Location-Specific Information Extraction from Unstructured Biodiversity Literature

Jeezer Lee
November 25, 2024

Outline

1. Overview
2. Current Progress
3. Next Activities

Why Location-Based Biodiversity Information

- Identifying Species distribution and regional habitat conditions
- more relevant area-specific conservation strategies compared to species-focused data
  - may guide efforts on endangered species and ecosystems on particular locations
- Links geographical contexts to biodiversity data to address further concerns like habitat loss and climate change impacts

Extracting such unstructured information can be achieved through Natural Language Processing

Given that we want to obtain biodiversity information with a focus on location, we want to:

- Extract meaningful entities - through Named Entity Recognition (NER)
- Determine relations among such entities - through Relation Extraction (RE)

Previous studies dealt with information extraction on Unstructured Biodiversity data, where most have explored NER and several attempts on RE. Also, most Information Extraction primarily focuses on other specific data, such as species occurrence extraction.

The lack of annotated datasets for RE is a challenge that invites further exploration and experimentation.

Hence, this study focuses on such gaps, particularly on location-based Information Extraction on biodiversity texts.

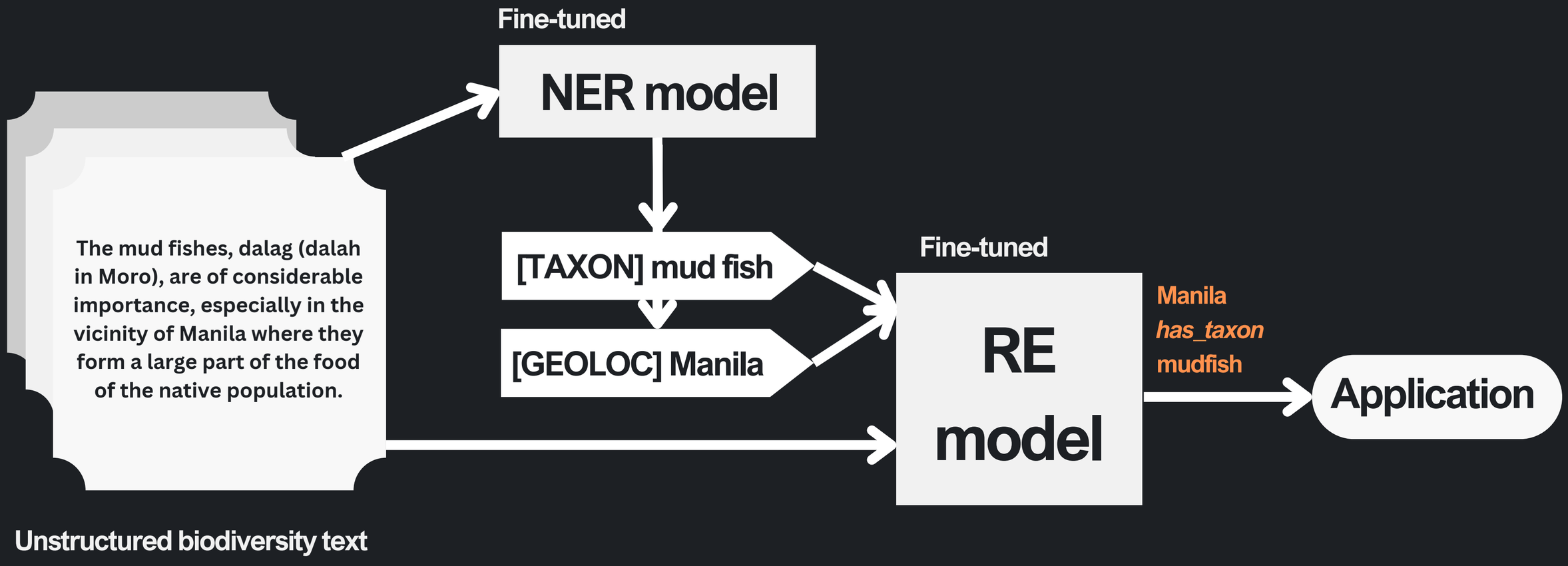
Objectives

1. Annotate existing NER-labeled biodiversity corpus for supervised relation extraction task

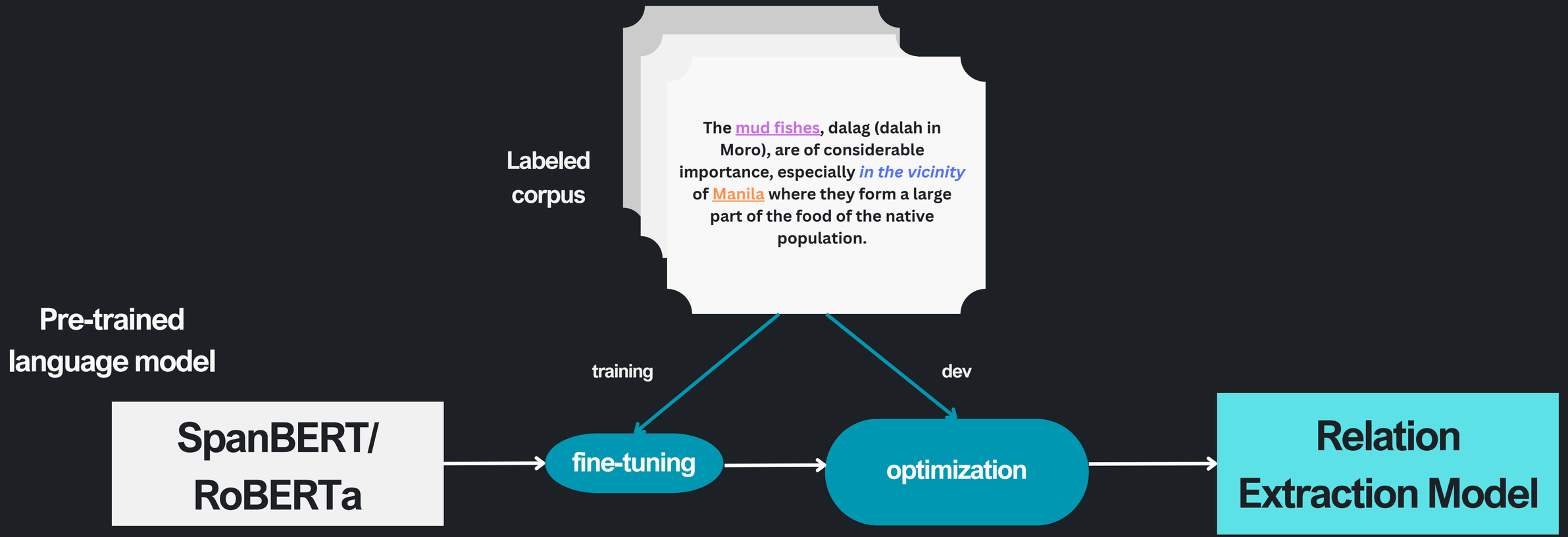
2. Train various RE models given the annotated dataset

3. Evaluate trained RE model performance and integrate model for application

Pipeline


Fine-tuned
NER model
The mud fishes, dalag (dalah in Moro), are of considerable importance, especially in the vicinity of Manila where they form a large part of the food of the native population.
[TAXON] mud fish
[GEOLOC] Manila
Fine-tuned
RE model
Manila
has_taxon
mudfish
Application
Unstructured biodiversity text

Model Training



PLMs

Transformer Architectures - (faster and deeper bidirectional context understanding) - good performance on sequence-to-sequence problems

BiodivBERT (NER)

- Followed the same pre-training procedure with BERT, but considered a great amount of biodiversity-related text.
- Approx. 5 Gb of training data
- Has shown good performance for biodiversity NER, where this study would continue to further optimize its hyperparameters (2024)

SpanBERT (RE)

- Aimed to improve BERT by using span-based masking (BERT uses token-based masking) for its masked language modeling (MLM) on pre-training
- Optimizes MLM by contiguous approach on masking with the Span Boundary Objective
- trained on Wikipedia and BookCorpus like BERT (16Gb)

RoBERTa (RE)

- Aimed to improve BERT by by removing Next Sentence Prediction component.
- Trained on a larger corpus, approx. 160Gb of data compared to BERT with 16 Gb, and trained on longer sequences

PLMs

SpanBERT

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$
$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$

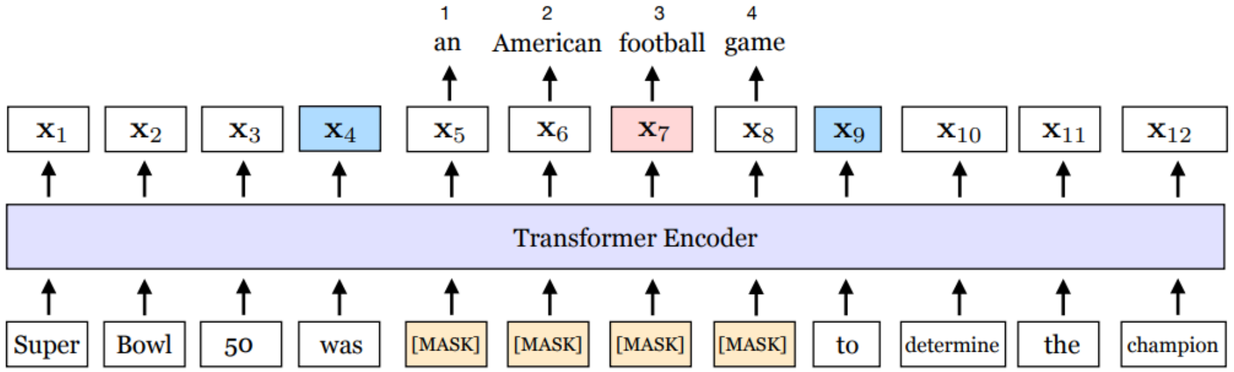


Figure 1: An illustration of SpanBERT training. The span *an American football game* is masked. The span boundary objective (SBO) uses the output representations of the boundary tokens, $\mathbf{x}_4$ and $\mathbf{x}_9$ (in blue), to predict each token in the masked span. The equation shows the MLM and SBO loss terms for predicting the token, *football* (in pink), which as marked by the position embedding $\mathbf{p}_3$, is the *third* token from x_4.

This approach forces the model to learn relationships between spans and their boundaries, which is critical for tasks like relation extraction.

On **pre-training**

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$
$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$

A span boundary objective (SBO) is introduced that involves predicting each token of a masked span using only the representations of the observed tokens at the boundaries.

I love [*chocolate and]* vanilla ice cream

I love [MASK MASK] vanilla ice cream

chocolate is predicted by

Span Boundary Objective:

- Left boundary: Embedding of "love"
- Right boundary: Embedding of "vanilla"
- Positional embedding: The relative position of "chocolate" in the span).

MLM full sentence context - embeddings of all tokens in the sequence

PLMs

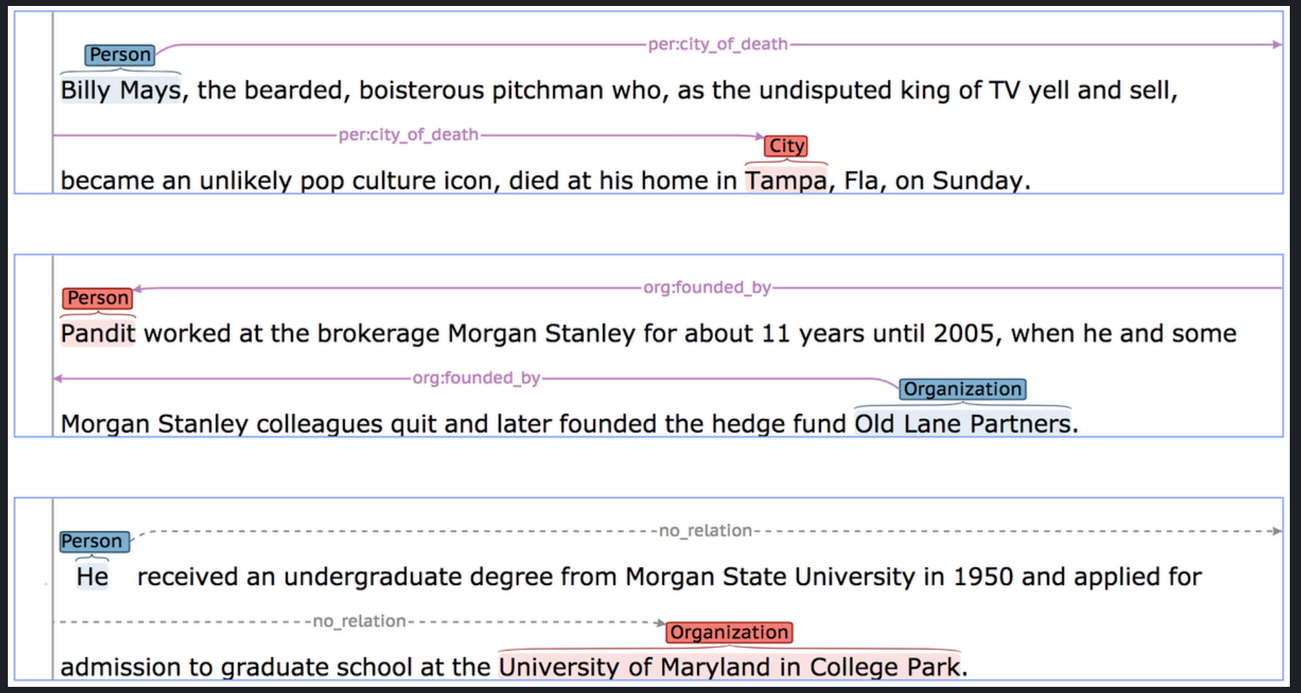
SpanBERT

2.4. Relation Extraction

	p	R	F1
BERT_{EM}(Soares et al., 2019)	–	–	70.1
BERT_{EM}+MTB*	–	–	**71.5**
Google BERT	69.1	63.9	66.4
Our BERT	67.8	67.2	67.5
Our BERT-1seq	**72.4**	67.9	70.1
SpanBERT	70.8	**70.9**	**70.8**

Test performance on the TACRED relation extraction benchmark

> *SpanBERT exceeds the reimplementation of BERT by 3.3% F1 and achieves close to the current state of the art.*

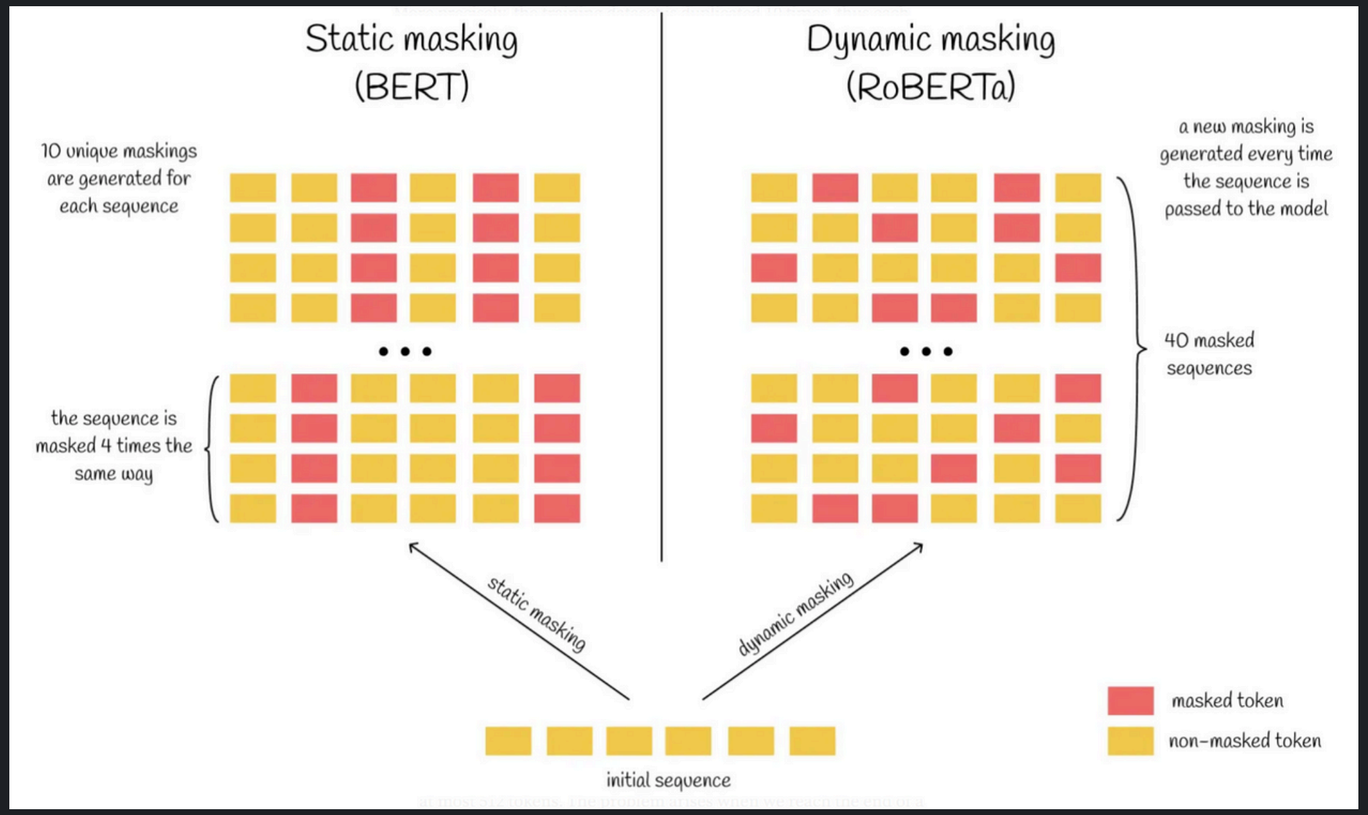


TACRED: Large Scale RE dataset - 41 relation types

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.

PLMs

RoBERTa



Dynamic masking

tokens to be masked are chosen randomly in every training epoch. The model is exposed to a wider range of masking patterns for the same input sentences, making it more robust to varied contexts; better generalizations

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

Dataset: COPIOUS (2019)

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 (1) taxon names, (2) geographical locations, (3) habitats, (4) temporal expressions, and (5) person names.

107004_English_78294_34354172_1912.txt

```
20 HOLLISTER.
Kerivouta whitehead! Thomas.
1894. Kerivoula whiteheadi Thomas, Ann. and Mag. Nat. Hist, VI, 14, 460.
Type locality. — Isabela, N. E. Luzon. Luzon (Thomas). Family MOLOSSID.
Genus CHAEREPHON Dobson.
1874. Chaerephon Dobson, Journ. Asiatic Soc. Bengal, 43, pt. 2, 144. Type. — Nyctinomiis johore
Bats with ears connected by a low band across crown; tragus very small; tail long, free from in
Chserephon plicatus (Buchanan).
1800. Vespertilio plicatus Buchannan, Trans. Linn. Soc, 5, 261. 1907. Chaerephon plicatus Ander
Type locality. — Puttahaut, Bengal. Philippine Islands (Dobson) ; Luzon (Elera)
Order CARNIVORA.
Family MUSTELID.
Genus MARTES Pinel.
1792. Martes Pinel, Actes Soc. Hist. Nat, Paris, 1, 55, footnote. Type. — Martes domestica Plne
Martens, the single supposed Philippine species about the size of a domestic cat; head and body
Martes henricii (Westerman).
1848. Mustela (Martes) henricii Westerman, Bijdragen tot de Dier-kunde, 1, 13.
Type locality. — Java. Sulu (Trouessart).  Sanchez, Los Mamiferos de Filipinas, does not allow
```

107004_English_78294_34354172_1912.ann

```
T1  Taxon 290 312   Nyctinomiis johorensis
T2  Taxon 651 670   Chserephon plicatus
T3  Taxon 1068 1085 Plnelmarfes foina
T4  Taxon 1444 1459 Martes henricii
T5  Taxon 14 33 Kerivouta whitehead
T6  Taxon 49 69 Kerivoula whiteheadi
T7  Taxon 49 90 Kerivoula whiteheadi Thomas, Ann. and Mag
T8  GeographicalLocation 133 140    Isabela
T9  GeographicalLocation 148 153    Luzon
T10 GeographicalLocation 155 160    Luzon
T11 Taxon 178 186   MOLOSSID
T12 Taxon 194 211   CHAEREPHON Dobson
T13 Taxon 219 236   Chaerephon Dobson
T14 Taxon 219 243   Chaerephon Dobson, Journ
T15 Taxon 290 319   Nyctinomiis johorensis DOBSON
T16 Taxon 651 681   Chserephon plicatus (Buchanan)
T17 Taxon 689 709   Vespertilio plicatus
T18 Taxon 689 719   Vespertilio plicatus Buchannan
T19 Taxon 689 726   Vespertilio plicatus Buchannan, Trans
T20 Taxon 753 772   Chaerephon plicatus
```

Dataset: COPIOUS (2019)

- Only has NER annotations for training, hence the project would also require and aim to obtain relation annotations for supervised training of a relation extraction model
- (*GeographicalLocation-Taxon, Geographical-Habitat*) relations

107004_English_78294_34354172_1912.txt

```
20 HOLLISTER.
Kerivouta whitehead! Thomas.
1894. Kerivoula whiteheadi Thomas, Ann. and Mag. Nat. Hist, VI, 14, 460.
Type locality. — Isabela, N. E. Luzon. Luzon (Thomas). Family MOLOSSID.
Genus CHAEREPHON Dobson.
1874. Chaerephon Dobson, Journ. Asiatic Soc. Bengal, 43, pt. 2, 144. Type. — Nyctinomiis johore
Bats with ears connected by a low band across crown; tragus very small; tail long, free from in
Chserephon plicatus (Buchanan).
1800. Vespertilio plicatus Buchannan, Trans. Linn. Soc, 5, 261. 1907. Chaerephon plicatus Ander
Type locality. — Puttahaut, Bengal. Philippine Islands (Dobson) ; Luzon (Elera)
Order CARNIVORA.
Family MUSTELID.
Genus MARTES Pinel.
1792. Martes Pinel, Actes Soc. Hist. Nat, Paris, 1, 55, footnote. Type. — Martes domestica Plne
Martens, the single supposed Philippine species about the size of a domestic cat; head and body
Martes henricii (Westerman).
1848. Mustela (Martes) henricii Westerman, Bijdragen tot de Dier-kunde, 1, 13.
Type locality. — Java. Sulu (Trouessart).  Sanchez, Los Mamiferos de Filipinas, does not allow
```

107004_English_78294_34354172_1912.ann

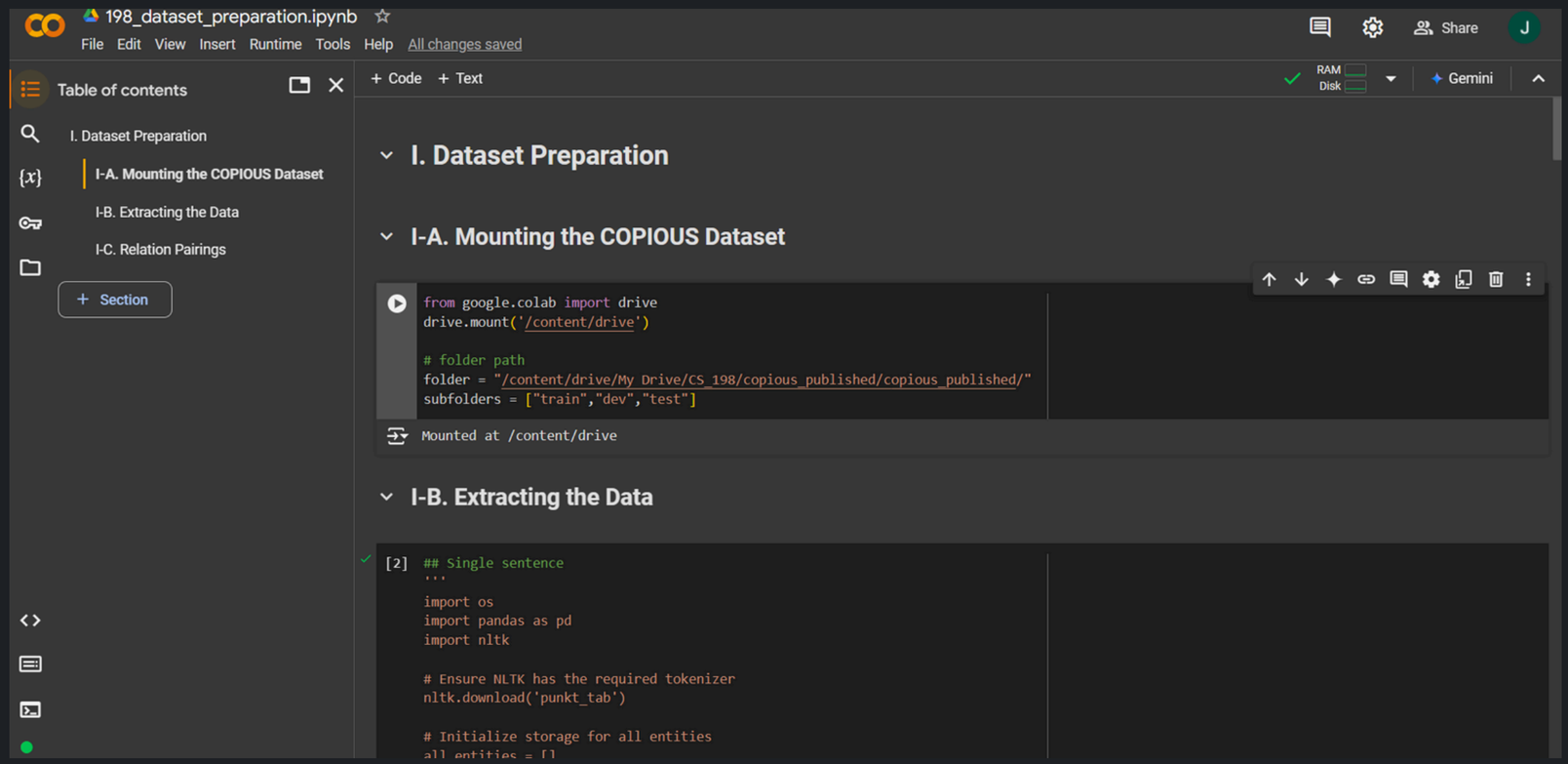
```
T1  Taxon 290 312   Nyctinomiis johorensis
T2  Taxon 651 670   Chserephon plicatus
T3  Taxon 1068 1085 Plnelmarfes foina
T4  Taxon 1444 1459 Martes henricii
T5  Taxon 14 33 Kerivouta whitehead
T6  Taxon 49 69 Kerivoula whiteheadi
T7  Taxon 49 90 Kerivoula whiteheadi Thomas, Ann. and Mag
T8  GeographicalLocation 133 140    Isabela
T9  GeographicalLocation 148 153    Luzon
T10 GeographicalLocation 155 160    Luzon
T11 Taxon 178 186   MOLOSSID
T12 Taxon 194 211   CHAEREPHON Dobson
T13 Taxon 219 236   Chaerephon Dobson
T14 Taxon 219 243   Chaerephon Dobson, Journ
T15 Taxon 290 319   Nyctinomiis johorensis DOBSON
T16 Taxon 651 681   Chserephon plicatus (Buchanan)
T17 Taxon 689 709   Vespertilio plicatus
T18 Taxon 689 719   Vespertilio plicatus Buchannan
T19 Taxon 689 726   Vespertilio plicatus Buchannan, Trans
T20 Taxon 753 772   Chaerephon plicatus
```

Current Progress

- Dataset Pre-processing for Annotations
- Dataset Annotation
- Model Training Preparation

Current Progress

Google Colab
google.com



Dataset Preparation

- Prepare for RE annotation
- Explore Dataset Distribution of entity and relations
  - stored in .xlsx files
  - Annotating the existence of a binary relation between two entities is suggested as the approach is simple

Current Progress

Dataset Distribution: Entities

Category	Train	Dev	Test
Taxon	9,357	1,548	1,322
Geographical Location	8,121	992	878
Person	2,479	180	230
Temporal Expression	1,800	157	253
Habitat	1,308	91	115

Current Progress

Dataset Distribution: Relation Pairs

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

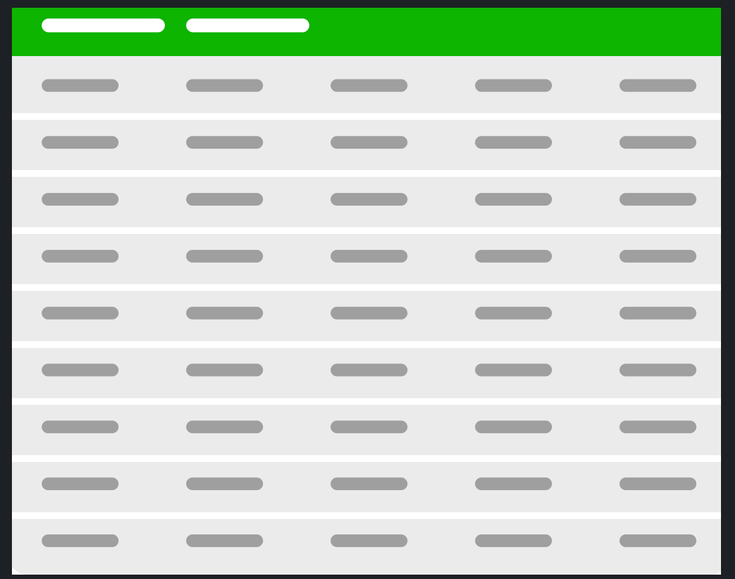
Current Progress

Dataset pre-processing for relation labeling

Google Colab
google.com

.txt file

.ann file



Spreadsheet format for relation labeling

Current Progress

Dataset pre-processing for relation labeling

.txt file

A BIOLOGICAL AND SYSTEMATIC STUDY OF PHILIPPINE PLANT GALLS 1
By Leopoldo B. Uichanco (From the College of Agriculture, University of the Philippines, Los Banos)
FIFTEEN PLATES INTRODUCTION
Galls are abnormal growths on the stems, leaves, roots, or other parts of plants, caused by the action of insects, arachnids, or fungi, or by unknown agencies. Just how these peculiar structural developments are brought about is still open to discussion and speculation, experimental proofs being, up to the present, too deficient to warrant our drawing any definite conclusion. These malformations have been ascribed to various causes, the more commonly accepted, in the absence of more reasonable, explanations being the following: 2 1, a severe mechanical injury to certain parts of the plant ; 2, a continuous mechanical irritation; 3, secretion of chemical stimulus by the causal animal or fungus. One, or a combination of two or all ...

.ann file

```
T1  GeographicalLocation 151 160    Los Banos
T2  Taxon 1314 1336 Aulax papaveris Perris
T3  Taxon 296 303   insects
T4  Taxon 305 314   arachnids
```

3-sentence chunks

(S1, S2, S3), (S2, S3, S4), (S3, S4, S5), ...

Spreadsheet rows for each entity, entity type, and chunks

File	Chunk	Entity ID	Entity Type	Entity Text	Spans
111121_English_36025311_1908	484SMITH. THE SULU ARCHIPELAGO. The Sulu Archipelago	T1	GeographicalLocation	Sulu Archipelago	36 52
111121_English_36025311_1908	THE SULU ARCHIPELAGO. The Sulu Archipelago is practicall	T2	Person	Becker	215 221
111121_English_36025311_1908	THE SULU ARCHIPELAGO. The Sulu Archipelago is practicall	T3	GeographicalLocation	Marongas Island	270 285

Current Progress

Dataset pre-processing for relation labeling

Spreadsheet rows for each entity, entity type, and chunks

File	Chunk	Entity ID	Entity Type	Entity Text	Spans
111121_English_36025311_1908	484SMITH. THE SULU ARCHIPELAGO. The Sulu Archipelago	T1	GeographicalLocation	Sulu Archipelago	36 52
111121_English_36025311_1908	THE SULU ARCHIPELAGO. The Sulu Archipelago is practicall	T2	Person	Becker	215 221
111121_English_36025311_1908	THE SULU ARCHIPELAGO. The Sulu Archipelago is practicall	T3	GeographicalLocation	Marongas Island	270 285

Filtering entities, pairing desired entities, readying a column for relation labeling

File	Entity 1	Entity 1 Type	Entity 2	Entity 2 Type	Chunk	relation
104421_English_33451123_1922-26.	Asia	GeographicalLocation	pantropic	Habitat	118, Loher 7213. Planted about r	
104421_English_33451123_1922-26.	Luzon	GeographicalLocation	GRAMINEAE	Taxon	i VOL GRAMINEAE 49 Luzon (Abra, Ilocos Norte, Pangas	
104421_English_33451123_1922-26.	Luzon	GeographicalLocation	open grasslands	Habitat	i VOL GRAMINEAE 49 Luzon (Abra, Ilocos Norte, Pangas	
104421_English_33451123_1922-26.	Abra	GeographicalLocation	open grasslands	Habitat	i VOL GRAMINEAE 49 Luzon (Abra, Ilocos Norte, Pangas	
104421_English_33451123_1922-26.	Ilocos Norte	GeographicalLocation	open grasslands	Habitat	i VOL GRAMINEAE 49 Luzon (Abra, Ilocos Norte, Pangas	

Current Progress

Dataset Feature

- Hence, sentences are further grouped into chunks to at least contain two desired entities to classify a binary relation

A	B	C	D	E	F	G	H	I	
influence	1	...	soil	density	varies	in	response	to	earth
	-	O	B-Quality	I-Quality	O	O	O	O	B-Ph
have	1	functional	richness	...	on	ecosystem	states	and	proce
	-	B-Quality	I-Quality	O	O	B-Environment	O	O	O
occur_in	1	...	and	diversity	of	chemolithoautotrophic	bacteria	in	saline
	-	O	O	O	O	B-Organism	I-Organism	O	B-En
have	1	ant	abundance	and	diversity	associated	with	natural	habita
	-	B-Quality	I-Quality	O	O	O	O	B-Environment	I-Env
	0	for	understanding	how	microbial	communities	degrade	plant	bioma
	-	O	O	O	B-Organism	I-Organism	O	B-Quality	I-Qua
	0	density	of	soil	invertebrates	in	response	to	earth
	-	B-Quality	O	O	O	O	O	O	B-Ph

BiodivNERE dataset

Sentence-level relations and annotations, short spans

- One such observation for the COPIOUS dataset is that two entities are not usually captured within a single sentence

The relevant predictions from Heaney et al. (1989) derived from our studies on Leyte and **Negros** are (1) in forested habitats, species richness and relative abundance of fruit bats should be highest in the **lowlands** and decline with increasing elevation, (2) species richness of fruit bats should be lower in **agricultural areas**, where geographically widespread species should predominate, (3) non-volant mammals should not exhibit highest species richness in **lowland forest,** (4) abundance of non-volant mammals should increase with elevation, and (5) endemic species of fruit bats and non-volant mammals should occur in primary forest, whereas non-endemics should predominate in disturbed habitats.

COPIOUS dataset

Entity pairs are not usually captured in a single sentence; longer spans

Current Progress

Dataset Annotation

- Orientation for relation labeling was conducted today (November 25, 2024)
  - Results are expected next week (December 2) to start model training

Relation Labels

has_taxon: Geolocation-Taxon pairs
has_habitat: Geolocation-Habitat pairs

Binary Classification of entity pair relation

Determine whether the sentence explicitly or implicitly established a **biological or ecological connection** with the entities

- *Tilapia thrives in lakes near Manila.”* → *has_taxon (Manila,Tilapia)*
- *Tilapia sold in China markets is imported from Manila.* → *no_relation (China,Tilapia)*

Current Progress

Dataset Annotation

- Two groups (A and B) of five CWTS students are annotating the data

- Group A & B (member 1)
- Group A & B (member 2)
- Group A & B (member 5)

train.xlxs	dev.xlxs	test.xlxs
2–625	2–27	2–75
626–1250	28–54	76–150
1251–1875	55–81	151–225
1876–2500	82–108	226–300
2501–3128	109–137	301–378

Two people would work independently on the same chunks per file to compare annotations

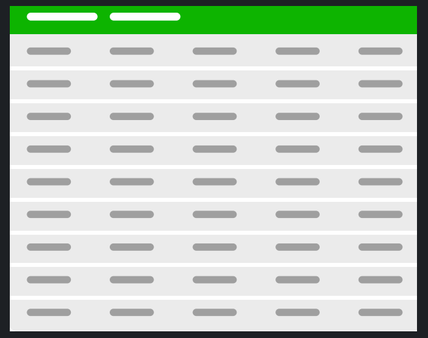
- for consolidation of final labels upon evaluation

Current Progress

Model Training Preparation

Google Colab
google.com

RE labeled dataset



Training code is being polished, prepared for labeled inputs and ready for training

```python
# Preprocessing
def preprocess_data(batch, tokenizer, model_name, label_to_id, max_seq_length):
    inputs = []
    for chunk, entity1, entity2, entity_type1, entity_type2 in zip(
        batch["chunk"], batch["entity1"], batch["entity2"], batch["entity_type1"], batch["entity_type2"]
    ):
        if "spanbert" in model_name:
            # SpanBERT-specific input formatting: Similar for BERT formattings
            input_text = chunk.replace(entity1, f"[{entity_type1}]").replace(entity2, f"[{entity_type2}]")
        else:
            # RoBERTa-specific input formatting: not really different since this was updated to follow the procedure for input formatt
            input_text = chunk.replace(entity1, f"[{entity_type1}]").replace(entity2, f"[{entity_type2}]")
        inputs.append(input_text)

    encodings = tokenizer(inputs, padding=True, truncation=True, max_length=max_seq_length)
    encodings["labels"] = [label_to_id[label] for label in batch["relation"]]
    return encodings
```

Next Activities

1. Annotation Evaluation and Consolidation - By December 2

2. Model Training and Fine-tuning - By December 3

3. Concerns with the amount of training data and model performance
   - Add more training data
   - Other models

4. Start building the NER segment for implementation purposes