CS388: Assignment-2

Shobhit Chaurasia
Department of Computer Science
UTEID: sc52987

1. INTRODUCTION

In this assignment, the goal is to compare and contrast the performance of HMM and CRF for the task of POS-tagging on atis and wsj dataset. Mallet's implementation of HMM and CRF was used. The first task was to write a parser to convert the dataset into the format accepted by Mallet. Secondly, accuracy specific to OOV words was calculated by building a vocabulary from the training set (using Java HashSet), and recording unseen words in the testing set. Thirdly, numerous experiments were run to analyze different aspects of the performance of HMM and CRF.

2. ORTHOGRAPHIC FEATURES

In addition to the POS-tags already present in the dataset, orthographic features [1] (see Table. 2) were added during the parsing step. CRF makes use of these additional features to improve its estimates. Standard HMM is not expressive enough to incorporate these features itself. I came up with a hack to emulate this, which is discussed later.

3. EXPERIMENTS

Table. 1 summarizes the results of HMM and CRF on atis and wsj corpus. The test accuracy of CRF is consistently higher than HMM. This is expected, because CRF is trained specifically for the classification task (i.e. it models the conditional probability $p(tag|word)$), while HMM attempts to model the full joint distribution $p(tag, word)$ by making stronger assumptions than CRF. A consequence of this is that HMM requires larger training set to achieve performance comparable to CRF.

3.1 Without orthographic features

The difference in the test accuracies of HMM and CRF is more profound on atis than wsj because atis is a fairly small dataset, which might not be sufficient for HMM to learn a good model. On the other hand, wsj is large, thereby better facilitating HMM's data intensive inference of joint distribution, leading to accuracy that is closer to CRF. The same argument holds for training accuracy as well.

HMM, being based on frequency count of words and tags, if forced to resort to smoothing or interpolation technique to handle OOV words. While begin effective, such techniques are a only a heuristic, and are less robust than optimization techniques for learning a weight vector for each feature (like in state transition). In addition to the general effectiveness of CRF, this could be a reason for its higher OOV accuracy.

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

HMM has a significantly lower training time than CRF. This is because supervised training of HMM involves only frequency counts of $(tag, word)$ pairs, while CRF uses optimization techniques like Gradient Descend, L-BFGS etc. to learn near-optimal weights. Such optimization techniques are iterative in fashion, and can take a long time to converge to the optimal solution.

3.2 With orthographic features

As evident from Table. 1, inclusion of orthographic features while training CRF leads to a significant improvement in its performance. This is because orthographic features are indeed indicative of the part of speech of the word (such as words with a first letter capitalized are more likely to be proper nouns). In particular, their inclusion leads to a drastic improvement in OOV accuracy (an increase of almost 25% in wsj), because these features make the Out-of-Vocabulary word appear "less unseen". For example, an OOV word "googling" on its own might not seem related to a known verb "flying" unless the POS-tagger identifies the common suffix "-ing", a feature that it can then utilize to make a more informed prediction.
What is worth noting is that while a hasty thought might lead someone to believe that training a CRF with additional orthographic features might be computationally more expensive, the numbers in Table. 1 tell a different story. The reason is that the additional features only increase the cost of vector arithmetic that is inherent to CRF by a small amount, while at the same time decreasing the training error rate, and thereby facilitating faster convergence of the optimization process.

3.3 Comparison of orthographic features

For the purpose of comparison of the effect of different set of orthographic features, CRF was trained on atis and wsj with one group of features at a time, where the features are grouped according to the groups in Table. 2. The results are shown in Table. 3 and 4. The Symbolic feature and the Plural suffix seem to have the maximum effect on the accuracy, especially on OOV words. The effectiveness of Symbolic feature on OOV words is intuitive; a significant number of OOV words tend to be proper nouns (starting with a capital letter), or words with hyphen (tree-like).

3.4 Increasing training data

Table. 5 shows the accuracy of HMM and CRF on wsj corpus after doubling the amount of training and test data. On comparison with the numbers in Table. 2, we can see that the accuracy of both HMM and CRF increase a couple

Metric	atis			wsj		
	HMM	**CRF**	**CRF + ortho**	**HMM**	**CRF**	**CRF + ortho**
Training accuracy (in %)	88.85	99.88	99.88	86.18	98.57	99.13
Testing accuracy (in %)	86.62	92.61	94.01	78.49	79.36	86.80
OOV accuracy (in %)	21.80	25.75	44.32	37.95	47.60	72.45
OOV percentage (in %)	2.97	2.97	2.97	15.33	15.33	15.33
Running time (in sec)	8	96	88	80	5607	4640

Table 1: Comparison of HMM, CRF, and CRF with orthographic features on atis and wsj.

Category	Feature
Symbolic	hyphen, starts with a digit, starts with caps, all caps
Noun suffixes	-ance, -er, -ism, -ist, -ment, -ness, -ship, -sion
Verb suffixes	-ize, -ise
Adjective suffixes	-able, -ible, -al, -esque, -ful, -ic, -ical, -ous, -ish, -ive, -less
Plural	-s

Table 2: Orthographic features used.

Feature	Train	Test	OOV
None	99.88	92.61	25.75
Noun	99.83	92.87	33.33
Adj.	99.83	93.00	33.33
Symbolic	99.83	99.22	41.67
Plural	99.83	93.57	41.67
All	99.88	94.01	44.32

Table 3: Comparison of different orthographic features on atis. Each row represents the effect of adding only that set of orthographic features.

of points, though at the cost of a 3-4 times increase in their training time.

3.5 Effect of number of iterations

HMM's inference procedure, being based on frequency counts, is not iterative. However, CRF uses an iterative optimization procedure to learn optimal weights. Fig. 1 shows the plot of accuracy of CRF on atis and wsj as the number of iterations are varied. As expected, all three accuracy metrics show a general increasing trend as the number of iterations are increased, though accuracy on OOV words seems to wander a bit.

3.6 Adding features directly to POS-tags

HMM does not directly support additional features. One way to emulate feature addition is to modify the POS-tag set by including features in the tag itself. For example, "Flying" is tagged with `VB-caps-ing`, instead of `VB` with features `-caps, -ing`. Table. 6 summarizes the result of running HMM and CRF on this modified POS-tagged dataset. The accuracy of both HMM and CRF drop because the modified POS-tag set becomes huge (282 POS-tags on wsj, up from 45) and brings about data sparsity. Moreover, in CRF, not only does the modified tag-set *not* emulate additional features, it makes inference harder. CRF on wsj with modified tag-set had not converged in about 13.5 hours, having completed only 19 iterations.

Feature	Train	Test	OOV
None	98.57	79.36	47.60
Noun	98.59	79.73	48.21
Adj.	98.60	79.80	48.33
Symbolic	99.13	83.22	58.80
Plural	98.68	82.73	58.24
All	99.13	86.80	72.45

Table 4: Comparison of different orthographic features on wsj. Each row represents the effect of adding only that set of orthographic features.

Metric	HMM	CRF	
		w/o ortho	w/ ortho
Training accuracy (in %)	88.70	99.40	99.39
Testing accuracy (in %)	83.35	84.22	89.38
OOV accuracy (in %)	39.57	50.03	73.34
OOV percentage (in %)	11.40	11.40	11.40
Running time (in sec)	221	21660	18176

Table 5: HMM, CRF, and CRF with orthographic features trained and tested on two sections of wsj.

Metric	HMM		CRF
	atis	wsj	atis
Training accuracy (in %)	86.66	77.47	99.89
Testing accuracy (in %)	82.43	70.66	91.15
OOV accuracy (in %)	14.93	27.66	18.21
OOV percentage (in %)	2.97	15.33	2.97
Running time (in sec)	21	2670	490

Table 6: HMM, CRF on dataset with modified POS-tags. CRF on wsj had not converged in 13.5 hours.

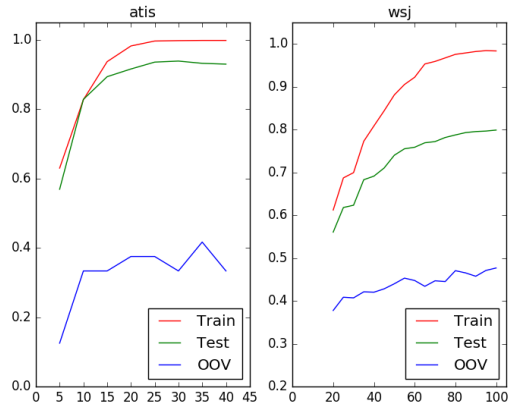

Figure 1: Plot of of CRF's accuracy as a function of the number of iterations.