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

Shiladitya Dutta, Degrees[1], Firstname B. Lastname, Degrees[2]
[1]Institution, City, State, Country (if applicable); [2]Institution, City, State, Country (if applicable)

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 SPARQL.

Web Crawler Component

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 title, abstract, DOI (if available), author(s), and publication date returned as basic metadata.

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 composition tree from which the subject(s), predicate, and object(s)/adjective(s) are recorded. Then co-reference resolution and pronomial anaphora occurs to recognize unique entities and ensure that their references are consistent throughout the graph. Once entities are identified, named entity extraction is completed by referencing the MeSH ontology(for biomedical terminology) and Wikipedia(for select named entities). Then, word sense disambiguation is performed on predicates and any remaining subjects to determine the WordNet synset for graph compression.

SPARQL Query Engine

The in-memory graphs are converted to turtle format. These turtle files are then stored in a DOORS directory via the Scribe API. When a SPARQL query is called the program retrieves the .ttl files from the DOORS directory and merges them into a singular in-memory graph. This singular graph can be queried via SPARQL.

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 the 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).