自動擷取中文典籍中人名之嘗試：以 PMI（Pointwise Mutual Information）斷詞於《資治通鑑》的應用為例

彭維謙[*]、劉士綱[**]、杜協昌[***]、翁稷安[****]、項潔[*****]

摘要

在數位人文領域裡，文字處理一直是重要的研究課題，如何針對文本特性找出其中的專有名詞更是學界所關注的重點項目之一。以歷史文本而言，「人」是極具代表性與研究價值的關鍵，因此在數位人文研究中，對於電子檔中人名的標記尤其重要。本文試圖提出一個自動演算法來擷取人名，首先利用 PMI（Pointwise Mutual Information）公式，對研究文本進行斷詞，配合規則找出候選人名，然後針對文本人的人名特性，進行人名驗證（The validation of the names）。本方法利用斷詞部份的處理來保證較高的召回率（Recall），並在人名驗證的部份進行準確度（Precision）的提高 ，比如加入人名的特徵、前後綴詞的機率分佈等，強化斷詞時發掘人名的機率。我們希望這會是個具通用意義的演算法，只要因應文本的特性略作微部調整，便可達到針對個別文本的最佳化。本論文將以《資治通鑑》為實驗對象，標記人名，進而探討本演算法的效能，以及討論其中的優點與缺點。

關鍵字：自動化、PMI、人名驗證、人名辨識、《資治通鑑》

[*] 國立臺灣大學資訊工程學系碩士生。
[**] 國立臺灣大學資訊工程學系碩士。
[***] 國立臺灣大學資訊工程系博士後研究員，Email：tu@turing.csie.ntu.edu.tw。
[****] 國立臺灣大學數位典藏研究發展中心碩士後研究員，E-mail：r90123005@ntu.edu.tw。
[*****] 國立臺灣大學資訊工程學系特聘教授，E-mail：jhsiang@ntu.edu.tw。

Automated Name-extraction in the Chinese Classics: Using PMI (Pointwise Mutual Information) Segmentation in "Zizhi Tongjian"

Wai-him Pang[*], Shih-gang Liu[**], Hsieh-chang Tu[***], Ghi-an Weng[****], Jieh Hsiang[*****]

Abstract

Text processing is an essential topic in the digital humanities research and among those, term-extraction is one of the most important topics. Therefore, it is important to extract person names correctly in historical materials as it is a highly representative symbol. This paper provides an automated name-extraction algorithm. We first segment the context into phrases by using the formula of PMI (Pointwise mutual information) and put them into the candidate list then we sort out the potential person names from the list. The segmentation is firstly used to improve the recall and the validation of names is for enhancing the precision. We target it is a general method that can be used to look for names in any kind of Chinese corpus. With this mainframe and minor adjustments may be applied in detail parts, this method can be optimized in any corpus. In this paper we uses "Zizhi Tongjian" as the experimental corpus. By identify the person names in the corpus, we discuss the advantages and disadvantages of this algorithm.

Keywords: Name-Extraction, Automation, PMI (Pointwise Mutual Information), Validation of Names, "Zizhi Tongjian"

[*] Graduate Student, Department of Computer Science and Information Engineering, National Taiwan University.
[**] Master, Department of Computer Science and Information Engineering, National Taiwan University.
[***] Postdoctoral Researcher, Department of CSIE, National Taiwan University. Email: tu@turing.csie.ntu.edu.tw.
[****] Research Associate, Research Center for Digital Humanities, National Taiwan University. E-mail: r90123005@ntu.edu.tw.
[*****] Distinguished Professor, Department of Computer Science and Information Engineering, National Taiwan University. E-mail: jhsiang@ntu.edu.tw.