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

Tabanao, Geilah | Pagdanganan, Andrew Miguel | Batista-Navarro, Riza | Gabud, Roselyn

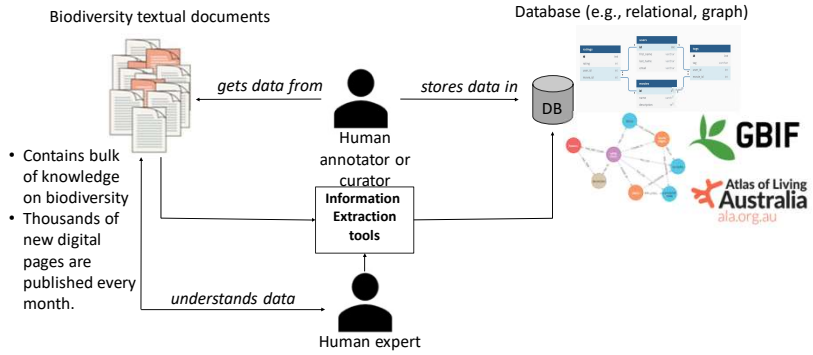
15th International SWAT4HCLS Conference
Semantic Web Applications and Tools for Health Care and Life Sciences
February 26-29, 2024 | Leiden, The Netherlands

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
[*] rsgabud@up.edu.ph

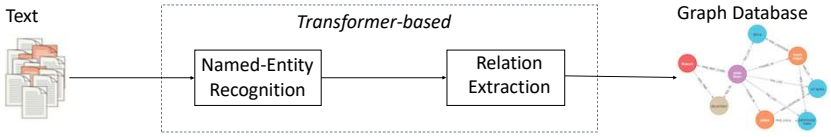
Background

1) Growing need for Information Extraction tools: from text to structured formats
2) 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.



Research Objectives

1) 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.
2) To employ the best performing NER model in a biodiversity Information Extraction pipeline, which was applied on the forestry compendium of the Centre for Agricultural and Biosciences International Digital Library (CABI).



Methodology

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.

Results

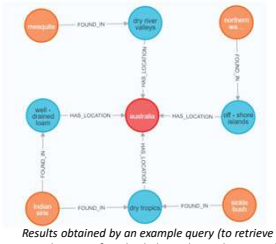
Amongst our fine-tuned models, DeBERTa obtained the best performance, with an F1-score of 84.18%. This is comparable and competitive with that obtained by a BERT model pretrained on domain-specific data.

F1-scores obtained by the BERT-based NER models on the COPIOUS test set.

NE Type	DistilBERT	ALBERT	BERT-base	RoBERTa	DeBERTa	BiodivBERT
Taxon	85.59	83.64	85.72	86.11	**87.60**	86.81
Geographic Location	85.62	84.16	86.74	**87.85**	87.58	86.74
Temporal Expression	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

Knowledge Graph Curation

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.



Results obtained by an example query (to retrieve species that were found in habitats located in Australia) that was run on the populated knowledge graph.

Conclusions

We demonstrated that general-domain, enhanced variants of the original BERT language model that were fine-tuned for the NER task on the COPIOUS corpus obtained performance that is comparable and competitive with that obtained by a BERT model pretrained on domain-specific data.

[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.