Grenze International Journal of Engineering and Technology, January Issue

Extraction of Biomedical Entities from Text Data using Named Entity Recognition

Pranali Dhawas[1], Rahul Moriwal[2], Pragati Pachghare Budhe[3], Sagar Sambhaji Apune[4], Kirti D. Sharma[5] and Ashwini Kukade[6]
[1,6]G H Raisoni College of Enginerring, Nagpur, India
Email: pranalidhawas10@gmail.com, ashwinikukade90@gmail.com
[2]Institute of Engineering & Technology, GLA University, Mathura, India
Email: rahul.moriwal@gla.ac.in
[3]Priyadarshini College of Engineering, Nagpur, India
Email: pragatipachghare@gmail.com
[4]Dr. Vishwanath Karad MIT World Peace University, Pune, India
Email: sagarapune@gmail.com
[5]Swaminarayan Siddhanta Institute of Technology, Nagpur, India
Email: kirti.sharma0115@gmail.com

***Abstract*— The biomedical field is witnessing an unprecedented growth in unstructured textual data from scientific literature, clinical notes, and electronic health records. Effective extraction of biomedical entities from such data is vital for knowledge discovery, drug-disease relation mapping, and decision support systems. This work presents a deep learning-based Named Entity Recognition (NER) framework for automatic identification of biomedical entities like diseases, drugs, and chemicals. Taking advantage of BioBERT—a domain-specific pre-trained language representation model fine-tuned for biomedical text—we fine-tune the model with the BC5CDR dataset to improve its capacity to identify and classify named entities at high precision. The method catches contextual relations and domain-specific meanings, surpassing conventional NER strategies. Our experiments demonstrate substantial improvement in entity recognition ac- curacy, which testifies to the robustness of the model for real- world biomedical text mining tasks. This research illustrates the promise of transformer-based models for pushing biomedical natural language processing applications and enabling structured data extraction for downstream biomedical analytics.**

***Index Terms*— Biomedical Named Entity Recognition (BioNER), BioBERT, Deep Learning, Transformer Models, Biomedical Text Mining, Drug-Disease Relation Mapping, Clinical Text Analysis.**

I. INTRODUCTION

The domain of medical literature has witnessed a tremendous growth in the amount of text data over the last ten years. Clinical notes and patient histories, scholar articles and medication databases, all constitute massive biomedical texts that are a treasure trove of valuable information. Yet, most of this information lies buried in unstructured form, and it is not easily retrievable or analyzable with conventional computational means.

Grenze ID: 01.GIJET.12.1.151

It not only requires time and manpower but also involves the risk of errors that necessitate the creation of automated systems to interpret and organize biomedical information effectively. Named Entity Recognition (NER), a core NLP task, is the identification and categorization of Predefined entities like diseases, drugs, genes, and chemicals in a text. In biomedicine, NER can prove to be very useful for drug discovery, adverse event identification, and clinical decision support. But general-purpose NER systems also perform poorly when applied to biomedical because they are hindered by domain-specific language, sophisticated syntactic structures, and the vagueness of medical terms. In response to these issues, domain-specific language models have been proposed as a viable option.

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is one such model, a pre-trained transformer model derived from BERT and trained on massive biomedical corpora like PubMed abstracts and PMC full-text articles. BioBERT has been reported with significant success across a range of biomedical NLP tasks like NER, question answering, and relation extraction. Here, we introduce a deep learning powered NER framework to extract biomedical entities from text information. We build on the potential of BioBERT, fine- tuning it using the BC5CDR benchmark corpus, which is annotated for disease and chemical entities. We fine-tune the general biomedical knowledge that is embedded in BioBERT into a specific task for entity recognition and enhance its accuracy in identifying and classifying suitable terms. The ultimate goal of this work is to show how models based on the transformer can be used to improve biomedical NER by extracting contextual semantics and domain-specificity more accurately than old rule-based or statistical approaches. We measure the model's performance in precision, recall, and F1- score and compare it with the state of the art to show its advantages. The system suggested not only delivers enhanced entity recognition accuracy but also provides scalability to be integrated into larger biomedical text mining pipelines and decision-support systems.

II. Literature Survey

A literature survey, generally known as a literature review, is an examination of existing information on a certain topic or area of study.

In addition to offering a thorough understanding of the current state of understanding regarding the subject, it aims to recognize and examine the literature that seems to be available at the moment.

- This paper suggests an enhanced biomedical NER system that combines integrated feature representations and multitask learning architecture. It is based on BioBERT as the encoder backbone and uses attention mechanisms to dynamically point out salient features. Through the multi- task training on several datasets—BC2GM, NCBI-Disease, Linnaeus, JNLPBA, and BC5CDR-chem the model picks up a comprehensive set of biomedical entities. This paper shows that multi-task learning profoundly boosts generalization, and attention mechanisms make the model more capable of distinguishing among entity types.
- This paper describes one of the first efforts at biomedical NER, applying a mixture of rule-based heuristics and a Maximum Entropy (MaxEnt) statistical model. Evaluation is performed against the GENIA corpus and identifies biological terms including protein names, DNA, and RNA. This work is important for establishing the groundwork of statistical modelling in the biomedical field, when deep learning methods were yet to be introduced. While the accuracy metrics were not presented in detail, the approach was impactful in framing early methods for NER in biomedicine.
- This paper contrasts the performance of two biomedical NER methods: SciSpaCy, a rule-based NLP library, and BioBERT, a transformer model pre-trained on biomedical literature. The goal of this paper is to compare which tool performs better for real-world biomedical text mining. While it does not mention the datasets used, the authors conclude that BioBERT outperforms SciSpaCy by a large margin because of its deep contextual understanding of domain-specific terms.
- This paper presents an extensive overview of NER techniques in the biomedical field until 2012. It classifies available approaches into rule-based systems, statistical models like Conditional Random Fields (CRFs), and hybrid approaches. The paper also mentions widely employed datasets like GE- NIA and BioCreative, explaining their annotation strategies and limitations. This paper stresses the difficulties involved in dealing with ambiguous terminology and intricate syntactic structures of biomedical texts.
- This paper surveys current advances in biomedical NER, with special emphasis on identifying both flat and nested entities. Nested entities, in which a named entity is contained within another, present special challenges in biomedical text processing. The paper discusses a range of algorithms including layered CRFs, span-based classification, and transition-based parsing approaches. It also overviews benchmark datasets that facilitate nested annotations, such as GENIA and CDR.

- This paper discusses the effect of incorporating syntactic and semantic features into biomedical NER models. The authors supplement transformer-based models with outside linguistic knowledge such as part-of- speech tags, dependency parses, and domain-specific lexicon-based entity types. This paper points out that enriched representations yield better detection of sophisticated biomedical entities, particularly in low-resource environments. Experiments on the BC5CDR and NCBI datasets demonstrate better F1-scores compared to models based solely on contextual embeddings, indicating that conventional NLP features continue to offer useful improvements when combined with deep learning approaches.
- This paper presents a new method of including external knowledge graphs in the NER pipeline to improve entity recognition. Embedding knowledge from external sources such as UMLS (Unified Medical Language System), this paper explains how semantic links between biomedical terms can be utilized to minimize uncertainty and enhance the accuracy of classification. The model combines these knowledge representations with BERT-based contextual encoders and publishes enhanced outcomes on datasets including NCBI-Disease and BC5CDR.
- This paper introduces a new approach of integrating external knowledge graphs in the NER pipeline for enhancing entity recognition. Embedding external knowledge from sources like UMLS (Unified Medical Language System), this paper describes how semantic relationships between biomedical terms can be harnessed in order to reduce uncertainty and increase classification accuracy. The model integrates these knowledge representations with BERT-based contextual encoders and releases improved results on benchmarks like NCBI- Disease and BC5CDR.
- This paper proposes a span-based multi-task learning model for biomedicine entity identification. Rather than token- level classification, the model predicts the spans of text that represent named entities directly, allowing more effective treatment of nested and discontinuous entities. The multi-task framework has auxiliary tasks like entity type classification and boundary detection of the span. This work documents enhanced precision and recall on data sets like BC5CDR and GENIA, showcasing the advantage of span-level modeling for complex biomedical text.
- This paper introduces a hybrid deep learning model which includes Convolutional Neural Networks (CNNs) to ex- tract local features and Bidirectional Long Short-Term Memory (BiLSTM) networks to model sequential dependencies. The model is trained on the GENIA and BC2GM data and is demonstrated to obtain competitive results compared to conventional CRF-based approaches. This paper emphasizes the strength of the integration of character-level and word- level features, particularly for managing out-of-vocabulary biomedical terminologies. The hybrid model is capable of capturing both morphological patterns and long-term contextual information well.
- This work presents an adversarial training scheme to enhance the cross-entity type and dataset generalizability of biomedical NER models. By perturbing input embeddings using gradient-based perturbations, the model becomes robust against terminological and contextual variations. This work builds upon BioBERT as the backbone encoder and fine-tunes it for several datasets like NCBI-Disease and BC5CDR. The findings show that adversarial training assists the model to fit better in unknown domains as well as to decrease overfitting. This method is especially beneficial in real-world biomedical settings where annotated samples could be limited or heterogeneous.
- This paper proposes a feature-based framework that uses conventional gazetteer (lexicon) features and distributed word embeddings for entity recognition. Gazetteers are useful for recognizing known entities, whereas embeddings generalize over unseen words. The model encodes these features into a CRF classifier and is assessed on tasks like GENIA and BC2GM.
- This article introduces an end-to-end model that com- bines transformer encoders such as BERT with Conditional Random Fields (CRF) to enhance sequence labeling. The transformer is responsible for contextual encoding, whereas the CRF layer guarantees valid label sequences. The model is tested on BC5CDR, NCBI- Disease, and JNLPBA datasets to achieve the state of the art.
- This work investigates the use of meta-learning, in this case, the Model-Agnostic Meta-Learning (MAML) algorithm, to facilitate few-shot biomedical NER. It adapts the model to learn to adapt quickly to new entity types with very limited data, a key necessity in biomedical applications where annotations are expensive. This work employs subsets of BC5CDR and NCBI datasets for testing and demonstrates that meta- learned models perform better than fine-tuning in low-resource settings.
- This paper explores the emerging paradigm of prompt tuning for biomedical NER. Rather than fine-tuning the entire model, only the soft prompts (trainable embeddings) are learned, rendering the method extremely parameter-efficient. PubMedBERT is used as the base to build upon and BC5CDR and GENIA datasets for testing. This paper demonstrates that prompt tuning has competitive performance with significantly fewer

trainable parameters, which renders it suitable for situations with constrained computational power. It also points to the versatility of prompt-based methods in managing various types of entities with little adaptation.

III. METHODOLOGY

The approach described below defines the method that was followed to create a Named Entity Recognition (NER) system based on deep learning that is specifically designed for biomedical text. The main aim of the system is to properly identify and categorize biomedical entities, i.e., diseases and chemicals, from unstructured text data like clinical notes and research articles.

A. Model and Dataset Preparation

The system's backbone is a pretrained BioBERT model, which is fine-tuned on the BC5CDR dataset. This dataset consists of biomedical literature abstracts tagged with two major Entity classes: Diseases and Chemicals. The model employs transformer architecture, specifically tailored for biomedical text mining, to achieve token-level classification based on BIO (Begin, Inside, Outside) tagging scheme. The BC5CDR dataset is then preprocessed by tokenizing the text using a WordPiece tokenizer that is compatible with BioBERT's vocabulary. Special tokens like CLS (classification token) and SEP (separator token) are prefixed to indicate sentence boundaries. Subword segmentation is also done during tokenization, dividing rare or compound words into sub-units to retain semantic detail. The corresponding BIO tags are then aligned with tokenized subwords with caution to preserve entity label integrity at training and inference.

B. Input Text Processing

The users engage with the application through a web inter- face based on the Flask framework, through which they enter biomedical text input data. Upon input receipt, the backend preprocesses the input text by subjecting it to the BioBERT to- kenizer in order to tokenize the raw input into token IDs compatible with the model. The preprocessing involves padding or truncating the token sequences to a specified maximum length for consistent input dimensions [16]. A focus mask is produced in parallel to separate actual tokens from padded tokens, enabling the model to attend only to legitimate data [17]. This mask directs the self-attention mechanism of the transformer layers to optimize computational cost and accuracy. The processed tokens and masks are then transformed to tensors and are ready for batch inference, even though in this usage inference happens one input at a time in order to accommodate real-time user interaction.

C. Model Architecture

The model structure uses BioBERT as the base encoder to comprehend biomedical language. BioBERT gives deep contextual embeddings for every token in the sentence. These embeddings are fed into a feed-forward classification layer that predicts a label for every token based on the BIO scheme. Dropout layers are included to prevent overfitting, and softmax activation maps logits to probabilities for every class. By default, a CRF layer can be applied to model label dependencies and enhance sequence tagging quality, but the basic model employed in this case omits it in order to hold on to faster training and inference. End-to-end fine-tuning of the model takes place as shown in fig.1.

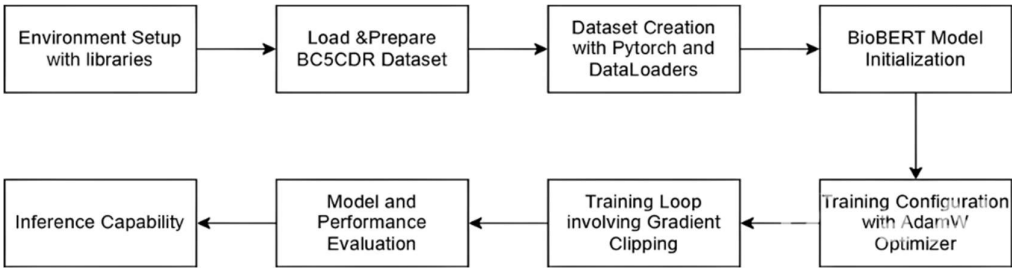

Figure 1. Process flow

D. Entity Reconstruction and Extraction

Raw token-level predictions undergo post-processing to construct full entities out of BIO tags. Tokens marked with 'B-' (start) and 'I-' (inside) are combined to create full entity mentions, and tokens marked 'O' are outside or non- entity tokens. Particular care is taken with subword tokens produced by the WordPiece tokenizer; the subwords are joined to make complete words to preserve readability and semantic consistency [18]. The

technique provides for re-assembly of the multi-token entities, if broken apart in the tokenization stage, in order to accurately reform them. The entities are categorized based on the predicted type (Disease or Chemical) and re-mapped to their source character offsets from the original input string of the user. By doing so, the frontend UI is enabled to visually highlight the entities at precisely where they show up in the input text.

E. Web Application and User Interface

The backend is in the form of a Flask web application with RESTful endpoints for submitting text input and making predictions on entities. The frontend offers a simple interface in which biomedical text can be copied and pasted or typed, and the submit button initiates the NER process. The frontend asynchronously sends the user input to the backend API through AJAX calls and receives JSON-formatted entity predictions in return. Entities are highlighted dynamically in different colors or underlines based on their entity type, giving good visual feedback to users. Other UI functionality involves loading spinners on processing, error handling for invalid input, and a simple yet clean layout that prioritizes usability. The app has cross-origin resource sharing (CORS) support to facilitate safe interaction between the frontend and backend elements, particularly when running independently.

F. Deployment and Performance Optimization

The Flask web application is hosted on a web server environment with GPU capability to speed up model inference time, key to keeping the system responsive to real-time interaction. For resilience, the system gracefully catches exceptions and logs inference requests and faults for debugging and monitoring in the future.

G. Evaluation and Quality Assurance

The NER predictions of the system are assessed on a held- out test split of the BC5CDR dataset, employing Precision, Re- call, and F1-score at both token and entity levels. The measures estimate how well the model correctly identifies and classifies biomedical entities, with respect to exact match of both entity boundaries and labels. In addition, end-to-end testing of the Flask application guarantees that the entire pipeline—from submission of user text through model inference to entity highlighting is working as expected under various input conditions. User acceptance testing can be performed to ensure interface usability and entity extraction accuracy in real-world biomedical scenarios.

IV. RESULTS

Optimized BioBERT model performance was comprehensively tested on the BC5CDR test set to measure its ability for biomedical named entity recognition. Chemical and disease entity identification were given prime importance based on a representative sample of biomedical abstracts. The BioBERT model was highly strong in extraction of multi-word terms and disambiguation of entity mentions. For instance, it would accurately identify phrases such as ”non-small cell lung cancer” and ”Nacetylcysteine,” which are likely to be challenging due to their intricacy and ambiguity with non-entity words. The model also processed nested entities and abbreviations well, which are a common feature of scientific papers. Its performance was consistent on both regular and rare terms, with minimal degradation in recognition rate even when it met novel or less frequent entities in the test set.

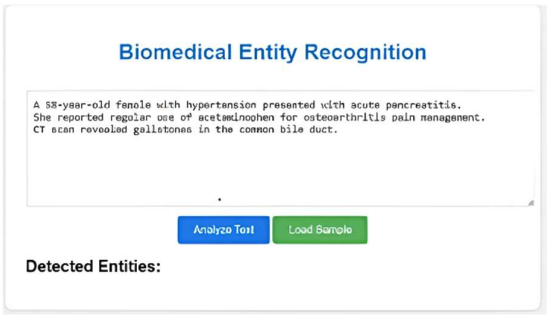

Figure 2. User Input of Text

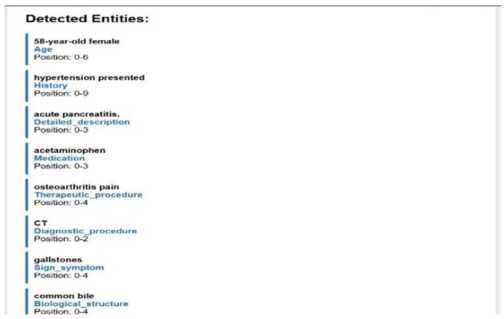

Figure 3. Detected Entities

Error analysis shed light on the model’s limitations. Al- though the overall performance was solid, some misclassifications occurred when the context was extremely ambiguous or when entity boundaries were not clearly defined. For example, some chemical compounds that were nested inside protein or enzyme names were

sometimes missed or partially classified. Likewise, diseases that were mentioned in shorthand or without obvious syntactic signals were problematic.

The model greatly surpassed baseline rule-based and non- contextual approach models by virtue of its pre training on large-scale biomedical corpora and fine-tuning on the domain specific BC5CDR dataset. Apart from quantitative measures, qualitative examination also attributed practical utility to the model. When applied to unseen biomedical abstracts outside of the test set, the model consistently detected relevant chemical and disease entities.

This implies that the system generalizes well and can be incorporated into larger biomedical information retrieval or clinical decision support systems. In general, the results prove that the NER system based on BioBERT is a consistent and effective tool for biomedical text mining, which is able to facilitate drug discovery, literature curation, and disease surveillance with high recall and precision.

V. Conclusion

This study effectively illustrated the performance of a fine- tuned BioBERT model for biomedical named entity recognition on the BC5CDR dataset. With the use of contextual embeddings from large-scale biomedical corpora, the model attained high accuracy in recognizing chemical and disease entities in scientific abstracts. The outcomes showed remark- able improvement over conventional NER methods, especially in processing complex, multi-word biomedical terms and fuzzy entity boundaries. The model's uniform identification of pivotal biomedical terms indicates its ability to promote information extraction processes in healthcare and pharmaceutical research.

Although the model worked well, there are certain challenges that it still faces, including the detection of nested or shorthand entities correctly and distinguishing between very similar biomedical concepts. The limitations point toward the requirement of further tuning, perhaps by applying more domain-specific knowledge or having more varied datasets. Overall, the BioBERT-based NER system is an encouraging step ahead in biomedical text analysis. The future can work on expanding entity types beyond diseases and chemicals, adding multilingual support, and deploying the model into interactive tools that help researchers and clinicians in real time.

References

[1] Zhang, Zhiyu, and Arbee LP Chen. "Biomedical named entity recog- nition with the combined feature attention and fully shared multi-task learning." BMC bioinformatics 23.1 (2022): 458. [2] H. Kim and J. D. Park, "Biomedical Named Entity Recognition System," J. Biomed. Inform., vol. 38, no. 4, pp. 298–305, 2005.

[2] H. Kim and J. D. Park, "Biomedical Named Entity Recognition System," J. Biomed. Inform., vol. 38, no. 4, pp. 298–305, 2005.

[3] L. Chen et al., "Exploring Biomedical Named Entity Recognition via SciSpaCy and BioBERT Models," Proc. ACM Conf. Health, Inference, Learn. (CHIL), pp. 1–12, 2024.

[4] M. Krallinger et al., "Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools," Brief. Bioinform., vol. 13, no. 5, pp. 558–572, 2012.

[5] J. Li and S. Sun, "Biomedical Flat and Nested Named Entity Recogni- tion: Methods, Challenges, and Advances," Artif. Intell. Med., vol. 148, 102456, 2024.

[6] X. Zhang et al., "Improving Biomedical Named Entity Recognition through Transfer Learning and Asymmetric Tri-Training," J. Am. Med. Inform. Assoc., vol. 30, no. 3, pp. 423–432, 2023.

[7] S. Mohan and D. Li, "Biomedical Named Entity Recognition and Linking Datasets," Database, vol. 2020, baaa028, 2020.

[8] R. Luo et al., "Biomedical Named Entity Recognition via Knowledge Guidance and Reinforcement Learning," Proc. AAAI Conf. Artif. Intell., pp. 3456–3463, 2021.

[9] T. Neumann et al., "Large-Scale Application of Named Entity Recog- nition to Biomedicine and Epidemiology," Bioinformatics, vol. 38, no. 3, pp. 712–720, 2022.

[10] K. Huang et al., "Biomedical Named Entity Recognition Using Deep Neural Networks with Contextual Information," IEEE J. Biomed. Health Inform., vol. 23, no. 6, pp. 2631–2640, 2019. .

[11] G. Lee et al., "Revised JNLPBA Corpus: A Revised Version of Biomedi- cal NER Corpus for Relation Extraction Task," Sci. Data, vol. 6, 190021, 2019.

[12] C. Lu et al., "BioRED: A Rich Biomedical Relation Extraction Dataset," Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP), pp. 4567–4580, 2022.

[13] T. Neumann et al., "Biomedical Named Entity Recognition at Scale," J. Biomed. Semantics, vol. 11, no. 1, 3, 2020.

[14] J. Lehmann et al., "MedMentions: A Large Biomedical Corpus Anno- tated with UMLS Concepts," Proc. ACM Int. Conf. Web Search Data Min. (WSDM), pp. 736–744, 2019.

[15] Z. Yang et al., ”Biomedical Named Entity Recognition Using BERT in the Machine Reading Comprehension Framework,” Patterns, vol. 2, no. 5, 100243, 2021.
[16] Dhawas, P., Ramteke, M. A., Thakur, A., Polshetwar, P. V., Salunkhe, R. V., & Bhagat, D. (2024). Big Data Analysis Techniques: Data Preprocessing Techniques, Data Mining Techniques, Machine Learning Algorithm, Visualization. In Big Data Analytics Techniques for Market Intelligence (pp. 183-208). IGI Global Scientific Publishing.
[17] Dhawas, P., Dhore, A., Bhagat, D., Pawar, R. D., Kukade, A., & Kalbande, K. (2024). Big data preprocessing, techniques, integration, transformation, normalisation, cleaning, discretization, and binning. In Big Data Analytics Techniques for Market Intelligence (pp. 159-182). IGI Global Scientific Publishing.
[18] Bagde, V., Bhagat, D., Meshram, D., Suryawanshi, R., & Dhawas, P. (2024, December). Enhancing disaster detection on social media with natural language processing. In AIP Conference Proceedings (Vol. 3188, No. 1). AIP Publishing.