A Knowledge Graph for Humanities Research

Pertsas, Vayianos
vpertsas@gmail.com
Athens University Of Economics and Business

Leontaridis, Panagiotis
leontpng@gmail.com
Athens University Of Economics and Business

Kasapaki, Marialena
kasapakimariael@gmail.com
Athens University Of Economics and Business

Constantopoulos, Panos
panosc@aueb.gr
Athens University Of Economics and Business

The steep increase of research publications in every major discipline makes it increasingly difficult for experts to maintain an overview of their domain, increases the risk of missing new work or reinventing solutions, and makes it harder to relate ideas from different domains. The latter becomes of high concern in multidisciplinary fields, like Digital Humanities, where maintaining a cross-disciplinary overview of what goes on in terms of research goals, activities and methods is even harder. This situation could be significantly alleviated by supporting information searches such as: *find all papers that address a given problem*; *how was the problem solved*; *which methods are employed by whom in addressing particular research goals*; etc. Answering queries like these essentially requires access to information that could be compiled interactively, or automatically extracted from research publications, finally offered in a structured form suitable for supporting semantic queries. Note that search engines widely used by researchers, such as Google Scholar [1], Scopus [2] or Semantic Scholar [3] mostly leverage bibliographic metadata, while knowledge expressed in the actual text is exploited mostly by matching query terms to documents. In this paper we present a Knowledge Graph (KG) specifically designed for matching the above information needs of researchers in Humanities. The KG is derived from a large multidisciplinary dataset from the JSTOR [4] repository, which has undergone various NLP processes so that information from each article could be properly extracted, combined with metadata and other information from the Web and finally transformed into RDF triples available as linked data. The entire process was driven by Scholarly Ontology (SO) (Pertsas *et al.* 2017), specifically designed for capturing research processes. A specialization of SO for DH is known as NeMO.

The original dataset consisted of 25,681 papers -produced by OCR on the original scanned files- from various disciplines such as Archeology, Paleontology, Social Sciences, Anthropology, etc. years 2000-2021. After a shallow rule-based cleaning to remove references sections and titles, the text was split into sentences using the SpaCy [5] NLP framework. However, since the full text of each paper was the outcome of OCR, it had to be appropriately cleaned from noise elements such as unrecognized characters, tables, footnotes, references, section headings, etc. Furthermore, text deriving from scanned two-column papers yielded incomprehensible material that also had to be identified and removed. To this end we trained a Deep Learning text classifier using Hugging-Face [6] BERT-base-uncased Transformer in order to recognize if a given sentence is proper or noisy, based on a manually curated dataset of 10,000 sentences (half of which were identified as clean and the rest as noise). Evaluation of our model yielded 95.8% F1 score for the text classification task. The classifier was then applied to the original dataset filtering 3,700,000 cleaned sentences.

Next, we trained three Deep Learning Entity Recognizers using Hugging-Face RoBERTa-base Transformers in order to identify and extract three core types of entities of SO, namely: 1) *Activities* (i.e. actual research processes or steps thereof, like an archeological excavation, an anthropological study, an experiment, etc. carried out by the researchers-authors of the paper); 2) *Methods* (i.e. procedures employed by researchers to carry out research activities, like an algorithm or a specific technique, which appear as named entities in text) and 3) *Goals* (i.e. the research tasks that were addressed by the researchers through their activities). Our training set consisted of 10,000 sentences, containing approximately 7200 Methods, 4200 Activities and 1800 Goals, deriving from 3,082 papers. It was manually annotated by 3 annotators who, after appropriate training, reached inter-annotator agreement higher than 85% (Kappa statistic) for every task. Evaluation of our classifiers yielded F1 scores: 87.4% for Methods, 81.7% for Activities and 88.2% for Goals respectively. Performance depends on the syntactic complexity of textual spans: Goals are syntactically clearer, while Activities in passive voice with the agent missing proved a major source of errors.

After entity extraction, additional post-processing rules were applied in order to infer semantic relationships among the extracted activities with methods and goals respectively. These rules are based on the SO definitions for *employs(Activity,Method)* and *hasGoal(Activity,Goal)* relationships and the proximity of their corresponding entities' manifestations in text. Specifically, for each *employs(Activity,Method)* the corresponding textual spans of activity and method must overlap, while for *hasGoal(Activity,Goal)*, co-appearance of the corresponding activity and goal in the same sentence is necessary. Evaluation of those rules on a test-set of 1000 cases for each relationship type, yielded F1 scores: 96.4% and 98.1% respectively.

Finally, we extracted information from metadata regarding the authors of the articles (further matched, when possible, with ORCID [7] using the provided API), publication information and author keywords. The KG was produced in RDF [8] data format, using the NIF [9] model for the URIs of the entities derived from text, which were then interrelated -when appropriate- and connected with those extracted from article's metadata. Through this procedure, each paper is transformed into approximately 200 triples, on average.

Generating KGs for scientific literature is an active research topic with many endeavors like (Steenwinckel *et al.* 2020, Färber/Michael 2019) focusing on interconnecting bibliographic information of papers and authors, while others like (Dessì *et al.* 2021, D'Souza et al. 2022) leveraging out-of-the-shelf NER solutions for extraction of *named* entities (e.g. material, task, dataset, etc.) in specialized domains of literature. To the best of our knowledge, our KG is currently the only effort that concurrently addresses the problems of extracting information from articles' full text, dealing with noisy OCRed material and semantically complex (and of variable length) entities like research activities and

goals, while focusing on the domain of Humanities research. In addition, the inference of relationships between the extracted entities allows for better understanding and representation of their semantic context (e.g. the research process during which a method was employed, the reason for its employment, etc.) making it possible to address complex queries such us the ones described above. Future work involves expanding our KG with recognition and extraction of other SO entities such as researchers' assertions based on the outcomes of their activities and information from citations, as well as entity linking based on knowledge bases of other repositories such as Wikidata, etc.

Notes

1. https://scholar.google.com/
2. https://www.scopus.com/home.uri
3. https://www.semanticscholar.org/
4. https://www.jstor.org/
5. https://spacy.io/
6. https://huggingface.co/
7. https://orcid.org/
8. https://www.w3.org/RDF/
9. https://www.w3.org/community/bpmlod/wiki/NIF_Web_Services

Bibliography

Dessì et al. (2021): Generating knowledge graphs by employing Natural Language Processing and Machine Learning techniques within the scholarly domain. FGCS, 116 pp.253–264.

D'Souza et al. (2022): Computer Science Named Entity Recognition in the Open Research Knowledge Graph (https://orkg.org/). ArXiv abs/2203.14579.

Färber, Michael. (2019): The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data. *SEMWEB*.

Pertsas et al. (2017): Scholarly Ontology: modelling scholarly practices, IJDL, Vol. 18(3), pp.173–190.

Steenwinckel et al. (2020): Facilitating the Analysis of COVID-19 Literature Through a Knowledge Graph. ISWC. https://doi.org/10.1007/978-3-030-62466-8_22