BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 8: e140428
doi: 10.3897/biss.8.140428

OPEN ACCESS

Conference Abstract

BiodiViz: Leveraging NER and RE for Automated Knowledge Graph Generation in Biodiversity Research

Angela Shannen S. Tan‡, Paul Michael C. Dimayuga‡, Roselyn Gabud‡,§

‡ University of the Philippines Diliman, Quezon City, Philippines
§ University of the Philippines Los Baños, Los Baños, Philippines

Corresponding author: Angela Shannen S. Tan (shannentan22@gmail.com)

Received: 28 Oct 2024 | Published: 29 Oct 2024

Citation: Tan ASS, Dimayuga PMC, Gabud R (2024) BiodiViz: Leveraging NER and RE for Automated Knowledge Graph Generation in Biodiversity Research. Biodiversity Information Science and Standards 8: e140428. https://doi.org/10.3897/biss.8.140428

Abstract

In biodiversity research, the integration of machine learning and data visualization is increasingly important for uncovering valuable insights from academic literature. This study introduces an innovative knowledge graph application, BiodiViz, designed to translate intricate text into intuitive visual representations, fostering a deeper comprehension of biodiversity relationships.

BiodiViz uses the top-performing Named Entity Recognition (NER) and Relation Extraction (RE) models to automatically generate a comprehensive knowledge graph for biodiversity research. The NER model extracts and categorizes entities like organisms, phenomena, and habitats, while the RE model identifies relationships such as "have," "occur in," and "influence" from the BiodivNERE dataset (Abdelmageed et al. 2022). These entities and relationships are organized into nodes and edges within a graph.

Researchers input text into BiodiViz, producing a visual knowledge graph that simplifies the analysis of complex biodiversity data, reducing manual effort and enhancing efficiency.

Named Entity Recognition & Relation Extraction

BiodiViz leverages advanced Bidirectional Encoder Representations from Transformers (BERT)-based Large Language Models (LLMs) (Rogers et al. 2020), fine-tuned specifically for NER and RE tasks using the BiodivNERE dataset. The fine-tuning process involved various models, including BERT (Devlin et al. 2019), ELECTRA (Clark et al. 2020), and BiodivBERT (Abdelmageed et al. 2023). These models were evaluated for performance using the results of their F1-score as the main metric, which is the harmonic mean of precision (the proportion of true positive results among all positive predictions) and recall (the proportion of true positive results among all actual positives), with BiodivBERT achieving an F1-score of 77.16% for the NER task, while BERT excelled in the RE task with an F1-score of 81.28%. Rigorous hyperparameter optimization further enhanced the performance of BiodivBERT in the RE task by 3.38%.

The BiodivNERE corpora by Abdelmageed et al. (2022) were used to fine-tune several models for NER and RE tasks in the biodiversity domain. The first corpus from the BiodivNERE corpora is BiodivNER, which is a gold standard dataset (manually labelled test corpora) for evaluating NER tasks. The fine-tuning process employed the token classification method from the Hugging Face library (Hugging Face 2023b), which assigns labels to each token in a sequence. Experiments were conducted with a batch size of four, meaning the model processes four examples/rows of data at a time before making an update to improve its learning. This is due to the constraints of the NVIDIA® GeForce RTX™ 3060 graphics processor. (NVIDIA 2024) Model performance was evaluated using the seqeval library (Nakayama 2018), focusing on accuracy, precision, recall, and F1 scores.

For text classification, the second corpus, BiodivRE, was utilized, following previous research recommendations to explore fine-tuning settings for BiodivBERT. Hyperparameter optimization (Feurer and Hutter 2019) was conducted using Hugging Face's Trainer API with an Optuna backend (Hugging Face 2023a), concentrating on learning rate and the number of training epochs (i.e., the number of complete passes through the entire dataset during model training).

The BiodiViz Knowledge Graph Application

The fine-tuned NER and RE models with the best F1-scores—BiodivBERT and BERT, respectively—were integrated into the knowledge graph application.

Fig. 1 illustrates the flowchart of the application pipeline. Each sentence in the input text will go through the NER model to identify and label the entities within the sentence. Subsequently, these labeled entities, together with the original sentence, will be input into the RE model. The RE model will analyze every pair of entities for a potential relation and output the type of relation they share. The application will then utilize this data to create a graph with appropriate labels and color-coding. An example of the application's user interface with the knowledge graph is shown in Fig. 2.

This study highlights the practical application of machine learning and data visualization in advancing biodiversity research, emphasizing the importance of developing user-friendly tools to support scientific exploration and discovery.

The BiodiViz application, including the code and resources, is available on GitHub*[1], providing an accessible tool for biodiversity researchers to streamline their analyses.

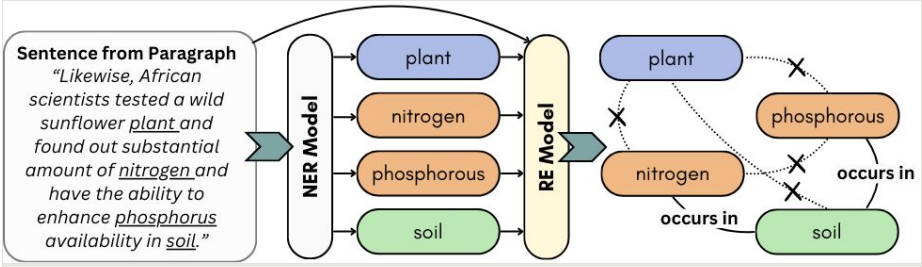

Figure 1.

Application pipeline with integrated named entity recognition (NER) and relation extraction (RE) models.

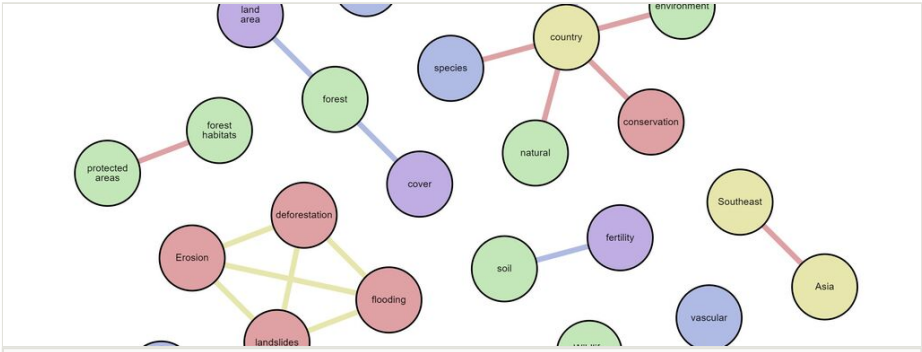

Figure 2.

Section of BiodiViz graph page with knowledge graph generated from input text

Keywords

information extraction, named entity recognition, relation extraction

Presenting author

Roselyn S. Gabud

Presented at

SPNHC-TDWG 2024

Acknowledgements

We would like to thank our advisers, Dr. Paul Regonia, Dr. Prospero Naval, and Carlo Raquel, for their utmost support and guidance throughout the duration of our project.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Abdelmageed N, Löffler F, Feddoul L, Algergawy A, Samuel S, Gaikwad J, Kazem A, König-Ries B, et al. (2022) BiodivNERE: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain. Biodiversity Data Journal 10 https://doi.org/10.3897/bdj.10.e89481
- Abdelmageed N, Löffler F, König-Ries B (2023) BiodivBERT: a Pre-Trained Language Model for the Biodiversity Domain. URL: https://www.researchgate.net/publication/372338738_BiodivBERT_a_Pre-Trained_Language_Model_for_the_Biodiversity_Domain
- Clark K, Luong M, Le Q, Manning C (2020) ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv https://doi.org/10.48550/arXiv.2003.10555
- Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv https://doi.org/10.48550/arXiv.1810.04805
- Feurer M, Hutter F (2019) Automated Machine Learning. The Springer Series on Challenges in Machine Learning, 3-33 pp. https://doi.org/10.1007/978-3-030-05318-5
- Hugging Face (2023a) Hyperparameter Search using Trainer API. https://huggingface.co/docs/transformers/en/hpo_train
- Hugging Face (2023b) Libraries. https://huggingface.co/docs/hub/en/models-libraries
- Nakayama H (2018) seqeval: A Python framework for sequence labeling evaluation. URL: https://github.com/chakki-works/seqeval
- NVIDIA (2024) GeForce RTX 3060 Family. https://www.nvidia.com/en-gb/geforce/graphics-cards/30-series/rtx-3060-3060ti/
- Rogers A, Kovaleva O, Rumshisky A (2020) A Primer in BERTology: What We Know About How BERT Works. Transactions of the Association for Computational Linguistics 8: 842-866. https://doi.org/10.1162/tacl_a_00349

Endnotes

*1 https://github.com/shannentan22/BiodiViz