柯南

C467

鄭子曰：欲查其人，先鑑其文；

用現代自然語言處理算法鑑定古文的作者

假定您在中央電視臺看到的某一篇新聞引用孔子的一句話，“孔子曰：脚踏實地者，難以穿上褲子也。”這句奇怪話肯定會引起您的注意，但您怎麽會調查該句話的來源何在抑或孔子到底有沒有説這句話呢？現代論文的作者會在論文或書上很明顯地寫上自己的名字，但如果找不到的話，就可以在網絡上進行搜索來找出更多的和作者相關的信息。可是古文不是這樣的，有的沒有記下作者的名字，有的古文大家說有兩三個作者。傳統的鑑定方法主要是深入的調查寫作風格（Juola, 2012），最近也有人把傳統的方法同統計算法以及自然語言處理的分析結合來提高效率同精確性。非自然語言處理、非計算密集型的、和很簡單的作者鑑定項目已經存在，包括荷蘭語、英語、希臘語、西班牙語、以及德文的文章（Halvani et.al., 2016），此外，也有現代漢語的作弊偵查的模型（Bao et.al., 2003），但沒有通過這樣通過自然語言處理的方法鑑定古文作者的項目。因此，該論文的研究目的是是否可以通過現代自然語言處理鑑定方法來鑑定古文作者。

如何鑑定作者可以分成幾個部分：通過可能的作者的組以及這些作者寫的幾篇文章來訓練作者鑑定程序，接下來分析目標文件。根據 Stamatatos 的《現代作者鑑定算法的眺望》，本領域有兩個主要的做法：以作者特徵為基礎的算法（profile-based）和以文本特徵為基礎的算法（instance-based）（也有一個混合算法），主要的不同在於第一個算法從作者的角度來應對問題，第二個算法從文章的角度來應對問題。以作作者特徵基礎的算法包括給每個作者構建一個“作

者輪廓”，然後比較目標文件以及這個作者輪廓。以文本特徵爲基礎的算法包括根據這些文章分別構建作者不同的寫作風格（Stamatatos, 2009）。本論文打算通過以作者簡介爲基礎的角度，并且利用 Sari 和 Stevenson（2016）所寫的《調查詞語嵌入及文字 n-元語法來做作者聚類》裏面所用的作者聚類算法，主要是文字 n-元語法。如果讀者對於詞語嵌入方法感興趣的話，可以看附錄。

討論本論文的算法之前需要先解釋幾個詞語。通用依存（Universal Dependencies 或 UD）指的是某些語言學家所製造的自然語言處理數據庫，我打算利用京都大學注釋的文言文數據庫。UD 所提供的 UDpipe 軟件主要有分詞和注釋功能；訓練模型之後，可以輸入原本的文言文，接下來它會自動注釋並輸出注釋的數據，以後可以用來分析。以上就是最基本的預處理的部分。n-元語法指的是把某個句子分成幾部分，比如説“我非常熱愛學中文”的 3 元語法就是“我非常”、“非常熱”、“常熱愛”、“熱愛學”等。這樣不只是分別看詞語，而是能夠看順序以及其他方面的信息。詞頻-逆文本頻率指數，或是説 TfIdf（Term frequency – inverse document frequency），主要是文本挖掘常用的加權技術，也就是説能夠分析並向量量化某篇文章，接下來這個向量會反映什麽詞語在該文章最重要等等。K-平均算法（k-means）是一種聚類分析方法，主要是輸入數據點之後，它會多次嘗試把它們分成幾個不同的組，最後會聚類成幾個最合適的組。F1 分數（BCubed F-Score）是一種聚類的統計分析算法，能夠輸出某個聚類的精確性。

本研究的算法如下：從 Ctexts 網站下載一批古文的文章，接下來利用通用依存（ Universal Dependencies treebanks）所提出的文言文訓練數據來訓練 UDpipe 的分詞及令牌的模型，這樣能夠預處理文章。預處理文章之後，可以摘錄文字 n-元語法並利用 SciKit 所提出的 TFIDF 向量量化機器來摘錄詞頻-逆文本頻率指數的向量。詞頻-逆文本頻率指數摘錄好了之後，就可以用 K-平均算法的成組算法來分析所下載的文章，進行作者聚類，接下來會利用 F1 分數來評估聚類算法的表現以及作者輪廓描述的質量。

至今，我只做完 n-元語法部分的分析，不過，K 平均算法還沒有完全優化；主要是因爲 K 平均算法輸出平均的 F1 分數是 0.3 左右，最低是 0.000 以及最高才是 0.64。但如果不標準化 TFIDF 向量，F1 分數會達到並超過所模仿的論文的 0.76。因此，兩個可以進步的地方在於輸入數據以及 K 平均算法的優化。首先，我所用的數據不夠健全；劉向寫的文件的比重多於其他的作者，倘若能夠獲得更多的輸入數據就應該能夠提高 K 平均算法的效率。其次，同一個參數的 K 平均算法會輸出不同的作者聚類；這樣其實是挺合理的，原因在於 K 平均算法每次會隨機選擇開始點，接下來會輸出不同的作者聚類，但整個項目每一部分有許多參數，因此有可能還沒有達到最有效的配置。回到原來的研究問題，我們可以通過現代的自然語言處理方式來鑑定古文的作者，這些自然語言處理的算法的潛力很大，優化的 K 平均算法可以比較準確的進行聚類。莊子的作品具有爭議性，古文學者同意內篇是莊子寫的，但外篇是否莊子寫的，還沒達到共識。如果用這個 N-元語法的 K 平均算法最好的表現，分析的三個莊子的外篇作品，其中兩個跟莊子內篇的幾個作品聚類在一起。總的來說，這些自然語言處理的作者鑑定算法具有很大潛力，可以用來補充傳統的鑑定法。

附錄:

Word2vec 是一個學習算法，根據輸入的文章計算出這些詞語展現在什麼樣的環境，然後用這些伴發統計數據做出詞語嵌入（word embeddings）。COS 相似性距離的公式（Cosine Similarity Distance）在詞語嵌入的範圍下很重要，主要是因爲詞語嵌入跟伴發統計數據有關係，因此如果用簡單的距離公式的話，就會説"的"和每一個詞語意義上很密切，但是這個餘弦相似性距離的公式會用兩個向量之間的角度來決定意義是否相同的。

倘若有時間的話，也會訓練一個古文的 word2vec 模型，並利用這個模型以及 K-平均算法重新分析並聚類。做好了之後，可以利用餘弦相似性距離的公式來進行言談意義分析。

文獻:

Bao, J. P., Shen, J. Y., Liu, X. D., & Song, Q. B. (2003). A survey on natural language text copy detection. Journal of software, 14(10), 1753-1760.

Halvani, O., Winter, C., & Pflug, A. (2016). Authorship verification for different languages, genres and topics. Digital Investigation, 16, S33-S43.

Juola, P., An Overview of the Traditional Authorship Attribution Subtask, Notebook for PAN at CLEF 2012.

Sari, Y., & Stevenson, M. (2016). Exploring Word Embeddings and Character N-Grams for Author Clustering. In CLEF (Working Notes) (pp. 984-991).

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, Xiaoyong Du, Analogical Reasoning on Chinese Morphological and Semantic Relations, ACL 2018.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, 60(3), 538-556.

Sturgeon, D., Chinese Text Project: a dynamic digital library of premodern Chinese, Digital Scholarship in the Humanities 2019/

Yasuoka, Koichi: Universal Dependencies Treebank of the Four Books in Classical Chinese, DADH2019: 10th International Conference of Digital Archives and Digital Humanities (December 2019).