Named Entity Recognition

RISE Research Institutes of Sweden

Jayant Yadav

1 Findings

1. System A and System B models (M7 and M8 respectively in table 3) show a 5% difference in F1 score on evaluation set. But when models were inferred on test dataset with scores given to exact entity match only, the entity-wise F1 score is negligible (refer table 1 and 2). It can be observed that reducing entities from the training set (ie. system B) did not give us a superior model than the model trained on all entities (ie. system A).

2. Fine-tuning benefited from larger batch size as it is evident from table 3, while moving from model M3 to M4. This can be attributed to the nature of MultiNERD data which has short sentences of 21.7 words (on average) and 1.5 named entities (on average) [1].

3. Special tokens (like $< s >$ and $< /s >$) and tokens obtained from the subsequent splitting of a token by the RoBERTa tokenizer were ignored during training and evaluating the F1 score. This was done by giving them a label of -100 (PyTorch ignores this index while calculating the loss function). This gave a slight improvement in score (Refer table 3. "Label all tokens" was set to False from model M5 to M6 which gives a bump of 1% F1 score) [2].

Entity	Precision	Recall	F1 score	Support
ANIM	0.71	0.77	0.739	1604
BIO	0.5	0.125	0.2	8
CEL	0.738	0.756	0.746	41
DIS	0.737	0.772	0.754	759
EVE	0.952	0.968	0.960	352
FOOD	0.679	0.545	0.605	566
INST	0.75	0.75	0.75	12
LOC	0.994	0.991	0.993	12024
MEDIA	0.940	0.969	0.954	458
ORG	0.977	0.981	0.979	3309
PER	0.992	0.995	0.993	5265
PLANT	0.617	0.730	0.669	894
MYTH	0.647	0.687	0.666	32
TIME	0.825	0.820	0.822	289
VEHI	0.812	0.812	0.812	32
Overall	0.939	0.947	0.943	

Table 1: Precision, Recall and F1 score metric on test dataset for System A

Entity	Precision	Recall	F1 score	Support
ANIM	0.70	0.765	0.731	1604
DIS	0.721	0.778	0.749	759
LOC	0.995	0.993	0.994	12024
ORG	0.976	0.983	0.979	3309
PER	0.992	0.995	0.993	5265
Overall	0.960	0.969	0.964	

Table 2: Precision, Recall and F1 score metric on test dataset for System B

2 Limitations

1. Since MultiNERD is not a gold-standard dataset, there might be a presence of illegal tags which were not checked and/or corrected in this assignment. An illegal tag can be an Inside (I-PER) tag not preceded by Beginning tag (B-PER).

2. No sampling was performed from the training set to balance the distribution of all labels (Refer to Figure 1). This can potentially lead the model to overfit or underfit to some labels.

3. We set the maximum sequence length to 512 even though the average sequence length in MultiNERD is 21.7 [1]. This can cause the model to pad significant portions of input seq with zeros. This is not only computationally inefficient but also irrelevant information to process. This also limits the batch size we can set, as a higher maximum sequence length would mean a smaller batch size that can fit in the memory. Model might also not generalize well on shorter sentences during testing, since it gets overfitted to longer sequences.

4. Since no confusion matrix was generated for our systems A and B, it is hard to tell which entities were being confused with one another. The MultiNERD paper [1] does mention about ANIM and PLANT being confused with FOOD, but we could not confirm that. Further, no Outside ('O') token scores were calculated since the HuggingFace *seqeval* metric package lacks that output [3].

5. Model 1-8 were not hyperpertuned for system B. The same configuration from System A model (M7) was used to fine-tune system B model (M8). A separate hyperparameter tuning might have given better results.

References

[1] Simone Tedeschi and Roberto Navigli. "MultiNERD: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition (and Disambiguation)". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Findings 2022. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 801–812. DOI: 10.18653/v1/2022.findings-naacl.60. URL: https://aclanthology.org/2022.findings-naacl.60 (visited on 11/28/2023).

[2] *notebooks/examples/token_classification.ipynb at main · huggingface/notebooks*. GitHub. URL: https://github.com/huggingface/notebooks/blob/main/examples/token_classification.ipynb (visited on 11/28/2023).

[3] *chakki-works/seqeval at Github*. GitHub. URL: https://github.com/chakki-works/seqeval (visited on 11/28/2023).

A Tables and Figures

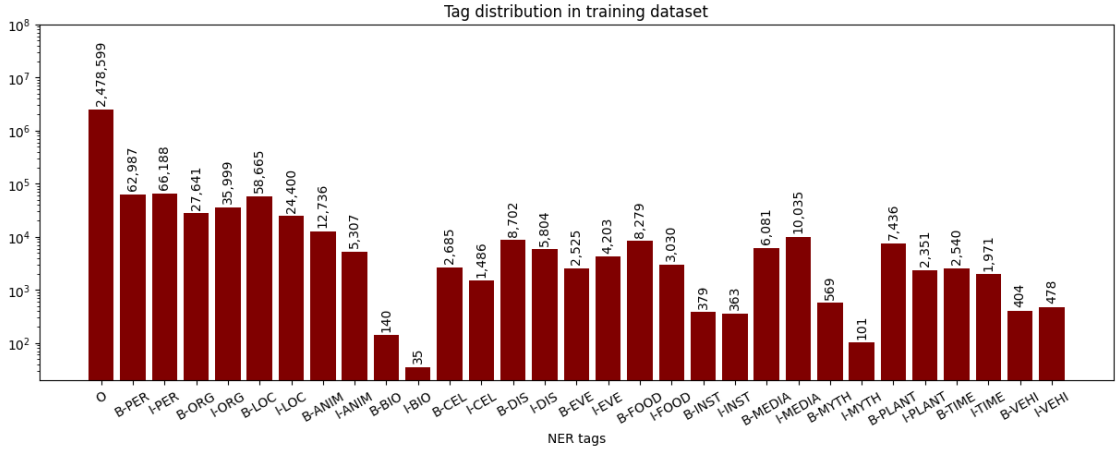

Figure 1: Count of named entities in Training dataset

Model No.	(System) HuggingFace Model	Vocabulary	Label all tokens	Batch \ Size	Learning Rate	Training Set	Epochs	Micro Avg. F1 (Evaluation set)
M1	(A) distilbert-base-uncased	Wikipedia, Bookcorpus	True	16	0.00002	10%	2	0.8792
M2	(A) bert-base-multilingual-cased	Wikipedia	True	16	0.00002	10%	0.55	0.8516
M3	(A) bert-base-multilingual-cased	Wikipedia	True	16	0.00002	10%	1.22	0.8767
M4	(A) bert-base-multilingual-cased	Wikipedia	True	32	0.00005	10%	2	0.8868
M5	(A) roberta-base	Wikipedia, Bookcorpus	True	32	0.00005	10%	1	0.8933
M6	(A) roberta-base	Wikipedia, Bookcorpus	False	32	0.00005	10%	1	0.9018
M7	(A) roberta-base	Wikipedia, Bookcorpus	False	32	0.00005	50%	1	**0.905**
M8	(B) roberta-base	Wikipedia, Bookcorpus	False	32	0.00005	50%	1	**0.9519**

Table 3: Configuration summary of various HuggingFace pretrained models versus micro average F1 score on the evaluation dataset

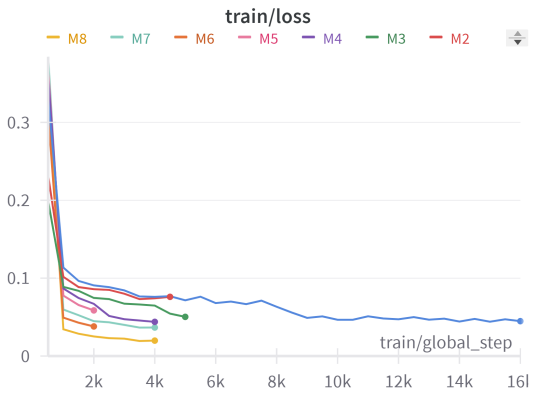

Figure 2: Training loss

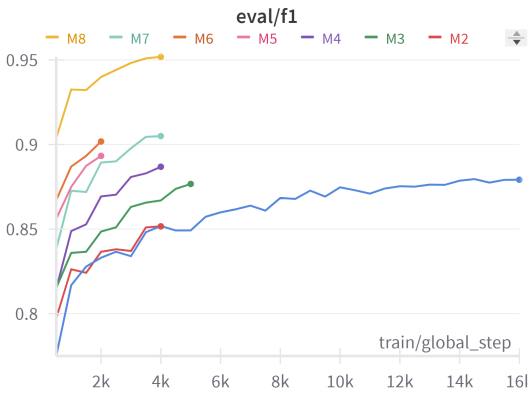

Figure 3: F1 score on validation dataset