Semantify CEUR-WS Proceedings: towards the automatic generation of highly descriptive scholarly publishing Linked Datasets

Francesco Ronzano, Gerard Casamayor del Bosque, and Horacio Saggion

TALN Research Group, Universitat Pompeu Fabra,
C/Tanger 122, 08018 Barcelona, Spain
{francesco.ronzano,gerard.casamayor,horacio.saggion}@upf.edu

Abstract. Rich and fine-grained semantic information describing varied aspects of scientific productions is essential to support their diffusion as well as to properly assess the quality of their output. To foster this trend, in the context of the ESWC2014 Semantic Publishing Challenge, we present a system that automatically generates rich RDF datasets from CEUR-WS workshop proceedings. Proceedings are analyzed through a sequence of processing phases. SVM classifiers complemented by heuristics are used to annotate missing CEUR-WS markups. Annotations are then linked to external datasets like DBpedia and Bibsonomy. Finally, the data is modeled and published as an RDF graph. Our system is provided as an on-line Web service to support on-the-fly RDF generation. In this paper we describe the system and present its evaluation following the procedure set by the organizers of the challenge.

Keywords: Semantic Web, Information Extraction, Scholarly Publishing, Open Linked Data

1 Information extraction and knowledge modeling approach: motivation and overview

The enhancement of scholarly publishing data by better structuring, interlinking and semantically modeling is one of the core objectives of **semantic publishing** [1] [2]. Semantic Web technologies are an enabling factor towards this vision [3]. They provide the means to structure and semantically enrich scientific publications so as to support the generation of Linked Data from them [4] [5]. In this context, recently, a few scientific publication repositories including DBLP[1], ACM[2] and IEEE[3] have been also published as Open Linked Data. In general, however, only basic bibliographic information is exposed that is too generic to

The work described in this paper has been funded by the the European Project Dr Inventor (FP7-ICT-2013.8.1 - Grant no: 611383).

[1] http://dblp.l3s.de/d2r/
[2] http://acm.rkbexplorer.com/
[3] http://ieee.rkbexplorer.com/

properly support the diffusion and the assessment of the quality of scientific publications.

With the purpose of experimenting with new approaches to generate rich and highly descriptive scholarly publishing Open Linked datasets, and in the context of the Task 1 of the ESWC2014 Semantic Publishing Challenge (SemPub Task 1), we present a system that automatically mines contents from the workshop proceedings of CEUR-WS Web portal and exports them as an RDF graph. SemPub Task 1 aims at enabling the computation of indicators of the quality of workshops. For this purpose participants are provided with a dataset of HTML documents with information about workshop proceedings indexed at CEUR-WS together with the PDF documents of the related papers. Microformats[4] and RDFa[5] annotations are available for some of these documents, and missing in others. The task requires extracting pieces of information from these textual and unstructured sources and model these contents as an RDF graph so as to enable the computation of the indicators of the quality of a workshop by means of SPARQL queries.

Our approach to Sem Pub Task 1 is based on the following core observations:
- Since 2010, HTML pages detailing the contents of individual proceeding volumes have been annotated with 20 microformat classes (CEURVOLEDITOR, CEURTITLE, CEURAUTHORS, etc.). The occurrences of each class provide a set of examples of relevant kinds of information that need to be extracted and modelled in SemPub Task 1. **This data can be exploited to train an automatic text annotation system**.
- There are several scholarly publishing resources accessible on-line where part of the information published by CEUR-WS proceedings is linked and partially replicated in a structured or semi-structured format. Some examples are Bibsonomy[6], DBLP, and Wiki CFP[7]. These resources can be exploited **to support the information extraction process and to make the RDF contents generated by our system strongly linked with related datasets**.

Starting from these observations, we have designed and implemented a data processing workflow to convert the CEUR-WS proceeding documents into rich RDF datasets.

2 System description

This Section describes in detail each phase of the processing workflow outlined in Figure 1. This workflow was implemented using the text processing pipeline provided by GATE Text Engineering Framework[8] [8], and complemented by

[4] A semantic markup approach that conveys metadata and other attributes in Web pages by existing HTML/XHTML tags.
[5] A semantic markup useful to embed RDF triples within XHTML documents.
[6] http://www.bibsonomy.org/
[7] http://www.wikicfp.com/cfp/
[8] https://gate.ac.uk/

external tools and interactions with on-line Web services and knowledge repositories[9].

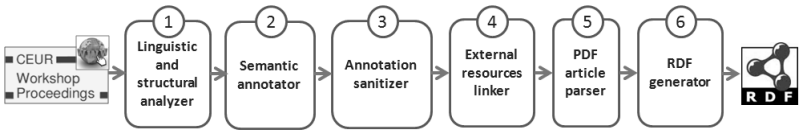

Fig. 1. CEUR-WS Proceeding to RDF: data processing work-flow

PHASE 1: Linguistic and structural analyzer: given a set of CEUR-WS proceeding Web pages, their contents are characterized by means of both linguistic features and structural information derived from HTML markup so as to support the execution of the following processing steps. Texts are split into lines containing homogeneous information according to a set of heuristics, then tokenized and POS-tagged using the information extraction framework ANNIE[10]. Names of paper authors, conference titles and acronyms are identified using gazetteers that were derived from the XML dump of DBLP and WikiCFP.

PHASE 2: Semantic annotator: this component automatically adds semantic annotations to the textual contents of HTML documents without semantic markups. The semantic annotator is based on a set of chunk-based and sentence-based Support Vector Machine (SVM) classifiers [9] trained with microformat annotations of the documents of 562 proceeding volumes that do have annotations. Considering the 14 most frequent microformat classes adopted by CEUR-WS, we compiled 14 training corpora. Each corpus includes all the documents that are annotated with the corresponding microformat class. Since editor affiliations constitute relevant information for SemPub Task 1 that is not marked by CEUR-WS microformat classes, we introduced a new dedicated type of annotation, CEURAFFILIATION. We created an additional training corpus by randomly choosing 75 proceedings among the previous set of volumes. We manually annotated these proceedings with editor affiliations, thus generating 252 training examples. The set of features used to characterize chunks and sentences are the linguistic and structural information added to each proceeding by PHASE 1. For each annotation type we trained a chunk-based and a sentence-based SVM classifier, evaluated both classifiers using 10-fold cross-validation over the related training corpus and chose the one with the best F1 score, achieving F1 values of 0.9 or greater for all annotation types.

PHASE 3: Annotation sanitizer: a set of heuristics are applied to fix cases when the annotation borders are incorrectly identified or to delete annotations that are not compliant with the normal sequence of annotations of a proceeding

[9] The system described in this paper can be accessed on-line at: http://sempub.taln.upf.edu/eswc2014sempub/ (password: ceurrdf2014).

[10] http://gate.ac.uk/sale/tao/splitch6.html

(e.g. editor affiliations annotated after the list of paper titles and authors). In addition, links between pairs of related annotations are created (e.g. editors and affiliations by means of their markups).

PHASE 4: External resources linker: existing annotations are enriched with additional information not covered internally by CEUR-WS and useful to calculate the workshop quality indicators. The Bibsonomy REST API is used to associate annotations of paper titles to Bibsonomy entries and to import the related BibTeX metadata and links to external repositories like DBLP. DBpedia Spotlight Web service [7] is used to identify the DBpedia resources corresponding to mentions of states, cities and organizations.

PHASE 5: PDF article parser: this component retrieves the PDF files of the papers of a proceeding and parses them by means of the Apache PDFBox Java library. The number of pages of each paper is extracted and mentions of states, cities and organizations in the PDF document are identified using the DBpedia Spotlight Web service.

PHASE 6: RDF generator: all the information gathered by the previous processing steps is aggregated and normalized so as to generate an Open Linked RDF dataset that is informative enough to compute the quality indicators defined in SemPub Task 1. The contents of each Proceeding are modeled by reusing and extending two widespread semantic publishing ontologies: the *Semantic Web for Research Communities Ontology* (prefix swrc) and the *Publishing Role Ontology* (prefix pro).

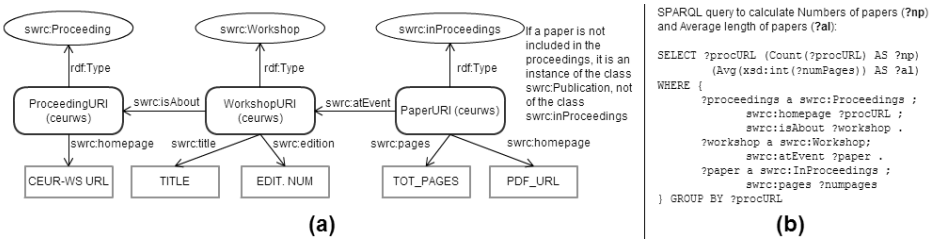

Fig. 2. (a) RDF model of papers presented at a workshop, included in a proceeding volume; (b) SPARQL query for **Numbers of papers** and **Average length of papers**

We defined - in a ceur-ws namespace - a few new classes and properties that were required to fully model the annotations produced by our system. For all these new resources we subclassed existing classes and properties in the ontologies before mentioned. The resulting RDF dataset imports the reused vocabularies and asserts all pertinent T-BOX axioms for the new concepts and properties.

In Figure 2 we provide an example of the SPARQL query formulated to compute two of the workshop quality indicators of SemPub Task 1, starting from the proceeding information modelled by the RDF graph shown in the same Figure.

3 Lessons learned and future work

Our system design has been motivated by the need of flexibility and robustness in the face of different ways in which information is written, structured or annotated in the input dataset. Although relying on external Web services provides us with useful information, the responsiveness of our Web interface is compromised by the lag-times introduced by the access to these external services. Overall, however, we think that the benefits outweigh the problems and that reusing existing datasets and services is good practice.

In the future we plan on increasing the flexibility of the information extraction procedure by doing further experimentation with statistical and trainable methods. Another venue of research is to increase the role of external services and datasets to inform and complement the content extraction process, specially in the face of missing annotations, ambiguous text or unstructured documents. Finally, we would like to generalize our approach so as to be able to extract and merge RDF datasets mined from distinct on-line Web Portals of academic documents and data.

References

1. Shotton, D.: Semantic publishing: the coming revolution in scientific journal publishing. Learned Publishing 22.2, pp. 85-94 (2009)
2. Eefke, S., and Van Der Graaf, M.: Journal article mining: the scholarly publishers' perspective. Learned Publishing 25.1, pp. 35-46 (2012)
3. Bizer, C.: Linking Data & Publications Expert Report, Global Research Data Infrastructure of European Union (2012)
4. Ciancarini, P., Di Iorio, A., Nuzzolese, A. G., Peroni, S., and Vitali, F.: Semantic Annotation of Scholarly Documents and Citations. In AI* IA 2013: Advances in Artificial Intelligence, pp. 336-347, Springer International Publishing (2013)
5. Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S. R., and Thorne, D.: Utopia documents: linking scholarly literature with research data. Bioinformatics, 26(18), pp. 568-574 (2010)
6. Shotton, D., Portwin, K., Klyne, G., and Miles, A.: Adventures in semantic publishing: exemplar semantic enhancements of a research article. PLoS computational biology, 5(4), e1000361 (2009)
7. Mendes, P. N., Jakob, M., Garca-Silva, A., and Bizer, C.: DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th International Conference on Semantic Systems, pp. 1-8, ACM (2011)
8. Cunningham, H., Maynard, D., Bontcheva, K., and Tablan,V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, ACL (2002)
9. Li, Y., Bontcheva, K. and Cunningham, H.: Adapting SVM for Data Sparseness and Imbalance: A Case Study on Information Extraction. Natural Language Engineering, Volume 15, pp. 241-271, Cambridge University Press (2009)