HMM with part-of-speech as features:

```
Confusion Matrix
--------------
      B      O      I
B 11551.0 539.0   827.0
O 425.0   19911.0 684.0
I 297.0   336.0   12976.0

B        precision 0.894248     recall 0.941172          F1 0.917110
O        precision 0.947241     recall 0.957904          F1 0.952543
I        precision 0.953487     recall 0.895700          F1 0.923690
accuracy rate = 0.934632
```

HMM with words themselves as features:

```
Confusion Matrix
--------------
      B      O      I
B 10834.0 383.0   835.0
O 1080.0  20031.0 1150.0
I 359.0   372.0   12502.0

B        precision 0.898938     recall 0.882751          F1 0.890771
O        precision 0.899825     recall 0.963677          F1 0.930657
I        precision 0.944759     recall 0.862981          F1 0.902020
accuracy rate = 0.912106
```

HMM with part-of-speech and words as features:

```
Confusion Matrix
--------------
      I      B      O
I 12685.0 348.0   420.0
B 732.0   11105.0 321.0
O 1070.0  820.0   20045.0

I        precision 0.942912     recall 0.875613          F1 0.908017
B        precision 0.913390     recall 0.904832          F1 0.909091
O        precision 0.913836     recall 0.964351          F1 0.938414
accuracy rate = 0.921949
```

The best model is the one using only the pos tags as features. The reason why it outperforms the other two may because that in the actual test set there are plenty of out-of-vocabulary words that are not in the training feature list, but since part of speech tags are limited, there is few new pos tags in the test set.