應用條件隨機場於中文未知詞之識別

Conditional Random Field for Chinese Unknown Word Recognition

[2]Xun Zhou, [1]Jing-Yun Zeng, [1,2,*]Ying-Chih Lin, [1]Ching-Ching Yang and [3]Chih-Chung Kao
[1]Department of Applied Mathematics,
[2]Industrial Ph.D. Program of Internet of Things,
Feng Chia University, Taichung, Taiwan, R.O.C.
[3]GEO Informatics Inc.
[*]yichlin@fcu.edu.tw

摘要

許多中文的自然語言處理研究及應用都植基於中文斷字斷詞系統之上，而中文斷詞的準確率則受到未知詞(unknown word)的影響。網路的興起使得網民能在網路媒體上貢獻的大量中文文章，造成許多新興詞彙的泛濫，且隨著社會與技術發展也衍生出許多新詞。這些不在詞庫裡的未知詞彙，除了大幅降低斷字斷詞結果的正確性外，也形成自動化文章分析的瓶頸。因此，本研究發展一套識別未知詞的工具，用自行建置的詞庫和語料庫，生成訓練語料以及未知詞的特徵集，訓練以條件隨機場的統計模型為基礎的工具，用來捕捉文章裡的未知詞。同時，這套工具裡使用的參數也會經過多次測試及調校，務求構建出一個實用的辨識工具，減緩未知詞在自動化文章分析過程中造成的阻礙。

1. 介紹

「詞」是一個獨立語意或語法功能的單位，自然語言處理(Natural Language Processing, NLP)領域已經發展出許多統計的語言模型，而這些模型大多是建立在詞的基礎上，因為詞是表達語義的最小單位。對於西方的拼音語言來說，詞之間有明確的分界符號(delimit)，利用統計和語言模型處理起來相對簡單；而對於東方語言來說，詞之間沒有明確的分界，於是在套用模型之前就需要先進行斷詞的動作。以中文斷詞而言，最簡單的做法是查字典，這種做法是從左到右掃描一個句子，遇到字典裡有的詞就標識出來，遇到複合詞（如「逢甲大學」）就找最長的詞來匹配(maximum matching)，如果遇到不認識的字串就分割成單字詞，直到處理完整個句子為止。

由於網路的盛行，網民們在各種網路媒體上貢獻大量的中文文章，導致各式各樣流行用語以及新興詞彙的泛濫，例如「柯 P」、「魯蛇」、「天龍人」等，這些詞彙因為不在既有的詞庫裡，所以很難被識別出來。這些不容易被識別的字串會被斷詞方法分成零散的字詞塊，除了大大降低斷字斷詞結果的準確性之外，也形成自動化文章分析的瓶頸；此外，各種人名、商品名、地名等難以正確識別的字串，也容易引導到錯誤的語意分析方向，比如「洗板」、「婉君」、「閃光」等詞彙，都被網民們延伸出不同於原詞彙的含意。

中文的資訊處理逐漸獲得廣泛重視，但是礙於上述未知詞(unknown word)的存在，對於開發自動化文章分析技術是一個挑戰。有鑑於此，本文擬開發一套識別未知詞的工具，以機器學習式的統計模型為基礎，透過自行建立的訓練語料與特徵集，訓練出一個有效辨識未知詞的模型。之後的實驗測試則以新聞及網路文章為樣本，進行參數調校以及模型微調，務求構建出一個實用的辨識工具，緩和未知詞在文章分析過程中的帶來的負擔。

2. 相關工作

在中文語言的處理中，斷字斷詞是相當基礎且重要的一個步驟，而未知詞是影響斷詞準確度最主要的因素之一，文獻指出其影響程度甚至超過歧異詞[1]。所謂中文未知詞，指的是不在詞庫內的中文詞，可能的原因是該詞彙是一個新詞彙、流行用語、或是不在詞庫裡的中文人名、外來譯名、地名、術語等專有名詞，這些詞可以是隨著技術與社會發展而出現的詞彙或流行用語，也能是在詞庫建構過程所遺漏的字詞。據統計，一篇文章中約有 3 ~ 5%的未知詞[2]，新聞類的文章更是遠高而此，而某些類型的未知詞詞構非常複雜，往往也不具有強烈的統計特性，因此識別未知詞一直是中文語言處理上一個重要且困難的研究課題。Chen and Bai 分析中研院的語料庫[3]，發現經常出現的未知詞類型有縮寫字（如中油、中科）、專有名詞（如物聯網、大數據）、衍生詞（如谷歌化、可看性）、複合詞（如書桌、搜尋法）以及數值（如二零一五年、十五巷）等，而從中研院語料庫內的三百萬字也統計出 14 個最常出現未知詞的類別，如圖 1 所示。

Category	Frequency	Meaning of Category
A	1453	/*non-predictive adjective*/
Na	34372	/*common noun*/
Nb	14813	/*proper noun*/
Nc	9688	/*location noun*/
Nd	2264	/*time noun*/
VA	6466	/*active intransitive verb*/
VC	8462	/*active transitive verb*/
VCL	811	/*active transitive verb with locative object*/
VD	448	/*ditransitive verb*/
VE	1051	/*active transitive verb with sentential object*/
VG	996	/*classificatory verb*/
VH	10492	/*stative intransitive verb*/
VHC	498	/*stative causative verb*/
VJ	1471	/*stative transitive verb*/
total:	93,285	

圖 1：中研院語料庫最常出現的 14 類未知詞

未知詞識別的方法大體上有以下三種：(1). 基於規則的方法。這種做法通常是依據字詞的內部結構，配合上下文資訊的特徵，以人工構造的方式建立規則，進而以規則匹配的方式發現未知詞[4]。這種方法對識別指定文章中特定類型的詞彙往往能有不錯的結果，但其缺點也不少，如建立規則的人力和時間成本過高、受參與專家的影響大、主觀性較強等。(2). 基於統計的方法。這類型的做法又可分成基於統計特徵和機器學習的兩類方法。基於統計特徵的方法主要依據字詞之間的關聯度，常用的特徵包括字詞頻率、T 檢驗、資訊熵(Entropy)等，選擇特徵值符合閾值的字詞作為處理結果[5]。這種方法不依賴句法和語意訊息，具有很好的通用性，但性能在很大程度上依賴語料庫的規模和挑選的字詞頻率。另一方面，基於機器學習的方法是通過學習訓練資料的特徵構造模型以識別未知詞，常用的有隱藏式馬可夫模型(HMM)、支持向量機(SVM) [6]、最大熵模型(ME)以及條件隨機場(Conditional Random Field, CRF)等機器學習演算法[7][8]。基於機器學習的優點與基於統計特徵的方法相似，缺點則是需要人工建立學習資料，性能受制於訓練集的規模和品質。整體而言，基於統計的方法能描述大部分語言現象，有很好的通用性，但是對於詞語搭配的描述精度卻不如基於規則的方法，容易漏掉低頻詞及錯誤識別高頻詞。(3). 混合式作法。混合式方法利用多個統計方法的規則來獲得較高的辨識效能[7][8]，然而這種作法在挑選規則組合時需要特別注意，避免規則互相牴觸或是過度識別等問題。

CRF 是一種切割和標示結構性資料的機率模型，常用於標注或分析序列資料，如自然語言文字或是生物序列等。最早由 Lafferty 等人於 2001 年提到 CRF 是利用無向圖(undirected graph)模型最佳化序列標注的結果，對於指定的節點輸入值，CRF 能夠計算指定節點輸出值上的條件機率，其訓練目標是使條件機率最大化[9]。透過依序標示特定部份的觀察值來定義條件機率分佈，與隱藏式馬可夫模型相比，CRF 在處理相同應用的概念比較簡單，對於輸入輸出序列分佈的要求比較寬鬆，因此也容易處理，減輕了 HMM 相依假設的影響程度。在許多不同領域都有 CRF 的應用，像是電腦視覺[10]、中文分詞[11]、淺層句法分析[12]、行為分析[13]等領域。

CRF 的結構以線性鏈(linear-chain)是最常見的特定圖，由指定的輸出節點按順序鏈接而成，一個線性鏈與一有限狀態機(finite state machine)相對應，可用於解決序列數據的標注問題。若 $x=x_1, x_2, \cdots, x_n$ 為一給定的觀測值序列(例如一個中文詞序列)，亦即無向圖模型中 n 個輸入節點上的值；$y = y_1, y_2, \cdots, y_n$ 為一個狀態序列，每個狀態均來自於一個有限狀態機的狀態集合，而每個狀態可對應到一個標記，y 的狀態序列對應到無向圖模型中 n 個輸出節點上的值。一個線性 CRF 把輸入序列 x 得到的狀態序列 y 的條件機率定義如下，其中 $Z_\Lambda(x)$為正規化因子，使得給定輸入所有可能的狀態序列之機率和為 1：

$$P_\Lambda(y|x) = \frac{1}{Z_\Lambda(x)} \exp\left(\sum_{i=1}^{n}\sum_{k}\lambda_k f_k(y_{i-1}, y_i, x, i)\right)$$

需要注意 CRF 模型中 $Z_\Lambda(x)$的計算量相當龐大，因為它涉及到所有狀態的組合，幸好在線性鏈模型的節點間沒有循環路徑，因此可透過動態規劃法快速地運算，且尋找最可能狀態序列的問題，也可以用同一技巧處理，常見的作法是利用維特比(Viterbi)演算法來處理[9]。上式中的 $f_k(y_{i-1}, y_i, x, i)$是一個特徵函數，λ_k 是從語料中訓練而得，做為特徵 f_k 的權重參數。當考慮觀察序列 x 的第 i 個觀察值時，就可以用特徵函數來衡量 $y_{i-1} \rightarrow y_i$ 狀態轉換過程的各種可能性，如果得到一個較大的正數，代表事件很有可能發生；反之，負數則表示事件傾向於不發生[9]。最早採用 CRF 對句子分析進行淺層分析(shallow parsing)的是 Sha and Pereira [14]，他們繼承了 Ratnaparkhi 的方法，只做了句子分析的第一層，即從詞到詞組的自動組合。由於改進了統計模型，因此句子的淺層分析的正確率高達 94% [14]。另一方面，目前也已經有許多實作通用 CRF 模型的工具套件[15]，其中 CRF++是許多人認為最有效率的一階線性 CRF 工具[16]。

3. 計算方法

3.1. 詞庫與斷字斷詞

在文章中根據撰寫者的不同，可能會有不同的習性有些人習慣使用全形或半形的字體。然而語料庫訓練集當中大多為單一型態，由於中央研究院現代漢語語料庫皆為全形，故本工作先將所有的收集資料標準化同一規格為全形。此外，對基於字串匹配的方法來說，詞庫扮演著相當重要的

角色，一個好的詞庫應該要能包含文章的大部分字詞，至少是包含常用的字詞。然而，過大的詞庫容易造成資源的浪費，而過於精簡的詞庫則會降低分詞的效果，因此詞庫的建構需要相當小心謹慎。我們所使用的詞庫是幾個部分拼湊而成，首先挑選開源斷詞程式 Jieba [17]使用的簡體詞庫，將該詞庫（共十萬餘詞）轉成繁體。Jieba 的詞庫附有詞性的標注，採用中國科學院計算技術研究所的漢語詞法分析系統(Institute of Computing Technology, Chinese Lexical Analysis System, ICTCLAS)漢語詞性標注集的方式，方便我們標示出名詞的詞語，並以中研院漢語平衡語料庫的詞類標記為主進行轉換[18]。以轉換後的 Jieba 詞庫為基礎，再添加一些常用詞彙整合成我們的詞庫，例如教育部成語典、地名、學術名詞、台灣名人、百家姓等，詞庫最後總計共十八萬餘詞。

3.2. 條件隨機場模型(CRF)

本研究採用 CRF 模型做為未知詞識別的主要框架。訓練方式採用外部特徵為主，包括未知詞前、後 1~2 字詞與其詞性。這些特徵能充分利用未知詞上下文知識與字詞的內部訊息，呈現出詞性、字詞間的上下文關係，以此識別新詞。另一方面，本文利用相當受歡迎且以 C++撰寫的開源工具 CRF++ [16]，這是許多人認為最有效率的一階線性 CRF 工具之一。CRF++不僅允許使用者重新定義特徵集合，也能給出最高機率的 N 個組合(N-best)，對於大型資料的訓練與測試也有較快速的擬牛頓法(quasi-newton algorithm)以及使用較少記憶體的設計。這些特色都有助於加速本文所建立模型的運算，並對運算結果有更多改進的空間。

3.3. 未知詞識別

影響 CRF 模型效能有許多原因，其中特徵的選取是重要的因素之一。在 NLP 應用中常用的有字詞、詞性、邊界以及前後綴等特徵，而與歐美語言的特質不同，在中文 NLP 任務中，邊界與前後綴做為特徵的價值往往不如字詞及詞性。詞性是詞的重要屬性，表示詞在語法意義上所屬的類別，幾乎決定了詞與詞之間的組合方式以及詞在句子中的位置，所以無論在中文還是歐美自然語言文本處理中都是最重要的特徵。然而，中文文章的詞與詞之間缺乏自然分隔，在無法準確切分詞語的情況下，詞性也就無用武之地，此時只能退而求其次，以字為特徵，但是單一個字既無法詮釋該詞的語意，也不能決定在句子中的位置和排列方式，這個特徵也就失效了。因此，本文在在序列的基礎上補充詞與詞性或類似的特徵，以提升模型的未知詞識別能力，並以開源工具 CRF++做為計算 CRF 模型的工具。整體流程如圖 2 所示，說明如下：

(1). 詞庫：利用前述做法，使用從網路以及其他管道所蒐集的中文字詞及詞性，建立自己的詞庫做為模型的初步訓練及測試，並視結果的優劣考慮是否採用「中華民國計算語言學學會」的中文詞庫（約 8 萬詞）[19]。

(2). 語料庫：語料庫中的字詞帶有詞性的標記，且經過專人校正過，具有極高的正確率，非常適合做為樣本的抽取。「中央研究院現代漢語語料庫」(Sinica Corpus)內含有五百萬詞，於 1997 年 10 月建立完成並開放線上使用[18]，可以做為初期語料庫。同樣地，視測試結果的優劣考慮是否購買新版的平衡語料庫。

(3). 建立訓練集語料：從前述步驟的清單中抽取指定類術語的語料，包含常見的名詞、動詞、形容詞、副詞、連接詞、介詞、語助詞等大類別，以此建立訓練集語料，對語料進行分詞和詞性標註。

(4). 語料庫測試：將語料庫分成三份，測試用三分之一、三分之二、全部的語料庫，比較其成效。以分析訓練集之輸入優劣。

(5). 資料格式轉換：轉換分詞和詞性標註後的測試、訓練集資料，添加 SBIEO 標記形成特徵模版(如後所述)，以便於符合 CRF++工具套件所要求的格式。

(6). 執行 CRF 運算：配置 CRF 特徵模板並結合訓練集得到的語料，分析輸入的文章，獲取 CRF++輸出的未知詞識別結果。

(7). 詞庫更新：辨識出來的未知詞經過人工確認無誤後，可以新增至詞庫並標示正確的詞性。

在 CRF++的標注集中，每個漢字在一個特定的詞語中都佔據著一個確定的構詞位置，也就是詞位。未知詞識別可視為詞位標注問題，亦即透過確定每個字的詞位標注來完成識別的工作。詞位標注(Inside Outside Beginning, IOB)格式是在計算語言學中，標記在片段語詞內字詞的標籤，有兩種廣泛使用的命名實體(named entity)片段表示法，Inside/Outside 表示法以及表示 Start/End 資訊。假如我們有一些詞語的樣本，表明哪些字詞(token)在詞語中的開始與結束位置，那麼 Start/End 的表示法會比較有用，因為能讓我們進行邊界的標定[20]。常用的詞位標注集有二字位、四字位、五字位及六字位標注集，本文採用五字位標注集，亦即 SBIEO 的標注方式，其中 S (Single)表示單字詞，B (Begin)

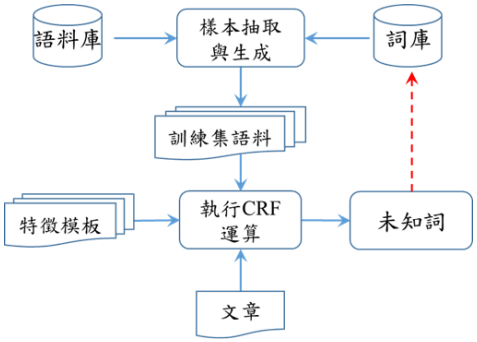

圖 2：未知詞識別流程

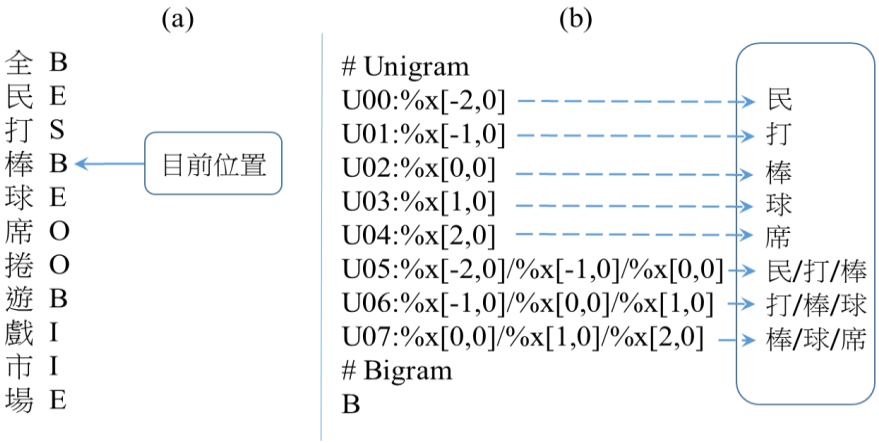

圖 3：訓練集(a)與特徵模板(b)的範例

表示該詞的第一個字，I (In)表示詞中間的字，E (End)表示詞的最後一個字，而 O (Out)則代表不屬於標注對象的字，底下是一個利用 SBIEO 標注方式的範例。

『全(B)民(E)打(S)棒(B)球(E)席(O)捲(O)遊(B)戲(I)市(I)場(E)』

在本文中，CRFs 與詞庫、語料庫相結合的模型採用字詞本身的特徵，且為了觀察字詞語上下文的關係，以便於發現其在字詞序列中的長距離關聯，所以將特徵視窗的大小向前後延展。然而，視窗越大將使模型的訓練時間變長，為了兼顧模型效能和運算時間，依經驗先設定視窗長度為 5，這個設定數值是鑒於一般未知詞的長度都不會太長。此外，特徵模板的格式則根據 CRF++官網的規定[16]，訓練集裡至少要有兩行（字詞與標注集），若有需要加入詞性標注，可添加在這兩行中間。圖 3 是一個訓練集與特徵模板的簡單範例，最後總共會生成 LxN 個特徵函數，其中 L 是指標注集的大小，也就是 5，而 N 指的是通過模板擴展出來的所有單個字串（特徵）的個數。

4. 實驗測試

本文採用中央研究院漢語平衡語料庫來進行 CRF 模型的訓練，本文將採用三種不同的訓練集來比較。分別為 1/3、2/3、全部的語料庫來進行訓練。訓練集的大小最直接影響的是訓練時間，1/3 的訓練時間大略 30 分鐘，而全部資料及的訓練大約需要 90 分鐘。訓練及測試介面如圖 4 所示。

圖 5 是一個斷字斷詞的結果，其中程式自動斷出來的字詞以斜線(/)分隔開，而以紅色底線顯示的字詞逮表的是程式自動判斷出來的未知詞。由這個實驗範例可以得知我們所發展的方法有不錯的效果，一些沒有在辭庫內的地名（如「龍崎區」、「大樹區」、「牛埔」等）、組織單位（如「發展司」、「水保局」），甚至是慣用語（如「臺泰」、「泰方」）

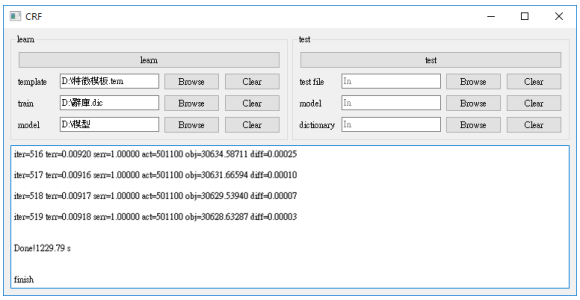

圖 4：進行訓練的程式介面

皆能識別出來。儘管能找出大部分的未知詞，但還

泰國/農業部/土地/發展司/一/行/5人/，/3月/8日/、/9日/由/農委會/水土/保持局/陪同/，/參訪/臺南市/龍崎區/「/牛埔/泥岩/水土/保持/教學/園區/」/及/高雄市/大樹區/「/龍目/社區/」/。
水保局/臺南/分局/表示/，/這/次/泰國/農技/專家/來/南部/參訪/，/屬於/臺泰/農業/合作/計畫/第2/階段/的/工作/項目/之/一/，/主要/目的/為/使/泰方/人員/瞭解/臺灣坡/地/開發/示範區/規劃/執行/及/泥岩/地形/整/治/水土/保持/工法/，/並/師法/水土/保持/戶外/教室/建置/與/運作/方法/，/另外/也/冀望/透過/農再/社區/示範/案例/參訪/，/使/泰方/人員/瞭解/我國/發展/農村/產業/環境/與/土地/利用/之/經驗/。

圖 5：測試範例的斷詞結果

是有需要改進的地方，例如：「臺灣坡/地」應該合併為一個詞彙、「龍目/社區」也應該合併為一個地名等。

為了更加了解本研究設計方法之效能，我們也廣泛地對未知詞的識別作精準度(Precision)與召回率(Recall)的測試比較。在下表的測試結果中，A_1 是第一篇測試文章，A_2 則是第二篇；而 T_{all} 代表使用整個語料庫，$T_{2/3}$ 則僅使用語料庫的 2/3。從表格中可得知語料庫大小並不太影響召回率，但會使得準確率下降，但從 F-measure 可得知語料庫大小相當重要，語料庫太小會使得 F-measure 大幅下滑。

	$\mathbf{A_1}T_{all}$	$\mathbf{A_2}T_{all}$	$\mathbf{A_1}T_{2/3}$	$\mathbf{A_2}T_{2/3}$	$\mathbf{A_1}T_{1/3}$	$\mathbf{A_2}T_{1/3}$
Precision	0.88	0.80	0.89	0.58	0.73	0.44
Recall	0.88	0.94	0.94	0.85	0.85	0.85
F-measure	0.88	0.86	0.91	0.69	0.79	0.58

表 1：A 表示為文章，T 表示為訓練集

5. 結論

中文斷字斷詞是許多中文的自然語言處理工作的基礎，而未知詞是降低中文斷詞準確率的一個重要因素。透過各種社交、網路媒體產生大量流行用語，且隨著社會與技術發展也衍生出許多新詞。這些不在詞庫裡的未知詞彙，除了大幅降低斷字斷詞結果的正確性外，也形成自動化文章分析的瓶頸。有鑒於此，本工作發展一套識別未知詞的工具，用自行建置的詞庫和語料庫，生成訓練語料以及未知詞的特徵集，訓練以條件隨機場的統計模型為基礎的工具，用來捕捉文章裡的未知詞。

從實驗的測試結果可知本研究所設計的方法，在準確率與召回率上皆有不錯的表現。未來除了繼續探討語料庫的大小、優劣對於未知詞識別的影響外，也考慮如果自動化訓練模型，能有效率地處理新產生語料庫進行再訓練的議題。

References

[1] C.N. Huang and H. Zhao (2007) “Chinese word segmentation: A decade review,” *Journal of Chinese Information Processing*, 21(3), pp. 8–19.

[2] K.-J. Chen and W.-Y. Ma (2002) “Unknown word extraction for Chinese documents,” *Proceedings of Coling*, pp.169-175.

[3] K.-J. Chen and M.-H. Bai (1998) “Unknown word detection for Chinese by a corpus-based learning method,” *Computational Linguistics and Chinese Language Processing*, 3(1), pp. 170–177.

[4] W.-Y. Ma and K.-J. Chen (2003) “A bottom-up merging algorithm for Chinese unknown word extraction,” *2nd SIGHAN Workshop on Chinese Language Processing*, pp. 31–38.

[5] X. Wen (2015) “New word identification for Chinese patents based on multiple statistic measures and pattern combination,” *International Conference on System Reliability and Information Technology*.

[6] Z. Lu, Z. Yan and J. Gu (2013) “A novel schema-oriented approach for Chinese new word identification,” *27th Pacific Asia Conference on Language, Information and Computation*, pp. 108–117.

[7] X. Jiang, Y. Cao and Z. Lu (2011) “Automatic recognition of Chinese unknown word for single-character and affix models,” *6th International Conference on Intelligent Systems and Knowledge Engineering*, 435–444.

[8] N. Xi et al. (2012) “Adapting conventional Chinese word segmenter for segmenting micro-blog text,” *2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pp. 63–68.

[9] J.D. Lafferty, A. McCallum and F.C. Pereira (2001) “Conditional random fields: probabilistic models for segmenting and labeling sequence data,” *18th International Conference on Machine Learning*, pp. 282–289.

[10] X. He, R.S. Zemel and M.A. Carreira-Perpinñán (2004) “Multiscale conditional random fields for image labeling,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

[11] H. Zhao, C.-N. Huang and M. Li (2006) “An improved Chinese word segmentation system with conditional random field,” *5th SIGHAN Workshop on Chinese Language Processing*, pp. 162–165, Sydney, Australia.

[12] F. Sha and F. Pereira (2003) “Shallow parsing with conditional random fields,” *Proceedings of HLT-NAACL 2003*, pp. 213–220.

[13] K. Bousmalis, S. Zafeiriou, L. Morency and M. Pantic (2013) “Infinite hidden conditional random fields for human behavior analysis,” *IEEE Transactions on Neural Networks and Learning Systems*, 24(1), pp. 170–177.

[14] F. Sha and F. Pereira, (2003) “Shallow parsing with conditional random fields,” *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Edmonton, Canada, pp. 213–220.

[15] Generic CRF tools, https://en.wikipedia.org/wiki/Conditional_random_field#Software

[16] CRF++: Yet Another CRF toolkit, https://taku910.github.io/crfpp/

[17] Jieba, https://github.com/fxsjy/jieba

[18] 中央研究院—現代漢語平衡語料庫，http://app.sinica.edu.tw/kiwi/mkiwi/

[19] 中華民國計算語言學學會，http://www.aclclp.org.tw/corp_c.php

[20] C. Sutton and A. McCallum, (2011) “An introduction to conditional random fields,” *Foundations and Trends in Machine Learning*, 4(4), pp. 267–373.