漢文文言文史文本的數位化、斷詞、分句與資訊擷取

Optical Character Recognition, Word Segmentation, Sentence Segmentation, and Information Extraction for Historical and Literature Texts in Classical Chinese

劉昭麟 Chao-Lin Liu
國立政治大學資訊科學系
Department of Computer Science
National Chengchi University, Taiwan
chaolin@g.nccu.edu.tw

摘要

文言文本是研究二十世紀之前中國歷史、社會與文學議題的主要素材。近年以來語言分析技術的進展大都集中於白話文本。在這一簡短的報告中，我們介紹一些應用自然語言處理、機器學習(包含深度學習)技術以從文言文資料源抽取有用資訊的數位人文研究議題[4]。

自動化的軟體服務，讓我們得以以較低的人力代價，從大量的歷史文獻中擷取傳記資料[8]，進而建構諸如 CBDB[2]、TBDB[1][14]、CBETA[3][6] 和 DocuSky[5] 等專業服務。例如，人物和地名是史學研究的重要根基[11][15]；在獲得大量的人物和地名等資訊之後，我們可以嘗試推測文言文的句法，進一步增進資訊擷取的能力[11]。諸如條件隨機場和深度學習等技術讓研究者比較容易發覺大量文獻中這一些命名實體的資訊；進階的關係推演技術，使得推論文本中人際關係和社會網絡的難度大幅降低[13]。

數位人文分析技術同時也讓我們可以透過分析大量的傳統漢詩來探索詩人內心的語言和文學世界[12]。從統計面來說，我們發現了歷代漢詩的用字分佈呈現了齊夫分佈。借助量化分析的取徑，我們也可以觀察詩人如何巧妙地安排和組合優美字詞，來營造讀者所感受到的美學意境。

以上提到的這一些研究，雖然都已經有不錯的進展，但是計算工具所傳回的結果，都還需要專業人員相當程度的精度來驗證和詮釋。原始資料源中的資料，不僅僅詞彙之間沒有明顯邊界，就連句子之間也沒有現代人所熟知的標點標記。要精確了解這一些連續漢字的古文並不容易。為了提高人類專家分析這一些語料的效能，為文言資料進行斷詞和分句是非常基礎的研究工作[7][9][10]。事實上，尚有許多文言語料目前只有紙本資料，還沒有完全的數位化；需要以人工繕打或者以計算方法進行文字辨識。由於文言文書印刷與書寫的方式、排版的方式有諸多變化，文本數位化的自動化仍然有許多工作仍待解決。

Abstract

Texts written in classical Chinese are the major sources for studying the Chinese history, society, and literature, particularly for the periods before the twentieth century. Recent advances in language technology focus mostly on the analysis of modern mandarin Chinese, or vernacular Chinese. In this brief presentation, we shall go through some digital-humanities topics of applying techniques of natural language processing and machine-learning methods, including deep learning, for extracting useful information from sources of classical Chinese texts [4].

The abilities to recognize and extract named entities (NEs) such as person and place names [15] and to infer about personal relationships and their social networks [13] are fundamental for assisting historical research. We may apply conditional-random-field models or deep learning methods to extract the NEs, and attempt to infer the grammar of classical Chinese when sufficient samples are available [11]. By reasonably automating the extraction of biographical information from historical documents [8], domain experts can build such research-oriented databases and platforms as CBDB [2], CBETA [3][6], DocuSky [5], and TBDB [1][14] at affordable costs.

With the availability of ample text collections of classical Chinese poems, we can explore the linguistic and literature worlds of poets with the help of computational tools [7][12]. We were not surprised to find that the distributions over Chinese characters in classical Chinese poems follow the Zipf's law. We may even study how poets organized their words to induce aesthetic imagery in the minds of their readers, from a certain analytical perspectives.

To verify the raw findings reported by algorithmic procedures still require non-negligible amount of close reading by human experts. We have not achieved the results mentioned completely automatically yet. The original classical Chinese texts do not have delimiters between consecutive words [7][10]. Unless added by human experts, most of the original classical texts do not use modern punctuation marks either, so we need to split sentence segments as well [9][10]. Furthermore, there are still a myriad of printed books in classical Chinese that need to be manually typed or algorithmically digitized, and optical character recognition for many ancient books remains a practical challenge due to the wide variety of page layouts and writing/printing styles.

致謝 (Acknowledgments)

Liu was supported in part by the contracts MOST-104-2221-E-004-005-MY3 and MOST-107-2200-E-004-009-MY3 of the Ministry of Science and Technology of Taiwan and in part by the mini-project 109H124D-09 of the National Chengchi University in Taiwan. The Chinese Biographical Database Project of Harvard University and the Harvard-Yenching library provide the data for some of the reported research, and we have discussed our work in OCR with Professor Donald Sturgeon of the Durham University and with a research team at the Academia Sinica Center for Digital Cultures (中研院數位文化中心) before.

References

[1] S.-b. Chang, T.-h. Lee, Y.-l. Lee, C.-j. Li, Y.-w. Ku, H.-R. Ke, and S.-H. Sie. From CBDB to TBDB: The "Treatise of Historical Figures" of the new edition of Changhua Local Gazetteer as a starting point, *J. of Digital Archives and Digital Humanities*, **2**:91–115, 2018.

[2] China Biographical Database Project (CBDB): https://projects.iq.harvard.edu/cbdb/home

[3] Chinese Buddhist Electrical Text Association (CBETA): https://www.cbeta.org/

[4] J. Hsiang. Editorial for the inaugural issue: From digital archiving to digital humanities, *J. of Digital Archives and Digital Humanities*, **1**:i–iv, 2018.

[5] I M. Hung, C. Hu, and J. Hsiang. Exploring Guangxu-era missionary activities in Taiwan from Chinese Recorder, Dan-Hsin Archives and Ming-Qing Taiwan Administrative Archives through DocuSky, *Proc. of the 2020 Int'l Conf. on Digital Humanities*, 2020.

[6] J.-J. Hung. CBETA research platform: A digital tool for studying Chinese Buddhist texts in the new era, *J. of Digital Archives and Digital Humanities*, **1**:149–174, 2018.

[7] C.-L. Liu and W.-T. Chang. Onto word segmentation of the Complete Tang Poems, *Proc. of the 2019 Intl Conf. on Digital Humanities*, 2019.

[8] C.-L. Liu, W.-T. Chang, T.-Y. Zheng, and P.-S. Chiu. Toward building chronicles from biographies in local gazetteers: An application of syntactic and dependency parsing, *Proc. of the 2019 Int'l Conf. on Digital Humanities*, 2019.

[9] C.-L. Liu and Y. Chang. Classical Chinese sentence segmentation for tomb biographies of Tang dynasty, *Proc. of the 2018 Int'l Conf. on Digital Humanities*, 231–235, 2018.

[10] C.-L. Liu, C.-T. Chu, W.-T. Chang, T.-Y. Zheng. When classical Chinese meets machine learning: Explaining the relative performances of word and sentence segmentation tasks, *Proc. of the 2020 Int'l Conf. on Digital Humanities*, 2020.

[11] C.-L. Liu, C.-K. Huang, H. Wang, and P. K. Bol. Mining local gazetteers of literary Chinese with CRF and pattern-based methods for biographical information in Chinese history, *Proc. of the 3rd Workshop on Big Humanities Data*, 2015 IEEE Int'l Conf. on Big Data, 1629–1638, 2015.

[12] C.-L. Liu, T. J. Mazanec, and J. R. Tharsen. Exploring Chinese poetry with digital assistance: Examples from linguistic, literary, and historical viewpoints, *J. of Chinese Literature and Culture* (a special issue on Digital Methods and Traditional Chinese Literary Studies), **5**(2):276–321, 2018.

[13] C.-L. Liu and H. Wang. Matrix and graph operations for relationship inference: An illustration with the kinship inference in the China biographical database, *Proc. of the 2017 Annual Meeting of the Japanese Assoc. for Digital Humanities*, 94–96, 2017.

[14] Taiwan Biography Database (TBDB): http://tbdb.ntnu.edu.tw/

[15] T.-H. R. Tsai, C.-H. Wu, P.-L. Pai, and I-C. Fan. Automatic labeled data generation for person named entity disambiguation on the Ming Shilu, *Proc. of the 2020 Int'l Conf. on Digital Humanities*, 2020.