# 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:      241        1780       13%
PER_Victim:      8          183        4%
PER_Accused:     121        804        15%
ORG_Victim:      7          51         13%
ORG_Accused:     174        618        28%
ORG_Others:      338        2455       13%
LOC_Accused:     2          167        1%
LOC_Others:      222        1748       12%
LOC_Event:       201        1326       15%
LOC_Victim:      0          60         0%

Average Accuracy: 9%
```

## Experiment 2: Features, 1st order markov order

**CRF**

1. Using NLTK

Feature:

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

Class            Matched    Total      %
--------------------------
PER_Others: 1351 1923 70%
PER_Victim: 36 329 10%
PER_Accused: 260 820 31%
ORG_Victim: 0 55 0%
ORG_Accused: 406 559 72%
ORG_Others: 1215 2490 48%
LOC_Accused: 0 155 0%
   LOC_Others: 1017 1795 56%
LOC_Event: 581 1362 42%
LOC_Victim: 0 54 0%

Average Accuracy: 32.9%
```

2. Using CRFSuite

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

Class    Matched Total %
--------------------------
PER_Others: 1233 1769 69%
PER_Victim: 74 306 24%
PER_Accused: 308 808 38%
ORG_Victim: 0 65 0%
ORG_Accused: 433 572 75%
ORG_Others: 1170 2091 55%
LOC_Accused: 1 163 0%
LOC_Others: 936 1721 54%
LOC_Event: 481 1151 41%
LOC_Victim: 0 20 0%

Avg Accuracy = 35.6%
```

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.