# 以語料為基礎的中文語篇結構關係自動標記

鄭守益 吳典松 梁婷
國立交通大學資訊科學與工程研究所
{gis93540, gis92807, tliang }@cis.nctu.edu.tw

## 摘要

語篇分析是文本理解中一項不可缺乏的工作，藉以釐清文章的論題或邏輯結構。本論文提出以語料為主的語篇分析方法，針對並列、承接、遞進、選擇、轉折、因果、條件、解證、目的等九種常見語篇類別，進行表層特徵收集及擴展，並制定標記規則，建立有效的自動標記程序。我們使用中研院平衡語料庫 3.0 版中的報導、傳記日記、散文、信函、評論、說明手冊等文類，共 7265 篇作為探勘語料，進行線索詞、連續詞性序列、特殊標點符號等語篇特徵之探勘。在實驗中，我們使用 100 篇平均字數為 1500 字的報紙社論進行效能評估，在句內的語篇標記部份，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%。我們相信此語篇標記的研究，有助於自動問答、作文評分、閱讀測驗、摘要和簡報系統等應用。

## 1. 緒論

語篇是指在特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[4]。國外語篇分析的相關研究，多以連貫理論為基礎[6]。在連貫理論中，一個語篇是由許多語篇片段組成，有不同的連貫關係，例如：評估、因果、描述、解釋、排列等等。Wolf 和 Gibson[15]曾發表以語料為主的語篇連貫研究，並提出圖形表示法以描述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的影響，用線索片語來當作語篇中的重要結構元素[8]，而不強調語用學理論及世界知識。例如 Sadao 和 Makoto [11]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[9] 提出所謂的重心理論，探討一段文章的內在結構中，其參照延續性及言談本身特點之間的關聯。

在中文語篇的研究中，黃國文[2]提出語篇特性分為銜接與連貫兩種。銜接關係可以用語法或詞彙連結，而連貫關係則是以語義做為連結。另一方面，胡壯麟[4]將語篇特性分為指稱性、結構銜接及邏輯連接。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。相對的，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。此外，程祥徽和田小琳[1]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。文獻上，中文語篇的計算模型甚少被提出。王元凱等人[14]曾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事件，用以表現語義重心的轉移。另外 Chan 等人[7] 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。

中文語篇的切分目前尚未有明確的定義，因此在本論文中，我們將依據 Marcu [10]所提出的定義，將語篇片段標記為不重疊的文本片段，並分別對分句間的句內語篇關係(以逗點做為分句的切分界線)和長句間的句間語篇關係(以冒號、句號、問號及驚嘆號為切分界線)提出有效的標記程序。

## 2. 語篇連貫關係分類

我們依據 [3] 和[15]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。在本論文中，我們暫不探討沒有明顯表層特徵的總分及連鎖關係，只對如下常見的九種語篇連貫關係提出自動標記程序：

表 1 語篇連貫關係類別

| 語篇類別 | 定義 |
|---|---|
| 並列關係 | 指表達幾件相關的事件，但彼此並不構成因果關係，也沒有語氣或語義上的轉折。 |
| 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，以及依事件發生的空間順序來進行敘述的事件。 |
| 選擇關係 | 含有從幾件事物中進行選擇的語義。 |
| 遞進關係 | 在連續片段中，具有後一個片段比前一個片段的語義層次更進一層關係的語篇視為遞進關係。 |
| 轉折關係 | 指前一片段的語義與後一片段相對或相反。 |
| 因果關係 | 使用兩個或兩個以上的片段來說明事件的原因及其結果。 |
| 條件關係 | 前一片段假設一種情況或提出一種條件，後一片段說明如果實現的話會產生的結果。 |
| 解證關係 | 前一片段提出一種看法、道理、事實、現象，後一片段加以解釋、說明、補充、引申的語篇。 |
| 目的關係 | 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。 |

## 3. 語篇線索詞探勘

語篇線索詞的探勘是以中研院平衡語料庫 3.0 版中的敘述型語料來進行的，主要步驟為現有線索詞收集、線索詞詞性篩選、成對線索詞組探勘、單一線索詞探勘及輔助特徵探勘。

### 3.1 線索詞收集與篩選

我們以「現代漢語」[3]所列出的語篇線索詞來當作查詢詞，在探勘語料中進行更多線索詞的收集。由於詞性影響到詞彙的語義或語法角色，因此這些詞在語料中的詞性若為下列的詞性，將不視為線索詞。

{Na, Nb, Nc, SHI, T, VA, VC, VCL, VD, V, VH, VJ, Nf}

### 3.2 成對線索詞組探勘

成對線索詞組的探勘包括四個主要步驟：

*步驟 1：設定抽取線索詞之範圍及位置*

從探勘語料中，我們隨機選取 24 個線索詞組進行例句搜尋，共收集 2300 個分句，由其統計分布得到線索詞多為語篇片段中的第 1 到第 12 個詞。

另一方面，從探勘語料的例句統計分布，我們也觀察到成對線索詞組，不論是句內或句間的語篇片段連結距離多為 3，亦即若詞組中的前詞在第一語篇片段，則其搭配詞多在接續的三個片段中出現。

*步驟 2：設定線索詞出現位置之權重*

語篇中的線索詞是否具有連結功能與其出現的位置有關，因此藉由觀察步驟 1 中線索詞分佈位置的統計資料，我們發現線索詞的分布近似於函數 $\frac{1}{x^3}$ (其中 $x$ 為線索詞的出現位置，$1 \le x \le 12$)，因此我們可以此函數來作為計算線索詞組連結強度時的權重，並進行正規化，使其權重值介於 0 與 1 之間。設線索詞在分句中出現的位置共有 $j$ 個($1 \le j \le 12$)，則由線索詞出現在分句內的位置分佈，我們可得到一正規化常數為 $D$：

$$D = \sum_{j=1}^{12} \frac{1}{j^3} = 1.2 \qquad (1)$$

因此，若線索詞出現在第 $j$ 個位置，其權重即為

$$w_j = \frac{1}{1.2 j^3} \qquad (2)$$

*步驟 3：計算線索詞組之連結強度 k*

我們以線索詞組$(T_h, T_i)$一起出現在語篇片段之間的頻率標準差倍數，作為其連結強度[12]，其中$T_h$為給定的線索前詞，$T_i$為$T_h$的第$i$個搭配詞，其計算公式如下：

$$k_i = \frac{f_i - \overline{f}}{\sigma} \tag{3}$$

其中搭配詞$T_i$出現在語篇片段的位置$j$ $(1 \le j \le 12)$之頻率$f_i$定義為：

$$f_i = \sum_{j=1}^{12} f_{i,j} w_j \tag{4}$$

其平均頻率$\overline{f}$以及標準差$\sigma$的計算公式如下：

$$\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f_i \quad ; \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f_i - \overline{f}\right)^2} \tag{5}$$

步驟 4：線索詞組篩選

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

### 3.3 單一線索詞探勘

中文語篇的成對線索詞，有時也可單獨出現，例如：

*例句 1：他(**不但**)吃米飯(A)，**也**吃牛排(B)。*

其中「不但」可以省略。另外中文線索詞在書寫的過程中，常會省略關聯前詞或後詞，如例句 2 中的「如果」和例句 3 中的「所以」常被省略：

*例句 2：(**如果**)我們這麼做，**可能**會導致環境的破壞。*

*例句 3：**因為**情勢如此變化，(**所以**)我們不得不做這樣的決定。*

因此，在語篇連貫關係辨識的過程中，有必要進行單一線索詞的收集及探勘工作。我們以檢驗成對線索詞組的例句進行辨識篩選，收集了 309 個單一線索詞。

### 3.4 輔助特徵探勘

我們參考 [1]及 [13]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，將視這些語篇片段具有「承接關係」。

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，將視這些語篇片段具有「並列關係」。

c. 語篇片段的末尾若出現標點符號「：」，將可判定其次一語篇片段為

「解證關係」。

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」。

## 4. 辨識及標記執行步驟

我們使用中央研究院的線上中文斷詞系統[1]，進行文本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。其步驟如下表所示：

表 2 語篇連貫關係辨識及標記步驟

| 階段 | 步驟 | 執行動作 |
|---|---|---|
| 一 | **成對線索詞組比對** | |
| | 1 | 以兩個分句為單位進行成對線索詞組比對。 |
| | 2 | 將比對後具有語篇連貫關係之兩個分句合併。 |
| | 3 | 判斷是否已合併完成，並進行語篇連貫關係標記。 |
| 二 | **單一線索詞比對** | |
| | 1 | 進行向前連結之單一線索詞比對及合併。 |
| | 2 | 進行向後連結之單一線索詞比對及合併。 |
| | 3 | 判斷是否已合併完成，並進行語篇連貫關係標記。 |
| 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。 |
| | 2 | 進行解證關係標點符號之比對及合併。 |
| | 3 | 進行相似句之比對及合併。 |
| | 4 | 進行特殊線索詞之比對及合併。 |

### 4.1 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：

*步驟 1：以兩個分句為單位進行成對線索詞組比對*

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。

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

*步驟 2：將比對後具有語篇連貫關係之兩個分句合併。*

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，其遞增變數為語篇連結門檻值$d$，處理同一片段的合併問題。第二個部分稱為橫向合併，其遞增變數為語篇片段數量$n$，以處理相鄰片段的合併問題。我們將依循以下規則：

規則 1：縱向合併時，同一片段若可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。

如例句 4 中語篇片段 (B)的線索詞「因為」可與線索詞「所以」連結，但 (C)與(E)卻同時出現該線索詞，因此，我們根據規則 1 取距離最小者，將(B)(C)兩個片段合併。

*例句 4：這種盲識她不覺得有必要去澄清(A)，**因為**知情的人知道真相為何(B)，**所以**她認為夫妻檔是利多於弊(C)，**因為**兩人志趣相投(D)，**所以**不但目標一致(E)，**而且**團結力量一定大(F)。*

規則 2：橫向合併時，相鄰片段若連續形成相同的語篇連貫關係時，則合併成為同一語篇在同一階層。

如例句 5 中語篇片段(C)、(D)、(E)、(F)經依循規則 1 進行縱向連結之後，將以線索詞「或」分別連結成 3 個選擇語篇(CD、DE 、EF)，然後，我們再依規則 2 進行橫向連結，將語篇段落合併在同一階層為一個大的語篇段落。

*例句 5：今年的元宵燈節十分熱鬧(A)，展出的花燈無奇不有(B)，**或**以造型取勝(C)，**或**以作工服人(D)，**或**色彩華麗(E)，**或**聲光迷人(F)。*

規則 3：橫向合併時，相鄰片段若連續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。

如例句 6 中語篇片段(A)和(B) 經依循規則 1 進行縱向連結之後，以線索詞組「不但…還」合併為遞進關係，(A)和(C)又以「因為…所以」合併為因果關係，然後，我們再依規則 3 進行橫向連結，將這兩個語篇段落以不同語篇合併在不同階層。

*例句 6：**因為**她**不但**嘴巴很壞(A)，**還**喜歡打人(B)，**所以**我們都不喜歡她(C)。*

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。

**4.2 單一線索詞比對**

使用單一線索詞來辨識語篇連貫關係時，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，我們設計三種屬性：

1. 連結方向

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。

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，並在設定的位置門檻值內，則設定為 0；若出現在語篇片段末尾，則設定為 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。

此階段所指的「單一線索詞」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，若在句內省略前詞而僅單用後詞，則其連結方向多為向前連結，反之亦然。此外也會有複合線索詞的出現，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)。*

在分句(B)中出現兩個單一線索詞，一個是「也」表示並列關係，另一個是「因為」表示因果關係。因此，我們將以下規則進行第二階段比對及合併：

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，則依以下詞性優先順序決定：

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 9 之線索詞「然而」。

*例句 9：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。*

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 10 之線索詞「即使」。

*例句 10：**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。*

### 4.3 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：進行連續時間詞及數詞定詞詞彙之比對及合併*

如前所述，對連續語篇片段，我們可利用時間詞來輔助辨識承接關係；用數詞標示並列關係。

*步驟 2：進行解證關係標點符號之比對及合併*

我們以冒號(：)作為輔助解證關係的辨識及標記。

*步驟 3：進行相似句之比對及合併*

我們採用之前所設計的中文句子相似度計算模組[5]，進行語篇中並列關係的探勘。此模組考量中文相似句中的語義和結構的相似度。從訓練語料中我們抽出 3000 對分句進行測試，部份結果如下：

表 3 中文相似句實驗範例

| 編號 | 前分句 | 後分句 | 相似值 |
|---|---|---|---|
| 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。我們將相似值的門檻值訂為 0.48，這個數值可以達到資料涵蓋率 80.45%，正確率 83.88%。

*步驟 4：進行特殊線索詞之比對及合併*

我們在語料中發現有兩種特殊線索詞。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。

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 11 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 11：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 12 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 12：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)。*

**4.4 標記範例與說明**

我們將輸入的文本自動標記出相應的語篇連貫關係。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。表 4 和圖 1 分別為語篇連貫關係符號表和標記結果範例。

表 4 語篇連貫關係標記符號表

| 符號 | 說明 |
|---|---|
| **@** | 做為分隔語篇段落的界線。 |
| **( )** | 標示語篇結構的左右邊界。 |
| **\|** | 表示在同一層的語篇片段。 |
| **D#,** | 標示語篇連貫關係之編號。 |

| | |
|---|---|
| **[]** | 標示語篇片段的左右邊界。 |
| **C#** | 標示分句語篇片段在整個長句裡的順序。 |
| **S#** | 標示長句語篇片段在整個文章裡的順序。 |
| **Theme** | 標示語篇連貫關係中的第一個語篇片段。 |
| **Rheme** | 以 Rheme 標示語篇連貫關係中的其他語篇片段。 |

*D8,( Theme: [C1:*尤其是除了金融與企業行為的管理以外，*]|D1,( Theme: [C2:*更是有許多限制與控管是針對個人而來的，*]| Rheme:D4,( Theme: [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，*]| Rheme: [C4:*以及開放大陸人士來台灣觀光的管理等等。*]))*

圖 1 語篇標記結果範例

我們將之轉換成樹狀圖，如下所示：

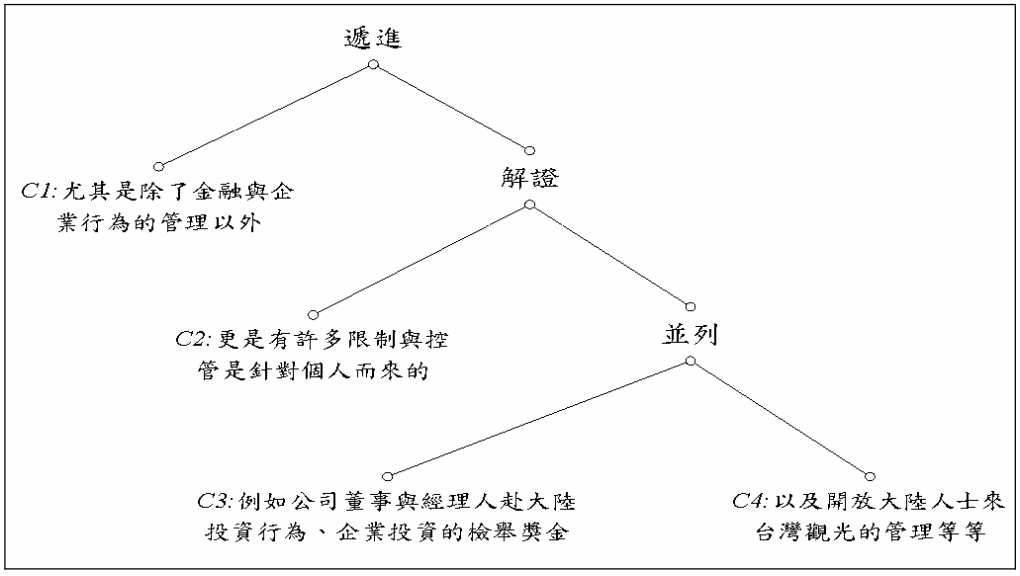


圖 2 語篇標記樹狀結構範例

## 5. 實驗設計與分析

我們分別從工商時報、中國時報、自由時報、聯合報、及經濟日報等主要的平面媒體電子報，收集 100 篇社論來檢驗本論文所提的標記程序效能，每篇的字數平均約為 1500 字。我們將系統的標記效能依表 5 的標記情況定義為：

表 5 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 |

$$\text{標記正確率 } P=\frac{a}{a+b+c} \tag{6}$$

$$\text{標記召回率 } R=\frac{a}{a+b+d} \tag{7}$$

$$\text{系統篩檢正確率 } FP=\frac{d+e}{c+e} \tag{8}$$

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%。

表 6 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% |

另外由表 6 可以看出，社論類的文章使用最多的語篇是遞進與轉折。次多者在句內是並列與條件，而句間則為解證與因果。相對於句間大量使用解證，句內較常使用的是條件語篇。至於在語篇的特徵上，我們也觀察到單一線索詞的使用率不論是在句內或句間的語篇辨識上都是最高的，分別有 77.43%及 75.63%。使用率偏低的特徵在句內是連續的數詞定詞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%。在句間語篇辨識上使用率較少的則是片段最後一個詞使用有解證涵意的線索詞，其比例只有 0.13%。

## 6. 結論與未來研究

本論文提出並實作一個中文語篇自動標記系統，經實驗數據的分析顯示，能有效地標記出並列、遞進、轉折等九類語篇連貫關係。其後續研究有下列幾個方向：

1. 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. 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. 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. 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


[1] 田小琳，”中學教學語法系統提要（試用）”，北京人民教育出版社，1984。

[2] 黃國文編著，”語篇分析概要”，北京商務印書館，1988。

[3] 程祥徽、田小琳，”現代漢語”，台北書林書店，1989。

[4] 胡壯麟，語篇的銜接與連貫，上海外語教育出版社，1994。

[5] 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.

[6] Allen, J., Natural Language Understanding, 2nd, Benjamid/Cummings, 1995.

[7] Chan, W. K., Lai, B. Y., Gao, W. J. and T'sou, K., "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, 2000.

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

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

[10] Marcu, D., “The rhetorical parsing of unrestricted texts: A surface-based approach.”, Computational Linguistics 26: 395-448, 2000.

[11] Sadao K., Makoto N. , “Automatic Detection of Discourse Structure by Checking Surface Information in Sentences”, COLING , pp.1123-1127, 1994.

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

[13] Tomohide S. and Sadao K., "Automatic Slide Generation Based on Discourse Structure Analysis", In Proceedings of Second International Joint Conference on Natural Language Processing (IJCNLP-05), Jeju Island, Korea, pp.754-766, 2005.

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

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