Perspective

Fei He, Kai Liu, Zhiyuan Yang, Mark Hannink, Richard D. Hammer, Mihail Popescu and Dong Xu*

Applications of cutting-edge artificial intelligence technologies in biomedical literature and document mining

https://doi.org/10.1515/mr-2023-0011
Received March 9, 2023; accepted May 29, 2023;
published online June 27, 2023

Abstract: The biomedical literature is a vast and invaluable resource for biomedical research. Integrating knowledge from the literature with biomedical data can help biological studies and the clinical decision-making process. Efforts have been made to gather information from the biomedical literature and create biomedical knowledge bases, such as KEGG and Reactome. However, manual curation remains the primary method to retrieve accurate biomedical entities and relationships. Manual curation becomes increasingly challenging and costly as the volume of biomedical publications quickly grows. Fortunately, recent advancements in Artificial Intelligence (AI) technologies offer the potential to automate the process of curating, updating, and integrating knowledge from the literature. Herein, we highlight the AI capabilities to aid in mining knowledge and building the knowledge base from the biomedical literature.

Keywords: artificial intelligence technologies; biomedical literature mining; pathway figure mining; text mining

***Corresponding author: Dong Xu**, Department of Electrical Engineer and Computer Science, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, 65211, MO, USA, E-mail: xudong@missouri.edu. https://orcid.org/0000-0002-4809-0514
Fei He, School of Information Science and Technology, Northeast Normal University, Changchun, Jilin Province, China; and Department of Electrical Engineer and Computer Science, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, USA
Kai Liu and Zhiyuan Yang, School of Information Science and Technology, Northeast Normal University, Changchun, Jilin Province, China
Mark Hannink, Department of Biochemistry, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, USA
Richard D. Hammer, Department of Pathology and Anatomical Sciences, University of Missouri, Columbia, USA
Mihail Popescu, Department of Health Management and Informatics, University of Missouri, Columbia, USA

AI-based text mining applications

The genes, proteins and their relationships are reported in the text of biomedical literature. AI-based text mining tools leverage Natural Language Processing (NLP) to facilitate entity recognition and relation extraction (Table 1, Figure 1).

Named Entity Recognition

Named Entity Recognition (NER), as a typical task of NLP, consists of labeling and identifying names of biological concepts, such as proteins, genes, chemical compounds, drugs, and diseases from a biomedical literature corpus. NER is instrumental in extracting key biological concepts from scientific articles, helping build biological ontologies and knowledge bases.

For example, given the sentence “Mutations in the BRCA1 are associated with an increased risk of breast and ovarian cancer”, an NER tool can tag the word “BRCA1” as a gene, and the words “breast cancer” and “ovarian cancer” as diseases. A major bottleneck of NER is that the same biomedical entity may be mentioned using nonstandard abbreviations and terminologies. For example, the transcription factor “C/EBP-β” is also known as “NF-IL6”; the protein “Arnt” is sometimes referred to as ‘HIF1-β’. Some entities also nest other entities. For instance, the protein entity “alanine aminotransferase” contains the chemical entity “alanine”. To address the above challenges, biomedical entity linking, aka entity normalization or entity grounding, may be used to map ambiguous entities to normalized, unique identifiers from an ontology, such as Gene Ontology.

Relation extraction

Building upon NER, Relation Extraction (RE) involves identifying relationships among the entities previously found. RE focuses on uncovering connections, such as protein-protein interactions, gene-disease association, genotype-phenotype

Table 1: AI-based methods for biomedical literature mining.

Methodology	Task	Source	Tool Url	Reference
Text mining	Named Entity Recognition (NER)	PubTator: a web-based text mining tool for assisting biocuration	https://www.ncbi.nlm.nih.gov/research/pubtator/	https://www.ncbi.nlm.nih.gov/pubmed/31114887
		LATTE: latent type modeling for biomedical entity linking	https://www.ncbi.nlm.nih.gov/research/pubtator/	https://ojs.aaai.org/index.php/AAAI/article/view/6526
		PubmedKB: an interactive web server for exploring biomedical entity relations in the biomedical literature	https://www.pubmedkb.cc/	https://www.ncbi.nlm.nih.gov/pubmed/35536289
	Relation Extraction (RE)	BERE: a novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories	https://github.com/haiya1994/BERE	https://www.nature.com/articles/s42256-020-0189-y/
		PALMER: improving pathway annotation based on the biomedical literature mining with a constrained latent block model	https://dongjunchung.github.io/palmer/	https://www.ncbi.nlm.nih.gov/pubmed/33008309
		GAIL: an interactive webserver for inference and dynamic visualization of gene-gene associations based on gene ontology guided mining of biomedical literature	https://chunglab.io/GAIL/	https://www.ncbi.nlm.nih.gov/pubmed/31260503
		Relation extraction for biological pathway construction using node2vec	https://github.com/eliorc/node2vec	https://www.ncbi.nlm.nih.gov/pubmed/29897325
		miRiaD: a text mining tool for detecting associations of microRNAs with diseases		https://www.ncbi.nlm.nih.gov/pubmed/27216254
	Pretrained models	BioBERT: a pre-trained biomedical language representation model for biomedical text mining	https://github.com/dmis-lab/biobert-pytorch	https://www.ncbi.nlm.nih.gov/pubmed/31501885
		PubMedBERT: domain-specific language model pretraining for biomedical natural language processing	https://microsoft.github.io/BLURB/	https://dl.acm.org/doi/10.1145/3458754
		SciFive: a text-to-text transformer model for biomedical literature	https://github.com/justinphan3110/SciFive	https://arxiv.org/abs/2106.03598
		BioGPT: generative pre-trained transformer for biomedical text generation and mining	https://github.com/microsoft/BioGPT	https://doi.org/10.1093/bib/bbac409
		BioReader: a text mining tool for performing classification of biomedical literature	http://www.cbs.dtu.dk/services/BioReader/	https://www.ncbi.nlm.nih.gov/pubmed/30717659
Image understanding	Gene and relation extraction	Pathway information extracted from 25 years of pathway figures		https://pubmed.ncbi.nlm.nih.gov/33168034/
		Identifying genes and their interactions from pathway Figures and text in biomedical articles.		https://ieeexplore.ieee.org/document/9669391
		Identifying genes in published pathway Figure Images		https://www.biorxiv.org/content/10.1101/379446v1
		Extracting molecular entities and their interactions from pathway figures based on deep learning		https://ieeexplore.ieee.org/document/8983234
	Figure classification	A novel figure panel classification and extraction method for document image understanding		https://pubmed.ncbi.nlm.nih.gov/24783406/
		Novel image features for categorizing biomedical images		https://ieeexplore.ieee.org/document/6392689/
		Figure classification in biomedical literature to elucidate disease mechanisms, based on pathways		https://pubmed.ncbi.nlm.nih.gov/20427165/

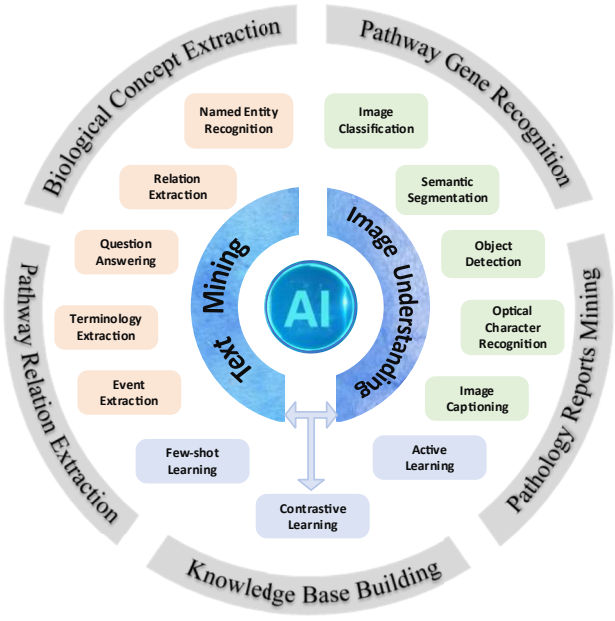

Figure 1: Scope of AI technologies in biomedical literature mining. The figure comprises two panels that showcase the different directions and applications of AI in text mining (on the left) and image understanding (on the right). The left panel features red boxes that outline the various tasks involved in text mining, while the right panel highlights green boxes that represent the tasks of image understanding in the context of biomedical literature mining. Additionally, the figure includes blue boxes that enumerate some promising AI advances aimed at addressing the limitations of current AI methods for biomedical literature mining. The figure's outer circle depicts some typical AI applications in biomedical literature mining. AI, Artificial Intelligence.

relations, chemical-protein interactions, and drug-drug interactions. RE is formulated by recognizing, in a given sentence, an entity pair and the relation type.

Representative text mining techniques

AI-based NER methods can learn the context and model the word semantics to differentiate biologically meaningful concepts from the rest of the words. For example, PubTator [1] is designed to tag words falling into six types of biological concepts, including genes/proteins, gene variants, diseases, chemicals, species, and others, from the abstract or full text of biological publications. Traditionally, the tagged bio-entities can be mapped to their standard forms by performing a fuzzy string match. At the same time, recent research improves this mapping in a latent space (embedding) built by a neural network [2]. Some other AI-based RE techniques employ contextualized representations of biomedical sentences to detect biomedical entity relationships. They extract and aggregate features of sentences from semantic and syntactic aspects, and from multiple views for identifying relations [3]. Along this line, recent research moves relationship mining from the sentence level to the document level to further enrich the knowledge extraction results [4].

Pretrained models

AI-based text mining often relies on robust semantic representations with pretrained models on large-scale web corpora by self-supervised learning. The pretrained models can be further fine-tuned with a relatively small dataset for a specific task. Some well-known NLP models include BERT, T5, and GPT. BioBERT [5] is a widely used domain-specific language representation model pre-trained on large-scale biomedical corpora (PubMed abstract and PMC full-text article) starting from the general BERT model. BioBERT is able to extend to biomedical NER, RE, and question answering (QA). Another tool PubMedBERT [6] uses PubMed's abstracts and PubMedCentral's full-text articles for pre-training from scratch. SciFive [7] is a domain-specific T5 model pre-trained on large biomedical corpora for text understanding tasks (i.e., NER, RE, and QA) and biomedical text generation. More recently, BioGPT [8] pre-trained the GPT-2 model with 15 million PubMed abstracts from scratch to generate fluent descriptions for biomedical terms.

AI-based pathway figure mining

In addition to text, the biomedical literature also contains valuable knowledge in the form of figures. Researchers often use diagrams, such as biological pathways, to summarize their findings in publications for molecular events leading to a biological process or disease. The advancements in AI-based image understanding technologies have improved our capacity for extracting entities and relationships from pathway diagrams, which may be used to complement the same knowledge extracted from text.

Bio-entity mining from pathway figures

Early AI-based methods extracted biomedical entities using Optical Character Recognition (OCR) techniques to recover the gene names from pathway figures. Due to the challenges of nonstandard abbreviations and terminologies, such an approach requires domain experts to manually create some entity normalization rules to ground the gene names. A study applied this method to the pathway figures from the

publications in the past 25 years and recognized thousands of genes missing from pathway databases [9].

Bio-interactions mining from pathway figures

Furthermore, our tool Pathway Curator [10, 11] was designed to extract molecular entities and their interactions from pathway figures. Our pipeline integrates an image understanding model and an image processing strategy to capture the locations, names, and interactions of pathway entities in the figure. The pipeline can recognize genes using symbols and gene relationships using arrows (for upregulation) or T-bars (for inhibition). Pathway Curator provides a complementary approach to text-mining in biological literature mining and a comprehensive view of a disease pathway across multiple publications. Our approach can be extended to other RE tasks for figures, such as microRNA-gene and chemical-protein interactions.

Challenges and outlook

Even though AI technologies, especially deep learning algorithms, have shown great capacity for curating biomedical entities and relationships in an automatic procedure, some limitations of AI technologies still hinder their replacement of manual literature curation:

(1) *Limited annotated data from the biomedical literature.* Annotating sufficient biomedical concepts and relationships for AI training is challenging due to the large volume of publications and the diversity of expressions used. The quantity and quality of labeled data play a crucial role in the robustness of AI models in biomedical literature mining.

(2) *Limitations of current AI's capacities in discovering objects from an established vocabulary.* Currently, most AI approaches are built on pre-defined corpora or pre-labeled datasets. This data dependency limits the ability of AI technologies to mine objects outside the vocabulary.

(3) *Limitations of current AI's capacities in handling inconsistencies from the literature.* The biomedical literature contains outdated or incorrect statements, which may mislead AI approaches.

The rapid development of AI technologies, particularly deep-learning methods, has created new opportunities for curating biomedical knowledge. OpenAI's recent AI-based chatbot, Chat Generative Pre-trained Transformer (ChatGPT), impressed users with its ability to write essays, answer questions, and mimic human conversation. With a comprehensive knowledge base, ChatGPT can be used to retrieve biomedical knowledge at a user's request, paving the way for more efficient and accurate knowledge mining in the biomedical domain. While ChatGPT currently often provides incorrect or unreproducible information, continuous upgrades and better prompt learning techniques offer the potential for more accurate and reliable biomedical knowledge mining. ChatGPT can also be more trained/aligned to target the biomedical field specifically for building hypotheses, finding new drug targets, and generating new small molecules and antibodies. It is promising to change the way of mastering knowledge and skills, assisting doctors to make clinical decisions, and reducing medical errors.

Several new AI approaches hold great potentials to advance literature mining from both text and figure modalities. Active learning allowing iteratively training models with newly labeled data offers the opportunity to gradually upgrade AI models against limited annotated data. Reinforcement Learning from Human Feedback (RLHF) enhances the robustness and generalization of AI algorithms by aligning predictions with human values and preferences. By incorporating human feedback, RLHF can improve the accuracy and reliability of AI predictions, going beyond the limitations of annotated data. In addition, contrastive learning for multiple modalities (i.e., text and image) enables one to learn the common (joint) semantic representations between corresponding text and image, e.g., a gene name 'AKT' in text and an image snippet containing 'AKT' for better performance. Additionally, meta-learning and few-shot learning strategies are also promising to generalize the AI technologies modeling on large-scale general corpora to the biomedical-specific domain.

In clinical practice, various clinical documents, such as Electronic Health Records (EHR) and Pathology Reports (PR), contain significant biomedical and pathological information that can benefit from applying AI technologies for large-volume curation. Several studies have developed AI-based mining tools for EHR [11] and PR [12–14], which utilize similar technologies as those used in literature mining to recognize diagnostic entities and relationships from unstructured text and biomedical images. Integrating the mining results from biomedical literature and clinical documents can facilitate clinical studies and precision medicine. This approach holds great promise for future medical research and patient care.

The rapid growth of biomedical literature presents both opportunities and challenges for biomedical knowledge mining. With more efforts to apply cutting-edge AI technologies to biomedical literature mining, the pace of related annotation, prediction, and knowledge base construction will be accelerated for biomedical research and clinical practices.

Research funding: This work was supported by the National Library of Medicine of the National Institute of Health (NIH) award number 5R01LM013392.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: Authors state no conflict of interest.

Informed consent: Not applicable.

Ethical approval: The local Institutional Review Board deemed the study exempt from review.

References

1. Wei C-H, Kao H-Y, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res 2013;41:W518–22.
2. Zhu M, Celikkaya B, Bhatia P, Reddy CK. LATTE: latent type modeling for biomedical entity linking. Proc AAAI Conf Artif Intell 2020;34:9757–64.
3. Hong L, Lin J, Li S, Wan F, Yang H, Jiang T, et al. A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories. Nat Mach Intell 2020;2: 347–55.
4. Nam JH, Couch D, da Silveira WA, Yu Z, Chung D. Palmer: improving pathway annotation based on the biomedical literature mining with a constrained latent block model. BMC Bioinf 2020;421:432.
5. Lee J, Yoon W, Kim S, Kim D, Kim S, Ho CS, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36:1234–40.
6. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for niomedical natural language processing. ACM Trans Comput Healthc 2021;3:1–23.
7. Phan LN, Anibal JT, Tran H, Chanana S, Bahadroglu E, Peltekian A, et al. SciFive: a text-to-text transformer model for biomedical literature. ArXiv 2021; abs/2106.03598.
8. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022;23:bbac409.
9. Hanspers K, Riutta A, Summer-Kutmon M, Pico AR. Pathway information extracted from 25 years of pathway figures. Genome Biol 2020;21:273.
10. He F, Thompson J, Mao Z, Ren Y, Nussbaum Y, Kholod O, et al. Identifying genes and their interactions from pathway figures and text in biomedical articles. IEEE Int Conf Bioinform Biomed 2021: 398–405.
11. He F, Wang D, Innokenteva Y, Kholod O, Shin D, Dong X. Extracting molecular entities and their interactions from pathway figures based on deep learning. IEEE Int Conf Bioinform Biomed 2019:1191–3.
12. Derington CG, Mueller SR, Glanz JM, Binswanger IA. Identifying naloxone administrations in electronic health record data using a text-mining tool. Subst Abuse 2021;42:806–12.
13. PericlesGiannaris S, Al-Taie Z, Kovalenko M, Thanintorn N, Kholod O, Innokenteva Y, et al. Artificial intelligence-driven structurization of diagnostic information in free-text pathology reports. J Pathol Inf 2020; 11:4.
14. Giannaris PS, Al-Taie Z, Kovalenko M, Hammer RD, Popescu M, Shin D. Informatics framework to identify consistent diagnostic techniques. IEEE Int Conf Bioinform Biomed 2019:1481–6.