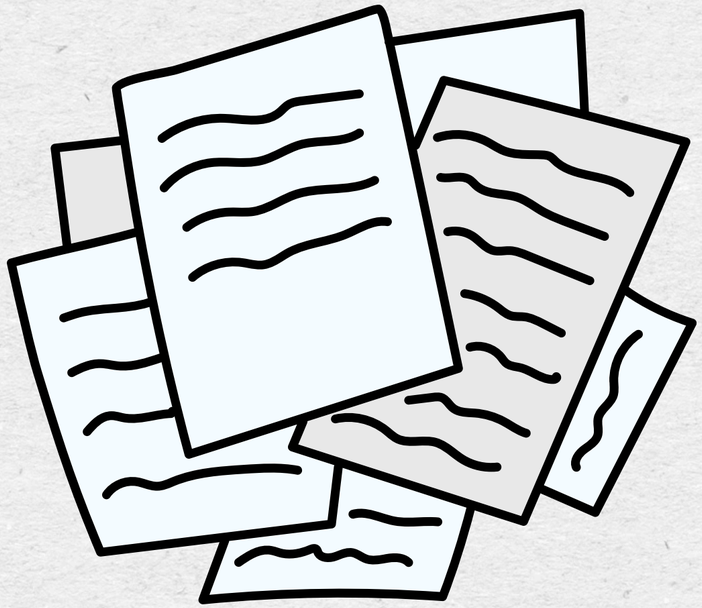
Towards the Curation of Environment-related Knowledge Graphs: Fine-tuning General-domain Language Models for Biodiversity Named Entity Recognition

Geilah T. Tabanao, Andrew Miguel V. Pagdanganan,
Riza Batista-Navarro, Roselyn S. Gabud

Biodiversity Textual Documents



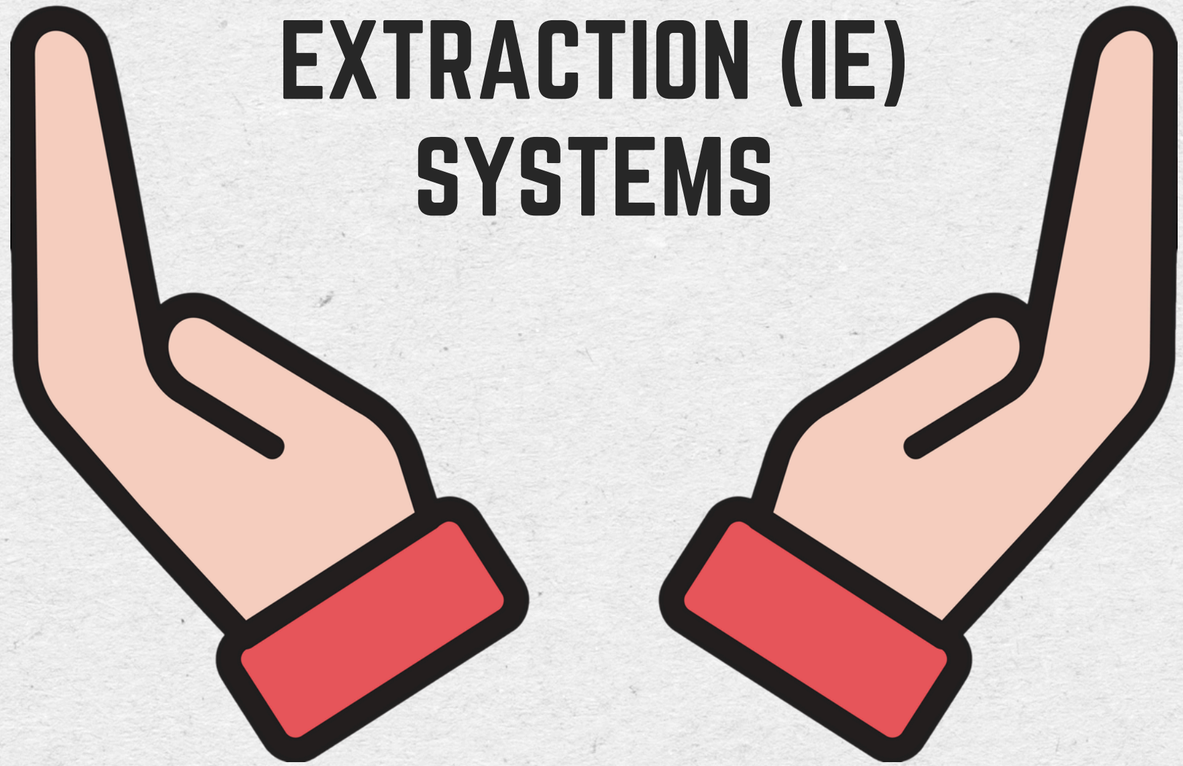
Contains bulk of knowledge on biodiversity

Millions are published

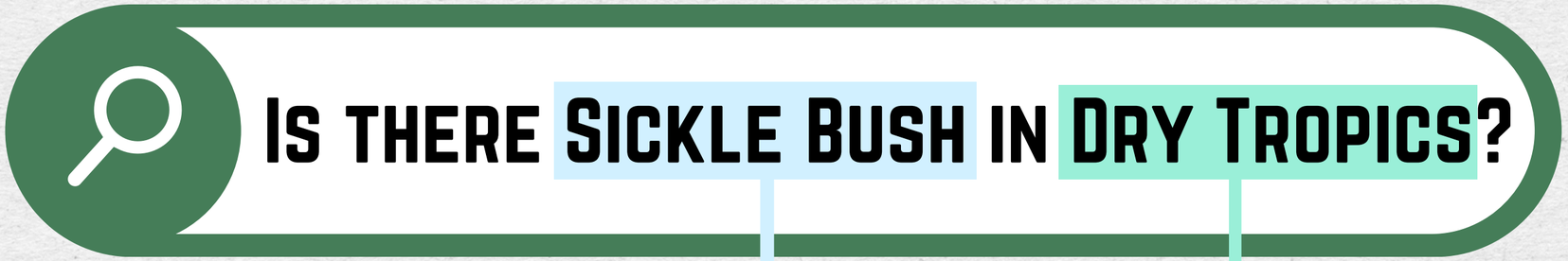
DATA WOULD HELP POLICY MAKERS FOR EFFECTIVE CLIMATE ACTION

HOW DO WE MAKE IT A STRUCTURED, ACCESSIBLE DATA?

THE USE OF INFORMATION EXTRACTION (IE) SYSTEMS



NAMED ENTITY RECOGNITION



RELATION EXTRACTION

OBJECTIVES

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

2. Employ best performing fine-tuned NER model to an IE pipeline

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

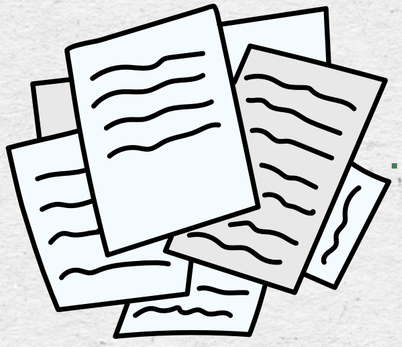
[1] Nhung T.H. Nguyen, Roselyn S. Gabud, and Sophia Ananiadou. COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. Biodiversity Data Journal, (7):e29626, January 2019. ISSN 1314-2828. doi: 10.3897/BDJ.7.e29626. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351503/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

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

- DeBERTa obtained the best performance, with an F1-score of 84.18%.

PIPELINE



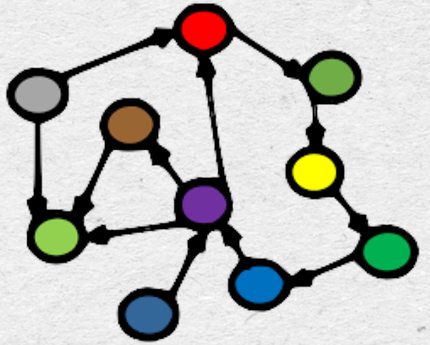
Corpus from CABI textual descriptions

NER

Fine-tuned DeBERTa NER model

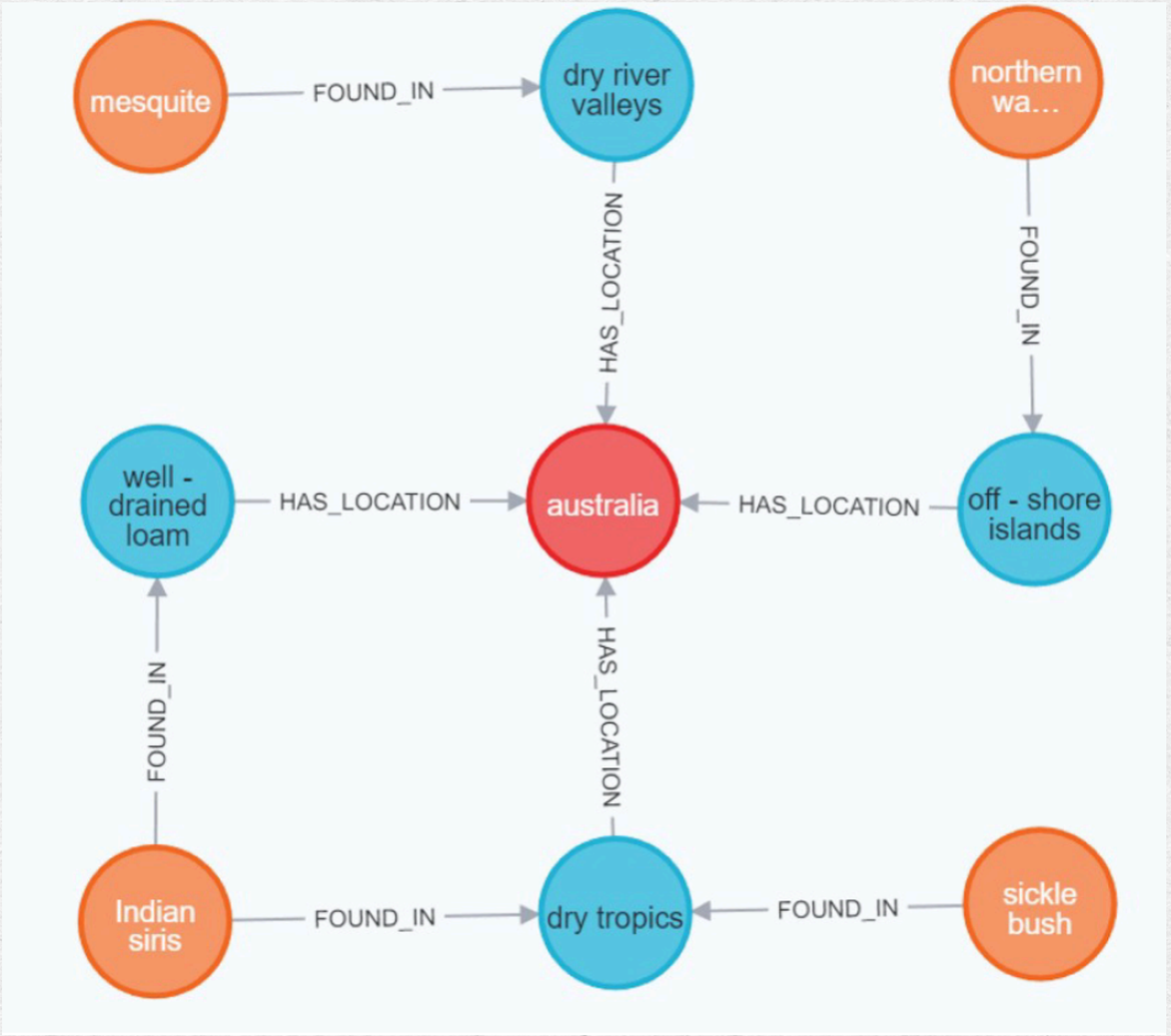
RE

Relation Extraction Model [1]



Neo4j Framework

[1] Roselyn Gabud, Portia Lapitan, Vladimir Mariano, Eduardo Mendoza, Nelson Pampolina, Maria Art Antonette Clariño, and Riza Batista-Navarro. A Hybrid of Rule-based and Transformer-based 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


mesquite
FOUND_IN
dry river valleys
northern wa…
HAS_LOCATION
FOUND_IN
well - drained loam
HAS_LOCATION
australia
HAS_LOCATION
off - shore islands
FOUND_IN
HAS_LOCATION
Indian siris
FOUND_IN
dry tropics
FOUND_IN
sickle bush

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.

Thank You!