Presentation Details
Title: Mining for Translated Texts of Late-Qing and Republican Period in Chinese Databases
Speaker: Mr. CAO Anran, Nason (The Chinese University of Hong Kong)
Abstract: There are several online Chinese databases available for researchers to mine for translated texts in late-Qing and Republican Period such as Duxiu Academic Search (讀秀學術搜索), Dacheng Data (大成故紙堆), National Index to Chinese Newspapers and Periodicals (全國報刊索引) and Han Tang Modern Newspaper (瀚堂近代報刊). This presentation will focus on the text capture process of corpus compilation. I will go through some of these Chinese databases for translated texts, compare features of different features and analyze the OCR problems existing for the databases and the text capture process. The efficiencies of four different OCR software packages will be compared and the best OCR software package for texts published in late-Qing and Republican Period will be recommended for the text capture process in corpus compilation.
Biography **Nason Cao** is now Research Assistant of Centre for Translation Technology (CTT), Department of Translation, CUHK. His research interests include comparative language studies, corpus linguistics and language policies in Chinese-speaking societies. He is currently working on several projects of CTT including Chinese Translations and Pseudo-translations, 1712-1840: A Three-way Comparable Corpus and Building and Using Bilingual Sentiment Corpora for Translation Research: A Pilot Study.