frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 23 May 2024
DOI 10.3389/frai.2024.1371411

OPEN ACCESS

EDITED BY
Lili Mou,
University of Alberta, Canada

REVIEWED BY
Mauajama Firdaus,
University of Alberta, Canada
Alaa Mohasseb,
University of Portsmouth, United Kingdom

*CORRESPONDENCE
Roselyn Gabud
✉ rsgabud@up.edu.ph

RECEIVED 16 January 2024
ACCEPTED 10 May 2024
PUBLISHED 23 May 2024

CITATION
Gabud R, Lapitan P, Mariano V, Mendoza E, Pampolina N, Clariño MAA and Batista-Navarro R (2024) Unsupervised literature mining approaches for extracting relationships pertaining to habitats and reproductive conditions of plant species. *Front. Artif. Intell.* 7:1371411. doi: 10.3389/frai.2024.1371411

Unsupervised literature mining approaches for extracting relationships pertaining to habitats and reproductive conditions of plant species

Roselyn Gabud[1,2]*, Portia Lapitan[3], Vladimir Mariano[4], Eduardo Mendoza[2,5,6,7], Nelson Pampolina[3], Maria Art Antonette Clariño[2] and Riza Batista-Navarro[2,8]

[1]Department of Computer Science, College of Engineering, University of the Philippines Diliman, Quezon City, Philippines, [2]Institute of Computer Science, College of Arts and Sciences, University of the Philippines Los Baños, Laguna, Philippines, [3]Department of Forest Biological Sciences, College of Forestry and Natural Resources, University of the Philippines Los Baños, Laguna, Philippines, [4]Young Southeast Asian Leaders Initiative (YSEALI) Academy, Fulbright University Vietnam, Ho Chi Minh City, Vietnam, [5]Mathematics and Statistics Department, De la Salle University, Manila, Philippines, [6]Center for Natural Science and Environmental Research, De la Salle University, Manila, Philippines, [7]Max Planck Institute of Biochemistry, Munich, Germany, [8]Department of Computer Science, University of Manchester, Manchester, United Kingdom

Introduction: Fine-grained, descriptive information on habitats and reproductive conditions of plant species are crucial in forest restoration and rehabilitation efforts. Precise timing of fruit collection and knowledge of species' habitat preferences and reproductive status are necessary especially for tropical plant species that have short-lived recalcitrant seeds, and those that exhibit complex reproductive patterns, e.g., species with supra-annual mass flowering events that may occur in irregular intervals. Understanding plant regeneration in the way of planning for effective reforestation can be aided by providing access to structured information, e.g., in knowledge bases, that spans years if not decades as well as covering a wide range of geographic locations. The content of such a resource can be enriched with literature-derived information on species' time-sensitive reproductive conditions and location-specific habitats.

Methods: We sought to develop unsupervised approaches to extract relationships pertaining to habitats and their locations, and reproductive conditions of plant species and corresponding temporal information. Firstly, we handcrafted rules for a traditional rule-based pattern matching approach. We then developed a relation extraction approach building upon transformer models, i.e., the Text-to-Text Transfer Transformer (T5), casting the relation extraction problem as a question answering and natural language inference task. We then propose a novel unsupervised hybrid approach that combines our rule-based and transformer-based approaches.

Results: Evaluation of our hybrid approach on an annotated corpus of biodiversity-focused documents 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. Our work shows that even without training models on any domain-specific labeled dataset, we are

able to extract relationships between biodiversity concepts from literature with satisfactory performance.

KEYWORDS

relation extraction, information extraction, unsupervised methods, rule-based methods, transformer models, biodiversity

1 Introduction

Plants provide food to humans and other terrestrial animals, are habitats to more than 80% of terrestrial species, are the source of clean air and water, and regulate climatic balance (UNDP, 2023). However, there has been a continuous decline in the world's forests and biodiversity (FAO, 2019; FAO and UNEP, 2020). A major contributor to this decline in natural resources is the growing population on Earth, currently around 8 billion, that increases global demand for food and other commodities. Climate change has also curtailed vegetation eventually affecting food production in a number of locations (Lobell et al., 2011; Ray et al., 2019).

In order to ensure that the benefits of land-based ecosystems will be enjoyed for generations to come, the United Nations Sustainable Development Goal (SDG) #15, Life on Land, focuses specifically on managing forests sustainably, halting and reversing land and natural habitat degradation, successfully combating desertification and stopping biodiversity loss (UNDP, 2023). Land restoration and rehabilitation require knowledge of plant species' reproduction and regeneration properties. Precise timing in fruit collection and knowledge of reproductive activities is necessary especially in the tropical regions (Luna-Nieves et al., 2017) where many tree seeds are recalcitrant (Barbedo et al., 2013) and have very short viability, lasting only a few weeks or months under normal conditions (Oshima et al., 2015). Reproductive activities such as flowering and fruiting in the case of plants, are timed events within the species' life cycle that are associated with seasonal timings (Amasino, 2010). As a life event, reproduction is studied in association with its coupling factor, i.e., time (Amasino, 2010; Ehrlén, 2015). Thus, it is impossible to understand reproduction without the context of time. Reproductive patterns may occur in irregular intervals for long periods of time, and may even be affected by habitat variations.

Understanding species' habitat preferences, i.e., the natural home or environment, is another aspect of successful reforestation (Poulin et al., 2013; Lelli et al., 2018; Staples et al., 2020). With the changing ecosystem affected by factors such as climate change, deforestation, and desertification (Cochard, 2001; Gebeyehu and Hirpo, 2019), a complete knowledge of "what grows where," that traditionally used to be geographic-based distribution of species, is now being augmented by more data-driven approaches that associate information on habitats with distribution data (Morueta-Holme and Svenning, 2018). Hence, information on plant species' reproductive conditions with associated temporal information, and habitats with associated geographic locations will aid in more informed reforestation and cultivation of land.

Sustainable management of land and vegetation must be fueled by long-term and broad-scale understanding of the biology underpinning the reproduction of plant species (Gabud et al., 2019), the foundation of which is biodiversity data available in either structured or unstructured form. There exist widely-used biodiversity databases that contain structured information on species and their occurrences such as the Global Biodiversity Information Facility (GBIF) with about two billion occurrence records (GBIF, 2023), and the Atlas of Living Australia (ALA) containing around 100 million occurrence records (ALA, 2023). Both GBIF and ALA use data standards such as Darwin Core (Wieczorek et al., 2012) to mobilize and deliver biodiversity data. Darwin Core is the internationally agreed data standard that includes a glossary of terms, i.e., fields or attributes, intended to facilitate the sharing of information on biodiversity by providing identifiers, labels, and definitions (TDWG, 2023). Darwin Core is primarily based on taxa, their occurrence, and related information which is reflected in the types of data stored in GBIF and ALA, e.g., species' occurrence data. Darwin Core includes terms that are used to represent information on species' reproductive condition and habitats, namely, `Reproductive Condition` and `Habitat`; they can, for example, be populated with values such as *"in bloom"* and *"oak savanna,"* respectively. While both GBIF and ALA publish data on species that include these two fields, a substantial number of their occurrence records have these fields blank and unpopulated. For example, in GBIF, the 61,392 occurrence records for *Dipterocarpaceae* (a family of tropical rainforest trees) provide information on reproductive condition and habitat for only 37,624 and 4,332 records, respectively. Meanwhile, in ALA, out of the 1,248 occurrences of a similar family, *Dipterocarpaceae Blume*, only 419 and 870 provide reproductive condition and habitat information, respectively.

Although there is a growing movement to make research data more reusable and accessible in structured form, scientific literature still remains a major repository for much of our knowledge about the natural world and represents centuries of investment (Thessen et al., 2012; Le Guillarme and Thuiller, 2022). This is because the research literature often contains detailed descriptions of the data collection methods and analytical procedures used, which can be essential for understanding and interpreting findings. For instance, detailed information on species reproduction and habitat is narrated well in literature, e.g., *"In 1981, flowering was heavy in Kepong and Pasoh, but weak in Gombak and Ampang"* and *"The main observation site was conserved forest at Dongmakhai."* Currently, over 60 million pages of legacy biology text are scanned and made available online through the Biodiversity Heritage Library (BHL, 2023) and thousands of new digital pages are published every month in open-access biology and ecology journals (Cornford et al., 2021). This represents an enormous amount of unstructured information that can potentially be exploited in data-driven studies. If tools are developed to automatically extract fine-grained information from these sources and feed such

information into open biodiversity databases in a structured form, then this information will be much more accessible and more useful for large-scale studies (Le Guillarme and Thuiller, 2022). This could potentially allow researchers to better understand the relationships between species and their habitats, and to develop more effective conservation strategies.

Information extraction (IE) is an umbrella term for tasks that seek to automatically extract structured information from unstructured text. With the exponential growth of digitized literature over the years, IE has become increasingly pertinent, due to its role in (semi-)automatically populating databases with content (Ravikumar et al., 2015; Lee et al., 2018; Paragkamian et al., 2022). Relation extraction (RE) is an IE task that is concerned with the identification of semantic relationships between entities or concepts in text. This task predicts whether a relationship holds between two entities (or concepts), based on the context of the sentence. For example, in the sentence *"The flowering commenced in July and continued until October 2001, with viable seed fall occurring from early December 2001 until early February 2002,"* a relation extraction system should be able to identify the relationship between the reproductive condition mention *"flowering"* and the temporal expression *"October 2001,"* but no relation between *"flowering"* and *"December 2001."* This information can prove to be crucial in analyzing the reproductive behavior of a species of interest. Harvesting these details from literature will enable big data-centric discovery focused on understanding plant species' reproductive patterns and habitats.

In this paper, we focus on extracting the following, with a view to enabling the curation of biodiversity databases: (1) relationships between expressions denoting a plant species' reproductive condition and the time or duration when those conditions hold, and (2) relationships between habitats and their geographic locations. Our work will thus ultimately support the analyses required for planning efforts in plant regeneration, and land restoration and rehabilitation. The main contributions of our work include: (1) the creation of a new RE corpus—drawn from the biodiversity-focused corpus presented in the work of Gabud et al. (2019)—in which the above-mentioned relationship types were manually labeled, thus providing a new gold standard dataset; and (2) a comparative evaluation and combination of two types of unsupervised approaches: rule-based and transformer-based methods for RE.

In the remainder of this paper, we first provide a review of prior work related to our study (Section 2). This is followed by a formalization of the problem we are aiming to solve (Section 3) and a description of the dataset we developed and used in our experiments (Section 4). Importantly, we present details of the various unsupervised RE approaches that we developed (Section 5), and the results of evaluating and combining them into one hybrid model (Section 6). We then analyze our results, discuss their implications and limitations (Section 7) before providing a summary of our findings and directions for future work (Section 8).

2 Related work

In this section, we provide a review of previously reported work on natural language processing and information extraction more specifically, that have been applied to the biodiversity domain and are considered to be relevant to our own work.

2.1 Text mining and natural language processing

The growing volume of scientific literature on biodiversity has led to a focus on the development of computational methods for extracting meaningful information from unstructured textual data (Farrell et al., 2022; Paragkamian et al., 2022). This computational task is known as text mining, and it has been used to identify trends, patterns, and relationships that would otherwise be difficult to detect. Text mining has successfully been applied to biodiversity literature (Batista-Navarro et al., 2016, 2017; Gabud et al., 2017; Parr and Thessen, 2018; Chaix et al., 2019; Lee et al., 2019; Page, 2019). There has been significant progress in the development of natural language processing (NLP) and machine learning (ML) algorithms that can be used to automatically annotate taxonomic text (Lücking et al., 2022), identify taxonomic names in text (Le Guillarme and Thuiller, 2022), link person names to biodiversity data (Groom et al., 2020), and extract phenotypic traits from text (Thessen et al., 2018). However, more tools and services are needed to scale up the accessibility of biodiversity data (Thessen et al., 2022).

2.2 Information extraction

Information extraction (IE) is a specific task within NLP that deals with the automatic process of extracting structured information from unstructured data sources such as scientific publications and books. One IE subtask that has gained much attention is Named Entity Recognition (NER), i.e., the task of identifying named entities such as taxonomic names. Most work on NER in the biodiversity domain have been focused on the extraction of taxonomic names. Early work on taxonomic NER systems were rule-based (Koning et al., 2005; Sautter et al., 2006) and made use of handcrafted rules based on regularities in taxon naming conventions to recognize mentions of taxa in text. Researchers also proposed dictionary-based approaches that match text against a predefined dictionary of species names (Leary et al., 2007; Gerner et al., 2010; Pafilis et al., 2013). However, taxonomic NER systems that were developed more recently employ either a deep neural network-based approach (Le Guillarme and Thuiller, 2022) or a hybrid approach combining ML with rules and dictionaries (Mozzherin, 2022; Thessen et al., 2022).

NER methods that consider other types of entities include the ML-based models developed by Ahmed et al. (2019) and Nguyen et al. (2019) for recognizing taxon names, person names, locations and temporal expressions in biodiversity literature. More recently, Mora-Cross et al. (2022) employed ML to extract plant phenological data from specimen labels, which is a textual document that follows a specific format. In our work, we assume that recognized named entities are provided as part of the input to our RE methods. While the development of biodiversity NER methods is outside of the scope of this paper, it is important to

note that pre-identified mentions of reproductive conditions, time, habitats, and locations are required for RE.

2.3 Relation extraction

Relation extraction (RE) is another IE task that involves detecting pre-defined semantic relationships between entities mentioned in text (Song et al., 2022). There are a few available methods for relation extraction in the biodiversity domain. The work by Chaix et al. (2019) focused on extracting *lives_in* relationships between bacteria species and the location where it lives, which can either be a habitat or a geographic entity. Nguyen et al. (2019), due to lack of gold standard annotated relations, employed an unsupervised pattern-based method to identify binary relations that pertain to the occurrence of species in specific geographic locations, habitats, or points of time. Kopperud et al. (2022) extracted related taxonomic names and geographic locations by training an ML-based relation classifier to determine if a given sentence explicitly stated or strongly implied that a given taxon was found in the mentioned location, or not.

RE approaches can be categorized into three: early rule-based, supervised, and zero-shot learning approaches. Early methods for extracting relations between entities were based on rule templates that were created by experts (Fundel et al., 2007; Zhang et al., 2009; Nguyen et al., 2015). These rules were designed to capture the syntactic patterns that are associated with different types of relations, as observed in a corpus. Rules have the advantage of being highly interpretable, as they can be easily understood by humans. However, rule-based methods have two main limitations: they can be time-consuming to create and they are domain-dependent. To alleviate the burden on human experts, methods to automatically construct rules were developed (Carlson et al., 2010; Zheng et al., 2019).

Supervised methods for RE have been extensively studied (Yan et al., 2021). Early work on superpervised methods trained an ML-based model on a manually labeled dataset, and then employed the model to classify the relation. Models can be feature-based (Miller et al., 2000; Kambhatla, 2004) or kernel-based (Zelenko et al., 2002; Culotta and Sorensen, 2004). The more advanced Deep Neural Network (DNN) methods have been shown to outperform traditional supervised methods for RE. These DNN models learn higher-order, abstract feature representations from sentences. With the emergence of DNNs, models that employ neural architectures such as convolutional neural networks (CNNs; Liu et al., 2013; dos Santos et al., 2015), recurrent neural networks (RNN; Zhang and Wang, 2015; Vu et al., 2016), 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; Lee et al., 2022) have been utilized for RE tasks. Like traditional ML-based models, DNN-based models learn features from data. This gives them strong generalization ability, adaptability, and scalability. However, training or fine-tuning them for downstream applications such as RE requires labeled data (Zhao et al., 2023). Since we have limited labeled data, supervised methods such as those mentioned above cannot be applied in our study.

In recent years, approaches to IE that are considered to be *zero-shot*, i.e., requiring no labeled data, have started gaining momentum (Liu et al., 2020; Cheng et al., 2021; Du and Cardie, 2021; Li et al., 2022); these leverage models that were originally trained for other tasks, e.g., machine reading comprehension (MRC), question answering (QA). 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). These are models that were pre-trained on large-scale corpora using unsupervised learning objectives such as masked language modeling. These PLMs can be 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. Zero-shot methods have some advantages over traditional RE methods, including: (a) the ability to generalize and extract previously unseen relations; and (b) significant reduction in the labeling cost associated with RE because only a small amount of labeled data is required, i.e., test samples for evaluating the model. However, as the models underpinning zero-shot methods were trained on out-of-domain data, they might struggle to perform well on RE for specialized domains (such as biodiversity). One way to alleviate this issue is by constructing a hybrid model that combines the strengths of different types of approaches, e.g., zero-shot and rule-based approaches. Our work thus explores the development and evaluation of such a hybrid approach to RE.

2.4 Annotated corpora

Text collections that include manually provided annotations (also known as gold standard corpora) are valuable resources for NLP research as they support the development and evaluation of various methods. In recent years, gold standard corpora drawn from the biodiversity domain have increasingly become more available.

The first proposed biodiversity corpora contain taxonomic name annotations to support taxonomic NER. These are Linnaeus-100 (Gerner et al., 2010) that consists of 100 randomly selected full-text documents from PubMedCentral (PMC), and Species-800 (Pafilis et al., 2013) that is based on 800 PubMed abstracts. The latter is a collection of 100 abstracts from each of the following eight subject areas: bacteriology, botany, entomology, medicine, mycology, protistology, virology, and zoology, and thus contains annotations that represent many taxonomic groups.

Meanwhile, biodiversity corpora that were introduced more recently cover multiple types of entities pertaining to information on species occurrence reported in biodiversity literature. BIOfid (Ahmed et al., 2019) is a collection of German documents drawn from historical scientific literature on the biodiversity of plants, birds, moth and butterflies, that were converted to plain text by

optical character recognition (OCR). It includes annotations of taxon names, locations, temporal expressions, person names, and organization names. Similarly, documents in the COPIOUS corpus (Nguyen et al., 2019) were annotated to capture taxon names, locations, temporal expressions, and person names, as well as mentions of habitats. Thus far, COPIOUS is the largest biodiversity corpus that consists of 668 documents downloaded from the Biodiversity Heritage Library (BHL) with over 26K sentences and more than 28K annotated entities. Parallel to the development of COPIOUS, Gabud et al. (2019) developed DipteroMine, a biodiversity corpus that contains additional annotations pertaining to reproductive information, i.e., the reproductive condition of species. It consists of abstract-length documents: 250 from BHL, 150 from journal articles, and 100 from government reports. The most recent addition to the list of biodiversity corpora is BiodivNERE (Abdelmageed et al., 2022). Drawn from biodiversity dataset metadata and abstracts, it comes with two datasets, one supporting NER and the other supporting Relation Extraction (RE). In terms of named entity annotations, it covers six entity types, i.e., organism, environment, quality, location, phenomena, and matter. The RE dataset, meanwhile, is based on a subset of the NER dataset and includes annotations of binary relationships between different entity types, such as *occur_in* (between organism and environment), *influence* (between organism and process) and *have/of* (between quality and environment).

This work makes use of the journal articles included in the DipteroMine corpus (Gabud et al., 2019), a gold standard corpus for biodiversity NER that was designed in accordance with the annotation scheme used in the COPIOUS project (Nguyen et al., 2019). In addition to taxonomic names, geographic locations, and temporal expressions, the DipteroMine corpus contains manual annotations pertaining to habitat and reproductive condition mentions which are relevant to our study. Table 1 shows a comparison of existing corpora containing manually annotated biodiversity-relevant named entities and relations.

3 Problem formulation

We now provide a formal definition of our relation extraction task. 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.

- **`has_time` relation**: This relation 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. In this type of relation, the temporal expression provides additional information on the reproductive condition.
- **`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

In this work, we utilized documents that were drawn from the DipteroMine NER gold standard corpus proposed by Gabud et al. (2019), which took inspiration from the design of the COPIOUS dataset (Nguyen et al., 2019). The DipteroMine corpus was developed to support the extraction of information on the distribution and reproductive patterns of forest tree species belonging to the *Dipterocarpaceae* family, more commonly known as dipterocarps. It consists of one- to two-paragraph documents that were manually selected from online environmental science and ecology journal repositories, e.g., Journal of Tropical Ecology, Journal of Ecology, Journal of Biosciences, and Forest Ecology and Management. These scholarly articles on dipterocarps were retrieved using keywords such as *"flowering," "fruiting," 'mass flowering," "phenology,"* and *"masting."* Based on this process, a total of 151 abstract-length documents were included in the said corpus.

In our RE work, we are particularly interested in capturing relationships between the following types of entities: habitat, geographic location, temporal expression, and importantly, reproductive condition. The lattermost type includes expressions that pertain to the reproductive status of dipterocarps, or seasonal events involving them, e.g., *"sterility," "budburst,"* and *"flowering."* Table 2 provides descriptions and examples for each of our entity types of interest. From the DipteroMine corpus, we selected sentences that contain at least one 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 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. One relation type is the `has_time` relation which holds between a reproductive condition mention (E_S) and a temporal expression (E_T). The other type is the `has_location` relation which holds between a habitat mention (E_S) and a geographic location (E_T). Our dataset contains all occurrences of the E_S and E_T pairs found in every sentence in the corpus. For instance, if a sentence contains one habitat mention and two geographic location mentions, two `has_location` pairs are generated as relation instances, as exemplified in Table 3.

Two annotators manually provided the label y for each data instance (I, E_S, E_T, y). Based on the sentence input I, y is set to 1 if the two entities E_S and E_T have a relationship, and set to 0 otherwise. One annotator is a Biology degree holder, while the other annotator is a Computer Science student. They worked on the annotation task independently. We then randomly split the set of annotated instances into a training set (70%), development set (10%), and test set (20%).

TABLE 1 Characteristics of existing corpora with manual annotations of biodiversity-related named entities and relations.

Corpus	Doc type	Docs	Sent	Words	NE type (count)	RE type (count)
Linnaeus	PMC full paper	100	17,580	502,507	Taxon (4,259)	NA
S800	PubMed abstract	800	8,064	201,981	Taxon (3,708)	NA
BIOfid	German historical literature	969	15,833	Undisclosed	Taxon (15,085) Loc (6,785) Time (5,197) Person (5,393) Org (1,085) Other (7,849)	NA
COPIOUS	BHL pages	668	26,277	298,230	Taxon (12,227) Geo Loc (9,921) Temp Exp (2,889) Habitat (2,210) Person (1,554)	NA
BiodivNERE	Dataset metadata, PubMed abstract	150	2,398	102,113	Organism (2,602) Loc (310) Env (1,666) Matter (1,053) Phenomena (724) Quality (3,627)	Org-Env (392) Org-Mat (189) Org-Qua (865) Env-Env (6) Env-Mat (267) Env-Phe (430) Env-Loc (46) Env-Qua (744) Phe-Phe (4) Phe-Loc (27) Phe-Qua (300) Other (730)
DipteroMine	Online journal	151	5,045	99,359	Taxon (1,460) Geo Loc (711) Temp Exp (787) Habitat (475) Person (36) Rep Cond (539) Hab Att (115) Hab AttVal (126)	Rep-Tem (1,404) Hab-Loc (413)

"NA" stands for "not applicable."

TABLE 2 Descriptions and examples of our biodiversity concept types of interest.

Concept	Description	Example
Habitat	Environments in which organisms live.	*"In the [lowland mixed dipterocarp forests] of Borneo the Dipterocarpaceae can comprise roughly 107 of species"*
Geographic Location	Any identifiable point or area in the planet, ranging from continents, major bodies of water, named landforms, countries, states, cities, and towns.	*"The main obervation site was conserved forest at [Dongmakhai] ([18deg20'03"N, 102deg30'5"E], 190m a.s.l.)"*
Reproductive Condition	Indicators of the specimens' reproductive behavior.	*"There were two [flowerings] in March to May, and one in August during this period."*
Temporal Expression	Spans of text pertaining to points in time.	*"Most fruit fall occurred from the [end of July] to [mid-August]."*

TABLE 3 Sample `has_location` relation data instances for a sentence (I, E_S, E_T, y).

	Sentence (I)	Habitat (E_S)	Geo. Location (E_T)	y
1	*"The main observation site was conserved forest at Dongmakhai (18deg20'03"N, 102deg30'5"E, 190m a.s.l.)."*	conserved forest	Dongmakhai	1
2	*"The main observation site was conserved forest at Dongmakhai (18deg20'03"N, 102deg30'5"E, 190m a.s.l.)."*	conserved forest	18deg20'03"N, 102deg30'5"E	1

Source of sentence: Kato et al. (2008).

5 Methods

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

5.1 Rule-based approaches

We designed two traditional rule-based approaches for RE. We based our rules on syntactic patterns observed in the sentences in our training corpus. The first one extracts relations based on the distance of the dependency between a given pair of entities, and the second one implements pattern-matching using regular expressions.

5.1.1 Dependency distance

Dependency distance refers to the number of edges traversed between the head and dependent words along the shortest path in the dependency parse tree of a sentence. In a dependency parse tree, each word or token in a sentence is represented as a node, and the syntactic relationships between words are represented as directed edges. In this work, we used the dependency distance between two entity spans, E_S and E_T, that are contained in an input sentence I, as basis for RE. We used the Stanford Dependency Parser[1] to generate the dependency parse tree of each of the sentences in our dataset. Given an input sentence I, that is a sequence of tokens $[t_0, t_1, ..., t_n]$, the parser returns a sentence's dependency parse tree in the CoNLL-U format,[2] a standard means for storing dependency and feature structures of sentences. We implemented a pre-processing step to make the tab-separated value (TSV) CoNLL-U file include additional columns corresponding to the manual annotations for named entities, i.e., mentions of habitats, geographic locations, reproductive conditions, and temporal expressions. If a token t_i in an input sentence I belongs to a named entity, we add the entity's type to the token's FEATS field, i.e., the sixth column in the CoNLL-U file that contains a list of morphological features. Specifically, we added an additional feature called *biodiv*, and set its value to the named entity type following the BIO (Beginning, Inside, Outside) sequence labeling format, e.g., *biodiv=B-Habitat* for the first token of a habitat entity. Figure 1 shows an excerpt from an example processed CoNLL-U file. This file was then used as input to *Grew*,[3] a graph rewriting system for manipulating linguistic representations (Bonfante et al., 2018; Guillaume, 2021). We selected this tool as it readily supports pattern matching over linguistic representations (including those written in the CoNLL-U format). Furthermore, it is well-documented and is continuously being maintained.

Grew comes with a component called *Grew-match* that executes queries over documents in a given corpus, whereby a query is in the form of a pattern and all items matching the pattern are returned. An input pattern describes the nodes (node clauses) and relations (edge clauses) that must be found in the text of the documents. A *node clause* is a node described by an identifier and some constraints in its feature structure; an example of a node clause is *bdREP [biodiv=B-ReproductiveCondition, upos=NN]*, where *bdREP* is the identifier, and *biodiv=B-Habitat, upos=NN* are the constraints that should be met. Meanwhile, an *edge clause* specifies the existence of an edge between two nodes with or without additional constraints. For example, *bdREP→bdTEMP* describes an edge from the node with identifier *bdREP* to the one with identifier *bdTEMP* without any additional constraint, while *bdREP -[nsubj|obj]→bdTEMP* specifies that the edge label (i.e., dependency type) should be either *nsubj* (nominal subject) or *obj* (direct object). We supplied patterns to *Grew-match* in order to extract related entities whose dependency distance is an integer n. To determine the best value for n, we performed experiments on our development set, as described in Section 6.2.1.

We handcrafted *Grew-match* patterns that made use of nodes that are biodiversity named entities and edges without additional constraints, to extract related entities. An example of a pattern that extracts entities with a dependency distance of 2 is shown on Figure 2. Together with an input sentence I containing a feature structure in CoNLL-U format, our created patterns, when fed into *Grew-match*, were able to analyze sentences and extract entities separated by a dependency distance of n. Using this method, we are able to associate habitats with their geographic locations (`has_location` relation), and determine a species' reproductive condition at a specific point in time (`has_time` relation). Figure 3 presents the `has_time` relations extracted from an example sentence, using the pattern (with dependency distance $n = 2$) shown in Figure 2.

5.1.2 Regular expression-based rules

Apart from handcrafting patterns based on dependency distances, we also created rules to extract related biodiversity entities by observing syntactic patterns, i.e., word order, in the sentences they appear in. These patterns were then captured by a set of regular expressions (regexes). 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 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*.

1 Available at: https://nlp.stanford.edu/software/stanford-dependencies.html.

2 https://universaldependencies.org/format.html

3 https://grew.fr/

```
1	There	_	EX	EX	sentid=Ashton_1989_p226.conll_0003|chunk=B-NP	2	expl	_	_
2	were	_	VBD	VBD	chunk=B-VP	0	root	_	_
3	two	_	CD	CD	chunk=B-NP	4	nummod	_	_
4	flowerings	_	NNS	NNS	biodiv=B-ReproductiveCondition|chunk=I-NP	2	nsubj	_	_
5	in	_	IN	IN	chunk=B-PP	6	case	_	_
6	March	_	NNP	NNP	biodiv=B-TemporalExpression|chunk=B-NP	4	nmod	_	_
7	to	_	TO	TO	chunk=B-PP	8	case	_	_
8	May	_	NNP	NNP	biodiv=B-TemporalExpression|chunk=B-NP	6	nmod	_	_
9	,	_	,	,	_	4	punct	_	_
10	and	_	CC	CC	_	4	cc	_	_
11	one	_	CD	CD	chunk=B-NP	4	conj	_	_
12	in	_	IN	IN	chunk=B-PP	13	case	_	_
13	August	_	NNP	NNP	biodiv=B-TemporalExpression|chunk=B-NP	11	nmod	_	_
14	during	_	IN	IN	chunk=B-PP	16	case	_	_
15	this	_	DT	DT	chunk=B-NP	16	det	_	_
16	period	_	NN	NN	chunk=I-NP	2	nmod	_	_
17	.	_	.	.	_	2	punct	_	_
```

FIGURE 1
An excerpt from an example CoNLL-U file that was used as input for Grew. Source of sentence: Ashton (1989).

```
pattern {
        bdREP [biodiv=B-ReproductiveState | I-ReproductiveState];
        bdTEMP [biodiv=B-TemporalExpression | I-TemporalExpression];
        bdREP -> N1;
        N1 -> bdTEMP;
}
```

FIGURE 2
Example *Grew-match* pattern with a dependency distance of 2.

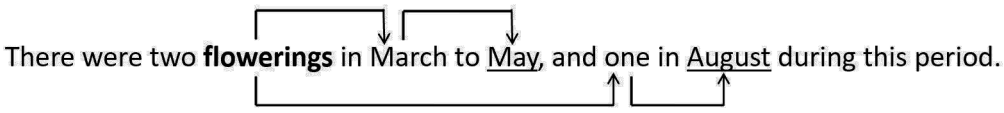

FIGURE 3
Related reproductive condition (in bold) and temporal expression (underlined) extracted using a *Grew-match* pattern with maximum dependency distance $n = 2$.

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

Token type	Symbol	Description	Entity type
Source	S	A token that belongs to a named entity category identified as a source category	Reproductive condition or habitat
Target	T	A token that belongs to a named entity category 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	

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. We formulated the two regular expressions above to extract related consecutive entities in a sentence. Figure 4 shows a sample sentence with a text span that matches regex rule 1 above.

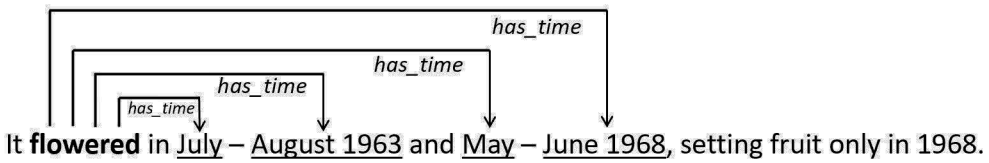

FIGURE 4
Example sentence with entity pairs that matched the rule `[S]+(o)?(To|Td|T)+`, where S corresponds to reproductive condition (in bold) which is the *source* entity, T corresponds to temporal expression (underlined) which is the *target* entity and o refers to *other* tokens. Source of example sentence: Medway (1972).

5.2 Transformer-based approaches

As mentioned in Section 2.3, the emergence of transformer-based models (Vaswani et al., 2017) has allowed researchers to cast RE as a natural language understanding problem such as question answering or machine reading comprehension. One of the key features of transformer-based models is its transfer learning capability, whereby they learn rich representations of natural language via pre-training on large-scale corpora using unsupervised learning objectives such as masked language modeling. The pre-trained models can then be fine-tuned on specific downstream tasks (e.g., question answering) using relatively smaller amounts of task-specific labeled data. Fine-tuning a pre-trained transformer often leads to better performance, faster convergence, and improved generalization compared to training models from scratch.

In this work, we cast our RE problem as: (1) a boolean question answering (boolean QA) task, and (2) a natural language inference (NLI) task, both employing transformer-based language models. Given an input sentence I and two entities E_S and E_T for which we wish to determine whether a relation holds, we systematically generate a passage-question pair or a premise-hypothesis pair, which serves as input to the boolean QA and NLI models, respectively. 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.

5.2.1 Boolean question answering

We determine the existence of a `has_time` relationship between a reproductive condition and a temporal expression, and a `has_location` relationship between a habitat and a geographic location using transformer-based models that were fine-tuned on the BoolQ dataset (Clark et al., 2019). BoolQ is a dataset that consists of *Yes*/*No* questions; it comes with 15,942 examples that are naturally occurring, i.e., generated in unprompted and unconstrained settings. Each example in BoolQ is a triple of the form *(question, passage, answer)*, where question is the *Yes*/*No* question, passage is the context for answering the question, and answer is either *Yes* or *No*. For our boolean QA task, we used two transformer models that were fine-tuned on BoolQ and are available in HuggingFace,[4] specifically, one based on RoBERTa (`roberta-base-boolq`)[5] and the other based on T5 (`t5-base-finetuned-boolq`).[6]

RoBERTa was introduced by Liu et al. (2019) and was built upon the original BERT model (Devlin et al., 2019) that was released by Google in 2018. Specifically, it introduced optimized key hyperparameters and was trained with much larger mini-batches. The `roberta-base-boolq` model that was fine-turned for boolean QA accepts a question and a passage as input, and returns probabilities for the *Yes* and *No* answers (i.e., the possible labels) as output.

Text-to-Text Transfer Transformer (T5) is an encoder-decoder model pre-trained on a mixture of unsupervised and supervised tasks; each task was converted into a sequence-to-sequence format, where each of the input and the output is a sequence of tokens (Raffel et al., 2020). T5 distinguishes between different NLP tasks by requiring an indicative prefix to be prepended to the input sequence. For example, T5, already fine-tuned for the machine translation and question answering tasks, interprets the prefixes *"translate English to German:"* and *"question: ... context: ..."* to mean that, respectively, the supplied input should be translated to German, and that the input is a question that needs to be answered based on the supplied passage (i.e., the context). The `t5-base-finetuned-boolq` model, fine-tuned on the BoolQ dataset, accepts as input a question and a passage with the *"question: ... context: ..."* prefix, and returns either *Yes* or *No* (as a sequence).

Given a data instance (I, E_S, E_T, y), the input sentence I is taken as the passage, while the question is created by populating either of the following question templates with E_S and E_T, depending on the types of the two entities:

- *Is there <habitat> in <geographic location>?*
- *Did <reproductive condition> event happen on <temporal expression>?*

Our question templates for the boolean QA models and examples with their corresponding outputs are shown in Table 5. We say that a relationship between two entities exists if the model's predicted class is *Yes*, otherwise the entities are considered to be unrelated.

4 https://huggingface.co/

5 Available at: https://huggingface.co/shahrukhx01/roberta-base-boolq.

6 Available at: https://huggingface.co/mrm8488/t5-base-finetuned-boolq.

TABLE 5 Examples of populated question templates for generating inputs (passages and corresponding questions) for the boolean QA model, together with the corresponding expected outputs.

Relation type	Question template	Examples		
		Passage/context	Question	Answer
`has_location`	Is there <**Habitat**> in <**Geographic Location**>?	*Bukit Sai and Lesong belong to the lowland dipterocarp forest types with D. aromatica being the predominant species.*	Is there **lowland dipterocarp forest** in **Bukit Sai**?	Yes
`has_time`	Did <**Reproductive Condition**> event happen on <**Temporal Expression**>?	*It flowered in July - August 1963 and May - June 1968, setting fruit only in 1968.*	Did **fruit** event happen on **August 1963**?	No

A relation holds between two given entities (in bold) only if the boolean QA model predicts *Yes* as the output label. Source of input sentences (passage): Medway (1972) and Lee (2000).

5.2.2 Natural Language Inference

We also cast our RE problem as a natural language inference (NLI) problem that we also addressed 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 T5 as our model, considering that NLI is one of the downstream NLP tasks for which T5 was already fine-tuned (Raffel et al., 2020). Specifically, we used the T5-large (`t5-large`) model[7] with 770 million parameters. Similar to our boolean QA methods, we systematically generated a premise-hypothesis pair which serves as input to the NLI model. Here, the input sentence I is taken as the premise, while the hypothesis is created by populating either of the following sentence templates with E_S and E_T:

- *The <habitat> was in <geographic location>.*
- *The <reproductive condition> event happened on <temporal expression>.*

Table 6 provides some example inputs for the T5-based NLI model, and its expected outputs. For our purposes, a relationship between two entities exists only if the model's predicted class is `entailment`; otherwise the entities are considered to be unrelated.

Due to variations in noun forms or verb tenses, the automatically generated hypothesis (for NLI) and question (for boolean QA) may not necessarily be grammatically correct; for instance, the example for the `has_time` relation in Table 6 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 for both boolean QA and NLI, as we expected the transformers-based models to be robust to such grammatical errors.

5.3 Hybrid approach: rules and transformers

In order to improve performance and reduce the 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. It is worth noting that between our two rule-based methods, the regex-based one was chosen over the one based on dependency distance, as the former does not require the optimization of any parameters, unlike the latter which relies on the careful selection of the value of n, the maximum dependency distance.

The instances that were 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_{comp}. The regular expression that was designed to extract E_{comp} is $(Et|E)\{2,\}$, where E is a named entity of a specific type, and t is any token. E_{comp} 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 section, we assess the reliability of our RE corpus and present the results of the experiments we performed to evaluate the performance of each of the RE approaches presented in Section 5.

6.1 Reliability of RE annotations

Our corpus consists of scholarly articles that contain manual annotations of entities of type geographic location, habitat, and

7 Available at: https://huggingface.co/t5-large.

TABLE 6 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.

Relation type	Hypothesis template	Examples		
		Premise	Hypothesis	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**.	entail-ment
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**.	contra-diction

A relation holds between two given entities (in bold) only if the NLI model predicts `entailment` as the output label.

TABLE 7 Frequency of instances for each relation type in our training (train), development (dev) and test sets.

Relation type	Train	Dev	Test
has_time	843	173	388
has_location	252	34	127

temporal expression. Additionally, it also contains annotations of spans of text pertaining to the reproductive condition of species. It is a subset of the corpus presented in Gabud et al. (2019). To facilitate the annotation of relations between entity pairs, we automatically identified (1) co-occurring mentions of reproductive condition and temporal expressions, and (2) co-occurring habitat and geographic location entities within every sentence in the corpus. Two annotators manually provided the label y for each data instance (I, E_S, E_T, y). Our senior annotator, a Biology degree holder, labeled all instances in the entire dataset, while the junior annotator, a Computer Science student, provided labels for only 30% of the same dataset. Our annotators manually determined whether a pair of co-occurring entities are semantically related to each other ($y = 1$). We calculated the agreement between our two annotators in terms of F1-score, and obtained an overall agreement of 95.87%. Specifically, 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 paper. The instances with disagreements were re-evaluated and re-labeled by the third annotator. We randomly split our dataset into proportions of 70, 10, and 20% to serve as our training, development, and test sets, respectively. Table 7 shows the number of instances for each relation type.

6.2 Relation extraction

In this section, we report the results of our experiments for evaluating performance of the unsupervised approaches we developed to extract `has_time` and `has_location` relations. We designed rules and templates based on our training set. The rules were then refined based on our held-out development (dev) set. Finally, the performance of the approaches was evaluated using the test set. We present the performance of each of our rule-based, transformer-based, and hybrid RE approaches in terms of precision, recall, F1-score, and Matthews correlation coefficient (MCC). Precision tells us how much we can trust the model when it predicts an example as positive, while recall measures the ability of the model to find all the positive examples in the dataset. The harmonic mean of precision and recall is the F1-score, which is useful in finding the best trade-off between the two values (Grandini et al., 2020). While the F1-score is widely used and is considered to be the standard performance metric for RE, MCC offers a more balanced assessment by considering all aspects of the confusion matrix, i.e., the number of true positives, false positives, false negatives, and true negatives (Chicco and Jurman, 2020). The range of MCC is $[-1, 1]$. It produces a high score only if good results were obtained for all of the four confusion matrix categories, proportional to the number of positive examples and the number of negative examples in the dataset, making it suitable for imbalanced datasets (Chicco and Jurman, 2020, 2023).

6.2.1 Rule-based approaches

As explained in the Methods section (Section 5.1), we designed two rule-based approaches for RE, i.e., patterns based on dependency distance and regular expression-based rules (regexes). The patterns based on dependency distances were implemented using the command line interface of the *Grew-match* package (version 1.11). Using these patterns, we extracted related biodiversity entities whose dependency distance is at most n, where $1 \leq n \leq 8$. For example, if n is set to 5, we extract relations between entities that have a dependency distance of 5 or less. It is worth noting that as the dependency distance increases, the extractions obtained by the dependency distance-based patterns become similar to those of a simple co-occurrence-based method that considers every pair of entities as related as long as they appear within the same sentence. For the purposes of our evaluation, we chose 4 as the value of n, as it yielded the highest F1-score for the `has_location` relation type on the development set. As described in Section 5.1, we also developed regular expression-based rules to extract related consecutive entities. The *re* Python module was used in implementing these regex-based rules.

Table 8 shows the precision, recall, F1-score and MCC of our two rule-based approaches, compared with a simple co-occurrence-based method as a baseline. The patterns based on dependency distance obtained F1-scores of 78.02 and 85.58% for the `has_location` and `has_time` relations, respectively, based on evaluation on our test set. These are lower than the F1-scores of a simple co-occurrence-based approach, which are 84.02% for the `has_location` relation, and 94.57% for the `has_time` relation.

However, it should be noted that the dependency distance-based method has a higher precision than the co-occurrence-based one; the recall of the latter is 100% only because it classifies all instances as positive relations and none as negative.

Among our rule-based methods, the regular expression-based rules obtained the better MCC for the `has_location` relation type, with an MCC value of 0.37. Furthermore, they obtained perfect precision (100%) for both relation types. This means that every relation predicted as a positive example by our regex-based rules is indeed a correct relation. However, such rules obtained poor recall, i.e., 33.91 and 36.96% for the `has_time` and `has_location` relations, respectively, implying that the regexes fail to extract majority of correct relations described in text; this resulted in the lowest F1-scores among the methods that we evaluated.

6.2.2 Transformer-based approaches

We performed experiments using more recent and state-of-the-art transformer-based models. As discussed in Section 5.2, we cast our relation extraction problem as boolean QA and NLI tasks, for which we created question and hypothesis templates. We employed these methods by building upon (1) RoBERTa-BoolQ, a RoBERTa model that was fine-tuned on the BoolQ dataset, (2) T5-BoolQ, a T5 model that was also fine-tuned on BoolQ, and (3) T5-Large NLI, a T5-large model that was already fine-tuned for the NLI task. All of these models are available in the Hugging Face platform. We used Google Colaboratory to apply these models to the RE problem in a zero-shot manner, i.e., without any fine-tuning on domain-specific data.

Table 9 shows the performance of the above-mentioned boolean QA and NLI models on RE. Applying the models and our templates on our test set gave us F1-scores that are higher than our dependency distance-based patterns and rule-based approaches. The F1-scores obtained by Roberta-boolQ and T5-boolQ are similar to, if not better than, the F1-scores of the co-occurrence-based method, with a difference of only <1 percentage point. For the `has_location` relation type, the T5-NLI model produced the highest F1-score, 84.75%, while for `has_time`, T5-NLI yielded the lowest F1-score, 86.98%, among the three transformer models. However, this model obtained the highest precision and MCC among the three models, i.e., F1-scores of 97.16% and 88.24%, and MCC values of 0.4 and 0.5 for the `has_time` and `has_location` relation types, respectively.

6.2.3 Hybrid approach: rules and transformers

In this section, we present the evaluation results for our hybrid approach, a two-step method comprised of our regular expression-based rules and transformer-based approaches (Rules + Transformer). The precision, recall, F1-scores, and MCC values of these methods are summarized in Table 10. This approach resulted in an increase in the MCC value of up to 0.06 for all the models. Additionally, it produced a slight improvement on the F1-scores of Roberta-boolQ and T5-boolQ. In the case of T5-NLI, the initial step of rule-based analysis improved the F1-score for `has_location` relation extraction from 84.75 to 85.39%, and from 86.98 to 89.61% for the `has_time` 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 inputs to the transformer models (Rules + Compound entities + Transformers), instead of separate single entities. Evaluating this approach on our test set, we determined that the T5-NLI model produced the highest F1-scores and MCC values among all the methods that we experimented with. The computed F1-score is 89.90% for the `has_location` relation type, and 96.75% for the `has_time` relation type, while the MCC is 0.58 and 0.64 for the `has_location` and `has_time` relation types, respectively.

7 Discussion

In this section, we analyze the performance of our unsupervised RE approaches for biodiversity entities. This is followed by a discussion of the quality of the annotations in our corpus, an overview of the implications of our results, as well as the limitations of our work.

7.1 Comparative analysis

Our regex-based approach is the most precise among all the approaches we developed in this study. However, it is also the approach that obtained the lowest recall. It misses to identify more than half of true relations in the test set. The regex-based approach is suitable for applications that cannot compromise high levels of precision, e.g., systems that support clinical decisions, or the automatic curation of databases. The main drawback of the regex-based approach is its dependency on syntactic similarity only, i.e., solely on similar word order patterns found within sentences.

The other rule-based approach we developed uses the dependency distance between entities as basis for relation extraction. This approach relies on the selection of a value for maximum dependency distance n, which determines the trade-off between higher recall and poorer precision. That is, a low value for n provides us with high precision. As we increase the value of n, recall increases as well. However, precision decreases and approaches the performance of a simple co-occurrence-based RE method. Among our approaches, we found that the value of n was most difficult to fine-tune.

The drawback of our rule-based approaches is their reliance on syntactic structure only. Both our rule-based approaches (i.e., the regexes and the patterns based on dependency distance) are quite sensitive to noisy data and do not consider any semantics. Any deviation from typical sentence structures would affect the performance of the rule-based RE method.

Among the approaches presented in this paper, our transformer-based approaches were most straightforward to implement. The formulation of question templates for the boolean QA models, and hypothesis templates for the NLI model are based on natural language and do not require know-how of the English grammar nor of programming. The transformer

TABLE 8 Precision (P), Recall (R), F1-score (F), and Matthews correlation coefficient (MCC) values obtained by rule-based approaches on the test set for `has_time` and `has_location` relation types.

Rule-based methods	has_time				has_location			
	P(%)	R(%)	F(%)	MCC	P(%)	R(%)	F(%)	MCC
Co-occurrence	89.69	100.00	94.57	0.00	72.44	100.00	84.02	0.00
Dependency distance ($n = 4$)	94.14	78.45	85.58	0.25	78.89	77.17	78.02	0.23
Regular expression rules	100.00	33.91	50.64	0.22	100.00	36.96	53.97	0.37

TABLE 9 Precision (P), Recall (R), F1-score (F), and Matthews correlation coefficient (MCC) values obtained by transformer-based approaches on the test set for `has_time` and `has_location` relation types.

Transformer models	has_time				has_location			
	P(%)	R(%)	F(%)	MCC	P(%)	R(%)	F(%)	MCC
Roberta-BoolQ	91.94	98.28	95.00	0.36	72.44	100.00	84.02	0.00
T5-BoolQ	93.04	95.98	94.48	0.39	72.58	97.83	83.33	0.02
T5-Large NLI	97.16	78.74	86.98	0.40	88.24	81.52	84.75	0.50

TABLE 10 Precision (P), Recall (R), F1-score (F), and Matthews correlation coefficient (MCC) values obtained by hybrid approaches on the test set for `has_time` and `has_location` relation types.

Hybrid approaches	Transformer models	has_time				has_location			
		P(%)	R(%)	F(%)	MCC	P(%)	R(%)	F(%)	MCC
Rules, Transformer	Roberta-BoolQ	92.00	99.14	95.44	0.41	72.44	100.00	84.02	0.00
	T5-BoolQ	93.15	97.70	95.37	0.45	72.58	97.83	83.33	0.02
	T5-Large NLI	97.31	83.05	89.61	0.45	88.37	82.61	85.39	0.52
Rules, Compound, Transformer	Roberta-BoolQ	91.80	99.71	95.59	0.43	72.44	100.00	84.02	0.00
	T5-BoolQ	91.40	97.70	94.44	0.27	72.36	96.74	82.79	-0.01
	T5-Large NLI	95.26	98.28	**96.75**	**0.64**	83.96	96.74	**89.90**	**0.58**

The best F1-scores and MCC values are shown in bold.

models paired with our question/hypothesis templates for RE provided us with F1-scores higher than our rule-based methods. For the `has_time` relations, the QA models produced high F1-scores. However, the NLI model obtained the highest precision, 97.16%, and highest MCC, 0.45. For the `has_location` relations, the NLI model is the most precise and has the highest F1-score. For both relation types, the NLI model obtained the lowest recall.

During method development and preliminary evaluation on our development set, we noticed the high precision of the regex-based approach and the transformer-based model's higher recall, compared to that of the regexes. This led us to the idea of combining the high-precision regex-based approach and the high-recall transformer-based approach. Thus, we developed a hybrid approach that is a two-step method comprised of the regex-based method followed by a transformer-based one. This hybrid approach increased the recall for `has_time` relations by up to 4.31 percentage points, and the recall for `has_location` relations by 1 percentage point. We inspected some instances which the hybrid approach failed to identify as a relation. We noticed that this approach failed to identify relations between entities that belong to an enumeration or a compound statement of entity mentions. For example, in the sentence "*Ashton (1989) record the extent of mass flowerings in peninsular Malaysia and Borneo for the period 1950–1983 based on state forest department records (Table 5),*"[8] 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-based rules to identify compound entities in sentences, as described in Section 5.3. Where they exist, these compound entities were used in populating the question and hypothesis templates, instead of individual named entities. The inclusion of regexes for compound entities in our hybrid approach significantly improved the recall of the NLI model by 14–15% points. This hybrid approach underpinned by the NLI model provided us with the highest F1-scores and MCC values among all the approaches we developed for both relation types.

8 Source: Appanah (1993).

7.2 Quality of annotations

Our results show that the annotations in our labeled corpus are reliable, given that a high level of agreement between the two annotators was obtained. Most of the disagreements were due to human errors that can be expected, e.g., missed relations when at least one of the entities belongs to an enumeration, or wrong interpretation of a complex sentence. Aside from these errors, there were very few instances of disagreements that were due to difficult cases that required deep knowledge of domain-specific terminology, e.g., when our junior annotator failed to determine that there is a semantic relationship between *"GF"* and *"February 2002"* in the sentence *"We quantified pre-dispersal seed predation for focal trees for all species of the genus Shorea section mutica having five or more qualifying individuals in February 2002, September 2002 and August 2005 (for the 2001, 2002, and 2005 GF events, respectively)."* Our more senior annotator (who has a Biology background) is more knowledgeable in reproductive conditions and was able to determine that, e.g., *"GF"* is an abbreviation for *general flowering*, and that *"dispersal"* implies the existence of fruit.

7.3 Theoretical and practical implications

From a theoretical perspective, our work in developing unsupervised RE methods for the biodiversity domain advances the NLP task of RE by proposing traditional rule-based, transformer-based, and a hybrid of rules and transformer models that extract related biodiversity entities without requiring a large amount of labeled training data. Our work has led to the development of unsupervised approaches for RE that can perform at a satisfactory level, according to the results of our evaluation. Our main methodological contribution to the NLP research community is the development of a two-step hybrid approach employing a high-precision regex-based method followed by a high-recall transformer-based NLI model, that obtains superior performance on the RE task.

From a practical perspective, our research has several applications. Firstly, our solution has the potential to support the curation of structured biodiversity data resources (e.g., databases, knowledge graphs) with finer-grained information on species' reproductive conditions and habitats. For instance, by extracting related mentions of reproductive conditions and temporal expressions from text, our method facilitates the inclusion of temporally specific reproductive conditions into a database or knowledge graph. This also applies to our other relation type of interest (i.e., the `has_location` relation type) that extracts geographically defined habitat information from text. Our approach can thus be applied to the enrichment of information in well-known biodiversity databases such as the Global Biodiversity Information Facility (GBIF) and the Atlas of Living Australia (ALA).

Furthermore, as part of a user-facing application, our RE approach can facilitate the extraction of detailed information (e.g., geographically broad-scale and long-term information on species' reproduction and habitats) from unstructured data, that can inform the decision-making of natural resource management (NRM) regulators. This can aid the implementation of data-driven approaches to land restoration and rehabilitation, which is part of the United Nations Sustainable Development Goal #15: Life on Land. Extracting `has_time` and `has_location` relations can help researchers in understanding the biology underpinning the proper timing for seed collection and seeding at suitable locations, which are important factors for effective regeneration and reforestation efforts. Once the extracted information has been represented in a structured form such as a knowledge graph, the entities and the relations between them can be easily visualized. This provides a more practical and intuitive way of viewing and analyzing reproductive condition—temporal expression and habitat—geographic location relations, compared to reading lengthy textual documents.

7.4 Limitations

To the best of our knowledge, the biodiversity-focused corpus proposed by Gabud et al. (2019) is the only publicly available corpus containing labels pertaining to reproductive condition, temporal expression, habitat and geographic location named entities. Hence, the methods we developed for extracting reproductive condition - temporal expression (`has_time`) and habitat - geographic location (`has_location`) relations were based on this corpus only. In this work, we show that despite the relatively small size of the RE-annotated corpus resulting from our work, we were able to develop unsupervised approaches for RE that perform at a satisfactory level. However, this corpus might be too small to support the training of traditional supervised machine learning-based RE models.

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 in text. We present our relation extraction corpus that contains manually labeled relations between mentions of reproductive conditions and temporal expressions (i.e., `has_time` relations), and between habitats and geographic locations (i.e., `has_location` relations). We used this corpus to experiment with our three unsupervised methods for relation extraction, namely, a rule-based approach, a transformer-based one, and a hybrid approach that combines the first two. Our rule-based approaches are based on patterns in the dependency distance between entities, and regular expressions that capture the word order of entities. They obtained the highest precision but were poor in terms of recall. For the transformer-based approach, we framed our relation extraction problem as boolean QA and NLI tasks. The methods we experimented with using this approach resulted in F1-scores higher than those obtained by the rule-based approach. We designed our two-step hybrid approach by combining our rules with our transformers models. A further improvement is the addition of using compound entities in generating the question (for boolean QA) or hypothesis (for NLI) instead of single entity mentions. Our hybrid approach composed of rules, compound entities, and a transformer-based

NLI model (built upon T5-large) produced the best performance, with a 96.75% F1-score for related reproductive conditions and temporal expressions, and a 89.90% F1-score for related habitats and geographic locations. Our work shows that even without a large training dataset, we have been able to extract `has_location`, and `has_time` relations from literature with satisfactory performance.

We consider our work to be a contribution toward large-scale studies on biodiversity that impacts sustainable life on land. This could eventually facilitate the development of a biodiversity database enriched with information on the habitats and reproductive conditions of species, extracted from literature. As part of our future work, we plan to analyze the extent to which our pre-processing steps (e.g., tokenization, dependency parsing) affects RE performance. Furthermore, we intend to implement an information extraction pipeline comprised of an NER tool and our hybrid RE approach that automatically identifies related biodiversity entities from text. We will explore how to incorporate such reproductive condition and habitat information from text into species occurrence data stored in databases such as GBIF and ALA.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

RG: Formal analysis, Investigation, Methodology, Software, Writing - original draft, Writing - review & editing. PL: Writing - review & editing, Data curation. VM: Formal analysis, Writing - review & editing. EM: Formal analysis, Writing - review & editing. NP: Writing - review & editing, Data curation. MC: Formal analysis, Writing - review & editing. RB-N: Conceptualization, Writing - original draft, Writing - review & editing, Formal analysis.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

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

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abdelmageed, N., Löffler, F., Feddoul, L., Algergawy, A., Samuel, S., Gaikwad, J., et al. (2022). BiodivNERE: gold standard corpora for named entity recognition and relation extraction in the biodiversity domain. *Biodiv. Data J.* 10:e89481. doi: 10.3897/BDJ.10.e89481

Ahmed, S., Stoeckel, M., Driller, C., Pachzelt, A., and Mehler, A. (2019). "BIOfid dataset: publishing a German Gold Standard for named entity recognition in historical biodiversity literature," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (Hong Kong: Association for Computational Linguistics), 871–880.

ALA (2023). *Atlas of Living Australia*. Available online at: https://www.ala.org.au (accessed December 05, 2023).

Amasino, R. (2010). Seasonal and developmental timing of flowering. *Plant J.* 61, 1001–1013. doi: 10.1111/j.1365-313X.2010.04148.x

Appanah, S. (1993). Mass flowering of dipterocarp forests in the aseasonal tropics. *J. Biosci.* 18:463.

Ashton, P. S. (1989). "Dipterocarp reproductive biology," in *Tropical Rain Forest Ecosystems*, eds. H. Lieth and M. J. A. Werger (Amsterdam: Elsevier Science Publishers), 226.

Barbedo, C. J., Centeno, D. d. C., and Ribeiro, R. d. C. L. F. (2013). Do recalcitrant seeds really exist? *Hoehnea* 40, 583–593. doi: 10.1590/S2236-89062013000400001

Batista-Navarro, R., Hammock, J., Ulate, W., and Ananiadou, S. (2016). "A text mining framework for accelerating the semantic curation of literature," in *Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science*, eds. N. Fuhr, L. Kovács, T. Risse, and W. Nejdl (Cham: Springer International Publishing), 459–462.

Batista-Navarro, R., Zerva, C., Nguyen, N. T. H., and Ananiadou, S. (2017). "A text mining-based framework for constructing an RDF-compliant biodiversity knowledge repository," in *Information Management and Big Data, Communications in Computer and Information Science*, eds. J. A. Lossio-Ventura and H. Alatrista-Salas (Cham: Springer International Publishing), 30–42.

BHL (2023). *Biodiversity Heritage Library*. Available online at: https://www.biodiversitylibrary.org/ (accessed November 29, 2023).

Bonfante, G., Guillaume, B., and Perrier, G. (2018). *Application of Graph Rewriting to Natural Language Processing*. Hoboken, NJ: John Wiley & Sons.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E., and Mitchell, T. (2010). Toward an architecture for never-ending language learning. *Proc. AAAI Conf. Artif. Intell.* 24, 1306–1313. doi: 10.1609/aaai.v24i1.7519

Chaix, E., Deléger, L., Bossy, R., and Nédellec, C. (2019). Text mining tools for extracting information about microbial biodiversity in food. *Food Microbiol.* 81, 63–75. doi: 10.1016/j.fm.2018.04.011

Cheng, J., Jiang, H., Yang, D., and Xiao, Y. (2021). A question-answering based framework for relation extraction validation. *arXiv [Preprint]. arXiv*:2104.02934.

Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* 21:6. doi: 10.1186/s12864-019-6413-7

Chicco, D., and Jurman, G. (2023). The Matthews correlation coefficient?(MCC) should replace the ROC-AUC as the standard metric for assessing binary classification. *BioData Min*. 16:4. doi: 10.1186/s13040-023-00322-4

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. (2019). BoolQ: exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv*:1905.10044.

Cochard, R. (2001). *Consequences of Deforestation and Climate Change on Biodiversity*. IGI Global. Available online at: https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60960-619-0.ch002

Cornford, R., Deinet, S., De Palma, A., Hill, S. L. L., McRae, L., Pettit, B., et al. (2021). Fast, scalable, and automated identification of articles for biodiversity and macroecological datasets. *Glob. Ecol. Biogeogr*. 30, 339–347. doi: 10.1111/geb.13219

Culotta, A., and Sorensen, J. (2004). "Dependency tree Kernels for relation extraction," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)* (Barcelona), 423–429.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv*:1810.04805

dos Santos, C., Xiang, B., and Zhou, B. (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)* (Beijing: Association for Computational Linguistics), 626–634.

Du, X., and Cardie, C. (2021). Event extraction by answering (almost) natural questions. *arXiv preprint arXiv*:2004.13625.

Ehrlén, J. (2015). Selection on flowering time in a life-cycle context. *Oikos* 124, 92–101. doi: 10.1111/oik.01473

FAO (2019). "The state of the world's biodiversity for food and agriculture," in *FAO Commission on Genetic Resources for Food and Agriculture Assessments*, eds. J. Bélanger and D. Pilling (Rome: FAO), 572.

FAO and UNEP (2020). *The State of the World's Forests 2020: Forests, Biodiversity and People*. Rome: FAO and UNEP.

Farrell, M. J., Brierley, L., Willoughby, A., Yates, A., and Mideo, N. (2022). Past and future uses of text mining in ecology and evolution. *Proc. Royal Soc. B* 289:20212721. doi: 10.1098/rspb.2021.2721

Fundel, K., Küffner, R., and Zimmer, R. (2007). RelEx—Relation extraction using dependency parse trees. *Bioinformatics* 23, 365–371. doi: 10.1093/bioinformatics/btl616

Gabud, R., Yap, S., Batista-Navarro, R., and Ananiadou, S. (2017). Developing a knowledge base on the habitats and reproductive conditions of Dipterocarps through information extraction. *Biodiv. Inform. Sci. Stand*. 1:e20066. doi: 10.3897/tdwgproceedings.1.20066

Gabud, R. S., Batista-Navarro, R. T., Mariano, V. Y., Mendoza, E. R., and Yap, S. L. (2019). Literature mining on dipterocarps: towards better informed natural regeneration and reforestation in Luzon, Philippines. *Tech. J. Philippine Ecosyst. Natl. Resour*. 29, 39-53.

GBIF (2023). *Global Biodiversity Information Facility*. Available online at: https://www.gbif.org/ (accessed January 08, 2024).

Gebeyehu, M. N., and Hirpo, F. H. (2019). Review on effect of climate change on forest ecosystem. *Int. J. Environ. Sci. Nat. Resour*. 17, 1–4. doi: 10.19080/IJESNR.2019.17.555968

Gerner, M., Nenadic, G., and Bergman, C. M. (2010). LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinformat*. 11:85. doi: 10.1186/1471-2105-11-85

Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv:2008.05756*. doi: 10.48550/arXiv.2008.05756

Groom, Q., Güntsch, A., Huybrechts, P., Kearney, N., Leachman, S., Nicolson, N., et al. (2020). People are essential to linking biodiversity data. *Database* 2020:baaa072. doi: 10.1093/database/baaa072

Guillaume, B. (2021). "Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, eds. D. Gkatzia and D. Seddah (Stroudsburg, PA: Association for Computational Linguistics), 168–175.

Kambhatla, N. (2004). "Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction," in *Proceedings of the ACL Interactive Poster and Demonstration Sessions* (Stroudsburg, PA), 178–181.

Kato, M., Kosaka, Y., Kawakita, A., Okuyama, Y., Kobayashi, C., Phimminith, T., et al. (2008). Plant–pollinator interactions in tropical monsoon forests in Southeast Asia. *Am. J. Bot*. 95:1377. doi: 10.3732/ajb.0800114

Koning, D., Sarkar, I. N., and Moritz, T. (2005). TaxonGrab: extracting taxonomic names from text. *Biodiv. Informat*. 2:17. doi: 10.17161/bi.v2i0.17

Kopperud, B. T., Lidgard, S., and Liow, L. H. (2022). Enhancing georeferenced biodiversity inventories: automated information extraction from literature records reveal the gaps. *PeerJ* 10:e13921. doi: 10.7717/peerj.13921

Le Guillarme, N., and Thuiller, W. (2022). TaxoNERD: deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods Ecol. Evol*. 13, 625–641. doi: 10.1111/2041-210X.13778

Leary, P. R., Remsen, D. P., Norton, C. N., Patterson, D. J., and Sarkar, I. N. (2007). uBioRSS: tracking taxonomic literature using RSS. *Bioinformatics* 23, 1434–1436. doi: 10.1093/bioinformatics/btm109

Lee, J. H., Kwan, Y. A., Park, J., Park, S. M., and Oh, J. S. (2019). Analysis of utilization of biological resources using text mining based on freshwater biodiversity information platform. *Biodiv. Inform. Sci. Stand*. 3:e37664. doi: 10.3897/biss.3.37664

Lee, K., Famiglietti, M. L., McMahon, A., Wei, C.-H., MacArthur, J. A. L., Poux, S., et al. (2018). Scaling up data curation using deep learning: an application to literature triage in genomic variation resources. *PLoS Comput. Biol*. 14:e1006390. doi: 10.1371/journal.pcbi.1006390

Lee, S. L. (2000). Mating system parameters of *Dryobalanops aromatica* Gaertn. f.(Dipterocarpaceae) in three different forest types and a seed orchard. *Heredity* 85:339. doi: 10.1046/j.1365-2540.2000.00761.x

Lee, Y., Son, J., and Song, M. (2022). BertSRC: transformer-based semantic relation classification. *BMC Med. Informat. Decision Mak*. 22:234. doi: 10.1186/s12911-022-01977-5

Lelli, C., Nascimbene, J., and Chiarucci, A. (2018). Are available vegetation data suitable for assessing plant diversity? A study case in the Foreste Casentinesi National Park (Italy). *Rendiconti Lincei. Scienze Fisiche e Naturali* 29, 355–362. doi: 10.1007/s12210-018-0681-z

Levy, O., Seo, M., Choi, E., and Zettlemoyer, L. (2017). Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv*:1706.04115.

Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., and Li, J. (2022). A unified MRC framework for named entity recognition. *arXiv preprint arXiv*:1910.11476.

Liu, C., Sun, W., Chao, W., and Che, W. (2013). "Convolution neural network for relation extraction," in *Advanced Data Mining and Applications, Lecture Notes in Computer Science*, eds. H. Motoda, Z. Wu, L. Cao, O. Zaiane, M. Yao, and W. Wang (Berlin; Heidelberg: Springer), 231–242.

Liu, J., Chen, Y., Liu, K., Bi, W., and Liu, X. (2020). "Event extraction as machine reading comprehension," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Stroudsburg, PA: Association for Computational Linguistics), 1641–165.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: a robustly optimized BERT pretraining approach. *arXiv preprint arXiv*:1907.11692.

Lobell, D., Schlenker, W., and Costa-Roberts, J. (2011). Climate trends and global crop production since 1980. *Science* 333, 616–620. doi: 10.1126/science.1204531

Lücking, A., Driller, C., Stoeckel, M., Abrami, G., Pachzelt, A., and Mehler, A. (2022). Multiple annotation for biodiversity: developing an annotation framework among biology, linguistics and text technology. *Lang. Resour. Eval*. 56, 807–855. doi: 10.1007/s10579-021-09553-5

Luna-Nieves, A. L., Meave, J. A., Morellato, L. P. C., and Ibarra-Manríquez, G. (2017). Reproductive phenology of useful Seasonally Dry Tropical Forest trees: guiding patterns for seed collection and plant propagation in nurseries. *For. Ecol. Manag*. 393, 52–62. doi: 10.1016/j.foreco.2017.03.014

Medway, L. (1972). Phenology of a tropical rain forest in Malaya. *Biol. J. Linnean Soc*. 4:128.

Miller, S., Fox, H., Ramshaw, L., and Weischedel, R. (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* (Stroudsburg, PA).

Mora-Cross, M., Morales-Carmiol, A., Chen-Huang, T., and Barquero-Prez, M. (2022). Essential Biodiversity Variables: extracting plant phenological data from specimen labels using machine learning. *Res. Ideas Outcomes* 8:e86012. doi: 10.3897/rio.8.e86012

Morueta-Holme, N., and Svenning, J.-C. (2018). Geography of plants in the new world: Humboldt's relevance in the age of big data 1. *Ann. Missouri Botan. Gard*. 103, 315–329. doi: 10.3417/2018110

Mozzherin, D. (2022). *Global Names Finder*. Available online at: https://finder.globalnames.org/ (accessed December 10, 2023)

Nguyen, N. T., Gabud, R. S., and Ananiadou, S. (2019). COPIOUS: a gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiv. Data J*. 7:e29626. doi: 10.3897/BDJ.7.e29626

Nguyen, N. T., Miwa, M., Tsuruoka, Y., Chikayama, T., and Tojo, S. (2015). Wide-coverage relation extraction from MEDLINE using deep syntax. *BMC Bioinformat*. 16:107. doi: 10.1186/s12859-015-0538-8

Oshima, C., Tokumoto, Y., and Nakagawa, M. (2015). Biotic and abiotic drivers of dipterocarp seedling survival following mast fruiting in Malaysian Borneo. *J. Trop. Ecol*. 31, 129–137. doi: 10.1017/S026646741400073X

Pafilis, E., Frankild, S. P., Fanini, L., Faulwetter, S., Pavloudi, C., Vasileiadou, A., et al. (2013). The species and organisms resources for fast and accurate identification of taxonomic names in text. *PLoS ONE* 8:e65390. doi: 10.1371/journal.pone.0065390

Page, R. (2019). Text-mining BHL: towards new interfaces to the biodiversity literature. *Biodiv. Inform. Sci. Stand.* 3:e35013. doi: 10.3897/biss.3.35013

Paragkamian, S., Sarafidou, G., Mavraki, D., Pavloudi, C., Beja, J., Eliezer, M., et al. (2022). Automating the curation process of historical literature on marine biodiversity using text mining: the DECO workflow. *Front. Mar. Sci.* 9:940844. doi: 10.3389/fmars.2022.940844

Parr, C. S., and Thessen, A. E. (2018). "Biodiversity informatics," in *Ecological Informatics: Data Management and Knowledge Discovery*, eds. F. Recknagel and W. K. Michener (Cham: Springer International Publishing), 375–399.

Poulin, M., Andersen, R., and Rochefort, L. (2013). A new approach for tracking vegetation change after restoration: a case study with peatlands. *Restorat. Ecol.* 21, 363–371. doi: 10.1111/j.1526-100X.2012.00889.x

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 167.

Ravikumar, K. E., Wagholikar, K. B., Li, D., Kocher, J.-P., and Liu, H. (2015). Text mining facilitates database curation-extraction of mutation-disease associations from bio-medical literature. *BMC Bioinform.* 16:185. doi: 10.1186/s12859-015-0609-x

Ray, D. K., West, P. C., Clark, M., Gerber, J. S., Prishchepov, A. V., and Chatterjee, S. (2019). Climate change has likely already affected global food production. *PLoS ONE* 14:e0217148. doi: 10.1371/journal.pone.0217148

Sautter, G., Böhm, K., and Agosti, D. (2006). A combining approach to find all taxon names (FAT). *Biodiv. Informat.* 3:34. doi: 10.17161/bi.v3i0.34

Song, X., Yu, H., Li, S., Hu, X., and Wang, H. (2022). "Overview of relation extraction," in *2021 International Conference on Advanced Computing and Endogenous Security* (New York, NY), 1–11.

Staples, T. L., Mayfield, M. M., England, J. R., and Dwyer, J. M. (2020). Comparing the recovery of richness, structure, and biomass in naturally regrowing and planted reforestation. *Restorat. Ecol.* 28, 347–357. doi: 10.1111/rec.13077

TDWG (2023). *List of Darwin Core Terms*. Available online at: https://dwc.tdwg.org/list/ (accessed November 20, 2023).

Thessen, A., Mozzherin, D., Shorthouse, D., and Patterson, D. (2022). Improving the discoverability of biodiversity data using the Global Names Finder. *Biodiv. Inform. Sci. Stand.* 6:e90026. doi: 10.3897/biss.6.90026

Thessen, A., Preciado, J., Jain, P., Martin, J., Palmer, M., and Bhat, R. (2018). Automated trait extraction using ClearEarth, a natural language processing system for text mining in natural sciences. *Biodiv. Inform. Sci. Stand.* 2:e26080. doi: 10.3897/biss.2.26080

Thessen, A. E., Cui, H., and Mozzherin, D. (2012). Applications of natural language processing in biodiversity science. *Adv. Bioinformat.* 2012:e391574. doi: 10.1155/2012/391574

UNDP (2023). *Goal 15: Life on land | Sustainable Development Goals | United Nations Development Programme*. Available online at: https://www.undp.org/sustainable-development-goals/life-on-land (accessed December 10, 2023).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems, Vol. 30*, eds. Von Luxburg et al. (New York, NY: Curran Associates, Inc).

Vu, T., Adel, H., Gupta, P., and Schütze, H. (2016). Combining recurrent and convolutional neural networks for relation classification. *arXiv preprint arXiv*:1605.07333.

Wang, L., Cao, Z., de Melo, G., and Liu, Z. (2016). "Relation classification via multi-level attention CNNs," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Stroudsburg, PA), 1298-1307.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., et al. (2012). Darwin core: an evolving community-developed biodiversity data standard. *PLoS ONE* 7:e29715. doi: 10.1371/journal.pone.0029715

Xiao, M., and Liu, C. (2016). "emantic relation classification via hierarchical recurrent neural network with attention," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (Osaka: The COLING 2016 Organizing Committee), 1254–1263.

Yan, Y., Sun, H., and Liu, J. (2021). "A review and outlook for relation extraction," in *Proceedings of the 5th International Conference on Computer Science and Application Engineering, CSAE '21* (New York, NY: Association for Computing Machinery), 1–5.

Zelenko, D., Aone, C., and Richardella, A. (2002). "Kernel methods for relation extraction," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing—Volume 10, EMNLP '02* (New York, NY: Association for Computational Linguistics), 71–78.

Zhang, C., Zhang, X., Jiang, W., Shen, Q., and Zhang, S. (2009). "Rule-based extraction of spatial relations in natural language text," in *2009 International Conference on Computational Intelligence and Software Engineering* (New York, NY), 1–4.

Zhang, D., and Wang, D. (2015). Relation classification via recurrent neural network. *arXiv preprint arXiv:* 1508.01006.

Zhao, Y., Yuan, X., Yuan, Y., Deng, S., and Quan, J. (2023). Relation extraction: advancements through deep learning and entity-related features. *Soc. Netw. Anal. Min.* 13:92. doi: 10.1007/s13278-023-01095-8

Zheng, S., Han, X., Lin, Y., Yu, P., Chen, L., Huang, L., et al. (2019). DIAG-NRE: a neural pattern diagnosis framework for distantly supervised neural relation extraction. *arXiv preprint arXiv*:1811.02166.

Zhu, H., Lin, Y., Liu, Z., Fu, J., Chua, T. S., and Sun, M. (2019). "Graph neural networks with generated parameters for relation extraction," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics), 1331–1339.