NLP Lab 7 - Structured Perceptron

Moritz Bohm, 140164022

Current Word-Current Label:

```
 ------------------------
Highest weighted features:
[[('1996_O', 3898.9138678223185), ('1_O', 1566.25), ('0_O', 1169.6603092783505), ('2_O', 1096.1199004975124), ('08_O', 765.5), ('3_
O', 735.9076923076923), ('1996_LOC', 725.0), ('6_O', 685.7146341463415), ('4_O', 610.3692307692307), ('AT_O', 505.5)]]
 ------------------------
F1_Score:
0.9464096385542169
 ------------------------
```

Making use of the current word-current labels over 10 passes yields very successful results, achieving an F1 score of 0.9464. This success is possibly down to the dataset containing a large proportion of “O” tags, and this is also shown in the 10 highest weighted features, which are almost entirely a word with an “O” tag.

Current Word-Current Label and Previous Label-Current Label:

```
 ------------------------
Highest weighted features:
[[('O_O', 13632.858212375859), ('None_O', 6860.5), ('LOC_O', 6092.017786187323), ('ORG_O', 5091.0085714285715), ('None_LOC', 5020.5
), ('1996_O', 3899.336222776032), ('None_ORG', 3560.5), ('O_ORG', 2035.414778325123), ('O_LOC', 1765.5), ('1_O', 1566.25)]]
 ------------------------
F1_Score:
0.9459277108433735
 ------------------------
```

Introducing previous label-current label into the feature set decreases the F1 score over 10 passes from just current word-current label, which is not expected but is most probably due to the fact that the dataset is skewed towards the “O” tag, and as such, introducing more context will reduce the F1 score. It is not surprising that the “O_O” and “None_O” are the two most highly weighted features. Upon closer analysis of the training data, “1996_O” is the most common word-label pair, which is, once again, reflected in the results by being the 7th highest weighted feature.

Other features which can be implemented:

Two additional features which could be implemented in the system are sub-word features, such as prefixes or suffixes, as well as label trigrams. Both of these improvements would give the perceptron additional context according to which it can predict the label sequence, and in doing so, would hopefully increase the F1 score due to the fact that it provides more context in the feature vector. Furthermore, it would also Due to time constraints, and efficiency concerns, I have chosen not implemented these features.