SH-NER

Annotation Guideline

Abstract

This annotation guideline supports the structured labelling of scientific publications within the computer science domain, with a focus on identifying key technological and research-related concepts. It outlines the definitions and boundaries of multiple entity types across hardware and software entities. The aim is to assist annotators in consistently tagging elements, to aid clarity and precision, examples from annotated texts accompany each entity definition. Annotators need to review the entire guideline before beginning, and to provide feedback on edge cases or ambiguities encountered during trial annotations.

Annotation Tool

Doccano 1.8.3

Hardware & Software

- ***Hardware Entity***

A **Hardware Configuration Entity** refers to explicit mentions of computational hardware components (e.g., GPUs, CPUs, or other processors) and their specifications within scientific texts. These entities document the physical hardware used to conduct experiments, simulations, or data processing in research.

Annotate when the text includes:

- Specific hardware names (e.g., “NVIDIA A100”, “Intel Xeon”)

- Hardware with specs (e.g., “32-core CPU”, “16GB VRAM GPU”)
- Annotate the full span if it includes useful context

Do Not Annotate:

- Vague mentions like “high-end server” or “powerful machine or we used GPU or CPU” without specific hardware reference
- Software, cloud platforms, or hosting environments

Examples:

- The experiments were run on a **single NVIDIA A100 GPU.**
  NVIDIA A100 GPU : Hardware entity
- We used a 64-core AMD EPYC processor for model training.
  64-core AMD EPYC processor : Hardware entity

- ***Cloud Platform***

The Cloud_Platform entity refers to explicitly named remote computing environments provided by third-party vendors that deliver scalable infrastructure (compute, storage, networking) via the internet. These platforms serve as virtualized hardware foundations for running machine learning experiments, large-scale data processing, or deployment of scientific applications.

Annotation Criteria:

- Annotate only direct mentions of specific cloud services or platforms.
- Include both branded names (e.g., AWS, Google Cloud)

Do Not Annotate:

- Do not annotate vague or descriptive uses without a clear platform reference (e.g., “the data was processed in the cloud” without naming a platform).
- The model was deployed to the cloud.” (No specific platform is mentioned)
- “Cloud-based processing was applied.” (No specific platform is mentioned)

Examples:

- Text: “We trained our model using **AWS** EC2 instances.”
- AWS : Cloud Platform
  EC2 : not Cloud Platform
  Note: EC2 (Elastic Compute Cloud) is a specific service offered under the AWS umbrella — it's more like a product or instance type rather than a standalone cloud platform.

- Text: “The experiments were executed on **Google Cloud** infrastructure.”
- Google Cloud : Cloud Platform
- Text: “**Azure** was used for scalable storage and training.”
- Azure : Cloud Platform

- Software Entity:

The Software Entity refers to specialized libraries, tool and frameworks that provide pre-written code and tools for specific tasks in software development, particularly in machine learning, data processing, computer vision, and natural language processing. These libraries help streamline and standardize processes by providing reusable functions, algorithms, and utilities.

Examples:

- Text: "The model was trained using **PyTorch** for deep learning."

PyTorch : Software Entity

- "We applied **TensorFlow** to deploy the neural network."
  TensorFlow : Software Entity
- "The data was pre-processed using **Scikit-learn**."
  Scikit-learn : Software Entity
- Text: "We used **Kubernetes** to manage containerized applications."
  Kubernetes : Software entity
- Text: "The data processing was done using **Hadoop** clusters."
  Hadoop : Software entity

Incorrect Annotation:

- "We used a machine learning library for the task." (Generic term, no specific library named)
- "The framework provides various data pre-processing functions." (No specific library or framework mentioned)

Resources

This guideline is inspired by:

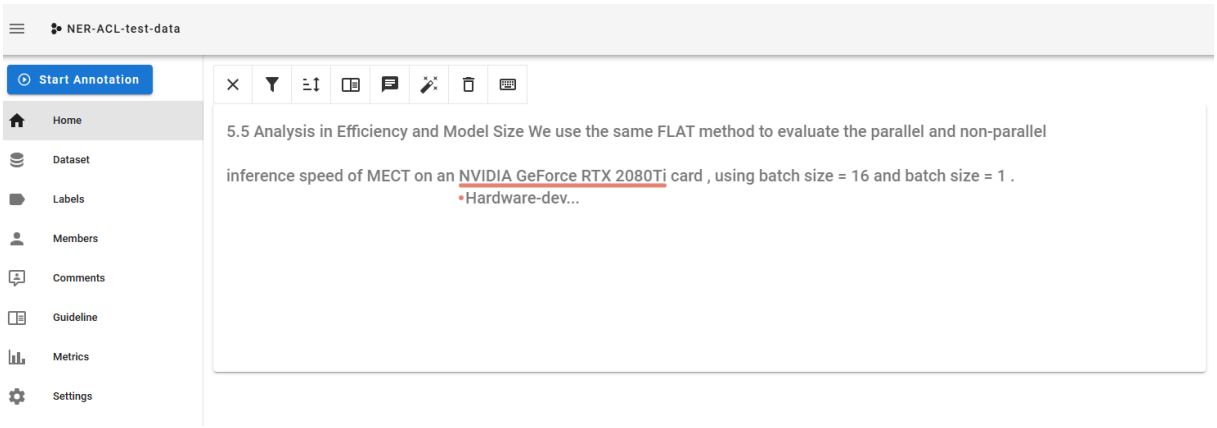
SciERC annotation guideline

- http://nlp.cs.washington.edu/sciIE/annotation_guideline.pdf

ACL TEC annotation guideline

- https://doi.org/10.13140/RG.2.1.1939.1446

Some annotation samples from our annotated dataset.



A.6 Infrastructure and Reproducibility We run all experiments on a single 32GB NVIDIA Tesla V100 GPU .
•Device-Count
•Device-Memor...
•Hardware-dev...

A.6 Infrastructure and Reproducibility We run all experiments on a single 32GB NVIDIA Tesla V100 GPU .
•Device-Count
•Device-Memor...
•Hardware-dev...

The whole experiment is carried out on 1 TITANX GPU .
•Device-Count
•Hardware-dev...

We trained BERT-base [Devlin et al. , , 2019] and RoBERTa-base [Liu et al. , , 2019] on this data for 10 epochs with early stopping , and a batch size of 8 × 2 gradient accumulation steps — all other parameters are defaults set by Huggingface [14] .
•Software-Ent...