Gaurav Nanda	Homework #2
Natural Language Processing	March 1, 2014

1 Problem Statement

There exist probabilistic sequence models that allow integrating uncertainty over multiple, inter-dependent classifications and collectively determine the most likely global assignment of Part of Speech Tags (POS). Two such standard models are Hidden Markov Model (HMM) and Conditional Random Field (CRF).

Our task is to evaluate HMM and CRF models on various parameters including training accuracy, testing accuracy, runtime, accurate tag estimation for Out of Vocabolary(OOV) Words and addition of orthographic features. Both of these models are to be evaluated on Airline Travel Information Service (ATIS) and Wall Street Journal (WSJ) data sets.

2 Testing Setup

2.1 Mallet Format Input

Two models have been compared using Mallet's CRF and HMM implementations. To convert Pen TreeBank's raw data to Mallet format, I have implemented RawToMallet.java. For adding orthographic features to tagged output file, one can pass a feature file as input with suitable flags to include or exclude a particular feature.

```
~/:>cat features
CAPS  true
SUFFIX false
PREFIX true
HYPHEN false
START_NUMBER false

~/:>java RawToMallet atis3.pos feature_mallet_files/atis3_caps_prefix.mallet ./features
```

2.2 Handling OOV Tokens

To estimate the count and testing accuracy of OOV Tokens, "evaluateInstanceList" function in "TokenAccuracyEvaluator.java" has been modified. During first iteration, all the training tokens are stored in a HashMap and all the test tokens are looked up in the Map to estimate the count and accuracy the OOV tokens.

3 Experiments

Various experiments have been performed to compare two models on various attributes. All the numbers computed for ATIS have been averaged over 10 random seeds. ATIS is using 80% of the input data for training and WSJ is using 50% data for training. WSJ (00 and 01) signifies data from 00 and 01 sections of WSJ. Similarly, WSJ (00, 01, 02 and 03) signifies data from 00, 01, 02 and 03 sections of the WSJ data set.

3.1 Overall Test Accuracy

Model	HMM	CRF	Total TokenCount
ATIS	0.8608	0.9216	3283
WSJ (00 and 01)	0.7859	0.8066	73576
WSJ (00, 01, 02 and 03)	0.8324	0.8423	155513

Table 1: Overall Test Accuracy HMM vs CRF.

Analysis

1. CRF model performs better than HMM model.

In general, generative models have to make strict independence assumptions on observations to achive tractability which reduces performance. For HMM model, this independence assumption is relaxed by arranging the output variables in a linear chain. But still to come up with a tractable model, it assumes that state depends only on its immediate predecessor, that is, each state y_t is independent of all its ancestors y_1, y_2, . . . , y_{t-2} given its previous state y_{t-1}. Second, an HMM assumes that each observation variable x_t depends only on the current state y_t. CRF does not make such assumptions and is a conditional probabilistic model.

Therefore, because of these assumptions, HMM does not perform as good as CRF model.

2. For WSJ corpus, testing accuracy keeps on increasing as we keep on increasing token count. With the addition of more data, models are better trained and perform better for test data.

3. In the case of ATIS, data is restricted to a limited domain, hence not diverse. Therefore, testing accuracy is high for both HMM and CRF. I performed another experiment, where even if we use only 20% of the data for training and rest for testing, we still get accuracy as high as 0.863.

3.2 Test accuracy for OOV items

Analysis

As expected, test accuracy for OOV is very less in comparison to the seen tokens. Because these tokens were not present during training, their observation probabilities would be very less (Primarily

Model	HMM	CRF	OOV Count	Testing Token Count	Percentage OOV
ATIS	0.2680	0.3058	25	641	3.9%
WSJ (00 and 01)	0.3819	0.4760	5735	37465	15.3%
WSJ (00, 01, 02 and 03)	0.3938	0.5017	9340	81939	11.3%

Table 2: OOV Test Accuracy HMM vs CRF.

dependent upon smoothing techniques used). However, CRF being a discriminative model still performs better than the HMM.

3.3 Training Accuracy

Model	HMM	CRF	Total TokenCount
ATIS	0.8886	0.9990	3283
WSJ (00 and 01)	0.8627	0.9953	73576
WSJ (00, 01, 02 and 03)	0.8877	0.9936	155513

Table 3: Training Accuracy HMM vs CRF.

Analysis

Training accuracy for CRF is much higher than HMM models. This is expected because CRF model uses L-BFGS optimization procedure to maximize the conditional log likelihood of the supervised training data. At the same time, CRF is more likely to overfit the training data also. This is the reason that training accuracy for ATIS is very high.

3.4 Run Time

There is no restriction on the number of iterations for these results (By default, CRF tagger runs for 500 iterations). However, CRF generally converged between 120-160 iterations.

Model	HMM (sec)	CRF (sec)	Training Tokens
ATIS	5	57	2642
WSJ (00 and 01)	167	13100	36111
WSJ (00, 01, 02 and 03)	609	41150	73574

Table 4: Run Time HMM vs CRF.

Analysis

CRF models has to perform complex optimizations to calculate its parameters. Therefore, it takes a very large of time to train CRF models. As the data keeps on growing , CRF model does not

scale well. As we doubled the size of WSJ training data, run time increased four times (However, this is also dependent upon the machine resources available at that time. But still gives a fair idea of increase in run time).

3.5 Adding Orthographic features

I have introduced 5 different types of orthographic features. Words that start with CAPS, list of top 80% Suffixes ["*s*","*ed*","*ing*","*ly*","*er*","*or*","*ion*","*ble*"] and Prefixis ["*un*","*re*","*in*","*im*","*dis*", "*en*","*non*","*over*","*mis*","*sub*"] (http://www.darke.k12.oh.us/curriculum/la/prefixes.pdf), words having hypen and words that start with a number.

Features	Training Accuracy	Testing Accuracy	OOV-Accuracy	RunTime (sec)
CAPS	0.9988	0.9377	0.3209	96
Hyphen	0.9987	0.9301	0.2519	106
Start Number	0.9987	0.9301	0.2519	98
Prefix	0.9988	0.9311	0.2726	106
Suffix	0.9988	0.9345	0.3198	109
All Except Suffix	0.9991	0.9452	0.4074	110
All Except CAPS	0.9991	0.9337	0.2962	112
All	**0.9988**	**0.9433**	**0.4164**	**92**
NONE	**0.9990**	**0.9216**	**0.3058**	**57**

Table 5: [ATIS] Effects of orthographic features on Overall Training Accuracy, Overall Testing Accuracy, OOV Accuracy and Runtime.

Features	Training Accuracy	Testing Accuracy	OOV-Accuracy	RunTime (sec)
CAPS	0.9953	0.7372	0.5283	16405
Hyphen	0.9954	0.8084	0.4886	18150
Start Number	0.9922	0.8030	0.4711	10009
Prefix	0.9952	0.8050	0.4781	19966
Suffix	**0.9954**	**0.8570**	**0.6668**	**17142**
All Except Suffix	0.9949	0.8332	0.5870	14498
All	**0.9949**	**0.8861**	**0.7819**	**16726**
NONE	**0.9953**	**0.8066**	**0.4760**	**13100**

Table 6: [WSJ] Effects of orthographic features on Overall Training Accuracy, Overall Testing Accuracy, OOV Accuracy and Runtime. Trained on WSJ_00 and Tested on WSJ_01.

Analysis

Training Accuracy

For Both ATIS and WSJ, there is hardly an change with the increase of orthographic features. Even without new features, training accuracy is more than 99.5%.

Testing Accuracy

For Both ATIS and WSJ, with addition of new features, testing accuracy increased considerably. This is major advantage of the CRF that we can easily add new features to the model and increase its testing accuracy.

To determine which feature contributed most to the increase in Accuracy, tests were run by including each of the feature separately. For ATIS, CAPS and Suffix contributed the most. For WSJ, Suffix contributed the most.

OOV Accuracy

With the increase of features, OOV Accuracy increased by a large number. For ATIS, CAPS contributed most to the OOV Accuracy increase and Suffix contributed most for WSJ.

RunTime

For a fixed number of iterations, with increase in the number of features, run time goes up.

We have set iteration count to 500 by default. With additional features, CRF seems to converge in lesser number of iterations. For instance for WSJ with all features it took 121 iterations to converge and without addition of any feature it took 150 iterations to complete. Therefore, increase in run time due to addition of features is somewhat compensated by lesser number of iterations.

3.6 Accuracy with Iterations

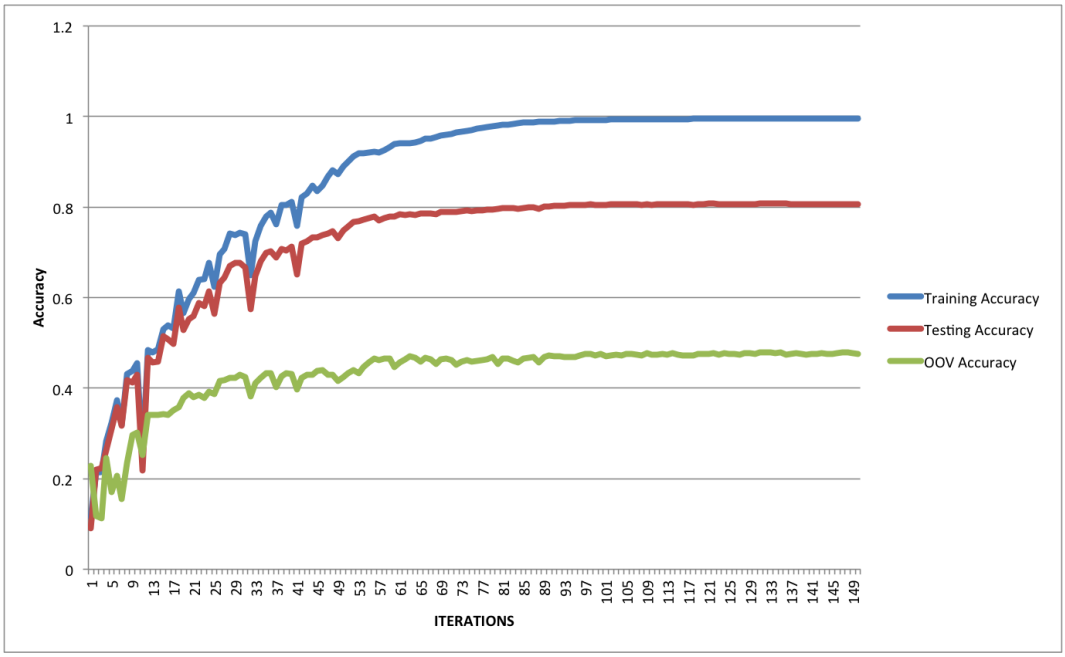

Figure 1: Accuracy with Iterations. [CRF with WSJ (00 and 01)]

As expected for CRFs, accuracy seem to increase with addition in the number of iterations. However, percentage increase after 100 iterations is very less.

For HMMs, number of iterations hardly make any change to accuracy. On the contrary, testing accuracy for all tokens and OOV tokens is maximum with only one iteration.