Assignment 3: NER

I used the plain CRF model for the task. I analysed the data manually to bring out features in data. So, basically all the features fed in were driven by intuitions. Initial features like POS tags, presence of @ etc gave F1 scores of the range 0.52 but on inserting the word itself as a feature, the model was boosted up to 0.76-0.77 macro-F1. Then to increase the score further, I added prefixes and suffixes of the original string of length <=3. This basically recognised the general nature of words. I added another feature that is the previous word and the next word in the sentence. Adding the length of the current word and position of the current word as features also helped in predictions. Positions might have served as information to entitity-position relationship i.e generally there is a Subject in the starting words and ends with an object.

The CRF library was available as sklearn_crfsuite libraries that helped ease of with the implementation. Bi-LSTM CRF didn't do well because of less annotated data available. Training embeddings depended a lot on the techniques used to fine tune the embeddings too and that is why results could have been very different. Moreover there were many misspelled words in the dataset due to which a simple feature based approach worked better.

-Atishya Jain
-2016CS50393