以語境判定中文未知詞詞類的方法

白明弘 陳超然 陳克健
中央研究院資訊科學研究所
e-mail: evan@iis.sinica.edu.tw, richard@iis.sinica.edu.tw, kchen@iis.sinica.edu.tw
Fax:(02)2788-1638

摘要

從中研院平衡語料庫估算，未知詞在實際的文章中約佔 3.51%，由於這些詞無法直接從辭典中獲得詞類訊息，所以必須以猜測的方式來獲得未知詞的詞類。在[Chen et al. 97]中曾經以詞首字及詞尾字和詞類的關係來猜測未知詞的詞類，其前三名猜測的覆蓋率可以達到約 96%，然而第一名召回的正確率只有 76%。本文將基於此一猜測方法，提出以語境規則來協助判定未知詞詞類的方法。在實際的測試中，第一名召回的正確率可以提升到 83.83%。

1. 緒論

在中文自然語言分析系統中，從辭典中查詢一個詞彙的詞類訊息是最基本的工作。然而，並不是所有的詞彙都可以在辭典中找到，這些辭典中找不到的詞稱爲未知詞。未知詞大部分是一些專有名詞或是複合詞等，因爲無法被窮舉，所以在一般的辭典中不會收錄。未知詞在實際文章中大約佔 3.51%，由於這些詞無法直接從辭典中獲得詞類訊息，而必須依賴詞彙本身的結構以及語境來猜測，所以詞彙的詞類猜測研究成爲不可避免的課題。在西方語言的研究當中，未知詞詞類的猜測方法十分依賴詞綴的構詞律[Mikheev 96]。但是在中文裡，詞彙之間沒有空白字元作區隔，而詞彙本身也沒有規則的詞尾變化，詞綴衍生詞的構詞律只能解決一小部分的問題而已。在[Chen et al. 97]中曾經提出以詞首字及詞尾字和詞類間的相互關係來猜測未知詞的詞類，這個方法猜測未知詞的效率，其前三名的覆蓋率可達約 96%，但是第一名召回的正確率卻降到 76%。爲了更準確地猜測未知詞的詞類，我們嘗試以這個猜測方法所猜測的前三名詞

類做爲候選詞類，而以語境規律從候選詞類中選擇最可能的未知詞詞類。

1.1 未知詞的型態

在中文未知詞的處理當中，不同構詞型態的未知詞，其處理的方式差異可能非常大。在目前所發表的論文當中，多半把不同型態的未知詞以不同的主題來研究。光是專有名詞一類就有專門探討中式人名識別[孫茂松等 94]、音譯人名識別[Lee 94]、以及組織名識別[陳信希等 94]等等的論文。我們從中央研究院平衡語料庫[Chen et al. 96]中，分析未知詞的構詞型態，歸納出未知詞最常發生的幾種類型，分列如下：

(a)略語：例如'中油(Nb)'，'台汽(Nb)'。略語的構詞律非常不規則，他們的詞首詞尾不大能夠反映出詞的意義[Huang 94]。

(b)專有名詞：例如'陳壽(Nb)'，'電機科(Nc)'，'香檳城(Nc)'，'微軟(Nb)'。專有名詞可以進一步分成三類，人名、地名以及組織名。在不同的類中，有不同的關鍵字可以識別。中國人名的姓氏的用字 90%集中在 114 個字裡[孫等 1994]，地區名常常以'市'、'鄉'等字爲結尾。而組織名則比較沒有規則，其構詞成分幾乎沒有任何限制。

(c)衍生詞：例如'電腦化(Vh)'。衍生詞有詞綴，是很好的識別指標。

(d)複合詞：例如'轉赴(VCL)'，'獲允(VE)'，'搜尋法(Na)'，'電腦桌(Na)'。複合詞的構詞比較複雜。

(e)數字型複合詞：例如'1986 年(Nd)'，'三千'，'19 巷(Nc)'。數字型複合詞的特點是他們都包含數字，例如日期、時間、電話號碼、地址、數字和定量式複合詞等等，都屬於這種型態。這類的複合詞比較規律，可以使用構詞律來識別。

然而不同類型之間的未知詞可能會有歧義或相同的構詞律，例如'陳年品'可能是一個普通名詞，也可能是人名。從構詞成分不容易區分，但是可能比較容易從未知詞出現的上下文中判別其正確的詞類。

1.2 語境和詞類的關係

在[Chen et al. 97]的研究中利用詞首和詞尾字與詞類間的關係，找出和詞首字與詞

尾字之間相關訊息量(mutual information)[Blahut 87, Su 96]及 Dice 測度(dice measure)[Su 96, Smajda 96]可以計算出相關性最強的詞類。前三名的猜測覆蓋率律可達 96%，然而第一名召回的正確率只有 76%。很顯然只靠詞首詞尾的訊息，不太容易正確的判定未知詞的詞類，但是可能很容易的去除大部分不可能的詞類。根據我們在語料庫中的觀察，未知詞的詞類和語境之間有某種程度的關係。例如'院長'和'所長'後面所接的未知詞 90%以上都是人名（Nb 類），'位於'後面所接的未知詞 85%以上都是地名（Nc 類），'民國'和'西元'後面所接的未知詞 99%以上都是時間名詞（Nd 類），類似的情形不勝枚舉，表一列出幾個例子。由於詞類和前後文語境之間有選擇性的關係，因此可以利用語境訊息，從未知詞的可能候選詞類中，進一步選出最可能的詞類，而不直接用候選詞類的第一名作爲未知詞的詞類猜測。採用語境規則進一步選擇詞類的方法，將遭遇兩個困難，第一個困難是如何找出有用的語境規則，由於語境的多樣性，不同的未知詞和語境之間有太多太複雜的搭配關係，不容易以人爲的方式歸納出規則，第二個困難是語境規則和候選詞類的原始權重如何調整，以找到最佳的平衡點。本論文提出一個從語料庫中自動抽取語境規則的方法，並且利用這些語境規則來協助判定未知詞的詞類。

語境與未知詞詞類	語境出現頻率	詞類相符的機率
公共 -> (Na)	12	0.916667
缺乏 -> (Na)	14	0.928571
主任 -> (Nb)	144	0.993056
市長 -> (Nb)	130	0.953846
位於 -> (Nc)	84	0.916667
民國 -> (Nd)	594	1.000000
西元 -> (Nd)	149	1.000000
(A) <- 麻痺	16	0.937500
(Na) <- 肝炎	106	0.990566
(Na)<- 們	37	0.972973
(Nb) <- 小姐	87	0.977011
(Nb) <- 將軍	54	0.981481
(Nb) <- 教授	155	0.851613

表一、語境與未知詞詞類之間的相關性

2. 詞類判定規則的取得

在本文中我們嘗試從語料庫中學習由語境來判定詞類的規則。訓練語料庫是使用中央研究院平衡語料庫 2.0 版，其中包含了 350 萬詞。語料庫中的每一個句子都已經經過分詞的處理，以空白字元隔開每一個詞，並且在每一個詞的後面有詞類標記。我們將語料庫分成兩個部分，其中的 300 萬詞當作規則的訓練資料，另外的 50 萬詞則當作測試資料。

2.1 規則的抽取

語境規則描述了一個未知詞詞類在實際語料庫中，與語境之間的關係。例如規則'所長->Nb'，說明了在語料庫中，所長一詞後面所接的未知詞應該是 Nb 類。本實驗抽取語境規則的方法是類似於[Brill 95, Chen 97]所使用的錯誤驅動學習法(error-driven learning method)，不同於 Brill 方法的地方在於，此一研究是用未知詞驅動而非錯誤驅動。在抽取規則之前，必須先設定規則的基本型態。本實驗總共使用了九種基本的型態。這九種型態如下所示：

==========

a. $word_{-1}$ -> category, 例如：'所長->Nb'，'很->VH'。
b. $word_{+1}$ -> category, 例如：'先生->Nb'，'警局->Nc'。
c. $category_{-2}$, $category_{-1}$ -> category, 例如：'A,Caa->A'，'Nb,Caa->Nb'。
d. $category_{+1}$, $category_{+2}$ -> category, 例如：'Caa,A->A'，'Caa,Nb->Nb'。
e. $category_{-1}$, $category_{+1}$ -> category, 例如：'VJ,VJ->Na'，'A,D->Na'。
f. $word_{-2}$, $category_{-1}$ -> category, 例如：'女,Na->Nb'，'新任,Na->Nb'。
g. $category_{+1}$, $word_{+2}$ -> category, 例如：'Cab,人->Nb'，'D,組成->Na'。
h. $word_{-2}$ -> category, 例如：'有效->Na'，'位於->Nc'。
i. $word_{+2}$ -> category, 例如：'報導->Nb'，'大小->Na'。

==========

其中 $word_{-i}$ 表示未知詞的前第 i 個詞，$word_{+i}$ 表示未知詞的後第 i 個詞，同樣的 $category_{-i}$ 表示未知詞的前第 i 個詞的詞類，$category_{+i}$ 表示未知詞後第 i 個詞的詞類。

在抽取規則的方法上，使用已經標記好的語料庫爲本，抽取的程序如下：

規則抽取程序：

1. 對已經標記好的訓練語料庫中的每一個詞，如果有辭典中找不到的詞
   1.1 依據該詞的前後文，以及該詞所標記的詞類，產生 9 類的規則

以下面的句子爲例：

職位(Na) 低(VH) 的(DE) 不(D) 具(VJ) 裁決權(Na) ，(COMMACATEGORY) '裁決權(Na)'一詞沒有收錄在辭典中，所以依據裁決權的前後文以及其詞類，產生規則：

a . 具->Na
b. ，->Na
c. D,VJ->Na
d. --
e. VJ,COMMACATEGORY->Na
f. 不,VJ->Na
g. --
h.不->Na
i. --

其中 d, g, i, 類型由於條件不足，不產生規則。

2.2 規則的評分

並不是每一條從語料庫中抽出來的規則都具有相同的判斷能力，語境規則只能判斷一個未知詞詞類在某個語境狀態下的可能程度。例如'院長'一詞後面所接的未知詞 93.33%是 Nb 類，仍有 6.66%可能是其他的詞類。因此，爲了給每一條規則評分，我們必須對抽出的規則做了一些統計，計算對於每--條規則在訓練語料庫中匹配到未知詞的語境之次數，以及計算其正確判斷詞類之次數。評分的公式如下所示：

Score-of-Rule(r,cate) = 規則 r 在語料庫中正確匹配 *cate* 的次數 / 規則 r 在語料庫中匹配的次數

每一條語境與未知詞詞類關係的規則的分數，代表在某一語境之下，對於某一未知詞詞類的支持度。

3. 未知詞詞類的判別

在判別未知詞的程序上，主要分成兩個步驟：第一個步驟是以(詞首字,詞類)及(詞尾字,詞類)的相關訊息來猜測未知詞詞類，並取其前三名的猜測作爲候選詞類，第二個步驟是使用語境與未知詞詞類關係的規則，從候選詞類中選出比較可靠的詞類。本文的焦點主要在探討第二個步驟，第一個步驟請參考[Chen 97]。一個未知詞在匹配語

境與未知詞詞類關係規則的時候，一般在每一類型的規則中，都會匹配到一條規則，而獲得一個分數。在方法上，我們是以獲得分數最高的的詞類爲未知詞的詞類。在實驗中，我們一共使用了如表三中所列的 9 種類型的規則，所以一個未知詞將會得到 9 個分數，再加上第一個步驟猜測時獲得的分數，則一共有 10 個分數(註：如果未知詞的語境在某一種類型的規則中並未出現，則所得分數爲零)。由於第一個步驟和第二個步驟的評分標準不同，並且 9 種規則對於詞類判斷的能力也不盡相同。所以在加總分數的時候，每一個分數都有一個權重。如何調整權重以找到最佳的平衡點，是實驗中的一項難題。

3.1 權重的調整

一個未知詞詞類和它的語境在一種類型的語境規律中，只會匹配到一條規則。所以對於未知詞的詞類 *cate* 而言，在語境規律類型 *j* 中只會匹配到一條規則假設爲 $Rule_n$，則語境規律類型 *j* 對於 *cate* 的評分可以表示成：

$$Score\text{-}of\text{-}Rule\text{-}Type_j(\text{cate}) = Score\text{-}of\text{-}Rule(Rule_n, \text{cate})$$

，如果將每一個分數乘上一個比重來求得總分，則一個未知詞總分的計算式可以表示爲：

$$score(cate) = MI(cate) + \sum_{j=1}^{9} W_j \cdot Score\text{-}of\text{-}Rule\text{-}Type_j(cate)$$

其中 *cate* 表示未知詞的候選詞類，MI 表示以(詞首字,詞類)及(詞尾字,詞類)相關訊息量賦予詞類 *cate* 的分數，W_j 表示第 *j* 類型規則的權重，$Score\text{-}of\text{-}Rule\text{-}Type_j(cate)$表示詞類 *cate* 在第 *j* 類型規則中所匹配到規則的分數，。在權重調整上，是以貪婪法(greedy method)來取得最佳值。首先，給每一個權重一個初始值，然後在訓練語料庫中測試，得到一個召回正確率值。其次，調整 W_1 使召回正確率增加，直到召回正確率不再有明顯的增加爲止。接著用同樣的方法調整 $W_2 \ldots W_9$，調整完 W_9 之後再從頭調整一次，調整後，如果召回正確率還有明顯的增加，就再從頭調整，如此重複調整，直到召回正確率沒有明顯的增加爲止。

3.2 判別的演算法

在本章的開頭已經說明了判別未知詞的主要步驟，在觀念上十分單純。然而在判別未知詞的過程中還將遇到一個難題，語境中如果包含了未知詞將使得語境的匹配變的複雜許多。下面是一個未知詞的語境中包含未知詞的例子：

另(Nes) 建議(VE) 由(P) 政風室(Nc,0.482;Na,0.223;VA,0.162) 閻琴南(Nb,0.427;Nc,0.259;Na,0.148) 負責(VL) 推動(VC) 小組(Na) 成立(VC)[+nom] 事宜(Na) 。

在上例中要以語境規則判斷未知詞'政風室'的詞類，必須使用到'閻琴南'的詞類，不幸的是'閻琴南'本身也是一個未知詞。

如果每一個詞都是未知詞，要真正找到最佳詞類猜測，必須考慮所有的詞類組合，即使採用 dynamic programming 的方法，依然有太多及複雜的計算，因此我們只用從左到右的依序處理方式，只以區域最佳解(local maximal)爲滿足，不追求真正的全域最佳解(global maximal)。由於在實驗中，語境只使用到前面兩個詞以及後面兩個詞，所以，在處理時，可以假設一個包含 5 個詞的詞窗：

$word_{-2}(cate_{-2})\ word_{-1}(cate_{-1})uword(cate_{11}..cate_{1i})word_{1}(cate_{11}..cate_{1j})\ word_{2}(cate_{21}..cate_{2k})$

假設 uword 爲即將處理的未知詞，由於處理未知詞的方向是由最左邊的詞一個接著一個向右處理，所以 $word_{-2}$ $word_{-1}$ 已經處理過，可以視爲已知詞。而 $word_1$, $word_2$ 都還沒有處理過，將其視爲具有 i 個及 j 個候選詞類的詞，如此一來，對於已知詞而言，其候選詞類的個數爲 1，對於未知詞而言其候選詞類的個數爲 3。於是我們可以把問題看成是一個最佳路徑選擇的問題。假設 $score_{m,n}(cate_l)$表示未知詞的候選詞類 l，在語境 $word_1$ 的詞類爲 $cate_m$， $word_2$ 的詞類爲 $cate_n$ 時，所獲得的分數，則 $cate_l$ 的最佳分數可以簡化爲：

$$score_{opt}(cate_l) = \underset{m,n}{MAX}\{score_{m,n}(cate_l)\}$$

最後在比較每個候選詞類的最佳分數，然後選擇分數最高的候選詞類爲未知詞的猜測詞類。

以上面的句子爲例，要處理未知詞'政風室'的時候，詞窗爲：

建議(VE)　由(P)　政風室(Nc,0.482;Na,0.223;VA,0.162)　閻琴南(Nb,0.427;Nc,0.259;Na,0.148)　負責(VL)

對於'政風室'的候選詞類 Nc 而言，分別假設'閻琴南'的詞類爲 Nb, Nc, Na 三種路徑做處理。

路徑 1. '建議(VE)　由(P)　Nc　閻琴南(Nb)　負責(VL)' 得到 $score_{1,1}$(Nc)

路徑 2. '建議(VE)　由(P)　Nc　閻琴南(Nc)　'負責(VL)' 得到 $score_{2,1}$(Nc)

路徑 3. '建議(VE)　由(P)　Nc　閻琴南(Na)　負責(VL)' 得到 $score_{3,1}$(Nc)

所以'政風室'候選詞類 Nc 的最佳分數爲：

$score_{opt}$(Nc) = $MAX\{score_{1,1}$(Nc), $score_{2,1}$(Nc), $score_{3,1}$(Nc)$\}$

同樣的方法可以得到'政風室'候選詞類 Na 及 VA 的分數 $score_{opt}$(Na)及 $score_{opt}$(VA)。最後比較 Nc, Na, VA 三個候選詞類的最佳分數，假設 Nc 的最佳分數最高，則選擇 Nc 爲'政風室'的詞類。

4. 實驗結果

本實驗的測試資料是使用中央研究院語料庫，中研院語料庫 2.0 版總共有 350 萬詞。其中的 300 萬詞當成訓練語料庫，另外 50 萬詞當做測試語料庫。而其中出現在 CKIP 辭典中的詞被視爲已知詞。目前 CKIP 辭典一共收錄了大約八萬目詞，每一個詞項都包含他的語法類別以及文法訊息。一個沒有收錄在 CKIP 辭典中的詞，如果也沒有被識別爲外來語(通常是文章中夾雜的英文字)的話，則被視爲是一個未知詞。在中央研究院語料庫中，總共有 52 種不同的詞類標記。而其中只有 14 種詞類具有較高的滋生力，其他的詞類通常是虛詞或是低滋生力的詞類。因此，我們的實驗只針對 14 種高滋生力的詞類。在訓練語料庫中，一共有 135896 個未知詞，而在測試語料庫中則有 21588 個未

知詞。表二列出 14 此種詞類，以及未知詞在各種詞類的頻率分佈情形。

Category	Training	Testing	Meaning of the Categories
A	1911	285	/*non-predictive adjective*/
Na	37646	5641	/*common noun*/
Nb	42853	6619	/*proper noun*/
Nc	16346	2242	/*location noun*/
Nd	11845	2037	/*time noun*/
VA	3985	656	/*active intransitive verb*/
VC	8757	1663	/*active transitive verb*/
VCL	1484	307	/*active transitive verb with locative object*/
VD	642	134	/*ditransitive verb*/
VE	991	257	/*active transitive verb with sentential object*/
VG	1675	295	/*classificatory verb*/
VH	5437	1073	/*stative intransitive verb*/
VHC	683	88	/*stative causative verb*/
VJ	1641	291	/*stative transitive verb*/

表二、未知詞在 14 種高滋生力詞類中的分佈情形

4.1 規則抽取的結果

我們從語料庫中自動抽取了 9 種不同型態的語境—未知詞詞類關係的規則，並且計算每一條規則在語料庫中匹配到未知詞語境的頻率，以及正確匹配的頻率。例如 $word_1$ -> category 類的規則中，'院長->Nb'一共匹配了 45 次，其中有 42 次是正確的匹配，亦即在訓練語料庫中，'院長'一詞後面接未知詞一共出現了 45 次，其中有 42 次所接的未知詞是 Nb 類。在附錄一中所列的是一些規則的樣本。在 300 萬詞的訓練語料庫中，去除匹配次數小於 3 次的規則之後，一共抽得 113327 條規則，各類型的規則數量分佈如表三所示。

規則類型	規則條數
a. $word_{-1}$ -> category	13450
b. $word_{+1}$ -> category	13572
c. $category_{-2}$ $category_{-1}$ -> category	7238
d. $category_{+1}$ $category_{+2}$ -> category	6802
e. $category_{-1}$ $category_{+1}$ -> category	8513
f. $word_{-2}$ $category_{-1}$ -> category	15943
g. $category_{+1}$ $word_{+2}$ -> category	15027
h. $word_{-2}$ -> category	16125
i. $word_{+2}$ -> category	16657
規則總數	113327

表三、各類型語境與未知詞詞類關係規則的數量分佈

4.2 詞類判別的結果

在探討結果以前先定義實驗的召回率、精確率以及覆蓋率：

召回率 = 未知詞詞類爲 cat 且被正確猜測的詞數 / 未知詞詞類爲 cat 的總詞數

精確率 = 未知詞詞類爲 cat 且被正確猜測的詞數 / 被猜測爲詞類 cat 的總詞數

前 n 名覆蓋率 = 未知詞詞類爲 cat 且正確詞類包含在猜測的前 n 名內 / 未知詞詞類爲 cat 的總詞數

表四是以 300 萬詞的訓練語料庫做內部測試(inside test)的結果。其中第三欄 MI(1) 的召回率是表示，只以詞首字—詞類及詞尾字—詞類關係猜測未知詞詞類，取其第一名爲未知詞詞類的召回率。第六欄的召回率則是以語境與未知詞詞類關係規則輔助猜測的召回率。

類別	未知詞數	MI(1)召回率	MI(1)精確率	MI(3)覆蓋率	召回率	精確率
A	1911	89.80%	34.74%	98.90%	78.23%	73.83%
Na	37646	69.77%	89.62%	97.20%	90.62%	87.87%
Nb	42853	85.59%	91.83%	97.57%	93.83%	92.06%
Nc	16346	85.05%	79.90%	97.92%	81.23%	92.43%
Nd	11845	97.89%	91.48%	99.21%	97.89%	96.65%
N	108690	90.45%	98.98%	99.21%	98.06%	97.91%
VA	3985	65.04%	43.78%	94.93%	60.43%	76.91%
VC	8757	67.08%	80.48%	98.16%	86.64%	85.91%
VCL	1484	95.89%	44.99%	99.53%	93.19%	68.94%
VD	642	95.17%	56.37%	99.69%	94.39%	84.17%
VE	991	92.23%	44.30%	98.99%	84.66%	67.28%
VG	1675	98.99%	82.24%	99.82%	98.69%	88.40%
VH	5437	71.20%	63.56%	94.74%	71.11%	87.51%
VHC	683	98.24%	57.84%	99.71%	96.78%	75.80%
VJ	1641	88.42%	50.50%	98.54%	85.74%	73.55%
V	25295	94.84%	75.75%	99.68%	91.73%	92.78%
總計	135896	80.37%		97.61%	89.11%	

表四、內部測試的結果

從表四的召回正確率來看，詞首—詞類及詞尾—詞類關係猜測的召回正確率爲80.37%，經過語境規則從前三名中輔助判定詞類召回正確率達到 89.11%，召回正確率提高了 8~9 個百分點。而其中 Na 類、Nb 類及 VC 類的召回率提升了，但是其他類的詞類召回率反而下降了。追究其原因發現 這些召回率下降的詞類，他們原本猜測的精確率都比語境規則判斷的精確率低很多，也就是說，用語境規則的猜測方式提高了各類詞類猜測的精確率，並且提高了整體的召回正確率。

類別	未知詞數	MI(1)召回率	MI(1)精確率	MI(3)覆蓋率	召回率	精確率
A	285	81.75%	27.74%	91.93%	63.16%	52.33%
Na	5641	65.48%	85.55%	96.10%	86.23%	82.52%
Nb	6619	82.14%	91.64%	96.90%	90.06%	90.46%
Nc	2242	83.99%	77.52%	96.16%	75.69%	86.54%
Nd	2037	96.02%	86.51%	97.89%	95.19%	92.55%
N	16539	88.87%	98.37%	98.85%	96.77%	96.76%
VA	656	53.96%	38.52%	89.94%	47.71%	64.94%
VC	1663	60.61%	76.60%	96.39%	79.49%	80.46%
VCL	307	92.83%	47.42%	98.37%	83.06%	63.43%
VD	134	94.03%	56.00%	96.27%	90.30%	79.61%
VE	257	86.77%	49.78%	97.28%	78.21%	67.91%
VG	295	95.59%	73.63%	99.32%	95.59%	79.21%
VH	1073	66.73%	59.42%	92.17%	62.44%	77.19%
VHC	88	89.77%	41.58%	94.32%	84.09%	51.39%
VJ	291	84.19%	47.12%	96.22%	75.26%	60.66%
V	4764	93.70%	76.87%	99.71%	89.67%	90.82%
總計	21588	76.53%		96.19%	83.83%	

表五、外部測試的結果

表五是外部測試的結果，以詞首—詞類及詞尾—詞類關係猜測的召回正確率爲76.53%，經過語境規則從前三名的猜測中判斷詞類，召回正確率提升到 83.83%，提升了約 7~8 個百分點。我們從表中發現跟內部測試類似的結果，Na 類, Nb 類,及 VC 類的召回率提升了，而其他詞類的召回率下降了，但是召回率下降的詞類，其精確率都有非常顯著的提升。

4.3 與二連詞(bigram)詞類標記模型的比較

在實驗中，我們也嘗試使用二連詞(bigram)的統計機率模型[Church 1993, Su 1996]，來輔助預測未知詞的詞類。二連機率模型的公式如下：

$$cat'_i = \arg\max_{cat_i} P(cat_i \mid cat_{i-1}) * P(word_i \mid cat_i)$$

二連機率 $P(cat_i|cat_{i-1})$

$P(cat_i|cat_{i-1})$和 P(cat|word)同樣是以中央研究院平衡語料庫的 300 萬詞估算而得。由於每一個未知詞的 P(word|cat)機率值無法從語料庫中估算，因此我們以 P(cat|word)取代。而 P(cat|word)的值是以不同詞類猜測值爲權重分配而得。例如未知詞'陳年品'以 Dice 測度前三名猜測依序爲 Nb、Na、以及 VC 類，其猜測值分別爲：

猜測值(Nb) = 8.48

猜測值(Na) = 7.327

猜測值(VC) = 2.956

因此我們假設：

P(Nb|陳年品) = 猜測值(Nb) / 總分 = 0.452

P(Na|陳年品) = 猜測值(Na) / 總分 = 0.390

P(VC|陳年品) = 猜測值(VC) / 總分 = 0.158

在實際演算過程中，我們是以 *W**log(P(cat|word))來取代未知詞 *word* 的 log(P(word|cat))值，*W* 爲調整的權重。對於未知詞而言，由於 $P(cat_i|cat_{i-1})$和 P(cat|word)兩個值的來源不同，在合併的時候，以 *W* 爲權重調整兩個值的比重以得到最佳的結果。實驗結果，同樣以中央研究院平衡語料庫另外的 50 萬詞爲測試語料庫，召回正確率從 76.53%提升到 79.97%，提升的幅度不如我們所提出的方法。

5. 結論與未來的研究

從語料庫自動產生判別規律的方法，經實際驗證是一種非常有效的方法，不但產生容易，而且有較佳的覆蓋率可以照顧到許多不同的類型。以未知詞的詞類判別而言，可以從語料庫中產生超過百萬條不同的規律，只是大部分是沒有什麼效果的。出現的頻率太低的規則可以被忽略而不會降低召回正確率。至於有些規則有包含關係，例如兩個語境相關，就可能包含於一個語境相關的規則中，只是它們各有不同的權重，因此並未被刪除。以語境與未知詞詞類關係的規則來輔助猜測詞類，大約可以提升 7~8

個百分點的召回正確率。距離前三名的覆蓋率還有一段距離，應該還有很大的提升空間。我們觀察所抽取的規則發現，有 50%的規則在語料庫中出現的頻率少於十次，並且在訓練語料庫中大部分的未知詞都是屬於 Na 類及 Nb 類，其他詞類的未知詞樣本太少，訓練出來的規則所具有的代表性不足，對於判斷的正確性有很大的影響。如何針對不同詞類調整權重，也是未來可能的研究。事實上，本研究所提出的方法，不僅適用於未知詞，也適用於任何具有多重詞類的已知詞的詞類判定上。因此，可以應用在詞類標記的工作上。此一方法和 Brill 所提出的方法，最大的不同是 Brill 的規律在給分時只有 0 和 1，而本文提出的方法，每一種規則有不同的權重，而且每一條規則有不同的給分。

參考文獻

陳信希、李振昌, 1994, "中文文本組知名之辨識" *Communications of COLIPS, VOL 4, NO 2, Page 131-142.*

孫茂松、黃昌寧、高海燕、方捷, 1994, "中文姓名的自動辨識" *Communications of COLIPS, VOL 4, NO 2, Page 143-149.*

Blahut, Richard E., 1987, Principles and Practice of Information Theory, *Addison-Wesley Publishing Company.*

Brill, Eric, 1995, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging." *Computational Linguistics Vol. 21, No. 4, pp. 543-566.*

Chen, C. J., M. H. Bai, K. J. Chen, 1997, "Category Guessing for Chinese Unknown Words." *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997: pp. 35-40. NLPRS '97 Thailand.*

Chen, H. H. and J. C. Lee, 1994, "The Identification of Organization Names in Chinese Texts." *Computer Processing of Chinese and Oriental Languages, Vol. 8, No. 1, June*

1994, pp. 75-85.

Chen, K. J., C. R. Huang, L. P. Chang & H.L. Hsu, 1996, "SINICA CORPUS: Design Methodology for Balanced Corpora." *Proceedings of PACLIC 11th Conference, pp. 167-176*

Chen, Keh-Jiann, Ming-Hong Bai, 1997, "Unknown Word Detection for Chinese by a Corpus-based Learning Method." *Proceedings of the 10th Research on Computational Linguistics International Conference, pp159-174.*

Huang, Chu-Ren, Wei-Mei Hong, and Keh-jiann Chen, 1994, "An Information Based Lexical Rule of Abbreviation." *the Proceedings of the Second Pacific Asia Conference on Formal and Computational.*

Church, K. W., & R. L. Merser, 1993, "Introduction to the Special Issue on Computational Linguistics Using Large Corpora." *Computational Linguistics, Vol. 19, #1, pp. 1-24*

Lee, J.C. , Y.S. Lee and H.H. Chen, 1994, "Identification of Personal Names in Chinese Texts." *Proceedings of 7th ROC Computational Linguistics Conference.*

Mikheev, A., 1996,"Unsupervised Learning of Word-Category Guessing Rules." *Proceedings of ACL-96.*

Smadja, F.A., K.R. McKeown, and V. Hatzivassiloglou, 1996, " Translating Collocations for Bilingual Lexicon: A Statistical Approach." *Computational Linguistics, Vol. 22, No. 1.*

Su, Keh-Yih, Tung-Hui Chiang, & Jing-Shin Chang, 1996," An Overview of Corpus-Based Statistics-Oriented Techniques for Natural Language Processing." *Computational Linguistics and Chinese Language Processing, vol. 1, no. 1, pp. 101-157.*

附錄一、語境—未知詞詞類關係規則的一些例子

a. $word_{-1}$ -> category
主任->Nb 144 0.993056
位於->Nc 84 0.916667
積極->VC 10 1.000000

b. $word_{+1}$ -> category
肝炎->Na 106 0.990566
般->Na 24 0.916667
女士->Nb 64 1.000000

c. $category_{-2}$, $category_{-1}$ -> category
Nb, PAUSECATEGORY->Nb 1267 0.930545
Nd, Caa->Nd 253 0.964427
Nh, Caa->Nb 84 0.916667

d. $category_{+1}$, $category_{+2}$ -> category
Caa, Nb->Nb 732 0.939891
Caa, Nd->Nd 313 0.913738
DASHCATEGORY, Nb->Na 191 0.989529
Nc, VE->Nb 583 0.914237

e. $category_{-1}$, $category_{+1}$ -> category
A, D->Na 82 0.975610
Neu, VG->Na 63 0.888889
A, P->Na 21 0.952381
Nb, VI->Na 13 0.923077

f. $word_{-2}$, $category_{-1}$ -> category
下午, Nd->Nd 72 0.972222
女, Na->Nb 32 0.937500
中心, Na->Nb 37 0.945946
平行式, Caa->A 4 1.000000

g. $category_{+1}$, $word_{+2}$ -> category
跟->Nb 16 0.937500
高興->Nb 13 0.923077
表示->Na 55 0.963636

h. $word_{-2}$ -> category
位於->Nc 59 0.915254
前任->Nb 14 0.928571
訂->Nd 21 0.952381

i. $word_{+2}$ -> category
掩埋場->Nc 20 0.950000
現任->Nb 13 0.923077

附錄二、測試結果的例子

第一行是經過自動斷詞與自動詞類標記後的結果，其中詞類欄內爲'?'的詞表示爲未知詞。第二行是經過詞首—詞類及詞尾—詞類關係猜測後的結果，每一個未知詞後面都有3個候選詞類，每一個候選詞類後面都有一個值，是猜測時所給的分數。第三行是經過語境—未知詞詞類關係規則判定詞類後的結果。

****

珍貴(VH) 的(DE) 古(VH) 陶壺(?) 並(D) 被(P) 視爲(VG) 貴族(Na) 的(DE) 傳家(VI) 之(DE) 寶(Na) ，

珍貴(VH) 的(DE) 古(VH) 陶壺(Na,0.421;Nb,0.215;VJ,0.159) 並(D) 被(P) 視爲(VG) 貴族(Na) 的(DE) 傳家(VI) 之(DE) 寶(Na) ，

珍貴(VH) 的(DE) 古(VH) 陶壺(Na) 並(D) 被(P) 視爲(VG) 貴族(Na) 的(DE) 傳家(VI) 之(DE) 寶(Na) ，

****

清代(Nd) 臺灣(Nc) 早期(Nd) 的(DE) 移墾(?) 者(Na) ,由於(Cbb) 其(Nep) 祖籍(Na) 之(DE) 不同(VH) ，

清代(Nd) 臺灣(Nc) 早期(Nd) 的(DE) 移墾(VCL,0.363; VC,0.287; VA,0.207) 者(Na) 由於(Cbb) 其(Nep) 祖籍(Na) 之(DE) 不同(VH) ，

清代(Nd) 臺灣(Nc) 早期(Nd) 的(DE) 移墾(VC) 者(Na) 由於(Cbb) 其(Nep) 祖籍(Na) 之(DE) 不同(VH) ，

****

另(Nes) 建議(VE) 由(P) 政風室(?) 閻琴南(?) 負責(VL) 推動(VC) 小組(Na) 成立(VC)[+nom] 事宜(Na) 。

另(Nes) 建議(VE) 由(P) 政風室(Nc,0.482;Na,0.223;VA,0.162) 閻琴南(Nb,0.427;Nc,0.259;Na,0.148) 負責(VL) 推動(VC) 小組(Na) 成立(VC)[+nom] 事宜(Na) 。

另(Nes) 建議(VE) 由(P) 政風室(Nc) 閻琴南(Nb) 負責(VL) 推動(VC) 小組(Na) 成立(VC)[+nom] 事宜(Na) 。

****

也(D) 肩負起(?) 更(D) 重大(VH) 的(DE) 任務(Na) 。

也(D) 肩負起(VJ,0.325;VC,0.293;VA,0.193) 更(D) 重大(VH) 的(DE) 任務(Na) 。

也(D) 肩負起(VC) 更(D) 重大(VH) 的(DE) 任務(Na) 。

****