AI-Driven Metadata Extraction from Scientific Texts for Enhanced Biomedical Data Reusability

Vigneshwaran Thangaraju
CGI Technology and Solutions
USA
vigsan714@gmail.com

S.Sangeetha
Artificial Intelligence & Data Science
PSG Institute of Technology and Applied Research
Coimbatore, India
0000-0003-2513-5502

2025 International Conference on Next Generation Computing Systems (ICNGCS) | 979-8-3315-9116-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICNGCS64900.2025.11183478

***Abstract*— Biomedical research has generated enormous volumes of data, yet the lack of structured, high-quality metadata often limits its full potential. Manual curation of metadata is labor-intensive, error-prone, and non-scalable, which hinders data integration and impedes data-driven discovery. In this paper, we propose an AI-driven framework for automatically extracting and enhancing metadata from unstructured scientific texts. Our approach leverages state-of-the-art deep learning techniques, including Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks, to identify and predict relevant metadata elements such as experimental conditions, disease categories, and biological sample types. Comprehensive evaluations on large-scale biomedical datasets, such as the Gene Expression Omnibus (GEO), demonstrate significant improvements in metadata quality and consistency. The proposed method not only facilitates data reuse and interoperability but also supports the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, thereby accelerating biomedical research and discovery.**

Keywords*— *Biomedical Data, Metadata Extraction, Deep Learning, Convolutional Neural Networks, BiLSTM, FAIR Principles, Natural Language Processing, Gene Expression Omnibus.

I. INTRODUCTION

The proliferation of biomedical datasets, fueled by highthroughput technologies, has revolutionized research by enabling extensive omics profiling. Public repositories now host diverse datasets (e.g., gene expression, microRNA, methylation), fostering data reuse and discovery.

However, the utility of these datasets is often hampered by poor metadata quality. Metadata—detailing experimental conditions, sample origin, and processing—are vital for achieving FAIR principles. Yet, inconsistent formats and uncontrolled vocabularies limit data interoperability and reuse.

Platforms like GEO accept metadata in varied formats (e.g., XML, CSV), typically as unstructured key-value pairs, which pose the following challenges:

- Spelling Errors: e.g., “diagnosis” vs. “diagonosis”
- Format Variation: e.g., “age (years)” vs. “age year”
- Term Mismatch: e.g., “disease” vs. “condition”
- Compound Keys: e.g., “tissue/cell line”

Such issues hinder metadata indexing, integration, and reuse. Manual correction is impractical at scale. To address this, a machine learning-driven framework is proposed to automate metadata extraction and normalization, with the following aims:

- Extract structured metadata from free-text biomedical sources
- Align entries with controlled vocabularies
- Enable scalable curation through deep learning

This facilitates improved search, semantic interpretation, and alignment with FAIR standards.

This work presents a deep learning pipeline that:

- Leverages CNN–BiLSTM to capture local and contextual features
- Applies domain-specific preprocessing for biomedical texts
- Validates metadata using RDF and SPARQL
- Outperforms traditional models (e.g., TF-IDF + SVM) in precision and recall

II. RELATED WORK

A. Metadata Challenges in Biomedical Repositories

Repositories such as BioSamples and GEO exhibit metadata inconsistencies due to unstandardized submissions. Prior solutions include clustering and crowdsourcing, though they often yield low agreement. TF-IDF and SVM-based classifiers improve precision but struggle with scalability and sparsity.

B. Biomedical Term Extraction Techniques

a) Topic Models.: LDA and NMF help discover latent metadata structures, though TF-IDF may downweight rare yet relevant terms.

b) Embeddings.: Word2Vec models trained on biomedical corpora improve semantic understanding. Preprocessing enhancements, such as sentence shuffling, further boost performance.

c) NER.: Deep models like LSTM-CRF accurately extract biomedical entities (e.g., genes, conditions) by leveraging domain-specific corpora.

d) Traditional ML.: Binary classifiers perform well for certain keys (e.g., genotype), but multi-label tasks remain more complex and error-prone.

Despite advances, metadata normalization and standardization remain open challenges, addressed here using a hybrid deep learning approach.

979-8-3315-9116-8/25/$31.00 ©2025 IEEE

Comparison with Baselines

The proposed model achieves the highest scores, validating its effectiveness in capturing both local and sequential patterns in biomedical text.

VI. CONCLUSION

In this paper, we presented a comprehensive framework for extracting metadata from scientific texts by integrating traditional multi-label classification techniques with state-of-the-art deep learning architectures, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units. Our approach leverages pre-trained word embeddings to capture the nuanced semantics of biomedical texts, thereby enhancing the quality and consistency of extracted metadata.

The experimental results indicate that deep learning models, particularly CNNs, are effective in capturing local textual patterns, while RNNs and LSTMs excel at modeling sequential dependencies within the text. However, the overall performance remains constrained by the limited availability of large, gold standard annotated datasets and the inherent semantic ambiguities present in unstructured biomedical text.

Future research should focus on expanding the annotated corpus to include additional metadata types such as organism, tissue, and cell line, as well as exploring advanced neural architectures and transfer learning techniques. Integrating external domain knowledge through ontologies and knowledge graphs may further improve the accuracy of metadata extraction, thereby advancing the FAIR principles of data reusability and interoperability in biomedical research.

REFERENCES

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[3] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. International Conference on Learning Representations (ICLR)*, 2013.

[5] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[6] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014. [7] Gene Expression Omnibus (GEO), https://www.ncbi.nlm. nih.gov/geo/.

[8] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, 2016.

[9] U.S. National Library of Medicine, "Unified Medical Language System (UMLS)," https://www.nlm.nih.gov/research/umls/.

[10] A. Brazma *et al.*, "Minimum information about a microarray experiment (MIAME) – toward standards for microarray data," *Nature Genetics*, vol. 29, no. 4, pp. 365–371, 2001.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[12] J. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. NAACL-HLT*, pp. 260–270, 2016.

[13] S. Badami, "Optimizing LSM Tree Operations with Deferred Updates: A Comparative Study," in *7th International Conference on Data Science and Information Technology (DSIT)*, 2024, pp. 1-9.

[14] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *European Conference on Machine Learning*, 1998, pp. 137–142.

[15] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. of NAACL-HLT*, 2016, pp. 260–270.

[16] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. of EMNLP*, 2014, pp. 1746–1751.