Comparative Analysis of Hidden Markov Model and Conditional Random Fields for Part of Speech Tagging

Harshal Priyadarshi[*]

University of Texas at Austin
harshal@cs.utexas.edu

Abstract

Tagging each word of a sentence with a part-of-speech based on its occurance in the sentence is called Part of Speech (POS) tagging. The problem is to efficiently do the POS Tagging of Linguistic corpuses. A small ATIS corpus (consisting of airline booking converstation), and a large WSJ corpus (Wall street journal article collection) were used to test the efficiency of two proposed algorithms, Hidden Markov Model(HMM), a generative model; and Conditional Random Fields(CRF), a discriminative model.HMM is proposed to perform relatively poor when compared to CRF, because it is a more general model, that uses the joint probability distribution to train the conditional POS Tagging model, while CRF, developed solely for conditional analysis performed way better because of its dedicated scope to the underlying problem.However CRF takes a lot more time to run as compared to HMM, because of the hard optimization problem. A set of orthographic features were used to train the CRF to get a drastic improvement in the accuracy over Out-of-Vocabulary(OOV) tokens.

I. Introduction

Part of Speech Tagging is an application to understand the syntactic meaning of a sentence. We will be using two methods, as mentioned for doing the POS tagging:

Hidden Markov Model

It is a generative model, which is used to model a series of discrete symbols, for instance a language. It takes as input language tokens (Eg- Words, Sentence) and outputs the underlying state of the symbol(Eg- POS Tag). However it does so by explicitly obtaining the joint distribution of the observation and labels. Thus, it can now model any conditional distribution, including the desired distribution of the POS tags given the sentence observation. However, it can also do the opposite, that given the tags, find the most likely sentence, which is irrelevant in our tagging problem. HMM, though is a jack of all trades, will be shown to be master of none.

Conditional Random Field

This is a recent alternative to the HMM, and is rather a discriminative approach,i.e. it can only model the conditional distribution of the POS Tags, given the input tokens.It is a one-trick pony. It uses a parameter (tokens and vocabulary tokens) based learning, which are tuned over the training data, by performing maximum log likelihood optimization on Logistic Regression model, using L-BFGS algorithm(Limited Memory Broyden-Fletcher-Goldfarb-Shanno).

II. Experiments

The following sub-categories of experimentation were performed:

I. ATIS

The results obtained after averaging the performance of CRF and HMM on ATIS dataset over 10 runs, are as shown below,

Table 1: *ATIS data - 10 splits avergage*

		Accuracy %			
Algo	Time(s)	Train	Test	OOV	OOV/Vocab
HMM	13	88.87	88.86	22.46	2.93%
CRF	126	99.86	92.72	28.71	3.12%

II. ATIS-Additional Features

For these experiments 5 additional features - prefix, suffix, isCapital, hasHyphen,beginsWithNumber were added. First all features were used to train the CRF, then all but one was uses, to know which feature played the significant role in performance improvement. Table 2 represents the results,

Table 2: *ATIS data - CRF model - 10 splits average*

		Accuracy %			
Features	Time(s)	Train	Test	OOV	OOV/Vocab
all_features	91	99.86	93.61	47.19	3.21%
No_prefix	89	99.86	93.58	48.08	3.17%
No_suffix	113	99.86	93.44	39.43	3.09%
No_capital	107	99.84	92.81	40.89	2.99%
No_hyphen	97	99.86	93.61	41	3.05%
No_number	94	99.86	93.62	47.19	3.25%

III. WSJ

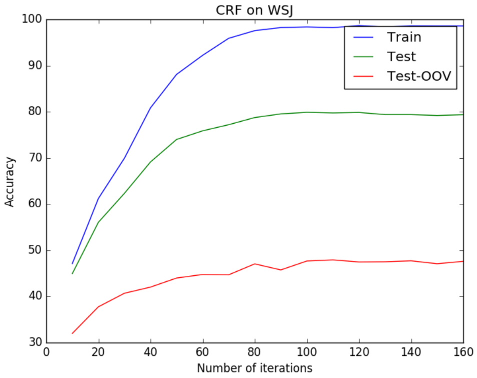
The HMM and CRF was applied to the big WSJ model with different amount of training and testing data, to check the performance w.r.t to the training size. The results are shown in Table 3. Algo(a/a) ← $(No.of TrainingSections/No.of TestingSections)$

[*] Masters in Computer Science UT Austin, Fall 2015

Table 3: *WSJ data*

		Accuracy %			
Algo(Train/Test)	Time(s)	Train	Test	OOV	OOV/Voc
HMM(1/1)	79	86.17	78.48	37.94	15.33%
HMM(2/2)	233	88.70	83.33	39.57	11.4%
HMM(4/4)	856	90.36	86.89	41.41	9.21%
HMM(8/8)	1543	92.22	90.08	43.31	7.37%
HMM(12/12)	3391	93.03	91.17	44.70	5.62%
CRF(1/1)	5612	98.57	79.36	47.60	15.33%
CRF(2/2)	28800	99.40	84.22	50.02	11.4%

Also the tests were done for the base case of CRF(1/1) for different number of iterations to dynamically show that the algorithm is converging fairly quickly. The result is below:



IV. WSJ- Additional Features

The same experiment as of part II was done for the WSJ data using CRF(1/1), whose results are in Table 4.

Table 4: *WSJ data - CRF model*

		Accuracy %			
Features	Time(s)	Train	Test	OOV	OOV/Vocab
all_features	4576	98.75	87.36	76.13	15.33%
No_prefix	4432	98.75	87.66	76.58	15.33%
No_suffix	4965	96.98	81.76	58.74	15.33%
No_capital	4721	93.08	81.59	67.50	15.33%
No_hyphen	4591	98.73	87.10	75	15.33%
No_number	4638	98.74	86.82	74.34	15.33%

III. Discussion

I. Overall Test Accuracy

Acording to Table 1 and 3, **for CRF**, overall test accuracy is around 93% on ATIS data and around to 79 - 84% on WSJ data. **For HMM**, it is around 89% on ATIS, and 78 - 91% on WSJ depending on the training size.
For the same training size, CRF performs better than HMM, which is justified because CRF is made for the sole purpose of modelling conditional distribution, and does it better than HMM, which uses joint distribution to get conditional.

II. OOV Test Accuracy

OOV test accuracy of **CRF** (28.71% on ATIS, and 47.60 - 50% on WSJ) is slightly better than the **HMM** model(22.46% on ATIS, and 37.94 - 44.70% for WSJ)
OOV test accuracy is very less compared to the seen tokens for both models, because since OOV tokens were absent during training, their observation probability is very less. However, CRF being a more specific, discriminative model, still performs better than more general HMM.

III. Training Accuracy

The training accuracy for CRF is close to 100 % for both ATIS and WSJ dataset.While, for HMM, it is around 89% on ATIS and 86 - 93% on WSJ. However, the difference between training and test accuracy is more pronounced for CRF than HMM, hinting towards overfitting.
This was expected, as CRF uses L-BFGS algorithm that maximizes the conditional log-likelihood of the training data, and maximum-log-likelihood training, is known to overfit.

IV. Runtime

HMM converged in about 84 iterations, while CRF took about 150 iterations to converge for WSJ(1/1). However, CRF takes 10 times more time on ATIS and around 70 - 123 times on WSJ compared to HMM time.
This is justified, because L-BFGS is a very complex optimization process, and a lot of parameter weights are to be updated. CRF model has poor scalability as well. With doubling of WSJ training data, the convergence time got 5 fold increment.

V. Orthographic Feature - Runtime and OOV

With addition of orthographic features, the runtime decreased (126 → 91 for ATIS; 5612 → 4576 for WSJ) significantly. This is justified, as adding more features provide more identifying information and thus training is faster.
The OOV accuracy improved drastically(28% → 47% for ATIS; 47% → 76% for WSJ). This is justified as now OOV words can be identified using these orthographic features, which have been trained by both the models. Eg. -ly suffix hints adverb, Capital words hint pronouns, etc.

VI. Best Features

The absence of best features deteriorates the OOV accuracy in Table 2 and 4.On both **ATIS** and **WSJ**, the best features were, **Suffix and Capital features**.
Suffix in general is a good feature, as it identifies with adjective and verb, which are common occurances in any sentence. The **isCapital** was good for ATIS, because most of the flight booking chat, include the Source and Destinations which are proper nouns.For WSJ most of data targets a Person, Place or Object each of which is a Noun. Capital identifies with Proper nouns.