ASHE Final Report

Prepared for Library Lab, November 16, 2012

Jonathan Kennedy, Juliane Schneider, Peter Rolla, Betsy Eggleston

I. SUMMARY

The Automatic Subject Heading Extraction (ASHE) project is designed to improve the discoverability of books that have not been fully cataloged (have no subject headings). Books are not full-text indexed in the HOLLIS catalog, so the presence of subject headings makes the difference between a book being used or sitting in storage, invisible. This is particularly true at the Countway Medical Library where subject heading (specifically, MeSH) searches are regularly used to find books, of which over 115,000 are stored off-site at the Harvard Depository. Among these, over 50,000 lack subject headings and are unlikely to be found by patrons. ASHE solves this problem by using the digital form of a text, available for approximately 10,000 of these titles, to extract relevant subject headings for addition to the catalog. The ASHE project has succeeded in developing a tool that is capable of extracting medical concepts from digital books and identifying the best matching MeSH heading. The tool improves upon previous efforts at concept identification by using tools and techniques from natural language processing.

II. ACCOMPLISHMENTS

This project developed the ASHE Cataloger, a tool that can successfully identify medical concepts within a text. The ASHE Cataloger is composed of three main components:

1. The natural language processing pipeline uses linguistic principles to parse free text into meaningful phrases that can be used for concept identification (see accompanying PDF for a visual description of the process). Components from the popular UIMA and OpenNLP libraries were used and original code was developed to organize the final candidate phrases. This improves upon past projects by using NLP tools and additional linguistic insights to efficiently narrow the search space.

2. The lookup algorithm is able to search the UMLS Metathesaurus (a large database aggregating many different biomedical ontologies) to identify concepts. Advanced features include the ability to lexically normalize terms for lookup using the NLM Lexical Tools, permuting phrases to broaden searching, and following the relationships between concepts to identify matching MeSH terms.

3. A scoring algorithm was developed to determine which MeSH headings were the best to select for addition to the catalog. Because the tool is able link synonymous terms to the same concept, it can accurately measure which concepts are the most prevalent within a document. The scoring algorithm uses the tf*idf calculation that is foundational to modern search engine technologies.

Significant experience was gained using important tools throughout the duration of this project, including: UIMA, OpenNLP, NLM Lexical Tools, MeSH, and the UMLS Metathesaurus.

Early results from this project have validated that meaningful concepts can be extracted using the OCR from scanned books. Using a subset of 20 titles (OCR was downloaded from the Internet Archive), we produced ranked subject headings for each title. Our initial analysis found that the top ranked headings are all relevant concepts with few exceptions. The tool produces many more candidate headings than a human cataloger would; in many cases, a human would only assigned one or two headings, while the tool may identify hundreds of concepts within a text. While this makes a one-to-one comparison challenging, our initial analysis has found that **the headings assigned by a human are usually found within the top 10 candidates produced by the tool**. We have also run an experiment using only the title and table-of-contents. While this method produces many fewer candidates, the results generally match the top scores using the full text; a result that may be of interest to future book scanning efforts.

Benefits of this project have also included the identification of additional uses for the tool and future collaboration opportunities. Specifically, the book scanning and transcription projects, also funded by Library Lab, each present additional opportunities for the use of this tool. Increased digitization efforts will magnify the benefits of the ASHE project by creating more material that can be automatically cataloged and discovered. Two collaboration opportunities have been discussed; 1. Methods developed for the ASHE project might be used to evaluate the quality of scanned OCR produced by various scanning technologies, and 2. ASHE might be extended to identify concepts beyond subject headings, such as names, locations and events.

The main project team was made up of a technical services librarian, a metadata librarian, and a software engineer. Over the course of the project, significant knowledge sharing occurred about cataloging methodologies, the intricacies of vocabularies and ontologies, and the opportunities created by computational methods. We believe this project is an ideal illustration of the value of increased collaboration between the information scientist and computer scientist within Harvard libraries.

III. CHALLENGES

While we believe our initial results have demonstrated the value of automated methods, more work is necessary to refine the results if the goal is to mimic a human cataloger. The project initially set out to compare the subject headings entered by a human expert to those generated by the tool for 100 titles. However, the results require further refinement before a one-to-one comparison will yield a meaningful result. The project team intends to continue working on the results and produce a final evaluation.

Certain technical challenges were identified up front by the project team and were incorporated into the parameters set for the project:

1. It was determined that only English language titles would be appropriate since the quality of the NLP tools may vary by language. However, it is thought that this is one area where automated methods can offer significant value, given the general difficulty of staffing to support many languages.

2. Qualifiers in MeSH represent broader concepts than a tool may be able to pick out of a document. For example, a book about the history of medicine might be cataloged as Medicine/History. It was determined that qualifiers would be ignored by the tool and not included in the comparison.

3. Since many of the scanned titles are from out-of-copyright works from the late 1800s and early 1900s, it was thought that the tool might underperform given the differences between historical and modern medical terminology. However, this does not seem to have limited the tool, possibly because the broad terms that would be used for cataloging have changed less over time than more specific terminology.

IV. NEXT STEPS

As time and funding permit, the project team will continue refining the results of the tool and coordinate with LTS to retrieve the rest of the relevant documents to run the final evaluation.

During the development of this project and through feedback from peers, the project team has identified numerous opportunities for developing advanced techniques to improve the results of the tool.

1. Processing prepositions and conjunctions is a complex problem as illustrated by the famous Groucho Marx quote, “One morning I shot an elephant in my pajamas.” Improved grammar parsing will allow the tool to recognize that “cancer of the lung” should be treated as a single concept or that “disorders of synovium, tendon, and bursa” should be treated as three concepts, all of which are disorders.

2. The ability to infer broader statements about a work is critical to match the results of a human expert. Since the ontologies used by the tool express rich relationships between concepts, this might be used to aggregate terms into broader categories for cataloging. For example, the tool may recognize that mention of many different kinds of heart disease should result in the general heading “Heart diseases.”

3. Disambiguation is a complex problem in the area of natural language processing. For example, the term “lupus” may be used to refer to many different representations of the disease. Developing methods to disambiguate between multiple possible matches will significantly reduce the chance of an inaccurate heading being applied.

4. Further analysis might reveal that some ontologies produce better

results than others, so we might restrict the lookup algorithm to only certain ontologies contained within the UMLS.

5. Incorporating other ontologies and vocabularies, such as the Library of Congress subject headings and authority records, will make the tool more generally useful and allow the project team to test the tool for other disciplines.

6. The addition of qualifiers was determined to be out-of-scope for the initial phase of this project, but the development of advanced grammar parsing may allow the tool to understand when to apply certain qualifiers.

7. OCR errors can create problems for automated tools, especially if there is a desire to work on smaller selections of text (i.e. just the table of contents). Methods of correcting OCR might be a useful future project.

8. While ASHE aims to reproduce the efficacy of the human expert, the project opens up new possibilities for improving search and discovery. Today, a user for searching information on “nitrogen imbalances” can’t simply search for those keywords; the user would have to first search for books with the heading “nutrition” or “metabolism” and then manually inspect those titles. Why not improve discovery by indexing all of the concepts extracted from a book (as secondary headings, or in some other index)?

V. PUBLICITY

Juliane Schneider and Jonathan Kennedy will present this work at the KLA (Kentucky Library Association) Academic and Special Library Sections and SLA (Special Libraries Association) Kentucky Chapter's 2013 Joint Spring Conference.

The project team intends to write a paper for publication once results permit an easy to understand one-to-one human-to-computer comparison.

Many thanks to Library Lab for their generous and continued support of innovation within the Harvard Libraries.