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

1 Problem Statement

There exist probabilistic sequence models that allow integrating uncertainty over multiple, interdependent 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 ATIS ans WSJ data sets from the Penn Treebank.

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 ten random seeds. ATIS is using 80% of data for training and WSJ is using 50% data for training.

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 very less diverese. Therefore, testing accuracy is high for both HMM and CRF. Even if we use only 20% of the data in training, we still get accuracy as high as 0.863.

3.2 Test accuracy for OOV items

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

Table 2: OOV Test Accuracy HMM vs CRF.

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, 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 RunTime

There is no restriction on the number of iterations for these results. However, CRF generally convegerd 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: RunTime 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.

3.5 Adding Orthographic features

I have introduced 5 different types of orthographic features. Words that start with CAPS, List of top 80% Suffixes and Prefixis (http://www.darke.k12.oh.us/curriculum/la/prefixes.pdf), words having hypen and tokens that start with a number.

Features	TrA	TsA	OOV-TsA	OOV-TsA	RunTime
CAPS					
Hyphen					
Start Number					
Prefix					
Suffix					
All Except Suffix					
All					

Table 5: [ATIS] Effects of orthographic features on Overall Training Accuracy (TrA), Overall Testing Accuracy (TsA), OOV Training Accuracy(OOV-TrA), OOV Testing Accuracy(OOV-TsA)

Features	TrA	TsA	OOV-TsA	OOV-TsA	RunTime
CAPS					
Hyphen					
Start Number					
Prefix					
Suffix					
All Except Suffix					
All					

Table 6: [WSJ] Effects of orthographic features on Overall Training Accuracy (TrA), Overall Testing Accuracy (TsA), OOV Training Accuracy(OOV-TrA), OOV Testing Accuracy(OOV-TsA)

Analysis

3.6 Accuracy with Iterations

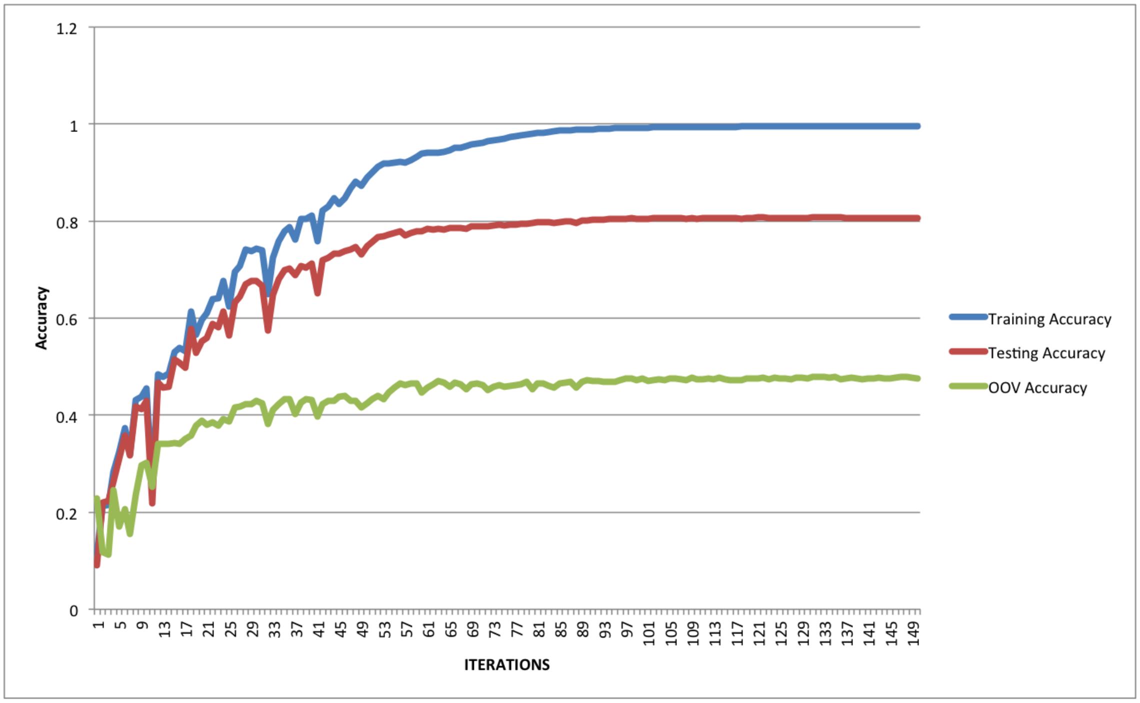

Figure 1: Accuracy with Iterations. [CRF with WSJ (00 and 01)]