DE GRUYTER
OPEN

Guest Editorial

Extraction and Evaluation of Knowledge Entities from Scientific Documents

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

[1]Nanjing University of Science and Technology, China
[2]GESIS - Leibniz-Institute for the Social Sciences, Germany
[3]Wuhan University, China
[4]University of Technology Sydney (UTS), Australia

Citation: Zhang, C.Z., Mayr, P., Lu, W., & Zhang, Y. (2021). Extraction and evaluation of knowledge entities from scientific documents. *Journal of Data and Information Science*, 6(3), 1–5. https://doi.org/10.2478/jdis-2021-0025

As a core resource of scientific knowledge, academic documents have been frequently used by scholars, especially newcomers to a given field. In the era of big data, scientific documents such as academic articles, patents, technical reports, and webpages are booming. The rapid daily growth of scientific documents indicates that a large amount of knowledge is proposed, improved, and used (Zhang et al., 2021). In scientific documents, knowledge entities (KEs) refer to the knowledge mentioned or cited by authors, such as algorithms, models, theories, datasets and software, diseases, drugs, and genes, reflecting rich resources in diverse problem-solving scenarios (Brack et al., 2020; Ding et al., 2013; Hou et al., 2019; Li et al. 2020). The advancement, improvement, and application of KEs in academic research have played a crucial role in promoting the development of different disciplines. Extracting various KEs from scientific documents can determine whether such KEs are emerging or typical in a specific field, and help scholars gain a comprehensive understanding of these KEs and even the entire research field (Wang & Zhang, 2020). KE extraction is also useful for multiple downstream tasks in information extraction, text mining, natural language processing, information retrieval, digital library research, and so on (Zhang et al., 2021). Particularly for researchers in artificial intelligence (AI), information science, and other related disciplines, discovering methods from large-scale academic literature, and evaluating their performance and influence have become increasingly necessary and meaningful (Hou et al., 2020).

There are four kinds of methods of KE extraction in scientific documents. They are manual annotation-based (Chu & Ke, 2017; Tateisi et al., 2014; Zadeh & Schumann, 2016), rule-based (Kondo et al., 2009), statistics-based (Heffernan & Teufel, 2018; Névéol, Wilbur, & Lu, 2011; Okamoto, Shan, & Orihara, 2017), and

[†] Corresponding author: Chengzhi Zhang (E-mail: zhangcz@njust.edu.cn).

the state-of-the-art one—deep learning-based (Paul et al., 2019; Yang et al., 2018), respectively.

Currently, KEs are evaluated via frequency or text content (Wang & Zhang, 2020). Some scholars analyzed KEs' influence using bibliometric indicators, e.g. the frequency of mentions, citations, and the usage in full text (Belter, 2014). Additionally, some studies also utilized text content to deeply explore the role, function, and relationship of KEs (Li & Yan, 2018; Li, Yan, & Feng, 2017; Wang & Zhang, 2020). Identifying the pattern of citations and the use of KEs through the content of academic papers is also on the trail (Yoon et al., 2019).

In recent years, the topic *Extraction and Evaluation of Knowledge Entities from Scientific Documents* has attracted the attention from the community. There are some conferences and workshops in line with this topic, such as the Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE) (Zhang et al., 2020), the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) (Cabanac et al., 2016), the Workshop on Mining Scientific Publications (WOSP, https://wosp.core.ac.uk/), the Workshop on AI + Informetrics (AII) (Zhang, et al., 2021), the Workshop on Scholarly Document Processing (SDP) (Chandrasekaran et al., 2020) and the Workshop on Natural Language Processing for Scientific Text (SciNLP, https://scinlp.org).

We are very grateful that there are seven contributions submitted to the special issue of *Journal of Data and Information Science* (*JDIS*) and five submissions are accepted after several rounds of peer-review and revisions.

The paper "Sentence, Phrase, and Triple Annotations to Build a Knowledge Graph of Natural Language Processing Contributions—A Trial Dataset" (D'Souza & Auer, 2021) normalized the NLPCONTRIBUTIONS scheme to a designed structure, which was directly extracted from natural language processing (NLP) articles. They demonstrated that the NLPCONTRIBUTIONGRAPH data integrated into the Open Research Knowledge Graph (ORKG), a next-generation KG-based digital library with intelligent computations, enabled over-structured scholarly knowledge to assist researchers in their daily academic tasks.

The paper "Automatic Keyphrase Extraction from Scientific Chinese Medical Abstracts Based on Character-Level Sequence Labeling" (Ding et al., 2021) proposed an automatic model of key-phrase extraction for Chinese medical abstracts, which combined sequence labeling formulation and pre-trained language model. Experiments compared word-level and character-level sequence labeling approaches on supervised machine learning models and BERT-based models. The experimental results show that the proposed character-level sequence labeling model based on BERT obtains F_1-score of 59.80%, getting 9.64% absolute improvement.

The paper “Content Characteristics of Knowledge Integration in the eHealth Field: An Analysis Based on Citation Contexts” (Wang et al., 2021) explored the content characteristics of knowledge integration in an interdisciplinary field—eHealth. Associated knowledge phrases (AKPs) shared between citing papers and their references were extracted from the citation contexts of eHealth papers by applying a stem-matching method. A classification schema that considers the functions of knowledge in the given domain was proposed to categorize the identified AKPs. The annotated AKPs reveal that different knowledge types have remarkably different integration patterns in terms of knowledge amount, the breadth of source disciplines, and the integration time lag.

The paper “A New Citation Recommendation Strategy Based on Term Functions in Related Studies Section” (Chen, 2021) proposed a term function-based citation recommendation framework to recommend articles for users. The author presented nine term functions, and among them, three were newly created and six were identified from the literature. The experiments show that the term function-based methods outperform the baselines, demonstrating its performance in identifying valuable citations.

The last paper, “Embedding-based Detection and Extraction of Research Topics from Academic Documents Using Deep Clustering” (Vahidnia, Abbasi, & Abbass, 2021) proposed a modified deep clustering method to detect research trends from the abstracts and titles of academic documents. The experimental results show that the modified DEC in conjunction with Doc2Vec can outperform other methods in the clustering task. Using the proposed method, the authors also show how the topics have evolved in the period of the recent 30 years, taking advantage of a keyword extraction method for cluster tagging and labeling, demonstrating the context of the topics.

Acknowledgments

We acknowledge support from the EEKE community, especially our authors and reviewers. We could never have finished such an excellent special issue without their devoted contribution and generous help.

References

Brack, A., D’Souza, J., Hoppe, A., Auer, S., Ewerth, R. (2020). Domain-Independent Extraction of Scientific Concepts from Research Articles. In: Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol 12035. Springer, Cham. https://doi.org/10.1007/978-3-030-45439-5_17

Belter, C.W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. PloS One, 9(3), Article e92590. https://doi.org/10.1371/journal.pone.0092590

Cabanac, G., Chandrasekaran, M., Frommholz, I., Jaidka, K., Kan, M., Mayr, P., & Wolfram, D. (2016). Report on the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016). SIGIR Forum, 50(2), 36–43.

Chandrasekaran, M.K., de Waard, A., Feigenblat, G., Freitag, D., Ghosal, T., Hovy, E., & Shmueli-Scheuer, M. (2020, November). Proceedings of the first workshop on scholarly document processing. Retrieved from https://www.aclweb.org/anthology/volumes/2020.sdp-1/

Chen H. (2021). A New Citation Recommendation Strategy Based on Term Functions in Related Studies Section. Journal of Data and Information Science, 6(3), 75–98. https://doi.org/10.2478/jdis-2021-0022

Chu, H., & Ke, Q. (2017). Research methods: What’s in the name? Library & Information Science Research, 39(4), 284–294. https://doi.org/10.1016/J.LISR.2017.11.001

D’Souza, J., & Auer, S. (2021). Sentence, Phrase, and Triple Annotations to Build a Knowledge Graph of Natural Language Processing Contributions—A Trial Dataset. Journal of Data and Information Science, 6(3), 6–34. https://doi.org/10.2478/jdis-2021-0023

Ding, L., Zhang, Z., Liu, H., Li, J., & Yu, G. (2021). Automatic Keyphrase Extraction from Scientific Chinese Medical Abstracts Based on Character-Level Sequence Labeling. Journal of Data and Information Science, 6(3), 35–57. https://doi.org/10.2478/jdis-2021-0013

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013). Entitymetrics: Measuring the impact of entities. PloS one, 8(8), e71416. https://doi.org/10.1371/journal.pone.0071416

Heffernan, K., & Teufel, S. (2018). Identifying problems and solutions in scientific text. Scientometrics, 116, 1367–1382. https://doi.org/10.1007/s11192-018-2718-6

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

Hou, L., Zhang, J., Wu, O., Yu, T., Wang, Z., Li, Z., & Yao, R. (2020). Method and dataset entity mining in scientific literature: A CNN+ Bi-LSTM model with self-attention. ArXiv Preprint. arXiv:2010.13583.

Li, K., & Yan, E. (2018). Co-mention network of R packages: Scientific impact and clustering structure. Journal of Informetrics, 12(1), 87–100. https://doi.org/10.1016/j.joi.2017.12.001

Li, K., Yan, E., & Feng, Y. (2017). How is R cited in research outputs? Structure, impacts, and citation standard. Journal of Informetrics, 11(4), 989–1002. https://doi.org/10.1016/j.joi.2017.08.003

Li, X., Rousseau, J.F., Ding, Y., Song, M., & Lu, W. (2020). Understanding Drug Repurposing From the Perspective of Biomedical Entities and Their Evolution: Bibliographic Research Using Aspirin. JMIR medical informatics, 8(6), e16739. https://doi.org/10.2196/16739

Kondo, T., Nanba, H., Takezawa, T., & Okumura, M. (2009). Technical Trend Analysis by Analyzing Research Papers’ Titles. In Proceedings of the 4th Language and Technology Conference. Poznan, Poland: Springer, 512–521. https://doi.org/10.1007/978-3-642-20095-3_47

Névéol, A., Wilbur, W., & Lu, Z. (2011). Extraction of data deposition statements from the literature: A method for automatically tracking research results. Bioinformatics, 27(23), 3306–3312. http://doi.org/10.1093/bioinformatics/btr573

Okamoto, M., Shan, Z., & Orihara, R. (2017). Applying Information Extraction for Patent Structure Analysis. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 989–992. https://doi.org/10.1145/3077136.3080698

Paul, D., Singh, M., Hedderich, M.A., & Klakow, D. (2019). Handling Noisy Labels for Robustly Learning from Self-Training Data for Low-Resource Sequence Labeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. 29–34. http://dx.doi.org/10.18653/v1/N19-3005

Tateisi, Y., Shidahara, Y., Miyao, Y., & Aizawa, A. (2014). Annotation of Computer Science Papers for Semantic Relation Extraction. In Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland: LREC, 1423–1429. http://www.lrec-conf.org/proceedings/lrec2014/summaries/461.html

Vahidnia, S., Abbasi, A., & Abbass, H. (2021).Embedding-based Detection and Extraction of Research Topics from Academic Documents Using Deep Clustering. Journal of Data and Information Science, 6(3), 99–122. https://doi.org/10.2478/jdis-2021-0024

Wang, S., Mao, J., Tang, J., & Cao, Y. (2021). Content Characteristics of Knowledge Integration in the eHealth Field: An Analysis Based on Citation Contexts. Journal of Data and Information Science, 6(3), 123–145. https://doi.org/10.2478/jdis-2021-0015

Wang, Y., & Zhang, C. (2020). Using the Full-text Content of Academic Articles to Identify and Evaluate Algorithm Entities in the Domain of Natural Language Processing. Journal of Informetrics, 14(4), 101091. https://doi.org/10.1016/j.joi.2020.101091

Yang, Y., Chen, W., Li, Z., He, Z., & Zhang, M. (2018). Distantly Supervised NER with Partial Annotation Learning and Reinforcement Learning. COLING. In Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New-Mexico, USA: Association for Computational Linguistics, 2159–2169. http://aclweb.org/anthology/C18-1183

Yoon, J., Chung, E., Lee, J.Y., & Kim, J. (2019). How research data is cited in scholarly literature: A case study of HINTS. Learned Publishing, 32, 199–206. https://doi.org/10.1002/leap.1213

Zadeh, B., & Schumann, A. (2016). The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods. In Proceedings of the Tenth International Conference on Language Resources and Evaluation. Portorož, Slovenia: LREC, 1862–1868. http://www.lrec-conf.org/proceedings/lrec2016/summaries/681.html

Zhang, C., Mayr, P., Lu, W., & Zhang, Y. (2020). Extraction and evaluation of knowledge entities from scientific documents: EEKE2020. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, 573–574. https://doi.org/10.1145/3383583.3398504

Zhang C., Mayr, P., Lu W., & Zhang Y. (2021). Editorial—Knowledge Entity Extraction and Text Mining in the Era of Big Data. Data and Information Management, 5(3), 309–311. https://doi.org/10.2478/dim-2021-0009