藏漢佛教語料品目之自動對列

陳光華[*]、闕慧貞[**]、李家名[***]、唐國銘[****]、黃乾綱[*****]

摘要

藏漢佛學研究的一項重要課題是從語言學和文獻學角度比對與勘定藏漢佛教文獻。佛學研究發展至今，卻依然未能建立佛教文獻文本的準確性和可靠性。在梵文原典散佚而所剩無幾的情況下，研究藏漢佛教文獻無疑是揭示藏漢譯文的種種闕漏，釐定藏漢文譯本的較為可行的道路。然而，藏譯或是漢譯佛經在千餘年的流傳過程，也出現了種種版本學上的問題，僅僅依靠個別譯本佛經本身的勘定難以解決問題，必須透過相應的不同譯本作為參照。然而，傳統上這樣的比對工作卻僅能仰賴佛教文獻學者親力親為，花費大量人力與時間，卻僅能進行小規模的研究工作。基於前述的現象，本研究發展一套自動對列藏漢語料的方法，在文獻的品目層次，對列藏文佛教文獻與漢文佛教文獻，以降低佛教文獻學者在整理研究文獻的時間成本與人力成本，而將研究重心放在電腦系統無法取代的文獻校勘與經典譯注。本研究主要係基於資訊檢索（Information Retrieval，簡稱 IR）及計算語言學（Computational Linguistics，簡稱 CL）的相關理論及技術，使用藏漢雙語詞典，建立向量空間的運算模型。實驗語料《法華經》的藏文版與漢文版分別取自臺北版之藏譯大藏經與 CBETA 版之漢譯大藏經。為了探討停用詞與雙語詞典對於運算模式的影響，本研究使用二部不同類型的雙語詞典：張怡蓀編《藏漢大詞典》通用綜合詞典和榊亮三郎整理之《翻譯名義大集》專業佛學詞典，並因應停用詞的使用與否發展二套運算模式。實驗結果顯示採用 vector-space model，搭配 CKIP 中文斷詞處理、使用專業佛學詞典的實驗設定，可以在前二個候選品目找到真正的對應品目；簡單的 n-gram matching 方法，搭配專業佛學詞典，平均而言，也可以在前三個候選品目找到真正的對應品目。這樣的實驗結果顯示專業藏漢佛學詞典對於處理不同譯本對列問題的重要性；此外，停用詞僅

[*] 國立臺灣大學圖書資訊學系教授，通訊作者，Email：khchen@ntu.edu.tw。
[**] 法鼓佛教學院碩士生。
[***] 國立臺灣大學工程科學暨海洋工程學系博士生。
[****] 國立臺灣大學工程科學暨海洋工程學系博士生。
[*****] 國立臺灣大學工程科學暨海洋工程學系副教授。

有在 n-gram matching 方法，有比較大的影響。綜言之，本研究的結論是不同語言譯本的佛教文獻品目層次的自動對列是可行的。

關鍵字：自動對列、佛教語料、CBETA、漢文、藏文

Automatic Chapter-level Alignment for Tibetan and Chinese Buddhist Texts

Kuang-hua Chen[*], Hui-chen Chueh[**], Chia-ming Lee[***],

Kuo-ming Tang[****], Chien-kang Huang[*****]

Abstract

One important research issue of Tibetan and Chinese Buddhist studies is the comparison and demarcation for Tibetan and Chinese Buddhist texts in perspectives of linguistics and philology. However, Buddhist studies still failed to establish the precision and reliability on its texts. In fact, few Sanskrit scriptures have been remained nowadays. The study on Tibetan and Chinese Buddhist texts with no doubt is one possibly practical way to identifying gaps and demarcating translations of Tibetan and Chinese. Buddhist texts have been spread for thousands of years. There exist a lot of problems due to different translations or versions. It is very difficult to solve problems by examining individual translation only. On the contrary, we have to iteratively investigate different translations or version as cross-references. However, such kind of work traditionally relied on Buddhist scholars themselves. It took a lot of human power and time but was only practical for small-scale researches. Based on the aforementioned phenomena, this study proposed an approach of automatic chapter-level alignment for Tibetan and Chinese texts. The purpose is to reduce the time cost and human cost in processing texts and to allow Buddhist scholars focusing on demarcation and annotation of Buddhist texts that cannot be done by computer systems. We applied Tibetan-Chinese dictionaries and built vector-space processing models based on related theories

[*] Professor, Department of Library and Information Science, National Taiwan University. Corresponding Author, Email: khchen@ntu.edu.tw.
[**] Master Student, Dharma Drum Buddhist College.
[***] Ph.D. Student, Department of Engineering Science and Ocean Engineering, National Taiwan University.
[****] Ph.D. Student, Department of Engineering Science and Ocean Engineering, National Taiwan University.
[*****] Associate Professor, Department of Engineering Science and Ocean Engineering, National Taiwan University.

and techniques of information retrieval (IR) and computational linguistics (CL). The Tibetan and Chinese testing Buddhist texts, Saddharma-puṇḍarīka sutra, were collected from Taipei edition of Tibetan Tripitaka of Saddharmapuṇḍarīka Databaseand and The Taishō Shinshū Daizōkyō of Chinese Tripitaka of CBETA, respectively. In addition, the effects of stop words and bilingual dictionary to the proposed approach were investigated. Two types of bilingual dictionaries, Tibetan-Chinese Great Dictionary by Zhang Yi-Sun (a general dictionary) and Mahāvyutpatti by Ryozaburo Sakaki (a professional dictionary), were used in this study. Two models with/without using stop words were implemented and then compared as well. The experimental results showed that the proposed model with CKIP segmentation tool and professional Buddhist dictionary demonstrated its satisfied performance in finding true aligned chapter within Top 2 candidates. In contrast, simple n-gram matching with professional Buddhist dictionary also returned true aligned chapter within Top 3 candidates. It concluded that an appropriate professional Buddhist dictionary had its key role in Buddhist chapter-level alignment. In addition, stop-word list only showed its effectiveness in simple n-gram matching. To sum up, automatic chapter-level alignment for Tibetan and Chinese Buddhist Texts is feasible.

Keywords: Automatic Alignment, Buddhist Texts, CBETA, Chinese, Tibetan