RoBERTa-Based Approach for Named Entity Recognition in Biological Documents

1st MKOUKA BTISSAM
dept. Computer Science
Caddi Ayyad University
Marrakech, Morocco
b.mkouka8805@uca.ac.ma

2nd ZAHEDI CHAYMAA
dept. Computer Science
Caddi Ayyad University
Marrakech, Morocco
c.zahedi2314@uca.ac.ma

3rd JIHAD ZAHIR
dept. Computer Science
Caddi Ayyad University
Marrakech, Morocco
j.zahir@uca.ac.ma

***Abstract*—In recent years, Natural Language Processing (NLP) has become a fundamental tool in various domains, including bioinformatics and biodiversity research. Tasks such as Named Entity Recognition (NER) have enabled the automatic extraction of domain-specific entities from unstructured text, facilitating the construction of enriched and structured knowledge bases. This work focuses on applying NER techniques to biological documents to identify and extract relevant biological entities—specifically those related to fish and aquatic species—in support of the Add-my-Pet (AmP) project. The AmP portal aims to compile referenced data on the energetics of animal species, enabling the estimation of Dynamic Energy Budget (DEB) model parameters and related traits. To automate and scale this effort, we fine-tuned the RoBERTa transformer-based language model on a curated corpus of biological texts, carefully taking into account the issue of imbalanced data by employing advanced loss functions. Our approach achieved a training loss of 0.02 and an F1-score of 0.899, demonstrating the model's effectiveness in accurately identifying domain-specific entities. These results highlight the potential of transfer learning in extracting structured information from biological literature, ultimately contributing to the enrichment of species databases and supporting comparative biological analysis.**

***Index Terms*—Natural Language Processing (NLP), Named Entity Recognition (NER), RoBERTa, Fine-Tuning**

I. Introduction

With the growing interest in biodiversity, ecology, and species-specific energetics, there is an increasing need for automated tools that can extract structured biological information from unstructured texts. Biological platforms such as Add-my-Pet (AmP) continuously collect extensive data related to animal species, particularly on parameters used to model their life-history traits through the Dynamic Energy Budget (DEB) theory. However, much of this data remains embedded in natural language documents, making automated information extraction a key challenge.

Named Entity Recognition (NER), a subfield of Natural Language Processing (NLP), aims to identify and classify textual segments into predefined categories such as names of organisms, numerical values, biological measurements, and other semantically meaningful labels. While NER has been successfully applied in general domains and the medical field, biological texts present unique challenges due to the complexity, diversity, and specificity of their terminology. Biological entities vary in form and context and often include domain-specific abbreviations or measurement units that are not well-covered by general-purpose models.

In this work, we focus on the development of a domain-adapted NER system capable of recognizing biological entities from unstructured textual data. Our target entities include species names ("Specie"), family classification ("Family"), and a series of DEB-related biological traits such as age at hatching ("ah"), age at birth ("ab"), length at puberty ("Lp"), and wet weight at different life stages ("Wwb", "Wwp", "Wwi"). These entities are crucial for enriching the AmP database with structured knowledge extracted from textual resources.

To address this, we fine-tuned the pre-trained RoBERTa model, leveraging its contextual understanding capabilities for domain-specific entity recognition. Unlike traditional rule-based or dictionary-based systems, transformer-based models offer a more robust and generalizable solution by learning from annotated corpora and capturing deep semantic patterns. Our approach achieves strong empirical results, with a final training loss of 0.02 and an F1-score of 0.899, confirming its suitability for biological NER tasks.

By automating the extraction of biological traits and species information, our method contributes to the scalability of biological data curation efforts and facilitates more efficient integration into structured databases like Add-my-Pet. This work demonstrates the potential of modern NLP techniques to advance biodiversity informatics and bio-ontological knowledge discovery.

II. Related Work

Named Entity Recognition (NER) has seen substantial advancement in the biomedical domain, evolving from traditional rule-based systems to sophisticated deep learning approaches. The introduction of transformer-based architectures revolutionized biomedical NER.BioBERT(Lee et al 2020) [1], pre-trained on PubMed and PMC articles, significantly improved performance across multiple biomedical NLP tasks, including named entity recognition. Similarly, SciBERT (2019) [2], trained on a large corpus of scientific texts, demonstrated strong performance in various scientific domains. BioALBERT (2020) [3], further improved model efficiency using parameter reduction techniques without sacrificing performance.

Another important recent study of interest is the work titled "NER-RoBERTa: Fine-Tuning RoBERTa for Named Entity Recognition (NER) within Low-Resource Languages" [4] published in 2024, which proposes a RoBERTa-based NER system tailored for Kurdish language. While both our system and the Kurdish NER (KNER) model share a common backbone—RoBERTa-base—their implementation contexts and challenges diverge significantly. The KNER system fine-tunes RoBERTa on a small corpus of approximately 69,000 tokens, focusing on general entity types such as `PERSON`, `LOCATION`, `NUMBER`, and `DATE`, without addressing class imbalance or disclosing architectural adaptations. In contrast, our system is designed for a high-resource domain-specific setting, over 1.3 million tokens. We introduce a custom focal loss function to mitigate extreme class imbalance, and achieve an F1-score of 0.899 across 21 fine-grained biological entity types. Although KNER reports slightly higher global F1 scores (92.1% and 92.9% using different tokenization strategies), these are attained on less complex datasets. Our results thus highlight the importance of tailored learning strategies—such as loss reweighting—when adapting transformer-based models to domain-specific, imbalanced corpora.

Despite these advancements, most of the work focuses on clinical or biomedical data. Few efforts have been made to apply NER techniques to biological datasets like those in zoology studies. Our work addresses this gap by fine-tuning RoBERTa on domain-specific documents related to the Add-my-Pet (AmP) initiative. This is a novel application of NER for enriching biological trait databases by extracting entities such as species names, wet weights, lengths, and age metrics.

III. Methodology

A. Dataset

1) Data Collection: The proposed approach began by collecting a comprehensive dataset to support the Named Entity Recognition (NER) system. We collected data for **2,895** fish species to build a corpus tailored to biological entities in the context of the Add-my-Pet (AmP) portal.

We initiated the process by scraping data from **Wikipedia**, specifically targeting the definition and description sections of pages related to fish species. Our focus was on two major taxonomic classes: Chondrichthyes and Actinopterygii, as these pages frequently contain rich textual contexts that include biological entities of interest—such as family names and physiological traits like age, length, and weight.

To further strengthen the dataset, we conducted a detailed examination of the Add-my-Pet (AmP) database to trace the origin of its species-related content. Our analysis revealed that the majority of biological information referenced in AmP entries originated from **FishBase** [5], a globally recognized repository for fish biodiversity. Based on this insight, we developed a scraping pipeline that systematically parsed the reference links provided in AmP entries—especially those pointing to FishBase—and programmatically verified their validity. If a reference link was valid, the corresponding species description was extracted. This process resulted in the collection of 1,914 valid FishBase texts, aligned with 2,886 Wikipedia entries previously gathered for the same species. This dual-source strategy resulted in a total of **5,905** textual samples, covering a wide range of biological and taxonomical information. The diversity and complementarity of these two sources enriched the training corpus, enabling our fine-tuned RoBERTa model to effectively learn and generalize the recognition of domain-specific biological entities.

2) Data Annotation: To build a high-quality training corpus for Named Entity Recognition (NER), we annotated the collected texts using an advanced large language model (LLM) available via the free-tier API of **DeepSeekR1** from Grok Cloud. This model was selected based on its strong performance in various NLP benchmarks, particularly as reported by Vellum, which ranks LLMs according to accuracy and robustness across tasks.

To reduce the prompt complexity and improve annotation consistency, the raw species descriptions were annotated directly within the text using an inline format. Specifically, entities were marked in the form [specie: ...], [family: ...], [ah:...]. This approach was chosen to make the simple manual verification easier, which was conducted by a PhD student with domain expertise. Inline annotations allowed quick scanning and correction without needing to cross-reference entity positions.

After manual validation, we developed a script to convert the inline annotations into the standard BIO format (Beginning–Inside–Outside), which is widely used for sequence labeling tasks. This conversion enabled the preparation of the dataset in a format compatible with the fine-tuning of transformer-based NER models, such as RoBERTa. The distribution of the annotated entity tags is shown in Table I.

TABLE I
DISTRIBUTION OF BIO-ANNOTATED ENTITY TAGS IN THE DATASET

Name of Tags	Count
O	1,261,627
B-SPECIE	4,126
I-SPECIE	3,980
B-FAMILY	3,864
I-FAMILY	60
B-AH	86
I-AH	156
B-AB	185
I-AB	167
B-AM	950
I-AM	1,166
B-LB	123
I-LB	321
B-LP	1,086
I-LP	2,788
B-LI	5,473
I-LI	11,763
B-Wwb	21
I-Wwb	25
B-Wwp	46
I-Wwp	67
B-Wwi	1,590
I-Wwi	3,515

O : stands for Outside, indicating a token that is not part of any named entity (Non-Named Entity).

B-XX (Begin) : marks the beginning of an entity of type XX, for example, B-SPECIE denotes the start of a species entity.

I-XX (Inside) : indicates a token inside an entity of type XX, continuing from a preceding B-XXX or I-XXX token.

The Table II summarized an example of the annotated corpus which includes two sentences with corresponding tokens and their tags.

TABLE II
SUMMARIZED EXAMPLE OF THE ANNOTATED CORPUS WITH BIO TAGS.

Sentence	Word	Tag
Sentence 1	The	O
Sentence 1	Rhizoprionodon	B-SPECIE
Sentence 1	oligolinx	I-SPECIE
Sentence 1	is	O
Sentence 1	a	O
Sentence 1	requiem	O
Sentence 1	shark	O
Sentence 1	of	O
Sentence 1	the	O
Sentence 1	family	O
Sentence 1	Carcharhinidae	B-FAMILY
Sentence 2	typically	O
Sentence 2	measure	O
Sentence 2	60–75	B-LI
Sentence 2	cm	I-LI
Sentence 2	in	O
Sentence 2	length	O

3) Data Preprocessing: The preprocessing of our dataset involved several critical steps to ensure the text was suitable for training the RoBERTa model for Named Entity Recognition (NER). Initially, we cleaned the text by removing unwanted artifacts, such as ¡think¿ tags generated by the DeepSeek R1 API during annotation, which could interfere with model training. Additionally, we removed reference annotations, such as [12], commonly found in Wikipedia and FishBase texts, to eliminate noise and maintain focus on biological entities. For tokenization, we employed **Byte-Pair Encoding (BPE)** to effectively handle sub-words and compound structures, aligning with RoBERTa's tokenization approach. This was implemented using the tokenizer **AutoTokenizer.from_pretrained(”FacebookAI/roberta-base”, add_prefix_space=True)** from Hugging face, ensuring that the tokenized output was compatible with the model's input requirements and optimized for recognizing entities like species names, families, and energetic traits.

B. Model Architecture

Our Named Entity Recognition (NER) system leverages the RoBERTa model, an optimized version of Bidirectional Encoder Representations from Transformers (BERT) [6] , RoBERTa is a transformers model pretrained on a large corpus of English data in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labelling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts, and its designed to improve performance through enhanced pre-training strategies. RoBERTa's architecture consists of stacked transformer layers, each comprising multi-head attention, add normalize, Feed-Forward Network (FFN), and a second add normalize layer, enabling robust contextual understanding of text as shown in the figure 1. For our task, we fine-tuned the roberta-base model using the AutoModelForTokenClassification from the Hugging Face library, configured with a custom number of labels corresponding to our entity set: ”Specie,” ”Family,” ”ah,” ”ab,” ”am,” ”Lb,” ”Lp,” ”Li,” ”Wwb,” ”Wwp,” and ”Wwi.” The model was initialized with id2label and label2id mappings to handle token classification and was moved to a CUDA device for efficient computation.

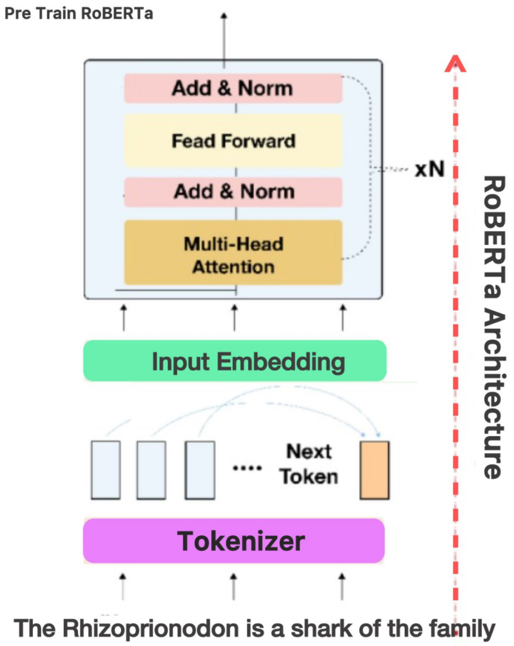

Figure 1: The architecture of the RoBERTa model.

To evaluate the model's performance, we implemented a custom `compute_metric` function using the `seqeval` library to calculate the F1 score, focusing on tokens with valid labels (excluding `-100`, which denotes ignored tokens). The function processes model logits and true labels, computing predictions via `argmax` along the label dimension, and returns the overall F1 score.

The training process utilized the `TrainingArguments` class with a setup of 5 epochs, a learning rate of 2e-5, a weight decay of 0.01, and batch sizes of 16 for both training and evaluation. We employed a `DataCollatorForTokenClassification` to handle dynamic padding of tokenized inputs, ensuring

compatibility with the model. The `Trainer` class from Hugging Face was used to orchestrate training, integrating the model, tokenizer, data collator, and tokenized datasets (`ds_tokenized["train"]` and `ds_tokenized["test"]`), with progress logged to Weights & Biases (wandb) every 10 steps.
To address the issue of data imbalance(as shown in Table I), particularly affecting rare entity labels, we incorporated the **Focal Loss** technique during training. As shown in Table III, this approach significantly improved performance on underrepresented classes, while maintaining strong generalization on the overall dataset.

IV. Experiments and Evaluation Results

To evaluate our Named Entity Recognition (NER) system, we initially fine-tuned the `RoBERTa-base` model on a dataset of 5,790 text samples sourced from Wikipedia and FishBase, encompassing 2,895 fish species from the *Chondrichthyes* and *Actinopterygii* classes. The dataset, comprising approximately **1.3 million** tokens, was tokenized using the RoBERTa tokenizer and split into 80% training and 20% testing sets.

In the first phase, we trained the model without addressing the issue of class imbalance. Using the standard cross-entropy loss, the model achieved a strong performance with an **F1 score of 0.871** and a loss of **0.029**, particularly on common entity types such as `SPECIE`, `FAMILY`, `LI`, and `LP`. However, performance remained limited for rare entity types such as wet weight at birth (`Wwb`, 21 occurrences) and age at hatching (`AH`, 86 occurrences).

To address the issue of class imbalance and to improve recognition of these underrepresented and difficult entities, we experimented with the **Focal Loss** function. Unlike standard or weighted cross-entropy losses, Focal Loss dynamically down-weights well-classified examples and concentrates the learning process on harder, misclassified ones. With this approach, the model achieved an improved **F1 score of 0.899** and a lower loss of **0.02**. These results confirm that Focal Loss enhances the model's ability to learn from sparse classes, without compromising the overall performance an shown in table III .

Loss Function	**F1 Score**	**Loss Value**
Cross-Entropy(default)	0.871	0.029
Focal Loss	**0.899**	**0.02**

TABLE III
Comparison of model performance with and without Focal Loss

The model was trained for 5 epochs using a learning rate of 2e-5, a batch size of 16, and a weight decay of 0.01 on a CUDA-enabled GPU. The final fine-tuned version was made publicly available on the Hugging Face Hub under the name `mkouka/ner_roberta_v10`. Additionally, we developed a web interface that allows users to input raw text and visualize extracted entities in context, supporting easier validation and application in ecological and taxonomic research.

V. Discussion and Comparison

Our NER system, designed for extracting fish-related entities from a domain-specific corpus of 5,790 samples (approximately 1.3 million tokens) sourced from Wikipedia and FishBase, shares architectural similarities with the Kurdish NER (KNER) approach proposed in the study **”NER-RoBERTa: Fine-Tuning RoBERTa for Named Entity Recognition (NER) within Low-Resource Languages”** [4]. Both systems fine-tune the RoBERTa-base model, leveraging its transformer-based architecture characterized by bidirectional encoding and self-attention mechanisms for effective token-level classification. However, the two approaches diverge significantly in implementation and context. The KNER study operates within a low-resource setting, utilizing a custom-annotated Kurdish corpus containing 69,309 tokens. While the authors mention modifying RoBERTa's architecture, specific technical details—such as additional layers, tuning strategies, or loss functions—are not disclosed. Moreover, the study does not explicitly address class imbalance, a common issue in NER for low-resource languages, which may limit its performance on underrepresented entity classes like `PERSON` or `LOCATION`. In contrast, our system introduces a custom focal loss function integrated into a modified `FocalTokenClassificationModel` to directly mitigate extreme class imbalance (e.g., 1,261,627 `O` tags versus only 21 `B-Wwb` tags). This adaptation contributes to a high overall F1 score of 0.899 and improves recognition of rare entities. Although the KNER system reports slightly higher global F1 scores—92.1% using Byte-Pair Encoding and 92.9% with SentencePiece tokenization—these metrics are achieved on a much smaller and less diverse dataset composed of general entity types. Our performance, while marginally lower in absolute terms, reflects a more challenging setting involving severe imbalance and fine-grained biological annotations across 21 specialized entity types. Our corpus is both larger and more domain-specific, annotated with 21 fine-grained entity types such as `SPECIE`, `FAMILY`, `Wwb`, and `AH`, offering a richer semantic structure compared to general-purpose NER datasets. In contrast, the KNER corpus focuses on general entities such as `PERSON`, `LOCATION`, `NUMBER`, and `DATE`, lacking the domain-specific granularity required in specialized applications like biological taxonomy. Furthermore, per-class F1 evaluations highlight the strength of focal loss in enhancing detection performance for low-frequency tags—a robustness that is critical in biological data annotation. This comparison underlines how adapting pre-trained language models like RoBERTa to specialized domains not only demands dataset-specific annotation but also tailored learning strategies such as loss reweighting. While the KNER system demonstrates the adaptability of RoBERTa to low-resource languages, our system shows how architectural and loss-based enhancements can significantly elevate NER performance in highly imbalanced, domain-constrained settings.

VI. Conclusion

In this work, we successfully tackled the challenge of named entity recognition (NER) on a highly imbalanced, domain-specific dataset by fine-tuning the `RoBERTa` model. The dataset, composed of multiple custom entity types relevant to biological and chemical nomenclature, posed a significant difficulty due to the overwhelming dominance of the non-entity tag "O".

To address this, we introduced **Focal Loss**, a loss function designed to focus learning on harder, misclassified examples and mitigate the effect of class imbalance. The enhanced model achieved a remarkable F1-score of 89,9%, reflecting its strong ability to generalize and correctly classify both frequent and rare entities. The effectiveness of our approach was confirmed through both quantitative evaluation and qualitative inspection of predicted entities in real documents.

Beyond experimentation, we pushed our trained model to the Hugging Face Hub to ensure public accessibility and reproducibility. Additionally, we developed a user-friendly interface that highlights named entities directly in uploaded or pasted documents, making our solution readily usable by researchers, domain experts, and developers.

This project not only demonstrates the power of transformer-based models in specialized NER tasks but also shows how thoughtful handling of class imbalance and user-centric deployment can significantly enhance the practicality of machine learning solutions.

References

[1] BioBERT: a pre-trained biomedical language representation model for biomedical text mining | bioinformatics | oxford academic.

[2] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text.

[3] Usman Naseem, Matloob Khushi, Vinay Reddy, Sakthivel Rajendran, Imran Razzak, and Jinman Kim. BioALBERT: A simple and effective pre-trained language model for biomedical named entity recognition.

[4] Abdulhady Abas Abdullah, Srwa Hasan Abdulla, Dalia Mohammad Toufiq, Halgurd S. Maghdid, Tarik A. Rashid, Pakshan F. Farho, Shadan Sh Sabr, Akar H. Taher, Darya S. Hamad, Hadi Veisi, and Aras T. Asaad. NER- RoBERTa: Fine-tuning RoBERTa for named entity recognition (NER) within low-resource languages.

[5] FishBase.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach.