Mining Classical Chinese Literature for Names and Transliterations: The Experiences of Taiwanese Digital Humanity Researchers

Richard Tzong-Han Tsai[1]
1 Department of Computer Science and Information Engineering, National Central University

Names of people, titles, locations, times, etc. are usually the main concepts in sentences. In natural language processing and information extraction, names are referred to as named entities (NE). The first step in extracting information from any literature is identifying the named entities in the text.

In classical Chinese literature, many NEs are transliterations, such as Buddhist terminology, geographical names, or the plethora of Western technical terms and proper nouns introduced in the 18th and 19th centuries. Transliterated terms are key records that show the contact and interchange between different cultures. However, they present certain problems from the perspective of text mining. For example, their formation is irregular and very different from the formation of regular Chinese NEs. Another difficulty for extracting NEs that applies to all Chinese literature is the lack of punctuation marks and word delimiters. In this talk, I will give an overview of the work in extracting Chinese NEs and transliterated NEs from classical Chinese literature carried out by Taiwanese researchers in recent years.