中文動詞名物化判斷的統計式模型設計

馬偉雲　黃居仁

中央研究院語言學研究所
ma@iis.sinica.edu.tw　churenhuang@gmail.com

摘要

中文動詞名物化的現象在中文的語法研究上一直是一個重要的課題。而對中文自然語言處理系統來說，自動判別句子當中的動詞是否名物化也是在剖析過程當中不可或缺的技術之一。一個動詞在句子當中所扮演的角色到底是單純的謂語，或是派生名詞，影響剖析結果甚鉅。由於中文動詞名物化時缺乏構形上的變化(zero-derivation)，因此判斷動詞是否名物化就必須仰賴動詞本身的內部語素結構、語意以及上下文方可得知。過去由於語料庫大小限制，欠缺足夠的名物化樣本及其語境可供建立統計式模型，因此前人多利用少數觀察到的語法規則企圖建立規則式模型來判斷名物化。舉例來說，動詞前或後出現“的”時，這個動詞即有很高的可能性是派生名詞。然而，較為複雜的名物化現象仍難以這些簡單的規則就能判定。

本論文是第一個嘗試以統計方式自動判斷中文動詞名物化的研究。利用大規模的帶名物化標記的語料庫，根據不同假設，訓練出各類統計式模型，自動判斷一個動詞在其語境當中是否名物化。實驗結果顯示出，表現最佳的統計模型對於派生名詞的包含率為 71.8%，準確率為 76.6%，F-Score 為 74.1%。我們也針對不同的統計式模型的表現作分析，發現整合派生名詞的動詞來源詞(verbal counterpart)的語法詞類(syntax category)訊息的模型，往往比未包含此訊息的模型表現要來得好。經由實際語料的分析，我們觀察到不同的動詞來源詞的語法詞類不僅僅在扮演謂語角色時語境不同，在扮演派生名詞的角色時，所搭配的語境有時也有極大的差異。這樣的差異性在設計名物化判斷系統上是不可欠缺的關鍵因素之一。

1. 簡介

中文動詞名物化的現象非常普遍[葉美利等 1992]，當一個動詞在句子中轉換成名詞的語法功能時，就稱此動詞名物化或是轉換成派生名詞了。然而，一個動詞要轉換到什麼程度才可稱為派生名詞，目前並沒有統一的看法。根據[詞庫小組技術報告 95-02/98-04]對於動詞名物化的定義。動詞名物化四種最常見的情況是：一.動詞出現在主語位置，如(1)(2)。二.動詞出現在虛化動詞(light verb)的賓語位置，如(3)(4)。三.動詞出現在名詞片語結構的中心語，如(5)(6)。四.動詞出現在名詞前修飾名詞，如(7)(8)。

(1)　打架(VA) [+nom] 是(SHI) 不(D) 對(VH) 的(DE)
(2)　上學(VA) [+nom] 幫助(VC) 我們(Nh) 學習(VC) 知識(Na)
(3)　進行(VC) 調查(VE)[+nom]
(4)　維持(VJ) 清潔(VH)[+nom]
(5)　學生(Na) 的(DE) 不(D) 合作(VH) [+nom]
(6)　他(Nh) 對(P) 國家(Na) 的(DE) 認同(VJ) [+nom]
(7)　主辦(VC) [+nom] 單位(Na)
(8)　吵架(VA) [+nom] 方式(Na)

[葉美利等 1992]提供了派生名詞更細緻的分類，將派生名詞依帶論元與否分為兩大類，並按所帶的論元及其體現將帶論元的派生名詞分成十個小類，且提供了語法表達模式。

自動判斷名物化與否的技術是中文自然語言處理重要的一環,對於句子或動詞片語的剖析更是關鍵。但由於中文動詞名物化時缺乏構形上的變化(zero-derivation)，缺乏英語 -ion、-ment、-ing 等動詞名物化標示，因此無法從構形上判斷，必須仰賴動詞本身的內部語素結構，語意，以及上下文方可得知。過去由於語料庫大小限制，欠缺足夠的名物化樣本及其語境可供建立統計式模型，因此前人多利用少數觀察到的語法規則企圖建立規則式模型來判斷名物化。如[Lin et al., 1997]分析包含派生名詞的名詞片語語法結構，建立一系列的語法規則作為剖析器的參數，當剖析完成時自然就決定了每個動詞名物化與否。這樣的作法好處是將名物化判斷和剖析整合在同一過程當中，考慮的範圍廣而全面，並且最後能夠得到完整的剖析結構。只是[Lin et al., 1997]所定義的包含派生名詞的名詞片語語法結構，必須後接“的”才算是包含派生名詞的名詞片語。並無法全面處理上述[詞庫小組技術報告 95-02/98-04]對於動詞名物化的定義。

相較於[Lin et al., 1997]，本論文的作法將名物化判斷從剖析當中分離出來，名物化判斷之後的結果才送交剖析器作進一步處理，這樣做有兩個理由：1.名物化判斷基本上屬於詞類標記(tagging)的問題，目前主流的剖析技術都是以詞類標記後的結果作為其輸入。2.雖然現今主流的統計式剖析技術，如“機率式上下文無關”(PCFG) 剖析器，理論上只要有大規模的帶名物化標記的語料庫，它也可以訓練出統計式模型來剖析，一併處理名物化判斷的問題。不過，因為影響剖析好壞的原因很多，並不只有動詞名物化這個原因。因此暫時排除剖析這個變因，可以使我們專注在名物化判斷這個主題。除了得到一個高準確率的模型之外，我們也希望能分析出影響動詞名物化的催化因素，提供語言學上的解釋及驗證。特別是在語言學的分析當中，[葉美利等 1992]、[Huang et al., 1994]等都提出轉換後的派生名詞和轉換前的來源詞 (verb counterpart) 有密不可分的關係，在上下文中往往可以找到相對應的論元成分。本論文進一步觀察到針對某些特定來源詞的語法詞類，他們轉換成派生名詞之後，其論元的位置通常也具有共通性。因此，我們設計了區分來源詞語法詞類的模型跟未區分的模型,希望能探究來源詞語法詞類在名物化當中所扮演的角色。

名物化判斷基本上可說是一個詞類標記(tagging)的問題，當我們判斷一個動詞在上下文中已經轉換成派生名詞時，我們可以給予一個通用的名物化詞類標記(如 Nv)取代原本的動詞詞類標記，另一個策略是在其原本的動詞詞類後面附加名物化特徵標記(如(VC)[+nom])。由於詞類標記的技術已經相當成熟,因此在本論文中首先會測試在詞類標記技術當中最廣為使用的隱藏式馬可夫模型(HMM model)，之後為了更進一步掌握動詞本身特性和上下文的因素，我們模仿在詞義區分當中廣為使用的貝氏分類器(Naïve Bayes Classifier)，提出另一種型態的統計式模型，並藉由實驗證明其可行性。在本論文的實驗討論當中，也分析了各類模型所隱含的語言學現象。

2. 系統描述

設計系統之前，首先我們必須先決定什麼是它的輸入以及什麼是它的輸出，才能明確規範出系統所要真正解決的問題。不同的輸入或輸出也會在在評估系統表現上產生極大的差異。若是單一限定在動詞名物化判斷這個主題，並打算測試其表現好壞，直覺上，輸入應是一個已經斷好詞，並且可能已經有一些現成的標記(如詞類標記)的句子，斷詞和標記都是正確無誤的。這樣的輸入經

過名物化模型的判斷，輸出時，系統在某些動詞上取代或標記上名物化的詞類或特徵。

輸入：學生(Na) 的(DE) 不(D) 合作(VH)
輸出：學生(Na) 的(DE) 不(D) 合作(Nv) / 學生(Na) 的(DE) 不(D) 合作(VH)[+nom]

但是以這樣的輸入來評估系統表現顯得不切實際，因為在中文處理當中，我們所面臨真實的輸入是一個未正確斷好詞以及標記的句子。但是若是採取一個連斷詞都不具備的輸入，系統又勢必承接許多斷詞所造成的錯誤，在評估時就難以評估出名物化判斷模組的好壞。因此，我們採取的策略是將斷詞這個變因在系統設計當中剔除。也就是說，在實際的中文處理上，斷詞由另一套斷詞模組負責，和名物化判斷無關，在評估名物化判斷的表現時，我們的輸入就是一句已經正確斷好詞的句子。我們將詞類標記和名物化判斷整併到一個模組當中，原因是這兩者在判斷上實際上密不可分，互為影響，一併設計並且評估它們是比較符合實際的作法。另外，由於在派生名詞的輸出上，附加名物化特徵的呈現方式(如(VH)[+nom])會比只標記名物化詞類(如 Nv)，更多了來源詞的語法詞類訊息(如 VH)，因此我們採用這樣的輸出模式。

輸入：學生 的 不 合作
輸出：學生(Na) 的(DE) 不(D) 合作(VH)[+nom]

我們採用[詞庫小組技術報告 95-02/98-04]對於動詞名物化的定義來設計系統，原因是這套定義十分簡明，並且我們所用來訓練的語料庫的名物化標記亦採用這套定義，如此可以產生具有一致性標注原則的訓練以及測試語料，來產生與評估我們的模型。

3. 統計模型

詞庫小組在 1997 年完成了帶詞類標記的平衡語料庫[詞庫小組技術報告 95-02/98-04]，並且利用附加名物化特徵的方式呈現派生名詞。語料庫的規模為五百萬詞。這樣的規模使得訓練統計式的名物化判斷模型成為可能。

名物化的催化因素大體上可分成兩大類，分別是動詞本身名物化的可能性以及上下文對此動詞名物化的影響力。根據不同假設，這兩大類催化因素又可繼續細分成更細部的催化因素，我們也據此提出兩種隱藏式馬可夫模型以及三種貝氏分類器。

值得一提的是，本論文並沒有使用目前在標記上與分類領域最新的統計式技術，如“條件式隨機場域”(Conditional Random Field)或是“支援向量機器”(Support Vector Machine)等，其原因並非認為它不適用於名物化判斷，而是本論文作為第一個針對統計式名物化判斷的研究，除了想開發出準確的系統之外，我們也想藉由各類的模型探討背後所隱含的語言學現象，找出名物化的催化因素。因此，根據不同假設，我們設計出比較符合直覺的統計式模型，且均以相連機率(Bigram Probability)為運算材料，在實驗分析上，這些機率值可以讓我們深入觀察或驗證名物化的催化因素，並使錯誤分析更容易。以下為本論文所使用的符號:

w_i	the word at position i
v_i	the verb at position i
t_i	the tag at position i
$v_i(t_i)[+nom]$	the nominalized verb and its tag at position i
$v_i(t_i)$	the non-nominalized verb and its tag at position i
$(t_i)[+nom]$	the nominalized-verb's tag at position i
(t_i)	the non-nominalized-verb's tag at position i
$v[+nom]$	a nominalized verb
v	a non-nominalized verb
$c(v_i):\{w_{i-1},t_{i-1},w_{i+1},t_{i+1}\}$	the context of

3.1. 隱藏式馬可夫模型

名物化判斷基本上可說是一個詞類標記的問題，而詞類標記的技術當中最廣為使用的就是隱藏式馬可夫模型。我們提出兩種不同類型的隱藏式馬可夫模型，未區分動詞來源詞語法詞類的模型，稱之為 HMM1，區分動詞來源詞語法詞類的模型，稱之為 HMM2。我們採用傳統的雙連式隱藏式馬可夫模型，以雙連詞類機率作為運算參數。換句話說，這樣的模型定義了判斷標的之上下文為前一個以及後一個語法詞類。

3.1.1 HMM-1

對每一個動詞的上下文來說，當我們假設一個特定的上下文對每一個動詞都具有同樣的名物化催化力時(如任一個動詞，無論它是什麼詞類，只要前接副詞-"地"，幾乎就可以確定此動詞不會轉換成派生名詞，而單純扮演謂語的角色)，我們就可以將所有派生名詞都視為同一個詞類，如 Nv，因此就相當於我們在做詞類標記的問題，只是詞類集多了一個成員-"Nv"。利用典型的隱藏式馬可夫模型，計算

$$\prod_{i=1}^{n}[P(w_i \mid t_i)\times P(t_i \mid t_{i-1})]$$

找出最佳的詞類序列，就可以得到名物化判斷的結果。不過這樣的結果還並不是最後的輸出，因為對於所判斷的派生名詞，我們仍然不知道他們的來源詞語法詞類。一個最簡單的作法，就是取他們在統計上頻率最高的來源詞語法詞類作為輸出。之所以可以這樣做，主要是由於一個動詞的名物化現象絕大多數只發生在它的一個特定動詞詞類身上。以下舉例說明整個流程：

輸入：學生 的 不 合作
經過 HMM 後：學生(Na) 的(DE) 不(D) 合作(Nv)
取頻率最高的來源詞語法詞類為輸出：學生(Na) 的(DE) 不(D) 合作(VH)[+nom]

3.1.2 HMM-2

對每一個動詞的上下文來說，當我們假設一個特定的上下文對不同動詞的詞類具有不同的名物化催化力時，在模型設計上就必須區分出上下文和不同動詞詞類之間的關係。以詞類標記的角度來

看，就相當於詞類集多了許多成員，如“(VC)[+nom]”，“(VA)[+nom]”，“(VH)[+nom]”…等等。利用典型的隱藏式馬可夫模型找出最佳的詞類序列就可得到最後輸出。以下舉例說明整個流程：

輸入：學生 的 不 合作
經過 HMM 後：學生(Na) 的(DE) 不(D) 合作(VH)[+nom]

3.2. 貝氏分類器

名物化判斷除了可以視為是一個詞類標記的問題，事實上它也可視為是一個二元分類的問題。根據本論文第二章所做的系統描述，當輸入是一個動詞時(如：合作)，系統要作的就是根據這個動詞的上下文，判斷這個動詞屬於哪一類-是單純謂語（如：合作(VH)，表示它未名物化且詞類為 VH)，還是派生名詞(如：合作(VH)[+nom]，表示它名物化且來源詞語法詞類為 VH)。

上述的分類法會碰到一個實際上的困難。我們知道上下文是很重要的分類依據，而上下文除了相鄰詞之外，相鄰詞的詞類標記也是很重要的線索，在只有相鄰詞而沒有相鄰詞的詞類標記的情況之下，建立統計模型時必定面臨資料稀疏(data sparseness)的問題。因此，我們將原始的詞類標記和名物化判斷分開處理，也就是說，先將輸入的句子以傳統的隱藏式馬可夫模型決定其每個詞的語法詞類標記，之後再針對其中每個動詞，分析其本身的名物化可能性以及包含相鄰詞詞類標記的上下文，作名物化判斷。這個流程的前半段是傳統的詞類標記問題，後半段是一個二元分類問題，即只有兩個分類類別，包含[+nom]或未包含[+nom]。以下舉例說明整個流程:

輸入：學生 的 不 合作
經過 HMM 後：學生(Na) 的(DE) 不(D) 合作(VH)
經過分類器後：學生(Na) 的(DE) 不(D) 合作(VH) [+nom]

我們提出三種不同類型的分類器，區分動詞來源詞但未區分其語法詞類的模型，稱之為 Classifier-1，未區分動詞來源詞但區分其語法詞類的模型，稱之為 Classifier-2，區分動詞來源詞且區分其語法詞類的模型，稱之為 Classifier-3。之所以設計這三種不同的分類器，主要目的除了希望得到一個最佳表現的分類器之外，也想藉此了解動詞來源詞的語法詞類是否在名物化判斷上扮演重要的角色。另外，跟隱藏式馬可夫模型不同的是，這三個分類器不僅考慮了判斷標的之前後語法詞類，也考慮了判斷標的之前後詞。

Bayes decision rule:
Given a verb v_i, *its tag* t_i and its *context* $c(v_i)$
if $P(v_i(t_i)[+nom] \mid v_i, t_i, c(v_i)) > P(v_i(t_i) \mid v_i, t_i, c(v_i))$, choose $v_i(t_i)[+nom]$
else choose $v_i(t_i)$

$$
\begin{aligned}
&P\left(v_i(t_i)[+nom] \mid v_i, t_i, c(v_i)\right) \\
&= \frac{P\left(v_i, t_i, c(v_i) \mid v_i(t_i)[+nom]\right)}{P\left(v_i, t_i, c(v_i)\right)} \times P\left(v_i(t_i)[+nom]\right) \\
&\cong P\left(v_i, t_i, c(v_i) \mid v_i(t_i)[+nom]\right) \times P\left(v_i(t_i)[+nom]\right) \\
&= P\left(c(v_i) \mid v_i(t_i)[+nom]\right) \times P\left(v_i(t_i)[+nom]\right) \\
&\cong \log\left(P\left(c(v_i) \mid v_i(t_i)[+nom]\right)\right) + \log\left(P\left(v_i(t_i)[+nom]\right)\right) \\
&= \log\left(P\left(\{w_{i-1}, t_{i-1}, w_{i+1}, t_{i+1}\} \mid v_i(t_i)[+nom]\right)\right) + \log\left(P\left(v_i(t_i)[+nom]\right)\right) \\
&= \log\left(P\left(w_{i-1} \mid v_i(t_i)[+nom]\right)\right) + \log\left(P\left(t_{i-1} \mid v_i(t_i)[+nom]\right)\right) + \\
&\quad \log\left(P\left(w_{i+1} \mid v_i(t_i)[+nom]\right)\right) + \log\left(P\left(t_{i+1} \mid v_i(t_i)[+nom]\right)\right) + \\
&\quad \log\left(P\left(v_i(t_i)[+nom]\right)\right) \\
&= \log\left(\alpha P\left(w_{i-1} \mid v_i(t_i)[+nom]\right) + \beta P\left(w_{i-1} \mid (t_i)[+nom]\right) + \gamma P\left(w_{i-1} \mid v[+nom]\right)\right) + \\
&\quad \log\left(\alpha P\left(t_{i-1} \mid v_i(t_i)[+nom]\right) + \beta P\left(t_{i-1} \mid (t_i)[+nom]\right) + \gamma P\left(t_{i-1} \mid v[+nom]\right)\right) + \\
&\quad \log\left(\alpha P\left(w_{i+1} \mid v_i(t_i)[+nom]\right) + \beta P\left(w_{i+1} \mid (t_i)[+nom]\right) + \gamma P\left(w_{i+1} \mid v[+nom]\right)\right) + \\
&\quad \log\left(\alpha P\left(t_{i+1} \mid v_i(t_i)[+nom]\right) + \beta P\left(t_{i+1} \mid (t_i)[+nom]\right) + \gamma P\left(t_{i+1} \mid v[+nom]\right)\right) + \\
&\quad \log\left(P\left(v_i(t_i)[+nom]\right)\right)
\end{aligned}
$$

Classifier-1: $\alpha = 0.8, \beta = 0, \gamma = 0.2$

Classifier-2: $\alpha = 0, \beta = 0.8, \gamma = 0.2$

Classifier-3: $\alpha = 0.5, \beta = 0.3, \gamma = 0.2$

The calculation $P\left(v_i(t_i) \mid v_i, t_i, c(v_i)\right)$ is similar to $P\left(v_i(t_i)[+nom] \mid v_i, t_i, c(v_i)\right)$

4. 實驗

為了比較上述模型的表現，本章提出我們的實驗方法。

4.1. 實驗環境

我們以詞庫小組所開發的平衡語料庫作為訓練以及測試的材料，這個語料庫是一個帶語法詞類標記的語料庫，並且利用附加名物化特徵的方式來呈現派生名詞(未附加名物化特徵的動詞即表示此動詞扮演單純的謂語角色)

如：他(Nh) 無法(D) **忍受(VK)** 學生(Na) 的(DE) 不(D) **合作(VH)[+nom]**

若以這個例子作為測試句，系統的輸入和輸出如下：

輸入：他 無法 **忍受** 學生 的 不 **合作**

輸出：他(Nh) 無法(D) **忍受(VK)** 學生(Na) 的(DE) 不(D) **合作(VH)[+nom]**

輸入是一個正確的斷詞結果(因為從語料庫得到)。輸出是帶語法詞類標記並附加名物化特徵的詞串，理想的輸出結果跟語料庫所標注的內容應該完全一樣，比較兩者即可評估模型的表現。表 1 是語料庫的詳細資料，“總詞數”表示語料庫中所有詞的個數總和，“動詞詞數”表示語料庫中所有動詞(即未名物化以及名物化的動詞)的個數總和，“名物化動詞詞數”表示語料庫中所有

名物化的動詞的個數總和，“名物化比率”為“名物化動詞詞數”除以“動詞詞數”的值。

表 1. 平衡語料庫的名物化情形

總詞數(W#)	動詞詞數(V#)	名物化動詞詞數(Nv#)	名物化比率(RofNv)
5821222	1205208	90951	7.5%

我們將平衡語料庫按照主題切分成六大類-文學、生活、社會、科學、哲學以及藝術。在實驗當中我們會個別測試這六大主題及其綜合表現。表 2 則顯示各主題的名物化情形。

表 2. 六大主題的名物化情形

	W#	V#	Nv#	RofNv
文學	939656	203863	5974	2.9%
生活	1061336	212387	14062	6.6%
社會	2330536	486901	47759	9.8%
科學	399934	77828	7739	9.9%
哲學	527602	112703	4869	4.3%
藝術	562158	111526	10548	9.5%

除了按照主題切分之外，我們也按照語式來切分平衡語料庫，在實驗當中我們會個別測試不同的語式-書面語、口語及其綜合表現。表 3 則顯示各語式的名物化情形。

表 3. 不同語式的名物化情形

	W#	V#	Nv#	RofNv
書面語	5193355	1086887	87130	8.0%
口語	627867	118321	3821	3.2%

藉由主題以及語式的分類，讓我們對不同主題和語式的名物化情形有所了解。從系統評估的角度來說，綜合評估之外再加上個別評估可以反映出現實中的不同需求及應用，測試出模型的強健性。

我們將平衡語料庫的 80%當作訓練語料，20%當作測試語料。訓練語料和測試語料當中的主題分佈比率或是語式分佈比率都跟原平衡語料庫相同。也就是說，以文學類為例，它的訓練語料當中，有四倍於其測試語料的文學類訓練語料，以及其餘五類的訓練語料。

4.2. 評量標準

實驗所關注的焦點是動詞名物化的判斷是否正確，我們以派生名詞召回率(recall)、準確率(precision)以及綜合兩者的 F-Score 來評量。

Recall(R)=Nv_Match# / Nv#

Precision(P)= Nv_Match# / Result_Nv#

F-Score(F)=2*R*P/(R+P)

上式的 Nv#是參考語料的名物化動詞詞數，Result_Nv#表示輸出的派生名詞個數，Nv_Match#表示“派生名詞相符”的個數。上述所謂“派生名詞相符”，指的是針對某一個動詞，參考語料和輸出結果都標明它是派生名詞。

4.3. 實驗結果

表 4. 不同主題的 HMM 測試結果

	HMM-1			HMM-2		
	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)
文學	47.7	62.2	54.0	52.3	66.0	58.3
生活	59.4	66.3	62.7	61.3	67.1	64.1
社會	69.4	67.3	68.3	71.6	69.4	70.5
科學	63.4	54.3	58.5	65.8	56.7	60.9
哲學	59.4	58.6	59.0	62.5	59.4	60.9
藝術	64.6	67.8	66.2	67.1	69.2	68.2
綜合	65.9	65.7	65.8	68.3	67.7	68.0

表 5. 不同主題的 Classifier 測試結果

	Classifier-1			Classifier-2			Classifier-3		
	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)
文學	53.6	75.7	62.8	50.7	73.1	59.9	51.2	79.1	62.1
生活	65.0	76.2	70.2	63.7	75.7	69.2	63.8	79.6	70.9
社會	75.9	75.6	75.7	76.3	75.9	76.1	75.7	78.2	76.9
科學	68.9	62.4	65.5	70.3	60.7	65.1	70.0	64.7	67.3
哲學	63.5	68.1	65.7	62.0	66.7	64.3	60.0	69.2	64.3
藝術	71.9	72.2	72.1	71.9	73.5	72.7	71.4	74.7	73.0
綜合	72.3	74.0	73.1	72.3	73.9	73.1	71.8	76.6	74.1

圖 1. 不同主題的 HMM 以及 Classifier 測試結果

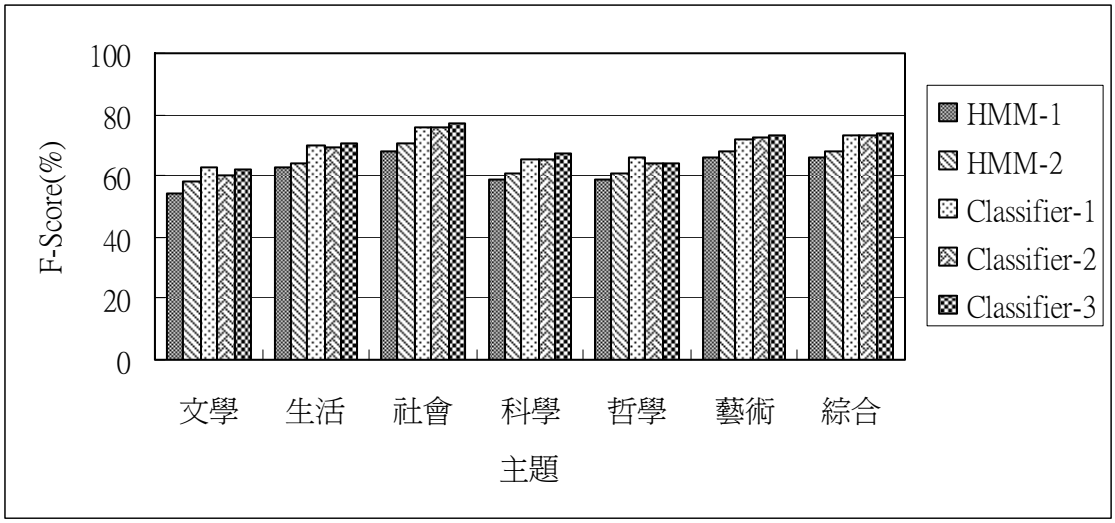


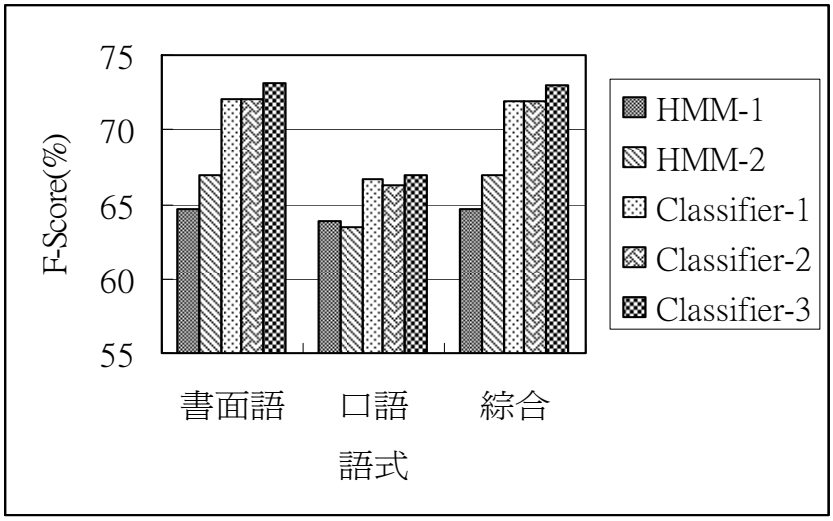
表 6. 不同語式的 HMM 測試結果

	HMM-1			HMM-2		
	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)
書面語	63.8	65.6	64.7	65.9	68.0	66.9
口語	63.4	64.1	63.8	65.0	61.9	63.4
綜合	63.8	65.6	64.6	65.9	67.8	66.9

表 7. 不同語式的 Classifier 測試結果

	Classifier-1			Classifier-2			Classifier-3		
	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)
書面語	69.5	74.7	72.0	70.1	74.0	72.0	69.3	77.4	73.1
口語	65.7	67.8	66.7	64.4	68.2	66.3	64.2	69.7	66.9
綜合	69.4	74.5	71.9	70.0	73.9	71.9	69.2	77.3	73.0

圖 2. 不同語式的 HMM 以及 Classifier 測試結果



4.4. 實驗討論

從主題的角度來看，表 2 顯示出，社會、科學以及藝術的名物化比率較高(分別為 9.8%、9.9%以及 9.5%)。從表 4 和表 5 我們可以大致歸納出名物化比率和測試結果的關係：無論是隱藏式馬可夫模型或是貝式分類器，在測試結果上都顯示出社會、藝術類擁有最佳的測試結果(分占第一、二名)，顯示出名物化比率越高，名物化的判斷越準確。例外的是科學類，雖然名物化比率(9.9%)比生活類(6.6%)高，但是所有模型的測試結果都顯示出科學類的判斷結果輸給生活類，而大致居於第四名的位置，這極可能跟語料的大小有關係(生活類的語料大約是科學類的 2.65 倍)。從計算語言學的角度來看，當我們能夠從訓練語料取得越多的判斷線索或是相關例證時，系統自然較能夠利用這些出現過的資料來判斷標的。因此，對統計式的名物化判斷模型來說，較高的名物化比率以及較大的語料會得到較準確的判斷結果。根據表 3、表 6 以及表 7，我們發現語式也有著一樣的關係：擁有較高名物化比率以及較大語料的書面語比低名物化比率以及較小語料的口語有

著更準確的判斷結果。

由以上的討論出發，若是我們想提升綜合、整體的名物化判斷結果，一個簡單直接的途徑就是增加那些低名物化比率的訓練語料，例如：文學類(名物化比率 2.9%)、口語類(名物化比率 3.2%)，來改善它們的判斷結果，而使綜合的判斷結果也跟著提升。

4.4.1 HMM-1 vs HMM-2

HMM-1 和 HMM-2 唯一的差別在於 HMM-1 未區分動詞來源詞語法詞類，而 HMM-2 區分了動詞來源詞語法詞類(請見 3.1 節)。對每一個動詞的上下文來說，HMM-1 是假設一個特定的上下文對每一個不同的來源詞都具有同樣的名物化催化力，HMM-2 則是假設一個特定的上下文對同一語法詞類的不同來源詞具有同樣的名物化催化力，而對不同語法詞類的來源詞具有不一樣的名物化催化力。

由表 4 及表 6 可以發現，HMM-2 大體上在每一個參考語料下的每一個評量指標(R、P、F)都比 HMM-1 來得高(唯一例外是口語的 P 及 F)。這告訴我們 HMM-2 的假設應該是比較正確的。來源詞語法詞類在名物化判斷上面的確扮演了重要的角色。

這樣的實驗結果也可以從語言學上找到解釋，[葉美利等 1992]以及[Huang et al., 1994]都指出：一個動詞，無論是單純謂語或者是派生名詞，其論元成分往往是一致的。在上下文中可以找到相對應的論元成分。由表 4 的實驗結果和實際的語料觀察，我們進一步地發現某些動詞的語法詞類，他們無論是扮演單純謂語角色或是轉換成派生名詞，相對應的論元成分往往在上下文中也具有相當固定的位置。舉例來說，動作不及物述詞(VA)，當它由單純謂語轉換成派生名詞時，單純謂語的主事者(agent)會從原本單純謂語的前面跑到派生名詞的後面緊鄰位置，當名詞片語的中心語(head word)，作為派生名詞的修飾對象。

如：學生(Na)　示威(VA) → 示威(VA)[+nom] 學生(Na)
　　學生(Na)　違規(VA) → 違規(VA)[+nom] 學生(Na)

另一方面，由於動作不及物述詞在單純謂語時，在語法上不緊鄰任何受詞，所以綜合上述，當我們看到一個 VA，其後緊鄰詞彙“學生”，幾乎可以斷定這個 VA 會轉換成派生名詞當“學生”的修飾語。

動作單賓述詞(VC)又是不同的情況，它在單純謂語時的主事者通常不會成為它在轉換成派生名詞後的修飾對象。

如：學生(Na)　攻擊(VC) 警察(Na) ╳→ 攻擊(VC)[+nom] 學生(Na)

“攻擊(VC)[+nom] 學生(Na)”是不合中文語法的。原因是，VC 在單純謂語時通常會後接一受詞，當我們看到一個 VC，其後緊鄰詞彙“學生”，通常會認為這個 VC 是單純謂語，而學生是它的受詞。也就是說，為了避免歧異，中文似乎不容許 VC 由單純謂語轉換成派生名詞去修飾原本單純謂語時的主事者。

因此，上述的觀察可以解釋為何同樣的上下文(如：後接“學生”)，針對相同來源詞語法詞類的不同動詞(如“示威(VA)”和“違規(VA)”)，有著相似的名物化催化力，而對不同的來源詞語法詞類(如：VA 和 VC)具有不同的名物化催化力。

4.4.2 HMM vs Classifier

由圖 1 和圖 2 可以看出 Classifier 模型均比 HMM 模型來得好。

HMM 的優點是它的目的是求整體最佳詞類序列，能夠將相鄰詞類之間的互相影響都考慮進去。但是缺點是未使用相鄰詞彙的訊息。Classifier 則正好相反，它的優點是它除了使用相鄰詞類作為上下文的元素之外，更把相鄰的詞一併考慮進來，使得所利用的上下文訊息更加豐富，而缺點是它是針對一個個的判斷標的，獨自得到個別的最佳判斷結果。而不是整體的最佳解。

由實驗結果看來，上下文的詞彙提供了名物化判斷更準確的鑑別能力以及更佳的強健性，在名物化判斷上是不可或缺的關鍵特徵。我們目前所提出的 Classifier 模型都只有使用左右相鄰(window size=1)的詞彙和詞類，未來還可再進一步觀察離判斷標的(window size>=2)較遠的詞彙及詞類對名物化判斷的影響。

4.4.3 Classifier-1 vs Classifier-2 vs Classifier-3

由表 5 和表 7 可以發現 Classifier-1 和 Classifier-2 的效果相當類似。

理論上來說，Classifier-1 應該比 Classifier-2 的表現好很多才對，因為在使用同樣上下文特徵的情況之下，Classifier-1 利用了來源詞本身這個更加精細的訊息，而 Classifier-2 卻只有使用來源詞語法詞類這個較粗糙的訊息。但是兩種模型所顯示的結果卻只有極些微的差距。事實上，這樣的結果再次印證了 4.4.1 節的討論。也就是說，一個特定的上下文對同一語法詞類的不同來源詞具有相似的名物化催化力，而對不同語法詞類的來源詞具有不一樣的名物化催化力。所以當我們用來源詞語法詞類代替來源詞本身時，其表現仍然維持幾乎一樣的水準。

因此我們可以推論，來源詞語法詞類可以提供一定程度的鑑別力以及優秀的強健性，來源詞本身則可以提供一定程度的強健性以及優秀的鑑別力，當我們把這兩樣訊息整合在同一個模型 Classifier-3 的時候，即得到了最好的表現結果以及最佳的強健性。

5. 結論

在本論文中，我們提出了兩種 HMM 模型以及三種貝氏分類器作為自動名物化判斷的系統，針對不同模型的表現，我們分析了表現差異的因素並從語言學上的角度加以驗證，其中，最值得注意的是，我們發現派生名詞的動詞來源詞(verbal counterpart)的語法詞類和其語境有密不可分的關係。藉著實際語料的分析，我們發現某些動詞語法詞類，他們無論是扮演單純謂語角色或是轉換成派生名詞，相對應的論元成分的位置在上下文中往往也遵循了某種固定模式。因此，當我們整合這個訊息到模型當中，即會使得系統的表現有顯著提升。表現最佳的統計模型對於派生名詞的包含率為 71.8%，準確率為 76.6%，F-Score 為 74.1%。

6. 未來研究方向

針對自動名物化判斷的未來研究主要有兩大方向，第一個方向是進一步使用範圍更大的語境資訊（如 window size 的加大）或者是更細緻的語法、語意資訊（如使用語意類別）。由於我們已經知道派生名詞的來源詞語法詞類在判斷上可以扮演關鍵角色，我們很好奇派生名詞的來源詞語意訊息(如語意類別)或者是詞構，是否也是重要的判斷因素。

另一個方向是當我們對自動名物化判斷的影響因素已經了解清楚後，最新的統計式技術，如“條件式隨機場域”(Conditional Random Field)或是“支援向量機器”(Support Vector

Machine)等，就可以嘗試來解這樣的問題。藉由最佳化的機器學習技巧，應該可以更準確的求得這些因素相互之間的關係以及使用比重。除了得到更佳的判斷結果之外，也可據此驗證或挖掘出更深入的名物化相關的語言現象。

7. 參考文獻

1. 葉美利、湯志真、黃居仁、陳克健,“漢語的動詞名物化初探—漢語中帶論元的名物化派生名詞, ROCLING V, pp177~193, 1992
2. 洪偉美、黃居仁、湯志真、陳克健,“中文派生詞的構詞規律初探”,第三屆世界華文教學研討會, 1991
3. Huang, Chu-Ren, Meili Yeh, and Li-Ping Chang, “A Corpus-based Study of Nominalization and Verbal Semantics: Two Light Verbs in Mandarin Chinese”, Proceedings of the Sixth North American Conference on Chinese Linguistics. Los Angeles: GSIL, USC. pp. 106-120, 1994.
4. Lin, Koong H.C., Von-Wun Soo, and Sandiway Fong, “Dealing with Nominalizations in Mandarin Chinese Using a Principles and Parameters Parser”, Computer Processing of Oriental Languages 11(3). pp. 291-307, 1998.
5. “中央研究院平衡語料庫的內容與說明”詞庫小組技術報告 95-02/98-04
6. Jane Grimshaw, Argument Structure, the MIT Press, 1990
7. Huang, Chu-Ren. “Mandarin Chinese NP de -- A Comparative Study of Current Grammatical Theories”, Special Publication No. 93 of the Institute of History and Philology, Academia Sinica. 1989.
8. Tsai, Yu-Fang and Keh-Jiann Chen, "Reliable and Cost-Effective Pos-Tagging", Proceedings of ROCLING XV, pp. 161-174, 2003
9. Tsai, Yu-Fang and Keh-Jiann Chen, "Context-rule Model for POS Tagging", Proceedings of PACLIC 17, pp. 146-151, 2003
10. Tsai, Yu-Fang and Keh-Jiann Chen, 2004, "Reliable and Cost-Effective Pos-Tagging", International Journal of Computational Linguistics & Chinese Language Processing, Vol. 9 #1, pp. 83-96, 2004.