自然語言處理與數位人文

Natural Language Processing for Digital Humanities

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

洪振洲 Jen-Joe Hung
法鼓文理學院佛教學系
Department of Buddhist Studies
Dharma Drum Institute of Liberal Arts, Taiwan
jenjou.hung@dila.edu.tw

張素玢 Su-bing Chang
國立臺灣師範大學臺灣史研究所
Graduate Institute of Taiwan History
National Taiwan Normal University, Taiwan
109682@ntnu.edu.tw

吳宛怡 Wu, wan-yi
香港理工大學中國文化學系
Department of Chinese Culture
The Hong Kong Polytechnic University, Hong Kong
wan.yi.wu@polyu.edu.hk

摘要

數位人文近十幾年以來在國際學術界已然蓬勃發展。相對之下，我國計算語言學界參與數位人文研究的程度並不如一般所預期。這一個座談會藉著簡短介紹四個具備數位人文屬性的研究工作，希望能夠吸引更多計算語言學界的學者參與數位人文的研究。中華電子佛典學會的大正藏網路資料庫 (CBETA) 的內容與附屬的服務都是免費的。洪振洲將分享建構這一個資料庫的過程中，語文分析科技包含人工智慧與自然語言處理技術等，的相關應用。歷史人物的傳記資料是歷史研究的重要基石，張素玢會介紹「臺灣傳記人物資料庫」 (TBDB) 的研究團隊從多個臺灣的地方志中抽取個人傳記資料來建構 TBDB 的經驗和展望。戲劇是中國傳統藝術的重要部分，相關資料以許多不同形式留存至今。吳宛怡將分享關於中國戲劇的研究經驗，並且討論語文分析技術對於各種戲劇研究的潛在貢獻機會。如果時間允許，劉昭麟將介紹漢文古文書數位化的文字辨識工作，

文言歷史文本的分句工作，諸如《全唐詩》、《全宋詩》和《全臺詩》等格律詩的斷詞工作，還有從文言史學資料中抽取有用資訊，例如個人傳記資料，的相關經驗。

Abstract

The research of digital humanities has flourished in the past decade internationally. In contrast, the participation of researchers of computational linguistics in domestic research projects remains less common than one may have anticipated. The goal of this panel is to introduce sample research projects of digital humanities to the community of computational linguistics, hoping to promote further cooperation between the two communities. The panel consists of four parts. Hung introduces the online repository of the Taishō Tripiṭaka that the Chinese Buddhist Electronic Text Association (often called CBETA) offers. Applications of language technology, including artificial intelligence and natural language processing, for the construction of CBETA will be discussed. Chang and her colleagues aim to build the Taiwan Biographical Database (TBDB), which, in the long term, will serve as part of the bedrock for historical studies about Taiwan. Experience about how the research team extracted and integrated the information from some collections of local gazetteers to build the TBDB will be discusses. Dramas are important part of Chinese arts. Relevant materials about dramas are available in some different databases and in different forms. Wu will share with us her study on Chinese dramas, and elaborates on the potential contributions of language technology to the studies of Chinese dramas. If time allows, Liu plans to outline his work on optical character recognition (OCR) for ancient Chinese documents, sentence segmentation for classical Chinese, word segmentation for classical Chinese poems, including the Tang and Song poems, and information extraction from historical documents in classical Chinese.