Triple Threat: Benchmarking spaCy, BioBERT, and BERT-Cased on Biomedical Entity Extraction

Dumitrache Flavian
dumitrache.flavian@s.unibuc.ro

Andrei-Virgil Ilie
andrei-virgil.ilie@s.unibuc.ro

Abstract

Named-entity recognition (NER) has an important role in finding the rich, unstructured information within clinical narratives, health records and biomedical literature. Identifying different entities like diseases, symptoms, medications, procedures and anatomical phrases or many other enable many different tasks such and not only: medical decision support and large scale health analytics. In this paper, we will present a comparison between spacy and two transformer based architectures(Vaswani et al., 2017): BioBERT, a domain-adapted language model pretrained on vast biomedical corpora (Lee et al., 2020) and BERT base NER, a ready to use NER model based on BERT (Devlin et al., 2018) trained on the (Sang and Meulder, 2003) dataset. All models were trained and evaluated on a medical NER dataset with diferent document types. We will report precision, recall and F1 scores across both models.

1 Introduction

With the accelerated integration of technology into healthcare, exponentially increased the amount of unstructured clinical data. Electronic Health records, discharge summaries to radiology reports and biomedical manuscripts pose a challenge, but also an opportunity for modern healthcare. These kind of data encapsulate detailed patient histories, research findings, treatment plans, but also complicate large-scale analysis and automated reasoning. One solution to this problem would be transforming raw clinical narratives into strucured knowledge using Named-Entity Recognition. NER is the task of finding and classifying spans of texts into predefined categories. These categories usually include: Name, Location, Currency, Date/Time, but there could also be domain specific entities such as: diseases, medications, symptoms, procedures etc.

Recently, transformer-based pretrained language models have set new benchmarks in many NLP tasks. In particular, BioBERT, which was pretrained on a biomedical corpora, has shown good performance by capturing domain specific semantics. NER Bert was fine-tuned on entity tagged data, tailored for entity extraction. On the other hand, spaCy's NER framework offers a faster, more rule oriented pipeline which uses neural networks and is widely used in the industry for real-time NER tasks. The accuracy on medical text might not match the accuracy of the transformer based modelsIn contrast, spaCy's NER framework offers a faster, rule-augmented pipeline that uses lightweight neural networks and is widely adopted in industry for real-time entity recognition tasks. While it may not match transformer-based models in accuracy on complex biomedical texts, it provides a practical trade-off between speed and performance, especially in resource-constrained settings.

In this paper, we will evaluate spaCy's NER capabilities with the transformer based approaches of BioBERT and NER Bert on an annotated medical NER dataset. By training both models under identical conditions, we will assess their performance with precision, recall and F1 score. Our aim is to find each model's strenghts and weaknesses.

The code for this project can be found at: https://github.com/flaviiG/bioNER-bc5cdr

2 Dataset

We chose bc5cdr corpus (Li et al., 2016) for our experiments. Before training the models, we performed an exploratory data analysis (EDA) on our dataset, to better understand its scale, composition and annotation characteristics.

The corpus contains $\tilde{1}$500 PubMed articles (Li et al., 2016), annotated for two entity types:

- Chemicals = substances with a defined molecular structure used in or resulting from chemical processes.

- Diseases = abnormal conditions impairing normal bodily functions, typically characterized by specific signs and symptoms.

2.1 Training Data

We have decided to train the models on the original train set + validation set. The training data has a total of 19343 entities with the following distribution:

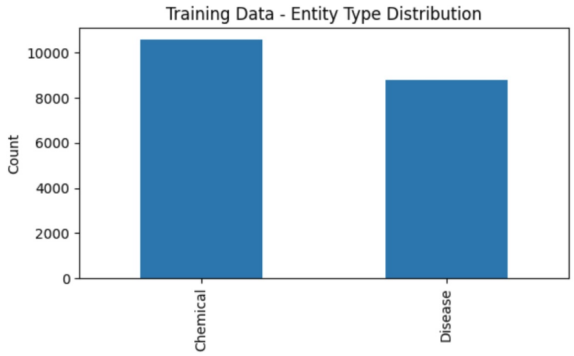

Figure 1: Training Data distribution

The top 10 most frequent entities in the dataset are:

1. cocaine - 143
2. toxicity - 138
3. seizures - 137
4. hypotension - 121
5. pain - 117
6. morphine - 108
7. lithium - 95
8. dopamine - 95
9. hypertension - 92
10. haloperidol - 80

In the training set, the majority of the entities span over only 1 word. The less common entity lengths are: 2 and 3 with some outliers of 4 - 6.

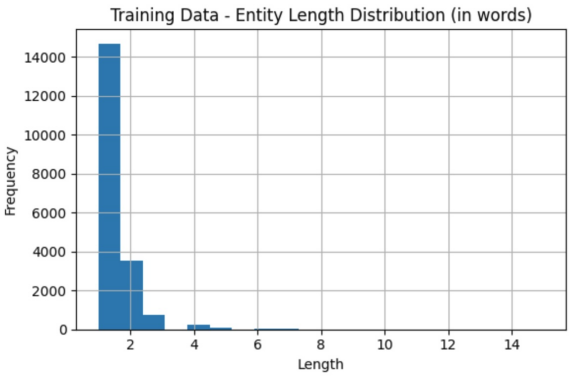

Figure 2: Training Entity Length distribution

2.2 Validation + Test

The validation set was composed on 50% of the original test set, and the rest of the original test set was used as test set.

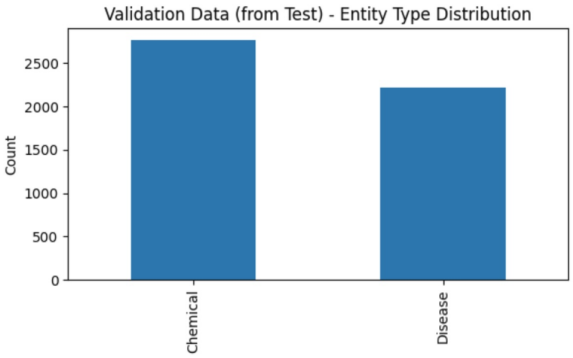

Figure 3: Dev Data distribution

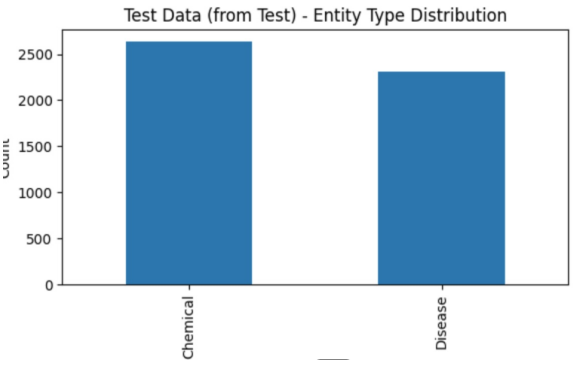

Figure 4: Enter Caption

The most frequent entities in the validation set are:

1. pilocarpine = 49
2. seizures = 44
3. pain = 38

4. doxorubicin = 36
5. propranolol = 33
6. cocaine = 32
7. cardiotoxicity = 31
8. levodopa = 29
9. creatinine = 28
10. bupivacaine = 28

The most frequent entities in the test set are:

1. cocaine = 60
2. seizures = 55
3. hypertension = 47
4. toxicity = 37
5. hypotension = 30
6. doxorubicin = 27
7. warafarin = 24
8. caffeine = 24
9. carditoxicity = 23
10. cisplatin = 22

The dataset is a little bit unbalanced, chemicals having more examples than the diseases entity.

3 Related Work

3.1 spaCy

spaCy (Honnibal et al., 2020) is an open-source NLP library offering fast, efficient pipelines for tokenization, POS tagging, dependency parsing and NER. Even though spaCy's pretrained models are more lightweight than large transformer architectures, they still achieve competitive results on general-domain texts and can easily be fine-tuned on cutom data. This makes spaCy a really useful tool for real-time applications and environments with limited computational resources.

3.2 BioBERT

BioBERT (Lee et al., 2020) is a variant of Google's BERT (Devlin et al., 2019) which was pre-trained on a large biomedical corpora of 4.5 billion words from PubMed abstracts and full-text articles which were published in PubMed Central. After that it was fine-tuned on downstream biomedical NLP tasks. The model outperforms the original BERT-Base architecture on a variety of benchmarks, including NER, relation extraction and question answering.

3.3 BERT Cased

BERT Cased (Devlin et al., 2019) refers to the original BERT (Bidirectional Encoder Representations from Transformers) model that preserves the case of words during both pre-training and downstream fine-tuning. Unlike its "uncased" counterpart—which lowercases all input text and strips accent markers—BERT Cased retains capitalization and diacritics, making it particularly suitable for tasks where case information carries important semantic or syntactic cues, such as named entity recognition (NER) and part-of-speech tagging. The cased variant is trained on the English Wikipedia and BookCorpus using a WordPiece tokenizer with a 30,000-token vocabulary. Although BERT Cased is often slightly outperformed by domain-adapted models in specialized settings, it serves as a strong general-purpose baseline, especially for languages and applications where casing impacts model performance.

4 Method

4.1 spaCy Model

In addition to transformer-based models, a spaCy pipeline was trained for the same biomedical NER task using the `en_core_web_lg` model and the `-optimize accuracy` training flag. This option instructs spaCy to prioritize accuracy over speed during optimization, which is well-suited for NER tasks with fine-grained entity boundaries.

4.1.1 Data Conversion

The BC5CDR dataset originally provides entity annotations using `offset` and `length`. For spaCy compatibility, these annotations were converted to the format required by spaCy using absolute `start` and `end` character indices. This ensures proper alignment with spaCy's tokenization and supports accurate entity span creation.

4.1.2 Training Performance

The training loss and evaluation F1 score over time are shown in Figures 5 and 6.

Figure 5: Training Loss over Time for spaCy

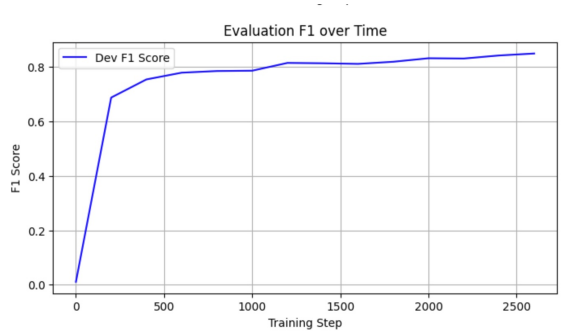

Figure 6: Evaluation F1 over Time for spaCy

We observe a rapid increase in F1 score early in training, followed by a steady convergence toward a final score above 0.85, indicating effective learning and stable generalization.

4.1.3 Evaluation Results

Final performance metrics for the spaCy model are listed in Table 1. The overall F1 score is 82.96, with particularly high performance on the `Chemical` class.

Entity	Precision	Recall	F1-score
Disease	78.42	78.39	78.40
Chemical	89.69	84.35	86.94
Overall (NER)	84.35	81.62	82.96

Table 1: Classification Report for spaCy NER Model

4.2 BERT Transformers

Two transformer models were fine-tuned for the NER task: **BioBERT** and **BERT-cased**. Both were trained under identical conditions for 10 epochs. Their training and evaluation losses, as well as F1 Score, Precision, and Recall, were tracked for comparison.

4.2.1 Data Conversion

The original BC5CDR dataset provides entity annotations using character-level offsets. For BERT-based training, the data was converted to the **CoNLL format**, which aligns tokens with IOB2 (Inside–Outside–Beginning) tags. Each token is paired with its corresponding entity tag, with document structure preserved for correct evaluation.

4.2.2 Training and Evaluation Loss

Figures 7 and 8 show the training and evaluation loss per epoch for both models.

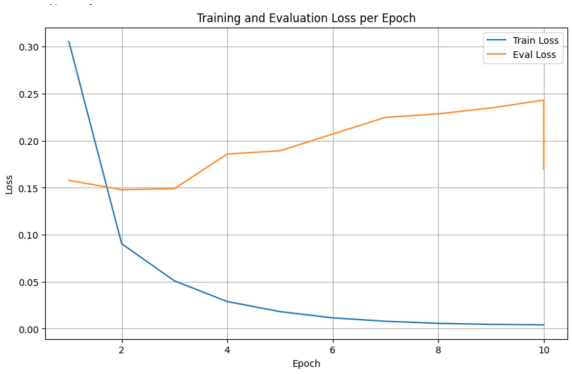

Figure 7: BioBERT Loss

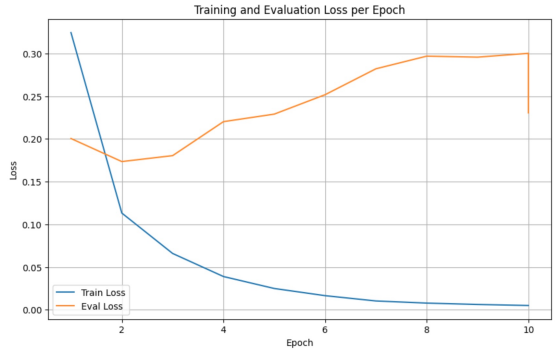

Figure 8: Bert-Cased Loss

As seen above, BioBERT maintains a more stable and lower evaluation loss compared to BERT-cased, but both models start overfitting after epoch 4.

4.2.3 Evaluation Metrics

Evaluation metrics per epoch are illustrated in Figures 9 and 10. BioBERT outperforms BERT-cased in all metrics (F1, Precision, and Recall) across almost all epochs.

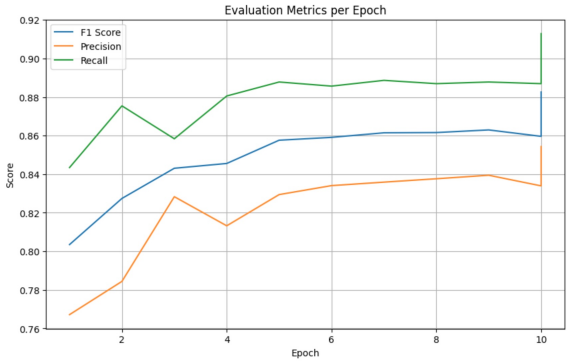

Figure 9: Evaluation Metrics per Epoch for BioBERT

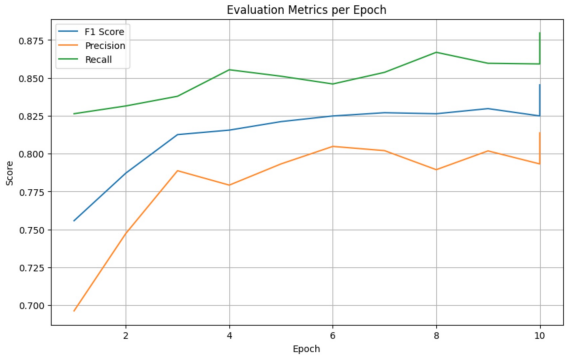

Figure 10: Evaluation Metrics per Epoch for BERT-cased

4.2.4 Classification Results

Tables 2 and 3 summarize the final test results after training.

Entity	Precision	Recall	F1-score	Support
CHEMICAL	0.92	0.95	0.93	1178
DISEASE	0.79	0.87	0.83	1141
Micro avg	0.85	0.91	0.88	2319
Macro avg	0.86	0.91	0.88	2319
Weighted avg	0.86	0.91	0.88	2319

Table 2: Classification Report for BioBERT

Entity	Precision	Recall	F1-score	Support
CHEMICAL	0.89	0.93	0.91	1178
DISEASE	0.74	0.83	0.78	1141
Micro avg	0.81	0.88	0.85	2319
Macro avg	0.82	0.88	0.85	2319
Weighted avg	0.82	0.88	0.85	2319

Table 3: Classification Report for BERT-cased

BioBERT achieves an overall accuracy of 0.969 and outperforms BERT-cased (accuracy 0.958) particularly in the recognition of the `DISEASE` entity, where the F1 score improves from 0.78 to 0.83. This suggests that BioBERT's domain-specific pretraining offers a significant advantage in biomedical entity recognition.

4.2.5 BERT Models Conclusion

Overall, BioBERT demonstrates better generalization and performance on biomedical NER tasks. It converges faster, maintains lower evaluation loss, and achieves higher F1 scores and accuracy on both entity types.

5 Conclusion

This study compared multiple approaches for Named Entity Recognition (NER) on the BC5CDR biomedical dataset, evaluating both a spaCy-based model and two transformer-based models: BioBERT and BERT-cased.

The spaCy pipeline, trained with the `en_core_web_lg` model and optimized for accuracy, achieved an overall F1 score of 82.96. It demonstrated solid performance, particularly on the `Chemical` entity class, suggesting spaCy can be an effective choice for NER in resource-constrained or speed-sensitive environments.

Transformer-based models, however, significantly outperformed the spaCy model. BioBERT, pretrained on biomedical corpora, yielded the highest overall performance with an F1 score of 0.88 and accuracy of 96.9%. It especially excelled at recognizing `Disease` entities, outperforming the general-purpose BERT-cased model, which achieved an F1 score of 0.85 and accuracy of 95.8%.

In conclusion, while spaCy offers a viable and efficient solution for biomedical NER, transformer-based models—particularly BioBERT—demonstrate superior performance due to their domain-specific pretraining. For high-stakes or precision-critical biomedical applications, BioBERT stands out as the most effective model among those evaluated.

6 Future Work

While this study demonstrates the superior performance of transformer-based models like BioBERT for biomedical NER, several promising directions remain for future work:

- **Integrating BioBERT with spaCy:** One compelling extension is to incorporate

BioBERT weights into a spaCy pipeline using the `spacy-transformers` extension. This would combine spaCy's efficient pipeline management and preprocessing tools with the deep contextual embeddings of BioBERT. Such integration could potentially offer both high accuracy and faster inference in production settings.

- **Exploring Other Domain-Specific Models:** Additional transformer models pretrained on biomedical or clinical text (e.g., SciBERT, PubMedBERT, or ClinicalBERT) could be evaluated on the BC5CDR dataset for a broader performance comparison across architectures.
- **Error Analysis and Span Boundary Refinement:** Further error analysis could guide improvements in span detection, particularly in edge cases with ambiguous or overlapping entities.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 106–111. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiao Li, Yifan Sun, Robert J. Johnson, Daniel Sciaky, Chih-hsuan Wei, Robert Leaman, Allan P. Davis, Christopher J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database (Oxford)*, 2016:baw068.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 6000–6010, Red Hook, NY, USA. Curran Associates, Inc.

A Authors Contributions

Contributions are as follow:

Dumitrache Flavian:

- Researched dataset
- Trained and compared BERT models

Andrei-Virgil Ilie:

- Trained spaCy model
- Managed documentation