RESEARCH

Annotating the SIRO model and discovering experimental protocols

Olga Giraldo[1*], Alexander García[1], Tazro Ohta[1,2] and Federico López[3]

*Correspondence:
ogiraldo@fi.upm.es
[1]Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain
Full list of author information is available at the end of the article

Abstract

We are manually annotating a corpus of 100 full-text experimental protocols in cell biology, neuroscience, developmental biology, microbiology and molecular biology. We are gathering the protocols from repositories like bio-protocols, Cold Spring Harbor Protocols and Nature protocols. Our annotation focuses on: the SIRO model, namely: i) the Sample tested, ii) the Instruments employed, iii) the Reagents used and, iv) the overall Objective of the protocol. The SIRO model represents the minimal information for describing an experimental protocol. Our manual annotation experience illustrates how to use annotations from domain experts in the enrichment of an ontology; we are identifying key terms in the text and relating these to concepts in the ontology. By the same token, our annotation experience is supporting our NLP infrastructure; the identification of specific facets, S I R O, allows us to enrich our gazetteers and better tailor our JAPE rules. Our approach supports answering queries such as we support queries such as "transformation protocol (objective)", of "*Saccharomyces cerevisiae* (sample)" by using "lithium acetate/single-stranded carrier (reagent)".

Keywords: semantic web; graph theory; ontologies; natural language processing; knowledge representation

Background

We are manually annotating a corpus of 100 full-text experimental protocols in cell biology, neuroscience, developmental biology, microbiology and molecular biology. We are gathering the protocols from repositories like bio-protocols, Cold Spring Harbor Protocols and Nature protocols. Our annotation task focuses on the annotation of four common elements found across our corpus of protocols. These elements are "the SIRO model", namely: i) the Sample tested, ii) the Instruments employed, iii) the Reagents used and, iv) the overall Objective of the protocol. The SIRO model represents the minimal information for describing an experimental protocol. SIRO extends and structures available metadata for experimental protocols, for instance, author, title, date, journal, abstract, and other properties that are available for published protocols.

For the annotation task, SIRO anchors the annotations. The SIRO model derives from the SMART Protocols ontology [1]. The analysis of the annotated entities allows us to identify terminology that widens the coverage and improves the accuracy of the ontology. The terminology thus gathered also enriches our gazetteers; these are pre-classified lists containing names of entities such as samples, instruments, reagents etc and metadata. These lists are used to find occurrences in text, e.g.

named entity recognition. The JAPE rules are implemented to recognize regular expressions in the documents; these rules use the gazetteers and the corresponding metadata to perform complex extraction of specific expressions, for instance actions over samples. In this way we support queries such as "transformation protocol (objective)", of "*Saccharomyces cerevisiae* (sample)" by using "lithium acetate/single-stranded carrier (reagent)". These kinds of queries combine terminology and make use of JAPE rules and gazetteers in GATE [2] – the NLP engine we are using. The information extraction is performed by linguistic pre-processing (tokenization, POS tagging), followed by a named entity recognition component that uses gazetteers and rule-based grammar techniques.

We are currently annotating 100 published protocols; our corpus of documents only has PDF files. For each protocol we have three expert annotators; annotations are kept private so that annotators can only see their own annotations. From our annotations we are calculating the inter-annotator agreements (overall fleis's kappa between 0.75 – 1.00). This allows us to know the agreement amongst annotators with respect to any given SIRO facet. We are exposing our annotations over a SPARQL endpoint; our annotation infrastructure was built upon Hypothesis. This is an open source community effort that facilitates the annotation of PDFs and HTML. We extended Hypothesis so that we could focus on the facets that we needed to annotate; we also removed some parts of the original interface and connected it to annotation services from BioPortal and GATE.

Our manual annotation experience illustrates how to use annotations from domain experts in the enrichment of an ontology; we are identifying key terms in the text and relating these to concepts in the ontology. By the same token, our annotation experience is supporting our NLP infrastructure; the identification of specific facets, S I R O, allows us to enrich our gazetteers and better tailor our JAPE rules. The resulting corpus of annotated documents has the added value of being verifiable; unlike other corpuses, ours is available over the annotation tool. This makes it easy for researchers to review the annotations and extend the corpus as needed; researchers can load the corpus over the tool or programmatically via the API.

The combination of semantic and NLP is the pillar for the development of a platform that works as a recommendation system of experimental protocols. We envision a platform that facilitates the generation of protocols that are to be born semantics as well as the registry of existing protocols that are NLP processed in order to make the semantics explicit. Experimental protocols are in this way machine processable and therefore an integral component of the web of data. This allows finding protocols according to specific samples, availability of equipment and reagents in the lab and the purpose of the protocol. It also makes it possible for researchers to reuse parts of published protocols keeping track of provenance; for recommendation systems to accurately "recommend" what protocols to use depending on specifics of the researchers. Moreover, experimental actions semantically represented could easily be consumed by software agents, e.g. from robots, and maintained by humans.

During BLAH2016 we want to i) share our experience, ii) share our code base, iii) discuss the RDF model for experimental protocols, iv) align our effort with PubAnnotation, perhaps generating a branch of PubAnnotation addressing experimental protocols, v) understand how could our linked data make use of annotations available in PubAnnotation, specify use cases and modify our model if needed.

Author details
[1]Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain. [2]Database Center for Life Science, Tokyo, Japan. [3]Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle, Cali, Colombia.

References

1. Giraldo, O., García, A., Corcho, O.: Smart protocols: semantic representation for experimental protocols. In: Proceedings of the 4th International Conference on Linked Science-Volume 1282, pp. 36–47 (2014). CEUR-WS. org
2. Cunningham, H., Tablan, V., Roberts, A., Bontcheva, K.: Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS Computational Biology **9**(2), 1002854 (2013). doi:10.1371/journal.pcbi.1002854