Semantic Publishing Challenge 2015: Supplementary Material

- Literature Management
- Semantic Publishing
- Semantic Computing
- NLP
- Text Mining

Overview

This page provides supplementary material for our submission to the Semantic Publishing Challenge 2015 co-located with the Extended Semantic Web Conference (ESWC 2015).

We present an automatic workflow that performs text segmentation and entity extraction from scientific literature to primarily address Task 2 of the challenge. The proposed solution is composed of two subsystems:

- A text mining pipeline, developed based on the GATE framework, which extracts structural and semantic entities, such as, authors' information and citations, from text and produces semantic (typed) annotations.
- A flexible exporting module that translates the document annotations into RDF triples according to a custom mapping file.

Data Model

Our text mining pipeline accepts documents in the training and evaluation datasets and generates semantic annotations as output. For each workshop proceeding, we created a *copus* that contains the documents within that proceeding's volume. Various entities desired for the challenge are extracted using multiple gazetteer (dictionary) lists and hand-crafted rules. Therefore, the challenge's designated entities are all in form of annotations and attached to each document using our custom PUBO vocabulary. Subsequently, the annotations are distinguished by their semantic types and features during the querying phase.

While the type of annotations, e.g. Affiliation, is determined by the text mining component, we still would like to have the flexibility to express the mapping of annotations to RDF triples and their inter-relations at run-time. This way, various representations of knowledge extracted from documents can be constructed based on the intended use case and customized without affecting the underlying syntactic and semantic processing components. We designed an RDF Mapper component in our text mining pipeline that accepts a mapping file as input and transforms the designated document’s annotations into their equivalent RDF triples. For each annotation type that is to be exported, the mapping file has an entry that describes: (i) the annotation type in the document and its corresponding semantic type, (ii) the annotation’s features and their corresponding semantic type, and (iii) the relations between exported triples and the type of their relation.

Following the best practices on Linked Open Dataset creation, we tried to reuse the existing Linked Open Vocabularies in our mapping file, as much as possible.

Queries for Task 2

Here, we provide a natural language description of each of our submitted queries.

Q2.1: Affiliations in a paper

Affiliations are organization named entities found in the Metadata body (i.e., everything from the beginning of each document until the Abstract) of each paper. All Affiliation entities are annotated with foaf:Organization as their semantic type. We also annotate each detected person's fullname in the Metadata body as an Author and use foaf:Person as its semantic type.

Next, we attempt to detect the relations between Authors and Affiliations using various heuristics (e.g., matching indices, offset proximity). Once a match is found, we relate the Author instance and its Affiliation using the rel:employedBy property.

Q2.2: Papers from a country

Our text mining pipeline attempts to guess the location of each Affiliation annotation in text. The geographical location of each annotation (typically, the country name) is related to each Affiliation instance using the gn:locatedIn property.

Essentially, Q2.2 is similar to the previous query, except that it is possible to filter the results by comparing the string value of the affiliation's location.

Q2.3: Cited Works

Cited works are Reference annotations in the document detected by the pipeline in the Reference body of each document. The query distinguishes Cited Works annotation in a given document using the swrc:Publication semantic type. For each matched instance, we return the title of the cited item.

Q2.4: Recent Cited Works

Q2.4 is similar to the previous query, except that it is possible to filter the results by comparing the value of the fabio:hasPublicationYear predicate of the cited works.

Q2.5: Cited Journal Papers

Q2.5 is similar to the Q2.3 query, except that it is possible to filter the results by comparing the string value of the affiliation's type from the set of {"journal","proceedings"}.

Q2.6: Research grants

Query 2.6 tries to retrieve all annotations in documents that are of type frapo:Grant.

Q2.7: Funding agencies

Query 2.7 tries to retrieve all annotations in documents that are of type frapo:FundingAgency.

Q2.8: EU projects

We do not address Query 2.8 of the challenge.

Q2.9: Related ontologies

We capture mentions of ontology names, including acronyms, camel case or concatenated ontology names in each paper using our text mining pipeline, but remove the ones that our outside the span of the Abstract section of each document before the exporting phase. Therefore, Q2.9 merely retrieves all annotations in documents that are of type owl:Ontology.

Q2.10: New ontologies

We distinguish the new ontology annotations in a document, only if they appear within the boundary of a Contribution annotation. We use our Metadiscourse Extraction component (see [1]) to annotate the sentences in the document's abstract. Then, we update the status of each Ontology annotation contained in a Contribution annotation to "new".

Therefore, Q2.20 is similar to Q2.9, except that we look for Ontology annotations that have "new" as value of their opmw:hasStatus predicate.

References

1. Sateli, B., and R. Witte, "What's in this paper? Combining Rhetorical Entities with Linked Open Data for Semantic Literature Querying", *Semantics, Analytics, Visualisation: Enhancing Scholarly Data (SAVE-SD 2015)*, Florence, Italy : ACM, pp. 1023–1028, 05/2015.

Source URL (retrieved on *2025-12-22 07:56*): https://www.semanticsoftware.info/sempub-challenge-2015