Automatic Subject Heading Extraction

Jonathan Kennedy, Sr. Software Engineer, Center for Biomedical Informatics
Juliane Schneider, Metatdata Librarian, Countway Library
Peter Rolla, Information Acquisitions & Management Librarian, Countway Library
Betsy Eggleston, Director of Collections & Knowledge Management, Countway Library

Final draft: changes to this document are in **bold**.

1. Introduction

Like the stacks of many Harvard libraries, much of the older print collection of the Countway Medical Library is stored off-site at the Harvard Depository. This collection is still valuable to researchers (especially those engaged in studying the history of medicine) but the location of these resources creates obstacles to discovery; researchers are not able to physically browse the collection so the HOLLIS catalog provides the only possible method of discovery. However, many of the records for these books are inadequate for discovery; specifically, many of the catalog records lack medical subject headings (MeSH) and thus are not findable via subject searches. Countway Library currently has 115,724 titles stored at the Harvard Depository and **45%** (**51,081**) of the bibliographic records for those titles lack MeSH headings. Because of the quantity of items involved it is unfeasible for library staff to deal with titles on an individual basis. Therefore, the authors propose the use of computational methods to automatically extract the appropriate subject headings using existing metadata and the electronic full-text where available from the Harvard-Google book-scanning project (or other sources). **There are 6,419 of these titles with local electronic copies and an additional estimated 2,500 available via Google Books, scanned from other libraries.**

While these materials may be discoverable from Google Books via full text searching, this full text is not indexed by HOLLIS. The use of subject headings for search remains an important method employed by patrons and library staff and has not been displaced by the popularity of full text indexes[1]. Modern information retrieval systems, including those relying on full text indexing, increasingly utilize structured and semantic information to improve full text searching with techniques like synonym expansion for improved recall and document clustering and "aboutness" measures for relevance ranking[2]. Semantic Web, Linked Open Data, and various text-mining initiatives also rely on the creation of robust metadata for holdings.

This project will provide an important opportunity to improve the discovery of existing library materials, **support the use of modern information retrieval**

[1] This topic was recently discussed on the hlcomms mailing list.
[2] PubMed search is a highly visible example of this in higher education.

and linked data techniques, and validate the use of automated methods for subject heading extraction.

2. Background

Medical term extraction (concepts, entities, etc.) from natural language (books, publications, etc.) is an important area of work in bioinformatics and medical informatics. Project authors, working for Countway Library of Medicine and the Center for Biomedical Informatics (CBMI), have recent experience with term extraction using biomedical ontologies (including MeSH). The Community Connects to Research[3] and Harvard Catalyst Clinical Trials[4] projects use natural language processing methods to identify medical conditions, researchers, locations, and institutions from free-form descriptions of clinical trials. This process allows the creation of synonymy and other relationships that are used to improve upon modern search technology.

3. Project

This project will create a tool to improve the completeness of existing catalog records by using proven techniques to automatically extract subject headings from digitized print collections. This will create the opportunity to evaluate proven methods and experiment with cutting edge techniques (i.e. measuring "aboutness"[5]) as they relate to cataloging and discovery.

Techniques that have been proven on CBMI/Countway projects will provide the basis for this work. Methods include the use of existing ontologies and lexical normalization to perform term matching in a body of text. Relationships within and between ontologies are used to connect matching phrases to a concept of interest. For this project, the authors will use the UMLS ontology (a collection of important biomedical ontologies published by the National Library of Medicine), which offers a source for term synonymy. The Northwestern project to map LCSH terms to MeSH terms will also be evaluated[6]. Project team will go further by also experimenting with proven NLP annotators to develop language models that may improve term identification accuracy.

Subject headings derived from this project may be added to HOLLIS using an automated batch load or existing manual entry methods.

[3] http://www.connecttoresearch.org/

[4] http://catalyst.harvard.edu/

[5] http://dl.acm.org/citation.cfm?id=1654763

[6] http://www.accessmylibrary.com/article-1G1-19422101/mapping-lcsh-and-mesh.html

4. Determining success

Project is expected to offer immediate benefits by using proven methods to add MeSH subject headings to existing records. This will improve the discoverability of **approximately 9,000** print materials and validate automated methods for extracting subject headings. **This proposal was shared with Scott Wicks, Head of Technical Services, who expressed excitement about the project and an interest in learning how much digitized text is necessary to produce competitive results, as this could improve the records of other Harvard assets that may not be immediately processed for a variety of reasons.**

An evaluation will be conducted to compare the tool**, using various selections of the text (full text, introduction-only, first *n* chapters only, etc.),** against existing manual methods to determine accuracy and coverage. A selection of print materials will be selected to undergo both manual and automatic subject heading extraction; experienced cataloging staff will evaluate the results.

5. Project staff

Jonathan Kennedy will perform the necessary software development for the project. Juliane Schneider will consult on ontology use and data quality. Peter Rolla will coordinate data integration and results analysis. Betsy Eggleston will provide senior oversight and consultation.

6. Budget

No physical resources will be required for this project; software can run on existing CBMI/Countway servers. Countway Library will support project coordination activities. Project will request a partial chargeback for development efforts.

8 weeks (2 months) x $85/hr x %75 = $20,400

7. Future work

Semantic understanding of natural language is a growing area of importance in business and IT and is making critical inroads into modern information retrieval technologies. Research into natural language processing is a diverse field with many areas of potential research. The authors hope that successful application of this technology to library systems will improve service to patrons, maintain library competitiveness, and generate further interest in the use of these technologies.