NER-MultiNERD-English

Test Set	Precision		Recall		F1		Support
	SystemA	SystemB	SystemA	SystemB	SystemA	SystemB	
PER	0.99	0.99	0.99	0.99	0.99	0.99	10530
ORG	0.97	0.96	0.97	0.97	0.97	0.96	6618
LOC	0.99	0.99	0.99	0.99	0.99	0.99	24048
ANIM	0.65	0.66	0.66	0.63	0.65	0.64	3208
DIS	0.71	0.71	0.65	0.66	0.68	0.68	1518
Overall	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**45922**

Table 1: Evaluation of SystemA (left) and SystemB (right) on test set. Mean values over two 5-epoch fine-tuning runs are reported.

Both systems are fine-tuned starting from a *Cased Multilingual DistilBERT* checkpoint. This base model is chosen in order to ensure quick training and evaluation times under limited resources and time. The fact that it is a multilingual model also makes the experiment easily repeatable for other languages.

As an evaluation metric *seqeval* is chosen. *seqeval* does not include the *O* tag while calculating the overall metrics. Thus, this metric allows us to compare system A and system B, which are trained with different datasets, based on common NER category performances. Also, we did not use the “strict” version of the metric which imitates the *conlleval*. Thus, predicting *B-PER* as *I-PER* is accepted as correct for the metric calculations.

According to the results, both system A and system B perform equally well on the common NER categories. This means that decreasing the complexity of the dataset by removing some of the NER categories does not benefit the rest of the categories under these experiment conditions.

Both models have their poorest performances in the ANIM and DIS categories. According to the MultiNERD dataset authors, the annotations for ANIM entities were poor which can explain the confusion of the models.