應用直方圖均化於統計式未知詞萃取之研究

Histogram Equalization for Statistical Unknown Word Extraction

陳弈璁　Yi-Cong Chen
國立台灣科技大學資訊管理學系
Department of Information Management
National Taiwan University of Science and Technology
m9709104@ntust.edu.tw

林伯慎　Bor-Shen Lin
國立台灣科技大學資訊管理學系
Department of Information Management
National Taiwan University of Science and Technology
bslin@cs.ntust.edu.tw

摘要

隨著人們的生活方式的演變以及資訊普及的加速，新事物、新觀念不斷的產生，新的詞彙自然而然地快速增加。因此，學習與辨識新詞彙是一個自然語言處理系統能與時俱進的重要能力。本論文利用統計式的機器學習方法，結合不同特性的統計特徵訓練出一個詞彙的分類器，進行詞彙的萃取與驗證。然而，自然語言處理技術的應用範疇非常廣，用來訓練或測試的語料庫其領域或大小也都不盡相同，這使得以統計爲基礎的方法，會產生訓練集與測試集的特徵分佈不匹配的問題。我們提出應用直方圖均化（Histogram Equalization）將描述長度增益（Description Length Gain）特徵值進行正規化，讓測試集與訓練集的特徵值分佈能互相匹配，解決語料庫大小或領域不同所造成特徵值範圍變動及分佈差異的問題。這使得本論文的統計式詞彙萃取方法更具有一般性，可以適用於不同領域的詞彙萃取。

我們使用SIGHAN2的繁體語料庫進行測試，在結合四種統計特徵，並且經過特徵值分佈正規化後，會有最佳的詞彙驗證效能。對於中研院資訊所組庫小組及香港城市大學所提供的語料庫，F-measure分別可以達到68.43%和71.40%。我們將此詞彙萃取方法應用於萃取新穎領域的未知詞時，發現本論文方法可以萃取出具有統計特性顯著但較難透過語意結構資訊萃取出來的未知詞，例如：「海角7號」、「金融海嘯」等專有名詞。但是相對地，因爲並未使用語意結構規則，於人名、地方名或組織名的未知詞萃取，則顯得能力較爲不足。我們並觀察到，本論文的統計萃取方法與上述兩套斷詞系統所萃取的未知詞之間具有良好的互補性，適當地將這些方法結合將可以達到截長補短的效果。

Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Pages 364-378, Puli, Nantou,Taiwan, September 2010.

Abstract

With the evolution of human lives and the accelerated spread of information, new things and concepts are generated quickly, and new words emerge every day. It is therefore important for natural language processing systems to identify new words. This paper used the scheme for Chinese word extraction based on machine learning approaches to combining various statistical features. Due to the broad areas for the natural language applications, however, it is quite probable that the mismatch of statistical characteristics between the training and the testing domains occurs, which degrades the performance for word extraction inevitably.

This paper proposes the scheme of utilizing the histogram equalization for feature normalization in statistical approaches. Through this scheme, the mismatch of the feature distributions for the training set and the testing set, with different sizes or in different domains, can be compensated. This makes the statistical approaches of unknown word extraction more robust for novel domains.

This scheme was tested on the corpora provided by SIGHAN2. The best results, 68.43% and 71.40% of F-Measure for the CKIP corpus and the HKCU corpus respectively, can be achieved with four features with normalization and histogram equalization. When applied to unknown word extraction in an novel domain, it can be found that this scheme is capable of identifying such pronouns as “Cape No. 7”(海角七號), “Financial Tsunami”(金融海嘯) and so on, which are not easy to be extracted by those approaches based on semantic characteristics. This scheme appears not good enough for extracting such new terms as the names of humans, places and organizations, in which the semantic structures are prominent. When compared with the results of unknown word extraction for two Chinese word segmentation systems, it can be observed that this scheme exhibits to be complementary with other approaches, and it is promising to combine approaches with different capabilities.

關鍵字：未知詞萃取、機器學習、多層次類神經網路、中文詞彙萃取、直方圖均化。

Keywords: Unknown Word Extraction, Machine Learning, Multilayer Perceptrons, Chinese Word Extraction, Histogram Equalization.

一、緒論

隨著人類生活方式的演變和資訊普及的加速，新的詞彙在網路與媒體上不斷快速地增加。這使得自然語言處理系統必須具有學習新詞的能力，才能與時俱進。例如，在中文斷詞系統中通常都會用到詞典；系統雖然可以盡可能地增加詞典中的詞彙數量，但是無論詞彙量有多大，都不可能含括所有可能用到的詞彙。這是因爲，自然語言處理的應用領域非常廣泛，自然語言本身就是隨著時間演進，在各種知識領域中都有獨特的、不斷新增的關鍵字詞或專有名詞等，這些不是系統設計者可以預先知道的。因此，斷詞系統的詞典內包含的詞彙不能一成不變，應該隨著處理的文章或相關領域做更新。如果自然語言處理系統能針對各種新領域自動萃取出未知的詞彙，對於系統的應用範圍或新領域的探索會有相當大的幫助。

一般萃取未知詞的方法主要有統計式與法則式兩大類。法則式的萃取方式，主要依據未知詞的種類，考慮詞彙的語意，而訂定特定的萃取規則。在文章中，常出現具有規則的

未知詞，包含了人名、專名詞、複合詞、數值等。例如：Sun 等人針對中文人名做辨識[1]。在統計式的萃詞方法中，運用人們的習慣，詞彙組成的特徵資訊，利用統計的方式計算出數值，判斷是詞彙的可能性，例如：利用字元組在文章中出現的次數，Lu 等人提出運用字元組出現頻率[2]，並且判斷字元組被另一個字元組包含時，兩字元組出現頻率是否相同，做爲刪除候選詞的決策條件。但是很難藉由某一種特徵值，就能完全的模擬詞彙的特性。在中文語言的處理系統中，萃取新詞，常運用在改善斷詞的準確性。例如 Hai zhao 利用常用的統計特徵[3]，計算字元組的特徵值，篩選出未知詞，適當的刪減詞量，運用在斷詞上，針對特徵值的數值，實驗出一個門檻值，做爲篩選未知詞的依據。但這樣的做法，經過多次的實驗，得到一個最佳的門檻值，會增加不確定因素，而且過度依賴特定語料庫的特性。此外，除了單純利用統計方式外，還有結合訂定法則的方式。例如：Ken-Jiann 等人在大量的語料庫中[4]，以字元爲觀點，找出規則的模組，標示未知詞的可能位置，針對被標示未知詞的部分，分別利用統計和法則，對於不同的種類型態的未知詞，用不同的規則篩選，其特點在於可以找出低詞頻的未知詞。

詞彙的組成方式非常的複雜，不論是統計式、法則式或是結合兩種特性，都必須建立許多的規則或是找出許多的特徵，也因爲使用不同的語料庫，所以找出來的規則，或是統計出來的特徵值，無法適用於其他的語料庫。因此有學者提出運用分類器的方式，萃取新詞，例如：梁婷等人利用構詞學的原理[5]，及非詞彙的篩選法則，將三音新詞篩選出來並且過濾掉大部分非詞彙的新詞，針對這些詞統計特徵，利用統計特徵值結合類神經網路來萃取新詞。GOH 等人針對字元在構成詞彙的特性[6]，例如字元的詞性，字元位置詞彙的位置。結合了鄰近字元的特性和支持向量機（Support Vector Machines , SVM）訓練一個分辨字元在詞彙中的位置，進而萃取出新詞。

爲了降低對於特定訓練語料庫的依賴性，並且能針對各種不同領域，萃取出詞彙。因此，本篇論文的研究，主要利用統計式的方法，探討如何結合不同特性的統計特徵，應用機器學習方法來萃取詞彙。針對統計特徵值的分佈進行直方圖均化（Histogram Equalization），使得測試特徵值分佈能與訓練特徵值分佈能互相匹配，解決語料庫大小或領域不同所造成特徵值範圍變動及分佈差異的問題，不必因爲領域的差異而重新訓練詞彙萃取模型，使得本論文的詞彙萃取方法更具一般性。

我們使用 SIGHAN2 的繁體語料庫進行詞彙萃取的測試，在結合 DLG、AV、Link、PreC 四種特徵時，並且利用直方圖均化的方法對於 DLG 特徵分佈進行正規化，可以使得 F-Measure 上升 8%左右。對於中研院資訊所詞庫小組及香港城市大學所提供的語料庫，F-Measure 分別可以達到 68.43%和 71.40%。最後將本論文詞彙萃取方法應用於新穎領域的資料，從新穎領域資料萃取出未知詞，並與中央研究院資訊所詞庫小組和中國科學院計算技術所提供的斷詞系統抽取的未知詞進行分析比較，我們發現本論文方法與其兩套斷詞系統具有互補的特性，可以萃取出具有強烈的統計詞彙特性且難以透過語意的方式萃取出來的未知詞。但是對於人名或地方名稱的未知詞萃取，則本論文方式萃取能力較不足。

二、統計特徵的計算

在計算統計特徵之前，我們先統計字元 n 連次數(character n-gram)，初步篩選出現次數大於等於 5 且長度小於等於 7 的字元組，作爲候選的未知詞，稱爲候選詞。所有候選詞所形成的集合，稱之候選詞集。針對這些候選詞，可計算下列各種不同特性的統計特徵：

1. 字元 n 連次數對數值（Logarithm of Character N-Gram, LogC）

$$LogC(T_i) = log\big(C(T_i)\big) \qquad \text{（公式 2.1）}$$

T_i: 第 i 個候選詞

$C(T_i)$: 候選詞 T_i 在所有文件中出現的總次數

詞彙本身就具有重複出現的特性，因此，若候選詞出現的次數愈高，愈有可能是詞彙。

2. 描述長度增益（Description Length Gain, DLG）

$$DLG(\mathrm{T}_i) = L(X) - L(X[@ \rightarrow T_i]) \qquad \text{（公式 2.2）}$$

$$L(X) = -|X| \sum_{x \in V} p(x) log_2 p(x)$$

X: 語料庫中的所有文句

$|X|$: 語料庫中所有文句的字元總數

V: 語料庫中所有字元所構成的集合

$L(\cdot)$: 語料庫的資訊量（亂度）

$X[@ \rightarrow T_i]$: 語料庫所有文句中，將候選詞 T_i 取代成"@"

描述長度增益特徵是由 Kit 等人所提出來的統計特徵[7]，主要概念是利用資料壓縮的程度來評估字元組是一個詞彙的可能性。公式 2.2 中的 $L(X)$爲語料庫含有 T_i 的資訊量，$L(X[@\rightarrow T_i])$則是將語料庫中所有出現的候選詞 T_i 取代爲"@"之後的資訊量。因此，$DLG(T_i)$表示 T_i 所產生的語料庫資訊量增益，可以反應出該候選詞對於整個語料庫資訊量的貢獻度。對語料庫資訊量貢獻度愈高的候選詞，愈可能是個詞彙。

3. 介接變異度（Accessor Variety, AV）

$$AV(T_i) = min\{ L_{AV}(T_i), R_{AV}(T_i)\} \qquad \text{（公式 2.3）}$$

$L_{AV}(T_i)$: 候選詞左邊相鄰不同字元的個數

$R_{AV}(T_i)$: 候選詞右邊相鄰不同字元的個數

介接變異度是由 Feng 等人提出[8]，用來衡量一個字元組獨立出現的程度。其主要想法是，若字元組前後可鄰接的不同字元數愈高，則該字元組愈可能是一個詞彙。反之，若字元組可鄰接字元數低的時候，顯示該字元組並不常被單獨使用，而是須伴隨其他特定字元一起被使用。因此，該字元組較可能只是一個詞彙的一部份，本身並非一個詞彙。

所以 $AV(T_i)$利用的是候選詞的上下文資訊，來反應該候選詞獨立出現的程度，獨立性愈高，表示愈像是一個詞彙。

4. 鏈結強度對數值（Logarithm of Total Links, Link）

在 LogC 特徵中只考慮了候選詞出現的次數，並未考慮候選詞的子結構（子字元組）對該候選詞是否爲詞彙的支持強度，我們因此提出了鏈結強度特徵。鏈結強度不僅考慮候選詞本身出現的次數，也考慮其內部結構所蘊含的支持強度。計算的方式是累計該候選詞內部所有可能子字元組的 n 連次數，如下列公式。

$$Link(T_i) = log\left(\sum_{k\leq l} \mathrm{C}(S(T_i;k,l))\right) \qquad \text{（公式 2.4）}$$

$S(T_i;k,l)$：從候選詞 T_i 中取出位置 k 到 l 的字元組。

以字元組「行政院長」爲例，其包含的子字串除了“行政院長”外還有“行政”、“行政院”、“政院”、等，累加所有子字串的出現次數，即可得到鏈結強度。

5. 字首分離度（PreC）

我們利用與字元組擁有相同的字首的其他字元組與其子字元組的資訊，加強此特徵值的可靠度；也就是透過具有相同字首的字元組，一起計算出字首的分離程度。如果字首的分離度愈大，則候選的字元組較可能不是詞彙。其計算公式如下：

$$\bar{C}(F) = \sum_{x \in S(F)} C(x_{1L}) \qquad \text{（公式 2.5）}$$

$$PreC(T_i) = \begin{cases} \frac{1}{|S(F)|}\bar{C}(F) & if\ |Ti| > 2 \\ C(T_i) & elsewhere \end{cases}$$

F：T_i 的字首

$S(F)$：以 F 字元爲字首的字元組所組成的集合，且字元組長度需大於 2

$|S(F)|$：S 集合的字元組總數

x_{1L}：不包含字首的 x 子字元組 $x[1:L]$

我們針對個別字首，先取出長度爲 3 到 7 的字元組，計算其移除字首後子字元組出現次數的平均，作爲該字首的分離度，若候選詞長度大於 2，輸出其字首分離度，小於等於 2 時，則以候選詞出現次數替代。例如：「在台北」，與其相同字首的字元組有「在拍攝」、「在學校」等，則分別統計子字元組「台北」、「拍攝」、「學校」的出現次數，計算其平均值，即可取得字首分離度。

由於前面所述的特徵都是由語料庫統計得到，這些特徵的數值會受語料庫大小的影響，而落在不同的範圍；如果測試和訓練的語料庫大小有明顯的差異時，則訓練和測試的統計特徵值將落在不同的動態範圍，這使得訓練出來的分類器無法對測試的資料做可

靠地分類。爲了解決這個問題，我們把上述的各個統計特徵以線性方式正規化到0至1之間，公式如下：

$$F(v) = \frac{v - Min(y)}{Max(y) - Min(y)} \qquad \text{（公式 2.6）}$$

v：輸入的特徵數值

y ：特徵值的種類

Min(y)：y特徵值中最小的數值

Max(y)：y特徵值中最大的數值

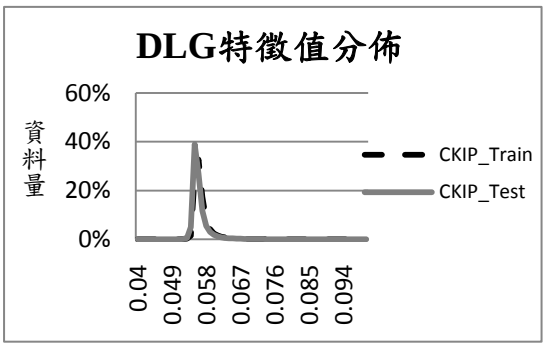
F(v)：特徵值 v 經過正規化後的輸出值

三、詞彙萃取的方法

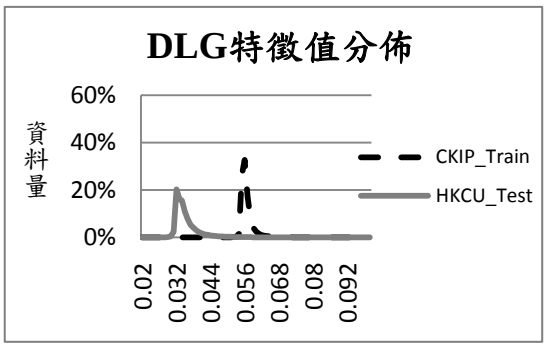
(一)、統計特徵分佈的問題

由於本論文使用的分類特徵都是由語料庫統計得到，特徵值分佈容易因爲語料庫的不同而有所差異。雖然利用第二章節中公式(2.6)，可以將數值範圍正規化到 0 至 1 之間，但是當分類器應用於跨領域的分類資料時，有些統計特徵值可能會因語料庫領域或大小有明顯的差異。這會造成訓練集與測試集的特徵分佈不匹配，而影響到分類的正確率。所以我們提出改進正規化的方法，將在下一節介紹。

我們以在 SIGHAN2 競賽中由中央研究院資訊科學研究所詞庫小組（Chinese Knowledge Information Processing Group, Institute of Information Science, Academia Sinica, 簡稱 CKIP）以及香港城市大學（HKCU）所提供的繁體語料庫爲例。首先，將 CKIP 提供的語料庫，隨機且平均地分成兩份語料庫，一份當作訓練語料庫，稱爲 CKIP_Train，另一份當作同領域的測試語料庫，稱爲 CKIP_Test。HKCU 的語料庫簡稱爲 HKCU_Test。因爲 CKIP_Train 語料與 CKIP_Test 語料是由 CKIP 提供的語料庫隨機等分的語料庫，爲相同領域的語料庫。HKCU_Test 語料相對於 CKIP_Train 語料，則屬於跨領域的語料庫。我們分別統計這三種語料庫中特徵值分佈，比較相同領域與跨領域的特徵值分佈。圖 3.1 是 DLG 統計特徵值的分佈，圖中的(a)顯示相同領域的語料庫 CKIP_Train 和 CKIP_Test 中的 DLG 分佈差異，圖(b)則顯示跨領域語料庫 CKIP_Train 和 HKCU_Test 中的 DLG 分佈差異。分別以語料庫的名稱命名分佈曲線，例如：圖(a)中的 CKIP_Train 曲線，表示從 CKIP_Train 語料庫中統計得到的 DLG 特徵值分佈。由圖 3.1(a)可以看出，從相同領域且資料量相近的語料庫，統計出來的 DLG 特徵值分佈，彼此只有些微差異，但是在圖 3.1(b)中，因爲不同領域的語料庫、資料量明顯的差異等原因，使得 DLG 特徵值分佈有明顯的差異。因此，使用 CKIP_Train 訓練出來的分類器，尚可被應用於分類 CKIP_Test 的語料；但是應用於跨領域的資料時，其統計分佈有明顯的差異，導致此分類器無法可靠地分類。所以我們針對 DLG 特徵值進一步做正規化，使其訓練資料與測試資料中的統計分佈，可以互相匹配。



(a) 相同領域分佈圖



(b) 跨領域分佈圖

圖 3.1 不同語料庫之 DLG 特徵值分佈比較圖

(二)、分佈正規化的法方介紹

訓練與測試特徵值分佈有明顯差異時，訓練出來的分類器就無法可靠地分類測試領域的資料。由於我們希望詞彙萃取技術可以用來探索未知的新領域，在新領域中特徵的統計分佈很可能不同於訓練領域，因此必須克服此問題。本論文分別爲使用標準差倍數法與直方圖均化法進一步正規化 DLG 數值。首先介紹標準差倍數正規化方法（Mean - Standard Deviation Weight, 簡稱 MSW），公式如下：

$$X_{d} = M_d + \sigma_d \left(\frac{X_s - M_s}{\sigma_s} \right) \qquad \text{（公式 3.1)}$$

d : destination 目標領域

S : source 來源領域

M_d：目標領域(訓練資料)特徵分佈中的平均數

M_S：來源領域(測試資料)特徵分佈中的平均數

σ_d：目標領域特徵分佈中的標準差

σ_S：來源領域特徵分佈中的標準差

X_S：來源領域的特徵值

X_d：轉換後的目標領域特徵值

標準差倍數正規化是一種線性調整的方法。轉換方式是以來源領域（測試資料）標準差爲衡量基準單位，計算特徵數值與其分佈平均值之間的正規化距離爲多少倍標準差，再換算成目標領域（訓練資料）的值。

接著介紹直方圖均化法（Histogram Equalization, 簡稱 HEQ）[9]，其轉換函式如下：

$$X_d = P(X_S) \cdot (X_{max} - X_{min}) + X_{min} \qquad \text{（公式 3.2）}$$

X_S：特徵值

X_d：均化後的數值

$P(X_S)$：特徵值之累積分佈函數(CDF)

$P_{EQ}(X)$：均化分佈之累積分佈函數（CDF）

X_{max}：特徵的最大值

X_{min}：特徵的最小值

圖 3.2 是 HEQ 正規化方法的示意圖，縱軸爲特徵 X 的累積分佈函數（Cumulative Distribution Function, CDF），橫軸爲特徵 X 的值。此轉換方式是利用累積分佈函數，將 X_S 轉換到均化分佈 $P_{EQ}(X)$上具有相同 CDF 值的特徵 X_d，也就是將 X 特徵的估測 CDF 分佈映射到線性 CDF 分佈的空間中。而線性 CDF 所對應的機率密度函數值（Probability Density Function, PDF)是一均勻分佈（uniform distribution），故稱爲「均化」。直方圖均化是一種單調（monotonic）的轉換方式，根據特徵值的資料量做非線性調整，調整後的數值能平均分佈於相同動態範圍中（X_{min} 到 X_{max} 之間）。使用 HEQ 正規化方法時，必須將訓練特徵和測試特徵都必須使用公式 3.2 進行均化，使統計分佈同時轉換至線性 CDF 分佈空間，讓彼此可以互相匹配，解決統計分佈差異造成分類器無法可靠地分類問題。

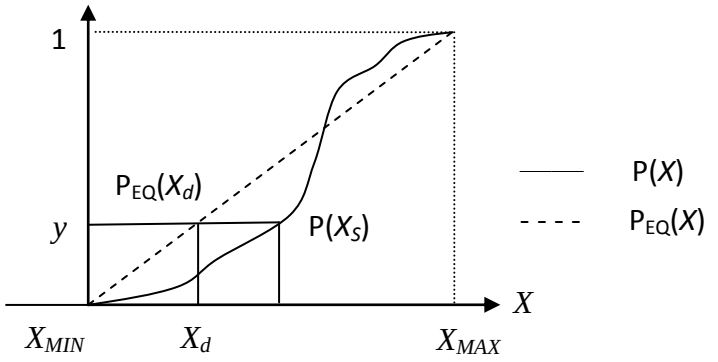


圖 3.2 HEQ 正規化方法示意圖

在進行標準差倍數正規化前，須先對訓練資料統計 M_d、σ_d 及對測試資料計算 M_S、σ_S。訓練資料計算出的 DLG 特徵不須做轉換，但測試資料之 DLG 特徵須根據公式 3.1，轉換至訓練資料的特徵空間。在進行直方圖均化正規化時，則是訓練資料和測試資料均須根據公式 3.2 轉換，但轉換所使用的 CDF 是分別由訓練資料與測試資料統計得到。所以標準差倍數法是將測試資料的特徵空間轉換至訓練資料的特徵空間；直方圖均化法則是將測試資料與訓練資料的特徵空間，同時轉換至線性的 CDF 分佈空間。

(三)、類神經分類器

詞彙組成的結構和使用方式非常複雜，單靠一種特徵值通常並不能做可靠的判斷，往往必須結合多種特徵，以發揮互補的功效。利用分類器可以彈性地結合不同的特徵，以達到更好的分類效果。因此，本論文中使用了多層次類神經網路分類器，進行詞彙的驗證。這是一種回歸的方法，可以實現非線性的分類。它的學習過程是以錯誤的倒傳遞方式，重複迭代網路權重值，使得總平方誤差最小化。

我們使用類神經網路分類器進行詞彙驗證的架構如圖 3.3 所示。以公式(2.1)~公式(2.5)

對所有候選詞計算 LogC、AV、Link、PreC、DLG，經公式(2.6)正規化，並且將 DLG 進一步做 HEQ 或 MSW 正規化，經特徵選取後作爲此分類器的輸入特徵。輸入特徵 $\boldsymbol{x}$ 可爲單一特徵或多維特徵。分類器的輸出 y 爲 0 到 1 之間的數值，經門檻測試，進行詞彙驗證，決策候選詞是否爲一詞彙。

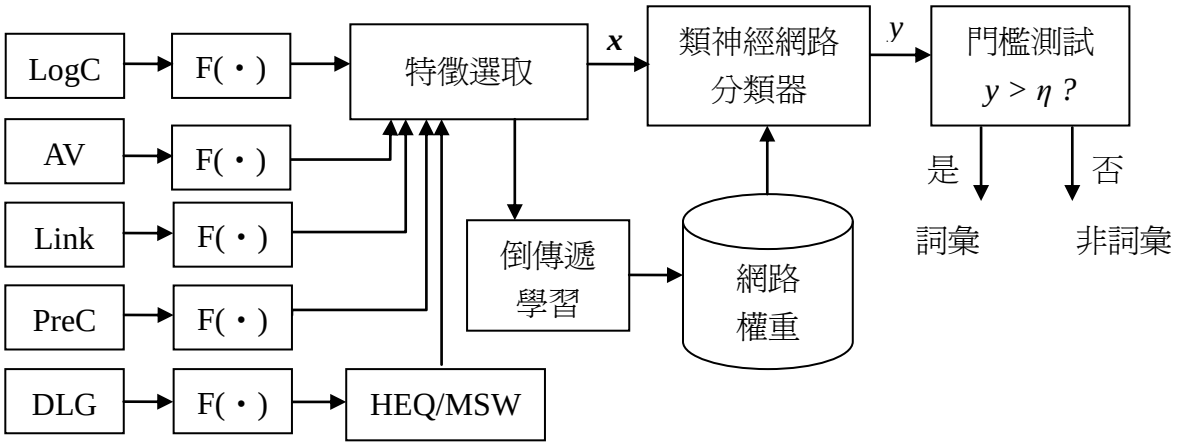


圖 3.3 應用類神經網路分類器於詞彙驗證架構

四、實驗分析

我們使用第三章第一節中所述的 CKIP_Train、CKIP_Test 及 HUCK_Test 語料庫進行實驗。這三個語料庫分別包含了 361,691、363,382、54,511 個句子。經過初步篩選候選詞分別得到 222,446、224,929、149160 個候選詞。並且利用個別語料庫斷詞文章中的詞彙，對於所有候選詞標記是否爲詞彙，所標記的詞彙數分別爲 33,429、33,661、22,913 個詞彙。關於實驗語料詳細資訊如表 4.1。

表 4.1 實驗語料詳細資訊表

語料庫簡稱	用途	句數	候選詞數	詞彙數
CKIP_Train	分類器訓練	361,691	222,446	33,429
CKIP_Test	相同領域的測試	363,382	224,929	33,661
HKCU_Test	不同領域的測試	54,511	149,160	22,913

首先，我們對於相同領域語料庫，進行四種以上的特徵組合實驗。以 CKIP_Train 訓練分類器，並以 CKIP_Test 進行測試。實驗結果如表 4.2。在表 4.2 中，ALL 代表使用了五種特徵值，其他則分別代表刪除其中一種特徵，例如：No_LogC 表示刪除 LogC 特徵，只使用 DLG、AV、Link 和 PreC 四種特徵。從表中可以看出，當結合 DLG、AV、Link、PreC 四種特徵時，會有最佳的偵測效能，其 F-Measure 可達到 60.03%。同時也表示了增加 LogC 特徵值並無法改善效能。雖然 LogC 是常被使用的特徵值，但是其他特徵的結合，足以取代掉 LogC 特徵。因此，在後續的實驗則不考慮 LogC 特徵值，只針對結合 DLG、AV、Link、PreC。

表 4.2 四種以上特徵之詞彙驗證效能表

特徵組合	F-Measure
No_LogC	60.03%
No_DLG	57.21%
No_AV	51.74%
No_Link	48.06%
No_PreC	53.19%
ALL	59.69%

接著，我們進行不同正規化方法的實驗。首先，我們對同領域的語料庫進行實驗，以 CKIP_Train 訓練分類器，以 CKIP_Test 進行測試。詞彙偵測效能改進效果如圖 4.1。圖中黑色實線曲線 No_Equ 是沒做正規化的效能曲線、灰色實線曲線 HEQ 表示針對 DLG 特徵值做完 HEQ 正規化後的效能曲線，灰色虛線曲線 MSW 表示針對 DLG 特徵值做完 MSW 正規化後的效能曲線。從圖中可以看到，對於 DLG 的特徵值做 HEQ 的調整後，可以使效能明顯提升，最佳的 F-Measure 從 60.02%上升至 68.43%。但經過 MSW 正規化後，其並沒有改進其效能。因爲相同領域的 DLG 統計分佈並無明顯差異，如圖 3.1(a)。

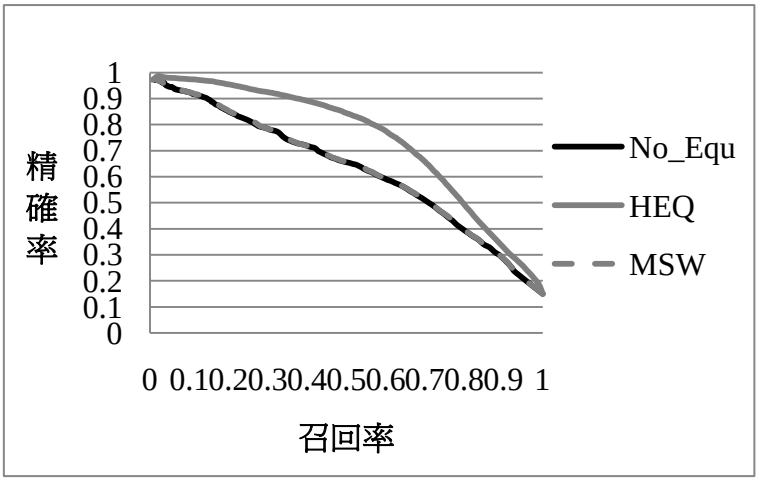


圖 4.1 相同領域之不同正規化方式的效能比較圖

接著，我們對跨領域語料庫進行實驗，以 CKIP_Train 訓練分類器，HKCU_Test 進行測試。實驗結果如圖 4.2。圖中的黑色實線曲線 No_Equ 是沒做正規化的效能曲線、灰色實線曲線 HEQ 表示針對 DLG 特徵值做完 HEQ 正規化後的效能曲線，灰色虛線曲線 MSW 表示針對 DLG 特徵值做完 MSW 正規化後的效能曲線。從圖中可以看出當 Recall 值較高時（門檻值 η 較低時)，經過 MSW 轉換的效能提升一些，這是因爲統計分佈差異，使得許多候選詞的決策數值 y（圖 3.3 中分類器產生的決策數值 y）偏小且皆相同，無法可靠分類的問題。所以經過正規化後，可以有效的識別這些候選詞。並且從圖中可以明顯的看出經過非線性 HEQ 轉換後，使得效能大幅的提升，最佳的 F-Measure 可達 71.40%。

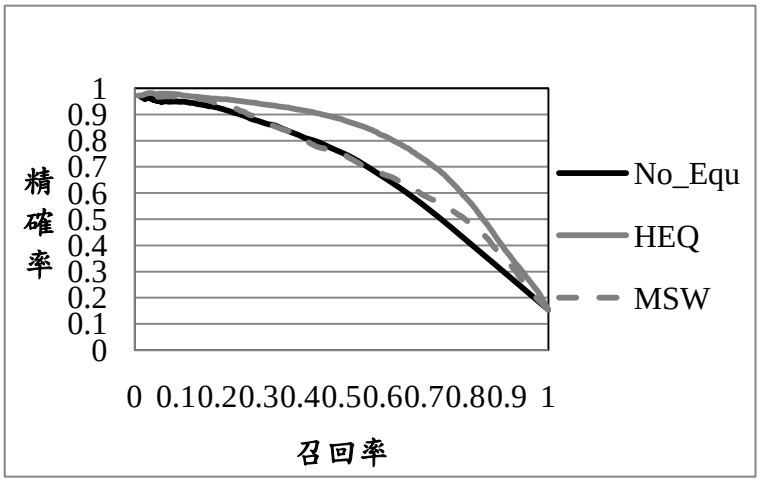


圖 4.2 跨領域之不同正規化方式的效能比較圖

根據上面結果的結果，我們發現 HEQ 可以讓訓練和測試的統計分佈互相匹配，應用在跨領域的資料，使得訓練模型更加一般化。不僅如此，也表示 DLG 特徵值適合利用 HEQ 對數值做非線性調整，使得 DLG 特徵值對於分類詞彙的特性也更加顯著，不論相同領域或跨領域測試，對於未知詞偵測效能的改進，均有顯著的效果。

五、應用於新穎領域的未知詞萃取

本章將探討前述方法應用於新穎領域的未知詞萃取。首先，我們從網路新聞網站中，以關鍵字「八八水災」搜尋 2009 年的相關新聞，當作新穎領域的測試語料庫（以下簡稱 UKW_Test），總共 32,207 句，依據詞頻及字元組長度初步篩選出候選詞集，總共有 81,447 個候選詞。接著針對各個候選詞計算 DLG、AV、Link、PreC 四種特徵數值，把特徵數值正規化到 0 至 1 之間，並且將 DLG 特徵值進行 HEQ 正規化，產生候選詞的特徵向量，在經過 CKIP_Train 語料庫訓練出來的詞彙萃取模型，輸出一個 0 至 1 的數值，此數值表示候選詞是一詞彙的可能性，依據此輸出數值將所有候選詞進行排序，接著篩選出前 10,000 個候選詞。

(一)、標記詞彙的方法

要對新穎領域分析詞彙萃取效能，最大的困難是在中文上詞彙並沒有共通的定義，不同的斷詞系統或詞典所使用的詞彙定義可能會有所出入，例如表 5.1 所示，此表列出部分 CKIP 與 HKCU 不同詞彙定義的例子，「詞彙」欄位表示對此字的組的詞彙定義，如在 CKIP 欄中「奶粉 錢」表示句子中的「奶粉錢」是由兩個詞彙「奶粉」與「錢」組成的；反之，在 HKCU 欄中，則認爲「奶粉錢」應該被視爲單一個詞彙。因此，我們的策略是使用兩個線上斷詞系統所共同產生的詞彙作爲「確定的答案」，再輔以人工標記。首先，我們將 UKW_Test，分別透過兩套具有偵測新詞能力的繁體斷詞系統進行斷詞。這兩套系統爲 CKIP 與中國科學院計算技術所（Institute of Computing Technology Chinsee Academy of Science, ICTCAS）所提供的斷詞服務。之後，將斷詞後產生的詞彙，視爲

斷詞系統產生的詞典，即爲這兩套系統各別萃取出的詞彙。這兩個詞典分別稱爲 CKIP_Dic 與 ICTCAS_Dic。

表 5.1　CKIP 與 HKCU 不同詞彙定義的範例表

詞彙		句子	
Ckip	HKCU	Ckip	HKCU
奶粉 錢	奶粉錢	**奶粉錢**也有點需要	爲了賺**奶粉錢**和教育基金
別 無 選擇	別無選擇	那自然**別無選擇**	除此**別無選擇**
混 日子	混日子	懶懶散散的**混日子**	以做肉串**混日子**
身 陷	身陷	則可能**身陷**其中無法自拔	**身陷**逃兵醜聞的韓星宋承憲
紐約 市長	紐約市長	**紐約市長**魯迪	朱利安尼當上**紐約市長**後

接下來我們針對從 UKW_Test 語料庫中初步篩選的候選詞集，各別利用上述產生的兩個詞典(CKIP_Dic 與 ICTCAS_Dic)進行詞彙的標記，標記的結果如圖 5.1(a)所示。圖中右邊圓型表示候選詞集存在於 CKIP_Dic 的詞彙集合(簡稱 CKIP_Words)，總共包含 11,290 個詞彙。左邊圓型表示候選詞集存在於 ICTCAS_Dic 的詞彙集合（簡稱 ICTCAS_Words），總共包含 10,642 個詞彙。我們將這兩集合的交集共 9,802 個詞，視爲「確定的詞彙」。利用這些正確的詞彙，對本論文篩選出來的 10,000 個詞彙進行過濾，總共有 6,577 個相同詞彙，若以「確定答案」9,802 個詞而言，本論文的召回率約 67.09%。剩於的 3,423 個候選詞，是本系統挑選出來，但不在「確定詞彙」中的。我們進一步以人工的方式進行詞彙標記，標記出 1,179 個詞彙，如圖 5.1(b)所示。透過上述標記詞彙的方法，本論文的詞彙萃取方法總共萃取出 7,756 個詞彙，精確率約 77.56%。

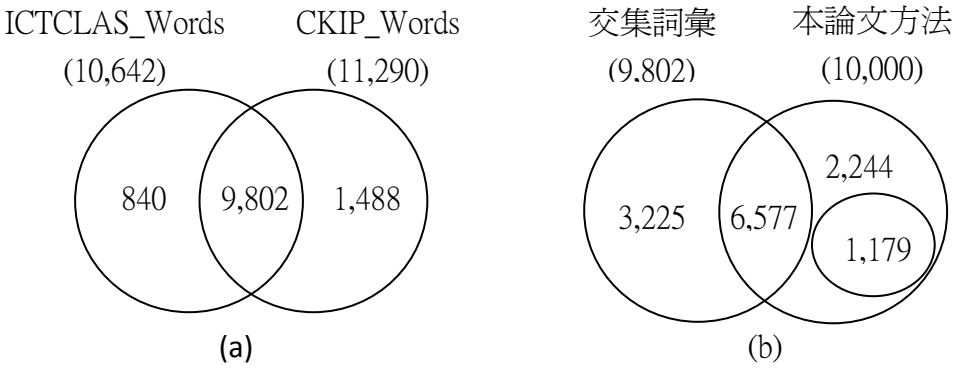


圖 5.1 新穎領域之詞彙標記方法示意圖

(二)、新穎領域之未知詞萃取分析

接下來我們對本論文詞彙萃取方法與兩套斷詞系統所萃取出來的未知詞，進行分析比較。本論文定義的未知詞即爲訓練語料庫（CKIP_Train 語料）中未出現的詞彙。因此我們將各方法萃取出來的詞彙與訓練語料庫的詞彙做比較，刪去訓練語料庫中已出現的詞彙，作爲各個方法萃取出的未知詞集合。如表 5.2 所示。本論文詞彙萃取方法篩選出 1,486 個未知詞，CKIP 斷詞系統篩選出 2,402 個未知詞，而 ICTCAS 斷詞系統篩選出 1,477

個未知詞。

表 5.2 各方法之萃取未知詞數表

方法	詞彙萃取數	未知詞數
本論文萃取方法	10,000	1,486
CKIP 斷詞系統	11,290	2,402
ICTCAS 斷詞系統	10,642	1,477

接著我們比較本論文方法與兩斷詞系統對於萃取未知詞的差異。圖如 5.2 所示。圖 5.4(a)中可看到，ICTCAS 斷詞系統抽出的未知詞集合，共有 1,477 個未知詞，而本論文抽取的 1,486 個詞中，有 567 個相同的詞彙。圖 5.2(b)中則看到 CKIP 斷詞系統抽出的未知詞集合，共有 2,404 個未知詞，本論方法抽出來的未知詞，有 911 個相同未知詞彙。

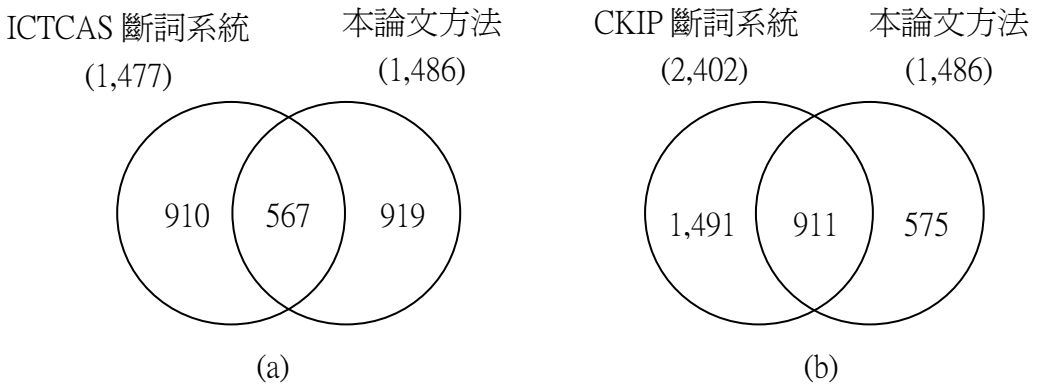


圖 5.2 未知詞差異示意圖

接著我們列出只有本論文方法有萃取出來，而其他兩套系統都沒有找到的幾個未知詞範例，如表 5.3 中的(a)所示。表 5.3(b)和表 5.3(c)則分別是 ICTCAS 斷詞系統與 CKIP 斷詞系統有找到的未知詞，而本論文方法並未找到的未知詞範例。從表 5.3(a)可以看出，有很多新產生出來的詞彙，例如：「蠟筆小新」、「金融海嘯」等專有名詞，其實很難透過語法的規則抽取出來，但它們具有明顯的統計特性，可以利用本論文的統計特徵抽取出來。CKIP 和 ICTCAS 斷詞系統結合了統計和語義的特性偵測未知詞，也未能抽出這些具的有統計特徵的未知詞，這顯示了本論文結合多種不同特性的統計特徵，的確可以更可靠地萃取出一些須依賴統計特徵的新穎詞彙。不過這兩套系統在特殊類別的未知詞抽取效能比本論文的方法佳，例如：人名、地名或是具有特殊文法結構的詞彙，如「蘇縣長」、「經發局」等。這是由於本論文是純粹以統計特徵爲主要萃取方法，並未使用任何文法規則。這樣的方法應該和以語法規則抽詞的架構產生很大的互補性。

表 5.3 三種方式的萃取未知詞範例

(a) 本論文方法

未知詞	
海角 7 號	功夫灌籃
小巨蛋	批踢踢
佳暮英雄	綠豆椪
蠟筆小新	焦糖哥哥
紙教堂	龍眼乾
語音信箱	金融海嘯
那瑪夏鄉	劍湖山

(b) ICTCAS 斷詞系統

未知詞	
陳添勝	新發大橋
林政助	二手衣
夢工場	簡志忠
南迴公路	消費券
光林村	梅山鄉
總執行長	泰武村
義賣品	馬總統

(c) CKIP 斷詞系統

未知詞	
救難隊	凱達格蘭
平安米	秀姑巒溪
馬政府	監察院長
蘇縣長	正大光明
秋節禮品	毀於一旦
頂呱呱	副駕駛
張瑞賢	經發局

六、緒論

本論文研究統計式的詞彙萃取方法，希望透過機器學習的方法，結合不同特性的統計特徵，萃取出未知的詞彙。適時更新自然語言處理系統中所使用的詞典，增進系統的處理效能。詞彙組成的結構和使用方式非常複雜，通常並不能單靠一種特徵值來做判斷，往往必須結合多種特徵。首先，我們針對四種以上的特徵組合，進行分析對於詞彙萃取的效能影響。實驗中顯示當結合DLG、AV、Link、PreC四種特徵值時，會有最佳的偵測效能。我們使用SIGHAN2競賽中的語料庫進行測試，對於中研院資訊所詞庫小組所提供的語料庫（CKIP_Test語料庫），其F-Measure的數值爲60.03%。

另外我們針對統計特徵值分佈可能會有不匹配的問題，提出了使用直方圖均化（Histogram Equalization）的正規化方法，使得測試與訓練特徵值分佈能互相匹配，解決語料庫大小或領域不同所造成特徵值範圍變動及分佈差異的問題。不必因爲領域的差異而重新訓練詞彙萃取模型，降低對於特定訓練語料庫的依賴性。使得本論文的詞彙萃取方法更具一般性，能針對各種不同領域，萃取出各個領域常使用的詞彙。對於中研院資訊所詞庫小組及香港城市大學所提供的語料庫（CKIP_Test語料庫、HCKU_Test語料庫），F-Measure分別可以達到68.43%和71.40%。同時我們也發現，針對DLG做直方圖均化，不論是在相同領域或跨領域的測試，均可以改進詞彙萃取的效能。

最後我們將詞彙萃取的方法應於用萃取新穎領域的未知詞，並與中央研究院資訊所詞庫小組和中國科學院計算技術所提供的斷詞系統抽取的未知詞進行分析比較，我們發現本論文方法與其兩套斷詞系統具有互補的特性，可以萃取出具有強烈的統計詞彙特性且難以透過語意的方式萃取出來的未知詞，例如：「海角7號」、「金融海嘯」等未知詞。

但是對於人名或地方名稱的未知詞萃取，則本論文方式萃取能力對於兩套斷詞系統而言顯得較不足。

參考文獻

[1] MS Sun, CN Hung, HY Gao, and JFang, "Identifying Chinese Name in Unrestricted Texts", *Chinese & Oriental Languages Information Processing Society*, 1994.

[2] Chunyu,Xueqiang Lu, Le Zhang, and Junfeng Hu, "Statistical Substring Reduction in Linear Time", *In Proceeding of the 1nd International Joint Conference on Natural Language Processing（IJCNLP）*, 2004.

[3] Zhao Hai and Kit Chunyu, "An Empirical Comparison of Goodness Measures for Unsupervised Chinese Word Segmentation with a Unified Framework", *In Proceedings of The 3nd International Joint Conference on Natural Language Processing（IJCNLP）*, 2008.

[4] Keh-Jiann Chen and Wei-Yun Ma, "Unknown Word Extraction for Chinese Documents", *In Proceedings of The 19nd International Conference on Computational Linguistics (COLING), Pages 169-175*, 2002.

[5] 梁婷, 葉大榮, 應用構詞法則與類神經網路於中文新詞萃取, *In Proceedings of Research on Computational Linguistics Conference XIII(ROCLING), Pages 21-40*, 2000.

[6] Goh Chooi Ling, Masayuki Asahara, and Yuji Matsumoto, "Chinese unknown word identification using character-based tagging and chunking", *In Proceedings of The 41nd Annual Meeting on Association for Computational Linguistics - Volume 2, Pages 197-200*, 2003.

[7] Chunyu Kit and Yorick Wilks, "Unsupervised Learning of Word Boundary with Description Length Gain", *In Proceedings of CoNLL99 ACL Workshop*, 1999.

[8] Haodi Feng, Kang Chen, Xiaotie Deng, and Weimin Zheng, "Accessor Variety Criteria for Chinese Word Extraction", *Computational Linguistics*, 2004.

[9] Luning Ji, Mantai Sum, Qin Lu, Wenjie Li, and Yirong Chen, "Chinese Terminology Extraction Using Window-Based Contextual Information", *In Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing, Pages: 62 – 74*, 2009

[10] Robert Hummel," Image enhancement by histogram transformation", *Comp. Graph. Image Process. , vol. 6, Pages 184-195*, 1977.