# Natural Language Processing and the Web

TECHNISCHE UNIVERSITÄT DARMSTADT

**Machine Learning Project - Documentation**

Benedikt Lins (1799381) und Stefan Thaut (1800351)
Department 20 - Computer Science
January 15, 2019

## 1 Foundations

In this Machine Learning Project we want to develop a named entity recognizer based on a machine learning approach. A named entity is a set of tokens, which form a name. Examples are "New York" or "Angela Merkel". We use a Conditional Random Field as the machine learning model.
Our training set contains a set of tokens with a chunk-annotation and the associated named entity class relating to the IOB-notation. We also have a german training set, that additionally contains a lemma for the given token. We also have a dev set, on which we evaluate our learned model.
An analysis for the english training set shows, that about 83 % of all tokens are no named entities, as we can see in figure 1.

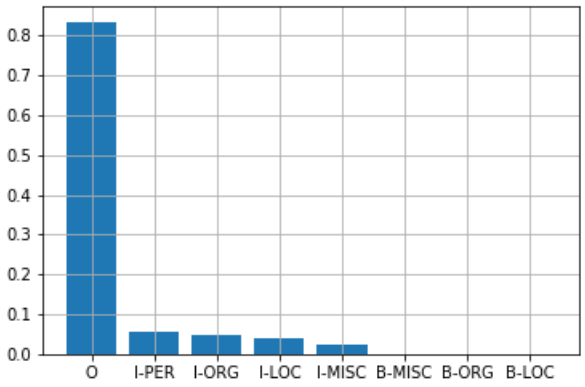



**Figure 1:** Class distribution of the english training set

Our CRF-model is based on the CRF-model we heard about in the fifth exercise of the lecture. So we already used the provided features:

- Is the first letter of a token capitalized? (Cap)
- Number of characters (NoOfChar)

In table 1 we can see the improvement of the measurements with this features in comparison to the baseline. The Micro F1 is increased by about just 0.035, but the Macro F1 is increased by about 0.15.

### 1.1 Evaluation

To have a baseline for an evaluation, with which we can compare our results, we added just a dummy feature to the CRF-model and did a testrun with the english trainingset and the english validationset as testset. The evaluationmetrics are shown in table 1.
To know which combination of features works best, we performed a grid search, to identify the features, that give an improvement in comparison to the baseline. The results are also shown in table 1 in appendix A.

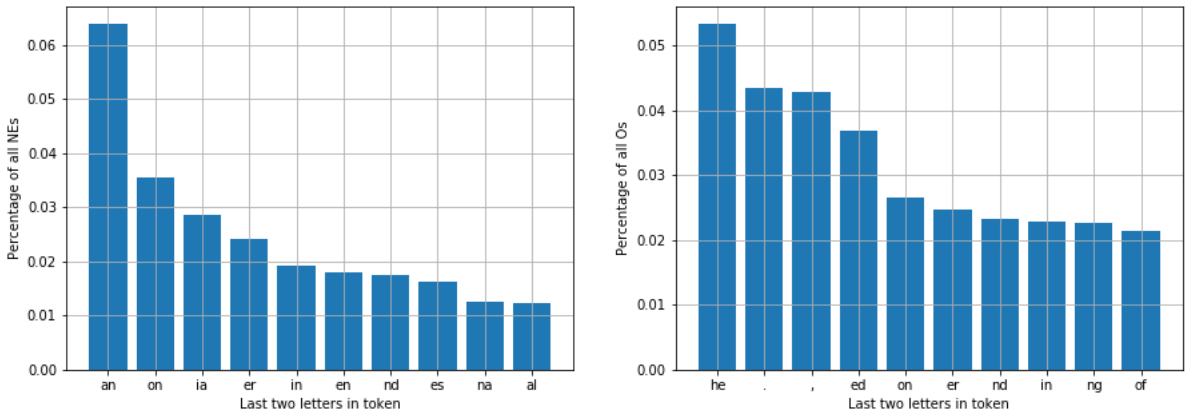



**Figure 2:** This figures show the ten most frequent ltc. The right side shows the ltc for the class *O* and the left side shows the ltc for all tokens, that are not in class *O*

In a first approach, we used every feature, that gives a positive improvement of the Micro and Macro F1. The evaluation of this setting results in an improvement of 0.00698 of the Micro F1 and 0.03180 of the Macro F1 in comparison to the baseline. For our interest, we added also the features, that give a negative improvement of the measures except the features belonging to the chunks. And remarkably this setting produces again a better result, that is in total a Micro F1 of 0.87529 and a Macro F1 of 0.32658.

## 2 Additional features

In this section, we describe additional features, with which we have tried to increase the measures.

### 2.1 Last two characters of a token

At first we have tried to use the last two characters (ltc) of a token as a feature. For the meaningfulness of this feature, we analysed the most frequent ltc for the class *O* and for the tokens, that are not in the class *O*. The ten most frequent ltc are presented in figure 2.
As we can see, about six percent of all named entities end with "an" and about five percent of all non named entities end with "he". In the next step it is important to find out, if some endings are typical for named entities or for non named entities. So we examined the appearances of the most frequent ltc of named entities in tokens that are non named entities and vice versa. The results for the ten most frequent ltc are shown in figure 3. The left side shows the comparison for the most frequent ltc of named entities. The blue curve represents the occurences of the ltc in named entities and the orange curve represents the occurences of the ltc in non named entities. So we can imply, that the greater the distance between the both curves is, the more specific the ltc is for the class. We identified the ltc "an", "ia", and "na" as typical for named entities and analogously the ltc "he", ".", ",", "ed" and "of" as typical for non named entities.

In a first testrun with this observations, we used only the ltc, that are typical for non named entities (ltcForO). As we can see in table 1, we do not get an improvement in comparison to the baseline with this new feature. When we add this features to the capital- and the number-of-char-feature, we even decrease the evaluation in comparison to the use of both of the features.

We assumed the insufficient classification as the reason for the lower evaluation. We determined just two classes: named entities and non named entities. But there are finer classes for the named entities. So in a second step we searched for the five most frequent ltc for each class and plotted it against the frequencies of the ltc in the other classes (see figure 4). The cell for a specific ltc in a specific class is brighter, the more typical the ltc is for this class (i.e. the more token in the class end with the ltc). Thus we can use a ltc as a feature, if a cell for this ltc is nearly white or if all other cells in the column of a ltc are black (i.e. this ltc does not occur in any other class).

### 2.2 POS-Tags

With the same strategy as in figure 4 we inspected the POS tags per each class. The relevant POS tags are plotted in figure 5a. All other POS tags are not on a significant level except for the class "O".

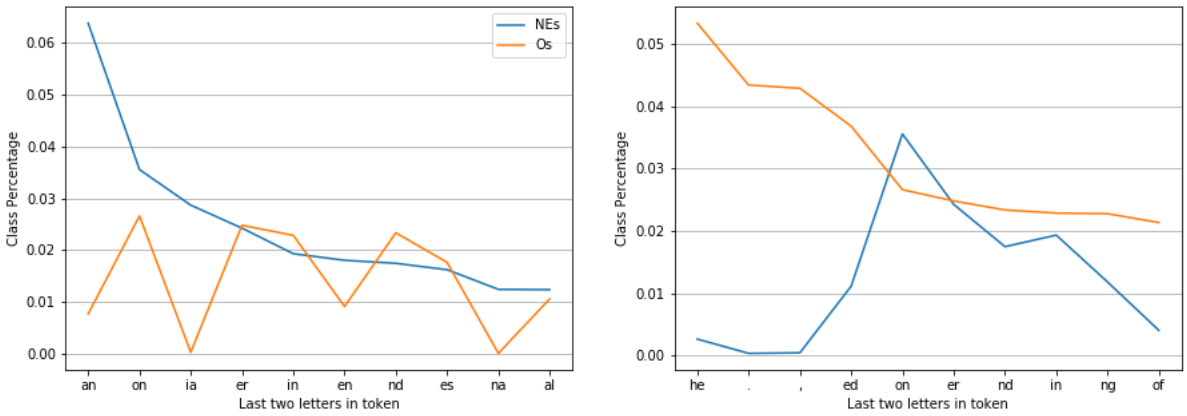



**Figure 3:** This figures show the percentages of the ten most frequent ltc in comparison of the classes named entity and non named entity

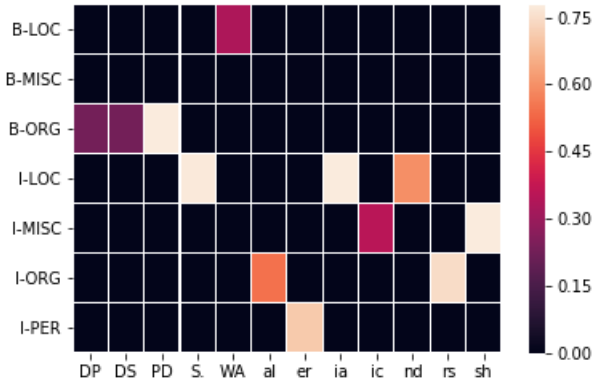
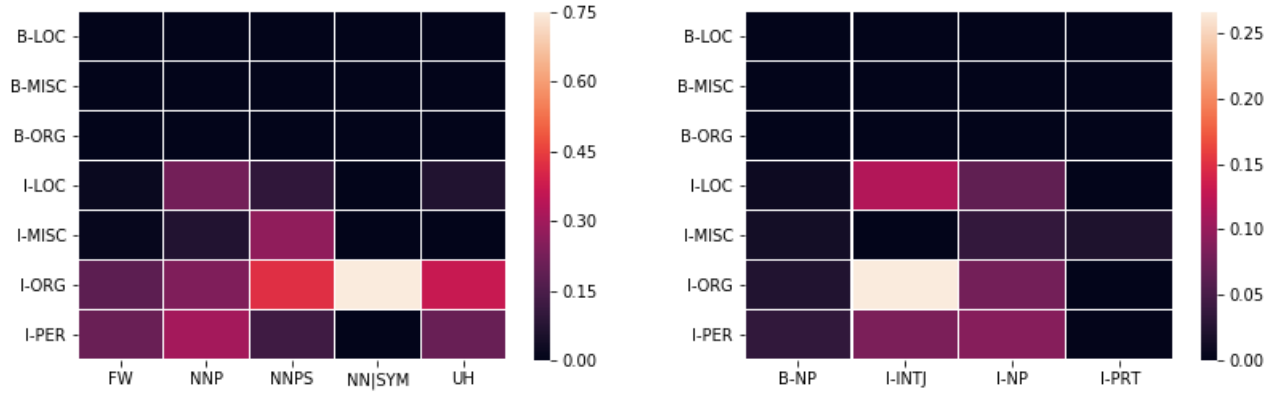



**Figure 4:** This figure show the percentages of the five most frequent ltc per classes

## 2.3 Chunks

As a last feature, we added the chunk annotation provided by the dataset. For this feature we had to add an additional annotation type. This type serves only for searching for it in the feature classes.
The relevant chunks are shown in figure 5b. We can see, that the most relevant chunk is "I-INTJ". But we have to consider, that the relative frequency of this chunk is only by 25 % in all classes. So this chunk rather represents the class "O".




**(a)** Relevant POS tags for each class

**(b)** Relevant chunks for each class

**Figure 5:** Distribution of the POS tags, resp. chunks for the classes

## 3 Results of the test set

In this section, we present our results of the evaluation on the finally given test set. In this run, we used the same configuration, we described at the end of section 1.1. Also we used the training set merged with the dev set as a new training set. This evaluation gives the following values:

$$\text{Micro F1:} = 0.89022 \tag{1}$$

$$\text{Macro F1:} = 0.44749 \tag{2}$$

## 4 Project structure

For a better understanding, we finally want to describe our project structure.

### 4.1 Data

We placed the data sourcefiles in the resources-folder in a folder called "data". We devided the files language dependend. So we have a folder for the german ("de") and a folder for the english ("en") files. Each of this folders contains the sourcefiles for the data, splitted into train (".train"), development (".dev") and test set (".test"). In the english folder, we also have a sourcefile, that contains the training data merged with the development data (".traindev").

### 4.2 Code

Our sourcecode is placed in the "java"-folder. We have the following packages:

- features: Contains our self designed features
- ner: Contains the CRFSuite for the NER-task
- reader: Contains the data reader for the data sourcefiles
- types: Contains the additional annotation types for the DKPro pipeline
- utils: Contains just the util class provided by the CRFSuiteDemo from the last exercise of the lecture

**Table 1:** Results of the evaluation. There is a "1", if the feature was used for the evaluation run. The captions of the first three columns are abbreviations (D = Dummy, FLC = FirstLetterCapital, NrC = NumberOfChars)

| D | FLC | NrC | er | rs | al | sh | ic | S. | ia | nd | PD | DP | DS | WA | UH | NNSYM | NNPS | NNP | I-NP | I-INTJ | Micro F1 | Macro F1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | | | | | | | | | 0,8253 | 0,11304 |
| | 1 | 1 | | | | | | | | | | | | | | | | | | | 0,86023 | 0,26933 |
| | 1 | 1 | 1 | | | | | | | | | | | | | | | | | | 0,860838 | 0,270758 |
| | 1 | 1 | | 1 | | | | | | | | | | | | | | | | | 0,860687 | 0,271021 |
| | 1 | 1 | | | 1 | | | | | | | | | | | | | | | | 0,861979 | 0,274005 |
| | 1 | 1 | | | | 1 | | | | | | | | | | | | | | | 0,86172 | 0,287434 |
| | 1 | 1 | | | | | 1 | | | | | | | | | | | | | | 0,859933 | 0,270984 |
| | 1 | 1 | | | | | | 1 | | | | | | | | | | | | | 0,860622 | 0,269557 |
| | 1 | 1 | | | | | | | 1 | | | | | | | | | | | | 0,862539 | 0,275036 |
| | 1 | 1 | | | | | | | | 1 | | | | | | | | | | | 0,861139 | 0,271407 |
| | 1 | 1 | | | | | | | | | 1 | | | | | | | | | | 0,86045 | 0,269773 |
| | 1 | 1 | | | | | | | | | | 1 | | | | | | | | | 0,860342 | 0,269396 |
| | 1 | 1 | | | | | | | | | | | 1 | | | | | | | | 0,860321 | 0,269303 |
| | 1 | 1 | | | | | | | | | | | | 1 | | | | | | | 0,860385 | 0,269374 |
| | 1 | 1 | | | | | | | | | | | | | | 1 | | | | | 0,8605146 | 0,2697153 |
| | 1 | 1 | | | | | | | | | | | | | | | 1 | | | | 0,8604716 | 0,2696812 |
| | 1 | 1 | | | | | | | | | | | | | | | | 1 | | | 0,8614191 | 0,2759899 |
| | 1 | 1 | | | | | | | | | | | | | | | | | 1 | | 0,8657047 | 0,3006439 |
| | 1 | 1 | | | | | | | | | | | | | | | | | 1 | | 0,8596963 | 0,2688408 |
| | 1 | 1 | | | | | | | | | | | | | | | | | | 1 | 0,8603424 | 0,2693733 |
| | 1 | 1 | 1 | 1 | 1 | 1 | | 1 | 1 | 1 | 1 | 1 | | 1 | | | | | | | 0,867212 | 0,301131 |
| | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | | | | | | | 0,867126 | 0,301915 |
| | 1 | 1 | | | | | | | | | | | | | 1 | 1 | 1 | 1 | | | 0,868978 | 0,309946 |
| | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | | | 0,875288 | 0,326584 |
| | 1 | 1 | | | | | | | | | | | | | | | | | 1 | 1 | 0,859782 | 0,268545 |
| | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | | 1 | 0,87503 | 0,326133 |