LAEDC

Recibido 29 Oct. 2025 *ReCIBE, Año 14 No. 3, Dic. 2025*

Aceptado 24 Nov. 2025

Biomedical Entity Recognition Using a Small Language Model

Reconocimiento de Entidades Biomédicas utilizando un Modelo de Lenguaje Pequeño

Amed Clavería Méndez[1]

Israel Roman Godínez[1]
israel.roman@academicos.udg.mx

Sulema Torres Ramos[1]

Stewart Rene Santos Arce[1]

[1]University of Guadalajara

RESUMEN
Esta investigación propone el desarrollo de un Modelo de Lenguaje Pequeño (SLM, por sus siglas en inglés) computacionalmente eficiente para el Reconocimiento de Entidades Nombradas Biomédicas (NER). Mediante el ajuste fino (fine-tuning) de un SLM preentrenado de dominio general (basado en Modelos de Espacio de Estados Selectivos, SSM) sobre textos biomédicos especializados, se busca lograr un rendimiento comparable al de modelos de mayor tamaño. Nuestro enfoque aborda las limitaciones de los modelos basados en Transformers, como BioBERT (con más de 110 millones de parámetros), al combinar el procesamiento de secuencias en tiempo lineal con técnicas de adaptación al dominio biomédico.

PALABRAS CLAVE: Modelo de Lenguaje Pequeño, Modelos de Espacio de Estados Selectivos, Reconocimiento de Entidades Nombradas Biomédicas.

ABSTRACT
This research proposes developing a computationally efficient Small Language Model (SLM) for Biomedical Named Entity Recognition (NER). By fine-tuning a general-domain pretrained SLM (based on selective state space models, SSMs) on specialized biomedical texts, we aim to achieve performance comparable to larger models. Our approach addresses limitations of Transformer-based models like BioBERT (110M+ parameters) by combining linear-time sequence processing with biomedical domain adaptation techniques.

KEYWORDS: Small Language Model, Selective State Space Models, Biomedical Named Entity Recognition.

I. INTRODUCTION

Biomedical Natural Language Processing (NLP) faces significant challenges due to the computational intensity of models like BioBERT (Lee et al., 2020) and ClinicalBERT (Huang et al., 2020), hindering deployment in low-resource settings. Small Language Models (SLM) emerges as alternatives by balancing performance and efficiency, demonstrating utility in specialized tasks like biomedical Named Entity Recognition (NER) (Marengo et al., 2024). Conventional Transformer-based implementations inherit critical deficiencies such as quadratic complexity, rigid processing, and hardware dependence (Gu & Dao, 2024). This work bridges that gap using an SSMs-based SLM (Mamba architecture), overcoming these barriers. We focus on adapting this model through fine-tuning for biomedical NER, demonstrating that our compact model can match the accuracy of larger counterparts without sacrificing accessibility.

II. MATERIAL & METHODS

Our experimental approach is structured in two interconnected phases, which are shown in Figure 1:

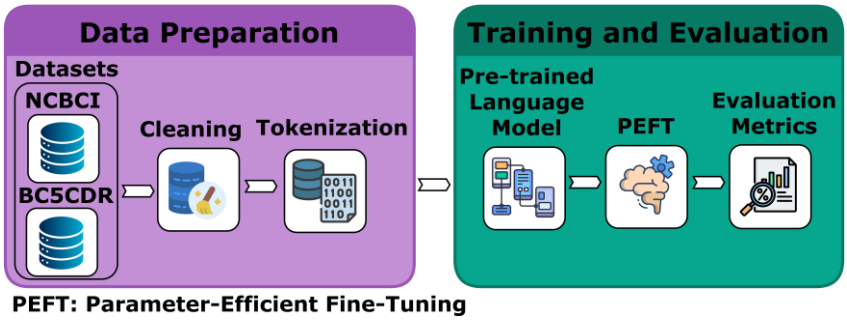

Figure 1. Proposed methodology for the design of our model.

Phase 1 - Data Preparation: We employed two benchmark datasets for medical entity recognition: the BC5-Disease corpus (1,500 PubMed documents with disease annotations) and NCBI-Disease (793 biomedical abstracts with disease entities). These datasets, structured into training, validation, and test subsets, underwent preprocessing, including cleaning to eliminate inconsistencies and noise, followed by tokenization and adaptation to the biomedical NER task to facilitate their use in subsequent phases.

Phase 2 - Training & Evaluation: Subsequently, starting from a general-domain pretrained Mamba-130M base model, we applied Parameter-Efficient Fine-Tuning (PEFT) for 10 epochs with a learning rate of 1×10^{-4} using the AdamW optimizer with progressive warmup (10% of initial steps), a batch size of 4, and sequences of up to 176 tokens. Evaluation addressed performance and efficiency (F1-score, RAM consumption), and robustness, through comparison with four Transformer-based biomedical NER models: BioBERT, ClinicalBERT, GatorTron, and Vanilla-ICL (Brown et al., 2020).

III. RESULTS

Our fine-tuned SSM model (MambaNER) demonstrated an optimal balance between precision and efficiency. It achieved F1-scores of 32.68% on NCBI-Disease and 57.80% on BC5-Disease, respectively. Compared with the best general-domain model, Vanilla-ICL, MambaNER lags only 10.8 percentage points on the NCBI-Disease dataset, despite Vanilla-ICL having approximately ten times more parameters. In contrast, MambaNER surpasses Vanilla-ICL on the BC5-Disease dataset by 13.9 percentage points. Regarding domain-specific models, MambaNER overcomes the best model, BioBERT, by 5.49 and 21.02 percentage points on both NCBI-Disease and BC5-Disease datasets, respectively. Additionally, MambaNER was trained using a GPU with five times less memory, made possible by its selective filtering mechanism that efficiently handles medical terminological variability. Results indicate that Mamba architecture benefits more from PEFT than Transformers-based models, as suggested by (Yoshimura et al., 2024). Table 1 summarizes comparative performance.

Table 1. Performance evaluation between MambaNER and the four baseline models in NCBI-Disease and BC5-Disease datasets.

Models	F1-Score (%)		*Parameters*	*RAM (GB)*
	NCBI-Disease	***BC5-Disease***		
Domain-specific models				
BioBERT	27.19	36.78	1.1×10^8	> 64
ClinicalBERT	8.27	19.96	1.1×10^8	
General-domain models				
GatorTron	35.00	26.97	3.45×10^8	> 64
Vanilla ICL	43.48	43.90	1.76×10^9	
MambaNER	**32.68**	**57.80**	**1.3×10^8**	**12**

IV. CONCLUSION

This work validates that SSMs resolve the precision-efficiency trade-off in biomedical NLP, enabling robust clinical performance with minimal computational resources. This innovation supports deployment on mobile devices and peripheral healthcare systems.

ACKNOWLEDGMENT

Amed Claveria Mendez thanks CONACyT (Consejo Nacional de Humanidades, Ciencias y Tecnologías) for the scholarship granted (CVU:1184252) to pursue his postgraduate studies at the University of Guadalajara.

REFERENCES

Brown et al., (2020). *Language Models are Few-Shot Learners* (No. arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Gu, A., & Dao, T. (2024). *Mamba: Linear-Time Sequence Modeling with Selective State Spaces* (No. arXiv:2312.00752). arXiv. https://doi.org/10.48550/arXiv.2312.00752

Huang et al., (2020). *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission* (No. arXiv:1904.05342). arXiv. https://doi.org/10.48550/arXiv.1904.05342

Lee et al., (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Marengo et al., (2024). *Benchmarking LLMs and SLMs for patient reported outcomes*. https://doi.org/10.48550/ARXIV.2412.16291

Yoshimura et al., (2024). *MambaPEFT: Exploring Parameter-Efficient Fine-Tuning for Mamba* (No. arXiv:2411.03855). arXiv. https://doi.org/10.48550/arXiv.2411.03855