Place Name Extraction from Field Notes Based on Text Analysis for Area Studies

Taizo Yamada
Historiographical Institute, The University of Tokyo
E-mail: t_yamada@hi.u-tokyo.ac.jp

The researcher in Area Studies needs to analyze various phenomena in an area. In order to promote the analysis, the data should have the elements concerning spatiotemporal data. Recently a variety of databases for Area Studies such as a catalogue of a material, image, movie and statistical data have been published. Most of the data has spatial and/or temporal data. On the other hand, the analysis of text data such as a field may be difficult for the researcher to use comparing to a catalogue or an image, because the data of traditional field note is plain and unstructured.

In the study we introduce the method of text analysis for field note which can extract a place name from field note. In order to use the method, we prepared structured text data which is converted from plain text of field note. A date element can be obtained easily from the structured data, because most of field note is constructed by a unit of a day. We attempted to extract the place name using information technique automatically from a sentence of field note. Using morphological analysis as an information technique, place names can be obtained from field note. Because in field note there are a lot of non-famous place names which is not registered in our gazetteer, we attempted the result of morphological analysis are improved by our rule based place name extraction based on machine learning.

Furthermore, we characterized a sentence in field note with a latent topic which is hidden in the field note and can be detected by LDA which is one of a topic model, in LDA each text can be represented as a mixture of various (latent) topics and each topic can be represented as a mixture of various terms. We show a prototype of a search system which the user can obtain a search result which has a place name and a latent topic which can categorize a text.