意見持有者辨識之研究

A Study on Identification of Opinion Holders

李佳穎　古倫維　陳信希
國立臺灣大學資訊工程學系
{cylee, lwku}@nlg.csie.ntu.edu.tw , hhchen@csie.ntu.edu.tw

摘要

意見持有者辨識是從意見句中擷取出表述意見的人或組織，本研究將意見持有者辨識分爲作者意見辨識及意見持有者標記兩部分，作者意見辨識使用支援向量機處理，意見持有者標記使用條件隨機域處理。本研究提出的方法應用在 NTCIR7 MOAT 繁體中文語料的效能達到 F 值 0.734，是採取機器學習方法的參賽隊伍中效能最佳者，也相當接近目前最佳系統的效能。對於意見持有者辨識語料中標記歧異的情形，本研究加以分析，並提出使用此語料來訓練模型的方法。

Abstract

The identification of opinion holders aims to extract entities that express opinions in opinion sentences. In this paper, the task of opinion holder identification is divided into two subtasks: the identification of author's opinions and the labeling of opinion holders. Support vector machine is adopted to identify author's opinions, and conditional random field model (CRF) is utilized to label opinion holders. The proposed method achieves an F-score 0.734 in NTCIR7 MOAT task at traditional Chinese side. The proposed method achieves the best performance among participants who adopted machine learning methods, and also this performance was close to the best performance in this task. In addition, the ambiguous markings of opinion holders are analyzed, and the best way to utilize the training instances with ambiguous markings is proposed.

關鍵詞：意見持有者辨識，意見探勘，條件隨機域，支援向量機

Keywords: opinion holders identification, opinion mining, CRF, SVM.

一、緒論

意見代表人們對某個議題的主觀想法，人們常透過文章表述意見。隨著 Web2.0 的崛起，網路上出現大量、免費與即時的資料，使用者對文章中的意見很感興趣，但卻無法大量閱讀數以千萬計的資料。意見探勘 (opinion mining) 的技術可以幫助使用者自動分析文章中的意見， Kim 和 Hovy[1] 在 2004 年提出意見中包括意見傾向 (opinion polarity)、意見強度 (opinion strength)、意見持有者 (opinion holder) 及評論目標 (opinion target) 四個要素。意見傾向描述此意見是正面、中立或負面，意見強度描述此意見的語氣強弱，表述此意見的人或組織稱爲意見持有者，而討論的主題則稱爲評論目標。以例句 1 爲例，此句的意見傾向爲正面、意見強度爲強烈、意見持有者爲王建民、評論

目標爲打棒球。意見持有者通常會以一或多個詞的形式出現在意見句中，我們將這些詞稱爲意見持有者的代表詞，但有時意見持有者不會以詞的形式出現在意見句中，例如例句 2 是作者根據例句 1「王建民」的意見推論的意見，例句 2 的意見持有者爲文章作者。

例句 1：王建民非常喜歡打棒球

例句 2：王建民應該也喜歡打網球

在意見探勘中，意見持有者辨識的技術對於了解有哪些人或組織在表述意見、某個人或組織在哪些議題中發表過意見及兩個人或組織發表過的意見是否相似等相關資訊特別重要。意見持有者辨識可應用於社群網路分析中，找出社群網路中是否存在著一些意見領袖，他們的意見常被引用，也會影響其他使用者的意見。意見持有者辨識也可以應用在意見問答系統中，找出某些意見是由哪些意見持有者提出的，並進而藉由意見持有者的權威性與可靠度來輔助判斷答案的權威性與可靠度。

意見持有者辨識主要有三大挑戰：同指涉解析、巢狀結構及處理歧異的標記。意見持有者有時會以代詞 (Anaphor) 的形式出現在文句中，並指涉到前面的先行詞 (Antecedent) ，例如例句 3 中的「雙方」即是指涉到「美國」與「中共」。

例句 3：

據媒體報導，美國在與中共討論簽署停止以核武相互瞄準協議的問題，貝肯說，雙方過去就曾討論此事，前任國防部長裴利在中共國防部長遲浩田於一九九六年十二月訪美時就曾提起，後來雙方在其他會議中也曾討論。

意見句有時會有巢狀結構 (nested structure) ，以例句 3 爲例，文章作者引述「媒體報導」的內容，「媒體報導」的內容又引述「貝肯」的發言，意見持有者常會是子句的主詞，判斷意見持有者是哪一層結構中的主詞也是意見持有者辨識的一個重要議題。

標記意見持有者辨識所用的語料時，有時會出現標記歧異，不同的標記者可能會認爲文句的意見持有者爲不同的實體。以例句 3 爲例，一位標記者認爲意見持有者爲「國防部發言人貝肯/貝肯」，另一位標記者卻認爲意見持有者爲『美「中」/雙方』，從文句內容來看，意見持有者爲「國防部發言人貝肯/貝肯」，但深究背後的意義，貝肯是轉述『美「中」/雙方』的意見，兩種說法都沒錯，端看標記者的認知，也因此意見持有者可能被多個標記者標記出不同的答案。，如何利用標記歧異的語料也是意見持有者辨識的一大挑戰。

二、相關研究

Pang 和 Lee[2] 整理出意見探勘領域中重要的研究，意見持有者辨識的研究剛開始起步，研究團隊使用的方法主要可分爲以經驗法則 (heuristic rule) 爲基礎與以機器學習爲基礎兩種。

(一)、以經驗法則爲基礎的方法

以經驗法則爲基礎的方法中，Yohei 等人[3] 先使用名詞片語與語法特徵值，透過支援向量機，將意見持有者分爲文章作者與非文章作者，接著再透過語法規則，選出最有可能的具名實體，做爲答案的意見持有者，他們主要專注於處理英文與日文的語料。Xu

和 Wong[4] 提出的方法是先解決同指涉問題，再使用經驗法則擷取出意見持有者，使用的規則與標點符號、連接詞、字首 (prefix) 、字尾 (suffix) 與表述關鍵字相關，Xu 和 Wong 的方法是目前中文意見持有者辨識中效能最佳的，在 NTCIR7 多語意見分析評比項目的繁體中文語料上， F 值可達到 0.825。

(二)、以機器學習爲基礎的方法

以機器學習爲基礎的方法中，許多研究團隊使用最大熵法 (maximum entropy) 、支援向量機演算法 (support vector machine algorithm) 與條件隨機域模型 (conditional random field model) 等分類器解決此問題。

Kim 和 Hovy[5] 以最大熵法從新聞語料的文句擷取出意見持有者與意見評論目標，他們先找出意見詞 (opinion words) 與進行語意角色標注 (semantic role labeling) ，再找出代表意見持有者與意見評論目標的語意角色。

使用支援向量機演算法的研究團隊中， Kim 等人[6] [7] 先將意見持有者分爲文章作者、有同指涉情形與沒有同指涉情形三種，再使用詞彙與語法特徵值 (syntactic features) ，透過支援向量機，選出最有可能的意見持有者，Kim 等人的方法是目前英文意見持有者辨識中效能最佳的，應用在 NTCIR7 多語意見分析評比項目的英文語料上，F 值可達到 0.346。 Wu 等人[8] 則使用詞彙與詞性特徵值，透過 L2-norm 線性核心支援向量機，以類似具名實體辨識的方法解決中文的意見持有者辨識問題。

使用條件隨機域模型的研究團隊中， Breck 與 Choi 等人[9] [10] 使用詞彙、語法、字典 (dictionary-based) 及依存關係 (dependency relation) 特徵值，透過條件隨機域模型標記出最有可能的意見持有者。相形之下， Meng 和 Wang[11] 使用詞彙、詞性及表述關鍵字 (operator) 特徵值，而 Liu 和 Zhao[12] 則使用詞性、語意、依存關係、位置 (position) 及前後文 (contextual) 特徵值，透過條件隨機域模型標記出最有可能的意見持有者。

三、意見持有者辨識方法

本研究將意見持有者辨識分爲作者意見辨識及意見持有者標記兩個主要工作。本研究提出的流程包括前置處理程序、作者意見辨識程序、意見持有者標記程序、後置處理程序及結果合併程序五個部份。

(一)、針對斷詞與詞性標記的特殊處理

前處理程序包括斷詞、詞性標記、具名實體辨識及特徵值擷取。本實驗使用的是羅[13]研發的斷詞及詞性標記系統，爲了能夠更準確的斷出與意見持有者相關的具名實體，我們修改斷詞系統的人名模組並引入字典資訊。我們發現外國人名容易出現斷詞錯誤，所以我們著手修改斷詞及詞性標記系統的人名模組，來處理日文姓名長度與中文姓名長度不同的問題，我們在原本人名模組使用的姓氏列表中加入日本常見姓氏，名字長度的限制也從兩個字放寬爲三個字，使得系統能正確斷出如「高村正彥」、「兒玉源太郎」、「鈴木」等日本人名。

我們另外加入了職稱名、職業名、日本常見姓氏及台灣公營企業列表等字典。本研究的具名實體辨識是使用查詢詞典的方法，將人名、地名及組織名分別標上標籤。

(二)、作者意見辨識

作者意見辨識的目的是辨識意見句之意見持有者是否爲文章作者。本研究把作者意見辨識的問題視爲二元分類問題 (binary classification problem)，使用支援向量機來處理，實作上使用的套裝軟體是 Chang 和 Lin[14] 開發的 LIBSVM 。

表一、作者意見辨識使用的特徵值

特徵值類別	特徵值代號	特徵值描述
詞彙相關資訊	fHasI	本句有沒有「我」
	fHasWe	本句有沒有「我們」
	fNumI	本句有幾個「我」
	fNumWe	本句有幾個「我們」
詞性相關資訊	fHasPronoun	本句有沒有代名詞
	fHasManPronoun	本句有沒有人稱代名詞
	fNumPronoun	本句有幾個代名詞
	fNumManPronoun	本句有幾個人稱代名詞
具名實體資訊	fHasPer	本句有沒有人名詞
	fHasLoc	本句有沒有地名詞
	fHasOrg	本句有沒有組織名詞
	fHasNa	本句有沒有普通名詞
	fHasNb	本句有沒有專有名詞
	fHasNc	本句有沒有地方名詞
	fNumLoc	本句有幾個地名詞
	fNumOrg	本句有幾個組織名詞
	fNumPer	本句有幾個人名詞
	fNumNa	本句有幾個普通名詞
	fNumNb	本句有幾個專有名詞
	fNumNc	本句有幾個地方名詞
標點符號資訊	fHasExclamation	本句有沒有驚嘆號，例如：「！」或「!」
	fHasQuestion	本句有沒有問號，例如：「？」或「?」
	fHasColon	本句有沒有冒號，例如：「：」或「:」
	fHasLeftQuotation	本句有沒有上引號，例如：『「』或「【」
	fHasRightQuotation	本句有沒有下引號，例如：『」』 或「】」
文句組成資訊	fNumChar	本句有幾個字
	fNumWord	本句有幾個詞
	fNumSubsen	本句有幾個子句
意見相關資訊	fOperator	本句有沒有某個表述關鍵字

表二、意見持有者標記使用的特徵值

特徵值類別	特徵值代號	特徵值描述
詞彙相關資訊	fWord	本詞
詞性相關資訊	fPOS	本詞的詞性
	fIsPronoun	本詞是不是代名詞
	fIsNoun	本詞是不是名詞
具名實體資訊	fIsPer	本詞是不是人名
	fIsLoc	本詞是不是地名
	fIsOrg	本詞是不是組織名
標點符號資訊	fAfterParen	本詞是否在下引號之後兩詞，例如：『」』 或「】」
	fBeforeColon	本詞是否在冒號之前兩詞，例如：「：」或「:」
文句組成資訊	fNearSenStart	本詞是否靠近句首
	fSenLen	本詞所在句中的詞數
	fWordOrder	本詞在句中的詞序
	fWordPerc	本詞在句中詞序的百分比
前後文相關資訊	fNearVerb	同句中最靠近本詞的動詞
	fNearVerbPOS	同句中最靠近本詞的動詞詞性
	fDistNearVerb	同句中本詞到動詞的最短距離
意見相關資訊	fHasOpKW	同句中有沒有表述關鍵字
	fHasPosKW	同句中有沒有正面意見詞
	fHasNegKW	同句中有沒有負面意見詞
	fHasNeuKW	同句中有沒有中立意見詞
	fNearOpKW	同句中最靠近本詞的表述關鍵字
	fNearPosKW	同句中最靠近本詞的正面意見詞
	fNearNegKW	同句中最靠近本詞的負面意見詞
	fNearNeuKW	同句中最靠近本詞的中立意見詞
	fNearOpKWPOS	同句中最靠近本詞的表述關鍵字的詞性
	fNearPosKWPOS	同句中最靠近本詞的正面意見詞的詞性
	fNearNegKWPOS	同句中最靠近本詞的負面意見詞的詞性
	fNearNeuKWPOS	同句中最靠近本詞的中立意見詞的詞性
	fDistOpKW	同句中本詞到表述關鍵字的最短距離
	fDistPosKW	同句中本詞到正面意見詞的最短距離
	fDistNegKW	同句中本詞到負面意見詞的最短距離
	fDistNeuKW	同句中本詞到中立意見詞的最短距離

作者意見辨識使用的特徵值主要可分爲詞彙、詞性、具名實體、標點符號、文句組成及

意見相關資訊六種類別，表一列出作者意見辨識所有使用的特徵值，其中詞性、文句組成、意見相關資訊及標點符號中的驚歎號相關特徵值爲本研究首先提出的。

文句組成相關特徵值包括作者意見的文句在文句長度上的資訊。意見相關資訊則包含表述關鍵字 (operator) ，希望了解作者發表的意見中是否較常使用特定的表述關鍵字。表述關鍵字是用來表達意見的詞，通常爲動詞，例如：「說」、「報導」及「主張」等。標點符號中的驚嘆號常用來表達個人情緒的反應。

(三)、意見持有者標記

意見持有者標記的目的是辨識出意見持有者的代表詞，本研究將意見持有者標記問題視爲二元分類問題，試著使用決策樹演算法 (Decision Tree Algorithm) 解決，實作上使用的套裝軟體是 Mierswa 等人[15] 開發的 RapidMiner 中的 CHAID 決策樹演算法，CHAID 爲使用卡方檢定 (CHI Square Test) 的剪枝決策樹 (Pruned Decision Tree) 。本研究也將意見持有者標記的問題視爲序列標記問題 (sequential labeling problem) ，使用 Lafferty 等人[16] 提出的條件隨機域模型來標記出意見詞有者所涵蓋的詞彙，實作上使用的套裝軟體是 Kudo[17] 開發的 CRF++ 。

意見持有者標記使用的特徵值主要可分爲主要可分爲詞彙、詞性、具名實體、標點符號、文句組成、前後文及意見相關資訊七種類別，表二列出意見持有者標記所有使用的特徵值，其中前後文、意見相關資訊及詞性中的本詞是不是代名詞或名詞特徵值爲本研究首先提出的。

前後文相關資訊的特徵值考慮意見持有者的代表詞是否會較常與某些動詞搭配使用。意見相關資訊的特徵值則包含句中與意見關鍵字：表述關鍵字、正面意見詞、負面意見詞及中立意見詞相關的特徵值。正面意見詞爲表達正面意見立場的詞：如「同意」、「相信」、「成功」等。負面意見詞爲表達負面意見立場的詞：如「不會」、「反對」、「指控」等。中立意見詞爲表達中立意見立場的詞：如「未置評」、「兩難」、「可能」等。

NTCIR7 多語意見分析評估項目的訓練集較小，我們引入 Blum 和 Mitchell[18] 提出的協同訓練 (co-training) 來改善效能。協同訓練是半監督式機器學習方法 (semi-supervised learning method) ，能結合標記資料與未標記資料一起訓練模型。本研究挑選 CRF 預測信心值較高的實例，以文句爲單位回饋到訓練語料中，藉此提升系統效能。

(四)、後置處理

後置處理包含意見持有者爲詞組時之特殊處理及具名實體修復兩個部份。意見持有者標記會標示出本詞是不是意見持有者的一部分，但意見持有者常會由多個詞組成，因此需要根據標記結果將他們組合起來。本研究使用五種標籤來標記意見持有者：意見持有者的首詞 (H) 、尾詞 (T) 、中間詞 (I) 、本身爲意見持有者的單詞 (S) 及非意見持有者 (O) ， 因此 CRF 標籤集由 HITSO 五種標籤排列組合而成。

CHAID 分類器產生的結果爲 YES 與 NO 標籤，代表本詞是不是意見持有者的一部分，如果 CHAID 分類器產生的結果全部爲 NO 標籤時，系統會將意見持有者設爲文章作者。本系統使用下列兩條規則將這些詞組合成詞組：

規則 1：將連續名詞組合起來。
例如：「印度 (Nc) 總統 (Na) 瓦希德 (Nb) 」將組合成「印度總統瓦希德」

規則 2：使用連接詞、「的」字及頓號「、」將連續名詞組合起來。
例如：「瓦希德 (N) 、 (PAUSECATEGORY)柯林頓 (N) 與 (Caa) 小淵惠三 (N) 」將組合成「瓦希德、柯林頓與小淵惠三」

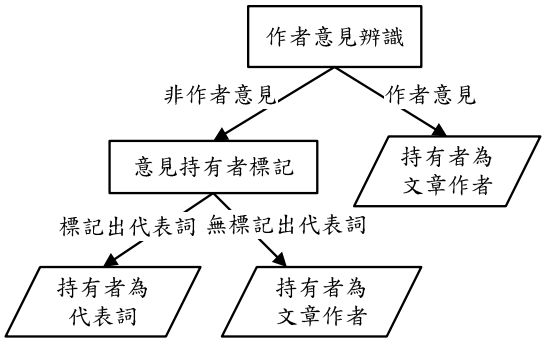
根據 CRF 分類器產生的結果標籤來組合意見持有者，組合規則爲先找到信心值最高的 H 標籤，再找到後面連續多個 I 標籤，最後再找到 T 標籤將這些詞組合起來。如果 CRF 分類器產生的結果全部爲 O 標籤時，系統會提報文章作者爲組合結果。

外國譯名的具名實體容易在斷詞時出現錯誤，這樣的錯誤可能造成辨識出不完整的意見持有者，因此我們可能需要修復意見持有者中的具名實體。本系統假設不完整的具名實體在文章中出現的頻率與完整的具名實體相同，所以本系統會將意見持有者標記結果跟前後字串接，測試組合成的新詞與原詞在文章中出現的頻率是否相同，相同則以新詞取代原詞。例句 4 爲具名實體修復過程的一例，原本意見持有者標記的結果是「蘇哈」，括號內的數字代表該詞在文章中出現的次數，修復後可輸出「蘇哈托」。透過這樣的具名實體修復方法，本系統可以將斷詞錯誤的具名實體修復爲完整的具名實體。

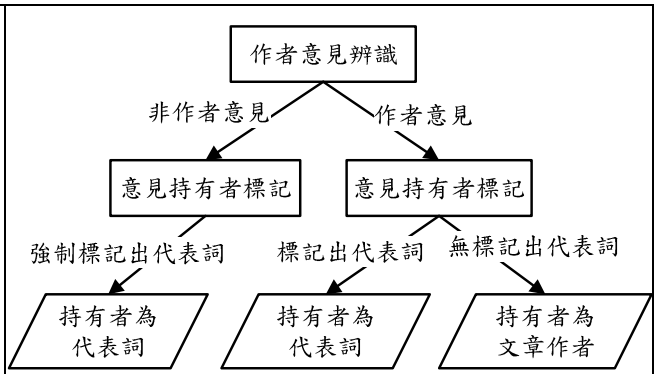
例句 4：印尼強人蘇哈托統治印尼卅二年
執行過程：蘇哈(16) → 蘇哈托(16) 頻率相同往後繼續
→ 蘇哈托統(1) 頻率不同改往前 → 人蘇哈托(3) 頻率不同結束

(五)、合併作者意見辨識與意見持有者標記之結果

作者意見辨識將意見句區分爲作者意見及非作者意見兩類，意見持有者標記則可標記出意見持有者的代表詞，本研究將此兩部分的結果合併，產生最後提報之意見持有者。本研究提出兩種結果合併策略，圖二、三爲結果合併策略 A 與 B 的示意圖。



圖一、結果合併策略 A



圖二、結果合併策略 B

結果合併策略 A 中，我們相信作者意見辨識判斷出的作者意見，作者意見辨識判斷出非作者意見的文句則透過意見持有者標記標出意見持有者的位置。結果合併策略 B 中，我們相信作者意見辨識判斷出的非作者意見，作者意見辨識判斷出是作者意見的部份，因爲不夠確定，再透過意見持有者標記二次檢查是否爲作者意見：如果沒有標出意

見持有者的代表詞，才提報該句的意見持有者是作者。作者意見辨識判斷出的非作者意見並沒有標記出意見持有者的代表詞爲何，所以我們透過意見持有者標記程序強制標記出代表詞，也就是從所有詞中找出最有可能代表意見持有者的詞。

透過前置處理、作者意見辨識、意見持有者標記、後置處理及結果合併五個程序，我們就可以提報最後辨識出之意見持有者。

四、實驗與討論

本節將介紹本實驗使用的語料與資源，並討論作者意見辨識實驗、意見持有者標記實驗及意見持有者辨識整體實驗的結果。

(一)、NTCIR 7 多語意見分析評比項目介紹

本實驗使用的語料爲 NTCIR 7 多語意見分析評比項目中繁體中文的語料庫， NTCIR (NII Test Collection for IR Systems) 是日本國家資訊研究所 (National Institute of Informatics，NII) 所策劃主辦的國際評比會議，是世界三大資訊檢索會議之一。多語意見分析評比項目 (Multilingual Opinion Analysis Task, MOAT) 是其中一個評比項目，Seki 等人[19] 對此評比項目有詳細的介紹。

多語意見分析評比項目提供英文、日文、繁體中文與簡體中文的語料庫，語料庫中提供相關句、意見句、意見傾向、意見強度、意見持有者與評論目標的標記。語料庫分爲訓練集與測試集， NTCIR7 訓練集包括 3 個主題、1,509 個文句、944 個意見句，NTCIR7 測試集包括 14 個主題、4,665 個文句、2,174 個意見句，參賽者們會以文句爲單位進行意見分析。因爲 NTCIR7 訓練集較小，本實驗的訓練語料加入 NTCIR6 意見分析試驗評比項目 (Opinion Analysis Pilot Task) 繁體中文的測試集，NTCIR6 意見分析試驗評比項目是 NTCIR 7 多語意見分析評估項目的前身，語料庫中提供相關句、意見句、意見傾向、意見持有者的標記。 NTCIR6 測試集包括 29 個主題、9,240 個文句、5,453 個意見句，我們利用語料庫中關於意見句與意見持有者的標記當作我們的實驗語料。

(二)、實驗資源

本研究使用的實驗資源包括意見詞詞典與具名實體詞典。意見詞詞典的部份包含標記者從 NTCIR 7 多語意見分析評比項目的訓練集中標記出表述關鍵字、正面意見詞、負面意見詞及中立意見詞等意見詞，也使用 Ku 和 Chen[20] 開發的台大意見詞詞典 (NTUSD) ，本研究將這些意見詞詞典應用於特徵值擷取。

本研究引入人名詞典、地名詞典及組織名詞典三類具名實體詞典，使用的詞典包含百萬人名字典、中文詞庫、中央社譯名檔、國立編譯館專業字典、日本常見七千個姓氏、教育部地名譯名詞典、外國地名譯名及台灣公營企業列表。這些具名實體詞典則應用於具名實體辨識。

(三)、作者意見辨識實驗

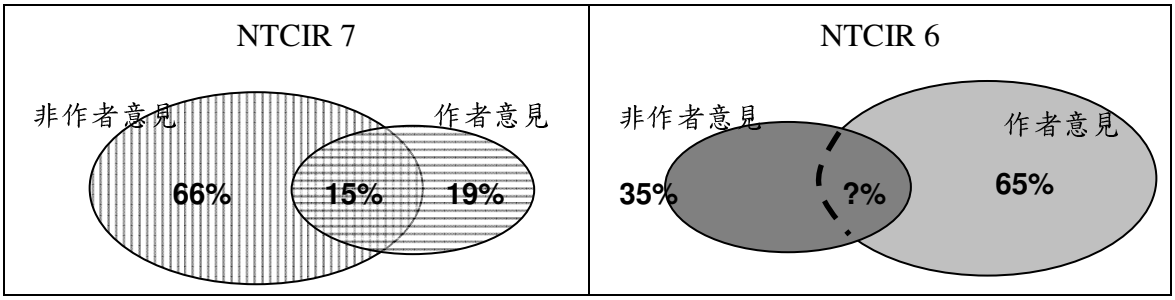
本實驗的目的是辨識意見句之意見持有者是否爲文章作者，本實驗使用的訓練語料是 NTCIR7 訓練集及 NTCIR6 測試集，測試語料是 NTCIR7 測試集中的寬鬆意見句。

NTCIR7 測試集中的意見句判定分爲嚴格意見句與寬鬆意見句，嚴格意見句的條件是三位標記者都將此句標記爲意見句，寬鬆意見句的條件則是三位標記者中，有兩位以上將此句標記爲意見句。

本實驗判斷此句的意見持有者是不是文章作者，也就是此句是不是作者意見，我們使用精確率 (Precision) 、召回率 (Recall) 、F 值 (F-score) 及正確率 (Accuracy) 來評估系統效能。我們研究語料之後發現部分文句會有作者意見標記歧異的情形，例如例句 5 的第二句，一種說法認爲意見持有者爲「臺灣」，另一種說法認爲意見持有者標記爲文章作者。

例句 5：
第一句：臺灣官方已排除核武選擇，並將最終安全託付於美國，
第二句：因爲臺灣了解，核武之路將根本破壞與美國之間的關係。

我們將標記的情形加以分析。圖三是語料庫中作者意見比例示意圖，我們發現在 NTCIR7 的語料庫中，被部分標記者標記爲作者意見，其他標記者標記爲非作者意見的標記歧異語料佔 15%，與被所有標記者標記爲作者意見的比例 (19%) 相距不遠。



圖三、語料庫中作者意見比例示意圖

在 NTCIR6 的語料庫中，不只是意見句，所有的文句都會被標上發表者，代表表示這個說法的人或組織， NTCIR6 的意見持有者標記方法是如果可以從文章中標記出意見持有者就標記，如果不能，則意見持有者爲文章作者，所以無法得知 NTCIR6 的語料庫中標記出現歧異的文句占全部文句的比例。

我們從實驗中發現使用意見句當訓練語料的 F 值爲 73.24%，比使用全部文句當訓練語料的 F 值高了 8.04%。使用 NTCIR 6 加 NTCIR 7 訓練集的 F 值爲 79.98%，也比單獨使用其中一個訓練集爲高，所以本實驗使用 NTCIR 6 加 NTCIR 7 訓練集中的意見句作爲訓練語料。

表三、標記歧異的語料對作者意見辨識效能的影響

設定	精確率	召回率	F 值	正確率
視爲作者意見	***69.68%***	***93.85%***	***79.98%***	***83.49%***
視爲非作者意見	64.87%	95.94%	77.40%	80.31%
不列入訓練語料	50.52%	91.53%	65.10%	77.28%

本實驗的實驗設定共有三種：將標記歧異的語料視爲作者意見、將標記歧異的語料視爲非作者意見、及標記歧異的語料不列入訓練語料。表三顯示在這三種實驗設定下，標記歧異的語料對作者意見辨識效能的影響。

實驗結果顯示，視爲作者意見的設定效能最佳： F 值爲 79.98%，其次爲視爲非作者意見的設定： F 值爲 77.40%，效能最差的是不使用標記歧異的語料： F 值爲 65.10%。從本實驗結果可以發現，使用標記歧異的語料能提升系統效能，將標記歧異的語料視爲作者意見加入訓練效能最佳。

(四)、意見持有者標記實驗

本實驗的目的是辨識出意見持有者的代表詞，以供後置處理程序組合成意見持有者。本實驗使用的訓練語料是 NTCIR7 訓練集，測試語料是 NTCIR7 測試集中的嚴格意見句與寬鬆意見句。本實驗將意見持有者標記的結果組合出最後的意見持有者，再使用正確數、錯誤數、 set F 值 (set F-score) 來評估系統效能。 NTCIR7 的評估方式會先評估每個參賽隊伍正確提報的意見句數，再評估每個參賽隊伍正確提報的意見持有者，參賽隊伍提報的意見句數等於提報的意見持有者數，也等於答案的意見持有者數，此時精確率、召回率、 F 值與正確率會相等，因此我們不使用精確率、召回率、 F 值，而使用 set F 值來評估效能， set F 值的定義如下：

$$setF\text{值} = \frac{\text{系統正確提報意見持有者的意見句數}}{\text{系統正確提報的意見句數}} \quad (1)$$

實驗比較兩種分類演算法：用來解決二元分類問題的決策樹演算法 CHAID ，及用來解決序列標記問題的條件隨機域模型 CRF 。 CHAID 使用的意見持有者標記標籤是 YES 與 NO 標籤， CRF 使用的意見持有者標記標籤是意見持有者的首詞 (H) 、意見持有者的中間詞 (I) 及非意見持有者組成詞 (O) 。表四顯示使用不同分類演算法對作者意見辨識效能的影響。

表四、分類演算法對意見持有者標記效能的影響

	分類演算法	正確數	錯誤數	set F 值
嚴格意見句	CHAID	564	605	48.16%
	CRF	817	351	69.89%
	CRF+CHAID	825	344	70.57%
寬鬆意見句	CHAID	981	967	50.31%
	CRF	1317	631	67.57%
	CRF+CHAID	1322	627	67.83%

嚴格意見句部份的評估中， CRF 的 set F 值爲 69.89%，比 CHAID 高 21.73%，效能好很多，寬鬆意見句部份的評估也可以得到類似的結果。接著我們將 CHAID 預測出來的結果當作 CRF 的一個特徵值再重新訓練模型，也可以小幅提升系統效能。原因可能因爲 CRF 使用的意見持有者標記標籤較多，增加的 H 標籤有助提升系統效能，也可能因爲 根據 CRF 標籤結合詞組的效能比 根據 CHAID 的結果再用規則連接的效能

好。

寬鬆意見句部份的評估中，本研究能達到的最佳效能是 set F 值 67.83%。在標記結果錯誤的實例中，有 13% (254 句) 是正確答案的意見持有者爲單詞或詞組，但系統標記出之單詞或詞組與正確答案不符，以下將分析標記結果錯誤的實例，並提出解決的方法。我們根據正確答案與系統標記出之答案的位置分析，將主要的標記錯誤分爲 6 類：

1.答案無關聯

無法判斷正確答案與系統標記出之答案之間的關聯，此類佔標記錯誤的 29.1%。

2.多擷取前後一詞

系統標記出之答案包含正確答案，但卻又多將前後一詞判斷爲意見持有者的一部分，例如：也許**蘇哈托**、**魯斯曼**日前、**他們**可以。範例中，**粗體字**代表正確答案，標底線的字代表系統標記出之答案，此類佔標記錯誤的 18.1%。

3.擷取出頭銜但未擷取出其後的人名

系統標記出之答案包含正確答案中之頭銜，但卻沒有擷取出頭銜後的人名，例如：**科索伏著名塞裔領袖特拉伊科維契**，此類佔標記錯誤的 8.3%。

4. 擷取出形容詞但未擷取出其後的普通名詞

系統標記出之答案包含正確答案中前面的形容詞但未擷取出其後的普通名詞，例如：**該裁決**、**國際停火觀察團**、**美國全國公共廣播電台**，此類佔標記錯誤的 7.5%。

5. 額外擷取出其他非答案詞

系統標記出之答案包含正確答案，但卻又額外擷取出其他的詞，例如：**狄蘭**在記者會、**他**祝賀巴勒斯坦的科學家，此類佔標記錯誤結果的 5.5%。

6. 具名實體擷取不完整

正確答案包含系統標記出之答案，但具名實體部份擷取得不完整，例如：**德州農工大學的複製專家韋斯特休生**、**車燈廠堤維西**，此類佔標記錯誤結果的 4.7%。

根據這些標記結果錯誤的類別，我們提出幾種方法來改善系統效能。大部分的類別都有意見持有者詞組中首詞、尾詞不明確的問題，所以我們提出增加意見持有者標籤的方法。從實驗中發現，使用 HIO 標籤集在嚴格意見句部份的評估中可得到最佳的 set F 值 70.57%。針對第 6 類具名實體擷取不完整的問題，我們也提出具名實體修復方法來解決這個問題，從實驗中發現，加入協同訓練與具名實體修復在嚴格意見句部份的評估中可得到最佳的 set F 值 72.03%，效能提升了 1.46%。

(五)、意見持有者辨識整體實驗

本實驗的目的是探討使用不同結果合併策略對意見持有者辨識效能的影響，實驗設定爲結果合併策略 A 與結果合併策略 B ，表五顯示使用不同結果合併策略的系統效能。

嚴格意見句部份的評估中策略 B 的 set F 值爲 73.40%，比策略 A 高了 2.48%，寬鬆意見句部份的評估也可以得到類似的結果。結果顯示作者意見辨識程序較擅長判斷非作者意見，可能與我們使用較多與非作者意見相關的特徵值有關，實驗結果也顯示策略 B 可以達到最佳效能。

表五、結果合併策略對意見持有者辨識效能的影響

	結果合併策略	正確數	錯誤數	set F 值
嚴格意見句	策略 A	829	340	70.92%
	策略 B	***858***	***310***	***73.40%***
寬鬆意見句	策略 A	1338	611	68.65%
	策略 B	***1372***	***576***	***70.40%***

(六)、與 NTCIR 7 參賽隊伍比較

我們將本系統效能與 NTCIR7 參賽隊伍的效能比較， NTCIR7 的評估方式分爲兩種，一種評估系統提報正確的意見句，也就是我們在意見持有者標記中使用的評估方式，另一種則評估 NTCIR7 語料中所有的意見句。

表六顯示本系統與 NTCIR 7 參賽隊伍的效能比較。 NTCIR7 意見持有者擷取評比項目的參賽隊伍包含香港中文大學、北京大學、龍捲風科技及台灣大學四隊，香港中文大學使用經驗法則方法、北京大學使用條件隨機域模型、龍捲風科技使用支援向量機、台灣大學使用決策樹演算法。

嚴格意見句部份，以系統提報正確的意見句評估，本系統的最佳效能爲 F 值 73.40%。香港中文大學的效能最佳： F 值爲 82.30%，比本系統高了 8.90%，但本系統的效能也比其他使用機器學習方法的隊伍高了 15.09%以上。以所有意見句數評估中，本系統與香港中文大學的效能差距拉近到 5.16%。比較嚴格意見句與寬鬆意見句的評估，可以發現本系統與其他系統不同，較擅長於辨識嚴格意見句的意見持有者，換句話說，本系統擅長於辨識出較無爭議的意見持有者，也就是較爲可靠的意見持有者。

表六、意見持有者辨識整體效能—與 NTCIR 7 參賽隊伍比較

	參賽隊伍	猜對意見句數	以系統猜對意見句數評估			以所有意見句評估		
			精確率	召回率	F 值	精確率	召回率	F 值
嚴格意見句	香港中文大學	757	***82.30%***	***82.30%***	***82.30%***	***19.88%***	***49.52%***	***28.38%***
	北京大學	880	57.84%	57.84%	57.84%	13.03%	40.53%	19.72%
	龍捲風科技	1213	54.91%	54.91%	54.91%	8.22%	52.95%	14.23%
	台灣大學	1169	48.16%	48.16%	48.16%	8.14%	44.90%	13.78%
	本系統	1169	**73.40%**	**73.40%**	**73.40%**	**12.38%**	**68.31%**	**20.97%**
寬鬆意見句	香港中文大學	1134	***82.54%***	***82.54%***	***82.54%***	***29.92%***	***43.05%***	***35.31%***
	北京大學	1364	58.72%	58.72%	58.72%	20.51%	36.84%	26.35%
	龍捲風科技	2070	56.47%	56.47%	56.47%	16.78%	40.02%	23.65%
	台灣大學	1948	50.31%	50.31%	50.31%	14.43%	53.73%	22.75%
	本系統	1948	**70.40%**	**70.40%**	**70.40%**	**19.80%**	**63.11%**	**30.15%**

五、結論與未來展望

本研究提出一個以機器學習方法爲基礎的意見持有者辨識方法，並且依照此方法實作出一套意見持有者辨識系統。

本研究根據意見持有者的分類將意見持有者辨識分爲作者意見辨識及意見持有者標記兩部分。在作者意見辨識中，本研究提出詞彙相關資訊、詞性相關資訊、具名實體資訊、關鍵符號資訊、文句組成資訊及意見相關資訊等特徵值，並使用支援向量機來解決此問題。在意見持有者標記中，本研究提出詞彙相關資訊、詞性相關資訊、具名實體資訊、關鍵符號資訊、文句組成資訊、前後文相關資訊及意見相關資訊等特徵值，並使用條件隨機域模型並提出不同的標記方式來解決此問題。本研究提出協同訓練來解決訓練語料過少的問題，並提出結果合併策略以提升意見持有者辨識效能。

本研究實作出一套意見持有者辨識系統。本系統在 NTCIR7 多語意見分析評比項目繁體中文語料中可以達到 F 值爲 0.734 的效能，是使用機器學習方法中效能最佳的，也相當接近目前最佳系統的效能。目前效能最佳的方法是使用經驗法則解決本問題，經驗法則較難重製與驗證，但研究者很容易就可以重製與驗證本研究提出的機器學習方法。本研究分析意見持有者辨識訓練語料中標記歧異的情形，並提出最佳的應用方式，本研究也分析系統辨識之錯誤結果，並以具名實體修復及意見持有者尾詞標記的方法來改善錯誤情況。

最終我們希望能將意見持有者辨識的結果與其它意見探勘的結果結合，整合成一套能自動擷取出意見句的意見傾向、意見強度、意見持有者及意見評論目標的意見探勘系統，以提供使用者更有用的資訊。

參考文獻

[1] S. M. Kim and E. Hovy. “Determining the sentiment of opinions.” Proceedings of the COLING conference, pp.1367-1374 , 2004

[2] B. Pang and L. Lee. “Opinion mining and sentiment analysis” Foundations and Trends in Information Retrieval, Vol. 2, pp. 1-135, 2008S. M. Kim and E. Hovy. “Determining the sentiment of opinions.” Proceedings of the COLING conference, pp.1367-1374 , 2004

[3] Y. Seki, N. Kando and M. Aono. “Multilingual opinion holder identification using author and authority viewpoints.” Journal of Information Processing and Management, pp. 189-199, 2009 [co-training]

[4] R. Xu and K. F. Wong. “Coarse-Fine opinion mining – WIA in NTCIR-7 MOAT task.” Proceedings of the Seventh NTCIR Workshop, pp. 307-313, 2008

[5] S. M. Kim and E. Hovy. “Extracting opinions, opinion holders, and topics expressed in online news media text.” Proceedings of the Workshop on Sentiment and Subjectivity in Text at the joint COLING-ACL conference, pp. 1-8, 2006

[6] Y. Kim, Y. Jung and S. H. Myaeng. “Identifying opinion holders in opinion text from online newspapers.” International Conference on Granular Computing, pp. 699-702, 2007

[7] Y. Kim, S. Kim and S. H. Myaeng. “Extracting topic-related opinions and their targets in NTCIR-7.” Proceedings of the Seventh NTCIR Workshop, pp. 247-254, 2008

[8] Y. C. Wu, L. W. Yang, J. Y. Shen, L. Y. Chen and S. T. Wu. “Tornado in multilingual opinion analysis: a transductive learning approach for Chinese sentimental polarity recognition.” Proceedings of the Seventh NTCIR Workshop, pp. 301-306, 2008

[9] E. Breck and Y. Choi and C. Cardie. “Identifying expressions of opinion in context.” Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 2683-2688, 2007

[10] Y. Choi, C. Cardie, E. Riloff and S. Patwardhan. “Identifying sources of opinions with conditional random fields and extraction patterns.” Proceedings of EMNLP conference, pp. 355-362, 2005

[11] X. Meng and H. Wang. “Detecting opinionated sentences by extracting context information.” Proceedings of the Seventh NTCIR Workshop, pp. 268-271, 2008

[12] K. Liu and J. Zhao. “NLPR at Multilingual Opinion Analysis Task in NTCIR7.” Proceedings of the Seventh NTCIR Workshop, pp. 226-231, 2008

[13] 羅永聖, “結合多類型字典與條件隨機域之中文斷詞與詞性標記系統研究” 國立台灣大學碩士論文, 2008

[14] C. C. Chang and C. J. Lin. “LIBSVM: a library for support vector machines”, 2001

[15] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz and T. Euler. “YALE: rapid prototyping for complex data mining tasks” In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 935-940, 2006

[16] J. Lafferty, A. McCallum, F. Pereira. “Conditional random fields: probabilistic models for segmenting and labeling sequence data” In Proceedings of International Conference on Machine Learning, pp. 282-289, 2001

[17] T. Kudo, “CRF++: yet another CRF toolkit.” http://crfpp.sourceforge.net/ , 2003

[18] A. Blum and T. Mitchell. “Combining labeled and unlabeled data with co-training.” Conference on Computational Learning Theory, pp. 92-100, 1998

[19] Y. Seki, D. K. Evans, L. W. Ku, L. Sun, H. H. Chen and N. Kando. ”Overview of multilingual opinion analysis task at NTCIR-7.” Proceedings of the Seventh NTCIR Workshop, pp.185-203, 2008

[20] L. W. Ku and H. H. Chen. “Mining opinions from the web: beyond relevance retrieval.” Journal of American Society for Information Science and Technology, pp. 1838-1850, 2007