BiodivNER: Dataset information

December 14, 2024

This supplementary document provides the definitions of the 6 named entity category contained in the dataset BiodivNER, which have been obtained from the explanations of each category provided by the dataset authors. In addition to these definitions, which have been used to design the prompts, 1 gives an overview of the frequency of NE instances per NE class and per data split. Tables 2 and 3 give basic statistical information about the average, median, maximum, and minimum length of sentences and tokens in each data split and in the whole dataset, alongside the total number of sentences and tokens per data split and in the complete dataset. Finally, Tables 4 and 5 list the ten most frequent and least frequent real-world instances of the 6 NE categories respectively. This data is presented per data split.

Definitions of NE classes

- **ORGANISM**: all individual life forms such as microorganisms, plants, animals, mammals, insects, fungi, bacteria etc.;
- **PHENOMENA**: occurring natural, biological, physical or chemical processes such as decomposition, colonisation, deforestation, as well as events, such as climate change;
- **MATTER**: chemical and biological compounds, and natural elements, such as carbon, sediment, sand;
- **ENVIRONMENT**: natural and man-made environments organisms live in, such as groundwater, garden, aquarium, mountain;
- **QUALITY**: data parameters measured or observed, phenotypes and traits, such as volume, age, structure, morphology;
- **LOCATION**: geographic location such as China, the United States etc.

Named entity class	Train	Development	Test	Total per category
ORGANISM	1977	164	281	2422
PHENOMENA	517	59	63	639
MATTER	471	41	292	804
ENVIRONMENT	1167	157	154	1478
QUALITY	2406	292	455	3153
LOCATION	170	20	32	222
Total per data split	6708	733	1277	8718

Table 1: Named entity instances per category and per data split in BiodivNER

Data split	Count	Average len.	Median len.	Maximum len.	Minimum len.
train	1828	39.03	28	1053	3
development	229	39.35	28	438	3
test	229	47.74	28	2047	5
total	2286	39.94	28	2047	3

Table 2: BiodivNER sentence features

Complete dataset information Sentence lengths: mean: 39.94, median: 28.0, max: 2047, min: 3. Number of total sentences is 2286 Token lengths: mean: 5.25, median: 4, max: 99, min: 1. Number of total tokens is 91293 The data split has a total of 8718 entities.

Data split	Count	Average len.	Median len.	Maximum len.	Minimum len.
train	71348	5.24	4	99	1
development	9012	5.12	4	65	1
test	10933	5.45	5	60	1
total	91293	5.25	4	99	1

Table 3: BiodivNER token features

Most frequent NE instances per category in BiodivNER			
NE class	**Train**	**Development**	**Test**
ORGANISM	species, tree, morphospecies, Arthropod species, Diptera, trees, plant, Species, Tree, bacteria	species, trees, tree, plant, plant species, earthworms, human, Species, flora, tree species	species, tree, plant, caterpillars, Arthropod species, morphospecies, trees, Plant, seedling, Species
ENVIRONMENT	soil, field, ecosystem, forest, Soil horizon, community, habitat, woody, land vegetation	field, ecosystem, soil, forest, community, land, soils, habitat, Vegetation, forests	field, soil, ecosystem, forest, land, habitat, community, soils, communities, Nature Reserve
LOCATION	China, country, Jiangxi Province, location, locations, Tübingen, Zhejiang Province, Mediterranean Sea, Germany, countries	China, Jiangxi Province, countries, location, Greenland, Tübingen, Netstal, Switzerland, country, Sweden	China, Lueneburg Scharnhorststr, country, New Zealand, Country, Zhejiang Province, Australia, USA, Tübingen Rümelinstr, Freiburg im Breisgau
MATTER	carbon, metal, water, woody debris, sand, wood, nitrogen, woody debris items, sediment, clay	sediment, water, oil palm, metal, carbon, maize, soya, bean, Coarse material, coarse sand	Chemical elements, Nitrogen, Carbon, carbon, woody debris, sediment, N, C, Ca, Fe
PHENOMENA	Precipitation, rainfall, climate change, sand loss, precipitation, pollination, conservation, CO2 emissions, planting, ice storm	Precipitation, conservation, rainfall, precipitation, extinction, planting, consumption, pollination, Biological activity, growth	Precipitation, rainfall, climate change, precipitation, fragmentation, growth, conservation, weather, rain events, mutualistic ant-hemipteran interactions
QUALITY	abundance, height, diameter, rainfall amount, species richness, Abundance, area, trait, soil properties, peak rainfall intensity	species description, abundance, rainfall amount, species richness, average rainfall intensity, height, biomass, density, Abundance, peak rainfall intensity	Phylogenetic biodiversity, abundance, rainfall amount, average rainfall intensity, Leaf stomata size, Abundance, species richness, area, peak rainfall intensity, biomass

Table 4: Ten most frequent NE instances in each category for every data split of BiodivNER

Least frequent NE instances per category in BiodivNER			
NE class	**Train**	**Development**	**Test**
ORGANISM	wildlife, raccoon, deer, hawk, chicken, rooster, dog, cat, squirrel, invertebrates	fishes, conifer plantations, Cunninghamia lanceolata, Pinus massoniana, Plant, caterpillars, Organism, Animal, open-habitat species, Dead wood	Cunninghamia lanceolata, Pinus massoniana, voucher specimens, fish, Herbivore, Herbivores, microbes, Animal, drought-sensitive species, rogenhoferi
ENVIRONMENT	above ground, Grasses, herbivore communities, green spaces, herb layer, lands, Tropical Agroforestry Landscapes, above-ground, harbor, sea ice cover	host, forest habitat, forest ecosystem, sites, subplots, broad-leaved forests, tree, broadleaf canopies, bare ground, dense conifer forests	nature, Island, plant communities, Park, vegetation layer, above ground, soil environment, bathyal habitat, mineral soil layers, organic soil layers
LOCATION	Laysan Island, Freiburg Germany, New York City, United States, Puerto Rico, U.S. Virgin Islands, North America, South-eastern China, Xingangshan, Tübingen Germany	China, Jiangxi Province, countries, location, Greenland, Tübingen, Netstal, Switzerland, country, Sweden	Alboran Sea, Aegean Sea, Lüneburg Germany, Leipzig, Deutscher Platz 5e Leipzig Germany, Freiburg, Tennenbacher Str, carbon, southeast China, Leipzig Deutscher Platz 5e Leipzig Germany
MATTER	terpenes, chemical elements, Soil lipid fractions, primary metabolites, nutrients, antioxidant, cellulosic, biofuels, flux, Nitrate	fine sand, below-ground carbon, biofuel, sunlight, medium silt, soil enzymes, nitrogen, phosphorus, potassium, extrafloral nectaries	Rock fragments, network, soil samples, clay, total clay, fine sand, oil palm, raindrops, CO2, stratum
PHENOMENA	Weather, death, impairment, forest, short dry season, fragmentation, tree planting, human, pressures, Forest restoration, ocean warming	intraspecific variation, trait evolution, Water consumption, summer, autumn, tree planting, climate change, Fertilization, Grazing, neighbourhood interactions	environmental change, treatment, throughfall, spring, summer, Climate Change, Rainfall, drought, mutualism, predation
QUALITY	fungal biomass, positioning, woody increment, species-level trait, trait covariations, biodiversity indices, shannon index, Altitude, forests cover, growth	mass, young, shrub species names, sulphur contents, needle area, Total needle surface area, Soil description, species, landscape heterogeneity, crown asymmetry	Species name, Capacity, landscape scales, successional age, diversity gradient, shrub position, thick layers, toxicity, pH, Microbial biomass

Table 5: Ten least frequent NE instances in each category for every data split of BiodivNER