DOI: 10.15439/2022F63
ISSN 2300-5963 ACSIS, Vol. 30

# Named Entity Recognition System for the Biomedical Domain

Raghav Sharma
Indian Institute of Technology Kharagpur
Kharagpur, India
Email: sharmaraghav.20@iitkgp.ac.in

Deependra Chauhan, Raksha Sharma
Indian Institute of Technology Roorkee
Roorkee, India
Email: {d_chouhan, raksha.sharma}@cs.iitr.ac.in

***Abstract*—The recent advancements in medical science have caused a considerable acceleration in the rate at which new information is being published. The MEDLINE database is growing at 500,000 new citations each year. As a result of this exponential increase, it is not easy to manually keep up with this increasing swell of information. Thus, there is a need for automatic information extraction systems to retrieve and organize information in the biomedical domain. Biomedical Named Entity Recognition is one such fundamental information extraction task, leading to significant information management goals in the biomedical domain. Due to the complex vocabulary (*e.g., mRNA*) and free nomenclature (*e.g., IL2*), identifying named entities in the biomedical domain is more challenging than any other domain, hence requires special attention. In this paper, we deploy two novel bi-directional encoder-based systems, *viz.*, BioBERT and RoBERTa to identify named entities in the biomedical text. Due to the domain-specific training of BioBERT, it gives reasonably good performance for the NER task in the biomedical domain. However, the structure of RoBERTa makes it more suitable for the task. We obtain a significant improvement in F-score by RoBERTa over BioBERT. In addition, we present a comparative study on training loss attained with ADAM and LAMB optimizers.**

## I. Introduction

INFORMATION extraction in the biomedical domain involves the identification of the independent pieces of information, for example, cause-effect arguments, causal triggers, adverse drug reaction, *etc*. Automated extraction of information from the biomedical text is an essential facilitator of clinical research and informed diagnosis [1], [2]. The presence of many domain-specific terminologies in biomedical literature makes information extraction a challenging task.

An entity is a word or a sequence of words in the text with a physical existence with different properties. Named entity recognition (NER) is a sub-task of information extraction that seeks to identify and classify named entities as predefined categories in unstructured text. NER always serves as the foundation for many natural language applications such as *question answering, text summarization*, and *machine translation*. Biomedical Named Entity Recognition (BioNER) is a task of identifying biomedical named entities such as *gene, disease, drug, species, etc.*, in the raw text. Because of the complexity of biomedical nomenclature, BioNER is a more challenging task than NER in general. A gene name often contains a mix of alphabet, digits, hyphens, and other characters, for example, *HIV-1*. The domain frequently uses abbreviations (“IL2” for “Interleukin 2”). In addition, the same biomedical named entities can be expressed in various forms. For example, gene names often contain alphabets, digits, hyphens, and other characters, thus having many variants (*e.g.*, “HIV-1 enhancer” versus “HIV 1 enhancer”). Moreover, many abbreviations (*e.g.*, “IL2” for “Interleukin 2”) have been used for biomedical named entities. Sometimes, the same entity can have very different aliases (*e.g.*, “PTEN” and “MMAC1” refer to the same gene) [1]. Another challenge of BioNER is the ambiguity problem. The same word or phrase can refer to more than one type of entities or does not refer to an entity depending on the context (*e.g.*, “TNF alpha” can refer to a protein or DNA).

Table I shows a few example sentences from the biomedical domain with the named entities and their types. Named Entity Recognition in the biomedical domain has been tried using various available methodologies and continues to be an active research topic due to the complexity and utility of the problem. BioBERT [3] is a language model trained on biomedical data to produce distributed representation of words. This paper presents a deep neural system for named entity recognition in the biomedical domain using BioBERT. Specifically, in this paper, we deploy two novel bi-directional encoder-based systems, *viz.*, BioBERT and RoBERTa to identify named entities in the biomedical text. Due to the domain-specific training of BioBERT, it gives reasonably good performance for the NER task in the biomedical domain. However, the supportive structure of RoBERTa makes it more suitable for the BioNER task than BioBERT.

## II. Related Work

Named Entity Recognition in the biomedical domain is a fundamental text mining task. It has attracted a lot of attention from researchers across different languages. Methodologies applied to this problem range from the traditional rule-based approaches to the most recent deep learning models. Due to the non-standard use of abbreviations, synonyms, synchronizations, ambiguities, and the frequent use of phrases to describe the entities, NER in the biomedical domain is still a challenging task [4].

Rule-based methods rely on hand-crafted rules to identify and classify named entities in text. An exhaustive lexicon almost always boosts the performance of these models. NER

TABLE I
EXAMPLES OF SENTENCES WITH THE NAMED ENTITY ANNOTATIONS

| S.No. | **Sentences with Annotations** |
|---|---|
| 1 | Identification of APC2, a homologue of the {adenomatous polyposis coli tumour}_*gene* suppresor. |
| 2 | {Methanoregula formicica}_*species* sp.nov., a methane-producing archaeon isolated from methanogenic sludge. |
| 3 | {IL-2}_*gene* gene expression and {NF-kappa B}_*protein* activation through {CD2B}_*antibody* requires reactive oxygen production by {5-lipoxygenase}_*protein*. |
| 4 | Assymetrical cell division was observed in rod-shaped cells. |

tools in the Biomedical domain rely on specific features to capture the characteristics of the different entity classes until recently. For instance, the suffix *-ase* is more frequent in protein names than in diseases; species names often consist of two tokens and have Latin suffixes; chemicals often contain specific syllabi like *methyl* or *carboxyl* [5]. However, handcrafted semantic and syntactic rules often make these models data specific. Any change in the source of data will drop the performance of the system [5], [6]. As a result, rule-based approaches lead to a high precision but low recall.

Advancements in supervised machine learning were also applied to generic NER. NER can be considered like a multiclass classification or sequence labeling task. The correct selection and engineering of features are vital to the model's performance based on them. Many machine learning models have been tried and researched based on these features. These include Hidden Markov Models (HMMs) [7], decision trees [8], SVMs [9] and Conditional Random Fields(CRFs). A major requirement for supervised machine learning models to perform well is the presence of sufficient labeled/structured data. However, the presence of labeled data is limited, leading to the rise of unsupervised learning approaches. These models tend to focus more on corpus statistics (e.g. IDF), terminologies, and syntactic knowledge KALM [10].

More recently, deep learning methods that can automatically develop and extract features from the raw text are used end-to-end for generic NER. These models generally use character or word-level embeddings such as Word2Vec and GloVe as their basic input. Various models based on CNNs and RNNs have been researched. However, the BiLSTM-CRF model [11] has been most commonly used. Transformer-based models [12] have proven to be superior in quality and also take less time to train. Based on transformers, several pre-trained language models have been released, which on fine-tuning give state-of-the-art performance on various end tasks. These include Generative Pre-trained Transformer (GPT) [13] (left to right architecture) and Bidirectional Encoder Representations from Transformers (BERT) [14] (takes both left and right context). Bio-BERT shows that pre-training BERT on biomedical data significantly improves its performance on end tasks in the biomedical domain. This paper uses BioBERT for named entity recognition in the biomedical domain. However, BioBERT takes a significant amount of time to train; we reduce the training time of BioBERT. We also modify the pre-training settings of BioBERT, which enables us to achieve better performance on the end task, that is, Named Entity Recognition in the biomedical text.

TABLE II
STATISTICS OF BIOMEDICAL NER DATASETS

| **Dataset** | **Entity Type** | **No. of annotations** |
|---|---|---|
| NCBI-Disease [16] | Disease | 6881 |
| BC5CDR [17] | Drug/Chem | 15411 |
| BC2GM [18] | Gene/Protein | 20703 |
| Species-800 [19] | Species | 3708 |

## III. DATASET

We preprocess the four datasets in the biomedical domain, *viz.*, NCBI-Disease, BC5CDR (drug/chem, disease), BC2GM, and Species-800. The preprocessing of the NCBI-Disease dataset results in fewer annotations than the original dataset because duplicate articles are removed from its training set. The *Species-800* dataset was preprocessed and split as per Pyysalo et al., [15]. The statistics of the biomedical NER dataset are listed in Table II.

## IV. METHODOLOGY

This paper presents a deep architecture for the named entity recognition in the biomedical domain. Our system deploys the representations of the words by a domain-specific language model, that is, BioBERT. We further optimize the system for the task using the LAMB optimizer. Furthermore, RoBERTa model is built on top of the BERT model. The architecture similarity with the BERT model makes RoBERTa model suitable for the named entity recognition task. This section describes the algorithm and its components.

### *A. BioBERT*

Text documents in the biomedical domain contain a considerable amount of domain-specific proper nouns, (*e.g., BRACA1*), which requires expertise in the domain to understand named entities. The general-purpose language representation models such as GloVe and Word2Vec give a poor performance for biomedical texts [20], [21]. The distribution of the words shifts from general domain corpora to biomedical corpora; hence direct application of generic word embeddings results in unsatisfactory performance [5], [15], [22]. BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is trained on the biomedical corpus. First, BioBERT is initialized with weights from BERT; BERT was pre-trained on general domain corpus to overcome the data sparsity problem and bring more coverage. The main advantage of BERT over previous language model approaches like combinations of LSTMs and CRF is that BERT has relatively a simple architecture based on bidirectional transformers. Based on its last layer representations, BERT computes only token level probabilities in the BIO2 format (Begin, Inside, Others). Then BioBERT is trained on biomedical corpora from PubMed. BioBERT is the first domain-specific BERT model which has been trained for biomedical-specific tasks [22].

### B. RoBERTa

RoBERTa stands for Robustly Optimized BERT Pre-training Approach. Most of the training procedure of BERT and RoBERTa is common. However, there are a few fundamental structural differences. This section presents the differences between the two models.

*1) Static Masking vs. Dynamic Masking:* BERT relies on randomly masking and predicting tokens. In the original BERT model, each sequence was masked in only ten different ways over 40 epochs. RoBERTa uses dynamic masking in which different mask is generated every time a sequence is fed to the model.

*2) Model Input Format and Next Sentence Prediction:* The BERT model is also trained for the Next Sentence Prediction (NSP) objective along with the masked language modeling objective. The objective of auxiliary Next Sentence Prediction loss is to determine whether the segments belong to the same or different documents. It is equiprobable for document segments to be sampled continuously from the same or distinct documents. RoBERTa, on the other hand, takes a different approach by ignoring the NSP loss. The input representation can be seen as packed with full sentences sampled contiguously from one or more than one documents. The maximum input length is set to 512 tokens.

*3) Training with Large Batches:* The same computational cost models can be made by increasing the batch size and decreasing the number of steps. The original BERT was trained with 256 sequence batch size for 1 million steps via gradient accumulation. This computational cost can approximate the training model for 125k steps with a batch size of 2k or 31k steps for 8k. It can be inferred from previous works done on neural networks that training the model with large mini-batches improves end-to-end performance. RoBERTa uses a batch size of 8k.

*4) Text Encoding:* The difference between the BPE vocabulary of original BERT and RoBERTa lies in the subword size, preprocessing of input, and tokenization rule. BERT uses 30k, whereas RoBERTa uses a more extensive vocabulary of 50k subwords. BERT does preprocess of input while RoBERTa expands vocabulary size without additional preprocessing. Roberta raises vocabulary size without other tokenization rules.

### C. LAMB Optimizer

This paper also shows the efficacy of the Large Batch Optimization (LAMB) algorithm with BioBERT for NER in the biomedical domain. LAMB helps to reduce the training time, and boost performance for text processing task [23]. Large batch training is the key to reducing deep neural networks' training time in a large distributed system. LAMB is a layer-wise adaptive large batch optimization technique. The generalization gap becomes a problem in the case of training large batches models. If direct optimization is performed, it may cause performance degradation. Devlin et al., [24] implemented BERT with a variant of ADAM optimizer, which uses ADAMs optimizer along with weight decay for training. LARS is another successful adaptive optimizer that has been used for large batch convolutional neural networks, but they are not effective for text processing tasks [23]. LAMB has shown superior performance across BERT and ResNet-50 training tasks with minimal hyperparameter tuning. Hence, we train BioBERT with the LAMB optimizer to optimize the training time. In addition, we show the superiority of the LAMB optimizer over the ADAM optimizer for the BioNER task (Section VI).

TABLE III
BIOBERT TOKEN LEVEL EVALUATION WITH ADAM OPTIMIZER

| Dataset | Precision | Recall | F-score | **Loss** |
|---|---|---|---|---|
| NCBI Disease | 88.8 | 91.8 | 90.2 | **33.71** |
| BC5CDR | 89.2 | 90.5 | 89.9 | **37.56** |
| BC2GM | 88.7 | 89.4 | 89.0 | **37.56** |
| Species-800 | 79.1 | 83.2 | 81.1 | **32.89** |

TABLE IV
BIOBERT TOKEN LEVEL EVALUATION WITH LAMB OPTIMIZER

| Dataset | Precision | Recall | F-score | **Loss** |
|---|---|---|---|---|
| NCBI Disease | 86.9 | 91.8 | 90.2 | **11.26** |
| BC5CDR | 89.2 | 90.5 | 89.9 | **10.74** |
| BC2GM | 88.7 | 89.4 | 88.88 | **17.17** |
| Species-800 | 79.1 | 83.2 | 80.28 | **12.52** |

TABLE V
BIOBERT ENTITY LEVEL EVALUATION

| Dataset | Precision | Recall | **F-score** |
|---|---|---|---|
| NCBI Disease | 86.92 | 89.27 | **88.08** |
| BC5CDR | 92.74 | 92.79 | **92.77** |
| BC2GM | 83.59 | 83.39 | **83.74** |
| Species-800 | 71.39 | 76.79 | **73.99** |

TABLE VI
ROBERTA ENTITY LEVEL EVALUATION

| Dataset | Precision | Recall | **F-score** |
|---|---|---|---|
| NCBI Disease | 87.32 | 88.84 | **88.07** |
| BC5CDR | 93.59 | 92.95 | **93.27** |
| BC2GM | 93.41 | 92.16 | **92.78** |
| Species-800 | 76.01 | 84.00 | **79.81** |

## V. Experimental Setup

The overall process can be divided into pre-training and fine-tuning BioBERT. The pre-training weights are taken from Cohen and Hunter [2]. The fine-tuning step is problem-specific. For example, the model needs to be fine-tuned for named entity recognition, relation extraction, question-answering, and tasks independently. We fine-tune the BioBERT model for our dataset's named entity recognition task. A batch size of $8$ was chosen for fine-tuning. The learning rate was set to $1e-5$, and the model was trained for $10$ epochs. F-score is computed at the token level, word level, and entity level, that is, phrase level.

## VI. Results

Results are focused on two aspects: the optimizer's performance during training and the F-score for the task. We compare the training Loss by ADAM optimizer and LAMB optimizer. ADAM optimizer is a frequently used optimizer for the classification task. Table III and Table IV present the F-Score obtained with BioBERT at token level with ADAM and LAMB respectively. The last column of Table III and Table IV shows the Loss attained during training with ADAM and LAMB, respectively. We observed a significant difference in the training Loss value with the LAMB optimizer; hence LAMB made the model converge in a significantly shorter time than ADAM. However, there is no significant difference in the F-score obtained with ADAM and LAMB. Table V and Table VI show the comparison between BioBERT and RoBERTa for NER across the four datasets from biomedical domain. We fine-tune both the models on our dataset for the named-entity recognition task. RoBERTa significantly improved the F-score for the named-entity recognition in the biomedical domain.

## VII. Conclusion

Named entity recognition in the biomedical domain is a challenging task considering the unconstrained nomenclature of the biomedical vocabulary. This paper presents a named-entity recognition system for the biomedical domain. We deploy two pre-trained language models for the task, *viz*., BioBERT and RoBERTa. Due to the domain-specific training of BioBERT, it gives reasonably good performance for the NER task in the biomedical domain. However, the structure of RoBERTa makes it more suitable for the task. Simple fine-tuning of RoBERTa on the dataset for BioNER boosts the results significantly. Additionally, we show a comparison between the training loss attained with ADAM and LAMB optimizers.

## References

[1] S. Ananiadou and J. Mcnaught, *Text mining for biology and biomedicine*. Citeseer, 2006.

[2] K. B. Cohen and L. Hunter, "Getting started in text mining," *PLoS computational biology*, vol. 4, no. 1, 2008.

[3] J. Lee, W. Yoon, and S. Kim, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, 2019.

[4] U. Leser and J. Hakenberg, "What makes a gene name? named entity recognition in the biomedical literature," *Briefings in Bioinformatics*, vol. 6, no. 4, p. 357–369, 2005.

[5] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, 2017.

[6] S. Eltyeb and N. Salim, "Chemical named entities recognition: a review on approaches and applications," *Journal of cheminformatics*, vol. 6, no. 1, pp. 1–12, 2014.

[7] D. M. Bikel, R. Schwartz, and R. M. Weischedel, "An algorithm that learns what's in a name," *Machine Learning*, vol. 34, no. 1-3, pp. 211–231, 1999.

[8] G. Szarvas, R. Farkas, and A. Kocsor, "A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms," *Discovery Science Lecture Notes in Computer Science*, p. 267–278, 2006.

[9] P. Mcnamee and J. Mayfield, "Entity extraction without language-specific resources," *proceeding of the 6th conference on Natural language learning - COLING-02*, 2002.

[10] A. Liu, J. Du, and V. Stoyanov, "Knowledge-augmented language model and its application to unsupervised named-entity recognition," *Proceedings of the 2019 Conference of the North*, 2019.

[11] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," 2015, cite arxiv:1508.01991. [Online]. Available: http://arxiv.org/abs/1508.01991

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018.

[14] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[15] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou, "Distributional semantics resources for biomedical text processing in proceedings of lbm. 2013," *Google Scholar*, pp. 39–44, 2013.

[16] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[17] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional lstm-cnns," 2015, cite arxiv:1511.08308Comment: To appear in Transactions of the Association for Computational Linguistics. [Online]. Available: http://arxiv.org/abs/1511.08308

[18] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, "Hierarchical attention networks for document classification." in *HLT-NAACL*, K. Knight, A. Nenkova, and O. Rambow, Eds. The Association for Computational Linguistics, 2016, pp. 1480–1489. [Online]. Available: http://dblp.uni-trier.de/db/conf/naacl/naacl2016.html#YangYDHSH16

[19] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/va challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 552–556, 2011.

[20] S. Pawar, R. Sharma, G. K. Palshikar, P. Bhattacharyya, and V. Varma, "Cause–effect relation extraction from documents in metallurgy and materials science," *Transactions of the Indian Institute of Metals*, vol. 72, no. 8, pp. 2209–2217, 2019.

[21] R. Sharma and G. Palshikar, "Virus causes flu: Identifying causality in the biomedical domain using an ensemble approach with target-specific semantic embeddings," in *International Conference on Applications of Natural Language to Information Systems*. Springer, 2021, pp. 93–104.

[22] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[23] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh, "Large batch optimization for deep learning: Training bert in 76 minutes," 2020.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.