**YAMADA — Detection**

<0 figures>

**Detection of People Relationships Using Topic Model from Diaries in the Medieval Period of Japan**

Yamada, T. and Inoue, S.

Analysis of relationships between persons is an important element for historical study. Traditionally, the analysis is processed by a historical researcher manually, because reading comprehension of historical materials and historical background are needed for the analysis. However, the analysis is based on subjective judgment, and finding a relationship that is yet unknown by the analyst is difficult. Furthermore, manual analysis cannot cope with a large amount of historical materials. We consider that a method for objective and automatic analysis is required.

In this study of analysis between persons, we introduce a method for detecting relationships between persons from historical materials. In the method, we detect co-occurrence for each text in historical materials as a relationship between persons. The score that indicates the co-occurrence is calculated according to the topics that are latent and hidden in the text. For the detection of the latent topic, we use LDA (Latent Dirichelt Allocation; Blei et al., 2003) as a topic model. In the study, LDA assumes that text has one or more latent topics, a topic has one or more personal names, and the topic is hidden in the text. The latent topic is calculated by personal name co-occurrence. When the latent topics are calculated, the following two output results can be obtained: relationships between texts and latent topics, and relationships between latent topics and personal names. Using LDA, a text can be shown by topic content ratio. By the process, because appearance of personal name can be indicated by allocation of the latent topics, relationship between personal names can be calculated objectively and automatically.

Since LDA is required for the vector of the personal name in the study, we introduce a personal extraction method from text of historical materials. The method is which personal name can be extracted according to sequence pattern matching, because we have no dictionary for the historical personal name, and morphological analysis against text of Japanese medieval material is very difficult. The sequence pattern means string sequence, which seems to personal name appearances. We prepare training data that includes patterns of personal name and non-personal name and test the data created by random sampling from texts. We confirm test data with the training data on whether the terms extracted by the sequence pattern are personal names. When the prediction failure pattern appears, the failure pattern is added into the set of the sequence pattern in correspondence with personal name, and then pattern matching is processed again. The precision of the personal name extraction can be improved until about 0.95 by performing the feedback several times. For the pattern matching, we use SVM (Support Vector Machine). Currently SVM is one of most superior supervised learning techniques.

We prototyped text search systems where a user can search text of ‘上井覚兼日記’ (*Uwaikakken nikki*), which is a diary (1574–1586) from the Japanese medieval period written by ‘上井覚兼’ (Uwai kakken), who is a senior statesman of ‘島津家’ (Shimadu family) of Japan. For the historical study of Kyushu (which is a local area of Japan) or 島津家 in the medieval period, the diary is one of the important historical materials and a Japanese national treasure. The diary is held by the Historiographical Institute (HI), the University of Tokyo, and the text data is listed in the database of *The Full-Text Database of the Old Japanese Diaries* published by HI. In the system, a user can obtain search results by traditional text search, a personal name in the text, and a personal name related to the personal name. The system can also show a graph indicating the relationship between personal names. Due to the presentation of related personal names, the system uses the detection method of relationship between personal names. And the system uses our personal name extraction method for presentation of a personal name.

**Reference**

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research,* **3** (March): 993–1022.