# Machine Learning for Named Entity Recognition and Classification

**J.W. Brook**

## Abstract

Fill in a short abstract for your final submission.

The final report has a strict maximum of 6 pages (12 columns). This does not include Acknowledgements or references (they may appear on Page 7). You can use appendices if you want to share more details.

## 1 Introduction

Length: ± 1 column for Introduction and Related Work sections combined.

Write a short introduction to your approach for the final submission. The introduction should include:

- A brief description of the task (not detailed)
- An outline of the ML approaches included in the report
- A brief summary of your results

*Assignment 3 / Final submission*

## 2 Related work

*Assignment 2*

## 3 Task and Data

Length: ± 1.5-2 columns for Task and Data.

### 3.1 Task

*Assignment 1*

This task involves Named Entity Recognition (NER), an NLP method that extracts information about named entities (i.e. proper nouns like names, locations, and organisations) from unstructured text through use of an automatic ML algorithm. The specific task is language-independent, meaning that language-specific rules for identifying Named Entities (NEs) may not be used. The data categorises named entities by Person (PER), Location (LOC), and Organisation (ORG), as well as employing a Miscellaneous (MISC) category for otherwise-uncategorised NEs. These categories are annotated according to the MUC (Message Understanding Conference) guidelines.

The CoNLL data representation aims to simply represent a collection of annotated sentences, with one word from said sentences per line, while empty lines represent sentence boundaries. Each line contains a word, as well as its part-of-speech (POS) tag, chunk tag, and NE tag.

*Assignment 2* Update if necessary

*Assignment 3* Update if necessary

### 3.2 Dataset and Distribution

*Assignment 1*

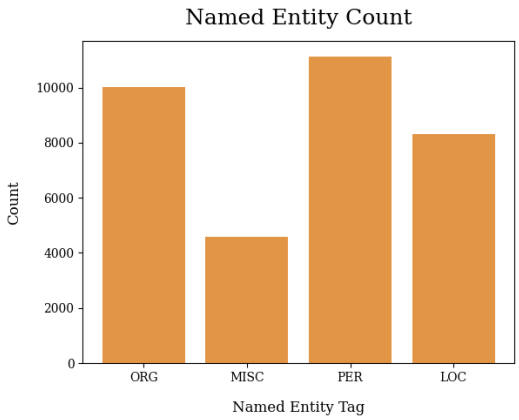


Figure 1: NE Tag Distribution

The distribution of NE tags in the training data is demonstrated in the above figure, from which we can see that we have relatively similar counts for PER (11128), ORG (10025), and LOC (8297), with significantly less MISC tags present (only 4593).

*Assignment 2* update if necessary

*Assignment 3* update if necessary

### 3.3 Preprocessing

*Assignment 1*

Thanks to the well-defined CoNLL formatting, very little preprocessing is necessary to train this NER model. The tokens, POS tags, and NE tags are all simply extracted from the structured data, while a boolean value is created for each word to record whether it is capitalised or not.

*Assignment 2* update (if applicable)

*Assignment 3* update (if applicable)

### 3.4 Evaluation Metrics

*Assignment 1*

The performance of the NER system is evaluated over tokens using three metrics: namely, precision, recall, and F1 score. These metrics are calculated for each possible tag value: ORG, PER, LOC, MISC, and O (no NE tag), and an average is taken. A confusion matrix is also created as a visual aid for better understanding of how individual tokens tend to be (mis)classified.

*Assignment 2 and 3*: Update if necessary

## 4 Models and Features

Length: ± 2 columns for Models and Features.

### 4.1 Models

*Assignment 1* Logistic Regression

The logistic regression model is an ML classification technique which can model the probability of an event taking place based on the log-odds of a linear combination of a collection of independent variables (features).

*Assignment 2* Alternative Methods (SVM, NB)

*Assignment 3* More advanced models (fine-tuning BERT, CRF for ReMA students only)

### 4.2 Features

*Assignment 1*

There are many features available which could be added to the model in order to improve its tagging accuracy. To this end, we propose three potentially useful features:

**Capitalisation:** Whether a word begins with a capital letter or not should be indicative of whether it may be an NE, as, in formal English at least, names of people, locations, and organisations are often capitalised.

**POS Tag:** Which POS the word belongs to (i.e. NNP, PRP, etc.) may also be very indicative of whether it is an NE. In the training data, more than 90% of NEs have the NNP POS tag

**Word Order:** The order in which a word appears in a sentence may also partially indicate its likeliness to have a given NE tag, as NEs are often the subject of a sentence, meaning they will often appear earlier.

*Assignment 2* More features

*Assignment 3* More advanced features

## 5 Experiments and Results

Length: ± 3 columns (including figures and tables) for Experiments and Results. Additional material can be provided in the Appendix.

### 5.1 Evaluation

*Assignment 1*:

The model's performance is evaluated token-wise using three metrics: precision, recall, and F1 score. The BIO labels are dropped for ease of interpretability, meaning B-PER and I-PER are merged into a single PER NE tag, for instance. This leaves us with 4 possible NE tags: ORG, PER, LOC, and MISC, with the O label representing no tag. Below is a table of the evaluation results for each tag, as well as average scores for the tags as a whole.

| | Precision | Recall | F1 Score |
|---|---|---|---|
| ORG | 0.86 | 0.79 | 0.82 |
| MISC | 0.85 | 0.70 | 0.77 |
| O | 0.99 | 0.99 | 0.99 |
| LOC | 0.82 | 0.65 | 0.72 |
| PER | 0.79 | 0.94 | 0.85 |
| Mean | 0.86 | 0.81 | 0.83 |

Table 1: Model Evaluation Scores

The confusion matrix below shows us that NE tags are most commonly misassigned to PER. The O tag is very rarely misassigned, with more than 99% of them being correct. Overall, these results are promising, and may be difficult to significantly improve upon in the future.

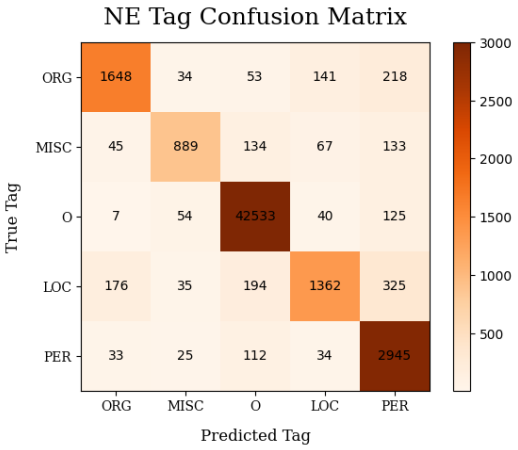


Figure 2: Confusion Matrix

*Assignment 2*: update

*Assignment 3*: update

### 5.2 Feature Ablation

*Assignment 3*: feature ablation for one system
Note: if you decide after your feature ablation that you would like to discuss this before the Evaluation, you may move this subsection.

## 6 Error Analysis

Length: ± 1.5-2 columns for Error Analysis.

*Assignment 3*: beyond the confusion matrix

## 7 Discussion

Length: ± 1 column for Discussion and Conclusion together.

*Assignment 3*

## 8 Conclusion

*Assignment 3*

## Acknowledgements

## References

## Theoretical Component

### Assignment 1

Machine Learning (ML) is a collection of data-driven approaches to solving problems using statistical techniques without human intervention. ML utilises algorithms that teach machines to learn and improve with data without explicit programming and can be used to solve real-world problems across various domains, e.g. natural language processing, computer vision, fraud detection, etc.

### Assignment 2

### Assignment 3

## NERC-research preparation

Precision, recall, and F-score are all evaluation metrics commonly used in ML, particularly in information retrieval tasks, to assess the performance of a model, in terms of how well the model does at correctly identifying relevant items while minimising false positives and false negatives.

Precision measures the proportion of correctly identified positive items out of all items predicted as positive, quantifying how accurate the model is when it predicts positive instances. High precision is important in e.g. email spam detection - when classifying emails as spam or not spam, high precision is crucial to avoid false positives, which would result in genuine emails being classified as spam and potentially being missed by the user.

Recall measures the proportion of correctly identified positive items out of all actual positive items, quantifying the ability of the model to find all positive instances in the dataset. High recall is important in e.g. COVID testing, where it is crucial to ensure that all positive instances (patients with the disease) are found, while false positives are less worrisome.

F-score is a single metric which combines precision and recall into a single value, providing a balanced measure of the model's performance by considering both false positives and false negatives.

## Time spent

Please use Table 2 to give an overview of the time you spent on each submission. You can modify the table (add new categories) if necessary.

| Week | Task | Time |
|---|---|---|
| 1 | watch videos | 30 minutes |
| 1 | understand labels in data | 1 hour |
| Total | 1h30min | |

Table 2: Time overview.

## A Example Appendix

This is an appendix. You can include additional analyses and tables here (e.g. samples you analyzed for your error analysis).