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

Shiladitya Dutta, Carl Taswell, M.D., Ph.D.
Brain Health Alliance, Ladera Ranch, CA, USA

Abstract: Semantic descriptions are a prerequisite for many of the goals laid out for the Nexus PORTAL-DOORS System, however it is labor-intensive to manually annotate resources from the open-web. To remedy this, we have created CoVaSEA: an automated web crawler/query engine that can search and expand NPDS metadata records. With this system, articles from external biomedical databases can have semantic descriptions extracted from them for use in a variety of computational tasks.

Introduction: The Nexus PORTAL-DOORS System (NPDS)[1] manages lexical and semantic renditions of resources. It has 3 main components: the Nexus diristry, PORTAL registry, and DOORS directory. In order to both expand the DOORS records along with providing a method to search NPDS with a SPARQL query, a semantic section of CoVaSEA(Concept-Validated Search Engine Agent) was developed. Acting as an addition to past lexical-based CoVaSEA systems[2], we hope to furnish a means of expansion and search for NPDS linked data descriptions.

Methods: CoVaSEA integrates several features which benefit NPDS including: (A) An implementation of SPARQL query based semantic search to allow retrieval and manipulation of DOORS linked data descriptions (B) Targeted web-crawling for relevant article metadata from external biomedical literature databases to expand NPDS records (C) Automated translation of free-form text abstracts 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 description to RDF, so that a SPARQL query engine can comb through the retrieved material. The web crawler component searches for articles and their metadata in biomedical literature databases 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 triples from the unstructured text of the articles' abstracts. First, constituency parsing is performed to create a tree from which the subject(s), predicate, and object(s) are recorded in a method similar to Rusu D et al.[3]. Then co-reference resolution and pronomial anaphora occur to recognize unique entities and ensure that their references are consistent. Once the logical-form triple post-processing is finished, 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 names. Finally, the RDF graph is converted to the turtle format, yielding a file that portrays the natural-language information in each abstract in a linked data format. 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.

Conclusion: Here we presented a system in which resources from the open-web can be translated into machine-understandable semantic information and be searched via SPARQL. CoVaSEA has the capability to both search "externally" with the web crawler for semantic data to expand the NPDS 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, thus can furnish large amounts semantic descriptions on a regular basis, CoVaSEA lays the groundwork for a variety of future NPDS applications for which linked data stores are a necessity.

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. Bae SH, Craig AG, Taswell C. Expanding Nexus Diristries of Dementia Literature with the NPDS Concept-Validating Search Engine Agent.
3. Rusu D, Dali L, Fortuna B, Grobelnik M, Mladenic D. Triplet extraction from sentences. InProceedings of the 10th International Multiconference" Information Society-IS 2007 Jul (pp. 8-12).