A Web Crawler and SPARQL Query Search Agent to Expand and Navigate NPDS Semantic Metadata Records

Shiladitya Dutta[1], **Carl Taswell, MD, PhD**[2]
[1]**Brain Health Alliance, Ladera Ranch, CA, USA**

Introduction

The Nexus PORTAL-DOORS System (NPDS)[1] manages lexical metadata and semantic descriptions of resources. It has 3 principal components: Nexus diristry, PORTAL registry, and DOORS directory. This project interfaces with the DOORS directory and the semantic metadata stored within it. Semantic metadata descriptions are a prerequisite for many of the goals laid out for NPDS such as automated meta-analysis. However, linked data semantic descriptions are time consuming to manually annotate. Thus, we have developed a system that can extract semantic metadata from a variety of biomedical resources and provide a method of searching through that metadata via a SPARQL query.

Web Crawler and SPARQL Query Engine

The web crawler component retrieves articles and article metadata from DOAJ, Elsevier ScienceDirect, CORE[2], and PubMed via REST API. The websites' databases are searched through via their inbuilt natural language search functionality. Each article has its basic metadata(title, abstract, author(s), etc.) returned. From the abstract, semantic triples are extracted, placed in a graph, and converted to turtle format on a document-to-document basis. These turtle formatted files are then stored in a DOORS directory via the Scribe API. When a SPARQL query is called the program retrieves the graphs from the database to be queried via the SPARQL query engine.

Triple Extraction Process

The program extracts RDF triples from the unstructured text of the articles' abstracts utilizing NLTK[3] and the Stanford Core NLP[4] modules. First, constituency parsing is performed to create a tree from which the subject(s), predicate, and object(s)/adjective(s) are recorded. Then co-reference resolution and pronomial anaphora occur to recognize unique entities and ensure that their references are consistent throughout the graph. Once the logical-form triples are compiled, they are converted from lexical-based triples to valid RDF by identifying each part of the subject-verb-object triples. This is accomplished by using various databases (i.e. MeSH) for field-specific "jargon" and select named entities, word sense disambiguation for standard words, and literals for numerals and unlinked names.

Conclusion

Here we presented a method for searching through and retrieving semantic metadata from the open web to expand the Nexus PORTAL-DOORS System (NPDS) records. In order to perform this task a pipeline was developed consisting of a web crawler component to retrieve article metadata from external databases, a triple extraction process to derive logical form triples from the unstructured text of each article's abstract, and a SPARQL query engine to facilitate semantic metadata retrieval. These components mesh together to create a system that can furnish large amounts of linked semantic metadata for use by NPDS with relatively low time investment.

References

1. Taswell C, TeleGenetics G, Ladera Ranch CA. PORTAL-DOORS infrastructure system for translational biomedical informatics on the semantic web and grid. Proceedings of the American Medical Informatics Association Summit on Translational Bioinformatics, San Francisco, CA. 2008 Mar:43.
2. Knoth P, Zdrahal Z. CORE: three access levels to underpin open access. D-Lib Magazine. 2012 Nov;18(11/12).
3. Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."; 2009 Jun 12.
4. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. InProceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations 2014 (pp. 55-60).