The 2017 Conference on Computational Linguistics and Speech Processing
ROCLING 2017, pp. 183-196

PTT 網站餐廳美食類別擷取之研究
A Study of Restaurant Information and Food Type Extraction from PTT

鍾智宇 Chih-Yu Chung
國立中央大學資訊工程學系
Department of Computer Science and Information Engineering
National Central University
tommychihyu@gmail.com

周建龍 Chien-Lung Chou
國立中央大學資訊工程學系
Department of Computer Science and Information Engineering
National Central University
formatc.chou@gmail.com

張嘉惠 Chia-Hui Chang
國立中央大學資訊工程學系
Department of Computer Science and Information Engineering
National Central University
chia@csie.ncu.edu.tw

摘要

隨著資訊科技與網際網路的快速發展，從自然語言中擷取所需資訊（Information Extraction）技術也愈顯重要，本研究希望針對國內最大的電子佈告欄系統 (BBS, Bulletin Board System) 「PTT」中的「Food」版發展出一套自動化擷取文章中餐廳相關資訊並判斷餐廳類別的方法，讓餐廳資訊的取得更加快速且便利。本文架構主要分為三個部分，第一部分為餐廳相關資訊擷取，透過 PTT Crawler 擷取 PTT Food 版上的文章進行格式化處理，並藉由關鍵字比對的方式擷取特定文章標題，以及正規表達式 (Regular Expression) 擷取內文包含的餐廳名稱、電話、地址及 URL 資訊。第二部分則是文章標題作為餐廳類別(例：咖啡、涮涮鍋、台式料理)的擷取來源，隨機挑選 10,000 筆標題資料針對隱含其中的餐廳類別進行人工標記；最後再透過 WIDM 實驗室研究室整合了條件式隨機域 (Conditional Random Field, CRF) 所開發的 WIDM NER TOOL 分別進行監督式學習與半監督式學習的實驗，並從實驗結果得知利用此法在餐廳類別的擷取可獲得不錯的效果。

Abstract

In this study, we hope to develop a system to automatically extract restaurant type from the FOOD board of PTT, the largest BBS web site in Taiwan. This paper is divided into three

parts. The first part is pre-processing, where we crawl articles from the PTT FOOD board and extract title、restaurant name、telephone、address and URL information via regular expressions. The second part is restaurant type labeling from title data. We used WIDM NER TOOL to train a model for restaurant type extraction. The last part of the article is experiment. We randomly selected 10,000 titles for manual labeling and testing. We used the labeled data for supervised learning and included unlabeled data for Semi-Supervised learning. Finally we got a good result using this method in restaurant type extraction.

關鍵詞：機器學習，Tri-Training，Distant Learning，命名實體辨識

Keywords: Machine Learning, Tri-Training, Distant Learning, Named Entity Recognition

一、 緒論

在資訊化技術及網際網路快速發展的今日，網路上豐富且大量的資料成為人們取得資訊的主要來源，其中美食餐廳資訊更是生活中不可或缺的部分，因此越來越多的美食評論網站、部落格…等隨之產生；然而這些資料來源的組成大多為知名或是大型的連鎖餐廳，無法涵蓋許多不具名但人氣度高的路邊攤小店，加上該類網站大多由商家透過程式設計者設計出帶有廣告性質的既定框架與內容，此外該類網站資料更新的頻率通常取決於管理者的維護頻度，因此資料更新的即時性往往跟不上消費者更新的速度。基於上述考量，本文以時下台灣最大的電子佈告欄系統 (Bulletin Board System, BBS) 「PTT 實業坊」作為研究的資料來源，希望設計出一套方法能自動擷取 PTT FOOD 版上不斷即時更新的文章內容，讓使用者能更快速便利得透過此方法獲取餐廳的相關資訊，並提供擷取後的資訊做為其他相關研究的興趣點(Point of Interest, POI)參考資料。

二、 相關研究

資訊擷取主要是從各種結構化與非結構化的文字中萃取出特定的資訊，而命名實體辨識 (Named Entity Recognition, NER)則屬於自然語言處理 (Natural Language Processing , NLP) 的領域。早期以 Rule Based Extraction Methods 為主要方向但耗費的成本通常較高，而機器學習法 (Machine Learning Based Method) [1,2] 是人工智慧的一個支領域且在資訊擷取領域廣泛的使用。目前大多數 NER 的相關研究方向以監督式學習為基礎，透過序列標記(Sequence Labeling)的方式來建立模型，主要可分為三種模型：第一種為隱藏式馬可夫模型 (Hidden Markov Model, HMM) [3]。第二種最大化熵馬可夫模型

(Maximum-Entropy Markov Model, MEMM) [4]，或稱為條件式馬可夫模型(Conditional Markov Model, CMM) [5]。最後一種則是條件式隨機域(Conditional Random Field, CRF)。

半監督式學習 (Semi-Supervised Learning) 所涉及的監督程度較小，常用於資料來源不易且昂貴時，常見的演算法有: Self-Learning [6]、Semi-Supervised- Support -Vector Machine (S^3VM) [7]、Co-Training [8]、Tri-Training[9] 等。關於半監督式學習的廣泛調查發現，並未有明確的實驗結果證明半監督式學習的效能優於監督式學習 [10]。

Co-Training 和 Tri-Training 在少量標記資料的分類相關研究上時常被提及。最早關於 Co-Training 的相關研究是由 Blum and Mitchell [8]所提出，而 Zhou[9]等人提出 Tri-Training 可視為 Co-Training 的改良，不同的地方在於 Tri-Training 使用三個分類器並且以投票 (Voting) 的機制來解決如何衡量分類記所標記的的答案可信度問題，並提供完整的數學證明與演算法，計算每回合自未標記資料 U 可取得的新資料量上限以及停止疊代條件。其實驗內容是比較在 12 個資料集執行 Tri-Training 的效果，首先對每個資料集取 25%的資料為測試資料，剩餘資料中再取 20%、40%、60%、80%的資料量為已標記資料 L 來訓練三個效能較差的分類器，再透過 Tri-Training 自剩餘資料的未標記資料 U 取得新的資料，這些新的資料再與 L 做聯集後重新訓練模型。透過此方式，經過數次疊代後可用較多的資料訓練並得到效能較好的模型。

Distant Supervision Learning 是透過啟發式規則 (Heuristics Rule) 所標記的小量資料或是小型的知識庫來訓練模型。舉例而言，Chou 等人[11]利用已知的 7,053 個人名於新聞網站收集 67,104 個句子，後續利用這些已知人名標記訓練資料，以及透過觀察而得的記者人名 Pattern 提升訓練資料品質，實驗結果 F-Measure 可達到 0.8689。

三、 設計與實作

本文欲擷取的資訊主要分成兩個部分，第一個部分是餐廳名稱、地址、電話、URL、標題等屬於半結構化的資料，可透過正規表達式擷取；第二部分為餐廳類別擷取，例如咖啡、義大利麵。

經由隨機挑選 1,000 筆 PTT 貼文(如圖 1)，經人工檢視並統計餐廳類別隱含於文章內文、標題、餐廳名稱的比例以決定擷取此項資訊的來源。經分析發現，雖然大部分的內文存

在著餐廳類別，但內文屬於敘述性文章，作者往往同時介紹多種該餐廳所販售的美食與特色，如此不具固定章法的文字敘述難以判斷何者為主要餐廳類別。例如一篇有關義大利麵館的文章，其內文可能同時提到其他食物或與其他餐廳的食物比較結果。因此，排除以內文作為擷取來源後，我們發現餐廳類別隱含在「文章標題」的比例為 72.5%，所以本研究以「文章標題」作為餐廳類別的擷取來源。

作者　zzzzzken (做正確的事)
標題　[食記] 台北萬華 現選現滷極神秘豬腳麵攤
時間　Wed Oct 18 16:29:38 2017

餐廳名稱：無名攤販
消費時間：2017年/10月
地址：台北市萬華區環河南路二段125巷7弄
電話：沒有電話
營業時間：0500AM ~ 1030AM
每人平均價位：110
可否刷卡：不行
有無包廂：沒有
推薦菜色：豬腳
官網：沒有

超級神秘的40年老攤販，在巷弄裡，還只賣早上，到底是何等運氣我才不小心發現，當時6年前我只是想找個豆漿喝，卻讓我念念不忘，想再訪又老是睡過頭。

圖 1 PTT 文章範例

本研究的系統架構圖如圖 2 所示，包含相關資訊擷取(3.1 節)、餐廳類別擷取(3.2)，實驗與效能評估(第四章)三大模組，各模組細節請參考對應章節。

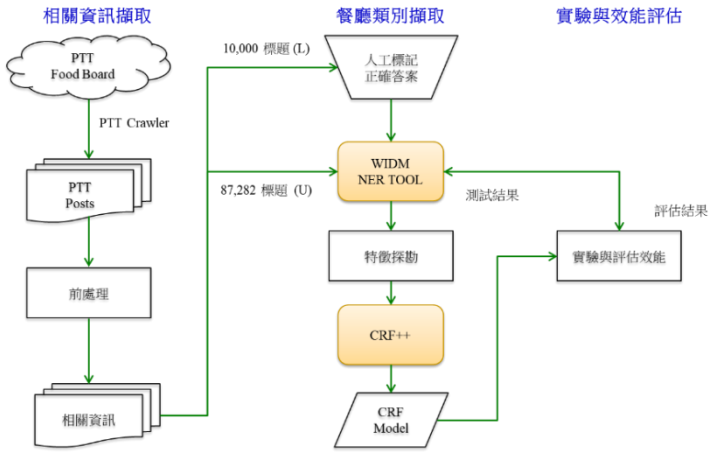


圖 2　系統架構圖

3.1　相關資訊擷取

相關資訊擷取包含三個程序，依序是網頁資料蒐集、資料前處理、餐廳相關資訊擷取。首先透過 PTT Crawler 將 PTT FOOD 版上的文章逐一下載後儲存於資料庫。第二部分則

為資料前處理，主要目的是去除雜訊並濾除無關的文章。第三部分為餐廳相關資訊擷取，此部分所擷取的資料屬於半結構化資訊，可透過資料分析獲取相關規則，再依規則撰寫正規表達式進行擷取，相關資訊包含了文章的「餐廳名稱」、「地址」、「電話」、「URL」以及「文章標題」。除少數文章內文電話格式不正確或根本缺少該項資訊外，大部份的相關資訊擷取不易出錯。

3.2 餐廳類別擷取

我們先針對「餐廳類別」加以定義；本文將「餐廳類別」定義為具有獨立意義且可讓消費者辨識餐廳類型的名稱，主要可分為兩大類：「可成為獨立類別的餐廳特徵」，例如台式料理、無菜單料理；或是「餐廳的主要販售商品」，例如咖啡、義大利麵。

但在人工標記檢視文章標題並標記答案時遇到許多模擬兩可的情況，例如當標題包含「日式涮涮鍋」時，其正確餐廳類別應為「日式涮涮鍋」、「涮涮鍋」還是「鍋」？為了解決此一問題，我們使用 CKIP [12]對大量文章標題斷詞並取得詞性，以斷詞結果決定那一些詞要不要再細分為更細的詞，而且經分析詞性後發現大部分的餐廳類別可由前綴詞(例：炒、牛肉、日式)、後綴詞(例：飯、麵、火鍋)與獨立美食(本身就是一種食物類別) 所組成，其中前後綴詞大多是由形容詞(A)或名詞(N)所構成，獨立美食則以名詞(N)為主，相關範例如下：

- 「前綴詞」＋「獨立美食」: 日式(A) + 烏龍麵(N)
- 「獨立美食」＋「獨立美食」: 咖哩(A) + 烏龍麵(N)
- 「獨立美食」＋「後綴詞」: 老虎(N) + 麵食館(N)
- 單一「獨立美食」: 鐵板燒(N)

因此在人工標記答案時我們透過 CKIP 取得斷詞結果與詞性作為人工標記之參考。例如原始文章的標題為「東海商圈－炒日式拉麵館」，此時「日式拉麵」、「拉麵」、「麵」都可作為餐廳類別，參考 CKIP 斷詞後的結果:「東海(N) 商圈(N) －(DET) 炒(Vt) 日式(A) 拉麵館(N)」並以最長詞彙優先標記的餐廳類別為：日式拉麵館。

3.3 特徵探勘與擷取

以 NER 的角度來看，我們欲擷取文章中的餐廳類別即為命名實體 (Named Entity)，在

3.2 節中，依據 CKIP 的分析結果得知文章標題具有一定的規則，亦即實體前後的詞是有規則的，同時實體本身也是有規律存在。因此我們應用 WIDM NER Tool 所設計 14 個特徵辨識標題中的實體，包括實體前後的詞(長度 1-3)，以及實體本身的前綴詞、後綴詞(長度 1-3)，再加上英文/數字以及符號，細節請參考表 1。

以常見實體的前綴詞為例，假設欲辨識的是人名，那麼前綴詞就是常見的姓(例：陳、林)，這一些常見的姓可自公開資料中取得 (例：榜單)，但本研究欲辨識的目標為餐廳類別，所以收集這些資料是有困難的。因此特徵探勘(Feature Mining)目的是分析訓練資料後擷取隱含於訓練資料中，實體前後方的詞及實體的前後綴詞，以作為餐廳類別辨識的辭典項目(Dictionary Terms)。

表 1 特徵值設計

ID	特徵	說明	長度	範例
1	Common Before_1	常見於實體前方的詞	1	的、式、大、小
2	Common Before_2	常見於實體前方的詞	2	手工、創意、經典
3	Common Before_3	常見於實體前方的詞	3	好吃的、古早味
4	Entity Prefix_1	常見實體的前綴詞	1	素、茶、烤、乾
5	Entity Prefix_2	常見實體的前綴詞	2	日本、日式、泰式、港式
6	Entity Prefix_3	常見實體的前綴詞	3	義大利、無國界
7	Entity Suffix_1	常見實體的後綴詞	1	鍋、粥、羹、凍
8	Entity Suffix_2	常見實體的後綴詞	2	料理、飲茶、火鍋
9	Entity Suffix_3	常見實體的後綴詞	3	自助餐、吃到飽
10	Common After_1	常見於實體後方的詞	1	館、屋、廳、亭、店
11	Common After_2	常見於實體後方的詞	2	大王、餐廳
12	Common After_3	常見於實體後方的詞	3	專賣店、專門店
13	English/ Number	數字與英文字母的混合	1	「A」、「F-15」
14	Symbol	半形或全形符號	1	「，」、「。」、「：」

我們分別針對辭典項目出現的「頻率 (Support)」以及「置信度 (Confidence)」作為自訓練資料中擷取項目的方法。以 Common Before 特徵為例，所謂「Support」意指該項目於訓練資料中出現於 Entity (餐廳類別)前方的次數，而 Confidence 定義為「出現於 Entity 前方的次數/該項目於訓練資料的總出現次數」。

舉例而言：「涮」此一項目出現於食物類別前的次數有 65 次，但是在整體訓練資料出現了 130 次，亦即此項目後連接一個 Entity 的機率只有 50%；但「炒」此項目雖然出現在

食物類別中的次數僅有 15 次，但在訓練資料中也是出現了 15 次，意即此項目後連接一個 Entity 的機率達 100%，代表這些項目出現的次數雖少，但只要一出現幾乎後方都會出現欲辨識的 Entity。所以若僅用次數作為挑選條件，容易忽略這些次數低但具有高度鑑別力的項目。

因此，我們先分析訓練資料，依據表 1 的設計，統計所有不同長度辭典項目(Dictionary Terms)的出現次數，以及該項目於訓練資料中出現的總次數。依據這兩個數據即可計算每個項目的 Support 與 Confidence 值，並輸出辭典項目。

- By Support：將 Support 由大到小排序後，輸出 Top n 項目
- By Confidence：將 Confidence 由大到小排序，輸出 Top n 項目

本研究使用 BISEO 標記法分別根據符號出現的位置標誌分為：開始 B (Beginning)、中間 I (Intermediate)、結尾 E (End)、獨立類別 S (Single) 以及不屬於餐廳類別的符號 O。

本研究使用 CRF 做為序列標記的模型，並採用 Taku Kudo 發展 CRF++ [14]工具將餐廳類別擷取問題轉換成序列標誌問題(Sequence Labeling Problem)，標題中的每一個符號有其對應的標記 (Label)，對此我們採用 BISEO 標記法分別根據符號出現的位置標誌分為：開始 B (Beginning)、中間 I (Intermediate)、結尾 E (End)、獨立類別 S (Single) 以及不屬於餐廳類別的符號 O (Other) 等五種類別，範例請參考下圖 3。

本研究使用中央大學資工系 WIDM 實驗室開發的 WIDM NER TOOL [13] 來實作，其結合了 [特徵擷取]、[辭典分析]、[訓練模型]、[實驗測試]...等功能，此工具可直接呼叫 CRF++作為訓練模型的工具。

「	0	0	0	0	0	0	0	0	0	0	0	0	0	1	O
桃	0	0	0	0	0	0	0	0	0	0	0	0	0	0	O
園	0	0	0	0	0	0	0	0	0	0	0	0	0	0	O
」	0	0	0	0	0	0	0	0	0	0	0	0	0	1	O
大	0	0	0	0	0	0	0	0	0	0	0	0	0	0	O
倉	0	0	0	0	0	0	0	0	0	0	0	0	0	0	O
日	0	0	0	1	1	1	0	0	0	0	0	0	0	0	B
本	0	0	0	0	0	0	0	0	1	0	0	0	0	0	I
料	0	0	0	0	0	0	0	1	0	0	0	0	0	0	I
理	0	0	0	0	0	0	1	0	0	0	0	0	0	0	E

圖 3 特徵矩陣範例

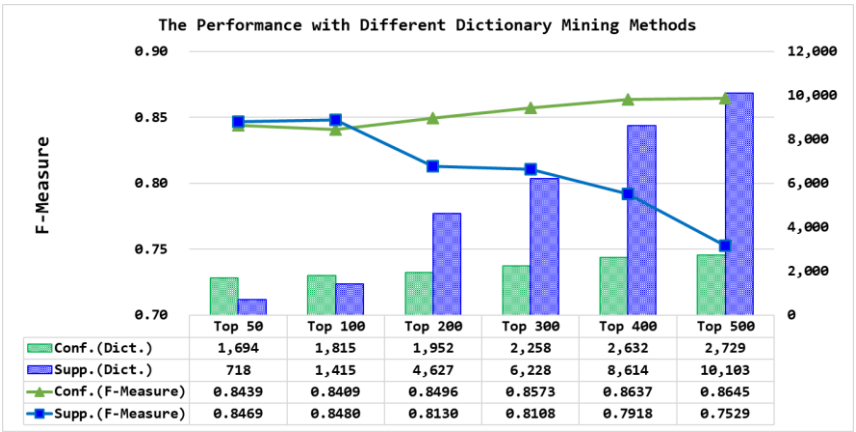
四、 實驗與效能評估

本研究使用人工標記答案的 10,000 筆資料 (L) 及未標記的 87,282 筆資料 (U) 作為基礎進行實驗；主要分成三個部份，特徵探勘(Features Mining)、監督式學習、半監督式學習三個部分，並以 5-fold cross validation 方式完成所有實驗。

第一個部分以監督式學習的方式搭配 10,000 筆已標記資料進行特徵探勘用以決定選取辭典項目的方式與參數。第二部分則進行監督式學習的基礎模型 (Basic Model) 實驗，最後半監督式學習實驗的部分包含(1)以 Basic Model 為基礎，搭配未標記資料 U (Un-Labeled Data) 進行 Tri-Training 實驗，用以測試加入自 U 選取的新訓練資料對系統效能提升的程度，以及(2)利用 L 中人工標記的餐廳類別作為已知實體(Known Entities，a.k.a. Seeds) 並依其出現的次數由高而低排序後進行 Distant Learning 實驗。

4.1 特徵探勘

首先以 10,000 筆已標記的資料 (L) 分成五等分取其中 8,000 筆作為訓練資料，2000 筆作為測試資料 (5-Fold CV)，將 By Confidence 與 By Support 兩種實驗方法所得的效能 繪成圖表 (如下圖 4) ，By Confidence 在 取 Top 500 的項目 時所得的 F-Measure 達到最高 0.8645，故選定以 By Confidence Top 500 作為選取辭典項目的方式進行後續實驗。



	Top 50	Top 100	Top 200	Top 300	Top 400	Top 500
Conf.(Dict.)	1,694	1,815	1,952	2,258	2,632	2,729
Supp.(Dict.)	718	1,415	4,627	6,228	8,614	10,103
Conf.(F-Measure)	0.8439	0.8409	0.8496	0.8573	0.8637	0.8645
Supp.(F-Measure)	0.8469	0.8480	0.8130	0.8108	0.7918	0.7529

圖 4 The Performance with Different Dictionary Mining Methods

4.2 監督式學習

完成辭典探勘實驗後，以 By Confidence Top 500 自 10,000 筆已標記資料 (L)擷取適當的辭典項目後，以 5-Fold CV 的方式進行學習曲線(Learning Curve)實驗，結果如圖 5 中

顯示，當訓練資料量達 8,000 筆時 F-Measure 可達 0.8645。

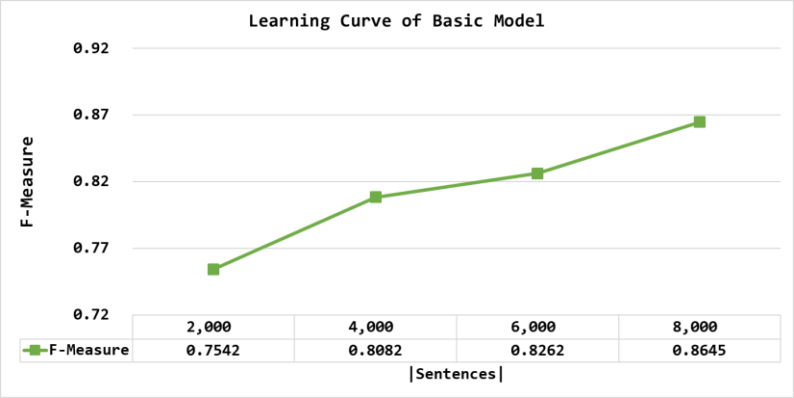


圖 5 Learning Curve of Basic Model

4.3 半監督式學習

4.3.1 Tri-Training

依照 4.2 節的實驗，我們自 PTT FOOD 版收集的未標記資料(U)共 87,282 筆，透過 Tri-Training 每一回合自 U 中選取部分資料，再利用已訓練完成的 NER 模型標記答案，這些新資料答案可能不正確(即包含 Noise)，隨著每回合加入更多的新訓練資料提升系統效能。為避免實驗數據受 U 中離群值 (Outliers) 影響，因此在每一次的 5-Fold-CV 中均再進行 5 次的 Tri-Training 並取其平均，共執行 5*5 = 25 次 Tri-Training，流程如圖 6。

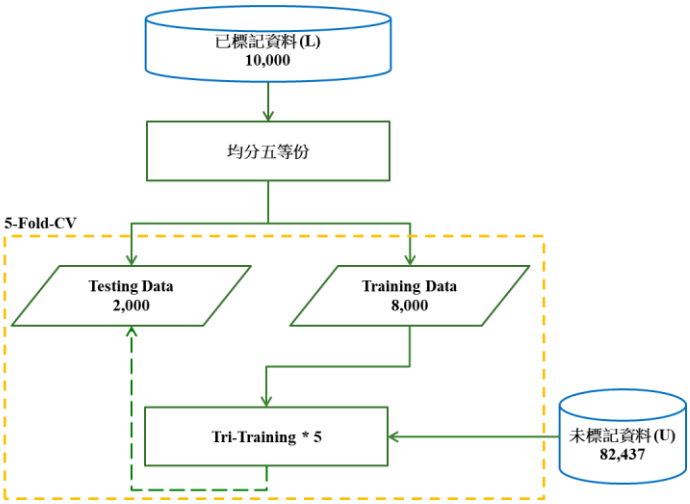


圖 6 Tri-Training 實驗流程圖

Tri-Training 的實驗結果如圖 7。從實驗數據中可發現，Tri-Training 平均所使用的訓練資料量 17,684 是 Basic Model 訓練資料量的 2.2 倍，而其平均 F-Measure 為 0.8685 較 Basic Model 的 0.8645 微幅提升 0.0040，結果顯示雖然 Tri-Training 效能僅微幅提升，但這也

達到 Tri-Training 從龐大的 U 中獲取新訓練資料來提升 NER 效能的目的。

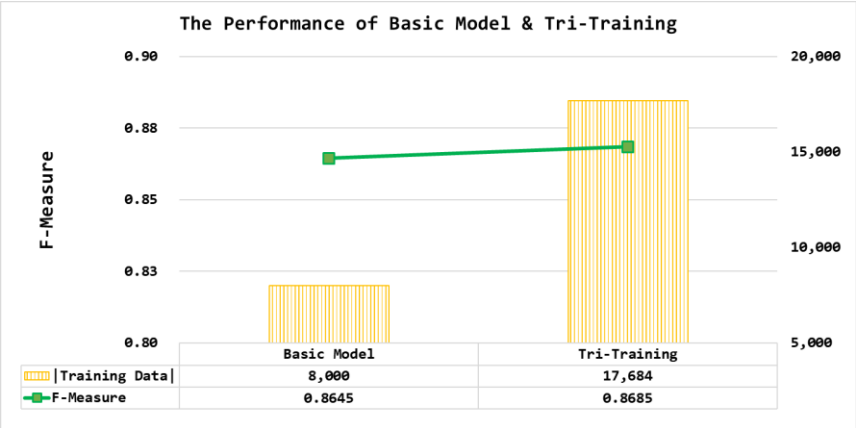


圖 7 The Performance of Basic Model & Tri-Training

4.3.2 Distant Learning

我們以 Basic Model 為基準，以 L 所包含的 1,563 個 Entities 為種子 (Seeds) 進行半監督式 Distant Learning 實驗，藉以測試以 Seeds 自動標記資料取代人工標記的實驗結果，實驗流程如下圖 8 所示。

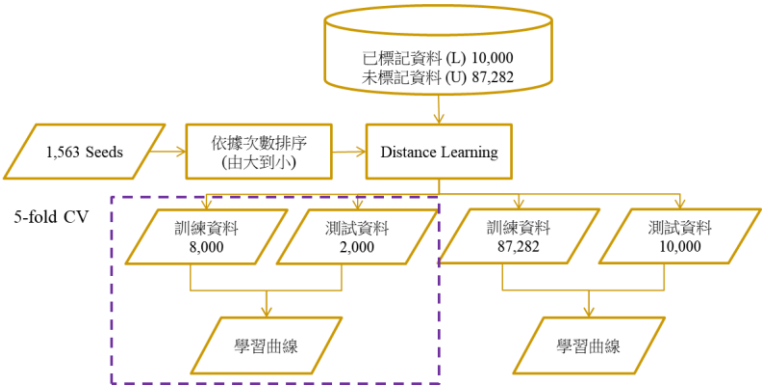


圖 8 Distant Learning 實驗流程圖

首先統計各 Seed 的出現次數並依照降冪排序後以向下累積的方式整理如下圖 9，圖中顯示 1,563 個 Seeds 於 L 中總出現次數為 7,961 次，而前 500 個 Seed 的累積出現次數已達 6,738 次，亦即前 31.9% 的 Seed 其出現次數占了總出現次數的 84.6%，由此顯示利用出現次數較高 (可視為較熱門) 的前幾個 Seed 即可完成相當大比例的資料標記，因此後續實驗的進行將依熱門程度排序後的 Seed 出現次數為組距，依順序由高至低進行自動標記，再將實驗結果依組距區分後繪製成學習曲線，後續觀察使用不同 Seed 量對系統效能的影響，最後再將實驗結果與監督式學習的 Basic Model 及 Tri-Training 實驗進行效能比較。

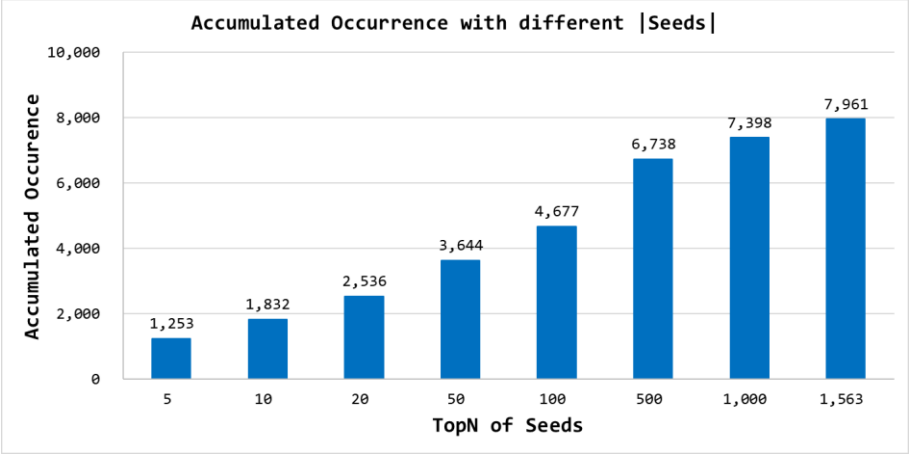


圖 9 Seeds 出現次數累計圖

實驗的第一個部分自 L 取 8,000 筆資料並移除答案後以 Seeds 標記後作為訓練資料，剩餘的 2,000 筆資料作為測試資料進行 5-Fold-CV 實驗，實驗結果取 F-Measure 的平均值整理如下圖 10。從結果可見在使用所有的 Seeds 時可於 8,000 筆資料中標記 5,958 筆最少包含一個 Entity 的資料 (positive example)，並移除未標記任何 Entity 的資料(negative example)後，F-Measure 為 0.8387，與人工標記進行 Basic Model 實驗的結果 0.8645 差距 0.0258; 由此可見利用 Distant Learning 的方式進行自動標記的結果其效能雖然略低於人工標記，但卻可大幅節省人工標記所需耗費的時間與成本。

實驗的第二個部分同樣依各組別的 Seeds 量對 87,282 筆未標記資料 (U) 進行自動標記後僅以 positive examples 作為訓練資料，實驗結果如下圖 11，其 F-Measure 在使用所有 Seeds 時可達到 F-Measure 0.8702，相較人工標記的結果，效能略微增加。由此顯示可透過 Distant Learning 標記大量訓練資料時，甚至可達到更好的效能。

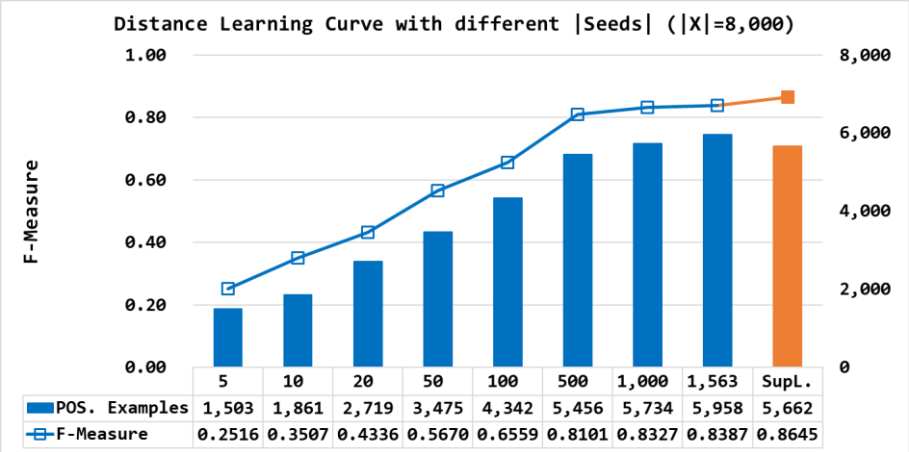


圖 10 Distance Learning Curve with different |Seeds| (|X|=8,000)

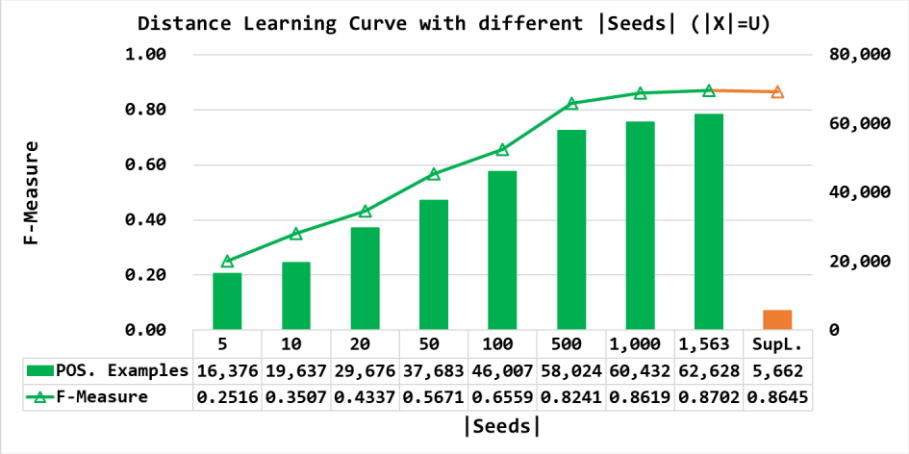


圖 11 Distance Learning Curve with different |Seeds| (|X|=U)

4.4 討論

首先我們利用 L 的 10,000 筆資料訓練 Basic Model，因為 L 的答案是經由人工標記的，因此可視為 F-Measure 0.8645 為這一份資料的效能上限(Upper Bound)，若希望再提升效能，提供更多的訓練資料是其中一個可行方法。

首先我們使用 Tri-Training 來達成此一目的，使用 L 訓練 Basic Model，再自 U 中挑選新的資料並使用 Basic Model 標記答案，此即為新的訓練資料。再以新的訓練資料重新訓練模型，此步驟重複執行數次，總訓練資料量亦隨之增加，但 Tri-Training 為了防止一個回合新增遠超過|L|的新增資料，因而陷入新增許多 Noise 的風險，故 Tri-Training 限定每一回合新增的訓練資料為一個|L|，依據我們的實驗數據得知，當 Tri-Training 停止時，總資料量將略少於|L|*2。

另外一個方法為 Distant Learning，因為是利用 Seeds 來標記訓練資料，所以必定會出現標記錯誤或未標記的 Noise，故需要收集大量的未標記資料以及 Seeds 來掩護 Noise 的負面影響。此方法優點是原理簡單，缺點是資料量以及 Seeds 的數量要求較高，為了降低 Noise 的影響，需一直收集新的資料以及 Seeds。若新資料/Seeds 的收集若沒有困難，或已知標記的訓練資料可以透過 Pattern 減少 Noise，則 Distant Learning 可說是一個簡單而又有效的方法。

整體而言，Tri-Training 僅需利用少數、品質高的訓練資料才可訓練效能不太差的 Basic Model，若品質太差則利用模型標記的答案則大多數是 Noise，則效能還是無法提升，

因此優點是訓練資料量的要求較低，因新增的資料是透過 NER 模型標記，故品質通常較 Distant Learning 為佳，故可使用較少的資料量得到效能差異不大的模型，缺點是每一回合需花費大量時間重複訓練模型。

由之前實驗結果得知 Tri-Training 共使用 17,684 筆資料其 F-Measure 為 0.8685，而 Distant Learning 使用 62,628 筆資料其 F-Measure 為 0.8702，總訓練資料量差異 3.54 倍，此結果正與上列兩個半監督學習方法的比較相符。

五、 結論

美食資訊與日常生活息息相關，建立一個即時且完整的 POI 資料庫讓使用者能更便利查詢在這波行動潮流中有著重要的地位。本文以 WIDM 實驗室所開發的 WIDM NER TOOL 結合 CRF++ Package 直接針對 PTT 內非結構化文章的簡短標題進行餐廳類別擷取，其結果可作為相關資訊系統、行動 App 的資料來源，未來亦可結合意見探勘 (Opinion Mining) 對該文章的內容分析，自動化給予該文章所提及的餐廳與美食評分，即可得到更完整的美食餐廳相關資訊資料庫。

參考文獻

[1] Dayne Freitag: Information Extraction from HTML: Application of a General Machine Learning Approach. AAAI/IAAI 1998: 517-523.

[2] Thomas G. Dietterich: Machine Learning for Sequential Data: A Review. SSPR/SPR 2002: 15-30.

[3] L. Satish and B.I. Gururaj. 1993. Use of hidden Markov models for partial discharge pattern classification. Electrical Insulation, IEEE Transactions on 28, 2 (Apr 1993), 172–182.

[4] Gideon S. Mann and Andrew McCallum. 2010. Generalized Expectation Criteria for Semi-Supervised Learning with Weakly Labeled Data. J. Mach. Learn. Res. 11 (March 2010), 955–984.

[5] Andrew Eliot Borthwick. 1999. A Maximum Entropy Approach to Named Entity Recognition. Ph.D. Dissertation. New York, NY, USA. Advisor(s) Grishman, Ralph.

AAI9945252.

[6] Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning Subjective Nouns Using Extraction Pattern Bootstrapping. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (CONLL'03). Association for Computational Linguistics, Stroudsburg, PA, USA, 25–32.

[7] Kristin P. Bennett and Ayhan Demiriz. 1999. Semi-supervised Support Vector Machines. In Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II. MIT Press, Cambridge, MA, USA, 368–374.

[8] Avrim Blum and Tom Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-training. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT' 98). ACM, New York, NY, USA, 92–100.

[9] Zhi-Hua Zhou and Ming Li. 2005. Tri-Training: Exploiting Unlabeled Data Using Three Classifiers. IEEE Trans. on Knowl. and Data Eng. 17, 11 (Nov. 2005), 1529–1541.

[10] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. 2006. Semi-Supervised Learning.

[11] Chien-Lung Chou and Chia-Hui Chang and Ya-Yun Huang, " Boosted Web Named Entity Recognition via Tri-Training", ACM Trans. Asian Low-Resour. Lang. Inf. Process. , Vol 16, pp. 10:1--10:23, December 2016.

[12] Ma, W.-Y., Chen, K.-J.: Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing - Volume 17, pp. 168-171. Association for Computational Linguistics, (2003)

[13] CRF++: Yet Another CRF toolkit：http://crfpp.sourceforge.net/