POS Tagging with HMM and CRF

Kunal Lad : KL28697

Overview

In this homework we explore the performance of Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) for POS tagging task on ATIS and WSJ using mallet implementation. We added a function to compute tokens in training data and are passing this to modified TokenAccuracyEvaluator class so that it can compute % OOV tokens and OOV accuracy for test data.

Experiments

MODEL	ATIS TRAIN	ATIS TEST	ATIS OOV	WSJ TRAIN	WSJ TEST	WSJ OOV
HMM	88.84	86.58	25.42	85.68	77.51	36.68
CRF (1 section)	99.81	92.61	27.17	99.51	79.77	46.74
CRF* (1 section)	99.87	94.19	51.21	99.54	88.45	78.76
CRF (2 sections)	-	-	-	99.44	85.54	49.66
CRF* (2 sections)	-	-	-	99.42	91.56	79.04

TABLE 1 : Performance Comparison of HMM and CRF on ATIS and WSJ

• How does the overall test accuracy of CRF and HMM differ (when using only tokens) ?
Results from table1 show that CRF outperforms HMM on both ATIS and WSJ test datasets. Reason for this is that HMMs are generative models which model full joint probability distribution whereas CRFs are discriminative model which maximizes log likelihood of the POS and are thus designed specifically for this task.

• How does the test accuracy for OOV items for CRF and HMM differ (when using only tokens)?
Again results from table1 show that CRF performs better than HMM on both ATIS and WSJ for OOV tokens but difference is more significant especially for CRF with orthographic features which gives 78.76% accuracy. The reason for this is that on encountering OOV token CRF use conditional probability based on other states but HMM some kind of smoothing to estimate the probability.

• How does the training accuracy of HMM and CRF differ and why?
We observe that training accuracy of CRF is significantly higher than HMM. This is because CRFs perform a complex optimization for a large number of parameter to maximize the log likelihood of label on the training set. HMM on other hand tries to model full joint distribution and with fewer parameters.

• How does the run time of HMM and CRF differ and why?

ITERATIONS	10	20	30	40	50	60
HMM	86.56 | 2.4s	86.56 | 2.42s	86.56 | 2.6s	86.44 | 2.71s	86.44 | 3.4s	86.44 | 3.74s
CRF (1 section)	82.82 | 16.77s	91.58 | 27.61s	92.87 | 36.86s	92.99 | 47.95s	92.99 | 57.01s	92.99 | 57.75s
CRF* (1 section)	85.51 | 16.77s	93.10 | 27.86s	93.57 | 38.48s	93.57 | 50.34s	93.80 | 60.71s	93.80 | 61.57s

TABLE 2 : Accuracy and Time vs No of Iterations on ATIS

As can be seen from table2 running time for CRF is orders of magnitude more than HMM for same number of iterations. The main reason for this is that HMM model joint distribution by performing simple aggregation of counts over the data using dynamic programming techniques like Viterbi whereas CRF

performs complex optimisation problem and learns in each iteration. Thus we see increase in CRF accuracy as number of iterations increases but HMM accuracy remain almost same. although results are reported on ATIS we observe similar pattern for WSJ. The running time for HMM on WSJ with 120 iterations is 1min:15s whereas CRF takes 1h:35min CRF* takes 2h:39min.

• How does adding orthographic features affect the accuracy (both overall and OOV) and runtime of the CRF and why?

Table 1 shows that adding orthographic features (number, hypen, caps, plural and 16 common english suffixes taken from here) gives a significant improvement in test accuracy and even more so for OOV tokens. Reason for this is that CRF with orthographic features is stronger model than CRF with extra states corresponding to orthographic features. Due to higher number of states the model is able to capture more information for OOV tokens and hence gives a huge boost for OOV. As is expected adding orthographic features increases the running time due to higher complexity of the model.

• Which features helped the most?

MODEL	WSJ TRAIN	WSJ TEST	WSJ OOV
CRF (Basic)	99.48	85.77	69.94
CRF (Basic + Suffix)	99.54	88.45	78.76

TABLE 3 : Accuracy for CRF with basic and all Orthographic features on WSJ

We performed an experiment in which we compared accuracy for CRF with basic orthographic features (caps, numbers, hypen and plural i.e. ends with s) and CRF with basic + top 16 common english suffixes. As expected Basic + Suffix model performs better than Basic since it is more powerful. We see that basic features has reasonable overall test accuracy (85.77 compared to 88.45 % for basic + suffix) but for OOV basic + suffix model gives a significant boost from 69.94 to 78.76 % This shows that suffixes capture a lot of information about the OOV tokens.

• How does performance vary with train-test split ?

MODEL	HMM	CRF (1 section)	CRF* (1 section)	% OOV
0.7 - 0.3	84.53	91.76	92.69	3.41
0.8 - 0.2	86.44	92.99	93.80	2.80
0.9 - 0.1	87.71	93.79	95.70	2.86

TABLE 4 : Overall Accuracy vs Proportion of Training Data on ATIS

On Atis (Table 4) We observe that overall accuracy increases with increase in training data and % OOV tokens is higher for 0.7 - 0.3 split due to less training data. We also performed similar experiment on WSJ where we compare results of training on 1 section and 2 sections (Table 1). Results on WSJ show similar pattern with higher accuracy for both CRF and CRF* when trained on 2 sections.

NOTE : Due to lack of space we only report results of some experiments in this report. Detailed results can be found in this sheet.