Toward using ontologies to improve results in searches for mental health information

Jonathan Bona[1,*], John Grohol[2], Meredith Zozus[1], Robert Zozus[3] and Mathias Brochhausen[1]

[1] Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, USA

[2] Psych Central.com, Newburyport, MA USA

[3] Clinical Psychologist in Private Practice; PsyUSA Network Curator, Little Rock, USA

ABSTRACT

We have implemented a proof of concept that uses select terms from the Consumer Health Vocabulary and the Human Disease Ontology to annotate and index articles about mental health from PsychCentral.com. This paper presents the approach used and preliminary results, which indicate that processing health-related documents with the use of biomedical ontologies and terminologies can make them easier to find using natural language search queries.

1 INTRODUCTION

As part of an ongoing project that aims to improve the ability of healthcare consumers to find, organize, and access information about their mental health concerns, we are developing ontology-driven tools for search, annotation, and exploration of consumer-oriented health content on the Internet. The major goal for this project is to improve healthcare consumers' access to information relevant to their health. This abstract reports on preliminary work that uses ontological and terminological resources to index consumer-oriented health content and make it more easily retrievable with natural language search queries that would not otherwise yield the same results. This can also serve to facilitate translation from consumer language-based queries to material using expert language, such as medical terminologies.

To demonstrate the usefulness of existing ontologies and terminologies for enhancing retrieval of consumer-oriented health texts in the mental health domain, we have developed an ontology-based natural language processing and indexing strategy and a proof of concept for a small domain, and tested it on a set of curated, consumer-oriented articles retrieved from Psych Central. Psych Central is the Internet's largest and oldest independent mental health social network, with over 450,000 content pages about mental health.

We selected the test case of searching for articles with information about *seasonal affective disorder* (SAD) among a curated set of several hundred consumer-oriented documents about depression. Seasonal affective disorder is cyclical depression that occurs only during certain times of year, most commonly in the winter. Some resources, such as the Human Disease Ontology (DO) treat "winter depression" and "seasonal affective disorder" as exact synonyms[1].

Because seasonal depression can occur in the spring/summer, *winter depression* might be more accurately modeled a subclass of *seasonal affective disorder*. To simplify we follow DO in treating these terms as synonyms here.

We also use terms from the OCHV (Amith & Tao, 2016), an OWL version of the Consumer Health Vocabulary (CHV) (Zeng & Tse, 2006), which is an open access effort to bridge the gap between healthcare consumers and professionals by linking everyday health terms to matching technical terms. Here we use the OCHV term labeled "seasonal affective disorder"[2] with alternate labels: "affective disorder seasonal", "depression seasonal", "seasonal affective disorder (SAD)", "seasonal affective disorders", "seasonal depression".

2 METHODS

2.1 Data collection and preparation

In collaboration with Psych Central, we downloaded 479 articles that appear in the site's library under *Articles on Depression*, or under one of a few related categories: *bipolar disorder*, *antidepressants*, *seasonal affective disorder*, *postpartum depression*. 22 of these 479 articles contain one or more of the exact phrases "seasonal affective disorder", "seasonal depression", or "winter depression". Of those, 14 (63.6%) contain only "seasonal affective disorder," so would not be immediately uncovered by exact searches the other terms. A Google search across all Psych Central pages shows that about 60% of pages containing any of these three terms actually contains only the term "seasonal affective disorder".

We used Python and its BeautifulSoup (Richardson, 2017) library to download and process these 479 depression-related articles, writing each page's content to a plain text file with most of the structure and formatting removed.

2.2 Named entity recognition and indexing

Using the selected OCHV and DO terms about *seasonal affective disorder*, we built a dictionary-based named entity recognizer (NER) using Apache's Java-based OpenNLP Toolkit (Apache Software Foundation, 2017b). This NER takes a text file and a dictionary as input and outputs a list of

* To whom correspondence should be addressed: jonathanbona@gmail.com

[1] http://purl.obolibrary.org/obo/DOID_0060167

[2] http://uth.tmc.edu/ontology/ochv#52085

positions in the text that match any of the terms, along with the URI(s) of the matching term(s).

The NER was run on each of the 479 articles, producing as output an XML file for each containing both the article's text and, as separate metadata fields, the URIs and labels of any of our SAD terms that the NER identified in the text.

These text and metadata files were then indexed using the open source search platform Apache Solr (Apache Software Foundation, 2017a). Solr prepares documents for fast retrieval by parsing their contents and indexing terms that appear therein. Solr can automatically perform some basic text processing tasks, but it does not have built-in ontology-based named entity recognition. Pro-processing documents with named entity recognition using our SAD terms, and using the result as document metadata allows Solr to index those terms along with those that appear in the source document itself.

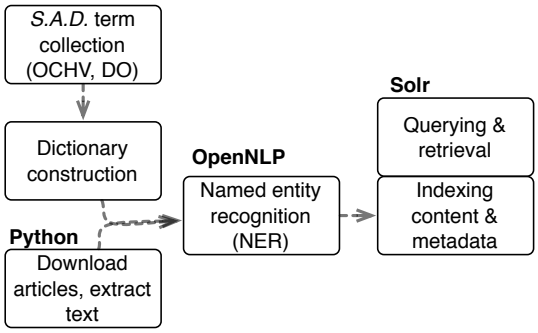

Figure 1: Indexing and searching articles using ontologies

3 RESULTS

The result of this process of document text extraction, ontology-driven named entity recognition processing, and combined text and ontology term metadata indexing is a set of articles stored in Solr for very fast retrieval that can be found using any of the terms that correspond to an entity that matches. For instance, for a document that uses only the phrase "seasonal affective disorder", our use of the NER tool prior to Solr indexing ensures that the metadata used to index that document includes its synonyms "winter depression," "seasonal depression," etc., and will thereby allow the document to be retrieved by queries that use any of these terms.

We tested the effectiveness of this by running queries for the exact phrases "seasonal affective disorder", "seasonal depression", and "winter depression," and comparing the results to similar queries run on a Solr instance that had been created from the original article files without the named entity recognition and annotation step. As expected, with the ontology terms added as metadata, all of the articles that contain any of these three terms are retrieved by any query that uses any of these terms. That is, articles containing only "seasonal depression" can be retrieved by a query that uses "winter depression" or "seasonal affective disorder" instead.

4 DISCUSSION AND FUTURE WORK

Pre-processing text with dictionary-based named entity recognition using ontology terms from resources such as DO and OCHV, and indexing the resulting annotations as metadata along with the original text using standard information retrieval tools, makes that content more easily retrievable using a wider variety of search terms. We expect this result to generalize beyond our specific test case of *seasonal affective disorder*, and beyond mental health. This approach can help realize the potential of consumer health vocabularies as a tool to bridge the gap between healthcare consumer language and technical language used by experts.

Using more terms even for this small test case (e.g. "weather related depression", "periodic depression"), would expand the possible queries that return relevant results using non-expert language to search for health information. We have focused here on exact search results, but Solr can also return partial matches. Having matched ontology terms as metadata with documents will also improve those results.

We will continue this work by using a larger, more general mental health test case requiring the use of many more ontology terms. We will also expand the set of documents used to include more content from Psych Central and other sources. We will investigate the use of relations between terms other than exact synonymy to allow, e.g., a search for "mood disorders" to yield content that mentions "seasonal affective disorder" even if it doesn't explicitly mention "mood disorder".

This approach might not easily scale to a very large set of terms. NER with this set of terms took less time per document than Solr's indexing, which is quite fast, but we don't know how this will change when working with many more terms.

The dictionary-based named entity recognition is a very simple NER approach that works well in this case in part because of the manual curation that has gone into creating the ontologies from which our terms were sourced. We will explore the use of more sophisticated NER.

REFERENCES

Amith, T., & Tao, C. (2016, January 20). Ontology of Consumer Health Vocabulary (OCHV). Retrieved from https://bioportal.bioontology.org/ontologies/OCHV

Apache Software Foundation. (2017a). Apache Solr Reference Guide Covering Apache Solr 6.5. Retrieved from https://www.apache.org/dyn/closer.lua/lucene/solr/ref-guide/

Apache Software Foundation. (2017b). OpenNLP. Retrieved from http://opennlp.apache.org/index

Richardson, L. (2017). Beautiful Soup. Retrieved from https://www.crummy.com/software/BeautifulSoup/

Zeng, Q. T., & Tse, T. (2006). Exploring and Developing Consumer Health Vocabularies. *Journal of the American Medical Informatics Association : JAMIA*, *13*(1), 24–29. https://doi.org/10.1197/jamia.M1761