SciVar Annotation Guidelines

updated V8 May 16, 2024

Introduction

In this task you will mark the occurrences of certain expressions as they appear in written texts. The expressions to be marked are names of various kinds of scientific entities and descriptions, geographical information (i.e., locations) and temporal expressions.

The annotation will be performed with Adobe Acrobat. Different colors, specified below, are used to highlight (i.e., annotate) each element type.

For some annotators, there will be an additional layer of annotation to capture scenario context or model card. These annotation tasks are described at the end of this document.

These guidelines were partially adapted from Ferro, Lisa (2006) "Super-Sized Named Entity Tag Set: Annotation Guidelines" version 1.0 and XXX and XXX (2023) "Guidelines for XXXXX PDF Annotation (version 1.0)."

Typographical Conventions

To avoid clutter, we have minimized the use of tags in our examples. In this document, we use plain square brackets to indicate the bounds of the tag: [Washington]. The context of the example will tell you which tag should be applied.

General Conventions

Annotators should prioritize precision over recall in their first round of annotation on each document. Annotations that are missed (i.e., elements that should be annotated but haven't been) can be corrected upon document review.

The task is to identify all instances of such expressions in each text, including in the title, abstract, and figure and table captions. Figures, tables, keywords, floating equations, acknowledgment sections, and references, however, are not annotated.

Nested Annotations

This task does not involve the annotation of "nested expressions." For example, there are some element types that overlap, such as in the instance of a "variable with value" entity type. These expressions contain a stand-alone variable and a stand-alone value. In these cases of overlapped annotation, the longest extent, the "variable with value" should be annotated. By chance, if there is a nested annotation within a longer annotation of another type, the nested annotation will be ignored as long as the nested element is irrelevant to the current annotation of focus. For example, the location expression "United States" is contained within the span of the Variable with Value annotation.

- the [estimated reproduction rate in the [United States] as a whole stood at around 2.5.]

Extents

The evaluation will use generous alignment standards that do not require exactly matching extents but it is preferable, though not mandatory, to exclude white space and punctuation when annotating.

Annotation Types

Annotators are asked to use assigned colors to highlight five different types of annotations: Variables with Values, Variable Descriptions, Locations and Temporal Contexts, and Scenario Card annotations. The guidelines below provide further instructions for each annotation type.

Variables with Values

This entity type captures variables with their numeric values. Values expressed as ranges should be annotated. To qualify as a Variable with Value, the expression must contain a number assigned to a simple expression.

- [growth rate of 0.01]
- [r0 = 1.2]
- [Reproduction numbers of COVID-19 vary in different studies and regions of the world (in addition over time) but have generally been found to be between 1.5 and 6.]
- the [estimated reproduction rate in the [United States] as a whole stood at around 2.5.]
- The number of [unquarantined infected cases was 1200].
- [Beta represents a value 1-3]

Do annotate a value expression as a Variable with Value even when the variable is implied, and not explicit. Annotate and then add a pop-up note to indicate the implied variable. For example, “334” would be annotated as a Variable with Value and then noted as “Implied variable: unquarantined infected cases”

- The number of [unquarantined infected cases was 1200]. [The number had been 334].

*The number refers to the unquarantined infected cases. As such, this is a way to handle coreference with implied variables.

Do **not** include confidence intervals in the extent of the variable with value expression:

- [the mean control reproductive number is 6.47] (95% CI 5.71-7.3)

Do **not** tag equations as variables with values.

- $I(t)=Io_e^{xt}$

These entity types are marked in **blue**.

Variable Descriptions

In the case of complex phrases, highlight the whole span of text that contains the complete information.

This entity type is highlighted in **yellow**.

- [lambda represents the infection coefficient]
- [infected (asymptomatic or pauci-symptomatic infected, undetected)]
- [*B* is the number of such variables]
- [*y* is the recovery rate constant]
- [*S* is the total number of infected]
- [normalized infection *i*]
- [I infections]
- [time T_d]=ln 2/α
- [H_p is the Hubble constant]
- The reproductive number
    - *Do not annotate this expression as it is just a description without a variable.
- Susceptible, Exposed, Infectious versus [S Susceptible], [E Exposed], [I Infectious]

Do **not** tag vacuous expressions as variable descriptions, such as:

- parameter v

Location Context

These are names and abbreviations of geographical locations and geo-political locations.

This entity type is highlighted in **green**.

- [Italy]
- The city of [Wuhan]
- The [Middle East]
- The [United States]
- the [United Kingdom]
- [California (CA)]
- [New York state]
- [Hubei province]
- three [US] States: [California], [New York], and [Texas]
- [US states]
- [US counties]
- [counties of Marion and Pickaway]
- [Lee, Collier, and Sarasota counties]
- [All counties excluding Marion and Pickaway counties]

Expressions in which place **names are listed in succession**, with or without a separating comma, are to be tagged as one LOCATION:

- [Boston, Massachusetts]
- [Washington, DC]
- [Wuhan, Hubei province, China]

Adjectival forms of Locations should **not** be tagged as Locations.

- American infection rates
- Chinese variant
- [US]-Russian

Facility names should **not** be tagged as Locations.

- Visited the [Wuhan Huanan Seafood Market]
- was applied for the first time at [John Hopkins University]

We will **not** annotate negated expressions such as “excluding Beijing” and “excluding Wuhan.”

Organization-Facility Metonyms, where the organization’s alias is based on a unique structure or facility in which the organization holds office should **not** be tagged as a Location unless it is made clear in the text that the entity is being referred to as a location, and not as an organization.

- The White House announced death rates
- The University of California released its newest space weather prediction model
- The outbreak was located at [University of California]

Organization-Location Metonyms, which reference political, military, athletic, and other organizations by the name of a city, country, or other associated location should **not** be marked as Locations:

- Imperial College London released a report
    - *Annotate nothing, not even London

Temporal context

This entity type can reference calendar dates, times of day, or durations (such as periods of hours, days, or even periods of centuries). The temporal context is extracted to tell us things about when the model parameters are valid (e.g., beta was 1.4 when the such-and-such variant was prevalent in March to May 2021).

Like Locations, this entity type is highlighted in **green**.

- [April 2020]
- [end of March]
- [10 o'clock EST]
- [11th of March]
- [mid-April]
- [early April]
- Published in [1927]
- Detected on [March 27, 2020]
- [Chinese Lunar New Year]
- [From late November to December 2019]
- [since January 1]
- [from January 1 to January 3, 1990]
- [27 March to 5 May]
- In the [1980s]
- [before January 30]
- the afternoon of [January 30] (afternoon is too relative to be considered a time)

Relative temporal expressions, such as "today" and "decline in week 22" are **not** to be annotated. Holidays, like "Spring Festival", are not to be annotated.

Other

The annotator may use **pink** highlighting to tag items that require discussion or the annotator believes that they could be relevant to another task.

Additional Annotation Tasks

Model card elements

Annotate model card elements only in those articles that describe a model in detail versus articles that simply cite a model as a tool in the research. For example, articles that describe the development of a model would be annotated with model card elements. However, articles that report results using a model or comparing models without adding sufficient detail to describe the models would **not** be annotated with model card elements.

The purpose of this task is to annotate specific attributes within the model card to facilitate the evaluation of the model. The model card contains the following attributes:

> *DESCRIPTION: Provide a concise one-sentence description of the model, typically found in the abstract.*
>
> *AUTHOR_INST: List the name of the institution that published the model.*
>
> *AUTHOR_AUTHOR: Include the name(s) of the author(s) associated with the model.*

AUTHOR_EMAIL: Specify the email address of the model's corresponding author.

DATE: Indicate the publication date of the model, which may correspond to the date of the associated paper.

SCHEMA: Offer a brief one-sentence description of the model's input and output schema.

PROVENANCE: Give a one-sentence summary of the model's training methodology.

DATASET: Describe in one sentence the dataset used for training the model, and provide a link if available.

COMPLEXITY: State the complexity level of the model.

USAGE: Summarize in one sentence the intended use case for the model.

LICENSE: Mention the licensing terms under which the model is distributed.

Model card attributes are highlighted in **red**.

- Carefully read through the document to identify any of the above attributes.
- When you find an attribute, right-click on the corresponding red-highlighted section and select "Open Pop-Up Note."
- In the pop-up note, enter the name of the identified attribute.
- It is possible that not all attributes will be present in the document. Annotate as many attributes as you can, and apply color-coding consistently to enhance visibility.

Scenario context

Scenario Context annotation happens after variable extraction tagging. To perform this annotation, follow these steps:

- Identify an annotation that has either or both *location* and *temporal* scenario context.
- Upon identification, right click on the highlighted annotation and choose *Open Pop-Up Note*.
- Use the text box to write **all** relevant contexts using the format below.
- Note that both types of context, i.e., location and temporal context, are optional. Write only the types that apply.
- For multiple contexts of the same type, use a comma to separate values.
- For adding two different context types, use a semicolon to separate. The order doesn't matter. The format is as follows:

When location and temporal context is expressed: location: loc1, loc2, loc3, ...;time: t1, t2, t3, ...

Example: location: new york; time: from 30 march to 5 may

When just temporal context is expressed: time: t1, t2, t3, ...