Evaluation of the English section of the MultiNERD dataset.

Adam Ek

1 Results

The results obtained from System A and B are shown in Table 1. We can note that system B, which uses a limited label set, performs in terms of precision, recall, and F1-score better than system A.

System A

	Accuracy	Precision	Recall	F1-score
Validation	0.978	0.844	0.819	0.828
Test	0.983	0.841	0.818	0.826

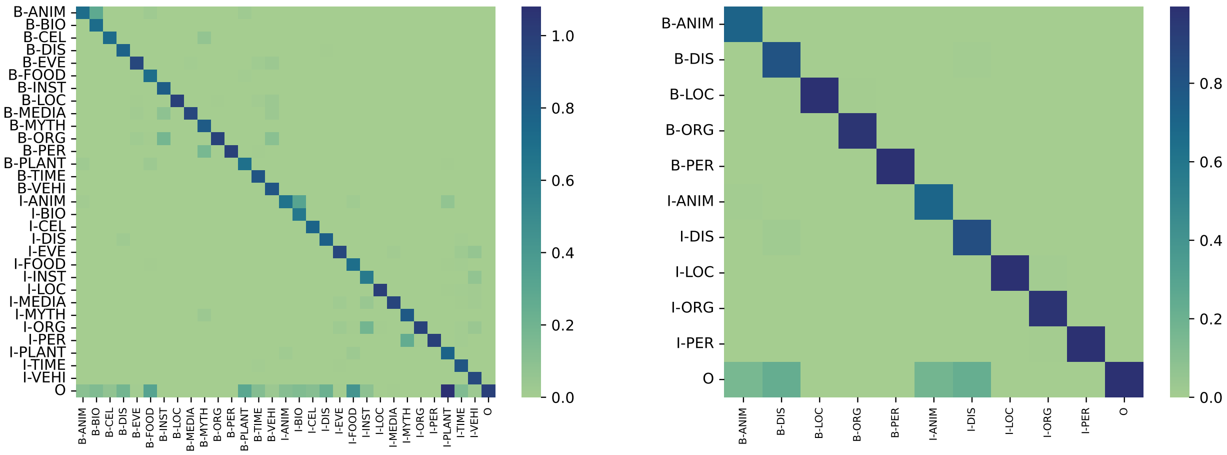
System B

	Accuracy	Precision	Recall	F1-score
Validation	0.989	0.916	0.905	0.91
Test	0.99	0.908	0.917	0.912

Table 1: Results on the test and validation split for System A and B. Each model was trained for 2 epochs.

We consider how often a label is incorrectly predicted in favor another label, this is shown for both systems in Fig. 1. From the confusion matrices we can observe that for system A, the most common misprediction for the `O`-tag is `I-PLANT`-tag. This indicates that either the model is not aware of the structure (I-tags occurring without a preceding B-tag), or that the model tends towards long plant names. For system B, the most common mistake is predicting `O` as `B-DIS` and `I-DIS`. Both of these error are likely due to the imbalanced dataset, i.e. there is not sufficient evidence for certain tags.

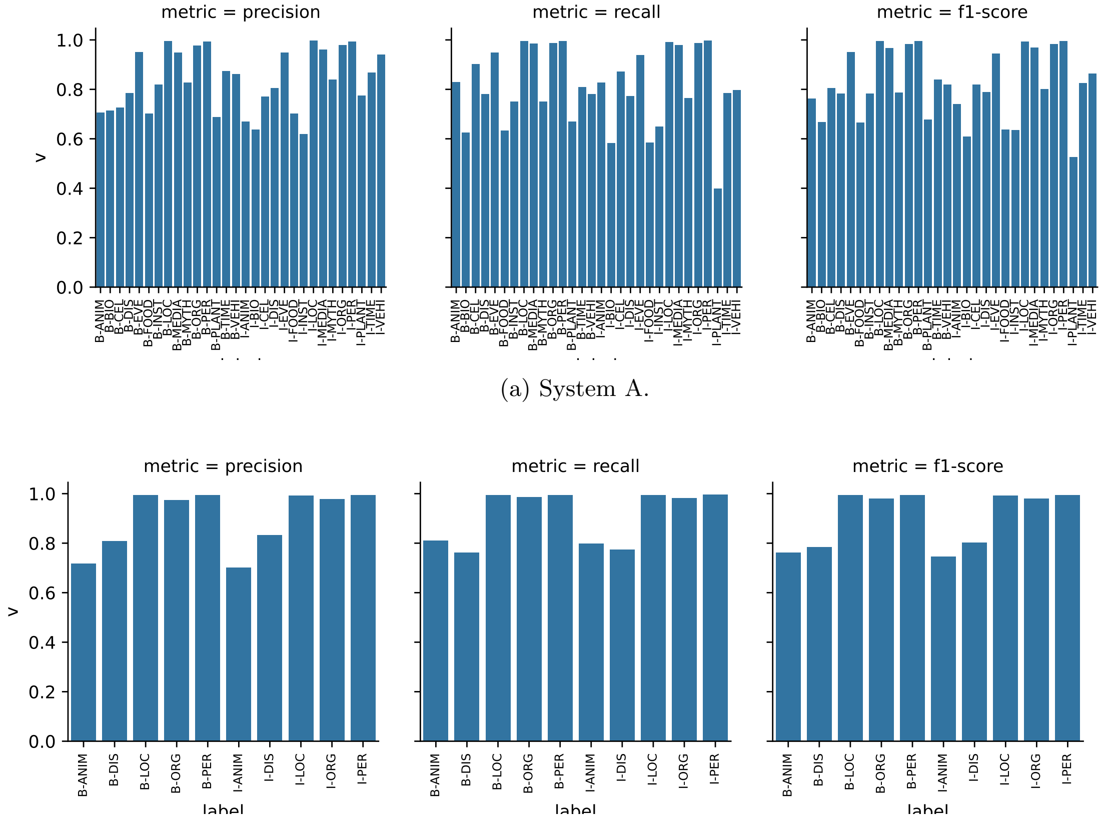
The metrics for each individual BIO-label is shown in Fig. 2. Generally, we can note that if a B-label which have a high score, the corresponding I-label also have a high score.



(a) System A.

(b) System B.

Figure 1: Confusion matrix for predictions on the test set.



(a) System A.

(b) System B.

Figure 2: Precision, Recall and F1-scores for the labels on the test.

2 Limitations

- No taking into account the distribution of the data (which is imbalanced). This can be done by various data augmentation methods, data sampling strategies, or class weights in the loss function (however, this in particular did not work very well in my experiments, it did increased the recall but decreased the precision significantly).

- We evaluating the models ability to predict BIO-structure in addition to the named entity categories. Both of these aspects are of interest, but should be separated in the evaluation. In this report, we consider BIO-labels.

3 Future work

- Explore Named Entity category prediction without the BIO-schema with, for example with a span-based model (Soares et al., 2019).

- Explore how composing BPE-tokens into word representations perform across labels and languages (Ek, Bernardy, 2020).

References

Ek Adam, Bernardy Jean-Philippe. Composing Byte-Pair Encodings for Morphological Sequence Classification // Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020). 2020. 76–86.

Soares Livio Baldini, FitzGerald Nicholas, Ling Jeffrey, Kwiatkowski Tom. Matching the blanks: Distributional similarity for relation learning // arXiv preprint arXiv:1906.03158. 2019.