Extraction and Evaluation of Knowledge Entities from Scientific Documents: EEKE2020

Chengzhi Zhang
Department of Information Management, Nanjing University of Science and Technology, China
zhangcz@njust.edu.cn

Philipp Mayr
GESIS – Leibniz Institute for the Social Sciences, Germany
Philipp.Mayr@gesis.org

Wei Lu
School of Information Management, Wuhan University, China
weilu@whu.edu.cn

Yi Zhang
Centre for Artificial Intelligence, University of Technology Sydney, Australia
Yi.Zhang@uts.edu.au

ABSTRACT

The goal of this workshop is to engage the related communities in open problems in the extraction and evaluation of knowledge entities from scientific documents. This workshop entitles this cutting-edge and cross-disciplinary direction *Extraction and Evaluation of Knowledge Entity* (EEKE), highlighting the development of intelligent methods for identifying knowledge claims in scientific documents, and promoting the application of knowledge entities. The website of this workshop is at: https://eeke2020.github.io/

ACM Reference Format:
Chengzhi Zhang, Philipp Mayr, Wei Lu, and Yi Zhang. 2020. Extraction and Evaluation of Knowledge Entities from Scientific Documents: EEKE2020. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20), August 1–5, 2020, Virtual Event, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3383583.3398504

1 INTRODUCTION

In the era of big data, massive amounts of information and data have dramatically changed human civilization. The broad availability of information provides more opportunities for people, but there has appeared a new challenge: how can we obtain useful knowledge from numerous information sources?

A *knowledge entity* is a relatively independent and integral knowledge module in a special discipline or a research domain [6]. As a crucial medium for knowledge transmission, scientific documents that contain a large number of knowledge entities attract the attention of scholars [8]. In scientific documents, knowledge entities refer to the knowledge mentioned or cited by authors, such as algorithms, models, theories, datasets and software, which reflect the various resources used by the authors in solving problems [9] [5]. Extracting knowledge entities from scientific documents in an accurate and comprehensive way becomes a significant topic. We may recommend documents related to a given knowledge entity (like the LSTM model) for scholars, especially for beginners in a research field. As an example, DARPA has recently launched the ASKE (Automating Scientific Knowledge Extraction) project[1], which aims to develop next-generation applications of artificial intelligence.

Therefore, the goal of this all-virtual workshop is to engage the related communities in open problems in the extraction and evaluation of knowledge entities from scientific documents. At present, scholars have used knowledge entities to construct general knowledge graphs [2] (e.g. Google Knowledge Graph[2]) and domain knowledge graphs (e.g. GeoNames[3]). Data sources for these approaches include text (news, policy files, emails, etc.) and multimedia (videos, images, etc.) data. Open Academic Graph (OAG)[4] is a large academic knowledge graph and includes about 700 million entities and 2 billion relationships. However, entities in OAG are mainly metadata of academic papers, not including *knowledge entities*.

Compared to existing research and workshops like Joint workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) [12], Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics (CLBib) [1], or the shared task ScienceIE [3] (a shared task at 2017 SemEval which deals with the extraction of keyphrases and relations from scientific publications[5]), this EEKE workshop aims to extract knowledge entities from scientific documents, and explore the features of entities to conduct practical applications. The results of this workshop are expected to provide scholars, especially early career researchers, with knowledge recommendations and other knowledge entity-based services.

2 OBJECTIVES AND TOPICS FOR EEKE2020

This workshop will be relevant to scholars in computer and information science, specialized in Information Extraction, Text Mining, NLP, IR and Digital Libraries. It will also be of importance for all stakeholders in the publication pipeline: implementers, publishers and policymakers.

EEKE2020 entitles this cutting-edge and cross-disciplinary direction *Extraction and Evaluation of Knowledge Entity*, highlighting the development of intelligent methods for identifying knowledge

JCDL '20, August 1–5, 2020, Virtual Event, China

ACM ISBN 978-1-4503-7585-6/20/08.
https://doi.org/10.1145/3383583.3398504

[1] https://www.darpa.mil/program/automating-scientific-knowledge-extraction
[2] https://developers.google.com/knowledge-graph
[3] http://www.geonames.org/
[4] https://www.aminer.cn/open-academic-graph
[5] https://scienceie.github.io/

claims in scientific documents, and promoting the application of knowledge entities. We invite stimulating research on topics including, but not limited to, methods of knowledge entity extraction and applications of knowledge entity. Specific examples of fields of interest include:

- Task and methodology from scientific documents [10]
- Model and algorithm extraction from scientific documents [14]
- Dataset and evaluation metrics extraction from documents [9]
- Software and tool extraction from scientific documents [4]
- Construction of a knowledge entity graph and roadmap [15]
- Knowledge entity summarization
- Relation extraction of knowledge entity [11]
- Construction of a knowledge base of knowledge entities
- Modeling function of knowledge entity citation [13] [16]
- Bibliometrics of knowledge entity [7]
- Application of knowledge entity extraction.

The workshop will be an **all-virtual event** and last one full day and specific activities include keynotes, paper presentations and a poster session. The tentative number of attendees is approx. 50. Cfp with all submission details is available via the EEKE website: https://eeke2020.github.io/, and notifications via emails, ACM SIGIR Mailing List and social media.

Two types of papers were presented: long papers and short papers. At the same time, we welcome submissions detailing original, early findings, works in progress and industrial applications of knowledge entities extraction and evaluation for a special poster session, possibly with a 2-minute presentation in the main session. Some research track papers will also be invited to the poster track instead. All submissions will be reviewed by at least two independent reviewers. Workshop proceedings will be deposited online in the CEUR workshop proceedings publication service. This way the proceedings will be permanently available and citable (digital persistent identifiers and long term preservation).

3 RELATED WORKSHOPS

There are some related workshops as follows.

- 4th Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019) at SIGIR 2019. The scope of the BIRNDL workshops (2016–2019)[6] were on full-text analysis, including multilingual analysis, IR methods for DL, and applications of citation-based NLP like summarization.
- 2nd Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics (CLBib 2017) brought together researchers to study the ways the ways bibliometrics can benefit from large-scale text analytics and sense mining of scientific papers, thus exploring the interdisciplinarity of Bibliometrics and NLP.
- 1st Workshop on Scholarly Document Processing (SDP 2020)[7] at EMNLP 2020. The scope of SDP is on natural language processing, information retrieval, and data mining problems in scientific documents and is a continuation of the BIRNDL and WOSP workshops.

[6]https://philippmayr.github.io/BIRNDL-WS/
[7]https://ornlcda.github.io/SDProc/

4 TENTATIVE SCHEDULE OF EVENTS

Along with research paper presentation (long and short), the workshop will host one or two keynotes and a poster session. The workshop will end with planning and discussions to decide on future directions and improvements to the workshop.

The all-virtual workshop EEKE2020 aims to be inclusive and diverse, in terms of both constituency and research. To this end, the open CFP will explicitly encourage female first authors and underrepresented groups.

REFERENCES

[1] Iana Atanassova, Marc Bertin, and Philipp Mayr. 2019. Editorial: Mining Scientific Papers: NLP-enhanced Bibliometrics. *Frontiers in Research Metrics and Analytics* (2019). https://doi.org/10.3389/frma.2019.00002

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 722–735.

[3] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 Task 10: ScienceIE-Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 546–555.

[4] Katarina Boland and Frank Krüger. 2019. Distant supervision for silver label generation of software mentions in social scientific publications. In *Proc. Joint Workshop Bibliometric-Enhanced Inf. Retrieval Nat. Lang. Process. Digit. Libraries*. 15–27.

[5] Arthur Brack, Jennifer D'Souza, Anett Hoppe, Sören Auer, and Ralph Ewerth. 2020. Domain-independent Extraction of Scientific Concepts from Research Articles. *arXiv preprint arXiv:2001.03067* (2020).

[6] Xiao Chang and Qinghua Zheng. 2007. Knowledge element extraction for knowledge-based learning resources organization. In *International Conference on Web-Based Learning*. Springer, 102–113.

[7] Ruiyi Ding, Yuzhuo Wang, and Chengzhi Zhang. 2019. Investigating Citation of Algorithm in Full-text of Academic Articles in NLP domain: A Preliminary Study. In *Proceedings of the 17th International Conference on Scientometrics and Informetrics*. 2726–2727.

[8] Ying Ding, Min Song, Jia Han, Qi Yu, Erjia Yan, Lili Lin, and Tamy Chambers. 2013. Entitymetrics: Measuring the impact of entities. *PloS one* 8, 8 (2013).

[9] Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5203–5213.

[10] Aleksandar Kovačević, Zora Konjović, Branko Milosavljević, and Goran Nenadic. 2012. Mining methodologies from NLP publications: A case study in automatic terminology recognition. *Computer Speech & Language* 26, 2 (2012), 105–126.

[11] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3219–3232.

[12] Philipp Mayr, Ingo Frommholz, Guillaume Cabanac, Muthu Kumar Chandrasekaran, Kokil Jaidka, Min-Yen Kan, and Dietmar Wolfram. 2018. Introduction to the Special Issue on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). *International Journal on Digital Libraries* (2018). https://doi.org/10.1007/s00799-017-0230-x

[13] Suppawong Tuarob, Sung Kang, Poom Wettayakorn, Chanatip Pornprasit, Tanakitti Sachati, Saeed-Ul Hassan, and Peter Haddawy. 2020. Automatic Classification of Algorithm Citation Functions in Scientific Literature. *IEEE Transactions on Knowledge and Data Engineering* (2020). https://doi.org/10.1109/TKDE.2019.2913376

[14] Yuzhuo Wang and Chengzhi Zhang. 2019. Finding More Methodological Entities from Academic Articles via Iterative Strategy: A Preliminary Study. In *Proceedings of the 17th International Conference on Scientometrics and Informetrics*. 2702–2703.

[15] Hanwen Zha, Wenhu Chen, Keqian Li, and Xifeng Yan. 2019. Mining Algorithm Roadmap in Scientific Publications. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1083–1092.

[16] He Zhao, Zhunchen Luo, Chong Feng, Anqing Zheng, and Xiaopeng Liu. 2019. A Context-based Framework for Modeling the Role and Function of On-line Resource Citations in Scientific Literature. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5209–5218.