Information Extraction of Location-Specific Data from Unstructured Biodiversity Literature

Jeezer Niño D. Lee,
Roselyn S. Gabud
Computer Vision & Machine Intelligence Group
Department of Computer Science
College of Engineering
University of the Philippines-Diliman

ABSTRACT

This study explores the extraction of location-specific relationships from unstructured biodiversity literature, specifically interested in the relations between geographical locations, species, and habitats. While named entity recognition (NER) strategies in biodiversity are well-established, rela- tion extraction (RE) remains to be further explored largely due to limited annotated datasets. This study fine-tunes two transformer models, SpanBERT and RoBERTa, with the goal of improving such models for domain-specific RE by extending the COPIOUS corpus with relation labels. With an F1 score of up to 69.49%, the results evaluation indicates that SpanBERT performs better than RoBERTa, demonstrating its potential for specialized biodiversity data extraction. These lay the groundwork for applications in biodiversity conservation and monitoring, with further work aiming for more precise data annotation quality and reliable model training.

1. INTRODUCTION

1.1 Problem Statement

Over the years, various natural language processing (NLP) methods have been developed to extract structured information from unstructured biodiversity data. These popular methods, namely Named Entity Recognition (NER) and Relation Extraction (RE), have been very reliable for identifying key concepts on such works - such as species, locations, and habitats. While NER tools are well-developed and studied, existing methods still struggle to accurately capture relationships between these entities, especially in biodiversity-specific contexts [5]. Previous approaches aimed at investigating relation extraction (RE) tasks on biodiversity literature commonly fall on pattern and rule-based approaches [3], with the matter being underexplored due to the lack of labeled training data for the tasks. This then suggests the probable lack of precision commonly required for domain-specific applications. Additionally, existing systems have primarily focused on species and organisms, rather than other significant entities in biodiversity like location and its descriptions.

This paper aims to address these gaps by focusing on the extraction of essential relationships involving geographical places mentioned within biodiversity data. The existence of models trained for this specific task remains very few and insufficient, leaving an area of further inquiry. From this, approaching such objectives involves the acquisition of a well-annotated dataset along with training models capable of handling such complexities associated with the RE task for biodiversity literature. From these models, several applications can be made that could support the understanding of biodiversity, for instance, the visualization of organisms and habitats as seen in a biogeography map that complements biodiversity studies and provides insight to broader ecological interactions.

1.2 Significance of the Study

Extracting specific location-based relationships, such as the presence of species X in location Y or location Y having habitat Z, remains a significant challenge while also offering valuable data for further analysis. This is particularly true in biodiversity contexts, where the relationships between entities like location, taxon, and habitat are complex and unique to its boundaries. For example, knowing a location's profile of species and their habitats helps inform conservation strategies, as it allows researchers to identify areas at risk, track biodiversity changes, and prioritize conservation efforts. Such extracted location-specific data can aid in local biodiversity monitoring, enabling policymakers and conservationists to make grounded decisions concerning protected areas due to native species occurrence, ecosystem management. This study is motivated by such reasons, also aiming to extend its findings to more practical biodiversity-related applications.

1.3 Scope of the Study

This paper primarily focuses on employing a supervised learning approach by fine-tuning several pre-trained models - SpanBERT and RoBERTa – with an annotated dataset having both entity and relation labels. While limited to these specific relations, it aims to improve the precision and applicability of RE methods for extracting meaningful biodiversity information for further downstream tasks. The scope is also constrained by the size and diversity of the dataset, which may affect the model's generalizability, yet its focus on location-based entities and their relationships will provide valuable insights into automated biodiversity data extraction. A geographical visualization of biodiversity using the trained models is also another aspect which the project considers to be under its scope.

2. METHODOLOGY

The project's methodology is generally divided into two parts: (1) the acquisition of the essential labeled dataset followed by the (2) fine-tuning of desired transformer models. The practical implementation of integrating such fine-tuned models is to be further explored and executed in the project's future.

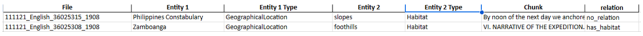
2.1 Dataset Acquisition

2.1.1 COPIOUS

This project utilizes COPIOUS, a gold-standard corpus specifically designed to extract species occurrences from biodiversity literature. Primarily intended for entity extraction tasks, the corpus includes manually annotated entities such as taxon names, geographical locations, habitats, temporal expressions, and person names. The corpus consists of 668 documents with the following entity count: 12,227 for *Taxon*, 12,227 for *Geographical Locations*, 2889 for *Person*, 2210 for *temporal expression*, and 1554 for *habitat.* The corpus integrates multiple entity types relevant to biodiversity, making it suitable for extracting complex relations between these identified species and their environments. Yet, COPIOUS does not have relation labels, which this project also needs to proceed with model training. Hence, the project also involves relation labeling, as it requires less expertise in validating its quality compared to entity labels which is within the available capabilities [1]

2.1.2 Dataset Preparation

The COPIOUS corpus consists of *.txt* and *.ann* files, containing the raw text and the entity labels on each of these separate files. From these, each pair is pre-processed and combined into a single *.xlsx* to be prepared for relation labeling. Each column of the *.xlsx* file consists of the following names: *file, entity id, entity type, span of the entity relative to the text, the entity text*, and the corresponding *sentence* where the particular entity occurs. To only observe the desired relations, the *person* and *temporal expression* entity types were disregarded, followed by the same large table to be further preprocessed - wherein two rows are merged given that they both have the same sentence field or column along with the condition that the individual sentence possesses at least two of the desired entities. This results to the final table with lesser records but additional two columns accounting for the second entity in the same sentence.



File	Entity 1	Entity 1 Type	Entity 2	Entity 2 Type	Chunk	relation
111121_English_36025315_1908	Philippines Constabulary	GeographicalLocation	slopes	Habitat	By noon of the next day we anchore	no_relation
111121_English_36025308_1908	Zamboanga	GeographicalLocation	foothills	Habitat	VI. NARRATIVE OF THE EXPEDITION.	has_habitat

Figure 1: sample rows of the annotated dataset

However, upon further inspection, the resulting *.xlsx* file contained no entries. Given this, the approach to the preparation was modified to include both the previous and next sentences of a specific sentence of interest - naming the three-sentence group as chunks. Hence, the earlier steps consisted of separating sentences in chunks (within the same file) and incorporating the same procedures afterwards. From these, a total of the desired entity pairings was obtained:

Dataset	Geolocation-Taxon	Geolocation-Habitat
train	2437	691
dev	106	30
test	313	64

Table 1: entity-pair count for COPIOUS

2.1.3 Relation Labeling

From the obtained relation pairings, three relation labels were utilized:

1. *has_taxon*: *GeographicalLocation-Taxon* entity pair with *valid* relations.
2. *has_habitat*: *GeographicalLocation-Habitat* entity pair with *valid* relations.
3. *no_relation*: *GeographicalLocation-Habitat* or *GeographicalLocation-Taxon* entity pair with *no valid* relation.

Two groups of students, with each having five members, were tasked to label each row by reading the given chunk and determining whether a valid relation between the two entities exists or not. The general guideline for annotating was determining whether a biological or ecological connection between the two entities exists. For example, the sentence, "Tilapia [Taxon] thrives in lakes near Manila [Geolocation]" implies an ecological connection between the two while "Tilapia [Taxon] sold in Manila [Geolocation] markets is imported from China" does not. Economic relations were also not considered as for some instances these do not have associations to biodiversity, such as, "Tilapia [Taxon] is a popular fish in Manila [Geolocation] markets.". The annotators were also instructed to solely base their labeling by inferring relation within the bounds of the chunk, which means that even though an instance occurs where the entity pair is determined to have a relation from general knowledge, they were advised to mark the row as no_relation. The setup for two groups was primarily to ensure data quality and consolidation given the broadness of the guidelines' interpretation. However, due to time constraints, these procedures would be implemented on a later date to observe and improve the training results.

2.2 Fine-Tuning for Relation Extraction

2.2.1 Preliminaries

The fine-tuning process of the succeeding models was accomplished using the Google Colab environment. The Tesla T4

GPU was primarily used given its well-suited capability for training large-scale transformer models. Additionally, some notable libraries that were also utilized in the pipeline are: (1) *Pandas* library for Dataframes, (2) *Dataset* library for a convenient dataset interface, (3) *Transformers* Library from *HuggingFace*, and (4) *Scikit-learn* library for its metrics in evaluations purposes.

2.2.2 SpanBERT

The chosen model 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 show potential in obtaining a better RE performance given its span-based pre-training approach, recording a maximum of 10 contiguous masked tokens during pre-training [4].

The procedure for fine-tuning the model follows from the same steps taken by its developers for RE training on the larger TACRED dataset. The input is preprocessed in a manner where the particular entity text is replaced by its entity type before proceeding to tokenization. For instance, the sentence "This feline is commonly found in Quezon City" with feline and Quezon City as Taxon and GeographicalLocation entities is replaced in the sentence by its entity type: This [Taxon] is commonly found in [GeographicalLocation]. This method was proposed as it allows the model to generalize better by learn the context of both entities -given the surrounding spans of words - to determine its relation rather than focus the attention on the specific entity itself, somewhat involving masks in the process.

The parameters for fine-tuning the pre-trained SpanBERT model is of as follows:

- Learning rate: *1e-6, 1e-5, 2e-5, 3e-5, 5e-5*
- Batch size: $16, 32$
- Number of epochs: 8
- Max length of input sequence: 128

These parameters are based on the implementations of SpanBERT's researchers themselves, only differing in the number of epochs which originally was 10 and reduced due to time and resource limitations

2.2.3 RoBERTa

Another model of interest is RoBERTa, which was developed improving 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 SpanBERT is 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].

	no_relation	*has_taxon*	*has_habitat*
train	1856	882	383
dev	54	57	25
test	171	159	47

Table 2: Summary of relation labels

Table 3: Base performance of both models for RE

Model	Accuracy	F1	Precision	Recall
Spanbert	0.397878	0.341392	0.378337	0.331494
Roberta	0.517241	0.316625	0.347314	0.384531

Similar parameters are also being experimented with RoBERTa in which the results are to be further discussed in the following section. The parameters for fine-tuning the pre-trained RoBERTa model is the same as that of SpanBERT for comparison purposes.

3. RESULTS AND DISCUSSION

3.1 Evaluation of Resulting Annotations

Given that only one group have accomplished the labeling task due to time constraints, the consolidation of labels has not actually been applied. However, fine-tuning the models would proceed given that we have a fully labeled relation dataset. A summary of the resulting dataset can be seen on Figure 2

Also, seven total typographical errors for the training set were observed, consisting of *has_relation* and *has_habitat* labels, which could have possibly affected the model training performance.

3.2 Base Case Performance

Both models were initially tested, obtaining the results shown in Table 3. These scores clearly suggests the models are needed to be trained for the desired RE task.

3.3 RE Fine-tuned Model Comparison Across Learning Rates and Batch Sizes

A grid search approach was employed to obtain the optimal hyperparameters for both fine-tuned SpanBERT and RoBERTa. 20 trials were executed in total for about approximately 4 hours, attempting various combinations of learning rate and batch size similar to how SpanBERT was

Table 4: Learning Rate: 5e-6, Batch Size: 16

Model	Accuracy	F1	Precision	Recall
Spanbert	0.610080	0.609613	0.672114	0.628516
Roberta	0.503979	0.536095	0.565303	0.568025

Table 5: Learning Rate: 5e-6, Batch Size: 32

Model	Accuracy	F1	Precision	Recall
Spanbert	0.618037	0.639279	0.647327	0.645820
Roberta	0.503979	0.540922	0.560856	0.567143

tried for TACRED [4]. Each model was also saved in persistent memory for possible future reference. Tables 3 to 12 shows the comparisons between the two models given the same hyperparameters.

In summary, SpanBERT fine-tuned with a learning rate of *1e-5* and a *batch size* of 16 has outperformed all other models from the experiments, achieving an approximate f1_score of 69.49%. The same model also outperformed the others in terms of accuracy and recall, recording the corresponding scores of 69.5% and 70.56% respectively as seen in Table 6. The same models precision score is relatively high, achieving a score of approximately 71.8%. However, another model has achieved the highest score of 73.65% compared to all other models as shown in Table 8,

In terms of the models themselves, SpanBERT has significantly outweighed RoBERTa for all trials. Most of the lowest scores from all the trials has been observed from this model which underscores the great performance SpanBERT has shown given the experiments.

Additionally, more tests were desired to be conducted, but due to time constraints, such further experimentation on optimizing hyperparameters are to be explored in the project's continuation.

3.4 Discussion of Results

The fine-tuning of SpanBERT for a relation extraction task has generally shown good performance, given that it has achieved a score close to the metrics it obtained from RE training using the multi-class dataset. SpanBERT's benchmarking for another RE dataset TACRED yielded an f1 score of 70.8% [4], which is not far from the project's fine-tuned instance. This also considers that TACRED consists of approximately 106,000 samples with 42 relation label types [4]. However, concerning the slightly decreased performance, various factors that affected the overall score could be suggested, some of these was the constraint on the number of samples the labeled COPIOUS has along with not having a definitive guarantee of its span-masking features - given that SpanBERT has been pre-trained with masking

Table 6: Learning Rate: 1e-5, Batch Size: 16

Model	**Accuracy**	**F1**	Precision	**Recall**
Spanbert	**0.694960**	**0.694874**	0.718031	**0.705587**
Roberta	0.564987	0.586345	0.618202	0.597872

Table 7: Learning Rate: 1e-5, Batch Size: 32

Model	Accuracy	F1	Precision	Recall
Spanbert	0.628647	0.649037	0.668118	0.649358
Roberta	0.511936	0.549581	0.571819	0.567701

Table 8: Learning Rate: 2e-5, Batch Size: 16

Model	Accuracy	F1	**Precision**	Recall
Spanbert	0.639257	0.644293	**0.736451**	0.666711
Roberta	0.615385	0.609280	0.710071	0.649167

Table 9: Learning Rate: 2e-5, Batch Size: 32

Model	Accuracy	F1	Precision	Recall
Spanbert	0.668435	0.671328	0.714159	0.676397
Roberta	0.530504	0.568701	0.582309	0.580610

Table 10: Learning Rate: 3e-5, Batch Size: 16

Model	Accuracy	F1	Precision	Recall
Spanbert	0.618037	0.647115	0.665318	0.640236
Roberta	0.514589	0.514748	0.570922	0.578177

Table 11: Learning Rate: 3e-5, Batch Size: 32

Model	Accuracy	F1	Precision	Recall
Spanbert	0.652520	0.661347	0.704509	0.669550
Roberta	0.580902	0.577510	0.677271	0.618389

Table 12: Learning Rate: 5e-5, Batch Size: 16

Model	Accuracy	F1	Precision	Recall
Spanbert	0.612732	0.639521	0.660586	0.624581
Roberta	0.527851	0.552241	0.575040	0.580132

Table 13: Learning Rate: 5e-5, Batch Size: 32

Model	Accuracy	F1	Precision	Recall
Spanbert	0.625995	0.621434	0.649202	0.622871
Roberta	0.557029	0.542575	0.670921	0.591148

a contiguous span of 10 words [4], where the training sample sentences possess entities frequently separated in longer spans of distances. Also, comparing these results to previous rule-based and hybrid approaches exhibits that the current iteration of the RE model still significantly underperformed relative to its initial objectives [3].

Regarding RoBERTa, on the other hand, its poor performance compared to SpanBERT is highly due to the number of samples provided for fine-tuning. RoBERTa is generally larger in various aspects compared to the typical BERT model, hence obtaining the maximum f1 score of 61% is not surprising. This suggestion is certainly given that a separate study has fine-tuned RoBERTa on two larger clinical datasets with approximately 28,000 to 35000 samples encompassing both binary and multiclass relations – with the fine-tuned model achieving 88% and 95% correspondingly for binary relations [7]. Hence, further experiments for fine-tuning RoBERTa might prove to be challenging given the small number of samples available for domain-specific training.

Another factor considered affecting the results is the training data itself. Inspection of the data reveals that each subdivision for training and testing does not reflect the same relation label distributions. A great class imbalance between the amount of entity pairs, along with some unwanted relation labels in the training set might have hindered the overall performance of the model after the fine-tuning process. Overfitting during optimization is also highly suspected, given the log results recorded with the dev or eval metrics achieving over 70% percent, yet surprisingly underperformed on the final test set. Despite these flaws, SpanBERT has shown good performance which shows promise, supposing data quality is further refined along with leveraging more samples for the whole dataset. The refining of annotations partnered with the redistribution of samples across the train, dev, and test subdivisions are also highly considered to maintain model training integrity and performance. Furthermore, other preprocessing techniques are also being considered besides the chunked approach, for instance, another strategy instead of type masking before tokenization to improve model performance. Additionally, more models are also being sought which perform efficiently considering the limited amount of samples the project has.

4. CONCLUSION

The project's current developments gives notable insights in the effec- tiveness of transformer models, particularly SpanBERT, for extracting meaningful relationships from biodiversity liter- ature. By focusing on location-based connections such as species occurrences and habitats, the research offers a prac- tical approach to processing ecological data that may be used for further downstream tasks. The fine-tuned SpanBERT model consistently outperformed RoBERTa in terms of the RE task, exhibiting competitive results despite the issues encountered regarding the small amount of training samples paired with unimproved annotation quality, These findings emphasize the importance of high-quality data and optimized methods for enhancing RE tasks in biodiversity. From these, improving dataset coverage, refining annotation strategies, and exploring additional model architectures will be critical to effectively extend the utility of these tools in biodiversity research and conservation efforts.

5. REFERENCES

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

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

[3] R. S. Gabud, P. Lapitan, V. Mariano, E. Mendoza, N. Pampolina, M. A. A. Clariño, and R. T. Batista-Navarro, A hybrid of rule-based and transformer-based approaches for relation extraction in biodiversity literature, *Proc. 2nd Workshop Pattern-Based Approaches to NLP in the Age of Deep Learning*, 2023, pp. 103–113, Assoc. Comput. Linguistics. doi: 10.18653/v1/2023.pandl-1.10.

[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. doi: 10.48550/arXiv.1907.10529.

[5] S. Wadhwa, S. Amir, and B. C. Wallace, Revisiting Relation Extraction in the era of Large Language Models, *arXiv preprint*, arXiv:2305.05003, 2024. doi: 10.48550/arXiv.2305.05003.

[6] Y. Liu, M. Ott, N. Goyal, et al., RoBERTa: A Robustly Optimized BERT Pretraining Approach, *arXiv preprint*, arXiv:1907.11692, 2019. doi: 10.48550/arXiv.1907.11692.

[7] X. Yang, Z. Yu, Y. Guo, J. Bian, and Y. Wu, Clinical Relation Extraction Using Transformer-based Models, *arXiv preprint*, arXiv:2107.08957, Jul. 2021. https://doi.org/10.48550/arXiv.2107.08957.