Relation Extraction From Unstructured Species Descriptions Using TaxonNERD and LLaMA 2 7B

Fabricio De Jesus Rios Montero[‡], Ervin Rodríguez[§], Maria Mora Cross[§]

‡ Computer science, Heredia, Costa Rica
§ Computer science, Alajuela, Costa Rica

Corresponding author: Fabricio De Jesus Rios Montero (fabrirm15@gmail.com)

Abstract

Ontologies are essential tools for organizing information on taxonomy, ecology, and inter-species relationships, helping to standardize ecological data and facilitate integration of large datasets. Combining ontologies with advanced Natural Language Processing (NLP) techniques, such as Named Entity Recognition (NER) and Relation Extraction (RE), has greatly improved the discovery of insights from unstructured scientific texts, particularly in biodiversity (Gabud et al. 2023, Abdelmageed et al. 2022, Hearst 1992).

This study combines ontologies and NLP to analyze complex trophic interactions among animal species (Gabud et al. 2023), using a dataset (National Biodiversity Institute of Costa Rica (INBio) 2015) containing species descriptions in English and Spanish. We applied TaxoNERD to identify taxonomic entities (Le Guillarme and Thuiller 2021) and we fine-tuned the Large Language Model Meta AI (LLaMA 2 7B) to extract feeding interactions and predator-prey relationships (CheeKean 2023), due to its effectiveness in handling complex language patterns and its adaptability to diverse scientific domains.

Our results (Fig. 1) showed a recall of 0.73 and a precision of 0.68, indicating that the model effectively identifies feeding relationships in most cases. However, the lower precision suggests that the model may still capture some unrelated interactions, highlighting an area for improvement to reduce false positives and increase accuracy (Touvron et al. 2023). Previous studies also emphasize the need for further refinement of relation extraction models to enhance accuracy (Mora-Cross et al. 2023). The structured dataset offers valuable insights into species' diets and roles, contributing to biodiversity research and conservation efforts (Mora-Cross et al. 2023, Touvron et al. 2023).

Moreover, this research highlights the potential of integrating AI-driven tools with ontological frameworks to manage and analyze biodiversity data at scale (Abdelmageed et al. 2022). By transforming unstructured text into structured data, we make ecological information more accessible, supporting better decision-making in conservation strategies (Abdelmageed et al. 2022, Hearst 1992). This approach scales well with the growing volume of biodiversity data, offering a more efficient and accurate method for

analyzing species interactions, which are crucial for ecosystem management and endangered species protection (Gabud et al. 2023).

Keywords

biodiversity, ontologies, Named Entity Recognition (NER), Relation Extraction (RE), LLaMA2-7b, feeding relationships

Presenting author

Fabricio Ríos Montero

Presented at

SPNHC-TDWG 2024

Acknowledgements

This work was made possible through the support provided by the Instituto Nacional de Costa Rica (ITCR), the International Development Research Center (IDRC) through the Central American Higher University Council (CSUCA), and the Costa Rican Innovation and Research Promoter of the Ministry of Science, Innovation, Technology, and Telecommunications (MICITT) of Costa Rica.

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
- CheeKean (2023) Understanding Llama2: KV Cache, Grouped Query Attention, Rotary Embedding and More. https://plainenglish.io/community/understanding-llama2-kv-cache-grouped-query-attention-rotary-embedding-and-more-9a79bd
- Gabud R, Lapitan P, Mariano V, Mendoza E, Pampolina N, Clariño MAA, Batista-Navarro R, et al. (2023) A Hybrid of Rule-based and Transformer-based Approaches for Relation Extraction in Biodiversity Literature. Proceedings of the 2nd Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning103-113. https://doi.org/10.18653/v1/2023.pandl-1.10

- Hearst MA, et al. (1992) Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th conference on Computational linguistics - 2 https://doi.org/10.3115/992133.992154
- Le Guillarme N, Thuiller W, et al. (2021) TaxoNERD: Deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. Methods in Ecology and Evolution 13 (3): 625-641. https://doi.org/10.1111/2041-210x.13778
- Mora-Cross M, Ulate W, Retana Chacón B, Biarreta Portillo M, Castro Ramírez JD, Chavarria Madriz J, et al. (2023) Structuring Information from Plant Morphological Descriptions using Open Information Extraction. Biodiversity Information Science and Standards 7 https://doi.org/10.3897/biss.7.113055
- National Biodiversity Institute of Costa Rica (INBio) (2015) *Atta*: Species Records from Costa Rica Documented between 1999 and 2015 [Dataset]. National Biodiversity Institute of Costa Rica. URL: https://docs.google.com/spreadsheets/d/1EVIjVQYE7gw5m4uGPWrn6rXIgcnXLkqr/edit?usp=sharing&ouid=112437040868151967020&rtpof=true&sd=true
- Touvron H, Bojanowski P, Caron M, Cord M, El-Nouby A, Grave E, Izacard G, Joulin A, Synnaeve G, Verbeek J, Jégou H, et al. (2023) ResMLP: Feedforward Networks for Image Classification With Data-Efficient Training. IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (4): 5314-5321. https://doi.org/10.1109/tpami.2022.3206148

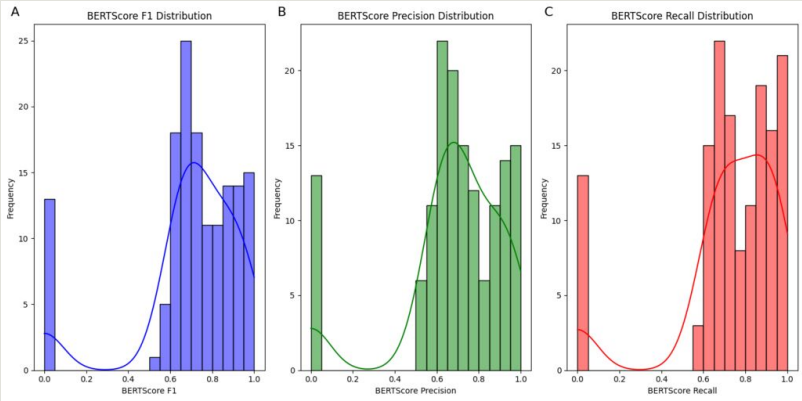

Figure 1.

Distribution of BERTScore (Bidirectional Encoder Representations from Transformers) metrics (F1, Precision, Recall) with most scores between 0.6 and 0.9. Outliers in Precision suggest areas for improving accuracy and reducing false positives.