Computational Linguistics and Chinese Language Processing
Vol. 19, No. 4, December 2014, pp. 1-18

POI 擷取:商家名稱辨識與地址配對之研究

POI Extraction from the Web: Store Name Recognition and Address Matching

林育暘、張嘉惠

Lin Yu-Yang* and Chang Chia-Hui*

摘要

行動化是 2014 的趨勢之一，而適地性服務（Location-based Service）在這波趨勢中具有至關重要的地位，因為裝置行動化的因素，大量查詢需求因此誕生，例如：路線導航、查詢附近餐廳、加油站等。適地性服務要能廣泛的提供服務，通常需要有一個完整的 POI（Point of Interest）資料庫，而整個網路就是最大的資訊來源。這些資料源自於網站管理者、群眾外包（crowdsourcing）或個人使用者所分享的資訊，包括了地址、名稱、電話、評論等資訊。現在雖然有各種擷取地址相關資訊的方法，但經常面臨無法取得明確 POI 的名稱，在資訊檢索上受到很大的限制。

在本篇論文中，我們提出一個商家名稱辨識的方法，藉由收集網路上包含地址的網頁，建立一個具有商家名稱與地址關聯性的資料庫，以提高地址相關資訊檢索的效果，讓使用者在使用行動裝置查詢時，能直接輸入店家名稱或關鍵字查詢地址之服務，提供便利的查詢功能。其中，在商家命名實體辨認上，本篇論文提出了商家與組織名稱在命名上的共同特性，利用此共同特性當作特徵加入 CRF 模型，以提供 N-Gram 與詞性之外的特徵。

關鍵詞：商家地理資訊檢索、商家名稱擷取、商家名稱與地址配對、序列標記、條件隨機域

*國立中央大學資訊工程學系
Department of Computer Science and Information Engineering, National Central University
E-mail: 101522052@cc.ncu.edu.tw; chchang8ncu@gmail.com

Abstract

Mobility is one of the trends in 2014. According to the report of IDC (International Data Corporation), the worldwide shipments of tablets have exceeded PCs in 2013 Quarter 4, while smart phones has already exceeded other devices in unit shipments and market ratio. With this trend, many location-based services (LBS) have been proposed, for example, navigation, searching restaurants or gas stations. Therefore, how to construct a large POI (Point-of Interest) database is the key problem. In this paper, we solve three problems including Taiwan address normalization, store name extraction, and the matching of addresses and store names. To train a statistical model for store name extraction, we make use of existing store-address pair to prepare training data for sequence labeling. The model is trained using common characteristics from store names in addition to POS tags. When testing on search snippets, we obtain 0.791 F-measure for store name recognition.

Keywords: POI, Store Name Extraction, Name-address Matching, Sequence Labeling, Conditional Random Field

1. 緒論

根據國際數據資訊 IDC 於 2013 年 9 月調查報告顯示，平板電腦的出貨量在 2013 年第四季首次超過個人電腦，而智慧型手機不論在出貨量或市佔率早就遠遠超過桌上型電腦和可攜式電腦的總和，IDC 甚至預測平板電腦的出貨量將在 2015 年超過桌上型電腦和可攜式電腦的總和。這顯示了行動裝置的普及是一種不可抵擋的趨勢。

行動裝置的普及造就大量地域性查詢的需求，其中最常見的一種查詢，就是尋找附近的餐廳或加油站，根據 Google 於 2013 年第一季台灣智慧型手機使用行為調查(多選題)，搜尋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%）。然而當使用者在電子地圖上搜尋這些地點名稱 (POI，Point of Interest)時，經常無法找到，因為電子地圖上雖有地點名稱標註，但是相關資訊不足，而這些資訊其實大多可以在網頁中找到。因此使用者大多必須開啟瀏覽器搜尋商家名稱找出地址，並把地址輸入至電子地圖查詢路線。但行動裝置螢幕小，且輸入文字不便利，如果要反覆查詢將是一件耗時耗力的工作。如果這時候有一個商家地理資訊系統能事先將網路上的商家資訊進行整合，最後提供一個 APP 直接讓使用者查詢，將可以大幅度減少使用者與裝置間的互動次數，有效的提供搜尋的便利性。

為建構商家地理資料庫，Chuang 等人(Chuang *et al*., 2014)對於包含地址網頁提出以廣度優先搜尋、黃頁爬蟲、與地址樣版查詢三種抓取程式，並利用 Chang 等人(Chang *et al*., 2012)的地址擷取程式，取得大量中文地址。Li 與 Chang(2009)並定義地址相關資訊擷取問題，希望藉此豐富每個 POI 的相關資訊，提高地理資訊檢索(Geographical Information

Retrieval, GIR)的召回率。然而不論是 Li 與 Chang(2009)或 Chang 等人(Chang *et al.*, 2012)或 Chuang 等人(Chuang *et al*., 2014)的相關資訊擷取方法都僅能從多筆地址網頁擷取資訊，所得資訊有限，對於單筆地址網頁的相關資訊擷取仍尚無研究。

本研究從地址擷取的角度出發做為商家辨識的標記，利用已抓取大量包含地址的網頁，先找出網頁中的地址，再藉由地址找出對應的商家名稱進行配對。換言之，給定一個已知地址，我們希望能透過網路資料擷取出該地點的名稱（如：商家名稱、政府單位…等）。舉例而言：當我們已有地址「新北市板橋區中山路二段 88 號 3F」，我們希望能知道這個地址對應的名稱「大鈞醫學美容診所」，如此即可進一步藉由地址、名稱並利用搜尋引擎收集更多額外商家資訊。這些額外資訊不僅可以有效提昇地圖上搜尋也就是地理檢索系統 GIS 的召回率，也可提昇商家分類的準確率(陳宜勤 等，2013)。

在辨識商家名稱的部分，本篇論文使用了條件隨機域 (Conditional Random Field)當作學習演算法。目前有許多關於中文組織名稱辨認的研究 (Zhang *et al*., 2007) (Yao, 2o11) (Ling *et al*., 2012) (Wu *et al*., 2008)，可以從新聞或一些較正式的文章中萃取出組織名稱，但是並沒有嘗試以一個 CRF-Model 直接對各種網站中的整個網頁內容進行中文組織名稱辨認。這兩者之間不同處在於新聞類文章屬於較正式的文章體裁，因此容易出現行政機關與正式的組織名稱，例如：行政院和維德食品有限公司，但是整個網路上商家組織名稱的命名方式傾向則不同，例如：吼牛排、努哇克咖啡、阿嬤祖傳菜包肉粽仙草…等，都是商家組織名稱。另外，一個完整的網頁內容有結構與非結構化的資訊交錯呈現，雖然結構化資訊會造成自然語言文字內容的破碎，但這些結構也隱含有可利用的資訊。

為了使商家辨識能以最少人力進行自動化學習，本研究使用自動標記方式建立訓練資料，我們先針對部份的黃頁網站(如 104 求職網、愛評網、工商名錄網站)撰寫 Parser 取得大量商家名稱與地址的組合，並以這些已經取得的商家名稱對網頁語料進行自動標記，再利用自動標記後的語料訓練 CRF 序列標記模型。然而一個地址可能出現在多個網頁之中，僅只仰賴其中一個網頁也有失之偏頗之慮，因此我們也收集了 Google Snippets 當作訓練資料進行比較。本篇論文的第二個主題則是商家地址的配對，由於一個網頁可能包含多個商家名稱，我們對網頁以簡單的規則進行分類後，使用了啟發式（heuristic）的配對規則，利用各類型的網站所具有的表達特性，對地址與商家名稱進行配對。

本研究承續 (Su, 2012) (Chuang *et al*., 2014)之研究，經由爬取網頁上包含地址的大量網頁（包括 Yellow Page 與 Surface Web）進行商家名稱擷取。其中 Yellow Page 提供了大量商家名稱以及地址與商家的配對資料，而 Surface Web 則利用 (Chang *et al.*, 2012)之地址擷取模型擷取出了可能含有台灣地址的網頁與地址清單。本篇論文以已知可能含有台灣地址的中文網頁、每筆網頁的地址清單、大量商家名稱清單以及已知的地址與商家名稱配對資料為基礎，提出了一個商家名稱擷取系統，方法分為三大步驟：地址網頁的前處理、商家名稱命名實體辨認、及地址－商家名稱匹配。本研究在三個模型聯合標記商家名稱的方式下，地址與商家名稱的平均配對正確率為 0.57。

本論文共有五個章節，第一節是緒論，說明研究動機與背景；第二節是相關研究，

介紹中文組織名稱辨認和地址相關資訊擷取的相關研究。第三節是方法，會詳細介紹如何對地址-網頁分類、中文組織名稱辨認以及地址與商家組織名稱的配對。第四節是我們針對現有的網頁中，依據我們的分類，每類隨機抽取網頁進行的實驗與結果分析。最後是我們的結論以及未來的展望。

2. 相關研究

擷取地址相關資訊牽涉到三個領域，資訊擷取（Information Extraction）、自然語言處理（Natural Language Processing）與資訊檢索（Information Retrieval）。這三者彼此間互相交錯，很難精確切割出各自所屬的範疇。大致上來說，資訊擷取主要是從各種結構化資料與非結構化文字萃取出特定資訊的方法，而自然語言處理則屬於人工智慧領域的一個分支，目的在於自動化的理解並處理人類所使用的語言。資訊檢索則是從大量資料中以機率統計模型對資料進行排序（rank）、建立索引，快速找出使用者目標文件的方法。

本研究相關的主要技術，分別為如何有效爬取包含地址之目標網頁、地址相關資訊擷取與命名實體辨認。地址相關資訊擷取是在得知地址資訊後，從含有地址的網頁中擷取出與該地址相關的資訊，如：電話、網址、電子郵件、評論…等資訊。命名實體辨認則是為了辨認文句所提到的特定種類概念，如：人名、地名、組織名稱。本章中將依序介紹這些技術的相關研究。

2.1 包含地址的網頁抓取與地理資訊檢索

這裡所謂的地理資訊檢索，是從網路上爬取包含地點或地址的網頁，萃取地理資訊並利用此資訊排序與建立索引，提供快速檢索的服務。目前的搜尋引擎，像是 Google 和 Yahoo 也分別從 2005 與 2002 年開始提供電子地圖的服務。而這些服務需要藉由使用者的標記等群眾外包的方式建立 POI 資訊。(Dirk & Susanne , 2007)等人提出了一個以位置資訊為基礎的搜尋引擎，可以自動從網路資源中取得與空間相關的文句，而在他們最近的研究中 (Ahlers, 2013a; 2013b)，則專注在如何從深度網頁例如黃頁與 Wikipedia 擷取出位置命名實體。由於地址是 POI 的明確指標，因此 Chuang 等人(Chuang *et al*., 2014)提出以廣度優先搜尋、黃頁爬蟲與地址樣版查詢三種策略爬取含有地址的網頁。實驗結果顯示雖然爬取黃頁網頁可以較快取得大量地址，然而地址樣版查詢可以補足黃頁涵蓋度不足之處，也是建立商家查詢服務不可或缺的方法。

2.2 地址與相關資訊擷取

地址擷取是因應地址資訊檢索所產生的需求，目的是從網路上大量的網頁中，擷取取出地址資訊，在 2009 年 Li 的研究中 (Li, 2009)，Li 以序列標記（Sequence Labeling）和 CRF 模型對美國地區的英文地址進行訓練與測試。Li 利用該地區地址的特性建立了 14 種特徵，並使用 BIEO 標記法，實驗結果 F-measure 達到了 0.913 的準確率。2011 年 Huang 延續了 Li 的研究 (Chang *et al*., 2012)，利用 17 種台灣地址特徵和 BIEO 及 IO 兩種標記法，其中 IO 標記法因為邊界偵測能力較弱，需搭配極大分數子序列（Maximal Scoring

Subsequence）進行修正。BIEO 標記法的實驗結果 F-measure 約在 0.96 至 0.99 之間，IO 標記法則在 0.94 至 0.96 之間。

相關資訊擷取是地址擷取的延伸研究，目的是針對已知的地址擷取出與該地址有關的訊息，如：電話、網址、電子郵件、評論…等資訊。主要的作法是針對已經成功擷取出的地址，找出可能的上下邊界、劃出資料範圍作為該地址的相關描述，可以視為一種深度網頁資料擷取（Deep Web Data Record Extraction）的一種特例。在 Li 的研究中，主要是把所有地址所在的文字葉節點（Text Leaf Node）當作起點，利用這些節點走訪至根節點過程中，Html Tag 的變化當作邊界點。但是 Li 的方法對於網頁中擁有兩種以上的地址相關資訊排版無法有效擷取，為了解決此問題，Huang 會先針對各地址路徑的相似度作出分類，再針對各類實行 Li 的方法。在最後英文地址相關資訊擷取的實驗中，Li 的相關資訊擷取的 F-measure 達到了 0.8689，而加入了 Huang 的改進則提昇 0.0233。

2012 年 Su (Su, 2012)發現 Li 與 Huang 的做法過度簡化各筆紀錄（Record）的產生模版（Template），Li 與 Huang 的做法中，只要模版中有任何一筆選擇性資料（Optional Data），就會發生連鎖錯誤。為了解決此問題，Su 將 2010 年 Wei Liu 所提出基於視覺（Vision-Based）的資料紀錄（Data Record）擷取演算法套用在地址相關資訊擷取的研究中，並重作 Li 的實驗，將 F-measure 由 0.7912 提昇至 0.9504。

Li、Huang 和 Su 的研究皆專注於資訊擷取的效果上，但其前提是網頁中存在多個地址字串。若提及地址相關資訊的網頁內不存在地址字串，則無法得知網頁內含與 POI 相關的資訊，更不可能有後續萃取資訊的過程。因此，本研究試圖擷取出地址的商家組織名稱，以利後續的相關資訊萃取與檢索。

2.3 中文組織命名實體辨認

命名實體辨認屬於資訊萃取與自然語言的一個共同分支，此研究起因於任何系統皆無法窮舉出所有的詞彙與代表的意義，因為再大的詞庫都會有沒收錄的詞彙（OOV word，Out-of-Vocabulary），且同樣的詞彙在不同的內容中很可能代表不同的意義。目前的主要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。

目前已經有許多中文組織名稱辨認的研究 (Zhang *et al*., 2007) (Yao, 2011) (Ling *et al*., 2012)，2007 年 Zhang 等人以人民日報 的新聞當作訓練資料，將數個 CRF 模型串連起來進行辨識，採用的特徵有：是否為前級輸出的各種命名實體（is Named-Entity）、常見的組織名稱開頭、內容與結尾、N 元文法（N-gram）。在 Zhang 所做的實驗中，F-measure 達到 0.9794。

2011 年 Yao (2011)則是將中文組織名稱分為三段：前置詞（Prefix words）+中間詞（middle words）+記號詞（mark words）（例如：中國+移動通訊+公司）且不採用現有的模型，使用自行設計的統計方法，考慮組織名稱的頻率、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」，最後的實驗使用了人

民網的語料進行訓練，以人民網、新華網 和北京郵電大學網站首頁的新聞 當作測試資料。平均準確率最高達到 0.959，平均召回值則達到 0.8724，皆超過隱藏馬可夫模型（HMM）與最大熵模型（ME）。

2012 年 Ling 等人 (Ling *et al*., 2012) 以規則式的辨認方法（Rule-based Named-Entity Recognition）辨識人民日報與新浪網的新聞，Ling 首先將語料經過斷詞並將中文組織名稱拆解為多個修飾詞（Modifiers）+核心特徵詞（Core Feature Word）。在統計訓練資料後，找出常用的核心特徵詞，建立核心特徵詞庫當作組織名稱的結尾，並找出 6 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。最後的實驗結果顯示，Ling 的方法的 F-measure 最高達到了 0.8573。

然而上述研究皆著重新聞語料之命名實體擷取，對於非新聞文件的一般網頁擷取並未著墨。事實上網頁的自由度使得命名實體擷取相對較為困難，這也是本篇論文的挑戰之處。

3. 商家名稱擷取與地址配對系統

本研究承續 (Su, 2012) (Chuang *et al*., 2014) 之研究以及 (Chang *et al*., 2012) 之地址擷取系統，經由爬取網頁上大量含有地址的網頁（包括 Yellow Page 與 Surface Web）進行商家名稱擷取。我們從這些網頁中過濾出含有台灣地址的可用網頁，進行商家名稱擷取，之後利用網站的特性如清單網頁、深度資訊網頁、註腳網頁、及自由文字網頁等為每一個地址配對商家名稱。

3.1 商家名稱辨認

本研究試圖對網頁內容擷取出所有的商家名稱，這裡所指的商家名稱涵蓋了各種範圍：明確的興趣點（POI，Point of Interest）、實際的組織名稱和產品的廠商名稱。目前在命名實體辨認的領域，通常使用序列標記法（Sequence Labeling）透過條件隨機域（CRF）模型進行辨認，然而監督式學習需仰賴大量的訓練資料，為減少人工標記的負荷，本文利用已知的商家名稱對網頁內容進行自動標記，並以標記後的網頁文字當作 CRF 的訓練資料。當 CRF 訓練完畢後，即可對網頁內容進行商家名稱辨識，建立商家名稱清單。下面將分別介紹本研究的自動標記、以及訓練資料的準備方式。

藉由 Web 上的黃頁網站所提供的商家資訊，我們可以取得「地址-商家名稱對」清單，對訓練網頁進行自動標記。然而由於網頁總數達 39.6 萬筆，而不重複的商家名稱總數高達 68.8 萬，基於執行時間的考量，無法對所有的網頁的每個句子都檢查是否存在已知商家名稱。因此，我們以每筆網頁已知的地址清單來加快標記速度：也就是說，系統只會依據網頁所擁有的地址查詢對應的商家名稱，並對網頁內容掃描這些對應的商家名稱是否存在，若存在就會以特殊的標籤（Tag）來標註這些商家名稱。圖 1 左圖即是自動標記自動產生訓練資料的流程圖。

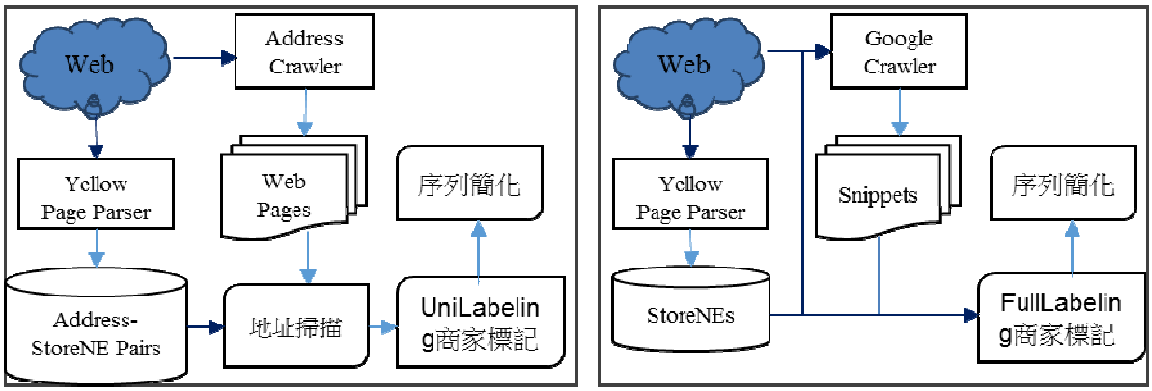


圖 1. 個別完整網頁的自動標記流程(左)與 Snippets 自動標記流程(右)

本研究另外以商家名稱當作關鍵字收集 Google 搜尋引擎提供的 20 筆 Snippets，並以所有的已知商家名稱對這些 Snippets 中的句子進行標記，試圖降低個別網頁資料的複雜度與標記不完整的問題。如圖一右圖所示，以 Google Snippets 為資料來源的處理流程，主要差異點在自動標記不僅只用單一的商家名稱來協助標記（稱之為 UniLabeling），而是採用所用商家名稱來進行標記（稱之為 FullLabeling），其餘皆與以整個網頁為資料來源的處理方式相同。訓練資料處理流程如下所述：

- **前處理**

針對每一個原始網頁後，本系統首先使用 Apache Tika™ (Apache License, 2004)將網頁內容連同標題擷取成文字內容後才進行後續步驟。為了使序列單元（Tokens）特徵的強度增強，系統會先將所有全形符號轉換成半形符號，圓弧型的括號「(、（、﴾、❨」統一轉成「(」，因為此種括號通常含有補充說明的意義。非圓弧型的括號「[、{、〔、｛、〈、…」統一轉成「[」，因為此種括號通常具有強調的意思。第二步是將換行符號、地址電話、時間…等以正規表示法取代成特殊的序列單元，這些取代動作能有效加強邊界特徵，縮短序列的長度，提昇辨識效果。

- **樣本序列**

完整的網頁內容與一般的文章相比，不同的地方在於網頁會利用結構化資訊、排版等表達方式將文字內容傳達給使用者，因此很少有完整的句子，而是直接把項目、名稱、屬性…等資訊以列表或依序列出等方式呈現。若我們採取傳統的句子樣本單元（Training or Testing Examples），進行訓練與測試，很難有好的成果。因此我們將網頁內容轉成文字後，移除空白類字元、以連續三個換行符號當作分隔符號（Delimiter），將文字切為許多區塊（Block），以含有商家名稱的區塊加上前後區塊，以連續三區塊為一個訓練樣本，這樣的好處是盡可能讓訓練樣本涵蓋商家名稱，也能有較多的非商家名稱範例。同樣地在測試時，也採用三行文字為一個單位當作樣本單元進行測試。

- **序列單元與標記**

一般說來，在人名辨識中，雖然人名有大量的組合與可能性，但是依然會有所謂的常用字，「菜市場名」就是一種很好的例子。但是商家組織名稱中除了結尾部份的常用詞外，在主要名稱上幾乎沒有任何規範，例如：「土地」、「阿嬤祖傳菜包肉粽仙草」中所有詞皆為常用詞彙，「18 度 c 巧克力工坊」、「591 租屋」為中英數字元交錯出現，「努哇克咖啡」、「蕾克爾烘培坊」為音譯詞，「蘆薈花園雲南食府」、「三峽歷史文物館」為地名。僅管如此，這些商家組織名稱的詞性卻有常見序列，如名詞+名詞或動詞、專有名詞+名詞或動詞、數字或英文+名詞或動詞 …等，所以詞性是一種不可忽略的重要特徵。因此在序列單元（Tokens）的選擇上，我們利用 Stanford Segmenter 及 POS Tagger 將網頁的文字內容經過斷詞及詞性（POS，Part of Speech）標記，以詞為單位進行訓練與測試。經過斷詞的序列，再以 B、I、E、O 四種標記代表商家名稱的起始、中間、結尾、以及非商家名稱。

- **特徵**

一般人在判斷一段文字是否是商家名稱時，會依靠兩類特徵，第一種是外部特徵（Outside Feature），這種特徵落在商家名稱的左右，但是此種特徵無法準確判斷商家名稱，只能進行推測上的輔助。第二種則是內部特徵（Inside Feature），內部特徵能提供強烈的判斷資訊，因為絕大多數的商家名稱都是由三個部份所組成：真名（Real Name）、產品或服務（Service or Product）、地標性詞彙（Landmark），舉例來說：「燦坤 3C 量販店」可以拆成「燦坤/3C/量販店」或「燦坤/3C 量販/店」。即使是非常短的商家名稱都會有這種結構，例如：「麗嬰房」是「麗/嬰/房」，其中嬰是指提供兒童用品。

我們統計已知的商家名稱，對組織、建築、房間、地標建立清單，例如：會、城、房、站…等，當每個序列單元（Tokens）是以此清單中的文字為結尾，就表示具有地標性詞彙的特徵（Landmark Feature）。另外，我們也收集了黃頁網站的服務、產品建立清單，如果序列單元（Tokens）含有此清單中的詞彙，就表示具有產品服務特徵（Service/Product Feature）。

當我們有了上述兩種特徵，問題就簡化成如何找出真名（Real Name）的部份，網頁內容與一般文章不同的地方在於名稱更傾向於單獨出現，而鮮少存在於一段完整的句子中 ，所以一段文字意思的起點就變成很重要的特徵：如果一個序列單元（Tokens）是樣本單元的起點或前一個序列單元屬於符號類，就具有開始特徵（Start Feature），反過來說，當序列單元是樣本單元的結尾或下一個序列單元屬於符號類，就具有結尾特徵（End Feature）。例如：網頁中的標語「[阿嬤祖傳菜包肉粽仙草]有阿嬤的精神傳承製作出客家傳統米食好滋味!」與網頁標題「阿嬤祖傳菜包肉粽仙草」中，前者的「阿嬤」的前一序列單元為符號，後者為序列單元的起點所以皆具有開始特徵，「仙草」則具有結尾特徵。系統最後選擇對商家名稱具有強烈判斷資訊的內部特徵加入訓練模型，所有原始特徵列於表 1。

表 1. 本研究所使用的原始特徵

NO.	Feature	Explanation
1	Token	個別詞 Individual Word, e.g. 591, 租屋
2	isPOS	詞性 Part of Speech, e.g. NR, NN, CD
3	isStart	樣本序列開頭 或 短語開頭
4	isSymbol	屬於符號詞, e.g. (, [, breakline, !, :
5	isService/Product	屬於服務/產品詞, e.g. 3C, 壽司, 出租, 通信
6	isLandmark	屬於地標詞, e.g. 廟, 莊, 公司, 店
7	isEnd	樣本序列結尾 或 短語結尾

3.2 地址-商家名稱匹配

當我們有了地址與商家名稱後,便可以開始進行配對。由於各類別的網頁特性差異很大，所以系統會針對各類別設計各自的啟發式（heuristic）的配對方式。首先我們依照網站將網頁分成不同群組，接著依網站中的地址資訊將網頁分成四類。

- 自然語言網頁：當地址字串所在的文字節點（Text Node）有超過 50 個字就會歸類至自然語言網頁（請參考圖 2），因為會這個長度相當於一小塊片段文字（Snippet）。位於此種網頁的商家名稱左右大多接有能意會到該處為商家名稱的訊息，例如：「走進 edia cafa 店裡一眼望去」、「我昨天去了燦坤 3C 買東西」。這也是網頁中唯一接近一般文章的類別。通常具有外部特徵 （Outside Feature）。
- 註腳資訊網頁：當一個「網站」內超過 80%的網頁都有相同的地址與文件物件樹路徑（DOM Tree Path），這些地址就會歸類至註腳資訊網頁。此類別中的所有網站，商家名稱周圍的文字資訊都有很高的相似度，經常會有：「本網站為…」「…版權所有」、®、©、地址、電話…，這些資訊在 N 元文法（N-Gram）的特徵上，能提供有用的資訊。
- 清單網頁：當一個網頁內包含超過 3 筆地址有相同的文件物件樹路徑（DOM Tree Path），這些地址就歸類為清單類型。清單型的商家名稱雖然不像自然語言網頁中，商家名稱的左右具有描述性的文字，但取而代之的是周圍具有換行符號、電話、地址、時間等資訊，藉由事先用正規表示法取代這些字串後，亦能利用 N-Gram 取得此特性。
- 深度資訊網頁(Detail Pages)：當一個網站內不同網頁的地址有相同的文件物件樹路徑（DOM Tree Path），但是地址字串卻不相同，這些地址就歸類為深度資訊網頁，當我們從多個網頁來看時，地址和商家名稱通常擁有同樣的文件物件樹路徑（DOM Tree Path），我們可以透過此特性進行商家名稱的修正。

圖 2. (a) 自然語言網頁範例 (b) 註腳網頁範例 (c) 清單網頁範例 (d)深度網頁範例

對於第一和第二類網頁而言，地址所對應的商家名稱通常落在：網頁標題、地址前、地址後或高頻商家名稱。若只有一個地址，則第一順位是網頁標題中的商家名稱。其次，以靠近地址的商家名稱為優先配對對象，「地址前」的配對方式是將地址與所在位置的前五行內的商家名稱列為配對候選者，而「地址後」則是將地址與所在位置的後兩行內的商家名稱列為配對候選者，當多個候選者距離相同時，會以網頁中出現較多次的商家名稱為優先，若次數完全相同則選擇位於地址前方的商家名稱。

至於第三和第四類網頁，因為網頁通常由模板（Template）和紀錄（Record）所組成，而相同類型的紀錄會放置在類似路徑下，所以存在一個專門的研究領域稱為 Wrapper Induction，目的是透過參考一個或多個網頁內容反向推導出模板與紀錄。本研究中使用了 TEX 作為輔助工具(Hassan & Sleiman, 2013)，TEX 是一個 Deep Web Crawling Tool，可以將多個網頁的原始檔文字內容當作輸入（作者稱為 TextSet），透過尋找各文件所擁有的共享樣式（Shared Pattern）當作紀錄的分隔點，經過反覆尋找共享樣式與切割後，找出最後的資料節點。藉由 TEX (Hassan & Sleiman, 2013) 擷取出網頁中具有同性質的資料節點，當有一定數量的同類節點被認為是商家名稱且商家名稱長度佔節點內容的 20%以上時，則把同類的非商家名稱節點也視為商家名稱進行配對。舉例而言，圖 2 中的「天天 100 剪髮」並沒有被 CRF 辨識出來，但是在同網站的其他網頁中，此節點的內容

「GM 造型館」、「肯特造型沙龍」…等已被成功辨識為商家名稱，所以系統也會將「天天 100 剪髮」視為商家名稱。本系統中，門檻值為 0.2，即該節點有 20%以上的內容被認為是商家名稱，則其餘網頁的該節點也會被認為是商家名稱。

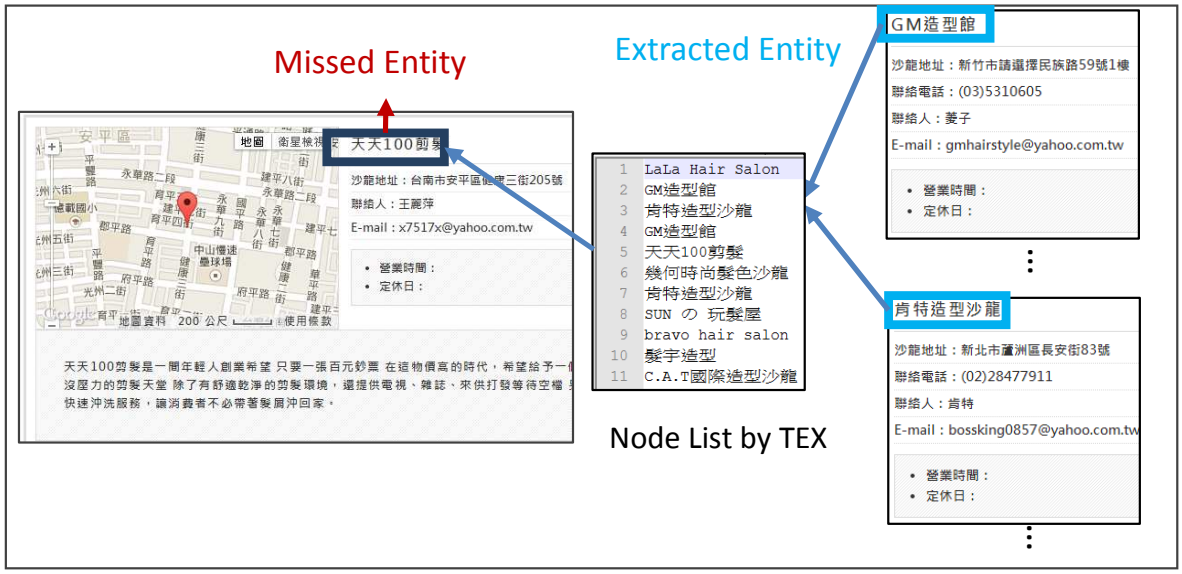


圖 3. 深度資料網頁配對範例

當我們利用路徑找出所有可能的商家名稱後，將開始進行實際配對。清單型網頁與深度資訊網頁的配對方式大致相同：以每筆地址的上方全部內容與下方兩行內當作配對候選，離地址近的優先配對，當距離相同時，以地址前方的商家名稱為優先。但清單型網頁會以地址為界線，在挑選配對候選者時，不會越過地址進行配對。

另外當我們以地址為關鍵字收集 Google Snippet 後，這些 Snippets 中的網頁結構資訊較弱，但是可以同時參考大量與近期相關的網頁提高可信度，所以當我們以商家名稱的 Snippet 訓練出 CRF 模型後，就直接以某地址為關鍵字所得到的所有 Snippets 中，出現最多次的商家名稱和該地址進行配對。

4. 實驗

因為本研究是先進行商家名稱辨認，再將地址與已知的商家名稱進行配對，所以實驗部份也依照這兩個階段來進行。第一階段的實驗為商家名稱辨識率，資料來源有兩種，第一種是以[9]所取得的約 50 萬個可能含有地址的網頁當作原始資料，在經過前處理後，過濾出約 39 萬個含有台灣地址的網頁，經過地址正規化後含 19 萬筆台灣地址（請參考表 2）。在經過網頁分類後，我們隨機挑選各類中的 100 個網站，每個網站中各隨機抽取 1 個網頁進行實驗，但 Detail Pages 因為配對方法需參考多個網頁，所以隨機挑選了 11 個網站，每個網站抽取 10 個網頁。最後對這 410 個網頁人工標記了 10,457 個商家名稱當作測試資料。而訓練資料則隨機挑選了 30,000 個訓練樣本，包含 51,775 個以自動標記法標記的商家名稱。

表 2. 以個別完整網頁為資料來源的訓練語料與測試資料

	Training Corpus		Testing Data				
	Raw	Preprocessing	FreeText	Foot	Detail	List	Sum
# Sites	-	13,224	100	100	11	100	311
# Web Pages	508,038	396,093	100	100	110	100	410
# Addresses	272,987	190,180	219	156	467	807	1,649
# Stores	-	-	1,841	1,975	3,855	2,786	10,457

第二種資料來源是使用 Google 搜尋引擎所取得的網頁內容片段（Snippets，請參考表 3），在訓練資料的部份，我們以 11,138 筆商家名稱進行查詢，以自動標記的方式產生了兩種訓練資料：SnippetUniLabeling 和 SnippetFullLabeling，在 SnippetsUniLabeling 中，我們僅以關鍵字的商家名稱對 Snippets 中的句子進行標記，共標記了 222,121 個商家名稱，而 SnippetsFullLabeling 中，則是以所有已知的商家名稱對 Snippets 中所有句子進行標記，共標記了 390,113 個商家名稱，藉由不同的標記方式產生不同程度的雜訊，以了解雜訊對辨識率的影響。在測試資料的部份則以 6,963 筆地址為關鍵字，收集每筆地址排名前 20 的搜尋結果（Snippets），以自動標記的答案進行最後 NER 效能評估。最後再對兩類資料進行交叉測試。

表 3. 以 Search Snippets 為資料來源的訓練資料與測試資料

	Training Data		Testing Data	
	# of Store Queries	Tag Stores	# of Address Queries	Stores (Auto Labeling)
Snippet Uni Labeling	11,138	222,121	6,963	70,449
Snippet Full Labeling	11,138	390,113	6,963	70,449

第二階段為地址與商家名稱配對的正確率，針對不同資料來源以各自的方式進行配對，第一種是針對不同網頁類別以各自的啟發式（heuristic）規則進行配對，第二種是以 Snippets 中各商家名稱的最高出現次數進行配對。

標記比對的評估方式如下：雖然我們有明確訂出商家組織名稱的判定規則，但很多時候依然難以準確定出邊界標準，例如：「飯店名稱：西門星辰大飯店」中，「西門」二字該不該列入商家名稱中有許多意見分歧的情況，由於商家名稱主要提供後續的地理資訊檢索，因此系統標記結果若能包含正確答案(Gold)，我們即認定正確，若是僅為正確答案的部份，則給 0~1 之間的分數，並依此分數計算 Precision、Recall、F-measure：

$$Gold, SysTag\text{比對分數} = \begin{cases} 1, & if\ SysTag\text{包含}Gold \\ \dfrac{TagNELength}{GoldNELength}, & if\ Gold\text{包含}SysTag \end{cases}$$

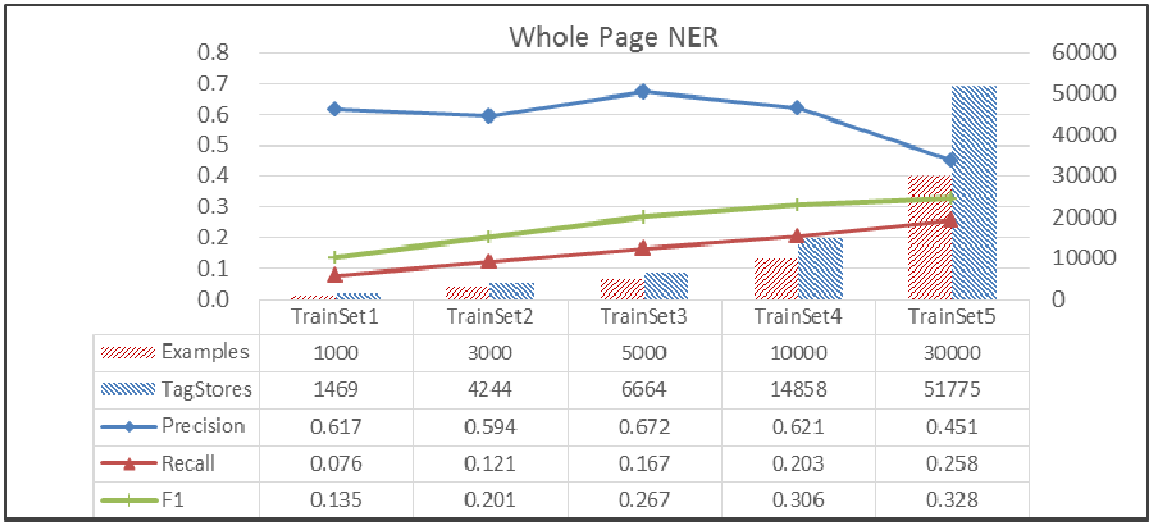
$$Precision = \frac{SysTag\text{辨識出的所有商家名稱與}Gold\text{進行比對的分數總和}}{SysTag\text{所辨識出的所有商家名稱數量}}$$

$$Recall = \frac{SysTag\text{辨識出的所有商家名稱與}Gold\text{進行比對的分數總和}}{\text{人工標記的所有商家名稱數量}}$$

此種評估方式，可以解決當 CRF 辨識出的商家名稱邊界包含地名、百年老店…等難以判定是否屬於商家名稱的一部分的問題。

4.1 商家名稱辨識率

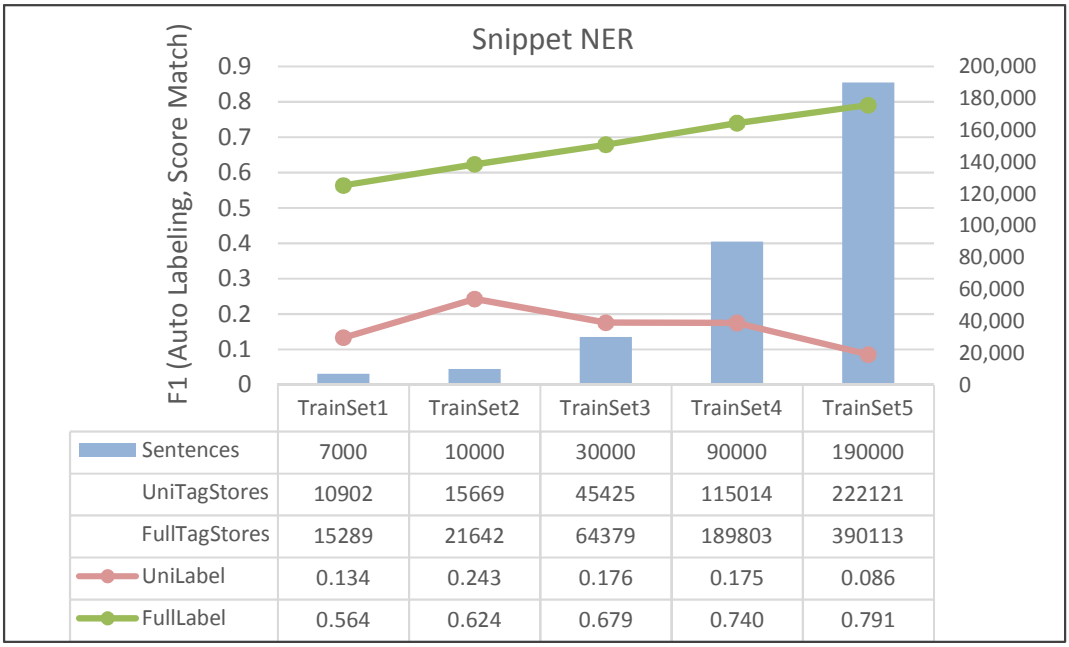
我們首先以個別完整網頁為資料來源，實驗了訓練資料數量對辨識效能的影響。接著我們以 Snippet 為資料來源，分別實驗了 Uni-Labeling 和 Full-Labeling 的效能，以了解在自動標記中，雜訊對辨識效能的影響，然後對兩種資料來源所訓練出的模型進行交叉測試，觀察不同來源的訓練資料所訓練出的模型，應用在不同測試資料時的表現。最後是本研究的在商家辨識部份的最後輸出。



	TrainSet1	TrainSet2	TrainSet3	TrainSet4	TrainSet5
Examples	1000	3000	5000	10000	30000
TagStores	1469	4244	6664	14858	51775
Precision	0.617	0.594	0.672	0.621	0.451
Recall	0.076	0.121	0.167	0.203	0.258
F1	0.135	0.201	0.267	0.306	0.328

圖 4. 完整網頁中，訓練資料數量對 F1 的影響

圖 4 是訓練資料數量對 Precision、Recall、F1 影響的趨勢圖，圖中顯示當訓練資料數量達到 30,000 樣本序列時，辨識效果依然只有 0.328，雖然 Recall 獲得提昇，但是 Precision 也較大幅的下降。主要的原因可能在於我們使用自動標記產生訓練資料時，並沒有使用所有已知的商家名稱進行標記，所以造成了大量標記錯誤（應為 B/I/E/S、卻標成 O）。因此在 Search Snippets 實驗中，我們嘗試探討降低語料複雜度與標記不完全兩個問題。



	TrainSet1	TrainSet2	TrainSet3	TrainSet4	TrainSet5
Sentences	7000	10000	30000	90000	190000
UniTagStores	10902	15669	45425	115014	222121
FullTagStores	15289	21642	64379	189803	390113
UniLabel	0.134	0.243	0.176	0.175	0.086
FullLabel	0.564	0.624	0.679	0.740	0.791

圖 5. 以 Snippets 為資料來源，雜訊與訓練資料數量對效能的影響

在 Snippets 方面的實驗，我們測試了訓練資料數量與標記品質對辨識效能的影響，以了解在自動標記中，雜訊對辨識效能的影響。如圖 5 所示，在 UniLabeling 模型中，當資料增加時，訓練資料含有的雜訊（標記不完全）更為嚴重，使得效能下降；而 FullLabeling 模型因為使用所有的商家名稱進行標記，所以雜訊大幅減少，在資料增加的情況下可大幅度提昇效能，FullLabeling 模型的效能最高為 0.791。

不過在 Search Snippets 的測試資料中並非使用人工標記的答案進行驗證，而是使用自動標記的答案。為了了解使用某一語料所訓練出的模型是否能應用在另一不同語料的測試資料，我們對個別網頁與 Search Snippets 進行了交叉測試，我們以完整網頁為訓練資料所訓練出的模型對 Snippet 的測試資料進行測試，同時也以 Snippet 中兩種訓練資料所訓練出的模型對 410 個網頁進行測試。

實驗結果如表 4 所示，圖中顯示不論是何種測試資料類型，由 SnippetFullLabeling 所訓練出的模型都具有比較好的辨識效果，甚至比個別完整網頁所訓練出的模型用在測試同類資料還要高，可見在自動標記中，只使用部份已知的商家名稱所產生的訓練資料，並不是一個好的方式，會大幅度受到雜訊與樣本數限制的影響。

綜合以上實驗結果來看，我們認為影響辨識效能的主要的原因有三個：第一是因為商家組織名稱屬於變異性較大的一種命名實體，在訓練階段中，資料的準備能否盡可能的涵蓋各類商家組織名稱的特性。第二，網頁屬於一種結構複雜的資料來源，以此種資料來源再以自動標記進行訓練，可能造成訓練樣本的品質不佳，因此對商家名稱這種變化性極大的命名實體，較難辨識出正確的答案，因此需要更多的特徵與提昇標記品質。

表 4. 交叉測試

	Whole Page	Search Snippets
Whole Page Model	0.305	0.473
Snippets Model (Full Labeling)	0.310	0.791

第三，當我們利用已知的商家名稱進行標記時，這些已知資料可能存在不正確、不齊全或是歧義性等問題，造成自動標記的第一次錯誤，而且擁有大量的已知名稱和網頁時，無法對所有網頁中的所有字串都檢查是否存在商家組織名稱，只能利用地址查詢是否存在對應的商家名稱，造成第二次的錯誤，若要在合理的執行時間內解決此問題，可能需要使用 Hadoop 或是其他分散式系統,以所有已知的商家名稱進行標記以提昇標記品質。

4.2 地址-商家名稱 匹配正確率

圖 6 是 SnippetFullLabeling 以不同訓練資料所訓練出的模型對配對正確率的影響，圖中顯示當 NER 的效能大幅提高時，Match 雖然跟著上升，但僅有微幅成長。而在完整網頁為資料的實驗中，雖然我們無法辨認出所有的商家名稱，但經由啟發式（heuristic）的配對規則，可以提昇配對的正確率。圖 7 是以完整網頁為資料來源，地址-商家名稱配對正確率的實驗結果。以單一類別來看，在深度資訊網頁的實驗中，利用文件物件樹路徑的相似度後，可以將配對準確率提昇至 0.951，平均正確率則為 0.573。

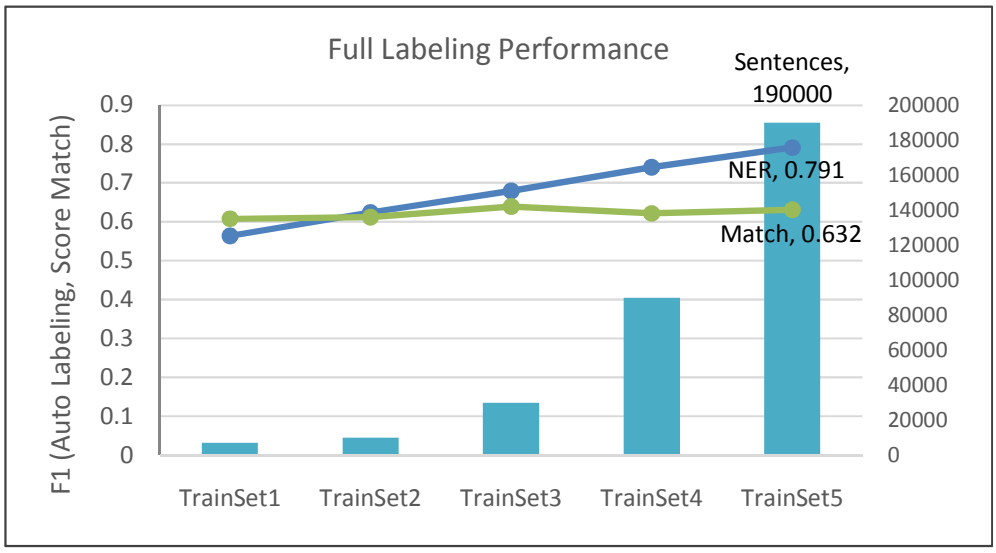


圖 6. SnippetFullLabeling 不同訓練資料數量的模型中，NER 對 Match 的影響

5. 結論

2014 是一個行動裝置的時代，大量的適地性服務（LBS）因此誕生，而 POI 資料庫在這波以行動裝置為主流的趨勢中具有至關重要的地位，建立一個完整的 POI 資料庫，可以讓使用者在地圖上提供更為便利的查詢。地址是 POI 的重要指標如果能找出地址所代表

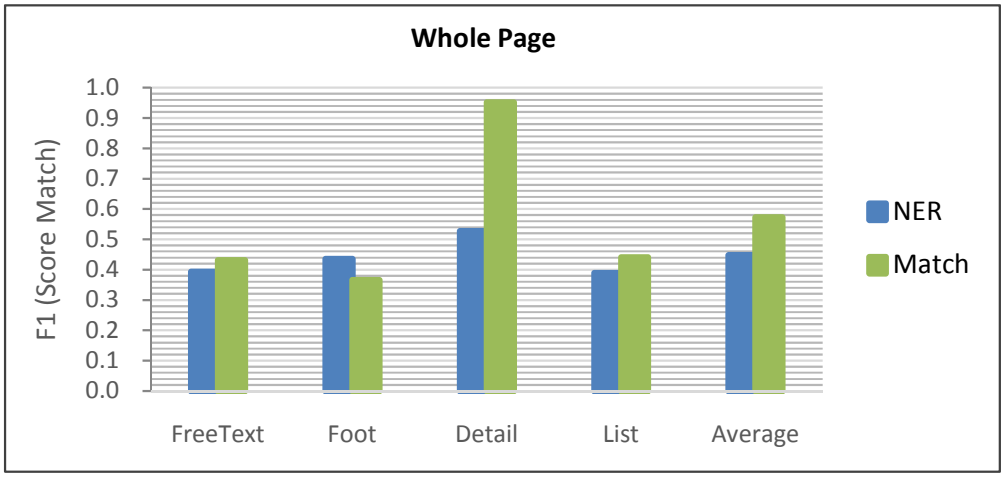


圖 7. 以完整網頁為資料來源的配對正確率（訓練樣本數：4,398）

的商家名稱，再以商家名稱當作搜尋引擎的關鍵字取得 POI 的相關資訊，就可以成功建立 POI 資料庫。

過去命名實體以新聞報導中的人名、地名、組織名擷取為主軸，目的在了解新聞中的事件，但對於網路上的興趣點 POI 的收集較少著墨。本研究試圖直接對整個網頁進行辨認，雖然受限於標記的不全，在命名實體辨認的效果並不好，但是在深度資訊網頁（也是含有最多地址的網頁類型）的地址-商家名稱配對中，利用網頁間的相似度可以取得 0.9514 的準確率，而平均正確率則為 0.5726。而 Google Snippets 的方法中，NER 效能最高為 0.791，配對正確率最高為 0.632。

在實驗過程中，我們發現啟發式的配對規則雖然可以提昇 Detail Pages 的配對正確率，但是其餘類型依然很仰賴命名實體的辨認結果。若要更進一步提昇商家名稱的辨識結果，我們覺得可以朝兩個方向進行，第一，必須將外部特徵加入特徵矩陣中，因為外部特徵雖然不能明確指出商家名稱，但是依然是進行推測的重要提示，在未來我們希望能把外部特徵和詞頻加入 CRF，提昇商家名稱的辨識效果。第三是利用分散式系統的速度，完整標記已知（大量）已知的商家名稱，解決自動標記產生的訓練資料品質不佳的問題。

參考文獻

Ahlers, D. (2013). Business entity retrieval and data provision for yellow pages by local search. *Integrating IR technologies for professional search, ECIR*, 2013.

Ahlers, D. (2013). Lo major de dos idiomas – cross-lingual linkage of geotagged Wikipedia articles. *Advances in Information Retrieval*, 2013, 668-671.

Apache Tika (2004). The Apache Software Foundation, [Online]. Available: http://tika.apache.org/.

Chang, C.-H., Huang, C.-Y., & Su, Y.-S. (2012). Chinese Postal Address and Associated Information Extraction. *The 26th Annual Conference of the Japanese Society for Artificial Intelligence*, 2012.

Chuang, H.-M., Chang, C.-H., & Kao, T.-Y. (2014). Effective Web Crawling for Chinese Addresses and Associated Information. in *EC-Web*, Munich, Germany, 2014.

Dirk, A., & Susanne, B. (2007). Location-based Web search. *Advanced Information and Knowledge Processing 2007*, 55-66.

GeoNames. [Online]. Available: http://www.geonames.org/.

Hassan, R. C., & Sleiman, A. (2013). TEX: An efficient and effective unsupervised Web information extractor. *Knowledge-Based Systems*, 2013, 109-123.

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

Li, S.-Y. (2009). *Application and Extraction of Postal Addresses and Related Information*, National Central University, 2009.

Ling, Y., Yang, J., & L. He. (2012). Chinese Organization Name Recognition Based on Multiple Features. *Pacific Asia conference on Intelligence and Security Informatics, 7299*, 136-144.

Liu, W., Meng, X., & Meng, W. (2010). ViDE: A Vision-Based Approach for Deep Web Data Extraction. *Transactions on Knowledge and Data Engineering*, *22*(3), 447-460.

The Stanford NLP (Natural Language Processing) Group. Stanford NLP Group, [Online]. Available: http://nlp.stanford.edu/software/segmenter.shtml.

Su, Y.-S. (2012). *Associated Information Extraction for Enabling Entity Search on Electronic Map*, National Central University, 2012.

Wu, C.-W., Tsai, R. T.-H., & Hsu, W.-L. (2008). Semi-joint labeling for Chinese named entity recognition. In Proceedings of the 4th Asia information retrieval conference, *4993*, 107-116.

X. Yao. (2011). A Method of Chinese Organization Named Entity Recognition Based on Statistical Word Frequency, Part of Speech and Length. *Broadband Network and Multimedia Technology (IC-BNMT)*, 637-641.

S. Zhang, & X. Wang. (2007). Automatic Recognition of Chinese Organization Name Based on Conditional Random Fields. *Natural Language Processing and Knowledge Engineering 2007*, 229-233.

教育部重編國語辭典修訂本－主站。中華民國教育部，[Online]. Available: http://dict.revised.moe.edu.tw/.

陳宜勤、賴郁婷、莊秀敏與張嘉惠(2013)。加入 Google Snippets 改善網頁商家多標籤分類, *The 18th Conference on Artificial Intelligence (TAAI 2013)*, 6-8.