Breathing Life into Digital Archives: Use of Natural Language Processing to Revitalize the Grey Literature of Public Health

Anne M. Turner
Division of Medical Informatics
and Outcomes Research,
Oregon Health & Science
University
Portland, Oregon
503-494-2388
turneann@ohsu.edu

Elizabeth D. Liddy
Center for Natural Language
Processing,
School of Information Studies,
Syracuse University,
Syracuse, New York
315-443-5484
liddy@syr.edu

Jana Bradley
School of Information Studies,
Syracuse University,
Syracuse, New York
315-443- 1797
jobradle@syr.edu

The goal of our 2-year Robert Wood Johnson-funded project is to apply Natural Language Processing (NLP) technology to improve access and use of the digitalized public health "grey" literature. Much public health information, such as meeting notes, think-tank reports, policy statements, and data sets, is not available through traditional commercial pathways and is considered grey or fugitive literature. Although grey literature documents are increasingly posted in digital archives on the Web, the unstructured and varied nature of grey literature makes accessing useful content difficult at best.

In an effort to help make the content of public health digital collections more accessible to public health providers, we will use proven NLP techniques to identify and extract key elements of digital documents. NLP techniques can be used to identify and tag key elements from full-text documents. Once tagged, the content of various documents can be extracted and summarized in tables and charts for comparison and review. The ability of NLP to recognize and represent both the explicit and implicit content of full text documents makes it a powerful technique for interpreting the language of text documents. Our NLP information access system has been used in other domains to extract individual entities and events, as well as draw relationships between entities and events to build a content representation.

We are currently analyzing grey literature documents from digital archives found on public health websites. One such archive is the New York Academy of Medicine's Grey Literature Report, a digital archive that provides links to agency reports pertaining to pubic health. Document titles are posted under the submitting agency but are not indexed or searchable.

The goal is to develop a model of public health interventions and identify key entities and events from these digital archives. Key elements may include type of study and population demographics as well as more traditional bibliographic elements such as author, title and publication date.

NLP technology will be used to search, identify and extract key elements based on the user's request. Key elements can be extracted across multiple documents for summary and comparison. For example, the user can extract key elements from annual reports about "teenage smoking cessation programs" to compare method of intervention, demographic population, and outcome. Such comparisons will help public health professionals to determine how a particular intervention fits with similar interventions reported in the grey literature. This system holds great promise for improving access to public health information through digital archives.

Categories and Subject Descriptors: H.3.7 [Digital Libraries]

General Terms: Human Factors, Documentation

Keywords: digital libraries, digital archives, natural language processing, public health

JCDL'02, July 13-17, 2002, Portland, Oregon, USA.
ACM 1-58113-513-0/02/0007.