

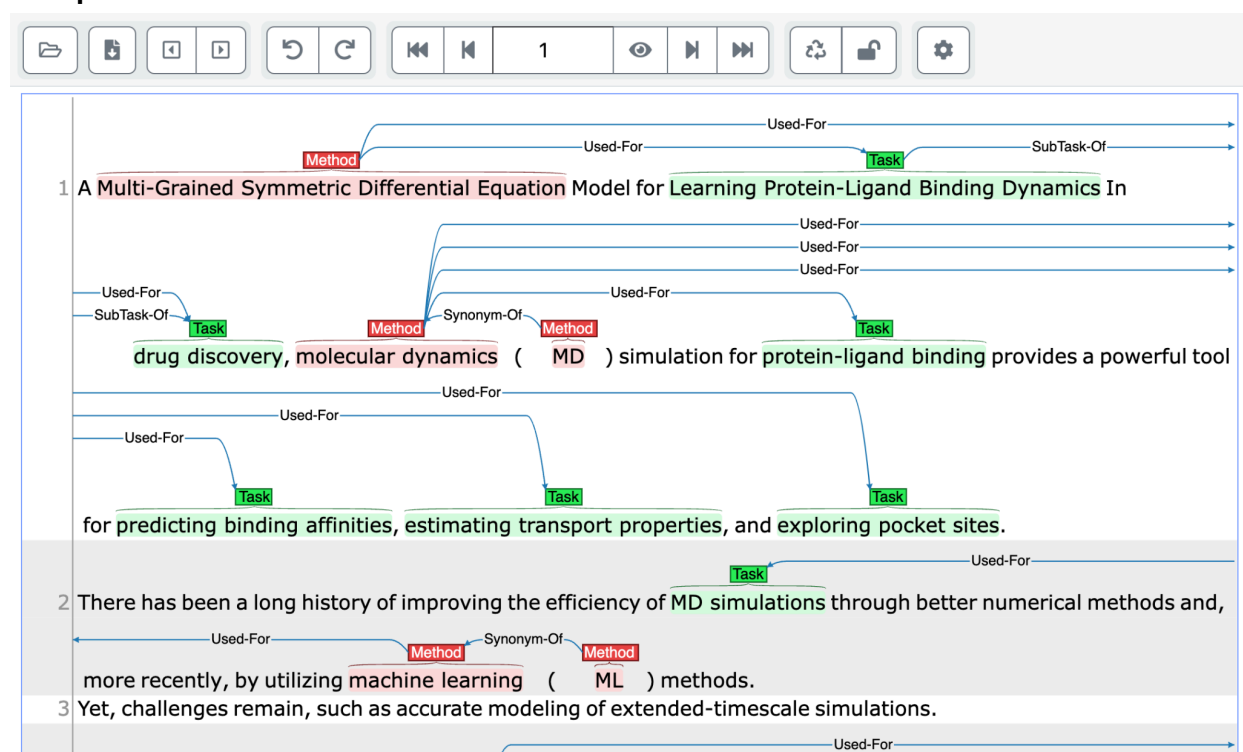
# SciER Annotation Guideline

## Abstract

This document is the annotation guideline for our joint entity recognition and relation extraction datasets for scientific documents in the Artificial Intelligence (AI) domain. We provide the definitions for 3 types of entities and 10 types of relations. Considering the difficulty of the annotation task, we will provide examples to help annotators have better understanding. The annotators should first read the whole guideline before labeling and report any questions during annotating.

# Annotation Tool

## INCEpTION 30.4



## Annotation tool interface

# Entity Annotation

## Entity Definition

Scientific entities in the machine learning (ML) or Artificial intelligence (AI) domains refer to key concepts or components that are integral to the structure and study of ML/AI papers. We follow the definition of entities/terms and build our annotation guides for NER based on the ACL RD-TEC Annotation Guideline [1], Papers With Code [2] and SciDMT [3]. We are interested in three specific entity types: Dataset, Task, and Method. Here are their explanations and definitions:

- **Dataset:** A realistic collection of data that is used for training, validating, or testing the algorithms. These datasets can consist of various forms of data such as text, images, videos, or structured data. For example, MNIST, COCO, AGNews, IMDb, etc.
- **Task:** A task in machine learning refers to the specific problem or type of problem that a ML/AI model is designed to solve. Tasks can be broad, like classification, regression, or clustering, or they can be very specific, such as Pedestrian Detection, Autonomous Driving, Sentiment Analysis, Named Entity Recognition and Relation Extraction.
- **Method:** A method entity refers to the approach, algorithm, or technique used to solve a specific task/problem. Methods encompass the computational algorithms, model architectures, and the training procedures that are employed to make predictions or decisions based on data. For example, Convolutional Neural Networks (CNNs),

## Notes

Considering that annotators may have varying understandings of the annotation details, we have defined a set of rules and notes to standardize the annotation process based on the annotation guideline of SciDMT [3].

### Cases of tokens to be ignored:

- We do not annotate **generics** and **determiners**. Generics cannot be used independently to refer to any specific entities, e.g., “This task”, “the dataset”, “a public corpus” etc. The determiners should not be part of an entity span. For example, given span “the SQuAD v1.1 dataset”, where the determiner “the” should be excluded the entity span. We refer ignoring
- **Minimum span principle.** Annotators should annotate only the minimum span necessary to

represent the original meaning of task/dataset/metric (e.g.: “The”, “dataset”, “public”, ‘method’, ‘technique’ are often omitted).

- **Partial Conjunction.** Annotators can ignore entities in conjunction since that token itself can

not represent an entity. For example, the span "Semeval 17, 18", we just need to annotate "semeval 17".

### Cases of tokens to be kept:

- **Factual entity.** Only annotate "factual, content-bearing" entities. Task, dataset, and method

entities normally have specific names and their meanings are consistent across different

papers. For example, the "CoNLL03", "SNLI" are factual entities, but the "a high-coverage sense-annotated corpus" is not a factual entity.

- **Include "corpus/dataset/benchmark"** when annotating dataset if these tokens are the

head noun of the dataset entity. For example: "ubuntu corpus", "SemEval-2010 Task 8 Dataset"

- **Abbreviation.** If both the full name and the abbreviation are present in the sentence,

annotate the abbreviation and its corresponding full name separately. For instance, "20-newsgroup (20NG)", we annotate "20-newsgroup" and "20NG" as two separate entities.

- **Complete Conjunction.** If the entity is in conjunction with other detected entities using 'and'

or comma and following the above rules for keeping (eg: factual entity), annotate them.

Example 1: "will be a useful tool in numerous research fields including video analysis, human inspired motion generation, learning by demonstration, intuitive human-robot interaction, and human behavior analysis." Each research field should be annotated as a task since we have detected 'human-robot interaction' and 'human behavior analysis' as a task.

Example 2: "effectiveness of VAC+GAN", 'VAC' and 'GAN' are labeled as two separate Methods.

Example 3: "Our method is combining VAC and GAN".

## Relation Annotation

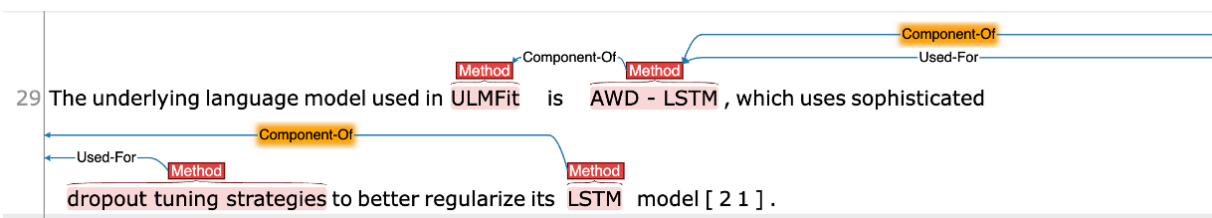
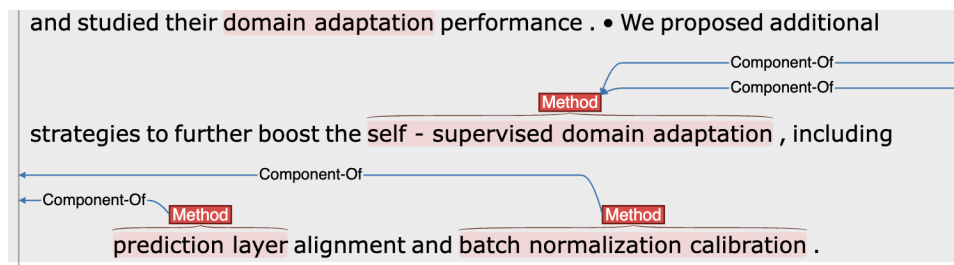
### Relation Definitions

Relation links cannot exceed the sentence boundary. We define 10 types of relationships for the Dataset, Method, and Task entities. We classify these relation types to two classes, 1) **specific entity-type relations** and 2) **general entity-type relations**.

**Specific entity-type relations:** These relations are defined with specific restrictions on the types of entities that can participate in the relationship. Each relation in this class explicitly involves particular entity types (like Dataset, Task, or Method), indicating that the relationship is only applicable between those specific types.

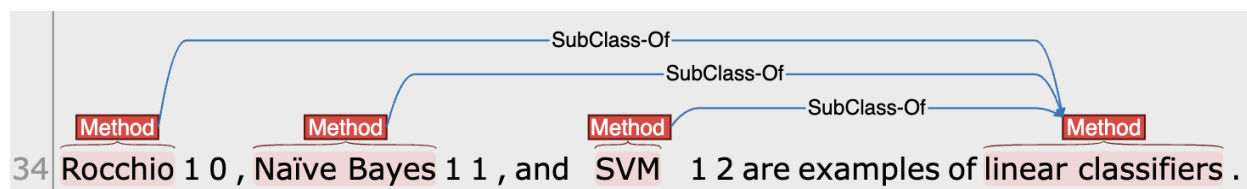
- **Part-Of (Method-Method)**

- Definition: This relationship denotes that one method is a component or a part of another method.
- Examples(*In the old version, I used the Component-Of as the relation name*):



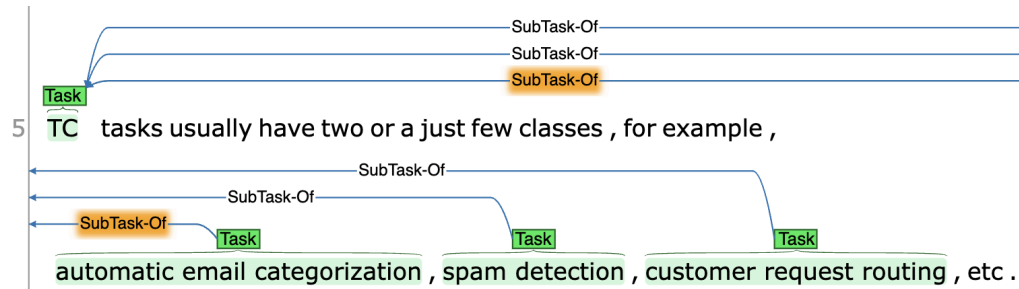
- **SubClass-Of (Method-Method)**

- Definition: Specifies that one method is a subclass or a specialized version of another method.
- Example:



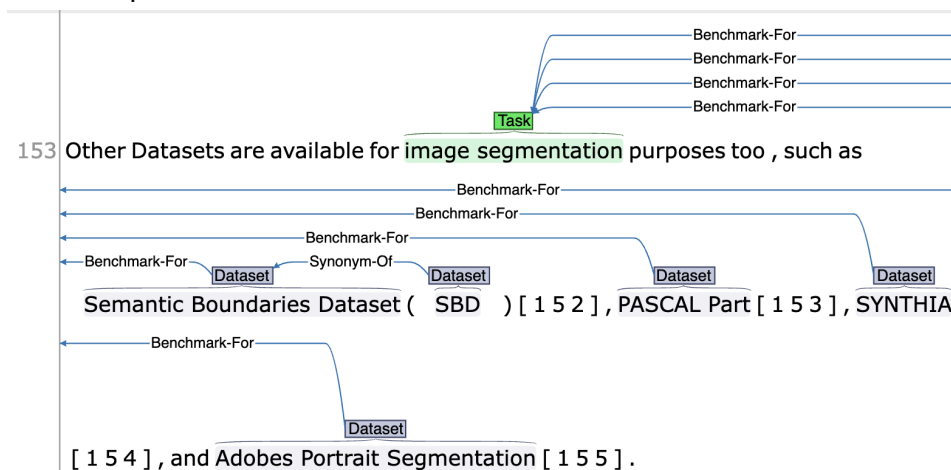
- **SubTask-Of (Task-Task)**

- Definition: Indicates that one task is a subset or a specific aspect of another broader task.
- Example:



- **Benchmark-For (Dataset-Task)**

- Definition: Shows that a dataset serves as a standard or benchmark for evaluating the performance of methods on a specific task.
- Example:



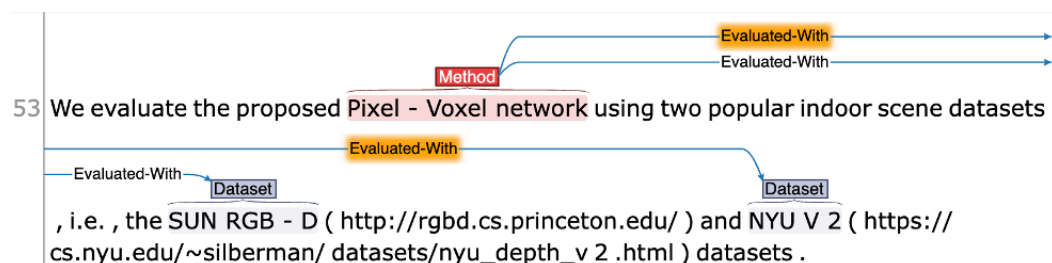
- **Trained-With (Method-Dataset)**

- Definition: Indicates that a method is trained using a specific dataset.
- Example:

19 Therefore , we pre - process the videos and extract frame - level features using a state - of - the - art deep model : the publicly available Inception network [ 4 ] trained on ImageNet [ 1 4 ] .

- **Evaluated-With (Method-Dataset)**

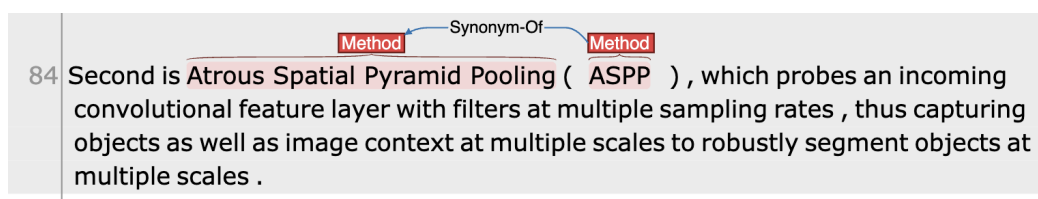
- Definition: This relationship denotes that a method is evaluated using a specific dataset to test its performance or conduct the experiments.



**General Entity-Type Relations:** These relations do not specify restrictions on the types of entities that can participate. This makes them more flexible and applicable to any two entities, whether they are of the same type or different types. This category allows for broader usage across different contexts and scenarios.

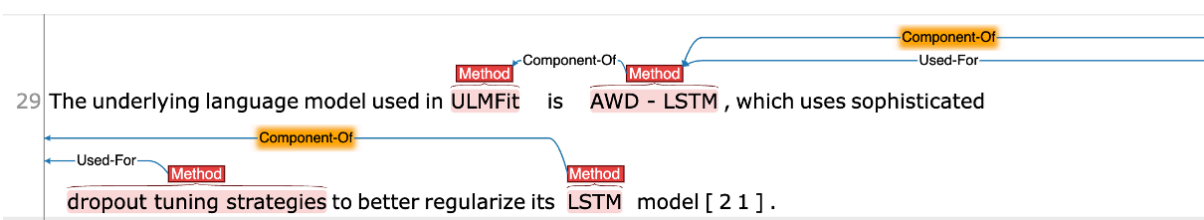
- **Synonym-Of**

- Definition: Indicates that two terms or entities are considered to have the same or very similar meaning, such as abbreviation.
- Example:

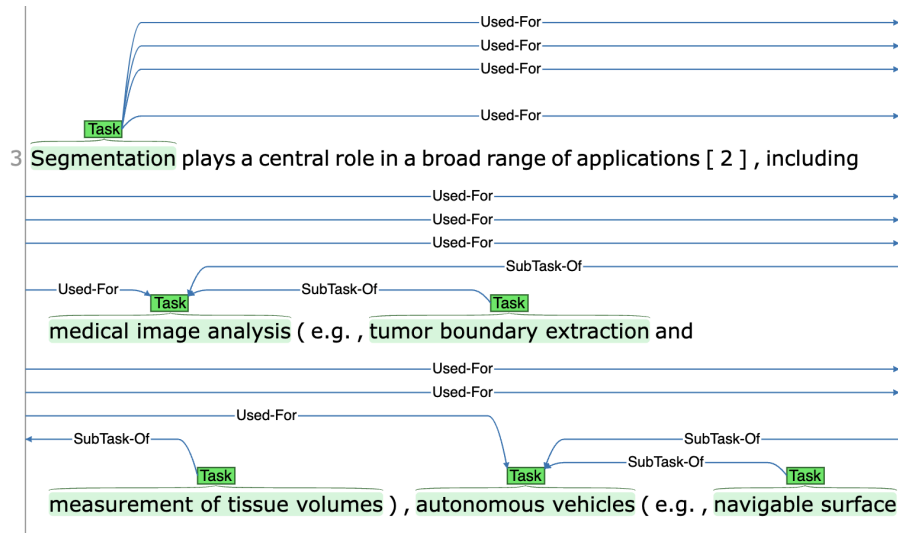


- **Used-For (修改 ! ! !)**

- Definition: Shows that one entity is utilized for achieving or performing another entity. This relation is highly flexible, allowing for generic relationships across diverse entities.
- Example:  
Example 1: where the “dropout tuning strategies” method is not a part of the method’s architecture but a method used for another method.

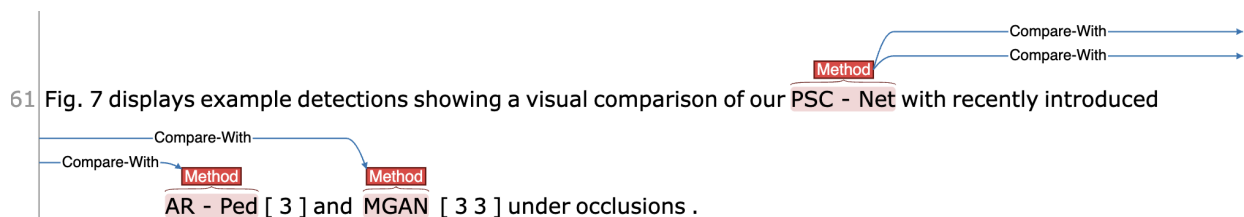
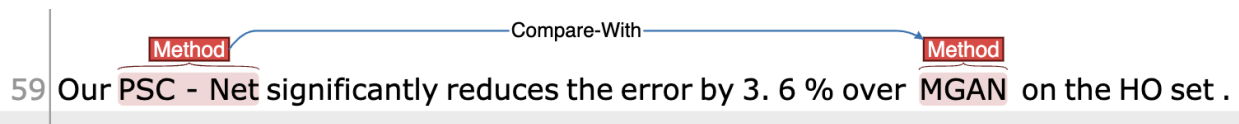


Example 2: in this example, “Segmentation” is a task used in many other tasks.



- **Compare-With**

- Definition: This relationship is used when one entity is compared with another to highlight differences, similarities, or both.
- Example:



## Notes

- **Negative Relations:** Please do not annotate negative relations. For example, X is not used in Y or X is hard to be applied in Y.
- **Check Entity Types for Specific Entity-Type Relations:** Verify that the entities involved in the relation match the prescribed types (e.g., Method-Dataset for Trained-With). Incorrect entity types should not be linked by these specific relations.

- **Direct Evidence:** Annotate a relationship only if there is direct evidence or clear implication in the text. Avoid inferring relationships that are not explicitly mentioned or clearly implied.
- **Consistency:** Ensure consistency in how relationships are annotated across different texts. If uncertain, refer back to the guideline definitions or consult with a supervisor.
- **Avoid Assumptions:** Do not make assumptions about relationships based on personal knowledge or external information. Rely solely on the information provided in the text

## References

- [1] QasemiZadeh, Behrang, and Anne-Kathrin Schumann. "The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods." Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016.
- [2] <https://paperswithcode.com/about>
- [3] SciDMT: A Large-Scale Corpus for Detecting Scientific Mentions. COLING-LREC 2024. <https://huggingface.co/datasets/jopan/SciDMT>