From Research Articles to Knowledge Graphs

Methods for Ontology-driven Knowledge Base Creation from Text

Vayianos Pertsas
Department of Informatics
Athens University of Economics and Business, vpertsas@aueb.gr

Panos Constantopoulos
Department of Informatics
Athens University of Economics and Business; Digital Curation Unit, Athena Research Centre,
panosc@aueb.gr

Abstract

Understanding and extracting knowledge contained in text and encoding it as linked data for the WEB is a highly complex task that poses several challenges, requiring expertise from different fields such as conceptual modeling, natural language processing and web technologies including web mining, linked data generation and publishing, etc. When it comes to the scholarly domain, the transformation of human readable research articles into machine comprehensible knowledge bases is considered of high importance and necessity today due to the explosion of scientific publications in every major discipline, that makes it increasingly difficult for experts to maintain an overview of their domain or relate ideas from different domains. This situation could be significantly alleviated by knowledge bases capable of supporting queries such as: find all papers that address a given problem; how was the problem solved; which methods are employed by whom in addressing particular tasks; etc. that currently cannot be addressed by commonly used search engines such as Google Scholar or Semantic Scholar.

This tutorial addresses the above challenge by introducing the participants to methods required in order to model knowledge regarding a given domain, extract information from available texts using advanced machine learning techniques, associate it with other information mined from the web in order to infer new knowledge and republish everything as linked open data on the Web. To this end, we will use a specific use case – that of the scholarly domain, and will show how to model research processes, extract them from research articles, associate them with contextual information from article metadata and other linked repositories and create knowledge bases available as linked data. Our aim is to show how methodologies from different computer science fields, namely natural language processing, machine learning and conceptual modeling, can be combined with Web technologies in a single meaningful workflow.

WWW '19, May 13-17, 2019, San Francisco, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320090

CCS CONCEPTS

• Computing methodologies~Information extraction • Computing methodologies~Ontology engineering • Computing methodologies~Learning paradigms

Author Keywords

Conceptual Modeling; Ontology Engineering; Ontology Population; Information Extraction from Text; Machine Learning Methodologies; Linked Data

ACM Reference format:

Vayianos Pertsas and Panos Constantopoulos. 2019. From Research Articles to Knowledge Graphs: Methods for Ontology-driven Knowledge Base Creation from Text. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3308560.3320090

1 Topic and Relevance:

In this tutorial we will present a variety of methods covering the entire workflow of transforming human readable text (in particular that of research articles) into machine comprehensible linked data available on the Web. To this end we will: 1) introduce the participants to essential instruments for knowledge organization such as taxonomies and ontologies and will show how these can be used in a specific use case -that of modeling scholarly work- and 2) tutor the participants into using various methods for ontology population / knowledge base creation -based on our use case scenario- that cover mining information from publication metadata, extracting information from text using advanced machine learning techniques and ontology-based guidance, and combining those with mined resources from linked data repositories such as DBpedia and ORCID to infer new knowledge. Both parts will include a theoretical overview to explain the concepts and methods presented, as well as a presentation of the actual implementations in the form of an ontology for the first part and a system implemented in Python for the second part.

Transforming human readable text into knowledge graphs consisting of linked data available on the web is an active

research topic of high relevance to the WWW community since it requires bridging different Web technologies with methods from other computer science fields such as machine learning, natural language processing and conceptual modeling. The use case of modeling research processes and creating knowledge bases from research articles underlines the importance of the task, given the constant increase in the production and publishing of scientific papers in every discipline. Assisting researchers in answering queries that are currently beyond the capabilities of standard search engines, such as Google Scholar, Scopus or Semantic Scholar, which mostly navigate through author or citation graphs, can improve access to scientific literature and enhance researcher productivity. The construction of knowledge bases capable of answering complex queries of the form "*who* has done *what, when, why, how* and with *what* results" requires encoding information about research processes in a manner that enables reasoning and guides the extraction of information from publications, as well as the direct documentation of research activities. To this end, the level of expertise (with parts of our work published in journals and conferences such as IJDL [1], TPDL [2] , and ISWC [3]) as well as the long experience of the tutors on the subject, along with the scope of the tutorial, covering the state of the art in methodologies from knowledge representation and organization to information extraction from text and knowledge base creation, qualifies for a high-quality introduction to the topic.

2 Duration:

The proposed duration of the tutorial is half day, divided into two parts, where the exact division of the content will be adapted on demand to the participants. The following subjects will be covered:

First Part:

- Elements of knowledge organization (Controlled vocabularies, Taxonomies, Ontologies)
- Elements of Linked Data and Semantic Web Data Models (RDF, RDFS, OWL)
- Methodology and design patterns for modeling a specific domain. Use case: Scholarly Ontology for modeling scholarly work

Second Part:

- Knowledge base creation via ontology population. Use case: Populating the core concepts of Scholarly Ontology
- Distant supervision techniques for training Named Entity Recognizers of non-common type
- Information extraction from text using syntactic analysis and ontology-based semantics
- Entity and Relation Extraction from text using advanced Machine Learning techniques: feature engineering, embeddings, token/text representation.
- Metadata association and Linked Data integration and generation

3 Interaction Style:

Participants will be introduced to the methods and concepts covered in this tutorial through thorough presentations followed by Q&As and discussion on the practicalities of each covered subject. No specific equipment is required for attending the tutorial.

4 Audience:

Our target audience includes researchers and practitioners in web technologies, machine learning and information modeling with an interest in natural language processing and linked data methodologies. A background in computer science and some general familiarity with relevant concepts (ontologies, taxonomies, distant supervision, entity and relation extraction from text, linked data generation and integration as well as machine learning methods) are welcome but not necessary, as a brief overview of the employed concepts and methods will be included in the tutorial. Our aim is not to delve into the details of those concepts and methods, but to show how to use them and how they can be integrated into a meaningful workflow.

We expect the audience to take away a substantial overview of the state-of-the art in methodologies covering the entire workflow from designing models that capture knowledge in a specific domain (in our case that will be the domain of scholarly work), to the creation of knowledge bases by populating these models with information extracted from text (using machine learning techniques) along with linked data or publication metadata mined from the web.

5 Previous Editions:

This tutorial has never been presented in its entirety before (both parts covering conceptual modeling AND knowledge base creation). However, variations of the first part (conceptual modeling of research practices in the Digital Humanities) have been presented in the past in workshops with high attendance (more than 30 participants).

6 Tutorial Material:

Material will include tutorial notes, RDF serializations of the conceptual models, trained machine learning classifiers and samples of the actual produced knowledge bases (in the form of RDF triples), to be provided in time. No copyright issues are involved.

7 Equipment:

No additional equipment is needed for this tutorial.

8 Organizers:

Vayianos Pertsas holds a Dipl. Eng. Degree in Electrical and Computer Engineering from the University of Patras and a PhD degree in Informatics from the Athens University of Economics and Business. His research interests evolve around conceptual modeling and ontology population using information extraction

techniques that mainly focus on leveraging linked data and its applications, along with NLP and machine learning techniques. He has worked in the development of the NeDiMAH Methods Ontology in the ESF-funded Network for Digital Methods in the Arts and Humanities and in its evolution and operationalization in the form of the Scholarly Ontology. He has authored papers appearing in IJDL, ISWC, TPDL and co-tutored various workshops.

Panos Constantopoulos is Professor in the Department of Informatics, Director of the MSc Programme in Digital Methods for the Humanities, and former Dean of the School of Information Sciences and Technology, Athens University of Economics and Business. He is also affiliated with the Information Management Systems Institute of the "Athena" Research Centre, where he heads the Digital Curation Unit. He has previously been Professor and Chairman in the Department of Computer Science, University of Crete (1986-2003). From 1992 to 2003 he was head of the Information Systems Laboratory and the Centre for Cultural Informatics at the Institute of Computer Science, Foundation for Research and Technology - Hellas.

His scientific interests include: knowledge representation and conceptual modelling, ontology engineering, semantic information access, process mining, knowledge management systems, decision support systems, cultural informatics, digital libraries, digital curation and preservation.

He holds a Diploma in Electrical and Mechanical Engineering from the National Technical University of Athens (1978), a Master of Science in Electrical Engineering from Carnegie-Mellon University (1979) and a Doctor of Science in Operations Research from Massachusetts Institute of Technology (1983).

He has been principal investigator or scientific responsible on the part of his affiliation in 40 national or international research and development projects, in 13 of which project coordinator. He is currently heading "APOLLONIS-Greek Infrastructure for Digital Arts, Humanities and Language Research and Innovation", a three year, 4M Euro project jointly advancing the Greek components of CLARIN and DARIAH.

He has published over 100 articles in scientific journals, conference proceedings or as book chapters.

References

[1] Pertsas, V., Constantopoulos, P.: Scholarly Ontology: modelling scholarly practices. International Journal on Digital Libraries. 18, 173–190 (2017). doi:10.1007/s00799-016-0169-3

[2] Pertsas, V., Constantopoulos, P.: Ontology-driven information extraction from research publications. Lecture Notes in Computer Science. 11057 LNCS, 241–253 (2018). doi:10.1007/978-3-030-00066-0_21

[3] Pertsas, V., Constantopoulos, P.: Ontology-Driven Extraction of Research Processes. 1–16. doi:10.1007/978-3-030-00671-6_10