中文名詞組的辨識：監督式與半監督式學習法的實驗
Chinese NP Chunking: Experiments with Supervised,and Semi-supervised Learning

林晏僖 Yen Hsi Lin
國立台灣大學資訊網路與多媒體研究所
r95944002@ntu.edu.tw

高照明 Zhao Ming Gao
國立台灣大學外國語文學系
zmgao@ntu.edu.tw

高成炎 Cheng Yan Kao
國立台灣大學資訊工程學系
cykao@csie.ntu.edu.tw

摘要

本文先利用 Taku Kudo 所發展的 SVM 工具 Yamcha 訓練中文名詞組辨識的初始模型，並嘗試以不同於多數文獻的 IOB 表示法及前二後二位置的語法標記資訊，找到適用於中文的參數。接著利用半監督式學習法中自我學習的概念，利用網路上未標記的資料，強化 supervised-learning 的模型。實驗結果證明，supervised learning 這個步驟裡，我們選用的參數比前人的更合適；而我們所提出的半監督式學習法，可以提昇辨識結果，特別是在動詞修飾名詞的情形，半監督式學習法可以大幅提高辨識的正確率。

Abstract

This paper utilizes Yamcha, a SVM tool designed by Taku Kudo, to train an NP-chunking model for Chinese. In addition to IOB and two words surrounding the focused word, we experimented on new features and exploited unlabeled data from web pages to enhance the previous model. Our experiments with supervised learning indicate that our chosen feature sets outperform those reported in previous studies. In addition, the proposed method of semi-supervised learning is proved to be effective in distinguishing a noun phrase from a verb phrase both consisting of V N combination, thus enhancing the overall accuracy.

關鍵詞：名詞組辨識、YamCha、監督式學習、半監督式學習

Keywords：NP-chunking、YamCha、supervised learning、semi-supervised learning

一、緒論

名詞組的辨識一直以來都是自然語言研究及其相關領域，如網路探勘（web mining）、文件分類（text categorization）等非常關鍵的一個步驟。在自動問答系統（question answering）中，關鍵詞多半以名詞搜尋爲主。在自然語言問答系統裡，名詞組的辨識也是不可或缺。我們每天都使用的搜尋引擎，大家輸入的關鍵詞以及搜尋引擎統計出來的熱門關鍵詞亦以名詞組居多；大至搜尋引擎背後大量的資料庫、小至普通文本建檔而成的資料庫，製作索引分類時，名詞組的使用也多於動詞、副詞；也有越來越多網頁都在針對網頁中的名詞組做自動偵測以及外部連結 ... 除了這些例子之外，語意角色標記（semantic role labeling）、專有名詞辨識（name entity identification）、文章處理中的回指（coreference），名詞組的辨識都是一個重要的步驟，因此有好的 NP chunker，可以改善許多 NLP 的研究成果以及相關應用。

在這篇論文中，我們先用 Taku Kudo 所提出利用 SVM 的演算法當作一開始的模型，除了許多參考文獻中常用的 IOB 標示法以及位置，我們還嘗試了以不同標示法以及加入不同位置的句子部份資訊當作特徵，證明對於中文的處理，不論是封閉或開放實驗中，IOE 表示法和加入前後兩個位置的詞及中研院簡化標記，是我們利用中研院句法樹庫 Sinica Treebank 所能得到最好的結果的參數。接著，我們利用一個沒有句法結構訊息的大型語料庫 word sketch engine 中的句子，加上半監督式學習法中，自我學習的概念，利用網路上大量未標記的網頁，來彌補 Sinica TreeBank 裡不足的訊息，改善利用監督式學習法實做出的 chunker。

實驗部份，除了封閉測試外，由於中研院樹庫圖中資料有限，我們額外收集了不同類型的句子當作開放測試的語料，以分別比較兩種作法在名詞組辨識的效果及限制。實驗結果顯示，我們選用的參數較前人選用的參數做出的模型在第一階段開放測試中高出了 16 個百分比，在第二個開放測試中也有 70%的 f-rate；加入 unlabeled data 這個步驟的半監督式學習法，也的確提昇監督式學習法的效果，使開放測試的 f-rate 提高至 78.79%，不但保存分類器的優點，也明顯提昇中文在難解的名物化歧義的名詞辨識結果。

接下來的章節中，第二章爲文獻回顧包含 Chunking, SVM, 半監督式學習法的基本介紹；第三章及第四章分別爲實驗方法說明和數據結果討論。最後爲結論與未來展望。

二、 文獻回顧

（一）、規則法

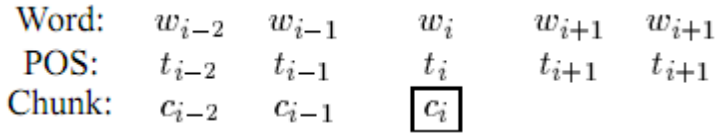
Abney(1995)利用了有限狀態機做出的規則式剖析器。他的實驗利用語意的訊息例如字形變化來當作特徵。他對英文及德文做了測試都成功並且快速的取出主要類型，包括動詞、名詞、介係詞的詞組。不過他還是強調選取的文法的重要性。Kinyon(2000) 提出一個適用於不同語言的 rule-based chunker。即使缺乏大量的訓練語料，只要有一些能夠辨別結構邊界的規則，就能使用 Kinyon 所提出的方法。Igor (2005)比較利用 NLTK 工具實做完成的 rule-based chunker 以及利用 TnT(Trigram and tag) 統計方式這兩種方法做出的 chunker 在名詞組和動詞組上的辨識效果，實驗證明近來鮮少被大家採用的規則方式實做出來的 chunker 不但沒有比利用統計的 chunker 遜色，甚至在召回率（recall）及 f-rate 的表現上要來的更好。在中文方面，雖然 Zhao 等(1999)對某些類型的詞組整理出結構的規則，但 Zhao 等(1999)還是捨棄規則式的作法，使用記憶基礎學習（memory-based learning）的方式。他們的實驗顯示若不加上詞彙本身的訊息，而只有利用詞性的訊息下，效果會比較差。

（二）、監督式學習及統計方法

在大規模語料庫建立之前，名詞組辨識常利用組成名詞組結構規律透過有限狀態機找出符合的模式 pattern，或從標記好詞性的語料庫以統計方式得到，或結合語言規律及語料庫統計；隨著賓州大學樹庫圖（University of Penn TreeBank）開放給大家使用之後，詞組辨識也朝向以機器學習的方法來解決：Skut and Brants(1998)、Koeling (2000) and Osborne (2000)使用最大熵演算法；Park and Zhang 採用規則以及記憶學習（memory-based learning, MBL）綜合的方式；Kudo and Matsumoto(2000,2001)利用 8 個 Support Vector Machine（SVM）系統投票（voting）的方式得出 chunking 模型，其他利用監督式學習（supervised learning）的方法還有 Hidden Markov Model(HMM)（Li (2004)）、transform-based learning(Ramshaw and Marcus (1995))這幾種，大都是利用語料的結構及前後語境的特徵得到的。這些演算法也早已被用在其他跟自然語言處理有關的議題上。

1 、Kudo 的支持向量機演算法

Kudo 等(2000) 第一個將 SVM 有效利用在詞組辨識作業上。它利用周圍的詞、這些詞的詞性以及預測的詞組類別當作訓練、預測過程中的特徵集，利用 SVM 對每個詞做標記的動作。要辨識第 i 個字的詞組類別 Ci，Kudo 採用了如圖一的特徵：



圖一、Kudo 提出的演算法中所使用的特徵[2]

Wi 是出現在第 *i*-th 個位置的詞, *Ti* 是 *Wi* 的詞性 而 *Ci* 是第 *i*-th 個字的詞組類別標記。另外，他們把特徵集中的 (*Ci+1*, *Ci+2*) 換成 (*Ci-1*, *Ci-2*) 以達到反向剖析的效果。由於在測試時，詞組類別標記這個特徵 (正向剖析：*Ci-1, Ci-2*；反向剖析：*Ci+1,Ci+2*) 並不是事先給定，而是利用當下模型決定的結果，因此被稱爲動態特徵；相對的 *Wi* 和 *Ti* 則爲靜態特徵。給定一個句子，例如：這/是/詞組/範例/標記，表一是對應的範例向量，其中 B 和 O 分別表示該詞是名詞組的開始或不在名詞組內。

表一、「這是詞組範例標記」在 Kudo 演算法中對應的範例向量

關注詞的類別	*Wi*	*Wi-2*	*Wi-1*	*Wi+1*	*Wi+2*	*Ti*	*Ti-1*	*Ti+1*
B	1:這	1:0	1:0	1:是	1:詞組	1:NES	1:0	1:SHI
O	1:是	1:0	1:這	1:詞組	1:範例	1:SHI	1:NES	1:NA
B	1:詞組	1:這	1:是	1:範例	1:標記	1:NA	1:SHI	1:NA

張席維、高照明與劉昭麟(2005)仿照 Taku 提出的這個演算法，以 Sinica TreeBank 當作語料，訓練中文的模型。雖然 Taku 在英文的實驗結果有 94%左右的正確率，張席維等只得到了 87.43%。

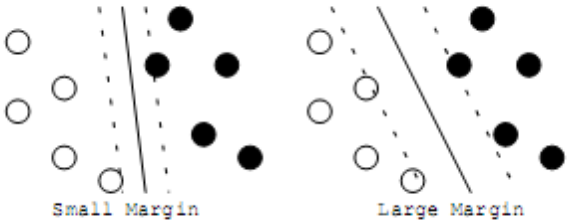
（三）、SVM 以及 YAMCHA

支持向量機（Support Vector Machine）是一個目前被廣泛運用在分類問題上的數學工具，是根據 Vapnik 的 max margin strategy 發展出來的分類器。相較於其他傳統分類器，如：決策樹學習(decision tree learning)、最大熵法(maximum entropy)等，SVM 有以下明顯的優點：

(1)即使在高維特徵向量空間下還是能產生好的效能。

(2)核心函數能將資料映射到更高維的空間而沒有增加計算複雜度。

支持向量機主要的想法是製造一個最佳的平面可以讓訓練範例向量分成兩個類別（positive and negative）並且把這個平面的邊界最大化。圖二中，黑色實線就是兩個可將資料分成兩類的平面，兩條虛線中間的距離就是邊界（margin），也就是 SVM 演算法試著最大化的目標。在虛線兩邊的點稱爲支持向量（support vectors），而且只有在訓練集中的支持向量會影響整個模型的結果。雖然 SVM 的分類準確度十分驚人，計算複雜度跟其他機器學習方法比起來也相對的高了許多。在需要龐大的訓練語料集的狀況下，利用 SVM 的訓練過程不但不夠有效率，甚至有可能因爲需要的訓練時間太久這種因素，實際情況下無法看到成果。



圖二、兩種可能將資料分開的超平面[2]

YamCha（Yet Another Multi-purpose Chunking Annatator）是 Taku Kudo 基於 Taku 等(2000)中的演算法，設計專門用在解決詞組辨識、詞性標記甚至文件分類等自然語言處理應用的工具。整個架構採用的分類方法是 SVM。跟單純的 SVM 分類器不同的地方是，Yamcha 要求的輸入檔案格式比較符合人直觀的想法，把需要做詞組分類的資料如同圖三，每個詞會利用到的特徵簡單的排列，直接交給 YamCha 去執行即可。如果不做任何參數的變動，這個工具就用 Taku 中一樣的預設值，把 (n-2, n-1, n, n+1, n+2) 位置上的字和特徵都當成關注詞(第 n 個詞)的特徵集去訓練。所以跟傳統分類器不同的地方，只在於 YamCha 幫使用者處理了資料格式的問題。另外值得注意的一點是，雖然 SVM 分類的效果非常的好，但其耗費的計算量以及時間也比其他分類器來的大，而 YamCha 在這點上做了改進，使得訓練時間以及分類時間都加速了至少三倍以上。

He	PRP	B-NP
reckons	VBZ	B-VP
the	DT	B-NP
current	JJ	I-NP
account	NN	I-NP
deficit	NN	I-NP

圖三、Yamcha 輸入資料的格式，與 CoNLL 2000 shared task 相同。其中 B, I, O 分別表示該詞是某種詞組的開始，內部，或不在詞組中。

（四）、半監督式學習法

如同表面上的字義一樣，半監督式學習法介於監督式學習法和非監督式學習法之間：利用大量未標記過的資料結合一些已經標記過的資料來做訓練的模型以解決資料量稀少及分散的問題。而在一個自然的考量下，如果有一樣數量標記過的資料（labeled data），我們是不是能利用大量容易取得的未標記過、未處理過的資料（unlabeled data）來建造一個更精確的分類器（classifier）？ 這個問題通常就會被歸類於半監督式學習（semi-supervised learning）。在真實生活中，標記資料不但耗費時間、人工、甚至金錢；相對的，未經過標記的資料量多而且隨手可得。因此在機器學習的領域上，如何利用未標記資料是一個重要的課題，例如：我們可以利用程式取得網路上大量的網頁存檔，卻需要人工處理才能做正確的分類；在語音辨識的研究中，要取得大量的錄音檔非常簡單，但是要去標記這些錄音檔則需要大量的時間和人力去打逐字稿，在這些情況下，如果 unlabeled data 確實能讓模型的效能提高，半監督式學習是非常的有幫助的。

在半監督式學習中，資料型態和不同演算法的模型的配對是十分重要的，不然也有可能造成反效果。EM with generative mixture model、self-training、co-training、TSVM 還有基於圖形的演算法都是 SSL 裡常用的方法並且各有優點。假如標記的類別可以把把資料分開得很明確，使用 EM with generative mixture model 比較好；如果特徵群就足夠隨著資料被分成兩半，那麼 co-training 的方法會比較適合，因爲這個演算法就是對分開的特徵做一些假設，並在不同的特徵集上使用不同的學習工具。若有相同的特徵的

點會被分在同一個類別，而目前的模型也無法被改進時， Mincut, Boltzmann Machine, Tree-based Bayes 等基於圖形的方法會比較適合。

最早引進利用未標記過的資料這個概念應用在分類上的也許就是自我學習這種演算法。一開始只利用少量標記過的資料作訓練，接著利用當下的決策函數（decision function） 從未標記的資料中找出符合的點，加進本來標記過的訓練集，重新訓練並找出新的決策函數。直到在未標記的資料中無法再找到可標示的資料或是整個情形超過一些閥值（threshold）。

Yarowsky（1995）是在提到自我學習時常被引用的有名的例子。Yarowsky 利用這個方法做分辨語意歧異度，跟只用標記過資料和監督式學習法的效果比起來改進很多。在一個篇章中，要如何決定 plant 這個字是植物的意思還是工廠的意思？Yarowsky 的假設是：

（1）一個詞在一章篇幅中只會有一個意思,，（2）一組搭配語中也只會有一個意思。他先從未標記過的資料中，取小量（約爲訓練集的 2%）的句子出來標記答案，例如：句子中的 plant 是植物則此句爲 class A，是工廠則爲 class B。根據（1）的假設，找出同類別句子中的搭配語，例如：一開始選定的 A 中的句子的 plant 旁邊都有 life 這個字，而 B 的 plant 旁邊都有 manufactoring 這個字，利用這個特徵到未標記過的資料中去找有同樣特徵的句子，收進訓練集裡。也同時利用一些決策函數在這些新標記好的句子中找新的搭配語。接著利用（2）的假設，如果同一篇文章中的多個句子都已經被歸到同一個類別，則同篇文章中剩下的句子都可以分到同一個類別，這個假設不但可以擴大訓練集，還可以修正在前面的步驟被歸類錯的句子。重複擴大訓練資料、利用新的資料訓練模型的步驟直到未標記資料的數量不再有太大的變化爲止。實驗結果證明，這個方法訓練出來的模型跟只利用標記過資料及監督式學習法的效果比起來，確實將效能提昇並且少了很多人工標記的動作。

三、作法說明

（一）、名詞組表示法

Inside/Outside： Ramshaw and Marcus(1995)提出並使用了下述三種 class（IOB）表示一個詞在詞組中的位置。

I：這個詞在某個詞組之中

O：這個詞不屬於任何詞組

B：這個詞是緊接著別的詞組的詞組開頭

這個表示法被 Tjong Kim Sang 稱爲 IOB1，另外他還提出了 IOB2/IOE1/IOE2：

IOB2 中 B 是任何詞組的開頭；

IOE1 中 E 是緊鄰著別的詞組的詞組結尾；

IOE2 中 E 是任何任何詞組的結尾；

表二是各種表示法的範例說明。另外還有 start/end 的表示法，由於在 Taku(2000)中實驗結果不佳，因此本實驗中並沒有用到，也不加以介紹。

表二、以各種表示法標記「這是詞組範例標記說明」

	IOB1	IOB2	IOE1	IOE2
這	I	B	I	E
是	O	O	O	O
詞組	I	B	I	I
範例	I	I	I	I

標記	I	I	E	E
說明	B	B	I	E

（二）監督式學習法：Supervised-learning

我們仿照[1]及[2]中的演算法（在 2.2.1 中也有描述），我們將前後兩個詞及詞性分類的訊息及前面兩個已經判別好的詞分類當作特徵集讓 SVM 分類器參考，利用 SVM 分類器判別每個詞是屬於 IOB 中的哪一類。例如「這是一個例句」中有哪些名詞組？要判別「一」這個詞時，將前後兩個詞及個別的詞性標記 這、NEP、是、SHI、個、NF、例句、NA、前面兩個已經判別出來的詞組分類 E（這）、O（是）以及本身一、NEU 這 12 個當作分類的特徵值。最後這五個詞分別的類別可能爲

這	是	一	個	例句
E	O	I	I	E

則可以利用 IOE 分出詞組間的邊界，判斷出此句的名詞組爲「這」及「一個例句」。 因此監督式學習法可說是希望能找出有最佳辨識效果的特徵集。

（三）半監督式學習法

從[1]及監督式學習法的實驗可以看出，只從樹圖資料庫的資訊加上查詢詞典的詞義對我們判斷動詞的名物化現象並沒有幫助，於是我們希望能利用外部的資源幫助我們獲得一些新的訊息，靈感來自於前面提到的 Self-training。
從訓練語料的句子裡，我們對動詞後面接一個名詞的組合做觀察之後發現一些規則，下表是例句及推測：

表三、語料中部份句子及其特性

例句	推測
導遊 發 給 每人 一 ***本*** *導覽 手冊*。	量詞的後面常常會接名詞組。
這個 報導 給予 我 們 無限 ***的*** *想像 空間*。	"的"的後面常常會接名詞組。
大家 看 ***了*** *宣導 短片 之*後 有 什麼 感想 呢？	時態詞的後面常會接名詞組。

表四、從語料庫推測出來的規則

可能出現在動詞後面的詞類	NEQA、NEP、NEU 等量詞
	NG : 時間後置詞
	NC : 位置詞
	NH : 代名詞

先從訓練語料中任選當修飾的動詞接名詞(例如：採購人員、運動精神)，以及動詞加上受詞(例如：祭拜祖先、來自家人) 的這兩種情形，各 100 組 Bi-gram，從 word sketch engine 中搜尋 50 句包含這些 bigram 的句子，如下面幾句是包含採購人員的句子：

...團體 *之 採購人員*。
...需要仰賴的不只有 *總務 採購人員 自身*。
...又蘊含 *了 採購人員 對* 他人和社會...
...大多數 *的 採購人員 會* 蕭規曹隨...

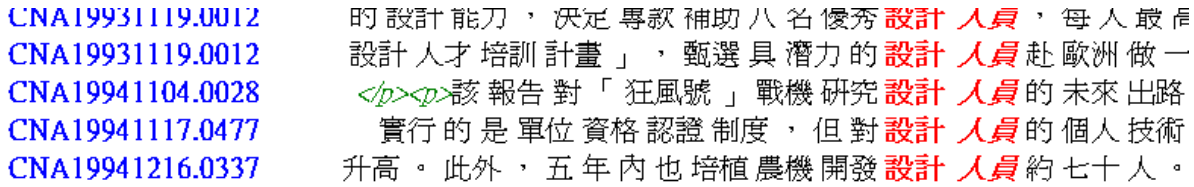
蒐集到某個數量的句子之後，我們對關注的詞組前後緊鄰的詞類做統計。即使收集來的句子前或後會連接許多不同的詞類，我們還是可以分別對這兩種情形歸納出一些可

能的共通點，如下表：

表五、在不同功用的動詞前的詞類比較

類型	可能出現在前面的詞類
動詞	D：副詞
修飾用的動詞	DE：的，之等、DI：時態標記、NEU, NF 等量詞

接著我們進行初步小規模的測試，從封閉測試語料中，選擇動詞後面立即接一個名詞(V1 N1)的組合，一樣的從 WSE 中蒐集句子，接著統計這些句子中緊鄰詞的詞類，再利用上表先前推測出的特徵判斷（V1 N1）是哪一種情形。例如我們蒐集包含「設計人員」的句子（圖四）：



圖四、WSE 中截取出來的例句

統計過後的結果，前面出現"DE"、"DI"及量詞的機會（次數）比出現副詞"D"的機會（次數）高，因此將設計人員歸類爲修飾用的動詞加名詞組合而成的複合詞（class A）。而包含「採購汽車」的句子中，前面緊鄰副詞"D"的機會比緊鄰"DE"、"Di"及量詞的高，則將其歸類於動詞接受詞這類（class B）。還有一種情形是，如「服務社會」這個詞組，在 WSE 裡找不到一個句子精確的包含這個 bi-gram 或是只收入了四五句相關的句子，也就是資料不足（class C）。

針對在 WSE 裡面，沒有出現過的詞語組合，例如：服務社會，或出現次數不夠多，拿資料來分析似乎不夠客觀的詞組，例如：紅燒牛肉，我們利用 google 搜尋引擎尋找內文裡有符合一樣詞組的網頁，得到如同 WSE 中包含關注詞組的短句。跟 WSE 一樣，我們收集前 50 個 google 傳回的 Snippet（各個網頁的摘要），只是在 WSE 裡的句子有做過斷詞以及有部份的標記，而從搜尋引擎得來的是沒被分析過的資料（raw-data）。從 Google 得到資料並經過中研院分詞程式處理後，我們利用跟在 WSE 中作法雷同，統計關注詞組前面的詞的類型。在我們剛剛提到的功能詞的長度大部分都是一個詞或兩個詞，因此我們查詢字典中，關注詞組前面的一個詞以及兩個詞的詞性，查看是否是符合我們假設中模式的詞並做統計。

由於這兩個利用外部語料庫資料的步驟，只能將某些 POS 本來屬於動詞的詞彙功能做分類，因此這個實驗的結果將被當成一個特徵加入 SVM 工具中一起訓練，SVM 工具還是整個標記名詞組過程的核心。

因此此時模型採用的特徵成爲 IOE, Wi-2 .. Wi+2, Pi-2..Pi+2, Vi, Ci-2..Ci-1,

$$\text{其中 Vi} = \begin{cases} \text{C, Wi 的 pos 不屬於動詞類別} \\ \text{A, Wi 是被名物化的動詞} \\ \text{B, Wi 是動詞} \end{cases}$$

四、實驗結果討論

（一）、實驗語料介紹

本文的實驗中，監督式以及半監督式學習法所需要的訓練語料，使用的是中研院中文句結構樹資料庫 Sinica Treebank3.1。這個資料庫的文章來源分別有直接從平衡語料庫

中取出的文章、國小課本、光華雜誌以及中研院語言所的語音平衡檔案，再經過電腦剖析及人工校對做成樹圖庫。全部包含了六個檔案，分別爲不同的背景、情境，共有65434 個中文樹圖、392237 個詞（平均一句包含了六個詞）。我們把資料庫中每個檔案的 70%取出整合來當作訓練語料、30%當作封閉測試的測試語料。
檔案中的句子都以下句的形式表示，除了可從中得知結構訊息之外還有中文的語意角色。句子中的詞類標記，是 CKIP 詞類標記，與中研院的字典所使用的詞性標記是同一種（另外有還有簡化標記、精簡標記兩種）。

> # S(agent:NP(Head:Nca:觀光局)|evaluation:Dbb:還|quantity:Daa:另|Head:VE12:安排|aspect:Di:了|theme:NP(property:NP(quantifier:DM:幾處|Head:Ncb:市郊)|property:Nv4:遊覽|Head:Nac:活動))# 。(PERIODCATEGORY)

在訓練語料中，約有 45000 個句子，依照 Church 的定義，所取出的 NP chunk 共有 65009 個，但是每個 chunk 平均只包含了 1.57 個詞。可見得在中研院樹庫圖裡標記得 np-chunk 還是以單詞居多。從上一個句子中，依照樹庫圖中的結構，「觀光局」「幾處市郊」「活動」都成爲一個單獨的 chunk，但少了我們認定的「遊覽活動」甚至有可能是「幾處市郊遊覽活動」。如果只利用 non-recursive chunk 的結構從 Treebank 自動抽取答案出來，會得到很不理想的訓練語料，因此在標記詞組答案上，我們對幾種結構做了修改，標記出較符合我們思維的答案，包含
(1) 名詞+的+名詞 ；(2) 形容詞+的+名詞；(3)量詞+的+名詞

（二）、相關資源介紹

在之後的實驗中，我們使用到一些廣爲人知的工具及資源：
（1）中研院斷詞程式：輸出包含分詞結果，與每一個詞的詞性標記，並可以處理未知詞。
（2）詞彙特性速描系統（Word Sketch Engine） [12]：這是一個包含了中文、英文、法文、德文等多種語言的大型語料庫，並且已經對這些語料做了同義詞、用語索引、搭配語分類的整理。這個語料庫中的句子，是標示好的資料，除了已經做好斷詞，也可以查看詞性標記。如下列兩句：

> 稅捐處 工商 稅科 、 財產稅科 、 稽徵 科 及 稅務 管理 科 等 依照 權責 ，將 分別 全面 查緝 逃漏稅 。
>
> 在*/P21* 賦稅*/Naeb* 方面*/Nac*，*/COMMACATEGORY* 查緝*/VC2* 逃漏稅*/Na* 及*/Caa* 進行*/VC2* 會計師*/Nab* 評鑑

在詞性標記這部份與中研院斷詞程式不同的地方在於，中研院斷詞程式似乎參考了句子的語境標示每個詞的 POS 而 WSE 沒有，所以在不屬於功能詞（function word）的部份，也就是一個詞可能會有多種詞性的情況下，WSE 的詞類標記的精確度會比較差，並且比較不適合拿來當參考。
（3）Google Soap API：讓使用者合法做關鍵字搜尋，每天最多可做 1000 筆的搜尋，並提供回傳網頁的相關資訊。。

（三）潛在問題

1. 相較於英文有 CoNLL2000 Shared Task 的 Chunking 規格、資料集和答案，中文這方面的訊息並不一致，多數還是從標記好的樹庫中抽取初以標記好的 chunk（Xia 等(2000)）或是自己再從樹圖標示的結構定義及標記 chunk 當作訓練語料的答案（Li (2003), Zhao and Huang (1999)）。若同 CoNLL2000 shared task 按照 Church(1996)在英文中的定義，將 chunk 視爲沒有包含其它種 chunk 的詞組，也就是不重疊（non-overlap）、不遞迴（non-recursive）

的詞組合而成的，即 NP-chunk 可簡單的想成不包含別種詞組的名詞組。但回頭看語料庫中標記好的 NP-chunk 有絕大部分是屬於單詞，反而不符合我們一般的思維（4.1 節），因此若利用語料庫標記好的 NP-chunk，除了很難將全部的答案改成如同我們所希望看到的規則，還會有下面提到的長詞組的問題。

2. 由於中文是一種沒有屈折語素（inflectional morpheme）的語言，例如英文中被動式的動詞會有一個變化型，轉爲名詞的用法則字尾變成 ing、加上 tion 等等，但是在中文裡則是加個「被」字以表達被動式，其他的情形在前後不一定有加入的關鍵詞，必須由對中文有一定了解程度的人自己對語境做推測來判斷每個詞的詞性及功用。例如：從下面這個句子

The experiment involved the *combining* of the two chemicals。

可以很清楚的看出 combining 是名詞的用法，但是在以下這兩個句子，無法直接看出進口和喜愛的詞性。

政府編定汽車管理制度使進口汽車得以合法化。 他深得學生的喜愛。

這也是張席維等、Ding 等(2005)中提及的名物化現象。由於在 Sinica TreeBank 裡有這種現象的詞組只有不到 3000 組，因此張席維等也根據實驗結果強調利用監督式學習法辨識中文的名詞組時，能否找出被名物化的動詞是一個提昇正確率的關鍵。

3. 從 Sinica TreeBank 中取出標記好的 chunk 的平均長度不到兩個詞，Cheng 等(2005)提到中文實際上有非常多由數個名詞組合而成得名詞組，例如：行政院/國家/科學/委員會、電腦/人體/模型...等等，這些在日常生活中都不是令人陌生的詞語。因此使用 Sinica TreeBank 當做一種 gold standard 或是訓練語料時，很難解決長詞的問題。

（四）開放測試集

由於上面說明了很多在訓練及封閉測試語料中無法觀察到的情形，因此我們利用開放測試的結果作爲不同方式設計的模型間比較的準則。在監督式及半監督式學習方法也各有一組的開放測試資料作爲該次實驗內的參數比較。監督式學習法中的開放測試語料，大部分是包含"形容詞接名詞"、"量詞接名詞"、"量詞接形容詞接名詞"、名詞中有所有格的句子，例如：這是最新的車款、事情發生在去年的夏天、班上有一名天才學生..等等；半監督式學習法的開放測試語料強調動詞的判別，因此測試語料中的名詞組包含一些已轉化爲別的作用的動詞，例如："他在拍賣網站上買東西"，"警察透過銷贓管道抓到小偷"等等。

實驗結果我們採用與 CoNLL 2000 shared task 一樣的評量方法，直接利用他們提供的評量工具，分別算出詞標記正確率（tag accuracy）以及詞組正確（Precision）、詞組召回率（Recall）以及 F-rate = 2PR/P+R，而 F-rate 還是爲主要考量。

（五）監督式學習法 Supervised-learning

在[1]中，作者利用 IOB 表示法來做名詞組的標記，並指出簡化標記及精簡標記對名詞辨識的影響度；當其餘的詞類使用大分類，而保留簡化標記的動詞次分類時可使學習效果提昇。樹庫中的 CKIP 詞類標記比簡化標記的分類更細，也代表 CKIP 的詞類標記透露出更多的語言訊息，這樣是否能讓 SVM 的學習效果更好呢？另外，由於中文的名詞詞組的中心語（Head）傾向出現在最後面（head-final），那麼 IOE 的表示法是否比 IOB 的恰當？因此我們先針對詞性標記以及名詞組表示法做選擇。(CKIP 詞類標記和簡化標記的對照和代表意義可參考 http://godel.iis.sinica.edu.tw/CKIP/paper/poslist.pdf)

表六是封閉測試的數據，從此表中可以發現這四個模型之間並沒有明顯數字上的差距，

因此我們轉向開放測試的結果。下列是一些開放測試的句子：

我們買了一張很貴的票。
我聘了一個很優秀的職員。
阿忠的那一間房子。

表六、比較 IOB,IOE,CKIP,simplified（簡化標記）四種特徵在封閉測試時的結果

feature combination	tag accuracy	precision	recall	f-rate
W(n-2 .. n+2) , P(n-2 .. n+2) in Simplified tagset , T(n-2 .. n-1), IOB	91.21%	84.85%	86.98%	85.90%
W(n-2 .. n+2) , P(n-2 .. n+2) in CKIP tagset, T(n-2 .. n-1), IOB	90.89%	84.44%	86.60%	85.50%
W(n-2 .. n+2) , P(n-2 .. n+2) in Simplified tagset, T(n-2 .. n-1), IOE	92.06%	84.65%	86.28%	85.46%
W(n-2 .. n+2) , P(n-2 .. n+2) in CKIP tagset, T(n-2 .. n-1), IOE	91.93%	84.34%	86.09%	85.20%
W(n-2 .. n+2) , P(n-2 .. n+2) V in CKIP, others in Simplified, T(n-2 .. n-1), IOE	92.11%	84.67%	86.23%	85.44%

由於[1]已經說明監督式學習訓練出來的模型對名物化詞類沒有好的辨識效果，因此在這部份的開放測試，我們先選擇一些句子裡有名詞組中基本形式，像量詞接名詞、形容詞接名詞、量詞接形容詞加名詞，加上一些包含長複合詞組以及有所有格的句子。表七是這個實驗的結果。

表七、IOB IOE 初步開放測試比較結果

feature combination	accruracy	precision	recall	f-rate
W(n-2 .. n+2) , P(n-2 .. n+2) in Simplified tagset , T(n-2 .. n-1), IOE	92.95%	86.67%	89.66%	88.14%
W(n-2 .. n+2) , P(n-2 .. n+2) in Simplified tagset , T(n-2 .. n-1), IOB	88.93%	68.75%	75.86%	72.13%

從表六中，我們可以看出（1）CKIP 詞類標記總共有多達約 230 個分類，簡化標記約有 45 個分類，因此不論是只有動詞採用 CKIP 的次分類或是整體的詞類都利用次分類來標記，可能由於 CKIP 詞類分項太細造成分類器中資料稀疏的問題，使得採用 CKIP 詞類標記的表現並沒有以用簡化標記的表現好，以及（2）雖然在封閉測試中兩種表示法精確度不相上下，從開放測試的表七來看，以 IOE 表示法的結果會比 IOB 表示法來的精確，所以我們固定這兩種參數來進行之後的實驗（包括半監督式學習法）。利用簡化標記的另一個好處是：當我們用中研院的斷詞程式對開放測試的資料做前處理時，得到的標記與訓練出來的模型使用的一致。由於我們從這個開放測試的結果發現：

1. 初始模型對長詞的偵測不太敏感：語料中有"一名騎機車的年輕人"及"一名高級官員"。這兩句中的「一名」都是某個 NP 的一部分，但是前者的詞組標記爲一 O/ 名 O 而後者爲一 I / 名 I。因此量詞後面被 tagging 過程考慮進來的詞性變得十分重要。由於我們初始模型只將前後各兩個詞加入特徵集，在開放測試裡「有一間很漂亮的教師休息室」的「一間」就有辨識錯誤的可能，因爲在量詞後面的兩個詞都還不見名詞的蹤影。
2. 即使有一些句子有著非常類似的形式（例如：很漂亮的衣服，很貴的票），但是模型輸出的結果卻不相同，而我們發現這或許是因爲 "很漂亮" 在訓練集中出現過爲名詞一部分的用法，而"貴"在訓練集裡只有當動詞用。

因此我們考慮：

1. 往前後看不同長度的詞及語意特徵作爲特徵集；
2. 某個位置的詞及語意特徵不必同時存在。

表八是特徵集的符號以及其代表的意義；

表八、實驗中採用的特徵代表符號以及相對的意義

代表符號	代表意義	代表符號	代表意義
Wn	第 n 個詞	IOB	利用 IOB 表示法標記名詞組
Ln	第 n 個詞的 POS	IOE	利用 IOE 表示法標記名詞組
Tn	第 n 個詞的 tag 標記	Hn	第 n 個詞在 hownet 中的義元
Simplified	簡化標記	CKIP	CKIP 標記
F	forward parsing	B	backward parsing

表九是不同模型所使用的特徵組合，由於這些組合在封閉測試的結果都十分相近，所以不將數據一一列出；

表九、模型及其利用的特徵對照表

model 1	F, Wi-2.. Wi+2, Pi-2..Pi+2 Ti-2..Ti-1
model 2	F, Wi-2.. Wi+2, Pi-2..Pi+2,Hi-2..Hi+2, Ti-2..Ti-1
model 3	F, Wi-4.. Wi+2, Pi-4..Pi+2 Ti-2..Ti-1
model 4	F, Wi-1.. Wi+1, Pi-1..Pi+1 Ti-2..Ti-1
model 5	F, Wi, Pi-2..Pi+2 Ti-2..Ti-1
model 6	B, Wi-2.. Wi+2, Pi-2..Pi+2 Ti-2..Ti-1
model 7	F, Pi-2..Pi+2 Ti-2..Ti-1

表十是各個模型開放測試的結果，從開放測試的輸出來看每個模型都有明顯不足的地方，「監察人員」這種型態的詞更沒有一個模型判斷正確。model 2 除了語料庫的內部訊息之外，加入了每個詞在 HowNet 中的義元（可以視爲語意特徵或類別）當做一個特徵，雖然比 model 1 的 F-rate 好上約 0.8 個百分比，訓練模型的時間卻也增加爲約 1.8 倍；model 3 參考前四個詞及後兩個詞，雖然對上面提到包含量詞的長名詞組合有些幫助，但是對其他問題沒有太大的影響；backward parsing(model6) 在 closed test 的部份，雖然有最高的 F-rate, 但是在開放測試的部份卻沒有特別好的表現。

表十、各模型開放測試比較結果

model	tag accuracy(%)	precision(%)	recall(%)	F-rate(%)
model 1	92.95	86.67	89.66	88.14
model 2	93.43	87.67	90.63	88.97
model 3	91.95	83.37	87.93	85.59
model 4	91.14	86.19	88.26	87.21
model 5	89.95	81.97	81.14	81.55
model 6	92.30	84.78	85.34	85.06
model 7	90.55	80.54	80.44	80.49

這階段開放測試的句子中，有很多詞（或詞性標記(pos)序列）是重複的，加入不同的搭配語或修飾語，或是以不同的語序組合，是測試模型對不同形式的語句的準確

度及穩定度。
除了受限於訓練語料這個問題之外，在訓練語料方面除了之前提到的缺點以外，歧義、名物化、未知詞這些實際生活中的現象，在語料庫裡是沒有標記的；由於語料庫的組成大多是長篇文章切成的句子，某些用法或語句，會因爲出自於同一篇文章而重複出現很多次，實際的訓練語料並不如統計過後數字上得多。雖然在這個語料庫裡共有六萬多句的句子，但是因爲分句是以標點符號爲原則，所以有很多句子其實只包含單詞，或只由名詞組成，或比正常情況書寫來的短，無法從中看出結構的訊息，反而可能變成分類器的雜訊（noise）。例如：「爸爸說：山路不難走」這個句子在語料庫中被分爲「爸爸/說」和「山路/不/難/走」兩句，但實際生活中多數的寫法還是會以長句子爲主。中研院句法樹庫中的詞及詞組標記由於經過人工校對，所以相當精確、可信度高，但是在實際測試時發現，有些詞的詞性標記利用中研院斷詞程式執行出來的結果，由於分詞程式分詞或標記錯誤或其它原因，並不會出現。例如，在語料庫中有一類 NV 的詞如電腦（NA）/打字（NV4）/及（CAA）/排版（NV4），而中研院斷詞系統並沒有辦法即時判斷出名物化(NV)的現象，因此斷詞後的結果爲「電腦(Na)　打字(VA)　及(Caa)　排版(VA)」；在訓練集中，有「兩（NEU）/者（NA）/同等重要」這樣的句子，中研院的線上程式斷詞的結果是，「兩者（NH）/同等重要」，顯現出開放測試和封閉測試的語料有一定的差異。
然而這個實驗的數據顯示，利用監督式學習法處理中文 NP-chunking 時，只有訓練語料中的資訊可以被利用的時候，IOE 名詞組表示法、簡化詞類標記、關注詞本身及前後兩個位置的詞，還有他們的詞性（pos）是最好的特徵組合。也因此我們固定這幾個特徵，當成下一個實驗的基本特徵集。

（六）實驗：半監督式學習法 Semi-supervised learning

我們對這個方法分別做了封閉及開放測試。由於此處著重在改進辨識名物化的動詞，這部份的開放測試語料除了上個實驗的測試語料，我們還加入包含不曾出現在訓練語料中的名物化動詞詞組的句子，佔全部測試語料的 50%。表十一是這個方法的實驗結果，supervised II 是將詞的前一個詞性單獨拿出拿作爲一個特徵。在封閉測試中，半監督式學習法實驗結果的 F-rate 爲 85.46%，雖然只比監督式學習作法的高出了 0.2%。但在開放測試的部份，半監督式學習法明顯比監督式學習法的 F-rate 高出了 8.79%之多。當我們只透過此實驗的作法描述中利用未標記資料判別封閉測試裡 VN Pair 中的動詞類別時約有八成的正確率（包含錯誤的被修正以及本來正確的維持正確），但是透過分類器的預測結果之後，VN pairs 這部份卻只有約 5%的改善，要如何有效應用分類器來突顯出新找到的特徵的重要性著實爲一個議題；由於轉爲名詞的動詞在樹庫圖中的詞類標記爲 NV 所以封閉測試語料中並不需要考慮表面上爲名詞接動詞的情形。開放測試中有 21 個名物化動詞的詞組，監督式學習法中正確判斷出六組，而半監督式學習法判斷出九組，剩餘的 12 組中有 7 組在作法描述中利用未標記資料判斷動詞類別時的分類是正確的，但經過分類器分類之後變成誤判的情形。

表十一、監督式及半監督式學習法實驗結果

	Tag accuracy	Precision	Recall	F-rate
封閉測試				
supervised	92.06%	84.65%	86.28%	85.46%
supervised II	91.76%	81.71%	86.05%	83.82%
semi-supervised	92.19%	84.85%	86.64%	85.73%
開放測試				
supervised	89.03%	67.31%	72.92%	70%

supervised II	83.83%	63.06%	69.23%	66%
semi-supervised	91.61%	76.47%	81.25%	78.79%

半監督式學習法中，我們只用了蒐集來的語料裡特定詞組前後的詞性做判別，但這與在監督式學習法裡將前後的詞性獨立分類(supervised II)作爲特徵有何不同呢? 由於在訓練語料中包含這些變化動詞的詞組約爲 5%，也就是訓練語料太少；另外句子中的前後詞性非常可能只是恰好在當句中出現在隔壁，而不是真正具有辨別作用的成分，如我們可能需要大量的 DE 或 DI 類的詞來佐證一個 VN 組合爲名詞的複合詞組，但是在語料庫中這個詞組的前面是個普通名詞，如同我們蒐集來的句子也有非常多是無法在我們統計過程中能被拿來參考的。因此半監督式學習法的成果還是比監督式學習法好。另外對於 WSE 和 Google 這兩個語料庫的比較上則各有優缺點。WSE 中的資料不但有斷詞還有詞類標記非常方便拿來直接使用，而從 Google 得到的則是 raw data，必須再經過斷詞的處理；WSE 中的句子經過挑選，形式較單純，而網路上的網頁還包含了關鍵字可能在標題、連結或搜尋到的網頁的重複性太高，可能有的特定的句子，現在很流行、或被很多人引用過，那麼出現在搜尋頁上很多筆都在描述同一筆資料，也就是同一個句子，而且在關注詞前面出現的詞，是在我們期待之外的 pattern、或無法讓我們利用的 pattern，分別形成錯誤資料太多、能用的資料太少的情況。因此無法在我搜尋的範圍內，有足夠符合我預期的模式的資料；在 WSE 中無法找到新的詞彙及不適合被收錄在語料庫中的詞彙，從 Google 得到的資料則沒有這個問題。

五、結論及未來展望

我們利用不同的特徵並將原本的模型加以改善過後，利用監督式學習法在小型開放測試有 70%的 f-rate，但在利用未標記過的網頁當作特徵之後，將模型的 f-rate 提昇至 78.79%，比原本高出了 8.79%。雖然模型的 performance 還是會受到訓練語料的影響使得結果不穩定，偶爾會有與預期不符合的情形發生，整體來說，我們提出的利用半監督式學習方法善用了網路上隨手可得的資源並且的確增進名詞辨識的效果。
本篇論文的貢獻在於：

(1)雖然詞組辨識一直是許多自然語言處理議題中的重要步驟，但在中文方面並沒有看到太多相關的研究。由於中文的名詞組結構相較於其它語言的名詞組結構都要複雜的多，因此本文只專注於名詞組的辨識，並且證明之前用在類似作法以及在其他語言中被普遍採用的特徵，並不完全適合用在中文語言上。而我們也找到了更適合的特徵。

(2)我們提出一個簡單的半監督式學習法，改善了監督式學習法中資料稀疏 (data sparseness) 及只依賴訓練語料時無法解決的問題。並且跟原本的 chunker 相比之下提高不少準確度及實用性。對於一些自然語言處理中，需要利用較高比例的名詞組的應用，例如：句子剖析 parsing、語意角色 semantic role labeling、文件分類 text categorization 等等，都有實質上的幫助。

未來我們希望能夠找到更好的特徵，加上監督式機器學習的方式，以解決更長的複合詞組以及包含有兩個以上名物化動詞的複合詞組。另外，由於網路上大量未標記過的資料隨手可得，因此我們也希望能提出非監督式(unsupervised)的演算法，以突破受限於少量人工標記過訓練語料的限制。

參考文獻

[1] 張席維，高照明，劉昭麟 （ 2005 ）利用向量支撐機辨識中文基底名詞組的初步研究。第十七屆自然語言與語音處理研討會。 pp. 317-332
[2] Kudo, Taku, and Matsumoto, Yuji. (2000). Use of Support Vector Learning for Chunk

Identification. In Proceedings of CoNLL-2000, pp. 142-144.
[3] Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (pp. 189–196).
[4] Kudo, Taku, and Matsumoto, Yuji. (2001). Chunking with Support Vector Machine. In Proceedings of NAACL 2001, pp. 192-199.
http://chasen.org/~taku/software/YamCha/
[5] Chang, Chih-Chung and Lin, Chih-Jen. (2004) LIBSVM -- A Library for Support Vector Machines.[On line]. Available.
http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[6] Guang-Lu Sun, Chang-Ning Huang, Xiao-Long Wang, and Zhi-Ming Xu .Chinese Chunking Based on Maximum Entropy Markov Models. Computational Linguistics and Chinese Language Processing Vol. 11, No. 2, June 2006, pp. 115-136
[7] R. K. Ando and T. Zhang. A high-performance semi-supervised learning method for text chunking. In Proceedings of the Annual Meetings of the Association for Computational Linguistics (ACL), pages 1-9. 2005
[8] Semi-supervised learning book
http://www.kyb.tuebingen.mpg.de/ssl-book/
[9] Xiaojin Zhu, Semi-supervised literature survey, December 14, 2007
http://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf
[10] Yuchang CHENG and Masayuki ASAHARA and Yuji MATSUMOTO,“Machine Learning-based Dependency Analyzer for Chinese”, Journal of Chinese Language and Computing 15 (1): (13-24) ,2005
[11] CoNLL 2000 Shared Task http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.tx
[12] The Sketch Engine http://www.sketchengine.co.uk