Towards the Curation of Environment-related Knowledge Graphs: Fine-tuning General-domain Language Models for Biodiversity Named Entity Recognition

Geilah T. Tabanao[1], Andrew Miguel V. Pagdanganan[1], Riza Batista-Navarro[2,3], 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

Overview

Motivation: Leveraging Information Extraction (IE) systems to convert unstructured scientific knowledge into structured, accessible data can enable policymakers to take more effective climate action.

Goal: To demonstrate that fine-tuning general-domain models is sufficient for extracting named entities related to species occurrence from biodiversity literature and to integrate our Named Entity Recognition (NER) model into a biodiversity Information Extraction (IE) pipeline applied to a forestry compendium.

Approach: Evaluate the Named Entity Recognition (NER) performance of Bidirectional Encoder Representations from Transformers (BERT) models fine-tuned on a domain-specific corpus.

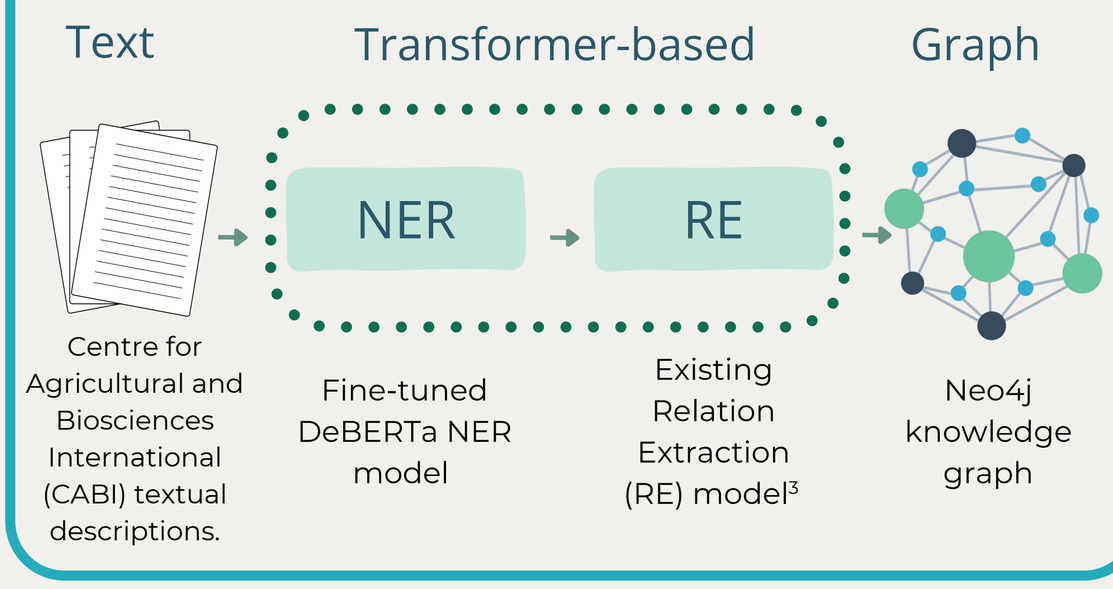
Methods

Fine-tune the selected BERT models on the COPIOUS corpus[1] using BiodivBERT's hyperparameters[2] :

- BERT-base (bert-base-cased)
- DistilBERT (distilbert-base-cased)
- ALBERT (albert-base-v2)
- RoBERTa (roberta-base)
- DeBERTa (deberta-v3-base)

Performance assessed using SeqEval framework

Pipeline



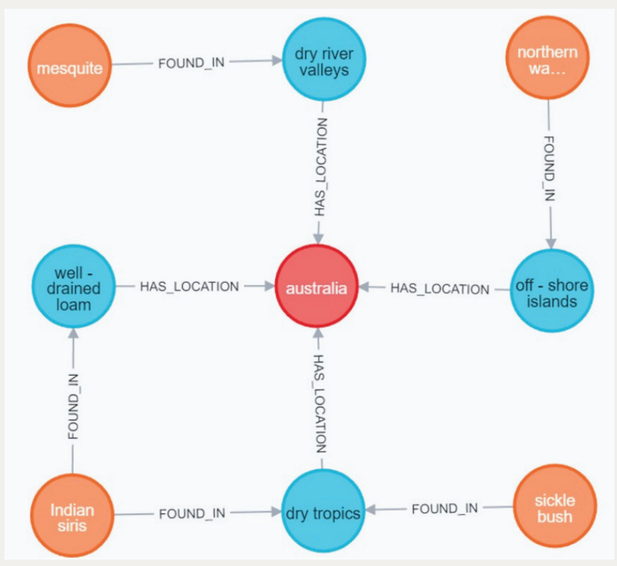
Results

Table 1: F1-scores obtained by the BERT-based NER models on the COPIOUS test set. Key: Geo Loc = Geographic Location; Temp Expr = Temporal Expression.

NE Type	DistilBERT	ALBERT	BERT-base	RoBERTa	DeBERTa	BiodivBERT
Taxon	85.59	83.64	85.72	86.11	**87.60**	86.81
Geo Loc	85.62	84.16	86.74	**87.85**	87.58	86.74
Temp Exp	78.11	73.58	81.50	79.58	70.28	**82.59**
Habitat	69.91	65.70	66.21	68.76	**70.93**	66.99
Person	64.71	63.24	69.15	65.88	68.08	**69.32**
OVERALL	82.77	80.67	83.51	83.87	84.18	84.23

The DeBERTa NER model demonstrated the best performance, obtaining a micro-averaged F1-score of 84.18% based on entity-level evaluation.

Knowledge Graph Curation



Sample Inferred Data:

Mesquites were found in dry river valleys in Australia

Conclusion

- General-domain BERT variants fine-tuned on domain-specific corpus for NER perform comparably to domain-specific BERT models.
- Integrated into an IE pipeline, our best NER model enabled knowledge graph curation, allowing retrieval and visualization of fine-grained information hidden in text.
- Future work will explore larger general-domain models to enhance biodiversity NER performance.

CITATIONS:

[1] Nhung T.H. Nguyen, Roselyn S. Gabud, and Sophia Ananiadou. COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. Biodiversity Data Journal, (7):e29626, January 2019. ISSN 1314-2828. doi: 10.3897/BDJ.7.e29626. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351503/

[2] Nora Abdelmageed, Felicitas Löffler, and Birgitta König-Ries. BiodivBERT: a Pre-Trained Language Model for the Biodiversity Domain. In Atsuko Yamaguchi, Andrea Splendiani, M. Scott Marshall, Chris Baker, Jerven T. Bolleman, Albert Burger, Leyla Jael Castro, Ole Eigenbrod, Sabine Österle, Martin Romacker, and Andra 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, pp.62–71. CEUR-WS.org, 2023. URL https://ceur-ws.org/Vol-3415/paper-7.pdf

[3] Roselyn Gabud, Portia Lapitan, Vladimir Mariano, Eduardo Mendoza, Nelson Pampolina, Maria Art Antonette Clariño, and Riza Batista-Navarro. A Hybrid of Rule-based and Transformerbased Approaches for Relation Extraction in Biodiversity Literature. In The Second Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning (Pan-DL), co-located with The 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 2023