Dataset Characterstics:

1. Tags and Their Frequencies

```
PER_Others 9087
PER_Victim 1281
PER_Accused 4338
ORG_Victim 400
ORG_Accused 2958
ORG_Others 11715
LOC_Accused 814
LOC_Others 8958
LOC_Event 6806
LOC_Victim 223
```

Hidden Markov Model Based Named Entity Recognition

Experiment 1: No Feature. 1st order markov order

Hidden states:
Output:
N-gram:
Dealing with Unknowns:
Probability Calculations:

```
<HiddenMarkovModelTagger 10 states and 20252 output symbols>

['ORG_Others', 'PER_Others', 'LOC_Event', 'PER_Victim', 'LOC_Others',
'PER_Accused', 'ORG_Victim', 'ORG_Accused', 'LOC_Accused', 'LOC_Victim']
```

For few documents, Number of tags and words are different.

```
======= Accuracy Class Wise =====================

Class            Matched    Total      %
--------------------------------------------
PER_Others:      156        1988       7.85%
PER_Victim:      21         368        5.71%
PER_Accused:     63         743        8.48%
ORG_Victim:      12         96         12.5%
ORG_Accused:     109        537        20.30%
ORG_Others:      246        2340       10.51%
LOC_Accused:     0          145        0.0%
LOC_Others:      198        1762       11.24%
LOC_Event:       154        1111       13.86%
LOC_Victim:      0 29 0.0%



Average Accuracy: 9.04%
```

CRF

1. Using NLTK

```
======= Accuracy Class Wise ====================

Class            Matched    Total      %
--------------------------
PER_Others: 1257 1702 73.85%
PER_Victim: 28 268 10.45%
PER_Accused: 205 599 34.22%
ORG_Victim: 0 81 0.0%
ORG_Accused: 379 546 69.41%
ORG_Others: 1289 2015 63.97%
LOC_Accused: 0 120 0.0%
LOC_Others: 961 1743 55.13%
LOC_Event: 475 1100 43.18%
LOC_Victim: 0 27 0.0%



Average Accuracy: 35.02%
```

2. Using CRFSuite

```
======= Accuracy Class Wise ====================

Class    Matched Total %
--------------------------
PER_Others: 1278 1669 76.57%
PER_Victim: 58 211 27.48%
PER_Accused: 291 504 57.74%
ORG_Victim: 3 45 6.67%
ORG_Accused: 431 528 81.63%
ORG_Others: 1277 1957 65.25%
LOC_Accused: 7 80 8.75%
LOC_Others: 1000 1625 61.54%
LOC_Event: 517 1168 44.26%
LOC_Victim: 2 41 4.88%

Avg Accuracy = 43.48%
```

3. Adding Entity Score to CRF with word score calculated using word frequency and position:

$$word_{score} = \log_{10}\left(\left(1 + word_{freq} / word_{total}\right) * \left(1 + word_{total} - word_{position}\right) / word_{total}\right)$$

```
======= Accuracy Class Wise ====================

Class    Matched Total %
--------------------------
PER_Others: 1174 1629 72% PER_Victim: 64 229 27%
PER_Accused: 355 884 40% ORG_Victim: 3 57 5%
ORG_Accused: 394 546 72% ORG_Others: 1137 2199 51%
LOC_Accused: 1 160 0% LOC_Others: 868 1640 52%
LOC_Event: 456 1106 41% LOC_Victim: 3 31 9% Avg Accuracy = 36.2%
```

LSTM

With word embedding created using Glove from dataset itself.

With pre treained Golve word embedding taken from Stannford NLP.

BLSTM+Softmax

1. With word embedding created using Glove from dataset itself.

```
======= Accuracy Class Wise =====================

Class    Matched Total %
--------------------------
PER_Others: 658 823 79%
PER_Victim: 49 182 26%
PER_Accused: 182 428 42%
ORG_Victim: 1 34 2%
ORG_Accused: 234 317 73%
ORG_Others: 655 1145 57%
LOC_Accused: 3 75 4%
LOC_Others: 528 907 58%
LOC_Event: 284 680 41%
LOC_Victim: 0 12 0%

Average Accuracy: 38.2%
```

2. With pre treained Golve word embedding taken from Stannford NLP.

```
Class    Matched Total %
--------------------------
PER_Others: 664 823 80%
PER_Victim: 50 182 27%
PER_Accused: 232 428 54%
ORG_Victim: 5 34 14%
ORG_Accused: 245 317 77%
ORG_Others: 679 1145 59%
LOC_Accused: 4 75 5%
LOC_Others: 492 907 54%
LOC_Event: 372 680 54%
LOC_Victim: 0 12 0%

Average Accuracy: 42.4%
```

BLSTM + CRF

1. With word embedding created using Glove from dataset itself.

```
======= Accuracy Class Wise =====================

Class    Matched Total %
--------------------------
PER_Others: 665 823 80%
PER_Victim: 28 182 15%
PER_Accused: 161 428 37%
ORG_Victim: 0 34 0%
```

```
ORG_Accused: 238 317 75%
ORG_Others: 662 1145 57%
LOC_Accused: 3 75 4%
LOC_Others: 550 907 60%
LOC_Event: 234 680 34%
LOC_Victim: 0 12 0%

Average Accuracy: 36.2%
```

2. With pre treained Golve word embedding taken from Stannford NLP.

```
======= Accuracy Class Wise ====================

Class            Matched    Total      %
------------------------------------------
PER_Others:      662        823        80%
PER_Victim:      42         182        23%
PER_Accused:     253        428        59%
ORG_Victim:      3          34         8%
ORG_Accused:     245        317        77%
ORG_Others:      688        1145       60%
LOC_Accused:     1          75         1%
LOC_Others:      489        907        53%
LOC_Event:       371        680        54%
LOC_Victim:      0          12         0%

Average Accuracy: 40%
```

Analysis of Dataset and Results

1. Context Keyword: Context keyword plays an important role in identifying the tag. In our dataset there is a high overlapping of surrounding words for ‘relevant’ and ‘non-relevant’ entities.