From Papers to Triples

Semantic Analysis of Scientific Literature

Motivation

Finding relevant scientific literature is one of the essential tasks researchers are facing on a daily basis. Digital libraries and web information retrieval techniques provide rapid access to a vast amount of scientific literature. However, no further automated support is available that would enable fine-grained access to the knowledge ‘stored’ in these documents. The emerging domain of Semantic Publishing aims at making scientific knowledge accessible to both humans and machines, by adding semantic annotations to content, such as a publication’s contributions, methods, or application domains. However, despite the promises of better knowledge access, the manual annotation of existing research literature is prohibitively expensive for wide-spread adoption. We argue that a novel combination of three distinct methods can significantly advance this vision in a fully-automated way: (i) Natural Language Processing (NLP) for Rhetorical Entity (RE) detection; (ii) Named Entity (NE) recognition based on the Linked Open Data (LOD) cloud; and (iii) automatic knowledge base construction for both NEs and REs using semantic web ontologies that interconnect entities in documents with the machine-readable LOD cloud.

Results

We present a complete workflow to transform scientific literature into a semantic knowledge base, based on the W3C standards RDF and RDFS. A text mining pipeline, implemented based on the GATE framework, automatically extracts rhetorical entities of type Claims and Contributions from full-text scientific literature. These REs are further enriched with named entities, represented as URIs to the linked open data cloud, by integrating the DBpedia Spotlight tool into our workflow. Text mining results are stored in a knowledge base through a flexible export process that provides for a dynamic mapping of semantic annotations to LOD vocabularies through rules stored in the knowledge base. We created a gold standard corpus from computer science conference proceedings and journal articles, where Claim and Contribution sentences are manually annotated with their respective types using LOD URIs. The performance of the RE detection phase is evaluated against this corpus, where it achieves an average F-measure of 0.73. We further demonstrate a number of semantic queries that show how the generated knowledge base can provide support for numerous use cases in managing scientific literature.

Availability

Development releases of individual components are available on our GitHub page under open source licenses at https://github.com/SemanticSoftwareLab.

Scholarly User Profiling

Motivation

Scientists increasingly rely on intelligent information systems to help them in their daily tasks, in particular for managing research objects, like publications or datasets. The relatively young research field of Semantic publishing has been addressing the question how scientific applications can be improved through semantically rich representations of research objects, in order to facilitate their discovery and re-use. To complement the efforts in this area, we propose an automatic workflow to construct semantic user profiles of scholars, so that scholarly applications like digital libraries or data repositories can better understand their users’ interests, tasks, and competences, by incorporating these user profiles in their design. To make the user profiles sharable across applications, we propose to build them based on standard semantic web technologies, in particular the Resource Description Framework (RDF) for representing user profiles and Linked Open Data (LOD) sources for representing competence topics. To avoid the cold start

problem, we suggest to automatically populate these profiles by analyzing the publications (co-)authored by users, which we hypothesize reflect their research competences.

Results

We developed an open source library, ScholarLens, which can automatically generate semantic user profiles. For modeling the competences of scholarly users and groups, we surveyed a number of existing linked open data vocabularies. In accordance with the LOD best practices, we propose an RDF Schema (RDFS) based model for competence records that reuses existing vocabularies where appropriate. To automate the creation of semantic user profiles, we developed a complete, automated workflow that can generate semantic user profiles by analyzing full-text research articles through various natural language processing (NLP) techniques. The input to our library is a set of research articles for a given user. A knowledge base in RDF format is then populated through our NLP pipeline with user profiles containing the extracted competences. We evaluated our system through two user studies, resulting in mean average precision (MAP) of up to 95%. As part of the evaluation, we also analyze the impact of semantic zoning of research articles on the accuracy of the resulting profiles. Finally, we demonstrate how these semantic user profiles can be applied in a number of use cases, including article ranking for personalized search and finding scientists competent in a topic – e.g., to find reviewers for a paper.

Availability

Development releases of ScholarLens are available on our GitHub page at https://github.com/SemanticSoftwareLab/ScholarLens.

Additional Information

- Semantic representation of scientific literature: bringing claims, contributions and named entities onto the Linked Open Data cloud
- Semantic User Profiles: Learning Scholars' Competences by Analyzing their Publications