**中文斷詞與詞性標記工具WECAn**

**WECAn: A Software Development Kit for Chinese Word Segmentation and Part-of-Speech Tagging**

張道行 Tao-Hsing Chang

國立高雄科技大學資訊工程系

Department of Computer Science and Information Engineering, National Kaohsiung University of Science and Technology

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

曾靖維 Jing-Wei Ceng

國立高雄科技大學資訊工程系

Department of Computer Science and Information Engineering, National Kaohsiung University of Science and Technology

[gg0939240695@gmail.com](mailto:gg0939240695@gmail.com)

**摘要**

本文說明WECAnAPI斷詞工具的功能、環境設定以及如何使用本工具，作為使用者參考。

1. **系統概觀**

斷詞及詞性標記系統是自然語言處理中非常基礎也十分重要的工具，WECAn (Word Extraction for Chinese Analysis)[1]能提供中文斷詞與詞性標記的功能，其作法是將一句文字或一篇文章分割為詞彙並標記每個詞的詞性。例如句子「今天天氣很好」經過斷詞與詞性標記後會得到：

今天(Nd) ＿天氣(Na)＿很(Dfa) ＿好(VH)

句子「今天天氣很好」被分割為詞彙「今天」、「天氣」、「很好」並分別給予詞性「Nd(時間詞)」、「Na(普通名詞)」、「VH(狀態不及物動詞)」。特別需要說明的是，WECAn針對詞類”b(未知詞)”做了禁止輸出的處理，不管是採用哪種參數、設定與方法，當計算結果為詞類”b”時，WECAn將以統計出來的可能詞性列表結合前後詞性運算後給予一個可能的詞性。

本論文在第2及第3節介紹系統所需資源檔的來源以及環境設定的步驟，在第4、5、6節介紹使用方法以及參數設定。

1. **引用資源**

WECAn繁體使用的字典來自國立臺灣師範大學數位學習實驗室的「TERMFINAL V2.1」版本，其參考中研院平衡語料庫、中文句結構樹資料庫及中文八萬目詞語料整理而成；簡體為北大樹庫與語委。因此使用WECAn者必須持有上述三項資料庫之授權。為了方便WECAn使用，我們修改了字典Cindy(繁體)、Simon(簡體)的部分資料，以下為主要分類：

1. 刪除WECAn起始字、字長等欄位
2. 刪除VR、VP、DM等詞性
3. 修改人為標記造成的錯誤詞性
4. 刪除數字與時間等詞彙
5. **資源檔**

WECAnAPI系統全部使用Visual Studio C# 撰寫，在函數運作前必須設定環境，需要使用的API與資源檔皆包裝在WECAnAPI.rar，詳細的資料內容如圖1所示。

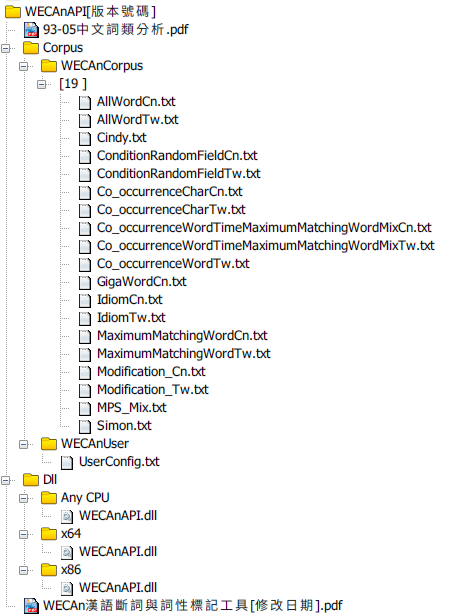


圖1 WECAnAPI壓縮檔詳細資料內容

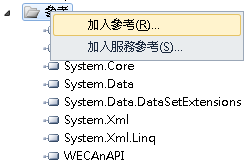
在函數呼叫前需要由使用者自行將WECAnCorpus與WECAnUser兩個資料夾複製到執行檔目錄或指定目錄下，WECAn所讀取的所有字典檔與設定檔皆為UTF-16(Little Endian)編碼。另外目錄WECAnUser下的UserWord.txt是可有可無的文字檔，關於UserWord.txt檔案的作用，將在第6節做更詳細的說明，其他目錄與純文字檔皆不可隨意更改名稱或刪除，而上述兩個目錄所在路徑未必要與執行檔相同目錄，設定方法在第5節將會詳細說明。

WECAnCorpus為函數所需的字典資源，而WECAnUser\UserConfig.txt為WECAnAPI的參數設定檔案，在第6節會有更詳細的說明。

WECAnAPI.dll是WECAnAPI主要的函數檔案，使用方式也在第5節有詳細的說明。

1. **參考WECAnAPI.dll**

在第3節資源檔環境設定完成後，使用者在呼叫WECAnAPI函數前須先設定參考WECAnAPI.dll，操作畫面如圖2。



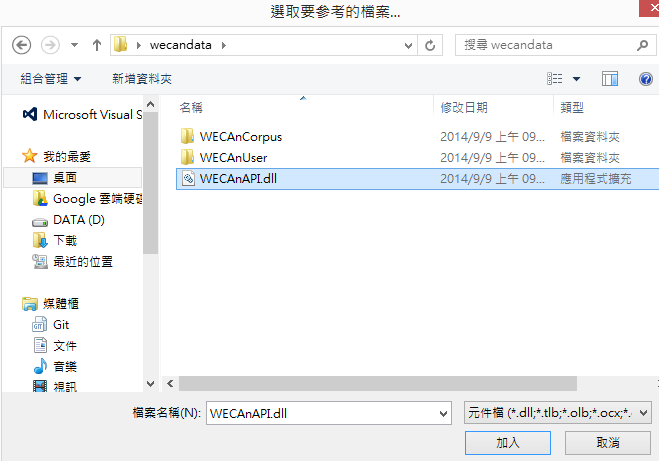


圖2 參考WECAnAPI.dll之步驟

1. **WECAnAPI.dll宣告與函數呼叫**

WECAnAPI.dll包含兩個主要函數: IOSegment()以及Segment()，分別供大量批次處理以及應用程式嵌入使用。而使用前請參考5.1之參考宣告方式後方可使用。

* 1. WECAnAPI.dll宣告

WECAnAPI根據資源檔路徑指定與否，分為兩種宣告方式。第一種為預設宣告，不輸入任何參數，其資源檔路徑就是執行檔所在目錄。例如：

WECAnAPI.WECAnAPI wecan = new

WECAnAPI.WECAnAPI();

第二種宣告方式可以指定資源檔目錄，假如欲指定目錄為"D:\WECAnDictionary"，則可宣告為：

WECAnAPI.WECAnAPI wecan = new

WECAnAPI.WECAnAPI

(@"D:\WECAnCorpus");

若宣告路徑為"D:\WECAnCorpus"，則第3節介紹的WECAnCorpus與WECAnUser資料夾需要複製到"D:\WECAnCorpus"內。

* 1. WECAnAPI.dll函數呼叫

在5.1節將物件wecan宣告完成後，其提供三種函數呼叫，第一種是使用文字檔案輸出入的方式進行斷詞，較適合大量資料斷詞。使用範例如下：

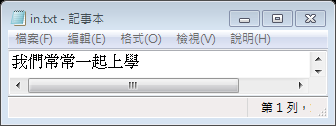
public void IOSegment (string input,

string output);

參數input是指定需要被斷詞的文字檔案名稱，參數output是指定放置斷詞與詞性標記結果的文字檔案名稱，例如：

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

指定"in.txt"檔案經過斷詞處理後，將斷詞與詞性標記結果放置"out.txt"。輸出入檔案格式如圖3，檔案模式預設是採用ANSI編碼格式，另有明確指定編碼的檔案模式請參閱附錄(二)。



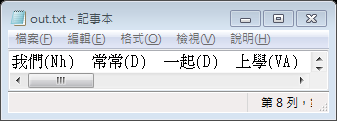


圖3 函數wecan.IOSegment()輸出入格式

第二種函數為即時斷詞函數，較適合使用者依需要在自己的程式中對單句使用，得到單句的立即反應。範例如下：

public void Segment( string input\_str,

List<string> output\_Segment,

List<string> output\_POS,

List<List<string>> segedSources)

第一個參數input\_str是需要斷詞的句子，第二個參數output\_Segment則是存放斷詞結果，第三個參數output\_POS則存放詞性標記結果，第四個參數segedSources則存放斷詞依據的字典檔來源檔名，第二及、第三個及第四個參數在經過函數wecan.Segment()處理之前，需宣告為空的變數型態，宣告範例如下：

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

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

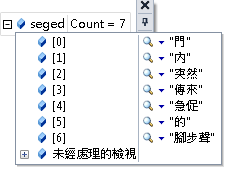
List< List<string>> segedSources =

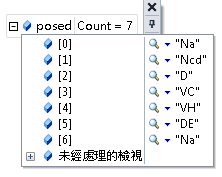
new List< List<string>> ();

宣告三個空的集合容器後再執行函數wecan.Segment()進行斷詞。例如：

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

句子「門內突然傳來急促的腳步聲」經過函數wecan.Segment()處理後，斷詞結果存放在seged，詞性標記結果存放在posed，斷詞依據的字典檔來源檔名存放在segedSources，其內容如圖4，所支援的多載請參閱附錄(二)。





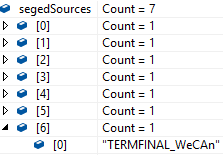


圖4 參數seged、posed、segedSources經函數wecan.Segment()之結果

※使用注意事項：無論是呼叫IOSegment或者Segment，都不應將整篇文章以單行形式餵入，否則會造成效能問題。請使用者務必在餵入大量資料之前，確保資料的分行正確。

由於斷詞的依據可能不來自字典檔，所以segedSources當中某一字詞所對應的子容器List<string>是有可能數量為0的，這表示這個字詞不存在於斷詞當下載入記憶體中的字典檔，可能是程式規則或計算模型而得到的詞，而一個字詞也可能來自於多個字典檔，例如Idiom與Cindy，而來源字串所代表的來源資訊請參照附錄(三)。

第三種是注音函數，由於字典中有包含各詞彙的注音資料，所以額外提供給使用者相關的函數做運用，範例如下：

public List<string> GetMPS(string str)

public List<List<string>> GetMPS

(List<string> str)

GetMPS()提供了兩種呼叫方式，方便使用者可以輸入單一詞彙或是一串斷詞的結果，例如：

wecan.GetMPS("注音");

List<string> list = new List<string> { "注音", "測試", "是", "一", "種", "方式" };

wecan.GetMPS(list);

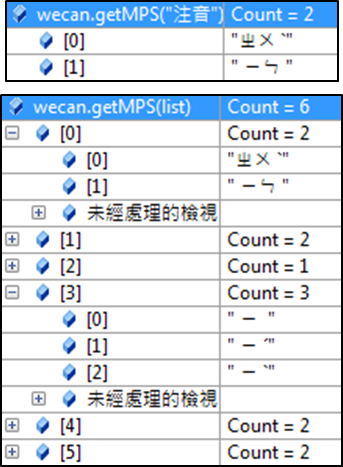


圖5 注音函數 GetMPS() 輸出格式

從圖5可以發現詞彙有幾個字就對應幾個注音的數量(Count)，但在一個字的詞彙中可能存在破音字，例如詞彙「一」就有三種注音「一 、一ˊ、一ˋ」，所以一個字的詞彙未必只回傳一個注音，而使用者需要自行決定適合此字彙的注音。

注音函數中還有提供全部詞彙對應注音的資料，由於某些使用者需要大量呼叫注音的資料，所以額外提供函數GetAllWord\_MPS()讓使用者直接取得注音資料的記憶體位置，使其呼叫的速度不受到函數呼叫而延遲，函數格式如下：

public Dictionary<string, List<string>>

GetAllWord\_MPS()

而回傳的注音格式則與函數GetMPS()回傳的格式類似，但資料是使用Dictionary的資料格式來暫存的，例如：

Dictionary<string, List<string>> MPS =

GetAllWord\_MPS();

MPS["一"];

MPS["一"]將會以List<string>格式回傳注音資料「一 、一ˊ、一ˋ」。

1. **UserConfig設定檔及參數說明**

WECAn可以由使用者決定要使用哪些字典檔作為斷詞的依據，也可以由使用者決定是否開啟或關閉系統的某些功能。上述決定的設定方式是透過修改目錄WECAnUser中的檔案UserConfig.txt達成。該檔案所含參數及設定方式可參考圖6，而各參數的意義及設定值在6.1到6.20節會詳細介紹。

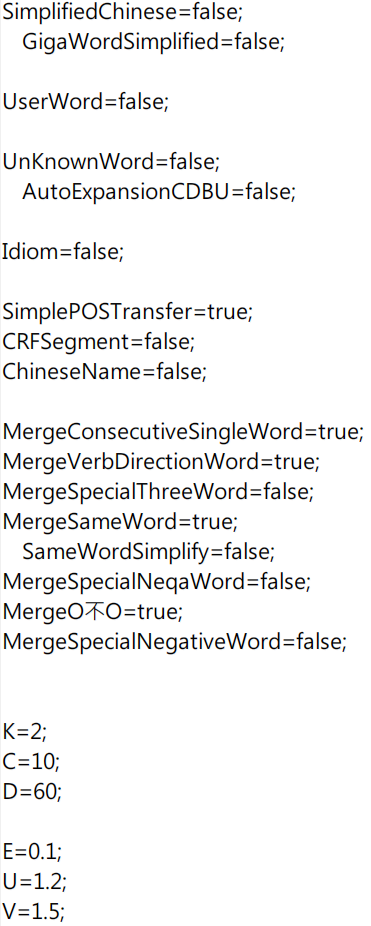


圖6 WECAnUser\UserConfig.txt示例

需特別注意的是，目錄WECAnUser以及文字檔UserConfig.txt都是固定的名稱，若更改檔名或刪除檔案將無法讀取使用者設定值，但與WECAnCorpus、WECAnAPI.dll不同的是，即便WECAnUser\UserConfig.txt未能被正確讀取到，仍然可以正常運作，系統將以預設狀態執行，預設狀態的參數值如圖6。因此使用者必須小心確認自訂的設定檔是否被正確讀取，以免誤解。

1. SimplifiedChinese

當參數SimplifiedChinese設定為true時，開啟純簡體斷詞模式，若不開啟設為false，則為純繁的形式。

1. GigaWordSimplified

GigaWordSimplified的參數用來指示是否載入語料檔GigaWord(簡)，值得說明的是，因為此資料檔為簡體專用，因此只有在參數SimplifiedChinese為true時，此參數才有效用。

1. UserWord

UserWord的參數值有true和false兩個選擇，若為true則會將目錄WECAnUser下的UserWord.txt文字檔載入字典中，WECAn所讀取的字典檔為UTF-16(Little Endian)編碼，這字典檔的功用在於能由使用者自行補充斷詞字典中不存在的詞彙，使斷詞更有擴充性。而擴充詞彙在UserWord.txt文字檔中的格式如圖7，字詞與字詞以分號(;)隔開做區隔，字詞與所擁有的欄位資訊之間以冒號(:)隔開，欄位資訊與欄位資訊之間以逗號(,)作區隔，支援”/\*”、”\*/”成對組成的區塊註解方式。使用者可以填入已知的資訊，也可以選擇不填入描述詞彙的資訊，則WECAn將會在執行時產生必須的資料。除了詞性與詞性頻率可供指定外，MPS表示注音欄位，若此欄的注音含有兩個以上的字，則每個字注音之間以全型的空白間隔開，strokes欄位表示總筆畫數。

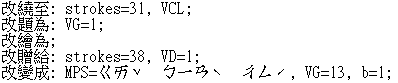


圖7 UserWord.txt檔案格式

若有不詳盡之處，可參照檔案” \WECAnCorpus\Cindy.txt”內的格式。若字詞包含分號(;)、冒號(:)、斜線(/)，須在上述字元前加上反斜線(\)來跳脫，否則會造成判讀錯誤。使用者辭典內的詞不一定會被斷在一起，因此加入6.11、6.12之強制合併、強制斷開功能。

1. UnKnownWord

UnKnownWord的參數值有true和false兩個選擇，若為true則設定為使用未知詞擷取技術修正斷詞結果。由於我們使用的未知詞擷取技術要仰賴足夠的資訊量才能擷取未知詞，建議在處理大量文章時使用此參數，可以有效處理未知詞彙。

1. UnKnownPoS

UnKnownPoS可以指定未知詞的詞性，預設b代表由WECAn自動計算合適的詞性；設定b以外則表示以指定詞性標記所有未知詞。

1. AutoExpansionCDBU

AutoExpansionCDBU的參數值有true和false兩個選擇，若為true則可以在自動偵測到未知詞時新增至6.3節所介紹的文字檔UserWord.txt，而參數UnKnownWord必須是true此參數才有作用，其新增的未知詞也必須在參數UserWord設定為true時才會載入字典作為斷詞的依據。

在新增未知詞到該文字檔前，將會搜尋該文字檔已記錄過的詞彙，若有重複將不再新增，且新增未知詞至該文字檔會受到檔案大小的限制，若UserWord.txt的檔案大小已超過16MB系統將拒絕新增當前搜尋到的未知詞，但即使檔案大小超過限制也不影響參數UserWord的運作，所以使用者在自行新增詞彙時並不受容量16MB的限制。自動新增未知詞的格式如圖8，系統會將每次(例如IOSegment()斷完一篇文章)斷詞完成後所認定的未知詞以時間區隔開(#開頭的句子為註解)。

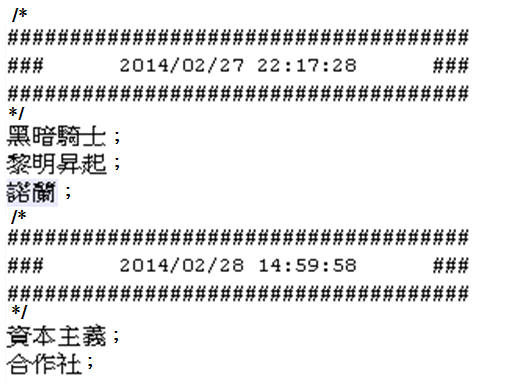


圖8 擴充未知詞至文字檔UserWord.txt之檔案格式

1. JustUnique

此參數在AutoExpansionCDBU設定為true時才有效，參數值有true和false兩個選擇，若為true則只會將彼此無父子句關係的未知詞加入UserWord.txt，假設偵測出”泛舟”、”舟哥”、”泛舟哥”，則只有”泛舟哥”會被寫入至UserWord.txt。

1. Idiom

Idiom的參數值有true和false兩個選擇，若為true則設定額外使用成語典斷詞，false則不使用。設定效果舉例如下：

關閉(false)：一(Neu)＿寸(Nf) ＿光陰(Na) ＿一(Neu) ＿寸(Nf) ＿金(Na)

開啟(true)：一寸光陰一寸金(b)

當此參數設定為false，由於系統不會載入成語典，而標準詞典中也沒有這個諺語，因此會將整句諺語拆解。反之則會將完整諺語視為單一詞。若斷詞的文件經常出現諺語或成語，例如學生作文，則建議將此參數設定為true。

1. SimplePOSTransfer

SimplePOSTransfer的參數值有true和false兩個選擇，若為true則設定輸出時詞性標記為簡易詞性標記，false為完整詞性標記。範例如下：

開啟(true)：飯前(Nd) ＿一(Neu) ＿粒(Nf)

關閉(false)：飯前(Ndc) ＿一(Neu) ＿粒(Nf)

詞彙「飯前」在字典中完整詞性為Ndc，若使用者只需要用到簡化詞性Nd，則可將此參數設定為true。

1. CRFSegment

CRFSegment 的參數值有true和false兩個選擇，若為true則設定使用CRF斷詞技術提高WordSegment斷詞的準確度。有些詞彙沒有出現在字典中，而CRFSegment是利用在訓練文章常出現的詞彙透過頻率計算來修正字典詞彙不足的問題。經過平衡語料測試，其正確率不如理論值來得高，故一般預設使用長詞優先法(MM)，建議將此參數設為false。

1. ChineseName

ChineseName 的參數值有true和false兩個選擇，若為true則使用自動擷取華人姓名技術修正斷詞結果。範例如下:

開啟(true)：蔣偉寧(Nb)＿擔任(VG)＿部長(Na)

關閉(false)：蔣(Nb)＿偉(b)＿寧(D)＿擔任(VG)＿部長(Na)

若將該參數設定為true，則華人姓名會被視為一個詞彙，如第一句的「蔣偉寧」。將該參數設定為false則如第二句。需要注意的是，WECAn同一篇文章最大人名上限是1000個，超過該數目將不再偵測新人名。由於此功能之正確率因輸入文件的特性不同而有所差異，雖然測試時其偵測率超過80%，但仍有一定機率誤判。通常若欲處理的資料為幾乎沒有人名作文類型，建議設為false；若資料為新聞類型，可將其設為true。

1. ForceMergeWord

若將該參數設定為true，在順逆向以長詞斷詞法斷詞後，如果字詞組合會出現在WECAnUser\ForceMergeWord.txt裡面，則將之合併，例如自定義字典有”很快結束”這個紀錄，則”將會很快結束。”經順向斷詞後為”將”、”會”、”很”、”快”、”結束”、”。”，再合併成”將”、”會”、”很快結束”、”。”，才進行詞性標記。強制合併自定義清單之格式為字詞接上分號表示一筆紀錄，若詞內有換行字元、空格、縮排字元(tab)都會被忽略。若字詞包含分號(;)、斜線(/)，須在上述字元前加上反斜線(\)來跳脫，否則會造成判讀錯誤。範例請見WECAnUser\ForceMergeWord.txt。

1. ForceSplitWord

若將該參數設定為true，在順逆向以長詞斷詞法斷詞後如果字詞出現在WECAnUser\ForceSplitWord.txt裡面，則將之依使用者定義拆開，例如自定義字典有”電路+板”這個紀錄，則”一塊電路板。”經順向斷詞後為” 一塊”、”電路板”、”。”，再拆開成” 一塊”、”電路”、”板”、”。”，才進行詞性標記。強制分開自定義清單格式同強制合併自定義清單，字詞強制斷開處以”+”號表示即可，若字詞包含分號(;)、加號(+)、斜線(/)，須在上述字元前加上反斜線(\)來跳脫，否則會造成判讀錯誤。範例請見WECAnUser\ForceSplitWord.txt。

1. ForceMergeNeuAndFW

當參數設為true時，會將順逆向以長詞斷詞法斷詞後出現的連續英數字(包含ASCII字元32~126)合併，例如”通上20A的電流”經順向斷詞後為”通”、”上”、”2”、”0”、”A”、”的”、”電流”、”。”，再合併為”通”、”上”、”20A”、”的”、”電流”、”。”，才進行詞性標記。

1. SeperateInSpace

在參數ForceMergeNeuAndFW設為true的條件下，此參數才會做用。當參數設為true時，強制合併連續英數不會包含空白，即在空白處斷為兩個詞，例如”Intel i7處理器” 經連續英數合併後為”Intel”、” ”、”i7”、”處理器”，才進行詞性標記。。

1. MixTwCn

當參數為true時，句子會逐字轉換為當前使用語言後進行斷詞及詞性標記，兩步驟完成後再轉換回原句子繼續之後流程，例如”出国期间院务由罗副院長代行。”會轉為”出國期間院務由羅副院長代行。”後進行斷詞及詞性標記。(因轉換為每句每字，因此會增加些許效能負擔，使用者須注意)。

1. MergeConsecutiveSingleWord

當參數設為true時，將會合併連續相同且長度為一的字，例如"嘿"、"嘿"、"嘿"會合成"嘿嘿嘿"。

1. MergeVerbDirectionWord

當參數設為true時，將會合併由動詞與方向詞組成的三字組，例如"放"、"回去"會合成"放回去"。

1. MergeSpecialThreeWord

當參數設為true時，將會合併組合後前兩字相同的三字組，例如"搖"、"搖頭"會合成"搖搖頭"。

1. MergeSameWord

當參數MergeSameWord設定為true時，將開啟重複字詞合併功能，例如"好棒"、"好棒"會合成"好棒好棒"。

1. SameWordSimplify

在參數MergeSameWord設為true的條件下，此參數才會做用。設為true時，兩組相同二字組將會減為一組，例如"好棒好棒"將會變成"好棒"。由於此功能會刪減輸出的詞彙，所以若非特殊需求，強烈建議設為false。

1. MergeSpecialNeqaWord

當參數設為true時，將會合併由Neu與後面為Nf或Neqa所組成的三字組，例如"一"、"顆顆"將合成"一顆顆"。

1. MergeO不O

當參數設為true時，將會合併”O不O”類的字詞，其中O為人工篩選的特定字詞，例如”看”、”不到”會合併成”看不到”。

1. MergeSpecialNegativeWord

當參數設為true時，出現以下情況將自動合併。

“不”+[任意詞]

例如：不能、不開心

[任意單字詞]+”不”+[任意單字詞]

例如：好不好、能不能

此功能若非特殊需求，強烈建議設為false。

1. 參數K、C、D、E、U、V

當參數UnKnownWord設定為true時，參數K、C、D、E、U、V對應下列條件式(7)中提及的。這6項參數會影響未知詞擷取的正確率，其參數值設定所造成的影響可以參考[2]。若要篩選某詞彙組合為未知詞，必須滿足[2]中提及的4種條件：

其中參數決定詞彙的最少字數，若設定為3就不會擷取兩個字的未知詞；參數為未知詞出現的頻率門檻，數值太低，將會認定過多非未知詞，數值高雖然較精確，但也較難篩選出低頻未知詞；參數與SPLR[2]與父子序列()[2]相關，詳細的設定方式應參考[2]。

1. **討論**

使用本工具前請先確實了解自身需求進行妥切的設定，否則斷詞效果可能不如預期。根據目前我們針對平衡語料庫的測試結果，斷詞正確率介於0.95至0.96 之間(包含標點符號)，詞性標記正確率達到0.95。

**參考文獻**

1. Chang, T. H., Sung, Y. T., and Lee, Y. T. 2012. A Chinese word segmentation and POS tagging system for readability research. In Proceedings of the 42nd Annual Meeting of the Society for Computers in Psychology. (Minnesota, USA, 2012.
2. Chang, T. H. and Lee, C. H. 2003. Automatic Chinese unknown word extraction using small-corpus-based method. In Proceedings of International Conference on Natural Language Processing and Knowledge Engineering. (Beijing, China, 2003), 459-464.
3. 詞庫小組，1993，中文詞類分析，中文詞知識庫小組技術報告 # 93-05，南港，中央研究院

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[3]

|  |  |  |
| --- | --- | --- |
| 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\*/ |
| 簡化標記 | 詳細標記 | 詞類標記 |
| 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\*/ |

附錄(二)WECAn供使用者呼叫的方法一覽

斷詞與詞性標註主要方法一覽

public void IOSegment(string inputFilePath, string outputFilePath)

//等同ANSI編碼的IOSegment\_ANSI

public void IOSegment\_ANSI(string inputFilePath, string outputFilePath)

//使用ANSI編碼檔案輸出入模式

public void IOSegment\_UTF8(string inputFilePath, string outputFilePath)

//使用UTF8編碼檔案輸出入模式

public void IOSegment\_Unicode(string inputFilePath, string outputFilePath)

//使用UTF-16(Little Endian)編碼檔案輸出入模式

public void Segment(List<string> input\_Segment, List<string> output\_POS)

//傳入斷詞好的句子去計算對應的詞性

public void Segment(string sentence, List<string> output\_Segment)

//輸入句子得到斷詞結果

public void Segment(string sentence, List<string> output\_Segment, List<List<string>> segedSources)

//輸入句子得到斷詞結果以及依據的斷詞語料來源

public void Segment(string sentence, List<string> output\_Segment, List<string> output\_POS)

//輸入句子得到斷詞結果以及對應的詞性結果

public void Segment(string sentence, List<string> output\_Segment, List<string> output\_POS, List<List<string>> segedSources)

//輸入句子得到斷詞結果、對應結果與依據的語料來源

public string GetCourpusPath() //取得當前語料路徑

public Dictionary<string, List<string>> GetAllWordMPS() //得到所有詞彙的注音集合

public List<List<string>> GetMPS(List<string> str) //查詢指定句子各詞彙的注音

public List<string> GetMPS(string str) //查詢輸入詞彙的注音

public List<string> GetAllWord(Set set) //取得當前使用所有字詞清單

public bool IsWord(Set userSet, string word) //根據設定檔，查詢word是否為一字詞

public bool IsUnicodeBasicMultilingualPlaneChineseChar(char \_char)

//檢查字元是否符合基礎多語言平面(BMP)字符定義範圍

public void ClearNameCache() //清除未知人名偵測快取

可直接對記憶體內設定值操作的方法一覽

public void SetSimplifiedChinese(bool enable)

public void SetGigaWordSimplified(bool enable)

public void SetUserWord(bool enable)

public void SetUnKnownWord(bool enable)

public void SetUnKnownPoS(string pos)

public void SetAutoExpansionCDBU(bool enable)

public void SetIdiom(bool enable)

public void SetSimplePOSTransfer(bool enable)

public void SetCRFSegment(bool enable)

public void SetChineseName(bool enable)

public void SetForceMergeWord(bool enable)

public void SetForceSplitWord(bool enable)

public void SetForceMergeNeuAndFW(bool enable)

public void SetSeperateInSpace(bool enable)

public void SetMergeConsecutiveSingleWord(bool enable)

public void SetMergeVerbDirectionWord(bool enable)

public void SetMergeSpecialThreeWord(bool enable)

public void SetMergeSameWord(bool enable)

public void SetSameWordSimplify(bool enable)

public void SetMergeSpecialNeqaWord(bool enable)

public void SetMergeO不O (bool enable)

public void SetMergeSpecialNegativeWord(bool enable)

public void SetK(int valueK)

public void SetC(int valueC)

public void SetD(int valueD)

public void SetE(double valueE)

public void SetU(double valueU)

public void SetV(double valueV)

附錄(三) segedSources內含字串所代表的來源意義

"UserWord"：

來自使用者所定義的詞彙 (\WECAnUser\UserWord.txt)

"WECAn-Nb"：

由WECAn未知人名偵測模組字詞

"Cindy"：

由Cindy2.1語料庫所生成的詞彙 (\WECAnCorpus\Cindy.txt)

"Simon"：

由國家漢語委員資料庫以及北京大學現代漢語樹庫所生成的詞彙(\WECAnCorpus\Simon.txt)

"IdiomTw"：

成語與諺語集的詞彙 (\WECAnCorpus\IdiomTw.txt)

"IdiomCn"：

成語與諺語檔由office word轉換為簡體的詞彙 (\WECAnCorpus\IdiomCn.txt)

"GigaWordCn"：

GigaWord(簡)語料庫所生成的詞彙 (\WECAnCorpus\GigaWordCn.txt)