* 1. Experiment

Applied CONLL evaluation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | BASIC FEATURES | | ADVANCED FEATURES | |
|  |  | LogReg | CRF | LogReg | CRF |
| Twitter\_dev.ner.pred | Accuracy | 95.54 | 95.77 | 95.76 | 96.07 |
| Precision | 49.61 | 60.61 | 48.9 | 62.50 |
| Recall | 16.89 | 26.81 | 23.86 | 33.51 |
| FB1 | 25.20 | 37.17 | 32.07 | 43.63 |
| Twitter\_dev\_test.ner.pred | Accuracy | 91.02 | 91.31 | 91.50 | 91.71 |
| Precision | 32.35 | 46.82 | 31.89 | 44.51 |
| Recall | 8.54 | 15.99 | 14.91 | 23.29 |
| FB1 | 13.51 | 23.84 | 20.32 | 30.58 |

According to the sheet, the result of advanced is higher both logistic regression and crf method. Compared CRF and logreg, the CRF method has better result and higher FB1 score, that indicate CRF is a better method which can give higher score, due to crf method based on conditional probability and it is a discriminative model. Let’s focus on FB1 score, which is significant to evaluate the features are good or not.

For witter\_dev.ner, it improved 6.8 in average.

For witter\_dev\_test.ner, it improved 6.7 in average.

That mainly due to the advanced one gives model more features and improved model’s behavior.

Applied python script evaluation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | BASIC FEATURES | | ADVANCED FEATURES | |
|  |  | LogReg | CRF | LogReg | CRF |
| Twitter\_dev.ner.pred | Token-wise accuracy | 95.54 | 95.77 | 95.76 | 96.07 |
| Token-wise F1 f(macro) | 21.58 | 29.56 | 23.51 | 29.29 |
| Token-wise F1 (micro) | 95.54 | 95.77 | 95.76 | 96.07 |
| Sentence-wise accuracy | 66.61 | 68.64 | 66.27 | 68.64 |
| Twitter\_dev\_test.ner,pred | Token-wise accuracy | 91.02 | 91.31 | 91.50 | 91.71 |
| Token-wise F1 (macro) | 10.92 | 17.98 | 17.36 | 23.45 |
| Token-wise F1 (micro | 91.02 | 91.31 | 91.50 | 91.71 |
| Sentence-wise accuracy | 48.65 | 50.50 | 49.36 | 52.20 |

According to the python script evaluation, advanced feature has better accuracy, I prefer analyzing F1(micro) because it can analyze the data in general better, macro will calculate F1 score of each class and then do average, which is not useful.