國 立 中 央 大 學

資 訊 工 程 學 系

碩 士 論 文

PTT 網站餐廳美食類別擷取之研究

研 究 生：鍾智宇

指導教授：張嘉惠 博士

中 華 民 國 106 年 6 月

國立中央大學圖書館
碩博士論文電子檔授權書

(104 年 5 月最新修正版)

本授權書授權本人撰寫之碩/博士學位論文全文電子檔(不包含紙本、詳備註 1 說明)，在「國立中央大學圖書館博碩士論文系統」。(以下請擇一勾選)

(V)同意 (立即開放)

(　)同意 (請於西元 _____年_____月_____日開放)

(　)不同意，原因是：_____

在國家圖書館「臺灣博碩士論文知識加值系統」

(V)同意 (立即開放)

(　)同意 (請於西元 _____年_____月_____日開放)

(　)不同意，原因是：_____

以非專屬、無償授權國立中央大學、台灣聯合大學系統圖書館與國家圖書館，基於推動「資源共享、互惠合作」之理念，於回饋社會與學術研究之目的，得不限地域、時間與次數，以紙本、微縮、光碟及其它各種方法將上列論文收錄、重製、與利用，並得將數位化之上列論文與論文電子檔以上載網路方式，提供讀者基於個人非營利性質之線上檢索、閱覽、下載或列印。

研究生簽名: 鍾 智 宇　　學號：995302023

論文名稱: PTT 網站餐廳美食類別擷取之研究

指導教授姓名：張嘉惠 博士

系所 ： 資訊工程 所 □博士班 ■碩士班

填單日期： 2017/7/24

備註：

1. 本授權書之授權範圍僅限**電子檔**，紙本論文部分依著作權法第 15 條第 3 款之規定，採推定原則即預設同意圖書館得公開上架閱覽，如您有申請專利或投稿等考量，不同意紙本上架陳列，須另行加填申請書，詳細說明與紙本申請書下載請至本館數位博碩論文網頁。
2. 本授權書請填寫並**親筆**簽名後，裝訂於各紙本論文封面後之次頁（全文電子檔內之授權書簽名，可用電腦打字代替）。
3. 讀者基於個人非營利性質之線上檢索、閱覽、下載或列印上列論文，應遵守著作權法規定。

摘要

隨著資訊科技與網際網路的快速發展加上行動裝置日漸普及化，從網路上獲取生活所需的資訊已成為趨勢主流，然而該如何從豐富且多樣化的大量資料中有效擷取有用的資訊成為一項重大的挑戰，因此資訊擷取（Information Extraction）技術逐漸成為熱門的研究議題，其內容主要是透過整理、篩選…等步驟將非結構化的資料加以整合成為結構化的資料，最後從中有效得擷取出有用的資訊。本研究希望透過資訊擷取技術中機器學習 (Machine Learning) 的方法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 版上的文章資訊存入資料庫中進行格式化處理，並以人工分析的方式瞭解資料的概貌，接著藉由關鍵字搜尋的方式掃描文章以擷取文章標題、餐廳名稱、電話、地址及 URL 資訊。第二部分則是進行餐廳類別擷取，藉由前處理作業時分析資料的結果得知 72.5% 的餐廳類別隱含在文章的標題中，因此以文章標題作為餐廳類別的擷取來源，透過 CKIP 系統進行斷詞後參考其結果隨機挑選 10,000 筆標題資料針對隱含其中的餐廳類別進行人工標記；最後再將標記後的資料透過 WIDM 研究室整合了條件式隨機域 (CRF, Conditional Random Field) 所開發的 WIDM_NER_TOOL 搭配 BIESO 標記法訓練模型。最後則是將標題資料輸入訓練好的模型後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

With the rapid development of Internet information technology and the popularity of mobile devices, access to information from web pages has become a trend, but how to extract useful information from rich and diverse information becomes a major challenge. The development of information extraction technology has gradually become a popular research topic, its main purpose is through the sorting、screening, unstructured information will be integrated into a structured data, and finally can effectively extract useful information. In this study, we hope to develop a system to automatically extract restaurant type from the FOOD board of PTT of the largest BBS web site in Taiwan through the Machine Learning Method in information extraction technology, so that users can get more convenient and fast access restaurant information

This paper is divided into three parts, the first part is pre-processing, we extract the articles from the PTT FOOD site by the PTT Crawler and then format the data; based on the extracted articles, we analysis of the keyword by statistical from the article to extract the Title、Restaurant Name、Telephone、Address and URL information; The second part is restaurant type extraction; by pre-processing analysis, we know that 72.5% of the restaurant type was implied in the title; we segmented the extracted title data through the CKIP System, and then refer to the results for manual labeling. We used WIDM_NER_TOOL which bundled CRF++ package to train the labeled data and BISEO markers to train an extraction model, the input data are used to capture the restaurant type after the model's testing process. The last part of the article is experiment, we used the labeled data for supervised learning and used unlabeled data for Semi-Supervised to evaluate system performance. Finally we got a good result from experiment result that used this method in restaurant type extraction.

目錄

圖目錄

表目錄

一、 緒論

1-1 研究動機

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，快速更新的資訊正不斷地在網路中流動及累積，再加上行動上網裝置日漸普及，使用者依賴行動裝置上網查詢所需的資訊已成為一種新的趨勢。根據美國市調公司尼爾森所發表的 [2009 年全球網路消費者調查] 顯示，近 7 成的消費者相信網路上的評價與建議。而根據 Google 2016 年公布的 「消費者洞察報告」指出 96% 的台灣人每天上網，此比率在亞太區中僅次於第一名的香港（97%），而台灣智慧型手機滲透率也創 82% 新高，其中 25 到 34 歲的年輕族群普及率更達 100%，由此可見網路資訊已逐漸成為消費者重要的參考訊息來源，其中美食餐廳資訊更是生活中不可或缺的部分，根據 OpenTable [2015 年科技與外出就餐調查] 顯示，87% 的消費者在外出用餐前會透過行動裝置上網搜尋餐廳；另一份由 Google 公布的 [臺灣與亞洲行動網路及使用者行為調查報告] 更指出，智慧型手機使用者最常搜尋的內容依序為產品資訊 60%，餐廳、酒館和酒吧 51%，旅遊 49%，工作機會 29% 以及購屋、租屋資訊 28%，由此可見超過半數的消費者越來越習慣透過網路尋找美食做為用餐的參考，也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，因此評論相對較不具客觀性；此外，該類網站資料更新的頻率通常取決於特定管理者的維護頻度，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 版上由使用者以不具語法規則的自然語言所撰寫且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 POI 參考資料。

1-2 研究背景與限制

「批踢踢實業坊」 (以下簡稱 PTT)（https://www.ptt.cc）是時下台灣最大的 BBS 網站，總註冊人數超過 150 萬，尖峰時段超過 15 萬人同時上線，每日超過 2 萬筆新文章以及超過 50 萬筆的推文被發表，由此可見其擁有龐大且即時的資訊；然而，站內的文章並沒有嚴格的編排規則與語法限制，文章標題及內容大多為作者透過口語化形式直覺得以自然語言所撰寫，因此要達到純文字的自動化資訊擷取存在相當大的困難度。

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「可成為獨立類別的餐廳特徵」或是「餐廳的主要販售商品」兩大類，針對此問題傳統自然語言的處理方式採用人工撰寫擷取規則程式，或是建立辭典撰寫程式進行比對，此法雖然可從比對結果擷取所需的資訊，但因中文美食命名並無一定的規則，加上數種美食名稱加以組合後又可成為一種獨立的美食種類，例如: 「咖哩」與「餡餅」 分屬兩種獨立類別的美食，可分別從辭典中查詢獲得個別的意義，但兩者組合後的「咖哩餡餅」又可成為另一種獨立意義的美食，由此可見在美食種類如此不斷推陳出新的情況下很難找到一部辭典能囊括所有的種類。基於上述兩個原因，我們嘗試以機器學習的方式整合分析非結構化的文章標題並加入特徵值後，以條件隨機域(Conditional Random Field，簡稱 CRF）的方式為基礎，透過中央大學 WIDM 實驗室所開發的 WIDM_NER_TOOL 針對序列性資料整合了 [特徵擷取]、[辭典分析]、[訓練模型]、[實驗測試]...等模組訓練與測試模型，進而設計出一套自動化擷取餐廳類別的方法，以解決傳統由人工自行撰寫擷取程式的困擾。

1-3 章節概要

本論文共有五個章節，第一章是緒論，主要說明本文的大綱，內容包含了研究動機與背景及相關限制；第二章是相關研究，討論機器學習與中文實體命名擷取領域的相關研究；第三章為設計與實作，內容詳細說明本文的研究架構與方法；第四章則是實驗，說明以 5-Fold-Cross-Evaluation 方式為基礎，並分別以 Partial Match 及 Exact

Match 搭配 Precision、Recall 計算出 F-Measure 為指標作為驗證結果，再分別針對監督學習、半監督式學習的實驗結果進行分析與討論；第五章則為本研究之結論與後續建議。

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，多數研究組織像是 Message Understanding Conference (MUC) 和 Text Retrieval Conference (TREC) 等已投入相關研究多年並累積相當的成果；而實體名稱擷取則屬於自然語言處理 (Natural Language Processing , NLP) 的領域，意指從非結構化的文件中識別實體名稱，並測試機器能夠理解自然語言所編寫的文件以及自動執行通常由人類執行的常規任務的程度。早期針對實體名稱擷取的研究通常以 Rule Based Extraction Methods 為主要方向，透過分析句法考慮語意，標記詞性並觀察字間文法結構找出規則建立辭典，再由人工撰寫程式進行比對來擷取資料；然而這樣的方式除了受限於程式開發者須對該領域有一定程度的了解外，當處理未曾出現過的資料或是規則時需要新增辭典或是修改程式，如此不僅耗時且所耗費的成本通常較高，因此讓機器從範例資料中透過邏輯規則與統計手法自動學習的演算法逐漸成為發展趨勢。

機器學習法 (Machine Learning Method) [1,2] 是人工智慧 (Artificial Intelligence , AI) 的一個支領域，在資訊擷取領域廣泛的使用，多年來已發展為多領域交叉學科，涉及逼近論、機率論、凸分析、統計學、計算複雜性理論等多門學科。有別於 Pattern-Based Method，機器學習法不僅不受限於程式撰寫者對該領域的了解程度，也不須侷限於字/辭典內容的豐富性，相較於 Pattern-Based Method 具有較高的彈性可以適用於不同變化格式及不具規則資料來源。目前機器學習已廣泛應用於資料探勘、自然語言處理、搜尋引擎、醫學診斷、生物特徵識別、和機器人等領域。近年來機器學習法用於處理自然語言實體命名識別的相關研究從早期將其歸類為分類問題，建立數種特徵作為屬性來訓練模型，像是 Support Vector Machine (SVM) 或 C4.5 決策樹 (C4.5 Decision Tree) 等演進為序列性標誌 (Sequential Labeling) 問題，常見的方法如隱藏式馬可夫模型 (Hidden Markov Model, HMM) [3]、最大化熵馬可夫模型(Maximum Entropy Markov Model, MEMM) [4] 以及條件式隨機域 (Conditional Random Field, CRF) [5] 。

過去針對餐廳類別或是美食名稱擷取的相關研究並不多，其中大多以 Pattern-Based Method 的方法進行，先從各大美食評論網站如：愛評網、大眾點評網...等蒐集美食資料後建立美食辭典，再經過程式比對找出所需的美食資訊；然而由於中文美食的種類繁多且不斷創新，加上本文擷取的對象「PTT FOOD」版內的文章並不具語法規則，因此不但難以找到一部辭典囊括所有的種類，而且要從中找出一套規則以 Pattern-Based Method 的方式撰寫程式也有一定的困難。因此本文把餐廳類別擷取問題視為實體名稱擷取類別，參考中文組織命名實體辨別的做法，試著利用機器學習的方式透過 WIDM_NER_TOOL 結合條件式隨機域（Conditional Random Fields, CRF）來實現。

2-1 中文組織命名實體辨認

實體名稱擷取（Named Entity Extraction)可歸屬於資訊萃取與自然語言領域的共同分支，早期對於實體名稱取的問題通常以「Rule-Based-Extraction-Method」的方式藉由人工訂定規則或蒐集辭庫並撰寫程式進行比對來完成，然而任何系統皆無法窮舉出所有的詞彙，而且相同的詞彙出現在不同的句子中也可能代表著不同的意義。因此，近來許多研究的方法是利用序列標記配合機率統計模型計算出最可能的標記。目前已有許多中文組織名稱辨認的研究 [6-8]，例如： 2007 年 Zhang 等人將數個 CRF 模型串連起來進行辨識，以人民日報的新聞內容作為訓練資料來源，採用「是否為前級輸出的命名實體（Is Named-Entity)」 、 「常見的組織名稱開頭」 、 「內容與結尾」、 「N 元文法（N-gram)」，其實驗結果的 F-measure 可達 0.9794。2011 年 Yao [9] 則是將中文組織名稱切分為三個部分的串接，分別是「前置詞(Prefix Words)」+「中間詞(Middle Words)」+ 「記號詞（Mark Words)」，例如：日式+ 涮涮鍋 + 餐廳，並使用自行設計的統計方法，將「組織名稱的頻率」、「詞性與長度」，配合以下假設進行計算：「記號詞能完全收錄」、「前置詞與中間詞為名詞、 形容詞、序數或位置等」、「記號詞大部分為名詞」和「組織名稱小於等於 10 個字」；最後利用人民網的語料進行訓練，並以人民網、北京郵電大學網站首頁、新華網的新聞當作測試資料，平均準確率最高達 0.959，平均召回值則達到 0.8724，這樣的結果皆超過隱藏馬可夫模型（HMM）與最大熵模型（ME)。2012

年 Ling 等人 [10] 以規則式實體名稱辨認法（Rule-Based-Named-Entity Recognition）來辨識人民日報與新浪網的新聞，語料在經過斷詞處理後，其中文組織名稱被拆解為數個修飾詞（Modifiers）+ 核心特徵詞。在統計訓練資料後找出常用的核心特徵詞並建立詞庫當作組織名稱的結尾，再找出種左邊界特徵（Left-Border Features）判斷組織名稱的起點。在取得組織名稱候選者之後，利用該系統的常見錯誤模式（Debugging Patterns）進行修正。最後的實驗結果的 F-measure 最高達到了 0.85。

2-2 監督式學習

大多數實體名稱識別的相關研究方向以監督式學習為基礎，透過序列標記的方式來建立模型，主要可分為三種: 第一種為隱藏式馬可夫模型 (HMM) 。HMM 是統計型式它用來描述一個含有隱含未知參數的馬爾可夫過程。在正常的馬爾可夫模型中，觀察者直接可見狀態，而狀態的轉換機率便是全部的參數。然而，在隱藏式馬爾可夫模型中對觀察者而言狀態並非直接可見，每一個狀態在可能輸出的符號上都有一機率分布。因此可將 HMM 視為是混合的模型，其中隱藏變量在馬爾可夫過程中具有關聯而非彼此獨立，因此輸出符號的序列能夠透露出狀態序列的一些信息。第二種最大化熵馬可夫模型 (Maximum-Entropy Markov Model, MEMM)，或稱為條件式馬可夫模型(Conditional Markov Model, CMM) [11]。 MEMM 合併了 HMM 與最大化熵模型的特徵，假設未知變量在馬爾科夫鏈中互相連接而不是彼此有條件地獨立的最大熵分類器的擴展。最後一種則是條件式隨機域（Conditional Random Field, CRF)。CRF 是一種以統計的概念建立模型的方法，其可以考慮上下文並基於有區別的無向性機率建立圖形模型，並使用特徵矩陣來對特徵之間的已知關係進行編碼並且推斷未知變量。它通常用於標記或解析順序數據，如自然語言處理（Natural Language Processing, NLP）或生物序列以及計算機視覺；舉例而言，CRF 經常應用於 NLP 中以預測序列輸入樣本的標籤序列。

本文利用中央大學 WIDM 實驗室所開發，以 CRF 作為機器學習演算法結合 Taku Kudo 研究發展的 CRF++ [12] 並整合了 [特徵擷取]、[辭典分析]、[訓練模型]、[實驗測試]...等功能的 WIDM_NER_TOOL [13] 來實作。CRF 為一個序列型式架構，保有

MEMM 的優點，常用於標註或分析序列資料如自然語言文字或是生物序列。MEMM 和 CRF 之間主要差異在於 MEMM 使用每一個狀態的指數模型來描述在給予前一個狀態時當前狀態的條件機率，而 CRF 則是採用單一指數模型來描述在給予觀察序列的條件下整個狀態序列的聯合機率，因此在不同的狀態中，不同的特徵函數所賦予的權重可以考慮到狀態彼此的情形。

CRF 與馬爾可夫隨機域 (Markov Random Field) 相同為無向性之模型，圖1 中的頂點代表隨機變數 Y，頂點間的連線代表隨機變數間的相依關係，在條件隨機域當中，隨機變數 Y 的分佈為條件機率，給定的觀察值則為隨機變數 X。原則上，CRF 的圖模型佈局可任意給定，鏈結式架構式一般較為常用的佈局方式，不論在訓練 (Training)、推論 (Inference)、或是解碼 (Decoding) 上，都存在有效率的演算法可供演算。CRF 與 HMM 常被一起提及，與 HMM 不同的是這個模型為非生成性 (Generative) 模型，也就是沒有觀察值是由狀態之模型所生成的假設，這樣的特點使得在模型中不需要估算狀態的機率分佈模型，而且 CRF 對於輸入和輸出的機率分佈，沒有如 HMM 那般強烈的假設存在。

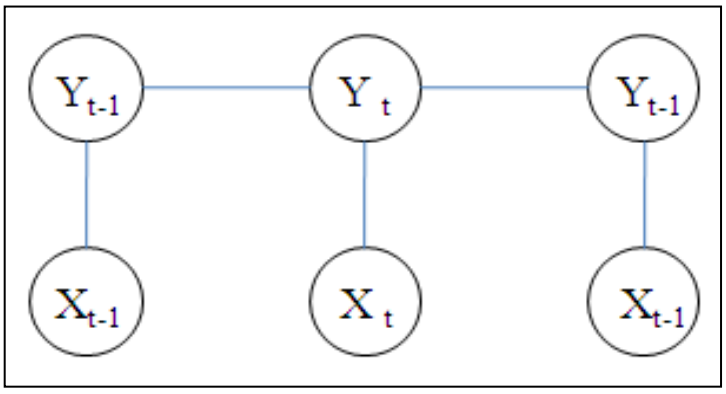


圖 1 線性鏈結構 CRF 架構

CRF 在 NLP 的各種應用最早由 John Lafferty 於 2001 年提出[14]，在詞性標註、中文斷詞、命名實體辨認方面有顯著的成果。目前許多關於中文組織名稱辨認的研究[15] [16] [17] [18]，可以從一些較正式的文章中萃取出組織名稱，然而這類文章大多具有較正式的書寫規範，組織名稱也大多具有較正式的命名結構，但對於餐廳類別而言並無一定的規範可言，例如：阿伯紅茶、巷口的超人氣牛排、阿草婆祖傳肉粽...等，都隱

含著餐廳類別；此外，本文的研究對象 PTT 網站上的網頁內容是由使用者以自然語言撰寫並無顯著的文法規則，版內文章內容具有結構與非結構化的資訊交錯呈現，其中隱含有可利用的資訊，因此本文希望透過 CRF 作為演算法，並參考中文組織命名實體辨識的相關研究來達成餐廳類別的擷取。

2-3 半監督式學習

半監督式學習 (Semi-Supervised) 所涉及的監督程度較小，意即所需的標記資料量相對較少，常用於標記資料來源不易且昂貴時，藉以降低取得訓練資料的成本。半監督式學習系統主要可藉由已知的範例從未標記的資料中尋找類似的實例並自動標記重新訓練模型以提高效能；透過不斷的重複此步驟，可累積大量的標記資料用來訓練模型，因此，半監督式學習也可解釋為利用未標記的數據來訓練模型的技術。近年來許多關於半監督式學習的方法陸續被提出，其中較為常用的包含了： Self-Learning (bootstraping) [19]、Expectation-Maximization-Based-Approach 、 Graph-Based-Approach 、 Semi-Supervised-Support -Vector Machine [20] 、Co-Training [21] 、 Tri-Training [22] 等。在 Self-Learning 中，首先利用已標記的範例訓練分類器，接著應用此分類器針對那些未標記的資料進行標記並使用新標記的範例子集（與原始標記的範例結合）重新訓練模型。Expectation-Maximization-Based-Approach 則是利用透過最大似然估計 (Maximum Likelihood Estimation, MLE) 或最大後驗估計 (Maximum A Posteriori Estimation, MAP) 所標記的資料來訓練初始模型，然後使用此模型來“猜測”未標記資料的欲標記內容後，再用這些資料來重新訓練模型並一直重複上述過程直到收斂為止。Graph-Based Approach 則是將無論是否已標記的每一筆資料都視為圖的頂點，再應用一種稱為圖形正規化 (Graph-Based-Regularization) 的方式來優化預測函數，並加以利用優化設計中的標籤平滑度屬性，不僅將標記數據的損失最小化而且也確保沿著圖形的標記和標記數據的平滑性。

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 [23]，然而從 Yu 和 Kubler 應用 EM-NB 分類器 (EM-Based

Naïve-Base)、S^3VM、Self-Training、Co-Training 等方法來進行見解探勘 (Opinion Miming) 的研究中顯示 [24]，半監督式學習其效能非常接近具有完整數據的監督式學習。雖然半監督式學習法的分類器已被廣泛的提出與應用，但是針對序列性標記問題卻鮮少被關注且設計的原理也有相當的差異。舉例而言，Ando 和 Zhang [25] 採用學習典範 (Learning Paradigm) 的方式從未標記的資料中自動建立輔助問題，因此可以由多個分類問題共享的公共結構中學習預測結構，這可以用於改善目標問題的性能。他們利用 CoNLL'03 實體命名資料集 (英文和德文) 來驗證他們的方法，實驗結果的 F-measure 達到 0.893 (英文) 以及 0.753 (德文)；此外，這樣的方法也應用在 CoNLL'00 資料集進行語法驗證，實驗結果的 F-measure 可達 0.944 (全部) 以及 0.947 (名詞短語)。

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 [21] 所提出，內容主要描述透過 Co-Training 的方式僅利用 12 筆已標記的資料來將完成網頁分成兩類。這樣的演算法需要利用已標記資料分別訓練兩個不同 view 的分類器。Nigam and Ghani [26] 的研究中使用與 Blum and Mitchell 相同的資料集 (WebKB course) 來進行實驗，結果證明當獨立的特徵集假設能被有效的應用時，可提高 Co-Training 的效能。

Tri-Training 可視為 Co-Training 的改良，不同的地方在於 Ttri-Training 使用三個分類器並且以投票 (Voting) 的機制來解決兩個分類器共同所共同標記的答案信心度的問題；其運作方式如下圖 2，在每一次的 Tri-Training 過程中，分類器 h_j 和 h_k 從原始的未標記集(U)中選擇一些樣本進行標記並將標記結果提供給分類器 h_i ($i, j, k \in \{1,2,3\}, i \neq j \neq k$) 。 其中 L^t_i 表示第 t 次 Tri-Training 過程中由 h_j 和 h_k 所標記的樣本。接著 L 與 L^t_i 的聯集 $(L \cup L^t_i)$ 則成為 h_i 的訓練資料並重新訓練 h_i；此外，在第 t 次 Tri-Training 過程中由 h_j 和 h_k 所標記的資料集 L^t_i 並不會被放入原始的已標記資料集 (L) 內，相反的，在第 t+1 次的 Tri-Training 過程中所有的 L^t_i 將被視為是未標記資料並且再次放入原始的未標記集 (U) 內。關於初始的三個分類器 h_i、h_j、h_k 則有數種的建構方式，Zhou.. 等採用自助抽樣法 (bootstrapped sampling) [22]，研究人員嘗試各種演算法並應用多個不同的 view 以及它們的組合來建構初始分類器，並可達 0.804 的精確度 [27]。

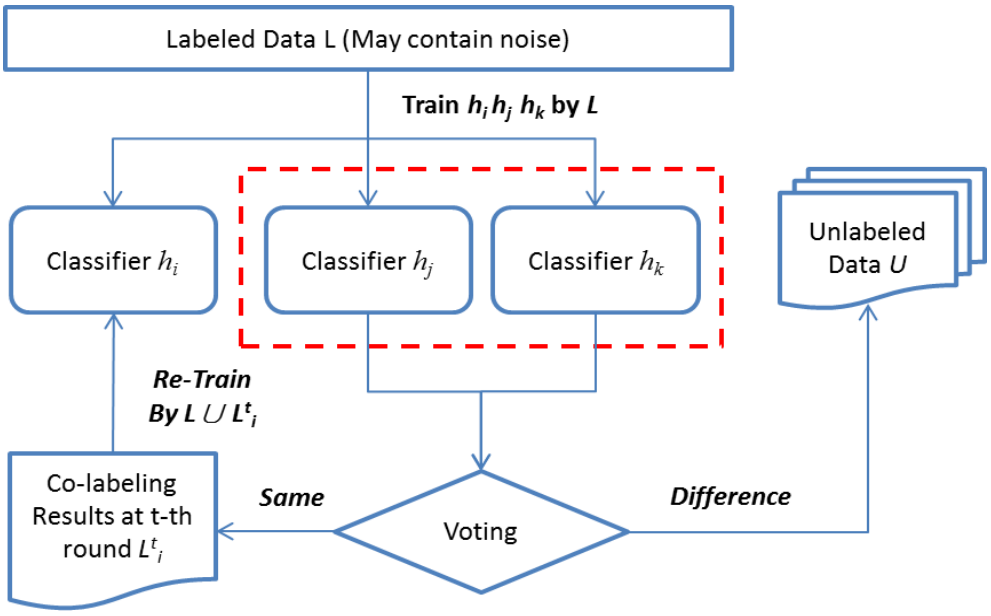


圖 2　Tri-training 流程示意圖

Distant Supervision Learning 也是半監督式學習的一種演算法，透過啟發式規則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，Snow 等人 [28] 利用人工建立的正規表達式來自動擷取實體之間上下位的配對以辨識實體之間的關係。他們使用從解析樹提取的依賴性路徑作為特徵來確定新聞文章中的兩個名詞之間是否為上位詞關係，其分類器利用從維基百科收集具有 200,000 個相關性路徑的擴展特徵詞典作為附加訓練數據，最佳實驗結果的 F-measure 達到 0.359。

Mintz 等人 [29] 提出了一種不需要已標記語料庫並且允許使用任意大小語料庫的替代模型，他們使用一種稱為 Freebase [30] 包含數千個已知的關係對的大型語義數據庫來進行 Distant Supervision Learning; 而他們發現在 Freebase 中出現過一次或多次的關係對也都能在大型未標記的語料庫中找到，進而提取相關特徵以訓練關係分類器。它們結合了監督式 IE（在機率分類器中組合 400,000 個雜訊模式特徵）和非監督式 IE（從任意域的 155 個大型語料庫截取大量關係）的優點以 0.676 的精度提取 102 個關係的 10,000 個實例。對於從非結構化的文本中擷取資訊而言一開始所選定的知識庫或資料集 (Data Set) 是相當重要的，Michelson 和 Knoblock [31] 提出了一種直接從文本本身構建引用集的方法，該技術建構元組 (tuples) 來表示文本中的實體以形成參考集。他們以九次的資訊截取任務評估他們的系統，主要從 Craiglist 版上的貼文中擷取包含了汽車、筆記型

電腦、滑雪板...等屬性名稱；實驗結果發現他們的方法在 F-Measure 上獲得改善，在性能上相較於其他法也具有競爭力。

Joohui An 等人 [32] 使用已知的種子 (Seeds) 實體作為查詢關鍵詞來收集包含實體實例的網頁，這種自動標記的語料庫其標記品質或許較手動標記稍差，但是其大小可以幾乎無限增加且無需任何人力入。他們使用了大約 28-51 次自動標記的命名實體，並在韓國人姓名的實驗上獲得了幾乎與手動標記相似的性能（0.85）。Rae 等人 [33] 提出了一種類似的方法透過從 Web 片段引導來增加訓練數據的數量，他們使用維基百科標題和社交媒體簽入標籤名稱（Foursquare 和 Gowalla）作為已知的 POI 來查詢搜索引擎以收集訓練句子。透過 10-Fold Cross-Validation，他們可以在自由文本中識別 POI，實驗的 F-Measure 值介於 0.557 和 0.915 之間。Fu 等人 [34] 提出了一種從英漢對話平行語料庫生成大規模中國 NER 訓練數據的方法。他們訓練了一個高性能的 NER 系統在英語語料庫中標記命名實體，並根據詞級校準和雙語語料庫標記中文語料庫。在他們的實驗中產生了一個含有 167,100 句子的中文 NER 語料庫，對於 863 評估語料庫有 67.89% 的 F-measure，對於 OntoNotes 語料庫則達到 73.20％的 F-Measure。

三、 設計與實作

本研究主要以 PTT 網站裡 FOOD 版的文章作為擷取對象，先利用爬蟲程式擷取版上文章並存入資料庫進行格式化後隨機以人工的方式檢視 1,000 筆資料進行結構分析以了解資料的概貌，再依據分析的結果設計本文的研究流程。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；分析結果顯示餐廳名稱、地址、電話這些資訊多半出現在文章的內文中與特定的描述詞相連結，並以符號或換行的方式加以區隔，因此針對這部分的資訊本文以文章內文作為擷取來源並透過關鍵字比對的方式擷取; 而 URL 則是由固定的字串加上文章編號所組成，最後標題的部分則是直接從原始文章的標題進行擷取即可，以上皆可透過規則撰寫程式來完成。第二部分為餐廳類別擷取，經由分析發現餐廳的類別可由文章標題中擷取的比例較高，然而文章標題的組成並無特定的格式，而是作者依其口語習慣以自然語言隨意寫下不具特定文法句型結構規則的短語，屬於非結構化的資訊來源。依照上述分析結果，設計本文的研究流程如下圖 3 所式，詳細內容分述如下。

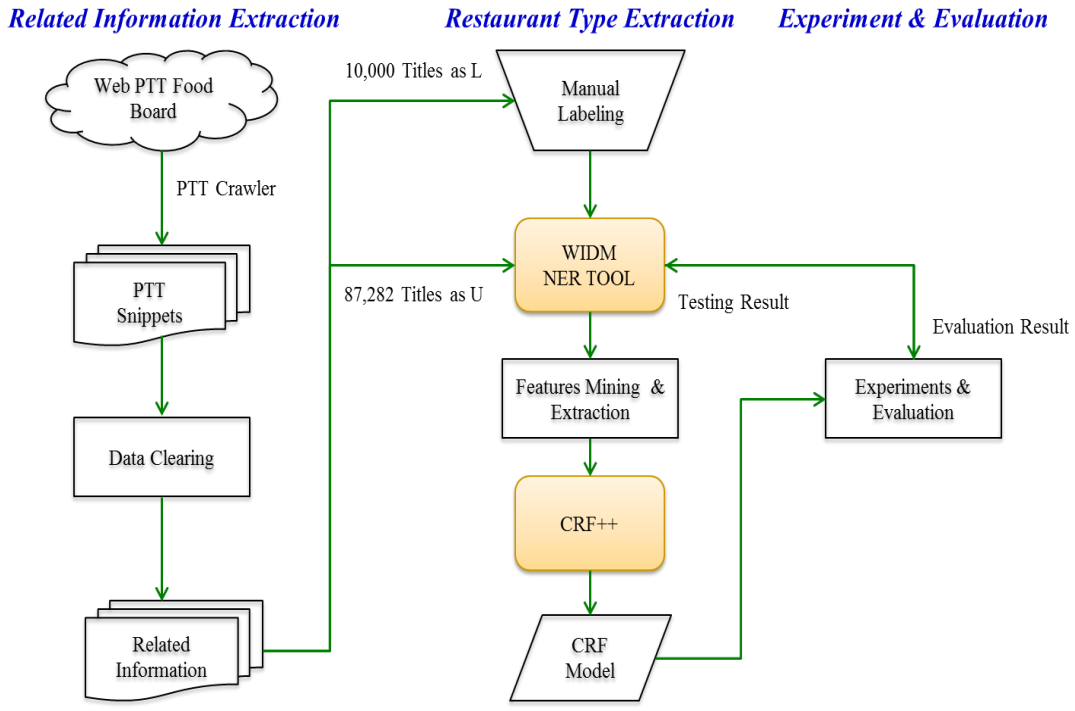


圖 3 研究流程圖

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、資料格式化、餐廳相關資訊擷取，如下圖4。首先在網頁資料擷取的部分是透過 PTT Crawler 程式將 PTT Food 版上的文章逐一擷取後儲存於資料庫，程式的主要運作流程為取得 FOOD 版的總頁數後從最後一頁（最新的文章）開始往前逐一擷取，再將擷取後的文章逐一以文章的編號命名儲存成.txt 檔， 並不斷重複上述動作以隨時自動更新最新的推文內容；以 2016/7/21 所擷取的總資料量共 111,157 筆作為本文研究資料來源。

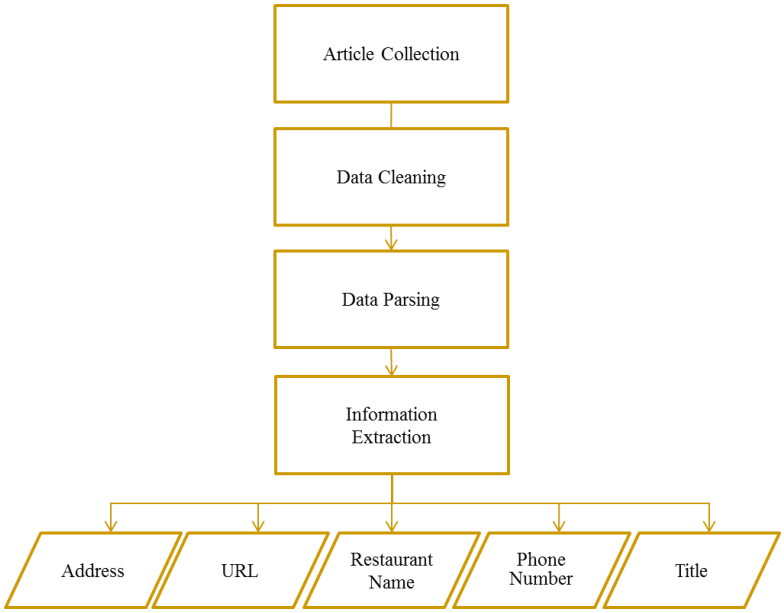


圖 4 相關資訊擷取流程圖

第二部分則為資料格式化，主要分為三個步驟，首先針對存成文字檔的資料進行格式轉換，原因在於擷取後的文章屬於 UTF-8 格式，而後續使用的程式格式需求為 ANSI，若未轉檔直接開啟會有亂碼產生，因此先透過程式自動將檔案逐一從 UTF-8 轉成 ANSI 檔，避免擷取時發生錯誤。接著進行資料的篩選，主要目的是濾除無關餐廳類別的文章，如: [廣告]、[系統公告]、[請益]...等，僅保留屬於 [食記] 的文章作為本文研究對象。 依 2016/7/11 所擷取的 111,157 筆資料中統計可得 [食記] 類別的發文共 97,282 筆。資料格式化的最後一個步驟則為去除資料中的雜訊，包含了空白、特殊字元、亂碼...等，以使資料的格式統一化便於後續程序應用。

前處理的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，再依規則撰寫程式進行擷取，其包含了文章的「餐廳名稱」、「地址」、「電話」、「URL」以及「標題」，擷取方式與來源分述如下:

- 「標題」: PTT 系統已為標題預設了固定的格式 (如圖5)，置於文章的第一行並開始於固定字元 「Title:」， 使用者所撰寫的標題內容則接續於後，最後再以固定字串「- 看板 Food - 批踢踢實業坊」為結尾，因此根據上述規則利用程式去除固定字串與字元後可直接擷取標題內容。

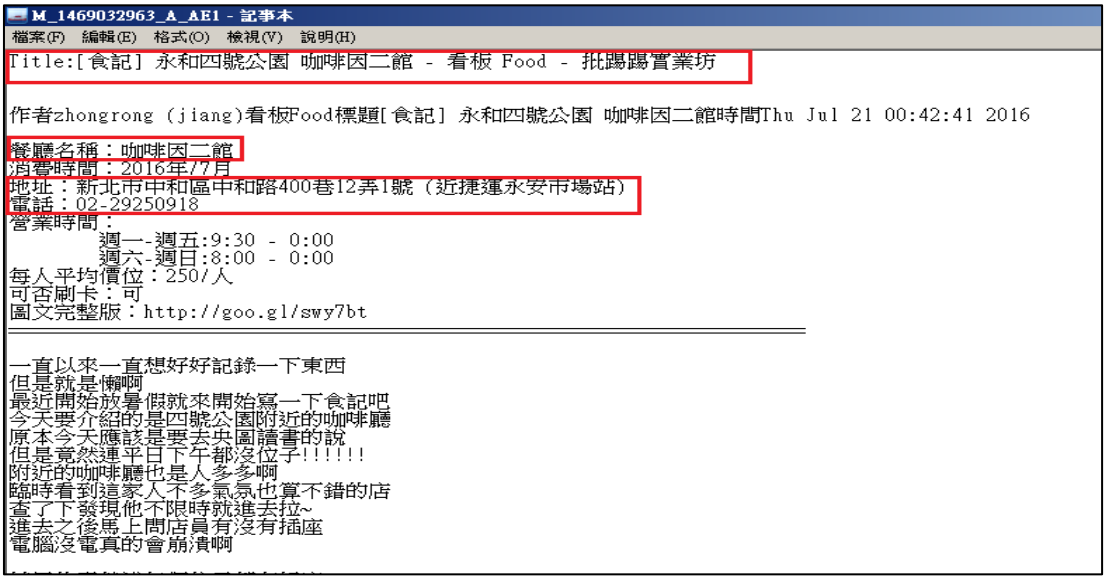


圖 5 餐廳名稱、地址、電話撰寫格式

- 「餐廳名稱」、「地址」、「電話」: 經隨機檢視 1,000 筆資料進行擷取來源分析（如表 1）後發現，近 90% 以上文章的「餐廳名稱」、「地址」、「電話」等資訊可從文章的內文中獲取，此外，由於 FOOD 版的發文規則已初步制定這些資訊的撰寫方式（如圖 5）， 所以在內文中可發現這些資訊通常個別存在於單獨的一行並且接續在特定的關鍵字後出現，僅少數文章將這些資訊夾雜於內文中，因此，我們透過關鍵字（如表 2）比對的方式撰寫程式對每一個文字檔的內文逐行進行比對，再將符合的該行擷取後儲存於資料庫。
- URL: Food 版內文章的 URL 是由既定的格式所組成，其格式為固定字串「https://www.ptt.cc/bbs/Food」+ 「文章編號」 + 固定字串 「.html」所組成，舉

例而言，若擷取的文章編號為 M.1469023425.A.A76，則文章的 URL 如下: https://www.ptt.cc/bbs/Food/M.1469023425.A.A76.html。

表 1 餐廳名稱、地址、電話擷取來源分析

來源 / 擷取資料	內文	標題
餐廳名稱	**92.40%**	**92.50%**
地址	**93.70%**	1.30%
電話	**91.30%**	0%

表 2 餐廳名稱、地址、電話擷取關鍵字

擷取資料	**關鍵字**
餐廳名稱	餐廳名稱、店名
地址	地址、位置
電話	電話、聯絡方式

3-2 餐廳類別擷取

在開始擷取的流程前我們先針對欲擷取的「餐廳類別」加以定義；本文中所擷取的餐廳類別屬於中文組織實體名稱擷取的領域，其定義不單只是指商店的名稱或是販售的食物，舉例而言: 餐廳名稱為「綠色空間」，單從字面意義上並無法辨識餐廳的類別，而「杜蘭朵法式餐廳」，其餐廳的類別為「法式餐廳」並非販售的食物；因此，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「可成為獨立類別的餐廳特徵」或是「餐廳的主要販售商品」兩大類，舉例如下：

- Title:[食記] 高雄玫瑰園*景觀餐廳* => **可成為獨立類別的餐廳特徵**
- Title:[食記] 月光奇想*歐風小館*—嘉義 =>**可成為獨立類別的餐廳特徵**
- Title:[食記] 新竹的深坑*臭豆腐* => **餐廳的主要販售商品**
- Title:[食記] 通化街的老店頭*排骨酥麵* => **餐廳的主要販售商品**

3-2-1 擷取來源

經由人工隨機挑選 1,000 筆資料分別統計餐廳類別隱含於內文、標題、餐廳名稱的比例以評估擷取來源，統計結果如下表3。經分析後發現，雖然大部分的內文都存在著餐廳類別的識別字，但由於內文屬於敘述性文章，作者往往同時介紹多種餐廳內所販售的美食與特色，如此多種類且不具固定章法的編寫方式難以由程式自動判斷何者為主要類別；舉例而言，「豐茗樓港式飲茶」的餐廳類別為「港式飲茶」，但作者在內文中會往往會針對覺得值得推薦的商品如湯包、蒸餃、燒賣、蛋塔...等逐一介紹但描述的方式並不一定是條列式敘述而大多是依作者的習慣隨意以不具規則的自然語言所編寫; 此外，作者在敘述中也常穿插著非關文章的餐廳類別，例如在某篇介紹港式飲茶的文章中提到「...蛋塔比起其他吃到飽餐廳裡的要好吃多了...」，在這個句子中包含了「吃到飽餐廳」與「蛋塔」兩種可作為餐廳類別的識別字，但「吃到飽餐廳」亦屬於一種獨立的餐廳類別卻與文章無關，由此例可見要從內文自動判斷文章的餐廳類別是「港式飲茶」或是「蛋塔店」還是「吃到飽餐廳」具有一定的困難度。

表 3 餐廳類別擷取來源分析

來源 / 擷取資料	內文	標題	餐廳名稱
餐廳類別	數種類別	**72.50%**	65.30%

排除以內文作為擷取來源後，從「標題」與「餐廳名稱」的分析結果 (如上表 1) 中發現，餐廳類別隱含在 「標題」 中的比例為 72.5% 略大於隱含在「餐廳名稱」中的65.3%，進而針對「餐廳名稱」與「標題」進行交叉分析後發現，92.5% 的餐廳名稱被隱含在標題中，意即大部分的餐廳名稱被包含在標題裡重複出現，例如：文章的餐廳名稱為「香港竹家莊」，而此名稱也重複出現在文章的標題「台北松山區-香港竹家莊港式料理」裡。基於上述原因，本文選定以文章的「標題」作為餐廳類別的擷取來源。

3-2-2 CKIP 斷詞與人工資料標記

詞是中文具有意義的最小單位，而中文斷詞的結果在處理中文組織名稱擷取時扮演著重要的角色。參考過去關於中文商家名稱辨識的相關研究中將商家名稱分段串接的方式，我們發現大部分的中文美食名稱的組成亦可拆分段來討論，種類可分為以下四類：

(1)「**前綴詞**」＋「**獨立美食類別**」：日式 + 雞排。

(2) **不可拆分的「獨立美食類別」**： 鐵板燒。

(3)「**獨立美食類別**」＋「**獨立美食類別**」：咖哩 + 雞排。

(4)「**獨立美食類別**」+ 「**後綴詞**」：火鍋 + 吃到飽。

進而依上述的四種類別加以詞性進行分析發現，前後綴詞大多是由形容詞 (A) 或名詞 (N) 所構成，而獨立美食類別則以名詞（N）為主，由此推論詞性 (POS，Part of Speech) 可做為輔助標記的重要特徵，因此在人工標記前我們透過中研院所提供的中文斷詞系統 (CKIP, Chinese Knowledge Information Processing Group) 將標題進行斷詞，獲得的斷詞結果與詞性作為後續人工標記之參考；如 :「大潤發中崙店美食街日式涮涮鍋吃到飽」，經斷詞後得到結果為：「大潤發(Nb)　中崙店(Nc)　美食街(Nc)　日式(A)　涮涮鍋(Na)　吃到(VC)　飽(VH)」。

參考標題資料的斷詞結果與詞性，我們以啟發式規則 (Heuristics Rule) 對標題進行人工標記以作為訓練與測試資料；啟發式規則所指的是在資料標記時常遇到難以決定所要標記類別的層級；舉例而言，標題內出現餐廳類別識別字「日式拉麵」，若以食物層別來看「日本拉麵」的上一層為「拉麵」，而「拉麵」的上一層為最頂層「麵」，這三層皆可用來標記為餐廳類別，但若僅標記「麵」或「拉麵」則所涵蓋的範圍太廣有失辨識的意義，使用者難以藉由辨識結果區分餐廳類別究竟是販售哪一種麵類；因此本文以最長詞彙標記法 (longest) 搭配詞性作為標記原則，舉例而言，若原始文章的標題為「東海商圈－炒日式拉麵館」，此時 「日式拉麵」、「拉麵」、「麵」都可作為標記的關鍵字，因此參考 CKIP 斷詞後的結果:「東海(N)　商圈(N) －(DET)　炒(Vt)　***日式(A)***　***拉麵館(N)***」並以最長詞彙標記法所標記的結果為: 東海商圈－炒「**日式拉麵館**」。

3-2-3 特徵擷取

本文自特徵擷取以後步驟皆透過 WIDM 實驗室所開發的 WIDM_NER_TOOL Package 來完成，首先在處理餐廳類別擷取之前，必須先建立用於訓練資料的特徵，我們將人工標記後的標題經過特徵擷取，並給予標題中的每一個符號對應的特徵值。而特徵的選擇是否恰當會影響訓練資料及擷取的結果。從人工標記時統計發現，隱含於標題中的餐廳類別及前後的詞性主要分為四大類，分別是前後方常見的名詞、形容詞、介係詞以及組成餐廳類別的前後綴詞，而上述這些類別皆有專用來描述餐廳的常見字詞，例如好吃的、平價的、美味的、古早味…等皆為常用來形容餐廳的形容詞。而館、亭、屋、專賣店等則為餐廳類別後方常連接的名詞。以上這些特徵均可提供有關餐廳類別的跡象，所以我們將上述類別再依字數加以區分，設計了 14 種特徵值 (如表 4) 來協助我們擷取餐廳類別。

表 4 特徵值設計

ID	代號	特徵	說明	長度	範例
1	BE1	Before_1	常見於 Entity 前方的詞	1	的、式、大、小
2	BE2	Before_2	常見於 Entity 前方的詞	2	手工、創意、經典
3	BE3	Before_3	常見於 Entity 前方的詞	3	好吃的、古早味、
4	PF1	Prefix_1	常見前綴詞	1	素、茶、烤、乾
5	PF2	Prefix_2	常見前綴詞	2	日本、日式、泰式、港式
6	PF3	Prefix_3	常見前綴詞	3	義大利、無國界
7	SF1	Suffix_1	常見後綴詞	1	鍋、粥、羹、凍
8	SF2	Suffix_2	常見後綴詞	2	料理、飲茶、火鍋
9	SF3	Suffix_3	常見後綴詞	3	自助餐、吃到飽
10	AF1	After_1	常見於 Entity 後方的詞	1	館、屋、廳、亭、店
11	AF2	After_2	常見於 Entity 後方的詞	2	大王、餐廳、
12	AF3	After_3	常見於 Entity 後方的詞	3	專賣店、專門店
13	Eng+Num	English/ Number	半形或全形字母與	1	「A」、「F-15」
14	Sym	Symbol	半形或全形符號	1	「，」、「。」、「：」

選定了特徵後，接著要決定選擇特徵值的方法。我們分別以特徵值出現的「頻率 (Support)」以及「置信度 (Conf.，Confidence)」來從辭典 (Dictionary) 中選取候選字串

作為特徵值用以訓練學習模型；所謂「Support」意指候選字串在已標記的訓練資料中被視為特徵值的次數，但若單純僅以次數作為考量，則忽略了候選字串出現在整體資料的比例，舉例而言(如圖 6)，在特徵 Prefix_1 中，候選字元「涮」雖然被視為特徵值的次數有 65 次，但是在整體資料中出現了 130 次，亦即被視為特徵值的機率只有 50%，而像「炒」、「碗」等字元雖然被視為特徵值的次數僅有 15 次，但在整體資料中也僅出現了 15 次，意即被視為特徵值的機率達 100%，代表這些字元出現的次數雖少，但只要出現幾乎都可被認定為特徵值，倘若僅用次數作為挑選條件，容易忽略這些次數低但機率高的候選字元。因此本文將 Confidence (「視為特徵值的次數」/「總出現次數」) 納入考量並在下一章節進行 Dictionary Terms Mining 實驗加以比較驗證。

Terms	Find Entity Feq.	Total Feq.	Confidence
咖	566	612	0.9248
牛	349	390	0.8949
日	361	379	0.9525
涮	65	130	0.5
豆	109	122	0.8934
海	102	115	0.887
下	109	109	1
韓	86	106	0.8113
炒	15	15	1
南	12	15	0.8
碗	15	15	1
撈	12	15	0.8
鰻	11	15	0.7333
飲	14	14	1

圖 6 特徵值選取方式 Frequency & Confidence 比較

3-2-4 訓練過程和測試過程

標題經過前置處理和特徵擷取過程後，接著進入學習模組來訓練模型與測試資料，而本研究使用條件式隨機域來做為序列標記的模型，並採用 CRF++ 工具來實現。

使用條件式隨機域的原因在於餐廳類別並沒有固定的長度與命名規則，雖然多數相關研究以美食辭典比對的方式，採用 Pattern Base Method 來擷取餐廳類別，但本研究考量多數餐廳類別在加以排列組合後常可產生新的類別，且類別中的字詞可選擇性出現，例如: 義大利麵與義麵所代表的是屬於同一種美食類別，因此本文採取序列性標誌 (Sequential Labeling) 的方式來進行，將餐廳類別擷取問題轉換成標誌問題(Labeling

Problem)，標題中的每一個符號有其對應的標誌 (Label)，這些符號的標誌同樣由出現的位置來決定，對此我們採用 BISEO 標記法，分別根據符號出現的位置標誌 B、I、S、E、O 五種標誌，根據符號在餐廳類別中出現的位置可以分為：位於類別的開始 (Beginning) 、位於類別的中間 (Intermediate)、位於類別的結尾 (End) 、獨立類別 (Single) 以及不屬於地址的符號 (Other) 等五種類別，也稱作「BISEO 分類問題」。

我們利用條件式隨機域當作類別標誌的方法，採用 Taku Kudo 研究發展的 CRF++ 工具來實作，因為 CRF++ 設計成為一個泛用的工具，使用者必須事先指定屬性樣版 (Feature Template)，該樣版描述了在訓練和測試時會用到的屬性，而樣板的類型有 Unigram 與 Bigram 兩種，其中 Bigram 樣板會自動產生目前的輸出符號和前一個輸出符號的結合。因此我們混合 Unigram 與 Bigram 兩種樣版來產生大量的不同屬性。

利用條件式隨機域訓練資料時，先分析標題中的每一個符號，求算對應的 14 種特徵值並標上正確標誌，最後將這些資料訓練成學習模型，利用此模型來測試資料。而測試資料時，則將標題片段透過學習模型測試，此模型會標示標題片段中每一個符號的預測標誌，其中我們使用 BISEO 標記法中的五種標誌來標示每一個符號：這五種標籤分別代表此符號出現在類別的起始位置、中間、結尾處以及不屬於餐廳類別的部分或是獨立類別，每一個符號有其對應的特徵集合和預測的標誌，如圖 7 所示。

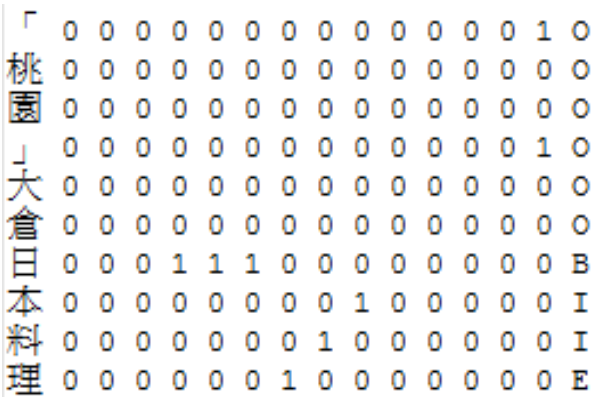


圖 7 各符號產生對應的特徵與標誌

四、實驗結果與分析

本章節主要的目的是針對本文所設計及訓練的模型利用已完成人工標記的 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;實驗主要分成三個部份,首先針對特徵值的設計與選取方式進行 Feature Mining 實驗,接著再依據實驗結果分別進行監督式學習與半監督式學習實驗,實驗方式除了對 U 進行 Distance Learning 的實驗外,其他皆採用 5-Fold CV (5-Fold Cross Evaluation) 的方式進行,並針對實驗結果分別以 Exact Match 以及 Partial Match 的評估其精確率 (Precision)、召回率 (Recall)、F-Measure,再將結果繪成圖表加以分析;以下 4.1 將說明評估方式,4.2 說明實驗設計,4.3 則為實驗結果與分析。

4-1 評估方式

評估方式採用 Exact Match 以及 Partial Match 兩種方式進行,主要原因在於餐廳類別的組成可具有選擇性,並非如同人名、地名..等專有名詞必需完全相符,有些字是否出現並不完全影響該餐廳類別的意義,舉例而言,人工標記結果為 「日式涮涮鍋吃到飽」,而系統標記結果為「涮涮鍋吃到飽」,雖然兩者標記結果並不完全相同,但對餐廳類別定義而言卻是相同的,僅在於標記精細的差異,若是以 Exact Match (如下圖 8) 的方式評估判定為 Fail 並不合理,因此本文一併採用 Partial Match 的方式評估,再將兩者結果加以比對。

Title	Manual Labeling Result	System Labeling Result	
大	O	O	
潤	O	O	
發	O	O	
中	O	O	
崙	O	O	
店	O	O	
美	O	O	
食	O	O	
街	O	O	
日	O	B	
式	O	I	
涮	B	I	
涮	I	I	
鍋	I	I	
吃	I	I	
到	I	I	
飽	E	E	Fail

圖 8 Exact Match 示意圖

評估指標採用機器學習、自然語言處理 (NLP) 的常用指標「精確率 (Precision)」、「召回率 (Recall)」以及「F1-Measure」來進行評估，相關代號說明如表 5 所示，定義與計算方式說明如下：

表 5 公式代號說明

	Relevant	Non-Relevant
Retrieved	TP：正確且被擷取的項目數	FP：錯誤但被擷取的項目數
Not Retrieved	FN: 正確但未被擷取的項目數	TN: 錯誤且未被擷取的項目數

(1) 精確率 (Precision)：「正確被擷取的項目」占「總擷取項目」的比例。

$$Precision = \frac{|TP|}{|TP + FP|}$$

(2) 召回率 (Recall)：「正確被擷取的項目」占「應該被擷取的項目」的比例。

$$Recall = \frac{|TP|}{|TP + FN|}$$

(3) F-Measure ：精確率和召回率的調和平均數。

$$\frac{2}{F1 - Measure} = \frac{1}{Precisionl} + \frac{1}{Recall}$$

$$F1 - Measure = \frac{2 * Precisionl * Recall}{Precisionl + Recall}$$

利用 Exact Math Measure 時，系統標記的結果須與人工標記的結果完全相同才可給分，否則判定為 Fail，計算方式如下:

(1) Precision: 分母為總擷取項目，分子為完全正確擷取的項目:

$$Precision = \frac{|Correctly\ identified\ entities|}{|Identified\ entities|}$$

(2) Recall: 分母為應該被擷取的項目，分子為完全正確擷取的項目:

$$Recall = \frac{|Correctly\ identified\ entities|}{|Real\ entities|}$$

(3) F-Measure: Precision 和 Recall 的調和平均數。

$$F - Measure = \frac{2PR}{P+R}$$

利用 Partial Math Measure 時，系統標記的結果只要部分與人工標記的結果相同即可部分給分，我們必須考慮各種 overlap 的情況，其計算方說明如下:

(1) Score-p: 分母為總擷取的 Token 數，分子為被擷取與正解交疊的 Token 數:

$$Score_p = \frac{|Overlap\ tokens|}{|Identified\ entity\ tokens|}$$

(2) Score-r: 分母為應擷取的 Token 總數，分子為被擷取與正解交疊的 Token 數:

$$Score_r = \frac{|Overlap\ tokens|}{|Real\ entity\ tokens|}$$

(3) Precision: 分母為總擷取項目，分子為 Score-p 的總和，計算方式為:

$$Precision = \frac{\sum Score_p}{|Identified\ entities|}$$

(4) Recall: 分母為應該被擷取的項目，分子為 Score-r 的總和，計算方式為:

$$Recall = \frac{\sum Score_r}{|Real\ entities|}$$

(5) F-Measure: 精確率和召回率的調和平均數，計算方式為:

$$F - Measure = \frac{2PR}{P+R}$$

4-2 實驗與分析

本文實驗主要分為 Feature Mining、監督式學習、半監督式學習三個部分，如下圖 9 所示。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 Feature Mining，內容包含兩個部分，首先進行「Dictionary Terms Mining」用以決定選取特徵值的方式，接著進行「CKIP Information Selection」，以實驗結果決定是否將 CKIP 斷詞後所得的資訊加入特徵值中。第二部分則依照 Feature Mining 的結果進行監督式學習的「Basic」實驗，並利用「Basic」實驗結果為基礎，搭配未標記資料 U (Un-Labeled Data) 進行「Tri-Training」實驗來測試當加入 U 後訓練資料增加對系統效能提升的程度。

最後半監督式學習實驗的部分則是利用「Basic」實驗所得的標記項目作為 Seeds 並依其出現的次數由高而低排序後進行「Distance Learning」實驗，分成兩個部分進行，首先以 Seeds 自動對 L (Labeled Data) 標記後作為訓練資料進行「Basic」實驗，並將實驗結果與監督式學習進行比較以檢視其效果是否與人工標記的效果接近；接著再利用 Seeds 對 U 自動標記後作為訓練資料，並以 L 作為測試資料以檢視加入 U 後訓練資料增加對效能的影響，詳述如下 。

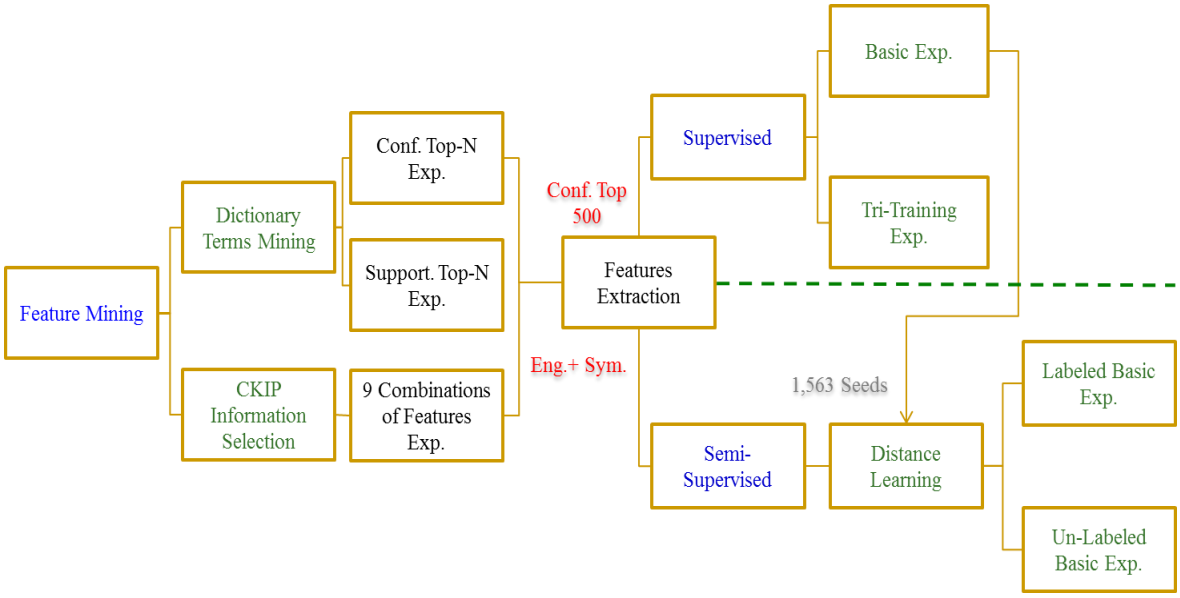


圖 9 實驗設計架構圖

4-2-1 Feature Mining

Feature Mining 的目的是透過實驗的結果來決定特徵值的設計與選取方式，主要分成 Dictionary Terms Mining 以及 CKIP Information Selection 兩個部分，分述如下：

1. Dictionary Terms Mining

在進行 Basic 實驗之前需先決定從辭典 (Dictionary) 特徵值的選取方式。本文分別針對特徵值 (Feature) 出現的 「頻率 (Support)」以及「置信度 (Confidence)」兩種選取特徵值的方式進行實驗，所謂頻率 (Support) 表示欲選取的候選字串在人工標記時被視為特徵值的次數，而「置信度 (Confidence)」則代表候選字串被視為特徵值的次數占總出現次數的比例。

實驗流程如下圖 10 所示，首先將 10,000 筆已標記的資料 (L) 分成五等分，其中 8,000 筆 (4 組) 作為訓練資料，2000 筆 (1 組) 作為測試資料；接著分別計算 Support (次數) 以及 Confidence (置信度)，再依照預設的最小限制量分別移除小於 Occur_Min (最小出現次數) 及 Confidence_Min (最小出現機率) 的資料以去除雜訊避免干擾實驗結果，最後再將濾除雜訊後的資料依 Support 及 Confidence 排序後取不同的組距資料分別交叉以 Exact Match 及 Partial Match 的方式進行 5-Fold-CV 實驗並取其平均值作為實驗結果，分析如下。

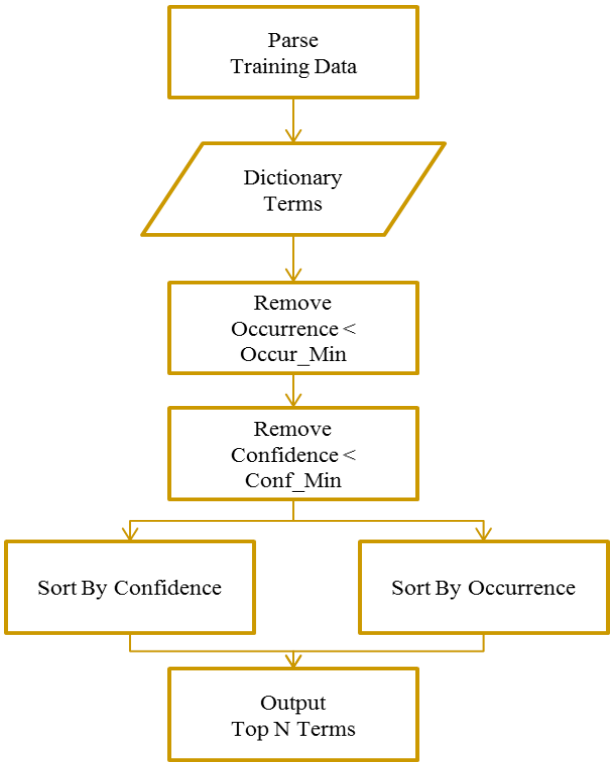
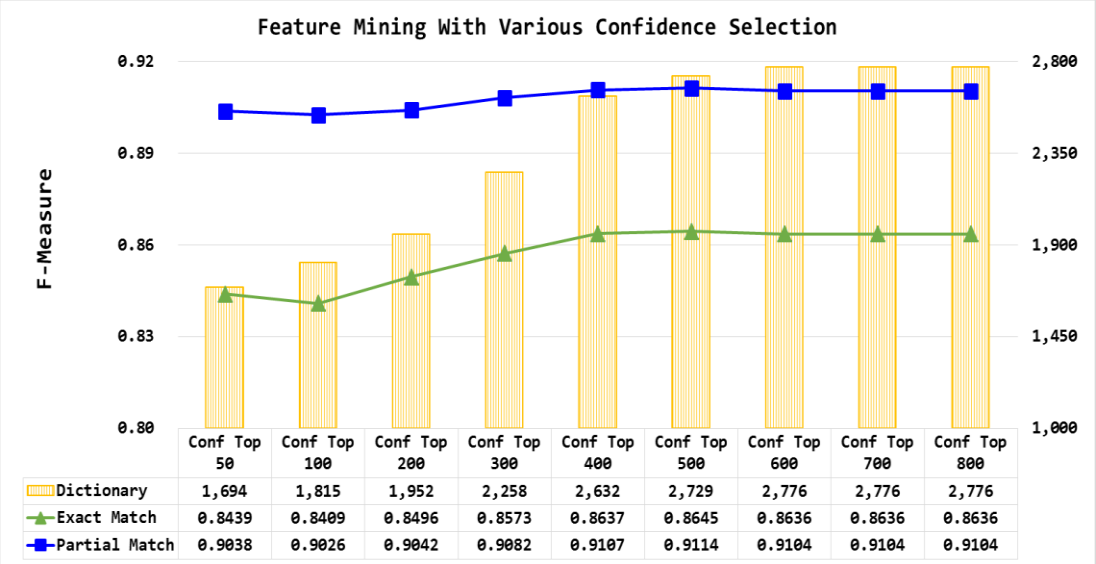


圖 10 Dictionary Terms Mining 實驗流程圖

圖 11 為分別使用 Exact Match 與 Partial Match 兩種方式針對不同的 Confidence 值進行實驗的結果，從圖中可發現，Exact Match 與 Partial Match 的 F-Measure 以及被選作特徵值的 Dictionary Terms 在 Top 500 前均隨著 N 的增加而呈現上升趨勢，並在使用 TOP 500 作為特徵值篩選條件時達到最高，F-Measure 分別為 0.8645 以及 0.9114，Dictionary Terms 則達到 2,729，然而當 N>500 後 F-Measure 與選作特徵值的 Dictionary Terms 則呈現下降後水平不變的趨勢，其原因在於當 N 超過 500 後雖然可納入的候選

字串變多，但由於其置信度過低，反而容易納入過多被視為雜訊的候選特徵值，由此可知，N 的選擇應以「最適」而非「最多」為原則。基於上述驗證結果，在 Confidence 的實驗中採用 TOP 500 作為擷取條件。



	Conf Top 50	Conf Top 100	Conf Top 200	Conf Top 300	Conf Top 400	Conf Top 500	Conf Top 600	Conf Top 700	Conf Top 800
Dictionary	1,694	1,815	1,952	2,258	2,632	2,729	2,776	2,776	2,776
Exact Match	0.8439	0.8409	0.8496	0.8573	0.8637	0.8645	0.8636	0.8636	0.8636
Partial Match	0.9038	0.9026	0.9042	0.9082	0.9107	0.9114	0.9104	0.9104	0.9104

圖 11 Feature Mining With Various Confidence Selection

圖 12 為分別使用 Exact Match 與 Partial Match 兩種方式針對不同的 Support 值進行實驗的結果，從實驗結果中可發現雖然被選作特徵值的 Dictionary Terms 隨著 N 的增加呈現上升趨勢，但 Exact Match 與 Partial Match 實驗結果的 F-Measure 卻除了在 N=100 前微幅從 0.8469 上升至最高 0.8480 以及 0.9084 上升至最高 0.9098 外，其他不同 Support 個數組別的 F-Measure 皆隨著 N 的增加呈現急下降趨勢，並在 N=500 時達到最低，分別是 Exact Match 的 0.7529 以及 Partial Match 的 0.7994。由此可見，以 Support 的方式從 Dictionary 中選取特徵值的效能並未與 N 呈正相關，而這樣的結果也呼應了我們再進行 Confidence 實驗時的推斷，頻率低的特徵值對訓練模型而言形同於雜訊，容易造成系標記時發生誤判的情況。

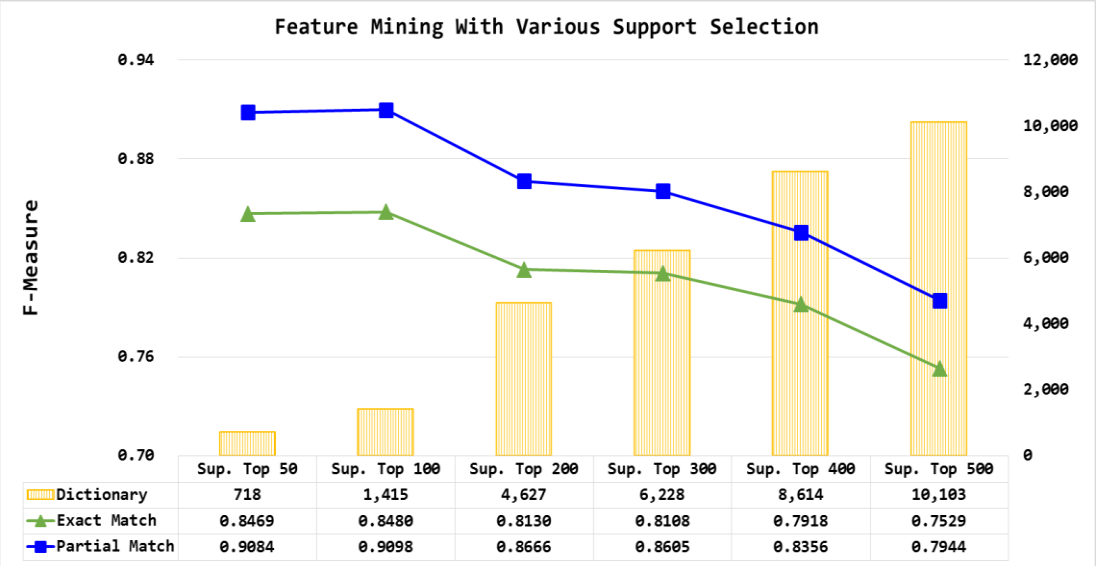
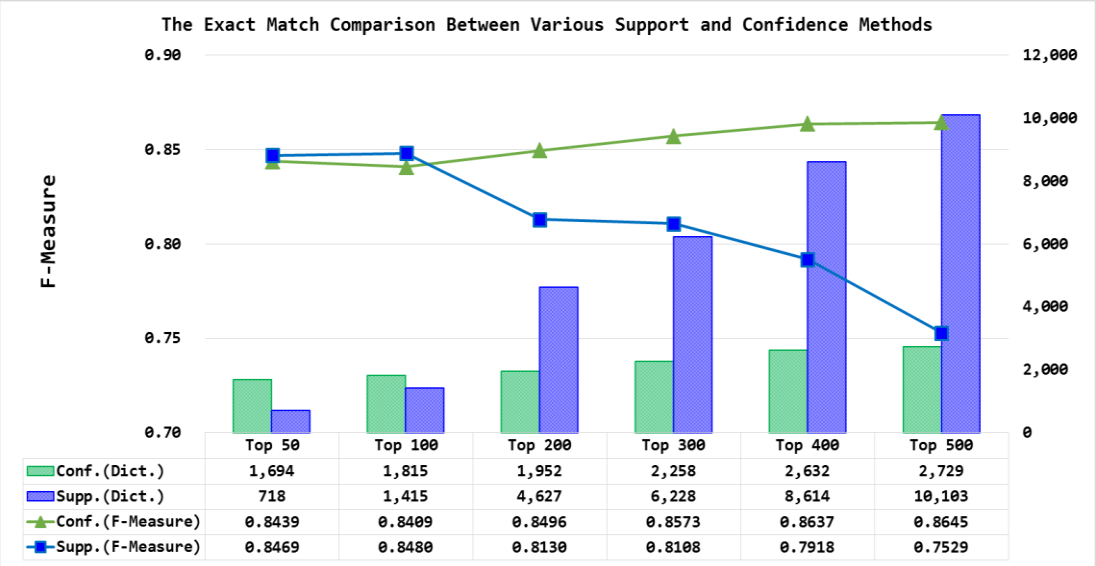


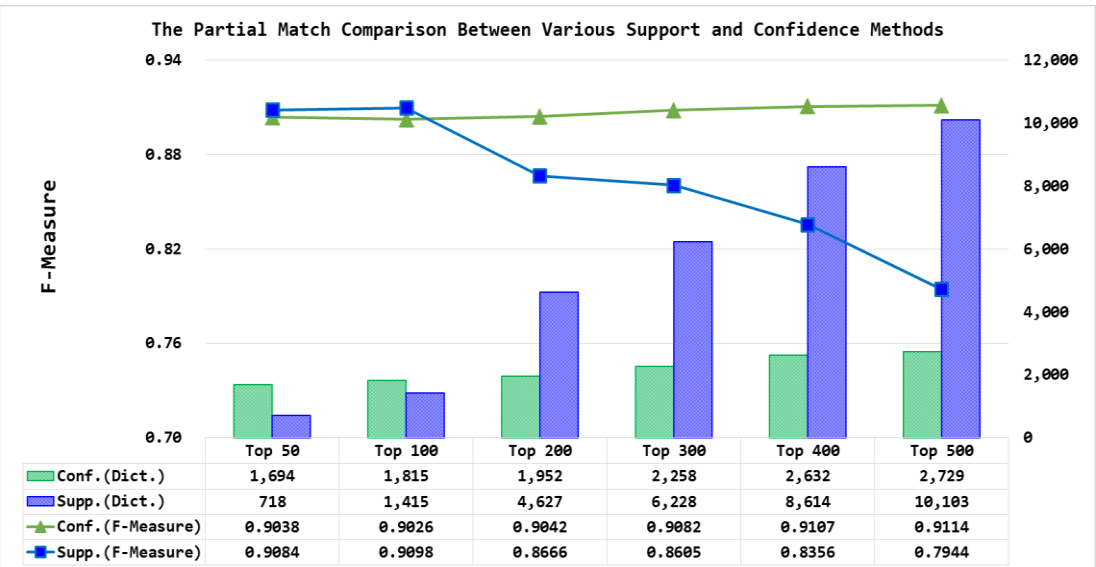
圖 12 Feature Mining With Various Support Selection

最後，我們將 Confidence 與 Support 兩種實驗方法所得的 F-Measure 繪成圖表(如下圖 13-14)；經分析顯示，Confidence 在 Top 500 時 Exact Match 與 Partial Match 所得的 F-Measure 達到最高 0.8645 與 0.9114，此結果均大於 Support 於 Top 100 達最高時的 F-Measure 0.8480 與 0.9098，且 Confidence 在 Exact Match 各組的 F-Measure 僅於 Top 50 與 100 時以 0.8439 與 0.8409 略低於 Support 的 F-Measure 0.8469 與 0.8480；而 Partial Match 所得的 F-Measure 亦有相同的現象，分別是 Confidence 以 0.9038 與 0.9026 略低於 Support 的 F-Measure 0.9084 與 0.9098; 以平均值來看 Confidence 在 Exact Match 與 Partial Match 的平均 F-Measure 分別是 0.8533、0.9068 均大於 Support 的 0.8106、0.8626。而 Confidence 的標準差分別為 0.0092、0.0035 均小於 Support 的 0.0327、0.0403。此外從圖中趨勢可見，Confidence 的 F-Measure 自 Top 100 後隨 N 的增加效能逐漸提升至最高 0.8645，但相反的 Support 卻呈現逐漸下降的趨勢至 Top 500 時降至最低 0.7529；綜合上述，本研究選定以 Confidence (置信度) Top 500 作為選取特徵值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 13 The Exact Match Comparison Between Various Support and Confidence Methods



	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)	0.9038	0.9026	0.9042	0.9082	0.9107	0.9114
Supp.(F-Measure)	0.9084	0.9098	0.8666	0.8605	0.8356	0.7944

圖 14 The Partial Match Comparison Between Various Support and Confidence Methods

2. CKIP Information Selection

在前處理步驟的人工標記時發現餐廳類別的組成其詞性隱含著「形容詞」+「名詞」與「名詞」+「名詞」的規則，且為了解決標記食物層級的問題因而參考 CKIP 斷詞後再經由人工修正的結果並以啟發式規則「longest」進行標記，因此在特徵選取前我們先進行 CKIP Information Selection 實驗來測試加入 CKIP 斷詞後所得的資訊對系統效能的影響。

實驗步驟如圖 15 所示，先將文章標題透過 CKIP 取得斷詞的結果 (Word Segmentation, WS) 以及詞性 (Part of Speech, POS)，再將 WS 以及 POS 結合原有的特徵 Eng. (英文字母+數字) 以及 Sym. (符號) 後組成 9 種組合進行實驗，分別是 「WS 」、「WS + Eng.」、「WS + Sym.」、「WS+ Eng + Sym」、「POS」、「POS + Eng」、「POS + WS」、「POS + Eng + WS」；實驗方式亦採 5-Fold CV 進行，選取 8,000 筆 (4 組) 作為訓練資料， 2000 筆 (1 組) 作為測試資料交叉進行 Exact Match 以及 Partial Match 實驗後再取其平均值，實驗結果分析於下。

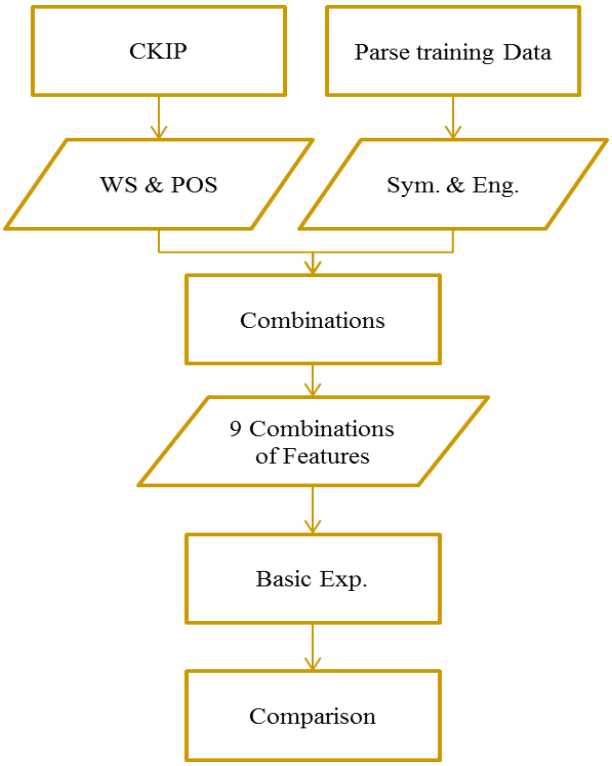
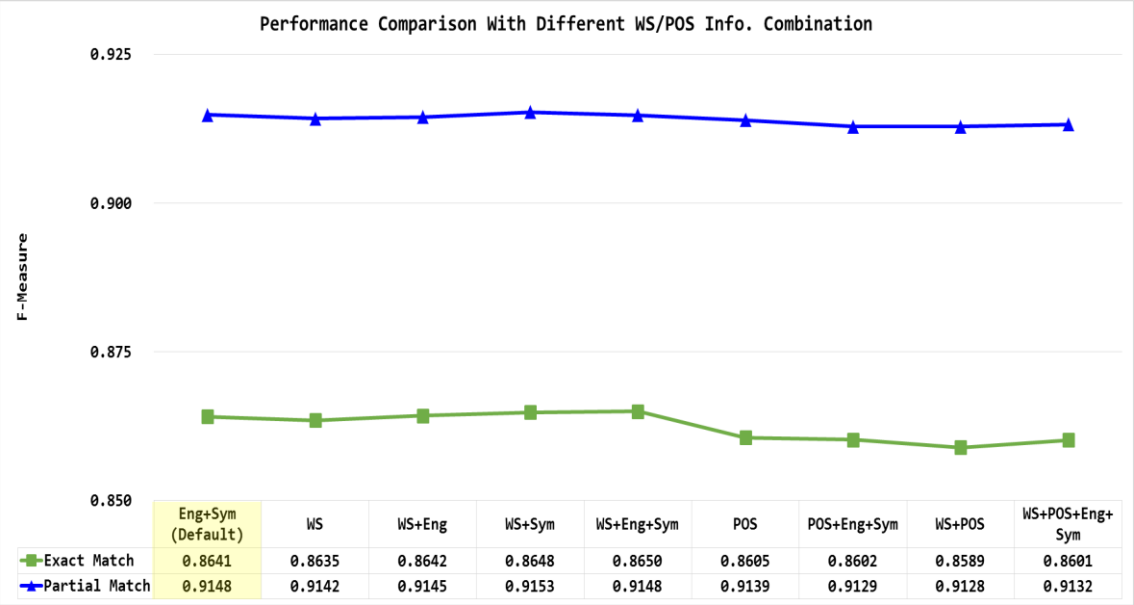


圖 15 CKIP Information Selection 實驗流程圖

從圖 16 實驗結果顯示，加入 POS 後組別的 F-Measure 無論透過 Exact Match 或是 Partial Match 皆小於原始「Eng + Sym」的 0.8641 以及 0.9148，因此我們先排除將 POS 加入特徵的可能性。分析 POS 導致系統效能變差的原因發現，主要原因在於人工標記時並非完全直接參照 CKIP 的斷詞結果，而是遇到食物層級問題時才進行參考; 此外部分 CKIP 斷詞的結果在參照前須先進行人工修正以符合標記需求，舉例而言，人工標記的參考規則為名詞(N) + 名詞(N) 或是形容詞(A)+名詞(N)，而「炒年糕」在 CKIP 的斷

詞結果為「炒 (V，動詞)」＋「年糕 (N，名詞)」，因此若未經過修正則「炒」就不會被標記，但對 longest 的標記原則而言「炒年糕」是一種獨立可區分類別的食物，因此在標記時會先修正 CKIP 的斷詞結果。

排除了 POS 後，針對加入「WS」的組合分析顯示，Exact Match 以及 Partial Match 的 F-Measure 均以「WS+Eng+Sym」組合最高，分別是 0.8650 以及 0.9148，其中，Exact Match 的 F-Measure 略高於原始「Eng + Sym」的 0.8641 但差異僅 0.0009，而在 Partial Match 時則是與原始值相同為 0.9148。 觀察趨勢發現，加入 WS 與 POS 後的各種組合其 Exact Match 以及 Partial Match 所得的 F-Measure 變化曲線平滑，其中加入 WS 的組合呈現持平或微幅升降的趨勢，而加入 POS 的組合其結果則均低於原始數據，由此可見 WS 與 POS 兩種特徵值對於提升系統效能並無太大的幫助，基於此結論本文在後續的實驗中將不加入 CKIP 的斷詞結果而僅以表 3 所設計的 15 種特徵值進行實驗。



	Eng+Sym (Default)	WS	WS+Eng	WS+Sym	WS+Eng+Sym	POS	POS+Eng+Sym	WS+POS	WS+POS+Eng+Sym
Exact Match	0.8641	0.8635	0.8642	0.8648	0.8650	0.8605	0.8602	0.8589	0.8601
Partial Match	0.9148	0.9142	0.9145	0.9153	0.9148	0.9139	0.9129	0.9128	0.9132

圖 16 Performance Comparison With Different WS/POS Info. Combination

4-2-2 Supervised Experiment

完成 Feature Mining 實驗後，依實驗結果以 10,000 筆已標記資料 (L) 搭配 87,282 筆未標記資料進行 Basic 以及 Tri-Training 實驗，分述如下：

1. Basic

依照 Dictionary Terms Mining 以及 CKIP Information Selection 的結果，我們以 Confidence Top 500 作為特徵值的篩選條件，並且以表 3 所設計的 15 種特徵值將 10,000 筆已完成人工標記的資料透過 5-Fold CV 的方式分別進行 Exact Match 以及 Partial Match 實驗，並將實驗結果取平均後以 X 軸為訓練資料量，Y 軸為 F-Measure 的平均值繪製成圖 17。結果顯示，當訓練資料量達 8,000 筆時 Exact Match 的 F-Measure 最高可達 0.8645 ; Partial Match 的 F-Measure 最高可達 0.9114；此外 Partial Match 實驗在各資料量組距的平均 F-Measure 均高於 Exact Match，這是由於 Partial Match 採部分給分的方式計算，如此的結果符合我們的預期也貼近本文所研究的「餐廳類別」其組成具有可選擇性的特色。

觀察曲線趨勢發現，兩種驗證方法所得的 F-Measure 皆隨著訓練資料量的增加而提升，其中 Exact Match 的提升斜率在 2,000 至 4,000 筆資料時較為顯著，從 0.7542 提升至 0.8082，當資料量增加至 8,000 筆時達最高 0.864。而在 Partial Match 方面的提升斜率趨勢與 Exact Match 相符但相較於 Exact Match 則較為趨緩，斜率在 2,000 至 4,000 筆資料時較為顯著，從 0.8474 提升至 0.8866，並在資料量增加至 8,000 筆時達最高 0.9114。由上述分析可知，系統的效能會隨著資料量的增加而提升，並且在訓練資料量達 8,000 筆時 F-Measure 達到最高。

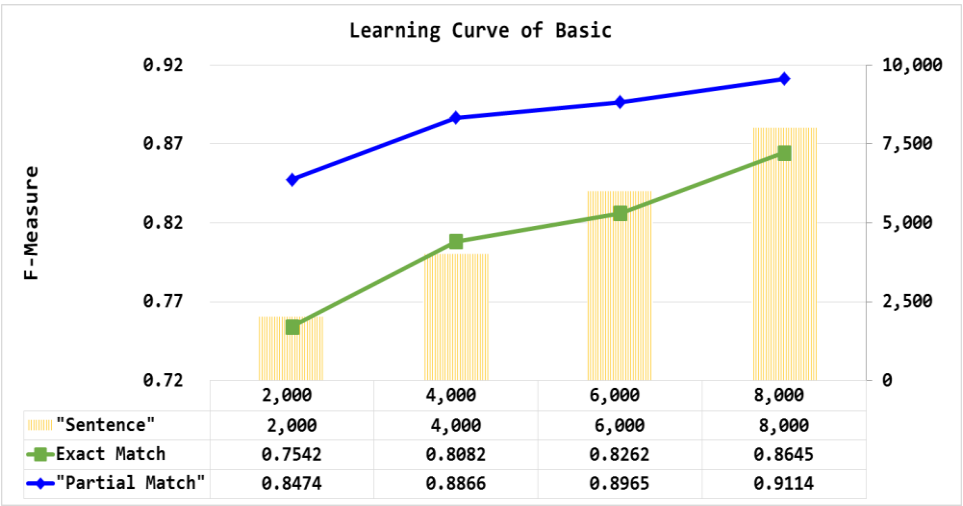


圖 17 Learning Curve of Basic

2. Tri-Training

依照上述 Basic 的實驗架構，我們再加入自 PTT FOOD 版上所擷取關於 [食記] 的未標記資料共 87,282 筆以 Tri-Training 的方式從 U (Un-Labeled Data) 中加入更多的訓練資料試著提升系統效能；實驗方式亦採 5-Fold-CV 的方式進行，其中考量到U的資料量較龐大且 Tri-Training 自 U 中選取資料採隨機的方式進行，為避免實驗數據受U中離群值(Outliers) 影響，因此在每一回合的 5-Fold-CV 中均再進行 5 次的 Tri-Training 並取其平均，總計實驗共執行 25 次 Tri-Training，流程如圖 18 所示。

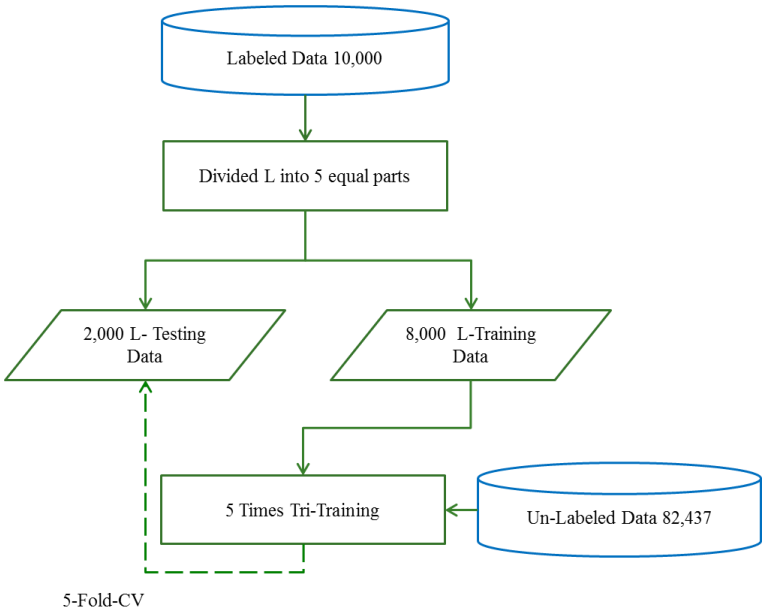


圖 18 Tri-Training 實驗流程圖

Tri-Training 的實驗數據結合 Basic 的實驗結果如圖 19。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 訓練資料量 8,000 筆的 2.2 倍，而其 Exact Match 的平均 F-Measure 0.8685 較 Basic 的 0.8645 微幅提升 0.0004，Partial Match 的平均 F-Measure 0.9168 較 Basic 的 0.9114 微幅提升 0.0054，結果顯示雖然 Tri-Training 的實驗結果相較於 Basic 僅微幅提升，但這也達到 Tri-Training 利用較少的已標記資料從龐大的 U 中獲取訓練資料來訓練模型的目的。

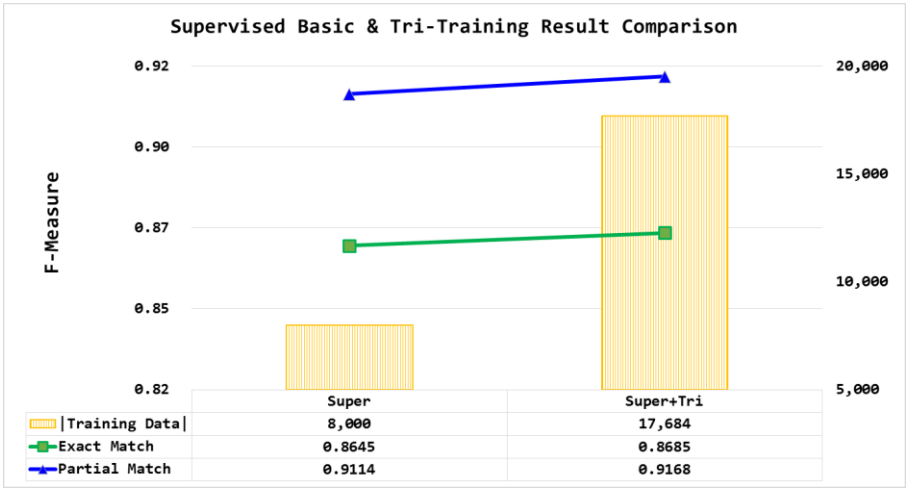


圖 19 Supervised Basic & Tri-Training Result Comparison

4-2-3 Semi-Supervised Experiment

我們以 Basic 實驗的結果為基準，利用人工標記 10,000 筆資料所得的 1,563 個 Entity 為種子 (Seeds) 進行半監督式 Distance 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；實驗分成兩個部分，分別針對 L (Labeled Data) 及 U (Un-Labeled Data) 自動標記後進行測試，實驗流程如下圖 20 所示。

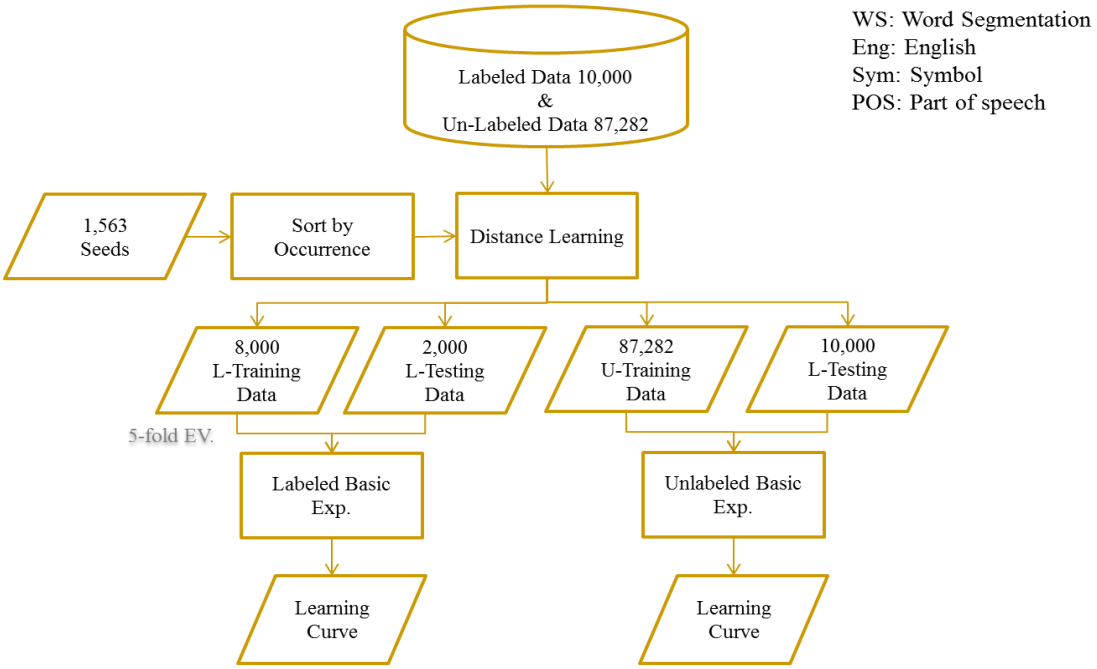


圖 20 Semi-Supervised 實驗流程圖

實驗首先統計各 Entity 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 21，圖中顯示 1,563 個 Seeds 的總出現次數為 7,961 次，而前 500 個 Entity 的累積出現次數已達 6,738 次，亦即前 31.9% 的 Entity 其出現次數占了總出現次數的 84.6%，由此顯示利用頻率較高的前幾個 Entity 作為 Seeds 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排序後的 Seeds 出現次數為組距，依順序由高至低取 Seeds 進行自動標記，再將實驗結果依組距區分後繪製成 Learning Curve 後觀察不同 Seeds 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 實驗及 Tri-Training 實驗進行比較。

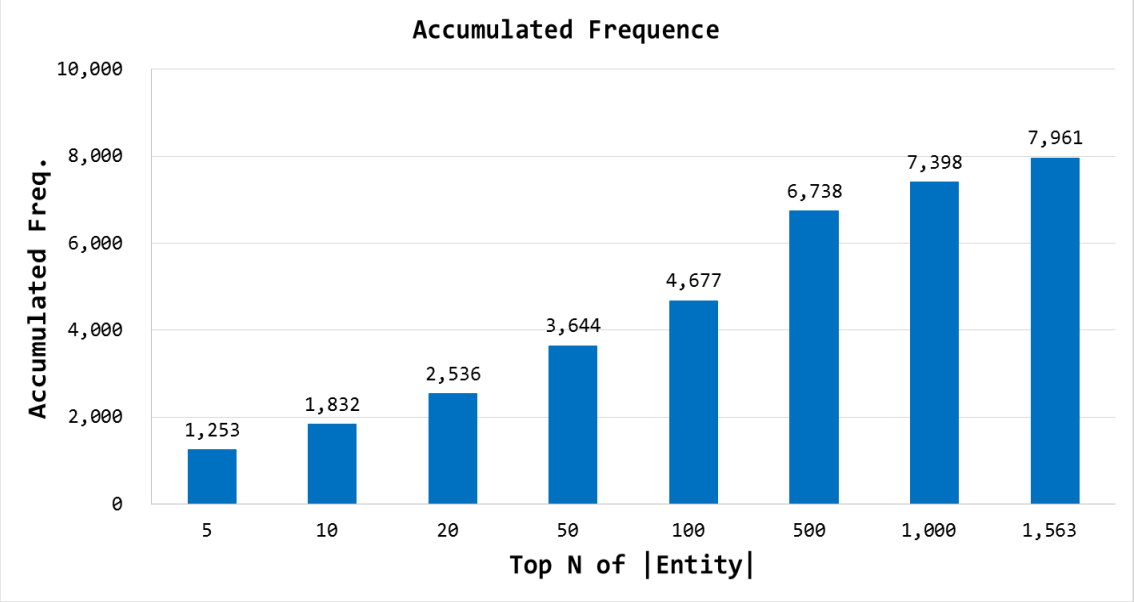
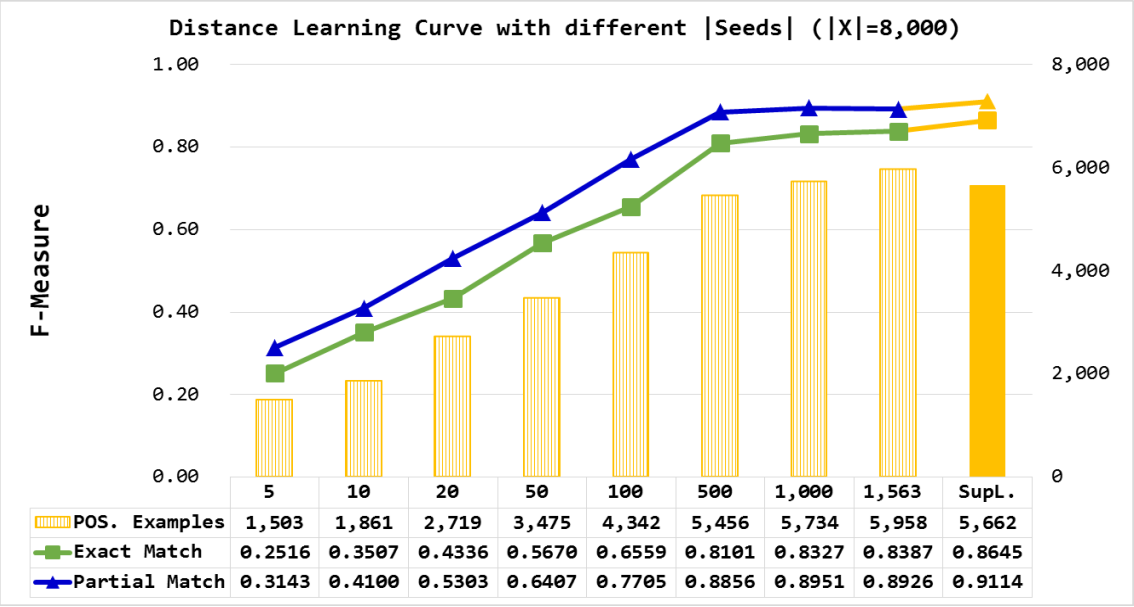


圖 21 Entity 出現次數累計圖

實驗的第一個部分先利用 10,000 筆已標記資料分成 8,000 筆測試資料以及 2,000 筆訓練資料進行 5-Fold-CV 實驗，首先清除 8,000 筆訓練資料的標記結果，再依照上述排序後的 Seeds 依組別對訓練資料進行自動標記後訓練模型，接著利用訓練後的模型對 2,000 筆測試資料交叉進行 Exact Match 以及 Partial Match 實驗，實驗結果取 F-Measure 的平均值整理如下圖 22。從結果可見 Exact Match 的 F-Measure 在 Seeds 量 1,563 時可達最高 0.8387，與人工標記進行 Basic 實驗的結果 0.8645 僅差距 0.025；而 Partial Match 的 F-Measure 亦在 Seeds 量 1,563 時可達最高 0.8926，與 Basic 的實驗結果 0.9114 僅差

距 0.018；由此可見利用 Distant Learning 的方式以 Seeds 進行自動標記的結果其效能雖然微幅低於人工標記，但卻可大幅節省人工標記所需耗費的時間與成本。此外，從實驗結果中可發現當 Seeds 量達到 500 時，Exact Match 的實驗結果 F-Measure 已可達 0.8101，Partial Match 可達 0.8856，意即利用 31% 的 Seeds 量即可分別達到 96% 及 99% 的效能，此結果呼應上述在實驗前先針對 Seeds 的出現頻率進行統計排序後利用頻率較高的前幾個 Seeds 即可完成相當大比例的資料標記，並且避免系統受到頻率較低被視為雜訊的 Seeds 所干擾而達到顯著的系統效能。

觀察曲線趨勢可發現，Exact Match 及 Partial Match 的 F-measure 皆隨著 Seeds 量的增加而提升，其中 Exact Match 從最低 0.2516 上升至最高 0.8387，而 Partial Match 從 0.3143 上升至最高 0.8926；觀察其趨勢線可發現，Exact Match 及 Partial Match 的 F-Measure 斜率皆自 Seeds 數量 500 後開始明顯變小呈現趨緩狀態，這樣的現象呼應上述當 Seeds 量達到前 500 時即可讓系統效能分別在 Exact Match 及 Partial Match 達到 96% 及 99% 的效能。

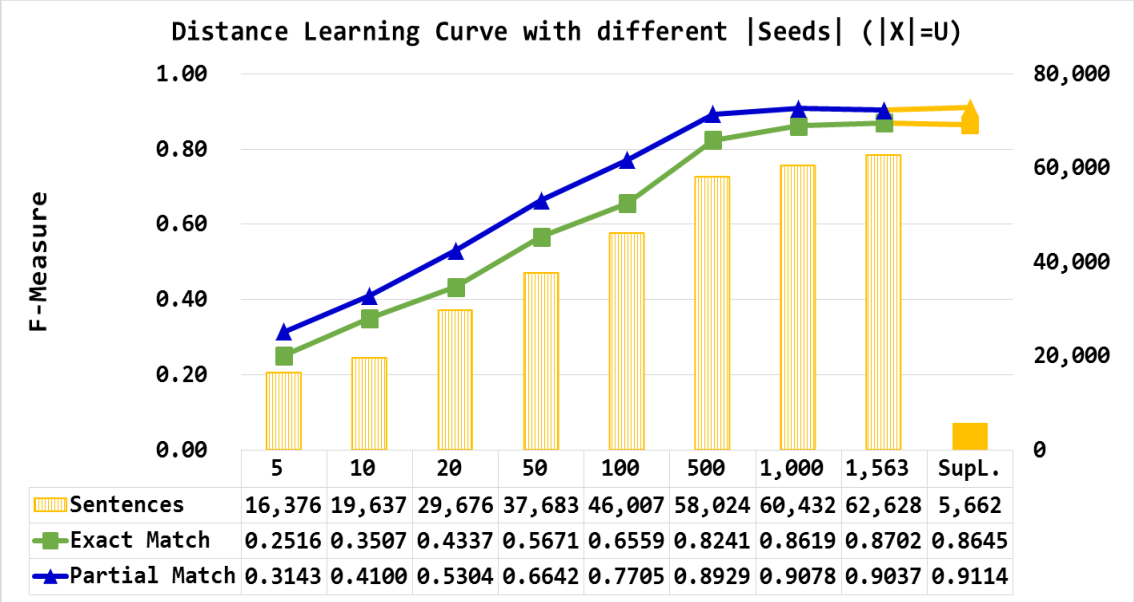


	5	10	20	50	100	500	1,000	1,563	SupL.
POS. Examples	1,503	1,861	2,719	3,475	4,342	5,456	5,734	5,958	5,662
Exact Match	0.2516	0.3507	0.4336	0.5670	0.6559	0.8101	0.8327	0.8387	0.8645
Partial Match	0.3143	0.4100	0.5303	0.6407	0.7705	0.8856	0.8951	0.8926	0.9114

圖 22 Distance Learning Curve with different |Seeds| (|X|=8,000)

實驗的第二個部分為利用排序後的 1,563 個 Seeds 依各組別的 Seeds 量對 87,282 筆未標記資料 (U) 進行自動標記以作為訓練資料訓練模型，接著利用 10,000 筆已標記

資料做為測試資料進行 Exact Match 以及 Partial Match 實驗，實驗結果取 F-Measure 的平均值整理如下圖 23，其 Exact Match 的 F-Measure 在 Seeds 量 1,563 時可達最高 0.8702 而 Partial Match 的 F-Measure 亦在 Seeds 量 1,563 時可達最高 0.9037。分析後發現，Distance Learning 的 Exact Match 及 Partial Match 實驗效能有著相同的趨勢，即其 F-measure 皆隨著 Seeds 量的增加而提升，且在 Seeds 量 500 後提升幅度趨於平緩並在 Seeds 量 1563 時 F-Measure 達到最高，而 U-Distance-Learning 在 Exact Match 及 Partial Match 的最高 F-measure 0.8702、0.9037 均微幅高於 L-Distance-Learning 的 0.8387、0.8926，由此顯示，利用 U 以 Seeds 自動標記後作為訓練資料使得訓練資料量增加對系統的效能可達微幅提升的效果。



	5	10	20	50	100	500	1,000	1,563	SupL.
Sentences	16,376	19,637	29,676	37,683	46,007	58,024	60,432	62,628	5,662
Exact Match	0.2516	0.3507	0.4337	0.5671	0.6559	0.8241	0.8619	0.8702	0.8645
Partial Match	0.3143	0.4100	0.5304	0.6642	0.7705	0.8929	0.9078	0.9037	0.9114

圖 23 Distance Learning Curve with different |Seeds| (|X|=U)

五、結論與未來工作

美食資訊與日常生活息息相關，而 2017 已進入人手一機的行動裝置的時代，大量的適地性服務（LBS）也隨之誕生，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。而過去關於美食類別或是餐廳類別擷取的相關研究，大多著墨於以美食辭典的方式在結構化的網頁中比對後進行資訊擷取，這樣的方式對於美食種類的多變性有所限制，而餐廳的類別也並不一定侷限於美食名稱，因此難以適用於非結構化即時更新的資料來源；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，並使用監督式學習與半監督式學習的方式分別進行 Basic 、Tri-Training 與 Distance Learning 實驗，整理如圖 24-25，由圖可見，利用監督式學習進行 Basic 實驗需藉由人力進行資料標記，在可標記的資料量有限的情況下，以 Basic 實驗的 F-Measure 0.8645、0.9114 為基礎，分別以 Tri-Training 及 Distance Learning 的方式增加訓練資料，結果顯示在 Exact Match 的實驗中 Tri-Training 與 Distance Learning-U 皆可微幅提升實驗的 F-Measure，最高可達 0.8702，而在 Partial Match 的實驗中雖然僅 Tri-Training 的 F-Measure 較 Basic 高，最高可達 0.9186，但 Distance Learning-U 的實驗結果也僅微幅低於 Basic 用人工標記的效能，但其可有效標記的資料量達 62,628 以占總資料量 87,282 的 71.7%，這樣的結果以十分接近在前處理時由人工分析 1,000 筆資料時所得的 72.5% 甚至高於人工標記 10,000 筆資料時所得的 70.6%，這樣的結果或許無法完全精確的辨識確切的美食類別，但對於餐廳類別關鍵字的判讀已可達一定的水準。

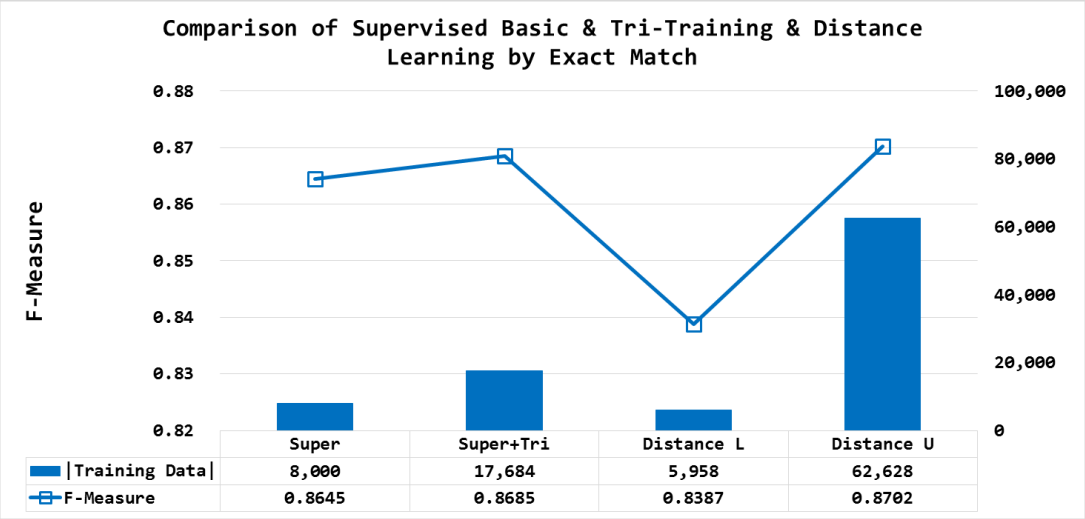


圖 24 Comparison of Basic & Tri-Training & Distance Learning by Exact Match

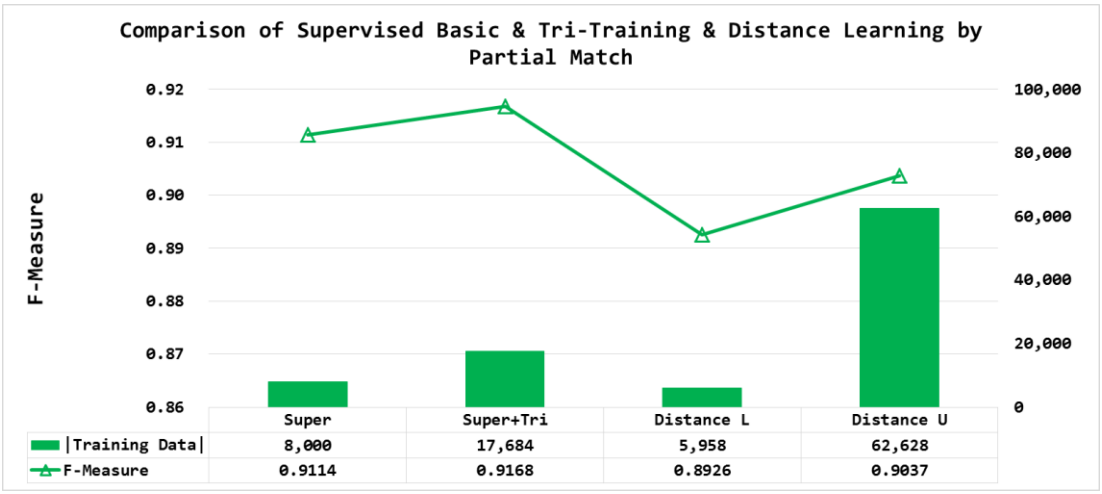


圖 25 Comparison of Basic & Tri-Training & Distance Learning by Partial Match

本論文的研究過程中由人工進行標記，並以監督式學習的方式完成訓練模型，雖然在訓練資料的品質上較為完善，但從實驗中我們發現，當訓練資料量越大時系統的辨識率也會跟著提昇，而人工標記相對耗時的缺點無法適用於龐大資料的處理，因此後續建議可利用非監督式學習並以自動標記的方式改善人工標記耗時的問題並達提升效能的目的。此外在 PTT Food 版內的文章並不完全是推薦文，其中亦存在著一定數量的反面意見文章，雖然本文已擷取出文章的 URL 可供使用者點選後閱讀以辨識，但若可再加入 Opinion Mining 的概念自動判斷文章所要傳達的正反向意見，並將本文所擷取的地址、電話...等相關資訊與地圖家以結合，相信可使系統更趨完備，提供使用者更便利的查詢介面。

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 McCallum and Wei Li. 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons. 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, 188–191.

[6] Z. Suxiang, Z. Suxian and W. Xiaojie, "Automatic Recognition of Chinese Organization Name Based on Conditional Random Fields," Natural Language Processing and Knowledge Engineering, pp. 229-233, 2007.

[7] Xiying, "A METHOD OF CHINESE ORGANIZATION NAMED ENTITIES RECOGNITION BASED ON STATISTICAL WORD FREQUENCY, PART OF SPEECH AND LENGTH," Broadband Network and Multimedia Technology (IC-BNMT), pp. 637-641, 2011.

[8] L. Yajuan, Y. Jing and H. Liang, "Chinese Organization Name Recognition Based on Multiple Features," Pacific Asia conference on Intelligence and Security Informatics, pp. 136-144, 2012.

[9] Y. Xiying, "A METHOD OF CHINESE ORGANIZATION NAMED ENTITIES RECOGNITION BASED ON STATISTICAL WORD FREQUENCY, PART OF SPEECH AND LENGTH," Broadband Network and Multimedia Technology (IC-BNMT), pp. 637-641, 2011.

[10] L. Yajuan, Y. Jing and H. Liang, "Chinese Organization Name Recognition Based on Multiple Features," Pacific Asia conference on Intelligence and Security Informatics, pp. 136-144, 2012.

[11] Andrew Eliot Borthwick. 1999. A Maximum Entropy Approach to Named Entity Recognition. Ph.D. Dissertation. New York, NY, USA. Advisor(s) Grishman, Ralph. AAI9945252.

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

[13] 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.

[14] L. D. John , M. Andrew and N. C. Fernando, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," ICML Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282-289, 2001.

[15] Z. Suxiang, Z. Suxian and W. Xiaojie, "Automatic Recognition of Chinese Organization Name Based on Conditional Random Fields," Natural Language Processing and Knowledge Engineering, pp. 229-233, 2007.

[16] Y. Xiying, "A METHOD OF CHINESE ORGANIZATION NAMED ENTITIES RECOGNITION BASED ON STATISTICAL WORD FREQUENCY, PART OF

SPEECH AND LENGTH," Broadband Network and Multimedia Technology (IC-BNMT), pp. 637-641, 2011.

[17] L. Yajuan, Y. Jing and H. Liang, "Chinese Organization Name Recognition Based on Multiple Features," Pacific Asia conference on Intelligence and Security Informatics, pp. 136-144, 2012.

[18] C.-W. Wu, R. T.-H. Tsai and W.-L. Hsu, "Semi-joint labeling for chinese named entity recognition," Proceedings of the 4th Asia information retrieval conference, pp. 107-116, 2008.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] Ning Yu and Sandra Kubler. 2010. Semi-supervised Learning for Opinion Detection.

[25] Rie Kubota Ando and Tong Zhang. 2005. A High-performance Semi-supervised Learning Method for Text Chunking. In Proceedings of the 43rd Annual Meeting on Association

for Computational Linguistics (ACL '05). Association for Computational Linguistics, Stroudsburg, PA, USA, 1–9.

[26] Kamal Nigam and Rayid Ghani. 2000. Analyzing the Effectiveness and Applicability of Co-training. In Proceedings of the Ninth International Conference on Informa675 tion and Knowledge Management (CIKM '00). ACM, New York, NY, USA, 86–93.

[27] Tri Thanh Nguyen, Le Minh Nguyen, and Akira Shimazu. 2008. Using Semi-supervised Learning for Question Classification. Information and Media Technologies 3, 1 (2008), 112–130.

[28] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning Syntactic Patterns for Automatic Hypernym Discovery. In Advances in Neural Information Processing Systems 17, L.K. Saul, Y. Weiss, and L. Bottou (Eds.). MIT Press, 1297–1304.

[29] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant Supervision for Relation Extraction Without Labeled Data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2 (ACL '09). Association for Computational Linguistics, Stroudsburg, PA, USA, 1003–1011.

[30] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD '08). ACM, New York, NY, USA, 1247–1250.

[31] Matthew Michelson and Craig A. Knoblock. 2009. Exploiting Background Knowledge to Build Reference Sets for Information Extraction. In Proceedings of the 21st International Jont Conference on Artifical Intelligence (IJCAI'09). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2076–2082.

[32] Joohui An, Seungwoo Lee, and Gary Geunbae Lee. 2003. Automatic Acquisition of Named Entity Tagged Corpus from World Wide Web. In Proceedings of the 41st

Annual Meeting on Association for Computational Linguistics - Volume 2 (ACL'03). Association for Computational Linguistics, Stroudsburg, PA, USA, 165–168.

[33] Adam Rae, Vanessa Murdock, Adrian Popescu, and Hugues Bouchard. 2012. Mining the Web for Points of Interest. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12). ACM, New York, NY, USA, 711–720.

[34] Ruiji Fu, Bing Qin, and Ting Liu. 2011. Generating chinese named entity data from a parallel corpus. In In Proceedings of 5th International Joint Conference on Natural Language Processing. 264–272.