Case Study: Named Entity Recognition (NER) Annotation

Overview

This case study documents a **manual Named Entity Recognition (NER) annotation project**. The project focuses on **span-based entity annotation**, schema discipline, and careful handling of ambiguous cases using professional annotation workflows.

All annotations were completed manually using **Label Studio**, with the goal of producing clean, machine-readable data suitable for AI training and evaluation.

Objective

The objective of this project was to:

- Demonstrate practical experience with NER annotation
- Apply a fixed entity schema consistently
- Handle edge cases and ambiguity using annotation comments
- Produce high-quality JSON outputs aligned with real-world NLP pipelines

Dataset

- **60 short texts**
- Mixed-domain content:
  - News-style sentences
  - Institutional references
  - Everyday language
- Sentences intentionally designed to include:
  - Clear entity mentions
  - Ambiguous cases
  - Titles vs proper names
  - Relative and absolute temporal expressions

Entity Schema

A fixed schema was defined and strictly followed throughout the project.

Entity Type	Description
Person	Named individuals (proper names only)
Organization	Companies, institutions, government bodies, media outlets
Location	Geographical locations (cities, countries, regions, landmarks)
Date	Absolute and relative temporal expressions

No additional entity types were introduced.

Tooling

- **Annotation platform:** Label Studio
- **Annotation type:** Manual, span-based
- **Input format:** CSV
- **Output format:** JSON

The labeling interface was configured to allow precise text span selection and optional inline comments.

Methodology

The annotation process followed these core principles:

- **Span precision:**
  Only exact entity names were annotated, without leading or trailing whitespace.
- **Schema discipline:**
  Entities were labeled strictly according to the predefined schema.
  If a mention did not clearly belong to one of the allowed entity types, it was excluded.
- **Context-aware decisions:**
  Entity labeling was based on how a term functioned in the sentence, not on recognizability alone.
- **Under-labeling preference:**
  In ambiguous cases, entities were excluded rather than labeled speculatively.

Handling Ambiguity

Special attention was given to ambiguous or borderline cases, including:

- Organization names used only as part of job titles
- Titles without explicit person names
- Generic nouns resembling entities
- Nested organization names containing location references
- Relative date expressions (e.g. *“last Friday”*, *“this week”*)

In such cases, short annotation comments were added to document the decision rationale.

Quality Control

Quality checks focused on:

- Correct entity types
- Clean span boundaries
- Consistent schema application
- Minimal but meaningful comments

Output Structure

Each annotated task includes:

- One or more span-based entity labels
- Character offsets for each span
- Entity type assignment
- Optional annotation comment
- Annotation metadata (including lead time)

The exported JSON is suitable for:

- NLP model training
- Evaluation datasets
- Further processing by ML pipelines

Skills Demonstrated

- Named Entity Recognition (NER)
- Span-based annotation
- Annotation guideline interpretation
- Ambiguity handling and documentation
- Label Studio configuration and use
- Quality-focused human-in-the-loop workflows