Statistical Approaches to NLP
PA 3 – HMM Writeup

José Andrés Molina

December 3, 2016

1 Just POS as Features

1.1 Confusion Matrix

	B	I	O
B	11368.0	841.0	531.0
I	278.0	12764.0	326.0
O	453.0	691.0	19535.0

1.2 Evaluation Metrics

	Precision	Recall	F_1
B	0.892307692308	0.939581783618	0.915334755828
I	0.954817474566	0.892837157247	0.922787738577
O	0.944678175927	0.957973715182	0.951279491612

Accuracy = 0.933314809669

2 Just Words as Features

2.1 Confusion Matrix

	B	I	O
B	10424.0	754.0	494.0
I	442.0	12645.0	615.0
O	1233.0	897.0	19283.0

2.2 Evaluation Metrics

	Precision	Recall	F_1
B	0.893077450308	0.861558806513	0.877035042699
I	0.922857976938	0.884513150532	0.903278805629
O	0.900527716808	0.945615927815	0.922521229518

Accuracy = 0.905208711822

3 Words and POS as Features

3.1 Confusion Matrix

	B	I	O
B	10680.0	643.0	417.0
I	426.0	12826.0	636.0
O	993.0	827.0	19339.0

3.2 Evaluation Metrics

	Precision	Recall	F_1
B	0.909710391823	0.882717579965	0.896010738705
I	0.923531105991	0.897174034695	0.910161793926
O	0.913984592845	0.948362102785	0.930856056413

Accuracy = 0.915745826832

4 Discussion

Of the three experiments performed on the HMM model, interestingly using just the Part of Speech tags as features performed the best at 93%. This is somewhat intuitive given that that the sequence of Part of Speech tags would provide good information as to where the noun phrases are. As expected, using just the words themselves as features performed the worst. They provide the least amount of information on where noun phrases are, and although the model still does well at 90% accuracy.

Unexpectedly, however, using the (Word, POS) tuples as features, performed between the two at 91.6% accuracy. I would've expected the more information and the combination of the two features would help the model and give it a higher accuracy than just POS tags, but this does not seem to be the case. This could possibly because this creates very sparse features that end up over-fitting and not generalizing well. This probably also makes much more use of the "UNK", or OOV feature, as there are probably many Word-POS combinations not modeled; whereas using just POS tags probably covers all possible POS tags that will show up in the test set as they are a much more restricted set.

From the above experiments, running just POS features is the only that passed the test of having all F_1 scores for each label over 90%. The B label in the other two experiments had slightly under that at 87.7% and 89.6% instead.