Computational Linguistics and Chinese Language Processing
Vol. 7, No. 1, February 2002, pp. 1-28

以構詞律與相似法爲本的中文動詞自動分類研究

A Hybrid Approach for Automatic Classification of Chinese Unknown Verbs

曾慧馨*、劉昭麟+、高照明**、陳克健*

摘　要

Abstract

本論文合併兩種方法預測未知動詞的詞類。第一種方法爲規則法，即從訓練語料中歸納出未知動詞組成的構詞規律，分成兩個主要的判斷方式：一、依照未知動詞的組成的關鍵字決定其分類。二、依照未知動詞的構成組合決定其分類。

關鍵字法首先將動詞依長度分爲四組。第一組爲二字詞、三字詞、四字詞、五字以上的詞彙。在對實際語料的觀察下，發現不同詞長的動詞結構相異，因此將語料依詞長分組。例如：三字詞可訓練出「好」、「出」兩條規則決定動詞的詞類，其他長度的未知動詞並沒有這兩條規則，另外「化」規則不適用於二字動詞。

規則法的第二部分爲依照構成組合決定其分類。在觀察未知動詞時，發現有部分未知動詞的組合很具有規律，我們就將訓練語料中未知動詞的組合做個歸納，得到九種組合。在十次實驗中，規則法可以處理的未知動詞平均約爲 23.19%，猜測正確的比例爲 91.67%。

二、相似法爲利用與未知動詞相似的例子來預測未知動詞的詞類。相似法主要利用知網與中央研究院中文句結構樹資料庫 1.0 作爲語意與詞類相似度測量的工具。藉由計算未知動詞與已知動詞的相似度來預測未知動詞的詞類，未知動詞的詞類爲與其相似度最高的相似例子的詞類。

* 中央研究院資訊所，曾慧馨 E-mail: huihsin@iis.sinica.edu.tw
陳克健 E-mail: kchen@iis.sinica.edu.tw
+ 政治大學資訊系 E-mail: chaolin@ nccu.edu.tw
** 台灣大學外文系 E-mail: zmgao@ccms.ntu.edu.tw

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。當兩個辭彙相似度高時，表示這兩個辭彙的詞類、語意類與結構必定相似。在十次實驗中，使用相似法預測動詞的正確率約爲 71.05%。

規則法的優點在於判斷正確率高，缺點爲可處理的未知動詞數量有限；相似法的優點爲可以處理大部分的未知動詞，但正確率不如規則法高。最後，我們結合這兩種處理方法來預測未知動詞的分類，將兩個方法同時應用在最後的測試語料中，規則法的正確率爲 87.25%，而相似法的正確率爲 65.04%，兩著者結合後的正確率爲 70.80%。

In this paper we present a hybrid approach for automatic classification of Chinese unknown verbs. The first method of the hybrid approach utilizes a set of morphological rules summarized from the training data, i.e. the set of compound verbs extracted from Sinica corpus, to determine the category of an unknown compound verb. If the morphological rules are not applicable, then the instance-based categorization using the *k*-nearest neighbor method for the classification is employed. It was observed that some suffix morphemes are frequently occurred in compound verbs and also uniquely determine the syntactic categories of the resultant compound verbs. By processing and calculating the training data, 15 suffix rules with coverage over 2% and category prediction accuracy higher than 80% were derived. In addition to the above type of morphological rules, the reduplication rules are also useful for category prediction, such as some famous Chinese reduplication rules, like “aa” in two characters word, “aab”, “abb” and “aab” in three characters word etc. For instance,“喝喝茶”has the same category as “喝茶,” and “研究研究” has the same category as“研究.” As a result, nine reduplication patterns are generated. Experimenting on the training data, it is found that the overall accuracy of the morphological rule classifier is 91.67% and its coverage is 23.19% only.

Since the coverage of the morphological rule classifier is low, an instance-based categorization method is employed to taking care the uncovered cases. The instance-based categorization utilizes similar examples to predict the category of an unknown verb. The lexical similarity was measured by both the semantic similarity and syntactic similarity. The semantic similarity between two words is measured by the semantic distance of their HowNet definitions and the syntactic similarity is measured by the distance of their syntactic categories. The distance between two syntactic categories is their cosine measure of their grammatical feature vectors derived from the Sinica Treebank. The category of an unknown

verb is predicted as the same as the examples, which are most similar to the unknown verb according to the above criteria of the similarity. For testing on the training data, the optimal accuracy of instance-based categorization is 71.05%, when the similar examples are from unknown verbs and verbs in the dictionary (known verbs).

Both the morphological rule classifier and the instance-based categorization have the advantages of not only predicting the syntactic categories of the unknown words but also recognizing their morphological structures and major semantic classes. The advantage of the morphological rule classifier is its higher accuracy and for the instance-based categorization is its higher coverage. However, both of the methods have their own drawback; the former cannot be applied to most unknown verbs, but the latter suffers from low accuracy. For open test, 1000 unknown verbs that are unseen in the training process were tested. The accuracy of the linguistic rule is 87.25%, and the instance-based categorization is 65.04%. Finally, the overall accuracy of the hybrid approach is 70.80%.

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爲未知詞的存在。現行的斷詞標記系統以辭典爲基礎輔以構詞的規則訊息進行斷詞標記，但因爲語言的特性之一「無窮盡的創造力」，無法窮舉出所有的辭彙；一本好的辭典也不應該無止盡的擴大所收錄的辭彙，因此如何辨識處理辭典中不存在的辭彙就成了一個重要的課題。本論文的目標即希望處理不存在辭典中的未知動詞。

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 等等]。僅有 Chen、Bai 與 Chen [1997] 利用字首(prefix)、字尾(suffix)的訊息處理全部的未知詞，正確率約爲 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%。在動詞分類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分類處理上，並且希望在未來將這種處理未知動詞的方法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，若動詞爲未知詞，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，我們認爲動詞自動分類研究至今無法提高正確率的重要原因爲動詞繁複的內部結構。

我們的目標爲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，封閉性詞類爲該分類中的辭彙不會增加，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 類分類中的動詞辭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 9 類爲動作不及物動詞(VA[1])、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爲不存在辭典中的辭彙。陳克健、陳超然 [1997] 分析未知詞的種類爲兩種，第一種爲封閉性，這一類型雖然在數量上可能爲無數個，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(數位)、二七八八三七九九(電話)等。第二類則爲開放性，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爲略語、專有名詞、衍生詞、複合詞與數字型複合詞。

未知動詞通常爲複合詞，由兩個以上的組成成分組合而成，我們稱這些組成成分爲詞基 (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爲動賓結構。雖然詞基是有限的，但是詞基與詞基的組合數量龐大，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規則法與相似法來判斷動詞的分類，規則法利用特定的關鍵詞與詞基的組成方式來預測未知動詞的分類。相似法則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。

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。未知動詞爲複合詞，通常由數個具有孳生性的詞基所組成，本身具有高透明性。

---

[1] 參見附錄一表格 12.中研院詞庫小組詞類標記[1993]。

[2] Sproat 與 Shih [1996] 稱內部的處理單位爲詞根(root)，Chen、Bai 與 Chen [1997] 稱處理的單位爲字首(prefix)與字尾(suffix)。我們則稱處理單位爲詞基(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爲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。

例如，未知動詞「求新」與「講錯」相對於列入辭典中的「忐忑」、「局促」這一類的辭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，我們認爲有三個因素可預測未知動詞的分類。一、語意。語意相近的辭彙，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，即一個語意類中的辭彙約有的詞類數量。因此我們認爲語意因素和辭彙的詞類有密切關聯。二、結構。結構通常會限定組成的詞類，若結構爲“VC+Na”的未知動詞，通常會組成 VA 詞類，因爲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，因此我們認爲結構會影響到動詞的詞類。三、關鍵詞。有些關鍵字可以直接的決定整個動詞的分類，如：若未知動詞的最後一個詞基爲「化」，該未知動詞即爲 VHC 類。軍國「化」爲 VHC 類。

在本篇論文中我們利用上述所提出的線索預測未知動詞所屬的詞類。

2. 方法與語料

我們提出兩種判斷動詞所屬分類的方法：規則法與相似法。規則法可分成兩個主要的判斷方式。一、依照未知動詞的組成的關鍵字決定其分類，如：軍國「化」爲 VHC 類。二、依照未知動詞的構成組合決定其分類，「aabb」的辭彙爲狀態類，如：平平靜靜；「abab」組合的辭彙爲動作類，如：準備準備。而相似法的處理方法爲尋找與未知動詞的相似詞，計算未知動詞與其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。

規則法的優點在於判斷正確率高，缺點爲可處理的未知動詞數量有限；相似法的優點爲可以處理大部分的未知動詞，但正確率不如規則法高。在本節中，我們首先介紹規則法，接著介紹相似法，最後結合這兩個處理方法來預測未知動詞的分類。

我們從中研院平衡語料庫中抽取出 10443 個不存在辭典中的未知動詞，保留 1000 個未知動詞爲最後評估本系統的測試語料。在規則法的實驗中，不斷從 9443 個訓練語料中重複的取出 1000 個未知動詞評估規則法的正確率與包含率。而在相似法調整比重的實驗中，爲了評估使用知網計算語意相似度，因此將擁有不存在於知網的詞基的詞彙刪除，剩 7535 個。剩餘的 7535 個未知動詞作爲相似法的訓練與建構中的測試語料，用來調整知網義原、詞類、語意所佔的比重。

3. 規則法

規則法可分成兩個主要的判斷方式。一、依照未知動詞的組成的關鍵字決定其分類。二、依照未知動詞的構成組合決定其分類。在本節中，我們將介紹我們如何從語料中尋找出這些規則。

3.1 規則訓練與觀察

規則法由兩個判斷方法組成，可依照未知動詞的組成的關鍵字決定其分類與依照構成組合決定其分類。首先，我們討論依照未知動詞的組成的關鍵字決定其分類的判斷方法。

我們將訓練語料中的動詞依照字數不同先分爲四組，參見表格 1。第一組爲二字詞、第二組爲三字詞、第三組爲四字詞、第四組爲五字以上的詞彙。依照字數不同分成四組的原因在於希望依照字數長短的不同訓練出規律。如：三字詞的「好」、「出」兩條規則，其他長度的未知動詞並沒有這兩條規則，「化」規則不適用於二字動詞等等。我們將語料分成四組，希望依照語料的特性做成歸納。

我們將二字詞的第一個詞基與第二個詞基當作我們歸納規則的訊息，想要從訓練語料中觀察當第一個詞基或第二個詞基出現時，爲某個詞類的正確率 (Accuracy) 有多少與這條規則對該詞類的包含率 (Coverage)[3] 有多大。以第一詞基爲關鍵詞可以訓練出來三千四百多條規則，以第二詞基爲關鍵詞訓練出來三千三百多條規則，但是僅有四條規則包含率大於 2%，正確率高於 80%。

我們設定正確率必須高於 80%與包含率必須大於 2%是找出預測詞類準確性較高的規則，因此將正確率訂高，維持規則的水平。限定包含率必須大於 2%的原因，我們不希望找出僅可以處理一個例子的規則，縱使這條規則的正確率爲 100%，這條規則也不具代表性。最後我們就設定了包含率爲 2%。在二字動詞的規則訓練部分我們就保留這 4 條規則做爲我們動詞自動分類中的規則。

第二組爲三字動詞。將經過內部斷詞的三字詞的第一個詞基與最後一個詞基拿來當我們歸納規則的訊息，每一個不同的詞基對我們來說都是一條規則，觀察該詞基出現時，對詞類辨識的正確率與對該詞類的包含率。在三字動詞最後一個詞基中，我們歸納出 1454 條規則，爲了控制規則的品質，我們訂立每條規則的包含率必須超過 2%，正確率必定高過 80%的限制，經過篩選後僅剩 9 條規則。

第三組爲四字詞。我們訓練出來的四字詞的規則僅有一條規則符合包含率高於 2%，正確率高於 80%的限制。第四組爲五字詞以上，訓練出來的五字詞以上之規則也僅有一條規則符合我們對包含率高於 2%，正確率高於 80%的限制。最後，我們使用表格 1 中所列出的關鍵字做爲規則法第一部份的規則。

---

[3] 計算包含率與正確率的公式定義如下：

包含率(後詞基爲詞類 i,詞類 i)=Freq(後詞基爲詞類 i)/Freq(詞類 i)

正確率(後詞基爲詞類 i,後詞基)=Freq(後詞基爲詞類 i)/Freq(後詞基)

i={VA、VC、VCL、VD、VE、VG、VH、VHC、VJ}

本文包含率與正確率的定義每個小節不盡相同,請參照該節的定義。

表格 1. *規則法之關鍵字*

字串長度	最後一個詞基	詞類	正確率	包含率
二字詞	給	VD	91.0%	19.7%
二字詞	予	VD	83.3%	4.8%
二字詞	為	VG	88.7%	22.0%
二字詞	成	VG	84.3%	36.5%
三字詞	好	VC	83.6%	5.9%
三字詞	出	VC	81.5%	22.7%
三字詞	給	VD	92.0%	82.4%
三字詞	予	VD	100%	2.4%
三字詞	為	VG	98.6%	52.1%
三字詞	成	VG	100%	40.0%
三字詞	化	VHC	95.3%	94.4%
三字詞	有	VJ	90.0%	13.5%
三字詞	於	VJ	82.2%	20.3%
四字詞	化	VHC	88.9%	76.1%
五字詞以上	化	VHC	100%	100%

規則法的第二部分爲依照構成組合決定其分類。在觀察未知動詞時，發現有部分未知動詞的組合很具有規律。表格 2 是我們觀察訓練語料後做出的歸納，以 a 與 b 代表未知動詞的內部詞基。在二字詞部分，有重疊的辭彙（aa），我們觀察到二字詞重疊的辭彙的詞類與單字詞相同，即 aa 未知動詞的詞類爲 a 的詞類，但是也有少數例外，如：亦亦、定定等。

三字詞的構成組合有 abb、aaa、aab 與 aba 四種，在這四種情況下，我們觀察到若 ab 原來就是一個辭彙，這該未知動詞的詞類與 ab 的詞類相同。若 ab 不存在於字典中，結構爲 aab 的話，猜測該未知詞爲 VA 類。aab 結構的未知動詞有部分多爲動賓結構，動賓結構的動詞部分，可以使用這種重複模式組成三字的動詞。而 abb 結構的未知動詞，若 ab 不存在於字典當中，傾向猜測該未知詞 VH 類。aba 結構的未知動詞，則猜測整個未知動詞的詞類與 a 的詞類相同。aaa 結構的未知動詞，猜測未知動詞的詞類與 a 詞基相同。

四字動詞的構成組合有 aabb、abab、a*X*a*Y* 與 *X*b*Y*b。aabb 的未知動詞分類與 ab 相同，若 ab 不存在辭彙的話，猜測爲 VH 類，abab 的未知動詞分類與 ab 相同，若 ab 不存在時，猜測爲 VA 類。

四字動詞大部分爲不及物，依據我們對語料的觀察發現 aabb 的大部分爲狀態類，abab 的結構大部分爲動作類，因此當 ab 不爲一辭彙，沒有詞類時，則猜測 abab 爲 VA 類。a*X*a*Y* 的結構中就算 a 原爲及物動詞，但是在這種結構下 a*X*a*Y* 就組成不及物動詞，而不及物動詞的動作或狀態特性則依照 *X* 與 *Y* 這兩個關鍵字來決定未知動詞應該狀態類或動作類的不及物動詞。*X*b*Y*b 的詞類與 b 相同，因爲這裏的 *X* 與 *Y* 多爲修飾語，修飾主要動詞 b。

表格 2. *規則法之組合規律*

構成組合	例子
aa	收收、伸伸、作作、改改、見見、念念、拌拌、玩玩、急急
abb	大喇喇、心浮浮、白晃晃、白森森、凶巴巴、死沈沈、死翹翹、灰撲撲、血油油、血糊糊、冷森森、冷酸酸
aba	好不好、快不快、拔一拔、肩並肩、夠不夠、硬不硬、遠不遠
aab	小小聲、按按摩、洗洗腳、洗洗澡、洗洗臉、吵吵架、抓抓癢、喝喝茶、泡泡水
aaa	對對對、嘻嘻嘻、羞羞羞
aabb	久久長長、大大方方、工工整整、分分合合、切切割割、切切實實、反反覆覆、太太平平、心心念念、扎扎實實、文文靜靜、方方正正、片片斷斷、仔仔細細、出出入入
abab	呼吸呼吸、奔走奔走、拉扯拉扯、欣賞欣賞、爭取爭取、拜會拜會、指點指點、研究研究、料理料理、消化消化、討論討論、參考參考、參觀參觀、商量商量、排解排解、教訓教訓、規劃規劃、勞動勞動、湊合湊合、測驗測驗
a*X*a*Y*	吃來吃去、忙進忙出、串來串去、吵來吵去、扭來扭去、找來找去、扯來扯去、改來改去、拉來拉去、拋來拋去、爬上爬下、爭來爭去、玩來玩去、直來直往、挖來挖去、挑來挑去、看來看去、穿來穿去、讓來讓去、鑽來鑽去、踢來踢去
*X*b*Y*b	左等右等、好說歹說、大紅特紅、左想右想、左謝右謝、東跑西跑、一看再看

3.2 規則法評量

我們從中研院平衡語料庫中抽取出的 9443 個訓練語料中反覆十次抽取 1000 個未知動詞作爲規則法的測試語料，可以處理的未知動詞約爲 23.19%，猜測正確的比例爲 91.67%。

正確率=猜測正確的動詞數量/可以使用規則處理的語料

包含率=可以使用規則處理的語料/全部的測試語料

表格 3. *規則法實驗評估*

測試	正確率	包含率
1	89.52%	24.80%
2	91.38%	23.20%
3	93.61%	21.90%
4	91.42%	23.30%
5	89.91%	22.80%
6	92.12%	24.10%
7	89.75%	24.40%
8	93.83%	22.70%
9	91.98%	21.20%
10	93.19%	23.50%
平均	91.67%	23.19%

4. 相似法

這節我們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爲組成成分屬於常用詞且語意明確，例如：試印、講完。這兩個辭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」等等各樣的辭彙。

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，但因爲語言的複雜度，無法將所有的規則條列出來。因此我們在這邊使用相似法，將訓練語料中的每個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」爲我們的未知動詞。未知動詞的第二個組成成分與訓練語料中的例子相同都爲「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。當兩個辭彙相似度高時，表示這兩個辭彙的詞類、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。

4.2 相似度測量

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

一、知網爲一雙語(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、知網管理程式等。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，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。

4.3 語意相似度測量

知網約選用了一千五百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，原因在於辭彙的多義性，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。

$$\mathrm{HowNetSimScore}(\mathrm{Word}_1, \mathrm{Word}_2) = \max_{x,y} \ \mathrm{EntrySimScore}(\mathrm{Word}_1\mathrm{Entry}_x, \mathrm{Word}_2\mathrm{Entry}_y)$$

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)，第二個以後的義原爲次要區分，與辭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辭彙之間的關係十分重要，必須與其它的次要義原分開計算。因此

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

知網中有描述義原與義原之間的階層關係的分類樹，我們利用這個描述義原關係的分類樹來幫助我們計算義原間的相似度。陳克健及陳超然 [1997:270] 認爲兩個語意類的

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 (System) 看成一個訊息系統，一個語意類 Sem (相當於知網中的義原) 的訊息量定義爲 Entropy(System)-Entropy(Sem)。我們在這邊使用他們的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爲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：

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

而第二部分的相似度的定義爲：

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

我們可以假設 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 $\text{Entropy}(\text{System})$ 是爲維持相似值介於 0 和 1 之間。

4.4 詞類相似度測量

我們將中研院中文句結構樹 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 的句結構樹可歸納出右邊的三條規則。從句結構樹中我們可以觀察到“quantity NP”唯一個父節點 (parent node) 可以衍生“Head_Neqa”這個子節點，我們就將這個句結構樹中的關係改寫爲規則，寫成爲 quantity NP→ Head_Neqa，並且統計每條規則出現的次數，將句結構樹中所有出現同樣規則做數量累計，作爲我們計算詞類相似度的變數。

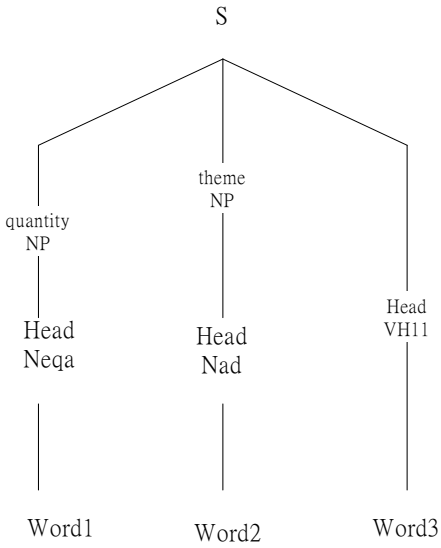

圖 1 中文句結構樹樹狀圖與歸納規則

每一個詞類(Category)的所定義的向量由各父節點與兄節點出現的頻率組成，該向量的組成成分的數量與排序是固定的，若該詞類的向量其中一個組成成分沒有出現過，則其值爲零，向量的組成成分爲先是各父節點的頻率，其次爲各兄節點的頻率。定義如下：

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

$$\overrightarrow{Category_i} = < freq(parent\ node_1), freq(parent\ node_2), ... freq(parent\ node_n),$$
$$freq(sibling\ node_1), freq(sibling\ node_2), ..., freq(sibling\ node_m) >$$

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。

$$CategorySore\left(\overrightarrow{Category_x}, \overrightarrow{Category_y}\right) = \frac{\overrightarrow{Category_x} \bullet \overrightarrow{Category_y}}{\left|\overrightarrow{Category_x}\right| * \left|\overrightarrow{Category_y}\right|}$$

我們列出部分 VH 類的動詞與各類動詞的相似度於表格 4。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 類兩者皆爲狀態動詞，他們的差別僅在於可接的論元數量。VI 類爲類單賓動詞，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 類的相似程度爲次高，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 4. *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

4.5 相似詞選取

在使用相似法來預測動詞分類的過程中，三個主要的步驟。一爲未知動詞的相似詞的選取，二爲測量未知動詞與相似詞的相似度，三爲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爲新的未知動詞預測詞類的依據，計算新的未知動詞($Word_{unknown}$)與訓練語料中動詞 ($Word_{known}$)的定義如下：

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

$$Sim(Word_{unknown},Word_{known}) = weight_1*Sim(wordbase_{1,unknown},wordbase_{1,known}) + weight_2*Sim(wordbase_{2,unknown},wordbase_{2,known}) + ... + weight_n*Sim(wordbase_{n,unknown},wordbase_{n,known})$$

若採用這種方法必須計算訓練語料中的每一個辭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或後詞基相同的相似詞爲計算標的。尋找到前詞基相同或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。計算兩個辭彙相似度的方法，如下：

$$Sim(Word_{unknown},Word_{known}) = \alpha*HowNetSimScore(Base_i,Base_j) + \beta*CategoryScore(category(Base_i),category(Base_j))$$

$\alpha+\beta=1$

$Word_{known}$ 爲相似詞

$Base_i$ 爲未知動詞與相似詞相異的詞基

$Base_j$ 爲相似詞與未知動詞相異的詞基

若未知詞爲「唱完」爲未知動詞，「講完」爲相似詞，即 $Base_i$ 爲「唱」，$Base_j$ 爲「講」。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。

4.6 相似法參數選定

相似法中需要討論下列三點。一、語意相似度測量的比重，即主要義原與次要義原比重的變化對正確率的影響。二、語意與詞類的比重，即語意分數（來自知網）與詞類分數的比重之變化對正確率的影響。三、K 值的變化，即相似例子數量多少對正確率的影響。

使用相似法預測未知動詞會出現兩種類型不能預測的未知動詞：一、找不到相似例子的未知動詞。二、未知動詞的詞基爲知網中未收錄的詞彙，因此無法計算相似度。

爲了尋找最佳的比重，在訓練語意的比重、語意與詞類比重與相似詞數量的實驗中，我們先將有詞基不存在於知網的未知動詞刪除，相似法不能預測的未知動詞僅剩下一種類型，即找不到相似例子的未知動詞。

本節對於正確率的定義爲：

正確率＝猜測正確的未知動詞/(1000-不能處理的未知動詞)

4.5.1 語意相似度比重調整

首先要固定兩個變數，語意與詞類的比重與 K 值大小，才能觀察出相似度比重的變化對正確率的影響。因此我們先給予 K=1，語意與詞類比重爲 1 與 0。我們從未知動詞語料中重複 10 次隨機取出 1000 個未知動詞作爲測試語料，其餘的未知動詞作爲訓練語料，計算 10 次所得到的正確率的平均，圖 2 爲依照相似度比重的變化對正確率的影響製成的圖表。10 次實驗詳細的數據請參見附錄二表格 13。

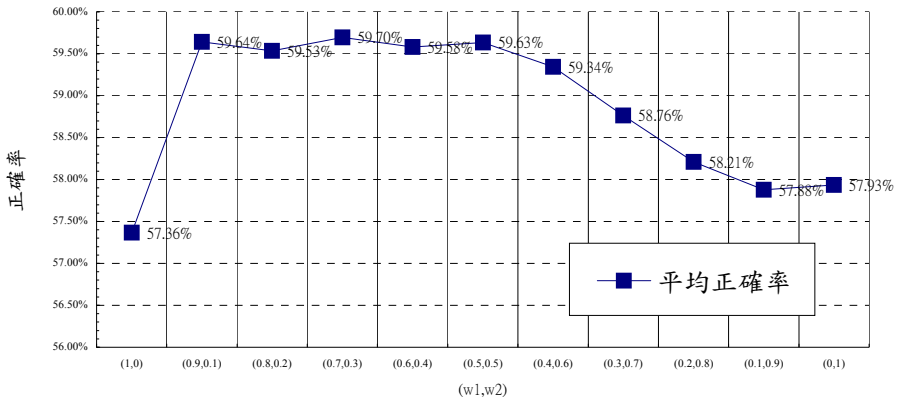

***圖 2** 語意相似度主要義原與次要義原(w1,w2)與正確率關係圖*

由上表可以看出主要義原的比重爲 0.7 與次要義原的比重爲 0.3 時可以得到最高的正確率 59.70%，因此在本實驗中我們使用 0.7 與 0.3 作爲主要義原與次要義原的比重。

4.5.2 語意與詞類比重評量

經由上節的實驗，我們將相似度比重設定 w_1 爲 0.7 與 w_2 爲 0.3 與 K=1。從訓練語料中隨機抽取 1000 個未知動詞重複處理 10 次，觀察語意與詞類比重的變化對正確率的影響。

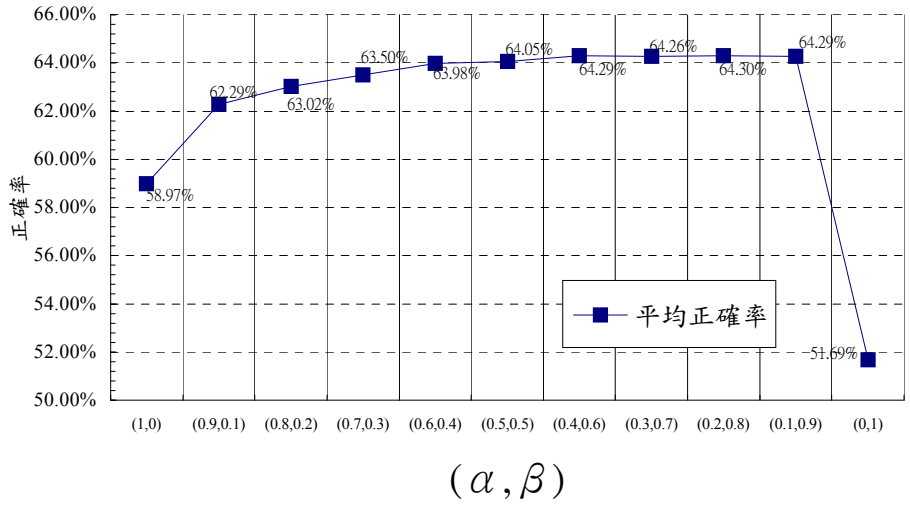

***圖 3** 語意與詞類比重(α, β)與正確率關係圖*

從圖 3 觀察出兩個現象：一、語意類分數所佔的比重越大，使得正確率越降低。二、當語意或詞類所佔的比例爲 0 時會造成正確率的驟降。語意相似度的比重爲 0.2 與詞類相似度的比重爲 0.8 時可以得到最高的正確率 64.30%，因此在本實驗中我們使用 0.2 與 0.8 作爲語意與詞類的比重。10 次實驗詳細的數據請參見附錄二表格 14。

4.5.3 K值變化[4]

基於上述的實驗訓練出來的結果，我們現在將語意相似度中主要義原與次要義原的比重設定爲 0.7 與 0.3，而語意與詞類的比重設定爲 0.2 與 0.8，觀察訓練語料大小與相似詞數量的變化對正確率的影響。

取 K 個相似詞的方法爲將尋找到相似詞先依照詞類分組。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。將未知動詞依照詞類分類的原因是避免受到量大的詞類的影響，而相似例子不足 K 個的未知動詞以其擁有相似例子數量計。

在本節我們設計了兩個實驗：第一個實驗的訓練語料純粹爲未知動詞（與上述調整比重的實驗相同）。第二個實驗則是將辭典中的動詞加入到訓練語料當中，觀察訓練語料的增大對正確率的影響，

當分類資料純粹來自訓練語料時，我們觀察到 K 值的增大對正確率有正面的影響。在十次的實驗中，平均約 K=9 時可以達到最佳的正確率 68.37%，參見圖 4，但 K 值若繼續增大使正確率下降，原因在於當 K 過大時，原本一些相似度較遠或相似度爲 0 的相似詞都會被納入計算，造成實驗的正確率降低。十次實驗詳細的數據請參見附錄二表格 15。

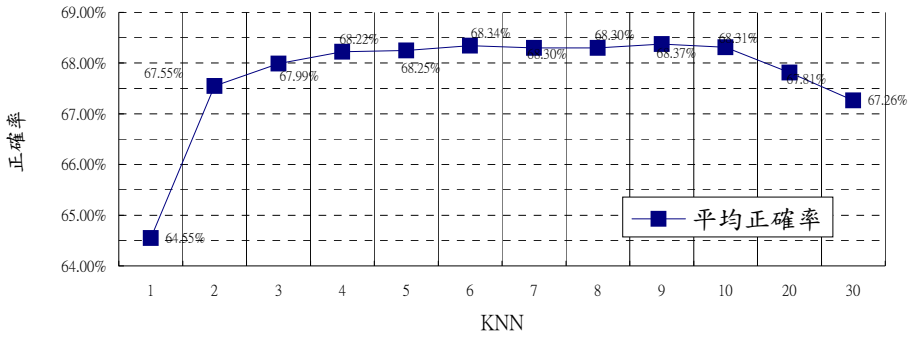

圖 4 相似詞數量(KNN)與正確率關係（訓練語料爲未知動詞）

[4] 這個值是 K-nearest neighbor (KNN)分類法的參數。

我們假設訓練語料的多少會對整體正確率有影響，因此我們將辭典中的動詞放入我們的訓練語料語料中，觀察訓練語料增多時正確率的變化。

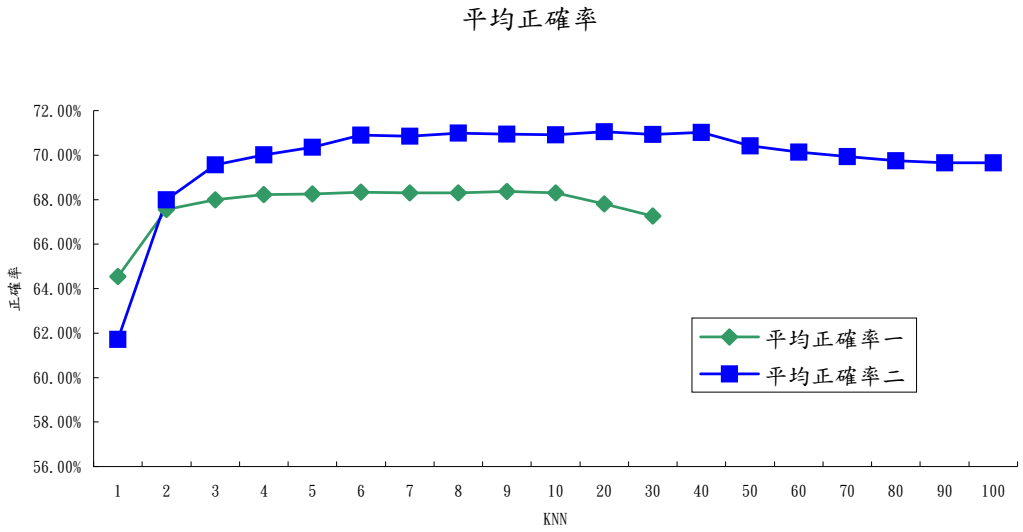

***圖 5** 相似詞數量(KNN)與正確率關係*

（平均正確率一訓練語料爲未知動詞平均正確率二訓練語料爲未知動詞與字典）

從圖 5 中我們觀察到，當訓練語料爲未知動詞加上字典時，正確率隨之增長。當 K=20 時，可以達到最高的正確率 71.05%。訓練語料爲未知動詞加上字典所得到的正確率比單使用未知動詞作爲訓練語料約提高 2.68%。足見當 K 值固定時，訓練語料增多，正確率會隨之提升。十次實驗詳細的數據請參見附錄二表格 16。

表格 5 爲以單純使用訓練語料與訓練語料加字典爲訓練語料所做實驗的結果。從兩個不同訓練語料的實驗中，我們觀察到當 K=1 時，使用未知動詞作爲訓練語料的正確率較使用未知動詞與字典作爲訓練語料的正確率高出 2.83%，參見表格 5。我們從字典中的詞彙特性來解釋 K=1 時以字典加上未知動詞所得的正確率較低的原因。字典中收錄的詞彙有一部份是不具有語意透明性，不能從字面上得出其語意，因此，當我們使用未知動詞與字典作爲訓練語料時，若我們僅取一個相似詞，很有可能取到這些不具有語意透明性的詞彙，干擾我們的判斷結果，就造成了當 K=1 時正確率較低的結果。

表格 5. *訓練語料爲未知動詞與未知動詞加上字典中詞彙比較表格*

未知動詞 KNN	平均正確率(10 次)	未知動詞+字典 KNN	平均正確率(10 次)
1	64.55%	1	61.72%
2	67.55%	2	68.00%
3	67.99%	3	69.56%
4	68.22%	4	70.01%
5	68.25%	5	70.35%
6	68.34%	6	70.90%
7	68.30%	7	70.85%
8	68.30%	8	70.99%
9	**68.37%**	9	70.95%
10	68.31%	10	70.92%
20	67.81%	**20**	**71.05%**
30	67.26%	30	70.93%

5. 綜合測試結果

本節以上述實驗所調整出來的比重，進行最後對本系統正確率的評估，本節的測試語料爲最初保留的最後測試語料。在實驗中我們得到當語意的比重爲 0.2，詞類爲 0.8，語意相似度內主要義原的比重爲 0.7，次要義原的比重爲 0.3，訓練語料爲未知動詞加上字典且 K=20 時可以達到最佳的正確率。我們以上述的比重作爲評估最後測試語料結果的比重，進行下列兩個實驗：一、以相似法處理測試語料。二、結合規則法與相似法處理測試語料。

我們使用相似法，配合上述調整出來的比重，處理先前所保留的 1000 筆資料，得到正確率爲 68.67%，但在這 1000 個未知動詞當中，有 52 個動詞無法處理，參見表格 6。

表格 6. *相似法實驗評估*

(w_1,w_2)	(α,β)	KNN	相似法正確率	無法處理的未知動詞
(0.2,0.8)	(0.7,0.3)	20	68.67%	52

結合前面所提到的規則法與相似法，先將這 1000 個未知動詞使用規則法處理，規則法無法處理的動詞再使用相似法處理，結合兩種方法預測未知動詞的詞類，結合後的正確率爲 70.80%，比單用相似法做預測的正確率高出 2.13%，無法處理的未知動詞也減

少至 31 個。

表格 7. 規則法結合相似法實驗評估

(w_1,w_2)	(α,β)	KNN	相似法正確率	規則法正確率	結合後正確率	無法處理的未知動詞
(0.2,0.8)	(0.7,0.3)	20	65.04%(467/718)	87.25%(219/251)	70.80%	31

6. 結果分析

6.1 錯誤分析

我們動詞分類的系統約有三成的錯誤率，因此我們在這一節分析這些預測錯誤的未知動詞的特性，作爲改進預測分類系統的參考。本節討論的未知動詞包括猜測錯誤的未知動詞與系統無法處理的未知動詞。

6.1.1 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未知動詞中，有一部份爲資料本身的問題，主要可分爲兩類。一、已詞彙化的詞語。二、標記有疑問的詞彙，可參見表格 8 中的例子。

猜測錯誤的例子爲較罕見的詞語，如：合祀、中邪與駁坎等。該詞彙無法從詞彙的組成成分觀察出該詞彙的意義來，這些不具語意透明性的詞彙的解決方式爲將這類詞彙新增入辭典中。其次爲標記有疑問的詞彙，如：「打暈」標記爲 VC 類，這些詞彙的標記可能爲語料庫中的錯誤。

表格 8. 猜測結果

	未知動詞	詞類標記
詞彙化詞彙（不具語意透明性）	合祀	VA
	中邪	VA
	駁坎	VH
標記有疑問詞彙	打暈	VC
	稱美	VC
	車拼	VC
	理直	VC

6.1.2 無法處理之未知動詞

在表格 9 中我們將不能處理的未知動詞爲兩類：一、無相似分數，找到相似詞，但是無法計算相似度。二、無相似詞彙，沒辦法在訓練語料中找到相似詞。

表格 9. *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、升進到、起酒疹、大收紅、上山下海、泫然淚下、遊手好閒、彙寄到、歸併到
無相似辭彙	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、松蘿垂挂、喁喁情話、萬民歸心、克己複禮

無法計算相似分數的原因在於知網沒有收錄該未知動詞組成的詞基，就算有相似詞也無法計算相似度。如：「泫然淚下」尋找到兩個相似詞---「泫然欲滴」與「泫然欲泣」。但是，「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。

在 4.5.3 節的兩個實驗中，觀察到當訓練語料僅爲未知動詞時，在十次的實驗中平均約有 64.7 個未知動詞無法辨識，但當訓練語料的數量增大時，不能處理的動詞數量便平均降低了 7.7 個。

表格 10. *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（沒有相似例子）
未知動詞	64.7
未知動詞＋字典	7.7

無法找到相似辭彙的未知動詞大多爲 VH 類的成語，如：「蕞爾小邦」。這一類的未知動詞處理方法爲收錄辭典。

6.2 語料分析

我們討論三個語料問題對於本實驗造成的影響。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、知網定義義原數量。

6.2.1 未知詞定義與抽取未知詞的方法

我們在本節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，本文未知詞的定義爲不存在於字典中的辭彙，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辭彙的語意，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、離去、遄飛、熏繞、絜靜、歍彼等。我們認爲解決這部分辭彙最好的方法就是將這一類型的辭彙全部收錄字典中。

6.2.2 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：「V+不了」這種結構，在 V 屬動作動詞的情況下，我們發現有

部分的標記人員將「V+不了」這種結構的動詞標記成 V 的類別，即仍屬動作動詞；另外有部分的人則將「V+不了」標記成爲一個狀態動詞，論元結構分類不改變。例如，「抵擋不了」標記爲 VJ 類(狀態單賓動詞)，「阻擋不了」標記爲 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爲「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，而另外一部分人認爲加上「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，我們無法從中明確的歸納出規則。標記不一致的原因主要來自於詞彙本身的多義與標記人員給予標記時的不完整與標記錯誤，而訓練語料中辭彙標記正確與否會直接的影響到相似詞法的正確率。雖然我們認爲這類型的辭彙的確很難去決定分類，但希望有個統一的規則，若這類型的標記爲多義，則希望標記人員將所有可能的詞類標記出，不僅僅是標記其中一種詞類，將這種類型的辭彙給予一致性的標記。我們也希望藉由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，使得語料庫標記更爲一致。

6.2.3 知網動詞定義義原數量

從表格 11 中可觀察到在知網中四個開放性詞彙的詞類擁有義原數量的比例。75%的動詞僅具有一個義原，這種情形不利於使用主要義原與次要義原來計算詞彙相似度的方法，因爲大部分的詞彙都僅具有一個主要義原，容易造成同義詞的產生，只要是主要義原相同，相似度即可達到最高的相似分數，這樣使得計算語意的方法變的不客觀，不易區別近義詞的距離。

表格 11. *知網中開放性詞類擁有義原數量分配表*

義原數量 / 詞類	1	2	3	4	5	6
ADJ	1.20%	1.68%	32.91%	62.68%	1.36%	0.16%
ADV	10.86%	2.86%	78.28%	7.99%	0.00%	0.00%
N	25.45%	26.19%	31.52%	13.28%	3.10%	0.41%
V	75.45%	7.78%	11.65%	4.37%	0.62%	0.13%

7. 結論與未來工作

本論文中利用規則法與相似法來判斷動詞的分類。首先，規則法可分成兩個主要的判斷方式。一、依照未知動詞的組成的關鍵字決定其分類，如：軍國「化」爲 VHC 類。二、依照未知動詞的構成組合決定其分類，「aabb」的辭彙爲狀態類，如：平平靜靜；「abab」組合的辭彙爲動作類，如：準備準備。其次，相似法則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。結合規則法與相似法的正確率爲 70.80%。

分析猜測錯誤的未知動詞中，部分的辭彙爲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辭彙增多時，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辭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。另外，使用知網的詞彙定義義原數量不多也影響到我們計算詞彙相似度。

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。

未來我們的工作將處理動詞的相似法應用到所有的未知詞上，同時也將使用相似法來幫助未知詞作語意的分類。相似法的好處在於不僅可以預測未知詞的詞類分類，同時也可以預測語意與結構分類。當兩個辭彙相似度高時，表示這兩個辭彙的詞類、語意類與結構必定相似，因此利用相似法不僅可以預測未知詞的分類，也可以幫助未知詞作語意上的自動分類，將未知詞與已存在的詞彙語意網路系統作連結。以相似法預測未知詞的詞類、語意類與結構不僅對斷詞標記系統有利，也將可應用到其他自然語言的系統上。

參考文獻

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。

中央研究院詞知識庫小組。«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。

中央研究院詞知識庫小組。«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»。臺北：臺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。<中文裏「動名」述賓結構與「動名」偏正結構的分析>，«第八屆計算語言學會論文集»，1996，頁 1-29。

陳克健、陳超然。<語料庫爲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»。臺北：學生書局，1988。

董振東、董強。知網---中文資訊結構庫。<http://www.keenage.com>，2000。

董振東、董強。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。

賴育升、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>，«第十三屆計算語言學會會議論文集»，2000，頁 135-156。

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 vol. 3 no. 1, 1998, pp. 27-44.

Chen, Keh-Jiann and Ming-Hong Bai. "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.

Resnik, Philip. "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 Probabilistic Model," Computational Linguistics 19, 1993, pp. 359-382.

附錄一

表格 12. *中研院詞庫小組詞類標記 [1993]*

詞類標記	說明	詞類標記	說明
A	非謂形容詞	Neu	數量定詞
Caa	對等連接詞	Nf	量詞
Cab	連接詞	Ng	後置詞
Cba	連接詞	Nh	代名詞
Cbb	關聯連接詞	P	介詞
D	副詞	SHI	是
Da	數量副詞	T	語助詞
DE	的、之、得、地	VA	動作不及物動詞
Dfa	動詞前程度副詞	VAC	動作使動動詞
Dfa	動詞後程度副詞	VB	動作類及物動詞
Di	時態標記	VC	動作及物動詞
Dk	句副詞	VCL	動作及物動詞＋地方賓語
FW	外文標記	VD	動作雙賓動詞
I	感嘆詞	VE	動作句賓動詞
Na	普通名詞	VF	動作謂賓動詞
Nb	專有名詞	VG	分類動詞
Nc	地方詞	VH	狀態不及物動詞
Ncd	位置詞	VHC	狀態使動動詞
Nd	時間詞	VI	狀態 類及物動詞
Nep	指定代詞	VJ	狀態及物動詞
Neqa	數量定詞	VK	狀態句賓動詞
Neqb	後置數量定詞	VL	狀態謂賓動詞
Nes	特指定詞	V_2	有

附錄二

表格 13. *語意相似度比重與正確率變化表*

語意相似度比重 (w_1,w_2)	1(%)	2(%)	3(%)	4(%)	5(%)	6(%)	7(%)	8(%)	9(%)	10(%)	平均(%)
(1,0)	57.60	56.66	57.13	58.64	58.05	57.19	59.31	56.16	56.42	56.50	57.36
(0.9,0.1)	60.93	58.23	59.23	61.21	59.48	60.22	60.97	58.15	58.22	59.79	59.64
(0.8,0.2)	60.93	58.23	59.23	61.31	59.05	60.11	60.65	57.93	58.43	59.47	59.53
(0.7,0.3)	61.68	58.34	59.23	61.31	59.37	60.32	60.76	58.15	58.64	59.15	59.70
(0.6,0.4)	61.14	58.45	59.55	60.99	59.70	60.00	60.76	57.93	58.54	58.73	59.58
(0.5,0.5)	61.14	58.13	59.01	61.53	59.59	60.22	60.65	58.04	58.85	59.15	59.63
(0.4,0.6)	61.14	58.02	58.91	60.56	59.37	59.57	60.23	57.83	58.75	59.05	59.34
(0.3,0.7)	59.31	57.48	59.23	59.81	58.50	59.89	59.18	57.93	57.90	58.41	58.76
(0.2,0.8)	58.34	57.59	58.91	58.84	57.64	59.68	57.81	57.51	58.32	57.46	58.21
(0.1,0.9)	57.70	57.05	58.48	59.05	57.31	59.24	57.38	57.51	58.11	56.93	57.88
(0,1)	56.90	57.04	58.84	58.54	57.21	58.72	57.67	57.32	59.03	58.07	57.93

表格 14. *語意與詞類相似度比重與正確率變化表*

語意與詞類比重 (α,β)	1(%)	2(%)	3(%)	4(%)	5(%)	6(%)	7(%)	8(%)	9(%)	10(%)	平均(%)
(1,0)	60.49	58.64	58.54	58.17	59.85	57.80	61.87	58.19	58.64	57.53	58.97
(0.9,0.1)	63.40	60.85	62.13	62.13	63.64	60.59	64.21	61.89	62.49	61.57	62.29
(0.8,0.2)	63.92	60.95	63.82	61.92	64.16	61.86	65.16	62.74	63.33	62.31	63.02
(0.7,0.3)	64.77	61.38	64.56	62.45	64.38	62.61	65.89	63.05	63.65	62.31	63.50
(0.6,0.4)	65.51	61.80	64.87	63.50	64.69	63.98	65.68	63.58	63.44	62.74	63.98
(0.5,0.5)	65.82	61.80	65.08	63.50	64.69	64.62	65.58	63.89	62.80	62.74	64.05
(0.4,0.6)	65.61	61.69	65.72	63.92	65.64	64.41	65.58	64.11	63.01	63.16	64.29
(0.3,0.7)	65.51	61.38	65.61	64.35	65.86	64.19	64.84	64.84	62.80	63.27	64.26
(0.2,0.8)	65.08	61.90	66.03	64.77	65.33	63.88	65.05	64.95	62.70	63.27	64.30
(0.1,0.9)	65.08	62.01	66.24	64.98	65.22	63.56	65.16	65.16	62.49	62.95	64.29
(0,1)	52.71	51.43	52.16	51.42	54.86	50.65	48.94	51.60	50.90	52.20	51.69

表格 15. *K 值與正確率變化表（訓練語料爲未知動詞）*

KNN	1(%)	2(%)	3(%)	4(%)	5(%)	6(%)	7(%)	8(%)	9(%)	10(%)	平均(%)
1	63.73	65.30	65.30	66.81	64.28	64.14	61.92	65.13	63.72	65.17	64.55
2	67.23	67.72	69.30	70.03	67.97	67.09	64.14	68.24	66.09	67.68	67.55
3	67.44	67.62	70.36	69.82	68.70	67.83	65.51	69.09	65.66	67.89	67.99
4	68.29	68.25	70.46	70.14	69.13	68.35	65.19	69.20	65.66	67.57	68.22
5	68.29	68.14	70.68	69.93	68.91	68.14	65.51	69.20	65.77	67.89	68.25
6	69.25	68.57	70.57	69.61	68.39	68.14	65.72	69.41	65.55	68.20	68.34
7	68.93	68.57	70.89	69.82	68.39	67.93	65.82	69.09	65.55	67.99	68.30
8	69.46	68.46	71.10	69.93	68.49	67.51	65.82	69.09	65.23	67.89	68.30
9	69.88	68.04	70.99	70.14	68.49	67.72	65.72	69.20	65.55	67.99	68.37
10	69.78	67.72	70.89	70.03	68.18	67.93	65.61	69.09	65.77	68.10	68.31
20	69.14	67.19	69.94	69.82	68.18	67.41	65.08	68.66	65.23	67.47	67.81
30	68.50	66.67	69.41	69.30	67.54	66.88	64.56	68.56	64.37	66.84	67.26

表格 16. *K 值與正確率變化表（訓練語料爲未知動詞）*

KNN	1(%)	2(%)	3(%)	4(%)	5(%)	6(%)	7(%)	8(%)	9(%)	10(%)	平均(%)
1	60.08	61.66	63.89	62.53	63.28	63.41	59.43	60.46	62.30	60.20	61.72
2	65.93	70.49	69.71	67.68	68.21	68.55	66.09	68.11	69.35	65.93	68.00
3	68.15	71.91	70.81	69.70	69.11	69.76	67.41	69.82	71.27	67.64	69.56
4	68.35	71.91	70.71	69.60	69.52	69.96	68.82	70.42	72.28	68.54	70.01
5	68.25	71.91	70.81	69.90	70.22	70.36	70.13	70.62	72.38	68.94	70.35
6	68.75	73.12	71.41	70.61	70.12	70.77	70.74	71.33	72.98	69.15	70.90
7	68.75	72.82	71.72	70.81	69.72	70.67	70.33	71.13	73.29	69.25	70.85
8	69.05	72.92	71.92	70.91	69.62	70.77	70.84	70.93	73.59	69.35	70.99
9	68.85	72.52	72.32	70.91	69.42	70.16	71.34	71.23	73.39	69.35	70.95
10	68.75	73.02	71.92	70.71	69.32	70.06	71.64	71.33	73.08	69.35	70.92
20	68.45	72.82	73.22	71.01	69.22	70.56	71.44	71.33	72.48	69.95	71.05
30	68.55	72.62	71.82	70.71	69.62	71.17	70.94	70.93	72.08	70.85	70.93
40	68.45	72.72	72.02	70.91	69.82	71.27	71.04	71.13	71.88	70.95	71.02
50	67.64	71.91	71.21	70.00	69.52	70.97	70.64	70.82	71.07	70.35	70.41
60	67.54	71.70	70.81	69.60	69.32	70.97	69.93	70.82	70.56	70.05	70.13
70	66.83	71.40	70.81	69.60	68.91	70.87	69.93	70.52	70.36	70.05	69.93
80	66.83	71.20	70.61	69.29	68.71	70.87	69.73	70.32	70.06	69.85	69.75
90	66.53	70.89	70.51	69.39	68.51	70.87	69.83	70.22	69.96	69.85	69.66
100	66.53	70.89	70.51	69.19	68.61	70.87	69.83	70.22	69.96	69.95	69.66