Recognition of Biodiversity-related Named Entities by Fine-tuning General-domain BERT-based Language Models

Geilah T. Tabanao[1], Andrew Miguel V. Pagdanganan[1], Riza Batista-Navarro[2,3] and Roselyn S. Gabud[1,2,*]

[1]*Department of Computer Science, University of the Philippines Diliman, Quezon City, Philippines*

[2]*Institute of Computer Science, University of the Philippines Los Baños, Laguna, Philippines*

[3]*Department of Computer Science, University of Manchester, UK*

Abstract

Named Entity Recognition (NER) is crucial for various Natural Language Processing (NLP) tasks, including uncovering insights from vast textual datasets. We evaluated Bidirectional Encoder Representations from Transformers (BERT) models pre-trained on general data, fine-tuning them on the COPIOUS dataset for biodiversity NER. Achieving the most optimal performance, our DeBERTa NER model was employed in a biodiversity Information Extraction pipeline, which was applied on the forestry compendium of the Centre for Agricultural and Biosciences International Digital Library. We demonstrate that the pipeline enables the enrichment of descriptive information on reproductive conditions and habitats of tree species.

Keywords

Named Entity Recognition, Biodiversity, Transformers, Information Extraction

1. Named Entity Recognition models

In the biodiversity domain, named entity recognition (NER) systems that can extract named entities relevant to the identification of species occurrence information (e.g., taxonomic names, geographic locations, temporal expressions, and habitats) in literature could form the basis of NLP applications such as the automatic curation of databases. A biodiversity occurrence database that has been curated with the support of NLP systems could potentially provide researchers with long-term, broad-scale data from the literature, that will then be readily available for analysis.

In this paper, we aim to investigate the performance of transformer models on the biodiversity NER task. Specifically, we sought to assess the NER performance of BERT models [1] that were pre-trained on massive amounts of general-domain data, when fine-tuned on a domain-specific corpus. To this end, we developed NER models by fine-tuning BERT-base, DistilBERT, ALBERT, RoBERTa, and DeBERTa models [2] on the COPIOUS dataset [3], the biggest annotated corpus relevant to species occurrence data. The results of the work by Abdelmageed et al. [4] showed

15th International SWAT4HCLS Conference, February 26–29, 2024, Leiden, The Netherlands

[*]Corresponding author.

geilahtabanao67@gmail.com (G. T. Tabanao); avpagdanganan@up.edu.ph (A. M. V. Pagdanganan); riza.batista@manchester.ac.uk (R. Batista-Navarro); rsgabud@up.edu.ph (R. S. Gabud)

CEUR Workshop Proceedings (CEUR-WS.org)

that this is the one dataset where pre-training a BERT model on domain-specific data, did not lead to any improved performance, thus prompting the question of whether other BERT-based models could perform better, even when pre-trained on general-domain data only. Amongst our fine-tuned models, DeBERTa obtained the best performance, with an F1-score of 84.18%. This is impressive, considering that this model was not pre-trained on domain-specific data.

2. Knowledge Graph Curation

A popular application of NER is the extraction of fine-grained information from text, that can then be leveraged to populate or curate structured databases. In this vein, we set out to explore the extent to which an Information Extraction pipeline underpinned by NER and relation extraction (RE), can curate a biodiversity-focused database, based on information buried within textual descriptions of various tree species in the Centre for Agricultural and Biosciences International (CABI) Digital Library.[1] Specifically, we integrated our best performing NER model into the pipeline, and applied an existing RE model to extract information on the habitats and reproductive conditions of species in the CABI Library forestry compendium.

Taking a corpus of CABI textual descriptions, our pipeline: (1) applies NER to extract mentions of geographic locations, habitats and temporal expressions; (2) applies RE to identify related habitats and geographic locations (i.e., habitat-geographic location relations) and related reproductive conditions and temporal expressions (i.e., reproductive condition-temporal expression relations); and (3) populates a graph database to store the related entities, to allow for querying and visualisation.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. URL: http://arxiv.org/abs/1810.04805. doi:10.48550/arXiv.1810.04805, arXiv:1810.04805 [cs].

[2] K. S. Kalyan, A. Rajasekharan, S. Sangeetha, AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing, 2021. URL: http://arxiv.org/abs/2108.05542. doi:10.48550/arXiv.2108.05542, arXiv:2108.05542 [cs].

[3] N. T. Nguyen, R. S. Gabud, S. Ananiadou, COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature, Biodiversity Data Journal (2019) e29626. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351503/. doi:10.3897/BDJ.7.e29626.

[4] N. Abdelmageed, F. Löffler, B. König-Ries, BiodivBERT: a Pre-Trained Language Model for the Biodiversity Domain, in: A. Yamaguchi, A. Splendiani, M. S. Marshall, C. Baker, J. T. Bolleman, A. Burger, L. J. Castro, O. Eigenbrod, S. Österle, M. Romacker, A. Waagmeester (Eds.), 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4HCLS 2023), Basel, Switzerland, February 13-16, 2023, volume 3415 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 62–71. URL: https://ceur-ws.org/Vol-3415/paper-7.pdf.

[1] https://www.cabidigitallibrary.org/