CS198 Project Proposal

Jeezer Niño D. Lee
Computer Vision and Machine Intelligence Group
Department of Computer Science
College of Engineering
University of the Philippines Diliman

I. Project Title

- Location-Specific Information Extraction from Unstructured Biodiversity Literature

II. Project Implementor

- Jeezer Niño D. Lee

III. Project Advisors

- Roselyn Gabud, MSc

IV. Significance of the Project

Biodiversity data encompasses a diverse array of interconnected information types, each with distinct definitions and complex relationships. Generally, biodiversity consists of terms about the variety of life from species to ecosystems [7]. Biodiversity data also captures the interconnected roles of environmental factors, such as climate, geographic location, and habitat conditions, which influence the distribution and survival of species. This information is important as it raises awareness about ecosystem dynamics along with monitoring purposes, enabling scientists to track changes in biodiversity and assess the impacts of climate change, habitat loss, and conservation efforts on various life forms.

Relation extraction (RE), a subtask of natural language processing (NLP), provides an effective method for addressing the challenges of extracting and organizing such significant terms from extensive biodiversity literature. This technique automates the identification of relationships between entities [8], such as location and its habitats and inhabiting species, enabling researchers to efficiently mine relevant information from text-heavy resources. This task is beneficial in a wide range of fields such as biodiversity, where there is a great wealth of scattered or unstructured literature. By leveraging relation extraction tools, researchers can compile data more effectively, contributing to the development of comprehensive databases that support scientific research, policymaking, and conservation strategies. With this, several models are to be explored in a supervised approach concerning relation extraction, as elaborated by the following literature review.

V. Problem Statement and Objectives

Previous studies dealt with information extraction on Unstructured Biodiversity data, where most have explored NER and several attempts on RE. Additionally, most Information Extraction tasks involving Machine and deep learning primarily focus on other specific data, such as species occurrence extraction and its related components. Given the lack of annotated datasets for RE is a challenge that initially obstructs further exploration and experimentation with the subject.

Hence, the project mainly aims to extract entities and their relations by training and fine-tuning several deep-learning and pre-trained models. Such associations are centered around the location entity, where the project targets the extraction of desired location-based descriptions concerning Location-Taxon and Location-Habitat relations. Also employing a supervised approach, the objectives of this study are outlined as follows:

- Annotate existing NER-labeled biodiversity corpus for supervised relation extraction task
- Train and explore various RE models given the annotated dataset
- Develop an insightful visual diagram that utilizes such entities and relations

VI. Review of Related Literature

Relation Extraction Fine-tuning

SpanBERT

SpanBERT is a variation of BERT that focuses on predicting contiguous spans instead of individual tokens. SpanBERT's unique span boundary objective (SBO) enables the model to learn representations better suited for span-based reasoning tasks, such as relation extraction. SpanBERT demonstrated superior performance over BERT on span-centric tasks, achieving new state-of-the-art results on the TACRED dataset for relation extraction and notable improvements in question answering (SQuAD) and coreference resolution. Its span-focused approach is particularly effective in handling entity relationships, where spans rather than individual tokens are central. This model shows potential in obtaining a better RE performance given its span-based pre-training approach. [4]

RoBERTa

RoBERTa, which was developed to improve BERT's pre-training approaches. This mode was pre-trained on a larger corpus, approx. 160Gb of data compared to BERT with 16 Gb with longer sequences while also disregarding the next sentence prediction aspect of vanilla BERT. This model is trained with dynamic masking and larger mini-batches, with notable benchmarks for NLP tasks like Question Answering and Natural Language Understanding [6]. For consistency, the same procedure used from Span BERTis implemented here for fine-tuning– substituting the specific entity for its type during preprocessing. This seems to be the most common approach with RE fine-tuning, given that another biodiversity dataset, BiodivNERE, follows the same masking format for its RE dataset [2].

Annotated Datasets for Biodiversity Literature

COPIOUS

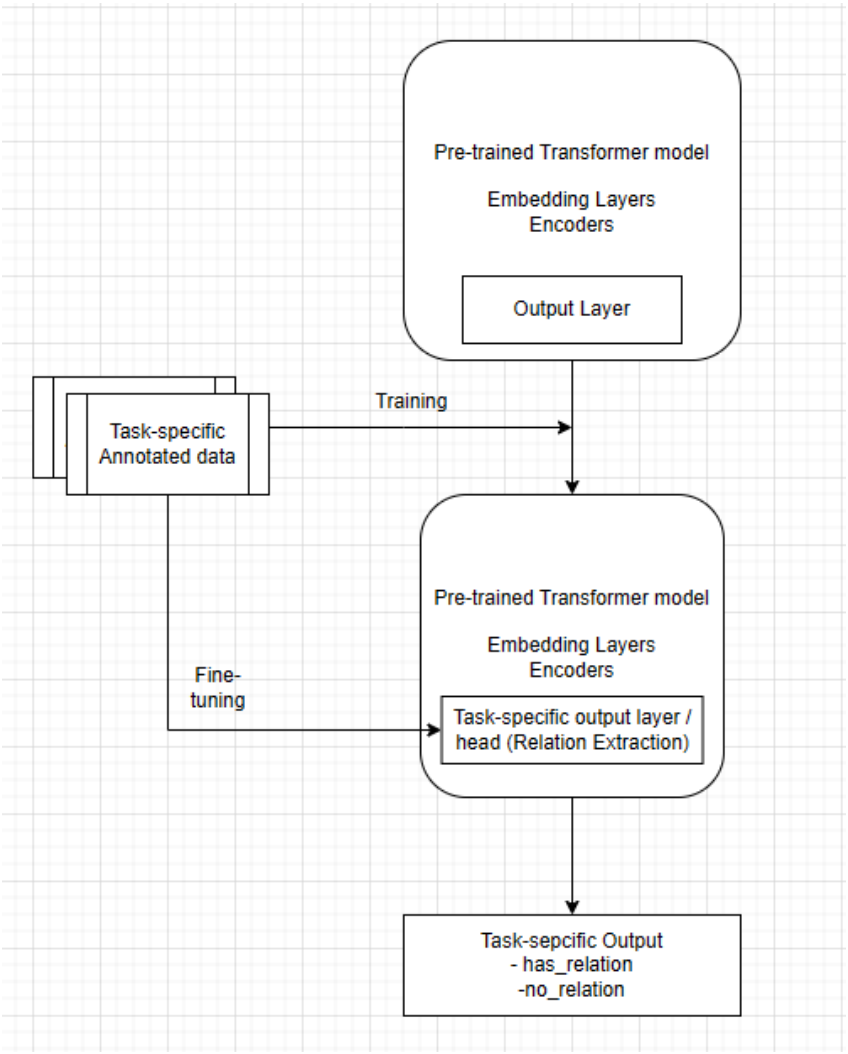
Nguyen et al. present COPIOUS, a gold-standard corpus specifically designed to extract species occurrences from biodiversity literature. Primarily aligned for entity extraction tasks, the corpus includes manually annotated entities such as taxon names, geographical locations, habitats, temporal expressions, and person names. COPIOUS integrates multiple entity types relevant to biodiversity, making it suitable for extracting complex relations between species and their environments. Entity relations, such as Taxon-GeographicalLocation and Taxon-Habitat entity pairs, are extracted using a pattern-based RE system. [1]

In addition, a study by Gabud et al. focuses on a tropical tree family, *Dipterocarpaceae*, known for its ecological and reproductive significance in the Philippines and Southeast Asia, which used a subset of the COPIOUS corpus for its data. The subset of 151 manually selected documents is made up of brief excerpts from journals related to environmental science and ecology, such as the *Journal of Tropical Ecology*, offering insights into species' habitats, geographic locations, reproductive conditions, and temporal expressions. Relation extraction was modeled in a zero-shot learning context instead of common approaches using pre-existing labeled training data, focusing on *has_time* and *has_location* relation links between Reproductive_Condition-Temporal_ Expression and Habitat-Geographic_Location entity pairs respectively. The nature of the subset is thus constrained but rich, allowing for fine-grained extraction of biodiversity-related relations while demonstrating the effectiveness of hybrid rule-based and transformer-based approaches [2].

BiodivNERE

In more recent years, Abdelmageed et al. developed *BiodivNERE*, a gold-standard corpus designed for both NER and RE in biodiversity research. The corpus is tailored to capture important relations between entities such as organisms, environmental concepts, geographical locations, and phenomena. Manual annotations of data were employed and verified by biodiversity experts, where entities relevant to ecological and environmental studies were determined and relations were determined from the existing ontology provided by *BiodivOnto*. Complex relationships such as *occur_in* (e.g., Organism in Environment) and *influence* (e.g., Organism influences Process) are identified, which are essential for understanding species interactions and environmental dependencies. The corpus is structured to aid in the development of NLP tools that can automatically analyze and process biodiversity data, providing a foundation for more intelligent biodiversity monitoring systems. Other significant entity labels include *Quality* and *Matter,* which pertain to data parameters, observation, and traits for the former and chemical, biological compounds, or elements for the latter. [2]. This dataset is essential as insights may be drawn apart from its approaches as this project suggests the annotation of relation labels for the former;y discussed corpus.

VII. Theoretical Framework



VIII. Proposed Methodology

- Dataset Collection
    - Dataset Preparation: COPIOUS
- Schema and Annotation Setup
    - Identification of entity types and labels
    - Definition of Relation Schema and labels
    - Annotator Orientation and Training
    - Annotation proper
- Label Validation
    - Consolidation of annotations & dataset splitting
    - Exploratory data analysis of the finalized annotated dataset
- Model Development and Implementation
    - Selection of NER models for entity extraction
    - Data Preprocessing for RE
    - Model Training for RE (for each suggested model)
- Evaluation and Validation
    - Cross-validation
    - Performance metrics
        - Precision
        - Recall
        - F1-score

IX. Major Activities

- Construction of an annotated dataset
    - Corpus Creation, Data Collection & Literature Review:
        - Dataset Preparation
    - Schema Development & Annotation:
        - Formulate annotation guidelines
        - Annotator Orientation and Training
        - Annotation Proper
        - Consolidation, analysis, and splitting of annotated data.
- Writing I
    - Conference Paper Writing
- Model Implementation
    - NER stage
        - Data Pre-processing for NER
        - Model Training
    - RE stage
        - Data Pre-processing for RE
        - Training of RE Models
- Model Testing & Evaluation:
    - Cross-validation techniques
    - evaluation metrics such as precision, recall, and F1-score.
    - hyperparameter tuning and adjustments
- Writing II
    - Conference Paper Writing
    - Thesis Writing
    - Finalize Thesis Writing
    - Revise Thesis

X. Resources Needed

Data:

- COPIOUS (with RE labels)

Software:

- Python, Hugging Face Transformers
- TensorFlow/PyTorch.
- NER/RE evaluation librarie - scikit-learn

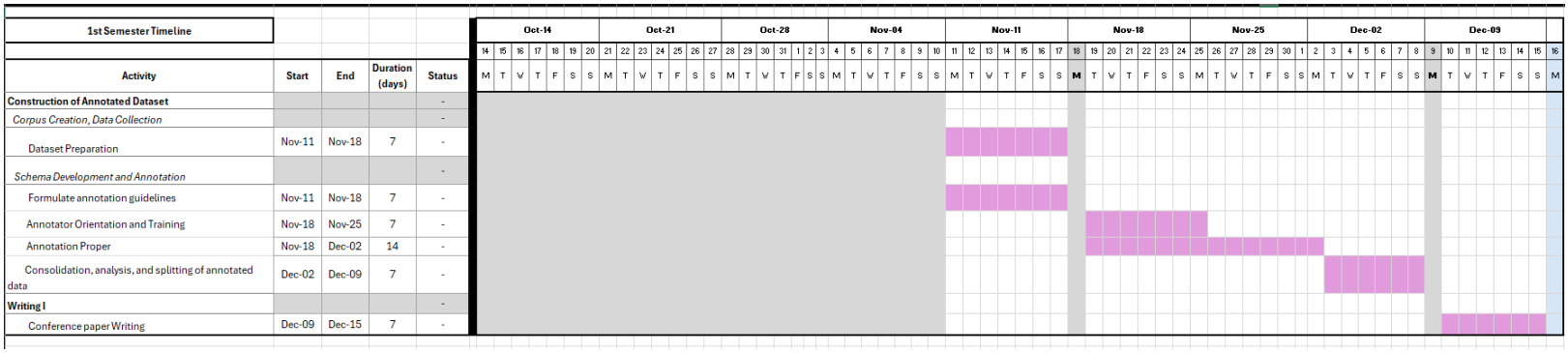
Hardware:

- High-performance GPU machine for RE training.

Human Resources:

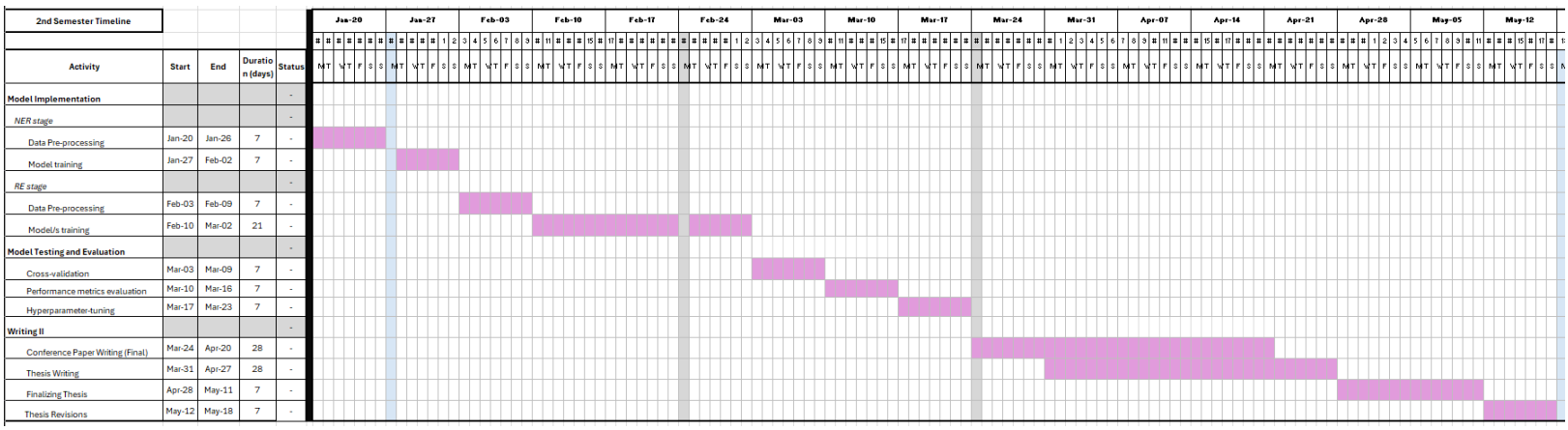
- Trained annotators.

XI. Timeline (1st and 2nd sem) – Gantt Chart



1st Semester Timeline

Activity	Start	End	Duration (days)	Status
Construction of Annotated Dataset				-
Corpus Creation, Data Collection				-
Dataset Preparation	Nov-11	Nov-18	7	-
Schema Development and Annotation				-
Formulate annotation guidelines	Nov-11	Nov-18	7	-
Annotator Orientation and Training	Nov-18	Nov-25	7	-
Annotation Proper	Nov-18	Dec-02	14	-
Consolidation, analysis, and splitting of annotated data	Dec-02	Dec-09	7	-
Writing I				-
Conference paper Writing	Dec-09	Dec-15	7	-



2nd Semester Timeline

Activity	Start	End	Duration (days)	Status
Model Implementation				-
NER stage				-
Data Pre-processing	Jan-20	Jan-26	7	-
Model training	Jan-27	Feb-02	7	-
RE stage				-
Data Pre-processing	Feb-03	Feb-09	7	-
Model/s training	Feb-10	Mar-02	21	-
Model Testing and Evaluation				-
Cross-validation	Mar-03	Mar-09	7	-
Performance metrics evaluation	Mar-10	Mar-16	7	-
Hyperparameter-tuning	Mar-17	Mar-23	7	-
Writing II				-
Conference Paper Writing (Final)	Mar-24	Apr-20	28	-
Thesis Writing	Mar-31	Apr-27	28	-
Finalizing Thesis	Apr-28	May-11	7	-
Thesis Revisions	May-12	May-18	7	-

XII. References

[1] Nguyen, N. T. H., Gabud, R. S., & Ananiadou, S. (2019). COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. Biodiversity Data Journal, 7, e29626. https://doi.org/10.3897/BDJ.7.e29626

[2] Abdelmageed, N., Löffler, F., Feddoul, L., Algergawy, A., Samuel, S., Gaikwad, J., Kazem, A., & König-Ries, B. (2022). BiodivNERE: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain. Biodiversity Data Journal, 10, e89481. https://doi.org/10.3897/BDJ.10.e89481

[3] Gabud, R. S., Lapitan, P., Mariano, V., Mendoza, E., Pampolina, N., Clariño, M. A. A., & Batista-Navarro, R. T. (2023). A hybrid of rule-based and transformer-based approaches for relation extraction in biodiversity literature. Proceedings of the 2nd Workshop in Pattern-Based Approaches to NLP in the Age of Deep Learning, 103–113. Association for Computational Linguistics.

[4] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving Pre-training by Representing and Predicting Spans," arXiv preprint, arXiv:1907.10529, 2020.

[5] S. Wadhwa, S. Amir, and B. C. Wallace, "Revisiting Relation Extraction in the Era of Large Language Models," arXiv preprint, arXiv:2404.12345, 2024.

[6] Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint, arXiv:1907.11692, 2019.

[7] “Biodiversity,” *Stanford Encyclopedia of Philosophy*, 2022. [Online]. Available: https://plato.stanford.edu/entries/biodiversity/

[8] Zhao, X., Deng, Y., Yang, M., Wang, L., Zhang, R., Cheng, H., Lam, W., Shen, Y., & Xu, R, A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, 56(11), Article 293, 2024. https://doi.org/10.1145/3674501:contentReference[oaicite:1]{index=1}.