**WeCanAPI\_v2.0.1310技術手冊**

張道行 Tao-Hsing Chang

許曜麒 Yao-Chi Hsu

國立高雄應用科技大學資訊工程系

Department of Computer Science and Information Engineering

National Kaohsiung University of Applied Sciences

[changth@kuas.edu.tw](mailto:changth@kuas.edu.tw)

[r1099108121@gmail.com](mailto:r1099108121@gmail.com)

摘要

1. WeCan斷詞系統介紹

斷詞系統是自然語言處理中非常基礎也十分重要的工具，很多的應用都與之相關。而斷詞系統「WeCan」提供了斷詞與詞性標記兩個主要的函數工具，而斷詞所用到的字典是來自國立臺灣師範大學所提供的「TERMFINAL」之V2\_1版本，其字典提供了227622個詞彙，為了配合WeCan斷詞所需調整了字典的內容，最後剩餘224535個詞彙，其調整的內容主要是修改人為標記錯誤的詞性。而WeCan運用其字典作為斷詞參考完成文句的分割，主要以長詞優先斷詞、CRF等方法並用，詞性標記則額外參考「SinicaTreebank\_3.1」的詞性頻率幫助提升準確度，而斷詞的詳細運作將在接下來的章節做解說。

WeCan由三個主要API組成，分別是WordSegment、SegmentException、UnKnownWordHandle。首先最核心的API－WordSegment，提供了WeCan斷詞系統的斷詞、詞性標記的函數。第二，SegmentException是處理WordSegment無法正確斷詞的特殊狀況，其處理對象都是非常顯著且變化較小的樣本，例如時間、數字、英文單字、疊字、華人姓名等等。第三，UnKnownWordHandle是藉由文章訓練可能的未知詞，且設定一些相關的門檻值來限制未知詞的篩選，以達到較高的準確率，所以在文章的資訊量必須要有一定的大小才能有顯著的效果。

1. 斷詞函數(WordSegment)
2. 長詞優先斷詞

本系統主要使用的斷詞方法為「長詞優先斷詞」，其方法與一般讀者在閱讀文章時對於所得到的資訊反射性地由較長的詞彙開始分割句子十分相似，例如：

中山大學校長的戴傳賢

→中山大學校長(X)

→中山大學校(X)

→中山大學(O)

中山大學　校長　的　戴傳賢

從較長的「中山大學校長」開始確認字典中是否存在其詞彙，逐一減少字詞直到詞彙存在或是剩餘單一字時，接著重複此步驟以完成整個句子的分割處理。

1. CRF斷詞

CRF斷詞的模型是訓練自「SinicaTreebank\_3.1」，CRF斷詞的準確度與效能方面雖然沒有優於長詞優先斷詞，但是仍然能解決部分長詞優先斷詞無法解決的問題，例如：

正解： 這種(DM)　 情景(Nad)　把(P07)　我(Nhaa)　嚇壞(VH22)　了(Ta)

CRF： 這種(b)　　 情景(Na)　 把(P)　 我(Nh)　 嚇壞(VHC)　 了(T)

長詞優先： 這(Nep) 種(VC) 情景(Na)　 把(P)　 我(Nh)　 嚇壞(VHC)　 了(T)

「這種」詞彙在字典檔中並不存在，但是CRF是經過統計訓練出來的資料，透過訓練資料提出「這」、「種」兩個字應該要被合併，但因長詞優先斷詞有較高的成功率，所以大部分的斷詞結果依舊是選擇長詞優先斷詞，少部分才選擇CRF斷詞。

1. 詞性標記

詞性標記的公式，主要使用到「詞性於句首、句尾之頻率」、「詞性總頻率」、「詞性對詞性的頻率」、「各詞彙對應詞性之頻率」等依據，使用改良過的貝氏機率公式來決定哪種組合的詞性機率最高，公式如下：

例句： 今天 天氣 很好

詞性： A B C

設詞性A、B、C為例句「今天天氣很好」的其中一組詞性組合，表示詞性A對詞性B的出現頻率，而「」為詞彙”天氣”以詞性B出現的頻率，故某些詞彙對應詞性所出現的頻率過高，容易影響其他頻率較小但可能為正確答案的詞性，因此透過ln來減少頻率過高所造成的影響。而多組詞性組合經過上述公式得到機率值後，選出機率最高的組合作為詞性標記。

1. 斷詞例外處理(SegmentException)
2. 長詞優先斷詞之例外

長詞優先斷詞因為符合常人閱讀文章時的習慣，所以鮮少有錯誤產生，但也因為是由前至後尋找詞彙是否存在，還是有特殊的情況須被重視：

(X)到處　有空　地

(O)到處　有　空地

雖然「空地」這個詞彙在字典裡是存在的，但是「有空」卻因為長詞優先斷詞是由前至後尋找而更早被斷詞，這種斷詞方法很有可能忽略更適合的詞彙，因此這種情況時需要特別處理。

1. 數字、時間、英文單字合併

在字典中並不會提供所有數字、英文單字的組合，所以在遇上數字、時間、英文單字時，總會被斷開，而數字的格式範圍較小，適合特別處理，例如：

九　乘　九　等於　八十一

我　出生於　一九九二年

也　就是　designin　事前　規劃　的　觀念

時間方面，就針對一些較常見的時間單位詞進行處理，例如年、月、日、時、點、分、秒等等，在數字處理完畢後合併。

1. 疊字

某些疊字的詞彙在字典中也是不存在的，因此為這些疊字詞彙訂定一些規則，對詞性、詞彙進行篩選以完成疊字處理，例如：

幼稚園　小朋友　又跑又跳

尋尋覓覓　，　冷冷清清　，　淒淒　慘慘　戚戚　。

與　師傅　切磋切磋　武藝

「又跑又跳」屬於第一與第三個字相同，且四個字皆被斷開，而四個字中未有出現N、DE、P等詞性，符合上述疊字條件則可將四個字斷為一個詞彙並賦予詞性「VA」；「淒淒」、「慘慘」、「戚戚」皆為兩個重複的單一字，其詞性皆不能有Nb、SHI、DE、T、P等等，則兩個重複的單一字即可被斷為一個詞彙並賦予詞性「VH」；「切磋切磋」為兩個相同的二字詞彙，其詞性皆不為Na，符合規定則可以合併成一個詞彙並賦予詞性「VC」。

1. 華人姓名

華人姓名絕大多數都是三個字或兩個字，且大部分的姓氏都可以由「百家姓」等資料提供，由於二字華人姓名斷詞效果較差，目前WeCan只提供三個字的華人姓名斷詞，例如：

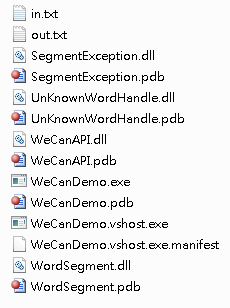
2012年　11月　，　蔣偉寧　擔任　教育部長　期間

「蔣偉寧」在未經過華人姓名處理前是三個分開的單一詞，其「蔣」為百家姓的姓氏之一，而「偉」、「寧」之詞頻皆小於門檻值，則可視為華人姓名處理。

1. 未知詞處理(UnKnownWordHandle)

未知詞正是字典檔未出現過的詞彙，但是在字典檔未出現過的詞彙且又不符合斷詞例外處理的詞，在沒有任何背景知識的支持之下，如何確定這類詞彙是否該分割？其實可以依據「出現的次數」來決定是否分割未知詞彙，將文章的每個句子當成「父句()」，而每個父句的排列組合為「子句()」，而又可分為左子句()、右子句()，表示在文章中出現的頻率，即可計算其SPLR(strict phrase likelihood ratio)之值：

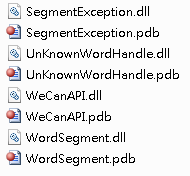
以及代表依照實驗結果得到的門檻值。符號ε、μ以及是依照實驗結果得到的容錯係數。若符合下列四種條件，即可成為未知詞：

1. WeCanAPI使用說明
2. 參考WeCanAPI.dll

WeCan API 是由三個API組成，分別是：

* + WordSegment
  + SegmentException
  + UnKnowWordHandle

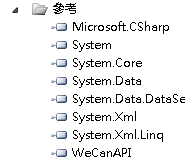
結合三個API之dll為WeCanAPI，圖一的八個檔案都是斷詞所需的API函數，必須與執行檔同目錄，如圖二。



圖二

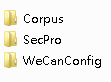
圖一

而使用斷詞函數，只需要參考WeCanAPI.dll即可，如圖三。



圖三

1. 字典檔

****

* + WordSegment.dll使用Corpus字典檔
  + SegmentException.dll使用SecPro字典檔
  + UnKnowWordHandle.dll未使用任何字典檔

預設的情況下，字典檔都與執行檔需要放在同一目錄下，而要額外設定字典檔的目錄，可以從WeCanAPI的建構子下設定，在接下來的章節有更詳細的說明。

1. WeCanAPI.dll函數呼叫

WeCanAPI提供兩種函數呼叫：

* public void IOSegment(string input, string output)
* public void Segment(string input\_str,

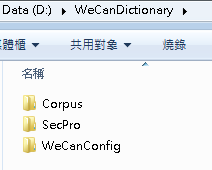
List<string> output\_Segment,

List<string> output\_POS)

而呼叫函數前，需先宣告變數：

WeCanAPI.WeCanAPI wecan = new WeCanAPI.WeCanAPI();

WeCanAPI.WeCanAPI wecan = new WeCanAPI.WeCanAPI(@"D:\WeCanDictionary");

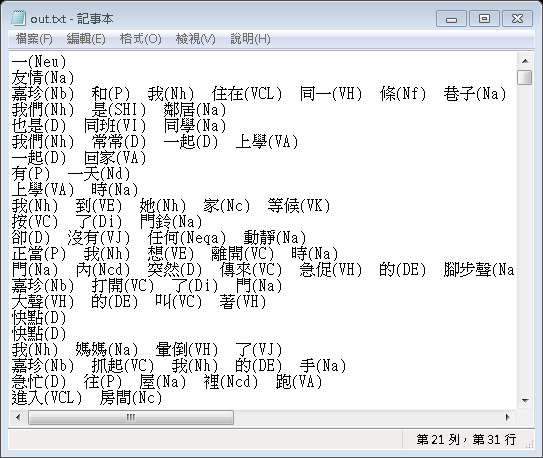
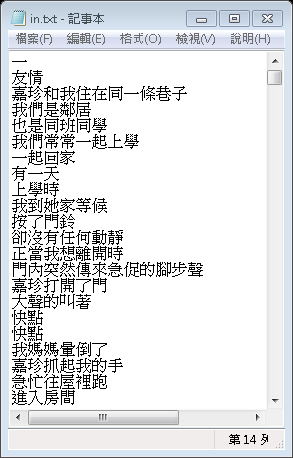
建構子中可以指定字典檔的位置，如圖四，不指定則與執行檔同目錄底下。

圖四

* IOSegment

IOSegment為指定輸入檔案(\*.txt)，而將斷詞結果輸出於指定檔名(\*.txt)，如圖五。

wecan.IOSegment("in.txt", "out.txt");



圖五

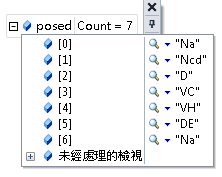
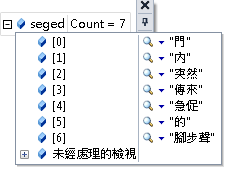
* Segment

Segment為即時呼叫的斷詞函數，需先new兩個List<string>讓函數放置斷詞結果，字串經過Sement()後可於seged、posed取得其斷詞結果與詞性標記結果，如圖六。

List<string> seged = new List<string>();

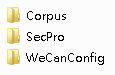
List<string> posed = new List<string>();

wecan.Segment("門內突然傳來急促的腳步聲", seged, posed);



圖六

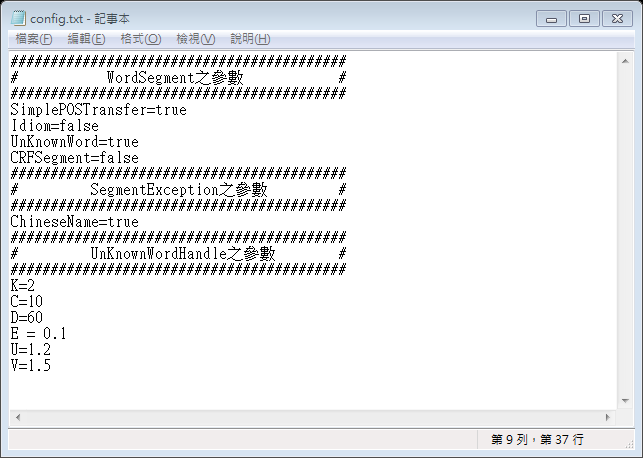
1. WeCanConfig設定檔說明

****

圖七

字典檔方面有Corpus、SecPro等資料夾分別由WordSegment.dll與SegmentException.dll讀取使用，而圖七WeCanConfig資料夾中有config.txt檔案，其檔案可以設定載入哪些字典檔以及某些斷詞功能是否開啟等等。資料夾WeCanConfig->config.txt都是固定的名稱，若進行更改將無法讀取，但是讀不到WeCanConfig->config.txt不代表斷詞系統無法運作，斷詞將會以預設狀態進行其功能(其預設狀態如圖八)。在config.txt中若某行以「#」開頭則此行為註解。

* **Corpus、SecPro為非常重要之字典檔資料夾，不可任意修改其檔案名稱。**
* **每個參數的大小寫皆是固定的，請依照下面圖片之大小寫輸入參數。**

****

**WordSegment之參數設定：**

圖八

1. SimplePOSTransfer = true/false；設定輸出時是否為簡易詞性標記(true)，否則為原始詞性標記(false)。
2. Idiom = true/false；設定是否載入四字詞字典，是為(true)，否(false)。
3. UnKnownWord = true/false；設定是否使用未知詞修正斷詞結果，是為(true)，否(false)。
4. CRFSegment = true/false；設定是否使用CRF斷詞輔助WordSegment斷詞準確度，是為(true)，否(false)。

**SegmentException之參數設定：**

1. ChineseName = true/false；是否使用華人姓名修正斷詞結果，是為(true)，否(false)。

**UnKnownWordHandle之參數設定：**

六個參數中()符合以上四個條件的詞彙即可列為未知詞，其六個門檻值可在設定檔config.txt中設定。

附錄(一) 48種簡化詞性標記與詳細詞性標記轉換表

－參考「中研院平衡語料庫」詞類標記表

|  |  |  |
| --- | --- | --- |
| 簡化標記 | 詳細標記 | 詞類標記 |
| A | A | /\*非謂形容詞\*/ |
| Caa | Caa | /\*對等連接詞，如：和、跟\*/ |
| Cab | Cab | /\*連接詞，如：等等\*/ |
| Cba | Cbab | /\*連接詞，如：的話\*/ |
| Cbb | Cbaa,Cbba,Cbbb,Cbca,Cbcb | /\*關聯連接詞\*/ |
| D | Dab,Dbaa,Dbab,Dbb,Dbc,Dc,Dd,Dg,Dh,Dj | /\*副詞\*/ |
| Da | Daa | /\*數量副詞\*/ |
| DE | DE | /\*的,之,得,地\*/ |
| Dfa | Dfa | /\*動詞前程度副詞\*/ |
| Dfb | Dfb | /\*動詞後程度副詞\*/ |
| Di | Di | /\*時態標記\*/ |
| Dk | Dk | /\*句副詞\*/ |
| FW | FW | /\*外文標記\*/ |
| I | I | /\*感嘆詞\*/ |
| Na | Naa,Nab,Nac,Nad,Naea,Naeb | /\*普通名詞\*/ |
| Nb | Nba,Nbc | /\*專有名稱\*/ |
| Nc | Nca,Ncb,Ncc,Nce | /\*地方詞\*/ |
| Ncd | Ncda,Ncdb | /\*位置詞\*/ |
| Nd | Ndaa,Ndab,Ndc,Ndd | /\*時間詞\*/ |
| Nep | Nep | /\*指代定詞\*/ |
| Neqa | Neqa | /\*數量定詞\*/ |
| Neqb | Neqb | /\*後置數量定詞\*/ |
| Nes | Nes | /\*特指定詞\*/ |
| Neu | Neu | /\*數詞定詞\*/. |
| Nf | Nfa,Nfb,Nfc,Nfd,Nfe,Nfg,Nfh,Nfi | /\*量詞\*/ |
| Ng | Ng | /\*後置詞\*/ |
| Nh | Nhaa,Nhab,Nhac,Nhb,Nhc | /\*代名詞\*/ |
| P | P | /\*介詞\*/ |
| SHI | SHI | /\*是\*/ |
| T | Ta,Tb,Tc,Td | /\*語助詞\*/ |
| VA | VA11,VA12,VA13,VA3,VA4 | /\*動作不及物動詞\*/ |
| VAC | VA2 | /\*動作使動動詞\*/ |
| VB | VB11,VB12,VB2 | /\*動作類及物動詞\*/ |
| VC | VC2,VC31,VC32,VC33 | /\*動作及物動詞\*/ |
| VCL | VC1 | /\*動作接地方賓語動詞\*/ |
| VD | VD1,VD2 | /\*雙賓動詞\*/ |
| VE | VE11,VE12,VE2 | /\*動作句賓動詞\*/ |
| VF | VF1,VF2 | /\*動作謂賓動詞\*/ |
| VG | VG1,VG2 | /\*分類動詞\*/ |
| VH | VH11,VH12,VH13,VH14,VH15,VH17,VH21 | /\*狀態不及物動詞\*/ |
| VHC | VH16,VH22 | /\*狀態使動動詞/ |
| VI | VI1,VI2,VI3 | /\*狀態類及物動詞\*/ |
| VJ | VJ1,VJ2,VJ3 | /\*狀態及物動詞\*/ |
| VK | VK1,VK2 | /\*狀態句賓動詞\*/ |
| VL | VL1,VL2,VL3,VL4 | /\*狀態謂賓動詞\*/ |
| V\_2 | V\_2 | /\*有\*/ |
| Nv | Nv1,Nv2,Nv3,Nv4 | /\*名物化動詞\*/ |
| b | b | /\*未知詞\*/ |