

2009 Chicago Colloquium on Digital Humanities & Computer Science: Abstract for Paper Presentation

Title: Who's Who in Your Digital Collection? Developing a Tool for Name Disambiguation and Identity Resolution

Authors: Jean Godby, Patricia Hswe, Judith Klavans, Hyoungtae Cho, Dan Roth, Lev Ratinov, and Larry Jackson

In the past five years or so, the problem space of automatically recognizing, extracting, and disambiguating named entities (e.g., the names of people, places, and organizations) from digitized text has received considerable attention in research produced by the library, computer science, and linguistics communities. Name identification and extraction tools, particularly when integrated with an authority file (e.g., WorldCat Identities, Wikipedia, etc.), can enhance reliable subject access for a document collection, improving on its discoverability by end-users.

For example, in the context of historical documents, the ability to find out who knew whom and why they were associated, as well as whether the individuals are actually the ones the user is seeking, cultivates a potential for further, value-adding analysis of the documents' content. Of late, named entity extraction applications have been featured in environments such as institutional repositories, where name authority control and disambiguation of names are crucial for indexing and citation purposes. Digital humanities projects have incorporated name extractor tools as well. The Perseus Digital Library (<http://www.perseus.tufts.edu/hopper/>) has Named Entity Search Tools that mine its collections for people, places, and even dates. The Metadata Offer New Knowledge (MONK) project (<http://www.monkproject.org/>) offers a workbench for textual analysis on multiple levels, including a tool for recognizing and extracting named entities in its collections (which consist of works of eighteenth- and nineteenth-century American literature and works by William Shakespeare). Named-entity extractors can also be found in cataloging utilities, such as the Computational Linguistics for Metadata Building (CLiMB) Toolkit (<http://www.umiacs.umd.edu/~climb/>), which addresses the "subject metadata gap" in visual resources cataloging by increasing subject access points for images of art objects (Klavans et al., 2009, p. 184).

The Extracting Metadata for Preservation (EMP) Project, funded by the National Digital Information Infrastructure and Preservation (NDIIPP) Program, is addressing this ongoing challenge of identifying proper names to improve the accuracy of end-user information access via web-based search and retrieval. As a three-way collaboration among the University of Illinois at Urbana-Champaign, OCLC, and the University of Maryland, EMP researchers bring multidisciplinary perspectives from the library, computer science, and linguistics communities to the problem of high-quality identification and disambiguation of names. Our work has three goals: 1) to advance the state of the art in automated name identification and disambiguation; 2) to link the outputs of these programs to longstanding efforts in the library community to manage names and identities in the published record; and 3) to lower the barrier of access to sophisticated text-processing tools that have broad applicability.

This paper reports on three main activities of the EMP project. First, we describe the open-source name extractor tool, developed by computer scientists at Illinois, which uses the approach of Machine Learning (Ratinov & Roth, 2009), and which is configured with a plug-in interface and performs favorably against previously available solutions. It has been enhanced with a utility that matches personal names with the correct entry in Wikipedia and development is underway to do the same with OCLC's WorldCat Identities (<http://www.worldcat.org/identities/>). Second, we present methods for evaluating the tool. Evaluation is essential for determining the performance of a name extractor tool on the different types of tasks required by library and museum applications, and we discuss the issues and decisions involved in raising the rate of precision (correctly identified named entities) in our results. Third, we demonstrate the use of this tool by integrating it into two applications developed at the collaborating institutions: the CLIMB Project, hosted at the University of Maryland; and the QuestionPoint knowledge base (<http://www.questionpoint.org/>), a database of questions and answers, hosted at OCLC, which represents interactions between reference librarians and library patrons. When enhanced with the EMP tools, these projects display improved handling of the personal names residing in their full-text databases, which is visible to users as clearer indexes, richer hyperlinks, and more focused search results.

Tools for recognizing and extracting named entities from unstructured text have become more prevalent, as applications such as the ClearForest Gnosis add-on for Firefox (<https://addons.mozilla.org/en-US/firefox/addon/3999>) and the AlchemyAPI (<http://www.alchemyapi.com/>) tool suite for content analysis, attest. Yet, unlike the application EMP is developing, these utilities are not open-source, which limits possibilities for widespread adoption in diverse contexts. Our research outcomes will be of interest to archivists, curators, and humanities scholars looking for practical and easy-to-use ways of discovering who's who in their digital resource collections.

References:

- Hickey, T. (2008, April). *VIAF and WorldCat Identities*. [PowerPoint slides.] Paper presented at the annual European Library Automation Group (32nd ELAG Library Systems Seminar), Wageningen, The Netherlands. Retrieved August 29, 2009, from http://library.wur.nl/WebQuery/file/formulier/profielelaglt_i00117278_001.ppt.
- Klavans, J. et al. (2009). Mining texts for image terms: the CLIMB project. In *Digital Humanities 2009. Conference Abstracts. University of Maryland, College Park, USA. June 22 – 25, 2009* (pp. 184-186). College Park, MD: Maryland Institute for Technology in the Humanities (MITH). Retrieved August 28, 2009, from http://www.umiacs.umd.edu/~jimmylin/publications/Klavans_etal_DH2009.pdf.
- Ratinov, L. & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *CoNLL-2009: Thirteenth Conference on Computational Natural Language Learning* (pp. 147-155). Boulder, CO: Association for Computational Linguistics. Retrieved August 29, 2009, from <http://aclweb.org/anthology-new/W/W09/W09-1119.pdf>.