PTT 災害事件擷取系統

蔣佳峰
國立中央大學資訊工程學系
joe22485457@gmail.com

張嘉惠
國立中央大學資訊工程學系
chiahui@g.ncu.edu.tw

劉致灝
國家災害防救科技中心
liuchihhao@ncdr.nat.gov.tw

摘要

台灣屬於較常遭受天然災害侵襲的國家，災害發生期間，須仰賴災區民眾的主動回報，若救災單位接聽人手不足，或將成為迅速掌握災情的窒礙。另一方面，隨著網路通訊的蓬勃發展，災害發生當下，災情資訊也可能在社群網路間流動。因此，我們另闢一個獲取災情資訊的管道：從社群媒體中獲取災情資訊。

我們建立一個 PTT 災害事件擷取系統，使用批踢踢實業坊做為資訊來源，使用命名實體辨識（Named entity recognition）擷取文章中之「災害名稱」、「災害地點」及「災情敘述」等災情資訊，以建立災害事件報告。使用條件隨機域（Conditional Random Field）做為演算法，以建立上述三個辨識模型。根據實驗結果顯示：經人工標記後的測試資料比較，各模型在 Exact Match 與 Partial Match 之 F-Measure 皆高於 0.7 與 0.75。

關鍵詞:Web Crawler; Article Classification; Information Extraction; Name Entity Recognition;

1. 緒論

台灣較常遭受天災侵襲，如夏秋之際的颱風與不定時的地震，當災害發生之時，民眾目前多半透過電話或傳真的方式，將相關急迫性之災情傳遞給救災單位。此外，較次要但不容忽視的資訊，例如：路樹傾倒、招牌掉落等，救災單位較難在第一時間內充分掌握。另一方面，由於網路的普及化，民眾在社群網站上(如：PTT-批踢踢實業坊[4]、Dcard 等)進行社交溝通聯繫，災情內容亦可能受到討論。如能善用社群網路資源，我們可以此做為新的管道，從中擷取有用資訊。

基此，我們建立另一套獲取地震、颱風等天然災害資訊的系統，透過民眾於社群媒體所發表的輿論內容，從中篩選相關災情資訊，藉以建立災情事件報告。此主題分為三個工作項目：1.資料蒐集：使用爬蟲從社群網站上獲取文章。2.文章分類：區分災害相關的文章，以節省後續執行時間。3.資訊擷取：從文章中擷取將災情資訊。

文章分類為使用目前分類效果最好的 SVM（Support Vector Machine）[6]，選取常見的關鍵字做為特徵，建立災害相關類別文章的分類器。

災情資訊則使用 WIDM 實驗室所開發的 NER_Tool[2]，以 CRF 作為演算法訓練模型，根據訓練資料中，命名實體的前後文作為特徵，以此辨識並擷取文章中的災情資訊。

而災情資訊的詳細項目則參考 Wang [8]於 2012 年的中文新聞事件擷取系統，定義「災害名稱」、「發生地點」及「災情敘述」做為關鍵資訊：。範例由圖 1 所示：

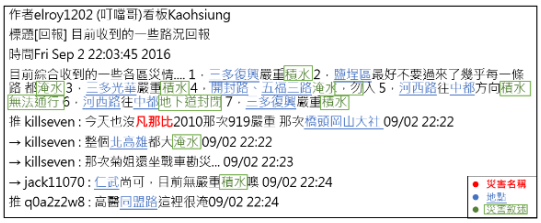

圖 1 PTT 災情回報示意圖

2.相關研究

2.1 災害事件

Takeshi Sakaki 等人[7]提出建立時間與空間的模型，偵測地震事件的發生時間與地點，對應於 Twitter 的文章發文時間點與發文者的 GPS 座標之間的關係。利用統計模型根據文章出現數量之變化，估算災害發生的時間點，並藉由地震等災害字詞獲取相關文章，建立以下三種特徵：統計特徵（文章字數、關鍵字出現於文章位置）、關鍵字特徵，以及查詢詞的前後詞特徵。依此做為分類器的特徵進行分類，並使用相關文章的 GPS 座標，以推估災害地點。

2.2 命名實體辨識

命名實體辨識（Named Entity Recognition）是資料擷取中的子項目，為從文章中找出人名、地點、組織、時間等命名實體的位置。NER 的實作，可區分為語言上的文法結構及機率統計模型的方式（例：機器學習）。使用文法基底的系統進行辨識，可得到較高的準確率，但召回率亦會相對性的變差。原因是經文法剖析過的句子，如果該句子結構完整即能被辨識出來，若為口語上或慣用句造成文法不合者，則無法被辨識。

NER 目前主要透過統計型的序列標記（Sequence Labeling）來達成。模型基於訓練資料原文、人工的標記結果，配合訓練句子本身的句型建立，常見的作法有隱馬爾可夫型（Hidden Markov Model, HMM）[1]及條件隨機域（Conditional Random Field）。本論文使用的 CRF，是一種模式識別及機器學習的建模方法，由 John Lafferty 等人[5]提出，用於分析序列資料，如自然語言或生物序列。CRF 是一種無向性的機率圖形模型，針對給定的句子，考慮相鄰的字與字之間，其標記結果是否有關係性，從中找出最佳的標記組合，透過訓練資料轉換成的編碼，觀察出已知上下文的關係，並建立出一致的解釋。

3. 系統架構與方法

整體系統大致由以下三種模組構成：（1）文章抓取模組，經由網路爬蟲對 PTT 網頁版之八卦板與其他看板，進行文章抓取。（2）文章分類模組，找出相關災害文章，以減少下一階段的資料處理。（3）災害事件擷取模組，將相關災害文章進行命名實體辨識，以獲得文章中提及的災害名稱、災害地點、災情敘述等命名實體，做為事件之重要資訊。系統架構如下所示：

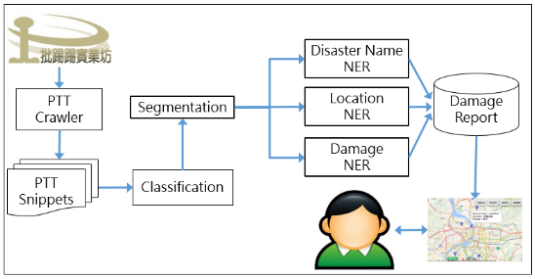

圖 2 系統架構圖

3.1 災害名稱的資料處理

災害名稱的命名實體集（Seedlist）之建立，乃是根據維基百科的颱風名稱表，及「台灣歷年重大災害事件」列表做為關鍵詞（例：九二一大地震、八七水災、台南地震），以獲得訓練文本。我們主要蒐集災情嚴重或範圍較廣的天災，例如地震及颱風。

災害名稱的命名與標記原則，參考自科技部國家災害防救科技中心所提供的天然災害事件名稱規則，以及維基百科的世界氣象組織命名表。

3.2 災害地點的資料處理

一般提到的地點大致分為兩類：具有街道路名的地點敘述或廣為人知的興趣點（Point Of Interest, POI）。本研究將這些敘述方式，依照敘述地點的範圍分為四類（如表 4）：（1）國家、島嶼名稱，例如菲律賓、東沙島、太平島等。（2）河川、海洋等水文敘述，例如曾文溪、秀姑巒溪、巴士海峽等。（3）行政區域及其詳細敘述，例如嘉義縣梅山鄉。（4）POI：大眾感興趣的知名景點、建築物、地標等，例如陽明山、中正紀念堂、桃園機場等。

災害地點的命名實體集（Seedlist）之資料來源，取自於中華郵政系統的街道路名、行政區名列表、維基百科的各國國名列表、台灣河流名稱列表，以及交通部觀光局的台灣旅遊景點列表，合計 29,091 筆命名實體內容，自資料庫獲得的文章總數 85,730 筆。

3.3 災情敘述的資料處理

災情敘述的命名實體之建立，透過 EMIC[3]提供近兩年，桃竹苗地區的民眾災情回報內容，以這些內容節選關鍵字獲取 2,303 篇文章。此階段的重點，在於如何正確地標記訓練資料。

因為災情敘述在書寫上自由度高，在標記訓練資料時，若採用自動標記，即以 Seedlist 內容做替代式標記時，文章中的災情敘述可能無法全數被標記。僅只是一字之差，仍無法被正確地標記。目前應對的辦法為，使用容許部分內容相異的對齊標記（Alignment Labeling）。

根據文章內容中的災情敘述，與 Seedlist 中命名實體的最小編輯距離，若計算結果低於門檻值時，將視為可以被標記的內容。透過此方式，可以減少相近的災情敘述`卻無法正確標記的問題。

3.2 文章分類

我們使用颱風、地震、土石流、水災、損失、豪大雨、淹水及火災等八個關鍵字，抓取合計 1,620 篇的 PTT 文章，經人工驗證後屬災害相關者為 931 篇，而非災害者則為 689 篇。

文章經過結巴斷詞（jieba）並去除低文件頻率（$df<5$）的詞類後，使用 SVM（Support Vector Machine）建立文章分類器，針對特徵的選取，使用以下三種方式實作：

1. 勝算比（Odds Ratio）
2. 卡方檢定（Chi-Square）
3. 資訊獲利（Information Gain）

3.3 災害事件擷取

我們使用中央大學 WIDM 實驗室所提出的 NER 模型建立工具，準備建立災害名稱、災害地點及災情敘述所需之訓練資料，利用關鍵詞查詢資料庫，以尋找相關的 PTT 文章內容，並透過自動標記獲得訓練文本，執行流程如圖 3 所示。

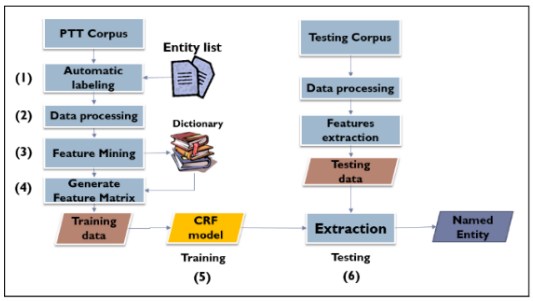

圖 3 NER_Tool 執行流程圖

（1）命名實體標記：使用 Seedlist 的內容，將訓練資料進行替代式標記。（2）文章前處理：文章中的符號正規化及句子切割。（3）字典建立：觀察訓練資料中，實體的內容及其上下文，建立數個字典做為特徵（4）特徵矩陣建立：使用上一步驟的字典，對於句子中的每一個字，我們依照各字典代表的性質，給與各式特徵。根據標記結果於矩陣最後一行，使用 B、I、E 表實體的開頭、中間、及結尾、S 為長度為一的實體、O 則為不相關。（5）模型訓練：使用 CRF++訓練辨識模型。（6）文章測試：測試資料於步驟 4 為起始，測試結果之最後一行即為模型判讀結果。

4.實驗與系統效能

本實驗共分為兩部分，包含文章分類及各辨識模型的相關實驗。

在文章分類的實驗中，我們準備人工確認的文章共計1,620 篇，其中災害相關文章 931 篇，非災害相關文章 689 篇。由於文本數較少，實驗結果皆經過五次交叉驗證的方式，以計算平均準確度（Average Accuracy）。

前述之三個模型皆以同一組文本測試，使用颱風、地震、土石流、水災、損失、豪大雨、淹水及火災等八個關鍵字，擷取 PTT 文章並人工標記，獲得 18,266 行測試資料。由於災害名稱與災情敘述以各自的 Seedlist 中命名實體標記結果，即字典標記（Dictionary Match）做為 baseline，而災害地點與 CKIP 的詞性標記結果(Nc Tag)、Stanford NER(LOC、GPE Tag)及字典標記做為比較。

4.1 文章分類的效能評估

我們根據勝算比、卡方檢定與資訊獲利的特徵選取方法，分別建立分類模型，由於測試資料較為少量，此部分皆使用五次交叉驗證進行，圖 4 為各方法的平均準確率（Average Accuracy）：

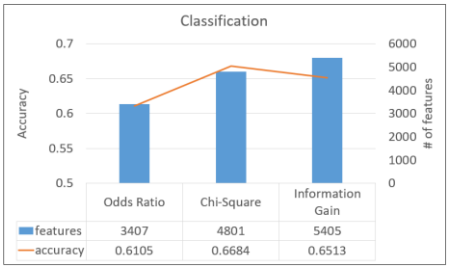

圖 4 三種特徵選取方法製作成 SVM 分類器的分類結果

由上圖可以得知，目前分類效果最高達到準確度=0.68，目前在此任務上達到的結果不到非常理想，問題主要為災情相關的文章，其真正討論到災情的內容，屬散佈於回文者的留言中，較明顯的例子為，即使文章內容同為轉載自中央氣象局的陸上颱風警報，若其中一篇的留言中出現災情敘述，在人工標記上會被認為一篇非災情相關，而另一篇則為是。

4.2 命名實體的效能評估

實體的算分公式，考量到部分標記資料的情形，以災害名稱為例：「九二一集集大地震」如果只標記出「集集大地震」，我們應該針對這個標記結果，給予部分分數。我們參考 Huang[8]的評估方法，對於每個辨識到的命名實體 e 與正確答案的命名實體 a，根據以下公式取得 F1-Measure。

$$P(e,a)=\frac{|e\cap a|}{|e|} \qquad R(e,a)=\frac{|e\cap a|}{|a|}$$

$$Precision=\frac{\sum P(e,a)}{|Identified\ \ entities|}$$

$$Recall=\frac{\sum R(e,a)}{|Real\ \ entities|}$$

$$F1-Score=\frac{2PR}{P+R}$$

4.3 災害名稱模型的效能評估

由圖 5 顯示，採用字典標記，相對於名稱模型擷取之結果較差。字典詞彙量的不足或新出現的災害名稱，將可能降低效能。相對的，有機會從上下文推論災害名稱的模型，辨識效果較佳，若測試資料中，未收錄於 Seedlist 中的災害名稱的數量增加時，其差異會更加明顯。

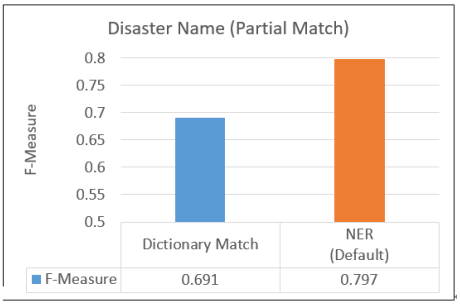

圖 5 災害名稱模型與其字典標記的效能比較

以下為標記錯誤範例（底線為人工標記答案，**粗體字**為系統答案）：(1)漏判：該命名實體未收錄至 Seedlist 中，且前後文內容不充分。例：「238.6mm 1966.8.16 (蒂斯颱風通過東海南部)」。

(2)誤判：敘述近似災害名稱。例：「推 Jasy:台東**蘭嶼地震**站南方 234.2 公里。」。因民眾討論之部分內容與災害名稱近似，目前主要依靠大量的負面範例(Negative examples)降低誤判率。

4.4 災害地點模型的效能評估

我們自郵政系統、維基百科與觀光局網站，獲取街道路名、國家名及台灣著名景點，合計 29,091 筆建立 Seedlist，並利用各命名實體做為關鍵字，獲取對應文章，目前蒐集 85,730 筆做為訓練資料。經由基本特徵建立的模型，與以下既有模型或工具比較其效能，詳如圖 6 所示：

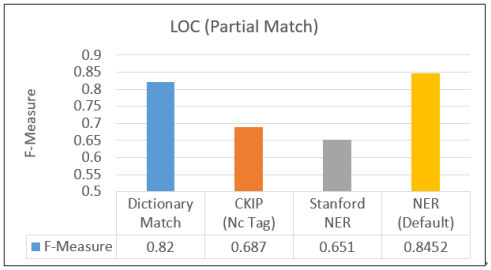

圖 6 災害地點與其他模型/方法的效能比較

CKIP 與 Stanford NER 主要因為大量的誤判，導致測試結果較差，常見的字眼如「家」、「國中」、「學校」等，在文義上雖有地點的特性而被標記，但不符災害名稱的定義而被視為誤判。

地點模型的判斷失誤，多半為未見過之 POI 或地點，且其上下文不足以判斷所致。或者標記答案近似地址敘述，因收錄的地點內容與地址相較，未能達到詳細程度，而無法完整標記。因本實驗較著重於辨識地點而非地址，故地址的標記情形，並非我們主要的探討對象。

4.5 災情敘述模型的效能評估

實驗為字典標記與對齊標記的效能比較，以及災情敘述模型與字典標記的效能比較。前者為比較字典標記與對齊標記所標記的訓練資料，探討不同標記方法的所得到的命名實體數量，所反映出測試效能的關係，後者為災情敘述模型與字典標記比較。

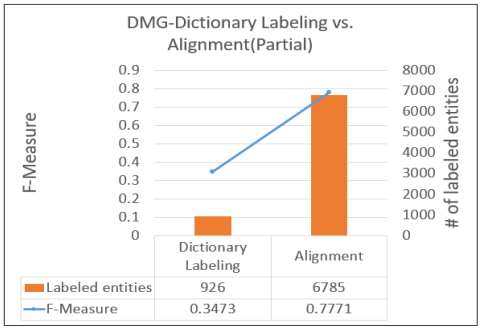

圖 7 字典標記與對齊標記的命名實體標記數量及其模型效能的比較

由圖 7 可知，使用同一組的 Seedlist 進行標記，對齊標記所能標記的命名實體數量遠超過字典標記，所反映出的效能差距也十分顯著。因字典標記的標記彈性較低，訓練資料中該被標記的命名實體，若與 Seedlist 收錄的命名實體僅有一字之差時，則無法成功標記，此標記缺陷因災情敘述的自由性而特別顯著。相對的，對齊標記可以減少漏標情形，進而提升訓練資料的完整度與測試效能。

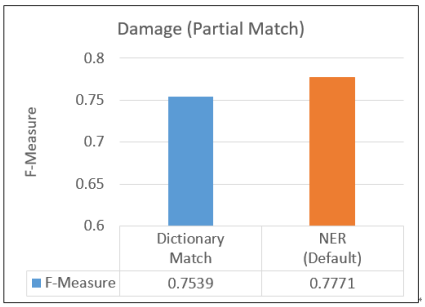

圖 8 災情敘述模型與字典標記的效能比較

圖 8 可以觀察出兩種方法得到之結果相當接近，這必須考慮到民眾在敘述災情上的手法：對於「淹水」這項災情，會額外說明淹到幾公分或淹到膝蓋等程度，相較於提到出「淹水」兩字的民眾來的少，程度詞的出現與否會影響到字典標記的結果，越是簡要的災情敘述，字典標記越有機會成功標記。另一方面，冗長的災情敘述對模型來說也提高誤判的可能，進而可能導致效能並未顯著提升，由此兩種情形有助於解釋到兩者效能接近的緣故了。

以下為標記錯誤的範例（底線為標記答案，**粗體字**為系統答案）：「將近兩千五百萬人流離失所。」、「中午彰化市區**淹水**最低處竟然高達 3 公尺如此的驚人！」。

在經過觀察測試後，推斷造成效能低落的原因，主要為訓練資料仍無法被標記完整。造成此結果的原因有二，其一為使用中文敘述災情時過於自由，即使透過對齊標記，若 Seedlist 未收錄相近的命名實體，也無法成功標記。其次為對齊標記自身的誤差，可能導致漏標情形。

5.結論

鑒於台灣較易遭受天然災害的侵襲，相關災情的蒐集非常重要，所幸現在的網路便利性高，災情資訊也很有機會從社群網站上獲得。為此，我們建立一個PTT災害事件擷取系統，可擷選出民眾發表的災情資訊。系統流程透過網路爬蟲，獲取大量文章，並經過分類獲取災害相關文章，重要的災情資訊則由事先訓練好的NER模型判讀。

目前分類用的特徵選取，採自動化的方法，雖減少人力成本，但由於文章內常含有許多雜訊，目前的分類效果不盡理想。未來將進一步以句子為單位進行分類，以求更好的分類結果。

災害名稱、災害地點及災情敘述的模型效能，在現階段仍有很大的進步空間，除了提升Seedlist收錄的實體數量，針對文章內容設計新的特徵也是未來的研究方向。而災情敘述所使用的訓練資料，特別需要更大的Seedlist做支援，以增加訓練資料的質量。

6.參考資料

[1] Blunsom, Phil. "Hidden markov models." *Lecture notes, August* 15 (2004): 18-19.

[2] Chou, Chien-Lung, Chia-Hui Chang, and Ya-Yun Huang. "Boosted Web Named Entity Recognition via Tri-Training." *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 16.2 (2016): 10.

[3] http://portal.emic.gov.tw/nfasso/action/ssoLogon.do

[4] https://zh.wikipedia.org/wiki/%E6%89%B9%E8%B8%A2%E8%B8%A2

[5] Lafferty, John, Andrew McCallum, and Fernando Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001): 282-289.

[6] Murty, Maddipati Narasimha, and Rashmi Raghava. *Support Vector Machines and Perceptrons: Learning, Optimization, Classification, and Application to Social Networks*. Springer, 2016.

[7] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." *Proceedings of the 19th international conference on World wide web*. ACM, 2010.

[8] Wang, Wei. "Chinese news event 5W1H semantic elements extraction for event ontology population." *Proceedings of the 21st International Conference on World Wide Web*. ACM, 2012.

[9] Y. Y. Huang, C.H. Chung, “A Tool for Web NER Model Generation Based on Google Snippets,” Proceedings of the 27th Conference on Computational Linguistics and Speech Processing, pp. 148–163, ROCLING, 2015.