A Hybrid of Rule-based and Transformer-based Approaches for Relation Extraction in Biodiversity Literature

Roselyn Gabud[a,b,*], **Portia Lapitan**[c], **Vladimir Mariano**[d], **Eduardo Mendoza**[b,e]
Nelson Pampolina[c], **Maria Art Antonette Clariño**[b], **Riza Batista-Navarro**[b,f]

[a]Department of Computer Science, University of the Philippines Diliman
[b]Institute of Computer Science, University of the Philippines Los Baños
[c]Department of Forest Biological Sciences, University of the Philippines Los Baños
[d]YSEALI Academy, Fulbright University Vietnam
[e]Max Planck Institute of Biochemistry, Munich, Germany
[f]University of Manchester, United Kingdom
[*]rsgabud@up.edu.ph

Abstract

Relation extraction (RE) is one of the tasks behind many relevant natural language processing (NLP) applications. Exploiting the information hidden in millions of scholarly articles by leveraging NLP, specifically RE, systems could benefit studies in specialized domains, e.g. biomedicine and biodiversity. Although deep learning (DL)-based methods have shown state-of-the-art performance in many NLP tasks including RE, DL for domain-specific RE systems has been hindered by the lack of expert-labeled datasets which are typically required to train such methods. In this paper, we take advantage of the zero-shot (i.e., not requiring any labeled data) capability of pattern-based methods for RE using a rule-based approach, combined with templates for natural language inference (NLI) transformer models. We present our hybrid method for RE that exploits the advantages of both methods, i.e., interpretability of rules and transferability of transformers. Evaluated on a corpus of biodiversity literature with annotated relations, our hybrid method demonstrated an improvement of up to 15 percentage points in recall and best performance over solely rule-based and transformer-based methods with F1-scores ranging from 89.61% to 96.75% for reproductive condition - temporal expression relations, and ranging from 85.39% to 89.90% for habitat - geographic location relations.

1 Introduction

Relation extraction (RE) is a Natural Language Processing (NLP) task that is concerned with the identification of binary semantic relationships between entities or concepts in text. RE predicts whether a relationship holds between two entities (or concepts), based on the context of the sentence. Many approaches to RE take as input a sentence together with pre-extracted named entities (within the sentence), and identify relations between those entities using heuristics or machine learning-based approaches (Detroja et al., 2023). For example, in the sentence *"Semangkok Forest Reserve is a designated hill dipterocarp forest conservation area located in Selangor state about 60 km north of Kuala Lumpur, while Pasoh Forest Reserve, is a designated lowland dipterocarp forest conservation area in Negeri Sembilan about 60 km east - southeast of Kuala Lumpur, Peninsular Malaysia"*,[1] an RE system should be able to identify the relationship of the geographic location entity "Semangkok Forest Reserve" with the habitat entity "hill dipterocarp forest", but no relation between "Semangkok Forest Reserve" and "lowland dipterocarp forest". This type of information is helpful in associating information on habitats with geographic distribution of species.

Extracting information from text is better guided by domain knowledge of the targeted use case (Chiticariu et al., 2013; Waltl et al., 2018; Wu et al., 2022). For example, in the biodiversity domain, methods for extracting a plant species' *geographic location* with related *habitat* information, and *reproductive condition* (i.e., reproductive status) with related *temporal expression* that exist in biodiver-

[1]Source: Tani, N., et al. (2016). Selective logging simulations and male fecundity variation support customisation of management regimes for specific groups of dipterocarp species in Peninsular Malaysia. Journal of Tropical Forest Science, p370.

Proceedings of the 2nd Workshop in Pattern-based Approaches to NLP in the Age of Deep Learning , pages 103–113

sity texts, are better crafted with the involvement of domain experts. Rule-based methods lend themselves well to domain-specific RE tasks (Aljamel et al., 2015; Peng et al., 2016; Wang et al., 2022). Aside from being highly interpretable, such methods define a set of rules manually created by domain experts and capture syntactic patterns that are associated with different types of relations observed in a corpus (Ravikumar et al., 2017; Korger and Baumeister, 2021; Xu et al., 2022). Advancements in machine learning (ML) and, more recently, deep learning (DL), have led to state-of-the-art performance in RE. ML and DL-based models learn features from data giving them strong generalization ability, adaptability, and scalability. However, the performance of ML- and DL-based methods relies on the availability of domain-specific annotated datasets; this assumption is not always viable for many specialised domains such as biodiversity, law or finance (Thomas and Sivanesan, 2022).

In this paper, we integrate the advantages of both rule-based and DL-based RE methods by developing a zero-shot hybrid RE approach. Our main contribution is a novel hybrid RE method that is underpinned by a two-step approach. In the first step, we hand-crafted rules that capture syntactic patterns, which were implemented based on regular expressions (regexes). The second step leverages a state-of-the-art transformer model, Text-to-Text Transfer Transformer or T5 (Raffel et al., 2020), for natural language inference (NLI). We created premise-hypothesis templates as input for the T5-based NLI model, to determine if a relation holds between a given pair of entities. Our method presents the following advantages over other RE methods: (1) improved performance over solely rule-based or DL-based methods; (2) reduced computational bottleneck since a substantial proportion of the relations are extracted using the more computationally efficient rule-based method; and (3) reduced labeling cost associated with dataset or corpus annotation.

We applied our method on documents drawn from the body of literature on biodiversity – a relatively under-resourced domain – focusing on two types of relations: (1) plant species' reproductive conditions and their related temporal expressions, and (2) habitats and their related geographic locations. Harvesting these details from biodiversity literature will enable data-driven discovery of plant species' reproductive patterns and habitats. This, in turn, will aid in more informed plans for reforestation and restoration of land.

In the remainder of this paper, we first provide a review of prior work related to our study (Section 2). This is followed by our problem formulation (Section 3) and a description of the dataset we developed and used in our experiments (Section 4). Then, we present details of the zero-shot hybrid RE approach that we developed (Section 5), and the results of evaluating the hybrid model (Section 6). We then analyze our results (Section 7) before providing a summary of our findings and directions for future work in Section 8.

2 Related Work

Existing RE methods can be categorized into two broad types: traditional and DL-based methods. Traditional methods use either rules or machine learning techniques (e.g., those based on statistical classifiers trained on hand-crafted features) to extract a set of predefined relations from a corpus (Detroja et al., 2023). Rule-based methods define rules, which are a set of hand-crafted extraction patterns typically created by domain experts (Agichtein and Gravano, 2000; Fundel et al., 2007; Zhang et al., 2009; Nguyen et al., 2015). These rules are based on manually identified syntactic patterns that are associated with different types of relations, as observed in a corpus. Rules have the advantage of being highly interpretable: they can be easily understood by humans, which makes them a good choice for tasks where it is important to explain the reasoning behind the system's output. However, rule-based methods have two main limitations: they can be time-consuming to create and they are domain-dependent (e.g., a rule-based system that is designed to identify relations in medical text may not necessarily be able to identify relations in financial text). Meanwhile, ML techniques for RE are based on the supervised training of a classification model on a dataset whereby relations have been manually annotated. There are feature-based methods (Miller et al., 2000; Kambhatla, 2004) that use selected syntactic and semantic features as the bases of similarity in training a classification model. There are also kernel-based methods (Zelenko et al., 2002; Culotta and Sorensen, 2004) that use kernel functions to determine similarity between two relation instance representations, together with a support vector machine (SVM) model as a classifier. Although ML-based RE methods

gained superiority in the past in terms of performance, their performance is greatly dependent on the set of selected features or the choice of kernel functions. As they are trained in a supervised manner, ML models also require labeled data which can be costly.

DL methods that have emerged more recently have been shown to outperform traditional methods for RE. These DL models learn higher-order, abstract feature representations from sentences that make them more generalizable, adaptable to new domains, and scalable. With the emergence of DL, models that employ neural architectures such as convolutional neural networks (CNNs) (Liu et al., 2013; dos Santos et al., 2015), recurrent neural networks (RNN) (Vu et al., 2016; Zhang and Wang, 2015), graph convolutional networks (GCN) (Zhu et al., 2019), attention-based neural networks (Wang et al., 2016; Xiao and Liu, 2016), and transformer-based language models (Vaswani et al., 2017; Le Guillarme and Thuiller, 2022) have been utilized for RE tasks. However, similar to traditional ML models, training or fine-tuning DL models for downstream applications such as RE also requires labeled data (Zhao et al., 2023).

In recent years, *zero-shot* transformer-based approaches to information extraction requiring no labeled data have become popular (Liu et al., 2020; Du and Cardie, 2021; Cheng et al., 2021; Li et al., 2022). For instance, Levy et al. (2017) reduced RE to the problem of answering simple reading comprehension questions. They mapped each relation type *R(x, y)* to at least one parameterized natural-language question q_x whose answer is *y*. For example, the relation *educated_at(x, y)* can be mapped to *"Where did x study?"* and *"Which university did x graduate from?"*. The success of these types of RE methods is primarily due to the significant developments in and availability of transformer-based pre-trained language models (PLMs) (Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020). PLMs were pre-trained on large-scale corpora using unsupervised learning objectives such as masked language modeling, and were then fine-tuned for downstream tasks, such as question answering (QA) and natural language inference (NLI), using relatively smaller amounts of task- or domain-specific labeled data (Devlin et al., 2019). Zero-shot transformer-based approaches to RE are based on the careful, systematic construction of inputs for PLMs, which then elicit a model prediction (i.e., a label) that can be mapped to a decision on whether a certain type of relation holds. Zero-shot methods significantly reduce the labeling cost associated with RE because only a small amount of labeled data is required, i.e., test samples for evaluating the model.

In this work, we developed a zero-shot approach to RE that is the first of its kind to be applied in the biodiversity domain. Our zero-shot methods for RE are based on rules and transformer models that – when combined – demonstrate superior performance, in comparison to the use of rules or transformers alone.

3 Problem Formulation

Given an input sentence I that is a sequence of tokens $[t_0, t_1, ..., t_n]$, a source entity $E_S = [t_i, ..., t_j]$ and a target entity $E_T = [t_u, ..., t_v]$, we treat the RE task as a binary classification task, whereby the input is the triple (I, E_S, E_T), and the output is $y \in \{0, 1\}$ where 1 indicates that a relationship from the source entity to the target entity ($E_S \rightarrow E_T$) exists, otherwise 0. In this work, we focus on the two relation types described below.

The `has_time` relation: This holds between a reproductive condition mention and a temporal expression, i.e., "*reproductive condition* `has_time` *temporal expression*", whereby the reproductive condition mention is considered to be the source and the temporal expression serves as the target.

The `has_location` relation: This holds between a habitat mention and a geographic location, i.e., "*habitat* `has_location` *geographic location*", whereby the habitat mention is considered to be the source and the geographic location is the target.

4 Dataset

To support the development of approaches to the above-mentioned problem, we utilized a corpus that is a subset[2] of the gold standard corpus for named entity recognition (NER) that was presented in the work of Gabud et al. (2019) and was designed in accordance to the annotation scheme used in the COPIOUS project (Nguyen et al., 2019). It contains information relevant to the occurrence and reproductive patterns of a tropical tree family, *Dipterocarpaceae* (dipterocarps). The corpus is

[2]Due to resource constraints, we made the decision at the beginning of the study to have only a limited number of documents manually annotated.

comprised of 151 manually selected one- to two-paragraph documents from online environmental science and ecology journal repositories, e.g., Journal of Tropical Ecology, Journal of Ecology, etc. For this RE work, we are particularly interested in the annotations of the following named entity (NE) types: habitat, geographic location, temporal expression, and reproductive condition. The descriptions and examples of these NE types are shown in Table 1. We selected sentences that contain at least one occurrence of an entity pair, i.e., either a pair of habitat and geographic location mentions, or a pair of reproductive condition and temporal expression mentions. We then produced relation annotations by creating data instances, each of which is in the form (I, E_S, E_T, y), where I is the input sentence, E_S is the source entity, E_T is the target entity, and y is the relation label which is set to 1 if a binary relation between the source and target entities hold, otherwise y is set to 0. As mentioned in the previous section, we decided to focus on two types of relations: `has_time` (which holds between a reproductive condition mention and a temporal expression), and `has_location` (which holds between a habitat mention and a geographic location). Table 2 shows the two data instances belonging to the `has_location` relation type, that were created from the following example sentence that contains one habitat and two geographic location entities: *"The main observation site was conserved forest at Dongmakhai (18deg 20' 03"N, 102deg 30' 5"E, 190m a.s.l.)."*

Concept	Description and Example
Habitat	Environments in which organisms live. e.g., *lowland mixed dipterocarp forests*
Geographic Location	Any identifiable point or area in the planet, ranging from continents, major bodies of water, named landforms, countries, etc. e.g., *Sabah*
Reproductive Condition	Indicators of the specimens' reproductive behaviour. e.g., *mast fruiting*
Temporal Expression	Spans of text pertaining to points in time. e.g., *mid-August*

Table 1: Descriptions and examples of our biodiversity entity types of interest.

Two annotators manually provided the label y for each data instance (I, E_S, E_T, y). Our more senior annotator, a Biology degree holder, labeled all the instances in the entire dataset, while a junior annotator, a Computer Science student, provided labels for 30% of the dataset only. They carried out the annotation task independently. We calculated the agreement between our two annotators in terms of F1-score, and obtained an overall agreement of 95.87%. The agreement for the `has_time` relation type is 94.36%, while that for the `has_location` type is 97.37%. We resolved the disagreements by involving a third annotator who is a co-author of this work. The instances with disagreements were re-evaluated and re-labeled by the third annotator. We randomly split our dataset into 70% training set, 10% development set, and 20% test set. Table 3 shows the number of instances for each relation type.

Habitat	Geo. Location
conserved forest	*Dongmakhai*
conserved forest	*18deg 20' 03"N , 102deg 30' 5"E*

Table 2: Example data instances of the `has_location` relation type based on the sentence, *"The main observation site was conserved forest at Dongmakhai (18deg 20' 03"N, 102deg 30' 5"E, 190m a.s.l.)."*

Relation Type	train	dev	test
`has_time`	843	173	388
`has_location`	252	34	127

Table 3: Frequency of instances for each relation type in our training (train), development (dev) and test sets.

5 Methods

In this section, we present our methods for extracting (1) related reproductive condition and temporal expressions (`has_time`), and (2) related habitat and geographic location mentions (`has_location`).

5.1 Regular Expression-based Rules

We created rules based on syntactic patterns (i.e., word order) that were observed in sentences, and implemented pattern-matching using regular expressions (regexes) to extract related biodiversity entities. Given an input sentence, I, that is a sequence of tokens $[t_0, t_1, ..., t_n]$, we firstly categorized every token t_i according to the following

Token type	Symbol	Description	Entity type
source	S	a token that belongs to a named entity type identified as a source category	reproductive condition or habitat
target	T	a token that belongs to a named entity type identified as a target category	temporal expression or geographic location
delimiter	d	a token that is a separator in an enumeration	comma or semicolon
other	o	any token that is neither a part of a named entity nor a delimiter	

Table 4: Types of tokens we designed for the regular expression-based rules.

types: *source*, *target*, *delimiter*, and *other* as shown in Table 4. We define *source* as a token that belongs to a named entity identified as a source entity type, i.e., either reproductive condition (for `has_time` relations) or habitat (for `has_location` relations). Meanwhile, *target* is a token that belongs to a named entity considered to be a target entity type, i.e., temporal expression (for `has_time` relations) or geographic location (for `has_location` relations). *Delimiter* is a token that acts as a separator in an enumeration, i.e., a comma or semicolon. Any token that is neither a part of a named entity nor a delimiter is categorized as *other*. We convert each token t_i into a character representation of the token's type. Hence, we convert a sentence into a string of characters, wherein each character is either `S` (source), `T` (target), `d` (delimeter), or `o` (other). We use this sequence of token types as input to our regex method implemented using Python's regular expression module, `re`. To extract relations, we created the following regex rules:

1. `[S]+(o)?(To|Td|T)+` – *source* token that may or may not be followed by one *other* token, then followed by one or more *target* tokens that may or may not be delimited by any token, and

2. `(?<!S)(To|Td|T)*T(o)?[S]+` – one or more *target* tokens that may or may not be delimited by any token that is not immediately preceded by a *source* token, and followed by a *source* token that may or may not be preceded by one *other* token.

The entity spans (i.e., source and target tokens) that match the patterns above are perceived to be related, and are given the value 1 for y. Figure 1 shows a sample sentence with a text span that matches regex rule 1 above.

5.2 Transformer-based Models

We cast our RE problem as a natural language inference (NLI) problem that we address using a transformer-based model. NLI is the task of determining whether a *hypothesis* is true (`entailment`), false (`contradiction`), or unverifiable (`neutral`) given a *premise* which corresponds to some known knowledge about the subject. We selected the Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020) as our model. Underpinned by a transformer encoder-decoder architecture (Vaswani et al., 2017), T5 casts various NLP tasks (e.g., machine translation, text classification, question answering) as a sequence-to-sequence learning problem, therefore producing outputs via text generation. NLI is one of the downstream NLP tasks that T5 was already fine-tuned on. We used the T5-large model[3] specifically, with 770 million parameters. Given an input sentence I and two entities E_S and E_T for which we wish to determine whether a relation holds,[4] we systematically generate a premise-hypothesis pair which serves as input to the NLI model. Specifically, the input sentence I is taken as the premise, while the hypothesis is created by populating either of the following sentence templates with E_1 and E_2:

- *The <habitat> was in <geographic location>.*

- *The <reproductive condition> event happened on <temporal expression>.*

[3] Available at https://huggingface.co/t5-large

[4] The two entities are considered only if one of them is a reproductive condition mention (E_S) and the other is a temporal expression (E_T), or if one of them is a habitat mention (E_S) and the other is a geographic location (E_T). This, respectively, means that we are aiming to determine if a `has_time` or `has_location` relation possibly holds between them.

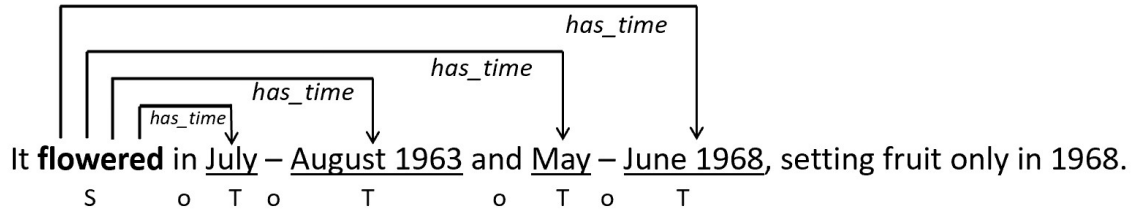

Figure 1: Example sentence with entity pairs that matched the rule `[S]+(o)?(To|Td|T)+`, where `S` corresponds to reproductive condition (in bold) which is the *source* entity, `T` corresponds to temporal expression (underlined) which is the *target* entity and `o` refers to *other* tokens. Source of example sentence: Medway, L. (1972). Phenology of a tropical rain forest in Malaya. Biological Journal of the Linnean Society, 4(2), p128.

Table 5 provides some example inputs for the T5-based NLI model, and its expected outputs. Due to variations in noun forms or verb tenses, the automatically generated hypothesis may not necessarily be grammatically correct; for instance, the example for the `has_time` relation in Table 5 would be more correct if it reads *"The fruiting event happened on August 1963"*. Nevertheless, we did not carry out any engineering on our templates to handle such variations, as we expected the transformers-based NLI model to be robust to such grammatical errors.

For our purposes, we say that a relationship between two entities exists if the model's predicted class is `entailment`, otherwise the entities are considered to be unrelated.

5.3 Hybrid Approach: Rules and Transformers

In order to improve performance and reduce required computational resources, we designed a two-step solution to our RE problem. Here, we combined our rule-based syntactic pattern matching and transformer-based approaches. The first step is to extract relations using our regex rules. These are the regular expressions we designed to extract consecutive entities in a sentence. The instances that were not identified as not pertaining to any relations using the first step, are fed into the second step. In this step, our transformer-based model is applied on the remaining instances. This step produces a set of related entities using less computational resources compared to running the transformer-based model on the entire dataset.

We investigated the incorporation of an enhancement to our hybrid approach: the use of compound entities in filling in the hypothesis templates instead of using single entity mentions, where applicable. We designed rules to identify multiple, consecutive entities in a given sentence that belong to the same entity type and thus comprise a compound entity $E_{compound}$. The regular expression that was designed to extract $E_{compound}$ is `(Et|E){2,}`, where `E` is a named entity of a specific type, and `t` is any token. $E_{compound}$ consists of consecutive entities belonging to the same entity type `E`, which may or may not be delimited by a token (`t`). For example, given the sentence *"It flowered in July - August 1963 and May - June 1968, setting fruit only on 1968."*, the reproductive condition is expressed by the mention *"flowered"* and the compound temporal expression is *"July - August 1963 and May - June 1968"*. Instead of populating a hypothesis template for every temporal expression, we formulated only one hypothesis: *"The flowered event happened on July - August 1963 and May - June 1968."*

6 Evaluation and Results

In this paper, we designed rules based on our training set. We tested and refined these rules on a held-out development (dev) set, and evaluated their performance using the test set. Table 6 presents the performance of our rule-based, transformer-based, and hybrid RE methods in terms of precision, recall and F1-score.

We applied the regular expression rules we created to extract related entities in a given sentence. This approach resulted in 100% precision for both relation types. This means that our regex-based rules can reliably identify positive samples, i.e., correct relations. However, this approach obtains poor recall, i.e., 33.91% and 36.96% for `has_time` and `has_location` relations, respectively, implying that our rules fail to identify many correct relations. This results in the lowest F1-scores (50.64% and 53.97%, respectively) among the methods we investigated. Even using a simple co-occurrence approach obtains much higher F1-scores: 94.57%

Relation Type	Hypothesis Template	Examples		
		Premise	**Hypothesis**	**NLI Output**
has_location	The **<habitat>** was in **<geographic location>**.	*Bukit Sai and Lesong belong to the lowland dipterocarp forest types with D. aromatica being the predominant species.*	*The **lowland dipterocarp forest** was in **Bukit Sai**.*	entailment
has_time	The **<reproductive condition>** event happened on **<temporal expression>**.	*It flowered in July - August 1963 and May - June 1968 , setting fruit only in 1968.*	*The **fruit** event happened on **August 1963**.*	contradiction

Table 5: Examples of populated hypothesis templates for generating inputs (premise-hypothesis pairs) for the NLI model, together with the corresponding expected outputs by the NLI model. A relation holds between two given entities (in bold) only if the NLI model predicts entailment as the output label. Source of the first input sentence (premise): Lee, S. L. (2000). Mating system parameters of Dryobalanops aromatica Gaertn. f.(Dipterocarpaceae) in three different forest types and a seed orchard. Heredity, 85(4), p339, and Medway, L. (1972). Phenology of a tropical rain forest in Malaya. Biological Journal of the Linnean Society, 4(2), p128.

for has_time and 84.02% for has_location relations.

To evaluate our transformer-based approach, we applied our chosen T5 model on the NLI task, building upon the Huggingface library[5]. Our evaluation on the test set yielded F1-scores higher than our rule-based method. For the has_location relation type, our transformer method produced an F1-score of 84.75%, which is slightly higher than that of the co-occurrence-based method (84.02%). However, the transformer-based method was outperformed by the co-occurrence-based one by 7.59 percentage points (86.98% vs 94.57% in terms of F1-score).

It is noticeable that combining our rule-based approach with the transformer model to form a hybrid approach improved the F1-score for the has_time relation type from 86.98% to 89.61%, and from 84.75% to 85.39% for the has_location relation type. Apart from improved performance, our hybrid approach is also more efficient, in that it requires the application of the more computationally expensive transformer models only on instances that were not classified by the rule-based approach as pertaining to relations.

We further improved our hybrid approach by using compound entities identified using regex rules in generating premise-hypothesis pairs for the transformer model, instead of separate single entities. We evaluated this method (referred to as 'hybrid + compound entities' in Table 6) on our test set and we observed that it obtained the highest F1-scores among all our investigated methods. Specifically, it led to an F1-score of 96.75% for has_time relations, and 89.90% for has_location relations.

7 Discussion

Our rule-based method is the most precise among all the methods we developed in this study. However, it is also the method that yielded the lowest recall, missing to identify more than half of true relations in the test set. The rule-based approach is suitable for applications that cannot compromise on precision, e.g., systems that support clinical decisions or automatic curation of databases. Its main drawback, however, is its reliance on syntactic similarity only, i.e., solely on patterns found within sentences. Thus, it is not robust to noisy data; any deviation from the expected sentence structure that is captured by the rules, would affect the performance of the method.

Among the methods presented in this paper, our transformer-based method is the most straightforward to implement. It is based on the population of natural-language hypothesis templates with named

[5]Available at https://github.com/huggingface/transformers

RE Approach	has_time			has_location		
	P	R	F1	P	R	F1
Co-occurrence	89.69%	100.00%	94.57%	72.44%	100.00%	84.02%
Regex-based rules	100.00%	33.91%	50.64%	100.00%	36.96%	53.97%
Transformer (T5)	97.16%	78.74%	86.98%	88.24%	81.52%	84.75%
Hybrid	97.31%	83.05%	89.61%	88.37%	82.61%	85.39%
Hybrid + compound entities	95.26%	98.28%	**96.75%**	83.96%	96.74%	**89.90%**

Table 6: Precision (P), Recall (R), and F1-score (F1) of our RE Methods on the test set for has_time and has_location relations.

entities, which are then fed to the NLI model (together with their corresponding premise). The transformer model paired with our hypothesis templates for RE provided us with F1-scores higher than those obtained by our rule-based method.

Our hybrid approach combines the strengths of the rule-based method (i.e., high precision) and the transformer-based model (i.e., high recall). This approach increased the recall for has_time relations by 4.28 percentage points and the recall for has_location relations by 1.09 percentage points, respectively. Error analysis of a small sample of instances from the development dataset showed that the hybrid method failed to identify relations that involve entities in an enumeration. For example, in the sentence *"Ashton et al (1988) record the extent of mass flowerings in peninsular Malaysia and Borneo for the period 1950 - 1983 based on state forest department records (table 5)"*,[6] the hybrid approach failed to determine that there is a relationship between *"mass flowerings"* and *"1983"*. Thus, as an enhancement to the hybrid method, we created regex rules to identify compound entities in sentences, as described in Section 5.3. Where they exist, these compound entities were used in populating the hypothesis templates, instead of individual named entities. With the incorporation of this step, the recall of the hybrid model was improved by 14-15 percentage points (i.e., 98.28% vs 83.05% for the has_time and 96.74% vs 82.61% for the has_location relations). This improved version of the hybrid approach ('hybrid + compound entities') provided us with the best F1-scores for both relation types (96.75% for has_time and 89.90% for has_location), among all the approaches we developed. These results demonstrate the role that rule-based methods can still play in complementing state-of-the-art DL approaches, i.e., transformers, enabling us to obtain optimal performance in RE.

[6]Source: Appanah, S. (1993). Mass flowering of dipterocarp forests in the aseasonal tropics. Journal of Biosciences, 18, p463.

8 Conclusions and Future Work

In this paper, we present our unsupervised relation extraction methods to extract relationships pertaining to habitats and reproductive conditions of plant species as described in text. These methods include: (1) regular expression-based rules; (2) transformer-based models for NLI; (3) a hybrid approach combining our rules and transformer model; and (4) an improved hybrid approach that captures compound entities. Our rule-based method undepinned by regexes obtained the highest precision but lowest recall. Meanwhile, our transformer-based method, which is based on the systematic generation of premise-hypothesis pairs as input for a T5-based NLI model, resulted in F1-score values higher than those produced by the regex rules. The strengths of the rule- and transformer-based methods are combined in our hybrid approach. With the incorporation of compound entities in the generation of NLI inputs, our hybrid approach produced the best performance, with F1-scores of 96.75% for the has_time relation type, and 89.90% for the has_location relation type. Our work shows that even without a large labeled training dataset, it is viable to extract – with satisfactory performance – relations between entities from biodiversity literature. This also shows that the combination of rules or pattern-based methods with state-of-the-art transformer models can lead to an improvement in the performance of RE, compared with a method that is solely based on transformers.

For our future work, we plan to compare our

hybrid approach with state-of-the-art zero-shot relation extraction methods, e.g., those proposed by Tran et al. (2022) and Najafi and Fyshe (2023), evaluating it on other datasets such as FewRel (Han et al., 2018) and WikiZSL (Chen and Li, 2021) which were drawn from the general domain. Furthermore, we will explore using other transformer-based models and formulating RE in terms of other downstream tasks, e.g., question answering. We also intend to integrate our hybrid approach into an application, i.e., an information extraction pipeline that can automate the curation of information from literature to populate a biodiversity-focused database.

Limitations

For this work, we focused on the requirements of a biodiversity-focused project, which is concerned with extracting information about the distribution and reproductive patterns of species in the *Dipterocarpaceae* (dipterocarps) family. We have evaluated the performance of our RE methods only on the dataset described above, and not on a wider range of datasets. The main reason for this is the lack of other datasets (drawn from the biodiversity domain) that are concerned with similar relation types. It is also worth noting that our RE methods are able to extract intra-sentential relations only, i.e., relations between entities within the same sentence.

Ethics Statement

For this work, we used an already existing dataset drawn from the biodiversity domain, that does not pose any data protection issues given that the documents comprising the corpus are all publicly available. All annotators volunteered to carry out the annotation task. They were trained and were explicitly informed that their work will be utilized in this study. Institutional ethical approval was not required as no personal data was collected, and the data that was being annotated did not contain any sensitive information.

Acknowledgements

We thank our annotators for their valuable work in annotating our RE dataset.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, DL '00, pages 85–94, New York, NY, USA. Association for Computing Machinery.

Abduladem Aljamel, Taha Osman, and Giovanni Acampora. 2015. Domain-Specific Relation Extraction - Using Distant Supervision Machine Learning:. In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 92–103, Lisbon, Portugal. SCITEPRESS - Science and and Technology Publications.

Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: Towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3470–3479, Online. Association for Computational Linguistics.

Jiayang Cheng, Haiyun Jiang, Deqing Yang, and Yanghua Xiao. 2021. A Question-answering Based Framework for Relation Extraction Validation. ArXiv:2104.02934 [cs].

Laura Chiticariu, Yunyao Li, and Frederick R. Reiss. 2013. Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 827–832, Seattle, Washington, USA. Association for Computational Linguistics.

Aron Culotta and Jeffrey Sorensen. 2004. Dependency Tree Kernels for Relation Extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 423–429, Barcelona, Spain.

Kartik Detroja, C. K. Bhensdadia, and Brijesh S. Bhatt. 2023. A survey on Relation Extraction. *Intelligent Systems with Applications*, 19:200244.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cícero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying Relations by Ranking with Convolutional Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 626–634, Beijing, China. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2021. Event Extraction by Answering (Almost) Natural Questions. ArXiv:2004.13625 [cs].

Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2007. RelEx—Relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

R. S. Gabud, R. T. Batista-Navarro, V. Y. Mariano, E. R. Mendoza, and S. L. Yap. 2019. Literature mining on dipterocarps: towards better informed natural regeneration and reforestation in Luzon, Philippines. *The Technical Journal of Philippine Ecosystems and National Resources*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

N. Kambhatla. 2004. Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Information Extraction.

Andreas Korger and Joachim Baumeister. 2021. Rule-based Semantic Relation Extraction in Regulatory Documents.

Nicolas Le Guillarme and Wilfried Thuiller. 2022. TaxoNERD: Deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods in Ecology and Evolution*, 13(3):625–641. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13778.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-Shot Relation Extraction via Reading Comprehension. ArXiv:1706.04115 [cs].

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2022. A Unified MRC Framework for Named Entity Recognition. ArXiv:1910.11476 [cs].

ChunYang Liu, WenBo Sun, WenHan Chao, and WanXiang Che. 2013. Convolution Neural Network for Relation Extraction. In *Advanced Data Mining and Applications*, Lecture Notes in Computer Science, pages 231–242, Berlin, Heidelberg. Springer.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event Extraction as Machine Reading Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Scott Miller, Heidi Fox, Lance Ramshaw, and Ralph Weischedel. 2000. A Novel Use of Statistical Parsing to Extract Information from Text. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Saeed Najafi and Alona Fyshe. 2023. Weakly-supervised questions for zero-shot relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3075–3087, Dubrovnik, Croatia. Association for Computational Linguistics.

Nhung T.H. Nguyen, Roselyn S. Gabud, and Sophia Ananiadou. 2019. COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiversity Data Journal*, (7):e29626.

Nhung TH Nguyen, Makoto Miwa, Yoshimasa Tsuruoka, Takashi Chikayama, and Satoshi Tojo. 2015. Wide-coverage relation extraction from MEDLINE using deep syntax. *BMC Bioinformatics*, 16(1):107.

Yifan Peng, Chih-Hsuan Wei, and Zhiyong Lu. 2016. Improving chemical disease relation extraction with rich features and weakly labeled data. *Journal of Cheminformatics*, 8(1):53.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv:1910.10683 [cs, stat].

K.E. Ravikumar, Majid Rastegar-Mojarad, and Hongfang Liu. 2017. BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences. *Database*, 2017:baw156.

Anu Thomas and Sangeetha Sivanesan. 2022. An adaptable, high-performance relation extraction system for complex sentences. *Knowledge-Based Systems*, 251:108956.

Van-Hien Tran, Hiroki Ouchi, Taro Watanabe, and Yuji Matsumoto. 2022. Improving discriminative learning for zero-shot relation extraction. In *Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*, pages 1–6, Dublin, Ireland and Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thang Vu, Heike Adel, Pankaj Gupta, and Hinrich Schütze. 2016. Combining Recurrent and Convolutional Neural Networks for Relation Classification.

Bernhard Waltl, Georg Bonczek, and Florian Matthes. 2018. Rule-based Information Extraction: Advantages, Limitations, and Perspectives. *Jusletter IT*, (4).

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation Classification via Multi-Level Attention CNNs. pages 1298–1307.

Xinzhi Wang, Jiahao Li, Ze Zheng, Yudong Chang, and Min Zhu. 2022. Entity and relation extraction with rule-guided dictionary as domain knowledge. *Frontiers of Engineering Management*, 9(4):610–622.

Lang-Tao Wu, Jia-Rui Lin, Shuo Leng, Jiu-Lin Li, and Zhen-Zhong Hu. 2022. Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web. *Automation in Construction*, 135:104108.

Minguang Xiao and Cong Liu. 2016. Semantic Relation Classification via Hierarchical Recurrent Neural Network with Attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1254–1263, Osaka, Japan. The COLING 2016 Organizing Committee.

Na Xu, Hong Chang, Bai Xiao, Bo Zhang, Jie Li, and Tiantian Gu. 2022. Relation Extraction of Domain Knowledge Entities for Safety Risk Management in Metro Construction Projects. *Buildings*, 12(10):1633. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2002. Kernel methods for relation extraction. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 71–78, USA. Association for Computational Linguistics.

Chunju Zhang, Xueying Zhang, Wenming Jiang, Qijun Shen, and Shanqi Zhang. 2009. Rule-Based Extraction of Spatial Relations in Natural Language Text. In *2009 International Conference on Computational Intelligence and Software Engineering*, pages 1–4.

Dongxu Zhang and Dong Wang. 2015. Relation Classification via Recurrent Neural Network. ArXiv:1508.01006 [cs].

Youwen Zhao, Xiangbo Yuan, Ye Yuan, Shaoxiong Deng, and Jun Quan. 2023. Relation extraction: advancements through deep learning and entity-related features. *Social Network Analysis and Mining*, 13(1):92.

Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. 2019. Graph Neural Networks with Generated Parameters for Relation Extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1331–1339, Florence, Italy. Association for Computational Linguistics.