Editorial

Open Access

Chengzhi Zhang*, Philipp Mayr, Wei Lu, Yi Zhang

Knowledge Entity Extraction and Text Mining in the Era of Big Data

https://doi.org/10.2478/dim-2021-0009

In the era of big data, tremendous amounts of information and data have drastically changed human civilization. The rapid growth in the number of documents generated everyday means that a large amount of knowledge is proposed, improved, and used. For readers, especially newcomers to a given field, excavating suitable knowledge entities from massive documents is time-consuming and labor-consuming, negatively impacting research efficiency. The broad availability of information provides more opportunities for people, but a new challenge has risen as well; that is, how to extract and use knowledge from numerous information resources, especially how to conduct knowledge extraction and text mining (TM) from massive documents in special domains.

A knowledge entity is a relatively independent and integral knowledge module in a special discipline or research domain (Chang & Zheng, 2007). In scientific documents, knowledge entities refer to the knowledge mentioned or cited by authors, such as algorithms, models, theories, datasets, and software (Wang & Zhang, 2018), and reflect various resources used by the authors in problem-solving (Zhang, Mayr, Lu, & Zhang, 2020; Hou, Jochim, Gleize, Bonin, & Ganguly, 2019; Brack, D'Souza, Hoppe, Auer, & Ewerth, 2020). Extracting knowledge entities from numerous information resources is useful for multiple downstream tasks in information extraction, TM, natural language processing (NLP), information retrieval, digital library research, and so on. Particularly, in the field of artificial intelligence (AI), information science, and some other related disciplines, discovering methods from a large scale of academic literature, and evaluating the performance and influence of such methods, have become increasingly necessary and meaningful (Hou et al., 2020). In 2019, the "Heart of Machine" launched the project "SOTA (state of the art) model"[1]. Targeting more than 100 tasks in machine learning research, the project obtained models, open datasets, evaluation indicators, and results from academic literature through manual annotation and named entity recognition, and provided open retrieval services for users. Defense Advanced Research Projects Agency (DARPA) has recently launched the Automating Scientific Knowledge Extraction (ASKE) project[2] to develop next-generation applications of artificial intelligence.

In parallel, deep learning techniques introduce new progresses to NLP and TM. Many kinds of neural network models, e.g., convolutional neural network (CNN), recurrent neural network (RNN), graph neural networks (GNN), and attention mechanism, have been widely involved in these tasks (Qiu et. al, 2020), particularly, text classification (Zhang, Zhao, & LeCun, 2015; Lai, Xu, Liu, & Zhao, 2015; Yao, Mao, & Luo, 2019; Liu & Guo, 2019) and clustering (Xu et al., 2015; Xu et al., 2017).

There are some conferences and workshops in line with this topic, such as 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), the Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE) (Zhang et al., 2020), the Workshop on AI + Informetrics (AII) (Zhang, Zhang, Mayr, & Suominen, 2021), and the Workshop on Scholarly Document Processing (SDP) (Chandrasekaran et al., 2020).

We are very grateful that three contributions were invited to the special issue of *Data and Information*

***Chengzhi Zhang,** Department of Information Management, Nanjing University of Science and Technology, Nanjing, China, Email: zhangcz@njust.edu.cn
Philipp Mayr, GESIS – Leibniz-Institute for the Social Sciences, Köln, Germany
Wei Lu, School of Information Management, Wuhan University, Wuhan, China
Yi Zhang, Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, Australia

[1] See the website: https://www.jiqizhixin.com/sota

[2] See the website: https://www.darpa.mil/program/automating-scientific-knowledge-extraction

Management (DIM). These three submissions were accepted after several rounds of peer-reviewing and revisions.

The paper “Discovering Booming Bio-entities and Their Relationship with Funds” (Tan, Zhang, Yang, Wu, & Xu, 2021) tracked the overall trends and changes in biomedical topics from 1988 to 2017. It collected funding information in the PubMed database and the website of the United States National Institutes of Health (USNIH), and extracted funding-related entities and research hotspots in the corresponding fields. This study provides new insights for research funding allocation, and may support the science policy and strategic management of stakeholders.

The paper “A Pattern and POS Auto-Learning Method for Terminology Extraction from Scientific Text” (Shao, Hua, & Song, 2021) proposed an unsupervised method based on sentence patterns and part of speech (POS) sequences extracted from scientific texts. The proposed method only requires a few initial learnable patterns to obtain initial terminological tokens and their POS sequences. Experiments on abstracts of articles in the Web of Science (WoS) database demonstrate its recognized performance.

The paper “Automatic Subject Classification of Public Messages in E-government Affairs” (Pan & Chen, 2021) touched upon the task of automatic classification using bi-directional long short-term memory (Bi-LSTM) network model based on attention mechanism. This paper used the Bi-LSTM algorithm to strengthen the relevance of messages before and after the training process, and introduced semantic attention to highlight the weight of important text features.

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. Many thanks also go to Dr. Wu Dan, Executive Editor-in-Chief of DIM, for the invitation, and to Dr. Haiying Huang, editor of DIM, for insightful advice.

References

Brack, A., D’Souza, J., Hoppe, A., Auer, S., & Ewerth, R. (2020) Domain-independent extraction of scientific concepts from research articles. In J. Jose et al. (Eds.) *Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science,* vol. 12035 (pp. 251-266). Cham: Springer. doi: 10.1007/978-3-030-45439-5_17

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). *ACM SIGIR Forum, 50*(2), 36-43.

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

Chang, X., & Zheng, Q. (2007). Knowledge element extraction for knowledge-based learning resources organization. In H. Leung, F. Li, R. Lau, & Q. Li (Eds.). *Advances in Web Based Learning – ICWL 2007. ICWL 2007. Lecture Notes in Computer Science, vol. 4823* (pp. 102-113). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-540-78139-4_10

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

Hou, Y., Jochim, C., Gleize, M., Bonin, F., & Ganguly, D. (2019). Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. *ArXiv Preprint.* arXiv:1906.09317

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 29, 2267-2273. doi: 10.5555/2886521.2886636

Liu, G., & Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing, 337*, 325-338. doi: 10.1016/j.neucom.2019.01.078

Pan, P., & Chen, Y. (2021). Automatic subject classification of public messages in e-government affairs. *Data and Information Management*, *5*(3), 336-347. doi: 10.2478/dim-2021-0004

Qiu, X., Sun, T., Xu, Y., Shao, Y. F., Dai, N., & Huang, X. J. (2020). Pre-trained models for natural language processing: *A survey. Science China Technological Sciences, 63*(10), 1872-1897. doi: 10.1007/s11431-020-1647-3

Shao, W., Hua, B., & Song, L. (2021). A pattern and POS auto-learning method for terminology extraction from scientific text. *Data and Information Management*, *5*(3), 329-335. doi: 10.2478/dim-2021-0005

Tan, F., Zhang, T., Yang, S., Wu, X., & Xu, J. (2021). Discovering booming bio-entities and their relationship with funds. *Data and Information Management, 5*(3), 312-328. doi: 10.2478/dim-2021-0007

Wang, Y., & Zhang, C. (2018). Using full-text of research articles to analyze academic impact of algorithms. In G. Chowdhury, J. McLeod, V. Gillet, & P. Willett (Eds) *Transforming Digital Worlds. iConference 2018. Lecture Notes in Computer Science 10766* (pp. 395-401). Cham: Springer. doi: 10.1007/978-3-319-78105-1_43

Xu, J., Wang, P., Tian, G., Xu, B., Zhao, J., Wang, F., & Hao, H. (2015). Short text clustering via convolutional neural networks. *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 62-69. doi: 10.3115/v1/W15-1509

Xu, J., Xu, B., Wang, P., Zheng, S., Tian, G., Zhao, J., & Xu, B. (2017). Self-taught convolutional neural networks for short text clustering. *Neural Networks, 88*, 22-31. doi: 10.1016/j.neunet.2016.12.008

Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. *Proceedings of the AAAI Conference on*

Artificial Intelligence, 33(01), 7370-7377. doi: 10.1609/aaai.v33i01.33017370

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.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *ArXiv Preprint*. arXiv:1509.01626

Zhang, Y., Zhang, C., Mayr, P. , & Suominen, A. (2021). *Preface to the 1st workshop on AI+ informetrics: Multi-disciplinary interactions on the era of big data*. Retrieved from http://ceur-ws.org/Vol-2871/preface.pdf