Assignment Report

Yini Chen

December 7, 2023

1 Model Setup and Experiments

For the fine-tuning tasks, the 'distilbert-base-uncased' model was used as it is a smaller version of BERT while its performance on many NLP tasks (especially classification tasks) remain excellent. The original MultiNERD dataset was preprocessed before the fine-tuning tasks.System A used only the English data from the dataset. As for system B, the data size is essentially the same as the original dataset, only that in the 'ner_tags", the value of excluded tags are set to zero. However, due to the limited access to resource (GPU on google colab pro), fine-tuning the model on full size data was not feasible for the given time frame, therefore for the experiments, 10% was subtracted from the shuffled original dataset for the training of system B. The data split for each system is demonstrated in Table1.

	train	validation	test
system A	262560	32820	32908
system B	276840	33480	33598

Table 1: data split after preprocessing

Apart from the system specific test set, both systems were also tested on a test set where all ner_tags and all language data were kept. The results in Table 1 indicate that system A suffered greatly from unseen tokens in other languages, while system B had better f1 score performance on the same test set.

	system A		system B	
all tags, all language, sampled 33958 test data	overall_precision	0.450669128	overall_precision	0.837256909
	overall_recall	0.229078899	overall_recall	0.626233413
	overall_f1	0.303756052	overall_f1	0.716531218
	overall_accuracy	0.864217339	overall_accuracy	0.93788818
5 ner_tags, all language, sampled 33958 test data			overall_precision	0.837256909
			overall_recall	0.815623615
			overall_f1	0.826298691
			overall_accuracy	0.973012677
all tags, only English, 32908 test data	overall_precision	0.916794872		
	overall_recall	0.928291404		
	overall_f1	0.922507321		
	overall_accuracy	0.981752246		

Table 2: experiment results

2 Limitations and future work

The experiments all had the same training epochs, learning_rate, and weight_decay. It will probably be worthwhile to explore more hyperparameters and find the optimal setting to achieve better model performance. Furthermore, the current results are presented in overall f1 and accuracy scores, while the seqeval metric supports detailed evaluation on each ner_tag. Considering the distribution of ner_tags in the dataset, a weighted evaluation such as macro f1 would have made more sense if the given time allows.