BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 7: e113055
doi: 10.3897/biss.7.113055

OPEN ACCESS

Conference Abstract

Structuring Information from Plant Morphological Descriptions using Open Information Extraction

Maria Auxiliadora Mora-Cross[‡], William Ulate[§,|], Brandon Sthuar Retana Chacón[‡], María Fernanda Biarreta Portillo[‡], Josué David Castro Ramírez[‡], Jose Alejandro Chavarria Madriz[¶]

‡ ITCR, Alajuela, Costa Rica
§ CRBio, Heredia, Costa Rica
| Missouri Botanical Garden, Saint Louis, MO, United States of America
¶ ITCR, Cartago, Costa Rica

Corresponding author: Maria Auxiliadora Mora-Cross (mariamoracross@gmail.com)

Received: 20 Sep 2023 | Published: 21 Sep 2023

Citation: Mora-Cross MA, Ulate W, Retana Chacón BS, Biarreta Portillo MF, Castro Ramírez JD, Chavarria Madriz JA (2023) Structuring Information from Plant Morphological Descriptions using Open Information Extraction. Biodiversity Information Science and Standards 7: e113055. https://doi.org/10.3897/biss.7.113055

Abstract

Taxonomic literature keeps records of the planet's biodiversity and gives access to the knowledge needed for research and sustainable management. The number of publications generated is quite large: the corpus of biodiversity literature includes tens of millions of figures and taxonomic treatments. Unfortunately, most of the taxonomic descriptions are from scientific publications in text format. With more than 61 million digitized pages in the Biodiversity Heritage Library (BHL), only 467,265 taxonomic treatments are available in the Biodiversity Literature Repository. To obtain highly structured texts from digitized text has been shown to be complex and very expensive (Cui et al. 2021). The scientific community has described over 1.2 million species, but studies suggest that 86% of existing species on Earth and 91% of species in the ocean still await description (Mora et al. 2011). The published descriptions synthesize observations made by taxonomists over centuries of research and include detailed morphological aspects (i.e., shape and structure) of species useful to identify specimens, to improve information search mechanisms, to perform data analysis of species having particular characteristics, and to compare species descriptions.

To take full advantage of this information and to work towards integrating it with repositories of biodiversity knowledge, the biodiversity informatics community first needs to

convert plain text into a machine-processable format. More precisely, there is a need to identify structures and substructure names and the characters that describe them (Fig. 1).

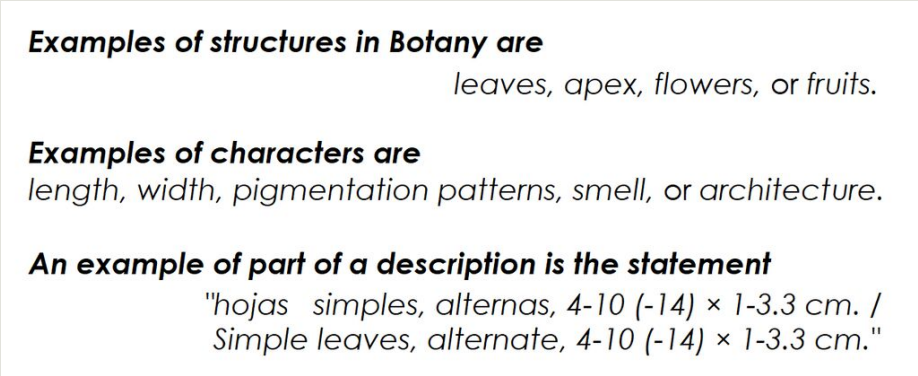

Figure 1.

Morphological descriptions may describe structures, substructures, characters, states, and relationships between structures (Mora and Araya 2018).

Open information extraction (OIE) is a research area of Natural Language Processing (NLP), which aims to automatically extract structured, machine-readable representations of data available in unstructured text; usually the result is handled as n-ary propositions, for instance, triples of the form <noun phrase, relation phrase, noun phrase> (Shen et al. 2022).

OIE is continuously evolving with advancements in NLP and machine learning techniques. The state of the art in OIE involves the use of neural approaches, pre-trained language models, and integration of dependency parsing and semantic role labeling. Neural solutions mainly formulate OIE as a sequence tagging problem or a sequence generation problem. Ongoing research focuses on improving extraction accuracy; handling complex linguistic phenomena, for instance, addressing challenges like coreference resolution; and more open information extraction, because most existing neural solutions work in English texts (Zhou et al. 2022).

The main objective of this project is to evaluate and compare the results of automatic data extraction from plant morphological descriptions using pre-trained language models (PLM) and a language model trained on data from plant morphological descriptions written in Spanish.

The research data for this study were sourced from the species records database of the National Biodiversity Institute of Costa Rica (INBio). Specifically, the project focused on selecting records of morphological descriptions of plant species written in Spanish.

The system processes the morphological descriptions using a workflow that includes phases like data selection and pre-processing, feature extraction, test PLM, local language

model training, and test and evaluate results. Fig. 2 shows the general workflow used in this research.

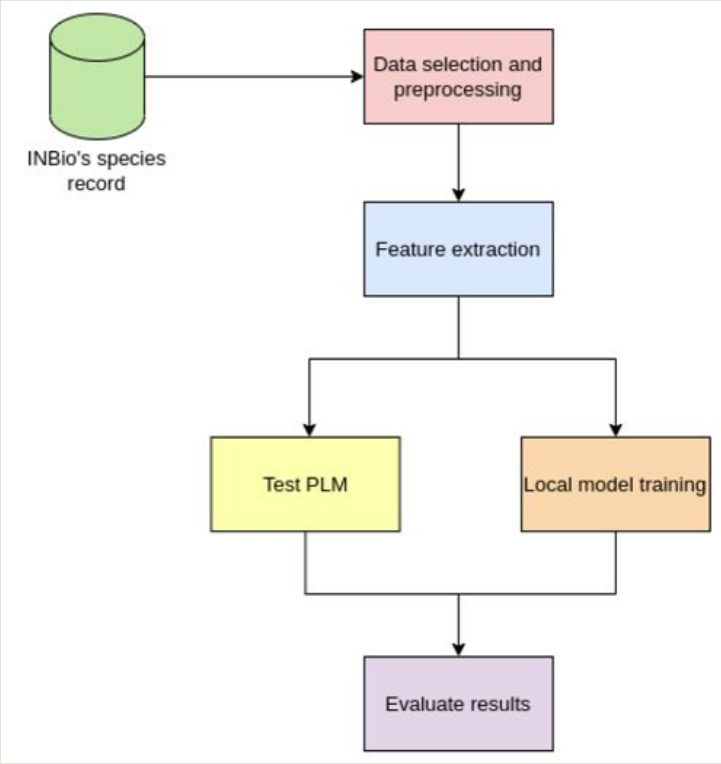

Figure 2.

Workflow of implemented algorithm.

Pre-processing and Annotation: Descriptions were standardized by removing special characters like double and single quotes, replacing abbreviations, tokenizing text, and other transformations.

Some records of the dataset were annotated with the ground-truth structured information in the form of triples that were extracted from each paragraph. Additionally, structured data from the project carried out by Mora and Araya (Mora and Araya 2018) were included in the dataset.

Feature extraction: The token vectorization was done using word embedding directly by the language models.

Test PLM: The evaluation process of PLM models used the zero-shot approach and involved applying the models to the test dataset, extracting information, and comparing it to annotated ground truth.

Local Language Model Training: The annotated data was split into 80% training data and 20% test data. Using the training data, a language model based on the Transformers architecture was trained.

Evaluate results: Evaluation metrics such as precision, recall, and F1 (a meaure of the model's accuracy) were calculated comparing the extracted information and the ground truth. The results were analyzed to understand the models' performance, identify strengths and weaknesses, and gain insights into their ability to extract accurate and relevant information. Based on the analysis, the evaluation process iteratively improved models results.

The main contributions of this project are:

- A Transformers-based language model to extract information from morphological descriptions of plants written in Spanish available on the project website.*[1]
- A corpus of morphological descriptions of plants, written in Spanish, labeled for information extraction, and made available on the project website.
- The results of the project, the first of its kind applied to morphological descriptions of plants written in Spanish, published on the project website.

Keywords

deep learning, natural language processsing, Spanish, large language models

Presenting author

María Auxiliadora Mora Cross

Presented at

TDWG 2023

Acknowledgements

The authors would like to thank ITCR's support to this project.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Cui H, Ford B, Starr J, et al. (2021) Author-Driven Computable Data and Ontology Production for Taxonomists. Biodiversity Information Science and Standards 5 https://doi.org/10.3897/biss.5.75741
- Mora C, Tittensor D, Adl S, Simpson AB, Worm B (2011) How Many Species Are There on Earth and in the Ocean? PLoS Biology 9 (8). https://doi.org/10.1371/journal.pbio.1001127
- Mora M, Araya J (2018) Semi-automatic Extraction of Plants Morphological Characters from Taxonomic Descriptions Written in Spanish. Biodiversity Data Journal 6 https://doi.org/10.3897/bdj.6.e21282
- Shen W, Yang Y, Liu Y (2022) Multi-View Clustering for Open Knowledge Base Canonicalization. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining https://doi.org/10.1145/3534678.3539449
- Zhou S, Yu B, Sun A, Long C, Li J, Sun J (2022) A Survey on Neural Open Information Extraction: Current Status and Future Directions. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence https://doi.org/10.24963/ijcai.2022/793

Endnotes

[*1] https://github.com/colibri-itcr