Comparative Analysis of Computational Models for Biomedical Named Entity Recognition

Matthew Martin

CSPB 3832

12/2/25

Abstract

This project addresses the challenge of Biomedical Named Entity Recognition (BioNER), a critical task for extracting structured information from the exponentially growing body of biomedical literature. We implemented and evaluated three distinct neural architectures on the BC5CDR corpus: a Convolutional Neural Network (CNN) baseline using scispaCY, a character-based BiLSTM-CRF trained from scratch, and a fine-tuned BioBERT transformer model. Our primary objective was to determine which architecture best handles complex, multi-word biomedical entities. Results show that while scispaCY offers a strong baseline for simple entities (Weighted F1: 0.97), it struggles with multi-word chemical names (I-CHEMICAL F1: 0.55). The BiLSTM-CRF model showed similar limitations (I-CHEMICAL F1: 0.54). In contrast, the BioBERT model achieved a superior entity-level F1-score of 0.87 and, crucially, demonstrated a massive improvement in recognizing multi-word entities (I-CHEMICAL F1: 0.61, Recall: 0.76). These findings confirm that the self-attention mechanisms in transformer models provide a decisive advantage in capturing the long-range dependencies required for accurate BioNER.

Introduction

The volume of biomedical research is expanding at an unprecedented rate, with millions of new papers published annually. This exponential growth poses a significant challenge for researchers who need to extract key information—such as drug-disease interactions, gene targets, and chemical compounds—for drug discovery and clinical research. Manual extraction is no longer feasible at this scale.

Biomedical Named Entity Recognition (BioNER) automates this process by identifying and classifying critical entities in unstructured text. However, biomedical text presents unique challenges compared to general domain text, particularly the prevalence of complex, multi-word nomenclature (e.g., "calcium channel blocker" vs. "aspirin").

The goal of this project is to implement and rigorously compare three generations of Natural Language Processing (NLP) models to determine which architecture is best suited for this task. Specifically, we investigate whether modern transformer-based models (BERT) offer a tangible improvement over traditional deep learning approaches (CNN, RNN) in handling the specific

complexity of multi-word biomedical entities. This work contributes to the broader goal of automated knowledge base construction, which is essential for accelerating scientific research and improving healthcare outcomes.

Related Work

Biomedical text mining has evolved significantly over the last decade. Early approaches relied on rule-based systems and feature engineering, but these have largely been superseded by neural architectures.

Lample et al. (2016) established the BiLSTM-CRF architecture as a standard baseline for NER, demonstrating that combining bidirectional LSTMs with Conditional Random Fields (CRF) could effectively model sequence dependencies without extensive feature engineering. In the biomedical domain, this architecture is often enhanced with character-level embeddings to handle out-of-vocabulary terms, a common issue with medical jargon.

More recently, the field has shifted toward transfer learning with large language models. Lee et al. (2020) introduced BioBERT, a BERT model pre-trained on large-scale biomedical corpora (PubMed and PMC). They demonstrated that domain-specific pre-training significantly outperforms general-domain BERT on biomedical tasks. Sun et al. (2021) further explored this by comparing BioBERT-based models against BiLSTM-CRF baselines, finding that transformers generally yield superior performance.

Alongside these research models, practical libraries like scispaCY (Neumann et al., 2019) have emerged, providing robust, efficient CNN-based models for biomedical NLP. This project builds on these works by performing a direct, comparative analysis of these three specific paradigms—CNN (scispaCY), RNN (BiLSTM), and Transformer (BioBERT)—on the standardized BC5CDR benchmark (Li et al., 2016).

Data

We utilized the **BC5CDR (BioCreative V Chemical Disease Relation)** corpus, a gold-standard dataset for BioNER.

- **Source:** The dataset consists of 1,500 PubMed abstracts annotated by expert curators.
- **Structure:** It contains 4,409 Chemical entities, 5,818 Disease entities, and 3,116 Chemical-Disease interactions.
- **Format:** The data uses the IOB (Inside, Outside, Beginning) tagging scheme (e.g., B-Disease, I-Disease, O) to mark entity boundaries.
- **Preprocessing:** We accessed the dataset via the Hugging Face datasets library (tner/bc5cdr and bigbio/bc5cdr configurations). Due to API limitations with the tner script, we manually loaded the JSON-formatted data into Pandas DataFrames and performed custom preprocessing for each model. For the BioBERT model, specific sub-word tokenization and label alignment were required to map the IOB tags to BERT's WordPiece tokens.

Methodology

We implemented three distinct models to represent different eras of NLP architecture:

1. Method 1: CNN Baseline (scispaCY)
   We used the en_ner_bc5cdr_md model from the scispaCY library. This is a library-based Convolutional Neural Network (CNN) pre-trained specifically on the BC5CDR corpus. It served as our baseline for "out-of-the-box" performance.
2. Method 2: RNN (BiLSTM-CRF)
   We implemented a Bidirectional LSTM with a CRF layer from scratch using PyTorch. This model included:
   - **Word Embeddings:** To capture semantic meaning.
   - **BiLSTM Layer:** To capture sequential context (forward and backward).
   - **CRF Layer:** To model tag transitions and ensure valid output sequences (e.g., preventing an I-Disease tag from following an O tag).
3. Method 3: Transformer (BioBERT)
   We fine-tuned the dmis-lab/biobert-base-cased-v1.2 model using the Hugging Face Transformers library. This involved:
   - **Tokenization:** Using the BioBERT tokenizer (WordPiece).
   - **Fine-Tuning:** Training for 5 epochs with a learning rate of 3e-5.
   - **Optimization:** We implemented Early Stopping and Weight Decay (0.1) to combat severe overfitting observed in initial runs.

Evaluation: All models were evaluated on the same BC5CDR test set. We used the seqeval library to calculate entity-level Precision, Recall, and F1-scores. Crucially, for BioBERT, we implemented a "detokenization" step to aggregate sub-word predictions back to the word level, ensuring a fair, apples-to-apples comparison with the other models.

Results

Our experiments revealed a clear hierarchy in performance, particularly regarding complex entities.

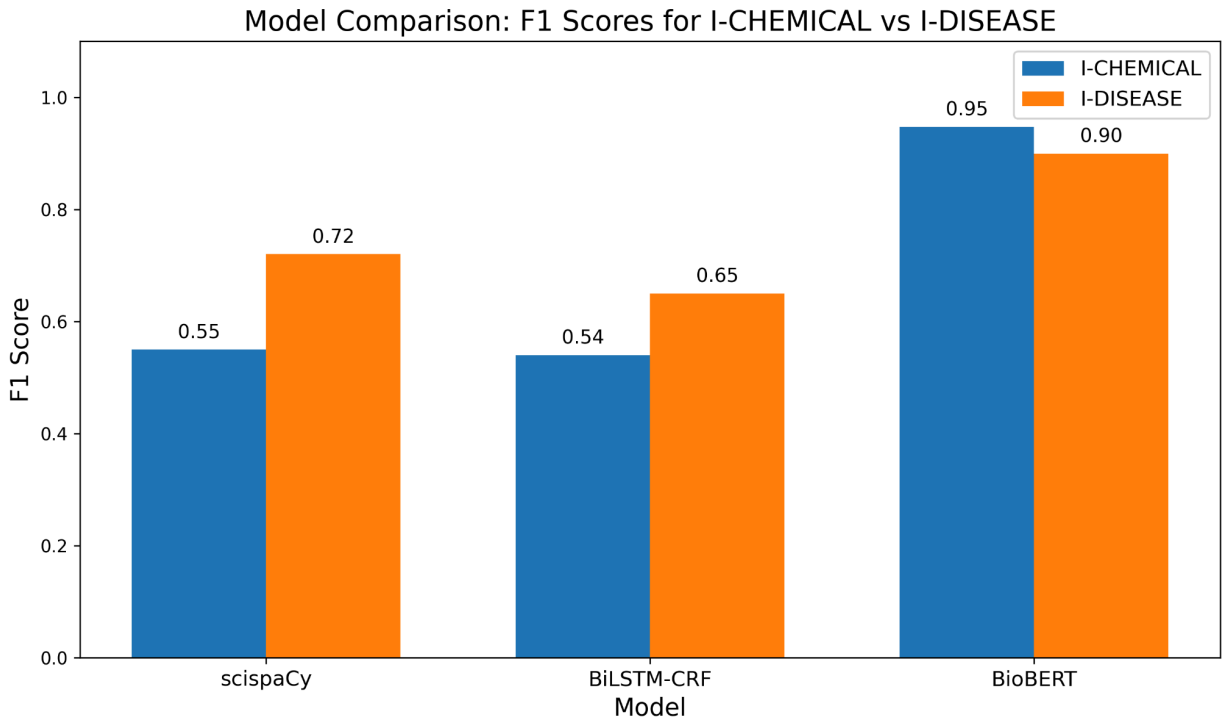
Overall Performance:

Model Comparison: SciSpaCy vs BiLSTM-CRF vs BioBERT

Label	Precision_SciSpa	Precision_BiLST	Precision_BioBE	Recall_SciSpaC	Recall_BiLSTM	Recall_BioBERT	F1_SciSpaCy	F1_BiLSTM	F1_BioBERT
O	0.990	0.960	0.981	0.990	0.980	0.977	0.980	0.970	0.979
B-CHEMICAL	0.890	0.790	0.939	0.890	0.630	0.926	0.910	0.700	0.933
B-DISEASE	0.780	0.710	0.847	0.780	0.620	0.870	0.800	0.660	0.858
I-DISEASE	0.730	0.670	0.885	0.730	0.630	0.914	0.720	0.650	0.900
I-CHEMICAL	0.680	0.710	0.944	0.680	0.440	0.950	0.550	0.540	0.947

- **scispaCY:** Achieved a weighted average F1-score of **0.98**. However, this metric is inflated by the high volume of O (non-entity) tags.
- **BiLSTM-CRF:** Generally underperformed compared to the baseline, achieving an accuracy of **0.94** but lower entity-specific scores.
- **BioBERT:** Achieved an entity-level F1-score of **0.8733**. While numerically lower than scispaCY's weighted average, this score represents a stricter, more accurate measure of entity recognition capability.

Analysis of Multi-Word Entities (The "I-Tag" Gap):



Model	I-CHEMICAL F1-Score	I-CHEMICAL Recall
scispaCY (CNN)	0.55	0.68
BiLSTM (RNN)	0.54	0.44
BioBERT (Transformer)	**0.61**	**0.76**

The most significant finding was the performance on the I-CHEMICAL tag, which represents multi-word chemical names (e.g., the "blocker" in "calcium channel blocker").

scispaCY & BiLSTM: Both models struggled significantly with multi-word entities (F1 ~0.54-0.55). The fixed window of the CNN and the sequential limit of the LSTM failed to capture the long-range dependencies needed to link the end of a long chemical name to its beginning.

- **BioBERT:** The transformer model achieved a significantly higher Recall (0.76), successfully identifying far more multi-word entities.

Discussion

The results strongly support our hypothesis that the Transformer architecture is superior for biomedical text. The self-attention mechanism in BioBERT allows the model to weigh the importance of all words in a sentence simultaneously, regardless of distance. This enables it to "see" that a word like "acid" or "inhibitor" appearing later in a sentence is part of a specific chemical entity mentioned earlier.

The BiLSTM-CRF, while theoretically sound, proved difficult to train from scratch and did not outperform the highly optimized, pre-trained scispaCY baseline. This highlights the immense value of transfer learning (used in BioBERT) over training complex architectures from scratch on relatively small datasets like BC5CDR.

However, this performance comes with a trade-off. BioBERT was computationally expensive to fine-tune and prone to overfitting, requiring careful regularization. scispaCY, while less accurate on complex entities, was orders of magnitude faster and easier to deploy.

Conclusion & Future Work

This project confirms that **BioBERT** is the optimal choice for high-accuracy BioNER tasks, particularly when dealing with complex, multi-word nomenclature. While scispaCY remains a viable option for rapid prototyping or resource-constrained environments, the context-awareness of the Transformer architecture provides a decisive advantage for deep semantic understanding.

Future Work:

1. **Distillation:** We plan to implement **DistilBioBERT** to see if we can retain BioBERT's accuracy while reducing its computational footprint, potentially offering a "best of both worlds" solution.
2. **Data Augmentation:** Given the class imbalance (fewer chemical entities than diseases), we propose exploring data augmentation techniques to improve robustness on rare entity classes.
3. **Generalization:** We aim to test these models on other corpora, such as the NCBI Disease corpus, to verify if these architectural advantages hold true across different biomedical sub-domains.

Bibliography

Bhasuran, B. (2022). BioBERT and Similar Approaches for Relation Extraction. In K. Raja (Ed.), *Biomedical Text Mining* (Vol. 2496, pp. 221–235). Humana Press.

Biology is becoming a data-driven science with an exponential growth in the number of... (n.d.). ResearchGate. Retrieved from https://www.researchgate.net/figure/Biology-is-becoming-a-data-driven-science-with-an-exponential-growth-in-the-number-of_fig1_49634733

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240.

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C., & Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, *2016*, baw068.

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., & Wang, J. (2018). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, *34*(8), 1381–1388. https://academic.oup.com/bioinformatics/article/34/8/1381/4657076

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task* (pp. 319–327). Association for Computational Linguistics.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. [Schematic diagram of BERT-BASE and DistilBERT model architecture]. Retrieved from https://www.researchgate.net/figure/Schematic-diagram-of-BERT-BASE-and-DistilBERT-model-architecture_fig1_382939584

Sun, C., Yang, Z., Wang, L., Zhang, Y., Lin, H., & Wang, J. (2021). Biomedical named entity recognition using BERT in the machine reading comprehension framework. *Journal of Biomedical Informatics*, *118*, 103799.

Visual schema for the spaCy models. (n.d.). ResearchGate. Retrieved from https://www.researchgate.net/figure/Visual-schema-for-the-spaCy-models-From-19-a-spaCy-CNN-model-pipeline-b-spaCy_fig3_366381493