中文動詞自動分類研究[1]

Automatic Classification of Chinese Unknown Verbs

曾慧馨、劉昭麟、高照明、陳克健

政治大學語言所、政治大學資訊系、台灣大學外文系、中央研究院資訊所

huihsin@iis.sinica.edu.tw;chaolin@cs.nccu.edu.tw;zmgao@ccms.ntu.edu.tw;kchen@iis.sinica.edu.tw

Abstract

We present a new method for automatic classification of Chinese unknown verbs. The method employs the instance-based categorization using the k-nearest neighbor method for the classification. The accuracy of the classifier is about 70.92%.

Keyword: unknown word, lexical similarity

1. 緒論

自然語言處理中重要的步驟是將中文文件斷詞並附加詞類標記；在斷詞標記的過程中會遇到的一個問題為未知詞的存在。現行的斷詞標記系統以辭典為基礎輔以構詞的規則訊息進行斷詞標記，但因為語言的特性之一「無窮盡的創造力」，無法窮舉出所有的詞彙；一本好的辭典也不應該無止盡的擴大所收錄的詞彙，因此如何辨識處理辭典中不存在的詞彙就成了一個重要的課題。

1-1 研究動機與目標

前人對於未知詞的探討重點集中在名詞細目的辨認上，如組織名、人名、地名辨識等(李振昌(1993)，李振昌、李御璽與陳信希(1994)人名辨識等等)。僅有 Chen、Bai 與 Chen(1997)利用前綴(prefix)、後綴(suffix)的訊息處理全部的未知詞，正確率約為 76%，而白明宏、陳超然與陳克健(1998)使用 Chen、Bai 與 Chen (1997)所提出的方法，再利用前後文的訊息來補強 Chen、Bai 與 Chen (1997)方法不足之處，將正確率提高至 83.83%。在動詞辨識正確結果不高的情況下，本論文將處理重心放在未知動詞的辨識處理上，並且希望將這種處理未知動詞的方法在未來可以轉移處理名詞與形容詞。

動詞不管在任何文法理論中，在剖析句子時都是位於最中心的部分，若動詞為未知詞，勢必將影響句子剖析的正確性。現代漢語的動詞結構繁複，內部規則複雜，若無足夠的語言訊息完全無法判斷其分類，我們認為動詞自動分類研究至

[1] 本論文中程式設計感謝馬偉雲、楊昌樺學長提供意見；兩位評審老師惠賜意見，特此致謝。

今無法提高正確率的主因為動詞繁複的內部結構。

我們的目標為將動詞自動分類到中研院詞庫小組(1993)的詞類架構上，動詞的詞類分類共有 15 類，但並非每一類都具有孳生性。有些類別如功能詞一般，屬於封閉性詞類，封閉性詞類為該分類中的詞彙不會增加，而在中研院詞庫小組的分類中 15 類中有 9 類是具有孳生性的分類；這 9 類分類中的動詞詞彙，會隨著語料庫的增長而增多，我們希望將未知動詞自動分類到這 9 類動詞分類中，這九類為動作不及物動詞(VA)、動作及物動詞(VC)、動作及物動詞＋地方賓語(VCL)、動作雙賓動詞(VD)、動作句賓動詞(VE)、分類動詞(VG)、狀態不及物動詞(VH)、狀態使動動詞(VHC)、狀態及物動詞(VJ)。

1-2 研究方法

本論文中未知詞的定義為不存在辭典中的詞彙。陳克健、陳超然(1997)分析未知詞的種類為兩種，第一種為封閉性，這一類型雖然在數量上可能為無數個，但是可用規則語法(Regular Expression)來產生與辨識，如：西元一九九九年(時間)、一千兩百七十二(數字)、二七八八三七九九(電話)等。第二類則為開放性，這一類的未知詞很難用規則語法來表達，複合詞即屬這一類。白明宏、陳超然與陳克健(1998)在分析中研院平衡語料庫後歸納出未知詞主要的分類為略語、專有名詞、衍生詞、複合詞與數字型複合詞。

未知動詞通常為複合詞，由兩個以上的組成成分組合而成，這種組成成分我們稱為詞基(base)[2]。趙元任(1968)、Li 與 Thompson (1981)與湯廷池(1988)提及漢語的複合詞具有特定的內部句法結構；如：「欺敵」，由「欺」與「敵」這兩個詞基組成，兩個詞基之間的關係為動賓結構。雖然詞基是有限的，但是詞基與詞基的組合數量龐大，且組合成分間的語意關係複雜，因此造成了我們無法將所有的未知動詞收錄進字典中。

在本論文中我們利用相似法來判斷動詞的分類，尋找未知動詞的相似詞，計算未知動詞與相似詞之間的相似度，再將這些相似詞依照詞類分組。從每個詞類當中取出 K 個相似詞出來，將這些相似詞的分數予以平均，得到未知動詞到每個詞類的平均距離，未知動詞的詞類即與其距離最相近的詞類。

---

[2]Sproat 與 Shih (1996) 稱內部的處理單位為詞根(root)，Chen、Bai 與 Chen (1997)稱處理的單位為前綴(prefix)與後綴(suffix)。我們則稱處理單位為詞基(base)，並採用 Katamba (1993:45) 對詞基(base)所下定義：“…a base is any unit whatsoever to which affixes of any kind can be added….In other words, all roots are bases. Bases are called stems only in the context of inflectional morphology.” 我們在此處決定使用詞基為我們切割的單位的原因在於詞基的定義較詞根 (root) 、詞幹 (stem) 寬鬆。未知動詞被我們斷詞系統切分出來很多單位，我們並不確定這些單位真正的意義，因此我們希望選用一個最寬鬆的定義可以涵蓋所有被斷詞系統所切分的單位。

1-3 語料分析與處理

我們在此介紹未知動詞的特性與可猜測未知動詞詞類的可能因素。首先，討論未知動詞的特性。未知動詞為複合詞，通常由數個具有孳生性的詞基所組成，本身語言具有高透明性。例如，未知動詞「求新」與「講錯」相對於列入辭典中的「忐忑」、「侷促」這一類的詞彙多具有語意透明性，並且可以從其組成成分預測出該詞的語意。

其次，我們認為有兩個因素可預測未知動詞的分類。一、語意。語意相近的詞彙，所屬的詞類應類似。我們將同義詞詞林中的語意類與中研院詞庫小組(1993)詞類作對應，中研院詞庫小組詞類有 45 類。平均來說，同義詞詞林一個語意類僅對應到詞庫小組 1.97 種詞類，即一個語意類中的詞彙共有的詞類數量。因此我們認為語意因素可左右詞彙的詞類。二、結構。結構通常會限定組成的詞類，若結構為“VC+Na”的未知動詞，通常會組成 VA 詞類，因為在這個未知動詞的內部結構中已經出現了一個普通名詞(Na)來滿足前面的動作及物動詞(VC)所要求的論元，在這種情形下通常會形成不及物動詞，因此我們認為結構會影響到動詞的詞類。

在本篇論文中我們利用這些線索尋找與未知動詞相似的詞彙，來預測未知動詞所屬的詞類。

2. 實驗方法

我們利用相似法來判斷動詞的分類，尋找未知動詞的相似詞，計算未知動詞與相似詞之間的相似度，再將這些相似詞依照詞類分組。從每個詞類當中取出 K 個相似詞出來，將這些相似詞的分數予以平均，得到未知動詞到每個詞類的平均距離，未知動詞的詞類即與其距離最相近的詞類。

2-1 相似法 (Instance-Based Categorization)

我們在這節說明如何使用相似法來預測動詞的分類。未知動詞的特性之一為組成成分屬於常用詞且語意明確，例如：試印、講完。這兩個詞彙都無法在辭典中查詢到，但我們卻很清楚的可以從字面上得知這兩個動詞的語意，而且這樣的組合方式是非常具有孳生性的，可以繼續孳生「唱完」、「說完」等等各樣的詞彙。

根據我們對未知動詞語料的觀察，未知動詞的組成雖然有一定的模式，但因為語言的複雜度，無法將所有的規則條列出來。因此我們在這邊使用相似法，將每個訓練語料中的未知動詞都當作是一條規則，當有新的未知動詞出現時，將其與所有的動詞做比較，測量新的未知動詞與訓練語料中的動詞的相似度，新的未知動詞與訓練語料中的動詞越相似時，新的未知動詞越有可能屬於與其相似動詞的詞類。例如：講完與唱完。若「講完」我們訓練語料中的動詞，「唱完」為我

們的未知動詞。未知動詞的第二個組成成分與訓練語料中的例子相同都為「完」，因此我們僅需要得知「講」與「唱」的相似度，若「講」與「唱」分屬的詞類相似度高，則表示「講」與「唱」的結構類似；若「講」與「唱」的語意相似程度高的話，則「唱完」的動詞分類則很可能與「講完」相同。

使用相似法的好處在於相似法所尋找的的相似詞，若相似度高的話，不僅可以預測詞類分類，同時也可以預測語意與結構分類。當兩個詞彙相似度高時，表示這兩個詞彙的詞類、語意類與結構必定相似。

我們在本節中首先介紹語意與詞類相似度的測量方法，接下來說明相似詞的選取與未知動詞詞類的預測。

2-2 相似度測量

在本論文中我們使用知網作為語意測量的工具，中央研究院中文句結構樹測量詞類相似度，介紹如下。

一、知網為一雙語(中文、英文)的知識性辭典，由董振東與董強編撰完成收錄約十一萬條詞條，知網系統中包含有中英雙語知識辭典、中文簡體知識辭典、中文繁體知識辭典、概念特徵、動態角色與屬性、詞類表、反義關係表、對義關係表、標示符號與說明、知網管理程序等。我們在本節當中將介紹如何使用知網計算語意相似度與評量方法。

二、中央研究院中文句結構樹資料庫 1.0 中包含了十個檔案，三萬八千七百二十五棵中文結構樹，含有二十三萬九千五百三十二個詞詞彙，每一句結構樹，標示漢語句法與語意訊息，詞類標記與斷詞標記系統四十五個標記不同，結構樹中的標記是由四十五個標記細分而成。在本節中我們利用中研院中文句結構樹測量詞類的相似度。

2-2-1 語意相似度測量

知網約選用了一千七百多個義原來定義中英雙語知識辭典中的每個詞，並且建有描述各個義原之間的關係的分類樹。例如：「讀書」一詞由「從事」、「學」與「教育」三個義原定義而成，知網中並有分類樹表示「從事」、「學」與「教育」三個義原之間的關係。

一般來說，一個詞在知網中可能擁有多個詞條，原因在於詞彙的多義性，因此我們在這邊定義兩個詞 $Word_1$,$Word_2$ 間的相似度相等於兩個詞各屬的詞條間最大相似度。

$$\mathrm{HowNetSimScore}(Word_1, Word_2) = \max \mathrm{HowNetSimScore}(Word_1 - Entry_x, Word_2 - Entry_y)$$

其次，每一個詞條可能由一到八個義原定義而成，如「讀書」一詞由「從

事」、「學」與「教育」三個義原定義而成，在知網標記義原的規則中，在詞條的所有定義義原中，第一個義原一定是主要意義分類，形成概念間的上下位關係(is-a relation)，第二個以後的義原為次要區分與詞彙之間的關係就不確定，依照知網標記決定。計算兩個詞條間相似度時主要義原與整個詞彙之間的關係十分重要，必須與其他的次要義原分開計算。因此

$$\begin{aligned}&\mathrm{HowNetSimScore}(\mathrm{Word}_1-\mathrm{Entry}_x,\mathrm{Word}_2-\mathrm{Entry}_y)\\&=w_1*\mathrm{PrimaryScore}(\mathrm{Sem}_{x,1}\cap\mathrm{Sem}_{y,1})\\&+w_2*\mathrm{SecondaryScore}((\mathrm{Sem}_{x,2}...\mathrm{Sem}_{x,n}),(\mathrm{Sem}_{y,2}...\mathrm{Sem}_{y,m}))\end{aligned}$$

知網中有描述義原與義原之間的階層關係的分類樹，我們在這邊利用描述義原之間關係的分類樹來幫助我們計算義原間的相似度。陳克健、陳超然(1997:270)認為兩個語意類的相似度在於兩個語意類在分類樹交集節點的語意訊息量(Information Content)，將整個詞分類架構看成一個訊息系統，一個語意類 Sem (相當於知網中的義原)的訊息量定義為 Entropy(System)-Entropy(Sem)。我們在這邊使用陳克健、陳超然(1997)計算語意訊息量的方法來計算知網中各義原的訊息量。

知網中兩個義原的相似度為這兩個義原所交集節點的語意訊息量，所得到語意訊息量越高表示這兩個義原越相似，因此第一部份的相似度定義如下：

$$\begin{aligned}&\mathrm{PrimaryScore}(\mathrm{Sem}_{x,1}\cap\mathrm{Sem}_{y,1})\\&=\mathrm{InformationContent}(\mathrm{Sem}_{x,1}\cap\mathrm{Sem}_{y,1})/\mathrm{Entropy}(\mathrm{System})\\&=(\mathrm{Entropy}(\mathrm{System})-\mathrm{Entropy}(\mathrm{Sem}_{x,1}\cap\mathrm{Sem}_{y,1}))/\mathrm{Entropy}(\mathrm{System})\end{aligned}$$

而第二部分的相似度的定義為：

$$\begin{aligned}&\mathrm{SecondaryScore}((\mathrm{Sem}_{x,2}...\mathrm{Sem}_{x,n}),(\mathrm{Sem}_{y,2}...\mathrm{Sem}_{y,m}))\\&=\left(\left(\sum_{i=2}^{n}\underset{j=\{1...m\}}{\mathrm{Max}}((\mathrm{InformationContent}(\mathrm{Sem}_{x,i}\cap\mathrm{Sem}_{y,j})/\mathrm{Entropy}(\mathrm{System}))/(n-1))\right)\right)\\&=\left(\left(\sum_{i=2}^{n}\underset{j=\{1...m\}}{\mathrm{Max}}((\mathrm{Entropy}(\mathrm{System})-\mathrm{Entropy}(\mathrm{Sem}_{x,i}\cap\mathrm{sem}_{y,j}))/\mathrm{Entropy}(\mathrm{System}))\right)/(n-1)\right)\end{aligned}$$

我們令(n>=m)，也就是第一個詞條的定義的義原多於或等於第二個詞條的義原，從第一個詞條中第二個義原開始，每個義原與第二個詞條中的每個義原計算相似度，第一個詞條中每個義原留下與第二個詞條義原相似分數最高的組合，將第一個詞條中每個義原得到的分數平均，就是我們所定義的第二部分的相似度。

以上兩式中各項皆除以 Entropy(System)是為維持相似值介於 0,1 之間。

2-2-2 詞類相似度測量

我們將 1.0 版中的句結構樹中歸納出規則，並統計每條規則出現的頻率，如圖 1 可歸納出右邊的三條規則，規則之前的數量表示規則出現的次數。下圖為中研院中文句結構樹的範例：

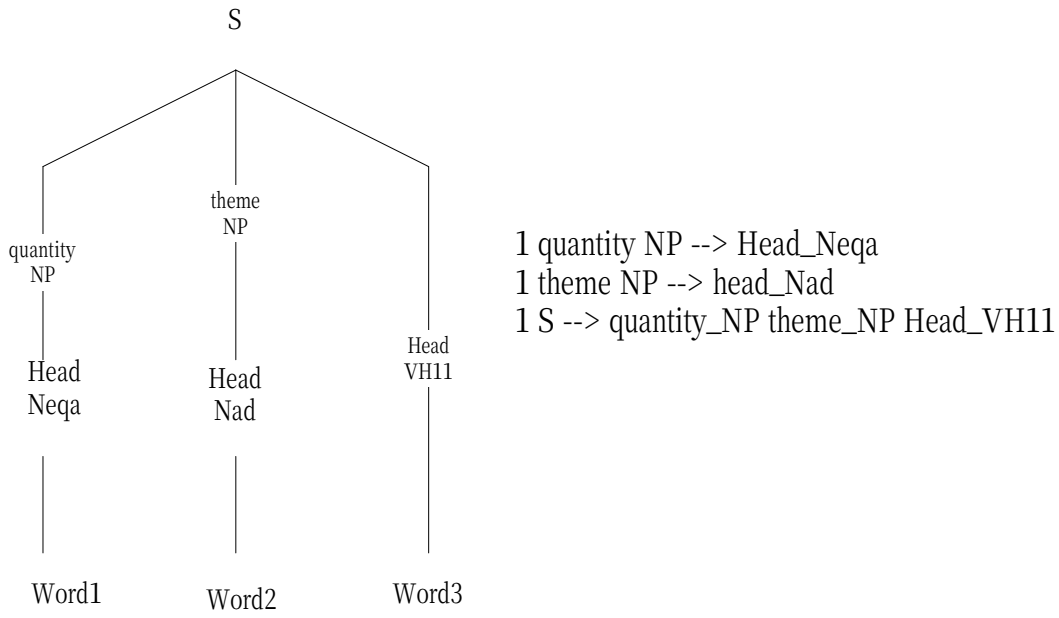

圖 1 中文句結構樹樹狀圖與歸納規則

每一個詞類的向量由各父節點與兄節點出現的頻率組成，先為放入各父節點的頻率，再依次放入兄節點的頻率，若該個節點沒出現在該詞類中，則放入為 0。定義如下：

i={VA, VAC, VB, VC, VCL,…Na, Nb….A,…P,…}

$\overrightarrow{\text{Category}_i}$ =< freq(parent node$_1$), freq(parent node$_2$),...freq(parent node$_n$), freq (sibling node$_1$), freq(sibling node$_2$),..., freq(sibling node$_m$) >

得到各個詞類的向量後，我們利用下列公式計算詞類與詞類之間的相似程度，所得的分數介於 0~1 之間，1 表示完全相同，0 表示完全不相同。

$$CategoryScore\left(\overrightarrow{Category_i},\overrightarrow{Category_j}\right)=\frac{\overrightarrow{Category_i}\bullet\overrightarrow{Category_j}}{\left|\overrightarrow{Category_i}\right|*\left|\overrightarrow{Category_j}\right|}$$

我們列出部分 VH 類的動詞與各類動詞的相似度於表格 1。除了 VH 類下的分類 VHC 類外，VH 類動詞與 VI 類相似程度最高，VH 類與 VI 類兩者皆為狀態動詞，他們的差別僅在於可接的論元數量。VI 類為類單賓動詞，基本上也是不及物動詞，但是 VI 類的動詞在語意上可接受一個論元，但該論元的位置不出現在動詞之後，通常使用一個介詞將論元引介出來。而 VH 類與 VA 類的相似程度為次高，VH 類與 VA 類同屬不及物動詞，他們的差別僅在於動作與狀態的區分。

表格 1　詞類相似度(部分)

詞類 1	詞類 2	相似度
VH	VA	0.674
VH	VC	0.611
VH	VD	0.643
VH	VE	0.540
VH	VG	0.591
VH	VH	1.000
VH	VI	**0.736**
VH	VJ	0.655
VH	VHC	0.852

2-3 相似詞的選取

在使用相似法來預測動詞分類的過程中，三個主要的步驟。一為未知動詞的相似詞的選取，二為測量未知動詞與相似詞的相似度，三為決定未知動詞的詞類。

首先，當一個新的未知動詞出現時，我們並不知道哪些訓練語料的動詞與新的未知動詞較相似，因此理論上我們必須計算每個訓練語料中的動詞與新的未知動詞的相似度，尋找出相似度較高的相似詞作為新的未知動詞預測詞類的依據，計算新的未知動詞與訓練語料中動詞的定義如下：

If Word= $wordbase_1$+$wordbase_2$+$wordbase_3$...+$wordbase_n$

$Sim(Word_{unknown},Word_{known})$

$=weight_1*Sim(wordbase_{1\text{-}unknown},wordbase_{1\text{-}known})$

$+weight_2*Sim(wordbase_{2\text{-}unknown},wordbase_{2\text{-}known})$

+...

$+weight_n*Sim(wordbase_{n\text{-}unknown},wordbase_{n\text{-}known})$

若採用這種方法必須計算訓練語料中的每一個詞彙與我們未知動詞的相似度，將會浪費許多不必要的計算時間，因此僅就訓練語料中與新的未知動詞前詞基相同與後詞基相同的相似詞為計算標的。尋找到前詞基相同與後詞基相同的相似詞後，第二步需計算這些選取出來的相似詞中與新的未知動詞詞基相異的部分的相似度。計算兩個詞彙相似度的方法，如下；

$$Sim(Word_{unknown},Word_{known})$$
$$=w_1*Score_1+w_2*Score_2$$
$$=w_1*HowNetSimScore(Base_i,Base_j)$$
$$+w_2*CategoryScore(category(Base_i),category(Base_j))$$

$Word_{known}$ 為相似詞
$Base_i$ 為未知動詞與相似詞相異的詞基
$Base_j$ 為相似詞與未知動詞相異的詞基

最後一個步驟是決定未知動詞的詞類。我們已有了一群相似詞，同時每個相似詞也有與未知動詞的相似分數。先將這些相似詞依照詞類分組，從每個詞類當中取出 K 個相似詞出來，將這些相似詞的分數予以平均，得到未知動詞到每個詞類的平均距離，未知動詞的詞類即與其距離最相近的詞類。我們將在下一節測試語意相似度中的比重、語意與詞類的比重以及 K 值的大小對正確率的影響。

3. 實驗結果

中研院平衡語料庫第三版五百萬詞內，未知詞經由人工標記為動詞者有 9170 個，佔 18.2%，名詞類有 40455 個，佔 80.3%，非謂形容詞有 761 個，佔 1.5%。本論文的處理範圍為這 9170 個未知動詞，即不存在辭典中的動詞。從中抽取出 1000 個動詞當作測試語料(Final Test Set)，其餘的未知動詞當作訓練語料(Training Set)，測試語料的正確答案為人工標記的詞類。

在相似法中需要討論下列三點。一、調整語意相似度中的主要義原與次要義原間的比重，二、調整語意與詞類兩種相似度的比重，三、調整 K 值的大小，使整個系統的正確率達到最佳狀態。

正確率的定義為：
正確率＝猜測正確的未知動詞/(1000-無法猜測的未知動詞)

3-1 語意比重相似度評量

我們首先要固定兩個變數，語意與詞類的比重與 K 值，才能觀察出相似度比重的變化對正確率的影響。因此先給予 K=1，語意與詞類比重為 1 與 0，依照相似度比重的變化對正確率的影響製成下表。

表格 2　相似度比重與正確率變化表

語意與詞類比重(語意,詞類)	語意相似度比重(w_1,w_2)	正確率
(1,0)	(1,0)	54.04%
(1,0)	(0.9,0.1)	57.58%
(1,0)	**(0.8,0.2)**	**57.70%**
(1,0)	(0.7,0.3)	57.58%
(1,0)	(0.6,0.4)	56.97%
(1,0)	(0.5,0.5)	56.85%
(1,0)	(0.4,0.6)	56.23%
(1,0)	(0.3,0.7)	55.87%
(1,0)	(0.2,0.8)	56.11%
(1,0)	(0.1,0.9)	56.11%
(1,0)	(0,1)	56.09%

由上表中可看出主要義原的比重為 0.8 與次要義原的比重為 0.2 時可以得到最高的正確率，因此在本實驗中我們使用 0.8 與 0.2 作為主要義原與次要義原的比重。

3-2 語意與詞類比重評量

我們將相似度比重設定 w_1 為 0.8 與 w_2 為 0.2，K=1，觀察語意與詞類比重的變化對正確率的影響。

表格 3 語意與詞類比重與正確率變化表

語意與詞類比重(語意,詞類)	語意相似度比重(w_1,w_2)	正確率
(1,0)	(0.8,0.2)	57.70%
(0.9,0.1)	(0.8,0.2)	58.41%
(0.8,0.2)	(0.8,0.2)	58.30%
(0.7,0.3)	(0.8,0.2)	58.85%
(0.6,0.4)	(0.8,0.2)	58.85%
(0.5,0.5)	(0.8,0.2)	59.40%
(0.4,0.6)	(0.8,0.2)	59.62%
(0.3,0.7)	(0.8,0.2)	60.40%

(0.2,0.8)	(0.8,0.2)	60.51%
(0.1,0.9)	**(0.8,0.2)**	**60.62%**
(0,1)	(0.8,0.2)	48.97%

經由上表的觀察，我們不能放棄詞類分數或語意分數，因為當詞類分數或語意分數比重為 0 時，所得到的正確率皆低。另外，我們發現約有一成的未知動詞無法猜測，原因在於沒有相似的例子或知網中沒有收錄該詞彙，造成沒有猜測結果的情形。

3-3 K 值變化

經由上述的實驗，我們現在將相似度比重設定為 0.8 與 0.2，而語意與詞類的比重設定為 0.9 與 0.1，來觀察 K 值的變化對正確率的影響。

表格 4　K-value 與正確率變化表

語意與詞類比重	語意相似度比重	K-value	正確率
(0.1,0.9)	(0.8,0.2)	1	60.62%
(0.1,0.9)	(0.8,0.2)	2	65.82%
(0.1,0.9)	(0.8,0.2)	3	66.70%
(0.1,0.9)	(0.8,0.2)	4	67.48%
(0.1,0.9)	(0.8,0.2)	5	67.15%
(0.1,0.9)	(0.8,0.2)	6	67.26%
(0.1,0.9)	(0.8,0.2)	7	67.59%
(0.1,0.9)	(0.8,0.2)	8	67.81%
(0.1,0.9)	(0.8,0.2)	9	68.14%
(0.1,0.9)	**(0.8,0.2)**	**10**	**68.25%**
(0.1,0.9)	(0.8,0.2)	20	68.14%
(0.1,0.9)	(0.8,0.2)	30	67.59%
(0.1,0.9)	(0.8,0.2)	40	67.48%
(0.1,0.9)	(0.8,0.2)	50	68.25%

在這邊我們發現，相似例子的多少對整體正確率的變化有影響，因此我們將字典中的動詞放入我們的訓練語料中，使得訓練語料增多，觀察正確率的變化。

從表格 5 中我們發現，當訓練語料為未知動詞加上字典中的動詞時，正確率

隨之增長。但在表格 5 與 4 的比較中發現，當 k=1 和 2 時，表格 5 的正確率比表格 4 的正確率低。我們認為可能的解釋原因在於字典中有一小部分的詞彙不具語意透明性，當我們僅尋找少數相似的詞彙時，容易造成誤判，使正確率降低造成的結果。

表格 5　K-value 與正確率變化表

語意與詞類比重	語意相似度比重	K-value	正確率
(0.1,0.9)	(0.8,0.2)	1	58.59%
(0.1,0.9)	(0.8,0.2)	2	65.12%
(0.1,0.9)	(0.8,0.2)	3	67.65%
(0.1,0.9)	(0.8,0.2)	4	67.97%
(0.1,0.9)	(0.8,0.2)	5	67.86%
(0.1,0.9)	(0.8,0.2)	6	68.49%
(0.1,0.9)	(0.8,0.2)	7	68.49%
(0.1,0.9)	(0.8,0.2)	8	68.49%
(0.1,0.9)	(0.8,0.2)	9	68.81%
(0.1,0.9)	(0.8,0.2)	10	68.91%
(0.1,0.9)	(0.8,0.2)	20	70.60%
(0.1,0.9)	(0.8,0.2)	30	70.50%
(0.1,0.9)	(0.8,0.2)	40	70.71%
(0.1,0.9)	**(0.8,0.2)**	**50**	**70.92%**
(0.1,0.9)	(0.8,0.2)	60	70.71%
(0.1,0.9)	(0.8,0.2)	70	70.60%
(0.1,0.9)	(0.8,0.2)	80	70.39%
(0.1,0.9)	(0.8,0.2)	90	70.18%
(0.1,0.9)	(0.8,0.2)	100	70.07%
(0.1,0.9)	(0.8,0.2)	200	68.81%

4. 結果分析

我們在本節分析猜測錯誤的未知動詞、相似法無法處理之未知動詞與我們蒐集語料的問題。

4-1 猜測錯誤之未知動詞分析

在我們實際觀察相似法猜測錯誤的例子當中，參見下表(詳列於在附錄中)，我們條列出來猜測錯誤的例子。從這些例子當中我們觀察幾個現象：一、大部分的猜測錯誤的例子為較罕見的詞語，如：言趣、黏結之類。使用者在使用這些詞語之時，若無語境實在也很難判斷出正確的分類。二、有部分的詞彙已經詞彙化了，如：高挑。該詞彙無法從詞彙的組成成分觀察出該詞彙的意義來。這些動詞的處理方式，不適用於以上所提出的方法。

表格 6　猜測結果

未知動詞	系統猜測詞類	正確詞類
言趣	VH	VA
捐掉	VC	VD
晨運	VC	VA
夷平	VH	VC
黏結	VC	VH
自屬	VC	VG
練唱	VC	VA

4-2 相似法無法處理之未知詞

我們以相似法在處理未知動詞之時，觀察到當訓練語料僅為未知動詞時，約有一成的未知動詞無法辨識。而當訓練語料的數量增大時，不能處理的動詞數量便降低了。

表格 7　無法處理之未知動詞數量變化

訓練語料	不能處理動詞的數量
未知動詞	96
未知動詞＋字典	34

我們可以從上表中發現，當我們的訓練語料增加時，的確可以縮減不能處理

的未知動詞。我們在下面分析這三十四個不能處理的未知動詞，並試著尋找解決無法預測詞類的問題。

表格 8　無法預測分類的未知動詞分類

無相似分數	潰腫起來、大挪移、一決勝負、直垂到、瀲灩、鬧雙胞,、大買單、下油鍋、昇進到、起酒疹、大收紅、上山下海、泫然淚下、游手好閑、匯寄到、歸併到、暗藏玄機、唉聲嘆氣、安天樂命、黯然無語、大飽耳福,、轉增資
無相似詞彙	蕞爾小邦、遊客如織、蝶躞、潸然淚下、叱吒、洞房花燭、吃飽喝足、上山下海、商調至、松蘿垂掛、喁喁情話、萬民歸心、克己復禮

我們可以從上表中大概歸納出未知動詞不能處理的情形。一、找到相似的詞彙，但無法計算出相似度。

如：「泫然淚下」尋找到兩個相似詞。

泫然欲滴,0,VH

泫然欲泣,0,VH

但是因為「淚下」與「欲滴」無法計算相似度，造成了無法判斷「泫然淚下」的詞類。二、完全無法找到相似詞彙，如：「蕞爾小邦」。不能處理的四字未知動詞大多為 VH 類的成語，因此我們猜測這些不能處理的四字動詞為 VH 類，三字動詞若為 V＋N 結構則為 VA 類，字尾為趨向補語的詞彙詞類與補語之前的動詞相同，剩餘的都猜測為 VH 類。

4-3 語料分析

我們討論三個語料問題。一、未知詞的定義與抽取未知詞的方法。二、中研院平衡語料庫中標記的一致性。三、中研院平衡語料庫標記的模糊地帶。

4-3-1 未知詞定義與抽取未知詞的方法

我們在這邊討論未知詞的定義與抽取出來的未知動詞所衍生的一些問題。首先，本文未知詞的定義為不存在於字典中的詞彙，並且假設未知詞應具有語意透明性，即我們可以從字面上得到該詞彙的語意，但是在我們所收集的未知動詞中，有一小部分並不屬於這種類型，例如：中的(一箭中的)、夯築、向邇、离去、遄飛、熏繞、絜靜、鴃彼等。

不具語意透明性詞彙的最佳的處理方式是將其增入辭典中，根據我們對語料

的觀察與分析，發現這類型詞彙出現的原因大多是作者引用到了非現代漢語的詞彙，或非常用詞與字，我們認為解決這部分詞彙最好的方法就是將這一類型的詞彙全部收錄字典中。

4-3-2 中研院平衡語料庫標記的一致性

在我們觀察訓練語料中，發現有標記不統一的現象，這讓我們很難將這一部份的語料歸納出任何的結論，例如：「$verb_i$+不了」這種結構，在 $verb_i$ 屬動作動詞的情況下，我們發現有部分的標記人員將「$verb_i$+不了」這種結構的動詞的分類標記成 $verb_i$ 的分類，即仍屬動作動詞，另外有部分的人則將「$verb_i$+不了」標記成為一個狀態動詞，論元結構分類不改變。如：「抵擋不了」標記為 VJ 類(狀態單賓動詞)，「阻擋不了」標記為 VC 類(動作單賓動詞)，但「抵擋」與「阻擋」在中研院詞庫小組詞知識辭典中的詞類皆屬 VC 類(動作單賓動詞)。

我們推測這樣的標記方法是部分標記人員認為「不了」會使整個動詞狀態化，但是不會改變整個動詞的論元結構，因此標記人員將這樣的組合給予狀態動詞，而另外一部分的認為加上「不了」後，並不會影響整個動詞的動作與狀態的分類，則給予該 $verb_i$ 原先的分類。

由於標記規則的不統一，我們無法從中歸納出任何規則，這些標記不一致的詞彙影響到我們使用相似詞判斷動詞分類的正確率。我們也認為這種類型的詞彙的確很難去決定分類，但希望有個統一的規則，可以將這種類型的詞彙給予一致性的標記。我們也希望藉由我們從這個角度的觀察與提出討論，爾後進一步修改中研院平衡語料庫中的詞類標記，使得語料庫標記更為一致。

4-3-3 中研院平衡語料庫標記模糊地帶

在我們觀察訓練語料時，發現有許多未知動詞具有兩種以上的分類，我們認為這一類的動詞出現的原因在於本身語意上的模糊，讓標記人員不易判斷該類動詞的所屬標記。下表 9 列出未知動詞具有兩種以上標記的詞彙，出現最多的為兼有動作與狀態的未知動詞，其次為及物性的模糊，如：及物與不及物的模糊與單賓與雙賓的模糊。

表格 9 標記模糊詞彙

動作與狀態模糊	出現次數	詞彙
兼有 VA 與 VH	43	一爭長短、三拖四拉、大放光明、大開眼界、小反彈、不忍卒說、天搖地動、斗轉星移、左等右等、交互為用、共霑法益、同床共枕、合作來合作去、在握、安享天年、行俠仗義、呼朋喝友、居無定所、忽前忽後、背井離鄉、家

		傳戶曉、站得住、笑吟吟、鬥鬧熱、剪徑伏擊、密鑼緊鼓、接應不暇、清火、連戰皆墨、速審速結、喧天動地、惡語相向、發人省思、亂停、滑漏過去、蓄勢待撲、蝕甚復圓、語冠全場、誤蹈法網、趕流行、撫今思昔、講道完、斷去、藏垢納汙
兼有 VC 與 VJ	17	出不了、交織出、伴隨有、抵免、抵抗不了、附設有、看不順眼、要不了、浪費掉、配置有、停放有、牽連到、脫離不了、著有、裝設有、過不了、應證
及物性模糊		
兼有 VA 與 VC	17	大吵、大降、生產出來、存下來、折回來、狂襲過來、延伸出來、拖出去、放飛、流傳下來、洗來洗去、捲來捲去、教改、淡出、聊開、散布開來、著墨、解解饞、蓋下來、聯合起來、隱去
兼有 VH 與 VJ	3	互敬、住不起、嚇慘、憨拙
兼有 VC 與 VD	13	付得起、保留給、致上、展現給、租到、退回給、配起來、推介給、移轉給、設計給、散播給、轉述給、轉移到

動詞分類的模糊性影響了正確率的高低，有些動詞可為動作也可為狀態，但是這裡所謂的正確答案並沒有把這類動詞所有可能的分類標記出來。在這種的情況下，若他們僅標記該動詞為動作動詞的一類，而我們僅猜測為狀態動詞的一類，使得正確答案與我們預測的分類不同。儘管我們預測了其中一項可能的分類，但仍須計算為猜測錯誤，使得我們的正確率降低。

例如：「站得住」一詞可以當 VA 類與 VH 類的動詞，但若標記者僅將「站得住」標記為 VH 類，而我們的系統卻猜測「站得住」為 VA 類，雖然我們的系統猜測正確，但是因為正確答案僅給予「站得住」一個可能的詞類，沒有標記出來第二個可能的詞類。這類標記模糊的動詞是我們系統無法提高正確率的原因之一。

這一類的詞彙大概的歸納為下列幾點。一、成語性的詞彙多具有 VA 類別與 VH 類別。二、特殊的後綴詞，如：「不了」與「有」。若該詞彙擁有這樣的後綴詞，就可能具有狀態與動作的身份。三、部分加了「趨向補語」的動詞，也可能具有不及物與及物兩種特性。四、加上了「給」為後綴的動詞，也具有單賓與雙賓的特性。

我們將來希望將這一部分的詞彙歸納出規則，若屬於這類型的動詞，我們就可以自動給予一種以上的分類。

5. 結論

我們本文中利用相似法來判斷動詞的分類，尋找未知動詞的相似詞，計算未知動詞與相似詞之間的相似度，再將這些相似詞依照詞類分組。從每個詞類當中取出 K 個相似詞出來，將這些相似詞的分數予以平均，得到未知動詞到每個詞類的平均距離，未知動詞的詞類即與其距離最相近的詞類，正確率為 70.92%。

分析猜測錯誤的未知動詞中，大部分的詞彙為比較罕見的用詞或是已經詞彙化的詞語，我們建議將這一部分的未知詞收錄於字典中。其次，將部分無法預測分類的未知詞收錄辭典中，如成語『蕞爾小邦』。另外，我們也期待當訓練語料增多與知網收錄詞彙增多時, 可以處理另外一部份目前無法得到相似分數或無法尋找到相似詞的未知詞。

相似法容易受到語料中錯誤訊息的干擾，因此中研院平衡語料庫中標記的不一致性與部分詞彙本身的模糊性都影響到我們未知動詞自動分類的正確率。我們希望中研院平衡語料庫中標記不一致的語料與標記模糊語料的處理方式能夠得到改善，也期待改善後的結果能夠影響我們動詞分類系統的效能。

6. 參考文獻

中文

中央研究院詞知識庫小組。«技術報告 9305：中文詞類分析»。南港：中央研究院詞知識庫小組，1993。

---。«技術報告 9601：『搜』文解字---中文詞界研究與資訊用分詞標準»。南港：中央研究院詞知識庫小組，1996。

---。«技術報告 9502/9804：中央研究院平衡語料庫的內容與說明»。修訂版。南港：中央研究院詞知識庫小組，1998。

白明弘、陳超然、陳克健。<以語境判定中文未知詞詞類的方法>，«第十一屆計算機語言學會論文集»，1998，頁 47-60。

李振昌。«中文文本專有名詞辨識問題之研究»。台北：台灣大學資訊工程研究所碩士論文，1993。

李振昌、李御璽、陳信希。«中文文本人名辨識問題之研究»，<第七屆計算機語言會會議論文集>，1994，頁 203-222。

李坤霖。«網際網路 FAQ 檢索中意圖萃取及語意比對之研究»。台南：成功大學資訊工程研究所碩士論文，2000。

陳克健、洪偉美。<中文裡「動名」述賓結構與「動名」偏正結構的分析>，«第八屆計算機語言學會論文集»，1996，頁 1-29。

陳克健、陳超然。<語料庫為本的中文複合詞構詞律模型研究>，«漢語計量與計算研究»，編輯：鄒嘉彥、黎邦洋、陳偉光、王士元，1997，頁 283-305。

梅家駒、竺一鳴、高蘊琦、殷鴻翔。«同義詞詞林»。香港：商務印書館，1984。

湯廷池。«漢語詞法句法論文集»。台北：學生書局，1988。
董振東、董強。知網---中文信息結構庫。<http://www.keenage.com>，2000。
---。事件關係與角色轉換庫。<http:// www. keenage.com >，2000。
趙元任。«中國話文法»。丁邦新譯。香港：中文大學，1980。
賴育昇、李坤霖、吳宗憲。«網際網路 FAQ 檢索中意圖萃取及語意比對之研究»。<第十三屆計算機語言學研討會>，2000，頁 135-156。

西文

Chen, Chao-Jan, Ming-Hung Bai and Keh-Jiann Chen. “Category Guessing for Chinese Unknown Words,” Proceedings of the Natural Language Processing Pacific Rim Symposium, 1997, pp. 35-40.

Chen, Keh-Jiann and Ming-Hong Bai. “Unknown Word Detection for Chinese by a Corpus-based Learning Method,” Computational Linguistics and Chinese Language Processing vol3 no. 1, 1998, pp. 27-44.

---. “Knowledge Extraction for Identification of Chinese Organization Names,” Proceedings of the second Chinese Language Processing Workshop, 2000, pp. 15-21.

Li, Charles and Sandra Thompson. “Mandarin Chinese: A Functional Reference Grammar”. Berkeley: University of California Press, 1981.

Resnik, Philip. “Using Information Content to Evaluate Semantic Similarity in a Taxonomy,” Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995, pp. 448-453.

---. “Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language,” Journal of Artificial Intelligence Research XI, 1998, pp. 95-130.

Resnik, Philip and Mona Diab. Measuring Verbal Similarity. Technical Report: LAMP-TR-047//UMIACS-TR-2000-40/CS-TR-4149/MDA-9049-6C-1250. University of Maryland, College Park, 2000.

Sproat Richard and Shilin Shih. “A Corpus-Based Analysis of Mandarin Nominal Root Compound,” Journal of East Asian Linguistics 5, 1996, pp. 49-71.

Weischedel, Ralph, Marie Meteer, Richard Schwartz, Lance Ramshaw and Jeff Palmucci. “Coping with Ambiguity and Unknown Words through Probalistic Model,” Computational Linguistics 19, 1993, pp. 359-382.

附錄

未知動詞	系統猜測詞類	正確詞類
迎來	VA	VC
言趣	VH	VA
捐掉	VC	VD
晨運	VC	VA
夷平	VH	VC
黏結	VC	VH
自屬	VC	VG
練唱	VC	VA
傾盡	VC	VH
夾處	VC	VJ
車拼	VA	VC
迷向	VC	VA
齊薰	VH	VC
走穩	VH	VA
攏好	VH	VC
耐燃	VC	VH
嗤鼻	VH	VA
申領	VH	VC
正價	VA	VH
堵回	VC	VCL
承標	VA	VC
高挑	VC	VH
漂忽	VJ	VA
大登	VH	VC
抽背	VA	VC
交詮	VE	VC
報來	VH	VC
吹響	VH	VC
中第	VH	VA
悠到	VCL	VC
轉劇	VA	VH
燒昏	VC	VHC
人治	VC	VH
自定	VH	VC
重登	VC	VCL
轉走	VC	VA
折拗	VC	VH
實徵	VC	VH
提味	VH	VA
湊熱	VH	VA
偵錯	VC	VA
曝身	VA	VH
避靜	VH	VA
擁至	VC	VCL
步踵	VCL	VC
自食	VA	VC
藏諸	VC	VCL
孝親	VA	VH
對招	VC	VA
營築	VH	VC
試走	VC	VA
給就	VC	VD
易記	VC	VH
無誨	VC	VH
寶肝	VH	VA
起腳	VH	VA
哭窮	VH	VA
時起	VC	VH
無念	VJ	VH
回傳	VD	VC
啞口	VA	VH
湧退	VC	VA
鼓足	VH	VC
互映	VA	VH
自野	VA	VH
正對	VH	VC
不輟	VH	VA
承續	VH	VC
板結	VC	VH
對中	VC	VJ
獨統	VA	VH
攘外	VH	VA
租住	VC	VCL
教懂	VH	VC
沾黏	VH	VC
抗震	VH	VA
並稱	VE	VG
削聳	VC	VH
運賣	VC	VD
增殖	VC	VA
惡用	VC	VJ
作美	VH	VA
明化	VHC	VH
撤到	VC	VCL
畫方	VH	VA
採到	VCL	VC
怒張	VH	VA
虛矯	VH	VC
為亂	VH	VA
酌進	VH	VC
轉遊	VA	VCL
移葬	VCL	VC
撲上	VC	VA
倒盡	VH	VJ
近身	VH	VA
自云	VA	VE
叮穩	VC	VA
廣作	VG	VC
唸作	VG	VC
雙殺	VC	VA
解壓	VC	VA
加烈	VC	VH
從眾	VA	VH
靜棲	VA	VCL
哭倒	VC	VA
收擔	VC	VA
悅動	VC	VH
跌斷	VC	VHC
互連	VH	VJ
觀照	VA	VC
拒容	VJ	VA
超產	VA	VC
大誇	VH	VC
盲動	VA	VH
整錯	VH	VC
共學	VC	VA
澆到	VC	VCL
搞懂	VH	VE
裝得了	VJ	VC
破記錄	VH	VA
斷不了	VH	VC
碰不得	VH	VC
管起來	VC	VE
定出來	VA	VC
問清楚	VC	VH
遞上來	VC	VA
妨害到	VJ	VC
運下來	VA	VC
大震撼	VC	VJ
侵占去	VA	VC
撿回來	VA	VC
沒法度	VC	VH
往回收	VC	VA
伸進來	VA	VC
哄上去	VA	VC
傾銷至	VC	VCL
抬回來	VA	VC
切進來	VA	VC
拗脾氣	VA	VH
合八字	VH	VA
行的通	VC	VH
巡禮完	VC	VCL

播下去	VA	VC
暗下來	VA	VH
當不了	VC	VG
惹不得	VH	VJ
漲退潮	VH	VA
跑進來	VA	VCL
走著瞧	VC	VA
自宣布	VA	VE
談戀愛	VH	VA
喜愛上	VC	VJ
大辯論	VC	VE
窩藏進	VC	VCL
穿進去	VA	VC
減輕到	VC	VH
溜進去	VA	VCL
搬下去	VA	VC
大販賣	VC	VD
實習用	VC	VH
激增到	VC	VJ
起不了	VH	VJ
減少到	VC	VH
大肚子	VA	VH
縮小到	VC	VH
經得住	VC	VJ
試開工	VC	VH
容不進	VC	VJ
拖出來	VA	VC
傳遞至	VCL	VC
量出來	VA	VC
久旱未雨	VH	VA
學有專攻	VC	VH
驚醒過來	VA	VHC
得意滿面	VA	VH
金盆洗手	VH	VA
易朝換主	VA	VH
照射下來	VH	VA
春光外泄	VC	VH

跌足搥嘆	VE	VH
蹈光養晦	VH	VA
展現出來	VA	VC
天人交戰	VH	VA
提報上來	VA	VC
言者諄諄	VA	VH
睽隔已久	VH	VA
糾結起來	VA	VH
隱善揚惡	VH	VA
生聚教養	VH	VA
增添進來	VA	VC
聽不進去	VC	VA
學無止境	VC	VH
塵灰撲撲	VC	VH
消受得起	VC	VJ
流於言表	VA	VH
大傷腦筋	VA	VH
表露出來	VA	VC
先馳得點	VH	VA
昏迷過去	VA	VH
應付得了	VJ	VC
大顯神威	VH	VA
搜竭枯腸	VH	VA
共體國艱	VH	VA
發監執行	VC	VA
拋頭露臉	VC	VH
乘興而去	VA	VCL
曬曬太陽	VC	VA
回復過來	VA	VH
傲霜鬥雪	VA	VH
以策萬全	VA	VH
雕梁畫棟	VA	VH
大行其道	VE	VH
持之有故	VH	VA
視而未見	VC	VA
竄高伏低	VH	VA
弄虛作假	VC	VA

堆積下來	VA	VC
決定不了	VC	VE
放縱出來	VC	VA
鴨子聽雷	VA	VH
經受不起	VC	VJ
藉題發揮	VC	VA
傳染開來	VC	VH
遙不可及	VJ	VH
滲濕開來	VC	VH
妙筆生花	VA	VH
期待出來	VC	VH
隨而俱之	VA	VH
碰觸不得	VH	VC
知恩圖報	VC	VH
生存下去	VA	VH
未述其詳	VH	VA
據實以告	VE	VA
另當別論	VA	VH
勞而不獲	VC	VH
再蹈覆轍	VH	VA