Homework 2 - Part-of-Speech Tagging with HMMs and CRFs

Abhishek Sinha

I. Problem Statement

The objective is to generate a Part-Of-Speech (POS) tag sequence for each sentence in the given datasets.

II. Approach

Two approaches were consider to solve the problem.

- **Hidden Markov Models**: The HMM was trained in a supervised fashion using MLE. POS tags for a sentence were estimated jointly using the Viterbi Algorithm.
- **Conditional Random Fields**: CRFs were also trained in a supervised fashion. Additionally, we also tried using orthographic features such as capitalization, hyphen and common English suffixes.

III. Experiments

The following experiments were tried

- **Only Tokens** In this experiment, only tokens were used as features. For the Atis dataset, the final results were generated by averaging the results from 10 runs in each of which the data was split into the 80-20 proportion using a different random seed. For the WSJ dataset, training was done on section 00 and the testing on section 01.
- **Tokens and Orthographic Features** These experiments were run only for CRFs since HMMs cannot make use of orthographic features. 3 sub-experiments were performed within this case.
  - Using only the prefix (caps, starting with a digit) and hyphen features.
  - Using only the suffix features. The list of suffixes tried were *”ible”, ”able”, ”ness”, ”ment”, ”ing”, ”ogy”, ”ion”, ”tion”, ”ity”, ”est”, ”ies”, ”ful”, ”ive”, ”ous”, ”ic”, ”al”, ”en”, ”es”, ”or”, ”ed”, ”ly”, ”er”, ”y”, ”s”*
  - Using all the orthographic features (prefix, hyphen and suffix).
- **Impact of number of iterations on CRFs and HMMs** The number of training iterations were varied using the *–iterations* parameter and the impact on training accuracy, test accuracy, OOV token accuracy and running time was studied.
- **Impact of more data on CRFs and HMMs** For CRFs 2 cases were tried - training on section 00 and testing on section 01, training on sections 00, 01 and testing on sections 02, 03. For HMMs, we also tried on training on first 3 and testing on next 3 sections, training on first 4 sections and testing on next 4 sections.

In the above experiments, only the ones involving CRFs on the WSJ dataset were run on Condor. The remaining experiments were run on the local system which has a 2.9 GHz Intel core i5 processor with 16Gb of RAM.

IV. Results

- **Only Tokens**

Atis

	CRF	HMM
Training Accuracy	99.87 ± 0.03	88.85 ± 0.31
Testing Accuracy	92.61 ± 0.64	86.62 ± 1.67
OOV Accuracy	25.3 ± 6.9	21.4 ±10.5
Percentage of OOV Tokens	2.97 ± 0.36	2.97 ± 0.36
Running Time (seconds)	123.5 ± 21.6	3.54 ± 0.08

WSJ

	CRF	HMM
Training Accuracy	99.5	86.43
Testing Accuracy	80.76	78.72
OOV Accuracy	47.65	37.87
Percentage of OOV Tokens	15.31	15.31
Running Time (minutes)	77	1.13

- **Tokens and Orthographic Features**

Atis

	Testing Accuracy	OOV Accuracy
Only Tokens	92.61 ± 0.64	25.3 ± 6.9
Prefix and Hyphen	93.23 ± 0.48	31.6 ± 5.3
Suffix	93.32 ± 0.83	39.6 ± 11.1
All Orthographic	94.31 ± 0.75	49.4 ± 9.11

The training accuracy for all was around 99.8% , running time around 125 seconds and 2.97% of tokens were out of vocabulary.

WSJ

	Testing Accuracy	OOV Accuracy
Only Tokens	80.76	47.65
Prefix and Hyphen	83.59	58.85
Suffix	86.11	67.70
All Orthographic	89.41	79.83

The training accuracy for all was around 99.5% , 15.13% of tokens were out of vocabulary. The running time for only tokens was 77 minutes, prefix 72 minutes, suffix 84 minutes and all orthographic features 68 minutes.

- **Impact of number of iterations**

CRFs on Atis

	Train	Test	OOV Acc.	T(sec)
5 iterations	62.23	61.22	13.7	13.03
10 iterations	83.28	79.75	22.4	19.21
20 iterations	98.57	91.36	26.4	31.34
40 iterations	99.87	92.55	25.3	56.63

In the above table Train is training accuracy, Test is testing accuracy, OOV acc. is out of vocabulary accuracy , T is the running time in seconds.
Similar results were also obtained for CRFs on the WSJ Dataset. For HMMs, there wasn't significant variance in performance with number of iterations. Hence, it is not show over here.

- **Impact of more data on CRFs and HMMs**
  The table below shows variation of performance of the 2 approaches upon training it on different number of sections of WSJ dataset. Note that the testing is also done on an equal number of sections.

HMM

	Train	Test	OOV Acc.	T(s)	OOV Perc.
1 sec.	86.43	78.72	37.87	68	15.31
2 sec.	88.89	83.4	39.37	214	11.40
3 sec.	89.68	85.81	40.80	396	8.59
4 sec.	90.56	87.08	41.94	617	7.554

All symbols have the same meaning as the above table. OOV Perc. is percentage of Out of Vocabulary Items.

CRF

	Train	Test	OOV Acc.	T(s)	OOV Perc.
1 sec.	99.50	80.76	47.65	77	15.31
2 sec.	99.45	84.41	50.89	328	11.40

The various terms have the same meaning as the above table except that the time is now measured in minutes.

V. Discussion

- **How does the overall test accuracy of CRF and HMM differ (when using only tokens) and why?** The overall test accuracy is higher for CRFs than HMM both on the Atis Dataset (92.61 vs 86.62) and WSJ Dataset (80.76 vs 78.72). The main reason for this is that CRF is a conditional, discriminative model and only tries to model the probability of the label sequence given the token sequence. HMM on the other hand is generative model which tries to model the joint distribution of token sequence and label sequence. Thus CRF is more suited for the task at hand which is to predict the label sequence given the token sequence whereas HMM unnecessarily tries to model extra things. One also observes that for the larger dataset (WSJ) , the difference is much smaller (2% vs 6% on the smaller dataset). This is because with more amount of data, it becomes much more viable to fit the generative model.
- **How does the test accuracy for OOV items for CRF and HMM differ (when using only tokens) and why?** For Atis dataset, CRF has a 4% higher OOV accuracy (25.4% vs 21.4%) whereas on the WSJ dataset they differ by almost 10% (47.65% vs 37.87%). The reasoning is again the same. CRF being a conditional discriminative model better models the dependencies. So it is able to better utilize the labels on adjacent tokens to predict the label for the OOV token.
- **How does the training accuracy of HMM and CRF differ and why?** The training accuracy of CRFs is almost 100% whereas that for HMM is 88.85% on the Atis dataset and 86.43%. The gap in the training and test accuracy is much larger for CRF than for HMM which indicates that it is overfitting. However, it still performs better than HMM because it is in a sense solving the more appropriate appropriate problem i.e. labelling a given token sequence rather than trying to jointly generate both the token and the label sequence.
- **How does the run time of HMM and CRF differ and why?** The run time for CRFs is much higher than that for HMMs (125.5 seconds vs 3.54 seconds on Atis and 77 minutes vs 1.13 minutes on WSJ). This is primarily because the training (parameter estimation) algorithm for CRFs is much slower. For HMMs, parameter estimation basically involves computing relative frequencies which can be done in 1 pass over the data. On the other hand, parameter estimation for CRFs involves an iterative algorithm (L-BFGS) which involves computing approximate Hessians over the entire data in each iteration.
- **How does adding orthographic features affect the accuracy (both overall and OOV) and runtime of the CRF and why?** Adding orthographic features leads to an improvement in both the OOV accuracy as well as the overall test accuracy. This is because now even for an out of vocabulary token we have some feature (could be prefix, hyphen or suffix) for which we have seen examples in training. Also many times, features like prefixes and suffixes are very discriminative in determining the POS tag. For examples, a caps at the beginning of the word very strongly indicates that is a proper noun and the suffix 'ly' is a strong indicator for an adverb.
- **Which features helped the most? (i.e. try only including some feature types and not others)** The suffix features helped more in comparison to the prefix features. This is many because there are fewer tokens starting with caps, digit or containing a hyphen in comparison to the number of tokens containing some of the popular suffixes like 'ing', 'ed', 's' .
- **What is the impact of number of iterations on the performance of CRF)** The number of iterations required for the convergence of CRF on Atis dataset was around 50. We see, that initially when we increase the number of iterations from 5 to 10 and then from 10 to 20 there is very steep increase in both the testing(from 61% to 80% to 91%) and the training accuracy . By 20 iterations, the accuracy is only 1% less than the accuracy attained at full convergence. On the other hand the time required is almost half of that required for full convergence (30 seconds vs 1 minute).
- **What is the impact of training data size on performance of CRFs and HMMs)** We see that both testing and OOV accuracy increase with increase in training data size for both CRFs and HMMs. We also see that the difference in accuracy between HMM and CRF decreases with increase in data (2% for 1 section vs 1% for 2 sections).