The 2015 Conference on Computational Linguistics and Speech Processing
ROCLING 2015, pp. 148-163

基於已知名稱搜尋結果的網路實體辨識模型建立工具

A Tool for Web NER Model Generation Using Search Snippets of Known Entities

黃雅筠　Ya-Yun Huang
國立中央大學資訊工程學系
Department of Computer Science and Information Engineering
National Central University
a2425320032002@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

周建龍　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
formatc.chou@gmail.com

摘要

在過去，命名實體辨識（NER）研究都以新聞報導等正式文章中的人名、地名、組織名稱為主，相對地以網路的非正式文章則著墨較少。因此，現有的辨識模組對於網頁內容的辨識效果顯得較差，當需要辨識網頁內容中的命名實體時，勢必要重新訓練辨識模組。然而，訓練一個模型的時間和人力成本非常高，包含前置的大量訓練資料準備、人工收集及標記答案，且為了提升模組辨識效果，必須要為資料做適當切割、符號統一、正規化，以及特徵值的設計、準備已知關鍵詞庫（Dictionary）等，工作非常瑣碎複雜。此外，對於不同語言或不同辨識主題則需重複上述工作。本論文的目的，期能解決上述命名實體辨識工作過於費力耗時的問題，經由給定已知實體名稱的搜尋結果來自動標記訓練資料，並結合 Chou 及 Chang [2]於 2014 年在網頁中文人名的辨識研究之 Tri-training 半監督式訓練架構來產生 NER 模組。實驗證實，使用本工具可以套用在不同語言及類型的命名實體辨識，在中文組織名稱辨識的效能可達到 86.1%，在日文組織名稱辨識的效能可達到 80.3%，在英文組織名稱辨識的效能可達到 83.2%，辨識不同主題的中文地點名稱辨識效能可達到 84.5%，另外，辨識較長的命名實體如中文地址及英文地址辨識效能也可達到 97.2%及 94.8%。

Abstract

Named entity recognition (NER) is of vital importance in information extraction and natural language processing. Current NER models are trained mainly on journalistic documents such as news articles. Since they have not been trained to deal with informal documents, the performance drops on Web documents, which may lack sentence structure and contain colloquial expression. Therefore, the State-of-the-art NER systems do not work well on Web

documents. When users want to recognize named entity from Web documents, they certainly have to retrain the new model. Retraining a new model is labor intensive and time consuming. The preparatory work includes preparing a large set of training data, labeling named entity, selecting an appropriate segmentation, symbols unification, normalization, designing feature, preparing dictionary, and so on. Besides, users need to repeat the previous work for different languages or different recognition types. In this research, we propose a NER model generation tool for effective Web entity extraction. We propose a semi-supervised learning approach for NER model training via automatic labeling and tri-training, which makes use of unlabeled data and structured resources containing known named entities. Experiments confirmed that the use of this tool can be applied in different languages for various types of named entities. In the task of Chinese organization name extraction, the generated model can achieve 86.1% F1 score on the 38,692 sentences with 16,241 distinct names, while the performance for Japanese organization name, English organization name, Chinese location name extraction, Chinese address recognition and English address recognition can be reached 80.3%, 83.2%, 84.5%, 97.2% and 94.8% F1-measure, respectively.

關鍵詞：命名實體辨識，協同訓練，Tri-Training

Keywords: Named Entity Recognition, Co-Training, Tri-Training.

一、 緒論

命名實體辨識是自然語言處理的一項重要基礎工作，其辨識正確率對後續的語意分析（Semantic Analysis）、機器翻譯（Machine Translation）等自然語言處理議題具重大的影響。在大量文字資料中，常有人名、地名、組織名等有意義的專有名稱出現，然而因應社會需要及科技發展，這些不斷被創造的詞彙，難以被單一詞庫所收藏，因此需有命名實體辨識以便擴充詞庫。不同類型的命名實體出現於語句中的位置、規則或詞性皆不相同，因此需要的特徵值也都不同。以中文組織名稱辨識為例，目前許多關於組織名稱辨認的研究，主要是從新聞或一些較正式的文章中訓練組織名稱擷取模型[7] [11] [13]，但是網路上商家組織名稱傾向較不正式的命名方式，例如：彼得公雞地中海餐廳、造紙龍手創館等，而新聞等較正式的體裁則容易出現公司行號與正規的組織名稱，如：伊甸基金會、國立中央大學、高鐵公司等，且網路上發表於論壇或社群媒體的文章語句結構與用字遣詞皆與正式文章不同，因此辨識效果不佳。如表一以及表二所示，我們利用 2,000 筆已知地址為查詢關鍵字，於 Google 搜尋結果片段（Search Snippets）中包含關鍵字的句子為測試資料，再使用 Stanford NER[1] (Named Entity Recognizer) 來做組織名稱辨識實驗，F1 效果只能達到 54.3%。另外，我們也利用 200 筆中文地點名稱為查詢關鍵字，利用 Google search snippets 包含關鍵字的句子為測試資料，同樣利用 Stanford NER 來做地點名稱辨識實驗，F1 效果僅達 20.1%。顯示現有的公開 NER 工具對於 Web 上非正式文章的命名實體辨識效果有限，並導致後續的相關研究效能有限。

命名實體辨識可視為序列標記（Sequence Labeling）的問題，故通常使用 Conditional Random Field（CRF）來解決此問題，CRF 為一機率架構的無向圖（Undirected Graphical）模型，常用於標注序列資料。我們利用開放的 CRF++[3]程式進行實驗，為了使 CRF 標記能有好的準確率，我們必須處理原始大量文字資料，包含人工收集答案、標記答案等，同時為了提升模組辨識效果也必須要為資料做適當切割、選擇斷詞工具、統一符號、數

[1] http://nlp.stanford.edu/software/CRF-NER.shtml

值正規化，以及準備具有鑑別度的特徵值、或設計已知辭典等。若要辨識不同的語言或不同類型的命名實體，就要重複以上的動作來完成工作，造成了不少人力與時間的浪費，因此在本篇論文中我們將以上的動作模組化，並將其整合成一個命名實體辨識模型的產生工具。

表一、以 Snippets 為測試資料對 Stanford NER 測試效能

Testing Data		
	Chinese Organization Name	Chinese Location Name
# Queries	2,000	200
# Sentences	38,692	2,638
# Distinct Entities	16,241	600

表二、Stanford NER 對 Snippets 為資料來源之辨識效果

	Task	Precision	Recall	F-measure
Stanford NER	Chinese Organization Name	0.518	0.542	0.530
	Chinese Location Name	0.215	0.188	0.201

使用本工具可方便的訓練不同語言、類型的命名實體辨識模組，我們使用欲辨識的命名實體列表為本工具的輸入，於網路收集大量的 Google 搜尋結果片段，透過自動標記（Automatic Labeling）與特徵值的準備，產生訓練資料。為了減少命名實體標記不完整的問題，以中文組織為例，我們不只利用單一的組織名稱來協助標記（稱之為 UniLabeling），也採用所有已知的組織名稱來進行標記（稱之為 FullLabeling）。因自動標記可能造成訓練資料品質不佳，因此我們採用自我測試（Self-Testing）能進一步改善資料品質，再藉由半監督式學習（Semi-supervised learning）方法，引入 Tri-Training 增加訓練資料量，提升辨識模型之正確率。

實驗顯示系統在中文組織名稱辨識部份以 Tri-Training 演算法確實使得 F-Measure 更進一步提升至 86.1%，而在日文組織名稱、而在英文組織名稱、中文景點名稱也可達到 80.3%, 83.2%, 84.5%效能；另外在長命名實體中文地址以及英文地址的擷取上，F-Measure 辨識效果也分別達到 97.2% 及 94.8%。

二、 相關研究

命名實體辨認屬於資訊擷取與自然語言處理的一個共同分支，也是許多應用領域的重要基礎工具，自非結構化文字中識別具有特定意義的命名實體，如人名、地名、組織名稱，亦或命名實體相關屬性如電子郵件、地址及專有名詞等，目前有許多中文組織名稱及中文人名辨識的研究，利用序列標記配合機率統計模型是主要辨識方式。

● 辨識正式文章中文命名實體

Zhang 等人[13]於 2007 年將多個 CRF 模型串連起來進行組織名稱辨識，採用的特徵值包含是否為前級輸出的各種命名實體、常見的組織名稱開頭、內容與結尾、N-gram。並以中文人民日報新聞稿當作訓練資料，其最終的中文組織名稱辨識 Recall 可以達到 88.78%，Precision 可達到 82.35%。

2011 年 Yao[11]將中文組織名稱分為三個部份包含前置詞(Prefix words)、中間詞(Middle words)、記號詞（Mark words），舉例來說：「中國移動通訊公司」可以拆成「中國+移

動通訊+公司」，考慮中文組織名稱的出現頻率、詞性與長度，並配合自行設計的統計方法。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，其中文組織名稱辨識 Recall 可以達到 87.24%，Precision 可達到 95.9%。

2012 年 Ling 等人[7]將中文組織名稱語料斷詞後拆解為多個修飾詞（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，並以特徵判斷組織名稱的起點。取得候選者之後，利用規則式的辨認方法（Rule-based Named-Entity Recognition）進行修正。最後的實驗結果顯示， F-measure 最高可達到 85.7%。

- 辨識非正式文章中文命名實體

目前已經有許多如上述在正式文章中的中文組織名稱辨認（CONER，Chinese Organization Named Entity Recognition）研究[7][11][13]，但用這類訓練資料產生的模型在網頁及社群媒體短文等非正式文章中的辨識效果較不理想。為了解決這個問題，Lin 等人在 2014 年[6]，以中華黃頁網站取得的商家名稱對網頁語料進行自動標記（Automatic Labeling），再利用自動標記後的語料訓練 CRF 序列標記模型。在包含地址網頁以及 Google 搜尋引擎進行查詢所回傳的搜尋結果片段兩種資料中使用所有商家名稱來進行標記（稱之為 Full-Labeling）來建立測試資料。在包含地址網頁中文商家名稱辨識 F-measure 僅達 39.8%；而搜尋結果片段的 F-measure 可達為 79.1%。我們認為前者效能不佳的原因，可能在於不同網頁的文句的變異較大且切割的困難，此外 Lin 等人[6]採用斷詞分析語句，但經過斷詞後邊界錯誤的問題會較為嚴重。在 Google 搜尋為資料來源部分，Lin 等人[6]採用完整 Google 搜尋結果片段進行訓練，過長的結果片段會致使訓練時間拉長，也難有好的辨識效果。

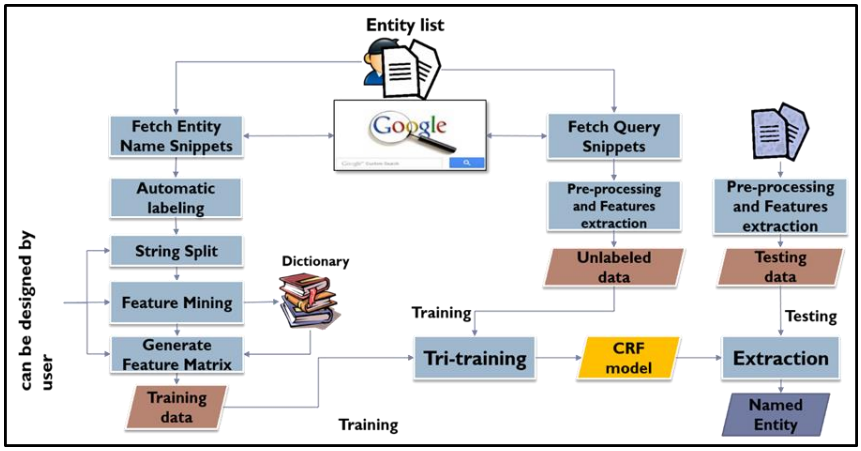
本篇論文延續 Chou 及 Chang [2]於 2014 在網頁中文人名的辨識研究中，為了解決訓練資料過少導致辨識效果不佳的問題，使用網路爬蟲於網際網路上自動收集大量包含中文姓名的資料，並自動標記已知的人名為答案，作為訓練資料之用。由於自動標記答案會有資料品質較差的固顧慮，為提升訓練資料的品質，Chou 等人也引入自我測試 Self-Testing 方法保留可信度較高的訓練資料；同時為了能利用未標記的資料，也改良 Zhou 等人於 2005 發表的原始半監督式 Tri-Training 演算法[12]，使得自未標記資料（U）中選取的新增訓練資料量足以對效能產生影響，以提升辨識正確率，最終 F-measure 可達到 91.3%。

三、 系統架構

我們設計的 Web NER 模組產生系統主要是接受使用者提供的命名實體列表，從 Google 搜尋結果片段中進行自動標記產生標記訓練資料，再藉由關鍵詞庫（Dictionary）建立、特徵值擷取（Feature Extraction），運用 CRF 建模。同時為彌補自動標記之不足，我們採用 Self-Testing 與半監督式 Tri-Training 訓練改善標記錯誤及標記不完全等問題，以達到辨識效能提升的目的。相較於收集訓練語句再由人工標記命名實體，由使用者提供大量命名實體範例再由系統從 Web 收集訓練語句並自動標記的成本相對少很多。系統架構如圖一，主要模組包括訓練資料收集模組、自動標記模組、特徵值擷取模組、Tri-Training 訓練模組及測試擷取模組，將於本章節中詳細描述。

3.1 資料收集與自動標記模組

為減少人工準備資料的負荷，本系統利用使用者輸入的命名實體列表作為 Google 搜尋的查詢詞，收集 Google 搜尋引擎回傳的前 N 筆搜尋結果片段。在本研究訓練資料準備部份使用 Google 搜尋引擎回傳的前 5 筆搜尋結果片段，而測試資料則收集 Google 搜尋引擎回傳的前 10 筆搜尋結果片段。



圖一、系統架構圖

- 自動標記命名實體

以往 CRF 序列標記模型的訓練資料皆為人工方式產生，雖然資料的品質可以信賴，但需花費大量的時間與人力。由於人工對搜尋結果片段進行答案標記成本過高，為此，本工具使用已知的命名實體作為答案，對搜尋結果片段內容進行自動標記，該標記即為欲擷取的目標，如此可以節省大量訓練資料標記成本。基於 Lin 等人[6]的研究顯示，使用單一的商家名稱來自動標記（稱之為 UniLabeling）的辨識效果較採用所用商家名稱來進行標記（稱之為 FullLabeling）要來的差，原因是在 UniLabeling 模型中，資料含有較多標記不完全的雜訊，使得效能下降；而 FullLabeling 模型使用所有的商家名稱進行標記，因此雜訊大幅減少。為減少雜訊影響，本系統採用 FullLabeling 的方式進行自動標記。

- 比對法標記長命名實體

自動標記的挑戰在於對於較長拼音文字的命名實體使用完全相配（Exact Match）並不能有效的標記。這是因為較長拼音文字的命名實體在不正式網頁文章中的書寫方式相較正式文章具有彈性，例如英文地址在正式書寫時會有固定格式、拼寫以及縮寫方式一致等規定。但我們利用“1131 Mountain Rd NW, Albuquerque, NM 87102”在 Google 搜尋時，雖然雙引號（“”）能限制搜尋結果都要有包含搜尋詞，但雙引號並不能保證搜尋結果片段內容中的命名實體與搜尋詞完全一致，搜尋結果片段中就算是在搜尋詞中穿插不同標點符號，或是沒有任何標點符號都會被搜尋出來。從圖二可以看到在 Google 搜尋前 10 筆結果片段就有 7 種不同寫法，而它們明顯都是表示此一地址。

```
1131 Mountain Rd NW , Albuquerque , NM 87102
1131 Mountain Rd NW Albuquerque NM 87102
1131 Mountain Rd NW - Albuquerque , NM 87102
1131 Mountain Rd Nw , Albuquerque , NM
1131 Mountain Rd NW Ste 2 , Albuquerque , NM 87102
1131 Mountain Rd. NW , Albuquerque NM 87102
1131 Mountain Rd NW Albuquerque , NM 87102
```

圖二、“1131 Mountain Rd NW, Albuquerque, NM 87102”在 Google 搜尋得到多種寫法

如使用完全相配方式來做自動標記，這些地址將沒辦法被標記出來。為了處理較長命名實體可能因標點及縮寫等問題無法被辨認出來的情形，我們使用排比（Alignment）的方式找出搜尋結果片段內容中可能的名命實體位置。在搜尋結果片段中找尋目標命名實體時，我們希望標記目標的命名實體在搜尋結果片段中與查詢詞相匹配的字越集中相鄰越好，因此我們設計了排比標記法（AlignmentLabeling）標記搜尋結果片段，再以排比搜尋結果片段以及搜尋詞所產生的相配 Match 及間隔 Gap 大小，做為我們判斷此一排比後的結果是否該標記的依據。如搜尋結果片段中與查詢詞經過排比後符合(1)相配字數大於命名實體長度 Len 減去間隔大小的一半，且(2)第一個排比配對到的字 $_1$到最後一個排比配對到的字 與命名實體查詢詞長度差距小於 3，則系統將會標記為出現範例。

$$(Mat\ h > \frac{Len - Ga}{2}) \text{ 且} (|(\quad - \quad _1) - \text{Len}| < 3)$$

然而排比標記法對於非拼音文字如中文應用的效果不如拼音文字。中文不同於英文，中文的縮寫是從長句子中取具有代表性的字出來，並且不會在單一命名實體中隨意加入標點符號。再者本系統在對 Google 搜尋時會使用雙引號，因此能確保搜尋結果片段中的長命名實體會與查詢詞完全一致，如此我們將可利用完全相配標記法（ExactMatchLabeling）正確且有效率的做自動標記。

3.2 字串切割與標記模組

在訓練資料的準備上，雖然可以採用完整 Google 搜尋結果片段做為樣本單元進行訓練，但過長的句子會致使訓練時間拉長，也難有好的辨識效果。但是搜尋結果片段中的網頁文章會有標點符號混用以及格式架構不嚴謹的問題，直接利用統一的切割方法將造成訓練樣本長度相差過大且品質不良。為準備適當長度的訓練句子，我們移除搜尋結果片段中的空白字元，利用自動標記的答案為基準取前後 W 字元為窗口大小，在我們實驗中，中文及日文設定 W 為 20，而英文則設 W 為 10，將文字切為許多區塊，以區塊為一個訓練樣本，最後去除重複的樣本，如此可使訓練樣本涵蓋命名實體，也能有適當的非命名實體範例。圖三為設定 W 為 20 的切割範例。

圖三、中文組織名稱辨識搜尋片段，取 N=20, 以詠展商行為基準

本系統採用不斷詞的中文字為基本處理單元 Token，避免樣本因為錯誤斷詞產生命名實體被分割成兩個詞的邊界錯誤的問題，減少錯誤累積。同時對於每一筆搜尋結果片段我們的系統會先將所有全形符號轉換成半形符號，如表三所示。

表三、全形符號轉換成半形符號範例

圓弧括號	非圓弧括號
(((==> (	[{ 「〔{〈『《【〔 { ==> [

答案標記方式我們選用 Start/End 標記法，此種標記法共有 5 個標記 B、I、E、S、O，依序表示命名實體的開始、中間、結束、單一序列單元以及非命名實體的序列單元，因為對開始和結束都給予不同的標記，可以提昇邊界的偵測效果。

3.3 特徵值擷取模組

特徵值的提取是訓練資料準備中非常重要的一步，常見的特徵是判定一個字是否為具有某種屬性，例如是否為數字或是百家姓等，因此準備相關詞庫是相當繁瑣的一環。一般說來，在判斷一段文字是否是特定命名實體時，會依靠兩類特徵，第一種是外部特徵（Outside Feature），這種特徵落在命名實體的左右，第二種則是命名實體的內部特徵（Inside Feature）。然而這些特徵往往必須要靠著熟悉語言或對該辨識領域了解的人來逐一產生，如我們要針對中文以外的語言進行辨識，關鍵詞庫就必須要由熟悉該國語言且有足夠背景知識的人員來準備。

為了使得本系統能夠避免這種語言能力及辨識主題上的限制達到通用的目的，我們的做法為統計字詞出現頻率，自動產生常見的關鍵詞庫。實務上，我們統計命名實體中的前一字、兩字及三字的頻率以及最後一字、兩字及三字的頻率，如表四中 ID 4~9。舉例而言，中文商家名稱最後一字常出現「廟」、「莊」、「店」等一字詞，或是「事務」、「數位」等兩字詞，又或是「基金會」、「雜貨店」等三字詞。我們也以命名實體出現在樣本中的位置為基準，統計出現在其前後方字、詞頻率，如表四中 ID 10~15 即為外部特徵值。我們利用自動選擇前 M 個常出現的字或詞來產生關鍵詞庫，在實驗章節將有針對關鍵詞庫大小對辨識效果的影響進行實驗。除上述 12 個自動產生之特徵值外，再加上針對辨識類別特別準備的特徵如縣市名稱及其簡稱、詞性（POS）tagging、是否為標點符號等特徵，此外因為在網頁中命名實體也常有單獨出現的情形，因此一段文字的起點就變成重要特徵，如果是樣本單元的起點或前一個字元屬於符號類，就具有開始特徵（Start Feature），當字元是樣本單元的結尾或下一個字元屬於符號類，就具有結尾特徵（End Feature），共 6 個預設特徵值。在不另外調整的情形本工具總共 18 個特徵值。

3.4 自我測試與協同訓練

我們使用開放且免費的 CRF++[3]程式做為序列標記模型訓練方法。由於本研究採用自動化的技術收集大量非結構化的資料以及自動標記產生的訓練資料，這些大量的訓練資料可能包含錯誤標記，為了提升訓練資料的品質，我們的工具設計在學習過程中可選擇使用 Self-testing 將雜訊移除提高訓練資料的品質。Self-testing 的實作方式是使用訓練完成的模型對訓練資料做測試並輸出機率，若該機率低於門檻值則認定該語句為雜訊，自訓練資料中移除，再以移除雜訊後的資料重新訓練模型，在本研究的實驗中設定機率為 0.7，其值可以視情況調整。

完成初步的資料品質提升後，工具會以 Self-testing 後之資料為基礎，加上 Chou 等人發表的 Tri-Training 演算法之改進[2]，應用未標記資料來改善效能。Tri-Training 的實作方

式是在學習過程中使用三個分類器，h_i、h_j 與 h_k $(i,j,k\in \quad ,1\leq \ , \ , \ \leq 3, i\neq \quad \neq \quad)$利用已標記的資料（L）訓練模型，並使用投票（Voting）想法挑選可信度較高的未標記資料（U）放到 L 集合中，稍後以新增資料後的 L 重新訓練分類器，疊代次數增加 L 集合所包含的訓練資料量亦隨之增加，使得分類器的辨識效能更進一步提升。

表四、自動產生的中文組織名稱辨識特徵值

ID	說明	長	範例
…	…	…	…
4	POI 中常見前方字	1	代、茶
5	POI 中常見前方詞	2	事務、數位
6	POI 中常見前方詞	3	多媒體、星巴克
7	POI 中常見倒數字	1	廟、莊、店
8	POI 中常見倒數詞	2	門市、公司
9	POI 中常見倒數詞	3	基金會、雜貨店
10	常見於 POI 前方的字	1	到、的
11	常見於 POI 前方的詞	2	推薦、加盟
12	常見於 POI 前方的詞	3	名稱：、店介紹
13	常見於 POI 後方的字	1	逛、是
14	常見於 POI 後方的詞	2	統編、營業
15	常見於 POI 後方的詞	3	高品質、營業項
…	…	…	…

在每一輪的疊代，Tri-Training 使用兩個模組 h_j 與 h_k 標記 U 中的資料，若兩模型答案一致，我們可以將此答案當作h 第 t 次疊代的新訓練資料 ，h 第 t 次疊代的訓練資料為 L ∪ 。若| |資料量過大，h 第 t 次疊代的錯誤率以 表示，前後次疊代間的錯誤率比例公式 | | < $^{-1}$ | $^{-1}$|將無法成立，此時則須對 做取樣動作，由

$$s= \left\lceil \frac{^{-1}| \ ^{-1}|}{} - 1 \right\rceil$$

公式計算可以自 隨機挑選 s 筆資料為新增的訓練資料，確保公式 | | < $^{-1}$ | $^{-1}$| 成立。Chou 等人[2]的改良演算法使得 Tri-Training 可適用於較大的資料集，避免原始 Tri-Training 在大量資料的情況下，僅可自 U 中選取少量資料作為新的訓練資料，對系統效能幾乎沒有影響的問題。

四、 實驗

本論文目的在完成一個不限語言、主題的 Web NER 模型自動產生工具，我們也將從實驗了解自動標記產生的訓練資基本效能（Basic）、透過 Self-Testing 資料過濾、以及 Tri-Training 等方法對於效能的影響。對於本系統所產生的特徵擷取方法，我們也將應用中文商家名稱辨識實驗比較人工準備關鍵詞庫及使用統計出現頻率的方式自動產生關鍵詞庫對於效能的影響。

由於在判定是否為正確答案時，有時會有難以準確定出邊界的可能，例如：「7-ELEVEN（行天門市）」中，「（行天門市）」可以視為包含在商家名稱之中，但若沒有標記出「（行天門市）」只有「7-ELEVEN」也不能算錯，因此對於每個辨識到的命名實體 e 與正確答案的命名實體 a，我們定義 P(e,a)、R(e,a)分數，再取平均值得到整體的 Precision、Recall。其定義如下：

- $(\ ,\ ) = \frac{|\ \cap\ |}{|\ |}$
- $(\ ,\ ) = \frac{|\ \cap\ |}{|\ |}$
- $Pr\ cision\ = \frac{\Sigma\ (\ ,\ )}{|\quad ntifie\quad ntitie\ |}$
- $ca\quad = \frac{\Sigma\ (\ ,\ )}{|\qquad ntitie\ |}$
- $-\ Measu\quad = \frac{2PR}{+}$

依照上述的評分公式，利用模型標記出來的答案（Identified entity）與正確答案（Real entity）間重疊的字數（Overlap tokens），分別除以標記答案長度和正確答案長度來給予部份正確的標記分數，此方法可以避免因為一兩個字的誤差而導致完全沒有分數的狀況。

4.1 實驗資料集

我們測試不同語言以及不同辨識主題的Web NER的辨識正確率，各個資料集如表五。

- 中文商家組織名稱辨識

我們透過中華黃頁[2]收集的11,138筆商家名稱，透過Google搜尋引擎進行查詢，取每筆搜尋前5個結果的搜尋結果片段，並以已知的商家名稱對搜尋結果片段中所有句子進行完全相配的FullLabeling標記產生已標記訓練資料（L）。未標記訓練資料（U）則使用50,000筆商家進行查詢，取每筆搜尋排名前10個結果的搜尋結果片段，共提取156,822個句子。測試資料則以另外2,000筆地址為關鍵字，收集排名前10個結果的搜尋結果片段，以人工的方式標記38,692個句子，標記出不重複的商家組織名稱共16,241個，最後使用此人工標記答案進行NER效能評估。

- 日文商家組織名稱辨識

我們透過iタウンページ[3]這個日本黃頁網站收集了10,000筆日文商家名稱，取每筆搜尋排名前5的搜尋結果片段，並對搜尋結果片段進行完全相配的FullLabeling標記產生訓練資料（L）。未標記資料（U）使用30,000筆商家名進行查詢，取每筆搜尋排名前10的搜尋結果片段，共提取88,074個句子。測試資料的部份則另外取200筆地址為關鍵字，收

[2] https://www.iyp.com.tw/
[3] http://itp.ne.jp/?rf=1

集每筆查詢排名前10的搜尋結果片段，以人工的方式標記測試資料共809個句子，共標記不重複的日文商家組織名稱438個。

- 英文商家組織名稱辨識

我們透過Yelp[4]收集的10,000筆商家名稱，透過Google搜尋引擎取得進行查詢，取每筆搜尋前5個結果的搜尋結果片段，並以已知的商家名稱對搜尋結果片段中所有句子進行完全相配的FullLabeling標記即為已標記訓練資料（L）。未標記訓練資料（U）則使用30,000筆商家進行查詢，取每筆搜尋排名前10個結果的搜尋結果片段，共提取100,182個句子。測試資料則以另外200筆地址為關鍵字，收集排名前10個結果的搜尋結果片段，以自動的方式標記941個句子，標記出不重複的商家組織名稱共465個，最後使用此自動標記答案進行NER效能評估。

- 中文地點名稱辨識

為了瞭解本工具辨識不同類別的能力，我們透過政府資料開放平台[5]收集了10,000筆臺灣地區地名資料，每筆取Google搜尋排名前5的搜尋結果片段，並以已知的地名對搜尋結果片段中所有句子進行完全相配的FullLabeling標記產生訓練資料（L）。未標記資料（U）使用30,000筆地名進行查詢，取每筆搜尋排名前10個結果的搜尋結果片段，供提取132,486個句子。測試資料另外取200筆地名為關鍵字，收集排名前10的搜尋結果片段，以人工的方式標記測試資料共2,638個句子，共標記不重複的臺灣地區地名600個。

- 中文地址辨識

為了瞭解長度較長的中文命名實體辨識效果，我們透過中華黃頁收集了1,800筆臺灣地址為搜尋關鍵字，每次取Google搜尋排名前5的搜尋結果片段，並以已知的地名對搜尋結果片段中所有句子進行完全相配的FullLabeling標記產生訓練資料（L）。未標記資料（U）使用10,000筆地址進行查詢，取每筆搜尋排名前10個結果的搜尋結果片段，供提取78,177個句子。測試資料另外取200筆中文商家組織名稱為關鍵字，收集排名前10的搜尋結果片段，以自動的方式標記測試資料共1,519個句子，共標記不重複的臺灣地區地址645個。

- 英文地址辨識

為了瞭解長度較長的英文命名實體辨識效果，我們透過Yelp收集了2,400筆美國地址為搜尋關鍵字，每次取Google搜尋排名前5的搜尋結果片段，並以已知的地名對搜尋結果片段中所有句子進行AlignmentLabeling比對並搭配UniLabeling產生訓練資料（L）。未標記資料（U）使用6,650筆地址進行查詢，取每筆搜尋排名前10個結果的搜尋結果片段，供提取49,851個句子。測試資料另外取200筆英文組織名稱為關鍵字，收集排名前10的搜尋結果片段，以自動的方式標記測試資料共652個句子，共標記不重複的臺灣地區地址257個。

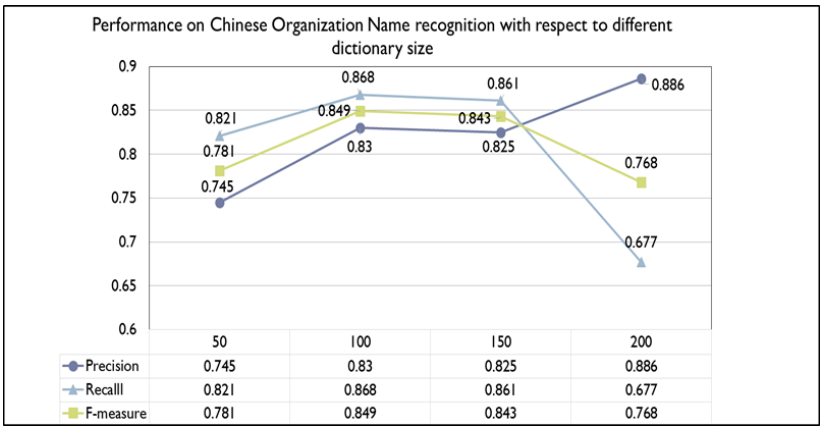
[4] http://www.yelp.com/
[5] http://data.gov.tw/?q=node/7063

表五、不同語言與不同辨識主題資料集

Item	Chinese Organization Name	Japanese Organization Name	English Organization Name	Chinese Location Name	Chinese address	English address
Source	中華黃頁	iタウンページ	Yelp	OpenData	中華黃頁	Yelp
Training: L	11,138	10,000	10,000	10,000	1,800	2,400
#Sentence	87,916	29,999	39,798	53,313	28,739	18,198
Training: U	50,000	30,000	30,000	30,000	10,000	6,650
#Sentence	156,822	88,074	100,182	132,486	78,177	49,851
Testing	2,000 addr	200 addr	200 addr	200 loc	200 organ	200 organ
#Sentence	38,692	809	941	2,638	1,519	652
#Distinct Entities	16,241	438	465	600	645	257

4.2 使用不同大小自動關鍵詞庫比較效能

以中文組織名稱為例，我們分別使用50、100、150、200個字或詞的自動產生內部特徵以及外部特徵建立關鍵詞庫，使用Self-testing將雜訊移除提高訓練資料的品質，實驗中假設低於0.7為雜訊將其去除，並以Self-testing後之資料為基礎進行Tri-Training演算法。各別關鍵詞庫大小之效能如圖四、使用不同大小自動關鍵詞庫比較效能，比較各資料集我們可以發現當使用大量字或詞的關鍵詞庫將導致Recall大幅降低。



圖四、使用不同大小自動關鍵詞庫比較效能

4.3 多種語言及辨識主題之 NER 效能

接下來的實驗中自動產生關鍵詞庫大小皆設為100，並以Self-testing低於0.7為雜訊去除後之資料為基礎，進行Tri-Training演算法。

- 短命名實體辨識效能

短命名實體辨識效能如表六，相較於中文及英文組織名稱辨識效能，日文的組織名稱辨識的F-measure稍低，我們猜測其原因可能在於日文屬於音節文字（Syllabary）是表音文字的一種，除了部分使用漢字外大部分使用平假名或片假名書寫，當在自動擷取外部與內部特徵時就會遇到僅取到部份拼音而不具有意義的問題。

我們也注意到在中文地點名稱辨識部分有很高的Precision，但Recall卻明顯較低。造成這個結果的原因是我們對於中文地點名稱有較廣泛的定義，例如：「高雄市」、「紫竹寺」、「平林里」、「狗母山」、「東石大橋」、「曹公圳」、「台北火車站」…等。因此我們在標記測試資料答案時的答案定以也較廣泛，但實際模組在標記時雖然能有高的準確率，但卻無法辨識所有類型的中文地點名稱。

表六、短命名實體之辨識效能採用自動產生之關鍵詞庫

	Chinese organization names	Japanese organization names	English organization names	Chinese location names
Precision	0.825	0.845	0.789	0.925
Recall	0.875	0.766	0.881	0.777
F-measure	0.849	0.803	0.832	0.845

- 長命名實體辨識效能

本系統對Google搜尋時雖使用雙引號，確保搜尋結果片段中的長命名實體會與查詢詞之字與字之間順序一致，但並不能保證搜尋結果片段內容中的命名實體與搜尋詞完全相同，因此搜尋結果片段中的搜尋詞中可能穿插不同標點符號。這對於使用ExactMatchLabeling標記出長的拼音文字命名實體是不容易的。因此為了標記出英文地址，我們使用AlignmentLabeling比對並搭配UniLabeling來標記查詢詞所在。但由中文地址在單一命名實體中不會隨意的插入標點符號與縮寫，因此我們可使用ExactMatchLabeling比對並搭配FullLabeling正確標記出中文地址。我們將會在後續實驗中比較AlignmentLabeling與ExactMatchLabeling之標記效果。對於長命名實體如英文地址及中文地址之辨識效能見表七。

表七、長命名實體之辨識效能採用自動產生之關鍵詞庫

	Chinese address	English address
Precision	0.997	0.938
Recall	0.948	0.958
F-measure	0.972	0.948

4.3 人工產生關鍵詞庫之 NER 效能

除自動產生關鍵詞庫之外，我們以中文組織辨識為例採用人工產生關鍵詞庫比較與系統自動產生詞庫效能的差異。特徵值包含人工收集的服務詞、產品詞以及地標詞詞庫，另外觀察常見於商家名稱前後的字詞產生更多詞庫。

表八顯示以Google搜尋結果片段辨識中文組織名為例，比較系統自動產生詞庫、人工產生關鍵詞庫與Stanford NER效能的差異。總體而言，雖然自動產生關鍵詞庫會導致Precision與F-measure降低，但卻能夠維持Recall水準甚至微幅提升。而辨識效果降低主要原因可能在於商家組織名稱屬於變異性較大的一種命名實體，資料能否盡可能的涵蓋各類商家組織名稱的特性是重要因素，而自動產生的關鍵詞庫相對於人工設計的關鍵詞庫包含較多的雜訊，且會有完全針對輸入的訓練資料設計等問題，但當訓練資料量夠大且商家類別多樣化時辨識應能再提升。

表八、以中文組織辨識為例比較系統自動產生詞庫、人工產生關鍵詞庫與 Stanford NER 效能的差異

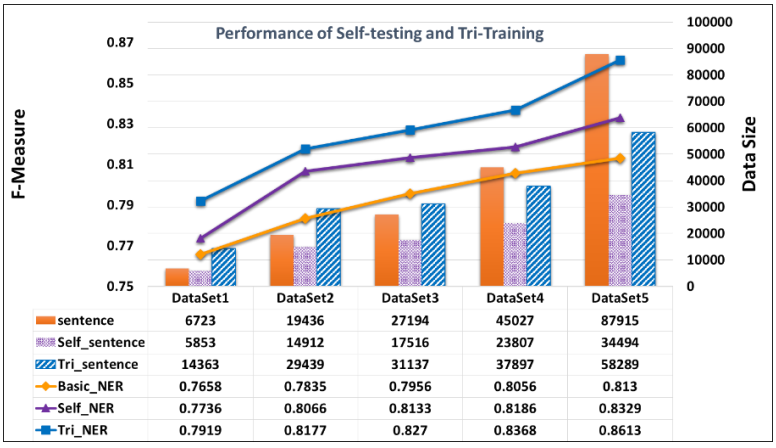
	Manual Dictionary	Automatic Dictionary	Stanford
Precision	0.8500	0.8249	0.529
Recall	0.8730	0.8753	0.557
F-measure	0.8613	0.8494	0.543

4.4 使用 Self-Testing 及 Tri-Training 後之 NER 效能提升

本實驗旨在了解使用Self-Testing以及Tri-Training產生之新辨識模型對Google搜尋結果片段NER效果的影響。我們以中文組織名稱辨識為例，將訓練資料分為五個資料集如表九。在中文組織名稱辨識人工產生關鍵詞庫的Self-Testing及Tri-Training實驗中，由圖五可以看到利用採用人工產生關鍵詞庫方式在Self-Testing以及Tri-Training的各個資料集大小辨識效果皆有提升，對於DS5提升幅度為4.83%，由0.8130達到0.8613。

表九、中文組織名稱辨識之已標記訓練資料（DS1~DS5）及未標記訓練資料（U）

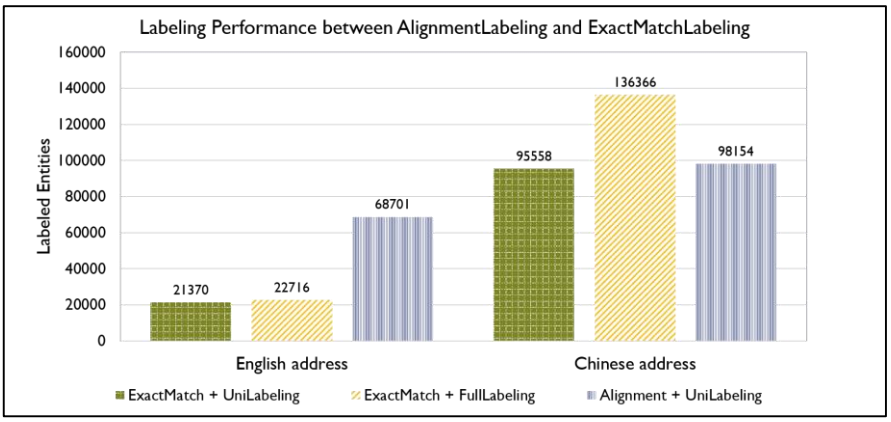
	DS1	DS2	DS3	DS4	DS5	Unlabeled
Query	1,000	3,000	4,000	6,000	11,138	50,000
Sentence	6,724	19,437	27,198	45,028	87,916	156,822



圖五、Basic、Self-Testing 及 Tri-Training 在中文組織辨識人工產生關鍵詞庫之效能

4.5 比較 ExactMatchLabeling 及 AlignmentLabeling 標記效果

圖六顯示英文及中文地址的標記效果。當使用ExactMatchLabeling比對搭配UniLabeling及FullLabeling產生訓練資料時，僅可從86,388筆搜尋結果片段中標記出21,370及22,313個英文地址。但當使用AlignmentLabeling比對搭配UniLabeling產生訓練資料時，共可標記出68,701個英文地址。另外，當使用ExactMatchLabeling比對搭配UniLabeling產生中文地址訓練資料時，已經可從108,435筆搜尋結果片段中標記出95,558個中文地址，若是採用FullLabeling，更可以標記出136,366個中文地址；因此使用AlignmentLabeling比對搭配UniLabeling標記出98,154個中文地址，未能勝過ExactMatchLabeling搭配FullLabeling的效果。



圖六、AlignmentLabeling 與 ExactMatchLabeling 之標記效能

不同於英文，通常中文並不會在單一命名時體中加入標點符號，因此我們可以利用ExactMatchLabeling標記出大量的長命名實體。除此之外，我們也發現使用AlignmentLabeling容易會在中文搜尋結果片段中標記出類似於中文地址的命名實體。例如，「彰化縣鹿港市場169號」並非是合法的台灣地址，但在AlignmentLabeling仍會被當作是目標給標記起來，此種錯誤會導致較低的準確率。

在表十中我們比較了 Alignment 搭配 UniLabeling 以及 Exact Match 搭配 FullLabeling 對於中文地址及英文地址的辨識影響。我們可以從表十中看出對於英文地址辨識使用 Alignment 搭配 UniLabeling 可以得到較好的 Recall 以及 F-measure(0.948)；然而在中文地址辨識，使用 Exact Match 搭配 FullLabeling 可以得到較好的 Recall 以及 F-measure(0.972)。

表十、長命名實體使用 Alignment + UniLabeling 及 ExactMatch + FullLabeling 之效能

	Alignment + UniLabeling		ExactMatch + FullLabeling	
Type	Chinese address	English address	Chinese address	English address
Labeled Entity	98,154	68,701	136,366	21,370
Precision	0.911	0.938	0.997	0.951
Recall	0.456	0.958	0.948	0.330
F-measure	0.607	0.948	0.972	0.490

五、 結論

訓練一個模型的時間和人力成本非常的高，包含前置的大量訓練資料準備、人工收集答案、標記答案，為了提升模組辨識效果而必須要為資料做適當優化，以及特徵值的設計、關鍵詞庫準備等，工作非常瑣碎複雜，且對於不同語言或不同辨識主題都要再重新設計特徵值。本研究期能設計一個使用Google搜尋結果片段之Web NER辨識模型的產生工具，不僅解決上述命名實體辨識過於耗時費力的問題，也能夠輕易地應用在不同的辨識類型、語言中，並希望達到良好的辨識效果。

在本系統我們使用自動標記的方式標記訓練資料而非使用人工標記答案，並且為了有效標記長的命名實體我們可以使用AlignmentLabeling增加標記到的命名實體數量。雖然自動標記可能包含雜訊，但我們因而能產生大量的已標記訓練資料。

外部特徵是進行命名實體辨認的重要輔助，而內部特徵能提供強烈的判斷資訊，我們利用頻率統計的方式能夠自動產生上述兩種特徵，並利用完整標記已知大量的命名實體與Self-Testing及Tri-Training演算法，使得辨識效能更進一步提升，解決訓練資料品質不佳的問題。

我們以中文之商家組織名稱辨識做測試，實驗顯示在中文組織名稱辨識部份以Tri-Training演算法確實使得辨識效能更進一步提升，F-Measure可由DS1的0.779提升至DS5的0.861，而在日文組織名稱、而在英文組織名稱、中文地點名稱、中文地址以及英文地址的F-Measure辨識效果依序可達80.3%, 83.2%, 84.5%, 97.2% 及 94.8%。

References

[1] D.-M. Bikel, S. Miller, R. Schwartz and R. Weischedel, "Nymble: a High-Performance Learning Name-finder", Applied natural language processing, pp. 194-201, 1997.

[2] C.-L. Chou, C.-H. Chang, S.-Y. Wu, " Semi-supervised Sequence Labeling for Named Entity Extraction based on Tri-Training: Case Study on Chinese Person Name Extraction," Semantic Web and Information Extraction, pp. 244-255, 2014.

[3] CRF++: Yet Another CRF toolkit, http://crfpp.googlecode.com/svn/trunk/doc/index.html 9-1541

[4] J. Lafferty, A. McCallum and F.C.N. Pereira, "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.

[5] C. Gu, X.-P. Tian, and J.-D Yu, "Automatic Recognition of Chinese Personal Name Using Conditional Random Fields and Knowledge Base," Mathematical Problems in Engineering, 2015.

[6] Y.-Y. Lin, C.-H. Chang, "Store Name Extraction and Name-Address Matching on the Web," Proceedings of the 26th Conference on Computational Linguistics and Speech Processing, pp. 91-93, 2014.

[7] Y. Ling, J. Yang and L. He, "Chinese Organization Name Recognition Based on Multiple Features," Pacific Asia conference on Intelligence and Security Informatics, pp. 136-144, 2012.

[8] W. Li, A. McCallum, "Semi-supervised sequence modeling with syntactic topic models,"

AAAI'05 Proceedings of the 20th national conference on Artificial intelligence - Volume 2, pp. 813-818, 2005.

[9] A. McCallum, W. Li, "Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons," Proceedings of the Seventh Conference on Natural Language Learning HLT-NAACL 2003 - Volume 4 (CONLL'03), pp. 188-191, 2003.

[10] 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.

[11] X. Yao, "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.

[12] Z.-H. Zhou, M. Li, "Tri-Training: Exploiting Unlabeled Data Using Three Classifiers", IEEE Transactions on Knowledge and Data Engineering archive, Volume 17 Issue 11, November 2005, Page 152.

[13] S. Zhang, S. Zhang and X. Wang, "Automatic Recognition of Chinese Organization Name Based on Conditional Random Fields," Natural Language Processing and Knowledge Engineering, pp. 229-233, 2007.