A3: NER on Real Estate Data

Applying Bi-LSTM CRF on the data did not give much reward because the number of words which were misspelled or out of vocabulary was too much. Training embeddings would require more data and much more effort. Thus a CRF with simple features, specific to the task were expected to work much better.

I used the CRF that was already present in the sklearn library and put a host of features for recognizing each class, which were related to the shape of the word. This gave a macro F-score of 0.72 only.

However, when the actual word to be predicted was put into the CRF, the macro F suddenly went up to near 0.78. This was because this was capturing all the substring features. On the other hand, once all the features related to word shapes that I had added were removed, I got an F-score near 0.81. Adding other features was hurting the performance.

The only additional features that helped increase performance were adding the substrings (of length <=4) of the current word, and the full words of the forward and the backward word.