DOI: 10.6245/JLIS.201804_44(1).0002

中文情感語意分析套件 CSentiPackage 發展與應用

陳韋帆
Bauhaus-Universität Weimar 博士生
E-mail: wei-fan.chen@uni-weimar.de

古倫維
中央研究院資訊科學所助研究員
E-mail: lwku@iis.sinica.edu.tw

關鍵詞：中文處理工具；意見分析；情感分析；社群媒體資料分析

【摘要】

近年來，意見與情感分析技術漸漸為大家所注意且因網路環境之餘意見發表的便利性，文本意見與情感分析技術的需求與應用也快速增加，然而在中文上，並沒有公開可利用的意見與情感分析工具。本文介紹一個供研究使用免費且公開的中文情感語意分析套件：CSentiPackage，套件中包含多個可以用於中文情感語意分析研究所需要的各式工具，例如中文意見、構詞詞典、中文意見樹庫、意見挖掘計分工具，及深度社群立場分析模型等。本文將詳細介紹各個工具的內容。欲取得 CSentiPackage 套件及其使用方法，可從網址 http://www.lunweiku.com/ 之 Lab Homepage 分頁進入申請下載。

緒論

意見與情感分析旨在探勘目標文本所表達的意見或情感及其程度。意見一般指的是某個發言者對某個目標的看法，可能是正面、負面、中立，以立場來說則是支持或反對；情感則是指人類感受到的情感，例如快樂、傷心、憤怒、厭惡、害怕等。意見與情感分析包含所有找出相關資訊並加以應用的技術。

意見與情感分析的研究，至今已約莫近十五年。從早期在意見分析上，只是做正面與負面文章的二元分類，中期技術進展到進入句子與詞彙層次，並且能夠分析情感能量的強度，近期更與社群媒體平台結合，與使用者資料共同運作，並引入了最新的深度學習技術，大大提高了技術效能，使其應用的可能性更高，應用的範圍也更廣。

意見與情感分析的技術，早期多是採用統計式或是簡單的機器學習模型。因為意見與情感分析牽涉到較為複雜且深入的語意，若未深入考慮到語意層次，即使使用較複雜的學習模型，效果未必能夠優於簡單的規則模型或關鍵字對應（keyword matching）模型。與自然語言處理的技術發展模式雷同，意見與情感分析的技術，也是分成統計式與機器學習兩線來發展，同時間混合型的模型也利用這兩線發展出的最新技術，疊加出更好的效能。近五年來，由於社群媒體的興起及網路的普及，一般人開始在網路上抒發各式各樣的意見，包括部落格的生活分享、開箱文、臉書粉絲團、論壇及商場的產品評價文，其他人也開始在網路上搜尋可靠的意見分享文，做為他們平常生活的購買行為與議題看法的參考。在這樣的環境下，意見與情感分析技術的重要性快速升溫並受到矚目，加上這兩年深度學習的技術帶動了技術的精進，使得意見與情感分析的效能更趨成熟，並能夠應用在各領域的文件分析工作中。

如上所述，意見與情感分析技術可用於各式文本的分析，最常見的就是產品評論的分析。意見分析的困難之處，除了需要語意上的深入理解之外，由於它本質上屬於監督式的問題，因此如果採用機器學習的方法，就需要正確標上意見或情感資訊的大量資料，才能夠訓練出效能好的模型。剛好產品評論資料上使用者所給的星級滿意程度（一般為一到五顆星），大量且快速地提供了研究所需的標準答案資訊，也就順勢成為意見與情感分析領域研究人員最常討論開發與測試技術的文類。另外一個在此領域蓬勃發展的則是輿情分析，由於關心選舉及公共事務的公民，經常於網路發表正反意見及各式討論且此類網路平台也逐漸增加，吸引了不同族群的使用者，雖然這類主題的文章並沒有平台相關功能的支援來得到標準答案，但由於應用端的需求，也促成許多研究人員投入人工標記資料的工作。

意見與情感分析也被用於許多資訊科學以外的不同領域，只要是文本的分析，牽涉到意見與情感的部分都可能可以應用。例如，在醫學方面，可應用此技術分析憂鬱症病患的手稿，偵測並自動警示病患可能的自殺傾向；在心理學治療方面，可藉由分析兩方交談的內容，自動加以介入做適當的調解；在政治學方面，可分析兩國間的相關新聞及評論，以得知兩國關係發展的參考資訊；在商業上分析出商品的優劣及優缺點可用於定價、廣告與行銷等等，顯見此技術也具有相當高的跨領域能力。

相關資源探討

意見與情感分析的資源包含語料庫、詞典，還有偏向應用層面、針對特定產品的評論等。最有名的是史丹佛大學開發的情感樹庫（Stanford Sentiment Treebank）〔1〕，目前已經整合進史丹佛大學的自然語言核心工具（Stanford CoreNLP）中，其相關程式碼也可自由下載；談到資源，很早就有哈佛大學建置的 General Inquirer Lexicon〔2〕，其包含出現在哈佛字典或

Lasswell 字典中一共 11,178 個字，每個字標註不只一個相關特性，例如正面、負面、主動、被動等；接著有仿照詞網（WordNet）擴展出來的 WordNet Affect（WN-Affect）〔3〕和 SentiWordNet〔4〕，前者將 WordNet 的每個同義詞集賦予一至多個情緒標籤，後者則是依同義詞集的正負面、主觀性等給予三種情緒分數，比起前者多了強度的資訊；EmoLex〔5〕則是加拿大國家研究院開發的情緒詞典，包含 14,182 個字，除了正負面以外，還有八種情緒四種強度的標註。另外，SenticNet〔6〕總共有一萬七千個概念，每個概念標註四個不同的情緒分數，並有綜合四個情緒分數產生之情緒極性。

應用方面，康乃爾大學的電影評論資料〔7〕是一被廣泛使用的資料庫，其包含情感極性資料庫、情感量表資料庫和主觀性資料庫，多用於情感分類的任務。史丹佛大學開發的 Large Movie Review Data〔8〕則適用於正面、負面之二元情感分析，比起前者擁有更大量的評論資料。德國 Darmstadt 服務評論語料庫〔9〕蒐集許多消費者的評論，並標註每個句子及整體評論的意見和情緒資訊。除了產品或電影評論之外，多相問答意見語料庫（MPQA）〔10〕則擁有豐富的新聞資料，並標註新聞中的事件、意見和情緒等其他狀態，用於問答系統之中。

意見與情感分析技術已經發展十數年，也受到學界與產業界的極大重視，意見與情感分析技術之應用性，使得具有資源的英文環境中，如產品評論、政府施政輿情分析等等應用，技術強度都大幅度增加，開發相關技術的新創公司在矽谷也如雨後春筍般出現並獲得成功。然而，相較於豐富的英文資源，中文資源顯得相當匱乏。在中文上雖然中國與台灣都各自有許多相關的技術發表於研究領域，卻幾乎沒有語料庫、資源與免費程式方便研究社群投入更大量的研究。中文資源的短絀，使得中文業界相關技術應用普遍低落，高端技術也掌握在少數特定公司手中，實在相當可惜。

有鑑於上述中文環境中情感與意見分析技術發展的困境，我們開發了中文情感語意分析套件（CSentiPackage，Chinese Sentiment Package）並開放供研究目的的自由免費下載。中文情感語意分析套件，是一個處理中文的語意分析領域可使用的套件。其中包含多個資料集、語意字典及語意分析工具。其中資料集的部分有中文構詞資料集（Chinese Morphological Dataset）及中文意見樹庫（Chinese Opinion Treebank）；語意字典則包括了台灣大學意見詞詞典（NTUSD）以及增廣意見詞詞典（ANTUSD）；語意分析工具則有用於中文意見分析的語意挖掘計分工具（CopeOpi）及一個深度社群立場分析模型（UTCNN）。以下將分段說明這些工具的內容及用法。

語意分析資料集

語意分析資料集可作為驗證實驗的資料集，本套件中包含了兩種不同的資料集，可供研究者在語意分析領域中作為實驗的素材。

中文構詞資料集（Chinese Morphological Dataset）

中文構詞資料集包含了超過八千個詞，是 ACBiMa（Advanced Chinese Bi-Character Word Morphological Analyzer，包含超過 11, 000 個詞）的前身，標記了詞的組合方式，包括有八種（Huang, T. H., Ku, L. W., & Chen, H. H., 2010；Ku, L. W., Huang, Ting-Hao (Kenneth) & Chen, H. H., 2010）。

1. 並列聯合（Parallel）。表示兩個字語意相當，或是可以對比，例如：財富、打罵、男女。另外如人人、謝謝等疊字詞也屬於這個類別。
2. 修飾，偏正（Substantive-Modifier）。表示前字是後字的修飾，例如：低級、痛哭。兩字都是名詞的情況也有可能屬於這類，例如衣櫃，其中衣修飾了櫃。
3. 主謂（Subjective-Predicate）。類似於一個句子當中的主詞和動詞結構，例如：心疼、氣虛。
4. 動賓，述賓（Verb-Object）。第一個字通常是對第二個字有影響的動詞，結構類似於一個動詞和一個受詞。例如：失控、免職。
5. 動補，述補（Verb-Complement）。第一個字一般而言是動詞而也有可能是形容詞，而第二個字用來解釋第一個字的不同情況。例如：看清、擊潰。值得注意的是此類是後修飾結構，而第二類則是前修飾。現代中文的雙字詞中，後修飾結構比前修飾要少的很多，而且大部分的後修飾結構都是動補。
6. 否定（Negation）。第一個字是否定字，例如：非、不、否、無。
7. 肯定（Confirmation）。第一個字是肯定字，例如：有。
8. 其他。不屬於上面七種的詞，也包括所謂的連綿詞，例如：披薩、阿媽、牛仔。

中文意見樹庫（Chinese Opinion Treebank）

中文意見分析樹庫於中文樹庫 5.1（Chinese Treebank）的資料上，額外標記了語意分析的資料。因此在使用上也需要搭配中文樹庫 （可於 Linguistic Data Consortium, LDC〔11〕取得）來使用，本套件中只包含了標記資料，而不包括原本中文樹庫中的文字及剖析樹等內容。其標記介面可參照圖 1，其中每個句子標記了：1）是否有意見（是／否），2）若有，其極性為何（正面／中性／反面），3）句子的類型（表述/狀態/動作）。同時，對於具有意見的句子，也標記了句中詞與詞的關係。

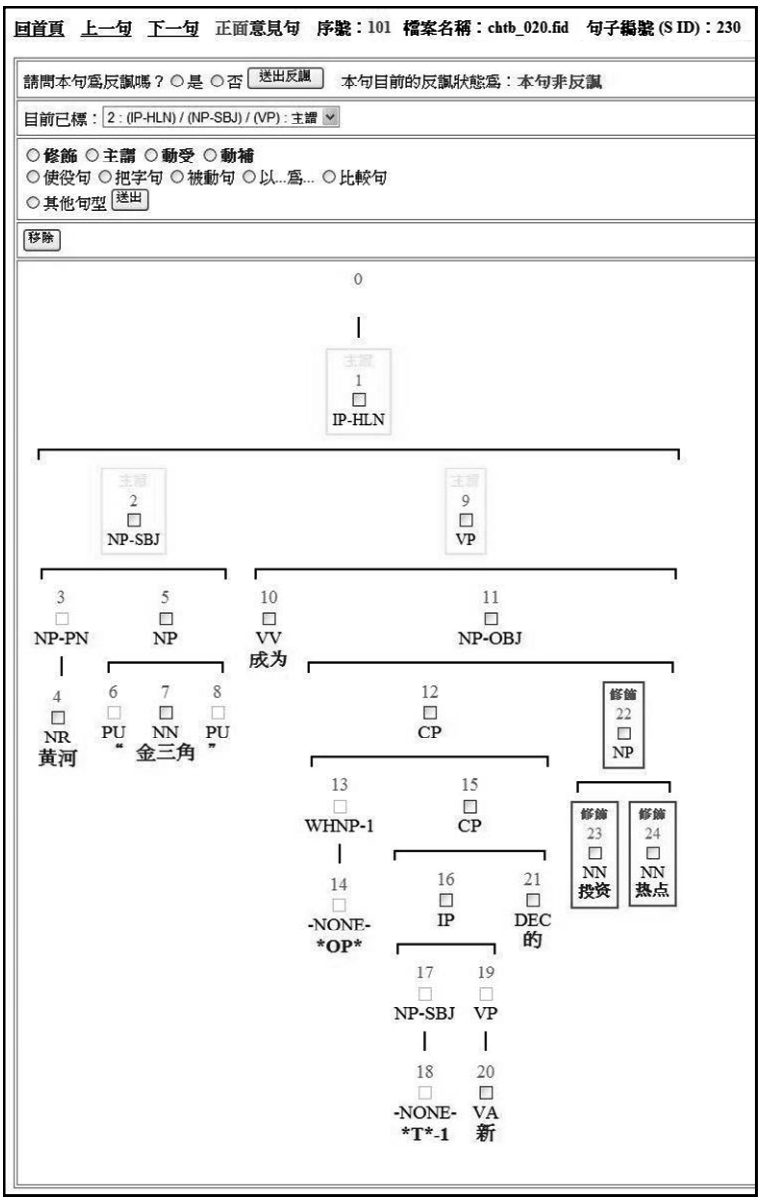

圖 1　**中文意見分析樹庫的標記介面**

詞與詞的詞間關係參照該句子的剖析樹加以標記，範例如圖 2。關係標記於一個三元結構（OLeft, ORight, t）之上，而標記的方法是以一個五維向量紀錄此三元結構的資訊，如（triID, OParent, OLeft, ORight, t），其中 triID 是該五元向量的序號；OParent 是三元結構的父節點；OLeft 和 ORight 則分別為左節點和右節點；t 則是此兩節點的關係，包括了修飾關係（Substantive-Modifier），主詞-謂詞（Subjective-Predicate），動詞-受詞（Verb-Object），動詞-補語（Verb-Complement）以及其他（Other）共五種。表 1 列出了幾個三元結構及其對應的五元向量的例子。

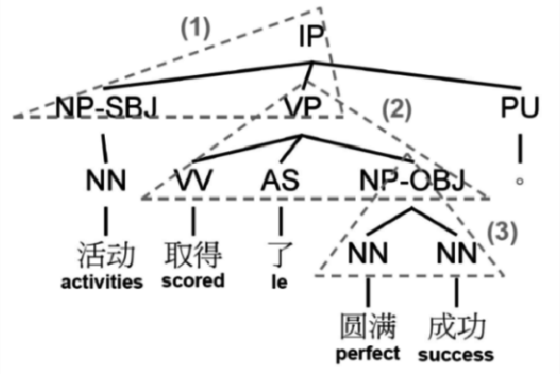

圖 2　一個剖析樹和其中包含的詞間關係標記

註：在本例中，包含了三個關係，分別為（1, IP, 活動, VP, Subjective-Predicate），（2, VP, 取得, NP-OBJ, Verb-Object）和（3, NP-OBJ, 圓滿, 成功, Substantive-Modifier）

表 1　**詞與詞關係的範例**

類型	範例
修飾關係（Substantive-Modifier）	高大的樓房
主詞-謂詞（Subjective-Predicate）	學習認真
動詞-受詞（Verb-Object）	恢復疲勞
動詞-補語（Verb-Complement）	收拾乾淨

情感語意字典

語意字典是語意分析的基礎工具之一，目前中文情感語意處理的工具仍相對缺乏，本套件中包含的兩個情感語意字典，台大意見詞詞典（NTUSD）以及增廣意見詞詞典（ANTUSD），是少數目前包含中文詞語的字典。

台大意見詞詞典

台大意見詞詞典（NTUSD，*National Taiwan University Sentiment Dictionary*）〔3〕是 2006 年公開的中文意見詞詞典，於學術用途可免費使用。詞典包括了簡體及繁體中文，共有 11,088 個語意詞，其中有 2,812 個正面詞及 8,276 個負面詞。詞字典中只標註了正面或是負面，並沒有強度的區分或是其他額外資訊。本套件中所包含的詞典內容如圖 3。

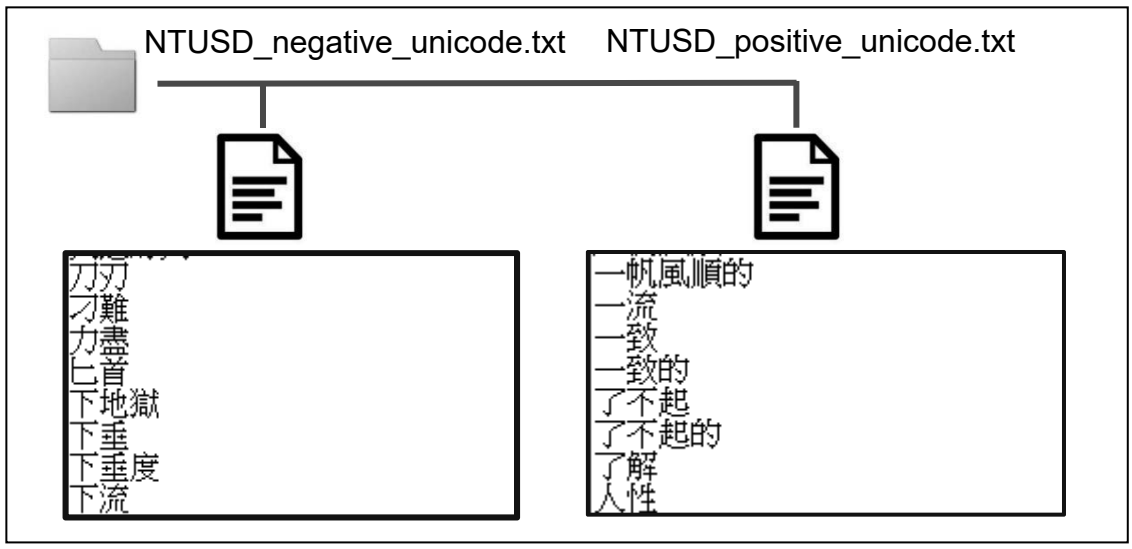

圖 3　**台大意見詞詞典的結構**

註：文件中包含兩個純文字文件，分別表列了負面詞和正面詞。

增廣意見詞詞典

增廣意見詞詞典（ANTUSD）〔4〕於 2016 年公開，是上述字典的增進版，並添加了更多的詞語來到 27,221 個。此字典有五種類別，分別為正面、中立、負面、非意見詞以及否定詞，其統計資料可見表 2。

表 2　**增廣意見詞詞典的統計資料**

類型	數量
正面	9,382
中立	16
負面	11,224
非意見詞	5,415
否定詞	612

對於每個詞，標註了六個欄位，分別是：CopeOpi 分數，這是由語意分析工具 CopeOpi 自動估計出來的語意分數、正面標註的數量、中立標註的數量、負面標註的數量、非意見詞標註的數量、非詞標註的數量（此標記對於處理自動斷詞的文本很有幫助），範例可見表 3。

表 3　**增廣意見詞詞典的標記範例**

	語意分數	正面	中立	負面	非意見詞	非詞
開心	0.434168	1	0	0	0	0
酣聲	0	0	0	1	3	0
憤怒	-0.80011	0	0	5	0	0

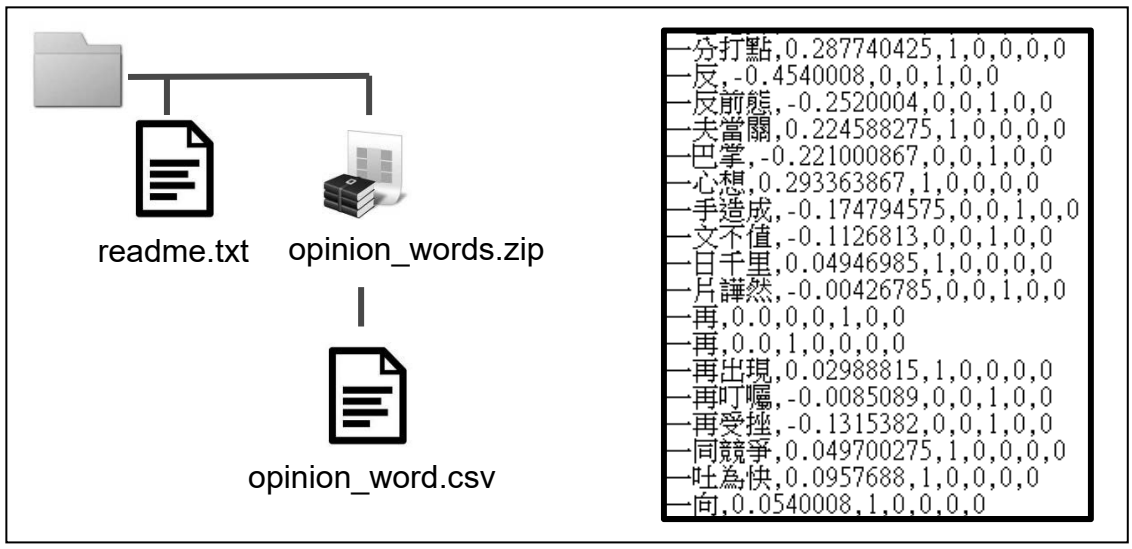

圖 4　**增廣意見詞詞典的結構**

註：文件中包含一個說明檔，說明各個欄位的意義。以及一個壓縮檔，內含一個 csv 格式的純文字文件，記錄了每個詞於各個欄位的值。

語意分析工具

目前已有許多用於語意分析的演算法或是工具，本套件中提供了兩個，一個是非監督式的，另一個則是監督式的，兩者都已經被驗證可用於語意分析的問題上且具有優秀的表現，實際測試的效能及與其他模型的比較可以參考相關論文（Chen, W. F., Ku, L. W., & Lee, Y. H., 2015；Chen, W. F. and Ku, L. W., 2016）。

CopeOpi

CopeOpi（Ku, L. W., Ho, H. W., & Chen, H. H., 2009）是一個統計性的語意分析工具，可給予一個句子或是一篇文章語意分數，此語意分數可以用來判斷這段文字的語意（正面或負面）以及強弱。相較於其他工具而言，CopeOpi 的優點在於它無須依靠訓練資料就可以單獨運作，而且執行迅速，是一個使用上相當便利的非監督式工具。除此之外使用者也可以自由地加入更多的語意字典來增強或是調適於不同領域的文章（目前使用的字典是前面提到的增廣意見詞詞典）。

CopeOpi 的基本概念來自於中文的構詞，考慮到每個字本身均帶有一定的語意訊息。例如，德、勝、高適合作為人名，而笨、悲、慘則顯然不適合。儘管可能有少數例外（如徐悲鴻），但我們能從大量的文本之中訓練出這些語意資訊：

如式（1），N_{c_i} 表示這個字 c_i 的負面分數。它是由該字出現於負面詞彙的比例正規化而來，其中 fn_{c_i} 為 c_i 這個字在負面詞彙出現的頻率，而 fp_{c_i} 則為 c_i 這個字在正面詞彙出現的頻率。同樣的，藉由將字出現於正面詞彙的比例正規化，我們也可以計算出他的正面分數 P_{c_i} 如式（2）。

$$N_{c_i} = \frac{fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}}{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j} + fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}} \quad (1)$$

$$P_{c_i} = \frac{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j}}{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j} + fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}} \quad (2)$$

接著，這個字的意向分數S_{c_i}就是正面分數減去負面份數，如式（3）。

$$S_{c_i} = (P_{c_i} - N_{c_i}) \quad (3)$$

一個詞的意向分數S_w就是這個詞中包含的字的意向分數的平均，如式（4）。

$$S_w = \frac{1}{p} \times \sum_{j=1}^{p} S_{c_j} \quad (4)$$

若更進一步考慮到詞內或是詞與詞間的結構，我們可以參考這些組成的方法來計算出詞分數或是句分數。詞內結構方面，在中文構詞資料集中有八種構詞的方法，我們可以根據不同的構詞方法來組合出詞分數，分別是：

(1) 並列，聯合（Parallel）。在此類中兩字的意思相當，因此詞分數為兩字的分數平均。

(2) 修飾，偏正（Substantive-Modifier）。由於前字是後字的修飾，故計算如式（5）。其中 S（C）為字 C 的意向分數，（參考式（3）），而 C1C2 則表示一個二字詞包含字 C1 與 C2。當兩字皆有分數時，若兩者皆正，則只取前字的分數（意即，只取修飾詞的分數），否則取前字的意向強度加上負號（換言之，兩字意向相反時必為負，而只取修飾詞的強度來用）；在只有其中一個字有分數的情況下，則直接使用該字的分數。

$$\begin{aligned} &\text{if } (S(C_1) \neq 0 \text{ and } S(C_2) \neq 0) \text{ then} \\ &\quad \text{if } (S(C_1) > 0 \text{ and } S(C_2) > 0) \text{ then } S(C_1C_2) = S(C_1) \\ &\quad \text{else } S(C_1C_2) = -1 \times |S(C_1)| \\ &\text{else } S(C_1C_2) = S(C_1) + S(C_2) \end{aligned} \quad (5)$$

(3) 主謂（Subjective-Predicate）。此類組成方式為主詞加上動詞，分數計算如式（6）。其中，當動詞（第二字）有分數時就只考慮動詞，反之則只考慮主詞。

$$\begin{aligned} &\text{if } (S(C_2) \neq 0) \text{ then } S(C_1C_2) = S(C_2) \\ &\text{else } S(C_1C_2) = S(C_1) \end{aligned} \quad (6)$$

(4) 動賓，述賓（Verb-Object）。此類組成方式為動詞加上受詞，計算方式如式（7）。其中，當兩字皆有分數時，詞分數是動詞強度以及兩字的意向相乘（負負得正，正正得正，正負得負）；只有一字有分數的話，那麼就採用該字的分數。

$$\begin{aligned}&\text{if } (S(C_1)\neq 0 \text{ and } S(C_2)\neq 0)\\&\quad\text{then } S(C_1C_2)=|S(C_1)|\times SIGN(S(C_1))\times SIGN(S(C_2))\\&\quad\text{else } S(C_1C_2)=S(C_1)+S(C_2)\end{aligned} \tag{7}$$

(5) 動補，述補（Verb-Complement）。由於結構類似，計算方式如同第（3）主謂類。

(6) 否定（Negation）。若第一個字是否定字（如「不」、「非」等），那就取第二字的分數加上負號；反之則取第一字的分數加負號。

$$\begin{aligned}&\text{if } (C_1\in NC) \text{ then } S(C_1C_2)=(-1)\times S(C_2)\\&\text{else } S(C_1C_2)=(-1)\times S(C_1)\end{aligned} \tag{9}$$

(7) 肯定（Confirmation）。若第一個字是肯定字（如「是」、「有」等），則就取第二字的分數；反之則取第一字的分數。

$$\text{if } (C_1\in PC) \text{ then } S(C_1C_2)=S(C_2) \text{ else } S(C_1C_2)=S(C_1) \tag{10}$$

(8) 其他。由於無法歸類結構資訊，故同第（1）類並列一樣單純將兩字的分數做平均。

同樣的概念亦可利用在詞與詞的結構（詞間結構）上，我們可利用中文意見樹庫的三元標記，套用計算式來得出句子的分數，以進一步的獲得更好的句意見分數。使用結構資訊約可提升正確率 1-2%，目前的 CopeOpi 分析工具未使用結構資訊，若希望整合使用者已有的詞內結構資訊以提升正確率，可來信與作者連繫；詞間結構資訊的利用較為複雜，目前具有此功能的版本尚未提供自由下載使用。一個輸入句的格式範例如下：

支持（VC）核能（Na），（COMMACATEGORY）支持（VC）核四（Nc），（COMMACATEGORY）享受（VJ）相對（VH）便宜（VH）的（DE）電價（Na）。（PERIODCATEGORY）

經過 CopeOpi 語意分析工具後的輸出範例如下：

支持/0.03811470000000001 核能/0.0 ，/0.0 支持/0.03811470000000001 核四/0 ，/0.0 享受/0.034075599999999984 相對/-0.042713250000000064 便宜/-0.3732806 的/0.0 電價/0.0 。/0.0

***Score=0.06759174999999995

UTCNN

UTCNN（Chen & Ku, 2016）是一個用於處理社群網路上語意分析的深度學習模型。此模型的目標在於能夠有效整合社群網路上多種不同的資訊，並利用深度學習來判斷一篇文章

的語意。以目前台灣廣泛使用的 Facebook 為例，當中可利用的資訊除了文章的內容之外，還包括了作者、按讚者、文章主題、對本文回覆的內容，發表回覆者等資訊。這些資訊都可以提供更多的線索來判斷一篇文章的語意，例如，有許多篇文章被同一位使用者按讚，可能代表這些文章具有類似的語意；或是從回覆中看到回覆者的是支持某個意見，且他反對本文的內容，也可能可以推斷本文的語意。

首先，一般應用深度學習來進行語意分析的模型，多是先由詞向量開始，堆疊之後將一篇文章表示成一個向量，例如：圖 5。在此卷積類神經網路（convolutional neural network）中，將每個詞向量（word embedding）排列起來，接著使用不同大小的卷積層，以獲得不同長度的語意（例如在本模型中使用長度為 1、2、3，那麼就可以來捕捉單字、雙字、三字的語意）。接著再經過若干的選擇層（pooling layer）來堆疊不同語意的資訊，就可以獲得文章的向量。

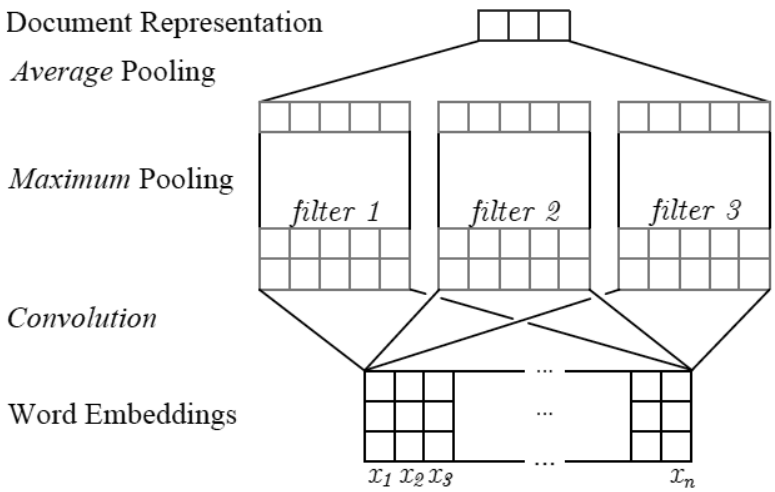

圖 5　**常見的卷積類神經網路**

接著我們希望可以考慮到使用者（包含作者及按讚者）和主題兩個資訊，來建立文章向量。如圖 6，在原本的詞向量上，額外增加一個語意轉換層，可用來捕捉不同使用者的語意偏好、或是用字差異，也用來捕捉不同主題的文字差異。獲得此新的轉換向量後，就視作一般的詞向量進行卷積類神經網路的操作。

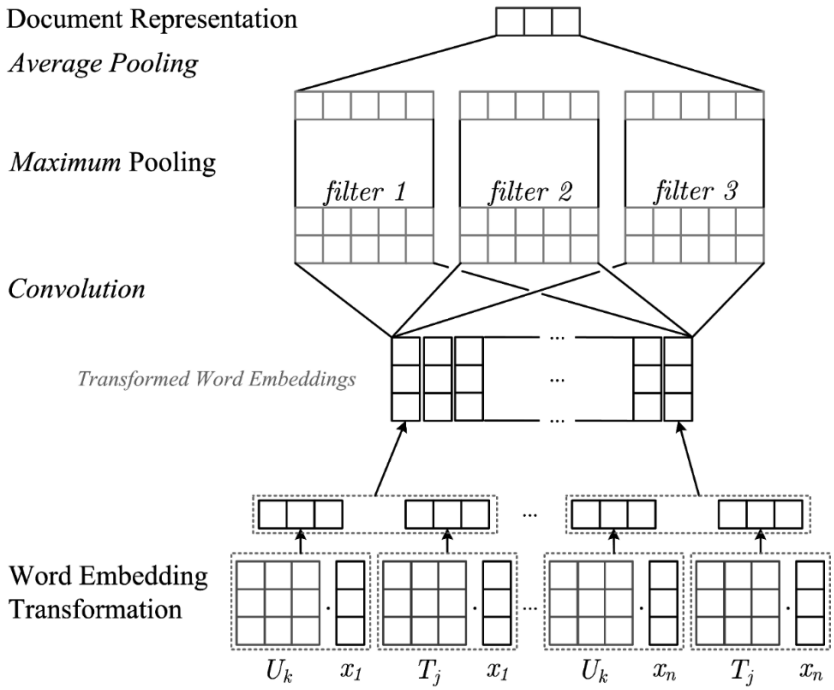

圖 6　**考慮使用者和主題文意之後的卷積類神經網路**

同時，我們也希望考慮不同使用者或是主題可能具有的不同語意偏好。例如，一個使用者發表的文章可能經常是正面的，或是某類主題經常是正面性的。這與上一個轉換層的不同在於，前者旨在捕捉文意上的差異，而這裡考慮的則是使用者與主題對於語意的偏好。如圖 7，在經過圖 6 的文章向量後，額外連接一個使用者和主題向量，來捕捉此資訊。

圖 7　**考慮使用者和主題語意之後的卷積類神經網路**

另外，一篇文章可能同時包含多個使用者（一個作者和多個按讚者），也可能包含多個主題（此主題是用主題模型，例如 Latent Dirichlet Allocation（LDA），自動判斷，我們選出前幾個主要的主題類別為一篇文章可能的主題），對於此，我們添加了一個最大選擇層（Maximum pooling layer），物理意義近似於用來選出最有代表性的使用者或是主題來判斷文章語意。

對於文章的回覆內容我們也使用如本文同樣的概念，考慮回覆者、回覆的主題，產生出一個回覆向量，接著再用一個最大選擇層來整合回覆的資訊供類神經網路來使用。

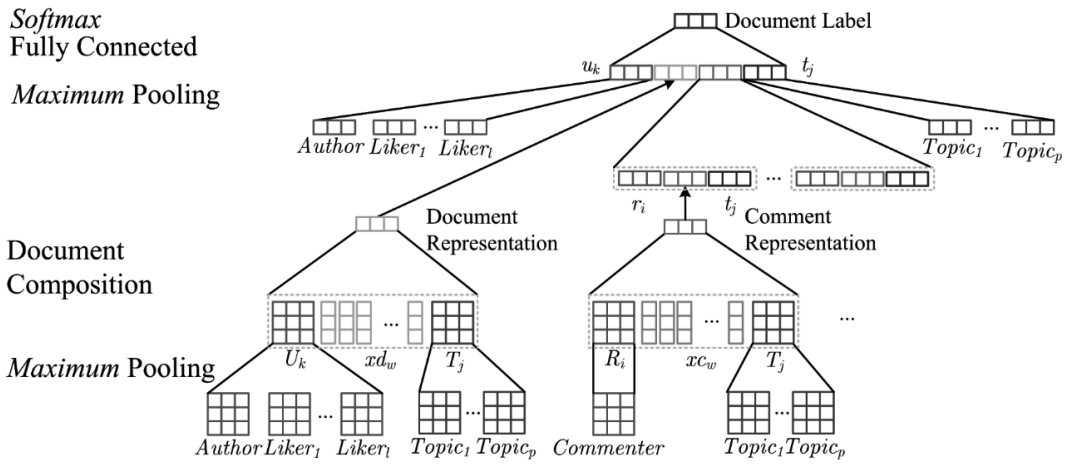

圖 8　**UTCNN 的完整模型**

註：考慮到使用者、主題、回覆的資訊，並將這些資訊用一個向量來表示，最後經過一個輸出層來判斷這篇文章的語意。UTCNN 的效能及與其他的模型比較可參考〔6〕。因為 UTCNN 是一個監督式的深度學習工具，使用 UTCNN 需要準備已標記的文本，同時需要 Tensorflow〔12〕及 Keras〔13〕環境。目前 UTCNN 的架構是固定的，使用時不開放設定深度學習的層數及輸出層使用的函數，模型中有一些維度參數可供設定，但建議採用經實驗測試後的最佳值，也就是模型預設值。

結語

本文介紹了中文情感語意分析套件 CSentiPackage，此套件包含的項目均為發展情感分析相關技術非常有用的資源，且使用上十分的簡易，同時用於研究目的話則開放自由免費下載，可提供給有志於在情感語意分析領域的研究者一個十分有效的資源，相關的演算法與效能皆可參考本文中所引用的參考文獻。本套件提供的 UTCNN 模型，更是一個不限語言皆可使用的模型，同時可以用在社群媒體與論壇之中，為情感與意見分析領域最具應用性的產品評論及輿情分析兩大議題上，提供了一個目前技術水準最高的工具。

藉由本文，我們希望能推廣這個套件給更多有興趣及需要的研究者或產品開發者，未來我們將持續在套件中增加相關的工具程式，以提供情感與意見分析技術最新與最完整的資源予中文世界。

誌謝

本論文中之研究部分由科技部編號 106-3114-E-468-001- 計畫補助支持，特此誌謝。

附註

[1] http://nlp.stanford.edu/sentiment/treebank.html

[2] http://www.wjh.harvard.edu/~inquirer/

[3] http://wndomains.fbk.eu/wnaffect.html

[4] http://sentiwordnet.isti.cnr.it/

[5] http://saifmohammad.com/WebPages/lexicons.html

[6] http://sentic.net/

[7] http://www.cs.cornell.edu/people/pabo/movie-review-data

[8] http://ai.stanford.edu/~amaas/data/sentiment

[9] http://www.ukp.tu-darmstadt.de/data/sentiment-analysis/darmstadt-service-review-corpus/

[10] http://mpqa.cs.pitt.edu/lexicons/

[11] https://www.ldc.upenn.edu/

[12] https://www.tensorflow.org/

[13] https://keras.io/

參考文獻

Chen, W. F., Ku, L. W., & Lee, Y. H. (2015). Mining Supportive and Unsupportive Evidence from Facebook Using Anti-Reconstruction of the Nuclear Power Plant as an Example. *In AAAI Spring Symposium on Socio-Technical Behavior Mining: From Data to Decisions*, 10-15.

Chen, W. F. & Ku, L. W. (2016). UTCNN: A Deep Learning Model of Stance Classification on Social Media Text. *In COLING*, 1635-1645.

Huang, T. H., Ku, L. W., & Chen, H. H. (2010). Predicting Morphological Types of Chinese Bi-Character Words by Machine Learning Approaches. *Proceedings of LREC*, 844-850.

Ku, L. W., Ho, H. W., & Chen, H. H. (2009). Opinion Mining and Relationship Discovery Using CopeOpi Opinion Analysis System. *Journal of the American Society for Information Science and Technology, 60*(7), 1486-1503.

Ku, L. W., Huang, Ting-Hao (Kenneth) & Chen, H. H. (2010). Construction of a Chinese Opinion Treebank. *Proceedings of LREC*, 1315-1319.

Ku, L. W., Liang, Y. T., & Chen, H. H. (2006). Opinion Extraction, Summarization and Tracking in News and Blog Corpora. *In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 100-107.

Wang, S. M., & Ku, L. W. (2016). ANTUSD: A Large Chinese Sentiment Dictionary. *Proceedings of LREC*, 2697-2702.

DOI: 10.6245/JLIS.201804_44(1).0002

Introduction to CSentiPackage: Tools for Chinese Sentiment Analysis

Wei-Fan Chen
Ph. D. Student, Webis group, Faculty of Media
Bauhaus-Universität Weimar, Germany
E-mail: wei-fan.chen@uni-weimar.de

Lun-Wei Ku
Assistant Research Fellow, Institute of Information Science
Academia Sinica, Taiwan (R.O.C.)
E-mail: lwku@iis.sinica.edu.tw

Keywords: Chinese Text Processing; Opinion Mining; Sentiment Analysis; Social Media Analytics

---

【Abstract】

Sentiment analysis determines the polarities and strength of the sentiment-bearing expressions, and it has been an important and attractive research area. In the past decade, resources and tools have been developed for sentiment analysis in order to provide subsequent vital applications, such as product reviews, reputation management, call center robots, automatic public survey, etc. However, most of these resources are for the English language. Being the key to the understanding of business and government issues, sentiment analysis resources and tools are required for other major languages, e.g., Chinese.To overcome this obstacle, we introduce CSentiPackage, where resources for retrieving sentiment from texts in the Chinese language, are provided. The related sentiment analysis technologies and datasets are described to give the readers the opportunities to use resources and tools to process Chinese sentiment texts from the very basic to the advanced, i.e., applying sentiment dictionaries, obtaining sentiment scores, and analyzing stance of social media posts using the deep learning model. The introduced resources and tools in this paper include NTUSD, ANTUSD, the Chinese Morphological Dataset, the Chinese Opinion Treebank, CopeOpi, and UTCNN. These resources are all available at http://academiasinicanlplab.github.io/ and they are free for the research purpose.

【Long Abstract】

Introduction

The rapid accumulation of data in social media (in million and billion scales) has imposed great challenges in information extraction, knowledge discovery, and data mining, and texts bearing sentiment and opinions are one of the major categories of user generated data in social media. Sentiment analysis is the main technology to quickly capture what people think from these text data, and is a research direction with immediate practical value in 'big data' era. Learning such techniques will allow data miners to perform advanced mining tasks considering real sentiment and opinions expressed by users in additional to the statistics calculated from the physical actions (such as viewing or purchasing records) user perform, which facilitates the development of real-world applications. However, the situation that most tools are limited to the English language might stop academic or industrial people from doing research or products which cover a wider scope of data, retrieving information from people who speak different languages, or developing applications for worldwide users.

More specifically, sentiment analysis determines the polarities and strength of the sentiment-bearing expressions, and it has been an important and attractive research area. In the past decade, resources and tools have been developed for sentiment analysis in order to provide subsequent vital applications, such as product reviews, reputation management, call center robots, automatic public survey, etc. The most well known tools and resources include (1) Stanford Sentiment Treebank[1], which is part of Stanford CoreNLP: the sources codes are available for download (2) General Inquirer Lexicon[2] which includes 11,178 words selected from the Harvard dictionary and the Lasswell dictionary: words are annotated with the labels such as positive, negative, active, passive, etc. (3) WordNet Affect (WN-Affect) [3]and SentiWordNet[4], which are constrcted based on WordNet: the former gives each synset in WordNet one or more emotion labels, while the latter gives three sentiment scores for the positive, negative, subjective properties. (4) EmoLex[5]: it includes 14,182 words which are labeled with positive, negative, one of eight emotions and one of four strength scales. (5) SenticNet[6]: it includes 17 thousands of concepts and each is annotated with 4 different sentiment scores. The final polarity of the concept is calculated considering these 4 scores. These materials show that the research of sentiment analysis is prosperous.

However, most of these resources, or even their derived tools, are for the English language. Being the key to the understanding of business and government issues, sentiment analysis resources and tools are required for other major languages, e.g., Chinese. To overcome this obstacle, we introduce CSentiPackage, where resources for retrieving sentiment from texts in the Chinese language, are provided. The related sentiment analysis technologies and datasets are described to give the readers the

opportunities to use resources and tools to process Chinese sentiment texts from the very basic to the advanced, i.e., applying sentiment dictionaries, obtaining sentiment scores, and analyzing stance of social media posts using the deep learning model.

CSentiPackage Resources and Tools

The introduced resources and tools in CSentiPackage include:

1. NTUSD (Ku, L. W., Liang, Y. T., & Chen, H. H., 2006): NTUSD is a Chinese sentiment dictionary. It contains a list of 2,812 positive words and 8,276 negative words. It is one of the most popular Chinese opinion dictionaries. Both traditional and simplified versions are available.

2. ANTUSD (Wang, S. M., and Ku, L. W., 2016): ANTUSD is an augmented version of NTUSD. A total of 9,382 positive, 16 neutral, 11,224 negative, 5,415 non-opinionated, and 612 negation words are included. Labels of each word were annotated by multiple annotators.. For each word, six values are given, and they are the sentiment score from CopeOpi, the number of positive labels, the number of neutral labels, the number of negative labels, the number of non-opinion labels, and the number of non-word labels. Both traditional and simplified versions are available.

3. The Chinese Morphological Dataset (Huang, T. H., Ku, L. W., & Chen, H. H., 2010): The Chinese Morphological Dataset contains more than 8,000 words and it is the previous version of ACBiMa (Advanced Chinese Bi-Character Word Morphological Analyzer, which contains more than 11, 000 words). A total of 8 morphological labels are annotated, including Parallel, Substantive-Modifier, Subjective-Predicate, Verb-Object, Verb-Complement, Negation, Confirmation, and Others.

4. The Chinese Opinion Treebank (Ku, L. W., Ting-Hao (Kenneth) Huang, & Chen, H. H., 2010): The Chinese Opinion Treebank is a dataset with opinion labels on Chinese Treebank 5.1. This dataset only contains the annotations themselves. The original sentences and parsed trees are not included. Therefore, users need to license the Chinese Treebank 5.1 first from Linguistic Data Consortium (LDC[7]) in order to use this dataset. In this dataset, each sentence is annotated with the opinionated label (yes/no), the polarity label (positive/neutral/negative), the type label (expression/status/action). If the sentence is labeled as opinionated, the relations between its composite words are also annotated.

5. CopeOpi (Ku, L. W., Ho, H. W., & Chen, H. H., 2009): CopeOpi is a statistical sentiment analysis tool. It provides the sentiment score for the given word, sentence, or document. This sentiment score can be used to determine the polarity of the sentiment (positive if the score is greater than zero, negative if the score is less than zero) and the strength of the sentiment (the absolute value of

the score denotes the degree of the strength). CopeOpi is a convenient un-supervised tool. Using CopeOpi, it is not necessary to re-train the model and it runs comparably efficient.

6. UTCNN (Chen, W. F. and Ku, L. W., 2016): UTCNN is a language independent deep neural network model used to determine the post stance in the forum-like platform. It considers the interactions between readers and the post content to boost up the performance. It has been tested on the Facebook posts and the CreateDebate benchmark dataset to show that it is the state of the art model for the stance classification. Users will need to prepare the training data to train their own model before using it. The default settings as well as sample training data are provided.

Conclusion

The resources introduced in this paper are all available at http://academiasinicanlplab.github.io/ and they are free for the research purpose. Through this paper, we hope to lower the entry barrier of developing technologies for Chinese sentiment analysis and encourage more researchers to step into this research area. In the future, we will keep developing related tools for CSentiPackage and provide the most complete resources for the Chinese sentiment analysis.

【Romanization of references is offered in the paper.】