A Web Crawler and SPARQL Query Search Agent to Expand and Navigate NPDS Brain Literature Records

Shiladitya Dutta, Carl Taswell M.D., Ph.D.

2019-September-4

Abstract: The Nexus-PORTAL-DOORS System (NPDS) manages independent repositories of lexical and semantic renditions of resources. Analogous to IRIS-DNS, NPDS is organized into 3 server types: PORTAL registries for lexical metadata, DOORS directories for semantic metadata, and Nexus diristries which act as hybrid PORTAL-DOORS servers. Many of the goals laid out for the Nexus PORTAL-DOORS System require records of semantic descriptions, however it is labor-intensive to manually annotate resources from the open-web. To remedy this, we have created CoVaSEA (Concept-Validated Search Engine Agent): an automated web crawler/query engine. The intention of CoVaSEA is to provide a method for a user to search and expand the semantic records of brain literature in the DOORS directories. To this end, CoVaSEA integrates multiple features which benefit NPDS including: (A) An implementation of SPARQL query based search to allow retrieval and manipulation of DOORS descriptions, (B) Targeted web-crawling for relevant articles from external biomedical literature databases to expand NPDS records, and (C) Automated translation of free-form text into RDF triples to derive the semantic representations of lexical data. The system consists of a pipeline in which a web crawler registers articles by converting the abstract text and descriptions to RDF, so that a SPARQL query engine can comb through the retrieved material. The web crawler searches for articles in 4 biomedical literature databases, PubMed, Elsevier ScienceDirect, DOAJ, and Springer, via REST API. Each article has its basic metadata (title, abstract, author(s), etc.) returned. In order to translate the lexical information in the abstract into semantic metadata, CoVaSEA develops RDF descriptions of the articles' abstracts. First, constituency parsing is performed to create a tree from which the logical-form triples are extracted by searching for the subject(s), predicate(s), and object(s). Then the triples are processed by coreference resolution and pronomial anaphora to ensure that unique entities have consistent references. Once the logical-form triple processing is finished, they are converted from lexical-based triples to valid RDF by assigning URI's to each part of the subject-predicate-object triples. This is accomplished by using various databases (i.e. MeSH) for field-specific terminology and select named entities, word sense disambiguation for standard words, and literals for numerals and names. Finally, the graph is converted to RDF, yielding a file that portrays the natural-language information in each abstract in a linked data format. These 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. As a whole, CoVaSEA presents the capability to both search "externally" with the web crawler for semantic data to expand the DOORS knowledge base and "internally" with SPARQL to provide a method to navigate the data inside the DOORS directory. With the distinct advantage that the system is automated, CoVaSEA can furnish large numbers brain-related literature descriptions on a regular basis, thus laying the groundwork for a variety of future NPDS applications for which semantic metadata stores are a necessity.