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

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。

$$\mathrm{HowNetSimScore}(Word_1, Word_2) = \max_{x,y} \mathrm{EntrySimScore}(Word_1Entry_x, Word_2Entry_y)$$

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

$$\begin{aligned}&\mathrm{EntrySimScore}(Word_1Entry_x, Word_2Entry_y)\\ &= w_1 * \mathrm{PrimaryScore}(Sem_{x,1} \cap Sem_{y,1})\\ &+ w_2 * \mathrm{SecondaryScore}((Sem_{x,2}...Sem_{x,n}),(Sem_{y,2}...Sem_{y,m}))\end{aligned}$$

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。

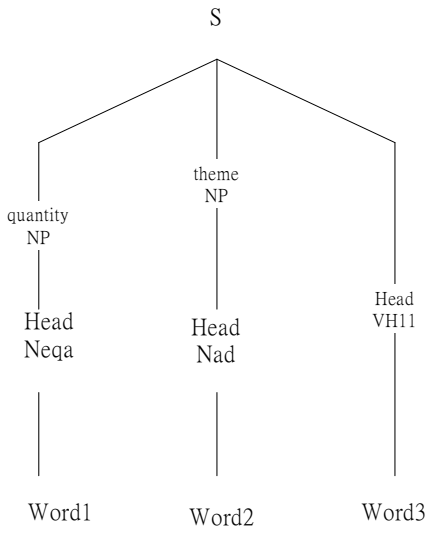


圖 1 中文句結構樹樹狀圖與歸納規則

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。

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

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。

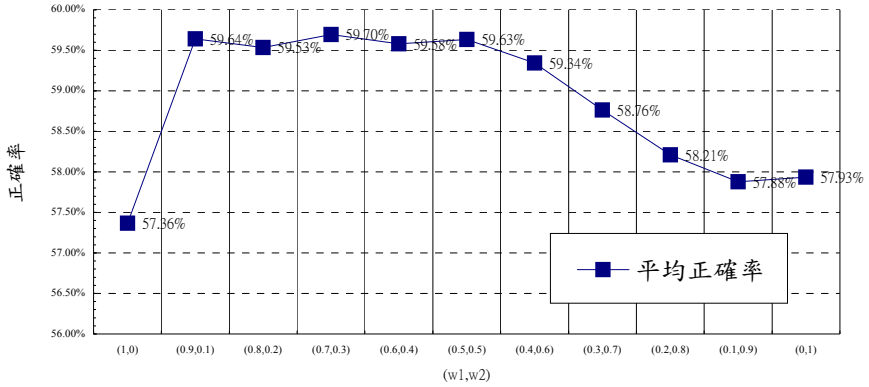


圖 2 *語意相似度主要義原與次要義原(w1,w2)與正確率關係圖*

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。

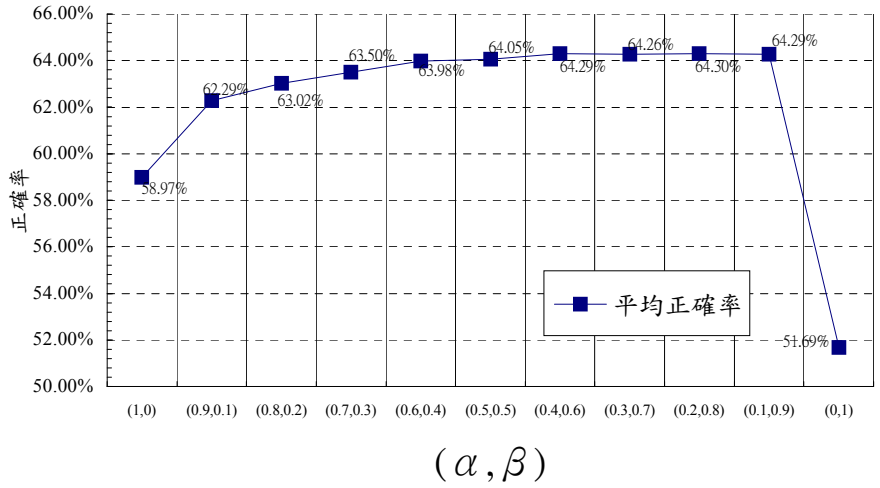


圖 3 *語意與詞類比重(α, β)與正確率關係圖*

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。

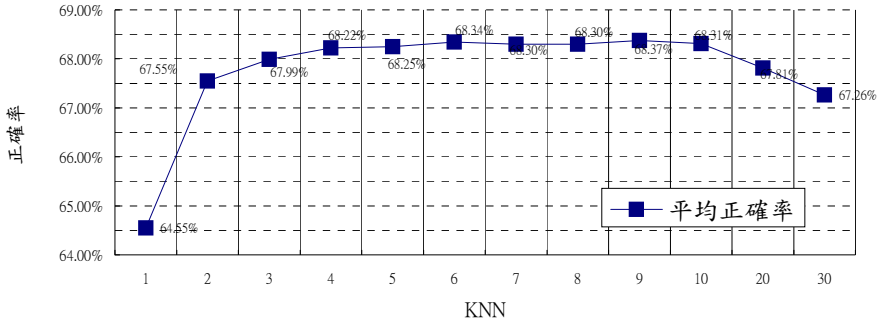


圖 4 相似詞數量(KNN)與正確率關係（訓練語料爲未知動詞）

[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。

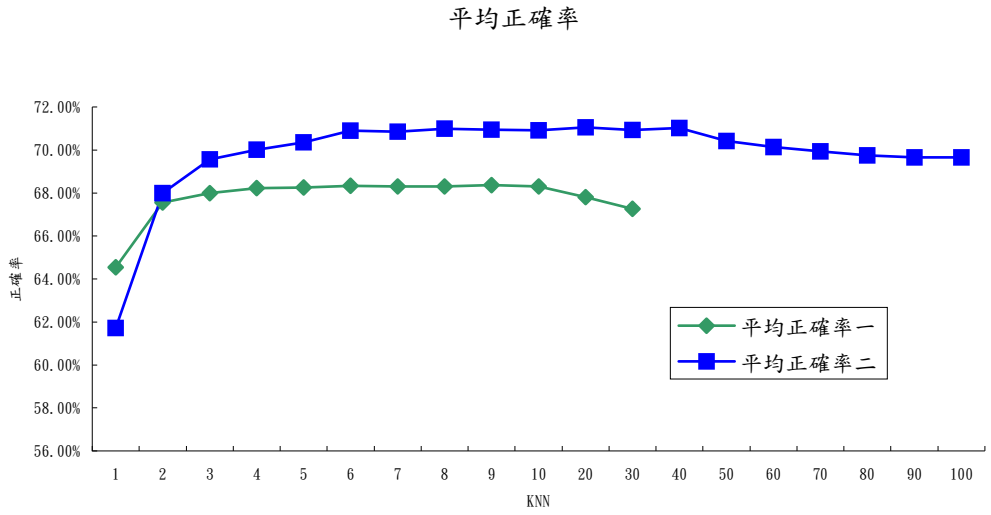


圖 5 *相似詞數量(KNN)與正確率關係*

（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、激灩、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、歈彼等。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

Computational Linguistics and Chinese Language Processing
Vol. 7, No. 1, February 2002, pp. 29-46

Word Sense Disambiguation and Sense-Based NV Event Frame Identifier

Jia-Lin Tsai*, Wen-Lian Hsu and Jeng-Woei Su

Abstract

Word sense is ambiguous in natural language processing (NLP). This phenomenon is particularly keen in cases involving noun-verb (NV) word-pairs. This paper describes a sense-based noun-verb event frame (NVEF) identifier that can be used to disambiguate word sense in Chinese sentences effectively. A knowledge representation system (the NVEF-KR tree) for the NVEF sense-pair identifier is also proposed. We use the word sense of Hownet, which is a Chinese-English bilingual knowledge-base dictionary.

Our experiment showed that the NVEF identifier was able to achieve 74.8% accuracy for the test sentences studied based only on NVEF sense-pair knowledge. By applying the techniques of longest syllabic NVEF-word-pair first and exclusion word checking, the sense accuracy for the same test sentences could be further improved to 93.7%. There were four major reasons for the incorrect cases: (1) lack of a bottom-up tagger, (2) lack of non-NVEF knowledge, (3) inadequate word segmentation, and (4) lack of a multi-NVEF analyzer. If these four problems could be resolved, the accuracy would reach 98.9%.

The results of this study indicate that NVEF sense-pair knowledge is effective for word sense disambiguation and is likely to be important for general NLP.

Keywords: word sense disambiguation, event frame, top-down identifier, Hownet

1. Introduction

Word sense disambiguation (WSD) has been a pervasive problem in natural language processing (NLP) since 1949 [Weaver 1949]. Word sense ambiguity (or lexical ambiguity), is generally classified into two types: syntactic and semantic ambiguity [Small *et al*. 1988, Krovetz *et al*. 1992]. Syntactic ambiguity is caused by differences in syntactic categories (e.g.

* Institute of Information Science, Academia Sinica, Nankang, Taipei, Taiwan, R.O.C.
E-mail: {tsaijl,hsu,raycs}@iis.sinica.edu.tw

"play" can occur as a *noun* or *verb*). Semantic ambiguity is caused by homonymy (e.g. "bank" in "*to put money in a bank*," "*the bank of a river*") or polysemy (e.g. "face" in "*human face*," "*face of a clock*"). Although many approaches have been adopted to disambiguate word sense, algorithms for word sense determination still are not reliable [Krovetz *et al*. 1992, Resnik *et al*. 2000]. Human beings usually can disambiguate word sense by using additional information from the speaker, the writer or the context. When out-of-context (or out-of-sentence) information is not symbolized and processed in the computer, WSD either becomes very difficult or, sometimes, impossible. Therefore, it is crucial to investigate what kind of knowledge is useful for WSD [Krovetz *et al*. 1992].

According to a study in cognitive science [Choueka *et al*. 1983], people often disambiguate word sense using only a few other words in a given context (frequently only one additional word). Thus, the relationships between one word and others can be effectively used to resolve ambiguity. Furthermore, from [Small *et al*. 1988, Krovetz *et al*. 1992, Resnik *et al*. 2000], most ambiguities occur with nouns and verbs, and the object-event (i.e. noun-verb) distinction is a major ontological division for humans [Carey 1992]. However, no clear data has been collected to support these claims. These observations motivated us to demonstrate through an experiment, how noun-verb (NV) relationships can be used to disambiguate word sense in Chinese sentences.

In this paper, we shall focus on word sense disambiguation involving NV word-pairs since these are most troublesome. Consider the following sentence: "**這輛車行駛順暢** (This car moves well)." In this sentence, we have two possible NV word-pairs, "**車-行駛** (car, move)" and "**車行-駛**(auto-shop, move)." It is clear that the permissible NV word-pair is "**車-行駛** (car, move)." We shall call such a permissible NV word-pair an *NV-event frame* (*NVEF*) word-pair. Using a collection of pre-learned NVEF word-pairs, we can identify the NVEF word-pair "**車-行駛**" from the sentence "**這輛車行駛順暢**." The word "**車**" in a dictionary can have three possible senses: 'surname' (noun), 'car' (noun) and 'turn' (verb). To resolve this ambiguity, we can use the pre-defined sense of the NVEF word-pair "**車-行駛**(car, move)" to determine that the correct sense of the Chinese word "**車**" is "car" in the above Chinese sentence.

In this paper, we shall show that knowledge of NVEF sense-pairs (to be defined in Section 2) can be effectively used to resolve word sense ambiguity. In the next section, we will propose an NVEF sense-pair identifier, which is based on pre-stored knowledge of NVEF sense-pairs. We use this NVEF sense-pair identifier to identify NVEF word-pairs in an input sentence and to determine the corresponding word senses. In Section 3, we will present and analyze the results of a WSD experiment on a set of test sentences using the NVEF sense-pair identifier. Finally, we will give conclusions and directions for future research in Section 4.

2. Development of an NVEF Sense-Pair Identifier

We use Hownet [Dong] as our system' s Chinese machine-readable dictionary (MRD). Hownet is a Chinese-English bilingual knowledge-base dictionary, which provides knowledge of the Chinese lexicon, parts-of-speech (POS) and word senses.

2.1 Definition of an NVEF Sense-Pair

The sense of a word is defined as its DEF (concept definition) in Hownet. Table 1 lists three different senses of the Chinese word "車 (*Che*/car/turn)." In Hownet, the DEF of a word consists of its main feature and secondary features. For example, in the DEF "character|文字,surname|姓,human|人,ProperName|專" of the word "車 (*Che*)," the first item "character|文字" is the main feature, and the remaining three items, "surname|姓," "human|人," and "ProperName|專," are its secondary features. The main feature in Hownet can inherit features in the hypernym-hyponym hierarchy. There are approximately 1,500 features in Hownet. Each of these features is called a *sememe*, which refers to the smallest semantic unit that cannot be further reduced.

Table 1. *Three different senses of the Chinese word "車(Che/car/turn)."*

C.Word [a]	E.Word [a]	Part-of-speech	Sense (i.e. DEF in Hownet)
車	*Che*	Noun	character|文字, surname|姓, human|人, ProperName|專
車	car	Noun	LandVehicle|車
車	turn	Verb	cut|切削

[a] C.Word refers to a Chinese word; E.Word refers to an English word

The Hownet dictionary used in this study contains 50,121 Chinese words, among which there are 29,719 nouns, 16,652 verbs and 16,242 senses (including 9,893 noun-senses and 4,440 verb-senses). Table 2 gives the statistics of the number of senses per Chinese word and the number of Chinese words per sense used in Hownet.

Table 2. *Statistics of the number of senses per Chinese word and the number of Chinese words per sense used in Hownet.*

Item [a]	Total	Noun	Verb
Maximum number of senses per Chinese word	27	14	24
Mean number of senses per Chinese word	1.24	1.14	1.23
Maximum number of Chinese words per sense	374	372	129
Mean number of Chinese words per sense	3.8	3.0	4.6

[a] Similar WordNet statistics can be found in [Voorhees 1993]. (WordNet is a trademark of Princeton University.)

Now, take the NV word-pair "車-行駛 (car, move)" for example. According to the sense of the Chinese word "車 (Che/car/turn)" and the sense of the Chinese word "行駛 (move)," the only permissible NVEF sense-pair for the NV word-pair "車-行駛(car, move)" is "LandVehicle|車"-"VehicleGo|駛." We call such a permissible NV sense-pair an NVEF *sense-pair* in this paper. Note that an NVEF sense-pair is a class that includes the permissible word-pair instance "車-行駛(car, move)."

2.2 Knowledge Representation Tree of NVEF Sense-Pairs

A knowledge representation tree (KR-tree) of NVEF sense-pairs is shown in Fig.1. There are two types of nodes in the KR-tree, namely, ***function nodes*** and ***concept nodes***. Concept nodes refer to words and features in Hownet. Function nodes are used to define the relationships between their parent and children concept nodes. If a concept node A is the child of another concept node B, then A is a subclass of B. Following this convention, we can omit the function node "subclass" (which should exist) between A and B. We can classify the noun-sense class (名詞詞義分類) into 15 subclasses according to their main features. They are "微生物 (bacteria)," "動物類 (animal)," "人物類 (human)," "植物類 (plant)," "人工物 (artifact)," "天然物 (natural)," "事件類 (event)," "精神類 (mental)," "現象類 (phenomena)," "物形類 (shape)," "地點類 (place)," "位置類 (location)," "時間類 (time)," "抽象類 (abstract)" and "數量類 (quantity)." Appendix A gives a sample table of 15 main features of nouns in each noun-sense subclass.

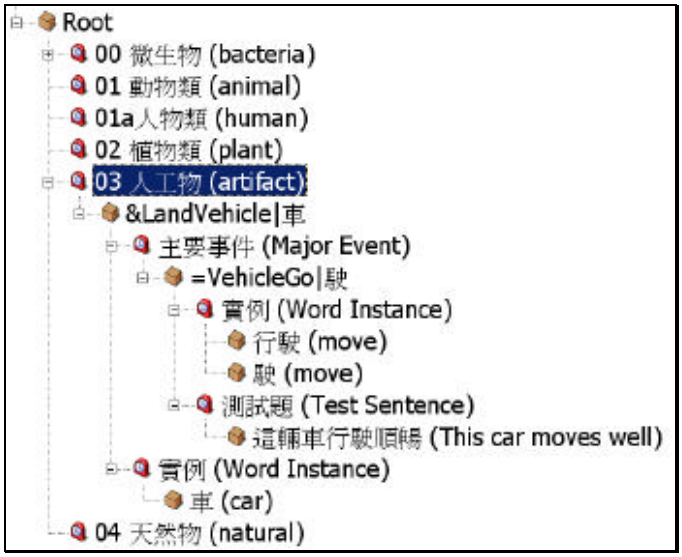


Figure 1 *An illustration of the KR-tree using "人工物 (artifact)" as an example noun-sense subclass. (The English words in parentheses are there for explanatory purposes only.)*

Three function nodes are used in the KR-tree as shown in Fig. 1:

(1) Major-Event (主要事件): The content of its parent node represents a noun-sense subclass, and the content of its child node represents a verb-sense subclass. A noun-sense subclass and a verb-sense subclass linked by a Major-Event function node is an NVEF subclass sense-pair, such as "&LandVehicle|車" and "=VehcileGo|駛" in Fig. 1. To describe various relationships between noun-sense and verb-sense subclasses, we have designed three subclass sense-symbols, in which "=" means "*exact*," "&" means "*like*," and "*%*" means "*inclusive*." An example using these symbols is given below.

Given three senses S_1, S_2 and S_3 defined by a main feature A and three secondary features B, C and D, let

S_1 = A, B, C, D,

S_2 = A, B, and

S_3 = A, C, D.

Then, we have that sense S_2 is in the "=A,B" *exact*-subclass; senses S_1 and S_2 are in the "&A,B" *like*-subclass; and senses S_1 S_2, and S_3 are in the "%A" *inclusive*-subclass.

(2) Word-Instance (實例): The content of its children are the words belonging to the sense subclass of its parent node. These words are self-learned by the NVEF sense-pair identifier according to the sentences under the Test-Sentence nodes.

(3) Test-Sentence (測試題): The content of its children is several selected test sentences in support of its corresponding NVEF subclass sense-pair.

2.3 Generation of NVEF Sense-Pairs

To speedup the creation of the KR-tree, an example-based algorithm is proposed to generate the KR-tree semi-automatically. This algorithm is described below.

Step 1. Select a noun-sense, such as "disease|疾病," in Hownet.

Step 2. Collect all Chinese polysyllabic words of the selected noun-sense. (Monosyllabic words are not considered at this stage.)

Step 3. Select those Chinese un-segmented sentences that include at least one word collected in Step 2 from the Sinica corpus (which is a Chinese corpus of two millions words [CKIP 1995]) or other domain specific collections. For example,

the Chinese sentence "醫生的職責就是預防與治療疾病 (A doctor's job is to prevent a disease and to cure the patient)" is a candidate sentence that includes the Chinese word "疾病 (disease)."

Step 4. Find all possible verb-senses from the sentences selected in Step 3 to form all possible verb-senses for the selected noun-sense. Calculate the frequency for each verb-sense.

Step 5. Sort all possible different verb-senses according to their corresponding frequencies from large to small. (See Fig. 2) Determine a cut-off frequency in the list. Among all verb-senses above the cut-off frequency, manually pick the permissible ones for the selected noun-sense. Meanwhile, determine their subclass sense-symbols (i.e. "&," "%" and "=".)

Step 6. Add these permissible NVEF subclass sense-pairs to the KR-tree.

Note that among the above steps, only step 5 requires human intervention. This step is quite laborious, but through learning, human involvement can be greatly reduced. Fig. 2 shows the top 5 possible verb-senses picked by the above algorithm for the noun-sense "disease|疾病" collected from 302 sentences in the Sinica corpus. In Fig. 2, the permissible verb-senses for the noun-sense "disease|疾病" are "cure|醫治" with a frequency of 24, "Cause Affect|傳染, medical|醫" with one of 23, "Result In|導致" with one of 19 and "obstruct|阻止" with one of 14. It is observed that, if the number of sentences collected in Step 3 is greater than 300, then the top 5 verb-senses will almost always form NVEF sense-pairs with the selected noun-sense.

Figure 2 *Top 5 possible verb-senses for creating permissible NVEF sense subclasses for the noun-sense "disease|疾病."*

2.4 A Primitive NVEF Sense-Pair Identifier

Based on the KR-tree, we shall develop a primitive NVEF sense-pair identifier as follows. For a given sentence, the algorithm will first identify all NVEF sense-pairs in the KR-tree that have corresponding NVEF word-pairs in the sentence. It will then arrange these NVEF sense-pairs and their corresponding NVEF word-pairs into a tree, called a *sentence-NVEF tree,* as shown in Fig. 3.

```
(A single-NVEF sentence)
+這輛車行駛順暢 (This car moves well)
+--+N1
+--+--+LandVehicle|車
+--+--+--+車 (car)
+--+V1
+--+--+VehicleGo|駛
+--+--+--+行駛 (move) (the longest syllablic word)
+--+--+--+駛 (move)

(A multi-NVEF sentence)
+趕馬入畜欄 (Drive the horses into the corral)
+--+N1
+--+--+livestock|牲畜
+--+--+--+馬 (horse)
+--+V1
+--+--+expel|驅趕
+--+--+--+趕 (drive)
+--+N2
+--+--+facilities|設施,space|空間,@foster|飼養,#livestock|牲畜
+--+--+--+畜欄 (corral)
+--+V2
+--+--+GoInto|進入
+--+--+--+入 (into)
```

Figure 3 *Two sentence-NVEF trees for the input Chinese sentences (a) "**這輛車行駛順暢**" (a single-NVEF sentence) and (b) "**趕馬入畜欄**" (a multi-NVEF sentence), respectively.*

A more formal description of the primitive NVEF sense-pair identifier is given below:

Step 1. Input a sentence.

Step 2. Generate all possible NV word-pairs of the input sentence.

Step 3. Check each NV word-pair got in step 2 to see if its corresponding NV sense-pairs can be matched to an NVEF subclass sense-pair in the KR-tree. If matches are found, then use the corresponding noun-senses and verb-senses to form the permissible NVEF sense-pairs, respectively, for this sentence.

Step 4. Arrange all permissible NVEF sense-pairs and their corresponding NVEF word-pairs in a sentence-NVEF tree.

A system overview of the primitive NVEF sense-pair identifier is given in Fig. 4.

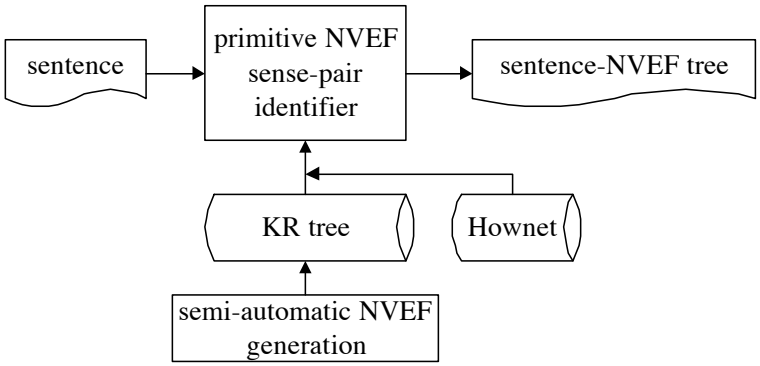


Figure 4 *System overview of the primitive NVEF sense-pair identifier.*

2.5 An NVEF Sense-Pair Identifier

In Fig. 3, the correct segmented results of the two Chinese sentences are "這/輛/車/行駛/順暢" and "趕/馬/入/畜欄," respectively. The upper part of Fig. 3 is a sentence-NVEF tree with a single NVEF sense-pair, "LandVehicle|車"- "VehicleGo|駛," which has two corresponding NV word-pairs, i.e. "車-行駛" and "車-駛." If we further apply the "*longest syllabic NVEF-word-pair first*" strategy (LS-NVWF), the incorrect NVEF word-pair "車-駛" will be successfully dropped. Note that the "longest syllabic word first strategy" is an effective technique for Chinese word segmentation [Chen *et al*. 1986]. The lower part of Fig. 3 is a sentence-NVEF tree with two NVEF sense-pairs including "expel|驅趕"-"livestock|牲畜" (NV word-pair is "馬-趕") and "facilities|設施, space|空間, @foster|飼養, #livestock|牲畜"-"GoInto|進入" (NV word-pair is "畜欄-入").

Another useful technique is to exclude certain nouns or verbs from the sentence-NVEF tree. A word with very low frequency as a noun or a verb is treated as a *word of exclusion* for the NVEF sense-pair identifier. Take the Chinese word "的 (of/target)" as an example. Its frequency as a noun or a verb is only 0.004% (computed according to the Sinica corpus). Thus, "的" becomes a word of exclusion. In our experiment, the *exclusion word list* (*EWL*) consists of those words whose frequencies as nouns or verbs are no greater than 5%. When an NVEF word-pair includes at least one exclusion word, its corresponding NVEF sense-pair is excluded from the sentence-NVEF tree. This process is called *EWL checking*. Appendix B

lists all of the exclusion words used in this experiment.

Thus, our final NVEF sense-pair identifier can be described as follows.

Step 1. Input a sentence.

Step 2. Generate all possible NV word-pairs of the input sentence. Exclude certain word-pairs based on EWL checking.

Step 3. Check each NV word-pair to see if its corresponding NV sense-pairs can be matched to an NVEF subclass sense-pair in the KR-tree. For each NV sense-pair that matches an NVEF subclass sense-pair in the KR-tree, use it to the set of permissible NVEF sense-pairs, respectively, for this sentence. Resolve conflicts using the LS-NVWF strategy.

Step 4. Arrange all permissible NVEF sense-pairs and their corresponding NVEF word-pairs in a sentence-NVEF tree.

A system overview of the NVEF sense-pair identifier is given in Fig. 5.

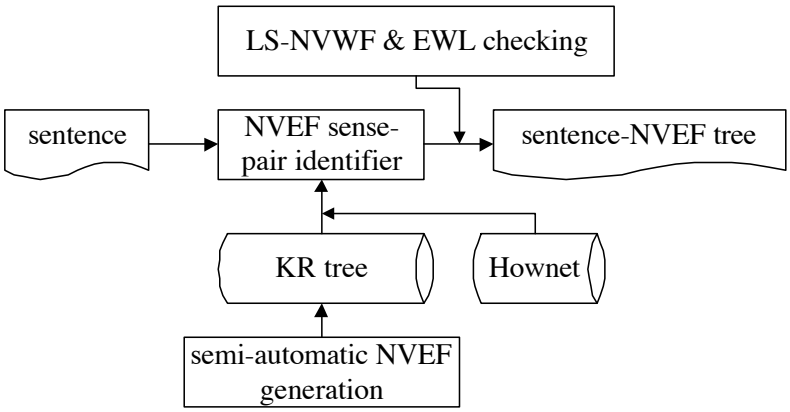


Figure 5 *A system overview of the NVEF sense-pair identifier.*

To evaluate the WSD performance of the NVEF sense-pair identifier, we will consider a WSD experiment in the next section.

3. The WSD experiment

Within a sentence, the number of available NVEF sense-pairs is finite. Consider the Chinese sentence "這輛車行駛順暢 (This car moves well)." Table 3 gives eight possible pairs of NVEF senses found in this sentence, but there is only one permissible NVEF sense-pair, "LandVehicle|車"-"VehicleGo|駛."

To evaluate the performance of WSD by using the NVEF sense-pair identifier with the KR-tee, we define the NVEF sense accuracy for a set of test sentences to be

NVEF sense accuracy = # of successful sentences / # of test sentences, (1)

where a sentence is *successful* if all NVEF sense-pairs and their corresponding NVEF word-pairs obtained from the NVEF sense-pair identifier are correct for this sentence. With the KR-tree, the WSD performance for the test sentences can be evaluated by computing the NVEF sense accuracy. This equation is designed from the viewpoint of natural language understanding. Since NVEF sense-pairs often represent a key feature in the meaning of a sentence, any incorrect NVEF sense-pair identification could result in misunderstanding this sentence.

Table 3. *Eight possible pairs of NVEF senses found in the Chinese sentence "***這輛車行駛順暢** *(This car moves well)."*

C.Word / Noun-sense	C.Word / Verb-sense	P.NVEF [a]
這 / time|**時間**	**行駛** / Go|**駛**	No
這 / time|**時間**	**車** / cut|**切削**	No
這 / time|**時間**	**車** / irrigate|**澆灌**	No
車 / LandVehicle|**車**	**行駛** / VehicleGo|**駛**	Yes
車 / character|**文字**,surname|**姓**,human|**人**,ProperName|**專**	**行駛** / VehicleGo|**駛**	No
車 / machine|**機器**	**行駛** / VehicleGo|**駛**	No
車 / part|**部件**,%tool|**用具**,#recreation|**娛樂**	**行駛** / VehicleGo|**駛**	No
車 / LandVehicle|**車**	**駛** / VehicleGo|**駛**	Yes

[a] P.NVEF represents a permissible NVEF sense-pair.

3.1 WSD Evaluation

The framework of WSD evaluation for the NVEF sense-pair identifier is as follows.

1. Select a set of Chinese test sentences from the Sinica Corpus [CKIP 1995] randomly.
2. Use the tool of example-based possible NVEF generation to search and create all permissible NVEF subclass sense-pairs found in these test sentences in the KR-tree.
3. Apply the NVEF sense-pair identifier to these test sentences and obtain their corresponding sentence-NVEF trees.
4. Compute the NVEF sense accuracy for the test sentences using Equation 1.

In this study, we analyzed 7.7% (=764/9,893) of the noun-senses in Hownet and created 4,028 NVEF subclass sense-pairs in the KR-tree. The minimum, maximum and mean of

characters per sentence (of the 445 Chinese test sentences) were 4, 24 and 11.5, respectively. In addition, the numbers of single-NVEF sentences and multi-NVEF sentences among the test sentences were 96 and 349, respectively.

We conducted the experiment in a progressive manner. The NVEF sense accuracy of the test sentences determined using the NVEF sense-pair identifier with only the knowledge of the KR-tree was 74.8% (see Table 4). When the strategy of adopting the longest syllabic NVEF-word-pair first (LS-NVWF) was used together with the NVEF sense-pair identifier, the NVEF sense accuracy reached 87.6%. When the exclusion word list (EWL checking) was adopted together with the NVEF sense-pair identifier, the NVEF sense accuracy reached 89.2%. When the techniques of both LS-NVWF and EWL checking were adopted with the NVEF sense-pair identifier (see Table 4), the NVEF sense accuracy improved to 93.7%. Meanwhile, along with the NVEF sense-pair identifier, the word-segmentation accuracy (for those ambiguous NVEF word-pairs) for these sentences was 99.6% (443/445). This result also supports the aforementioned claim that the NVEF word-segmentation accuracy was better than the NVEF sense accuracy. Appendix C presents two successful and one unsuccessful sentence-NVEF trees obtained in this experiment.

Table 4. *Results of the WSD experiment for 445 Chinese un-segmented test sentences.*

# of NVEF	NVEF sense accuracy	Using LS-NVWF [a]	Using EWL[b]	Using Both[c]
4,028	74.8%(333/445)	87.6%(390/445)	89.2%(397/445)	93.7%(417/445)

[a] "Using LS-NVWF" represents NVEF sense accuracy using LS-NVWF with the NVEF sense-pair identifier.
[b] "Using EWL" represents NVEF sense accuracy using EWL checking with the NVEF sense-pair identifier.
[c] "Using Both" represents NVEF sense accuracy using both LS-NVWF and EWL checking with the NVEF sense-pair identifier.

3.2 An Analysis of the Unsuccessful Cases

Although the NVEF sense accuracy could reach 93.7% when the techniques of both LS-NVWF and EWL checking were adopted with the NVEF sense-pair identifier, there was still a room for improvement. Below, we have classified the reasons behind the unsuccessful cases into four major types:

(1) ***Lack of a bottom-up tagger:*** There are many specific linguistic units, such as names, addresses, determinative-measure compounds, etc. in sentences which need to be recognized in order to supplement the NVEF sense-pair identifier (which works in a top-down fashion). In this study, 6 sentences were unsuccessful for this reason. Although the techniques of LS-NVWF and EWL checking inadvertently resolved these cases, this is still a potential problem.

(2) ***Lack of Non-NVEF knowledge:*** Consider the Chinese sentence, "**太太要掌握先生的荷包** (A wife wants to take her husband's wallet into her hands)." There were three different noun-senses of the Chinese word, "**先生**(teacher/doctor/husband)," which could form an NVEF sense-pair with the verb-sense "**掌握** (take..into one's hands)." To get the correct noun-sense "**先生** (husband)" for this sentence, we need the knowledge of a noun-noun (NN) sense-pair, such as "**太太** (wife)"-to-"**先生** (husband)," or other contextual information. This knowledge is not available from the KR-tree and needs to be collected separately. In this study, 15 sentences were unsuccessful for this reason, and this problem could not be resolved using the technique of LS-NVWF or EWL checking.

(3) ***Inadequate word segmentation:*** Consider the Chinese sentence, "**他以滿分得到冠軍**(He obtained the championship with a full mark)." There were two possible verbs with the same verb-sense "**分得** (obtain)" and "**得到** (obtain)" that could form NVEF sense-pairs with the noun-sense "**冠軍** (champ)." In this case, we have two conflicting NVEF sense-pairs and need a better segmentation algorithm to determine that the correct verb are "**得到** (obtain)" for this sentence (the correct segmented result of this sentence is "**他/以/滿分/得到/冠軍**"). In this study, 3 sentences were unsuccessful for this reason, and this problem could not be resolved using the technique of LS-NVWF or EWL checking.

(4) ***Lack of a multi-NVEF analyzer:*** Consider the Chinese sentence "**搭飛機離開台北** (Take airplane to leave Taipei)." The NVEF sense-pair identifier detected that there were three NVEF sense-pairs: N_1-V_1: [N_1= **飛機** (airplane),V_1= **搭** (take)], N_2-V_2: [N_2= **台北** (Taipei),V_2=**離開** (leave)], and N_3-V_3: [N_3=**飛機** (airplane),V_3=**離開** (leave)] in the sentence. In this case, N_1-V_1 and N_2-V_2 can be used to construct a permissible bi-NVEF sequence V_1-N_1V_2-N_2 , which will compete with the NVEF sense-pair N_3-V_3 . Currently, such cases are not analyzed since our system does not yet have the knowledge of permissible multi-NVEF sense-pairs. In this study, 5 sentences were unsuccessful for this reason, and this problem could not be resolved using the technique of LS-NVWF or EWL checking.

If these four problems could be resolved, the NVEF sense accuracy could be improved to (417+15+3+5) / (445) = 98.9%.

Based on this experiment, we find that our NVEF sense-pair identifier has the potential to provide the following information for a given sentence: (1) main verbs, (2) nouns, (3) NVEF word-pairs, (4) NVEF sense-pairs, (5) NVEF phrase-boundaries, and (6) the initial relationship among multi-NVEF sense/word-pairs. A correct NVEF sense-pair will naturally

include the correct NVEF word-pair for word segmentation. However, the converse is not true. That is, a correct NVEF word-pair cannot guarantee that the corresponding NVEF sense-pair is permissible. Thus, the NVEF word-segmentation accuracy is normally better than the NVEF sense accuracy.

4. Conclusions and Directions for Future Research

In this paper, we have described an NVFE sense-pair identifier which we attempted to use to disambiguate word sense in Chinese sentences. A WSD experiment was conducted using the NVEF sense-pair identifier with the KR-tree. The knowledge in the KR-tree was created with the help of a semi-automatic NVEF generation tool.

Based on current techniques, our experiment showed that the NVEF sense accuracy reached 93.7% and the NVEF word-segmentation accuracy 99.6%. We have indicated, in Section 3, several ways to further improve the performance of our system, some of which are currently being studied.

Our experiment indicated that NVEF sense-pair knowledge can be used effectively to achieve NVEF word-sense disambiguation in Chinese sentences. It also supports the claim in [Choueka *et al*. 1983] that people usually disambiguate word sense using only a few words (frequently only one word) in the given context. We are particularly pleased to note that the NVEF knowledge can achieve high accuracy in NVEF word-segmentation since correct word-segmentation is one key to a successful Chinese NLP [Slocum *et al*. 1985].

Although we have a semi-automatic NVEF generation tool, it was still a laborious task to create our current level of NVEF knowledge, which constitutes only 7.7% of the entire NVEF knowledge. Hence, a systematic method for fully automatic NVEF knowledge generation is highly desired. Furthermore, we will try to develop a combined top-down and bottom-up NVEF sense-pair identifier that can address the issues involved in the four unsuccessful cases described in Section 3.

We plan to create a full fledged KR-tree so that we can investigate the robustness of the sense-based approach for monolingual and bilingual (e.g. English-Chinese) WSD. The study of NVEF will also be extended to noun-noun pairs, noun-adjective pairs and verb-adverb pairs. Another related research goal is to apply the NVEF sense-pair identifier to other fields of NLP, in particular, document classification, information retrieval, question answering and speech understanding.

Acknowledgements

We are grateful to the our colleagues in the Intelligent Agent Systems Lab., Li-Yeng Chiu, Mark Shia, Gladys Hsieh, Masia Yu and Yi-Fan Chang, who helped us create all the necessary NVEF knowledge for this study. We would also like to thank Prof. Zhen-Dong

Dong for providing us with the Hownet dictionary.

Reference

Carey, S., "The origin and evolution of everyday concepts (In R. N. Giere, ed.)," *Cognitive Models of Science*, Minneapolis: University of Minnesota Press, 1992.

Chen, C. G., K. J. Chen and L. S. Lee, "A model for Lexical Analysis and Parsing of Chinese Sentences," *Proceedings of 1986 International Conference on Chinese Computing*, Singapore, 1986, pp.33-40.

Choueka, Y. and S. Lusignan, "A Connectionist Scheme for Modeling Word Sense Disambiguation," *Cognition and Brain Theory*, 6 (1) 1983, pp.89-120.

CKIP. *Technical Report no. 95-02, the content and illustration of Sinica corpus of Academia Sinica*. Institute of Information Science, Academia Sinica, 1995. http://godel.iis.sinica.edu.tw/CKIP/r_content.html

Dong, Z. and Q. Dong, Hownet, http://www.keenage.com/

Krovetz, R. and W. B. Croft, "Lexical Ambiguity and Information Retrieval," *ACM Transactions on Information Systems*, 10 (2), 1992, pp.115-141.

Resnik, P. and D. Yarowsky, "Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation," *Natural Language Engineering*, 5 (3), 2000, pp.113-133.

Slocum, J. and C. F. Justus, "Transportability to Other Language: The Natural Language Processing Project in the AI Program at MCC," *ACM Transactions on Office Information Systems*, 3(2) 1985, pp.204-230.

Small, S., and G. Cottrell, and M. E. Tannenhaus, *Lexical Ambiguity Resolution*, Morgan Kaufmann, Palo Alto, Calif., 1988.

Voorhees, E. M., "Using WordNet to Disambiguate Word Senses for Text Retrieval," *ACM-SIGIR*, 1993, pp. 171-180.

Weaver, W., *Translation*. Mimeographed, 1949, pp. 12.

Appendix A. A Sample Table of the Main Features of Nouns and their corresponding Noun-Sense Classes

An example Main Feature	Noun-sense Class
bacterial微生物	微生物
AnimalHumanl動物	動物類
humanl人	人物類
plantl植物	植物類
artifactl人工物	人工物
naturall天然物	天然物
factl事情	事件類
mentall精神	精神類
phenomenal現象	現象類
shapel物形	物形類
InstitutePlacel場所	地點類
locationl位置	位置類
attributel屬性	抽象類
quantityl數量	數量類

Appendix B. Exclusion Word List

I. Monosyllabic exclusion words

/之/的/不/與/兩/再/以/了/較/就/次/得/於/已/把/都/太/一/某/最/
/內/均/原/由/被/全/初/及/將/該/總/塊/項/和/二/從/三/凡/尚/前/
/十/極/番/元/件/甚/因/甲/向/才/四/本/若/先/便/五/粒/常/卅/後/
/左/曾/竟/廿/八/支/六/著/首/剛/應/篇/能/七/終/依/位/暫/共/須/
/中/九/時/可/俱/整/謹/宜/邊/往/批/夥/在/唔/年/諸/略/束/特/磅/

II. Polysyllabic exclusion words

/所以/不能/不會/是否/之間/終於/不必/唯一/西方/恐怕/連續/
/必須/不妨/大家/不得/一旦/初步/據說/看來/全面/臨床/無數/
/依法/國立/過度/突然/通常/一同/單一/大力/純粹/大都/當然/
/種種/大概/國有/順便/總是/不再/默默/無不/那麼/黑白/個人/
/四處/自行/恰好/終究/最佳/一心/十分/甚為/私立/一起/可以/
/多元/所有/依然/現成/正好/針對/一般/難怪/等到/到底/應該/
/貿然/獨家/原先/根據/微微/不勝/國產/整整/衷心/好些/安然/
/慈善/為什麼/一下子/一塊兒/非正式/

Appendix C. Three sentence-NVEF trees used in this study

I. Successful sentence-NVEF tree

+林震南舉起手中煙袋 (Lin Cheng-Nan picks up the pipe on his hand.)
+--+N1
+--+--+tool|用具, *addict|嗜好
+--+--+--+煙袋 (pipe)
+--+V1
+--+--+lift|提昇
+--+--+--+舉起 (pick up)

II. Successful sentence-NVEF tree

+我所下的部分結論 (Parts of conclusion I have given.)
+--+N1
+--+--+thought|念頭, $decide|決定
+--+--+--+結論 (conclusion)
+--+V1
+--+--+announce|發表/V
+--+--+--+下 (give)

III. Unsuccessful sentence-NVEF tree that includes one incorrect word sense: "**樹** (put)"

+**樹上掛著黃絲帶** (Yellow silk ribbons are hung on the tree.)

+--+N1

+--+--+tool|**用具**, linear|**線**, *fasten|**拴連**, *decorate|**裝飾**

+--+--+--+**絲帶** (silk ribbon)

+--+V1

+--+--+put|**放置**

+--+--+--+**樹** (put)

+--+N2

+--+--+tool|**用具**, linear|**線**, *fasten|**拴連**, *decorate|**裝飾**

+--+--+--+**絲帶** (silk ribbon)

+--+V2

+--+--+ hang|**懸掛**

+--+--+--+**掛** (hang)

Computational Linguistics and Chinese Language Processing
Vol. 7, No. 1, February 2002, pp. 47-78

一種基於知網的語義排歧模型研究

A Study of Semantic Disambiguation Based on HowNet

楊曉峰*，李堂秋*

Yang Xiaofeng, Li Tangqiu

摘　　要

本文提出了機器翻譯中句法分析的一種語義排歧模型，該模型以《知網》爲主要語義知識源。《知網》是一個以漢語和英語的詞語所代表的概念爲描述物件，以揭示概念與概念之間以及概念所具有的屬性之間的關係爲基本內容的常識知識庫,它爲我們的排歧提供了豐富的語義資訊。排歧模型結合了基於規則及基於統計的方法，應用於分析所產生的中間結構中，從“優選”的角度進行詞義及結構的排歧。

排歧模型首先利用大規模的語料庫獲取義原的同現集合，該語料庫未進行任何的語義標誌，因此獲取過程是無指導的。然後它根據轉換模板構造出義原的語義限制規則。《知網》中的詞語義項由義原組成，義項的語義限制規則可以由其構成義原的語義規則得到。

在語義排歧階段，我們首先確定輸入句的每個實義詞的上下文相關詞集。由於實義詞的語義關係在對當前句子的語法結構確定及各詞語詞義的選擇起著相當重要的作用，我們對一個句子的評價就建立在對該句中實義詞的評價基礎之上。把詞語的當前上下文相關詞集與詞語各義項的限制規則所描述語義特徵資訊進行比較，根據比較的相似度選擇最合適的義項。同時將相似度的最大值作爲該詞語的評價值。中間分析結果中各實義詞的評價分值可以成爲評價此中間結果的依據，以此在多個中間結構中選出最佳的結果。這樣，我們在解決詞義歧義的基礎上同時也解決了結構歧義。

本文所提出的語義排歧模型已在機器翻譯系統中具體地實現。實驗例句的測試表明該排歧模型對解決句法分析中的辭彙歧義、結構歧義是有效的，並且優於傳統的 YES/NOT 的方法。

* 廈門大學計算機系，廈門，361005
Department of Computer Science, Xiamen University, Xiamen ,361005

本文首先提出了排歧模型的主要思想，並簡要介紹了《知網》。然後給出了從語料庫中抽取義原同現資訊及將其轉化成語義限制規則的方法。接著文章詳細介紹了排歧演算法，包括構建上下文相關詞集，義原間、語義規則和上下文詞集間的相似度計算。最後文章給出了模型的試驗實例結果。

關鍵字: 語義排歧、知網、中間語言、相似度、模式匹配、語料庫、語義限制規則、語義環境

Abstract

This thesis presents a description of a semantic disambiguation model applied in the syntax parsing process of the machine translation system.

The model uses *Hownet* as its main semantic resource, which is a common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. It can provide rich semantic information for our disambiguation.

The model makes the word sense and structure disambiguation in the way of "preferring". "preferring" is applied in the results produced by the parsing process. It combines the rule-based method and statistic based method.

First we extract from a large the co-occurrence information of each sense-atom. The corpus is untagged so the extracting process is unguided. We can construct restricted rules from the co-occurrence information according to certain transfer template. The semantic entry of a word in the Hownet is made of sense-atoms, so we can make out the restricted rules for each entry of any word.

During the course of disambiguation, the model constructs the context-related words set for each notational word in the input sentence. The semantic collocation relations between notional words can play a very important role in the syntax structure disambiguation. Our evaluation of some candidates is based on the degree of tightness of match between notional words in the structure. We compare the context-related words set of the word in the current structure with all the restricted rules of the word in the lexicon, and find the best match. Then the entry with the best match is taken as the word's explanation. And the degree of similarity shows how the word in the structure matches with other notional words in it, so it can be taken as the reference of the notional words. Because the discrepancy of different candidate parses of a structure, the same word has different content-related words set, and so will get different scores. We can calculate the best match according to

the score of all the notional words of the sentence. In this way we can solve the most of word sense disambiguation and structural disambiguation at the same time.

The semantic disambiguation model proposed in this thesis has been implemented in MTG system. Our experiment shows that the model is very effective for this purpose. And it is obviously more tolerant and much better than traditional YES or NO clear cut method.

In this thesis we first put forward the general idea of the method and give a brief introduce to *the Hownet Dictionary*. Then we give the methods of extracting co-occurrence information for each sense-atom from the corpus and transferring this information to restricted rules. Then the algorithm of disambiguation is proposed with detail, which includes constructing context-related words set, the calculation of the similarity between atom-senses, and between restricted-rules and the context-related sets. The experiment result given in the end of the paper shows that the method is effective.

Keywords: Word Sense Disambiguation, *Hownet,* InterLigua, Sense Atom, Corpus, Semantic Environment

1. 前言

1.1 文本分析的歧義消解問題

歧義是自然語言中普遍存在的現象。其研究可追溯至古希臘時期的亞裏斯多德，他在《工具論・辯謬篇》中就探討了自然語言的歧義問題。1930 年，恩普森(W.Empson)發表了《歧義的七種類型》(Seven Types of Ambiguity)一書，開始從語言理論的角度研究歧義問題。科艾(J.G.Kooij)於 1971 年發表的《自然語言的歧義》則標誌著自然語言的歧義進入了系統化的研究階段。在現代語言學的發展史上，“歧義問題總是成爲某個新的語言學派崛起時向傳統陣地進擊的突破口”[呂叔湘 1984, 馮志偉 1995]

歧義就是同一形式與不同的意義産生聯繫。在自然語言處理中，歧義是一個不能回避而且也無法回避的問題，它成爲自然語言的自動分析的巨大障礙之一。

漢語的歧義一般可以分爲以下兩種類型[馮志偉 1995, 苑春法 等]：

(1) 詞的多義，又稱辭彙歧義，即同一詞語可能具有多個不同的義項；如“打”一詞在“打字”、“打酒”、“打球”、“打地基”、“打人”中就有不同的意義；

(2) 短語的同形異構，又稱結構歧義，即同種組合卻含有不同的句法功能結構。如“VP+的+是+NP”就是一個有歧義的結構：

“扮演的是一個演員”

這句話可以理解爲“一個演員扮演了劇中某個角色”(“扮演的”是施事)，也可以理解爲“被扮演成一個演員”(“扮演的”的是受事)。

還有“N1+N2+N3”，可以被理解爲((N1+N2)+N3)，也可以理解爲(N1+(N2+N3))。這類的歧義結構在漢語中有很多，它一直是語法學家研究的熱點問題。

在機器翻譯中，辭彙歧義表現爲譯文會有多種的選擇，而結構歧義表現爲句法分析中，一個詞語或片語可能會產生一個以上結構不同的分析結果。

歧義在特定的語義及常識下，並不一定都能夠成立，例如在“打球”中，根據“打”的受事物件我們可以知道“打”只能選擇“Play”的譯文；而在“反對的是戰爭”中，我們也可以知道“戰爭”是反對的受事體而不可能是施事體。消除這類不符合語義知識的歧義過程稱爲歧義的消解，也稱排歧。我們在機器翻譯的句法分析過程中，必須要引入語義的知識才能夠更好地完成歧義的自動消解。

1.2 歧義消解的方法

歧義消除的方法多種多樣，有的方法在辭彙量、詞法結構和句型上對源語言文本進行限制，從而避免大多數的分析及選詞上的歧義；有些學者提出利用語義關聯網進行排歧[Dan Roth 1998]；還有的研究人員嘗試了基於學習的自動消歧方法，如 DAN ROTH 使用 Winnow 學習方法來進行拼寫校正、語料標注[趙鐵軍 等 2000]。面前市場上流行的“雅信 CAT 漢英雙向翻譯系統”，則採用了用戶互動式消歧方法，讓用戶自己來決定結構及譯詞的選擇。

辭彙消歧是結構消歧的基礎。大多數的排歧方法都是以多義詞的詞義排歧爲切入點。短語、句子、篇章都是由最基本的詞語構成，如果一個句法結構中的詞語意義尚且不能確定，整個結構的意義的把握更無從談起了。同時，詞義選擇需要足夠多的語義知識及上下文知識，而這些知識也爲進一步解決結構歧義問題提供了良好的依據。本文的研究也主要是從詞義消歧入手，在詞義消歧的過程中進行結構的消歧。

詞義消歧方法分爲三類：基於 AI 的方法，基於知識的方法，基於語料庫的方法。按詞義消歧的智慧程度又可分爲有指導與無指導的方法[Wilks *et al.* 1998, Philip *et al.*, 董振東 等]。

基於 AI 的方法包括符號主義方法和連接主義方法。如利用神經網路等進行詞義選擇(Collins)。這類的方法在實際上對語言理解並不實用。

基於知識的方法主要包括基於義類詞類詞典和基於規則的方法。前者的代表爲西班牙研究者基於 WordNet 提出的應用概念密度的詞義消歧方法。而後者則在基於轉換的機器翻譯系統中被廣泛地使用，如 Wilks 提出的應用選擇限制來詞義消歧。

基於語料庫的方法分爲基於統計和基於實例的兩種方法。基於統計的方法經常統計詞與詞、詞義與詞義的搭配，利用搭配消除歧義。基於實例的方法是根據輸入句與實例

的相似度計算來選擇最佳的匹配。這類方法中比較成功的系統有新加坡的LEXAS(Hwee)。

以上的方法各有其優缺點。基於知識的方法可以很好地處理確定的，大粒度的知識，語義和語法的知識比較豐富，但這些知識通常由專家組織，因此有很大的主觀性，並且知識的一致性、擴充性、完備性都難以很好地實現；而基於統計的方法可以較好地處理語言中的不確定的、小粒度的知識，靈活性好，但卻難以反映自然語言中具有普遍性的語法規律和語義知識。

1.3本文的主要工作

本文主要研究目標是在機器翻譯的文本分析中如何引入語義知識進行有效的詞義消歧，進一步進行結構消歧。在課題的研究過程中，本文主要在以下幾個方面進行了探討：

1：利用《知網》爲語義知識源，從《知網》中抽取出必要的語義資訊，並將之轉化成爲方便系統實現的表示結構。

2：從基於優選的角度對分析生成的中間語言進行排歧處理。本文先利用大規模的語料庫獲取義原的同現集合，並根據轉換模板構造出義原的初始限制規則，再通過手工的方式對初始規則進行修改與調整，以得到一個較完善的規則集。義項的語義限制規則可以由其構成義原的語義規則得到。排歧演算法將義項的語義規則與義項所在的語義環境進行相似度的計算，並根據計算結果進行義項選擇和結構的語義搭配的評價，從而進行詞義排歧與結構排歧。本文提出了這種排歧方法的詳細演算法與實現步驟。

3：將上述的思想及演算法在機器翻譯系統中具體地實現。

2. 系統結構與基礎知識

2.1 系統結構描述

結構消歧的難度很大，各種各樣的歧義結構還有待語法學家的進一步發現與總結。而辭彙消歧是結構消歧的基礎。短語、句子、篇章都是由最基本的詞語構成，如果一個句法結構中的詞語意義尚且不能確定，整個結構的意義就無法把握，排歧則更無從談起了。同時，詞義選擇需要足夠多的語義知識及上下文知識，而這些知識也爲進一步解決結構歧義問題提供了良好的依據。本文中語義排歧模型的指導思想就是首先解決多義詞的辭彙歧義，然後在此基礎之上進行結構歧義的消解。爲此本文中提出了一種語義排歧的模型，該模型以《知網》爲主要語義資源，以“優選”的方法來實現詞義與結構的消歧。

“優選”運用于分析所生成中間語言中，這種方法把詞語的當前語境與詞語各義項的限制規則所描述語義特徵資訊進行比較，根據比較的相似度選擇最合適的義項。同時將相似度的最大值作爲該詞語的評價值。中間分析結果中各實義詞的評價分值可以成爲

評價此中間結果的依據,以此在多個中間結構中選出最佳的結果(注意本文中提到的“最佳”都是相對於演算法而言的，是在當前演算法下最好的解，但這並不一定總是實際正確的解）。這樣，我們在解決詞義歧義的基礎上同時也解決了結構歧義。“優選”排歧將基於規則與基於統計的排歧方法相結合，排歧中所使用的義項限制規則的獲取方法是從大規模的語料庫中統計出義原的同現集合，再按一定的轉換模板半自動地生成的。使用這種方法可以大大減少手工編制規則的工作量。

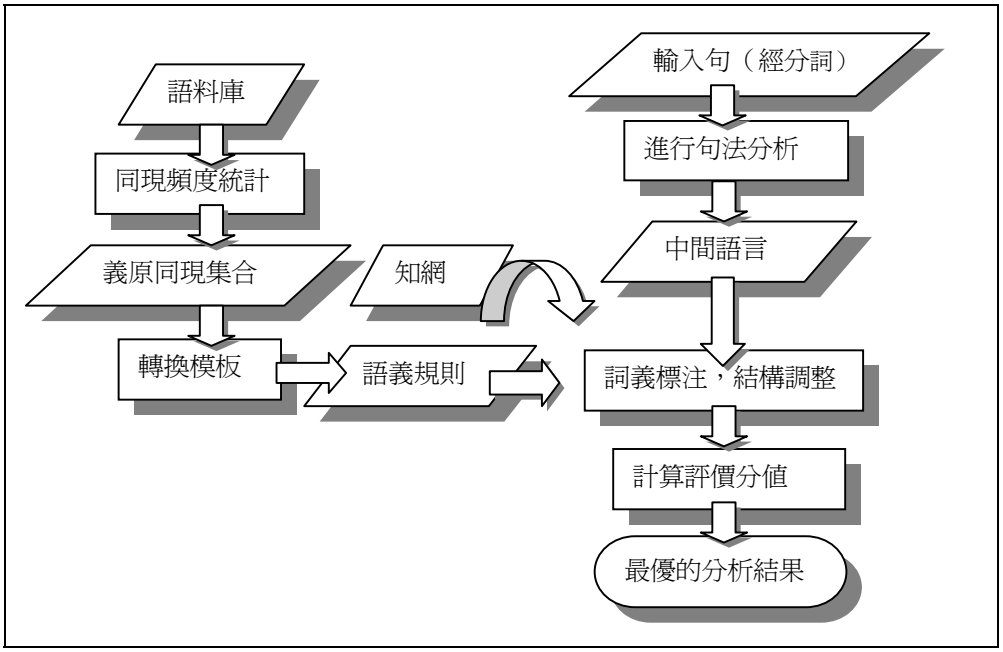


***圖 2.1** 含有詞義排歧模組的 PARSING 流程*

含有詞義排歧模組的漢英機譯系統分析部分的工作流程如圖 2.1 所示。

2.2 《知網》介紹

《知網》(英文名稱 *HowNet*)是其創建人董振東先生花費逾十年研究心血的重要成果。《知網》是一個以漢語和英語的詞語所代表的概念爲描述物件，以揭示概念與概念之間以及概念所具有的屬性之間的關係爲基本內容的常識知識庫，它是一個網狀的有機的知識系統[李涓子 等 1999]。

語義詞典是知網系統的基礎文件。在這個文件中每一個詞語義項的概念及其描述形成一個記錄。目前詞典中提供漢英雙語的記錄，每一種語言的每一個記錄都主要包含 4 項內容。其中每一項都由兩部分組成，中間以“=”分隔。每一個“=”的左側是資料的功能變數名稱，右側是資料的值。它們排列如下：

NO.= 詞或短語序號

[W_X= 詞或短語
G_X= 詞或短語的詞性
E_X= 詞或短語的例子]+
DEF= 概念定義

其中的 W_X、G_X、E_X 構成每種語言的記錄，X 用以描述記錄所代表語種，X 爲 C 則爲漢語，爲 E 則爲英語。每個詞語由 DEF 來描述其概念定義，DEF 的值由若干個義原及它們與主幹詞之間的語義關係描述組成。義原是知網中最基本的、不易於再分割的意義的最小單位，知網通過對約六千個漢字進行考察和分析來抽取了 800 多個義原，並總結了如部分、主體、客體、從屬、時空、材料等若干種義原間的語義關係，這些關係在知網中用義原前附加如“%”、“@”、“$”等相對應符號來表示，因此我們把這些語義關係稱之義原的字首語義關係，而對應的符號爲義原的字首語義關係符。這些符號的意義在《知網》中有詳細的定義說明，表 2.1 列舉了本文出現的一些字首符號及其對應語義關係。具體的符號定義可參看文獻 9。

表 2.1 *義原的部分字首符號及其對應語義關係*

字首符號	對應的語義關係
#	相關
%	部分
$	事件的受事、目標、所有
*	事件的施事、體驗者、工具
+	蘊含
&	從屬
~	可能性
@	時空
?	材料
^	否定

下面我們用形式化的語言來對 DEF 進行定義：

DEF = [Mark]Atom[,[Mark]Atom]*

Mark = * | @ | ? | ! | ~ | # | $ | % | ^ | &

ATOM = $atom_1|atom_2|\ldots|atom_k$

所有這些義原及其關係應能對《知網》中出現的任何詞語概念進行定義。

下例是動詞“打”作“打球”解的義項在詞典中的定義：

W_C=打
G_C=V
E_C=~網球，~牌，~秋千，~太極，球~得很棒
W_E=play

G_E=V
E_E=
DEF=exercise|鍛練,sport|體育
通過 DEF 的定義可以知道在“打球”中“打”和“體育”與“鍛練”有關。

又例如“面”這個詞語包含有如下兩個義項：

W_C=面
G_C=N
W_E=noodles
G_E=N
DEF= food|食品

W_C=面
G_C=N
W_E= face
G_E=N
DEF= part|部件,%AnimalHuman|動物,skin|皮

第一個義項的“面”作“麵條”解，它的義項定義是“fool|食物”；第二個義項的“面”作“臉面”解，它的義項定義說明在這個義項中“面”是動物的部件，是“皮”。

除了語義詞典外，知網還提供了義原分類樹，分類樹把各個義原及它們之間的聯繫以樹的形式組織在一起，父子結點的義原具有上下位的關係。我們可以通過義原分類樹計算義原間的語義距離。在知網中存在 ENTITY、EVENT 等幾棵分類樹，如下圖是詞網中表示事件義原關係的 EVENT 分類樹：

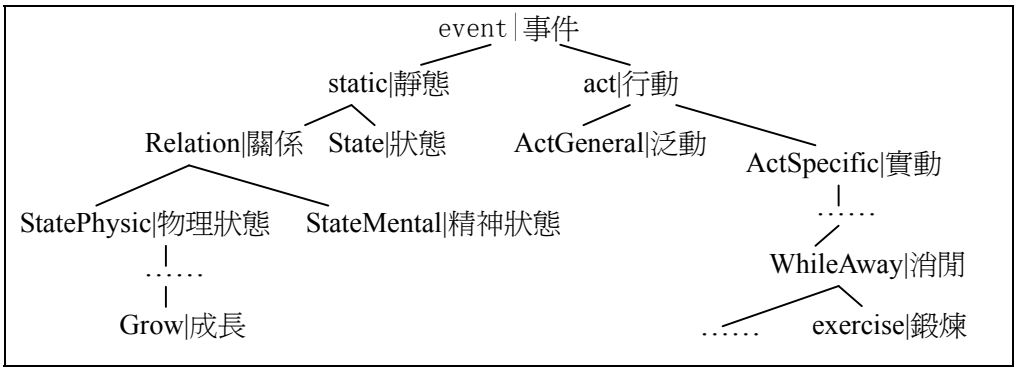


圖 2.2　EVENT 義原分類樹的樹結構表示

知網的詞典用文本的方式保存的，要事先把它轉換成方便系統實現的格式，考慮到系統用 LISP 實現，本文把詞典用表的形式來表示。詞條中除 DEF 外的項都被表示爲以該項爲首位元素的子表，而 DEF 項被拆分成一系列以語義關係爲首位元素的子表，DEF 中所有的具有相同語義關係的義原都歸於相應的子表中。如果義原無字首語義關係，可以都歸爲 PROPERTY 子表。例如我們可以把“面”的義項 2 表示爲：

```
( (W_C  面)(G__C N)(W_E face)(G_E N)
   ( (PROPERTY part|部件 skin|皮)
     (PARTOF AnimalHuman|動物) )
 )
```

3. 基於優選的詞義排歧

3.1“優選”演算法的總體思路

本演算法將基於規則與基於統計的排歧方法相結合。首先利用統計的方法從未標注的大規則語料庫中抽取出義原的同現語義資訊，並根據轉換模板構造出義原的初始限制規則，再通過手工的方式對初始規則進行修改與調整，以得到一個較完善的規則集，繼而獲得詞語義項的語義限制規則。在對一個中間結構進行詞義排歧的過程中，我們按一定的演算法確定各詞語的上下文語境，將該詞語各義項的限制規則和其語境一一進行相似度比較，根據比較的結果確定它及其語境中各詞語的意義。最大的相似度可作該詞語的評價值，而中間結構的評價值就是根據其構成詞語的評價值計算得到的。在最後的結構排歧中，演算法以評價值大小爲優選依據從多個侯選的中間結構中挑出一個最優的結果。這樣我們在解決詞義的消歧的同時也可解決結構的消歧。

3.2 從語料庫中獲取義原的同現集合

語言學家 FIRTH 在對詞義辨識的描述中提出“觀其伴而知其意”(You shall know a word by the company it keeps)，他認爲，詞語的意義只能在上下文中才能得以辨識。如果一個詞語的某一詞義在語料庫中出現多次，我們在其出現的上下文中可以發現某些詞語出現的頻率很高，而這些詞語與該詞義之間有著比較密切的搭配的關係。詞義由義原構成，同樣地，與一個特定義原在上下文中共同出現次數越多的義原也與該義原有越密切的語義關聯，而這些的搭配義原就可以爲我們制定這個義原的語義限制規則提供很好的參考依據。動詞、形容詞的詞義對句子的語義影響最大，而動詞、形容詞的主要構成是動作類義原和屬性值及數量值類義原，因此我們希望得到這兩類義原的同現集合。

義原 a 的同現集合定義爲：

$S(a)=\{(b\ Prob(a,b)) \mid b \in ATOMSET\}$

集合 S(a)中的每個元素都是一個二元組(b Prob)，它的第一個分量是義原 b，第二個分量 Prob(a,b) 表示 a,b 義原的同現概率，即含有分別 a,b 義原的詞語在同一指定上下文窗口中出現的可能性。ATOMSET 是《知網》中所有義原的集合。

“同現”是針對一定詞語範圍而言的，我們把這一範圍稱爲上下文窗口，如果兩個義原所在的詞語在同一上下文窗口中出現，我們稱這兩個義原在當前前窗口中“同現”。上下文窗口的長度選取也是影響到統計結果好壞的重要因素。窗口選取得太小，例如只選擇當前詞語的前後一兩個詞是不夠的，漢語中的許多複雜句式語義相關詞的距離比較長，如

“人口普查的資料對國家制定長遠的經濟規劃具有十分重要的意義”

在這句話核心詞語是“資料”、“具有”、“意義”，它們之間相間的詞語數多達 6 個。如果視窗過小，句子中的關鍵搭配資訊就很可能超出了視窗的範圍而不能被統計。但反之如果窗口選擇太大，窗口內會含有過多與搭配不相關的詞語，如果把它們都進行統計，也會導致統計結果的不準確。按照一般的經驗，本文選取的窗口大小爲前後 7 個詞[梅家駒 1999]。

爲降低統計雜訊，視窗內詞語的選擇也應當考慮詞性、語法等資訊，應儘量選取存在語法關係的辭彙。本文是根據當前詞語的詞性與其他詞性可能組成的各種句法關係來選擇視窗的詞語。動詞及形容詞可以與其他詞性組成的句法關係如表所示[李涓子 等 1999]：

表 3.1 *不同詞性的可能搭配組合表*

詞性	相對位置	詞性組合	詞法關係
動詞	前	名+~	主謂關係
		動+~	並列關係
		名+~	定中關係
	後	~+名	動賓關係
		~+名	定中關係
		~+動	並列關係
		~+形	動補關係
形容詞	前	名+~	主謂關係
		形+~	並列關係
	後	~+名	定中關係
		~+動	並列關係

本文主要從搭配的角度來進行排歧，因此暫不考慮並列的語法關係。這樣從表中可以看到，對動詞來說，前窗口應選擇詞性爲名詞的詞,後視窗應爲詞性集合{名詞,形容詞}中的詞。形容詞的前後窗口都是詞性爲名詞的詞。詞語的前後出現的詞語所起的語法及語義關係差別較大，所以有必要就詞語的前後窗口分別選擇同現詞語，統計出來的義原同現集合也應相應地分成兩部分。我們令 S-(A)、S+(A)分別代表義原 A 的前、後同現集合。

我們的統計語料來源是《讀者 20 年文集》，規模大小約爲 1,100 萬字。語料庫沒有經過任何的語義標注，因此我們無法確定語料庫中一個句子裏某個詞語的詞義究竟是什么，如果把它的所有的義項都參與同現統計，勢必會使統計結果含有較大的語義噪音。爲此上下文窗口中的詞語應選取無詞義歧義的，即只對單義詞進行處理，而不考慮多義詞。注意這裏的“單義”是相對於詞性而言的，詞語可能具有多個詞性，如果詞語 W 在某個詞性 C 中的意義是唯一的，我們就稱 W 在詞性 C 下是單義的。在漢語的詞語中，單義詞的數量占很大的比重，在《知網》中的總詞語數爲 53332，而單義詞數目爲 47188，

占總數的 88.47%；而從實際的運用來考查，單義詞出現的頻度也不低。經統計，在語料庫中，單義詞的數目占 44.18%。這表明單義詞的搭配資訊具有相當的代表性，利用單義詞的搭配資訊來進行多義詞詞義選擇的思想是完全可行的。

義原的同現集合獲取過程是無指導的。但是我們事先要對統計語料庫進行自動分詞及詞性標性，才能正確地選擇詞語的上下文窗口。

演算法 3.1：義原的同現集合獲取演算法

定義：設 Total 爲單義詞的計算器; 陣列 COUNT-F,COUNT-B 是兩個二維陣列，COUNT-F [a,b]是 b 對於 a 的前同現次數；而 COUNT-B [a,b]是 b 對於 a 的後同現次數。

初始化：Total=0;

for each a∈動作類義原集∪屬性值類義原集∪數量值類義原集,

```
for each b∈AtomSet
   Count-F[a,b]=Count-B[a,b]=0;
```

處理過程：

對於語料庫中出現每個單義詞語 W, Category(W) ∈{v,adj }，執行：

{ Total=Total+1;
確定 W 的義原定義集合爲 Atoms(W)；
確定 W 的前、後窗口 Window-F(W)，Window-B(W)
for each a∈Atoms(W)，處理前同現集合：
{
for each b∈ $\bigcup_{k\in Window-F(W)} Atoms(k)$
COUNT-F[a,b]=COUNT-F[a,b]+1；
}
for each a∈Atoms(W)，處理後同現集合：
{
for each b∈ $\bigcup_{k\in Window-F(W)} Atoms(k)$ 都做
COUNT-B[a,b]=COUNT-B[a,b]+1；
}
}

在對語料庫中的句子處理完畢後，可以計算出每個義原的同現義原集

S-(A)=

{(b freq(a,b)) | freq(a,b)= γ_1*count-F[a,b]/Total, b∈ATOMSSET }

S+(A)=

{(b freq(a,b)) | freq(a,b)= γ_2*count-B[a,b]/Total, b∈ATOMSSET }

其中 γ_1, γ_2 分別是前後同現集合中同現概率的放大係數。太小的概率難以表現同現義原間的頻度差異，演算法將概率值同時乘以一個放大係數，這並不影響它們之間的大小關係，而且適當地放大了相互的差異值。

統計語料庫經自動獲取得到的動作類義原同現集合 703 條，屬性值類及數量值義原同現集合 446 條。下例分別是動作類義原“EAT|吃”和屬性值類義原“HAPPY|福”的前後同現集合：

EAT|吃 ：

前同現集合：

{(HUMAN|人 0.56) (THIRDPERSON|他 0.41) (MALE|男 0.28) (MASS|衆 0.23)
(PLACE|地方 0.22) (BIRD|禽 0.20) (PROPERNAME|專 0.20) …}

後同現集合：

{(MEDICINE|藥物 0.52)(PART|部件 0.38)(ATTRIBUTE|屬性 0.26)
(HUMAN|人 0.20) (DESIRED|良 0.20) (FOOD|食品 0.15) …}

HAPPY|福 ：

前同現集合：

{(HUMAN|人 0.56) (THIRDPERSON|他 0.41) (MALE|男 0.28) (MASS|衆 0.23)
(PLACE|地方 0.22) (BIRD|禽 0.20) (PROPERNAME|專 0.20) …}

後同現集合：

{(MEDICINE|藥物 0.52)(PART|部件 0.38)(ATTRIBUTE|屬性 0.26)
(HUMAN|人 0.20) (DESIRED|良 0.20) (FOOD|食品 0.15) …}

通過對同現集合的分析可以發現，集合中有些義原的意義很接近，語義相似度比較小。我們希望能把同現義原集進行聚義處理，即語義接近的義原合併成一個義原，使同現集合的長度得以減少。下面我們給出聚類的演算法：

演算法 3.2 同現集合聚義演算法

設 S 爲義原 Atom 經統計得到同現集合；

初始化新集合 L=Φ;

將 S 中的義原按所在分類樹的層數高低排序，使上位義原總是位於其下位義原之前；

WHILE(S ≠ Φ)

{

從 S 中 POP 出首位元素 A =(a $Prob_a$);

T = 0;

for each B∈S,B=(b $Prob_b$)

{ if 義原 a,b 的相似度 r 小於閾值

{ T=T+$Prob_b$*r;

將 B 從 S 中移出;

}

}

將元素(a T)PUSH 至集合 L;

}

將 L 中的義原按同現概率的數值從大到小進行排序，截取集合前 10%位元的義原；

用集合 L 代替原同現集合 S。

演算法先將同現集合中的義原按其在分類樹的層數排序，這樣可以保證在一個語義相似的義原集群中，義原總是向著最上位的義原，即最抽象的概念進行聚義。由於同現概率高的義原對語境的描述起著更爲重要的影響，因此演算法在聚義完畢後，又將同現集合按義原同現概率的大小排序，捨棄了小同現概率的義原，只保留序列前部 10%的元素。經過演算法 3.2 的處理，新的同現集合的長度將比原來的有顯著減少。

3.3 產生語義限制規則

3.3.1 義原限制規則的定義

利用用演算法 3.1 可以得到每個動作類及屬性值、數量值類義原的上下文同現集合。實際上到這一步我們已經可以根據同現集合與測試句中詞語的上下文語境進行相似度的計算，將相似度最高的義項作爲該詞的在當前句子中的詞義。這種方法對於簡單的句式結構有比較高的正確率，但對於具有特殊的語法性質的詞語則排歧結果不是很理想。如對動詞來說[楊曉峰 2001]，帶複雜賓語(如小句賓語和兼語賓語)的多義詞的詞義排歧結果會差於帶簡單賓語的多義詞。這是由於在複雜語式裏中心詞語與搭配詞語距離較遠，搭配詞語或是超出中心詞語的上下文窗口範圍，或是與中心詞語之間有過多的干擾詞語。

在機器翻譯的語法分析階段成生了源語句的中間語言，它可以更爲準確地描述詞語所在上下文的語境；同時我們也爲義原定義語義限制規則，它描述了含有該義原的詞語的期望出現的語義環境。這樣我們可以根據詞語實際所處的語義環境與義原規則中描述的語義環境進行相似度的計算，將比較結果作爲詞義排歧的依據。

本文中採取的中間語言模型在第二章中已給出了詳細的說明。如測試句：

“我怕你把筆尖給弄斷了。”

經分析的中間語言框架爲：

```
(   (CROOT 怕) (CAT V)
   (AGENT ((HUMAN +) (CAT PRON)(AGREE SG) (PERSON FIRST) (CROOT 我)))
   (THEME (    (CROOT 弄)
               (AGENT ( (HUMAN +)(CAT PRON)
                        (AGREE SG) (PERSON SECOND) (CROOT 你)))
               (THEME ((CAT N)  (CROOT 筆尖)))
               (RESULT ((CROOT 斷) (CAT V)))
          ))
)
```

在格結構中動詞或形容詞的語義環境是由 Agent、Theme、Result、Clause 等格資訊來描述。由於缺少語義資訊，在分析階段給出的語義格資訊並不一定正確，我們可以在詞義排歧階段進行自動的語義格調整。

具體的義原的語義限制規則的模型表示爲：

Rule = (SenseAtom Rule-Items)

Rule-Items=(Logic-Op {(Rule-Items)}+) | {(Sem-Case Logic-Item)}+

Logic-Item = (Logic-Op {Logic-Item}+) | { Sense-Item}+

Sem-Case = Agent | Theme | CO-THEME | Result | Clause | · · ·

Logic-Op = *AND* | *OR* | *NOT*

Sense-Item=*NOT* | (SenseAtom Prob) | (Relation {(SenseAtom Prob)}+)

SenseAtom= EAT|吃 | HAPPY|福 | · · ·

從上面的形式化描述中我們可以看出義原的語義規則與第三章的模式規則有相似之處。規則是用表的形式來表示的，表的首位元素指明規則所屬的義原。表的第二項即表的規則體。規則體由一套或多套子規則體構成。每一套子規則又是由一系列格的限制描述組成。格限制描述說明了含有當前義原的詞語在實際句子的語義環境中希望出現特徵資訊，包括義原及其同現概率，這些義原可能會帶上語義關係成分如 Partof，Material 等。格限制描述這些特徵資訊的邏輯語義關係。這些邏輯關係用*AND*、*OR*、*NOT*等符號來表示。*AND*表示在在實際格中出現的詞語義項應同時滿足指定的特徵資訊；*OR*表示只要能滿足特徵資訊的任意一項即可；而*NOT*則規定詞語不應出現任何指定的特徵資訊。邏輯運算不僅可以作用在格描述的特徵資訊上，也可以作用於不同套子規則中。在這一級上的邏輯算符一般使用*OR*，表示規則可以選用給出的若干套子規則中任意一套。帶有邏輯運算的限定規則可以表述各種複雜的語義資訊。

3.3.2 義原語義限制規則的生成

爲了將 3.2 中得到的義原同現集合的資訊充分運用到義原的限制規則定義中，本文採用了以下的轉換模板，它可根據同現集合構造出一個初始的義原語義規則：

（1）對於屬性值、數量值類義原，前同現集合中實體類的義原作爲“EXPERIENCE”格。
（2）對於屬性值、數量值類義原，後同現集合中實體類的義原作爲“THEME”格；
（3）從動作類義原的前後同現集合中取出“implement|器具”及其所有的下位義原，將它們作爲“INSTRUMENT”格；
（4）從動作類義原的前後同現集合中取出“earth|大地”、“place|地方”、“space|空間”及其所有的下位義原，將它們作爲“LOCATION”格；
（5）從動作類義原的前後同現集合中取出“time|時間”及其所有的下位義原，將它們作爲“TIME”格；
（6）從動作類義原的前後同現集合中取出“degree|程度”、“range|幅度”、“frequency|頻率”及其所有的下位義原，將它們分別作爲“DEGREE”、“RANGE”、“FREQUENCEY”格；
（7）從動作類義原的後同現集合中取出的屬性值、數量值類義原，將它們作爲“RESULT”格；
（8）將動作類義原的前同現集合餘下的實體類義原作爲“AGENT”格；
（9）將動作類義原的後同現集合中的餘下的實體類義原作爲 THEME 格；

例如對於動作類義原“eat|吃”，利用轉化模板得到的義原語義限制規則爲：

```
(eat|吃
 (AGENT (*OR* (HUMAN|人  0.56) (THIRDPERSON|他  0.41) (MALE|男  0.28)
              (MASS|衆  0.23) (PLACE|地方  0.22) (BIRD|禽  0.20)
              (PROPERNAME|專  0.20) … )
      )
 (THEME (*OR* (MEDICINE|藥物  0.52)(PART|部件  0.38) (HUMAN|人  0.20)
              (FOOD|食品  0.15) … )
 (RESULT (*OR* ((ATTRIBUTE|屬性  0.26)  (DESIRED|良  0.20) )
)
```

轉換模板爲我們構造了一個初始規則庫，規則庫中定義了義原 Agent、Theme、Result、Instrument 等格的限制描述。對於語法性質簡單的義原，這些格描述已經足夠。但是對於大多數義原而言，這些自動生成的規則就過於簡單了，因此我們需要在初始規則的基礎上手工對其進行修改與調整，加上必要的格描述、剔除錯誤的特徵義原。例如，對於如“URGE|促使”，由該義原定義的詞語如“促使”、“推動”、“鼓勵”等詞語，一般都帶有兼語，因此我們要爲其增加兼語表示的格(EVENT)的限制描述；還有如由“GIVE|給”定義的詞語，一般都帶有雙賓語，我們也要爲“GIVE|給”定義 CO-THEME(間接賓語)格的限制。另處我們也需爲“EXPECT|期望”等義原規則定義 CLAUSE 格描述。

在修改義原的語義限制規則時，本文從語料庫中爲每個義原選取一定數量的例句，

這些例句中都含有由該義原定義的詞語。我們根據參考例句對初始規則進行修改完善。由於義原的語義規則規模有限，並且事先有一個自動生成的初始規則集，因此手工制定與修改規則所花費的工作量並不大。通過在初始規則的基礎上進行人工調整的方法，我們可以得到一個較爲完善的語義規則集。

3.3.3 義項語義限制規則的確定

在爲動作類義原及屬性值、數量值類義原定義限制規則後，依據《知網》中對義項的一些規定，我們可以自動地生成動詞或形容詞的任意一個義項的語義限制規則。

對於動作類動詞，DEF 項的第一位置只能是事件類規定的主要特徵；因此可以直接將義項第一位置上的義原的限制規則作爲該義項的限制規則。

打 1：buy|買,commercial|商
打 2：exercise|鍛練,sport|體育
吃 1：eat|吃
吃 2：destroy|消滅,military|軍

在上面幾個動詞義項，我們分別將“buy|買”、“exercise|鍛練”、“eat|吃”、“destroy|消滅”的限定規則作爲打 1，打 2，吃 1，吃 2 的限定規則。

而對於形容詞，它們的義項主要由屬性值類義原及數量類義原構成。“屬性值”是所有屬於屬性值概念的唯一的主要特徵，“數量值”是所有屬於數量值概念的唯一的主要特徵，它們分別是形容詞的各義項的首位標識；屬性值類義原和數量值類義原除首位標識外必須還包含有一個次要特徵。在第二位元上一定要標注該屬性值或數量值所指向的屬性或數量特徵；而通常絕大多數情況下在第三位置上標注該屬性值或數量值的具體值，而這些具體值正是我們所感興趣的。有時在 DEF 第三位置後還有一些輔助特徵，它們只是進一步對關鍵義原進行補充說明，對義項的語義影響很小，因此我們在定義形容詞的限制規則中，只需考慮第三位的義原。例如下面是幾個形容詞的義項定義：

巨大 1：DEF=aValue|屬性值,size|尺寸,big|大

巨大 2：DEF=QValue|數量值,amount|多少,many|多

香 1：DEF=aValue|屬性值,circumstances|境況,flourishing|興,desired|良

香 2 ：DEF=aValue|屬性值,odor|氣味,fragrant|香,desired|良

在上面的例子中，我們將選擇“big|大”、“many|多”、“flourishing|興”、“fragrant|香”的義原限制規則作爲對應形容詞義項的語義限制規則。

對於量詞我們也可以自動生成限制規則，規則中規定量詞修飾的詞的語義特徵。在《知網》中，名量詞的定義裏用“&”標注其指向的屬性或事物的類型；例如：

本： DEF=NounUnit|名量,&publications|書刊
輛： DEF=NounUnit|名量,&LandVehicle|車

於是我們就可以直接把“&”標注的義原作爲義項的限定規則。如上例中，義項“本 1”與“輛 1”的規則分別爲

本 1：(THEME publications|書刊)
輛 1：(THEME LandVehicle|車)

與此同時，我們還可以爲某些語義格定義通用的規則，以處理中間結構中不能根據搭配關係進行排歧的詞語。這些語義格通常含有較明顯的語義特徵，如 LOCATION 格中一般含有地點等資訊，而 TIME 格中一般含有時間等資訊，因此，我們可以爲這兩個格定義通用的限定規則如下：

LOCATION：(*OR* PLACE 地方)
TIME： (*OR* 時間)

3.4 詞義排歧演算法的詳細描述

上一節介紹了如何獲得詞語義項的語義限制規則。本節將討論如何運用義項的語義規則在給定的中間結構中進行詞義排歧。

3.4.1 確定詞語的語義環境

本文第二章給出了機譯系統中間語言的表示方法，爲了方便地獲得某一實義詞的上下文相關詞，我們先將中間語言轉放成爲依存關係樹的形式。樹的根結點是句子的核心詞，其他受核心詞支配的附屬成分就作爲根結點的子樹，這些子樹分別也是以各附屬成分爲根結點而建立起來的依存關係樹。例如輸入句“維修/圖書館/的/空調”的兩個可能的句法結構樹如下：

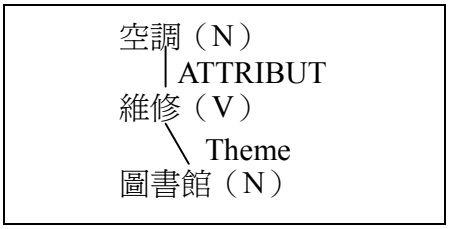


圖 3.1 ((維修 圖書館)的 空調)的樹結構

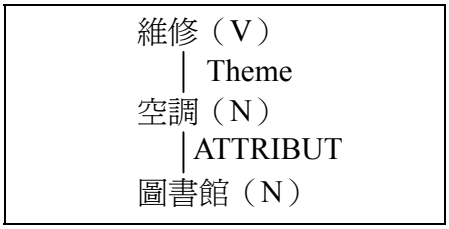


圖 3.2 (維修(圖書館的 空調))的樹結構

在確定詞語的上下文語境前我們定義詞語的限定關係：設有詞語 A，B，如果在句子中 A 修飾、支配 B，則稱在當前句中 A 是 B 的限定詞，B 是 A 的被限定詞。這裏的限定關係與傳統的依存關係有所不同：依存語法認爲發生依存關係的一對詞中，如果詞 A 修飾詞 B，則 B 爲主詞，A 爲從屬詞，A 是 B 的附屬成分，在依存關係樹中體現爲 A 是 B 的子結點；而在我們定義的限定關係中被限定詞則是限定詞的修飾支配物件。如對於主謂賓結構的句子來說，主動詞就是句子的詞語主語、賓語的限定詞；ADJ＋NP 的形

容詞短語中 ADJ 是 NP 的限定詞。如果限定詞與被限定詞之間是偏正的修飾關係，如 ADJ+NP,NP+NP 等，限定詞充當 ATTRIBUTE、MANNER 等附屬成分，則在依存關係樹中體現爲限定詞是被限定詞的子結點；如果它們之間是主謂結構，如 NP+VP+NP(主謂賓)、NP+ADJP 等，則限定詞在樹中充當被限定詞的父結點。

我們之所以要對詞語區分限定詞與非限定詞是由於這兩類詞的評價演算法不相同。根據定義，限定詞在詞語之間的搭配關係中起著主要的修飾支配作用，它的義項選擇及評價分數影響著當前分析結構的總體評價值。在具體的演算法實現上，限定詞的評價是通過各個義項的實例集與它的被限定詞語集的比較獲得的；而非限定詞的評價則要在其限定詞義項確定之後進行，選取與這個義項的實例的最高比較值作爲評價依據。對一棵關係樹進行評價就是建立在對句中充當限定詞的詞語結點的評價上的[詹衛東 1996]。

對於一個可充當限定詞的詞語，我們定義它的上下文語境，也稱爲上下文窗口，是所有被其限定的詞語並上所具有的格資訊。給定句子的一個句法結構，通過該結構對應的依存樹，我們可以很方便地得到各詞的上下文語境。下面是詞語在當前依存樹中的語境選取原則：

(1) 如果詞語是動詞，則選取其子結點的 AGENT、THEME、MANNER、TENSE、TOOLS、LOCATION、TIME 等附屬成分；
(2) 如果詞語爲其父結點的 PROPOSTION(從句)、ATTRIBUTE 等附屬成分，則需並上父結點並將其格注爲 Parent；
(3) 如果動詞的語義格爲 EVENT，這時動詞做兼語句裏的兼語，此時動詞所在兼語句中的賓語應充當其 AGENT 格；
(4) 如果動詞的語義格爲 THEME，這時動詞做賓語或小句賓結構的賓語從句。如果該動詞的 AGENT 語義格不存在，則加入父結點語境中的 AGENT 格。如果該動詞是被動語態，要將 AGENT 格改爲 THEME 格；
(5) 如果動詞的語義格爲 SUBEVENT，這時動詞做連動句，則加入父結點的主語詞及其所充當的語義格。如果該動詞是被動語態，要將 AGENT 格改爲 THEME 格；
(6) 如果詞語是形容詞且爲 ATTRIBUTE 格，則選取其父結點並將其語法格注爲 THEME；否則如果形容詞爲 PREDICATE 格，即形容詞作表語，則選取兄弟結點中的 Experiencer 格；
(7) 如果詞語是量詞且爲 Quatity 格，則選取其父結點並將其語法格注爲 THEME。

3.4.2 詞語的評價值計算

在 3.3.3 中介紹了動詞、形容詞及量詞的義項的語義限定規則獲取方法。上一節描述如何得到一棵中間語言的結構依存樹中各支配詞的上下文語境。現在可以根據詞語所在語義環境及其各義項的限定規則中所描述的語義環境進行相似度的計算。

1)詞語義項與規則的格限制描述的相似度

《知網》中提供了動作類，實體類、屬性值類等義原分類樹，義原間的語義距離與義原

在分類樹對應結點間的最短路徑的邊數有關：最短路徑越長，則表示兩個義原的語義距離越遠，它們的相似程度越小。並且，位於分類樹下部的一對父子結點對應的義原間的語義距離應當小於位於其上的任意一對父子。因此，在計算語義距離時應對分類樹上的邊加權。此外，計算中還應體現出不同類型的義原(如實體類與動作類)之間語義的不可比性。根據上述分析，義原間語義距離和語義相似度的定義如下：

設義原 A、B，它們分別位於分類樹的第 a 和第 b 層(規定根結點爲第 0 層)，它們的最近共同祖先(可以是 A 或 B 本身)位於第 c 層，則 A 與 B 的語義距離爲：

$$\text{DISTANCE-ATOM}(A,B)=\begin{cases}\text{MAXVALUE} & \text{;如果义原A,B不在同一棵分类树上}\\ (\sum_{i=c+1}^{a}Weight(i)+\sum_{j=c+1}^{b}Weight(j))/2 & \text{;否则}\end{cases} \quad \textbf{(3.1)}$$

其中權值函數應是一個單調遞減的函數。設當前分類樹的樹高爲 Depth，我們定義的權值函數如下：

$$Weight(i) = 2*(Depth - i) / (Depth*(Depth+1)) \quad \textbf{(3.2)}$$

而義原 A，B 的語義相似度爲：

$$\text{SIM-ATOMS}(A,B)=\begin{cases}0 & \text{;如果义原A,B不在同一棵分类树}\\ (1-\text{DISTANCE-ATOM}(A,B))*100 & \text{;否则}\end{cases} \quad \textbf{(3.3)}$$

其中 MAXDISTANCE 是義原 A 所在分類樹的最長的語義距離。

以上定義的語義距離與語義相似度中的 A，B 是可交換的，即 A 與 B 的語義相似度等於 B 與 A 的語義相似度。但是我們將會看到，SIM-ATOMS 中比較的是模式義原 A 與實際義原 B 之間的相似度。如果實際義原 B 是模式義原 A 的下位，則它們的語義距離應比較小，如果規則義原在實際義原的下位或是它們只是擁有某個相同祖先的兩個結點，則語義距離應較大。這樣我們應對語義距離函數修改如下：

$$\text{DISTANCE-ATOM}(A,B)=\begin{cases}\text{MAXVALUE} & \text{;如果义原A,B不在同一棵分类树上}\\ (m*\sum_{i=c+1}^{a}Weight(i)+\sum_{j=c+1}^{b}Weight(j))/2 & \text{;否则}\end{cases} \quad \textbf{(3.1')}$$

在實際計算中，我們可以讓 m 的值取得足夠大，使得當義原 B 是義原 A 的子孫結點時語義距離較小，否則將得到一個較大的語義語離。根據實驗，M 值設定爲 5 能夠取得比較好的排歧效果。

2)詞語義項與規則的格限制描述的相似度

義項的語義限制規則中定義了某一語義格可能出現的特徵義原的邏輯組合，而詞語的義項是由義原及語義關係構成的。我們希望能能夠判定一個義項滿足規則的格限制描述的

程度，即義項與規則的格描述的相似度。

設某義項語義規則中規定了某格的限制描述爲 C，
$C=(OP\ (R_1\ S_1)\ (R_2\ S_2)\ldots(R_m\ S_m)\ ,CR_1,CR_2,CR_z)$。
其中對於 $1\le i\le m$,$R'_i \in RelationSet$，S_i 是描述中字首語義關係爲 R'_i 的特徵義原。描述中可能含有無字首語義關係的義原，它們本身就是格描述的特徵語義屬性，爲了保持計算的統一性，不妨假定這些義原的字首語義關係爲 Property。OP 是邏輯組合運算符，包括 *AND*,*OR*,*NOT*，它們的功能在前面已說明過。$CR_i(1\le i\le z)$是帶有邏輯組合運算符的嵌套語義描述集。

現有一詞語義項 $Entry=((R'_1\ E_1)\ (R'_2\ E_2)\ldots(R'_n\ E_n))$ 對於 $1\le i\le n$ 有 $E_i \subset AtomSets$，$R_i \in RelationSet$。同理，對於義項中無字首語義關係的義原，我們也假定其字首語義關係爲 Property。

設(R S)是 C 中的一個元素，其中 $S=(\ (atom_1\ prob_1)\ (Atom_2\ Prob_2)\ldots(Atom_m\ Prob_m)\ )$，定義函數

$$SIM\text{-}ENTRY\text{-}RELATIONITEM((R\ ,S)\ ,Entry)=$$
$$\begin{cases} \underset{(atom\ prob)\in S, a\in E'_i}{MAX} SIM\text{-}ATOMS(atom,a)\times prob & ;\text{如果存在}i,1\le i\le n,\text{有}R'_i=R \\ 0 & ;\text{否則} \end{cases} \quad (3.4)$$

要得到 C 與 Entry 的相似度，可以將 OP 的運算元依次與 Entry 進行相似度的計算，並根據 OP 的值從計算結果中挑選出合適的結果。注意對於 $CR_i(1\le i\le z)$，需要遞迴地進行計算。

我們定義集合 RS 是 OP 的每一個運算元與 Entry 比較的相似度集合，即

$$RS=\{SIM\text{-}ENTRY\text{-}RELATIONITEM(\ (R_i\ ,S_i),\ Entry)\mid 1\le i\le m\}$$
$$\cup\{\ SIM\text{-}Entry\text{-}SC(Entry\ ,\ CR_j)\ \mid\ 1\leqslant i\leqslant z\ \} \quad (3.5)$$

如果 OP 爲*AND*，根據 3.3.1 中的定義，只要 RS 集合中有一個元素的比較值較小，返回的值就應該小，此時應從 RS 中選取一個最小值；如果 OP 爲*OR*，則表示集合可取任何一比較值，這樣此時應從 RS 中選取一個最大值；如果 OP 爲*NOT*，表示集合中只有一個元素，並且如果該元素值比較大，返回的值反而應該小，反之元素的值比較小的話，返回的值應該大。因此，我們可以用 1 與元素值的差來計算比較相似度。

綜上所述，可以定義 C 與 Entry 的相似度爲

$$SIM-ENTRY-SC(Entry,CaseRule)=\begin{cases} \mathbf{MAX}\ RS & ;\text{当OP= }\mathbf{*OR*}\text{ 时} \\ \mathbf{MIN}\ \ RS & ;\text{当OP= }\mathbf{*AND*}\text{ 时} \\ \mathbf{1-MAX}\ RS & ;\text{当OP= }\mathbf{*NOT*}\text{ 时} \end{cases} \quad (3.6)$$

3)詞語的評價演算法

演算法 3.3：詞語的評價演算法

輸入：詞語 Word 及 Word 的上下文語境 ENV，其中

ENV＝((CASE'$_1$ WORD$_1$ CAT$_1$)(CASE'$_2$ WORD$_2$ CAT$_2$)…(CASE'$_n$ WORD$_n$ CAT$_n$))

輸出：Word 及 Wordi(1≤i≤n)在 ENV 中的最佳義項及評價值，它們以如下表的形式輸出：

((Word Mark BestEntry)(word$_1$ Mark$_1$ BestEntry$_1$)…(word$_n$ Mark$_n$ BestEntry$_n$))

定義：EnvBestMark,EnvBestEntry 分別記錄 word$_i$ 在 ENV 中的當前評價值及最佳義項；

WordBestMark,WordBestEntry 分別記錄 word 的當前評價值及最佳義項；

二維陣列 Best-Mark-Entry[word$_i$,Entry]用以記錄 word$_i$ 對於 Word 的 Entry 義項的評價值及最佳義項；

函數 GET-ENV-CASE-WORD(ENV，CASE)用以返回 ENV 中格爲 CASE 的項；

步驟：for each Entry1 in Entries-of-Word(Word)

```
{ 令累計總分數 T=0；
  確定 Entry1 的語義限制規則集合 RuleSet;
  for each R∈RuleSet
   { 設 R=((CASE1  CASESC1)(CASE2 CASESC2) …(CASEm CASESCm))；
     for each (CASE CASESC) in R
     { 設 GET-ENV-CASE-WORD(ENV，CASE)的返回值爲(CASE Wordi CATi)；
       EnvBestMark=MINIMUM;
      for each Entry2 of Entries-of-Word(Wordi)
        if  Entry2 的詞性爲 CATi  then
            { curMark=SIM-ENTRY-SC(Entry2,CASESC);
              if curMark>EnvBestMark then
               {  EnvBestMark=curMark;
                  EnvBestEntry=Entry2;
               }
            }
        Best-Mark-Entry[wordi,Entry1] = (wordi EnvBestMark EnvBestEntry);
        T = T + EnvBestMark;
     }
   T=T / |R|;
   if  T> WordBestMark then
      WordBestMark=T; WordBestEntry=Entry1;
}}
```

令 Result = {(Word WordBestMark WordBestEntry)}

U {Best-Mark-Entry[w,WordBestEntry] | (CASE w CAT) ∈ENV}

返回 Result;

演算法首先確定 word 各義項的語義規則集，注意義項的語義規則可能有多套。將

各規則中格的限制描述與 word 的語義環境 ENV 中各詞語的義項進行相似度的計算，這些詞語相對於 word 當前義項的最大相似度值及對應義項記錄於一個二維陣列中。接著演算法根據 ENV 中各詞語的最大相似度計算 word 當前義項的評價值。最後，演算法把 word 的具有最大評價值的義項作爲 word 的最佳義項 BestEntry，這一最大評價值作爲 word 的評價值。同時，演算法從二維陣列中取出 ENV 中各詞語相對於 BestEntry 的最大相似度及義項，並將其作爲該詞語在 word 的語境中的評價值及最佳義項。

3.4.3 詞語的義項選擇

在演算法 3.3 給出了詞語的評價值計算方法。該演算法可以得到中間結構中各限定詞詞的評價值及最佳義項，同時還可以確定受該詞語限定的各詞語在該詞語的語境中的評價值及最佳義項。本節中介紹如何在中間語言結構中進行詞義的選擇，即詞義排歧。

本文按照詞性的語法特性來以一定的先後順序對不同詞語進行義項選擇。一般來說，動詞的搭配關係對結構的語義語義影響最大，其次是形容詞。根據這一原則，我們提出了以下詞義選擇的步驟演算法：

1：創建一個 RESULT 哈希表，它的入口是結構依存樹中各結點對應的詞語，每個入口項的值就是入口詞語的侯選義項集合及結點的評價分值。如果詞語在分析階段已經計算出義項集及評價分數，就將它們設爲相應的項值，否則初始的候選義項爲詞語的所有義項，而詞語評價分值爲 MINIUM。

2：按至底向上的順序對結構依存樹中的動詞結點進行詞義選擇，將選擇結果集合(可能含有一個或多個的元素)及評價分值填入 RESULT 哈希表的相應項中。對受該動詞限定的所有詞語，我們可以得到在該動詞下的評價分值及最優義項，將這個評價分值與 RESULT 表中對應項的評價值進行比較，如果更大，則將此最優義項集合與分值替代原來的內容。

3：與第 2 步類似，對樹中所有形容詞及其限定詞進行詞義選擇並更改相應的 RESULT 表項。

4：與第 2 步類似，對樹中所有量詞及其限定詞進行詞義選擇並更改相應的 RESULT 表項。

5：對依存樹中還未進行過詞義排歧的結點，如果其所在語義格有通用的限制規則，則根據通用規則進行詞義選擇，更改相應的 RESULT 表項。

6：對於依存樹的每個詞語結點，到 RESULT 表查出對應項的義項集合，將這些義項的所有英文對譯詞作爲分析結構的詞語譯文。

3.4.4 詞義排歧演算法示例

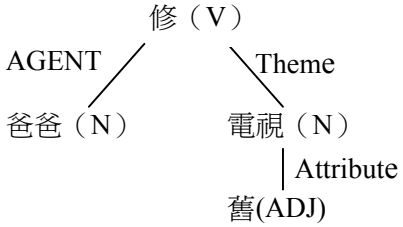
下面我們來看看如何運用本章給出的演算法進行詞義的排歧，設有輸入句

“爸爸正在修那台舊電視呢。”

分析後得到的中間結果爲：

```
((CAT V) (CROOT 修)
 (AGENT  ((CAT N) (CROOT 爸爸)))
 (THEME ((CAT N) (CROOT 電視)
          (ATTRIBUTE  ( (CAT ADJ) (CROOT 舊)))
         )))
```

(1) 將中間結果表示爲依存關係樹的形式：



(2) 確定動詞的上下文語義環境：

```
(修 ((AGENT 爸爸 N) (THEME 電視 N)))
(舊 ((THEME 電視 N)))
```

(3) 確定中間語言結構中各語義的義項規則

表 3.2 *中間語言結構中各語義的義項資訊表*

詞語	義項資訊		
	義項定義	義項規則	對應譯文
修	PutInOrder|整理	((AGENT(*OR* (HUMAN|人 0.21)..) (THEME(*OR* (SPACE|空間 0.33) (FACILITIES|設施 0.27) …)))	prune, trim
	Repair|修理	((AGENT(*OR* (HUMAN|人 0.23)..) (THEME(*OR* (PART|部件 0.37) (TOOL|用具 0.26) …)))	mend, repair, overhaul
	Study|學	((AGENT(*OR* (HUMAN|人 0.31)..) (THEME(*OR* (EDUCATION|教育 0.14) (KNOWLEDGE|知識 0.09) …)))	cultivate, study
爸爸	Human|人, family|家, Male|男	NIL	DAD, father, papa
電視	Image|圖像, Shows|表演物	NIL	television
	Tool|用具	NIL	television -set , TV-set
舊	Used|舊	(THEME (*OR* (clothing|衣物 0.29) (TOOL 用具 0.23) …))	old, used, worn
	Past|昔	(THEME (*OR* (TIME|時間 0.29) …)	old, past, bygone
	Original|原	(THEME (*OR* (PHYSICAL|物質 0.21) …)	former, onetime

(4) 對“修”各義項及其語義環境的詞語進行評價及選詞。結果如下：

表 3.3 “修”各義項在語境中的評價分值最對應義項

義項	語義環境中的詞語評分				義項評價分數
	爸爸		電視		
	評價分	最佳義項	評價分	最佳義項	
PutInOrder|整理	100	Human|人	83.92	tool|用具	91.96
Repair|修理	100	Human|人	91.07	tool|用具	95.53
Study|學	94.64	Human|人	0	NIL	47.32

從表中可得到“修”最大義項的評價分數爲 95.53，對應的最佳義項是“Repair|修理”

與該義項對應的“爸爸”在當前環境中的評價分數爲 100,最佳義項爲“Human|人”

“電視”在當前環境中的評價分數爲 91.07,最佳義項爲“tool|用具”

(5) 對“舊”及其語義環境的詞語進行評價及選詞。結果如下：

表 3.4 “修”各義項在語境中的評價分值最對應義項

義項	語義環境中的詞語評分		義項評價分數
	電視		
	評價分	最佳義項	
used|舊	100	tool|用具	100
past|昔	19.64	tool|用具	19.64
original|原	75.26	tool|用具	75.26

從表中可得到“舊”最大義項的評價分數爲 100，對應的最佳義項是“used|舊”

與該義項對應的“電視”在當前環境中的評價分數爲 100,最佳義項爲“tool|用具”

(6) 確定各詞語的詞義：

表 3.5 句子中各詞語的對應譯文

詞語	對應譯文
修	mend, repair,overhaul
爸爸	DAD, father, papa
電視	television-set ,TV-set
舊	old, used, worn

3.5 對中間結果進行語義處理

上一節中我們對詞義的排歧的演算法作了詳細的介紹。在詞義排歧的基礎上，我們可以進一步進行中間語言的結構排歧。爲了提高結構排歧的正確率，在詞義排歧的過程中，應該要同時對中間結果進行語義處理，調整句法分析錯誤的語義框架，或確定中間結構

中某些複雜的語義修飾關係。本節將給出中間結果的語義調整及語義關係確定的處理演算法。

3.5.1 調整分析結果的語義框架

3.4 中介紹了如何利用詞語所處的語義環境中與義項的規則進行相似度比較進行詞義排歧。語義環境由受轄詞及其語義格構成。當時我們假設了中間語言結構給出的是正確的語義格資訊，然而在實際的句法分析過程中，由於缺少語義資訊，中間結構的語義格並一定總是正確的。如下面的例句：

“被子叠得整整齊齊”

“我把被子叠得整整齊齊”

在第一個句子中，形式主語“被子”就是“叠”的受事體和邏輯賓語，因爲根據常識，“被子”是不可能折叠什么其他的東西，它只能是被叠。這類形式主語爲受事體的句子可以看成是“被”字句的省略形式，如這句話可以寫成“被子被叠得整整齊齊”。而第二個句子中“我”是“叠”的形式主語，也是施動者和邏輯主語，因爲一般情況下人不會被叠而可能是叠別的東西。在句法分析時不能獲得這些語義的資訊，因此難以正確地決定施事者與受事體，第一句的分析結果很可能會出現如下的錯誤：

ENV: (叠 (AGENT 被子)(RESUTL 整整齊齊))

在詞義排歧時獲得的語義資訊使我們能夠對結果的語義框架進行適當的調整。

“被子”的義項爲：

被子 1：tool|用具,*cover|遮蓋,#sleep|睡

而“叠”的義項的限制規則爲

叠 1：((AGENT (HUMAN|人 0.426))(THEME (TOOL|用具 0.138)))

它規定了 Agent 人的語義特徵是 HUMAN|人，如果用這一限制描述與“被子”的義項進行比較值會得到一個比較低的相似度值。這啓發我們可以採用如下的語義格測試調整方法：

設詞語 Word 的語義環境爲

$W=((CASE_1\ word_1)(CASE_2\ word_2)\ldots(CASE_i\ word_i)\ldots(CASE_n\ word_n))$

W 的一個義項規則爲 R；

如果詞語含有(CHANGECASE $CASE_i$ $CASE'_i$)的警告標記，同時，W 中存在 $Case_i$ 格且不存在 $Case'_i$ 格，則我們分析 $Case_i$ 的詞語 $Word_i$ 在 W 下的評價分值，如果小於某一預定義的最低閾值，則我們試著將 W 替換爲

W’=((CASE$_1$ word$_1$)(CASE$_2$ word$_2$)…(CASE’$_i$ word$_i$)…(CASE$_n$ word$_n$))

並再次將 W 與義項進行相似度計算，得到了一個新的 Word 的評價分值，如果新的評價分值大於預定義的調整閾值，我們就認爲在分析結構中的詞語 Word$_i$ 的語義格應調整成爲 Case’$_i$。

如處理“被子疊得整整齊齊”一句時，可以首先在語法分析階段的語法詞典中爲“疊”這樣具有主語充當受事體的動詞做上標誌(CHANGECASE AGENT THEME)，預選定義最低閾值與調整閾值分別爲 20 與 85。

在語義排歧時我們可以發現原語義環境 ENV 中詞語“被子”在“疊”下的評價分值很低，我們用新的語義環境

ENV’：(疊 (THEME 被子)(RESUTL 整整齊齊))

重新與“疊 1”的規則進行比較計算，調整後“被子”在“疊”下的評價分值達到了 100 分，高於調整閾值，我們就可以認爲在此結果框架中的 AGENT 應調整爲 THEME。

3.5.2 確定分析結果的語義關係

漢語中存在某些特殊的句型，如“VP+NP”、“VP+的+是+NP”或“VP+的+NP”，或是“NP+VP”、“NP+的+VP” 等偏正結構，其中的 NP 在 VP 的語義環境中所充當的語義關係格不是固定的。我們把 NP 當作 VP 語義環境的 Parent 格詞語。NP 可以作 VP 的任何一個不存在的語義成分。例如：

1：裝修圖書館的工人整整忙了一天。(“工人”做“裝修”的 AGENT)
2：天下哪有白吃的午餐？(“午餐”作“吃”的 THEME)
3：我們以前住的地方現在是片廣場了。(“地方”作“住”的 LOCATION)
4：她的打擾會使他感到厭煩。（“她”作“打擾”的 AGENT）
5：感謝你對藝術事業的大力支持！(“事業”作“支援”的 THEME)

從上面的例子我們可以看到，NP 在 VP 中可以充當的成分是很豐富的[1]。在進行詞義選擇時，我們根據詞語的語義環境與義項限制規則進行相似度計算，需要確定這個 NP 在 VP 語義環境中的語義格。

與 3.5.1 的思路相同，我們也可以採用測試比較法來解決 NP 的成分確定問題。下面主要針對“VP+DE+NP”或“VP+NP”類的定中結構進行討論，對於“NP+VP”或“NP+的+VP”型的結構，也可同樣進行處理。

設 VP+DE+NP 或 VP+NP 結構中 VP 的中心動詞 Word 的語義環境爲

W＝((PARENT PWORD)(CASE$_1$ word$_1$)……(CASE$_n$ word$_n$))

其中 PWORD 是 VP 短語的修飾詞語。

現有 W 的一個義項規則 R：

$R=((CASE'_1 \quad CASESC_1)(CASE'_2\ CASESC_2)\ \ldots(CASE'_m\ CASESC_m))$

我們可以將 W 中的 PARENT 格依次替換爲在 R 中存在而在 W 中不存在的格，當然替換的格必須不違反 W 中的某些對成分要求的語法限制，例如當 W 是不及物動詞時，Parent 就不能替換成爲 THEME 格。Parent 替換後得到的新語義環境 W'與 R 進行相似度的計算。設當 Parent 換成 $Case_i$ 時取得最大的相似度值 V，且 V 大於一個預定義的閾值，則我們可以確定 Parent 在 VP 中充當 $Case_i$ 的語義成分。

如果$\{CASE'_1, Case'_2, \ldots,Case'_n\}-\{Case_1,Case_2,\ldots,Case_m\}=\Phi$，或者 V 小於閾值，則表明被修飾的 NP 在 VP 不能充當合適的語義成分，或語義成分不能確定。這種情況如

1：他中獎的消息立刻傳開了。(“消息”爲“他中獎”的同指)

2：創作方法很重要。 (“方法”實際上是“創作”的“方式”成分，但創作的規則中沒有對“方式”的限制描述)

對於上述這兩種的情況，可以將詞語與通用格規則進行匹配，具有最大匹配相似度所對應的格可作爲被修飾詞在 VP 中所充當的語義成分。

3.6 利用詞義排歧進行結構排歧

一個測試句經過語法分析後有可能產生多個中間結果，這就需要對中間結果進行評價，從中優選出一個最優結果。即進行結構排歧。

漢語的結構歧義錯綜複雜，許多的漢語言文學研究學者都對其進行深入的研究，並總結了許多的歧義短語組合格式。

本文對結構歧義的消除演算法是建立在詞義排歧的基礎之上。最佳的中間結果應是最符合語義與常識的，而中間結果“優選”的原則也應是選擇最滿足語義的結構。在“約束”排歧及本章前面介紹的詞義排歧中，我們對詞語進行義項選擇的同時還得到了詞語的評價分值，它們反映了詞語在當前語境中符合語義的程度。前文中定義了詞語的限定關係，那些起限定作用的詞語在當前結構中起著關鍵的搭配作用，它們對結構的語義具有最直接的影響，因此我們在對一個中間結果進行詞義排歧後，可以將排歧得到的各限定詞的評價分值的總和作爲當前分析結果的評價值。評價值的大小是優選的根據，具有最高評價值的中間分析結構就作爲最終的結構排歧結果。

例如：“vp+np+的+np”結構可能存在兩種歧義：

1： (vp (np 的 np)) 如“修理爸爸的自行車”

2： ((vp np) 的 np) 如“修理自行車的爸爸”

這兩句話可能存在 4 種歧義結構，對它們分別進行詞義排歧，得到的評價情況如下：

表 3.6 各分析結構的評價分數

分析結構	各詞語評價分數			分析結構評價分數
	修理	爸爸	自行車	
((修理 爸爸)的 自行車)	0	0	25.0	0.0
(修理 (爸爸 的 自行車))	89.28	0	89.28	89.28
((修理 自行車) 的 爸爸)	94.64	100	89.28	94.64
(修理(自行車 的 爸爸))	0	0	0	0·0

從表中可以確定(修理 (爸爸 的 自行車))與((修理 自行車)的 爸爸)分別是句 1 與句 2 的最佳分析結果。

3.7 實驗結果及討論

3.7.1 實驗結果

在義原的同現集合獲取中我們使用《讀者 20 年文集》作爲統計語料庫，測試時也使用同類的語料。測試目的是檢驗本文提出的排歧演算法的是否有效，同時通過中間語言的結構優取及調整的正確率來考查詞義評價模型是否合理。我們從語料庫中選取了 2，000 個測試句進行排歧實驗。下表是測試的指標及測試結果：

表 3.7 測試結果

測試指標	指標描述	測試值
詞義排歧的正確率	詞義判斷正確的詞語數/測試語料中歧義詞的總數	0.92
結構排歧的正確率	結構選擇正確的句數/測試語料中有多個候選分析結構的句數	0.82
結構調整的召回率	進行結構調整的結果數/測試語料中需進行調整的分析結果數	0.95
結構調整的正確率	經自動調整後正確的結果數/進行自動結構調整的結果數	0.98

測試結果表明，利用語料庫的同現義原來構造義原的語義限制規則，並以此進行詞義排歧的思想是合理的。並且在詞義排歧的過程中同時也能實現高正確率的結構排歧、結構調整。

3.7.2 演算法存在的問題

(1)：演算法找出的義原規則是具有普遍搭配關係的詞義排歧規則，我們通過義原的限制來構造義項的規則，有可能義原規則限制的粒度比義項應有語義限制大；這就造成了產生的義項規則語義粒度過大。例如象“找”一詞，它在作“找了兩元錢”時的義項的

定義是“return|還”，這一義項中的搭配受事體一般來說是和錢財有關的，但“return|還”的義原規則中的受事體的語義特徵不一定是和錢有關，這就有可能導致如“找書”中的“找”也會誤選爲“Return”。

(2)：《知網》中詞語含有的一些文言用法對干擾了詞義的正確選擇。如“去”一詞，在《知網》中有“leave|離開”的義項解釋，這種用法在現代文中很少出現，而它的存在對會影響“去”的常見義項“go|去”的選擇。

3.8 小結

本演算法具有如下特點：

(1) 在《知網》中詞語的義項由多個義原定義，不象傳統的分類詞典中義項只是一個類代碼，這樣對詞語義項的意義描述更加全面，豐富。

(2) 利用義原的規則與當前詞語所在語義環境進行相似度的比較進行詞義排歧，可以提高複雜句式結構的詞義排歧正確率。

(3) 將排歧知識建立在義原的基礎上，義原的數目是有限的，這樣避免了手工編制大規模詞義排歧知識的繁重勞動。同時義原的排歧知識是參考義原的同現集合，而這一集合是通過對語料庫無指導學習獲取的。這樣知識的獲取的工作量進一步的減少了。

(4) 利用詞語義項的評價演算法，在詞義排歧過程中可以對中間語言的語義框架做適當調整。

(5) 在進行詞義排歧的同時解決多個分析結果的結構排歧。

4. 總結

漢英機器翻譯中要解決分析產生的辭彙歧義、語義歧義，得到一個比較好的句法分析結果，必須要引進語義知識，進行語義分析。本文構造了一個用於機器翻譯文本分析的語義排歧模型，它能夠結合語義知識進行有效的詞義消歧，進一步進行結構消歧。

本文提出的語義排歧系統有以下特點：

1：在句法分析過程中進行的約束排歧大大減少了中間語言的生成數目，減輕了針對中間結果進行的“優選”工作。

2：在“優選”的方法中我們利用了從語料庫中構造義項語義限制規則的方法，減輕了人工制定語義知識的工作量。同時語義規則也避免出現單純利用統計進行語義排歧時對複雜句式的處理效果不理想的現象。

3：排歧過程中給出的義項評價值可以使得系統在完成排歧的同時可以進行別的與

語義有關的處理，如語法格的調整等。

本文所提出的語義排歧模型已在機器翻譯系統中具體地實現。實驗例句的測試表明這一排歧對處理辭彙歧義、結構歧義是有效的。

由於研究的時間關係，本文的排歧模型從功能上來看還只是一個實驗系統，還有不少可以改進的地方。例如：

1：本文主要是在詞歧消歧的基礎上進行結構消歧，因此只考慮了常用的幾種歧義格式，還有很多其他的歧義結構有待總結與處理。同時一些如“吃食堂”，“吃大餐”等具有特殊的語義格轉換歧義的現象，需要進一步的深入研究。

2：本文的語義排歧模型主要針對實義詞進行，對於虛詞詞語歧義問題沒有過多地考慮。

3：本文處理都是單句漢語的歧義排歧，未涉及到篇章級的上下文語境理解。而實際上不少歧義需要放在上下文中才能得以消除。

以上這些問題，都有待在後繼的工作中不斷地加以補充與改進，使用語義排歧模型更加有效、實用。

參考文獻

馮志偉 <論歧義結構的潛在性>《中文資訊學報》，1995，第 9 卷(4)

呂叔湘 <歧義類型>《中國語文》1984 年第 5 期

馮志偉 <歧義消解策略初探>《計算語言學進展與應用》清華大學出版社，1995

苑春法，黃錦輝，李文捷 <基於語義知識的漢語句法結構排歧> 《中文資訊學報》第 13 卷第 1 期

Dan Roth. “Learning to Resolve Natural Language Ambiguities: A Unified Approach.” AAAI-98, 1998

趙鐵軍等《機器翻譯原理》·哈爾濱：哈爾濱工業大學出版社，2000

Wilks, Y. Stevenson, M. “Word Sense Disambiguation Using Optimized Combinations of Knowledge Sources”, In *Proceedings of joint COLING-ACL'98*. 1998. Montreal, Canada.

Philip Resnik ,David Yarowsky. A Perspective on Word Sense Disambiguation Methods and Their Evaluation. *Proceedings of the SIGLEX Workshop* "Tagging Text with Lexical Semantics: What, why and how?", pp. 79-86, Washington, D.C.

董振東，董強·《知網》·http://www.how-net.com

李涓子，黃昌寧 <基於轉換的無指導詞義標注方法>《清華大學學報》（自然科學版），1999 年第 39 卷(7)

梅家駒 《現代漢語搭配辭典》漢語大詞典出版社. 1999 年 12 月第 1 版

李涓子，黃昌寧 ＜一種無指導的詞義排歧模型＞《計算語言學文集》北京：清華大學出版社，1999

楊曉峰，李堂秋，洪青陽 ＜基於實例的漢語句法結構分析歧義消解＞《中文資訊學》報，2001 年第 15 卷

詹衛東 ＜現代漢語 VP 結構定界各結構關係判定＞《北京大學碩士學位論文》．北京大學，1996

Computational Linguistics and Chinese Language Processing
Vol. 7, No. 1, February 2002, pp. 79-90

基於文本概念和 kNN 的跨語種文本過濾

Cross-Language Text Filtering Based on Text Concepts and kNN

蘇偉峰*，李紹滋*，李堂秋*，尤文建*

Weifeng Su, Shaozi Li, Tanqiu Li, Wenjian You

摘　　要

本文介紹一個可以從中文或英文大量的資訊中過濾出用戶的興趣所在的文檔的模型，用一簇可分義原向量空間的向量來表示用戶所感興趣的文本，然後把需要處理的文本也表示成一個可分義原空間中的一個向量，在向量空間中與 k 個最相近的向量進行計算，從而決定是否將該文本呈現給用戶。實驗證明，這是一個比較好的過濾方法。

關鍵字：可分義原、向量空間、kNN、文本表示、知網

Abstract

The WWW is increasingly being used source of information. The volume of information is accessed by users using direct manipulation tools. It is obviously that we'd like to have a tool to keep those texts we want and remove those texts we don't want from so much information flow to us. This paper describes a module that sifts through large number of texts retrieved by the user.

The module is based on HowNet, a knowledge dictionary developed by Mr. Zhendong Dong. In this dictionary, the concept of a word is divided into sememes. In the philosophy of HowNet, all concepts in the world can be expressed by a combination　more than 1500 sememes. Sememe is a very useful concept in settle the problem of synonym which is the most difficult problem in text filtering. We classified the set of sememes into two sets of sememes: classfiable sememes and unclassficable semems. Classfiable sememes includes those sememes that are more

* 廈門大學計算機系
Department of Computer Science, Xiamen University , Xiamen，361005
Weifeng Su: waveletsu@263.net

useful in distinguishing a document's class from other documents. Unclassfiable sememes include those sememes that have similar appearance in all documents. Classfiable includes about 800 sememes. We used these 800 classficable sememes to build Classficable Sememes Vector Space(CSVS).

A text is represented as a vector in the CSVS after the following step:

1. text preprosessing: Judge the language of the text and do some process attribute to its language.
2. Part-of-Speech tagging
3. keywords extraction
4. keyword sense disambiguation based on its environment by calculating its classifiable sememes relevance with it's environment's classifiable sememes. We add the weight of a semantic item if there are classifiable sememes the same as classifiable sememe in the its environment word's semantic item. This is not a strict disambiguation algorithm. We just adjust the weights of those semantic items.
5. Those keywords are reduced to sememes and the weight of all keywords 's all semantic items 's classifiable sememes are calculated to be the weight of its vector feature.

A user provides some texts to express the text he interested in. They are all expressed as vectors in the CSVS. Then those vectors represent the user's preference. The relevance of two texts can be measured by using the cosine angle between the two text's vectors. When a new text comes, it is expressed as a vector in CSVS too. We find its k nearest neighbours in the texts provided by the user in the CSVS . Calculating the relevance of the new text to its k nearest neighbours and if it is bigger than a certain valve, than it means it is of the user's interest if smaller, it means that it is not belong to the user's interesting. The k is determined by calculated every training vector its neighbours.

Information filtering based on classifiable sememes has several advantage:

1. Low dimentional input space. We use 800 sememes instead of 10000 words.
2. Few irrelevant feature after the keyword extraction and unclassifiable sememes's removal.
3. Document vector's feature's weight are big.

We made use of documents from eight different users in our experiments. All these users provides texts both in Chinese and English. We took into account the user's feedback and got a result of about 88 percent of recall and precision. It demonstrates that this is a success method.

Keywords: Classfiable Sememe, Vector Space, kNN, Text Representation, HowNet

1.引言

隨著因特網和其他在線資訊資源的迅猛發展，大量的資訊朝人們湧來，據統計美國每個上班人員平均每天收到 80 封電子郵件，當然裏面包含大量的無用的垃圾郵件，顯然如果對這些郵件逐一查看要花費很多時間，而且有可能激發某些病毒郵件從而破壞電腦系統，同樣的情景出現在許多辦公環境中，許多人希望能在許多的歷史或者別人送達的電子文本當中由電腦自動挑出自己最感興趣的內容。

文本過濾是自動分挑出有用的文本的一種很重要的方法。文本過濾是指從大量的源資訊中過濾出那些最符合用戶需求的資訊傳送給用戶，而跨語種文本過濾是指源資訊中包含多種語言（比如英語、漢語等），或者某個文本中就含有多種語言，從中過濾出用戶所需要的文本，過濾出的文本可能也是多種語言的。在沒有國界的因特網上，跨語種過濾出所要的資訊就顯得更爲重要。在把大量的資訊送給用戶之前過濾掉那些用戶不感興趣的東西，這比在有條件後，翻譯成某種語言過後再進行過濾更能省掉用戶大量的精力和時間，跨語種過濾系統對於那些對需要這一語種的資訊而又對該語言掌握得不好的用戶特別重要。

在跨語種文本過濾方面，人們已經摸索出了許多方法來實現不同語種之間的相互轉換形式。最初人們是提出一種基於控制辭彙的方法[TRANSLIB 1995]，即把文本表示成一些固定的詞，用戶的需求也表示成這些固定辭彙，然後進行匹配。這個方法最大的缺陷是辭彙必須在可管理的範圍之內，而一旦辭彙超出可管理的範圍，則其召回率和精確率則迅速下降，而且如何把文本表示爲辭彙目前也沒有一個很好的方法。

接著又有人提出基於字典的方法[L. Ballesteros 1996]，就是編輯一本多語字典把某種語言的文本表現形式通過翻譯表示成另一種語言的表現形式，從而使那些單種語言上的文本過濾技術可以應用於多語言的文本過濾，這個方法理論上是有可能的，但是有兩個方面的原因卻限制了它的應用。首先是一詞多義的現象，在翻譯中一個詞可能翻譯成幾個意思，若幾個意思全都採用則大大降低了精確率，若採用某一個意思，則有可能降低召回率，或者根本就選擇錯誤而導致召回率極低。第二是一義多詞的現象，由於不同的作者可能用不同的詞來表達同一個意思而導致召回率下降。

本文提出一種新的思路，我們不從詞這一級來分析概念，而是把詞所包含的概念進行分解，再對分解過後的概念進行分析，從而得到文本的主題和性質。其實類似的思想

在一些自然學科當中經常用到，比如我們分析某種物質的性質時，我們經常從其構成的分子或原子水平的性質進行分析然後再得到物質的性質。

2. 過濾模型的系統結構

我們採用的技術主要是向量空間模型，即文本表示成爲向量空間中的一個向量，向量空間的優點是將文本內容轉換成易爲數學處理的向量方式，使得各種相似運算和排序成爲可能。因此 ，在文本檢索、文本過濾和文本摘要等方面獲得廣泛應用，取得了良好效果。本文所提出的基於向量空間的文本過濾模型可以用於對中文和英文的文本進行過濾。其基本思想是首先利用用戶所提供的材料來獲取用戶的模板，然後利用用戶模板來判斷某一文件是否與用戶模板相近。

我們採用了董振東先生所研製的《知網》[董振東 等]，該系統帶有 53000 個中文片語和 57000 英語單詞。《知網》是一個以漢語和英語的詞語所代表的概念爲描述物件，以揭示概念與概念之間以及概念所具有的屬性之間的關係爲基本內容的常識知識庫，《知網》採用義原來表示概念，義原是最基本的、不易於再分割的意義的最小單位，我們設想所有的概念都可以分解成各種各樣的義原。董振東先生提取出了 1500 多個義原，並用它們的組合來表示世上所有的概念，比如它是這樣注釋“扭虧爲盈”的：

DEF=alter|改變,StateIni=InDebt|虧損,StateFin=earn|賺。

即是指“扭虧爲盈”是一種“改變”，其起始狀態是“虧損”，最終狀態是“賺”。

把概念分解成爲義原可以極大限度地解決一義多詞的問題，比如“電腦”、“電腦”、“computer”這三個詞，在《知網》裏均定義爲“computer|電腦”，這樣我們就可以把它們視爲概念等同的三個詞語。其實從這個層面上來理解，我們可以把某個詞的中英文意思同樣看作是一個一義多詞的一種形式，這樣只要解決好了排岐的問題，我們並不需要特殊處理就可以解決跨語種的問題，因而從義原這一個層面上來說我們的方法可以說是一種與語言無關的方法。

我們把義原繼續分爲兩類：可分義原和不可分義原。把義原分爲可分義原與不可分義原是從以下兩方面考慮的：

- 某一義原若是在某一類主題的文檔當中出現頻度越高，則認爲該義原與這一類主題越有關係。
- 某一義原若是在語料庫中所有的文檔當中出現的頻度越高，則認爲該義原在區分主題的作用越差。

因此不可分義原是指那些比較常見，沒法用來指出該概念一些特有的性質的義原，而可分義原則是指那些能表示該概念的重要的可與別的概念相區分的義原。若我們不排除掉不可分義原，則由於不可分義原的較高的出現頻率，就有可能誤導我們。可分義原在本系統中占著重要的地位。

可分義原和不可分義原的粗略判定方法如下：

從語料庫中隨機抽取 500 篇各類的文章，在對這些文章進行分詞之後，把詞分解成相應的義原，並對這些義原進行統計，若某個詞有岐義，則把該詞的所有意義進行歸一併將其義原加入統計，設定一高通篩檢程式，對於義原的統計值高於某個值的列爲不可分義原，其餘的義原定爲可分義原。

本文下面所用到的技術對於中文文檔和英文文檔同時適用，若是有不同的地方則會分別指出。

3. 過濾模型的設計和實現

3.1 文本表示方法

我們採用的技術是向量空間模型，文本表示爲向量空間中的一個向量。向量空間表示爲 $\vec{D}$ ，而每一個分量 d_i 是知網中的一個可分義原，那文本就表示成向量 $\vec{V}$ ，其分量 v_i 爲對應於 d_i 的值，若文本中沒有包含 d_i ，則 v_i =0。

然而並非文件當中所有的詞都用於構造文本向量，只有那些最能代表文件所要表達的意思的詞也就是關鍵字彙可被用來構造向量。我們可以採用統計的方法來決定哪些辭彙是關鍵字彙，還有，由於辭彙的岐義，我們也要作一定程度上的排岐。文本表示方法可歸納如下：

文本預處理。對於中文文本包括詞的切分、詞性標注，而對於英文文本，則只進行詞性標注。

關鍵字提取。在英語文本中去除所有屬於下列的單詞：冠詞（如 a, the, an）、介詞或連接主句和從句的副詞（如 in ,to ,of）、情態動詞（如 would, must）和連接詞（如 and）等，在中文文本中去除所有的虛詞，這樣在文本中就剩下主要的詞像名詞、動詞、形容詞和副詞，形成關鍵字序。我們也可以給各種詞性的詞賦予不同的權值來表示它們不同的重要性，一般而言，名詞要賦以最大的權值。對於那些在標題、首段、末段、段首、段尾出現的詞語也可以增加其權重。我們也可以設一個閥值，把那些出現頻率低於該頻率的詞去除。

關鍵字概念排岐。過多的岐義會損害我們向量表示該文本的效果，尤其當某個詞在該文本當中佔有比較重要的地位時。排岐的基本思想是根據上下文詞的義原對該詞爲某一意思進行概率統計。其主要思想是：在一篇文章當中，某個詞會對上下文的用詞產生影響，通過上下文可以判定某個詞的意思從而進行排岐，在本模型下，著重考慮其上下文當中其他的關鍵字的義原與該詞的義原有無以下情況：

a.有相同可分義原，

b.材料-成品關係，

c.施事/經驗者/關係主體-事件關係，

d.受事/內容/領屬物等-事件關係，

e.工具-事件關係，

f.場所-事件關係，

g.時間-事件關係，

h.事件-角色關係，

i. 相關關係。

如果其上下文的某個關鍵字當中有個可分義原與該詞的某個意思的某一可分義原有上述關係，則增加該意思的權重。

在 **W** 中，對某個詞 *w*，在以其爲中心的窗口寬度爲 n 的字串表示爲：

$$w_1 w_2 \ldots w_{n/2} w w_{n/2+1} \ldots w_{n-1}$$

對於 *w* 在知網中的每一個意思，賦予一個初權 k，調節詞 *w* 每一個意思的權值的方法的僞代碼演算法 1 所示

演算法 1 ：詞的義原的權值的調節

W_I—窗口中除去 *w* 的第 I 個詞

S_{IJ}—窗口中除去 *w* 的第 I 詞的第 J 個意思

CS_{IJK}--窗口中除去 *w* 的第 I 詞的第 J 個意思的第 K 個可分義原

WS_J—詞 *w* 的第 J 個意思

WCS_{JK}--詞 *w* 的第 J 個意思的第 K 個可分義原

Weight(WCS_J)—詞 *w* 的第 J 個意思的權值

```
FOR I=1 TO n-1     //對於窗口中除了w 外的每一個詞
   FOR J=1 TO    (WI的意思數目)
       FOR K=1 TO    ( SIJ的可分義原數目)
          FOR M=1 TO    (詞 w 的意思數目)
              FOR O=1 TO    (WSJ的可分義原數目)
              IF CSIJK  與  WCSJK  有上述關係 THEN Weight(WSJ)= Weight(WSJ)
                                                    +1
                 ENDIF
             ENDFOR
          ENDFOR
       ENDFOR
   ENDFOR
ENDFOR
```

由此，詞語的那些與上下文相關的意思都通過增加權值而得到加強，當然我們還要對此進行歸一化處理，其歸一化的公式如下所示：

$$wt(WS_i) = \frac{Weight(WS_i)}{\sum_i Weight(WS_i)} \qquad (1)$$

其中 i 是該詞的意思的序號。

文本表示成一向量。在經過了關鍵字提取和排岐之後，我們把這些關鍵詞根據其義原權值按照知網裏的單詞定義分解成爲義原，並在去除了不可分義原之後，我們採用演算法 2 中的方法計算各可分義原，文件就表示成了可分義原空間中的一個向量。

演算法 2 把一個文件表示成可分義原空間的一個向量演算法

V_K— 向量中的分量的值

SM_{IJK}—第 I 個關鍵字第 J 個意思的第 K 個可分義原

Weightof(SM)—某個可分義原的標量值

$wt(S_{IJ})$－第 I 個關鍵字第 J 個意思的權值

給向量的每個分量值賦初值 0

```
FOR I:=1 TO   (關鍵字的數目)
  FOR J=1 TO   (第 I 個關鍵字的意思總數)
    FOR K=1 TO (第 I 個關鍵字第 J 個意思)
      Weightof(SMK)= Weightof(SMK)+wt(SIJ)
    ENDFOR
  ENDFOR
ENDFOR
```

3.2 用戶模板表示

首先用戶提供 m 篇其所感興趣的文檔，爲了增加用戶興趣的文本在向量空間中的密度，一般要求 m>50，採用上文所述的方法把這些文本表示爲可分義原空間中的向量，這些向量就成了代表該用戶興趣的示例，我們稱其爲用戶示例。在進行文本過濾時，我們就是從用戶示例中找出 k 個與正在過濾的文本最爲鄰近的向量作爲鄰居向量進行分析。

3.3 文本相似度的計算

至此，文本已表示成可分向量空間中的一個向量，兩個文本的相似度可以通過公式（1）中的余弦值表示，其值越大，則表示這兩個文本的主題越相似，我們認爲他們是越相近鄰居：

$$\cos(a) = \frac{(V_{user}, V_{text})}{|V_{user}||V_{text}|} \qquad (2)$$

其中 (V_{text1}, V_{text2}) 是指用戶向量和文本向量的內積，$|V_{text}|$ 表示文本向量的標量。

在文本過濾當中，我們採用了 k 個最近鄰居（kNN）的方法：對於某一輸入文本 s，按照上面所述的方法將其表示爲可分義原空間的向量，在用戶示例中，利用公式（2）挑選出 k（k<<m）個與之最相近的鄰居文本，根據公式（3）計算它與這 k 個文本的相似程度 *Si*，其值越高，則我們 r 認爲它越是用戶所感興趣的文本。

$$Si = \sum_{i=1}^{k} S^2(\cos(a_i)) \tag{3}$$

其中

$$S(x) = \begin{cases} 0 & \text{当 } x<h \text{ 时} \\ x & \text{当 } x>=h \text{ 时} \end{cases}$$

在所需過濾的所有文本當中，我們可以根據 *Si* 來進行相關度排序反饋給用戶，也可以設一閥值 *t*，當某文本與用戶需求的相關度大於 *t* 時則認爲該文本符合用戶需求，把文本按相關度大小的順序返回給用戶，把低於該值的所有文本去除或存在某處以備用戶在有空時處理。我們可以把用戶的回饋考慮進去，若用戶認爲幾乎所有我們所過濾出的文件都是他所感興趣的，則我們可調低 *t* 值，反過來，若有很多文本不符合用戶的興趣，則我們調高 *t* 值。

3.4 文本類別的歸類

我們採用 kNN 的方法。首先我們訓練的時候，我們把這些已經分好類的按是否爲用戶的需要全部按上述方法表示成可分義原向量空間的向量，對一新進來的一個新的文本，我們採用上面的方法轉化爲可分義原向量空間中的空間向量，假設爲 d，從中找出 k 個與其最爲鄰近的向量，然後檢查這 k 個已經確定好類別的向量的類別作爲這個向量的類別。這 k 個向量的權重可以通過其與 d 的相近程度進行賦值。

kNN 是一個基於範例的學習法，其主要的計算量是從向量空間中找出 k 個最近的鄰居時間複雜度爲 ***O(L*N)***，其中 ***L*** 是可分向量空間的可分義原數目，***N*** 爲可分向量空間中的訓練文本的數量。

k 值的確定方法：

我們主要採用登山法來確認 k 值，在訓練文本全部表示成向量空間的向量以後，按下面演算法進行計算：

演算法 3 kNN 中的 k 的計算演算法

```
biggestequal:=0
bigestk：=0；
給向量的每個分量值賦初值 0
FOR k:=（一個>1 的小整數）TO  (一個大整數)
  km:=0;
  FOR I=1 TO  (訓練文本的數目)
      對於第 I 個訓練文本，計算 k 個最近鄰居，並利用 k 個鄰居的類別判定
      第 I 個文本的類別，如果相等，則 km:=km+1;
  ENDFOR
```

```
If km>biggestequal then
Begin
    biggestequal:=km;
    bigestk:=k;
end;
ENDFOR
```

4. 過濾模型的實驗結果及實驗分析

我們獲得了八個用戶的實驗資料，這八個用戶都提供了他所感興趣的內容相近的中英文文本各 60 篇作爲相關文本，另外提供 1000 篇其他內容的文本作爲干擾文本，其中中英文各 500 篇，對於每個用戶，我們使用從其所提供的相關文本隨機抽取中英文文本各 30 篇構造其用戶模板，其餘的相關文本與干擾文本混雜一起構成了測試集，我們就想從其中過濾出那些相關文本。

我們使用了兩個參數來評價我們的模型：召回率和精確率。召回率是指我們過濾出的相關文本占所有相關文本的比率，精確率是指在我們所有過濾出的文本當中，相關文本所占的比率，一般而言，召回率上升，則精確率會下降，而精確率上升，則召回率會下降。

表 1 就是我們實驗的結果，結果表明用該方法進行過濾的方法效果非常好，精確率很高，在實際應用當中，我們還可以把用戶反饋的情況考慮進去，形成可根據用戶的興趣改變而把改變用戶模板向量從而改變選擇的文本的自適應系統。

		User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	Average
召回率(%)	English	88.7	90	90	89	86	87	92	91	89.2
	Chinese	86.6	91.5	86	85	84	87	90.6	90	87.6
精確率(%)	English	86	88.6	85	88.7	87.5	88.5	84.7	90	88.5
	Chinese	82	85.4	85	87.6	84.2	86.3	88.6	86.8	87.5

***表 1** 使用該方法的八個用戶的召回率和精確率*

我們可以從以下幾方面來分析這個過濾模型產生較好結果的原因：

1. **低維分析空間**：所有的概念都被分解成義原，只須在可分義原空間中計算相似程度，這樣我們就只要計算 600 個左右的可分義原而不是 100000 個左右的中英文

單詞，如此降低維數可極大地提高召回率，還有，可以降低計算複雜度。

2. **相關分量值較大**：比如在一篇病人上醫院去看病的文本裏，可能會會出現許多類似“病人”、“醫生”、“醫院”、“治療”等片語，這些詞都包含有“醫治”等義原，從而使“醫治”這個義原分量的值比較大，這樣就能突出本文的所要講述的內容主要是關於醫療這一方面的，這有助於提高精確率與召有率。
3. **干擾項較少**：經過了關鍵字提取、詞語排岐和不可分義原的去除後，所剩下的義原大多與文本有重要的聯繫，而與文本相關度較少的其他分量的值相比之下明顯較小。

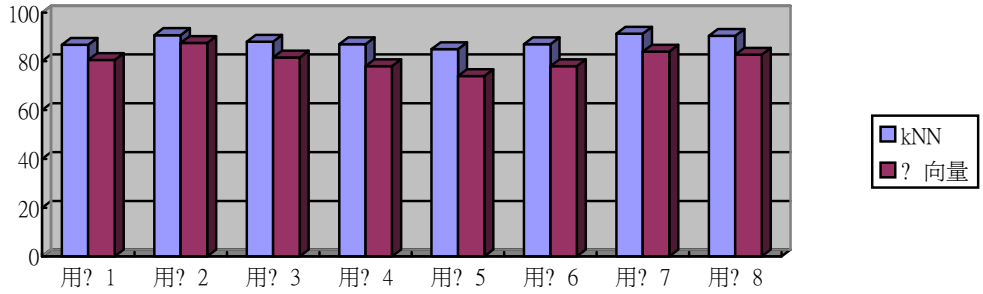
在我們以前的工作當中，我們把用戶表示成爲**一個**向量，並以用戶向量與文本向量的夾角來表示文本與用戶的相關性，而採用了 kNN 技術，可在以下這些方面體現出其優勢：

1. 首先對於某一個用戶可能有比較廣泛的興趣，則取其平均向量可能會導致比較大的誤差。
2. 對於同一個領域，不同體裁的文章其在向量空間當中也可能有較大的差距，取平均向量也會造成較大的誤差。
3. 如果用戶興趣產生變化，平均向量的改變較爲遲緩，並且在這個過程當中也有較大的誤差。

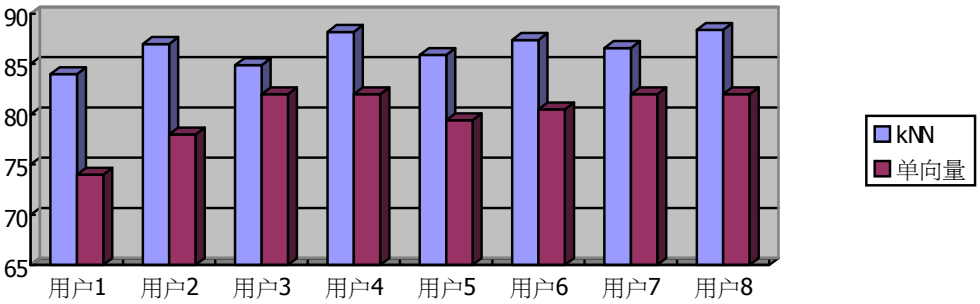
而 kNN 則恰恰相反，

1. 若用戶有比較廣泛的興趣，則在向量空間當中形成不同簇的向量，就可有不同的鄰居。
2. 對於同一領域而不同體裁的文章，也可在向量空間中形成不同簇的向量，構成不同的鄰居。
3. 若用戶興趣發生變化，只要再次提供新的所興趣的文本，在向量空間當中幾乎不受舊的向量的影響，且可保留舊的向量以備另用。

其優勢可在圖 1 和圖 2 體現出來。



图*圖 1 kNN 和單向量來表示用戶需求的召回率的比較*



图*圖 2 kNN 和單向量來表示用戶需求的精確率的比較*

5. 結束語

從網路資訊服務需求出發，我們認爲有必要對資訊源的資訊進行過濾。本文提出了一個在可分義原空間中採用向量空間模型的方法進行文本過濾的模型，理論和實驗均表明，該模型具有比較好的過濾效果，從速度和服務性能上達到了較好程度。

在模型的實現過程中，我們發現把這種方法與關鍵字的方法相結合在相當程度上會提高過濾的性能，這將是我們下一步研究的目標。

參考文獻

TRANSLIB. “Advanced Tools for Accessing Multilingual Library Catalogues.” *Technical Report*, Deleveralbe D.1.4:Evaluation of Tools.Knowledge S.A., June 1995.

L.Ballesteros,W.B. Croft. “Dictionary-based methods for cross-lingual information retieval.” *Proc. Of the 7 th Int. DEXA Conference on Database and Expert Systems Applications* ,1996.

董振東、董強 《知網》 http://www.keenage.com/html/index.html

Douglas W.Oard, Gary Marchionini, “A Conceptual Framework for Text Filtering.” http://citeseer.nj.nec.com

張月傑、姚天順 ＜基於特徵相關性的漢語文本自動分類模型的研究＞《小型微型電腦系統》，1998 年第 8 期

A.T.Armapatzis and Th.P. van der Weide and C.H.A.Koster and P.van Bommel. “Texts Filtering using Linguistically-Motivated Indexing Terms.” http://citeseer.nj.nec.com

Anandeep S.Pannu and Katia Sycara. “A Learning Personal Agent for Texts Filtering and Notification.” http://citeseer.nj.nec.com

James Allen, *Natural Laguage Understanding.* The Benjamin/Cumming Publishing Company, Inc.

Thorsten Joachims. “Texts categorization with support vector machines: Learning with many relevant features.” http://citeseer.nj.nec.com

Douglas W.Oard and Nicholas DeClaris. “On Automatic Filtering of Multilingual.” http://citeseer.nj.nec.com

Ellen Riloff and Wendy Lehnert. “Information extraction asbasis forhigh-precision textclassification.” *ACM Trams-actions on Information System*, vol. 12, No 3, July 1994

Eui-Hong(Sam)Han , Geoge Karypis and Vipin Kumar. Text Cateorization Using Weight Adjusted k-Nearest Neighbor Classification. http://citeseer.nj.nec.com

蘇偉峰、李紹滋、李堂秋、尤文建 ＜可分義原向量空間中的跨語種文本過濾模型＞《自然語言理解與機器翻譯》2001 年