# Project Plan - NERC of different granularities

Group 10: Laura Komorek, Oliver Jung, Thorsten Trinkaus

February 9, 2026

## 1 Task Description

The aim of our project is to investigate Named Entity Recognition and Classification (NERC) at different levels of granularities. To get a good understanding of not only the capabilities, but also the limitations of pre-trained language models (PLMs) in the context of NERC, we will increase the granularity of the entity types from coarse-grained up to ultra-fine-grained. This section includes a summary of our main focus, the phenomena we try to investigate, and the hypotheses we want to test.

### 1.1 Project Focus

Our focus for this project lies on the performance of modern language models, such as masked language models (MLMs), T5, and NLI-based approaches, to investigate how well they can handle ever growing label spaces, increasing ambiguity, as well as context dependencies. To achieve this, we will investigate the following phenomena:

- The effect of label granularity on model performance.
- Generalization capabilities across datasets with different type hierarchies.
- Robustness to multi-word named entities and implicit entity types.
- The information that the models rely on for making predictions.

### 1.2 Project Hypotheses

Over the course of this project, we would like to test the following hypotheses:

1. The model performance (especially when probing) degrades as the label granularity increases.
2. Fine-tuning improves performance but still shows confusions among semantically simular fine entity types.
3. Seq2Seq models (e.g., T5) outperform standard MLMs, such as BERT, on multi-word named entities duo to span reconstruction.
4. NLI-based formulations are more roobust to label granularity and rare entity types than direct classification.
5. Models heavily rely on surface-level lexical cues and dataset-specific biases, rather than deeper semantic understanding of the context.

## 2 Methods

To acomplish our project goals, we plan to utilize and compare several modeling paradigms for NERC across datasets with different levels of granularity, by combining quantitative evaluation with qualitative model inspection. This section should give an overview of the methods relevant to this project:

- **Fine-Tuned Approaches**
  - token-level and span-level classification

  - Seq2Seq generation (T5)

- **Probing Approaches**
  - Masked Language Modeling (MLM)
  - span masking (SpanBERT, T5)

- **NLI-based Setting**
  - recasting entity types as a NLI task using class-specific hypotheses

Our probing and fine-tuning strategies include masking the entity mention and predicting the entity type for MLMs, masked span reconstruction with entity type generation for T5, and recasting for NLI in the form of

**PREMISE:** *sentence containing the entity*
**HYPOTHESIS:** *the entity is of type X.*

The mesurements we will use to evaluate model performance include standard NERC metrics, such as precision, recall, and F1-score, as well as cross-dataset comparisons of evaluation scores, error and confusion analyses, and performance breakdowns by entity type, frequency, and type granularity.

# 3 Datasets and Models

## 3.1 Datasets

The following datasets should provide a good coverage of different levels of granularity for our experiments:

- **OntoNotes/CoNLL-2003**: coarse-grained entity types
- **FIGER**: 112 fine-grained entity types
- **Ultra-Fine-Entity Typing**: 10k+ ultra-fine-grained entity types

Interesting dataset specifics and statistics to investigate beforehand are class distribution (head. vs. tail types), the proportion of multi-word entity mentions, and the degree of context-dependence for entity type disambiguation. Additional datasets may be of interest, such as Entity type hierachies (e.g., FIGER, Ultra-Fine) and pre-trained NLI datasets (e.g., MNLI).

## 3.2 Models

We plan to experiment with the following models:

- **BERT/RoBERTa**: MLM-based
- **SpanBERT**: span-level predictions
- **T5**: Seq2Seq generation and masked span reconstruction

# 4 Experiments (at planning phase)

Our target is to evaluate context masking vs entity masking, foiling tests with closely related types, and zero-shot vs fine fine-tuned performance across the different datasets. To make our experiments interpretable, we first have to define the evaluation metrics, the baseline, and the specific experimental setup, which this section should acomplish.

### 4.1 Evaluation Metrics

- Precision, Recall, F1-score (both micro and macro)
- Accuracy (for OntoNotes)
- Recall at k (for ultra-fine-grained setting)
- Mention frequency vs performance analysis (bias detection)
- Case studies of specific examples (right for the right reason?)

### 4.2 Baseline

- Majority class baseline (predicting the most frequent entity type)
- Linear classifier with frozen embeddings (e.g., BERT embeddings with a linear layer on top)

### 4.3 Experimental Setup

- Unified training / dev / test splits across datasets (if possible)
- Comparision of:
  - probing vs fine-tuning
  - MLM vs T5 vs NLI-based approaches

### 4.4 Experimental Variations

- Entity type granularity (coarse vs fine vs ultra-fine)
- Masking strategies (entity masking vs context masking)
- Input representation and prompting templates

### 4.5 Stress Testing and Data Manipulation

To test the robustness of the models and to gain insight into their decision-making processes and biases, we plan to also perform some stress tests , such as:

- Comparing the original mention strings to masked versions to see if the model can predict from context alone.
- Deleting or shuffling parts of the context to see how much the model relies on specific cues.
- Counterfactual checks by swapping entity mentions across different contexts.

### 4.6 Expected Outcome

We expect to collect empirical evidance on the limitations of current PLMs in ultra-fine entity typing, to get an insight into the strengths and weaknesses of different modeling paradigms, and to identify dataset biases and heuristic model behaviors.

## 5 Planned Steps and Task Distribution

Especially at the beginning of the project, we will ensure a mutual understanding of the task, the datasets, and models, which requires a lot of colloboration and comunication. Once general agreement and understanding is established, we will work in parallel and the workload will be distributed in a manner that allows each team member to work on different model architectures, datasets, or experimental variations simultaneously. To ensure a complete and connected project, we will have regular meetings to discuss problems and findings. Therefore, the exact allocation of said tasks will be refined iteratively and adapted based on intermediate results and challenges during the project.