CS 388: Natural Language Processing
Homework 2: Part-of-Speech Tagging with HMMs and CRFs

Manu Agarwal

March 5, 2016

Introduction

Part-of-Speech Tagging is the process of associating every word in a text (corpus) with a part-of-speech based on the definition as well as the context of the word. Developing a deterministic procedure for this task is hard as some words can represent more than one part of speech at different times. Several approaches have been proposed to do POS tagging such as Hidden Markov Models (HMMs), dynamic programming models, Baum-Welch algorithm etc. In this assignment, we explore the performance of HMMs and Conditional Rnadom Fields (CRFs) on the POS tagging task.

HMMS and CRFs

Hidden Markov Models are probabilistic generative models for sequences. They take as input a set of observed symbols (tokens of language in this case) and infer a set of latent states (parts of speech in this case) that produced that symbol. Probabilistic transitions are assumed from one state to another over time. They model the full joint distribution of observations and labels and use Bayesian inference to determine the most likely set of states given the symbols.

Conditional Random Fields, on the other hand, are discriminative undirected probabilistic models that model conditional distribution of labels given a sequence of tokens. They are specifically designed and trained to maximize performance of classification. Because of their design, CRFs generally perform better on classification than generative models when given a reasonable amount of training data.

Table 1 and Table 2 show a detailed comparison of the performance of HMM and CRF using Mallet's implementation. As is evident from the table, the overall test accuracy of CRF is significantly higher than that of HMM. This is understandable as CRFs are specifically designed for classification tasks as opposed to HMMs that can handle a variety of tasks. Thus, CRFs are more likely to perform better on the task of POS tagging which requires conditional probability estimation.

If we compare the test accuracy of OOV items, the difference between the performance of CRF and HMM becomes all the more significant. CRF performs much better on this task than HMM. This is again attributable to the fact that CRFs are designed to maximize performance of classification. One important point to observe is that CRF using orthographic features performs exceptionally well on OOV items on both ATIS and WSJ as compared to both CRF without using orthographic features as well as HMM. In fact, on WSJ, the OOV accuracy goes as high as 0.8. For the purposes of this experiment, I used 20 most common English suffixes along with three orthographic features - caps, number and hyphen. This shows that such extra features play a key role particularly in predicting the POS for unseen words. We further tested what extra features help the most. **??** shows the comparison between the performance of CRF with orthographic features such as caps, num and hyphen vs CRF with morphological features such as suffixes. We see that morphological features slightly outperform orthographical features. However, this comparison might be slightly biased towards orthographical features as we took 20 suffixes as opposed to three orthographical features. Adding orthographic features also led to approximately 25% increase in time.

The training accuracy of CRF is better than that of HMM. However, if we observe deeply, the difference between the training and testing accuracy of CRF is more pronounced than the difference between the training and testing accuracy of HMM. This suggests that HMM is more robust to overfitting than CRF.

The run time of HMM is remarkably better than that of CRF. This is due to the fact that CRF performs multiple optimizations during the training phase to infer its parameters. However, HMM does not have to infer as many paramters and trains almost 30-40 times faster than the CRF.

We also tried changing the number of iterations on both the CRF and HMM. There was no marked difference in the performance of HMM with increase in number of iterations. However, CRF saw an increase in the performance with the rate of growth declining as we increased the number of iterations.

We also tried training both the HMM and CRF on two sections of WSJ and tested on two other sections. Both the training and testing accuracies increased by almost 0.03 points when we included training on two other sections on HMM. For CRF, the training accuracy was already 0.996 which did not see much of an improvement. However, the testing accuracy saw a jump of almost 0.06 points. We also tried training HMM on half the sections of WSJ and testing on the other half. This saw an increase of almost 0.05 points in both training and testing accuracy. Due to lack of space, we have not put tables for variation of number of iterations and training on more sections of WSJ.

Model	Corpus	Training acc.	Testing acc.	OOV acc.	% OOV tokens	time
HMM	ATIS	0.888	0.866	0.218	0.028	2.95
CRF	ATIS	0.999	0.926	0.257	0.028	56.08
CRF*	ATIS	0.998	.931	0.498	0.028	62.98
HMM	WSJ	0.857	0.775	0.402	0.16	81
CRF	WSJ	0.995	0.797	0.501	0.16	4416
CRF*	WSJ	0.996	0.906	0.8	0.16	5562

Table 1: Comparison of training, testing and OOV accuracies of HMM and CRF on ATIS and WSJ. This table has entries for ATIS corresponding to training-proportion of 0.8 and test-proportion of 0.2. For WSJ, this table contains entries corresponding to training on section 00 and testing on section 01. For ATIS, results have been averaged over 10 runs using random seeds.

Model	Train acc (0.7)	Test acc (0.7)	OOV acc (0.7)	% OOV tokens	Train acc (0.9)	Test acc (0.9)	OOV acc (0.9)	% OOV tokens
HMM	0.881	0.857	0.200	0.034	0.894	0.868	0.232	0.028
CRF	0.999	0.922	0.273	0.034	0.999	0.928	0.265	0.028
CRF*	0.998	.926	0.386	0.034	0.998	0.952	0.58	0.028

Table 2: Comparison of training, testing and OOV accuracies of HMM and CRF on ATIS. Note: Train acc (0.7) means training-proportion of 0.7 has been used.

Model	Train accuracy	Test accuracy	OOV accuracy	Time (in sec)
CRF^1	0.996	0.859	0.6	5540
CRF^2	0.996	0.881	0.696	5583

Table 3: Comparison of training, testing and OOV accuracies of CRF on WSJ using different orthographic features. Note: CRF^1 denotes using caps, num and hyphen. CRF^2 denotes using 20 most common English suffixes.