4. EVALUATION METRICS

Python Script Result Evaluation

		BASIC FEATURES		ADVANCED FEATURES	
		LogReg	CRF	LogReg	CRF
Twitter_dev.ner	Token-wise accuracy	95.5361012395	95.7701308832	95.8568085291	96.0648348791
	Token-wise F1 (macro)	21.5780375334	29.5648858833	26.96900893088	32.4084058773
	Token-wise F1 (micro)	95.5361012395	95.7701308832	95.8568085291	96.0648348791
	Sentence-wise accuracy	66.6101694915	68.6440677966	67.4576271186	69.6610169492
Twitter_dev_test.ner	Token-wise accuracy	91.0152104705	91.3070392642	91.6519278387	91.9083834454
	Token-wise F1 (macro)	10.9195384447	17.9817691763	17.8909073962	21.644445822
	Token-wise F1 (micro	91.0152104705	91.3070392642	91.6519278387	91.9083834453
	Sentence-wise accuracy	48.6486486486	50.4978662873	50.4978662873	52.347083926

Conlleval Script Result Evaluation

		BASIC FEATURES		ADVANCED FEATURES	
		LogReg	CRF	LogReg	CRF
Twitter_dev.ner	Accuracy	95.54	95.77	95.86	96.06
	Precision	49.61	60.61	46.68	59.36
	Recall	16.89	26.81	24.66	34.85
	FB1	25.20	37.17	32.74	43.92
Twitter_dev_test.ner	Accuracy	91.02	91.31	91.65	91.91
	Precision	32.35	46.82	35.22	46.08
	Recall	8.54	15.99	17.39	23.76
	FB1	13.51	23.84	23.38	31.35

The CONLL evaluation script is basically a phrase chunking evaluation system. Both the CONNL and python evaluation are good in certain applications.

The CONNL script is essentially good for benchmarking. It provides cumulative performance measures (such as accuracy, precision, recall, FB1) for all classes of NER. Hence, here a single performance value accounts for it over all classes. Individually for each class it too gives the precision, recall, and f1-score.

However, the python evaluation is good for feature engineering. Here, the cumulative performance measures like token wise accuracy, token wise F1 (micro), token wise F1 (macro) and sentence wise accuracy are got for all classes. For each class too (including the B- and I- tags), the recall, precision and F1 is got. Hence, for a class level detailed analysis, and for finding out the precision, recall and F1 for each possible type of NER tag, this metric seems more apt.

However, both evaluation scripts are apt in their own way based on the application.