漢文文獻之外來語音譯詞擷取方法

王昱鈞[1]、呂翊瑄[**]、蔡宗翰[***]、劉青峰[****]、金觀濤[*****]、劉昭麟[******]

摘要

在古文獻研究中，文獻中所出現之音譯詞是不同文化接觸交流之重要紀錄。特別於十八、十九世紀時期，由於交通與工業技術的突破性成長，促使了東西方文化之間產生史所未見的接觸與衝擊。在此時期的東方漢文古文獻中，針對西方諸國帶入的人事物與新概念皆嘗試以音譯詞的方式進行翻譯，成為研究該時期之語言與文化接觸之珍貴資料。然漢文文獻卷秩浩繁，為協助文史學家深入研究，利用資訊技術自大量的漢文文本中擷取出來自外來語言之音譯詞已成為迫切的需求。本論文提出一整合的音譯詞擷取方法，主要分為詞彙抽取、候選詞過濾與音譯詞分類排序三部分。在詞彙抽取上，我們採用後綴數組抽詞方法對於整部漢文文獻擷取出可能的詞彙。後綴數組抽詞之候選詞結果仍多非音譯詞，故先以規則式方法刪除顯非音譯詞之候選詞。另再以同時期之數部傳統文學文獻進行後綴數組抽詞，將其抽詞結果與原文獻之抽詞結果進行差集運算，藉以過濾更多非音譯詞之詞彙。為提供文史學家更佳之研究工具，針對抽詞過濾結果之候選詞，結合候選詞於文獻原文語句中出現之特性，進行候選詞的分類排序，以篩選較具研究價值之音譯詞。我們以十九世紀介紹域外諸國之漢文文獻《海國圖志》一書作為實驗資料來評估實驗結果，為驗證我們方法可有效抽取出之大量之音譯詞，以前人所整理之音譯詞集作為答案評估召回率，其召回率達 76.54%。為評估音譯詞排序之有效性，以人工標註排序之結果計算其平均準確率，其準確率達 96.14%，此皆說明該音譯詞擷取方法為相當有效的。

關鍵字：音譯詞、詞彙擷取、漢文文獻處理

[1] 國立臺灣大學資訊工程研究所博士研究生，Email：d97023@csie.ntu.edu.tw。
[**] 元智大學資訊工程學系大學部專題生，Email：s983322@mail.yzu.edu.tw。
[***] 元智大學資訊工程學系副教授，通訊作者，Email：thtsai@saturn.yzu.edu.tw。
[****] 香港中文大學當代中國文化研究中心研究員，Email：qingfeng@cuhk.edu.hk。
[*****] 國立政治大學講座教授，Email：gtqf1908@gmail.com。
[******] 國立政治大學資訊科學系教授，Email：chaolin@nccu.edu.tw。

Transliteration Extraction Methods for Historical Chinese Literature

Yu-chun Wang[2], Yi-hsuan Lu[**], Richard Tzong-han Tsai[***], Qing-feng Liu[****], Guan-tao Jin[*****], Chao-lin Liu[******]

Abstract

Transliteration terms in classical literature are key records that show the contact and interchange between different cultures. Especially in pre-modern period, from 18th to 19th century, a number of Chinese scholars compiled many classical Chinese books to introduce new ideas and technologies from the West. Many new concepts and named entities are transliterated into Chinese. However, a lot of transliterations in the pre-modern Chinese books are different from the ones in modern Chinese. Therefore, extracting transliterations from classical Chinese literature precisely is important to humanity researchers. We propose a transliteration extraction method for classical Chinese literature, comprising three modules: term extraction, term filtering, and candidate ranking. First, we adopt suffix array method to extract potential terms from the text. Then, the extracted terms are filtered by rules and complement set from different literature texts to obtain the possible transliteration candidates. Next, these candidates are classified by a maximum entropy model to select the most plausible transliteration candidates. We take a famous classical Chinese literature "Illustrated Treatise on the Maritime Kingdoms" to evaluate our method. The recall of extracting transliteration is

[2] Ph.D Candidate, Department of Computer Science and Information Engineering, National Taiwan University, Email: d97023@csie.ntu.edu.tw.
[**] Undergraduate student, Department of Computer Science and Engineering, Yuan Ze University, E-mail: s983322@mail.yzu.edu.tw.
[***] Associate Professor, Department of Computer Science and Engineering, Yuan Ze University, Corresponding Author, E-mail: thtsai@saturn.yzu.edu.tw.
[****] Researcher, Institute of Chinese Studies, The Chinese University of Hong Kong, E-mail: qingfeng@cuhk.edu.hk.
[*****] Chair Professor, National Chengchi University, E-mail: gtqf1908@gmail.com.
[******] Professor, Department of Computer Science, National Chengchi University, E-mail: chaolin@nccu.edu.tw.

up to 76.54% on the ground truth created by former researchers and the average precision is up to 96.14% on our human-tagged data set. The results show our method can extract and rank the transliterations effectively.

Keywords: Transliteration Extraction, Term Extraction, Historical Chinese Literature Processing