筑波大学
University of Tsukuba

The 132nd RCKC Colloquium

Human History Project

Speaker: Ichiro Fujinaga

Associate Professor, Music Technology Area at the Schulich School of Music, McGill University

Human History Project is a large, long-term enterprise, which aims to build a distributed international database of documented human history using Natural Language Processing (NLP) tools and Linked Open Data (LOD) to model historical data. Exploiting the ever-increasing availability of historical documents and recent improvements in optical character recognition (OCR), this project aims to create, automatically, an economically feasible digital prosopographical database, which will include series of events (relationships between named entities). Even with the current state-of-the-art OCR and NLP technologies, however, there are still some errors for which we plan to deploy crowd- or expert-sourcing techniques for corrections. For this we are developing a JavaScript-based online editor to correct errors. The results are stored in the quad RDF (Resource Description Framework) format, which then can be searched via SPARQL.

In a pilot NEH-funded project entitled “Digital Prosopography of Renaissance Musicians,” we are creating a framework that can answer questions not easily answered by Google-like searches or traditional means. For example, which printers in Venice in the 1530s were publishing books of music? Which foreign musicians visited Venice in 1538? Did composer A and composer B live in Venice in 1538? Were there musicians working in Venice from 1535–1540 who performed music by both of these composers?

We have experimented with the named-entity extraction of the GATE (General Architecture for Text Engineering) system using biographical entries on ten Renaissance composers from three different sources: Wikipedia, Oxford Music Online, and the 1911 edition of Grove’s Dictionary of Music and Musicians. The total of 5,441 entities were extracted with the accuracy of 99.24% precision and 98.9% recall. It should be noted, however, that it took over three hours to manually verify and correct the output from the thirty articles; confirming the need for efficient and economical means of correction.

It is hoped that as more historical documents are digitized and as the NLP technologies improve, a wealth of historical information, which was available but extremely difficult to extract, can be more easily searched and retrieved.

The seminar will be presented in English.

No charge to participate and No reservation is needed.

Anyone is WELCOME!

Date

Thursday, March 17th, 2016

13:00 – 14:00 Seminar
14:00 – 15:00 Discussion *

* Participation in the discussion is optional.

Venue

Meeting Room for Joint Research 1
on the 3rd floor of ULIS bldg.
in Kasuga Area, University of Tsukuba

Research Center for Knowledge Communities, University of Tsukuba
http://www.kc.tsukuba.ac.jp/index.html

Email: kc-office@ml.cc.tsukuba.ac.jp
Tel: 029-859-1524 (Ext. 81524)

知的コミュニティ基盤研究センター
Research Center for Knowledge Communities