Extracting Reproductive Condition and Habitat Information from Text Using a Transformer-based Information Extraction Pipeline

Roselyn Gabud[‡,§], Nelson Pampolina[§], Vladimir Mariano[|], Riza Batista-Navarro[¶,§]

‡ University of the Philippines Diliman, Quezon City, Philippines
§ University of the Philippines Los Baños, Los Baños, Philippines
| Fulbright University Vietnam, Ho Chi Minh, Vietnam
¶ University of Manchester, Manchester, United Kingdom

Corresponding author: Roselyn Gabud (rsgabud@up.edu.ph)

Abstract

Understanding the biology underpinning the natural regeneration of plant species in order to make plans for effective reforestation is a complex task. This can be aided by providing access to databases that contain long-term and wide-scale geographical information on species distribution, habitat, and reproduction. Although there exists widely-used biodiversity databases that contain structured information on species and their occurrences, such as the Global Biodiversity Information Facility (GBIF) and the Atlas of Living Australia (ALA), the bulk of knowledge about biodiversity still remains embedded in textual documents. Unstructured information can be made more accessible and useful for large-scale studies if there are tools and services that automatically extract meaningful information from text and store it in structured formats, e.g., open biodiversity databases, ready to be consumed for analysis (Thessen et al. 2022).

We aim to enrich biodiversity occurrence databases with information on species reproductive condition and habitat, derived from text. In previous work, we developed unsupervised approaches to extract related habitats and their locations, and related reproductive condition and temporal expressions (Gabud and Batista-Navarro 2018). We built a new unsupervised hybrid approach for relation extraction (RE), which is a combination of classical rule-based pattern-matching methods and transformer-based language models that framed our RE task as a natural language inference (NLI) task. Using our hybrid approach for RE, we were able to extract related biodiversity entities from text even without a large training dataset.

In this work, we implement an information extraction (IE) pipeline comprised of a named entity recognition (NER) tool and our hybrid relation extraction (RE) tool. The NER tool is a transformer-based language model that was pretrained on scientific text and then fine-tuned using COPIOUS (Conserving Philippine Biodiversity by Understanding big data;

Nguyen et al. 2019), a gold standard corpus containing named entities relevant to species occurrence. We applied the NER tool to automatically annotate geographical location, temporal expression and habitat information contained within sentences. A dictionary-based approach is then used to identify mentions of reproductive conditions in text (e.g., phrases such as "fruited heavily" and "mass flowering"). We then use our hybrid RE tool to extract *reproductive condition - temporal expression* and *habitat - geographical location* entity pairs. We test our IE pipeline on the forestry compendium available in the CABI Digital Library (Centre for Agricultural and Biosciences International), and show that our work enables the enrichment of descriptive information on reproductive and habitat conditions of species. This work is a step towards enhancing a biodiversity database with the inclusion of habitat and reproductive condition information extracted from text.

Keywords

relation extraction, biodiversity

Presenting author

Roselyn Gabud

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Gabud R, Batista-Navarro R (2018) Extracting granular information on habitats and reproductive conditions of Dipterocarps through pattern-based literature analysis. In ICEI 2018: 10th International Conference on Ecological Informatics-Translating Ecological Data into Knowledge and Decisions in a Rapidly Changing World.
- Nguyen N, Gabud R, Ananiadou S (2019) COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. Biodiversity Data Journal 7 https://doi.org/10.3897/bdj.7.e29626
- Thessen A, Mozzherin D, Shorthouse D, Patterson D (2022) Improving the discoverability of biodiversity data using the Global Names Finder. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.90026