自然語言處理技術於數位人文領域的機會與挑戰 --以佛教經典研究為例

The Opportunities and Challenges of Natural Language Processing Technology in the Field of Digital Humanities – Taking the Study of Buddhist Scriptures as an Example

洪振洲 Jen-Jou 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

摘要

佛教自東漢傳入中國以來，歷經長時間的發展，成為人民主要的信仰之一。佛教在中國發展的千年時間中，不僅發展出獨特的漢傳佛教風格，同時間，佛教也融合成為了中華文化的一部分。而佛教傳入中國後，另一個重要的產出，便是引發了歷經千年的佛經翻譯活動，因此產生了大量的漢文佛典文獻。這些佛教典籍，經由歷代僧人精心創作、發展、甄別、校對、編排與整理，形成我們今日所見的漢文大藏經。漢文大藏經中內收錄的大量佛典文獻衍生出多元並存的部類與思想體系，表現於外在書目有著一定的規範結構組織，其編輯的內在理路，包括編藏目的、典籍選編入藏的標準、編纂藏經方法，通常也反映著一個時代的知識構成與跨時代的知識演變。

隨著國家推動數位典藏發展，漢文大藏經及相關之大量佛教典籍已陸續數位化，並製造出多項佛學資料庫。此項成果，讓佛學、人文學科以及各領域的研究者可以前所未見的便利方式，取得佛經內容與相關參考資料。然而，人文學者過去以傳統方式運用網路環境蒐集資料、進行研究工作，如果研究主題涉及寬廣的背景知識，以往只能以關鍵詞搜尋到單一而零碎的訊息，需要將許多碎片化的資料來源，逐項解讀、判斷與比對，才能整合組成具有關連意義的背景資料。隨著大量資料在網際網路湧現，以關鍵詞搜尋資料的傳統方式，不僅耗費大量人工時間，也不免出現毫無意義或重覆訊息，乃至天壤地別的錯誤結果。

人工智慧領域中之自然語言處理之研究不斷推進，對於文字內容裡的重要資料擷取、語法及語意分析、文本生成與自動問答等技術持續發展，並且隨著深度學習方法的導入，在許多應用上得到令人欣喜的突破。然此類自然語言處理之技術發展，由於語料來源與商業利益之考量，多半仍集中於現代語言之處理，對於古典語言文獻處理上仍未得到廣泛的關注，也缺乏合適之自然語言處理工具與訓練語料集。如能將現前已深度發展之自然語言處理技術，帶入漢文佛典文獻之處理，前述之人文學者進行研究時所遭遇之問題將可望得到有效之解決。目前應用人工智慧技術於佛典數位資料研究，雖有獲取少數成功案例，但仍屬萌芽階段。本演講將描述佛教研究領域的數位資料建置現況、簡單說明

傳統佛教研究學者關心之研究議題，以及自然語言處理技術應用於此研究領域的發展與成果。藉此希望引發更多共鳴，以吸引更多計算語言學界的研究者參與數位人文的研究。

Abstract

Since Buddhism was introduced into China in the Eastern Han Dynasty, it has undergone long-term development and has become one of the main beliefs of the people. During the millennia of development of Buddhism in China, not only did it develop a unique style of Chinese Buddhism, but at the same time, Buddhism also became part of Chinese culture. After Buddhism was introduced to China, another important effect was to trigger a millennium-long process of Buddhist scripture translation, which resulted in a large number of Chinese Buddhist texts. These Buddhist scriptures, through the careful creation, development, screening, proofreading, arrangement and collation by monks of the past dynasties, form the Chinese canon that we see today. The large number of Buddhist texts included in the Chinese canon derives from the coexistence of diverse divisions and ideological systems, which is reflected in the external bibliography having a certain standardized structure and organization, and the internal rationale of its editing, including the purpose of the collection, the standards of the selection and inclusion of scriptures, and the method of compiling canonical scriptures usually reflects the knowledge structure of one era and the evolution of knowledge across the ages.

As the country promotes the development of digital collections, the Chinese canon and a large number of related Buddhist scriptures have been digitized one after another, and a number of Buddhist studies databases have been created. This achievement allows researchers in Buddhism, the humanities, and various other fields to access the content of Buddhist scriptures and related reference materials in a convenient way that has never been seen before. However, humanities scholars use traditional methods to collect data and conduct research in the Internet environment. If the research topic involves broad background knowledge, in the past, only single and fragmented pieces of information could be searched for by using keywords, and many fragmented pieces of data were needed. Sources, item-by-item interpretation, assessment, and comparison can be integrated to form relevant background information. With the emergence of a large amount of information on the Internet, the traditional way of searching for information with keywords not only consumes a lot of time, but also inevitably produces meaningless or repeated messages, and even widely erroneous results.

Natural language processing research in the field of artificial intelligence is constantly advancing. The technology of capturing important data in text content, grammar and semantic analysis, text generation, and automatic question answering continues to develop, and with the introduction of deep learning methods, it has been applied in many ways to provide gratifying breakthroughs. However, the technological development of this type of natural language processing, due to the source of the corpus and the consideration of commercial interests, is mostly still focused on the processing of modern languages. The processing of classical language documents has not yet received widespread attention, and there is a lack of suitable natural language processing tools and training corpus. If the previously developed natural language processing technology can be brought into the processing of Chinese Buddhist texts,

the foregoing problems encountered by humanities scholars in their research will be expected to be effectively solved. At present, the application of artificial intelligence technology in the research of digital data of Buddhist scriptures has obtained a few successful cases, but it is still in its infancy. This lecture will describe the current situation of digital data construction in the field of Buddhist research, and briefly explain the research topics that traditional Buddhist scholars are concerned with, as well as the development and results of natural language processing technology in this research field. It hopes to arouse greater interest and attract more researchers in computational linguistics to participate in the research of digital humanities.

致謝 (Acknowledgments)

This research was supported in part by the contracts MOST-106-2420-H-655-001-MY3 and MOST-106-2420-H-655 -002 -MY3 of the Ministry of Science and Technology of Taiwan.