New Frontiers of Scientific Text Mining: Tasks, Data, and Tools

Xuan Wang
xwang174@illinois.edu
University of Illinois at Urbana-Champaign

Hongwei Wang
hongweiw@illinois.edu
University of Illinois at Urbana-Champaign

Heng Ji
hengji@illinois.edu
University of Illinois at Urbana-Champaign

Jiawei Han
hanj@illinois.edu
University of Illinois at Urbana-Champaign

ABSTRACT

Exploring the vast amount of rapidly growing scientific text data is highly beneficial for real-world scientific discovery. However, scientific text mining is particularly challenging due to the lack of specialized domain knowledge in natural language context, complex sentence structures in scientific writing, and multi-modal representations of scientific knowledge. This tutorial presents a comprehensive overview of recent research and development on scientific text mining, focusing on the biomedical and chemistry domains. First, we introduce the motivation and unique challenges of scientific text mining. Then we discuss a set of methods that perform effective scientific information extraction, such as named entity recognition, relation extraction, and event extraction. We also introduce real-world applications such as textual evidence retrieval, scientific topic contrasting for drug discovery, and molecule representation learning for reaction prediction. Finally, we conclude our tutorial by demonstrating, on real-world datasets (COVID-19 and organic chemistry literature), how the information can be extracted and retrieved, and how they can assist further scientific discovery. We also discuss the emerging research problems and future directions for scientific text mining.

ACM Reference Format:
Xuan Wang, Hongwei Wang, Heng Ji, and Jiawei Han. 2022. New Frontiers of Scientific Text Mining: Tasks, Data, and Tools. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3534678.3542606

1 TARGET AUDIENCE AND PREREQUISITES

This tutorial is intended for researchers and practitioners in data mining, text mining, graph mining, natural language processing, machine learning, and their applications to other domains. While the audience with a good background in the above areas would benefit most from this tutorial, we believe the material to be presented would give general audience and newcomers a complete picture of the important research topics in scientific text mining. Our tutorial is designed as self-contained, so no specific background knowledge is assumed of the audience.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3542606

2 TUTORS

Xuan Wang is a Ph.D. candidate at Computer Science Department, University of Illinois at Urbana-Champaign. Her research focuses on mining and constructing structured knowledge from massive unstructured corpora with minimum human supervision, emphasizing applications to biological and health sciences. She is the recipient of YEE Fellowship Award in 2020-2021 at UIUC. She has delivered tutorials in IEEE-BigData'19 and WWW'22.
Hongwei Wang is a postdoctoral researcher at Computer Science Department, University of Illinois Urbana-Champaign. His research interests include machine learning and data mining, particularly in graph representation learning mechanisms, algorithms, and their applications in real-world data mining scenarios such as knowledge graphs, recommender systems, social networks, and sentiment analysis. He was one of the recipients of 2020 CCF (China Computer Federation) Outstanding Doctoral Dissertation Award and 2018 Google Ph.D. Fellowship. He has delivered a tutorial in WWW'22.
Heng Ji is a Professor at Computer Science Department of University of Illinois Urbana-Champaign, and an Amazon Scholar. Her research interests focus on Natural Language Processing, especially on Multimedia Multilingual Information Extraction, Knowledge Base Population and Knowledge-driven Generation. She was selected as "Young Scientist" and a member of the Global Future Council on the Future of Computing by the World Economic Forum in 2016 and 2017. The awards she received include "AI's 10 to Watch" Award by IEEE Intelligent Systems in 2013, NSF CAREER award in 2009, Google Research Award in 2009 and 2014, IBM Watson Faculty Award in 2012 and 2014 and Bosch Research Award in 2014-2018, and Amazon AWS Award in 2021. She has delivered more than 15 conference tutorials (e.g., ACL'18, 21, EMNLP'21, and KDD'21).
Jiawei Han is Michael Aiken Chair Professor, Department of Computer Science, University of Illinois at Urbana-Champaign. His research areas encompass data mining, text mining, data warehousing, and information network analysis, with over 1000 research publications. He is Fellow of ACM, Fellow of IEEE, and received numerous prominent awards, including ACM SIGKDD Innovation Award (2004) and IEEE Computer Society W. Wallace McDowell Award (2009). He delivered 50+ conference tutorials or keynote speeches (e.g., SIGKDD 2017-2021 tutorials and WSDM 2018 keynote).

3 TUTORIAL OUTLINE

- Introduction
  - Overview of Scientific Text Mining and Unique Challenges
- Scientific Information Extraction
  - Fine-Grained Scientific Named Entity Recognition (NER) [4, 6, 7, 23, 24, 27, 28, 30]

    - Knowledge-Enhanced Fine-Grained Scientific NER
    - Ontology-guided Fine-Grained Scientific NER
  - Scientific Relation Extraction [7, 9, 26, 30]
    - Abstract Meaning Representation Guided Biomedical Relation Extraction
    - Meta-Pattern Guided Open Relation Extraction
  - Scientific Event Extraction [7, 30]
- Text Mining for Scientific Discovery
  - Textual Evidence Discovery from Scientific Literature [1, 12, 17–19, 22, 25, 29]
    - What is textual evidence discovery?
    - Textual Evidence Discovery in COVID-19 Literature
  - Scientific Topic Contrasting for Drug Discovery [8, 10, 13, 14, 16, 21]
    - What is scientific topic contrasting?
    - Scientific Topic Contrasting for Drug Discovery
  - Chemical-Reaction-Aware Molecule Representation Learning [2, 3, 5, 11, 15, 20]
    - What is Molecule Representation Learning (MRL)?
    - Chemical-Reaction-Aware MRL
- System Demonstrations and Future Directions
  - System Demos (COVID-KG and ReactionTracker) [21]
  - Research Problems and Future Directions

ACKNOWLEDGMENTS

This material is based upon work supported in part by the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, US DARPA KAIROS Program No. FA8750-19-2-1004, U.S. DARPA AIDA Program No. FA8750-18-2-0014, Air Force Nos. FA8650-17-C-7715 and FA8750-20-2-10002, SocialSim Program No. W911NF-17-C-0099, and INCAS Program No. HR001121C0165, and National Science Foundation IIS-19-56151, IIS-17-41317, and IIS 17-04532. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of DARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

REFERENCES

[1] Alexis Allot, Qingyu Chen, Sun Kim, Roberto Vera Alvarez, Donald C Comeau, W John Wilbur, and Zhiyong Lu. 2019. LitSense: making sense of biomedical literature at sentence level. *Nucleic acids research* (2019).

[2] Seyone Chithrananda, Gabe Grand, and Bharath Ramsundar. 2020. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *ArXiv preprint* abs/2010.09885 (2020).

[3] Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230* (2020).

[4] Jiayuan He, Dat Quoc Nguyen, Saber A Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoyan Fang, Hiyori Yoshikawa, et al. 2020. Overview of ChEMU 2020: named entity recognition and event extraction of chemical reactions from patents. In *CLEF*. Springer, 237–254.

[5] Sabrina Jaeger, Simone Fulle, and Samo Turk. 2018. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling* 58, 1 (2018), 27–35.

[6] Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. CHEMDNER: The drugs and chemical names extraction challenge. *Journal of cheminformatics* 7, 1 (2015), S1.

[7] Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. 2021. Joint Biomedical Entity and Relation Extraction with Knowledge-Enhanced Collective Inference. In *ACL*. 6248–6260.

[8] Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, et al. 2020. Gaia: A fine-grained multimedia knowledge extraction system. In *ACL*. 77–86.

[9] Qi Li, Xuan Wang, Yu Zhang, Fei Ling, Cathy Wu H, and Jiawei Han. 2018. Pattern Discovery for Wide-Window Open Information Extraction in Biomedical Literature. In *BIBM*. 420–427.

[10] David A Liem, Sanjana Murali, Dibakar Sigdel, Yu Shi, Xuan Wang, Jiaming Shen, Howard Choi, John H Caufield, Wei Wang, Peipei Ping, et al. 2018. Phrase mining of textual data to analyze extracellular matrix protein patterns across cardiovascular disease. *American Journal of Physiology-Heart and Circulatory Physiology* 315, 4 (2018), H910–H924.

[11] Emily K Mallory, Ambika Acharya, Stefano E Rensi, Peter J Turnbaugh, Roselie A Bright, and Russ B Altman. 2018. Chemical reaction vector embeddings: towards predicting drug metabolism in the human gut microbiome. In *PSB*. 56–67.

[12] Hans-Michael Müller, Eimear E Kenny, and Paul W Sternberg. 2004. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.* 2, 11 (2004), e309.

[13] Thomas Rebele, Fabian Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. 2016. YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames. In *ISWC*. Springer, 177–185.

[14] Xiang Ren, Jiaming Shen, Meng Qu, Xuan Wang, Zeqiu Wu, Qi Zhu, Meng Jiang, Fangbo Tao, Saurabh Sinha, David Liem, et al. 2017. Life-inet: A structured network-based knowledge exploration and analytics system for life sciences. In *ACL*. 55–60.

[15] Stefano Rensi and Russ B Altman. 2017. Flexible analog search with kernel PCA embedded molecule vectors. *Computational and structural biotechnology journal* 15 (2017), 320–327.

[16] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *WWW*. 243–246.

[17] Raphael Tang, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly Bootstrapping a Question Answering Dataset for COVID-19. *arXiv preprint arXiv:2004.11339* (2020).

[18] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics* 16, 1 (2015), 138.

[19] Marco A Valenzuela-Escárcega, Özgün Babur, Gus Hahn-Powell, Dane Bell, Thomas Hicks, Enrique Noriega-Atala, Xia Wang, Mihai Surdeanu, Emek Demir, and Clayton T Morrison. 2018. Large-scale automated machine reading discovers new cancer-driving mechanisms. *Database* 2018 (2018).

[20] Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin D Burke. 2021. Chemical-Reaction-Aware Molecule Representation Learning. *ICLR* (2021).

[21] Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Ranran Haoran Zhang, Weili Liu, et al. 2021. COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation. In *NAACL*. 66–77.

[22] Xuan Wang, Yingjun Guan, Weili Liu, Aabhas Chauhan, Enyi Jiang, Qi Li, David Liem, Dibakar Sigdel, John Caufield, Peipei Ping, and Jiawei Han. 2020. EVIDENCEMINER: Textual Evidence Discovery for Life Sciences. In *ACL*. 56–62.

[23] Xuan Wang, Vivian Hu, Xiangchen Song, Shweta Garg, Jinfeng Xiao, and Jiawei Han. 2021. ChemNER: Fine-Grained Chemistry Named Entity Recognition with Ontology-guided Distant Supervision. In *EMNLP*. 5227–5240.

[24] Xuan Wang, Xiangchen Song, Bangzheng Li, Kang Zhou, Qi Li, and Jiawei Han. 2020. Fine-Grained Named Entity Recognition with Distant Supervision in COVID-19 Literature. In *BIBM*. 491–494.

[25] Xuan Wang, Yu Zhang, Aabhas Chauhan, Qi Li, and Jiawei Han. 2020. Textual Evidence Mining via Spherical Heterogeneous Information Network Embedding. In *BigData*. 828–837.

[26] Xuan Wang, Yu Zhang, Qi Li, Yinyin Chen, and Jiawei Han. 2018. Open Information Extraction with Meta-pattern Discovery in Biomedical Literature. In *BCB*. 291–300.

[27] Xuan Wang, Yu Zhang, Qi Li, Xiang Ren, Jingbo Shang, and Jiawei Han. 2019. Distantly supervised biomedical named entity recognition with dictionary expansion. In *BIBM*. 496–503.

[28] Taiki Watanabe, Akihiro Tamura, Takashi Ninomiya, Takuya Makino, and Tomoya Iwakura. 2019. Multi-Task Learning for Chemical Named Entity Recognition with Chemical Compound Paraphrasing. In *EMNLP-IJCNLP*. 6244–6249.

[29] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research* 41, W1 (2013), W518–W522.

[30] Zixuan Zhang, Nikolaus Parulian, Heng Ji, Ahmed Elsayed, Skatje Myers, and Martha Palmer. 2021. Fine-grained Information Extraction from Biomedical Literature based on Knowledge-enriched Abstract Meaning Representation. In *ACL*. 6261–6270.