廣義知網詞彙意見極性的預測

Predicting the Semantic Orientation of Terms in E-HowNet

李政儒 Cheng-Ru Li, 游基鑫 Chi-Hsin Yu, 陳信希 Hsin-Hsi Chen

國立台灣大學資訊工程系

Department of Computer Science and Information Engineering

National Taiwan University

#1, Sec.4, Roosevelt Road, Taipei, 10617 Taiwan

crlee@nlg.csie.ntu.edu.tw, jsyu@nlg.csie.ntu.edu.tw, hhchen@ntu.edu.tw

摘要

詞彙的意見極性是句子及文件層次意見分析的重要基礎，雖然目前已經存在一些人工標記的中文情緒字典，但如何自動標記詞彙的意見極性，仍是一個重要的工作。這篇論文的目的是爲廣義知網的詞彙自動標記意見極性。我們運用監督式機器學習的方法，抽取不同來源的各種有用特徵並加以整合，來預測詞彙的意見極性。實驗結果顯示，廣義知網詞彙意見極性預測的準確率可到達 92.33%，這個結果跟人的標記準確率不相上下。

Abstract

The semantic orientation of terms is fundamental for sentiment analysis in sentence and document levels. Although some Chinese sentiment dictionaries are available, how to predict the orientation of terms automatically is still important. In this paper, we predict the semantic orientation of terms of E-HowNet. We extract many useful features from different sources to represent a Chinese term in E-HowNet, and use a supervised machine learning algorithm to predict its orientation. Our experimental result showed that the proposed approach can achieve 92.33% accuracy, which is comparable to the accuracy of human taggers.

關鍵詞：廣義知網，情緒分析，情緒字典, 語義傾向, 向量支援機

Keywords: E-NowNet, Sentiment Analysis, Sentiment dictionary, Semantic orientation, SVM

一、緒論

情緒分析（Sentiment Analysis）在現今的網路世界中，有許多實際且重要的運用，例如從網路的評論文章中分析消費者對產品的評價，或分析消費者對產品性能的關注焦點等等。不管對句子或文件層次的情緒分析，意見詞詞典都是一個重要的資源。通常意見詞詞典是用人工來收集詞彙，並用人工標記詞彙的各種情緒屬性，包括主客觀（subjective or objective）、極性（orientation/polarity)、）及極性的強度（strength）[1]。這些情緒屬性對不同的應用有不同的重要性，標記難度也各不相同，通常詞彙的極性是最容易進行標記的屬性。

標記情緒屬性時，研究者可以從零開始收集詞彙以建立意見詞詞典，如台大意見詞詞典 NTUSD[2]。在另一方面，也有研究者嘗試爲自然語言處理中的許多現存的資源，添加情緒屬性，如 SentiWordNet[3]。但現有資源的語彙量通常很大，如 WordNet 3.0 就包括 206,941 個不同的英文字義（word-sense pair），要全部用人工進行標記之成本太高。因此，通常的作法是少量標記一些詞彙，再用機器學習方法，爲剩下的詞彙進行自動標記，雖然自動標記的準確率不如人工標記，但對一般應用有某種程度的幫助。

在中文自然語言處理，NTUSD 是一部重要的意見詞詞典，但此詞典只包括詞彙及極性的資訊。另一方面，董振東先生和陳克健教授所建立的知網[4]和廣義知網[5]，是重要的語意資源。對於每個詞彙，都用有限的義原給予精確的定義，但這些定義卻缺乏情緒的語意標記。因此，如何自動爲廣義知網加上情緒標記，成爲一個重要的課題，也是本研究的目的。

本研究提出爲廣義知網加上情緒標記的方法，首先利用 NTUSD 跟廣義知網詞彙的交集建立標準答案集，再由標準答案集訓練出分類器，爲其他廣義知網詞彙進行標記。如何有效的運用監督式機器學習演算法，如何爲詞彙抽取出有用的特徵，是主要的挑戰議題。在此研究中，我們有系統的嘗試抽取各種不同的詞彙特徵，最後得到跟人工標記準確率不相上下的分類器。

第二節介紹廣義知網、及英文和中文相關的情緒屬性標記研究，第三節介紹從 E-HowNet 及 Google Chinese Web 5-gram 抽取特徵的方法，第四節呈現各種實驗的結果及分析，包括跟 NTUSD 人工標記的比較，最後總結論文的成果。

二、相關研究

董振東先生於 1998 年創建知網（HowNet），並在 2003 年，跟中央研究院資訊所詞庫小組在 2003 年，將中研院詞庫小組詞典（CKIP Chinese Lexical Knowledge Base）的詞條跟知網連結，並作了一些修改，最後形成廣義知網（Extended-HowNet, E-HowNet）。詞庫小組修改並擴展知網原先的語義義原角色知識本體，建構出廣義知網知識本體（Extended-HowNet Ontology），並用這些新的語義義原，以結構化的形式來定義詞條，詞條定義式的例子如圖一。

有關情緒屬性標記的研究，我們分爲英文及中文來討論。在英文方面，最早是由 Hatzivassiloglou & McKeown[6]在 1997 年針對形容詞所做的研究，他們所用的形容詞分別有正面詞 657 個及負面詞 679 個，該論文依據不同的實驗設定，監督式機器學習的

準確率（Accuracy）由 82% 到 90%。之後陸續有不同的研究，所用多爲半監督式機器學習的演算法[7-9]，效能從 67%到 88%不等，但因爲這些演算法所用的資料集並不相同，實驗過程及評估標準也不一樣，（有用 Accuracy、Precision、或 F-Measure），所以效能沒有辦法直接比較。

```
<Word item = "汽油">
   <WordFreq>15</WordFreq>
   <WordSense id="1">
      <English>gasoline</English> <Phone>ㄑㄧˋ ㄧㄡˊ</Phone> <PinYin>qi4 you2</PinYin>
      <SyntacticFunction> <POS>Naa</POS> <Freq>15</Freq> </SyntacticFunction>
      <TopLevelDefinition>{material|材料:attribute={StateLiquid|液態},telic={burn|焚燒}}</TopLevelDefinition>
      <BottomLevelExpansion>{material|材料:attribute={StateLiquid|液態},telic={burn|焚燒}}</BottomLevelExpansion>
   </WordSense></Word>
```

圖一、「*汽油*」的廣義知網定義式

在中文的情緒屬性標記相關研究，Yuen et al.[10]2004 年利用 Turney & Littman[7] 的半監督式機器學習演算法，在正面詞 604 個及負面詞 645 個的資料集上做實驗，得到最高的成績是 80.23%的精確度及 85.03%的召回率。之後從 2006 到 2010 年，陸續的研究使用不同的資料集，用不同類型的機器學習演算法來處理這個問題[11-14]，所得到的效能在不同的指標（Accuracy、Precision、或 F-Measure）下，從 89%到 96%不等。因爲基準不同，這些效能一樣沒有辦法直接比較，但相較於英文，成績則明顯提高。

三、特徵抽取及機器學習演算法

由於我們運用監督式機器學習演算法來訓練分類器，最重要的問題是爲詞彙抽取出有用的特徵。在此論文中，我們分別從 E-HowNet 及 Google Chinese Web 5-gram 這兩個來源抽取兩大類的特徵，接著將這兩個來源的特徵組合訓練分類器。此外，我們也嘗試使用組合式的監督式機器學習演算法（ensemble approach），來更進一步得到更高的效能，以下我們分別詳細介紹。

（一）、基礎義原特徵

從 E-HowNet 抽取的特徵稱之爲基礎義原特徵，也就是對每一個 E-HowNet 的詞彙 i，用一向量 $V_i = (w_{i,j}) = (w_{i,1}, w_{i,2}, \ldots, w_{i,n})$ 表示，其中 n 爲向量的維度。

由於每一詞彙的每一個語意（sense）都有一個結構化的定義式，而且定義式中都用義原來進行定義，公式 (1) 定義 V_i 中每個特徵的權重：

$$w_{i,j} = \begin{cases} 1, & \text{如果定義式 i 中出現義原 j} \\ 0, & \text{不出現義原 j} \end{cases} \qquad (1)$$

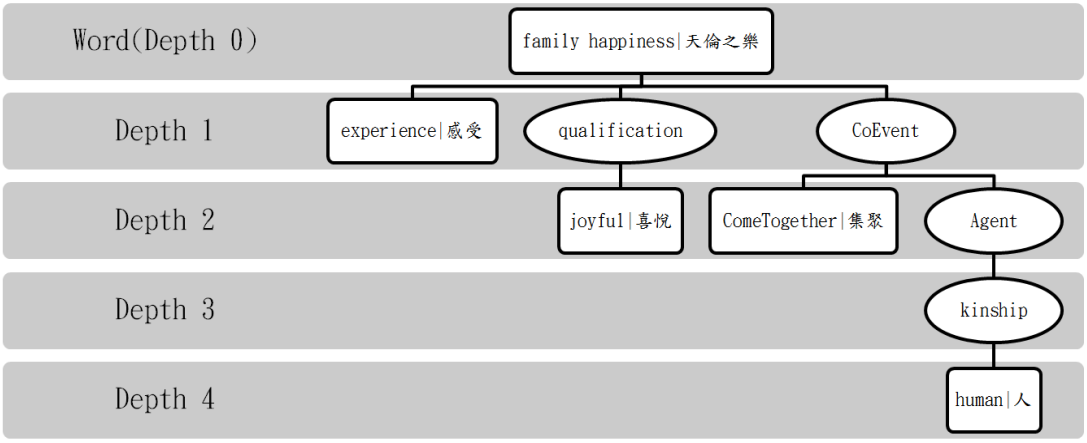
以圖一「*汽油*」這個詞彙爲例，其定義式中出現了義原 material，所以它的值 $w_{\text{汽油, material}}$

就會是 1，其他沒出現的義原，值就會是 0。我們共使用了 2567 個義原來當特徵。

廣義知網的詞彙有歧異性，也就是每個詞彙可能有許多語意。而詞彙的第一個語意，是出現頻率最高的語意（除了四個詞彙例外），所以我們用詞彙的第一個語意來抽取特徵。只從詞彙的一個語意抽取特徵，而不把該詞彙所有的語意放在一起，代表這種方法可爲不同的語意給出不同的極性預測。只是由於目前 NTUSD 極性標記只到詞彙的層級，所以無法對語意的層級進行極性預測。但只要有語意層級的極性標記，我們這種做法可馬上套用。

1、基礎義原特徵加權值

除了公式 (1) 的方式外，我們可以利用更多 E-HowNet 的特性，來抽取出有用的資訊。一個可能的方式是定義式中的結構，如果把定義式展開，會得到如圖二的樹狀結構。在這樹狀結構中，義原所在的深度是一個有用的資訊，因此我們仿照劉群&李素建[15]的公式，將深度的資訊當作權重引入公式 (1)，得到公式 (2)。



圖二、「*天倫之樂*」定義式的樹狀表示

$$w_{i,j} = \begin{cases} \dfrac{1}{1+\alpha \times d_{i,j}}, & \text{如果定義式 i 中出現義原 j} \\ 0, & \text{不出現義原 j} \end{cases} \qquad (2)$$

公式 (2) 中，α 是可調的參數，$d_{i,j}$ 是詞彙 i 跟義原 j 的距離，這可用義原 j 的深度表示。調整公式 (2) 中的 α ，讓我們可以實驗那一種方式，才應給較高的權重：

（可能一）$\alpha < 0$：深度越深，表示該義原有較多資訊，應給較高權重。

（可能二）$\alpha > 0$：深度越深，表示該義原有較少資訊，應給較少權重。

由於 $\alpha < 0$ 時，$w_{i,j}$ 可能變爲負值，所以最小的 α 設爲 -0.05。另外，當 $\alpha = 0$，公式 (2) 會等於公式 (1)，所以我們在做實驗時，只要使用公式 (2) 即可。

2、加入否定關係調整特徵的加權值

在計算義原深度時，可能會經過帶有否定意義的關係，例如「*一事無成*」定義式中有

「{*not*({*succeed*|*成功*})}」，可以發現 *succeed* 被 *not* 所修飾。這時，義原 *succeed* 的權重用負值來表示可能會更好，因此我們將否定的概念引入公式 (3) 如下：

$$w_{i,j} = \begin{cases} \dfrac{Neg_{i,j}}{1+\alpha \times d_{i,j}}, & \text{如果定義式 i 中出現義原 j} \\ 0, & \text{不出現義原 j} \end{cases} \qquad (3)$$

其中，$Neg_{i,j}$ 表示義原 j 是否有被否定意義的關係所修飾，若有則 $Neg_{i,j}$ 爲 -1，若無則 $Neg_{i,j}$ 爲 $+1$。另外，如果樹狀結構上面的義原被否定意義的關係所修飾，這否定意義會傳遞到下面的義原。

（二）、語篇（context）特徵

廣義知網雖然有嚴謹的定義式可用以表示詞彙，但是有四個缺點，造成只用義原當特徵無法正確獲得詞彙的極性。

第一個缺點是詞彙所標的義原量太少，因爲詞彙是用人工標示義原，所以無法給予很多標示。這表示詞彙擁有的資訊量有限，會造成分類器無法有效學習。第二個缺點是義原數量太少，這會造成語義的劃分不夠精確，無法顯示出真實的語義差別，例如「*明哲保身*」跟「*見風轉舵*」的定義式都是「*{sky|狡}*」，但「*明哲保身*」是正面意見，「*見風轉舵*」卻是負面意見。第三個缺點是廣義知網定義標準的差異，例如，專有名詞在廣義知網中會用客觀的義原來定義，但該專有名詞經過使用，卻可能會引起人的正反情緒，這種差異會引入程度不等的雜訊到分類器中。第四個缺點是廣義知網尚未對所有詞彙標上定義式，例如「*乾淨俐落*」在廣義知網及 NTUSD 中都出現，但廣義知網卻沒有標上定義式。

因此我們引入語篇的特性，從該詞彙在語言中的實際使用情況，抽取出詞彙的特徵，來補償這些缺點。我們使用 Liu et al.[16] 所建立的 Google Web 5-gram Version 1，來抽取語篇特徵。Google Web 5-gram 是 Google 從網路中收集大量的簡體中文網頁，並經過處理所建立的資源。他們收集了 882,996,532,572 個字符（token），共 102,048,435,515 個句子，經過斷詞後建成 n-gram。n-gram 的 n 從 1 到 5，並且只保留頻率大於 40 的 n-gram。Google Web 5-gram 的例子如圖三所示。

```
恐吓 或 辱骂 他人 </s>  796466
恐吓 或 辱骂 他人 内容  173
恐吓 或 过度 兴奋 或    251
恐吓 或 非法 骚扰 侵犯  574
恐吓 或 非法 骚扰 有    200
恐吓 或 非法 骚扰 的    4463
恐吓 或 非法 骚扰 等    705
恐吓 或 骚扰 侵犯 他人  95
```

圖三、Google Web 5-gram 資料範例

上圖中，表示「*恐吓 或 非法 骚扰 的*」這一 5-gram 共出現了 4463 次。從圖中我們也可看到，Google Web 5-gram 是簡體中文，但廣義知網爲繁體中文，所以我們先將廣義

知網用 Microsoft Word 翻譯爲簡體中文，之後才使用 Google Web 5-gram 這一語料庫。語料庫在使用時，只用 5-gram 的部分來抽取特徵。

1、Google Web 5-gram 特徵抽取

我們使用特徵跟詞彙的同出現（co-occurrence）次數做爲特徵值，以圖三爲例，如果詞彙是「*恐吓*」，以「*非法*」當特徵值，則同出現次數會將所有「*恐吓*」及「*非法*」一同出現的 5-gram 次數相加。在上面的例子中，「*恐吓*」及「*非法*」的同出現次數爲 574+200+4463 + 705=5942 次。

另外，由於廣義知網跟 Google Web 5-gram 的斷詞標準並不一致，所以在處理時把 Google Web 5-gram 的空白去掉，直接找出「*詞彙*」跟「*特徵*」這兩字串是否同時出現，來計算次數，這樣可以避免斷詞標準不一所產生的問題。例如「*一事無成*」在 Google Web 5-gram 中被斷成四個獨立的詞，將空白去掉就可以正確比對到。

因爲這裡的詞彙集合就是等待標示極性的詞，所以我們只要指定特徵的集合包括那些詞，就可算出表示詞彙 i 的向量 $V_i = (c_{i,j}) = (c_{i,1}, c_{i,2},\ldots, c_{i,\mathrm{m}})$。其中，m 是特徵集合的大小，$c_{i,j}$ 是「*詞彙 i*」跟「*特徵 j*」這兩字串同出現的次數。在我們的實驗中，共嘗試了十種不同的特徵集合，分別是廣義知網的名詞、廣義知網的動詞、廣義知網的副詞、廣義知網的形容詞、廣義知網所有詞彙、Google Web 5-gram 最常出現的 5000 詞、Google Web 5-gram 最常出現的 5000 詞（但詞彙長度最少爲 2）、Google Web 5-gram 最常出現的 10000 詞、Google Web 5-gram 最常出現的 10000 詞（但詞彙長度最少爲 2）、以及 NTUSD 完整版。

2、Google Web 5-gram 特徵值處理

用 $V_i = (c_{i,1}, c_{i,2},\ldots, c_{i,\mathrm{m}})$ 的方式來表示的缺點，是 $c_{i,j}$ 的值變化的範圍會非常大，最小爲 40，最大會到上千萬。這在機器學習中，通常需要做進一步的處理才會有比較好的結果。我們實驗了兩個不同的方法來處理這一問題：第一種是一般的餘弦標準化（cosine-normalization），將原本的向量 V_i 用公式 (4) 處理；第二種是 Esuli & Sebastiani[1] 所提的餘弦標準化 TFIDF （cosine-normalized TF-IDF），他們用該方法來處理 WordNet 中的詞彙的權重，如公式 (5) 所述。

$$CosNorm(V_i) = \frac{V_i}{\sqrt{\sum_{1\le k\le m} c_{i,k}^2}} \quad \in \Re^m \tag{4}$$

$$CosNorm(TFIDF_i) = \frac{TFIDF_i}{\sqrt{\sum_{1\le k\le m} tfidf_{i,k}^2}} \quad \in \Re^m \tag{5}$$

$$TFIDF_i = (tfidf_{i,1}, tfidf_{i,2}, ..., tfidf_{i,m})$$

$$tfidf_{i,j} = tf_{i,j} * idf_j$$

$$tf_{i,j} = \frac{c_{i,j}}{\text{特徵 } j \text{ 總出現次數}} = \frac{c_{i,j}}{\sum_{k \in D} c_{k,j}}$$

$$idf_j = \log(df_j)^{-1} = \log \frac{|D|}{|\{i : c_{i,j} > 0, \forall i \in D\}|}$$

公式 (5)中 D 表示文件的集合，此處把詞彙 i 當成文件，特徵 j 當成 term。

公式 (4) 的標準化可以讓所有詞彙的向量等長，消掉次數變化過大的缺點。公式 (5) 的想法則認爲特徵 j 的權重，應該先跨詞彙進行標準化（normalization），所以 $tf_{i,j}$ 會除以特徵 j 的總出現次數，另外再考慮特徵 j 的稀有度，所以乘上 idf_j，最後再讓所有詞彙的向量等長。我們會在後面的實驗中，比較這兩種不同權重處理方式的效能。

（三）、不同特徵的組合

我們用了基礎義原特徵 $(w_{i,1}, w_{i,2}, \ldots, w_{i,n}) = (w_{i,j})$ ，及語篇特徵 $(c_{i,1}, c_{i,2}, \ldots, c_{i,m}) = (c_{i,j})$ 來表示詞彙 i。如果想同時使用這兩種特徵中的資訊，一種直觀的方式，是將兩種特徵表示方式混合，用 $V_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,n}, c_{i,1}, c_{i,2}, \ldots, c_{i,m})$ 來表示。由於基礎義原特徵及語篇特徵都有許多不同的變形，我們無法一一嘗試所有可能的組合，所以會先分別用實驗找出最好的基礎義原特徵 $(w_{i,j})$ 及語篇特徵 $(c_{i,j})$，再把兩種特徵混合來進行實驗。我們沒有對混合後的向量做其它的處理，只是直接串接成爲 n+m 維向量。

（四）、組合式的監督式機器學習演算法（ensemble approach）

由於廣義知網詞彙的每一個意義（sense）都標有詞性，而且我們用了很多不同的特徵集合，這表示我們會有很多不同的分類器。如果依不同詞性選擇做得最好的分類器，則可以有更好的效能。例如，如果分類器 A 的總體效能不是最好，但如果它對名詞做的效能是最好的，也許拿它來預測名詞的極性會更準確，依此類推。我們把廣義知網的詞性，分爲名詞、動詞、副詞、形容詞及其他共五類，分別選在該類別預測最好的分類器來預測。這作法是一種常見的組合不同分類器的策略（ensemble approach），我們也會對此進行實驗，來觀察效能。

四、實驗與分析

（一）、實驗資料與實驗設定

本研究使用國立台灣大學意見詞詞典完整版（NTUSD）、與廣義知網的交集，作爲實驗資料，這兩個資料集的詞彙數如表一。資料集 E-HowNet∩NTUSD 會作爲標準答案集，在我們所看的相關論文中，這個答案集的大小是最大的一個。實驗使用標準答案集其中的 80% 爲訓練資料集，其餘 20%爲測試資料集，並依照實驗資料的詞性分布以及語意極性分布作分層抽樣（stratified sampling）。

表一、廣義知網、NTUSD、以及交集的資料筆數

資料集	正面	負面	總數
E-HowNet	N/A	N/A	88,127
NTUSD	21,056	22,750	43,806
E-HowNet∩NTUSD	5,346	6,256	11,602

分層抽樣詳細的作法如下：先將資料依照五種詞性以及兩種極性分成十個子集合，再針對每個子集合取其中 80%作爲訓練資料，另外 20%作爲測試資料。由於我們的資料量夠多，所以可以使用這種抽樣。這種抽樣主要的好處在於我們更容易對測試結果進行更多的分析，我們把分層抽樣的結果列於表二。

表二、訓練資料的詞性以及傾向分布

詞性		全部資料集		正面傾向	訓練資料集		測試資料集	
		正面	負面	百分比	正面	負面	正面	負面
名詞	2,040	931	1,109	45.64%	745	887	186	222
動詞	9,056	4,134	4922	45.65%	3,307	3,938	827	984
副詞	383	206	177	53.79%	165	142	41	35
形容詞	74	45	29	60.81%	36	23	9	6
其他	49	30	19	61.22%	24	15	6	4
總數	11,602	5,346	6,256	46.08%	4,277	5,005	1,069	1,251

本研究使用 Chang & Lin[17] 所發布的 LIBSVM 支援向量機，來當監督式機器學習演算法，使用 radial basis function (RBF) kernel function， RBF 的兩個手動參數 cost c 與 gamma g 以網格搜尋（Grid Search）的方式尋找最佳參數值 (c, g)，搜尋範圍 $c \in \{2^{-5}, 2^{-3}, 2^{-1}, ..., 2^{15}\}$、$g \in \{2^{-15}, 2^{-13}, 2^{-11}, ..., 2^{-3}\}$，總共 110 組參數，取五疊交叉驗證(5-fold cross validation）中平均準確率最高的參數。

我們使用預測準確率（accuracy）來比較分類器間的效能，這是看訓練出的分類器在測試資料集中的成績，而分類器會對測試資料集中的所有詞彙都進行極性的預測。另外，使用 McNemar 檢定[18]來測試分類器的效能差距是否爲顯著，顯著水準設定爲 0.95。

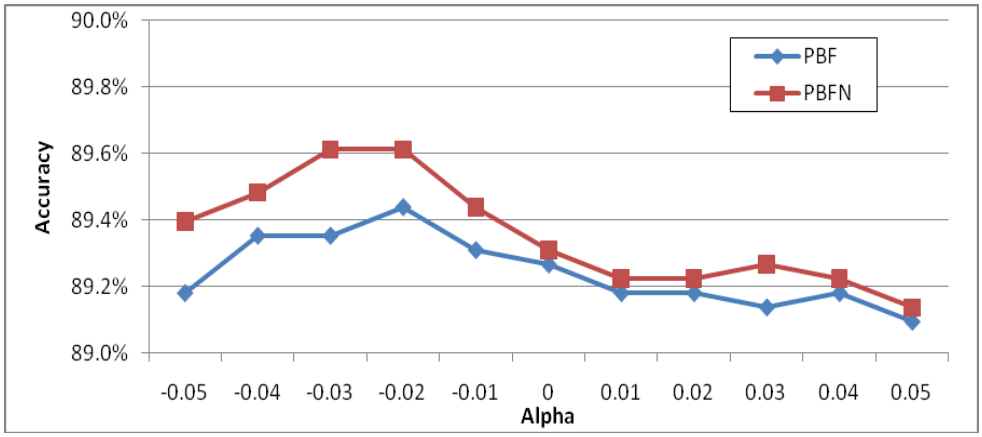
McNemar 檢定將測試資料依照兩個分類器（以下稱爲分類器 A 與分類器 B）的標記，分成四組並計數。其中測試樣本數即爲下面 $n_{1,1}$、$n_{0,1}$、$n_{1,0}$、$n_{0,0}$ 四個數字的總合，在虛無假設（null hypothesis）中，兩個分類器應具有相同的錯誤率，也就是 $n_{0,1}=n_{1,0}$ 。

$n_{1,1}$：	分類器 A 與分類器 B 皆正確標記的樣本數	$n_{0,1}$：	分類器 A 標記錯誤，但分類器 B 標記正確的樣本數
$n_{1,0}$：	分類器 B 標記錯誤，但分類器 A 標記正確的樣本數	$n_{0,0}$：	分類器 A 與分類器 B 皆錯誤標記的樣本數

McNemar 檢定建構在卡方適合度檢定 （$\chi 2$ test goodness of fit）上，整理而得的檢定值爲 $\frac{(|n_{0,1}-n_{1,0}|-1)^2}{n_{0,1}+n_{1,0}}$，此檢定值在 $n_{0,1}+n_{1,0}$ 夠大的時候會趨近於自由度爲 1 的卡方分配，因此在顯著水準（significant level）爲 0.95 時，此值若大於 $\chi^2_{1,0.95}=3.8415$，則拒絕虛無假設。我們用 (McNemar 檢定結果, p-value) 來顯示我們的檢定結果，例如檢定結果 (1.50, 0.22) 表示，McNemar 檢定結果爲 1.50 < 3.84，所以沒有通過 McNemar 檢定，p-value 爲 0.22。

（二）、基礎義原特徵的效能

圖四爲基礎義原方法在不同 α 值所得到的預測準確率，其中公式 (2) 的結果是 PBF（Prime-Based Feature）那條折線，最佳的 α 值爲 −0.02，準確率爲 89.4397%。當 PBF 中 α = 0，該結果即爲公式 (1) 的結果。公式 (3) 的結果是 PBFN（Prime-Based Feature with Negation）那條折線，最佳的 α 值爲 −0.02 及 −0.03，準確率爲 89.6121%。



圖四、廣義知網特徵於不同 α 值的效能比較

我們從圖四可以看出，描述 PBFN 的折線在所有的 α 值下，準確率皆略高於 PBF，但是兩個最大值（α = −0.02）的差距僅 0.1724%，此差距爲不顯著，檢定結果 (1.50, 0.22)。由於 α < 0 有最佳效能，這表示深度較深給較高權重，該義原有較好的特徵，可以給分類器學習。

（三）、語篇特徵的效能

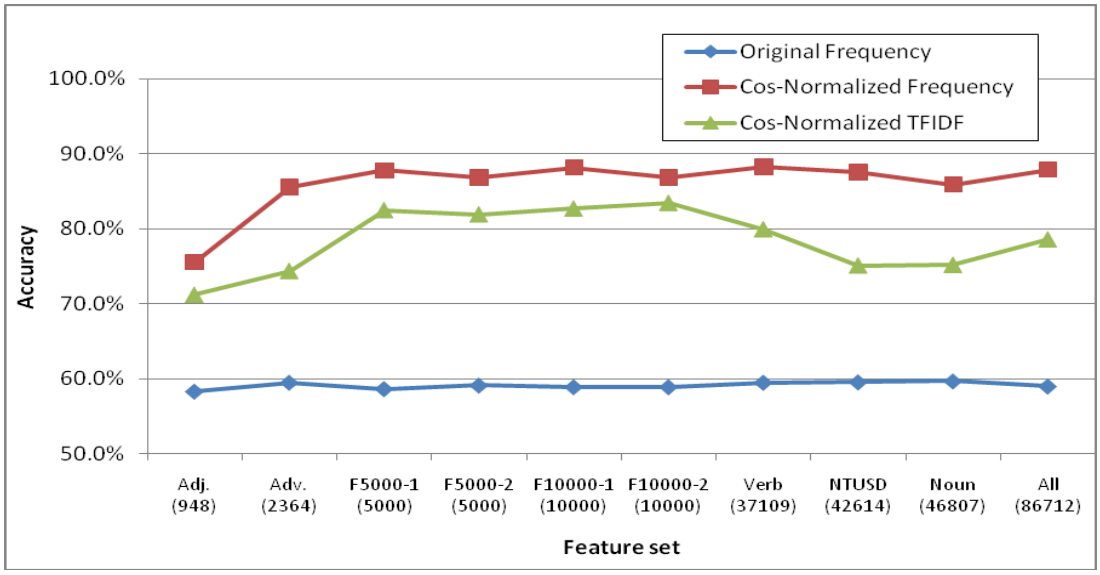
語篇特徵使用十組特徵集的名稱，以及特徵數量，如表三所示。在表中，我們使用特徵

集代號來代表該特徵集。十組特徵集中，最少的是 *Adj* 的特徵集，只有 948 個詞，最多的是 *All* 的特徵集，有 86,712 個詞。

表三、語篇特徵所使用的特徵集與其特徵數

特徵集	特徵集代號	特徵數
廣義知網名詞	*Noun*	46,807
廣義知網動詞	*Verb*	37,109
廣義知網副詞	*Adv.*	2,364
廣義知網形容詞	*Adj.*	948
廣義知網所有詞彙	*All*	86,712
最常出現 5000 詞	*F5000-1*	5,000
最常出現 5000 詞（長度≧2）	*F5000-2*	5,000
最常出現 10000 詞	*F10000-1*	10,000
最常出現 10000 詞（長度≧2）	*F10000-2*	10,000
NTUSD（完整版）	*NTUSD*	42,614

我們使用三種不同的加權方式得到的預測準確率如圖五，圖中我們也把特徵集的特徵數由左至右由小到大排列。



圖五、使用語篇特徵時的預測效能

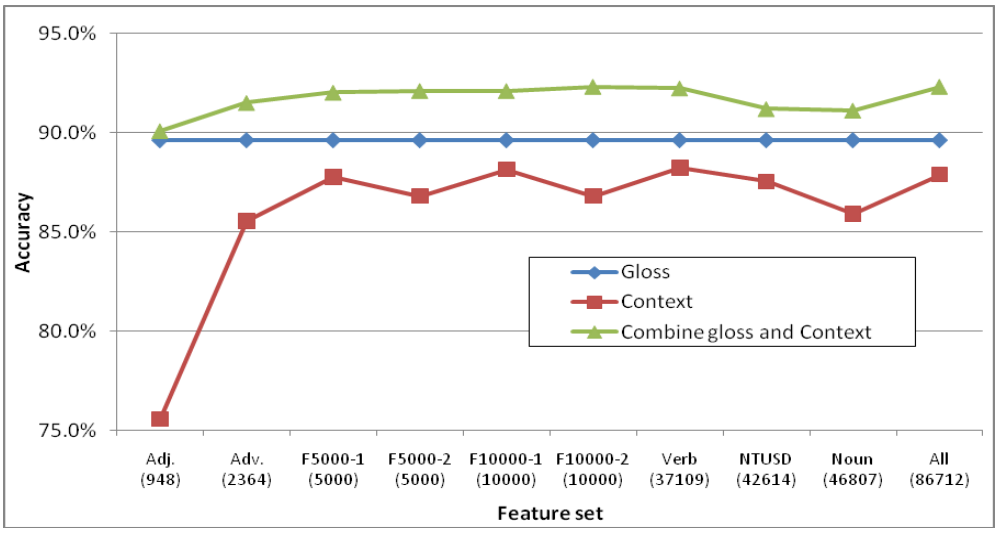
從圖五可以看出，沒有標準化的原始頻率的最佳準確率爲 59.70%，使用的特徵集爲「廣義知網名詞」，其效能最差且差距很大。餘弦標準化 TFIDF 的效能排在中間，最佳準確率爲 83.41%，使用的特徵集爲「最常出現 10000 詞」。而經過餘弦標準化的特徵值則可以得到最佳效能，其最佳準確率爲 88.23%，此時使用的特徵集爲「廣義知網動詞」，此效能跟其他兩者的差距爲顯著，檢定結果 (4.61, 0.03)。

圖五中特徵集的個數，並沒有絕對的影響，但若個數太少，如特徵個數小於 2364 個，則效能會明顯變差。圖四中的最佳值 PBFN（$\alpha = -0.02$）爲 89.61%，特徵個數爲 2,567 個，這個值比圖五中的最佳值 88.23%還要大，這表示廣義知網中的特徵比較準確，但這差距爲不顯著，檢定結果 (2.49, 0.11)。

（四）、組合不同特徵的效能

組合特徵時，因爲餘弦標準化有最好的效能，所以語篇特徵選擇餘弦標準化後的十組特徵集，分別與廣義知網特徵效能最好的 $PBFN_{\alpha=-0.03}$ 組合，來訓練分類器，分類器預測準確率如圖六。其中廣義知網特徵的特徵集效能爲固定，因此以水平直線表示（gloss 那條折線）。組合而成的特徵集，以「*語篇特徵集代碼*+$PBFN_{\alpha=-0.03}$」加以命名，例如「*F10000-2*+$PBFN_{\alpha=-0.03}$」表示「最常出現 10000 詞（長度≧2）」跟「$PBFN_{\alpha=-0.03}$」兩個特徵集的組合。

我們從圖六可以看出，將廣義知網特徵與外部語料特徵組合之後，準確率都有顯著提升，提升後的最高準確率爲 92. 3276%，使用「廣義知網所有詞彙 *All*+$PBFN_{\alpha=-0.03}$」和「最常出現 10000 詞（長度≧2） *F10000-2*+$PBFN_{\alpha=-0.03}$」爲特徵集時皆有相同的準確率。上圖中，「廣義知網所有詞彙 *All*」準確率從 88.23%提升至 92.33%時，此差距爲顯著，檢定結果 (32.14, $1.4*10^{-8}$)。



圖六、廣義知網、語篇特徵、與組合特徵的準確率比較

（五）、組合式的監督式機器學習演算法效能

在圖六中，組合出的特徵集有十個，所以共有十個分類器，每個分類器在訓練時，對不同詞性有不同的效能，我們將這十個分類器對於每個詞性的標記效能整理成表四。表四中的特徵集代號是「*語篇特徵集代碼*+$PBFN_{\alpha=-0.03}$」的簡寫，因爲使用相同的 $PBFN_{\alpha=-0.03}$，所以將其忽略。「總體效能」是指分類器訓練時的整體效能。表中，一欄中最佳的

標記效能以**粗體字**表示。

表四、訓練資料集中，組合特徵對不同詞性的標記準確率

特徵集代號	總體效能	訓練資料集中，依詞性分別計算的準確率				
		名詞	動詞	副詞	形容詞	其他
Adj.	94.3223%	95.9559%	94.2167%	89.9023%	93.2203%	82.0513%
Adv.	95.3243%	96.5074%	95.2795%	92.1824%	91.5254%	84.6154%
F5000-1	96.1000%	97.3039%	96.0110%	92.8339%	94.9153%	89.7436%
F5000-2	97.2635%	98.0392%	97.1705%	**96.0912%**	94.9153%	**94.8718%**
F10000-1	96.2400%	97.3652%	96.1767%	92.8339%	94.9153%	89.7436%
F10000-2	**97.5005%**	**98.2843%**	**97.4189%**	**96.0912%**	94.9153%	**94.8718%**
Verb	96.5632%	97.5490%	96.5079%	94.4625%	91.5254%	89.7436%
NTUSD	96.8218%	97.3039%	96.8254%	95.1140%	93.2203%	**94.8718%**
Noun	96.8541%	98.1005%	96.6460%	**96.0912%**	**96.6102%**	89.7436%
All	96.4124%	97.4265%	96.3699%	93.1596%	94.9153%	89.7436%

表四中我們可以發現，訓練時，*F10000-2*+PBFN$_{\alpha=-0.03}$ 有最高的總體效能，其各詞性效能除了形容詞外，多是最好；考量到資料集中形容詞的數量並不多，這表示組合多個分類器後，效能的提昇空間可能有限。表四中另一個值得注意的一點是訓練資料集的內部測試效能（inside test）*F10000-2*+PBFN$_{\alpha=-0.03}$ 的 97.5005% 跟實際測試效能 92. 3276% 相比，降低了 5.31%，這降低幅度並不大，顯示這分類器的 generalization 能力不錯，這也是使用 Google Web 5-gram 的優點，它可產生較強健（robust）的分類器[19]。

我們在表四中選不同詞性做得最好的分類器來組合，如果效能相同，則選特徵數量較少的那一個分類器，因爲特徵數較少通常在未看過的資料集會做得較好。組合出的分類器我們稱爲 *EnsembleClassifier*，其結果列在表五，其中 *F10000-2*+PBFN$_{\alpha=-0.03}$ 於各詞性的標記效能也列出來比較。

表五、組合分類器於各詞性的標記效能及比較

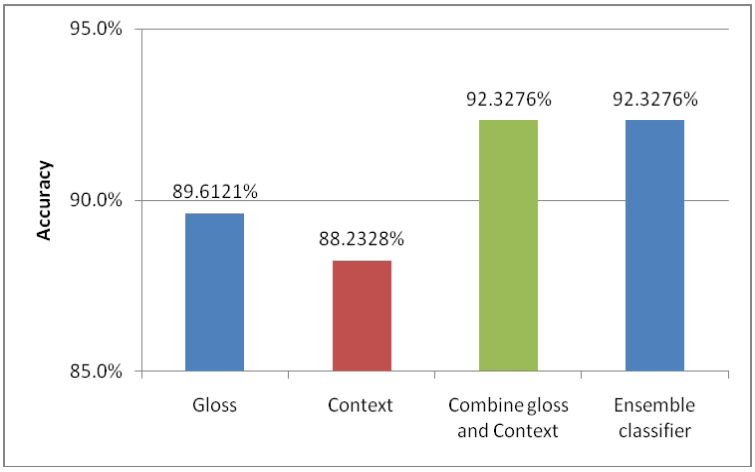
分類器 / 詞性	*F10000-2*+PBFN$_{\alpha=-0.03}$ 分類器於各詞性的標記效能			組合分類器 *EnsembleClassifier* 於各詞性的標記效能				
	正確個數	錯誤個數	準確率	使用的分類器	正確個數	增減	錯誤個數	準確率
名詞	371	37	90.9314%	*F10000-2*	371	(+0)	37	90.9314%
動詞	1,681	130	92.8216%	*F10000-2*	1,681	(+0)	130	92.8216%
副詞	67	9	88.1579%	*F5000-2*	69	(+2)	7	90.7895%
形容詞	14	1	93.3333%	*Noun*	12	(-2)	3	80.0000%
其他	9	1	90.0000%	*F5000-2*	9	(+0)	1	90.0000%
總數	2,142	178	92.3276%		2142	(+0)	178	92.3276%

表五中，我們也列出每種詞性做錯與做對的個數，並以 *F10000-2*+PBFN$_{\alpha=-0.03}$ 分類器爲基準，看組合後的分類器，在各詞性中做對做錯的次數的增減，用括號來標出增減的數量。

EnsembleClassifier 所得成績跟 *F10000-2*+PBFN$_{\alpha=-0.03}$ 相同，這表示目前的分類器組合方式，無法提升效能。

（六）、相關研究效能比較

我們總結前面各種不同的實驗結果，畫成圖七，來方便我們比較效能。其中，gloss 表基礎義原特徵 PBFN$_{\alpha=-0.03}$，最好的效能到 92.3276%。



圖七、四種方法效能比較

由於我們使用 NTUSD，我們想看看 NTUSD 人類標記的效能跟我們分類器的效能有何差異。在 Ku & Chen [2]的研究中，聘請標記者對舊版 NTUSD 進行標記。舊版 NTUSD 爲經過翻譯的 General Inquirer（GI）與 Chinese Network Sentiment Dictionary（CNSD）的組合，每個詞彙都有人工的意見標記。該研究中標記者的最佳標記效能與本研究的比較如表六，從表六中可以看出，本研究所產生的自動標記演算法達到了接近人類標記的效能。

表六、NTUSD 標記者與本研究標記效能比較

分類器	Recall	Precision	F-Measure
F10000-2+PBFN$_{\alpha=-0.03}$	92.36%	92.20%	92.27%
三人中最佳的人類標記者	96.58%	88.87%	92.56%

表六中，人類標記者的 Recall 及 Precision 取自 Ku & Chen [2]。*F10000-2*+PBFN$_{\alpha=-0.03}$

的預測結果爲 (True Positive, False Positive, True Negative, False Negative) = (TP, FP, TN, FN) = (968, 77, 1174, 101)，其中 Positive 表正面極性。我們分別對正負面極性計算 Recall、Precision 及 F-Measure (R^+、P^+、F^+、R^-、P^-、F^-)，其中，P^+=TP/(TP+FP)、R^+=TP/(TP+FN)、F^+= $2P^+R^+/(P^++R^+)$、P^-=TN/(TN+FN)、R^-=TN/(TN+FP)、F^-= $2P^-R^-/(P^-+R^-)$，最後系統的 Recall=$(R^++R^-)/2$、Precision=$(P^++P^-)/2$ 及 F-Measure = $(F^++F^-)/2$ = (91.58% + 92.95%)/2 = 92.27%。由計算中我們可以看到，我們的系統對負面極性做得較好，而且因資料集有較多的負面詞彙，所以最後的準確率 92.33% 比 F^+ 高。

五、結論

本研究使用了 Google Web 5-gram Version 1 來抽取語篇特徵，並加上來自 E-HowNet 的基礎義原特徵，用監督式機器學習的方法，來預測 E-HowNet 詞彙的意見極性。雖然單獨使用不同的特徵已經可以接近 90% 的準確率，但如果把兩種特徵都加以使用，分類器的極性預測的準確率可到達 92.33%，這個結果跟人的標記準確率不相上下；以這種方式建立的分類器，可用來自動標記 E-HowNet 詞彙的意見極性。

我們希望在未來能把這種方式，往不同的方向擴展，來給予 E-HowNet 詞彙更多意見的屬性，這包括對詞彙標記主客觀的屬性及正負面傾向的強度等。除此之外，因爲 E-HowNet 詞彙有許多不同的詞性，我們也希望能把我們的方法，運用詞性的層次來進行標記。藉由提供更精確的字彙意見標記資訊，來支援句子及文件層次的意見分析。

致謝

Research of this paper was partially supported by National Science Council (Taiwan) under the contract NSC 98-2221-E-002-175-MY3.

參考文獻

[1] A. Esuli and F. Sebastiani, “Determining the semantic orientation of terms through gloss classification,” In *Proceedings of* CIKM-05, pp. 617–624, 2005..

[2] L.-W. Ku and H.-H. Chen, “Mining opinions from the Web: Beyond relevance retrieval,” *Journal of the American Society for Information Science and Technology*, vol. 58, no. 12, pp. 1838-1850, 2007.

[3] A. Esuli and F. Sebastiani, “SentiWordNet: A publicly available lexical resource for opinion mining,” In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 06)*, pp. 417–422, 2006.

[4] Z. Dong and Q. Dong, *HowNet and the Computation of Meaning*. World Scientific, 2006.

[5] 陳克健, 黃淑齡, 施悅音, 和 陳怡君, “多層次概念定義與複雜關係表達－繁體字知網的新增架構,” *漢語詞彙語義研究的現狀與發展趨勢國際學術研討會，北京大學*, 2004.

[6] V. Hatzivassiloglou and K. R. McKeown, “Predicting the semantic orientation of adjectives,” In *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics*, pages 174–181, Madrid, ES, 1997. Association for Computational Linguistics.

[7] P. D. Turney and M. L. Littman, “Measuring praise and criticism: Inference of semantic orientation from association,” *ACM Transactions on Information Systems,* 21(4):pp. 315–346, 2003.

[8] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, “Using WordNet to measure semantic orientation of adjectives,” In *Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation*, vol. 4, pp. 1115–1118, Lisbon, PT,2004.

[9] A. Esuli and F. Sebastiani, “Determining term subjectivity and term orientation for opinion mining,” 2006, pp. 193-200.

[10] R. W. M. Yuen, T. Y. W. Chan, T. B. Y. Lai, O. Y. Kwong, and B. K. Y. T’sou, “Morpheme-based derivation of bipolar semantic orientation of Chinese words,” 2004, pp. 1008-1014.

[11] J. Yao, G. Wu, J. Liu, and Y. Zheng, “Using bilingual lexicon to judge sentiment orientation of Chinese words,” 2006, pp. 38-43.

[12] D. Li, Y.-tao Ma, and J.-li Guo, “Words semantic orientation classification based on HowNet,” *The Journal of China Universities of Posts and Telecommunications*, vol. 16, no. 1, pp. 106-110, 2009.

[13] Z. Han, Q. Mo, M. Zuo, and D. Duan, “Efficiently identifying semantic orientation algorithm for Chinese words,” presented at the *International Conference on Computer Application and System Modeling*, 2010, vol. 2, pp. 260-264.

[14] B. Lu, Y. Song, X. Zhang, and B. Tsou, “Learning Chinese polarity lexicons by integration of graph models and morphological features,” *Information retrieval technology*, pp. 466-477, 2010.

[15] 刘群 and 李素建, “基于《知网》的词汇语义相似度计算,” *第三届汉语词汇语义学研讨会*, 2002.

[16] F. Liu, M. Yang, and D. Lin, “Chinese Web 5-gram Version 1.” Linguistic Data Consortium, Philadelphia, 2010.

[17] C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*. 2001.

[18] T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” *Neural computation*, vol. 10, no. 7, pp. 1895-1923, 1998.

[19] S. Bergsma, E. Pitler, and D. Lin, “Creating robust supervised classifiers via web-scale N-gram data,” in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010, pp. 865-874.