Abu Ahmad, Raia et al.

Conference Paper — Published Version

NFDI4DS Shared Tasks

Suggested Citation: Abu Ahmad, Raia et al. (2023) : NFDI4DS Shared Tasks, In: Klein, Maike et al. (Ed.): INFORMATIK 2023, ISBN 978-3-88579-731-9, Gesellschaft für Informatik e.V., Bonn, pp. 931-935, https://nextcloud.gi.de/s/onnyxKSQoFHdqar

NFDI4DS Shared Tasks

Raia Abu Ahmad[1] Ekaterina Borisova[1] Georg Rehm[1] Stefan Dietze[2] Saurav Karmakar[2]
Wolfgang Otto[2] Jennifer D'Souza[3] Fidan Limani[4] Ricardo Usbeck[5]

Abstract: Shared tasks have proven to be successful in proposing innovative solutions for challenging research problems. The NFDI4DS consortium plans to host various shared tasks to tackle problems under the umbrella of scholarly information processing. We discuss three shared tasks in detail: Field of Research Classification, Software Mention Detection, and Tracking State-of-the-Art in Empirical AI. We also briefly mention other shared tasks planned to be released in the future.

Keywords: NFDI; NFDI4DS; Shared Tasks

1 Introduction

Shared tasks are scientific competitions in which teams attempt to find efficient solutions to a specific problem using shared data and evaluation measures [Pa17]. The goal is to objectively and directly compare different methods for tackling the same problem by using gold-standard data and common performance measures. Shared tasks are usually organised either at conferences or workshops such as the Conference on Natural Language Learning[6] or the International Workshop on Semantic Evaluation[7], or by companies (e. g., Kaggle[8]).

In recent years, shared tasks have been very successful in advancing state-of-the-art methods and standards to solve challenging problems [FU18]. In most shared tasks, the gold-standard dataset is made publicly available after the competition, thus providing valuable resources for the research community [Pa17]. Recent work has also introduced guidelines to ensure transparency and reproducibility of shared tasks in order to benefit scientific progress [Es21]. By utilising competitiveness among participants, shared tasks have proven to significantly encourage the development of novel and innovative solutions [Pa17].

[1] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Berlin, Germany
ekaterina.borisova@dfki.de, raia.abu_ahmad@dfki.de, georg.rehm@dfki.de
[2] GESIS Leibniz Institut für Sozialwissenschaften, Germany
stefan.dietze@gesis.org, saurav.karmakar@gesis.org, wolfgang.otto@gesis.org
[3] TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany, jennifer.dsouza@tib.eu
[4] ZBW – Leibniz-Informationszentrum Wirtschaft, Kiel, Germany, f.limani@zbw-online.eu
[5] Universiät Hamburg, Germany, ricardo.usbeck@uni-hamburg.de
[6] https://conll.org
[7] https://semeval.github.io
[8] https://www.kaggle.com

This paper presents three shared tasks that are part of the NFDI for Data Science and AI (NFDI4DS) project.[9] We also briefly introduce tasks currently under development. All shared tasks mentioned in Section 2, 3 and 4 are planned to start in 2024.

2 FoRC: Field of Research Classification for Scholarly Publications

The fast-growing pace of research has given rise to repositories and digital libraries that capture and manage scientific knowledge. For automated processes, the information of which scientific field a publication belongs to can assist downstream tasks. However, existing classification systems are limited either in terms of the used taxonomy, or in terms of the classification model itself.

We propose the Field of Research Classification (FoRC) shared task, which consists of two sub-tasks. **Subtask I** is a multi-class problem for general fields of research (FoR). This subtask uses an existing FoR taxonomy from the Open Research Knowledge Graph (ORKG) [Au20]. The dataset was constructed by fetching scientific publications from ORKG, arXiv, Crossref[10] API, and S2AG[11]. The subtask I dataset consists of 59,500 scholarly papers in English, classified into 123 FoR over five high-level fields and four hierarchical levels. **Subtask II** is a fine-grained multi-label classification problem that deals with multiple sub-fields within one specific FoR, i. e., computational linguistics (CL). This subtask will use a hierarchical taxonomy of ca. 170 CL sub-fields structured into three hierarchical levels that was constructed using a topic modeling approach. The dataset is currently being developed by a team of six annotators and will contain 1,500 manually labeled articles from the ACL Anthology.

3 SOMD: Software Mention Detection in Scholarly Publications

Across all disciplines science has become increasingly data-driven, leading to additional needs regarding software for collecting, processing and analysing data. Transparency about software used as part of the scientific process is crucial to ensure reproducibility and to understand provenance of research data and insights. Furthermore, understanding software usage, citation habits, and their evolution over time within and across disciplines can shape the understanding of the evolution of disciplines, the influence of software on scientific impact, and the emerging needs for computational support.

Given the scale and heterogeneity of software citations, robust methods are required, able to detect and disambiguate mentions of software and related metadata. However, despite the existence of software citation principles [SKN16], software mentions in scientific articles

[9] https://www.nfdi4datascience.de
[10] https://www.crossref.org
[11] https://www.semanticscholar.org/product/api

are usually informal and incomplete. Initial pipelines for the automated extraction through deep learning-based NLP pipelines [Is22] usually rely on fine-tuned or otherwise supervised models and sufficient ground truth data [Sc21].

The SOMD shared task will build on a corpus and annotation framework and existing baselines (SoMeSci [Sc21]). In particular, we consider three different subtasks. **Subtask I** will deal with recognising Software from individual sentences. At the same time software mentions shall be classified according to its mention type, e.g., mention, usage, or creation and the software type, e.g., application, programming environment, and plugin or package. **Subtask II** will deal with extracting additional information for each software according to the SoMeSci schema. And **Subtask III** will deal with classifying relations to other recognised entities. This includes versions and developers, but also URLs or Host applications for plugins. Evaluation will be based on exact matches rather than partial matches.

4 SOTA? Tracking the State-of-the-Art in AI Scholarly Publications

The growing number of publications poses a major challenge [Ji10]. How can we stay updated with the fast-paced research progress? The current format of scholarly communication, with results buried in unstructured PDFs, hinders comprehension and limits machine usability. One potential solution is to represent research results in structured and semantic formats within knowledge graphs (KG) of scientific knowledge [Sh09] such as the ORKG[12]. When it comes to AI research, the leaderboards construct of information organisation provides an overview of the state-of-the-art (SOTA) by aggregating results from multiple studies addressing the same research challenge. A leaderboard typically comprises a task (T), a dataset (D), evaluation metrics (M), and scores obtained by the model (S).

The SOTA shared task would facilitate reaping diverse machine learning observations on a relatively non-trivial task as the automated mining of leaderboards for empirical AI research. **Subtask I** will deal with TDM extraction, **Subtask II** with TDMS extraction, and **Subtask III** with extracting URLs of source-code from publications. The dataset for each subtask is extracted from the community-annotated Leaderboards on PapersWithCode (PwC).

5 Future Work

Four additional shared tasks are currently under development:

1. Question Answering (QA): There is currently no natural language interface that allows users to access scholarly data, e. g., papers, authors, models, or datasets. We will introduce a shared task at ISWC 2023[13] with two subtasks. SciQA [Au23] focuses

[12] https://orkg.org
[13] https://kgqa.github.io/scholarly-QALD-challenge/2023/

on QA over scholarly data using the ORKG. DBLP-QUAD involves KG question answering over the DBLP Knowledge Graph using DBLP-QUAD [Ba23].

2. Machine Learning Model Detection: Machine learning model (MLModel) mentions in scholarly articles are currently unable to be detected. We are preparing a manually annotated gold standard dataset as part of the GESIS Scholarly Annotation Project (GSAP), comprised of annotating MLModel family, dataset model family, and their relationships. The GSAP data will be used in two subtasks: HowMLMod will deal with detecting how an MLModel is mentioned in an article. AncesterMLMod will figure out which MLModel architecture is used in specific MLModel mentions.

3. Dataset Mention Detection: In Machine Learning research, ensuring the traceability of datasets is vital due to issues like biases, data leakage, train-test contamination, and ethical concerns. Reuse and overlap of datasets further complicate matters. We will propose two shared tasks. The first focuses on detecting dataset mentions that are aligned with mentions of ML models. The second task emphasizes traceability by detecting relationships between datasets. It aims to uncover how datasets are reused or transformed, unraveling the dataset's history.

4. Standards, FAIR principles, and FAIR Digital Objects (FDO): The project targets different types of research artefacts, including the ones generated from the shared tasks, meant to benefit the NFDI4DS research community. Our goal is to evaluate the adoption of the FDO model as means of elevating these artefacts into "actionable knowledge units" [DSKW20], able to support existing or enable new use cases. FAIR principles [Wi16] are a key component for this model, and this "dependency" implies that evaluating their adoption for the NFDI4DS resources remains an important aspect for this shared task. Finally, there are tasks across the project, such as the adoption of metadata standards, Knowledge Organization Systems, machine-readable representation, to name just a few, that provide a valuable input to and will be considered in their supporting role to addressing the challenges of FDO adoption.

6 Conclusion

This paper presents three of NFDI4DS's shared tasks, as well as four additional ones currently under development. The shared tasks tackle challenging problems in the field of scholarly information processing. All tasks plan to release publicly available datasets and follow guidelines of transparency and reproducibility.

Acknowledgements

This work has received funding through the German Research Foundation (DFG) project NFDI4DS (no. 460234259).

Bibliography

[Au20] Auer, Sören; Oelen, Allard; Haris, Muhammad; Stocker, Markus; D'Souza, Jennifer; Farfar, Kheir Eddine; Vogt, Lars; Prinz, Manuel; Wiens, Vitalis; Jaradeh, Mohamad Yaser: Improving Access to Scientific Literature with Knowledge Graphs. Bibliothek Forschung und Praxis, 44(3):516–529, 2020.

[Au23] Auer, Sören; Barone, Dante A.C.; Bartz, Cassiano; Cortes, Eduardo G.; Yaser, Mohamad Jaradeh; Karras, Oliver; Koubarakis, Manolis; Mouromtsev, Dmitry; Pliukhin, Dmitrii; Radyush, Daniil; Shilin, Ivan; Stocker, Markus; Tsalapati, Eleni: SciQA Benchmark: Dataset and RDF Dump. 2023.

[Ba23] Banerjee, Debayan; Awale, Sushil; Usbeck, Ricardo; Biemann, Chris: DBLP-QuAD: A Question Answering Dataset over the DBLP Scholarly Knowledge Graph. CoRR, abs/2303.13351, 2023.

[DSKW20] De Smedt, Koenraad; Koureas, Dimitris; Wittenburg, Peter: FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications, 8(2):21, 2020.

[Es21] Escartín, Carla Parra; Lynn, Teresa; Moorkens, Joss; Dunne, Jane: Towards Transparency in NLP Shared Tasks. arXiv preprint arXiv:2105.05020, 2021.

[FU18] Filannino, Michele; Uzuner, Özlem: Advancing the State of the Art in Clinical Natural Language Processing through Shared Tasks. Yearbook of medical informatics, 27(01):184–192, 2018.

[Is22] Istrate, Ana-Maria; Li, Donghui; Taraborelli, Dario; Torkar, Michaela; Veytsman, Boris; Williams, Ivana: A Large Dataset of Software Mentions in the Biomedical Literature. 2022.

[Ji10] Jinha, Arif E.: Article 50 Million: An Estimate of the Number of Scholarly Articles in Existence. Learned Publishing, 23(3):258–263, 2010.

[Pa17] Parra Escartín, Carla; Reijers, Wessel; Lynn, Teresa; Moorkens, Joss; Way, Andy; Liu, Chao-Hong: Ethical Considerations in NLP Shared Tasks. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. Association for Computational Linguistics, Valencia, Spain, pp. 66–73, April 2017.

[Sc21] Schindler, David; Bensmann, Felix; Dietze, Stefan; Krüger, Frank: SoMeSci—A 5 Star Open Data Gold Standard Knowledge Graph of Software Mentions in Scientific Articles. In: Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21). Association for Computing Machinery, Virtual Event, QLD, Australia, November 2021.

[Sh09] Shotton, David: Semantic Publishing: The Coming Revolution in Scientific Journal Publishing. Learned Publishing, 22(2):85–94, 2009.

[SKN16] Smith, Arfon M; Katz, Daniel S; Niemeyer, Kyle E: Software Citation Principles. PeerJ Computer Science, 2:e86, 2016.

[Wi16] Wilkinson, Mark D; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E et al.: The FAIR Guiding Principles for Scientific Data Management and Stewardship. Scientific data, 3(1):1–9, 2016.