Preface to the 3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents at JCDL 2022

Chengzhi Zhang[1], Philipp Mayr[2], Wei Lu[3], Yi Zhang[4]

[1] *Nanjing University of Science and Technology, No. 200, Xiaolinvgwei, Nanjing, 210094, China*
[2] *GESIS - Leibniz-Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne, 50667, Germany*
[3] *Wuhan University, Luojiashan, Wuhan, 430072, China,*
[4] *Australian Artificial Intelligence Institute, University of Technology Sydney, 15 Broadway, Ultimo, NSW, Australia*

Abstract

The 3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE 2022) was held online at the ACM/IEEE Joint Conference on Digital Libraries (JCDL) 2022. The goal of this workshop series (https://eeke-workshop.github.io/) is to engage the related communities in open problems in the extraction and evaluation of knowledge entities from scientific documents. Topics of this proceedings include extraction method of knowledge entity, application of knowledge entity extraction, knowledge entity and bibliometrics.

Keywords

Knowledge entity, entity extraction, entity evaluation, scientific document

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 a new challenge is rising: how can we obtain useful knowledge from numerous information sources [1]? Knowledge entities in scientific documents may include method entities [2], tasks, dataset and metrics [3], software and tools [4], etc. Knowledge entity application includes the construction of a knowledge entity graph and roadmap [5], modeling functions of knowledge entity citations [6], etc.

The 3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE 2022) was affiliated with the ACM/IEEE Joint Conference on Digital Libraries (JCDL) on June 23~24, 2022. This workshop engaged related communities in open problems in the extraction and evaluation of knowledge entities from scientific documents. Participants presented studies on identifying knowledge entities, exploring features of various entities, analyzing the relationship between entities, and constructing extraction platforms or knowledge bases. This workshop provided scholars, especially early career researchers, with knowledge recommendations and other knowledge entity-based services [7, 8].

2. Overview of the papers

There were 24 papers submitted for peer-review to this workshop. Out of these, 17 papers were accepted for this volume, 6 as regular papers, 7 as short papers and 4 as poster. In addition, the workshop featured two keynote talks touching on different fields of EEKE studies and applications. All workshop contributions have been documented on the workshop website (https://eekeworkshop.github.io/2022/).The following section briefly lists the 2 keynotes and the 17 accepted submissions.

3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2022), June 24-25, 2022, Cologne, Germany and Online
EMAIL: zhangcz@njust.edu.cn (Chengzhi Zhang); philipp.mayr@gesis.org(Philipp Mayr); weilu@whu.edu.cn(Wei Lu); yi.zhang@uts.edu.au(Yi Zhang)

2.1 Keynotes

Two keynotes were presented at EEKE2022.

Professor Alan Porter (George Institute of Technology, USA) and Mr. Nils Newman (Search Technology Inc., USA) delivered the first talk on 'What knowledge to extract from "Tech Mining"'.

Alan and Nils introduced "Tech Mining" -- text analyses of R&D Information to gain useful intelligence on advancing sciences and technologies. This specialty applies some of the same tools as

EEKE, as they illustrated by comparing their topical emphases. They presented the generation of tech emergence scores as an illustration of an advanced Tech Mining analytical capability. Some lessons learned in the development and applications of Tech Mining may suggest EEKE possibilities that we hope to discuss.

Professor Daqing He (University of Pittsburgh, USA) delivered the second talk on Keyphrases as Knowledge Units for Text-based Applications.

Natural language text is the main form of communication in various domains such as scholarly communication, student instructions, and healthcare. Keyphrases in the form of noun phrases are often identified and extracted as the knowledge unit for representing the content of natural language text, and they take various roles in contributing downstream tasks. However, some important relevant issues still exist and have not been solved appropriately, such as the characteristics of keyphrases, their roles in knowledge exchange, and their usages in different domains. In this talk, Daqing presented several research projects he and his team conducted on exploring keyphrases as knowledge units, and their applications to different domains. His talk covered keyphrase generation from academic papers using deep learning methods, keyphrase representation as the knowledge units in textbooks for supporting students' learning, and keyphrase identification in the form of chief complaint recognition from clinical reports for representing patients' symptoms and diseases. This talk highlighted the importance of keyphrases in natural language text and illustrated appropriate technologies for fulfilling keyphrase's knowledge unit roles in various application domains.

2.2 Research papers and posters

The 17 papers were presented in 4 sessions.

2.2.1 Session 1: Extraction method of knowledge entity

This session highlights extraction methods of knowledge entities including framework developments and the incorporations of some machine learning methods.

Yongqiang Ma, Jiawei Liu, Wei Lu and Qikai Cheng proposed a metrics-driven mechanism and a knowledge extraction pipelinebased on a pre-trained model.

Liangping Ding, Zhixiong Zhang and Huan Liu proposed a bootstrapped model for recognizing Chinese biomedical name entities, in which Lexicons are facilitated.

Yujie Zhang, Rujiang Bai and Ling Kong introduced a novel framework using causality to extract scientific literatures.

Hao Wang, Xian-Ling Mao and Heyan Huang developed a semisupervised transfer learning framework for extracting low resource entities and relations in scientific domains.

2.2.2 Session 2: Application of knowledge entity extraction

This session demonstrates some applications of knowledge entity extraction.

Nina Smirnova and Philipp Mayr evaluated embedding models for automatic extraction and classification of acknowledged entities in scientific documents.

Bikun Chen, Kuan Bai and Yuxin Liu investigated the topic distribution of China's data governance policies by using a full-text highlighted clue word approach.

Chaoyu Gao, Tianxing Wu, Shengqi Jing and Yuxiang Wang presented a study on medical schema matching using knowledge graph embedding techniques.

Chuhan Wang, Tongyang Zhang, Yi Bu and Jian Xu exploited the research diversity of scholars based on the multi-dimensional calculation of entities.

Xiang Shi, Zikun Feng, Jiawei Liu, Qikai Cheng and Wei Lu proposed a framework for automatically constructing a Technology Function Matrix (TFM) that requires only a small amount of labeled data.

2.2.3 Session 3: Knowledge entity and bibliometrics

This session collects contributions on bibliometric studies 、 facilitating knowledge entities and knowledge graph techniques.

Dongin Nam, Jiwon Kim, Jeeyoung Yoon, Chaemin Song, Seongdeok Kim and Min Song proposed a solution for characterizing knowledge entities in citation sentences.

Mengjia Wu, Yi Zhang, Mark Markley, Caitlin Cassidy, Nils Newman and Alan Porter proposed a research framework to assist scientists in identifying, retrieving, and visualizing the emerging Covid-19 knowledge.

Lu Huang, Xiaoli Cao, Hang Ren and Tianbin Xing created a solution of detecting technological recombination by using semantic analysis and dynamic network analysis.

Tingting Ma, Ruiping Cheng, Hongshu Chen and Xiao Zhou developed a hybrid approach to identify and forecast technological opportunities based on topic modeling and sentiment analysis.

2.2.4 Session 4: Poster

This session mainly demonstrates research approaches and examples of knowledge extraction and knowledge graph building.

Xin An, Mengmeng Zhang and Shuo Xu developed a corpus for entity recognition on the COVID-19 full-text literature.

Zhongyi Wang, Jing Chen, Jiangping Chen and Haihua Chen focused on the extraction of interdisciplinary topics and their evolution.

Yuchen Yan and Chong Chen proposed a solution to annotate the function and topics of scientific papers, and constructed a knowledge graph, named SciGraph.

Wenjiao Zheng and Bolin Hua applied a machine learning approach to extract entities for understanding the dynamics of science and technology policies.

3. Outlook and further reading

The EEKE workshops have achieved great success and received significant attentions from related research communities. The outcomes of this workshop series contributed novel technological developments and empirical practices and insights to the literature. The EEKE2022 organization committee is editing a Special Issues in Scientometrics. For more information, please see https://eeke-workshop.github.io/2022/si-eeke.html.

4. References

[1] Zhang C., Mayr P., Lu W., & Zhang Y. (2022). JCDL2022 Workshop: Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2022). In: Proceedings of the 22th ACM/IEEE Joint Conference on Digital Libraries (JCDL2022), Cologne, Germany. https://doi.org/10.1145/3529372.3530917

[2] Wang, Y., Zhang, C., & Li, K. (2022). A review on method entities in the academic literature: extraction, evaluation, and application. Scientometrics, 127(5): 2479–-2520. https://doi.org/10.1007/s11192-022-04332-7

[3] Hou, Y., Jochim, C., Gleize, M., Bonin, F., & Ganguly, D. (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. http://doi.org/10.18653/v1/P19-1513

[4] Boland K., & Krüger F. 2019. Distant supervision for silver label generation of software mentions in social scientific publications. In Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries. 15--27. http://ceur-ws.org/Vol-2414/paper3.pdf

[5] Zha H., Chen W., Li K., & Yan X. 2019. Mining algorithm roadmap in scientific publications. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 1083--1092. http://doi.org/10.1145/3292500.3330913

[6] Tuarob, S., Kang, S. W., Wettayakorn, P., Pornprasit, C., Sachati, T., Hassan, S. U., & Haddawy, P. (2019). Automatic classification of algorithm citation functions in scientific literature. IEEE Transactions on Knowledge and Data Engineering, 32(10), 1881--1896. https://doi.org/10.1109/TKDE.2019.2913376

[7] Zhang C., Mayr P., Lu W., & Zhang Y. (2020). Extraction and Evaluation of Knowledge Entities from Scientific Documents: EEKE2020. In: Proceedings of the 20th ACM/IEEE Joint Conference on Digital Libraries (JCDL2020), Wuhan,

China. https://doi.org/10.1145/3383583.3398504

[8] Zhang C., Mayr P., Lu W., & Zhang Y. (2021). Preface to the 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents at JCDL 2021. In: Proceedings of the 1st Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents co-located with the ACM/IEEE Joint Conference on Digital Libraries in 2021 (JCDL 2021), Virtual Event. https://dblp.org/rec/conf/jcdl/Zhang0LZ21.html