Natural Language Processing and the Web

TECHNISCHE UNIVERSITÄT DARMSTADT

Machine Learning Project - Documentation

Benedikt Lins (1799381) und Stefan Thaut (1800351)
Department 20 - Computer Science
December 13, 2018

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.
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

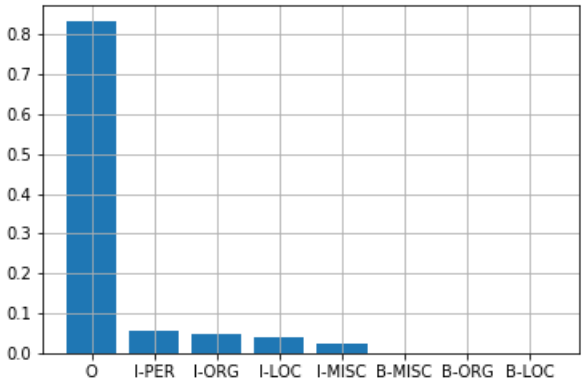

Figure 1: Class distribution of the english training set

2 Additional features

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 3.
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 **??**. 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.

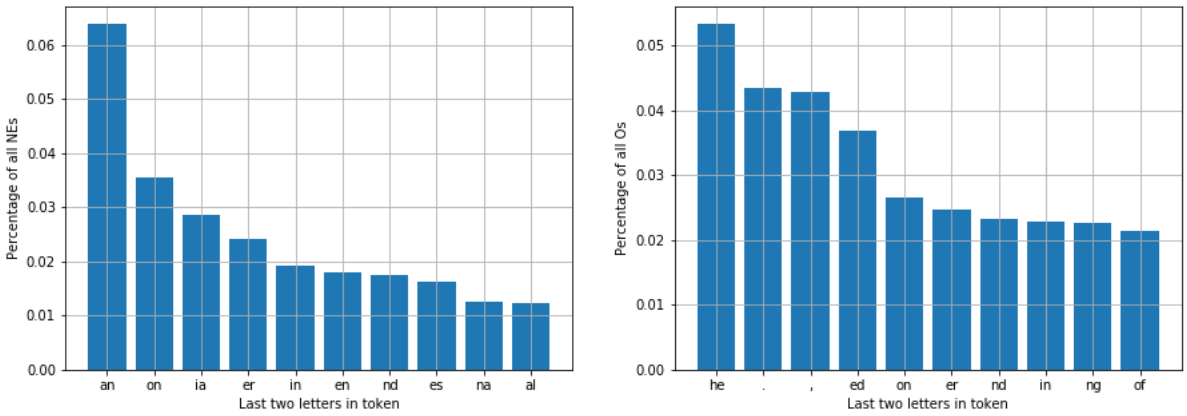

Figure 2: This figures show the ten most frequent ltc. The right side shows the ltc for the class *O* and the left side shows the ltc for all tokens, that are not in class *O*

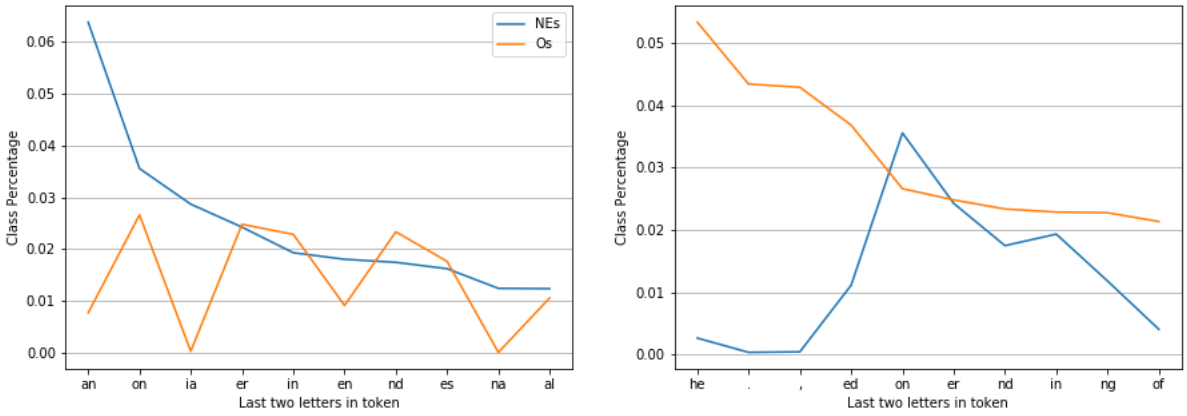

Figure 3: This figures show the percentages of the ten most frequent ltc in comparison of the classes named entity and non named entity