CS388 Natural Language Processing
Homework 2: Part-of-Speech Tagging with HMMs and CRFs

Jianyu Huang

March 7, 2015

1 Introduction

We would like to explore the performance of Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) on the Part-of-Speech (POS) tagging task, based on some real-world data from the Penn Treebank[2]. I adopt the implementation of HMM and CRF from the Mallet (Machine Learning for LanguagE Toolkit) package[3] as our code base.

The evaluation part in Mallet provides the overall training and testing accuracy after sequence labeling with HMM/CRF. Beyond that, I change the training and testing code in Mallet to also measure accuracy specifically for out-of-vocabulary items (OOV) [2].

Finally, I compare the performance of CRF model on the data with extra orthographic features, including capitalizations, hyphens, common English suffixes, etc.

2 Implementation

1. In the first step I covert the raw data from the Penn Treebank format to the Mallet format. I modify the POSTaggerFile.java in the first assignment to do the preprocessing by splitting all valid tokens and replacing every backslash with a space.

2. To measure accuracy specifically for out-of-vocabulary (OOV) items, I record the number of all seen training instances in TokenAccuracyEvaluator class with HashSet during the first training iteration and reference it during the testing iterations. To be specific, the method evaluateInstanceList is modified to count the total number of OOV items in training phase and the number of correct OOV items in testing phase.

3. To add extra orthographic features to the CRF model, I implement POSextraFeatures.java to detect seven most common features in the preprocessing stage. The features I added are list in Table 1. (Noun suffixes, verb suffixes and adjective suffixes are from [1])

Label	Feature
gerund	end with ′ing′
plural	end with ′s′
hythen	contains ′-′
caps	start with upper case characters
noun	end with ′-acy, -al, -ance, -ence, -dom, -er, -or, -ism, -ist, -ity, -ty, -ment, -ness, -ship, -sion, -tion′
verb	end with ′-ate, -en, -ify, -fy, -ize, -ise′
adj	end with ′-able, -ible, -al, -esque, -ful, -ic, -ical, -ious, -ous, -ish, -ive, -less, -y′

Table 1: Extra Orthographic features in CRF modeling

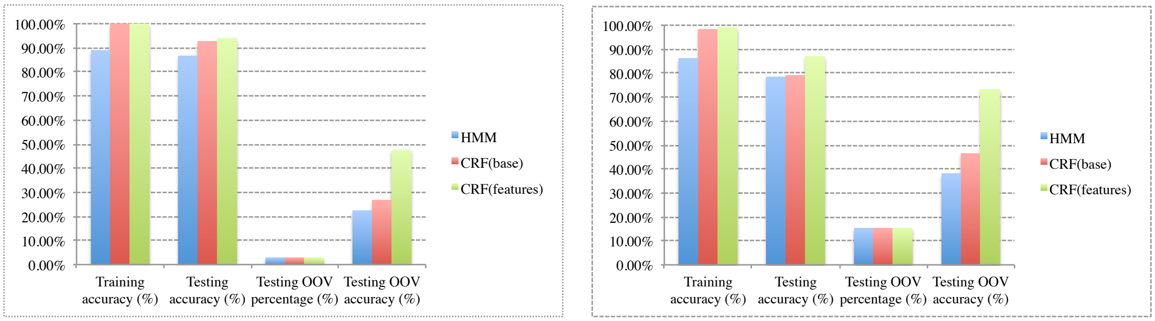
4. Two datasets are used for the experiment. In the small corpus *atis*, we train on 80% of the data and test on the remaining 20%, and we average the results over 10 random training/test splits. In the large corpus *wsj*, we use section 00 for training and section 01 for testing.

3 Experiment

1. The results for HMM, CRF using only tokens and CRF with extra orthographic features on *atis* and *atis* corpora are shown in Table 2 and Figure 3.

	HMM		CRF		CRF with extra features	
	atis	wsj	atis	wsj	atis	wsj
Training accuracy (%)	88.85%	86.18%	99.88%	98.36%	99.88%	99.19%
Testing accuracy (%)	86.59%	78.54%	92.62%	79.16%	94.00%	86.93%
OOV percentage (%)	2.94%	15.31%	2.94%	15.33%	2.94%	15.33%
Testing OOV accuracy (%)	22.41%	38.03%	26.77%	46.53%	47.51%	73.30%
Running time(s)	5.239	78	94.701	7740	92.405	6584

Table 2: Results of HMM, CRF using only tokens and CRF with extra orthographic features



(a) atis

(b) wsj

Figure 1: Results of HMM, CRF(base) and CRF(features) on *atis* and *wsj* dataset

2. The results for adding extra orthographic features to the data are shown in Table 3 and Figure 2.

		raw	gerund	caps	plural	hyphen	noun	verb	adj	All
atis	Training accuracy (%)	99.88%	99.83%	99.83%	99.83%	99.83%	99.83%	99.83%	99.83%	99.88%
	Testing accuracy (%)	92.62%	93.11%	92.99%	93.57%	92.99%	93.22%	93.11%	92.99%	94.00%
	OOV percentage (%)	2.94%	2.80%	2.80%	2.80%	2.80%	2.80%	2.80%	2.80%	2.94%
	Testing OOV accuracy (%)	26.77%	37.50%	37.50%	41.67%	33.33%	45.83%	37.50%	33.33%	47.51%
	Running time(s)	94.7008	97.78	93.867	99.918	101.494	101.79	100.146	87.959	92.4047
wsj	Training accuracy (%)	98.36%	98.61%	99.14%	98.35%	98.61%	98.63%	98.62%	98.62%	99.19%
	Testing accuracy (%)	79.16%	80.44%	82.57%	81.93%	79.98%	80.20%	79.66%	79.84%	86.93%
	OOV percentage (%)	15.33%	15.33%	15.33%	15.33%	15.33%	15.33%	15.33%	15.33%	15.33%
	Testing OOV accuracy (%)	46.53%	50.76%	56.36%	57.77%	48.89%	50.06%	47.66%	47.94%	73.30%
	Running time(s)	7740	8202	12390	7645	8044	6499	6876	7394	6584

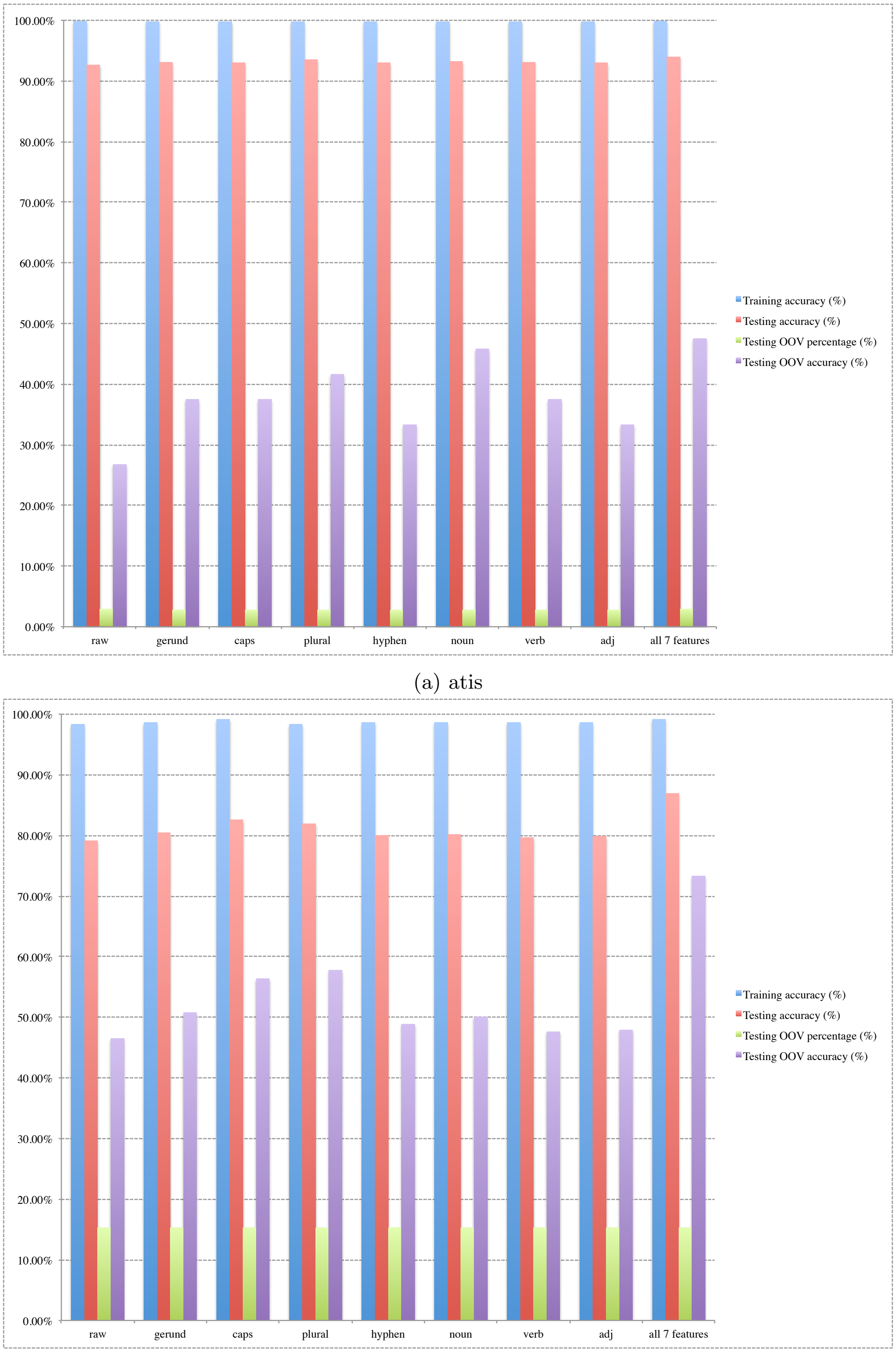
Table 3: Results of adding extra orthographic features to CRF

3. In addition to the basic experiment, I also finish some extension experiments, which is included in the Appendix section.

4 Discussion

4.1 Comparison between HMM and CRF

1. For both dataset, the training accuracy, the overall testing accuracy, the OOV testing of CRF are all greater than those of HMM. There are several reasons:



(a) atis

(b) wsj

Figure 2: Results of adding extra orthographic features to CRF model on *atis* and *wsj* dataset

- The discriminative models are more accurate than generative models at sequence labeling.
- CRF is modeled by maximizing conditional likelihood $p(y|x)$ while HMM is simply trained by maximizing likelihood of data $p(x, y)$.
- The feature dependency is considered by feature weighs in CRF, while the features are assumed independent in HMM.

2. For both dataset, the run time of CRF is much longer than that of HMM (roughly 20x for *atis*, and 80x for *wsj*).
    - In each iteration CRF takes more time than HMM
        - CRF models the dependencies between each state, so the optimization procedure in each iteration will take more time.
        - The sequence tagging results are inference of both training and testing corpora in each iteration in CRF, while that is done only after convergence in HMM.
    - CRF takes more iterations to converge than HMM.
        - The convergence speed in CRF is slower than HMM.

4.2 Analysis of Adding Orthographic Features

1. For both dataset, adding orthographic features increase the accuracy (both overall and OOV) and decrease running time.
    - Orthographic features provide valuable information on OOV words for sequence tagging so that the accuracy is increased.
    - Orthographic features provide additional information to help the optimization converge faster so that the iteration times decreased. Thus finally the run time is decreased.
2. From the comparison in Figure 2, we can see that *noun* is the most useful feature for improving the performance on *atis* dataset, while *plural* feature helps the most important feature on *wsj* dataset.

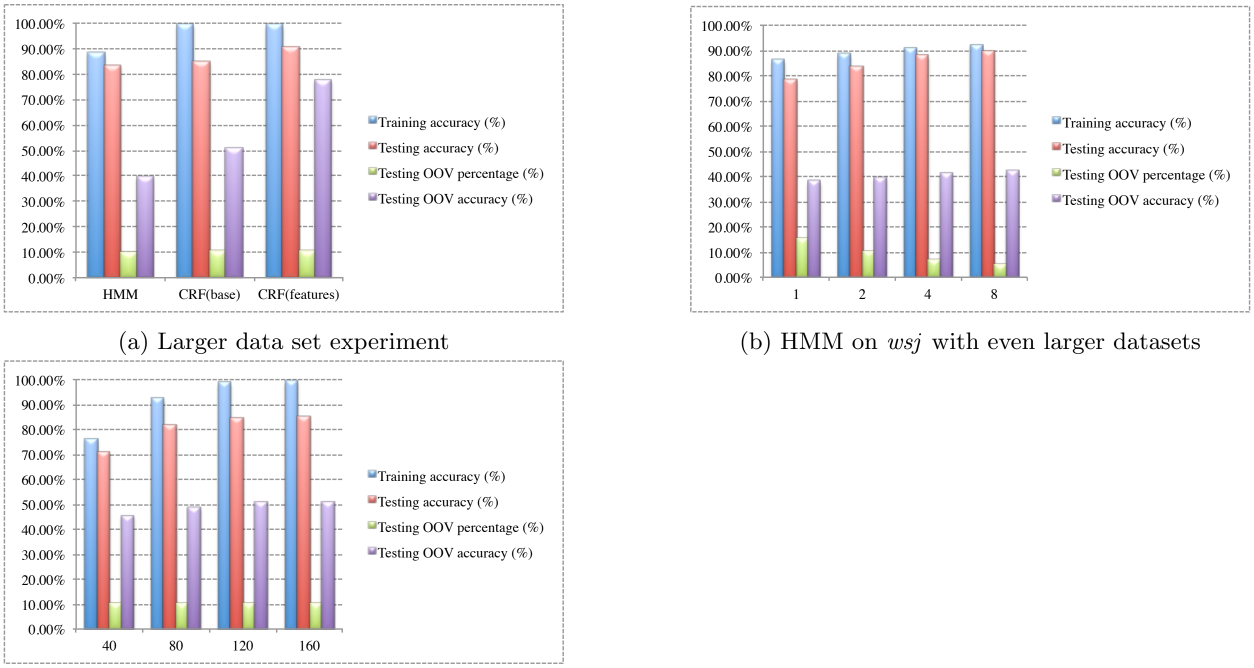
5 Conclusion

Through this assignment, We explore the performance of Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) on the Part-of-Speech (POS) tagging task. It turns out that CRF is more accurate than HMM for POS tagging, but cost more time.

6 Apendix

6.1 Extension Experiments

1. Larger data set experiment
   I train on section 02, 03 and test on section 04, 05 for *wsj* dataset.
   Observation: with larger training set, the result is more accurate. Still, CRF with extra orthographic features has the best performance.
2. HMM on *wsj* with even larger datasets
   I test HMM over even larger data set (the *x-axis* represents how many sections I use to train and test).
   Observation: accuracy increases and OOV percentage decreases over larger data.
3. Change the number of iteration
   I sample the interation number for the first extension experiment, CRF using only tokens over *wsj* dataset (the *x-axis* indicates the number of iterations).
   Observation: more iteration means more accuracy and the improvement in early iterations is larger than that in later iterations.



(a) Larger data set experiment

(b) HMM on *wsj* with even larger datasets

(c) Change the number of iteration

Figure 3: Extension Experiments

6.2 Testing result

Two datasets are used for the experiment. In the small corpus *atis*, we train on 80% of the data and test on the remaining 20%, and we average the results over 10 random training/test splits. In the large corpus *wsj*, we use section 00 for training and section 01 for testing. All the results are stored in "hw2.xlsx" file.

7 Reference

[1] Common suffixes. *http://grammar.about.com/od/words/a/comsuffixes.htm.*

[2] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[3] Andrew K McCallum. Mallet: A machine learning for language toolkit. 2002.