Preface to the Special Issue of the TAL Journal on Scholarly Document Processing

Florian Boudin* — **Akiko Aizawa****

* *JFLI, CNRS, Nantes University, France*
** *National Institute of Informatics, Japan*

ABSTRACT. *The rapid growth of scholarly literature makes it increasingly difficult for researchers to keep up with new knowledge. Automated tools are now more essential than ever to help navigate and interpret this vast body of information. Scientific papers pose unique difficulties, with their complex language, specialized terminology, and diverse formats, requiring advanced methods to extract reliable and actionable insights. Large language models (LLMs) offer new opportunities, enabling tasks such as literature reviews, writing assistance, and interactive exploration of research. This special issue of the TAL journal highlights research addressing these challenges and, more broadly, research on natural language processing and information retrieval for scholarly and scientific documents.*

RÉSUMÉ. *La croissance rapide de la littérature scientifique rend de plus en plus difficile pour les chercheurs de suivre l'évolution des connaissances. Le recours à des outils automatisés est aujourd'hui indispensable pour naviguer et interpréter cette immense masse d'informations. Les articles scientifiques posent des difficultés uniques en raison de leur langage complexe, de leur terminologie spécialisée et de leurs formats variés, ce qui nécessite des méthodes avancées pour extraire des informations fiables et exploitables. Les grands modèles de langage (LLMs) ouvrent de nouvelles perspectives, permettant des tâches telles que les revues de littérature, l'assistance à la rédaction et l'exploration interactive des travaux scientifiques. Ce numéro spécial de la revue TAL met en lumière des recherches qui s'attaquent à ces défis et, plus largement, des recherches sur le traitement automatique des langues et la recherche d'information appliqués aux documents scientifiques et académiques.*

KEYWORDS: *Scholarly Document Processing, Natural Language Processing for Science, Large Language Models (LLMs).*

MOTS-CLÉS : *Traitement automatique de documents scientifiques, Traitement Automatique des Langue pour la science, Grands modèles de langues (LLMs).*

1. Introduction

The volume of scholarly literature is expanding rapidly. A compelling example is the ACL Anthology[1], a repository for scientific contributions within the fields of computational linguistics and Natural Language Processing (NLP), which recently surpassed 100,000 papers, doubling its size in just four years (Bollmann *et al.*, 2023). As the rate of publication continues to accelerate, researchers and institutions face increasing challenges in keeping up with the flood of new knowledge. This highlights the critical need for automated methods to help navigate, understand and distill the growing body of scientific information.

To address this pressing challenge, researchers across various fields—including computational linguistics, NLP, text mining, information retrieval, digital libraries and scientometrics—have dedicated significant efforts into developing methods and resources designed to process scientific documents. This led to a surge in publications on the matter, alongside the successful hosting of numerous international events, such as the workshops *Scholarly Document Processing (SDP)* (Ghosal *et al.*, 2024), *SCIentific DOCument Analysis (SCIDOCA)* (Nguyen and Matsumoto, 2024), *Natural Language Processing for Scientific Text (SciNLP)* (Cohan *et al.*, 2021) and *Bibliometric-enhanced Information Retrieval (BIR)* (Frommholz *et al.*, 2024).

At the national level in France, scholarly document processing is also gaining momentum. This interest is exemplified by the success of the workshop *Analyse et Recherche de Textes Scientifiques (ARTS)*[2] (Boudin *et al.*, 2023), held at the TALN-CORIA 2023 conference. The event, which saw the presentation of 12 papers and attracted over 40 participants, highlighted the relevance of the topic and prompted the call for this special issue of the TAL journal.

Scientific papers present unique challenges for document processing methods due to their inherent complexity. They are characterized by intricate technical language, discipline-specific terminology, distinct structural conventions, and frequent use of mathematical expressions, all of which pose significant challenges for current methods (Ramesh Kashyap *et al.*, 2023). Additionally, the multi-modal nature of scientific papers, with their tables, figures and diagrams, further complicates their processing (Shen *et al.*, 2022). Beyond these document-level challenges, effective methods should also account for features present at the collection level, such as citation networks, and leverage rich metadata, including authors, keywords, and publication venues, each introducing its own set of difficulties.

Developing methods to extract reliable, valuable and verifiable information from scientific papers is crucial for many downstream applications, including retrieval (Boudin *et al.*, 2020; Wang *et al.*, 2023), recommendation (Kreutz and Schenkel, 2022; Huang *et al.*, 2024), summarization (Luo *et al.*, 2023), question-answering (Saikh *et al.*, 2022; Auer *et al.*, 2023) and document understanding (Wright

---

1. `https://aclanthology.org/`
2. `https://arts2023.sciencesconf.org/`

and Augenstein, 2021; Veyseh *et al.*, 2021). With the rise of large language models (LLMs) and their enhanced ability to analyze and synthetize insights across multiple scientific papers, new applications are continuously emerging. Promising developments include accelerating scientific discovery (Zhang *et al.*, 2024b), generating novel research directions (Wang *et al.*, 2024), reviewing of the literature (Newman *et al.*, 2024), assisting scientific writing (Jourdan *et al.*, 2024) and enabling interactive exploration of papers (Zheng *et al.*, 2024).

LLMs are also being developed for specialized scientific domains, such as healthcare and medicine (Labrak *et al.*, 2024) or material sciences (Zhang *et al.*, 2024a). These domain-specific models assist experts and researchers with complex tasks, including drug discovery (Savage, 2023), diagnosis generation (Abdullahi *et al.*, 2024), and science education (Cooper, 2023).

Efforts are also underway to reduce the growing computational and environmental costs associated with training and deploying LLMs (Hershcovich *et al.*, 2022; Sadat Moosavi *et al.*, 2023). At the same time, ethical concerns are being addressed, with research focused on the responsible use of LLMs in science, including issues of bias, fairness, and transparency in AI-driven research (Peled-Cohen *et al.*, 2024).

This special issue of the TAL journal focuses on research addressing these challenges, with an emphasis on *NLP and information retrieval for scholarly and scientific documents*.

2. Call, Reviewing and Selection of Papers

The call for submissions to this special issue of the TAL journal on scholarly document processing was announced in December 2023, and the submission platform[3] closed in March 2024. The scope of relevant topics extended beyond NLP and information retrieval tasks, tools, and resources designed for scientific documents, encompassing areas such as bibliometrics, scientometrics, citation analysis and recommendation, claim verification, plagiarism detection and scientific writing assistance.

A total of five articles were submitted (two in French and three in English) by authors from Iran, India and France. Each article was reviewed by three experts, two members of the scientific committee and one member of the Editorial Board. In May 2024, the Editorial Board and the guest editors met to discuss the first round of reviews and notified the authors of the outcomes. One paper was selected for a second round of reviews and was ultimately accepted, resulting in a selection rate of 20%.

---

3. `https://tal-65-2.sciencesconf.org/`

3. Accepted paper

This issue of the TAL journal features one paper: *Évaluation de la qualité de rapport des essais cliniques avec des larges modèles de langue* (Evaluating clinical trials research article quality with large language models) by Mathieu Laï-king and Patrick Paroubek. The paper focuses on the biomedical domain, specifically investigating the use of LLMs to evaluate the quality of Randomized Controlled Trials (RCTs), a type of clinical research article. The authors frame the evaluation task as a question-answering problem, prompting LLMs to assess articles according to the Consolidated Standards of Reporting Trials (CONSORT) framework. Through extensive experiments, the study demonstrates the high effectiveness of LLMs, achieving an accuracy of up to 85%, thus paving the way for advancements in automating quality assessment in clinical research.

4. References

Abdullahi T., Singh R., Eickhoff C., "Learning to Make Rare and Complex Diagnoses With Generative AI Assistance: Qualitative Study of Popular Large Language Models", *JMIR Med Educ*, vol. 10, p. e51391, Feb, 2024.

Auer S., Barone D. A., Bartz C., Cortes E. G., Jaradeh M. Y., Karras O., Koubarakis M., Mouromtsev D., Pliukhin D., Radyush D. *et al.*, "The sciqa scientific question answering benchmark for scholarly knowledge", *Scientific Reports*, vol. 13, n⁰ 1, p. 7240, 2023.

Bollmann M., Schneider N., Köhn A., Post M., "Two Decades of the ACL Anthology: Development, Impact, and Open Challenges", *in* L. Tan, D. Milajevs, G. Chauhan, J. Gwinnup, E. Rippeth (eds), *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, Association for Computational Linguistics, Singapore, p. 83-94, December, 2023.

Boudin F., Daille B., Dufour R., El O., Houbre M., Jourdan L., Kooli N. (eds), *Actes de CORIA-TALN 2023. Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023*, ATALA, Paris, France, 6, 2023.

Boudin F., Gallina Y., Aizawa A., "Keyphrase Generation for Scientific Document Retrieval", *in* D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 1118-1126, July, 2020.

Cohan A., Dasigi P., Hope T., Lo K., Mohan S., Wade A., Wang L., Williams I., Zhang D. (eds), *Proceedings of the 2nd Workshop on Natural Language Processing for Scientific Text (SciNLP 2021)*, Virtual, August, 2021.

Cooper G., "Examining science education in ChatGPT: An exploratory study of generative artificial intelligence", *Journal of Science Education and Technology*, vol. 32, n⁰ 3, p. 444-452, 2023.

Frommholz I., Mayr P., Cabanac G., Verberne S. (eds), *Proceedings of the 14th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2024)*, Glasgow, Scotland, March, 2024.

Ghosal T., Singh A., Waard A., Mayr P., Naik A., Weller O., Lee Y., Shen S., Qin Y. (eds), *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, Association for Computational Linguistics, Bangkok, Thailand, August, 2024.

Hershcovich D., Webersinke N., Kraus M., Bingler J., Leippold M., "Towards Climate Awareness in NLP Research", *in* Y. Goldberg, Z. Kozareva, Y. Zhang (eds), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, p. 2480-2494, December, 2022.

Huang Z., Tang D., Zhao R., Rao W., "A scientific paper recommendation method using the time decay heterogeneous graph", *Scientometrics*, vol. 129, n⁰ 3, p. 1589-1613, 2024.

Jourdan L., Boudin F., Hernandez N., Dufour R., "CASIMIR: A Corpus of Scientific Articles Enhanced with Multiple Author-Integrated Revisions", *in* N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (eds), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, ELRA and ICCL, Torino, Italia, p. 2883-2892, May, 2024.

Kreutz C. K., Schenkel R., "Scientific paper recommendation systems: a literature review of recent publications", *International journal on digital libraries*, vol. 23, n⁰ 4, p. 335-369, 2022.

Labrak Y., Bazoge A., Morin E., Gourraud P.-A., Rouvier M., Dufour R., "BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains", *in* L.-W. Ku, A. Martins, V. Srikumar (eds), *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics, Bangkok, Thailand, p. 5848-5864, August, 2024.

Luo Z., Xie Q., Ananiadou S., "CitationSum: Citation-aware Graph Contrastive Learning for Scientific Paper Summarization", *Proceedings of the ACM Web Conference 2023*, WWW '23, Association for Computing Machinery, New York, NY, USA, p. 1843–1852, 2023.

Newman B., Lee Y., Naik A., Siangliulue P., Fok R., Kim J., Weld D. S., Chang J. C., Lo K., "ArxivDIGESTables: Synthesizing Scientific Literature into Tables using Language Models", *in* Y. Al-Onaizan, M. Bansal, Y.-N. Chen (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Miami, Florida, USA, p. 9612-9631, November, 2024.

Nguyen M. L., Matsumoto Y. (eds), *Proceedings of the Eighth International Workshop on SCIentific DOCument Analysis (SCIDOCA 2024)*, Hamamatsu, Shizuoka, Japan, August, 2024.

Peled-Cohen L., Calderon N., Lissak S., Reichart R. (eds), *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, Association for Computational Linguistics, Miami, FL, USA, November, 2024.

Ramesh Kashyap A., Yang Y., Kan M.-Y., "Scientific document processing: challenges for modern learning methods", *Int. J. Digit. Libr.*, vol. 24, n° 4, p. 283–309, mar, 2023.

Sadat Moosavi N., Gurevych I., Hou Y., Kim G., Kim Y. J., Schuster T., Agrawal A. (eds), *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, Association for Computational Linguistics, Toronto, Canada (Hybrid), July, 2023.

Saikh T., Ghosal T., Mittal A., Ekbal A., Bhattacharyya P., "Scienceqa: A novel resource for question answering on scholarly articles", *International Journal on Digital Libraries*, vol. 23, n° 3, p. 289-301, 2022.

Savage N., "Drug discovery companies are customizing ChatGPT: here's how", *Nat Biotechnol*, vol. 41, n° 5, p. 585-586, 2023.

Shen Z., Lo K., Wang L. L., Kuehl B., Weld D. S., Downey D., "VILA: Improving Structured Content Extraction from Scientific PDFs Using Visual Layout Groups", *Transactions of the Association for Computational Linguistics*, vol. 10, p. 376-392, 2022.

Veyseh A. P. B., Dernoncourt F., Nguyen T. H., Chang W., Celi L. A., "Acronym Identification and Disambiguation Shared Tasks for Scientific Document Understanding", 2021.

Wang J. A., Wang K., Wang X., Naidu P., Bergen L., Paturi R., "Scientific Document Retrieval using Multi-level Aspect-based Queries", *in* A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (eds), *Advances in Neural Information Processing Systems*, vol. 36, Curran Associates, Inc., p. 38404-38419, 2023.

Wang Q., Downey D., Ji H., Hope T., "SciMON: Scientific Inspiration Machines Optimized for Novelty", *in* L.-W. Ku, A. Martins, V. Srikumar (eds), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Bangkok, Thailand, p. 279-299, August, 2024.

Wright D., Augenstein I., "CiteWorth: Cite-Worthiness Detection for Improved Scientific Document Understanding", *in* C. Zong, F. Xia, W. Li, R. Navigli (eds), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, Online, p. 1796-1807, August, 2021.

Zhang H., Song Y., Hou Z., Miret S., Liu B., "HoneyComb: A Flexible LLM-Based Agent System for Materials Science", *in* Y. Al-Onaizan, M. Bansal, Y.-N. Chen (eds), *Findings of the Association for Computational Linguistics: EMNLP 2024*, Association for Computational Linguistics, Miami, Florida, USA, p. 3369-3382, November, 2024a.

Zhang Y., Chen X., Jin B., Wang S., Ji S., Wang W., Han J., "A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery", *in* Y. Al-Onaizan, M. Bansal, Y.-N. Chen (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Miami, Florida, USA, p. 8783-8817, November, 2024b.

Zheng Y., Sun S., Qiu L., Ru D., Jiayang C., Li X., Lin J., Wang B., Luo Y., Pan R., Xu Y., Min Q., Zhang Z., Wang Y., Li W., Liu P., "OpenResearcher: Unleashing AI for Accelerated Scientific Research", *in* D. I. Hernandez Farias, T. Hope, M. Li (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Miami, Florida, USA, p. 209-218, November, 2024.