

語法規律的抽取及普遍化與精確化的研究

Grammar Extraction, Generalization and Specialization

謝佑明 楊敦淇 陳克健

中央研究院資訊科學研究所
{morris, ydc, kchen}@iis.sinica.edu.tw

摘要. 相較於傳統PCFG的CNF處理，在本篇論文中，我們提出二元化句法規則產生模式。並且深入探討其語法普遍化與精確化方法對中文剖析器的影響。實驗設計從中研院中文句結構樹中依不同的語法抽取原則，抽取出不同的語法規律集合，來剖析三份測試語料並評估結果。觀察結果試著去尋找出有效的語法普遍化及精確化方法，得到覆蓋率高且精確的句法規則，以加強中文剖析器的剖析效能。剖析精確率的實驗結果，從基本普遍化語法的81.45%增加到精確化語法的86.14%。(關鍵詞: 覆蓋率、語法歧義、句法剖析、語法抽取)

1 緒論

自然語言處理的過程中，句法剖析(parsing)是一個核心處理過程。在過去研究中，剖析器(parser)利用從樹庫(treebank)中訓練出的probabilistic context-free grammar(以下簡稱PCFG)，對句子剖析是很常用的技術。在英文的部份，因為有大量的英文樹庫資料，利用PCFG剖析英文句子都會有不錯的效果，現有資料顯示約可至九成，還進一步的做到詞彙化剖析(lexicalized parsing)[6]。相對於有限的中文句結構樹庫，非詞彙化剖析(unlexicalized parsing)是一個研究的開始。在本篇論文中，研究如何從有限的中研院中文句結構樹庫(Sinica Treebank)中，抽取最佳的PCFG，使得抽取出的語法規律有較佳的覆蓋率(coverage)及較低的語法歧義(ambiguity)。我們同時建立一個符合需求的中文剖析器，從中文句結構樹庫抽取出不同的語法規律集合，來剖析三份測試語料並評估結果。從這些實驗中，觀察結果試著去尋找出有效的語法普遍化(generalization)及精確化(specialization)方法，得到覆蓋率高且精確的句法規則，以加強中文剖析器的剖析效能。

在最後章節中，我們也探討到未來進一步的研究與實驗方向，如何整合句法與語意訊息讓剖析器有效解決句法結構歧義的問題。

1.1 中文句結構樹庫簡介

中央研究院中文句結構樹庫由中研院詞庫小組於2000年開始建制。目前的版本是 2.0 (9個檔案)，其中包含有 38,944 句結構樹與240,979詞。每一個句結構樹都有標註”詞、細詞類、語法結構與語意角色”訊息。一般看到的句結構樹是只有標註語法結構訊息，較少有語意角色訊息。Chen等[4]提到了語意角色的定義與考量，想利用單純句法限制去定義中文的關係是困難的。在中文句結構樹中，特別的是加上了語意的訊息，意指在瞭解到每一個成員(constituent)與其它成員的關係。舉例如下：

他 叫 李四 撿 球。 Ta jiao Li-si jian qiu. "He asked Lisi to pick up the ball." <i>S(agent:NP(Head:Nhaa:他) Head:VF2:叫 goal:NP(Head:Nba:李四) theme:VP(Head:VC2:撿 goal:NP(Head:Nab:球)))</i>
--

圖1. 他叫李四撿球

圖1表示完整的中文句結構樹內容，標示了詞組結構(phrase structure)規則及語法與語意關係。對於句中標示每個詞的語法詞類意義，詳細定義與例子說明可參照[1]的技術報告手冊。

1.2 研究方法

二個操作策略：(1)將語法規律普遍化 及 (2)將語法規律精確化。普遍化的結果是增加語法覆蓋率，但同時可能產生的副作用就是降低了精確性與增加歧義性。一般的普遍化方法是放鬆詞類限制及增加不同的詞組規律，反之精確化的方式是增加詞類及特徵(feature)限制，使語法規律更精確。由於普遍化和精確化的操作互為制衡，常有過與不及的情形發生，如何拿捏控制找到有效的普遍化及精確化的方法是本論文探討的重點。

在語法規律精確化方面，Charniak[2,3]與Johnson[8]都有相同的結論，就是，對於非詞彙化PCFG在每個詞組節點加上子節點成員或是上層節點的特徵訊息，對剖析的精確率會有所提升。Zhang等[12]也提到，在詞組節點上加入上層詞組詞類與出現位置順序同樣也會對剖析結果有一些改進。Klein和Manning[9]與Collins[6]則提到對CFG加上中心語，也是有所幫助。

我們研究的方法就是試著從訓練語料中，依不同的語法抽取原則與語言學特性，抽取出不同的PCFG，再對測試語料進行剖析與結果評估。觀察結果並調整語法抽取原則，再次產生PCFG。從反覆的實驗中，找出較佳的句法規則來。

在普遍化的方法中，詞類限制放鬆在 Chen和Hsieh[5]有基本的研究結果。在語法覆蓋率增加方面，本文提出二元化句法規則產生方法，3.2節有詳細說明。在精確化方面，我們嘗試了不同的附加特徵，其中以中心語及最左成份特徵對剖析結果有最明顯的改進。其它特徵的實驗結果也會在第6節討論中說明。

2 系統設計架構

整個系統流程包含句法抽取、剖析與結果評估三大部份，圖2表示了整個的系統操作流程。對於每個部份的處理內容與方式，將分三小節來說明。

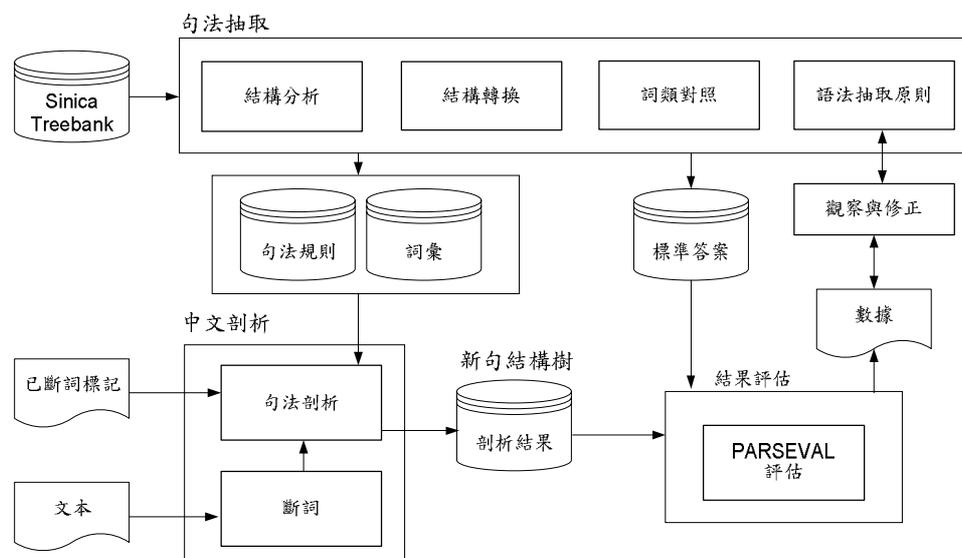


圖2. 系統架構圖

2.1 句法抽取

主要動作就是從樹庫中依語法抽取原則抽取出我們要的句法規則，並進行部份規則的調整與計算出相關的機率值，目的在供剖析器作為中文剖析的依據。說明如下：

- 1 從中研院中文句結構樹庫整理出完整且可以處理的樹
- 1 依語法抽取規則自動抽取出我們要的句法規則

- l Charniak[3]提到很直覺的統計公式，對每條句法規則計算其機率值。

$$\hat{P}(N^i \rightarrow X^j) = \frac{C(N^i \rightarrow X^j)}{\sum_k (N^i \rightarrow X^k)}$$

- l 建立虛規則列表，以解決Top-Down模式的中文剖析問題。如：起始符號為 S 的問題。加入 S' → VP、S' → NP...等。
- l 去除 A → A 的規則，避免出現遞迴(recursive)的情形，導致剖析器出現問題。

2.2 中文剖析器

PCFG的句法分析中，最廣泛採用的為LR與Chart演算法。而Earley's Algorithm是在英文剖析中，最常用被拿來實作的一個演算法[7]。它的處理特性為上到下(top-down)、預測(predictive)與左到右(left-right)，當應用在中文時，我們保有原有的predict、scan與complete這三個分析動作，同樣是分析到有節點到S'為止。為了符合我們對文本輸入格式¹與產生句結構樹結構²的需求，對剖析器增加了資料輸入格式分析與處理虛規則的能力。

2.3 評估模型

最常用來評估樹結構括號(bracket)與詞組標記(label)好壞的評估模型是PARSEVAL[7]。如果樹結構括號有錯，詞組詞類標記位置也會是錯的。所以，當括號的精確率愈高，連帶詞組標記的精確率也會愈高。我們採用如下評估項目與公式如下：

- l 詞組標記精確率 LP(Labeled Precision)

$$LP = \frac{\# \text{ label correct constituents in parser's parse of testing data}}{\# \text{ label constituents in parser's parse of testing data}}$$

- l 詞組標記召回率LR(Labeled Recall)

$$LR = \frac{\# \text{ label correct constituents in parser's parse of testing data}}{\# \text{ label constituents in treebank's parse of testing data}}$$

- l 詞組標記效能評估LF(Labeled F-measure)

$$LF = \frac{LP * LR * 2}{LP + LR}$$

- l 括號精確率BP(Bracketed Precision)

$$BP = \frac{\# \text{ bracket correct constituents in parser's parse of testing data}}{\# \text{ bracket constituents in parser's parse of testing data}}$$

- l 括號召回率BR(Bracketed Recall)

$$BR = \frac{\# \text{ bracket correct constituents in parser's parse of testing data}}{\# \text{ bracket constituents in treebank's parse of testing data}}$$

- l 括號效能評估BF(Bracketed F-measure)

$$BF = \frac{BP * BR * 2}{BP + BR}$$

另外，我們增加幾個觀察變數，以加強瞭解句法規則有用程度。說明如下：

RC-Type：句法規則覆蓋率—類型(type coverage of rules)

RC-Token：句法規則覆蓋率—頻率(token coverage of rules)

PA：句子可以剖析的比率

¹ 輸入格式為： 他(Nh) 叫(VF) 李四(Nb) 撿(VC) 球(Na) ，(COMMACATEGORY)

² 輸出格式為：#1.1..[O] S(NP(Head:Nh:他)|Head:VF:叫|NP(Head:Nb:李四)|VP(Head:VC:撿|NP(head:Na:球)))#，(COMMACATEGORY)

- PC：句子剖析中，樹結構全對的比率
- LF-1：僅對可剖析的句結構樹進行 LF 的評估。
- BF-1：僅對可剖析的句結構樹進行 BF 的評估。

3 語法普遍化

Chen和Hsieh[5]抽出的句法規則，它是保有完整的詞組規則，我們稱為長詞組規則。對於長詞組規則的低覆蓋率問題，我們歸納出幾個可以對語法普遍化的方法，除了將詞類做粗細的變化，也試著將長詞組規則進行二元化的轉換。

3.1 詞類簡化

規律右邊詞類簡化意義在使一條規律能通用在更多不同的詞彙上，換句話說一個詞能夠應用到更多條句法規則上。詞類愈細，可以用到的句法規則就愈少，反之就愈多。而過度的詞類簡化會造成精確率的下降，太細的詞類，雖然提升剖析精確率，卻也降低語法覆蓋率，使得無法剖析的句子增加。詞類簡化例子如圖3所示，至於詞類簡化的原則，可以參照附錄1的詞類對照表。

原始詞組規則	$S \rightarrow \text{agent:NP time:Dd manner:Dh Head:VA4}$
細詞類(Sinica Treebank詞類)	$S \rightarrow \text{NP Dd Dh VA4}$
粗詞類(斷詞標記詞類)	$S \rightarrow \text{NP D D VA}$
最粗詞類	$S \rightarrow \text{NP D D V}$

圖3. 詞類簡化例子

從Chen和Hsieh[5]所做的實驗結果來看，很明顯地證明上述的看法。從細詞類層級到最粗的單一詞類層級，抽出的語法覆蓋率從82%增到92%；但是語法歧義也從每一詞類平均適用規律132條增加到835條；而且對訓練語料的剖析正確率不升反降平均從70%降到62%。剖析結果精確率和句法規則的覆蓋率與精確性有相當程度的互動關係。因此，本論文探討如何提高句法規則的覆蓋率與精確性以增加剖析效能。其中面臨的一個最重大問題就是，句法規則愈細，可能因訓練資料量的不足，而無法產生高覆蓋率且精確的語法規律。於是，對長詞組規則的簡化是有必要的。

3.2 二元化句法規則

從過去的研究中可以知道，句法規則的操作直接影響到剖析的結果，因此句法規則的呈現方式在句結構樹中是一個可調整的變數。長詞組規則其機率值較低，且覆蓋率也會較低。在縮短詞組規則的研究中，Manning[7]提到Chomsky normal form(CNF)的處理方式，主要在增加剖析器的剖析速度。經由CNF處理後的語法與原來的語法是完全相同的，並不會改變句法規則的覆蓋率。因而，我們提出的二元化句法規則方法，其擁有CNF增加剖析速度的特性，又有增加語法覆蓋率與句法延伸性³。其語法上的意義在於，同一詞組下的附加成份(adjuncts)在二元化的語法規則中可以任意出現或不出現。

CNF句法的基本定義格式： $A \rightarrow a$ 或 $A \rightarrow B C$ 。A、B與C為非終端符號(non-terminal)，a為終端符號(terminal)。 \rightarrow 左邊的成員我們稱之為LHS(left-hand side)， \rightarrow 右邊成員稱之為RHS(right-hand side)。在RHS中，最多為二個成員。非終端符號在句結構樹中為詞組詞類，如VP、NP、S...等；終端符號為詞的詞類，如Nh、Nb...等。簡單的CNF在英文的做法是分析每條詞組規則，從最左開始，反覆向右，兩兩切割規則，產生新的非終端節點，並標記Rx，直到每條規則都符合上述的定義。二元化句法規則與CNF的差異，可以從表1得知。

³ 所謂延伸性，如表 1(b)的二元化句法規則所示， $VP \rightarrow D VP$ 與 $VP \rightarrow D VA$ 這二條句法規則中的 D 即擁有延伸性。如 $S \rightarrow NP D D D VA$ 在二元化後，也是只有這二條句法規則，D 是可多可少的成份。

表格1. CNF與二元化句法規則比較

原始句法規則：(以斷詞標記詞類層級為例) $S \rightarrow NP\ D\ D\ VA$	
(a) CNF 結果	(b) 二元化句法規則
$S \rightarrow NP\ R0$	$S \rightarrow NP\ S$
$R0 \rightarrow B0\ R1$	$S \rightarrow D\ S$
$R1 \rightarrow B0\ B1$	$S \rightarrow D\ VA$
$B0 \rightarrow D$	
$B1 \rightarrow VA$	

二元化句法規則的產生方式與CNF類似，不同處是在分析每條詞組時，從最右節點開始向左，兩兩切割詞組，不僅考慮到區域(local)的新二元化成員與詞組詞類，也考慮到原本整體(global)詞組詞類與中心語訊息。新規律的左邊成份Rx的操作，R必須維持原有詞組名外，x部份表示俱有附加特徵區別多樣化，在不同的語法抽取原則下，x也會選有不同的特徵值，而抽出的句法規則也就不同，這都會影響句法剖析的結果。舉例說明，我們對現有中文句結構樹庫中每一顆樹進行二元化的轉換，例1為原來的句結構樹；例2為二元化不加任何特徵；例3 Rx為詞組加中心語特徵；例4 Rx為詞組加成員特徵。

$S(NP(Nh:我們) D:常常 D:一起 VA:上學)$	(例1)
$S^-(NP^-(Nh:我們) S^+(D:常常 S^+(D:一起 VA:上學)))$	(例2)
$S_{VA}^-(NP_{Nh}^-(Nh:我們) S_{VA}^+(D:常常 S_{VA}^+(D:一起 VA:上學)))$	(例3)
$S_{NP,S}^-(NP_{Nh}^-(Nh:我們) S_{D,S}^+(D:常常 S_{D,VA}^+(D:一起 VA:上學)))$	(例4)

註：S-為終節點(原節點)，S+中間節點(虛節點)。

另外，在二元化句法規則中，我們建立特徵的觀念。特徵會因語法抽取的定義不同而有所改變，一般用在語法精確化的處理上。如果不加任何特徵只處理詞類的部份，就是最原始的語法普遍化。第4節將討論如何運用有語法支持的特徵。

4 語法精確化

上節討論的語法普遍化，雖然覆蓋率會有所提升，但句法剖析的精確率卻因語法歧義的增加而降低。語法精確化目的在提高句法剖析的精確率，句法規則覆蓋率又不會降低太多。主要研究內容為語法精確化的操作方式與特徵的選擇。

4.1 詞類精確化

在斷詞標記的詞類中，中研院詞庫小組[1]對詞類有完整的定義與說明。基本上相同語法行為的詞被歸為同一詞類，詞類精確化目的是針對某些詞類在句法剖析時易造成混淆的情況，進行更細的分類，使句法剖析有較好的結果。例如從樹庫我們觀察到DE詞類在PCFG中的問題。在中文詞中，詞類標記為DE的有“得、地、之、的”四種。而從各別的語法特性，進一步瞭解到不應該在PCFG的規則中混在一起。根據觀察與分析得到可以分成三組：“的、之”、“得”與“地”。因此，我們將DE這類精確化後成DE、DE1與DE2，舉例說明如下：

<p>的、之：炙熱的太陽、言下之意。</p> <p>得：說得容易做得難</p> <p>地：高興地唱歌</p>
--

4.2 特徵限制

本論文研究了許多不同組合的特徵限制，並歸納出提了二個最重要的特徵限制：中心語訊息與詞組成份訊息，目的都在使語法歧義減少，增加剖析器對歧義結構的鑑別能力。在特徵的限制上，中心語是很重要的訊息。在中文句結構樹中，每一種詞組NP、VP、S、GP...等都各自不同中心語詞類，其語法規律也不同。因此，將詞組的中心語詞類加入二元化句法規則是非常重要，相信對於句法剖析的精確率會有所提升。例子5說明了中心語加入的結果。

句結構樹：VP(D:終於|VC:到|Di:了)

二元化：VP_{VC}⁻(D:終於|VP_{VC}⁺(VC:到|Di:了))

(例5)

另外，參考Johnson[8]提到在詞組上加入子節訊息能有較佳的結果，於是，我們提出二元化過程中，對每個二元化詞組加入了最左成份的詞類訊息。第5節實際的實驗結果顯示中心語及最左成份的特徵對語法規率精確性的確有幫助，且不會降低太多語法覆蓋率，對整體的剖析結果有顯著的改善。

5 實際操作

本實驗方式在圖2中有一個語法抽取原則處理部份，依不同的方法，產生不同的句法規則，剖析器再依此句法規則進行測試語料的剖析，並評估剖析的結果。反覆的實驗與觀察，找出最好用的語法抽取原則出來。

訓練語料與測試語料的介紹：中研院中文句結構樹庫的訓練語料統計情形，包含37,889句數、235,669詞數與6.22每句平均詞數訊息。我們從不同領域中取三份測試語料(Sinica、光華雜誌與南一課本)進行實驗。表說明了這三份語料的句長分佈、句數與難易程度。

這三份測試語料句法難易程度有所不同。Sinica語料與訓練語料句法較為相近的，程度屬適中。南一課本為小學正式上課教材，句子不會太長，程度屬簡單。光華雜誌為市面上出版書目，句法嚴謹，程度屬較難。

表2. 三份測試語料內所包含句子的資訊

測試語料	難易度	0-5 詞句數	6-10 詞句數	11 詞以上句數	總計
Sinica	適中	612	385	124	1,027
光華	較難	428	424	104	956
南一	簡單	1,159	566	25	1,750

5.1 詞類普遍化

Chen和Hsieh[5]的實驗顯示一但語法覆蓋率超過90%，增加訓練資料量，很難快速的增加語法的覆蓋率。因此，在有限的訓練語料下，操作語法規律的普遍化是有其必要性。對於詞類普遍化的做法，本文3.1節中已有相關實驗的結果討論。在我們的實驗中，詞類對照的部份，同樣參照附錄1的對照原則，例子如圖3所示。分別對這三份測試語料實作的結果如表3所示，以最粗詞類層級的語法覆蓋率98.303%為最高，但其剖析效能只有74.29%，且最粗詞類並無法對詞性有效區分；另外，細詞類層級的語法覆蓋率約91%，明顯較其它二組低，剖析效能為76%，也不是最好的；因此，粗詞類層級(斷詞標記詞類)的實驗結果有其操作空間，往後實驗目標是將長詞組規則二元化以提高語法覆蓋率，並試著從語法精確化的過程中，在覆蓋率減少的容許範圍內，提升剖析效能。

表3. 詞類普遍化實驗

	細詞類層級			粗詞類層級			最粗層級		
	Sinica	光華	南一	Sinica	光華	南一	Sinica	光華	南一
RC-Type	63.015	73.194	71.806	74.681	82.154	79.877	83.269	86.581	86.330
RC-Token	86.358	91.195	92.928	93.223	96.033	96.506	97.289	98.303	98.679

PA	96.88	98.22	98.06	99.32	99.69	99.66	100	100	99.49
LF	76.84	75.11	82.42	82.80	77.82	84.04	81.22	74.29	80.97
BF	82.19	81.22	86.70	87.92	84.27	88.56	86.00	80.87	85.47
LF-1	79.31	76.48	84.05	83.37	78.07	84.33	81.22	74.29	81.38
BF-1	84.83	82.69	88.41	88.52	84.53	88.86	86.00	80.87	85.91
PC	46.25	40.48	57.89	58.13	46.44	63.66	59.69	39.85	61.14

5.2 二元化句法規則 vs. 長詞組規則

從前面章節瞭解二元化句法規則的優點，而長詞組規則為語法抽取的最原始規則。有關二元化的例子，可參照3.2節說明，本節就實作與結果進行討論。

實驗結果如表4所示。二元化句法規則擁有99%的高語法覆蓋率，對於測試語料的句子，幾乎都可以剖析出來，但是，LF-1與LF反而下降了約0.5%到1.8%。句法規則不加任何特徵限制，則將可以剖析出所有的樹，PA值大於99.9%，但過度普遍化卻造成了剖析的精確率的下降。接下來的研究就在於如何去提升剖析的精確率。

在剖析速度的分析方面，二元化句法規則的剖析時間明顯的較長詞組規則快，這是句法剖析過程中，二元化句法規則可以比長詞組規則較快去過濾掉不必要的句法候選規則，減少運算量。

表4. 長詞組規則與二元化句法規則實驗結果

	(a)長詞組規則			(b)二元化句法規則		
	Sinica	光華	南一	Sinica	光華	南一
RC-Type	74.681	82.154	79.877	96.154	94.657	94.761
RC-Token	93.223	96.033	96.506	99.590	99.362	99.548
PA	99.32	99.69	99.66	99.90	99.90	100
LF	82.80	77.82	84.04	81.45	76.16	83.50
BF	87.92	84.27	88.56	87.71	83.58	88.38
LF-1	83.37	78.07	84.33	81.53	76.24	83.50
BF-1	88.52	84.53	88.86	87.79	83.67	88.38
PC	58.13	46.44	63.66	52.29	42.36	60.00

5.3 特徵限制：中心語

二元化過程中，我們對詞組部份加入了中心語特徵，能保有原詞組的語法特性，又可避免句法規則的過度普遍化。舉例來說：“一輛大型玩具機車”的長句法規則為 $NP \bar{a} DM A Na Na$ ，在二元化後為 $NP^- \bar{a} DM NP^+$ ， $NP^+ \bar{a} A NP^+$ 與 $NP^+ \bar{a} Na Na$ ，加入中心語特徵限制後為 $NP_{Na}^- \bar{a} DM NP_{Na}^+$ ， $NP_{Na}^+ \bar{a} A NP_{Na}^+$ 與 $NP_{Na}^+ \bar{a} Na Na$ 。從表5的結果中知道，在加入中心語特徵後，語法覆蓋率下降0.3%左右，而剖析的精確率確提升1.4%左右。證明了在二元化句法規則中加入中心語的特徵是有用的。

表5. 中心語特徵限制實驗結果

	中心語		
	Sinica	光華	南一
RC-Type	95.824	94.124	94.456
RC-Token	99.273	99.026	99.334
PA	99.61	99.90	99.94
LF	82.62	77.50	84.52
BF	88.65	84.47	89.02
LF-1	82.95	77.58	84.57
BF-1	89.00	84.56	89.08
PC	60.18	45.40	64.80

5.4 特徵限制：最左成份

另一個特徵限制的實驗，是在詞組上加註左邊成份的特徵。這個特徵的意義是說，成份和成份之間有順序及搭配的限制。例如上節例句‘一輛大型玩具機車’中的附加成份定量詞(DM)、形容詞(A)及名詞修飾語(Na)有順序上的限制，通常定量詞一定是在最外層。同前一節的例子，在詞組節點加入最左成份特徵限制後的二元化句法規則為中心語特徵限制後為 $NP_{Na,DM}^- \hat{a} DM NP_{Na,A}^+$ ， $NP_{Na,A}^+ \hat{a} A NP_{Na,Na}^+$ 與 $NP_{Na,Na}^+ \hat{a} Na Na$ 。表6分別顯示不同的實驗結果：(a)單獨加左成份與(b)加中心語和左成份。

從表6(b)與表4(b)來看語法覆蓋率，因特徵限制的增加，減少至98.5%左右。剖析的精確率則增加2%到4%不等。觀察測試語料的特性，在Sinica增加4.2%，其它光華與南一測試語料分別為2.88%與1.91%。這說明了訓練語料與Sinica測試語料句法結構較為相似，其它則不然。表6(a)與表6(b)的差異在中心語加入與否，從數據看來，加入中心語的PC值有些許提升，整體剖析效能則差不多。

表6. 最左成份實驗結果

	(a) 左成份			(b) 中心語 + 左成份		
	Sinica	光華	南一	Sinica	光華	南一
RC-Type	95.633	94.939	94.900	93.977	92.988	92.557
RC-Token	99.236	99.106	99.345	98.602	98.500	98.810
PA	99.90	99.79	100	99.42	99.79	99.94
LF	85.62	79.82	85.74	85.64	79.04	85.41
BF	89.75	85.64	89.52	89.84	85.29	89.42
LF-1	85.70	79.98	85.74	86.14	79.21	85.46
BF-1	89.83	85.82	89.52	90.37	85.47	89.47
PC	64.65	48.54	65.83	65.73	49.06	66.23

5.5 詞類精確化：“DE” 處理

從4.1的說明中，知道DE可分為三組語法特性。這樣的特徵限制，屬於區域限制，目的在於降低詞類DE造成的混淆。在我們實作中，為了鑑別不同的詞所代表的 DE，因此給予DE不同的編號。例如：“的、之”標示為DE、“得”標示為DE1、“地”標示為DE2。我們設計三種實驗方式 (a)單獨在二元化後加入DE特徵；(b)依續5.3節的研究，再加入DE特徵；(c)依續5.4的研究，再加入DE特徵。實驗結果如表7所示。表7(a)與表4(b)的實驗差別在DE特徵限制的部份，剖析精確率在光華語料中有0.55%的提升，其它則不太明顯。

表7(c)有最佳的剖析結果，表7(b)相較於表5的數據，也只是微幅的提升。因此，從數據看來，DE的加入，或多或少對剖析結果還是有影響的。

表7. ‘DE’ 實驗結果

	(a) DE			(b) 中心語+DE			(c) 中心語+左成份+DE		
	Sinica	光華	南一	Sinica	光華	南一	Sinica	光華	南一
RC-Type	96.161	94.368	94.611	95.829	93.920	94.352	93.982	92.840	92.478
RC-Token	99.590	99.330	99.536	99.273	98.994	99.322	98.602	98.470	98.798
PA	99.90	99.90	100	99.61	99.90	99.94	99.42	99.79	99.94
LF	81.33	76.71	83.59	82.71	77.82	84.62	85.63	79.43	85.50
BF	87.78	84.19	89.00	88.87	84.89	89.67	89.97	85.73	89.98
LF-1	81.41	76.79	83.59	83.04	77.90	84.66	86.13	79.60	85.55
BF-1	87.87	84.28	89.00	89.21	84.98	89.72	90.49	85.91	90.03
PC	52.29	42.47	60.23	60.27	45.40	65.03	65.73	49.06	66.57

6 討論

在實作語法精確化的過程中，我們嘗試了不同的附加特徵，其中以第5節探討的最左成份及中心語特徵對剖析結果有最明顯的改進。在本節我們將一一討論其它的實驗結果。

語法抽取規律：詞組角色“Head/head”的保留。不同於圖3的例子，我們對於詞組角色為Head/head的部份進行保留並抽取規律與實驗。從結果得知，雖然整體剖析效能提高1%左右，但是語法覆蓋率卻降低。因此，在我們第5節的實驗中，以覆蓋率高的詞組角色“Head/head”不保留為實驗基礎。

二元化詞組的變化：原節點不加特徵限制。原節點的意思說明請參考3.2節，在我們最早的實驗中，對於原節點的部份，並不加入任何特徵限制訊息，保有原來詞組詞類，只專注在虛節點的處理上。後來改進對原節點同步處理時，發現整體的剖析效能明顯地提升。因此，我們在實驗過程中，談到語法精確化部份，都是同步對原節點與虛節點處理。

二元化方向。從左向右二元化，在實際剖析上，很明顯地剖析速度比較慢，主要原因是剖析器運用為由上到下的句法規則連結，而剖析方向為從左到右，對一個要處理的詞類，向右二元化的句法規則產生的規則候選集較向左句法規則來的多，剖析運算量也比較多，這就是我們不採用的原因。另外一個二元化的方向，是以中心語為開始點，先向右二元化，再向左二元化。從實際剖析結果來看，並沒有比向左二元化來的明顯較好，且建立的方法又繁雜許多，於是採用不用採它。

特徵限制：加入上層詞組詞類訊息。Johnson[8]與Zhang[12]的內容中提到加入上層詞組詞類訊息可以提升剖析效能。我們實作將上層詞組詞類加入特徵限制後，因語法規律更為精確，覆蓋率明顯地下降，LF-1值確實有些許提高，但從整體剖析效能來看影響不太。因此，該方法並不適合我們使用。

角色普遍化。中文句結構樹的角色，是一個可以嘗試普遍化之處，我們試著觀察哪些角色可以代替詞類或詞組用在語法規律上。以time為例，在time:Dd或time:NP的語法規律相近，可以用time代替達到角色普遍化目的。從剖析結果看來，每一個動作的語法覆蓋率都有些許提升，合併使用多個角色普遍化時提升更多，但是整體對於剖析效能卻影響不太。此外，如果選錯角色進行普遍化，卻造成剖析效能明顯下降。

7 結論與未來研究

本論文中，我們建構一完整的實驗平台，其中包含符合語料輸入輸出格式與處理需求的中文剖析器。剖析器所參考的句法規則，從最原始的長詞組規則到二元化句法規則，從沒有加入任何特徵的語法普遍化到加上特徵限制的語法精確化，每一個動作都有相關的實驗結果評估。從實驗數據得知，建構二元化句法規則的方法與特徵的選擇，對中文剖析有直接的影響，而二元化句法規則則比長詞組規則較好的剖析結果，剖析速度也相對快些。當句法規則愈精確化，其句法規則覆蓋率會降低一些與剖析效能會有所成長；句法規則的普遍化增加了語法歧義，可以利用其它的知識來幫忙解決結構歧義的問題。You和Chen[11]的研究說明了詞與詞相關訊息(word-association)的應用，不僅僅是可以對剖析好的樹結構進行語意角色的指定，也可以用來幫忙解決歧義句結構的問題。因此，在剖析器的設計上，可以進一步的將詞與詞相關的訊息整合進來，也就是用PCFG來提供剖析器句法結構的選擇，再利用句法規則的機率和詞與詞相關的機率來指導結構的挑選。

中心語加左成份似乎並沒有比僅加左成份更好，主要是中心語限制降低了語法的覆蓋率，相信在更大的訓練語料下，有中心語特徵的語法規律會比不採用中心語特徵的語法規律表現得更好。並且經由中心語的限制可以增加剖析的效率，因為和輸入句不相關的規律，也就是說語法規律指定的中心語詞類不在句中出現，此規律可以提早刪除。

從實驗的結果可以得到另一個結論，特定領域的句法規則學習是有必要的，不同領域的文章會有其專門句法表示，比如，新聞類、財經類或教材課文類的文本，句法上彼此會有些許的不同；最好的方法就針對特定領域處理，比如要剖析財經類文章，就用從財經領域抽取的句法規則來剖析，以取得較好的剖析結果。另外，Verdú-Mas、Calera-Rubio與Carrasco[10]提出一個平滑技術(smoothing techniques)是將二種較粗與較細的句法規則混合使用，以提升剖析的能力，這是我們未來可以參考實驗的方向。

致謝

本論文的研究得到國科會計畫編號 NSC93-2752-E-001-001-PAE 及 NSC93-2422-H-001-0004 的部分補助。

參考文獻

- [1] 中央研究院詞庫小組, “中文詞類分析 (三版).” CKIP Technical Report No.93-05.
- [2] Charniak, E. and Carroll, G., “Context-sensitive statistics for improved grammatical language models.” In *Proceedings of the 12th National Conference on Artificial Intelligence*, AAAI Press, pp. 742-747, Seattle, WA, 1994.
- [3] Charniak, E., “Treebank grammars.” In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pp. 1031-1036. AAAI Press/MIT Press, 1996.
- [4] Chen, Feng-Yi, Tsai, Pi-Fang, Chen, Keh-Jiann, and Huang, Chu-Ren, “Sinica Treebank.” *Computational Linguistics and Chinese Language Processing*, 4(2):87-103, 2000.
- [5] Chen, Keh-Jiann and Hsieh, Yu-Ming, “Chinese Treebanks and Grammar Extraction.” the *First International Joint Conference on Natural Language Processing (IJCNLP-04)*, March 2004.
- [6] Collins, Michael, “Head-Driven Statistical Models for Natural Language parsing.” Ph.D. thesis, Univ. of Pennsylvania, 1999.
- [7] Manning, Christopher D. and Schutze, Hinrich, “Foundations of Statistical Natural Language Processing.” the MIT Press, Cambridge, Massachusetts, 1999.
- [8] Johnson, Mark, “PCFG models of linguistic tree representations.” *Computational Linguistics*, Vol.24, pp.613-632, 1998.
- [9] Klein, Dan and Manning, Christopher D., ”Accurate Unlexicalized Parsing.” *Proceeding of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 423-430, July 2003.
- [10] Verdú-Mas, Jose L., Calera-Rubio, Jorge and Carrasco, Rafael C., ”Smoothing Techniques for Tree-k-Grammar-Based Natural Language Modeling.” *IbPRIA* , pp. 1057-1065, 2003.
- [11] You, Jia-Ming and Chen, Keh-Jiann, “Automatic Semantic Role Assignment for a Tree Structure.”the *Third SIGHAN Workshop on Chinese Language Processing*,” July 2004.
- [12] Zhang, Hao, Liu, Qun , Zhang, Kevin, Zou, Gang and Bai, Shuo, “Statistical Chinese Parser ICTPROP.” Technology Report, Institute of Computing Technology, 2003.

附錄1 Syntactic Category Mapping

Level2	Level3	Level4	Nc*	Nc	N	V_2	V_2	V
Caa	Caa	C	Ncd*	Ncd	Ncd	VA*	VA	V
Cab	Cab	C	Nd*	Nd	N	VA2	VAC	V
Cba	Cba	C	Nep	Nep	Ne	VB*	VB	V
Cbaa	Cbb	C	Neqa	Neqa	Ne	VC1	VCL	V
Cbab	Cba	C	Neqb	Neqb	Ne	VC*	VC	V
Cbba	Cbb	C	Nes	Nes	Ne	VD*	VD	V
Cbbb	Cbb	C	Neu	Neu	Ne	VE*	VE	V
Cbca	Cbb	C	Nf*	Nf	N	VF*	VF	V
Cbcb	Cbb	C	Ng	Ng	Ng	VG*	VG	V
D*	D	D	Nh*	Nh	N	VH*	VH	V
Dab	Da	D	Nv1	Nv	N	VH16	VHC	V
DE	DE	DE	Nv2	Nv	N	VH22	VHC	V
Dfa	Dfa	D	Nv3	Nv	N	VI*	VI	V
Dfb	Dfb	D	Nv4	Nv	N	VJ*	VJ	V
Dk	Dk	D	P*	P	P	VK*	VK	V
I	I	I	T*	T	T	VL*	VL	V
Na*	Na	N	V_11	SHI	V	DM	DM	DM
Nb*	Nb	N	V_12	SHI	V	Di	Di	D