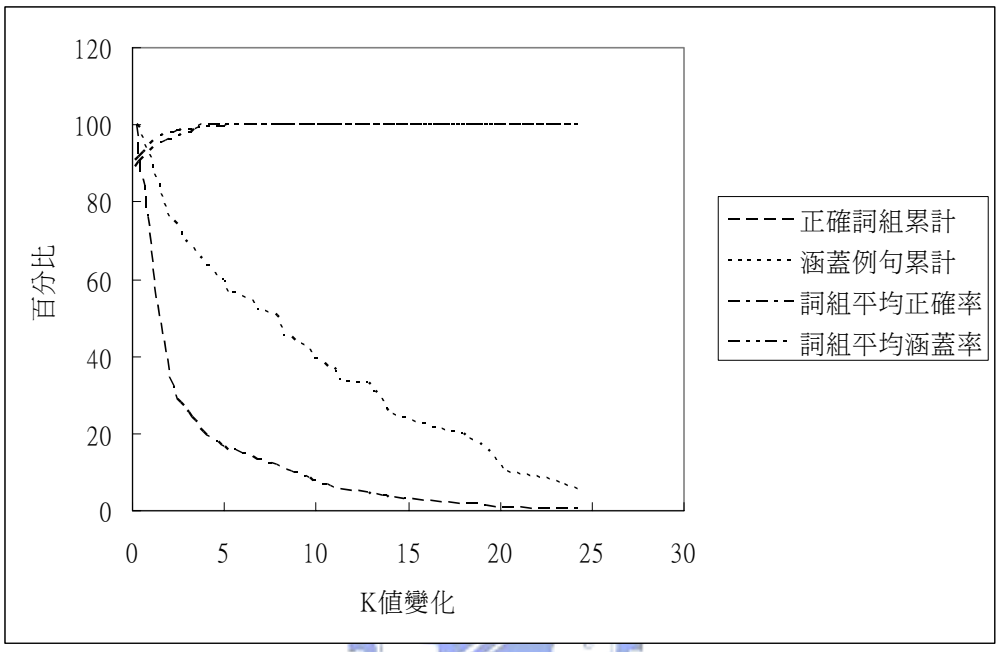

圖 2-5 句內線索詞組 K 值變化圖

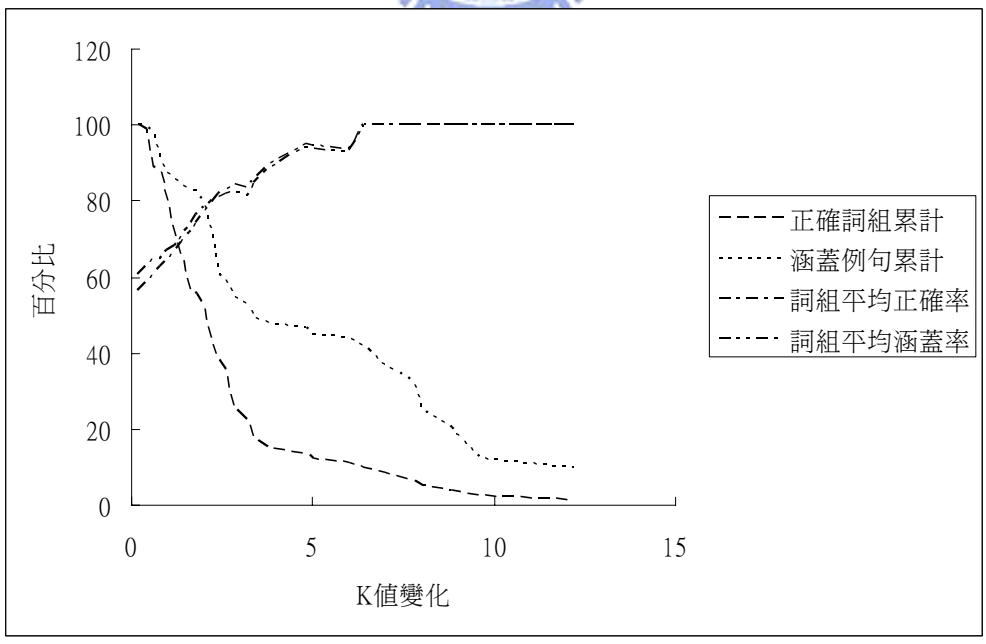

圖 2-6 句間線索詞組 *k* 值變化圖

k 值除了在詞組擴充時可作為門鑑值以自動濾除可能錯誤的詞組之

外，亦可應用在語篇標記時，用來判斷出現詞組中之任一詞彙或單一線索詞與未知線索詞的可能連接強度。

2.4.5 單一線索詞探勘

中文語篇的線索詞可分為成對及單一兩種形式，有些成對線索詞因語氣的輕重不同，有時也可單獨出現，例如：

*例句 5：他**不但**吃米飯(A)，**也**吃牛排(B)。*

藉助「不但…也」這對線索詞組，可將例句中之(A)及(B)兩個片段判定為遞進關係，若改寫成：

*例句 6：他吃米飯(A)，**也**吃牛排(B)。*

則因「也」這個線索詞的單獨出現，而變成並列關係。而中文線索詞在書寫的過程中，常會省略關聯前詞，而單用關聯後詞，如例句 7 也可改寫為例句 8 的形式：

例句 7：如果我們這麼做，可能會導致環境的破壞。

例句 8：我們這麼做，可能會導致環境的破壞。

另外，也有某些情況會省略關聯後詞，而單用關聯前詞，如例句 9 也可改寫為例句 10 的形式：

例句 9：因為情勢如此變化，所以我們不得不做這樣的決定。

例句 10：因為情勢如此變化，我們不得不做這樣的決定。

除此之外，解證及目的這兩種語篇的線索詞都是單獨出現，例如：

*例句 11：同時也談到科學的發現不能設計或預期，**也就是說**(A)，我們應該努力創造良好的科學研究條件與環境(B)，真正培養努力鑽研的科學家，這才能使科學方面經常有若干新的創獲。*

藉由「也就是說」，我們可以將例句中之(A)及(B)這兩個分句片段判定為解證關係。因此，在語篇連貫關係辨識的過程中，除了成對的線索詞組之外，也有必要進行單一線索詞的收集及探勘工作。

單一線索詞主要分為三類：

1. 成對線索詞組的省略

由於人們在使用語言有時會為了增進溝通效率或因應語氣的輕重不同而有簡省詞彙的趨向，而在語篇線索詞的使用上也具有這樣的特性，因此，我們假設成對線索詞皆可分別單獨使用。

2. 語篇線索詞特性

解證及目的兩種語篇的線索詞都是單獨出現[程祥徽與田小琳 ‘89]，因此，我們也收集了屬於這兩個語篇的單一線索詞。

3. 特殊語篇線索詞

我們由已知的線索詞，透過 HOWNET[Dong and Dong, ‘99]中的 DEF 欄位，進行語料的觀察發現，還有一些線索詞可以幫助我們判斷語篇片段之間的關係，但是卻未被語言學者提出，例如：我們發現當動作句賓動詞(VE)出現在分句片段末尾位置時，具有連接兩個語篇片段成為解證關係的特性，例如：

*例句 12：關於公司的前景，張總經理**表示**，未來將以生物科技搭配醫療器具的生產為主。*

由以上三種來源，我們以人工的方式進行辨識篩選，以達成探勘的目的，共收集了 309 個單一線索詞，其中第一類線索詞有 65 個，第二類有 60 個，第三類有 184 個。

使用單一線索詞來辨識語篇連貫關係時，還需要考慮連結方向、涵

蓋範圍以及出現位置等三個問題。因此，我們設計了以下屬性：

1. 連結方向

此屬性分為兩種情況：若由線索詞向後連結次一片段，則將此值設為 1，若為向前連結前一片段，則設為-1。如下圖所示：

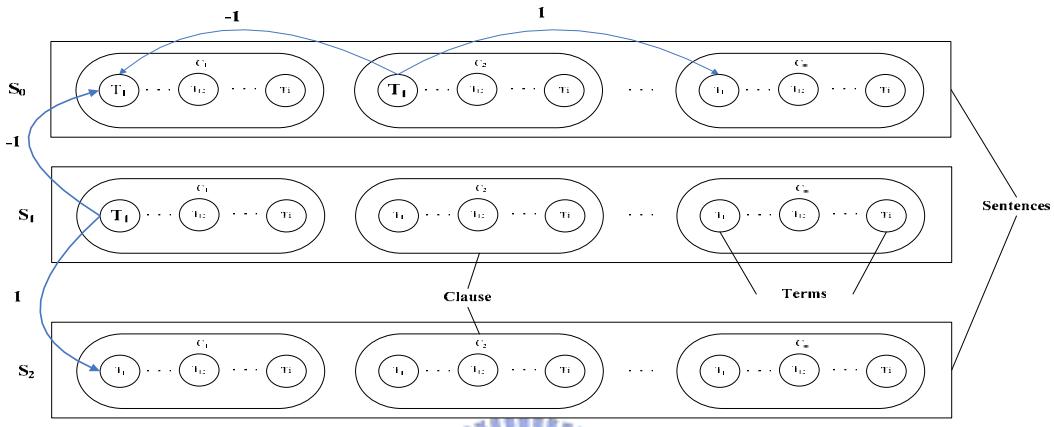

圖 2-7 連結方向示意圖

2. 出現位置

線索詞出現的位置可分為兩種，一為出現在語篇片段的前半部份，並在我們所設定的位置門檻值內的位置，我們將此值設定為 0；另外則為出現在語篇片段末尾，我們將此值設定為 1。至於出現於中間位置的線索詞，我們則忽略不計。

3. 適用片段種類

可同時使用在句內及句間的線索詞，則此值設定為 1；反之若只能使用在句內，則設定為 0。

至於單一線索詞的涵蓋範圍，我們則不另外設定屬性標示。我們的假設是，單一線索詞的連結涵蓋範圍，為鄰近的前一或次一單獨語篇片段或具有某些語篇連貫關係的語篇段落。如果我們能設計良好的語篇標記優先順序及合併規則，則涵蓋範圍的問題將自動解決，例如：

*例句 13：他**不但**是個品學兼優的好學生(A)，**而且**還熱心助人(B)，**所以**我們班的同學都很喜歡他(C)。*

在例句 9 裡的三個語篇片段中，我們先以成對線索詞「不但...而且」合併(A)、(B)兩個語篇片段，成為新的語篇片段群(AB)，然後再依我們的假設使用「所以」這個單一線索詞向前合併(AB)，如此我們便不需要去設定可能錯誤的涵蓋門檻值，但是這個方法的正確率，取決於對各種語篇連貫關係標示的涵蓋率，及合併規則的正確率。

2.4.6 輔助特徵探勘

我們為了提高語篇辨識的涵蓋率，因此參考程祥徽與田小琳[‘89]及 Tomohide 與 Sadao[‘05]的研究，設定了如下四種輔助特徵：

a. 當具有時間詞(Nd)詞性的詞彙，例如：「今天...明天」，出現在連續的語篇片段時，則可判定這些語篇片段具有「承接關係」，例如：

*例句 14：**今天**我預習了國、英、數三個基本科目，**明天**我將繼續把理化、生物等科目也預習一遍。*

b. 當具有數詞定詞(Neu)詞性的詞彙，例如：「第一...第二」，出現在連續的語篇片段時，則可判定這些語篇片段具有「並列關係」，例如：

*例句 15：**第一**、我們要振興經濟，**第二**、我們要防止舞弊...*

c. 語篇片段的末尾若出現標點符號「：」，則可判定其次一語篇片段為「解證關係」，例如：

例句 16：IPv6 具備下列各項特性：1.較大的位址空間，2. 整合認證及安全的機制，3. 較佳的路由效率及最佳化。

d. 若相似的語篇片段連續出現時，則可以將這些語篇片段判定為

「並列關係」，例如：

例句 17：紅的像火，粉的像霞，白的像雪。

第三章　語篇辨識及標記

3.1 名詞定義與標記符號說明

實驗過程中所需使用之相關名詞說明如下：

1. 語篇片段：分成長句及分句語篇片段。
2. 句間關係：存在於當語篇片段單位為長句時。
3. 句內關係：存在於當語篇片段單位為分句時。
4. 語篇段落：內含數個語篇片段，並至少已合併一個或以上之語篇連貫關係的長句群或分句群稱之。
5. 待處理文本：分成句內關係比對的長句和句間關係比對的整篇文章。

我們在剖析的過程中，依據所制定的各種比對及合併的原則，將輸入的文本自動標記出相應的語篇連貫關係，因此每一個語篇段落都標記有語篇連貫關係之類型。若某語篇段落內含兩個或以上之語篇片段時，則依規則，標記為樹狀結構，而段落與段落間的結構關係，則不予辨識，若某一段落只有單一片段則不予標記，以下為語篇連貫關係符號表：

表 3-1 語篇連貫關係標記符號表

符號	適用關係	說明
@	全部	以“@”置於語篇之間，做為分隔語篇段落的界線。
( )	全部	語篇的組合成分為複雜結構，以“(”、“)”標示其結構的左右邊界。
|	全部	分隔在同一層次上的成分結構。
D#,	全部	以”D”作為開頭的數字，為語篇連貫關係之編號，如表 3-2 所示。
[]	全部	以“[”、“]”標示語篇片段的左右邊界。
C#	句內	以”C”作為開頭的數字，為分句語篇片段在整個長句裡的排列順序。
S#	句間	以”S”作為開頭的數字，為長句語篇片段在整個文章裡的排列順序。
Theme	全部	以 Theme 標示語篇連貫關係中的第一個語篇片段。
Rheme	全部	以 Rheme 標示語篇連貫關係中的其他語篇片段。

表 3-2 語篇連貫關係符號表

關係編號	連貫關係	符號
1	並列	Coordinate
2	承接	Continue
3	選擇	Option
4	遞進	Forward
5	轉折	Disjunctive
6	因果	Cause and Effect
7	條件	Conditions
8	解證	Elaboration
9	目的	Goal
10	其他	Other

例句 1 經過電腦標記產生的句內語篇連貫關係結構為例句 2，其中每一個語篇片段皆以“|”分開：

例句 1：立委或輿論如果將關切重點放在蔡英文是否聰明抑或生澀，以及致電目的是關切審查程序抑或實質內容關說，可能模糊了焦點或偏離主題，對台灣經濟發展無

甚幫助。

*例句 2：**D1,([C1:**立委或輿論如果將關切重點放在蔡英文是否聰明抑或生澀，**]|D7,([C2:**以及致電目的是關切審查程序抑或實質內容關說，**]|[C3:**可能模糊了焦點或偏離主題，**]))@[C4:**對台灣經濟發展無甚幫助。**]***

例句 3 經過電腦標記產生的句間語篇連貫關係結構為例句 4，其中每一個語篇片段皆以"|"分開：

例句 3：行政院副院長蔡英文一通關切環評進展的電話，竟然引發多名環評委員發表聲明，譴責行政院高層干預中部科學園區環評審查。然而，中部科學園區、國光石化及台塑大煉鋼廠案所涉及的環境評估、經濟發展及社會觀感，及其背後關鍵的政府基本政策與選擇，遲早政府必須對外說清楚、講明白。

*例句 4：**D5,([S1:**行政院副院長蔡英文一通關切環評進展的電話，竟然引發多名環評委員發表聲明，譴責行政院高層干預中部科學園區環評審查。**]|[S2:**然而，中部科學園區、國光石化及台塑大煉鋼廠案所涉及的環境評估、經濟發展及社會觀感，及其背後關鍵的政府基本政策與選擇，遲早政府必須對外說清楚、講明白。**])***

3.2 辨識及標記執行步驟

我們使用中央研究院所開發之線上中文斷詞系統[1]，進行文本之斷詞及詞性標記的工作。並將語篇辨識及標記的工作分為三個階段，分別依線索詞及輔助特徵的優先順序進行比對，其整體步驟如下表所示：

[1] 請參閱網址：http://ckipsvr.iis.sinica.edu.tw/

表 3-3 語篇連貫關係辨識及標記步驟

階段	步驟	執行動作	適用關係
一	**成對線索詞組比對**		
	1	以二字組為單位進行成對關鍵詞比對。	全部
	2	將比對後具有語篇連貫關係之二字組合併。	全部
	3	判斷是否已合併完成，並進行語篇連貫關係標記。	全部
二	**單一線索詞比對**		
	1	進行向前連結之單一關鍵詞比對及合併。	全部
	2	進行向後連結之單一關鍵詞比對及合併。	句內
	3	判斷是否已合併完成，並進行語篇連貫關係標記。	全部
三	**輔助特徵及特殊單一線索詞比對**		
	1	進行連續 Nd 及 Neu 詞彙之比對及合併。	全部
	2	進行解證關係標點符號之比對及合併。	全部
	3	進行相似句之比對及合併。	句內
	4	進行特殊線索詞之比對及合併。	全部

執行上述步驟時，我們將比對之線索詞分布位置門檻值及語篇連貫關係連結範圍門檻值設為 3。此門檻值之限制並不包括第三階段之特殊單一線索詞之比對工作，以下分別說明各階段之流程。

3.2.1 成對線索詞組比對

由於成對線索詞組本身即具有排除語篇連貫關係歧義性，及明顯合併範圍和方向之特性，因此我們將其列為第一優先比對的特徵，此階段分為三個步驟，茲詳細說明如下：

步驟 1：以二字組為單位進行成對線索詞比對。

假設某待處理文本所含之語篇片段數量為 n，語篇連結門檻值為 d，則我們可據此產生一個長度為 n 的輸入陣列及 $n \times d$ 之比對結果矩陣，若為句間比對，則以每一長句第一分句為輸入之比對片段。將陣列輸入系統，並依序增加 d 值進行成對線索詞比對，其演

算法如下圖所示：

輸入：由待處理文本所形成之長度為 n 的陣列 *InputContextArr[n]*
輸出：內含以 bi-gram 為單位比對後之 $n \times d$ 結果矩陣。*IPKResultMir[n,d]*

1. *FOR i=1 TO Min(n −1, d)*
2. *FOR j=1 TO n*
3. 分別挑選第 j 個語篇片段及第 j+i 個語篇片段的詞彙進行比對。
4. 若比對成功，則以命中之語篇編號及 *i,j* 之值產生合併字串，填入結果矩陣。
5. 輸出結果矩陣。

圖 3-1 成對線索詞比對演算法

步驟 2：將比對後具有語篇連貫關係之二字組合併。

我們將合併之過程分為兩個部份，第一個部分稱為縱向合併，其遞增變數為門檻值 d ，此部分主要是處理同一片段的合併問題。第二個部分稱為橫向合併，其遞增變數為 n ，此部分主要是處理相鄰片段的合併問題，我們將依循以下規則：

規則 1：同一片段若同時與兩個以上之片段形成語篇連貫關係時，只保留距離最小者。

規則 2：相鄰片段若形成相同的語篇連貫關係時，合併成為同一語篇在同一階層。

規則 3：相鄰片段若形成不同之語篇連貫關係時，以向左合併為原則，合併成為不同語篇不同階層。

其演算法如下圖所示：

輸入：成對線索詞比對結果矩陣 *IPKResultMir[3,n]*
輸出：成對線索詞合併結果陣列 *MergerResultArr[n]*
&&Merge_Step1
1. *FOR i=1 TO Min(n-1,d)*
2. *FOR j=1 TO Min(n-1,d)*
3. 依序比對 *IPKResultMir[j,i]*的合併段落，取距離最小者，並置入 *MergerResultArr[i]*

&&Merge_Step2
4. *FOR i=1 TO n*
5. 若 *MergerResultArr[i]*及 *MergerResultArr[i+1]*為不同語篇，則
6. 合併成為不同語篇不同階層，填入 *MergerResultArr[i]*
7. 否則，若 MergerResultArr[i]及 MergerResultArr[i+1]為相同語篇，則
8. 合併成為同一語篇在同一階層，填入 *MergerResultArr[i]*
9.輸出結果陣列 *MergerResultArr[n]*

圖 3-2 成對線索詞合併演算法

步驟 3：判斷是否已合併完成，並進行語篇連貫關係標記。

若輸入文本已合併為單一語篇段落，則對照語篇連貫關係符號表進行語篇標記後跳出比對流程，若尚未合併為單一語篇段落，則繼續第二階段之比對工作。如例句 4 為已合併之單一語篇段落，例句 5 為標記後之結果：

例句 4：D3,([C1:想在黑白分明的領域中取得「兼顧」的可能，]|D5,([C2:或者說是政策的妥協點，]|[C3:其實是不切實際的幻想。]))

例句 5：Option:(Theme:[C1:想在黑白分明的領域中取得「兼顧」的可能，]| Rheme:Disjunctive:(Theme:[C2:或者說是政策的妥協點，]| Rheme:[C3:其實是不切實際的幻想。]))

3.2.2 單一線索詞比對

此階段所指的「單一線索詞」是不含第二章所指之特殊線索詞的子

集合，包括成對線索詞組的省略詞，及解證與目的關係中的一般線索詞共 244 個，其屬性值為(-1,0,0)、(-1,0,1)、(1,0,0)。

根據我們的觀察，若在句內省略前詞單用後詞，則其連結方向大部分為向前連結，反之亦然。但若在句間則為單用後詞居多，如例句 6：

*例句 6：雲林縣此舉，除了財政拮据之外，還夾雜著對大規模企業「本縣拉屎，他處下蛋」的忿懣與積怨，因此高舉防治污染大旗，以環境保護為名義徵稅。**然而**，純就租稅體制而言，雲林縣此舉並不符合稅制的基本邏輯。*

單一線索詞比對的另外一個問題就是會出現複合語篇線索詞的情況，如例句 7 所示：

*例句 7：他會這麼做(A)，多少**也因為**還愛著你(B)。*

在(B)中出現兩個單一線索詞，一個是表示並列關係的「也」，另外一個為表示因果關係的「因為」。綜上所述，我們將依循以下規則進行第二階段各步驟之比對及合併：

規則 4：若比對單一線索詞時，同一語篇片段出現兩個以上之候選線索詞，則依以下優先順序決定： Cbb> Caa>Cab>Cba>D>Da>Dk>P

規則 5：單一線索詞連結時須避免將內含輔助特徵及特殊線索詞之語篇片段合併。如例句 8 中所出現之線索詞「或」，不應合併(A)，因其包含了特殊線索詞「宣示」。

*例句 8：我們建議政府儘快明白**宣示**(A)，**或**為政治、經濟問題(B)，國家永續發展問題，何者才是政府的最大關切？*

規則 6：若向前合併之單一線索詞單獨出現在第一分句，則為句間線索詞，不與句內連結。如例句 10 之線索詞「然而」。

規則 7：若向後合併之單一線索詞單獨出現在第一分句，則為

句內線索詞，不與句間連結，如例句 9 之線索詞「即使」。

*例句 9：**即使**真的應將污染性企業產值列入分配因素考量，亦不應只涵蓋石化工業，高污染產業還有很多。*

單一線索詞比對共分為 3 個步驟，茲分別說明如下：

步驟 1：進行向前連結之單一關鍵詞比對及合併

我們優先比對使用率較高的關聯後詞，及解證與目的關係中的一般線索詞，其演算法如下圖所示：

輸入：第一階段輸出之合併結果陣列 *MergerResultArr[n]*
　　由待處理文本所形成之長度為 *n* 的陣列 *InputContextArr[n]*
輸出：單一線索詞比對及合併結果陣列 *MergerResultArr2[n]*

&&Map_Step1
1.若為句內關係比對，則
2. *KeyWordArr[]*←關連詞屬性值為(-1,0,0)及(-1,0,1)之單一線索詞集合
3.若為句間關係比對，則
4. *KeyWordArr[]*←關連詞屬性值為(-1,0,1)之單一線索詞集合
5. *FOR i= 2 TO n*
6. 從 *InputContextArr[i]* 中，提取候選詞彙與 *KeyWordArr[]* 進行比對
7. 若比對成功，則
8. 若有兩個或以上之候選詞命中，則依規則 4 篩選
　　&&進行向前合併流程
9. 檢查是否 InputContextArr[BackSearchC]中是否含有輔助特徵的關鍵詞
10. 若無，則以向左合併為原則，填入 *MergerResultArr2[BackSearchC]*
11. 檢查是否已合併前一語篇片段或段落，若是，則跳出
12.輸出結果陣列 *MergerResultArr2[n]*

圖 3-3　向前合併線索詞比對及合併演算法

步驟 2：進行向後連結之單一關鍵詞比對及合併

此步驟為比對其他向後連結之單一線索詞，但句間不作此項比對流程，其演算法如下圖所示：

輸入：步驟 1 輸出之合併結果陣列 *MergerResultArr2[n]*
　　由待處理文本所形成之長度為 *n* 的陣列 *InputContextArr[n]*
輸出：單一線索詞比對及合併結果陣列 *MergerResultArr3[n]*
&&Map_Step2
1.若為句內關係比對，則
2.　KeyWordArr[]←關連詞屬性值為(1,0,1)及(1,0,0)之單一線索詞集合
3.若為句間關係比對，則
4.　跳過此步驟
5. *FOR i= (n-1) TO 1　STEP -1*
6.　從 *InputContextArr[i]* 中，提取候選詞彙與 *KeyWordArr[]* 進行比對
7.　若比對成功，則
8.　　若有兩個或以上之候選詞命中，則依規則 4 篩選
　　&&進行向後合併流程
9.　　*FOR j = i+1 TO Min(n,i+d)*
10.　　檢查 *InputContextArr[BackSearchC]* 中是否含有輔助特徵的關鍵詞
11.　　若無，則合併後填入 *MergerResultArr2[BackSearchC]*
12.　　若已合併後一語篇段落，則跳出
13.輸出結果陣列 *MergerResultArr3[n]*

圖 3-4 向後合併線索詞比對及合併演算法

步驟 3：判斷是否已合併完成，並進行語篇連貫關係標記

3.2.3 輔助特徵及特殊單一線索詞比對

此階段我們總共設定了四種輔助特徵及兩種特殊線索詞比對，共分為 4 個步驟：

步驟 1：進行連續 Nd 及 Neu 詞彙之比對及合併

根據我們的觀察，當文本中有某些具有特定詞性的詞彙，連續出現於語篇片段中，則通常這些語篇之間都具有特定的結構關係，因此，我們設定了兩種詞性進行比對。第一種是承接關係的時間連續性，我們利用詞性標記裡的時間詞來輔助我們辨識並標記此種關

係；第二種是並列關係的數詞標示慣例，其演算法如下：

輸入：第二階段之結果陣列 *MergerResultArr3[n]*
　　由待處理文本所形成之長度為 *n* 的陣列 *InputContextArr[n]*
輸出：第三階段之結果陣列 *MergerResultArr4[n]*

1. 若比對連續時間詞，則 *KeyWor*←*"Nd"*
2. 若比對連續數詞定詞，則 *KeyWord*←*"Neu"*
3. *FOR i=1 TO n*
4. 　若 *MergerResultArr3[i]*尚未合併為語篇段落，則
5. 　　若 i >1 則
6. 　　從 *InputContextArr[i]*及 *InputContextArr[i-1]*取出候選詞性進行比對
7. 　　若出現連續連續的 *KeyWord*，則
8. 　　*StartPoint*←*i*
9. 　　*MergerStr*←進行語篇合併
10. 　否則
11. 　　從 *InputContextArr[i]*及 *InputContextArr[i+1]*取出候選詞性進行比對
12. 　　若出現連續連續的 *KeyWord*，則
13. 　　*StartPoint*←*i*
14. 　　*MergerStr*←進行語篇合併
15. 若 i==n 或次一比對語篇已合併，則
16. 　EXIT

&&搜尋並列句群第一分句應合併的位置

17.*FOR i = StartPoint TO 1 STEP -1*
18. 若 *MergerResultArr3[i]*中的語篇段落包含 *InputContextArr[StartPoint]*，則
19. 　*MergerResultArr4[i]*←(*MergerStr* 與 *MergerResultArr3[i]*進行合併)
20.若已合併完成，則
21. 輸出結果陣列 *MergerResultArr4[n]*
22.否則
23. 　繼續步驟 2

圖 3-5 連續特定詞性詞彙之比對及合併演算法

步驟 2：進行解證關係標點符號之比對及合併

某些標點符號也具有輔助語篇標記的功能，因此我們以引號(：)作為輔助解證關係的辨識及標記工作，其演算法請參閱步驟 4。

步驟 3：進行相似句之比對及合併

我們採用鄭守益與梁婷[‘05]的中文句子相似度計算模組，此模組以聚合規則相似度和組合規則相似度來計算中文句子的相似程度。使用兩個句子中所含的詞彙之同義或近義詞，並以改良式編輯距離計算的方法，設計新的權重配置比例、候選句篩選原則，來計算聚合語義的相似度。同時，也使用全域匹配(Global Alignment)及局部匹配(Local Alignment)的策略，求取兩句在詞性序列性質上的結構相似度。我們從實驗語料庫中抽出 3000 對分句進行測試，其結果如下：

表 3-4 中文相似句實驗範例

編號	前分句	後分句	Sim
1	刀魚說生命的顏色是白色的	蚯蚓說生命的顏色是紅色的	1.00
2	久之則漸似矣	久之則愈似矣	1.00
3	法名傳縈	字雪個	1.00
4	能捉的都被捉了	該殺的都被殺了	1.00
5	自一以分萬	自萬以治一	1.00
6	錯開順序	顛倒方向	1.00
7	有一點不凡	有一點叛逆	1.00
8	第一是人文之美	第二是人格之美	1.00
9	先是綠色的葉片	後是白色的花朵	0.84
10	從以前的希特勒、史達林	到近代的馬可仕、哈珊	0.77

由上表觀察，編號 1~8 為並列例句，9~10 為承接例句。我們在實驗中亦發現，相似度大的句子幾乎都為並列結構，只有極少數例句為承接。因此，本系統將相似度高的分句優先判定為並列，本步驟不適用於比對句間關係。

另外，由下圖之結果我們將相似值(Sim)的門檻值訂為 0.48，這個數值可以達到資料涵蓋率 80.45%，正確率 83.88%。

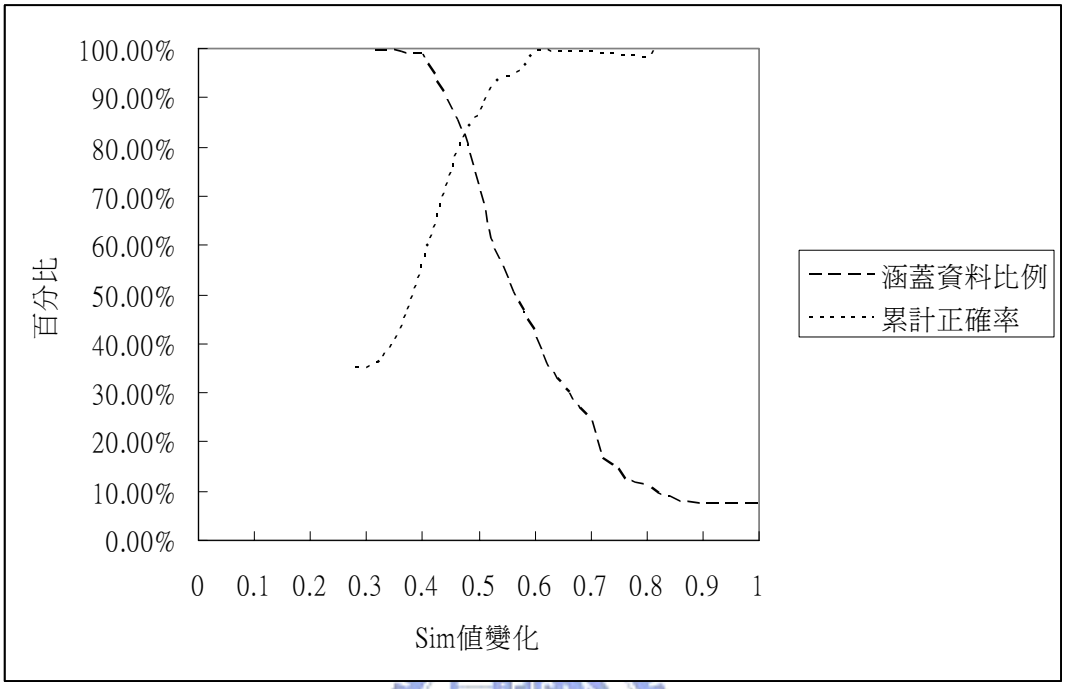

圖 3-6 中文相似句測試圖

本步驟之演算法如下所示：

輸入：上一步驟之結果陣列 *MergerResultArr4[n]*
　　由待處理文本所形成之長度為 *n* 的陣列 *InputContextArr[n]*
輸出：本步驟結果陣列 *MergerResultArr5[n]*
相似度比對門檻值 Sim←0.48

1.*FOR i=1 TO n*
2.　若 *MergerResultArr3[i]*尚未合併為語篇段落，則
3.　　若 i >1 則
4.　　從 *InputContextArr[i]*及 *InputContextArr[i-1]*取出候選詞性進行比對
5.　　若出現連續連續的 *KeyWord*，則
6.　　*StartPoint*←*i*
7.　　*MergerStr*←進行語篇合併
8.　　否則
9.　　從 *InputContextArr[i]*及 *InputContextArr[i+1]*取出候選詞性進行比對
10.　　若出現連續連續的 *KeyWord*，則
11.　　*StartPoint*←*i*
12.　　*MergerStr*←進行語篇合併
13.　若 *i==n* 或次一比對語篇已合併，則
14.　　*EXIT*
&&搜尋並列句群第一分句應合併的位置
15.*FOR i = StartPoint TO 1 STEP -1*
16.　若 *MergerResultArr4[i]*中的語篇段落包含 *InputContextArr[StartPoint]*，則
17.　　*MergerResultArr5[i]*←(*MergerStr* 與 *MergerResultArr4[i]*進行合併)
18.若已合併完成，則
19.　輸出結果陣列 *MergerResultArr5[n]*
20.否則
21.　繼續步驟 4

圖 3-7 相似的語篇片段比對演算法

步驟 4：進行特殊線索詞之比對及合併

我們在語料的觀察中發現，有些線索詞的詞性與一般研究中所發現的線索詞並不一樣，例如本研究提出兩種特殊線索詞。這兩種線索詞的共同特性是，都出現在語篇片段的末尾，涵蓋範圍比一般的線索詞要大；不同之處則在於連結的方向，一個向前，一個往後。因此，我們將分為兩個階段來比對這兩種線索詞，並將涵蓋範圍擴

大 3 個語篇分段或段落。

第一種為列舉線索詞，此種線索詞的連結方向往前，所連結之語篇連貫關係為並列，屬性值為：(-1,1,0)，僅適用於句內關係的比對，共收錄 5 筆資料。如例句 10 中的「等等」，即可將(C)、(D)、(E)三個語篇片段合併為並列關係。

*例句 10：環保局秘密提前啟用本垃圾場(A)，將垃圾灰燼進場掩埋(B)，原承諾之八十三年元月十五日啟用前對南港居民做簡報(C)，提出污染防治保證書(D)，及有效管理辦法及罰則**等等**(E)，均未兌現(F)。*

第二種為動詞線索詞，此種線索詞的連結方向往後，所連結之語篇連貫關係為解證，屬性值為：(1,1,1)，共收錄 57 筆資料。如例句 11 中的「宣示」，即可將(B) 與(C)、(D)、(E)、(F) 五個語篇片段合併為解證關係。

*例句 11：西方人士說(A)，這份文件**宣示**(B)，一個歐洲關係新時代已開始(C)，各國將不再相互仇恨(D)，轉而建立夥伴關係(E)，並伸出友誼之手(F)。*

本步驟演算法如下：

輸入：上一步驟之結果陣列 *MergerResultArr5[n]*
　　由待處理文本所形成之長度為 *n* 的陣列 *InputContextArr[n]*
輸出：本步驟結果陣列 *MergerResultArr6[n]*

&&Process_Step1
1. *KeyWordArr[]*←關連詞屬性值為(-1,1,0)之單一線索詞集合
2.*FOR i=1 TO n*
3.　從 *InputContextArr[i]*取出語篇片段與 *KeyWordArr[]*進行比對
4.　若比對成功，則
5.　　*FOR j=i+1 TO Min(n,i+d)*
6.　　若 *MergerResultArr5[j]*尚未合併成段落，則
7.　　*MergerStr*←將 *MergerResultArr5[j]*與 *MergerStr* 合併為[其他關係]
　　MergerResultArr6[i]← *MergerStr* 與 *MergerResultArr5[i]*合併為並列關係
&&Process_Step2
8. *KeyWordArr[]*←關連詞屬性值為(1,1,1)之單一線索詞集合
9.*FOR i=1 TO n*
10.　從 *InputContextArr[i]*取出語篇片段與 *KeyWordArr[]*進行比對
11.　若比對成功，則
12.　　*FOR j=i+1 TO Min(n,i+d)*
13.　　若 *MergerResultArr5[j]*尚未合併成段落，則
14.　　*MergerStr*←將 *MergerResultArr5[j]*與 *MergerStr* 合併為[其他關係]
　　&&搜尋解證句群第一分句應合併的位置
15.　　*FOR k= i TO 1 STEP -1*
16.　　若 *MergerResultArr5[k]*中的語篇段落包含 *InputContextArr[i]*，則
17.　　*MergerResultArr6[k]*←(*MergerStr* 與 *MergerResultArr5[k]*合併為解證關係)
18.輸出結果陣列 *MergerResultArr6[n]*

圖 3-8 特殊線索詞比對及合併演算法

3.3 標記範例與說明

本研究之標記結果可分為可完全標記及無法完全標記兩種情形，茲說明如下：

3.3.1 可完全標記之情況

例句 12：尤其是除了金融與企業行為的管理以外，更是有許多限制與控管是針對個人而來的，例如公司董事與經理人赴大陸投資行為、企業投資的檢舉獎金，以及開放大陸人士來台灣觀光的管理等等。

例句 12 經過斷句及 POS 標記處理後，我們將得到如下資料：

C1：尤其(D) 是(SHI) 除了(P) 金融(Na) 與(Caa) 企業(Na) 行為(Na) 的(DE) 管理(Na) 以外(Ng) ，(COMMACATEGORY)

C2：更(D) 是(SHI) 有(V_2) 許多(Neqa) 限制(Na) 與(Caa) 控管(VC) 是(SHI) 針對(P) 個(Nf) 人(Na) 而(Cbb) 來(VA) 的(T) ，(COMMACATEGORY)

C3：例如(P) 公司(Nc) 董事(Na) 與(Caa) 經理人(Na) 赴(VCL) 大陸(Nc) 投資(VC) 行為(Na) 、(PAUSECATEGORY) 企業(Na) 投資(Na) 的(DE) 檢舉(VC) 獎金(Na) ，(COMMACATEGORY)

C4：以及(Caa) 開放(VC) 大陸(Nc) 人士(Na) 來(VA) 台灣(Nc) 觀光(VA) 的(DE) 管理(Na) 等等(Cab) 。(PERIODCATEGORY)

圖 3-9　例句 12 斷句及 POS 標記結果

接著我們進行語篇標記，其標記流程如下：

表 3-5 例句 12 語篇標記流程

	階段編號	特徵編號	關連詞	合併段落一	合併段落二	連貫關係
標記流程	2	2	以及 Caa	*[C3]*	*[C4]*	並列
	2	2	例如 P	*[C2]*	*D1,([C3]|[C4])*	解證
	2	2	更 D	*[C1]*	*D8,([C2]|D1,([C3]|[C4]))*	遞進

由上述之標記過程，我們可以得到樹狀結構之語篇標記結果如下：

Forward:(Theme:[C1]| Rheme:Elaboration:(Theme:[C2]| Rheme:Coordinate:(Theme:[C3]| Rheme:[C4])))

圖 3-10 例句 12 語篇標記結果

經過語篇符號轉換後，我們將獲得如下之標記結果：

遞進*[Forward]*:(*[C1:*尤其是除了金融與企業行為的管理以外，*]*|解證*[Elaboration]*:(*[C2:*更是有許多限制與控管是針對個人而來的，*]*|並列*[Coordinate]*:(*[C3:*例如公司董事與經理人赴大陸投資行為、企業投資的檢舉獎金，*]*|*[C4:*以及開放大陸人士來台灣觀光的管理等等。*])))*

圖 3-11 例句 12 語篇標記轉換結果

我們將之轉換成樹狀圖，如下所示：

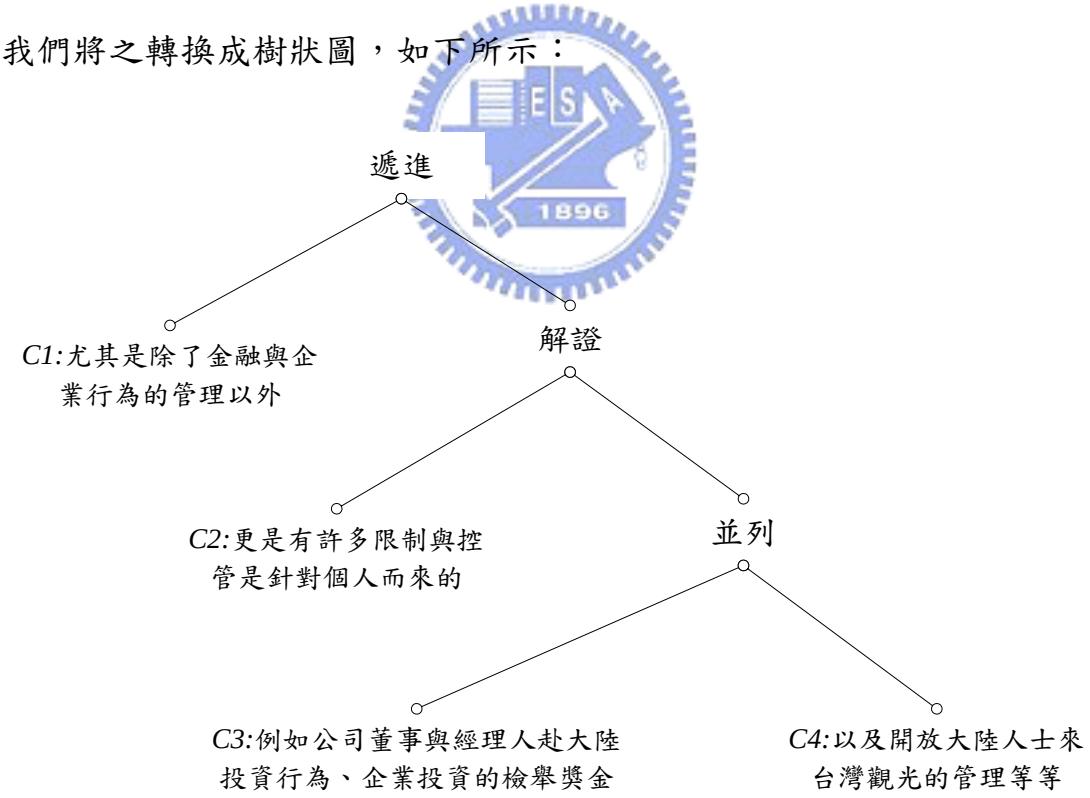

圖 3-12 例句 12 語篇標記樹狀結構

3.3.2 未能完全標記之情況

例句 13：自從雲林縣提出課徵碳稅的構想後，已引起中央與地方政府間的立場對立，財政部長呂桔誠公開表示，不會准予「雲林縣碳稅自治條例」依法備查的同意權。

例句 13 經過斷句及 POS 標記處理後，我們將得到如下資料：

C1：自從(P) 雲林縣(Nc) 提出(VC) 課徵(VC) 碳(Na) 稅(Na) 的(DE) 構想(Na) 後(Ng) ，(COMMACATEGORY)

C2：已(D) 引起(VC) 中央(Nc) 與(Caa) 地方(Na) 政府(Na) 間(Ng) 的(DE) 立場(Na) 對立(VH) ，(COMMACATEGORY)

C3：財政部長(Na) 呂桔誠(Nb) 公開(VHC) 表示(VE) ，(COMMACATEGORY)

C4：不會(D) 准予(VE) 「(PARENTHESISCATEGORY) 雲林縣(Nc) 碳(Na) 稅(Na) 自治(VA) 條例(Na) 」(PARENTHESISCATEGORY) 依法(D) 備查(VH) 的(DE) 同意權(Na) 。(PERIODCATEGORY)

圖 3-13 例句 13 斷句及 POS 標記結果

接著我們進行語篇標記，其標記流程如下：

表 3-6 例句 13 語篇標記流程

標記流程	階段編號	特徵編號	關聯前詞	關連後詞	合併段落一	合併段落二	連貫關係
	1	1	自從 P	已 D	*[C1]*	*[C2]*	承接
	3	7	表示 VE		*[C3]*	*[C4]*	解證

由上述之標記過程，我們可以得到樹狀結構之語篇標記結果如下：

Continue:(Theme:[C1]| Rheme:[C2])@Elaboration:(Theme:[C3]| Rheme:[C4])

圖 3-14 例句 13 語篇標記結果

經過語篇符號轉換後，我們將獲得如下之標記結果：

承接*[Continue]:([C1:*自從雲林縣提出課徵碳稅的構想後，*]|[C2:*已引起中央與地方政府間的立場對立，*])@* 解證*[Elaboration]:([C3:*財政部長呂桔誠公開表示，*]|[C4:*不會准予「雲林縣碳稅自治條例」依法備查的同意權。*])*

圖 3-15 例句 13 語篇標記轉換結果

我們將之轉換成樹狀圖，如下所示：

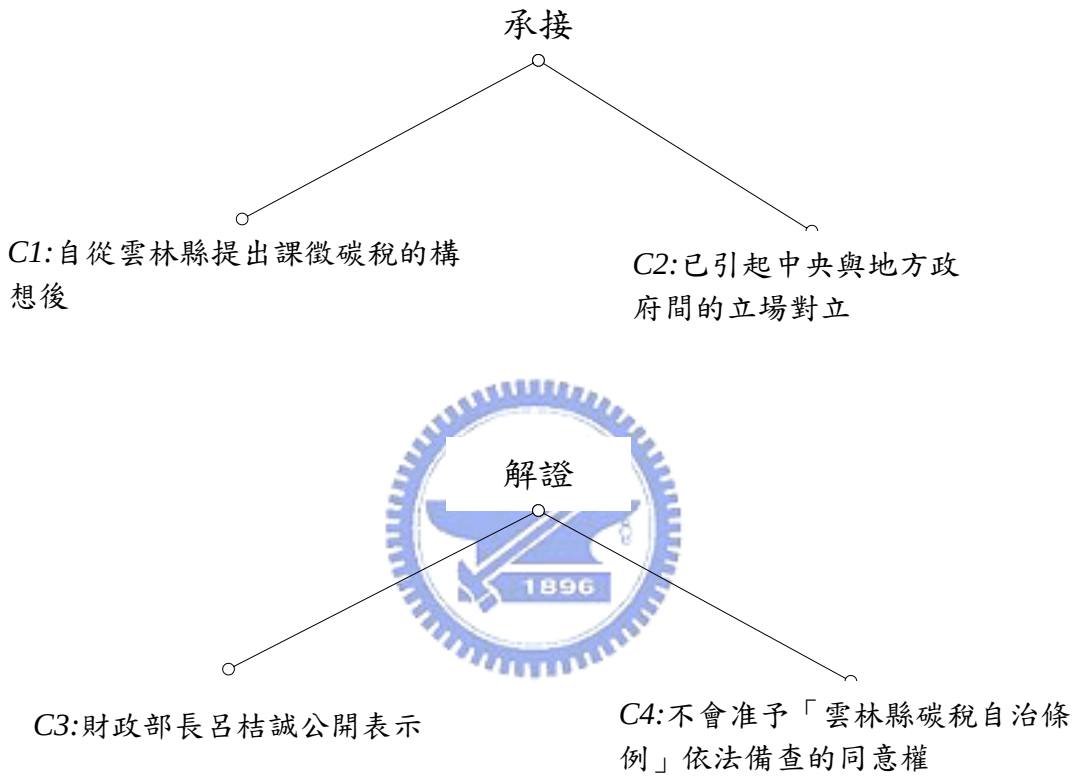

圖 3-16 例句 13 語篇標記樹狀結構

第四章　實驗設計與分析

4.1 實驗語料使用

我們從各家主要的平面媒體電子報，收集 100 篇社論來檢驗系統標記的效能，每篇的字數平均約為 1500 字，內容均為政治、經濟及民生議題，詳細數量及來源如下表所示：

表 4-1 實驗語料明細表

文章來源	篇數	長句數	分句數	平均字數
工商社論	23	536	2423	1548
中時社論	35	1067	3856	1527
自由社論	2	55	214	1558
經濟日報社論	21	522	2348	1548
聯合報社論	19	554	2085	1424
總計	100	2734	10926	

依據 2.1 節的切分規則，我們將上述語料共切分為句內 10926 個分句，句間 2734 個長句。

4.2 實驗結果

我們將系統的標記效能定義為：

表 4-2 可能的標記情況

	應標記	不應標記
正確標記	a	none
錯誤標記	b	c
未標記	d	e

$$\text{系統正確率 } P = \frac{a}{a+b+c} \tag{4.1}$$

$$\text{系統召回率 } R = \frac{a}{a+b+d} \tag{4.2}$$

$$\text{系統篩檢正確率 } FP = \frac{d+e}{c+e} \tag{4.3}$$

在我們的實驗中，句內標記正確率可達到 91%，召回率是 95%，篩檢正確率是 98%。另外，句間標記正確率可達到 86%，召回率是 93%，篩檢正確率是 95%。

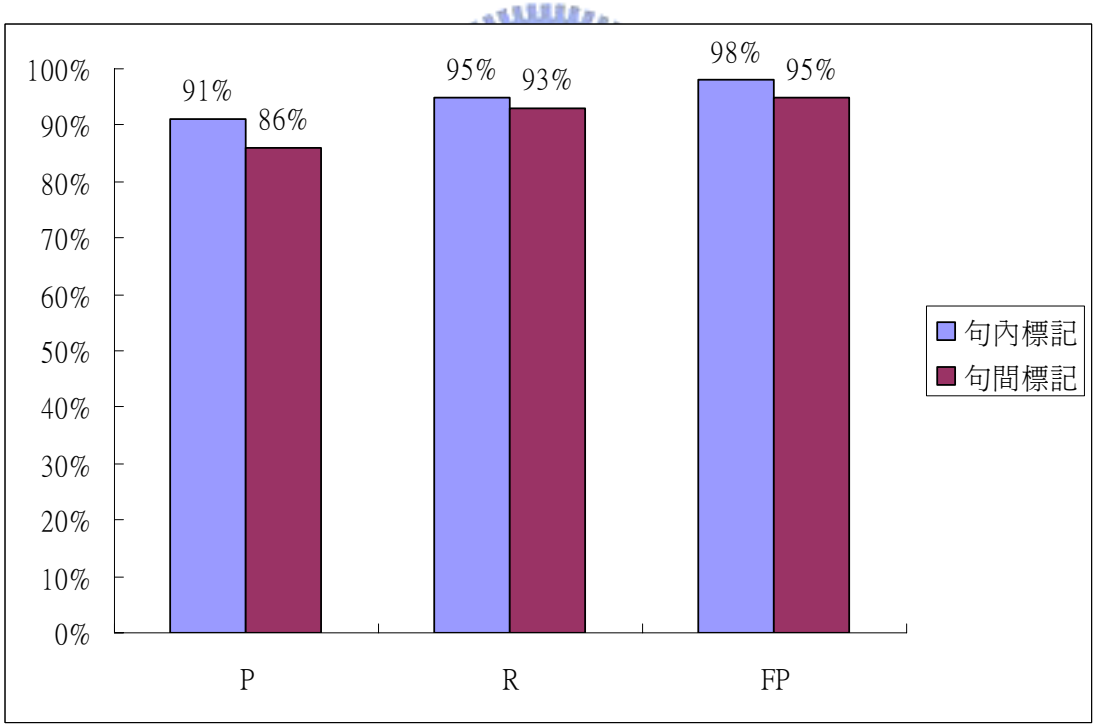

圖 4.1 標記結果

表 4-3 標記情況統計表

標記類別	適用關係	數量	百分比
a	句內	2570	33.16%
	句間	755	37.62%
b	句內	67	0.86%
	句間	56	2.79%
c	句內	183	2.36%
	句間	68	3.39%
d	句內	72	0.93%
	句間	4	0.20%
e	句內	4859	62.69%
	句間	1124	56.00%
總計	句內	8634	100.00%
	句間	14693	100.00%

由表 4-4 可以看出，社論類的文章使用最多的語篇是遞進與轉折，其比例分別是句內 20.27%、27.35%以及句間 13.11%、36.69%，次多者句內和句間有所不同，句內為並列與條件，比例分別為 17.20%、14.05%，而句間則為解證與因果，比例分別為 20.53%、9.53%。相對於句間大量使用解證，句內則較少使用，而句內較常使用的條件語篇，在句間則很少使用。

表 4-4 語篇數量分佈統計表

語篇編號	語篇種類	適用關係	數量	百分比
1	並列	句內	442	17.20%
		句間	65	8.61%
2	承接	句內	99	3.85%
		句間	57	7.55%
3	選擇	句內	85	3.31%
		句間	18	2.38%
4	遞進	句內	521	20.27%
		句間	99	13.11%
5	轉折	句內	703	27.35%
		句間	277	36.69%
6	因果	句內	192	7.47%
		句間	70	9.27%
7	條件	句內	361	14.05%
		句間	14	1.85%
8	解證	句內	136	5.29%
		句間	155	20.53%
9	目的	句內	31	1.21%
		句間	0	0%
總計		句內	2570	100%
		句間	755	100%

為了觀察各種表層特徵在系統進行辨識時，所使用的情形，我們將第 2 章所歸納的特徵表列如下：

表 4-5 表層特徵整理

階段編號	特徵編號	特徵	適用語篇
1	1	成對關鍵字	全部
2	2	單一關鍵字	全部
3	3	連續的 Nd	承接
3	4	連續的 Neu	並列
3	5	上下兩片段相似	並列
3	6	片段最後一個詞使用列舉涵義的詞。	並列
3	7	片段最後一個詞使用解證涵義的詞。	解證
3	8	長句最後一片段的最後一個標點符號出現[：]	解證

由表 4-6 我們可以看出，不論是句內或句間單一線索詞的使用率都是最高的，分別達到 77.43%及 75.63%，而句內使用率偏低的是連續的 Neu 以及最後一個詞使用有列舉涵義的詞，其比例分別只有 0.58%、0.62%，而句間使用率較少的則是片段最後一個詞使用有解證涵意的詞，其比例只有 0.13%，我們認為是因為長句再表達解證義涵時，通常都會使用「：」來表達，其比例有 13.25%。

表 4-6 句內表層特徵使用數量統計

階段編號	特徵編號	適用單位	數量	比例
1	1	句內	398	15.49%
		句間	50	6.62%
2	2	句內	1990	77.43%
		句間	571	75.63%
3	3	句內	32	1.25%
		句間	23	3.05%
3	4	句內	15	0.58%
		句間	10	1.32%
3	5	句內	32	1.25%
		句間	0	0%
3	6	句內	16	0.62%
		句間	0	0%
3	7	句內	87	3.39%
		句間	1	0.13%
3	8	句內	0	0%
		句間	100	13.25%
總計		句內	2570	100.00%
		句間	755	100.00%

4.3 標記情形分析

我們將系統的標記情形以人工進行細分類，並分析其錯誤標記或無法辨識的原因，茲分別說明如下：

表 4-7 標記情形分類

編號	名稱
0	標記正確
1	線索詞出現位置超出比對範圍
2	線索詞連結方向有歧義
3	線索詞功能有岐義
4	相似句門檻值誤差
5	詞性優先順序誤差
6	詞性標註錯誤
7	語篇涵蓋範圍誤差
8	語篇切分錯誤
9	線索詞省略現象
10	未處理的語篇連貫關係
11	標點符號使用習慣
12	未收錄線索詞
13	未收錄輔助特徵

1. 線索詞出現位置超出比對範圍

我們為了提高系統標記的正確率，將比對線索詞的出現位置門檻值設定為 3，因此會造成無法正確比對線索詞的情形發生，此類錯誤的比例在句內為 0.13%，句間則為 0.05%，如例句 1(B)的並列線索詞「也(D)」，因為出現在分段中的第四個位置，因此會造成系統的標記錯誤。

*例句 1：這種作法在公司法上站不住腳(A)，在性質上**也**是延宕*

蹉跎(B)

2. 線索詞連結方向有歧義

在實際的語料中，我們觀察到有些線索詞除了一般的連結方向之外，有時也會因應語意的需求，而出現不同的連結情況，因此會造成無法正確連結語篇片段，此類錯誤的比例在句內為 0.18%，句間則為 0.45%，如例句 2(B)的因果線索詞「因為(Cbb)」的連結方向，在此長句中應為向前連結。

*例句 2：陳木在遭公股解任(A)，大概是**因為**財政部認為他續任董座不利於該公司之健全經營(B)。*

3. 線索詞功能有歧義

此類錯誤有四種主要的類型：第一種是某些線索詞具有可連結詞彙或連結語篇的特性，因此會造成系統的標記錯誤，主要是選擇關係的線索詞，如例句 3(C)的選擇線索詞「或(Caa)」，其連結對象應為前後兩個動詞片語，而不是(B)(C)之間的語篇連貫關係。

*例句 3：即便央行提出浮動利息的建議(A)，在自由利率的體制下(B)，採行**或**不採行仍然得由各銀行自行決定(C)。*

第二種是具有重複出現以表示並列關係的線索詞，有時也可當成數量名詞使用，如例句 4(A)(B)的並列線索詞「一(Neu)」，其語法功能為普通名詞，而並非語篇線索詞：

例句 4：「紅唇族」平均每吃一顆檳榔(A)，其醫療費用支出較不吃檳榔民眾要多出零點一元(B)。

第三種是具有表示時間承接關係的線索詞，當其單獨出現在句內時，有時為連結句間語篇，有時則只是表示語氣的結尾，並未構成承接語篇，如例句 5(A)的承接線索詞「現在(Nd)」，其為連接句

間語篇連貫關係之線索詞，但卻會造成(A)(B)兩個句內語篇的連結錯誤。

*例句 5：**現在**(A)，會計年度第一季尚未結束(B)，雲林縣就擬議課徵「碳稅」(C)，可見地方財政問題之嚴重(D)。*

第四種是因為前方或後方語篇片段中，含有具功能歧義的特殊線索詞，因而造成系統未進行合併，如例句 6「交代(VE)」的功能為動名詞，並不具備連接語篇的功能，卻使得系統合併時產生錯誤。

*例句 6：開發金經營者的作為是否違反公司法三六九條之四、二六七條或其他公司治理原則，也是三、五天就能釐清之事，金管會實應以最快速度調查，並在最短時間內給人民一個**交代**。*

以上四類錯誤的比例在句內為 2.36%，句間則為 3.39%。

4. 相似句門檻值誤差

此類錯誤乃因為相似句門檻值的設定，造成非並列關係的語篇片段被標記，或應為並列關係而未被標記，此類錯誤的比例在句內為 0.45%，句間則為 0.05%，如例句 7 (A)(B)應為並列關係，但因為其比對結果未達門檻值，而未能成功標記：

例句 7：這般無限上綱的行政裁量空間(A)，這般無所不管的「積極管理」(B)，比當年的「戒急用忍」還要可怕(C)。

5. 詞性優先順序誤差

設定詞性的優先順序，雖能幫助我們減低錯誤率，但還是有部份語篇片段，無法完全適用。其錯誤主要是線索詞詞性優先順序的誤判，造成系統無法正確比對，此類錯誤的比例在句內為 0.09%，句間則為 0.05%，如例句 8(B)的線索詞「連(Cbb)」其優先順序大於真正的遞進關係線索詞「甚至(D)」，而造成功標記錯誤：

例句 8：針對最近事故連連(A)，甚至連金管會都放話不排除約談其負責人(B)。

6. 詞性標註錯誤

此錯誤為中研院斷詞系統的 POS 標記錯誤所導致，此類錯誤的比例在句內為 0.04%，句間則未發現，如例句 9(B)的詞彙「一意想以後現代」被標成「一(Neu)　意想(b)　以後(Nd)　現代(Nd)」，造成系統誤以「以後(Nd)」將(A)(B)合併成承接關係：

例句 9：但台灣若想東施效顰(A)，一意想以後現代解構民族榮光云云，可得三思(B)。

7. 語篇涵蓋範圍誤差

我們將單一線索詞及特殊線索詞的涵蓋範圍門檻值設定為 3，因此會造成無法正確連接語篇的情形發生，此類錯誤的比例在句內為 0.59%，句間則為 2.29%，如例句 10(A)的轉折線索詞應將(A)(B)(C)(D)四個語篇片段合併，但卻只合併了(A)(B)(C)：

例句 10：雖然我們可以自我安慰(A)，東亞自由貿易區因成員眾多關係繁雜(B)，彼此利害、立場又很難統一(C)，完成困難度頗高(D)；但在全球性區域經貿組織不斷出現壓力下(E)，以及中、日分別積極推動與區內國家或東協建立經貿合作關係的激烈競爭(F)，出現突破性結果的可能性不能排除(G)。

8. 語篇切分錯誤

我們在切分的過程中發現有以下兩種錯誤發生。

第一種是包含在一對引號(「」)之中的語句，有時是一個完整的語篇片段，應被系統分析及標記，但有時只是某語篇片段的一個名詞或形容詞組，如例句 11「」所包含的語句乃形容忿懣與積怨的形容詞組，並非完整之語篇片段：

例句 11：還夾雜著對大規模企業「本縣拉屎，他處下蛋」的忿懣與積怨

第二種錯誤是由於語料中有時會夾雜表示金額的數字，其中會使用「,」來斷開數字，在切分的過程中，為避免非中文字的雜訊干擾，因此會將所有的非中文符號全部轉成全形來處理，因此會造成系統的誤判，而將數字間的分隔符號斷成語篇片段，如例句 12 的「8,878」、「7,404」兩個數字，將造成此種錯誤。

例句 12：但 1987 年每人 GDP，香港則高達 8,878 美元居首、新加坡 7,404 美元居次。

9. 線索詞省略現象

在實驗結果的觀察中，我們發現語篇片段之間，有時會有省略線索詞的現象。這種現象在社論這種文類中似乎特別常見，此類現象的比例在句內為 7.22%，句間則為 3.89%。如例句 13(C)便是省略了線索詞「並且」，因此造成系統無法將(B)(C)兩個片段連結成並列關係。

例句 13：情勢如此發展(A)，不但出乎蔡英文意料之外(B)，對以照顧地方民眾基本利益起家的民進黨政府而言(C)，亦是相當難堪(D)。

10. 未處理的語篇連貫關係

由於我們的研究僅處理第二章所定義的九種語篇連貫關係，因此會出現不在處理範圍內的語篇片段出現，此類現象的比例在句內為 36.63%，句間則為 46.79%。如例句 14(A)(B)兩個片段應為指代關係：

例句 14：這個事件可能模糊了焦點或偏離主題(A)，對台灣經濟發展無甚幫助(B)。

11. 標點符號使用習慣

在寫作中文文章之時，有時會因為構句較為複雜或語氣停頓的因素，而標上標點符號，但這些標點有時並非一個完整的語義表達片段，因此會造成系統無法標記，此類現象的比例在句內為15.40%，句間則為 0.1%。如例句 15 (A)中的「可惜」應可與(B)片段合併，而(B)片段也應可與(C)片段合併。

*例句 15：**可惜**(A)，讓積穢現形的(B)，不是像李子春這樣的體制內改革者(C)，而竟然是媒體的狗仔隊(D)，這更顯示司法改革的內部動力實在太過微弱(E)。*

12. 未收錄線索詞

我們雖然在 2.4 節時，做過初步的線索詞探勘，但是由於我們的收集範圍不夠，以及因為我們的實驗是採開放測試，因此會出現未收錄線索詞影響系統標記的現象，此類現象的比例在句內為2.85%，句間則為 3.09%。如例句 16(E)中的「乃至」並未被系統收錄，以致無法將(D)(E)合併為遞進關係：

*例句 16：我們建議政府儘快明白宣示政策的走向(A)，如此一來(B)，類似的爭議才有可能停歇(C)，對立、抗爭(D)，**乃至**衝突的危機才有可能化解(E)。*

13. 未收錄輔助特徵

有某些較為複雜的語篇片段，需要更多的輔助特徵才能進行標記，如例句 17(C)(D)與(E)(F)兩個語篇段落，應為並列的兩個相似片段群，但是由於目前相似句比對只針對連續的分句，對於句群之間的比對無法處裡。此類現象的比例在句內為 0.58%，句間則為 2.14%。

例句 17：地方財政苦不堪言(A)，但是(B)，各縣市政府、縣市議會(C)，卻蓋得一棟比一棟金碧輝煌(D)，縣市政府與議會首長(E)，公務轎車一輛比一輛高級豪華(F)。

由上之說明，我們統計出本次實驗中各項標記情況的數量及比例，茲表列如下：

表 4-8 句內標記情形數量統計表

編號	名稱	適用單位	數量	百分比
0	標記正確	句內	2570	33.16%
		句間	755	37.62%
1	關聯詞出現位置超出比對範圍	句內	10	0.13%
		句間	1	0.05%
2	關聯詞連結方向有歧義	句內	14	0.18%
		句間	9	0.45%
3	關聯詞功能有岐義	句內	183	2.36%
		句間	68	3.39%
4	相似句門檻值誤差	句內	35	0.45%
		句間	1	0.05%
5	詞性優先順序誤差	句內	7	0.09%
		句間	1	0.05%
6	詞性標註錯誤	句內	3	0.04%
		句間	0	0.00%
7	語篇涵蓋範圍誤差	句內	46	0.59%
		句間	46	2.29%
8	語篇切分錯誤	句內	24	0.31%
		句間	2	0.10%
9	關聯詞省略現象	句內	560	7.22%
		句間	78	3.89%
10	未處理的語篇連貫關係	句內	2839	36.63%
		句間	939	46.79%
11	標點符號使用習慣	句內	1194	15.40%
		句間	2	0.10%
12	未收錄關聯詞	句內	221	2.85%
		句間	62	3.09%
13	未收錄輔助特徵	句內	45	0.58%
		句間	43	2.14%
總計		句內	7751	100%
		句間	2007	100%

第五章　結論

本論文提出並實際製作了一個中文語篇自動標記系統，經實驗數據的分析顯示，能正確的標記出並列、遞進、轉折等九類語篇連貫關係。本論文從研究、設計到製作，可以歸納出幾個主要的成果與貢獻：

1. 針對語料中的語篇線索詞進行研究與觀察，使資訊科學研究者可以更了解線索詞在實際語料中的分布特性，有助於後續研究的進行。

2. 利用語篇的分布特性進行初步的探勘，並驗證我們所提出的抽取演算法之效能，的確可以幫助我們抽取出可堪使用的語篇線索詞。

3. 運用可擴充的規則模組，成功的在實際語料中標記出九種語篇連貫關係，可節省大量人工標記的時間成本。

4. 完成中文語篇自動標記系統之雛形，並可運用於中文作文自動批改系統，以辨識學生的語篇運用能力。

本論文的後續研究有下列幾個方向：

1. 未知線索詞的自動擴充

可利用同義詞或近義詞，搭配連結強度來自動抽取更多的線索詞，以提高系統的資料涵蓋率。

2. 輔助特徵的研究

由於語篇的結構有時十分複雜，因此需要找尋更多的輔助特徵，來協助系統標記語篇。

3. 未知語篇的定義與研究

可進行更多位之語篇的定義與研究，以利提高系統的資料涵蓋率。

4. 建立語篇自動辨識模型

可利用機器學習及建立語義概念網路的方式，來幫助系統辨識語義的轉折，並可利用統計模型來進行語篇的自動辨識。

參考文獻

田小琳(1984)，中學教學語法系統提要（試用），北京人民教育出版社

黃國文(1988)，語篇分析概要，編著，北京商務印書館

程祥徽、田小琳(1989)，現代漢語，台北書林書店

周國正（1993）語法句群與篇章句群，語文建設通訊，41 期

胡壯麟(1994)，語篇的銜接與連貫，上海外語教育出版社

林孝璘(2002)，台灣華文逗句號研究，國立新竹師範學院臺灣語言與語文教育研究所，碩士論文

楊遠(1962)，標點符號研究，香港天健出版社，頁 3，15-31

曹逢甫（1995），主題在漢語中的功能研究（A functional study of topic in chinese:The first step toward discourse analysis）:謝天蔚譯，北京語文出版社

鄭守益,梁婷, 中文句子相似度之計算與應用,第十七屆自然語言與語音處理研討會, Tainan, Taiwan, 2005 Proceedings of ROCLING XVII pp. 113-124.

Allen, J., (1995), Natural Language Understanding, 2nd, Benjamid/Cummings.

Burstein, J., Kukich, K., Wolff, S., Lu, C. and Chodorow, M. (1998), “Enriching Automated Scoring Using Discourse Marking”, In the Proceedings of the Workshop on Discourse Relations & Discourse Marking, Annual Meeting of the Association of Computational Linguistics, August, 1998.

Chan, W. K., Lai, B. Y., Gao, W. J. and T'sou, K.,(2000), "Mining Discourse Markers for Chinese Textual Summarization." In Proceedings of the 6th Applied Natural Language Processing Conf. and the North American Chapter of the Association for Computational Linguistics. Workshop on Automatic Summarization,Seattle, Washington, 29 April to 3 May.

Dong, Z. D., Dong, Q., (1999), “HowNet“, http://www.keenage.com

Grosz, B. J. and C: L. Sidner,(1986), “Attention, intentions, and the structure of discourse”, Computational Linguistics, vol. 12, no. 3, pp. 175-204.

Grosz, B. J., A. K. Joshi, and S. Weinstein,(1995), “Centering: a framework for modeling the local coherence of discourse”, Computational Linguistics, vol. 21, no. 2, pp. 203-225.

Guenthner, F., H. Lehmann, and W. Schonfeld,(1986), “A theory for the

representation of knowledge", JBM Journal of Research and Development, vol. 30, no. 1, pp. 39-56,January.

Hirschberg, J. and D. Litman,(1993), "Empirical studies on the disambiguation of cue phrases" Computational Linguistics, vol. 19, no. 3, pp. 501-530.

Hobbs, J. R. Literature and Cognition, (1990), CSLI Press, Stanford, California.

Hovy, E. and E. Maier.,(1995),"Parsimonious or profligate: How many and which discourse relations?", Technical report, University of Southern California.

Hsu, W. L., Y. S. Chen, and Y. K. Wang,(1998), "A context sensitive model for concept understanding", Proceedings of Third Int. Conf. on Information-Theoretic Approachesto Logic, Language, and Computation.

Hearst, M. A.,(1997),"TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages", Computational Linguistics, 23 (1), 33-64, March.

Halliday, M. A. K. & Hasan, R., Coherence in English, London: Longman.,1976

Kamp, H.,(1981), "A theory of truth and semantic representation: Formal Methods in the Study of Language", MC TRACT 135, J. A. G. Groenendijk, T. M. V. Janssen, and M. B. J. Stokhof (Eds.), Amsterdam, p. 277.

Lin, K. H. C. and V. W. Soo,(1993), "Toward discourse-guided theta-grid parsing for Mandarin Chinese --a preliminary report", Proceedings of ROCLING Il, pp. 259-270.

Li, S., Zhang, J.,(2002), "Semantic Computation in Chinese Question-Answering System", Journal of Computer Science and Technology, 17(6): 933

Marcu, D., (2000), "The rhetorical parsing of unrestricted texts: A surface-based approach.", Computational Linguistics 26: 395-448

Sadao K., Makoto N.,(1994), "Automatic Detection of Discourse Structure by Checking Surface Information in Sentences", COLING , pp.1123-1127

Smadja, F., (1993), "Retrieving collocations from text: Xtract", Computational Linguistics, 19(1): 143-177

Tomohide S. and Sadao K.,(2005), "Automatic Slide Generation Based on Discourse Structure Analysis", In Proceedings of Second International Joint Conference on Natural Language Processing (IJCNLP-05),

pp.754-766, Jeju Island, Korea.

Wang, Y. K., Y. S. Chen, and W. L. Hsu, (1998), “Empirical study of Mandarin Chinese discourse analysis: an event-based approach,” to appear in 10th IEEE Int’l Conf. on Tools with Artificial Intelligence (ICTAI’98).

Wolf, F. and Gibson, E.,(2005),”Representing discourse coherence: A corpus-based analysis”, Computational Linguistics, 31(2): 249-287.