Proceedings of the 2nd International Workshop on Semantic Technologies and Deep Learning Models for Scientific, Technical and Legal Data co-located with the Extended Semantic Web Conference 2024

Rima Dessi[1], Danilo Dessi[2], Francesco Osborne[3], and Hidir Aras[1]

[1] FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Germany
[2] GESIS Leibniz Institute for the Social Sciences, Cologne, Germany
[3] Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom
firstname.lastname@fiz-karlsruhe.de, danilo.dessi@gesis.org, francesco.osborne@open.ac.uk

1 Organizing Committee

- Rima Dessi, FIZ Karlsruhe, Germany.
- Danilo Dessi, GESIS Leibniz Institute for the Social Sciences, Cologne, Germany.
- Francesco Osborne, The Open University, Milton Keynes, United Kingdom.
- Hidir Aras, FIZ Karlsruhe, Germany.

2 Program Committee

- Rubén Alonso, Télécom Paris, R2M Solution Srl, Italy.
- Ahmad Alrifai, FIZ Karlsruhe, Germany.
- Davide Buscaldi, Sorbonne Paris North University, France.
- Pablo Calleja, Polytechnic University of Madrid, Spain.
- Mathieu DÁquin, LORIA, University of Lorraine, France.
- Rene Hackl-Sommer, DeepL, Germany.
- Inma Hernandez, University of Seville, Spain.
- Fabian Hoppe, VU Amsterdam
- Mirko Marras, University of Cagliari, Italy.
- Giacomo Medda, University of Cagliari, Italy.
- Angelo Antonio Salatino, The Open University, United Kingdom.
- Lise Stork, Vrije Universiteit Amsterdam, the Netherlands.
- Sabine Wehnert, Otto-von-Guericke-Universität Magdeburg, Germany.
- Lei Zhang, FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany.

3 Preface

The rapid growth of online available scientific, technical, and legal data such as patents, reports, articles, etc. has made the large-scale analysis and processing of such documents a crucial task. Today, scientists, patent experts, inventors, and other information professionals (e.g., information scientists, lawyers, etc.) contribute to this data every day by publishing articles, writing technical reports, or patent applications. It is a challenging task to process, analyze, and explore these documents due to their length, the use of domain-specific vocabulary, and the complexity introduced by targeting various scientific fields and domains. These semi-structured types of documents cover unstructured textual parts and structured parts such as tables, mathematical formulas, diagrams, and domain-specific information such as chemical names, bio-sequences, etc. Such kind of information brings complexity in processing such documents.

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

In order to benefit from the scientific-technical knowledge present in such documents, e.g., for decision-making or for professional search and analytics, there is an urgent need for analyzing, enriching, and linking such data by employing state-of-the-art Semantic Web technologies and AI methods. However, as they are heterogeneous and are written using domain-specific terminology applying the existing semantic technologies is not straightforward.

To address the challenges mentioned above, Semantic Web Technologies, Natural Language Processing (NLP) techniques, and Deep Neural Networks (DNN) must be leveraged in order to provide efficient and effective solutions for creating easily accessible and machine-understandable knowledge of science and industry.
To this end, the goal of the organized workshop[4] was to provide a meeting forum for people from academia as well as industry to come together and discuss topics such as the application of Semantic Web Technologies and Deep Learning Models to scientific, technical, and legal data. Further, the primary objective of the workshop was to promote collaboration among the participants and exchange ideas. The workshop started with a keynote entitled "Understanding Scientific and Societal Adoption of Scientific Knowledge and Resources Through NLP and Knowledge Graphs" by Prof. Dr. Stefan Dietze. An invited talk was also given on "Semantic Web and Machine Learning Systems for Intelligent Systems in Complex Domains" by Prof. Dr. Marta Sabou. These talks led to very insightful discussions within the community.

Overall, the workshop's success can be demonstrated by the high number of participants and submissions. Further, during the workshop, many participants joined the discussions, asked questions, and exchanged ideas about the application of Semantic Web Technologies and Machine Learning models on Scientific, Technical, and Legal Data. We believe this workshop helped participants build a new network and encourage future projects related to the mentioned topics. We definitely plan to organize the 3rd edition of this workshop.

May 2024 Rima Dessì, Danilo Dessì, Francesco Osborne, and Hidir Aras

Contents

Keynote Talk

Keynote by *Prof. Dr. Stefan Dietze*

Keynote on **Understanding Scientific and Societal Adoption of Scientific Knowledge and Resources Through NLP and Knowledge Graphs** .

Keynote Abstract:
Scientific discourse is scattered across unstructured scholarly publications and increasingly takes place online, e.g. in news or social media. Understanding the state-of-the-art in specific research fields, involved data, software, or methods, and their impact on both science and society requires substantial efforts and has become increasingly challenging. At the same time, societal debates about topics such as COVID or climate change have demonstrated the impact of science discourse on public opinion, policies, and society as a whole. This talk will provide an overview of a range of works that use deep learning-based NLP, such as PLMs and LLMs, to construct and use knowledge graphs about scientific discourse. These include, on the one hand, approaches that extract metadata about scholarly entities, such as code, data, tasks or machine learning models from scientific publications to enable machine-interpretable research information and understand dependencies between scholarly artefacts. On the other hand, we introduce NLP methods and knowledge graphs that enable an understanding

[4] https://semtech4stld.github.io/

of societal discourse about science, e.g. on Twitter/X, and facilitate interdisciplinary research into (mis-)representation and -information of scientific claims and findings in societal debates.

Invited Talk

Invited Talk by *Prof. Dr. Marta Sabou*

Invited Talk on **Semantic Web and Machine Learning Systems for Intelligent Systems in Complex Domains**.

Invited Talk Abstract: Creating intelligent applications that valorize complex domain data such as in the scientific, technical, and legal domain often calls for solutions that combine learning and symbolic artificial intelligence (AI) methods. In line with such developments, in the first part of this talk, we focus on describing a new sub-area of AI that focuses on combining Machine Learning components with techniques developed by the Semantic Web community—Semantic Web Machine Learning (SWeML). We report on the results of a systematic mapping study during which we analysed nearly 500 papers published in the past decade in this area, where we focused on evaluating architectural and application-specific features of such systems. In the second part of the talk, we describe the development and evaluation of a concrete SWeML system that aims to extract key elements from official Austrian permits, including the Issuing Authority, the Operator of the facility in question, the Reference Number, and the Issuing Date. We hope that our lessons learned both about this area as a whole (through the survey of SWeML systems) and the concrete system we built will provide inspiration for researchers and practitioners working with such complex data as in the legal domain and beyond.

Paper Session I

GerPS-NER: A Dataset for Named Entity Recognition to Support Public Service Process Creation in Germany .
Leila Feddoul, Sarah T. Bachinger, Clara Lachenmaier, Sebastian Apel, Pirmin Karg, Norman Klewer, Denys Forshayt, Robin Erd and Marianne Mauch

Automating Citation Placement with Natural Language Processing and Transformers.
Davide Buscaldi, Danilo Dessì, Enrico Motta, Marco Murgia, Francesco Osborne and Diego Reforgiato Recupero

Combining Knowledge Graphs and Large Language Models to Ease Knowledge Access in Software Architecture Research.
Angelika Kaplan, Jan Keim, Marco Schneider, Anne Koziolek and Ralf Reussner

Paper Session II

Extracting license information from web resources with a Large Language Model.
Enrico Daga, Jason Carvalho and Alba Catalina Morales Tirado

ChatGPT vs. Google Gemini: Assessing AI Frontiers for Patent Prior Art Search Using European Search Reports.
Renukswamy Chikkamath, Ankit Sharma, Christoph Hewel and Markus Endres

Bridging the Innovation Gap: Leveraging Patent Information for Scientists by Constructing a Patent-centric Knowledge Graph .
Hidir Aras, Rima Dessi, Farag Saad and Lei Zhang

Investigating Environmental, Social, and Governance (ESG) Discussions in News: A Knowledge Graph Analysis Empowered by AI .
Simone Angioni, Sergio Consoli, Danilo Dessì, Francesco Osborne, Diego Reforgiato Recupero and Angelo Salatino.

PRICER: Leveraging Few-Shot Learning with Fine-Tuned Large Language Models for Unstructured Economic Data.
Matt Murtagh, Declan O'Sullivan and Pj Wall.