利用向量支撐機辨識中文基底名詞組的初步研究

張席維 高照明 劉昭麟

台大資工系 台大外文系 政大資科系
b91083@csie.ntu.edu.tw zmgao@ntu.edu.tw chaolin@nccu.edu.tw

-

摘要

本文實做 Kudo and Matsumoto (2000, 2001)以向量支撐機（SVM）辨識基底名詞組（base NP）演算法。我們以中央研究院中文句結構樹資料庫 Sinica Treebank 3.0 的 80%作爲訓練語料,20%作爲測試語料,並比較以 Sinica Treebank 三種不同的詞性標記集訓練出來的 SVM 的辨識率（簡化標記，精簡標記，及簡化標記的大類）。實驗的結果顯示具備詳細次分類的簡化標記的辨識率最高，在封閉測試的 F-measure 爲 87.43%，初步小規模開放測試的 F-measure 爲 78.79%。詳細次分類的標記集的名詞組辨識率較高的原因是中文某些類別的動詞能夠修飾名詞，因此沒有詳細次分類的詞類標記集無法區別那些類別的動詞可以修飾名詞。與英文日文高達 94%以上的辨識率相比較，SVM 在中文基底名詞組辨識的效果並不理想，我們認爲中研院句法樹的表示法與中文本身的特性是造成辨識率不夠高的主要原因。

1. 前言

名詞組的辨識與標示（NP Chunking）是自然語言處理（NLP）的一個重要研究議題 (Ramshaw and Marcus (1995), Kudo and Matsumoto (2000, 2001))，無論是句法處理中的剖析(parsing)語意處理中的語意角色的標示（semantic role labeling）及篇章處理中的回指(co-reference)與連貫性（coherence），其它領域如資訊檢索(information retrieval)資訊擷取（information extraction）文件探勘（text mining）文件分類，與文件自動摘要都需要名詞組的辨識，例如在資訊檢索中最常被檢索的大都是名詞組（特別是人名，地名，組織名等所謂的 name entity），因此在文件或網頁中自動辨識名詞組並建立索引以方便檢索分類及自動摘要是智慧型資訊處理極爲重要的一環。

一般名詞組的辨識指的是基底名詞組（base NP），也就是將名詞組下面又包含名詞組的複雜名詞組（如關係子句及名詞組並列結構(NP conjunction)）排除在外。目前英文名詞組的辨識正確率可以達到 94%以上(Kudo and Matsumoto (2000, 2001))，但中文名詞組的辨識至今只有少數零星的研究。本文採用監督式機器學習（supervised learning）嘗試以向量支撐機(SVM, support vector machine)透過中研院句法樹庫實做 Kudo and Matsumoto (2000, 2001)所提出的演算法。本研究的主要目的在於（一）探討中文基底名詞組辨識的重要特徵（二）評估各種基底名詞組辨識的 SVM 表示法與其限制（三）從語言學的觀點分析影響中文基底名詞組辨識率的原因。

2. 文獻回顧

在大規模語法樹庫還沒有建立之前，名詞組辨識常將組成名詞組結構的規律透過有限狀態機（finite state machines）去找出符合名詞組的 pattern (Voutilainen (1993)) 或從標記好詞性的語料庫以統計的方式得到（Church (1988)），或結和語言規律和語料庫統計（Chen and Chen (1994)）。自從賓州大學大規模的英文語法樹庫(Penn Treebank)建構完成後 (Marcus, Santorini and Marcinkiewicz (1993)),絕大多數的名詞組辨識研究是以機器學習（machine learning）的方法透過語法樹庫裡面的語法結構及前後語境的特徵得到。運用機器學習辨識名詞組的方法大致可分爲 HMM (hidden Markov model)，transformation-based（Ramshaw and Marcus （1995））, memory-based （Veenstra （1998）, Tjong Kim Sang and Veenstra （1999） Argamon, Dagan and Krymolowski（1998））, maximum entropy （Skut and Brants（1998））, 及 SVM (Kudo and Matsumoto, 2000, 2001)等方法。上述幾種的方法都是監督式學習。HMM (hidden Markov model)使用統計的方法在 finite state machine 的 transition function 之上加上語料庫的統計結果。transformation-based learning 由現有的語料庫訓練出 transformational rules, 再利用這些規則對測試資料作 parse。HMM，transformation-based learning, memory-based learning 在自然語言處理中已被廣泛應用。 SVM 則是一種較新的 machine learning 技術,近幾年逐漸被應用到自然語言處理的各項研究議題,。

上述這些演算法針對英文 Wall Street Journal Corpus 訓練得到的結果顯示,精確率(precision)與召回率(recall)大都超過 90%, 其中以 SVM （Kudo and Matsumoto（2001））的效果最好，精確率(precision)與召回率(recall)都超過 94% （http://staff.science.uva.nl/~erikt/research/np-chunking.html）。

中文名詞組辨識的研究起步較晚，迄今只有零星的研究，還沒有針對同一個語料庫的大規模的測試與比較。例如中國大陸學者 Zhao and Huang (1998)提出以語料庫統計結合規律，利用 minimum description length principle (MDL)得到 quasi-dependency strength 加上規律來得到 base NP。這種採用非監督式機器學習（unsupervised learning）的方法，在封閉測試(close test)和開放測試(open test)中分別有 91.5% 和 88.7%的精確率。本研究使用 SVM 演算法來辨識中文基底名詞組（base NP），以便瞭解影響中文基底名詞組辨識的最重要特徵究竟有哪些，以及 SVM 在中文基底名詞組辨識的效果與限制。

3. 中文句法樹庫

由於 SVM 是監督式學習的演算法，我們必須擁有中文句法樹庫(treebank)的資料才能訓練出辨識名詞組的程式。目前我們擁有的兩個句法樹庫資料,一個是中央研究院中文句結構樹資料庫（Sinica Treebank 3.0）(http://www.aclclp.org.tw/use_stb_c.php)，另一個爲美國賓州大學中文句法樹庫（Penn Chinese Treebank 4.0）(http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T05)。兩者在語言，語料來源，語料庫大小，標記集，標記單位，標記訊息，及依據的語言學理論都不相同。下面是兩者的比較。

表（一）Sinica Treebank 3.0 與 Penn Chinese Treebank 4.0 的綜合比較

	語言	語料來源	語料大小	標記集	句法樹的性質	標記的訊息	所採用的語法理論
Sinica Chinese Treebank 3.0	繁體中文	台灣的報紙	290144 個詞，54902 棵結構樹	CKIP Tagset 包含未簡化標記，簡化標記及精簡標記三種	以標點符號分隔的詞組	語法及語意（包括詞組結構，中心語，修飾語，及語意角色等）	Information-based Case Grammar (ICG)
Penn Chinese Treebank 4.0	簡體中文	大部分爲中國大陸新華社新聞，少部分爲香港新聞及台灣光華雜誌社文章	404156 個詞，15162 棵結構樹	只有一個標記集。與英文 Penn Treebank 部分相同，另有部分標記是專爲中文設計	大部分爲完整的句子	語法結構與語法功能（主詞，受詞）	大致上採用 Chomsky 的 Government and Binding Theory 的詞組結構理論但額外加註語法功能的訊息。

表（二）Sinica Treebank 3.0 與 Penn Chinese Treebank 4.0 的部分例子

Sinica Treebank 的格式與例子	#PP(Head:P50:除了|DUMMY:GP(DUMMY:NP(Head:Nab:排鼓)|Head:Ng:以外))# #PP(Head:P21:在|DUMMY:GP(DUMMY:NP(Head:Nad:政治)|Head:Ng:上))# #NP(quantifier:NP(Head:Neqa:這些)|predication:VP・的(head:VP(Head:VL4:令 |goal:NP(Head:Nab:人)|theme:VP(Head:VK1:討厭))|Head:DE:的)|Head:Nac:戲)# #S(theme:VP(Head:VH11:這樣)|Head:VH11:好|particle:Ta:了)# #S(result:Cbca:而|theme:NP(quantifier:NP(Head:Nes:該)|Head:Ncb:校)|evaluation:Dbb:也 |Head:VK2:需要|goal:NP(quantifier:DM:一個|property:Nac:英文|Head:Nab:老師))#
Penn Chinese Treebank 的格式與例子	((IP (NP-PN-SBJ (NR 上海) (NR 浦东)) (VP (VP (LCP-TMP (NP (NT 近年)) (LC 来)) (VP (VCD (VV 颁布) (VV 实行)) (AS 了) (NP-OBJ (CP (WHNP-1 (-NONE- *OP*)) (CP (IP (NP-SBJ (-NONE- *T*-1)) (VP (VV 涉及) (NP-OBJ (NP-APP (NN 经济) (PU 、) (NN 贸易) (PU 、) (NN 建设) (PU 、) (NN 规划) (PU 、) (NN 科技) (PU 、) (NN 文教) (ETC 等)) (NP (NN 领域))))) (DEC 的))) (QP (CD 七十一) (CLP (M 件))) (NP (NN 法规性) (NN 文件))))) (PU ，) (VP (VV 确保) (AS 了) (NP-OBJ (DNP (NP (NP-PN (NR 浦东)) (NP (NN 开发))) (DEG 的)) (ADJP (JJ 有序)) (NP (NN 进行))))) (PU 。)))

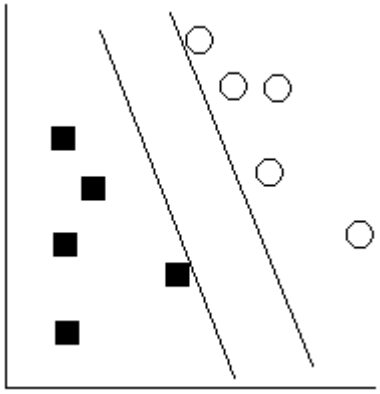
Sinica Treebank 與 Penn Chinese Treebank 最大的差別在於結構樹的語法單位不同。前者以標點符號作爲分隔不同結構樹的單位，因此一個結構樹很多時候只是一個詞組（如 PP, NP）而不是一個完整的句子。而後者除小部分結構樹是句子的片段(以 FRAG 標示)大部分的結構樹是完整的句子(sentence)(以 IP 標示)。另外 Sinica Treebank 語法結構採取中心語主導原則 (Head-Driven Principle)，註明中心語(Head)和其他成分（如附加語）的語法和語意訊息，表達出句子中詞和詞之間的語法結構和語意角色關係（http://godel.iis.sinica.edu.tw/CKIP/treebank.htm），而 Penn Chinese Treebank 並沒有中心語與語意角色的訊息，而是在詞組上加註如主詞 SBJ 受詞 OBJ 等語法功能的方式來取代。

我們選擇 Sinica Treebank 作爲訓練語料的主要原因在於做開放測試時我們需要一個與訓練語料採用同樣標記集的分詞與詞性標記程式。如果採用 Sinica Treebank，開放測時我們可以使用中研院的線上分詞與詞性標注程式 http://ckipsvr.iis.sinica.edu.tw/ 做爲我們 SVM 演算法的輸入資料。該線上程式的準確率相當高，特別是對於辭典未收錄詞的分詞與詞性標注的準確率比其它類似程式要高，採用該程式可以有效降低因爲一開始分詞與詞性標注錯誤而導致後面 SVM 演算法判斷錯誤的機率。另外由於 Sinica Treebank 有未簡化標記，簡化標記及精簡標記三種標記集，相較於 Penn Treebank 只有一種標記集，Sinica Treebank 的三種不同的標記集可以作爲不同的特徵。除此之外只有 Sinica Treebank 有標示語意角色的訊息，未來如果我們要以 SVM 來訓練標記語意角色，Sinica Treebank 顯然是較佳的選擇。

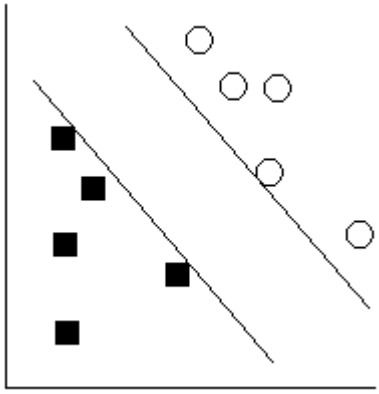
4. SVM 簡介

SVM 是較新的 machine learning 技術 (Boser, Guyon, and Vapnik (1992), Corts and Vapnik (1995)) 它使用一些策略來最大化具有不同特徵的資料中間的界限，並針對未知資料的特徵來判斷它屬於哪個類別。SVM 已在文件分類（Joachims （1998） Taira and Haruno （1999））以及名詞組標示 (Kudo and Matsumoto （2000, 20001）) 取得超越其它作法的準確性，而近幾年應用在自然語言處理的各個議題的研究更是方興未艾，如未知詞辨識（unknown word guessing）(Nakagawa, Kudo, and Matsumoto (2001)）詞性標注(part of speech tagging)Nakagawa, Kudo, and Matsumoto (2002)， Giménez J esús and Márquez Lluís (2004))句法依存關係辨識(dependency analysis)(Kudo and Matsumoto (2000))詞義辨別與標注(word sense disambiguation and sense tagging) (Cabezas, Resnik, and Stevens (2001))語意剖析(semantic parsing) (Pradhan et al. (2004) Sun and Jurafsky (2004))等都取得不錯的成果。

SVM 是一個分類用的 machine。請參照圖（一，二），



圖一



圖二

SVM 找出兩種資料 (黑色方形與白色圓形) 中間的界限, 圖一，圖二顯示出可能的兩種分割方式, 顯然的, 後者的切割方式是較佳的 (兩種資料的界線爲兩平行線之中線), 而 SVM 以滿足下面條件

$\min\Phi(\omega)= (1/2)|\omega|^2$

找出最佳平面 (即在線性可分的情況下, 可視爲解二次規畫的問題), 而此可由拉格朗日乘子法（Lagrange multiplier）求解.

由於很多的問題常常並不是線性可分的 (如我們的詞組切割), 這個時候 SVM 在比現有資料更高的向量空間 H 使用線性分類函數 $\Phi: R^d \rightarrow H$ 將 x 對應到高維空間, 便可在此以不破壞資料特徵亦不增加複雜度的方式對其進行分類.

在轉換的過程中, 我們會使用一 kernel function: $K(x_i,x_j) = \Phi(x_i)^T\Phi(x_j)$ 來實現非線性變換後的線性分類, 而使用不同的 kernel function 對不同的資料會有不同的效果.

以下爲一個簡單的 SVM 運作方式

給定一個訓練的資料集合:

(x_i, y_i) { i= 1, 2, ..., l; x_i 屬於 R^n; y_i 屬於 { 1, -1 }}

其中 l 爲訓練之資料數, x_i 爲一個 n 維向量, y_i 則是其類別 (分爲正類別 1 與負類別 -1) SVM 找到正類別與負類別中之最大的界限, 即解決下面的最佳化問題的解答

$\min_{w,b,e}(1/2)w^Tw + C\Sigma_{i=1}^{l}e_i$ 使得

$y_i(w^T\Phi(x_i) + b) >= 1- e_i, e_i >= 0$

x_i 經由 Φ 函數被對應到一個更高維的向量空間 H 之後 SVM 於此找到不同類別之間最大的界限; $K(x_i,x_j)$ 爲 Kernel function.

5. 詞性標記集與中文句法樹庫

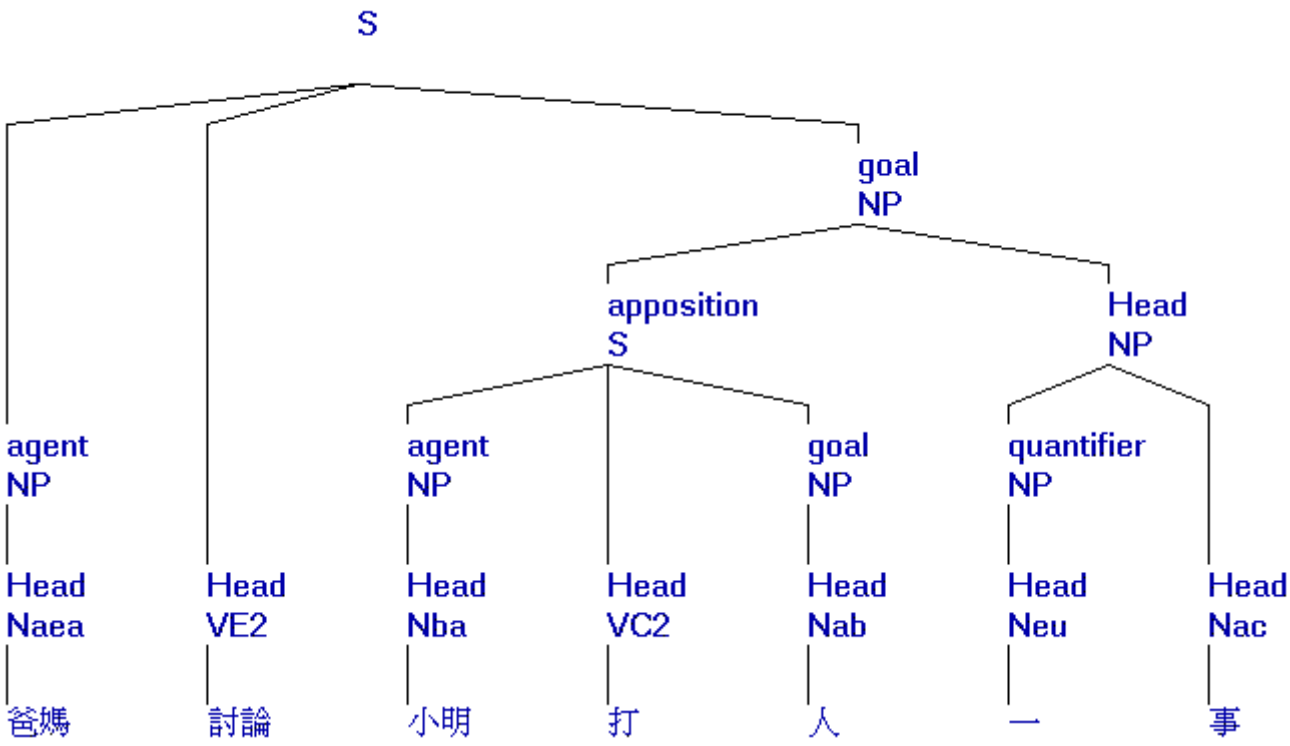
我們的實驗使用中研院語法樹庫 Sinica Treebank。中研院的詞性標記集及每個標記代表的語言學涵義如表(三)。（參考 http://www.sinica.edu.tw/SinicaCorpus/modern_c_wordtype.html）中研院詞知識庫小組所出版的「中文詞類分析」技術報告所提出的中文詞類的分類比表（三）的簡化詞類更細，但爲了顧及實用性中研院的漢語平衡語料庫所用的詞類標記爲已經經過合併的簡化詞類。我們可以看出即使是簡化詞類，連接詞，名詞，動詞，副詞每一項都有不少的次分類。以動詞爲例除了先分成動作與狀態兩大類之外，另外又根據動詞所帶的論元（argument）數目與種類各自分爲若干小類。中研院另外又將簡化詞類做進一步的合併形成所謂的精簡詞類。在簡化詞類裡面的動詞原先有 16 類但在精簡標記裡面只剩及物與不及物動詞 2 類。

表(三) 中研院的詞性標記集

簡化詞類	代表意義	精簡詞類	簡化詞類	代表意義	精簡詞類	簡化詞類	代表意義	精簡詞類
A	非謂形容詞	A	Nb	專有名稱	N	VB	動作類及物動詞	Vi
Caa	對等連接詞	C	Nc	地方詞	N	VC	動作及物動詞	Vt
Cab	連接詞，如：等等	POST	Ncd	位置詞	N	VCL	動作接地方賓語動詞	Vt
Cba	連接詞，如：的話	POST	Nd	時間詞	N	VD	雙賓動詞	Vt
Cbb	關聯連接詞	C	Nep	指代定詞	DET	VE	動作句賓動詞	Vt
D	副詞	ADV	Neqa	數量定詞	DET	VF	動作謂賓動詞	Vt
DE	的, 之, 得, 地	T	Neqb	後置數量定詞	POST	VG	分類動詞	Vt
Da	數量副詞	ADV	Nes	數量副詞	DET	VH	狀態不及物動詞	Vi
Dfa	動詞前程度副詞	ADV	Neu	數詞定詞	DET	VHC	狀態使動動詞	Vt
Dfb	動詞後程	ADV	Nf	量詞	M	VI	狀態類	Vi

	度副詞						及物動詞	
Di	時態標記	ASP	Ng	後置詞	POST	VJ	狀態及物動詞	Vt
Dk	句副詞	ADV	Nh	代名詞	N	VK	狀態句賓動詞	Vt
FW	外文標記	FW	SHI	外文標記	Vt	VL	狀態謂賓動詞	Vt
I	感嘆詞	T	T	語助詞	T	V_2	有	Vt
NAV	名謂詞	NAV	VA	動作不及物動詞	Vi			
Na	的, 之, 得, 地	N	VAC	動作使動動詞	Vi			

而 NP, VP 等詞組的判斷標準亦採用中研院句法樹庫的資料做爲我們測試的標準, （圖三）」是一個範例樹圖: (取自 http://godel.iis.sinica.edu.tw/CKIP/treebank/apposition.htm)



（圖三）中研院句法樹庫範例

如（圖三）所示，中研院的中文句法樹庫的 terminal node 是詞，詞上方有詞性標記和中心語（head）這類的語法訊息，構成詞組的結點(node)有詞組標記和語意角色等語意訊息。我們的焦點是 NP,也就是由 ”爸媽”, ”小明”, ”人”, ”一”, ”一事”組成的詞組。”小明打人一事”這類名詞組因爲包含其它的名詞組，不屬於基底名詞（base NP），

所以不在我們的討論之列。

6. 以 SVM 辨識中文名詞組的實作與實驗結果

訓練語料由於採取中研院的句法樹庫所以句子已經分詞並標注詞性。我們以 (Kudo and Matsumoto (2000, 2001) 的經驗做爲名詞組的辨識基礎。第一次實驗以 (I,O,B)三個標記分類:

這個方法以三個 class （I,O,B）表示一個詞在詞組中的位置:

I: 詞在詞組之中
O: 詞在詞組之外
B: 緊接著一個詞組之詞組的開頭

此種方法被 Tjong Kim Sang 稱爲 IOB1 表示法, 另外還有 IOB2, IOE1, IOE2,
在此不多加詳述.

Start/End
最初被用在日本語的作業上 (Uchimoto et al.（2000)）, 也就是 S，E，加上 I,O,B,共五個 class:

B: 多詞詞組的開頭
E: 多詞詞組的結尾
I: 詞在多詞詞組中
S: 單詞詞組
O: 詞在詞組之外

以下爲兩者之範例標記:

	Inside/Outside	Start/End
這	I	S
是	O	O
詞組	I	B
標記	I	I
範例	I	E
說明	B	S

一開始, 我們簡單的將測試資料排列成 7 維的向量, Wordi 是 i 位置的詞，POSi 是 i 位置詞的標記，加上前後各兩個詞的標記:

Wordi POS(i-2) POS(i-1) POSi POS(i+1) POS(i+2)

這裡根據詞，詞的標記, 和前面後面各兩個詞的標記來做分類。上面的範例向量表示如下:

I 1:這 2:0 3:0 4:N 5:S 6:N
O 1:是 2:0 3:N 4:S 5:N 6:V
I 1:詞組 2:N 3:S 4:N 5:V 6:N
I 1:標記 2:S 3:N 4:V 5:N 6:V
I 1:範例 2:N 3:V 4:N 5:V 6:0
B 1:說明 2:V 3:N 4:V 5:0 6:0

中研院中文句結構樹資料庫有 54,902 棵中文結構樹，290144 個詞。我們用樹庫的 80％做爲訓練的資料,20％做爲測試資料。由於不知道訓練語料對於 SVM 而言是否足夠，我們第一次實驗採用最少的特徵，採用的向量爲大類詞性 (即 N, V, P, ...)，也就是只看中研院詞類簡化標記的第一個字母所形成的大類。這比精簡標記的類別更少。我們使用現有的 SVM Tool: LIBSVM (Chang and Lin （2004)）作爲工具。SVM kernel function 爲 $K(x_i,x_j) = (\gamma\, x_i^T x_j + r)^d, \gamma > 0$ 以多項式趨近.實驗結果列於表（四）的第一列.

由表（四）可見辨識的成果並不好。 從語言學的角度來分析，中文名詞組的辨識比英文困難原因在於中文的動詞可以修飾名詞，例如投資大眾，建設公司,流浪教師等。這些詞沒有任何構詞上的特徵或證據可以視爲名物化（nominalization），因此詞性標記程式很難將這些詞判斷成名詞。由於中文的動詞可以修飾名詞使得自動辨識中文名詞組變得相當困難。不過我們仔細觀察後可以發現並不是所有的中文動詞都可修飾名詞，例如 VD(雙賓動詞), VK(狀態句賓動詞), VG(分類動詞)等這些類的動詞很少有修飾名詞的例子。由於我們採用的向量爲大類詞性 (即 N, V, P, ...), 動詞次分類這個重要特徵沒有考慮進去，因此實驗的結果非常不理想。如下面的例子:

可能(D) 代表(VK) 台灣(Nc) 人民(Na) 對(P) 朝野(Na) 政黨(Na) 傳達(VD) 訊息(Na)

程式抽取出來的 NP chunks 爲: “台灣人民”, “朝野政黨傳達訊息”; 顯然的 “傳達” 並不應該出現在 NP chunk 之中, 而就我們給予 SVM 的資料來看, 這邊並沒有明顯的訊息可以得知其不適用 (我們給予 SVM 的資料爲 “傳達(V)”), 而如 VH 等靜態動詞之類的動詞, 卻又常常出現在 NP 之中, 同樣標示爲 V。 由於我們採取簡化詞類標記的第一個字母的大類來表示，在缺乏動詞次分類訊息特徵的情形下使得實驗結果非常不理想.

因此, 我們保留將簡化標記動詞次分類的特徵，其它詞性則仍然使用大類, 結果如表（四）第二列所顯示，改良的方法在精確率上提升了 23%以上, 召回率也提升了 6%以上，雖然還不是非常好, 但顯示了詞性標記的選擇（有無動詞次分類的訊息）是影響 SVM 效果的重要的特徵.

表（四）動詞次分類訊息對 SVM 的影響

	Precision	Recall
(1)取簡化標記詞性第一個字母做大部分類	54.99%	53.17%
(2)動詞採用簡化標記細部分類其餘詞性取第一個字母大部分類	78.18%	59.33%

無論是精確率或召回率，我們實驗的結果與 Kudo and Matsumoto (2000,2001)發表的結果 (94%) 差了一大段距離; 可以改進的地方如下:

IOB tag, 我們的實驗只採取了 I/O 兩種 tag, 這在當兩個 chunk 緊連的時候會是一個致命的問題 (無法確認 chunk 的終結點)。修改 tag, 使用 IOB 與 Start/End 將可提升辨識率.

由目前的經驗得知, 好的詞性分類有助於準確度的提升。所謂好的詞性分類是指透過細部的詞性分類將能名詞組內部與外部兩種不同的特徵顯示出來，而將無助於此項辨識工作的詞性細部分類精簡成大類。如此透過 SVM 演算法可以提升名詞組的辨識精確率。

kernel function 與其微調的參數是影響 SVM 準確度的一大原因, 預期將會使用 linear, polynomial, radial basis function, sigmoid... 等等函數來做逼近, 並嘗試採用 cross validation 來尋找最佳參數.

目前面對的問題還有一點爲: 訓練的時間太久. 一個約 8,000 詞的訓練資料約需要花費 4 分鐘, SVM 之 time complexity 約爲 $O(n^2)$, 也就是說若有一 300,000 詞之訓練資料, 將需要花費約三天以上的時間訓練, 如此一來, 對於要使用 cross validation 將會是一大挑戰, 因此會嘗試使用 scaling 的方式來減少所需要訓練的時間.

YAMCHA (http://chasen.org/~taku/software/YamCha/)是 Taku Kudo 專門爲 NP Chunking 所設計的 SVM 工具，因此比一般性 SVM 工具（SVM Tool: LIBSVM (Chih-Chung Chang and Chih-Jen Lin, 2004)）方便實做。

YAMCHA 與 libsvm 的最大不同點在於:

a) Dynamic programming

b) Kernel Function

由於 libsvm 本身的限制, 我們很難能即時的將 chunking 的結果應用在下面一個未知 chunking 的判斷. 舉例而言, 之前的句子:

	Inside/Outside
這	I
是	O
詞組	I
標記	I

範例　　I

說明　　(B)

當 SVM 要判斷"說明"這個詞的 tag 時，它會去參考"標記"與"範例"的詞與詞性;原來的設計並未考慮到它們的 IOB tag，而由於中文（其實任何語言應該都一樣）有前後相依性，因此把 IOB tag 計算在內，會是一個適當而重要的特徵.

YAMCHA （Kudo and Matsumoto (2000,2001)）使用 IOB tag 代替 IO tag 方面，由於 B tag 表示了一個緊鄰之前 NP-chunk 的開頭，解決了兩個相鄰 NP-chunk 的分類問題.

另外 Kudo and Matsumoto (2000,2001)使用 voting 來提升辨識效果。voting 在很多應用中經常被使用。我們有許多不同的標記集，和不同方向的 parsing 方式（backward 即將所有的詞顛倒排列後做訓練與測試），藉著由不同標記集和不同的 parsing 方向訓練出來的 SVM 模型，可以採用其 Accuracy 之分數來統計未知詞組的得分. 這種方法可以避開某些詞性標記或者是 parsing 方向的盲點，以提升準確度。

另外從我們第一次的實驗結果得知動詞次分類訊息是一個影響 SVM 效果的重要的特徵。忽略動詞次分類的訊息會使辨識效果差很多。我們希望能從實驗數據中比較使用簡化詞類和精簡詞類是否會有很大的差別。

Kudo and Matsumoto (2000)以資訊檢索常用的 F measure 作爲評估系統的標準。F = (2* precision * recall) / (precision + recall)。由於 precision 高時則 recall 低，而 recall 高時則 precision 低，F measure 同時考慮 precision 與 recall，成爲評估時的綜合指標。

表（五）是我們利用 YAMCHA 實作 Base-NP chunking 所得到的結果。

表（五）不同的標記集和 parsing 方向的辨識率

	Precision	Recall	F measure
簡化詞類 (Forward)	86.48% (10360/11980)	88.41% (10360/11716)	87.43%
簡化詞類 (Backward)	86.29% (9983/11569)	85.21% (9983/11716)	85.74%
精簡詞類 (Forward)	87.34% (8789/10063)	75.02% (8789/11716)	80.71%
精簡詞類 (Backward)	84.88% (8651/10192)	73.84% (8651/11716)	78.98%
Vote using Accuracy Rate	88.71% (10048/11327)	85.76% (10048/11716)	87.21%

從表（五）可以觀察到 F measure 最高的是簡化詞類 forward parsing，使用 voting 並沒有提升 F measure，這是不是與訓練語料量不夠大有關，或其它因素造成，還是意味著中文只要 forward parsing 就能得到最好的效果不需要 backward parsing 和 voting，這些都有待進一步研究。值得注意的是在召回率（recall）方面簡化標記比精簡標記高 12 個百分點以上，原因是簡化標記具有 16 個動詞次分類而精簡標記動詞只有及物和不及物兩個次分類。由於精簡標記沒有足夠詳細的次分類的特徵，導致不少基底名詞組被誤判成動詞組。如果拿表（五）最好的結果與第一次的實驗結果表（四）比較，精確率提高了 10 個百分點，召回率則提高了 26 個百分點，這顯示

dynamic programming 和使用 IOB 與 Start/End 發揮了功用。雖然與英文的 95% F measure 仍有一大段差距，但是辨識效能已經大幅度的提升。

中研院的句法樹庫經過人工檢查，所以很少有錯誤。但開放測試時由於輸入的句子必須經過分詞和詞性標注（此部分透過中研院詞庫小組的線上分詞與詞性標注系統 (http://ckipsvr.iis.sinica.edu.tw/)），而分詞與詞性標注這兩個過程都有可能出錯，因此可以預期在開放測試時辨識的正確率會比封閉測試差，我們初步小規模的開放測試證實了這個預測。精確率與召回率分別爲 81.25%與 76.47%，F measure 則爲 78.79%。

7. 觀察到的問題

我們將訓練出來的模型實際使用在名詞組辨識時發現了幾個原來並未考慮到的問題。例如當我們實際測試如下的資料:

這(Nep) 是(SHI) 一(Neu) 個(Nf) 公平(VH) 的(DE) 審判(Na)

名詞組辨識的結果爲:
這(Nep)
一(Neu)

我們預期的結果:
這(Nep)
審判(Na)

我們發現了這個奇怪的結果之後，做了很多的測試，發現 Neu 類型詞皆會呈現單獨的 chunk，而 DE 之後則完全不會有 chunk. 這個問題在我們回頭檢查中研院句法樹庫時有了答案.

base-NP 就定義來看，爲 non-recursive NP chunk. 而在中研院句法樹庫中, DE 開頭的句子本身會成一個 (NP・的) 的 structure，而其餘接在 DE 詞 (如"的", "之") 之後的詞組皆不會成爲單獨的 NP chunk，也因此在上面我們所期望的"審判", 並沒有被抽取出來. 換句話說，在我們給予的訓練資料裡，就已經沒有將其標記爲 NP chunk 的例子了.

由於中研院句法樹庫特殊的標記方式，目前的情況是無法完全得到我們"看似"base-NP 的詞組分類結果，雖然 SVM 正確的分解出來它所判斷的詞組，但並不完全是我們想要的，這個問題目前似乎還沒有較好的解決方法.

檢視開放測試的資料，我們發現造成 SVM 判斷錯誤的主要有兩種情形。一種是 Nd 類的名詞不修飾名詞而當副詞，例如「以後(Nd) 經濟部(Nc) 」及「未來(Nd) 台灣(Nc) 水(Na) 」都被誤判成名詞組，另一種錯誤則是動詞修飾名詞例子，例如「重要(VH) 工作(Na)」這個名詞組並沒有被辨識出來。理論上中研院句法樹庫中 VH 類動詞修飾名詞的例子非常多，SVM 應該可以辨識出這樣的結構，實際上卻沒有辨識出來，造成此類錯誤的原因

還需要進一步研究。

8. 結論與未來的研究

我們的實驗中顯示動詞次分類的訊息對於提昇基底名詞組辨識的精確率與召回率而言是一個重要的特徵。我們的實驗間接證明中研院簡化標記中詳細動詞次分類訊息在中文自然語言處理上的優點。該分類系統先將動詞分成動作及狀態兩大類，再依據動詞的論元結構(argument structure)詳細分類。10 幾年前中研院詞知識庫小組院設計這套詞類標記系統的語言學家和計算語言學家或許只是單純從句法學與語意學的觀點提出這樣的分類系統。從事中文自然處理的研究人員對如此龐大詳細的詞類標記系統的必要性或許會質疑。但從我們的實驗中發現，中研院詳細的詞類次分類至少在動詞的次分類方面的確爲解決名詞組辨識的問題預先鋪路。機器學習演算法固然可以從訓練資料中自動學習，但是沒有專家的知識來分辨重要的特徵，仍然無法得到良好的效果。自然語言處理研究如果要有更進一步的發展，單靠機器學習演算法是不夠的，結合語言規律與知識是必然要走的路。

我們的實驗也顯示中研院句法樹庫某些結構表示法不利於我們辨識基底名詞組,加上中文的動詞可以修飾名詞造成同一個 SVM 演算法的辨識率比英文日文低許多。

近期我們會嘗試結合更多的語言知識及語言特徵，例如 Zhao and Huang (1998)以語料庫統計結合規律的方式來提升辨識率。我們相信類似的研究不僅有助於解決名詞組辨識的問題，對中文詞類標記集與句法樹庫的設計與修正也能提供回饋。

致謝

本研究得到國科會專題研究計畫 93-2411-H-002-013 「詞彙語意關係之自動標注─以中英平行語料庫爲基礎(3/3) 」94-2411-H-002-043 「中英平行句法樹庫的建立與英漢結構對應演算法的研究」及 94 年度國科會大專學生參與專題研究計畫「利用 SVM 標示中文名詞組的研究」經費補助，特此致謝。

參考文獻

Argamon, Shlomo, Dagan, Ido, and Krymolowski,Yuval (1998). A Memory-Based Approach to Learning Shallow Natural Language Patterns. In Proceedings of the 17th international conference on Computational linguistics, Vol. 1, pp. 67 - 73 , Montreal, Quebec, Canada.”

Brill, Eric and Ngai, Grace (1999), Man vs. Machine: A Case Study in Base Noun Phrase Learning. In Proceedings of ACL'99, pp. 65-72, University of Maryland, MD, USA.

Boser, E. Bernhard, Guyon, Isabelle, and Vapnik, Vladimir. (1992). A Training Algorithm for Optimal Margin Classifiers. COLT: pp. 144-152

Cabezas, Clara, Resnik, Philip, and Stevens,Jessica. (2001). Supervised Sense Tagging using Support Vector Machines. Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2), Toulouse, France, 5-6 July 2001.

Cardie, Claire and Pierce, David (1998). Error-Driven Pruning of Treebank Grammars for Base Noun Phrase Identification. In Proceedings of COLING-ACL'98, pp. 218-224, Montreal, Canada.

Chang, Chih-Chung and Lin, Chih-Jen. (2004) LIBSVM -- A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Chen, Kuang-hua and Chen, Hsin-Hsi (1994). Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation, In Proceedings of ACL-94, Las Cruses, NM, USA.

Church, K. (1988) A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. *Second Conference on Applied Natural Language Processing*, Austin , Texas , pp. 136-143.

Corte, Corinna, and Vapnik, Vladimir (1995). Support-Vector Networks. Machine Learning 20(3), pp. 273-297.

Giménez J esús and Márquez Lluís (2004). SVMTool: A general POS tagger generator based on Support Vector Machines Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal. 2004 .

Joachims, Thorsten. (1998) *Text Categorization with Support Vector Machines: Learning with Many Relevant Features.* Proceedings of the European Conference on Machine Learning (ECML), Springer, 1998.

Hsu, Chih-Wei, Chang, Chih-Chung, and Lin, Chih-Jen. (2004). A Practical Guide to Support Vector Classification.

Kudo, Taku, and Matsumoto, Yuji. (2000). Use of Support Vector Learning for Chunk Identification. In Proceedings of CoNLL-2000, pp. 142-144.

Kudo, Taku, and Matsumoto, Yuji (2000). Japanese Dependency Analysis Based on Support Vector Machines, EMNLP/VLC 2000

Kudo, Taku, and Matsumoto, Yuji. (2001). Chunking with Support Vector Machine. In Proceedings of NAACL 2001, pp. 192-199.

Marcus, Mitchell P., Santorini, Beatrice and Marcinkiewicz, Mary Ann. (1993) Building a large annotated corpus of English: the Penn Treebank, Computational Linguistics, 19:2. vol. 19, no. 2, pp. 313–330.

Pradhan, Sameer, Ward, Wayne, Hacioglu, Kadri, Martin, James H.and Jurafsky, Daniel. (2004). Shallow Semantic Parsing Using Support Vector Machines. In Proceedings of NAACL-HLT 2004,pp. 233-240..

Nakagawa, Tetsuji, Kudo, Taku, and Matsumoto, Yuji. (2001). Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines. NLPRS, pp. 325-331

Nakagawa, Tetsuji, Kudo, Taku, and Matsumoto, Yuji. (2002). Revision Learning and its Application to Part-of-Speech Tagging. In Proceedings of ACL 2002, pp. 497-504.

NP Chunking. http://staff.science.uva.nl/~erikt/research/np-chunking.html

Ramshaw, Lance A., and Marcus, Mitchell P.. (1995). Text Chunking Using Transformation-based Learning. In Proceedings of the Third ACL Workshop on Very Large Corpora, pp. 82-94, Cambridge MA, USA.

Skut, Wojciech and Brants, Thorsten. (1998) A Maximum-Entropy Partial Parser for Unrestricted Text. In Proceedings of the Sixth Workshop on Very Large Corpora, pp. 143-151, Montreal, Canada.

Sun, Honglin and Jurafsky, Daniel. 2004. Shallow Semantic Parsing of Chinese. In Proceedings of NAACL-HLT 2004, pp.192-199.

Taira, Hirotoshi, Haruno, Masahiko（1999）: Feature Selection in SVM Text Categorization. AAAI/IAAI 1999, pp. 480-486.

Tjong Kim Sang, Erik F. and Veenstra, Jorn (1999). Representing Text Chunks. In Proceedings of EACL'99, 173-179, Bergen, Norway.

Tjong Kim Sang, Erik F. (2002) Memory-Based Shallow Parsing. Journal of Machine Learning Research, Vol. 2, pp. 559-594.

Uchimoto, Kiyotaka, Ma, Qing, Murata, Masaki, Ozaku, Hiromi, Isahara, Hitoshi. (2000) Named entity extraction based on a maximum entropy model and transformation rules. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Hong Kong, pp. 326 – 335.

Veenstra, Jorn. (1998). Fast NP chunking using memory-based learning techniques, In F. Verdenius and W. van den Broek eds.,Proceedings of BENELEARN-98, pp. 71-79, Wageningen, The Netherlands.

Voutilainen, A. (1993) NPtool, a Detector of English Noun Phrase. In Proceedings of the First Annual Workshop on Very Large Corpora, pp. 48-57.

YamCha: Yet Another Multipurpose CHunk Annotator http://chasen.org/~taku/software/YamCha/

Zhao,Jun and Huang, Changning. (1998). A Quasi-Dependency Model for Structural Analysis of Chinese BaseNPs. In Proceedings of COLING-ACL 98, pp. 1-7 , Montreal, Canada.

中文詞類分析 (1988). 中央研究院詞知識庫小組技術報告,台北。

中研院詞知識庫小組中文斷詞系統(包含未知詞擷取與標記) http://ckipsvr.iis.sinica.edu.tw/

中文句結構樹資料庫」(Sinica Treebank Version 3.0). 中華民國計算語言學會
http://www.aclclp.org.tw/use_stb_c.php

史忠植 (2003). 知識發現, 清華大學出版社,北京.