Enhancing Semantic Analysis of Pathology Reports

Philip E. Whalen*, Aditya Trilok Muralidharan, Jonathan R. Kiddy, and William D. Duncan*
Department of Bioinformatics & Biostatistics
Roswell Park Comprehensive Cancer Center
Buffalo, New York, USA

Keywords: ontology, natural language processing, named entity recognition, pathology report

Abstract

Pathology reports play an essential role in cancer treatment and research. They contain vital findings about a patient's cancer, such as cell histology and molecular markers, that are used to diagnose the type of cancer, determine treatment options, and enhance our understanding of the nature of the disease. At Roswell Park Comprehensive Cancer Center,[1] pathology reports are stored mainly as unstructured text in a relational database.[2] To find information within pathology reports, we use either string matching methods (e.g., regular expressions), or the TIES[3] natural language processing (NLP) program. The drawback of string matching is that string variations need to be accounted for in order to find information. For instance, a search for patients whose tumors lack estrogen receptor (ER) proteins will also have to search for strings matching 'ER negative', 'estrogen receptor negative', 'hormone status negative', and the like.

TIES addresses some of this variability by mapping multiple strings to the same ontology class, but some searches consistently do not perform well.[4] Moreover, a class identified by TIES is not linked to the formal axioms that define the class, which prevents researchers from fully leveraging the formal relations that hold between classes within an ontology. For example, the formal definition of Medullary Breast Carcinoma (C17965[5]) in the NCI Thesaurus (NCIt) [1] includes the axiom:

Disease_Mapped_To_Gene some 'BRCA1 Gene'

But, this axiom is not accessible within TIES, and, thus, you are not able to query for other cancers that are also mapped to the BRCA1 gene (such as hereditary prostate carcinoma) and investigate commonalities between them.

To address these shortcomings, we are developing an ontology that we currently call the 'Document Content Ontology'[6] (DCO) to represent the terms, words, word contexts, and their positions (i.e., indexes) within the documents. That is, we are using the DCO to represent the content of the document and where the content is located. It is important to make clear that while we are using an NLP program for named entity recognition (described in what follows), we are not developing an NLP program. Rather, we are augmenting the output of the NLP program so that we can more fully leverage the axioms contained within an ontology.

A high-level summary of the DCO is illustrated in **Figure 1**. Documents contain (i.e., has part) terms, and terms, which are composed of one more words, have meanings that are specified using the semantic type, sematic label, and semantic source annotation properties to reference an ontology class. The literal value data property associates the actual data (e.g., strings) with the term, and the polarity annotation represents whether the term has a positive or negative connotation (e.g., the patient does not have breast carcinoma). In some cases, we also represent the word context: the group of words surrounding the word or words that constitute a term. Word contexts are useful in aiding NLP programs to disambiguate the sense in which a word is being used. For brevity, not all properties and classes are discussed. Full details are available at https://github.com/RoswellParkResearch/document-content-ontology.

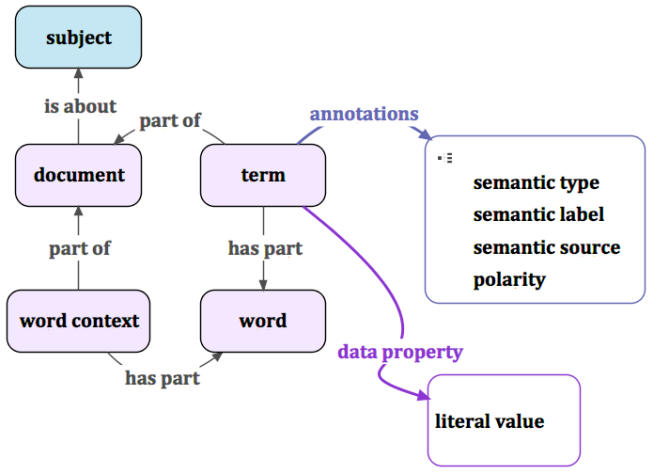

Figure 1: Architecture of the Document Content Ontology

We are aware that a number of other ontologies (such as the Information Artifact Ontology and Semanticscience Integrated

[1] https://www.roswellpark.org
[2] The database does contain some structured fields, but we find that most researchers are interested in the information contained in the unstructured text.
[3] http://ties.dbmi.pitt.edu
[4] TIES consistently fails to identify findings that the cells in the tumor lack estrogen-receptor proteins.
[5] IRI: http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C17965
[6] The name of the ontology may change!

* corresponding author

Ontology) have terms similar to ours. However, these ontologies carry with them metaphysical commitments, such as a document being a type of generically dependent continuant. Since we are just beginning to develop the DCO, we wish (for now) to remain agnostic concerning such commitments.

At present, we are using the DCO to structure output from the Noble Coder Named Recognition Engine [4]. Noble takes a document as input and outputs a file containing information about (named) entities identified within the document as well as the associated ontology classes that specify the meanings of the named entities. For example, if the Noble program determines that some text within a document refers to ductal breast carcinoma, Noble associates this text with the NCIt class 'Ductal Breast Carcinoma' (C4017[7]).

We translate the output of Noble into OWL and load it along with the full ontology of association classes (which we call a named entity's semantic type) into a GraphDB[8] semantic triple store. This allows us to simultaneously query for pathology reports having a specified named entity and the ontology for other entities related to the named entity.

Figure 2 shows an example of a SPARQL query to find documents that contain a term whose semantic type references ontology terms that represent diseases that are mapped to the BRCA1 gene. This method of querying pathology reports using a named entity's semantic type as well as the ontology's formal structure provides better coverage for finding relevant pathology reports than simple named entity recognition alone.

```
PREFIX disease_mapped_to_gene:
  <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#R176>
PREFIX brca1_gene:
  <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C17965
select distinct ?entity_uri ?entity_name where {
  ## Entity is a subclas of the anonymous class:
  ## Disease_Mapped_To_Gene some 'BRCA1 Gene'
  ?entity_uri
      rdfs:label ?entity_name;
      rdfs:subClassOf
          [rdf:type owl:Restriction;
             owl:onProperty disease_mapped_to_gene:;
             owl:someValuesFrom brca1_gene:] .
  ## Documents that have a term with semantic type
  ## that are subclasses of the entity
  ?document rdf:type dco:document;
            dco:has_part ?term .
  ?term dco:semantic_type ?entity_uri }
```

Figure 2: Query for documents containing terms whose semantic type has been mapped to the BRCA 1 gene.

In addition to leveraging the ontologies axioms, we can also examine the word context surrounding a term. For instance, this is useful for addressing the aforementioned issue of searching pathology reports in which the cells are found to be ER negative.

ACKNOWLEDGMENT

We gratefully acknowledge support from the Roswell Park Comprehensive Cancer Center Biomedical Data Science Shared Resource that is funded in part by Cancer Center Support Grant NCI P30CA16056.

REFERENCES

[1] Wright LW. (2007). NCI Thesaurus: A semantic model integratingcancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1): 30-43. doi: 10.1016/j.jbi.2006.02.013. PMID: 16697710.

[2] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc.* 2010;17(5):507–13. doi: 10.1136/jamia.2009.001560.

[3] Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, Hua Xu. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *JAMIA*. doi: 10.1093/jamia/ocx132.

[4] Tseytlin E, Mitchell K, Legowski E, Corrigan J, Chavan G, Jacobson RS. NOBLE - Flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics*. 2016 Jan 14;17:32. doi: 10.1186/s12859-015-0871-y.

[7] IRI: http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C4017
[8] http://graphdb.ontotext.com